Computational strategies to investigate the genetic cause of human eye disease

Wesley Andrew Goar

University of Iowa
COMPUTATIONAL STRATEGIES TO INVESTIGATE THE GENETIC CAUSE OF HUMAN EYE DISEASE

by

Wesley Andrew Goar

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 2019

Thesis Supervisors: Professor Todd Scheetz Professor Val Sheffield
To my wife, Jenna, and my parents for always supporting me through my educational journey
ACKNOWLEDGEMENTS

I have been blessed to have a wonderful graduate education at the University of Iowa. Drs. Todd Scheetz and Val Sheffield have been amazing mentors that have taught me so much more than just how to be a good scientist. I would also like to thank the members of my thesis committee Drs. Ben Darbro, Rob Mullins, and Elliott Sohn, for their advice and teaching during my graduate studies. The Institute for Vision Research has been a phenomenal place to work with many collaborators including Drs. Michael Anderson, John Fingert, and Ed Stone.

I had the privilege of working in two amazing labs filled with great people that made my graduate studies enjoyable. I am especially grateful to Dr. Adam DeLuca for all his help regarding any next-generation sequencing questions I had, the time we spent together optimizing the latest sequencing pipeline, and all of the random factoids I learned from him during our chats. I’d also like to thank Dr. Scott Whitmore who was an invaluable reference for information on R information and RNA-sequencing. My friends in both labs provided me with many laughs and support during my graduate work.

I am especially thankful to my wonderful family for their support throughout my life. My parents taught me how to work hard and finish what I set out to do. They have always believed in me and loved me unconditionally. I could not have asked for better parents to help me reach my goals as a scientist and a human being.

Lastly, I would like to thank my beautiful wife, Jenna, for her love, patience, encouragement, and support. She is my best friend, my favorite person, my rock, and my companion.
ABSTRACT

It is estimated that 4000 genetic diseases/syndromes affect humans with one third of these diseases involving the eye. Many eye disorders, such as age-related macular degeneration that affects an estimated 170 million elderly adults worldwide, are associated with genetic variants. Since the conception of the human genome project we have learned a great deal about the genetic make-up of the human race and have identified over ~20,000 genes. Over 270 of these genes have been implicated in retinal diseases alone with many more genes involved in other forms of ocular disease. Though we have made a great deal of progress in understanding the genetics of eye disease, there remain many eye diseases with significant evidence of genetic components for which a disease-causing gene has not been identified.

In my thesis research, I utilized computational tools and strategies to analyze microarrays and whole-exome sequencing to investigate the genetic causes of three different eye diseases. First, I utilized a combination of familial analyses and whole-exome sequencing to study the genetic cause of Keratoconus, a progressive cornea abnormality that can lead to distorted vision and light sensitivity. Second, I analyzed three different cohorts of patients with Bardet-Biedl syndrome (BBS), a syndromic retinopathy leading to blindness, using whole-exome sequencing to identify both known and novel genetic causes of BBS. Finally, I performed the largest whole-exome sequencing study at the time for Pigment Dispersion Syndrome (PDS), a disorder associated with glaucoma, and identified variants within previously established candidate genes and a novel candidate gene that is now the subject of further scientific investigation.

By using computational tools and strategies in tandem with high-quality bench research performed by fellow lab members, we have identified both candidate and known eye disease-causing genes/mutations and furthered the goal to cure blindness.
A poll performed by the Association for Research in Vision and Ophthalmology showed that Americans think that of the five senses the loss of vision would have the largest impact on their lives. Further studies have confirmed the belief of those polled by showing a decrease in the quality of life within people affected by vision loss. Millions of people worldwide are affected by eye diseases that are associated with, or directly caused by, genetic factors.

I used computational tools and strategies to investigate the genetic cause of three different eye diseases; 1) keratoconus, a progressive cornea abnormality that can lead to distorted vision and light sensitivity 2) Bardet-Biedl syndrome, a syndromic retinopathy leading to blindness and 3) pigment dispersion syndrome, a disorder associated with glaucoma. The overarching goal was to analyze the DNA of patients with these diseases to identify changes in their genetic code that could be causing the disease presentation. Since there are hundreds of thousands of such changes within each individual, it is crucial to apply strategies to remove the majority of the genetic changes that are benign and identify the changes that need further investigation. Using a combination of state-of-the-art high-throughput sequencing technology, genetic databases, familial information, clinical diagnoses, and bioinformatic tools, I was able to identify new regions of interest for three families with keratoconus. I also identified both known and novel genetic causes of disease in Bardet-Biedl syndrome patients. Finally, I identified alterations in known candidate genes and a novel candidate gene for pigment dispersion syndrome within the whole exome sequencing cohort to date.
# TABLE OF CONTENTS

LIST OF TABLES ......................................................................................................................... xi

LIST OF FIGURES ..................................................................................................................... xiii

CHAPTER 1: INTRODUCTION .................................................................................................... 1

CHAPTER 2: COMPUTATIONAL TECHNOLOGIES ................................................................. 6

- Introduction ......................................................................................................................... 6
- Linkage Analysis ................................................................................................................. 6
  - Software ......................................................................................................................... 8
- Regions of Homozygosity ................................................................................................. 9
  - Software ....................................................................................................................... 10
- Linkage Disequilibrium ................................................................................................. 10
- Whole Exome Sequencing .............................................................................................. 11
  - Sequence Generation .................................................................................................... 12
  - Sequence Alignment and Processing .......................................................................... 12
  - Calling and Annotating Variants ............................................................................... 13
  - Variant Filtering ......................................................................................................... 13
  - Experimental Confirmation and Validation .............................................................. 16
  - Software and Resources ............................................................................................. 16
- Copy Number Variant Calling ......................................................................................... 19
  - CoNIFER ..................................................................................................................... 19
  - CLAMMS ..................................................................................................................... 20
- Variant Pathogenicity Prediction .................................................................................... 20
  - PolyPhen2 .................................................................................................................. 21
  - SIFT ............................................................................................................................ 21
  - CADD ......................................................................................................................... 21
CHAPTER 3: IDENTIFYING THE GENETIC CAUSE OF KERATOCONUS IN THREE BEDOUIN FAMILIES

Abstract

Introduction

Materials and Methods

Collection of Bedouin samples

Affymetrix Genome-Wide SNP 5.0 Array

Identification of regions consistent with segregation

Identification of Runs of Homozygosity

Whole Exome Sequencing

Sequence Alignment and Processing

Calling and Annotating Variants

Copy Number Variant (CNV) analysis

Variant Filtering

Variant Prioritization

Experimental Confirmation

Results

Variants in KC-associated Genes (Tier 1)

Regions Consistent with Segregation

Plausible Variants in Shared Regions Consistent with Segregation (Tier 2)

Plausible Variants in Shared Regions Consistent with Segregation in Two Families (Tier 3)
CHAPTER 4: IDENTIFYING THE GENETIC CAUSE OF BARDET-BIEDEL SYNDROME IN THREE DIFFERENT COHORTS OF PATIENTS THROUGH WHOLE EXOME SEQUENCING.

Abstract ................................................................................................................................50

Introduction ..............................................................................................................................51

Materials and Methods .........................................................................................................57

DNA Sequencing ................................................................................................................57

Sequence Alignment and Processing ................................................................................57

Calling and Annotating Variants ......................................................................................57

Variant Filtering ................................................................................................................58

Variant Prioritization ........................................................................................................58

Experimental Confirmation .............................................................................................59

Project 1: Reanalysis of 26 BBS patients ..............................................................................59

Results ....................................................................................................................................59

Whole Exome Sequencing Results ................................................................................59

Copy Number Variant Results ..........................................................................................62

Discussion .............................................................................................................................63

Complete BBS Genotypes ................................................................................................63

Incomplete BBS Genotypes ..............................................................................................64

Rare Mutations in Genes with BBS-associated Phenotypes .............................................65
LIST OF TABLES

Table 1. Genes associated with Keratoconus...................................................................................... 32

Table 2. Variants identified in KC-associated genes. ........................................................................ 38

Table 3. Genomic space and number of genes identified through linkage analysis. Depicted above are the number of mega bases and percentage of the total genome in each family found to be consistent with segregation. The table also shows the number of genes in each of those regions. Through linkage analysis we were able to substantially reduce the number of genes and fraction of the genome we were inspecting to greatly increase our power when analyzing data........................................................................................................................................... 39

Table 4. Progression in filtering of variants within the three exomes. By utilizing our linkage analysis results in tandem with our whole exome sequencing filtering, we were able to reduce the total number of plausible variants significantly. Filtering includes removal of common variants identified in databases such as gnomAD and 1000 Genomes................................................................. 40

Table 5. Number of plausible variants in regions consistent with segregation between two families. Shown are the number of variants identified in regions consistent with segregation between 2 families. We then identified specific variants shared between 2 families (row 3) in those regions for manual inspection using IGV and variant effect prediction with PolyPhen2, SIFT, and CADD. Following filtering, we deemed there to be no shared plausible variants. ......................................................................................................................... 41

Table 6. Shared runs of homozygosity between families. Depicted above are the runs of homozygosity that were shared between at least 2 of the families............................................. 41

Table 7. Tier 5 variants. ................................................................................................................ 42

Table 8. Conservation of Amino Acids in COL4A1 ........................................................................ 44

Table 9. Conservation of Amino Acids in COL5A1 ........................................................................ 46

Table 10. Conservation of Amino Acids in MPDZ. ..................................................................... 48

Table 11. BBS genes and their HGNC symbols ............................................................................ 51

Table 12. BBS10 and BBS7 variants within previously unsolved cases........................................ 60

Table 13. Reported variants from diseases with shared BBS phenotypes................................. 61

Table 14. Overlapping phenotypic features between BBS and other diseases........................... 62

Table 15. Variant information from Mexican BBS patient. ........................................................... 69

Table 16. Variants in BBS genes. ................................................................................................. 73
Table 17. Mutations in \( KCNV2 \) within our Puerto Rican BBS cohort. ................................. 75
Table 18. Study cohort statistics. .................................................................................................. 86
Table 19. Primary analysis results. ............................................................................................... 87
Table 20. Clinical features of patients with variants found in the primary analysis.................. 88
Table 21. Tier 2 analysis results. .................................................................................................. 91
Table 22. Top 10 genes from Tier 3 analysis (SKAT-O) ............................................................. 92
Table 23. Reported \( MRAP \) variants. ............................................................................................. 93
Table 24. Intrinsic component of the plasma membrane................................................................. 95
Table 25. Clinical features of patients with \( MRAP \) mutations...................................................... 99
LIST OF FIGURES

Figure 1. Two generation pedigree segregating an autosomal dominant trait. Alleles at 2 marker loci are designated A and B are shown. Squares indicate males; circles, females; open symbols, normal phenotype; and solid symbols, disease phenotype. ............................................. 7

Figure 2. Depiction of the 2nd half of the whole exome sequencing pipeline. In this depiction, Haplotype Caller creates 3 slice gVCFs for samples 1-3. All slice 1 gVCFs are combined into a single batch 1 gVCF and are joint genotyped using Genotype gVCFs. Following the genotyping all the batch files are merged into a master VCF and then demultiplexed back into separate VCF files for each sample (1-3) in preparation for annotation. .............................................. 15

Figure 3. Visualization of a Burden-based analysis................................................................. 23

Figure 4. Visualization of a SKAT-based analysis................................................................. 24

Figure 5. Venn Diagram depicting the overlap between the candidate genes and a group of genes from a specific pathway................................................................. 26

Figure 6. Hypergeometric test formula.................................................................................... 27

Figure 7. Difference between a normal eye and a Keratoconus eye........................................ 30

Figure 8. Pedigree of Family 1. ............................................................................................ 32

Figure 9. Pedigree of Family 2. ............................................................................................ 33

Figure 10. Pedigree of Family 3. ........................................................................................... 33

Figure 11. Reported CNV in two families. Depth of coverage is shown for all three KC samples for which exome data was obtained. Based upon normalized coverage levels, the sample from Family 3 is clearly missing most or all of exons 7 and 8 of gene, and the sample from Family 1 appears to have substantially increased coverage, consistent with a duplication................................................................................................................................. 43

Figure 12. Genotyping results for C > T variant in COL4A1 within Family 2. The variant was identified in the female labeled Exome but was not found in any of the other affected or unaffected members of the family. ........................................................................................................................................... 45

Figure 13. Segregation of the C > T variant in COL5A1 within Family 2. All family members with DNA available for testing presented with this variant.................................................. 47

Figure 14. Family genotypes for the MPDZ variant within Family 3. .................................... 48

Figure 15. Fundus Photographs of BBS Patient. ................................................................. 53

Figure 16. Homozygous deletion of exons 3 and 4 in two BBS patients. Shown is the total read coverage for a region in BBS4. The top and bottom tracks shown are BBS patients with
homozygous deletion of exons 3 and 4 in \textit{BBS4} as seen by the lack of reads in that section. The middle track is another BBS patient without the deletion.

Figure 17. Visualization of variant phase using IGV. Using IGV, we were able to show that each parent carried one of the mutations before Sanger sequencing confirmation. Each sequencing read shown that overlaps both variants location provides evidence supporting only one of the variants.

Figure 18. Pedigree depicting the segregation of the two BBS3 mutations. Sanger sequencing confirmed the carrier status for the L177F allele in the unaffected mother and the G167R in the unaffected father of the proband. Segregation through the family was fully confirmed when the unaffected brother of the proband was confirmed to carry only the L177F allele.

Figure 19. Apparent Homozygous deletion of exons 7 & 8 in three Puerto Rican BBS samples. IGV depiction of homozygous deletion of 2 exons in PR BBS samples. This deletion was not identified by CLAMMS in any of the other PR samples or non-BBS samples that were from the same NGS batch. Sanger sequencing of the region failed to confirm the deletion.

Figure 20. Overview of SKAT-O data preparation pipeline. Steps taken to prepare the data for SKAT-O analysis.

Figure 21. Model of MRAP function. Shown is a schematic of MRAP protein dimerizing and then binding to MC2R within the Endoplasmic Reticulum. The complex then is transported to the Golgi apparatus to be glycosylated (green circles) and then to the plasma membrane where it can interact with ACTH.
CHAPTER 1: INTRODUCTION

According to the World Health Organization, approximately 1.3 billion people worldwide are affected by some form of visual impairment. A recent study estimates that 217 million people have moderate to severe visual impairment and another 36 million people are blind\(^1\). Vision loss has a significant impact on those affected, their families, and society as a whole. Studies have shown a decrease in the quality of life for people with vision loss including loss of independence, decreased mobility, increased chance of injury, decreased mental health, decreased social function, and a lower level of education\(^2\text{-}^5\).

Along with a decrease in general quality of life, there is a significant financial cost associated with vision loss. In 2013 it was estimated that the overall financial impact of vision loss in the United States of America was approximately 139 billion dollars\(^6\). While a significant portion of vision loss worldwide is considered to be avoidable, many people are afflicted with genetic forms of eye disease such as Stargardt disease and retinitis pigmentosa. Recently, Stone \textit{et al.} were able to make population estimations regarding certain inherited retinal diseases within the United States\(^7\). They estimated that 1/10,000 people suffer from Stargardt disease caused by mutations in a single gene, \textit{ABCA4}, which is roughly 32,500 people in the United States\(^7\). There are many other genes, both known and unknown, that cause syndromic and non-syndromic eye disease. Additional research is needed to identify and develop treatments to eradicate blindness in those affected by these diseases.

Genes are segments of DNA that are the blueprints for proteins, which build, regulate, and maintain the body. When the code for a gene is compared against the consensus sequence and
sequence differences are identified, such differences are known as DNA variants. Not all DNA variants are deleterious, they can affect the protein created from the gene in positive, neutral or negative ways. Every human has millions of variants when compared to one another, or to the reference human genome. The Human Genome Project was performed in order to cost effectively identify all human genes and advance the understanding of genetics through the completion of multiple goals. First, the human genome project set out to create a genetic map of the genome by identifying markers that are important to detect genetic variation among individuals. Second, the project had a goal to make a physical map of the human genome by using short DNA sequences that are unique known as sequence-tagged sites. Third, the project sought to sequence 95% of the gene-containing portion of the human genome with an accuracy of 99.99% and was able to sequence 99% of the gene-containing portion of the human genome. Fourth, the project had a goal to decrease the cost of sequencing and increase the speed and efficiency of sequencing. Fifth, the project wanted to identify at least 100,000 mapped human single nucleotide polymorphisms (SNPs), but ended up identifying 3.7 million SNPs. Sixth, it had a goal to identify as many full-length human cDNAs as possible. In March of 2003 the project reported the identification of 15,000 full-length human cDNAs. Seventh, complete the genome sequencing for *E. coli*, *S. cerevisae*, *C. elegans* and *D. melanogaster* to help further medical research. This goal was completed by April of 2003 with drafts completed for another four model organisms. Eighth, promote the development of new genomic-scale technologies to obtain and analyze data. Finally, with the advances in genetics having an impact on individuals and society, the project sought to understand the consequences of genomic research through the development of the Ethical, Legal, and Social Implications (ELSI) program. Thousands of individuals DNA were sequenced in order to develop a reference
genome showing which nucleotide was at a specific location in the majority of people sequenced. Today, we continue to update the human reference genome as more and more human genomes are sequenced and databases established to provide us with a more accurate depiction of the variance within the human genome\textsuperscript{16}.

Humans have 46 chromosomes that contain almost all of their genes, the other 37 genes are located within the mitochondrial DNA passed to a child by their mother. When a child is conceived it receives half of its DNA from each of its parents (23 chromosomes). Two of the total 46 chromosomes are known as the sex (X and Y) chromosome and determine the sex of the individual. Females have two copies of the X chromosome while males have one copy of the X chromosome and one copy of the Y chromosome. The inheritance of genes (traits) was first identified by Gregor Mendel when studying the inheritance of different traits within pea plants. Though Mendel didn’t know it at the time he was studying variants within individual genes that caused specific traits within the pea plant. After performing his experiments, Mendel came up with three laws or principles. Mendel’s first law, the law of segregation, states that genes come in pairs and each parent will pass one of the two copies of each gene (known as an allele) to an offspring. Which allele is passed from parent to offspring is up to chance (random). Mendel’s second law, the law of independent assortment, states that different pairs of alleles are passed onto offspring independently of each other. The inheritance of genes at a given location in the genome does not influence the inheritance of genes at another location. Finally, the law of dominance states that when mating occurs between two organisms with a different trait each offspring will exhibit the trait of one parent only. If a dominant factor is present in the individual, the dominant trait will manifest in the offspring in the heterozygous state. A recessive trait will
only manifest if the offspring inherits the recessive allele for a given gene from both parents in the homozygous state.

Today, diseases that are caused by variants within a single gene are referred to as Mendelian disorders or monogenetic disorders. There are three main patterns of inheritance within monogenetic disorders; 1) autosomal recessive; 2) autosomal dominant and 3) sex-linked (or X-linked). Autosomal recessive disorders are caused when both copies of DNA inherited from the parents are required to have a disease-causing mutation (variant that is known to be harmful) before the disease/trait is expressed. Autosomal dominant disorders are caused when a single copy of the DNA carrying a disease-causing mutation is sufficient to cause disease. Diseases that follow a sex-linked pattern of inheritance have mutations on the X or Y chromosome. Diseases such as Bardet-Biedl syndrome\textsuperscript{17}, Stargardt disease\textsuperscript{18}, sickle-cell anemia\textsuperscript{19}, and cystic fibrosis\textsuperscript{20} are examples of monogenetic disorders.

However, not all genetic diseases are inherited in a Mendelian fashion. More complicated patterns of inheritance include complex (multifactorial) disorders and chromosomal disorders. Complex disorders are caused by a combination of less detrimental variants in genes in tandem with environmental factors. Examples of complex disorders are heart disease\textsuperscript{21}, diabetes\textsuperscript{22}, some forms of glaucoma\textsuperscript{23-25} and age-related macular degeneration\textsuperscript{26-28}. Chromosomal disorders are caused by an excess or loss of genes due to structural changes within chromosomes. The most well-known example of a chromosomal disorder is Down syndrome, which is caused by an extra copy of chromosome 21\textsuperscript{29}. Some chromosomal disorders involve only a portion of a chromosome being duplicated or deleted. These disorders are known as microduplication and
microdeletion syndromes. With the advent of novel methods, such as the use of chromosomal microarrays, the discovery of such syndromes has greatly increased.

Inherited eye diseases can fall into any of the three classifications - Mendelian, complex or chromosomal. Similarly, each of these diseases may be syndromic (systemic) or non-syndromic (isolated to the eye). Eye disease is typically separated by the section of the eye that is being affected such as the retina (photoreceptors), ganglion cells and optic nerve, or the cornea. The first gene known to cause retinitis pigmentosa was identified in 1990 and we now know over 270 genes implicated in monogenic retinal disease alone. The ability to identify the genetic cause of eye disease has greatly improved since the 1990s, with Stone et al. reporting that they were able to identify 76% of the disease-causing genotypes within a cohort of 1000 consecutive families with retinal disease seen by a single clinician.

However, while there has been a considerable amount of effort and progress in studying Mendelian, complex, and chromosomal eye disorders, there is still much to learn. With the development of next-generation sequencing and the field of bioinformatics we have gained many tools and techniques to study genetic disease. In this thesis, I will discuss the computational strategies that my collaborators and I have made to further understand the genetics of three different eye diseases; 1) keratoconus, 2) Bardet-Biedl syndrome, and 3) pigment dispersion syndrome.
CHAPTER 2: COMPUTATIONAL TECHNOLOGIES

Introduction
Identifying the causes of genetic eye disease requires both traditional laboratory bench research, as well as computational technologies and strategies. Whether the disease-causing genes are known or yet to be discovered, it is critical to have a strong depth of knowledge concerning the current technologies available to provide the best possible chance of accurately identifying the genetic cause of disease for each individual patient. In this chapter, I present the computational technologies and analytic strategies employed during my thesis work, discuss the proper utilization of these technologies.

Linkage Analysis
Mendelian diseases are caused by mutations in DNA that are inherited by a child from their parent(s). Linkage analysis is a strategy used to identify the genomic region containing a mutation segregating within a family, and to refine the chromosomal location of the disease-causing gene. This strategy is based on the concept of genetic linkage, in which some traits and genetic markers do not segregate independently, i.e. they are encoded by DNA sequences that are very close to each other and are therefore frequently inherited together, resulting in an exception to Mendel’s second law. Genetic linkage was first described by Thomas Hunt Morgan when he described traits in Drosophila that did not segregate independently. The unit of measurement for linkage is called the centimorgan, in honor of Morgan. When genes or DNA markers are close enough to one another they are rarely separated during meiosis, and hence they segregate together more often than would be predicted by chance, and thus are commonly inherited together from parent to offspring. Thus the polymorphisms in the human genome can
be used as genetic markers to identify genetic regions that are linked with a disease allele within individual families. This concept is visualized in Figure 1 with two genetic markers, A and B, each of which have two possible alleles – A1 and A2 for marker A, and B1 and B2 for marker B. Because humans are diploid, each person has two copies of each autosomal chromosome, and hence two copies of each gene. As shown in Figure 1, markers A and B are genetically linked, that is they are physically close enough to each other that they do not segregate independently during meiosis. Thus, the alleles of these DNA polymorphisms are inherited together more often than would be predicted by chance – A2 with B1, and A1 with B2. In addition, Figure 1 shows that the disease or phenotype annotated on the pedigree, shaded in black, co-segregates with both the A2 and B1 alleles, in an autosomal dominant pattern of inheritance.

Figure 1. Two generation pedigree segregating an autosomal dominant trait. Alleles at 2 marker loci are designated A and B are shown. Squares indicate males; circles, females; open symbols, normal phenotype; and solid symbols, disease phenotype.
To determine whether a polymorphism is segregating with a disease phenotype more frequently than would occur by chance, statistical analysis is required. The logarithm of the odds (LOD) score is used as the measure the likelihood of linkage\textsuperscript{32,33}. A LOD score of 3.0 or higher is accepted as evidence for linkage and a LOD score of lower than -2 is accepted as evidence against linkage. LOD score analysis is highly dependent on correct definition of the model parameters, which themselves rely upon accurate clinical data to assign affectation status, mode of inheritance, disease gene frequency, and penetrance. In order to identify possible regions of linkage across the genome, companies have made genotyping chips with hundreds of thousands of single nucleotide polymorphisms (SNP) spanning the genome.

One such chip is the Affymetrix Genome-Wide SNP 5.0 Array (Affymetrix; Santa Clara, California) that contains nucleic acid probes designed to genotype 500,568 single nucleotide polymorphisms that can be used to identify regions linked to a disease phenotype. Genome-wide SNP genotypes can be obtained using the BRLMM-P algorithm within the Affymetrix Power Tools suite (Affymetrix (2009) Affymetrix Power Tools). The genotypes can then be used in a parametric linkage analysis using programs such as MERLIN\textsuperscript{34}.

Software

**Affymetrix Power Tools**

Affymetrix Power Tools (APT) is a software package used for the analysis and data handling of GeneChip arrays\textsuperscript{35}. APT can be used to analyze gene expression on a standard microarray or perform genotype calling on genotyping arrays. It provides a wide depth of customizable
parameters to be used to fit the needs of the user, including support for a wide variety of genotyping arrays developed by Affymetrix.

**MERLIN**

MERLIN is a software package used for pedigree analysis\textsuperscript{34}. MERLIN provides analysis capabilities for a wide variety of genetic analyses, including haplotyping, genotype error detection, linkage analysis, and genome-wide association analysis.

**Regions of Homozygosity**

Homozygosity by decent mapping is a method used to discover loci containing genes with an autosomal recessive pattern of inheritance within consanguineous families\textsuperscript{36}. This strategy is designed to identify regions of shared homozygosity between affected family members\textsuperscript{37} and has been used to identify numerous disease-causing genes\textsuperscript{38-42}. In particular, our lab has mapped and eventually identified multiple Bardet-Biedl syndrome (BBS) genes through the use of homozygosity by descent mapping\textsuperscript{43-45}. In 1995, the Sheffield lab was able to utilize the principles of homozygosity by descent to develop a novel streamlined method of genetic mapping know as DNA pooling using DNA samples from a consanguineous Bedouin family with BBS\textsuperscript{43}. This approach makes the assumption that the affected individuals within the family shared a common chromosomal region inherited from a common ancestral founder. This allowed the lab to simplify the analysis of the screen by looking for band shift patterns from the PCR products of short tandem repeat markers among three separate pools of DNA from parents, unaffected siblings, and affected patients. This pooling strategy reduced the number of PCR products that needed to be evaluated by 10-fold and eliminated the need to score individual
genotypes. With this strategy, the lab identified the locus for \textit{BBS4} and subsequently, many additional disease loci. Homozygosity by decent mapping was originally used with panels of approximately 400 markers, and the increased marker density afforded by modern genotyping platforms with hundreds of thousands of SNPs can be used to greatly increase the precise regions of homozygosity shared among affected family members. Identifying regions of homozygosity gives researchers a powerful and economic strategy to identify the linked region in consanguineous families, after which the gene and causative mutation can be identified with targeted sequencing. The software package PLINK\textsuperscript{46} can be used to identify these regions of homozygosity once genotypes have been derived from microarray genotyping arrays\textsuperscript{47}.

\textbf{Software}

\textbf{Plink}

Plink is a software package used for the analysis of genome-wide association study (GWAS) data. PLINK is designed to use genotype data extracted from microarray genotyping arrays. PLINK can also be used to identify regions of autozygosity, identity by descent and other effects expected in pedigrees.

\textbf{Linkage Disequilibrium}

Linkage disequilibrium is defined as alleles at two or more loci occurring together with a greater frequency than expected by chance within a given population. Linkage disequilibrium is influenced by factors such as genetic selection, the rate of genetic recombination, mutation rate, and genetic drift. When a disease mutation arises from a founder within a population, the disease mutation will be in linkage disequilibrium with alleles close to the gene containing the mutation.
This concept allows researchers to identify genomic regions potentially harboring a disease-causing gene or genetic risk factor by identifying alleles that are shared by affected members of a population more frequently than predicted by chance compared to non-affected individual in the population. In such a situation, the disease would be stated to be in linkage disequilibrium with the genetic marker(s). Such a situation could arise, for example, in an isolated population where an ancestral founder of the population contributed to the prevalence of the disease by being the source of the mutation. This would allow researchers to look for potential disease-causing mutations within regions of linkage disequilibrium that are shared among multiple consanguineous families from the same geographic region because the families may have a shared common ancestor that passed on the disease mutation.

**Whole Exome Sequencing**

Whole exome sequencing (WES) is the high-throughput sequencing of the human exome, which consists of every transcribed exon in the human genome. The exome represents approximately 1-2% of the human genome (30-60 Mb) and is the protein-coding region of the genome. The ability to quickly and accurately analyze the exome of patients suffering from many disorders has revolutionized the field of precision medicine. However, the amount of data produced through WES can make it challenging to properly call and annotate variants and requires multiple filtering steps to remove false positive variants. In order to take advantage of the latest sequencing technologies I recently built a new WES pipeline with the help of Dr. Adam DeLuca. In this section I will describe the pipeline that we are currently using at the Institute for Vision Research that has been highly optimized for use on the University of Iowa’s high-performance compute cluster “Argon” and provide a schematic in Figure 2 for visualization. In brief, the WES
process consists of exome enrichment and sequencing; preprocessing and alignment to a reference genome; alignment refinement; variation calling and annotation. After the comprehensive set of annotated variants have been generated, they are filtered and interpreted in a patient-specific manner.

**Sequence Generation**

Exome libraries were generated using commercially-available exome enrichment and library construction kits from Agilent and Nimblegen. Libraries were sequenced using an Illumina HiSeq 2500 or 4000 to obtain a depth of at least 80x over approximately 90% of the targeted regions.

**Sequence Alignment and Processing**

Before aligning to the genome, we use the MarkIlluminaAdapters function in Picard tools\textsuperscript{48} to remove any existing adapter-trimming tags created by Illumina sequencing thus reducing the amount of sequence that is not meaningful to the individual exome, thereby improving the overall sequence quality. The sequence is then aligned to the human reference genome GRCh37 using the Burrows-Wheeler Aligner\textsuperscript{49}. Duplicate sequences, which are often artifacts of the sequencing library preparation, are removed using the MarkDuplicates function in Picard tools\textsuperscript{48}. Indel realignment and base quality score recalibration are performed using the Genome Analysis Toolkit version 3 (GATK3)\textsuperscript{50} producing a recalibrated BAM (binary alignment map) file for each sample.
**Calling and Annotating Variants**

Variants are called using the GATK Haplotype Caller, which accepts the recalibrated BAM as input and produces a file of variations from the reference genome in the gVCF format. To reduce the overall computational time required to process a batch of exomes, we optimized this process by splitting each sample into 256 slices that can be run concurrently on our available 256 compute threads. This results in 256 slice gVCF files containing variants for each sample that are then pooled together using the CombineGVCF’s tool in GATK with their matching slice gVCFs from the other samples within the batch to create a total of 256 slice gVCFs containing information for every sample run within the batch. Joint genotyping, i.e. genotyping across multiple samples at the same time, is performed for all variants on each of the 256 slice gVCFs using the GATK GenotypeGVCFs tool to produce the best genotyping results. All 256 slice VCFs are merged into a single whole-batch VCF and indexed using the BCFTools concat and index functions. Individual samples are then normalized to make the VCF into multi-line variants and separated into their own VCF files using BCFtools. Allele frequencies from the genome aggregation database (gnomAD) are annotated to variants along with a custom sequence analysis and annotation regarding the variant using an in-house annotation pipeline (Institute for Vision Research, Iowa City, IA).

**Variant Filtering**

A typical exome sequencing result produces more than one hundred thousand variations. Thus, it is critical to refine the number of variants for potential experimental follow-up. Our filtering strategy consists of multiple steps to reduce the number of false positive variants. 1) The removal of low-quality variants as defined by those with a GATK quality score of less than 25, a quality
by depth score of less than one, a minimum depth of coverage of 10x, or fewer than 20% of
reads supporting the variation. 2) The removal of variants with a minor allele frequency greater
than the prevalence of the disease of interest within the general population using data from
gnomAD\textsuperscript{54} when the inheritance pattern is dominant. When the pattern of inheritance is
recessive, a stricter minor allele frequency is applied. To exclude systematically called variants,
we also remove variants that represent more than 10% of alleles within our local variant database
of over 1500 exomes. 3) Functional filtering is used to remove variants unlikely to have an effect
on the resultant protein. This includes synonymous variants outside of splice regions, as well as
intrinsic variants not involved in splicing. 4) Finally, family-based filtering can be applied to
remove variants that are not consistent with disease segregation. Due to the variability of familial
data, this step is performed outside of our automated pipeline and specific criteria are determined
for each case individually.
Figure 2. Depiction of the 2nd half of the whole exome sequencing pipeline. In this depiction, Haplotype Caller creates 3 slice gVCFs for samples 1-3. All slice 1 gVCFs are combined into a single batch 1 gVCF and are joint genotyped using Genotype gVCFs. Following the genotyping all the batch files are merged into a master VCF and then demultiplexed back into separate VCF files for each sample (1-3) in preparation for annotation.
Experimental Confirmation and Validation

Once the variants have gone through our initial filtering and annotation, there are a few more steps required before classifying the variant as a plausible candidate for disease presentation. First, the variant must be confirmed to exist in the proband. Second, the variant is assessed to determine if it co-segregates with disease in the family of the proband. Third, the variant must not be found at a frequency above disease prevalence within ethnically-matched control samples. Finally, the variant must functionally impair the gene, or be identified in a statistically significant number of additional families if functional data is unavailable.

Software and Resources

FASTQ File Format

The FASTQ file format is standard file format for storing base calls and quality scores from next-generation sequencers. We routinely perform paired-end sequencing for our exomes with the forward and reverse reads being stored in separate FASTQ files.

Picard

Picard is a set of tools within the Genome Analysis Toolkit used for manipulating SAM and BAM files. Picard can be used to effectively remove the Illumina adapters from the sequence by assigning them a quality score of 0 so they do not affect the mapping of reads when performing read alignment. Picard is also routinely used to remove redundant reads from SAM and BAM files to eliminate PCR artifacts created in the exome library preparation.
**BAM File Format**

The sequence alignment/map (SAM) file format is used to store sequence data and alignment information for short read sequence. The binary version of a SAM format file is known as a BAM file. BAM file format has become the standard for storing aligned reads within next-generation sequencing projects.

**Burrows-Wheeler Aligner**

The Burrows-Wheeler Aligner (BWA) is a next-generation sequence alignment program. BWA utilizes the Burrows-Wheeler transform to align paired-end short sequence reads to the human reference genome. BWA accepts FASTQ format files and produces aligned sequence in the BAM file format.

**Genome Analysis Toolkit**

The Genome Analysis Toolkit (GATK) is a software package used for the processing, realignment and variant calling of whole exome sequencing data. The GATK IndelRealigner and BaseQualityScoreRecalibration tools are used for quality control and best mapping. The GATK HaplotypeCaller is used for calling variants within exome sequencing projects. The GATK Genotype gVCF tool is used for joint genotyping of multiple samples by producing genotype likelihoods and then re-genotyping the samples.
**BCFtools and VCF File Format**

The variant call format (VCF) is the standard file format used by variant calling programs when reporting variants discovered during next-generation sequencing projects\(^{52}\). VCF is flexible file format that can handle custom annotation of variants within one or more samples. The file format allows for the addition of further annotation and tags to be assigned to each variant. BCFtools\(^{56}\) is a software package that provides a plethora of tools to manipulate VCF files and is a marked improvement in speed over older tools that used to be included in our pipeline.

**Genome Aggregation Database**

The Genome Aggregation database (gnomAD) is a data set of 125,748 exomes and 15,708 genomes from unrelated individuals sequenced as part of various disease-specific and population studies\(^{57}\). All of the raw data collected has been reprocessed through the same pipeline and have undergone joint variant calling to increase consistency across projects. This data is used to identify minor allele frequencies (frequency the second most common allele occurs in a specific population) for our sequencing data in order to filter out common variants within specific populations. Population-level allele frequencies are available for European (non-Finnish), African, European (Finnish), Latino, Ashkenazi Jewish, East Asian, and South Asian populations.

**BEDTools**

BEDTools is a program used to identify overlapping genomic intervals in BED or VCF file format\(^{58}\). A BED file is a tab delimited file format used to define genomic intervals\(^ {59,60}\). BEDTools can identify intersects between multiple BED files, merge overlapping features
between BED files, sort BED files, and even subtract features found in one BED file from another BED file.

**Integrative Genomics Viewer**

Integrative Genomics Viewer (IGV) is a tool that is used to visually analyze genomic data such as BAM files\(^6\). IGV allows users to visually inspect files from multiple samples to make comparisons regarding ethnic polymorphisms and sequencing artifacts. It also allows users to identify and visually inspect copy number variants, gene expression data, and even phasing if the two variants are close enough to each other to be seen on different sequencing reads as seen in Chapter 4.

**Copy Number Variant Calling**

Copy number variants (CNVs) are structural variants in DNA altering the number of copies of specific regions of DNA. These variants can be duplications or deletions of regions of DNA that can span more than 1kb of nucleotides in the genome. CNVs can be inherited similarly to single nucleotide polymorphisms but can also spontaneously arise during the development of an embryo. CNVs are known to cause disease, including specific microduplication and deletions syndromes, and the identification and interrogation of CNVs is critical in our research\(^6^2-6^4\).

**CoNIFER**

CoNIFER is a tool designed to identify CNVs using exome sequencing data\(^6^5\). CoNIFER combines read depth from multiple exomes with singular value decomposition methods to
discover CNVs. One major drawback we identified while using CoNIFER was the inability of
the algorithm to identify CNVs that were smaller than two exons in length as demonstrated in
chapter 4 of this thesis.

**CLAMMS**

CLAMMS is an algorithm designed to identify CNVs using exome sequencing data\(^6\)\(^6\).

CLAMMS also utilizes data from numerous exomes, but instead creates a reference panel with a
model for each exome capture region. CNVs are called using a hidden Markov model for each
individual sample after the reference panel is created. CLAMMS was able to successfully
identify a two-exon deletion in *BBS4* and has now become our default choice for CNV calling
within our exome data.

**Variant Pathogenicity Prediction**

With each person containing thousands of variants that have no effect on their disease
presentation, it is critical to have strategies and tools available to filter out variants that are
unlikely to affect disease presentation in patients. Many tools have been designed to predict the
pathogenicity of variants identified in sequencing projects. We frequently use three different
algorithms to help us make decisions regarding variants identified through our exome sequencing
pipeline. The predictions from these algorithms are used in tandem to overcome potential bias or
weakness that each algorithm alone would suffer from.
PolyPhen2
PolyPhen2 is a software and web server that predicts the possible impact of single nucleotide polymorphisms on the stability and function human proteins based on structural and comparative evolutionary considerations.\textsuperscript{67,68}

SIFT
SIFT is a software designed to predict the pathogenicity of a variant based upon sequence homology and was first released in 2003.\textsuperscript{69} The algorithm makes the assumption that highly-conserved amino acids between species are more likely to damage the function of a protein.

CADD
CADD is another pathogenicity prediction software that was designed to utilize data from over 60 different sources including but not limited to, surrounding sequence context, gene model annotations, evolutionary constraint, epigenetic measurements, and functional predictions to produce a CADD score from a machine learning model.\textsuperscript{70} The machine learning model is trained on over 30 million variants and hundreds of features and was first available for use in 2014.\textsuperscript{71}

Variant Burden Analysis
A Mendelian disease can be classified as a disease that can be caused by a single locus and can be tracked through inheritance patterns within families. Diseases that tend to be sporadic or that don’t follow the standard inheritance models may be considered as complex disease. Complex diseases may be caused by a burden of rare variants or a variant in tandem with an environmental
effect. With the continual decrease of next-generation sequencing cost there has been significant effort devoted to developing statistical methods for testing the association between rare variants and complex disease in order to identify the cause of complex diseases\textsuperscript{72-74}.

The first type of test developed from this effort is known as a burden test. Burden tests combine rare variants in a genetic region into a single burden variable and test the association between that variable and the disease phenotype. Common burden tests include the cohort allelic sum test\textsuperscript{75}, the combined multivariate and collapsing method\textsuperscript{76}, and the nonparametric weighted sum test\textsuperscript{77}. Burden tests often make the assumption that all rare variants within the genomic region are causal and affect the phenotype in a similar manner which causes them to have a significance loss in statistical power when those assumptions are violated\textsuperscript{78,79} and can be visualized in Figure 3.
To address the potential shortcomings of burden tests, kernel-based methods like the sequence kernel association test (SKAT) were developed. Instead of assigning a cumulative score for the variants, SKAT combines individual variant scores with weights to allow for both protective and deleterious variants. Although SKAT does not suffer from assumptions made by burden tests, it
can be less powerful when a large portion of the rare variants are actually casual and influence the phenotype in the same manner\textsuperscript{79,80} (Figure 4). SKAT was also designed for large sample studies and can produce false-positives for small association studies\textsuperscript{80,81}.

Figure 4. Visualization of a SKAT-based analysis.
To overcome the limitations from both types of tests, SKAT-O was designed to combine the strengths of burden tests and kernel-based methods like SKAT by creating a weighted average of both types of tests\textsuperscript{82}. It was also designed to have analytical adjustment method to calculate the reference distribution for the small sample and control type I error (false positives). These features allow researchers to perform variant burden analysis on smaller exome sequencing cohorts and have the possibility of identifying genomic regions that contribute to the phenotype of interest.

**SKAT Package**

The SKAT package is an R package that allows users to test for association between SNP sets and continuous/binary phenotypes while adjusting for covariates and kinships\textsuperscript{83}. Users can choose between standard SKAT analysis and SKAT-O analysis. The package requires a genotype matrix containing genotypes from all the samples being studied and vectors of continuous or binary phenotypes. The package will accept files that have been transformed using PLINK making it easy to create all the necessary files from VCF format. It is also necessary to include a sets file containing information on all of the SNPs that the user is interrogating with the algorithm.

**Over-representation Analysis**

Over-representation analysis (ORA) is a method that measures the percentage of genes in a pathway or group that have been perturbed. When performing variant burden analyses it can be used to identify pathways or groups that may been more important to the disease of interest by performing ORA on the top results from burden analysis as depicted in Figure 5.
ORA is calculated using the hypergeometric test to evaluate the significance of enrichment for a specific category or term in the gene set the user provides. Imagine there are \( n \) genes in the gene set (A) identified by the user and \( m \) genes in the reference gene set (B). Further imagine that there are \( k \) genes in A and \( j \) genes in B for a specific pathway or GO category (C). If B has genes that are also in A the hypergeometric test in Figure 6 is used to evaluate the significance of enrichment for C in the candidate gene set A.
WebGestalt was originally released in 2005 and is a free to use web-based set of tools for functional enrichment analysis. WebGestalt can perform over-representation analysis (ORA), gene set enrichment analysis and network topology-based analysis. It also supports analysis for 12 different model organisms and 324 sets of gene identifiers from various databases and technology platforms.

Conclusion

In this chapter I discussed the tools and technologies that I applied during my thesis research that will be discussed in Chapters 3, 4, and 5. As sequencing technology continues to improve and decrease in cost, more and more researchers will participate in next-generation sequencing projects. It is important for both wet and dry lab researchers to understand the power and limitations of these technologies to ensure they design their sequencing projects with the highest efficiency and best possible chance of acquiring a significant finding. A thoughtful, well designed experiment is still and always will be the most important component to answering scientific questions.
CHAPTER 3: IDENTIFYING THE GENETIC CAUSE OF KERATOCONUS IN THREE BEDOUIN FAMILIES

Abstract

Purpose. Keratoconus (KC) is the most common corneal dystrophy, affecting 1 in 2000 people worldwide. Despite the prevalence of this disease, researchers have been unable to identify a confirmed genetic cause of KC. There is strong evidence supporting a more frequent occurrence of KC coupled with a more aggressive onset of disease within Israeli and Saudi Arabia populations. Knowing this, we sought to identify a genetic cause of KC by performing linkage coupled with whole exome sequencing within three Bedouin families with a history of KC.

Methods. We obtained a total of 16 samples from 3 different Bedouin families with a history of KC. These samples were genotyped with the Affymetrix Genome-Wide SNP 5.0 microarray platform. We analyzed the genotype data for regions consistent with segregation within each of the families using MERLIN, under a dominant model of inheritance. Regions of homozygosity within each of the families were identified using PLINK. One affected KC sample from each of the families was chosen for whole exome sequencing. Variants were filtered based upon predicted function, variant quality, and population prevalence (gnomAD) to create a variant list. Variant analysis was performed using a prioritization strategy in order to reduce the false genotyping rate.

Results. Three plausible variants were identified in previously reported KC-associated genes within two of our families. However, these variants did not segregate properly through the affected/unaffected members of the families, suggesting they are not responsible for KC in these families. We also failed to identify a plausible variant or gene that was shared between at least two of the three families within regions consistent with segregation or regions of homozygosity.
In addition, no single gene with plausible disease-causing variants was shared across all three families. We were able to identify 6 plausible variants shared between 2 of the 3 families that have been shared with our collaborator in Israel for further investigation.

**Conclusions.** Our data supports the genetic heterogeneity of KC even within relatively isolated populations. Though we did not identify the causative variant/gene of KC, we were able to utilize linkage analysis to significantly reduce the genomic space when performing a gene discovery project. This strategy shows the continued importance of collecting familial data when studying diseases without a known genetic cause.

**Introduction**
Keratoconus (KC) is the most common corneal dystrophy and a major cause of blindness worldwide\(^{86,87}\). The primary characteristic of KC is the progressive thinning of the cornea (75% loss) which leads to corneal protrusion (conical shape) due to a pressure on the weakened cornea by the normal intraocular pressure levels in the eye\(^{86-92}\)(Figure 7). Patients with KC are often diagnosed with myopia, irregular astigmatism, distortion in vision, and sensitivity to light\(^{93-96}\). The loss of visual acuity is the major outcome of KC development due to changes in the corneal curvature which affects the refraction of light and the its transmission onto the retina\(^{87,88}\). KC can also lead to the development of acute corneal edema and scar formation\(^{97-99}\). Examination of KC diagnosed eyes show a severe loss of corneal collagen fibrils, membrane abnormalities of the keratocytes and fragmentation of the Bowmans’ membrane\(^{90,100-102}\). Slit-lamp examination can identify Vogt’s striae, which present as fine vertical white lines in the deep stroma and Descemet’s membrane that disappear upon pressure. Finally, examination of the basal layers of the corneal epithelium can identify iron deposits within a KC eye.
KC occurs in all ethnic groups, but there is an increase in the prevalence (25/100,000-2300/100,000) of KC in Asians populations (China, India, Pakistani, and Bangladesh) compared to Caucasians (3.3/100,000)\textsuperscript{88-92}. There is also an increase in the prevalence of KC within Saudi Arabia and Israel (20/100,000)\textsuperscript{103,104}. While KC is generally thought to develop within the teens and early twenties with progression lasting until the fourth decade of life, patients from China, Israel and Saudi Arabia have an earlier age of presentation and a more severe phenotypic presentation than other populations\textsuperscript{103-106}. The combination of an earlier onset and increased severity requires corneal replacement surgery at an earlier age and may lead to the patient having multiple surgeries\textsuperscript{107}.

KC is considered a complex disorder with evidence supporting both autosomal recessive and autosomal dominant modes of inheritance\textsuperscript{3,26-31}. The percentage of familial KC varies greatly depending on the report, spanning from only 5% all the way to 23%, providing further evidence
that KC is complex and multifactorial\textsuperscript{86,108,109}. There are also conflicting reports concerning the ratio of males to females with KC, though the majority of the reports support an equal occurrence within both males and females\textsuperscript{88,90,91,110}. Autosomal dominant KC presents with multiple phenotypes that have incomplete penetrance making it even more challenging to track within families\textsuperscript{111,112}. There is also evidence that environmental and lifestyle factors play a role in the development of KC. However, the prevalence of KC in first degree relatives of KC patients and in monozygotic twins further supports the genetic component of KC and disputes the cause being solely due to environment or lifestyle. Finally, an increase in corneal curvature, which is highly heritable, has been associated with KC and further supports a complex model of inheritance.

While KC can present without any other comorbidities, there is ample evidence supporting the association of KC with both genetic and non-genetic factors. Associations between contact lens use, eye rubbing, allergies, UV light, spring conjunctivitis, atopic disorders, trauma and KC have been established\textsuperscript{113-121}. KC is also associated with genetic diseases such as Apert syndrome, Down syndrome, Leber congenital amaurosis, Rieger syndrome, Marfan syndrome, joint hypermobility, pseudoxanthoma elasticum, retinitis pigmentosa, GAPO syndrome, mitral valve prolapse and others\textsuperscript{116,122-130}. Linkage analysis studies and genome wide association studies (GWAS) have been performed to identify genes/loci associated with KC. These results have been widely disputed as candidate genes/loci in one study may not verify in another study. The lack of verification through replication may be due to different ethnicities/environmental variable between the cohorts. All reported KC associated genes are shown in Table 1. These data further strengthen the hypothesis that KC is a complex genetic disease that requires further study.
Table 1. Genes associated with Keratoconus.

<table>
<thead>
<tr>
<th>COL1A1</th>
<th>COL4A1</th>
<th>COL4A3</th>
<th>COL5A1</th>
<th>COL6A1</th>
<th>COL8A1</th>
<th>COL8A2</th>
<th>CRB1</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRX</td>
<td>DOCK9</td>
<td>FNDC3B</td>
<td>FOXO1</td>
<td>HGF</td>
<td>LOX</td>
<td>MIR184</td>
<td>MPDZ</td>
</tr>
<tr>
<td>PRDM5</td>
<td>RAB3GAP1</td>
<td>RXRA</td>
<td>SOD1</td>
<td>SPARC</td>
<td>TGFB1</td>
<td>VSX1</td>
<td>ZNF469</td>
</tr>
</tbody>
</table>

Knowing that there is continued confusion surrounding the genetic cause of KC and that KC is more frequent and more aggressive in Israel, we sought to find a genetic cause of KC by performing whole exome sequencing in three Bedouin families (Figure 8, Figure 9, Figure 10) with a family history of KC.

Figure 8. Pedigree of Family 1.
Figure 9. Pedigree of Family 2.

Figure 10. Pedigree of Family 3.
Materials and Methods

Collection of Bedouin samples
A total of sixteen DNA samples from 3 Bedouin families with a history of KC and their pedigrees were collected by Dr. Imtirate and Dr. Pavari at the Ben-Gurion University of the Negev in Beersheba Israel.

Affymetrix Genome-Wide SNP 5.0 Array
Genome-wide SNP genotypes were obtained for all samples with the Affymetrix 5.0 array (Affymetrix Inc, Santa Clara, CA) using the manufacturer’s recommended protocols. The BRLMM-P (Affymetrix (2007) BRLMM-P: a Genotype Calling Method for the SNP 5.0 Array) algorithm within the Affymetrix Power Tools suite (Affymetrix (2009) Affymetrix Power Tools) was used to calculate the individual genotypes for each of the samples using default parameters\textsuperscript{131}.

Identification of regions consistent with segregation
Parametric linkage analysis was performed within each family using MERLIN with default parameters as described previously\textsuperscript{34}. A dominant model of inheritance was employed for the analysis following the clinician’s preliminary diagnosis. Regions consistent with segregation (Logarithm of odds scores > 0) were identified within each family. Overlapping regions consistent with segregation between families was identified using the BEDtools intersect function\textsuperscript{58} on the BED files generated by MERLIN for the three families.
Identification of Runs of Homozygosity

Runs of homozygosity were identified for each family using the PLINK toolset\textsuperscript{46}. The default parameters for the runs of homozygosity algorithm were used except for increasing the homozygous window to 1000 kb and the number of heterozygotes allowed in a window to 10, based on the known consanguinity within two of the three families.

Whole Exome Sequencing

One exome of an affected member from each of the three families with Keratoconus were collected and processed using the procedure outlined below. Commercially available exome capture reagents (Agilent Sure Select v5) were used to capture the exome, followed up by sequencing at a minimum of 80x depth using an Illumina HiSeq 2500 at the University of Iowa’s DNA core. This relatively high depth enables detection of both single nucleotide variants and copy number variants.

Sequence Alignment and Processing

DNA sequence reads were converted to SAM file format and Illumina adapters removed from the sequence using the MarkIlluminaAdapter function in Picard tools (http://broadinstitute.github.io/picard) package. The trimmed sequence was aligned to the human genome reference genome GRCh37 using the Burrows-Wheeler aligner\textsuperscript{49} and duplicate reads were removed using the MarkDuplicates function in Picard tools. Indel realignment and base quality score recalibration were performed using GATK3\textsuperscript{51}.
Calling and Annotating Variants

Variants were called using the GATK Haplotype Caller,\textsuperscript{51} and joint genotyping was performed using the GATK GenotypeGVCFs tool\textsuperscript{51} to produce the best genotyping results. Individual samples were then normalized and separated into their own VCF file using BCFtools\textsuperscript{53}. Allele frequencies from the 1000 Genomes Project\textsuperscript{132}, dbSNP\textsuperscript{133}, and the genome aggregation database\textsuperscript{54} (gnomAD) were annotated to variants along with a custom sequence analysis and annotation regarding the variant using an in-house annotation pipeline (Institute for Vision Research, Iowa City, IA).

Copy Number Variant (CNV) analysis

CNVs were identified by utilizing the program, Copy Number Inference From Exome Reads (CoNIFER)\textsuperscript{65} on our exome sequencing data. CoNIFER analyzed the 32 exomes that were sequenced as a batch at the University of Iowa’s DNA core, including the 3 Keratoconus exomes, for CNV’s using the default parameters. CNV’s identified by CoNIFER were visually inspected using the Integrative Genomics Viewer\textsuperscript{61} and then molecularly confirmed using Sanger sequencing\textsuperscript{134}.

Variant Filtering

Variants with a minor allele frequency of greater than 1% in the 1000 Genomes project\textsuperscript{132} or gnomAD\textsuperscript{54} were judged to be too common to cause KC and were excluded from the analysis.
Variants with a GATK quality score of less than 25, or a quality by depth score of less than 1 were also excluded from the analysis.

**Variant Prioritization**

Variants were analyzed using a tiered testing strategy to reduce the false genotype rate. Tier one analysis only analyzed variants in genes previously reported as associated with KC. Tier two analysis included variants from regions consistent with segregation in all three families identified through our linkage analysis using a dominant model of inheritance. Tier three analysis consisted of variants within regions consistent with segregation in two of the three families using a dominant model of inheritance. The fourth tier of analysis included variants within the three regions of homozygosity shared between the families. Finally, our fifth tier of analysis included variants that were shared between at least 2 of the 3 samples. Variants were analyzed for disease prediction using PolyPhen2, SIFT, and CADD and allele frequency within gnomAD and our own cohort of over 1500 exomes.

**Experimental Confirmation**

Candidate variants identified by exome sequencing were confirmed to exist in the proband by Sanger sequencing. When possible, candidate variants were also confirmed through Sanger sequencing within family members to confirm segregation of the variant with disease in the family. Candidate variants identified within ethnically matched controls through visual inspection using the Integrative Genome Viewer (IGV) were regarded as non-disease causing and removed from the candidate list.
Results

Variants in KC-associated Genes (Tier 1)

We identified 5 different heterozygous variants within KC-associated genes from our exome data (Table 2). Each variant passed the minor allele frequency filter, however the population we were interrogating is underrepresented in gnomAD. Therefore, it is critical to manually inspect the variants using IGV along with ethnically matched controls to identify ethnic polymorphisms. All variants in Table 2 passed visual inspection using ethnically matched controls. However, the variants in \textit{COL6A1} and \textit{COL8A2} were predicted to be benign by PolyPhen2\textsuperscript{67} and CADD\textsuperscript{70} and were excluded from further analysis.

Table 2. Variants identified in KC-associated genes.

<table>
<thead>
<tr>
<th>SAMPLE</th>
<th>Gene</th>
<th>Genomic Change</th>
<th>Coding Change</th>
<th>Protein Change</th>
<th>gnomAD Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>KER-4</td>
<td>\textit{COL4A1}</td>
<td>chr13:110839625:G&gt;A</td>
<td>c.1588C&gt;T</td>
<td>p.Pro530Ser</td>
<td>1.04E-03</td>
</tr>
<tr>
<td>KER-4</td>
<td>\textit{COL5A1}</td>
<td>chr9:137694837:C&gt;T</td>
<td>c.3110C&gt;T</td>
<td>p.Thr1037Met</td>
<td>8.05E-05</td>
</tr>
<tr>
<td>KER-16</td>
<td>\textit{COL6A1}</td>
<td>chr21:47417656:C&gt;G</td>
<td>c.1504C&gt;G</td>
<td>p.Pro502Ala</td>
<td>Not reported</td>
</tr>
<tr>
<td>KER-16</td>
<td>\textit{COL8A2}</td>
<td>chr1:36564213:C&gt;G</td>
<td>c.1069G&gt;C</td>
<td>p.Gly357Arg</td>
<td>3.77E-04</td>
</tr>
<tr>
<td>KER-16</td>
<td>\textit{MPDZ}</td>
<td>chr9:13188953:A&gt;T</td>
<td>c.2194T&gt;A</td>
<td>p.Ser732Thr</td>
<td>1.21E-03</td>
</tr>
</tbody>
</table>

Regions Consistent with Segregation

From our linkage analysis we were able to identify regions consistent with segregation in each family and shared between the families as depicted in Table 3. Using a dominant model of inheritance based upon the referring clinician’s report, we were able to identify a total of 25 million nucleotides (Mb) shared between the 3 families that were consistent with segregation.
<table>
<thead>
<tr>
<th>Family</th>
<th>Mb</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family 1</td>
<td>379 (11.7%)</td>
<td>3750</td>
</tr>
<tr>
<td>Family 2</td>
<td>728.786 (22.5%)</td>
<td>6241</td>
</tr>
<tr>
<td>Family 3</td>
<td>1685 (52%)</td>
<td>13929</td>
</tr>
<tr>
<td>Shared</td>
<td>25 (0.77%)</td>
<td>245</td>
</tr>
</tbody>
</table>

Table 3. Genomic space and number of genes identified through linkage analysis. Depicted above are the number of mega bases and percentage of the total genome in each family found to be consistent with segregation. The table also shows the number of genes in each of those regions. Through linkage analysis we were able to substantially reduce the number of genes and fraction of the genome we were inspecting to greatly increase our power when analyzing data.

**Plausible Variants in Shared Regions Consistent with Segregation (Tier 2)**

In order to reduce the false genotype rate when performing gene discovery analyses, we used our linkage results to reduce the genomic space we needed to interrogate. Our linkage results left us with only 245 genes (1.2% of all genes) to interrogate for variants reported by our whole exome sequencing pipeline. Of those 245 genes we identified plausible (rare and non-silent) variants within only 13 (5.3%) of those genes.

Table 4 shows the power of using linkage analysis to reduce the number of plausible variants to investigate when performing gene discovery experiments. Though we were able to reduce the genomic space we investigated, there were no shared single nucleotide variants or CNV’s between the three families within these regions that passed visual inspection and functional effect predictive algorithms.
Table 4. Progression in filtering of variants within the three exomes. By utilizing our linkage analysis results in tandem with our whole exome sequencing filtering, we were able to reduce the total number of plausible variants significantly. Filtering includes removal of common variants identified in databases such as gnomAD and 1000 Genomes.

Plausible Variants in Shared Regions Consistent with Segregation in Two Families (Tier 3)

We were able to identify plausible variants in regions consistent with segregation within two of the families. Plausible variants that were shared between two of the families were identified and went through further filtering and inspection mentioned previously. After this process, none of the variants remained as candidates for further testing. This filtering strategy is depicted in Table 5.
Table 5. Number of plausible variants in regions consistent with segregation between two families. Shown are the number of variants identified in regions consistent with segregation between 2 families. We then identified specific variants shared between 2 families (row 3) in those regions for manual inspection using IGV\textsuperscript{61} and variant effect prediction with PolyPhen2\textsuperscript{67}, SIFT\textsuperscript{69}, and CADD\textsuperscript{70}. Following filtering, we deemed there to be no shared plausible variants.

<table>
<thead>
<tr>
<th>Shared between families</th>
<th>1 &amp; 2</th>
<th>1 &amp; 3</th>
<th>2 &amp; 3</th>
</tr>
</thead>
<tbody>
<tr>
<td># Plausible variants (PV)</td>
<td>22</td>
<td>66</td>
<td>62</td>
</tr>
<tr>
<td>#PV shared</td>
<td>0</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>#PV shared after filtering</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Variants Found within Runs of Homozygosity within Each Family (Tier 4)**

Knowing that two of the families were consanguineous, we also assessed if any runs of homozygosity (RoH) were present within the families and checked for variants within any shared RoHs. We identified four RoHs that were shared between two families and none between all three families (Table 6). However, after applying our filtering criteria there were no plausible variants within these regions.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Start</th>
<th>End</th>
<th>Overlapping Families</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr5</td>
<td>144539153</td>
<td>144588121</td>
<td>Family1 &amp; Family3</td>
</tr>
<tr>
<td>chr8</td>
<td>8791114</td>
<td>9358381</td>
<td>Family1 &amp; Family3</td>
</tr>
<tr>
<td>chr17</td>
<td>3635768</td>
<td>3900911</td>
<td>Family2 &amp; Family3</td>
</tr>
<tr>
<td>chr21</td>
<td>30179629</td>
<td>31268570</td>
<td>Family1 &amp; Family3</td>
</tr>
</tbody>
</table>

Table 6. Shared runs of homozygosity between families. Depicted above are the runs of homozygosity that were shared between at least 2 of the families.
Variants Shared Between at Least 2 of the Families (Tier 5)

Finally, we interrogated all rare variants (including those outside the segregating regions) reported through our exome sequencing pipeline that were shared among at least 2 of the 3 samples. For each of the 143 shared variants identified, we used IGV\textsuperscript{61} to analyze the BAM files to manually validate the variants against ethnically matched controls and used gnomAD to ensure that the variants were not commonly observed. We also used PolyPhen\textsuperscript{67}, SIFT\textsuperscript{69}, and CADD\textsuperscript{70} to predict the pathogenicity of the variant to further reduce the number of candidate variants for future exploration. There were no variants shared between all three families that survived all of our filtering criteria. A total of six variants passed all of our filtering and are reported in Table 7.

<table>
<thead>
<tr>
<th>Found In</th>
<th>Gene</th>
<th>Coding Change</th>
<th>Protein Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family 1 &amp; 2</td>
<td>CC2D2B</td>
<td>c.766C&gt;A</td>
<td>p.Pro256Thr</td>
</tr>
<tr>
<td>Family 1 &amp; 3</td>
<td>CD2AP</td>
<td>c.902A&gt;T</td>
<td>p.Lys301Met</td>
</tr>
<tr>
<td>Family 1 &amp; 2</td>
<td>DDX60L</td>
<td>c.4573delC</td>
<td>p.Leu1525CysfsTer2</td>
</tr>
<tr>
<td>Family 1 &amp; 3</td>
<td>PKD1L2</td>
<td>c.404C&gt;T</td>
<td>p.Pro135Leu</td>
</tr>
<tr>
<td>Family 1 &amp; 2</td>
<td>RGS20</td>
<td>c.1123A&gt;C</td>
<td>p.Lys375Gln</td>
</tr>
<tr>
<td>Family 1 &amp; 2</td>
<td>RIOXI</td>
<td>c.716A&gt;G</td>
<td>p.Gln239Arg</td>
</tr>
</tbody>
</table>

Table 7. Tier 5 variants.
We also checked for any CNVs that were shared between all three families using CoNIFER\textsuperscript{65}. Of particular interest was a region with a duplication in family 1 and a deletion in family 3 (Figure 11). Investigation of this region using IGV\textsuperscript{61} led to the discovery of a homozygous deletion of two exons in family 2. Sanger sequencing of this area failed to validate the homozygous deletion. We believe this is caused by the relatively low complexity of the region, leading to aberrant calls by CoNIFER\textsuperscript{65}.

![Figure 11. Reported CNV in two families. Depth of coverage is shown for all three KC samples for which exome data was obtained. Based upon normalized coverage levels, the sample from Family 3 is clearly missing most or all of exons 7 and 8 of gene, and the sample from Family 1 appears to have substantially increased coverage, consistent with a duplication.](image)

**Discussion**

**Variants in KC-associated Genes**

I identified three variants in genes that have been previously reported as associated with KC within 2 of our KC samples. Specifically, within our exome sample from family 2, I identified a
single heterozygous variant in both \textit{COL4A1} and \textit{COL5A1}. I also identified a heterozygous variant in \textit{MPDZ} within our family 3 exome sample. \textit{COL4A1} encodes the alpha-1 subunit of collagen type IV. Type IV collagen forms a meshwork instead of fibrillin structures and \textit{COL4A1} is expressed in the cornea\textsuperscript{135}. While \textit{COL4A1} has been implicated in KC, it is also reported to cause porencephaly 1 and schizencephaly, brain small vessel disease with or without ocular anomalies, hereditary angiopathy with nephropathy, aneurysms, and muscle cramps, susceptibility to intracerebral hemorrhage, and tortuosity of retinal arteries\textsuperscript{136-146}. Each of these disorders are reported to have autosomal dominant patterns of inheritance with variants at amino acid 519, 528, and 538 having been implicated as disease-causing in the aforementioned diseases. The variant we identified has been classified as a variant of unknown significance but is a proline that is highly conserved through all species based upon Homologene\textsuperscript{147} (Table 8) and is predicted as probably damaging using PolyPhen2 and CADD\textsuperscript{67,71}. However, Sanger sequencing within the family showed that the variant (T > C) did not segregate properly with disease as seen in Figure 12, thus eliminating it as a potential cause of keratoconus in this family.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|c|c|c|c|}
\hline
\textbf{Species} & \textbf{AA 525} & \textbf{AA 526} & \textbf{AA 527} & \textbf{AA 528} & \textbf{AA 529} & \textbf{AA 530} & \textbf{AA 531} & \textbf{AA 532} & \textbf{AA 533} & \textbf{AA 534} & \textbf{AA 535} \\
\hline
H. sapiens & G & A & K & G & E & P & G & E & F & Y & F \\
\hline
P. troglodytes & G & A & K & G & E & P & G & E & I & Y & F \\
\hline
M. mulatta & G & A & K & G & E & P & G & E & I & Y & F \\
\hline
C. lupus & G & A & K & G & E & P & G & E & I & Y & F \\
\hline
B. taurus & G & A & K & G & E & P & G & E & I & Y & F \\
\hline
M. musculus & G & A & K & G & E & P & G & E & I & F & F \\
\hline
R. norvegicus & G & A & K & G & E & P & G & E & I & F & F \\
\hline
G. gallus & G & A & K & G & E & P & G & D & F & T & Y \\
\hline
\end{tabular}
\end{table}
We also identified a heterozygous p.Thr1037Met mutation in COL5A1 within the same sample. Variants in COL5A1 has been reported to cause Ehlers-Danlos syndrome which can have an effect on the corneal structure\textsuperscript{148-150,151} and is inherited in an autosomal dominant pattern of inheritance. While our variant has not been reported as a pathogenic variant, the threonine is conserved through mammals (Table 9), has a maximum minor allele frequency of 8/100,000 in gnomAD, and is predicted to be probably damaging by PolyPhen2 and CADD\textsuperscript{67,71}. Sanger sequencing confirmed the presence of the variant not only within all affected members of the family but also the unaffected sibling (Figure 13). The likelihood that both other affected members of the family would carry this variant is a 1 in 4 chance. Unfortunately, we are missing

<table>
<thead>
<tr>
<th></th>
<th>D. rerio</th>
<th>G</th>
<th>A</th>
<th>Q</th>
<th>G</th>
<th>E</th>
<th>P</th>
<th>G</th>
<th>D</th>
<th>I</th>
<th>F</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>X. tropicalis</td>
<td>G</td>
<td>A</td>
<td>K</td>
<td>G</td>
<td>E</td>
<td>P</td>
<td>G</td>
<td>E</td>
<td>V</td>
<td>F</td>
<td>F</td>
<td></td>
</tr>
</tbody>
</table>


Figure 12. Genotyping results for C > T variant in COL4A1 within Family 2. The variant was identified in the female labeled Exome but was not found in any of the other affected or unaffected members of the family.
the age of the other affected sibling. That information would help us judge if the unaffected sibling is too young to have developed KC as observed in this family, or if they are truly unaffected and expected to remain unaffected. Based upon expected age-of-onset for KC in the United States, the unaffected sibling might still develop KC. However, KC is frequently observed developing at an early age in some populations in the Middle East (Michael Wagoner, personal communication). Thus, if the affected sibling of unknown age is younger than the unaffected sibling, it is likely that the affection status of the unaffected sibling is correct, and this variant does not cause keratoconus. However, if the other affected sibling is older than the unaffected sibling, it is possible that the unaffected sibling has not yet presented with keratoconus and may develop keratoconus in the future. Until we have further information regarding the sibling’s age or are able to perform sanger sequencing on DNA from the father, we are unable to reliably comment on the affect this variant has within the family.

<table>
<thead>
<tr>
<th>Species</th>
<th>AA1032</th>
<th>AA1033</th>
<th>AA1034</th>
<th>AA1035</th>
<th>AA1036</th>
<th>AA1037</th>
<th>AA1038</th>
<th>AA1039</th>
<th>AA1040</th>
<th>AA1041</th>
<th>AA1042</th>
</tr>
</thead>
<tbody>
<tr>
<td>H. sapiens</td>
<td>A</td>
<td>G</td>
<td>K</td>
<td>E</td>
<td>G</td>
<td>T</td>
<td>K</td>
<td>G</td>
<td>D</td>
<td>P</td>
<td>G</td>
</tr>
<tr>
<td>M. musculus</td>
<td>A</td>
<td>G</td>
<td>K</td>
<td>E</td>
<td>G</td>
<td>T</td>
<td>K</td>
<td>G</td>
<td>D</td>
<td>P</td>
<td>G</td>
</tr>
<tr>
<td>R. norvegicus</td>
<td>A</td>
<td>G</td>
<td>K</td>
<td>E</td>
<td>G</td>
<td>T</td>
<td>K</td>
<td>G</td>
<td>D</td>
<td>P</td>
<td>G</td>
</tr>
<tr>
<td>G. gallus</td>
<td>T</td>
<td>G</td>
<td>K</td>
<td>E</td>
<td>G</td>
<td>T</td>
<td>K</td>
<td>G</td>
<td>D</td>
<td>P</td>
<td>G</td>
</tr>
<tr>
<td>D. rerio</td>
<td>A</td>
<td>G</td>
<td>K</td>
<td>E</td>
<td>G</td>
<td>A</td>
<td>K</td>
<td>G</td>
<td>D</td>
<td>P</td>
<td>G</td>
</tr>
<tr>
<td>X. tropicalis</td>
<td>A</td>
<td>G</td>
<td>K</td>
<td>E</td>
<td>G</td>
<td>S</td>
<td>K</td>
<td>G</td>
<td>D</td>
<td>P</td>
<td>G</td>
</tr>
</tbody>
</table>

In our sample from Family 3 we identified a heterozygous variant in *MPDZ*. This gene has been reported to cause Congenital Hydrocephalus 2 with or without brain or eye anomalies which has an autosomal pattern of inheritance$^{152-154}$. While this variant affects a highly conserved amino acid (Table 10) and is predicted to be damaging, this variant is only on a single allele and thus can’t cause an autosomal recessive disorder. More importantly, this sample was not reported to have hydrocephalus and thus would not fit the phenotype variants in this gene are reported to cause. In addition, the variant did not segregate within the family properly according to the Sanger sequencing results. The unaffected mother and sister carry the variant while the affected father does not (Figure 14) effectively ruling out the variant as the cause of disease in this family.

Figure 13. Segregation of the C > T variant in *COL5A1* within Family 2. All family members with DNA available for testing presented with this variant.
<table>
<thead>
<tr>
<th>Species</th>
<th>AA 727</th>
<th>AA 728</th>
<th>AA 729</th>
<th>AA 730</th>
<th>AA 731</th>
<th>AA 732</th>
<th>AA 733</th>
<th>AA 734</th>
<th>AA 735</th>
<th>AA 736</th>
<th>AA 737</th>
</tr>
</thead>
<tbody>
<tr>
<td>H. sapiens</td>
<td>V 1</td>
<td>I 1</td>
<td>I 1</td>
<td>R 1</td>
<td>S 1</td>
<td>L 1</td>
<td>V 1</td>
<td>P 1</td>
<td>G 1</td>
<td>G 1</td>
<td></td>
</tr>
<tr>
<td>P. troglodytes</td>
<td>V 1</td>
<td>I 1</td>
<td>I 1</td>
<td>R 1</td>
<td>S 1</td>
<td>L 1</td>
<td>V 1</td>
<td>P 1</td>
<td>G 1</td>
<td>G 1</td>
<td></td>
</tr>
<tr>
<td>M. mulatta</td>
<td>V 1</td>
<td>I 1</td>
<td>I 1</td>
<td>R 1</td>
<td>S 1</td>
<td>L 1</td>
<td>V 1</td>
<td>P 1</td>
<td>G 1</td>
<td>G 1</td>
<td></td>
</tr>
<tr>
<td>C. lupus</td>
<td>V 1</td>
<td>V 1</td>
<td>I 1</td>
<td>R 1</td>
<td>S 1</td>
<td>L 1</td>
<td>V 1</td>
<td>P 1</td>
<td>G 1</td>
<td>G 1</td>
<td></td>
</tr>
<tr>
<td>B. taurus</td>
<td>V 1</td>
<td>V 1</td>
<td>I 1</td>
<td>R 1</td>
<td>S 1</td>
<td>L 1</td>
<td>V 1</td>
<td>P 1</td>
<td>G 1</td>
<td>G 1</td>
<td></td>
</tr>
<tr>
<td>M. musculus</td>
<td>V 1</td>
<td>V 1</td>
<td>I 1</td>
<td>R 1</td>
<td>S 1</td>
<td>L 1</td>
<td>V 1</td>
<td>P 1</td>
<td>G 1</td>
<td>G 1</td>
<td></td>
</tr>
<tr>
<td>R. norvegicus</td>
<td>V 1</td>
<td>V 1</td>
<td>I 1</td>
<td>R 1</td>
<td>S 1</td>
<td>L 1</td>
<td>V 1</td>
<td>P 1</td>
<td>G 1</td>
<td>G 1</td>
<td></td>
</tr>
<tr>
<td>G. gallus</td>
<td>I 1</td>
<td>I 1</td>
<td>L 1</td>
<td>I 1</td>
<td>R 1</td>
<td>S 1</td>
<td>L 1</td>
<td>V 1</td>
<td>P 1</td>
<td>G 1</td>
<td></td>
</tr>
<tr>
<td>D. rerio</td>
<td>V 1</td>
<td>I 1</td>
<td>V 1</td>
<td>I 1</td>
<td>R 1</td>
<td>S 1</td>
<td>L 1</td>
<td>V 1</td>
<td>P 1</td>
<td>G 1</td>
<td></td>
</tr>
<tr>
<td>X. tropicalis</td>
<td>V 1</td>
<td>I 1</td>
<td>V 1</td>
<td>I 1</td>
<td>R 1</td>
<td>S 1</td>
<td>L 1</td>
<td>V 1</td>
<td>P 1</td>
<td>G 1</td>
<td></td>
</tr>
</tbody>
</table>

Table 10. Conservation of Amino Acids in *MPDZ*.

Figure 14. Family genotypes for the *MPDZ* variant within Family 3.
Variants from Tier 2, 3, 4, and 5 analysis

We were unable to identify any plausible variants shared among all three of the KC families in areas consistent with segregation or within runs of homozygosity. However, we were able to identify 6 single nucleotide variants that were shared between two of the families that have been reported to our collaborator for further follow up through the collection of additional family members. We also identified a false positive CNV result that was shared between two families. This aberrant CNV was called by CoNIFER due to sequencing artifacts resulting from the low complexity of the region in question.

Conclusions

Though we were unable to definitely identify the genetic causes of KC in the families studied, we were able to demonstrate the power of using linkage analysis within families to reduce the number of plausible variants and thus decrease the false genotype rate by identifying specific intervals of genomic space to interrogate. While next-generation sequencing continues to decline in cost, it is important to remember that there are many variants within the human genome that are benign. It is, therefore, still crucial to ascertain high quality clinical and family data when possible, especially when performing novel gene discovery experiments, to reduce the false genotype rate and have the greatest chance to identify the true genetic cause of disease in patients.
CHAPTER 4: IDENTIFYING THE GENETIC CAUSE OF BARDET-BIEDL SYNDROME IN THREE DIFFERENT COHORTS OF PATIENTS THROUGH WHOLE EXOME SEQUENCING.

Abstract

Purpose. Bardet-Biedl Syndrome (BBS) is an autosomal recessive, genetically heterogenous, pleiotropic ciliopathy caused by at least 22 genes. The fundamental features of BBS include retinal degeneration, obesity, polydactyly, renal anomalies, hypogonadism, and cognitive impairment. To identify the genetic cause of BBS in three different cohorts of BBS patients, we performed next-generation sequencing (NGS) and bioinformatic analyses.

Methods. Whole exome sequencing was performed on a total of 36 samples from three different cohorts of BBS patients. Variants within the BBS samples were identified and analyzed using our NGS pipeline and specific filtering criteria. Only variants in known BBS genes were evaluated during the primary analysis. A secondary analysis was performed on each sample if the sample had an incomplete BBS genotype. The secondary analysis evaluated variants in a list of genes known to cause eye-disease assembled by the Institute for Vision Research at the University of Iowa.

Results. We identified complete genotypes for 4 previously unsolved BBS patients in our first cohort consisting of 26 probands. We also identified 8 individuals with variants in genes that are known to cause BBS-like phenotypes within the first cohort. In our second cohort consisting of 4 individuals, we identified two novel ARL6 mutations that properly segregate through the family. Finally, we identified 3 samples from our third cohorts of 9 Puerto Rican BBS patients with two different mutations in KCNV2, a known eye disease gene.

Conclusions. By performing NGS analysis of BBS patients, we identified the genetic cause of BBS in multiple patients, including novel mutations in known BBS genes. We also identified
multiple patients with complete genotypes for genes that are known to cause BBS-like phenotypes. Whether these genes are novel BBS genes remains to be seen as we perform functional testing to assess their role in cilia.

**Introduction**

Bardet-Biedl Syndrome (BBS) is a genetically heterogeneous, autosomal recessive disorder that results in multiple phenotypic traits\(^1\)-\(^3\). BBS is caused by mutations in genes associated with proper function of primary cilia,\(^17\) of which 22 have been identified to date (Table 11). Homozygous or compound heterozygous mutations within *BBS1*, *BBS10*, *BBS2*, *BBS9*, and *MKKS* (BBS6) are known to cause the majority of clinically diagnosed BBS cases\(^155,7\). BBS has a prevalence of ~1:160,000 in Switzerland and Tunisia, ~1:18,000 in Newfoundland, and ~1:36,000 within Arab populations in Kuwait\(^156,157,158,7\).

<table>
<thead>
<tr>
<th>BBS1 (BBS1)</th>
<th>BBS2 (BBS2)</th>
<th>BBS3 (ARL6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBS4 (BBS4)</td>
<td>BBS5 (BBS5)</td>
<td>BBS6 (MKKS)</td>
</tr>
<tr>
<td>BBS7 (BBS7)</td>
<td>BBS8 (TTC8)</td>
<td>BBS9 (PTHB1)</td>
</tr>
<tr>
<td>BBS10 (BBS10)</td>
<td>BBS11 (TRIM32)</td>
<td>BBS12 (BBS12)</td>
</tr>
<tr>
<td>BBS13 (MKS1)</td>
<td>BBS14 (CEP290)</td>
<td>BBS15 (WDPCP)</td>
</tr>
<tr>
<td>BBS16 (SDCCAG8)</td>
<td>BBS17 (LZTFL1)</td>
<td>BBS18 (BBIP1)</td>
</tr>
<tr>
<td>BBS19 (IFT27)</td>
<td>BBS20 (IFT74)</td>
<td>BBS21 (C8ORF37)</td>
</tr>
<tr>
<td></td>
<td>BBS22 (SCAPER)</td>
<td></td>
</tr>
</tbody>
</table>

The primary phenotypic features of BBS are retinal degeneration, postaxial polydactyly, obesity, hypogonadism, renal anomalies, and cognitive impairment\(^159,160\) \(^149\)-\(^151\). Secondary phenotypic features include anosmia, strabismus, cataracts, astigmatism, brachydactyly/syndactyly,
orodental abnormalities, craniofacial dysmorphism, diabetes mellitus, ataxia, mild hypertonia, cardiovascular anomalies, hepatic involvement, Hirschsprung disease, developmental delay, behavior abnormalities, and speech delay\textsuperscript{2,10,11}. A clinical diagnosis of BBS is defined as having four primary features or three primary and two secondary features\textsuperscript{155}.

Retinal degeneration, obesity, and polydactyly are the most common features of BBS patients. While polydactyly is present in approximately 95\% of BBS patients\textsuperscript{161}, it occurs in approximately 1:700-1000 births within the general population\textsuperscript{162} and is easily remedied at birth. The majority (~72-92\%) of BBS patients develop obesity in early childhood with it worsening throughout their life\textsuperscript{155}. BBS is the second most common syndromic retinal degeneration and has been extensively studied\textsuperscript{7,163}. Often, patients are diagnosed with retinitis pigmentosa (RP) before they are 10 years old and reach legal blindness between their third and fourth decade of life\textsuperscript{164}. The first visual complaint is often night blindness (nyctalopia) followed by loss of peripheral vision resulting in a constricted visual field\textsuperscript{165-167}. The decrease in peripheral vision is caused by the progressive loss of the rod photoreceptors\textsuperscript{168}. As the disease progresses, the patient also experiences death of cone photoreceptors, which causes a decrease in visual acuity and a reduction of color vision\textsuperscript{168}. Unlike standard RP, BBS patients are usually considered to have an early macular involvement\textsuperscript{158,165,166,168,169} (Figure 15).
By performing segregation analysis of large pedigrees and consanguineous families the autosomal recessive pattern of inheritance in BBS has been firmly established. It has been reported that homozygous or compound heterozygous mutations in \textit{BBS1}, \textit{BBS10}, \textit{BBS2}, \textit{BBS9}, and \textit{MKKS (BBS6)} cause the majority of BBS cases\textsuperscript{7,155}. However, there is some dispute regarding the pattern of inheritance and the possibility of a “triallelic” inheritance pattern caused by three different mutant alleles at two different loci\textsuperscript{170}. Three pathogenic alleles were identified in \textit{BBS2} and \textit{BBS6} within BBS patients along with unaffected individuals carrying two \textit{BBS2} variations. These data led to the proposal of triallelic inheritance within BBS patients, prompting other groups to interrogate their own cohorts for evidence supporting or contradicting this hypothesis. These studies failed to yield evidence supporting triallelic inheritance, but provided further evidence supporting the autosomal recessive inheritance pattern with \textit{BBS}\textsuperscript{171-173}. 

Figure 15. Fundus Photographs of BBS Patient.

Right (20/100 BCVA) and left (20/80 BVCA) fundus photographs of a 47-year old male with BBS. Patient was molecularly confirmed to have a homozygous M390R mutation in \textit{BBS1}. The patient presents with macular atrophy and peripheral bone spicule pigmentation. Figure courtesy of Medical Research Archives\textsuperscript{17} and Ed Stone. BCVA = Best Corrected Visual Acuity.
While, there is a lack of further evidence supporting triallelic inheritance, it has been proposed that increased mutational load (e.g., a third mutation in a different BBS gene) can result in a more severe phenotype in BBS patients\(^1\). Multiple BBS genes have been reported to cause other diseases besides BBS and it is possible that mutations in a second BBS gene may further modify the phenotype. This has been shown in both animal models and human data with a \(Bbs14\) (\(Cep290\)) homozygous mutant mouse displaying increased obesity and an accelerated rate of retinal degeneration when \(Bbs4\) has a single functional allele\(^2\). Homozygous mutations in \(BBS7\) were identified within affected members of a BBS family, with a heterozygous null mutation also being identified in a member of the family with a more severe phenotype\(^3\). These data highlight the importance of continuing to study these genes to further elucidate their molecular functions and interactions with other genes.

A considerable amount of effort has been made to understand the molecular and cellular function of BBS proteins. Eight of the BBS proteins (\(BBS1, BBS2, BBS4, BBS5, BBS7, BBS8, BBS9,\) and \(BBS18\)) form a complex known as the BBSome, which is important to protein trafficking within cilia and intracellularly\(^4\). Three other BBS proteins (\(BBS6, BBS10, BBS12\)) are crucial to a chaperonin complex that is required for proper BBSome assembly\(^5\). \(BBS3, BBS14,\) and \(BBS17\) allow for the BBSome to move to and within the cilia\(^6\). The rest of the BBS gene products (\(BBS11, BBS13, BBS15-16, BBS19-21\)) are not closely associated with the BBSome and its proper function at this point. The role these genes play in cilia have yet to be fully characterized and are the subject of further investigation.
There is also evidence that BBS proteins are involved with proper protein trafficking within the connecting cilium and outer segment (OS) of photoreceptors\textsuperscript{203,204} thus leading to photoreceptor disease when BBS proteins are not fully functional. Previous animal models of BBS with retinal phenotypes have rhodopsin localization defects with the photoreceptors. In a healthy eye, rhodopsin is localized to the OS of rod photoreceptors and performs the first step in phototransduction. In BBS mutant mouse models, rhodopsin has severely diminished levels in the OS, but is instead localized to the inner segment (IS) and cell bodies of the outer nuclear layer\textsuperscript{203,205-208} (ONL). Study of the retina in BBS mouse models has also shown OS structural anomalies and the progressive loss of the OS, IS, and ONL while the inner retinal layers remain normal and functioning\textsuperscript{205,209}.

Further evidence supporting the trafficking role of BBS proteins within the OS was provided by Datta \textit{et al.} by investigating the OS proteome in \textit{Bbs17} mutant mice\textsuperscript{204}. The researchers showed an enrichment in 138 proteins within the OS of the mutant mice, whereas only 8 proteins showed decreased OS localization. Analysis of wild type (WT) mice showed not only a lack of enrichment for many of these proteins within the OS but a complete absence of many of the identified proteins within the OS of WT mice. Many of the proteins identified in the mutant OS are typically found within the IS, ONL, and synaptic terminal where they are involved with retinal function\textsuperscript{204}. This evidence suggests that BBS proteins are required for proper OS protein trafficking and that photoreceptor cell death in BBS mutants may be caused by an accumulation of proteins within the OS and the decrease in those proteins function within the IS, ONL, and synaptic terminal.
Further research supports the role of BBS proteins and trafficking by studying the effects mutant BBS proteins have on the leptin receptor\textsuperscript{210} and the insulin receptor\textsuperscript{211}. Mislocalization of ciliary receptors has also been associated with BBS mutations\textsuperscript{192,212-214}. The majority of these receptors are G-protein coupled receptors that are critical to cAMP and phosphatidylinositol signaling\textsuperscript{215}. While the knowledge gap has been reduced concerning BBS and why mutations in BBS genes cause disease, there is still much to learn about the molecular function of these proteins and how we can develop possible treatments for BBS patients.

A primary focus of the Sheffield lab for more than 25 years has been to identify novel BBS genes. Over the years the lab has independently identified ten novel BBS genes and has been instrumental in studying the disease through cell, mouse, and zebrafish models\textsuperscript{182,183,195,202,216,217}. Mutations in $BBS1$ and $BBS10$ account for approximately 40\% of all BBS cases, with the other 19 genes accounting for another 40\%, leaving 15-20\% of BBS cases without a molecular diagnosis\textsuperscript{155,218}. Knowing this, we embarked on multiple different BBS projects using next-generation sequencing (NGS) to identify novel BBS mutations and genes.

**Project 1.** With advances in next-generation sequencing analysis tools we re-analyzed 26 previously unsolved/solved BBS cases.

**Project 2.** Whole exome sequencing analysis of a Mexican patient diagnosed with BBS.

**Project 3.** Whole exome sequencing analysis of a Puerto Rican BBS cohort.
Materials and Methods

**DNA Sequencing**

Whole exome sequencing was performed on patients’ DNA as previously described\(^{219}\). Sonification was used to fragment the DNA, followed by library preparation and sample bar-coding performed with Agilent or Nimblegen kits. Paired-end sequencing was performed using an Illumina HiSeq platform to obtain a depth of at least 80x over approximately 90% of the targeted regions.

**Sequence Alignment and Processing**

DNA sequence reads were first converted to the SAM file format, followed by removal of the Illumina sequencing adapters using the MarkIlluminaAdapter function in the Picard tools package (http://broadinstitute.github.io/picard). Sequences were aligned to the human genome reference genome GRCh37 using the Burrows-Wheeler aligner,\(^ {49}\) and duplicate reads were removed using the MarkDuplicates function in Picard tools. Indel realignment and base quality score recalibration were performed using GATK3\(^ {51}\) resulting the final variant call ready alignment files..

**Calling and Annotating Variants**

Variants were called for each individual using the GATK Haplotype Caller\(^ {51}\). Joint genotyping on each cohort was performed using the GATK GenotypeGVCFs tool\(^ {51}\), to produce the best genotyping results. Individual samples were then normalized and separated into their own VCF
file using BCFtools\textsuperscript{53}. Variants were annotated with observed allele frequencies from the 1000 Genomes Project\textsuperscript{132}, dbSNP\textsuperscript{133}, and the genome aggregation database\textsuperscript{54} (gnomAD). Interpretations of the consequence of the variation on the gene in which it appeared were also annotated with our in-house sequence analysis and annotation pipeline (Institute for Vision Research, Iowa City, IA).

**Variant Filtering**

Variants with a minor allele frequency of greater than 1% in the 1000 Genomes project\textsuperscript{132} or gnomAD\textsuperscript{54} were judged to be too common to cause BBS and were excluded from the analysis. Variants with a GATK quality score of less than 25, or a quality by depth score of less than 1 were also excluded from the analysis.

**Variant Prioritization**

Variants were analyzed using a tiered testing strategy to reduce the false genotype rate. Tier one analysis only analyzed variants in known BBS genes. If tier one analysis yielded an incomplete genotype, tier two analysis included variants from known eye disease-causing genes with prioritization given to known ciliopathy-causing genes. Variants were analyzed for disease prediction using PolyPhen2\textsuperscript{67}, SIFT\textsuperscript{69}, and CADD\textsuperscript{70} and allele frequency within gnomAD and our own cohort of over 1500 exomes.
**Experimental Confirmation**

Candidate variants identified by exome sequencing were confirmed to exist in the proband by Sanger sequencing\textsuperscript{134}. When possible, candidate variants were also confirmed through Sanger sequencing within family members to verify segregation of the variant with disease in the family. Candidate variants identified within ethnically matched controls through visual inspection using the Integrative Genome Viewer\textsuperscript{61} (IGV) were regarded as non-disease causing and removed from the candidate list.

**Project 1: Reanalysis of 26 BBS patients**

**Results**

**Whole Exome Sequencing Results**

To determine the efficacy of our new exome sequencing pipeline and identify new candidate variants we reanalyzed 26 BBS patients’ whole exome sequencing data using our latest NGS pipeline. BBS samples were sent to the Sheffield lab with research consent by referring clinicians across the country. All samples were screened for the most common mutations in \textit{BBS1}, \textit{BBS10}, and \textit{BBS7} prior to whole exome sequencing. The majority of BBS samples sent to the lab are molecularly confirmed through this screen (90%) and do not need whole exome sequencing.

All samples sent for whole exome sequencing were negative for common BBS mutations in \textit{BBS1}, \textit{BBS10}, and \textit{BBS7}. Candidate mutations within 16 of the 26 patients (61.5\%) sent for
whole exome sequencing were identified using our NGS pipeline. New candidate BBS mutations were identified in 9 of the 26 patients (34.6%). We were unable to identify candidate variants within BBS genes in 9 of the BBS cases (34.6%) However, we did identify two different variants in *BBS10* in a previously unsolved patient giving them a complete BBS genotype (Table 12). We also found a homozygous variant in *BBS7* that causes a premature termination in exon 6 (Table 12). Our NGS pipeline identified all mutations that were previously called and molecularly confirmed within our 9 solved cases.

Table 12. BBS10 and BBS7 variants within previously unsolved cases.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Genotype</th>
<th>Genomic Mutation</th>
<th>Protein Mutation</th>
<th>PolyPhen2</th>
<th>SIFT</th>
<th>gnomAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBS10</td>
<td>Heterozygote</td>
<td>chr12:76739717:C&gt;T</td>
<td>p.Gly683Asp</td>
<td>possibly damaging</td>
<td>tolerated</td>
<td>7.96E-06</td>
</tr>
<tr>
<td>BBS10</td>
<td>Heterozygote</td>
<td>chr12:76741994:G&gt;A</td>
<td>p.Arg49Trp</td>
<td>probably damaging</td>
<td>deleterious</td>
<td>5.37E-05</td>
</tr>
<tr>
<td>BBS7</td>
<td>Homozygote</td>
<td>Chr4:122776691:TC-&gt;T</td>
<td>p.Glu185LysfsTer9</td>
<td>N/A</td>
<td>N/A</td>
<td>Not reported</td>
</tr>
</tbody>
</table>

Eight of the previously unsolved cases had candidate variants in genes known to cause other diseases that share phenotypes with BBS (Table 13 and Table 14). Sanger sequencing was performed to confirm the presence or absence of each the reported variants.
Table 13. Reported variants from diseases with shared BBS phenotypes.

<table>
<thead>
<tr>
<th>Patient</th>
<th>Gene</th>
<th>Genotype</th>
<th>Complete Genotype</th>
<th>Protein Change</th>
<th>Reported Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
<td>ALMS1</td>
<td>Homozygote</td>
<td>X</td>
<td>p.Asn3085Thrfs7er19</td>
<td>Alstrom Syndrome</td>
</tr>
<tr>
<td>Patient 1</td>
<td>BBS12</td>
<td>Heterozygote</td>
<td></td>
<td>p.Phe22Ser</td>
<td>Bardet-Biedl Syndrome</td>
</tr>
<tr>
<td>Patient 1</td>
<td>CEP290</td>
<td>Heterozygote</td>
<td></td>
<td>p.Asp1413His</td>
<td>Bardet-Biedl Syndrome</td>
</tr>
<tr>
<td>Patient 2</td>
<td>KIAA0586</td>
<td>Heterozygote</td>
<td></td>
<td>p.Glu301Stop</td>
<td>Joubert Syndrome 23, Short-rib thoracic dysplasia 14 with polydactyly</td>
</tr>
<tr>
<td>Patient 2</td>
<td>KIAA0586</td>
<td>Homozygote</td>
<td>X</td>
<td>p.Ser1010Leu</td>
<td>Joubert Syndrome 23, Short-rib thoracic dysplasia 14 with polydactyly</td>
</tr>
<tr>
<td>Patient 3</td>
<td>IFT140</td>
<td>Compound Heterozygote</td>
<td>X</td>
<td>p.Ala820Thr, p.311Stop</td>
<td>RP 80, Short-rib thoracic dysplasia 9 with or without polydactyly</td>
</tr>
<tr>
<td>Patient 4</td>
<td>CDH23</td>
<td>Heterozygote</td>
<td></td>
<td>p.Arg1771Gln</td>
<td>Usher Syndrome</td>
</tr>
<tr>
<td>Patient 4</td>
<td>USH1G</td>
<td>Heterozygote</td>
<td></td>
<td>p.Arg189Gln</td>
<td>Usher Syndrome</td>
</tr>
<tr>
<td>Patient 4</td>
<td>HMCN1</td>
<td>Heterozygote</td>
<td></td>
<td>p.Arg1522Cys</td>
<td>Macular Degeneration</td>
</tr>
<tr>
<td>Patient 5</td>
<td>HMCN1</td>
<td>Heterozygote</td>
<td></td>
<td>p.Gly3566Asp</td>
<td>Macular Degeneration</td>
</tr>
<tr>
<td>Patient 6</td>
<td>CDH23</td>
<td>Compound Heterozygote</td>
<td>X</td>
<td>p.Ser436Asn, p.Asn1098Ser</td>
<td>Usher Syndrome</td>
</tr>
<tr>
<td>Patient 7</td>
<td>ANTXR1</td>
<td>Homozygote</td>
<td>X</td>
<td>p.Arg480Cys</td>
<td>Gapo Syndrome</td>
</tr>
<tr>
<td>Patient 8</td>
<td>WFS1</td>
<td>Heterozygote</td>
<td></td>
<td>p.Glu394Val</td>
<td>Wolfram Syndrome</td>
</tr>
</tbody>
</table>
Table 14. Overlapping phenotypic features between BBS and other diseases.

<table>
<thead>
<tr>
<th></th>
<th>Retinal Degeneration</th>
<th>Obesity</th>
<th>Postaxial Polydactyly</th>
<th>Cognitive Impairment</th>
<th>Renal anomalies</th>
<th>Hypogonadism</th>
<th>Ataxia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alstrom Syndrome</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Joubert Syndrome</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRTD9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Usher Syndrome</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gapo Syndrome</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wolfram Syndrome</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Copy Number Variant Results

By integrating the copy number variant (CNV) caller CLAMMS\textsuperscript{66} into our NGS pipeline we were able to identify two patients with homozygous deletion of exons 3 and 4 in $BBS4$(Figure 16). This mutation was first identified in 2001 by our lab in two distinct families, one from Italy and one from Israel\textsuperscript{181}. This variant was previously unobservable due to technical limitations of the computational model employed by CoNIFER, the CNV caller used in earlier versions of our exome analysis pipeline.
Figure 16. Homozygous deletion of exons 3 and 4 in two BBS patients. Shown is the total read coverage for a region in BBS4. The top and bottom tracks shown are BBS patients with homozygous deletion of exons 3 and 4 in *BBS4* as seen by the lack of reads in that section. The middle track is another BBS patient without the deletion.

**Discussion**

**Complete BBS Genotypes**

With our primary analysis, we identified a total of 9 new mutations in BBS genes previously unidentified through our exome sequencing analysis. As mentioned previously, these contributed to identifying complete BBS genotypes in four previously unsolved BBS patients. One of the patients has compound heterozygote mutations in *BBS10* that are extremely rare and predicted to damage the protein. The second patient has a homozygous deletion of a single nucleotide causing a frameshift and premature stop in exon 6 of *BBS7*. This mutation has not been reported in either gnomAD\(^\text{54}\) or ClinVar\(^\text{220}\) and terminates the protein earlier than most of the reported variants in *BBS7*. The final two patients have homozygous deletions of exons 3 and 4 in *BBS4*, which has been reported to cause disease. Given this information, we believe that the cause of disease in these four patients has been firmly established.
Incomplete BBS Genotypes

Five patients were classified with an incomplete or unconvincing BBS genotype by our reanalysis. BBS is an autosomal recessive disease and requires two different non-functional alleles of the same gene to cause disease presentation. Two of the patients had mutations in a single allele which alone doesn’t cause disease. These patients are great candidates for additional investigation to ensure the coding regions of the putative disease-causing genes are fully covered, followed by whole genome sequencing to determine if a second mutation can be identified in the non-coding portion of those genes.

Another patient carries single allele mutations in BBS1 and BBS10. While we are currently calling this patient as unsolved, there is evidence in zebrafish that partial knockdown of two BBS genes in combination can cause phenotypic effects. However, only interactions between BBS1-8 were studied so further testing would be required. This patient is also a good candidate for whole genome sequencing screen for large-scale changes (e.g. rearrangements, inversions) in BBS1 or BBS10, resulting in a complete autosomal recessive genotype.

In a seventh patient, a homozygous variant was identified within an intron of BBS5. We do not believe this variant causes disease because it is outside of the canonical splice region and would likely have no effect on the protein structure. Finally, a patient (Patient 1) was found to have an incomplete BBS genotype with single allele mutations in BBS12 and CEP290 (BBS14). But this patient also has a homozygous variant in ALMS1 which causes a premature stop mutation in the protein likely causing Alstrom syndrome. This patient will be discussed further in the next section.
Rare Mutations in Genes with BBS-associated Phenotypes

Patient 1 was particularly interesting, as they have a homozygous single nucleotide deletion causing a premature termination within exon 10 of the \textit{ALMS1} gene, resulting in the lack of the ALMS motif from the protein. Alstrom syndrome (ALMS) is an autosomal recessive disorder with many similarities to BBS. ALMS patients have progressive cone-rod dystrophy, childhood obesity, type 2 diabetes mellitus, sensorineural hearing loss, cardiomyopathy, renal failure, and dysfunction of the pulmonary, hepatic and urologic systems\textsuperscript{222,223}. However, phenotypic features of ALMS patients do not include cognitive deficits, polydactyly, or hypogonadism\textsuperscript{224}. While BBS patients first experience peripheral vision loss, ALMS patients typically first notice central vision loss due to a retinal lesion. With many phenotypic features shared between the two disorders it can be a challenge for physicians to correctly distinguish between BBS and ALMS. With the homozygous premature stop in \textit{ALMS1}, which has been confirmed through Sanger sequencing, it is likely that this patient actually has ALMS instead of BBS.

We identified two different mutations in Patient 2 that highly suggest that the patient has Joubert syndrome instead of BBS. The patient carries a homozygous p.Ser1010Leu and heterozygous p.Glu301Stop in \textit{KIAA0586 (JBTS23)}. The p.Glu301stop mutation has not been reported in gnomAD and the p.Ser1010Leu is very rare with an allele frequency of 6.06E-05. PolyPhen2 and SIFT both predict the p.Ser1010Leu mutation to be damaging to the protein. Joubert syndrome and BBS are both ciliopathies with multiple shared phenotypes. Features of Joubert syndrome include the “molar tooth sign” (a radiographic abnormality of the cerebellum), hypotonia, abnormal eye movements, ataxia, mild to severe developmental delay, retinal degeneration, cystic kidney disease, nephronophthisis, polydactyly, and endocrine defects\textsuperscript{225-233}. At this time there are at least 34 genes associated with Joubert syndrome including two BBS genes\textsuperscript{234}. With
so many overlapping phenotypes between the two syndromes, it is possible that the patient has Joubert syndrome and was incorrectly diagnosed as BBS, or this gene also causes BBS similar to $\textit{MKS1}$ ($\textit{BBS13}$) and $\textit{CEP290}$ ($\textit{BBS14}$). To address this, we have reached out to the clinician that sent us this sample for more clinical information.

The analysis of Patient 3 identified a compound heterozygote mutation in $\textit{IFT140}$ which is known to cause RP and Short-rib thoracic dysplasia 9 with or without polydactyly (SRTD). Neither the p.Ala820Thr or the p.Tyr311Stop mutation were found in gnomAD, however the p.Ala820Thr mutation is predicted as probably damaging/deleterious by PolyPhen2 and SIFT. The alanine and surrounding 3 amino acids on either side are conserved through every species on HomoloGene$^{147}$ which supports the deleterious classification. Multiple publications have reported $\textit{IFT140}$ as the causative gene for SRTD with varying severity of the phenotypes within patients$^{235-237}$, although each patient had renal findings and RP. $\textit{IFT140}$ has also been associated with non-syndromic RP$^{238,239}$. Without having access to the patient’s clinical findings, we cannot definitively state whether or not this patient truly has SRTD or non-syndromic RP, though the genotype strongly supports this diagnosis.

Patient 4 carries single allele mutations in $\textit{CDH23}$, $\textit{USH1G}$, and $\textit{HMCN1}$. $\textit{CDH23}$ and $\textit{USH1G}$ are both known to cause Usher syndrome which causes early RP and mild to severe hearing loss. However, Sanger sequencing failed to confirm the presence of the $\textit{CDH23}$ mutation. $\textit{USH1G}$ has not been demonstrated to be the cause of the patient’s phenotype since it contains only one heterozygous variant and Usher syndrome has an autosomal recessive inheritance pattern requiring homozygous or compound heterozygous variants in the same gene. It is also unlikely
that \textit{HMCNI} is causing the patients disease because mutations in this gene cause macular degeneration instead of RP.

In patient 6, I identified compound heterozygous mutations in CDH23, which is known to cause Usher syndrome type 1D\textsuperscript{240,241}. Usher syndrome is the most common cause of syndromic RP\textsuperscript{7} and is inherited in an autosomal recessive fashion. Usher syndrome type I is characterized by congenital hearing loss accompanied by vestibular dysfunction and the development of RP within the first decade of life\textsuperscript{242}. Both mutations, p.Ser436Asn and p.Asn1098Ser, were predicted to be damaging to the protein according to PolyPhen2 and SIFT\textsuperscript{67,69}. These variants were reported with allele frequencies of 0.0038 and 0.0049 making them rare enough to be considered for recessive disease. Neither of these mutations are reported as disease-causing in ClinVar. Most importantly, Sanger sequencing of the variants failed to confirm their existence in this patient, moving this patient back into the unsolved group.

Finally, we identified a homozygous mutation in \textit{IQCB1} in patient 7. Mutations in \textit{IQCB1} cause Senior-Loken syndrome\textsuperscript{243}, a ciliopathy with renal and retinal abnormalities shared with BBS. However, even though the p.Ala309Ser variant has a minor allele frequency of 0.002476, it is predicted to be benign and most likely not the cause of disease in this patient. This patient also carries a homozygous p.Arg480Cys mutation in \textit{ANTXR1}, which causes Gapo syndrome\textsuperscript{244}. This mutation is predicted to be damaging by PolyPhen2 and SIFT and is highly conserved. This mutation is also reported to be very rare with a minor allele frequency of 4.64E-05 in gnomAD. Gapo syndrome is characterized by optic atrophy, hypogonadism, minor skeletal abnormalities, growth retardation, alopecia, cardiomyopathy, and can also present with glaucoma and
keratoconus²⁴⁵-²⁵⁰. While there is variability in the presentation of cardinal features of Gapo syndrome, it is unlikely that a patient diagnosed with BBS actually has Gapo syndrome.

**Conclusion**

The reanalysis of 26 putative BBS patients clinically diagnosed with our updated NGS pipeline led us to identify complete genotypes in four previously unsolved patients. We also identified mutations in genes with BBS-like phenotypes in 8 additional patients. These data reaffirm the need for high-quality clinical information being ascertained by the referring clinician and then passed on to the bioinformatician performing the analysis to reduce the false genotype rate and chance of misdiagnosing the causative disease-gene for the patient. As NGS cost continues to decrease and become a more readily available tool for clinicians it is critical that we continue to practice hypothesis-based testing to correctly identify the cause of disease within patients.

**Project 2: Whole exome sequencing analysis of a Mexican BBS patient**

**Results**

**BBS Patient Background**

A clinical geneticist in Mexico sent the Sheffield lab a DNA sample from a nine-year-old girl presenting with congenital nystagmus, obesity, postaxial polydactyly (left foot) and mild RP. The patient presented with normal intelligence and excelled in school. The patient has unaffected parents and an unaffected sibling.
**BBS Screening Panel Results**

The referring clinician asked us to screen *BBS3* and *BBS4* due to reports of normal intelligence in cases with *BBS3* and *BBS4* mutations. We first performed our normal screen looking for the most common mutations in *BBS1, BBS10, and BBS7*. After the screen returned negative results, the sample was sent for whole exome sequencing.

**Whole Exome Sequencing Results**

Whole exome sequencing analysis identified two heterozygous missense mutations in the BBS gene *ARL6 (BBS3)* (Table 15).

<table>
<thead>
<tr>
<th>Gene</th>
<th>Coding Change</th>
<th>Protein Change</th>
<th>Observed/Covered</th>
<th>Max gnomAD</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>ARL6</em></td>
<td>c.529C&gt;T</td>
<td>p.Leu177Phe</td>
<td>0.42</td>
<td>0.023%</td>
</tr>
<tr>
<td><em>ARL6</em></td>
<td>c.499G&gt;A</td>
<td>p.Gly167Arg</td>
<td>0.565</td>
<td>0.0033%</td>
</tr>
</tbody>
</table>

Both variants identified are very rare according to gnomAD and required further exploration. The c.529C>T variant was predicted by PolyPhen2 to be possibly damaging and SIFT to be deleterious. The c.499G>A variant was predicted by PolyPhen2 to be probably damaging and SIFT to be deleterious. Both variants were confirmed with Sanger sequencing of the patient DNA and using IGV, we were able to show that each variant came from one parent (Figure 17). We also had access to parental and sibling DNA to confirm with Sanger sequencing that each parent carried a single variant and that the unaffected sibling only carried one of the variants (Figure 18). Neither of these variants have been reported in the literature or in the ClinVar
database, though variants at amino acid position 169 and 170 have been previously shown to be disease-causing\textsuperscript{100,251}.

Figure 17. Visualization of variant phase using IGV. Using IGV, we were able to show that each parent carried one of the mutations before Sanger sequencing confirmation. Each sequencing read shown that overlaps both variants location provides evidence supporting only one of the variants.
Figure 18. Pedigree depicting the segregation of the two BBS3 mutations. Sanger sequencing confirmed the carrier status for the L177F allele in the unaffected mother and the G167R in the unaffected father of the proband. Segregation through the family was fully confirmed when the unaffected brother of the proband was confirmed to carry only the L177F allele.

Discussion

Previously Unreported Variants in BBS3 Cause BBS

Our lab originally identified the locus for and subsequently the gene, *ARL6 (BBS3)*, which is a cause of BBS\(^{243,244}\). Since then, *BBS3* has been studied extensively with over 30 variants being reported in ClinVar\(^{244-249}\). Through whole exome sequencing analysis, we were able to identify two novel mutations in *ARL6* that cause BBS. The ability to confirm segregation within affected and unaffected members of the probands family was critical to establish confidence in the genotype to phenotype diagnosis. This project and its findings reiterate the power of procuring familial DNA when studying genetic disease. The ability to confirm the segregation of mutations within a family, following the correct mode of inheritance, provides very strong evidence that
causative mutations have been identified. These data can then be used to research and develop therapeutic treatments for genetic disease.

Project 3: Whole exome sequencing analysis of Puerto Rican BBS patients

Results

BBS Screen Panel Results
A collaborator from Puerto Rico sent us 10 patients clinically diagnosed with BBS based on standard clinical criteria\textsuperscript{155} to sequence for mutations in BBS genes. The specific clinical phenotypes for the patients were not sent to us at the time of testing. Using our standard screening panel, we identified one sample with a homozygous M390R mutation in \textit{BBS1}. We also identified another sample with a heterozygous M390R mutation in \textit{BBS1}. No other mutations on our screening panel were identified within the Puerto Rican cohort.

Whole Exome Sequencing Results
Nine samples were sent for whole exome sequencing at the University of Iowa DNA core. Heterozygous mutations in BBS genes were identified in 5 of the 9 samples within the cohort (Table 16). All 5 of the samples were reported with an incomplete genotype in BBS genes prior to CNV analysis.
Table 16. Variants in BBS genes.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Gene</th>
<th>Variant</th>
<th>Genotype</th>
<th>gnomAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>PR-1</td>
<td>BBS1</td>
<td>Met390Arg</td>
<td>Heterozygous</td>
<td>0.001570</td>
</tr>
<tr>
<td>PR-2</td>
<td>IFT74 (BBS20)</td>
<td>Met222Thr</td>
<td>Heterozygous</td>
<td>0.0005024</td>
</tr>
<tr>
<td>PR-3</td>
<td>MKKS (BBS6)</td>
<td>c.1161+3A&gt;G</td>
<td>Heterozygous</td>
<td>0.00003892</td>
</tr>
<tr>
<td>PR-4</td>
<td>MKKS (BBS6)</td>
<td>c.1161+3A&gt;G</td>
<td>Heterozygous</td>
<td>0.00003892</td>
</tr>
<tr>
<td>PR-6</td>
<td>BBS1</td>
<td>Glu549Stop</td>
<td>Heterozygous</td>
<td>0.00001591</td>
</tr>
</tbody>
</table>

Copy Number Variant Analysis

Using the copy number variant caller CLAMMS\textsuperscript{66}, we identified a homozygous deletion of exons 7 and 8 in \textit{GBP3} in samples PR-2, PR-3, and PR-4 (Figure 19). However, Sanger sequencing was able to amplify the entire region in questions, indicating that the 2-exon deletion was spurious. Further analysis of the region using the UCSC genome browser revealed a large number of reported structural variants within the region supporting the variability in our sequencing results.
Whole Exome Sequencing Findings in Eye-disease Genes

We were unable to identify a complete genotype in BBS genes within our PR BBS samples. We expanded our search to our list of known eye-disease genes. We identified 3 samples with at least 2 mutant alleles in the retinal cone dystrophy gene, $KCNV2^{252-254}$. Two of our samples were homozygous for p.Lys260Stop, a known disease-causing mutation, with another sample being heterozygous for the same mutation (Table 17). We also identified the $KCNV2$ p.Asp140Glu mutation in each of the aforementioned samples. Sanger sequencing confirmed the presence of each of these mutations within their respective samples.
Table 17. Mutations in KCNV2 within our Puerto Rican BBS cohort.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Genotype</th>
<th>Gene</th>
<th>Coding Change</th>
<th>Protein Change</th>
<th>gnomAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>PR-3</td>
<td>Homozygote</td>
<td>KCNV2</td>
<td>c.420C&gt;A</td>
<td>p.Asp140Glu</td>
<td>0.000004079</td>
</tr>
<tr>
<td>PR-3</td>
<td>Homozygote</td>
<td>KCNV2</td>
<td>c.778A&gt;T</td>
<td>p.Lys260Stop</td>
<td>0.00003335</td>
</tr>
<tr>
<td>PR-4</td>
<td>Homozygote</td>
<td>KCNV2</td>
<td>c.420C&gt;A</td>
<td>p.Asp140Glu</td>
<td>0.000004079</td>
</tr>
<tr>
<td>PR-4</td>
<td>Homozygote</td>
<td>KCNV2</td>
<td>c.778A&gt;T</td>
<td>p.Lys260Stop</td>
<td>0.00003335</td>
</tr>
<tr>
<td>PR-9</td>
<td>Heterozygote</td>
<td>KCNV2</td>
<td>c.420C&gt;A</td>
<td>p.Asp140Glu</td>
<td>0.000004079</td>
</tr>
<tr>
<td>PR-9</td>
<td>Heterozygote</td>
<td>KCNV2</td>
<td>c.778A&gt;T</td>
<td>p.Lys260Stop</td>
<td>0.00003335</td>
</tr>
</tbody>
</table>

Discussion

Incomplete BBS Genotypes

We identified one allele findings in 5 of our PR samples. Samples PR-1 and PR-6 both have heterozygous known disease-causing mutations in BBS1. Both of these samples are prime candidates for whole genome sequencing to identify a non-coding mutation in BBS1 that would cause both copies of the protein to be defective. Sample PR-2 has a single p.Met222Thr mutation in IFT74 (BBS20) that is predicted as possibly damaging by PolyPhen2 and deleterious by SIFT. This mutation has an allele frequency of 0.0003112 in the Latino population of gnomAD and would be considered rare enough to cause BBS. Further investigation of the mutation and its effects on cilia in human cells are needed to determine its pathogenicity. This sample is also a good candidate for whole genome sequencing in search of a second, functional but non-coding, mutation in IFT74 to complete the genotype for an autosomal recessive mode of inheritance. Samples PR-3 and PR-4 have heterozygous c.1161+3A>G variants in MKKS (BBS6) that may affect splicing. At this time, we believe these variants are benign and that mutations in KCNV2 are the cause of the retinal phenotype in PR-3 and PR-4.
**KCNV2 – A New BBS Gene?**

Through our whole exome sequencing analysis, we identified two samples (PR-3 and PR-4) with 2 homozygous mutations and another patient (PR-9) with compound heterozygous mutations in *KCNV2*. It is likely that samples PR-3 and PR-4 are related based upon the sample names provided by our collaborator. We have not been able to confirm this or collect further phenotypic information regarding these samples due to challenges in communication with our collaborator since Hurricane Maria in Puerto Rico. The p.Lys260Stop is a mutation reported to cause retinal cone dystrophy 3B\(^{253}\) (RCD3B) also known as cone dystrophy with supernormal rod response (CDSRR). Characteristics of RCD3B include supernormal rod responses, photophobia, myopia, astigmatism, reduced color vision, nystagmus, development of nyctalopia (night-blindness) in later life and abnormal electroretinography (ERG) responses.

While two of our samples have two mutant p.Lys260Stop alleles, they also have two mutant p.Asp140Glu alleles which have not been previously reported. Sample PR-9 has one allele findings for both the p.Lys260Stop and p.Asp140Glu mutations potentially giving them a complete genotype if the p.Asp140Glu mutation is pathogenic and the two mutations are on separate alleles. The putative mutation is predicted to be damaging to the protein and the variant is incredibly rare with a maximum allele frequency of 0.00002940 in the Latino population. This putative mutation has not been observed in any of our prior exomes, therefore we performed Sanger sequencing in our PR samples to determine if this putative mutation is an ethnic polymorphism and found no other samples with the mutations.
BBS eye phenotypes also include reduced color vision, nystagmus, nyctalopia, abnormal ERG response and loss of visual acuity similar to RCD3B. Historically, this collaborator sent us large BBS families with accurately diagnosed BBS patients that led to the identification of BBS loci (BBS1 and BBS7), making it likely these are truly BBS patients. Receiving the clinical phenotypic information such as the presence of polydactyly would provide further support that \( KCNV2 \) also causes BBS. Multiple BBS genes have been reported to cause non-syndromic RP such as \( C8orf37 \) (BBS21), which was originally known to only cause non-syndromic RP before being identified as a BBS gene\(^{202,255,256} \). As we follow up with the referring clinician and perform cell studies, it is possible that \( KCNV2 \) is the next BBS gene to be identified.

**Conclusion**

In this chapter we discussed three different BBS projects that utilized whole exome sequencing to identify candidate disease-causing mutations. We identified complete genotypes in BBS genes and known eye-disease genes in our samples. With the continued decrease in cost for next-generation sequencing, it is critical that we continue to have high quality clinical phenotyping and a pre-existing hypothesis. We were able to identify the causative mutations in BBS genes for multiple patients due to a pre-existing hypothesis. We also identified and confirmed novel mutations in \( ARL6 \) (BBS3) that cause the patients phenotype by leveraging the power of segregation analysis within the affected patient’s family. We also established \( KCNV2 \) as a candidate BBS gene, awaiting further clinical information which may support or contradict the BBS diagnosis within the three patients with \( KCNV2 \) mutations.
CHAPTER 5: A CASE/CONTROL WHOLE EXOME SEQUENCING STUDY TO INVESTIGATE GENES INVOLVED IN PIGMENT DISPERSION SYNDROME

Abstract

Purpose. Pigment Dispersion Syndrome (PDS) is an eye condition affecting approximately 2.5% of the population of the United States. Pigment escapes from the iris, gathers within the trabecular meshwork and causes an increase in intraocular pressure (IOP). Elevated IOP leads to optic nerve damage and permanent vision loss. Although there is evidence for a genetic cause of PDS, researchers have yet to identify a gene responsible for disease presentation. To address this critical knowledge gap, we assembled 210 PDS subjects and 362 control subjects, and performed the largest next-generation sequencing (NGS) study of PDS to date.

Methods. Whole exome sequencing was performed on the PDS study cohorts. Variants within the 210 PDS subjects and 362 control subjects were identified and analyzed using our NGS pipeline and specific filtering criteria. The primary analysis evaluated loss of function variants within five candidate genes known to cause PDS in mice. Secondary analyses were performed to 1) evaluate loss of functions variants within an additional 21 candidate genes; 2) evaluate loss of function variant burden within the whole exome; and 3) evaluate rare non-silent variant burden within the whole exome.

Results. Four isolated loss of function variants within our 5 candidate genes were identified within the case and control cohorts. Three of the variants were identified in PDS patients, while the fourth variant was detected within a control subject. Our secondary analysis of 21 candidate genes identified two loss of functions variants within PDS subjects and one loss of function variant within a control subject. Secondary analysis failed to identify any suggestive loss of function variant burden within the whole exome. However, secondary analysis for rare non-silent variants identified a gene with 9 variants in the case cohort and none in the control cohort.
Conclusions. By performing the largest NGS analysis of PDS to date, we identified multiple variants within candidate genes that are of significant interest. We also identified a new gene of interest that is currently under further investigation in regard to its contribution to PDS.

Introduction
Pigment Dispersion Syndrome (PDS) is an eye condition found in approximately 2.5% of the general population of the United States. The condition is characterized by the abnormal release of pigment from the iris into the aqueous humor. PDS is most prevalent among 20 to 40 year-old myopic Caucasians and is associated with a concave configuration of the iris, which may promote pigment shedding due to mechanical rubbing of the posterior iris pigmented epithelium against the lens zonules. Once the pigment is released from the iris it circulates through the aqueous humor and accumulates within different sections of the eye including 1) along the suspensory zonules as they insert into the lens (Scheie stripe or Zentmayer ring); 2) on the anterior surface of the iris; 3) on the corneal endothelium (Krukenberg spindle); and 4) in the trabecular meshwork. It is hypothesized that accumulation of pigment increases intraocular pressure (IOP) by damaging trabecular meshwork endothelial cells and decreasing the outflow of the aqueous humor. The risk for patients with PDS to developed elevated IOP and secondary glaucoma, known as pigmentary glaucoma (PG), is estimated to increase with time, being 10% at 5 years after diagnosis, and 15% at 15 years, with rates as high as 50% being reported.

The major clinical features of PDS are inherited in a Mendelian fashion among inbred dogs, mice, and in human pedigrees, suggesting that PDS has a strong genetic component.
Although several regions of the genome have been linked to autosomal dominant familial cases of PDS \(^{268,271}\), no causative genes have been identified in humans. However, recent studies using mice suggest several genes involved in melanin synthesis as potential candidates for human disease. Mutations in \textit{Tyrp1} and \textit{Gpnmb} on the DBA/2J inbred mouse strain cause progressive release of pigment from the posterior iris (iris pigment epithelium), increased IOP, and optic nerve degeneration \(^{267,272,273}\). The protein products of \textit{Tyrp1} and \textit{Gpnmb} are found within melanosomes (lysosome-like organelles that synthesize and store melanin) and share common motifs with other melanosomal proteins such as tyrosinase, the enzyme that catalyzes the first step of melanin synthesis \(^{274}\). It is hypothesized that mutations in \textit{Tyrp1} and \textit{Gpnmb} within DBA/2J mice compromise the structural integrity of melanosomes, leading to the release of cytotoxic intermediates of melanin synthesis, thus damaging the iris pigment epithelium\(^{272}\).

\textit{TYRP1} (tyrosinase-related protein 1) is the most prevalent melanosomal glycoprotein \(^{275}\) and is a component within the membrane-bound protein complex required for the stabilization of tyrosinase \(^{276}\). \textit{GPNMB} (glycosylated protein nmb) is also a glycosylated transmembrane protein believed to be involved with melanogenesis. Within a small number of human PDS cases \(^{272,277}\), the human orthologues to the mouse genes, \textit{TYRP1} and \textit{GPNMB}, have been screened for mutations, but, no disease-causing variations have been identified.

Mutations in \textit{Tyrp1} and \textit{Gpnmb} in DBA/2J mice that cause pigmentary glaucoma are also associated with abnormal coat color. This led us to examine other coat color defect mouse strains for iris pigment dispersion and glaucoma similar to the DBA/2J mice. Previous work identified three substrains of C57BL/6J mice that have mutations in genes known to influence pigmentation and coat color \(^{278}\). The beige substrain has a mutation in \textit{Lyst} (lysosomal trafficking
regulator gene) that causes pronounced pigment dispersion and greatly enlarged melanosomes\(^ {278-280}\). The \textit{nm2798} and \textit{vitiligo} substrains of C57BL/6J also display pigment dispersion due to mutations in \textit{Dct} and \textit{Mitf}, which are genes involved in melanin synthesis\(^ {278}\). Mutations in the protein product of these genes is hypothesized to hinder enzymatic steps in melanogenesis and cause a buildup of highly cytotoxic intermediates\(^ {281}\). The release of iris pigment into the aqueous humor may be caused by cell death within the iris due to the release of these toxic intermediates by damaged melanosomes.

Due to the studies mentioned, we hypothesized that mutations in \textit{TYRP1}, \textit{GPMB}, \textit{LYST}, \textit{DCT}, and \textit{MITF} genes may be associated with PDS in humans. To test this hypothesis, we screened a cohort of 210 patients with PDS and 362 control subjects for mutations within these 5 candidate genes using whole exome sequencing. We also performed secondary analyses of variants detected within a second tier of candidate genes and throughout the whole exome.

### Materials and Methods

#### Patient Enrollment

This study was approved by the Institutional Review Board at the University of Iowa and adhered to the tenets of the Declaration of Helsinki and the ARVO statement on human subjects. A complete eye examination including slit lamp examination, gonioscopy, and ophthalmoscopy was performed on each study subject. Infra-red iris transillumination was conducted on some of the subjects to identify subtle sign of iris transillumination defects. To be diagnosed with PDS, patients had to have two of the four principal features of PDS. These are 1) radial iris
transillumination defects, 2) a Krukenberg spindle, 3) a Scheie stripe, and 4) moderate to heavy pigmentation of the trabecular meshwork (≥2+ pigmentation on gonioscopy). DNA was obtained from blood samples using standard techniques previously described\textsuperscript{282}.

**DNA Sequencing**

Whole exome sequencing was performed on patients’ DNA as previously described\textsuperscript{219}. Sonification was used to fragment the DNA, and library preparation and sample bar-coding was performed with KAPA reagents (KAPA Biosystems). The exome capture was conducted using SeqCap VCRome probes from Nimblegen/Roche. Paired-end sequencing was performed using an Illumina HiSeq2500.

DNA sequence reads were aligned to the human reference genome GRCh37 using the Burrows-Wheeler aligner (BWA)\textsuperscript{49}. Sequence variants were identified using the Genome Analysis Tool Kit version 3 (GATK)\textsuperscript{283} and an in-house sequence analysis and annotation pipeline.

**Variant Filtering and Mutation Analysis Strategy**

Variants with a reported minor allele frequency greater than 2.5\% in gnomAD\textsuperscript{54} or present in more than 2.5\% of our control subjects were deemed to be too common to cause PDS and were excluded from the analysis. To reduce the load of false-positive variants in the analysis, two variant filters were used to remove low-quality variations. First, a minimum of 20\% of the total overlapping reads were required to support all variant calls. A second quality filter was applied, based upon our experience with repetitive DNA, which excluded any in-frame variations supported by fewer than 35\% of the reads spanning the variation.
The frequency of mutations detected in PDS cases (n=210) and controls (n=362) was then compared. The primary analysis (Tier 1) was a comparison of the frequency of loss of function mutations within our five candidate genes (TYRP1, GPNMB, LYST, DCT, and MITF). Loss of function variants are defined as nonsense mutations (premature termination), frameshift mutations, and canonical-splice mutations. We also conducted three additional secondary analyses: Tier 2) comparison of loss of function variants among an additional 21 genes known to influence pigmentation; Tier 3) gene-wise variant burden of loss of functions variants throughout the whole exome; and Tier 4) gene-wise variant burden of all rare non-silent variants throughout the whole exome.

**Gene-Wise Rare Variant Burden Association Test**

To test for gene-based rare variant burden within our cohort we used the SKAT-O algorithm\(^{82}\). Variants with a minor allele frequency greater than 2.5% in gnomAD\(^ {54}\) were excluded from the analysis, since variants with frequencies greater than 2.5% were consider too great to cause this disorder. Variants with a total depth of less than 5720 (average of 10X coverage for 572 samples) were excluded with BCFtools\(^{56}\), due to an inflated of false-variations called due to lack of sufficient coverage across the samples. These data were also normalized (if indels are left aligned and parsimonious), and multi-allelic variants were excluded using BCFTools\(^{56}\). Variants from all 572 samples processed by our pipeline were stored in a master variant call file (VCF) before analysis. In order to prepare our data for the SKAT-O analysis we used the tool PLINK\(^{46}\) to transform the variants from VCF format to PLINK’s binary-file format. Figure 20 depicts the preparation of our data for SKAT-O analysis.
Following a model of prioritizing loss of function variants first, the Tier 3 analysis consisted of SKAT-O analysis of only loss of function variants (as defined previously) according to RefSeq\textsuperscript{284} transcripts. Finally, the Tier 4 analysis was SKAT-O analysis of all rare non-silent variants according to RefSeq\textsuperscript{284} transcripts. Both Tier 3 and Tier 4 analysis were conducted using the SKAT-O R package on the University of Iowa’s high-performance compute cluster (Argon).

**Statistics**

For the primary analysis (Tier 1) the frequency of loss of functions variants in each of the five Tier 1 candidate genes (\textit{TYRP1, GPNMB, LYST, DCT,} and \textit{MITF}) was compared between the
PDS and control cohorts using Fisher’s exact test. A Bonferroni corrected p-value of 0.05 / 5 genes examined = 0.01 was used a threshold for significance.

For Tier 2 analysis the frequency of loss of function variants between the PDS and control cohorts using Fisher’s exact test. P-values < 0.05 (uncorrected) were considered suggestive (hypothesis generation). No statistical adjustments for multiple measurements or assertions about statistical significance were made for these secondary analyses.

For the Tier 3 and Tier 4 analyses p-values were calculated by the SKATBinary algorithm using efficient resampling, quantile adjusted moment matching, moment matching, adaptive efficient resampling, and a hybrid approach. Uncorrected P-values were calculated by the SKATBinary algorithm for hypothesis generation.

**Over-representation Analysis**

To test for over-representation of specific pathways or groups within our SKAT-O results that me be important to PDS we utilized the online tool WebGestalt. Genes with a p-value of less than 0.005 as calculated by SKAT-O were analyzed for over-representation within GeneOntology groups and KEGG, Panther, and Reactome pathways through the overrepresentation enrichment analysis tool on WebGestalt.
Results

Cohort of PDS Patients

A cohort of 210 patients were diagnosed with PDS based on the presence of two or more of the classic features of PDS: iris transillumination defects, dispersed iris pigment on the cornea (Krukenberg spindle), lens zonules (Scheie Stripe), and in the trabecular meshwork. A total of 148 (70%) of 210 PDS patients were diagnosed pigmentary glaucoma due to the presentation of optic cupping and glaucomatous visual field defects. The clinical features of these patients are shown in Table 18.

Table 18. Study cohort statistics.

<table>
<thead>
<tr>
<th></th>
<th>210</th>
<th>362</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age*</td>
<td>59.6 ± 13.5</td>
<td></td>
</tr>
<tr>
<td>Gender (female %)</td>
<td>33.80%</td>
<td>56.40%</td>
</tr>
</tbody>
</table>

Signs of PDS detected

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Krukenberg spindle</td>
<td>175 (83%)</td>
</tr>
<tr>
<td>Heavy pigmentary of trabecular meshwork</td>
<td>198 (94%)</td>
</tr>
<tr>
<td>Iris transillumination defects</td>
<td>180 (86%)</td>
</tr>
<tr>
<td>Scheie stripe</td>
<td>35 (17%)</td>
</tr>
<tr>
<td>Pigmentary glaucoma</td>
<td>148 (70%)</td>
</tr>
</tbody>
</table>

*age at blood draw. Data available for n=190 PDS patients (90%)
Primary Analysis: Evaluation of 5 Candidate Genes for Loss of Function Variations in PDS Cases and Controls

Whole exome sequencing of 210 PDS patients and 362 controls was performed to study the genetic cause of PDS. We approached our study using a tiered testing strategy with our primary analysis of designed to evaluate only the best 5 previously identified candidate genes to give us the greatest power to achieve statistical significance. Mutations in *TYRP1, GPNMB, LYST, DCT,* and *MITF* have been associated with pigment dispersion in mice. We compared the frequency of loss of function variants in these genes between our PDS patients and controls. A single loss of function variant was detected after our filtering strategies in *TYRP1, GPNMB, LYST,* and *DCT,* with none identified in MITF (Table 19). Mutations in *TYRP1, LYST,* and *DCT* were identified within a PDS patient, and the *GPNMB* variant was found in one of our controls. Statistical analysis of each of these variants demonstrated no association with PDS in our cohort (p > 0.01). Each of these variants are described in more detail below.

Table 19. Primary analysis results.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Encoded protein</th>
<th>Instances</th>
<th>Instances</th>
<th>Allele frequency</th>
<th>dbSNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPNMB</td>
<td>p.Lys107fsTrpTer6</td>
<td>0</td>
<td>1</td>
<td>0.12%</td>
<td>rs758729806</td>
</tr>
<tr>
<td>LYST</td>
<td>P.Ser2834Ter</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TYRP1</td>
<td>p.Glu139Ter</td>
<td>1</td>
<td>0</td>
<td>0.00%</td>
<td>rs749735228</td>
</tr>
<tr>
<td>TYRP2</td>
<td>Gln175Ter</td>
<td>1</td>
<td>0</td>
<td>0.014%, 0.18%</td>
<td>rs145067828</td>
</tr>
</tbody>
</table>
Table 20. Clinical features of patients with variants found in the primary analysis.

<table>
<thead>
<tr>
<th>Patient</th>
<th>GGA-1302-1</th>
<th>GGA-2572-1</th>
<th>GGA-1599-1</th>
<th>NL-941-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene</td>
<td>LYST</td>
<td>TYRP1</td>
<td>TYRP2</td>
<td>GPNMB</td>
</tr>
<tr>
<td>Age at diagnosis (years)</td>
<td>27</td>
<td>34</td>
<td>40</td>
<td>71</td>
</tr>
<tr>
<td>Gender</td>
<td>Male</td>
<td>Male</td>
<td>Male</td>
<td>Female</td>
</tr>
<tr>
<td>Race / Ethnicity</td>
<td>Caucasian</td>
<td>Caucasian</td>
<td>African American</td>
<td>Caucasian</td>
</tr>
<tr>
<td>Spectacle refraction (right)</td>
<td>-6.25 + 2.75 x 171</td>
<td>-7.00</td>
<td>-4.00 +2.00 x 098</td>
<td>+2.25</td>
</tr>
<tr>
<td>Spectacle refraction (left)</td>
<td>-6.00 + 1.00 x 114</td>
<td>- 5.50</td>
<td>-3.25 +0.75 x 096</td>
<td>+2.25</td>
</tr>
<tr>
<td>Axial eye length OD (mm)</td>
<td>NA</td>
<td>26.70</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Axial eye length OS (mm)</td>
<td>NA</td>
<td>26.74</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Central Corneal Thickness OD</td>
<td>555 µm</td>
<td>533 µm</td>
<td>592 µm</td>
<td>NA</td>
</tr>
<tr>
<td>Central Corneal Thickness OS</td>
<td>560 µm</td>
<td>517 µm</td>
<td>581 µm</td>
<td>NA</td>
</tr>
<tr>
<td>Signs of PDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Krukenberg’s spindle</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Pigmentation of trabecular meshwork</td>
<td>1+</td>
<td>4+</td>
<td>2+</td>
<td>NA</td>
</tr>
<tr>
<td>Iris transillumination defects</td>
<td>YES</td>
<td>YES</td>
<td>YES*</td>
<td>NA</td>
</tr>
<tr>
<td>Scheie stripe</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>NA</td>
</tr>
<tr>
<td>Pigmentary glaucoma</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Laser trabeculoplasty</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Trabeculectomy</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Family history</td>
<td>YES</td>
<td>YES</td>
<td>NO</td>
<td>NO</td>
</tr>
</tbody>
</table>

*with infrared videography
LYST
We detected a nonsense mutation in the LYST gene, c.8501C>A, p.Ser2834Ter, in PDS patient GGA-1302-1 (Table 19). This mutation causes a premature termination in the protein and was not detected in any other samples in our cohort from Iowa or in gnomAD. This truncating mutation encodes a protein that lacks terminal functional BEACH and WD40 domains. Clinical features of patient GGA-1302-1 are described in Table 20.

TYRP1
We detected a heterozygous 4bp duplication in the TYRP1 gene in PDS patient GGA-2572-1 (Table 19). The duplicated sequence, c.410_413dupGTAA, results in premature termination at codon 139 (p.Glu139Ter). This mutation is also identified as rs749735228 and was not observed in any other samples in our cohort from Iowa. However, it was identified in gnomAD with a minor allele frequency of 0.0004% (1/245542 alleles). The c.410_413dupGTAA mutation occurs upstream of region of the TYRP1 gene that encodes the melanosomal domain and enzymatic activity. Clinical features of patient GGA-2572-1 are described in Table 20.

DCT/TYRP2
We detected a nonsense mutation in the DCT/TYRP2 gene, c.523C>T, Gln175Ter in PDS patient GGA-1599-1 (Table 20). This C to T change results in a premature termination and has been previously identified as rs145067828. This mutation was only detected in this single patient from our Iowa cohort and has a minor allele frequency of 0.007% (16/245984 alleles) in gnomAD. However, the African subset of the gnomAD population was responsible for all instances of this variant (www.ncbi.nlm.nih.gov/snp/rs145067828). Clinical features of patient GGA-1599-1 are described in Table 20.
**GPNMB**

We detected a 14bp deletion in the *GPNMB* gene, c.310_323delAGATGCCAAAAGGA, in a patient with no known signs of PDS from our control cohort, NL-941-1 (Table 19). This complex *GPNMB* variant encodes an abnormal protein with a frameshift mutation that results in a premature termination, p.Lys107fsTrpTer6. This mutation previously described as rs758729806, was not detected in any of our PDS cohort and only once in our control cohort from Iowa. This variant has a minor allele frequency of 0.06% (147/246194 alleles) in gnomAD. Clinical features of patient NL-941-1 are described in Table 20.

**Tier 2 Analysis: Evaluation of 21 Candidate Genes for Loss of Function Variations in PDS Cases and Controls**

Our first secondary analysis (Tier 2) included 21 additional genes to test for associations between our PDS and control cohorts. These genes (*AGRP*, *ATP6V1A*, *ATP6V1B2*, *COMT*, *CYB5R3*, *EN2*, *GPR143*, *LAMP1*, *LAMP2*, *MC1R*, *MIF*, *MLANA*, *MNX1*, *PAXIP1*, *PMEL*, *POMC*, *RACK1*, *SHH*, *SLC45A2*, *TYR*, and *VAT1*) were chosen based on their function in melanin synthesis and melanosome biology²⁸⁵⁻²⁹¹.

Whole exome sequencing data from 210 PDS patients and 362 control subjects was analyzed for loss of function mutations in these 21 genes using the same filtering strategy employed in the primary analysis (Table 21). We detected individual heterozygous mutations in *MC1R*, *SLC45A2*, and *TYR*. An isolated frameshift mutation was identified in the *MC1R* gene (p.Ile180HisfsTer59) and an isolated nonsense mutation in *TYR* (p.Arg116Stop) were detected.
within the PDS cohort. The frameshift variant in \textit{SLC45A2} (p.Gly89AspfsTer24) was identified in a single control subject. These rare mutations were present in gnomAD and were not associated with PDS in our cohort (uncorrected p > 0.05).

Table 21. Tier 2 analysis results.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Mutation in the encoded protein</th>
<th>Instances</th>
<th>Genotype frequency</th>
<th>Genotype frequency</th>
<th>Genotype frequency</th>
<th>dbSNP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>\textit{MC1R}</td>
<td>p.Ile180HisfsTer59</td>
<td>1</td>
<td>0.48%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>\textit{SLC45A2}</td>
<td>p.Gly89AspfsTer24</td>
<td>0</td>
<td>0.00%</td>
<td>1</td>
<td>0.28%</td>
</tr>
<tr>
<td></td>
<td>\textit{TYR}</td>
<td>p.Arg116Stop</td>
<td>1</td>
<td>0.48%</td>
<td>0</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

**Tier 3 Analysis: Genome-Wide Loss of Function Variant Burden Test**

A genome-wide collection of loss of function variants (nonsense, frameshift, and canonical-splice) in all known genes that passed our previously described filtering criteria were collected and used in a mutation burden analysis. A total of 11,767 genes were reported with loss of function variants in the PDS patient and control subject cohorts. SKAT-O was performed for each gene with a genome-wide adjusted threshold of $4.2 \times 10^{-6}$ used to gauge interest in further
examination of the reported genes with mutation burden. The top ten results from our Tier 3 analysis are reported in Table 22.

Table 22. Top 10 genes from Tier 3 analysis (SKAT-O)

<table>
<thead>
<tr>
<th>Gene</th>
<th>ABI3BP</th>
<th>USP25</th>
<th>ADGRG7</th>
<th>CD1C</th>
<th>MRPL43</th>
<th>TNFAIP3</th>
<th>SEMA4G</th>
<th>COL4A1</th>
<th>SLC36A1</th>
<th>ABCA1</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-value</td>
<td>1.34E-04</td>
<td>4.18E-04</td>
<td>6.68E-04</td>
<td>9.90E-04</td>
<td>1.02E-03</td>
<td>1.03E-03</td>
<td>1.45E-03</td>
<td>1.63E-03</td>
<td>1.70E-03</td>
<td>1.88E-03</td>
</tr>
</tbody>
</table>

**Tier 4 Analysis: Genome-Wide Rare Non-Silent Variant Burden Test**

As a final analysis, all rare non-silent variants that passed our previously mentioned filtering criteria were used in a mutation burden analysis. Variants were identified in 17,253 genes in our PDS patient and control cohorts. To compare the mutation burden (missense, nonsense, frameshift, in-frame insertions and deletions, and canonical-splice variants) between the PDS patient and control cohorts, we employed the SKAT-O algorithm for each gene. Our SKAT-O analysis identified two genes, AP3B2 and MRAP, with p-values of 5.36 x 10^-5 that suggested potential associations with PDS. The variants in these genes were identified and Sanger sequencing was performed for to confirm the variant status in each of the reported subjects. While the variants within AP3B2 failed to confirm, the variants within MRAP all confirmed, prompting further investigation into MRAP.

We identified nine instances of six unique heterozygous non-silent variants in MRAP (melanocortin 2 receptor accessory protein) in 210 PDS cases and no variants in 362 control subjects (p = 0.000054). All MRAP variants identified in our cohort are shown in Table 23 with their reported frequency according to gnomAD.
Table 23. Reported MRAP variants.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Variants</th>
<th>Coding Change</th>
<th>gnomAD%</th>
<th>Protein Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>GGA-2442-1</td>
<td>chr21:33671285:G&gt;A</td>
<td>c.3G&gt;A</td>
<td>0.005661</td>
<td>NM_206898:p.Met1Ile</td>
</tr>
<tr>
<td>GGA-1649-1</td>
<td>chr21:33671300:C&gt;A</td>
<td>c.18C&gt;A</td>
<td>0.0343</td>
<td>NM_206898:p.Asn6Lys</td>
</tr>
<tr>
<td>GGA-2571-1</td>
<td>chr21:33671300:C&gt;A</td>
<td>c.18C&gt;A</td>
<td>0.0343</td>
<td>NM_206898:p.Asn6Lys</td>
</tr>
<tr>
<td>GGA-1599-1</td>
<td>chr21:33678992:G&gt;A</td>
<td>C.148G&gt;A</td>
<td>0.3854</td>
<td>NM_206898:p.Val50Met</td>
</tr>
<tr>
<td>GGA-1710-1</td>
<td>chr21:33686912:G&gt;T</td>
<td>C.257G&gt;T</td>
<td>0.2842</td>
<td>NM_206898:p.Arg86Leu</td>
</tr>
<tr>
<td>GGA-1423-1</td>
<td>chr21:33686956:G&gt;A</td>
<td>c.301G&gt;A</td>
<td>NA</td>
<td>NM_206898:p.Ala101Thr</td>
</tr>
<tr>
<td>GGA-2263-1</td>
<td>chr21:33686957:C&gt;T</td>
<td>c.302C&gt;T</td>
<td>0.3949</td>
<td>NM_206898:p.Ala101Val</td>
</tr>
<tr>
<td>GGA-2517-1</td>
<td>chr21:33686957:C&gt;T</td>
<td>c.302C&gt;T</td>
<td>0.3949</td>
<td>NM_206898:p.Ala101Val</td>
</tr>
<tr>
<td>GGA-1285-1</td>
<td>chr21:33686957:C&gt;T</td>
<td>c.302C&gt;T</td>
<td>0.3949</td>
<td>NM_206898:p.Ala101Val</td>
</tr>
</tbody>
</table>

The most notable of the variants identified is the c.3G>A (rs80358231) that removes the start codon p.Met1Ile and may adversely affect gene expression. The p.Met1Ile variant was identified in one of our 210 PDS patients (0.48%) and has a maximum allele frequency of 0.0093% in the non-Finnish European population according to gnomAD. We identified a c.18C>A (rs138040820) variant within 2 PDS patients (0.95%) and no control subjects that creates a p.Asn6Lys amino acid change in the protein. This variant has a maximum allele frequency of 0.055% within the other population and is predicted to be deleterious by SIFT\(^69\), probably damaging by PolyPhen2\(^67\). The c.148G>A (rs75858661) variant was found in a single PDS patient (0.48%) and no control subjects. This variant produces an amino acid change of p.Val50Met and has a maximum allele frequency of 3.9% in the African population as reported.
by gnomAD. However, our cohorts are primarily Western European which has a reported allele frequency of 0.014% in gnomAD. This variant is predicted as deleterious by SIFT\textsuperscript{69}, and probably damaging according to PolyPhen2\textsuperscript{68}. Another notable variant, c.302C>T (rs140113354) was identified within 3 PDS patients (1.4%) and no control subjects, and causes a conservative amino acid substitution (p.Ala101Val). This variant has a maximum allele frequency of 1.8% in the South Asian population in gnomAD and is predicted to be tolerated (low-confidence) by SIFT\textsuperscript{69} and possibly damaging by PolyPhen2\textsuperscript{68}. All of the MRAP variants were detected at a higher frequency within our PDS cohort than in either our control cohort or gnomAD.

**Over-representation Analysis**

While no gene was significant according to Bonferroni correction, our goal in these secondary analyses was to generate new hypotheses. To generate further hypotheses, we evaluated the 94 genes with a p-value less than 0.005 from our SKAT-O analysis for over-representation of pathways and gene ontology terms. Using WebGestalt\textsuperscript{84} we found a significant enrichment (FDR = 3.5 \times 10^{-3}) of Tier 4 analysis genes annotated as intrinsic components of the plasma membrane (GO:0031226). These results are presented in Table 24.
Table 24. Intrinsic component of the plasma membrane.

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Gene Name</th>
<th>Entrez Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADAM8</td>
<td>ADAM metallopeptidase domain 8</td>
<td>101</td>
</tr>
<tr>
<td>ACKR2</td>
<td>atypical chemokine receptor 2</td>
<td>1238</td>
</tr>
<tr>
<td>SLC5A8</td>
<td>solute carrier family 5 member 8</td>
<td>160728</td>
</tr>
<tr>
<td>SLC16A11</td>
<td>solute carrier family 16 member 11</td>
<td>162515</td>
</tr>
<tr>
<td>F10</td>
<td>coagulation factor X</td>
<td>2159</td>
</tr>
<tr>
<td>PLXNB2</td>
<td>plexin B2</td>
<td>23654</td>
</tr>
<tr>
<td>GHR</td>
<td>growth hormone receptor</td>
<td>2690</td>
</tr>
<tr>
<td>SIGLEC9</td>
<td>sialic acid binding Ig like lectin 9</td>
<td>27180</td>
</tr>
<tr>
<td>GPR3</td>
<td>G protein-coupled receptor 3</td>
<td>2827</td>
</tr>
<tr>
<td>NPBWR2</td>
<td>neuropeptides B and W receptor 2</td>
<td>2832</td>
</tr>
<tr>
<td>GPR39</td>
<td>G protein-coupled receptor 39</td>
<td>2863</td>
</tr>
<tr>
<td>NOX4</td>
<td>NADPH oxidase 4</td>
<td>50507</td>
</tr>
<tr>
<td>SLC39A4</td>
<td>solute carrier family 39 member 4</td>
<td>55630</td>
</tr>
<tr>
<td>PCDHB12</td>
<td>protocadherin beta 12</td>
<td>56124</td>
</tr>
<tr>
<td>PCDHA12</td>
<td>protocadherin alpha 12</td>
<td>56137</td>
</tr>
<tr>
<td>PCDHA11</td>
<td>protocadherin alpha 11</td>
<td>56138</td>
</tr>
<tr>
<td>PCDHA10</td>
<td>protocadherin alpha 10</td>
<td>56139</td>
</tr>
<tr>
<td>RASA1</td>
<td>RAS p21 protein activator 1</td>
<td>5921</td>
</tr>
<tr>
<td>BST2</td>
<td>bone marrow stromal cell antigen 2</td>
<td>684</td>
</tr>
<tr>
<td>TPO</td>
<td>thyroid peroxidase</td>
<td>7173</td>
</tr>
</tbody>
</table>
Discussion

This study was designed to maximize the chance of finding a PDS causing gene in humans by prioritizing genes that have been shown through mouse models to have features of PDS. It is critical when performing large scale sequencing analyses to have defined testing strategies to reduce the false discovery rate and increase the possibility of identifying a true causative variant or gene for your disease of interest. Loss of function variants were prioritized within our analysis since they are more likely to have a detrimental effect on the protein structure and interactions.

Three of the candidate genes we examined TYRP1, DCT (TYRP2) and GPNMB share significant homology with one another and with TYR (Tier 2 candidate). TYR, TYRP1, and DCT (TYRP2) each have a signal peptide sequence domain to direct intracellular localization to the melanosomes (cytoplasmic); a transmembrane domain that spans the melanosome membrane; copper binding structures; and a catalytic domain. Both variants identified in TYRP1 and DCT (TYRP2) are found in a region of the protein with unknown function or domain. All variants identified from our primary analysis and our Tier 2 analysis are subject to further analysis. Our Tier 3 analysis failed to identify anything that peaked our interest for further investigation.

The Tier 4 analysis identified MRAP as an interesting gene that required further exploration. Mutations in MRAP or melanocortin 2 receptor accessory protein have been previously associated with familial glucocorticoid deficiency (FGD). This association was originally identified by Metherell et al. when they studied a group of individuals with FGD without a mutation in the only known causative gene at the time, MC2R\textsuperscript{292}. By using linkage analysis, they mapped the region of the genome associated with FGD and then conducted expression analysis of 30 candidate genes to identify MRAP as the causative gene. MRAP is the first G protein
coupled receptor accessory protein shown to cause human disease and can be attributed to at least 20% of FGD cases\textsuperscript{293-295}. All reported mutations in \textit{MRAP} linked to FGD are predicted to encode a severely truncated protein or no protein at all\textsuperscript{293,296}.

\textit{MRAP} transports melanocortin-2 receptor (\textit{MC2R}) from the endoplasmic reticulum (interior) of the cell to the cell surface where \textit{MC2R} is activated if bound to \textit{MRAP} (Figure 21).

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{MRAP_function.png}
\caption{Model of MRAP function. Shown is a schematic of MRAP protein dimerizing and then binding to MC2R within the Endoplasmic Reticulum. The complex then is transported to the Golgi apparatus to be glycosylated (green circles) and then to the plasma membrane where it can interact with ACTH.}
\end{figure}

MC2R binds to adrenocorticotropic hormone (ACTH) and initiates the production of glucocorticoids. Homozygous variants in \textit{MRAP} have been previously implicated in glucocorticoid deficiency 2 (FGD). ACTH and glucocorticoids form a negative feedback loop
with ACTH stimulating the production of glucocorticoids which in turn, inhibit ACTH production. Patients with FGD have increased levels of ACTH due to the reduction in glucocorticoids which can be easily tested for.

Interestingly, individuals with FGD are hyperpigmented along with mice overexpressing alpha melanocyte-stimulating hormone (α-MSH), an agonist of MC1R. MC1R is one of 5 melanocortin receptor proteins including MC2R within humans and is involved with the regulation of pigmentation. As mentioned previously, we identified a heterozygous premature stop mutation within a PDS patient who also has exfoliation glaucoma from our Tier 2 analysis. Both ACTH and α-MSH are products of the gene POMC, with α-MSH actually being a derivative of the first 13 amino acids within the ACTH sequence.

The mutations in MRAP identified within our cohort have all been confirmed as heterozygous, which fit the proposed model of autosomal dominant inheritance for PDS. Seven of the nine PDS patients identified with MRAP mutations have pigmentary glaucoma which is equivalent to the overall amount of the PDS cohort with pigmentary glaucoma (Table 25). We also performed sequencing of MRAP within a replication cohort of 88 PDS patients and 85 controls. We identified 3 more patients and no control samples with MRAP variants furthering strengthening our hypothesis that variants in MRAP could be important to the development of PDS.
<table>
<thead>
<tr>
<th>MRAP mutation (aa)</th>
<th>Diagnosis</th>
<th>KS OD</th>
<th>KS OS</th>
<th>TID OD</th>
<th>TID OS</th>
<th>Pigment in TM OD</th>
<th>Pigment in TM OS</th>
</tr>
</thead>
<tbody>
<tr>
<td>p.Ala101Val</td>
<td>PG</td>
<td>Yes (1+,2+)</td>
<td>Yes    (1+,2+)</td>
<td>Moderate</td>
<td>Moderate</td>
<td>4+ (Intense)</td>
<td>4+ (Intense)</td>
</tr>
<tr>
<td>p.Ala101Thr</td>
<td>PG</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Only on infrared</td>
<td>Only on infrared</td>
<td>3+ (Moderate)</td>
<td>2+ (Mild)</td>
</tr>
<tr>
<td>p.Val50Met</td>
<td>PG</td>
<td>Yes (1+,2+)</td>
<td>Yes (1+,2+)</td>
<td>Trace</td>
<td>Trace</td>
<td>3+ (Moderate)</td>
<td>3+ (Moderate)</td>
</tr>
<tr>
<td>p.Asn6Lys</td>
<td>PG</td>
<td>Yes (1+,2+)</td>
<td>Yes (1+,2+)</td>
<td>Unknown</td>
<td>Unknown</td>
<td>3+ (Moderate)</td>
<td>3+ (Moderate)</td>
</tr>
<tr>
<td>p.Arg86Leu</td>
<td>PDS + OHT</td>
<td>Yes (1+,2+)</td>
<td>Unknown</td>
<td>Unknown</td>
<td>Unknown</td>
<td>4+ (Intense)</td>
<td>4+ (Intense)</td>
</tr>
<tr>
<td>p.Ala101Val</td>
<td>PG w/o OHT</td>
<td>Yes (1+,2+)</td>
<td>Yes (1+,2+)</td>
<td>None</td>
<td>None</td>
<td>3+ (Moderate)</td>
<td>3+ (Moderate)</td>
</tr>
<tr>
<td>p.Met1Ile</td>
<td>PDS</td>
<td>Unknown</td>
<td>Yes (1+,2+)</td>
<td>Trace</td>
<td>Trace</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>p.Ala101Val</td>
<td>PG</td>
<td>Yes (1+,2+)</td>
<td>Yes (1+,2+)</td>
<td>Moderate</td>
<td>Moderate</td>
<td>4+ (Intense)</td>
<td>3+ (Moderate)</td>
</tr>
<tr>
<td>p.Asn6Lys</td>
<td>PG</td>
<td>Yes (1+,2+)</td>
<td>Yes (1+,2+)</td>
<td>Trace</td>
<td>Trace</td>
<td>4+ (Intense)</td>
<td>3+ (Moderate)</td>
</tr>
</tbody>
</table>

Table 25. Clinical features of patients with MRAP mutations.

(KS = Krukenberg spindle, TID = Transillumination Defect, TM = Trabecular Meshwork)
It is possible that a reduction of MC2R protein available to bind ACTH could lead to an increase in α-MSH levels; inducing pigment production in patients with MRAP mutations without causing a large enough reduction in glucocorticoids to cause FGD. On the other hand, heterozygous variants found in MC1R have been shown to cause a reduction in pigment, an increased chance of developing freckles that fade with reduction to sun exposure (ephelides) and skin cancer\textsuperscript{300,301}. Until further functional analysis and replication is performed, it is difficult to determine the real role that variants within MRAP and MC1R are playing within our PDS cohort.

**Conclusion**

Glaucoma is a heterogeneous disease that can be caused through multiple different mechanisms. This study has identified variants in known candidate genes and novel candidate genes that require further study to truly elucidate their function within human glaucoma. MRAP represents a promising candidate with variants being identified in nine cases of PDS, seven of which were diagnosed with PG. If through validation these variants are deemed to cause PDS or PG, it would represent the genetic cause of approximately 4% of PDS cases.
CHAPTER 6: CONCLUSION

Bioinformatics is the science of collating and analyzing complex biological data. This term first appeared in publications in 1978 and has since become an integral component in biomedical research. During my graduate studies I utilized many different computational strategies and bioinformatic tools to analyze the genetics of three very different human eye diseases furthering my development as a bioinformatician. Each project taught me important lessons that have made me a better researcher and a firm believer of the necessity of strong collaborations between clinicians, basic science researchers and bioinformaticians when investigating the genetics of human disease.

Though I did not definitively identify the genetic cause of keratoconus within the three Bedouin families, we were able to show the utility and importance of continuing to collect high-quality family data and DNA. By performing linkage analysis within each family, I was able to identify regions of the genome that likely contain the disease-causing mutation or risk alleles. The major limitation of this project was the number of samples that were collected from each family. We were not able to obtain DNA from the parents of our largest family, which limited the power of the linkage analysis. However, due to the principle of linkage disequilibrium I was able to explore combining the results from the three families further narrowing the genomic space of interest under the hypothesis of a common founder in the Bedouin population. By combining linkage analysis with exome sequencing we were able to remove a significant portion of the variants identified in each family, and thus to greatly reduce our chance at discovering a false positive result. I identified variants that were shared between at least two of the three families that are the subject of further investigation by our collaborator in Israel. This project powerfully
demonstrated the incredible utility that family data can provide when trying to identify the disease-causing mutation for a disease without a known genetic cause.

Next, I performed bioinformatic analyses on three distinct cohorts of Bardet-Biedl syndrome (BBS) patients that lead us to identify both known and novel causes of disease. The first cohort is a collection of unrelated patients that were contributed by multiple collaborators, and which had been previously screened for mutations in known BBS genes. From this cohort of previously unsolved BBS patients, I identified complete genotypes in known BBS genes. Two of those four patients had two exon deletions in *BBS4* that had been missed by the previous CNV analysis program, CoNIFER, but which were called by CLAMMS, that we recently began using for CNV identification. I also identified 8 individuals with mutations in genes that are known to cause BBS-like phenotypes. Four of these individuals have very convincing complete genotypes in genes that cause BBS-like phenotypes. Unfortunately, detailed clinical information is lacking for these individuals at this time, so we cannot definitively state whether: 1) this represents a new BBS genes, 2) the patients were misdiagnosed, or 3) the variants identified are not meaningful to the patient’s diagnosis and the causative mutations have yet to be found.

The second BBS cohort was a family of 4 with a single affected child. After a preliminary mutation screen failed to identify the most common BBS mutations, whole exome sequencing analysis of the affected patient was performed. This identified two novel heterozygous mutations in the BBS gene *ARL6 (BBS3)* within the patient. Sanger sequencing of the other family members confirmed segregation, with each parent carrying one of the ARL6 alleles, which were passed to the affected patient, and the unaffected sibling also carrying only one of the ARL6
alleles. The availability of DNA from the other family members provided the opportunity to confirm that these mutations in a known BBS gene are in trans, and therefore represent a complete genotype. Thus, this is the likely cause of BBS within this family.

The final BBS cohort was a collection of Puerto Rican samples that were diagnosed with BBS. A preliminary screen of the most common BBS mutations identified a homozygous genotype of the most common BBS mutation, M390R in BBS1, within a single sample. Heterozygous variants in previously described BBS genes were identified within 5 of the other 9 samples. Given the lack of complete genotypes within known BBS genes in the 9 samples, I expanded my search into other known eye disease-causing genes. This led us to identify two samples (familial) which were homozygous for both a premature stop mutation (Lys260Stop) and a missense Asp140Glu mutation in KCNV2. Another sample (unrelated) contained heterozygous Lys260Stop and Asp140Glu mutations in KCNV2. This gene is known to cause retinal cone dystrophy 3B and has not been previously associated with BBS. While the Lys260Stop has been reported as disease-causing for KCNV2 patients, both reports only provide clinical information regarding the visual phenotype\textsuperscript{253,303}. If those patients had other clinical BBS phenotypes that weren’t reported, or if the ethnic background of our samples has a modifier effect, it is possible that KCNV2 could be a novel BBS gene. This would not be unprecedented, as Sanchez et al. (2016) were able to show that perturbations in the expression of voltage-gated potassium channel gene KCNH1 affected ciliogenesis\textsuperscript{304}. Therefore, it is possible that mutations in KCNV2, which is also a voltage-gated potassium channel modifier, may lead to changes in ciliogenesis that could cause BBS phenotypes.
There is precedence for genes initially being identified as the cause of one disorder and subsequently being found to cause BBS. For example, \textit{C8ORF37 (BBS21)}\textsuperscript{202,255,256} was originally identified as a non-syndromic retina disease-causing gene before being identified as a BBS gene. In addition, the \textit{MKKS} gene was first identified as the cause of McKusick-Kaufman syndrome and subsequently found to cause BBS (\textit{BBS6}). We have reached out to our collaborator in Puerto Rico for more comprehensive clinical information regarding the three samples. If the patients are confirmed to meet the diagnostic criteria of BBS, \textit{KCNV2} could be considered a BBS gene.

Finally, I performed the largest (n=210 probands) whole exome sequencing study of pigment dispersion syndrome (PDS). Using a tiered testing strategy, I first interrogated loss of function variants in 5 candidate genes known to cause PDS in mouse models. Three isolated loss of function variants within our candidate genes were identified within the PDS cohort that are being studied further. A secondary analysis identified two loss of function variants in two of the additional 21 melanosome related genes within the PDS cohort. I also performed variant burden analysis for all rare non-silent variants within the whole exome. This led to the identification of a novel candidate gene, \textit{MRAP}, in which 9 heterozygous variants were identified in the PDS cohort, and none were found in the control cohort. Homozygous mutations in \textit{MRAP} are known to cause familial glucocorticoid deficiency which also presents with hyper-pigmentation. It is possible that heterozygous mutations in \textit{MRAP} may cause PDS, which is believed to be inherited in an autosomal dominant fashion. To investigate this further a replication cohort was evaluated, looking for variants in \textit{MRAP} within an additional 88 PDS cases and 85 controls. Three additional PDS cases were identified with \textit{MRAP} mutations and none of the additional control
subjects contained any of the mutations. At this time \textit{MRAP} is the subject of further investigation to determine its role in PDS development.

The ability to analyze genome-scale data was critical to each of these projects and was possible due to the decreasing cost of next-generation sequencing (NGS). The continued development of sequencing technologies over the last 20 years has given us the ability to sequence the entire genome of a human for \$1600 at the Iowa Institute of Human Genetics compared to the estimated \$300 million for the initial draft of the human genome. This has allowed for the development and implementation of precision genomic medicine within the medical field\textsuperscript{305}. A clinician has the ability to order whole genome sequencing for their patient if desired, providing an extensive amount of data regarding the patient’s genetic makeup.

As the cost of obtaining a patient’s genome continues to decrease and becomes ever more cost-efficient, the most important component of precision medicine is the actual clinician performing all of their diagnostic tests and then establishing a clinical diagnosis or hypothesis. Every human contains hundreds of thousands of variants when compared to the human reference genome that are benign and have no functional effect\textsuperscript{306}. The average person also carries single copies of multiple different disease-causing variants which, if present in both alleles, would be lethal\textsuperscript{307,308}. While many strategies have been developed to filter out a significant portion of the benign variants, there are still too many plausibly disease-causing variants to robustly assert their relevance without limiting the set of genes within the analysis, either with a strong clinically-based pre-test hypothesis or the ability to perform segregation analysis within family members of the patient. Stone \textit{et al.} referred to this dilemma as the “false genotype rate” (FGR), which is
defined as “the frequency with which one would encounter a plausibly disease-causing recessive or dominant complete genotype when sequencing the coding regions of a specific set of genes in a healthy person.” Simply put, ordering an exome or genome for a single patient without a strong hypothesis concerning the differential diagnosis or likely genes causing the disorder in the specific patient would provide far too many plausible disease-causing variants to have confidence in the result. In fact, Stone et al. determined that a pre-test hypothesis of 10 genes or less was required to keep the FGR less than 5% within their cohort of patients. Thus, without a strategy to limit the set of genes to be evaluated, incorrect reports regarding the genetic cause of patients’ diseases will be provided – the antithesis of precision medicine.

NGS is a powerful tool when properly applied, that has allowed for the discovery of many disease-causing genes, particularly when used in tandem with strategies to reduce the genomic space of interest. Before the development of NGS, diseases were first mapped to a specific locus within families which were then interrogated for a variant that segregated properly through the family. As we continue into the era of precision medicine it is critical that we continue to perform hypothesis driven science and not use NGS as a crutch. We can achieve this goal by training high-quality teams of clinicians, basic scientists, and bioinformaticians that can work together to help better the lives of patients.
REFERENCES


111


123. AUSTIN MG, SCHAEFER RF. Marfan's syndrome, with unusual blood vessel manifestations. *AMA Arch Pathol.* 1957;64(2):205-209.


