# DISEASE GENE DISCOVERY BY LINKAGE MAPPING AND EXOME ANALYSIS

by

Esra Yıldız Bölükbaşı

B.S., Molecular Biology and Genetics, Boğaziçi University, 2013

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in Molecular Biology and Genetics

Boğaziçi University

2018

*To my dear Family and*

*beloved Serhat*

# ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and respect to my thesis supervisor Prof. Aslıhan Tolun for her guidance, endless support and adding value to my life throughout my study in Kommagene Lab and for widening my horizon in our daily conversations about science and life. I also would like to extend my appreciation to the members of my thesis committees Prof. Nesrin Özören, Prof. Eda Tahir Turanlı, Assoc. Prof. Sibel Uğur İşeri, Prof. Beki Kan, Prof. O. Uğur Sezerman and Prof. Uğur Özbek for devoting their time to evaluate this work.

I would like to thank Assoc. Prof. Sajid Malik and Dr. Sara Mumtaz from Pakistan, Assoc. Prof. Şevket Özkaya and Assoc. Prof. Selvi Aşker for the blood samples, clinical evaluations and contributions to the articles. I thank all the families who participated in the study.

I would like to express my deep gratitude to Çiğdem Köroğlu and Özgecan Ayhan for their extensive teaching, guidance and sharing their experience in lab work and to my friends who were more than colleagues to me, Nehir Mavioğlu, Yeşerin Yıldırım, Gökhan Nalbant, Çağla Çakmak and İlker Karacan primarily for their help and great friendship; we shared a lot, and they have always been a great motivation for me.

I also would like to thank deeply my family for their endless support, patience and assistance throughout my entire life and for loving me unrequitedly. I am grateful to my husband Serhat who has supported me in following my dreams to become a scientist, believing in me all the time and sharing his life with me.

I sincerely thank MBG family members for friendship, sharing their experience and for their help in laboratory courses.

This work was supported by Boğaziçi University Research Fund (7695 and 10860) and the Scientific and Technological Research Council of Turkey (114Z829). I was supported as a fellow of the latter grant from 2016 to 2018.

# ABSTRACT


## DISEASE GENE DISCOVERY BY LINKAGE MAPPING AND EXOME ANALYSIS

Disease gene identification is an important area in genetics. Consanguineous marriages are very common in some populations and lead to emergence of rare diseases. To uncover the functions of our genes and the molecular bases of diseases, novel disease genes need to be identified. Later the effect of the mutation on the protein could be investigated by various analyses. Also, discovery of a new disease gene can be a glimmer of hope for the patients who are hoping for definite diagnosis and development of therapies. The steps of the disease gene identification is mapping the locus by linkage analysis or homozygosity mapping, identifying the causative mutation by exome sequence analysis and relating the mutation to the disease pathology, thus uncovering the molecular pathogenesis. Possible or known effect of the mutation can be investigated via computational algorithms and in the light of the information in published studies and databases.

In this thesis study, causative genes were searched in eight consanguineous families afflicted with six different recessive diseases. The two identified genes and the identified mutations are *PDIA3* (p.Cys57Tyr) in Syndromic Intellectual Disability family and *CEP19* (p.Tyr65*) in Bardet-Biedl Syndrome family. Besides, in BBS family possible modifiers *GLI1* p.Gly274Arg, *CCDC28B* p.Phe110Phe, MKKS/*BBS6* p.Ile339Val, *C8ORF37* p.Ala178Val and *TMEM67* p.Asp799Asp were identified/detected. In Isolated Intellectual Disability family, missense *PTRHD1* p.Cys52Tyr mutation confirmed that the gene is responsible for ID. Gene expression assay revealed wide expression in brain. In Intellectual Disability and Hypothyroidism family missense *TPO* p.Asp240Gly was identified. Intronic splicing c.6375-1G>C in *SPTBN2* in Spinocerebellar Ataxia family unraveled the basis of dominant vs recessive effects of the variants in this gene. Synonymous p.Ile382Ile in *TNKS2* in one of the Pleuroparenchymal Fibroelastosis families was identified as the strongest candidate. No candidate variant could be identified in the other two families, indicating genetic heterogeneity for this disease.

# ÖZET


## BAĞLANTI HARİTALAMASI VE EKSOM ANALİZİ İLE HASTALIK GENİ KEŞFİ

Hastalık geni keşfi, genetik araştırmaları içinde önemli bir yere sahiptir. Bazı toplumlarda akraba evliliklerinin çok yaygın olması nedeniyle, nadir hastalıklar sık ortaya çıkar. Genlerimizin işlevlerini ve kalıtsal hastalıkların moleküler nedenlerini bulmak için yeni hastalık genleri keşfedilmesi gerekmektedir. Daha sonra mutasyonun protein üzerindeki etkisi çeşitli analizlerle irdelenebilir. Ayrıca, hastalık geninin keşfi kesin tanıyı olanaklı kılar ve tedavi geliştirilmesini bekleyen hastalar için bir umut ışığı olabilir. Hastalık geni keşfindeki adımlar şöyledir: bağlantı analizi ile hastalık geninin bölgesinin belirlenmesi, eksom dizileme analizi ile aday bölgede hastalığa neden olan mutasyonun bulunması ve mutasyonun hastalık patogenezi ile ilişkilendirilerek moleküler düzeyde sebebin bulunması. Mutasyonun olası ya da bilinen etkisi yayınlanmış çalışmalar ile veritabanlarındaki bilgiler ışığında bilişimsel algoritmalar kullanılarak araştırılır.

Bu tez çalışmasında, akraba evliliği yapmış sekiz ailedeki altı ayrı hastalık için gen arandı. Bulunan iki gen yenidir ve belirlenen mutasyonlar şunlardır: Sendromik Mental Retardasyon için *PDIA3* p.Cys57Tyr mutasyonu ve Bardet-Biedl Sendromu için *CEP19* p.Tyr65*. BBS ailesinde ayrıca muhtemel düzenleyici olarak *GLI1* p.Gly274Arg, *CCDC28B* p.Phe110Phe, *MKKS/BBS6* p.Ile339Val, *C8ORF37* p.Ala178Val ve *TMEM67* p.Asp799Asp bulundu. Bu sendromda bu güne kadar bildirilen en fazla mutasyon yükünü bu ailenin taşıdığı gösterildi. Sendromik Olmayan Mental Retardasyon için bulunan *PTRHD1* p.Cys52Tyr mutasyonu genin mental retardasyondan sorumlu olduğunu kesinleştirdi. Mental Retardasyon ve Hipotriodizm için *TPO* p.Asp240Gly mutasyonu bulundu. Spinoserebellar Ataksi için kırpılmayı etkileyen *SPTBN2* c.6375-1G>C mutasyonu bu gendeki mutasyonların neden bazılarının baskın, diğerlerinin çekinik olduğuna ışık tuttu. Gerçekleştirilen anlatım analizi genin beyinde yaygın anlatımı olduğunu gösterdi. Plöroparenkim Fibroelastozisi ailelerinden birinde en güçlü aday olarak *TNKS2* p.Ile382Ile varyantı belirlendi. Diğer iki ailede aday bir varyant saptanmaması, bu hastalıkta genetik heterojenlik olduğuna işaret etti.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ACRONYMS/ABBREVIATIONS

| | |
|---|---|
| Ala | Alanine |
| APS | Ammonium peroxodisulphate |
| Asp | Aspartic acid |
| Arg | Arginine |
| Asn | Asparagine |
| Asp | Aspartic acid |
| BAM | Binary alignment/map |
| bp | Base pair |
| BSA | Bovine serum albumin |
| BWA | Burrows-wheeler aligner |
| cDNA | Complementary deoxyribonucleic acid |
| Chr | Chromosome |
| cM | Centimorgan |
| CNV | Copy number variation |
| Cont. | Continued |
| Cys | Cysteine |
| DGV | Database of Genomic Variants |
| $dH_2O$ | Distilled water |
| DNA | Deoxyribonucleic acid |
| dNTP | Deoxynucleotide |
| EDTA | Ethylenediaminetetraacetate |
| ENCODE | Encyclopedia of DNA Elements |
| EVS | Exome Variant Server |
| ExAC | Exome Aggregation Consortium |
| GATK | Genome Analysis Tool Kit |
| Gln | Glutamine |
| Gly | Glycine |
| gnomAD | Genome Aggregation Database |
| HCiE | Homozygosity comparison in Excel |

| | |
|---|---|
| HGVS | Human Genome Sequence Variation Society |
| His | Histidine |
| HRM | High resolution melting |
| IBD | Identical by descent |
| ID | Intellectual Disability |
| IDT | Integrated DNA Technologies |
| IGV | Integrative Genomics Viewer |
| Ile | Isoleucine |
| Indel | Insertion-deletion |
| i.e. | id est (that is) |
| kb | Kilobase pair |
| Leu | Leucine |
| LOD | Logarithm of odds |
| Lys | Lysine |
| MAF | Minor allele frequency |
| Mb | Mega base |
| mg | Milligram |
| min | Minute |
| ml | Milliliter |
| µg | Microgram |
| µl | Microliter |
| mRNA | Messenger ribonucleic acid |
| NCBI | National Center for Biotechnology Information |
| ng | Nanogram |
| NGS | Next-generation sequencing |
| NHLBI | National Heart, Lung and Blood Institute |
| nM | Nanomolar |
| OMIM | Online Mendelian Inheritance in Man |
| PAR | Pseudoautosomal region |
| PCR | Polymerase chain reaction |
| Phe | Phenylalanine |
| Pro | Proline |
| QPCR | Quantitative polymerase chain reaction |

| | |
|---|---|
| RNA | Ribonucleic acid |
| Rpm | Revolutions per minute |
| SA | South Asians |
| SAM | Sequence Alignment/Map |
| SDS | Sodium dodecyl sulfate |
| SNP | Single nucleotide polymorphism |
| SSCP | Single Strand Conformational Polymorphism |
| TBE | Tris, boric acid, EDTA |
| TE | Tris, EDTA |
| ter | Terminus |
| TEVD | Turkish Exome Variant Database |
| Tyr | Tyrosine |
| UCSC | University of California Santa Cruz |
| UIP | Usual Interstitial Pneumonia |
| URL | Uniform Resource Locator |
| UTR | Untranslated region |
| Val | Valine |
| X | Stop codon |
| YCGA | Yale Center for Genome Analysis |

# 1. INTRODUCTION

In this thesis study, causative genes in eight consanguineous families afflicted with six different recessive diseases were searched by applying linkage analysis and exome sequencing methods. Five of the families are from Pakistan and had inherited disorders either with novel manifestations or assessed as not linked to a locus of a similar known disease. Those diseases manifest with intellectual disability (ID), including the Bardet-Biedl syndrome and spinocerebellar ataxia. The remaining three families are Turkish and afflicted with pleuroparenchymal fibroelastosis (PPFE).

## 1.1. Intellectual Disability (ID)

Intellectual disability (ID) is a neurodevelopmental disorder that affects 1-3% of the population (Leonard and Wen, 2002). It is considered as an important medical problem and characterized by limitations in intellectual functioning and adaptive behavior (Vissers *et al.*, 2016). Its severity is highly variable from mild to profound. While most of the cases have genetic basis, some exogenous factors can also play a role in the development of ID. Mostly ID is recognized in early childhood and accompanied by developmental delay. It can manifest as an isolated or syndromic condition with neurological features such as autism and epilepsy or other findings such as skeletal or hand/foot malformations. Due to extreme genetic heterogeneity, although many genes associated with ID have been identified, there are still many cases awaiting genetic diagnosis (Mefford *et al.*, 2012). Previously, the most common causes of ID were Fragile X and Down syndrome (Pieretti *et al.*, 1991). With the introduction of the microarray technology, genome-wide studies can now be applied, and over 700 genes have been identified for all X-linked, autosomal dominant and recessive ID. It has been reported that de novo mutations play roles in 13-35% of severe ID cases (Vissers *et al.*, 2016). Despite many causative genes for X-linked ID have been discovered by now, just a small portion for autosomal ID could be identified
.

It is hypothesized that the proteins encoded by ID genes play roles in one or more shared pathways or functional modules, either by direct interactions or as part of more complex interaction networks. These can be cellular processes such as neurogenesis, neuronal migration, synaptic function, or regulation of transcription or translation (van Bokhoven, 2011). The RAS–MAPK (mitogen-activating protein kinase) pathway which is a metabolic pathway that regulates growth factors and embryological development is an example of a pathway associated with ID. Proteins encoded by genes linked to some well-known ID syndromes such as Noonan syndrome and Costello syndrome are included in RAS–MAPK pathway. Another example is RHO GTPase pathway, with a role in cellular functions such as morphogenesis of dendritic spines which are crucial for learning and memory.

### 1.2. Spinocerebellar Ataxia (SCA)

Ataxia is a neurodegenerative disorder that can be defined as uncoordinated and poor movement due to neurodegeneration of the cerebellum and other parts of the nervous system such as pons, the basal ganglia and the cerebral cortex (Velazquez-Perez *et al.*, 2017). It is mostly characterized by unsteady gait, but it can also affect fingers, hands, arms, speech (dysarthria) and eye movements (nystagmus). Ataxia can be in pure form or part of a neurodegenerative disease. Hereditary forms are generally caused by dysfunction of the cerebellum which is responsible for the control of motor movements by transmitting input from sensory systems of spinal cord and brain. The input is integrated into motor activity in the brain (Manto, 2008). Ataxia can be sporadic or inherited in an autosomal dominant, autosomal recessive or X-linked manner. Even though the genetic causes are different, many symptoms are common for all types.

Cerebellar ataxia (also called spinocerebellar ataxia, SCA) is a slowly progressive type of ataxia due to shrinkage or atrophy of the cerebellum. It is caused by mutations in genes that are crucial for brain function. More than 40 types SCA have been described, and many of them are dominantly inherited, early onset and caused by repeat expansions (OMIM). Also, there are 24 types of the disease that are inherited recessively, namely, the

“spinocerebellar ataxia, autosomal recessive”, diseases (OMIM). Spinocerebellar ataxia-14 (SCAR14; MIM 605361) is one of the autosomal recessive types which can be defined as severe early-onset gait ataxia, delayed psychomotor development, eye movement abnormalities such as nystagmus, cerebellar atrophy evident on brain imaging, and intellectual disability (Lise *et al.*, 2012).

*Spectrin, Beta, Non-Erythrocytic 2* (*SPTBN2*) mutations are associated with SCAR14. The encoded beta-III spectrin together with alpha subunits constitutes a spectrin protein which is a membrane scaffold protein. Spectrins were discovered in erythrocytes (Perrotta *et al.*, 2008). They are present also in brain and known to be important in cerebellar functions (Goodman *et al.*, 1995). They play an important role in anchoring and stabilizing membrane spanning proteins within specific subdomains of the plasma membrane. Beta-III spectrin is a 2,390 amino acid protein with three parts. These are N-terminal domain containing actin/ARP1 binding site, 17 spectrin repeats and Pleckstrin Homology (PH) domain in the C-terminal that recruits proteins to membranes by binding phosphatidylinositol lipids within biological membranes (Lise *et al.*, 2012).

Heterozygous mutations in *SPTBN2* can cause spinocerebellar ataxia type 5 (SCA5) which is an autosomal dominant, slowly progressive adult onset disease. In two families with more severe childhood-onset SCAR14 homozygous mutations were reported. In one of the families the symptoms are developmental delay, nystagmus, convergent squint and cognitive impairment, and in the other family, wide-based gait, developmental delay and pyramidal signs (Elsayed *et al.*, 2014; Lise *et al.*, 2012). Morphological abnormality of neurons in brain was observed in *beta-III spectrin* knockout mice, evidence that *SPTBN2* mutation is the basis of cognitive impairment in humans. It was reported that the knockout mice had cerebellar ataxia and a progressive loss of cerebellar Purkinje cells (Perkins *et al.*, 2010). In MRI of patients with homozygous *SPTBN2* mutation, cerebellar atrophy was observed.

## 1.3. Bardet-Biedl Syndrome (BBS)

One well-studied rare syndrome is Bardet-Biedl syndrome (BBS). It is a pleotropic ciliopathy with variable clinical features. The primary features are retinal dystrophy, central obesity, post-axial polydactyly, urino-genital abnormalities, learning difficulties and mental retardation. There are several secondary features such as speech disorders, developmental delay, dental anomalies, brachydactyly/syndactyly, diabetes, neurological malfunctioning and behavioral impairments. Mostly BBS manifests with at least three primary and two secondary features mentioned above (Forsythe and Beales, 2013; Schaefer *et al.*, 2014). Its prevalence is 1 - 9 in 1 million (Orphanet). BBS symptoms develop gradually throughout childhood, and polydactyly may be the only obvious feature at the time of birth. Hence, the condition is diagnosed rather late in childhood in most of the patients. There is substantial inter- and intra-familial clinical variability as well as reduced penetrance (Forsythe and Beales, 2013; Katsanis, 2004). Furthermore, genetically BBS is one of the most heterogeneous syndromes exhibiting wide locus and allelic heterogeneity.

Most of the BBS types are inherited autosomal recessively, and digenic or triallelic inheritance is also reported. Curiously, polyallelic recessive segregation has been indicated by inter- and intra-familial variable expressivity which is due to the involvement of other modifier genes (Katsanis *et al.*, 2001; Khan *et al.*, 2016). It has been suggested that the extent of complexity in phenotypic severity and mode of inheritance is dependent on the dose of the contribution of a third mutation in a secondary gene (Katsanis *et al.*, 2001). There are 24 known genes associated with BBS (BBS1—BBS21; OMIM: PS209900) including three modifier genes (one in both groups) plus the gene that we identified with studies included in this thesis (Katsanis, 2004; Khan *et al.*, 2016; Yildiz Bolukbasi *et al.*, 2018).

Obesity is one of the hallmarks of BBS, and its incidence varies from 72-86% in the cases reported (Moore *et al.*, 2005). Interestingly, birth weight is usually within the normal range and obesity develops in infancy/childhood, which may be of truncal or generalized type.

The primary molecular defect that underlies BBS is dysfunction of primary cilia due to problems in the assembly or function of the BBSome complex which is composed of proteins that seven of the 24 BBS genes encode. Many of the proteins encoded by known BBS genes are also ciliary and localize to centrosome and primary cilia. BBSome plays a role in primary cilia biogenesis by transporting signaling receptors to and from cilia (Scheidecker *et al.*, 2014).

One gene associated with obesity is *19-kd Centrosomal Protein* (*CEP19)*, encoding a ciliary protein localized to the centrosome and primary cilia. As an animal model, mouse knockout had been generated; it was morbidly obese, hyperphagic, glucose intolerant and insulin resistant as in humans homozygous for the gene mutation identified (Shalata *et al.*, 2013).

### 1.4. Pleuroparenchymal Fibroelastosis (PPFE)

Pleuroparenchymal fibroelastosis (PPFE) is an interstitial (the supportive framework throughout the lung, composed of connective tissue) lung disease characterized by predominantly upper-lobe fibrosis (Amitani *et al.*, 1992). It is a rare idiopathic disease with initial symptoms of dyspnea (respiratory distress), dry cough and chest pain due to pneumothorax in some patients. Although it was observed that PPFE is a slowly progressive disorder (Watanabe, 2013), some reports claim that in some cases it is rapidly progressive (Nakatani *et al.*, 2015). Etiology of the disease is unknown, but chemotherapy, drugs, auto-immunity, recurring infections and previous lung or bone marrow transplantation are suggested to be predisposing factors. The majority of the cases are not familial and thus considered not to be due to genetic factors; hence, they are considered as idiopathic PPFE (Cheng and Chuah, 2016). Its prevalence is <1 in 1 million (Orphanet).

PPFE manifests also with flattened thoracic cage, and restricted ventilation is another symptom (Watanabe, 2013). Although multidisciplinary approach such as surgical lung biopsy and CT-guided transthoracic core lung biopsy is needed for diagnosis, the definite feature is upper zone fibrosis of the visceral pleura, prominent homogeneous subpleural

intra-alveolar fibrosis (connective tissue deposition in the intra-alveolar spaces) with alveolar septal elastosis, and sparing of the parenchyma away from the pleura (Kusagaya *et al.*, 2012). PPFE can be misdiagnosed as usual interstitial pneumonia (UIP). In contrast to PPFE which affects the upper lobes, in UIP lower lobes are affected and the original parenchymal structure is destroyed whereas in PPFE alveolar structure is preserved or thickened (Cheng and Chuah, 2016).

Four telomere-related genes have been associated with sporadic and familial idiopathic pulmonary fibrosis which is a type of interstitial lung disease which is a similar disease with PPFE after telomere shortening was observed in patients (Cronkhite *et al.*, 2008; Stuart *et al.*, 2015). In 115 patients with variable interstitial lung disease including PPFE Newton and colleagues (2016) found mutations in one of four telomere maintenance machinery genes, namely, telomerase reverse transcriptase (*TERT*), telomerase RNA component (*TERC*), regulator of telomere elongation helicase 1 (*RTEL1*) and poly(A)-specific ribonuclease (*PARN*) (Newton *et al.*, 2016). This study linked telomere-related genes to pulmonary fibrosis. *TERT* mutations were found in 19 of the 106 patients (~18%), the largest number was in familial pulmonary fibrosis (Diaz de Leon *et al.*, 2010). In the analysis of *TERT* and *TERC* in patients with familial pulmonary fibrosis, more mutations were found in *TERT* (Garcia, 2011)

## 1.5. Linkage Analysis

The first step in disease gene search is the identification of the disease locus. The most common method is linkage mapping whenever a large number of members of the family participate in the genetic study. Genetic linkage is the tendency of alleles that are close to each other on the same chromosome to be inherited together in meiosis (Rimoin *et al.*, 2013). During chromosomal crossover in meiosis, genetic markers which are close to each other are less likely to segregate to homologous chromatids, that is, when the two genes are physically close to each other, the chance of recombination is low (Griffiths *et al.*, 2000). When two genes/loci are not close to each other or are on different chromosomes, then they are considered as unlinked. During meiosis, the probability of

recombination between those unlinked genes/loci is 50%. This probability of crossover which is defined as "recombination fraction" is denoted θ (theta). The maximum value for θ is 0.5, indicating that two loci do not cosegregate due to being either far apart on the same chromosome or on different chromosomes. The unit of recombination frequency used to define genetic distance on a chromosome is centimorgan (cM). One cM corresponds to roughly 1% probability of recombination and 1Mb on the physical map.

In disease gene mapping, different types of genetic markers such as single nucleotide polymorphisms (SNPs), short tandem repeats (STRs or microsatellites), variable number of tandem repeats (VNTRs) and restriction fragment length polymorphisms (RFLPs, a kind of SNP) are used to track recombination. The most commonly used type of genetic marker for linkage analysis is SNP, since a large number of markers (300k to 2.5M) can be analyzed easily using microarrays. Genotype data have easy online accessibility but have the drawback of being not very informative, since a SNP has only two alleles. Computer based programs can analyze a cluster of them together as a haplotype and thus overcome the drawback.

Linkage analysis is a statistical method used to map trait loci that are inherited according to Mendelian rules by using genome-wide markers in a family or even by non-Mendelian rules. The two types of analysis are parametric linkage analysis for a Mendelian form of inheritance and model-free (non-parametric) analysis for complex traits (Rimoin *et al.*, 2013). LOD (logarithm of odds) score is calculated to evaluate the results of a parametric linkage analysis (Morton, 1955).

Parametric linkage analysis is the most common approach to map genes for Mendelian diseases in large families. Parameters such as type of inheritance, penetrance and disease-allele frequency are taken into account. LOD score represents the logarithms of the odds ratio which is the ratio of the probability of a locus (marker set) being linked to the probability of not linked (Risch, 1992). The two hypotheses of no linkage to the trait (the null hypothesis, $\theta=0.5$) and linkage present (alternative hypothesis, $\theta <0.5$) are tested by LOD score analysis. A LOD score of 3 is generally considered as a threshold for significance of linkage for autosomal traits, and thus the null hypothesis is rejected. It indicates that the observed linkage most likely has not occurred by chance in 1000 to 1

odds (Morton, 1955). It is not always the case since not all kinships in a family are introduced due to limitations of the computer programs, and thus the real score could be lower than 3. A LOD score below -2 indicates no linkage to the locus investigated, and LOD sores between 3 and -2 are inconclusive.

LOD score calculations can be performed as either a two-point or a multipoint analysis. In the former, co-segregation of one marker at a time with the trait is investigated, whereas in the latter form more than one marker is included in the calculations. Linkage analysis is performed mostly by multipoint analysis to evaluate several SNP markers together as a haplotype.

Several software such as Allegro, SimWalk, GeneHunter is available for linkage analysis. easyLINKAGE package is a platform that converts all input information to a standard form for linkage analysis (Hoffmann and Lindner, 2005).

### 1.6. Homozygosity Mapping

Homozygosity mapping is another strategy to identify disease locus in consanguineous families afflicted with recessive diseases. It is the most common method whenever DNA samples available are mostly from affected members of the family. The haplotype in the region of the causative mutation is likely to be inherited from a recent common ancestor in an inbred population or a consanguineous family (Lander and Botstein, 1987). The regions of homozygosity shared by affected individuals can be detected by investigating SNP genotype data in MS excel or using appropriate software such as online HomozygosityMapper. In order to determine whether the homozygosity in a region is possibly due to a haplotype that is identical by descent (IBD), i.e. originated from the same ancestral chromosome, haplotypes are constructed. IBD means that the identical parental haplotypes have descended from the same recent common ancestor.

## 1.7. Whole Exome Sequencing (WES)

At the locus identified or the candidate loci detected, all genes are evaluated to assess whether any of them can be a good candidate to be the disease gene. If none of the genes is assessed as possibly associated with the disease, then WES is launched. After the development of massively parallel nucleic acid sequencing or next generation sequencing (NGS), our vision and knowledge about human genome has been enlarged. The NGS technology includes several applications such as whole genome sequencing (WGS), whole exome sequencing (WES), whole transcriptome analysis (RNA-seq), genome-wide profiling of epigenetic marks (methyl-seq) and chromatin structure (ChIP-seq).

Human exome analyzed using microarrays constitutes less than 2% of the genome and includes nearly 180,000 exons and approximately 30 million bases (Ng *et al.*, 2009). Mutations in these regions are more likely to be disease-causing as compared to mutations in other parts of the genome (Stenson *et al.*, 2009). Thus, WES, with its high-throughput capacity and low cost, represents an enriched subset of the genome and is a highly preferred, efficient and powerful method for both Mendelian and complex disease studies. As compared to Sanger sequencing, it is faster and more cost effective (Ku *et al.*, 2011). For a monogenic disease, by trio sequencing where parents and the affected infant/child are genotyped, the causative mutation can be distinguished among many variants detected at the same time. This strategy is especially useful for detecting de novo causal mutations and mutations in recessive diseases that develop later than infancy/childhood. Thus, exome sequencing has become a very powerful tool in identifying new Mendelian disease genes also in small families and has been successful in approximately 60% of the projects (Gilissen *et al.*, 2012).

Besides the many advantages of WES, there are some drawbacks. Due to a higher base calling error rate per nucleotide (5%) as compared to Sanger sequencing (near zero), variants of interest need to be validated by Sanger sequencing (Matullo *et al.*, 2013). Another drawback is that some exons in a candidate region might not be covered in the capture chip. Thus, it is possible to miss some of the variants possibly including the causative variant.

To explain briefly how exome sequencing is performed, DNA is sheared into fragments, and targeted regions (exons and flanking regions) are selectively hybridized to biotinylated oligonucleotide probes. Magnetic beads bind to those probes, and the captured regions are amplified. As a last step, the amplification products are subjected to sequencing on next generation sequencing platforms, creating short (80-100 nucleotides) sequence reads. By bioinformatics tools, reads are aligned to the reference genome. Sequences different from the reference sequence, i.e. variants, are called. Among all variants called, candidate variants are obtained following a prioritization strategy and subsequent filtering.

## 1.8. Disease Gene Identification Strategy

The general approach applied in this thesis work can be summarized in two steps: First, candidate gene locus/loci are determined via SNP genotyping and subsequent linkage analysis or homozygosity mapping by online tools or ocular investigation. Then, the causative mutation is searched in the exome sequence results at those loci. The strategy that is followed to identify the genes responsible for the disease afflicting a family is described below.

### 1.8.1. Initial Work

- For the family that will be studied, phenotype of each individual is defined.
- Pedigree of the family is constructed.
- Blood samples are collected from family members after written informed consent has been obtained according to the regulations of the institutional ethical review board.
- Whole-genome scan (SNP genotyping) is performed for all affected and a few unaffected individuals in the family.

### 1.8.2. Linkage Analysis and Homozygosity Mapping

- Linkage analysis is performed with genotype data by using Allegro program to calculate LOD scores. For large pedigrees (>20 bits), LOD score calculations are first performed using a simplified pedigree (≤20 bits).
- For regions with relatively high LOD scores (>2) fine mapping is performed with the actual pedigree and including 1-cM flanking regions. If pedigree is larger than 20 bits, SimWalk program is used.
- Regions with LOD scores above the cut-off are listed and then investigated in Excel displaying the genotype data to determine whether the region is IBD.
- Haplotypes at the candidate loci are constructed by including 0.5-1.0-cM flanking regions using Allegro program in order to investigate possible IBD.
- Alternatively, Homozygosity Detector program on Genome Studio can be used to detect homozygous regions in case the locus has been missed by linkage analysis or linkage analysis was not feasible.
- Candidate regions are investigated in OMIM/Map viewer Morbid/Disease Map to find out whether there are any similar phenotypes or candidate or associated genes in the regions.

### 1.8.3. Deletion-Duplication Analysis

- Deletion-duplication analysis is performed using the cnvPartition (v3.2.0) CNV Analysis Plug-in for Genome Studio for all family members to investigate whether there are any deleted or duplicated regions shared by patients.
- If any such duplicated or deleted region is detected, Database of Genomic Variants (DGV) is investigated to see whether any overlapping deleted or duplicated region is reported.

### 1.8.4. Searching for Candidate Genes and the Causative Variant by Exome Sequence Analysis in a Recessive Disease

- Whole exome sequencing (Illumina Omniexpress-24 chip) is performed for one or more affected individuals in the family.
- The list of variants is obtained from Macrogen, or raw data which are obtained as "fastq" file are aligned to the reference genome by BWA, variant calling is performed by SamTools and variants are annotated by ANNOVAR.
- Variants in the candidate regions or at the identified disease locus are listed.
- All rare/novel deleterious homozygous mutations are evaluated to see whether any of them possibly resides at a candidate locus.
- At the candidate loci presence of the variants that are listed with frequencies 0.99 or 1.0 in human reference sequence are searched in the exome file, because the reference sequence lists the common variant as the rare variant. If the patient has the actual reference base, it is listed as a variant in the exome sequence data. If the variant is not listed in the patient's exome data, then the patient may be having the rare variant.
- Candidate variants are obtained after the elimination of unlikely variants, i.e. minor allele frequency (MAF) >0.01 or 0.005, intergenic and alt base/total base ratio <0.6.
- Filtered variants are investigated on IGV (Integrative Genomics Viewer) to confirm that the candidate variant is read as indicated in the annotation file.
- EVS (Exome Variant Server, ~6500 exomes), ExAC (Exome Aggregation Consortium), gnomAD (Genome Aggregation Database), 1000 Genomes and TEVD (Turkish Exome Variant Database, ~1,150 exomes) are investigated for the presence and MAF of the variant, if not specified on the list obtained from wAnnovar.
- Remaining variants are checked in-lab exome data (56 exomes), i.e., other exome files in our laboratory. Variants present in at least one of the files are filtered out.
- Whether the variants are in the Pakistani and Icelandic loss of function (LoF) mutation lists are investigated (Gudbjartsson *et al.*, 2015; Narasimhan *et al.*, 2016; Saleheen *et al.*, 2017).

- The significance of the gene is evaluated by investigating how heavy the mutation load is as reported in databases such as ExAC and EVS and published articles.
- Priority is given to exonic and splicing variants.

### 1.8.5. Evaluation of the Candidate Variants

- Computational algorithms that predict mutation severity, such as Mutation Taster, SIFT, PolyPhen-2, PROVEAN, SpliceFinder, NNSplice and UMD-Predictor are employed to assess whether the variant is potentially harmful to the protein. Some of these online tools are not applicable to all types of variants.
- If a mutation is nonsynonymous, how the amino acid substitution affects the protein is evaluated by using online tools mentioned above and comparing the biochemical features of the native and substituted amino acids. Also, conservation of the altered residue among species is evaluated by using the HomoloGene database and UCSC 100 vertebrates track.
- If a mutation is synonymous, difference in codon usage is investigated.
- If a mutation is nonsense, how much of the protein is truncated and the features of the truncated part/domain are investigated.
- If the mutation is splicing, how it affects splicing of the pre-mRNA is evaluated via online tools such as Splice Finder and NNSplice.
- For mutations at UTRs, ENCODE is investigated for any regulatory element in that region. For a 5' UTR variant, whether the variant is in the promoter, enhancer or silencer is investigated. For a 3' UTR variant, whether it is in the polyadenylation signal site or the cleavage site is evaluated.
- If a mutation is intronic, its position in other isoforms is evaluated in UCSC ENCODE database.
- Which tissues the gene is expressed in is investigated using the Expression Atlas, Human Protein Atlas, Unigene etc.

### 1.8.6. Validation of the Candidate Variants

- The candidate variant is validated by Sanger sequencing in the patient whose exome file is available, and later some other family members and a control sample are tested.
- Available family members are screened for the variant by HRM (High resolution Melting) or SSCP (Single Strand Conformational Polymorphism) analysis to investigate segregation with the disease.
- A population sample is tested to make sure that the variant is not frequent in the population. This was applied only to IID Family; for others databases such as ExAC and gnomAD that include exome and genome data of at least 10,000 Pakistanis in the South Asian samples. Candidate variants are also interrogated in TUBITAK (MAM) for the Turkish population.
- The possible effect of the mutation is investigated by literature search or with further molecular studies with a collaborator.
- mRNA level assay using quantitative real-time PCR can be performed if needed to assess whether the variant possibly impairs the transcription of the gene.

# 2. PURPOSE

Among the vast number of genomic variants in a patient, to identify the one which is responsible for the disease is a challenge that requires various strategies that depend on the size of the family, inheritance pattern, severity of the phenotype and population frequency of the variant. Just the disease gene identification itself is a very long process, scientifically important and merits publication. Genetic studies substantially increase our understanding of cellular mechanisms, occasionally uncover novel such mechanisms and unravel molecular pathogenesis underlying diseases.

The purpose of this thesis study was to localize and identify the genes responsible for six rare diseases in the eight study families. In the families with different diseases, besides identifying the disease genes, various strategies were used for further investigations. In IID family, by using relative quantification, expressions of the gene in different tissues was investigated, and in BBS family mutations in all BBS-related genes were evaluated with the aim of assessing the effect of the mutation load on disease severity.

# 3. MATERIALS

## 3.1. Subjects

Genetic studies on eight families are included in this thesis study. Five of the families are Pakistani, and the last three are Turkish. All are consanguineous. Blood or DNA samples of Pakistani families as well as the clinical evaluations were obtained from Assoc. Prof. Sajid Malik at Quad-i Azam University in Pakistan. Blood samples of two of the Turkish Families and the clinical evaluations were obtained from Prof. Şevket Özkaya at Bahçeşehir University, and blood samples for the remaining Turkish family were obtained from Assoc. Prof. Selvi Aşker at Van Yüzüncü Yıl University. Informed consent was obtained from/for participants in accordance with the regulations of the Ethical Review Committee of Quaid-i-Azam University and Boğaziçi University Institutional Review Board for Research with Human Participants, both of which approved the study protocol.

### 3.1.1. Isolated Intellectual Disability (IID) Family

The family is from Southern Punjab in Pakistan. The first-cousin parents have four sibs afflicted with mild to moderate ID who additionally present speech delay, stuttering and certain early onset behavioral problems. Pedigree of the family is presented in Figure 3.1. The early onset behavioral problems include attention deficit, seclusion, hyperactivity, apraxia of speech and stuttering. Affected individuals are able to perform simple tasks and take personal care but require some aid and supervision in more complex daily activities. SNP genotype data of parents and five sibs were generated. DNA sample of affected individual 402 was subjected to exome sequencing. SNP genotype data of another unaffected sib were incompatible with both parents, and thus we did not include his data in the study.

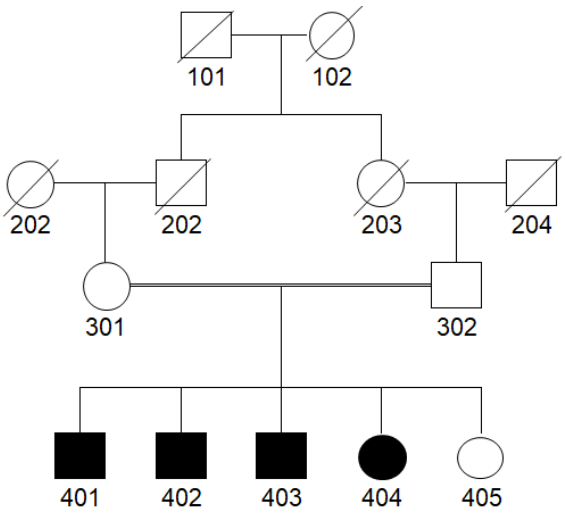



Figure 3.1. Pedigree of IID family. SNP genotype data were generated for parents and sibs.

### 3.1.2. Intellectual Disability and Hypothyroidism (IDH) Family

The major clinical features of the disease in this family are severe ID, developmental delay, speech and hearing problems, low vision, self-mutilation, aggressive behavior and hypothyroidism. However, severe ID is present in only two individuals (311 and 325). Pedigree of the family is presented in Figure 3.2. Symptoms of hypothyroidism emerged in infancy or early childhood in all affected subjects. Phenotypes of affected individuals were variable ranging from decreased activity, excessive sleeping, recurrent or prolonged jaundice (yellow discoloration of the skin), poor feeding and weight gain, constipation, protruding tongue and hypotonia (low muscle tone). DNA samples of four affected and eight unaffected individuals were available for the study. SNP genotyping was performed for three affected and two unaffected individuals, and exome sequencing for affected 322.

### 3.1.3. Syndromic Intellectual Disability (SID) Family

In total six individuals in two branches of the family are afflicted with a severe syndrome manifesting with intellectual disability (ID), developmental delay, facial dysmorphism, speech and hearing problems, low vision, self-mutilation, and aggressive behavior. They are completely dependent for all daily functions. Other clinical features

observed not in all affected subjects include muscle wasting, bed-ridden posture, lower-limb weakness, walking problem, loud and hoarse voice, dry and coarse skin, and unable to control urine and defecation (enuress and encopresis).

DNA sample of four affected and 11 unaffected members of the family were available. Pedigree of the family is presented in Figure 3.3. SNP genotype data were generated for four affected and three unaffected members. Exome sequencing was performed for patient 502.

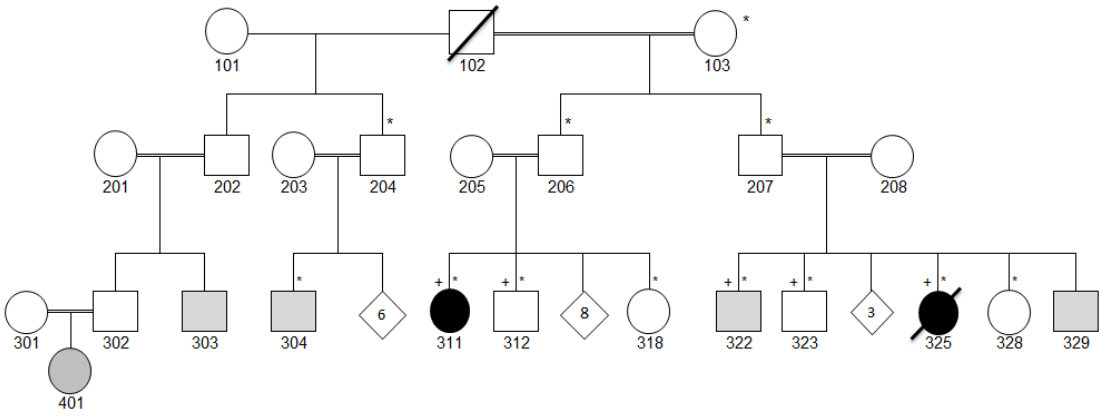



Figure 3.2. Pedigree of IDH family. DNA was available for individuals marked *. SNP genotype data were generated for individuals marked +. Black shading indicates hypothyroidism and ID and gray shading hypothyroidism.

### 3.1.4. Spinocerebellar Ataxia (SCA) Family

The large consanguineous family is from Southern Punjab in Pakistan and is afflicted with intellectual disability (ID), delayed developmental landmarks, and limb and gait ataxia as primary features. DNA samples of 16 individuals (eight affected and eight unaffected) from five branches of the family were available. Pedigree of the family is presented in Figure 3.4. All nine affected individuals and their eight unaffected relatives were physically examined by local physicians. All affected individuals exhibited signs of cerebellar ataxia such as dysmetria (medical condition that a person cannot judge the distance and thus cannot measure range of motion), dysdiadokinesia (inability to perform a rapid change of motion) and inability to tandem walk. Other common findings were attention deficit, aggressive behavior, poor eye contact and climacophobia (fear of

climbing stairs). Expressive aphasia and dysarthria were evident in seven of them. High arched palate and tremor were also observed as minor symptoms in some of the subjects. SNP genotyping was performed for seven affected and three unaffected individuals, and exome sequencing was performed for patient 412.

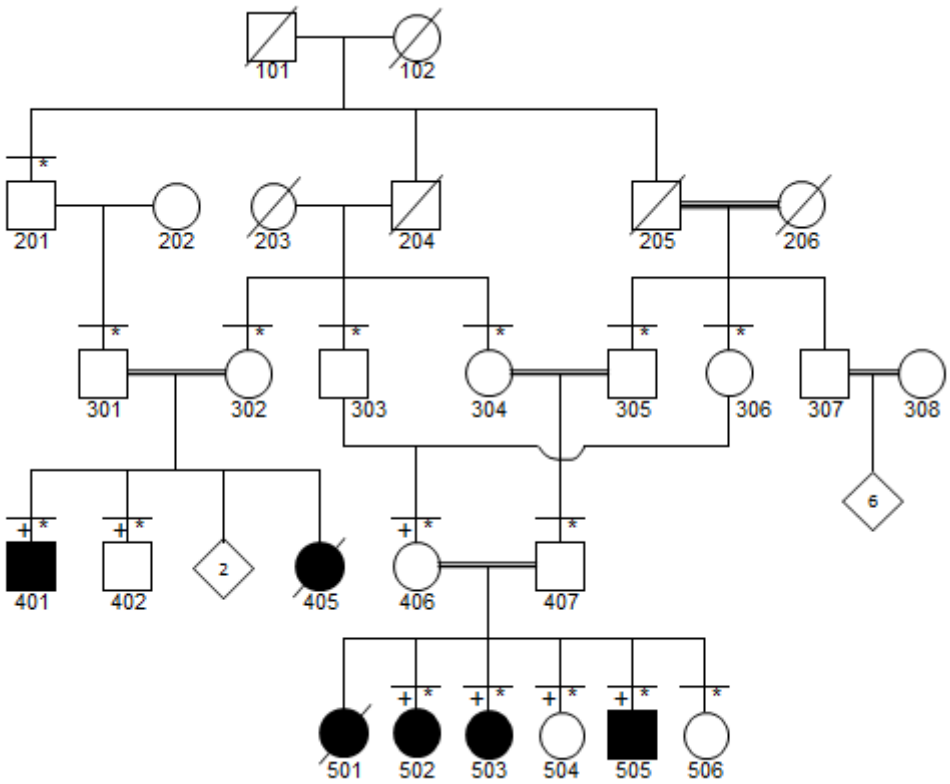



Figure 3.3. Pedigree of SID family. A horizontal line above the symbol indicates that the individual was physically examined. *Available for all genetic studies, +Available for mutation analysis.

### 3.1.5. Bardet-Biedl Syndrome (BBS) Family

This family is afflicted with Bardet-Biedl Syndrome (BBS), a group of autosomal recessive ciliopathies. Patients had postaxial polydactyly plus variable other clinical features including rod-cone dystrophy, obesity, intellectual disability, renal malformation, developmental delay, dental anomalies, speech disorder, and enlarged fatty liver. Clinical findings are compiled in Table 3.1.

DNA samples of seven affected and six unaffected individuals from the two branches of the family were subjected to SNP genotyping as shown in Figure 3.5. Exome sequencing was performed for two patients (503 and 509) from different branches.

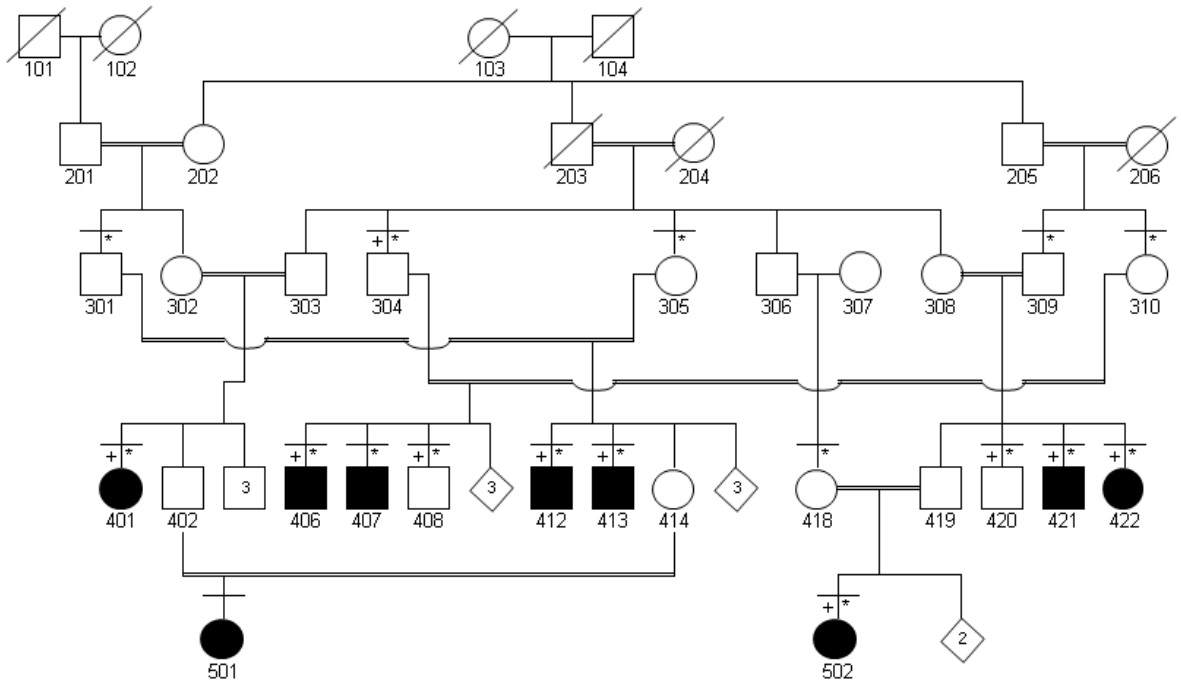



Figure 3.4. Pedigree of SCA family. Individuals included in the study are marked with asterisks, those included in SNP genotyping are marked with plus signs, and a horizontal line above a symbol indicates that the individual is clinically evaluated.

### 3.1.6. Pleuroparenchymal Fibroelastosis (PPFE) Families

These families are Turkish, parents are either first cousins or double first cousins, and initial diagnosis was pleuroparenchymal fibroelastosis (PPFE). Diagnosis was made by lung biopsy in PPFE1 and PPFE3 families. For PPFE1 blood samples of parents and three affected and one unaffected sibs were available. SNP genotyping was performed for all of them (Figure 3.6). Also, a cousin was subjected to SNP genotyping. Afterwards her data was not included in the analyses since the clinical evaluation was inconclusive. The two affected brothers had interlobular septal and pleural thickening and some other features compatible with PPFE. Elder brother also had low tidal volume. The brothers were lost due to respiratory failure. The elder sister, even though symptom-free, was initially diagnosed

with PPFE due to slight opacities in the lung parenchyma although she had no interlobular septal thickening.

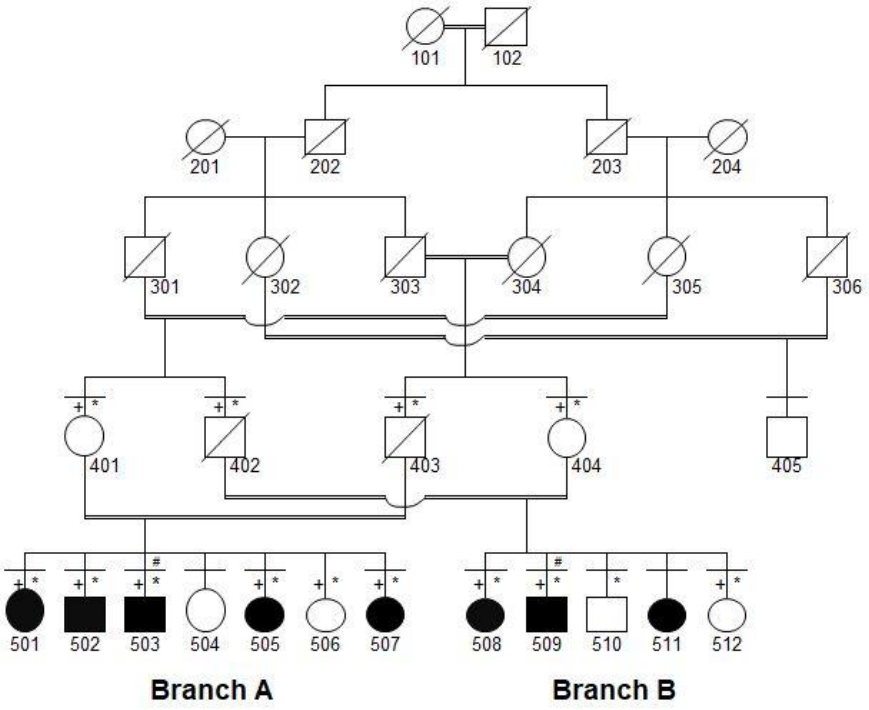



Figure 3.5. Pedigree of BBS family. A horizontal line above the symbol indicates that the individual is physically examined. *Available for genetic studies, +Available for mutation testing only, #Exome sequenced individuals.

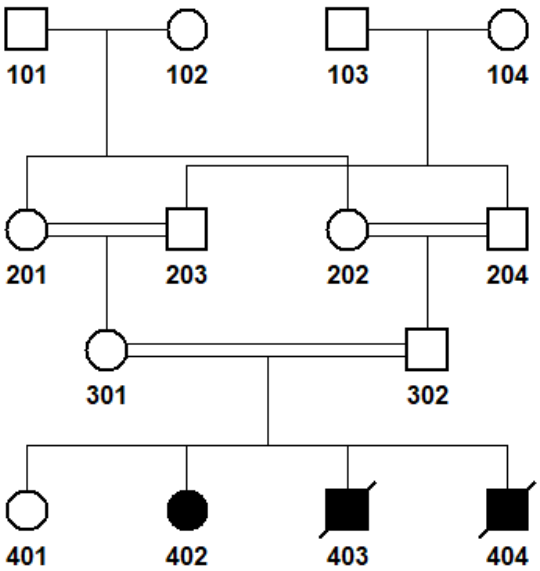



Figure 3.6. Pedigree of PPFE1 family.

Table 3.1. Characteristics of BBS family members.

| Features | Patients | | | | | | | | Unaffected relatives | | | | | | | | Concordance in patients investigated |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **501** | **502** | **503** | **505** | **507** | **508** | **509** | **511** | **401** | **402** | **403** | **404** | **504** | **506** | **510** | **512** | |
| Sex | F | M | M | F | F | F | M | F | F | M | M | F | F | F | M | F | |
| Age (years)* | 28 | 25 | 25 | 22 | 16 | 25 | 24 | 18 | 52 | 61 | 50 | 44 | 22 | 18 | 20 | 16 | |
| **Clinical findings** | | | | | | | | | | | | | | | | | |
| ***Primary features*** | | | | | | | | | | | | | | | | | |
| Rod-cone dystrophy | - | N/A | +, RP | +, ONA | N/A | + | +, early features | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 4/7 |
| Polydactyly in hands/feet | B/A | -/A | A/A | B/A | B/- | B/- | A/MA | -/A | - | - | - | - | - | - | - | - | 6/8, 6/8 |
| Obesity (BMI)** | Over-weight (25.6) | Over-weight (26.3) | - (20.0) | Over-weight (27.0) | - (23.0; 74th centile) | + (37.5) | + (36.3) | + (33.6) | - (20.6) | + (30.4) | - (23.5) | - (23.5) | Over-weight (28.8) | Over-weight (25.5) | - (22.5) | - (24.5; 85th centile) | 3/8 |
| Intellectual disability | - | - | ++ | ++ | - | + | + | - | - | - | - | - | - | - | - | - | 4/8 |
| Renal anomaly | - | N/A | Par. D | - | N/A | Par D | - | N/A | N/A | N/A | - | N/A | N/A | N/A | N/A | N/A | 2/5 |
| ***Secondary features*** | | | | | | | | | | | | | | | | | |
| Enlarged fatty liver | - | N/A | - | + | N/A | + | + | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 3/5 |
| Physical disability; neuromotor problem | - | - | ++, # | - | - | - | - | - | - | - | - | - | - | - | - | - | 1/8 |
| Speech disorder | - | - | + | + | - | - | - | - | - | - | - | - | - | - | - | - | 2/8 |
| Aggressive behavior | - | - | - | - | - | - | + | - | - | - | - | - | - | - | - | - | 1/8 |
| Dental anomalies | - | - | + | + | - | - | + | - | - | - | - | - | - | - | - | - | 3/8 |
| Developmental delay | - | - | ++ | ++ | - | - | + | - | - | - | - | - | - | - | - | - | 3/8 |
| Diabetes | - | N/A | - | N/A | N/A | + | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 1/3 |
| Syndactyly | - | - | - | - | - | - | 2/3 toes, R. | - | - | - | - | - | - | - | - | - | 1/8 |
| ***Others*** | | | | | | | | | | | | | | | | | |
| Exotropia of right eye | + | N/A | - | + | N/A | - | + | N/A | - | - | - | - | - | - | - | - | 3/7 |
| Shortness of breath | + | - | + | + | + | + | + | N/A | - | - | - | - | - | - | - | - | 6/7 |
| Diverse | RU | - | X | - | - | - | Y | - | Goiter | - | - | - | - | - | - | - | - |

*, age at the time of last examination

**for ages 20 years and over, normal BMI is 18.5-24.9, overweight 25-30, obese >30, and morbidly obese >40. BMI for ages below 20 was calculated using https://nccd.cdc.gov/dnpabmi/Calculator.aspx?CalculatorType=Metric as recommended by http://www.who.int/growthref/who2007_bmi_for_age/en/

A, bilateral postaxial type A; , bilateral postaxial type B; bilateral MA, mesoaxial; -, feature absent; +, feature present; ++, severe phenotype; #, walks with support; N/A, phenotype not ascertained; RP, retinitis pigmentosa early features; ONA, optic nerve atropy R, right; ParD, Paranchymal disease; RU, Retroverted uterus; X, Sparse hair; enuresis; enlarged head; frontal bossing, Y, Overriding 6th toe, left; hyperphagy; insomnia; low hairline, hypertensive

For PPFE2, initial diagnosis was PPFE by clinical evaluation; however, later diagnosis became uncertain. Biopsy could not be performed. Samples of three affected siblings were subjected to SNP genotyping and of 402 to exome sequencing (Figure 3.7).

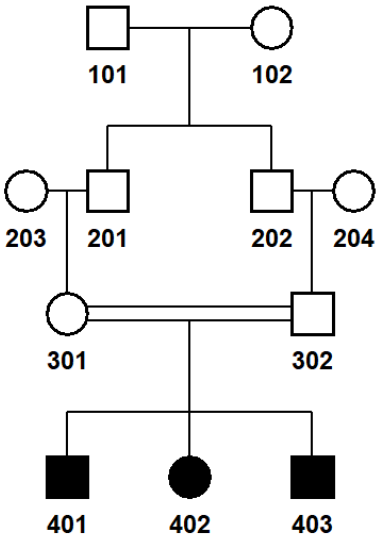



Figure 3.7. Pedigree of PPFE2 family.

In PPFE3 family, two brothers and a sister were diagnosed with PPFE by biopsy. Pedigree of the family is presented in Figure 3.8. One affected (401) individual afflicted with PPFE were subjected to exome sequencing. The other affected brother and the sister had died of PPFE. There are two unaffected siblings.

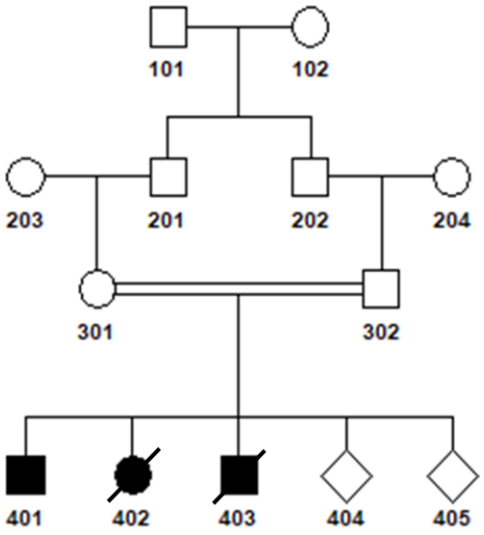



Figure 3.8. Pedigree of PFFE3 family.

## 3.2. Chemicals

All chemicals used throughout this study were purchased from Merck (Germany), Sigma (USA), Riedel de-Häen (Germany), Carlo Erba (Italy) or Biochrom (Germany), unless stated otherwise in the text.

### 3.2.1. DNA Extraction from Blood

All buffers and solutions used in DNA extraction from peripheral blood are given in Table 3.2.

Table 3.2. List of buffers and solutions used in DNA extraction from blood samples.

| Buffer/Solution | Ingredients |
|---|---|
| Red Blood Cell Lysis Buffer | 155 mM $NH_4Cl$, 0.1 mM $Na_2EDTA$ (pH 7.4) and 10 mM KHCO3 |
| Nucleus Lysis Buffer: | 400 mM NaCl, 2 mM $Na_2EDTA$, 10 mM Tris (pH 8.2) |
| Proteinase K | 20 mg/ml proteinase K in $dH_2O$ |
| Sodiumdodecylsulfate (SDS) | 10% SDS (w/v) in $dH_2O$ |
| Ammonium Acetate ($NH_4Ac$) | 7.5 M $CH_3COONH_4$ in $dH_2O$ |
| EtOH | Absolute Ethanol |
| TE Buffer | 20 mM Tris-HCl (pH 8.0), 1 mM EDTA |

### 3.2.2. Polymerase Chain Reaction (PCR)

The list of buffer and solutions used in polymerase chain reaction (PCR) is given in Table 3.3.

Table 3.3. List of ingredients used in polymerase chain reaction (PCR).

| Chemical | Ingredients |
|---|---|
| dNTP | 12.5 mM each of dATP, dCTP, dGTP, and dTTP in $dH_2O$ (Roche, Germany and Fermentas, Lithuania) |
| Taq Polymerase | Produced and extracted from bacteria (*Thermus aquaticus*) |
| Primers | Ordered from Macrogen Inc. |
| 10X PCR Buffer | 20 mM $MgSO_4$, 100 mM KCl, 100mM $(NH_4)_2SO_4$, 1%Triton X-100, 1 mg/ml BSA, 200 mM Tris-HCl (pH 8.8) |

### 3.2.3. Agarose Gel Electrophoresis

The list of materials used for agarose gel electrophoresis is given in Table 3.4.

Table 3.4. List of chemicals, buffers and solutions used in agarose gel electrophoresis.

| Chemical | Ingredients |
|---|---|
| Agarose | Agarose (Pronadisa, Spain) |
| Ethidium Bromide | 10 mg/ml in $dH_2O$ |
| 0.5X TBE Buffer | 20 mM EDTA, 0.89 M boric acid, 0.89 M Trizma base (pH 8.3) |
| 6 X Loading Dye | 50% Glycerol, 60 mM EDTA, 2.5 mg/ml bromophenol blue and/or 2.5 mg/ml xylene cyanol, 10 mM Tris-HCl (pH 7.6) |

### 3.2.4. High Resolution Melting (HRM) Analysis

The list of materials used for high resolution melting (HRM) analysis is given in Table 3.5.

Table 3.5. List of chemical used in high resolution melting (HRM) analysis.

| Chemical | Company |
|---|---|
| LightCycler 480 High Resolution Melting Kit | Roche, Germany |
| $MgCl_2$ | Roche, Germany |

### 3.2.5. Single Strand Conformational Polymorphism (SSCP) Analysis

The list of materials used for single strand conformational polymorphism (SSCP) analysis is given in Table 3.6.

Table 3.6. List of buffers and solutions used in strand conformational polymorphism (SSCP) analysis.

| Buffer/Solution | Content |
|---|---|
| Acrylamide-bisacrylamide (Stock) | 40% (37.5:1) in $dH_2O$ |
| 10X TBE buffer | 20 mM EDTA, 0.89 M boric acid, 0.89 M Trizma base (pH8.3) |
| Glycerol | 50% glycerol in $dH_2O$ |
| APS | 10% ammonium peroxydisulfate |
| TEMED | N,N,N,N-tetramethylethylenediamine |
| 10X Page Dye | 95% formamide, 20 mM EDTA, 0.05% xylene cyanol, 0.05% bromophenol blue |

### 3.2.6. Silver Staining

The list of materials used for Silver Staining is given in Table 3.7.

Table 3.7. List of solutions used in silver staining. .

| Name | Ingredients |
|---|---|
| Staining solution | 0.1% $AgNO_3$ in $dH_2O$ |
| Developing solution | 1.5% NaOH, 0.015% formaldehyde, 0.01% $NaBH_4$ in $dH_2O$ |

## 3.3. Kits

A list of commercial kits used in this study is given in Table 3.8.

Table 3.8. List of commercial kits used.

| Name | Used for | Company, Country |
|---|---|---|
| High Resolution Melting Kit (LightCycler 480) | Heteroduplex analysis (mutation screening) with LightCycler 480 device | Roche, Germany |
| Accumelt HRM SuperMix (LightCycler 480) | Heteroduplex analysis on LightCycler 480 | Quanta Biosciences, USA |
| SYBR Green I Master Kit | Relative quantification on LightCycler 480 and PCR reactions | Roche, Germany |

### 3.4. Oligonucleotide Primers

Oligonucleotide primers were designed with Primer3 software, and whether they form hairpins or form duplexes was checked by OligoCalc tool. Primer specificity was analyzed via UCSC in silico PCR tool on UCSCS Genome Browser. They were purchased from Macrogen Inc. (South Korea). Primers were dissolved in $dH_2O$ and used as 10 μM dilutions for PCR reactions and sequencing. Primer sequences are given in Table 4.4.

### 3.5. DNA Molecular Weight Markers

pUC19 DNA/*Msp*I marker and GeneRuler 1-kb DNA ladder were purchased from Fermentas (Lithuania). 100 bp DNA ladder was purchased from Biomatik (Canada).

### 3.6. Electronic Databases

A list of electronic databases, computational algorithms, and online tools and their URLs and descriptions used in this study is given in Table 3.9.

Table 3.9. List of electronic databases, computational algorithms, online tools and bioinformatics tools.

| Name (URL) | Description |
|---|---|
| **Databases** | |
| Online Mendelian Inheritance in Man (OMIM) (http://www.ncbi.nlm.nih.gov/Omim) | Database of human genes and genetic phenotypes |
| UCSC Genome Browser (http://genome.ucsc.edu) | Provides genome sequence. Used to analyze genomic data |

Table 3.9. List of electronic databases, computational algorithms, online tools and bioinformatics tools (cont.).

| Name (URL) | Description |
|---|---|
| **Databases** | |
| Ensembl Genome Browser (http://www.ensembl.org/index.html) | Database for vertebrate genomes. Provides multiple alignments, regulatory function predictions and collection of disease data |
| NCBI Gene (http://www.ncbi.nlm.nih.gov/gene) | Database of refseq gene data |
| NCBI HomoloGene (http://www.ncbi.nlm.nih.gov/homologene/) | Detects homology among species and provides multiple protein alignments |
| NCBI PubMed https://www.ncbi.nlm.nih.gov/pubmed/ | Archive of biomedical and life sciences literature |
| NCBI UniGene http://www.ncbi.nlm.nih.gov/unigene/ | Expression data of genes in various tissues |
| UniProt (Universal Protein Resource) http://www.uniprot.org/ | A database of protein sequence and structural and functional information |
| GeneDistiller 2014 http://www.genedistiller.org/ | Lists genes in a given region or by keywords |
| Exome Aggregation Consortium (ExAC) (http://exac.broadinstitute.org/) | Collection of exome data of 60,706 individuals. Provides frequencies of variants |
| Genome Aggregation Database (gnomAD) (http://gnomad.broadinstitute.org/) | Data set of 123,136 exome sequences and 15,496 whole-genome sequences |
| dbSNP http://www.ncbi.nlm.nih.gov/SNP/ | Database of single nucleotide polymorphisms |
| Database of Genomic Variants (DGV) http://projects.tcag.ca/variation/ | Catalogue of structural variation (SV) found in the human genome |
| Human Gene Mutation Database (HGMD) http://www.hgmd.org/ | Database of published gene defects responsible for human inherited diseases |
| NHLBI Exome Variant Server http://evs.gs.washington.edu/EVS/ | Exome sequence data of 6503 individuals and allele frequencies |

Table 3.9. List of electronic databases, computational algorithms, online tools and bioinformatics tools (cont.).

| Name (URL) | Description |
|---|---|
| **Online Tools / Computational Algorithms** | |
| Mutalyzer<br>http://www.lovd.nl/mutalyzer | Checks the best description of a variant according to HGVS |
| Mutation Taster<br>http://www.mutationtaster.org/ | Evaluates and predicts possible impact of a nucleotide change for all types of variants |
| Polymorphism Phenotyping (PolyPhen-2)<br>http://genetics.bwh.harvard.edu/pph | Predicts the effect of a missense mutation on the protein |
| Sorting Intolerant From Tolerant (SIFT)<br>http://blocks.fhcrc.org/sift/SIFT.html | Predicts the effect of a missense mutation on the protein |
| Protein Variation Effect Analyzer (PROVEAN)<br>http://provean.jcvi.org/index.php | Predicts the effect of an amino acid substitution or indels on a protein |
| Human Splicing Finder 2.4.1<br>http://www.umd.be/HSF/ | Identifies potential splice sites and branch points in a given sequence |
| UMD-Predictor<br>http://umd-predictor.eu/ | Predicts pathogenicity of mutations |
| NNSplice<br>http://www.fruitfly.org/seq_tools/splice.html | Analyzes the structure of donor and acceptor sites |
| GenScript<br>https://www.genscript.com/tools/codon-frequency-table | Codon usage frequency table tool |
| **Primer Design** | |
| Primer3<br>http://frodo.wi.mit.edu/primer3 | Used to design primers in a given sequence to amplify the region |
| Oligo Calc (Oligonucleotide Properties Calculator)<br>http://www.basic.northwestern.edu/biotools/oligocalc.html | Calculates melting temperatures of a primer and predicts potential self-annealing sites or hairpin structures |
| UCSC In-Silico PCR<br>http://rohsdb.cmb.usc.edu/GBshape/cgi-bin/hgPcr | Displays the regions on genome to be amplified for given primers |

## 3.7. Bioinformatics Tools

A list of bioinformatics tools used for exome sequence analysis and their URLs and descriptions used in this study is given in Table 3.10.

Table 3.10. Bioinformatics tools and their descriptions.

| Name (URL) | Description |
|---|---|
| **SNP Genotyping Data Analysis** | |
| Illumina GenomeStudio Software http://www.illumina.com/informatics/sequencing-microarray-data-analysis/genomestudio.ilmn | Software that visualizes and analyzes the data generated on Illumina sequencing and array platforms. |
| cnvPartition Plug-in v3.2.0 http://support.illumina.com/downloads/cnvpartition_plug-in_v320_for_genomestudio.ilmn | A software library that works with Illumina GenomeStudio data analysis software. It calculates copy numbers with confidence scores and detects CNV regions. |
| **Exome Sequence Data Analysis** | |
| BWA (Burrows-Wheeler Aligner) http://bio-bwa.sourceforge.net/ | Software for alignment of sequencing reads to the reference genome |
| SAMTools (Sequence Alignment/Map Tools) http://samtools.sourforge.net | Software for manipulating alignments in SAM format. It performs sorting, merging and indexing of alignments, generates per-position information in the pileup format, and performs variant calling. |
| ANNOVAR http://www.openbioinformatics.org/annovar/ | Performs functional annotations of genetic variants detected via next-generation sequencing |
| BEDTools http://code.google.com/p/bedtools/ | A package of tools for comparing genomic features such as computing read coverage in next-generation sequencing data |
| wANNOVAR http://wannovar.wglab.org/ | Online tool to annotate exome sequence data after obtaining vcf format |
| GATK (Genome Analysis Tool Kit) http://www.broadinstitute.org/gatk/ | A package of NGS data analysis tools including coverage analysis, SNP/indel calling and local realignment |
| Picard http://picard.sourceforge.net/ | Software for manipulating SAM (Sequence Alignment/Map) files using Java-based command-line |
| Integrative Genomics Viewer (IGV) https://www.broadinstitute.org/igv/home | Software for visualizing sequence read alignments and read coverage in BAM files |

## 3.8. Equipment

The list of equipment used in this study is given in Table 3.11.

Table 3.11. The list of equipment.

| Equipment | Type |
|---|---|
| Computer | Ubuntu 11.10 Operating System, Gigabyte X58A-UD5 Motherboard, Intel i7 960 (8X) processor, Kingston 1333 Mhz RAM, CPU @ 3.20GHz, 12328MB Memory |
| Centrifuges | MiniSpin Plus (Eppendorf, Germany) |
| | Allegra X-22R (Beckman Coulter, USA) |
| | J2-MC (Beckman Coulter, USA) |
| | Universal 16R (Hettich, Germany) |
| Documentation System | GelDoc Documentation System with Quantity One 4.6.9 Analysis Software (BioRad, USA) |
| Electrophoresis | Horizontal DNA Electrophoresis Gel Box (Bio-Rad, USA) |
| | Primo Minicell Horizontal Gel System (Thermo Scientific, USA) |
| | DCode Universal Mutation Detection System (Bio-Rad, USA) |
| Power Supplies | Fotoforce 250 Electrophoresis Power Supply (Fotodyne, USA) |
| | P250A Power Supply (Sigma-Aldrich, USA) |
| | Power Pac Model 3000 (Bio-Rad, USA) |
| Thermal Cyclers | LightCycler 480 (Roche, Germany) |
| | T100 ThermalCycler (Bio-Rad, USA) |
| Spectrophotometers | NanoDrop 1000 (Thermo Scientific, USA) |
| | Colibri Microvolume Spectrometer (Titertek Berthold, Germany) |

# 4. METHODS

Reference assembly hg19/GRCh37 was used throughout the study**.**

## 4.1. DNA Extraction from Blood Samples

DNA was extracted from peripheral blood samples. Blood samples (max. 4 ml) were collected in blood collection tubes with $K_2$EDTA to prevent coagulation. For each individual one or two tubes of blood were obtained. The blood sample was transferred to a clean falcon tube. After adding three ml of cell lysis buffer for each ml of blood sample to lyse the plasma membrane, the sample was kept at 4$^o$C for 10 minutes. It was centrifuged at 5000 revolution per minute (rpm) at 4$^o$C for 10 minutes. The supernatant was discarded, and to the pellet 10 ml of cell lysis buffer was added. After dissolving the pellet by vortexing, it was centrifuged again at 5000 rpm at 4$^o$C for 10 minutes, and the supernatant was discarded. The pellet was washed with 1-2 ml of cell lysis buffer. To the pellet, 0.3 ml nucleus lysis buffer per ml of initial blood sample was added, and it was dissolved. At this stage, it may be kept at -20$^o$C until the next step. After thawing, 5 µl of Proteinase K (20 mg/ml) and 8 µl of 10% SDS were added per one ml of initial blood volume and incubated at 56°C for three hours or at 37°C overnight to digest nuclear proteins. In order to salt out proteins, 0.9 ml NH4Ac (9.5 M) per ml of nucleus lysis buffer used was added, and the tube was shaken vigorously. The sample was centrifuged at 10000 rpm for 25 min at room temperature to pellet the proteins. The supernatant was transferred into a sterile 50 ml falcon tube, and two volumes of cold absolute ethanol was added to precipitate out the DNA. After inverting tube a few times, DNA was fished out with a sterile micropipette tip and placed into an eppendorf tube. The tube was kept open to let ethanol evaporate, and DNA was dissolved in TE buffer or sterile d$H_2O$ (50-200 µl, according to the amount of DNA). After measuring DNA concentration, DNA was stored at -20$^o$C for further analyses.

## 4.2. Identification of Disease Loci

### 4.2.1. Single Nucleotide Polymorphism (SNP) Genotyping

Whole genome single nucleotide polymorphism (SNP) genotype data for members of all families were generated at Yale Center for Genome Analysis (YCGA, USA) by using Illumina microarrays. Types of microarrays used for each family and their features are given in Table 4.1. The members of study families who have SNP genotype data are either mentioned in the text or shown on the pedigrees in section 3.1. Genotype data were analyzed by using Illumina Genome Studio v.1.02 Genotyping software. Data were copied to Excel, and each genotype (AA, AB and BB) were colored differently.

In some families, different microarrays needed to be used for different members whenever not all samples were genotyped together. In such a situation, common markers were filtered and haplotypes were evaluated.

Table 4.1. Types of microarrays, features and families.

| Name | # SNP Markers | Spacing Mean / Median | Family |
|---|---|---|---|
| HumanOmniExpress-12 BeadChip Kit | App. 710,000 | 4.0 / 2.1 | IID, SCA, IDH, BBS |
| Infinium OmniExpress-24 Kit | 713,599 | 4.08 / 2.22 | IID, IDH, SID, SCA, BBS, PPFEs |

### 4.2.2. Linkage Analysis and Homozygosity Mapping

Linkage analysis was performed for all families to detect the homozygous regions that were identical by descent (IBD) and that possibly harbored the disease gene by using easyLINKAGE v5.08 software, which is an open source graphical user interface. It contains several programs such as PedCheck to detect pedigree and genotyping errors and Allegro, GeneHunter for small (bit size <20), and SimWalk to calculate LOD scores for

large pedigrees. The size of the pedigree is calculated by using the formula 2n-f, where n and f are the numbers of non-founders (individuals whose parents are shown on the pedigree) and founders (whose parents are not shown) in bit as a unit, respectively. In this study, LOD scores were calculated with Allegro or SimWalk in easyLINKAGE v5.08 using the SNP genotype data, a pedigree file and a marker file.

Homozygosity mapping was performed by Homozygosity Comparison in Excel (HCiE) by investigating differently colored genotypes to detect homozygous regions shared by affected individuals and the inspection of genotype sharing among them. Homozygosity Detector program that shows the homozygous regions as blocks on an individual's SNP genotype data in Illumina Genome Studio was also used to confirm homozygous regions in IID and BBS families. Thresholds applied were 30 consecutive markers and minimal size of 200kb.

For all linkage analysis autosomal recessive inheritance, unless otherwise stated, and a disease frequency of 0.0001 were assumed. Haplotypes were constructed using Allegro and visualized via HaploPainter 029.5. Details of the process for each family are presented below.

4.2.2.1. IID Family. Linkage analysis was performed in several steps since some of the members were genotyped later. Initially four affected siblings were genotyped (Figure 3.1), and linkage analysis was performed with the generated data, with 0.07-Mb spacing and 100 marker sets. Regions of shared homozygosity were evaluated in Excel for possible IBD. Genes and phenotypes mapped to those regions were investigated in OMIM Morbid Map to identify the homozygous regions possibly IBD, haplotypes were constructed with Allegro.

A new linkage analysis was performed, this time including also unaffected sister 405, with 0.01 Mb spacing and 30 marker sets. Homozygous regions are listed. Due to an error, linkage analysis could not be performed for the X-chromosome. Genotypes were inspected in Excel, and the regions of shared hemizygosity in affected brothers and heterozygosity in the affected sister were detected. Analysis was repeated assuming that parents were first cousins; initially we had applied linkage analysis assuming that they were second cousins.

Lastly, SNP genotype data were performed for mother and father, and a new linkage analysis was performed.

A linkage analysis to detect regions possibly harboring a candidate modifier gene was performed, assuming two severely affected individuals (402 and 403) as affected and the other two as unaffected.

4.2.2.2. IDH Family. Initially linkage analysis was performed using the genotype data of two affected siblings (322 and 325) with 0.01 spacing and 30 marker sets. Later an affected cousin (304) and two unaffected siblings (312 and 323) were genotyped, and the data were added to the previous files. Linkage analysis was performed using all the genotype data. Parents were assumed as first double-cousins, but consanguinity among the grandparents was ignored, because the capacity of the program was not sufficient. Regions of homozygosity were listed. To not miss any homozygous regions, Homozygosity Detector was applied. The genotypes of affected individuals in those regions were evaluated for possible IBD.

Linkage analysis was performed also for the ID trait using genotype data of two ID patients (311 and 325) and two unaffected siblings. Regions yielding LOD scores >2 were investigated in Excel to delineate the regions of shared homozygosity.

4.2.2.3. SID Family. SNP genotyping was performed for the core family and one affected (401) and one unaffected cousin (402). Linkage analysis was performed several times with different parameters (spacing and marker sets) separately for the two branches of the family. The regions with LOD scores >2.5 were further evaluated in Excel for possible IBD.

Linkage analysis was performed again to make sure that we had not missed any candidate regions. It was performed in several steps (Analysis A, B and C). SNP data of all participants could not be included in linkage analyses A and B due to the limitations of the program, and analysis C took a very long time and thus could not be applied to the whole genome. In linkage analysis A, markers were selected with 0.01-Mb spacing, and sets of 30 markers were used in Allegro. The pedigree used is presented Figure 4.1. Linkage analysis

B was performed for the larger core family for the chromosomes yielding multipoint LOD scores >2.8 in linkage analysis A, i.e. chromosomes 1, 6, 7, 8, 13, 15, 17, 19 and 21. Simplified pedigree is presented Figure 4.2. Chromosomes 1, 8, 13, 17 and 21 were eliminated, because the LOD scores decreased. Regions yielding LOD scores >2.5 (on chromosomes 6, 7, 15 and 19) were investigated on MS Excel for possible identity by descent and for maximal sizes of the shared homozygosity. The region on chromosome 7 and four of the six regions on chromosome 15 were eliminated because the sizes of the shared homozygosity regions were too small (<500 kb). We investigated the genotypes in the remaining homozygous regions (>500 kb). In the two regions on chromosome 15, patients and no others shared the homozygosity possibly identical by descent but not in the regions on chromosomes 6 and 19.

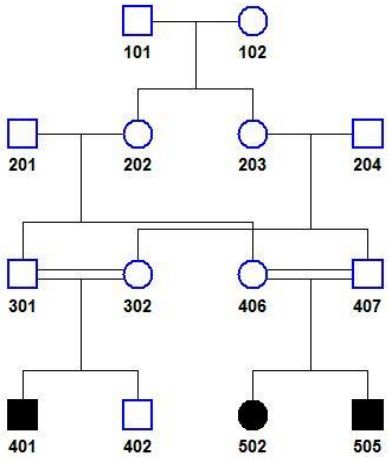



Figure 4.1. Simplified partial pedigree used for linkage analysis A.

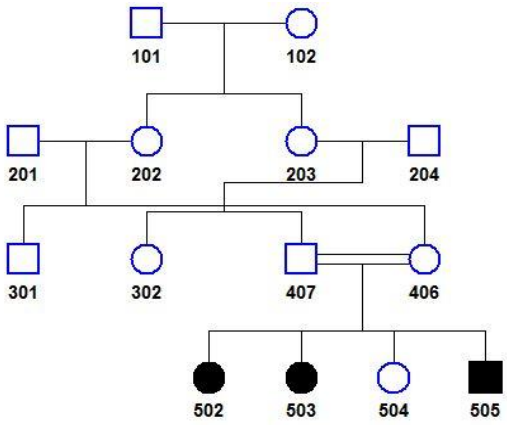



Figure 4.2. Simplified partial pedigree used for linkage analysis B.

Final linkage analysis (analysis C) was performed for the two regions on chromosome 15q with SimWalk including all available SNP data and with 10 marker sets. The simplified pedigree used is presented Figure 4.3.

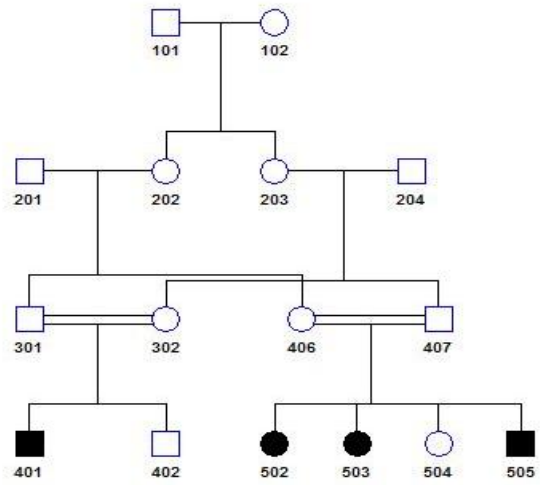



Figure 4.3. Simplified pedigree used for linkage analysis C.

4.2.2.4. SCA Family. Linkage analysis was performed in several steps since some family members were genotyped later and also to not exceed the capacity of the program. Initial linkage analysis was performed using the genotype data of three affected individuals (406, 412 and 413 in Figure 3.4) from two branches of the kindred, with 0.01 spacing and 30 marker sets. The regions with shared homozygosity, which are candidates to harbor the disease gene, were determined.

To confirm that the regions that were found in the first linkage analysis are indeed candidate loci, genotypes aligned in excel were investigated for shared homozygosity in the four patients, and the regions where patients did not share homozygous genotypes were eliminated. In the remaining regions linkage was performed with SimWalk (0.01 spacing, 30 marker sets) for homozygosity regions that yielded high LOD scores (>3), using the genotype data of four affected individuals and the actual pedigree; SNP markers in homozygosity regions plus markers in 1-Mb flanking regions were utilized. Later DNA samples of three more affected and three unaffected relatives were subjected to SNP genotyping. To evaluate whether the final five homozygous regions are really candidates,

another linkage analysis was carried out by including the data for the newly genotyped samples and using program Allegro. A part of the pedigree including the data of the three affected and two unaffected individuals 412 (A), 413 (A), 301 (UA, newly added), 406 (A) and 408 (UA, newly added) were used for linkage analysis to not exceed the capacity of the program. The regions were evaluated in Excel.

4.2.2.5. BBS Family. To investigate whether the disease in the family maps to a known BBS locus, linkage analysis was performed with the initial SNP genotype data of three affected individuals, and no relevant phenotype was found in the candidate regions. Exome sequencing was performed for a patient from another branch of the family, but the patient was later assessed to have a different clinical phenotype. Thus, linkage analysis and exome sequencing were needed to be performed again disregarding that individual. Some other members of the family including the parents in both branches were genotyped and added to the previous haplotype files. Linkage analysis was performed only in branch A (Figure 3.5) with 0.01 spacing and 30 marker sets, and genotypes in the region for the whole family were evaluated in Excel.

Finally, linkage analysis was performed separately for the two branches of the family since the whole pedigree exceeded the capacity of the software (Allegro). Genotype data of all affected individuals and parents but not unaffected siblings were included in order to not disregard regions possibly harboring candidate variants that have reduced penetrance. LOD scores for the two branches were added. For the only IBD region, linkage analysis was performed separately but including all members in the two branches.

Linkage analysis assuming dominant inheritance was performed separately for the two branches of the family. For regions yielding maximal multipoint LOD scores >1.6, haplotypes were constructed. A last linkage analysis was performed to finalize the candidate regions. Of those regions, the ones with LOD scores >3 cumulatively (for branches A plus B) were determined. Fine mapping with no spacing and using 10 marker sets was performed. The regions with LOD scores <3 were eliminated.

4.2.2.6. PPFE Families. PPFE1: Linkage analysis was performed several times since another individual was subjected to SNP genotyping later and the status of one individual was changed from affected to unaffected. Initial linkage analysis was performed assuming the eldest sibling as unaffected since we had been so informed with 0.01 spacing and 30 marker sets. Meanwhile, blood sample of affected cousin was obtained, her DNA was SNP genotyped, and linkage analysis was performed again. Due to the information that in 'unaffected sibling' some lesions had begun to arise, we had to assume her as affected. Additionally, we ascertained that the 'affected cousin' was not affected. Thus, linkage analysis was performed with genotype data of four siblings, assuming all of them affected. We were later informed that the eldest sister (401) that we had been told was unaffected initially seemed to not have lesions. Finally, we considered the results in the initial analysis in which this sister was assumed unaffected.

PPFE2: Multipoint linkage analysis was performed using SNP data of the three affected siblings with 0.01 spacing and sets of 30 markers in Allegro.

## 4.3. Copy Number Variation (CNV) Analysis

In order to detect copy number variations (CNVs) in genotyped individuals, CNV analysis was performed for all families that we had SNP genotype data for via cnvPartition v3.1.6 plug-in on Illumina GenomeStudio v.1.02. In each family any deleted or duplicated regions shared by patients were noted.

## 4.4. Whole Exome Sequence Analysis

Exome enrichment and sequencing was performed commercially at Macrogen Inc. (South Korea) for families IID, SID, IDH, SCA and BBS, and at Yale Center for Genomic Analysis (YCGA, USA) for the second sample from BBS family as well as samples of PPFE families. At Macrogen Inc. samples were subjected to exome capture with an Agilent SureSelect Target Enrichment kit (Agilent, USA) and at YCGA with xGen capture

kit from IDT (USA). Sequencing was performed on Illumina HiSeq 2000 and HiSeq 4000 platforms, respectively.

### 4.4.1. Exome Enrichment and Sequencing

Exome sequencing technique is composed of several steps including DNA fragmentation, library preparation, exome enrichment, cluster generation and sequencing, which require special ingredients and equipment. The technique is described briefly below.

For library preparation 5-10 μg genomic DNA is required. DNA is sheared randomly, and their single strand overhangs are converted to blunt end to obtain double stranded DNA fragments. These 300-400 bp fragments are ligated with adaptors. PCR reactions are performed by using primers specific to adaptors. Denatured DNA is hybridized with biotinylated capture probes of targeted regions (protein coding and flanking sites, i.e., the exome), they were pulled out using streptavidin beads that bind biotinylated probes. After the elution and PCR steps, the exome enrichment part is completed.

The generated DNA strands are hybridized onto the flow cell containing forward and reverse primers and amplified by the bridge amplification technique. The amplified fragments are sequenced by using fluorescently labeled nucleotides on a sequencer.

### 4.4.2. Analysis of Exome Sequence Results

Bioinformatics analysis results of exome sequencing that were performed at Macrogen Inc. were provided as a list of annotated variants in Excel format obtained by using standard parameters. However, I also performed bioinformatics analyses starting from the raw data (fastq. files) to be able to compare results and to not miss any variants that might have escaped detection due to application of a single method only. In addition to that we needed to perform additional analyses such as computation of the coverage of exons or visualizing aligned sequence reads using Integrated Genome Viewer (IGV). All bioinformatics analyses were performed on the Linux operating system.

Briefly, raw data (in "fastq" files) were analyzed using software Burrows-Wheeler Aligner (BWA), Sequence Alignment/Map Tools (SAMTools) and ANNOVAR. Paired-end reads in fastq file were aligned to reference sequence (Genome Reference Consortium Human Build 37 - GRCh37/hg19) downloaded from UCSC Genome Bioinformatics site by using BWA, and the final alignment file was generated in SAM (Sequence Alignment/Map) format. Files in SAM format were converted to BAM (Binary Alignment/Map) format which is also used in several programs as an input file by using SAMTools. After sorting and indexing the BAM files, variant calling was performed to list nucleotides differing from the reference genome sequence, including indels by SAMTools. Variants were annotated by using ANNOVAR program, and the list of annotated variants including for each variant the chromosomal location, base change, depth, frequencies obtained from 1000 Genomes, dbSNP ID (if any), and region (exonic, intronic, splicing, UTR, non-coding RNA or intergenic). In addition, for exonic variants type of the change (synonymous, nonsynonymous/missense, frameshift, stopgain/nonsense or stoploss) were investigated in Excel.

The reads aligned to the reference sequence were visualized by Integrative Genome Viewer (IGV) software using sorted.bam files. BEDTools software was used to detect the coverage of each exon in targeted regions. Coverage data were compared between two exome data files that were generated in the same batch to determine whether there are any deleted or duplicated exons.

Command lines used in the bioinformatics analyses are given in Table 4.2.

.

Table 4.2. Command lines used in the bioinformatics analysis of exome sequencing data and their functions.

| Command Line | Function |
|---|---|
| cat chrfile.fa > all_chr.fasta | Concatenates all chromosomes of the reference genome sequence |
| cat chrfile.fa > all_chr.fasta | Concatenates all chromosomes of the reference genome sequence |
| bwa index -a bwtsw all_chr.fasta | Indexes the reference genome |
| bwa aln -n 0.01 -t 8 all_chr.fasta file1.fastq > file1.sai | Aligns paired-end sequence reads to the reference genome |
| bwa sampe all_chr.fasta file1.sai file2.sai file1.fastq file2fastq > file.sam | Generates file in SAM format using paired-end read files |
| samtools view -bS file.sam > file.bam | Converts SAM to BAM format |
| samtools sort file.bam | Sorts the BAM file |
| samtools index file.sorted.bam | Indexes the sorted BAM file |
| samtools pileup -vcf all_chr.fasta file.sorted.bam > file.pileup.txt | Calls variants |
| java -jar ReorderSam.jar I=file.bam O=file_karyo.bam R=ucsc.hg19.fasta | Creates a karyotypically ordered BAM file using Picard-Tools software |
| Java –jar GenomeAnalysisTK.jar –T RealignerTargetCreator –R ucsc.hg19.fasta –I file_karyo.sorted.bam –o file_realign.intervals | Determines the intervals to be realigned |
| java -jar -I file_karyo.sorted.bam -R ucsc.hg19.fasta -T IndelRealigner -targetintervals file_realign.intervals -o file_realigned.bam | Makes local realignment |
| java -jar AddOrReplaceReadGroups.jar I=file_realigned.bam O=file_realigned.fixed.bam SORT_ORDER=coordinate RGID=file RGLB=file RGPL=illumine RGPU=file RGSM=file CREATE INDEX=True VALIDATION STRINGENCY=LENIENT | Converts the re-aligned.bam file to an input file compatible to UnifiedGenotyper, a variant calling tool in GATK, using Picard-Tools. |
| java -jar GenomeAnalysisTK.jar -R ucsc.hg19.fasta -T UnifiedGenotyper -I file_realigned.fixed.bam -o file_GATK.vcf | Calls variants using GATK |

Table 4.2. Command lines used in the bioinformatics analysis of exome sequencing data and their functions (cont.).

| Command Line | Function |
|---|---|
| convert2annovar.pl file.pileup.txt -outfile file_annovar.pileup.txt | Converts the SAMTools output file to a file compatible to ANNOVAR |
| convert2annovar.pl file_GATK.vcf -format vcf4 -outfile fileGATK_annovar.pileup.txt | Converts the GATK output file to a file compatible to ANNOVAR |
| annotate_variation.pl -buildver hg19 file_annovar.pileup.txt humandb | Region based annotation |
| annotate_variation.pl -filter -dbtype snp131 file_annovar.pileup.txt | Filter based annotation |
| bamToBed -i file.bam > file.bed coverageBed -a file.bed -b targetedRegions.bed > file_coverage.txt | Computes the coverage for each exon in the target regions of the exome chip |

The results of exome sequencing data for one patient each from BBS family and PPFE families were performed in YCGA and analyzed following the newly established GATK pipeline in our laboratory, presented in Table 4.3.

Table 4.3. Command lines of newly established GATK pipeline used in bioinformatics analysis of exome sequencing data and their functions. Name of the file is given as "File".

| Command Line | Function |
|---|---|
| bwa mem -M -t 8 ucsc.hg19.fasta File_1.fastq File_2.fastq > File.sam | Aligns paired-end reads to the reference genome |
| samtools fixmate -O bam File.sam File.bam | Cleans up unusual FLAGs such as read pairing information in the sam records created by BWA |
| samtools sort -O bam -o File_sorted.bam File.bam | Sorts the bam file with chromosomal coordinates |
| java -jar picard.jar AddOrReplaceReadGroups I=File_sorted.bam O=File_sorted.RG.bam RGID=1 RGLB=lib RGPL=illumina RGPU=index RGSM=File_sorted.RG.bam | Assigns read groups such as library, platform, sample names in a bam file |
| java -jar picard.jar BuildBamIndex I=File_sorted.RG.bam O=File_sorted.RG.bam.bai | Generates bam index file that allows fast look up of bam file |

Table 4.3. Command lines of newly established GATK pipeline used in bioinformatics analysis of exome sequencing data and their functions. Name of the file is given as "File" (cont.).

| Command Line | Function |
|---|---|
| java -Xmx2g -jar GenomeAnalysisTK.jar -nt 8 -T RealignerTargetCreator -R ucsc.hg19.fasta -I File_sorted.RG.bam -o File_sorted.RG.bam.intervals -known 1000G_phase1.indels.hg19.vcf -known Mills_and_1000G_gold_standard.indels.hg19.vcf | Determines intervals needed for local realignment |
| java -Xmx32g -jar GenomeAnalysisTK.jar -T IndelRealigner -R ucsc.hg19.fasta -I File_sorted.RG.bam -targetIntervals File_sorted.RG.bam.intervals --consensusDeterminationModel USE_READS -known 1000G_phase1.indels.hg19.vcf -known Mills_and_1000G_gold_standard.indels.hg19.vcf -LOD 0.4 -o File_sorted.realigned.bam | Performs the realignment around previously defined intervals |
| java -jar /home/aslitolun/Downloads/picard.jar BuildBamIndex I=File_sorted.realigned.bam O=File_sorted.realigned.bam.bai | Generates bam index file that allows fast look-up of bam file |
| java -Xmx4g -jar GenomeAnalysisTK.jar -nct 8 -T BaseRecalibrator -R ucsc.hg19.fasta -knownSites dbsnp_138.hg19.vcf -knownSites 1000G_phase1.indels.hg19.vcf -I File_sorted.realigned.bam -o File_recal_data.table | Detects systematic errors in base quality scores |
| java -Xmx4g -jar GenomeAnalysisTK.jar -nct 8 -T PrintReads -R ucsc.hg19.fasta -I File_sorted.realigned.bam -BQSR File_recal_data.table -o File_sorted.realigned.recal.bam | Compensates previously generated covariates by adjusting quality scores |
| java -Xmx4g -jar picard.jar MarkDuplicates REMOVE_DUPLICATES=TRUE INPUT=File_sorted.realigned.recal.bam OUTPUT=File_sorted.realigned.recal.rmdup.bam METRICS_FILE=File_marked_dup_metrics.txt | Marks and removes PCR duplicates |
| java -jar picard.jar BuildBamIndex I=File_sorted.realigned.recal.rmdup.bam O=File_sorted.realigned.recal.rmdup.bam.bai | Generates bam index file that allows fast look up of bam file |
| java -jar GenomeAnalysisTK.jar -T HaplotypeCaller --genotyping_mode DISCOVERY -R ucsc.hg19.fasta -I File_sorted.realigned.recal.rmdup.bam --dbsnp dbsnp_138.hg19.vcf -stand_call_conf 30 -o File_raw_variants.vcf | Detects germline SNPs and InDels |

## 4.5. Candidate Genes and Mutation Screening

Exome sequence data were filtered and evaluated according to the criteria explained in Section 1.8.4 to detect candidate variants that might possibly underlie the disease in the family. Since errors can occur in exome sequence analysis such as inaccuracies during sequencing the reads, alignment of the reads to the reference genome and variant calling, candidate variants need to be validated by Sanger sequencing after amplification with PCR. After designing primers that can amplify the site of a candidate variant, PCR was performed. PCR products were analyzed by Sanger sequencing to validate the variant. Family members were later screened to assess segregation of the trait.

### 4.5.1. Designing Primers

A pair of primers was designed to amplify a region of approximately 180-260 bp encompassing the site of the variant. The fasta sequence of the region was obtained from NCBI MapViewer. Primer3 software was used to select the best primer pair among those sequences. Potential hairpin structures and annealing of forward and reverse primers were checked using OligoCalc tool. Whether the primers amplify only the desired region was checked via in silico PCR at UCSC database.

### 4.5.2. Polymerase Chain Reaction (PCR) Amplifications

For a PCR reaction, 1X PCR buffer, 400 nM of each primer, 0.2 μM of each dNTP, 30-70 ng of genomic DNA, 0.2 U Taq DNA polymerase and sufficient distilled water were used in a total volume of 25 μl. PCR protocol was as follows: An initial denaturation step at 95°C for 3 min, then 35 cycles of denaturation for 30 sec at 95°C, annealing at an optimal temperature between 64 and 51°C for 30-40 sec, elongation for 30 sec to 1 min at 72°C, and a final extension step for 6 min at 72°C. The primer sequences to amplify the sites of variants and annealing temperatures are given in Table 4.4.

Table 4.4. PCR conditions to amplify the sites of candidate variants.

| Gene, variant | Primer sequences (5' → 3') | Product Size (bp) | Annealing Temperature ($^{o}$C) |
|---|---|---|---|
| **IID Family** | | | |
| *PTRHD1* (NM_001013663) c.155G>A | F: ACCTCGAGGACCACTTTGC<br>R:GGTCCTGGTACAATACTTGGTG<br>R2: ATGCACCGGGGAGTAGGT | 186/254 | 63 → 57 |
| *PTRHD1*_exp (NM_001013663) F: Exon1 R: Exon 1-2 | F:AAAGGATCTATCACAAGCTCCG<br>R:ATCTGGGGCCTCGAGGAC | 166 | 63 → 57 |
| *AGBL5* (NM_021831) c.2619delT | F: GCCCCTGCCTTTTCTCCTAT<br>R: CCCACTGTTACCCCATCTTG | 192 | 63 → 56 |
| *MAP3K15* (NM_001001671) c.776A>G | F: TGGATATCACGGTAGGACAGG<br>R: GTGCGTGCAGATGTGACTG | 214 | 63 → 56 |
| **IDH Family** | | | |
| *TPO* (NM_001206744) c.719A>G | F:GACATGTCATTCAAGTTTCAAATG<br>R:CATGGGTTTTGGTTCTCACA | 178 | 59.3 |
| **SID Family** | | | |
| *PDIA3* (NM_005315) c.170G>A | F: GGAAGTGTCTACTAGCTCAAAGG<br>R: CCATGTAACAAAGCTGAGACAAC | 272 | 62 → 57 |
| **SCA Family** | | | |
| *SPTBN2* (NM_006946) c.6375-1 G>C | F: GGTGGAGGGGCTACGACT<br>R: CCAGACAGCTTCTGACACCA | 245 | 63 → 57 |
| *SPTBN2* (NM_006946) c.6375-1 G>C (shorter) | F: GCACAGAGGGACAGTGGG<br>R: ACGGGTGAGAGCCAGGAT | 192 | 63 → 57 |
| *SPTBN2*_exp (NM_006946) F: exon 29-30 R: exon 34 | F: ACTGGCTTCAGCTGGTTTTG<br>R: CTGACTCGGTAGACCTGCT | 588 | 64 → 56 |
| *SPTBN2*_exp (NM_006946) F: exon 28-29 R: exon 34-35 | F: GGCCGAGGAGATCTCAGAGA<br>R: CAGGACCTGTTGGCAGCC | 789 | 64 → 56 |

Table 4.4. PCR conditions to amplify the sites of candidate variants (cont.).

| Gene, variant | Primer sequences (5' → 3') | Product Size (bp) | Annealing Temperature (°C) |
|---|---|---|---|
| **BBS Family** | | | |
| *CEP19* (NM_032898) c.194_195insA | F: TCCCGTTGAATTTGTTCCATTG R: TGTTCTGTTTTGGCTGTTGCAG | 243 | 63 → 57 |
| *GLI1* (NM_005269) c.820G>C | F: GGAGTCGTGGAAGAGGAACA R: GTGCACTTGTGTGGCTTCTC | 246 | 63 → 55 |
| *CCDC28B* (NM_024296) c.330C>T | F: CCTTCCTGACCGAGGTGACT R: TCCCTGGAACGAAGCCC | 236 | 61 → 54 |
| *MKKS* (NM_018848) c. 1015A>G | F: AGCTGCAGATTGTTGCTTCA R: TGGGGCTTTTATGTTGGCTA | 230 | 63 → 55 |
| *C8ORF37* (NM_177965) c.533C>T | F: TCTGTCCGAATGTGCAGTCT R: CAACATGCCAGAATTTCACAA | 177 | 63 → 55 |
| *TMEM67* (NM_153704) c.2397T>C | F: GCAGATGAGTTGCTATTTGCTTC R:CTTCTCAACTTAAAAACAAAAAGATG | 202 | 61 → 54 |
| *NPHP4* (NM_015102) c.1852G>A | F: CTGTAGCACCATCTCGTTGCT R: TCAGCTCTGCAGGGCAG | 179 | 61 → 54 |
| **PPFE1 Family** | | | |
| *FAM35A* (NM_019054) c.540_541insCC | F: CGCATTTCACCGAAGAAGA R: TTCATGATATTCTGTTGGCACA | 227 | 52 |
| *TNKS2* (NM_025235) c.1146A>T | F: CTTGATTGTTCTTGATTGGAAAC R: TCCAAAAGAGTAAAATCTCAATGAT | 228 | 52 |
| *CC2D2B* (NM_001159747) c.903T>G | F: GAGCGAGACTTCGTCTCAAAA R: GCTTCCTATGCCAAATTCCA | 229 | 59.8 |

## 4.6. Analysis of PCR products

To assess amplification efficiency and whether unspecific products were present, PCR products were resolved by electrophoresis on 2% agarose gels containing 10 mg/ml ethidium bromide. Prior to loading, 5 µl of the product was mixed with 1µl of 6X loading dye. Either a set of pUC19 restriction enzyme fragments or a 1-kb marker ladder was loaded as a marker to determine the size of the PCR product. Electrophoresis was

performed in 0.5X TBE buffer at 120 volts for 15 minutes, and the fragments were visualized under ultraviolet light using BIORAD GelDoc Documentation system and Quantity One 4.6.9 software.

### 4.6.1. Sequencing of PCR products

PCR products that supposedly amplified the target regions were sequenced at Macrogen Inc. (Amsterdam) or MedSanTek (Turkey). Firstly PCR products were purified from primers, nucleotides, salts and polymerase enzyme and then sequenced using fluorescently labeled dNTPs. The products are resolved by capillary gel electrophoresis and excited labeled DNA fragments are detected by light sensor. The results were provided as "ab1" files to visualize electrophoretograms with Chromas Lite 2.01 software and in "fasta" format to facilitate comparison to the human genome using UCSC BLAT. The exact region of amplification and any differences from the reference sequence could be determined.

### 4.6.2. High Resolution Melting Curve Analysis

High resolution melting (HRM) curve analysis was attempted to screen members of IID and BBS family for only one candidate variant, but it could not be optimized.

### 4.6.3. Single Strand Conformational Polymorphism Analysis

Single Strand Conformational Polymorphism (SSCP) analysis is a method used to determine the segregation of the candidate variants in all study families and also for population screening for the causative variant in IID family. This method enables us to detect even a single nucleotide difference in single DNA strands since it leads to a different folding conformation and thus a different migration rate on denaturing polyacrylamide gels. Migration patterns for homozygous wildtype, homozygous mutant and heterozygous strands for a variant are generally different. Even the two complementary strands generally migrate differently.

In the present study, SSCP analysis was used to screen family members for the candidate variants identified/detected in *PTRHD1* in IID family, in *TPO* in IDH family in *PDAI3* in SID Family, in *SPTBN2* in SCA Family, in *CEP19*, *GLI1*, *CCDC28B*, *MKKS*, *TMEM67*, *C8ORF37* and *NPHP4* in BBS family, and in *TNKS2* and *FAM35A* in PPFE1 family to investigate segregation. In addition, 200 unrelated individuals from Pakistani population were screened for the *SPTBN2* variant by SSCP. Sequences of the primers used for amplification, PCR conditions and SSCP conditions are summarized in Table 4.5.

Table 4.5. PCR and SSCP conditions for variant testing.

| **Primer Name** | **Primer Sequence (5' → 3')** | **Product Size (bp)** | **SSCP Conditions** |
|---|---|---|---|
| *SPTBN2* c.6375-1 G>C | F: GGAGGCCGAGCACTCG<br>R: GCAGCCCGGGAGCAG | 245 | Gels: Gly(+), 4°C, 4W, 20 h |
| *PTRHD1* c.155G>A | F: ACCTCGAGGACCACTTTGC<br>R2: ATGCACCGGGGAGTAGGT | 254 | Gels: Gly(-), 4°C, 5W, 16 h |
| *TPO* c.719A>G | F:GACATGTCATTCAAGTTTCAAATG<br>R:CATGGGTTTTGGTTCTCACA | 178 | Gels: Gly(+), 4°C, 4W, 20 h |
| *CEP19* c.194_195insA | F: TCCCGTTGAATTTGTTCCATTG<br>R: TGTTCTGTTTTGGCTGTTGCAG | 243 | Gels: Gly(-), RT, 10W, 5 h |
| *GLI1* c.820G>C | F: GGAGTCGTGGAAGAGGAACA<br>R: GTGCACTTGTGTGGCTTCTC | 246 | Gels: Gly(-), 4°C, 4W, 20 h |
| *CCDC28B* c.330C>T | F: CCTTCCTGACCGAGGTGACT<br>R: TCCCTGGAACGAAGCCC | 236 | Gels: Gly(-), 4°C, 4W, 20 h |
| *MKKS* c. 1015A>G | F: AGCTGCAGATTGTTGCTTCA<br>R: TGGGGCTTTTATGTTGGCTA | 230 | Gels: Gly(-), 4°C, 4W, 18 h |
| *C8ORF37* c.533C>T | F: TCTGTCCGAATGTGCAGTCT<br>R: CAACATGCCAGAATTTCACAA | 177 | Gels: Gly(-), RT, 4W, 16 h |
| *TMEM67* c.2397T>C | F: GCAGATGAGTTGCTATTTGCTTC<br>R:CTTCTCAACTTAAAAACAAAAAGATG | 202 | Gels: Gly(+), 4°C, 4W, 20 h |
| *NPHP4* c.1852G>A | F: CTGTAGCACCATCTCGTTGCT<br>R: TCAGCTCTGCAGGGCAG | 179 | Gels: Gly(-), RT, 4W, 16 h |
| *PDIA3* c.170G>A | F: GGAAGTGTCTACTAGCTCAAAGG<br>R: CCATGTAACAAAGCTGAGACAAC | 272 | Gels: Gly(+), 4°C, 10W, 5 h |

Gly, Glycerol; (+), with; (-), without; RT, room temperature (15-24 °C); W, watt; h, hour

Table 4.5. PCR and SSCP conditions for variant testing (cont.).

| Primer Name | Primer Sequence (5' → 3') | Product Size (bp) | SSCP Conditions |
|---|---|---|---|
| *FAM35A* c.540_541insCC | F: CGCATTTCACCGAAGAAGA<br>R: TTCATGATATTCTGTTGGCACA | 227 | Gels: Gly(-), RT, 4W, 20 h |
| *TNKS2* c.1146A>T | F: CTTGATTGTTCTTGATTGGAAAC<br>R: TCCAAAAGAGTAAAATCTCAATGAT | 228 | Gels: Gly(-), RT, 4W, 20 h |
| *CC2D2B* c.903T>G | F: GAGCGAGACTTCGTCTCAAAA<br>R: GCTTCCTATGCCAAATTCCA | 229 | Gels: Gly(-), RT, 4W, 20 h |

Gly, Glycerol; (+), with; (-), without; RT, room temperature (15-24 $^{o}$C); W, watt; h, hour

## 4.7. Assessment of Expression Levels of *PTRHD1* by Quantitative PCR (QPCR)

A real-time PCR was performed to investigate expression levels of the candidate gene *PTRHD1* in IID family, using an intron-spanning primer. Relative quantification of *PTRHD1* transcripts was performed in Light Cycler 480 (Roche, Germany) using cDNA generated from commercial total RNA extracted from 17 different tissues: frontal cortex, parietal cortex, occipital cortex, corpus callosum, cerebellum, brain stem, substantia nigra, pons, putamen, spinal cord, liver, blood, skeletal muscle, adipose, bone marrow and testis. For real time PCR reactions, a forward primer specific to exon 1 and a reverse primer specific to the junction of exon 1 and exon 2 were used. Housekeeping gene *BETA TUBULIN* (*TUBB*) was the reference gene, and a forward primer specific to exon 1 and a reverse primer specific to exon 3 were used. Sequences of the primers are presented in Table 4.6. The reaction mix was prepared using 10 µl Light Cycler 480 SYBR Green Master mix, 200 nM of each primer pair, 60 ng cDNA and distilled water up to a total volume of 20 µl. All reactions were performed in triplicate. Reaction conditions are as follows: 10 minutes initial denaturation at 95°C and 40 to 45 cycles of denaturation for 10 seconds at 95°C, 15 seconds annealing at the appropriate annealing temperature (65°C to 57°C), and elongation for 9-11 seconds at 72°C. Transcript levels were normalized to *TUBB* via advanced relative quantification analysis on the LightCycler 480 Relative Quantification Software (Roche), using relative standard curves of serial dilutions of

cDNAs to calculate PCR efficiencies. Cycle threshold ($C_T$) values and $2^{-\Delta\Delta C_T}$ method (Livak *et al*., 2001) calculations were performed using Light Cycler 480 Relative Quantification software, and the graphs showing ± standard errors of the mean (SEM) were generated by using Excel.

Table 4.6. PCR conditions for the relative quantification assay.

| **Gene, variant** | **Primer sequences (5' → 3')** | **cDNA Product Size (bp)** | **Annealing Temperature** |
|---|---|---|---|
| *PTRHD1*_exp<br>F: Exon1<br>R: Exon 1-2 | F:AAAGGATCTATCACAAGCTCCG<br>R:ATCTGGGGCCTCGAGGAC | 166 | 63 → 57 |
| *TUBB_exp*<br>F: Exon1<br>R: Exon3 | F: ACATCCAGGCTGGTCAGTG<br>R: AAAGGACCTGAGCGAACAGA | 226 | 63 → 57 |

# 5. RESULTS

## 5.1. Isolated Intellectual Disability (IID) Family

### 5.1.1. Linkage Analysis

Linkage analysis was performed in several steps since some members were subjected to SNP genotyping later than the first four samples. Initial linkage analysis was performed with the generated data of the four affected siblings (three brothers and a sister), with 0.07 Mb spacing and 100 marker sets. It yielded only two regions with maximal LOD scores >2.0. Homozygous regions were evaluated for possible IBD, and the two candidate regions detected are presented in Table 5.1. Those regions were investigated in OMIM Morbid Map, and no relevant phenotype was found. To detect the homozygous regions possibly IBD, haplotypes were constructed by Allegro.

Table 5.1. Regions of shared homozygosity assessed as IBD for patients of IID Family, detected by initial linkage analysis.

| Chr | Maximal homozygosity | | Flanking SNPs (start-end) | | LOD score | Size (bp) |
|---|---|---|---|---|---|---|
| X | 13,562,317 | 26,140,796 | rs5935601 | rs5986310 | 2.36 | 12,578,479 |
| 2 | 23,577,934 | 30,557,749 | rs2723137 | rs7604316 | 3.58 | 6,979,815 |

Linkage analysis was performed again including the genotype of unaffected sister, and using markers with 0.01 Mb spacing and sets of 30. Several regions were obtained with maximal LOD scores >3. All those regions were evaluated in Excel and those not IBD were eliminated. Possibly IBD regions are listed in Table 5.2 and Table 5.3. One reason for obtaining results different from the previous analysis is that in the region on chromosome 2, unaffected sister (included later) is homozygous after position 25,487,658. The reason for the region on chromosome 2: 21694451-22302013 being in Table 5.2 but not in Table 5.1 that the region escaped detection due to its small size on investigation in Excel. However, as in the final linkage analysis that included the SNP genotype data of parents did not yield LOD scores >2 in the region, the region was eliminated.

Due to an error, linkage analysis could not be performed for the X-chromosome. Instead, haplotypes were inspected in Excel, and regions of shared hemizygosity in affected males plus heterozygosity in the affected sister are detected (Table 5.3).

Table 5.2. Homozygous regions detected in the second linkage analysis.

| Chr | Maximal homozygosity | | Flanking SNP (start-end) | | Size (bp) | LOD score |
|---|---|---|---|---|---|---|
| 2 | 21694451 | 22302013 | rs11897825 | rs328539 | 607,562 | 3.7 |
| 2 | 23577934 | 25487658 | rs2723137 | rs6546045 | 1,909,724 | 3.7 |

Table 5.3. Regions with shared hemizygosity in the X-chromosome in affected brothers. Affected sister could be carrying the candidate disease haplotype in those regions.

| Maximal homozygosity | | Flanking SNPs (start-end) | | Size (bp) | Sisters |
|---|---|---|---|---|---|
| 3582703 | 11296936 | rs11152537 | rs765480 | 7714233 | Same heterozygosity |
| 13562317 | 26140796 | rs5935601 | rs5986310 | 12578479 | All sibs are homozygous |
| 26140796 | 29015651 | rs5986310 | rs4308905 | 2874855 | Same heterozygosity |
| 119399097 | 139943799 | rs11260237 | rs6634206 | 20544702 | Different heterozygosity |

Linkage analysis was performed again because we learned that parents were first cousins; we had applied linkage analysis assuming that they were second cousins. Mother and father were subjected to SNP genotyping, and a new linkage analysis yielded two regions with LOD score>2 and a maximal multipoint LOD score of 3.25 for only one region (Figure 5.1). The reason for a lower LOD score for the region on chromosome 2 compared to first linkage analysis is the closer consanguinity in the parents.

The disease locus was reevaluated to determine the exact region due to the discrepancy between the constructed haplotype and the results of the ocular examination of the SNP genotype data. We determined that the region is the gene locus as was determined already, and it was delineated by 23,577,934 bp (rs2723137) and 25,487,658 bp (rs6546045) and 1,909,724-bp long at 2p24.1-p23.3.

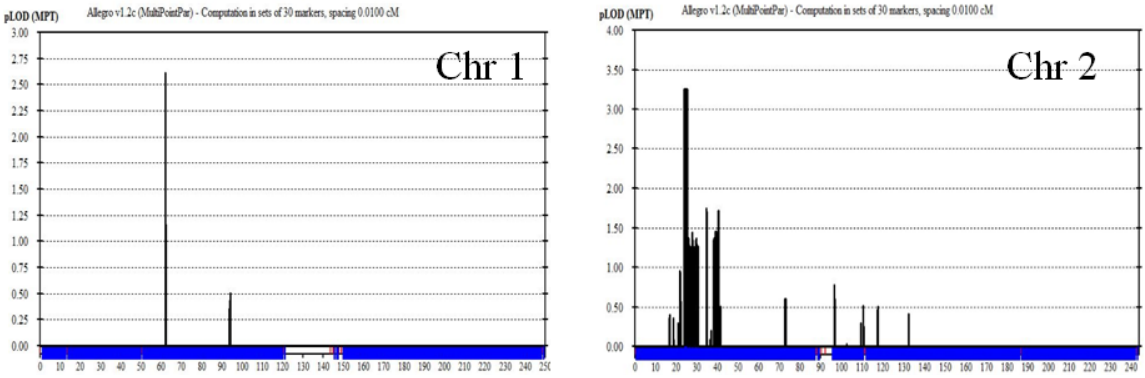



Figure 5.1. Multipoint LOD score graphics for chromosomes with LOD score > 2 in the final linkage analysis.

### 5.1.2. Deletion-Duplication Analysis

Deletion-duplication analysis was performed for all chromosomes and participants, and no duplication or deletion common to all patients was found. A shared homozygosity region in four affected siblings was detected by cnvPartition only at 2p23 (Figure 5.2). The genotypes of all affected individuals in the region homozygous in 402 (telomeric to the region of shared homozygosity) is also possibly IBD, as all affected sibs have the same genotype on the excel sheet.

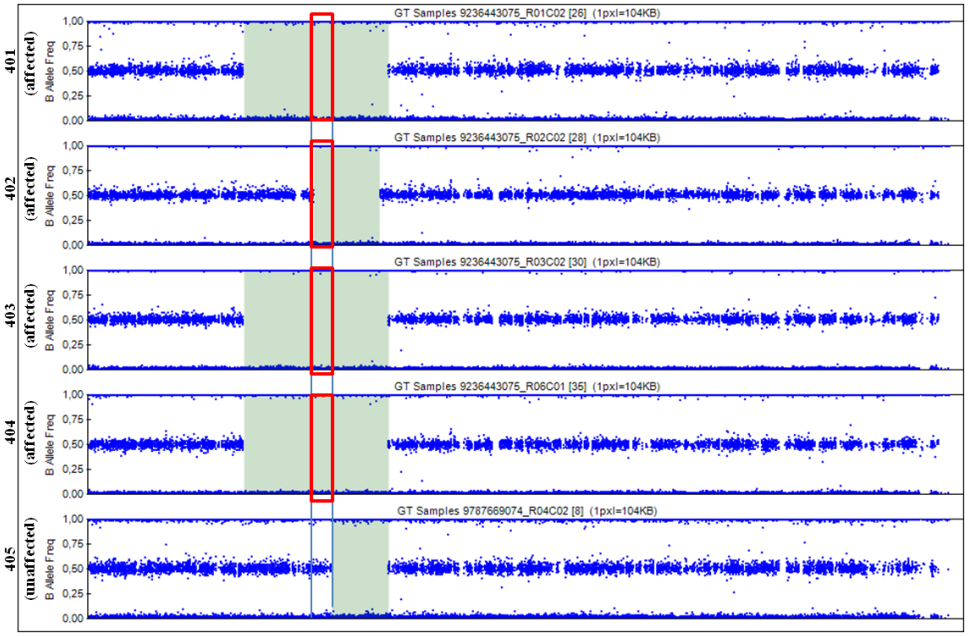



Figure 5.2. CNV partition results on chromosome 2 showing the homozygosity region (23.5-25.4 Mb) shared by affected siblings but not unaffected sibling.

### 5.1.3. Exome Sequence Analysis

DNA sample of a patient (402) was subjected to exome sequencing. Variants with allele frequencies >0.01 and ratios of the number of reads with the variant to the number of reads with the reference sequence <60% were filtered out. Candidate variants are presented in Table 5.4.

We evaluated the *PTRHD1* variant on the Genome Browser and assessed that it is a good candidate. The mutation was predicted as damaging by bioinformatics tools. The highest frequency of the variant in ExAC is in South Asian sample as 0.00006 and world-wide 0.000008. The other rare exonic variant at the candidate locus was *ITSN2* c.2245G>A with highest frequency in ExAC as 0.00006 in South Asian samples (gnomAD highest in SA 0.00003), but it was predicted as benign by online tools. No interaction was found between PTRHD1 and ITSN2 in GeneMania and STRING databases, indicating that the variant in *ITSN2* is most probably not a modifier.

### 5.1.4. Validation of the Variant

The *PTRHD1* variant was validated by Sanger sequencing (Figure 5.3). Since HRM analysis could not be optimized, SSCP analysis was applied. As a result, all patients had the same pattern, indicative of shared homozygosity for the variant. Mother (301), father (302) and unaffected sister (405) were all heterozygous. We screened in total 205 Pakistani samples since ExAC was not public yet, and the pattern for one of them seemed heterozygous. Sanger sequencing revealed heterozygosity for c.127C>A (rs545737721), a known frequent SNP (frequency 0.01 in ExAC SA).

Table 5.4**.** Candidate variants in the exome file of patient 402 in the identified gene region 23,577,934 bp - 25,487,658 bp at 2p24.1-p23.3.

| Start | Ref | Alt | Depth | | Gene | Freq* | dbSNP | Region | Change in Protein | Prediction Algorithms | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Tot | Alt | | | | | | Mutation Taster | SIFT | PolyPhen | Splice Finder |
| 24306117 | - | T | 16 | 16 | *TP53I3* | - | - | Intronic | - | Polym | NA | NA | No significant splicing motif alteration |
| 24307230 | A | G | 21 | 17 | *TP53I3* | 0.0011(SA) | rs376092956 | UTR5 | - | Polym | NA | NA | Creation of an intronic ESE site. Probably no impact on splicing. |
| 24307240 | G | A | 15 | 14 | *TP53I3* | - | - | UTR5 | - | Polym | NA | NA | Creation of an intronic ESE site. Probably no impact on splicing. |
| 24494647 | C | T | 24 | 24 | *ITSN2* | 0.00006 (SA) | rs751750609 | Exonic | c.2245G>A p.Glu749Lys | Polym | Tolerated | Benign | Alteration of an exonic ESE site. Potential alteration of splicing. |
| 25016092 | C | T | 110 | 110 | *PTRHD1* | 0.00006 (SA) | - | Exonic | c.155G>A p.Cys52Tyr | Disease causing | Damaging | Probably damaging | No significant splicing motif alteration |
| 25498465 | G | C | 76 | 76 | *DNMT3A* | - | - | Intronic | - | Disease Causing | NA | NA | Creation of an intronic ESE site. Probably no impact on splicing. |

*Highest frequency in ExAC or gnomAD SA, South Asian; Polym, polymorphism; NA, not applicable.

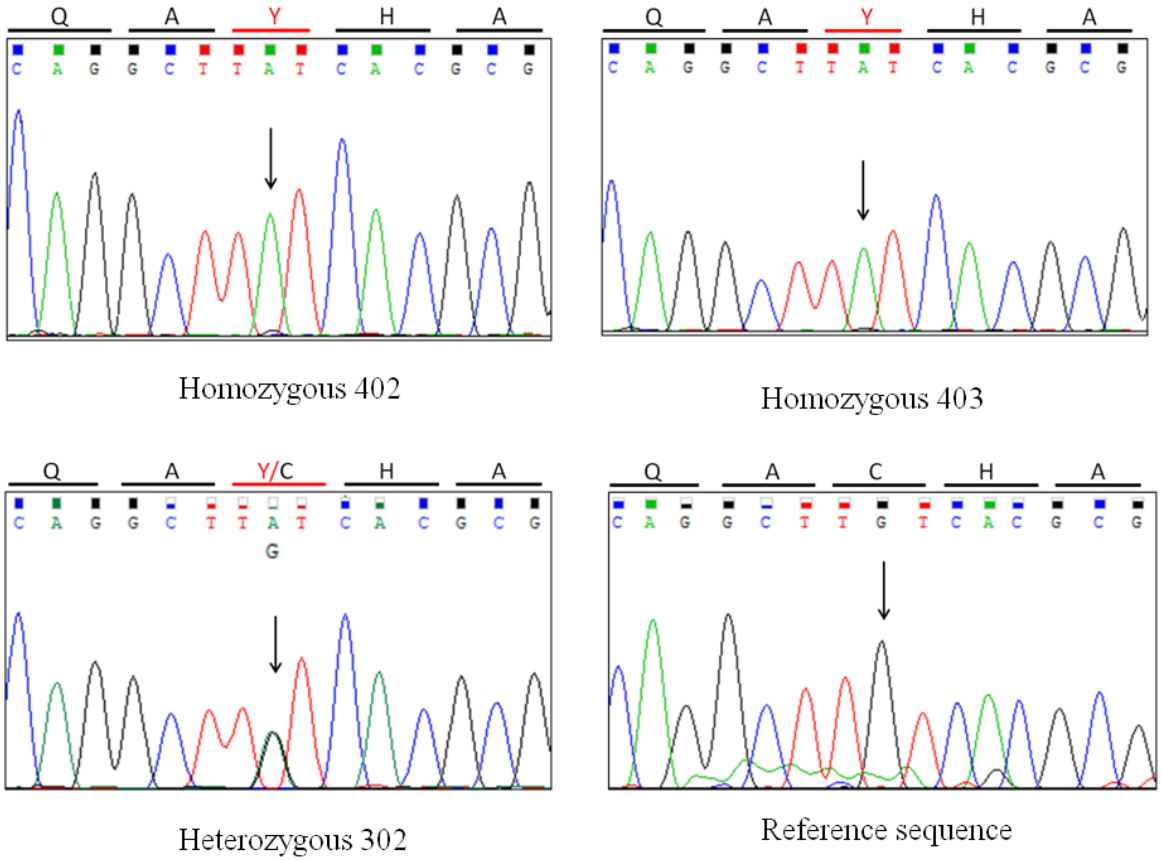



Figure 5.3. Electrophoretograms of *PTRHD1* c.155G>A (p.Cys52Tyr) in affected brothers 402 (exome sequenced) and 403, father 302 and a control sample.

### 5.1.5. Relative Expression of *PTRHD1* in Various Tissues

The gene is reported with wide tissue expression in Expression Atlas. We investigated its expression in adult brain and found expression in all regions assayed (Figure 5.4). Relative quantification of *PTRHD1* transcripts was performed using cDNA generated from commercial total RNA extracted from 17 different tissues: frontal cortex, parietal cortex, occipital cortex, corpus callosum, cerebellum, brain stem, substantia nigra, pons, putamen, spinal cord, liver, blood, skeletal muscle, adipose, bone marrow and testis. Housekeeping gene *TUBB* was used as the reference gene.

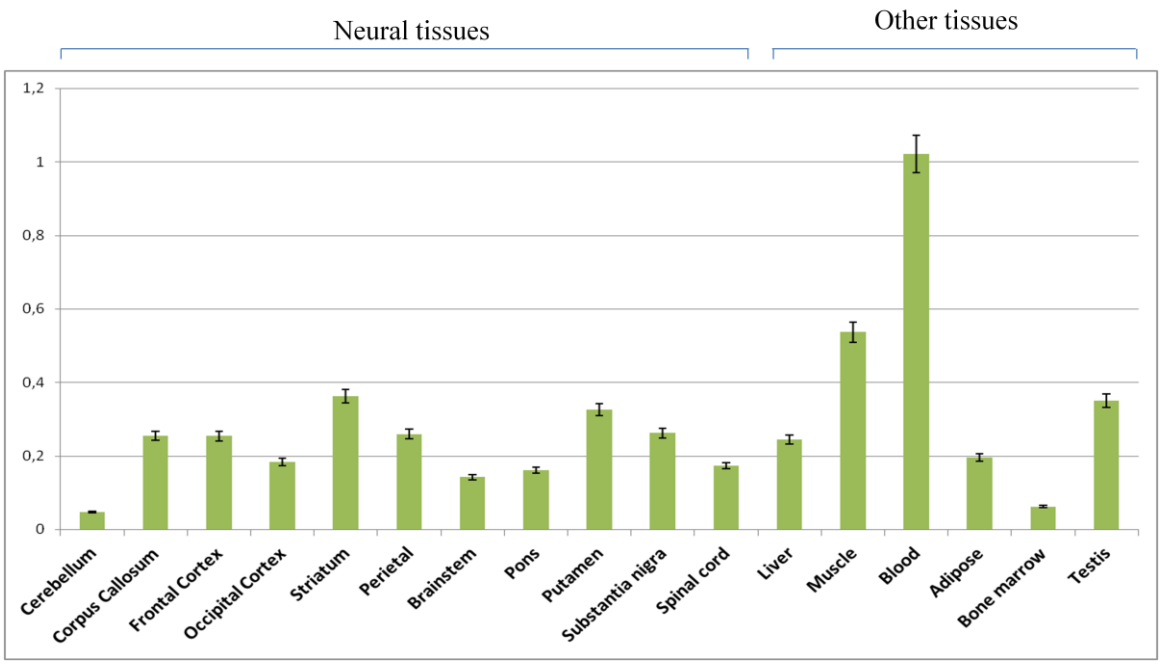



Figure 5.4. Expression of *PTHRD1* in various tissues.

## 5.2. ID and Hypothyroidism (IDH) Family

### 5.2.1. Linkage Analysis

Linkage analysis was performed in several steps since not all members of the family were subjected to SNP genotyping together. In the final linkage analysis all available SNP genotype data were included (Figure 3.2); previous linkage analysis results were ignored. Parents were assumed as first-cousins because the capacity of the program was not sufficient. Linkage analysis yielded a maximal LOD score 3.55 for nine regions with sizes >300 Mb, as listed in Table 5.5. Candidate regions obtained by linkage analysis were evaluated in Excel for IBD, eliminating the regions on chromosomes 3, 4, 7, 8 and 14 (Figure 5.5). Phenotypes mapped to the remaining loci on chromosomes 2, 11, 15 and 17 were evaluated in OMIM data base, and no relevant phenotype was found.

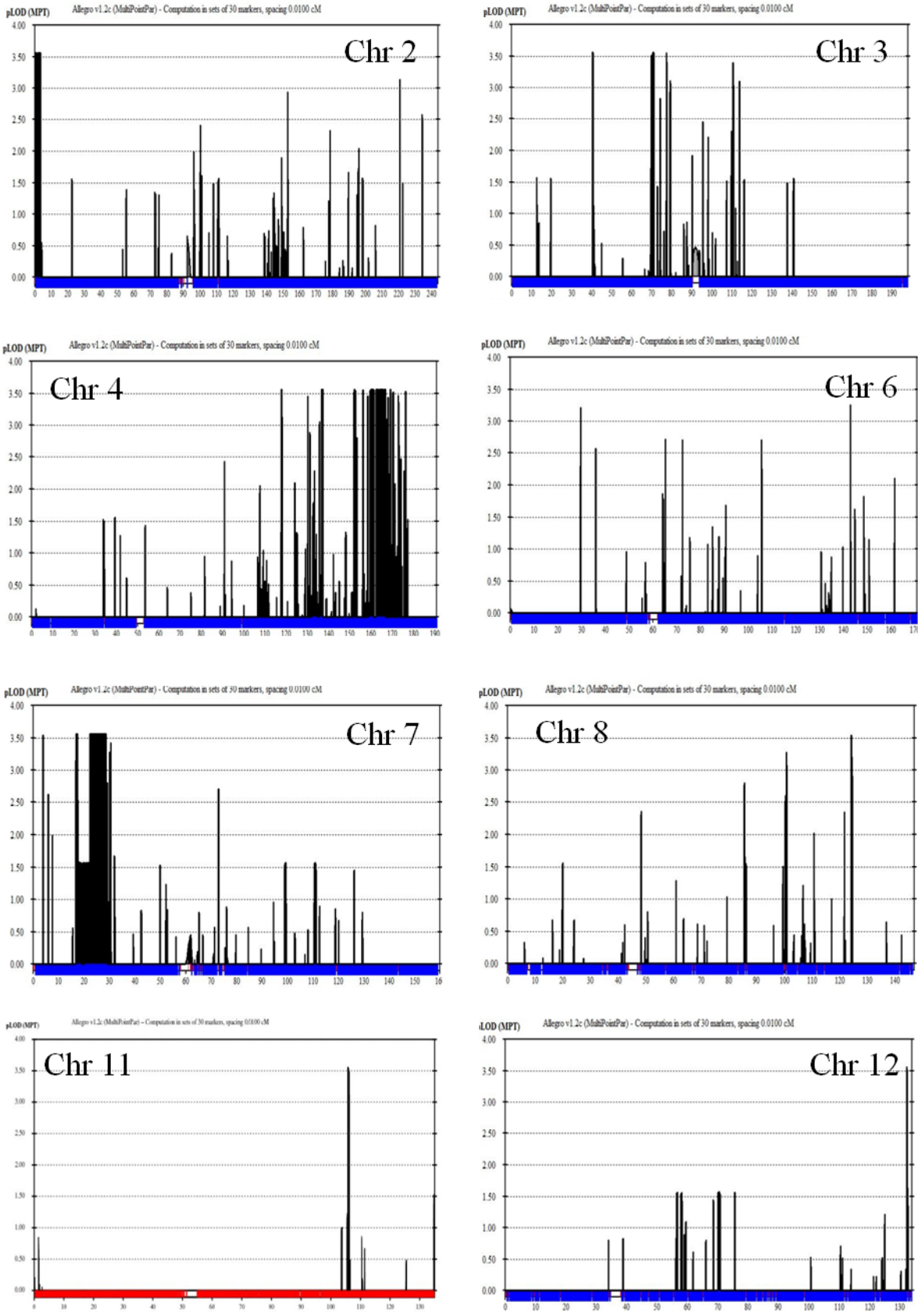



Figure 5.5. Multipoint LOD score graphs for IDH Family. Only chromosomes with LOD scores >3 are presented.

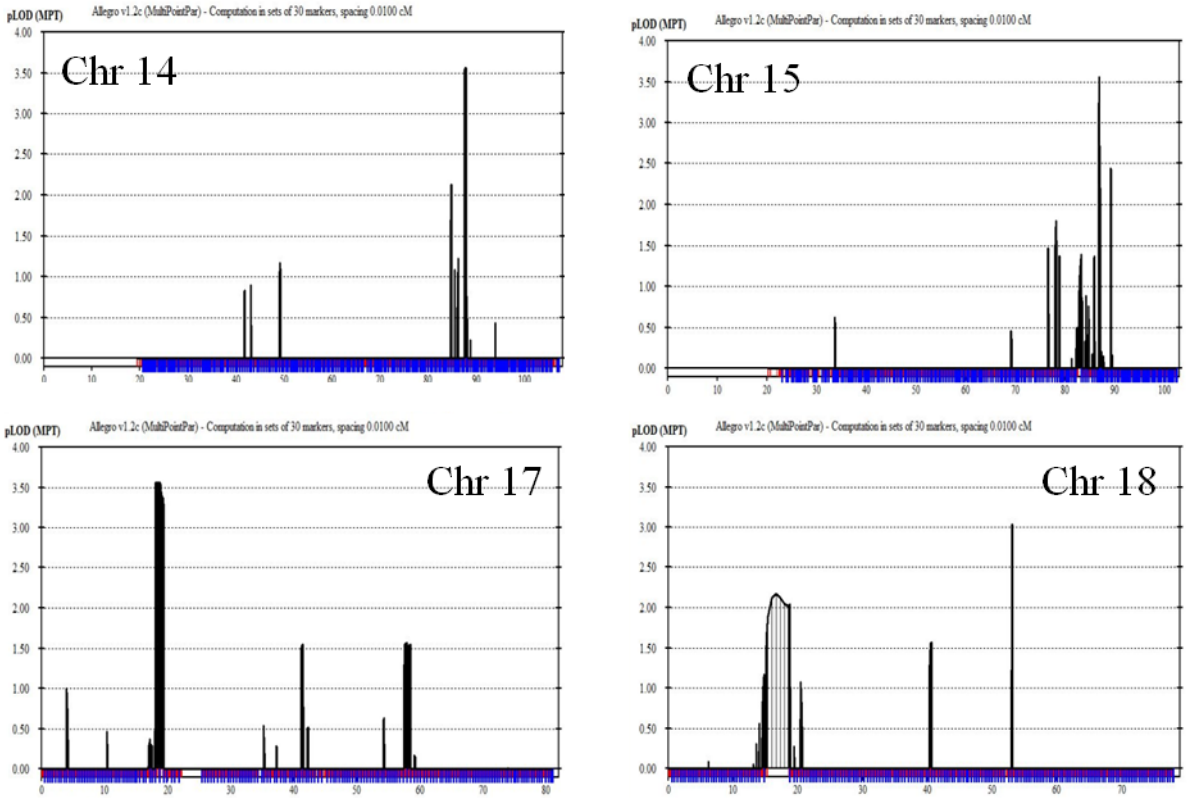



Figure 5.5. Multipoint LOD score graphs for Family 11. Only chromosomes with LOD scores >3 are presented (cont.).

Table 5.5. Maximal homozygosity regions obtained by linkage analysis and evaluation for IBD in Excel.

| Chr | Maximal | | SNP | | Size (bp) | LOD score | Remarks |
|---|---|---|---|---|---|---|---|
| | Start | End | Start | End | | | |
| 2 | 18674 | 3433368 | rs10195681 | rs10165836 | 3414694 | 3.55 | IBD |
| 17 | 17970229 | 18855611 | rs2955372 | rs2076562 | 885382 | 3.55 | IBD |
| 7 | 16731253 | 17522503 | rs696282 | rs17137763 | 791250 | 3.55 | Not IBD |
| 4 | 136764416 | 137408872 | rs12645397 | rs6535149 | 644456 | 3.55 | Not IBD |
| 3 | 40283255 | 40885879 | rs1880763 | rs627998 | 602624 | 3.55 | Not IBD |
| 14 | 87518037 | 88063317 | rs8016092 | rs1887350 | 545280 | 3.55 | Not IBD |
| 11 | 105503658 | 105923373 | rs4754133 | rs1939801 | 419715 | 3.55 | IBD |
| 15 | 86720146 | 87119340 | rs7169987 | rs2448928 | 399194 | 3.55 | IBD |
| 8 | 124068128 | 124411678 | rs7826050 | rs34573020 | 343550 | 3.54 | Not IBD |

After reevaluation of the family in a recent visit by our Pakistani collaborators, we learned that patient 322 for whom we had the exome data did not have ID and all ID patients also had hypothyroidism. We then decided to search for a new gene responsible for the ID in the family. Linkage analysis was performed using the genotype data of two ID patients and two unaffected siblings, and many regions were obtained with a maximal LOD score of 2.95 in nearly all chromosomes. Those regions were investigated in Excel to determine the exact limits of the shared homozygous regions. Seven regions >1 Mb were obtained.

### 5.2.2. Deletion-Duplication Analysis

Deletion-duplication analysis was performed again with 1-probe order by cnvPartition. No shared deletion-duplication was found.

### 5.2.3. Exome Sequence Analysis

DNA sample of a patient (322) was subjected to exome sequencing. Variants with frequencies <0.01 in the candidate regions (Table 5.5) were evaluated (Table 5.6). The ones that are found in-lab exome files were eliminated. The only candidate variant remained was *TPO* c.719A>G (p.Asp240Gly) at 2p25.3. This gene is associated with hypothyroidism, and in some patients ID was also observed (Mittal *et al.*, 2016). The variant is novel and predicted as disease causing by Mutation Taster, deleterious by SIFT and probably damaging by PolyPhen-2.

Table 5.6. Candidate variants in the regions listed in Table 5.5.

| Chr | Start | Ref | Alt | Alt/Tot* | Gene | Freq** | dbSNP | Change | Remarks*** |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 283238 | - | T | 5/5 | *FAM150B* | - | - | intronic | IOF |

* Alt/Total Depth

*Freq, highest frequency in ExAC or gnomAD

**Whether found in other in-lab exome files (IOF) or not (NIOF).

Table 5.6. Candidate variants in the regions listed in Table 5.5 (cont.).

| Chr | Start | Ref | Alt | Alt/ Tot* | Gene | Freq** | dbSNP | Change | Remarks*** |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 676480 | - | AAGT | 42/42 | *TMEM18* | - | rs74164457 | intronic | IOF |
| **2** | **1459954** | **A** | **G** | **84/84** | ***TPO*** | **-** | **-** | **exonic** | **NIOF** |
| 2 | 1696516 | CA | - | 29/29 | *PXDN* | - | rs35513437 | intronic | IOF |
| 2 | 3358462 | TGAAGTTAGATTTA | - | 6/6 | *TSSC1* | 0.00011 | rs11269308 | intronic | IOF |
| 17 | 18034185 | - | GT | 4/4 | *MYO15A* | - | - | intronic | IOF |
| 17 | 18047673 | - | C | 9/9 | *MYO15A* | - | rs11421905 | intronic | IOF |
| 17 | 18047774 | - | C | 97/97 | *MYO15A* | 0.8 | rs11421904 | intronic | IOF |
| 17 | 18286917 | - | G | 3/3 | *EVPLL* | - | rs112094897 | intronic | IOF |

* Alt/Total Depth

**Freq, highest frequency in ExAC or gnomAD

***Whether found in other in-lab exome files (IOF) or not (NIOF).

### 5.2.4. Validation of the Variant

The *TPO* variant was validated by Sanger sequencing in patient (322) with exome file and an unaffected brother (323) (Figure 5.6). Family members were tested for the variant by SSCP analysis. The pattern of the bands was not as we expected. Therefore, all patients and unaffected siblings that we have DNA samples from (304, 311, 312, 318, 322, 323, 325 and 328) were subjected to Sanger sequencing. *TPO c.*719A>G variant was shown to segregate with the disease for hypothyroidism traits. Patients were homozygous, and all unaffected siblings were heterozygous. Final pedigree with genotypes for *TPO* variant is presented in Figure 5.7.

In summary, only one candidate loci were detected and only one candidate variant was identified that segregated hypothyroidism. The disease in the family is compatible with biallelic *TPO* mutation.

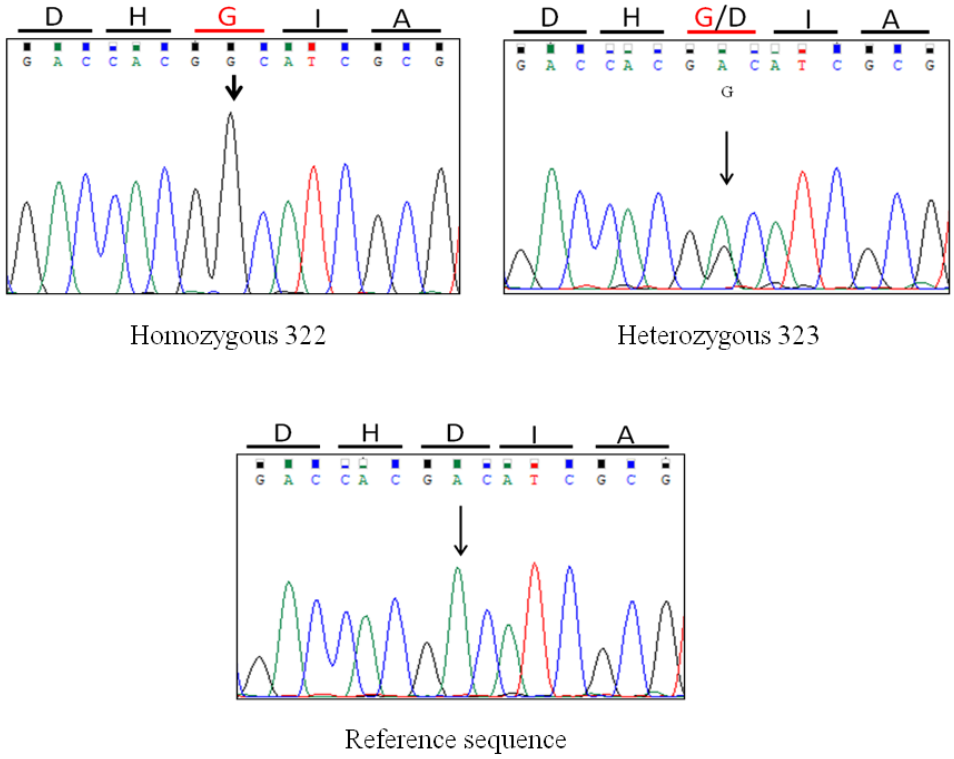



Figure 5.6. Electrophoretograms showing mutation *TPO c.*719A>G (p.Asp240Gly).

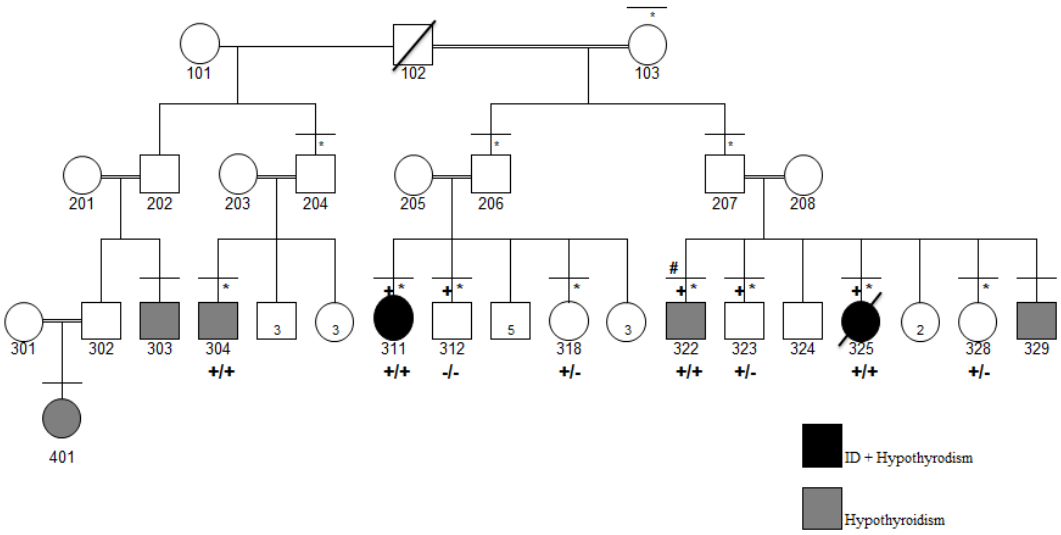



Figure 5.7. Final pedigree of IDH Family showing also genotypes for *TPO* c.719A>G variant. DNA was available for individuals marked *. SNP genotype data were generated for individuals marked +. # marked individual was subjected to exome sequencing.

## 5.3. Syndromic Intellectual Disability (SID) Family

### 5.3.1. Linkage Analysis

Initial multipoint linkage analysis was performed separately for the two branches of the family. The regions with LOD score >2.5 were evaluated (at chromosomes 7, 8, 15 and 19). Regions 7q22.2-22.3, 8q24.23-24.3 and 19p13.2 were eliminated since haplotypes were not IBD. The only remaining region is approximately 2.2 Mb at 15q15.2-21.1 between nucleotides 43,336,670 (rs1197547) and 45,095,902 (rs1288092), and it did not harbor any known genes possibly related to ID.

Further linkage analyses were performed to confirm that we have not missed any candidate regions. The analyses were performed in several steps (Analysis A, B and C) (Section 4.2.2.3). SNP data of all participants could not be included in linkage analyses A and B due to the limitations of the program. In linkage analysis A, markers were selected with 0.01-Mb spacing and sets of 30 markers were used in Allegro. The pedigree used is presented in Figure 4.1. Linkage analysis B was performed for the larger core family for the chromosomes yielding multipoint LOD scores >2.8 in linkage analysis A, i.e. chromosomes 1, 6, 7, 8, 13, 15, 17, 19 and 21 (Figure 5.8). Simplified pedigree is presented in Figure 4.2. Chromosomes 1, 8, 13, 17 and 21 were eliminated, because the new LOD scores were lower. Regions yielding LOD scores >2.5 (on chromosomes 6, 7, 15 and 19) were investigated in Excel for possible identity by descent and the sizes of the homozygosity regions (Figure 5.9). The region on chromosome 7 and four of the six regions on chromosome 15 were eliminated because of the small size (<500 kb). We investigated the genotypes in the remaining homozygous regions (>500 kb). Patients and no others shared the homozygosity possibly identical by descent in the two regions on chromosome 15 but not in the regions on chromosomes 6 and 19. As a result, all those regions were eliminated except for 15q15.1-21.1 and 15q22.31.

Final linkage analysis (analysis C) was performed for the two regions on chromosome 15q with SimWalk including all available SNP data and with 10 marker sets. The simplified pedigree used is presented Figure 4.3. At 15q15.1-21.16 only one region

>150 kb yielded the maximal multipoint LOD score of 3.85 (Figure 5.10). The maximal homozygosity was approximately 2.2 Mb, between rs17767270 (nucleotide 42898612) and rs1288092 (45090821). At 15q22.31 only one region yielded a maximal LOD score >3.8. It was 665 kb between rs639812 (65619797) and rs16949219 (66285299).

None of the rare (frequency <0.005) or novel variants in the 665-kb region at 15q22.31 were in the homozygous state; they were all heterozygous. Therefore, this region was evaluated as noninformative with respect to the SNP data and eliminated.

### 5.3.2. Exome Sequence Analysis

DNA sample of affected individual 502 was subjected to exome sequencing. All candidate variants in the identified gene region at 15q15.1-21.16 were evaluated, and the only exonic candidate variant was *PDIA3* c.170G>A, p.Cys57Tyr (NM_005315). The variant segregates with the disease and is not listed in dbSNP, Exome Variant Server, ExAC and gnomAD database. A comparison using HomoloGene indicated that the substituted residue is in a region of seven amino acids that are absolutely conserved across all species, including *Drosophila, C. elegans* and plants. Online mutation prediction tools PolyPhen-2, Mutation Taster and SIFT all predicted the variant as pathogenic (Table 5.7).

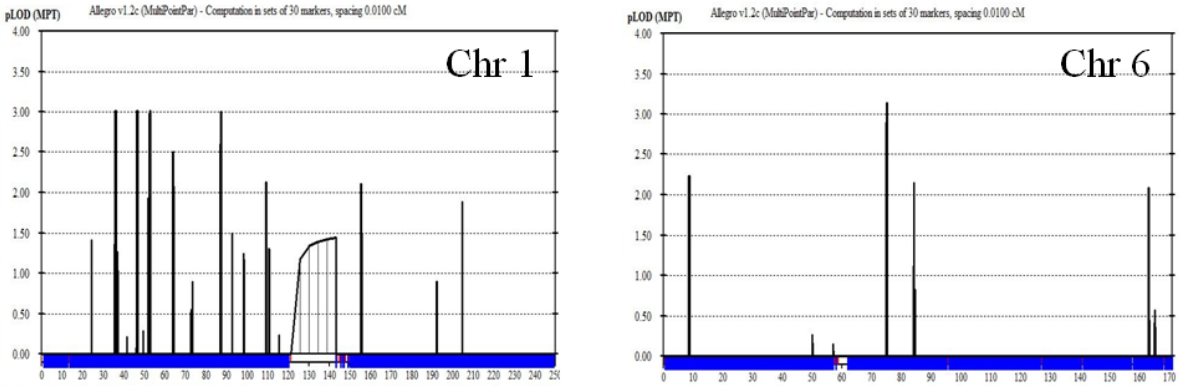



Figure 5.8. Multipoint LOD score graphs for Linkage A. Only chromosomes with LOD scores >2.8 are presented.

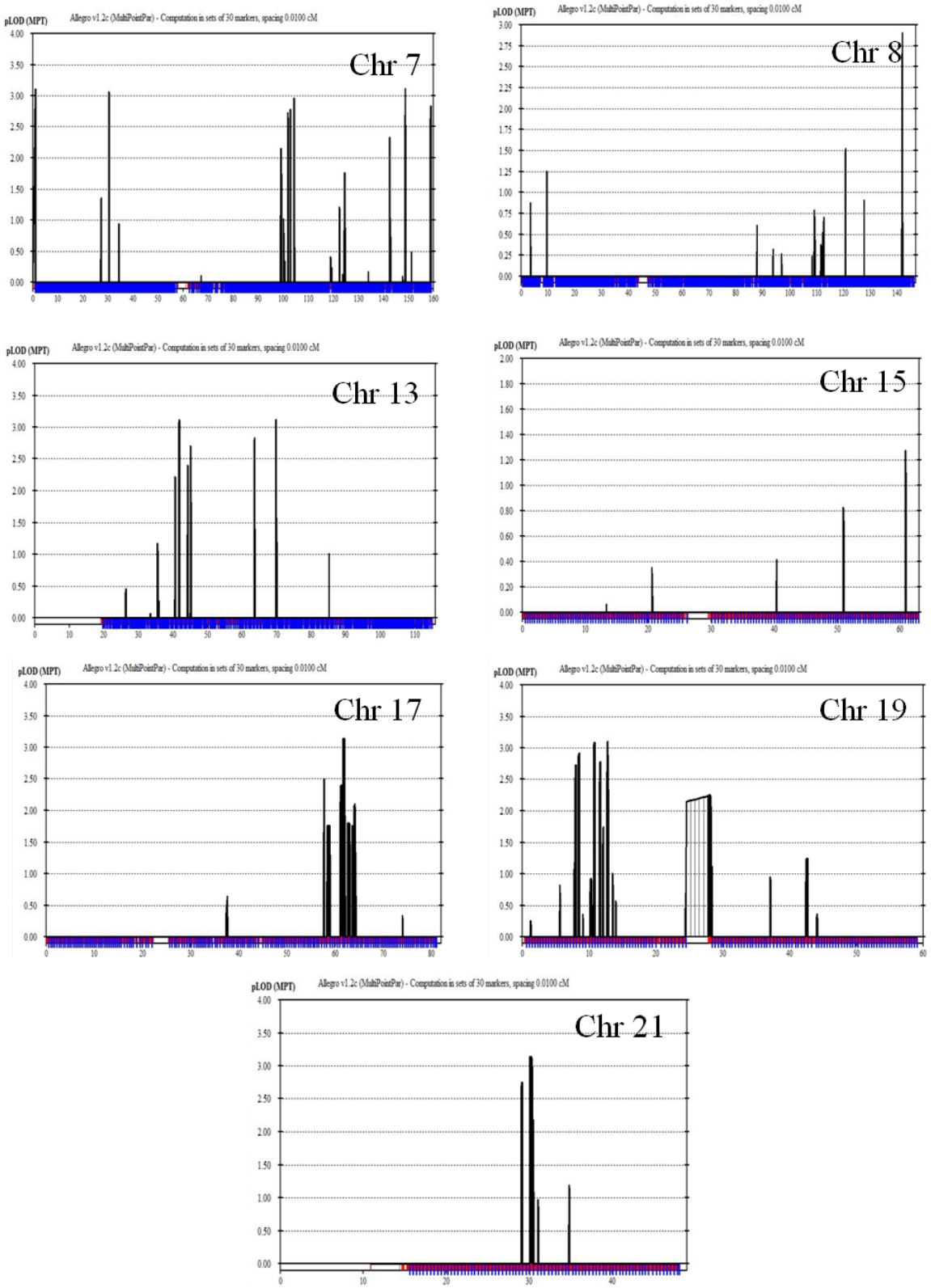



Figure 5.8. Multipoint LOD score graphs for Linkage A. Only chromosomes with LOD scores >2.8 are presented (cont.).

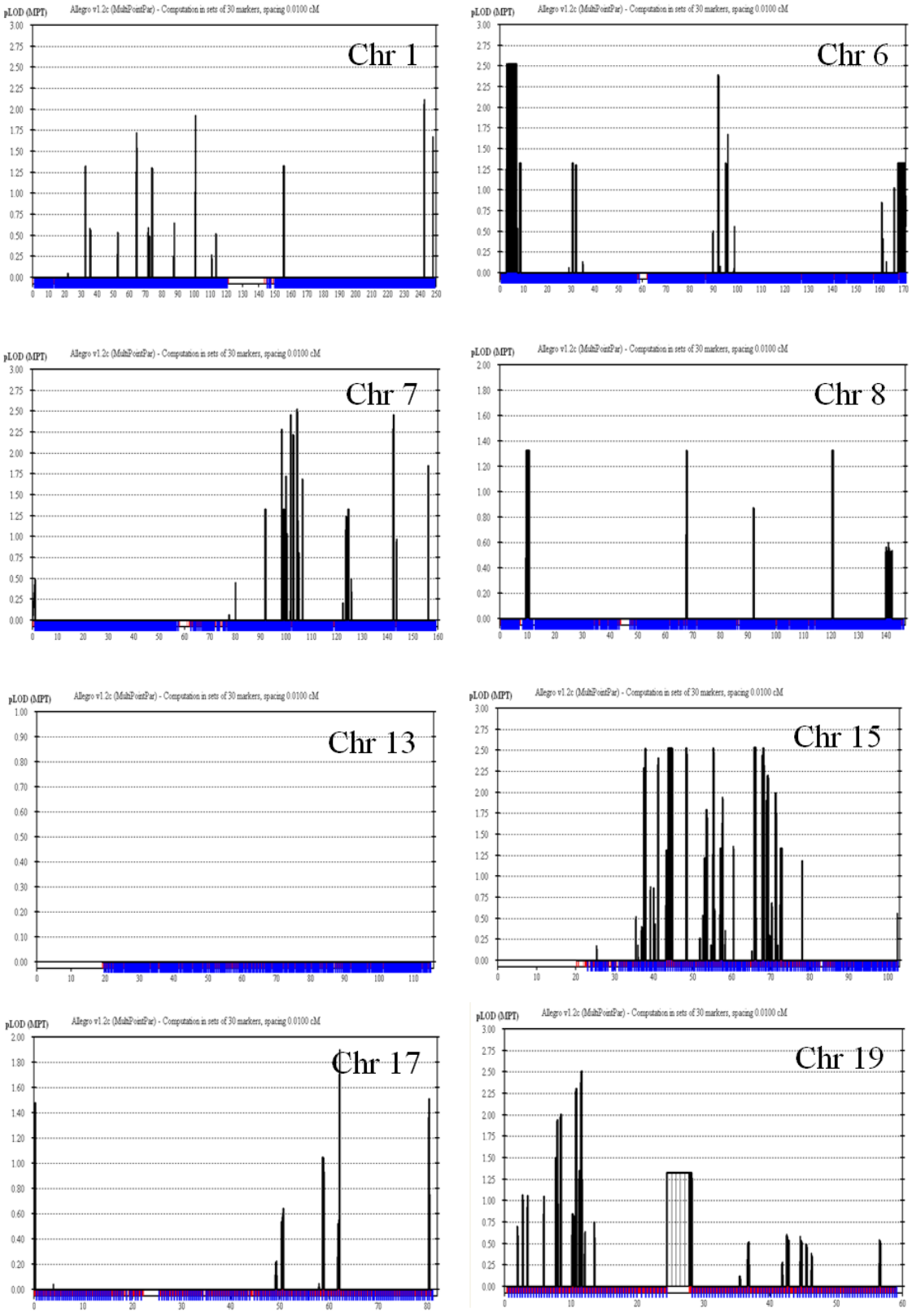



Figure 5.9. Multipoint LOD score graphs for Linkage B.

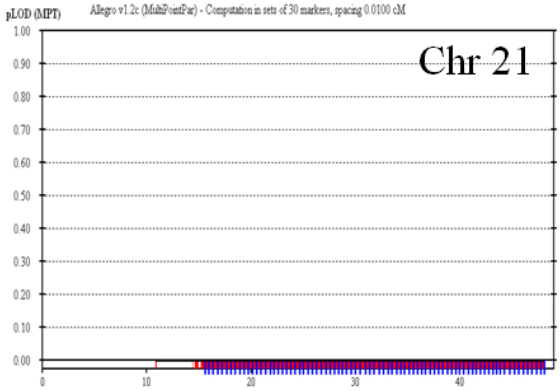



Figure 5.9. Multipoint LOD score graphs for Linkage B. (cont.)

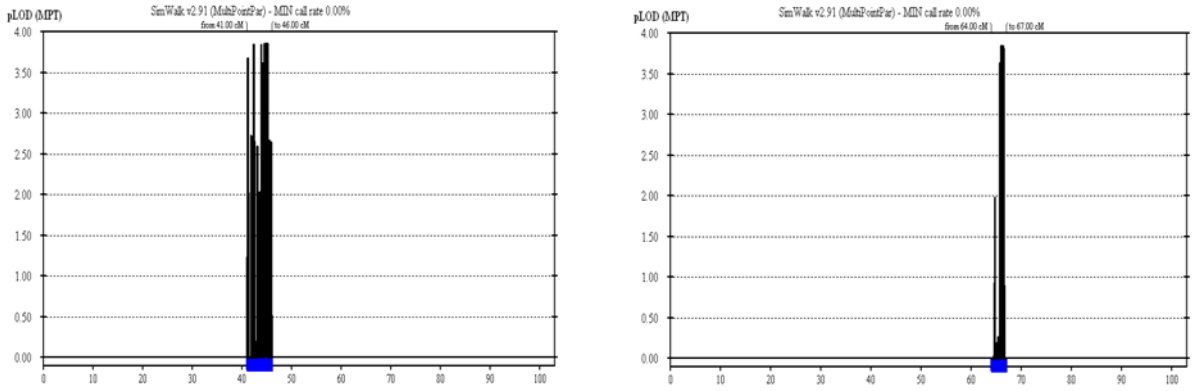



Figure 5.10. Multipoint LOD score graphs for Linkage C.

Table 5.7. Predictions of potential damage to protein of variant *PDIA3* c.170G>A (p.Cys57Tyr; NM_005315) using computational algorithms.

| **ALGORITHM** | **PREDICTION** |
|---|---|
| Mutation Taster | Disease causing |
| PROVEAN | Deleterious |
| SIFT | Damaging |
| PolyPhen-2 | Probably Damaging |
| UMD-Predictor | Pathogenic |
| Splice Finder | Creation of an exonic ESS site<br>Potential alteration of splicing |
| NNSPLICE | No splice site |

### 5.3.3. Validation of the Variant and Family Screening

Variant is validated by Sanger sequencing in family members (Figure 5.11).

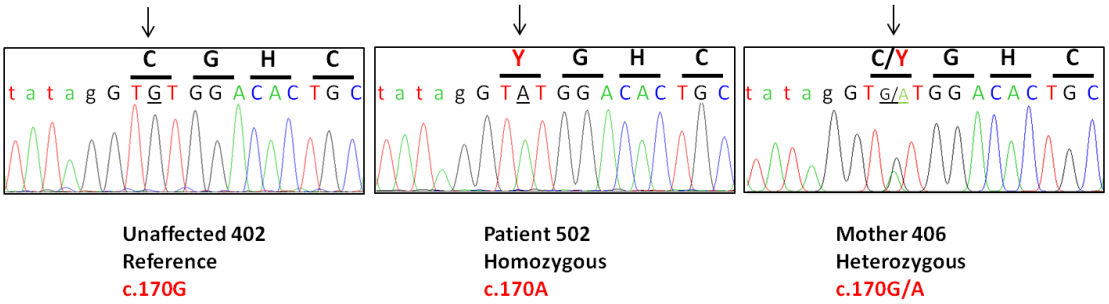



Figure 5.11. Electrophoretograms showing mutation c.170G>A in *PDIA3*.

## 5.4. Spinocerebellar Ataxia (SCA) Family

### 5.4.1. Linkage Analysis and Homozygosity Mapping

In order to identify the disease locus in this family afflicted with spinocerebellar ataxia, linkage analysis was performed in several steps to not exceed the capacity of Allegro program and since some individuals were subjected to SNP genotyping later than the first three patients.

Initial linkage analysis was performed using the genotype data of two affected brothers, an affected cousin (406, 412 and 413 in Figure 3.4), and the maximal LOD score was 3.00 (significant). Whether any disease with similar characteristics was reported in any of the candidate regions in database OMIM was investigated, and no such disease was found. To confirm that the regions that were found were indeed candidate loci, genotypes were investigated in Excel in all patients, and regions where patients did not share homozygosity were eliminated. In the remaining five regions linkage was performed with SimWalk (0.01 spacing and 30 marker sets) using the genotype data of 4 patients (including also patient 502) and the actual pedigree; SNP markers in those candidate regions plus markers in 1-Mb flanking regions were utilized. The regions, their lengths and maximal LOD scores are 41,614,306 bp (rs7930633) – 69,172,091 bp (rs603965),

27,557,785-bp long and LOD score 4.52 on chromosome 11, 96,167,389 bp (rs9842685) – 97,404,612 bp (rs719278), 1,237,223-bp long and LOD score 4.18 on chromosome 3, 111,841,592 bp (rs3177979) – 113,024,793 bp (rs12580178), 1,183,201-bp long and LOD score 4.53) on chromosome 12, 104,213,502 bp (rs11191291) – 105,225,201 bp (rs3014192), 1,011,699-bp long and LOD score 4.5) on chromosome 10, and 58,592,033 bp (rs6503977) – 59,255,231 bp (rs917571), 663198-bp long and LOD score 4.0 on chromosome 17 (Figure 5.12).

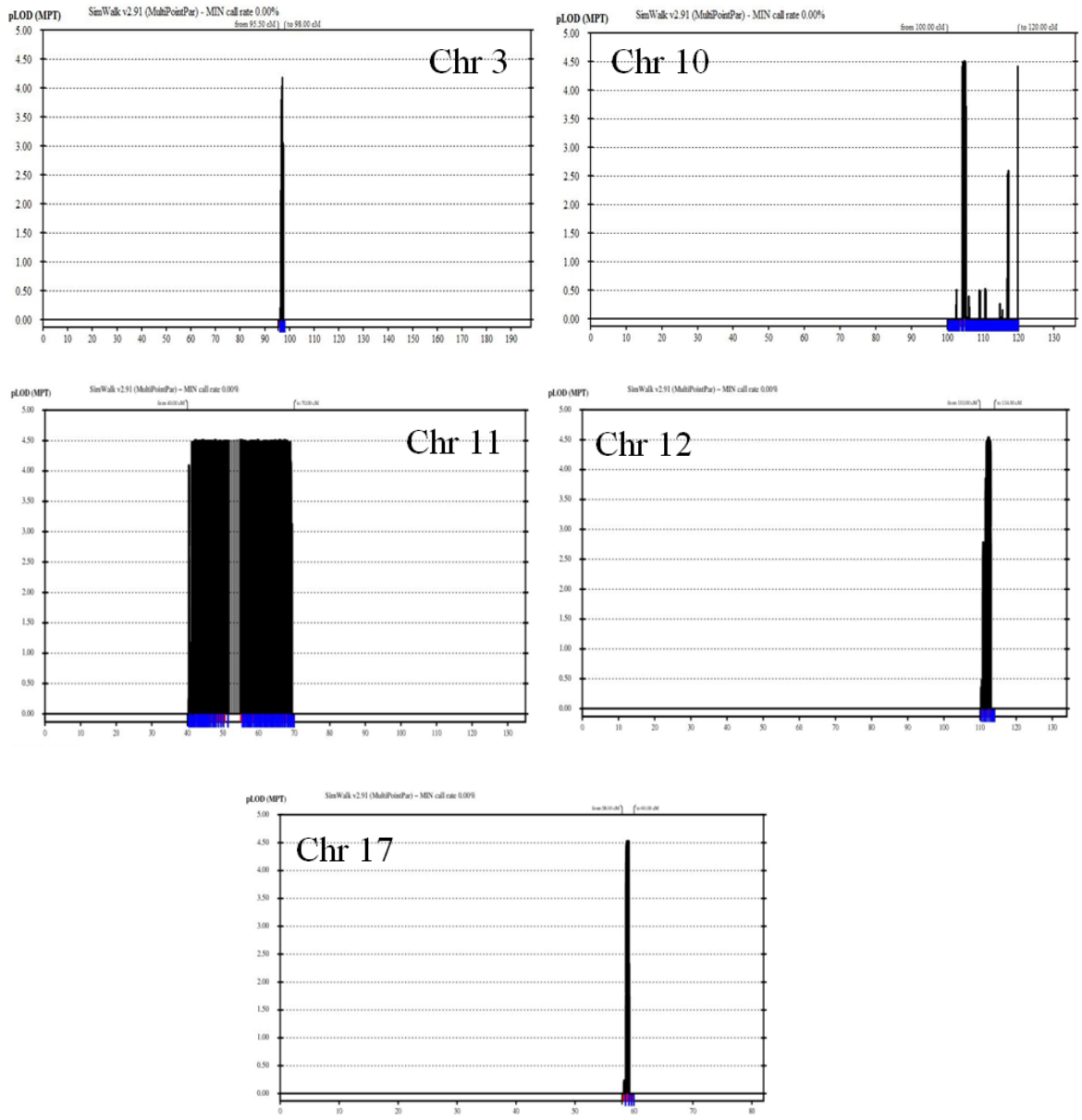



Figure 5.12. Multipoint LOD score graphics performed by SimWalk for the candidate regions detected in the initial linkage analysis.

Later SNP genotype data were generated for three more patients and their three unaffected relatives. To evaluate whether the five homozygous regions are really candidates, a third linkage analysis was carried out by including the newly generated genotype data for additional participants and using Allegro program. A part of the pedigree including the data of the three affected and two unaffected individuals, 412 (A), 413 (A), 301 (UA, newly added), 406 (A) and 408 (UA, newly added) were used for linkage analysis not to exceed the capacity of the program. Linkage analysis yielded 13 regions with LOD scores >3. Nine of them including the five regions that are mentioned above as a result of the first linkage analysis were eliminated since not all affected individuals were homozygous in those regions. Further, the regions on chromosome 8, 15 and 19 were eliminated because affected brothers 412 and 413 did not share genotypes. The only possible candidate region remained was again 41,614,306 bp (rs7930633) – 69,172,091 bp (rs603965), 27,557,785-bp long (LOD score 3.73) at 11p11.2-q13.2 (Figure 5.13).

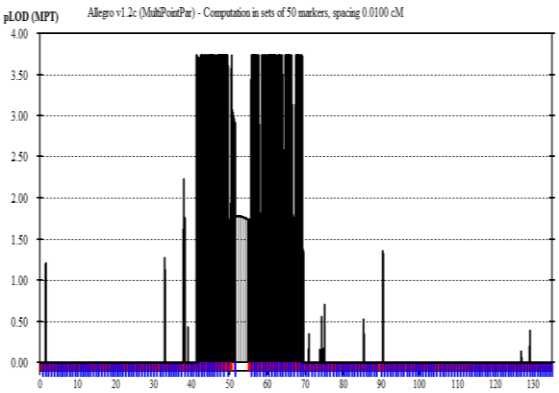



Figure 5.13. Multipoint LOD score graphic for chromosome 11 in the final linkage analysis.

### 5.4.2. Deletion-Duplication Analysis

Deletion-duplication analysis was performed for all chromosomes and family members with SNP genotype data using cnvPartition. No duplication or deletion that was common to all patients was found.

On the genotype excel sheet, one possible homozygous deletion was detected at 17q25.3; a SNP was no-call in all patients but was read for the other family members,

indicative of a shared deletion. This no-call SNP (rs9988824) is flanked by SNPs rs1879929 (59,184,192 bp) and rs1453547 (59,189,912 bp). In that region, there were no genes except for a part of an olfactory receptor gene (*ORA5A2*). In the database of genomic variants a deletion spanning the no-call SNP was reported. The deletion was not rare, found in 8/95 people, indicating an allele frequency of 8/190 or >0.04 (Wong *et al.*, 2006). Hence, we concluded that this deletion cannot be the causative mutation.

### 5.4.3. Evaluation of Exome Sequence Results

We searched for the putative disease mutation at the identified locus in the exome file of patient 412. All candidate variants selected as described in section1.8.4. (strategy) in the identified gene region at 11p11.2-q13.2 were evaluated (Table 5.8). The ones found in-lab exome files were eliminated. The variants in *MPEG1* (Mutation Taster: disease causing, SIFT: deleterious, and PolyPhen: probably damaging) and *SPTBN2* (Mutation Taster: disease causing) were considered candidate mutations. *SPTBN2* variant is not reported in ExAC or gnomAD database. Primers were designed to validate the *SPTBN2* variant in patient 412. Priority was given to that variant because the gene is associated with spinocerebellar ataxia. Nonsynonymous variant *ACTN3* p.Arg211Gln has a frequency of 0.03 in South Asian samples in ExAC, indicating that it is not rare. Variant *MPEG1* c.671C>T (p.Ser224Phe) is also eliminated since the frequency in South Asians (ExAC) is 0.006, again not rare. No interaction was found among *ACTN3*, *MPEG1* and *SPTBN2* (GeneMANIA).

Whether the variants with frequencies >0.99 in the region at 11p11.2-q13.2 were present in the exome files was investigated. There were twenty five such variants in the region, and three of them were not present in the patient exome data. However, they were not in all of the 56 other in-lab exome files, either (Table 5.9).

Table 5.8. Candidate exonic and splicing variants at the disease locus 11p11.2-q13.

| Start | Ref Base | Alt Base | Depth (Alt/Total) | Gene | Region | Change* | Remarks** |
|---|---|---|---|---|---|---|---|
| 46342261 | - | G | 105/105 | *CREB3L1* | Exonic splicing | Frameshift_ insertion | IOF |
| 47788670 | GGT GGT | - | 6/6 | *FNBP4* | Exonic | Nonframeshift_deletion | IOF |
| 58979668 | G | A | 28/29 | *MPEG1* | Exonic | Nonsyn | Disease causing. ExAC South Asians 0.0067 |
| 60609676 | C | T | 32/32 | *CCDC86* | Exonic | Nonsyn | Polymorphism |
| 61257355 | G | C | 35/35 | *PPP1R32* | Exonic | Nonsyn | IOF |
| 66322675 | G | A | 88/88 | *ACTN3* | Exonic | Nonsyn. | Polymorphism |
| 66455551 | C | G | 59/59 | *SPTBN2* c.6375-1G>C | Intronic splicing | Splicing | Disease causing |

* nonsyn, nonsynonymous**Whether found in other in-lab exome files (IOF), and if not, prediction of pathogenicity with Mutation Taster

Table 5.9. The exonic variants with frequencies >0.99 in the region at 11p11.2-q13.2 that are not present in patient's exome data

| Start | Ref Base | Alt Base | Hom/het | Gene | Change* | Freq (ExAC-SA) | Present in #of in-lab exome files |
|---|---|---|---|---|---|---|---|
| 49974371 | G | A | Hom | *OR4C13* | Nonsyn | 0.999 | 8 |
| 62292882 | G | T | Hom | *AHNAK* | Nonsyn | 1 | 44 |
| 64809090 | T | G | Hom | *SAC3D1* | Nonsyn | 0.997 | 39 |

As the mutation we identified in *SPTBN2* alters the last nucleotide of intron 31 which is always a G in human genes, we predicted that it would most probably affect splicing. In order to show aberrant splicing, the exclusion of exon 32, we tried to analyze *SPTBN2* transcripts. Since databases report expression of the gene in blood, cDNA was generated from total RNA obtained from patient blood cells.

The splicing *SPTBN2* variant was investigated by online tools. Mutation Taster predicted it as disease causing due to a splice site change that caused the acceptor site to be broken. If that was the case, we would expect that exon 32 would be excluded. Exclusion of exon 32 leads to the synthesis of 137 non-native amino acids and a premature termination codon due to the frameshift, and protein would be truncated (Figure 5.14). Terminal 265 native amino acids would be deleted as designated below (Figure 5.15). Alternatively, Splice Finder predicted that the acceptor site was altered most probably, affecting splicing by activating a cryptic site seven nucleotides downstream. Then, those seven nucleotides would be included in the intron causing a frameshift that would lead to the synthesis of 189 non-native amino acids and premature termination. NetGene2 and NNSPLICE detected the known splice site in the reference sequence but did not detect any potential splice site in the mutant sequence. SIFT, Polyphen-2, PROVEAN and UMD-Predictor could not be applied since the variant is intronic (Table 5.10).

1937 RDVSSADLVIKNQQGIKAEIEARADRFSSCIDMGKELLARSHYAAEEISEKL
SQLQARRQETAEKWQEKMDWLQLVLEVLVFGRDAGMAEAWLCSQEPLVRSAEL
GCTVDEVESLIKRHEAFQKSAVAWEERFCALEKLTALEEREKERKRKREEEERRK
QPPAPEPTASVPPGDLVGGQTASDTTWDGPCWDNRDLSTAASPKGRDLAQGTKP
MGPGERGRPGLGARPHLQCPRAGLPSQPMLPPCRLEAQSHLRPGADGGDAVPQA
GDGGLREEGCQQVLAERVLCPAAWEPRLLQGCQGSQRGSAIPRRSACQPGQGPG
QRRL 2262

Figure 5.14. The 137 residues in red are the deduced non-native amino acids encoded in the patients.

Reference sequence
1937 RDVSSADLVIKNQQGIKAEIEARADRFSSCIDMGKELLARSHYAAEEISEKL
SQLQARRQETAEKWQEKMDWLQLVLEVLVFGRDAGMAEAWLCSQEPLVRSAEL
GCTVDEVESLIKRHEAFQKSAVAWEERFCALEKLTALEEREKERKRKREEEERRK
QPPAPEPTASVPPGDLVGGQTASDTTWDGTQPRPPPSTQAPSVNGVCTDGEPSQPL
LGQQRLEHSSFPEGPGPGSGDEANGPRGERQTRTRGPAPSAMPQSRSTESAHAATL
PPRGPEPSAQEQMEGMLCRKQEMEAFGKKAANRSWQNVYCVLRRGSLGFYKDA
KAASAGVPYHGEVPVSLARAQGSVAFDYRKRKHVFKLGLQDGKEYLFQAKDEA
EMSSWLRVVNAAIATASSASGEPEEPVVPSTTRGMTRAMTMPPVSPVGAEGPVVL
RSKDGREREREKRFSFFKKNK 2390

Figure 5.15. The 265 residues highlighted are deduced to be deleted in the patients.

In conclusion, the disease was mapped to 11p11.2-q13.2 with a maximal multipoint LOD score of 4.52. The only candidate variant is splice mutation is *SPTBN2* c.6375-1G>C which is predicted as disease causing and segregates in the family.

Table 5.10. Summary of the predictions of mutation severity by online tools.

| **ONLINE TOOL** | **PREDICTION** |
|---|---|
| Mutation Taster | Disease causing |
| Splice Finder | Acceptor site is altered, most probably affecting splicing |
| NetGene2 | Splice site is abolished |
| NNSPLICE | Splice site is abolished |
| SIFT | Not applicable since the variant is intronic |
| PolyPhen-2 | Not applicable since the variant is intronic |
| PROVEAN | Not applicable since the variant is intronic |
| UMD-Predictor | Not applicable since the variant is intronic |

### 5.4.4. Validation and Screening for the Variant

*SPTBN2* variant was validated by Sanger sequencing (Figure 5.16), and then all family members were screened for it by SSCP analysis. The results were as expected: All affected individuals but no one else carried the mutation in the homozygous state (Figure 5.18).

We attempted many times to amplify *SPTBN2* transcripts in patient and heterozygous sibling blood samples, but we could not obtain any product specific to the gene. We could not amplify *SPTBN2* specific sequences from commercial blood cDNA, either. Since the gene is expressed in cerebellum according to expression databases, we used commercial cerebellar cDNA and were able to amplify the target gene region, indicating that the PCR strategy was working (Figure 5.17). We deduced that the expression of the gene in blood cells was very low or zero.

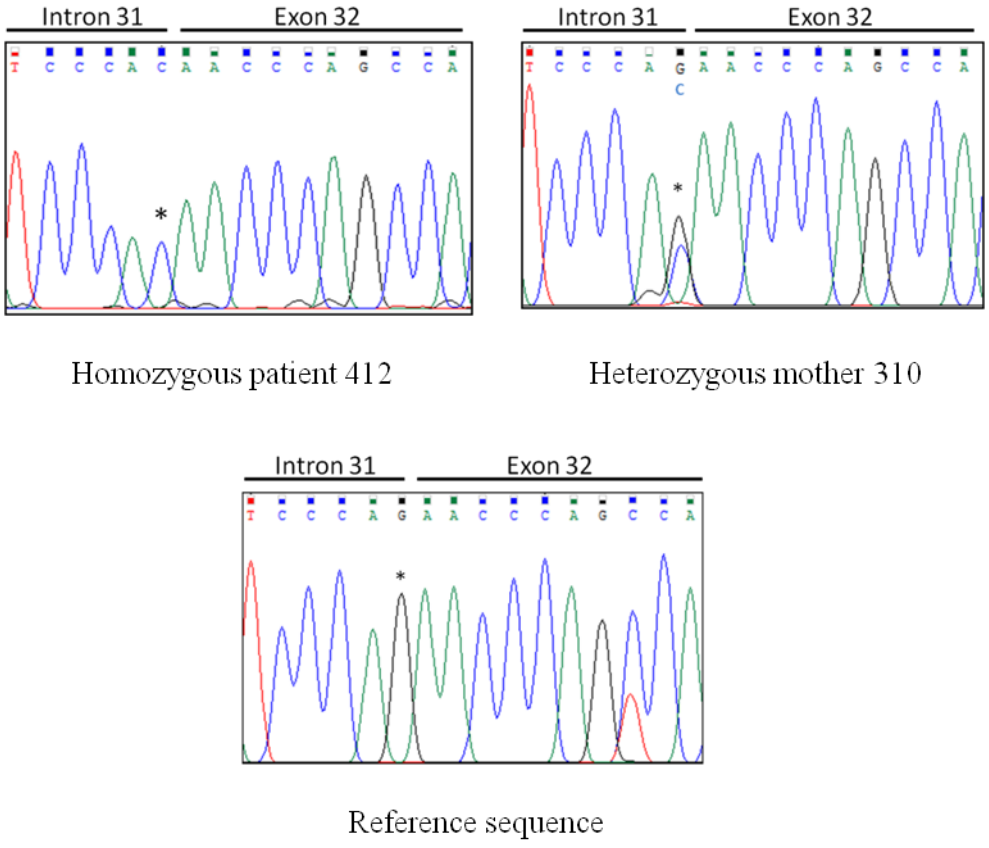



Figure 5.16. Electrophoretograms for Sanger sequencing for *SPTBN2* c.6375-1G>C.

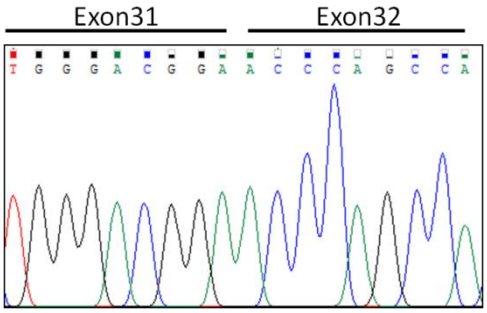



Figure 5.17. Sanger sequencing result across the junction of *SPTBN2* exon 31 and exon 32 using cerebellum cDNA as template.

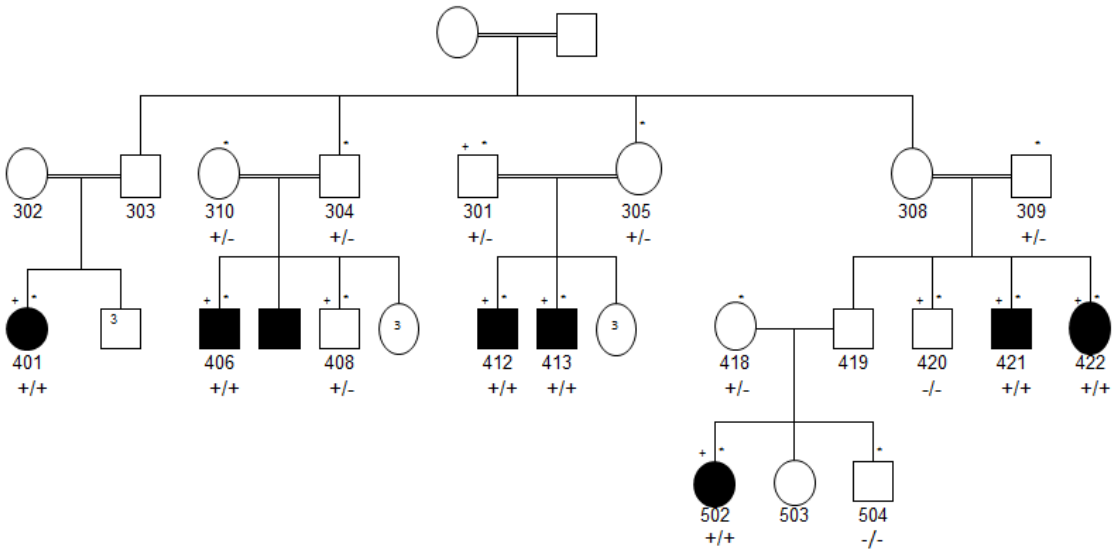



Figure 5.18. Pedigree of SCA family showing *SPTBN2* c.6375-1G>C genotypes.

## 5.5. Bardet-Biedl Syndrome (BBS) Family

### 5.5.1. Linkage Analysis

To investigate whether the disease in the family maps to a known BBS locus, linkage analysis was performed with the genotype data of four affected and three unaffected family members including parents (assumed as second cousins) for only branch

A (except 507). Linkage analysis yielded 12 regions with maximal LOD scores >3. Genotypes in those high LOD regions and BBS genes regions for the whole family including branch B were evaluated in MS Excel. Except for the BBS genes on chromosomes 1 and 17 (*MKS1* and *CCDC28B*, respectively), patients were heterozygous at all other BBS loci. The homozygous regions for these genes were very small, 328,382 bp and 112,928 bp, respectively. Thus, all BBS loci were excluded. All regions except the region at 3q39 were eliminated since they were not IBD. Thus, the only candidate disease region was at 3q29, between 193,587,435bp (rs2630239) and 198,159,313 (rs1147240), a 4,629,472 bp region yielding a LOD score of 3.69.

Finally, linkage analysis was performed separately for the two branches of the family including all patients but not unaffected individuals in order not to disregard regions possibly harboring candidate variants with reduced penetrance. As using the whole pedigree would have exceeded the capacity of the software (Allegro), LOD scores were added for the two branches, and only one region >1-Mb and with high LOD score (> 2) was detected (Figure 5.19, Figure 5.21). One unaffected brother shared the homozygous genotype in the region (Table 5.11). For this region at 3q29 the only IBD region, linkage analysis was performed separately but including all members of the branches. This region is 5.7-Mb and yielded a maximal cumulative LOD score of 5.65 (3.73 in branch A and 1.92 in branch B) (Figure 5.22).

Further linkage analysis was performed to finalize the candidate regions. The regions and their LOD scores were re-evaluated, and the ones with LOD scores >3 cumulatively for branch A and B were determined. Fine mapping with no spacing and 10 marker sets was performed (Figure 5.23). Regions with LOD scores <3 were eliminated (Table 5.11).

Table 5.11. Maximal shared IBD homozygosity regions at loci yielding relatively high LOD scores.

| **Chr** | **Nucleotide (SNP)** | | **Size (bp)** | **Total LOD score** | **Fine Mapping** | **Homozygosity shared by*** |
|---|---|---|---|---|---|---|
| | **Start** | **End** | | | | |
| 2 | 17557962 (rs1401705) | 18058862 (rs13395300) | 500,900 | 2.0 + 1.8 | 2.4 + 1.8 | 7 A, 3 UA |

*A, affected; UA, unaffected

Table 5.11. Maximal shared IBD homozygosity regions at loci yielding relatively high LOD scores (cont.).

| Chr | Nucleotide (SNP) | | Size (bp) | Total LOD score | Fine Mapping | Homozygosity shared by* |
|---|---|---|---|---|---|---|
| | Start | End | | | | |
| 2 | 18693427 (rs7606752) | 18973451 (rs11892969) | 280,024 | 2.0 + 1.8 | 2.4 + 1.8 | 7 A, 3 UA |
| 3 | 193587435 (rs2630239) | 198159313 (rs1147240) | 5,735,582 | 3.5 + 1.8 | 3.6 + 1.8 | 7 A |
| XY | 155020688 (rs28814596) | 155236747 (rs2981828) | 216,059 | 3.5 + -3.0 | 0.2 + 0.6 | 4 A |

*A, affected; UA, unaffected

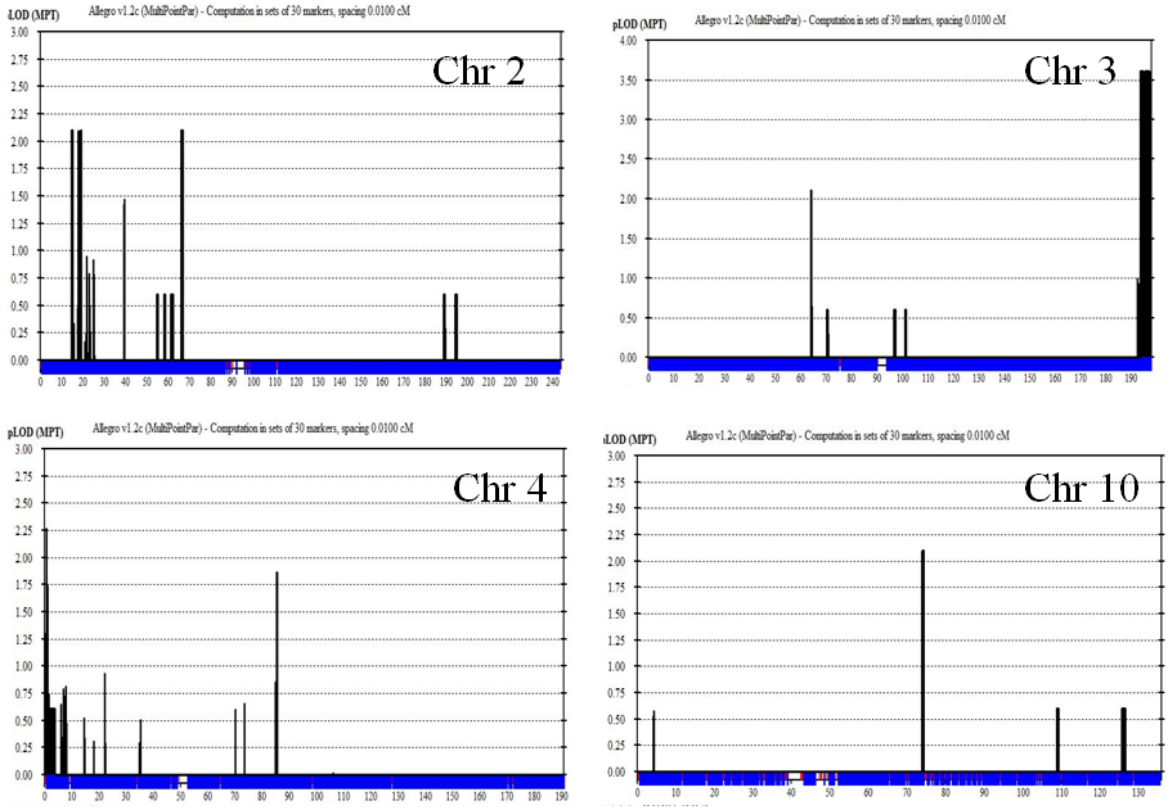



Figure 5.19. Multipoint LOD score graphs for chromosomes yielding LOD scores >2 for branch A of the family.

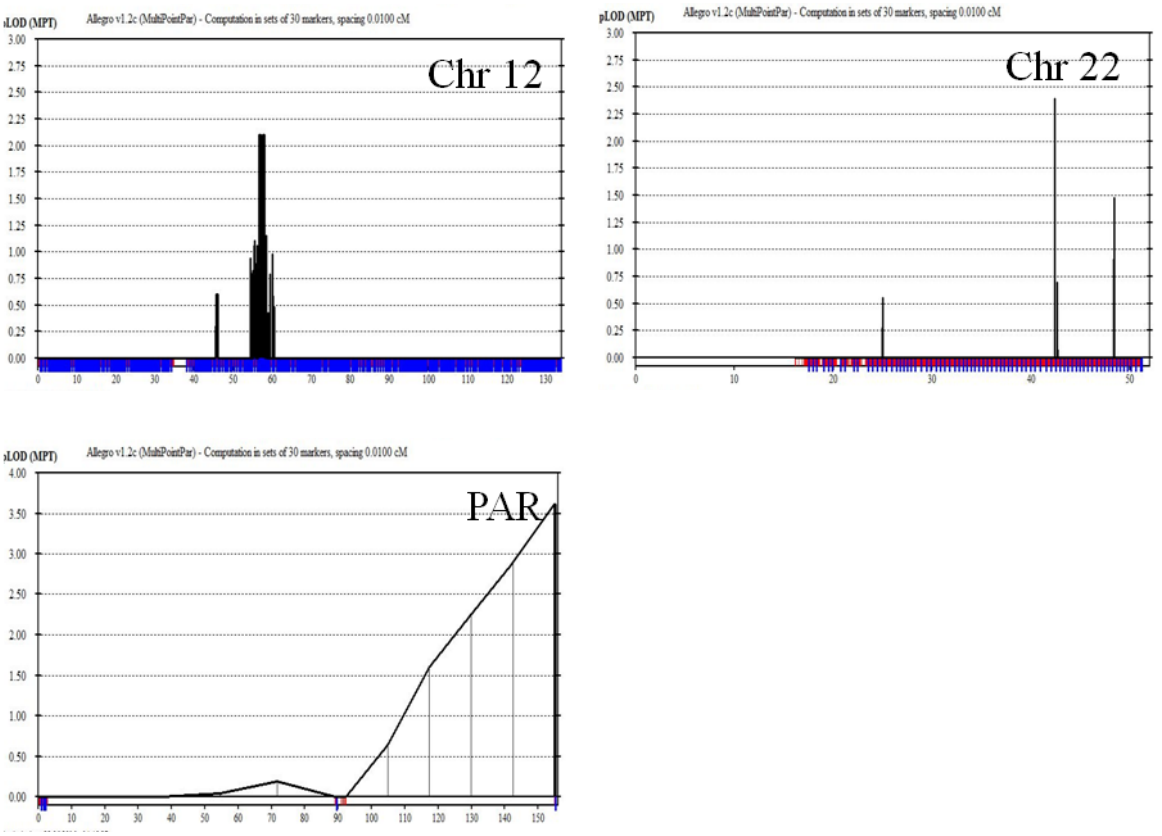



Figure 5.20. Multipoint LOD score graphs for chromosomes yielding LOD scores >2 for branch A of the family.

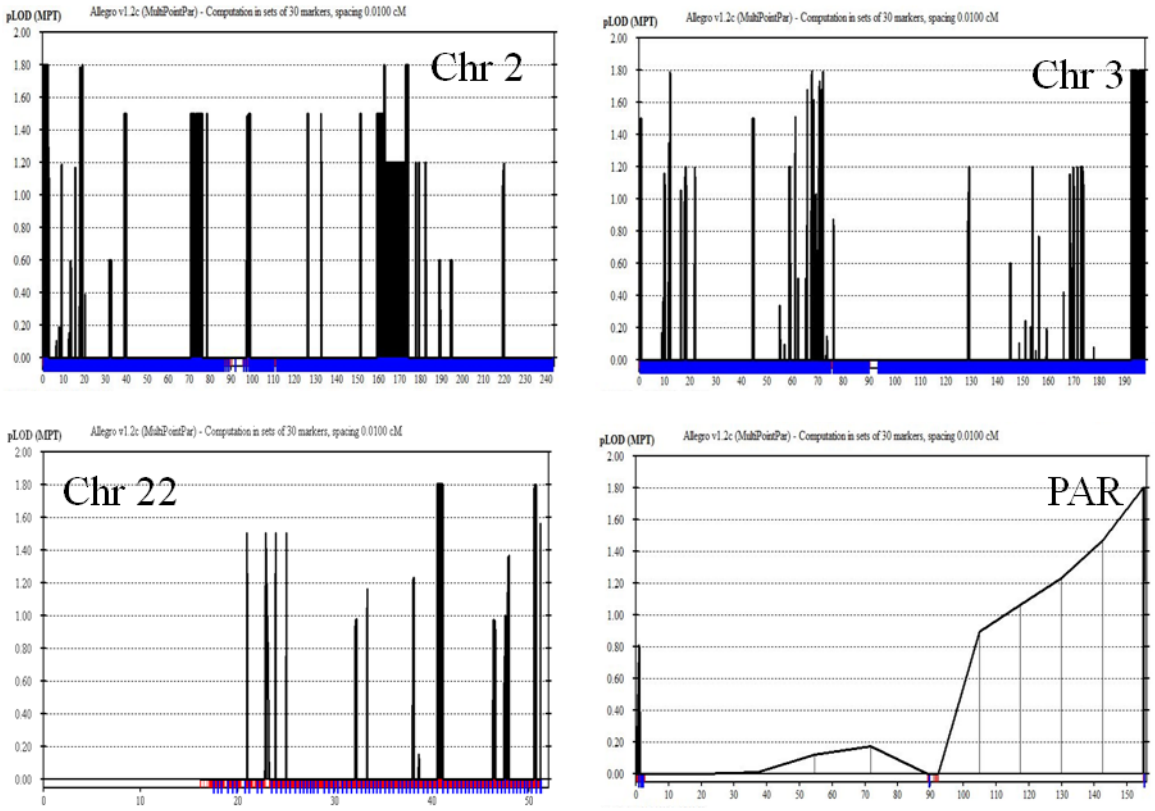



Figure 5.21. Multipoint LOD score graphs for branch B of the family for chromosomes with high LOD scores (>2) regions shared by branch A.

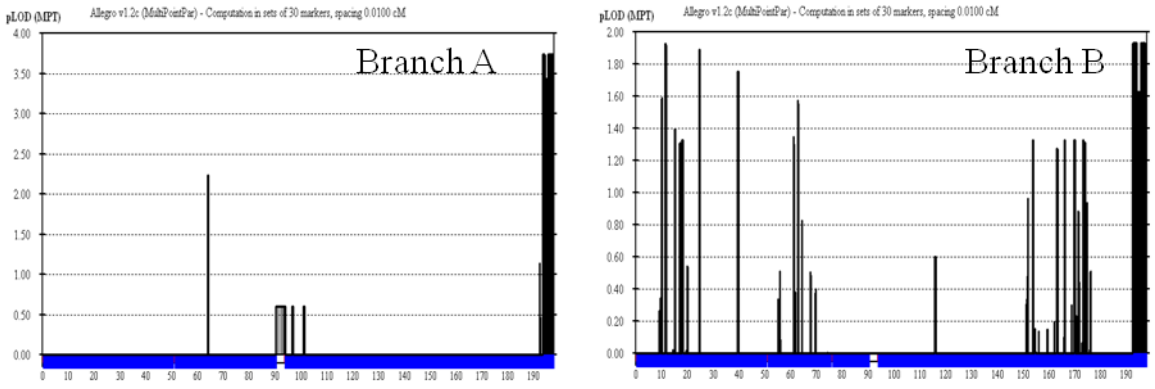
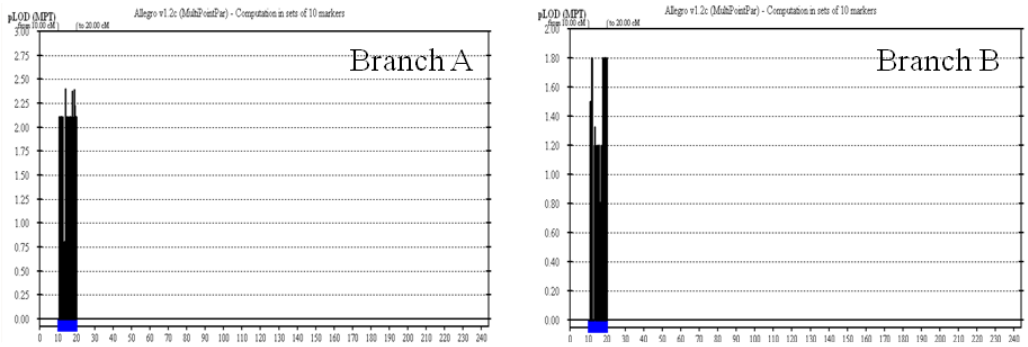
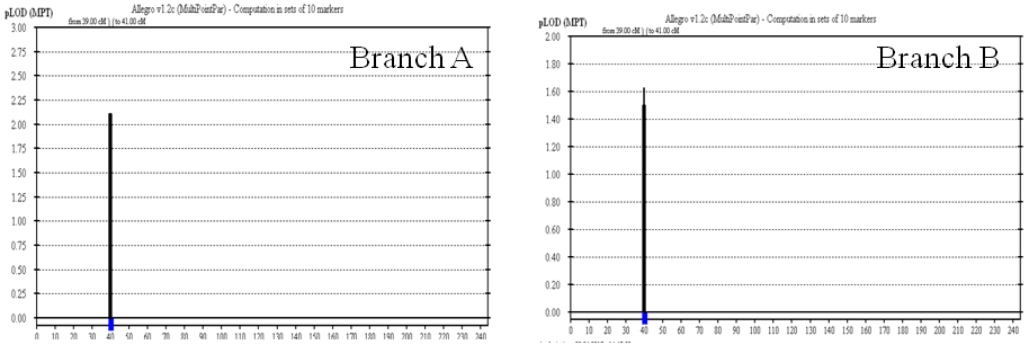


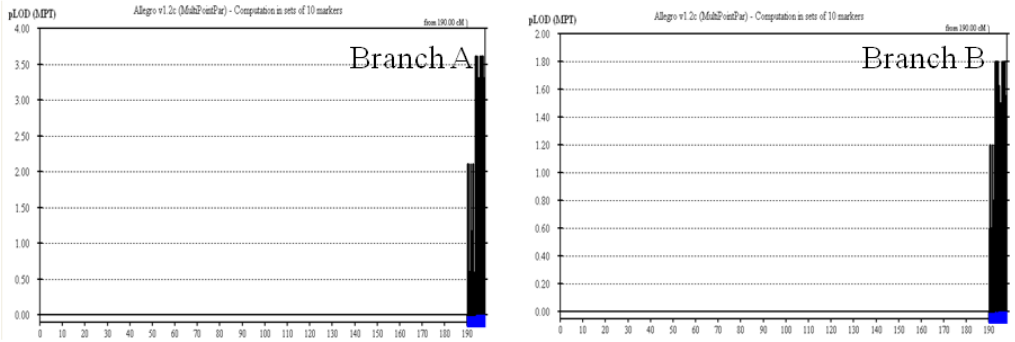

Figure 5.22. Final multipoint LOD score graphs for chromosome 3 for the two branches of the family. Genotype data of all participants were included.



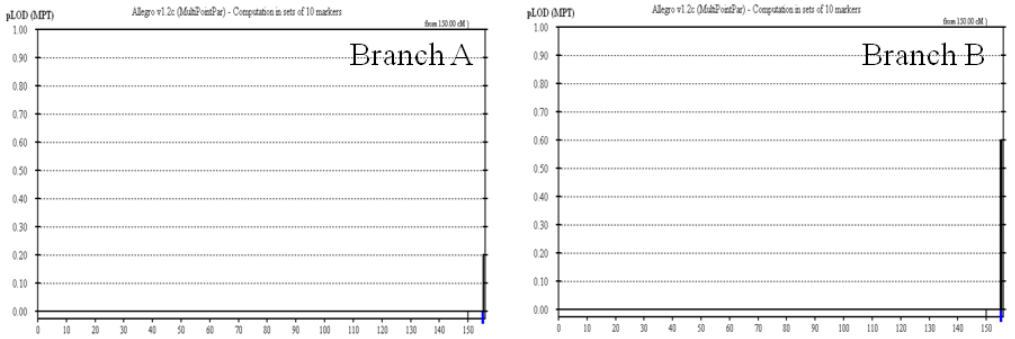

Chromosome 2, fine mapping at interval 10-20 Mb




Chromosome 2, fine mapping at interval 39-41 Mb




Chromosome 3, fine mapping at interval 190 Mb-ter




Chromosome XY, fine mapping at interval 150 Mb-ter

Figure 5.23. Multipoint LOD score graphs for fine mapping. It was performed for chromosomal intervals (plus 1-Mb flanking regions) yielding high (>3) cumulative LOD scores listed in Table 5.11, using all markers.

Linkage analysis was performed assuming a dominant inheritance model and 80% penetrance with genotype data of branch A only, as the whole pedigree exceeded the capacity of Allegro. The maximal LOD score was 1.6 for only regions at 8p23.1 and 12p11.2. For the chromosomes of those regions, linkage analysis was performed for branch B also. The results were negative LOD scores, eliminating both regions.

No deletion or duplication that is common to patients was found by deletion-duplication analysis.

### 5.5.2. Exome Sequence Analysis

DNA sample of 503 was subjected to exome sequencing. A sample from the other branch (509) was also subjected to exome sequencing later. In the two exome files, we investigated with priority the homozygous rare (MAF <0.01) and novel exonic and splicing variants that could potentially affect protein function and were within the identified gene region (Table 5.12).

The homozygous variant in *CEP19* is the best candidate to underlie the disease because the gene encodes a ciliary protein, as other BBS genes do, and is associated with morbid obesity. Since the variant causes truncation of the protein (deletion of 103 of the total 167 amino acids, 61%), it is deduced to be deleterious to protein function. It is not listed in the TEVD, 1000 Genomes and the ExAC data base, containing many Pakistani samples.

In *GLI1*, which was in another region of homozygosity shared by some of the affected individuals, there was a missense homozygous variant (c.820G>C; p.Gly274Arg) predicted as disease causing. Since there are several modifiers in BBS, we thought that this variant could be a modifier allele.

Table 5.12. Possibly damaging variants in the shared homozygosity regions yielding high LOD scores presented in Table 5.11.

| Chr | Start | Ref | Alt | Depth | | Gene | Freq* | dbSNP | Change | Mutation Taster | SIFT | Poly-Phen2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Tot | Alt | | | | | | | |
| In the exome file of individual 503 | | | | | | | | | | | | |
| 3 | 195506302 | G | T | 4 | 4 | *MUC4* | 0.0039 | - | Nonsyn | Polym | Del | Benign |
| 3 | 196434731 | - | T | 103 | 92 | *CEP19* | - | - | Stopgain | Disease causing | NA | NA |
| In the exome file of individual 509 | | | | | | | | | | | | |
| 3 | 195506302 | G | T | 136 | 129 | *MUC4* | 0.0039 | - | Nonsyn | Polym | Del | Benign |
| 3 | 195510653 | A | C | 18 | 14 | *MUC4* | 0.0002 | - | Nonsyn | Polym | Del | Benign |
| 3 | 196434731 | - | T | 51 | 50 | *CEP19* | - | - | Stopgain | Disease causing | NA | NA |

* Frequency in ExAC South Asian samples.

NA, not applicable; P, probably; Polym, polymorphism; Del, deleterious.

I also searched for any candidate variants in the 23 BBS related genes with MAF <0.05. In total 25 exonic variants in known BBS genes in the two exome files were found, and I considered the 3 with frequencies <0.05 (Table A.1). They were found in different combinations in BBS subjects and unaffected relatives. One is the rare splicing variant *CCDC28B* c.330C>T (formerly c.430C>T; p.Phe110Phe; rs41263993) (NM_024296); in the heterozygous state it was reported as a modifier of BBS1 (Badano *et al.*, 2006; Bin *et al.*, 2009). Another is the very rare missense variant *MKKS/BBS6* c.1015A>G (p.Ile339Val; rs137853909) (NM_018848) also reported in the heterozygous state as a modifier (Hichri *et al.*, 2005; Slavotinek *et al.*, 2002). Variant *C8ORF37* c.533C>T (p.Ala178Val; rs375314973) (NM_177965) was the last rare BBS gene variant detected.

Variants in the regions obtained by dominant linkage analysis were also evaluated. The ones with MAF <0.01 and number of the reads with the variant allele over the number of total reads >0.25 were filtered. No candidate variant was found.

The two exome files did not list any possibly deleterious rare or novel variants in postaxial polydactyly genes, namely, *GLI3, ZRS/SHH, ZNF141*, *MIPOL1* and *GLI2.*

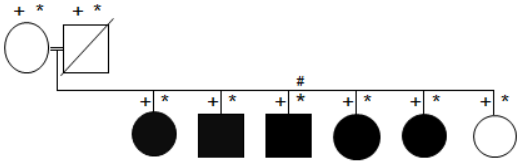
Lastly, we considered all rare (frequency <0.005) variants which were potentially deleterious to protein function in the two exome files separately as well as those shared by the two exome files for their possible contribution to the various findings unique to 503 and the aggressive behavior in 509. Two of the mutations shared by the exome files are heterozygous and in genes reported as susceptibility genes. *HTR2B* [MIM: 601122] nonsense c.738A>T (p.Lys246*; rs564438018) could be associated with aggressive behavior in individual 509; ExAC reports the highest frequency of 0.00163 in South Asians. Another nonsense mutation c.274C>T (pQ20*; rs79874540), was reported to predispose to impulsivity in Finns but not Japanese (Bevilacqua *et al.*, 2010; Tsuchimine *et al.*, 2013). Its frequency is highest in Finns (0.01482). Novel c.1966A>T (p.Ile656Phe) in *CACNA1F* [MIM: 300110] could have contributed to myopia in individual 503, as the gene is a susceptibility gene for myopia (Sun *et al.*, 2015).

### 5.5.3. Validation of the Variants and Population Screening

All six candidate variants were validated by Sanger sequencing. HRM analysis was attempted for only *CEP19* but could be optimized. In total 65 control samples from Pakistani population were screened by SSCP analysis for the *CEP19* variant, and none carried it. At that time ExAC database which contains many Pakistani exomes in the South Asian samples was made public. It did not list the variant, and hence population screening was not continued.

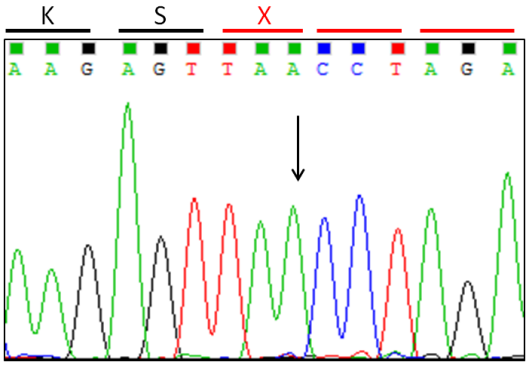
Family members were screened by SSCP analysis for the variants in *CEP19*, *GLI1*, *CCDC28B*, *MKKS*, *C8ORF37* and *TMEM67*. *CEP19* variant segregates with the trait. *GLI1*, *CCDC28*, *MKKS*, *C8ORF37* and *TMEM67* variants were not carried by all patients, as assessed by SSCP (Figure 5.24).

For each six variant 3-5 samples were validated by Sanger sequencing (Figure 5.25, Figure 5.26, Figure 5.27, Figure 5.28, Figure 5.29 and Figure 5.30). Genotype data for the six variants for family members are shown on the pedigree below.

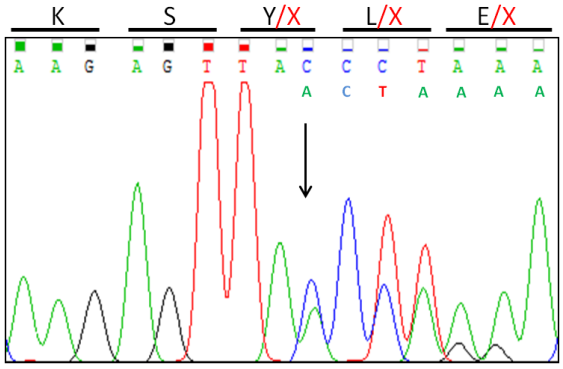


| Variants | 401 | 403 | 501 | 502 | 503 | 505 | 507 | 506 | 404 | 402 | 508 | 509 | 510 | 512 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *CEP19* c.194_195insA | +/- | +/- | +/+ | +/+ | +/+ | +/+ | +/+ | -/- | +/- | +/- | +/+ | +/+ | +/- | +/- |
| *GLI1* c.820G>C | +/- | +/- | +/+ | +/- | +/+ | +/+ | +/+ | -/- | +/- | -/- | -/- | +/- | +/- | -/- |
| *CCDC28B* c.330C>T | +/- | +/- | +/- | +/+ | +/- | +/- | -/- | +/- | +/- | -/- | +/- | -/- | +/- | -/- |
| *MKKS/BBS6* c.1015A>C | +/- | +/- | +/- | -/- | +/- | +/- | +/- | +/- | -/- | +/- | -/- | +/- | -/- | +/- |
| *C8ORF37* c.533C>T | -/- | +/- | +/- | +/- | +/- | -/- | +/- | +/- | -/- | -/- | -/- | -/- | -/- | -/- |
| *TMEM67* c.2397T>C | -/- | +/+ | +/- | +/- | +/- | +/- | +/- | +/- | +/- | -/- | -/- | +/- | +/- | -/- |

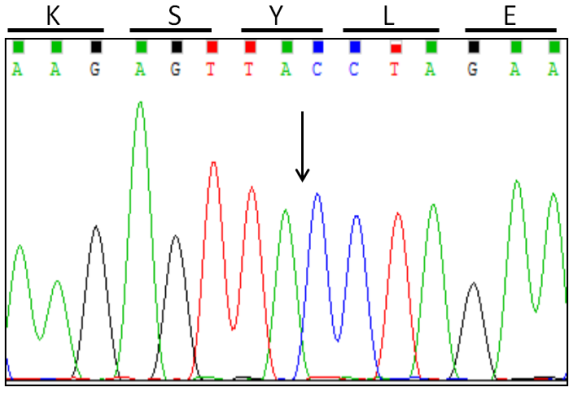
Figure 5.24. Pedigree of the family with genotypes for six variants.*Available for genetic studies, +Available for mutation analysis, #Exome sequenced individuals.




Homozygous 503 (Affected)

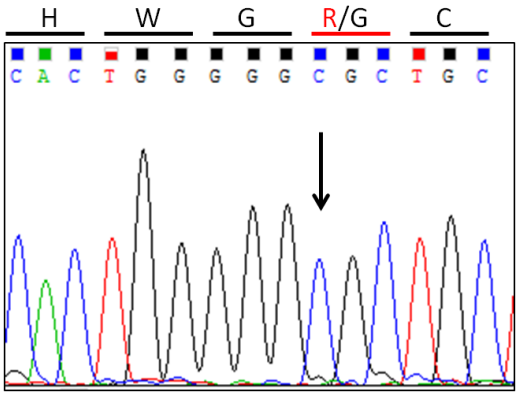



Heterozygous 401 (Unaffected)

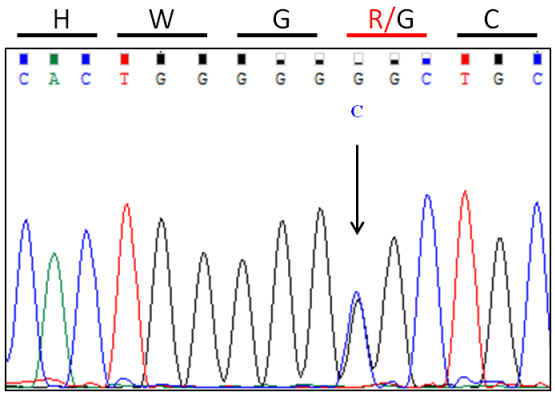



Reference 506 (Unaffected)

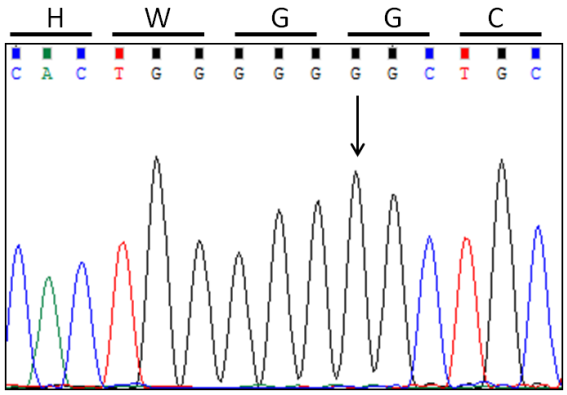
Figure 5.25. Electrophoretograms showing variant *CEP19* c.194_195insA, p.Tyr65*




Homozygous 503 (Affected)

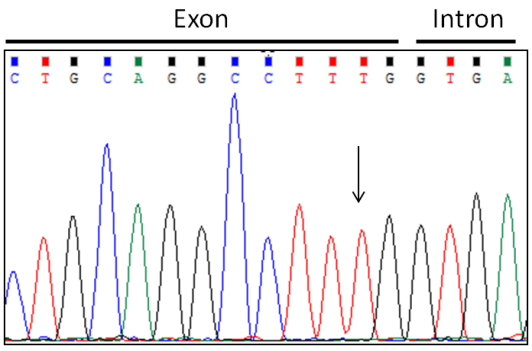



Heterozygous 509 (Affected)

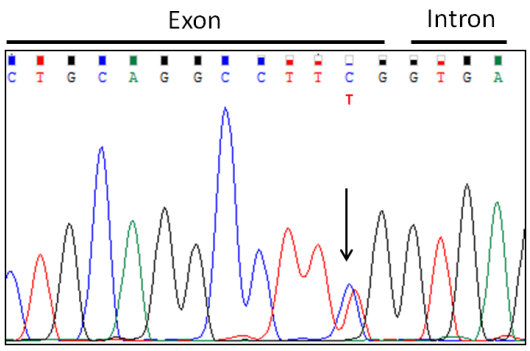



Reference 508 (Unaffected)

Figure 5.26. Electrophoretograms showing *GLI1* c.820G>C, p.Gly274Arg

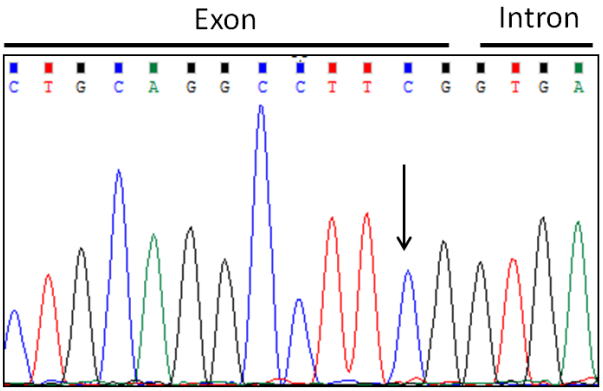



Homozygous 502 (Affected)



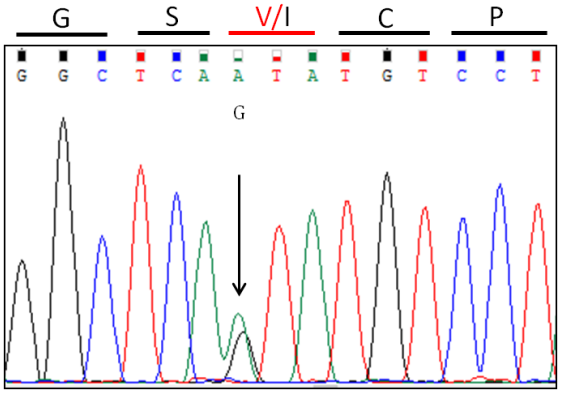

Heterozygous 503 (Affected)

Figure 5.26. Electrophoretograms showing CCDC28B c.330C>T, p.Phe110Phe.

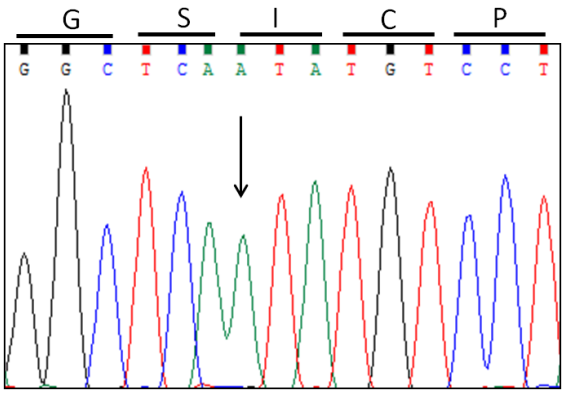



Reference 507 (Affected)

Figure 5.27 Electrophoretograms showing *CCDC28B* c.330C>T, p.Phe110Phe (cont.).

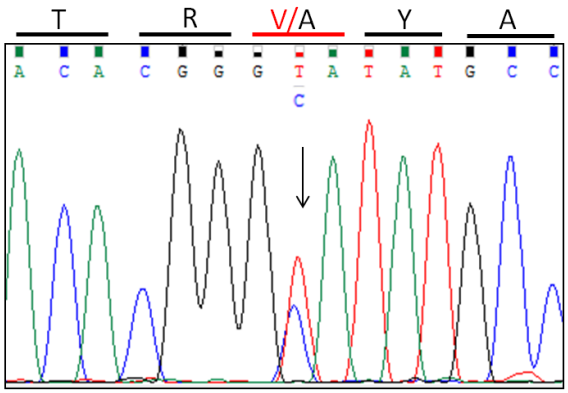



Heterozygous 503 (Affected)

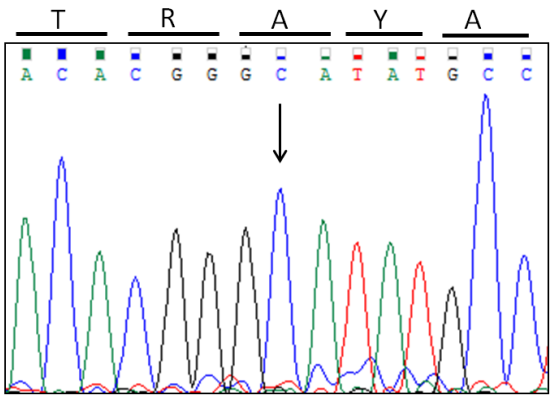



Reference 510 (Unaffected)

Figure 5.28. Electrophoretograms showing variant *MKKS* c.1015A>G, p.Ile339Val.




Heterozygous 503 (Affected)




References 509 (Affected)

Figure 5.29. Electrophoretograms showing variant *C8ORF37* c.533C>T, p.Ala178Val

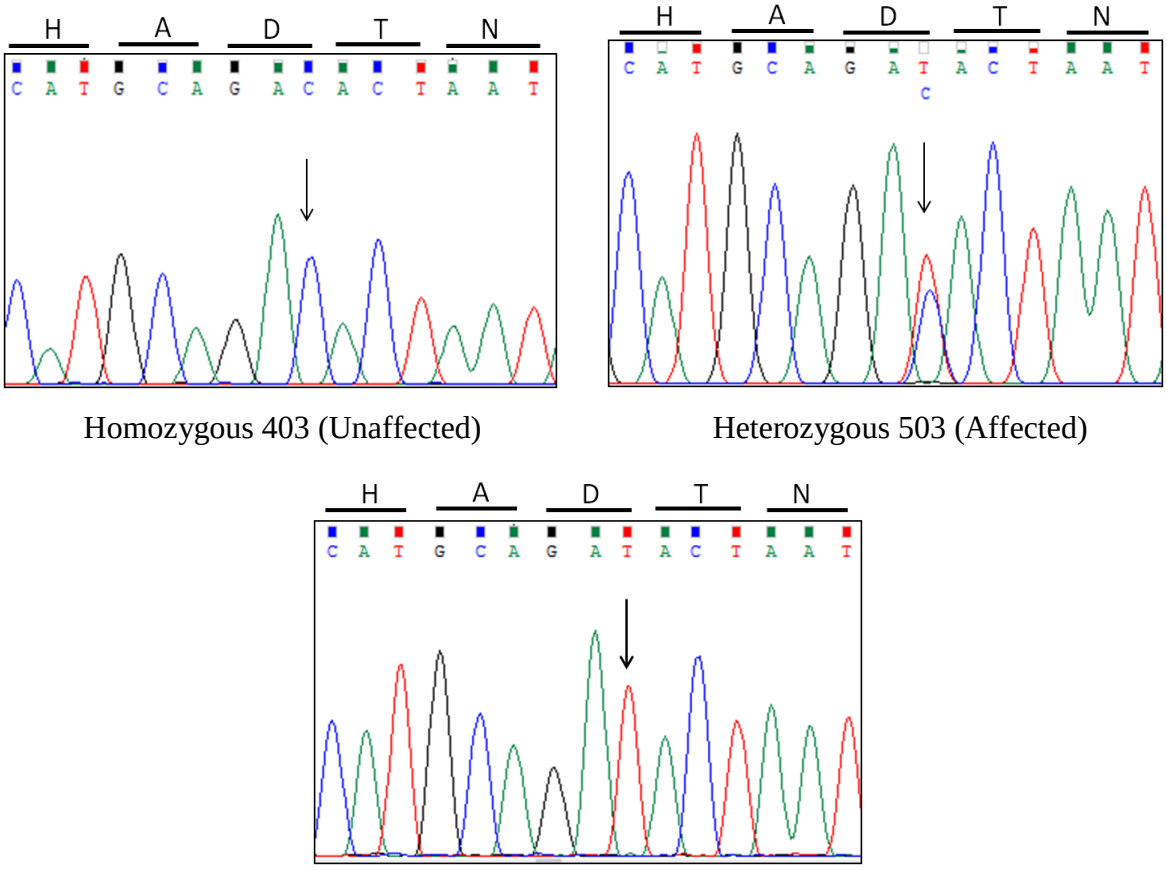



Figure 5.30. Electrophoretograms showing variant *TMEM67* c.2397T>C, p.Asp799Asp

## 5.6. Pleuroparanchymal Fibroelastosis (PPFE) Families

### 5.6.1. Linkage Analysis

For PPFE1 family, the linkage analysis with one unaffected and 3 affected siblings yielded nine regions with LOD scores >2 and sizes >300kb, and the maximal LOD score was 2.52 (Figure 5.31). Those nine regions were evaluated in Excel, and two of them were eliminated since genotypes of not all patients were homozygous in those regions. The remaining regions are listed with maximal regions of homozygosity in Table 5.13.

Additionally, we were told that in 'unaffected sister' some lesion had begun to arise. Thus, linkage analysis was performed with genotype data of four siblings, assuming all of them affected. Two regions with LOD scores >2 were obtained. However, later we were told that that sister was not affected, and thus initial linkage analysis results were considered.

Table 5.13. Homozygous regions detected in the initial linkage analysis of PPFE1 family.

| Chr | Maximal | | SNP (start-end) | | Size (bp) | LOD score |
|---|---|---|---|---|---|---|
| 10 | 82736647 | 95124673 | rs4511237 | rs787640 | 12388026 | 2.52 |
| 17 | 30033822 | 32505722 | rs7222438 | rs7218501 | 2471900 | 2.52 |
| 7 | 147759461 | 149795559 | rs10485844 | rs2628988 | 2036098 | 2.52 |
| 17 | 28550814 | 29738601 | rs2020936 | rs178843 | 1187787 | 2.52 |
| 20 | 60339324 | 61160251 | rs6142852 | rs6062235 | 820927 | 2.22 |
| 3 | 56151970 | 56556790 | rs2200432 | rs1491161 | 404820 | 2.50 |
| 3 | 3759119 | 4108261 | rs769781 | rs666412 | 349142 | 2.40 |

For PPFE2 family, multipoint linkage analysis was performed using SNP data of the three siblings. All chromosomes yielded many regions with a maximal LOD score of 2.40. All regions with sizes >200 kb with a maximal LOD score of 2.40 were investigated for possible IBD and listed (nearly 120). The regions with size >1Mb is listed in Table 5.14.

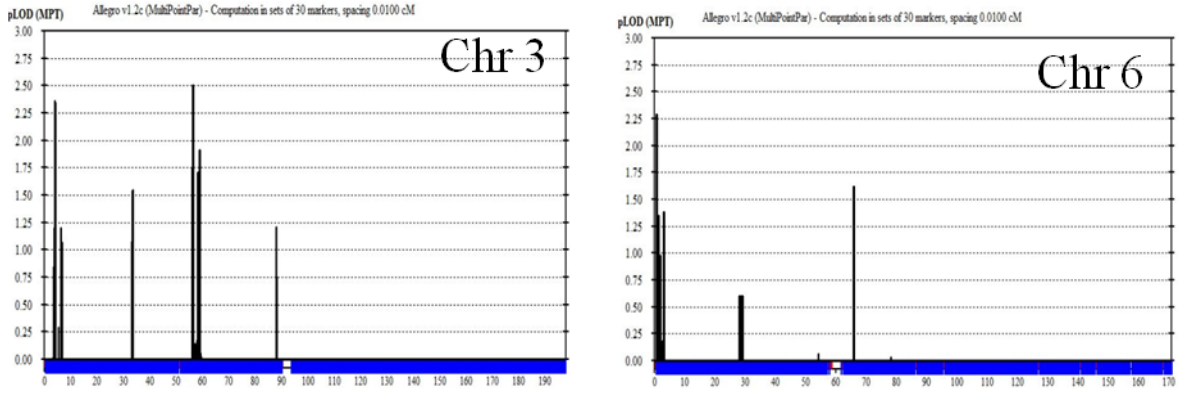



Figure 5.31. Multipoint LOD score graphs for PPFE1 family with three affected and one unaffected sibling. Only chromosomes with maximal LOD scores >2 are presented.

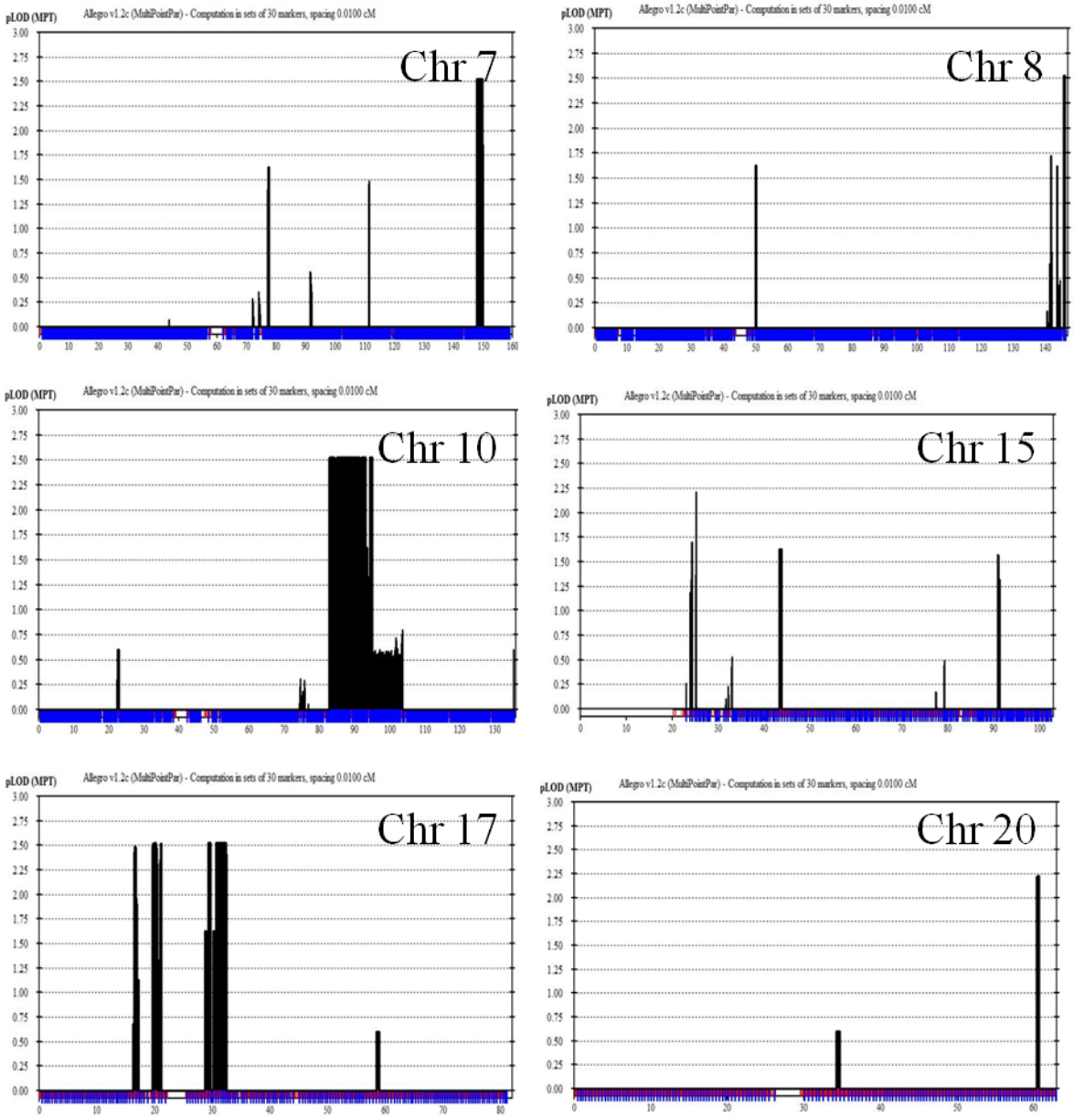



Figure 5.30. Multipoint LOD score graphs for PPFE1 family with three affected and one unaffected sibling. Only chromosomes with maximal LOD scores >2 are presented (cont.).

Table 5.14. Homozygous regions with sizes >1 Mb detected by ocular investigation in Excel of PPFE2 genotypes.

| Chr | Maximal | | SNP (start-end) | | Size (bp) |
|---|---|---|---|---|---|
| 12 | 70944620 | 83595716 | rs2584026 | rs983059 | 12651096 |
| 12 | 25192784 | 31288955 | rs2101302 | rs11051266 | 6096171 |

Table 5.14. Homozygous regions with sizes >1 Mb detected by ocular investigation in Excel of PPFE2 genotypes (cont.).

| Chr | Maximal | | SNP (start-end) | | Size (bp) |
|---|---|---|---|---|---|
| 5 | 129659729 | 131336287 | rs2108425 | rs440970 | 1676558 |
| 3 | 95116949 | 96681220 | rs501118 | rs9824190 | 1564271 |
| 1 | 49216371 | 50598368 | rs11586980 | rs4307603 | 1381997 |
| 7 | 113029219 | 114310053 | rs1524443 | rs12705971 | 1280834 |
| 2 | 96057011 | 97139307 | rs2320433 | rs6711452 | 1082296 |
| 18 | 64232938 | 65286754 | rs11875080 | rs9636020 | 1053816 |
| 3 | 44134579 | 45154357 | rs6797157 | rs12494480 | 1019778 |

For PPFE3 family linkage analysis was not performed since none of the family members was subjected to SNP genotyping.

### 5.6.2. Coverage Analysis

By CovBed analysis, whether any exons of the genes in the candidate regions had not been targeted or not sequenced in the exome sequencing process was investigated for all three families. Coverage data of each exome file were compared with those of another file generated in the same batch. I found that in the candidate regions all exons had been sufficiently covered.

### 5.6.3. Exome Sequence Analysis and Validation of Candidate Variants

In PPFE1 family, DNA sample of an affected sibling (402) was subjected to exome sequencing. Variants in the candidate regions obtained by the second linkage analysis (assuming eldest sister as affected were evaluated first. Among the five candidate variants priority was given to the *CC2D2B* variant. *CC2D2B* c.903T>G (p.Asn301Lys) was chosen as the strongest candidate since it is rare and predicted as damaging. In addition, its paralogue *CC2D2A* is associated with Meckel and Joubert syndromes, which include

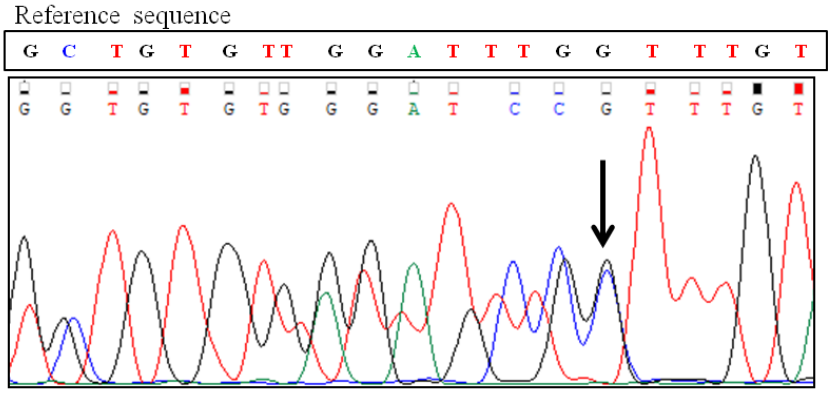
respiratory system problems. After the eldest sister was evaluated as unaffected, the region was not a candidate anymore, and thus this variant was disregarded.

All rare and possibly homozygous exonic and splicing variants in the regions presented in Table 5.13 and in others (<1Mb) with LOD scores >2 were evaluated. In total there were four candidate variants, all exonic (Table 5.15). Affected sister, unaffected mother and a control sample were tested for *FAM35A* (NM_019054; c.540_541insCC, p. Val181Profs*10). Mother and even affected sister were found heterozygous. Although it seems complicated since sequencing was performed with reverse primer and the reverse-complement version was evaluated, it can be observed on electrophoretograms that each base had shifted by two positions (Figure 5.32). *TNKS2* (NM_025235; c.1146A>T, p.Ile382Ile) was validated in the exome-sequenced patient by Sanger sequencing, and all family members were screened by SSCP analysis. As shown on the pedigree in Figure 5.33, affected siblings were homozygous and unaffected individuals were heterozygous for the variant. Genes of the remaining two variants are paralogs and reside in a duplicated region (>2500 bp). Thus, I could not design primers to amplify specifically the region of each variant. Except for *FAM22A*, all other genes are expressed in lung.

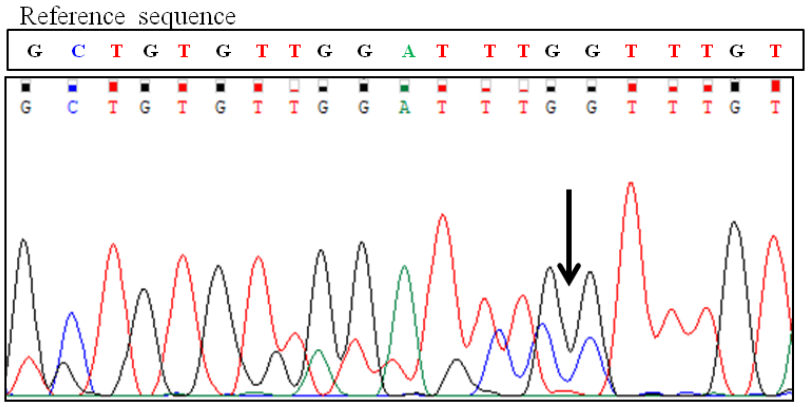
Table 5.15. Candidate variants in regions listed in Table 5.13. Only exonic and splicing variants were selected. All variants are on chromosome 10 and not found in the in-lab exome files. PolyPhen-2 could not be applied since it is suitable for only missense variants.

| **Start** | **End** | **Ref** | **Alt** | **Gene** | **Exonic** | **ExAC Highest** | **Status** | **IGV** | **Prediction Algorithms** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | **Mutation Taster** | **SIFT** | **Provean** |
| 88911651 | 88911651 | - | CC | *FAM35A* | Frm ins | - | hom | 45/45 | Disease causing | Causing NMD | NA |
| 88991897 | 88991898 | CT | - | *FAM22A* | Frm del | - | het | 22/32 | Disease causing | Causing NMD | NA |
| 89124115 | 89124116 | CT | - | *FAM22D* | Frm del | - | het | 15/40 | Disease causing | Causing NMD | NA |
| 93590721 | 93590721 | A | T | *TNKS2* | Syn | 0.0034 (SAS) | hom | 56/56 | Disease causing | Tolerated | Neutral |

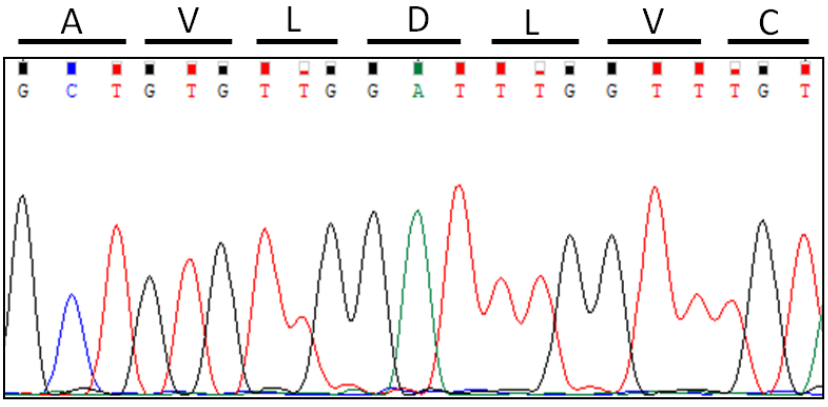
Frm ins, Frameshift insertion; Syn, Synonymous; IGV, Alt depth/total depth; NMD, nonsense mediated decay.




Heterozygous Affected Sister (402)




Heterozygous Mother




Reference Sequence

Figure 5.32. Electrophoretograms showing mutation *FAM35A* c.540_541insCC (p. Val181Profs*10; NM_019054). Sequencing was performed with reverse primer and reverse complement is displayed.

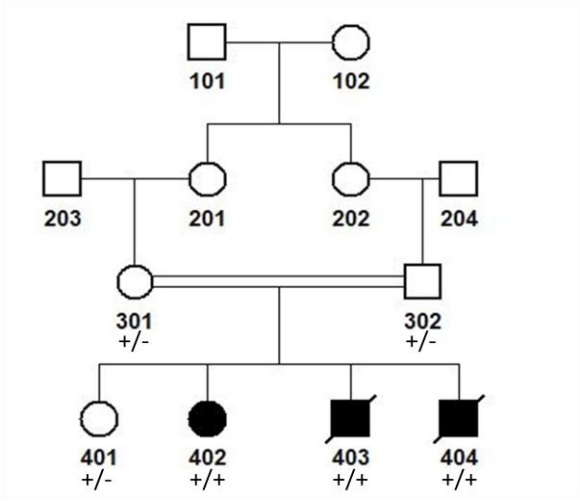



Figure 5.33. *TNKS2* c.1146A>T genotypes for PPFE1 family members.

For PPFE2 family, all variants in the candidate regions were listed, the exonic and splicing variants with MAF <0.01 or uncertain frequency were selected. Those present in in-lab exomes were eliminated. No candidate variant was found.

For PPFE3 family we had a blood sample of only one affected brother. His DNA sample was subjected to exome sequencing. All variants found in candidate genes in PPFE1 and in idiopathic pulmonary fibrosis (IPF) related genes (*ABCA3, ATP11A, DKC1, DPP9, DSP, EIF2AK4, FAM111B, FAM13A, MARS, MUC5B, OBFC1, SFTPA1, SFTPA2, SFTPC* and *TOLLIP*) were listed, and the ones with MAF <0.01 were evaluated. No candidate variant was found; the regions for all such variants were heterozygous. In evaluating for compound heterozygosity, two heterozygous variants with MAF <0.01 were found in *ATP11A*, but both were intronic.

The genes with rare variants in both PPFE1 and PPFE3 families were listed. In those 61 genes variants with MAF <0.01 were evaluated. There were no genes with candidate variants in both families.

We investigated all variants in the four telomere-related genes associated with PPFE in the exome files of patients from the three families. Only one heterozygous nonsynonymous variant (NM_002582: c.G1690A, p.V564I) was found, in *PARN* in family

PPFE1, which is predicted as benign by all prediction tools. No candidate variant was found in families PPFE2 or PPFE3.

In PPFE1, seven regions were considered as candidate regions (Table 5.13), and variants *FAM35A* c.540_541insCC and *TNKS2* c.1146A>T were detected in those regions. *TNKS2* was the strongest candidate. For PPFE2 although linkage analysis yielded many regions, no candidate variant was found. In PPFE3 since we do not have any data for the deceased affected siblings, only some variants were evaluated on exome data of a patient, and no candidate variant was detected.

### 5.6.4. Deletion Duplication Analysis

Deletion-duplication analysis was performed for all chromosomes and family members with SNP genotype data using the cnvPartition. No duplication or deletion common to patients was found in any of the families.

# 6. DISCUSSION

## 6.1. Isolated Intellectual Disability (IID) Family

Disease locus was mapped to 2p24.1-p23.3 (LOD score 3.25) by linkage analysis, and the single candidate variant *PTRHD1* c.155G>A (p.Cys52Tyr) was found to segregate with the disease in the family. The frequency of the variant in ExAC South Asian sample is 0.00006 and world-wide 0.000008. 52Cys is conserved across all vertebrates. Cysteine which is nonpolar and neutral is substituted with polar and neutral tyrosine, most probably affecting protein folding.

Structural model using SWISSMODEL was built for both the wild type and the mutant PTRHD1 by our collaborator Ute Woehlbier in Chili. Cys52Tyr amino acid change has an apparent impact on the predicted structure of the protein. PTRHD1 may play a role in synaptic translation, and its mutant form may lead to a problem there and consequently problems in learning and memory can arise (Hibaoui *et al.*, 2014).

Recently, an Iranian family afflicted with recessive ID and parkinsonism was reported with p.His53Tyr in *PTRHD1* (Khodadadi *et al.*, 2017). Also the same variant (c.155G>A) that we detected in the Pakistani family was reported in another Iranian family afflicted with parkinsonism as a second possible contributor to the disease (Jaberi *et al.*, 2016).

Although little is known about the relation of *PTRHD1* with ID, deletions at 2p23 locus that harbors *PTRHD1* was associated with several ID syndromes. The ubiquitin proteasome system plays a role in brain development and synaptic plasticity (Jarome *et al.*, 2013). The PTH2 domain, the family which PTRHD1 belongs to, is a ubiquitin-like (UBL) domain-binding protein in yeast that participates in the ubiquitin-proteasome pathway and suppresses ubiquitin-mediated degradation (Ishii *et al.*, 2006). In the light of this knowledge, we can claim that *PTRHD1* mutation could cause ID and Parkinsonism (Khodadadi *et al.*, 2017). Additionally, the variant in our family was also reported in

another family associated with early-onset Parkinsonism and cognitive dysfunction, but instead the *ADORA1* mutation in the family was claimed to be the causative mutation because it is located at the Parkinson's disease locus PARK16 and the encoded protein has role as an adenosine receptor in brain function and neuronal activity (Jaberi *et al.*, 2016). However, no deleterious variant was found in our patient's exome file. We concluded that *PTRHD1* is the causative gene for ID and either the symptoms of parkinsonism will develop much later in this family as compared to the reported families or parkinsonism is not an obligatory feature for PTRHD1 deficit. Other findings in our patients were early onset behavioural problems including attention deficit, seclusion, hyperactivity, apraxia of speech and stuttering.

A *PTRH2* mutation is reported in novel infantile-onset multisystem disease with ID, microcephaly, progressive ataxia and muscle weakness (Hu *et al.*, 2014). The amino acid sequences of *PTRH2* and *PTRHD1* were aligned and amino acid change Cys52Tyr in *PTRHD1* and the mutation Gln85Pro in *PTRH2* were found to be within two amino acids from each other. Network analysis was performed with integrated software Ingenuity by Ute Woehlbier, displaying indirect and direct interactions with PTRHD1. There is one direct interaction, with C6ORF89, and two indirect interactions, with MOV10 and SHMT2. The genes of these proteins have been related to ID disorders in different studies (Liu *et al.*, 2015). We hypothesize that if the three proteins in the network of PTRHD1 are related to ID, it is possible that PTRHD1 is also associated with ID as it might interact with these proteins or affects similar pathways. *PTRHD1* is expressed in many organs, and we found wide expression in adult brain.

The family was re-visited to evaluate patients for Parkinsonism by our Pakistani collaborators, who confirmed that the patients have early onset mental retardation but still no symptoms of Parkinsonism at the age of 38 for the eldest patient. The reported patients with *PTRHD1* mutation showed symptoms of the disease latest at age 27 years. Unlike the reported families, the family we present is sufficiently large to facilitate mapping of the disease gene and thus excluding any possible contribution of mutations at other loci to the phenotype.

Our findings confirm that *PTRHD1* mutation causes ID of variable severity but show that parkinsonism might not be evident till towards the end of the fourth decade, if it develops at all. Three of our patients are at ages 25 – 38 years. It indicates that it may not be causing Parkinsonism at least till the fifth decade. We propose that families afflicted with idiopathic ID especially with behavioural problems could benefit from testing for *PTRHD1* mutation. A manuscript was prepared and submitted to Journal of Medical Genetics as a short report. It was rejected for the reason that the patients are too young to develop Parkinsonism. However, later the clinicians stated that the patients are older than 30, and thus the manuscript will be submitted after revisions.

## 6.2. Intellectual Disability and Hypothyroidism (IDH) Family

In the beginning the information we had about the family was that there are three patients with ID, and linkage analysis was performed accordingly using all SNP genotype data. Four candidate regions yielded a maximal LOD score of 3.55. In those regions, the only candidate variant was novel *TPO* c.719A>G, p.Asp240Gly at 2p25.3. *Thyroid peroxidase (TPO)* is associated with hypothyroidism, and in some patients ID was also observed (Mittal *et al.*, 2016). Afterwards, we ascertained that all patients have hypothyroidism and just two patients have severe ID (311 and 325). Final version of the pedigree with TPO variant genotypes is presented in Figure 5.7.

Hypothyroidism is a systemic metabolic disorder in which less thyroid hormone is produced from the thyroid gland. The disease can be congenital or due to various other reasons such as an autoimmune disease, radiation treatment, medicine and surgical removal of thyroid gland. Congenital type can be caused by defects in thyroid hormone biosynthesis, and it is mostly inherited in a recessive manner (Grasberger and Refetoff, 2011). The most frequent causes of congenital hypothyroidism (CH) is mutations in *TPO* (Cangul *et al.*, 2013). TPO has a key role in thyroid hormone synthesis by oxidizing the iodide in thyroid cells (Park and Chatterjee, 2005). Total iodide organification defect (TIOD) is one of the causes for the increase in TSH (thyroid stimulating hormone)

indicating that hypothyroidism and the responsible gene for this defect is *TPO (*Bikker *et al.,* 1997*).*

After the recent visit and a clinical re-evaluation of the family, it was revealed that patient 322 that we had exome file for does not have ID. We have decided to search for a new gene which was responsible for ID in the family. Thus, linkage analysis was performed with two ID patients and two unaffected siblings. Many regions were found with a maximal LOD score of 2.95. These regions were evaluated in OMIM and no ID phenotype was found. Since we did not have a sufficient amount of DNA sample of an ID patient, we were unable to generate exome data for an ID patient.

Congenital hypothyroidism is a leading cause of intellectual disability, but it is mostly treatable with hormone replacement (Wheeler *et al.*, 2012). Thyroid hormones affect the development of central nervous system in fetus and in later stages of life (Wheeler *et al.*, 2011). A family afflicted with hypothyroidism and ID was reported with homozygous missense mutation in TPO (p.Arg412His) *(*Mittal *et al., 2016).* However, the case may be different in underdeveloped countries where the opportunities for new born testing could be limited and the disease may not be detected thus not preventing ID. In our study family, in addition to two patients with ID, three patients have developmental delay (322, 329 and 401), and among them two (322 and 329) are also slow learners and have reduced motor and cognitive skills.

We present a family with novel missense mutation p.Asp240Gly in *TPO* causing hypothyroidism and ID in variable severity. This study strengthens the role of *TPO* in hypothyroidism and suggests further phenotypic heterogeneity for congenital ID in this family. If we could have obtained a DNA sample for one of the ID patients, we could have investigated whether there was any other mutation leading to ID in the family.

## 6.3. Syndromic Intellectual Disability (SID) Family

In this Pakistani family afflicted with a severe syndrome manifesting with intellectual disability (ID), developmental delay and facial dysmorphism, multipoint linkage analysis yielded only one chromosomal region with a maximal LOD score 3.85 with shared homozygosity in the patients possibly due to identity by descent from a common ancestor. This approximately 2.2-Mb region at 15q15.2-21.1 between nucleotides 43,336,670 (rs1197547) and 45,095,902 (rs1288092) did not harbor any known genes possibly related to ID. At the locus only one candidate variant was identified, a homozygous missense c.170G>A (p.Cys57Tyr) in *PDIA3*.

The substitution of sulfur containing cysteine with hydrophobic tyrosine is expected to alter the protein structure, as cysteine can be involved in bonding which is important in maintaining the native 3-D structure of the protein. *PDIA3* encodes Endoplasmic reticulum resident protein 57 (ERp57), that promotes disulfide bonds in glycoproteins (Jessop *et al.*, 2007). It is a protein of the endoplasmic reticulum and interacts with other proteins to modulate the folding of newly synthesized proteins (Santana-Codina *et al.*, 2013). Impaired bone formation was reported in Pdia3 deficient mice (Wang *et al.*, 2014).

As the candidate *PDIA3* variant changes the third base in the exon, we investigated its potential to cause aberrant splicing. Both Mutation Taster and Splice finder predicted an alteration of the splice site. If the splice site is abolished, then exon 2 would be excluded, and 79 bases would be missing due to the frameshift. The protein is deduced to be truncated after the synthesis of 20 nonnative amino acids. However, the protein was found full-length and the same amount of that in wild type cells in the western analysis performed in knockin mouse by our colleagues.

The possible effect of the variant on protein was investigated. ERp57 has 505 amino acids and four TRX-like domains which are a-a' and b-b' that are thioredoxin active and inactive sites, respectively (Ferrari and Soling, 1999). Within "a" and "a'" domain, PDIA3 has thioredoxin-like boxes (Cys-Gly-His-Cys motifs) like in many other oxireductases (Hettinghouse *et al.*, 2017). This motif spans residues 57-60 and 406-409. The first motif is

totally conserved among species having the motif at that region and the second motif is totally conserved among species (HomoloGene). In our study family, the causative variant causes the substitution of the first cysteine of the first motif with tyrosine, disrupting the highly conserved CGHC motif.

Several studies performed by our collaborators in Chile to show the pathogenic properties of PDIA3$^{Cys57Tyr}$. It was shown that disruption of Cys57 residue caused aggregation of the protein in the cell culture. Thus, the mutant protein was overexpressed in zebrafish and mice. In the knockin axonal disorganization was observed in zebrafish embryos. Behavioral analysis was performed on knockin mouse and intellectual disability was observed.

### 6.4. Spinocerebellar Ataxia (SCA) Family

In this family afflicted with spinocerebellar ataxia, the disease was mapped to 11p11.2-q13.2 with a maximal multipoint LOD score of 4.52. The only candidate variant was splice mutation c.6375-1G>C in *SPTBN2*. The family was screened by SSCP analysis, and the mutation was found to segregate with the disease. Six heterozygous *SPTBN2* mutations have been reported to cause SCA5 (MIM 600224), a late onset, progressive, dominant, pure cerebellar ataxia (Elsayed *et al.*, 2014; Jayadev and Bird, 2013). Two homozygous *SPTBN2* mutations were reported in more severe childhood-onset spinocerebellar ataxia type 14 (SCAR14; MIM605361). The features of the disease in the first family were wide-based gait, tremor, developmental delay and pyramidal signs and in the second family developmental delay, nystagmus, convergent squint and cognitive impairment (Elsayed *et al.*, 2014; Lise *et al.*, 2012). The family we studied is afflicted with recessive complex ataxia different than in those in the two previously reported families.

In the presented family and the two reported families with homozygous *SPTBN2* mutation (Elsayed *et al.*, 2014; Lise *et al.*, 2012) the common clinical features are early onset ataxia, broad-based gait, dysdiadochokinesia, developmental delay and cerebellar

atrophy. However our study family has some unique features such as high arched palate, absence of nystagmus, climacophobia, and certain behavioral abnormalities. Climacophobia were unlikely to be related to *SPTBN2* mutation since the behavioral abnormalities might be unrelated to or caused by the intellectual disturbance. High arched palate has not been reported in other families. Two associated genes (*ASXL3* and *IRF4*) with high arched palate were found by literature search, but in patient's exome data no rare pathogenic variant was found in these two genes. Thus, we came up with that the high arched palate could be due to a defect in an unknown gene and climacophobia might be caused by the intellectual disturbance.

It was concluded that the intronic-splicing mutation that we identified in a gene already associated with complex ataxia is pathogenic since the last nucleotide G in introns and the upstream A are highly conserved. Thus, the causative mutation underlying the disease in the presented family is this variant which was also predicted as disease causing by computational algorithms. It is predicted to either lead to nonsense-mediated decay of the mutant mRNA (Nagy and Maquat, 1998) or exon 32 could be excluded. Exon 32 resides between the ankyrin binding site and the PH domain. If exon 32 is excluded, it would cause the premature termination after the synthesis of 137 non-native amino acids, and thereby the deletion of the terminal 265 native amino acids. The deletion of just the terminal PH domain caused by the mutation is deduced to not prevent tetramerization, but the mutant tetramer would very likely be completely functionless and have no ability to bind the target membrane proteins. Since data bases report expression of the gene in blood, cDNA was obtained from patient blood to analyze the mRNA. We attempted to amplify *SPTBN2* transcripts in patient and heterozygous sibling blood samples several times, but we could not obtain any product specific to the gene.

The five reported dominant *SPTBN2* mutations (SCA5) are expected to change mildly the structural of large protein. In the heterozygote, a case of SCA5, mutant tetramers would be defective but possibly competing with the native tetramers for binding the target membrane proteins. Thus, the protein would be nonfunctional. In contrast, in a recessive case the deletion of just the terminal PH domain in the presented family is deduced not to prevent tetramerization, and mutant tetramer likely has no function and cannot bind the membrane proteins. In a heterozygote (a carrier for SCAR14) the native

tetramers could provide adequate function; or more likely, the mutant transcript undergoes nonsense-mediated decay, not allowing the synthesis of any mutant protein (Nagy and Maquat, 1998), and then all tetramers would be wild-type. The findings expand the phenotypic variability of SCAR14 and could benefit families if children with complex cerebellar ataxia are tested for *SPTBN2* mutation.

The manuscript was accepted and published in the *American Journal of Medical Genetics* as a clinical report: Yıldız Bölükbaşı, E., Afzal, M., Mumtaz, S., Ahmad, N., Malik, S., & Tolun, A. (2017). Progressive SCAR14 with unclear speech, developmental delay, tremor, and behavioral problems caused by a homozygous deletion of the SPTBN2 pleckstrin homology domain. *American Journal of Medical Genetics Part A*, *173*(9), 2494-2499. (Abstract is present in Appendix B)

### 6.5. Bardet-Biedl Syndrome (BBS) Family

The inbred family has nine members afflicted with polydactyly; some of those members also have obesity, severe intellectual disability, speech problems and/or developmental delay. Linkage mapping localized the disease locus to a 5.7-Mb region at 3q29 with a maximal LOD score of 5.65. Exome sequencing identified homozygous truncating mutation c.194_195 (p.Tyr65*) in *CEP19,* encoding a ciliary protein, in all nine affected individuals as well as in one unaffected individual segregating in the family.

The region of the *CEP19* mutation we identified is highly conserved among species. The whole protein sequence is also highly conserved: Conservation among *Homo sapiens*, chimpanzee (*Pan troglodytes*) and Rhesus macaque (*Macaca mulatta*) is 100%, whereas between homo sapiens and mice (*Mus musculus*) is 97% and brown rat (*Rattus norvegicus*) is 86%. No orthologue of *CEP19* was found in *Drosophila*, yeast or C. *elegans* (Shalata *et al.*, 2013). There is just one published article on a mutation in this gene. The variant mentioned in the article is c.244C>T (p.arg82*), which results in morbid obesity and azoospermia. As an animal model, a mouse knockout was generated. Homozygous Cep19-knockout mice were morbidly obese, hyperphagic, glucose intolerant, and insulin resistant

(Shalata *et al.*, 2013). *CEP19* is not expressed in kidney, and the expression level in brain is lower than that of *BBS1* and *BBS10*. Also, there may be some other mutations in some other genes that contribute to different clinical findings.

Two very recent studies were investigated to strengthen the explanation of how a *CEP19* mutation can cause BBS. CEP19 interacts with RABL2B which is GTPase and binds to IFT complexes. Rabl2 knockout mouse displays features including infertility, obesity, polydactyly and retinal degeneration which are characteristic of ciliopathies. After it is recruited to the ciliary base, CEP19 captures RABL2B, and they together enter the primary cilium (Kanie *et al.*, 2017). Nishijima and colleagues showed that due to a 47-amino acid truncation, mutant CEP19 loses its ability to binding of RABL2B protein (Nishijima *et al.*, 2017). We can propose that Tyr65* truncation in CEP19 in our patients impairs the binding function of the centrosomal protein FGFR1OP and thus prevents the entry of RABL2B into primary cilium. Another *CEP19* mutation p.Arg82* was reported as causing morbid obesity, but none of our patients is morbidly obese. Moreover, morbid obesity is not considered a feature of BBS.

In several studies, it has been reported that in BBS cases disease severity and expressivity can be affected by a third trans mutation through an epistatic effect (Badano *et al.*, 2006; Fan *et al.*, 2004; Leitch *et al.*, 2008; Lindstrand *et al.*, 2016). Thus, we thought that besides the *CEP19* variant, variant(s) in a known or novel BBS gene can be modifying the phenotype in this family. *GLI1,* in which we detected a missense variant, encodes a transcription factor that localizes to the primary cilium and nucleus (Haycraft *et al.*, 2005). It is a part of the sonic hedgehog (SHH) pathway, and the rare homozygous mutation in our four patients (501, 593, 505 and 507) was hypothesized as a possible modifier of disease severity or responsible for polydactyly since it interacts with GLI3 (STRING tool), defects in which cause postaxial polydactyly types 1A and B or preaxial polydactyly type IV. In addition, a relation between defects in SHH signaling in the developing limb and BBS patients with postaxial polydactyly was suggested (Tayeh *et al.*, 2008). After our article was published, a report was published showing that *GL11* inactivation can cause various conditions including postaxial polydactyly (Palencia-Campos *et al.*, 2017). However, as only four of our seven patients tested were homozygous for the *GLI1* variant and two others plus three unaffected relatives were heterozygous. Two of our four

homozygous patients (503 and 508) were severely affected whereas the other two (501 and 507) were not so severe. The patient who does not carry the mutation at all (508) was also rather severely affected. Thus, a genotype-phenotypes correlation for this variant was not observed in our study family

Possible consequences of *GLI1* p.Gly274Arg was analyzed by Ute Woehlbier in Chili using three dimensional structure modeling. It was observed that the protein gained a positive charge in the second zinc finger domain of the protein consisting of five zinc finger domain due to change of a neutral, nonpolar glycine to a polar, hydrophilic arginine. The mutation seems to block the DNA binding site severely impairing DNA binding site (Infante *et al.*, 2015).

In total 25 exonic variants were found in known BBS genes in the two exome files, and we considered the four with frequencies <0.05 (Table A.1). They were found in different combinations in BBS subjects and unaffected relatives. One is the rare splicing variant *CCDC28B* c.C330T (p.Phe110Phe; rs41263993); in the heterozygous state it was reported as a modifier of BBS1 (Badano *et al.*, 2006; Bin *et al.*, 2009). Another is the very rare missense variant *MKKS/BBS6* c.A1015G (p.Ile339Val; rs137853909), also reported in the heterozygous state as a modifier in five patients but affected siblings did not carry it (Hichri *et al.*, 2005; Slavotinek *et al.*, 2002). It is predicted as polymorphism by Mutation Taster and tolerated by SIFT, and reported in ClinVar as likely benign. Residue Ile339 is not highly conserved in mammals; conservation of the altered amino acids across species and the sequences around the mutation was not high as assessed using UCSC Genome Browser Conservation Track. Variant *C8ORF37* c.C533T (p.Ala178Val; rs375314973) is predicted as disease causing by Mutation Taster, deleterious by SIFT and possibly damaging by PolyPhen-2. ExAC reports its frequency as 0.006275 (41 alleles in 6534, three homozygotes) in Europeans and as 0.00604 (98 alleles in 16224, two homozygotes) in South Asians. It might be a modifier allele in our patients as well, as some heterozygotes for the allele are severely affected and noncarriers have milder phenotypes. *C8ORF37* Ala178 is completely conserved across species and is within a stretch of seven residues that are very highly conserved. The substitution is predicted as damaging by computational algorithms. Synonymous *TMEM67* (BBS 14, modifier) c.2397T>C (p.Asp799Asp) (NM_153704 is predicted it as disease causing by Mutation Taster due to "donor

marginally increased"; SIFT and PolyPhen-2 are not applicable for synonymous changes. Its frequency in ExAC South Asians is 0.02291 (378 alleles in 16498). Since the protein structures were not available, a prediction of a potential structural and/or functional impact of these variant could not be made.

*CCDC28B* variant identified was found to enhance the use of a cryptic splice acceptor site and the premature termination codon thus created led to nonsense mediated decay of the mRNA. Heterozygous variant was found in 3 of the 64 unrelated BBS patients and 4 of 274 controls in a study (Badano *et al.*, 2006). No affected individual had homozygous or compound heterozygous mutations in that gene. Individuals carrying that allele were more severely affected (Badano *et al.*, 2006). According to ExAC, the frequency in South Asians is 0.011, higher than in Europeans, Finns and Africans. This variant is also found in EVS with a frequency of 0.019 in East Asians. An online tool (Mutation taster) predicts it as disease causing. Current designation given by OMIM is c.330C>T, as ours. According to Ensemble, cDNA position is 430. Perhaps the previous reports had referred to Ensemble. We cannot speculate that there is a genotype-phenotype correlation for the *CCDC28B* variant, because although patient 502 was a homozygote he had a rather mild phenotype with polydactyly only in the feet and no ID, developmental delay or obesity.

The genotype of unaffected individual 512 for six variants was the same as that of her affected brother 509. This was unusual, thus we suspected that the sample could have been mixed up with that of that patient. We decided to do SSCP analysis for the *CEP19* variant for a newly obtained sample together with the old sample of individual 512 to compare the genotypic data to that of individual 509. We observed that the results for the new 512 sample were different from that of 509. Therefore, the sample was subjected to Sanger sequencing for all six variants, and the genotype was found different for three of them. The updated results are presented on Figure 5.24.

In each BBS family a different feature can be prominent. In our study family, the prominent feature is polydactyly which all patients have. The first primary feature in BBS is vision lost, and in our four of five patients investigated rod-cone dystrophy was detected.

Obesity is second prominent feature that is seen in 72-92% of the cases, but only 50% of our patients are obese.

Complex inheritance of BBS due to total mutational burden was purposed by Badano and colleagues (Badano *et al.*, 2006), and also in another study, different mutational loads in cilia-associated genes among siblings was shown as the cause of phenotypic variability (Cardenas-Rodriguez *et al.*, 2013). Therefore, we intended to analyze whether the mutation burden is higher in our patients than the control group. For this purpose, we listed and evaluated exonic and splicing variants with frequencies <0.05 in BBS-related genes plus *GLI,* as we had already done for our BBS patients, in the exome data of ten unrelated Pakistani individuals as a control group. Average of the mutated alleles in BBS-related genes were 4.43 per patient (31/7), 2.67 (8/3) per unaffected sibling and 1.10 (11/10) per control. We applied unpaired one-tailed Student's t-test to the patient and a control group, and t value was calculated as 4.58 and p value as 0.00018. Thus, the mean of the variants in patients was extremely significantly higher ($p < 0.001$) as compared to the mean of the control group.

In conclusion, we identified *CEP19* as the gene responsible for the BBS afflicting the family. In some of the BBS subjects, we also detected, in different combinations, rare splicing mutation *CCDC28B* c.C330T (at 1p35) and missense *MKKS*/*BBS6* p.Ile339Val (20p12.2), both of which are known modifiers that increase BBS severity, p.Ala178Val in *C8ORF37* (*BBS21*) and p.Gly274Arg in *GLI1* (12q13), that encodes a transcription factor localized to cilia and nucleus and takes part in the sonic hedgehog pathway as a nuclear mediator, in embryogenesis. The consanguineous family we studied is very large; this enabled us to investigate for such modifier mutations. *GLI1* p.Gly274Arg is possibly a modifier allele but in homozygous state only. There is ample evidence that support functional redundancy in ciliary genes as well as the hypothesis that an increase in the ciliary mutation load renders the biallelic mutations penetrant. Among our patients the most severe case carries all the modifier alleles we detected and thus has the highest BBS mutational burden reported to date, and the other BBS relatives carry at least one modifier allele. We did not find a genotype-phenotype correlation for *CCDC28B* c.C330T mutation. Our study, while not refuting oligogenic inheritance (as no patient is without a modifier allele), does not support modifier hypothesis as the most severe and a milder case both

have the heaviest mutational load. Identification of *CEP19* as responsible for this novel Bardet-Biedl syndrome expands BBS genes and phenotype, and corroborates that uncovering the genetic etiology of BBS is a challenge.

The manuscript of this study was published in *Journal of Medical Genetics* as an original article: Bölükbaşı, E. Y., Mumtaz, S., Afzal, M., Woehlbier, U., Malik, S., & Tolun, A. (2018). Homozygous mutation in CEP19, a gene mutated in morbid obesity, in Bardet-Biedl syndrome with predominant postaxial polydactyly. *Journal of Medical Genetics*, 2018 Mar;55(3):189-197. doi: 10.1136/jmedgenet-2017-104758. (Abstract is present in Appendix B)

### 6.6. Pleuroparanchymal Fibroelastosis (PPFE) Families

We studied three families with PPFE. In the first family, a synonymous mutation in *TNKS2* (NM_025235; c.1146A>T, p.Ile382Ile) which was predicted as disease causing was identified, but for the other two families causative mutations could not be detected.

For PPFE1 family, due to uncertainty of the eldest sister phenotype, analyses were performed several times. Finally, she was assumed as unaffected and all results were re-evaluated. Among the four candidate variants (Table 5.15), the variants in *FAM35A* and *TNKS2* were validated in the patient with the exome data. The heterozygous variants in *FAM22A* and *FAM22D* were the same and could not be validated since these genes are paralogs and reside in a duplicated region (>2500 bp). Primers to amplify the sites of the variants mapped to seven different regions on genome (In-silico PCR, UCSC Genome Browser). Two other candidate variants screened in the family by SSCP analysis. While *TNKS2* variant segregates with the disease, banding patterns for the variant in *FAM35A* was not as expected. Unaffected sister and mother were both found heterozygous for the variant, and hence it was eliminated. The only candidate remained was the variant in *TNKS2*.

Although *TNKS2* variant is synonymous, it was predicted as disease causing by Mutation taster. The codon AUA which is rare (16%) for isoleucin is converted to a more commonly used codon AUU (36%) (Genscript). This case might be affecting the ribosome stalling and thus the protein folding (Yu *et al.*, 2015). The amino acid is highly conserved among species. *Tankyrase 2* (*TNKS2*) encodes TRF1-interacting ankyrin related ADP-Ribose polymerase 2. Tankrases are polymerases that catalyze the ADP-ribosylation of target proteins and also have a role in the telomerase. Cook and colleagues in 2002 showed that TNKS2 interacts with telomeric repeat binding protein 1 (TERF1) which is an inhibitor of telomerase, at telomeres with its PARP (poly(ADP-ribose) polymerase) activity and inhibits its binding to telomere (Cook *et al.*, 2002). Four telomere-related genes have been associated with sporadic and familial idiopathic pulmonary fibrosis which is a similar disease with PPFE due to telomere shortening, including *TERF1* and *PARP* (Cronkhite *et al.*, 2008; Stuart *et al.*, 2015).

For PPFE2 family, since linkage analysis was performed only with three affected individuals, it yielded many shared regions of homozygosity, but none of the variants in those regions was a good candidate. Since the phenotype was similar with PPFE1 family, the variants in the candidate genes of PPFE1 evaluated. Patient of family PPFE2 did not carry any homozygous rare variants in PPFE1 candidate genes *FAM35A*, *FAM22A*, *FAM22D* and *TNKS2*, and only small regions were homozygous in those regions. Afterwards, the clinician informed us that the phenotype is not clear and may be different from that of the first family.

A new Turkish family (PPFE3) afflicted with the same disease was included in the studies. We obtained blood samples of one affected brother and generated exome data. The other affected brother and the sister had died of PPFE. There are two unaffected siblings. In this family, exome data were generated for the surviving affected sibling without prior SNP genotyping. All variants found in candidate genes in PPFE1 and additionally all variants in idiopathic pulmonary fibrosis (IPF) related genes (*ABCA3, ATP11A, DKC1, DPP9, DSP, EIF2AK4, FAM111B, FAM13A, MARS, MUC5B, OBFC1, SFTPA1, SFTPA2, SFTPC* and *TOLLIP*) were listed upon the suggestion of the clinician, and the ones with MAF <0.01 were evaluated. Neither compound heterozygous variants nor a homozygous

candidate variant was found. The regions for all such variants were heterozygous according to exome file.

Newton and colleagues analyzed 115 patients with variable interstitial lung diseases and found mutations in one of four genes, i.e., telomerase reverse transcriptase (*TERT*), telomerase RNA component (*TERC*), regulator of telomere elongation helicase 1 (*RTEL1*) and poly(A)-specific ribonuclease (*PARN*). The mutations in these four genes cause "telomeropathy" and thus abnormalities in systems such as pulmonary fibrosis, bone marrow dysfunction, liver cirrhosis and early graying. Telomere shortening in circulating leukocytes affects the lung cells and causes pathogenesis (Newton *et al.*, 2016). As a result of investigation of all variants in these four telomere-related genes in the patient exome data of three families, only one heterozygous no synonymous variant (NM_002582: c.G1690A, p.V564I) was found, in *PARN* in family PPFE1, which is predicted as benign by all prediction tools. No candidate variant was found in families PPFE2 or PPFE3.

We thought that the disease may be due to a mutation in predisposition gene for PPFE. For this reason, the common genes with rare variants for PPFE1 and PPFE3 families were listed. In those 61 genes all variants with MAF <0.01 were evaluated. There were no genes with candidate variants in both families.

# 7. CONCLUSION

In this thesis study, causative genes in eight consanguineous families afflicted with six different recessive diseases were searched. In Isolated Intellectual Disability family, *PTRHD1* that was already reported with ID and parkinsonism was found as the causative gene. However, although one of the patients in our study family is over the age of 35, Parkinsonism is not evident. A manuscript was prepared and submitted to Journal of Medical Genetics as a short report. It was rejected for the reason that the patients are too young to develop Parkinsonism. However, it was revealed that the patients are older than 30 ages, and thus the manuscript will be submitted after revisions. In Intellectual Disability and Hypothyroidism family, a novel mutation in *TPO* gene which is already associated with hypothyroidism and ID was identified. In some patients ID was observed and severity was variable. In the patients with more severe ID, we would have liked to investigate whether there was any causative mutation for that, but since we did not have enough DNA sample, exome sequence analysis could not be performed. In Syndromic Intellectual Disability family, a novel missense mutation in *PDIA3,* a novel gene for ID, was identified. The manuscript is ready for submission. In Spinocerebellar Ataxia Family, a novel mutation in *SPTBN2* which was associated with SCAR14 was identified, expanding the phenotypic variability of SCAR14. This study was published in the *American Journal of Medical Genetics* as a clinical report. In Bardet Biedl Syndrome family, a novel gene *CEP19* was identified as a causative gene. In addition, possible modifiers were identified in *GLI1* and in BBS related genes *CCDC28B, MKKS/BBS6, C8ORF37* and *TMEM67*. This study was published in *Journal of Medical Genetics* as an original article. Synonymous mutation in *TNKS2* in PPFE1 family was identified as the strongest candidate. In PPFE2 and PPFE3 families no candidate variant could be identified.

# REFERENCES

Amitani, R., A. Niimi, and F. Kuse, 1992, "Idiopathic pulmonary upper lobe fibrosis (IPUF)", *Kokyu,* Vol.11, pp.693-699.

Badano, J. L., C. C. Leitch, S. J. Ansley, H. May-Simera, S. Lawson, R. A. Lewis, P. L. Beales, …, N. Katsanis, 2006, "Dissection of epistasis in oligogenic Bardet-Biedl syndrome", *Nature,* Vol.439, No.7074, pp.326-330.

Bevilacqua, L., S. Doly, J. Kaprio, Q. Yuan, R. Tikkanen, T. Paunio, …, D. Goldman, 2010, "A population-specific HTR2B stop codon predisposes to severe impulsivity", *Nature,* Vol.468, No.7327, pp.1061-1066.

Bikker, H., F. Baas, and J. J. De Vijlder, 1997, "Molecular analysis of mutated thyroid peroxidase detected in patients with total iodide organification defects", *Journal of Clinical Endocrinology & Metabolism,* Vol.82, No.2, pp.649-653.

Bin, J., J. Madhavan, W. Ferrini, C. A. Mok, G. Billingsley, and E. Heon, 2009, "BBS7 and TTC8 (BBS8) mutations play a minor role in the mutational load of Bardet-Biedl syndrome in a multiethnic population", *Human Mutation,* Vol.30, No.7, pp.E737-746.

Cangul, H., Z. Aycan, A. Olivera-Nappa, H. Saglam, N. A. Schoenmakers, K. Boelaert, …, E. R. Maher, 2013, "Thyroid dyshormonogenesis is mainly caused by TPO mutations in consanguineous community", *Clinical Endocrinoogyl (Oxford),* Vol.79, No.2, pp.275-281.

Cardenas-Rodriguez, M., D. P. Osborn, F. Irigoin, M. Grana, H. Romero, P. L. Beales, and J. L. Badano, 2013, "Characterization of CCDC28B reveals its role in ciliogenesis and provides insight to understand its modifier effect on Bardet-Biedl syndrome", *Human Genetics,* Vol.132, No.1, pp.91-105.

Cheng, S. K., and K. L. Chuah, 2016, "Pleuroparenchymal Fibroelastosis of the Lung: A Review", *Archives of Pathology & Laboratory Medicine,* Vol.140, No.8, pp.849-853.

Cook, B. D., J. N. Dynek, W. Chang, G. Shostak, and S. Smith, 2002, "Role for the related poly(ADP-Ribose) polymerases tankyrase 1 and 2 at human telomeres", *Molecular and Cellular Biology,* Vol.22, No.1, pp.332-342.

Cronkhite, J. T., C. Xing, G. Raghu, K. M. Chin, F. Torres, R. L. Rosenblatt, and C. K. Garcia, 2008, "Telomere shortening in familial and sporadic pulmonary fibrosis", *American Journal of Respiratory and Critical Care Medicine.,* Vol.178, No.7, pp.729-737.

Diaz de Leon, A., J. T. Cronkhite, A. L. Katzenstein, J. D. Godwin, G. Raghu, C. S. Glazer, …, C. K. Garcia, 2010, "Telomere lengths, pulmonary fibrosis and telomerase (TERT) mutations", *PLoS One,* Vol.5, No.5, pp.e10680.

Elsayed, S. M., R. Heller, M. Thoenes, M. S. Zaki, D. Swan, E. Elsobky, …, H. J. Bolz, 2014, "Autosomal dominant SCA5 and autosomal recessive infantile SCA are allelic conditions resulting from SPTBN2 mutations", *European Journal of Human Genetics,* Vol.22, No.2, pp.286-288.

Fan, Y., M. A. Esmail, S. J. Ansley, O. E. Blacque, K. Boroevich, A. J. Ross, …, M. R. Leroux, 2004, "Mutations in a member of the Ras superfamily of small GTP-binding proteins causes Bardet-Biedl syndrome", *Nature Genetics,* Vol.36, No.9, pp.989-993.

Ferrari, D. M., and H. D. Soling, 1999, "The protein disulphide-isomerase family: unravelling a string of folds", *Biochemical Journal,* Vol.339 ( Pt 1), pp.1-10.

Forsythe, E., and P. L. Beales, 2013, "Bardet-Biedl syndrome", *European Journal of Human Genetics,* Vol.21, No.1, pp.8-13.

Garcia, C. K., 2011, "Idiopathic pulmonary fibrosis: update on genetic discoveries", *Proceedings of the. American Thoracic Society,* Vol.8, No.2, pp.158-162.

Gilissen, C., A. Hoischen, H. G. Brunner, and J. A. Veltman, 2012, "Disease gene identification strategies for exome sequencing", *European Journal of Human Genetics*, Vol.20, No.5, pp.490-497.

Goodman, S. R., W. E. Zimmer, M. B. Clark, I. S. Zagon, J. E. Barker, and M. L. Bloom, 1995, "Brain spectrin: of mice and men", *Brain Research Bulletin,* Vol.36, No.6, pp.593-606.

Grasberger, H., and S. Refetoff, 2011, "Genetic causes of congenital hypothyroidism due to dyshormonogenesis", *Current Opinion in Pediatrics*, Vol.23, No.4, pp.421-428.

Griffiths, A. J., J. H. Miller, D. T. Suzuki, R. C. Lewontin, and W. M. Gelbart, 2000, *An introduction to genetic analysis,* New York.

Gudbjartsson, D. F., P. Sulem, H. Helgason, A. Gylfason, S. A. Gudjonsson, F. Zink, …, K. Stefansson, 2015, "Sequence variants from whole genome sequencing a large group of Icelanders", *Scientific Data* , Vol.2, pp.150011.

Haycraft, C. J., B. Banizs, Y. Aydin-Son, Q. Zhang, E. J. Michaud, and B. K. Yoder, 2005, "Gli2 and Gli3 localize to cilia and require the intraflagellar transport protein polaris for processing and function", *PLoS Genetics,* Vol.1, No.4, pp.e53.

Hettinghouse, A., R. Liu, and C. J. Liu, 2017, "Multifunctional molecule ERp57: From cancer to neurodegenerative diseases", *Pharmacology & Therapeutics*, Vol.181, No.1, pp.34-48

Hibaoui, Y., I. Grad, A. Letourneau, F. A. Santoni, S. E. Antonarakis, and A. Feki, 2014, "Data in brief: Transcriptome analysis of induced pluripotent stem cells from monozygotic twins discordant for trisomy 21", *Genomics Data,* Vol.2, pp.226-229.

Hichri, H., C. Stoetzel, V. Laurier, S. Caron, S. Sigaudy, P. Sarda, …, H. Dollfus, 2005, "Testing for triallelism: analysis of six BBS genes in a Bardet-Biedl syndrome family cohort", *European Journal of Human Genetics,* Vol.13, No.5, pp.607-616.

Hoffmann, K., and T. H. Lindner, 2005, "easyLINKAGE-Plus--automated linkage analyses using large-scale SNP data", *Bioinformatics,* Vol.21, No.17, pp.3565-3567.

Hu, H., M. L. Matter, L. Issa-Jahns, M. Jijiwa, N. Kraemer, L. Musante, …, A. M. Kaindl, 2014, "Mutations in PTRH2 cause novel infantile-onset multisystem disease with intellectual disability, microcephaly, progressive ataxia, and muscle weakness", *Annals of Clinical and Translational Neurology,* Vol.1, No.12, pp.1024-1035.

Infante, P., M. Mori, R. Alfonsi, F. Ghirga, F. Aiello, S. Toscano, …, L. Di Marcotullio, 2015, "Gli1/DNA interaction is a druggable target for Hedgehog-dependent tumors", *EMBO Journal,* Vol.34, No.2, pp.200-217.

Ishii, T., M. Funakoshi, and H. Kobayashi, 2006, "Yeast Pth2 is a UBL domain-binding protein that participates in the ubiquitin-proteasome pathway", *EMBO Journal,* Vol.25, No.23, pp.5492-5503.

Jaberi, E., M. Rohani, G. A. Shahidi, S. Nafissi, E. Arefian, M. Soleimani, …, E. Elahi, 2016, "Mutation in ADORA1 identified as likely cause of early-onset parkinsonism and cognitive dysfunction", *Movement Disorders,* Vol.31, No.7, pp.1004-1011.

Jarome, T. J., J. L. Kwapis, J. J. Hallengren, S. M. Wilson, and F. J. Helmstetter, 2013, "The ubiquitin-specific protease 14 (USP14) is a critical regulator of long-term memory formation", *Learning & Memory,* Vol.21, No.1, pp.9-13.

Jayadev, S., and T. D. Bird, 2013, "Hereditary ataxias: overview", *Genetics in Medicine,* Vol.15, No.9, pp.673-683.

Jessop, C. E., S. Chakravarthi, N. Garbi, G. J. Hammerling, S. Lovell, and N. J. Bulleid, 2007, "ERp57 is essential for efficient folding of glycoproteins sharing common structural domains", *EMBO Journal*, Vol.26, No.1, pp.28-40.

Kanie, T., K. L. Abbott, N. A. Mooney, E. D. Plowey, J. Demeter, and P. K. Jackson, 2017, "The CEP19-RABL2 GTPase Complex Binds IFT-B to Initiate Intraflagellar Transport at the Ciliary Base", *Devolepmental Cell*, Vol.42, No.1, pp.22-36 e12.

Katsanis, N., 2004, "The oligogenic properties of Bardet-Biedl syndrome", *Human Molecular Genetics*, Vol.13 Spec No 1, pp.R65-71.

Katsanis, N., S. J. Ansley, J. L. Badano, E. R. Eichers, R. A. Lewis, B. E. Hoskins, …, J. R. Lupski, 2001, "Triallelic inheritance in Bardet-Biedl syndrome, a Mendelian recessive disorder", *Science*, Vol.293, No.5538, pp.2256-2259.

Khan, S. A., N. Muhammad, M. A. Khan, A. Kamal, Z. U. Rehman, and S. Khan, 2016, "Genetics of human Bardet-Biedl syndrome, an updates", *Clinical Genetics*, Vol.90, No.1, pp.3-15.

Khodadadi, H., L. J. Azcona, V. Aghamollaii, M. D. Omrani, M. Garshasbi, S. Taghavi, …, C. Paisan-Ruiz, 2017, "PTRHD1 (C2orf79) mutations lead to autosomal-recessive intellectual disability and parkinsonism", *Movement Disorders*, Vol.32, No.2, pp.287-291.

Ku, C. S., N. Naidoo, and Y. Pawitan, 2011, "Revisiting Mendelian disorders through exome sequencing", *Human Genetics*, Vol.129, No.4, pp.351-370.

Kusagaya, H., Y. Nakamura, M. Kono, Y. Kaida, S. Kuroishi, N. Enomoto, …, K. Chida, 2012, "Idiopathic pleuroparenchymal fibroelastosis: consideration of a clinicopathological entity in a series of Japanese patients", *BMC Pulmonary Medicine*, Vol.12, pp.72.

Lander, E. S., and D. Botstein, 1987, "Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children", *Science,* Vol.236, No.4808, pp.1567-1570.

Leitch, C. C., N. A. Zaghloul, E. E. Davis, C. Stoetzel, A. Diaz-Font, S. Rix, …, N. Katsanis, 2008, "Hypomorphic mutations in syndromic encephalocele genes are associated with Bardet-Biedl syndrome", *Nature Genetics,* Vol.40, No.4, pp.443-448.

Leonard, H., and X. Wen, 2002, "The epidemiology of mental retardation: challenges and opportunities in the new millennium", *Mental Retardation and Developmental Disabilities Research Reviews,* Vol.8, No.3, pp.117-134.

Lindstrand, A., S. Frangakis, C. M. Carvalho, E. B. Richardson, K. A. McFadden, J. R. Willer, …, N. Katsanis, 2016, "Copy-Number Variation Contributes to the Mutational Load of Bardet-Biedl Syndrome", *American Journal of Human Genetics,* Vol.99, No.2, pp.318-336.

Lise, S., Y. Clarkson, E. Perkins, A. Kwasniewska, E. Sadighi Akha, R. P. Schnekenberg, …, A. H. Nemeth, 2012, "Recessive mutations in SPTBN2 implicate beta-III spectrin in both cognitive and motor development", *PLoS Genetics,* Vol.8, No.12, pp.e1003074.

Liu, Y. F., S. M. Sowell, Y. Luo, A. Chaubey, R. S. Cameron, H. G. Kim, and A. K. Srivastava, 2015, "Autism and Intellectual Disability-Associated KIRREL3 Interacts with Neuronal Proteins MAP1B and MYO16 with Potential Roles in Neurodevelopment", *PLoS One,* Vol.10, No.4, pp.e0123106.

Manto, M., 2008, "The cerebellum, cerebellar disorders, and cerebellar research--two centuries of discoveries", *Cerebellum,* Vol.7, No.4, pp.505-516.

Matullo, G., C. Di Gaetano, and S. Guarrera, 2013, "Next generation sequencing and rare genetic variants: from human population studies to medical genetics", *Environmental and Molecular Mutagenesis,* Vol.54, No.7, pp.518-532.

Mefford, H. C., M. L. Batshaw, and E. P. Hoffman, 2012, "Genomics, intellectual disability, and autism", *New England Journal of Medicine,* Vol.366, No.8, pp.733-743.

Mittal, K., M. A. Rafiq, R. Rafiullah, R. Harripaul, H. Ali, M. Ayaz, …, M. Ayub, 2016, "Mutations in the genes for thyroglobulin and thyroid peroxidase cause thyroid dyshormonogenesis and autosomal-recessive intellectual disability", *Journal of Human Genetics,* Vol.61, No.10, pp.867-872.

Moore, S. J., J. S. Green, Y. Fan, A. K. Bhogal, E. Dicks, B. A. Fernandez, …, P. S. Parfrey, 2005, "Clinical and genetic epidemiology of Bardet-Biedl syndrome in Newfoundland: a 22-year prospective, population-based, cohort study", *American Journal of Medical Genetics Part A,* Vol.132A, No.4, pp.352-360.

Morton, N. E., 1955, "Sequential tests for the detection of linkage", *American Journal of Human Genetics,* Vol.7, No.3, pp.277-318.

Nagy, E., and L. E. Maquat, 1998, "A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance", *Trends in Biochemical Sciences,* Vol.23, No.6, pp.198-199.

Nakatani, T., T. Arai, M. Kitaichi, M. Akira, K. Tachibana, C. Sugimoto, …, Y. Inoue, 2015, "Pleuroparenchymal fibroelastosis from a consecutive database: a rare disease entity?", *European Respiratory Journal,* Vol.45, No.4, pp.1183-1186.

Narasimhan, V. M., K. A. Hunt, D. Mason, C. L. Baker, K. J. Karczewski, M. R. Barnes, …, D. A. van Heel, 2016, "Health and population effects of rare gene knockouts in adult humans with related parents", *Science,* Vol.352, No.6284, pp.474-477.

Newton, C. A., K. Batra, J. Torrealba, J. Kozlitina, C. S. Glazer, C. Aravena, …, C. K. Garcia, 2016, "Telomere-related lung fibrosis is diagnostically heterogeneous but uniformly progressive", *European Respiratory Journal,* Vol.48, No.6, pp.1710-1720.

Ng, S. B., E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham, C. Lee, …, J. Shendure, 2009, "Targeted capture and massively parallel sequencing of 12 human exomes", *Nature,* Vol.461, No.7261, pp.272-276.

Nishijima, Y., Y. Hagiya, T. Kubo, R. Takei, Y. Katoh, and K. Nakayama, 2017, "RABL2 interacts with the intraflagellar transport-B complex and CEP19 and participates in ciliary assembly", *Molecular Biology of the Cell,* Vol.28, No.12, pp.1652-1666.

Palencia-Campos, A., A. Ullah, J. Nevado, R. Yildirim, E. Unal, M. Ciorraga, …, V. L. Ruiz-Perez, 2017, "GLI1 inactivation is associated with developmental phenotypes overlapping with Ellis-van Creveld syndrome", *Human Molecular Genetics,* Vol.26, No.23, pp.4556-4571.

Park, S. M., and V. K. Chatterjee, 2005, "Genetics of congenital hypothyroidism", *Journal of Medical Genetics,* Vol.42, No.5, pp.379-389.

Perkins, E. M., Y. L. Clarkson, N. Sabatier, D. M. Longhurst, C. P. Millward, J. Jack, …, M. Jackson, 2010, "Loss of beta-III spectrin leads to Purkinje cell dysfunction recapitulating the behavior and neuropathology of spinocerebellar ataxia type 5 in humans", *Journal of Neuroscience,* Vol.30, No.14, pp.4857-4867.

Perrotta, S., P. G. Gallagher, and N. Mohandas, 2008, "Hereditary spherocytosis", *Lancet,* Vol.372, No.9647, pp.1411-1426.

Pieretti, M., F. P. Zhang, Y. H. Fu, S. T. Warren, B. A. Oostra, C. T. Caskey, and D. L. Nelson, 1991, "Absence of expression of the FMR-1 gene in fragile X syndrome", *Cell,* Vol.66, No.4, pp.817-822.

Rimoin, D. L., R. E. Pyeritz, and B. Korf, 2013, *Emery and Rimoin's essential medical genetics,* Elsevier.

Risch, N., 1992, "Genetic linkage: interpreting lod scores", *Science,* Vol.255, No.5046, pp.803-804.

Saleheen, D., P. Natarajan, I. M. Armean, W. Zhao, A. Rasheed, S. A. Khetarpal, …, S. Kathiresan, 2017, "Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity", *Nature,* Vol.544, No.7649, pp.235-239.

Santana-Codina, N., R. Carretero, R. Sanz-Pamplona, T. Cabrera, E. Guney, B. Oliva, …, A. Sierra, 2013, "A transcriptome-proteome integrated network identifies endoplasmic reticulum thiol oxidoreductase (ERp57) as a hub that mediates bone metastasis", *Molecular & Cellular Proteomics,* Vol.12, No.8, pp.2111-2125.

Schaefer, E., J. Lauer, M. Durand, V. Pelletier, C. Obringer, A. Claussmann, …, H. Dollfus, 2014, "Mesoaxial polydactyly is a major feature in Bardet-Biedl syndrome patients with LZTFL1 (BBS17) mutations", *Clinical Genetics,* Vol.85, No.5, pp.476-481.

Scheidecker, S., C. Etard, N. W. Pierce, V. Geoffroy, E. Schaefer, J. Muller, …, H. Dollfus, 2014, "Exome sequencing of Bardet-Biedl syndrome patient identifies a null mutation in the BBSome subunit BBIP1 (BBS18)", *Journal of Medical Genetics,* Vol.51, No.2, pp.132-136.

Shalata, A., M. C. Ramirez, R. J. Desnick, N. Priedigkeit, C. Buettner, C. Lindtner, …, J. A. Martignetti, 2013, "Morbid obesity resulting from inactivation of the ciliary protein CEP19 in humans and mice", *American Journal of Human Genetics,* Vol.93, No.6, pp.1061-1071.

Slavotinek, A. M., C. Searby, L. Al-Gazali, R. C. Hennekam, C. Schrander-Stumpel, M. Orcana-Losa, …, L. G. Biesecker, 2002, "Mutation analysis of the MKKS gene in McKusick-Kaufman syndrome and selected Bardet-Biedl syndrome patients", *Human Genetics,* Vol.110, No.6, pp.561-567.

Stenson, P. D., E. V. Ball, K. Howells, A. D. Phillips, M. Mort, and D. N. Cooper, 2009, "The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics", *Human Genomics,* Vol.4, No.2, pp.69-72.

Stuart, B. D., J. Choi, S. Zaidi, C. Xing, B. Holohan, R. Chen, …, C. K. Garcia, 2015, "Exome sequencing links mutations in PARN and RTEL1 with familial pulmonary fibrosis and telomere shortening", *Nature Genetics,* Vol.47, No.5, pp.512-517.

Sun, W., L. Huang, Y. Xu, X. Xiao, S. Li, X. Jia, …, Q. Zhang, 2015, "Exome Sequencing on 298 Probands With Early-Onset High Myopia: Approximately One-Fourth Show Potential Pathogenic Mutations in RetNet Genes", *Investigative Ophthalmology & Visual Science,* Vol.56, No.13, pp.8365-8372.

Tayeh, M. K., H. J. Yen, J. S. Beck, C. C. Searby, T. A. Westfall, H. Griesbach, …, D. C. Slusarski, 2008, "Genetic interaction between Bardet-Biedl syndrome genes and implications for limb patterning", *Human Molecular Genetics,* Vol.17, No.13, pp.1956-1967.

Tsuchimine, S., N. Yasui-Furukori, A. Kaneda, and S. Kaneko, 2013, "Differential effects of the catechol-O-methyltransferase Val158Met genotype on the cognitive function of schizophrenia patients and healthy Japanese individuals", *PLoS One,* Vol.8, No.11, pp.e76763.

van Bokhoven, H., 2011, "Genetic and epigenetic networks in intellectual disabilities", *Annual Review of Genetics,* Vol.45, pp.81-104.

Velazquez-Perez, L. C., R. Rodriguez-Labrada, and J. Fernandez-Ruiz, 2017, "Spinocerebellar Ataxia Type 2: Clinicogenetic Aspects, Mechanistic Insights, and Management Approaches", *Frontiers in Neurology,* Vol.8, pp.472.

Vissers, L. E., C. Gilissen, and J. A. Veltman, 2016, "Genetic studies in intellectual disability and related disorders", *Nature Review Genetics,* Vol.17, No.1, pp.9-18.

Wang, Y., A. Nizkorodov, K. Riemenschneider, C. S. Lee, R. Olivares-Navarrete, Z. Schwartz, and B. D. Boyan, 2014, "Impaired bone formation in Pdia3 deficient mice", *PLoS One,* Vol.9, No.11, pp.e112708.

Watanabe, K., 2013, "Pleuroparenchymal Fibroelastosis: Its Clinical Characteristics", *Current Respiratory Medicine Reviews*, Vol.9, pp.299-237.

Wheeler, S. M., M. P. McAndrews, E. D. Sheard, and J. Rovet, 2012, "Visuospatial associative memory and hippocampal functioning in congenital hypothyroidism", *Journal of the International Neuropsychological Society*, Vol.18, No.1, pp.49-56.

Wheeler, S. M., K. A. Willoughby, M. P. McAndrews, and J. F. Rovet, 2011, "Hippocampal size and memory functioning in children and adolescents with congenital hypothyroidism", *Journal of Clinical Endocrinology & Metabolism*, Vol.96, No.9, pp.E1427-1434.

Yildiz Bolukbasi, E., S. Mumtaz, M. Afzal, U. Woehlbier, S. Malik, and A. Tolun, 2018, "Homozygous mutation in CEP19, a gene mutated in morbid obesity, in Bardet-Biedl syndrome with predominant postaxial polydactyly", *Journal of Medical Genetics*, Vol.55, No.3, pp.189-197.

Yu, C. H., Y. Dang, Z. Zhou, C. Wu, F. Zhao, M. S. Sachs, and Y. Liu, 2015, "Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding", *Molecular Cell*, Vol.59, No.5, pp.744-754.

# APPENDIX A: TABLE FOR BBS FAMILY

Table A.1. All exonic and splicing variants in known BBS-related genes in exome files of patients 503 and 509.

| Disease | Gene* | Change | | Transcript ID | Hom/Het | Freq** | Mutation Taster | SIFT | PolyPhen-2 | ClinVar | In patient |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | DNA | Protein | | | | | | | | |
| BBS 1 | *BBS1* | c.1413C>T | p.L471L | NM_024649 | Hom | 0.210 | Polymorphism | Tolerated | NA | - | 503, 509 |
| | | c.378G>A | p.L126L | NM_024649 | Hom | 0.211 | Polymorphism | Tolerated | NA | - | 503, 509 |
| BBS 1 modifier | *CCDC28* | c.330C>T (splicing) | p.F110F | NM_024296 | Het | 0.007 | Disease causing | Tolerated | NA | BBS, modifier | 503 |
| BBS 2 | *BBS2* | c.367A>G | p.I123V | NM_031885 | Het | 0.249 | Polymorphism | Tolerated | Benign | - | 503 |
| | | c.209G>A | p.S70N | NM_031885 | Hom | 0.990 | Polymorphism | Tolerated | Benign | - | 503, 509 |
| BBS 3 | *ARL6* | - | - | - | - | - | - | - | - | - | - |
| BBS 4 | *BBS4* | c.545T>C | p.I182T | NM_001252678 | Het | 0.430 | Polymorphism | Tolerated | Benign | NP, BBS | 503 |
| BBS 5 | *BBS5* | - | - | - | - | - | - | - | - | - | - |

*Genes are from OMIM and Lindstrand et al., 2016. ** Frequency in ExAC South Asian samples. NA, not applicable; P, pathogenic; NP, non-pathogenic; NS, not specified; CI, conflicting interpretations of pathogenicity.

Table A.1. All exonic and splicing variants in known BBS-related genes in exome files of patients 503 and 509 (cont.).

| Disease | Gene* | Change | | Transcript ID | Hom/Het | Freq** | Mutation Taster | SIFT | PolyPhen-2 | ClinVar | In patient |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | DNA | Protein | | | | | | | | |
| BBS 6 | *MKKS* | c.1015A>G | p.I339V | NM_018848 | Het | 0.006 | Polymorphism | Tolerated | Benign | Unknown | 503, 509 |
| BBS 7 | *BBS7* | - | - | - | - | - | - | - | - | - | - |
| BBS 8 | *TTC8* | - | - | - | - | - | - | - | - | - | - |
| BBS 9 | *BBS9* | c.1363G>A | p.A455T | NM_001033605 | Het | 0.23 | Polymorphism | Tolerated | Benign | - | 503 |
| BBS 10 | *BBS10* | c.1616C>T | p.P539L | NM_024685 | Het | 0.091 | Polymorphism | Tolerated | Benign | NP, NS | 509 |
| BBS 11 | *TRIM32* | - | - | - | - | - | - | - | - | - | - |
| BBS 12 | *BBS12* | c.1872A>G | p.Q624Q | NM_152618 | Het | 0.134 | Polymorphism | Tolerated | NA | NP, NS | 503 |
| | | c.1410C>T | p.C470C | NM_152618 | Het | 0.135 | Polymorphism | Tolerated | NA | NP, NS | 503 |
| | | c.1157G>A | p.R386Q | NM_152618 | Het | 0.462 | Polymorphism | Tolerated | Benign | NP, NS | 503 |
| | | c.1380G>C | p.V460V | NM_152618 | Het | 0.142 | Polymorphism | Tolerated | NA | NP, NS | 503 |
| | | c.1399G>A | p.D467N | NM_152618 | Het | 0.134 | Polymorphism | Tolerated | Benign | NP, NS | 503 |
| BBS 13 | *MKS1* | - | - | - | - | - | - | - | - | - | - |

*Genes are from OMIM and Lindstrand et al., 2016. ** Frequency in ExAC South Asian samples. NA, not applicable; P, pathogenic; NP, non-pathogenic; NS, not specified; CI, conflicting interpretations of pathogenicity.

Table A.1. All exonic and splicing variants in known BBS-related genes in exome files of patients 503 and 509 (cont.).

| Disease | Gene* | Change | | Transcript ID | Hom/Het | Freq** | Mutation Taster | SIFT | PolyPhen-2 | ClinVar | In patient |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | DNA | Protein | | | | | | | | |
| BBS 14 | *CEP290* | c.2268A>G | p.S756S | NM_025114 | Hom | 0.888 | Polymorphism | Tolerated | NA | NP, NS | 503, 509 |
| | | c.2512A>G | p.K838E | NM_025114 | Het | 0.092 | Polymorphism | Tolerated | Benign | NP, NS | 509 |
| | | c.4119A>G | p.K1373K | NM_025114 | Het | 0.081 | Disease causing | Tolerated | NA | Probably NP, NS | 509 |
| BBS 14, modifier | *TMEM67* | c.1810A>G | p.I604V | NM_153704 | Het | 0.681 | Polymorphism | Tolerated | Benign | Probably NP, NS | 503, 509 |
| | | c.2397T>C | p.D799D | NM_153704 | Het | 0.022 | Disease causing | Tolerated | NA | NP, NS | 503, 509 |
| BBS 15 | *WDPCP* | - | - | - | - | - | - | - | - | - | - |
| BBS 16 | SDCCAG8 | c.1134A>T | p.E378D | NM_006642 | Het | 0.456 | Polymorphism | Tolerated | Possibly damaging | - | 503, 509 |
| | | c.1725G>A | p.E575E | NM_006642 | Het | 0.4147 | Polymorphism | Tolerated | NA | - | 503, 509 |
| BBS 17 | *LZTFL1* | c.736G>A | p.D246N | NM_020347 | Het | 0.0750 | Polymorphism | Deleterious | Possibly damaging | - | 509 |
| BBS 18 | *BBIP1* | - | - | - | - | - | - | - | - | - | - |
| BBS 19 | *IFT27* | - | - | - | - | - | - | - | - | - | - |

*Genes are from OMIM and Lindstrand et al., 2016. ** Frequency in ExAC South Asian samples. NA, not applicable; P, pathogenic; NP, non-pathogenic; NS, not specified; CI, conflicting interpretations of pathogenicity.

Table A.1. All exonic and splicing variants in known BBS-related genes in exome files of patients 503 and 509 (cont.).

| Disease | Gene* | Change | | Transcript ID | Hom/Het | Freq** | Mutation Taster | SIFT | PolyPhen-2 | ClinVar | In patient |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | DNA | Protein | | | | | | | | |
| BBS 20 | *IFT74* | c.1790C>T | p.T597I | NM_001099222 | Het | 0.2620 | Polymorphism | Tolerated | Benign | - | 509 |
| | | c.670T>C | p.F224L | NM_001099222 | Het | 0.1260 | Polymorphism | Tolerated | Benign | - | 509 |
| BBS21 | *C8ORF37* | c.533C>T | p.A178V | NM_177965 | Het | 0.0005 | Disease causing | Deleterious | Possibly damaging | - | 503 |
| Possible BBS gene | *ALMS1* | c.36_41del | p.12_14del | NM_015120 | Het | 0.0529 | Polymorphism | - | NA | - | 509 |
| | | c.4176A>G | p.Q1392Q | NM_015120 | Het, Hom | 0.1588 | Polymorphism | Tolerated | NA | - | 503, 509 |
| | | c.5623A>G | p.I1875V | NM_015120 | Het, Hom | 0.1591 | Polymorphism | Tolerated | Benign | - | 503, 509 |
| | | c.8567A>G | p.N2856S | NM_015120 | Het, Hom | 0.1592 | Polymorphism | Tolerated | Benign | - | 503, 509 |
| | | c.1174C>T | p.R392C | NM_015120 | Het, Hom | 0.1680 | Polymorphism | Tolerated | Benign | - | 503, 509 |
| | | c.12172C>T | p.L4058L | NM_015120 | Het | 0.1573 | Polymorphism | Tolerated | NA | - | 503, 509 |
| | | c.2187C>T | p.F729F | NM_015120 | Het | 0.6670 | Polymorphism | Tolerated | NA | - | 503 |
| | | c.9558C>T | p.T3186T | NM_015120 | Het, Hom | 0.1543 | Polymorphism | Tolerated | NA | - | 503, 509 |

*Genes are from OMIM and Lindstrand et al., 2016. ** Frequency in ExAC South Asian samples. NA, not applicable; P, pathogenic; NP, non-pathogenic; NS, not specified; CI, conflicting interpretations of pathogenicity.

Table A.1. All exonic and splicing variants in known BBS-related genes in exome files of patients 503 and 509 (cont.).

| Disease | Gene* | Change | | Transcript ID | Hom/Het | Freq** | Mutation Taster | SIFT | PolyPhen-2 | ClinVar | In patient |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | DNA | Protein | | | | | | | | |
| Possible BBS gene | *ALMS1* | c.12086G>A | p.R4029K | NM_015120 | Het | 0.6694 | Polymorphism | Tolerated | Benign | - | 503 |
| | | c.4241G>C | p.G1414A | NM_015120 | Het, Hom | 0.1585 | Polymorphism | Tolerated | Benign | - | 503, 509 |
| | | c.6851G>C | p.R2284P | NM_015120 | Het, Hom | 0.1592 | Polymorphism | Tolerated | Benign | - | 503, 509 |
| | | c.8478G>T | p.R2826S | NM_015120 | Het, Hom | 0.1591 | Polymorphism | Tolerated | Benign | - | 503, 509 |
| | | c.2012T>G | p.V671G | NM_015120 | Hom | 0.8266 | Polymorphism | Tolerated | Benign | - | 503, 509 |
| | | c.6209T>C | p.I2070T | NM_015120 | Het, Hom | 0.1428 | Polymorphism | Tolerated | Benign | - | 503, 509 |
| | | c.6333T>A | p.S2111R | NM_015120 | Het, Hom | 0.1593 | Polymorphism | Tolerated | Benign | - | 503, 509 |
| Possible BBS gene | *NPHP1* | c.273C>T | p.T91T | NM_001128179 | Het | 0.0016 | Polymorphism | Tolerated | NA | - | 509 |
| | | c.468G>A | p.E156E | NM_001128179 | Hom | 0.5047 | Polymorphism | Tolerated | NA | NS | 503 |
| Possible BBS gene | *NPHP4* | c.2818-2T>A splicing | - | NM_015102 | Het | 0.8305 | Polymorphism | NA | NA | NP, NS | 503 |
| | | c.3570A>G | p.E1190E | NM_015102 | Het | 0.4711 | Polymorphism | Tolerated | NA | NP, NS | 503 |

*Genes are from OMIM and Lindstrand et al., 2016. ** Frequency in ExAC South Asian samples. NA, not applicable; P, pathogenic; NP, non-pathogenic; NS, not specified; CI, conflicting interpretations of pathogenicity.

Table A.1. All exonic and splicing variants in known BBS-related genes in exome files of patients 503 and 509 (cont.).

| Disease | Gene* | Change | | Transcript ID | Hom/Het | Freq** | Mutation Taster | SIFT | PolyPhen-2 | ClinVar | In patient |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | DNA | Protein | | | | | | | | |
| Possible BBS gene | *NPHP4* | c.2802C>T | p.R934R | NM_015102 | Het | 0.4962 | Polymorphism | Tolerated | NA | NP, NS | 509 |
| | | c.2643G>A | p.A881A | NM_015102 | Het | 0.1486 | Polymorphism | Tolerated | NA | NP, NS | 509 |
| Possible BBS gene | *RPGRIP1L* | c.3696C>T | p.D1232D | NM_001127897 | Het | 0.0837 | Polymorphism | Tolerated | NA | Probably NP, NS | 509 |
| | | c.2971G>A | p.G991S | NM_001127897 | Het | 0.1383 | Polymorphism | Tolerated | Benign | Probably NP, NS | 503, 509 |
| | | c.685G>A | p.A229T | NM_001127897 | Het | 0.0681 | Disease causing | Tolerated | Benign | CI | 503, 509 |
| Possible BBS gene | *TTC21B* | c.826A>G | p.T276A | NM_024753 | Hom | 0.9999 | Polymorphism | Tolerated | Benign | NS | 503, 509 |
| | | c.1695C>T | p.Y565Y | NM_024753 | Hom | 0.3331 | Polymorphism | Tolerated | NA | NS | 503 |
| | | c.601G>A | p.V201M | NM_024753 | Hom | 0.6666 | Polymorphism | Deleterious | Probably damaging | NS | 509 |
| | | c.2175T>C | p.F725F | NM_024753 | Hom | 0.3336 | Polymorphism | Tolerated | NA | NS | 503 |

*Genes are from OMIM and Lindstrand et al., 2016. ** Frequency in ExAC South Asian samples. NA, not applicable; P, pathogenic; NP, non-pathogenic; NS, not specified; CI, conflicting interpretations of pathogenicity