QUADRUPEDAL GAIT IN HUMANS: IDENTIFICATION AND PARTIAL CHARACTERIZATION OF A NOVEL GENE WD REPEAT DOMAIN 81 (*WDR81*)

A DISSERTATION SUBMITTED TO
THE DEPARTMENT OF MOLECULAR BIOLOGY AND GENETICS
AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

By
Süleyman İsmail Gülsüner
November, 2011

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of doctor of philosophy.

Prof. Dr. Tayfun Özçelik (Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of doctor of philosophy.

Assist. Prof. Dr. Özlen Konu

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of doctor of philosophy.

Assoc. Prof. Dr. Ali Osmay Güre

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of doctor of philosophy.

Assist. Prof. Dr. Ayşe Begün Tekinay

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of doctor of philosophy.

Prof. Dr. Haluk Topaloğlu

Approved for the Graduate School of Engineering and Science:

Prof. Dr. Levent Onural
Director of the Graduate School

ABSTRACT

QUADRUPEDAL GAIT IN HUMANS: IDENTIFICATION AND PARTIAL CHARACTERIZATION OF A NOVEL GENE WD REPEAT DOMAIN 81 (*WDR81*)

Süleyman İsmail Gülsüner
Ph.D. in Molecular Biology and Genetics
Supervisor: Prof. Dr. Tayfun Özçelik
November, 2011

Identification of disease genes responsible for cerebellar phenotypes provides mechanistic insights into the development of cerebellum. Neural pathways involved in bipedal gait in humans is not completely understood. Cerebellar ataxia, mental retardation, and disequilibrium syndrome (CAMRQ) is a rare neurodevelopmental disorder accompanied by quadrupedal gait, dysarthric speech and cerebellar hypoplasia. A large consanguineous family exhibiting this rare disorder was investigated in this study. Disease locus was mapped to a 7.1 Mb region on chromosome 17p by genetic analysis. Targeted capture and massively parallel DNA sequencing using the DNA of three affected and two carrier individuals enabled the identification of a novel variant, p.P856L, in a predicted transcript of WD repeat domain 81 gene (*WDR81*). Several exclusion filters including segregation analysis, identification of rare polymorphisms, extended pedigree screen and bioinformatics evaluation was performed. Expression analysis revealed highest levels of transcripts in cerebellum and corpus callosum. In mouse brain *Wdr81* RNA was observed in cerebellum, especially in Purkinje cell layer. The major structural abnormalities of the patients were atrophy of superior, middle and inferior cerebellar peduncles and corpus callosum. These findings are compatible with the expression pattern of the gene. Analysis of the developing mouse brain revealed that, the expression pattern of the gene was correlated with those involved in neuronal differentiation. This study was one of the first examples of the utility of next generation sequencing in discovery of genes associated with Mendelian phenotypes.

Keywords: Quadrupedal locomotion, CAMRQ2, Unertan syndrome, next generation sequencing.

ÖZET

İNSANDA EL-AYAK ÜZERİNDE YÜRÜME: WD REPEAT DOMAIN 81 (*WDR81*) GENİ'NİN TANIMLANMASI VE KISMİ KARAKTERİZASYONU

Süleyman İsmail Gülsüner
Moleküler Biyoloji ve Genetik, Doktora
Tez Yöneticisi: Prof. Dr. Tayfun Özçelik
Kasım, 2011

Serebellar ataksi, mental retardasyon ve dengesizlik sendromu (CAMRQ) insanda el-ayak üzerinde yürüme, dizartrik konuşma ve serebellar hipoplazi ile giden nadir bir sinir gelişimsel hastalıktır. Bu çalışmada, bu hastalıktan etkilenmiş akraba evliliği bulunan geniş bir aile incelenmiştir. Bağlantı incelemeleri ve homozigotluk haritalaması ile hastalık lokusu kromozom 17p13.1-13.3 üzerinde bulunan 7.1 Mb'lık bir bölgeye haritalanmıştır. Fenotipten sorumlu mutasyonu saptamak için, hastalık lokusu üç adet etkilenmiş birey ve iki adet taşıyıcının DNA örnekleri kullanılarak dizi yakalama mikrodizinleri ile yakalanmış ve yeni nesil DNA dizilemesi ile dizilenmiştir. Aile içi yayılım incelemeleri, nadir polimorfizmlerin belirlenmesi, genişletilmiş soyağacının taranması ve tahmin araçları gibi birçok süzgeç kullanılarak, WD repeat domain 81 (*WDR81*) geninin tahmini ifade varyantında p.P856L mutasyonu tanımlanmıştır. İfade incelemeleri insan beyninde en yüksek ifadenin serebellum ve korpus kallozumda olduğunu göstermiştir. Fare beyin dokularında yapılan incelemelerde *Wdr81* RNA'sı serebellumda, özellikle Purkinje hücrelerinde gözlenmiştir. Hastalarda saptanan belirgin yapısal bozukluklar üst, orta ve alt serebellar pedünküller ve korpus kallosumda saptanmıştır ve bu genin ifade paterni ile uyumludur. Gelişen fare beyni ifade mikrodizin incelemeleri sonucu, genin embyonik günler arasındaki ifade paterninin nöronal farklılaşmada yer alan genlere paralel olduğu belirlenmiştir. Bu çalışma Mendel kalıtımı gösteren hastalıklarda sorumlu genin bulunması amaçlı yeni nesil dizileme teknolojilerinin kullanılmasının ilk örneklerinden biridir.

Anahtar sözcükler: El-ayak üzerinde yürüme, CAMRQ2, Ünertan sendromu, yeni nesil dizileme.

ANNEM VE BABAM
HÜSEYİN VE GÖNÜL GÜLSÜNER'e

Acknowledgement

I would like to thank and express my sincerest gratitude to my supervisor Tayfun Özçelik for his guidance, encouragement, and support throughout my thesis work. I have learned a lot from his scientific and personal advices.

It is a pleasure to express my gratitudes to Iclal Özçelik for her kind support and motivated ambiance.

It is my pleasure to express my thanks to Dr. Mary-Claire King and Dr. Nurten Akarsu for innumerable discussions, suggestions and cheerful ambiance always motivated me during our studies.

I wish to express my thanks to Dr. Ayse Begum Tekinay for her help in determining the expression patterns of *WDR81* in mouse brain.

I would like to thank Dr. Huseyin Boyacı and Katja Doerschner for their brain imaging studies.

I would also like to thank Dr. Murat Günel and Dr. Kaya Bilgüvar for Illumina sequencing experiments.

I would like to thank Dr. Salim Çıracı for providing access to computer resources.

I would like to thank Emre Onat, Merve Aydın and Gülşah Dal for their help in population screening studies.

I am indebted to Hilal Ünal for her persistant support, patience and confidence in me. I would also like to thank Hilal, for her invaluable contributions to the mouse studies.

I'd like to dedicate this thesis to my father Hüseyin and my mother Gönül Gülsüner, my sisters Eda and Gülnur, and my grandmother Ayse who always supported, encouraged and guided me. It is impossible to express my endless love and thanks to my family. I will forever be grateful to them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Quadrupedal Locomotion in Humans

The life of humankind begins with walking on all fours and continue with upright posture (bipedal locomotion).[1] Human bipedal locomotion is unique among living primates. Although comprehensive studies provided valuable information about walking and upright posture [1, 2, 3, 4, 5], mechanistic insights into the developmental processes and their genetic determinants are poorly understood. Functional brain imaging studies of healthy individuals suggested that, cerebral cortex, occipital cortex, basal ganglia and cerebellum are the crucial brain regions that control locomotor activities.[6, 7]

The cerebellum has a particular role in timing and controlling of complex patterns of muscle movements, and important for balance and bipedal locomotion. Developmental disorders of the cerebellum affecting gait and locomotion are rare and genetically heterogeneous. These clinical traits are characterized by loss of balance and motor coordination. Hereditary syndromes involving cerebellar abnormalities are of interest since they provide functional insights into regulation, maintenance and organization of the motor system.[8] Cerebellar ataxia, mental retardation, and disequilibrium syndrome (CAMRQ) is a genetically heterogeneous disease characterized by cerebellar hypoplasia, mental retardation,

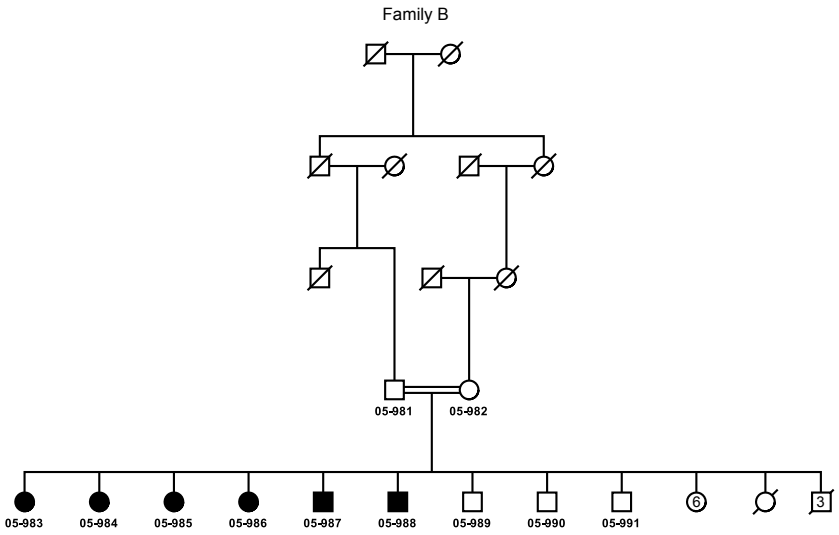


Figure 1.1: Pedigree of family B. Six of the 19 children of a first cousin marriage are affected by CAMRQ2.

dysarthric speech and quadrupedal locomotion. The syndrome was first described by Tan [9] in a large consanguineous family named family B (Figure 1.1). The first family was initially studied and the disease locus was mapped to a 7.1 Mb region on chromosome 17p (CAMRQ2, MIM 613227).[10] Subsequent analysis of the family using next generation sequencing of the entire disease locus, conservation analysis, presence of polymorphic stop codons in the variant sites, and genotyping of the control group and ethnically matched healthy individuals lead to the discovery of a missense mutation in WD repeat protein 81 (*WDR81*).[11]

1.1.1 Families described in Turkey

The first family described in the literature with quadrupedal gait is Family B. It consists of six affected and 13 healthy children who are the product of a consanguineous marriage (Figure 1.1).[9] Family B lives in south-eastern Turkey. The main characteristics of the affected individuals are cerebellar hypoplasia, dysarthric speech, mental retardation, truncal ataxia and quadrupedal locomotion. The pedigree shows an autosomal recessive mode of inheritance. The index

case is now a 33 years old man. He is mentally retarded based on Mini Mental State Examination Test. Another brother and a sister are bipedal and they have similar neurological findings with a milder degree of mental retardation. All affected individuals can move around freely. In resting state, patients are able to stand upright, but they quickly returned to the quadrupedal position for walking. Affected individuals have dysarthric speech with limited vocabulary. They can understand and respond to simple questions in their own language and can express their basic needs.[9, 10]

During the course of our studies three additional families have been identified in Turkey with slightly different clinical characteristics (Table 1). Family A is a consanguineous family from south-eastern Turkey with seven affected individuals (Figure 1.2). Family D is another consanguineous family from western Turkey with 3 affected individuals (Figure 1.3).[12, 13] Initial genetic analysis identified a nonsense mutation (p.R257X) and a single nucleotide deletion (c.2339delT) in Very Low Density Lipoprotein Receptor (*VLDLR*) gene in these families respectively (CAMRQ1 [MIM 224050]).[12]

Another consanguineous family affected by CAMRQ is residing in southern Turkey (Family C). The disease is observed in four individuals in three branches of the pedigree (Figure 1.4).[14] One of the affected individuals is bipedal ataxic albeit exhibiting quadrupedal locomotion in childhood. The other affected individual was also a quadruped, but he completely lost his ambulation now. Remaining two affected individuals exhibit quadrupedal locomotion. Both affected individuals had severe mental retardation and mild cerebellar hypoplasia. The initial genetic mapping studies and sequence analysis exclude the previously reported gene loci in this family.[15]

1.2 Cerebellum and Motor Coordination

In the last million years of evolution the size of the cerebellum and cerebral cortex dramatically increased. Through out evolution, cerebellum enable both

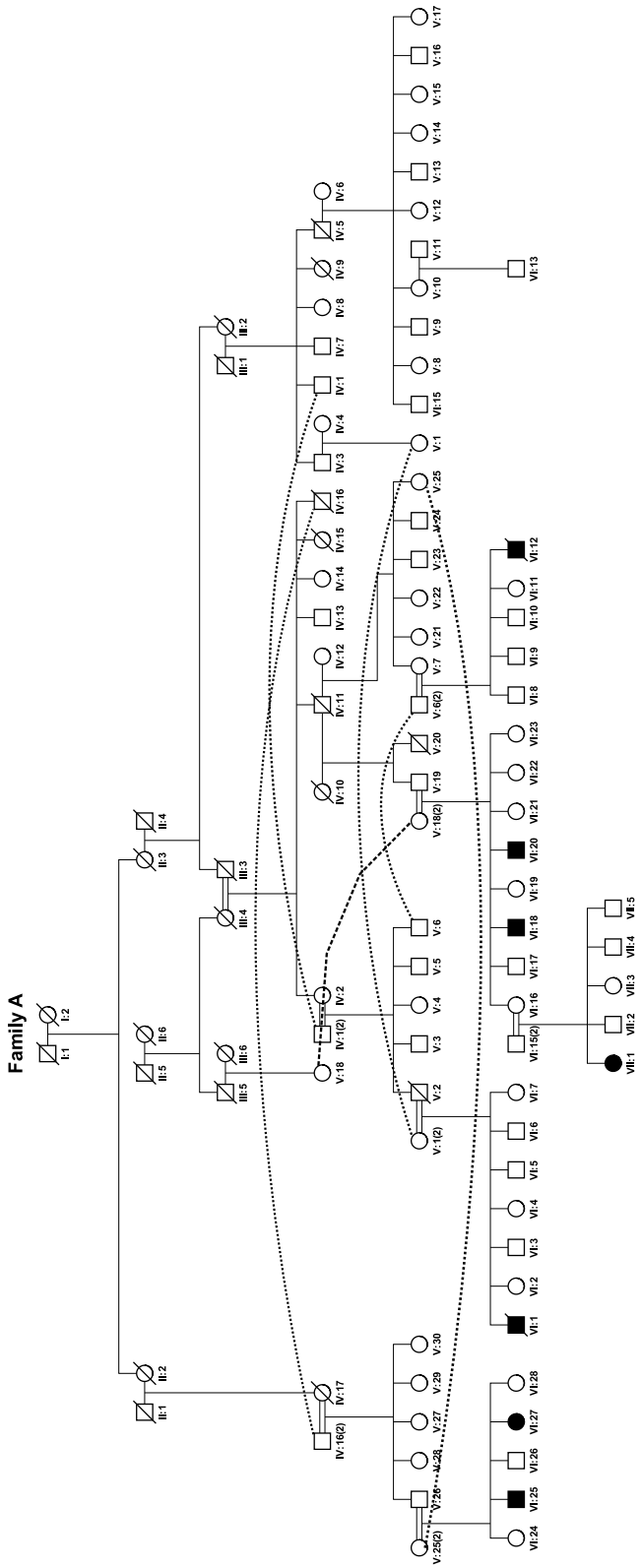


Figure 1.2: Pedigree of family A.

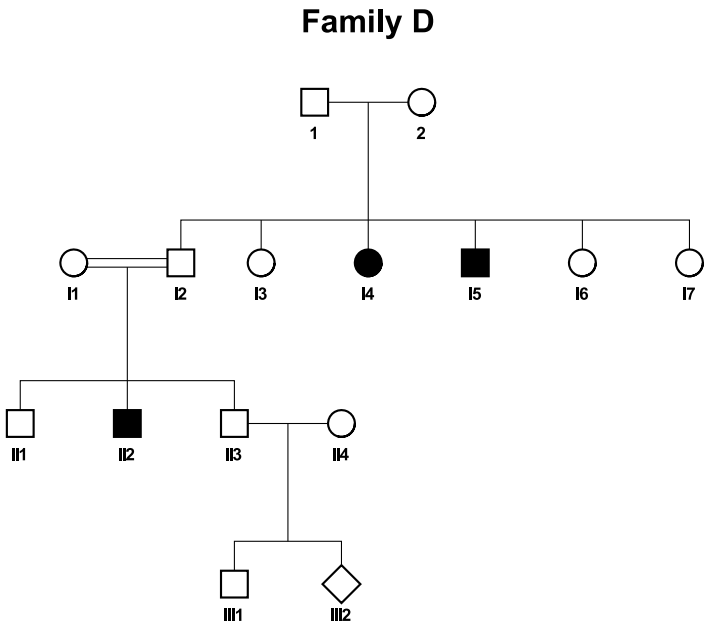


Figure 1.3: Pedigree of family D.

Table 1.1: Clinical characteristics of families.[12]

	Family A	Family B	Family C	Family D
Locus	9p24	17p	Not 9p or 17p	9p24
Gene	*VLDLR*	*WDR81*	–	*VLDLR*
Gait	Quadrupedal	Quadrupedal	Quadrupedal	Quadrupedal
Speech	Dysarthric	Dysarthric	Dysarthric	Dysarthric
Hyptonia	–	–	–	–
B.c.n.	Normal	Cvs defect	Pvs defect	Not done
M.r.	Profound	Severe to profound	Profound	Profound
Ambulation	Delayed	Delayed	Delayed	Delayed
T. ataxia	Severe	Severe	Severe	Severe
Low. leg ref.	Hyperactive	Hyperactive	Hyperactive	Hyperactive
Up. ext. ref.	Vivid	Vivid	Vivid	Vivid
Tremor	Very rare	Mild	Present	Absent
Pes-planus	Present	Present	Present	Present
Seizures	Very rare	Rare	Rare	Absent
Strabismus	Present	Present	Present	Present
Inf. cereb.	Hypoplasia	Hypoplasia	Mild hypoplasia	Hypoplasia
Inf. vermis	Absent	Absent	Normal	Absent
Cort. gyri	M.s.	M.s.	M.s.	M.s.
Corp. col.	Normal	Reduced	Normal	Normal

Abbreviations used in this table: B.c.n., barany caloric nystagmus; M.r., mental retardation; T. ataxia, truncal ataxia; Low. leg ref., lower leg reflexes; Up. ext. ref., Upper extremity reflexes; Inf. cereb., inferior cerebellum; Cor., cortical; Corp. col., corpus callosum; Cvs, Central vestibular system; Pvs, Peripheral vestibular system.

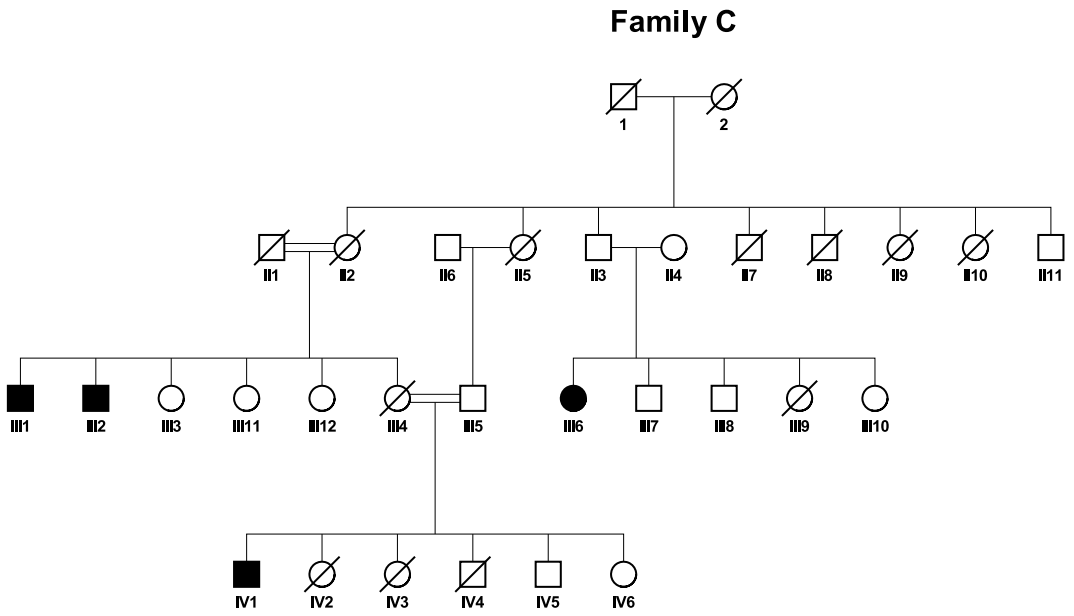


Figure 1.4: Pedigree of family C.

motor and mental capabilities, which provided humans great advantages in adaptation. Another adaptive advantage could be the possible role of the cerebellum in combining motor function of articulation to the mental functions that controls language and speak.[16] However, the hypothesis of the linguistic processing by the cerebellum has not been yet elucidated. Most of the knowledge regarding the functions of the cerebellum has been obtained from damaged cerebellar structures and brain imaging studies. Recent improvements in mouse genetics, brain imaging techniques, and genomic approaches have led to the identification of several genes underlying human cerebellar malformations.[17] Hence, those studies have accelerated our understanding of the development and functions of the cerebellum.

Cerebellum is essential for normal motor function. It coordinates timing, progression, and intensity of motor movements. It plays critical roles, especially in rapid movements like typing, running and talking. Complete loss of the cerebellum does not cause complete loss of the muscle movements, but total incoordination of these activities. Functionally, cerebellum is placed in a central position between the output signals of motor system to peripheral muscles and

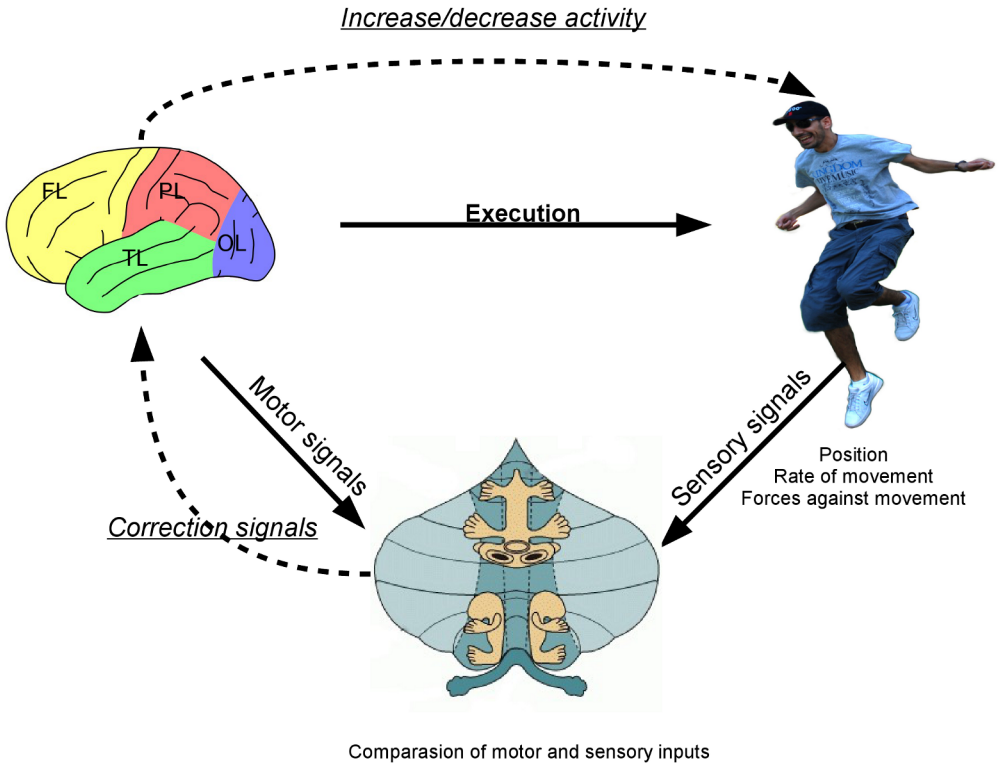


Figure 1.5: Adjustment of motor system. (The picture of the cerebellum in this figure is taken and modified from Purves et. al [19] with permission).

input signals from peripheral organs to central nervous system (Figure 1.5). Sensory signals such as, rate of movement, forces that act on the movement and the current position of the extremities send information to cerebellum. In addition, cerebellum compares two signals; motor signals of the desired movement and sensory signals from actualized movement. If the result is not favoured, cerebellum sends correction signals to the motor system, and motor system replies by increasing or decreasing the activation of targeted muscles.[18] The cerebellum also plans the next sequential movement, and directs the cerebral cortex while the first movement is still active.

1.2.1 Anatomical and functional areas of the cerebellum

Cerebellum is located in the bottom of the brain, and is divided into three anatomical regions: the anterior lobe, the posterior lobe and the fluccolonoduler lobe. These lobes divide the organ from top to bottom in humans. The fluccolonoduler lobe is often called as *vestibulocerebellum* and has distinct functions and connections. It controls the body equilibrium together with the vestibular system.[20] Remaining anterior and posterior lobes are functionally organized along the longitudinal axis. From the posteroinferior view, cerebellum is divided into two hemispheres by a narrow strip called *vermis*. Each side of the vermis consists of two large cerebellar hemispheres, and each hemisphere is divided into intermediate and lateral zones (Figure 1.6). As in other brain regions, a topographical body is represented in the cerebellum. Axial body, hips, shoulders and necks lie in the vermis. The distal part of the body -limbs, hands and feet- lie in the intermediate zones of the cerebellar hemispheres. These regions receive afferent nerve signals from the respective body parts and corresponding topographical motor areas of the brain. The intermediate zone of the hemispheres are also known as *spinocerebellum* and responsible for the adjustment of body and limb movements.[18] The lateral zone does not represent any topographical body section. It constitutes the *cerebrocerebellum* and this region receives signals from the cerebral cortex and sends output mainly to ventrolateral thalamus and red nuclei. It plays crucial roles in planning the next sequential movements.[21]

1.2.2 Cellular components

Cerebellar cell types, cell layers, axon types and tracts are the major components of the cerebellum. There are three major types of cell in the cerebellum; Purkinje cells, deep nuclei cells and granule cells. One purkinje cell and a deep nucleus constitutes a functional unit. Cerebellar cortex has three cell layers; the molecular layer, Purkinje cell layer and granule cell layer. The cerebellar deep nuclei, which sends the output signals to the nervous system are located under these cortical layers (Figure 1.7). Mossy fibers and the climbing fibers are the major tracts of

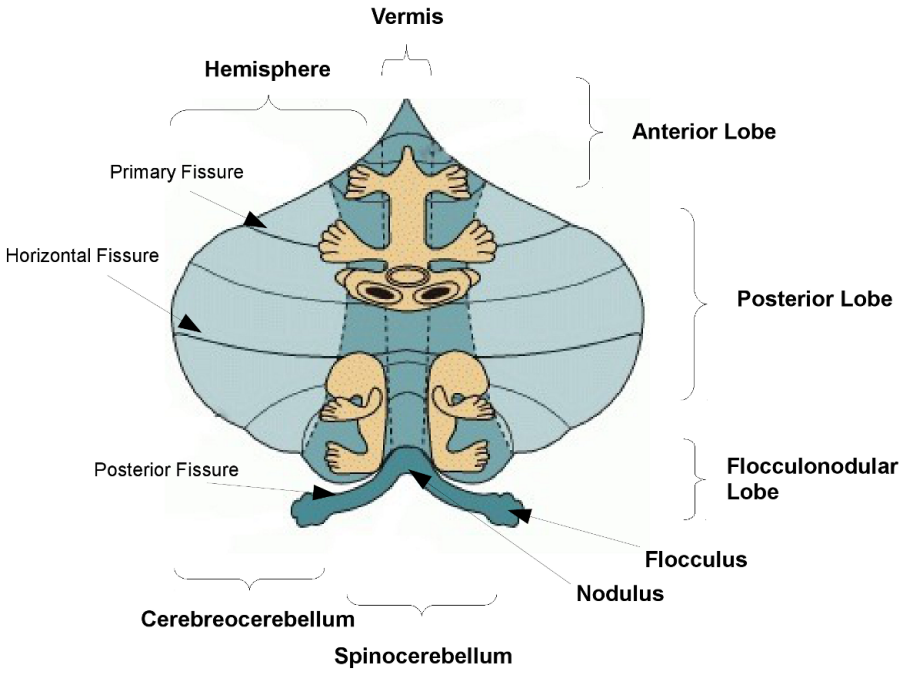


Figure 1.6: Functional and anatomical parts of the cerebellum. (The picture of cerebellum in this figure is taken and modified from Purves et. al [19] with permission).

cerebellar circuitry.[18]

Cerebellar cortex consists of nearly 30 million functional units. Deep nuclei cells are under the influence of both excitatory and inhibitory signals. The excitatory signals arise from the connections of afferent fibers coming from different brain regions and body parts. Mossy fibers and climbing fibers are the two types of afferent inputs. Climbing fibers arise from inferior olivaris of medulla. They make contacts with deep nuclei cells and extend to the molecular layer of cerebellar cortex to make synapses with dentrites and soma of Purkinje cells. Each climbing fiber makes contact with five to ten Purkinje cells. The mossy fibers arise from all other fibers that come from multiple sources and make contact with deep nuclei cells. In contrast to climbing fibers, mossy fibers reach to the granule layer and make synapses with thousands of granule cells. These granule cells send small axons to molecular layer and they divide in two branches to constitute parallel fibers. In the molecular layer, thousands of parallel fibers make synapses with Purkinje cells.[18, 20]

As summarized in Figure 1.7 deep nuclei cells are always activated by climbing and mossy fibers. Purkinje cells are also activated by those fibers and send inhibitory signals to deep nuclei cells. When motor system is activated, the deep nuclei cells are inhibited by Purkinje cells after a short delay with a negative feedback signal. This negative feedback mechanism prevent the muscle movement from over reaction and oscillation.

Besides from Purkinje, deep nuclei and granular cells, there are three types of additional cells located in the cerebellum. These are the stellate and basket cells in the molecular layer, and Golgi cells under the parallele fibers. All three types of cells are activated by parallele fibers. Stellate and basket cells then send inhibitory signals to Purkinje cells by lateral inhibition to sharpen the signals. In contrast, golgi cells inhibit granule cells, and provide fine-tuning of movements and prevents errors.

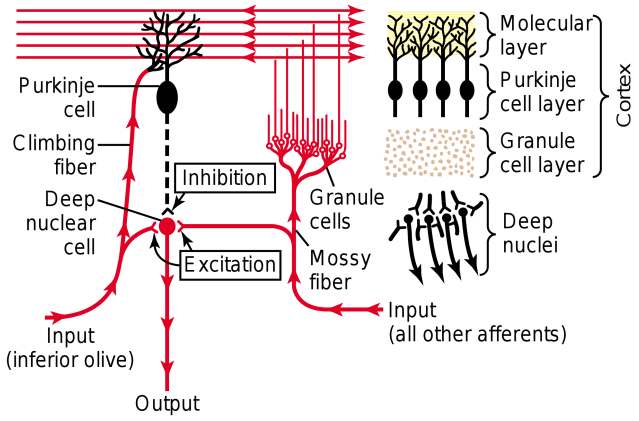


Figure 1.7: Neuronal circuit of the cerebellum. (Copyright Elsevier 2011. From Guyton et. al. 2006 [18] with permission).

1.2.3 Neuronal circuits of the cerebellum

As mentioned above, the major functional units of the cerebellum are Purkinje and deep nuclei cells. They receive input signals from various sources. In cerebellum, input signals arise from two major sources: 1) Afferent pathways from different brain regions and 2) Afferent sensory tracts from peripheral parts of the body. The first pathway activates deep nuclei cells and results in an increased output from these cells. Then, Purkinje cells balance the activation by negative feed-back. The peripheral sensory pathway provides information about position, rate of movement and the forces against movement. This information is used to adjust the movement by increasing and decreasing the activation signals from deep nuclei cells.

1.2.3.1 Afferent pathways from brain

There are four major pathways, which send signals from various brain regions to the cerebellum:

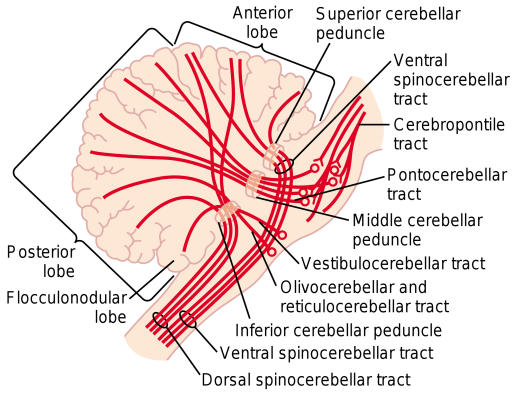


Figure 1.8: Principal afferent tracts to the cerebellum. (Copyright Elsevier 2011. From Guyton et. al. 2006 [18] with permission).

- The *Corticopontocerebellar* pathway originates from the cerebral motor, premotor and somatosensory cortices. They terminate in the lateral parts of the opposite cerebellar hemispheres by pontocerebellar tracts.
- The *olivocerebellar* fibers arise from cerebral motor cortex, basal ganglia, reticular formation and spinal cord. They reach to all parts of the cerebellum, after passing the inferior olive.
- The *vestibulocerebellar* tract originates from the vestibular apparatus and vestibular nuclei of the brain stem and terminate at the fluccolonoduler lobe.
- The *reticulocerebellar* fibers extend to the vermis from the reticular formation of the brain stem.

1.2.3.2 Afferent pathways from periphery

Two major tracts by which cerebellum receives sensory signals from peripheral body regions are the *dorsal* and the *ventral spinocerebellar* tracts (Figure 1.9). The most rapid conduction in nervous system occurs at the spinocerebellar tracts by which the cerebellum is informed of the rapid changes in the muscle movements instantaneously.

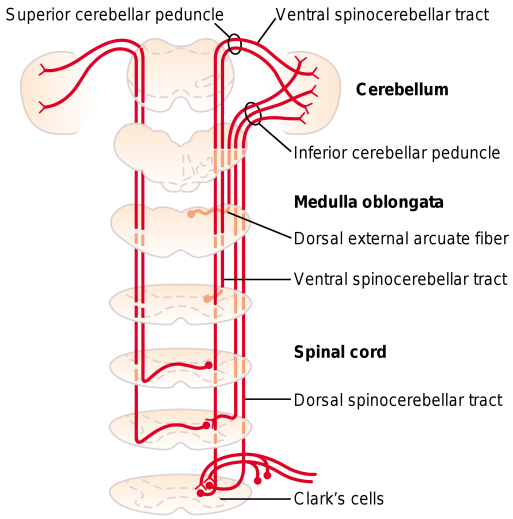


Figure 1.9: Spinocerebellar tracts. (Copyright Elsevier 2011. From Guyton et. al. 2006 [18] with permission).

- *Dorsal spinocerebellar* tract carries the signals that originate mainly from muscle spindles. Signals from tactile receptors of the skin, joint receptors, somatic receptors and golgi tendon organs are also transmitted through the dorsal spinocerebellar tract. The signals arise from these peripheral parts of the body to inform the cerebellum about the degree of muscle contraction and tension of the tendons, position of the body parts, rates of the movements, and the forces against the movements. Dorsal spinocerebellar tract reaches to the intermediate zones of the cerebellum and the vermis by passing through the inferior peduncle.
- *Ventral spinocerebellar* tract transmits information from the motor signals which arrive to the anterior horns of the spinal cord. The signals reach both sides of the cerebellum by passing through superior peduncle.

1.2.3.3 Efferent pathways

All efferent signals received by the cerebellum arrive at one of the cerebellar nuclei and its associated area in the cerebellar cortex. After a fraction of time, cerebellar

nucleus receives inhibitory signals from the cerebellar cortex as mentioned above. By this balancing system, adjusted output signals leave the cerebellum and spread through other regions of the brain. There are three major efferent pathways that originate from three different deep cerebellar nuclei (Figure 1.10).

- *Fastigial nucleus*: Signals from the vermis are transferred to the medulla and pons by passing through the fastigial nucleus. This pathway works together with vestibular nuclei and equilibrium apparatus to control body equilibrium. This circuit also helps coordinate the attitudes associated with posture of the body together with reticular formation.

- *Interposed nucleus*: The signals from intermediate regions of the hemispheres are transmitted to the ventroanterior and ventrolateral nuclei of the thalamus $\longrightarrow$ cerebral cortex $\longrightarrow$ midline structures of the thalamus $\longrightarrow$ basal ganglia $\longrightarrow$ reticular formation and red nuclei of the brain stem. This pathway mainly contributes to the coordination of the reciprocal contractions of the antagonist and agonist muscles in the peripheral parts of the extremities.

- *Dentate nucleus*: Cerebellum helps coordinate sequential movements through pathways arising from lateral hemispheres. These signals exit from the dentate nucleus and reach to the cerebral cortex by passing ventrolateral and ventroanterior nuclei of the thalamus.

1.3 Dysfunction of Cerebellum and Ataxias

The major consequences of cerebellar dysfunction are motor related and vary based on the localization of the damage.[20] One of the important aspects of cerebellar dysfunction is that, if the deep nuclei cells are not damaged, the undamaged parts of the cerebellum can compensate for motor functions in slow movements.[18] One of the most important symptoms of cerebellar diseases is ataxia.[20] The term is generally used to indicate uncoordinated movement. The

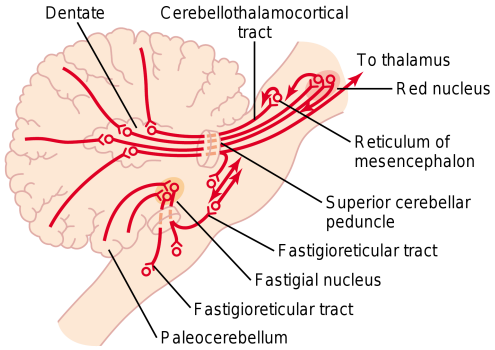


Figure 1.10: Efferent tracts to the cerebellum. (Copyright Elsevier 2011. From Guyton et. al. 2006 [18] with permission).

most common causes of ataxia are, damaging of cerebellar and spinal structures, atrophy of cells in the cerebellum or its connections and cerebellar degeneration.

The diseases accompanied by ataxia can be divided into three groups as i. acquired ataxias, ii. inherited ataxias and iii. degenerative ataxias. Acquired ataxias are mostly caused by traumatic events, strokes and intoxications. The most common form of toxic ataxia is caused by alcohol induced degeneration of the cerebellum.[22] Degenerative ataxias consist of multiple system atrophy (MSA) and idiopathic late-onset cerebellar ataxia (ILOCA).[22] MSA is the most common cause of the nonhereditary degenerative ataxia.[23]

Hereditary ataxias are caused by defective genes and follow the characteristics of Mendelian inheritance. These diseases are characterized by degeneration of the cerebellum and spinocerebellar tracts. Various symptoms in central and peripheral nervous system can accompany ataxias.[24]

Clinical presentation of different inherited ataxias caused by different genes may show significant similarities including histopathological findings. Therefore, it is difficult to make a classification based on clinical findings.[25] Thus, it is preferred to group inherited ataxias according to mode of inheritance as autosomal dominant, autosomal recessive, X-linked and mitochondrial.[26] Common genetic causes of autosomal recessive ataxias are summarized in Table 1.2

Table 1.2: Summary of common autosomal recessive ataxia syndromes.[22]

Type	Protein	Locus
Degenerative ataxias		
Friedreich ataxia	Frataxin	9q13
Charlevoix-Saguenay spastic ataxia	Sacsin	13q12
Mitochondrial recessive ataxic syndrome (MIRAS)	Polymerase γ	
Marinesco-Sjogren syndrome	SILI	5q32
Congenital ataxias		
Joubert syndrome (JBTS)	AHII, Nephrocystin-6	11p12, 12q21
Cayman ataxia	Cayataxin	19p13.3
Metabolic ataxias		
Ataxia with isolated vitamin E deficiency (AVED)	a-TTP	8q13
Refsum's disease	Phytanoyl-CoA hydroxylase	10pter-p11.2
Niemann-Pick type C	NPC1 protein	18q11-121
Ataxias with DNA repair defects		
Ataxia telangiectasia	ATM	11q22.3
Ataxia with oculomotor apraxia	Aprataxin, Senataxin	9p13, 9q34
Spinocerebellar ataxia with axonal neuropathy (SCAN1)	Tyrosyl-DNA phosphodiesterase	114q31

There is no specific treatment for ataxias. Current therapies generally aim to diminish disease severity.

1.4 Disease Gene Identification in Autosomal Recessive Disorders

Consanguinity has been recognized as a risk factor for rare diseases since the beginning of the previous century. Garrod reported notable excess of consanguineous marriages among the parents of alkaptonuria patients in 1902.[27] This phenomenon was later described with Mendelian inheritance. A consanguineous marriage is defined as the union of individuals descended from the same ancestor. Consanguinity refers to the amount of shared genetic material between individuals. The most common form of consanguineous marriage is between first degree cousins. In theory, such a couple shares 1/8 of their genetic material inherited from a common ancestor. Descendents of a consanguineous marriage inherit half of these shared alleles and 1/16 of their loci are homozygous. Therefore, assembly of two recessive disease alleles is more frequently observed in such families.

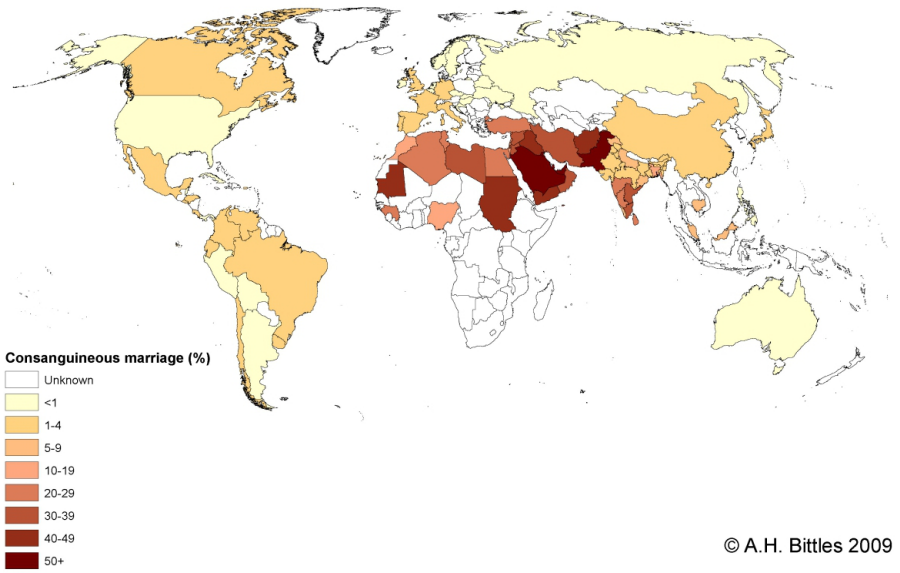


Figure 1.11: Prevalence of consanguineous marriages. (`http://www.consang.net`, with permission of A.H. Bittles).

Rate of consanguineous marriages in a population directly affects the frequency of recessive diseases. With the increasing levels of consanguineous marriages, the frequency of individuals affected by a recessive disorder is proportional to the frequency of disease allele. However, in randomly mating populations the frequency of affected individuals is proportional to the square of the frequency of the disease allele.[28] In some populations more than 50% of the marriages are consanguineous.[29] Highest levels of consanguineous marriages are observed in Suudi Arabia, Pakistan and southern and eastern rims of the Mediterranean basin (Figure 1.11, please see `http://consang.net` for further information).

In consanguineous families with recessive disorders, the regions adjacent to the disease causing mutation will preferentially be identical by descent leading to a stretch of homozygosity. Therefore, homozygosity mapping has the potential to identify the disease locus even in a single small consanguineous family.[30] In the presence of high density SNP genotyping data, detecting shared homozygous regions provides a powerful strategy.[31]

Table 1.3: Candidate gene prioritization tools.

GeneSeeker[32]	Expression and phenotypic data from mouse and human.	`http://www.cmbi.ru.nl/GeneSeeker`
Prioritizer[33]	Based on KEGG [34], Reactome [35], and HPRD [36] data.	`http://www.prioritizer.nl`
DGP[37]	Based on protein sequence properties (protein length, conservation, phylogenetic relationship and paralogy patterns).	`http://cgg.ebi.ac.uk/services/dgp`
PROSPECTR[38]	Based on sequence features (gene length, protein length, and homology).	`http://www.genetics.med.ed.ac.uk/prospectr`
SUSPECTS[39]	Combinatory approach using InterPro data, GO terms and expression data with the PROSPECTR classifier.	`http://www.genetics.med.ed.ac.uk/suspects`

The rate limiting step of disease gene identification studies lies between mapping of the disease locus and identification of the mutation. The disease locus could be several megabases long and contain hundreds of genes. Comprehensive analysis of the whole locus is time consuming and expensive. Publicly available data derived from experimental analysis and advances on bioinformatics tools enable to combine positional cloning and functional prediction approaches. Several tools have been developed to prioritize the candidate genes by their probability of involvement in a disease phenotype (Table 1.3).

However, candidate gene prioritization for the identification of the culprit gene could be complicated by the presence of hypothetical and/or uncharacterised genes in a given interval. Lack of functional information on a particular protein or pathway may further complicate the situation. In such cases a brute-force approach is required to identify the disease causing mutation. Availability of targeted capture of the genome and next generation sequencing approaches greatly facilitated disease gene identification studies.[40, 41, 42] On average, sequencing the whole exome of an individual yields 20,000 variants. More then 95% of those variants are known polymorphisms detected in healthy populations. The major challenge of this technology is pinpointing the disease causing variant among the many rare and neutral variants. Identifying the disease causing mutation among those variants depends on the estimation of the deleteriousness of variants. As

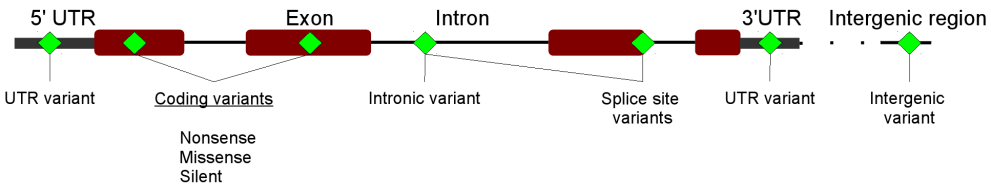


Figure 1.12: Classification of the variants according to functional regions.

in the candidate gene prioritization approach, several methods can be applied to prioritize causal variants.[43, 44]

Identification of a disease gene depends on several steps of filtering and stratification. First variants that are not compatible with the inheritance of the disease could be excluded. Segregation analysis of the variants in the family members and population screening are the primary steps for the exclusion of neutral variants. Identification of the same or different mutations in other families who are affected by the disease is the most powerful step. [44] However, in several recessive conditions the phenotypes could be extremely rare.[11] The second step in such instances is based on the stratification of the variants according to their functional impacts (Figure 1.12). By using publicly available curated databases, role of the genes in biological pathways or their interactions with genes that are known to cause similar phenotypes can be identified as in the candidate gene approaches (Table 1.3). In addition, several tools have been developed to predict the impact of the mutation. Most of these approaches use conservation as the measure of deleteriousness. The most preferable tools are summarized in Table 1.4.

1.5 Outline of the Thesis

In the next chapter, overview of the methods used in this study is given. The clinical description of the affected family and results of MRI studies were initially described in Chapter 3. Then, genetic analysis of the disease locus and results of the candidate gene sequencing approaches to find the culprit gene are discussed.

Table 1.4: Deleteriousness prediction tools for variants.

Tool	Method	Web address
DNA sequence based		
phastCons [45]	Evol	http://compgen.bscb.cornell.edu/phast/
GERP [46]	Evol	http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html
phyloP [47]	Evol	http://compgen.bscb.cornell.edu/phast/
Protein sequence based		
MAPP [48]	Evol, biochem	http://mendel.stanford.edu/SidowLab/downloads/MAPP/
PhD-SNP [49]	Evol, biochem	http://gpcr2.biocomp.unibo.it/~emidio/PhD-SNP/PhD-SNP_Help.html
SIFT [50]	Evol, biochem	http://sift.bii.a-star.edu.sg/
PANTHER [51]	Evol, biochem	http://www.pantherdb.org/
MutationTaster [52]	Evol, biochem, str	http://www.mutationtaster.org/
polyPhen [53]	Evol, biochem, str	http://genetics.bwh.harvard.edu/pph2/
SNAP [54]	Evol, biochem, str	http://www.rostlab.org/services/SNAP/
SNPs3D [55]	Evol, biochem, str	http://www.snps3d.org/

Abbreviations used in this table: Evol, Evolutionary; biochem, biochemical; str, structural.

In remaining parts of the Chapter 3, targeted next generation sequencing approach and initial identification and partial characterization of the disease gene will be described. Finally we conclude the thesis with Chapter 4.

Chapter 2

Materials and Methods

2.1 Recruitment of the Family and Control group

Family B from Turkey, afflicted by cerebellar ataxia, mental retardation, and disequilibrium syndrome (CAMRQ), was investigated in this study. Ancestors of the Family B migrated from a village of Syria to Turkey in early 1950s. Approximately 240 individuals spanning seven generations could be ascertained and blood samples were obtained from 177 individuals belonging to five generations.

Approximately 549 individuals, who were not affected and did not have any family history about movement disorders, were enrolled in the study as control group. Two cohorts were used to investigate presence of candidate mutations in the unrelated individuals with similar phenotypes. The first cohort consisted of 58 patients with cerebellar phenotypes with or without quadrupedal locomotion. The second cohort of 750 patients had structural cortical malformations or degenerative neurological disorders.

All clinical investigations performed were compatible with the Helsinki Declaration (`http://www.wma.net`). The study was approved by institutional review boards of Bilkent, Hacettepe, Başkent and Çukurova Universities (decisions BEK02, 28.08.2008; TBK08/4, 22.04.2008; KA07/47, 02.04.2007 and 21/3,

08.11.2005, respectively). Written informed consent forms were obtained from the all participants prior to the study.

Functional brain imaging studies were performed by National Research Center for Magnetic Resonance (UMRAM), Bilkent University, Ankara, Turkey. A total of 18 participants were subjected to the study. Six of the participants were from the affected family, including the carrier father (05-981), one female sibling who is homozygous for the wild type allele (10-033) and four affected siblings (05-984, 05-986, 05-987, 05-988). Remaining seven male and seven female individuals were age and sex matched controls without signs of any neurological phenotype and movement disorders. Prior to scanning, family members were sedated in order to eliminate movements. Sedation was performed by using intravenous midazolam (2 mg per subject), which was followed by propofol (120 mg) and fentanyl (50 mg).

2.2 DNA and RNA Samples

Peripheral blood samples of affected and healthy individuals were taken by venopuncture, collected in K3-EDTA containing tubes and transferred to the laboratory with cold chain conditions. They were separated into 1.5 µL aliquots in eppendorf tubes and stored at −80 ℃.

2.2.1 DNA isolation from blood samples

DNA were isolated from 200 µL of peripheral whole blood samples using Nucleospin^{TM} Blood Kit (Macherey-Nagel Inc., PA, USA) according to the protocol from the manufacturer. For the next generation sequencing experiments, Phenol-Chloroform DNA extraction method [56] was used to obtain high quality and high concentration of DNA. Quality and quantity of the DNA was measured by using NanoDrop^{TM} ND-1000 UV-Vis Spectrophotometer (NanoDrop, DE, USA) and PicoGreen method [57]. DNA quantities and qualities were verified by densitometric agarose gel electrophoresis.

2.2.2 RNA isolation and cDNA synthesis

RNA samples were isolated with Trizol reagent (Invitrogen, CA, USA) using 1 mL of fresh peripheral blood samples of the patients, the carriers and healthy individuals according to manufacturer's protocol. Commercially available total RNA from different human tissues were obtained to analyse the expression pattern of the disease causing gene (Clontech, CA, USA and Agilent Technologies, CA, USA). Qualities and concentrations of the RNA samples were measured by using NanoDrop^{TM} ND-1000 UV-Vis Spectrophotometer (NanoDrop, DE, USA) and Agilent Bioanalyzer 2100 (Agilent Technologies, CA, USA). cDNA was prepared from 1 µg DNaseI (Fermentas, NY, USA) digested RNA samples using First Strand cDNA Synthesis kit (MBI Fermentas, NY, USA). Random hexamer primers were used in cDNA synthesis reactions.

2.3 Genetic Mapping

2.3.1 Pedigree construction and analysis

Medical and familial histories of Family B were obtained from healthy members of the family. For the construction of an extended pedigree, 177 relatives of Family B were visited and family and medical histories were obtained from each individual. Pedigrees of each family were drawn on site. Haplopainter [58] and Inkscape (`http://inkscape.org/`) softwares were used to construct the extended pedigree. Pedigree analysis were performed according to characteristics of Mendelian diseases.[59]

2.3.2 Array based genotyping

Two obligate carrier parents, three healthy and six affected siblings were selected for whole-genome single nucleotide polymorphism (SNP) genotyping. GeneChip 10K Xba Affymetrix arrays (Affymetrix Inc, CA, USA) were used to genotype

11 individuals according to manufacturer's recommendations. Normalization of the raw data and genotype calling were performed by Affymetrix Genotype Console Software v2.1 (Affymetrix Inc, CA, USA). SNP genotype data were used to perform parametric linkage analysis and homozygosity mapping.

Furthermore, two affected individuals (05-984, 05-987) were genotyped using high resolution Illumina 300 Duo v2 BeadChip (Illumina Inc, CA, USA). Images were normalized and genotypes were called using Bead Studio (Illumina Inc, CA, USA). Illumina data were used to confirm homozygous regions and to calculate next generation data statistics as mentioned in sections 2.3.4 and 2.5.3.

Mendelian errors, missing genotype rates, sex and inbreeding statistics were calculated by using PLINK whole genome association analysis toolset v1.07.[60]

2.3.3 Genetic linkage analyses

Merlin V1.01 software [61] was used to perform parametric linkage analysis. PED, MAP and DAT files were generated from genotype data. PED files consisted of the pedigree structure and the genotype data of each individual. DAT files contains labels and orders of SNPs and the location of affection status in PED files. MAP files indicate the chromosomal locations and cM distance information of each marker. Contents of these files are summarized in Figure 2.1. A model file was prepared to input the inheritance model. Analysis was performed using standard Merlin parameters for parametric linkage analysis.

2.3.4 Homozygosity mapping and haplotype analysis

The flanking DNA regions harbouring the disease causing mutations segregate with the phenotype in the pedigrees of the families affected by monogenic diseases. Genetic linkage analysis is a powerful approach to identify the critical region segregating with the disease. However, significant linkage can be reached

MAP file

CHROMOSOME	MARKER	POSITION
1	SNP_A-1509443	3.28
1	SNP_A-1518557	8.61
1	SNP_A-1517286	11.17
1	SNP_A-1516024	11.31
1	SNP_A-1514538	11.7
1	SNP_A-1516403	11.94
1	SNP_A-1518687	13.28
1	SNP_A-1509959	13.32
1	SNP_A-1515791	14.1

DAT file

A	RARE_DISEASE
M	SNP_A-1509443
M	SNP_A-1518557
M	SNP_A-1517286
M	SNP_A-1516024
M	SNP_A-1514538
M	SNP_A-1516403
M	SNP_A-1518687
M	SNP_A-1509959
M	SNP_A-1515791

PED file

1	981	0	0	1	1	1	1	2	2	2	2
1	982	0	0	2	1	1	1	1	1	2	2
1	983	981	982	2	2	1	1	1	2	2	2
1	984	981	982	2	2	1	1	1	2	2	2
1	985	981	982	2	2	0	0	1	2	2	2
1	986	981	982	2	2	0	0	1	2	0	0
1	987	981	982	1	2	1	1	1	2	2	2
1	988	981	982	1	2	1	1	1	2	2	2
1	989	981	982	1	1	1	1	1	2	2	2
1	990	981	982	1	1	1	1	1	2	2	2
1	991	981	982	1	1	1	1	1	2	2	2

Pedigree structure

Genotype data

Affection (1: Unaffected, 2: Affected)

Sex (1: Male, 2: Female)

Figure 2.1: Contents of linkage files.

on the availability of large informative pedigrees. In an affected child of a consanguineous marriage, the flanking region spanning the disease causing mutation is almost homozygous by descent. In the presence of high density SNP genotyping data, detecting the shared homozygous regions provides a powerful strategy in the identification of the region which contains the disease causing mutation in a recessive condition.[31]

Homozygosity mapping analysis was performed to identify the shared homozygous regions in the affected individuals of the family B and to verify the candidate locus determined by linkage analysis. Since there were not any suitable software for homozygosity mapping, homozygous regions in 10K SNP data were detected by using a custom Perl script (algorithm summarized in Figure 2.2). In addition, regions were analysed with a spread sheet software visually. While 10K SNP microarrays contain only 10,204 SNPs with a mean inter-marker spacing of 258 Kb, a total of 318,238 SNPs with a mean inter-marker spacing of 8 Kb are printed on Illumina 300 Duo microarrays. Thus, two DNA of affected individuals were genotyped by using Illumina 300 Duo v2 BeadChip and the data were used to analyse unrepresented regions and to confirm the minimal critical region. Illumina SNP data was analysed by using HomozygosityMapper software [62]. Since it is expected to observe genotype errors in a high throughput microarray data and false heterozygotes can cause overlooking of homozygous blocks, analyses were repeated with different numbers of heterozygotes allowed in a homozygous window.

To verify and saturate the candidate locus, the family was further genotyped with polymorphic microsatellite markers (STR)(Appendix A). Then, the candidate region was visually analysed by constructing the haplotypes of the family with the genotype information obtained from STR analysis and SNP genotyping.

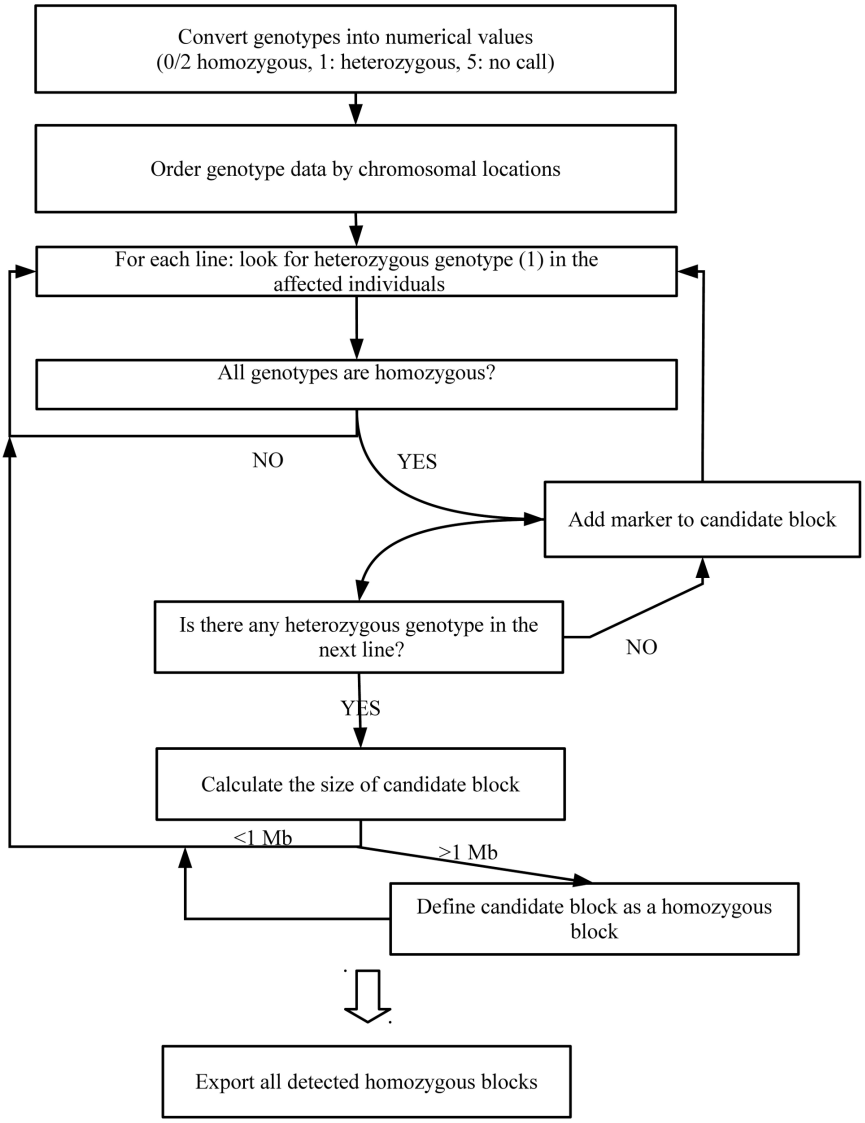


Figure 2.2: Homozygosity mapping algorithm. A Perl script was developed to detect shared homozygous regions in affected individuals.

2.4 Candidate Gene Analysis

2.4.1 Candidate gene prediction

The disease locus identified by linkage and homozygosity mapping analysis could contain a large number of genes. In the absence of high throughput sequencing technologies, selection and sequencing of the candidate genes within the disease interval is the the rate-limiting step of disease gene identification. Availability of experimental data and advances on bioinformatics tools enable to combine positional cloning and functional prediction approaches. The main aim of these approaches is prioritize the genes by their probability of involvement in a disease phenotype.

In an attempt to identify candidate genes in the minimal critical region, a combined bioinformatics approach was conducted and three different screening strategies were employed. These are systems biology techniques, bipartite distribution predictions and hybrid techniques.

- **Systems biology techniques:** *VLDLR* is the first gene implicated in CAMRQ. The protein encoded by *VLDLR* is involved in Reelin pathway as a receptor of Reln to regulate Dab1 tyrosine phosphorylation and microtubule function in neurons. Therefore, first, experimentally validated proteins that are interacting with VLDLR up to 3 connection levels were identified by using APID (Agile Protein Interaction DataAnalyzer)[63]. Second, corresponding genes of those proteins were fetched from ENSEMBL database (`http://www.ensembl.org`). Then, a subset o genes located in the critical region encoding the proteins which are interacting with VLDLR are determined. In addition, genes of the Reelin pathway were reviewed by using the GeneAssist^{TM} Pathway Atlas.
- **Bipartite distribution predictions:** All genes within the disease locus were analysed for their probability of being involved in a hereditary disease using two different tools. DGP tool (Disease Gene Prediction)[37], uses parameters -conservation, phylogenetic extent, protein length and paralogy-

that have been shown to follow specific trends in the already known disease genes. The second tool, Prospectr (PRiOrization by Sequence & Phylogenetic Extent of CandidaTe Regions)[64] ranks candidate genes with sequence features, sharing interpro domains, GO terms or similar expression profiles with any given gene or group of genes responsible for similar phenotypes. For both tools, *VLDLR* and other genes responsible for cerebellar hypoplasia were used as model genes.

- **Hybrid techniques:** The third tool, SUSPECTS prioritizes disease genes using a combination of genotype-phenotype mapping method based on disease-gene-associated keywords from InterPro and GO, and expression libraries. This tool uses the PROSPECTR Boolean classifier.[65]

2.4.1.1 Trinucleotide repeat containing genes

A variety of neurological disorders, including several forms of mental retardation, fragile X syndrome, Huntington's disease, Friedreich's ataxia and the inherited ataxias, have been characterized based on the presence of unstable expansions of trinucleotides.[66, 67] Hence, a bioinformatics approach was developed to predict the possible genes which have the probability of harbouring trinucleotide repeat expansion mutations.

For that purpose, previously reported disease causing nucleotide repeat mutations were determined (Table 2.1).[68, 69] Then, sequence data of 5' and 3' untranslated regions (UTR), exons, introns and 5' and 3' flanking regions of the genes within the disease locus was obtained from ENSEMBL database (`http://www.ensembl.org`). A Perl script was developed to scan and report the position information of the nucleotide motifs in the sequence data sets.

2.4.2 Mutation search

Selected candidate genes were screened by sequencing all coding regions by conventional Sanger method. Regions with nucleotide repeat motifs were amplified

Table 2.1: Genes and diseases with trinucleotide repeat expansion [Genes and diseases with trinucleotide repeat expansion.]

Gene	**Disease**	**Repeat**	**Repeat location**
AR	SBMA	CAG	Coding region
ARX	X-linked MR	GCG	Coding region
ATN1	DRPLA	CAG	Coding region
ATXN1	SCA1	CAG	Coding region
ATXN10	SCA10	ATTCT	Intron
ATXN2	SCA2	CAG	Coding region
ATXN3	SCA3	CAG	Coding region
ATXN7	SCA7	CAG	Coding region
ATXN8	SCA8	CTG / CAG	UTRs, coding region
CACNA1A	SCA6	CAG	Coding region
DM1	DM1	CTG	3'UTR
FMR1	FMR1	CGG	5'UTR
FMR1	FXTAS	CGG	5'UTR
FMR2	FMR2	GCC	5'UTR
FXN	FRDA	GAA	Intron
HTT	HD	CAG	Coding region
JPH3	HDL2	CTG	3'UTR, coding region
PABPN1	OPMD	GCG	Coding region
PPP2R2B	SCA12	CAG	Promoter, 5'UTR
TBP	SCA17	CAG	Coding region
ZNF9	DM2	CTG	Intron

Abbreviations of disease names: DM, myotonic dystrophy; DRPLA, dentatorubral-pallidoluysian atrophy; FMR, fragile X mental retardation syndrome; FRDA, Friedreich's ataxia; FXTAS, fragile X tremor ataxia syndrome; HD, Huntington's disease; HDL2, Huntington's disease-like 2; MR, mental retardation; OPMD, oculopharyngeal muscular dystrophy; SBMA, spinal and bulbar muscular atrophy; SCA, spinocerebellar ataxia.

and analysed by gel electrophoresis using the DNA of affected and healthy individuals.

2.4.2.1 Primers

PCR primers for all consensus splice sites and exons of the candidate genes were designed by using web based Primer3 software (`http://frodo.wi.mit.edu/`). All primers were verified by in-silico PCR (`http://genome.ucsc.edu/cgi-bin/hgPcr`) and BLAT (`http://genome.ucsc.edu/cgi-bin/hgBlat`) tools. Primers were purchased from Iontek Inc. (Istanbul, Turkey).

2.4.2.2 PCR conditions

75-150 ng of DNA samples were used as template for polymerase chain reaction (PCR). The optimal PCR conditions consisted of 2.5 µL PCR buffer (10X), 1.5 µL $MgCl_2$ (25 mM), 0.3 µL dNTPs (10mM), 1 µL (10pmol/µL) from each primer and 0.2 µL Taq DNA Polymerase (5U/µL) (MBI Fermentas, NY, USA). PCR reaction volumes were adjusted to 25 µL with ${}_{dd}H_20$ in standart reactions. When necessary, final volume was increased up to 50 µL. Reactions were performed in Techne^{TM} TC-512 thermal cycler. Reaction conditions were 5 min initial denaturation at 94 ℃, followed by 35 cycles of 94 ℃ for 30 sec, 56 ℃ - 64 ℃ for 30 sec and 72 ℃ for 30 sec and 5 min final extension at 72 ℃. PCR conditions were optimized when necessary by addition of PCR additives (DMSO, BSA) and by changing reaction conditions.

2.4.2.3 Agarose gel electrophoresis

Agarose (Basica LE, EU) was dissolved in 1X TAE buffer with a final percentage of 1%. 30 ng/ µL ethidium bromide was used as fluorescent tag. PCR products were mixed with 6X loading dye solution prior to loading onto agarose gel. pUC Mix Marker 8 and Mass Ruler DNA Ladder (MBI Fermentas, NY, USA) were

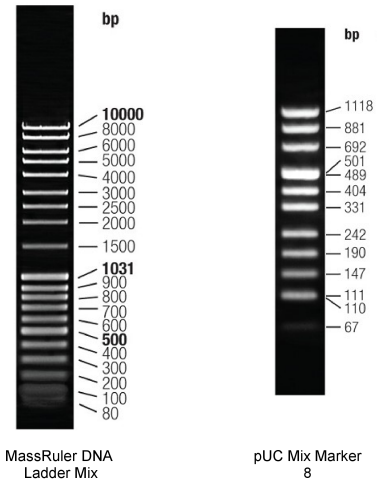


Figure 2.3: DNA Markers used in this study. Sizes of the fragments were shown. MassRuler DNA Ladder: 10 µL per lane, 1% agarose gel, 1X TAE 7 V/cm, 45 minutes. pUC Mix Marker 8: 0.5 µg per lane, 1.7% agarose gel, 1X TBE, 5 V/cm, 1.5 hours. (`http://www.fermentas.com/en/support/printed-media`).

used as DNA markers (Figure 2.3). Electrophoresis running times and voltages were determined according to the size of PCR amplicons. Gels were visually analysed and images were captured by using GelDoc imaging system (Bio-Rad, CA, USA) and MultiAnalyst software version 1.1 (Bio-Rad, CA, USA).

2.4.2.4 Polyacrylamide gel electrophoresis (PAGE)

10X TAE and 10% Ammonium persulfate was mixed with 30% acrylamide:bisacrylamide solution (29:1) and TEMED was added to the solution. 10 µL of each sample was loaded into gels. Gels were run at 15 W for 2-4 hours according to the length of the gel. Then, they were stained with EtBr for 10 min and destained in $_{dd}H_20$ for 5 min. Gels were visually analysed and imaged by using GelDoc imaging system (Bio-Rad, CA, USA) and MultiAnalyst software version 1.1 (Bio-Rad, CA, USA).

2.4.2.5 Sequencing reactions

PCR products were purified by using MinElute^{TM} 96 UF PCR Purification Kit (Qiagen, MD, USA). Purified products were sequenced using forward and reverse primers on an ABI 3130 XL capillary sequencing instrument (Applied Biosystems). Purification and sequencing steps were carried out by Refgen Corp. (Ankara, Turkey).

2.4.2.6 Data analysis

Raw sequence data files were obtained as AB1 sequence trace format. Each Sanger sequence trace file were aligned to the corresponding reference sequence and analysed by CLCBio Main Workbench (CLC bio, Denmark).

2.5 Targeted next generation sequencing

Next generation sequencing technologies were proceeded to use in 2007. This technology has not been used to identify a disease causing mutation until the period of this study. The disease locus contained more than 150 genes and candidate gene approaches were limited. Therefore sequencing the entire region was the best strategy to identify disease causing mutation. For this purpose, two different sequencing platforms, 454 GS FLX and Illumina Genome Analyser, were used.

2.5.1 Probe design and production

For 454 platform, critical region was targeted by NimbleGen 385K microarrays. 7,437 probes were designed to target the critical region by using the Sequence Search and Alignment by Hashing Algorithm (SSAHA)[70]. Due to the length of the region, a Perl script was developed in order to understand if all coding

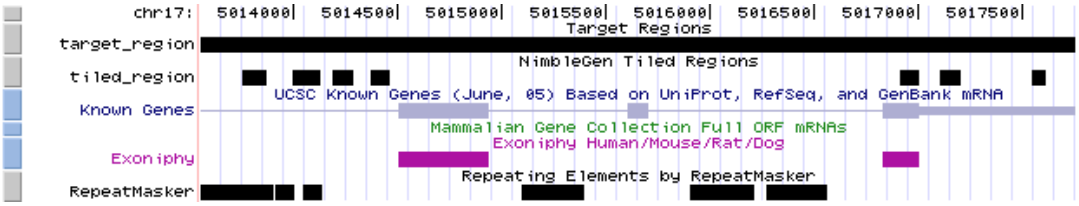


Figure 2.4: Locations of the probes on the targeted region. Probes were visually analysed by the Genome Browser (`http://genome.ucsc.edu`). An exon laying inside a gap and two exons partially covered in this region are shown.

regions were targeted by designed probe set. For this purpose, exon coordinates (n=1,848) of each alternative transcript variant were obtained from the ENSEMBL database. Regions within the start-end coordinates and 200 bp flanking sites of each probe were accepted as under coverage. Partially covered exons and completely non-covered exons were identified with the algorithm summarized in Figure 2.5. Probe coordinates were converted to BED format to visually analyse the probe locations by Genome Browser. As an example, Genome Broswer display of Exon 37 of USP6 gene, located in a 1,213bp gap was shown in Figure 2.4.

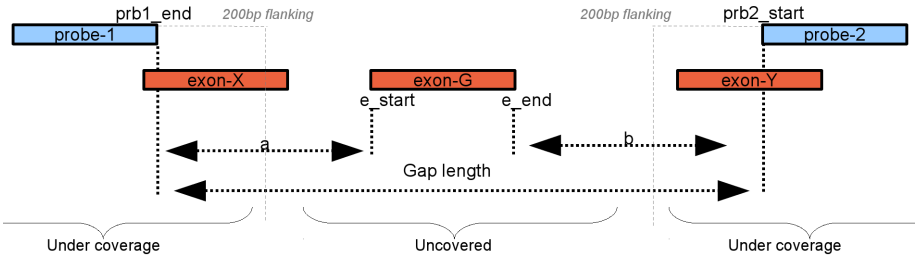
It was determined that, a total of 32 exons were located within the gaps. Eight exons were identified as completely not covered by the probe design (Table 2.2). Those regions were re-analysed by the SSAHA algorithm using less stringent parameters. Then, they were added to the chip design. Microarrays were produced with 7,464 unique probes and a total probe length of 4,853,455 bp targeting a 7.1 Mb region on chromosome 17p.

For Illumina sequencing, a total of 6,184,539 base pair long unique probes were designed using the SSAHA [70]. Probes were printed on a NimbleGen HD2 2.1M sequence capture microarray which targeted an extended region of 9 Mb spanning the disease locus.

Table 2.2: Exons located in the gaps between capture probes

Transcript	a	prb1_end	e_start	e_end	prb2_start	b	Gap
Partially covered exons							
AC087392.10-201	49	614,728	614,777	614,995	615,017	22	289
GLOD4-202	131	630,543	630,674	630,728	630,746	18	203
METT10D-201	235	2,304,970	2,305,205	2,305,243	2,305,418	175	448
AC015799.23-201	2	2,407,516	2,407,518	2,407,604	2,408,096	492	580
PAFAH1B1-201	140	2,511,868	2,512,008	2,512,036	2,512,100	64	232
ALOX15-001	432	4,487,829	4,488,261	4,488,421	4,488,545	124	716
USP6-203	169	4,969,671	4,969,840	4,969,957	4,970,193	236	522
USP6-206	134	4,971,389	4,971,523	4,971,658	4,971,707	49	318
USP6-205	141	4,977,321	4,977,462	4,977,569	4,977,629	60	308
USP6-206	230	4,980,429	4,980,659	4,980,753	4,980,785	32	356
USP6-201	6	4,985,438	4,985,444	4,985,518	4,985,710	192	272
USP6-204	35	4,989,389	4,989,424	4,989,598	4,990,222	624	833
USP6-206	653	4,989,389	4,990,042	4,990,193	4,990,222	29	833
USP6-206	54	4,991,048	4,991,102	4,991,213	4,991,284	71	236
USP6-205	682	4,991,893	4,992,575	4,992,786	4,992,926	140	1,033
USP6-205	406	5,011,545	5,011,951	5,012,142	5,012,303	161	758
USP6-204	37	5,014,461	5,014,498	5,014,934	5,016,911	1,977	2,450
ZNF594-201	10	5,026,052	5,026,062	5,026,122	5,026,559	437	507
ZNF594-203	12	5,026,052	5,026,064	5,026,142	5,026,559	417	507
ZNF594-203	114	5,026,052	5,026,166	5,026,167	5,026,559	392	507
C17orf87-201	636	5,064,328	5,064,964	5,065,073	5,065,091	18	763
RPAIN-201	5	5,267,809	5,267,814	5,267,818	5,268,446	628	637
AC004148.1-201	10	5,269,389	5,269,399	5,269,524	5,269,667	143	278
RPAIN-202	110	5,271,189	5,271,299	5,271,308	5,271,464	156	275
SLC13A5-201	207	6,539,443	6,539,650	6,539,659	6,539,705	46	262
Completely non-covered exons							
AC130689.8-201	230	1,707,746	1,707,976	1,708,188	1,708,436	248	690
PAFAH1B1-202	449	2,509,031	2,509,480	2,509,506	2,509,874	368	843
OR1D4-201	11,489	2,901,224	2,912,713	2,913,705	2,919,998	6,293	18,774
USP6-202	213	4,976,019	4,976,232	4,976,454	4,976,819	365	800
USP6-206	209	4,986,207	4,986,416	4,986,521	4,988,128	1,607	1,921
USP6-201	1,145	5,014,461	5,015,606	5,015,698	5,016,911	1,213	2,450
AC004148.1-201	348	5,273,948	5,274,296	5,274,302	5,274,524	222	576

Abbreviations of the table header are described in Figure 2.5.




```
If ( prb2_start – prb1_end > 200 ) :
      Selected as a gap.
      Are there any exons within this gap:

      if (    e_start > prb1_end  AND e_end   < prb2_start):

              if ( a > 200 and b > 200):

                    Exon-G (The exon is uncovered)

      if ( e_start < prb1_end AND
           e_end  – 200 > prb1_end AND
           e_end + 200 < prb2_start )

                    exon-X (Partially covered)

      if ( e_end > prb2_start AND
           e_start - 200 > prb1_end AND
           exon_start + 200 < probe-2_start )

              exon-Y (Covered by a probe)
```

Figure 2.5: Algorithm of the identification of the exons which are not covered by capture design. Exon-X is covered, exon-G in non-covered and exon-Y is partially covered in this example.

2.5.2 Sequence capture and sequencing

25 µg of DNA sample from each of the two obligate carrier parents (05-981, 05-982) and two affected siblings (05-985, 05-987) were captured with NimbleGen 385K microarrays. Sequence capture was performed by NimbleGen facility (Roche NimbleGen, WI, USA) according to manufacturer's protocol (`http://www.nimblegen.com/`). Captured DNA samples were subjected to standard procedures for 454 GS FLX sequencing with Titanium series reagents. Four full 454 GS FLX runs were conducted for each sample. Another affected individual (05-987) was captured by NimbleGen HD2 2.1M using 5 µg DNA. Captured DNA samples were subjected to standard procedures, and then sequenced by Illumina Genome Analyser IIx platform (`https://icom.illumina.com/`).

2.5.3 Sequence analysis

2.5.3.1 Mapping and annotation

SFF, FNA ve QUAL files were created from 454 GS FLX raw data. The sequence fragments obtained from next-generation sequence data is called "read". FNA file is a FASTA file consists of the each individual sequence read with a header line containing the code and length of the read. Corresponding quality score values for each base in the sequence reads are calculated with Phred basecalling algorithm [71] and are stored in the QUAL files (Figure 2.6). SFF (Standard Flowgram Format) file is the equivalent of the trace file of Sanger sequence data and contains information on the signal strength for each flow. First level analysis were performed using these files obtained from the sequence data of four individuals. First, sequence data was mapped to the hg18 reference human genome sequence by using the gsMapper module of Newbler software (454 Life Sciences, CT, USA) using standard genomic sequence mapping parameters. Variants were identified with ALLDiffs (All Differences) and more stringent HCDiffs (High-Confidence Differeneces) approaches by gsMapper.Ref [72] The criteria of the two approaches are:

- AllDiffs: There must be at least two non-duplicate reads that
    1. show the difference,
    2. have at least 5 bases on both sides of the difference,
    3. have few other isolated sequence differences in the read.
- HCDiffs: A difference is considered High-Confidence
    1. If at least 3 reads match the conditions listed in AllDiffs criteria,
    2. with at least one aligned in the forward direction and,
    3. at least one aligned in the reverse direction.

Annotations were performed using refGene table of UCSC Genome Browser (NCBI36/hg18) and novel variants were reported based on the SNPs included

```
>F0FQDYC01BUNW3 length=106 xy=0642_0773 region=1 run=R_2009_08_07_13_57_24_
ACATCACAGACACACACATACACACACACACAGACACACACACGCTCACACACACACACA
CAGAGACACGCACACACATCATATATACAGGCACACACAGGAGTAGAG
>F0FQDYC01DRODC length=51 xy=1428_1166 region=1 run=R_2009_08_07_13_57_24_
GTTTGGAAGTAGAGAGGGCTGCTCAGAGGAAGCGACGTCTGAGTTGGGTTT
>F0FQDYC01C7EUN length=45 xy=1197_1885 region=1 run=R_2009_08_07_13_57_24_
ACATGCATGCACACACACACACACAGCACACACTCTCACACTTAC
>F0FQDYC01CJRIC length=106 xy=0928_0370 region=1 run=R_2009_08_07_13_57_24_
TTTAAGTTGCTGAATCGTATTCCATCGAATGACTGTACTACAGTTTATTCACTCACCTAT
TGAAGGATATCTTGGTTGCTTCCAGTTGTTGACAATTTTGAATAAA
```

```
>F0FQDYC01BUNW3 length=106 xy=0642_0773 region=1 run=R_2009_08_07_13_57_24_
37 40 37 37 37 38 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 39 39 39 40 40
40 40 40 40 40 40 40 40 40 40 40 40 39 40 40 40 40 40 40 40 37 37 37 37
39 40 33 34 34 32 38 37 37 39 40 36 34 32 27 27 28 32 38 38 36 36 25 25 25
27 32 21 19 19 19
>F0FQDYC01DRODC length=51 xy=1428_1166 region=1 run=R_2009_08_07_13_57_24_
40 35 35 35 21 21 20 19 28 32 35 40 40 31 27 16 16 18 27 27 36 34 34 31 36
40 37 34 34 34 34 35 39 40 40 40 40 40 40 40 40 40 40 36 36 29 15 15 14 14
14
>F0FQDYC01C7EUN length=45 xy=1197_1885 region=1 run=R_2009_08_07_13_57_24_
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 28 24 24 24 37 35 23
20 18 18 18 30 36 33 30 32 29 29 36 38 38 37 32 23 22 22 30
>F0FQDYC01CJRIC length=106 xy=0928_0370 region=1 run=R_2009_08_07_13_57_24_
19 19 19 18 18 40 30 30 37 38 38 38 34 34 39 40 40 40 40 40 40 40 40 40
40 39 39 39 39 40 40 40 40 40 40 40 40 40 40 27 20 20 20 20 20 22 37 40
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 39 36 36 36
40 40 36 34 30 30 30 30 39 40 40 40 40 40 40 40 40 40 34 34 28 28 28 28 36
36 36 40 33 33 33
```

Figure 2.6: Contents of FNA and QUAL files. FNA files are fasta files containing sequence read and information of each read. The quality scores of each bases are stored at the QUAL files.

in the reference SNP129 database. Identified variants were saved in the AllDiff.txt and HCDiff.txt files (Figure 2.7). Variant informations were extracted from AllDiff and HCDiff files by using UNIX command line tool "grep" and saved as tab-delimited text files. Coordinate information data were further used to annotate variants by using custom annotation pipeline. Alignments were analysed visually by a text editor. Sequence statistics for 454 data was provided by Newbler software. All analysis were repeated with latest version of genome assembly upon its availability.

Illumina sequence data was mapped to the human genome by using two different aligner. Maq performs ungapped alignment of single-end reads [73] and was used to align sequence data to reference genome for single nucleotide variant (SNV) detection. For detection of small insertion and deletions, a gapped global alignment tool -BWA [74]- was used to align sequence data.

```
>chr17  1577570 1577570 C       T       40      97%

 Reads with Difference:
 chr17                  1577548+ -AGGCCTGTCTCCCA-GGGCC-T-GC-CCCC-ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 1577601
                                                          *****
 F0FQDYC01D19XL              30- -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCC                          1
 F0FQDYC01AUCEP             354+ -AGGCCTGTCTCC-A-GGGCC-T-GCTCC---ACC-T                        381
 F0FQDYC01DMQ9K  (2)         39- -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGC               1
 F0FQDYC01BXNYN             467+ -AGGCCTGTCTCC-A-GG-CC-T-GCTCCC--ACC-T-GCCCAAGCCAG-T          506
 F0FQDYC02HEPII  (2)        377+ -AGGCCTGTCTCCCA-GGGCC-T-GCTCC---ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 429
 F0FQDYC01CCRBV              68- -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 15
 F0FQDYC01DE4XH  (3)         80- -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 27
 F0FQDYC01DVYWM             100- -AGGCCTGTCTCCCAAGGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 46
 F0FQDYC01A6M1G             113- -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 60
 F0FQDYC01ER3N5             328+ -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACC-T-GCCCAAGCCAGCTTCTCAGCCCC 380
 F0FQDYC01BB9C5             191- -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 138
 F0FQDYC01CSB25             212- -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 159
 F0FQDYC02F3WMJ             265+ -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 318
 F0FQDYC02JZXM6             238+ -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 291
 F0FQDYC01BQ3DP             230+ -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 283
 F0FQDYC02H5LC3  (2)        259- -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 206
 F0FQDYC02FELXN  (2)        276- -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 223
 F0FQDYC02JUS63             184+ -AGGCCTGTCTCCCA-GGGCCCT-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 238
 F0FQDYC02GOFEB  (3)        179+ -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACC-T-GCCCAAGCCAGCTTCTCAGCCCC 231
 F0FQDYC02GSTNM             170+ -AGGCCTGTCTCCCA-GGGCC-T-GCTCC---ACC-T-GCCCAAGCCAGCTTCTCAGCCCC 221
 F0FQDYC02II68K             162+ -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC-G-CCCT-GCCCAAGCCAGCTTCTCAGCCCC 215
 F0FQDYC02GZUT6             178- -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 125
 F0FQDYC02FLUZN  (2)        152+ -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 205
 F0FQDYC01ANQ8J             150+ -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 203
 F0FQDYC02HBEYK             142+ -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 195
 F0FQDYC01EGF27             187- -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 134
 F0FQDYC01C0T8K              33- -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GC                      1
 F0FQDYC02HH4N2             115+ -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 168
 F0FQDYC01DYVWJ             104+ -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 157
 F0FQDYC02HAYSG              96+ -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 149
 F0FQDYC02HV1ZL  (2)        336- CAGGCCTGTCTCCCA-GGGCC-T-GCTCC---ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 283
 F0FQDYC02IDJ4E  (2)        406- -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 353
 F0FQDYC01BVZLI  (2)         82+ -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 135
 F0FQDYC02J2TB3             317- -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 264
 F0FQDYC02H862P              53+ -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 106
 F0FQDYC01AE7I1              35+ -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCC- 87
 F0FQDYC01CXEA5              23+ -AGGCCTGTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 76
 F0FQDYC02FMQMW             353-        GTCTCCCA-GGGCC-T-GCTCCC--ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 306
 F0FQDYC01DZQ2V             403-                    CC-TCGCTCC---ACCCTCGCCCAAGCCAGCTTCTCAGCCC- 367
                                                          *****

 Other Reads:
                                                          *****
 F0FQDYC02H2ND2             149+ -AGGCCTGTCTCCCA-GGGCC-T-GC-CCCC-ACCCT-GCCCAAGCCAGCTTCTCAGCCCC 202
                                                          *****
```

Figure 2.7: Representation of a variant in a diff file. In this figure, a header line starting with ">" and alignment of the region is shown. The position of the variant is chr17:1,577,570. This variant is detected in 97% of 40 the reads.

Analysis of the alignment data and identification of the variants were performed by using Samtools.[75] Annotation of the Illumina variants was performed by using annotation pipeline of Gunel's group, Yale University.

Alignment data of each platform was converted to tab-delimited IGV files to analyse visually by using Integrative Genomics Viewer (IGV).[76]

Fold enrichment of the targeted region was calculated with the formula $\frac{\sum \frac{REMTrm}{STrm}}{\sum \frac{RMG}{SG}}$ as described previously (REMTrm: Number of reads mapped to target region, STrm: size of target region, RMG: number of reads mapped outside of the target region, SG: size of human genome).[77] Coordinates of the variants were converted to hg19 genome assembly to re-analyse using updated annotation

datasets. Mapping analysis were repeated using hg19 version of genome assembly upon availability.

2.5.3.2 Coverage calculations

To reveal non-covered functional regions and possible deletions, functional coverage statistics were calculated by the following steps.

- For 454 data sequence read depths of each base were extracted from alignment info files of four individuals.
- For Illumina data, sequence read depths were obtained from alignment files by using pileup function of Samtools.
- For each base position mean values of the total read depths were calculated for the sequence data of three affected individuals.
- Bases were classified as non-covered ($<$2X mean read depth), low-coverage (2-3X mean read depth) and covered ($\geq$ 4X).
- Using the start-end coordinates of exons and untranslated regions of each transcript, non-covered and low-coverage regions were annotated.
- A total statistics indicating the coverage percentages of each functional genomic unit was calculated.
- To reveal if a non-covered region does actually indicate a pathogenic deletion, alignments of heterozygous parents were analysed.
- Non-covered and low-coverage regions were further analysed by alignment analysis of the next generation sequence data and Sanger sequencing to reveal any possible mutations in these gaps.

2.5.3.3 Genotype calling thresholds

In autosomal recessive disease gene identification studies, determination of the genotype status of the variants is one of the critical steps for exclusion of the variants by segregation analysis. Thus, optimization of the heterozygosity thresholds is crucial. As a first level genotype designation, a rigid heterozygosity threshold (30-70%) was used to maximize sensitivity for detection of homozygous variants. Using this threshold, variants which were detected as more than 70% of the reads were called as homozygous, and the variants which were detected as between 30-70% of the reads were classified as heterozygous. A total of 1004 SNPs within the targeted region were represented on Illumina 300 Duo v2 SNP microarrays. To optimize genotype calling thresholds, genotypes of the 1004 Illumina SNPs were compared with the Illumina 300 Duo v2 SNP microarray data and error rates were calculated in the genotype data obtained by different heterozygosity thresholds.

Mendelian errors (ie. AA x AA = AB) could be another method to optimize genotype calling parameters. For this purpose, first, genotypes of individuals were determined using different heterozygosity thresholds (ie: 30-70%, 20-80%, 10-90%, 30-90%) and variants were converted and stored as PED and MAP files. Then, Mendelian error rates were calculated by using PLINK software for each threshold.

2.5.3.4 Functional classification

An annotation pipeline was developed to further analyse functional consequences of the intronic and intergenic variants in terms of hypothetical genes and splicing variants. ENSEMBL GENES and VARIATION tables for hg18 human genome assembly were extracted from ENSEMBL54 database by using MartView interface of BIOMART data-mining tool (`http://www.ensembl.org`) to build a custom database. A Perl script was developed to annotate the variants using the custom annotation database. The annotated data were merged with the variant data annotated by Newbler software using refGene table.

Novel variants were filtered and then, variants in consistent with the Mendelian transmission of disease allele were selected (ie. variants detected as heterozygous in parents and homozygous in affected sibling). The selected variants were binned into 4 groups as i. transcribed and consensus splice site, ii. 5' & 3'UTRs, iii. intronic and iv. all remaining variants.

SNPs which were detected as homozygous in carrier parents were excluded. For the remaining SNPs which were compatible with the Mendelian transmission of disease allele, population frequencies were obtained from public databases. Variants which have not been detected as homozygous in healthy population were considered as a potential disease causing mutation.

2.6 Identification of disease causing mutation

2.6.1 Segregation analysis

Each functional novel variant was verified by Sanger sequencing in two affected and two carrier individuals. To reveal the segregation status, all variants were genotyped in all family members by using the appropriate genotyping methods (Table 2.3) as described below.

2.6.2 Population screening

Functional variants were further genotyped in control groups in order to exclude rare polymorphisms from disease causing mutation. For this purpose, three control groups were used: i. 214 unrelated healthy controls (428 chromosomes), 50 of whom were sampled from the same region of Turkey as Family B, ii. independent series of 400 individuals of various European and Middle Eastern ancestries, iii. 177 members from the kindred of family-B spanning 5 generations. Genotyping experiments were performed using restriction digestion or allele specific PCR assays.

The variants were further analysed in 1000 genomes (`http://www.1000genomes.org`) and ESV datasets (`http://evs.gs.washington.edu/EVS/`). 1000genomes data were downloaded from the ftp server of the project (`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/`). Genotype data were then analysed by VCF tools[78] and custom Perl scripts.

2.6.2.1 Restriction fragment length polymorphism

Restriction enzymes were determined by using NEBcutter.[79]. The regions containing the variants were amplified using the appropriate primers (Table 2.3) as described in 2.4.2.2. Restriction enzyme digestion was carried out with 5 µL of PCR products. 2 µL of restriction buffer, 1 Unit of restriction enzyme (MBI Fermentas, NY, USA) and 12.80 µL of $_{dd}H_20$ were mixed with PCR products. Restriction digestion reaction mixtures were incubated at 37℃, for 4 hours or overnight. 10 µL of each digestion product was analysed by 1.0-2.0% agarose gel.

2.6.2.2 Allele specific PCR

For the variants, which were not located inside a restriction enzyme recognition site, allele specific PCR (AS-PCR) primers were designed for wild-type and mutant alleles (Table 2.3). Standard multiplex PCR reactions were performed with the addition of a reference primer as internal control. PCR products were then analysed on 1.0-2.0% agarose gel electrophoresis.

2.7 Screening the candidate genes in disease cohorts

Candidate genes were further screened in different cohorts of patients with neurodevelopmental phenotypes for whom the genetic aetiology is unknown. The

Table 2.3: Primers and methods for genotyping of candidate variants

Locus	Forward	Reverse	Size	Method
Sequencing primers				
WDR81_V1	TGCCTACTCCACAGGGACA	CTGCCACAGAGCAAACACC	429	Sanger
MYBBP1A_V15	AGCCGCACCAAGACCATC	CAGCCCACATTCCACTCC	290	Sanger
ZNF594_361	GGTGTGACCTCTGGCTGAA	TCATTTGGCGCACAGCTT	426	Sanger
Genotyping primers and methods				
WDR81_WT	TGTCTCCCAGGGCCTGCC	GGACATGAGTGAGAGCACGA	250	AS-PCR
WDR81_MUT	TGTCTCCCAGGGCCTGCT	GGCAACAGGCTCAAACAGAT	301	AS-PCR
MYBBP1A_V15	AGCCGCACCAAGACCATC	CAGCCCACATTCCACTCC	290	HpaII RD
ZNF594_361	GGTGTGACCTCTGGCTGAA	TCATTTGGCGCACAGCTT	426	BglII RD

Abbreviations used in this table: RD, Restriction Digestion; AS-PCR, Allele specific polymerase chain reaction.

genes were sequenced in the first cohort consisted of 58 probands 12 had cerebellar hypoplasia (See Appendix A for primer lists). Another cohort consisted of 750 patients with structural cortical malformations or neurodegenerative diseases with unknown aetiology. In the second cohort, homozygous regions spanning the linkage interval were examined using the genotype data based on Illumina Human 370 Duo or 610 Quad BeadChips.

2.8 Functional characterization

2.8.1 Evidence of *WDR81* transcript

The mutation responsible for the disease was detected in the first exon of a predicted transcript of *WDR81*. To reveal the expression of the transcript and other possible transcript variants which may contain mutation site, an exon-spanning RT-PCR approach was performed. For this purpose, cDNA of human adult cerebellum was amplified using a forward primer 280 bp upstream from the mutation site and reverse primers for each of the exons (Appendix A).

2.8.2 Quantitative PCR analysis

The PCR primers flanking the mutation site were designed by Primer3 software (`http://frodo.wi.mit.edu/`) and were verified by in-silico PCR (`http://genome.ucsc.edu/cgi-bin/hgPcr`) and BLAT (`http://genome.ucsc.edu/cgi-bin/hgBlat`) tools. Primers were purchased from Iontek Inc. (Istanbul, Turkey). Primers for *WDR81* (WD40 repeat protein 81) were: F-5'-GCAAGCTGGACCAACTGTTT-3' and R-5'-GGGAAGTAGGGTGGGAAGG-3'. *GAPDH* (glyceraldehyde-3-phosphate dehydrogenase) was used as reference gene and primers were: F-5'-AGGTGAAGGTCGGAGTCAAC-3' and R-5'-GGGTCATTGATGGCAACA-3'. *GAPDH* primers were obtained from Dr. Önder Bozdoğan, Bilkent, Ankara.

Real-time quantitative PCR reactions were performed using SYBR Green Mix (Bio-Rad, CA, USA) according to the standard protocols. 25 µL of PCR reaction mixtures were prepared by 12.5 µL SYBR Green mix, 0.5 µL of each primer (10 pmol/µL), 6.5 µL $_{dd}H_20$ and 5 µL of 1:5 diluted cDNA.

Reactions were performed in triplicate with the following PCR conditions: 10 min initial denaturation at 94℃, followed by 40 cycles of 94℃ for 30 sec, 60℃ for 30 sec and 72℃ for 30 sec and 5 min final extension at 72℃. Melting curve analysis was performed following amplification step by raising temperature from 55℃ to 94℃ with 0.5℃ increments at each 15 seconds. Data were analysed using the Pfaffl method [80] and normalized according to the *GAPDH* expression. PCR products were further analysed on 1.5% agarose gel.

2.8.3 In situ hybridization

In order to reveal the expression pattern of the *Wdr81* in mouse brain, in situ hybridization experiments were performed. Experiments were carried out at Nanobiotechnology labs, UNAM, Turkey. The probes targeting the mutation region in human patient were prepared from mouse genomic DNA by amplification of the region. The PCR amplicons cloned into pCR®4-TOPO plasmids

(Invitrogen, CA, USA) for the antisense probes. To obtain sense probes, the same region was cloned into a modified pSK vector. Then, riboprobes were synthesized by using digoxigenin labeled NTPs. Mini Quick Spin DNA columns (Roche, Germany) were used to purify riboprobes.

Procedures were approved by the Animal Ethics Committee of Bilkent University. Animals were group housed in a 12-hour dark, 12-hour light cycle. Animals were anesthetized by intraperitoneal administration of Ketamine (40-50 mg/kg) and Xylazine (5 mg/kg). Then, mice were perfused with 4% paraformaldehyde by injecting into left ventricule. Brains of decapitated animals were removed and post-fixed in 4% paraformaldehyde for 1 hour at 4℃. The brains were then placed in sucrose solution at 4℃. Serial sagittal sections of P7 brain were cut at 20 microns with cryostat (Leica) and directly mounted onto HistoBond (+) slides. Sections were incubated overnight in hybridization buffer at 60℃. The hybridization buffer contains 50% formamide, 5xSSC, 5x Denhardt's reagent, 50 µg/ mL Heparin, 500 µg/ mL herring sperm DNA, and 250 µg/ mL yeast tRNA. Sections were then washed with 50% formamide and 2X SSC at 60℃ for 90 min. Then, anti-Dig Fab fragments conjugated to Alkaline Phosphatase and NBT/BCIP substrate mixture was used to detect the probes.

2.9 Functional prediction

2.9.1 Functional prediction of the variants

Disease causing probability of the variants were predicted by using SIFT[50] and PolyPhen[53] tools. Coordinates of the variants were converted into latest human genome assembly version (hg19) by using liftOver (`http://genome.ucsc.edu/cgi-bin/hgLiftOver`) to analyse in up-to-date versions of prediction software and datasets. Genomic Evolutionary Rate Profiling (GERP)[46] scores were obtained from UCSC Genome Browser allHg19RS_BW track. Domains of the protein product of candidate gene were predicted by using PFAM protein domain search module of CLCMain Workbench V5.0 (CLCBio, Inc.)

and ScanProsite[81]. Then, possible effects of the mutation was evaluated. Membrane spanning domains were predicted using TMpred software (`http://www.ch.embnet.org/software/TMPRED\_form.html`). Homologous proteins were identified and alignments were done by using CLCMain Workbench using appropriate modules.

2.9.2 Data mining from published expression datasets

Expression of the candidate gene in fetal brain tissues was evaluated using the published microarray data sets of E9.5, E11.5 and E13.5 mouse brain tissues (GSE8091)[82] from the GEO database (`http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi`). GeneSpring GX V11.1 software (Agilent Technologies, Inc.) was used to analyse data-sets. First, data were grouped into day groups. Standard quality control and filtering analysis were performed with appropriate modules of the software (`http://www.chem.agilent.com/cag/bsp/products/gsgx/manuals/GeneSpring-manual.pdf`). One-way ANOVA Test was used to filter differentially expressed genes within groups (Bonferroni corrected $p < 0.001$). Genes that correlated highly with the gene of interest (R=0.95-1.0) was identified by the "Find Similar Entity Lists" module. Then, functional annotation clustering of the subset of correlated genes was performed by DAVID tools [83].

A web-based data-mining engine NextBio[84] was used to further investigate the differential expression of the candidate gene in GEO datasets. Significant association was found in the analysis of GSE4175[85] dataset.

2.10 Chemicals, reagents and enzymes

2.10.1 Enzymes

All the restriction digestion enzymes and Thermus aquaticus polymerase enzymes were purchased from MBI Fermentas (NY, USA).

2.10.2 Solutions and buffers

1X TAE (Tris-acetic acid-EDTA):	40mM Tris-acetate, 2 nM EDTA, pH 8.0
Ethidium bromide:	10mg/ml in water (stock solution) 30 ng/ml (working solution)
Agarose gel loading buffer (6X):	15% ficoll 0.05% bromophenol 0.05% xylene cyanol
Acrylamide:bisacrylamide (30%):	29.5 gr acrylamide 0.44 gr bisacrylamide ddH2O to 100 ml
SSC (20X):	175.32 gr Sodium Chloride 88.23 gr Sodium Citrate $_{dd}H_20$ to 1 lt pH to 7.0

2.10.3 Chemicals and reagents

Reagent/Chemical	Company
Agarose	Basica LE, EU
Acetic acid	Sigma, MO, USA
Acrylamide	Sigma, MO, USA
Ammonium persulfate	Carlo Elba, Italy
Bisacrylamide	Sigma, MO, USA
Bromophenol blue	Sigma, MO, USA
dNTPs	Fermentas, NY, USA
EDTA	Fermentas, NY, USA
Ethanol	Merck, Germany
Ethidium bromide	Sigma, MO, USA
TEMED	Sigma, MO, USA
Tris-Base	Bio-Rad, CA, USA
Trizol reagent	Invitrogen, CA, USA
Formamide (Deionized)	Ambion, TX, USA
Denhardt's reagent	Invitrogen, CA, USA
Tween-20	Sigma, MO, USA
Heparin	Sigma, MO, USA
Herring sperm DNA	Invitrogen, CA, USA
Bakers yeast RNA	Sigma, MO, USA
Anti-Digoxigenin-AP, Fab fragments	Roche, Germany
NBT	Roche, Germany
BCIP	Roche, Germany
CHAPS	Sigma, MO, USA
NH4OAc	Ambion, TX, USA

2.11 Reference sequences used in this study

Accession codes for WDR81 orthologs are:

Ailuropoda melanoleuca:	XP_002918082
Callithrix jacchus:	XP_002747874
Danio rerio:	XP_001921778
Equus caballus:	XP_001502383
Gallus gallus:	XP_415806
Monodelphis domestica:	XP_001371487
Mus musculus:	NP_620400
Oryctolagus cuniculus:	XP_002718930
Pan troglodytes:	XP_523527
Pongo abelii:	XP_002826860
Rattus norvegicus:	NP_001127832
Sus scrofa:	XP_003131868
Taeniopygia guttata:	XP_002194363
Tetraodon nigroviridis:	CAG08933
Xenopus (Silurana) tropicalis:	XP_002937192

Chapter 3

Results

3.1 Clinical assessment of the affected family

The affected individuals originated from a consanguineous family in south-eastern Turkey. Six of the 19 children exhibited mental retardation, dysarthric speech, truncal ataxia. Five of the affected siblings were exhibited quadrupedal locomotion. One affected male was bipedal but exhibited similar neurological symptoms. Another affected sibling was deceased at age of 26. However, the phenotype of the deceased individual was not clear. Quadruped patients were well balanced and did not show any ataxic movements while walking on all fours (Figure 3.1). Their ulnar palms make contact with ground. In resting state, patients were able to stand upright but they quickly returned to the quadrupedal position for walking. Affected individuals have dysarthric speech with limited vocabulary. Mini Mental State Examination Test was used to asses their mental abilities and revealed severe mental retardation. They understood and respond to simple questions in their own language, but they did not show consciousness on time and place. Neurological examination of the index case revealed cerebellar ataxia, intentional tremor without pyramidal signs, bilateral dysmetria, dysdiadochokinesia and mildly decreased muscle tones. Brain imaging using cranial magnetic

Figure 3.1: Quadrupedal locomotion by a male patient.

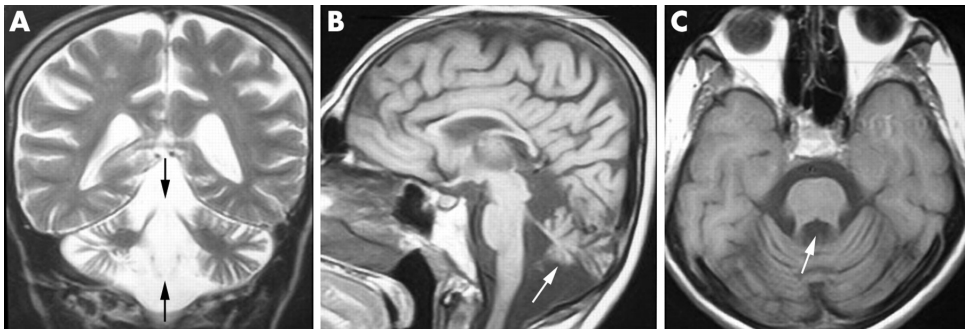


Figure 3.2: MRI of a male patient. Coronal (A), sagittal (B), and transverse (C) sections are shown. A and B arrows shows cerebellar hypoplasia and midline clefting. Note hypoplasia of the corpus callosum (B) and enlargment of the fourth ventricle (C). (From Turkmen et al. 2006 [10] by permission)

resonance revealed a number of anatomical abnormalities including, cerebellar hypoplasia, complete absence of inferior vermis, enlarged fourth nucleus, and mild hypoplasia of the corpus callosum (Figure 3.2). Obligate carrier father and an unaffected carrier brother had normal brain.[9, 10]

Further characterization of structural abnormalities of the brain of the affected individuals was carried out by magnetic resonance imaging and morphometric analyses. Volume in the cerebellum, corpus callosum, cortex and white matter was found to be reduced (Figure 3.3A). There were significant structural differences in motor areas precentral gyrus and BA6, and motor speech areas pars

opercularis and pars triangularis (Figure 3.3B). There was also moderate to high atropy in superior, middle and inferior cereballar peduncles revealed by diffusion tensor imaging (DTI).(Figure 3.4)

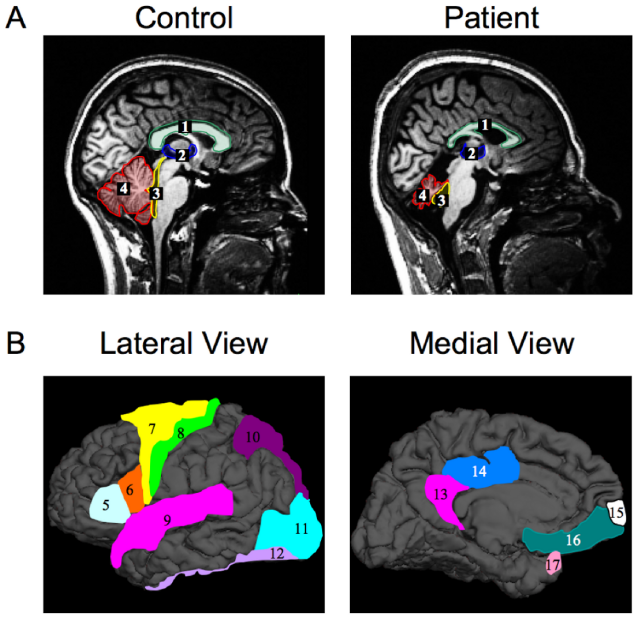


Figure 3.3: Morphological analysis of brain from affected and unaffected individuals. (A) Midsagittal MRI scans of a healthy control individual (left) and affected relative from Family B (right). The highlighted regions show areas where volumetric differences are readily visible: corpus callosum (1), third ventricle (2), fourth ventricle (3), and cerebellum (4). (B) Cortical regions with significant differences in morphometric parameters are displayed on a reference cortex, from lateral and medial view: BA45 (5), BA44 (6), BA6 (7), precentral (8), superior temporal (9), superior parietal (10), lateral occipital (11), fusiform (12), isthmus cingulated (13), posterior cingulated (14), frontal pole (15), medial orbitofrontal (16), and temporal pole (17).(From Gulsuner et. al. 2011 [11] by permission)

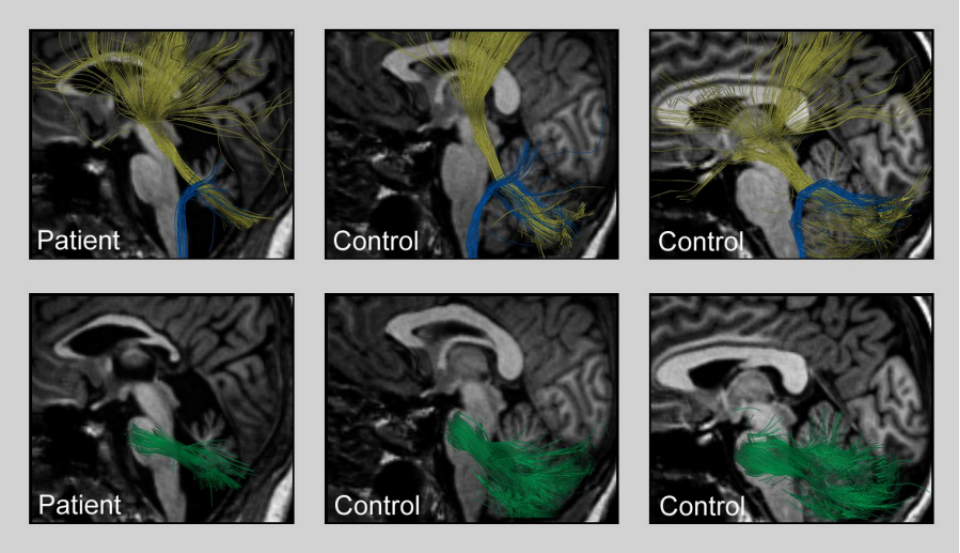


Figure 3.4: Diffusion tensor imaging (DTI) and fiber tractography. Fibers of the inferior middle and superior cerebellar peduncles of 1 female patient and 2 controls are shown. Peduncles of the patient are atrophic compared to healthy controls; this is most obvious in the middle cerebellar peduncle. (From Gulsuner et. al. 2011 [11] by permission).

3.2 Genetic mapping

Pedigree analysis of the family showed autosomal recessive inheritance of the disease. Previously, disease locus was mapped to a 7.1 Mb region on chromosomes 17p between markers D17S1866 and D17S960.[10] Linkage analysis with autosomal recessive model using Affymetrix 10K Xba GeneChip SNP data confirmed the previously reported locus on chromosome 17p (Figure 3.5 and Appendix B). Minimal critical region was further saturated and confirmed by genotyping with microsatellite markers and haplotype analysis (Figure 3.6).

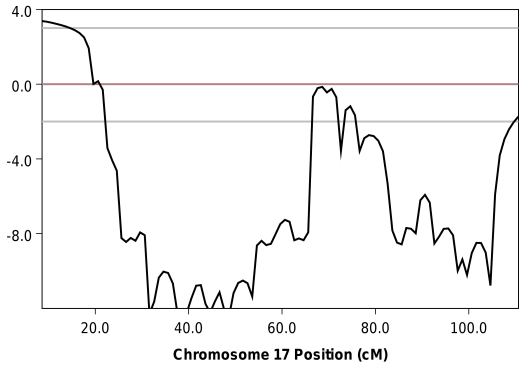


Figure 3.5: Parametric linkage analysis revealed a single locus on chromosome 17p. See appendix B for the remaining chromosomes.

However, genotype data of chromosomes 1, 2, 5 and 14 was not informative for linkage analysis and those chromosomes could not be excluded (Appendix B). In addition, homozygosity mapping analysis using 10K SNP data resulted with three homozygous loci on chromosomes 8, 14 and 17. The 10K SNP microarrays contain low number of SNPs (n=10,204) and do not represent all regions of human genome. In addition, genotype data of the family contains several missing genotypes (ie. 05-985: 22%, 05-986: 19%) and had relatively low confidence scores. Thus, analysis was repeated with high resolution Illumina 300 Duo v2 BeadChip (SNP count=318,328) using DNA of two affected individuals (05-984 and 05-987). All chromosomal regions except the minimal critical region on chromosome 17p were excluded by homozygosity mapping analysis using Illumina data. The minimal critical region was detected between 114,669-6,917,703 base

pairs (hg19) on chromosome 17 (Figure 3.7).

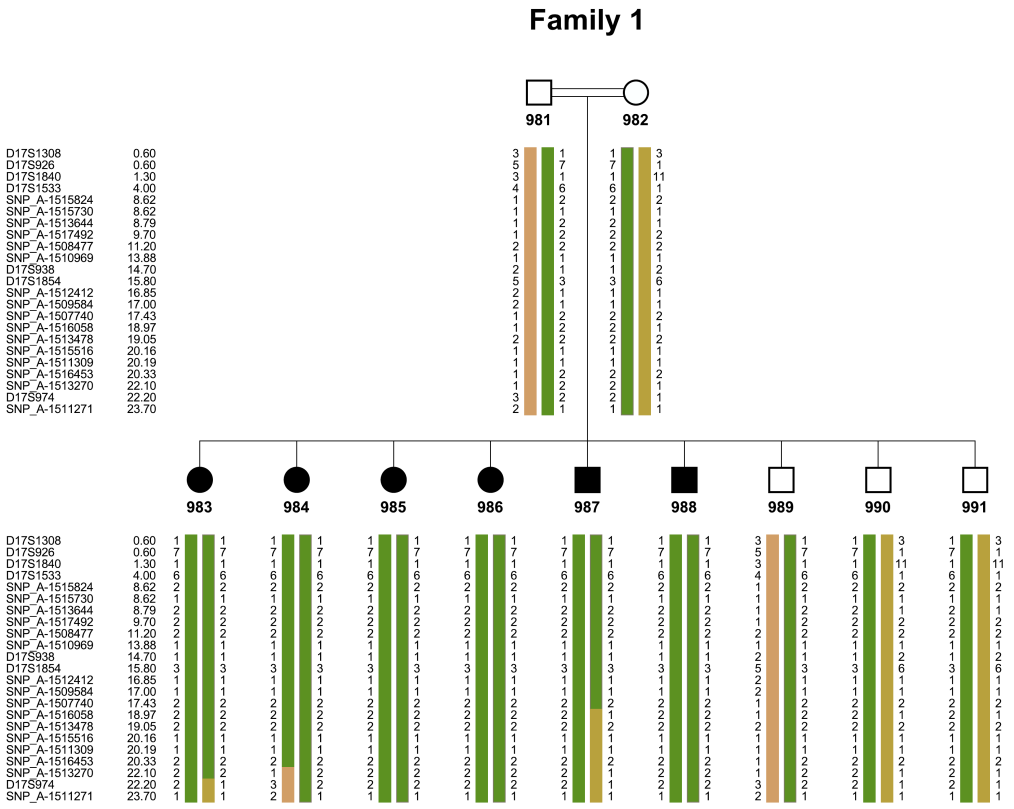


Figure 3.6: Haplotype analysis of minimal critical region on chromosome 17p. Haplotypes of the family confirmed the disease locus.

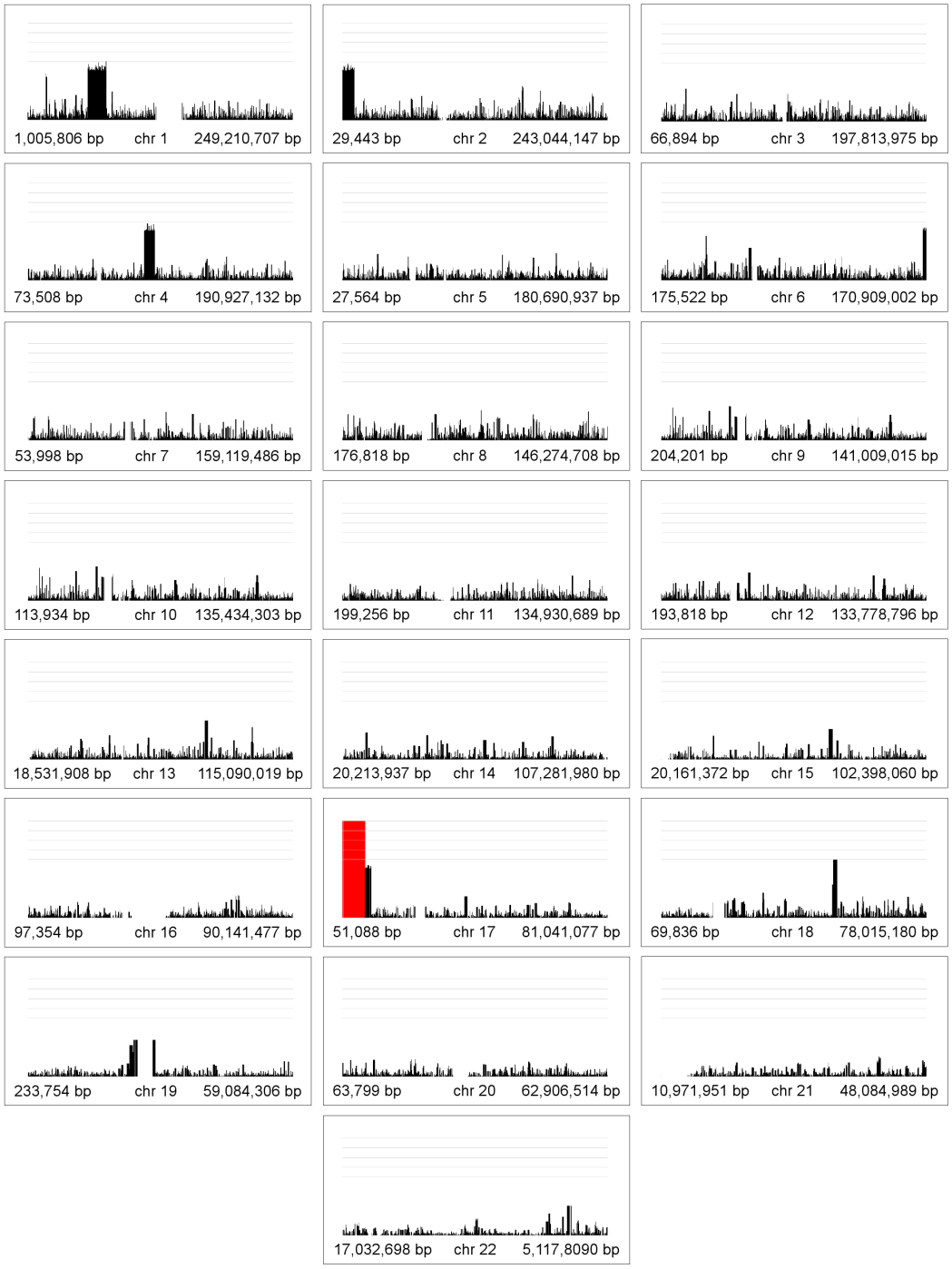


Figure 3.7: Homozygosity mapping analysis. Genotype data was obtained from 05-984 and 05-987 using Illumina 300 Duo v2 BeadChip microarrays. Only one single region was detected (chr17:114,669-6,917,703). Analysis was performed by using HomozygosityMapper software. Y-axis indicates genome-wide homozygosity scores (max=500). Red bars refers to the homozygous regions.

3.3 Candidate gene sequencing

The minimal critical region on chromosome 17p (D17S1866-D17S960) consisted of 214 genes (Appendix C.1) including 151 protein coding and 23 RNA genes (GRCh37.p5) with a total of 482 transcripts. Regarding to the huge number of genes which segregate with the disease in the family, a candidate gene sequencing approach was realized and three different screening strategies were employed.

CAMRQ is a genetically heterogeneous disease and the first gene implicated in CAMRQ is *VLDLR*.[12] It is suggested that, genetically heterogeneous diseases could be raised from disruptions of the same or related biological pathways.[86] Hence, first an interaction analysis of VLDLR protein was employed. A total of 2139 proteins were found to be interacted with VLDLR up to three connection levels by APID (Agile Protein Interaction DataAnalyzer). Probably 10 proteins directly interact with VLDLR which constitutes the first level. Approximately an additonal 81 proteins interact with those 10 at the first level, and finally 2139 proteins found to be interacting with the 81 genes of the second level. Next, the corresponding genes of those 2139 proteins were merged with those genes in the 17p13 interval. This analysis yielded 20 genes, of which 16 were known to have cerebellar expression. Two of them, *CRK* and *LIS1* (*PAFAH1B1*), encode for proteins that directly interact with DAB1, which is a key protein of Reelin pathway (Figure 3.8 and Table 3.1) Furthermore, we found an additional protein (KIF1C) in Reelin Pathway using GeneAssist^{TM} Pathway Atlas.

Next, DGP tool (Disease Gene Prediction)[37] revealed *CRK* as one of the most likely candidates (Table 3.2). The second tool, Prospectr (PRiOrization by Sequence & Phylogenetic Extent of CandidaTe Regions) identified *SCARF1* as the candidate with the highest weight score (Table 3.2).

In conclusion, these combined bioinformatics approaches resulted in the identification of *CRK*, *LIS1* (*PAFAH1B1*), *KIF1C*, and *SCARF1* as candidate genes.

Exons, exon/intron boundaries and untranslated regions of four candidate genes were sequenced using 77 primer pairs. 98.7% (76/77) of the sequencing

Table 3.1: Proteins interacting with VLDLR.

Ensembl Gene ID	Location	Gene symbol	UniProt/Swiss-Prot ID	C.Exp
ENSG00000108953	1,194,595	YWHAE	1433E_HUMAN	+
ENSG00000167193	1,272,226	CRK	CRK_HUMAN	+
ENSG00000174238	1,368,037	PITPNA	PIPNA_HUMAN	+
ENSG00000074660	1,483,903	SCARF1	SREC_HUMAN	+
ENSG00000174231	1,500,674	PRPF8	PRP8_HUMAN	+
ENSG00000167711	1,593,070	SERPINF2	A2AP_HUMAN	+
ENSG00000132383	1,680,095	RPA1	RFA1_HUMAN	+
ENSG00000007168	2,443,686	PAFAH1B1	LIS1_HUMAN	+
ENSG00000205813	3,128,014	OR3A2	OR3A2_HUMAN	–
ENSG00000083457	3,564,672	ITGAE	ITAE_HUMAN	–
ENSG00000074755	3,854,489	ZZEF1	ZZEF1_HUMAN	–
ENSG00000129219	4,657,392	PLD2	PLD2_HUMAN	+
ENSG00000141503	4,728,437	MINK1	MINK1_HUMAN	+
ENSG00000185245	4,776,680	GP1BA	GP1BA_HUMAN	+
ENSG00000108518	4,789,723	PFN1	PROF1_HUMAN	+
ENSG00000108561	5,276,676	C1QBP	C1QBP_HUMAN	+
ENSG00000091592	5,345,858	NLRP1	NALP1_HUMAN	–
ENSG00000108590	6,487,789	MED31	MED31_HUMAN	–
ENSG00000108839	6,840,128	ALOX12	LOX12_HUMAN	+
ENSG00000132535	7,033,933	DLG4	DLG4_HUMAN	+

Abbreviations used in this table: C.exp, cerebellar expression.

Table 3.2: Disease gene prediction. Scores calculated using DGP, Prospectr and Suspects.

Gene	VLDLR	Ataxia	Cerebellar	Sum	DGP	D.P.
SCARF1	63,04	12,86	12,71	88,61	0,34	No
PITPNA	50,3	12,68	17,26	80,24		
TRPV1	49,87	15	14,52	79,39	0,66	No
ASGR1	40,39	10,21	10,06	60,66		
RABEP1	39,52	12,08	13,4	65	0,62	No
CLEC10A	39,39	9,2	9,05	57,64		
TM4SF5	15,18	15,18	15,18	45,54	0,59	No
CHRNE	14,14	17,5	16,62	48,26	0,7	Yes
PRPF8	13,33	13,33	13,33	39,99	0,75	Yes
KIF1C	13,3	11,5	10,2	35	0,5	No
ACADVL	13,24	13,52	13,24	40		
C1QBP	13,17	13,17	13,17	39,51	0,6	No
SLC25A11	13,08	16,05	17,85	46,98	0,53	No
CTNS	13,01	13,01	13,01	39,03	0,73	Yes
P2RX5	12,26	11,81	11,84	35,91	0,53	No
RNF167	12,17	12,17	12,17	36,51		
PSMB6	12,03	12,68	12,03	36,74	0,54	No
MYO1C	11,88	14,81	11,88	38,57	0,52	No
RPA1	11,87	16,03	14,58	42,48	0,56	No
DVL2	11,81	11,81	11,81	35,43		
SKIP_HUMAN	11,48	12,6	12,79	36,87		
SLC13A5	11,48	12,44	12,28	36,2	0,6	No
DLG4	11,09	12,03	10,73	33,85		
SERPINF1	10,72	10,27	10,3	31,29	0,34	No
SPAG7	10,7	10,7	10,7	32,1	0,41	No
TAX1BP3	10,7	10,7	10,7	32,1		
UBE2G1	10,65	10,65	10,65	31,95	0,31	No
PLD2	10,56	15,42	20,73	46,71	0,47	No
ALOX15	10,38	10,66	10,38	31,42	0,54	No
ASPA	10,36	10,36	10,36	31,08	0,67	Yes
ARRB2	10,33	10,33	10,33	30,99	0,5	No
DULLARD	10,33	10,33	10,33	30,99		
CAMTA2	10,26	10,26	10,26	30,78	0,52	No
ATP2A3	10,2	14,38	19,99	44,57	0,47	No
MINK1	9,4	14,02	16,61	40,03		
CRK	9,38	10,19	10,47	30,04	0,75	No
GABARAP	9,32	12,07	11,43	32,82		
ENO3	9,2	10,98	9,2	29,38	0,74	No
PFN1	9,2	9,23	9,2	27,63	0,34	No
NUP88	9,04	9,04	9,04	27,12	0,38	No
SRR	8,94	9,84	8,94	27,72	0,53	No
CAMKK1	8,94	13,1	12,27	34,31	0,57	No
ALOX12	8,86	9,79	8,86	27,51	0,6	No
GP1BA	8,71	10,54	8,71	27,96	0,7	Yes
GPS2	8,29	8,72	9,58	26,59		
ABR	7,71	10,91	18,25	36,87	0,55	No
GEMIN4	7,38	7,38	7,38	22,14	0,2	No
PAFAH1B1	7,24	11	10,15	28,39	0,74	Yes
YWHAE	6,65	6,65	6,65	19,95	0,7	Yes
MYBBP1A	6,55	6,55	6,55	19,65	0,4	No
USP6	4,96	4,96	4,96	14,88	0,39	No
ITGAE					0.69	No

Abbreviations used in this table: D.P., disease phenotype.

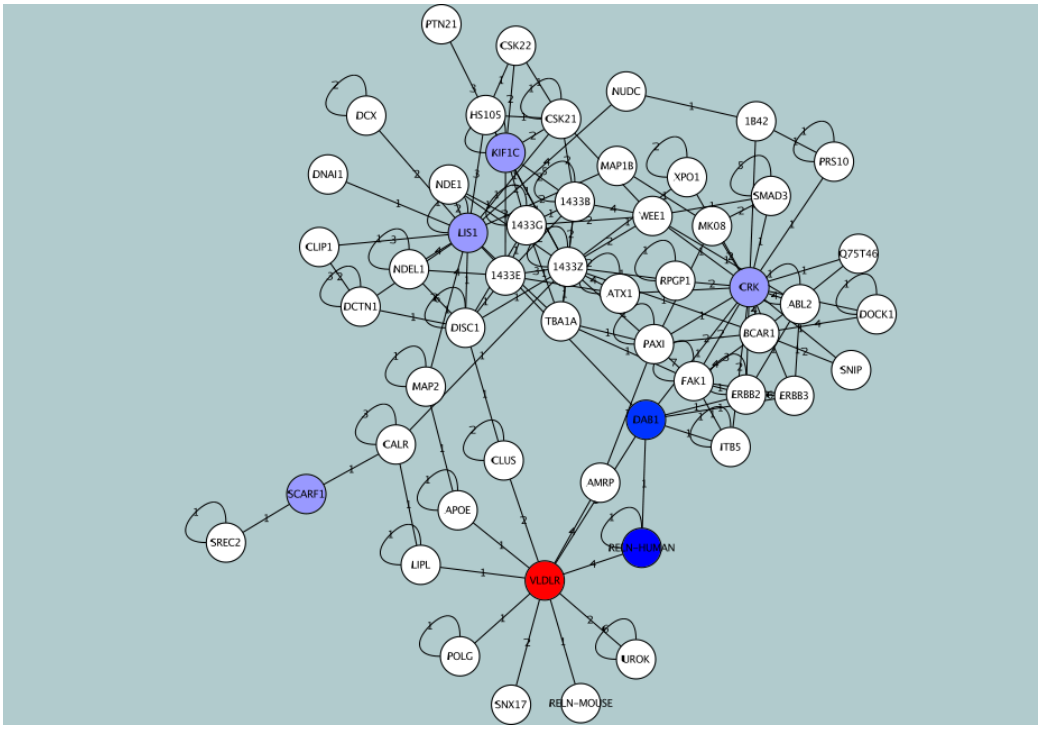


Figure 3.8: APID interaction analysis. An example output of interaction analysis is shown. *SCARF1*, *CRK*, *LIS1* (*PAFAH1B1*), and *KIF1C* are located in the minimal critical region. *KIF1C* is obtained from GeneAssist^{TM} Pathway Atlas manually added to the interactome figure.

reactions were successfully realized. Prior to sequencing all PCR products were analysed using PAGE electrophoresis to identify any possible small insertions and deletions. PCR amplicons were then sequenced using forward and reverse primers. However, no mutation that segregated with the disease was detected.

Expansion of several nucleotide repeats are commonly observed in neurological phenotypes. To rule out this mechanism as the cause of disease in family B, a nucleotide motif search in minimal critical region was realized. A total of eight nucleotide repeats were detected in the intragenic regions (Table 3.4). However, gel electrophoresis following the amplification of the regions did not reveal any change in the size of the repeats.

Table 3.3: Candidate gene sequence analysis.

Gene	Exon #	Total reaction	Sanger
KIF1C1	23	26	100%
CRK	3	12	91.6%
SCARF1	11	15	100%
PAFAH1B1	11	24	100%

Table 3.4: Repeat motifs in the minimal critical region.

Gene	Repeat type	Repeat location
WDR81	GAA * 5	Intron 2-3
HIC1	CGG * 7	Exon 2
RUTBC1	CGG * 7	Exon 1
MNT	CAG * 4	Exon 6
PAFAH1B1	CGG * 8	- 49 bp
GARNL4	CGG * 4	- 19 Mb
GARNL4	AGG * 11	- 19 Mb
TRPV1	CTG * 8	Intron 1-2

3.4 Targeted next generation sequencing of the critical region

The rate limiting step of the disease gene identification is between the identification of disease locus and finding the disease causing mutation. In most cases, disease locus contains hundreds of genes and candidate gene approaches fail to identify the culprit gene. A brute-force approach to sequence all genes in the minimal critical region with conventional approaches is very expensive and requires both extensive human labour and time. The minimal critical region in family B contains 482 transcripts with a total length of 515,569 bp (GRCh37.p5) which can be sequenced with at least 1000 PCR primer pairs and 4000 sequencing reactions to completely rule out all coding regions. Therefore it was decided that, using next generation sequencing approaches to sequence the 7.1 Mb region was the most cost-efficient strategy to find the culprit gene.

3.4.1 Capture and sequencing of the locus

Genomic DNA was obtained from blood samples of individuals by phenol-chloroform and ethanol precipitation method. Quality and quantity of the DNA samples were measured by spectrophotometry, PicoGreen method and analytical gel electrophoresis (Figure 3.9, Table 3.5). Then, the minimal critical region at chr17 (82,514-7,257,922 bp according to hg19) was captured using the DNA of two affected (05-985, 05-987) siblings and two obligate carrier parents (05-981 father, 05-982 mother) by Nimblegen 385K microarrays. 362 to 403 million base pairs (46.4X mean haploid coverage) were sequenced from captured DNA of each individual with 454 GS FLX using Titanium series reagents. Approximately 78.7% of the reads from each sample fell within the targeted region, which represents 1,275 to 2,247 fold enrichment. On average, 99.4% of all targeted bases were covered by at least 4 reads, which is used as the threshold for the detection of homozygous variants (Table 3.6). The statistics of the 454 sequencing is summarized in Table 3.7. An extended region (chr17:0-9,059,279) was captured by Nimblegen HD2 2.1M sequence capture microarrays using the DNA from another affected sibling (05-984). Captured DNA was then, sequenced on an Illumina Genome Analyzer IIx. 2.98 billion bases and 40.3 million reads were obtained. 27.88% of the reads were mapped to the targeted region. The calculated enrichment was 123 folds. 99.6% of the bases were covered by at least 4 times (Table 3.6).

Table 3.5: DNA concentrations for targeted capture.

Sample	Conc. (ng/µL)	Total DNA (µg)	260/280	260/230
05-981	310	34.1	1.90	2.22
05-982	277.4	47.16	1.90	2.17
05-985	349.9	52.49	1.86	2.14
05-987	338.3	40.6	1.90	2.17

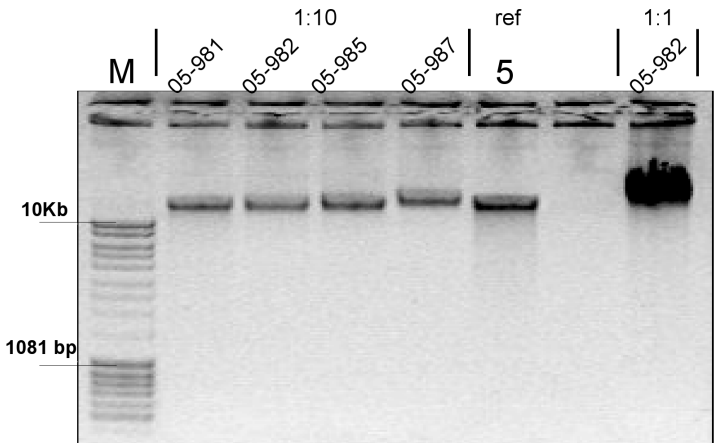


Figure 3.9: Analytic agarose gel electrophoresis. 1% Agarose Gel Electrophoresis 90V, 40 min. M, MassRuler DNA Ladder, Mix as marker; 2-5, DNA samples diluted 1:10 in TE; 6: 50 ng/ul reference DNA; 7, 05-982 undiluted. 1 µL of DNA and 1 µL Loading Buffer was loaded into each lane. Gel image was captured with BioRad Gel Doc 2000 system, DNA quantitation was performed using BioRad Multi Analyst 1.1 software.

Table 3.6: Coverage of the target region.

Depth	05-981	05-982	05-985	05-987	05-984
0	0.16	0.17	0.17	0.22	0.19
≥ 2	99.73	99.67	99.72	99.64	99.74
≥ 4	99.47	99.35	99.48	99.39	99.59
≥ 10	98.31	97.65	98.28	98.02	99.07
≥ 15	96.72	95.14	96.58	96.26	98.44
≥ 20	94.29	91.23	94.05	93.79	97.67
≥ 30	86.81	78.36	86.07	85.71	95.51

Table 3.7: Next generation sequencing statistics.

	05-981	**05-982**	**05-985**	**05-987**
Mapped bases (Mb)	403.5	362.7	410.2	399.2
% of total bases	99.4	99.1	99.3	99.4
Mapped reads (x106)	1.20	1.78	1.21	1.13
% of total reads	98.3	97.6	98.3	98.3
Target base coverage (%)	96.9	96.6	96.7	96.7
Fold enrichment	1963	1275	1622	2247
Mean coverage	48.4X	40.5X	47.4X	48.9X
All variants	26,751	27,471	24,115	24,690
dbSNP + CEU[a]	11,674	11,703	8991	9106
Novel variants	15,077	15,768	15,124	15,584
High confidence variants	11,942	12,801	9805	9592
dbSNP + CEU[a]	10,443	10,426	8150	8235
Novel heterozygous	678	1020	421	332
Novel homozygous	821	1355	1234	1025
Novel shared variants		1119		
Coding regions (G1)		20		
5' and 3' UTR (G2)		15		
Intronic (G3)		689		
Intergenic (G4)		395		
Mendelian compatibility (G1)		3		
Population screening		1		

[a]: CEU population of 1000Genomes variants. For a list of missense variants see Appendix D.

3.4.2 Variant calling and error rates

For 454 reads, variants were detected using Roche GSMapper software (Roche Inc), which uses two different approaches.[72] A conservative method was used to detect high confidence variants (HCDiff) and a less stringent method (AllDiff) was used to detect all possible sequence variants. Reads mapped to the outside of the target region were discarded from analysis. Across all samples 7287 shared and 18,410 total high confidence variants were detected (Table 3.7). As a first level genotype designation, we used a rigid heterozygosity threshold (30-70%) to maximize sensitivity for detection of homozygous variants. According to this threshold variants detected more than 70% of the reads designated as homozygous. When compared to the Illumine SNP genotyping data, the detection of alleles was achieved for 99.5% for sample 05-987. Across 1004 SNPs in the targeted region, five SNPs (0.49%) were found to be inconsistent with Illumina

SNP genotyping data (Table 3.8). Analysis of alignments revealed that, two of the five SNPs (rs719913, rs231674; 0.2%) were false positives and were found to be in a homopolymeric region. One SNP (rs8067500, 0.1%) was a false negative variant and one SNP (rs8068519, 0.1%) was determined as read error with a quality score of zero. Remaining one SNP (rs314239) was found to be a genotyping error of Illumina SNP genotyping platform. Alignments of the erroneous SNPs were given at Appendix E. As expected heterozygous SNPs were negligible in the affected DNA. 23/25 SNPs were called as heterozygous by both genotyping and sequencing platforms (Table 3.9). Those SNPs were located in the borders of homozygous block thus narrowed the homozygous block into 6.74 Mb. Alignment analysis of remaining 2 SNPs (rs719913, rs231674) which were detected only with sequencing platform, were identified as variant detection errors (Appendix E).

Table 3.8: Insconsistent SNPs between 454 and Illumina 300Duo V2 genotyping platforms[a]

dbSNP id	Position	Illumina 300 Duo	GS FLX	Read	Percent
rs8068519	1,020,053	TT	GG	53	100
rs719913	1,674,048	GG	AG	35	69
rs231674	3,137,283	CC	TC	27	67
rs314239	6,954,089	GA	GG	56	100
rs8067500	7,113,355	CT	TT	34	100

[a]: 5/1004 SNPs were found to be inconsistent with Illumina SNP genotyping data. Alignment analysis reveal that, 3/5 of the inconsistencies were resulted from variant detection errors (rs719913, rs231674, rs8068519), 1/5 was resulted from sequencing error (rs8068519) of sequencing platform. Remaining one SNP was found to be an error of genotyping platform (rs314239). See Appendix E for alignments.

Table 3.9: Heterozygous SNPs in the critical region[a].

		Illumina 300Duo V2		454 FLX NGS	
dbSNP id	**Position (hg19)**	**Genotype**	**Score**	**Genotype**	**Read / %**
rs7217872	chr17: 88,988	CT	0.63	CT	25 / 68
rs4617924	chr17: 114,669	CA	0.86	CA	84 / 50
rs719913	chr17: 1,727,298	GG	0.88	AG	35 / 69
rs231674	chr17: 3,190,533	CC	0.86	TC	27 / 67
rs7338	chr17: 6,917,703	CT	0.85	CT	27 / 48
rs10521149	chr17: 6,932,140	AC	0.8	AC	46 / 46
rs312456	chr17: 6,938,335	AG	0.81	AG	27 / 63
rs13342692	chr17: 6,946,287	TC	0.82	TC	32 / 34
rs364569	chr17: 6,979,179	AG	0.81	AG	66 / 64
rs444207	chr17: 6,982,444	AG	0.86	AG	47 / 40
rs8078782	chr17: 6,996,653	CA	0.91	CA	26 / 54
rs3760352	chr17: 7,019,878	AG	0.87	AG	52 / 46
rs2062737	chr17: 7,031,638	AG	0.45	AG	46 / 46
rs12150124	chr17: 7,036,251	TC	0.93	TC	33 / 42
rs314245	chr17: 7,040,878	CT	0.92	CT	51 / 47
rs1237044	chr17: 7,042,246	TC	0.86	TC	23 / 57
rs3826408	chr17: 7,101,292	CT	0.9	CT	60 / 52
rs390200	chr17: 7,109,995	AG	0.71	AG	29 / 38
rs2074222	chr17: 7,129,974	AG	0.86	AG	34 / 38
rs222851	chr17: 7,139,238	GA	0.83	GA	17 / 59
rs222852	chr17: 7,140,606	AG	0.82	AG	61 / 43
rs1215	chr17: 7,163,350	AG	0.83	AG	25 / 64
rs4796407	chr17: 7,245,371	AG	0.91	AG	55 / 51
rs11657406	chr17: 7,249,870	GT	0.75	GT	41 / 37
rs2292068	chr17: 7,253,659	TC	0.89	TC	14 / 57

[a]: One affected individuals (05-987) DNA was genotyped by Illumina 300Duo V2. 23/25 SNPs were called as heterozygote by both genotyping and sequencing platforms. 2/25 SNPs were determined as variant detection error (rs719913, rs231674). See Appendix E for alignments of rs719913 and rs231674.

Next, Mendelian inconsistencies were calculated in the HCDiff variants. 8,308 variants were included into analysis, which were detected in both parent and at least one offspring. The Mendelian error rate (ie. ABxBB=AA) was 0.39% (32/8118) at >5X and 0.26% (18/6938) at >15X with 30-70% heterozygosity threshold. Applying a coverage-dependent genotype calling approach [72] did not change the overall rate of Mendelian errors. Using different cut-off values between 20-30% as lower threshold and 65-80% as upper threshold, 64 erroneous variants were observed (Appendix F). It is suggested that, Mendelian errors could be used as a determinant of genotyping errors.[87] Using both genotyping data and Mendelian errors, 0.3% error rate was determined on overall sequence data.

3.4.3 Analysis of sequence gaps

To reveal non-covered coding regions and possible deletions, alignment data of three affected individuals were merged to obtain a mean value of read depth for each nucleotide position. Then, using the custom annotation pipeline coverage statistics were calculated for each of the functional genomic element. Combined sequence data of the 3 affected siblings yielded at least 4-fold coverage of 99.78% of all coding base pairs, 95.32% of intronic and UTR base pairs, and 91.36% of intergenic base pairs. The remaining 0.22% of coding regions with less than four-fold coverage was supposed to be candidate regions for disease causing deletions (Table 3.10). Those regions were compared to that of heterozygous parents to detect Mendelian inheritance consistency (Figure 3.10). None of the parents was detected as a heterozygous carrier for a possible deletion. The regions were detected as under low coverage in all of the sequenced individuals regardless of their affection status. Analysis of the regions by Sanger sequencing to detect any overlooked variants resulted with no mutation.

3.4.4 Variant annotation and filtering

To minimize the risk of erroneously overlooking significant transcript variants a custom annotation pipeline was developed using ENSEMBL data sets. A custom

Table 3.10: Coding regions with limited coverage. Genomic coordinates, gap sizes, GC percentage, accession code, gene symbol and coverage status of the coding regions with limited coverage.

Position (hg19)	**Size**	**GC %**	**Acc. code**	**Gene**	**Status**[a]
chr17:263352-263385	34	73.5%	NM_001013672	*C17orf97*	L.C.
chr17:882713-882715	3	100%	NM_022463	*NXN*	L.C.
chr17:1174341-1174383	43	86%	NM_001164405	*BHLHA9*	L.C.
chr17:2966802-2966821	20	45%	NM_014566	*OR1D5*	L.C.
chr17:2966822-2966893	72	52.8%	NM_014566	*OR1D5*	N.C.
chr17:2966894-2966901	8	50%	NM_014566	*OR1D5*	L.C.
chr17:4402342-4402343	2	50%	NM_001124758	*SPNS2*	N.C.
chr17:4402344-4402351	8	50%	NM_001124758	*SPNS2*	L.C.
chr17:6459704-6459726	23	73.9%	NM_001165966	*PITPNM3*	L.C.
chr17:6459710-6459724	15	86.7%	NM_001165966	*PITPNM3*	N.C.

[a]: Non-covered: <2X sequence depth. Low coverage: 2-3X sequence depth. Abbreviations used in this table: L.C., Low coverage; N.C., Non-covered.

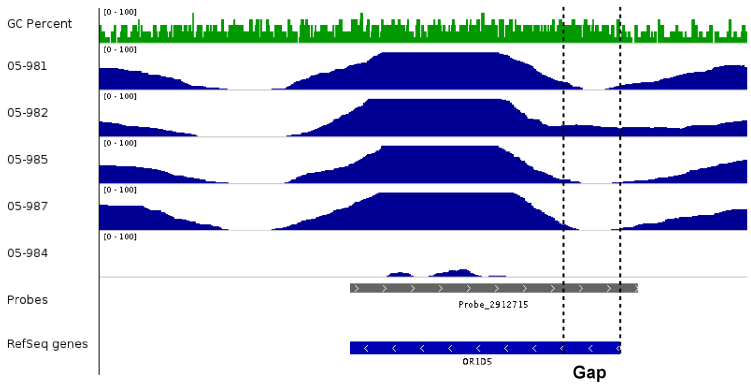


Figure 3.10: Segregation of a candidate deletion. Alignment of the depth graphs of all sequenced individuals for the gap in *OR1D5* pseudogene did not reveal any segregated deletion.

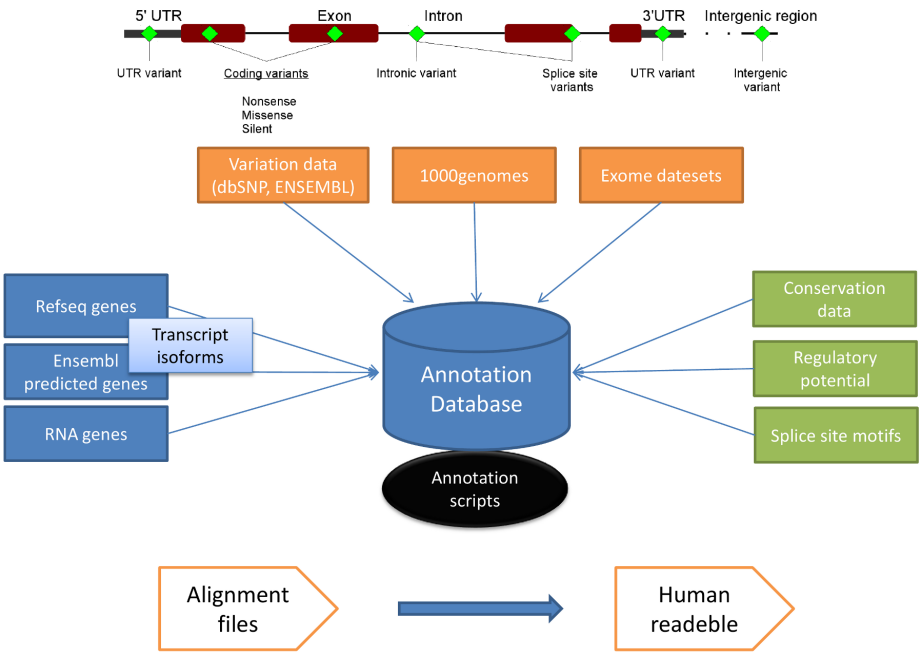


Figure 3.11: Custom annotation pipeline. Several data sources used to create a custom annotation database. Variants were identified and annotated by using custom scripts.

database was created using the genomic coordinates of protein coding genes, hypothetical genes, processed transcripts, miRNAs and other RNA genes and pseudogenes. The database also contains the coordinate information of all exons of all alternative transcripts variants. To annotate previously reported SNPs an additional database was created using the ENSEMBL Variation table. In addition variants were compared to 1000genomes variants [88] which were not added to dbSNP database. All variants were re-analysed using the updated versions of datasets upon their availability. Pipeline is summarized in Figure 3.11.

3.5 Identification of the disease causing variant

The key challenge of using the next generation sequencing technologies in the identification of novel disease genes is how to identify the disease causing mutation

among the background of neutral polymorphisms and sequencing errors.[44] With the use of high throughput sequencing, the challenge shift from the prioritization of the candidate genes to whistling down the list of candidate variants. Several steps of filtering and stratification approaches were carried out to narrow down the list of variants into single disease causing mutation. These approaches can be summarized as:

- Exclusion of the variants which were not compatible with the recessive inheritance model.
- Screening the novel variants to distinguish the rare polymorphisms from disease causing mutation.
- Stratification of the novel variants:

  Predicted roles of the genes in biological pathways or their interactions with genes that are known to cause similar phenotypes.

  Stratification of the candidate variants on the basis of their impact on the protein product and their deleteriousness.

  Conservation analysis based on the observation that mutations in the highly conserved regions are likely to be damaging.

3.5.1 Analysis and filtering SNPs

As the first step of discrete filtering, previously reported SNPs were excluded as neutral polymorphisms. This method assumes that the set of SNPs in db-SNP database does not contain any pathogenic alleles. However there are two possibilities that; i. dbSNP database can contain a small number of pathogenic allele contaminants, ii. For recessive diseases, carrier status will not result in a phenotype, thus the disease alleles can be observed in control group and reported as polymorphisms. To exclude all SNPs as disease causing mutation, segregation of nonsynonymous SNPs were analysed in the sequence data. Seventeen SNPs were compatible with the Mendelian transmission of the disease allele -detected

Table 3.11: Functional SNPs compatible with the Mendelian transmission of the disease allele.

SNP	Gene	Pos (hg19)	R/V	Annot	(AA/AB/BB)
rs9905106	*MYO1C*	chr17:1,373,518	T/C	Q826R	14.7/38.7/46.6
rs2070862	*SERPINF2*	chr17:1,648,294	C/T	A2V	61.2/34.2/4.7
rs216195	*SMG6*	chr17:2,203,167	T/G	K294Q	42.4/39.6/18.0
rs41333251	*OR1E2*	chr17:3,337,057	A/G	C27R	8.2/33.4/58.4
rs55916885	*TRPV1*	chr17:3,495,391	T/C	Q85R	96.1/3.8/0.1
rs150857	*SHPK*	chr17:3,526,637	C/T	E215K	61.2/30.4/8.4
rs2976230	*ITGAE*	chr17:3,631,241	A/G	V1019A	57.8/29.3/12.9
rs7214723	*CAMKK1*	chr17:3,775,848	T/C	E375G	44.8/39.2/16.0
rs34457931	*SPNS3*	chr17:4,352,636	G/A	G293R	83.2/15.9/0.8
rs3809849	*MYBBP1A*	chr17:4,458,598	G/C	Q8E	61.3/34.6/4.1
rs7216284	*GGT6*	chr17:4,463,699	G/A	R40W	68.2/27.6/4.2
rs1052748	*PLD2*	chr17:4,720,469	C/T	T577I	52.9/35.8/11.3
rs12761	*RPAIN*	chr17:5,326,145	C/G	N103K	36.6/40.3/23.1
rs2304977	*KIAA0753*	chr17:6,513,329	G/A	P566L	54.0/37.1/8.9
rs61146770	*KIAA0753*	chr17:6,524,298	T/A	E375D	48.2/42.4/9.4
rs16956264	*FBXO39*	chr17:6,683,684	A/G	Y166C	75.0/22.8/2.2
rs7213731	*FBXO39*	chr17:6,690,164	A/G	I363M	71.9/25.4/2.7

Abbreviations used in this table: R/V, Reference/Variant allele; Annot, Annotation at amino acid level. Frequencies of three different genotypes were given in the last column.

as heterozygous in parents and homozygous in affected siblings-. Frequencies of each allele were obtained from 1000genomes data [88] (October 2011 Intergrated Variant Set release #ICHG2011). Analysis of 1091 genomes revealed that all 17 SNPs were detected as homozygous state in at least one healthy individual (Table 3.11).

3.5.2 Functional classification of the variants

After the exclusion of previously reported dbSNP variants, 1119 shared novel variants were selected and binned into 4 groups (G) using annotation data. G-1 includes all variants in the amino acid coding regions of exons, consensus splices-site regions, and RNA genes (n=20). G-2 contains changes in 5'UTRs and 3'UTRs (15). G-3 contains the introns (689). All remaining variants were classified as G-4 (395). Next, variants which were compatible with the autosomal recessive inheritance model were selected. In this model variants detected homozygous in obligate carrier parents were filtered out. All functional and Mendelian compatible variants were confirmed by using Sanger sequencing (Figure 3.13). Please see Figure 3.12 for the summary of variant classification.

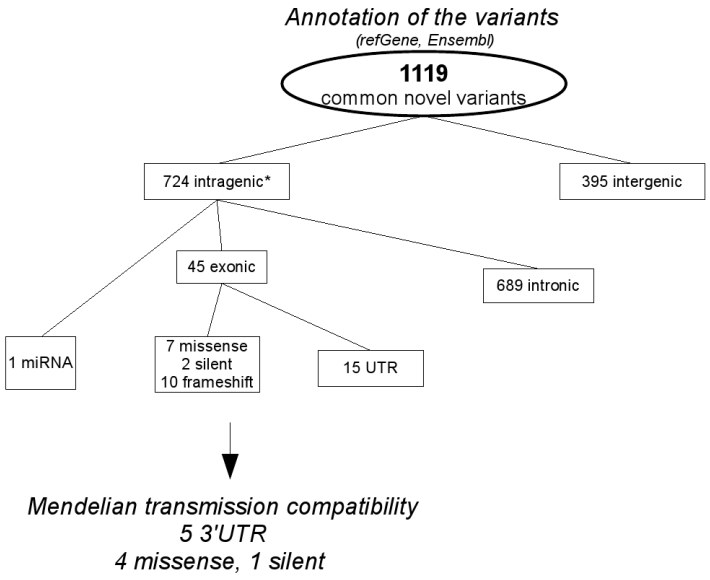


Figure 3.12: Flow chart of variant classification.

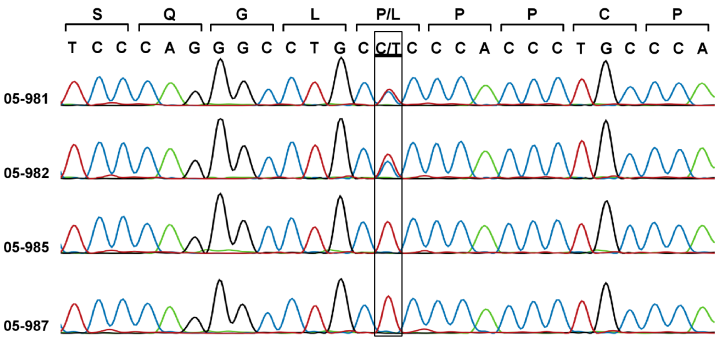


Figure 3.13: Segregation and Sanger confirmation of *WDR81* p.P856L variant.

Table 3.12: Novel UTR variants co-inherited with the disease.

Gene	Pos (hg19)	Type	Ref	Var	05-981	05-982	05-985	05-987
ELP2P	chr17:656,577	3'UTR	C	G	11 (0.36)	14 (0.29)	13 (0.62)	9 (1.00)
HIC1	chr17:1,962,684	3'UTR	A	-	17 (0.53)	9 (0.67)	16 (1.00)	23 (0.96)
USP6	chr17:5,076,345	3'UTR	G	C	23 (0.35)	21 (0.52)	26 (1.00)	22 (1.00)
ZNF594	chr17:5,083,351	3'UTR	A	T	18 (0.61)	12 (0.33)	12 (1.00)	12 (1.00)
ZNF594	chr17:5,083,804	3'UTR	G	C	24 (0.25)	14 (0.57)	19 (1.00)	21 (1.00)

3.5.3 Exclusion of the variants

Five of the 15 UTR variants were compatible with the recessive inheritance in the family (Table 3.12). They were analysed in detail including evaluation of protein interactions and regulatory motif searches. None of them was found to interact with previously identified genes with cerebellar phenotypes including CAMRQ associated VLDLR and CA8. According to regulatory region analysis, nucleotide substitutions did not have any significant effect on the regulatory motifs.

Genotypes of four missense variants (WDR81 p.P856L, MYBBP1A p.R671W, ZNF594 p.L639F and PELP1 p.T402I) and one silent variant (MYO1C p.E932E) were co-segregated with the disease in the family. Five variants in coding regions were detected homozygous in parents, and 10 variants were detected as false positive variants (Table 3.13). Thus, these 15 variants were excluded from analysis.

Alignment analysis of the silent variant in *MYO1C* (p.E951) revealed that, nucleotide change occurs in several species (Opossum, X. tropicalis, and chicken) without affecting the splicing pattern of the gene (Figure 3.14). PELP p.I401 variant is located in intron 10-11 of *PELP1* gene. This intron was retained in the alternative isoform (ENST00000301396, Q8IZL8-2) of the gene and annotated as an exon in the ENSEMBL database (Figure 3.15). These data strongly suggested that these two variants were neutral polymorphisms.

MYBBP1A p.R671W was excluded as the disease causing mutation based on genotypes of controls (Table 3.14). In 214 unrelated healthy controls, 13 individuals were heterozygous for the variant with a calculated allele frequency of 0.016 and homozygous frequency of approximately 1 in 4000, which was far

Table 3.13: List of novel variants in coding regions.

Gene	Pos (hg19)	Type	Ref	Var	Annot	05-981	05-982	05-985	05-987
Candidate variants									
MYO1C	chr17:1,371,325	Silent	C	T	p.E932E	42 (0.55)	24 (0.54)	26 (1.00)	39 (1.00)
WDR81	chr17:1,630,820	Missense	C	T	p.P856L	41 (0.51)	33 (0.52)	40 (0.97)	53 (1.00)
MYBBP1A	chr17:4,448,967	Missense	G	A	p.R671W	29 (0.52)	21 (0.48)	32 (0.97)	29 (1.00)
ZNF594	chr17:5,085,637	Missense	G	A	p.L639F	39 (0.54)	50 (0.56)	38 (0.97)	37 (1.00)
Excluded by homozygous parents									
C17orf97	chr17: 263,287	Missense	A	G	p.D218G	15 (1.00)	13 (1.00)	9 (1.00)	14 (1.00)
C17orf97	chr17: 263,316	Silent	T	C	p.D228H	13 (0.77)	6 (1.00)	4 (0.75)	8 (0.87)
AC090282.6	chr17:3,132,364	miRNA	-	C	-	39 (0.95)	37 (1.00)	34 (0.97)	39 (0.97)
TRPV1	chr17:3,477,172	Frameshift	-	G	p.A620fs	28 (1.00)	17 (1.00)	21 (1.00)	39 (1.00)
ASGR2	chr17:7,012,080	Frameshift	-	T	p.H65fs	29 (1.00)	35 (0.54)	36 (1.00)	41 (0.54)
Retained intron									
PELP1	chr17:4,578,955	Missense	G	A	T402I	46 (0.61)	37 (0.51)	45 (1.00)	60 (1.00)
False positive variants									
RPH3AL	chr17:169,221	Frameshift	T	-	p.D114fs	46 (0.17)	29 (0.28)	31 (0.23)	51 (0.18)
SLC43A2	chr17:1,479,053	Frameshift	C	-	p.V519fs	41 (0.15)	33 (0.21)	35 (0.23)	38 (0.16)
PRPF8	chr17:1,563,808	Frameshift	G	-	p.T1568fs	83 (0.13)	65 (0.20)	66 (0.17)	85 (0.07)
OR1E2	chr17:3,336,820	Frameshift	A	-	p.Y106fs	75 (0.09)	57 (0.19)	52 (0.12)	74 (0.16)
ITGAE	chr17:3,664,330	Frameshift	-	C	p.E192fs	40 (0.15)	29 (0.10)	35 (0.11)	43 (0.07)
USP6	chr17:5,036,205	Missense	A	C	p.K66Q	50 (0.20)	42 (0.17)	42 (0.19)	67 (0.31)
USP6	chr17:5,036,210	Missense	T	G	I68M	50 (0.20)	42 (0.17)	43 (0.23)	68 (0.35)
DHX33	chr17:5,353,585	Frameshift	C	-	p.E383fs	52 (0.15)	39 (0.18)	58 (0.12)	46 (0.09)
NLRP1	chr17:5,418,095	Frameshift	T	-	p.G1467fs	75 (0.19)	54 (0.22)	50 (0.20)	77 (0.16)
KIAA0753	chr17:6,513,435	Frameshift	T	-	p.K531fs	47 (0.17)	32 (0.22)	37 (0.16)	41 (0.15)

Abbreviations used in this table: Annot, protein level annotation.

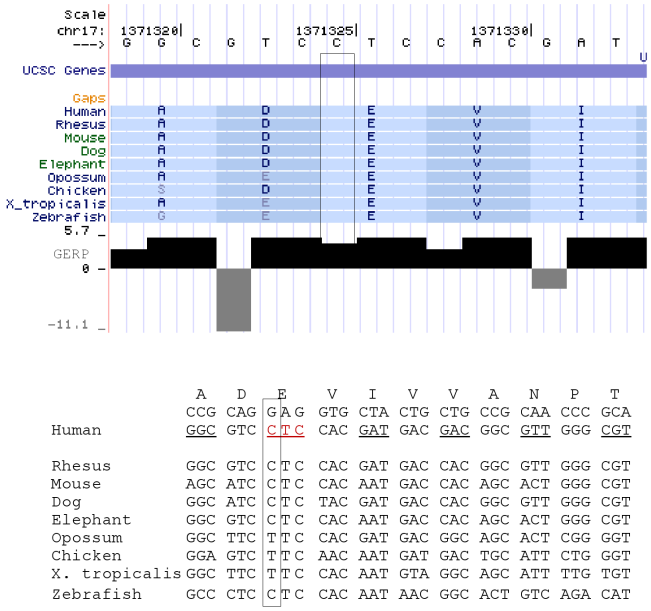


Figure 3.14: Alignment analysis of *MYO1C* variant.

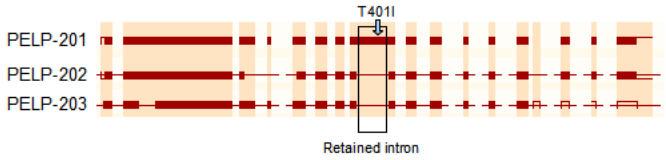


Figure 3.15: Retained intron of *PELP1*.

Table 3.14: Allele frequencies of MYBBP1A p.R671W variant.

Population	A	B	MAF	Ind(#)
Turkish	1213	15	1.22	614
1000genomes	2182	2	0.09	1092
European American	2609	11	0.42	1310
African American	2090	2	0.09	1046

Abbreviations used in this table: MAF, minor allele frequency; Ind (#), Number of individuals genotyped.

higher than the frequency of CAMRQ2. In an additional series of 400 individuals of various European and Middle Eastern ancestries, two were homozygous for MYBBP1A p.R671W. Neither of these two homozygotes had any signs consistent with CAMRQ2. The variant was further analysed in publicly available exome sequence datasets. In 2620 healthy controls of European ancestries, 11 of them was heterozygous for the minor allele (NHLBI Exome Sequencing Project). Therefore, MYBBP1A p.R671W variant was excluded as the allele responsible for the disease.

ZNF594 p.L639F was excluded as the causal mutation based on population screening and conservation considerations. The variant was genotyped in 1098 chromosomes (549 healthy individuals) and detected as heterozygous state in four individuals with an allele frequency of 0.364%. Analysis of various exome and whole genome data (n=4550) revealed that, the variant was present in the different populations (0.183% in HapMap populations, 0.737% in European American NHLBI Exomes, 0.310% in African American NHLBI Exomes). The ZNF594 p.L639F allele was observed with minor allele frequency of 0.415% in 5099 individuals. The Hardy-Weinberg consistency of the variant and the frequency in European American individuals strongly suggested that the variant could be a rare polymorphism. Furthermore, the variant was present in three of 17 species,

Table 3.15: Allele frequencies of ZNF594 p.L639F variant.

Population	A	B	MAF	Ind(#)
Turkish	1094	4	0,364	549
1000genomes	2180	4	0,183	1092
European American	2154	16	0,737	2170
African American	1284	4	0,310	1288
Total	6712	28	0,415	5099

Abbreviations used in this table: MAF, minor allele frequency; Ind (#), number of individuals genotyped.

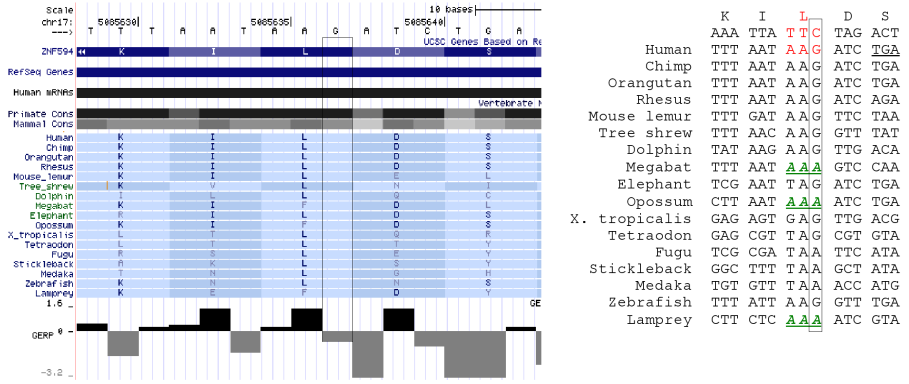


Figure 3.16: Alignment analysis of *ZNF594* variant.

which is also suggesting that phenylalanine at this site would also not be damaging in humans (Figure 3.16). A negative GERP score (-0.665) for the mutated nucleotide indicates that this site is probably evolving neutrally. The effect of amino acid change was evolved by using SIFT [50] and PolyPhen [53] tools and predicted as "benign" (PSIC score difference: 0.301) by PolyPhen and "damaging low confidence" (SIFT Score: 0.04) by SIFT. In addition, human ZNF594 harbors nonsense mutation at site near the ZNF594 p.L639F variant in the HapMap series. rs116878311 at residue 681 (ZNF594 p.Q681X) appeared in one of the 120 CEU controls and rs114754534 at residue 684 (ZNF594 p.E684X) observed in four of 118 Yoruban controls with an allele frequency of 0.034. Considering the genotyping data together with conservation data, exome sequence data and mutation load, ZNF594 p.L639F variant was excluded as the cause of disease (Table 3.15).

3.5.4 WDR81 p.P856L as the disease causing mutation

The remaining variant, which was segregated in the family (Figure 3.13 and 3.17, WDR81 p.P856L, was not observed among the 549 individuals of the control group. The variant is identified in a predicted alternative transcript of *WDR81*. The region containing the mutation was not completely represented in any exome capture array or library. Thus, the region can not be thoroughly analysed using exome sequencing datasets.

To exclude the possibility of homozygous individuals for *WDR81* variant, an extended genealogy of Family B was constructed revealing high levels of consanguinity in a large subset of individuals. Genotyping of WDR81 p.P856L in 177 individuals ruled out the presence of homozygotes (Figure 3.18A). However, 27/177 carriers were identified (Figure 3.18B) and genetic counseling is in progress. A carrier frequency of 14.8% and a mutation frequency of 10.7% were observed in this large genealogy. It was noted that the parents of the affected individuals are the only carrier members married to each other. Genotyping of neighbouring ZNF594 p.L639F and MYBBP1A p.R671W variant in the same genealogy suggested a single ancestral founder haplotype carrying the mutation (Figure 3.18C).

The status of *WDR81* was evaluated in 2 different cohorts of patients with neurodevelopmental/cerebellar phenotypes in which the causative mutations are not known. The first cohort consisted of 750 individuals afflicted by cortical malformations of the brain or degenerative neurological phenotypes. According to the genotyping data based on Illumina Human 370 Duo or 610K Quad BeadChips, none of the patients with cerebellar or ataxia phenotypes harboured a homozygous region ($\geq$ 2.5 cM) surrounding the *WDR81* locus. Exome sequencing data of these individuals did not reveal any mutations in the sequenced regions of *WDR81* gene. The second cohort consisted of 58 probands of which 12 had cerebellar phenotype with or without quadrupedal locomotion. Analysis of the coding regions of the *WDR81* in these patients by using Sanger sequencing did not reveal any mutations.

WDR81 is highly conserved protein throughout vertebrates. Particularly the

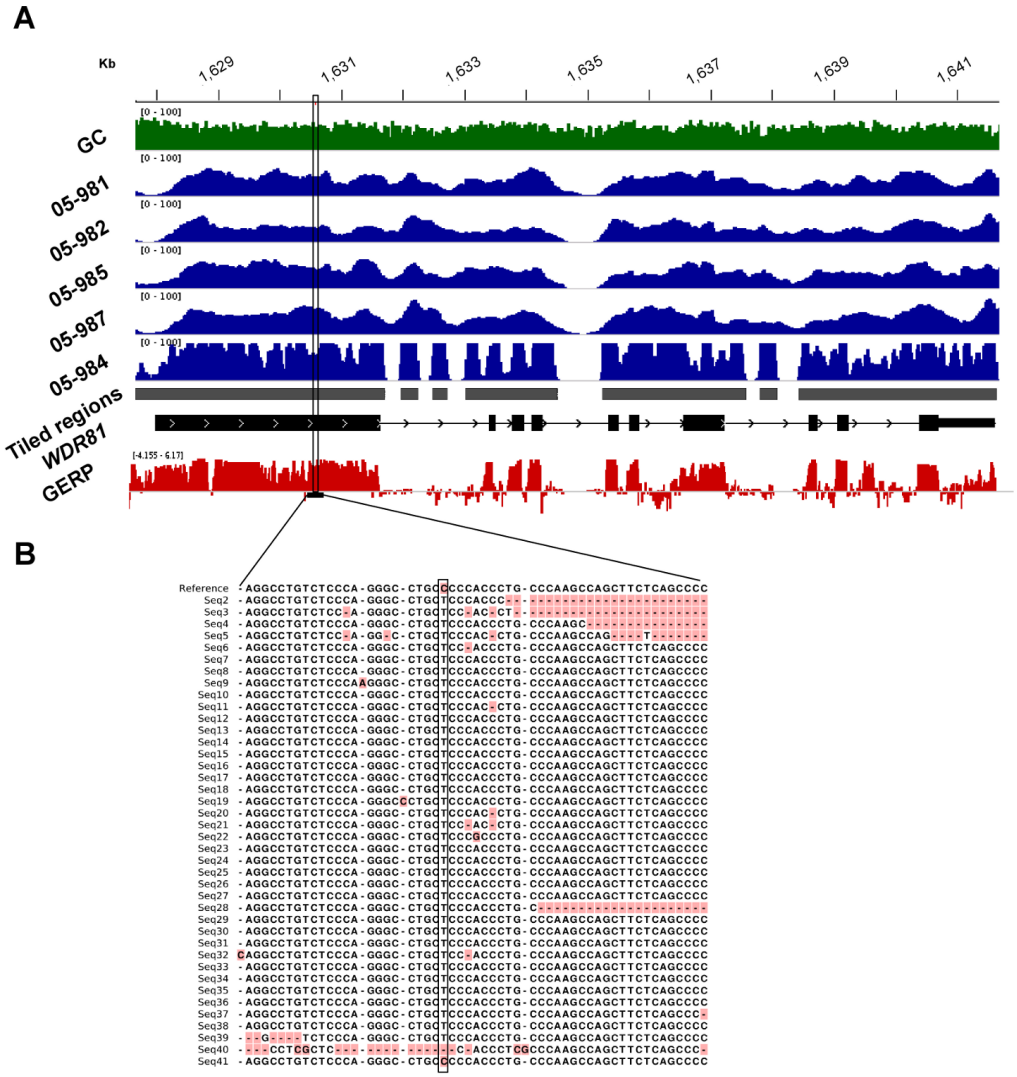


Figure 3.17: NGS coverage statistics and conservation of *WDR81*. GC content, NGS read depths, Nimblegen target probes, exon-intron structure and GERP conservation scores are shown in (A). Alignments of sequence reads for affected individual 05-985

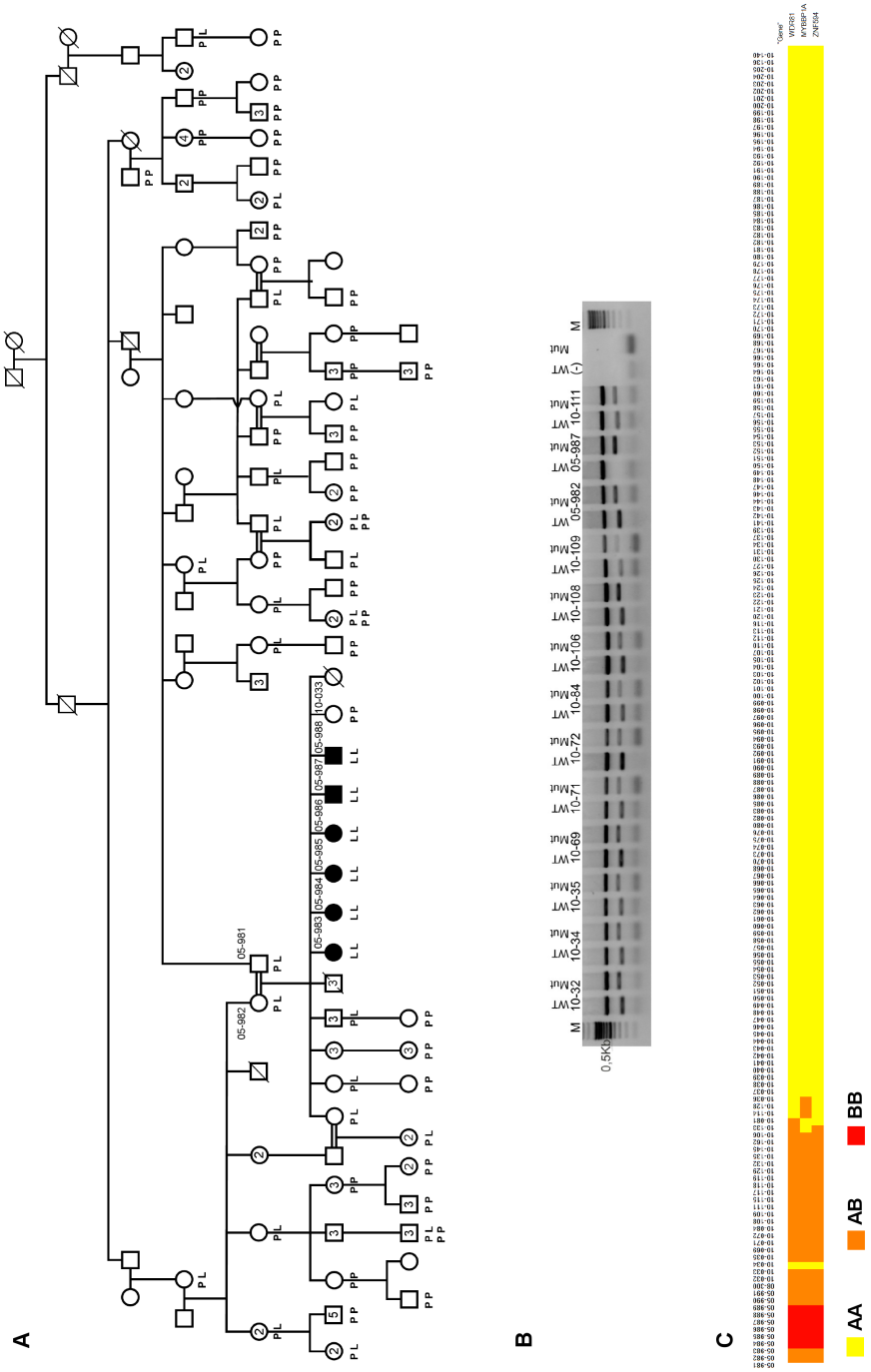


Figure 3.18: Genotyping *WDR81* p.P856L in the extended pedigree.(A) Family B with affected individuals indicated by filled symbols and genotypes shown for WDR81 p.P856L (From Gulsuner et. al. 2011 [11] by permission).(B) 1.5% agarose gel electrophoresis of AS-PCR genotyping poducts of WDR81 p.P856L variant carriers.(C) Lower figure show the genotypes of three variants in all individuals, which suggests a single ancestral founder gaplotype carrying the mutation.

mutated proline residue is completely conserved (Figure 3.20). The gene has no nonsense polymorphisms in any sequenced species. WDR81 p.P856L was predicted to be "damaging" (SIFT score: 0) by SIFT, "probably damaging" (PSIC score difference: 2.724) by PolyPhen, and "under evolutionary constraint" (GERP score: 5.68) by GERP (Figure 3.17). Together with the population data, conservation considerations and the exclusion of the remaining variants, WDR81 p.P856L is concluded as the mutation responsible for the disease.

3.6 Characterization of WDR81

WDR81 p.P856L mutation (chr17:1,630,820 [hg19]) is located in the first exon of the longest isoform of *WDR81* (ENST00000409644, NM_001163809.1, NP_001157281.1). The isoform consisted of 10 exons (Figure 3.19). The protein product of *WDR81* is highly conserved throughout vertebrates. In particular the proline residue is completely conserved in all known sequences, including the most distantly related ortholog Tetraodon nigroviridis WDR81 which is 47.8% identical, 57.2% similar, and has a distance score of 0.76 compared to the human protein (Figure 3.20). Human *WDR81* isoform 1 encodes 1941 amino acids. The protein consisted of a BEACH domain at amino acids 352-607, a major facilitator superfamily (MFS) domain, and six WD-repeats. The WDR81 p.P856L lies in the MFS domain. According to membrane spanning domain prediction analysis, WDR81 is transmembrane protein with six membrane spanning domains. The most N-terminal domain is located at residues 45-66 and the remaining five domains are located at the C-terminus of the protein (residues 980-1815) (Figure 3.19).

At present, the biological function of WDR81 is not known and a knock out mouse model is not available. However some clues about its function can be derived from its protein domains. BEACH (Beige and Chédiac Higashi) domain containing proteins have been implicated in membrane trafficking [89], synapse morphogenesis [90] and lysosomal axon transport [91].

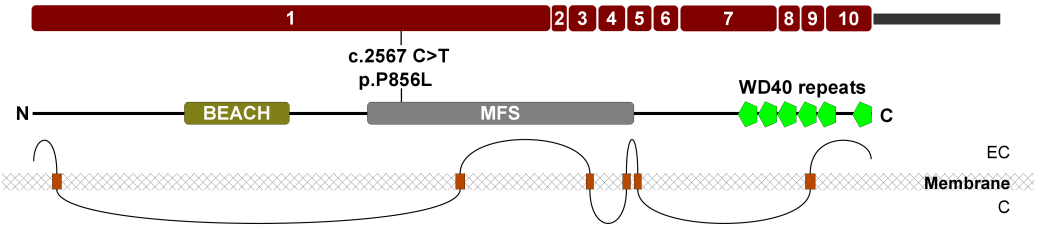


Figure 3.19: Exon-intron structure, protein domains and membrane spanning domains of *WDR81*. (From Gulsuner et. al. 2011 [11] by permission).

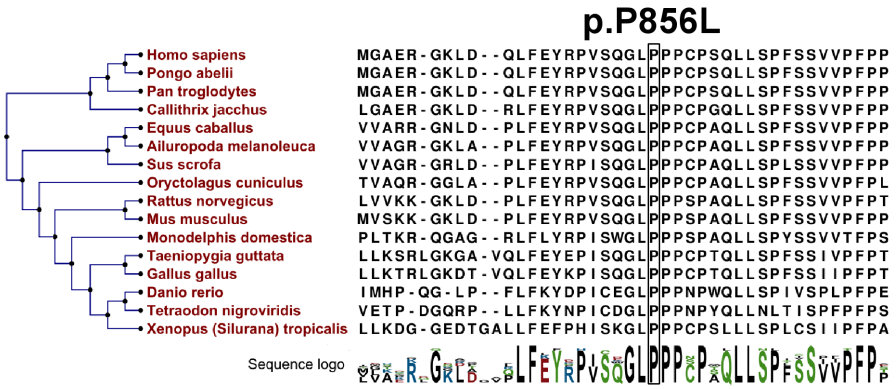


Figure 3.20: Conservation of WDR81 among several species

Figure 3.21: Proteins with BEACH domain and WD repeats.

The protein show similarity with a group of proteins with BEACH domain and WD40 repeats such as NSMAF (neutral sphyngomyelinase activation associated factor), NBEA (neurobeachin) and LYST (lysosomal trafficking regulator) (Figure 3.21). LYST gene contains a BEACH domain, seven WD40 repeats and additional HEAT/ARM repeats [92]. Nearly all *LYST* mutations lead to protein truncation and cause Chediak-Higashi syndrome (CHS), which is characterized by accumulation of giant intracellular vesicles leading to defects in the immune and blood systems [93]. Two missense *LYST* mutations was reported so far [94]. Interestingly, these patients did not represent any immunological involvement but neurological symptoms. A mouse mutant model ($Lyst^{Ing3618}/Lyst^{Ing3618}$) harbour a missense mutation in WD40 domain of *LYST* gene. Purkinje cell degeneration and age dependent impairment of motor coordination without signs of lysosomal defects and immunological symptoms were characteristics of these animals [93].

3.6.1 Expression of *WDR81*

To reveal the full length transcript of the gene and to analyse its unpredicted transcript isoforms containing the mutation site, an exon spanning RT-PCR approach was conducted. Each exon was analysed by amplification of human cerebellar cDNA using a forward primer targeting the mutation site (280 bp upstream) and reverse primers targeting each of the exons. Single bands with appropriate sizes to the gene structure was obtained for all exons (Figure 3.22).

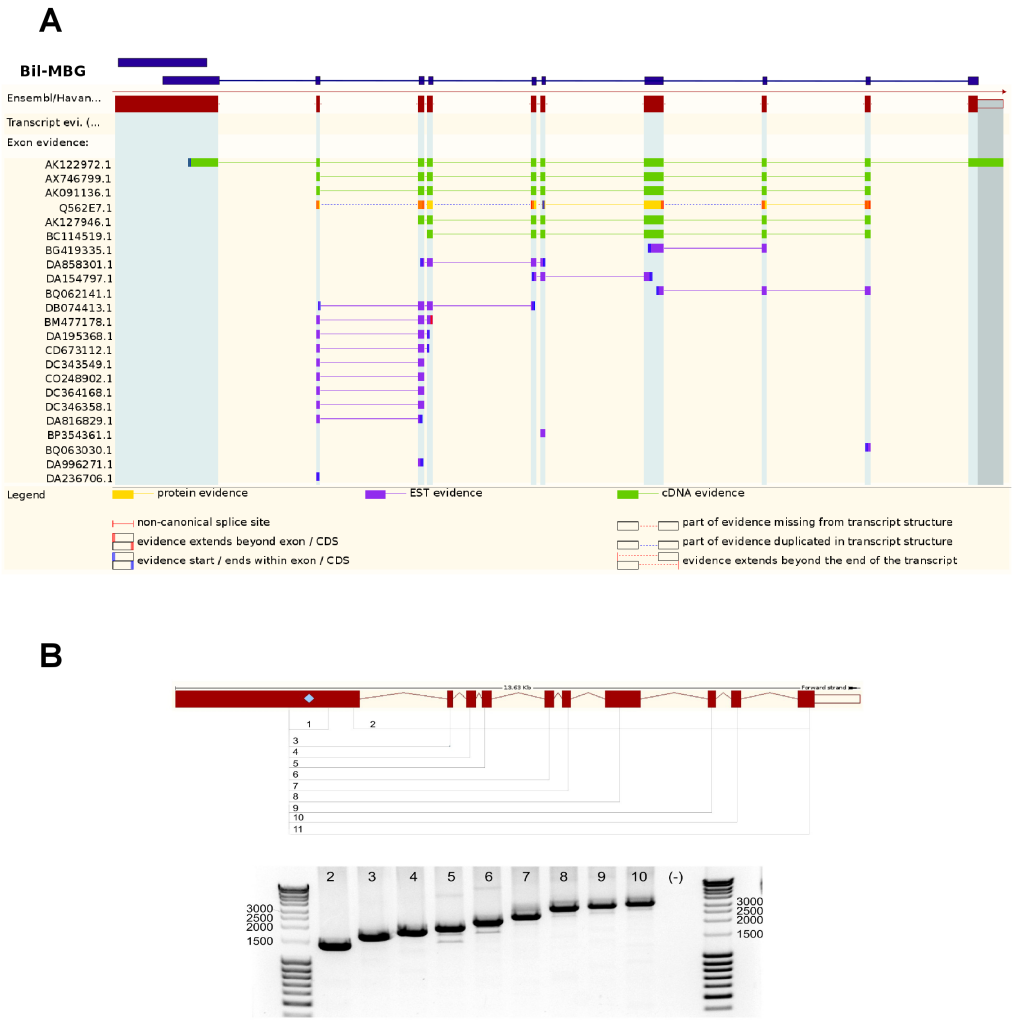


Figure 3.22: Evidences of predicted transcript of *WDR81*. (A) Evidences are based on EST and cDNA level libraries. Full length transcript containing exon 1 was not observed in any libraries (http://www.ensembl.org). Blue bar at the top of the figure is shown the evidences found by RT-PCR experiments. (B) Exon spanning RT-PCR strategy revealed the presence of the isoform in human cerebellum. Mutation site is shown with a blue box.

Expression of the ENST00000409644 transcript of *WDR81* was analysed by real-time reverse transcriptase PCR (RT-PCR) in a multiple human tissue panel. The flanking region containing the mutation in exon 1 was amplified in expression analysis. Human *WDR81* was detected in all human tissues, with the highest levels of expression in cerebellum and corpus callosum (Figure 3.23). Main anatomical abnormalities were shown in these two brain regions.

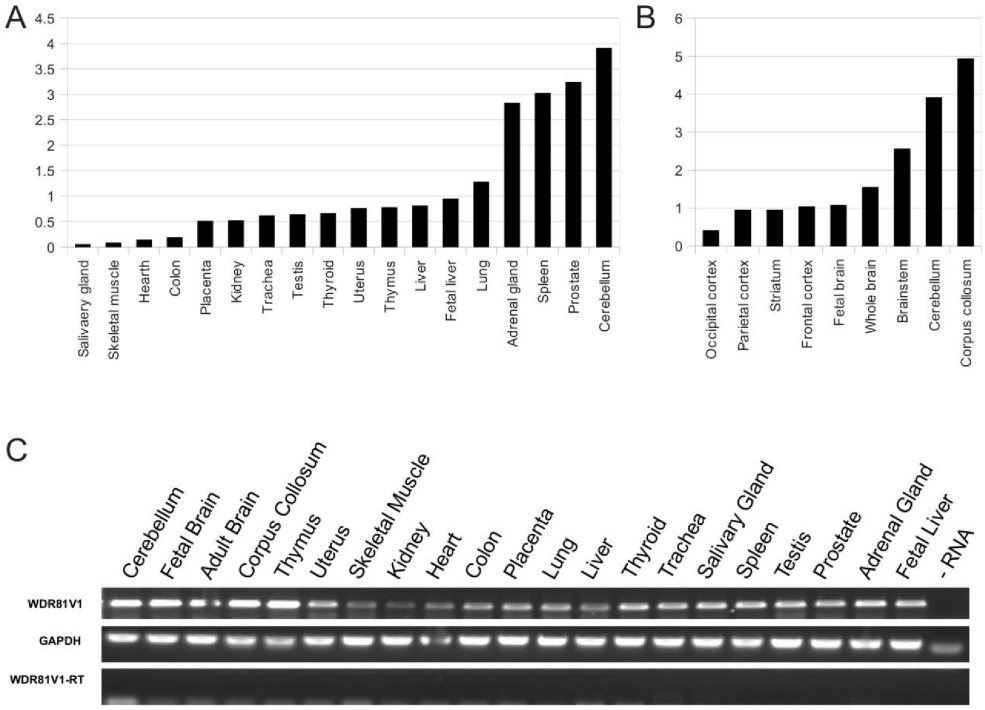


Figure 3.23: Expression of *WDR81* in different tissues. The region containing the mutation was expressed in all tissues(A,B). The highest expression levels in brain tissues were observed in corpus callosum and cerebellum. Multiple tissue panel Agarose gel electrophoresis of qPCR product shows no correspanding band in DNAseI digested -RT samples (C). (WDR81V1: 429bp, GAPDH 143bp).

3.6.1.1 In-situ hybridization analysis of WDR81 in mouse brain

Alignment of human and mouse WD repeat domain 81 proteins (1,941 and 1,934 predicted residues, respectively) showed 87.5% identity and 90.7% similarity. Mouse *Wdr81* maps to chromosome 11 within conserved synteny of a part of

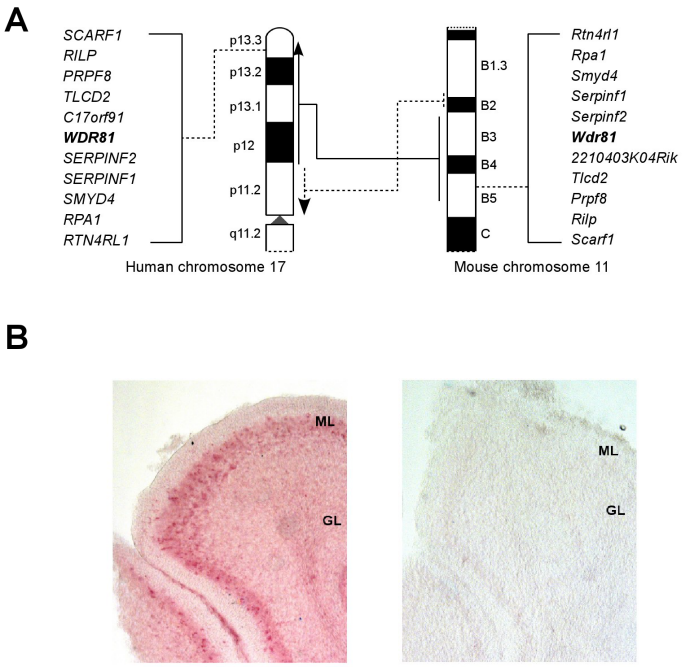


Figure 3.24: (A) Physical location of WDR81/Wdr81 in human and mouse chromosomes. (B) In situ hybridization of mouse embryonic brain revealing increased expression of Wdr81 in purkinje cells and molecular layer of cerebellum (left). No hybridization was observed with the sense probe (right). ML: Molecular Layer, GL: Granular Layer. (From Gulsuner et. al. 2011 [11] by permission).

human chromosome 17p, which also includes the 6.9 Mb CAMRQ 2 linkage interval (Figure 3.24A). Based on the similarity of sequences and conserved synteny, probes that contain the mutation site in human patients were prepared from mouse genomic DNA and specific expression pattern of *Wdr81* gene in mouse brain was analyzed. At P7, *Wdr81* expression was observed in Purkinje cell layer in the cerebellum (Figure 3.24B). No hybridization was observed with the sense probe. These results are consistent with the proposed function of WDR81 protein in mediating motor behavior.

3.6.1.2 Western blot analysis of human WDR81

ENST00000409644 transcript encodes a 212 kDa protein. Anti-WDR81 antibody recognizes exon-7 of the transcript and is predicted to target 212, 96, 77, 75, and 63 kDa splice variants of the gene. Western blot analysis using protein medley from adult brain showed bands approximately corresponds to the 75, 77 and possibly 96 kDa protein variants with additional bands present (Figure 3.25). However no bands was observed at 212 kDa sizes with current antibody. This results suggested the possibility that protein level expression of *WDR81* could be specific to fetal tissues and additional biochemical studies are required to characterize the protein product of *WDR81*.

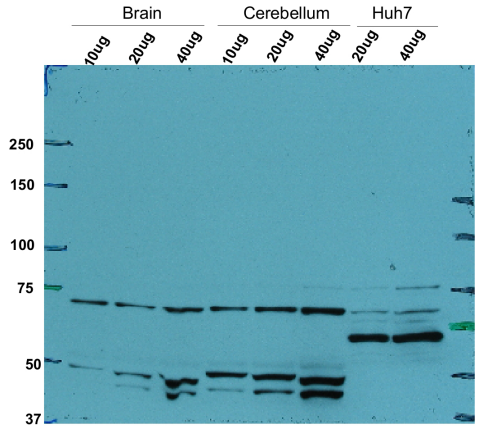


Figure 3.25: Western blot analysis of *WDR81* protein products.

3.6.2 Effect of the mutation in gene expression

The effect of the mutation in expression levels of the gene was investigated by comparing the expression levels between affected individuals, mutation carriers and wild type allele carriers. However, no significant change was observed (Figure 3.26). It could be due the fact that, *WDR81* expression levels were vary among healthy individuals.

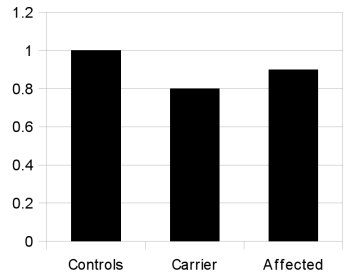


Figure 3.26: Effect of the mutation on *WDR81* expression levels.

3.6.3 Annotation clustering of developing mouse brain expression profiles

As a first step to decipher the function of WDR81, expression data set of early embryonic mouse brain tissues (GSE8091)[82] were evolved. Data were grouped by embryonic days, then differentially expressed genes within day groups were filtered (One-way ANOVA Test Bonferroni corrected $p<0.001$, n=3,611). A total of 670 genes were identified whose expression profiles were significantly correlated with that of WDR81 (R=0.95-1.0). DAVID functional annotation clustering tool [83] was used evaluate the predicted functions of those genes. Functional annotation (Appendix G suggested that positively correlated genes were enriched for those neuron projection proteins (Benferroni adjusted P = 2.3E-11), axonogenesis proteins (Benferroni adjusted P = 1.3E-9) and cell morphogenesis proteins involved in neuron differentiation (Benferroni adjusted P = 3.7E-9). Several genes linked to human diseases with neuronal phenotypes, such as PRNP, DCX and EL1CAM were found in these enrichment sets (Table 3.16). *WDR81* was co-expressed with mental retardation, quadrupedal gait and cerebellar hypoplasia associated gene *ATP8A2* gene (unpublished) suggesting that two genes could represent similar developmental pathways.

Furthermore a data mining analysis was performed to obtain clues to understand possible roles of *WDR81*. In a study that aims genome scale identification of membrane-associated human mRNAs (GSE4175) [85], WDR81 transcript was

Table 3.16: Neuronal phenotype genes with correlated expression profiles with WDR81.

Symbol	Name	Disease	OMIM	R
LTM2B	integral membrane protein 2B	Cerebral Amyloid Angiopathy, Itm2b-Related, 1	176500	0.981
PRNP	prion protein	Gerstmann-Straussler Disease; GSD	137440	0.993
LOX	lysyl oxidase	Menkes Disease	309400	0.995
DCX	Doublecortin	Lissencephaly 1; LIS1	607432	0.982
L1CAM	L1 cell adhesion molecule	MASA Syndrome	303350	0.984

found to be increased in membrane associated RNA in contrast to cytoplasmic RNA (4.14 folds [p=0.03] and 1.78 folds [p=0.0002] in two different datasets), raising the possibility that WDR81 may encode a membrane-associated protein.

Chapter 4

Discussion

Human disease gene identification has been greatly facilitated with the targeted enrichment of the genome and next-generation sequencing.[41] However, finding a causative mutation in the absence of 2 or more independent cases remains a challenge. This is particularly true when, there are multiple variants without any obvious effect on protein and the candidate mutation is identified in an uncharacterised gene.

Cerebellar ataxia, mental retardation, and dysequilibrium syndrome (CAMRQ) is a genetically heterogeneous and extremely rare disease characterized by congenital cerebellar hypoplasia, cerebellar ataxia and mental retardation with or without quadrupedal locomotion.[9, 10] CAMRQ was first described in a unique consanguineous family. In the context of this study, we describe the identification of a missense mutation in family B associated with CAMRQ2. The minimal critical region, which consisted of approximately 192 genes, was captured and sequenced using the DNA samples of three affected and two carrier individuals. The coverage analysis revealed that more then 99% of the coding bases were sequenced. Approximately 26,000 variants were detected by mapping the sequence data to the human reference genome. Variants were annotated by using a custom annotation pipeline. Three functional variants (WDR81 p.P856L, MYBBP1A p.R671W, ZNF594 p.L639F), which are consistent with

the Mendelian inheritance of the disease allele, was found by segregation analysis. High quality targeted next generation sequencing excluded the possibility of other disease related variants in the minimal critical region.

The causative mutation was identified by using several filtering steps. After segregation analysis, previously unreported rare polymorphisms were filtered out by population screening, analysis of the exome data of control populations, analysis of conservation at candidate variant sites and presence of polymorphic nonsense variants near the mutation site. In addition, an extended kinred of the family spanning five generations was constructed.

We conclude that, the WDR81-P856L mutation is the cause of CAMRQ2 in Family B. The p.P856L mutation was identified in the first exon of the longest isoform of *WDR81* gene. Expression of this hypothetical transcript variant was confirmed by RT-PCR analysis. Expression pattern of the transcript was analysed by quantitative RT-PCR experiments in a multiple human tissue panel and by in situ hybridization of mouse brain sections. Expression of *Wdr81* was observed in the Purkinje cell layer of the mouse cerebellum. Highest levels of expression in brain was detected in the cerebellum and corpus callosum, where major structural abnormalities was observed.

At present, the biological function of WDR81 is not known and a knock out mouse model is not available. But the function of the protein can be predicted from its protein domains. Bioinformatics tools predicted that the WDR81 protein is a membrane protein and composed of a BEACH domain, 6 WD40-repeats and a major facilitator superfamily domain in which the mutation resides.

WD repeats are short, approximately 40 amino acid motifs, which typically contain Trp-Asp (WD) dipeptide at the C-terminus. WD repeats are present in approximately 122 different genes (`http://www.genenames.org`) and are implicated in a variety of functions ranging from transcription regulation to cell cycle. They are thought to form a beta-propeller structure that serves as a platform for the assembly of protein complexes. Specific functions of the proteins are determined by the remaining sequence.[95, 96, 97] Some WD-repeat proteins (*SEC13* and *RACK1*) consisted of only WD-repeat domain[98]. According to

experimental data, WD repeat domains themselves have important functions. [99] There are several WD-repeat genes implicated in inherited diseases. The first WD-repeat gene identified as a cause of human disease is lissencephaly-1 (*LIS1/PAFAH1B1*). Mutations in *LIS1* gene appears to cause lissencephaly in humans.[100] Recently, mutations in WD repeat-containing protein 62 (*WDR62*) were identified as the cause of a wide spectrum of cerebral cortical malformations including microcephaly [41] and hypoplasia of the corpus callosum[101].

BEACH domain is found in eukaryotic proteins with diverse cellular functions ranging from membrane trafficking [89], to synapse morphogenesis [90] and lysosomal axon transport [91]. The function of the BEACH domain is still to be identified. The BEACH domain is generally followed by WD40 repeats. The BEACH domain was first described in the *LYST* (*CHS*) gene encoding lysosomal trafficking regulator protein. The mouse homologous of the LYST protein is Beige. The name for BEACH stands for the mouse Beige and human CHS (Chediak-Higashi syndrome) proteins.

LYST gene contains a BEACH domain followed by seven WD40 repeats and additional HEAT/ARM repeats [92]. Nearly all *LYST* mutations cause protein truncation and lead to CHS. CHS is an autosomal recessive condition characterized by accumulation of giant intracellular vesicles leading to defects in the immune and blood systems [93]. Consequently patients exhibit recurrent infections, albinism and progressive neuropathy.[92, 102] Two missense mutations in the *LYST* (R1563H and V1999D) have been reported so far.[94] The affected individuals had various neurological symptoms but did not have an infection history. The $Lyst^{Ing3618}/Lyst^{Ing3618}$ mice which harbours a missense mutation in WD40 domain, presents purkinje cell degeneration and impairment of motor coordination without signs of immunological symptoms. [93].

Subsequently, several proteins with BEACH domain and WD40 repeats were identified including NSMAF (neutral sphyngomyelinase activation associated factor) and NBEA (neurobeachin). NBEA has a brain specific expression. Mutations in this gene are implicated in a form of autism.[103]

CAMRQ has been linked to at least three different chromosomes. The first

gene on chromosome 9p24 was identified as *VLDLR* and the second mutation was identified in *CA8* gene. In this study, *WDR81* was identified as the third gene responsible for CAMRQ. Highest levels of expression of *WDR81* was observed in human cerebellum and corpus callosum. In mouse brain, *Wdr81* transcript was observed in Purkinje cell layer of cerebellum. These findings are consistent with our observations of major structural abnormalities of the patients were in these regions. These observations suggest a possible role for *WDR81* in motor behaviour. Further work will be required to understand the normal biological function of WDR81 and the role of the mutation in causing cerebellar hypoplasia and quadrupedal locomotion. Genomic analysis of Family B demonstrates that WDR81 is highly likely to be critical to these developmental processes.

Chapter 5

Future prospects

Cerebellar ataxia and mental retardation with or without quadrupedal locomotion (CAMRQ) has linked to three different chromosomes with gene identification studies. After the identification of *VLDLR* and *CA8* mutations as the cause of CAMRQ, a missense mutation was identified in *WDR81*. In addition, several families were identified and their genetic studies are ongoing. In developmental diseases which show genetic heterogeneity, identification of several genes would be interesting regarding to reveal a common pathway underlying the same phenotype.

In our study, the disease causing mutation was identified in a hypothetical transcript isoform of an uncharacterised gene. Therefore characterization of the gene structure and identification of the tissue specific novel transcript isoforms would be the first step of further studies. One of the most important approaches in the functional characterization of a new gene is using animal models. Knocking down a gene with a human phenotype would provide us to analyse and reveal the function of the gene in specific developmental processes. Therefore, functional Zebrafish studies have been developed in order to identify the function of the gene in brain development.

After the identification of a causative mutation in an inherited disease, DNA-based diagnostics could become possible. The mutation in *WDR81* can be recognized by AS-PCR method which makes it possible to screen a large number of individuals in an economical way. Our population screening efforts already indicated several carriers from the extended kindred of Family B. Genetic counselling is in progress for this large family.

Identification of other mutations in the same gene in similar phenotypes would also be interesting. For that purpose, the most efficient way is including all regions of this gene in exome enrichment libraries. By this way, it would be possible to analyse contributions of other variants of the gene to health and disease.

Bibliography

[1] B. G. Richmond and D. S. Strait, "Evidence that humans evolved from a knuckle-walking ancestor," *Nature*, vol. 404, pp. 382–385, 2000.

[2] Z. Alemseged, F. Spoor, W. Kimbel, R. Bobe, D. Geraads, D. Reed, and J. Wynn, "A juvenile early hominin skeleton from dikika, ethiopia," *Nature*, vol. 443, pp. 296–301, 2006.

[3] D. M. Bramble and D. E. Lieberman, "Endurance running and the evolution of homo," *Nature*, vol. 432, pp. 345–352, 2004.

[4] F. Spoor, B. Wood, and F. Zonneveld, "Implications of early hominid labyrinthine morphology for evolution of human bipedal locomotion," *Nature*, vol. 369, pp. 645–648, 1994.

[5] B. Wood, "Palaeoanthropology: A precious little bundle," *Nature*, vol. 443, pp. 278–281, 2006.

[6] H. Fukuyama, Y. Ouchi, S. Matsuzaki, Y. Nagahama, H. Yamauchi, M. Ogawa, J. Kimura, and H. Shibasaki, "Brain functional activity during gait in normal subjects: aspect study," *Neurosci Lett*, vol. 228, pp. 183–189, 1997.

[7] S. Morton and A. Bastian, "Mechanisms of cerebellar gait ataxia," *Cerebellum*, vol. 6, pp. 79–165, 2007.

[8] B. L. Fogel and S. Perlman, "Clinical features and molecular genetics of autosomal recessive cerebellar ataxias," *Lancet Neurol*, vol. 6, pp. 245–257, 2007.

[9] U. Tan, "Unertan syndrome: A new theory on the evolution of human mind," *Neuroquantology*, vol. 4, pp. 250–255, 2005.

[10] S. Turkmen, O. Demirhan, K. H. an A. Diers, C. Zimmer, K. Sperling, and S. Mundlos, "Cerebellar hypoplasia and quadrupedal locomotion in humans as a recessive trait mapping to chromosome 17p," *J Med Genet*, vol. 43, pp. 461–464, 2006.

[11] S. Gulsuner, A. B. Tekinay, K. Doerschner, H. Boyaci, K. Bilguvar, H. Unal, A. Ors, O. E. Onat, E. Atalar, A. N. Basak, H. Topaloglu, T. Kansu, M. Tan, U. Tan, M. Gunel, and T. Ozcelik, "Homozygosity mapping and targeted genomic sequencing reveal the gene responsible for cerebellar hypoplasia and quadrupedal locomotion in a consanguineous kindred," *Genome Res*, vol. Epub ahead of print, 2011.

[12] T. Ozcelik, N. Akarsu, E. Uz, S. Caglayan, S. Gulsuner, O. E. Onat, M. Tan, and U. Tan, "Mutations in the very low-density lipoprotein receptor vldlr cause cerebellar hypoplasia and quadrupedal locomotion in humans," *Proc Natl Acad Sci U S A*, vol. 105, pp. 4232–4236, 2008.

[13] U. Tan, "Unertan syndrome: review and rep ort of four new cases," *Int J Neurosci*, vol. 118, pp. 211–225, 2008.

[14] U. Tan, "Evidence for unertan syndrome and the evolution of the human mind," *Int J Neurosci*, vol. 116, pp. 763–774, 2006.

[15] S. Gulsuner, K. Bilguvar, M. Tan, U. Tan, M. Gunel, and T. Ozcelik, "Targeted next generation sequencing identifes a mutation associated with cerebellar hypoplasia and mental retardation with quadrupedal locomotion. european human genetics conference 2010, gothenburg, sweden, june 12 - 15," *Eur J Hum Genet*, vol. 18, p. S1:C12.3, 2010.

[16] H. Leiner, "Solving the mystery of the human cerebellum," *Neuropsychol Rev*, vol. 20, pp. 229–235, 2010.

[17] S. A. Sajan, K. E. Waimey, and K. J. Millen, "Novel approaches to studying the genetic basis of cerebellar development," *Cerebellum*, vol. 9, pp. 272–283, 2010.

[18] A. C. Guyton and J. E. Hall, *Textbook of Medical Physiology.* Elsevier, 11th ed., 2006.

[19] D. Purves and S. Williams, *Neuroscience.* Sinauer Associates, Inc., 4nd ed., 2007.

[20] E. R. Kandel and J. H. Schwartz, eds., *Principles of Neural Science*, ch. The cerebellum, pp. 502–522. Elsevier, 2nd ed., 1985.

[21] K. RE, *Concise Text of Neuroscience.* Lippincott Williams and Wilkins, 2nd ed., 2000.

[22] M. Manto and D. Marmolino, "Cerebellar ataxias," *Curr Opin Neurol*, vol. 22, pp. 419–429, 2009.

[23] S. Gilman, G. Wenning, P. Low, D. Brooks, C. Mathias, J. Trojanowski, N. Wood, C. C. C, A. Durr, C. Fowler, H. Kaufmann, T. Klockgether, A. Lees, W. Poewe, N. Quinn, T. Revesz, D. Robertson, P. Sandroni, K. Seppi, and M. Vidailhet, "Second consensus statement on the diagnosis of multiple system atrophy," *Neurology*, vol. 71, pp. 670–676, 2008.

[24] S. Di Donato, C. Gellera, and C. Mariotti, "The complex clinical and genetic classification of inherited ataxias ii. autosomal recessive ataxias," *Neurol Sci*, vol. 22, pp. 219–228, 2001.

[25] S. Banfi and H. Zoghbi, "Molecular genetics of hereditary ataxias," *Baillieres Clin Neurol*, vol. 3, pp. 281–95, 1994.

[26] C. Mariotti and S. Di Donato, "Cerebellar/spinocerebellar syndromes," *Neurol Sci*, vol. 22, pp. S88–92, 2001.

[27] A. Garrod, "About alkaptonuria," *Med Chir Trans*, vol. 85, pp. 69–78, 1902.

[28] T. Ozcelik, M. Kanaan, K. Avraham, D. Yannoukakos, A. Megarbane, G. Tadmouri, L. Middleton, G. Romeo, M. King, and E. Levy-Lahad, "Collaborative genomics for human health and cooperation in the mediterranean region.," *Nat Genet*, vol. 42, pp. 641–645, 2010.

[29] A. Bittles, "Consanguinity and its relevance to clinical genetics," *Clin Genet*, vol. 60, pp. 89–98, 2001.

[30] E. Lander and D. Botstein, "Homozygosity mapping: a way to map human recessive traits with the dna of inbred children.," *Science*, vol. 236, pp. 1567–1570, 1987.

[31] E. Lander and D. Botstein, "Homozygosity mapping: a way to map human recessive traits with the dna of inbred children," *Science*, vol. 236, pp. 1567–1570, 1987.

[32] M. van Driel, K. Cuelenaere, P. Kemmeren, J. Leunissen, H. Brunner, and G. Vriend, "Geneseeker: extraction and integra- tion of human disease-related information from web-based genetic databases," *Nucleic Acids Res*, vol. 33, pp. W758–W761, 2005.

[33] L. Franke, H. van Bakel, L. Fokkens, E. de Jong, M. Egmont-Petersen, and C. Wijmenga, "Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes," *Am J Hum Genet*, vol. 78, pp. 1011–1125, 2006.

[34] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori, "The KEGG resource for deciphering the genome," *Nucleic Acids Res*, vol. 32, pp. D277–D280, 2004.

[35] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. Déustachio, E. Schmidt, B. de Bono, B. Jassal, G. Gopinath, G. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein, "Reactome: a knowledgebase of biological pathways," *Nucleic Acids Res*, vol. 33, pp. D428–D432, 2005.

[36] S. Peri, J. Navarro, R. Amanchy, T. Kristiansen, C. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T. Gandhi, M. Gronborg, N. Ibarrola, N. Deshpande, K. Shanker, H. Shivashankar, B. Rashmi, M. Ramya, Z. Zhao, K. Chandrika, N. Padma, H. Harsha, A. Yatish, M. Kavitha, M. Menezes, D. Choudhury, S. Suresh, N. Ghosh, R. Saravana, S. Chandran, S. Krishna, M. Joy, S. Anand, V. Madavan, A. Joseph, G. Wong, W. Schiemann, S. Constantinescu, L. Huang, R. K. ravi Far,

H. Steen, M. Tewari, S. Ghaffari, G. Blobe, C. Dang, J. Garcia, J. Pevsner, O. Jensen, P. Roepstorff, K. Deshpande, A. Chinnaiyan, A. Hamosh, A. Chakravarti, and A. Pandey, "Development of human protein reference database as an initial platform for approaching systems biology in humans," *Genome Res*, vol. 13, pp. 2363–2371, 2003.

[37] N. Lopez-Bigas and C. Ouzounis, "Genome-wide identification of genes likely to be involved in human genetic disease," *Nucleic Acids Res*, vol. 32, pp. 3108–3114, 2004.

[38] E. Adie, R. Adams, K. Evans, D. Porteous, and B. Pickard, "Speeding disease gene discovery by sequence based candidate prioritization," *BMC Bioinformatics*, vol. 6, p. 55, 2005.

[39] E. Adie, R. Adams, K. Evans, D. Porteous, and B. Pickard, "Suspects: enabling fast and effective prioritization of positional candidates," *Bioinformatics*, vol. 22, pp. 773–774, 2006.

[40] S. Ng, E. Turner, P. Robertson, S. Flygare, A. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E. Eichler, M. Bamshad, D. Nickerson, and J. Shendure, "Targeted capture and massively parallel sequencing of 12 human exomes," *Nature*, vol. 461, pp. 272–276, 2009.

[41] K. Bilguvar, A. Ozturk, A. Louvi, K. Kwan, M. Choi, B. Tatli, D. Yalnizoglu, B. Tuysuz, A. Caglayan, S. Gokben, H. Kaymakcalan, T. Barak, M. Bakircioglu, K. Yasuno, W. Ho, S. Sanders, Y. Zhu, S. Yilmaz, A. Dincer, M. Johnson, R. Bronen, N. Kocer, H. Per, S. Mane, M. Pamir, C. Yalcinkaya, S. Kumandas, M. Topcu, M. Ozmen, N. Sestan, R. Lifton, M. State, and M.Gunel, "Whole-exome sequencing identifies recessive wdr62 mutations in severe brain malformations," *Nature*, vol. 467, pp. 207–210, 2010.

[42] A. Chakravarti, "Genomic contributions to mendelian disease," *Genome Res*, vol. 21, pp. 643–644, 2011.

[43] G. Cooper and J. Shendure, "Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data," *Nat Rev Genet*, vol. 12, pp. 628–640, 2011.

[44] M. Bamshad, S. Ng, A. Bigham, H. Tabor, M. Emond, D. Nickerson, and J. Shendure, "Exome sequencing as a tool for mendelian disease gene discovery," *Nat Rev Genet*, vol. 12, pp. 745–755, 2011.

[45] A. Siepel, G. Bejerano, J. Pedersen, A. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. Hillier, S. Richards, G. Weinstock, R. Wilson, R. Gibbs, W. Kent, W. Miller, and D. Haussler, "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes," *Genome Res*, vol. 15, pp. 1034–1050, 2005.

[46] E. Davydov, D. Goode, M. Sirota, G. Cooper, A. Sidow, and S. Batzoglou, "Identifying a high fraction of the human genome to be under selective constraint using gerp++," *PLoS Comput Biol*, vol. 6, p. e1001025, 2010.

[47] K. Pollard, M. Hubisz, K. Rosenbloom, and A. Siepel, "Detection of non-neutral substitution rates on mammalian phylogenie," *Genome Res*, vol. 20, pp. 110–121, 2010.

[48] E. Stone and A. Sidow, "Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity," *Genome Res*, vol. 15, pp. 978–986, 2005.

[49] E. Capriotti, R. Calabrese, and R. Casadio, "Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information," *Bioinformatics*, vol. 22, pp. 2729–2734, 2006.

[50] H. S. Ng PC, "Predicting deleterious amino acid substitutions," *Genome Res*, vol. 11, pp. 863–874, 2001.

[51] P. Thomas, M. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan, and A. Narechania, "Panther: a library of protein families and subfamilies indexed by function," *Genome Res*, vol. 13, pp. 2129–2141, 2003.

[52] J. Schwarz, C. Rodelsperger, M. Schuelke, and D. Seelow, "Mutationtaster evaluates disease-causing potential of sequence alterations," *Nature Methods*, vol. 7, pp. 575–576, 2010.

[53] S. Sunyaev, V. Ramensky, I. Koch, and B. P. W. Lathe 3rd, Kondrashov AS, "Prediction of deleterious human alleles," *Hum Mol Genet*, vol. 10, pp. 591–597, 2001.

[54] Y. Bromberg and B. Rost, "Snap: predict effect of non-synonymous polymorphisms on function.," *Nucleic Acids Res*, vol. 35, pp. 3823–3835, 2007.

[55] P. Yue, Z. Li, and J. Moult, "Loss of protein structure stability as a major causative factor in monogenic disease," *J Mol Biol*, vol. 353, pp. 459–473, 2005.

[56] J. Sambrook, E. Fritsch, and T. Maniatis, *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor, N.Y: Cold Spring Harbor Laboratory Press, 2nd ed., 1989.

[57] V. Singer, L. Jones, S. Yue, and R. Haugland, "Characterization of picogreen reagent and development of a fluorescence-based solution assay for double-stranded dna quantitation," *Anal Biochem*, vol. 249, pp. 228–238, 1997.

[58] H. Thiele and P. Nürnberg, "Haplopainter: a tool for drawing pedigrees with complex haplotypes," *Bioinformatics*, vol. 21, pp. 1730–1732, 2004.

[59] R. Nussbaum, R. McInnes, and H. Willard, *Thompson & Thompson Genetics in Medicine*. Philadelphia, Pa: W.B. Saunders, 7th ed., 2007.

[60] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. de Bakker, M. Daly, and P. Sham, "Plink: a tool set for whole-genome association and population-based linkage analyses," *Am J Hum Genet*, vol. 81, pp. 559–575, 2007.

[61] G. Abecasis, S. Cherny, W. Cookson, and L. Cardon, "Merlin-rapid analysis of dense genetic maps using sparse gene flow trees," *Nat Genet*, vol. 30, pp. 97–101, 2002.

[62] M. Seelow, M. Schuelke, F. Hildebrandt, and P. Nürnberg, "Homozygositymapper - an interactive approach to homozygosity mapping," *Nucleic Acids Res*, vol. 37, pp. W593–W599, 2009.

[63] C. Prieto and J. De Las Rivas, "Apid: Agile protein interaction dataanalyzer," *Nucl. Acids Res.*, vol. 34, pp. W298–W302, 2006.

[64] E. Adie, R. Adams, K. Evans, D. Porteous, and B. Pickard, "Speeding disease gene discovery by sequence based candidate prioritization," *BMC Bioinformatics*, vol. 6, p. 55, 2005.

[65] E. Adie, R. Adams, K. Evans, D. Porteous, and B. Pickard, "Sus-pects: enabling fast and effective prioritization of posi-tional candidates," *Bioinformatics*, vol. 22, pp. 773–774, 2006.

[66] H. Orr and H. Zoghbi, "Trinucleotide repeat disorders," *Annu Rev Neurosci*, vol. 30, pp. 575–621, 2007.

[67] S. Mirkin, "Expandable dna repeats and human disease," *Nature*, vol. 447, pp. 932–940, 2007.

[68] C. Pearson and J. C. K. Nichol Edamura, "Repeat instability: mechanisms of dynamic mutations," *Nat Rev Genet*, vol. 6, pp. 729–742, 2005.

[69] A. La Spada and J. Taylor, "Repeat expansion disease: progress and puzzles in disease pathogenesis," *Nat Rev Genet*, vol. 11, pp. 247–258, 2010.

[70] Z. Ning, A. Cox, and J. Mullikin, "Ssaha: A fast search method for large dna databases," *Genome Res*, vol. 11, pp. 1725–1729, 2001.

[71] B. Ewing and P. Green, "Base-calling of automated sequencer traces usingphred. II. error probabilities," *Genome Res*, vol. 8, pp. 186–194, 1998.

[72] D. Hedges, D. Burges, E. Powell, C. Almonte, J. Huang, S. Young, B. Boese, M. Schmidt, M. Pericak-Vance, E. Martin, X. Zhang, T. Harkins, and S. Züchner, "Exome sequencing of a multigenerational human pedigree," *PLoS One*, vol. 4, p. e8232, 2009.

[73] H. Li, J. Ruan, and R. Durbin, "Mapping short dna sequencing reads and calling variants using mapping quality scores," *Genome Res*, vol. 18, pp. 1851–1858, 2008.

[74] D. R. Li H, "Fast and accurate long-read alignment with burrows-wheeler transform," *Bioinformatics*, vol. 26, pp. 589–595, 2010.

[75] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, and R. D. G. Abecasis, "The sequence alignment/map format and samtools," *Bioinformatics*, vol. 25, pp. 2078–2079, 2009.

[76] J. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. Lander, G. Getz, and J. Mesirov, "Integrative genomics viewer," *Nat Biotechnol*, vol. 29, pp. 24–26, 2011.

[77] A. Rehman, R. Morell, I. Belyantseva, S. Khan, E. Boger, M. Shahzad, Z. Ahmed, S. Riazuddin, S. Khan, S. Riazuddin, and T. Friedman, "Targeted capture and next-generation sequencing identifies c9orf75, encoding taperin, as the mutated gene in nonsyndromic deafness dfnb79," *Am J Hum Genet*, vol. 86, pp. 378–388, 2010.

[78] P. Danecek, A. Auton, G. Abecasis, C. Albers, E. Banks, M. DePristo, R. Handsaker, G. Lunter, G. Marth, S. Sherry, G. McVean, and R. D. 1000 Genomes Project Analysis Group, "The variant call format and vcftools," *Bioinformatics*, vol. 27, pp. 2156–2158, 2011.

[79] T. Vincze, J. Posfai, and R. Roberts, "Nebcutter: a program to cleave dna with restriction enzymes," *Nucleic Acids Res*, vol. 31, pp. 3688–3691, 2003.

[80] M. Pfaffl, "A new mathematical model for relative quantification in real-time rt-pcr," *Nucleic Acids Res*, vol. 29, p. e45, 2001.

[81] A. Gattiker, E. Gasteiger, and A. Bairoch, "Scanprosite: a reference implementation of a prosite scanning tool," *Applied Bioinformatics*, vol. 1, pp. 107–108, 2002.

[82] D. Hartl, M. Irmler, I. Römer, M. Mader, L. Mao, C. Zabel, M. de Angelis, J. Beckers, and J. Klose, "Transcriptome and proteome analysis of early

embryonic mouse brain development," *Proteomics*, vol. 8, pp. 1257–1265, 2008.

[83] W. Huang da, B. Sherman, and R. Lempicki, "Systematic and integrative analysis of large gene lists using david bioinformatics resources," *Nature Protocols*, vol. 4, pp. 44–57, 2009.

[84] I. Kupershmidt, Q. Su, A. Grewal, S. Sundaresh, I. Halperin, J. Flynn, M. Shekar, H. Wang, J. Park, W. Cui, G. Wall, R. Wisotzkey, S. Alag, S. Akhtari, and M. Ronaghi, "Ontology-based meta-analysis of global collections of high-throughput public data.," *PLoS One*, vol. 5, p. e13066, 2010.

[85] M. Diehn, R. Bhattacharya, D. Botstein, and P. Brown, "Genome-scale identification of membrane-associated human mrnas," *PLoS Genet*, vol. 2, p. e11, 2006.

[86] J. McClellan and M. K. MC, "Genetic heterogeneity in human disease," *Cell*, vol. 141, pp. 210–217, 2010.

[87] R. Lemmers, P. van der Vliet, R. Klooster, S. Sacconi, J. D. P. Camaño, L. Snider, K. Straasheijm, G. van Ommen, G. Padberg, S. T. D.G. Miller, R. Tawil, R. Frants, and S. van der Maarel, "A unifying genetic model for facioscapulohumeral muscular dystrophy.," *Science*, vol. 329, pp. 1650–1653, 2010.

[88] 1000 Genomes Project Consortium, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, pp. 1061–1073, 2010.

[89] R. Khodosh, A. Augsburger, T. Schwarz, and P. Garrity, "Bchs, a beach domain protein, antagonizes rab11 in synapse morphogenesis and other developmental events," *Development*, vol. 133, pp. 4655–4465, 2006.

[90] X. Wang, F. Herberg, M. Laue, C. Wullner, B. Hu, E. Petrasch-Parwez, and M. Kilimann, "Neurobeachin: A protein kinase a-anchoring, beige/chediak-higashi protein homolog implicated in neuronal membrane traffic," *J Neurosci*, vol. 20, pp. 8551–8565, 2000.

[91] A. Lim and R. Kraut, "The drosophila beach family protein, blue cheese, links lysosomal axon transport with motor neuron degeneration," *J Neurosci*, vol. 29, pp. 951–963, 2009.

[92] D. Ward, G. Griffiths, J. Stinchcombe, and J. Kaplan, "Analysis of the lysosomal storage disease Chediak-Higashi syndrome," *Traffic*, vol. 1, pp. 816–822, 2000.

[93] M. Rudelius, A. Osanger, S. Kohlmann, M. Augustin, G. Piontek, U. Heinzmann, G. Jennen, A. Russ, K. Matiasek, G. Stumm, and J. Schlegel, "A missense mutation in the wd40 domain of murine lyst is linked to severe progressive purkinje cell degeneration," *Acta Neuropathol*, vol. 112, pp. 267–76, 2006.

[94] M. Karim, K. Suzuki, K. Fukai, J. Oh, D. Nagle, K. Moore, E. Barbosa, T. Falik-Borenstein, A. Filipovich, Y. Ishida, S. Kivrikko, C. Klein, F. Kreuz, A. Levin, H. Miyajima, J. Regueiro, C. Russo, E. Uyama, O. Vierimaa, and R. Spritz, "Apparent genotype-phenotype correlation in childhood, adolescent, and adult chediak-higashi syndrome," *Am J Med Genet*, vol. 108, pp. 16–22, 2002.

[95] T. Smith, C. Gaitatzes, K. Saxena, and E. Neer, "The wd repeat: a common architecture for diverse functions," *Trends Biochem Sci*, vol. 24, pp. 181–185, 1999.

[96] R. R. Li D, "Wd-repeat proteins: structure characteristics, biological function, and their involvement in human diseases," *Cell Mol Life Sci*, vol. 58, pp. 2085–2097, 2001.

[97] T. Smith, "Diversity of wd-repeat proteins," *Subcell Biochem*, vol. 48, pp. 20–30, 2008.

[98] K. Saxena, C. Gaitatzes, M. Walsh, M. Eck, E. Neer, and T. Smith, "Analysis of the physical properties and mol-ecular modeling of sec13: a wd repeat protein involved in vesicular traffic," *Biochemistry*, vol. 35, pp. 15215–15221, 1996.

[99] B. Chang, M. Chiang, and C. Cartwright, "The inter-action of src and RACK1 is mediated by the SH2 domain of src and by phosphotyrosines in the sixth wd repeat of RACK1, and is enhanced by activation of PKC and tyrosine phosphory-lation of RACK1," *J Biol Chem*, vol. 276, pp. 20346–20356, 2001.

[100] N. Lo, C. Chong, A. Smith, W. Dobyns, R. Carrozzo, and D. Ledbetter, "Point mutations and an intragenic deletion in lis1, the lissencephaly causative gene in isolated lissencephaly sequence and miller-dieker syndrome," *Hum. Mol. Genet.*, vol. 6, pp. 157–164, 1997.

[101] V. Bhat, S. Girimaji, G. Mohan, H. Arvinda, P. Singhmar, M. Duvvari, and A. Kumar, "Mutations in wdr62, encoding a centrosomal and nuclear protein, in indian primary microcephaly families with cortical malformations," *Clin Genet*, p. ahead of print, 2011.

[102] W. Introne, R. Boissy, and W. Gahl, "Clinical, molecular, and cell biological aspects of chediak-higashi syndrome," *Mol. Genet. Metab*, vol. 68, pp. 283–303, 1999.

[103] D. Castermans, V. Wilquet, E. Parthoens, C. Huysmans, J. Steyaert, L. Swinnen, J. Fryns, W. Van de Ven, and K. Devriendt, "The neurobeachin gene is disrupted by a translocation in a patient with idiopathic autism," *J. Med. Genet*, vol. 40, pp. 352–356, 2003.

Appendix A

Primer List

Table A.1: Primers for STR genotyping.

Exon	Direction	Sequence	Tm	Length
D17S849	F	CAATTCTGTTCTAAGATTATTTTGG	64	251-261
	R	CTCTGGCTGAGGAGGC	54	
D17S926	F	GCAGTGGGCCATCATCA	54	243-260
	R	CCGCAGAAGGCTGTTGT	54	
D17S1840	F	GCCTGGGCGACAGAGTGA	60	173-225
	R	TGGGGCAGACTTGGTCCTT	60	
D17S1533	F	AGTGGCTCTAGAGTCAGACAGC	68	210-226
	R	CTGAGGCTCACCTAGGATGA	62	
D17S525	F	GCCCACTCTTAATGAGTTCCC	64	110
	R	TCTTGCAGAGGCAGGAAGTCA	64	
D17S1798	F	CAGTGAAATGCAATGTGATG	56	244-258
	R	ATGCCCAGCCTGTGTTAG	56	
D17S1854	F	TTTGGGAGGTCACAGACATTC	62	92-108
	R	CCTTGCTCTTAGGATTTGAGGA	64	
D17S938	F	CCGGATTGCTACACCTAAAT	58	164-182
	R	AACAGTCTCTNCTGGAGCAG	58	

Table A.2: Primers for candidate gene sequencing.

Exon	Direction	Sequence	Tm	Length
CRK				
1_1	F	TTTCTCCTCCAATTCTGTCTGG	Tm	491
	R	TACCAGCTACTCCGCTCCTC	Tm	
1_1A	F	CCTTGCCCTTCTCCCAAT	Tm	247
	R	GCTCCTCCGAGTCGAAGTT	Tm	
1_2	F	GCTGCTGTGAAGCTGAAACC	Tm	303
	R	CCGCTGCTGTTGATGATGTAG	Tm	
1_3	F	GAGGAGCGGAGTAGCTGGTA	Tm	543
	R	AGCTCTTTGTCCCCACGTT	Tm	
2_1	F	CAGGAGTTTGTGGTGAGAGGAT	Tm	276
	R	TCTGGAAACTGGTTCTATCAACG	Tm	
2_2	F	CGTTGATAGAACCAGTTTCCAGA	Tm	273
	R	TACCTCCAATCAGAGCCGATAC	Tm	
2_3	F	GATGATTCCAGTCCCTTACGTC	Tm	347
	R	CCATCTACAACATCCCAATGC	Tm	
3_1	F	GTGCCTAACCAGAGAATGGAGT	Tm	379
	R	AAGACTAGGAATGGAGCAGTTCA	Tm	
3_2	F	TGTAATGGCAAACGAGGTCACT	Tm	416
	R	TAACACTCCTTCCTGTCCATCG	Tm	
3_3	F	TAGTGGTCGTGCCTGTGTGTAT	Tm	460
	R	GGTACATCCTGCTTTTCACCTG	Tm	
3_4	F	CCCCAAACCTGTAAGTAGGTGA	Tm	343
	R	AGAATTTGACCCCAAAACCTTC	Tm	
3_5	F	AAGGACCAATCACCTCTGATATTC	Tm	352
	R	CCTCCCAGGTTGAAGCAA	Tm	
3_6	F	AAATTCTGCCCTCTGCATTCT	Tm	472
	R	GCATTACTCTCCCCATGAGAAA	Tm	
KIF1C				
1	F	CCGACGTTTCCCTGTTCTT	Tm	493
	R	CAACCCGGCACATTCCTC	Tm	
2	F	GGAACCACAGGTGCCTTCTA	Tm	209
	R	AGCCATATTCCAGCTTGCAG	Tm	
3	F	GTCCCTGCAGTGGTTCTGAC	Tm	282
	R	CCCAGCATTACCCTCCTCTT	Tm	
4	F	AAATTAGCTTTGGTGGCTGGAG	Tm	363
	R	TCCTCCCAATTCACCCTGACTA	Tm	

5	F	GGGTGGTAAGGACAATGATGAG	Tm	471
	R	GCTTGCAGTCAACCTCTATCTCC	Tm	
6	F	CTGAGGTTGTTTTCCCATCTG	Tm	295
	R	GTCTGATTCCAGCGACAGC	Tm	
7	F	TGCGTTGTCATTGTTTTTGC	Tm	297
	R	ATCCTGCCTCTCACCTTCC	Tm	
8	F	GGAAGGTGAGAGGCAGGAT	Tm	320
	R	TGACGGAAAAACAGGGTAGG	Tm	
9-10-11	F	ACCCGCACCTTATCTCCTG	Tm	600
	R	GATTTGGGATTCTCGGCATA	Tm	
12	F	GACTCCCCTACCTCAGACCA	Tm	259
	R	TCCTGGCACTCTTCCTATGC	Tm	
13	F	CAGGAGTTCAGAGGCGAGTTG	Tm	377
	R	AAAGGCAAGTGAGGAGGGATG	Tm	
14	F	TGGAGTTCAAGTAGAGAGACTGG	Tm	538
	R	GCGTAAGGTGGGTAGGTTTC	Tm	
15	F	GTCCCAGACCATCCTGAAAC	Tm	332
	R	TCCTCACTCACCTTCTTTGGA	Tm	
16	F	ACGGGCAGTGAAGGTAAGAC	Tm	259
	R	GACAGAAATGATGGAGGGACA	Tm	
17	F	GGATGACGATGGGAGGAGTA	Tm	466
	R	GTTGTCTCGCCCTAACCAAA	Tm	
18	F	TCTGGCACATGCTCACAAAC	Tm	250
	R	CCTCTTTATATGGCTGCACTCC	Tm	
19	F	CACCTTCCTCCAGGGTCTAAT	Tm	151
	R	TGCGACACATCTTCTACCTGA	Tm	
20	F	GCTGATGGGAAGCGAGTTTAC	Tm	316
	R	GTATCTGACAAGAGGAGGACACG	Tm	
21	F	TGTCCTCCTCTTGTCAGATACTC	Tm	302
	R	TGGATTCTTCCTGTAGTCACATC	Tm	
22_1	F	TCCCGTCACTATCTCTGTTGTC	Tm	532
	R	ACAGCTCCCGCATCTTGA	Tm	
22_2	F	AAGATGCAGGCGGTGAAG	Tm	460
	R	TGAGATGGAAGGTTCAGACTCC	Tm	
23_1	F	AGGATTCAGCTCTCCTCACATC	Tm	241
	R	AAGACGACCACGACGGAAG	Tm	
23_2	F	GAATGAAGAAGGTGGTGAGGTC	Tm	546
	R	TGGGGATAAGAGTTGTGCTTTT	Tm	
23_3	F	CCCTCACGACTGCAAGCTAC	Tm	528

	R	CCTACTCGGACCTTCTCTCTCA	Tm	
23_4	F	CTTGCTAGGAGAAGGGAAGACG	Tm	240
	R	AGCAGCGTTTGTGGGACA	Tm	
23_5	F	AGGAGCAAACCAAAGTGAAGAG	Tm	427
	R	AGACCTGTGCCATTGCCTTAT	Tm	
PAFAH1B1				
1	F	CTCTTCCTGGCGGGTCTG	Tm	481
	R	GCGGTGGAGGAGACAGAG	Tm	
2	F	GACAAAACAGTATGGTTTTTGACA	Tm	485
	R	AAGAGACCTCCCAAAGCTGTA	Tm	
3	F	TTTCTTCAGAATAGAAATGAGGTCTT	Tm	237
	R	CAAATGACAAAATTGTGCGTAA	Tm	
4	F	TCCCAAAGTGCTGGGATTAC	Tm	
	R	CTAGGTCATACTTTTCATGAAGCAC	Tm	
5	F	GGACCAAGACAAATTCACTGC	Tm	564
	R	TCAGCCTCCCGTTAAGTCTG	Tm	
6	F	AAAGGAGTGATGGAGTTGGTG	Tm	247
	R	ACTGTTAAAGCACTATCCTCTACCC	Tm	
7	F	GACTATTTCCTTGTGGGTTGTTG	Tm	428
	R	GGCTGGTCTTCAATTCCTGA	Tm	
8	F	CTGGGAAGTGTCCTGATGATT	Tm	380
	R	TGCATTAACAGCCCTGAAAC	Tm	
9	F	CCGATCGATCATCTGAGGTT	Tm	678
	R	CACCGACTGGCTCTACAGATT	Tm	
10	F	TGAACCACCTAAGAACAAGGA	Tm	590
	R	TGGCAAAGTCCACACGATTA	Tm	
11_1	F	CAGGGTCTTGCTCTGTCACC	Tm	-
	R	GATTCAACTTCTGCCTTCTTAGG	Tm	
11_10	F	AGGAGGAAGCAAGCGTGTC	Tm	381
	R	TTGTCCTTATCCCAGGCAAG	Tm	
11_11	F	CTGTCCTGTTGATGCCCTTT	Tm	351
	R	ACGCGGTGTCTAAACCAGTC	Tm	
11_12	F	TGAAAACAGACATTCCAGTGC	Tm	352
	R	GAAACAGAAAGAGGCGAACG	Tm	
11_13	F	CGTTCGCCTCTTTCTGTTTC	Tm	345
	R	CAGCCGTGACCTGTAAATCC	Tm	
11_14	F	CTCCTGGTAGAGAAGCGACAT	Tm	576
	R	ATCAGCATTTCCCCTGTGC	Tm	
11_2	F	TCCTCCCTCTTTTCCTCTGG	Tm	385

	R	ACCACCACTTTTCCCCAAAT	Tm	
11_3	F	TGCTATCTGTTGGTGCCTGA	Tm	270
	R	ACCACTAGCCTCTGCGTGAT	Tm	
11_4	F	ATCACGCAGAGGCTAGTGGT	Tm	301
	R	GCACATTGCACATCCCTAAA	Tm	
11_5	F	TCTCCCCATTGAGTGTGTCA	Tm	354
	R	CCCAGGACTAAGGACATGGA	Tm	
11_6	F	GGAGTGGGAGGAATACAGCA	Tm	394
	R	GAGGGGGACATAACGGAAGT	Tm	
11_7	F	CAGCCGAAGGAAAATCACTT	Tm	364
	R	TTTGTCTTTTGCAGGAACCA	Tm	
11_8	F	GGAAAATTGGTTCCTGCAAA	Tm	315
	R	CAGTCAGAAGAGCTAAGGGAAAA	Tm	
11_9	F	TATTGCCTCACAACCCTGCT	Tm	354
	R	CGTTTCCTCCTCCTTTGTCA	Tm	
SCARF1				
1	F	CTCAGGAAAAGCCCTGAGAG	Tm	294
	R	GTCTCCTCCTTCCCCTAACG	Tm	
2	F	CTGGGGACCCTAGCTGTAAA	Tm	332
	R	CGTCACAGGAGAAACCCTGA	Tm	
3	F	CTCCCGTCTTTCCAACCTG	Tm	353
	R	GTGCTTTGTGGATGTGGGTA	Tm	
4	F	GACGCAGGGTAAGGGTCTG	Tm	654
	R	CGCCTGCTCACTTGTGTC	Tm	
4A	F	CGGATGTTCTGAGGGTCTGT	Tm	891
	R	CCGCTCACCAGTTTCCAC	Tm	
5	F	CGTGGAGTACAGGGGTAGGA	Tm	330
	R	GTACCCCACCCTGAACAGAA	Tm	
6	F	TCTGTCACCTGCTGGATTTG	Tm	377
	R	TCAGGAAATGCTGCACAGAG	Tm	
7	F	GCCTCTCCATGTCACTCCAT	Tm	473
	R	GGTCAGGAGCTTGAGACCAG	Tm	
8	F	TAGGAGACAGGAAGGCCTGA	Tm	400
	R	GCACTGCCAACTACCTCCAT	Tm	
9	F	ACCAGTGTCCCTACCCTCCT	Tm	351
	R	CCTCATCTGCCTCTCCAGAC	Tm	
10	F	GTCACCGTGGGTCTCATGTA	Tm	467
	R	ACTAAAGAGGGGGTCGTGGT	Tm	
11_1	F	CTCCAGCCAAGCTCTTCTGT	Tm	482

	R	CCAGCATGTAGATCGTGGTC	Tm	
11_2	F	GAGCACGTGGAAGCCATT	Tm	433
	R	GTTCCTCCTGCCTTTCCTCT	Tm	
11_3	F	GGCAGTCAGACCAGAGGAAG	Tm	432
	R	GGACAGCAGAGCAAAAGTCC	Tm	
11_4	F	GAAAAGGCAAGGCTTCCAG	Tm	
	R	ACGAGGTCAGGAGATCGAGA	Tm	
11_5	F	GAGACAGAATCTCCCTCTGTCC	Tm	587
	R	CTCTCTCTTGACCCACAGACG	Tm	

Table A.3: Primers for nucleotide repeat expansion screening.

Locus	Direction	Sequence	Tm	Length
WDR81	F	AGGTTGCAATGAACTGAGATTGT	Tm	209
	R	GTGGGTCTTCTGAAACTCCTTAAC	Tm	
MNT	F	CACCTAAGCTGAGCCATCGT	Tm	425
	R	CACAGTGATGTGCCCAATG	Tm	
LIS1	F	ctcttcctggcgggtctg	Tm	405
	R	CTCTTCCGTCCTTCACTCC	Tm	
HIC1	F	CCTGGTGCTGGACTTCATCT	Tm	283
	R	GTAGCAGGCCTGGATGACC	Tm	
TRPV1	F	TGGAACAGTAATCTCCCTAACTTC	Tm	203
	R	TGTAGAGGTTACAGTGAGCTGAGA	Tm	
RUTBC1	F	ctccggctgtgggaag	Tm	308
	R	ggggttcttgactcgacttt	Tm	
GARNL4	F	ctcaggacccgctgtacc	Tm	405
	R	agcgaagggcacctcagt	Tm	

Table A.4: Primers for mutation screening in cohorts.

Exon	Direction	Sequence	Tm	Length
WDR81				
1_1	F	AGCCACGTGCGCTTGTTT	Tm	570
	R	GTCCTGACTTCCGCTTCTCC	Tm	
1_2	F	GGGAAGGCGCTCTCAGAA	Tm	391
	R	GAAGCGAGTGAGGGTCTCAG	Tm	
1_3	F	CCACGTGGTGGCCCTAGT	Tm	590
	R	TGAAGGTGACCACGTCATGT	Tm	

1_4	F	GTGAAACTACCCAATGCCCT	Tm	658
	R	TGCTTATCCCCCTTGTTGAG	Tm	
1_5	F	CGCATCAGCAACTTCCACTA	Tm	582
	R	AAACGTGAGGTCGATCCA	Tm	
1_6	F	TCCCTCTATCTTCCGCTC	Tm	856
	R	AAACACCATCTCTGCCAA	Tm	
1_7	F	TGCCTACTCCACAGGGACA	Tm	429
	R	CTGCCACAGAGCAAACACC	Tm	
1_8	F	TGCACAGATTCATCCTCCTG	Tm	512
	R	AATCTCCTCCCCAAAAGCAC	Tm	
1_9	F	TGCAGGCATTTCTCACTCAC	Tm	769
	R	CTCCCAAAGTGCTGGGAATA	Tm	
2-3	F	GAGCCCCTGTTATCTGAGGA	Tm	597
	R	GGGCAGGTACTGGTAGGTGA	Tm	
3-4	F	CCGGCAACATCTACCAGAAG	Tm	559
	R	TCTTGACACCCAGTCCAGTG	Tm	
5-6	F	GCCTGGATGACAGAGCAA	Tm	574
	R	AACACCTTCTGCAGCTCGTC	Tm	
6	F	TGCCACAGGTGGTCTTCTCT	Tm	511
	R	CAAGGCTGTCTGTCCTCTCC	Tm	
7_1	F	CACAACGGAATGAATCCACA	Tm	597
	R	AAGTGAAAGTGGGCATCCTG	Tm	
7_2	F	GCATTCAGATCCCCAATGAC	Tm	580
	R	AAACTCCGCAGCATTACAGC	Tm	
8	F	CGGTGAGTTGGGGGATTAG	Tm	519
	R	CCACCTTCCCATCTCACACT	Tm	
9	F	AAGCCAGGATGTTGTTCTGG	Tm	564
	R	CTCCCTGCCCAGTTAGTCCT	Tm	
10_1	F	ACTGGGTTACGGATGCTGAG	Tm	328
	R	GCCGTACAGGTCAAAGGTGT	Tm	
10_2	F	AGCCCACCCATCACTACAAG	Tm	256
	R	GCCTCAGTCTATGCCAGGAG	Tm	
ZNF594				
1_1	F	CGCTAGTGAAAGCCACATCTC	Tm	583
	R	GGTTTGAACTCCTGTTGAAGG	Tm	
1_2	F	TGCAGGAGAAAGCTCCCATA	Tm	578
	R	GATGTGGGACAAGGTGTGAA	Tm	
1_3	F	CACAGTAGGGGGAAGCCATA	Tm	347

	R	TCTCCAGCATGCAATCTCTG	Tm	
1_4	F	TCAGGCAGCATTCTCACCTT	Tm	676
	R	GTGCGCCAAATGAAGAGC	Tm	
1_5	F	GCCAGAGGTCACACCTTGTT	Tm	204
	R	GCAGGCTCTGATGTTTGAGG	Tm	
1_6	F	GGTGTGACCTCTGGCTGAA	Tm	426
	R	TCATTTGGCGCACAGCTT	Tm	
1_7	F	GTGGGAAATCTTTTAGGGGT	Tm	583
	R	CCCCACATTTATTGCATACG	Tm	

Table A.5: Primers for exon skipping RT-PCR.

Exon	Direction	Primer sequence	Tm	Size
Exon 1	F	ATGGCCCAGGGCAGCGGG	64	1098
	R	CAGCTGCATGAGGTAGTGGA	62	
Exon 1-2	F	GTGAAACTACCCAATGCCCT	60	3115
	R	CAGCGGACCATCTTGCAGGC	66	
Exon 1-2	F	CTGCAGTGCCTACTCCACAG	64	1406
	R	CAGCGGACCATCTTGCAGGC	66	
Exon 1-3	F	GGGTGTCCTATTGGCAGAGA	62	1630
	R	GGGCAGGTACTGGTAGGTGA	64	
Exon 1-4	F	GGGTGTCCTATTGGCAGAGA	62	1774
	R	CAGGATGTCCATGAGTGTGG	62	
Exon 1-5	F	GGGTGTCCTATTGGCAGAGA	62	1892
	R	TGAGGCTGATGGTTTTCACA	58	
Exon 1-6	F	GCAAGCTGGACCAACTGTTT	60	1961
	R	GGAGAAGGGCACGTAGATTG	62	
Exon 1-7	F	GGGTGTCCTATTGGCAGAGA	62	2412
	R	TTCTCAGGCCGAGAGTCATT	60	
Exon 1-8	F	GGGTGTCCTATTGGCAGAGA	62	2976
	R	AAGCGCAGGGTAGAGTCAGA	62	
Exon 1-9	F	GGACACACTCCTGCAGATGA	62	2966
	R	CCATGAAGCCTGAGGAGAAG	62	
Exon 1-10	F	GGGTGTCCTATTGGCAGAGA	62	3316
	R	GCCGTACAGGTCAAAGGTGT	62	

Table A.6: Primers for quantitative RT-PCR.

Exon	Direction	Primer sequence	Tm	Size
WDR81	F	GCAAGCTGGACCAACTGTTT	60.3	115

	R	GGGAAGTAGGGTGGGAAGG	60.7	
GAPDH	F	AGGTGAAGGTCGGAGTCAAC	59.2	100
	R	GGGTCATTGATGGCAACA	58.8	

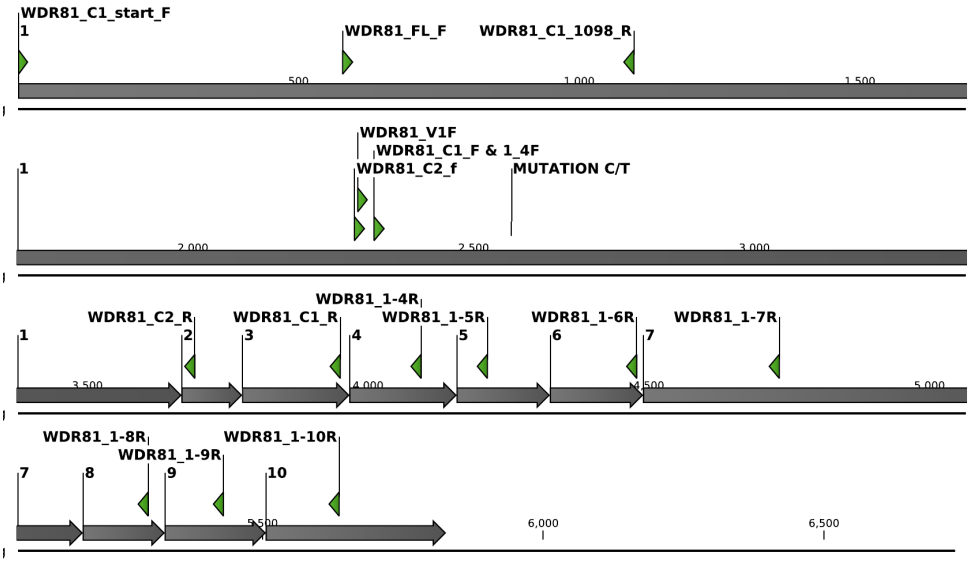


Figure A.1: Locations of the primers for exon skipping RT-PCR.

Appendix B

Parametric linkage results

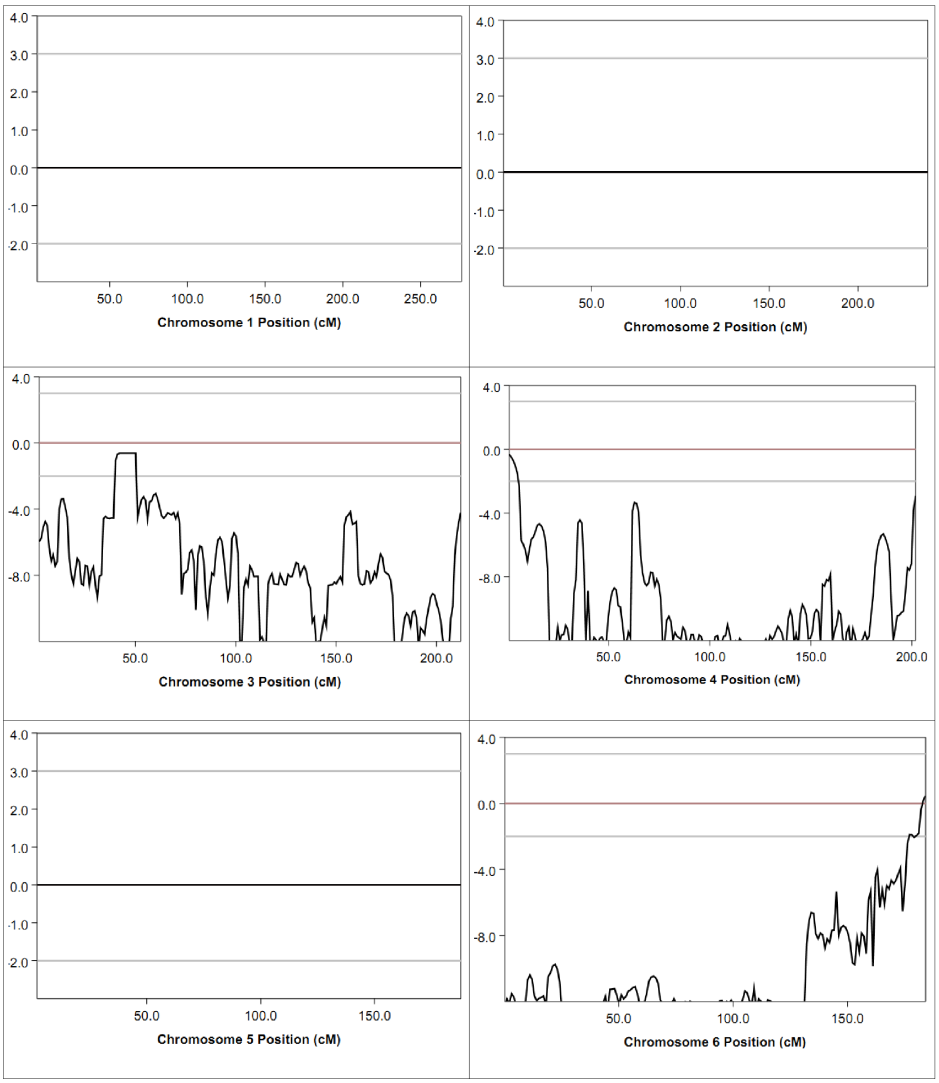


Figure B.1: Linkage results-1. Chromosome 1-6.

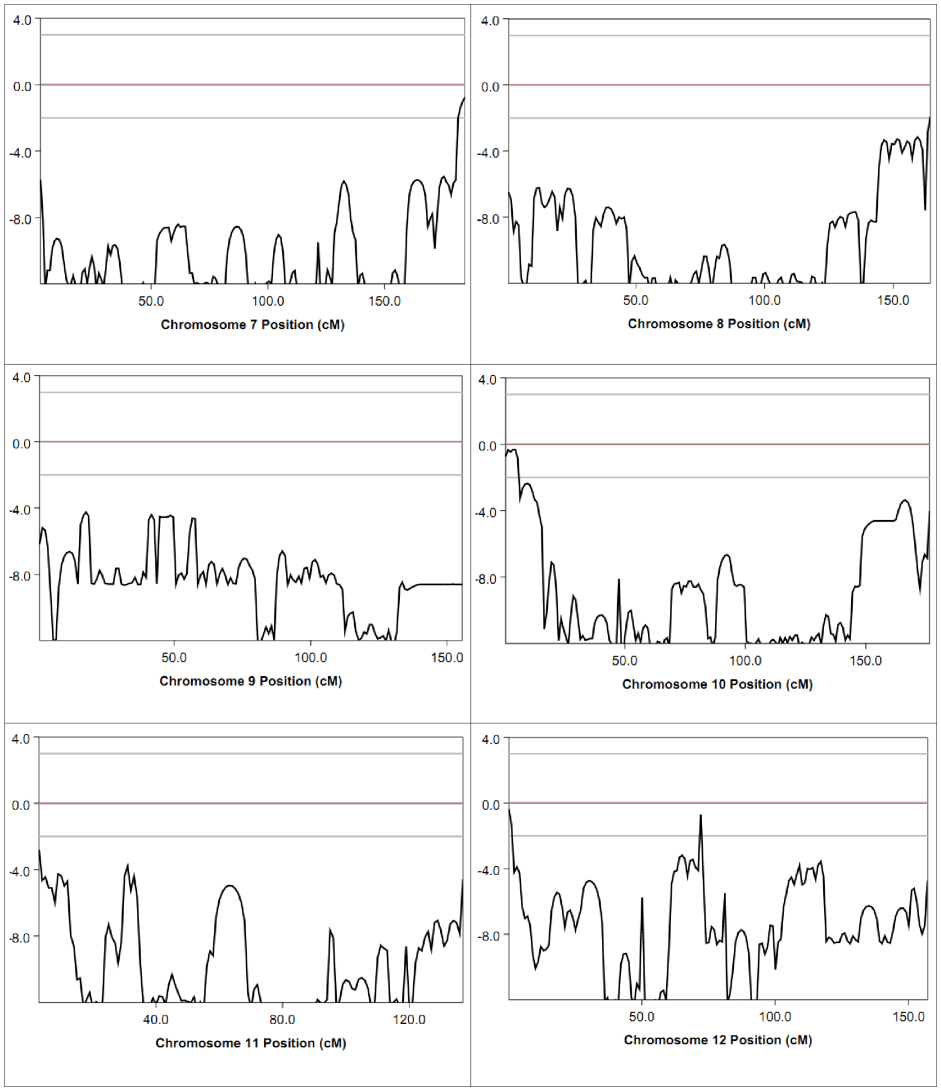


Figure B.2: Linkage results-2. Chromosome 7-12.

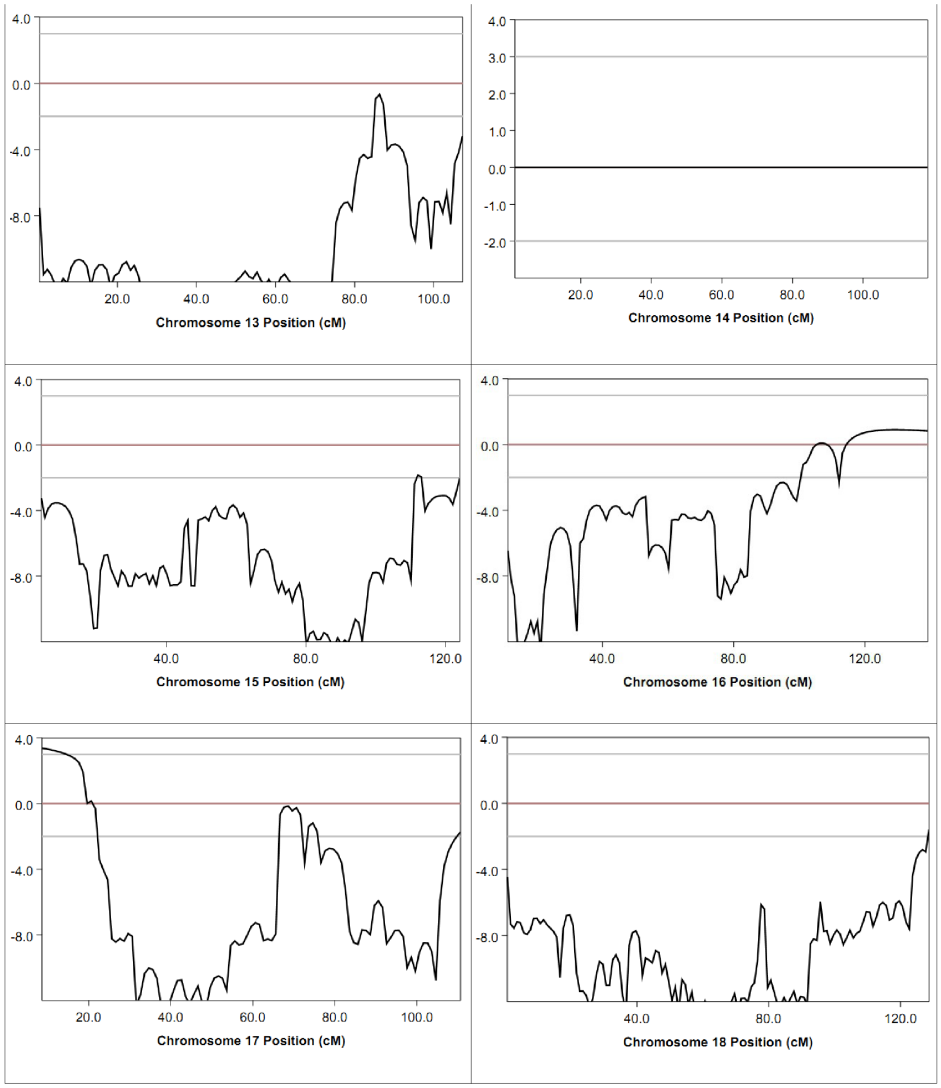


Figure B.3: Linkage results-3. Chromosome 13-18.

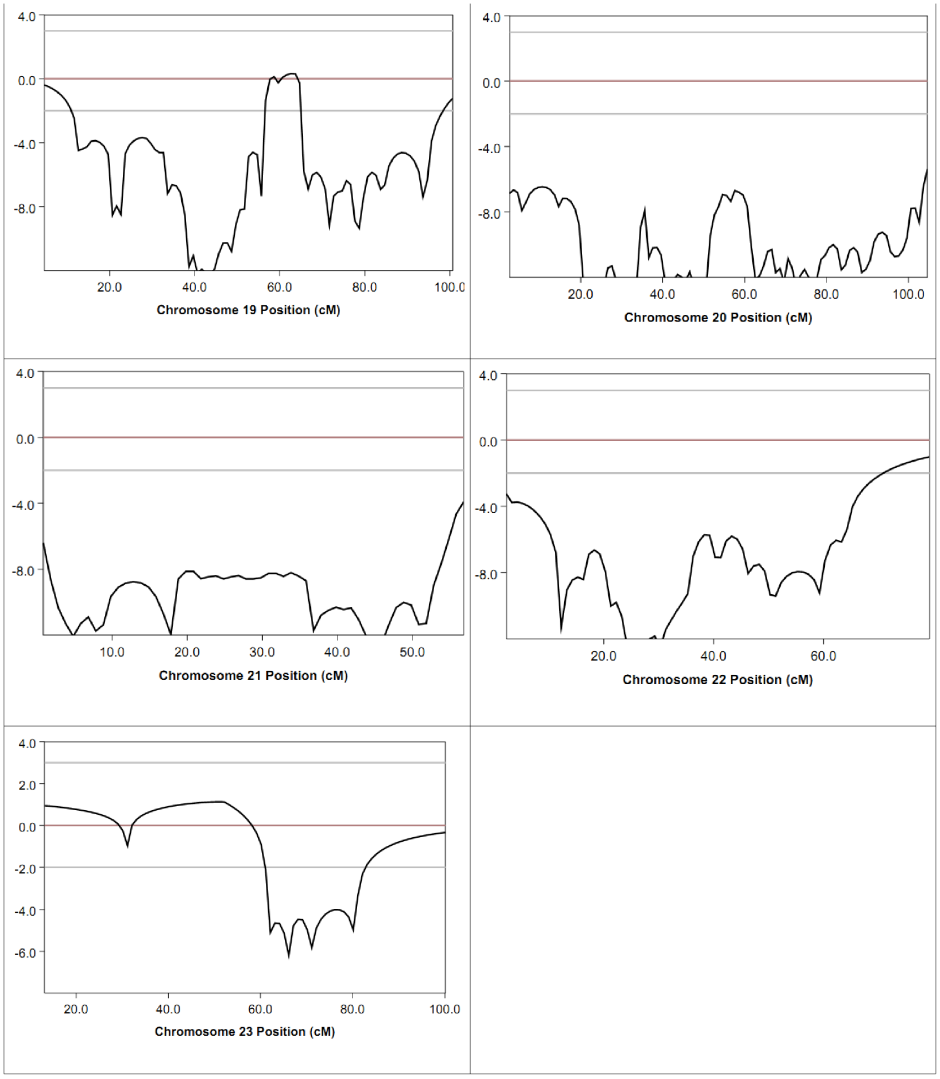


Figure B.4: Linkage results-4. Chromosome 19-23.

Appendix C

Genes located at the critical region

Table C.1: Genes located at the critical region.

Chr	Start	End	Biotype	Status	Name	Transcripts
17	62293	202888	protein_coding	KNOWN	*RPH3AL*	3
17	254332	263814	processed_transcript	PUTATIVE	*AC108004.3*	1
17	260118	264367	protein_coding	KNOWN	*C17orf97*	2
17	289769	295730	protein_coding	KNOWN	*FAM101B*	1
17	411908	618096	protein_coding	KNOWN	*VPS53*	6
17	635847	646074	protein_coding	KNOWN	*FAM57A*	3
17	647661	655501	protein_coding	KNOWN	*GEMIN4*	2
17	655900	658576	processed_transcript	NOVEL	*AC087392.3*	2
17	662550	686505	protein_coding	KNOWN	*GLOD4*	4
17	685513	695749	protein_coding	KNOWN	*RNMTL1*	1
17	702581	883010	protein_coding	KNOWN	*NXN*	3
17	900357	905388	protein_coding	KNOWN	*TIMM22*	1
17	906758	1090616	protein_coding	KNOWN	*ABR*	6
17	1085187	1088380	processed_transcript	PUTATIVE	*AC016292.1*	1
17	1173853	1174754	protein_coding	KNOWN	*BHLHA9*	1
17	1182957	1204281	protein_coding	KNOWN	*TUSC5*	1
17	1247566	1303505	protein_coding	KNOWN	*YWHAE*	8
17	1323983	1359552	protein_coding	KNOWN	*CRK*	2
17	1367480	1395995	protein_coding	KNOWN	*MYO1C*	6
17	1397872	1420182	protein_coding	KNOWN	*INPP5K*	16
17	1420225	1421389	processed_transcript	NOVEL	*AC100748.2*	1
17	1421287	1466110	protein_coding	KNOWN	*PITPNA*	3

17	1477666	1532180	protein_coding	KNOWN	*SLC43A2*	3
17	1519486	1519589	snoRNA	NOVEL	*snoU13*	1
17	1537152	1549042	protein_coding	KNOWN	*SCARF1*	3
17	1549444	1553352	protein_coding	KNOWN	*RILP*	1
17	1553923	1588154	protein_coding	KNOWN	*PRPF8*	2
17	1611064	1613651	protein_coding	KNOWN	*TLCD2*	1
17	1614805	1619504	lincRNA	KNOWN	*C17orf91*	1
17	1617191	1617285	miRNA	KNOWN	*MIR22*	1
17	1619817	1641893	protein_coding	KNOWN	*WDR81*	15
17	1646130	1658560	protein_coding	KNOWN	*SERPINF2*	6
17	1665253	1680859	protein_coding	KNOWN	*SERPINF1*	1
17	1682829	1733170	protein_coding	KNOWN	*SMYD4*	4
17	1733266	1802848	protein_coding	KNOWN	*RPA1*	1
17	1837978	1928178	protein_coding	KNOWN	*RTN4RL1*	1
17	1921177	1924000	processed_transcript	NOVEL	*AC099684.1*	1
17	1933431	1946724	protein_coding	KNOWN	*DPH1*	1
17	1945277	1946479	protein_coding	KNOWN	*OVCA2*	1
17	1953202	1953302	miRNA	KNOWN	*MIR132*	1
17	1953565	1953674	miRNA	KNOWN	*MIR212*	1
17	1958393	1962980	protein_coding	KNOWN	*HIC1*	2
17	1963133	2207069	protein_coding	KNOWN	*SMG6*	4
17	2116991	2117922	processed_transcript	PUTATIVE	*AC090617.1*	1
17	2118776	2119309	processed_transcript	PUTATIVE	*AC130689.5*	1
17	2135974	2136904	processed_transcript	NOVEL	*AL450226.2*	1
17	2207248	2228554	protein_coding	KNOWN	*SRR*	1
17	2225992	2240678	protein_coding	KNOWN	*TSR1*	2
17	2232413	2232507	snoRNA	KNOWN	*SNORD91B*	1
17	2233570	2233664	snoRNA	KNOWN	*SNORD91A*	1
17	2240806	2284344	protein_coding	KNOWN	*SGSM2*	2
17	2287354	2304412	protein_coding	KNOWN	*MNT*	2
17	2310279	2318731	protein_coding	KNOWN	*AC006435.1*	1
17	2319352	2415200	protein_coding	KNOWN	*METTL16*	4
17	2496923	2588909	protein_coding	KNOWN	*PAFAH1B1*	3
17	2558973	2559076	snoRNA	NOVEL	*snoU13*	1
17	2574352	2577029	lincRNA	KNOWN	*AC005696.2*	1
17	2592680	2614927	protein_coding	KNOWN	*KIAA0664*	3
17	2651372	2651476	miRNA	KNOWN	*MIR1253*	1
17	2699732	2941034	protein_coding	KNOWN	*RAP1GAP2*	4
17	2995352	2996290	protein_coding	KNOWN	*OR1D2*	1
17	3029878	3030875	protein_coding	KNOWN	*OR1G1*	1

17	3100813	3101742	protein_coding	KNOWN	*OR1A2*	1
17	3118915	3119844	protein_coding	KNOWN	*OR1A1*	1
17	3132283	3132372	miRNA	NOVEL	*AC090282.1*	1
17	3143970	3144908	processed_transcript	KNOWN	*OR1D4*	2
17	3181193	3182268	protein_coding	KNOWN	*OR3A2*	1
17	3194929	3195876	protein_coding	KNOWN	*OR3A1*	1
17	3300398	3301704	protein_coding	KNOWN	*OR1E1*	1
17	3323862	3324827	protein_coding	KNOWN	*OR3A3*	1
17	3336164	3337135	protein_coding	KNOWN	*OR1E2*	2
17	3343308	3417146	protein_coding	KNOWN	*SPATA22*	4
17	3377404	3406713	protein_coding	KNOWN	*ASPA*	2
17	3413796	3461289	protein_coding	KNOWN	*TRPV3*	3
17	3468743	3500392	protein_coding	KNOWN	*TRPV1*	5
17	3511893	3539616	protein_coding	KNOWN	*SHPK*	1
17	3539762	3564836	protein_coding	KNOWN	*CTNS*	9
17	3566196	3571976	protein_coding	KNOWN	*TAX1BP3*	1
17	3566357	3599488	processed_transcript	NOVEL	*RP11-48B14.2*	2
17	3572090	3572962	protein_coding	KNOWN	*TMEM93*	2
17	3575748	3599698	protein_coding	KNOWN	*P2RX5*	12
17	3617920	3704537	protein_coding	KNOWN	*ITGAE*	1
17	3627211	3630067	protein_coding	KNOWN	*GSG2*	1
17	3714460	3749540	protein_coding	KNOWN	*C17orf85*	2
17	3763609	3796338	protein_coding	KNOWN	*CAMKK1*	4
17	3799891	3819960	protein_coding	KNOWN	*P2RX1*	1
17	3822869	3867736	protein_coding	KNOWN	*ATP2A3*	8
17	3907739	4046314	protein_coding	KNOWN	*ZZEF1*	1
17	3960050	3960170	rRNA	NOVEL	*5S_rRNA*	1
17	4017902	4018004	snoRNA	NOVEL	*snoU13*	1
17	4046462	4060989	protein_coding	KNOWN	*CYB5D2*	1
17	4066665	4167274	protein_coding	KNOWN	*ANKFY1*	3
17	4089621	4089719	misc_RNA	NOVEL	*Y_RNA*	1
17	4096063	4096160	misc_RNA	NOVEL	*Y_RNA*	1
17	4172512	4269969	protein_coding	KNOWN	*UBE2G1*	1
17	4337219	4391498	protein_coding	KNOWN	*SPNS3*	2
17	4400689	4402969	processed_transcript	NOVEL	*AC118754.4*	1
17	4402129	4443228	protein_coding	KNOWN	*SPNS2*	1
17	4442192	4458681	protein_coding	KNOWN	*MYBBP1A*	3
17	4460224	4463876	protein_coding	KNOWN	*GGT6*	3
17	4487294	4511614	protein_coding	KNOWN	*SMTNL2*	2
17	4523044	4523147	snRNA	NOVEL	*U6*	1

17	4534214	4545589	protein_coding	KNOWN	*ALOX15*	2
17	4574680	4607632	protein_coding	KNOWN	*AC091153.1*	3
17	4578674	4600129	processed_transcript	NOVEL	*AC091153.4*	1
17	4607525	4608824	processed_transcript	NOVEL	*RP11-314A20.2*	1
17	4613784	4624795	protein_coding	KNOWN	*ARRB2*	4
17	4634723	4636888	protein_coding	KNOWN	*MED11*	1
17	4636828	4643223	protein_coding	KNOWN	*CXCL16*	1
17	4643319	4649411	protein_coding	KNOWN	*ZMYND15*	2
17	4675187	4686508	protein_coding	KNOWN	*TM4SF5*	1
17	4688581	4689729	protein_coding	KNOWN	*VMO1*	4
17	4692307	4693591	protein_coding	KNOWN	*GLTPD2*	1
17	4699439	4701798	protein_coding	KNOWN	*PSMB6*	1
17	4710391	4726727	protein_coding	KNOWN	*PLD2*	1
17	4736635	4801355	protein_coding	KNOWN	*MINK1*	4
17	4801069	4806369	protein_coding	KNOWN	*CHRNE*	1
17	4802713	4806227	protein_coding	KNOWN	*C17orf107*	2
17	4835592	4838325	protein_coding	KNOWN	*GP1BA*	2
17	4840425	4843463	protein_coding	KNOWN	*SLC25A11*	2
17	4843423	4848518	protein_coding	KNOWN	*RNF167*	1
17	4848947	4852309	protein_coding	KNOWN	*PFN1*	1
17	4851387	4860426	protein_coding	KNOWN	*ENO3*	17
17	4862523	4871132	protein_coding	KNOWN	*SPAG7*	1
17	4871287	4890960	protein_coding	KNOWN	*CAMTA2*	5
17	4876146	4877976	processed_transcript	NOVEL	*AC004771.2*	1
17	4891426	4900905	protein_coding	KNOWN	*INCA1*	2
17	4901243	4931696	protein_coding	KNOWN	*KIF1C*	1
17	4922509	4923388	processed_transcript	PUTATIVE	*AC109333.10*	1
17	4935895	4955304	protein_coding	KNOWN	*GPR172B*	3
17	4981754	4999668	protein_coding	KNOWN	*ZFP3*	1
17	5008930	5026397	protein_coding	KNOWN	*ZNF232*	2
17	5015227	5017672	processed_transcript	NOVEL	*AC012146.7*	1
17	5019733	5078329	protein_coding	KNOWN	*USP6*	5
17	5082830	5095178	protein_coding	KNOWN	*ZNF594*	3
17	5113705	5138137	protein_coding	KNOWN	*C17orf87*	2
17	5185558	5289129	protein_coding	KNOWN	*RABEP1*	6
17	5289346	5323000	protein_coding	KNOWN	*NUP88*	2
17	5322961	5336340	protein_coding	KNOWN	*RPAIN*	7
17	5328469	5336196	protein_coding	KNOWN	*AC004148.1*	1
17	5336099	5342471	protein_coding	KNOWN	*C1QBP*	1
17	5344232	5372236	protein_coding	KNOWN	*DHX33*	2

17	5374571	5389537	protein_coding	KNOWN	*DERL2*	1
17	5390247	5394134	protein_coding	KNOWN	*MIS12*	1
17	5402748	5487832	protein_coding	KNOWN	*NLRP1*	6
17	5402759	5404465	processed_transcript	KNOWN	*AC055839.1*	1
17	5417529	5417590	snRNA	KNOWN	*RNU7-31P*	1
17	5973934	6027745	protein_coding	KNOWN	*WSCD1*	2
17	6110483	6110589	snRNA	NOVEL	*U6*	1
17	6327057	6338519	protein_coding	KNOWN	*AIPL1*	4
17	6347761	6354382	protein_coding	KNOWN	*FAM64A*	2
17	6354583	6459877	protein_coding	KNOWN	*PITPNM3*	2
17	6481646	6544247	protein_coding	KNOWN	*KIAA0753*	3
17	6504264	6504371	rRNA	NOVEL	*5S_rRNA*	1
17	6544078	6547920	protein_coding	KNOWN	*TXNDC17*	1
17	6546635	6554954	protein_coding	KNOWN	*MED31*	1
17	6554971	6556593	protein_coding	KNOWN	*C17orf100*	1
17	6588032	6616740	protein_coding	KNOWN	*SLC13A5*	3
17	6659156	6678962	protein_coding	KNOWN	*XAF1*	5
17	6679551	6690965	protein_coding	KNOWN	*FBXO39*	1
17	6701312	6735080	protein_coding	KNOWN	*TEKT1*	2
17	6888442	6915653	protein_coding	NOVEL	*AC027763.2*	2
17	6899384	6914055	protein_coding	KNOWN	*ALOX12*	3
17	6915736	6918199	protein_coding	KNOWN	*RNASEK*	8
17	6915954	6920839	protein_coding	NOVEL	*C17orf49*	5
17	6917814	6920844	protein_coding	KNOWN	*AC040977.1*	9
17	6919137	6922949	processed_transcript	NOVEL	*MIR497HG*	1
17	6920934	6921020	miRNA	KNOWN	*MIR195*	1
17	6921230	6921341	miRNA	KNOWN	*MIR497*	1
17	6926369	6933219	protein_coding	KNOWN	*BCL6B*	2
17	6939477	6943438	protein_coding	KNOWN	*SLC16A13*	1
17	6944949	6947242	protein_coding	KNOWN	*SLC16A11*	2
17	6977857	6983600	protein_coding	KNOWN	*CLEC10A*	2
17	7004641	7018292	protein_coding	KNOWN	*ASGR2*	5
17	7076750	7082883	protein_coding	KNOWN	*ASGR1*	2
17	7093209	7123369	protein_coding	KNOWN	*DLG4*	14
17	7120444	7128587	protein_coding	KNOWN	*ACADVL*	5
17	7126616	7126698	miRNA	KNOWN	*MIR324*	1
17	7128660	7137868	protein_coding	KNOWN	*DVL2*	1
17	7138350	7142825	protein_coding	KNOWN	*PHF23*	3
17	7143333	7145772	protein_coding	KNOWN	*GABARAP*	1
17	7146910	7155280	protein_coding	KNOWN	*CTDNEP1*	1

17	7155372	7163250	protein_coding	KNOWN	*C17orf81*	4
17	7163223	7166512	protein_coding	KNOWN	*CLDN7*	3
17	7169374	7169485	misc_RNA	NOVEL	*Y_RNA*	1
17	7185053	7191576	protein_coding	KNOWN	*SLC2A4*	2
17	7191571	7,197876	protein_coding	KNOWN	*YBX2*	1
17	7210318	7215782	protein_coding	KNOWN	*EIF5A*	5
17	7214643	7222394	protein_coding	KNOWN	*GPS2*	4
17	7218951	7232638	protein_coding	KNOWN	*NEURL4*	2
17	7239848	7254797	protein_coding	KNOWN	*ACAP1*	1
17	7255208	7258258	protein_coding	KNOWN	*KCTD11*	1

Appendix D

Missense variants

Table D.1: Missense variants identified in the next generation sequence data of 05-985.

Chr	Pos	Ref	Var	Reads	Freq	Ref aa	Var aa	Gene
17	169,221	T	-	28	25%	D	X	*RPH3AL*
17	177,276	T	C	39	5%	Q	R	*RPH3AL*
17	263,359	-	G	34	6%	S	G	*LOC400566*
17	263,463	T	C	31	6%	R	R	*LOC400566*
17	293,199	A	G	42	5%	L	P	*FAM101B*
17	371,634	-	C	43	5%	K	R	*VPS53*
17	401,917	A	G	47	4%	L	P	*VPS53*
17	477,178	-	G	47	4%	F	F	*VPS53*
17	505,868	T	C	55	4%	N	D	*VPS53*
17	505,883	A	G	58	5%	S	P	*VPS53*
17	591,486	A	G	69	3%	I	V	*FAM57A*
17	591,532	A	GT	60	3%	D	G	*FAM57A*
17	594,990	-	G	47	4%	L	L	*GEMIN4*
17	595,007	A	G	46	4%	L	P	*GEMIN4*
17	595,037	T	A	48	4%	E	V	*GEMIN4*
17	595,214	-	G	43	7%	V	A	*GEMIN4*
17	595,278	A	G	50	4%	F	L	*GEMIN4*
17	595,401	-	C	55	4%	R	G	*GEMIN4*
17	595,764	A	GAC	44	5%	LS	LV	*GEMIN4*
17	596,161	A	G	48	6%	F	F	*GEMIN4*
17	596,453	T	C	49	4%	N	S	*GEMIN4*

17	596,459	TCT	ATCC	49	4%	ED	ED	*GEMIN4*
17	596,626	-	C	41	7%	R	R	*GEMIN4*
17	596,655	-	G	47	4%	T	P	*GEMIN4*
17	596,691	A	G	45	4%	F	L	*GEMIN4*
17	597,213	A	G	53	4%	S	P	*GEMIN4*
17	597,260	-	G	50	6%	Q	P	*GEMIN4*
17	597,487	A	G	61	3%	F	F	*GEMIN4*
17	597,601	A	GT	70	3%	F	L	*GEMIN4*
17	597,698	A	G	67	4%	L	P	*GEMIN4*
17	597,941	A	T	76	3%	L	Q	*GEMIN4*
17	597,974	A	G	80	2%	L	S	*GEMIN4*
17	621,404	A	G	37	5%	L	L	*GLOD4*
17	626,894	A	CG	61	3%	Y	Y	*GLOD4*
17	626,974	A	G	67	3%	Y	H	*GLOD4*
17	632,211	T	C	36	6%	K	R	*GLOD4*
17	632,215	A	G	37	5%	F	L	*GLOD4*
17	641,537	C	T	66	3%	A	A	*RNMTL1*
17	641,605	A	G	59	3%	E	G	*RNMTL1*
17	641,606	A	G	59	3%	E	E	*RNMTL1*
17	669,521	A	TG	36	6%	F	S	*NXN*
17	675,971	A	GT	62	3%	V	D	*NXN*
17	675,983	T	C	61	5%	N	S	*NXN*
17	848,775	T	C	47	4%	V	A	*TIMM22*
17	848,845	T	C	40	5%	P	P	*TIMM22*
17	859,673	T	C	18	11%	K	E	*ABR*
17	861,847	-	G	28	7%	V	A	*ABR*
17	862,678	T	AC	44	5%	K	S	*ABR*
17	863,153	A	G	51	4%	L	P	*ABR*
17	900,116	T	C	37	5%	K	R	*ABR*
17	900,587	A	-	37	24%	Y	X	*ABR*
17	907,029	A	G	49	4%	L	P	*ABR*
17	907,032	T	C	46	4%	K	R	*ABR*
17	907,080	A	G	39	5%	F	S	*ABR*
17	923,622	-	G	36	6%	V	A	*ABR*
17	1,215,084	T	C	41	5%	K	R	*YWHAE*
17	1,286,752	-	G	41	7%	L	P	*CRK*
17	1,318,075	C	T	23	96%	E	E	*MYO1C*
17	1,318,325	A	G	36	6%	L	P	*MYO1C*
17	1,319,627	T	C	43	5%	K	R	*MYO1C*
17	1,320,518	T	C	12	17%	K	R	*MYO1C*

17	1,320,704	A	G	20	10%	F	L	*MYO1C*
17	1,321,351	-	A	41	5%	P	L	*MYO1C*
17	1,322,194	T	CA	48	4%	CI	CC	*MYO1C*
17	1,322,235	A	G	48	4%	V	A	*MYO1C*
17	1,322,259	A	G	53	4%	F	S	*MYO1C*
17	1,328,736	T	C	45	4%	D	G	*MYO1C*
17	1,332,571	-	C	15	20%	D	G	*MYO1C*
17	1,346,171	T	C	36	8%	D	G	*SKIP*
17	1,346,191	-	A	37	8%	S	S	*SKIP*
17	1,359,381	-	T	44	5%	V	D	*SKIP*
17	1,371,684	T	C	35	6%	R	R	*PITPNA*
17	1,384,244	T	C	52	4%	N	D	*PITPNA*
17	1,385,263	C	-	24	37%	W	X	*PITPNA*
17	1,435,993	T	C	35	6%	D	G	*SLC43A2*
17	1,466,746	G	A	41	7%	N	N	*SLC43A2*
17	1,494,976	A	G	32	6%	I	T	*SCARF1*
17	1,496,653	A	G	36	6%	W	R	*RILP*
17	1,501,166	T	C	30	7%	N	S	*PRPF8*
17	1,503,612	T	C	57	3%	I	V	*PRPF8*
17	1,503,890	-	C	63	3%	E	G	*PRPF8*
17	1,504,009	A	G	72	3%	G	G	*PRPF8*
17	1,508,341	T	C	54	4%	I	V	*PRPF8*
17	1,508,595	T	C	53	4%	N	S	*PRPF8*
17	1,510,779	A	G	58	3%	F	S	*PRPF8*
17	1,510,833	T	C	66	3%	K	R	*PRPF8*
17	1,510,847	T	C	68	3%	E	E	*PRPF8*
17	1,511,085	T	C	73	3%	H	R	*PRPF8*
17	1,511,100	T	C	74	3%	E	G	*PRPF8*
17	1,511,426	-	C	66	3%	E	E	*PRPF8*
17	1,511,445	A	G	66	3%	L	P	*PRPF8*
17	1,511,783	G	C	72	4%	S	R	*PRPF8*
17	1,523,533	C	T	59	3%	V	V	*PRPF8*
17	1,525,836	-	G	63	3%	D	D	*PRPF8*
17	1,525,841	T	CA	64	3%	FM	FC	*PRPF8*
17	1,526,294	T	C	46	4%	K	E	*PRPF8*
17	1,526,560	T	C	41	5%	N	D	*PRPF8*
17	1,527,177	T	C	48	4%	Q	R	*PRPF8*
17	1,530,790	-	G	68	6%	S	P	*PRPF8*
17	1,530,845	T	C	72	3%	K	K	*PRPF8*
17	1,531,687	-	G	38	5%	M	T	*PRPF8*

17	1,531,732	T	C	39	5%	Y	C	*PRPF8*
17	1,531,909	A	G	52	4%	V	A	*PRPF8*
17	1,532,013	AG	-	47	21%	P	X	*PRPF8*
17	1,532,250	-	T	64	3%	L	L	*PRPF8*
17	1,532,270	T	C	63	5%	I	V	*PRPF8*
17	1,534,531	T	C	14	14%	K	E	*PRPF8*
17	1,563,788	T	C	46	4%	T	A	*C17orf91*
17	1,578,176	C	G	22	9%	A	G	*WDR81*
17	1,583,983	T	C	31	6%	F	F	*WDR81*
17	1,584,040	C	TA	31	6%	CV	CS	*WDR81*
17	1,587,473	T	C	50	4%	L	P	*WDR81*
17	1,587,556	G	A	47	4%	G	S	*WDR81*
17	1,595,794	T	AC	24	8%	C	T	*SERPINF2*
17	1,597,465	GA	-	25	8%	VK	XR	*SERPINF2*
17	1,621,123	T	C	32	6%	L	L	*SERPINF1*
17	1,625,160	T	C	31	6%	D	D	*SERPINF1*
17	1,625,226	T	CA	28	7%	DS	DI	*SERPINF1*
17	1,626,679	T	C	51	4%	F	S	*SERPINF1*
17	1,626,694	A	G	55	4%	D	G	*SERPINF1*
17	1,627,243	-	C	41	5%	L	S	*SERPINF1*
17	1,627,453	-	C	48	4%	L	P	*SERPINF1*
17	1,627,456	T	C	48	6%	F	S	*SERPINF1*
17	1,636,877	T	C	47	4%	N	D	*SMYD4*
17	1,637,610	G	A	29	17%	R	C	*SMYD4*
17	1,649,970	C	-	50	12%	V	X	*SMYD4*
17	1,650,001	A	G	53	4%	P	P	*SMYD4*
17	1,650,042	T	C	55	4%	R	G	*SMYD4*
17	1,650,073	T	A	53	4%	E	D	*SMYD4*
17	1,650,104	A	G	51	4%	L	P	*SMYD4*
17	1,650,131	C	AT	42	5%	S	N	*SMYD4*
17	1,650,155	A	G	41	5%	L	P	*SMYD4*
17	1,650,230	T	C	35	6%	N	S	*SMYD4*
17	1,650,246	T	C	34	6%	S	G	*SMYD4*
17	1,650,303	T	C	37	5%	I	V	*SMYD4*
17	1,650,446	T	C	51	4%	H	R	*SMYD4*
17	1,650,563	-	T	49	4%	G	E	*SMYD4*
17	1,650,588	T	C	48	4%	K	E	*SMYD4*
17	1,650,805	A	G	52	6%	A	A	*SMYD4*
17	1,650,837	T	C	55	4%	K	E	*SMYD4*
17	1,650,982	A	GT	38	5%	C	*	*SMYD4*

17	1,654,753	A	G	65	3%	S	P	*SMYD4*
17	1,662,030	T	CA	22	9%	A	A	*SMYD4*
17	1,722,489	C	AG	32	6%	P	S	*RPA1*
17	1,729,087	T	CA	86	2%	PL	PT	*RPA1*
17	1,729,098	T	C	85	2%	V	A	*RPA1*
17	1,729,271	T	C	80	4%	S	P	*RPA1*
17	1,729,320	A	-	85	11%	N	X	*RPA1*
17	1,729,436	A	G	70	3%	K	E	*RPA1*
17	1,729,648	A	TG	57	3%	T	C	*RPA1*
17	1,729,720	-	G	51	4%	T	A	*RPA1*
17	1,738,723	A	TG	73	3%	D	V	*RPA1*
17	1,747,188	T	C	76	3%	V	A	*RPA1*
17	1,786,883	T	-	31	23%	E	X	*RTN4RL1*
17	1,787,091	A	G	50	4%	C	R	*RTN4RL1*
17	1,787,526	G	A	39	5%	R	W	*RTN4RL1*
17	1,787,771	T	C	42	5%	Y	C	*RTN4RL1*
17	1,787,850	A	G	37	5%	C	R	*RTN4RL1*
17	1,883,600	A	G	34	6%	K	R	*DPH1*
17	1,886,636	A	G	31	6%	D	G	*DPH1*
17	1,890,641	-	A	56	4%	E	E	*DPH1*
17	1,892,846	GT	-	41	5%	V	X	*OVCA2*
17	1,893,116	T	C	48	4%	Y	H	*OVCA2*
17	1,915,149	T	-	31	16%	D	X	*SMG6*
17	1,915,633	T	C	63	3%	K	R	*SMG6*
17	1,918,917	A	G	47	4%	F	S	*SMG6*
17	1,931,965	A	G	50	4%	D	D	*SMG6*
17	1,931,980	-	C	51	4%	E	E	*SMG6*
17	1,935,842	A	T	59	3%	V	E	*SMG6*
17	1,935,852	C	-	61	13%	G	X	*SMG6*
17	1,935,928	G	A	62	3%	C	C	*SMG6*
17	2,022,716	T	-	40	15%	K	X	*SMG6*
17	2,038,519	T	C	65	3%	Q	Q	*SMG6*
17	2,133,754	T	C	63	3%	Y	C	*SMG6*
17	2,147,908	T	C	49	4%	E	G	*SMG6*
17	2,148,055	A	G	55	5%	I	T	*SMG6*
17	2,148,999	A	G	48	4%	S	P	*SMG6*
17	2,149,046	A	G	49	4%	L	P	*SMG6*
17	2,149,355	T	C	55	4%	D	G	*SMG6*
17	2,149,496	A	-	50	24%	L	X	*SMG6*
17	2,149,808	T	C	40	5%	N	S	*SMG6*

17	2,149,853	A	G	44	5%	L	S	*SMG6*
17	2,149,998	T	C	56	4%	S	G	*SMG6*
17	2,150,294	-	G	57	9%	V	A	*SMG6*
17	2,150,314	A	TG	58	3%	S	S	*SMG6*
17	2,150,505	A	G	62	3%	C	R	*SMG6*
17	2,150,636	-	A	61	3%	P	L	*SMG6*
17	2,165,675	A	G	39	5%	H	R	*SRR*
17	2,171,445	A	G	21	10%	T	T	*SRR*
17	2,173,260	A	G	52	4%	K	K	*SRR*
17	2,173,262	T	C	52	4%	L	P	*SRR*
17	2,173,787	T	C	60	3%	F	S	*SRR*
17	2,173,795	G	AC	61	3%	V	T	*SRR*
17	2,173,809	A	TG	59	3%	VK	VE	*SRR*
17	2,174,808	T	C	86	2%	R	G	*TSR1*
17	2,179,408	-	G	65	3%	L	L	*TSR1*
17	2,180,598	T	C	65	3%	T	T	*TSR1*
17	2,180,624	-	G	66	3%	V	L	*TSR1*
17	2,182,218	T	C	70	3%	R	R	*TSR1*
17	2,182,339	T	C	53	4%	D	G	*TSR1*
17	2,183,691	T	C	58	3%	R	G	*TSR1*
17	2,183,740	-	C	64	3%	V	V	*TSR1*
17	2,184,519	C	T	53	4%	M	I	*TSR1*
17	2,184,655	-	C	41	5%	Q	R	*TSR1*
17	2,185,727	A	G	39	5%	H	H	*TSR1*
17	2,213,045	-	TG	31	10%	A	V	*SGSM2*
17	2,213,094	C	-	49	35%	P	X	*SGSM2*
17	2,214,221	A	G	32	6%	E	G	*SGSM2*
17	2,214,948	T	CA	45	4%	GG	GR	*SGSM2*
17	2,217,448	A	G	36	6%	K	K	*SGSM2*
17	2,221,310	T	A	41	5%	A	A	*SGSM2*
17	2,225,553	T	C	38	8%	D	D	*SGSM2*
17	2,237,259	C	T	24	8%	V	M	*MNT*
17	2,237,967	T	C	28	7%	I	V	*MNT*
17	2,238,070	T	C	43	5%	E	E	*MNT*
17	2,245,321	G	-	22	14%	A	X	*MNT*
17	2,270,142	-	C	46	4%	N	D	*METT10D*
17	2,270,821	A	G	44	5%	I	T	*METT10D*
17	2,270,826	T	C	46	4%	E	E	*METT10D*
17	2,323,639	T	C	27	7%	E	G	*METT10D*
17	2,327,887	A	G	69	3%	C	C	*METT10D*

17	2,520,311	T	C	63	3%	C	C	*PAFAH1B1*
17	2,520,342	TT	CTC	60	3%	F	L	*PAFAH1B1*
17	2,524,132	A	G	36	6%	R	G	*PAFAH1B1*
17	2,545,081	T	C	26	8%	Y	C	*KIAA0664*
17	2,545,522	G	A	28	7%	L	L	*KIAA0664*
17	2,841,387	-	C	32	6%	F	S	*GARNL4*
17	2,848,325	A	G	59	3%	T	A	*GARNL4*
17	2,848,376	G	A	57	3%	V	I	*GARNL4*
17	2,855,432	A	G	41	5%	E	G	*GARNL4*
17	2,912,824	-	T	23	9%	M	I	*OR1D4*
17	2,912,975	A	G	41	5%	L	P	*OR1D4*
17	2,913,119	-	G	52	4%	F	S	*OR1D4*
17	2,942,262	A	G	65	3%	L	P	*OR1D2*
17	2,942,722	A	G	31	6%	L	L	*OR1D2*
17	2,942,848	T	C	34	6%	N	D	*OR1D2*
17	2,976,708	T	C	70	3%	E	E	*OR1G1*
17	2,976,753	T	C	80	2%	V	V	*OR1G1*
17	2,976,766	A	G	80	2%	M	T	*OR1G1*
17	2,976,767	T	A	82	4%	M	L	*OR1G1*
17	2,976,803	T	AC	77	3%	ST	SV	*OR1G1*
17	2,976,825	-	T	79	3%	C	*	*OR1G1*
17	2,976,828	A	G	78	3%	F	F	*OR1G1*
17	2,976,853	-	G	84	2%	V	A	*OR1G1*
17	2,976,864	-	G	85	2%	H	H	*OR1G1*
17	2,976,971	-	G	77	9%	I	L	*OR1G1*
17	2,977,003	A	G	82	2%	V	A	*OR1G1*
17	2,977,236	A	G	70	3%	Y	Y	*OR1G1*
17	2,977,247	C	T	66	3%	V	I	*OR1G1*
17	2,977,399	A	G	51	4%	L	P	*OR1G1*
17	2,977,433	-	A	46	7%	L	F	*OR1G1*
17	3,048,004	T	C	30	7%	S	P	*OR1A2*
17	3,048,334	T	-	53	9%	Y	X	*OR1A2*
17	3,065,852	T	C	33	6%	L	P	*OR1A1*
17	3,066,106	T	-	49	8%	S	X	*OR1A1*
17	3,066,109	T	AC	49	4%	W	T	*OR1A1*
17	3,066,263	A	G	53	4%	Y	C	*OR1A1*
17	3,066,379	A	G	56	4%	T	A	*OR1A1*
17	3,066,507	A	T	45	4%	P	P	*OR1A1*
17	3,128,065	A	G	35	6%	D	D	*OR3A2*
17	3,141,747	T	C	60	5%	S	G	*OR3A1*

17	3,141,754	G	A	60	3%	I	I	*OR3A1*
17	3,141,776	-	G	52	4%	I	T	*OR3A1*
17	3,141,916	C	AT	47	4%	R	R	*OR3A1*
17	3,141,956	A	GT	46	4%	V	D	*OR3A1*
17	3,142,010	-	AC	42	5%	G	G	*OR3A1*
17	3,142,033	A	T	40	5%	N	K	*OR3A1*
17	3,142,056	A	GG	39	5%	LS	LP	*OR3A1*
17	3,142,089	A	G	37	5%	F	L	*OR3A1*
17	3,142,100	A	G	37	5%	V	A	*OR3A1*
17	3,142,329	A	G	21	10%	C	R	*OR3A1*
17	3,160,461	-	C	55	4%	L	P	*OR3A4*
17	3,160,674	T	C	50	4%	F	S	*OR3A4*
17	3,160,705	G	A	54	4%	L	L	*OR3A4*
17	3,160,713	T	C	54	4%	I	T	*OR3A4*
17	3,160,763	A	G	53	4%	S	G	*OR3A4*
17	3,160,919	C	T	52	4%	Q	*	*OR3A4*
17	3,248,279	-	G	13	15%	M	T	*OR1E1*
17	3,271,405	T	-	37	14%	V	X	*OR3A3*
17	3,293,249	T	C	49	4%	K	R	*SPATA22*
17	3,319,329	G	A	38	5%	A	V	*SPATA22*
17	3,326,330	-	G	55	4%	R	G	*ASPA*
17	3,326,404	T	AC	67	3%	CD	*R	*ASPA*
17	3,326,419	TTT	CTTC	61	3%	IF	IF	*ASPA*
17	3,368,703	G	A	62	6%	L	F	*TRPV3*
17	3,368,876	-	C	52	4%	E	G	*TRPV3*
17	3,378,110	T	C	40	5%	M	V	*TRPV3*
17	3,382,808	T	AC	37	5%	RM	RV	*TRPV3*
17	3,393,641	T	-	49	18%	K	X	*TRPV3*
17	3,421,667	A	T	40	5%	W	R	*TRPV1*
17	3,423,880	T	C	28	7%	P	P	*TRPV1*
17	3,423,921	-	G	28	100%	A	P	*TRPV1*
17	3,427,767	C	T	24	8%	L	L	*TRPV1*
17	3,439,959	T	C	56	5%	E	G	*TRPV1*
17	3,442,140	T	C	33	100%	Q	R	*TRPV1*
17	3,442,149	A	G	33	6%	I	T	*TRPV1*
17	3,442,302	G	A	34	6%	S	F	*TRPV1*
17	3,471,292	-	C	29	7%	R	G	*SHPK*
17	3,473,501	A	G	39	5%	G	G	*SHPK*
17	3,473,520	T	C	38	5%	K	R	*SHPK*
17	3,474,114	G	A	65	3%	T	T	*SHPK*

17	3,474,211	A	G	65	3%	V	A	*SHPK*
17	3,474,255	-	G	60	10%	G	G	*SHPK*
17	3,480,249	T	C	43	5%	Q	Q	*SHPK*
17	3,490,304	A	G	54	4%	K	E	*CTNS*
17	3,490,306	A	G	54	4%	K	K	*CTNS*
17	3,497,520	T	C	54	4%	V	A	*CTNS*
17	3,510,386	A	G	39	5%	R	G	*CTNS*
17	3,566,795	A	G	54	4%	I	T	*ITGAE*
17	3,566,821	T	C	59	3%	K	K	*ITGAE*
17	3,573,399	A	G	40	5%	C	R	*ITGAE*
17	3,574,845	A	G	35	6%	T	T	*GSG2*
17	3,575,463	T	AC	56	4%	FG	LR	*GSG2*
17	3,575,536	T	C	53	6%	L	L	*GSG2*
17	3,575,572	A	G	55	4%	I	V	*GSG2*
17	3,595,967	A	G	34	6%	L	P	*ITGAE*
17	3,607,751	CA	-	35	11%	IG	HX	*ITGAE*
17	3,609,577	-	C	41	7%	Q	R	*ITGAE*
17	3,610,226	TT	ATA	43	5%	EK	DI	*ITGAE*
17	3,611,079	-	C	23	43%	E	G	*ITGAE*
17	3,611,090	CTTG	TCTCTC	17	12%	DK	ER	*ITGAE*
17	3,611,097	TC	CTCT	18	17%	EE	ER	*ITGAE*
17	3,613,943	A	G	42	5%	L	L	*ITGAE*
17	3,613,970	T	C	43	5%	T	T	*ITGAE*
17	3,663,160	C	AT	37	5%	R	H	*C17orf85*
17	3,663,278	T	C	31	6%	K	E	*C17orf85*
17	3,664,485	T	C	40	5%	K	E	*C17orf85*
17	3,668,601	GA	-	42	5%	VP	CX	*C17orf85*
17	3,672,066	T	C	53	4%	R	G	*C17orf85*
17	3,679,399	A	G	60	3%	V	A	*C17orf85*
17	3,712,272	T	C	17	12%	E	E	*CAMKK1*
17	3,722,652	AA	GAG	37	5%	DF	DS	*CAMKK1*
17	3,734,427	C	TG	42	5%	G	A	*CAMKK1*
17	3,753,291	T	C	39	5%	Y	C	*P2RX1*
17	3,753,366	A	G	37	8%	V	A	*P2RX1*
17	3,775,472	-	C	42	5%	P	P	*ATP2A3*
17	3,780,450	T	C	21	10%	N	D	*ATP2A3*
17	3,785,278	A	G	32	6%	F	S	*ATP2A3*
17	3,797,866	-	T	49	4%	G	G	*ATP2A3*
17	3,856,965	T	C	50	4%	K	E	*ZZEF1*
17	3,864,411	A	G	71	4%	W	R	*ZZEF1*

17	3,864,416	T	C	72	3%	Q	R	*ZZEF1*
17	3,867,538	A	G	57	3%	L	P	*ZZEF1*
17	3,867,690	C	T	76	3%	L	L	*ZZEF1*
17	3,867,880	A	T	68	4%	L	*	*ZZEF1*
17	3,869,730	-	G	24	8%	F	S	*ZZEF1*
17	4,986,475	-	G	19	11%	E	G	*USP6*
17	4,992,723	T	C	16	12%	G	G	*USP6*
17	5,025,866	-	G	12	17%	S	P	*ZNF594*
17	5,026,361	G	A	13	85%	L	F	*ZNF594*

Appendix E

Alignments of erroneous SNPs

Description of the header line: Chromosome, start and end coordinates, reference and variant allele, read depth, percent of the variant reads, gene name and SNP id.

```
>chr17  1020053 1020053 G       T       44      50%       ABR     rs8068519

Reads with Difference:
chr17                   1020033+ T-GTT-T-GT-GTG-TGTGTGT-G-T--G-G-G-G-GGGGG-TG-T-GTTT-GTGTA
                                                          ***
F0FQDYC02H20MN  (2)         314- T-GTT-T-GT-GTGC-GTGTGT-G-T--G--TG-G-G-GGGGG-TG-T-GTTT-GTGTA 274
F0FQDYC02GQH1S  (2)         204- T-GTT-T-GT-GTG-TGTGTGT-G-T--G--TG-G-G-GGGGG-TG-T-GTTTTGTGTA 163
F0FQDYC02G8SVX  (2)         199- T-GTT---GT-GTG-TGTGTGT-G-TN----TG-G-G-GGGGG-TG-T-GTTTTGTGTA 159
F0FQDYC02JFNUI  (2)         282- T-GTT---GT-GTG-TGTGTGT-G-T--G--TG-G-G-GGGGG-TG-T-GTTT-GTGTA 243
F0FQDYC02JH42Z  (2)         358- T-GTT-T-GT-GTG-TGTGTGT-G-T--G--TG-G-G-GGGG--TG-T-GTTT-GTGTA 319
F0FQDYC01DO4YY              402- TTGTT---GTTGTG-TGTGTGT-G-T--G--TG-G-G-GGGGG-TG-T-GTTT-GTGTA 361
F0FQDYC01AZ4BO  (2)         220- T-GTT---GT-GTG-TGTGTGT-G-T--G--TG-G-G-GGGGG-TG-T-GTTT-GTGTA 181
F0FQDYC01BL46L  (3)         322- T-GTT---GT-GTG-TGTGTGT-G-T--G--TG-G-G-GGGGG-TG-T-GTTT-GTGTA 283
F0FQDYC01EFLQ8  (4)         342- T-GTT---GT-GTG-TGTGTGT-G-T--G--TG-G-G-GGGGG-TG-T-GTTT-GTGTA 303
F0FQDYC02IXIJQ              376- T-GTT-T-GT-GTG-TGTGTGT-G-T--G--TG-G-G-GGGGGC-G-T-GTTT-GTGTA 336
F0FQDYC02HWWYR  (4)         316- T-GTT-T-GT-GTG-TGTGTGT-G-T--G--TG-G-G-GGGG--TG-T-GTTT-GTGTA 277
F0FQDYC02FNZWQ  (2)         297- TTGTT---GT-GTG-TGTGTGT-G-T--G--TG-G-G-GGGG--TG-T-GTTT-GTGTA 258
F0FQDYC01EV69B              187- TTGTT---GTTGTG-TGTGTGT-G-T--G--TG-G--CGGGGG-TG-T-GTTT-GTGTA 146
F0FQDYC02HU6ZD              427- T-GTT-T-GT-GTGC-GTGTGT-GGT--G--TG-G-G-GGGGG-TG-T-GTTT-GTGTA 386
F0FQDYC01CJB61  (2)         328- T-GTT---GT-GTG-TGTGTGT-G-TN----TG-G-G-GGGGG-TG-T-GTTT-GTGTA 289
F0FQDYC01EGKNI              387- T-GTT-T-GT-GTG-TGTGTGT-G-T--G--TG-G-G-GGGGG-TG-T-GTTT-GTGTA 347
F0FQDYC01ESFUD              104-         GT-GTG-TGTGTGT-G-T--G--TG-G-G-GGGGG-TG-T-GTTT-GTGTA 69
F0FQDYC02GSNHH              148-         GT-GTG-TGTGTGT-G-TN----TG-G-G-GGGGG-TG-T-GTTT-GTGTA 113
F0FQDYC01EGC72              175-         GT-GTG-TGTGTGT-G-TN----TG-G-G-GGGGG-TG-T-GTTT-GTGTA 140
F0FQDYC01CNDMS              203-            GTG-TGTGTGT-G-TN----TG-G-G-GGGGG-TG-T-GTTT-GTGTA 170
```

```
F0FQDYC01BADXZ                1+                    TGT-G-T--G--TG-G-G-GGGG--TG-T-GTTT-GTGTA 26
F0FQDYC02F7QA9                1+                     GT-G-T--G--TG-G-G-GGGG--TG-T-GTTT-GTGTA 25
                                                            ***
Other Reads:
                                                            ***
F0FQDYC01EJZYE  (3)         435+ T-GTT-TTGT-GTG-TGTGTGT-G-T--G-G-G--TG-GGG                   462
F0FQDYC02J0ZCM              406+ T-GTT-T-GT-GTG-TGTGTGT-G-T--G-G-G--TG-GGG                   432
F0FQDYC02FY9HN              317+ T-GTT-T-GT-GTG-TGTGTGT-G----G-G----TG-GGGG--TG-T-GTT--GTGT  352
F0FQDYC02IIVR8              254+ T-GTT-T-GT-GTG-TGTGTGT-G-T--G-G-G--TG-GGGGG-TG-TTGTT--GTGTA 294
F0FQDYC01BPU42              247+ T-GTT-T-GT-GTG-TGTGTGT-GGT--G-G-G-GTG-GGGG--TG-TTGTT--GTGTA 288
F0FQDYC02IDUBZ              240+ T-GTT-T-GT-GTG-TGTGTGT-G-T--G-G-G--TG-GGGGG-TG-T-GTTT-GTGTA 280
F0FQDYC02H22EG              200+ T-GTT-T-GT-GTG-TGTGTGT-GGT--G-G-G-GTG-GGGG--T               230
F0FQDYC02IBWSQ              168+ T-GTT-T-GT-GTG-TGTGTGT-G-T--G-G-G--TG-GGGG--TG-T-GTTT-GTGTA 207
F0FQDYC01CHGUE  (2)         152+ T-GTT-T--T-GTG-TGTGTGT-G-T--G-GTG-G-G-GG                   177
F0FQDYC01B3UB6              101+ T-GTT-T-GT-GTG-TGTGTGT-GGT--G-G-G-GTG-GGGGG-TGGTTGTT--GTGTA 144
F0FQDYC01DT0D9              334- T-GTT---GT-GTG-TGTGTGT-G-TNTG-G-G-G-G-GGGG--TG-T-GTTTTGTGTA 293
F0FQDYC01ASE2Z  (2)         353- T-GTT---GT-GTG-TGTGTGT-G-TNTG-G-G-G-G-GGGG--TG-T-GTTT-GTGTA 313
F0FQDYC02IG3BI              255- TTGTT---GT-GTG-TGTGTGT-G-T--G-GTG-G-G-GGG---TG-T-GTTT-GTGTA 216
F0FQDYC02JG0C0              335- T-GTT-T-GT-GTG-TGTGTGTN--T--G-G-G-G-G-GGG---TG-T-GTTT-GTGTA 297
F0FQDYC01CIQ3U              414- T-GTTAT-GT-GTG-TGTGTG--G-T--G------TG-GGGGG-TGGT-GTT--GTGTA 376
F0FQDYC02G0YXZ              233- T-GTT---GT-GTG-TGTGTGT-G-T-TG-G-G-G-G-GGGG--TG-T-GTTT-GTGTA 194
F0FQDYC01ARJ01              306- T-GTT-T-GT-GTG-TGTGTGT-G-T--GTG-G-G-G-GGGGG-TG-T-GTTT-GTGTA 265
F0FQDYC01CJCIV  (2)         225- TTGTT---GT-GTG-TGTGTGTN--T--G-G-GNG-G-GGGG--TG-T-GTTTTGTGTA 184
F0FQDYC01BBBIL              238- T-GTT---GTTGTG-TGTGTGT-G-TNTG-G-G-G-G-GGGG--TG-T-GTTT-GTGTA 197
F0FQDYC01EJGSP  (3)         472- T-GTT---GT-GTG-TGTGTGT-G-TNTG-G-G-G-G-GG----TG-T-GTT--GTGTA 435
F0FQDYC01D3KMU              374- T-GTT-T-GT-GTG-TGTGTGT-GGTNTG-G-G-G-G-GGGG--TG-T-GTTT-GTGTA 332
F0FQDYC01EVGTH              165- T-GTT-T-GT-GTG-TGTGTGT-G-T-TG-G-G-G-G-GGGG--TG-T-GTTT-GTGTA 125
                                                            ***
```

```
>chr17  1674048 1674048 A       G       35      69%     SMYD4   rs719913

Reads with Difference:
chr17                   1674028+ -CT-AAAAGTTT-C--AT-AC-TTTG-AAGGGGG-TTT-C-TGTTACTTGAAACTGAAGAGT
                                                          *******
F0FP3GZ02HTLGU               40- ACTTAAAAGTTT-C--ATTAC-TTTGGA-GGGGGGTTT-C-TGTTAC                1
F0FP3GZ01BS7IC               39- -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGGGTTT-C-TGTTACTTG             1
F0FP3GZ01EIGXW  (2)          66- -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGGGTTT-C-TGTTACTTGAAACTGAAGAGT 16
F0FP3GZ01CPUS7              137- -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGGGTTT-C-TGTTACTTGAAACTGAAGAGT 87
F0FP3GZ01B4Z0J  (3)         105- -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGGGTTT-C-TGTTACTTGAAACTGAAGAGT 55
F0FP3GZ02G43D8              106- -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGGGTTT-C-TGTTACTTGAAACTGAAGAGT 56
F0FP3GZ02F97YL              359+ -CT-AAAAGTTTTC--AT-AC-TTTG-A-GGGGGGTTT-C-TGTTACTTGAAACTGAAGAGT 410
F0FP3GZ02G6C28              154- -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGGGTTT-C-TGTTACTTGAAACTGAAGAGT 104
F0FP3GZ02FVPX1              173- -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGGGTTT-C-TGTTACTTGAAACTGAAGAGT 123
F0FP3GZ02I2STE               82- -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGGGTTT-C-TGTTACTTGAAACTGAAGAGT 32
F0FP3GZ01EB5CH              286+ -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGGGTTT-C-TGTTACTTGAAACTGAAGAGT 336
F0FP3GZ01AE1AM              250- -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGGGTTT-C-TGTTACTTGAAACTGAAGAGT 200
F0FP3GZ02G555J              306- -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGGGTTT-C-TGTTACTTGAAACTGAAGAGT 256
F0FP3GZ02GN74R              181+ -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGGGTTT-C-TGTTACTTGAAACTGAAGAGT 231
F0FP3GZ02F6CL5               95- ACTTAAAAGTTT-C--ATTAC-TTTGGA-GGGGGGTTTTC-TGTTACTTGAAACTGAAGAGT 40
F0FP3GZ02HOT1Y              323- -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGGGTTT-C-TGTTACTTGAAACTGAAGAGT 273
F0FP3GZ02FHOIO               92- -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGGGTTT-C-TGTTACTTGAAACTGAAGAGT 42
F0FP3GZ01D3VJT  (2)         153+ -CT-AAAAGTTT-C--AT-ACCTT-G-A-GGGGGGTTT-C-TGTTACTTGAAACTGAAGAGT 203
F0FP3GZ02I6H2O  (3)         389- -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGGGTTT-C-TGTTACTTGAAACTGAAGAGT 339
F0FP3GZ01CCYJ0              320- -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGGGTTT-C-TGTTACTTGAAACTGAAGAGT 270
F0FP3GZ02F9G5Y               73+ -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGGGTTT-C-TGTTACTTGAAACTGAAGAGT 123
F0FP3GZ01C6OOB  (3)         476- -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGGGTTT-C-TGTTACTTGAAACTGAAGAGT 426
F0FP3GZ01AH7EU               47+ -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGGGTTT-C-TGTTACTTGAAACTGAAGAGT 97
F0FP3GZ01CLLDZ              451- -CT-AAAAGTTT-C-TAT-AC-TTTG-A-GGGGGGTTT-CGTGTTACTTGAAACTGAAGAGT 399
                                                          *******
Other Reads:
                                                          *******
F0FP3GZ01EFZ3U  (3)         421+ -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGG-TTT-C-TGTTACTTGAAACTGAAGAGT 470
F0FP3GZ01BC6I3              421+ -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGG-TTT-C-TGT                   452
F0FP3GZ02GRKMX              418+ -CT-AAAAGTTT-C--AT-AC-TT-G-A-GGGG--TTT-C-TGTTACTTGAAACTG       459
F0FP3GZ02IWH9E              391+ -CT-AAAAGTT--C--AT-AC-TTTG-A-GGGGG-TTT-C-TGTTACTTGAAACTGAAGAGT 439
F0FP3GZ02H2AXO              293+ -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGG-TTT-C-TGTTACTTGAAACTGAAGAGT 342
F0FP3GZ01DWJUI              292+ -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGG-TTT-C-TGTTACTTGAAACTGAAGAGT 341
F0FP3GZ01EPZTZ              269+ -CT-AAAAGTT--CC-AT-AC-TTTG-A-GGGGG-TTT-C-TGTTACTTGAAACTGAAGAGT 318
F0FP3GZ01EU0W9              260+ -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGG-TTT-C-TGTTACTTGAAACTGAAGAGT 309
F0FP3GZ01CPUFG              105+ -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGG-TTT-C-TGTTACTTGAAACTGAAGAGT 154
F0FP3GZ01DECCD               67+ -CT-AAAAGTTT-C--AT-AC-TTTG-A-GGGGG-TTT-C-TGTTACTTGAAACTGAAGAGT 116
F0FP3GZ01CLHG3              276- ACTTAAAAGTTT-C--ATTAC-TT-G-A-GGGGG-TTT-C-TGTTACTTGAAACTGAAGAGT 225
                                                          *******
```

```
>chr17  3137283 3137283 T       C       27      67%       rs231674

Reads with Difference:
chr17                   3137257+ T-AATTTCTCATCA-TCC-ACCCCCCTCC-TACCCACCCAC-ATTT-CT-GATAC-A-T-C
                                                            **
F0FP3GZ01BGBHH               77- T-AATTTCTCATCA-TCC-ACCCCC-TCCC-ACCCACCCAC-ATTT-CT-GATAC-A-T-C 28
F0FP3GZ01B7BUJ  (2)          71- T-AATTTCTCATCA-TCC-ACCCCC-TCCC-ACCCACCCAC-ATTT-CT-GATAC-A-T-C 22
F0FP3GZ02J093O              125- T-AATTTCTCATCA-TCC-ACCCCCCTCCC-ACCCACCCAC-ATTT-CT-GATAC-A-T-C 75
F0FP3GZ01B4505              138- T-AATTTCTCATCA-TCCCACCCCC-TCCC-ACCCACCCAC-ATTT-CT-GATAC-A-T-C 88
F0FP3GZ01AS2WA              277+ T-AATTTCTCATCA-TCC-ACCCCCCTCCC-ACCCACCCAC-ATTT-CT-GATAC-A-T-C 327
F0FP3GZ02JK1G0              268+ T-AATTTCTCATCA-TCC-ACCCCCCTCCC-ACCCACCCAC-ATTT-CT-GATAC-A-T-C 318
F0FP3GZ01BU4OI  (2)         262+ T-AATTTCTCATCA-TCC-ACCCCCCTCCC-ACCCACCCAC-ATTT-CT-GATAC-A-T-C 312
F0FP3GZ02HFHU9              250+ T-AATTTCTCATCA-TCC-ACCCCCCTCCC-ACCCACCCAC-ATTTTCT-GATAC-A-T-C 301
F0FP3GZ02F8B31  (2)         195+ T-AATTTCTCATCA-TCC-ACCCCCCTCCC-ACCCACCCAC-ATTT-CT-GATAC-A-T-C 245
F0FP3GZ02GGWQX              136+ T-AATTTCTCATCA-TCC-ACCCCCCTCCC-ACCCACCCAC-ATTT-CT-GATAC-A-T-C 186
F0FP3GZ01E3VLU              122+ T-AATTTCTCATCA-TCC-ACCCCCCTCCC-ACCCACCCAC-ATTT-CT-GATAC-A-T-C 172
F0FP3GZ01ECLIR               59- T-AATTTCTCATCA-TCC-ACCCCC-TCCC-ACCCACCCAC-ATTT-CT-GATAC-A-T-C 10
F0FP3GZ02JLMOE              411- T-AATTTCTCATCA-TC--ACCCCC-TCCC-ACCCACCCAC-ATTT-CT-GATAC-AAT-C 362
F0FP3GZ02FTDZB              433- T-AATTTCTCATCA-TCC-ACCCCC-TCCC-ACCCACCCAC-ATTT-CT-GATAC-A-T-C 384
F0FP3GZ01C0LE5              131- TTAATTTCTCATCACTCCCACCCCC-TCCC-ACCCACCCAC-ATTT-CT-GATAC-A-T-C 79
F0FP3GZ02FN8ZA  (2)         387- T-AATTTCTCATCA-TC--ACCCCC-TCCC-ACCCACCCAC-ATTT-CT-GATAC-A-T-C 339
F0FP3GZ02JWQF7              339- T-AATTTCTCATCA-TCC-ACCCCCCTCCC-ACCCACCCAC-ATTT-CT-GATACCA-T-C 288
F0FP3GZ01A2HTG              429- T-AATTTCTCATCACTCC-ACCCCCCTCCC-ACCCACCCAC-ATTT-CT-GATAC-A-T-C 378
                                                            **
Other Reads:
                                                            **
F0FP3GZ02INZFQ              460+ T-AATT-CTCATCA-TCC-ACCCCC-TCC--ACCCACCCAC-ATTT-CT-GATAC-A-T-C 507
F0FP3GZ02HB4IP              448+ T-AATTTCTCATCA-TCC-ACCCCCCTCC--ACCCACC-AC-ATTT-CT-GATAC-A-T-C 496
F0FP3GZ01DU9FN  (2)         245+ T-AATTTCTCATCA-TCC-ACCCCC-TCC--ACCCACCCAC-ATTT-CT-GATAC-A-T-C 293
F0FP3GZ01DGCGJ  (2)         232+ T-AATTTCTCATCA-TCC-ACCCCC-TCC--ACCCACCCAC-ATT--CTTGATAC-A-T-C 280
F0FP3GZ01CRL5N              318- T-AATTTCTCATCA-TCC-ACCCCC-TCC--ACCCACCCAC-ATTT-CT-GATAC-A-TTC 269
F0FP3GZ02HZ95E              177+ T-AATTTCTCATCA-TCC-ACCCCC-TCC--ACCCACCCAC-ATTT-CT-GATAC-A-T-C 225
F0FP3GZ01A4SLV              179+ T-AATTTCTCATCA-TCC-ACCCCC-TCC--ACCCACCCACCATTT-CT-GATAC-A-T-C 228
F0FP3GZ01EC9G3  (2)         170+ T-AATTTCTCATCA-TCC-ACCCCC-TCC--ACCCACC-AC-ATTT-CT-GATAC-A-T-C 217
F0FP3GZ01A6YLU              140+ T-AATTTCTCATCA-TCC-ACCCCC-TCC--ACCCACCCACCATTT-CTTGATAC-A-T   189
                                                            **
```

```
>chr17  6954089 6954089 G       A       43      93%       ASGR2   rs314239
Reads with Difference:
chr17                   6954067+ CTTC-TCA-TTCCT-A-T-C-ACCCCCC--GA-CCC-A-GCAGGACCACAT-CCTGGGGAG
                                                           ***
F0FQDYC02HJGY9  (2)          31- CTTC-TCA-TTCCT-A-T-C-ACCCCC--A-A-CCC-A-GCAG                   1
F0FQDYC01D3YSY              107- CTTC-TCA-TTCCT-A-T-C-ACCCCC--A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 60
F0FQDYC01AM8HL              338+ CTTC-TCA-TTCCT-A-T-CCACCCCCC-A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 387
F0FQDYC01EAHI7  (2)         127- CTTC-TCA-TTCCT-A-T-C-ACCCCC--A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 80
F0FQDYC02HZD0P              139- CTTC-TCA-TTCCT-A-T-C-ACCCCCC-A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 91
F0FQDYC01BRF9N              270+ CTTC-TCA-TTCCT-A-T-C-ACCCCC--A-A-CC--A-GCAGGACCACAT-CCTGGGGAG 316
F0FQDYC02I070H              252+ CTTC-TCA-TTCCT-A-T-CCACCCCCCTA-A-CC--A-GC                    282
F0FQDYC01DZ0B5  (3)          56- CTTC-TCA-TTCCT-A-T-C-ACCCCC--A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 9
F0FQDYC01AXT1O  (3)          58- CTTC-TCA-TTCCT-A-T-C-ACCCCC--A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 11
F0FQDYC01CD0A4              220+ CTTC-TCA-TTCCT-A-T-C-ACCCCC--A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 267
F0FQDYC01DU32S               58- CTTC-TCA-TTCCT-A-T-C-ACCCCCC-A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 10
F0FQDYC01DUOH0              208+ CTTC-TCA-TTCCT-A-T-C-ACCCCC--A-A-CC--A-GCAGGACCACAT-CCTGGGGAG 254
F0FQDYC01BICEF  (2)         205+ CTTC-TCA-TTCCT-A-T-CCACCCCCC-A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 254
F0FQDYC01CZ0UY              197- CTTC-TCA-TTCCT-A-T-C-ACCCCC--A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 150
F0FQDYC01DX3HQ  (2)         296- CTTC-TCA-TTCCT-A-T-C-ACCCCC--A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 249
F0FQDYC01BMP0I              195- CTTC-TCA-TTCCT-A-T-C-ACCCCCC-A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 147
F0FQDYC01B9YEG              112- CTTC-TCA-TTCCT-A-T-C-ACCCCCC-A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 64
F0FQDYC02F0RIK              203- CTTC-TCA-TTCCT-A-T-C-ACCCCC--A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 156
F0FQDYC01CUTDX              141- CTTC-TCA-TTCCT-A---CCACCCCCC-A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 93
F0FQDYC02F8SV1              226- CTTC-TCA-TTCCT-A-T-C-ACCCCC--A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 179
F0FQDYC01BF8OJ               77- CTTC-TCA-TTCCT-A-T-C-ACCCCC--A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 30
F0FQDYC01BQOA7  (3)         113- CTTC-TCA-TTCCT-A-T-C-ACCCCCC-A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 65
F0FQDYC01DL14J              219- CTTC-TCA-TTCCT-A-T-C-ACCCCC--A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 172
F0FQDYC02H4BNQ              238- CTTC-TCA-TTCCT-A-T-C-ACCCCC--A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 191
F0FQDYC01CNS32  (2)         159+ CTTC-TCA-TTCCT-A-T-C-ACCCCC--A-A-CC--A-GCAGGACCACAT-CCTGGGGAG 205
F0FQDYC01C0H0H               43- CTTC-TCA-TTCCT-A-T-C-ACCCCCC-A-A-CCC-A-GCAGGACCACAT-CCT       1
F0FQDYC02GQL72              303- CTTC-TCA-TTCCT-A-T-C-ACCCCC--A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 256
F0FQDYC02IWTBZ  (2)         203- CTTC-TCA-TTCCT-A-T-C-ACCCCCC-A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 155
F0FQDYC01AXE9S              267- CTTC-TCA-TTCCT-A-T-C-ACCCCCC-A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 219
F0FQDYC02G9JZQ  (2)         355- CTTC-TCA-TTCCT-A-T-C-ACCCCC--A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 308
F0FQDYC01ALP0Z              127+ CTTC-TCA-TTCCT-A-T-C-ACCCCC--A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 174
F0FQDYC02HMEG6              128- CTTC-TCAATTCCT-A-T-C-ACCCCC--A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 80
F0FQDYC01D2TDN              308- CTTC-TCA-TTCCT-A-T-C-ACCCCC--A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 261
F0FQDYC01AHLKN               75+ CT-CCTCA-TTCCT-A-T-C-ACCCCCC-A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 123
F0FQDYC02IXD9Z               65+ CTTC-TCA-TTCCT-A-T-C-ACCCCC--A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 112
F0FQDYC02IGZFV              404- CTTC-TCA-TTCCT-ACT-C-ACCCCC--A-A-CCC-A-GCAGGACCACATAC-TGGGGAG 356
F0FQDYC01A48QN              366- CTTC-TCA-TTCCTTACT-C-ACCCCCC-A-A-CCCCACGCAGGACCACAT-CCTGGGGAG 314
F0FQDYC01CB2WJ               20+ CTTC-TCA-TTCCT-A-T-C-ACCCCCC-A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 68
F0FQDYC02G69IX  (2)         483- CTTC-TCA-TTCCT-ACT-C-ACCCCCC-A-A-CCC-ACGCAGGACCACAT-CCTGGG-AG 434
F0FQDYC02FNMUO              307-       CA-TTCCT-A-T-C-ACCCCC--A-A-CCC-A-GCAGGACCACAT-CCTGGGGAG 265
                                                           ***
Other Reads:
F0FQDYC02IYA6S               37- CTTC-TCA-TTCCT-A-T-C-ACCCCC--AG--CCC-A-GCAGGACCAC            1
F0FQDYC01EBQAE               59+ CTTC-TCA-TTCCT-A-T-C-ACCCCCC-AG--CCC-A-GCAGGACCACAT-CCTGGGGAG 107
F0FQDYC02J1MRN              194- CTTC-TCA-TTCCT-A-TN--ACCCCC----AACCC-A-GCAGGACCACAT-CCTGGGGAG 147
                                                           ***
```

```
>chr17  7113355 7113355 C       T       42      69%       rs8067500

Reads with Difference:
chr17                    7113328+ TC-TCAA-CCTTGA-CACTGTT-G-AC-ATTTT--CTCCCC-T-GCTATCTCTT-GGG-CT
                                                             ***
F0FQDYC02IQEPI               392+ TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTTT--TCCCCN--GCTATCTCTT-GGG-CT 440
F0FQDYC01DZH9J               267+ TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTTT--TCCCCN--GCTATCTCTT-GGG-CT 315
F0FQDYC02JYP5H               289- TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTTT--TCCCC-T-GCTATCTCTT-GGG-CT 241
F0FQDYC02F4HB9               345- TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTTT--TCCCC-T-GCTATCTCTT-GGG-CT 297
F0FQDYC01C0VNZ  (2)          171+ TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTTT--TCCCC-T-GCTATCTCTT-GGG-CT 219
F0FQDYC01EK5KZ               194- TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTTT--TCCCC-T-GCTATCTCTT-GGG-CT 146
F0FQDYC01DYLVR  (2)          264- TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTTTT-TCCCC-T-GCTATCTCTT-GGG-CT 215
F0FQDYC01DVC3W               158+ TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTTT--TCCCC-T-GCTATCTCTT-GGG-CT 206
F0FQDYC01CAX1J               188- TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTTT--TCCCC-T-GCTATCTCTT-GGG-CT 140
F0FQDYC01EKBZV               347- TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTTT--TCCCC-T-GCTATCTCTT-GGG-CT 299
F0FQDYC01ASREN               125+ TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTTT--TCCCC-T-GCTATCTCTT-GGG-CT 173
F0FQDYC01EIZCS               243- TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTTT--TCCCC-T-GCTATCTCTT-GGG-CT 195
F0FQDYC01CAEMP               273- TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTTT--TCCCC-T-GCTATCTCTT-GGG-CT 225
F0FQDYC01B71EU               312- TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTTT--TCCCC-T-GCTATCTCTT-GGG-CT 264
F0FQDYC02IIQLM               315- TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTTT--TCCCC-T-GCTATCTCTT-GGGGCT 266
F0FQDYC01BRYZ6  (3)          110+ TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTTT--TCCCC-T-GCTATCTCTT-GGG-CT 158
F0FQDYC01B72AL  (2)          260- -C-TCA-GCCTTGA-CACTGTT-G-AC-ATTTTT--TCCCC-T-GCTATCTCTT-GGG-CT 213
F0FQDYC02I449V  (2)          299- TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTTT--TCCCC-T-GCTATCTCTT-GGG-CT 251
F0FQDYC01A8JFC               457- TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTTT--TCCCC-T-GCTATCTCTT-GGG-CT 409
F0FQDYC01EEGO5               458- -CC-CA-GCCTTGA-CACTGTT-G-AC-ATTTTT--TCCCC-T-GCTATCTCTT-GGG-CT 411
F0FQDYC01ERC64               340- TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTTT--TCCCC-T-GCTATCTCTTTGGG-CT 291
F0FQDYC02JJ9KR                 8+ TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTTT--TCCCC-T-GCTATCTCTT-GGG-CT 56
F0FQDYC01BGW21               304- TC-TCA-GCCTTGA-CACTGTTTG-AC-ATTTTT--TCCCC-T-GCTATCTCTT-GGG-CT 255
F0FQDYC01BW8T1                 7+ TC-TCA-GCCTTGAACACTGTT-G-AC-ATTTTT--TCCCC-T-GCTATCTCTT-GGG-CT 56
F0FQDYC02J51PV                76- TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTTT--TCCCC-T-GCTATCTCTT-GGG-CT 28
F0FQDYC01DF2K7               407-  C-TCA-GCCTTGA-CACTGTT-G-AC-ATTTTT--TCCCC-T-GCTATCTCTT-GGG-CT 360
F0FQDYC02GVMD1                 1+          CCTTGA-CACTGTT-G-AC-ATTTTT--TCCCC-T-GCTATCTCTT-GGG-CT 43
F0FQDYC01DC1W8               462-                  ACTGTT-GTACCATTTTT--TCCCC-T-GCTATCTCTT-GGG-CT 425
F0FQDYC01ERL9F               172-                     GTT-G-AC-ATTTTT--TCCCC-T-GCTATCTCTT-GGG-CT 140
                                                             ***
Other Reads:
F0FQDYC01B36YZ               202- TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTT---TCCCC-T-GCTATCTCTT-GGG-CT 155
F0FQDYC02G9GPW               231+ TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTT---TCCCC-T-GCTATCTCTT-GGG-CT 278
F0FQDYC02HDQHG               209- TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTT---TCCCC-T-GCTATCTCTT-GGG-CT 162
F0FQDYC02J6NK1               153+ TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTT---TCCCC-T-GCTATCTCTT-GGG-CT 200
F0FQDYC02HRGJ9               119+ TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTT---TCCCC-T-GCTATCTCTT-GGG-CT 166
F0FQDYC01DYQRG               113+ TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTT---TCCCC-T-GCTATCTCTT-GGG-CT 160
F0FQDYC01D3GVX               194- TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTT---TCCCC-T-GCTATCTCTT-GGG-CT 147
F0FQDYC01BEJPW               351- TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTT---TCCCC-T-GCTATCTCTT-GGG-CT 304
F0FQDYC01AZIAC  (2)          434- TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTT---TCCCC-T-GCTATCTCTT-GGG-CT 387
F0FQDYC01DT9TI               224- TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTT---TCCCC-T-GCTATCTCTT-GGG-CT 177
F0FQDYC01DWXUC                53+ TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTT----CCCC-TTGCTATCTCTT-GGG-CT 100
F0FQDYC01B55KO               104- TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTT---TCCCC-T-GCTATCTCTT-GGG-CT 57
F0FQDYC02I1W94                43+ TC-TCA-GCCTTGA-CACTGTT-G-AC-ATTTT---TCCCC-T-GCTATCTCTT-GGG-CT 90
                                                             ***
```

```
>chr17  1674048 1674048 A       G       31      71%    SMYD4   rs719913

Reads with Difference:
chr17                    1674023+ CATAG-CT-AAAAGTTTCAT-ACTTT--GAAGGGGG--T-TTCTGTTACTT-GAAACTGAA
                                                          *******
F0FQDYC01AVMWN                49- CATAGACTTAAAAGTTTCATTACTTT--GA-GGGGGG-T-TTCTGTTACTT-GAA       1
F0FQDYC01DKDWN                56- CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGGG-T-TTCTGTTACTT-GAAACTGAA 5
F0FQDYC02JXYBK  (2)          230+ CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGGG-T-TTCTGTTACT-CGAAACTGAA 281
F0FQDYC01AS3DN                89- CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGGG-T-TTCTGTTACTT-GAAACTGAA 38
F0FQDYC01A3HLJ                91- CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGGG-T-TTCTGTTACTT-GAAACTGAA 40
F0FQDYC01DKPX6               130- CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGGGCT-T-CTGTTACTT-GAAACTGAA 79
F0FQDYC02F5WLR  (3)          186- CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGGG-T-TTCTGTTACTT-GAAACTGAA 135
F0FQDYC01CTYKF  (2)          195- CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGGG-T-TTCTGTTACTT-GAAACTGAA 144
F0FQDYC01CDEUQ               212- CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGGG-T-TTCTGTTACTT-GAAACTGAA 161
F0FQDYC01B2SGS               234- CATAG-CT-AAAAGTTTCAT-ACTTTN--A-GGGGGG-T-TTCTGTTACTT-GAAACTGAA 183
F0FQDYC01CW50U               297- CATAG-CT-AAAAGTTTCAT-ACTTTN--A-GGGGGG-T-TTCTGTTACTT-GAAACTGAA 246
F0FQDYC01B0CFM  (2)          222+ CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGGG-T-TTCTGTTACTT-GAAACTGAA 273
F0FQDYC01B18W1               214+ CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGGG-T-TTCTGTTACTT-GAAACTGAA 265
F0FQDYC01B7OKY               284- CATAG-CT-AAAAGTTTCAT-ACTTTN--A-GGGGGG-T-TTCTGTTACTT-GAAACTGAA 233
F0FQDYC02FOEWJ               330- CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGGG-T-TTCTGTTACTT-GAAACTGAA 279
F0FQDYC01CJCSY               129+ CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGGG-T-TTCTGTTACTT-GAAACTGAA 180
F0FQDYC02JFEIA                85- CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGGG-T-TTCTGTTACTT-GAAACTGAA 34
F0FQDYC01BSFK1               308- CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGGG-T-TTCTGTTACTT-GAAACTGAA 257
F0FQDYC02HRIZF                41+ CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGGG-T-TTCTGTTACTT-GAAACTGAA 92
F0FQDYC02IRR5J                38+ CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGGG-T-TTCTGTTACTT-GAAACTGAA 89
F0FQDYC02HB0P9  (2)          323- CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGGG-T-TTCTGTTACTT-GAAACTGAA 272
F0FQDYC02F9DVI               517- CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGGG-T-TTCTGTTACTT-GAAACTGAA 466
                                                          *******
Other Reads:
                                                          *******
F0FQDYC02JGU6J               345+ CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGG--TCT-CTGTTACTT-GAAACTGAA 395
F0FQDYC01AX8CD               295+ CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGG--T-TTCTGTTACTT-GAAACTGAA 345
F0FQDYC01CBKKR               262+ CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGG--T-TTCTGTTACTT-GAAACTGAA 312
F0FQDYC01DL1UE               236+ CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGG--T-TTCTGTTACTT-GAAACTGAA 286
F0FQDYC01AND8O  (2)          254+ CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGG--T-TTCTGTTACTT-GAA-CTGAA 303
F0FQDYC02FWOXX               167+ CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGG--T-TTCTGTTACTT-GAAACTGAA 217
F0FQDYC02FPKQL  (3)          339- CATAG-CT-AAAAGTTTCAT-ACTTTN--A-GGGGG--T-TTCTGTTACTT-GAAACTGAA 289
F0FQDYC02HRLTH                61+ CATAG-CT-AAAAGTTTCAT-ACTTT--GA-GGGGG--T-TTCTGTTACTT-GAAACTGAA 111
F0FQDYC01DCGDP               405- CATAG-CT-AAAAGTTTCAT-ACTTT-TGA-GGGGG--T-TTCTGTTACTT-GAAACTGAA 354
                                                          *******
```

```
>chr17  3137283 3137283 T       C       23      74%       rs231674

Reads with Difference:
chr17                   3137255+ AGTAATTTCTCATC-ATCCACCCCCCTCC--TACCC-ACCCACATTT-CTGAT-AC-AT
                                                                 ***
F0FQDYC02GWKGT  (2)          71- AGTAATTTCTCATC-ATCCACCCCC-TCCCC-ACCC-ACCCACATTT-CTGAT-AC-AT 20
F0FQDYC01DRM0P               64- AGTAATTTCTCATC-ATCCACCCCCCTCCC--ACCC-ACCCACATTT-CTGAT-AC-AT 13
F0FQDYC02GM4GA              310+ AGTAATTTCTCATC-ATCCACCCCCCTCCC--ACCC-ACCCACATTT-CTGAT-AC-AT 361
F0FQDYC02HP5FW               61- AGTAATTTCTCATC-ATCCACCCCCCTCCCC-ACCC-ACCCACATTT-CTGAT-AC-AT 9
F0FQDYC02FUDMH              142- AGTAATTTCTCATC-ATCCACCCCCCTCCCC-ACCCCACCCACATTT-CTGAT-AC-AT 89
F0FQDYC02F4B40              286+ AGTAATTTCTCATC-ATCCACCCCCCTCCC--ACCC-ACCCACATTT-CTGAT-AC-AT 337
F0FQDYC01EHPPY              207- AGTAATTTCTCATCG-TC-ACCCCC-TCCC--ACCC-ACCCACATTT-CTGAT-AC-AT 158
F0FQDYC01DS55U              284- -GTAATTTCTCATC-ATCCACCCCCCTCCC--ACCC-ACCCACATTT-CTGAT-AC-AT 234
F0FQDYC01D7A67  (2)          209+ AGTAATTTCTCATC-ATCCACCCCCCTCCC--ACCC-ACCCACATTT-CTGAT-AC-AT 260
F0FQDYC01CNKJ5              154+ AGTAATTTCTCATC-ATCCACCCCCCTCCC--ACCC-ACCCACATTT-CTGAT-AC-AT 205
F0FQDYC02JA5X6              113+ AGTAATTTCTCATC-ATCCACCCCCCTCCC--ACCC-ACCCACATTT-CTGAT-AC-AT 164
F0FQDYC02FP69L               62+ AGTAATTTCTCATC-ATCCACCCCCCTCCC--ACCC-ACCCACATTT-CTGAT-AC-AT 113
F0FQDYC01AZGHI              418- AGTAATTTCTCATC-ATCCACCCCC-TCCC--ACCC-ACCCACATTT-CTGAT-AC-AT 368
F0FQDYC02GRQA3  (2)          219- AGTAATTTCTCATC-ATC-ACCCCC-TCCC--ACCC-ACCCACATTT-CTGAT-AC-AT 170
F0FQDYC02FKXF2               15+ AGTAATTTCTCATC-ATCCACCCCCCTCCC--ACCC-ACCCACATTT-CTGAT-AC-AT 66
F0FQDYC01BSQ2D              357-  GTAATTTCTCATC-ATCCACCCCC-TCCC--ACCC-ACCCACATTT-CTGAT-AC-AT 308
F0FQDYC02G8M9K              429-                  CCACCCCC-TCCC--ACCC-ACCCACATTT-CTGAT-AC-AT 395
                                                                 ***
Other Reads:
                                                                 ***
F0FQDYC02GBTUK              348+ AGTAATTTCTCATC-ATCCACCCCCCTCC---ACCC-ACCCACATTTTCTGAT-AC-AT 399
F0FQDYC01DY22T              220- AGTAATTTCTCATC-ATC-ACCCCC-TCC---ACCC-ACCCACATTT-CTGAT-AC-AT 172
F0FQDYC02HPK3P              208+ AGTAATTTCTCATC-ATCCACCCCC-TCC---ACCC-ACC-ACATTTTCTGATTACCAT 259
F0FQDYC01D86AX              197+ AGTAATTTCTCATC-ATCCACCCCC-TCC---ACCC-ACC-ACATTT-CTGATTAC-AT 246
F0FQDYC01DIP65  (2)          165+ AGTAATTTCTCATC-ATCCACCCCC-TCC---ACCC-A                      196
F0FQDYC01A4J86              242- AGTAATTTCTCATC-ATC-ACCCCC-TCC---ACCC-ACCCACATTT-CTGAT-AC-AT 194
                                                                 ***
```

Appendix F

Mendelian errors

Table F.1: Mendelian errors with different cut-off values[a]

	05-981			05-982			05-985			05-987		
Pos.	G.	D.	F.	G.	D.	F.	G.	D.	F.	G.	D.	F.
83,173	2	57	100	2	41	100	2	51	100	1	48	52
122,807	2	29	72	2	37	81	1	19	63	1	32	69
122,813	2	29	72	2	36	83	1	19	63	2	32	72
123,233	2	4	75	2	5	80	1	6	50	2	14	71
123,234	2	6	83	2	5	80	1	6	50	2	15	73
134,431	2	35	89	2	19	89	2	33	79	2	37	95
157,037	2	15	87	2	8	100	2	11	82	1	13	69
164,514	1	30	67	1	29	69	1	20	65	1	23	57
165,403	1	29	69	2	18	78	1	21	52	1	33	52
166,814	2	10	90	2	6	100	2	5	80	2	10	90
182,990	2	36	83	2	39	82	2	31	71			
203,333	1	40	67	2	30	73	1	35	49	1	30	63
237,217	2	12	92	2	18	83	2	10	80	2	16	81
251,752	1	47	66	2	40	72	1	41	59	1	56	48
566,592	2	41	85	2	33	82	1	49	69	2	46	78
678,908	2	42	100	2	42	100	1	38	66	2	47	100
691,574	2	23	96	2	18	83	2	7	100	2	25	76
691,576	1	22	68	2	18	72	2	7	86	1	26	54
1,023,057	2	11	91	1	10	70	1	17	59	2	7	86

1,132,247	2	6	100	2	7	86	2	9	78	2	10	90
1,324,628	2	42	88	2	33	97	2	44	80	2	35	86
1,705,407	1	33	70	1	29	69	2	25	76	1	29	62
1,849,505	1	41	66	2	37	76	1	36	61	2	47	77
1,885,471	1	21	67	2	12	83	1	14	57	1	17	53
2,154,360	2	12	92	2	20	90	1	20	70	1	22	64
2,154,377	2	10	90	2	19	89	1	20	65	1	23	57
2,154,381	2	10	90	2	19	84	1	19	58	1	24	54
2,154,388	2	11	73	2	18	83	1	19	53	1	24	50
2,154,391	2	11	82	2	18	83	1	19	63	1	24	50
2,154,419	2	7	71	2	14	79	1	21	48	1	24	46
2,502,102	2	38	100	2	26	92	1	31	68	2	39	92
2,523,841	2	46	83	2	41	83	2	46	78	1	39	69
2,801,993	2	17	94	2	19	89	2	12	75	2	17	100
2,869,983	2	42	71	2	24	71	1	35	63	2	42	81
3,138,363	2	22	82	2	28	82	1	21	67	2	20	95
3,150,616	2	13	85	2	8	100	2	11	91	2	8	75
3,154,197	2	30	87	2	20	95	2	22	100	2	27	74
3,204,089	2	12	83	2	14	93	2	15	100	2	9	78
3,407,986	2	42	83	2	31	97	2	44	75	2	35	77
3,422,735	2	22	82	2	20	95	2	25	76	1	22	68
3,745,996	2	15	80	2	13	85	1	16	56			
3,992,558	2	29	72	2	21	90	2	38	71	1	39	69
3,995,787	1	52	67	1	47	70	1	45	62	2	48	71
4,445,431	2	55	82	2	22	86	2	36	83	2	35	80
4,522,032	2	29	86	2	11	91	1	23	70	2	31	94
4,533,315	2	14	86	2	9	100	2	12	92	2	13	77
4,565,544	2	20	85	2	20	80	2	24	83	1	18	56
4,677,997	2	21	81	2	18	72	1	22	68	1	22	59
5,016,562	2	10	100	2	10	90	2	10	80	2	16	94
5,085,205	2	16	94	2	14	79	1	20	60	2	30	83
5,102,912	2	45	91	2	42	90	2	35	80	2	37	92
5,508,616	2	17	71	2	14	93	2	16	100	1	22	68
5,631,114	2	53	91	2	41	88	2	50	88	1	50	70
5,686,057	2	4	75	2	4	100	2	7	71	1	8	62

5,690,146	2	19	95	2	17	100	2	16	100	1	19	68
5,690,164	2	20	90	2	21	86	2	16	100	1	20	70
5,690,173	1	16	69	1	18	67	1	16	56	1	15	53
5,713,997	2	40	92	2	31	84	1	38	66	1	42	69
5,752,670	2	17	94	2	7	100	2	6	100	2	9	78
5,756,144	1	16	69	1	27	67	1	17	53	1	26	69
5,982,779	2	39	79	2	25	80	1	52	65	2	48	77
6,297,160	2	17	88	2	12	92	1	25	68	2	20	80
6,581,945	2	17	100	2	13	92	1	13	62	2	17	100
6,629,270	2	14	79	2	9	89	1	13	69	2	11	100
6,915,766	2	20	95	2	19	100	2	35	97	2	19	100

[a]: Genotypes: 0, AA; 1, AB; 2, BB. G. Abbreviations: D., Depth; F., Freq

Appendix G

Functional annotation clusters

Table G.1: Top seven functional annotation clusters

Cluster 1	**E. Score: 10.09**	**Count**	**P**	**Benjamini**
GOTERM_BP_FAT	axonogenesis	29	6.3E-13	1.3E-9
GOTERM_BP_FAT	cell morphogenesis involved in neuron diff.	30	1.7E-12	1.8E-9
GOTERM_BP_FAT	cell morphogenesis involved in diff.	32	3.0E-12	2.1E-9
GOTERM_BP_FAT	neuron projection morphogenesis	29	4.4E-12	2.3E-9
GOTERM_BP_FAT	cell projection morphogenesis	31	4.5E-12	1.9E-9
GOTERM_BP_FAT	neuron projection development	32	6.4E-12	2.2E-9
GOTERM_BP_FAT	cell part morphogenesis	31	1.6E-11	4.8E-9
GOTERM_BP_FAT	cell morphogenesis	37	4.6E-11	1.2E-8
GOTERM_BP_FAT	cell projection organization	37	1.2E-10	2.7E-8
GOTERM_BP_FAT	neuron development	35	1.7E-10	3.5E-8
GOTERM_BP_FAT	neuron differentiation	41	3.8E-10	7.4E-8
GOTERM_BP_FAT	cellular component morphogenesis	38	4.4E-10	7.8E-8
GOTERM_BP_FAT	axon guidance	17	1.3E-7	1.4E-5
GOTERM_BP_FAT	cell motion	31	4.9E-6	4.5E-4
Cluster 2	**E. Score: 9.76**	**Count**	**P**	**Benjamini**

GOTERM_CC_FAT	neuron projection	38	7.2E-14	2.3E-11
GOTERM_CC_FAT	axon	20	6.4E-9	2.9E-7
GOTERM_CC_FAT	cell projection	50	1.2E-8	4.0E-7
Cluster 3	**E. Score: 7.64**	**Count**	**P**	**Benjamini**
GOTERM_BP_FAT	cell adhesion	50	4.7E-10	7.7E-8
GOTERM_BP_FAT	biological adhesion	50	4.9E-10	7.5E-8
SP_PIR_KEYWO.	cell adhesion	38	1.1E-8	1.4E-6
GOTERM_BP_FAT	cell-cell adhesion	21	1.2E-4	8.7E-3
Cluster 4	**E. Score: 5.76**	**Count**	**P**	**Benjamini**
SP_PIR_KEYWO.	signal	160	1.2E-9	2.3E-7
SP_PIR_KEYWO.	glycoprotein	178	6.2E-8	4.9E-6
UP_SEQ_FEATURE	signal peptide	160	6.1E-7	6.6E-4
SP_PIR_KEYWO.	disulfide bond	126	2.5E-6	1.1E-4
UP_SEQ_FEATURE	glycosylation site:N-linked (GlcNAc...)	167	1.7E-4	3.1E-2
UP_SEQ_FEATURE	disulfide bond	117	1.5E-3	1.6E-1
Cluster 5	**E. Score: 5.57**	**Count**	**P**	**Benjamini**
GOTERM_CC_FAT	synapse	41	3.3E-12	5.3E-10
GOTERM_CC_FAT	synapse part	32	1.9E-11	1.2E-9
SP_PIR_KEYWO.	synapse	27	1.9E-8	1.9E-6
GOTERM_CC_FAT	synaptic vesicle	15	9.3E-8	2.9E-6
SP_PIR_KEYWO.	cell junction	36	2.2E-7	1.3E-5
GOTERM_CC_FAT	cell junction	41	2.9E-7	8.4E-6
GOTERM_CC_FAT	clathrin-coated vesicle	16	3.9E-6	1.0E-4
GOTERM_CC_FAT	coated vesicle	16	3.4E-5	8.3E-4
GOTERM_CC_FAT	cytoplasmic vesicle	33	1.3E-3	1.8E-2
GOTERM_CC_FAT	vesicle	33	1.8E-3	2.1E-2
SP_PIR_KEYWO.	cytoplasmic vesicle	18	2.6E-3	4.3E-2
GOTERM_CC_FAT	cytoplasmic membrane-bounded vesicle	26	7.2E-3	6.2E-2
GOTERM_CC_FAT	membrane-bounded vesicle	26	8.6E-3	7.1E-2
Cluster 6	**E. Score: 5.05**	**Count**	**P**	**Benjamini**
GOTERM_CC_FAT	extracellular matrix	35	4.9E-9	2.6E-7
GOTERM_CC_FAT	proteinaceous extracellular matrix	34	6.5E-9	2.5E-7
SP_PIR_KEYWO.	extracellular matrix	26	7.7E-8	5.1E-6
GOTERM_CC_FAT	extracellular region part	49	1.2E-4	2.1E-3
SP_PIR_KEYWO.	Secreted	75	1.6E-4	3.8E-3
GOTERM_CC_FAT	extracellular matrix part	12	4.2E-4	6.6E-3

SP_PIR_KEYWO.	collagen	11	5.6E-4	1.1E-2
GOTERM_CC_FAT	extracellular region	81	3.7E-3	3.6E-2
Cluster 7	**E. Score: 4.22**	**Count**	**P**	**Benjamini**
GOTERM_BP_FAT	vesicle-mediated transport	37	2.1E-6	2.0E-4
GOTERM_BP_FAT	endocytosis	20	1.5E-5	1.3E-3
GOTERM_BP_FAT	membrane invagination	20	1.5E-5	1.3E-3
GOTERM_BP_FAT	membrane organization	23	1.1E-4	8.3E-3
SP_PIR_KEYWO.	Endocytosis	9	1.4E-2	1.4E-1

PUBLICATIONS

Mutations in the very low-density lipoprotein receptor *VLDLR* cause cerebellar hypoplasia and quadrupedal locomotion in humans

Tayfun Ozcelik*†‡, Nurten Akarsu§¶, Elif Uz*, Safak Caglayan*, Suleyman Gulsuner*, Onur Emre Onat*, Meliha Tan‖, and Uner Tan**

*Department of Molecular Biology and Genetics, Faculty of Science and ‡Institute of Materials Science and Nanotechnology, Bilkent University, Ankara 06800, Turkey; §Department of Medical Genetics and ¶Gene Mapping Laboratory, Department of Pediatrics, Pediatric Hematology Unit, Ihsan Dogramaci Children's Hospital, Hacettepe University Faculty of Medicine, Ankara 06100, Turkey; ‖Department of Neurology, Baskent University Medical School, Ankara 06490, Turkey; and **Faculty of Sciences, Cukurova University, Adana 01330, Turkey

Edited by Mary-Claire King, University of Washington, Seattle, WA, and approved January 16, 2008 (received for review October 22, 2007)

Quadrupedal gait in humans, also known as Unertan syndrome, is a rare phenotype associated with dysarthric speech, mental retardation, and varying degrees of cerebrocerebellar hypoplasia. Four large consanguineous kindreds from Turkey manifest this phenotype. In two families (A and D), shared homozygosity among affected relatives mapped the trait to a 1.3-Mb region of chromosome 9p24. This genomic region includes the *VLDLR* gene, which encodes the very low-density lipoprotein receptor, a component of the reelin signaling pathway involved in neuroblast migration in the cerebral cortex and cerebellum. Sequence analysis of *VLDLR* revealed nonsense mutation R257X in family A and single-nucleotide deletion c2339delT in family D. Both these mutations are predicted to lead to truncated proteins lacking transmembrane and signaling domains. In two other families (B and C), the phenotype is not linked to chromosome 9p. Our data indicate that mutations in *VLDLR* impair cerebrocerebellar function, conferring in these families a dramatic influence on gait, and that hereditary disorders associated with quadrupedal gait in humans are genetically heterogeneous.

genetics | Unertan syndrome

Obligatory bipedal locomotion and upright posture of modern humans are unique among living primates. Studies of fossil hominids have contributed significantly to modern understanding of the evolution of posture and locomotion (1–5), but little is known about the underlying molecular pathways for development of these traits. Evaluation of changes in brain activity during voluntary walking in normal subjects suggests that the cerebral cortices controlling motor functions, visual cortex, basal ganglia, and the cerebellum might be involved in bipedal locomotor activities (6). The cerebellum is particularly important for movement control and plays a critical role in balance and locomotion (7).

Neurodevelopmental disorders associated with cerebellar hypoplasias are rare and often accompanied by additional neuropathology. These clinical phenotypes vary from predominantly cerebellar syndromes to sensorimotor neuropathology, ophthalmological disturbances, involuntary movements, seizures, cognitive dysfunction, skeletal abnormalities, and cutaneous disorders, among others (8). Quadrupedal locomotion was first reported when Tan (9, 10) described a large consanguineous family exhibiting Unertan syndrome, an autosomal recessive neurodevelopmental condition with cerebellar and cortical hypoplasia accompanied by mental retardation, primitive and dysarthric speech, and, most notably, quadrupedal locomotion. Subsequent homozygosity mapping indicated that the phenotype of this family was linked to chromosome 17p (11). Thereafter, three additional families from Turkey (12–14) and another from Brazil (15) with similar phenotypes have been described, and video recordings illustrating the quadrupedal gait have been made (10–12). Here, we report that *VLDLR* is the gene responsible for the syndrome in two of these four Turkish families and report additional gene mapping studies that indicate the disorder to be highly genetically heterogeneous.

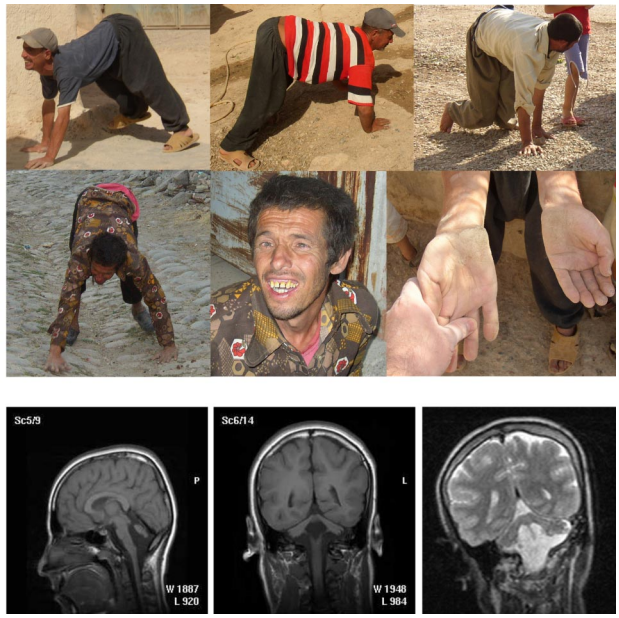


Fig. 1. Phenotypic (*A*) and cranial radiologic (*B*) presentation of quadrupedal gait in families A and D. (*A*) Affected brothers VI:20 and VI:18 and cousin VI:25 in family A (*Upper*) and the proband II:2 in family D (*Lower*) display palmigrate walking. This is different from quadrupedal knuckle-walking of the great apes (2). The hands make contact with the ground at the ulnar palm, and consequently this area is heavily callused as exemplified by VI:20. Strabismus was observed in all affected individuals. (*B*) Coronal and midsagittal MRI sections of VI:20, demonstrating vermial hypoplasia, with the inferior vermial portion being completely absent. Inferior cerebellar hypoplasia and a moderate simplification of the cerebral cortical gyri are noted. The brainstem and the pons are particularly small (*Left* and *Center*). Similar findings are observed for II:2 (*Right*).

Author contributions: T.O., N.A., and U.T. designed research; T.O., N.A., E.U., S.C., S.G., and O.E.O. performed research; T.O., N.A., E.U., S.C., S.G., and M.T. analyzed data; and T.O., N.A., and U.T. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

†To whom correspondence should be addressed. E-mail: tozcelik@fen.bilkent.edu.tr.

This article contains supporting information online at www.pnas.org/cgi/content/full/0710010105/DC1.

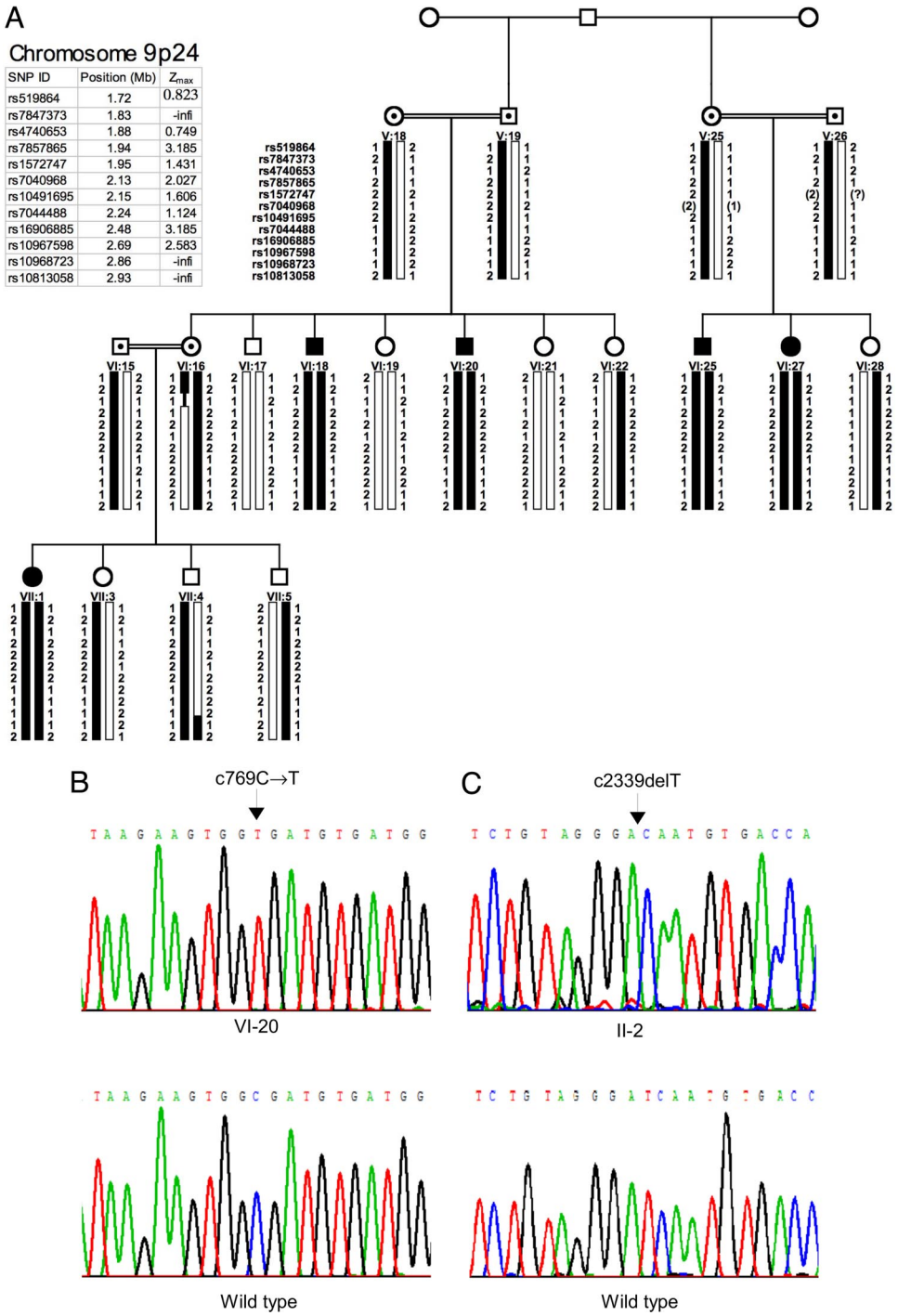


Fig. 2. Homozygosity mapping of cerebellar hypoplasia and quadrupedal locomotion to chromosome 9p24 (*A*) and identification of the *VLDLR* c769C → T mutation in family A (*B*) and of the *VLDLR* c2339delT mutation in family D (*C*). (*A*) Pedigree of family A; filled symbols represent the affected individuals. Squares indicate males, and circles indicate females. Black bars represent the haplotype coinherited with the quadrupedal phenotype in the family. Recombination events in individuals VI:16 (obligate carrier) and VII:4 (normal sibling) positioned the disease gene between markers rs7847373 and rs10968723. Physical positions and pairwise lod scores for each marker are shown on the upper left. Z_{max} represents the maximum lod score obtained at $\theta = 0.00$ cM. (*B* and *C*) Sequences of critical regions of *VLDLR* for wild-type and homozygous mutant genotypes.

Results

The proband of Family A (12) is a 37-year-old male with habitual quadrupedal gait (Fig. 1*A Upper Left* and Fig. 2*A*, VI:20). He did not make the transition to bipedality during his childhood despite the efforts of his healthy parents. He has dysarthric speech with a limited vocabulary, truncal ataxia, and profound mental retardation. He was not aware of place or of the year, month, or day. His MRI brain scan revealed inferior cerebellar and vermial hypoplasia, with the inferior vermial portion being completely absent. Whereas corpus callosum appeared normal, a moderate simplification of the cerebral cortical gyri accompanied by a particularly small brainstem and the pons was observed (Fig. 1 *B Left* and *Center*). Subsequently, we studied the proband's affected brother and cousin (Fig. 1*A Upper Center* and

Table 1. Physical, radiological, and genetic characteristics of the Turkish families in this study and of Hutterite family DES-H (27)

Characteristic	Family A	Family B	Family C	Family D	DES-H
Chromosomal location	9p24	17p	Not 9p or 17p	9p24	9p24
Gene and mutation	*VLDLR* (c769C → T)	Unknown	Unknown	*VLDLR* (c2339delT)	Deletion including *VLDLR* and *LOC401491*
Gait	Quadrupedal	Quadrupedal	Quadrupedal	Quadrupedal	Bipedal
Speech	Dysarthric	Dysarthric	Dysarthric	Dysarthric	Dysarthric
Hypotonia	Absent	Absent	Absent	Absent	Present
Barany caloric nystagmus	Normal	Cvs defect	Pvs defect	Not done	Not done
Mental retardation	Profound	Severe to profound	Profound	Profound	Moderate to profound
Ambulation	Delayed	Delayed	Delayed	Delayed	Delayed
Truncal ataxia	Severe	Severe	Severe	Severe	Severe
Lower leg reflexes	Hyperactive	Hyperactive	Hyperactive	Hyperactive	Hyperactive
Upper extremity reflexes	Vivid	Vivid	Vivid	Vivid	Vivid
Tremor	Very rare	Mild	Present	Absent	Present
Pes-planus	Present	Present	Present	Present	Present
Seizures	Very rare	Rare	Rare	Absent	Observed in 40% of cases
Strabismus	Present	Present	Present	Present	Present
Inferior cerebellum	Hypoplasia	Hypoplasia	Mild hypoplasia	Hypoplasia	Hypoplasia
Inferior vermis	Absent	Absent	Normal	Absent	Absent
Cortical gyri	Mild simplification	Mild simplification	Mild simplification	Mild simplification	Mild simplification
Corpus callosum	Normal	Reduced	Normal	Normal	Normal

Cvs, central vestibular system; Pvs, peripheral vestibular system.

Upper Right and Fig. 2*A*, VI:18 and VI:25) and other branches of the family living in nearby villages in southeastern Turkey. All affected individuals were offspring of consanguineous marriages (Fig. 2*A*). With the exception of one female (VII:1), who was an occasional biped with ataxic gait, all affected persons in family A had quadrupedal locomotion.

The proband of family D (14) is a 38-year-old male (Fig. 1*A Lower Left* and *Center*). Like all other quadrupedal individuals in these families, he did not make the transition to bipedality during his early childhood. He is profoundly retarded and exhibits dysarthric speech along with truncal ataxia. His MRI brain scan images are consistent with moderate cerebral cortical simplification and inferior cerebellar and vermial hypoplasia (Fig. 1*B Right*). The 65-year-old aunt and 63-year-old uncle of the proband are both mentally retarded and continue to walk on their wrists and feet despite their advanced ages. The family is consanguineous; all relatives were raised in neighboring villages on the western tip of the Anatolian peninsula.

All patients in these four families had significant developmental delay noted in infancy (Table 1). They sat unsupported between 9 and 18 months, and began to crawl on hands and knees or feet. Whereas normal infants make the transition to bipedal walking in a short period, the affected individuals continued to move on their palms and feet and never walked upright. All patients had severe truncal ataxia affecting their walking patterns. They can stand from a sitting position and maintain the upright position with flexed hips and knees. However, they virtually never initiate bipedal walking on their own and instead ambulate efficiently in a quadrupedal fashion. All patients had hyperactive lower leg and vivid upper extremity reflexes. Normal tone and power were observed in motor examination. All affected persons were mentally retarded to the degree that consciousness of place, time, or other experience appeared to be absent. However, no autistic features were expressed. The affected individuals all had good interpersonal skills, were friendly and curious to visitors, and followed very simple questions and commands. Additional clinical information on families A and D is provided in supporting information (SI) Table 2.

To identify the chromosomal locale of the gene or genes responsible for this phenotype, we carried out genome-wide linkage analysis and homozygosity mapping in families A–C (see *Materials and Methods* below). Although the families lived in isolated villages 200–300 km apart and reported no ancestral relationship, the rarity of the quadrupedal gait in humans led us to expect a single locus shared by affected individuals in all families. Instead, the trait mapped to three different chromosomal locales. In family A, linkage analysis and homozygosity mapping positioned the critical gene on chromosome 9p24 between rs7847373 and rs10968723 in a 1.032-Mb region (Fig. 2*A* and SI Fig. 4). In family B, the trait mapped to chromosome 17p13, confirming a previous study (11). In family C, highly negative logarithm of odds (lod) scores were obtained for both chromosomes 9p24 and 17p13 (SI Figs. 5 and 6); gene mapping in this family is ongoing. In family D, polymorphic markers from the critical intervals of chromosomes 9p24 and 17p13 were genotyped, and homozygosity was detected with markers on 9p24. Together, these results indicate that the syndrome including quadrupedal gait, dysarthric speech, mental retardation, and cerebrocerebellar hypoplasia is genetically heterogeneous.

The chromosome 9p24 region linked to the trait in families A and D includes *VLDLR*, the very low-density lipoprotein receptor. We hypothesized that a gene involved in neural development, cell positioning in brain, and cerebellar maturation could be involved in the pathogenesis of quadrupedal gait. In addition, cerebellar hypoplasia with cerebral gyral simplification was shown to be associated with a genomic deletion that includes *VLDLR* (16). We therefore considered *VLDLR* (17) to be a prime positional candidate for our phenotype and sequenced the gene in genomic DNA from probands of the four families (SI Table 3). The *VLDLR* sequence of affected members of family A was homozygous for a nonsense mutation in exon 5 (c769C → T; R257X) (Fig. 2*B*). The *VLDLR* sequence of the proband of family D was homozygous for a single-nucleotide deletion in exon 17 resulting in a stop codon (c2339delT; I780TfsX3) (Fig. 2*C*). *VLDLR* sequences excluded the possibility of compound heterozygosity in families B and C (SI Fig. 7). In families A and D, homozygosity for the *VLDLR* mutations was perfectly coinherited with quadrupedal gait (SI Figs. 8 and 9). Both mutations were absent from 100 unaffected individuals who live in the same local areas of southeastern and western Turkey as families A and D (SI Fig. 10).

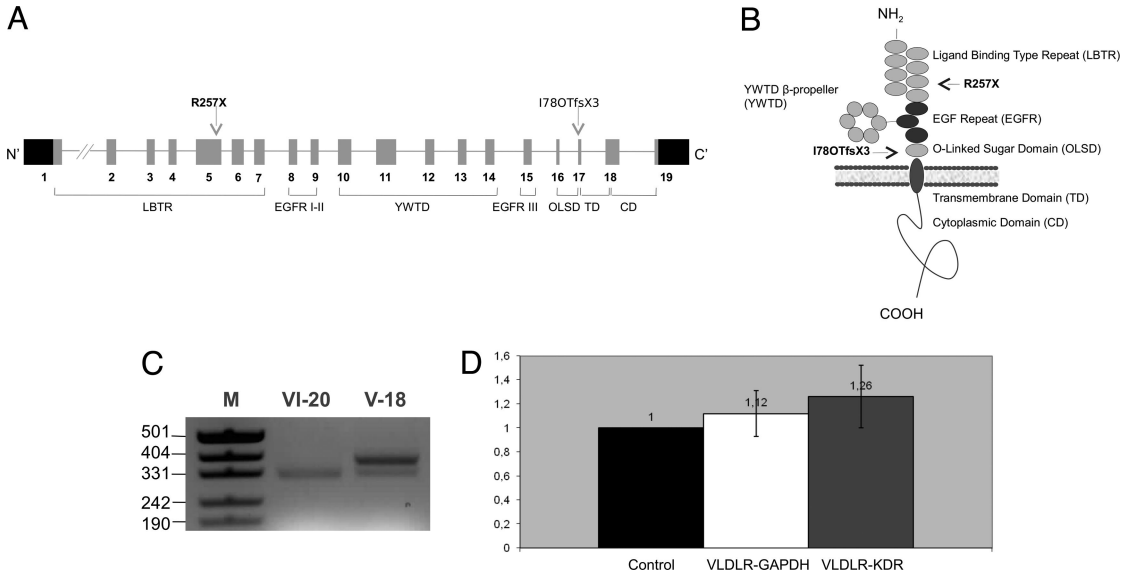


Fig. 3. Functional domains of *VLDLR* with positions of the mutations relative to the exons (*A*), domains (*B*), and the analysis of VLDLR transcript (*C* and *D*). (*A*) The gene consists of 19 exons. Arrows indicate the locations of the mutations. (*B*) VLDLR consists of ligand-binding type repeat (LBTR), epidermal growth factor repeat (EGFR) I–III, YWTD β-propeller (YWTD), *O*-linked sugar domain (OLSD), transmembrane domain (TD), and cytoplasmic domain (CD) (34) (www.expasy.org/uniprot/P98155). (*C*) Restriction-based analysis with HphI revealed the presence of only the mutant (347 bp) and both the mutant and wild type (396 and 347 bp; please note that the 49-bp fragment is not visible) VLDLR transcripts in patient VI:20 and carrier V:18 (both from family A), respectively. M is a DNA size marker. (*D*) Quantitative RT-PCR analysis of VLDLR transcript from peripheral blood samples of all probands in families A and D and controls was performed. Relative expression ratios were normalized according to the housekeeping gene *GAPDH* (glyceraldehyde-3-phosphate dehydrogenase) and the endothelial marker *KDR* (kinase insert domain receptor). ΔCt values were calculated from duplicate samples and were converted to linear scale (35). Control denotes ''VLDLR expression in controls,'' VLDLR-GAPDH denotes ''VLDLR expression in patients normalized to GAPDH,'' and finally VLDLR-KDR denotes ''VLDLR expression in patients normalized to KDR.''

VLDLR_R257X is in the ligand-binding domain, and *VLDLR_I780TfsX3* is in the *O*-linked sugar domain of the VLDLR protein (Fig. 3 *A* and *B*). Mutant VLDLR transcripts were expressed in endothelial cells from blood of affected individuals (Fig. 3*C*), and in these cells, levels of mutant and wild-type transcript expression appeared approximately equal (Fig. 3*D*; please also see *SI Text*). Because the stop codons of both mutations are located in the extracellular domain of VLDLR (Fig. 3*B*), the encoded mutant proteins could not be inserted into the membrane and could not function as receptors for reelin.

We propose *VLDLR*-associated Quadrupedal Locomotion (VLDLR-QL) or Unertan Syndrome Type 1 to describe the phenotype of families A and D.

Discussion

The identification of these *VLDLR* mutations provides molecular insight into the pathogenesis of neurodevelopmental movement disorders and expands the scope of diseases caused by mutations in components of the reelin pathway (18). Reelin is a secreted glycoprotein that regulates neuronal positioning in cortical brain structures and the migration of neurons along the radial glial fiber network by binding to lipoprotein receptors VLDLR and APOER2 and the adapter protein DAB1 (19). In the cerebellum, reelin regulates Purkinje cell alignment (20), which is necessary for the formation of a well defined cortical plate through which postmitotic granual cells migrate to form the internal granular layer (21). Homozygous mutations in the reelin gene (*RELN*) cause the Norman–Roberts type lissencephaly syndrome, associated with severe abnormalities of the cerebellum, hippocampus, and brainstem (OMIM 257320) (22). Mutation of *Reln* in the mouse results in the *reeler* phenotype and disrupts neuronal migration in several brain regions and gives rise to functional deficits such as ataxic gait and trembling (23). In contrast, mice deficient for *Vldlr* appear neurologically normal (24), but the cerebellae of these mice are small, with reduced foliation and heterotopic Purkinje cells (17).

In humans, homozygosity for either of two *VLDLR* truncating mutations leads to cerebrocerebellar hypoplasia, specifically vermial hypoplasia, accompanied by mental retardation, dysarthric speech, and quadrupedal gait. In the Hutterite population of North America, homozygosity for a 199-kb deletion encompassing the *VLDLR* gene leads to a form of Disequilibrium Syndrome (DES-H, OMIM 224050), characterized by nonprogressive cerebellar hypoplasia with moderate-to-profound mental retardation, cerebral gyral simplification, truncal ataxia, and delayed ambulation (16). The designation Disequilibrium Syndrome was originally given to cerebral palsy characterized by a variety of congenital abnormalities, including mental retardation, disturbed equilibrium, severely retarded motor development, muscular hypotonia, and perceptual abnormalities (25, 26). Neither DES-H nor other disequilibrium syndromes have been reported to include quadrupedal gait. The movement of most DES-H patients was so severely affected that independent walking was not possible. Those who could walk had a wide-based, nontandem gait (27).

The neurological phenotypes in the Turkish families and in the Hutterite families appear similar, with the most striking difference being the consistent adoption of efficient quadrupedal locomotion by the affected Turkish individuals (Table 1). In our view, the movement disorder described for the Hutterite patients may be a more profound deficit, with the patients perhaps lacking the motor skills for quadrupedal locomotion. The 199-kb deletion in DES-H encompasses the entire *VLDLR* gene and part of a hypothetical gene. *LOC401491*, the hypothetical gene, is an apparently noncoding RNA that shares a CpG island and likely promoter with *VLDLR*, and is represented by multiple alternative transcripts expressed in brain. It has been suggested that the DES-H phenotype could be the result of deletion of *VLDLR* or both *VLDLR* and the neighboring gene (16).

It has been suggested that in the Turkish families, lack of access to proper medical care exacerbated the effects of cerebellar hypoplasia, leading to quadrupedality. Although it may be true that family B lacked proper medical care, families A and D had consistent access to medical attention, and both families actively sought a correction of quadrupedal locomotion in their affected children. An unaffected individual in family A is a physician who was actively involved in the medical interventions. In family D, the proband's mother sought a definitive diagnosis and correction of the proband's quadrupedal locomotion from private medical practices and from two major academic medical centers. The parents in family A discouraged quadrupedal walking of their affected children, but without success. We conclude that social factors were highly unlikely to contribute to the quadrupedal locomotion of the affected individuals.

In conclusion, we suggest that *VLDLR*-deficiency in the brain at a key stage of development leads to abnormal formation of the neural structures that are critical for gait. Given the heterogeneity of causes of quadrupedal gait, identification of the genes in families B and C promises to offer insights into neurodevelopmental mechanisms mediating gait in humans.

Materials and Methods

Study Subjects. Parents of patients and other unaffected individuals gave consent to the study by signing the informed consent forms prepared according to the guidelines of the Ministry of Health in Turkey. The Ethics Committees of Baskent and Cukurova Universities approved the study (decision KA07/47, 02.04.2007 and 21/3, 08.11.2005, respectively).

Genome-Wide Linkage Analysis. Linkage analysis was performed by SNP genotyping with the commercial release of the GeneChip 250K (NspI digest) or 10K Affymetrix arrays as described (28). In addition, genotype data were analyzed by hand to identify regions of homozygosity. The parametric component of the Merlin package v1.01 was used for the multipoint linkage analysis assuming autosomal recessive mode of inheritance with full penetrance (29, 30). The analysis was carried out along a grid of locations equally spaced at 1 cM. Haplotype analysis was performed on chromosomal regions with positive lod scores (Fig. 2*A* and SI Figs. 4–6). Pairwise linkage was analyzed by using the MLINK component of the LINKAGE program (FASTLINK, version 3) (31–33). Markers D17S1298 (3.51 Mb) and D9S1779 (0.4 Mb), D9S1871 (3.7 Mb) were used to test for homozygosity to chromosomes 17p13 and 9p24, respectively.

Mutation Search. Sequencing primers were designed for each *VLDLR* exon by using Primer3, BLAST, and the sequence of NC_000009. DNA from all of the probands was sequenced in both directions by using standard methods. The mutations in exons 5 (c769C → T) and 17 (c2339delT) were detected in all affected (homozygous) and carrier (heterozygous) individuals of families A and D, respectively. The c769C → T mutation creates a restriction site for the enzyme HphI (5′-GGTGA(N)8 ↓ 3′), and the c2339delT mutation abolishes a restriction site for the enzyme MboI (5′-G↓ ATC-3′). Assays using these restriction enzymes were developed to test for the mutations in all four families and in 200 healthy controls from the Turkish population. Restriction based mutation and quantitative RT-PCR analyses of *VLDLR* transcript in patients and controls was also performed (please see *SI Text* relating to Fig. 3 *C* and *D*).

ACKNOWLEDGMENTS. We thank the patients and family members for their participation in this study, E. Tuncbilek and M. Alikasifoglu for providing the microarray facility at Hacettepe University, Iclal Ozcelik for help in writing the manuscript, and Mehmet Ozturk for support. This work was supported by the Scientific and Technological Research Council of Turkey Grant TUBITAK-SBAG 3334, International Centre for Genetic Engineering and Biotechnology Grant ICGEB-CRP/TUR04-01 (to T.O.), and by Baskent University Research Fund KA 07/47 and TUBITAK-SBAG-HD-230 (to M.T.).

1. Spoor F, Wood B, Zonneveld F (1994) *Nature* 369:645–648.
2. Richmond BG, Strait DS (2000) *Nature* 404:382–385.
3. Bramble DM, Lieberman DE (2004) *Nature* 432:345–352.
4. Alemseged Z, Spoor F, Kimbel WH, Bobe R, Geraads D, Reed D, Wynn JG (2006) *Nature* 443:296–301.
5. Wood B (2006) *Nature* 443:278–281.
6. Fukuyama H, Ouchi Y, Matsuzaki S, Nagahama Y, Yamauchi H, Ogawa M, Kimura J, Shibasaki H (1997) *Neurosci Lett* 228:183–186.
7. Morton SM, Bastian AJ (2007) *Cerebellum* 6:79–86.
8. Fogel BL, Perlman S (2007) *Lancet Neurol* 6:245–257.
9. Tan U (2005) *Neuroquantology* 4:250–255.
10. Tan U (2006) *Int J Neurosci* 116:361–369.
11. Turkmen S, Demirhan O, Hoffmann K, Diers A, Zimmer C, Sperling K, Mundlos S (2006) *J Med Genet* 43:461–464.
12. Tan U, Karaca S, Tan M, Yilmaz B, Bagci NK, Ozkur A, Pence S (2008) *Int J Neurosci* 118:1–25.
13. Tan U (2006) *Int J Neurosci* 116:763–774.
14. Tan U (2008) *In J Neurosci* 118:211–225.
15. Garcias GL, Roth MG (2007) *Int J Neurosci* 117:927–933.
16. Boycott KM, Flavelle S, Bureau A, Glass HC, Fujiwara TM, Wirrell E, Davey K, Chudley AE, Scott JN, McLeod DR, Parboosingh JS (2005) *Am J Hum Genet* 77:477–483.
17. Trommsdorff M, Gotthardt M, Hiesberger T, Shelton J, Stockinger W, Nimpf J, Hammer RE, Richardson JA, Herz J (1999) *Cell* 97:689–701.
18. Tissir F, Goffinet AM (2003) *Nat Rev Neurosci* 4:496–505.
19. Hiesberger T, Trommsdorff M, Howell BW, Goffinet A, Mumby MC, Cooper JA, Herz J (1999) *Neuron* 24:481–489.
20. Miyata T, Nakajima K, Mikoshiba K, Ogawa M (1997) *J Neurosci* 17:3599–3609.
21. Wechsler-Reya RJ, Scott MP (1999) *Neuron* 22:103–114.
22. Hong SE, Shugart YY, Huang DT, Shahwan SA, Grant PE, Hourihane JO, Martin ND, Walsh CA (2000) *Nat Genet* 26:93–96.
23. D'Arcangelo G, Miao GG, Chen SC, Soares HD, Morgan JI, Curran T (1995) *Nature* 374:719–723.
24. Frykman PK, Brown MS, Yamamoto T, Goldstein JL, Herz J (1995) *Proc Natl Acad Sci* 92:8453–8457.
25. Hagberg B, Sanner G, Steen M (1972) *Acta Paediat Scand* 61(Suppl. 226):1–63.
26. Sanner G (1973) *Neuropaediatrie* 4:403–413.
27. Glass HC, Boycott KM, Adams C, Barlow K, Scott JN, Chudley AE, Fujiwara TM, Morgan K, Wirrell E, McLeod DR (2005) *Dev Med Child Neurol* 47:691–695.
28. Matsuzaki H, Dong S, Loi H, Di X, Liu G, Hubbell E, Law J, Berntsen T, Chadha M, Hui H, *et al.* (2004) *Nat Methods* 1:109–111.
29. Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) *Nat Genet* 30:97–101.
30. Abecasis GR, Wigginton JE (2005) *Am J Hum Genet* 77:754–767.
31. Lathrop GM, Lalouel JM (1984) *Am J Hum Genet* 36:460–465.
32. Cottingham RW, Jr, Idury RM, Schaffer AA (1993) *Am J Hum Genet* 53:252–263.
33. Schaffer AA, Gupta SK, Shriram K, Cottingham RW, Jr (1994) *Hum Hered* 44:225–237.
34. Herz J, Bock HH (2002) *Annu Rev Biochem* 71:405–434.
35. Pfaffi MW (2004) in *A-Z of Quantitative PCR*, ed Bustin S (International University Line, La Jolla, CA), pp 89–120.

LETTER

Reply to Herz *et al*. and Humphrey *et al*.: Genetic heterogeneity of cerebellar hypoplasia with quadrupedal locomotion

Mutations in the very low-density lipoprotein receptor VLDLR are responsible for cerebellar hypoplasia with quadrupedal gait (1). The most likely mechanism leading to this phenotype is that VLDLR deficiency in the brain at a key stage of development precludes the normal formation of neural structures critical for gait. Quadrupedal gait is an integral part of VLDLR-associated cerebellar hypoplasia syndrome in these families (1, 2). It is not necessary to invoke an "epiphenomenon" or "unfavorable environmental conditions" to explain the phenotype (3), but rather simply considering clinical heterogeneity in the context of genomic understanding of complex traits is sufficient.

Disequilibrium syndrome was first described by the Swedish neuropediatrician Bengt Hagberg and colleagues (4) as a form of cerebral palsy characterized by a variety of congenital abnormalities. Subsequently, Schurig *et al*. (5) described, in the North American Hutterite population, inherited cerebellar disorder with mental retardation, the genetic basis of which proved to be homozygous deletion of the VLDLR gene and the adjacent noncoding LOC401491 sequence (6). Based on the phenotypic similarities of the Swedish and Hutterite patients, the acronym DES-H [disequilibrium syndrome-Hutterites, Online Mendelian Inheritance in Man (OMIM) accession no. 224050] was adopted for this syndrome (6).

Our results (1) and those of others (7) extend these findings to different VLDLR mutations leading to cerebellar hypoplasia and related disequilibrium features, including in some families bipedal gait (5, 6), in other families quadrupedal gait (1, 8), and in another family "gait ataxia" (7). Additional kindreds with disequilibrium syndrome and quadrupedal gait have been described in Brazil (9) and Iraq (10). It will be interesting to know whether mutations responsible for the phenotype in these families lie in the VLDLR gene or in one of the other loci linked to this genetically heterogeneous phenotype (1).

The comments of Humphrey *et al*. (11) address three fundamental features of genomic analysis of human traits: allelic heterogeneity, genotype–phenotype correlations, and variable expression.

Allelic heterogeneity—the expression of the same phenotype due to different mutations in a gene—is characteristic of virtually all human genetic disease. For example, homozygosity for any of >300 different mutations in the LDL receptor leads to hypercholesterolemia. It was to be expected, therefore, that in different families different mutations in VLDLR would lead to a phenotype comprising cerebellar hypoplasia with quadrupedal gait. It would not be expected that quadrupedalism would be present only in the presence of one "specific mutation."

The converse observation, of a correlation between genotype and phenotype, is also characteristic of inherited human disease. Different mutations in the same gene frequently lead to different clinical phenotypes. Contrary to the statement of Humphrey *et al*. (11), the Hutterite families in North America and families A and D in Turkey do not carry "the same homozygous mutation." The Hutterite mutation is a complete genomic deletion of VLDLR; the mutations in Turkish families A and D are, respectively, a nonsense mutation and a single-base-pair deletion leading to a frame shift in VLDLR. It is not surprising, therefore, that features of the cerebellar hypoplasia syndrome, including presence or absence of quadrupedal walking, differ among families with different mutations in the gene.

Third, variable expression of a phenotype is frequently observed even among persons with the same mutation in a critical gene. Variable expression may be due to differences in genetic background of the individual, to differences in environmental exposures, or to chance. Among affected individuals in families A and D, none displays exclusively bipedal locomotion; two affected individuals can walk bipedally for short distances but prefer quadrupedal locomotion (1, 8).

Finally, the use of a walking frame to assist bipedalism in affected individuals (12) does not demonstrate that the cause of quadrupedalism was "local cultural environment." Wearing eyeglasses assists persons with myopia. Should we then conclude that near-sightedness is caused by "local cultural environment"?

Some descriptions by the press of Turkish families with cerebellar hypoplasia and quadrupedal gait have portrayed the affected individuals as doomed to quadrupedal gait by the religious beliefs of their parents (13). We hope that future descriptions of these families will conform to standards reflected in recent genomic analyses of their disorder.

Tayfun Ozcelik*[*†‡], Nurten Akarsu[§¶], Elif Uz*, Safak Caglayan*, Suleyman Gulsuner*, Onur Emre Onat*, Meliha Tan[‖], and Uner Tan*

**Department of Molecular Biology and Genetics, Faculty of Science, and †Institute of Materials Science and Nanotechnology, Bilkent University, Ankara 06800, Turkey; §Department of Medical Genetics and ¶Gene Mapping Laboratory, Department of Pediatrics, Pediatric Hematology Unit, Ihsan Dogramaci Children's Hospital, Hacettepe University Faculty of Medicine, Ankara 06100, Turkey; ‖Department of Neurology, Baskent University Medical School, Ankara 06490, Turkey; and **Faculty of Sciences, Cukurova University, Adana 01330, Turkey*

1. Ozcelik T, *et al.* (2008) Mutations in the very low-density lipoprotein receptor VLDLR cause cerebellar hypoplasia and quadrupedal locomotion in humans. *Proc Natl Acad Sci USA* 105:4232–4236.
2. Tan U (2005) A new theory on the evolution of human mind. Unertan syndrome: Quadrupedality, primitive language, and severe mental retardation. *NeuroQuantology* 4:250–255.
3. Herz J, Boycott KM, Parboosingh JS (2008) "Devolution" of bipedality. *Proc Natl Acad Sci USA* 105:E25.
4. Hagberg B, Scanner G, Steen M (1972) The dysequilibrium syndrome in cerebral palsy. Clinical aspects of treatment. *Acta Paediatr Scand* 61(Suppl 226):1–63.
5. Schurig V, Van Orman A, Bowen P (1981) Nonprogressive cerebellar disorder with mental retardation and autosomal recessive inheritance in Hutterites. *Am J Med Genet* 9:43–53.

www.pnas.org/cgi/doi/10.1073/pnas.0804078105

6. Boycott KM, *et al.* (2005) Homozygous deletion of the very low density lipoprotein receptor gene causes autosomal recessive cerebellar hypoplasia with cerebral gyral simplification. *Am J Hum Genet* 77:477–483.
7. Moheb LA, *et al.* (2008) Identification of a nonsense mutation in the very low-density lipoprotein receptor gene (VLDLR) in an Iranian family with dysequilibrium syndrome. *Eur J Hum Genet* 16:270–273.
8. Turkmen S, *et al.* (March 26, 2008) Cerebellar hypoplasia, with quadrupedal locomotion, caused by mutations in the very low-density lipoprotein receptor gene. *Eur J Hum Genet*, 10.1038/ejhg.2008.73.
9. Garcias GL, Roth MG (2007) A Brazilian family with quadrupedal gait, severe mental retardation, coarse facial characteristics, and hirsutism. *Int J Neurosci* 117: 927–933.
10. Fletcher M (October 17, 2007) Life on all fours. *Times Online*. Available at www.timesonline.co.uk/tol/life_and_style/health/article2671426.ece.
11. Humphrey N, Mundlos S, Turkmen S (2008) Genes and quadrupedal locomotion in humans. *Proc Natl Acad Sci USA* 105:E26.
12. Harrison J, Holt S (2006) *The Family That Walks on All Fours* (BBC, London).
13. Ahuja A (2007) We're all made with quadrupedal walking ability. *Times Online*. Available at http://women.timesonline.co.uk/tol/life_and_style/women/the_way_we_live/article2671044.ece.

Author contributions: T.O., N.A., E.U., S.C., S.G., O.E.O., M.T., and U.T. wrote the paper.

The authors declare no conflict of interest.

‡To whom correspondence should be addressed. E-mail: tozcelik@bilkent.edu.tr.

Homozygosity mapping and targeted genomic sequencing reveal the gene responsible for cerebellar hypoplasia and quadrupedal locomotion in a consanguineous kindred

Suleyman Gulsuner, Ayse Begum Tekinay, Katja Doerschner, et al.

Genome Res. published online September 1, 2011
Access the most recent version at doi:10.1101/gr.126110.111

Research

Homozygosity mapping and targeted genomic sequencing reveal the gene responsible for cerebellar hypoplasia and quadrupedal locomotion in a consanguineous kindred

Suleyman Gulsuner,[1] Ayse Begum Tekinay,[2] Katja Doerschner,[3,4] Huseyin Boyaci,[3,4] Kaya Bilguvar,[5,6,7] Hilal Unal,[2] Aslihan Ors,[4] O. Emre Onat,[1] Ergin Atalar,[4,8] A. Nazli Basak,[9] Haluk Topaloglu,[10] Tulay Kansu,[11] Meliha Tan,[12] Uner Tan,[13] Murat Gunel,[5,6,7] and Tayfun Ozcelik[1,2,14]

[1]*Department of Molecular Biology and Genetics, Faculty of Science, Bilkent University, Ankara 06800, Turkey;* [2]*Institute of Materials Science and Nanotechnology, Bilkent University, Ankara 06800, Turkey;* [3]*Department of Psychology, Faculty of Economics, Administrative and Social Sciences, Bilkent University, Ankara 06800, Turkey;* [4]*National Research Center for Magnetic Resonance, Bilkent University, Ankara 06800 Turkey;* [5]*Department of Neurosurgery, Yale University School of Medicine, New Haven, Connecticut 06510, USA;* [6]*Department of Neurobiology, Yale University School of Medicine, New Haven, Connecticut 06510, USA;* [7]*Department of Genetics, Center for Human Genetics and Genomics and Program on Neurogenetics, Yale University School of Medicine, New Haven, Connecticut 06510, USA;* [8]*Department of Electrical and Electronics Engineering, Faculty of Engineering, Bilkent University, Ankara 06800, Turkey;* [9]*NDAL Laboratory, School of Arts and Sciences, Bogazici University, Istanbul 34342, Turkey;* [10]*Department of Pediatric Neurology, Ihsan Dogramaci Children's Hospital, Ankara 06100, Turkey;* [11]*Department of Neurology, Hacettepe University Faculty of Medicine, Ankara 06100, Turkey;* [12]*Department of Neurology, Baskent University Faculty of Medicine, Ankara 06490, Turkey;* [13]*Department of Physiology, Cukurova University Faculty of Medicine, Adana 01330, Turkey*

The biological basis for the development of the cerebro-cerebellar structures required for posture and gait in humans is poorly understood. We investigated a large consanguineous family from Turkey exhibiting an extremely rare phenotype associated with quadrupedal locomotion, mental retardation, and cerebro-cerebellar hypoplasia, linked to a 7.1-Mb region of homozygosity on chromosome 17p13.1–13.3. Diffusion weighted imaging and fiber tractography of the patients' brains revealed morphological abnormalities in the cerebellum and corpus callosum, in particular atrophy of superior, middle, and inferior peduncles of the cerebellum. Structural magnetic resonance imaging showed additional morphometric abnormalities in several cortical areas, including the corpus callosum, precentral gyrus, and Brodmann areas BA6, BA44, and BA45. Targeted sequencing of the entire homozygous region in three affected individuals and two obligate carriers uncovered a private missense mutation, WDR81 p.P856L, which cosegregated with the condition in the extended family. The mutation lies in a highly conserved region of WDR81, flanked by an N-terminal BEACH domain and C-terminal WD40 beta-propeller domains. WDR81 is predicted to be a transmembrane protein. It is highly expressed in the cerebellum and corpus callosum, in particular in the Purkinje cell layer of the cerebellum. *WDR81* represents the third gene, after *VLDLR* and *CA8*, implicated in quadrupedal locomotion in humans.

[Supplemental material is available for this article.]

Developmental abnormalities of the cerebellum are a rare and genetically heterogeneous group of disorders characterized by loss of balance and coordination. Identification of the genes responsible for these disorders provides mechanistic insights into the regulation of neuronal development, differentiation, morphogenesis, migration, and organization (Fogel and Perlman 2007). These genes can be identified by exploiting targeted genomic sequencing in combination with linkage analysis and homozygosity mapping (Ropers 2007; Bilguvar et al. 2010). We applied this approach to the analysis of cerebellar hypoplasia and quadrupedal locomotion in an extended consanguineous family from southern Turkey.

Multiple families have been reported with cerebellar ataxia, mental retardation, and disequilibrium syndrome (CAMRQ) (Tan 2006; Turkmen et al. 2006, 2009; Moheb et al. 2008; Ozcelik et al. 2008; Kolb et al. 2010). All the reported CAMRQ families are consanguineous with recessive inheritance of their condition. Clinical characteristics vary slightly among the families. In four families from Turkey and Iran, the condition is due to homozygosity for mutations in the *VLDLR* gene encoding the very low density lipoprotein receptor (CAMRQ1 [MIM 224050]). Each of these four families harbors a different *VLDLR* mutation. In a fifth family, from Iraq, the condition is due to homozygosity for a missense mutation in the *CA8* gene encoding carbonic anhydrase VIII (CAMRQ3 [MIM 613227]). In Family B, the first family described in the literature

[14]Corresponding author.
E-mail tozcelik@bilkent.edu.tr.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.126110.111.

and also referred to as Uner Tan syndrome (Tan 2006), homozygosity mapping revealed a 7.1-Mb interval on chromosome 17p13, containing 192 genes and at least 20 pseudogenes, that segregates with the disease (CAMRQ2 [MIM 610185]) (Turkmen et al. 2006; Ozcelik et al. 2008). In order to identify the mutation responsible for CAMRQ2 in Family B, we targeted and fully sequenced the 7.1-Mb genomic interval and evaluated all variation in the region.

Results

Description of the affected family

Family B came to medical attention because of the unusual form of locomotion in five of the 19 siblings. A detailed clinical description, including video recordings and genetic mapping, was published elsewhere (Tan 2006; Turkmen et al. 2006; Ozcelik et al. 2008). Pedigree analysis suggested autosomal recessive inheritance. Linkage analysis and homozygosity mapping revealed a single locus on chromosome 17p between D17S1866 and D17S960. Illumina 300 Duo v2 BeadChip SNP genotype data of two of the affected individuals (05-984 and 05-987) revealed a single 6.8-Mb homozygous stretch between markers rs4617924–rs7338 (chr17: 114,669–6,917,703) and confirmed that chromosome 17p is the only region of interest (Supplemental Fig. 1).

The phenotype was further characterized by magnetic resonance imaging (MRI) and morphometric analyses (Fig. 1). The most dramatic morphological differences were significant reductions in volume in the cerebellum and corpus callosum of the patient's brain (Fig. 1A). Both the cortex and the white matter of the cerebellum were significantly smaller in the patients. In contrast, the volume occupied by the caudate nucleus was significantly larger. Significant structural differences were also detected in the motor areas precentral gyrus and BA6 (increased mean curvature and gray matter volume) and motor speech areas pars opercularis and pars triangularis (increased cortical thickness and mean curvature) (Fig. 1B). A detailed account of the morphometric analyses is presented in Supplemental Figure 2 and Supplemental Table 1. Diffusion tensor imaging (DTI) and fiber tractography revealed moderate to high atrophy in superior, middle, and inferior cerebellar peduncles (Supplemental Fig. 3).

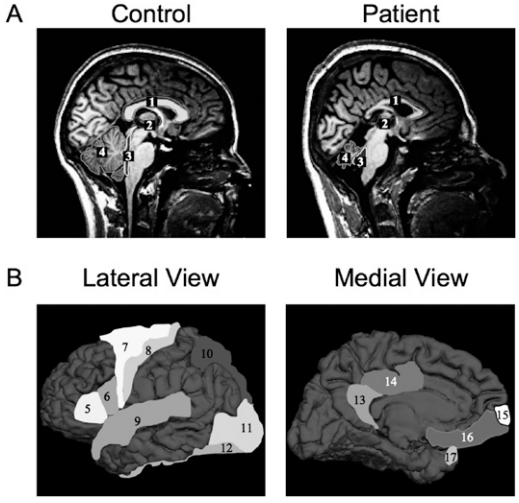


Figure 1. MRI-based morphological analysis of brain from affected and unaffected individuals. (*A*) Midsagittal MRI scans of a healthy control individual (*left*) and affected relative from Family B (*right*). The highlighted regions show areas where volumetric differences are readily visible: corpus callosum (1), third ventricle (2), fourth ventricle (3), and cerebellum (4). (*B*) Cortical regions with significant differences in morphometric parameters are displayed on a reference cortex, from lateral and medial view: BA45 (5), BA44 (6), BA6 (7), precentral (8), superior temporal (9), superior parietal (10), lateral occipital (11), fusiform (12), isthmus cingulated (13), posterior cingulated (14), frontal pole (15), medial orbitofrontal (16), and temporal pole (17). Additional details are provided in Supplemental Figure 2 and Supplemental Table 1.

Targeted next-generation sequencing of the critical region

The critical region at chr17: 82,514–7,257,922 (hg19) was captured by NimbleGen 385K microarrays and sequenced with 454 Life Sciences (Roche) GS FLX in DNA of two of the affected individuals (05-985, 05-987) and two of the unaffected obligate carrier parents (05-981 father, 05-982 mother). An average of ~400 Mb, yielding 46.3× haploid coverage, was sequenced from the captured DNA of each individual. An average of 79% of all reads from each sample mapped to the target region, representing 1275-fold to 2247-fold enrichment (Supplemental Table 2). On average, 99.4% of all targeted bases were covered by at least four reads (Supplemental Table 3).

In a parallel experiment, the same region from the DNA of another affected sibling (05-984) was captured with NimbleGen HD2 2.1M sequence capture microarrays and sequenced on an Illumina Genome Analyzer IIx. The captured region was enriched 123-fold, with 2.98 billion bases and 40.3 million reads obtained and 28% of reads mapped to the targeted region; 99.6% of targeted bases were covered by at least four reads. Combined sequence data for the three affected siblings yielded at least a fourfold coverage of 99.78% of all coding base pairs, 95.32% of intronic and UTR base pairs, and 91.36% of intergenic base pairs. The remaining 0.22% of coding regions with less than fourfold coverage was analyzed by Sanger sequencing (Supplemental Table 4).

With the 454 GS FLX platform, a total of 18,410 different variants were detected at high confidence (defined as in Hedges et al. 2009) in at least one sample (Supplemental Table 2). No additional functional variants were detected with the Illumina sequencing platform. Comparison of the sequence data from both platforms with Illumina 300 Duo v2 SNP genotype data indicated that the alleles were detected with sensitivity and specificity >99%. Heterozygous SNPs detected at the borders of the homozygous blocks of the affected individuals narrowed the region of homozygosity to 6.74 Mb (Supplemental Table 5). The Mendelian error rate, an indicator of call errors (Hedges et al. 2009), was calculated as 0.3%.

Of the 18,410 high-confidence variants, 17,281 were reported by dbSNP. For each nonsynonymous SNP compatible with the Mendelian transmission of the disease allele, the frequencies of homozygotes for each allele were accessed from public databases. With one exception, homozygosity at both alleles had been reported in control populations. The one exception, rs55916885, was at a nonconserved site and was predicted as tolerated by SIFT (Ng and Henikoff 2001) and Polyphen-2 (Sunyaev et al. 2001). Based on these observations, all previously reported nonsynonymous variants were excluded (Supplemental Table 6).

Of the 18,410 high-confidence variants, 1119 variants were both novel vis-á-vis dbSNP132 and present in both the affected siblings and their obligate carrier parents. These 1119 novel shared variants were classified by genomic context: coding sequence or flanking splice junctions (n = 20), 5′ UTR or 3′ UTR (n = 15), intronic (n = 689), or intergenic (n = 395). The 20 variants in the

coding sequence or flanking splice junctions were genotyped in the family to evaluate cosegregation with the phenotype (Supplemental Table 7). Genotypes of three missense variants were consistent with the recessive inheritance of the disease allele in Family B: WDR81 p.P856L, MYBBP1A p.R671W, and ZNF594 p.L639F (Table 1). Of the 15 5′/3′ UTR variants, five cosegregated with the disease phenotype. Therefore, they were carried to a more detailed analysis, including evaluation of the protein interactions. None was found to interact with previously identified genes with cerebellar phenotypes, including CAMRQ-associated *VLDLR* and *CA8* (Supplemental Table 8).

Identification of disease causing variant

MYBBP1A p.R671W could be excluded as the causal mutation for the disorder of Family B based on the genotypes of controls (Supplemental Table 9). In 214 unrelated healthy controls (428 chromosomes), 50 of whom were sampled from the same region of Turkey as Family B, 13 individuals were heterozygous for MYBBP1A p.R671W. This carrier frequency yields an allele frequency of 0.016 and an expected frequency of homozygotes of about one in 4000, far higher than the frequency of CAMRQ2, which occurs in only one extended family. In a second, independent series of 400 individuals of various European and Middle Eastern ancestries, *MYBBP1A* was fully sequenced in the context of whole-exome sequencing. Of these 400 individuals, two were homozygous for MYBBP1A p.R671W. Neither of these two homozygotes had any signs consistent with CAMRQ2. MYBBP1A p.R671W was therefore excluded as the allele responsible for the disorder of Family B.

ZNF594 p.L639F could be excluded as the causal mutation for the disorder based on conservation considerations. Residue 639 of ZNF594 is not well conserved: Two of 16 species sequenced have phenylalanine (F) at the orthologous site, strongly suggesting that phenylalanine at this site would also not be damaging in humans. A negative GERP score (−0.665) for the mutated nucleotide indicates that this site is probably evolving neutrally (Davydov et al. 2010). The variant is predicted as "benign" (PSIC score difference, 0.301) by PolyPhen (Sunyaev et al. 2001) and "damaging low confidence" (SIFT score, 0.04) by SIFT (Supplemental Table 10; Ng and Henikoff 2001). In addition, the human *ZNF594* gene harbors polymorphic nonsense mutations at sites near the missense at L639F. ZNF594 p.E684X appeared in four of 118 Yoruban controls (rs114754534; allele frequency, 0.034), and ZNF594 p.Q681X appeared in one of 120 CEU controls (rs116878311; allele frequency, 0.0083) in the HapMap series.

In contrast, WDR81 p.P856L (Fig. 2A,B; Supplemental Fig. 4) is both rare and alters a highly conserved site. This missense did not appear in any of the 549 individuals of the control series. WDR81 is a highly conserved protein throughout vertebrates, with no polymorphic stops in any sequenced species. In particular, proline at residue 856 is completely conserved in all known sequences of the WDR81 protein (Fig. 2C).

The extended genealogy of Family B revealed consanguinity in several branches of the kindred (Fig. 2D), whose ancestors have migrated from a village on the Syrian side of the border with Hatay, Turkey, in the early 1950s. Approximately 240 individuals spanning seven generation could be ascertained. WDR81 p.P856L was genotyped in 177 members of the kindred spanning five generations. A single union of heterozygous carriers, 05-981 × 05-982, was observed whose children include the affected individuals of this study. None of the 172 unaffected individuals in the kindred is homozygous for WDR81 p.P856L. Genetic counseling is in progress for the 27 members of the family who are heterozygous carriers of the mutation. The status of *WDR81* was evaluated in two different cohorts of the patients with neurodevelopmental/cerebellar phenotypes for whom the underlying genetic cause is still unknown. The first cohort consisted of 750 patients with structural cortical malformations or degenerative neurological disorders. By using the whole-genome genotyping data based on Illumina Human 370 Duo or 610K Quad BeadChips, we did not identify any patient with a cerebellar phenotype or ataxia phenotype to harbor a homozygous interval (≥2.5 cM) surrounding the *WDR81* locus. Exome sequencing of the same group did not reveal any mutations, including compound heterozygous substitutions. In the second cohort of 58 probands, 12 had cerebellar hypoplasia with or without quadrupedal gait. No additional mutations in *WDR81* were identified by Sanger sequencing of the entire coding regions.

Characterization of *WDR81*

WDR81 p.P856L at chr17: 1,630,820 (hg19) lies in exon 1 of *WDR81* isoform 1 (ENST00000409644, NM_001163809.1, NP_001157281.1), the longest isoform of *WDR81*, containing 10 exons and encoding 1941 amino acids (Fig. 2A). Proline at this site was present in all species analyzed (Fig. 2C), including the most distantly related sequenced ortholog, the *Tetraodon nigroviridis* WDR81 protein, which is 47.8% identical and 57.2% similar and has a distance score of 0.76 compared with the human protein. WDR81 p.P856L was predicted to be "damaging" (SIFT score, 0) by SIFT (Ng and Henikoff 2001), "probably damaging" (PSIC score difference, 2.724) by PolyPhen (Sunyaev et al. 2001), and "under evolutionary constraint" (GERP score, 5.68) by GERP (Davydov et al. 2010).

The function of WDR81 is unknown, but clues can be derived from its structure. The conserved region of WDR81 that includes P856 is flanked on the N-terminal side by a BEACH (Beige and Chediak-Higashi) domain at amino acids 352–607. BEACH proteins

Table 1. Missense variants co-inherited with cerebellar hypoplasia and quadrupedal locomotion in Family B[a]

					No. and percentage of variant reads			
Gene	Position (hg19)	Ref	Var	Effect	05-981[b]	05-982[b]	05-985[c]	05-987[c]
WDR81	chr17: 1,630,820	C	T	P856L	41 (51%)	33 (52%)	40 (97%)	53 (100%)
MYBBP1A	chr17: 4,448,967	G	A	R671W	29 (52%)	21 (48%)	32 (97%)	29 (100%)
ZNF594	chr17: 5,085,637	G	A	L639F	39 (54%)	50 (56%)	38 (97%)	37 (100%)

Ref indicates reference nucleotide; Var, variant nucleotide.
[a]Coding regions, consensus splices-sites, and RNA genes.
[b]Carrier.
[c]Affected individual.

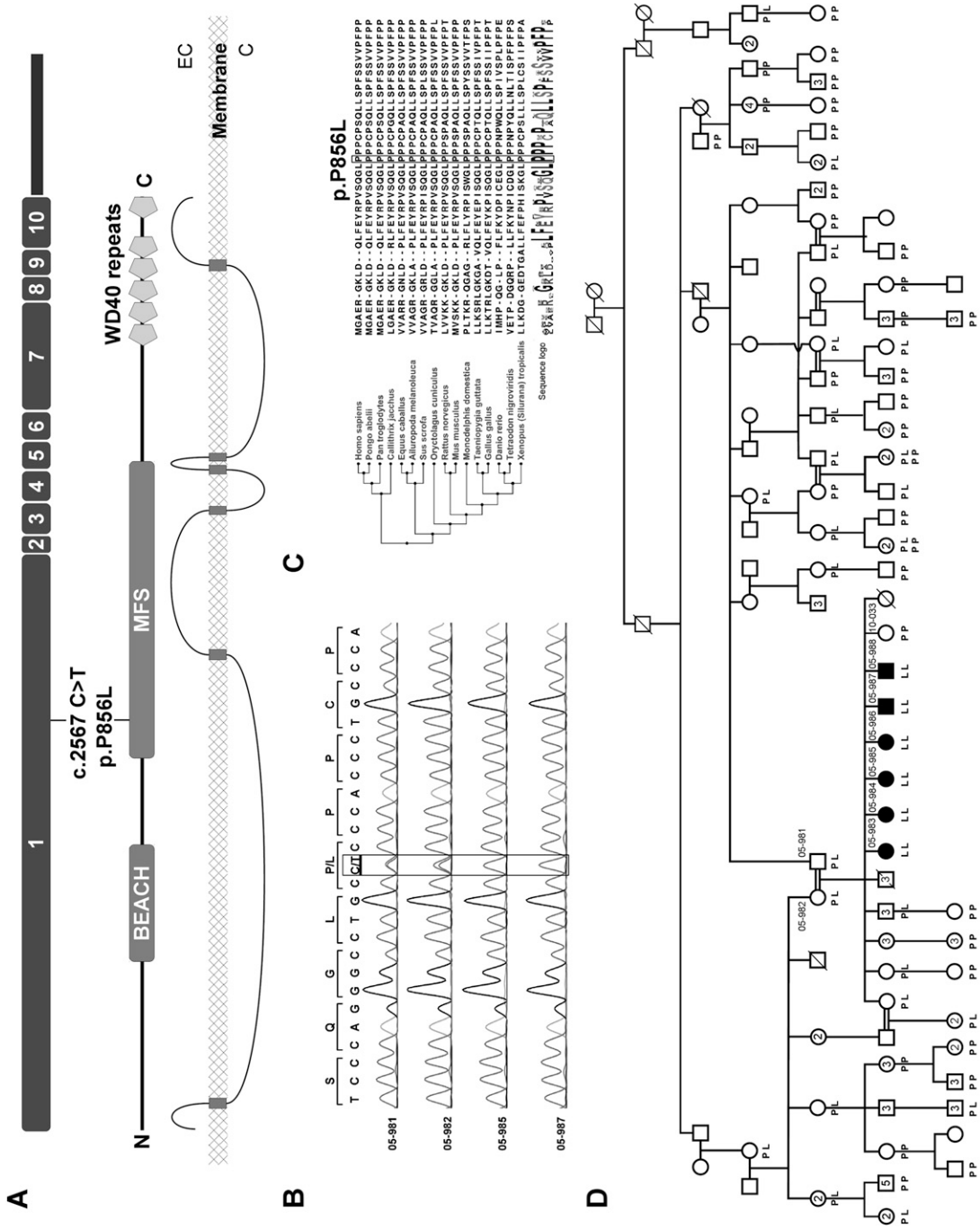


Figure 2. Identification of the *WDR81* mutation. (*A*) Genomic structure, predicted protein, and predicted transmembrane domains of *WDR81* gene [(EC) extracellular, (C) cytosol]. (*B*) Confirmation of the missense mutation c.2567C>T (p.P856L) in *WDR81* isoform 1 by Sanger sequencing. (*C*) Sequence homology of WDR81 protein in vertebrates. The box indicates the mutant amino acid. (*D*) Family B with affected individuals indicated by filled symbols and genotypes shown for WDR81 p.P856L.

have been implicated in membrane trafficking (Wang et al. 2000), synapse morphogenesis (Khodosh et al. 2006), and lysosomal axon transport (Lim and Kraut 2009). A BEACH domain is the major structural feature of neurobeachin, a scaffolding protein disrupted in a patient with autism (Volders et al. 2011). WDR81 p.P856L lies in a major facilitator superfamily (MFS) domain, a region characteristic of solute carrier transport proteins (Saier et al. 1999). The C terminus of WDR81 is composed of six WD-repeats that are likely constituents of a beta-propeller. Based on analysis by TMpred (www.ch.embnet.org/software/TMPRED_form.html), WDR81 is a transmembrane protein with six membrane-spanning domains, the most N-terminal at amino acids 45–66 and the other five at the C terminus of the protein, between amino acids 980 and 1815 (Fig. 2A). Supporting the likelihood that WDR81 is a transmembrane protein is the observation that *WDR81* transcript expression is increased in membrane-associated RNA in contrast to cytoplasmic RNA (4.14 folds, $P = 0.03$, and 1.78 folds, $P = 0.0002$ in Gene Expression Omnibus [GEO] [http://www.ncbi.nlm.nih.gov/geo/] data set GSE4175) (Diehn et al. 2006).

In order to assess a possible role for WDR81 in regulating motor behavior, we evaluated the expression profiles of human and mouse *WDR81/Wdr81* isoform 1 in the brain. Human *WDR81* isoform 1 transcript was expressed in all the tissues evaluated (Supplemental Fig. 5). In particular, all the brain tissues were positive for the transcript, with highest levels of expression in the cerebellum and corpus callosum (Fig. 3A). In the mouse brain at post-partum day P7, *Wdr81* expression was observed in Purkinje cell layer in the cerebellum (Fig. 3B,C). The cerebellum is a crucial regulatory center for motor function.

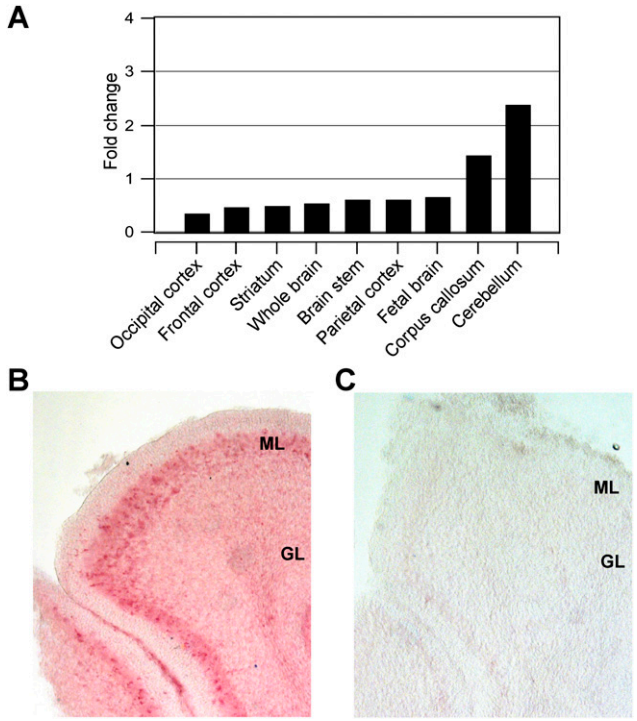


Figure 3. Expression pattern of *WDR81* in brain. (*A*) Expression in human brain with highest levels in cerebellum and corpus callosum. (*B*) In situ hybridization of mouse embryonic brain revealing increased expression of *Wdr81* in purkinje cells and molecular layer of cerebellum. (*C*) No hybridization was observed with the sense probe. (ML) Molecular layer, (GL) granular layer.

We examined the expression of *WDR81* in the context of expression profiles of the early embryonic mouse brain (GSE8091) (Hartl et al. 2008). Differentially expressed genes within the day groups were filtered (one-way ANOVA test Bonferroni-corrected $P < 0.001$, $n = 3611$). From these profiles, we identified the subset of genes whose expression was highly correlated with that of *WDR81* ($R > 0.95$, $n = 670$) and then used DAVID tools (Huang et al. 2009) to evaluate the predicted functions of this subset of genes. The subset of genes coexpressed with *WDR81* was enriched for those involved in neuronal differentiation and neuronal projection, axonogenesis, and cell morphogenesis (Bonferroni-corrected *P*-values 2.3×10^{-11}, 1.3×10^{-9}, and 3.7×10^{-9}, respectively). Among the genes coexpressed with *WDR81* were those encoding prion protein, doublecortin (responsible for lissencephaly), and *L1CAM* (responsible for MASA syndrome) (Supplemental Table 11). *WDR81* is not coexpressed with *VLDLR* and *CA8*, raising the possibility that *WDR81* represents a different developmental regulatory pathway.

Discussion

The identification of genes responsible for human disease has been greatly facilitated with new technologies, particularly the targeted enrichment of the genome by solution capture, followed by genomic sequencing (Bilguvar et al. 2010). Despite these advances, demonstrating the causality for a mutation in the absence of two or more independent cases remains a challenge. This is particularly true when multiple variants, none of them with obvious effect on protein function, cosegregate with the phenotype in the family; the candidate gene encodes a previously uncharacterized protein with multiple isoforms, of which the critical mutation is on only one; and the candidate mutation is a missense. However, unique families and uncharacterized proteins exist, and precisely because of this reason, it becomes imperative to fully exploit genetics and genomics approaches to distinguish the causative mutation.

We describe here the discovery of a mutation associated with an extremely rare and genetically heterogeneous autosomal recessive phenotype in a unique consanguineous family (Tan 2006). The putative causative mutation could be distinguished from previously unknown rare polymorphisms in the same genomic region by analysis of conservation at all candidate variant sites, by the presence of polymorphic stops in the critical region of another candidate gene, and by genotyping ethnically matched unaffected individuals who would not be expected to carry homozygous mutations at the mutant site. We conclude that the WDR81 p.P856L mutation is the cause of cerebellar hypoplasia associated with quadrupedal locomotion in Family B.

WDR81 is an uncharacterized gene. It shows similarity with a host of genes, including *NSMAF* (neutral sphyngomyelinase activation associated factor), *NBEA* (neurobeachin), and *LYST* (lysosomal trafficking regulator). The *LYST* gene contains HEAT/ARM repeats, a BEACH domain, and seven WD40 repeats (Ward et al. 2000). Nearly all reported *LYST* mutations result in protein truncation and lead to Chediak-Higashi syndrome (CHS), which is characterized by accumulation of giant intracellular vesicles leading to defects in the immune and blood systems (Rudelius et al. 2006). Two patients with missense *LYST* mutations have been reported (Karim et al. 2002). Interestingly, these patients presented with neurological symptoms without immunological involvement. The $Lyst^{Ing3618}/Lyst^{Ing3618}$ mutant mouse harbors a missense mutation in the WD40 domain. Purkinje cell degeneration accompanied by age-dependent impairment of motor coordination without

signs of lysosomal deficiency in immunological organs were characteristics of these animals (Rudelius et al. 2006).

Expression of *WDR81* at high levels in the human cerebellum and corpus callosum and in the Purkinje cell layer of the mouse cerebellum is consistent with our observations of major structural abnormalities in these regions of the brain of affected individuals. Together, these observations suggest a possible role for WDR81 in motor behavior. Further work will be required to understand the normal biological function of WDR81 and the role of the mutation in causing cerebellar hypoplasia and quadrupedal locomotion. Genomic analysis of Family B demonstrates that WDR81 is highly likely to be critical to these developmental processes.

Methods

Human subjects

The institutional review boards of Bilkent, Hacettepe, Baskent, and Cukurova Universities approved the study (decisions: BEK02, 28.08.2008; TBK08/4, 22.04.2008; KA07/47, 02.04.2007; and 21/3, 08.11.2005, respectively). Written informed consent, prepared according to the guidelines of the Ministry of Health in Turkey, was obtained from all family members and control group subjects prior to the study. A total of 18 subjects participated in MRI scans. Six of them were from Family B, including four affected siblings (05-984, 05-986, 05-987, 05-988), one normal female sibling homozygous for the wild-type allele of the WDR81 p.P856L variant (10-033), and their carrier father (05-981). The remaining 14 participants were age- and sex-matched healthy controls. The two male patients (age, mean ± SD = 37.00 ± 4.24) were matched to seven male controls (age, mean ± SD = 35.14 ± 5.76), and the two female patients (age, mean ± SD=27.00 ± 4.24) were matched to seven female controls (age, mean ± SD = 28.57 ± 3.64). Family B members were scanned under sedation. For the healthy controls, no sedation was performed. Sedation was achieved by initial administration of midazolam (2 mg per subject), which was followed by propofol (120 mg) and fentanyl (50 mcg) administration intravenously. Hypnosis level was adjusted by 20 mg injections of propofol approximately every 10 min to eliminate somatic responses such as slight movements. Blood oxygen level and heart rate were monitored during the entire procedure. Eyelash reflexes were absent at all times. Neuromuscular blockade was not used.

Next-generation sequencing

NimbleGen 385K microarrays were produced to capture the critical region at chr17: 82,514–7,257,922 (hg19) using 7464 unique probes with a total probe length of 4,853,455 bp. Sequence Search and Alignment by Hashing Algorithm (SSAHA) (Ning et al. 2001) was used to determine probe uniqueness by NimbleGen (Roche NimbleGen). Sequence capture was conducted by the NimbleGen facility using 25 μg of input DNA. Captured DNA samples were subjected to standard sample preparation procedures for 454 GS FLX sequencing with Titanium series reagents. Four full 454 GS FLX runs were conducted for two affected individuals (05-985, 05-987) and their unaffected obligate carrier parents (05-981 father, 05-982 mother). Sequence data were initially mapped to human genome reference sequence and annotated using the GSMapper software package (Roche). Fold enrichment of the target region was calculated with the formula $\sum$REMTrm/STrm: $\sum$RMG/SG as described previously (REMTrm, number of reads mapped to target region; STrm, size of target region; RMG, number of reads mapped outside of the target region; SG, size of human genome) (Rehman et al. 2010). Variants were identified with ALLDiff and more stringent HCDiff approaches (Hedges et al. 2009). Annotation of variants was made by GSMapper software using the refGene table of the University of California, Santa Cruz (UCSC) Genome Browser (Fujita et al. 2010). Ensembl 62 genome annotation data for hg19 human genome assembly were extracted using the BIOMART data-mining tool for further analysis of intronic and intergenic variants in terms of hypothetical genes and splicing variants (Flicek et al. 2011). Novel variants were reported based on the SNPs included in the reference SNP database. For Illumina sequencing, a total of 6,184,539-bp-long unique probes were designed to target a 9-Mb genomic region spanning the disease locus (chr17:0–9,059,276; hg19) using a custom NimbleGen HD2 2.1M sequence capture microarray. Another affected individual was sequenced with the Illumina Genome Analyzer IIx. Illumina sequence data were mapped to the reference genome using MAQ tools (Li et al. 2008), and single nucleotide variants were determined with Samtools (Li et al. 2009). To determine indels, data were mapped with BWA (Li and Durbin 2010) and analyzed with Samtools. Sequence data were visually analyzed using the Integrative Genomics Viewer (IGV) (Robinson et al. 2011).

Array based genotyping

We conducted Illumina 300 Duo v2 BeadChip for two affected individuals (05-984, 05-987) according to the manufacturer's recommendations (Illumina). The image data were normalized, and the genotypes were called using data analysis software (Bead Studio, Illumina). Sex, inbreeding, and sibship were confirmed. The Mendelian compatibility of sequence variants was analyzed with PLINK (Purcell et al. 2007).

DNA sequencing

Confirmation of novel variants identified by next-generation sequencing was done with conventional capillary sequencing. The Primer3 software (Rozen and Skaletsky 2000) was used to design PCR primers for the amplification of candidate variants (Supplemental Table 12). Products were analyzed via gel electrophoresis and were sequenced using forward and reverse primers on an ABI 3130 XL capillary sequencing instrument (Applied Biosystems). Sanger sequence trace files were analyzed with the CLCBio Main Workbench software package (CLCBio Inc.).

Population screening

To distinguish the disease-causing variant from novel polymorphisms, a population screening approach was conducted for each candidate variant. Allele-specific PCR (AS-PCR) and restriction fragment length polymorphism (RFLP) analyses were performed (Supplemental Table 12) on 1098 chromosomes from a healthy control population. In addition, the first-, second-, and third-degree relatives of the affected family, amounting to 177 individuals, were sampled for genotype analysis. Sanger sequencing was performed to confirm all of the variants detected in the normal population using the above-mentioned methods. Racial distribution of the control group was 100% Caucasian, including 22% from southeastern Turkey.

Quantitative real-time RT-PCR analysis of *WDR81* expression

First-strand cDNA was prepared from multi-tissue RNA panels (Clontech: 636567, 636643; Agilent: 540007, 540117, 540137, 540157, 540053, 540005, 540143, 540135) with RevertAid kit and random hexamer primers (Fermentas; K1622) after DNase I (Fermentas; EN0521) digestion. The PCR primers located in exon 1 and flanking the mutation site were designed using Primer3 soft-

ware (Supplemental Table 13; Rozen and Skaletsky 2000). SYBR Green real-time PCR were realized according to standard protocols (BioRad; 170-8882) with 100% PCR efficiency. Each assay included minus RT and nontemplate controls. C_t values were normalized to *GAPDH* as an internal control. The data were analyzed using the Pfaffl method (Pfaffl 2001).

In situ hybridization

In order to examine the specific expression pattern of *Wdr81* gene in the mouse brain, probes that contain the mutated region in human patients were prepared by PCR amplification of the region from mouse genomic DNA and subsequent cloning into plasmids. The riboprobes were synthesized by using Dig-labeled NTPs, and in situ hybridization experiments were performed as described (Tekinay et al. 2009). The Animal Ethics Committee of Bilkent University approved procedures for the tissue extraction and for in situ hybridization tests. Animals were group housed in a 12-h dark, 12-h light cycle. Embryo and P7 brain sections were prepared as described (Gong et al. 2003). Twenty-micrometer sagittal sections were taken with a cryostat (Leica). The antisense probe was prepared by PCR amplification from the mouse genomic DNA and subsequent cloning into pCR4-TOPO vector (Invitrogen). A modified version of pSK vector was used for cloning the sense probe of the same region. Digoxigenin (Dig)-labeled riboprobe was transcribed using Dig-NTP in the transcription reaction. Riboprobes were purified with Mini Quick Spin DNA columns (Roche) prior to hybridization. Sections were incubated at 60°C overnight in hybridization buffer containing 50% formamide, 5× SSC, 5× Denhardt's reagent, 50 μg/mL heparin, 500 μg/mL herring sperm DNA, and 250 μg/mL yeast tRNA. Hybridized sections were washed for 90 min with 50% formamide and 2× SSC at 60°C. Probes were detected with anti-Dig Fab fragments conjugated to alkaline phosphatase and NBT/BCIP substrate mixture (Tekinay et al. 2009).

Bioinformatics analyses

Homozygosity mapping analysis was performed using HomozygosityMapper software (Seelow et al. 2009). SIFT (Ng and Henikoff 2001) and PolyPhen (Sunyaev et al. 2001) tools were used to predict the functional impact of the variants. Genomic Evolutionary Rate Profiling (GERP) scores for each variant were obtained from the UCSC Genome Browser allHg19RS_BW track (Davydov et al. 2010). The PFAM protein domain search module of CLCMain Workbench V5.0 (CLCBio, Inc.) and ScanProsite (Gattiker et al. 2002) tools were used to predict domains and possible effects of the variant on protein product. Membrane spanning domains were predicted using TMpred software (www.ch.embnet.org/software/TMPRED_form.html). Homology searches were performed with CLCMain Workbench using appropriate modules (reference sequence accession codes for WDR81 orthologs are *Ailuropoda melanoleuca*, XP_002918082; *Callithrix jacchus*, XP_002747874; *Danio rerio*, XP_001921778; *Equus caballus*, XP_001502383; *Gallus gallus*, XP_415806; *Monodelphis domestica*, XP_001371487; *Mus musculus*, NP_620400; *Oryctolagus cuniculus*, XP_002718930; *Pan troglodytes*, XP_523527; *Pongo abelii*, XP_002826860; *Rattus norvegicus*, NP_001127832; *Sus scrofa*, XP_003131868; *Taeniopygia guttata*, XP_002194363; *Tetraodon nigroviridis*, CAG08933; *Xenopus* [*Silurana*] *tropicalis*, XP_002937192). Published microarray data sets of E9.5, E11.5, and E13.5 mouse brain tissue (GSE8091) were downloaded from the GEO database (http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi) (Hartl et al. 2008) and processed with GeneSpring GX V11.1 software (Agilent Technologies). Data sets were grouped within day groups, and standard quality control and filtering analysis were performed (http://www.chem.agilent.com/cag/bsp/products/gsgx/manuals/GeneSpring-manual.pdf). Differentially expressed genes within the day groups were filtered using a one-way ANOVA test (Bonferroni-corrected $P < 0.001$). Genes that correlated with *Wdr81* ($R = 0.95 - 1.0$) were obtained using the "Find Similar Entity Lists" module of the software. Functional annotation clustering was performed using the obtained gene list by DAVID tools (Huang et al. 2009). *WDR81* differential expression in the GEO data sets was further investigated using the NextBio System, a web-based data-mining engine (Kupershmidt et al. 2010), and the GSE4175 (Diehn et al. 2006) data set was selected as a significant difference in membrane-associated RNA versus cytoplasmic RNA comparisons. Ensembl identifiers of the candidate genes and transcripts are as follows: *WDR81* [ENSG00000167716; ENST00000409644], *MYBBP1A* [ENSG00000132382; ENST00000254718], and *ZNF594* [ENSG00000180626; ENST00000399604].

MRI data acquisition and structural analysis procedures

MRI data were acquired using a three Tesla scanner (Magnetom Trio, Siemens AG) with a 12-channel phase-array head coil. A high-resolution T1-weighted three-dimensional (3D) anatomical-volume scan was acquired for each participant (single-shot turbo flash; voxel size = $1 \times 1 \times 1$ mm^3; repetition time [TR] = 2600 msec; echo time[TE] = 3.02 msec; flip angle = 8°; field of view [FOV] = 256×224 mm^2; slice orientation = sagittal; phase encode direction = anterior-posterior; number of slices = 176; acceleration factor [GRAPPA] = 2). DTI data were acquired using a single-shot spin-echo EPI with a parallel imaging technique GRAPPA (acceleration factor 2). The sequence was performed with 30 gradient directions, and the diffusion weighting b-factor was set to 800 sec/mm^2 (TR, 6400 msec; TE, 88 msec; in-plane resolution, 1 mm × 1 mm; slice thickness, 3.0 mm; 50 transverse slices; base resolution, 128 × 128). Structural analyses were performed with the Freesurfer image analysis package (http://surfer.nmr.mgh.harvard.edu/). The analyses involved intensity normalization, removal of nonbrain tissue, subcortical segmentation (Fischl et al. 2002), and identification of the white matter/gray matter boundary upon which cortical reconstruction and volumetric parcellation were performed. The cortex was then registered to a spherical atlas and parceled into units according to the gyral and sulcal structure based on the Desikan-Kilinay Atlas (Desikan et al. 2006) and the Destrieux Atlas (Destrieux et al. 2010). Next, using the same software, we performed morphometric analyses of cortical thickness, mean curvature, surface area, and volume for each unit of parcellation and computed the group differences. Significant differences between the groups are determined using two-tailed unpaired *t*-tests at an alpha level of 0.05. Fiber tracking was performed in MedINRIA (Toussaint et al. 2007). Fibers with FA < 0.3 were excluded from the analysis. Region of interests (ROIs) were drawn manually over cross-sections of superior, middle, and inferior cerebellar peduncles, using the MRI Atlas of Human White Matter as a reference (Oishi et al. 2010). ROIs were drawn at approximately corresponding locations for the patients and healthy controls. Fiber tracts were first limited to pass through these ROIs and were then subsequently refined using a recursive tracking technique (Toussaint et al. 2007). T1-weighted images were coregistered with DWI data using FSL (Smith et al. 2004; Woolrich et al. 2009). Final tracts were manually overlaid onto high-resolution T1-weighted images for illustration purposes.

Data access

Sequence data of the homozygous region has been deposited at the DNA Data Bank of Japan (DDBJ; http://www.ddbj.nig.ac.jp/) under accession no. DRA000432. SNP genotype data have been deposited at the European Genome-Phenome Archive (EGA; http://www.

ebi.ac.uk/ega/), which is hosted at the EBI, under accession no. EGAS00000000099.

Acknowledgments

We thank Dr. Mary-Claire King for innumerable discussions, suggestions, and critical reading of the manuscript. We also thank the members of Family B and their relatives for cooperation in this study. Dr. Alper Iseri and Dr. Bayram Kerkez kindly provided technical and logistic support. This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK-SBAG 108S036 and 108S355) and the Turkish Academy of Sciences (TUBA research support) to T.O., and the European Commission (PIRG-GA-2008-239467) and TUBA-GEBIP award to H.B.

Authors' contributions: S.G., A.B.T., K.D., H.B., and T.O. conceived and designed the experiments. S.G., H.U., K.D., and H.B. performed the experiments. S.G., A.B.T., K.D., H.B., K.B., H.U., A.O., E.A., T.K., M.G., and T.O. analyzed the data. O.E.O., A.N.B., H.T., M.T., and U.T. contributed patient materials. S.G. and T.O. wrote the paper.

References

Bilguvar K, Ozturk AK, Louvi A, Kwan KY, Choi M, Tatli B, Yalnizoglu D, Tuysuz B, Caglayan AO, Gokben S, et al. 2010. Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature* **467:** 207–210.

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* **6:** e1001025. doi: 10.1371/journal.pcbi.1001025.

Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, et al. 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31:** 968–980.

Destrieux C, Fischl B, Dale A, Halgren E. 2010. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage* **53:** 1–15.

Diehn M, Bhattacharya R, Botstein D, Brown PO. 2006. Genome-scale identification of membrane-associated human mRNAs. *PLoS Genet* **2:** e11. doi: 10.1371/journal.pgen.0020011.

Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S, et al. 2002. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron* **33:** 341–355.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, et al. 2011. Ensembl 2011. *Nucleic Acids Res* **39:** D800–D806.

Fogel BL, Perlman S. 2007. Clinical features and molecular genetics of autosomal recessive cerebellar ataxias. *Lancet Neurol* **6:** 245–257.

Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D. 2010. The UCSC Genome Browser database: update 2011. *Nucleic Acids Res* **39:** D876–D882.

Gattiker A, Gasteiger E, Bairoch A. 2002. ScanProsite: a reference implementation of a PROSITE scanning tool. *Appl Bioinformatics* **1:** 107–108.

Gong S, Zheng C, Doughty ML, Losos K, Didkovsky N, Schambra UB, Nowak NJ, Joyner A, Leblanc G, Hatten ME, et al. 2003. A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* **425:** 917–925.

Hartl D, Irmler M, Romer I, Mader MT, Mao L, Zabel C, de Angelis MH, Beckers J, Klose J. 2008. Transcriptome and proteome analysis of early embryonic mouse brain development. *Proteomics* **8:** 1257–1265.

Hedges DJ, Burges D, Powell E, Almonte C, Huang J, Young S, Boese B, Schmidt M, Pericak-Vance MA, Martin E, et al. 2009. Exome sequencing of a multigenerational human pedigree. *PLoS ONE* **4:** e8232. doi: 10.1371/journal.pone.0008232.

Huang DW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4:** 44–57.

Karim MA, Suzuki K, Fukai K, Oh J, Nagle DL, Moore KJ, Barbosa E, Falik-Borenstein T, Filipovich A, Ischida Y, et al. 2002. Apparent genotype–phenotype correlation in childhood, adolescent, and adult Chediak–Higashi syndrome. *Am J Med Genet* **108:** 16–22.

Khodosh R, Augsburger A, Schwarz TL, Garrity PA. 2006. Bchs, a BEACH domain protein, antagonizes Rab11 in synapse morphogenesis and other developmental events. *Development* **133:** 4655–4665.

Kolb LE, Arlier Z, Yalcinkaya C, Ozturk AK, Moliterno JA, Erturk O, Bayrakli F, Korkmaz B, DiLuna ML, Yasuno K, et al. 2010. Novel VLDLR microdeletion identified in two Turkish siblings with pachygyria and pontocerebellar atrophy. *Neurogenetics* **11:** 319–325.

Kupershmidt I, Su Q J, Grewal A, Sundaresh S, Halperin I, Flynn J, Shekar M, Wang H, Park J, Cui W, et al. 2010. Ontology-based meta-analysis of global collections of high-throughput public data. *PLoS ONE* **5:** e13066. doi: 10.1371/journal.pone.0013066.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26:** 589–595.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18:** 1851–1858.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079.

Lim A, Kraut R. 2009. The *Drosophila* BEACH family protein, blue cheese, links lysosomal axon transport with motor neuron degeneration. *J Neurosci* **29:** 951–963.

Moheb LA, Tzschach A, Garshasbi M, Kahrizi K, Darvish H, Heshmati Y, Kordi A, Najmabadi H, Ropers HH, Kuss AW. 2008. Identification of a nonsense mutation in the very low-density lipoprotein receptor gene (VLDLR) in an Iranian family with dysequilibrium syndrome. *Eur J Hum Genet* **16:** 270–273.

Ng PC, Henikoff S. 2001. Predicting deleterious amino acid substitutions. *Genome Res* **11:** 863–874.

Ning Z, Cox A, Mullikin J. 2001. SSAHA: A fast search method for large DNA databases. *Genome Res* **11:** 1725–1729.

Oishi K, Faria AV, van Zijl PCM, Mori S. 2010. *MRI atlas of human white matter*, 2nd ed. Elsevier, Amsterdam.

Ozcelik T, Akarsu N, Uz E, Caglayan S, Gulsuner S, Onat OE, Tan M, Tan U. 2008. Mutations in the very low-density lipoprotein receptor VLDLR cause cerebellar hypoplasia and quadrupedal locomotion in humans. *Proc Natl Acad Sci* **105:** 4232–4236.

Pfaffl MW. 2001. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res* **29:** e45. doi: 10.1093/nar/29.9.e45.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81:** 559–575.

Rehman AU, Morell RJ, Belyantseva IA, Khan SY, Boger ET, Shahzad M, Ahmed ZM, Riazuddin S, Khan SN, Riazuddin S, et al. 2010. Targeted capture and next-generation sequencing identifies C9orf75, encoding Taperin, as the mutated gene in nonsyndromic deafness DFNB79. *Am J Hum Genet* **86:** 378–388.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29:** 24–26.

Ropers HH. 2007. New perspectives for the elucidation of the genetic disorders. *Am J Hum Genet* **81:** 199–207.

Rozen S, Skaletsky HJ. 2000. Primer3 on the WWW for general users and for biologist programmers. In *Bioinformatics methods and protocols: Methods in molecular biology* (ed. S Krawetz, S Misener), p. 365. Humana Press, Totowa, NJ.

Rudelius M, Osanger A, Kohlmann S, Augustin M, Piontek G, Heinzmann U, Jennen G, Russ A, Matiasek K, Stumm G, et al. 2006. A missense mutation in the WD40 domain of murine Lyst is linked to severe progressive Purkinje cell degeneration. *Acta Neuropathol* **112:** 267–276.

Saier MH Jr, Beatty JT, Goffeau A, Harley KT, Heijne WH, Huang SC, Jack DL, Jähn PS, Lew K, Liu J, et al. 1999. The major facilitator superfamily. *J Mol Microbiol Biotechnol* **1:** 257–279.

Seelow D, Schuelke M, Hildebrandt F, Nürnberg P. 2009. HomozygosityMapper: an interactive approach to homozygosity mapping. *Nucleic Acids Res* **37:** W593–W599.

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TE, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, et al. 2004. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* **23:** S208–S219.

Sunyaev S, Ramensky V, Koch I, Lathe W 3rd, Kondrashov AS, Bork P. 2001. Prediction of deleterious human alleles. *Hum Mol Genet* **10:** 591–597.

Tan U. 2006. A new syndrome with quadrupedal gait, primitive speech, and severe mental retardation as a live model for human evolution. *Int J Neurosci* **116:** 361–369.

Tekinay AB, Nong Y, Miwa JM, Lieberam I, Ibanez-Tallon I, Greengard P, Heintz N. 2009. A role for LYNX2 in anxiety-related behavior. *Proc Natl Acad Sci* **106:** 4477–4482.

Toussaint N, Souplet JC, Fillard P. 2007. MedINRIA: Medical image navigation and research tool by INRIA. In *Proceedings of MICCAI'07*

Workshop on Interaction in Medical Image Analysis and Visualization. Brisbane, Australia. Lecture Notes in Computer Science, Vol. 4791. Springer, Berlin.

Turkmen S, Demirhan O, Hoffmann K, Diers A, Zimmer C, Sperling K, Mundlos S. 2006. Cerebellar hypoplasia and quadrupedal locomotion in humans as a recessive trait mapping to chromosome 17p. *J Med Genet* **43:** 461–464.

Turkmen S, Guo G, Garshasbi M, Hoffmann K, Alshalah AJ, Mischung C, Kuss A, Humphrey N, Mundlos S, Robinson PN. 2009. CA8 mutations cause a novel syndrome characterized by ataxia and mild mental retardation with predisposition to quadrupedal gait. *PLoS Genet* **5:** e1000487. doi: 10.1371/journal.pgen.1000487.

Wang X, Herberg FW, Laue MM, Wullner C, Hu B, Petrasch-Parwez E, Kilimann MW. 2000. Neurobeachin: a protein kinase A-anchoring, beige/Chediak-higashi protein homolog implicated in neuronal membrane traffic. *J Neurosci* **20:** 8551–8565.

Ward DM, Griffiths GM, Stinchcombe JC, Kaplan J. 2000. Analysis of the lysosomal storage disease Chediak–Higashi syndrome. *Traffic* **1:** 816–822.

Woolrich MW, Jbabdi S, Patenaude B, Chappell M, Makni S, Behrens T, Beckmann C, Jenkinson M, Smith SM. 2009. Bayesian analysis of neuroimaging data in FSL. *Neuroimage* **45:** S173–S186.

Volders K, Nuytens K, Creemers JW. 2011. The autism candidate gene neurobeachin encodes a scaffolding protein implicated in membrane trafficking and signaling. *Curr Mol Med* **11:** 204–217.

Received May 11, 2011; accepted in revised form August 23, 2011.

PERMISSONS FOR COPYRIGHTED MATERIAL

SINAUER ASSOCIATES, Inc. • Publishers • P.O. Box 407 • Sunderland, MA 01375-0407

Telephone: (413) 549-4300
Fax: (413) 549-1118
E-mail: orders@sinauer.com

PERMISSIONS AGREEMENT

October 26, 2011
Revised: October 27, 2011

Permission granted to:
Suleyman Gulsuner
Bilkent University
Faculty of Science
Department of Molecular Biology and Genetics
Ankara, NA 06800
TURKEY

FAX: 011 90 312 266 50 97

Material to be reproduced:
Purves et al: *Neuroscience, Second Edition*
Figure 19.1, page 410 and Figure 19.5, page 413

To be reproduced and or modified in the work:
Suleyman Gulsuner's PhD Thesis entitled "Quadrupedal locomotion in human" to be published by Bilkent Unviersity

Sinauer Associates owns copyright to the material described above and hereby grants permission for the one-time use of the material as specified, and for nonexclusive world rights provided that full and appropriate credit is given to the original source and that the work is for NON-COMMERCIAL use only. Please request permission for further use in subsequent editions, translations, or revisions of the work.

Sherri L. Ellsworth
Permissions Coordinator

Oct. 27, 2011
Date

Please acknowledge your acceptance of these terms by signing one copy of this form and returning it to Sinauer Associates. Permission Agreement is not valid until signed by applicant and received by Sinauer Associates.

Signature of Applicant

Oct. 28, 2011
Date

Dear MD Suleyman Ismail Gulsuner

We hereby grant you permission to reprint the material below at no charge in your thesis subject to the following conditions:

1. If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies.

2. Suitable acknowledgment to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

"This article was published in Publication title, Vol number, Author(s), Title of article, Page Nos, Copyright Elsevier (or appropriate Society name) (Year)."

3. Your thesis may be submitted to your institution in either print or electronic form.

4. Reproduction of this material is confined to the purpose for which permission is hereby given.

5. This permission is granted for non-exclusive world English rights only. For other languages please reapply separately for each one required. Permission excludes use in an electronic form other than submission. Should you have a specific electronic project in mind please reapply for permission.

6. Should your thesis be published commercially, please reapply for permission.

*6 figures only.

Yours sincerely,

Emma

Emma Williams
:: Rights Associate :: Global Rights :: Elsevier
T +44 (0) 1865 843841 :: F +44 (0)1865 853333
e.williams@elsevier.com

-----Original Message-----
From: gulsuner@bilkent.edu.tr [mailto:gulsuner@bilkent.edu.tr]
Sent: 12 September 2011 23:53
To: Rights and Permissions (ELS)
Subject: Obtain Permission

This Email was sent from the Elsevier Corporate Web Site
and is related to Obtain Permission form:

Product: Customer Support
Component: Obtain Permission
Web server: http://www.elsevier.com
IP address: 139.179.138.110
Client: Mozilla/5.0 (Windows NT 6.1; WOW64; rv:6.0.2) Gecko/20100101 Firefox/6.0.2
Invoked from:
http://www.elsevier.com/wps/find/obtainpermissionform.cws_home?isSubmitted=yes&navigateXmlFileName=/store/p65idstarget/act/framework_support/obtainpermission.xml

Request From:
MD Suleyman Ismail Gulsuner
Bilkent University
Dept. of Molecular Biology and Genetics
06800
Ankara
Turkey

Contact Details:
Telephone:
Fax:
Email Address: gulsuner@bilkent.edu.tr

To use the following material:
ISSN/ISBN:
Title: TEXTBOOK OF MEDICAL PHYSIOLOGY
Author(s): Dr. John E. Hall
Volume: NA
Issue: NA
Year: 2006
Pages: 555 - 713
Article title: The Nervous System

How much of the requested material is to be used:
Figures 45-2, 45-3, 56-4, 56-5, 56-6, 56-7

Are you the author: No
Author at institute: No

How/where will the requested material be used: [how_used]

Details:

I'd like to use the figures in my PhD dissertation.

Additional Info:

[acronym]
- end -

Subject:	Re: Reprint Request of the Global Prevalence
From:	"Alan Bittles" <abittles@ccg.murdoch.edu.au>
Date:	Tue, October 25, 2011 10:35 pm
To:	gulsuner@bilkent.edu.tr (more)

Dear Suleyman,
Many congratulations on the publication of your excellent paper, and our best wishes for the successful presentation and defence of your PhD thesis. We do feel free to use our Global Prevalence of Consanguinity plot in your thesis presentation.
With best regards,
Alan Bittles.

BMJ PUBLISHING GROUP LTD. LICENSE
TERMS AND CONDITIONS
Oct 12, 2011

This is a License Agreement between Bilkent University ("You") and BMJ Publishing Group Ltd. ("BMJ Publishing Group Ltd.") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by BMJ Publishing Group Ltd., and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	2766520125686
License date	Oct 12, 2011
Licensed content publisher	BMJ Publishing Group Ltd.
Licensed content publication	Journal of Medical Genetics
Licensed content title	Cerebellar hypoplasia and quadrupedal locomotion in humans as a recessive trait mapping to chromosome 17p
Licensed content author	S Türkmen, O Demirhan, K Hoffmann, A Diers, C Zimmer, K Sperling, S Mundlos
Licensed content date	May 1, 2006
Volume number	43
Issue number	5
Type of Use	Thesis/Dissertation
Requestor type	Individual
Format	Print and electronic
Portion	Figure/table/extract
Number of figure/table/extracts	2
Will you be translating?	No
Circulation/distribution	50
Title of your thesis / dissertation	QUADRUPEDAL GAIT IN HUMANS, IDENTIFICATION AND PARTIAL CHARACTERIZATION OF A NOVEL GENE WD REPEAT PROTEIN 81 (WDR81)
Expected completion date	Dec 2011
Estimated size(pages)	150
BMJ VAT number	674738491
Billing Type	Invoice
Company	Bilkent University
Billing address	Bilkent University Faculty of Science ankara, NA 06800 Turkey
Customer reference info	
Permissions Cost	0.00 USD
VAT (0.0%)	0.00 USD
Total	0.00 USD
Terms and Conditions	BMJ Group Terms and Conditions for Permissions

GENOME RESEARCH

COPYRIGHT TRANSFER AGREEMENT

Publisher: Cold Spring Harbor Laboratory Press
Journal: *Genome Research*

It is herein agreed that: The **Author** grants and assigns exclusively to the Publisher (Cold Spring Harbor Laboratory Press) for its use any and all rights of whatsoever kind or nature now or hereafter protected by the Copyright Laws of the United States and all foreign countries in all languages in and to the identified article, including all subsidiary rights. The Author warrants that his contribution is original and that he has full power to make this grant. The Author signs for and accepts responsibility for releasing this material on behalf of any and all co-authors.

In return for these rights: The **Publisher** agrees to publish the identified article at its own cost and expense, with the exception of retaining the option to charge the Author for author changes in type and page charges. It is understood that the Author will receive no royalty or other monetary compensation for use of his material. The Publisher also agrees to make all reasonable efforts to act in the interests of the Author and his employer, as well as in its own interest. In particular, the Publisher requires that the consent of the first-named author be sought as a condition in granting permission to others to reprint all or portions of this article. The Publisher also hereby grants to the Author, and to the employer for whom the work was performed, royalty-free rights to this article as stipulated under Returned Rights below.

Returned Rights: The Author and employer are herein given royalty-free permission to:

1. Retain all proprietary rights other than copyright, such as patent rights.
2. Reuse all or portions of the above paper in other works, provided that the source and the Publisher's copyright are indicated.
3. Reproduce, or have reproduced, the above paper for the author's personal use or for company use provided that

    a. the source and the Publisher's copyright are indicated,
    b. the copies are not used in a way that implies Publisher's endorsement of a product or service of an employer, and
    c. the copies per se are not offered for sale.

4. Make limited distribution of all or portions of the above paper prior to publication.
5. In the case of work performed under US Government contract, the Publisher grants the US Government royalty-free permission to reproduce all or portions of the above paper, and to authorize others to do so, for US Government purposes.

In the event the above article is not accepted and published by the Publisher, or is withdrawn by the Author prior to receipt of galley proofs, this agreement becomes null and void.