Exome sequencing and disease-network analysis of a single family implicate a mutation in *KIF1A* in hereditary spastic paraparesis.

Yaniv Erlich[1,+,*], Simon Edvardson[2,+], Emily Hodges[3], Shamir Zenvirt[2], Pramod Thekkat[3], Avraham Shaag[2], Talya Dor[2]
Gregory J. Hannon[3], Orly Elpeleg[2]

1 Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142.

2 Monique and Jacques Roboh Department of Genetic Research, the Department of Genetic and Metabolic Diseases, Hadassah, Hebrew University Medical Center, 91120 Jerusalem, Israel.

3 Watson School of Biological Sciences, Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, One Bungtown Road, Cold Spring Harbor, NY 11724.

+ These authors contributed equally to this work

* To whom correspondence should be addressed (yaniv@wi.mit.edu)

Keywords: Exome sequencing, hereditary spastic paraparesis, disease-netwrok analysis

Running title: Sequencing and disease analysis implicate a HSP mutation

Abstract

Whole exome sequencing has become a pivotal methodology for rapid and cost-effective detection of pathogenic variations in Mendelian disorders. A major challenge of this approach is determining the causative mutation from a substantial number of bystander variations that do not play any role in the disease etiology. Current strategies to analyze variations have mainly relied on genetic and functional arguments such as mode of inheritance, conservation, and loss of function prediction. Here, we demonstrate that disease-network analysis provides an additional layer of information to stratify variations even in the presence of incomplete sequencing coverage, a known limitation of exome sequencing. We studied a case of Hereditary Spastic Paraparesis in a single inbred Palestinian family. HSP is a group of neuropathological disorders that are characterized by abnormal gait and spasticity of the lower limbs. Forty five loci have been associated with HSP and lesions in 20 genes have been documented to induce the disorder. We used whole exome sequencing and homozygosity mapping to create a list of possible candidates. After exhausting the genetic and functional arguments, we stratified the remaining candidates according to their similarity to the previously known disease genes. Our analysis implicated the causative mutation in the motor domain of *KIF1A*, a gene that has not yet associated with HSP, which functions in anterograde axonal transportation. Our strategy can be useful for a large class of disorders that are characterized by locus heterogeneity, particularly when studying disorders in single families.

The datasets are available on http://cancan.cshl.edu/hsp/ and on dbGAP (http://www.ncbi.nlm.nih.gov/gap).

Introduction

Whole exome sequencing has ushered in a renaissance in identifying pathogenic variations in monogenic diseases. The approach enables a rapid and cost-effective detection from a small number of individuals, and has proved useful for a wide range of clinical settings (Choi et al. 2009; Ng et al. 2009; Krawitz et al. 2010; Ng et al. 2010a; Ng et al. 2010b; Pierce et al. 2010; Walsh et al. 2010). A major challenge of whole exome sequencing is determining the causative mutation from a substantial number of bystander variations that do not play a role in the disease etiology. The common strategy to filter out the bystander variations is based on a systematic rejection of variations according to genetic and functional arguments, narrowing down the candidate list until the causative variation is isolated. One class of genetic arguments that has been widely used is rejection of variations that are not shared between multiple cases or that do not follow the assumed mode of inheritance (Ng et al. 2009; Krawitz et al. 2010; Ng et al. 2010b). Another class of genetic arguments asserts that harmful mutations must be rare due to purifying selection and accordingly reject variations that have been catalogued in dbSNP or 1000 Genomes as these are assumed to be relatively common. Functional arguments focus on the impact of the variation on the protein, either by analyzing biochemical and structural features (Adzhubei et al. 2010) or by measuring the extent of multi-species conservation at the variation site (Cooper et al. 2010).

Monogenic disorders in isolated inbred families cast a unique setting for identifying the causative variations. Homozygosity mapping can quickly identify regions that are identical by decent, refining the search area to regions that typically span several megabases (Lander and Botstein 1987). However, this setting brings several challenges. First, bystander variations inside the homozygous region are (almost) always homozygous; therefore, they are not amenable for rejection based on mode of inheritance (Choi et al. 2009; Walsh et al. 2010). Second, the disease-harboring region is identical between the affected siblings and sequencing multiple cases from the same family adds minimal information beyond the data from the homozygosity mapping. Third, inbred unions are more prevalent in non-Western societies, such as North Africa, the Middle East, and Central Asia (Bittles 2001). These ethnic groups tend to be less represented in variation repositories (Carlson et al. 2003; Via et al. 2010), reducing the power of eliminating common polymorphisms.

In this study, we present disease-network analysis as an additional layer of arguments to stratify variations. Accumulating lines of evidence have shown that genes that are associated with phenotypically close disorders are prone to have similar molecular signatures (Goh et al. 2007; Feldman et al. 2008; Wu et al. 2008). These include similar expression profiles, participation in the same signaling or

metabolic pathways, or sharing similar protein domains. Thus, it is appealing to reject variations by comparing the molecular signature of their harboring genes to the signatures of closely related disease genes.

Unlike most of the rejection arguments above, disease-based arguments do not require sequencing information *per se*. This provides a means to exclude positions that were not covered or miscalled, a common caveat in whole exome sequencing (Ng et al. 2010a). Disease-network analysis has been used for gene prioritizing as the first step of gene-centric studies, when sequencing was a limiting factor (Ropers 2007). One drawback of this approach is the requirement to prioritize a large number of genes which increases the amount of false positives. In our approach, the analysis is executed at the last stage on a minimal list of genes, increasing the specificity of the method.

Results

Description of the affected family

We evaluated the performance of our strategy in a case of three brothers aged 20 (Patient II-4), 15 (Patient II-5), and 14 (Patient II-7) years, who presented to our clinic with a chief complaint of abnormal gait. The parents were of Moslem Palestinian origin and denied consanguinity; however both originated from the same village. They and the other children were reportedly healthy at 16 to 23 years old (**Figure 1**). The perinatal course and early development of all patients were uneventful and walking appeared at around one year of age. Stiff legs and a slowly progressive gait disturbance became evident at two years but both participated successfully in sport activities throughout childhood. Further aggravation of symptoms was noted at 10-13 years of age and the patients could no longer play soccer or walk long distances. Lower limb sensation, sphincters, upper limbs and intellectual functions were intact throughout childhood.

Patient II-4, when first examined at age 20, was unable to run. His gait was stiff with equinos-gait, scissoring and crouching. He had increased reflexes in his lower limbs, plantar reflexes were extensor and bilateral ankle clonus was elicited. Bilateral spasticity on ankle dorsiflexion, hip adduction and knee extension was 3 according to the modified Ashworth scale. The rest of the physical examination was unremarkable. Patient II-5, a 15-year-old male, was noted to have difficulty in running. The tendon reflexes were increased in his lower limbs and the plantar reflexes were extensor but ankle clonus was not elicited. Spasticity was noted on ankle dorsiflexion and hip adduction up to a modified Ashworth score of 2. The rest of the physical examination was normal. Patient II-7, at 14 years, was able to run slowly and with a tendency to tip-toeing. Physical examination disclosed increased reflexes

in the lower limbs, ankle clonus and extensor plantar responses. Reflexes were normal in the upper limbs. Spasticity was evident in both lower limbs at rest with a modified Ashworth score of 3 on ankle dorsiflexion. Knee flexion and hip adduction were less affected with an Ashworth score of 2. No weakness, wasting, decreased vibration sense or pinprick sensation was noted, and there were no sphincter disturbances or cerebellar signs. Brain MRI in all patients was normal. EMG and NCV in II-4 and II-5 were normal.

Based on these results, the patients were diagnosed with pure type of Hereditary Spastic Paraparesis (HSP). HSP is a group of genetic disorders resulting in axonal degradation of the corticospinal tract (Reid 2003; Dion et al. 2009). They are characterized by a progressive lower-extremity spastic weakness, hypertonic urinary bladder disturbance, mild diminution of lower-extremity vibration sensation and, occasionally, of joint position sensation. As of Sep 2010, 45 HSP loci and 20 HSP-related genes have been identified (**Supplemental Table 1**), showing autosomal dominant, autosomal recessive, and X-linked inheritance patterns (Dion et al. 2009). HSPs are subdivided into pure and complicated forms, where the complicated forms involve additional pathologies, such as mental retardation, ocular signs, and skin abnormalities. However, some loci are associated with both pure and complicated form of HSPs (Reid 2003; Dion et al. 2009).

Rejecting involvement of known HSP genes or X-linked disorders

We determined the complete exonic sequences of patient II-5 and his parents using array-based hybrid selection and Illumina sequencing (see **Methods**) and we genotyped the whole genome of Patient II-4 and Patient II-5 using GeneChip Human Mapping 250K Nsp Array of Affymetrix, as previously described (Edvardson et al. 2007).

We first evaluated the possibility that the disorder is caused by known HSP genes. Patient II-4 and II-5 shared 610Mbase (23%) of identical autosomal segments, and 75Mbase (50%) of identical segments on the chromosome X, as expected from full siblings. The shared autosomal segments contained 4 known HSP genes: *ATL1*, *GJC2*, *HSPD1*, and *SPG20*. The sequencing coverage was 99% for the coding regions of these genes. We could not find any homozygous variation or compound heterozygous variations in the genes that can explain the disorder (**Supplemental data and Supplemental Table 2**). In addition, a search for an X-linked mutation did not find any potential variation (**Supplemental data and Supplemental Table 3**).

We also compared a list of homozygous SNPs from the entire autosome of patient II-5 to more than 40,000 known pathogenic variations in HGMD (Stenson et

al. 2003) revision 2009.2. This did not identify any previously known variation that can account for the observed phenotype. Our systematic analysis indicates that the disorder is caused by a new autosomal gene.

Analyzing homozygous regions

Homozygous runs have an overwhelming probability to carry disease mutations in inbred families (Lander and Botstein 1987). The array data revealed 4 homozygous runs that are shared between the brothers and longer than 1Mbase. The longest segment spanned more than 2.5Mb at the tip of chromosome 2 (**Table 1**). The length of homozygous runs is inversely correlated with the number of meiosis events from the common ancestor (Clark 1999). Therefore, the longest segment is prone to hold the disease mutation with a probability that is non-proportional to its relative length (Woods et al. 2006). In order to maximize the sensitivity of our analysis, we included all four homozygous runs, with the expectation that the longest region will harbor the disease mutation.

The four homozygous runs contained 44 genes that consist of 76,588 bases that are either translated or reside in splicing sites. The sequencing data of patient II-5, the mother, and the father covered 90%, 90%, and 89% of these bases, respectively. Importantly, when we examined the 7.8K bases that were not covered, we found that 5.6K bases were part of exons that were not included in the design of the capture array. All of these exons reside in the longest homozygous run on the tip of chromosome 2. We did not exclude the positions that were not covered by sequencing, and we treated them as hypothetical suspects until proven innocent.

We applied a series of exclusion steps to uncover the causative mutation: (a) we rejected positions that matched the reference, reducing the candidate list to 8083 positions, out of which 213 positions were variants detected using the sequencing data (b) by analyzing the sequencing data from the parents, we excluded positions that were homozygous in one of the two parents (c) variations that were documented in dbSNP130 or in the 1000 Genomes project were excluded (d) we rejected synonymous substitutions (e) we removed positions that were not subject to even a modest purifying selection, as indicated by a score of zero or less in GERP (Cooper et al. 2005). 5098 positions in 15 genes passed this series of exclusion criteria, all of which reside in the longest homozygous run (**Table 2**). However, only 5 positions were variations that were called according to the sequencing information, whereas the rest of the candidate positions were unknowns, mostly due to the absence of a corresponding probe in the array design.

We turned to disease-network analysis to identify the pathogenic variation. We composed a list of all the known genes that are associated with a pure type of

HSP (**Table 3**), and measured their signature similarity to the 15 candidate genes. To increase the robustness of our analysis, we used three algorithms: SUSPECTS (Adie et al. 2005) (**Supplemental Table 4**), Toppgene (Chen et al. 2007) (**Supplemental Table 5**), and Endeavour (Aerts et al. 2006) (**Supplemental Table 6**). Each of these algorithms relies on a distinct combination of features and metrics to characterize the signature of similar disease-causing genes. SUSPECTS mainly uses sequence features, such as GC content and gene length, Toppgene integrates mouse phenotype data, and Endeavour fuses multiple data sources, including gene ontology, text mining, and expression data. The results were univocal: *KIF1A* was the top candidate in all algorithms (**Table 4**). This was supported by gene interaction, motif score, annotation similarity, sequence similarity to other disease genes, and mouse phenotype data.

We repeated the Endeavour analysis with a training set that did not include *KIF5A* in order to circumvent possible biases due to high similarity to *KIF1A*. Again, *KIF1A* was the top gene. To check the robustness of the results, we randomly selected subsets of five training genes and repeated the Endeavour analysis for the entire candidate list. In 9 out of 10 times *KIF1A* was the top gene. Finally, we trained Endeavour with the entire set of 20 genes that cause either complicated or pure form of HSP. *KIF1A* was also the top gene in this setting.

KIF1A had complete sequencing information, and the only variation that passed the multiple exclusion criteria was Ala255Val. As a complementary approach, we performed a loss of function analysis on the 5 variations that passed the elimination process. We used MutationTaster (Schwarz et al. 2010), Polyphen (Adzhubei et al. 2010), and SIFT (Ng and Henikoff 2003) to classify the variations. All of these analysis methods implicated Ala225Val in *KIF1A* as a harmful mutation. The results on the other variations were not consistent between the tools, except a variation in *HDLBP* that was also found harmful (**Figure 2**). Using Sanger sequencing, we found that the putative variation in *HDLBP* was a sequencing error (**Supplemental Figure 1**). Therefore, the only harmful variation according to our stringent analysis is Ala255Val in *KIF1A*.

We validated the *KIF1A* mutation by Sanger sequencing, confirming that the three patients are homozygous to this variation, whereas the parents and four unaffected siblings were heterozygous (**Supplemental Figure 2**). There is less than 0.5% chance of this segregation pattern in the children at random. To determine the carrier rate, we genotyped 573 anonymous individuals of the same ethnic origin and found 3 carriers, indicating a carrier rate of 1:191 in this population (95% confidence interval: 0.06%-1.15%), as expected from a rare disease caused by a relatively ancient founder in a genetic isolate.

The autosomal-recessive HSP30 (MIM 610357) had been previously linked to a 5.1 cM interval on chromosome 2q37.3 that encompasses *KIF1A gene* (Klebe et al. 2006). The maximum multipoint LOD score was between markers D2S2338 and D2S2585 (chr2: 238,514,684-242,575,273). The four reported HSP30 patients presented in adolescence with spastic gait, distal wasting, sensory neuropathy and cerebellar ataxia with mild diffuse cerebellar atrophy. In our patients, the disease presented in infancy and has only involved spastic gait without additional abnormalities. As noted above, it has been shown that different lesions in same HSP associated genes can create a range of phenotypes and account for both pure and complicated forms of HSP. Therefore, our data could suggest that *KIF1A* accounts for HSP30.

Characterization of KIF1A

KIF1A consists of 48 exons that encode a 1791 amino acid kinesin (**Figure 3a**). Residues 1-367 of the protein form the motor domain and Ala 255 is part of a nine amino acids stretch that is fully conserved throughout fungi, nematode, insect, and vertebrates (**Figure 3b**). It is adjacent to the tip of loop L11 (Kull et al. 1996), a key structural element in the catalytic core of the protein (Hirokawa et al. 2009). **Figure 3c** presents the location of Ala 255 on a 3D model of *KIF1A* motor domain based on the PDB entry 1VFZ (Nitta et al. 2004).

Discussion

We used genetic, functional, and disease-network analysis arguments to associate a mutation in a novel gene, *KIF1A*, with a pure form of HSP cases in a single inbred family. Our combined approach enabled us to examine *all* of the translated positions in the genome, including positions that were not covered due to incomplete design of the capture array, and to systematically reject all but one position in *KIF1A*.

Kinesins are a large superfamily of molecular motors; they use microtubules as a ‘rail’ to transport cargo along, and chemical energy of ATP to drive conformational changes that generate motile force. Active transport of proteins along directional cytoskeletal filaments to their appropriate destination using molecular motors is most prominent in polarized cells, including neurons, and is fundamental for neuronal function and survival because most of the proteins required in the axon and nerve terminals need to be transported from the cell body. There are 45 mammalian *KIF* genes and the encoded proteins are classified into 15 kinesin families, based on phylogenetic analyses; *KIF1A* belongs to the kinesin 3 family. The 15 families are grouped according to the position of the motor domain in the molecule. Hence, *KIF1A*

is an N-kinesin as its motor domain, which associates with microtubules, is in the aminoterminal region. Like other N-kinesins, *KIF1A* drives the microtubule plus end towards the axon terminal, thus powering anterograde transport. At the axon terminus, neurotransmitters containing synaptic vesicles are produced by endocytosis. These vesicles contain proteins that have been transported to the plasma membrane in synaptic vesicle precursors. The kinesin 3 family motors *KIF1A* and *KIF1Bβ* transport synaptic vesicle precursors that contain the synaptic vesicle proteins, synaptophysin, synaptotagmin and the small GTPase RAB3A (Hall and Hedgecock 1991; Okada et al. 1995; Hirokawa et al. 2009). At the axon terminal, RAB3 controls the exocytosis of synaptic vesicles.

Kinesins have long been implicated in the pathogenesis of axonal neuropathies. A loss-of-function mutation in the ATP binding consensus of *KIF1B* motor domain was associated with Charcot-Marie-Tooth neuropathy type 2A (Zhao et al. 2001). Mutations in the motor domain of *KIF5A* account for SPG1038 and heterozygous missense mutations in *KIF21A* are the cause of congenital fibrosis of the extraocular muscle type 1 (Yamada et al. 2003). A mutation in *unc-104*, the *C. elegans* ortholog of *KIF1A*, is associated with decreased transport of synaptic vesicle precursors in the axons (Otsuka et al. 1991). Knockout of *Kif1a* in mice is lethal soon after birth due to massive axonal and neuronal cell body degeneration at the central nervous system, accompanied by motor and sensory disturbances, more pronounced in the hindlimbs. Transport of synaptic vesicle precursor proteins, not only those carried by *Kif1a*, is decreased. This is evident by the abnormal clustering of vesicles in the cell bodies of the neurons, their reduced accumulation in ligation experiments, and the reduced number of synaptic vesicles. This finding was unexpected, as the transport of synaptic plasma membrane precursors seems intact. Nonetheless, the reduction in synaptic vesicles was associated with a decrease in nerve terminal number (Yonekawa et al. 1998). Knockout of the *Kif1b* gene in mice had a similar phenotype, with lethality in the perinatal period and severe neuronal degeneration and synaptic dysfunction (Zhao et al. 2001). Mice that were heterozygous for the *Kif1b* gene are viable, but suffer from progressive peripheral neuropathy. These studies and the present report, suggest that although redundantly transporting the same cargoes, *Kif1a* and *Kif1b* cannot compensate for each other; an abnormally low level of either one of them results in neuronal phenotype.

Disease-network analysis adds a complementary layer of information to the sequencing data. It provides a means to integrate prior knowledge on the expected molecular signature of the disease from closely associated conditions and to reject bystander variations. The disease-network analysis approach has been mainly used to prioritize gene lists as the first step of candidate gene studies. In this study, we

used gene prioritization as the final step, after exhausting genetic arguments. We found that the analysis is robust across different algorithms and random subsets of training disease genes.

The OMIM database describes more than 200 Mendelian conditions that show locus heterogeneity with a mixture of known and unknown loci. Our combined approach provides a rich layer of information for medical sequencing of those conditions.

Methods

Human Samples

All experiments involving DNA of the patients and their relatives were approved by the Hadassah Ethical Review Committee.

Illumina Sequencing

Blood derived DNA samples were fragmented and immortalized by ligation to standard Illumina adapters, followed by PCR-based enrichment (Hodges et al. 2009). Sequences corresponding to human exons were enriched by hybridization to an Agilent 1 Million Feature Array (Ng et al. 2009) and sequenced using the 76 paired-end standard Illumina kit.

We used Bowtie (Langmead et al. 2009) to align the reads to the human genome reference version NCBI36/hg18. We employed the following strategy to align the reads: align pair-end reads allowing up to 2 mismatches and suppress multiple mappers; pair-end reads that failed to align in the previous round were broken to single end reads and re-aligned again allowing up to 2 mismatches and no multiple mapprers; reads that failed to align in the previous round were trimmed by 10bp and realigned again. We repeated the trimming and realignment up to a length of 36nt, while suppressing multiple mappers in every step. In total, we accumulated 3.6Gbase of sequencing information for the affected child, 4.8Gbase for the mother, and 2.2Gbase for the father. The average fold coverage of patient II-5 was 31 and the median was 25 for autosomal coding regions as defined in refseq hg18; the average fold coverage of the mother was 57 and the median was 49; the average fold coverage of the father was 20 and the median was 17 (**Supplemental Figure 3**)

We called SNPs with SNVmix (Goya et al. 2010) using the default parameters and we instructed the program to also report positions without variations using a small change in the source code (available by request from the authors). Reads that failed to align with Bowtie were re-aligned with BWA allowing up to 3 indels that are not 5bp from the end of the read (BWA aln –i5 –n3). We used VarScan (Koboldt et al. 2009) to call indels using the following command parameters: --min-coverage 3 --

min-var-freq 0.3 --min-reads2. Homozygous indels were called only when 90% of the sequence reads reported the non-reference allele.

By comparing the genotyping results of array to the sequencing data of patient II-5, we estimate that the sensitivity of calling non-reference homozygous SNPs was 99% and 94% for heterozygous SNP. The false discovery rate was 1.8% for homozygous SNPs and 1.4% for heterozygous SNPs.

Finding Identical Segments and Homozygous Regions using Genotype Arrays

In order to find identical segments, we compared the genotyping results of the two brothers using a sliding window of 100 SNPs (~1cM), allowing up to 3% errors. Overlapping windows that passed the threshold were merged into one segment. We used plink (Purcell et al. 2007) to identify homozygous SNPs using the following parameters: --homozyg --homozyg-group --homozyg-window-het 0. To enhance the accuracy, we excluded discordant SNPs in the identical segments from the input ped file.

Exclusion Process.

Positions that were homozygous in the parents (wild-type or mutated) were excluded from the analysis. In order to increase the sensitivity, we used parental data to extrapolate about 450 positions that were called as heterozygous in one of the parents and were not covered in the patient. In all those cases, we assumed a worst case scenario, and treated them as homozygous variations in the patient. All of these variations were excluded based on other criteria.

SeattleSeq Annotation (http://gvs.gs.washington.edu/SeattleSeqAnnotation/HelpAbout.jsp) was used to analyze the variations. Endeavour (http://homes.esat.kuleuven.be/~bioiuser/endeavour/tool/endeavourweb.php) was trained with all possible features except BLAST. Toppgene (http://toppgene.cchmc.org/prioritization.jsp) and SUSPECTS (http://www.genetics.med.ed.ac.uk/suspects/) were used with the default training parameters.

Carrier Rate Determination

We employed the TaqMan Allelic Discrimination method with the forward primer 5'-AAATCAGCCTGGTGGACCTG-3', reverse primer 5'-CCTGGCCCCTACCTTGAG-3', and the following reporter probes: wild-type, VIC probe 5'-TGGAGTCAGCCCGCTC-3'; mutant, FAM probe 5'-TGGAGTCAACCCGCTC-3'. The analysis was performed on a 7900HT Real-Time PCR System with SDS version 2.3 software (Applied

Biosystems). With this method, heterozygosity for the mutation was detected in three of 573 anonymous individuals of the same ethnic origin.

Access to the Sequencing and Genotyping Datasets

The datasets of this study are available on dbGAP (http://www.ncbi.nlm.nih.gov/gap) and http://cancan.cshl.edu/hsp/. Please refer to the website for the Terms and Conditions.

Acknowledgments

We thank Melissa Gymrek for useful comments. This work was supported by a kind gift from Kathryn W. Davis and by the Joint Research Fund of the Hebrew University and Hadassah Medical Organization. Y.E is an Andrea and Paul Heafy Family Fellow of the Whitehead Institute for Biomedical Research. G.J.H is an investigator of the Howard Hughes Medical Institute.

Figure Legends:

Figure 1: Pedigree of the affected family. Sex of the non-affected individuals was randomized to prevent identification of the family.

Figure 2:
Loss of function prediction – *KIF1A* scores in all the tools. Orange – the subset of variations that were predicted as harmful by SIFT. Pink – harmful variations by PolyPhen. Green – harmful variations by MutationTaster. Blue – High quality variations based on SNVmix scores. The HDBLP variation was later proved to be a sequencing error. *KIF1A* is the only harmful variation.

Figure 3:
Ala255Val is a mutation in the protein motor area of *KIF1A*. (a) Schematic representation of *KIF1A* gene, the exons that encode the motor domain (green), and the location of the mutation. (b) The amino acid sequence of *KIF1A* homologs in the vicinity of Ala255 (yellow). Positions that are labeled with a star are fully conserved between human to fungi (c) A 3D model of *KIF1A* motor domain. Ala 255 is highlighted. The nucleotide binding pocket (N.B.P) and the magnesium stabilizer are found in close proximity to the mutation.

Supplemental Figure Legends:

Supplemental Figure 1: Sanger sequencing excludes the putative SNP in *HDLBP*. The putative variation in *HDLBP* (chr2: 241827802) was called with very low confidence using the high throughput sequencing data. Sanger sequencing of patient II-5 confirmed that the position (arrow) is intact.

Supplemental Figure 2: Sanger sequencing validates the *KIF1A* mutation. The upper panel presents the sequencing results of a healthy control. The middle panel shows a homozygous variation in location chr2: 241371863 of patient II-5. The lower panel shows that the mother is a carrier of the variation.

Supplemental Figure 3: Sequencing coverage of hg18 autosomal coding regions. (a) Patient II-5 (b) Father (c) Mother.

Tables:

Table 1:

Chr.	Start	Stop	Homo. SNPs	Size	Number of genes	Coding positions
2	240,066,688	242,650,580	132	2,584K	36	61,483
3	103,303,758	104,720,622	125	1,417K	1	1,373
2	234,755,156	235,930,278	111	1,175K	2	3,506
10	100,073,394	101,203,270	119	1,129K	5	10,226
			Total	**6,306K**	**44**	76,588

Table 1: Homozygous regions larger than >1,000K that are shared between II-4 and II-5**.**

Table 2:

Exclusion Method	Total		Sequenced	
	Genes	Positions	Genes	Positions
Candidates	44	76588	41	68718
...not homozygous WT	40	8083	35	213
… AND parental is not homozygous	24	7232	17	39
… AND not in dbSNP/1000 Genomes	18	7184	7	12
… AND not synonymous changes	17	7028	6	8
… AND conserved	15	5098	5	5
… disease-network analysis	**1 (*KIF1A*)**	**1 (*KIF1A*)**	-	-
... loss of function analysis	-	-	**1(*KIF1A*)**	**1 (*KIF1A*)**

Table 2: A description of the rejection process of the positions in the homozygous regions. Total – all possible positions, including positions without sequencing information. Sequenced – only positions that were covered in patient II-5.

Table 3:

Gene name	OMIM	Phenotype
CYP7B1	603711	SPG5, Bile acid synthesis defect
HSPD1	118190	SPG13, hypomelinating leukodystrophy
KIAA0196	610657	SPG8
KIF5A	602821	SPG10
NIPA1	608145	SPG6
PLP1	300401	SPG2, Pelizaeus-Merzbacher disease
PNPLA6	603197	SPG39
REEP1	609139	SPG31
SPAST	604277	SPG4
ATL1	606439	SPG3A
ZFYVE27	610243	SPG33

Table 3: A list of genes that are known to be associated with pure type of HSPs and used to obtain the disease signature.

Table 4:

Gene	Rank			
	SUSPECTS	**Toppgene**	**Endeavour**	**Combined**
KIF1A	***1***	***1***	***1***	***1***
D2HGDH	2	2	3	2
ATG4B	5	3	5	3
HDLBP	3	11	4	4-5
PASK	9	7	2	
ING5	4	8	8	6
SNED1	12	5	6	7
DTYMK	10	9	7	8
OR6B2	11	10	9	9-12
OR6B3	13	6	11	
AQP12B	7	12	13	
AQP12A	6	13	14	
LOC728846	14	4	15	13-14
THAP4	8	14	12	
LOC643905	15	15	10	15

Table 4: The rank of the candidate genes in different disease prediction algorithms and the combined results. *KIF1A* was scored as the top candidate in all 3 prediction algorithms.

References:

Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS. 2005. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics* **6**: 55.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods* **7**(4): 248-249.

Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B et al. 2006. Gene prioritization through genomic data fusion. *Nat Biotechnol* **24**(5): 537-544.

Bittles A. 2001. Consanguinity and its relevance to clinical genetics. *Clin Genet* **60**(2): 89-98.

Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, Nickerson DA. 2003. Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat Genet* **33**(4): 518-521.

Chen J, Xu H, Aronow BJ, Jegga AG. 2007. Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics* **8**: 392.

Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloglu A, Ozen S, Sanjad S et al. 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* **106**(45): 19096-19101.

Clark AG. 1999. The size distribution of homozygous segments in the human genome. *Am J Hum Genet* **65**(6): 1489-1492.

Cooper GM, Goode DL, Ng SB, Sidow A, Bamshad MJ, Shendure J, Nickerson DA. 2010. Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat Methods* **7**(4): 250-251.

Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**(7): 901-913.

Dion PA, Daoud H, Rouleau GA. 2009. Genetics of motor neuron disorders: new insights into pathogenic mechanisms. *Nat Rev Genet* **10**(11): 769-782.

Edvardson S, Shaag A, Kolesnikova O, Gomori JM, Tarassov I, Einbinder T, Saada A, Elpeleg O. 2007. Deleterious mutation in the mitochondrial arginyl-transfer RNA synthetase gene is associated with pontocerebellar hypoplasia. *Am J Hum Genet* **81**(4): 857-862.

Feldman I, Rzhetsky A, Vitkup D. 2008. Network properties of genes harboring inherited disease mutations. *Proc Natl Acad Sci U S A* **105**(11): 4323-4328.

Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. 2007. The human disease network. *Proc Natl Acad Sci U S A* **104**(21): 8685-8690.

Goya R, Sun MG, Morin RD, Leung G, Ha G, Wiegand KC, Senz J, Crisan A, Marra MA, Hirst M et al. 2010. SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics* **26**(6): 730-736.

Hall DH, Hedgecock EM. 1991. Kinesin-related gene unc-104 is required for axonal transport of synaptic vesicles in C. elegans. *Cell* **65**(5): 837-847.

Hirokawa N, Nitta R, Okada Y. 2009. The mechanisms of kinesin motor motility: lessons from the monomeric motor KIF1A. *Nat Rev Mol Cell Biol* **10**(12): 877-884.

Hodges E, Rooks M, Xuan Z, Bhattacharjee A, Benjamin Gordon D, Brizuela L, Richard McCombie W, Hannon GJ. 2009. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat Protoc* **4**(6): 960-974.

Klebe S, Azzedine H, Durr A, Bastien P, Bouslam N, Elleuch N, Forlani S, Charon C, Koenig M, Melki J et al. 2006. Autosomal recessive spastic paraplegia (SPG30) with mild ataxia and sensory neuropathy maps to chromosome 2q37.3. *Brain* **129**(Pt 6): 1456-1462.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**(17): 2283-2285.
Krawitz PM, Schweiger MR, Rodelsperger C, Marcelis C, Kolsch U, Meisel C, Stephani F, Kinoshita T, Murakami Y, Bauer S et al. 2010. Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. *Nat Genet*.
Kull FJ, Sablin EP, Lau R, Fletterick RJ, Vale RD. 1996. Crystal structure of the kinesin motor domain reveals a structural similarity to myosin. *Nature* **380**(6574): 550-555.
Lander ES, Botstein D. 1987. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* **236**(4808): 1567-1570.
Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**(3): R25.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.
Ng PC, Henikoff S. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**(13): 3812-3814.
Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC et al. 2010a. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* **42**(9): 790-793.
Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA et al. 2010b. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**(1): 30-35.
Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE et al. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**(7261): 272-276.
Nitta R, Kikkawa M, Okada Y, Hirokawa N. 2004. KIF1A alternately uses two loops to bind microtubules. *Science* **305**(5684): 678-683.
Okada Y, Yamazaki H, Sekine-Aizawa Y, Hirokawa N. 1995. The neuron-specific kinesin superfamily protein KIF1A is a unique monomeric motor for anterograde axonal transport of synaptic vesicle precursors. *Cell* **81**(5): 769-780.
Otsuka AJ, Jeyaprakash A, Garcia-Anoveros J, Tang LZ, Fisk G, Hartshorne T, Franco R, Born T. 1991. The C. elegans unc-104 gene encodes a putative kinesin heavy chain-like protein. *Neuron* **6**(1): 113-122.
Pierce SB, Walsh T, Chisholm KM, Lee MK, Thornton AM, Fiumara A, Opitz JM, Levy-Lahad E, Klevit RE, King MC. 2010. Mutations in the DBP-deficiency protein HSD17B4 cause ovarian dysgenesis, hearing loss, and ataxia of Perrault Syndrome. *Am J Hum Genet* **87**(2): 282-288.
Reid E. 2003. Science in motion: common molecular pathological themes emerge in the hereditary spastic paraplegias. *J Med Genet* **40**(2): 81-86.
Ropers HH. 2007. New perspectives for the elucidation of genetic disorders. *Am J Hum Genet* **81**(2): 199-207.
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**(3): 559-575.
Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. 2010. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* **7**(8): 575-576.
Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN. 2003. Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* **21**(6): 577-581.
Via M, Gignoux C, Burchard EG. 2010. The 1000 Genomes Project: new opportunities for research and social challenges. *Genome Med* **2**(1): 3.

Walsh T, Shahin H, Elkan-Miller T, Lee MK, Thornton AM, Roeb W, Abu Rayyan A, Loulus S, Avraham KB, King MC et al. 2010. Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82. *Am J Hum Genet* **87**(1): 90-94.

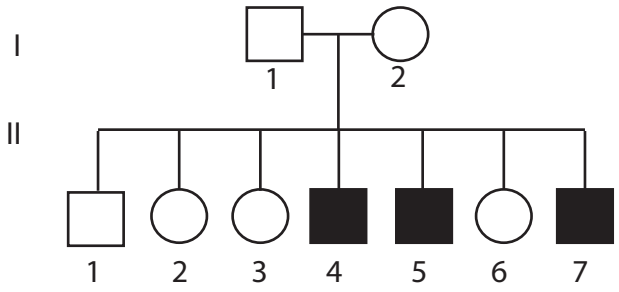
Woods CG, Cox J, Springell K, Hampshire DJ, Mohamed MD, McKibbin M, Stern R, Raymond FL, Sandford R, Malik Sharif S et al. 2006. Quantification of homozygosity in consanguineous individuals with autosomal recessive disease. *Am J Hum Genet* **78**(5): 889-896.

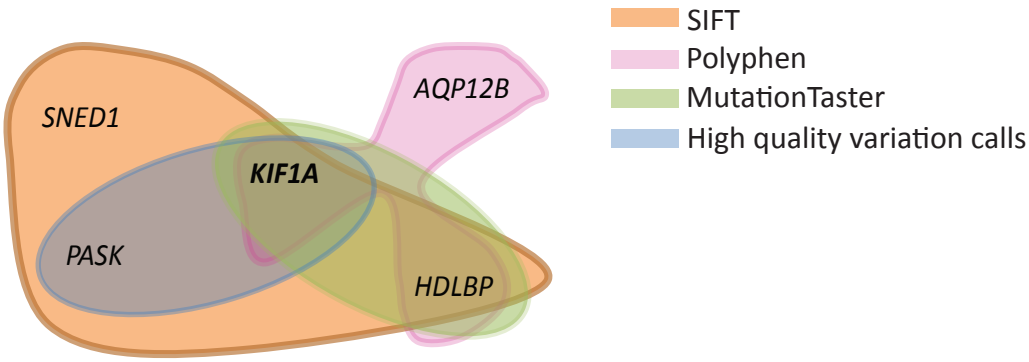
Wu X, Jiang R, Zhang MQ, Li S. 2008. Network-based global inference of human disease genes. *Mol Syst Biol* **4**: 189.

Yamada K, Andrews C, Chan WM, McKeown CA, Magli A, de Berardinis T, Loewenstein A, Lazar M, O'Keefe M, Letson R et al. 2003. Heterozygous mutations of the kinesin KIF21A in congenital fibrosis of the extraocular muscles type 1 (CFEOM1). *Nat Genet* **35**(4): 318-321.

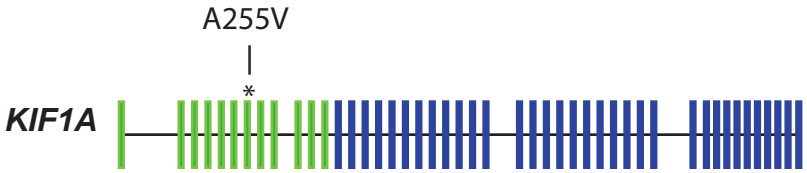
Yonekawa Y, Harada A, Okada Y, Funakoshi T, Kanai Y, Takei Y, Terada S, Noda T, Hirokawa N. 1998. Defect in synaptic vesicle precursor transport and neuronal cell death in KIF1A motor protein-deficient mice. *J Cell Biol* **141**(2): 431-441.

Zhao C, Takita J, Tanaka Y, Setou M, Nakagawa T, Takeda S, Yang HW, Terada S, Nakata T, Takei Y et al. 2001. Charcot-Marie-Tooth disease type 2A caused by mutation in a microtubule motor KIF1Bbeta. *Cell* **105**(5): 587-597.

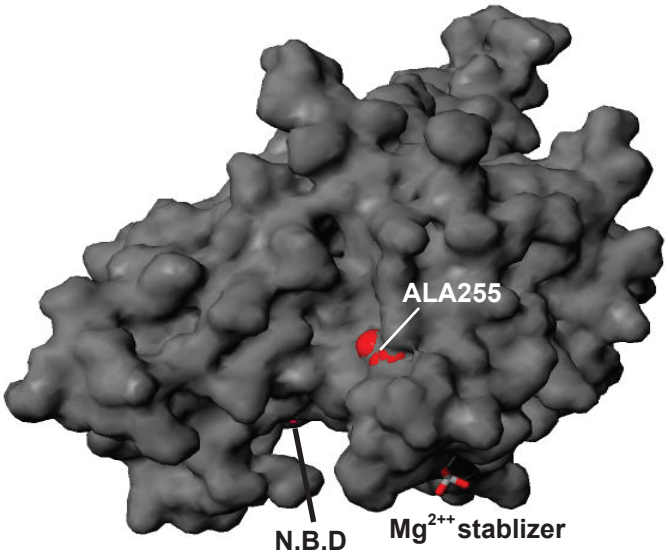

I
II
1
2
1
2
3
4
5
6
7


SNED1
AQP12B
KIF1A
PASK
HDLBP
SIFT
Polyphen
MutationTaster
High quality variation calls

a



b



c

```
Fully conserved positions      *           **  * ********** **** * ****
Homo sapiens                   QKRHDAETNITTEKVSKISLVDLAGSERADSTGAKGTRLKE
Bos taurus                     QKRHDAETNITTEKVSKVSLVDLAGSERADSTGAKGTRLKE
Mus musculus                   QKRHDAETNITTEKVSKISLVDLAGSERADSTGAKGTRLKE
Gallus gallu                   QKRHDAETDITTEKVSKISLVDLAGSERADSTGAKGTRLKE
Danio rerio                    QKQHDNDSENTTEKVSKISLVDLAGSERADSTGAKGTRLKE
Drosophila melanogaster        QRRHDLMTNLTTEKVSKISLVDLAGSERADSTGAKGTRLKE
Anopheles gambiae              QKRQDRMTSLETEKVSKISLVDLAGSERADSTGAKGTRLKE
Caenorhabditis elegans         QKRHCADSNLDTEKHSKISLVDLAGSERANSTGAEGQRLKE
Magnaporthe oryzae             QKSFDVETNMAMEKVAKISLVDLAGSERATSTGATGARLKE
Neurospora crass               QKRFDPETKMEMEKAAKISLVDLAGSERATSTGATGARLKE
```