Supplemental Methods and Data

A novel missense mutation in *CCDC88C* activates the JNK pathway and causes a dominant form of spinocerebellar ataxia

Ho TSOI, Allen C.S. YU, Zhefan S. CHEN, Nelson K.N. NG, Anne Y.Y. CHAN, Liz Y.P. YUEN, Jill M. ABRIGO, Suk-Ying TSANG, Stephen K.W. TSUI, Tony M.F. TONG, Ivan F.M. LO, Stephen T.S. LAM, Vincent C.T. MOK, Lawrence K.S. WONG, Kwok-Fai LAU, Jacky C.K. NGO, Ting-Fung CHAN and H.Y. Edwin CHAN

Supplemental Methods

Steps in filtering and prioritizing variants

A union set of 328,328 raw variants was identified among the six individuals using GATK UnifiedGenotyper 2.5. The raw variants contain a significant portion of off-target calls, which are outside of the 62 Mb targeted regions as defined by Illumina Truseq Exome Enrichment kit [1]. Sequences outside of the targeted enrichment regions have very low sequencing coverage, therefore the variants within off-target regions were filtered out due to poor quality.

77,841 variants remained upon removing the off-target variant calls. We then proceed to perform variant quality filtering according to the version 4 of GATK best practice for variant detection [2], finally obtaining 70,047 filtered variants. Snpeff [3] was used to annotate the genes (Ensembl GRCh37 release 75) affected by the variants, as well as their predicted functional impact and minor allele frequency in population genetics databases. Following linkage analysis, 615 variants were identified within the regions that showed linkage (see below).

Among the 615 variants in linked regions, 34 of them matched the observed autosomal dominant inheritance pattern (Figure 1A). The Illumina Truseq Exome Enrichment kit does cover 88.3% of exons, promoters, UTRs, microRNAs, and other noncoding RNAs as recorded in RefSeq. We revisited the variant calls and found no evidence of any non-common variants (MAF < 0.005) within the promoter, UTRs, microRNAs, or other noncoding RNAs that fits the observed co-segregation pattern.

Accordingly, synonymous mutations and noncoding mutations were discarded while missense,

splice site mutations, and insertions/deletions were kept, obtaining 13 variants for further analysis. To further exclude common variants, which are unlikely to be causative, we excluded variants with a minor allele frequency greater than 0.005 according to online databases including dbSNP (version 138) [4], 1000 Genomes Project (phase I release version 3) [5], HapMap release 28 [6], and NHLBI Exome Sequencing Project (ESP6500SI-V2) [7]. Variants labelled as pathogenic in these databases were not removed. Only three heterozygous candidate variants remained after this filtering step (Table S4).

Workflow of linkage analysis

Linkdatagen [8] was used to select SNP markers from the 70,047 filtered variants for genetic linkage analysis. Default parameters of Linkdatagen were used, except for changing the population to "Han Chinese in Beijing" from HapMap phase 2. The tool returned 7,443 markers with an average heterozygosity of 0.45 and a frequency of 1 per 0.3 cM, followed by genetic linkage analysis using MERLIN [9]. The genetic map coordinates (cM) was converted to hg18 genome coordinates by linear interpolation of HapMap phase 2 genetic map, followed by lifting over to hg19 genome coordinates by using UCSC liftOver tool. Four linkage regions with log of odds (LOD) score > 2 were identified and they are chr11: 70342417-78811197, chr14:87749314-92755402, chr18:67083133-74211893, and chr20:1189334-2796007.

Supplemental Data

Table S1. Sequencing statistics of the whole-exome sequencing data

	II:1	II:3	II:4	II:5	II:7	II:8
Total reads (in millions)	87.48	59.07	55.92	65.13	62.80	53.74
Total bases (Gbp)	8.84	5.97	5.65	5.64	6.34	5.43
Total post-filtered reads (in millions)	82.36	51.32	48.50	56.06	54.19	46.39
Total post-filtered bases (Gbp)	8.30	5.17	4.89	5.65	5.46	4.67

Statistics of sequencing data from Axeq Technologies before and after reads filtering are listed. To improve accuracy of genotyping, adapter sequences, low quality terminal bases, ambiguous bases, and un-paired singletons were removed from raw sequencing data, using fastq-mcf [10].

Table S2. Quality metrics for exome variant identification of the 6 samples

	II:1	II:3	II:4	II:5	II:7	II:8
Pre-filtering count	275,651	267,977	261,247	265,231	257,778	256,405
Pre-filtering union count	328,328					
Pre-filtering on target count	77,841					
Post-filtering count	70,047					
ts/tv	2.49					
snps/indels	7.52					
Singletons	7,986					
% variants in dbSNP 138 or 1000 Genome DB (phase I release version 3)	96.81					

Abbreviations: ts/tv: Transition/Transversion ratio, which is an indicator of variant filtration effectiveness. N/S: Nonsynonymous / Synonymous variants ratio. Singletons: Found in single sample only. Variant filtering was based on version 4 of the Genome Analysis Toolkit best practice for variant detection.

Table S3. Post-filtering variant prioritization steps

Filters	Number of variants
Located within linked regions (LOD ≥ 2)	615
Conformed to dominant inheritance pattern	34
Missense mutations, splice site mutations, or insertions/ deletions	13
A. Not a common variant	3
B. Predicted to be pathogenic	1
C. Expression patterns	1
D. PubMed search	1
A ∩ B ∩ C ∩ D	1

A—Minor Allele Frequency > 0.005 in dbSNP (version 138), 1000 Genomes Project (phase I release version 3), HapMap release 28, or NHLBI Exome Sequencing Project (ESP6500SI-V2)

B—All functional predictors (SIFT [11], Mutationassessor [12], MutationTaster [13], Polyphen-2 [14] and PROVEAN [15]) predicted the variant as probably damaging or damaging

C—Transcripts Per Million (TPM) > 0 in the brain EST profile of NCBI UniGene.

D—PubMed search of neurologically related citations.

A ∩ B ∩ C ∩ D—Intersection of filter condition A, B, C and D

Table S4. Candidate mutations of SCA40

Location	Ref	Obs	Gene	Mutation	Mutation Assessor	Mutation Taster	Polyphen2 (HumVar)	PROVEAN	SIFT	ABA	HBT	UniGene
chr11: 74413901	C	T	*CHRDL2*	NM_015424: c.G1058A:p.R353H	N (1.590)	N (0.692519)	D (0.910)	N (-1.371)	D (0.030)	1.80	5.96	0
chr14: 90651043	G	A	*KCNK13*	NM_022054: c.G923A:p.R308Q	N (0.895)	N (0.999994)	N (0.006)	N (-1.303)	N (0.146)	2.06	5.25	0
chr14: 91787600	C	T	*CCDC88C*	NM_001080414: c.G1391A:p.R464H	P (2.885)	D (1.0)	P (0.557)	D (-3.813)	D (0.001)	3.30	6.24	2

Abbreviations: Ref—Reference allele, Obs—Observed Allele, ABA—Allen Brain Atlas, HBT—Human Brain Transcriptome, UniGene—brain EST profile of the NCBI UniGene build 236 database.

Functional impact scores were classified as N—Neutral, P—Probably disease causing or D—Disease causing according to the documentation of individual tools. For MutationAssessor [12], PROVEAN [15] and SIFT [11], numbers in the brackets denotes the prediction scores. For MutationTaster [13] and Polyphen 2 [14], numbers in the brackets denotes the confidence of prediction, where a higher value denotes a higher confidence. For Allen Brain Atlas (ABA) and Human Brain Transcriptome (HBT), the average $\log_2$ signal intensities across all samples in cerebellum were shown. For NCBI UniGene, the normalized Transcripts Per Million (TPM) value from the brain EST profile is shown.

Figure S1. Results of parametric genetic linkage analysis

Heterozygous SNPs found in both whole-exome sequencing samples and HapMap phase 2 Chinese Han population were selected such that linkage equilibrium was attained at the frequency of 1 SNP per 0.3 cM. MERLIN was used for multipoint parametric linkage analysis, where a rare dominant disease model with disease allele frequency of 0.00001 was specified.

Figure S2. Overexpression of wildtype (WT) and mutant (MT) forms of CCDC88C, CHRDL2 and KCNK13 proteins in HEK293 cells. Myc-tagged wild type (WT) and mutant (MT) *CCDC88C* (GenBank accession number: NM_001080414), *CHRDL2* (GenBank accession number: NM_015424.4) and *KCNK13* (GenBank accession number: NM_022054) cDNA sequences were synthesized from GenScript USA Inc., and then subcloned into *pcDNA3.1* expression vector. All three sets of WT and MT expression constructs (1 µg) were independently used to transfect HEK293 cells. Cells were harvested 24 hours after transfection and the expression of the myc-tagged proteins were detected by anti-myc antibody 71D10 (1:1,000; Cell Signaling Technology). Anti-JNK 3708 (1:1,000, Cell Signaling Technology) and anti-p-JNK 5136 (1:1,000; Cell Signaling Technology) antibodies were used to detect endogenous JNK. Tubulin was used as loading control and was detected using anti-beta tubulin antibody E7 (1:10,000; Developmental Studies Hybridoma Bank). The experiment was repeated for at least three times. Only representative blots are shown.

Figure S3. Wildtype (WT) and mutant (MT) forms of *CCDC88C* induced dose-dependent phosphorylation of JNK in HEK293 cells. To overexpress CCDC88C protein in HEK293 cells, different amount (0.2 – 1.0 µg) of the *pcDNA3.1* WT and MT *CCDC88C* expression constructs

were used independently to transfect HEK293 cells. "-" denotes untransfected control. Cells were harvested 24 hours after transfection and expression of the CCDC88C proteins were detected by anti-myc antibody 71D10 (1:1,000; Cell Signaling Technology). Anti-JNK 3708 (1:1,000, Cell Signaling Technology) and anti-p-JNK 5136 (1:1,000; Cell Signaling Technology) antibodies were used to detect the total and phosphorylated form of endogenous JNK respectively. Tubulin was used as loading control and was detected using anti-beta tubulin antibody E7 (1:10,000; Developmental Studies Hybridoma Bank). The experiment was repeated for at least three times. Only representative blots are shown.

Figure S4. Overexpression of wildtype (WT) and mutant (MT) forms of *CCDC88C*, and knockdown of endogenous *CCDC88C* expression in HEK293 cells. (A) Overexpression of CCDC88C proteins in HEK293 cells. Both WT and MT *CCDC88C* expression constructs (0.5 µg) were used to transfect HEK293 cells. Cells were harvested 24 hours after transfection and expression of the CCDC88C proteins were detected by anti-myc antibody 71D10 (1:1,000; Cell Signaling Technology). (B) Small interfering RNA (siRNA) treatment reduced endogenous CCDC88C protein expression. HEK293 cells were treated with 5 pmol of ON-TARGETplus *CCDC88C* siRNA L-033364-00-0005 (Dharmacon) or control (ctrl) siRNA (Dharmacon). Protein expression of endogenous CCDC88C was detected by anti-CCDC88C antibody A302-951A (1:1,000; Bethyl Laboratories). Tubulin was used as loading control in all experiments and was detected using anti-beta tubulin antibody E7 (1:10,000; Developmental Studies Hybridoma Bank). The experiment was repeated for at least three times. Only representative blots are shown.

Figure S5. Domain organization of CCDC88C and the location of the p.R464H mutation.

CCDC88C consists of a N-terminal HOOK domain, a central coiled-coil region, a C-terminal disorder region, and a PDZ-binding motif (Gly-Cys-Val). The R464H missense mutation is located within the HOOK domain region. Domain information of the protein was fetched from Pfam [16] and Uniprot [17].

Figure S6. Subcellular localization of wildtype (WT) and mutant (MT) forms of CCDC88C proteins. HEK293 cells were transfected with 1 µg of WT or MT *CCDC88C* expression construct. Cells were harvested 24 hours after transfection and subcellular localization of the CCDC88C proteins was determined using an Olympus FV-1000IX81-TIRF confocal microscope. Primary antibody used was rabbit anti-myc antibody 71D10 (1:500; Cell Signaling Technology) and secondary antibody used was FITC-conjugated goat anti-rabbit antibody (1:500; Zymed). Nuclei were stained with Hoechst 33342 (1:400; Life Technologies). Scale bar represents 5 µm. The experiment was repeated for at least three times. Only representative images are shown.

Supplemental References:

1. Illumina. TruSeq Exome Enrichment Kit Support-Downloads. Secondary TruSeq Exome Enrichment Kit Support-Downloads 2013. http://support.illumina.com/sequencing/sequencing_kits/truseq_exome_enrichment_kit/downloads.ilmn.
2. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;**43**(5):491-8.
3. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 2012;**6**(2):80-92.
4. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001;**29**(1):308-11.
5. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature 2010;**467**(7319):1061-73.
6. International HapMap C, Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, Dermitzakis E, Bonnen PE, Altshuler DM, Gibbs RA, de Bakker PI, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Yu F, Chang K, Hawes A, Lewis LR, Ren Y, Wheeler D, Gibbs RA, Muzny DM, Barnes C, Darvishi K, Hurles M, Korn JM, Kristiansson K, Lee C, McCarrol SA, Nemesh J, Dermitzakis E, Keinan A, Montgomery SB, Pollack S, Price AL, Soranzo N, Bonnen PE, Gibbs RA, Gonzaga-Jauregui C, Keinan A, Price AL, Yu F, Anttila V, Brodeur W, Daly MJ, Leslie S, McVean G, Moutsianas L, Nguyen H, Schaffner SF, Zhang Q, Ghori MJ, McGinnis R, McLaren W, Pollack S, Price AL, Schaffner SF, Takeuchi F, Grossman SR, Shlyakhter I, Hostetter EB, Sabeti PC, Adebamowo CA, Foster MW, Gordon DR, Licinio J, Manca MC, Marshall PA, Matsuda I, Ngare D, Wang VO, Reddy D, Rotimi CN, Royal CD, Sharp RR, Zeng C, Brooks LD, McEwen JE. Integrating common and rare genetic variation in diverse human populations. Nature 2010;**467**(7311):52-8.
7. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S, Bustamante CD, Bamshad MJ, Akey JM, Broad GO, Seattle GO, Project NES. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 2012;**337**(6090):64-9.
8. Smith KR, Bromhead CJ, Hildebrand MS, Shearer AE, Lockhart PJ, Najmabadi H, Leventer RJ, McGillivray G, Amor DJ, Smith RJ, Bahlo M. Reducing the exome search space for mendelian diseases using genetic linkage analysis of exome genotypes. Genome Biol 2011;**12**(9):R85.
9. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 2002;**30**(1):97-101.
10. Aronesty E. ea-utils : "Command-line tools for processing biological sequencing data". Secondary ea-utils : "Command-line tools for processing biological sequencing data"

2011. http://code.google.com/p/ea-utils.
11. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;**31**(13):3812-4.
12. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 2011;**39**(17):e118.
13. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 2010;**7**(8):575-6.
14. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet 2013;**Chapter 7**:Unit7 20.
15. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One 2012;**7**(10):e46688.
16. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD. The Pfam protein families database. Nucleic Acids Res 2012;**40**(Database issue):D290-301.
17. UniProt C. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res 2014;**42**(Database issue):D191-8.