Human Molecular Genetics, 2019, Vol. 28, No. 14 2309–2318

doi: 10.1093/hmg/ddz063
Advance Access Publication Date: 1 April 2019
General Article

GENERAL ARTICLE

Rare variants in MYH15 modify amyotrophic lateral sclerosis risk

Hyerim Kim[1,2], Junghwa Lim[1], Han Bao[1], Bin Jiao[1], Se Min Canon[3], Michael P. Epstein[1], Keqin Xu[1], Jie Jiang[5], Janani Parameswaran[5], Yingjie Li[3], Kenneth H. Moberg[5], John E. Landers[6], Christina Fournier[3,4], Emily G. Allen[1], Jonathan D. Glass[3], Thomas S. Wingo[1,3,4,*] and Peng Jin[1,*]

[1]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322, USA, [2]Cancer Biology Program, Emory University, Atlanta, GA 30322, USA, [3]Department of Neurology, Emory University School of Medicine, Atlanta, GA 30322, USA, [4]Division of Neurology, Atlanta VA Medical Center, Decatur, GA 30033, USA, [5]Department of Cell Biology, Emory University and Emory University School of Medicine, Atlanta, GA 30322, USA and [6]Department of Neurology, University of Massachusetts Medical School, Worcester, MA 01655, USA

*To whom correspondence should be addressed at: Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322, USA. Tel: +1 4047274905 and +1 4047273729; Fax: +1 4047273728 and +1 4047275408; Email: thomas.wingo@emory.edu and peng.jin@emory.edu

Abstract

Amyotrophic lateral sclerosis (ALS) is a fatal neurological disorder characterized by progressive muscular atrophy and respiratory failure. The G_4C_2 repeat expansion in the *C9orf72* gene is the most prevalent genetic risk for ALS. Mutation carriers (C9ALS) display variability in phenotypes such as age-at-onset and duration, suggesting the existence of additional genetic factors. Here we introduce a three-step gene discovery strategy to identify genetic factors modifying the risk of both C9ALS and sporadic ALS (sALS) using limited samples. We first identified 135 candidate genetic modifiers of C9ALS using whole-genome sequencing (WGS) of extreme C9ALS cases diagnosed ∼30 years apart. We then performed an unbiased genetic screen using a *Drosophila* model of the G_4C_2 repeat expansion with the genes identified from WGS analysis. This genetic screen identified the novel genetic interaction between G_4C_2 repeat-associated toxicity and 18 genetic factors, suggesting their potential association with C9ALS risk. We went on to test if 14 out of the 18 genes, those which were not known to be risk factors for ALS previously, are also associated with ALS risk in sALS cases. Gene-based-statistical analyses of targeted resequencing and WGS were performed. These analyses together reveal that rare variants in *MYH15* represent a likely genetic risk factor for ALS. Furthermore, we show that *MYH15* could modulate the toxicity of dipeptides produced from expanded G_4C_2 repeat. Our study presented here demonstrates the power of combining WGS with fly genetics to facilitate the discovery of fundamental genetic components of complex traits with a limited number of samples.

Received: January 3, 2019. **Revised:** March 14, 2019. **Accepted:** March 21, 2019

© The Author(s) 2019. Published by Oxford University Press. All rights reserved.
For Permissions, please email: journals.permissions@oup.com

Introduction

Amyotrophic Lateral Sclerosis (ALS) is a complex neurodegenerative disease that can develop at any age, but most commonly occurs between the ages of 40 and 70 years (at a mean age of 55 years) (1). This rare neurological disorder is characterized by progressive degeneration of the upper and lower motor neurons and leads to weakness and death an average of 2–5 years after initial clinical symptoms develop (2,3). Approximately 5–20% of ALS patients exhibit a discernible family history defined as familial ALS (fALS) (4). Genetic factors are considered as obvious drivers for the pathogenesis in fALS cases (5), but a number of twin and other large-scale genomic studies have also shown a substantial genetic contribution in sporadic ALS (sALS), estimating ~60% of the heritability of sALS (6–8). Given this, many genetic studies have been conducted to understand the genetic etiology of ALS and have identified rare genetic variants in multiple genes such as superoxide dismutase 1 (*SOD1*), *FUS* and TAR DNA binding protein (*TARDBP*) (3,9).

Among the known pathogenic mutations, the recently identified hexanucleotide (G_4C_2) repeat expansion in the *C9orf72* gene is the most common genetic cause of ALS (C9ALS), although this mutation has an intermediate effect on ALS risk compared to traditional pathologic mutations (3,10). C9ALS has a wide range of phenotypic variability in terms of age-at-onset, duration and regions of motor neuron involvement (11,12), suggesting the burden of genetic variants in multiple genes may contribute to modulating ALS risk even in the patients sharing the same genetic alteration (3,13,14). However, the successful identification of novel genetic components involved in ALS pathogenesis is limited by only using genome-wide association study (GWAS) or whole-genome sequencing (WGS) if pathogenic mutations have a moderate or small effect on ALS risk (3).

In this study, we hypothesized that a gene differentially identified among C9ALS groups who have extremely distinct age-at-onset can be a novel genetic factor implicated in ALS. Using functional screening, we were able to prioritize candidate genes more biologically relevant to ALS causality. Here we performed a hypothesis-driven genetic association study using WGS to identify novel genetic candidates associated with ALS risk (step 1), followed by a genetic screen using a *Drosophila* model stably expressing the G_4C_2 repeat expansion (step 2). Prioritized candidate genes were further assessed by a candidate gene association study using sALS cases and non-ALS controls (step 3), consequently leading to the identification of rare variants in *MYH15* as a novel genetic factor of ALS. Furthermore, we show that *MYH15* could modulate the toxicity of dipeptides produced from the expanded G_4C_2 repeat. Our data together demonstrate the utility of combining WGS with fly genetics to facilitate the discovery of fundamental genetic components of complex traits with a limited number of samples.

Results

Candidate gene discovery through WGS

To facilitate the identification of novel genetic factors of ALS risk with a limited sample size, we employed a stepwise approach of candidate gene selection based on the assumption that genetic risk factors can be identified in even a small number of ALS patients who have the same G_4C_2 repeat expansion but develop clinical symptoms at different ages (Fig. 1). Therefore, in the discovery phase, we performed WGS on two distinct age-at-onset groups of four unrelated G_4C_2 repeat expansion carriers. Two of the individuals developed ALS at 31 and 41 years old [and referred to here as young ALS (YALS)], and the other two individuals developed ALS at 72 years old [and referred to as old ALS (OALS)] (Supplementary Material, Table S1). To identify disease-relevant variants, we selected rare and deleterious sites that were unique to either the YALS or OALS groups on the basis of the following criteria: variants that had a minor allele frequency (MAF) $<1\%$ in the Genome Aggregation Database (gnomAD) (15) and variants that had a combined annotation-dependent depletion (CADD) (16) score higher than 10 (17,18). In total, we identified 190 variants (159 variants from YALS, 31 variants from OALS) and 135 unique genes (105 genes from YALS and 30 genes from OALS) (Supplementary Material, Table S2).

Drosophila genetic screen

Given transgenic fly lines expressing G_4C_2 repeats display progressive neurodegeneration in eye and motor neurons similar to ALS patients (19,20), we performed a genetic screen using a transgenic fly expressing 30 repeats of G_4C_2 under a *GMR-Gal4* driver (eye-specific) as reported previously (19) and tested whether the 135 selected genes could modulate G_4C_2 repeat-associated toxicity (Fig. 1). Of the 135 genes, 89 genes (65.9%) had a functional homolog in the fly genome with at least a moderate rank score according to the *Drosophila* RNAi Screening Center (DRSC) Integrative Ortholog Prediction Tool [DIOPT, Version 6.0.2 (June 2017)] (21), (Fig. 1; Supplementary Material, Table S3). A total of 90 RNAi lines corresponding to 49 fly genes were crossed with the $(G_4C_2)_{30}$ repeat transgenic line to determine the genetic interaction between the G_4C_2 repeat and candidate genes (20) (Fig. 1, Supplementary Material, Table S4). All RNAi lines crossed with flies carrying the *GMR-Gal4* driver alone showed no pathological eye findings (data not shown). However, 11 RNAi/G_4C_2 lines suppressed the $(G_4C_2)_{30}$-related toxicity, and 7 lines showed an evident enhancement of the disrupted eye morphology accompanied by severe necrosis (Fig. 2A, Table 1). Thin-section analysis of $(G_4C_2)_{30}$ flies crossed with suppressors verified the recovery of photoreceptor cells and fewer vacuolated materials compared to the $(G_4C_2)_{30}$ flies itself (Fig. 2B). The genes identified in this screening are involved in various cellular functions including cell adhesion, DNA or RNA binding, and regulation of oxidative stress (Table 1). Interestingly, the WGS and *Drosophila* screen identified homeodomain interacting protein kinase 2 (*HIPK2*) as a candidate gene of interest (Table 1; Supplementary Material, Table S2). This gene was recently implicated in ALS neurodegeneration, lending validity to our approach (22).

Targeted resequencing of prioritized gene lists

To further understand the contribution of the genes resulting from WGS analysis and functional screening to ALS without the G_4C_2 repeats and known ALS associated genes, we tested whether individuals with sALS were enriched for rare, likely deleterious variants at selected genes compared to non-ALS controls. For a novel gene finding, we excluded 4 candidate genes which were already known for ALS association in the literature [*DBF4*, early growth response 3 (*EGR3*) and *HIPK2*] and other neurological disorders (*FXR2*), and then performed targeted resequencing that focused on exonic regions of 14 candidate genes and 5 known ALS-associated genes [(i.e. *FUS*, granulin precursor (*GRN*), *SOD1*, *TARDBP* and TANK-binding kinase 1 (*TBK1*)] in a col-

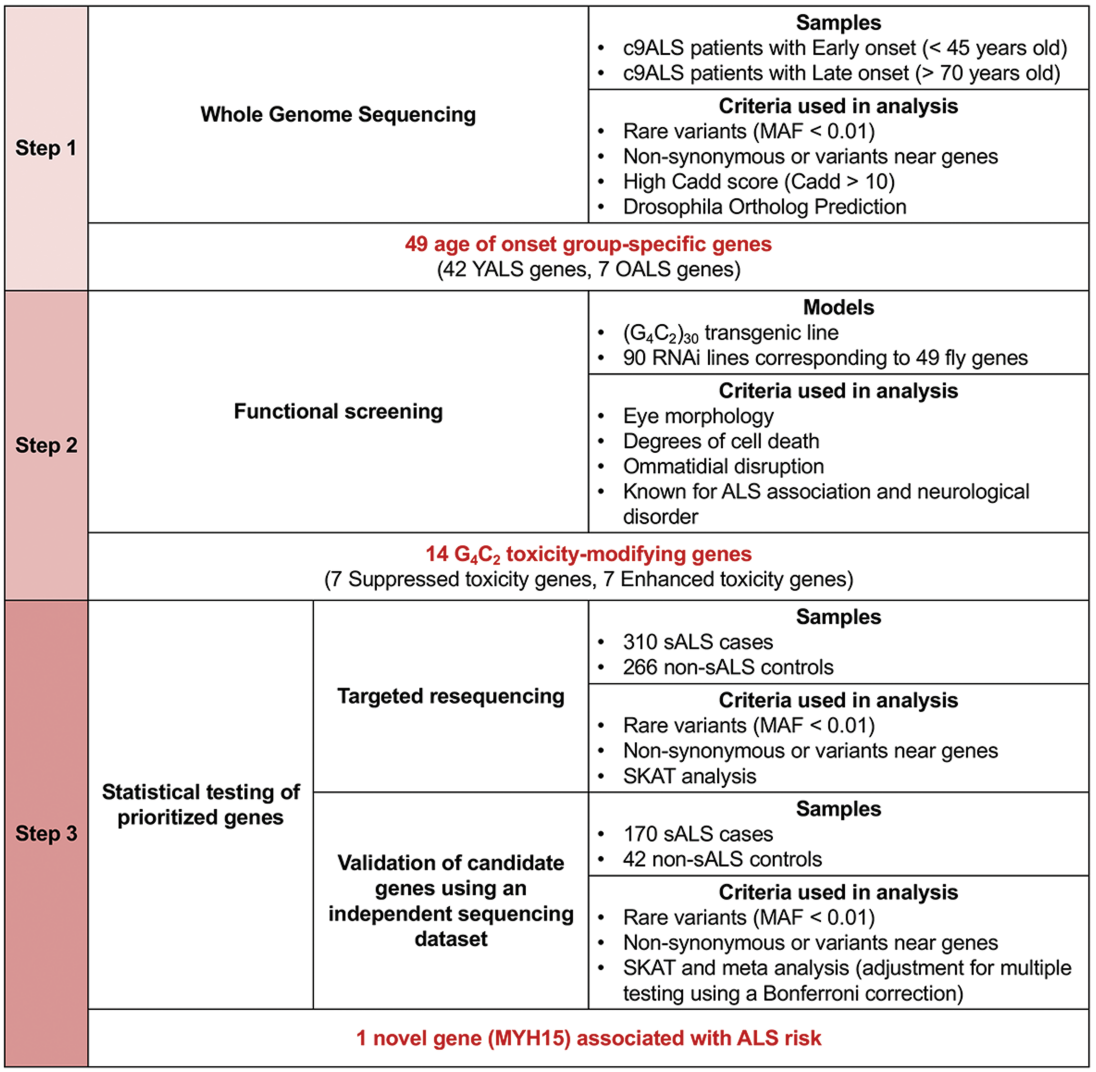

Figure 1. Three-step strategy to identify genetic factors associated with ALS risk using a hypothesis-driven and targeted genetic association study (steps 1 and 3) and fly genetics (step 2).

lection of ALS subjects who were negative for the G_4C_2 expanded repeat (12) (Fig. 1; Table 1; Supplementary Material, Table S5). After filtering outliers from principal-component analysis (PCA) and samples with low sequencing quality, 489 samples (272 sALS and 217 non-ALS) were included in the analysis (Supplementary Material, Fig. S1A and B). We filtered the 1447 variants by those predicted to cause coding changes (e.g. missense, nonsense and frameshift mutations) and having MAF $<1\%$ among controls (Supplementary Material, Fig. S1A). Two known pathogenic mutations in *SOD1* (Ile114Thr) and *TARDBP* (Gly287Ser mutation) were identified in ALS cases (*SOD1* carrier: female, the age of onset: 36 years; *TARDBP* carrier: female, the age of onset: 76.3 years) (23); these cases were excluded from further analysis. We performed sequence kernel association test (SKAT) analysis (24,25) using all variants from each gene, controlling for population structure using eigenvectors from PCA. Using bootstrap to estimate empirical *P*-values, we found two genes, *DLG2* (10 variants; $P=0.04180$) and *MYH15* (16 variants; $P=0.01950$), which showed suggestive evidence of association with ALS (Fig. 3A; Table 2; Supplementary Material, Tables S6 and S8). *MYH15* also showed suggestive association with ALS in the unified rare variant association test, optimized SKAT (SKAT-O; $P=0.03697$, Table 2).

To examine whether the *DLG2* and *MYH15* genes that showed suggestive association in the targeted resequencing dataset replicated in a validation dataset, we obtained WGS data of 212 people recruited at Emory University School of Medicine (Fig. 1). We followed the same quality control procedures as were used for the targeted resequencing (Supplementary Material, Fig. S1C and D), which resulted in the removal of four individuals (two were outliers for ancestry and two were carriers of known pathogenic mutations in *SOD1*). The same selection criteria for variants were applied, which were tested for the resequencing experiment. We performed gene-based analysis using SKAT and SKAT-O and found significant association between ALS and *MYH15* (six variants; $P=0.01233$ and 0.01708, respectively) after adjustment for multiple testing using a Bonferroni correction (Fig. 3B; Table 2; Supplementary Material, Table S7). We observed no association between ALS and *DLG2* (two variants; $P=0.30207$; Table 2; Supplementary Material, Tables S7 and S8).

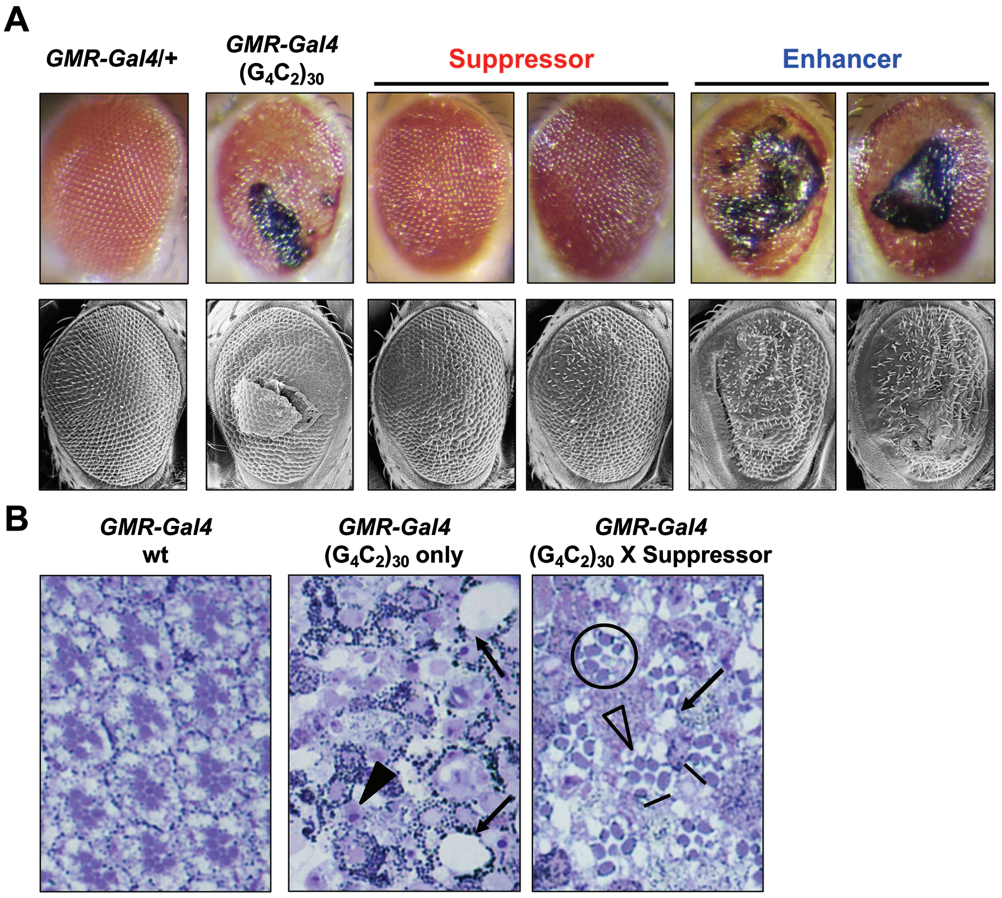

Figure 2. Functional screen identifies multiple genetic modifiers of $(G_4C_2)_{30}$ toxicity. (**A**) The expression of $(G_4C_2)_{30}$ driven by the *GMR-Gal4* driver causes rough eye phenotypes as shown in light microscope image and ommatidial disruption as shown in scanning electron microscope image. In screening the flies, we selected genes with rescued phenotypes as a suppressor and genes with aggravate phenotypes as an enhancer. (**B**) A representative thin section for GMR-Gal4 flies with either $(G_4C_2)_{30}$ alone or both the $(G_4C_2)_{30}$ and RNAi of suppressor genes. The $(G_4C_2)_{30}$ flies showed a loss of photoreceptor cells (arrowheads) and vacuolated material (arrows). However, $(G_4C_2)_{30}$ flies crossed with RNAi of suppressor exhibited rescued phenotypes regarding recovered photoreceptor cells (circle) and smaller size of vacuolated material (arrows) although there are an abnormal number of photoreceptor cells (open arrowheads) and polarity defects (bars).

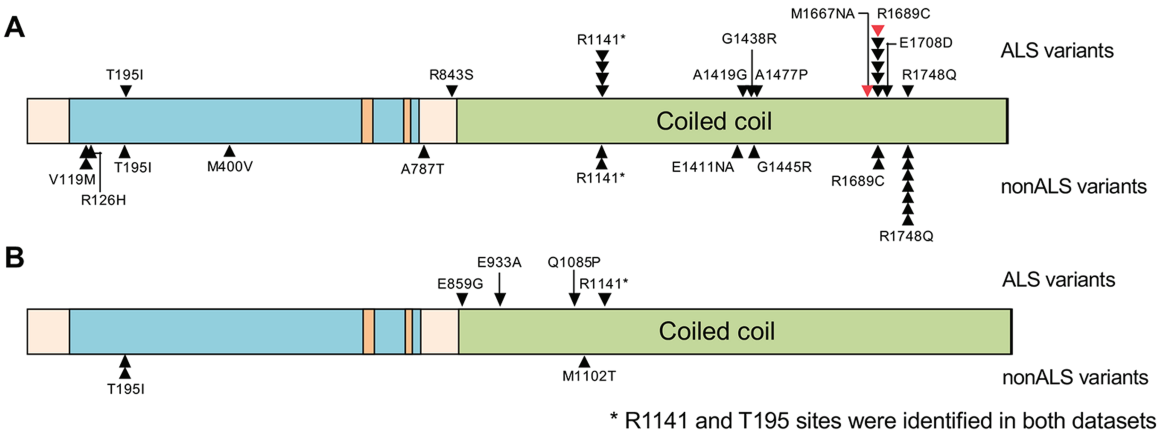

Figure 3. Coding variants of *MYH15* identified in either ALS cases or controls during the targeted resequencing (**A**) and validation dataset (**B**). Black arrow indicates heterozygous variant while red arrow indicates homozygous variant.

Meta-analysis

To improve statistical power, we performed a SKAT-based meta-analysis of the two independent datasets using the MetaSKAT (26) package. Since all samples used in this meta-analysis had the same ethnicity, we expected homogeneous genetic effects across the samples. The genomic coordinates from the WGS dataset were converted from hg19 to hg38 using LiftOver to pool individual-level genotype data from the two datasets mapped

Table 1. The 18 candidate genes in the table either suppress or enhance the neuronal toxicity, which resulted in a repeat expansion

Patient type	Gene symbol	Fly ortholog	Fly screening result	Targeted resequencing	Biological function
YALS	ABCC2	MRP	**Enhancer**	Yes	Protein transporter and regulation of oxidative stress
YALS	MYH15	Mhc	**Enhancer**	Yes	Tight junction pathway
YALS	PLEKHG2	GEFmeso	**Enhancer**	Yes	Postsynaptic signaling pathway
YALS	PPARD	Eip75B	**Enhancer**	Yes	Peroxisome
YALS	SVEP1	uif	**Enhancer**	Yes	Cell adhesion process
YALS	UTP20	CG4554	**Enhancer**	Yes	18 s rRNA processing
OALS	CDK11A	Pitslre	**Enhancer**	Yes	Cell cycle and apoptosis
YALS	CELF5	bru-3	**Suppressor**	Yes	mRNA editing and translation
YALS	DBF4	chif	**Suppressor**	No	Cell Cycle Checkpoints in DNA replication
YALS	DLG2	dlg1	**Suppressor**	Yes	Postsynaptic signaling pathway
YALS	EGR3	sr	**Suppressor**	No	Transcriptional regulator in mitogenic stimulus
YALS	FAM98B	CG5913	**Suppressor**	Yes	tRNA processing and gene expression
YALS	FXR2	Fmr1	**Suppressor**	No	RNA binding
YALS	HIPK2	Hipk	**Suppressor**	No	Endoplasmic reticulum (ER) stress
YALS	HK3	Hex-A	**Suppressor**	Yes	Metabolism
YALS	PDK3	Pdk	**Suppressor**	Yes	Metabolism
OALS	KDM2A	Kdm2	**Suppressor**	Yes	Epigenetic modification
OALS	KIF27	cos	**Suppressor**	Yes	Cytokinesis

Among the 18 genes, 14 genes which are previously unknown for ALS and other neurological association were validated by targeted resequencing.

Table 2. Gene-based analysis of rare variants for targeted resequencing dataset, replication (WGS) dataset, and meta-analysis which combines two datasets

Group	Genes	Targeted resequencing			Replication (WGS)			Meta-Analysis		
		variant number	SKAT P-value	SKAT-O P-value	variant number	SKAT P-value	SKAT-O P-value	variant number	SKAT P-value	SKAT-O P-value
Suppressors	DLG2	10	0.0418	0.09971	2	0.30207	0.30207	12	0.23962	0.37744
Enhancers	MYH15	16	0.0195	0.03697	6	0.01233	0.01708	20	0.02511	0.0472

Empirical P-values based on resampling techniques are provided.

with different assemblies (27). Consistent with SKAT results for each dataset, *MYH15* showed borderline significant association with ALS in the SKAT test (20 variants; adjusted $P = 0.02511$, Table 2). Two variants in *MYH15* were shared in *MYH15* by both datasets, one of which (rs61744539; R1141*) is a nonsense mutation potentially leading to downregulation of *MYH15* gene expression (Fig. 3; Supplementary Material, Table S8).

MYH15 modulates dipeptide-mediated toxicity associated with G4C2 repeat expansion

The G_4C_2 expansions exert neuronal toxicity through direct accumulation of G4C2 containing RNA transcripts (19,28) and repeat-associated non-AUG (RAN) translated dipeptide repeat proteins (29–31). Of the dipeptides, the arginine-rich proteins, proline-arginine (PR) and glycine-arginine (GR), leads to a significant decrease in survival and aggregate eye phenotypes in the *Drosophila* model (32). To determine whether *MYH15* could modulate PR-mediated toxicity, we performed a functional assay using a transgenic fly expressing 36 repeats of either PR or GR under *GMR-Gal4* driver (32). The knockdown of *MYH15* *Drosophila* ortholog, *Mhc*, resulted in enhancing retinal toxicity when crossing with both G_4C_2-repeat line and PR repeat line (Fig. 4A). We also observed substantial lethality in GR repeat line when crossed with *Mhc*-KD line (data not shown). In addition to a fly model, the downregulation of *Myh15* in Neuro2A cell line could enhance poly(PR)- and poly(GR)-mediated cell toxicity (33) (Fig. 4B). Recent report identified the moderate interaction between PR50 and *MYH9* as well as *MYH10* (34), consistent with our findings. Given that myosin heavy chain genes are involved in vesicle transport (35), *MYH15* can potentially modulate PR aggregate-mediated toxicity via the impairment of vesicle trafficking.

Discussion

Here we present a three-step strategy to identify ALS risk-associated genes by integrating fly genetics with WGS (Fig. 1). Our hypothesis is that genetic factors modulating phenotypic variability of G_4C_2 expansion carriers are associated with ALS risk. As such, initial candidate genes were selected by WGS on four unrelated G_4C_2 expansion carriers who developed ALS ~30 years apart, identifying 135 potential risk genes (step 1). To prioritize candidate genes from WGS of a small number of C9ALS individuals, we used a *Drosophila* genetic screen to test for genetic interactions between candidate genes and the G_4C_2 model of neurodegeneration (step 2). Through this unbiased screen, we identified novel genetic interactions as well as a known interaction with G_4C_2 toxicity (HIPK2), which supports that our approach is suitable for novel gene identification. Finally, rather than sequencing all genes most of which are

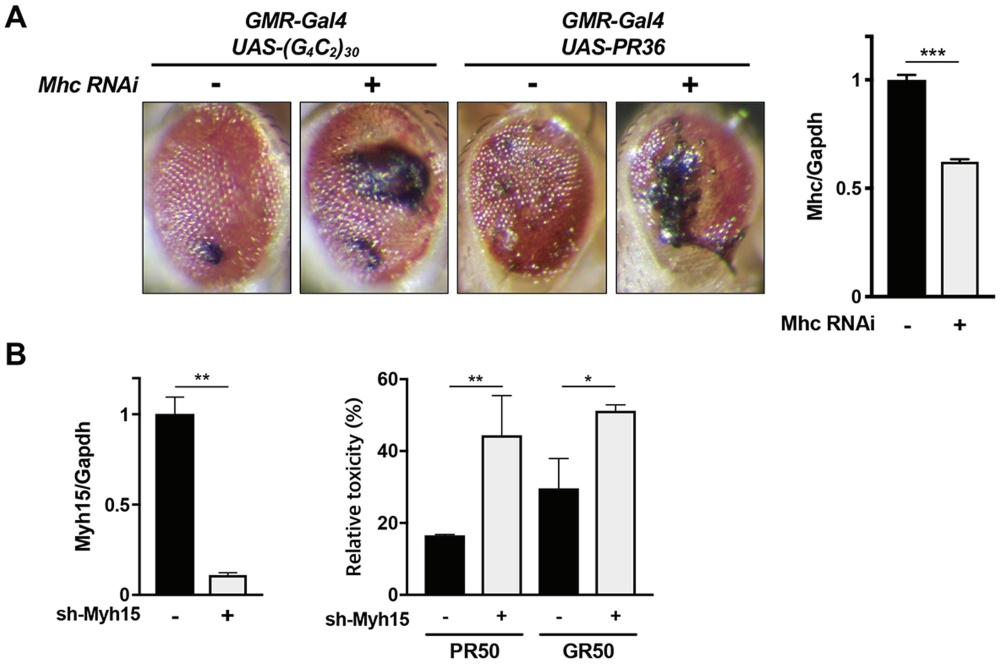

Figure 4. *MYH15* is a potential genetic modifier of dipeptide-mediated toxicity. (**A**) Left: transgenic lines expressing either $(G_4C_2)_{30}$ or $(PR)_{36}$ under *GMR-Gal4* driver cause progressive neurodegeneration in eye. Both transgenic lines displayed aggravate phenotypes when crossing with a RNAi line of *Mhc*, a drosophila ortholog of *MYH15*, implying *MYH15* can modify RNA- and dipeptide-mediated toxicity. Right: the knockdown efficiency of *Mhc* RNAi lines crossed with *ELAV-Gal4* driver was confirmed by quantitative RT–PCR (qPCR). (**B**) Left: relative *Myh15* expression after siRNA treatment (50 nM) in Neuro2A cell line. Right: relative cell viability measured 3 days after plasmid and siRNA co-transfection. Control: GFP, PR50: GFP-$(PR)_{50}$, GR50: GFP-$(GR)_{50}$.

irrelevant to ALS risk, only targeted candidate genes were analyzed to investigate their association with ALS without any known pathogenic mutations in *C9orf72*, *FUS*, *GRN*, *SOD1*, *TARDBP* and *TBK1* (step 3). Gene-based statistical testing of targeted resequencing and WGS on sALS cases and controls suggests rare variants in *MYH15* represent a likely genetic risk factor for ALS. A further functional assay revealed that *MYH15* can a genetic modifier of dipeptide-mediated toxicity of C9ALS (Fig. 4).

MYH15, myosin heavy chain 15, was a recently characterized as a slow-type myosin involved in muscle contraction and cytoskeleton remodeling (36,37). Well-known class-II *MHC* genes are considered to be divergent products from an ancestral gene through the series of gene duplications due to structural similarity of myofilaments within the same class genes (36,38). However, *MYH15*, along with *MYH14* and *MYH16*, displayed unrelated structural features from classical *MYH* genes. In particular, loop domains of *MYH15* are highly divergent; for instance, the N-terminal positive charge cluster in loop 1 is lost and there is no matched sequence in loop 2 (36). In addition, the unexpected large size of *MYH15* (>142 000 bp) provides greater possibility of having genetic variants in both exons and introns (36). Indeed, recent genetic studies identified a common male-specific association of single-nucleotide polymorphisms (SNPs) in *MYH15* with increased coronary microvascular dysfunction risk the nominal association of variants in *MYH15* with increased risk of stroke and coronary heart disease (39,40). One of the notable variants in *MYH15* identified in aforementioned studies is Thr1125Ala (rs3900940), which is located in the coiled-coil tail domain of *MYH15* (39). In our study, aside from R1748Q (rs56118396), variants identified in the ALS population of our study are distributed within the rod-like tail sequence while variants at N-terminus skew toward the non-ALS population (Fig. 3). Given that the combination of van der Walls forces and electrostatic interactions between proper amino acids is critical for home-dimerization of the tail domain, the disruption of the coiled structure resulting from nonsynonymous variants is likely associated with ALS progression.

In addition to genetic association studies in cardiovascular diseases, the association of variants in *MYH15* was investigated in a study of common mental disorders, schizophrenia and bipolar disorder (41). GWAS was performed a statistical association analysis using two independent data sets: the International Schizophrenia Consortium data set including 3322 schizophrenia cases and 3587 controls from the same ethnic population and the Genetic Association Information Network (42) data set, including 1351 cases and 1378 controls. Among associated genes involved in a tight junction pathway, a SNP in *MYH15* (rs16854665, MAF = 0.1287, gnomAD) displayed statistical significance in both studies (41), suggesting that genetic variants in *MYH15* can be associated with other brain disorders. In addition, *MYH15* is highly expressed in brain–spinal cord tissues compared to other organs (Supplementary Material, Fig. S2) (43). However, there is no previous evidence about the implication of *MYH15* variants in ALS pathogenesis. This is a first report that links *MYH15* to ALS.

In summary, we have identified *MYH15* as a potential genetic factor associated with ALS risk. Our analyses demonstrate that the combination of WGS with fly genetics facilitates the discovery of fundamental genetic components of complex traits with a limited number of samples.

Materials and Methods

Study subjects

In the discovery phase, we performed WGS on DNA samples from four unrelated patients carrying the G4C2 expansion mutation. We included two patients with an early age of onset (31.3

and 41.7 years old; YALS) and two with late age of onset (72.4 and 72.9 years old; OALS). All patients in this phase are unrelated to each other (Supplementary Material, Table S1). In the replication phase, 576 samples, including 310 sALS patients and 266 unaffected individuals, were used for targeted resequencing (Supplementary Material, Table S5). Validation of candidate genes using an independent WGS dataset was done with 170 sALS patients and 42 non-ALS controls. The protocols and consent forms for enrollment were approved by the Institutional Review Board at Emory University. Written and informed consent were obtained for all participants.

Genotyping G_4C_2 repeat size of study subjects

Genomic DNA from human white blood cells was extracted with the Gentra Puregene kit (Qiagen, Hilden, Germany) according to the manufacturer's protocols. The C9orf72 hexanucleotide repeat for study subjects was determined using the repeat-primed protocol, as described previously (12). Briefly, four primers (two forward primers, one reverse primer and a fluorescently labeled primer) were used for PCR amplification of DNA. Amplified products incorporating a fluorescently labeled primer were separated using a capillary electrophoresis DNA system (ABI3730; Thermo Fisher, Waltham, MA, USA). Based on a cutoff of 30 repeats as a positive indicator and DNA from a C9Pos control from Coriell Institute for Medical Research (6769B1), the status of C9Pos for each sample was determined using amplified fragment length polymorphism analysis in GeneMarker software (Softgenetics, State College, PA, USA).

WGS and candidate gene identification

The Hudson Alpha Institute for Biotechnology provided sequencing services (Huntsville, AL, USA). Raw sequencing data were aligned to the hg38 build of the human genome using PEMapper, and variants called using PECaller with default settings (18). Variant annotation and summary sequencing statistics were performed using Bystro (44). To identify candidate ALS phenotypic modifying genes, rare genetic variants (MAF < 0.01) commonly found in either YALS or OALS were considered along with a CADD phred-scaled score above 10 (16). We selected 89 candidate modifiers (67 genes from YALS and 22 genes from OALS), which have *Drosophila* orthologues searched by DRSC DIOPT (21), to determine the genetic interaction between G_4C_2 repeat-associated ALS and genes with rare, potentially damaging variants in both groups of patients (Supplementary Material, Table S3).

Genetic screen using fly model

The G_4C_2 repeat stable line was established by crossing a *GMR-Gal4* driver with a UAS-$(G_4C_2)_{30}$ repeat transgene (19). The RNAi lines were obtained from either the Bloomington Stock Center or the Vienna *Drosophila* RNAi Center (Supplementary Material, Table S4). The knockdown efficiency of RNAi lines crossed with the *ELAV-Gal4* driver was measured by quantitative RT–PCR (qPCR) (Fig. 4A, right). To determine the genetic interaction between G_4C_2 repeat and candidate genes, eye phenotypes of RNAi lines mated with the G_4C_2 repeat stable line were compared with the eye phenotype of the G_4C_2 repeat stable line, and images were obtained by light microscopy. Eye phenotypes in the figure are representative images of functional screening. All crosses were conducted at 25°C, replicated three times to validate the specific phenotype, and a minimum of 10 flies were used to determine phenotypic change. Scanning electron microscopy images of whole flies were obtained after dehydrating them in an ethanol gradient (25%, 50%, 75% and 100%) followed by incubation with hexamethyldisilazane for 1 h (Electron Microscopy Sciences, Hatfield, PA, USA). After the removal of all chemicals by drying overnight in the fume hood, the flies were coated with argon gas under an electric field and analyzed with a Topcon DS-130F and DS-150F Field Emission Scanning Electron Microscope. For further morphological analysis to confirm the recovery of organized ommatidia in the case of RNAi showing suppressed toxicity when crossing with the G_4C_2 repeat stable line, thin-section analysis of adult *Drosophila* eyes was conducted according to standard protocols (45). In brief, fly heads were exposed to 2% glutaraldehyde in 0.1 M PO_4 on ice followed by 2% OsO_4 in 0.1 M PO_4 on ice. After dehydration in an ethanol gradient (30%, 50%, 70%, 80%, 90% and 100%), 100% ethanol was replaced with propylene oxide, and an equal volume of resin was added. The fixed heads were transferred to a silicone rubber flat mold for embedding with resin. One µm sections were mounted on glass slides and stained with toluidine blue.

Targeted resequencing

Genomic DNA was extracted from white blood cells of 310 ALS patients and 266 non-ALS subjects using the Gentra Puregene kit (Qiagen, Hilden, Germany) according to the manufacturer's protocols. For targeted resequencing, two sets of primers were designed by using the multiplex primer design software with >90% coverage for each gene (46) and optimal multiplex design for the Access Array System (Fluidigm, South San Francisco, CA, USA). The first set was designed to capture 14 candidate genes including *DLG2*, *MYH15*, *KIF27* and *ABCC2*, and 5 known ALS genes (*GRN*, *SOD1*, *FUS*, *TARDBP* and *TBK1*) (Table 1). The second set covered 400 ancestrally informative and 25 common X chromosome markers. The samples were randomly plated concerning affectation, sex and age to minimize batch effects. Sequence capture was performed using the Access Array with 48 samples per batch according to the manufacturer's protocol. All samples were barcoded according to the manufacturer's protocol and 250 bp paired-ended sequencing was performed on an Illumina MiSeq.

Base calling and quality control

Mapping of raw targeted resequencing reads to the hg38 of the human genome was performed with PEMapper followed by variant calling using PECaller with default values (18). Variants were annotated and summarized using Bystro (44). As a quality control, samples with apparently different ethnicity according to demographic information were removed ($n = 18$) (Supplementary Material, Fig. S1A). Using unlinked ancestrally informative markers for PCA with EIGENSOFT, we excluded samples whose eigenvectors were >6 standard deviation (SD) away from the mean ($n = 25$) (Supplementary Material, Fig. S1A and B) (44). Samples within batches having amplicons with >3 SD missing sites and batches with >3 SD sample failure were eliminated from further analysis. Moreover, samples with >3 SD excess heterozygosity or genotype rate <95% were further dropped ($n = 44$) (Supplementary Material, Fig. S1A). Two samples with known ALS-associated mutations in *TARDBP* and *SOD1* were excluded from further analysis as well. In total, 270 ALS and 217 non-ALS samples were used for further statistical

analysis. Variants that failed Hardy–Weinberg filtering at 10^{-7} and >1% of MAF were excluded.

Genotype identification of target genes from replication dataset

The WGS replication dataset is based on whole genome sequencing with approximately 100 maxdepth of coverage for each chromosome and mapped to hg19 of the human genome. Individual 212 vcf files were obtained and combined using bcftools with −0 flag. Variants of merged vcf files were intersected using intersectBed, with genomic regions of interest converted from the hg38 to the hg19 using LiftOver to obtain variants of targeted genes and PCA markers (27). A total of 571 variants for PCA markers and 1294 variants for targeted genes were identified resulting in 208 samples for analysis. Through PCA, samples of outliers ($n=2$) with known ALS-associated mutations in *SOD1* were excluded from the further analysis (Supplementary Material, Fig. S1C and D).

Statistical analysis

We performed gene-based testing of rare variants in the targeted resequencing and the whole genome sequencing datasets using SKAT and SKAT-O implemented in the R package SKAT v1.2.1. We adjusted for population stratification by incorporating the top 2 eigenvectors from PCA as covariates within the analysis. For multi-allelic sites, the two minor alleles were combined to convert the site to a bi-allelic site using a custom R script ($n=9$) prior to analysis. Since our genetic interaction screening with a *Drosophila* model is based on the interaction of genes, not regulatory regions, we focused our analysis on those that alter coding sequence including missense or nonsense changes. In addition, we employed a CADD phred-scaled score above 20 (16), which is a more rigorous way to identify genetic variants leading to protein changes. We derived *P* values for SKAT/SKAT-O adjusted for a sample size of less than 2000 and binary traits with asymptotic and efficient resampling methods (24,25). In the targeted resequencing project, we used an unadjusted Type-I error rate of 0.05 to identify genes with suggestive evidence of association with ALS risk. For those genes passing this suggestive threshold, we interrogated replication using Emory ALS WGS dataset and identified those genes significantly associated with ALS risk using a Type-I error rate adjusted for multiple testing based on a Bonferroni correction.

For Meta-analysis for the gene-based association test, we used MetaSKAT (26), v0.6.0, with individual level genotype data of targeted resequencing and Emory ALS WGS dataset. We adjusted for the top 2 eigenvectors of PCA within each dataset. The genomic coordinates of the Emory ALS WGS dataset was converted from hg19 to hg38 using LiftOver to unify genotype assembly (27).

Toxicity assay using poly-dipeptides constructs

To determine the role of *Myh15* in mammalian system, we generated the constructs expressing 50 repeats of either PR or GA peptides (33). CellTiter-Blue Cell viability Assay (Promega, Madison, WI, USA) was used to assess cell viability on 3 days after transfection with poly-dipeptide constructs (150 ng) and siRNA (50 nM). Briefly, 20 µl of solution was added to each well directly 1 h before measurement. The fluorescence was measured using FLUOstar Omega (BMG Labtech, Ortenberg, Germany) microplate reader. All measurements were taken in triplicate and each experiment was replicated at least three times.

Data availability

Targeted resequencing data from this study have been deposited in the NCBI Sequence Read Archive under accession number SRP136672. Whole genome sequencing data and supporting data are available on request from the corresponding author.

Supplementary Material

Supplementary Material is available at *HMG* online.

Acknowledgements

We thank Hong Yi and Jeannette V. Taylor (The Robert P. Apkarian Integrated Electron Microscopy Core (IEMC), Emory University) for helping Scanning Electron Microscope (SEM) imaging of *Drosophila* eyes. We also appreciate core members of the Emory Integrated Genomic core (EIGC) for running MiSeq and helpful discussion about sample preparation. Support for patient recruitment and genetic analysis is provided to JDG and JEL by the ALS Association and the Muscular Dystrophy Association.

Conflict of Interest statement. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Veterans Health Administration.

Funding

This work was supported in part by the National Institutes of Health (NS051630, NS091859 and NS097206 to P.J.; NS073873 to J.E.L.; AG056533 to T.S.W.; AG025688), the Veterans Health Administration (BX001820 to T.S.W.) and the American ALS Association (to J.E.L. and J.D.G).

References

1. Taylor, J.P., Brown, R.H. Jr. and Cleveland, D.W. (2016) Decoding ALS: from genes to mechanism. *Nature*, **539**, 197–206.
2. Robberecht, W. and Philips, T. (2013) The changing scene of amyotrophic lateral sclerosis. *Nat. Rev. Neurosci*., **14**, 248–264.
3. Al-Chalabi, A., van den Berg, L.H. and Veldink, J. (2017) Gene discovery in amyotrophic lateral sclerosis: implications for clinical management. *Nat. Rev. Neurosci*., **13**, 96–104.
4. Cirulli, E.T., Lasseigne, B.N., Petrovski, S., Sapp, P.C., Dion, P.A., Leblond, C.S., Couthouis, J., Lu, Y.F., Wang, Q., Krueger, B.J. *et al*. (2015) Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science*., **347**, 1436–1441.
5. Byrne, S., Heverin, M., Elamin, M., Bede, P., Lynch, C., Kenna, K., MacLaughlin, R., Walsh, C., Al Chalabi, A. and Hardiman, O. (2013) Aggregation of neurologic and neuropsychiatric disease in amyotrophic lateral sclerosis kindreds: a population-based case-control cohort study of familial and sporadic amyotrophic lateral sclerosis. *Ann. Neurol*., **74**, 699–708.
6. Wingo, T.S., Cutler, D.J., Yarab, N., Kelly, C.M. and Glass, J.D. (2011) The heritability of amyotrophic lateral sclerosis in a clinically ascertained United States research registry. *PLoS One*, **6**, e27985.

7. Al-Chalabi, A., Fang, F., Hanby, M.F., Leigh, P.N., Shaw, C.E., Ye, W. and Rijsdijk, F. (2010) An estimate of amyotrophic lateral sclerosis heritability using twin data. *J. Neurol. Neurosurg. Psychiatry*, **81**, 1324–1326.
8. McLaughlin, R.L., Vajda, A. and Hardiman, O. (2015) Heritability of amyotrophic lateral sclerosis: insights from disparate numbers. *JAMA Neurol.*, **72**, 857–858.
9. Geevasinga, N., Menon, P., Ozdinler, P.H., Kiernan, M.C. and Vucic, S. (2016) Pathophysiological and diagnostic implications of cortical dysfunction in ALS. *Nat. Rev. Neurol.*, **12**, 651–661.
10. Haeusler, A.R., Donnelly, C.J. and Rothstein, J.D. (2016) The expanding biology of the C9orf72 nucleotide repeat expansion in neurodegenerative disease. *Nat. Rev. Neurosci.*, **17**, 383–395.
11. Renton, A.E., Chio, A. and Traynor, B.J. (2014) State of play in amyotrophic lateral sclerosis genetics. *Nat. Neurosci.*, **17**, 17–23.
12. Umoh, M.E., Fournier, C., Li, Y., Polak, M., Shaw, L., Landers, J.E., Hu, W., Gearing, M. and Glass, J.D. (2016) Comparative analysis of C9orf72 and sporadic disease in an ALS clinic population. *Neurology*, **87**, 1024–1030.
13. Chi, S., Jiang, T., Tan, L. and Yu, J.T. (2016) Distinct neurological disorders with C9orf72 mutations: genetics, pathogenesis, and therapy. *Neurosci. Biobehav. Rev.*, **66**, 127–142.
14. Pang, S.Y., Hsu, J.S., Teo, K.C., Li, Y., Kung, M.H.W., Cheah, K.S.E., Chan, D., Cheung, K.M.C., Li, M., Sham, P.C. *et al.* (2017) Burden of rare variants in ALS genes influences survival in familial and sporadic ALS. *Neurobiol. Aging*, **58**, 238 e9–238 e15.
15. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285–291.
16. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J. and Kircher, M. (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, **47**, D886–D894.
17. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M. and Shendure, J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
18. Johnston, H.R., Chopra, P., Wingo, T.S., Patel, V., International Consortium on Brain and Behavior in 22q11.2 Deletion Syndrome, Epstein, M.P., Mulle, J.G., Warren, S.T., Zwick, M.E. *et al.* (2017) PEMapper and PECaller provide a simplified approach to whole-genome sequencing. *Proc. Natl. Acad. Sci. USA*, **114**, E1923–E1932.
19. Xu, Z., Poidevin, M., Li, X., Li, Y., Shu, L., Nelson, D.L., Li, H., Hales, C.M., Gearing, M., Wingo, T.S. *et al.* (2013) Expanded GGGGCC repeat RNA associated with amyotrophic lateral sclerosis and frontotemporal dementia causes neurodegeneration. *Proc. Natl. Acad. Sci. USA*, **110**, 7778–7783.
20. Freibaum, B.D., Lu, Y., Lopez-Gonzalez, R., Kim, N.C., Almeida, S., Lee, K.H., Badders, N., Valentine, M., Miller, B.L., Wong, P.C. *et al.* (2015) GGGGCC repeat expansion in C9orf72 compromises nucleocytoplasmic transport. *Nature*, **525**, 129–133.
21. Hu, Y., Flockhart, I., Vinayagam, A., Bergwitz, C., Berger, B., Perrimon, N. and Mohr, S.E. (2011) An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinformatics*, **12**, 357.
22. Lee, S., Shang, Y., Redmond, S.A., Urisman, A., Tang, A.A., Li, K.H., Burlingame, A.L., Pak, R.A., Jovicic, A., Gitler, A.D. *et al.* (2016) Activation of HIPK2 promotes ER stress-mediated Neurodegeneration in amyotrophic lateral sclerosis. *Neuron*, **91**, 41–55.
23. Kabashi, E., Valdmanis, P.N., Dion, P., Spiegelman, D., McConkey, B.J., Vande Velde, C., Bouchard, J.P., Lacomblez, L., Pochigaeva, K., Salachas, F. *et al.* (2008) TARDBP mutations in individuals with sporadic and familial amyotrophic lateral sclerosis. *Nat. Genet.*, **40**, 572–574.
24. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.
25. Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., Christiani, D.C., Wurfel, M.M. and Lin, X. (2012) Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.*, **91**, 224–237.
26. Lee, S., Teslovich, T.M., Boehnke, M. and Lin, X. (2013) General framework for meta-analysis of rare variants in sequencing association studies. *Am. J. Hum. Genet.*, **93**, 42–53.
27. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
28. Kumar, V., Hasan, G.M. and Hassan, M.I. (2017) Unraveling the role of RNA mediated toxicity of C9orf72 repeats in C9-FTD/ALS. *Front. Neurosci.*, **11**, 711.
29. Wen, X., Tan, W., Westergard, T., Krishnamurthy, K., Markandaiah, S.S., Shi, Y., Lin, S., Shneider, N.A., Monaghan, J., Pandey, U.B. *et al.* (2014) Antisense proline-arginine RAN dipeptides linked to C9ORF72-ALS/FTD form toxic nuclear aggregates that initiate in vitro and in vivo neuronal death. *Neuron*, **84**, 1213–1225.
30. Lopez-Gonzalez, R., Lu, Y., Gendron, T.F., Karydas, A., Tran, H., Yang, D., Petrucelli, L., Miller, B.L., Almeida, S. and Gao, F.B. (2016) Poly(GR) in C9ORF72-related ALS/FTD compromises mitochondrial function and increases oxidative stress and DNA damage in iPSC-derived motor neurons. *Neuron*, **92**, 383–391.
31. Mori, K., Weng, S.M., Arzberger, T., May, S., Rentzsch, K., Kremmer, E., Schmid, B., Kretzschmar, H.A., Cruts, M., Van Broeckhoven, C. *et al.* (2013) The C9orf72 GGGGCC repeat is translated into aggregating dipeptide-repeat proteins in FTLD/ALS. *Science.*, **339**, 1335–1338.
32. Mizielinska, S., Gronke, S., Niccoli, T., Ridler, C.E., Clayton, E.L., Devoy, A., Moens, T., Norona, F.E., Woollacott, I.O.C., Pietrzyk, J. *et al.* (2014) C9orf72 repeat expansions cause neurodegeneration in Drosophila through arginine-rich proteins. *Science.*, **345**, 1192–1194.
33. Zhang, Y.J., Gendron, T.F., Grima, J.C., Sasaguri, H., Jansen-West, K., Xu, Y.F., Katzman, R.B., Gass, J., Murray, M.E., Shinohara, M. *et al.* (2016) C9ORF72 poly(GA) aggregates sequester and impair HR23 and nucleocytoplasmic transport proteins. *Nat. Neurosci.*, **19**, 668–677.
34. Lee, K.H., Zhang, P., Kim, H.J., Mitrea, D.M., Sarkar, M., Freibaum, B.D., Cika, J., Coughlin, M., Messing, J., Molliex, A. *et al.* (2016) C9orf72 dipeptide repeats impair the assembly, dynamics, and function of membrane-less organelles. *Cell*, **167**, 774–788e17.
35. Hirokawa, N., Niwa, S. and Tanaka, Y. (2010) Molecular motors in neurons: transport mechanisms and roles in brain function, development, and disease. *Neuron*, **68**, 610–638.
36. Desjardins, P.R., Burkman, J.M., Shrager, J.B., Allmond, L.A. and Stedman, H.H. (2002) Evolutionary implications of three

novel members of the human sarcomeric myosin heavy chain gene family. *Mol. Biol. Evol.*, **19**, 375–393.
37. Barany, M. (1967) ATPase activity of myosin correlated with speed of muscle shortening. *J. Gen. Physiol.*, **50**, 197–218.
38. Cope, M.J., Whisstock, J., Rayment, I. and Kendrick-Jones, J. (1996) Conservation within the myosin motor domain: implications for structure and function. *Structure*, **4**, 969–987.
39. Luke, M.M., O'Meara, E.S., Rowland, C.M., Shiffman, D., Bare, L.A., Arellano, A.R., Longstreth, W.T. Jr., Lumley, T., Rice, K., Tracy, R.P. *et al.* (2009) Gene variants associated with ischemic stroke: the cardiovascular health study. *Stroke*, **40**, 363–368.
40. Yoshino, S., Cilluffo, R., Best, P.J., Atkinson, E.J., Aoki, T., Cunningham, J.M., de Andrade, M., Choi, B.J., Lerman, L.O. and Lerman, A. (2014) Single nucleotide polymorphisms associated with abnormal coronary microvascular function. *Coron. Artery Dis.*, **25**, 281–289.
41. O'Dushlaine, C., Kenny, E., Heron, E., Donohoe, G., Gill, M., Morris, D., International Schizophrenia Consotrium and Corvin, A. (2011) Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility. *Mol. Psychiatry*, **16**, 286–292.
42. Ha Thi, B.M., Campolmi, N., He, Z., Pipparelli, A., Manissolle, C., Thuret, J.Y., Piselli, S., Forest, F., Peoc'h, M., Garraud, O. *et al.* (2014) Microarray analysis of cell cycle gene expression in adult human corneal endothelial cells. *PLoS One*, **9**, e94349.
43. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., GTEx Consortium, Nicolae, D.L. *et al.* (2015) A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.*, **47**, 1091–1098.
44. Shetty, A.C., Athri, P., Mondal, K., Horner, V.L., Steinberg, K.M., Patel, V., Caspary, T., Cutler, D.J. and Zwick, M.E. (2010) SeqAnt: a web service to rapidly identify and annotate DNA sequence variations. *BMC Bioinformatics*, **11**, 471.
45. Moberg, K.H., Bell, D.W., Wahrer, D.C., Haber, D.A. and Hariharan, I.K. (2001) Archipelago regulates Cyclin E levels in Drosophila and is mutated in human cancer cell lines. *Nature*, **413**, 311–316.
46. Wingo, T.S., Kotlar, A. and Cutler, D.J. (2017) MPD: multiplex primer design for next-generation targeted sequencing. *BMC Bioinformatics*, **18**, 14.