Table 2 Mutation and clinical summary of Japanese HSP patients

	Mutation	*Amino-acid change*	*Detecting method*	*Family history*	*Age at onset of index patient*	*Clinical phenotype*		*Reference*
SPG4								
Exon 1	c.139 A>T	p.K47*	R	AD	40s	Pure	Novel	
Exon 1	c.155 A>G	p.Y52C	R	Sporadic with consanguineous parents	49 y.o.	Pure	Novel	
Exon 1	c.283_323 del	p.A95Afs	D	AD	40 y.o.	Pure	Novel	
Exon 1	c.343_352 dup	p.V118Afs	D	AD	35 y.o.	Pure	Novel	
Exon 2	c. 422–425 delAGAA	p.Q141fs	D	AD	36 y.o.	Pure	Novel	
Exon 2	c. 422–425 delAGAA	p.Q141fs	D	AD	51 y.o.	Pure	Novel	
Exon 2	c. 422–425 delAGAA	p.Q141fs	D	Sporadic	35 y.o.	Pure	Novel	
Exon 3	c.532 C>T	p.Q178*	R	AD	33 y.o.	Pure	Novel	
Exon 5	c.734 C>G	p.S245*	R	AD	teens	Pure	Known	35
Exon 5	c.838 C>T	p.Q280*	R	AD	~6 y.o.	Pure	Novel	
Exon 6	c.871 delG	p.G291Vfs	D	AD	20 y.o.	Pure	Novel	
Intron 6	c.1005-2 A>G	IVS6-2A>G	R	AD	2 y.o.	Pure	Known	4
Exon 7	c. 1014 delT	p.A338Afs	D	AD	40s	Pure	Novel	
Exon 8	c.1105 A>C	p.T369P	R	AD	38 y.o.	Pure	Novel	
Exon 8	c.1141 C>T	p.F381L	R	Sporadic	<6 y.o.	Pure	Known	4
Exon 8	c.1141 C>T	p.F381L	R	AD	late 50s	Pure	Known	4
Intron 8	c.1173+1 G>A	IVS8+1G>A	R	AD	46 y.o.	Pure	Known	3
Exon 11	c.1378 C>T	p.R460C	R	AD	27 y.o.	Pure	Known	36
Exon 12	c.1426_1427 delGG	p.G476Rfs	D	AD	39 y.o.	Pure	Novel	
Intron 12	c.1493+2 T>C	IVS12+2T>C	R	Sporadic	40 y.o.	Pure	Known	35
Exon 13	c.1504 A>T	p.K502*	R	AD	30 y.o.	Pure	Novel	
Exon 13	c.1507 C>T	p.R503W	R	AD	~ 10 y.o.	Pure	Known	37
Exon 15	c.1646 insT	p.L549Lfs	D	AD	34 y.o.	Pure	Novel	
Exon 15	c.1646 insT	p.L549Lfs	D	AD	47 y.o.	Pure	Novel	
Exon 15	c.1646 T>C	p.L549P	R	AD	< 15 y.o.	Pure	Novel	
Exon 15	c.1688 G>A	p.R562Q	R	AD	~10 y.o.	Pure	Known	38
Exon 17	c.1741 C>T	p.R581*	R	AD	14 y.o.	Pure	Known	39
Promoter~intron 1	del Chr2:32136286–32145830 (9545 bp)	Del ex1	aCGH	AD	40 y.o.	Pure	Novel	
Intron 1~3′ region	(>170 kb)	Del ex2-17	aCGH	Affected sibling	24 y.o.	Pure	Novel	
Intron 16~3′ region	del Chr2:32290425–32231940 (58 482 bp)	Del ex17	aCGH	AD	58 y.o.	Pure	Novel	
Intron 16~3′ UTR	del Chr2:32229622–32234715 (5094 bp)	Del ex17	aCGH	AD	52 y.o.	Pure	Novel	
Exon 4~intron 7	dup Chr2:32177411–32199467 (22 057 bp)+insAGT	Tandem duplication (part of ex4-ex7)	aCGH	AD	< 6 y.o.	Pure	Novel	
SPG3A								
Exon 12	c.1243 C>T	p.R415W	R	AD	12 y.o.	Pure	Known	40
Exon 12	c.1483 C>T	p.R495W	R	Sporadic	~12 y.o.	Pure	Known	41
SPG8								
Exon 13	c.1749 A>C	p.R583S	R	AD	50 y.o.	Pure	Novel	
Intron 10~exon15	del Ch8:126 138 189–126 142 822 (4634 bp)	Del exon11-15	aCGH	AD	64 y.o.	Pure	Novel	
SPG17								
Exon 2	c. 107 G>A (c. 299 G>A)	p.C36Y (p.C100Y)	R	AD	~10 y.o.	Complicated	Novel	
SPG31								
Exon 2	c.87 insA	p.K30Kfs	D	AD	8 y.o.	Pure	Novel	
Intron 3~intron 5	del Ch2: 86 326 358–86 338 428 (12 064 bp)	Del exon 4-5	aCGH	AD	~12 y.o.	Pure	Novel	

Table 2 (Continued)

	Mutation	*Amino-acid change*	*Detecting method*	*Family history*	*Age at onset of index patient*	*Clinical phenotype*	*Reference*
SPG11							
Intron 18	c.3291+1 G>T	IVS18+1G>T (homozygous)	D	Consanguinity affected siblings	20 y.o.	Complicated Known	42
Intron 18	c.3291+1 G>T	IVS18+1G>T (homozygous)	D	Consanguinity affected sibling	25 y.o.	Complicated Known	42
Intron 8 and intron 38	c.1735+2 delT, c.6999+5 delG	IVS8+2 delT, IVS38+5 delG	D, D	Sporadic	22 y.o.	Complicated Novel	
Exon 20 and exon 28	c.3491G>A, c.4840T>A	p.W1164*, p.K1614*	D, D	Sporadic	18 y.o.	Complicated Novel	
Intron 7~intron 8 and exon 25	del Chr15:42709367–42715955 (6589 bp), c.4426 insAT	Del exon8, p.C1476Yfs	aCGH, D	Affected sibling	2 y.o.	Complicated Novel	
SPG21							
Exon 4	c.322 G>C	p.A108P (homozygous)	R	Familial	60 y.o.	Complicated Novel	

Abbreviations: aCGH, array-based comparative genomic hybridization analysis; AD, autosomal dominant; D, direct nucleotide sequence analysis; Del, deletion; dup, duplication; y.o., years old; R, resequencing microarray analysis; UTR, untranslated region.
All the patients presented the pure form. +1 of nucleotides is the first A of the start codon (ATG). The NCBI36/hg18 assembly is used as the reference genome.

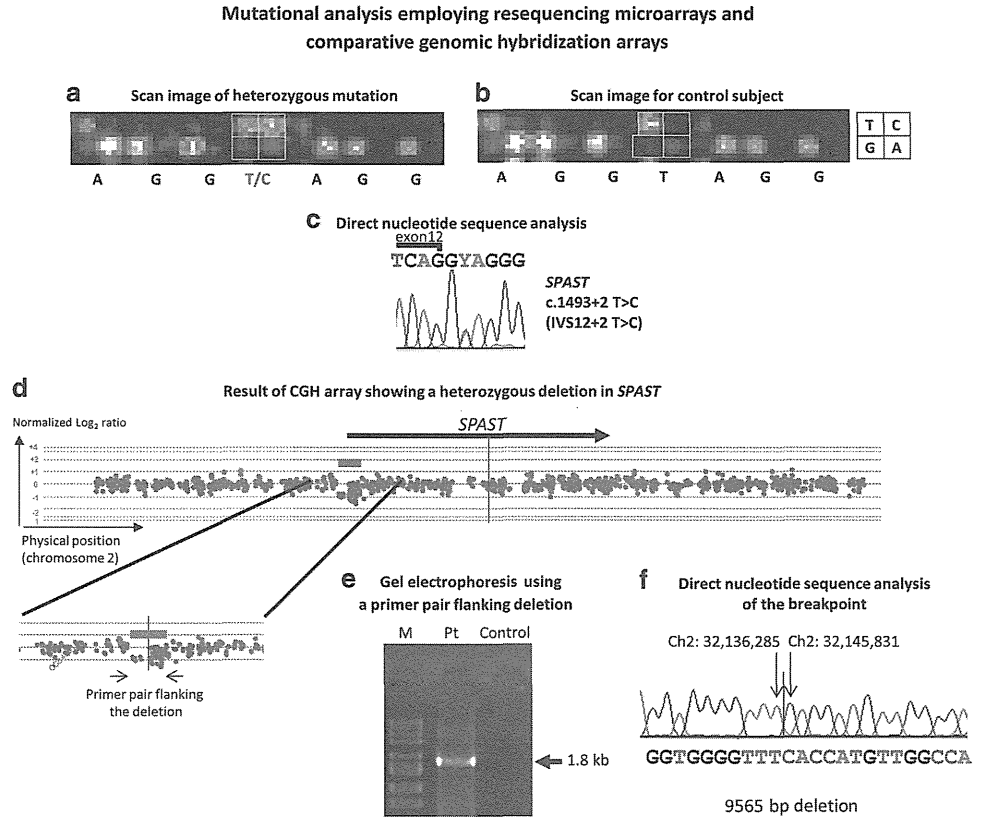

Figure 2 Mutational analysis using resequencing microarrays and comparative genomic hybridization microarrays. (**a**) This figure shows a scan image obtained by resequencing microarray analysis (TKYPD03) of a sporadic HSP patient. Each tile in a square indicates one of the four nucleotides. Depending on the nucleotide of each allele, each quadrant provides a fluorescent signal. As shown in a square that corresponds to the position of c.1493+2 of *SPAST*, the upper left tile and the upper right tile, which correspond to T and C, respectively, provided similarly intense hybridization signals. The signal pattern indicates the existence of the T allele (wild type) and the C allele (variant) in that position. (**b**) Scan image of the same positions of the resequencing microarray as those in panel (**a**) obtained from a mutation-negative patient, where only the upper left tile corresponding to 'T' gives an intense fluorescent signal. (**c**) Heterozygous c.1493+2 T>C mutation confirmed by direct nucleotide sequence analysis, which is expected to disrupt the consensus splice donor site. (**d**) Example of comparative genomic hybridization analysis. The vertical axis indicates the $\log_2$ ratio of hybridization signal intensities obtained from a patient with SPG4 and a male control subject. The horizontal axis indicates the physical position of oligonucleotide probes. If copy number variations do not exist, the $\log_2$ ratios of the hybridization signal intensities are expected to be near 0. In the region indicated by an orange bar, the $\log_2$ ratio of hybridization signal intensities is approximately −1, which indicates a heterozygous deletion (halved gene dosage) in *SPAST*. (**e**) PCR analysis using primers flanking the deletions revealed that the truncated band corresponding to 1.8 kb was detected only in the patient. No PCR product was detected in a control, because the distance between the primer pair was too long to amplify (~11 kb). (**f**) Direct nucleotide sequence analysis determines breakpoints with a deletion size of 9565 bp.

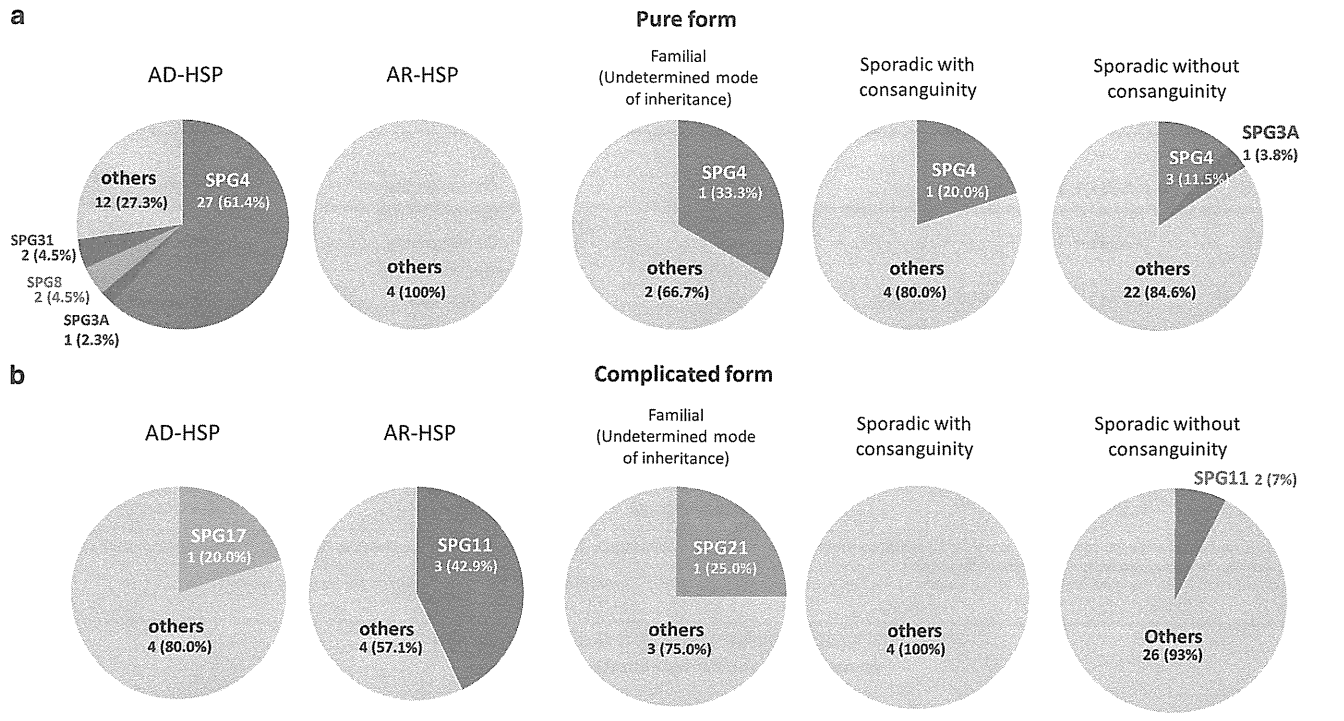

Figure 3 Relative frequencies of individual HSP types in groups classified on the basis of the clinical presentations and mode of inheritance. The figure shows the relative frequencies of individual HSP types in our cohort. (**a**) Pure form and (**b**) complicated form. The family history of each subgroup is indicated above the figures. Mutations were detected in a total of 67.3% of all the AD-HSP patients or 72.7% of the patients with pure-form AD-HSP. Focusing on sporadic HSP patients, six patients (four SPG4, one SPG3A and one SPG11) were identified, which accounted for 9.8% (6/61). Of note, *SPAST* mutations were present in 6.6% of all sporadic HSP patients, and particularly in 12.9% (4/31) of sporadic pure-form HSP patients, suggesting the usefulness of mutational analysis of *SPAST* in sporadic cases, particularly in patients with the pure form. Others, patients with unidentified mutation.

callosum and cognitive impairment, 35.7% (5/14) carried *SPG11* mutations.

Molecular and clinical spectra of individual HSP types

SPG3A. We found two patients with SPG3A carrying previously reported mutations (Table 2). Although both patients with SPG3A showed basically pure-form HSP with juvenile onset, one patient showed hypesthesia and hypalgesia in the distal lower limbs accompanied by decreased vibratory sensation in all extremities.

SPG4. Of the 32 patients with SPG4, 24 (75%) had nonsense, frameshift or large deletion/duplication mutations leading to truncated proteins, which were distributed throughout the genes (Supplementary Figure S2). On the other hand, seven out of the eight missense mutations were located in the AAA domain (ATPase associated with various cellular activities). We found a novel mutation (p.Y52C) outside the AAA domain. Note that large deletions/duplications in *SPAST* were detected by aCGH analysis, and small deletion mutations were detected by Sanger sequencing analysis in 22.7% (5/22)[22] and 45.5% (10/22) of AD-HSP patients, respectively, in whom no mutations were detected by the resequencing microarray analysis. The ages at onset of patients with SPG4 showed two peaks, in the teens and in 40 s (Supplementary Figure S3A). The types of the mutation in *SPAST* and age at onset did not correlate (Supplementary Figure S3B).

SPG8. We found a large deletion in *KIAA0196*, which has not been described to date. The breakpoints of the large deletion in *KIAA0196* are located in intron 10 and exon 15 (Figures 4a–c). RT-PCR and direct nucleotide sequence analyses revealed that exons 10–15 were deleted in cDNA, predicting a premature termination codon (Figures 4d and e). There are only three missense mutations reported to date, and in a previous paper, it was proposed that haploinsufficiency is the disease-causing mechanism of SPG8 on the basis of experiments using zebrafish.[20] The large deletion in *KIAA0196* detected in the present study further supported a disease mechanism of haploinsufficiency and indicate a necessity of screening for rearrangements of *KIAA0196* in AD-HSP. SPG8 has been reported to be an 'aggressive' subtype of HSP and the disease onset is in the 20s or 30s.[20] In contrast, two patients with SPG8 found in the study had adult-onset or late-onset HSP.

SPG11. The five patients with SPG11 showed complicated-form HSP with cognitive impairment and a thin corpus callosum. Notably, rearrangement in *SPG11* was found in a patient, and aCGH analysis was helpful for accurate diagnosis of the patient. The age at onset ranged from 2 to 25 years. Although SPG11 is allelic to juvenile amyotrophic lateral sclerosis (ALS5),[23] none of the patients showed the ALS phenotype.

SPG17. A novel *BSCL2* (NM_032667) p.C36Y substitution (which can also be called p.C100Y in NM_001122955 because there are two known start codons) was found in one AD-HSP patient. He suffered from early-onset spastic paraparesis with mild mental retardation and did not show amyotrophy. Clinical and genetic data of other family members were not available. C36 is conserved among species and is located in the first transmembrane domain,[24] raising a possibility that p.C36Y can change the function of seipin, the protein product of *BSCL2*. Because only p.N88S and p.S90L of seipin have been

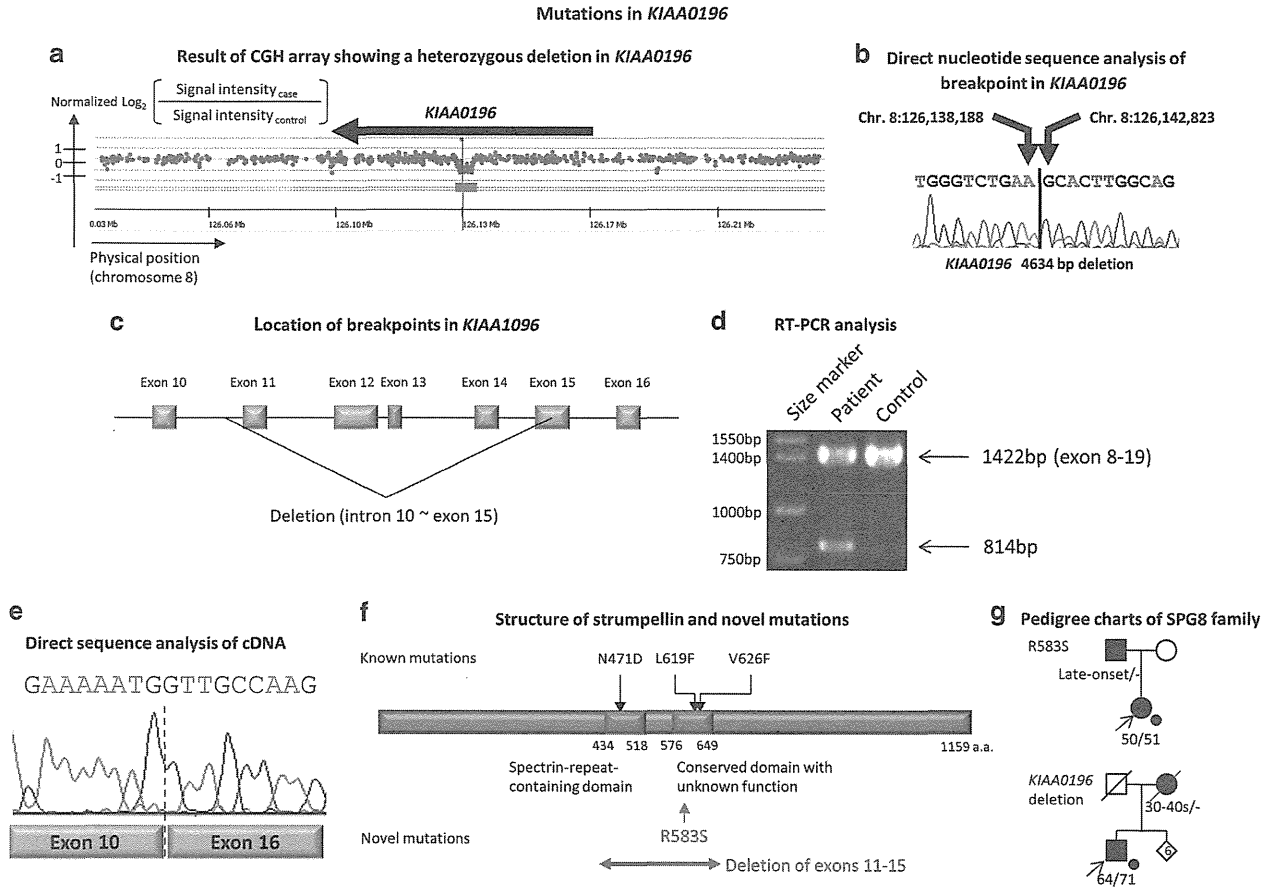

Figure 4 Mutations in *KIAA0196*. (**a**) Result of CGH array showing a heterozygous deletion in KIAA0196. An orange bar shows heterozygous deletion. (**b**) Direct nucleotide sequence analysis of the breakpoint in KIAA0196, which shows 4634 bp deletion. (**c**) Schematic presentation of the exon–intron structure of *KIAA0196*. The deletion detected by the array CGH analysis is shown. (**d**) RT-PCR analysis of species of RNAs extracted from the patient with the *KIAA0196* deletion and a control. In the control, only a single band with the expected size corresponding to 1422 bp was observed, while a truncated band with the size corresponding to 814 bp in addition to PCR products corresponding to 1422 bp was observed in the patient. (**e**) Direct nucleotide sequence analysis of the truncated PCR products revealed that exons 11–15 were absent in the *KIAA0196* mRNA as a result of a deletion in *KIAA0196*. (**f**) Schematic representation of strumpellin, the protein product of *KIAA0196*, and the mutations identified in patients with SPG8. The position of the large deletion (deletion of exons 11–14 and a part of exon 15) and the novel mutation found in the present study are shown (red). Previously reported mutations in *KIAA0196* (p.N471D, p.L619F and p.V626F) are located in the spectrin-repeat-containing domain (amino acids 434–518) or the conserved domain with unknown function (amino acids 576–649). The novel mutation (p.R583S) found in the present study is also located in the conserved domain with unknown function. (**g**) Pedigree charts of the Japanese SPG8 families. Age at onset and age at examination are indicated.

described in Silver syndrome/SPG17, we still need to be cautious about the pathogenicity of p.C36Y substitution.

SPG21. We found a novel homozygous amino-acid substitution (p.A108P) in *SPG21* encoding maspardin in a family with late-onset complicated-form HSP (Figure 5, Supplementary Tables S1 and S2). The two patients managed to walk with a cart or a cane in their 70s and 60s. In addition to cognitive decline, callosal disconnection syndrome, such as ideomotor apraxia predominantly of the left hand, agraphia of the left hand and constructional impairment predominantly on the right side, was observed, which was mild but progressed over 5 years in the index patient. There were no extrapyramidal signs, cerebellar signs or bulbar symptoms, as reported in the original family with an *SPG21* mutation.[25] Magnaetic resonance imaging of the index patient showed progressive thinning of the corpus callosum and predominantly frontotemporal atrophy (Figure 5e–i). ^{123}I-*N*-isopropyl-*p*-iodoamphetamine single-photon emission computed tomography revealed decreased blood flow in the frontal and temporal cortices (Figure 5j).

This family is the first family with SPG21 identified outside the Amish population.[25] Intriguingly, compared with the original Mast syndrome family with an *SPG21* mutation, the ages at onset of HSP symptoms in the patients in the new SPG21 family were strikingly late. Although characteristics such as cognitive decline and a thin corpus callosum were shared in common, characteristic clinical signs in Mast syndrome such as bulbar, extrapyramidal and cerebellar signs were not found in the new family (Supplementary Table S2), thus presenting dissimilar phenotypes. Because the mutation detected in the new family is a missense mutation (p.A108P) next to the active site of the alpha/beta-hydrolase domain (S109), dysfunction of alpha/beta-hydrolase activity of maspardin seems to be related to pathogenicity.

SPG31. The two novel mutations in *REEP1* were a frameshift (insertion of A) and a large deletion (Table 2), suggesting haploinsufficiency as the disease-causing mechanism. A large deletion detected in the study demanded a screening of rearrangement of *REEP1* in the diagnosis of SPG31. These two patients with SPG31 had

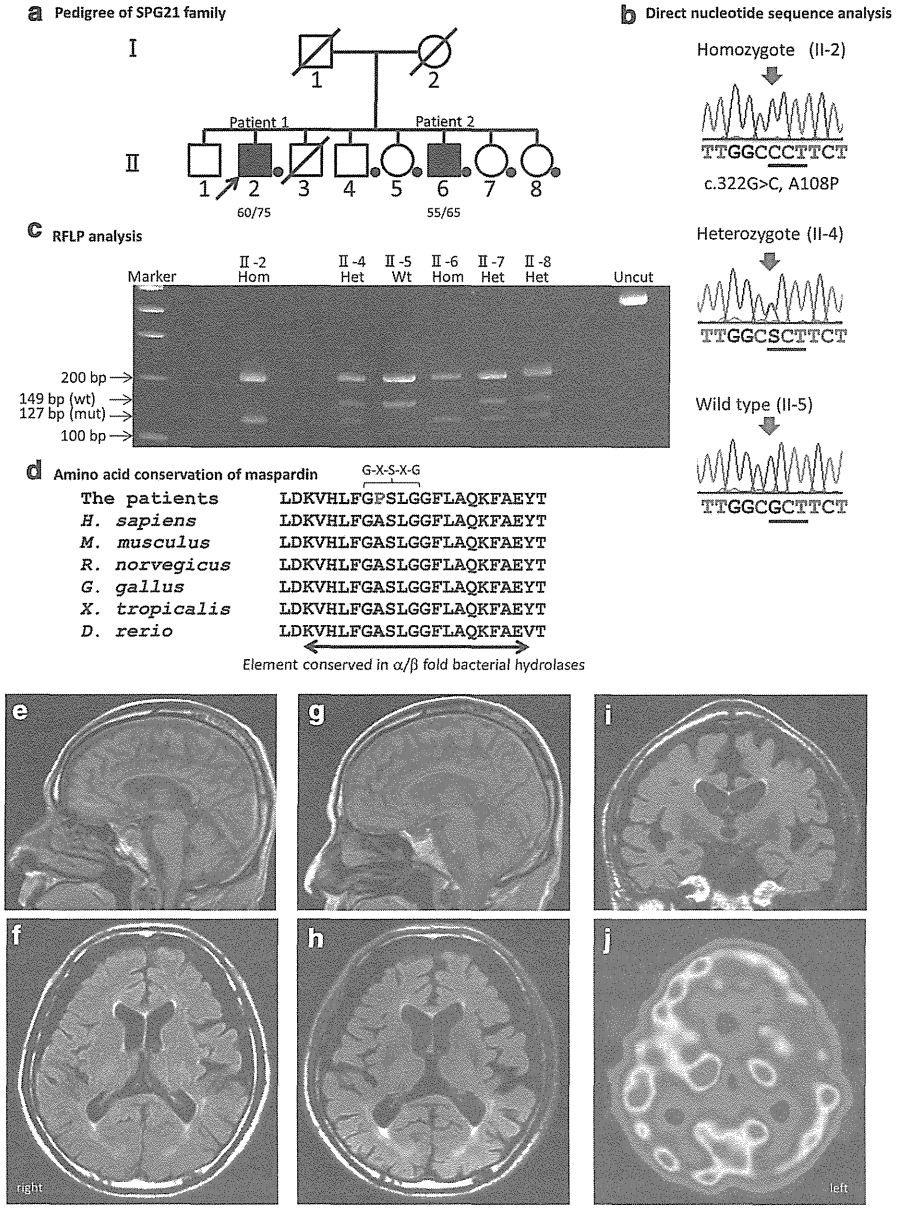

Figure 5 A family with SPG21 and molecular genetic analysis. (**a**) Pedigree tree of the family. Squares indicate males and circles indicate females. Black squares are affected members and the index patient (II-2) is indicated by an arrow. Symbols with a diagonal line indicate deceased members. Members with dots allowed us neurological and genetic examinations. (**b**) Electropherograms of the family members carrying homozygous c.322G>C mutation (II-2), heterozygous c.322G>C mutation (II-4) and wild-type allele (II-5). (**c**) PCR-restriction fragment length polymorphism (RFLP) analysis of family members. The uncut PCR fragment length is 344 bp. With *Hae*III digestion, the wild-type allele shows fragment sizes of 149 and 195 bp, whereas the mutant allele shows fragment sizes of 127, 22 and 195 bp. (**d**) Comparison of amino-acid sequence of ACP33/maspardin among species. A108 is located in the α/β-fold bacterial hydrolase domain, which is highly evolutionally conserved. The G-X-S-X-G motif at the nucleophile elbow is also shown. (**e** and **f**) A sagittal T1-weighted image (**e**) and a transverse fluid-attenuated inversion recovery (FLAIR) image (**f**) of patient 1 at the age of 70 years show a thin corpus callosum and mildly atrophic cerebrum. Atrophy in the brainstem and cerebellum is not observed. (**g**–**i**) A sagittal T1-weighted image (**g**), a transverse FLAIR image (**h**) and a coronal FLAIR image (**i**) of patient 1 at the age of 75 years shows progressive thinning of the corpus callosum mainly in the trunk and progressive atrophy of the cerebrum, which is marked in the frontal and temporal lobes. Slight white matter changes are observed around the lateral ventricles. Atrophy in the brainstem and cerebellum is not observed. (**j**) ^{123}I-*N*-isopropyl-*p*-iodoamphetamine single-photon emission computed tomography (SPECT) at the age of 75 years shows decreased blood flow in the frontal and temporal cortices. Wt, wild type; mut, mutant; Het, heterozygote; Homo, homozygote.

pure-form HSP and their disease started in their early teens, compatible with previous reports.[9,21]

Sporadic HSP. As much as 11.1% (7/63) of the patients with sporadic HSP were revealed to have mutations in the genes for monogenic diseases. Among sporadic pure-form HSP patients, 12.9% (4/31) had *SPAST* mutations, and 6.3% (2/32) of sporadic complicated-form HSP patients had *SPG11* mutations.

DISCUSSION

We herein described a comprehensive mutational analysis of as many as 16 causative genes of HSP and applied it to the mutational analysis

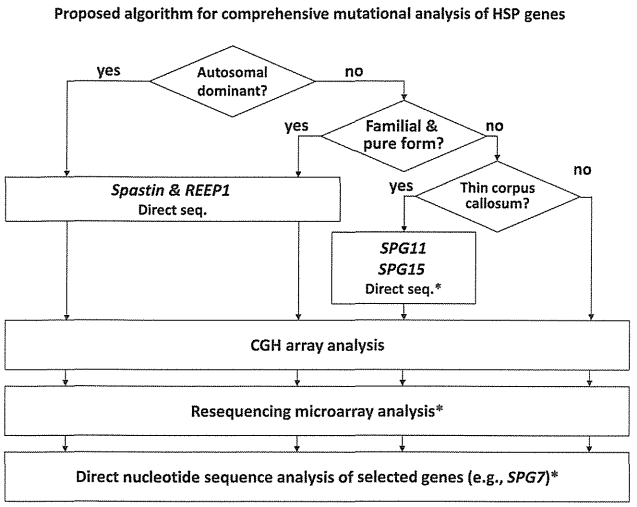

Figure 6 Proposed algorithm for comprehensive mutational analysis of HSP genes. Considering the types and frequencies of mutations in individual SPG genes, we propose an efficaceous strategy for a large-scale mutational analysis of HSP at the time of the study. In patients with ADHSP patients and in familial pure HSP patients with an unknown mode of inheritance, direct nucleotide sequence analysis of *spastin* and *REEP1* followed by CGH analysis is recommended, considering the relatively high frequency of small insertions/deletions in *spastin* and *REEP1* and large deletions/duplications in *spastin*. In patients with thin corpus callosum and/or cognitive dysfunction, *SPG11* and *SPG15* should be analyzed first. Next step is CGH analysis followed by resequencing microarray analysis, because throughput of CGH analysis is higher than that of resequencing microarray analysis. *In these days, these stages can be replaced by whole genome or exome sequencing. Direct seq., direct nucleotide sequence analysis.

of 129 Japanese HSP patients. An epidemiological study[26] based on the Registry of the Ministry of Health, Labour and Welfare, Japan in 2002 reported about 500 HSP patients. Although there remains a possibility that some patients may have not been registered for various reasons, the collection of 129 patients should represent a substantial proportion of Japanese HSP patients. In the 129 HSP patients, we identified 49 mutations, 32 of which were novel. Resequencing microarray and aCGH analyses were proved to be efficacious methods to detect nucleotide substitutions and large duplications/deletions, respectively. Indeed, the fact that we did not find additional base substitution mutations of *SPAST* and *REEP1* in AD-HSP patients by direct sequence analysis, for whom mutations were not detected by resequencing microarrays, indicates a false-negative rate of resequencing microarray analysis was low, if any, by tuning up by our algorithm (a computer program). However, note also that both resequencing microarray and aCGH analyses did not detect small insertion/deletion mutations, and direct nucleotide sequence analysis was needed to detect them. Our results revealed that the combination of these technologies, including resequencing microarray, aCGH, and direct nucleotide sequence analyses, are essential to detect various kinds of mutations, including base substitutions, and insertions/deletions of various sizes with high sensitivities.

Given the results of this study, we propose an algorithm for a comprehensive mutational analysis for HSP. To analyze genes that have relatively frequent small insertion/deletion mutations (for example, *SPAST*, *REEP1*, *SPG11* and *SPG15*), direct nucleotide sequence analysis is the first priority. To analyze genes in which most of the mutations are nucleotide substitutions (for example, *ATL1*, *NIPA1*, *KIF5A*, *KIAA0196*, *HSPD1* and *BSCL2*), resequencing microarray analysis is highly suitable. Considering the throughput, direct nucleotide sequence analysis becomes more laborious as the number of exons to be sequenced increases. In contrast, it is not the case for resequencing microarray and CGH array analyses. That is, the time required for analysis remains constant with increasing number of genes or exons to be sequenced until a limit determined by the structure of arrays. We propose a strategy of utilizing high-throughput microarray techniques and minimizing the use of time-consuming direct nucleotide sequence analysis considering the molecular epidemiology and the mutation types in individual genes (Figure 6). Although there remains a possibility that uncommon mutations (for example, insertions/deletions of intermediate length) or uncommon presentation (for example, *SPAST* mutation in a family having apparently autosomal recessive mode of inheritance or *SPG11* mutations in a pseudoautosomal dominant family) are missed and it might introduce some bias, the algorithm should be highly useful for the efficient identification of the majority, if not all, of the mutations responsible for HSP.

Utilizing the technologies, we elucidated molecular epidemiology of HSP in the Japanese population. Interestingly, the study revealed that the overall trend of molecular epidemiology of AD-HSP/AR-HSP in the Japanese population is similar to those in the Caucasian populations reported previously.[3,5,6,20,21,27] In contrast, considerable differences in the epidemiology of spinocerebellar ataxias[26] or amyotrophic lateral sclerosis (especially in those who have hexanucleotide repeat expansion mutation in *C9ORF72*)[28,29] have been demonstrated, which presumably reflect founder effects.[29,30] Thus, the similarity in the molecular epidemiology of HSP irrespective of ethnicity suggests that contribution of founder effects is limited in HSP.

We did not find causative mutations in 16 AD-HSP, 8 AR-HSP and 5 familial HSP patients. Although we cannot completely exclude the possibility of false-negative results in our analyses, we assume that these undiagnosed patients would have mutations in causative genes that have recently been identified after the study (*RTN2* or *GBA2*, for example) or mutations in as yet unidentified causative genes.

The extent to which mutations of causative genes account for apparently sporadic HSP is an important but unsolved issue. We found that 7 out of the 62 sporadic HSP patients (11.1%) had mutations of genes for HSP. In particular, we found that SPG4 and SPG11 are relatively frequent in sporadic pure-form HSP and complicated-form HSP patients, respectively. The findings indicate that careful genetic counseling of such patients and families will be required.

With recent progresses in massively parallel sequencing technologies, exome and targeted sequencing are now becoming a robust method for high-throughput resequencing analysis at a relatively reasonable cost.[31–34] Detection of large insertion/deletion mutations based on the short reads generated by next-generation sequencers, however, is still a challenging task. It is of note that a substantial proportion (7/49, 14.3%) of mutations found in the study were insertions/deletions detected by aCGH analysis. Thus, combining multiple technologies, as we did in the study, is indispensable to detect as many mutations as possible even in the next-generation sequencer era. In addition, information on the relative frequencies of HSP types and on the distribution of various types of mutations in each HSP gene as shown in the study is helpful for making strategies for mutational analyses.

In summary, we elucidated the molecular epidemiology of HSP in the Japanese population combining multiple technologies of

resequencing microarray, aCGH and Sanger sequencing. The study contributed to further broadening the clinical and mutational spectra of HSP.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank all the patients and their families for their participation in this study. We also thank the many doctors who kindly provided clinical information and the blood and brain samples of the participants. This work was supported in part by KAKENHI (Grant-in-Aid for Scientific Research) on Priority Areas, Innovative Areas, the Global COE program, and Scientific Research (A) from the Ministry of Education, Culture, Sports, Science and Technology of Japan and a Grant-in-Aid for 'the Research Committee for Ataxic Diseases' of the Research on Measures for Intractable Diseases from the Ministry of Health, Welfare and Labour, Japan. HI was supported by a Research Fellowship of the Japan Society for the Promotion of Science for Young Scientists.

1 Fink, J. K. The hereditary spastic paraplegias: nine genes and counting. *Arch. Neurol.* **60,** 1045–1049 (2003).
2 Finsterer, J., Löscher, W., Quasthoff, S., Wanschitz, J., Auer-Grumbach, M. & Stevanin, G. Hereditary spastic paraplegias with autosomal dominant, recessive, X-linked, or maternal trait of inheritance. *J. Neurol. Sci.* **318,** 1–18 (2012).
3 Fonknechten, N., Mavel, D., Byrne, P., Davoine, C. S., Cruaud, C., Bönsch, D. *et al.* Spectrum of SPG4 mutations in autosomal dominant spastic paraplegia. *Hum. Mol. Genet.* **9,** 637–644 (2000).
4 McDermott, C. J., Burness, C. E., Kirby, J., Cox, L. E., Rao, D. G., Hewamadduma, C. *et al.* Clinical features of hereditary spastic paraplegia due to spastin mutation. *Neurology* **67,** 45–51 (2006).
5 Namekawa, M., Ribai, P., Nelson, I., Forlani, S., Fellmann, F., Goizet, C. *et al.* SPG3A is the most frequent cause of hereditary spastic paraplegia with onset before age 10 years. *Neurology* **66,** 112–114 (2006).
6 Klebe, S., Lacour, A., Dürr, A., Stojkovic, T., Depienne, C., Forlani, S. *et al.* NIPA1 (SPG6) mutations are a rare cause of autosomal dominant spastic paraplegia in Europe. *Neurogenetics* **8,** 155–157 (2007).
7 Elleuch, N., Depienne, C., Benomar, A., Hernandez, A. M., Ferrer, X., Fontaine, B. *et al.* Mutation analysis of the paraplegin gene (SPG7) in patients with hereditary spastic paraplegia. *Neurology* **66,** 654–659 (2006).
8 Arnoldi, A., Tonelli, A., Crippa, F., Villani, G., Pacelli, C., Sironi, M. *et al.* A clinical, genetic, and biochemical characterization of SPG7 mutations in a large cohort of patients with hereditary spastic paraplegia. *Hum. Mutat.* **29,** 522–531 (2008).
9 Beetz, C., Schüle, R., Deconinck, T., Tran-Viet, K. N., Zhu, H., Kremer, B. P. *et al.* REEP1 mutation spectrum and genotype/phenotype correlation in hereditary spastic paraplegia type 31. *Brain* **131,** 1078–1086 (2008).
10 Goizet, C., Boukhris, A., Mundwiller, E., Tallaksen, C., Forlani, S., Toutain, A. *et al.* Complicated forms of autosomal dominant hereditary spastic paraplegia are frequent in SPG10. *Hum. Mutat.* **30,** E376–E385 (2008).
11 Warrington, J. A., Shah, N. A., Chen, X., Janis, M., Liu, C., Kondapalli, S. *et al.* New developments in high-throughput resequencing and variation detection using high density microarrays. *Hum. Mutat.* **19,** 402–409 (2002).
12 Barrett, M. T., Scheffer, A., Ben-Dor, A., Sampas, N., Lipson, D., Kincaid, R. *et al.* Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. *Proc. Natl Acad. Sci. USA* **101,** 17765–17770 (2004).
13 Arai, N., Kishino, A., Takahashi, Y., Morita, D., Nakamura, K., Yokoyama, T. *et al.* Familial cases presenting very early onset autosomal dominant Alzheimer's disease with I143T in presenilin-1 gene: implication for genotype-phenotype correlation. *Neurogenetics* **9,** 65–67 (2008).
14 Takahashi, Y., Seki, N., Ishiura, H., Mitsui, J., Matsukawa, T., Kishino, A. *et al.* Development of a high-throughput microarray-based resequencing system for neurological disorders and its application to molecular genetics of amyotrophic lateral sclerosis. *Arch. Neurol.* **65,** 1326–1332 (2008).
15 Seki, N., Takahashi, Y., Tomiyama, H., Rogaeva, E., Murayama, S., Mizuno, Y. *et al.* Comprehensive mutational analysis of LRRK2 reveals variants supporting association with autosomal dominant Parkinson's disease. *J. Hum. Genet.* **56,** 671–675 (2011).
16 Cutler, D. J., Zwick, M. E., Carrasquillo, M. M., Yohn, C. T., Tobin, K. P., Kashuk, C. *et al.* High-throughput variation detection and genotyping using microarrays. *Genome Res.* **11,** 1913–1925 (2001).
17 Mitsui, J., Takahashi, Y., Goto, J., Tomiyama, H., Ishikawa, S., Yoshino, H. *et al.* Mechanisms of genomic instabilities underlying two common fragile-site-associated loci, PARK2 and DMD, in germ cell and cancer cell lines. *Am. J. Hum. Genet.* **87,** 75–89 (2010).
18 Maeda-Hashimoto, M., Mitsui, J., Soong, B. W., Takahashi, Y., Ishiura, H., Hayashi, S. *et al.* Increased gene dosage of myelin protein zero causes Charcot-Marie-Tooth disease. *Ann. Neurol.* **71,** 84–92 (2012).
19 Silver, J. R. Familial spastic paraplegia with amyotrophy of the hands. *Ann. Hum. Genet.* **30,** 69–75 (1966).
20 Valdmanis, P. N., Meijer, I. A., Reynolds, A., Lei, A., MacLeod, P., Schlesinger, D. *et al.* Mutations in the KIAA0196 gene at the SPG8 locus cause hereditary spastic paraplegia. *Am. J. Hum. Genet.* **80,** 152–161 (2007).
21 Züchner, S., Wang, G., Tran-Viet, K. N., Nance, M. A., Gaskell, P. C., Vance, J. M. *et al.* Mutations in the novel mitochondrial protein REEP1 cause hereditary spastic paraplegia type 31. *Am. J. Hum. Genet.* **79,** 365–369 (2006).
22 Beetz, C., Nygren, A. O., Schickel, J., Auer-Grumbach, M., Bürk, K., Heide, G. *et al.* High frequency of partial SPAST deletions in autosomal dominant hereditary spastic paraplegia. *Neurology* **67,** 1926–1930 (2006).
23 Orlacchio, A., Babalini, C., Borreca, A., Patrono, C., Massa, R., Basaran, S. *et al.* SPATACSIN mutations cause autosomal recessive juvenile amyotrophic lateral sclerosis. *Brain* **133,** 591–598 (2012).
24 Ito, D. & Suzuki, N. Seipinopathy: a novel endoplasmic reticulum stress-associated disease. *Brain* **132,** 8–15 (2009).
25 Simpson, M. A., Cross, H., Proukakis, C., Pryde, A., Hershberger, R., Chatonnet, A. *et al.* Maspardin is mutated in mast syndrome, a complicated form of hereditary spastic paraplegia associated with dementia. *Am. J. Hum. Genet.* **73,** 1147–1156 (2003).
26 Tsuji, S., Onodera, O., Goto, J. & Nishizawa, M. Study Group on Ataxic Diseases.. Sporadic ataxias in Japan—a population-based epidemiological study. *Cerebellum* **7,** 189–197 (2008).
27 Stevanin, G., Santorelli, F. M., Azzedine, H., Cutinho, P., Chomilier, J., Denora, P. S. *et al.* Mutations in SPG11, encoding spatacsin, are a major cause of spastic paraplegia with thin corpus callosum. *Nat. Genet.* **39,** 366–372 (2007).
28 Ishiura, H., Takahashi, Y., Mitsui, J., Yoshida, S., Kihira, T., Kokubo, Y. *et al.* C9ORF72 repeat expansion in amyotrophic lateral sclerosis in the Kii peninsula of Japan. *Arch. Neurol.* **69,** 1154–1158 (2012).
29 Majounie, E., Renton, A. E., Mok, K., Dopper, E. G., Waite, A., Rollinson, S. *et al.* Frequency of the C9orf72 hexanucleotide repeat expansion in patients with amyotrophic lateral sclerosis and frontotemporal dementia: a cross-sectional study. *Lancet Neurol.*. **11,** 323–330 (2012).
30 Cossée, M., Schmitt, M., Campuzano, V., Reutenauer, L., Moutou, C., Mandel, J. L. *et al.* Evolution of the Friedreich's ataxia trinucleotide repeat expansion: founder effect and premutations. *Proc. Natl Acad. Sci. USA* **94,** 7452–7457 (1997).
31 Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl Acad. Sci. USA* **106,** 19096–19101 (2009).
32 Ku, C. S., Cooper, D. N., Polychronakos, C., Naidoo, N., Wu, M. & Soong, R. Exome sequencing: dual role as a discovery and diagnostic tool. *Ann. Neurol.* **71,** 5–14 (2012).
33 Ishiura, H., Sako, W., Yoshida, M., Kawarai, T., Tanabe, O., Goto, J. *et al.* The TRK-fused gene is mutated in hereditary motor and sensory neuropathy with proximal dominant involvement. *Am. J. Hum. Genet.* **91,** 320–329 (2012).
34 Mitsui, J., Matsukawa, T., Ishiura, H., Higasa, K., Yoshimura, J., Saito, T. L. *et al.* CSF1R mutations identified in three families with autosomal dominantly inherited leukoencephalopathy. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **159B,** 951–957 (2012).
35 Lindsey, J. C., Lusher, M. E., McDermott, C. J., White, K. D., Reid, E., Rubinsztein, D. C. *et al.* Mutation analysis of the spastin gene (SPG4) in patients with hereditary spastic paraparesis. *J. Med. Genet.* **37,** 759–765 (2000).
36 Falco, M., Scuderi, C., Musumeci, S., Sturnio, M., Neri, M., Bigoni, S. *et al.* Two novel mutations in the spastin gene (SPG4) found by DHPLC mutation analysis. *Neuromuscul. Disord.* **14,** 750–753 (2004).
37 Depienne, C., Tallaksen, C., Lephay, J. Y., Bricka, B., Poea-Guyon, S., Fontaine, B. *et al.* Spastin mutations are frequent in sporadic spastic paraparesis and their spectrum is different from that observed in familial cases. *J. Med. Genet.* **43,** 259–265 (2006).
38 Meijer, I. A., Hand, C. K., Cossette, P., Figlewicz, D. A. & Rouleau, G. A. Spectrum of SPG4 mutations in a large collection of North American families with hereditary spastic paraplegia. *Arch. Neurol.* **59,** 281–286 (2002).
39 Patrono, C., Scarano, V., Cricchi, F., Melone, M. A., Chiriaco, M., Napolitano, A. *et al.* Autosomal dominant hereditary spastic paraplegia: DHPLC-based mutation analysis of SPG4 revealed eleven novel mutations. *Hum. Mutat.* **25,** 506 (2005).
40 D'Amico, A., Tessa, A., Sabino, A., Bertini, E., Santorelli, F. M. & Servidei, S. Incomplete penetrance in an SPG3A-linked family with a new mutation in the atlastin gene. *Neurology* **62,** 2138–2139 (2004).
41 Dürr, A., Camuzat, A., Colin, E., Tallaksen, C., Hannequin, D., Coutinho, P. *et al.* Atlastin1 mutations are frequent in young-onset autosomal dominant spastic paraplegia. *Arch. Neurol.* **62,** 962–966 (2004).
42 Kim, S. M., Lee, J. S., Kim, S., Kim, H. J., Kim, M. H., Lee, K. M. *et al.* Novel compound heterozygous mutations of the SPG11 gene in Korean families with hereditary spastic paraplegia with thin corpus callosum. *J. Neurol.* **256,** 1714–1718 (2009).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)

BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 6 2014, pages 815–822
doi:10.1093/bioinformatics/btt647

Genetics and population analysis

Advance Access publication November 8, 2013

Rapid detection of expanded short tandem repeats in personal genomics using hybrid sequencing

Koichiro Doi[1], Taku Monjo[1,2], Pham H. Hoang[1,2], Jun Yoshimura[1], Hideaki Yurino[1], Jun Mitsui[3], Hiroyuki Ishiura[3], Yuji Takahashi[3], Yaeko Ichikawa[3], Jun Goto[3], Shoji Tsuji[3] and Shinichi Morishita[1,*]

[1]Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, Chiba 277-8562, [2]Department of Information and Communication Engineering, Faculty of Engineering and [3]Department of Neurology, Graduate School of Medicine, The University of Tokyo, Tokyo 113-8655, Japan

Associate Editor: Jeffrey Barrett

ABSTRACT

Motivation: Long expansions of short tandem repeats (STRs), i.e. DNA repeats of 2–6 nt, are associated with some genetic diseases. Cost-efficient high-throughput sequencing can quickly produce billions of short reads that would be useful for uncovering disease-associated STRs. However, enumerating STRs in short reads remains largely unexplored because of the difficulty in elucidating STRs much longer than 100 bp, the typical length of short reads.

Results: We propose *ab initio* procedures for sensing and locating long STRs promptly by using the frequency distribution of all STRs and paired-end read information. We validated the reproducibility of this method using biological replicates and used it to locate an STR associated with a brain disease (SCA31). Subsequently, we sequenced this STR site in 11 SCA31 samples using SMRT™ sequencing (Pacific Biosciences), determined 2.3–3.1 kb sequences at nucleotide resolution and revealed that (TGGAA)- and (TAAAATAGAA)-repeat expansions determined the instability of the repeat expansions associated with SCA31. Our method could also identify common STRs, (AAAG)- and (AAAAG)-repeat expansions, which are remarkably expanded at four positions in an SCA31 sample. This is the first proposed method for rapidly finding disease-associated long STRs in personal genomes using hybrid sequencing of short and long reads.

Availability and implementation: Our TRhist software is available at http://trhist.gi.k.u-tokyo.ac.jp/.

Contact: moris@cb.k.u-tokyo.ac.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 17, 2013; revised on October 20, 2013; accepted on November 4, 2013

1 INTRODUCTION

Many genetic disorders are caused by or associated with short tandem repeats (STRs), repetitive elements of 2–6 nt. Regarding the mechanism underlying the phenomenon of repeat expansion, unusual structural features of repeat-containing regions that affect cellular replication, repair and recombination are thought to induce frequent replication slippage, thereby expanding repeats (Mirkin, 2007). STRs have been found in a variety of genomic regions. Huntington's disease is associated with expansion of the triplet repeat $(CAG)_n$ (polyglutamine runs in proteins) in the coding region of huntingtin (The Huntington's Disease Collaborative Research Group, 1993), where $n<28$ in normal samples, $n=28$–35 in intermediate cases, $n=36$–40 in reduced penetrance and $n>40$ in full penetrance (Walker, 2007). Spinal and bulbar muscular atrophy is also associated with (CAG) repeats in one exon (La Spada *et al.*, 1991).

In addition to exons, STRs have been observed in a variety of genomic regions such as untranslated regions (UTRs), introns and promoters. Fragile-X syndrome is associated with (CGG) repeat in the 5′-UTR (Kremer *et al.*, 1991; Sherman *et al.*, 1985; Verkerk *et al.*, 1991) and myotonic dystrophy type 1 (DM1) with (CTG) repeat in the 3′-UTR (Brook *et al.*, 1992; Mahadevan *et al.*, 1992). In introns, spinocerebellar ataxia type 10 (SCA10) is associated with (ATTCT) repeat (Matsuura *et al.*, 2000), myotonic dystrophy type 2 (DM2) with (CCTG) repeat (Liquori *et al.*, 2001), amyotrophic lateral sclerosis/frontotemporal dementia (ALS/FTD) with (GGGGCC) repeat (DeJesus-Hernandez *et al.*, 2011; Orr, 2011; Renton *et al.*, 2011) and SCA36 with (GGCCTG) repeat (Kobayashi *et al.*, 2011). Consequently, whole-genome sequencing capable of observing non-exonic regions is required to characterize STRs peculiar to a personal genome.

Several expanded repeats in RNA, such as CUG, CCUG, CAG, CGG, AUUCU and UGGAA, are associated with hereditary diseases and are known to accumulate in nuclear RNA foci in which several proteins are sequestrated in the process of foci formation (for a review see Wojciechowska and Krzyzosiak, 2011). These RNA foci are thought to have a negative effect on host cells, leading to disorders in cellular pathways (Wojciechowska and Krzyzosiak, 2011).

To search personal genomes for STRs, the most cost-efficient way would be to resequence an entire personal genome and to collect billions of short reads of ~100 bp in length using available high-throughput sequencers. However, the infeasibility of obtaining longer reads at reasonable cost might lead to the failure to detect important STRs because expandable repeats associated with diseases can sometimes be quite long [e.g. $(ATTCT)_n$, $n=800$–4500 in SCA10 and $(CCTG)_n$, $n=$~5000 in DM2] and are much longer than 100 bp, the typical length of short reads,

*To whom correspondence should be addressed

making the identification and location of long STRs in a personal genome non-trivial.

Another serious problem is that STRs have several variants with many mutations. The spontaneous mutation rate of STRs, 3.78×10^{-4} to 7.44×10^{-2} in the human Y-chromosome (Ballantyne *et al.*, 2010), is far higher than the rate of copy number variation, 1.7×10^{-6} to 1.2×10^{-4} (Lupski, 2007), and the reported average rate of *de novo* single-nucleotide variation, 1.18×10^{-8} (SD $= 0.15 \times 10^{-8}$) (Conrad *et al.*, 2011) and 1.20×10^{-8} (Kong *et al.*, 2012). The ultrahigh mutation rate of STRs is thought to be a major force driving genetic variation producing a variety of STRs with differences often specific to personal genomes. Therefore, detecting various STRs by processing billions of short raw reads is fundamental to the analysis of personal genomes.

Several software programs list STRs, such as Tandem Repeat Finder (Benson, 1999), Mreps (Kolpakov *et al.*, 2003), ATRHunter (Wexler *et al.*, 2005), IMEx (Mudunuri and Nagarajaram, 2007) and T-reks (Jorda and Kajava, 2009) (for a recent review that compares these programs, see Lim *et al.*, 2013); however, these conventional programs are designed to retrieve STRs from nearly complete or draft long genomes and are not intended for processing billions of short reads in a reasonable amount of time. Another problem involved in handling short reads is the difficulty of determining the accurate positions of STRs in the genome because reads filled with STRs are not included in the genome or often map to multiple locations. The problem is solvable in some cases when a flanking region around an STR in a read is long enough to map to a unique position (Fig. 1B). To resolve these special cases, Gymrek *et al.* developed the program lobSTR (Gymrek *et al.*, 2012), which improves the efficiency of this process by selecting ~240 000 candidate regions harboring STRs in the human genome. Owing to severe restrictions in potential STR regions, however, we might overlook novel STRs hidden in numerous short reads because known STRs associated with diseases are frequently much longer than 100 bp, the typical length of short reads produced by high-throughput sequencers (Fig. 1C).

Here, we propose a new cost-efficient method for calculating a comprehensive collection of STRs that are longer than short reads by inspecting the frequency distribution of STRs in short reads. To approximate the locations of such STRs, we use paired-end sequencing to facilitate locating the opposite end of the read with the focal STR in a pair, thereby narrowing down the location of the focal STR. Finally, we present a statistical procedure for selecting STRs that are significantly expanded in the case sample.

2 METHODS

2.1 Non-redundant representation of STRs

Our goal was to enumerate all possible instances of STRs with 2–6-base-long repeat units efficiently. In general, our algorithm can detect repeat units of an arbitrary length without sacrificing computational time. We also present an example of disease-associated STRs with a 10-base repeat unit in SCA31 (Sato *et al.*, 2009). Care is required to avoid double counting identical STR occurrences characterized by more than one STR pattern. To remove redundancy, the basic unit of an STR should be minimized; e.g. the repeat unit of ACACACAC is AC rather than ACAC. Another reduction method is to merge occurrences of the reverse complement of an STR into the set of the focal STR. Therefore, we call the repeat unit representative if it is not a repeat of a shorter unit and is the first lexicographical motif when all possible shifts of the motif and its reverse complement are considered. Supplementary Table S1 presents the numbers of representative repeat units with typical examples.

2.2 Efficient algorithm for listing approximate STRs in billions of short reads

STRs are inherently 'approximate' in the sense that some unit occurrences are allowed to contain a small number of mutations (Ballantyne *et al.*, 2010). Listing approximate STRs, however, becomes computationally intractable because its time complexity grows exponentially in the maximum number of allowed mutations (Domanic and Preparata, 2007; Pellegrini *et al.*, 2010). Therefore, we use a heuristic approach to this problem. We first identify 'exact' STRs with no mutations in each short read using an efficient O(n log n)-time algorithm (Main, 1989), where n is the length of the read. A *repetition* is any non-empty string of the form $(p)_m q$, where p, a non-empty string, is called the unit of the repetition, $m \geq 2$, and q is a prefix of p. For example,

$$(\mathrm{CAG})_3\mathrm{CA} = \mathrm{CAGCAGCAGCA}$$

is a repetition of the form $(p)_m q$, where $p = \mathrm{CAG}$, $m = 3$ and $q = \mathrm{CA}$, a prefix of p. A repetition is *maximal* if it is not a proper substring of a repetition that has the same unit. For example, consider the following:

$$(\mathrm{CAG})_2\mathrm{CA}(\mathrm{CAG})_2\mathrm{CA} = \mathrm{CAGCAGCACAGCAGCA}$$

$(\mathrm{CAG})_2\mathrm{CA}$, a repetition with unit CAG, is maximal. In addition, the entire string is also a maximal repetition with unit $(\mathrm{CAG})_2\mathrm{CA}$. Listing all maximal repetitions is sufficient to identify all occurrences of STRs. We performed the following steps to retrieve STRs from each read.

(1) Enumerate all maximal repetitions in a read using Main's O(n log n)-time algorithm, where n is the length of the read (Main, 1989). More precisely, in 1984, Main and Lorentz designed an algorithm for enumerating all repetitions of the form xx (Main and Lorentz, 1984). In 1989, Main modified the algorithm to calculate maximal repetitions accurately (Main, 1989), and this is the version that we used to implement our system.

(2) For each maximal repetition Y, identify the minimum unit U such that U is not a repetition and Y is a concatenation of multiple occurrences of U and a prefix of U. For example, when $Y = (\mathrm{CAG})_6\mathrm{CA}$, $U = \mathrm{CAG}$.

(3) An approximate repetition is a substring such that its alignment with repetition $(U)_m$ is decomposed into series of exact matches of length $|U|$ or more, and neighboring series must have only one mismatch, one insertion or one deletion between them in the alignment, where $|U|$ indicates the length of U. We calculate an approximate repetition by extending a maximal (exact) repetition in both directions in a greedy manner. For example, given

CGCCCGCAGCGCAT$(\mathrm{CAG})_6$CATCAGGGA,

we can extend repetition $(\mathrm{CAG})_6\mathrm{CA}$ to the underlined substring,

CGCCCGCAGC-GCAT$(\mathrm{CAG})_6$CATCAGGGA,

where bold letters represent mismatches and '-' indicates a deletion. In this way, we retrieve an approximate STR that is not necessarily an exact repeat of the minimum unit U but may contain mismatches and indels.

(4) A read may contain multiple overlapping STRs with the same unit. If two overlap, eliminate the shorter one. If both are of the same length, select one arbitrarily.

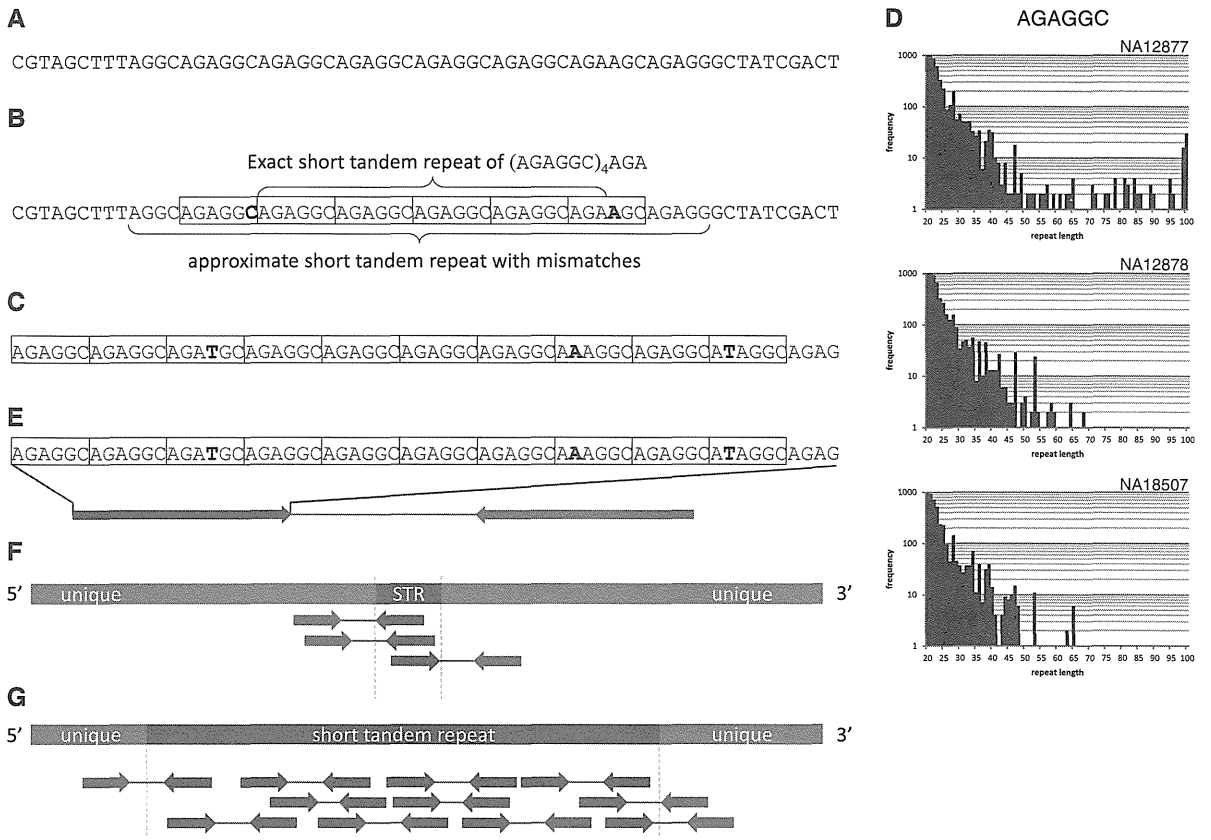

Fig. 1. Sensing and locating STRs in short reads. (**A**) An original short read. (**B**) An approximate STR (AGAGGC)*n* ($n=6$) in the short read. The central four copies of AGAGGC are an exact STR with no mutations, whereas the flanking copies contain the mutations shown in bold letters. If one of the regions (black) surrounding the STR aligns in a unique position, the STR can be located in the genome. (**C**) A read occupied by an approximate STR. (**D**) Sensing STRs from frequency distributions of (AGAGCC)*n* in NA12877 (father of the HapMap CEU trio), NA12878 (mother) and NA18507 (an African male). The *x*-axis is the lengths of STR occurrences detected in a read, and the *y*-axis is the frequency of reads containing STR occurrences of the length indicated on the *x*-axis. Note that 100-bp long STR occurrences are frequent in NA12877, whereas no STR occurrences of length >70 bp are observed in samples NA12878 and NA18507. (**E**) When a read is filled with an STR (red), we attempt to anchor the other end read (blue) to a unique position unambiguously. (**F** and **G**) An STR is located easily if its location can be sandwiched using information on paired-end reads. The length of an STR of length <100 bp is easily estimated (F), whereas determining the length of a much longer STR is non-trivial (G). We need to use third-generation sequencers, such as PacBio RS, with the capability of reading DNA fragments having a length of thousands of bases

The algorithm is able to process 10 million reads of length 100 bases in ~1700 s on a Xeon X5690 with a clock rate of 3.47 GHz (Supplementary Fig. S1). As the computational time is proportional to the number of reads, ~47 h is required to process 1 billion 100-bp reads, confirming the practicality of the method for processing real human resequencing data.

2.3 Sensing expanded STRs by analyzing the frequency distributions of STRs

The computational efficiency of our program facilitates the generation of frequency distributions of all approximate STRs in reads according to their lengths, as illustrated in Figure 1D. We used three samples of the whole-genome resequencing data downloaded from http://www.illumina.com/platinumgenomes/with accession numbers NA12877 (father of the HapMap CEU trio), NA12878 (mother) and NA18507 (an African male). We assumed that short reads were of length 100 bp, which is the typical length of reads output by cost-efficient high-throughput sequencers as of 2013. Although the length will likely increase in the near future, extending our procedure to process longer reads is straightforward because our algorithm runs in O(n log n)-time for processing reads of any length n as stated in the previous subsection. Comparing the distributions of more than one sample sometimes uncovers such a remarkable STR for which occurrences of length 100 bp are frequent in one sample (e.g. NA12877), but are absent in the other two samples, NA12878 and NA18507 (Fig. 1D), suggesting the presence of a long AGAGGC repeat in the former sample (Fig. 1D).

2.4 Reproducibility of detecting STR expansions for independent biological replicates

One might be concerned that despite the presence of a 100-bp long STR in a sample, our method might fail to report this with some probability. We examined this concern using two biological replicates collected independently from an identical DNA sample. The two replicates were independent datasets of 100-bp reads sequenced from the same DNA sample, NA12878, using an Illumina HiSeq2000 (Supplementary Table S2). One dataset was collected by DePristo *et al.* (2011) and the other dataset was downloaded from Illumina's platinum genome Web site (http://www.illumina.com/platinumgenomes/). We applied our method to both biological replicates (Supplementary Table S2) and examined whether 100-bp

occurrences of individual STRs were present simultaneously in both. We identified 60 STRs with 100 bp occurrences in one ($n=13$, 21.7%) or both ($n=47$, 78.3%) replicates of NA12878 (Supplementary Table S3). Of the 13 STRs with no counts in one replicate, 12 had one or two occurrences in the other replicate and the remaining one had four in the other. If an STR occurrence in the genome is short (e.g. 100 bp in length), failure to observe the STR has a high probability (e.g. 50% for 50-fold coverage of reads assuming the random collection of reads). Therefore, our method outputs essentially consistent results for the two biological replicates.

This analysis also indicated that the failure to detect 100-bp occurrences of an STR did not imply the absence of a 100-bp expansion of the STR in the focal personal genome. To be certain of its absence, we examined if the frequency distribution of lengths of STR occurrences was informative. Supplementary Figure S2 presents the frequency distributions of the 13 STRs in the two biological replicates. In most of the 13 STRs, when one biological replicate had 100-bp occurrences of an STR, the other replicate had occurrences of length >90 bp, although for two STRs, the longest occurrences were ~60 bp, which might stem from factors such as amplification bias and variation in sequencing coverage. Therefore, the absence of >60-bp STR occurrences does not necessarily deny the existence of 100 bp expansions of the STR in the genome.

2.5 Locating long expansions of STRs in the human genome

The genomic positions of each uncovered STR in a read remain to be determined. The problem is solvable if one of the two regions flanking an STR maps to a unique position (Fig. 1B), the method used in lobSTR (Gymrek *et al.*, 2012). Otherwise, we attempt to use information on paired-end reads, the two ends of an identical DNA fragment such that their typical average length ranges from 300 to 350 bp with an average standard deviation of ~10%. When one end-read is filled with an STR, we test whether the other end maps to a unique position in the genome using the Burrows–Wheeler Alignment Maximal Exact Matches algorithm (BWA-MEM), a tool for aligning reads with the genome (Li, 2013). If the test is successful, we can approximate the position of the STR from the location of the other end (Fig. 1E). An STR can be located if its location can be sandwiched using information on paired-end reads (Fig. 1F and G). An STR shorter than 100 bp is easier to determine (Fig. 1F), whereas estimating the lengths of longer STRs becomes more difficult (Fig. 1G). We will discuss this issue later in the text.

2.6 TRhist: a tool for sensing and locating STRs from billions of short reads

To assist in the correct positioning of STRs, for a read with an STR instance, our program outputs the repeat unit, length of the STR, number of mutations in the STR, flanking regions surrounding the STR and other paired-end read. With this information, the user can align the flanking regions and other end read to the reference to locate the STR in the genome. Our TRhist program is available at http://trhist.gi.k.u-tokyo.ac.jp/.

2.7 SMRT™ sequencing of expanded STRs

Successful identification of an accurate position for one end provides useful input for other analytical methods, such as repeat-primed polymerase chain reaction (PCR) (Warner *et al.*, 1996) and SMRT™ sequencing (Eid *et al.*, 2009; Loomis *et al.*, 2013), to estimate or determine long expansions of STRs. In particular, SMRT™ sequencing is capable of reading DNA fragments of average length ~5 kb as of 2013 (Fig. 1G). Using this emerging technology, Loomis *et al.* reported the first sequence, 750 CGG repeats, for fragile X syndrome (Loomis *et al.*, 2013). Using SMRT™ sequencing, we amplified the repeat region associated with SCA31 using PCR primers 1.5k-ins-F (5′- ACTCCAACTGGGAT GCAGTTTCTCAAT-3′) and 1.5k-ins-R (5′- TGGAGGAAGGAAAT CAGGTCCCTAAAG-3′).

We will describe the analysis in the Section 3. PCR was performed in a final volume of 50 μl containing 0.2 μM of each primer, 200 μM of each dNTP, 1 mM MgCl2, 1.25 U of PrimeSTAR HS DNA polymerase (Takara Bio, Otsu, Japan) and 100 ng of genomic DNA. The PCR profile comprised an initial denaturing at 95°C for 5 min followed by 30 cycles at 95°C for 20 s and 68°C for 8 min. The PCR product was purified on 0.8% agarose gels and converted to the proprietary SMRTbell™ library format using an RS DNA Template Preparation Kit 2.0 (Pacific Biosciences, Menlo Park, CA). Briefly, the PCR product was end-repaired, and hairpin adapters were ligated using T4 DNA ligase. Incompletely formed SMRTbell™ templates were degraded with a combination of exonuclease III and VII. The resulting DNA templates were purified using SPRI magnetic beads (AMPure; Agencourt Bioscience, Beverly, MA). Annealing was performed at a final template concentration of 5 nM, with a 20-fold molar excess of sequencing primer. The annealing reaction was carried out for 2 min at 80°C with slow cooling to 25°C. Annealed templates were stored at −20°C until polymerase binding. The DNA polymerase enzymes stably were bound to the primed sites of the annealed SMRTbell™ templates using the DNA Polymerase Binding Kit 2.0 (Pacific Biosciences). SMRTbell™ template (3 nM) was incubated with polymerase in the presence of phospholinked (Pacific Biosciences) nucleotides for 4 h at 30°C. Following incubation, the samples were stored at 4°C. Sequencing was performed within 36 h of binding. Samples were sequenced using commercial sequencing chemistry. Sequencing data were collected on a PacBio RS (Pacific Biosciences) for 90 min. Given PacBio RS-filtered subreads, we used the SMRT Pipe, P_ErrorCorrection module to generate corrected reads. Subsequently, we assembled these corrected reads using RS_CeleraAssembler to obtain contigs.

3 RESULTS

Here, we demonstrate the utility of an *ab initio* procedure for sensing, locating and sequencing STRs that are significantly expanded in the case sample.

Locating candidate STR positions

Select positions where STR occurrences are expanded significantly in the case sample in the following manner:

(1) Locate occurrences of each candidate STR in both the case and control samples by anchoring paired-end reads such that one end has a ≥50-bp occurrence of the STR and the other end maps to a unique position.

(2) Group paired-end reads anchored in a neighborhood (within ~300 bp, the average insert size of paired-end reads) into one cluster (Fig. 2).

(3) In each cluster, generate the frequency distribution of STR occurrences according to their lengths ranging from 50 to 100 bp (Fig. 2). If an STR in the cluster is significantly longer than 100 bp, the frequency of 100-bp occurrences in reads, denoted by f_{100}, becomes significantly greater than the frequencies of those shorter than 100 bp (Fig. 2B). We test this hypothesis statistically by checking if f_{100} is an outlier in the frequency distribution with the Smirnov–Grubbs' test. We calculate the t-score, $(f_{100}-\mu)/\sigma$, where μ and σ are the mean and standard deviation of the frequency distribution, respectively, and obtain the probability (P-value) that the t-score exceeds a

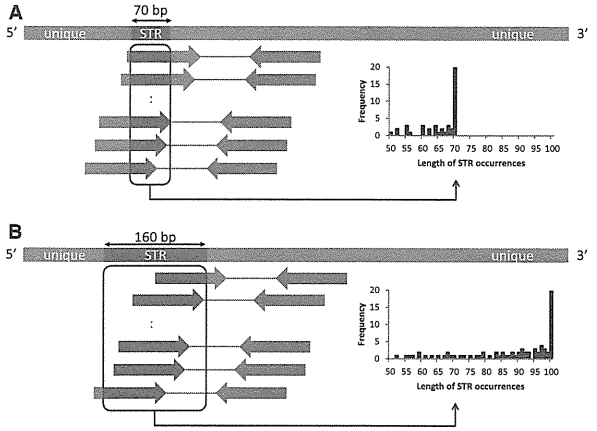

Fig. 2. Select positions where STR occurrences are expanded significantly. (**A**) We generate the frequency distribution of lengths of STR occurrences in paired-end reads. This picture shows the case of a 70-bp long STR. The histogram of the frequency distribution peaks at 70 bp. (**B**) When the STR is 160-bp long, the distribution has a significant peak at 100 bp. We test if the peak is a significant outlier in the frequency distribution using the Smirnov–Grubbs' test

threshold according to the Smirnov–Grubbs' test. For example, the $P < 5 \times 10^{-9}$ when the t-score is >5.27.

(4) We consider ~10 million non-overlapping regions of length 300 bp (the average insert size of paired-end reads) in the human genome. We perform multiple hypothesis testing using the Bonferroni correction to test if each 300-bp region has a significant STR expansion in the case sample at a significance level of 5% divided by 10 million (i.e. 5×10^{-9}). We select positions such that $P < 5 \times 10^{-9}$ in the case sample but no 100-bp STR occurrences are present in any of the control samples. We can relax the condition to consider more candidates with less evidence.

Sequencing candidate STR positions

SMRT™ sequencing of expanded STRs is performed using information on the boundaries of individual STR positions.

3.1 A rare STR significantly expanded in the case sample

To demonstrate the effectiveness of this approach, we first examined a well-characterized case sample, SCA31 (Sato *et al.*, 2009), which contains long expansions of two STRs, (AAAATAGAAT) repeat and (AATGG) repeat, in the introns of genes BEAN1 and TK2 (Chr.16 66,524,303 in hg19), where the reference genome has an (AAAAT) repeat.

We resequenced the genome of a sample from an individual whose parent is a case of SCA31 using an Illumina HiSeq2000 (Supplementary Table S4). All primary sequencing data of the SCA31 sample will be made available under controlled access through the DNA Databank of Japan (DDBJ; accession number JGAS00000000002). We examined whether we could find these STRs with no prior information. We applied the *ab initio* procedure to SCA31 as the case sample, and NA12877, NA12878 and NA18507 as control samples (Fig. 3A). Our procedure detected only one STR; AAAATAGAAT ($P = 1.07 \times 10^{-19}$).

Figure 3B shows the frequency distributions of the (AAAATAGAAT) repeat, supporting the presence of long occurrences of the STR in SCA31 and the absence of long occurrences of length >60 bp in the other control samples. Supplementary Figure S3A shows the distributions of the (AATGG) repeat, but the difference between SCA31 and the other samples was unclear because the (AATGG) repeat is enriched in human centromeres (Grady *et al.*, 1992). Therefore, our *ab initio* analysis suggests that long occurrences of the (AAAATAGAAT) repeat characterize SCA31, consistent with reported observations (Sato *et al.*, 2009). Arguably, we could detect the (AAAATAGAAT) repeat as an approximate (AAAAT) repeat because the last half, AGAAT, is identical to (AAAAT), except for the second base G; therefore, we analyzed the frequency distribution of the (AAAAT) repeat to determine the remarkable expansion of the (AAAAT) repeat in SCA31. This failed due to numerous long instances of the (AAAAT) repeat in all samples (Supplementary Fig. S3B). This example indicates the importance of looking at STRs of repeat units longer than 2–6-base units, to determine expansions of STRs associated with cases.

We also examined the frequency distributions of other well-characterized repeats, such as the (GGGTTA) repeat in telomeres (Supplementary Fig. S3C), (CAG) repeat encoding polyglutamine stretches in protein coding regions (La Spada *et al.*, 1991; The Huntington's Disease Collaborative Research Group, 1993; Walker, 2007 and Supplementary Fig. S4A), (CCTG) repeat associated with myotonic dystrophy type 2 (DM2; Liquori *et al.*, 2001 and Supplementary Fig. S4B) and (ATTCT) repeat associated with spinocerebellar ataxia type 10 (SCA10; Matsuura *et al.*, 2000 and Supplementary Fig. S4C). For the last three repeats, no significant differences were detected between SCA31 and the three control samples, suggesting that these three repeats are not associated with SCA31.

Using paired-end reads with AAAATAGAAT repeats at their 5′ ends and uniquely mapped reads at their 3′ ends, we could determine the 3′ end of the insertion. Figure 4A shows how we locate a ~2.5–3.8 kb insertion of the repeat associated with the SCA31 sample (Sato *et al.*, 2009).

We sequenced the repeat region in 11 SCA31 samples using SMRT™ sequencing. We designed a pair of PCR primers around the candidate repeat region in the SCA31 sample the right boundary of which could be determined. As illustrated in Figure 4A, we could identify the right boundary in the reference genome because the right ends of many paired-end reads mapped to the downstream region of the right boundary, whereas the left ends did not. We could sequence the candidate repeat region. Supplementary Table S6 presents the statistics of filtered subreads, corrected subreads and assembled contigs. Previously, Sato *et al.* estimated a 2.5–3.8 kb insertion of the following form for an SCA31 sample (Sato *et al.*, 2009):

$$\mathrm{TCACTAAAA(TAGAA)}_i\mathrm{(TGGAA)}_j$$
$$\mathrm{(TAGAA|TGGAA|TGGGAA)}_k$$
$$\mathrm{(TAGAA)}_l\mathrm{(TAAAATAGAA)}_m,$$

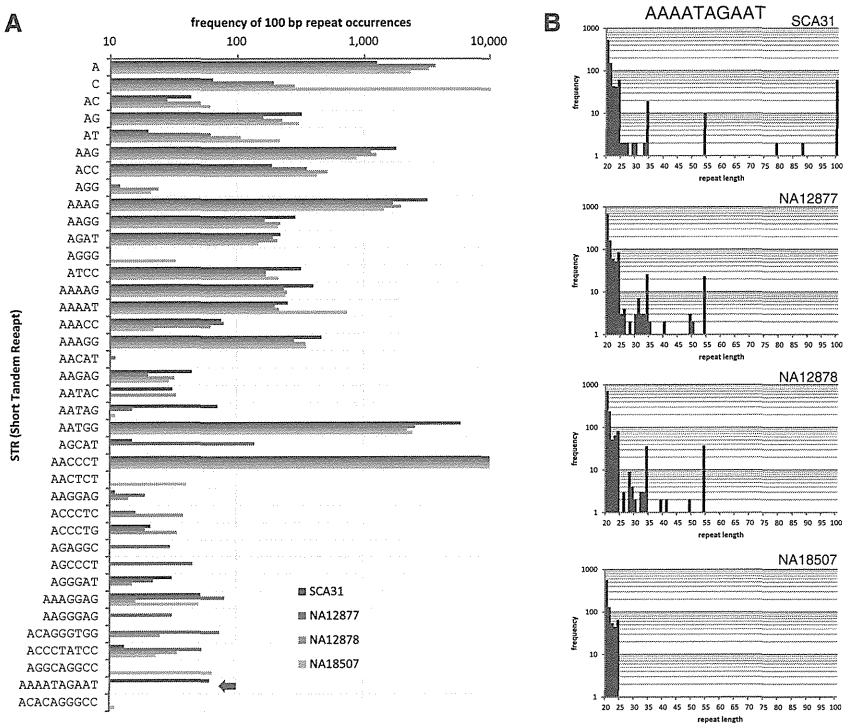

Fig. 3. Sensing expanded STRs associated with SCA31. (**A**) Frequencies of 100-bp STRs that have >10 occurrences in one of SCA31, NA12877, NA12878 or NA18507. For example, the arrow in the second lowest row shows that the (AAAATAGAAT) repeat is expanded only in SCA31. Our *ab initio* procedure analyzes this bar chart and selects STRs that are significantly abundant in the case sample (e.g., SCA31) but absent in all of the control samples. The bar chart is also useful for confirming the abundance of (AATGG) and (AACCCT) repeats, equivalent to the (GGGTTA) repeat, where the former and latter motifs are known to be enriched in centromeres and telomeres, respectively. (**B**) Frequency distributions of the (AAAATAGAAT) repeat. SCA31 has many 100-bp occurrences, whereas no occurrences of length >55 bp were observed in NA12877, NA12878 and NA18507

where (TAGAA | TGGAA | TGGGAA)$_k$ is a series of k occurrences of TAGAA, TGGAA and TGGGAA. In their sample, they determined that $i=2$, $k=10$ and $l=46$, but left j and m undetermined because both appeared to be extremely long. In our 11 SCA31 samples, we could determine the values of j and m. We found that the numbers of individual repeats varied markedly ($i=1\sim2, j=220\sim321, k=9\sim13, l=42\sim78$ and $m=90\sim118$) and the insertion size ranged from 2350 to 3088 b (Fig. 4C and Supplementary Table S5), demonstrating the instability of the STR expansion in SCA31. In particular, two STRs, (TAGAA)j and (TAAAA TAGAA)m, form ~90% of the entire repeat expansion, and the values of j and m are positively correlated (correlation coefficient $r=0.70$), implying that these two values are the determinants of the instability of the repeat expansions in SCA31 (Fig. 4D). In all samples, the repeat expansion was present in one allele, but was absent in the other. Note that the numbers of STR units might not be exact because PCR for repeat regions can introduce more replication errors than those produced by bacterial DNA replication (Loomis *et al.*, 2013).

3.2 Common STRs significantly expanded in the case sample

We also applied our procedure to the SCA31 data, and examined common STRs, AAAG, ATCC, AAAAG, AATAG and AATGG, present in both the case and control samples but significantly expanded in the case sample. We identified STR expansions at 11 genomic locations that were significantly expanded in the case sample ($P<5\times10^{-19}$ and Supplementary Fig. S5). We then used SMRT^{TM} sequencing to confirm the four expanded STRs in the case sample that were significantly longer than the corresponding STR occurrences in the reference genome (Fig. 5 and Supplementary Fig. S6). No false-positive expansions were found in this experiment, suggesting that the false-positive rate of the procedure is generally low.

4 DISCUSSION

STRs in personal genomes remain largely uncharacterized. We proposed a novel method for listing long approximate STRs with mutations in personal genomes using a massive number of short reads of length ~100 bp. Here, we discuss some situations in which detecting a long expansion of STRs specific to disease samples is inherently problematic. As genomic regions of GC content >70% are difficult to cover with an ample number of Illumina reads, our method is unlikely to detect long expansions of STRs with high GC contents. STRs in reads originating in centromeres, telomeres or retrotransposons are too numerous to map to unique genomic positions. As illustrated in

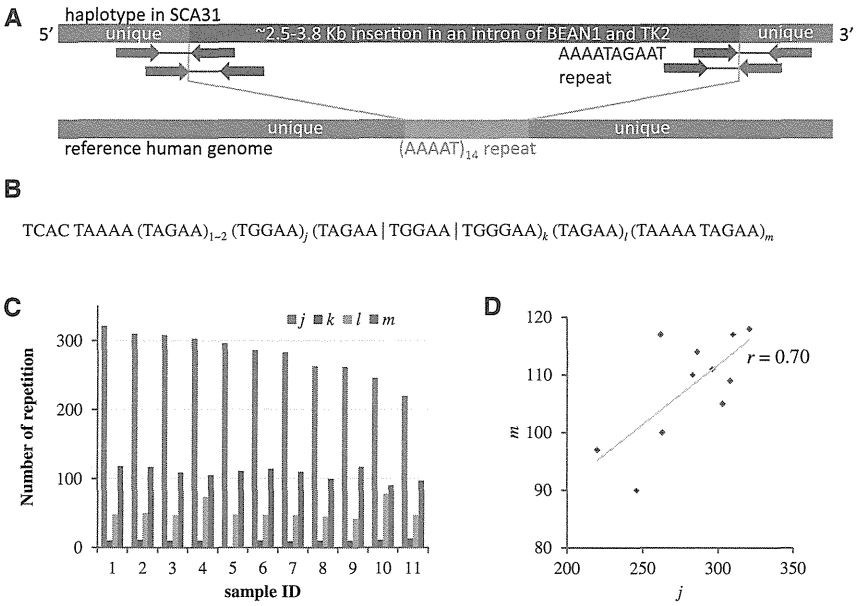

Fig. 4. Locating and sequencing expanded STRs associated with SCA31. (**A**) A real example from SCA31. One haplotype contains a ~2.5–3.8 kb insertion at Chr.16 66 524 303 in hg19 in an intron of BEAN1 and TK2. The right boundary of the insertion could be identified using paired-end reads with AAAATAGAAT repeats at their left ends and uniquely mapped reads at their right ends. The lower bar illustrates the reference genome (hg19) with an AAAAT repeat. (**B**) A form of expanded repeat associated with SCA31 samples. The values of *i*, *j*, *l* and *m* vary in the individual SCA31 samples. (**C**) We determined the values of *i*, *j*, *l* and *m* in 11 SCA31 samples using SMRT^{TM} sequencing. This shows that ~90% of the repeat expansion are $(TAGAA)_j$ and $(TAAAA\ TAGAA)_m$. (**D**) The values of ***j*** and ***m*** are positively correlated ($r = 0.70$). These two values are the determinants of the instability of the repeat expansions in SCA31

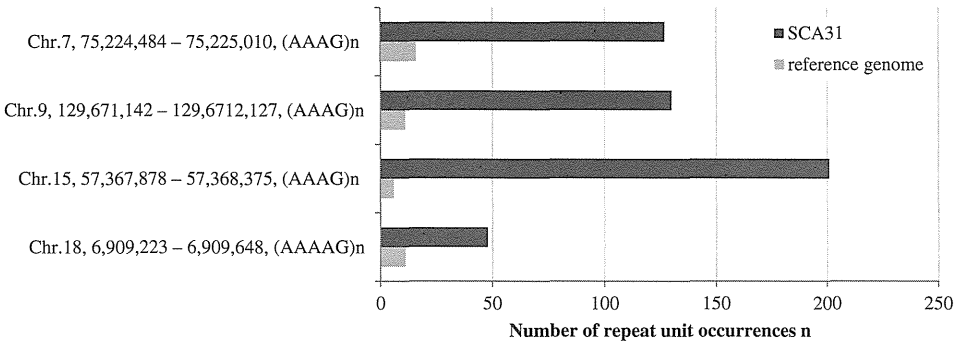

Fig. 5. Sizes of the common STRs, (AAAG)*n* and (AAAAG)*n*, at four genomic positions in the SCA31 sample and reference genome. Note that individual STR occurrences are significantly expanded in the SCA31 sample. The PCR primers used for amplifying individual regions and the sequences of amplicons can be found in Supplementary Figure S6

Supplementary Figure S3, massive numbers of long expansions of these STRs can be found in any sample.

We also presented an *ab initio* procedure for detecting significant expansions of STRs in case samples that are absent in control samples via comparisons between the frequency distributions of STRs in case and control samples. We demonstrated the potential applicability of this method using three publicly available control samples. To exploit this approach, however, constructing a large-scale database of the frequency distributions of STRs collected from a number of control samples is necessary.

The variety of expanded STRs of length >1 kb in disease remains unexplored. Also, examining whether expansions of STRs are more pronounced in germline and somatic cells would be intriguing. Thus, after locating STRs, sequencing expanded STRs is a promising direction of study. For this purpose, SMRT^{TM} sequencing enables the sequencing DNA fragments averaging ~5 kb long as of 2013. Using SMRT^{TM} sequencing, we were able to determine a divergent set of 2.3–3.1 kb STR sequences in 11 SCA31 samples, showing the instability of STR expansions. Analysis of the stability of STR expansions

in germline and somatic cells of a specific disease might eventually lead to the recognition of a functional role of STRs.

In the near future, the typical lengths of short reads in the majority of commercial sequencers should increase to 150–500 bases. Our method is ready to process longer reads in a straightforward manner. Furthermore, our method was designed so that it could output STRs of repeat units of any length, and we presented an illustrative case in which detecting STRs of a 10-base repeat unit from an SCA31 sample was essential. Our program will serve as a valuable tool for discovering unknown STRs in a variety of diseases, even with future advances in sequencing technology.

ACKNOWLEDGEMENTS

The authors are grateful to an anonymous reviewer for suggesting a way of detecting expansions of common STRs and for many valuable comments on the manuscript.

Funding: Grant-in-Aid for Scientific Research on Innovative Areas (22129008, 221S0002 to S.M.) (in part) from the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT) and by the Global COE program (Deciphering Biosphere from Genome Big Bang) to S.M. from the MEXT.

Conflict of Interest: none declared.

REFERENCES

Ballantyne,K.N. *et al.* (2010) Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. *Am. J. Hum. Genet.*, **87**, 341–353.

Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.

Brook,J.D. *et al.* (1992) Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. *Cell*, **69**, 385.

Conrad,D.F. *et al.* (2011) Variation in genome-wide mutation rates within and between human families. *Nat.Genet.*, **43**, 712–714.

DeJesus-Hernandez,M. *et al.* (2011) Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron*, **72**, 245–256.

DePristo,M.A. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, **43**, 491–498.

Domanic,N.O. and Preparata,F.P. (2007) A novel approach to the detection of genomic approximate tandem repeats in the Levenshtein metric. *J. Comput. Biol.*, **14**, 873–891.

Eid,J. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.

Grady,D.L. *et al.* (1992) Highly conserved repetitive DNA sequences are present at human centromeres. *Proc. Natl Acad. Sci. USA*, **89**, 1695–1699.

Gymrek,M. *et al.* (2012) lobSTR: a short tandem repeat profiler for personal genomes. *Genome Res.*, **22**, 1154–1162.

Jorda,J. and Kajava,A.V. (2009) T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. *Bioinformatics*, **25**, 2632–2638.

Kobayashi,H. *et al.* (2011) Expansion of intronic GGCCTG hexanucleotide repeat in NOP56 causes SCA36, a type of spinocerebellar ataxia accompanied by motor neuron involvement. *Am. J. Hum. Genet.*, **89**, 121–130.

Kolpakov,R. *et al.* (2003) mreps: Efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res.*, **31**, 3672–3678.

Kong,A. *et al.* (2012) Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature*, **488**, 471–475.

Kremer,E.J. *et al.* (1991) Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. *Science*, **252**, 1711–1714.

La Spada,A.R. *et al.* (1991) Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature*, **352**, 77–79.

Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*.

Lim,K.G. *et al.* (2013) Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance. *Brief. Bioinform.*, **14**, 67–81.

Liquori,C.L. *et al.* (2001) Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. *Science*, **293**, 864–867.

Loomis,E.W. *et al.* (2013) Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. *Genome Res.*, **23**, 121–128.

Lupski,J.R. (2007) Genomic rearrangements and sporadic disease. *Nat. Genet.*, **39**, S43–S47.

Mahadevan,M. *et al.* (1992) Myotonic dystrophy mutation: an unstable CTG repeat in the 3′ untranslated region of the gene. *Science*, **255**, 1253–1255.

Main,M.G. (1989) Detecting leftmost maximal periodisities. *Discrete Appl. Math.*, **25**, 145–153.

Main,M.G. and Lorentz,R.J. (1984) an O(n log n) algorithm for finding all repetitions in a string. *J. Algorithm.*, 422–432.

Matsuura,T. *et al.* (2000) Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. *Nat. Genet.*, **26**, 191–194.

Mirkin,S.M. (2007) Expandable DNA repeats and human disease. *Nature*, **447**, 932–940.

Mudunuri,S.B. and Nagarajaram,H.A. (2007) IMEx: Imperfect Microsatellite Extractor. *Bioinformatics*, **23**, 1181–1187.

Orr,H.T. (2011) FTD and ALS: genetic ties that bind. *Neuron*, **72**, 189–190.

Pellegrini,M. *et al.* (2010) TRStalker: an efficient heuristic for finding fuzzy tandem repeats. *Bioinformatics*, **26**, i358–i366.

Renton,A.E. *et al.* (2011) A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron*, **72**, 257–268.

Sato,N. *et al.* (2009) Spinocerebellar ataxia type 31 is associated with "inserted" penta-nucleotide repeats containing (TGGAA)n. *Am. J. Hum. Genet.*, **85**, 544–557.

Sherman,S.L. *et al.* (1985) Further segregation analysis of the fragile X syndrome with special reference to transmitting males. *Hum Genet*, **69**, 289–299.

The Huntington's Disease Collaborative Research Group. (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. *Cell*, **72**, 971–983.

Verkerk,A.J. *et al.* (1991) Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell*, **65**, 905–914.

Walker,F.O. (2007) Huntington's disease. *Lancet*, **369**, 218–228.

Warner,J.P. *et al.* (1996) A general method for the detection of large CAG repeat expansions by fluorescent PCR. *J. Med. Genet.*, **33**, 1022–1026.

Wexler,Y. *et al.* (2005) Finding approximate tandem repeats in genomic sequences. *J. Comput. Biol.*, **12**, 928–942.

Wojciechowska,M. and Krzyzosiak,W.J. (2011) Cellular toxicity of expanded RNA repeats: focus on RNA foci. *Hum. Mol. Genet.*, **20**, 3811–3821.

Please cite this article in press as: Takahashi et al., *ERBB4* Mutations that Disrupt the Neuregulin-ErbB4 Pathway Cause Amyotrophic Lateral Sclerosis Type 19, The American Journal of Human Genetics (2013), http://dx.doi.org/10.1016/j.ajhg.2013.09.008

REPORT

ERBB4 Mutations that Disrupt the Neuregulin-ErbB4 Pathway Cause Amyotrophic Lateral Sclerosis Type 19

Yuji Takahashi,[1] Yoko Fukuda,[1] Jun Yoshimura,[2] Atsushi Toyoda,[3] Kari Kurppa,[4,5] Hiroyoko Moritoyo,[6] Veronique V. Belzil,[7] Patrick A. Dion,[7,8] Koichiro Higasa,[2] Koichiro Doi,[2] Hiroyuki Ishiura,[1] Jun Mitsui,[1] Hidetoshi Date,[1] Budrul Ahsan,[1] Takashi Matsukawa,[1] Yaeko Ichikawa,[1] Takashi Moritoyo,[6] Mayumi Ikoma,[9] Tsukasa Hashimoto,[9] Fumiharu Kimura,[10] Shigeo Murayama,[11] Osamu Onodera,[12] Masatoyo Nishizawa,[12] Mari Yoshida,[13] Naoki Atsuta,[14] Gen Sobue,[14] JaCALS,[15] Jennifer A. Fifita,[16,17,18] Kelly L. Williams,[16,17,18] Ian P. Blair,[16,17,18] Garth A. Nicholson,[16,17] Paloma Gonzalez-Perez,[19] Robert H. Brown, Jr.,[19] Masahiro Nomoto,[6] Klaus Elenius,[4,20] Guy A. Rouleau,[7,21,22] Asao Fujiyama,[3] Shinichi Morishita,[2] Jun Goto,[1] and Shoji Tsuji[1,23,*]

Amyotrophic lateral sclerosis (ALS) is a devastating neurological disorder characterized by the degeneration of motor neurons and typically results in death within 3–5 years from onset. Familial ALS (FALS) comprises 5%–10% of ALS cases, and the identification of genes associated with FALS is indispensable to elucidating the molecular pathogenesis. We identified a Japanese family affected by late-onset, autosomal-dominant ALS in which mutations in genes known to be associated with FALS were excluded. A whole- genome sequencing and parametric linkage analysis under the assumption of an autosomal-dominant mode of inheritance with incomplete penetrance revealed the mutation c.2780G>A (p. Arg927Gln) in *ERBB4*. An extensive mutational analysis revealed the same mutation in a Canadian individual with familial ALS and a de novo mutation, c.3823C>T (p. Arg1275Trp), in a Japanese simplex case. These amino acid substitutions involve amino acids highly conserved among species, are predicted as probably damaging, and are located within a tyrosine kinase domain (p. Arg927Gln) or a C-terminal domain (p. Arg1275Trp), both of which mediate essential functions of ErbB4 as a receptor tyrosine kinase. Functional analysis revealed that these mutations led to a reduced autophosphorylation of ErbB4 upon neuregulin-1 (NRG-1) stimulation. Clinical presentations of the individuals with mutations were characterized by the involvement of both upper and lower motor neurons, a lack of obvious cognitive dysfunction, and relatively slow progression. This study indicates that disruption of the neuregulin-ErbB4 pathway is involved in the pathogenesis of ALS and potentially paves the way for the development of innovative therapeutic strategies such using NRGs or their agonists to upregulate ErbB4 functions.

Amyotrophic lateral sclerosis (ALS) is a devastating neurological disorder in which the degeneration of motor neurons leads to progressive weakness and wasting of limb, bulbar, and respiratory muscles. Familial ALS (FALS) comprises 5%–10% of ALS cases, and the remaining cases are simplex cases of ALS (SALS). To date, more than 20 genes have been shown to be associated with ALS,[1] and these account for 75% of FALS and 14% of SALS cases.[2] Mutations that are found in FALS-associated genes but that are also identified in individuals with SALS are considered mutations with reduced penetrance or de novo mutations. Further discovery of genes associated with FALS is indispensable to elucidating the molecular backgrounds of both FALS and SALS.

Identification of genes associated with familial diseases has been accomplished through identification of the disease loci on the chromosomes by linkage analysis of large pedigrees and subsequent positional cloning of the genes. The majority of the FALS pedigrees, however, are not large and do not have multiple affected members as a result of the poor prognosis of the disease and the late age of onset, which makes it difficult to sufficiently narrow the candidate regions by linkage analyses and means that it takes a tremendous effort to identify the genes associated with FALS. The recent development of massively parallel

[1]Department of Neurology, Graduate School of Medicine, The University of Tokyo, Tokyo 113-8655, Japan; [2]Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, Chiba 277-8561, Japan; [3]Comparative Genomics Laboratory, National Institute of Genetics, Shizuoka 411-8540, Japan; [4]Department of Medical Biochemistry and Genetics, University of Turku, Turku, 20014, Finland; [5]Turku Doctoral Programme of Biomedical Sciences, Turku, 20520, Finland; [6]Department of Neurology and Clinical Pharmacology, Ehime University Hospital, Ehime 791-0295, Japan; [7]Research Center, Centre Hospitalier Universitaire Sainte-Justine, Université de Montreal, Montreal, QC, H3T 1C5, Canada; [8]Department of Pathology and Cellular Biology, Université de Montréal, Montreal, QC H3T 1C5, Canada; [9]Department of Neurology, National Hospital Organization Ehime National Hospital, Ehime 791-0281, Japan; [10]Division of Neurology, First Department of Internal Medicine, Osaka Medical College, Osaka 569-8686, Japan; [11]Department of Neurology and Neuropathology and Brain Bank for Aging Research, Tokyo Metropolitan Geriatric Hospital and Institute of Gerontology, Tokyo 173-0015, Japan; [12]Department of Neurology, Brain Research Institute, Niigata University, Niigata 951-8520, Japan; [13]Department of Neuropathology, Institute for Medical Science of Aging, Aichi Medical University, Aichi 480-1195, Japan; [14]Department of Neurology, Nagoya University Graduate School of Medicine, Nagoya 466-8560, Japan; [15]Japanese Consortium for Amyotrophic Lateral Sclerosis Research; [16]Northcott Neuroscience Laboratory, ANZAC Research Institute, Sydney, New South Wales 2139, Australia; [17]Sydney Medical School, University of Sydney, New South Wales 2006, Australia; [18]Austrian School of Medicine, Macquarie University, Sydney, New South Wales 2109, Australia; [19]Department of Neurology, University of Massachusetts Medical School, Worcester, MA 01655-0318, USA; [20]Department of Oncology, Turku University Hospital, Turku 20521, Finland; [21]Montreal Neurological Institute, McGill University, Montreal, QC H3A 2B4, Canada; [22]Department of Neurology and Neurosurgery, McGill University, Montreal, QC H3A 2B4, Canada; [23]Medical Genome Center, The University of Tokyo Hospital, The University of Tokyo, Tokyo 113-8655, Japan
*Correspondence: tsuji@m.u-tokyo.ac.jp
http://dx.doi.org/10.1016/j.ajhg.2013.09.008.

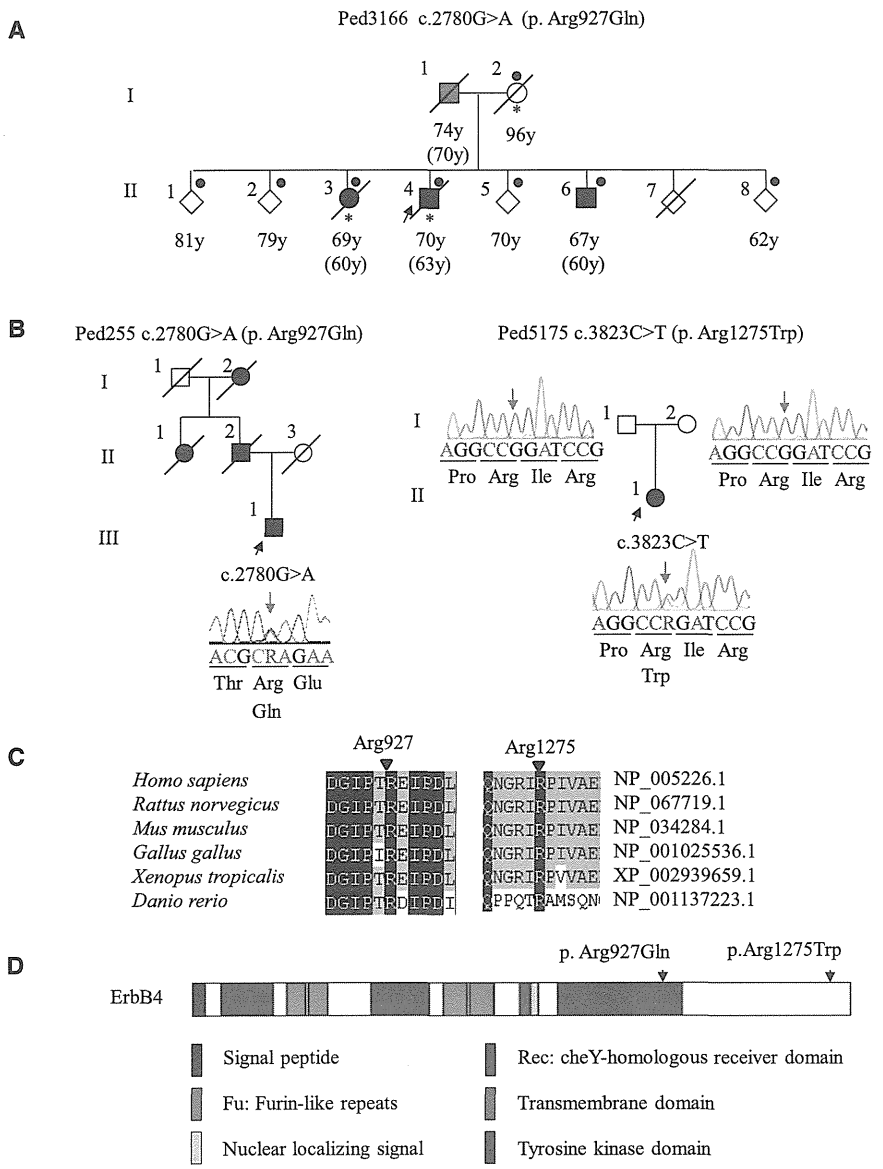

Figure 1. Pedigrees of ALS and Characterization of Mutations

(A) Pedigree charts of the index family. Filled symbols indicate affected individuals. The arrow indicates the proband. For confidentiality purposes, all unaffected siblings are indicated by diamonds. Dots or asterisks indicate individuals included in the linkage study or WGS, respectively. Age at present or age at death is shown under each individual, and ages at onset are shown in parentheses. The box with gray shading indicates that the individual's clinical information obtained from the family members strongly supports the diagnosis of ALS, although detailed neurological evaluations have not been conducted for this individual.

(B) Additional Canadian (Ped255) and Japanese (Ped5175) pedigrees with *ERBB4* mutations. The electropherograms of mutational data are shown beside each member. Nucleotide colors correspond to the colors in the electropherograms. The amino acids are designated below the nucleotide sequences. The blue arrows indicate the nucleotide positions of the mutations. In the electropherograms (Ped5175), nucleotide sequences of the reverse complementary strand are shown.

(C) Amino acid conservation. The amino acids Arg927 and Arg1275 are highly conserved among species.

(D) The protein structure along with the locations of amino acid substitutions are shown; amino acid substitutions are indicated by arrows. The amino acid substitution p. Arg927Gln resides in the tyrosine kinase domain, which mediates the key functions of ErbB4. The amino acid substitution p. Arg1275Trp resides in the C-terminal domain in the vicinity of multiple phosphorylation sites, which mediate downstream signaling pathways.

sequencing technologies has allowed us to overcome the difficulty by means of whole-genome sequencing (WGS) or exome analysis.

We identified a Japanese family with three affected siblings presenting with late-onset ALS (Figure 1A and Table 1). The familial history indicated that the mode of inheritance is probably an autosomal-dominant one. Mutational analysis of the proband (II-4) employing direct nucleotide sequence analysis, a microarray-based resequencing, or a repeat-primed PCR analysis excluded *SOD1*[MIM 147450], *ALS2*[MIM 606352], *DCTN1*[MIM 601143], *CHMP2B*[MIM 609512], *ANG*[MIM 105850], *TARDBP*[MIM 605078], *FUS*[MIM 137070] and *C9ORF72* [MIM 1614260] as the genes associated with FALS.[3,4] To identify a gene associated with FALS, we applied WGS in combination with a linkage analysis to the pedigree. Written informed consent was obtained from all the participants. This study was approved by the institutional review board at the University of Tokyo.

WGS was performed on three individuals (I-2, II-3 and II-4, as shown in Figure 1A) in the index pedigree. Paired-end DNA libraries were generated and subjected to massively parallel sequencing with a GAII Illumina Genome Analyzer in accordance with the manufacturer's instructions. The short read sequences obtained were aligned to the reference genome (NCBI37/hg19 assembly) via the Burrows-Wheeler Aligner.[5] Downstream analyses in which potential PCR duplicates were removed were processed with SAMtools.[6] Aligned reads were viewed on an Integrative Genomics Viewer.[7] Genomic sequence variations were identified with the SAMtools pileup command and annotated with Refseq, dbSNP135, 1000 Genomes, personal genome databases, the NHLBI GO Exome Sequencing Project (NHLBI-ESP) database, and an in-house variant database containing 41 whole genomes and 1,408 exomes in the Japanese population. The numbers of nonsynonymous variants that were identified in individuals I-2, II-3, and II-4 but that were not present in any of the

Please cite this article in press as: Takahashi et al., *ERBB4* Mutations that Disrupt the Neuregulin-ErbB4 Pathway Cause Amyotrophic Lateral Sclerosis Type 19, The American Journal of Human Genetics (2013), http://dx.doi.org/10.1016/j.ajhg.2013.09.008

Table 1. Clinical Characteristics of Affected Individuals

Pedigree Number	Pedigree 3166				Pedigree 255	Pedigree 5175
Ethnicity	Japanese				Canadian	Japanese
Inheritance	familial (autosomal dominant)				familial (autosomal dominant)	simplex
Mutation	c.2780G>A				c.2780G>A	c.3823C>T
Amino acid substitution	p. Arg927Gln				p. Arg927Gln	p. Arg1275Trp
Members	I-1	II-3	II-4 (proband)	II-6	III-3	II-1
Age at onset	70	60	63	60	67	45
Initial symptoms	bulbar	N.D.	upper limbs	respiration	upper limbs	upper limbs
Diagnostic criteria[a]	N.D.	N.D.	definite	definite	probable	probable
Progression	unable to walk after 3 years	ventilator -dependent after 5 years, locked-in state after 8 years	locked-in state after 5 years	ventilator- dependent after 1 year, locked-in state after 5 years	slow progression that significantly decelerated and finally stopped after 8 years	wheelchair- bound, MRS 1-2/5 in upper extremities after 5 years
Cognitive function	N.D.	N.D.	normal	normal	N.D.	normal
Age at death	74	69	70	66	N/A	N/A

Abbreviations are as follows: N.D., not described; MRS, manual muscle testing rating scale; and N/A, not applicable.
[a]El Escorial and Airlie House revised criteria.

databases (hereafter, variants not found in the databases are referred to as "novel") were 411, 404, and 382, respectively (Table S1). No novel nonsynonymous variants in genes known to be associated with FALS were included. Among the identified variants, 57 were identified both in the proband and in the affected sibling, but not in the mother, and were subjected to further analysis.

The individuals indicated by dots in Figure 1A were genotyped with Genome-Wide Human SNP Array 6.0 (Affymetrix). Linkage analysis and haplotype reconstruction were conducted with the pipeline software SNP-HiTLink[8] and Allegro version 2[9] under the assumption of an autosomal-dominant mode of inheritance and a disease-allele frequency of 0.000001. Parametric multipoint linkage analysis under the assumption of complete penetrance revealed three loci spanning 23.6 Mb on chromosomes 1, 6, and 13, having a maximum LOD score of 1.8 (Figure S1; penetrance = 1.0), and containing 88 annotated genes. However, no novel nonsynonymous variants were identified in the candidate regions. We then considered the possibility of reduced penetrance. When penetrance was reduced to 0.8 (Figure S1), seven additional loci had LOD scores > 0.7 and were thus shown to support linkage; these loci contained 809 annotated genes. Three heterozygous novel nonsynonymous variants were identified in these regions; among these variants, only c.2780G>A (p. Arg927Gln; dbSNP SubSNP ID ss831884245) substituting glutamine for arginine at codon 927 (p. Arg927Gln) in v-erb-a erythroblastic leukemia viral oncogene homolog 4 (avian) (*ERBB4* [MIM 600543; RefSeq accession number NM_005235.2]) was not present in 477 controls (Table S2). When we allowed further reduced penetrance, we identified 19 additional loci with LOD > 0; these loci contained 1,265 annotated genes. In these regions, we identified seven heterozygous novel nonsynonymous variants, among which three variants in *OR2D3* (RefSeq NM_001004684.1), *FTCD* (MIM 606806; RefSeq NM_206965.1), and *TJP2* (MIM 607709; RefSeq NM_001170414.2) were not present in 477 controls (Table S2). *OR2D3* is an olfactory receptor gene; the substituted amino acid in *OR2D3* is not conserved, and the substitution is predicted as benign by PolyPhen-2 analysis. *FTCD* and *TJP2* are associated with autosomal-recessive glutamate formiminotransferase deficiency (MIM 229100) and familial hypercholanemia (MIM 607748), respectively, and heterozygous carriers have not been described as exhibiting ALS. Taken together, the results pointed to c.2780G>A in *ERBB4* as the most likely pathogenic mutation.

We used a direct nucleotide sequence analysis method to conduct mutational analysis of *ERBB4* in 364 FALS and 818 SALS individuals by using an ABI 3100 sequencer and BigDye Terminator ver3.1 (Applied Biosystems). We used the ExonPrimer website to design oligonucleotide primers (Table S3). The mutation c.2780G>A was also identified in one Canadian FALS individual (Figure 1B). Unfortunately, DNA from other family members was not available to confirm segregation. To investigate a possibility that the c.2780G>A mutation identified in the Japanese and Canadian families is a common founder mutation, we compared the haplotypes with the c.2780G>A mutation in *ERBB4* of the Japanese and Canadian families (Figure S2). Different SNPs were observed 14 kbp and 5 kbp centromeric and telomeric to the mutation, respectively, indicating that disease haplotypes of the Japanese and Canadian families are different and that

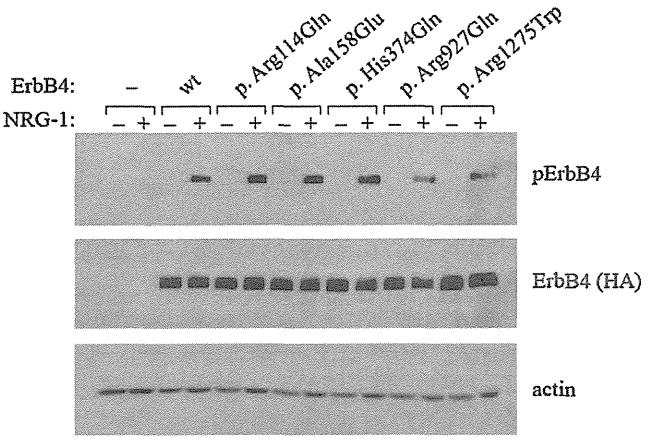

Figure 2. Functional Analysis of Wild-Type and Mutant ErbB4 upon Neuregulin-1 Stimulation

COS-7 cells transfected with an empty-vector control or plasmids encoding either wild-type (wt) or mutant HA-tagged ErbB4 (p. Arg114Gln, p. Ala158Glu, p. His374Gln, p. Arg927Gln, or p. Arg1275Trp) were stimulated with or without NRG-1, and the autophosphorylation activity of ErbB4 was analyzed by immunoblot analysis with antibodies against phospho-ErbB4 (Tyr1284) (Cell Signaling) and HA tag (Abcam), respectively. For loading controls, immunoblotting was performed with an anti-actin antibody (Santa Cruz Biotechnology). Three amino acid substitutions, including p.Arg114Gln, p.Ala158Glu, and p.His374Gln (rs760369), identified through mutational analysis of FALS and SALS individuals, were included in autophosphorylation assay. The substitutions p.Arg114Gln and p.Ala158Glu were not considered to be relevant to ALS because neither recurrence nor cosegregation was confirmed.

mutation occurred independently. We identified a de novo mutation of c.3823C>T (dbSNP SubSNP ID ss831884246), substituting tryptophan for arginine at codon 1275 (p. Arg1275Trp), in a Japanese SALS individual (Figure 1B) in whom a biological parent-descendant relationship was confirmed (Table S4) by the PLINK[10] algorithm. These mutations were neither present in the 477 Japanese controls nor registered in the in-house database containing 41 whole genomes and 1408 exomes, the 1000 Genomes database, or the NHLBI-ESP database, containing 6503 exomes. Furthermore, c.2780G>A was not present in 190 Canadian controls. The identification of c.2780G>A in two independent families of different ethnic backgrounds strongly supported c.2780G>A as the causative mutation for ALS. Given that de novo mutation rates have been estimated to be 1.20×10^{-8} per nucleotide per generation[11] and less than one nonsynonymous single-nucleotide variant (SNV)/generation,[12] the observation of the de novo mutation further supports the idea that c.3823C>T is likely to be the causative mutation for ALS in this individual. The mutation's substituted arginine residues, Arg927 and Arg1275, are highly conserved among species (Figure 1C), and the substitutions are predicted to be probably damaging by PolyPhen-2 analysis. The amino acid residue Arg927 resides in a tyrosine kinase domain, which is essential for the receptor tyrosine kinase activity, and Arg1275 is located in a C-terminal domain in the vicinity of multiple phosphorylation sites, which mediate downstream signaling pathways (Figure 1D). The clinical presentations of these ALS individuals with the *ERBB4* mutations are summarized in Table 1. The common clinical characteristics of the individuals included both upper and lower motor-neuron involvement diagnosed as definite or probable ALS according to El Escorial and Airlie House revised criteria, relatively slow disease progression, and no obvious cognitive impairment. The individuals with the c.2780G>A mutation were characterized by relatively late onset (the ages at onset ranged from 60–70 years) and a slightly reduced penetrance. In contrast, the individual with the c.3823C>T mutation was characterized by early onset (45 years of age).

ErbB4 is a member of the epidermal growth factor (EGF) subfamily of receptor tyrosine kinases (RTKs). It forms a homodimer or a heterodimer with ErbB2 or ErbB3 and is activated upon binding of neuregulins (NRGs) to the extracellular ligand-binding domain of ErbB4.[13] Activation of ErbB4 is mediated by increased tyrosine kinase activity upon NRG binding, resulting in autophosphorylation of the C-terminal tail.[14] To determine how the two mutations identified in the ALS individuals affect ErbB4 functions, we investigated the autophosphorylation of ErbB4 in cells expressing either wild-type or mutant (c.2780G>A or c.3823C>T) *ERBB4* in the presence of NRG-1. The *ERBB4* mutations were introduced into the pBABE-puro*ERBB4JM-aCYT-2HA* plasmid encoding HA-tagged ErbB4 JM-a CYT-2[15] by site-directed mutagenesis according to the protocol described in the Phusion Site-Directed Mutagenesis Kit (Thermo Fisher Scientific). After mutagenesis, all the constructs were verified by sequencing. The plasmids were transiently transfected into COS-7 cells via FuGENE 6 transfection reagent (Roche) in accordance with the manufacturer's instructions. Transfected cells were starved of serum overnight and stimulated with 0 or 50 ng/ml NRG-1 (R&D Systems) for 10 min at 37°C. After stimulation, the cells were lysed, and samples equivalent to 50 μg of total protein were separated through 8% SDS-PAGE gels. For detectiin of ErbB4 phosphorylation and total ErbB4 protein levels, immunoblotting was performed antibodies against phospho-ErbB4 (Tyr1284) (Cell Signaling) and HA-tag (Abcam), respectively. The two amino acid substitutions, p. Arg927Gln and p. Arg1275Trp, showed a clearly reduced autophosphorylation of ErbB4 (Figure 2). On the basis of these genetic and functional data, we concluded that the two mutations are causative mutations for ALS (ALS19).

This study revealed that a reduced autophosphorylation of ErbB4 upon NRG-1 stimulation is involved in the pathogenesis of ALS. *Erbb4* is specifically expressed in the soma of large motor neurons of the rat spinal cord.[16] The lack of *Erbb4* is embryonically lethal in mice, which displayed the derangement of motor-neuron axon guidance and pathfinding during embryogenesis.[17] Heterozygous-null mice showed a reduced body weight and delayed motor development, and brain-specific conditional knock-out mice

demonstrated reduced spontaneous motor activity and grip strength of the hindlimbs.[18] Mice lacking cysteine-rich domain (CRD) isoforms of *Nrg-1* (*CRD-NRG-1*$^{-/-}$) die perinatally as a result of respiratory failure, lack detectable limb movement, and exhibit a loss of ~60% of spinal motor neurons.[19] Similarly, motor and sensory neuron-specific conditional *Nrg-1* knockout mice die at birth and showed marked retraction of motor-neuron axons.[20] Furthermore, a decrease in the amount of CRD-NRG-1 has been detected in the spinal motor neurons in FALS and SALS individuals and *Sod1* mutant mice at disease onset,[21] raising the possibility that disruption of the NRG-ErbB pathway is commonly involved in the motor-neuron degeneration underlying ALS. This study provides insight into ALS pathogenesis and is expected to pave the way for the development of innovative therapeutic strategies such as using NRGs or their agonists to upregulate ErbB4 functions.

Supplemental Data

Supplemental Data include two figures and four tables and can be found with this article online at http://www.cell.com/AJHG/.

Consortia

Consortium members of JaCALS include Ryoichi Nakamura, Hazuki Watanabe, Yuishin Izumi, Ryuji Kaji, Mitsuya Morita, Kotaro Ogaki, Akira Taniguchi, Ikuko Aiba, Koichi Mizoguchi, Koichi Okamoto, Kazuko Hasegawa, Masashi Aoki, Akihiro Kawata, Imaharu Nakano, Koji Abe, Masaya Oda, Masaaki Konagaya, Takashi Imai, Masanori Nakagawa, Takuji Fujita, Hidenao Sasaki, and Masatoyo Nishizawa.

Acknowledgments

We thank all the family members for participating in this study. This study was supported in part by KAKENHI (Grants-in-Aid for Scientific Research on Innovative Areas [22129001 and 22129002]) to S.T.; the Global COE Program from the Ministry of Education, Culture, Sports, Science, and Technology of Japan, and a grant-in-aid (H23-Jitsuyoka [Nanbyo]-Ippan-004) from the Ministry of Health, Labour, and Welfare, Japan to S.T. We acknowledge support to R.H.B. from ALS Therapy Alliance, Project ALS, P2ALS, the Angel Fund, the Pierre L. de Bourgknecht ALS Research Foundation, the Al-Athel ALS Research Foundation, the ALS Family Charitable Foundation, and grant 1R01NS050557 from the National Institute of Neurological Disorders and Stroke of the National Institutes of Health and support to G.A.N. from the MND Research Institute of Australia. P.G.-P. was supported by the Alfonso Martin Escudero Foundation (Madrid).

Received: May 12, 2013
Revised: August 26, 2013
Accepted: September 13, 2013
Published: October 10, 2013

Web Resources

The URLs for data presented herein are as follows:

1000 Genomes Project Database, http://www.1000genomes.org/
dbSNP135, http://www.ncbi.nlm.nih.gov/projects/SNP/
ExonPrimer, http://ihg.gsf.de/ihg/ExonPrimer.html
NCBI37/hg19 assembly, http://genome.ucsc.edu/
NHLBI GO Exome Sequencing Project (NHLBI-ESP), https://esp.gs.washington.edu/drupal
Online Mendelian Inheritance in Man (OMIM), http://www.omim.org/
Personal genome databases, http://www.sequenceontology.org/resources/10Gen.html
PLINK algorithm, http://pngu.mgh.harvard.edu/purcell/plink/
PolyPhen-2, http://genetics.bwh.harvard.edu/pph2/
Refseq, http://www.ncbi.nlm.nih.gov/projects/RefSeq/
UCSC Human Genome Browser, http://genome.ucsc.edu/

Accession Numbers

The dbSNP accession numbers for the c. 2780G>A and c. 3823C>T mutations reported for *ERBB4* in this paper are ss831884245 and ss831884246, respectively.

References

1. Al-Chalabi, A., Jones, A., Troakes, C., King, A., Al-Sarraj, S., and van den Berg, L.H. (2012). The genetics and neuropathology of amyotrophic lateral sclerosis. Acta Neuropathol. *124*, 339–352.
2. Andersen, P.M., and Al-Chalabi, A. (2011). Clinical genetics of amyotrophic lateral sclerosis: what do we really know? Nat Rev Neurol *7*, 603–615.
3. Takahashi, Y., Seki, N., Ishiura, H., Mitsui, J., Matsukawa, T., Kishino, A., Onodera, O., Aoki, M., Shimozawa, N., Murayama, S., et al. (2008). Development of a high-throughput microarray-based resequencing system for neurological disorders and its application to molecular genetics of amyotrophic lateral sclerosis. Arch. Neurol. *65*, 1326–1332.
4. Ishiura, H., Takahashi, Y., Mitsui, J., Yoshida, S., Kihira, T., Kokubo, Y., Kuzuhara, S., Ranum, L.P., Tamaoki, T., Ichikawa, Y., et al. (2012). C9ORF72 repeat expansion in amyotrophic lateral sclerosis in the Kii peninsula of Japan. Arch. Neurol. *69*, 1154–1158.
5. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.
6. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.
7. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26.
8. Fukuda, Y., Nakahara, Y., Date, H., Takahashi, Y., Goto, J., Miyashita, A., Kuwano, R., Adachi, H., Nakamura, E., and Tsuji, S. (2009). SNP HiTLink: a high-throughput linkage analysis system employing dense SNP data. BMC Bioinformatics *10*, 121.
9. Gudbjartsson, D.F., Thorvaldsson, T., Kong, A., Gunnarsson, G., and Ingolfsdottir, A. (2005). Allegro version 2. Nat. Genet. *37*, 1015–1016.
10. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.