Linkage analysis

We extracted genomic DNAs from blood samples in the two affected individuals (II-1 and II-2) (figure 1C). Then, we conducted multipoint parametric linkage analysis using a pipeline software, SNP high-throughput linkage analysis system (SNP HiTLink).[6] We can directly import SNP chip data for the Mapping 100k/500k array set and Genome-Wide Human SNP array V.6.0 (Affymetrix, Santa Clara, California, USA) and pass to a multipoint parametric linkage analysis program, Allegro, with this system.[7] We calculated Parametric logarithm of the odds (LOD) scores using Allegro V.2 with the parameter setting of an AR model with 100% penetrance.

Exome sequencing

We extracted Genomic DNA from leucocytes from one patient (II-1) and then sheared. We purified the sample using Agencourt AMPure XP beads. We prepared an adaptor-ligated library and clustered on the cBOT system (Illumina, San Diego, California, USA). We performed exon capture with a SureSelect Human All Exon 50 Mb Kit (Agilent). We carried out paired-end sequencing on an Illumina Hiseq 2000, which generated 101 bp reads. We aligned the sequences with the human genome reference sequence (hg19 build, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/) using a Burrows-Wheeler Aligner for sequence alignment, variant calling and annotation. We carried out substitution calling with a Genome Analysis Toolkit (GATK, http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit). SNP calls were made with a GATK Unified Genotyper, and indel calls were made with a GATK IndelGenotyper V2. We performed SNP calling with reference to dbSNP135 (ftp://ftp.ncbi.nlm.nih.gov/snp/orgasms/human_9606/ASN1_flat/) and 1000 Genomes (ftp://1000genomes.ebi.ac.uk/vol1/ftp/release/20110521). All variants were annotated with reference to consensus coding sequences (CCDS) (NCBI release 20111122) and RefSeq (UCSC dumped 20111122).

Validation of mutations by Sanger sequencing

We amplified exon 12 of *LYST* and flanking intronic sequences using the genomic DNA of the two patients using a MJ Research PTC-100 Thermal Cycler (Marshall Scientific, Brentwood, New Hampshire, USA). The primer sequences were as follows: LYST ex12-F: agg aat gct gat atg tgt ggg; LYST ex12-R: cac att ttt acg gct caa gga. We performed Sanger sequencing according to an established standard protocol on an Applied Biosystems (ABI) 3730 capillary sequencer (Applied Biosystems, Carlsbad, California, USA). We also analysed genomic DNA samples from 200 Japanese subjects without apparent neurological disorders as controls.

The institutional review boards of the Jichi Medical University, the Shinshu University, the University of Tokyo and the University of Yamanashi approved this study.

RESULTS

Clinical features

The proband (II-2) was a 53-year-old man who was admitted to our hospital because of gait disturbance and slowly progressive weakness of the lower extremities. He was experiencing difficulties in walking up stairs and standing upright at age 48 years. Later he walked with a cane due to unsteadiness. His Mini-Mental State Examination (MMSE) score was 25/30. On neurological examination, he showed bilateral leg spasticity, and bilateral iliopsoas muscle weakness and atrophy. Tendon reflexes were exaggerated in all extremities except for diminished ankle jerks. He showed extensor bilateral plantar responses. We could observe subtle ataxia in the upper extremities. Brain MRI showed mild cerebellar atrophy, and spinal MRI revealed mild thoracic cord atrophy (figure 2A). Mild decreases of motor and sensory nerve conduction velocities in the lower extremities were revealed by a nerve conduction study, but SNAPs were not evoked in the sural nerves. Motor-evoked potential examination disclosed marked delay of the central motor conduction times in the corticospinal tracts. His eyes exhibited no areas of hyperpigmentation or hypopigmentation on ophthalmological examination. Pigmentary abnormalities of the skin were not observed. Peroxidase staining of peripheral blood revealed giant granules in granulocytes (figure 2B) and mild reduction of natural killer cell activity.

The second case, an older brother of the proband (II-1), showed almost the same clinical presentations except for the leg tonus. He suffered from gait difficulty at age 58 years and became unable to walk for a long time at age 62 years. A year later he walked with a cane and was admitted to our hospital for slowly progressive gait disturbance. His MMSE score was 16/30. Neurological examination at age 63 years showed bilateral iliopsoas muscle weakness and atrophy, but leg spasticity was not noted. Tendon reflexes were diminished in all extremities with extensor plantar responses bilaterally. Mild ataxia in the upper extremities was observed. Brain MRI showed mild cerebellar atrophy, whereas spinal MRI revealed no spinal cord atrophy. A nerve conduction study showed normal motor and sensory nerve conduction velocities in the upper and lower extremities, whereas the amplitudes of compound muscle action potentials and sensory nerve action potentials were decreased in all extremities, and F-wave conduction velocities were decreased in the lower extremities. Motor-evoked potential examination revealed prolongation of the central motor conduction times in the corticospinal tracts. Needle EMG showed chronic neurogenic patterns in his legs. A sural nerve biopsy disclosed a decreased number of nerve fibres of large diameter and residual axonal swelling, but no formation of onion bulb-like structure. Peripheral blood examinations showed peroxidase-positive giant granules in granulocytes and reduced natural killer cell activity.

Identification of candidate chromosome areas

We found linkages to a several parts of chromosomes 1 (rs3914503-rs2186205), 2 (rs12614444-rs4035021), 11 (rs11235880-rs120435) and 17 (rs7216464-rs4128515, rs7217461-rs11079098), with maximum cumulative LOD scores of 1.8 (figure 1A). These four chromosome parts did not contain previously reported HSP loci.

Exome sequencing allowed identification of the candidate gene substitutions

Exome sequencing covered 98.65% of the target region with an average sequence depth of 106.24X. All coding sequences in the four candidate chromosomal areas were covered at least eight times. The presence of consanguinity and two asymptomatic parents supported that the patients have homozygous disease-causing mutations. We selected all coding variants in the homozygous regions and then filtered them by discarding variants documented in the dbSNP135 and the 1000 Genomes Project, and heterozygous and synonymous substitutions. We identified three novel homozygous, non-synonymous single nucleotide variants: c.4189T>G (p.F1397V) in *LYST* in the chromosome 1 candidate area (figure 1B), c.5135G>A (p.G1712E) in *FAT3* in the chromosome 11 candidate area and

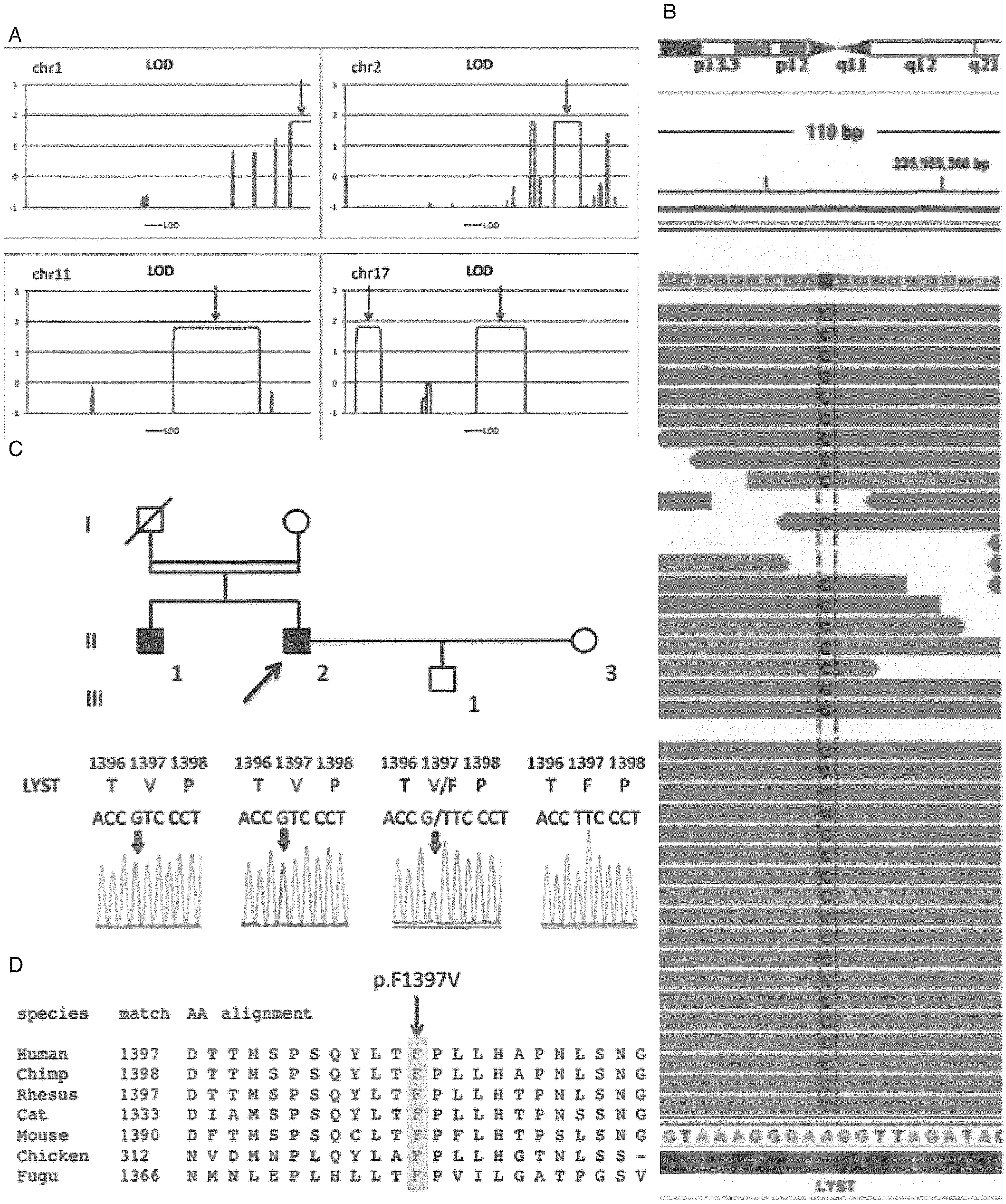

Figure 1 Linkage analysis and mutation of the *lysosomal trafficking regulator* (*LYST*) gene in the family. (A) Linkage analysis involving single nucleotide proteins revealed the highest logarithm of the odds (LOD) scores (about 1.8) in parts of chromosomes 1, 2, 11 and 17 (arrows). These four areas were thought to be candidate areas in which the causative gene was located. (B) Exome sequencing using one patient's DNA identified the mutation of the *LYST* gene on chromosome 1. This mutation was demonstrated with Integrative genomics viewer (IGV).[26] (C) Family tree and Sanger sequencing validation. Sanger sequencing confirmed the homozygous missense mutation (c.4189T>G, p.F1397V) of the *LYST* gene identified in the proband and the affected brother. This mutation was co-segregated with the disease in this family. (D) This amino acid substitution (p.F1397V) is located at a highly conserved residue within the concanavalin A (ConA)-like lectin domain (amino acid numbers 1390–1691) of the LYST protein.

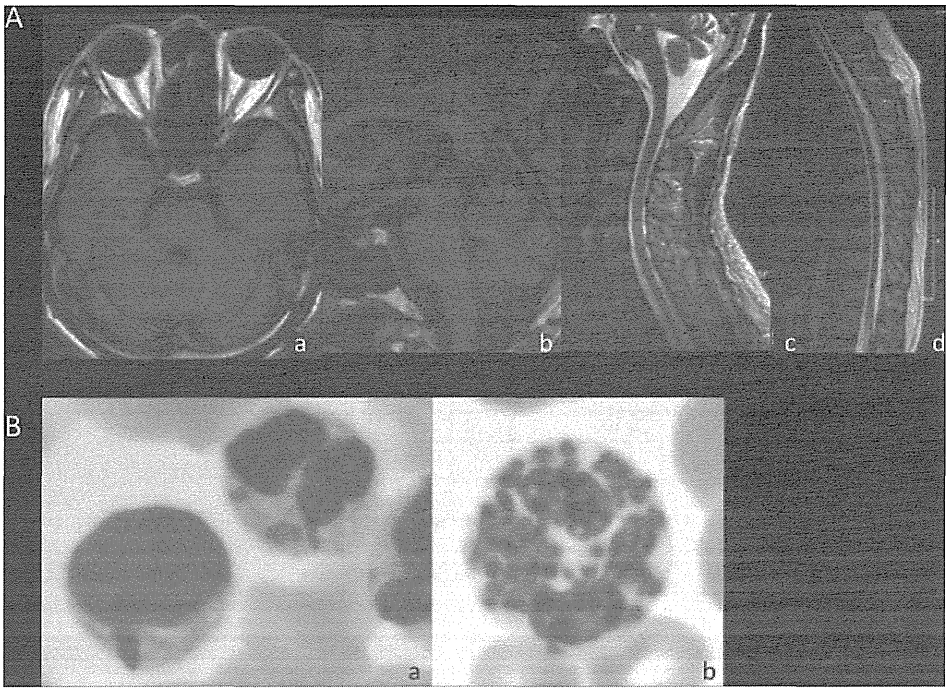

Figure 2 MRI and peripheral blood findings in the proband (II-2). (A) Brain and spinal MRI of the proband. Brain T1-weighted MRI (a, b) showed mild cerebellar atrophy, and spinal T2-weighted MRI (c, d) disclosed mild thoracic cord atrophy. (B) Peripheral blood leucocytes findings in the proband (May–Giemsa stain (a) and peroxidase stain (b)). We could observe large granules in the proband's leucocytes and large peroxidase-positive ones in granulocytes.

c.1045T>C (p.F349L) in *ANGPTL5* in the chromosome 11 candidate area with reference to dbSNP135. These three variations were validated by Sanger sequencing. No such variations were detected in the chromosome 2 and 17 areas. Subsequently, the two candidate gene variants in *FAT3* and *ANGPTL5* could be excluded as polymorphisms because these variants of the two genes were not co-segregated in the normal family members. The another candidate gene variant of *LYST* was predicted to be a functionally deleterious mutation with the prediction programs (PROVEAN, Polyphen-2 and Mutation Taster) and confirmed to be a homozygous missense mutation (c.4189T>G, p. F1397V) on Sanger sequencing in the two patients, II-1 and II-2 (figure 1C). This missense mutation was co-segregated within the family members (figure 1C) and not found in 200 Japanese control genomic DNA samples. This mutation is located at a highly conserved residue (figure 1D) within the concanavalin A (ConA)-like lectin domain (amino acids numbers 1390–1691).[8]

DISCUSSION

The present two patients exhibited spastic paraplegia, peripheral neuropathy and mild cerebellar ataxia with AR transmission. Autosomal recessive hereditary spastic paraplegia (AR-HSP) with cerebellar ataxia and neuropathy is considered to be SPG7 with the *Paraplegin* gene alteration linked to chromosome 16q24.3,[9] SPG21 with the *Maspardin* gene mutation linked to chromosome 15q22.31,[10] SPG27 linked to chromosome 10q22.1–q24.1[11] and SPG30 with the *KIF1A* mutation linked to chromosome 2q37.3.[12 13] However, linkage analysis did not show all reported HSP gene locus linkages, and whole exome sequence analysis did not disclose the *Paraplegin, Maspardin* and *KIF1A* gene mutations. According to these results, we could conclude that the causative gene in this family was not one of the previously reported HSP ones.

Through linkage analysis of the two patients' DNA and whole exome sequencing using one individual's DNA, we could identify a novel homozygous missense mutation in the *LYST* gene. This homozygous mutation was shared by the two patients. We considered the *LYST* gene mutation was causative of the neurological deficits in these two patients because it was co-segregated within their family members, located at a highly conserved amino acid, and not found in the normal controls. Moreover, large granules in leucocytes and reduced natural killer cell activity could support the diagnosis of CHS.

CHS is a rare, AR early-onset disorder characterised by severe immune deficiency, frequent bacterial infections, skin pigmentation or albinism, a bleeding tendency and progressive neurological dysfunction in most cases.[14] It is often complicated by a lymphoproliferative condition called the 'accelerated phase'. A classic diagnostic feature of CHS is enlarged granules in leucocytes, melanocytes, platelets and so forth. Most cases present in early childhood with haematological dysfunction, whereas a small number of cases with the adult form of CHS predominantly exhibit slowly progressive neurological dysfunction without apparent immunodeficiency or a bleeding tendency. Neurological involvement in CHS can include parkinsonism,[15] dementia,[16] cerebellar ataxia, peripheral neuropathy and spastic paraplegia.[17] Although the neurological phenotypes of our cases resembled those previously reported,[17] the main symptom in those patients was cerebellar ataxia that was more severe than that in our cases.

The gene responsible for CHS was identified in 1996, and was called *LYST*.[5 18] The *LYST* gene is a large gene that has 51 coding exons and an open reading frame of 11 403 kb.[6] The LYST protein, which is a large, putative cytosolic protein of 425 kDa (3801 amino acids), is ubiquitously expressed and involved in control of the exocytosis of secretory lysosomes.[5 19] The LYST protein has a BEACH (named after BEige And

Chédiak–Higashi) domain (amino acid numbers 3132–3422),[5] Trp-Asp (WD) 40 repeats (amino acid numbers 3477-3778) and a ConA-like lectin domain (amino acid numbers 1390–1691).[8] The LYST protein has been proposed to act as a scaffold protein in the mediation of fusion or a fission event of vesicles.[20] The mutation in this family (p.F1397V) is located within the ConA-like lectin domain. This domain could be involved in oligosaccharide binding associated with protein trafficking and sorting along the secretory pathway.[8]

Recently, *Drosophila* with a gene mutation of an *LYST* homologue was revealed to exhibit impaired autophagy.[21] The loss of function of some HSP-related proteins, TECPR2[22] and spastizin,[23] caused autophagic dysfunction and induced spastic paraplegia. Therefore, autophagic impairment might have resulted in spastic paraplegia in the CHS patients.

Karim *et al*[24] found apparent genotype–phenotype correlations in CHS, that is, that severe childhood CHS involved a functionally null mutation, whereas missense mutations were seen only in the two later-onset forms. They reported four missense mutations, two of which are located in the ConA-like lectin domain. Our cases correspond to late-onset, slowly progressive neurological CHS with a missense mutation of the *LYST* gene. According to the information on the Japanese cases in the literature,[24] the Japanese adult CHS cases with *LYST* missense mutations (R1563H or V1999D) showed spastic paraplegia, gaze nystagmus and diminished ATRs. Thus, their phenotypes were similar to those of our cases. Moreover, as far as we know, this family had one of the oldest adult CHS cases (onset of 58 years) with a *LYST* gene mutation in the literature. To date, a 57-year-old man has been reported who suffered from sensorimotor polyneuropathy and muscle wasting with a heterozygous *LYST* gene mutation (p.Y2026X).[25]

In this family, the proband showed spastic paraplegia dominantly as well as neuropathy and mild cerebellar ataxia, whereas the brother mainly showed peripheral neuropathy with a positive Babinski sign, cerebellar ataxia and dementia. These two patients did not exhibit parkinsonism. The phenotypic variety in this family might be explained by environmental factors or other modifier gene mutations.

In summary, we could diagnose these patients as having adult CHS presenting spastic paraplegia with neuropathy and cerebellar ataxia. As far as we know, this family includes one of the oldest adult CHS cases in the literature. The clinical spectrum of CHS is broader than previously recognised, and this family shows phenotypic variability. Adult CHS must be considered in the differential diagnosis of AR-HSPs. The linkage analysis and exome sequencing were useful for identifying the causative mutation in this family.

Acknowledgements We thank Professor Miyajima (Department of Neurology, Hamamatsu Medical University, Hamamatsu, Japan) for the Japanese adult CHS patient's information in a reference.[24]

Contributors HS was responsible for conception and design of the work, acquisition, analysis or interpretation of data, drafting the work and revising the work critically for important intellectual content. JH, TN and MN were responsible for acquisition and analysis of data. IN was responsible for analysis or interpretation of data. MY, KN, KY, S-il, HI, YF and YT were responsible for acquisition, analysis or interpretation of data. JG, ST and YT were responsible for conception and design of the work, interpretation of data and revising the work critically for important intellectual content.

Funding This work was supported by a Grant-in-Aid for Scientific Research (C) (23591253 to HS) and MEXT KAKENHI Grant Number 221S0002 from The Ministry of Education, Culture, Sports, Science and Technology in Japan. This work was also supported by a grant from the Research Committee for Ataxic Diseases (YT and HS) of the Ministry of Health, Labor and Welfare, Japan.

Competing interests None.

Patient consent Obtained.

Ethics approval The Institutional Review Boards of the Jichi Medical University, Shinshu University, University of Tokyo, and University of Yamanashi.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement Exome sequencing data are available (DDBJ Sequence Read Archive (DRA) accession number DRA000961).

REFERENCES

1 Harding AE. Classification of the hereditary ataxias and paraplegias. *Lancet* 1983;1:1151–5.
2 Hazan J, Fonknechten N, Mavel D, *et al*. Spastin, a new AAA protein, is altered in the most frequent form of autosomal dominant spastic paraplegia. *Nat Genet* 1999;23:296–303.
3 Stevanin G, Santorelli FM, Azzedine H, *et al*. Mutations in SPG11, encoding spatacsin, are a major cause of spastic paraplegia with thin corpus callosum. *Nat Genet* 2007;39:366–72.
4 Salinas S, Proukakis C, Crosby A, *et al*. Hereditary spastic paraplegia: clinical features and pathogenetic mechanisms. *Lancet Neurol* 2008;7:1127–38.
5 Nagle DL, Karim MA, Woolf EA, *et al*. Identification and mutation analysis of the complete gene for Chediak-Higashi syndrome. *Nat Genet* 1996;14:307–11.
6 Fukuda Y, Nakahara Y, Date H, *et al*. SNP HiTLink: a high-throughput linkage analysis system employing dense SNP data. *BMC Bioinformatics* 2009;10:121.
7 Gudbjartsson DF, Thorvaldsson T, Kong A, *et al*. Allegro version 2. *Nat Genet* 2005;37:1015–16.
8 Burgess A, Mornon JP, de Saint-Basile G, *et al*. A concanavalin A-like lectin domain in the CHS1/LYST protein, shared by members of the BEACH family. *Bioinformatics* 2009;25:1219–22.
9 Casari G, De Fusco M, Ciarmatori S, *et al*. Spastic paraplegia and OXPHOS impairment caused by mutations in paraplegin, a nuclear-encoded mitochondrial metalloprotease. *Cell* 1998;93:973–83.
10 Simpson MA, Cross H, Proukakis C, *et al*. Maspardin is mutated in mast syndrome, a complicated form of hereditary spastic paraplegia associated with dementia. *Am J Hum Genet* 2003;73:1147–56.
11 Meijer IA, Cossette P, Roussel J, *et al*. A novel locus for pure recessive hereditary spastic paraplegia maps to 10q22.1-10q24.1. *Ann Neurol* 2004;56:579–82.
12 Klebe S, Azzedine H, Durr A, *et al*. Autosomal recessive spastic paraplegia (SPG30) with mild ataxia and sensory neuropathy maps to chromosome 2q37.3. *Brain* 2006;129:1456–62.
13 Erlich Y, Edvardson S, Hodges E, *et al*. Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1A in hereditary spastic paraparesis. *Genome Res* 2011;21:658–64.
14 Kaplan J, De Domenico I, Ward DM. Chediak-Higashi syndrome. *Curr Opin Hematol* 2008;15:22–9.
15 Pettit RE, Berdal KG. Chediak-Higashi syndrome. Neurologic appearance. *Arch Neurol* 1984;41:1001–2.
16 Uyama E, Hirano T, Ito K, *et al*. Adult Chediak-Higashi syndrome presenting as parkinsonism and dementia. *Acta Neurol Scand* 1994;89:175–83.
17 Sheramata W, Kott HS, Cyr DP. The Chediak-Higashi-Steinbrinck syndrome. Presentation of three cases with features resembling spinocerebellar degeneration. *Arch Neurol* 1971;25:289–94.
18 Barbosa MD, Nguyen QA, Tchernev VT, *et al*. Identification of the homologous beige and Chediak-Higashi syndrome genes. *Nature* 1996;382:262–5.
19 Huynh C, Roth D, Ward DM, *et al*. Defective lysosomal exocytosis and plasma membrane repair in Chediak-Higashi/beige cells. *Proc Natl Acad Sci USA* 2004;101:16795–800.
20 Tchernev VT, Mansfield TA, Giot L, *et al*. The Chediak-Higashi protein interacts with SNARE complex and signal transduction proteins. *Mol Med* 2002;8:56–64.
21 Rahman M, Haberman A, Tracy C, *et al*. Drosophila mauve mutants reveal a role of LYST homologs late in the maturation of phagosomes and autophagosomes. *Traffic* 2012;13:1680–92.
22 Oz-Levi D, Ben-Zeev B, Ruzzo EK, *et al*. Mutation in TECPR2 reveals a role for autophagy in hereditary spastic paraparesis. *Am J Hum Genet* 2012;91:1065–72.
23 Vantaggiato C, Crimella C, Airoldi G, *et al*. Defective autophagy in spastizin mutated patients with hereditary spastic paraparesis type 15. *Brain* 2013;136:3119–39.
24 Karim MA, Suzuki K, Fukai K, *et al*. Apparent genotype-phenotype correlation in childhood, adolescent, and adult Chediak-Higashi syndrome. *Am J Med Genet* 2002;108:16–22.
25 Mottonen M, Lanning M, Baumann P, *et al*. Chediak-Higashi syndrome: four cases from Northern Finland. *Acta Paediatr* 2003;92:1047–51.
26 Robinson JT, Thorvaldsdottir H, Winckler W, *et al*. Integrative genomics viewer. *Nat Biotechnol* 2011;29:24–6.

Autosomal-recessive complicated spastic paraplegia with a novel *lysosomal trafficking regulator* gene mutation

Haruo Shimazaki, Junko Honda, Tametou Naoi, et al.

J Neurol Neurosurg Psychiatry published online February 12, 2014
doi: 10.1136/jnnp-2013-306981

Notes

Journal of Human Genetics (2014) 59, 163–172
© 2014 The Japan Society of Human Genetics All rights reserved 1434-5161/14
www.nature.com/jhg

ORIGINAL ARTICLE

Molecular epidemiology and clinical spectrum of hereditary spastic paraplegia in the Japanese population based on comprehensive mutational analyses

Hiroyuki Ishiura[1], Yuji Takahashi[1], Toshihiro Hayashi[1], Kayoko Saito[2], Hirokazu Furuya[3], Mitsunori Watanabe[4], Miho Murata[5], Mikiya Suzuki[6], Akira Sugiura[7], Setsu Sawai[8,9], Kazumoto Shibuya[10], Naohisa Ueda[11,12], Yaeko Ichikawa[1], Ichiro Kanazawa[13], Jun Goto[1] and Shoji Tsuji[1]

Hereditary spastic paraplegia (HSP) is one of the most genetically heterogeneous neurodegenerative disorders characterized by progressive spasticity and pyramidal weakness of lower limbs. Because >30 causative genes have been identified, screening of multiple genes is required for establishing molecular diagnosis of individual patients with HSP. To elucidate molecular epidemiology of HSP in the Japanese population, we have conducted mutational analyses of 16 causative genes of HSP (*L1CAM, PLP1, ATL1, SPAST, CYP7B1, NIPA1, SPG7, KIAA0196, KIF5A, HSPD1, BSCL2, SPG11, SPG20, SPG21, REEP1* and *ZFYVE27*) using resequencing microarrays, array-based comparative genomic hybridization and Sanger sequencing. The mutational analysis of 129 Japanese patients revealed 49 mutations in 46 patients, 32 of which were novel. Molecular diagnosis was accomplished for 67.3% (33/49) of autosomal dominant HSP patients. Even among sporadic HSP patients, mutations were identified in 11.1% (7/63) of them. The present study elucidated the molecular epidemiology of HSP in the Japanese population and further broadened the mutational and clinical spectra of HSP.

Journal of Human Genetics (2014) **59,** 163–172; doi:10.1038/jhg.2013.139; published online 23 January 2014

Keywords: array-based comparative genomic hybridization; hereditary spastic paraplegia; resequencing microarray

INTRODUCTION

Hereditary spastic paraplegia (HSP) is a neurodegenerative disorder characterized by progressive lower limb spasticity and pyramidal weakness, which is one of the most genetically and clinically heterogeneous disorders.[1,2] HSP is clinically divided into two forms, pure and complicated forms, depending on whether the neurological symptoms are basically confined to spasticity and pyramidal weakness of the lower limbs or accompanied by additional neurological symptoms such as cognitive dysfunction, cerebellar signs, optic atrophy, retinitis pigmentosa, amyotrophy and peripheral neuropathy. HSP is characterized by enormous genetic heterogeneity; to date, more than 50 genetic loci (SPG1-57) and 37 causative genes have been identified: *L1CAM* (SPG1), *PLP1* (SPG2), *ATL1* (SPG3A), *SPAST* (SPG4), *CYP7B1* (SPG5A), *NIPA1* (SPG6), *SPG7* (SPG7), *KIAA0196* (SPG8), *KIF5A* (SPG10), *SPG11* (SPG11), *RTN2* (SPG12), *HSPD1* (SPG13), *SPG15/ZFYVE26* (SPG15), *BSCL2* (SPG17), *ERLIN2* (SPG18), *SPG20* (SPG20), *SPG21* (SPG21), *DDHD1* (SPG28), *KIF1A* (SPG30), *REEP1* (SPG31), *ZFYVE27* (SPG33), *FA2H* (SPG35), *PNPLA6* (SPG39), *SLC33A1* (SPG42), *GJC2* (SPG44), *GBA2* (SPG46), *AP4B1* (SPG47), *KIAA0415* (SPG48), *TECPR2* (SPG49), *AP4M1* (SPG50), *AP4E1* (SPG51), *AP4S1* (SPG52), *VPS37A* (SPG53), *DDHD2* (SPG54), *C12ORF65* (SPG55), *CYP2U1* (SPG56), and *TFG* (SPG57).

Because of the limited availability of information on genotype–phenotype correlations and locus heterogeneity, it is often difficult to prioritize genes for the mutational analysis of HSP. Therefore, it is essential to incorporate knowledge of the molecular epidemiology of HSP and relative frequencies of the types of mutations (substitution, insertion/deletion or rearrangement) in each gene into the algorithm of molecular diagnosis of HSP. We also need to be aware that different

[1]Department of Neurology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan; [2]Institute of Medical Genetics, Tokyo Women's Medical University, Tokyo, Japan; [3]Department of Neurology, Neuro-Muscular Center, National Omuta Hospital, Fukuoka, Japan; [4]Department of Neurology, Institute of Brain Science, Hirosaki University Graduate School of Medicine, Aomori, Japan; [5]Department of Neurology, National Center of Neurology and Psychiatry, Tokyo, Japan; [6]Department of Neurology, Higashisaitama Hospital, National Hospital Organization, Saitama, Japan; [7]Department of Neurology, Shizuoka Institute of Epilepsy and Neurological Disorders, Shizuoka, Japan; [8]Department of Molecular Diagnosis, Graduate School of Medicine, Chiba University, Chiba, Japan; [9]Division of Laboratory Medicine and Clinical Genetics, Chiba University Hospital, Chiba, Japan; [10]Department of Neurology, Graduate School of Medicine, Chiba University, Chiba, Japan; [11]Department of Neurology, Chigasaki Municipal Hospital, Kanagawa, Japan; [12]Department of Neurology, Yokohama City University School of Medicine, Kanagawa, Japan and [13]Graduate School, International University of Health and Welfare, Tokyo, Japan
Correspondence: Dr S Tsuji, Department of Neurology, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8655, Japan.
E-mail: tsuji@m.u-tokyo.ac.jp

Received 13 September 2013; revised 16 November 2013; accepted 29 November 2013; published online 23 January 2014

methodologies are required to detect each type of mutations with high sensitivities. Although there have been studies on molecular epidemiology focusing on selected causative genes in large case series,[3–10] there have been only few studies based on comprehensive mutational analyses focusing on multiple genes as well as various types of mutation. Thus, the comprehensive molecular epidemiology of HSP is largely unestablished.

For these reasons, a comprehensive mutational analysis of multiple genes is necessary to efficiently provide molecular diagnosis for individual HSP patients and, furthermore, to clarify the molecular epidemiology of HSP. To accomplish high sensitivities for detection of various kinds of mutations, we have conducted comprehensive mutational analyses incorporating custom-made resequencing microarrays,[11] which enable comprehensive detection of single-nucleotide variations, custom-made comparative genomic hybridization (CGH) microarrays,[12] which enable efficient detection of large insertion/deletion variants, and Sanger sequencing, which enables detection of small insertions/deletions in addition to single base substitutions. We herein describe molecular epidemiology and the clinical spectrum of HSP based on a large-scale comprehensive mutational analysis of 129 Japanese patients with various forms of HSP.

SUBJECTS AND METHODS

Patients

One hundred and twenty-nine Japanese patients (75 male and 54 female) with a clinical diagnosis of HSP were enrolled in the study, including 45 patients who visited the University of Tokyo Hospital and 89 patients referred to our Department of Neurology, the University of Tokyo Tokyo, Japan for the molecular diagnosis of HSP from various regions in Japan between 1994 and July 2007. Genomic DNAs were prepared from peripheral blood leukocytes or an autopsied brain (1 patient) using a standard procedure. Written informed consent was obtained from all the participants or their family members. The study was approved by the institutional review board of the University of Tokyo.

Outline of mutational analysis system

The outline of the comprehensive mutational analysis is shown in Figure 1. All the samples were first subjected to resequencing microarray analysis for analyzing 13 causative genes of HSP. Among the patients in whom mutations were not detected by resequencing microarray analysis, direct nucleotide sequence analysis of *SPAST* and *REEP1* was carried out in patients with autosomal dominant HSP (AD-HSP) and sporadic pure-form HSP; in patients with a thin corpus callosum and cognitive impairment, direct nucleotide sequence analysis of *SPG11* was carried out. For those in which mutations were not detected by any of these methods, array-based CGH (aCGH) analysis was carried out.

Mutational analysis using custom-made resequencing microarrays

We developed resequencing microarrays using GeneChip CustomSeq (Affymetrix, Santa Clara, CA, USA).[11] We utilized custom-designed microarrays of the 30-kb format that contain tiled sequences for *SPAST* (NM_014946.3) and *ATL1* (NM_015915) (TKYPD01), those for *SPG7* (NM_003119) (TKYALS01), those for *L1CAM* (NM_000425) and *PLP1* (NM_000533) (TKYAD01) and those for *NIPA1* (NM_144599), *KIF5A* (NM_004984) and *SPG20* (NM_015087) (TKYPD02), as previously described.[13–15] In this study, we additionally designed two 50-kb-format microarrays. One was TKYPD03, which contained tiled sequences for *SPAST*, *ATL1* and *REEP1*, and the other was TKYALS02, which contained tiled sequences for 10 causative genes (*L1CAM* (NM_000425), *PLP1* (NM_000533), *NIPA1* (NM_144599), *SPG7* (NM_003119), *KIAA0196* (NM_014846), *KIF5A* (NM_004984), *HSPD1* (NM_002156), *BSCL2* (NM_001122955), *SPG20* (NM_015087) and *SPG21* (NM_016630)), enabling mutational analysis of 13 causative genes of HSP. Experiments were performed following the manufacturer's instructions (Supplementary Information S1). Data were analyzed using GeneChip DNA Analysis Software version 2.0 (GDAS2.0) for 30-kb-format microarrays,[16] or updated GeneChip Sequence Analysis Software version 4.0 (GSEQ4.0) adjunctively with a custom-designed program (Supplementary Information S2). All the called mutations were verified by direct nucleotide sequence analysis. The frequencies of detected nonsynonymous variations in the populations were checked using dbSNP (http://www.ncbi.nlm.nih.gov/snp/), the 1000 genomes database (http://www.1000genomes.org/) and by screening of >150 control chromosomes by direct nucleotide sequence analysis or restriction fragment length polymorphism analysis.

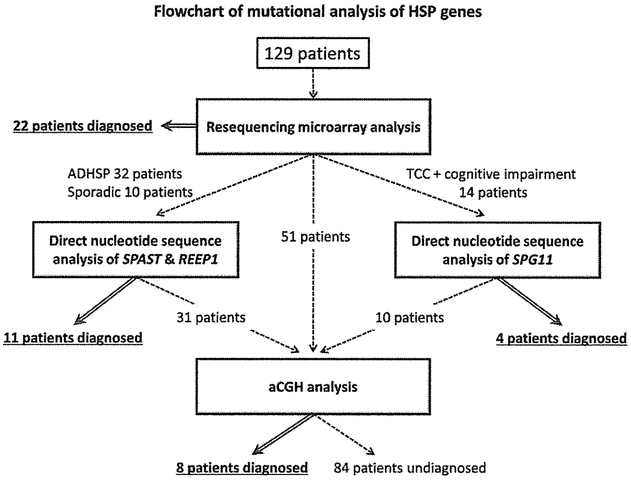

Figure 1 Flowchart of mutational analysis of HSP genes. We first performed resequencing microarray analysis, which could analyze 13 causative genes of HSP. Then, samples of mutation-negative AD-HSP and 10 sporadic pure HSP patients were subjected to direct nucleotide sequence analysis of *SPAST* and *REEP1*, because small insertion/deletion mutations are relatively frequent in these genes. Samples of mutation-negative HSP patients with a thin corpus callosum and cognitive impairment were subjected to direct nucleotide sequence analysis of *SPG11*. Finally, aCGH analysis was performed in the 92 mutation-negative patients.

Mutational analysis using aCGH

We custom-designed a CGH microarray (Agilent Technology, Santa Clara, CA, USA) in the 8 × 15 K format.[12] The genomic sequences of 16 causative genes of HSP and the flanking regions (*L1CAM*, *PLP1*, *ATL1*, *SPAST*, *CYP7B1*, *NIPA1*, *SPG7*, *KIAA0196*, *KIF5A*, *SPG11*, *HSPD1*, *BSCL2*, *SPG20*, *SPG21*, *REEP1* and *ZFYVE27*) were tiled on the array (Supplementary Information S1). The average distance between probes was 200 bp. When insertion/deletion mutations were detected by the aCGH analysis, breakpoints were determined by PCR analysis using primer pairs flanking the breakpoints and direct nucleotide sequence analysis.[17,18]

Direct nucleotide sequence analysis

Direct nucleotide sequence analysis was performed using ExoSAP-IT (USB, Cleveland, OH, USA), a BigDye Terminator v3.1 kit, and XTerminator using an ABI PRISM3100 sequencer (Applied Biosystems, Foster City, CA, USA). Primers and the amplification condition are described in Supplementary Information S1.

RESULTS

Demographic characteristics of patients

The demographic characteristics of the 129 HSP patients enrolled in this study are summarized in Table 1. The ages at onset of the patients classified on the basis of the clinical form of HSP (Supplementary Figure S1A) revealed a bimodal distribution in patients with pure-form HSP, whereas one large peak for juvenile onset and one small

Table 1 Characteristics of the HSP patients

No. of HSP patients		129
Male		75 (58.1%)
Female		54 (41.9%)
Male: female		1.4: 1
Ages at onset		0–70 y.o. (30 ± 20 y.o.)
Clinical phenotype		
Pure		82 (63.5%)
Complicated		47 (36.4%)
Family history		
Positive family history		
ADHSP		49
	Pure	44 (89.8% = 44/49)
	Complicated	5 (10.2% = 5/49)
ARHSP		11
	Pure	4 (36.4% = 4/11)
	Complicated	7 (63.6% = 7/11)
Familial (undetermined mode of inheritance)		6
	Pure	3 (50% = 3/6)
	Complicated	3 (50% = 3/6)
Sporadic		63
Sporadic with consanguinity		9
	Pure	5 (55.6% = 5/9)
	Complicated	4 (44.4% = 4/9)
Sporadic without consanguinity		54
	Pure	26 (48.1% = 26/53)
	Complicated	28 (51.9% = 28/53)

peak of adult to late onset were observed in patients with complicated-form HSP. Focusing on the mode of inheritance (Supplementary Figure S1B), the ages at onset of AD-HSP patients and sporadic HSP patients showed a similar bimodal distribution, whereas those of autosomal recessive HSP (AR-HSP) patients showed a skewed distribution.

We found 14 patients with complicated-form HSP with a thin corpus callosum and cognitive impairment. There were no patients with AD-HSP with motor neuropathy clinically diagnosed as Silver syndrome.[19]

Mutational analysis by resequencing microarray analysis

All the samples were first subjected to resequencing microarray analysis (Figure 1). The analysis detected 22 mutations, all of which were nucleotide substitutions (Table 2). Representative resequencing microarray data on a heterozygous mutation in *SPAST* (c.1493 + 2 T > C) are shown (Figures 2a–c). Using GDAS2.0 or GSEQ4.0, the overall call rate was about 90%. Except for the nucleotides for which base calling was difficult because of high GC contents, G/C stretches or locally repetitive polymorphic sequences (such as a GCG stretch in exon 1 of *NIPA1*), the overall rate of base calling using GDAS/GSEQ4.0 in combination with visual inspection was as high as 99.9%.

The custom-designed program detected one mutation (c.1741C > T, p.R581* in *SPAST*), which was not detected by GSEQ4.0, increasing the sensitivity of mutation detection (Supplementary Information S2). No additional base substitutions were detected by the subsequent direct nucleotide sequence analysis of *SPAST* and *REEP1* in AD-HSP patients.

Mutational analysis by Sanger sequencing

We applied Sanger sequencing of *SPAST* and *REEP1* in patients with AD-HSP and sporadic pure-form HSP, in which mutations were not detected by the resequencing microarray analysis, mainly to detect insertion/deletion mutations that are common in these diseases (Figure 1). We detected 10 small insertion/deletion mutations (1–41 bp) in *SPAST* and 1 insertion in *REEP1* (Table 2).

In patients with a thin corpus callosum and cognitive impairment among the cases in which mutations were not detected by the resequencing microarray analysis, we then applied Sanger sequencing of *SPG11* (Figure 1). We found homozygous or compound heterozygous mutations of *SPG11* in five patients with family histories consistent with the autosomal recessive mode of inheritance. In a sporadic HSP patient, we found compound heterozygous mutations of c.1735 + 2 delT and c.6999 + 5 delG, both of which were considered pathogenic because no other pathogenic alleles were detected, and *in silico* analysis of splicing scores of c.6999 + 5 delG showed scores that decreased from 10.1 to 7.1 (http://rulai.cshl.edu/new_alt_exon_db2/HTML/score.html) and from 1.0 to 0.62 (http://www.fruitfly.org/seq_tools/splice.html). Another patient had only one null allele (p.R2031*) in *SPG11* and no other pathogenic alleles were found.

Detection of large deletions and duplications by aCGH analysis

We applied aCGH analysis in patients in whom mutations were not detected by resequencing microarray analysis and direct nucleotide sequencing analysis (Figure 1). Representative results for a heterozygous large deletion in *KIAA0196* are shown in Figure 2d, in which the breakpoint sequence was clearly determined using an appropriate primer pair (Figure 2e).

In total, we identified 7 large deletions (4 in *SPAST*, 1 in *REEP1*, 1 in *KIAA0196*, and 1 in *SPG11*) and 1 duplication in *SPAST* in 92 patients examined (Table 2). All the breakpoints but one (in which the deletion was beyond the tiled sequences on the array) were unequivocally determined at the nucleotide level. The sizes of the deletions/duplication ranged from 4634 bp to > 170 kb.

Five (four *SPAST* deletions and one *SPG11* deletion) of the seven deletion breakpoints were inside *Alu* sequences. Among the breakpoint sequences of the remaining deletions, the deletion in *REEP1* and the deletion in *KIAA0196* showed microhomology of 3 bp. The breakpoint sequence of the *SPAST* duplication showed no homologous sequences. In total, five of the eight breakpoints (62.5%) were inside *Alu* sequences.

Molecular epidemiology of HSP in Japanese population

In summary, we found 49 mutations in 46 patients (Table 2). The relative frequencies of individual HSP types classified on the basis of the clinical presentations and the mode of inheritance are summarized in Figures 3a and b.

Focusing on all AD-HSP patients, SPG4 (55.1%) was the most frequent. SPG3A (2.0%), SPG8 (4.1%), and SPG31 (4.1%) were relatively rare in Japanese HSP patients. The frequency of SPG3A is lower than that in the Caucasian populations.[5] The frequencies of SPG8 (4.1%) and SPG31 (4.1%) in AD-HSP are comparable to those reported in Caucasian populations.[20,21]

In the AR-HSP group, we found five families with SPG11 and one family with SPG21. Among the 14 patients with a thin corpus

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.

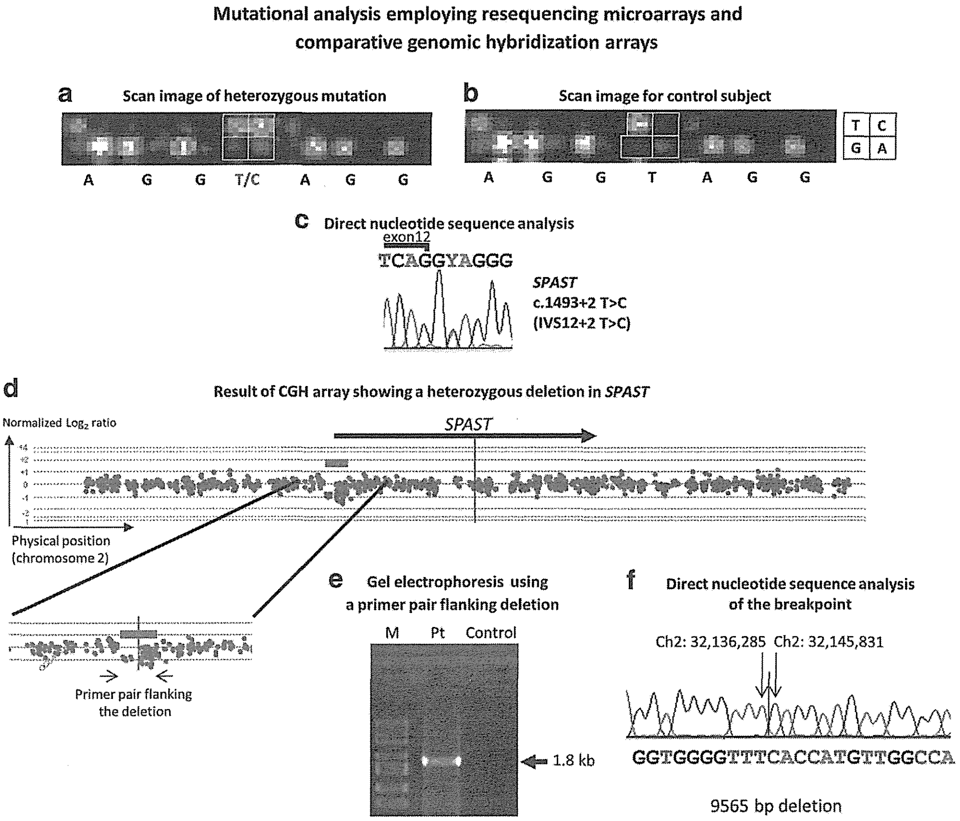
Mutational analysis employing resequencing microarrays and comparative genomic hybridization arrays

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.

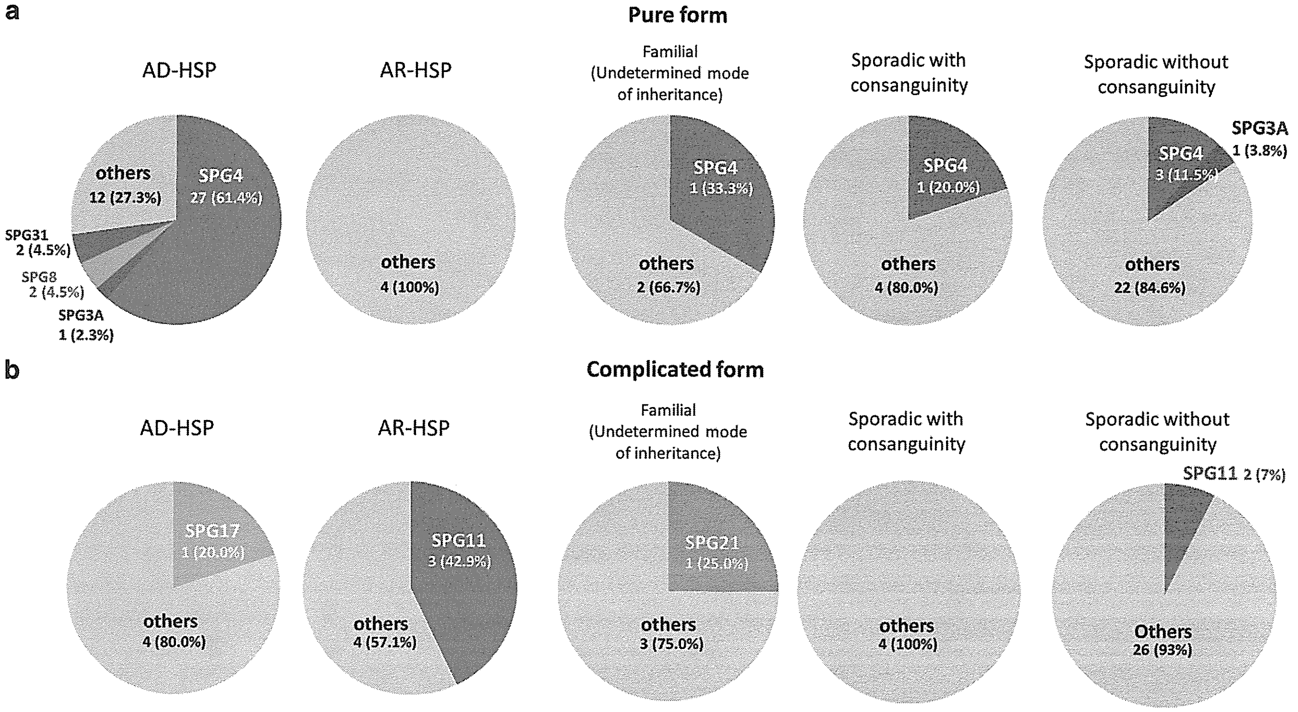

Figure 3 Relative frequencies of individual HSP types in groups classified on the basis of the clinical presentations and mode of inheritance. The figure shows the relative frequencies of individual HSP types in our cohort. (**a**) Pure form and (**b**) complicated form. The family history of each subgroup 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

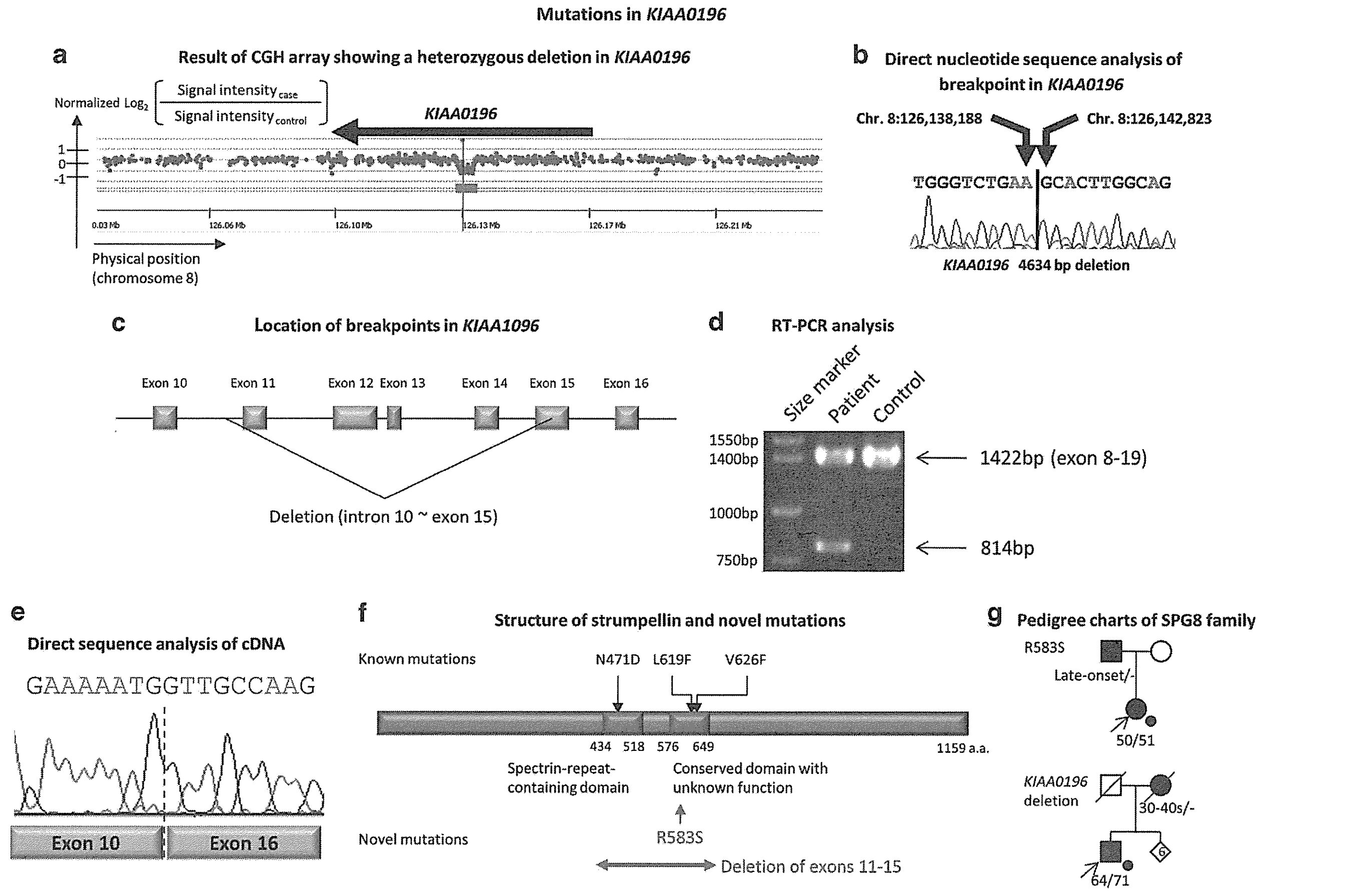
Mutations in *KIAA0196*
a
Result of CGH array showing a heterozygous deletion in *KIAA0196*
Normalized Log2 [Signal intensity case / Signal intensity control]
KIAA0196
1
0
-1
0.03 Mb
126.06 Mb
126.10 Mb
126.13 Mb
126.17 Mb
126.21 Mb
Physical position (chromosome 8)
b
Direct nucleotide sequence analysis of breakpoint in *KIAA0196*
Chr. 8:126,138,188
Chr. 8:126,142,823
TGGGTCTGAA
GCACTTGGCAG
KIAA0196 4634 bp deletion
c
Location of breakpoints in *KIAA1096*
Exon 10
Exon 11
Exon 12
Exon 13
Exon 14
Exon 15
Exon 16
Deletion (intron 10 ~ exon 15)
d
RT-PCR analysis
Size marker
Patient
Control
1550bp
1400bp
1000bp
750bp
1422bp (exon 8-19)
814bp
e
Direct sequence analysis of cDNA
GAAAAATGGTTGCCAAG
Exon 10
Exon 16
f
Structure of strumpellin and novel mutations
Known mutations
N471D
L619F
V626F
434
518
576
649
1159 a.a.
Spectrin-repeat-containing domain
Conserved domain with unknown function
R583S
Novel mutations
Deletion of exons 11-15
g
Pedigree charts of SPG8 family
R583S
Late-onset/-
50/51
KIAA0196 deletion
30-40s/-
6
64/71

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

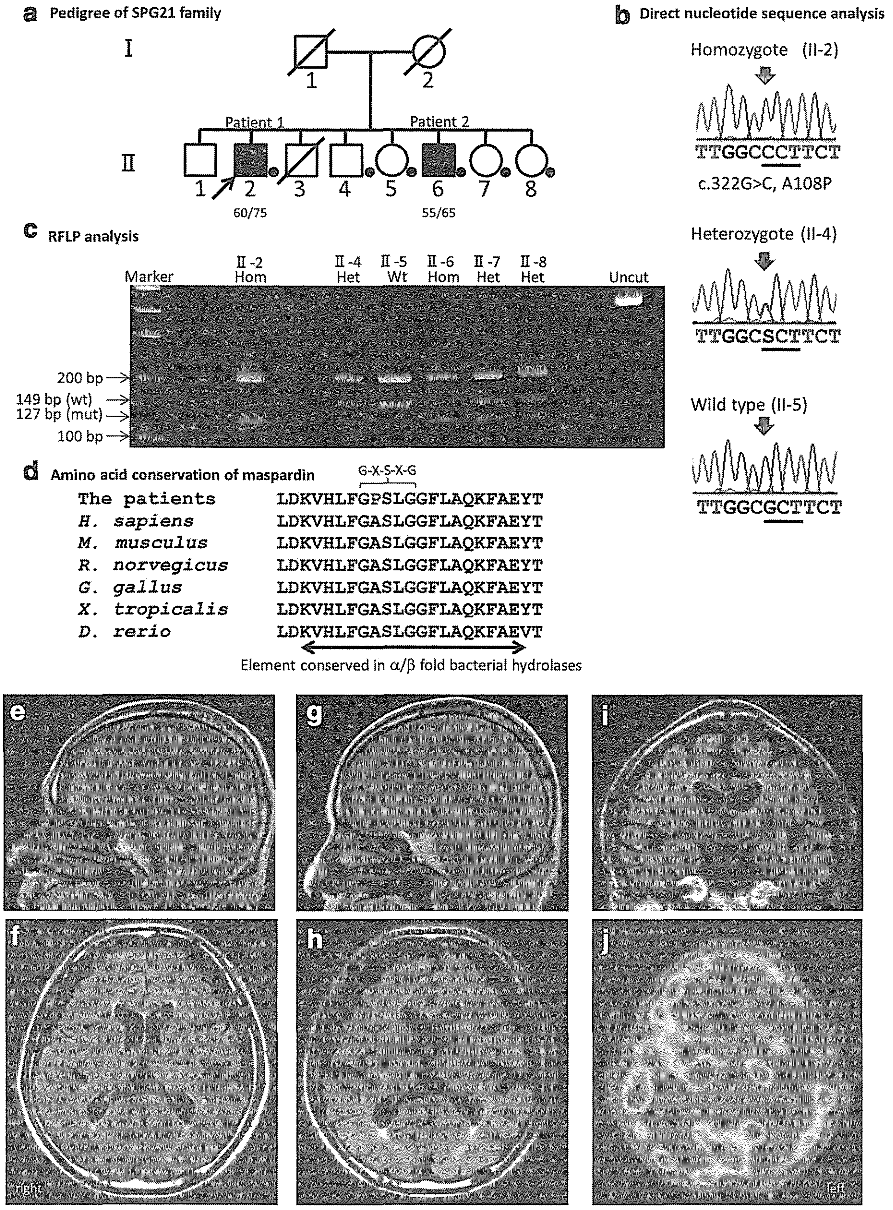
a Pedigree of SPG21 family
I
1
2
Patient 1
Patient 2
II
1
2
3
4
5
6
7
8
60/75
55/65
b Direct nucleotide sequence analysis
Homozygote (II-2)
TTGGCCCTTCT
c.322G>C, A108P
Heterozygote (II-4)
TTGGCSCTTCT
Wild type (II-5)
TTGGCGCTTCT
c RFLP analysis
Marker
II-2 Hom
II-4 Het
II-5 Wt
II-6 Hom
II-7 Het
II-8 Het
Uncut
200 bp
149 bp (wt)
127 bp (mut)
100 bp
d Amino acid conservation of maspardin
G-X-S-X-G
The patients LDKVHLFGPSLGGFLAQKFAEYT
H. sapiens LDKVHLFGASLGGFLAQKFAEYT
M. musculus LDKVHLFGASLGGFLAQKFAEYT
R. norvegicus LDKVHLFGASLGGFLAQKFAEYT
G. gallus LDKVHLFGASLGGFLAQKFAEYT
X. tropicalis LDKVHLFGASLGGFLAQKFAEYT
D. rerio LDKVHLFGASLGGFLAQKFAEVT
Element conserved in α/β fold bacterial hydrolases
e
g
i
f
h
j
right
left

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

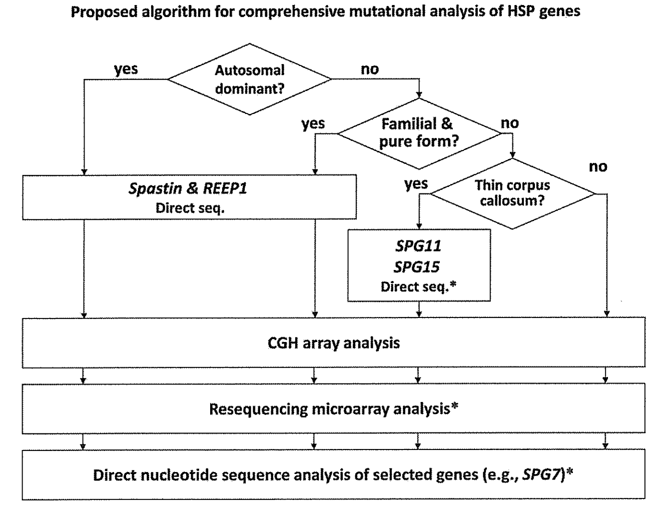
Proposed algorithm for comprehensive mutational analysis of HSP genes
Autosomal dominant?
yes
no
Familial & pure form?
yes
no
Thin corpus callosum?
yes
no
Spastin & REEP1
Direct seq.
SPG11
SPG15
Direct seq.*
CGH array analysis
Resequencing microarray analysis*
Direct nucleotide sequence analysis of selected genes (e.g., *SPG7*)*

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.* Atlastinl 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^{TM} 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 $(CCTG)_n$, $n=\sim5000$ 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} ($\mathrm{SD} = 0.15 \times 10^{-8}$) (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:

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

(CAG)$_2$CA, a repetition with unit CAG, is maximal. In addition, the entire string is also a maximal repetition with unit (CAG)$_2$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(CAG)$_6$CATCAGGGA,

   we can extend repetition (CAG)$_6$CA to the underlined substring,

   CGCCCGCAGC-GCA**T**(CAG)$_6$CA**T**CAGGGA,

   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.

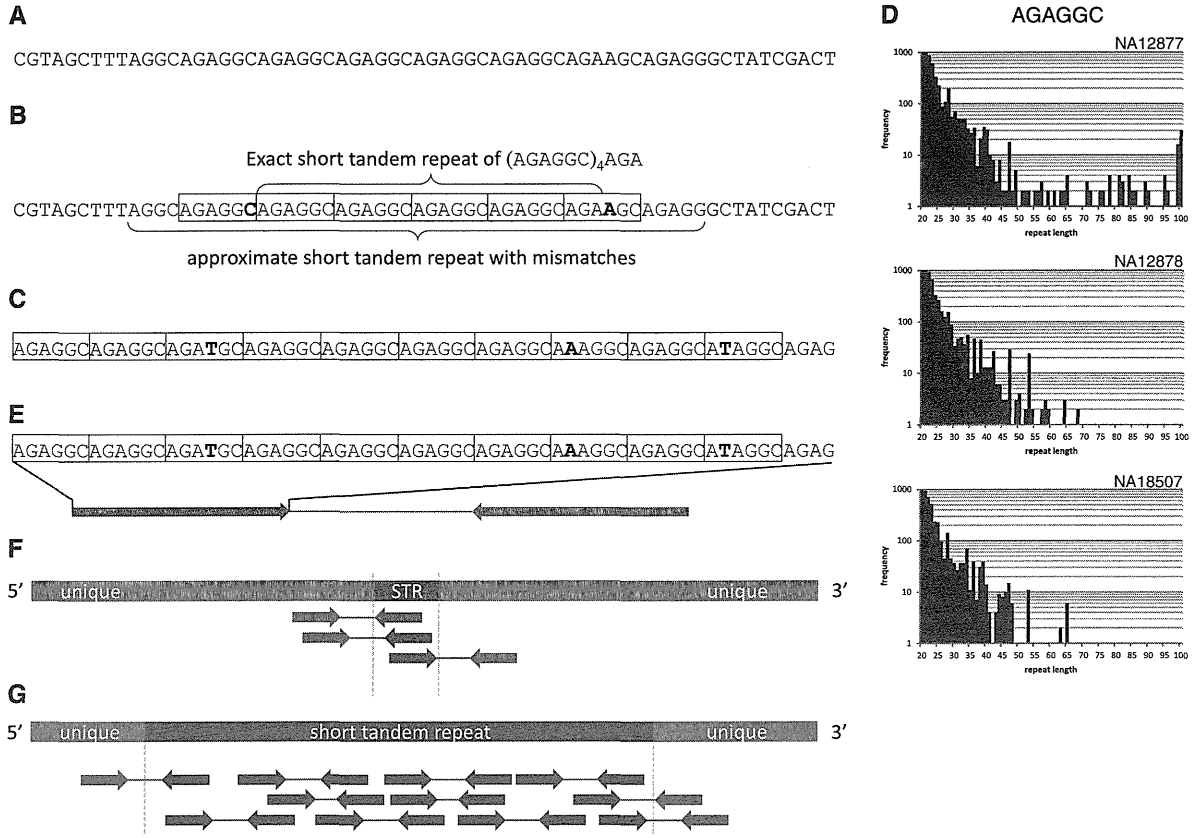
A
CGTAGCTTTAGGCAGAGGCAGAGGCAGAGGCAGAGGCAGAGGCAGAAGCAGAGGGCTATCGACT
B
Exact short tandem repeat of (AGAGGC)4AGA
CGTAGCTTTAGGCAGAGGCAGAGGCAGAGGCAGAGGCAGAGGCAGAAGCAGAGGGCTATCGACT
approximate short tandem repeat with mismatches
C
AGAGGCAGAGGCAGATGCAGAGGCAGAGGCAGAGGCAGAGGCAAAGGCAGAGGCATAGGCAGAG
D
AGAGGC
NA12877
NA12878
NA18507
frequency
1000
100
10
1
20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
repeat length
E
AGAGGCAGAGGCAGATGCAGAGGCAGAGGCAGAGGCAGAGGCAAAGGCAGAGGCATAGGCAGAG
F
5′
unique
STR
unique
3′
G
5′
unique
short tandem repeat
unique
3′

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^{TM} 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^{TM} sequencing (Eid *et al.*, 2009; Loomis *et al.*, 2013), to estimate or determine long expansions of STRs. In particular, SMRT^{TM} 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^{TM} 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^{TM} 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^{TM} 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^{TM} templates using the DNA Polymerase Binding Kit 2.0 (Pacific Biosciences). SMRTbell^{TM} 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