OPEN

Citation: Human Genome Variation (2017) **4**, 17032; doi:10.1038/hgv.2017.32
Official journal of the Japan Society of Human Genetics

www.nature.com/hgv

ARTICLE

Deep sequencing reveals variations in somatic cell mosaic mutations between monozygotic twins with discordant psychiatric disease

Yoshiro Morimoto[1,2], Shinji Ono[1,2], Akira Imamura[1], Yuji Okazaki[3], Akira Kinoshita[2], Hiroyuki Mishima[2], Hideyuki Nakane[4], Hiroki Ozawa[1], Koh-ichiro Yoshiura[2] and Naohiro Kurotaki[1]

Monozygotic (MZ) twins have been thought to be genetically identical. However, recent studies have shown discordant variants between them. We performed whole-exome sequencing (WES) in five MZ twin pairs with discordant neurodevelopmental disorders and one healthy control MZ twin to detect discordant variants. We identified three discordant variants confirmed by deep sequencing after analysis by personalized next-generation sequencing (NGS). Three mutations in *FBXO38* (chr5:147774428;T > G), *SMOC2* (chr6:169051385;A > G) and *TDRP* (chr8:442616;A > G), were detected with low allele frequency of mutant alleles on deep sequencing, suggesting that these loci are mosaic due to somatic mutations in a developmental stage. Our results suggest that deep sequencing analysis would be an adequate method to detect discordant mutations in candidate genes responsible for heritable diseases.

Human Genome Variation (2017) **4,** 17032; doi:10.1038/hgv.2017.32; published online 27 July 2017

INTRODUCTION

Neurodevelopmental disorders including schizophrenia, autism spectrum disorder (ASD) and gender dysphoria (GD) are thought to be multifactorial diseases. Genetic factors as well as environmental factors including life events are involved in their pathogenesis.

Schizophrenia is a chronic, debilitating psychiatric illness with a worldwide prevalence of 1%. A number of genetic studies have shown that schizophrenia has a high heritability and is strongly influenced by genetic factors. Several twin studies show that the concordance rate for schizophrenia between MZ twins is 41–79%, and 0–14% between dizygotic twins.[1] ASD is a heterogeneous, behaviorally defined, neurodevelopmental disorder that occurs in 1 in 150 children, characterized by deficits in communication, impaired social interaction and restricted patterns of behavior. Twin studies of ASD have demonstrated a significantly higher concordance rate for MZ twins (70–90%) than for dizygotic twins (0–10%).[2,3] Previously referred to as gender identity disorder, GD is a condition where a person experiences discomfort or distress due to a mismatch between their biological sex and gender identity. Twin studies of GD also show high concordance rates in MZ twins.[4] The higher concordance rate in MZ rather than dizygotic twins in these conditions suggests a contributory role for hereditary factors. A number of family and twin studies have shown that genetic factors play an important role in the onset of various neurodevelopmental disorders.

Phenotypically discordant MZ twins are interesting resources for genetic studies in diseases with a high concordance rate, and can make twin studies helpful in identifying the causative genes for heritable diseases. Identifying the molecular genetic differences between discordant MZ twins can help to elucidate the molecular mechanisms that underlie phenotypic discordance. To date, only a few studies have successfully detected genetic differences such as repeat length, single nucleotide variants (SNVs) and copy number variants (CNVs) between discordant MZ twins.[5–7]

A number of genetic studies for neurodevelopmental disorders have been reported; however, because of genetic heterogeneity, their pathogenesis is still unclear. To identify causative genes for neurodevelopmental disorders, we performed WES and deep sequencing in five pairs of MZ twins with discordant neurodevelopmental disorders and one healthy control MZ twin. Here, we describe the results of WES and deep sequencing to identify discordant variants between MZ twins.

MATERIALS AND METHODS

Subjects

Subjects were three pairs of MZ twins with discordant schizophrenia, one pair of MZ twins with discordant GD, one pair of MZ twins with discordant ASD and their parents, one pair of healthy MZ twins and two unrelated healthy volunteers as normal controls. In all twin pairs, the onset of neuropsychiatric conditions had occurred 10 years earlier. Detailed information was summarized in Table 1. All patients were diagnosed with psychiatric diseases according to the International Classification of Diseases, revision 10 and Diagnostic and Statistical Manual of Mental Disorders, fifth edition by two experienced psychiatrists. Peripheral blood samples, 10 ml each, were collected after obtaining written informed consent, and genomic DNA was extracted from blood lymphocytes using QIAamp DNA Mini Kit (QIAGEN, Hilden, Germany). Experimental procedures were approved by the Committee for the Ethical Issues on Human Genome and Gene Analysis at Nagasaki University.

[1]Department of Neuropsychiatry, Unit of Translation Medicine Nagasaki University Graduate School of Biomedical Sciences, Nagasaki, Japan; [2]Department of Human Genetics, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki, Japan; [3]Department of Psychiatry, Koseikai Michino-o Hospital, Nagasaki, Japan and [4]Unit of Rehabilitation Science, Department of Psychiatric Rehabilitation Science, University Graduate School of Biomedical Sciences, Nagasaki, Japan.
Correspondence: K Yoshiura (kyoshi@nagasaki-u.ac.jp)

Received 30 January 2017; revised 23 April 2017; accepted 20 June 2017

Table 1. Summary of samples in this study

Sample ID	*Phenotype*	*Sex*	*Age at sample collection (years)*
ASD-A	Autism spectrum disorder	Male	11
ASD-B	Normal	Male	11
ASD-Father	Normal	Male	50
ASD-Mother	Normal	Female	45
GD-A	Gender disorder (male-to-female)	Male	30
GD-B	Normal	Male	30
Sc-11-A	Schizophrenia	Male	41
Sc-11-B	Normal	Male	41
Sc-21-A	Schizophrenia	Male	38
Sc-21-B	Normal	Male	38
Sc-51-A	Schizophrenia	Male	41
Sc-51-B	Normal	Male	41
13A	Normal	Male	Unknown
13B	Normal	Male	Unknown
Control-1	Normal	Male	31
Control-2	Normal	Male	38

Abbreviations: ASD, autism spectrum disorder; GD, gender dysphoria.

Whole-exome sequencing

For WES, subjects' genomic DNA samples were enriched using SureSelect Human All Exon V5 (Agilent Technologies, Santa Clara, CA, USA) according to manufacturer's protocol. Enriched genomic libraries were sequenced with HiSeq2500 (Illumina, CA, USA).

Sequencing data analysis

Fastq format files were generated by the bcl2fastq software for HiSeq2500 and by the MiSeq Reporter software for MiSeq (Illumina). The Novoalign software (Novocraft, Selangor, Malaysia) was used to align reads on the hg19/GRCh37 human reference genome sequence. Aligned reads were sorted by the Novosort software (Novocraft) and were subjected to marking of PCR and optical duplication by MarkDuplicates in the Picard tools package (http://broadinstitute.github.io/picard/).

The Genome Analysis Toolkit[8] was used to perform local realignment (Genome Analysis Toolkit IndelRealigner) and variant call (Genome Analysis Toolkit HaplotypeCaller) implemented an in-house workflow management tool.[9] WES data was applied to eXome-Hidden Markov Model[10] analysis for detecting discordant CNVs within the MZ twins.

SNVs and insertion/deletion (Indels) were annotated using the ANNOVAR software.[11] Mutations that met the following criteria were selected as 'deleterious' mutations: (1) mutations: lead stop gain, stop loss, nonsynonymous or splice site mutations according to GENCODE basic version 19 downloaded from the UCSC genome browser; (2) alternative allele frequencies at mutation loci $\leqslant 0.5\%$ in the following databases: 1000 genome project all populations data released October 2014;[12] NIH NHLBI 6515 exome data (http://evs.gs.washington.edu/EVS/); Exome Aggregation Consortium 65000 exome data;[13] Human Genomic Variation Database exome data of 1,208 individuals collected in Japan;[14] SNV allele frequency collected from whole-genome sequencing data of 2,049 healthy Japanese population (https://ijgvd.megabank.tohoku.ac.jp) and (3) mutations not included in the UCSC segmental duplication region. To find genomic differences in discordant MZ twins, SNVs and indels were selected based on the following criteria: (1) the patient of the twin pair is not a reference homozygote, (2) a sibling of the twins is a reference homozygote and (3) a deleterious mutation. Average mean depth of all samples was 88.93 and average coverage rate of $> \times 10$ depth regions was 97.53%. Mean depth and coverage rate of all samples were summarized in Supplementary Table 4.

Sanger sequencing

We designed primers for discordant genotype loci between twins in WES. All primers were designed using Primer3.[15] Genomic DNA (5 ng) was amplified using KOD FX Neo polymerase (TOYOBO, Osaka, Japan) by PCR in 20 μl reactions. PCR mixtures were thermal cycled 32 times in the following conditions after initial denaturation at 94 °C for 2 min: denaturation 98 °C for 10 s, annealing 60 °C for 15 s, extension 68 °C for 30 s and final extension at 68 °C for 5 min. The PCR products were purified with the AMPureXP (Agencourt, Beverly, MA, USA) and sequence reactions were performed using the BigDye Terminator Cycle Sequencing Kit v3.1 (Applied Biosystems, Foster City, CA, USA) according to the manufacturer's instructions. After purification using the CleanSEQ (Agencourt), the products were separated on an ABI Genetic Analyzer 3130 (Applied Biosystems) and the electropherograms were evaluated using the 4Peaks software (http://nucleobytes.com/4peaks/).

Deep sequencing

PCR primers amplifying all candidate loci were designed using Primer3. Amplification was performed as for Sanger sequencing described above. PCR amplicons were mixed in every sample. Mixed PCR amplicons was sonicated to a size of 180 bp using Covaris E220 (Covaris, Wobum, MA, USA). End-repair, A-tailing and adapter ligation were prepared according to manufacturer's protocol for KAPA HTP Library Preparation Kits for Illumina Platforms. After adapter ligation, samples were amplified using KAPA HiFi HotStart Ready Mix (KAPA Biosystems, Wilmington, MA, USA) by PCR with indexed primers. Samples were thermal cycled seven times under the following conditions after initial denaturation at 98 °C for 45 s: denaturation 98 °C for 15 s, annealing 60 °C for 30 s, extension 72 °C for 30 s and final extension at 72 °C for 1 min. Prepared amplicon libraries were sequenced by MiSeq (Illumina). Mutant alleles of candidate loci were counted using Integrative Genomics Viewer.[16] Mean depths of all samples were summarized in Supplementary Table 5. To conform the discordant mutant alleles of three loci (chr5:147774428; T > G, chr6:169051385; A > G, chr8:442616; A > G) detected in GD twins, we performed resequencing analysis by Miseq (Illumina) using newly designed PCR primers and high fidelity Taq polymerase (KAPA HiFi Hot Start Ready Mix). Read depths of resequencing analysis was summarized in Supplementary Table 6.

Discordant CNV validation

A Droplet Digital PCR (ddPCR) QX200 system (Bio-Rad Laboratories, Hercules, CA, USA), was used to perform ddPCR to confirm copy number change. PCR primers amplifying all three candidate loci and reference primers within the *GNAS* and *RPPH1* genes were designed using Primer3. Ten nanograms of genomic DNA and 2 μM primer pair were added to the 20 μl PCR mixture containing 10 μl 2 × EvaGreen Supermix (Bio-Rad). Subsequently, 70 μl Droplet Generator Oil for EvaGreen (Bio-Rad) was added and droplet generation was performed using a DX200 Droplet Generator following manufacturer's protocol. The micro-reactor emulsion was thermal cycled 40 times under the following conditions after initial denaturation at 95 °C for 30 s: denaturation 95 °C for 30 s, annealing and extension 62 °C for 2 min, signal stabilization 4 °C for 5 min, then 90 °C for 5 min. Amplified droplet number was counted using a QX200 Droplet Reader (Bio-Rad). Absolute droplet numbers of *RPPH1* gene were used to calculate the copy number of three candidate loci with *GNAS* gene as control. Absolute droplet numbers of target loci were divided by that of *RPPH1* gene, then multiplied by two to calculate the relative copy number.

RESULTS

Evaluation of single-variant nucleotides and short indels

We performed WES to identify discordant heterozygous SNVs and short indels present in affected individuals. Clinical phenotypes of all samples are summarized in Table 1. A total of 72 discordant SNVs and indels calls were detected on WES (Supplementary Table 1). We then validated these calls by capillary sequencing and deep sequencing using MiSeq (Illumina).

A total of 22 loci could not be amplified by Sanger sequencing due to technical difficulties. Of the 50 loci validated by Sanger sequencing, 11 were concordant heterozygous variants (both affected and co-twin have same variants) and 39 loci were concordant with reference alleles. Sanger sequencing did not identify any obvious discordant heterozygous variants (SNVs/indels) (Supplementary Table 2).

Subsequently, we performed deep sequencing to more accurately evaluate the allele frequency of mutant alleles in these

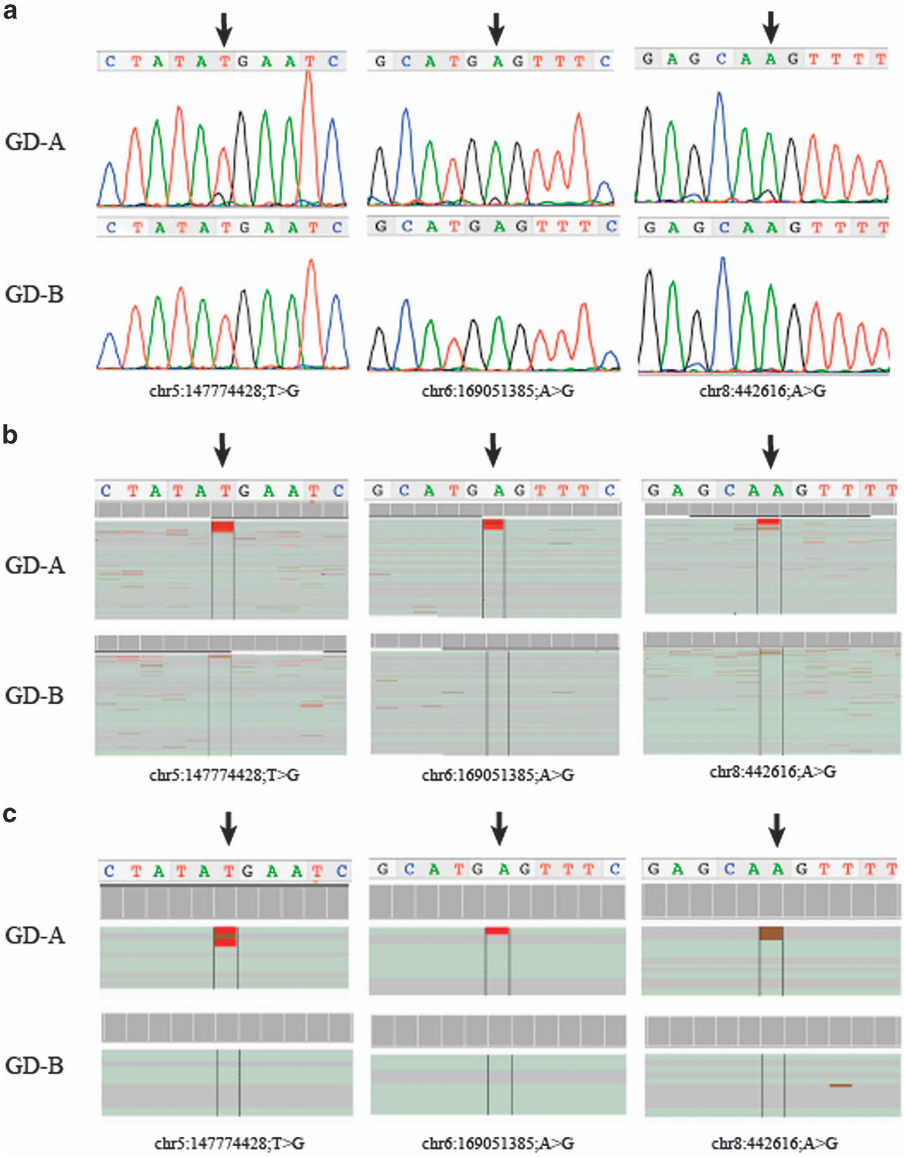

Figure 1. Discordant SNVs in GD MZ twin. (**a**) Electropherograms of discordant SNVs in GD twin. Mutant bases of GD-A barely confirmed in electropherograms. (**b**) Aligned reads visualized by IGV of discordant SNVs in GD twin. Mutant bases in aligned reads are displayed red and brown. Aligned reads show mutant bases clearly despite low allele frequency. (**c**) Aligned reads of resequencing analysis. Mutant bases of GD-A are conformed. Abbreviations: GD, gender dysphoria; IGV, Integrative Genomics Viewer; MZ, monozygotic; SNV, single-variant nucleotide.

Table 2. Summary of candidate discordant CNVs on XHMM analysis

Sample	*Interval*	*CytoBand*	*XHMM analysis*	*Relative copy number (ddPCR/mean)*
ASD-A	chr8:144887409–145012081	8q24.3	Reference	1.81548204
ASD-B	chr8:144887409–145012081	8q24.3	Duplication	1.916876751
ASD-A	chr9:139902662–139912822	9q34.3	Reference	2.236417293
ASD-B	chr9:139902662–139912822	9q34.3	Duplication	2.237732824
GD-A	chr14:106330725–106405832	14q32.33	Reference	4.088276371
GD-B	chr14:106330725–106405832	14q32.33	Deletion	4.226054708

Abbreviations: ASD, autism spectrum disorder; CNV, copy number variation; GD, gender dysphoria; XHMM, eXome-Hidden Markov Model.

samples. Eight loci could not be sequenced. Of the 64 loci for which the allele frequency of mutant alleles could be evaluated, 61 did not show any difference of allele frequency between twins, while 3 showed clear discordant mutation allele frequency dominantly in patients. All three discordant loci were in the GD-affected twin. We confirmed the results by resequencing analysis using newly designed PCR primers and high fidelity Taq polymerase (Figure 1, Supplementary Table 2).

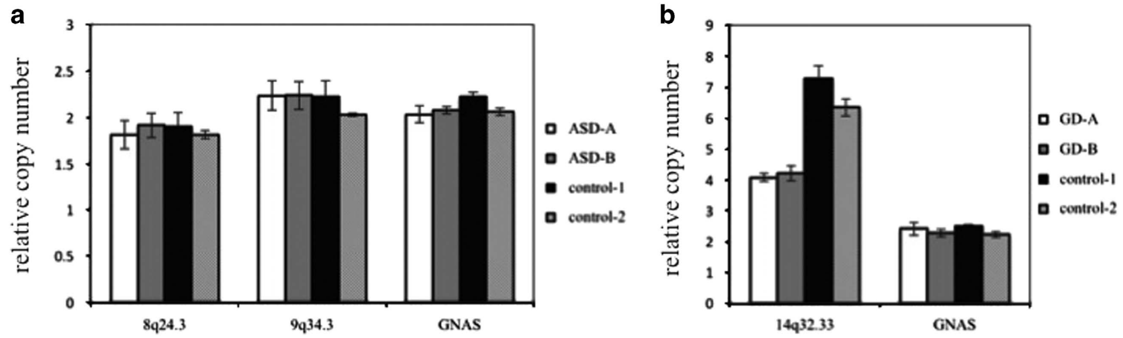

Figure 2. Discordant CNV analysis. (**a**) Relative copy numbers of 8q24.3, 9q34.3, and *GNAS* gene as control. All loci of all samples show ~2 copies. No difference in copy number was confirmed between ASD-A and ASD-B. (**b**) Relative copy number of 14q32.33 and *GNAS* gene. Variable copy number is shown among samples at 14q32.33 (GD-A; 4.088, GD-B; 4.226, control-1; 7.286, control-2; 6.349). No difference in copy number was confirmed between GD-A and GD-B. Abbreviations: ASD, autism spectrum disorder; CNV, copy number variation; GD, gender dysphoria.

Table 3. *In silico* analysis of discordant SNVs in GD twin

Gene	*Variants*	*cDNA change*	*Protein change*	*Polyphen2*		*PROVEAN*	*SIFT*
				HDivPred	*HVarPred*		
FBXO38	chr5:147774428;T > G	c.89T > G	p.Met30Arg	Possibly damaging	Benign	Deleterious	Damaging
SMOC2	chr6:169051385;A > G	c.965A > G	p.Glu322Gly	Probably damaging	Probably damaging	Deleterious	Damaging
TDRP	chr8:442616;A > G	c.341T > C	p.Leu114Pro	Probably damaging	Probably damaging	Deleterious	Damaging

Abbreviations: GD, gender dysphoria; *SMOC2*, *SPARC*-related modular calcium-binding 2; SNV, single-variant nucleotide; *TDRP*, testis development-related protein.

Analysis of CNV

CNV was suspected for three candidate loci on eXome-Hidden Markov Model analysis (Table 2), so we validated these candidate CNV loci by ddPCR. However, there was no CNV in the three candidate loci (Figure 2). Due to CNV from 4–7 copies among samples at 14q32.33, structural variations such as duplication would likely exist in this region.

In silico analysis of mutations with discordant allele frequency

We used the Polyhen_2,[17] SIFT[18] and PROVEN[19] programs to confirm whether these three variants affect protein function. All programs determined these three variants were 'damaging' (Table 3).

DISCUSSION

In this study, we report the successful detection of genomic differences between MZ twins discordant for GD. Three genomic changes dominant in affected individuals were confirmed by 'deep sequencing' and subsequent WES in MZ twin pairs. To the best of our knowledge, some studies have successfully identified somatic mutations between discordant MZ twins by several methods including Sanger sequencing, pyrosequencing and fluorescence *in situ* hybridization.[6,20,21] However, many twin studies could not identify genomic differences between MZ twin pairs. We previously reported a DNA microarray-based comparison between three pairs of discordant MZ twins with schizophrenia and we also were unable to identify any differences.[22] Two recent studies also attempted unsuccessfully to confirm discordant mutations by WES and Sanger sequencing in discordant MZ twin pairs.[23,24] Our results from deep sequencing show that variant allele frequencies of all three variants were ~10%, suggesting a small percentage of somatic cell mosaicism with mutation, too small to be identified by Sanger sequencing. Our findings suggest that somatic cell mosaicism may contribute to obscure genomic differences between MZ twins. Deep sequencing facilitates the detection of genomic differences even with low allele frequency in somatic mosaicism by read depth count between MZ twins using NGS.[25] Given that our deep sequencing results could also detect low allele frequency in somatic mosaicism, deep sequencing is an adequate method for detecting somatic mosaic mutations.

Nonetheless, deep sequencing also has some limitations when used to detect low allele frequency mutations. We found one locus that appeared as a low-frequency allele mutation on deep sequencing analysis in both twins and their father (Supplementary Figure 1A, Supplementary Table 3). However, Sanger sequencing showed that locus to be a heterozygous mutation. This highlights the phenomenon whereby heterozygous mutations appear as low allele frequency mutations on NGS data analysis. Therefore, when analyzing NGS data, it is important to compare allele frequency with other samples for detecting somatic mosaic mutation. Also, two loci showing low allele frequency mutations among all ASD family members (Supplementary Figure 1B, Supplementary Table 3). It is highly likely that those loci were false-positive calls caused by nonspecific PCR amplification or alignment errors when using NGS.

Twin studies of GD demonstrated higher concordance rates in MZ and suggested a strong heritable component to GD.[4,26,27] Postmortem brain studies in GD have shown the size of the bed nucleus of the stria terminalis in male-to-female transsexuals was closer to typical female size than to male size,[28,29] but the volume and number of neurons were similar to that of a biological male.[30] These results suggest some neurological etiology underlying the development of GD. Even with low-level mosaicism in peripheral blood samples, somatic mutations can cause neurological diseases such as epilepsy and intellectual disability.[31] In silico analysis showed that the effects of all three mutations were 'damaging'. F-box protein 38 (*FBXO38*) gene are known as modulator of

Küppel-like transcription factor 7 and the missense substitution of *FBXO38* gene cause spinal muscular atrophies.[32] In addition, homozygous mutation in *SPARC*-related modular calcium-binding 2 (*SMCO2*) gene cause sever developmental dental defect[33] and the potential role of fetal gonad in developmental stage are reported.[34] Specifically, testis development-related protein (*TDRP*) gene was found to be important in spermatogenesis and male infertility,[35,36] and may be related to maleness. In this context, any of 3 genes could be disease-causing genes.

Recently, Ye *et al* performed whole-genome sequencing in two pairs of MZ twins, 40 and 100 years old, in two independent NGS experiments. They showed that somatic mutations detected between MZ twins were related to aging but the number of such variants were hardly found even in 100-year-old MZ twins.[37] Given that the GD twin in this study was 30 years old and the allele frequencies were different in MZ twins, the 3 mosaic mutations would have occurred in the developmental stage, not just as a consequence of aging. The result that we didn't detect any discordant somatic mutations in other elder twins may support this idea. Thus, further investigations including functional analysis and epidemiological analysis are needed to verify the significance of the mutations found in this study.

We also performed CNV detection by eXome-Hidden Markov Model analysis and validated candidate loci using the ddPCR system between MZ twin pairs. The ddPCR system yields absolute quantification and can easily detect small fold changes in CNVs, and is thus a suitable method for detecting the small fold of change in CNVs of somatic cell mosaic events.[38] Given that the results of our CNV analysis showed no differences in CNVs between discordant MZ twins, discordant CNVs between MZ twin pairs would be a rare event. Nonetheless, it is necessary to investigate CNV events because they would have a considerable impact on the biological process. Bruder *et al* reported a study in 19 pairs of MZ twins with either concordant or discordant phenotype using two platforms for genome-wide CNVs analyses. They showed the presence of CNVs within both groups.[6] Because these CNVs could be somatic cell mosaic mutations, mutant allele frequencies may be very low in samples and exact mosaic frequencies should be measured in affected tissues including brain tissue in the future.

In conclusion, we analyzed discordant variants in MZ twin pairs and one healthy control MZ twin using WES, and detected three somatic cell mosaic mutations in the GD twins by deep sequencing. It might be important to consider the presence of somatic cell mosaic mutations when confirming data obtained from WES. Furthermore, these somatic mosaic mutations might help to elucidate the pathogenesis of GD.

AUTHOR CONTRIBUTIONS

Conception and design of the study: K-IY, NK, AI, YO, HN, HO. Analysis and interpretation of data: YM, SO, AK, HM. Critical revision of the article for important intellectual content: YM. Final approval of the article: K-IY.

COMPETING INTERESTS

The authors declare no conflict of interest.

PUBLISHER'S NOTE

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1 Shih RA, Belmonte PL, Zandi PP. A review of the evidence from family, twin and adoption studies for a genetic contribution to adult psychiatric disorders. *Int Rev Psychiatry* 2004; **16**: 260–283.

2 Steffenburg S, Gillberg C, Hellgren L, Andersson L, Gillberg IC, Jakobsson G *et al.* A twin study of autism in Denmark, Finland, Iceland, Norway and Sweden. *J Child Psychol Psychiatry* 1989; **30**: 405–416.

3 Bailey A, Le Couteur A, Gottesman I, Bolton P, Simonoff E, Yuzda E *et al.* Autism as a strongly genetic disorder: evidence from a British twin study. *Psychol Med* 1995; **25**: 63–77.

4 Coolidge FL, Thede LL, Young SE. The heritability of gender identity disorder in a child and adolescent twin sample. *Behav Genet* 2002; **32**: 251–257.

5 Kondo S, Schutte BC, Richardson RJ, Bjork, Knight AS, Watanabe Y *et al.* Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. *Nat Genet* 2002; **32**: 285–289.

6 Bruder CE, Piotrowski A, Gijsbers AA, Andersson R, Erickson S, Diaz de Atahi T *et al.* Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *Am J Hum Genet* 2008; **82**: 763–771.

7 Herdermann-van den Enden AT, Maaswinkel-Mooij PD, Hoogendoorn E, Willenmsen R, Maat-Kievit JA, Losekoot M *et al.* Monozygotic twin brothers with the fragile X syndrome: different CGG repeats and different mental capacities. *J Med Genet* 1999; **36**: 253–257.

8 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010; **20**: 1297–1303.

9 Mishima H, Sasaki K, Tanaka M, Tatebe O, Yoshiura K. Agile parallel bioinformatics workflow management using Pwrake. *BMC Res Notes* 2011; **4**: 331–338.

10 Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, Ruderfer DM *et al.* Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet* 2012; **91**: 597–607.

11 Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from next-generation sequencing data. *Nucleic Acids Res* 2010; **38**: e164.

12 1000 Genomes Projects Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM *et al.* A global reference for human genetic variation. *Nature* 2015; **526**: 68–74.

13 Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T *et al.* Analysis of protein-cording genetic variation in 60,706 humans. *Nature* 2016; **536**: 285–291.

14 Higasa K, Miyake N, Yohimura J, Okamura K, Niihori T, Saitsu H *et al.* Human genetic variation database, a reference database of genetic variations in Japanese population. *J Hum Genet* 2016; **61**: 547–553.

15 Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 2000; **132**: 365–386.

16 Robinson JT, Thovaldsdottir H, Wickler W, Guttman M, Lander ES, Getz G *et al.* Integrative genomics viewer. *Nat Biotechnol* 2011; **29**: 24–26.

17 Adzhubei IA, Schimidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* 2010; **7**: 248–249.

18 Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009; **4**: 1073–1081.

19 Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* 2012; **7**: e46688.

20 Galetzka D, Hansmann T, El Hajj N, Weis E, Irmscher B, Ludwig M *et al.* Monozygotic twins discordant for constitutive BRCA1 promoter methylation, childhood cancer and secondary cancer. *Epigenetics* 2012; **7**: 47–54.

21 Razzaghian HR, Shahi MH, Forsberg LA, de Ståhl TD, Absher D, Dahl N *et al.* Somatic mosaicism for chromosome X and Y aneuploidies in monozygotic twins heterozygous for sickle cell disease mutation. *Am J Med Genet* 2010, A **152A**: 2595–2598.

22 Ono S, Imamura A, Tasaki S, Kurotaki N, Ozawa Z, Yoshiura K *et al.* Failure to confirm CNVs as of aetiological significance in twin pairs discordant for schizophrenia. *Twin Res Hum Genet* 2010; **13**: 455–460.

23 Zhang R, Thiele H, Bartmann P, Hilger AC, Berg C, Herberg U *et al.* Whole-Exome Sequencing In Nine Monozygotic Discordant Twins. *Twin Res Hum Genet* 2016; **19**: 60–65.

24 Lyu N, Guan LL, Ma H, Wang XJ, Wu BM, Shang FH *et al.* Failure to identify somatic mutations in monozygotic twins discordant for schizophrenia by whole exome sequencing. *Chin Med J* 2016; **129**: 690–695.

25 Li R, Montpetit A, Rousseau M, Wu SY, Greenwood CM, Spector TD *et al.* Somatic point mutations occurring early in development: a monozygotic twin study. *J Med Genet* 2014; **51**: 28–34.

26 Bailey JM, Dunne MP, Martin NG. Genetic and environmental influences on sexual orientation and its correlates in an Australian twin sample. *J Pers Soc Psychol* 2000; **78**: 524–536.

27 Lippa R, Hershberger S. Genetic and environmental influences on individual differences in masculinity, femininity, and gender diagnosticity: Analyzing data from a classic twin study. *J Personality* 1999; **67**: 127–155.

28 Zhou JN, Hofman MA, Gooren LJ, Swaab DF. A sex difference in the human brain and its relation to transsexuality. *Nature* 1995; **378**: 68–70.

29 Savic I, Garcia-Falgueras A, Swaab DF. Sexual differentiation of the human brain in relation to gender identity and sexual orientation. *Prog Brain Res* 2010; **186**: 41–62.

30 Segal NL. Two monozygotic twin pairs discordant for female-to-male transsexualism. *Arch Sex Behav* 2006; **35**: 347–358.

31 Poduri A, Evrony GD, Cai X, Walsh CA. Somatic mutation, genomic variation, and neurological disease. *Science* 2013; **341**: 1237758.

32 Sumner CJ, d'Ydewalle C, Wooley J, Fawcett KA, Hernandez D, Gardiner AR *et al*. A dominant mutation in FBXO38 causes distal spinal muscular atrophy with calf predominance. *Am J hum Genet* 2013; **93**: 976–983.

33 Bloch-Zupan A, Jamet X, Etard C, Laugel V, Muller J, Geoffroy V *et al*. Homozygosity mapping and candidate prioritization identify mutations, missed by whole-exome sequencing, in SMOC2, causing major dental developmental defects. *Am J hum Genet* 2011; **89**: 773–781.

34 Pazin DE, Albrecht KH. Developmental expression of Smoc1 and Smoc2 suggests potential roles in fetal gonad and reproductive tract differentiation. *Dev Dyn* 2009; **238**: 2877–2890.

35 Wang X, Jiang H, Zhou W, Zhang Z, Yang Z, Lu Y *et al*. Molecular cloning of a novel nuclear factor, TDRP1, in spermatogenic cells of testis and its relationship with spermatogenesis. *Biochem Biophys Res Commun* 2010; **394**: 29–35.

36 Mao S, Wu F, Cao X, He M, Liu N, Wu H *et al*. TDRP deficiency contributes to low sperm motility and is a potential risk factor for male infertility. *Am J Transl Res* 2016; **8**: 177–187.

37 Ye K, Beekman M, Lameijer EW, Zhang Y, Moed MH, van den Akker EB *et al*. Aging as accelerated accumulation of somatic variants: whole-genome sequencing of centenarian and middle-aged monozygotic twin pairs. *Twin Res Hum Genet* 2013; **16**: 1026–1032.

38 Abyzov A, Mariani J, Palejev D, Zhang Y, Haney MS, Tmasini L *et al*. Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. *Nature* 2012; **492**: 438–442.

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Information for this article can be found on the Human Genome Variation website (http://www.nature.com/hgv)