Identification of somatic mutations in postmortem human brains by whole genome sequencing and their implications for psychiatric disorders

Masaki Nishioka, MD, PhD,1,2,3 Miki Bundo, PhD,1,4,5 Junko Ueda, MS, 6 Fumiki Katsuoka, PhD,7 Yukuto Sato, PhD,7 Yoko Kuroki, PhD,7,8 Takao Ishii, MD, PhD,9 Wataru Ukai, PhD,9 Shigeo Murayama, MD, PhD,10 Eri Hashimoto, MD, PhD,9 Masao Nagasaki, PhD,7 Jun Yasuda, MD, PhD,7,9 Kiyoto Kasai, MD, PhD,2 Takafumi Kato, MD, PhD,2 and Kazuya Iwamoto, PhD,1,5

Departments of 1Molecular Psychiatry, 2Neuropsychiay, Graduate School of Medicine, 3Division for Counseling and Support, The University of Tokyo, Tokyo, 4PRESTO, Japan Science and Technology Agency, Saitama, 5Department of Molecular Brain Science, Graduate School of Medical Sciences, Kumamoto University, Kumamoto, 6Laboratory for Molecular Dynamics of Mental Disorders, RIKEN Brain Science Institute, Saitama, 7Department of Integrative Genomics, Tohoku Medical Megabank Organization, Tohoku University, Sendai, 8Department of Genome Medicine, National Research Institute for Child Health and Development, Tokyo, 9Department of Neuropsychiatry, School of Medicine, Sapporo Medical University, Sapporo, and 10Department of Neuropathology, Tokyo Metropolitan Institute of Gerontology, Tokyo, Japan

Aim: Somatic mutations in the human brain are hypothesized to contribute to the functional diversity of brain cells as well as the pathophysiology of neuropsychiatric diseases. However, there are still few reports on somatic mutations in non-neoplastic human brain tissues. This study attempted to unveil the landscape of somatic mutations in the human brain.

Methods: We explored the landscape of somatic mutations in human brain tissues derived from three individuals with no neuropsychiatric diseases by whole-genome deep sequencing at a depth of around 100. The candidate mutations underwent multi-layered filtering, and were validated by ultra-deep target amplicon sequencing at a depth of around 200 000.

Results: Thirty-one somatic mutations were identified in the human brain, demonstrating the utility of whole-genome sequencing of bulk brain tissue. The mutations were enriched in neuron-expressed genes, and two-thirds of the identified somatic single nucleotide variants in the brain tissues were cytosine-to-thymine transitions, half of which were in CpG dinucleotides.

Conclusion: Our developed filtering and validation approaches will be useful to identify somatic mutations in the human brain. The vulnerability of neuron-expressed genes to mutational events suggests their potential relevance to neuropsychiatric diseases.

Key words: human brain, neuron, psychiatric disorder, somatic mutation, whole genome sequencing.

Recent studies have revealed the existence of somatic mutations in the brains of individuals with and without neuropsychiatric diseases. Various types of mutations, including single nucleotide variants (SNV), copy number variations (CNV), and retrotransposons, have been reported in human brains with no neuropsychiatric disease. The number of somatic SNV in a single neuron has been...
estimated to be around 1500 by whole-genome sequencing (WGS) of whole-genome-amplified single-neuron genomes.\(^1\) Large somatic CNV encompassing more than 1 Mb were detected in 41% of neurons by WGS analysis of whole-genome-amplified single-neuron genomes.\(^2\) New insertions of retrotransposons, especially the long interspersed nuclear element 1 (LINE1), have been detected by single-cell analysis as well as target enrichment sequencing of bulk brain tissues.\(^3\)–\(^6\)

Somatic mutations in the relevant genes have been reported in brain samples derived from patients with neuropsychiatric diseases. Several studies have reported somatic mutations in AKT3, PIK3CA, and MTOR in patients with hemimegalencephaly or cortical dysplasia with epilepsy.\(^7\)–\(^12\) and increased LINE1 copy number has been reported in brains of patients with Rett syndrome, ataxia-telangiectasia, and schizophrenia.\(^13\)–\(^15\) Particularly, somatic mutations in MTOR were shown to be causative of epilepsy in a mouse model.\(^9\) Three somatic mutations in CACNA1C, SCN1A, and SETD2 were reported in brain samples derived from patients with autism spectrum disorder or fragile X syndrome.\(^16\) Somatic CNV were reported in the brains of patients with schizophrenia,\(^17\)–\(^18\) and somatic mutations in MAPT were reported in the brains of patients with Alzheimer’s disease.\(^19\) These studies suggested possible links between somatic mutations and psychiatric disorders as well as brain malformation diseases. However, the landscape of somatic mutations in the human brain and their relevance to individual phenotypes, including psychiatric disorders, remain to be explored.

Here, we explored somatic SNV from high-depth WGS data derived from human brain bulk tissues. Previous translational research utilized whole-exome sequencing or target amplicon sequencing (TAS) of bulk tissues in investigating somatic mutations, but the genomic regions of interest are not evident in most psychiatric disorders. We adopted WGS to detect somatic mutations at the genome scale. The candidate mutations identified at the genome scale underwent multilayered filtering and extensive validation by ultra-deep TAS. We successfully identified 31 somatic SNV via brain subregion-specific, as well as brain cell type-specific, analyses of three individuals with no neuropsychiatric diseases. We also characterized the landscape of somatic mutations in the brain, revealing the vulnerability of neuron-expressed genes to mutational events as well as temporal and spatial variation in somatic mutations.

The vulnerability of neuron-expressed genes to mutational events suggests that these somatic mutations are prone to increased risks of dysfunctions in neuronal networking, which can underlie the pathogenesis of psychiatric disorders by altering the structure or function of coded proteins or expression profiles of molecules. Our approach can be utilized for clinical investigation in large cohorts of patients with psychiatric disorders, complementing existing methods of single-neuron genomics. Herein, we investigate the technical validity of identification of somatic mutations in bulk postmortem brain tissue, comparing with previous single-cell-based approaches, and discuss the implications for future psychiatric research.

**METHODS**

To identify somatic SNV in the human brain, we explored candidate somatic mutations from WGS data on bulk brain tissues. The candidates were validated by TAS. The workflow is illustrated in Figure S1. Detailed protocols, analytical parameters, and analytical processes are described in File S1, Figures S1–S6, and Tables S1–S14 in Supporting Information. We designed and executed this study according to the Declaration of Helsinki (World Medical Association) and the guidelines provided by the Ministry of Education, Culture, Sports, Science, and Technology and the Ministry of Health, Labour, and Welfare of the Japanese government. The ethics committees of the University of Tokyo (Graduate School of Medicine), RIKEN, Tohoku Medical Mega-bank Organization, Sapporo Medical University, Tokyo Metropolitan Institute of Gerontology, and Kumamoto University approved this study.

**Samples**

The postmortem brain and liver samples were derived from three men with no neuropsychiatric diseases (details in Table S1). The postmortem tissue samples were obtained from the brain bank of Stanley Medical Research Institute (CL_WGS_set) or Tokyo Metropolitan Institute of Gerontology (NeuN_WGS_set). Informed consent had been obtained from the next of kin in these institutes. For the CC_WGS_set, written informed consent was obtained from the next of kin at Sapporo Medical
University. Neuronal and non-neuronal nuclei were isolated from one frontal cortex sample by NeuN-based nuclei sorting with an anti-NeuN antibody (Millipore, Billerica, MA, USA, #MAB377X, RRID: AB_2298772), according to the method described by Iwamoto et al.\textsuperscript{20}

WGS

We performed WGS on genomic DNA (gDNA) extracted from the brain and liver samples. Based on the tissue types and the sequencing strategy, we grouped the samples into three sets: CL\_WGS\_set (cortex and liver samples from a 68-year-old man sequenced on a HiSeq 2500 [Illumina, San Diego, CA, USA]), NeuN\_WGS\_set (neuronal and non-neuronal nuclei, and liver samples from a 78-year-old man sequenced on a HiSeq 2500), and CC\_WGS\_set (cortex and cerebellum samples from an 84-year-old man sequenced on a HiSeq X; Table 1). We prepared the sequencing libraries for the CL\_WGS\_set and NeuN\_WGS\_set, using the TruSeq DNA PCR-Free Sample Prep Kit (Illumina, San Diego, CA, USA) according to the manufacturer’s protocol. These libraries were prepared without polymerase chain reaction (PCR) to preclude PCR-related errors. The prepared libraries were sequenced on the HiSeq 2500 with Rapid Run PE mode at Tohoku Medical Megabank. We minimized the problem of carry-over (DNA contamination) by indexing. We prepared sequence libraries from the CC\_WGS\_set with the TruSeq Nano DNA sample Prep kit (Illumina), according to the manufacturer’s protocol. The prepared libraries were sequenced on HiSeq X (Illumina) at RIKEN GENESIS (Tokyo, Japan). HiSeq X has no carry-over (DNA contamination) problem by default.

Quality control and alignment

We performed sequence alignment and quality control (QC) lane-by-lane on RHEL 6.2 or CentOS 6.5 (Linux kernel 2.6.32) at the University of Tokyo. The sequence data were quality controlled by FastQC-0.11.2 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and Trimmomatic-0.32.\textsuperscript{21} After QC, the sequence data were subjected to BWA-0.7.12\textsuperscript{22} alignment to GRCh37 + decoy reference genome provided by Broad Institute (ftp://ftp.broadinstitute.org/). The alignment data (BAM file) were quality controlled by Picard Deduplication (Picard-1.102, https://github.com/broadinstitute/picard), GATK Indel Realignment (GATK-3.2-2\textsuperscript{23}), and GATK Base Recalibration (GATK-3.2-2), in this order. We merged the BAM files for each lane into a single BAM file, and performed Picard Deduplication on the merged BAM files in the CC\_WGS\_set. We then performed GATK Indel Realignment on the merged BAM files. The read data with mapping quality (mapQ) values \( \geq 1 \) were selected from the BAM files by SAMtools-0.1.19.\textsuperscript{24} The quality of sequencing

| Sample data | Sequence data |
|-------------|---------------|
| **Set**     | **ID**        | **Sample**     | **Sex** | **Age (years)** | **Sequencer** | **PCR** | **Read length** | **Read pair** | **Raw DP** | **QC DP** | **10x Cov. (%)** |
| CL\_WGS\_set| AL30\_cortex  | Frontal cortex | Male | 68 | HiSeq 2500 | – | 162 | 945708090 | 97.7 | 84.2 | 98.0 |
|             | AL30\_liver   | Liver          | –    | 162 | HiSeq 2500 | – | 162 | 951672483 | 98.3 | 81.5 | 98.0 |
| NeuN\_WGS\_set| Y8763\_NeuN+ | Neuronal nuclei | Male | 78 | HiSeq 2500 | – | 162 | 940935271 | 97.2 | 83.5 | 98.6 |
|             | Y8763\_NeuN- | Nonneuronal nuclei | –    | 162 | HiSeq 2500 | – | 162 | 943326695 | 97.4 | 85.1 | 98.6 |
| CC\_WGS\_set| S6\_cortex   | Frontal cortex | Male | 84 | HiSeq X | + | 150 | 1282279777 | 122.6 | 76.3 | 98.3 |
|             | S6\_cerebellum| Cerebellum     | –    | 150 | HiSeq X | + | 150 | 1257772546 | 120.3 | 74.1 | 98.2 |

\( ^\dagger \)Coverage (%) of alignment data to nonN bases in whole genome with depth \( \geq 10 \) after quality control.

DP, average depth of alignment data; PCR, PCR amplification process in sequence library preparation; QC DP, DP after quality control.

© 2017 The Authors. Psychiatry and Clinical Neurosciences published by John Wiley & Sons Australia, Ltd on behalf of Japanese Society of Psychiatry and Neurology.
data depended on the sequencing platform. The sequence data quality of the CC_WGS_set (HiSeq X) was lower than that of the CL_WGS_set and NeuN_WGS_set (HiSeq 2500), as shown in Figure S2. The parameters were fine-tuned to optimize the final depth and base-call quality for somatic mutation detection in each set, as described in Figure S3.

Detection of somatic mutation candidates

We ran MuTect-1.1.525 and Strelka-1.0.1426 on the BAM files after QC. MuTect and Strelka called somatic mutation candidates by comparing target and control sample BAM files. In the CL_WGS_set, we performed one type of analysis: comparing AL30_cortex as the target sample to AL30_liver as the control sample. In the NeuN_WGS_set, we performed four types of analyses: two analyses comparing Y8763_NeuN+ as the target sample to Y8763_NeuN- or Y8763_liver as the control samples, and two other analyses comparing Y8763_NeuN- as the target sample to Y8763_NeuN+ or Y8763_liver as the control samples. In the CC_WGS_set, we performed two types of analyses: one analysis comparing S6_cortex as the target sample to S6_cerebellum as the control sample, and another analysis comparing S6_cerebellum as the target sample to S6_cortex as the control sample. We selected MuTect and Strelka from the available somatic mutation detection software because a previous study reported their superior sensitivity, especially the superiority of MuTect, over other software.27 The MuTect developers estimated MuTect sensitivity as 74.3% of the permitted allele fraction upper limit as 15% to preclude false positives due to germline SNV. All the samples underwent Strelka analysis with default parameters. The default parameters in Strelka analysis assume that the control sample has no somatic mutations identical to the target sample. The candidate sites where the allele fraction in the target sample was less than the allele fraction in the control sample were also excluded.

Filtering of somatic mutation candidates

The human genome has genomic regions enriched with sequences homologous to other regions, where accurate alignment is difficult with the current short-read technology.26 Among the candidates identified with MuTect, we excluded those located in multiregion copies (MCR), which we defined as ‘track: RepeatMasker, Interrupted Repeats, Microsatellite, Segmental Dups (Segmental Duplication), Self Chain, Simple Repeats’ in ‘group: Repeats’ in the hg19 assembly in the UCSC Genome Browser29 to preclude misalignment due to homologous sequences (1.63 Gb in total). We also excluded candidates near (±10 bp) insertion or deletion (INDEL) sites (called by GATK-3.2-2), because accurate alignment around INDEL is also difficult. We excluded candidates in suspected multicopy regions (SMCR, Table S2), as discussed in File S1.

After the initial filtering, we manually selected the candidates by IGV-2.3.4030 visualization, and performed a TAS validation experiment on the CL_WGS_set to decide the filtering parameters. Based on the results of TAS, we set the filtering parameters as follows: average base quality of candidate (BQ) ≥ 21, coverage depth at candidate site (DP) ≥ 40 (≥50 in the CC_WGS_set), and second-highest score of UCSC BLAT using the sequence around candidate site (±150 bp) < 150. The top score of the genomic region around the candidate site is always 300. The threshold of <150 excluded the candidate sites that clearly had homologous sequence sites in other genomic regions. The DP threshold in the CC_WGS_set (HiSeq X) was set higher, because sequence quality was lower than that of other sets (HiSeq 2500; Fig. S2).

Reliable base-call supporting somatic mutation (RBSM), used to focus on reliable sequence reads containing somatic mutations, were defined as a somatic mutation candidate call with all the following features: (i) base quality ≥ 25; (ii) read mapQ ≥ 30; (iii) not within 10 bp of the edge of read; (iv) not having minor mismatch or INDEL of ±15 bp on the same read; (v) no XA (alternative alignment) tags; (vi) not having soft-clip on the same read; and (vii) not explained by INDEL realignment errors. Difficult sequencing context
(SeqContext) was defined as short tandem repeat sites or poly-A sites. We then selected as the candidates either high-confidence (HC) sites, defined by RBSM count ≥ 2 and SeqContext ≠ difficult, or low-confidence (LC) sites, defined by RBSM count = 1 and SeqContext ≠ difficult or by RBSM count ≥ 2 and SeqContext = difficult.

**TAS**

We designed PCR primers for the candidate sites, as well as primers for dbSNP sites, as sequencing controls. We confirmed that primers yielded single-banded PCR amplicons of the expected sizes, and then prepared sequencing libraries with two rounds of PCR from 5–10 ng of DNA for each site (Fig. S4). We sequenced the library by Illumina MiSeq, using v3 reagents with a theoretical depth of 200 000 at each site. The primers used in the TAS are listed in Tables S3 and S4.

The TAS data were stringently quality controlled by Trimmomatic-0.32. After the QC steps, we aligned the reads to GRCh37 + decoy by BWA-0.7.12 and selected mapQ ≥ 60 reads. We then counted the base-calls with base quality ≥ 20 and calculated the AAF with respect to total base-calls on the candidate sites. The candidates with AAF > 0.316% (assumed sequence error rate, which is equivalent to a base-call quality of 25) by TAS analysis were considered validated somatic mutations.

**Statistical analysis**

Pearson’s correlation, $\chi^2$, and Fischer’s exact tests were performed using R-3.2.3 (https://www.r-project.org/). The $P$-values and false discovery rates (FDR) in GO analysis were calculated in ToppGene. We analyzed the sensitivity and specificity of our filtering processes with ROCR-1.0-7 (http://rocr.bioinf.mpi-sb.mpg.de/).

**Availability of data and material**

The WGS and TAS data from the CL_WGS_set, NeuN_WGS_set, and CC_WGS_set are available from the corresponding authors upon reasonable requests.

**RESULTS**

**Sequence data and QC**

We acquired deep-sequenced WGS data from the brain samples from three men with no neuropsychiatric diseases (Table 1 and Table S1). Because sequence quality differed between sequencers (Fig. S2), we adjusted the QC parameters on each platform (Fig. S3). After QC, drastic reductions in average read depths of raw data were observed in the WGS dataset with HiSeq X (from 121.5 to 75.2). In contrast, slight reductions were observed in the WGS dataset sequenced on HiSeq 2500 with PCR-free library preparation (from 97.0 to 83.0; Table 1 and Fig. S2). We eliminated the base-calls with low base quality to preclude false positives from sequencing errors, assuming sequencing errors to be a major source of false positives.

**Detection and validation of somatic mutations in the brain**

We explored somatic mutations in the frontal cortex of a 68-year-old man using relaxed criteria, and then refined the filtering parameters from the TAS validation results for subsequent analysis. MuTect detected 21 312 putative somatic mutation candidates in the frontal cortex, using the liver as the control (CL_WGS_set; Table S5). Among the putative candidates, 99.1% were in the MCR and were thus excluded from further analysis (MCR filtering). Further exclusion of the candidates around INDEL sites (INDEL filtering) or on the SMCR (SMCR filtering) resulted in 125 putative candidates.

We then broadly selected 34 candidates for TAS validation by manually excluding the candidates with apparently poor sequence quality or poor alignment (for example, candidates with average base quality of < 20 or average mapping quality of < 30). Among the 34 candidates tested, we successfully validated six candidates by TAS (Table 2). By comparing the analytical parameters of WGS data between successfully TAS-validated and unvalidated candidates (Table S6), we defined the parameters for

© 2017 The Authors. Psychiatry and Clinical Neurosciences published by John Wiley & Sons Australia, Ltd on behalf of Japanese Society of Psychiatry and Neurology
additional filtering after MCR, INDEL, and SMCR filtering (Table S5, see Methods). The further filtering focused on removing candidates whose surrounding genomic regions showed local homologies to other genomic regions, introducing the QC of sequence reads containing somatic mutations, and considering sequence context, such as poly-A and short tandem repeat sites. Finally, we selected the HC and LC candidates. Reanalysis of the 125 putative candidates by the established pipeline resulted in six HC and two LC candidates (Table S5).

In the NeuN_WGS_set, we ran MuTect on prefrontal neuronal or non-neuronal WGS data from a 78-year-old man. Applying the described filtering to the putative candidates, we obtained 14 HC and seven LC candidates in neurons, and 12 HC and two LC candidates in non-neuronal cells from four types of analyses (two analyses comparing NeuN+ to NeuN- or liver, and two analyses comparing NeuN- to NeuN+ or liver; Table S5). Some sites overlapped and were detected multiple times. In total, we detected 16 HC and seven LC sites as somatic mutation candidates. Subsequent TAS validated 11 HC sites and one LC site (Table 2 and Table S7).

In the CC_WGS_set, we ran MuTect on the cortex or cerebellum WGS data from an 84-year-old man. After MCR, INDEL, and SMCR filtering, 1332 and 1329 putative candidates remained in the cortex and cerebellum, respectively (Table S8). These numbers were much higher than those in the CL_WGS_set and NeuN_WGS_set, probably because the sequence quality of the CC_WGS_set was worse than that of the CL_WGS_set and NeuN_WGS_set (Fig. S2). Accordingly, we further separated the candidates into two groups based on the AAF difference (dAAF) between the cortex and cerebellum. In the dAAF ≥ 5% group, we obtained 13 HC and 44 LC sites. We performed TAS validation on the 13 HC sites and randomly selected 20 of the 44 LC sites. TAS validated seven HC candidates but failed to validate any of the LC sites (Table S9). Post-hoc inspection revealed that all the validated sites had BQ ≥ 25 in the WGS data. Therefore, we modified the filtering parameter to BQ ≥ 25. We next applied this filtering to the dAAF < 5% group, and obtained 12 HC and six LC candidates (Table S8). Because no LC sites were validated in the dAAF ≥ 5% group, we performed TAS on the HC sites, and successfully validated six sites (Table S10). In total, we validated 13 HC candidates in the CC_WGS_set (Table 2).

Overall, we identified 31 somatic mutations in the brains of three subjects (Table 2).

Robustness of TAS validation

We performed TAS on standard samples simulating somatic mutations with AAF of 1, 2.5, 5, and 10%, using seven dbSNP sites. The theoretical and experimental AAF were highly correlated (Pearson’s \( r = 0.969 \), Fig. 1a). We also performed TAS on a cumulative total of 35 heterozygous and 23 homozygous SNP. The average experimental AAF on heterozygous and homozygous SNP sites was 49.61 ± 0.96% and 99.98 ± 0.01%, respectively. The consistency between the theoretical and experimental AAF and the small standard deviations indicated the validity of AAF calculated by TAS.

The correlation between AAF in WGS and TAS on 31 somatic mutation sites was moderately high (Pearson’s \( r = 0.696 \), Fig. 1b). To check the reproducibility of AAF calculated by TAS, we performed a second TAS experiment on the same 31 somatic mutation sites using independent PCR and independent library preparation with the same DNA. The correlation between AAF in the first and second TAS was high (Pearson’s \( r = 0.987 \), Fig. 1c), demonstrating the robustness of AAF calculated by TAS. One example of somatic mutation identified by WGS and TAS is illustrated in Figure 2a.

The counts of reference and alternate base-calls obtained by TAS in the cortex significantly differed from the base-call counts in the liver in the CL_WGS_set by Fischer’s exact test (\( P \)-value < 2.2 × 10^{-16}). A significant difference in the base-call counts between target and control samples was also observed in the NeuN_WGS_set and the CC_WGS_set, except for chr22:45393311 G > A in NeuN_WGS_set (Table S11). The TAS-calculated dAAF (AAF differences) between target and control samples (NeuN vs liver in the NeuN_WGS_set) were generally larger than the dAAF between the first and second experiment at the same sites (Fig. S5). These results demonstrated that most of the AAF were not likely to be derived from systematic errors in sequencing.

Twenty-six of 31 TAS validated mutations were not found in either the large-scale genomic database of 1000 Genomes Project Phase 3\(^6\) or the Integrative Japanese Genome Variation Database\(^37,38\) which is a comprehensive variation database of 3554 Japanese genomes (Table S12). Six of 31 mutations are annotated in dbSNP150 (https://www.
Table 2. Somatic mutations validated by TAS in brain samples

| Chr | Position | Ref | Alt | Allele fraction (%) by TAS (WGS) | Gene | SO |
|-----|----------|-----|-----|----------------------------------|------|----|
|     |          |     |     | CL_WGS_set AL30_cortex AL30_liver |      |    |
| 13  | 72252357 | G   | A   | 14.038 (13.3) 10.317 (5.9)     | DACH1 | Sequence_feature |
| 1   | 211248507| G   | A   | 9.657 (7.7) 11.277 (5.9)       | KCNH1 | Sequence_feature |
| 1   | 47016199 | G   | A   | 3.050 (8.3) 4.574 (0)          | KCN   | Sequence_feature |
| 5   | 137203810| G   | A   | 1.897 (4.9) 0.500 (0)          | MYOT  | Intron_variant  |
| 6   | 164440297| G   | A   | 1.710 (4.1) 0.028 (0)          | MYOT  | Intergenic_region |
| 4   | 117374794| C   | T   | 0.736 (5.9) 0.028 (0)          |       | Intergenic_region |
|     |          |     |     | NeuN_WGS_set Y8763_NeuN+ Y8763_NeuN- Y8763_liver |      |    |
| 2   | 80253532 | C   | T   | 10.112 (11.1) 7.417 (7.8)      |       |    |
| 21  | 23955109 | T   | C   | 7.052 (8.1) 7.484 (7.8)        |       |    |
| 7   | 14747647 | G   | C   | 8.260 (5.5) 7.072 (5.0)        |       |    |
| 15  | 99259587 | G   | A   | 8.328 (14.4) 6.578 (4.5)       |       |    |
| 2   | 206520964| T   | G   | 8.261 (6.5) 6.786 (4.5)        |       |    |
| 3   | 192234799| C   | T   | 8.041 (5.3) 7.403 (7.3)        |       |    |
| 22  | 45393311 | G   | A   | 7.062 (6.8) 7.280 (6.7)        | RP4-753M9.1 | Downstream_gene_variant |
| 18  | 65440162 | G   | A   | 7.066 (8.9) 6.649 (4.9)        | RP11-638L3.1 | Intron_variant |
| 20  | 45395811 | G   | T   | 5.484 (7.6) 4.517 (4.7)        | IGF1R | Sequence_feature |
| 22  | 41257815 | G   | A   | 2.718 (3.7) 4.133 (8.4)        | DNAJB7 | Missense_variant (p.Arg62Trp) |
| 20  | 44667433 | G   | A   | 1.265 (4.7) 1.461 (0.9)        | SL12A5 | Upstream_gene_variant |
| 4   | 138417312| G   | A   | 1.592 (4.9) 1.169 (2.2)        |       |    |
|     |          |     |     | CC_WGS_set S6_cortex S6_cerebellum |      |    |
| 8   | 109128812| A   | T   | 10.276 (5.4) 13.863 (10.8)     | COL2A1 | Sequence_feature |
| 12  | 48380047 | C   | G   | 9.068 (6.6) 13.120 (5.4)       |       |    |
| 18  | 66437389 | C   | T   | 9.891 (7.6) 12.954 (11.8)      |       |    |
| 1   | 47281838 | G   | C   | 12.376 (14.5) 9.528 (7.0)      | CYP4B1 | Upstream_gene_variant |
| 1   | 239023569| C   | G   | 8.648 (6.9) 12.313 (17.7)      |       |    |
| 12  | 43522425 | C   | A   | 3.890 (5.1) 6.360 (8.1)        |       |    |
| 2   | 137814327| C   | T   | 5.419 (4.0) 6.356 (8.9)        | THSD7B | Synonymous_variant (p.Cys159Cys) |
| 1   | 240935724| C   | T   | 5.325 (6.3) 5.875 (4.3)        | PRKRIPR8 | Upstream_gene_variant |
| 2   | 202878979| G   | T   | 4.857 (2.0) 5.422 (8.9)        |       |    |
| 4   | 126987958| G   | A   | 3.609 (4.0) 5.273 (7.4)        | RP11-318L4.1 | Upstream_gene_variant |
| 18  | 76867501 | C   | T   | 2.771 (6.9) 1.383 (0)          | ATP9B | Sequence_feature |
| 2   | 102612321| C   | T   | 0.022 (0) 1.772 (9.2)          | IL1R2 | Upstream_gene_variant |
| 3   | 29527382 | C   | T   | 1.510 (4.8) 0.160 (0)          | RBMS3 | Sequence_feature |

The list is sorted by the order of alternate allele fraction in each set. The genomic positions are based on the reference genome GRCh37. The amino acid change described according to the Human Genome Variation society. Non-protein-coding sequences (e.g., non-coding RNA) are indicated by parentheses. See the details in Table S12.

SO, sequence ontology categories defined by the Sequence Ontology project (http://www.sequenceontology.org/); TAS, target amplicon sequence.
TAS-validated mutations (one overlap) are low (0.0001–0.0072; Table S12).

Characteristics of somatic mutations in the brain

Among the 31 TAS-validated somatic mutations, five sites were near heterozygous SNP. We confirmed that all five somatic mutations occurred in only one allele in the WGS and TAS data (Fig. 2b, Table S13), supporting that they were somatically acquired mutations, not sequencing artifacts.

Consistent with a previous study,1,39,40 21 of 31 brain somatic mutations (67.7%) were C > T transitions, which highly deviates from the expected probability of 16.7% (Pearson’s χ²-test P-value = 2.34 × 10⁻¹⁴, Table 2). Eleven of 21 C > T transitions (52.4%) were in CpG dinucleotides, showing the enrichment of brain somatic mutations in CpG dinucleotides. The enrichment implicates the involvement of cytosine methylation in the mechanism of somatic mutation, as reported in Lodato et al.1

Somatic mutations were detected in 15 genes and six transcripts. Among the somatic mutations found in the genes, only one had an effect on the amino acid sequence (chr22:41257815 G > A, missense variant in DNAJB7 [DnaJ Heat Shock Protein Family Member B7]). This missense variant was inferred to be damaging to protein function by SIFT, PRO-VEAN, and MutationTaster. GO analysis revealed that somatic mutations were enriched in neuron-expressed gene categories (Table 3, Table S14). Seven GO categories, most of which were neuron- and axon-related, were statistically significant after Bonferroni correction (Bonferroni corrected P-value < 0.05, Table 3).

The receiver–operator curve (ROC) of the HC sites to BQ thresholds in the brain samples showed that the sensitivity was 100% with BQ threshold ≥ 25 (specificity = 23.5%, Fig. S6). The difference between the true positive and false positive rates was the highest with BQ threshold ≥ BQ 28 (sensitivity: 89.7%, specificity: 82.4%). The ROC of the HC sites to DP thresholds showed that DP is not an appropriate selection parameter in HC candidates.

Detection of somatic mutation by Strelka

Strelka and further filtering procedures yielded HC and LC candidates (Tables S5 and S8). All TAS-validated somatic mutations detected by Strelka were included among those detected by MuTect (Tables S6, S7, and S9).

DISCUSSION

We successfully identified 31 somatic mutation sites in the human postmortem brains from three men with no history of neuropsychiatric diseases, and confirmed them by independent TAS analysis. The calibration experiment, small standard deviations of dbSNP sites, and the reproducibility of AAF by TAS indicated the reliability of our results.

In this study, we ran MuTect with a permitted AAF of 15% in control tissues. Potential somatic mutations with an AAF over 15% in control tissues could not be detected in our analysis. Increasing the permitted AAF in control tissues would improve sensitivity; however, this would also increase the false positives by detecting germline SNV. Almost all initial candidates detected by MuTect were removed by excluding candidates located in the MCR. Somatic mutations occurring in such regions would be missed in our analysis. Such sites are difficult to analyze with other technologies, including droplet digital PCR, Sanger sequencing, and pyrosequencing.

Several HC candidates were not validated by TAS, even after applying the filtering procedure. The ROC to BQ thresholds in HC candidates indicated that BQ is one of the major sources of false positives. Post-hoc analysis revealed that setting the BQ threshold ≥ 25 decreased the number of false positives without decreasing the number of true positives, and that setting the BQ threshold ≥ 28 increased the specificity with the highest difference between the true positive rate and the false positive rate. DP threshold had little effect on sensitivity and specificity after selecting HC candidates. The false positives could have arisen for several other reasons: possible sample contamination, PCR errors during library preparation, sequencing errors, and false alignment of sequence reads.

Sample contamination during sequencing is reported to be one reason for false positives.41,42 In our study, efforts were made to minimize sample contamination in the CL_WGS_set and the NeuN_WGS_set by introducing independent sample indexing. The data obtained by HiSeq X (CC_WGS_set) were, in principle, free from sample contamination. Slight sample contamination during sample collection and gDNA extraction cannot be
Figure 1. Validity of target amplicon sequencing (TAS) of somatic mutations in the brain. (a) Theoretical and experimental alternate allele fractions (AAF) of the simulated somatic single nucleotide variants (SNV) calculated by TAS. The theoretical (x-axis) and experimental (y-axis) AAF (%) of the simulated somatic SNV showed high correlation (Pearson’s r = 0.969, P-value < 2.2 × 10^{-16}). Simulated somatic SNV were prepared by mixing gDNA derived from the peripheral blood samples obtained from two Japanese men, with different genotypes (i.e., one homozygous, and the other heterozygous) at seven dbSNP sites (rs1125662, rs11505418, rs11249930, rs10904247, rs3848097, rs4902177, and rs10207). The regression line was y = 1.39x + 0.62 and the standard error of residuals was 1.27 (%). The standard error of 1.27 (%) corresponded with the 95% confidence interval of ±2.49% in TAS quantification. (b) Correlation between the AAF calculated by whole-genome sequencing (WGS) and TAS. The AAF by WGS and TAS were moderately correlated (Pearson’s r = 0.696, P-value = 5.87 × 10^{-12}). (c) Correlation between the AAF in the initial and the repeated TAS experiments. The AAF in the first and the second experiments were highly correlated (Pearson’s r = 0.987, P-value < 2.2 × 10^{-16}). (b,c) (●) CL_WGS_set. (●) NeuN_WGS_set. (●) CC_WGS_set.
totally excluded in five mutations found in the large-scale variant database, although the allele frequency was low (Table S12). Twenty-six of 31 validated mutations were not found in the large-scale variant database (1000 Genomes Project and Integrative Japanese Genome Variation). Although these databases have not captured all human genome variations, the absence of these 26 variants indicates that the possibility of sample contamination was low.

The fact that the CC_WGS_set (PCR+) had more unvalidated candidates than the CL_WGS_set or the NeuN_WGS_set (PCR-) suggests that PCR errors during library preparation were one of the main factors contributing to false positives. Oxidative DNA damage is reported to result in erroneous calls for somatic mutations, which are G > T or C > A deriving from 8-oxoguanine and PCR procedures during library preparation.43 None of the TAS-validated somatic mutations could be such errors, because
these mutations were confirmed from the DNA template prior to library preparation. In addition, the CL_WGS_set and the NeuN_WGS_set were free from such errors with no PCR procedures in library preparation. However, we cannot exclude such errors in the CC_WGS_set that included PCR procedures. Despite multilayered filtering, some unvalidated candidates of G > T or C > A might be the result of oxidative DNA damage.

If the Phred-scale base quality is not always accurate, sequencing errors can induce false positives. Illumina HiSeq and MiSeq are different sequencing machines, but adopt similar chemical and optical systems. TAS-validated mutations could be the results of systematic errors common to HiSeq and MiSeq. If the systematic sequencing errors induce false positives, control samples should have the same false positives at similar AAF. We observed significant difference in AAF between sample tissues or cell types except for chr22:45393311 (Table S11, Fig. S5), indicating the low possibility of false positives due to sequencing errors. The alternate allele at chr22:45393311 could be false positives from systematic sequencing errors. False alignment of homologous sequences is a common problem of current short-read sequencing technologies, and cannot be completely ruled out in our study. However, this was less likely than other reasons, because we excluded genomic regions having homologous sequences in the reference genome.

Although the theoretical and experimental AAF in simulation samples showed high correlation, the experimental AAF was 1.39 times greater than the theoretical AAF. This is likely to be due to a technical problem in the manual procedures that one sample was mixed with a larger amount than the theoretical volume when making somatic mutation simulation samples. The variance of experimental AAF may derive from PCR amplification bias. The DNA quantity used in this calibration experiment was 7 ng (corresponding to approximate 1000 cells). We performed TAS from 5–10 ng of DNA, as the amount of brain tissue was limited. Starting with a small amount of DNA is likely to cause increased PCR amplification bias. Using a larger quantity of DNA would increase the precision of TAS.

After the technical discussion mentioned above, we believe that PCR-free sample preparation is effective in reducing the number of false positives and the threshold of BQ ≥ 25 (BQ ≥ 28 in weighing on specificity) is suitable for candidate filtering in detecting somatic mutations from bulk brain tissues. Considering that the sensitivity to all the potential somatic mutations depends on the depth of coverage, a higher depth of coverage is desirable. In our study, somatic mutation detection required control tissues from the same individual. Bioinformatics analysis with no control samples would be useful in a situation where control samples are not available.

Single-cell analysis is an excellent way to study somatic mutations, especially for somatic mutations occurring at later developmental stages. However, single-cell analysis has technical problems of false positives due to the spontaneous deamination of C > U in cell lysis during single-cell preparation.

---

**Table 3. Results of GO analysis of validated somatic mutations in brain samples**

| Category | ID    | Name                          | Bonferroni corrected P-value | Hit count in query list | Hit count in genome | Hit in query list |
|----------|-------|-------------------------------|-------------------------------|------------------------|---------------------|------------------|
| CC       | GO:0032589 | Neuron projection membrane | 5.88 × 10⁻⁴                           | 3                      | 49                  | KCNH1,SLC12A5,CNTNAP2 |
| CC       | GO:0043025 | Neuronal cell body            | 5.74 × 10⁻³                           | 5                      | 620                 | KNCN,KCNH1,IGF1R,SLC12A5,CNTNAP2 |
| CC       | GO:0030673 | Axolemma                      | 8.83 × 10⁻³                           | 2                      | 19                  | KCNH1,CNTNAP2     |
| CC       | GO:0044297 | Cell body                     | 9.88 × 10⁻³                           | 5                      | 695                 | KNCN,KCNH1,IGF1R,SLC12A5,CNTNAP2 |
| CC       | GO:0031256 | Leading edge membrane         | 1.68 × 10⁻²                           | 3                      | 150                 | KCNH1,SLC12A5,CNTNAP2 |
| MF       | GO:0019838 | Growth factor binding         | 2.02 × 10⁻²                           | 3                      | 142                 | COL2A1,IL1R2,IGF1R |
| CC       | GO:0036477 | Somatodendritic compartment   | 2.85 × 10⁻²                           | 5                      | 871                 | KNCN,KCNH1,IGF1R,SLC12A5,CNTNAP2 |

See the details in Table S14. CC, cellular component; GO, gene ontology; MF, molecular function.
and whole-genome amplification (WGA) errors. Our analysis with WGS on bulk tissues followed by TAS validation provided a complementary approach for the detection of somatic mutations occurring in early development with several advantages: (i) no false positives from cell lysis or WGA/PCR errors (when using PCR-free library preparation); (ii) estimating the allele fractions in the target tissues; and (iii) availability in clinically oriented research using a large number of samples and tissues. Single-cell analysis, which requires huge resources of sequencing and sample preparation, is difficult to scale to a large number of individuals. WGS on bulk tissues has an advantage in detecting reliable somatic mutations occurring early in development, which should have stronger effects on individual phenotype than those occurring later in development.

The tissue distribution of somatic mutations and their AAF should reflect the timing of occurrence of mutations during development.47,48 We identified three somatic mutations (chr6:164440297, chr4:117374794, and chr4:138417312) that were present in the cortex but not in the liver, suggesting that they had occurred after the differentiation of the ectoderm and endoderm. Similarly, two somatic mutations were found in the cerebellum but not in the cortex (chr2:102612321) or vice versa (chr3:29527382), suggesting their occurrence after the development of the cortex and cerebellum. In contrast, 26 of 31 mutations did not show tissue-specificity, suggesting that most of the mutation events occurred at a very early developmental stage, possibly shared among various tissues. We did not detect brain cell-type-specific somatic mutations in this study. All 12 somatic mutations detected in the NeuN_WGS_set were found in both neuronal and non-neuronal cells with very similar AAF levels (Pearson’s r = 0.946, P-value = 3.41 × 10⁻⁶) and 11 of 12 were shared with the liver. These results suggest that somatic mutations caused by brain cell-type-specific events, such as activity-dependent insults to the neuronal genome, demethylation processes possibly involving base excision repair, or proliferation-associated mutations in non-neuronal cells, were difficult to detect in our study. Such mutations probably occurred at a later stage of development, and AAF were expected to be below the detection limit of 0.316% in this study.

The 31 brain somatic mutations detected in three men with no neuropsychiatric diseases were predicted not to have protein-disrupting effects except for one missense variant in DNAJB7, which is consistent with the fact that the samples were derived from men with no psychiatric diseases. However, the 31 somatic mutations were enriched in neuron-expressed genes, suggesting the vulnerability of neuronal genes to mutational events during development, consistent with Lodato et al. This is likely to be attributable to the longer sequences of neuron-expressed genes compared to others.52 If the vulnerability of neuron-expressed genes was generalized to the general population, the mutations in neuron-expressed genes should have the possibility to occur as protein-disrupting mutations during neurodevelopment in some cases. Mutations resulting in embryonic lethality or severe congenital diseases cannot exist in the germ-line genome, but they may exist as somatic mutations, possibly resulting in relatively less severe malfunctions, including some neuropsychiatric diseases. Somatic mutations could also modify brain functions and dysfunctions in addition to the individual’s germline genomic features. Rare and common germline variants additively contribute to autism spectrum disorder (ASD).53 Somatic mutations would also contribute additively to brain function and dysfunction. Despite large-scale genomic studies, the etiology of psychiatric disorders has not been fully elucidated. Somatic mutations are one of the candidate mechanisms expected to explain the remaining part of the etiology.

We detected somatic mutations with the AAF of 0.5–14.0%. Somatic mutations in the relevant genes in the affected brain from patients with brain malformation were reported with AAF of 1–13% in cortical dysplasia or 8–40% in hemimegalencephaly. These AAF are in parallel with the AAF of brain somatic mutations of 0.5–14.0% in our study, although our samples were derived from individuals with no history of neuropsychiatric diseases. When applied to brains from patients with neuropsychiatric diseases, our approach has the capability to detect somatic mutations with similar ranges of allele fractions, which could potentially contribute to individual neuropsychiatric phenotypes. Several studies detected somatic mutations (postzygotic mutations) in the known risk genes for ASD with the allele fractions ≥ 5% in blood cells from patients with ASD. The relevance of the genes and the large allele fractions indicate early occurrence in development, shared between the brain and the blood. The contribution of such somatic mutations to ASD was
estimated to be 3–5%. However, these studies did not show the existence of somatic mutations in the brain. Detecting somatic mutations with such estimated effects in the brain would be valuable in future studies on neuropsychiatric genetics. Our study lays the groundwork for future investigation of somatic mutations occurring early in development in human brain samples.

We identified 31 reliable somatic mutations by WGS of bulk brain tissues from three individuals with no neuropsychiatric diseases. Our approach to detect somatic mutations by conventional WGS of bulk brain tissues is applicable in detecting somatic mutations occurring early in development in clinically oriented research, complementing basic neuroscientific investigation with single-cell genomics. Further studies on postmortem brain samples from patients with neuropsychiatric diseases will clarify the physiological and pathophysiological roles of somatic mutations in the human brain.

ACKNOWLEDGMENTS

This research was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology (24116005 to T.K., 24116009 and 15H04891 to K.I, and 15K09801 to M.B). This research is partially supported by the Advanced Genome Research and Bioinformatics Study to Facilitate Medical Innovation (GRIFIN) from the Japan Agency for Medical Research and Development (AMED; T.K). Caucasian postmortem samples were donated by the Stanley Medical Research Institute, courtesy of Drs Michael B. Knable, Serge Weis, E. Fuller Torrey, Maree J. Webster, and Robert H. Yolken. We thank Keiko Tateno and Naomi Nakai-Inagaki for performing WGS. We also thank the Research Resources Center at the RIKEN BSI for the cell-sorting analysis.

DISCLOSURE STATEMENT

M. Nishioka, M.B., and K.I. belonged to the Department of Molecular Psychiatry at the University of Tokyo, which was an endowment department by Dainippon Sumitomo Pharma and Yoshitomiya-kin. T.K. received a grant from Takeda Pharmaceutical outside this work. The companies had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The other authors declare no conflict of interest in this study.

AUTHOR CONTRIBUTIONS

Author contributions are as follows: M. Nishioka, M.B., K.K., T.K., and K.I. designed the research procedures. T.I., W.U., E.H., S.M., and T.K. collected human postmortem brains. M.B. performed cell sorting. F.K., Y.S., Y.K., M. Nagasaki, and J.Y. performed whole genome sequencing. M. Nishioka and J.U. analyzed the whole genome and exome sequence data. M. Nishioka and M.B. performed PCR and target amplicon sequencing. M. Nishioka performed the subsequent analysis. M.B., T.K., and K.I. checked the results. M. Nishioka, M.B., T.K., and K.I. prepared the manuscript.

REFERENCES

1. Lodato MA, Woodworth MB, Lee S et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. Science 2015; 350: 94–98.
2. McConnell MJ, Lindberg MR, Brennand KJ et al. Mosaic copy number variation in human neurons. Science 2013; 342: 632–637.
3. Evrony GD, Cai X, Lee E et al. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell 2012; 151: 483–496.
4. Baillie JK, Barnett MW, Upton KR et al. Somatic retrotransposition alters the genetic landscape of the human brain. Nature 2011; 479: 534–537.
5. Upton KR, Gerhardt DJ, Jesuadian JS et al. Ubiquitous L1 mosaicism in hippocampal neurons. Cell 2015; 161: 228–239.
6. Evrony GD, Lee E, Mehta BK et al. Cell lineage analysis in human brain using endogenous retroelements. Neuron 2015; 85: 49–59.
7. Poduri A, Evrony GD, Cai X et al. Somatic activation of AKT3 causes hemispheric developmental brain malformations. Neuron 2012; 74: 41–48.
8. Lee JH, Huynh M, Silhavy JL et al. De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. Nat. Genet. 2012; 44: 941–945.
9. Lim JS, Kim WI, Kang HC et al. Brain somatic mutations in MTOR cause focal cortical dysplasia type II leading to intractable epilepsy. Nat. Med. 2015; 21: 395–400.
10. Nakashima M, Saito H, Takei N et al. Somatic mutations in the MTOR gene cause focal cortical dysplasia type IIb. Ann. Neurol. 2015; 78: 375–386.
11. Jansen LA, Mirzaa GM, Ishak GE et al. PI3K/AKT pathway mutations cause a spectrum of brain malformations from megalencephaly to focal cortical dysplasia. Brain 2015; 138: 1613–1628.
12. Mirzaa GM, Campbell CD, Solovieff N et al. Association of MTOR mutations with developmental brain disorders,
including megalencephaly, focal cortical dysplasia, and pigmentary mosaicism. *JAMA Neurol.* 2016; 73: 836–845.

13. Muotri AR, Marchetto MC, Coufal NG et al. L1 retrotransposition in neurons is modulated by MeCP2. *Nature* 2010; 468: 443–446.

14. Coufal NG, Garcia-Perez JL, Peng GE et al. Ataxia telangiectasia mutated (ATM) modulates long interspersed element-1 (L1) retrotransposition in human neural stem cells. *Proc. Natl. Acad. Sci. U. S. A.* 2011; 108: 20382–20387.

15. Bundo M, Toyoshima M, Okada Y et al. Increased L1 retrotransposition in the neuronal genome in schizophrenia. *Neuron* 2014; 81: 306–313.

16. D’Gama AM, Pochareddy S, Li M et al. Targeted DNA sequencing from autism spectrum disorder brains implicates multiple genetic mechanisms. *Neuron* 2015; 88: 910–917.

17. Sakai M, Watanabe Y, Someya T et al. Assessment of copy number variations in the brain genome of schizophrenia patients. *Mol. Cytogenet.* 2015; 8: 46.

18. Kim J, Shin JY, Kim JJ et al. Somatic deletions implicated in functional diversity of brain cells of individuals with schizophrenia and unaffected controls. *Sci. Rep.* 2014; 4: 3807.

19. Sala Frigerio C, Lau P, Troakes C et al. On the identification of low allele frequency mosaic mutations in the brains of Alzheimer’s disease patients. *Alzheimers Dement.* 2015; 11: 1265–1276.

20. Iwamoto K, Bundo M, Ueda J et al. Neurons show distinctive DNA methylation profile and higher interindividual variations compared with non-neurons. *Genome Res.* 2011; 21: 688–696.

21. Bolger AM, Lohse M, Usadel B. Trimmmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 2014; 30: 2114–2120.

22. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; 25: 1754–1760.

23. McKenna A, Hanna M, Banks E et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010; 20: 1293–1303.

24. Li H, Handsaker B, Wysoker A et al. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009; 25: 2078–2079.

25. Cibulskis K, Lawrence MS, Carter SL et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* 2013; 31: 213–219.

26. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RR. Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* 2012; 28: 1811–1817.

27. Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics* 2014; 15: 244.

28. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nat. Rev. Genet.* 2011; 13: 36–46.

29. Kent WJ, Sugnet CW, Furey TS et al. The human genome browser at UCSC. * Genome Res.* 2002; 12: 996–1006.

30. Robinson JT, Thorvaldsdottir H, Winckler W et al. Integrative genomics viewer. *Nat. Biotechnol.* 2011; 29: 24–26.

31. Cingolani P, Platts A, Wang L et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 2012; 6: 80–92.

32. Chen J, Bardes EE, Aronow BJ, Jegga AG. TopGen suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* 2009; 37: W305–W311.

33. 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.

34. 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.

35. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: Mutation prediction for the deep-sequencing age. *Nat. Methods* 2014; 11: 361–362.

36. 1000 Genomes Project Consortium, Auton A, Brooks LD et al. A global reference for human genetic variation. *Nature* 2015; 526: 68–74.

37. Yamaguchi-Kabata Y, Nariai N, Kawai Y et al. iJGVD: An integrative Japanese genome variation database based on whole-genome sequencing. *Hum. Genome Var.* 2015; 2: 15050.

38. Nagasaki M, Yasuda J, Katsukawa F et al. Rare variant discovery by deep whole-genome sequencing of 1070 Japanese individuals. *Nat. Commun.* 2015; 6: 8018.

39. Lynch M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. U. S. A.* 2010; 107: 961–968.

40. Rahbari R, Wuster A, Lindsay SJ et al. Timing, rates and spectra of human germline mutation. *Nat. Genet.* 2016; 48: 126–133.

41. Cibulskis K, McKenna A, Fennell T, Banks E, DePristo M, Getz G. ContEst: Estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* 2011; 27: 2601–2602.

42. Jun G, Flickinger M, Hetrick KN et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* 2012; 91: 839–848.

43. Chen L, Liu P, Evans TC Jr, Ettwiller LM. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science* 2017; 355: 752–756.

44. Huang AY, Xu X, Ye AY et al. Postzygotic single-nucleotide mosaicisms in whole-genome sequences of clinically unremarkable individuals. *Cell Res.* 2014; 24: 1311–1327.

© 2017 The Authors. Psychiatry and Clinical Neurosciences published by John Wiley & Sons Australia, Ltd on behalf of Japanese Society of Psychiatry and Neurology
SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article at the publisher’s web-site:

File S1. Detailed protocols in our experiments are described.

Figure S1. Workflow for the detection of somatic single-nucleotide variants (SNVs) from whole genome sequencing (WGS) data.

Figure S2. Representative examples of WGS data quality control.

Figure S3. Software parameters in our analysis pipeline.

Figure S4. Workflow of target amplicon sequence (TAS) validation.

Figure S5. AAF differences at 31 TAS-validated sites.

Figure S6. Receiver–operator curve (ROC) according to the average base quality or depth of HC somatic SNV candidates.

Table S1. Sample data in detail.

Table S2. Suspected multicopy region (SMCR).

Table S3. Common primers used in the library preparation (TAIL-PCR for MiSeq TAS).

Table S4. Primers used in TAS validation experiments for each genomic site.

Table S5. Summary of parameters employed for detecting and selecting somatic mutation candidates in CL_WGS_set and NeuN_WGS_set.

Table S6. Alternate allele fractions of somatic mutation candidates in CL_WGS_set.

Table S7. Alternate allele fractions of somatic mutation candidates in NeuN_WGS_set.

Table S8. Summary of parameters employed for detecting and selecting somatic mutation candidates in CC_WGS_set.

Table S9. Alternate allele fractions of somatic mutation candidates in CC_WGS_set (1st).

Table S10. Alternate allele fractions of somatic mutation candidates in CC_WGS_set (2nd).

Table S11. Fisher’s exact test comparing reference and alternate base-calls obtained by target amplicon sequence (TAS).

Table S12. TAS-validated somatic mutations and allele frequency in database.

Table S13. Validated somatic mutations and nearby SNP sites.

Table S14. Gene ontology analysis of the somatic mutations in brain samples.