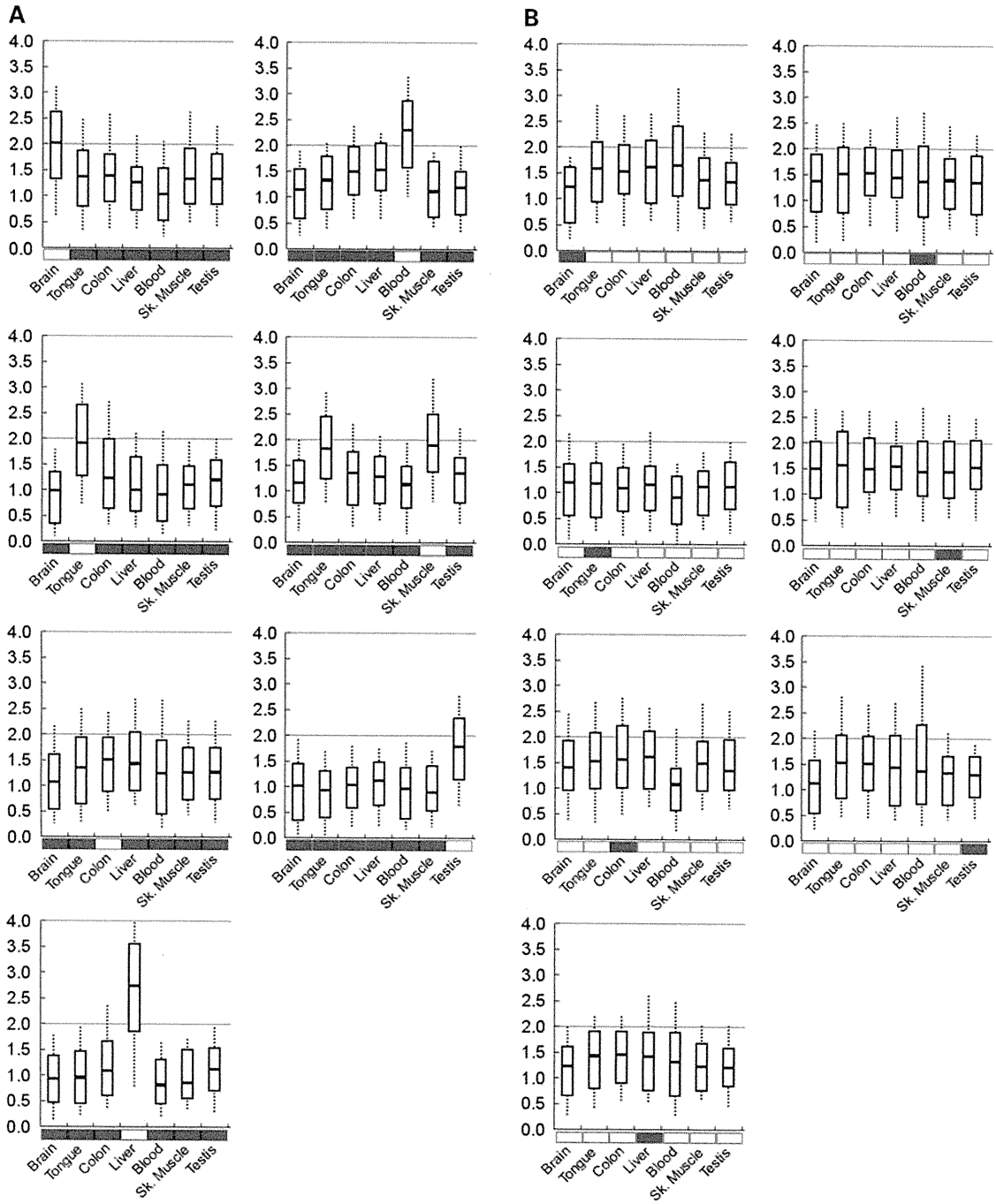

Figure 3. The gene expression level of tissue-specific differentially methylated genes. Shown box plots (from 25th percentile to the 75th percentile with heavy lines at the median) represent average gene expression levels (the log scale of the GeneChip score) of tissue-specific hypomethylated genes (**A**) and tissue-specific hypermethylated genes (**B**), for each tissue. The dotted lines extend above and below the box to show the first and ninth deciles. Black and white boxes below the bar graphs represent hypermethylation and hypomethylation of the given tissue, respectively.

this concept was true for some validated examples, it cannot adequately explain the global control of gene expression. In fact, consistent with the previous studies (6,10), we observed that most CpG island promoters are invariably unmethylated among normal tissues. In contrast with tissue-specific hypermethylation in CpG island promoters, tissue-specific hypomethylation in CpG-poor promoters has been underestimated so far and is significantly associated with the tissue phenotype.

These observations raise a new question about the molecular mechanism of tissue-specific hypomethylation established during terminal differentiation. Promoter demethylation in the differentiated cells is an old concept (29,30), but it has been forgotten while mammalian DNA demethylase was yet to be discovered. Now, two types of mechanisms for DNA demethylation, namely active demethylation and passive demethylation, are widely accepted for mammals (31,32).

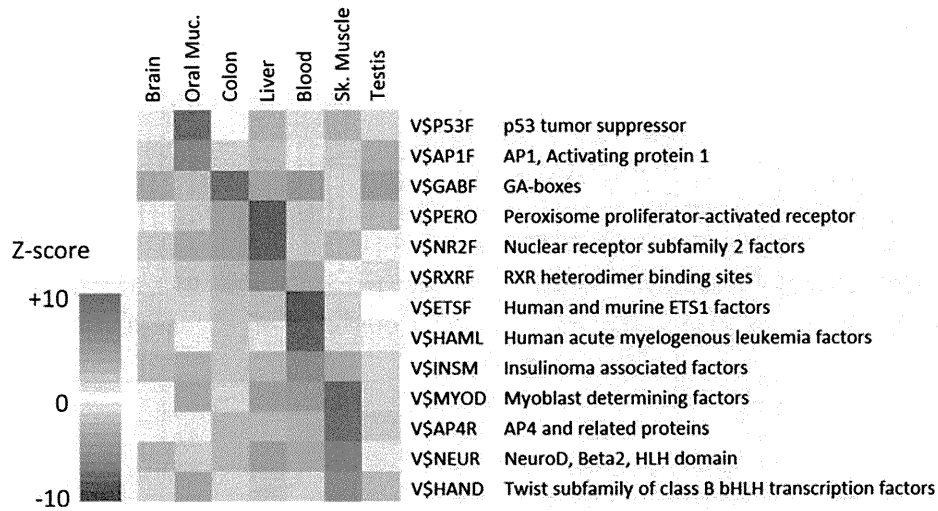

Figure 4. Enrichment of transcription factor recognition motifs in the tissue-specific hypomethylated regions. Each row represents a cis-regulatory module family with significant over-representation relative to a random set of mammalian promoters (Z-score > 8.0). Each column represents a tissue type. Four tissues (oral mucosa, liver, blood and skeletal muscle) show some specific enrichment of their master regulators binding motifs, respectively.

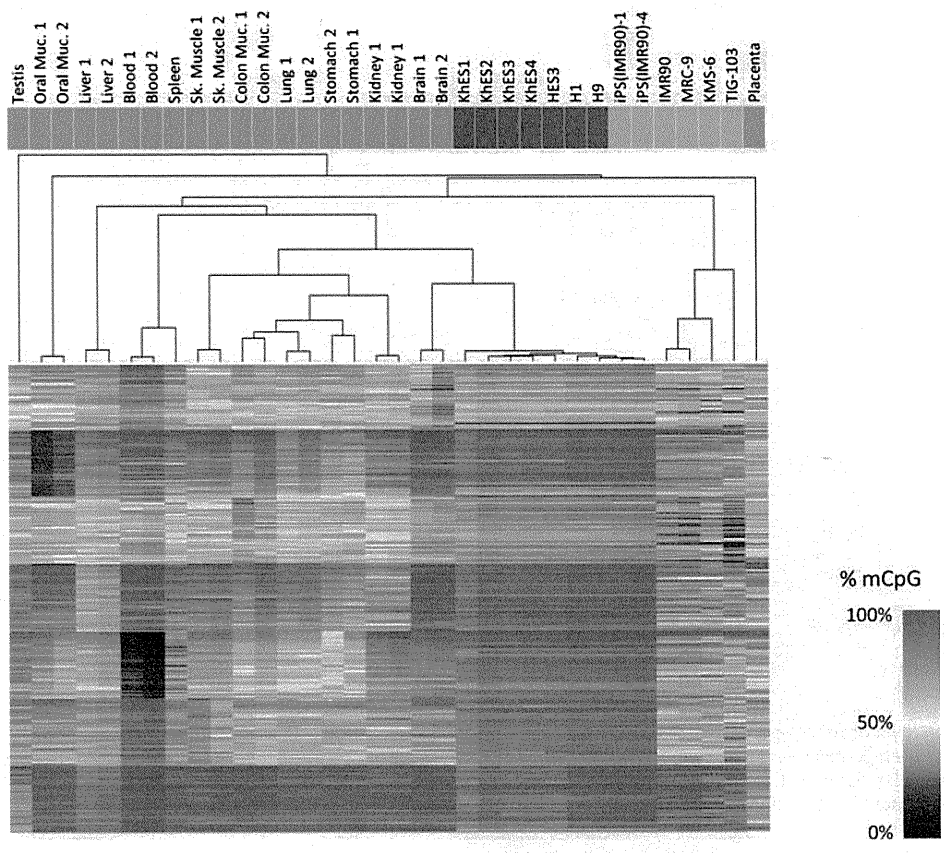

Figure 5. Hierarchical clustering analysis of human somatic tissues and normal cells. The dendrogram in the upper panel was obtained on the basis of the representative gene sets of tissue-specific hypomethylation using average linkage correlation. Each row represents a CpG locus (250 tissue-specific hypomethylation for each) and each column represents a sample. The colored boxes above the dendrogram indicate the nature of the samples; human somatic tissues (blue), human ES cells (red), human iPS cells (orange) and human primary fibroblast (green). The color scale bar at the right side shows the percentage of the methylation level (0–100%).

Active demethylation is observed in the paternal genome of an embryo during the first few days (33,34). In this process, demethylation occurs globally except for the limited foci such as imprinting control regions and centromeric and pericentromeric heterochromatin (35). Although recent reports suggested the ten-eleven translocation (TET) family proteins,

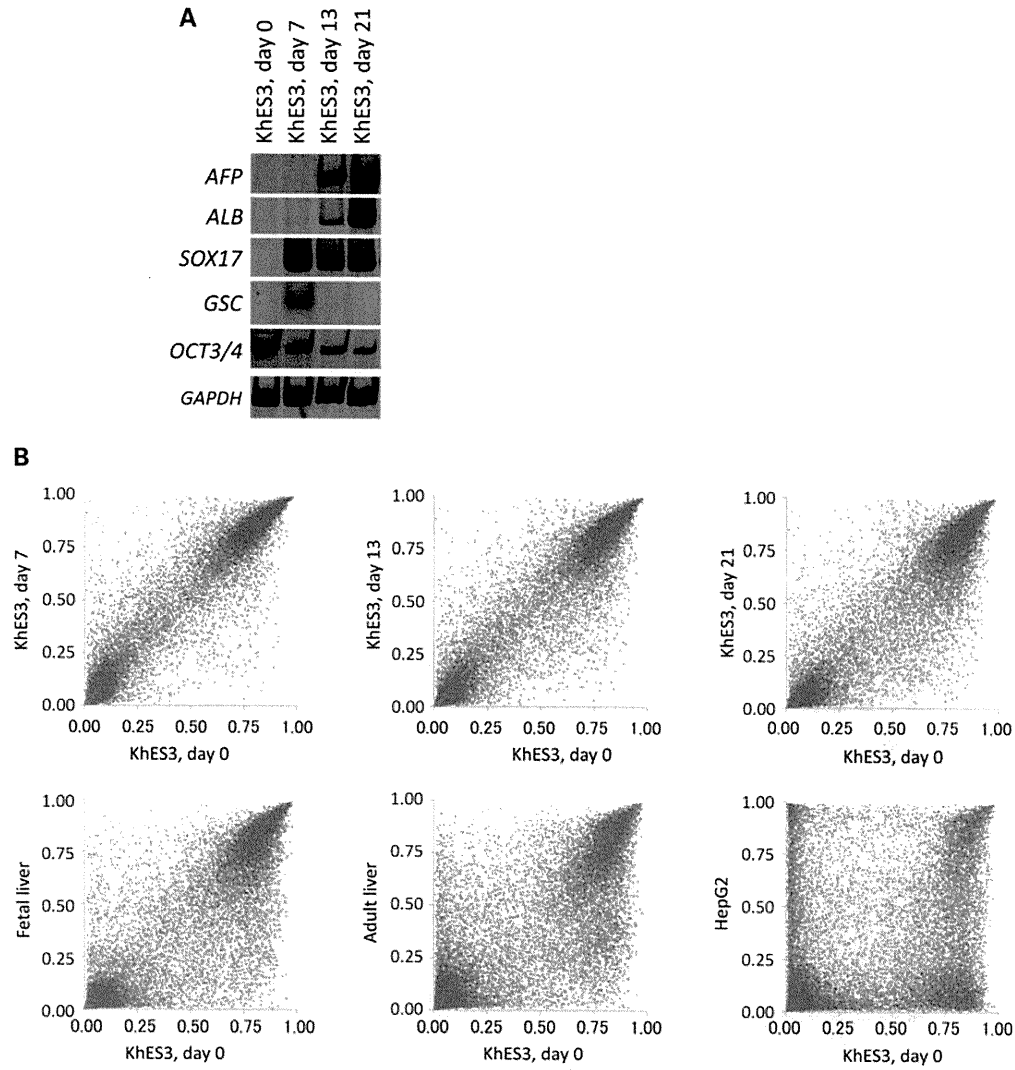

Figure 6. *In vitro* demethylation of liver-specific hypomethylated genes during hepatic differentiation (**A**) RT–PCR analysis of endodermal and hepatic differentiation markers in ES cells and differentiated cells (**B**) Global comparison among undifferentiated ES cells and differentiated cells, human fetal liver, adult liver and HepG2 cells. Liver-specific hypomethylated genes are indicated as red dots, overlapping with the others (blue). (**C**) Examples of gradually demethylated genes during *in vitro* differentiation into hepatic lineages. The bar graphs show the methylation levels of the genes that show gradual demethylation (~20% decrease) in day 21 of *in vitro* differentiation. (**D**) The liver-specific hypomethylated region around the *APOA1* gene. In the upper panel of the UCSC browser, nine black boxes indicate the position of PCR amplicons in a MassARRAY analysis. The methylation levels around the *APOA1* gene among ES cells and adult liver tissues are shown in the lower panel.

TET1, TET2 and TET3, are candidate proteins responsible for the erasure process through an oxidative demethylation pathway (32,36), further investigations are needed. The unexpected dynamics of DNA methylation during cellular differentiation might give us an important clue to elucidate the mechanism of cell fate determination during embryogenesis.

An alternative explanation for the tissue-specific demethylation seen in CpG-poor promoters is passive demethylation, which is usually observed in asymmetric cell division or highly proliferating cells like cancer cells. Inhibiting maintenance of cytosine methylation of the template strand could result in dilution of methylation in differentiated daughter cells. According to this scenario, transcription factor-related inhibition of DNA methyltransferase at the timing of cell division might be necessary because the developmental hypomethylation we observed here occurs not in a genome-wide manner but in a regional manner. Indeed, the enrichment of transcription factor-binding motifs is seen at the demethylated regions in a tissue-specific manner. Recently, it was shown that mitotically retained transcription factors are associated with the asymmetric cell division in some contexts (37,38). If sustained binding of transcription factors inhibits propagation of DNA methylation into the newly synthesized strand, transcription factor-driven demethylation will be inherited in proliferating cells. In our study, we examined *in vitro* differentiation in a series of promoters and found that a wave of demethylation develops from the TSS of *APOA1* and *ITIH3* promoters. Once the binding of transcription factors at demethylated regions induces gene expression in the tissue progenitor cells, sustained induction in response to appropriate extrinsic stimuli may result in loss of propagation of DNA methylation marks in the promoter regions for

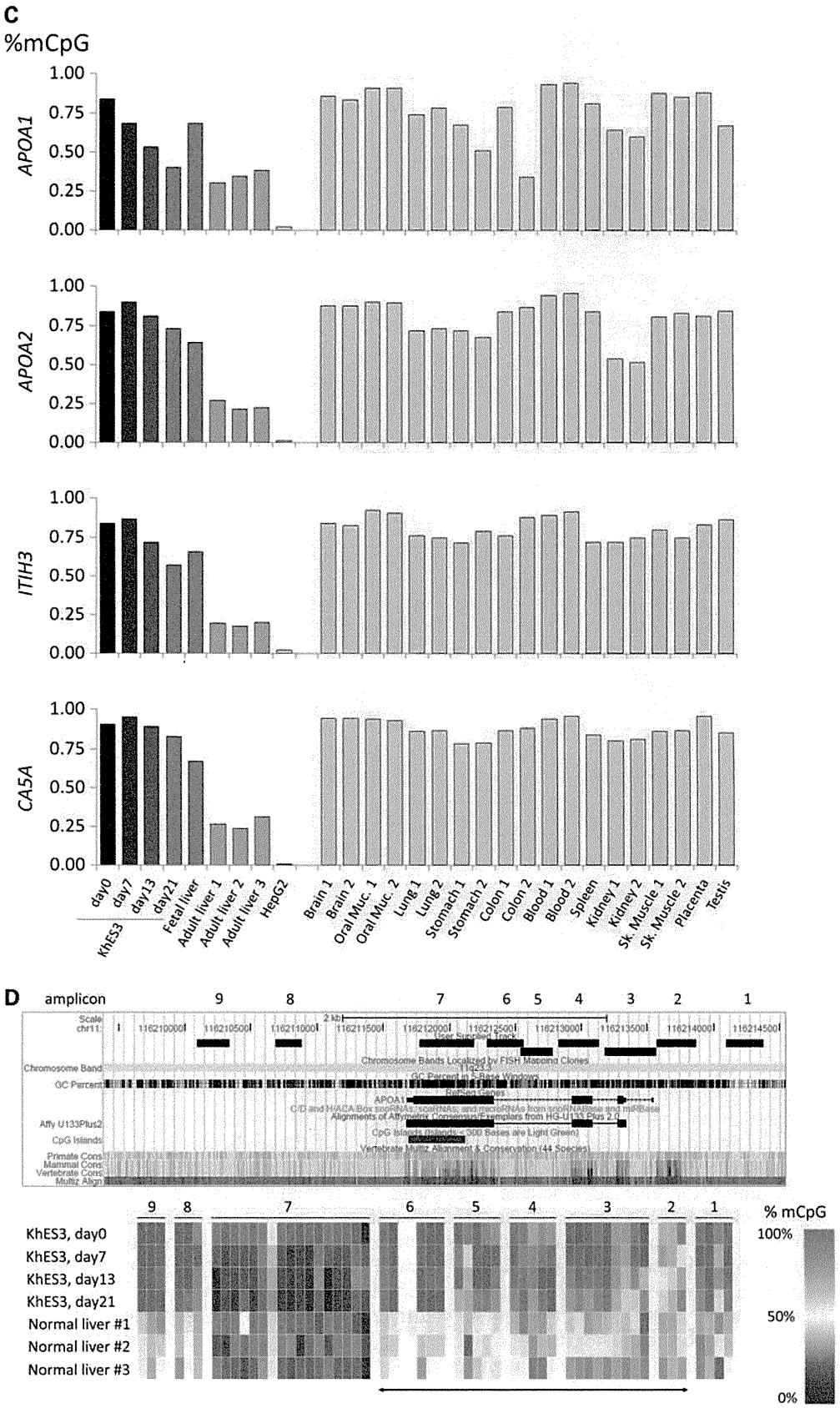

Figure 6. (Continued).

long-lasting maintenance of a transcriptionally active state. Subsequently, in this model, chromatin conformation changes in terminally differentiated cells would expand the demethylated regions and contribute to the establishment of stable and highly efficient expression of specific gene subsets.

Growing evidence suggests that forced induction of master regulator genes has the potential to change the fate of lineage-restricted cells, even in terminally differentiated cells (39–41). We identified restoration of methylation during reprogramming into iPS cells. The feasibility of cell reprogramming suggests that differentiated cells still have much more plasticity in the epigenetic status including DNA methylation than we had expected. Further analysis of methylation changes might provide novel insight into mechanisms that will generate a transcriptional repertoire for variable cell lineages and give us useful clues to control cell fate fixation, which might be applicable for regenerative medicine.

MATERIALS AND METHODS

Genomic DNA from human normal tissues

Frozen tissues of the brain, lung, liver and kidney were obtained from surgical specimens. Patients undergoing surgical resection at the Tokyo University General Hospital provided tissue after obtaining informed consent. Buccal swabs of oral mucosa, peripheral blood and placental tissue were from healthy volunteers. This study was certified by the Ethics Committee of Tokyo University. Genomic DNA from these clinical samples was extracted using the QIAamp DNA Mini Kit (QIAGEN). Genomic DNA of further individuals was purchased from BioChain (details are listed in Supplementary Material, Table S1). For the methylation-negative control, totally unmethylated genomic DNA was synthesized by a whole-genome amplification system, GenomiPhi (GE healthcare). For a positive control, fully methylated genomic DNA was generated by *Sss*.I CpG methylase (New England Biolabs) treatment of lymphocyte DNA.

Human ES cell lines

Human ES cell lines, KhES1, KhES2, KhES3, KhES4, were established and maintained as described previously (42). Human ES cell lines (H1, H9) and human iPS cell lines [iPS(IMR90)-1 and iPS(IMR90)-4] were obtained from WiCell Research Institute. HES3 cell line was obtained from ES Cell International.

Briefly, undifferentiated human ES cells were maintained on a feeder layer of MEF in DMEM/F12 (Sigma) supplemented with 20% KSR, l-Glu, NEAA and β-ME under 3% CO_2. To passage ES cells, ES cell colonies were detached from the feeder layer by treatment with 0.25% trypsin and 0.1 mg/ml of collagenase IV in PBS containing 20% KSR and 1 mm of $CaCl_2$ at 37°C for 5 min, followed by the addition of culture medium. ES cell clumps were disaggregated into smaller pieces by gentle pipetting.

An *in vitro* differentiation experiment was performed following the reported method, with some modification (43). Briefly, KhES3 cells were cultured in differentiation medium [RPMI supplemented with human recombinant activin A (100 ng/ml) and defined FBS]. FBS concentrations were 0% for the first 24 h, 0.2% for the second 48 h and 2.0% for subsequent days of differentiation. Media were replaced every 2 days with fresh differentiation medium supplemented with growth factors. ES cells were cultured in differentiation medium (DMEM supplemented with 10% KSR, Dex and HGF) for up to 30 days.

Methylation profiling

Methylation status was analyzed using HumanMethylation27 BeadChip (Illumina). Genomic DNA for methylation profiling was quantified using the Quant-iT dsDNA BR Assay Kit (Invitrogen). Five hundred nanograms of genomic DNA was bisulfite-converted using an EZ DNA Methylation Kit (Zymo Research). The converted DNA was amplified, fragmented and hybridized to a BeadChip according to the manufacturer's instructions. The raw signal intensity for both methylated (M) and unmethylated (U) DNA was measured using a BeadArray Scanner (Illumina). The methylation level of the each individual CpG is obtained using the formula (M)/(M)+(U)+100 by the GenomeStudio (Illumina).

Quantitative methylation analysis using the MassARRAY system

Bisulfite treatment of genomic DNA was performed using an EZ Methylation Kit (Zymo Research). Primer sequences are given in Supplementary Material, Table S4. This system utilizes MALDI-TOF mass spectrometry in combination with RNA base-specific cleavage (MassCLEAVE). A detectable pattern is analyzed for the methylation status. Mass spectra were acquired using a MassARRAY Compact MADLI-TOF (Sequenom) and spectra's methylation ratios were generated using Epityper software v1.0 (Sequenom).

Bisulfite sequencing

Bisulfite sequencing analysis was performed as described previously (44). Bisulfite treatment of genomic DNA was performed using an EZ Methylation Kit (Zymo Research). All primer sequences and melting temperatures for the polymerase chain reaction (PCR) are given in Supplementary Material, Table S4. PCR amplicons were subcloned into the pGEM-T vector (Promega). Clones were sequenced using PRISM3100 Sequencer (Applied Biosystems).

RNA extraction and gene expression microarray analysis

Genome-wide analysis of mRNA expression levels using U133plus2.0 human expression array® (Affymetrix) was done essentially as described previously (45). Briefly, total RNA was isolated using TRIzol reagent (Invitrogen), according to the manufacturer's instructions. One microgram of RNA was used for the generation of double-stranded cDNA with the SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen) according to the manufacturer's protocol. Double-stranded cDNAs were hybridized to the microarray.

Reverse transcription–polymerase chain reaction analysis

RNA extraction and reverse transcription–polymerase chain reaction (RT–PCR) were done as described (46). Total RNA was extracted using TRI Reagent (Sigma-Aldrich) or the RNeasy micro-kit (Qiagen) and then treated with DNase (Sigma-Aldrich). Three micrograms of RNA was reverse-transcribed using Moloney Murine Leukemia Virus reverse transcriptase (Toyobo, Japan) and oligo(dT) primers (Toyobo). The primer sequences are shown in Supplementary Material, Table S4. The PCR conditions for each cycle were as follows: denaturation at 96°C for 30 s, annealing at 60°C for 2 s and extension at 72°C for 45 s. RT–PCR products were separated by 5% non-denaturing polyacrylamide gel electrophoresis, stained with SYBR Green I (Molecular Probes), and visualized using a Gel Logic 200 Imaging System (Kodak).

Definition of probe classes and promoter classes

We classified 27 578 probes into three categories: HCG, ICG and LCG. Each probe position was defined with respect to the position of a given CpG site. We determined the GC content and the ratio of observed versus expected CpG dinucleotides in a surrounding 500 bp window. The CpG ratio was calculated using the following formula: (number of CpGs × number of bp) (number of Cs × number of Gs). Three categories of probes were determined as follows: (i) HCGs (8098 probes) covering a 500 bp area with a CpG ratio above 0.75 and GC content above 55%; (ii) LCGs (8374 probes) excluded from a 500 bp area with a CpG ratio above 0.48; and (iii) ICGs (11 106 probes) that could not be categorized as either HCGs or LCGs.

Clustering analysis

To analyze the similarity of the methylation levels among human somatic tissues, ES cells and iPS cells, we used the data set of tissue-specific hypomethylation selected in Figure 2A for the cluster analysis. We applied a hierarchical clustering algorithm using the uncentered correlation coefficient as the measure of similarity and average linkage clustering (47) and visualized the dendrogram and the heatmap using TreeView (48).

GO functional annotation analysis

GO functional annotations for differentially hypomethylated and hypermethylated gene sets were performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID) Bioinformatic Resources v6.7 (http://niaid.abcc.ncifcrf.gov/ home.jsp). The lists of 250 gene symbols that show specific hypermethylation or hypermethylation for each tissue were submitted and DAVID default population background (*Homo sapiens*) was chosen to detect significantly over-represented GO biological processes (GOTERM BP-FAT). *P*-values were calculated by a modified Fisher's exact test and adjusted for multiple hypotheses testing using Bonferroni correction. The three GO terms with the most significant *P*-value and the number of genes involved in the term were listed for each tissue.

Enrichment analysis of transcription factor-binding motifs

To determine over-represented transcription factor-binding sites in tissue-specific hypomethylated and hypermethylated regions, sequences around the probe within a 500 bp window were screened for the presence of binding sites using Genomatix RegionMiner (http://www.genomatix.de, matrix library version 7.1). The number of binding site motifs was determined and over-representation over the background of random mammalian promoter sequences was calculated as the *Z*-score. Transcription factor families with a Z-score greater than 8.0 were considered highly significant. The *Z*-scores of these representative TF modules are visualized in the heatmap using TreeView (48).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

We are grateful to Hiroko Meguro for microarray experiment, Kaoru Nakano for MassARRAY analysis, Elodie Lebredonchel for bisulfite sequencing experiment and Michael Jones for critical reading of the manuscript.

Conflict of Interest statement. None declared.

FUNDING

This work was mainly supported by a Grant-in-Aid for Scientific Research (S) 20221009 (H.A.) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, and the Program of Fundamental Studies in Health Sciences of the National Institute of Biomedical Innovation (NIBIO), Japan.

REFERENCES

1. Bird, A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.*, **16**, 6–21.
2. Bernstein, B.E., Meissner, A. and Lander, E.S. (2007) The mammalian epigenome. *Cell*, **128**, 669–681.
3. Li, E., Bestor, T.H. and Jaenisch, R. (1992) Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, **69**, 915–926.
4. Okano, M., Bell, D.W., Haber, D.A. and Li, E. (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, **99**, 247–257.
5. Jackson-Grusby, L., Beard, C., Possemato, R., Tudor, M., Fambrough, D., Csankovszki, G., Dausman, J., Lee, P., Wilson, C., Lander, E. *et al.* (2001) Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. *Nat. Genet.*, **27**, 31–39.
6. Rakyan, V.K., Down, T.A., Thorne, N.P., Flicek, P., Kulesha, E., Graf, S., Tomazou, E.M., Backdahl, L., Johnson, N., Herberth, M. *et al.* (2008) An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). *Genome Res.*, **18**, 1518–1529.
7. Khulan, B., Thompson, R.F., Ye, K., Fazzari, M.J., Suzuki, M., Stasiek, E., Figueroa, M.E., Glass, J.L., Chen, Q., Montagna, C. *et al.* (2006)

Comparative isoschizomer profiling of cytosine methylation: the HELP assay. *Genome Res.*, **16**, 1046–1055.
8. Shen, L., Kondo, Y., Guo, Y., Zhang, J., Zhang, L., Ahmed, S., Shu, J., Chen, X., Waterland, R.A. and Issa, J.P. (2007) Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. *PLoS Genet.*, **3**, 2023–2036.
9. Straussman, R., Nejman, D., Roberts, D., Steinfeld, I., Blum, B., Benvenisty, N., Simon, I., Yakhini, Z. and Cedar, H. (2009) Developmental programming of CpG island methylation profiles in the human genome. *Nat. Struct. Mol. Biol.*, **16**, 564–571.
10. Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V.K., Attwood, J., Burger, M., Burton, J., Cox, T.V., Davies, R., Down, T.A. *et al.* (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.*, **38**, 1378–1385.
11. Illingworth, R., Kerr, A., DeSousa, D., Jorgensen, H., Ellis, P., Stalker, J., Jackson, D., Clee, C., Plumb, R., Rogers, J. *et al.* (2008) A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol.*, **6**, e22.
12. Laird, P.W. (2010) Principles and challenges of genome-wide DNA methylation analysis. *Nat. Rev. Genet.*, **11**, 191–203.
13. Waterland, R.A., Kellermayer, R., Rached, M.T., Tatevian, N., Gomes, M.V., Zhang, J., Zhang, L., Chakravarty, A., Zhu, W., Laritsky, E. *et al.* (2009) Epigenomic profiling indicates a role for DNA methylation in early postnatal liver development. *Hum. Mol. Genet.*, **18**, 3026–3038.
14. Saxonov, S., Berg, P. and Brutlag, D.L. (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl Acad. Sci. USA*, **103**, 1412–1417.
15. Irizarry, R.A., Ladd-Acosta, C., Carvalho, B., Wu, H., Brandenburg, S.A., Jeddeloh, J.A., Wen, B. and Feinberg, A.P. (2008) Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res.*, **18**, 780–790.
16. Barrera, L.O., Li, Z., Smith, A.D., Arden, K.C., Cavenee, W.K., Zhang, M.Q., Green, R.D. and Ren, B. (2008) Genome-wide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs. *Genome Res.*, **18**, 46–59.
17. Bibikova, M., Le, J., Barnes, B., Saedinia-Melnyk, S., Zhou, L., Shen, R. and Gunderson, K.L. (2009) Genome-wide DNA methylation profiling using infinium assay. *Epigenomics*, **1**, 177–200.
18. Jones, P.A. and Takai, D. (2001) The role of DNA methylation in mammalian epigenetics. *Science*, **293**, 1068–1070.
19. Walsh, C.P. and Bestor, T.H. (1999) Cytosine methylation and mammalian development. *Genes Dev.*, **13**, 26–34.
20. Baek, D., Davis, C., Ewing, B., Gordon, D. and Green, P. (2007) Characterization and predictive discovery of evolutionarily conserved mammalian alternative promoters. *Genome Res.*, **17**, 145–155.
21. Kadonaga, J.T. (1998) Eukaryotic transcription: an interlaced network of transcription factors and chromatin-modifying machines. *Cell*, **92**, 307–313.
22. Yang, A., Zhu, Z., Kapranov, P., McKeon, F., Church, G.M., Gingeras, T.R. and Struhl, K. (2006) Relationships between p63 binding, DNA sequence, transcription activity, and biological function in human cells. *Mol. Cell*, **24**, 593–602.
23. Shiraki, N., Umeda, K., Sakashita, N., Takeya, M., Kume, K. and Kume, S. (2008) Differentiation of mouse and human embryonic stem cells into hepatic lineages. *Genes Cells*, **13**, 731–746.
24. Takahashi, K. and Yamanaka, S. (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, **126**, 663–676.
25. Wernig, M., Meissner, A., Foreman, R., Brambrink, T., Ku, M., Hochedlinger, K., Bernstein, B.E. and Jaenisch, R. (2007) *In vitro* reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature*, **448**, 318–324.
26. Cedar, H. and Bergman, Y. (2009) Linking DNA methylation and histone modification: patterns and paradigms. *Nat. Rev. Genet.*, **10**, 295–304.
27. Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Paabo, S., Rebhan, M. and Schubeler, D. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.*, **39**, 457–466.
28. Reik, W. (2007) Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, **447**, 425–432.
29. Bergman, Y. and Mostoslavsky, R. (1998) DNA demethylation: Turning genes on. *Biol. Chem.*, **379**, 401–407.
30. Eden, S. and Cedar, H. (1994) Role of DNA methylation in the regulation of transcription. *Curr. Opin. Genet. Dev.*, **4**, 255–259.
31. Ooi, S.K.T. and Bestor, T.H. (2008) The colorful history of active DNA demethylation. *Cell*, **133**, 1145–1148.
32. Wu, S.C. and Zhang, Y. (2010) Active DNA demethylation: many roads lead to Rome. *Nat. Rev. Mol. Cell Biol.*, **11**, 607–620.
33. Mayer, W., Niveleau, A., Walter, J., Fundele, R. and Haaf, T. (2000) Embryogenesis: demethylation of the zygotic paternal genome. *Nature*, **403**, 501–502.
34. Oswald, J., Engemann, S., Lane, N., Mayer, W., Olek, A., Fundele, R., Dean, W., Reik, W. and Walter, J. (2000) Active demethylation of the paternal genome in the mouse zygote. *Curr. Biol.*, **10**, 475–478.
35. Reik, W., Dean, W. and Walter, J. (2001) Epigenetic reprogramming in mammalian development. *Science*, **293**, 1089–1093.
36. Ito, S., D'Alessio, A.C., Taranova, O.V., Hong, K., Sowers, L.C. and Zhang, Y. (2010) Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature*, **466**, 1129–1133.
37. Young, D.W., Hassan, M.Q., Yang, X.-Q., Galindo, M., Javed, A., Zaidi, S.K., Furcinitti, P., Lapointe, D., Montecino, M., Lian, J.B. *et al.* (2007) Mitotic retention of gene expression patterns by the cell fate-determining transcription factor Runx2. *Proc. Natl Acad. Sci. USA*, **104**, 3189–3194.
38. Zaidi, S.K., Young, D.W., Montecino, M.A., Lian, J.B., van Wijnen, A.J., Stein, J.L. and Stein, G.S. (2010) Mitotic bookmarking of genes: a novel dimension to epigenetic control. *Nat. Rev. Genet.*, **11**, 583–589.
39. Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K. and Yamanaka, S. (2007) Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, **131**, 861–872.
40. Vierbuchen, T., Ostermeier, A., Pang, Z.P., Kokubu, Y., Sudhof, T.C. and Wernig, M. (2010) Direct conversion of fibroblasts to functional neurons by defined factors. *Nature*, **463**, 1035–1041.
41. Ieda, M., Fu, J.-D., Delgado-Olguin, P., Vedantham, V., Hayashi, Y., Bruneau, B.G. and Srivastava, D. (2010) Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell*, **142**, 375–386.
42. Suemori, H., Yasuchika, K., Hasegawa, K., Fujioka, T., Tsuneyoshi, N. and Nakatsuji, N. (2006) Efficient establishment of human embryonic stem cell lines and long-term maintenance with stable karyotype by enzymatic bulk passage. *Biochem. Biophys. Res. Commun.*, **345**, 926–932.
43. D'Amour, K.A., Agulnick, A.D., Eliazer, S., Kelly, O.G., Kroon, E. and Baetge, E.E. (2005) Efficient differentiation of human embryonic stem cells to definitive endoderm. *Nat. Biotech.*, **23**, 1534–1541.
44. Hayashi, H., Nagae, G., Tsutsumi, S., Kaneshiro, K., Kozaki, T., Kaneda, A., Sugisaki, H. and Aburatani, H. (2007) High-resolution mapping of DNA methylation in human genome using oligonucleotide tiling array. *Hum. Genet.*, **120**, 701–711.
45. Hippo, Y., Watanabe, K., Watanabe, A., Midorikawa, Y., Yamamoto, S., Ihara, S., Tokita, S., Iwanari, H., Ito, Y., Nakano, K. *et al.* (2004) Identification of soluble NH2-terminal fragment of glypican-3 as a serological marker for early-stage hepatocellular carcinoma. *Cancer Res.*, **64**, 2418–2423.
46. Shiraki, N., Yoshida, T., Araki, K., Umezawa, A., Higuchi, Y., Goto, H., Kume, K. and Kume, S. (2008) Guided differentiation of embryonic stem cells into Pdx1-expressing regional-specific definitive endoderm. *Stem Cells*, **26**, 874–885.
47. Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
48. Saldanha, A.J. (2004) Java Treeview–extensible visualization of microarray data. *Bioinformatics*, **20**, 3246–3248.

Identification of Novel Schizophrenia Loci by Homozygosity Mapping Using DNA Microarray Analysis

Naohiro Kurotaki[1]*, Shinya Tasaki[1], Hiroyuki Mishima[2,3], Shinji Ono[1], Akira Imamura[1], Taeko Kikuchi[1], Nao Nishida[4], Katsushi Tokunaga[4], Koh-ichiro Yoshiura[2], Hiroki Ozawa[1]

1 Department of Neuropsychiatry, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki, Japan, **2** Department of Human Genetics, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki, Japan, **3** Nagasaki University Global Center of Excellence Program, Nagasaki, Japan, **4** Department of Human Genetics, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

Abstract

The recent development of high-resolution DNA microarrays, in which hundreds of thousands of single nucleotide polymorphisms (SNPs) are genotyped, enables the rapid identification of susceptibility genes for complex diseases. Clusters of these SNPs may show runs of homozygosity (ROHs) that can be analyzed for association with disease. An analysis of patients whose parents were first cousins enables the search for autozygous segments in their offspring. Here, using the Affymetrix® Genome-Wide Human SNP Array 5.0 to determine ROHs, we genotyped 9 individuals with schizophrenia (SCZ) whose parents were first cousins. We identified overlapping ROHs on chromosomes 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 16, 17, 19, 20, and 21 in at least 3 individuals. Only the locus on chromosome 5 has been reported previously. The ROHs on chromosome 5q23.3–q31.1 include the candidate genes histidine triad nucleotide binding protein 1 (*HINT1*) and acyl-CoA synthetase long-chain family member 6 (*ACSL6*). Other overlapping ROHs may contain novel rare recessive variants that affect SCZ specifically in our samples, given the highly heterozygous nature of SCZ. Analysis of patients whose parents are first cousins may provide new insights for the genetic analysis of psychiatric diseases.

Citation: Kurotaki N, Tasaki S, Mishima H, Ono S, Imamura A, et al. (2011) Identification of Novel Schizophrenia Loci by Homozygosity Mapping Using DNA Microarray Analysis. PLoS ONE 6(5): e20589. doi:10.1371/journal.pone.0020589

Editor: Xiang Yang Zhang, Baylor College of Medicine, United States of America

Received November 28, 2010; **Accepted** May 6, 2011; **Published** May 31, 2011

Funding: NK was supported in part by grants from Grant-in-Aid for Scientific Research (No. 19591363) and by grants from the Mitsubishi Pharma Research Foundation. KY was supported in part by Grant-in-Aid for Scientific Research from the Ministry of Health, Labor, and Welfare, from the Takeda Scientific Foundation, and from the Naito Foundation. This work was also supported by Nagasaki University Global COE program, global strategic center for radiation health risk control. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: naokuro@nagasaki-u.ac.jp

Introduction

Schizophrenia (SCZ) is categorized as a severe chronic debilitating psychosis that affects approximately 1% of the global population. Although genetic factors are reported to contribute to the disease and multiple responsible loci have been identified from linkage analysis and case-control association studies, there have been few reproducible results to date [1].

Morrow et al. (2008) [2] suggested that homozygosity mapping is a powerful tool not only for investigating single gene defects but also for rare genomic variants in complex traits. They observed homozygous deletions in patients with autistic disorders and concluded that genomic alterations might be a subset of disease-causing mutations in chromosomal regions. The increased susceptibility to SCZ observed in consanguineous families suggests that genomic recessive variations may be involved in its etiology. [3–5] Considering this and other results, we hypothesized that homozygosity mapping, including identical by descent (IBD) analysis, would be a highly constructive method for identifying the loci responsible for SCZ.

We hypothesized that runs of homozygosity (ROHs) could contribute to SCZ by a recessive effect. We use the term "ROH" [6] instead of loss of heterozygosity (LOH) for regions where homozygous genotypes are contiguous because LOH implies heterozygous deletions or hemizygosity, while ROH suggests consecutive homozygous regions. Recessive effects are obtained by genetic variations including single nucleotide variations, small insertions/deletions, structural variations, and chromosomal rearrangements. These variations may affect amino acid sequences or the control of gene expression, including small RNA expression.

Here, we describe a homozygosity mapping strategy that consisted of 2 stages (Figure 1). The first stage aimed to find the appropriate size threshold for autosomal ROHs that would distinguish ROHs specifically existing in the offspring of first-cousin marriages from those that commonly exist in the offspring of non-consanguineous marriages. By comparing the size distribution of ROHs between the offspring of first-cousin marriages and non-consanguineous marriages, we concluded that ROHs >2.1 Mb in size in the offspring of consanguineous marriages can be assumed to be IBD segments from an individual 3 generations before. The second stage aimed to find shared ROHs among patient with SCZ using 2 models. In Model I, an autosomal ROH size threshold was applied to filter out smaller ROHs. Larger ROHs were assessed to find overlaps among the patients. In Model II, after filtering by the ROH size threshold, ROHs shared by the siblings of patients and ROHs of other patients were assessed to find overlaps. The overlapping ROHs we identified potentially contain SCZ causative regions that are specific to our samples because of the heterogeneous nature of SCZ.

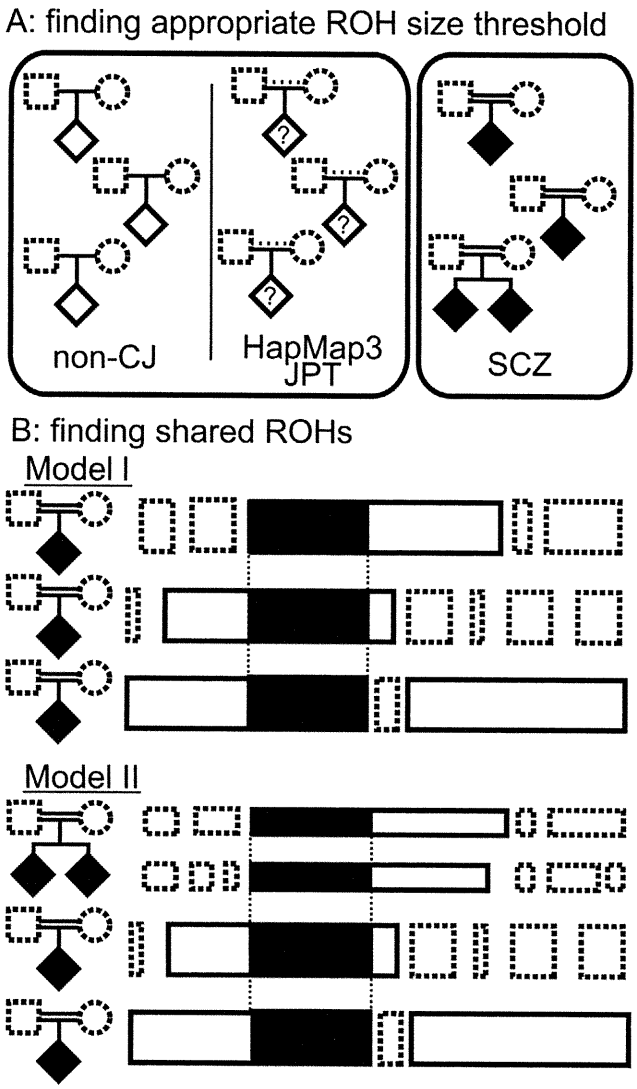

Figure 1. Two-stage design of this study. A, the first stage was to find an appropriate autosomal run of homozygosity (ROH) size threshold to distinguish specific ROHs from the offspring of first-cousin marriages from ROHs in the offspring of non-consanguineous marriages. The size distribution of ROHs in our non-consanguineous Japanese (non-CJ) and schizophrenia (SCZ) samples was compared. Non-CJ samples are the offspring of non-consanguineous marriages that were validated by interview. Here, SCZ samples were used as the offspring of first-cousin marriages regardless of phenotype. Samples from parents were not used in this study (dashed squares and circles). To confirm our strategy, we also assessed HapMap3 JPT samples, which do not have information for phenotypes or family consanguinity (dashed and solid lines between parents). B, the second stage was to find shared ROHs among the SCZ samples as patients with schizophrenia. In Model I, an autosomal ROH size threshold was applied to filter out smaller ROHs (dashed open boxes). Larger ROHs (solid open boxes) were assessed to find overlaps among patients (solid boxes). In Model II, after filtering by the ROH size threshold, ROHs shared by the siblings of patients and ROHs of other patients were assessed to find overlaps. In this study, the gender of the samples was not matched (diamonds) because we only evaluated autosomal ROHs.
doi:10.1371/journal.pone.0020589.g001

Materials and Methods

1. Samples

A total of 9 subjects with SCZ (3 males and 6 females, aged 31–56 years) (SCZ individuals) were recruited to this study after being diagnosed as having typical paranoid schizophrenia by a certified psychiatrist (N.K.) using the *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition, Text Revision (DSM-IV-TR) and the *Structured Clinical Interview for DSM-IV Axis I Disorders* (SCID). The study received ethics approval from the Committee for Ethical Issues on Human Genome and Gene Analysis at Nagasaki University, Japan. All of the patients were from the main islands of Japan, excluding Okinawa. We obtained written informed consent from all participants. The consanguineous patients were from 8 first-cousin marriages. Seven individuals (patients a to g) were unrelated and 2 were siblings (patients h-1 and h-2). We also recruited 92 healthy individuals from non-consanguineous marriages (non-CJ individuals) from the main islands of Japan, excluding Okinawa. We confirmed consanguinity by interview. We did not match for gender in the SCZ and non-CJ individuals because we only intended to analyze autosomal chromosomes.

After obtaining written informed consent, genomic DNA was isolated from peripheral blood. We did not collect blood samples from the patients' parents, except for 1 patient, or siblings; however, we confirmed that they had no history of psychiatric illness, with the exception of the older brother of patient g, by direct interview or from the medical records of the other related individuals.

Furthermore, we also assessed the International HapMap Project [7] phase 3 data of the Japanese in Tokyo (HapMap3 JPT) to evaluate the non-CJ individuals. Raw signal intensity files (CEL files) obtained using Affymetrix Genome-Wide Human SNP Array 6.0 (Affy6.0) were downloaded from http://www.hapmap.org/.

2. Microarray analysis

We performed genome-wide SNP genotyping of 9 SCZ samples and 92 non-CJ samples using the Affymetrix Genome-Wide Human SNP Array 5.0 (Affy5.0) according to the manufacturer's instructions. Our microarray data is MIAME compliant and the raw data has been deposited in the CIBEX database (CIBEX accession number: CBX141).

3. ROH detection

We generated the CHP genotype files from the CEL signal intensity files using the BRLMM-P genotype calling program [8,9]. For the detection of ROHs, we analyzed the CHP files with a hidden Markov model (HMM)-based ROH detection function of the Partek® Genomics Suite (Partek GS) software version 6.5 build 6.11.0207 (Partek, St. Louis, MO, USA). We applied the following default HMM parameters: max probability = 0.99, genomic decay = 0 (disabled), genotype error = 0.01, and default frequency = 0.3. We did not adopt the baseline files.

Detected ROHs were statistically analyzed and visualized (Figures 2 and 3; Tables 1 and 2) by using in-house scripts written in the R language [10]. The optimization of histogram bandwidths and the estimation of the probability density distributions were performed using the "KernSmooth" package of R [11].

Furthermore, to validate the data quality of our non-CJ samples, we also compared our data to HapMap3 JPT. Affy6.0 raw signal intensity data in CEL files were subjected to allele calling using Birdseed software version 2 [12]. SNP genotypes of shared loci between Affy6.0 and Affy5.0 were extracted and processed as well as the non-CJ and SCZ datasets to detect ROHs.

4. Detection of potential genetic loci for SCZ by overlapping ROHs

To detect the overlapping ROHs among the SCZ dataset, the identified ROHs were filtered by a size threshold on Partek GS,

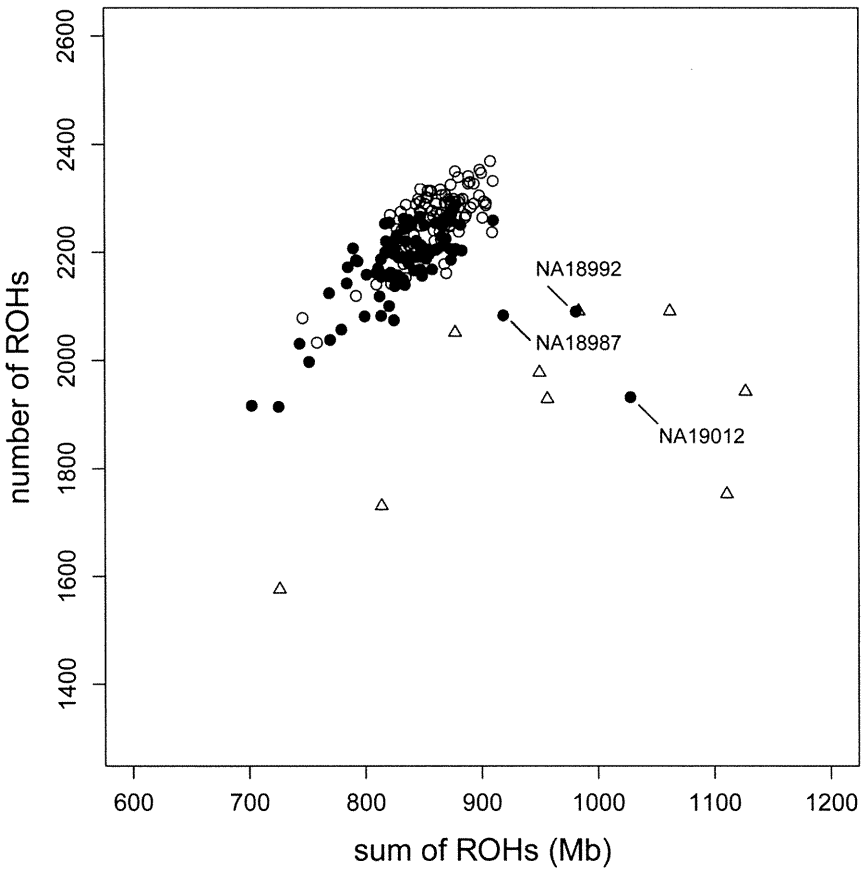

Figure 2. Distribution of the size and number of individual autosomal runs of homozygosity (ROHs). Sums and total numbers of individual ROHs are shown by circles and triangles indicating unrelated Japanese individuals (non-CJ: 92 samples) and the offspring of first-cousin marriages with schizophrenia (SCZ: 9 samples), respectively.
doi:10.1371/journal.pone.0020589.g002

analyzed using an in-house Ruby script (available on request) to generate a table of overlapping ROHs, and visualized with Partek GS. Then, we extracted the loci shared among more than 3 unrelated individuals (Model I) (Table S1). Furthermore, on the basis of the hypothesis that concordant sibling cases share causal loci, we detected the loci shared among 2 sibling cases (h-1 and h-2) (Model II) and found the ROHs that were shared by 1 or more of the unrelated samples (Table 3).

Results

1. Determination of the ROH size threshold discriminating the offspring from non-consanguineous and first-cousin marriages

We genotyped 440 794 SNPs in each individual. Genotype calling rates for each sample ranged from 97.23–98.83% and their call rates were high and accurate enough for their subsequent evaluation. We utilized the data from 92 non-CJ and 91 HapMap3 JPT samples in addition to the data from 9 SCZ individuals.

Our homozygosity mapping strategy utilized differences in the length distribution of ROHs between offspring from consanguineous and non-consanguineous marriages. Individuals from consanguineous families are expected to have an increased number of longer ROHs containing autozygous segments. These segments were also expected to be discriminated by their length from ROHs containing homozygous segments by chance or by linkage disequilibrium (LD). To demonstrate the strategy, we performed detailed comparisons of the length distribution of ROHs between the non-CJ, HapMap3 JPT, and SCZ datasets.

We initially plotted the total number and size of ROHs in the non-CJ, HapMap3 JPT, and SCZ datasets (Figure 2). The non-CJ and HapMap3 JPT datasets clustered together, except for 3 individuals in HapMap JPT. These 3 outlier individuals, NA18987, NA18992 [13], and NA19012 [14], have been assumed to be from consanguineous families; indeed, the distribution of these samples was similar to that of our offspring from first-cousin marriages (Figure 2).

We then analyzed the length distribution of ROHs in the non-CJ and SCZ datasets. Bar plot histograms of the length of ROHs were obtained and the probability density curves were estimated by the "KernSmooth" package in R (Figure 3A–D). Descriptive statistics of these plots are also shown in Table 1. Both datasets produced bell curve-like distributions in the $\log_{10}$ scale on the x-axis to indicate the length of each ROH; however, the SCZ dataset showed a secondary peak in the larger ROH region. We expected that the autozygous region from the founders of the third ancestral generation (great-grandparents) would be larger in the SCZ dataset than in the non-CJ dataset, in whom LD may encompass ROHs by chance. The proportion of larger ROHs in the SCZ dataset was clearly higher than in the non-CJ dataset. As we can expect that 1/16 of the whole genome in the offspring of first-cousin marriages would be autozygous regions from their great-grandparents, we highlighted the graphs in Figure 3B and

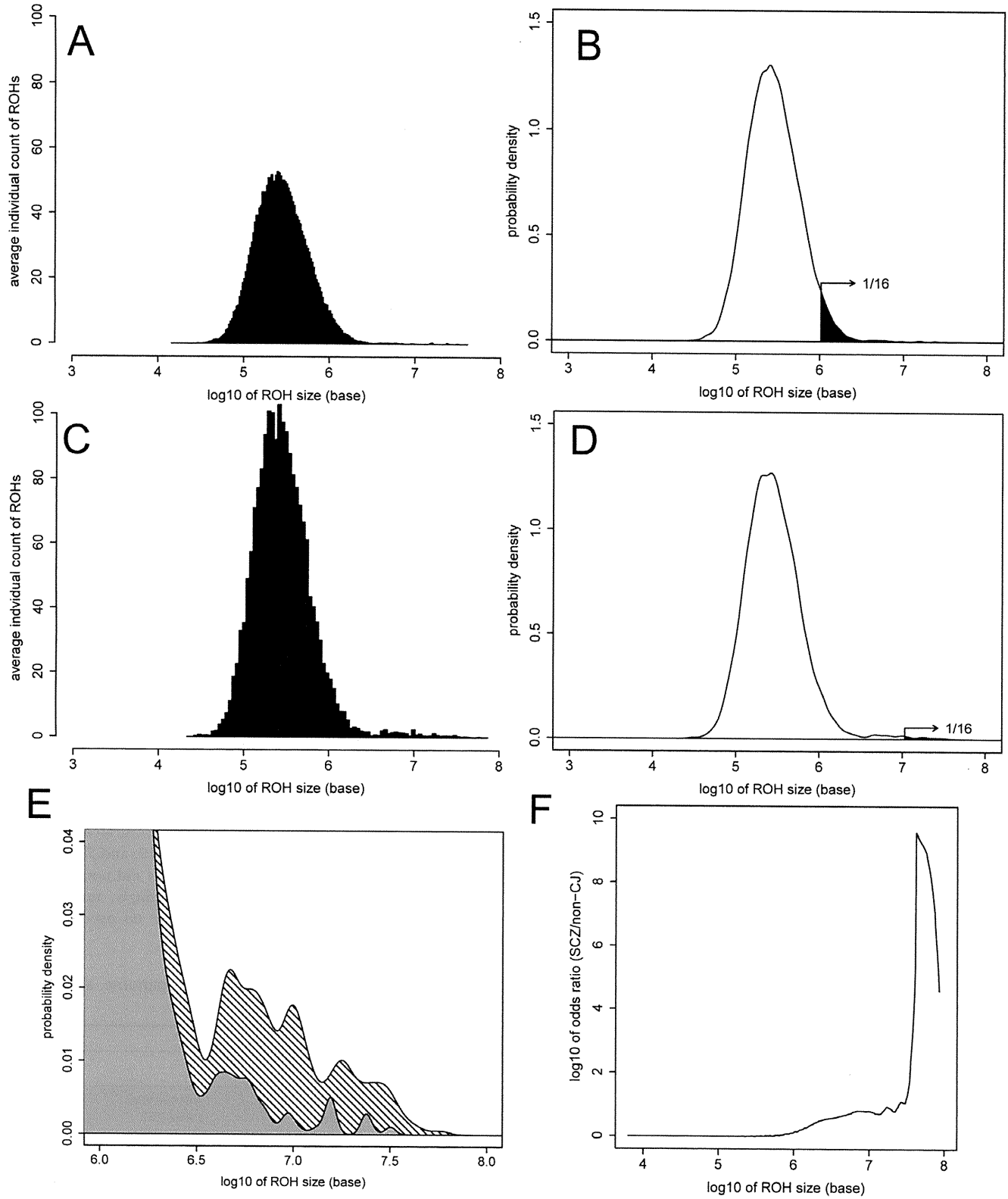

Figure 3. Size distribution of autosomal runs of homozygosity (ROHs). In the size distribution plot of non-consanguineous Japanese (non-CJ; A and B) and schizophrenia (SCZ; C and D) samples, the x-axis indicates the ROH size (log_{10} scale). A and C, individual average frequency of the ROHs as histograms. B and D, estimated probability density corresponding to each histogram. Black areas shows 1/16 (6.25%) of autosomes, which is equivalent to the expected sum of autozygous regions in the offspring of a first-cousin marriage. E, enlarged overlap of B (gray) and D (hatched). F, SCZ/non-CJ odds ratio plot. X-axis indicates the size of the ROHs (log_{10} scale). Y-axis (log_{10} scale) indicates the ratio of areas exceeding the given ROH size threshold in the estimated probability distributions of the SCZ and non-CJ datasets.
doi:10.1371/journal.pone.0020589.g003

Table 1. Autosomal runs of homozygosity (ROHs) size distribution, where descriptive statistics of ROH sizes were detected with Partek GS.

Dataset	N	Minimum[a]	Mode[b]	Maximum[c]	Average sum[d]
HM3JPT[e]	88	19 750 (14)	256 499 (27)	32 000 000 (1252)	831 159 144
non-CJ[f]	92	18 160 (14)	248 288 (27)	32 250 000 (1921)	859 784 793
SCZ[g]	9	27 380 (14)	258 488 (38)	57 810 000 (9896)	956 266 858

[a]Minimum ROH size in all individuals from each dataset.
[b]Mode ROH size in all individuals from each dataset.
[c]Maximum ROH size in all individuals from each dataset.
[d]Average sum is the average total ROH size per individual from each dataset.
[e]The International HapMap Project phase 3 Japanese in Tokyo. Three samples, NA18987, NA18992, and NA19012, of 91 samples are omitted because they are potentially the offspring of a consanguineous marriage.
[f]Non-consanguineous Japanese.
[g]Schizophrenia.
Numbers are in bases, and the numbers in parentheses are the included probe sets.
doi:10.1371/journal.pone.0020589.t001

3D at the point where the total sum of length in the upper tail of the ROH distribution reaches 179.2 Mb, which is 1/16 of the 2 867 732 772 bases total size of the autosomal haploid genome, according to the statistics from the NCBI Build 36.1 assembly (2006) [16]. This analysis suggested that it is highly probable that the longer ROHs would be inherited from the great-grandparents; however, it should be mentioned that genomic regions with less recombination tend to have longer ROHs.

To show further differences in the probability density distribution of the SCZ and non-CJ individuals, we also plotted an SCZ/non-CJ odds ratio (OR) plot (Figure 3F and Table 2), which indicates the ratio of probability for the existence of ROHs in each dataset over a given threshold length. To determine the overlapping ROH regions shared among the SCZ dataset, we adopted OR = 3.0 and the corresponding threshold of 2 137 962 bases to ensure practical power and to detect smaller IBD regions by recombination.

2. Determination of potential SCZ genetic loci by overlapping ROHs

The sum lengths of the overlapping regions among 0–7 independent family patients are shown in Figure 4, and the calculated percentage sum length among a given number of patients and more in the autosomal genome were as follows: 100%, 51.7%, 13.6%, 6.0%, 1.9%, 1.3%, and 0.6%. Considering the statistics, we adopted a minimum of 3 patients for identifying candidate loci. Figure 5 shows a schema of the overlapping ROHs within autosomes and their positions are summarized in Table S1.

Overlapping ROHs found in 3 or more SCZ individuals on chromosomes 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 16, 17, 19, 20, and 21 (Figure 5A) suggested that many loci are potentially associated with SCZ in our patients. Only the locus on chromosome 5 has been reported in a previous linkage analysis of SCZ [16]. The ROHs were expanded by the analysis of 4 additional individuals; however, no additional loci were detected (data not shown). The locus on chromosome 5q23.3–q31.1 included the regions containing the histidine triad nucleotide binding protein 1 (*HINT1*) and acyl-CoA synthetase long-chain family member 6 (*ACSL6*) genes. Our results suggest that recessive variants of these candidate genes could be involved in the pathogenesis of SCZ in our patients.

In the analysis of 2 siblings (h-1 and h-2) from a first-cousin marriage, we searched for the ROH regions shared by the siblings as a single gene defect. The detection of loci shared by the siblings and 1 or more unrelated individuals demonstrated ROHs on chromosomes 1, 5, 7, 8, 10, 12, 13, 16, 17, 19, and 21 that might be causative for SCZ (Figure 5B). Those loci did not include any previously reported candidate genes. Interestingly, among the loci detected in Figure 5A and 5B, there were no overlapping loci identified in this study.

Table 2. Thresholds, individual average sums of runs of homozygosity (ROHs), its ratio in the autosomal genome, and the individual average encompassed number of ROHs corresponding to the odds ratios.

Odds ratio	Threshold (base)	Non-CJ[a] dataset			SCZ[b] dataset		
		sum (base)	Autosomal ratio (%)	# of ROHs	sum (base)	Autosomal ratio (%)	# of ROHs
1.3	1 000 000	185 411 092	6.5	93.2	420 200 807	14.7	123.6
2.000	1 548 817	110 468 918	3.9	30.6	341 258 405	11.9	52.7
3.0	2 137 962	81 383 855	2.8	13.8	309 296 125	10.8	33.8
4.0	3 630 781	65 633 075	2.3	7.6	288 028 919	10.0	25.4
5.0	5 128 614	53 423 627	1.9	4.7	263 695 116	9.2	19.8
10.0	24 547 089	7 925 263	0.3	0.3	85 167 811	3.0	2.7

[a]Non-consanguineous Japanese.
[b]Schizophrenia.
doi:10.1371/journal.pone.0020589.t002

Table 3. Novel loci identified in this study that are different from those in Table S1, for the segments overlapping in more than 1 unrelated individual and the common regions between the 2 siblings (cases h-1 and h-2).

Chromosome	Start	End	Samples	# Samples[a]	Length	Cytoband
1	146258078	148749860	h-1, h-2, a	3	2491783	1q21.1-q21.2
5	45437574	49631829	h-1, h-2, d	3	4194256	5p12-q11.1
5	117360252	120214932	h-1, h-2, f	3	2854681	5q23.1
5	120214932	122586267	h-1, h-2, f, g	4	2371336	5q23.1-23.2
7	57594442	62282881	h-1, h-2, b, f	4	4688440	7p11.2-q11.21
8	129121122	131617749	h-1, h-2, b	3	2496628	8q24.21-q24.22
8	132434559	139244531	h-1, h-2, b	3	6809973	8q24.22-24.23
10	37363792	37599485	h-1, h-2, e	3	235694	10p11.21
10	37599485	37874740	h-1, h-2, e, g	4	275256	10p11.21
10	37874740	42217616	h-1, h-2, c, e, g	5	4342877	10p11.21-q11.21
12	33982292	36255461	h-1, h-2, a, d	4	2273170	12p11.1-q11
13	35366458	43580724	h-1, h-2, g	3	8214267	13q13.3-14.11
16	28924029	29606107	h-1, h-2, c	3	682079	16p11.2
16	29606107	29657036	h-1, h-2, c, f	4	50930	16p11.2
16	29657036	29680943	h-1, h-2, c, d, f	5	23908	16p11.2
16	29680943	31277953	h-1, h-2, b, c, d, f	6	1597011	16p11.2
16	34467305	34647935	h-1, h-2, a, b, c, d, f, g	8	180631	16p11.1
16	34647935	45122807	h-1, h-2, a, c, d, f, g	7	10474873	16p11.1-q11.2
16	45122807	47094922	h-1, h-2, a, b, c, d, f, g	8	1972116	16q11.2-q12.1
17	29659797	32811528	h-1, h-2, a	3	3151732	17q12
19	37676724	40349191	h-1, h-2, a	3	2672468	19q13.11-13.12
21	19821557	20188026	h-1, h-2, g	3	366470	21q21.2

[a]Number of individuals (including h-1 and h-2) who shared the region; for example, 5 indicates that 3 other individuals shared the common region of the 2 siblings.
doi:10.1371/journal.pone.0020589.t003

Discussion

1. Samples

We recruited 9 offspring from first-cousin marriages (SCZ) and 92 from non-consanguineous marriages (non-CJ). As shown in Figure 2, our non-CJ dataset and publicly available HapMap3 JPT datasets showed a common cluster, except for the presence of 3 outliers that have been reported to be potentially from consanguineous families [14,15]. This concordance suggests that our experimental quality and data processing approaches were appropriate. In this study, we analyzed a limited number of samples; however, homozygosity mapping was a reasonable strategy to adopt because it requires relatively smaller number of samples than case-control studies. We did not use samples from the parents of patients in this study because these are not very informative in our strategy. On the other hand, affected and unaffected siblings in single families are strongly informative in homozygosity mapping, and we are continuously recruiting additional siblings for future study.

2. ROH analysis

Most of the previous homozygosity mapping studies were based on genotypes derived from microsatellites or simple tandem-repeat polymorphisms (STRP). The highly polymorphic nature of multi-allelic STRP markers is suitable to cover the whole genome with a fewer numbers of markers. However, recent DNA microarray technologies have enabled massive genome-wide SNP genotyping to be performed in a short time. The problem with homozygosity mapping based on SNPs is the accurate detection of regions with ROHs. As SNPs have a less informative biallelic nature, using the naïve definition of an ROH as just a contiguous homozygous region may skew the detection of ROHs because of the frequency of low minor allele SNPs, genotyping errors, and "no-call" SNPs.

The solution to this problem using the Affymetrix Human Genotyping 500K arrays and Illumina Infinium HumanHap300v2 arrays was the application of ROH detection bins sliding through each chromosome to filter out low SNP density bins and to allow the small number of heterozygous SNPs and no-call SNPs to be placed in a bin [6,17]. An alternative method to detect ROHs is to adopt an HMM. Partek GS software implements the HMM-based "LOH detection" algorithm. A similar algorithm is also implemented in the Affymetrix GeneChip Chromosome Copy Number Analysis Tool (CNAT), as described in the CNAT user guide [18]. The HMM-based algorithm of these tools takes not only the information of adjacent SNP genotypes but also the heterozygosity of SNPs as a reference baseline calculated from the genotyping results in the reference samples or the *a priori* default frequency. This method is expected to more accurately detect ROH regions that reflect actual recombination.

Selection of the reference population for the baseline data is crucial for the HMM-based detection of ROHs. If the reference population is carefully selected to match the background of the case population, the baseline generated from the observation of actual SNPs in the reference population can omit ROHs resulting from LD and regions with low SNP density, such as centromeres. However, if strict matching of the used population background is difficult, use of the fixed default heterozygous frequency, whose

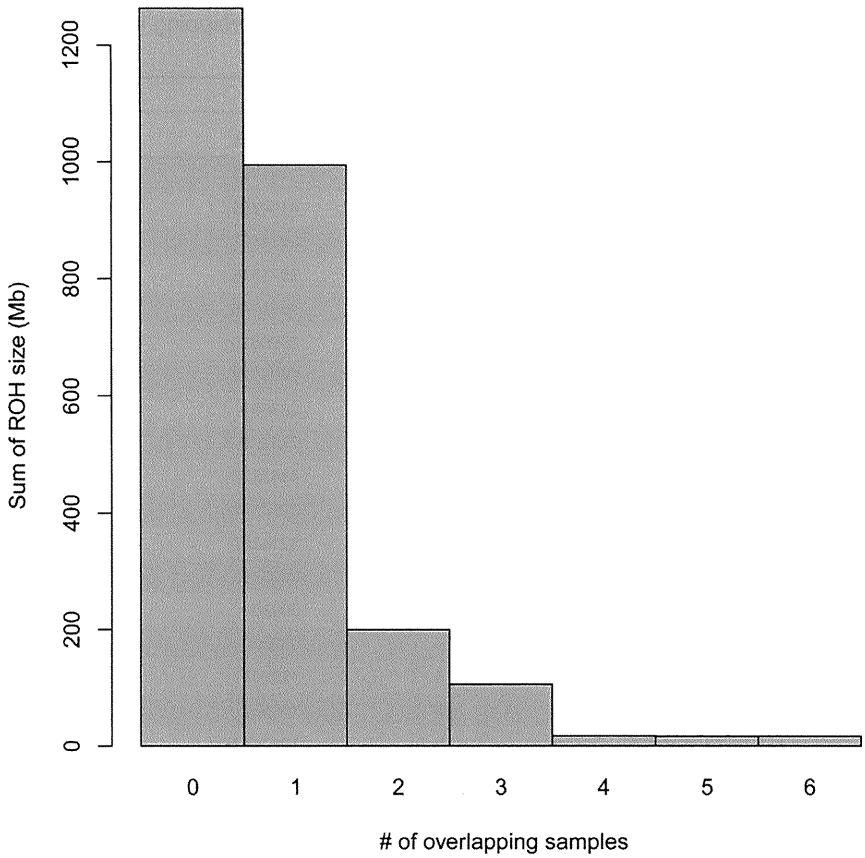

Figure 4. Sum of run of homozygosity (ROH) lengths and number of overlapping patients, excluding patient siblings. Y-axis indicates the sum of ROH lengths shared by a given number of patients. The zero column indicates the sum of ROHs not shared by any of the samples.
doi:10.1371/journal.pone.0020589.g004

default value is 0.3, still has the advantage of minimizing false-negatives in the detection of ROHs.

To determine the optimal length threshold of ROHs to extract autozygous segments from whole ROHs, we adopted OR = 3 for the analysis. This approach may work well when a large enough reference sample is available. When a reference population is not available, a threshold where the sum of the ROH length in the upper tail of its distribution is equal to the theoretical autozygous length of a genome, that is, 1/16 of a genome in the offspring of a first-cousin marriage, could be another option. In our SCZ dataset, the threshold using this approach was approximately 10.6 Mb.

Our results demonstrated obvious differences in the proportion of the length distribution of ROHs between the non-CJ and SCZ datasets. A recent report on European populations, including endogamy subpopulations, has shown that a higher proportion of individuals in endogamy subpopulations have ROHs longer than 1.5 Mb compared with other subpopulations [6]. Our scatter plot of the individual total number and size of ROHs (Figure 2) is not fully in agreement with this previous report, although the non-CJ dataset made a cluster and showed a positive correlation (Pearson product-moment correlation coefficient r = 0.773), and the SCZ dataset was scattered and showed a weak positive correlation (r = 0.432). This may be explained by the fact that the previous report excluded ROHs <500 kb to ignore ROHs that potentially resulted from LD and removed hemizygous deletions of ROHs. In this study, we did not adopt a strategy to filter ROHs by their size before the analyses because the discrimination of autozygous regions and LD simply by size is essentially impossible. Adopting a baseline file derived from a strictly matched population in the HMM-based detection of ROHs can be used instead. Additionally, differences in genotyping platforms with different SNP densities may affect the size distribution of ROHs. Although our data from the sparser Affymetrix Genotyping 10k SNP panel produced a similar bell curve-like ROH size distribution, the whole curve was shifted to the right (data not shown).

The size distribution of ROHs for a given population is affected by its inbreeding coefficient (F). Studies of consanguineous marriages in subpopulations from Japan during the 1980s compared the F values for Japan (F = 0.00134) to those in Kuwait (F = 0.0219), India (F = 0.02313), England (F = 0.00017), and the United States (F = 0.00003) [19,20]. These reports have also shown that despite the decrease in consanguineous marriages in Japan, local subpopulations have higher F-values. The same tendency has also been shown by a genealogical study that estimated inbreeding rates in large and semi-isolated populations on the basis of historical changes in population size [21]. Recently, the importance of studying endogamous populations has been stressed [22]; however, populations with intermediate F-values have advantages for our homozygosity mapping approach. This approach uses the differences in the size distribution of ROHs in a case population consisting of offspring from consanguineous marriages and a control population consisting of offspring from non-consanguineous marriages. A high F-value population may not have clear distribution differences between cases and controls. On the other hand, finding a sufficient number of cases in low F-value populations may not be easy. From this standpoint, an intermediate F-value population, such as the Japanese population,


A
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
B
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

Figure 5. Overlapping autosomal runs of homozygosity. Each autosome is shown horizontally with the number of overlapping samples (upper) and chromosome ideograms (lower). Centromeres are shown by hatched boxes. A, overlapping segments shared among 1 (isolated) to 7 samples in a total of 9 patient samples. B, overlapping segments shared by 2 patient siblings (h-I and h-II) and an additional 1–4 patient samples.
doi:10.1371/journal.pone.0020589.g005

represents an interesting dataset for our homozygosity mapping approach, as was shown previously in the Costa Rican population [23,24].

We presented here some threshold lengths of ROHs to detect IBD regions from great-grandparents. As recombination will, of course, occur everywhere by chance, small autozygous regions could be overlooked with the threshold shown here. However, no systematic analyses have so far identified IBD regions in consanguineous marriages by whole-genome SNP typing. Our method shown here, which 1) detects longer ROHs in each individual and 2) aligns ROHs and identifies overlapping regions, will be helpful for autosomal recessive disorders and also for complex disorders resulting from rare variants. If collecting patients in geographically and historically isolated areas is possible, this homozygosity mapping approach is likely to be successful. Nonetheless, the effectiveness of homozygosity mapping for complex disorders remains controversial [4]. We believe that we can uncover new candidate loci through the application of whole-genome SNP typing to homozygosity mapping because of its high density genomic coverage and high-throughput ability.

3. Possible novel loci for schizophrenia

We identified several putative SCZ loci that are presented in Figure 5 and Tables S1 and 3. In our study, we assumed the 2 models outlined in the Methods section. Model I was designed to find shared causal loci among unrelated individuals. For the other model, we hypothesized that the siblings shared the same causal loci; thus, Model II was designed to find the common loci between the siblings and unrelated individuals.

For Model I (Table S1), from the analysis of 7 unrelated individuals, the loci included the 5q23.3–q31.1 region that was previously identified by linkage analysis in the Irish population [16]. Among the genes that mapped to 5q23.3–q31.1, *HINT1* [25,26] and *ACSL6* [27] were previously reported to be possibly associated with SCZ. In patients from consanguineous families we analyzed, homozygous genomic variations may be causative for the disease (Figure 5 and Tables S1 and 3). However, small sample size in our study may be a limiting factor to generalize such conclusions. In common diseases such as psychiatric disorders, including SCZ and bipolar disorders, especially in familial cases or in cases from relatively isolated areas, rare variants possibly contribute more than common variants to the disease phenotype [28–30].

On the basis of the rare variant-common disease hypothesis, it is appropriate that the genetic etiology between sibling cases and other unrelated cases may be various. In addition, our results suggested that multiple loci influenced the susceptibility to SCZ, as other reports have suggested [31].

We presented here the systematic analyses of the homozygosity mapping method using whole genome SNP typing, and we identified ROHs that potentially contain SCZ causative recessive regions that are shared among our samples. When we explain SCZ as a result of the homozygous state of rare variant mutations, the number of overlapping individuals may be challenging, as it is possible that each individual has a different variation. The heterogeneity of SCZ may explain the lack of overlap for our results with previously reported regions [32,33]; moreover, our methodology has a limitation for detecting causative genes that are included in shorter ROHs by chance. We have shown that the Affymetrix Genome-Wide Human SNP Array 5.0 or 6.0 could be applied to special cases including first-cousin marriages to identify genomic variations. Increasing number of samples obtained from patients from consanguineous families with SCZ is important to make our results more meaningful. Furthermore, we plan to analyze genetic variants in updated ROHs by the next-generation sequencing technologies.

Supporting Information

Table S1 Novel loci identified in this study.
(DOC)

Acknowledgments

We thank Dr. Haruko Ichinose from Eijinkai Ariake Hoyouin Hospital and Drs. Takehito Sakai and Sumihisa Honda from Nagasaki University for their assistance. We also thank Dr. Pawel Stankiewicz from Baylor College of Medicine for a critical review of the manuscript.

Author Contributions

Conceived and designed the experiments: NK KI-Y HO. Performed the experiments: NK ST HM. Analyzed the data: SO AI TK NN KT. Contributed reagents/materials/analysis tools: SO AI TK NN KT. Wrote the paper: NK KI-Y HO.

References

1. Burmeister M, McInnis MG, Zöllner S (2008) Psychiatric genetics: progress amid controversy. Nat Rev Genet 9: 527–540.
2. Morrow EM, Yoo SY, Flavell SW, Kim TK, Lin Y, et al. (2008) Identifying autism loci and genes by tracing recent shared ancestry. Science 321: 218–223.
3. Bulayeva KB (2006) Overview of genetic-epidemiological studies in ethnically and demographically diverse isolates of Dagestan, Northern Caucasus, Russia. Croat Med J 47: 641–648.
4. Rudan I, Campbell H, Carothers AD, Hastie ND, Wright AF (2006) Contribution of consanguinuity to polygenic and multifactorial diseases. Nat Genet 38: 1224–1225.
5. Mansour H, Fathi W, Klei L, Wood J, Chowdari K, et al. (2010) Consanguinity and increased risk for schizophrenia in Egypt. Schizophr Res 120: 108–112.
6. McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, et al. (2008) Runs of homozygosity in European populations. Am J Hum Genet 83: 359–372.
7. The International HapMap Consortium (2003) The International HapMap Project. Nature 426: 789–796.
8. Hong H, Su Z, Ge W, Shi L, Perkins R, et al. (2008) Assessing batch effects of genotype calling algorithm BRLMM for the Affymetrix GeneChip Human Mapping 500 K array set using 270 HapMap samples. BMC Bioinformatics 9: S17.
9. BRLMM-P: a Genotype Calling Method for the SNP 5.0 Array. Available: http://www.affymetrix.com/support/technical/whitepapers/brlmmp_whitepaper.pdf. Accessed 14 Mar 2011.
10. R Development Core Team () R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2006). Available: http://www.R-project.org/. Accessed 14 Mar 2011.
11. Wand MP, Johnes MC (1995) Kernel Smoothing. Chapman and Hall, London.
12. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, et al. (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 40: 1253–1260.
13. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437: 1299–1320.
14. Yang H, Chang L, Huggins RM, Chen C, Mullighan CG (2011) LOHAS: loss-of-heterozygosity analysis suite. Genet Epidemiol. In press.

15. USCS Genome Bioinformatics, Golden Path Statistics, NCBI Build 36.1 assembly, March 2006 (hg 18). Available: http://genome.ucsc.edu/goldenPath/stats.html#hg18. Accessed 14 Mar 2011.
16. Straub RE, MacLean CJ, Ma Y, Webb BT, Myakishev MV, et al. (2002) Genome-wide scans of three independent sets of 90 Irish multiplex schizophrenia families and follow-up of selected regions in all families provides evidence for multiple susceptibility genes. Mol Psychiatry 7: 542–559.
17. Lencz T, Lambert C, DeRosse P, Burdick KE, Morgan TV, et al. (2007) Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. Proc Natl Acad Sci U S A 104: 19942–19947.
18. Affymetrix GeneChip Chromosome Copy Number Analysis Tool (CNAT) Version 4.0 User Guide (2007) Affymetrix Inc.
19. Imaizumi Y (1986) A recent survey of consanguineous marriages in Japan. Clin Genet 30: 230–233.
20. Al-Awadi SA, Moussa MA, Naguib KK, Farag TI, Teebi AS, et al. (1985) Consanguinity among the Kuwaiti population. Clin Genet 27: 483–486.
21. Pattison JE (2004) A comparison of inbreeding rates in India, Japan, Europe and China. Homo 55: 113–128.
22. Editorial (2006) The germinating seed of Arab genomics. Nat Genet 38: 851.
23. McInnes LA, Service SK, Reus VI, Barnes G, Charlat O, et al. (2001) Fine-scale mapping of a locus for severe bipolar mood disorder on chromosome 18p11.3 in the Costa Rican population. Proc Natl Acad Sci U S A 98: 11485–11490.
24. Mathews CA, Reus VI, Bejarano J, Escamilla MA, Fournier E, et al. (2004) Genetic studies of neuropsychiatric disorders in Costa Rica: a model for the use of isolated populations. Psychiatr Genet 14: 13–23.
25. Chen Q, Wang X, O'Neill FA, Walsh D, Kendler KS, et al. (2008) Is the histidine triad nucleotide-binding protein 1 (HINT1) gene a candidate for schizophrenia? Schizophr Res 106: 200–207.
26. Chen X, Wang X, Hossain S, O'Neill FA, Walsh D, et al. (2006) Haplotypes spanning SPEC2, PDZ-GEF2 and ACSL6 genes are associated with schizophrenia. Hum Mol Genet 15: 3329–3342.
27. Luo XJ, Diao HB, Wang JK, Zhang H, Zhao ZM, et al. (2008) Association of haplotypes spanning PDZ-GEF2, LOC728637 and ACSL6 with schizophrenia in Han Chinese. J Med Genet 45: 818–826.
28. O'Donovan MC, Craddock NJ, Owen MJ (2009) Genetics of psychosis; insights from views across the genome. Hum Genet 126: 3–12.
29. Schork NJ, Murray SS, Frazer KA, Topol EJ (2009) Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev 19: 212–219.
30. Gorlov I, Gorlova O, Frazier M, Spitz M, Amos C (2011) Evolutionary evidence of the effect of rare variants on disease etiology. Clin Genet 79: 199–206.
31. Ioannidis JPA, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. Nat Genet 29: 306–309.
32. Girard SL, Xiong L, Dion PA, Rouleau GA (2011) Where are the missing pieces of the schizophrenia genetics puzzle? Curr Opin Genet Dev. In press.
33. Glessner JT, Hakonarson H (2009) Common variants in polygenic schizophrenia. Genome Biol 10: 236.

The Breast Journal

LETTER TO THE EDITOR

Intracystic Papillary Carcinoma of Breast Harbors Significant Genomic Alteration Compared with Intracystic Papilloma: Genome-wide Copy Number and LOH Analysis Using High-Density Single-Nucleotide Polymorphism Microarrays

To the Editor:

Intracystic papillary breast tumors (ICPT) consist of benign papillomas, carcinomas in situ, and carcinomas with invasion, and they account for approximately 10% of benign breast tumors and less than 1% of malignant tumors, respectively (1,2). In breast lesions, indication for surgery is usually determined by pathological diagnosis together with radiologic findings, but differential, preoperative diagnosis of papillary carcinoma from papilloma is very difficult, even following needle biopsy (3) because of their nonspecific radiologic characteristics and their modest cytological and histologic appearance (4). To avoid excessive surgical intervention, another diagnostic procedure needs to be developed.

Cytogenetic studies of breast papillary tumors are limited, and cytogenetic differences between papillomas and papillary carcinomas are still controversial. Tsuda et al. (5,6) reported that papillary carcinomas have frequent changes in gene copy number and loss of heterozygosity (LOH), whereas papillomas did not show any gene copy number alteration or LOH at 16q and 1q. Boecker et al. (7) also reported that conventional comparative genomic hybridization (CGH) did not reveal any gene copy number change in papillomas. On the other hand, Lininger et al. (8) and Cristofano et al. (2) demonstrated that LOH at 16p or 16q was frequent in both papillomas and papillary carcinomas.

Address correspondence and reprint requests to: Koh-ichiro Yoshiura, MD, PhD, Department of Human Genetics, Nagasaki University Graduate School of Biomedical Science, 1-12-4 Sakamoto, Nagasaki 852-8523, Japan, or e-mail: kyoshi@nagasaki-u.ac.jp.

DOI: 10.1111/j.1524-4741.2011.01110.x

The Breast Journal, Volume 17 Number 4, 2011 427–430

The purpose of this study was to determine the profile of genomic alterations in breast ICPT and to explore the possibility of using high-density oligonucleotide SNP arrays as the basis of a novel diagnostic method of ICPT. Ten formalin-fixed paraffin-embedded (FFPE) breast ICPT were obtained from the Department of Pathology, Nagasaki University Hospital. The samples included five benign papillomas (Pap), three papillary carcinomas in situ (PurePC), and two papillary carcinomas with invasion (PCinv). Pathological diagnosis was independently determined by two pathologists. Clinicopathological findings of these tumors are provided in Fig. 1 and Table 1.

Extracted DNA from each sample was processed following the manufacturer's protocol and hybridized on Affymetrix GeneChip Genome-Wide Human SNP Array 5.0® (Affymetrix, Santa Clara, CA, USA). The QC call rates, which is an index measuring the quality of a SNP microarray experiment, obtained from the FFPE samples were from 70.75% to 91.93%, with a mean of 80.72% (Table 1), which was comparable to the results from former cytogenetic studies using DNA extracted from FFPE samples(9–11).

Copy number change and LOH analyses (called here SNPaCGH) were conducted using the Partek Genomics Suite (PGS) version 6.3 (Partek, St. Louis, MI, USA). To estimate the total rate of a copy number changed region, each segment amplified or lost was summed and divided by 2,829 Mb, which is the total Mb in the genome, excluding heterochromatic, centromeric, and telomeric regions not covered by probes. Similarly, to estimate the total rate of genomic alteration, the sum of segments with copy number change and copy number neutral loss of heterozygosity (CNLOH) was divided by 2,829 Mb. To validate the

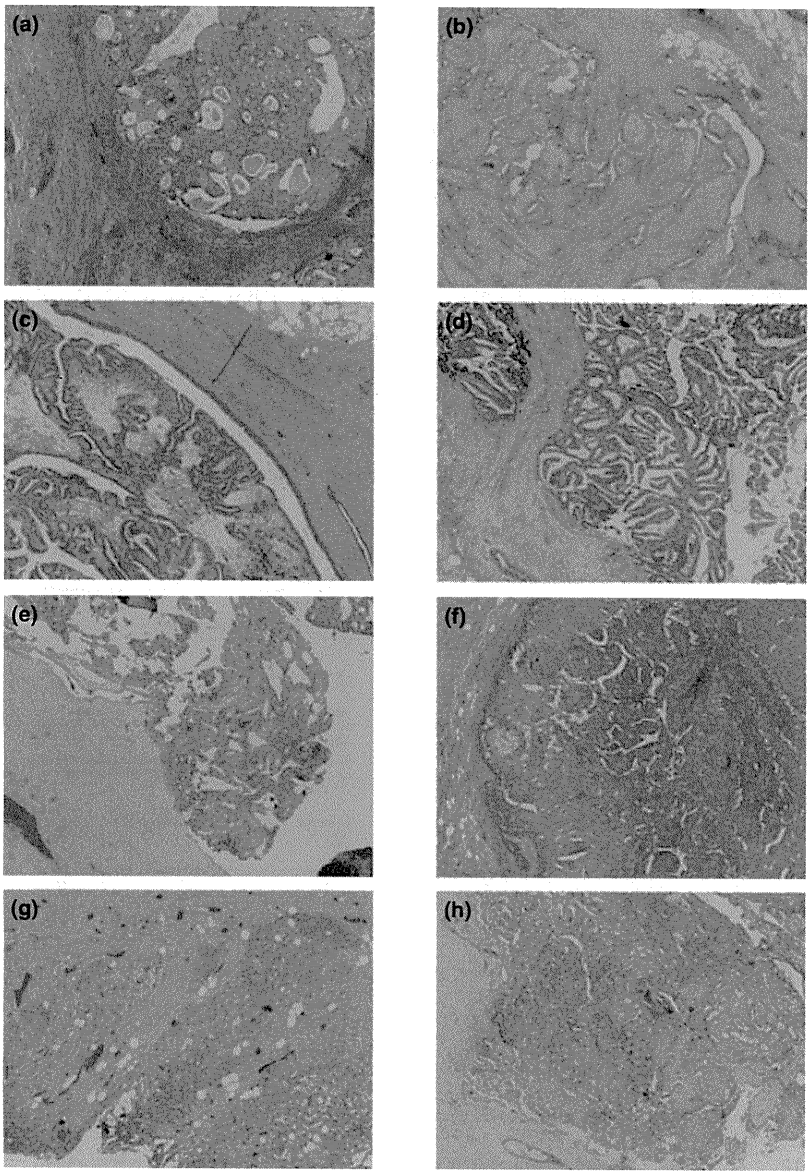

Figure 1. Hematoxylin-eosin stain in intracystic papillary tumors (Original magnification × 40). a–c, Intracystic papilloma (a: case 1, b: case 2, c: case 3). d–f, Intracystic papillary carcinoma in situ (d: case 6, e: case 7, f: case 8). g and h, Intracystic papillary carcinoma with invasion (g: case 9, h: case10).

Table 1. Characteristics of ten intracystic papillary lesions

	Clinicopathologic findings							Genetic findings		
Case	Diagnosis	Age	Size of cyst (mm)	MMG	US	FNAC	Receptor status	QC call rate (normal/tumor)	Ratio of CNC	Ratio of genomic alteration
1	Pap	43	80	Category 3	Category 3	Class 2	ER(+)	75.9%/82.9%	0.14%	0.24%
2	Pap	38	10	Category 1	Category 3	Class 3	NaN	83.4%/80.4%	0.66%	0.69%
3	Pap	49	25	Category 3	Category 3	Class 3	NaN	86.2%/86.5%	1.60%	1.60%
4	Pap	38	70	Category 3	Category 3	Class 2	NaN	89.9%/87.9%	0%	11.8%
5	Pap	49	75	Category 3	Category 3	Class 2	NaN	91.9%/89.8%	0%	0%
6	PurePC	61	31	Category 4	Category 3	Class 4	ER(+), HER2(1 +)	75.7%/76.2%	11.3%	24.1%
7	PurePC	58	49	Category 3	Category 4	Class 4	ER(+)	79.7%/70.8%	0.41%	8.83%
8	PurePC	43	16	Category 2	Category 4	Class 4	ER(+), HER2(1 +)	77.2%/79.9%	12.0%	13.2%
9	PCinv	60	96	NaN	Category 4	Class 1	ER(−), HER2(1 +)	71.6%/73.9%	16.6%	53.1%
10	PCinv	72	19	Category 4	Category 4	Class 5	ER(+), HER2(1 +)	82.0%/72.6%	16.0%	17.6%

Pap: intracystic papilloma, Pure PC: intracystic papillary carcinoma in situ, PC inv: intracystic papillary carcinoma with invasion, MMG: the mammographic features evaluated according to the Breast Imaging-Reporting and Data System (BI-RADS) of the American College of Radiology, US: the ultrasonographic features evaluated according to diagnostic guideline of the Japanese Association of Breast and Thyroid Sonology (JABTS), FNAC: the cytological features of fine needle aspiration cytology, ER: the status of estrogen receptor, HER2: the status of HER2/neu receptor, CNC: copy number change, genomic alteration: copy number change and copy neutral loss of heterozygosity, NaN: not analyzed.

copy number change identified by SNPaCGH, quantitative PCR assays were performed on a LightCycler® 480 Real-Time PCR System (Roche Diagnostics, Mannheim, Germany) at four selected loci, including independent genes (Table S1).

In SNPaCGH analysis, substantial divergence was observed between each ICPT subtype (Fig. 2). The mean rate of copy number change was 0.48% (from 0.0% to 1.60%), 7.89% (from 0.41% to 12.0%), and 16.3% (from 16.0% to 16.6%) in Pap, PC, and PCinv, respectively. The mean rate of genomic alteration (including copy number change and CNLOH) was 2.87% (from 0.00% to 11.8%), 15.4% (from 8.83% to 24.1%), and 35.3% (from 17.6% to 53.1%) in Pap, PC, and PCinv, respectively (Table 1). Malignant tumors (PurePC and PCinv) showed significantly more copy number changes and genomic alterations (copy number change and CNLOH) than benign tumors (Pap) (Wilcoxon's rank sum test, $p = 0.036$, 0.016, respectively) and these differences correlated with their malignant phenotype (Kruskal–Wallis' chi-squared test, $p = 0.046$, 0.043, respectively). The real time qPCR analysis to validate the copy number state in SNPaCGH demonstrated sufficient specificity, and thus all loci showing alteration in SNPaCGH were confirmed by real-time qPCR (Table S1). On the other hand, at 31 loci from the ten samples, where SNPaCGH showed the copy number state as disomy, ten loci were revealed to

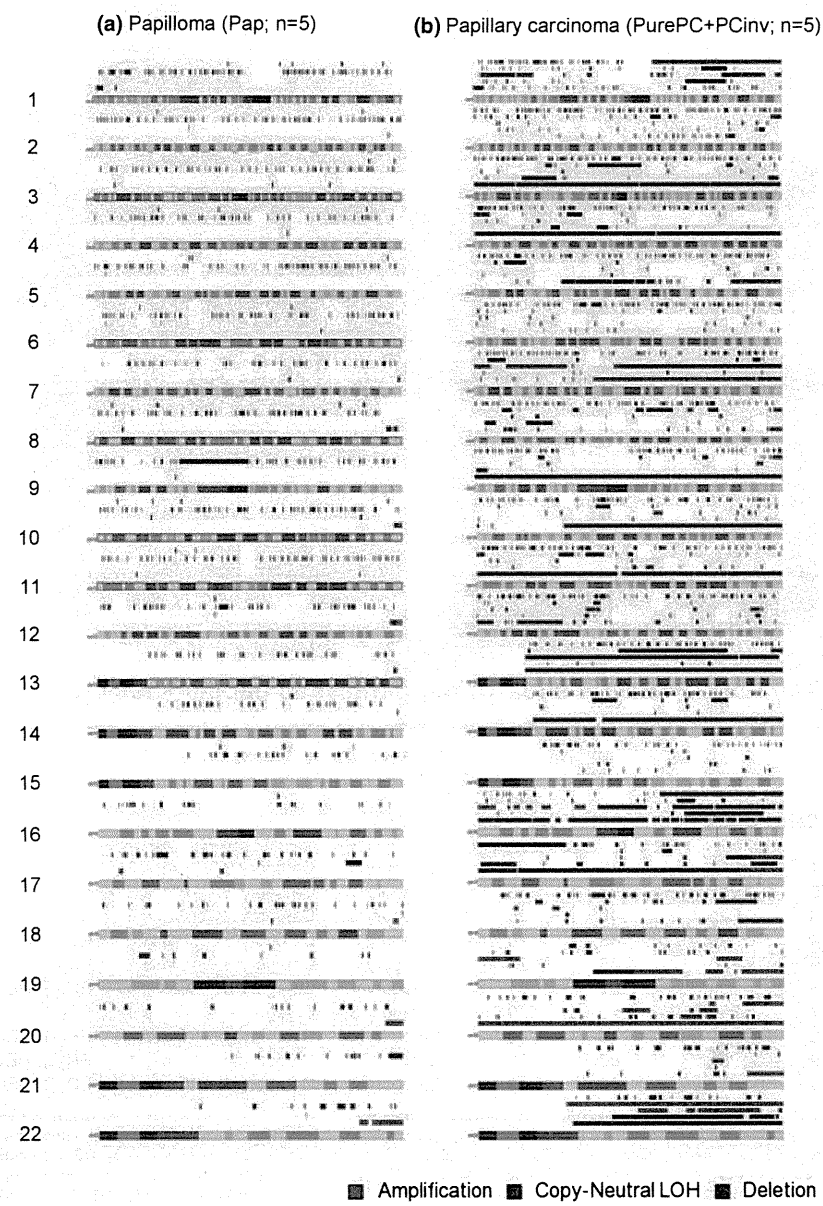

Figure 2. Graphic display of whole genomic alterations in papilloma (a) and papillary carcinoma (b). The color bar over each chromosome indicates copy number amplification (green color bars), copy-neutral LOH (blue color bars), and deletion (brown color bars) for each case. Papilloma includes five cases of Pap (a), and papillary carcinoma includes three cases of PurePC and two of PCinv (b).