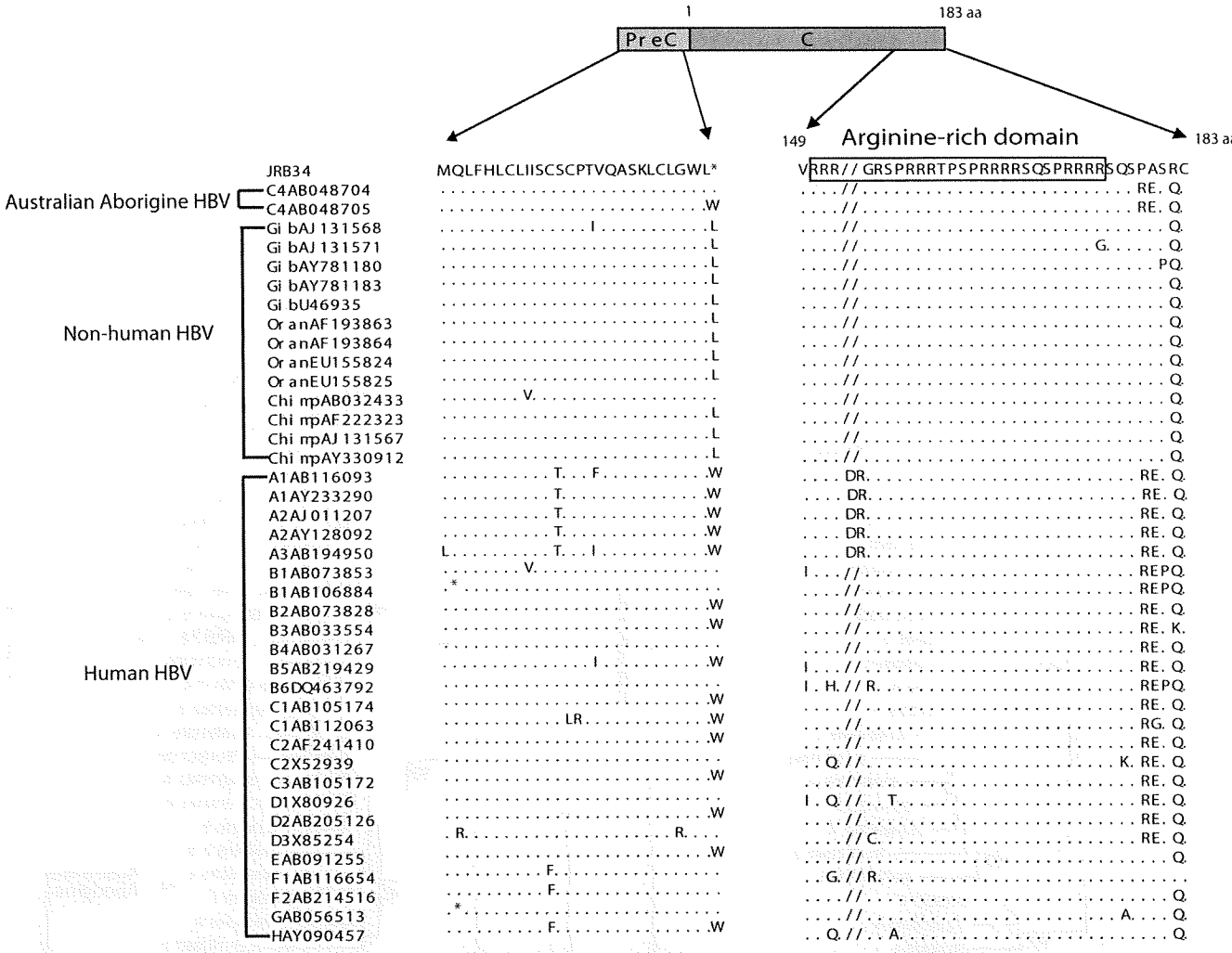

FIG. 4. Comparison of the amino acid sequence in the preC gene and carboxy-terminal amino acid sequences in the C gene of HBV isolates of various genotypes. The sequence of the HBV/J isolate (JRB34) is indicated at the top. Dots represent amino acids shared by JRB34, and a dash indicates the deletion of an amino acid. The sequence of the arginine-rich domain bearing the binding site with HBV DNA is boxed.

EMBL/DDBJ/GenBank database entries, the HBV/J strain was positioned distinctively from all known human genotypes (data not shown). It was closest to the cluster formed by gibbon- and orangutan-derived strains. However, including recombinant strains in such analyses may significantly affect the overall phylogenetic topology. This possibility was ruled out by reconstruction of the phylogeny using nonrecombinant HBV strains that further confirmed the phylogenetic peculiarity of the studied JRB34 strain (see Fig. S1 in the supplemental material). A total of 44 representative reference strains were further selected for establishing the consistency. Thus, phylogenetic topology indicating genotype-specific clustering is shown in the Fig. 1. Hence, using various sets of references, we confirmed that genotype J undoubtedly differed phylogenetically from all other known genotypes.

Lack of significant evidence of recombination with other human or ape genotypes in genotype J. To investigate possible recombination in the JRB34 genome, a window scanning analysis of aligned HBV genomes was performed by means of Simplot and Simmonics software packages. Both Bootscanning by SimPlot and GroupScanning by Simmonics showed similar output results. However, the methodological approach is different between these two software packages; GroupScanning provides more robust analysis of the phylogenetic relation between the examined strain and clusters of reference strains, whereas SimPlot does this comparison between the examined strain and parametrically generated consensus of the reference strains. The results obtained by SimPlot therefore can be significantly affected by selected parameters for the generation of consensus. This is especially undesirable when a new genotype strain (for which no references are available among known genotypes) is being analyzed (40). Figure 2 shows genome-wide distance scanning and GroupScanning plots for the JRB34 strain in comparison with a reference set consisting of 228 nonrecombinant HBV isolates retrieved from the public database (the phylogenetic tree is shown in Fig. S1 in the supplemental material). It is evident that the JRB34 strain was divergent from all known genotypes, and the closest genetic neighbors were estimated by distance and phylogenetic association scanning were the gibbon genotype (in preS, S, and P

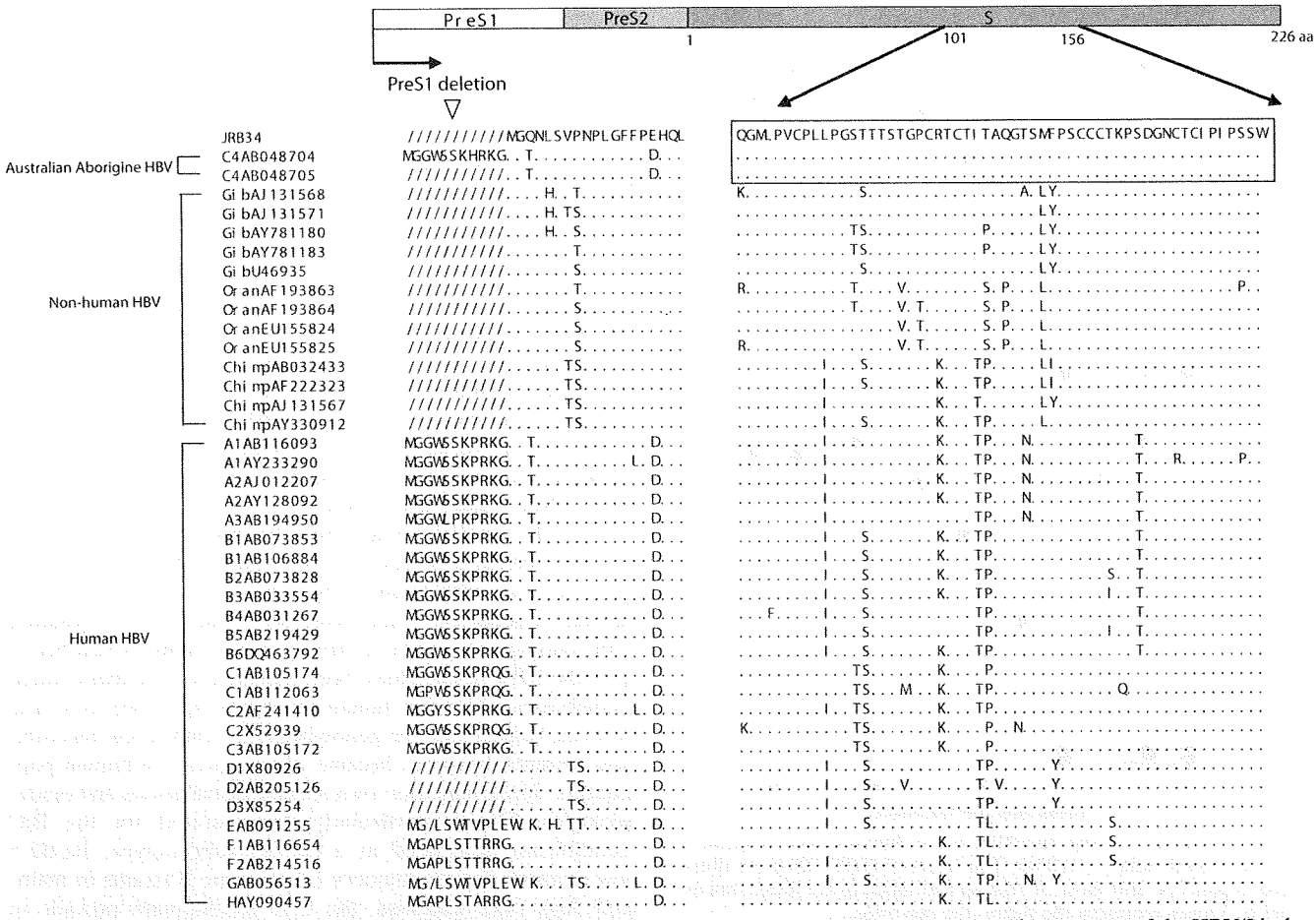

FIG. 5. Comparison of amino acid sequences of the preS/S gene among HBV isolates of various genotypes. The sequence of the HBV/J isolate (JRB34) is indicated at the top. Dots represent amino acids shared by JRB34, and a dash indicates the deletion of an amino acid. The sequence from positions 101 to 156 forming loops, bearing the common antigenic determinants of HBsAg, is boxed.

genes) and genotype C (in the core gene). However, no significant evidence of recombination between these two ape and human genotypes was revealed by the used methods. Homology scan carried out by SimPlot using the same set of reference sequences gave concordant results.

Phylogenetic analyses of the four open reading frames. Phylogenetic relationship between the JRB34 strain and other genotypes was further analyzed in four open reading frames. In the small S gene, subgenotype C4 recovered from Australian aborigines (43) changed its phylogenetic topology from the branch of human genotypes to a branch intermediate between orangutan and gibbon strains (Fig. 3A). Remarkably, genotype J and C4 strains joined together to create a clade between orangutan and gibbon strains. In contrast, genotype J clustered with human genotypes in the phylogenetic analysis of the C gene and was closely related to genotype C; it took a position outside genotype I strains, however (Fig. 3B). Genotype J was closer to gibbon and orangutan genotypes in the phylogenetic trees constructed on P and large S genes (data not shown), demonstrating its topology similar to that in the analysis of the entire genome (Fig. 1).

Amino acid sequence of the HBV/J isolate. The amino acid sequence of HBV/J was compared against those of other genotypes over three different areas of the genome. The amino acid sequence in the preC gene and arginine-rich domain in the carboxy-terminal sequence in the C gene were well conserved by genotype J (Fig. 4). In the preS1 region, genotype J had a deletion of 11 aa as gibbon and chimpanzee genotypes (Fig. 5). This deletion was shared by one of the two HBV/C4 isolates from Australian aborigines, as well as all HBV/D isolates. Amino acid sequence in the S gene of genotype J was the same as those of aborigine isolates of subgenotype C4; they would share antigenic epitopes of HBsAg. Amino acids at codons 122 and 160 were arginine (with G as nt 365) and lysine (with G as nt 479), respectively, which was consistent with subtype *ayw* of HBsAg from this patient (27).

Five domains (A to E) of DNA polymerase/reverse transcriptase in the P gene were preserved well in HBV/J, and it did not have mutations in the Tyr-Met-Asp-Asp motif in the domain C that determines the sensitivity to lamivudine (data not shown). HBV/J possessed A1762T/G1764A double mutations in the core promoter and G1896A stop codon mutation in the preC region, which was compatible with an HBeAg-minus phenotype of HBV recovered from the patient positive for anti-HBe.

Infection with HBV/J in chimeric mice with the liver repopulated for human hepatocytes. Two chimeric mice that had been transplanted with human hepatocytes were inoculated with 10^4 HBV DNA copies of genotype J. In both mice, HBV

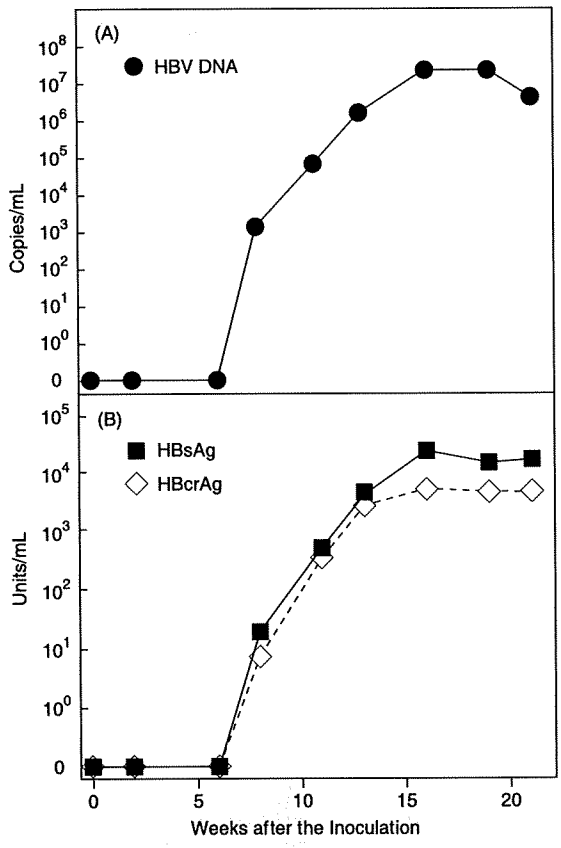

FIG. 6. Markers of HBV infection in two chimeric mice inoculated with the HBV/J isolate (JRB34). The levels of HBV DNA are illustrated in panel A, and those of HBsAg and HBcrAg are illustrated in panel B. Values represent the means for two mice.

DNA in a high titer (10^5 copies/ml) appeared in the circulation at week 7, plateaued at high levels (10^6 to 10^8 copies/ml), and stayed detectable until 22 weeks of observation after the inoculation (Fig. 6A). HBsAg and HBcrAg became detectable at week 7 and kept increasing in concentrations until week 15 when they reached a plateau at high levels (Fig. 6B). HBV strains recovered from mice at the last day of follow-up were identical in the complete genome sequence to the JRB34 strain used for inoculation.

The liver from chimeric mice infected with HBV/J was stained for HBcAg by immunofluorescence (Fig. 7A). The staining for HBcAg was confined to areas where mouse liver had been replaced for human hepatocytes, and the same areas were stained for human albumin (Fig. 7B). Colocalization of HBcAg and human hepatocytes was demonstrated by double staining for HBcAg and human albumin (Fig. 7C). Finally, expression and replication of the JRB34 strain were confirmed by successful detection of cccDNA and HBV RNA in the liver tissue from both sacrificed mice (see Fig. S2A and B in the supplemental material).

DISCUSSION

An HBV isolate (JRB34) was recovered from a male, 88-year-old Japanese patient with HCC and sequenced over the entire genome. In the full-genome sequence, the JRB34 strain had 10.9 to 15.7% divergence from 1,440 HBV strains retrieved from the DDBJ/EMBL/GenBank. The divergence exceeds 8% that has been defined originally for distinguishing between four genotypes (A to D) (29) and later for an additional four genotypes (E to H) (3, 26, 42). Phylogenetically, the sequence of JRB34 was closer to ape than human HBV genotypes. No significant evidence of recombination with eight known human and four ape genotypes was revealed by the GroupScanning analysis (40) and phylogenetic analyses. These lines of evidence have qualified the JRB34 strain to represent a possible new HBV genotype. To further confirm the epidemiological significance of this strain, capable of establishing new infections, two chimeric mice were each inoculated with 10^4 copies of JRB34 HBV DNA. They both were successfully infected with sharp increases in HBV DNA and HBsAg in serum several weeks after the inoculation. Replication in the chimeric mice was also confirmed by detection of cccDNA and HBV RNA in their liver tissues.

Recently, an HBV isolate from Vietnam (VH24 [accession no. AB231908]) was reported as a ninth human genotype (I) (12). However, VH24 differed by only 7.0% ± 0.4% from HBV isolates of genotype C and possessed complex recombination with genotypes A and G in three genomic areas. A number of sporadic HBV isolates have been reported to date that contain recombination between human genotypes (4, 24, 40), as well as between human and ape genotypes (21). Only a few recombinant variants, however, became widely spread in human populations, developing their own specific distributions and epidemiologies. This is particularly demonstrated for the B/C recombinant designated as a distinct subgenotype; Ba/B2-5 now accounts for the majority of genotype B strains in mainland Asia (44). Likewise, the C/D recombinant prevails in Tibet and northern China (50). To avoid assigning a new genotype for every newly discovered sporadic recombinant HBV variant, evidence of intergenotypic recombination should be carefully eliminated (14). However, in some cases, designation of a new genotype is proposed by a potential epidemiological significance of a novel genetic variant. Recently, a study carried out in Laos described a number of strains closely related phylogenetically with the Vietnamese genotype I strains, thereby suggesting their epidemiological significance (31). The JRB34 strain documented in the present study was genetically and phylogenetically distinct from any previously published strains, including those of genotype I from Vietnam and Laos. To avoid possible misconceptions in the future, the strain is provisionally designated genotype J.

HBV of distinct genotypes can infect great apes in the wild, including chimpanzee, gorilla, orangutan and gibbons (9, 20, 37, 51). HBV genotypes of chimpanzee and gorilla, as well as those of orangutan and gibbon, cocluster in agreement with their geographical distribution in Africa and Southeast Asia, respectively (41). Genotype J represented by the JRB34 strain clustered with gibbon/orangutan genotypes. In a phylogenetic analysis of the S region/gene sequence, JRB34 belonged to a nonhuman HBV group but was closely related to an HBV isolate of subgenotype C4 (AB048704) recovered from an Australian aborigine; C4 is most divergent from other subgenotypes of genotype C (43). In the phylogenetic analysis of the C gene, however, JRB34 clustered with human genotypes and closely related to genotype C, including C4, and was positioned

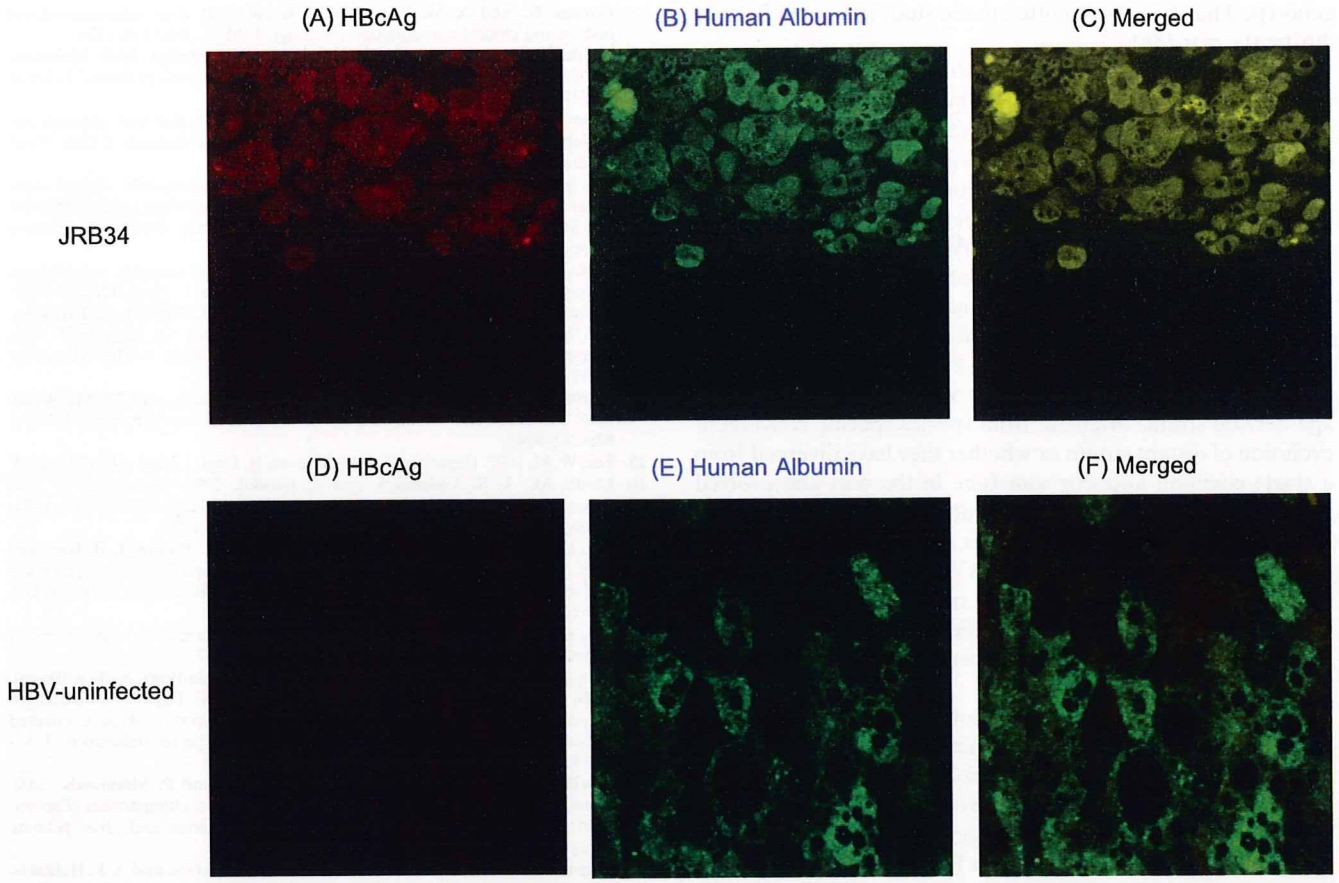

FIG. 7. (A and B) Immunofluorescent staining of a frozen liver section of a chimera mouse inoculated with the HBV/J isolate (JRB34). HBcAg is stained in panel A, and human albumin is stained in panel B. (C) Colocalization of HBcAg and human albumin is revealed by double staining. (D to F) HBV-uninfected mouse liver shows that only human albumin is stained.

outside genotype I strains (Fig. 4). Taken together, genotype J is phylogenetically close to gibbon/orangutan genotypes in the entire genome and to genotype C (C4 in particular) in the S and C genes. However, despite observed interchangeable relatedness with gibbon and genotype C/I strains, no strong evidence of recombination was confirmed in the JRB34.

In the sequence of C gene, carboxyl-terminal arginine-rich region, required for binding with HBV DNA, was preserved in JRB34. It had the G1896A stop codon in the precore region that aborts the translation of HBeAg (5, 30) and A1762T/G1764A double mutations in the core promoter that interfere with the transcription of HBeAg by downregulating preC mRNA (28, 45); they are compatible with the $HBeAg^-$ anti-HBe^+ phenotype of the patient from whom JRB34 was isolated. Since the double mutations are detected frequently in HBV DNA sequences from patients with HCC (17, 33), it could be implicated in hepatocarcinogenesis of the patient from whom JRB34 was isolated. It is not certain, however, if precore and core-promoter mutations had existed in HBV transmitted to the patient who is presumed to have been infected 60 years ago. Since amino acid sequences constituting antigenic loops of HBsAg (6) were the same as those of Australian aborigine isolates of C4, they would share antigenic epitopes of HBsAg. The amino acids at codons 122 and 160 were arginine (with G at nt 365) and lysine (with G at nt 479), respectively (27), in agreement with subtype *ayw* of HBsAg from this patient. Five domains (A to E) of DNA polymerase/reverse transcriptase in the P gene were preserved well in HBV/J, and it did not have mutations in the Tyr-Met-Asp-Asp motif in the domain C that determines the sensitivity to lamivudine (2).

How and when the patient contracted infection with HBV/J is not certain. It is very unlikely, however, that he acquired infection in Japan via perinatal or horizontal transmission. There are no wild primates in Okinawa, where the patient was originally from, and the prevalent human HBV genotypes are limited to B (60%), C (39%), and sporadic cases of A (1%) (32). Furthermore, HBV/J was not found among patient's family members who are currently alive (data not shown). The phylogenetic position within open reading frames of JRB34 in between gibbon/orangutan genotypes and human genotype C gives a clue where and when the patient had contracted HBV infection. He was drafted to Borneo during World War II (1939 to 1945); the island in the Southeast Asia is inhabited by gibbons and orangutans and has a local population mainly infected with genotypes B or C. Zoonotic infection of HBV has been previously reported (11, 46), and HBV of genotype E was recovered from a chimpanzee captured in West Africa where this genotype is common. There is a possibility that JRB34 of

genotype J had been transmitted to the study patient in Borneo during the war (38).

The origin of genotype J in gibbon/orangutan or human inhabitants in Borneo is not certain but very likely. HBV DNA and/or HBsAg was detected in 26% (55/213) and 20% (58/297) of gibbons and orangutans, respectively, captured in Southeast Asia (38). HBV is also endemic in people living there, with a prevalence of HBsAg at 2 to 8%. There would be high chances for cross-species transmission of HBV where it prevails both in human beings and nonhuman primates. Phylogenetic analysis for close relationship between human and nonhuman HBV genotypes has indicated geographical influence rather than association with particular species (41).

It is remains to be determined whether genotype J and ape-derived strains originate from species-specific convergent evolution of distant strains or whether they have diverged from a single common ancestor sometime in the past and evolved independently thereafter. The validity of cross-species infection or species-specific evolution for genotype J would be verified by sequence analysis of HBV DNA from gibbons and humans living in Borneo. If they turn out to be the same, cross-species infection will be justified. Should genotype J be restricted to human beings, in converse, species-specific infection will be confirmed.

In conclusion, a novel HBV genotype was identified in the Ryukyu isolate and provisionally named genotype J. Phylogenetic analyses over the full-length sequence and open reading frames indicate a close relationship of genotype J with gibbon/orangutan genotypes and human genotype C. The index patient would have been infected with HBV/J while he resided in Borneo inhabited by gibbons and orangutans. Although only one HBV isolate of genotype J (JRB34) has been identified, this may be only the tip of an iceberg. It would be worthwhile to examine the genotype of HBV infecting people and gibbons, as well as orangutans, living in Borneo and neighboring countries for mapping the epidemiology of genotype J and finding any clinical relevance.

ACKNOWLEDGMENTS

This study was supported in part by a grant-in-aid from the Ministry of Health, Labor and Welfare of Japan and a grant-in-aid from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

REFERENCES

1. **Abe, A., K. Inoue, T. Tanaka, J. Kato, N. Kajiyama, R. Kawaguchi, S. Tanaka, M. Yoshiba, and M. Kohara.** 1999. Quantitation of hepatitis B virus genomic DNA by real-time detection PCR. J. Clin. Microbiol. **37:**2899–2903.
2. **Allen, M. I., M. Deslauriers, C. W. Andrews, G. A. Tipples, K. A. Walters, D. L. Tyrrell, N. Brown, L. D. Condreay, et al.** 1998. Identification and characterization of mutations in hepatitis B virus resistant to lamivudine. Hepatology **27:**1670–1677.
3. **Arauz-Ruiz, P., H. Norder, B. H. Robertson, and L. O. Magnius.** 2002. Genotype H: a new Amerindian genotype of hepatitis B virus revealed in Central America. J. Gen. Virol. **83:**2059–2073.
4. **Bollyky, P. L., and E. C. Holmes.** 1999. Reconstructing the complex evolutionary history of hepatitis B virus. J. Mol. Evol. **49:**130–141.
5. **Carman, W. F., M. R. Jacyna, S. Hadziyannis, P. Karayiannis, M. J. McGarvey, A. Makris, and H. C. Thomas.** 1989. Mutation preventing formation of hepatitis B e antigen in patients with chronic hepatitis B infection. Lancet **ii:**588–591.
6. **Carman, W. F., A. R. Zanetti, P. Karayiannis, J. Waters, G. Manzillo, E. Tanzi, A. J. Zuckerman, and H. C. Thomas.** 1990. Vaccine-induced escape mutant of hepatitis B virus. Lancet **336:**325–329.
7. **Fung, S. K., and A. S. Lok.** 2004. Hepatitis B virus genotypes: do they play a role in the outcome of HBV infection? Hepatology **40:**790–792.
8. **Ganem, D., and A. M. Prince.** 2004. Hepatitis B virus infection–natural history and clinical consequences. N. Engl. J. Med. **350:**1118–1129.
9. **Grethe, S., J. O. Heckel, W. Rietschel, and F. T. Hufert.** 2000. Molecular epidemiology of hepatitis B virus variants in nonhuman primates. J. Virol. **74:**5377–5381.
10. **Hannoun, C., H. Norder, and M. Lindh.** 2000. An aberrant genotype revealed in recombinant hepatitis B virus strains from Vietnam. J. Gen. Virol. **81:**2267–2272.
11. **Hu, X., A. Javadian, P. Gagneux, and B. H. Robertson.** 2001. Paired chimpanzee hepatitis B virus (ChHBV) and mtDNA sequences suggest different ChHBV genetic variants are found in geographically distinct chimpanzee subspecies. Virus Res. **79:**103–108.
12. **Huy, T. T. T., T. N. Trinh, and K. Abe.** 2008. New complex recombinant genotype of hepatitis B virus identified in Vietnam. J. Virol. **82:**5657–5663.
13. **Kimura, T., A. Rokuhara, Y. Sakamoto, S. Yagi, E. Tanaka, K. Kiyosawa, and N. Maki.** 2002. Sensitive enzyme immunoassay for hepatitis B virus core-related antigens and their correlation to virus load. J. Clin. Microbiol. **40:**439–445.
14. **Kurbanov, F., Y. Tanaka, A. Kramvis, P. Simmonds, and M. Mizokami.** 2008. When should "I" consider a new hepatitis B virus genotype? J. Virol. **82:**8241–8242.
15. **Lee, W. M.** 1997. Hepatitis B virus infection. N. Engl. J. Med. **337:**1733–1745.
16. **Lindh, M., A. S. Andersson, and A. Gusdal.** 1997. Genotypes, nt 1858 variants, and geographic origin of hepatitis B virus: large-scale analysis using a new genotyping method. J. Infect. Dis. **175:**1285–1293.
17. **Liu, C. J., B. F. Chen, P. J. Chen, M. Y. Lai, W. L. Huang, J. H. Kao, and D. S. Chen.** 2006. Role of hepatitis B viral load and basal core promoter mutation in hepatocellular carcinoma in hepatitis B carriers. J. Infect. Dis. **193:**1258–1265.
18. **Liu, C. J., J. H. Kao, and D. S. Chen.** 2005. Therapeutic implications of hepatitis B virus genotypes. Liver Int. **25:**1097–1107.
19. **Lole, K. S., R. C. Bollinger, R. S. Paranjape, D. Gadkari, S. S. Kulkarni, N. G. Novak, R. Ingersoll, H. W. Sheppard, and S. C. Ray.** 1999. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J. Virol. **73:**152–160.
20. **MacDonald, D. M., E. C. Holmes, J. C. Lewis, and P. Simmonds.** 2000. Detection of hepatitis B virus infection in wild-born chimpanzees (*Pan troglodytes verus*): phylogenetic relationships with human and other primate genotypes. J. Virol. **74:**4253–4257.
21. **Magiorkinis, E. N., G. N. Magiorkinis, D. N. Paraskevis, and A. E. Hatzakis.** 2005. Re-analysis of a human hepatitis B virus (HBV) isolate from an East African wild born *Pan troglodytes schweinfurthii*: evidence for interspecies recombination between HBV infecting chimpanzee and human. Gene **349:** 165–171.
22. Reference deleted.
23. **Miyakawa, Y., and M. Mizokami.** 2003. Classifying hepatitis B virus genotypes. Intervirology **46:**329–338.
24. **Morozov, V., M. Pisareva, and M. Groudinin.** 2000. Homologous recombination between different genotypes of hepatitis B virus. Gene **260:**55–65.
25. **Norder, H., A. M. Courouce, P. Coursaget, J. M. Echevarria, S. D. Lee, I. K. Mushahwar, B. H. Robertson, S. Locarnini, and L. O. Magnius.** 2004. Genetic diversity of hepatitis B virus strains derived worldwide: genotypes, subgenotypes, and HBsAg subtypes. Intervirology **47:**289–309.
26. **Norder, H., A. M. Courouce, and L. O. Magnius.** 1994. Complete genomes, phylogenetic relatedness, and structural proteins of six strains of the hepatitis B virus, four of which represent two new genotypes. Virology **198:**489–503.
27. **Okamoto, H., M. Imai, F. Tsuda, T. Tanaka, Y. Miyakawa, and M. Mayumi.** 1987. Point mutation in the S gene of hepatitis B virus for a *d/y* or *w/r* subtypic change in two blood donors carrying a surface antigen of compound subtype *adyr* or *adwr*. J. Virol. **61:**3030–3034.
28. **Okamoto, H., F. Tsuda, Y. Akahane, Y. Sugai, M. Yoshiba, K. Moriyama, T. Tanaka, Y. Miyakawa, and M. Mayumi.** 1994. Hepatitis B virus with mutations in the core promoter for an e antigen-negative phenotype in carriers with antibody to e antigen. J. Virol. **68:**8102–8110.
29. **Okamoto, H., F. Tsuda, H. Sakugawa, R. I. Sastrosoewignjo, M. Imai, Y. Miyakawa, and M. Mayumi.** 1988. Typing hepatitis B virus by homology in nucleotide sequence: comparison of surface antigen subtypes. J. Gen. Virol. **69**(Pt. 10):2575–2583.
30. **Okamoto, H., S. Yotsumoto, Y. Akahane, T. Yamanaka, Y. Miyazaki, Y. Sugai, F. Tsuda, T. Tanaka, Y. Miyakawa, and M. Mayumi.** 1990. Hepatitis B viruses with precore region defects prevail in persistently infected hosts along with seroconversion to the antibody against e antigen. J. Virol. **64:** 1298–1303.
31. **Olinger, C. M., P. Jutavijittum, J. M. Hubschen, A. Yousukh, B. Samountry, T. Thammavong, K. Toriyama, and C. P. Muller.** 2008. Possible new hepatitis B virus genotype, southeast Asia. Emerg. Infect. Dis. **14:**1777–1780.
32. **Orito, E., T. Ichida, H. Sakugawa, M. Sata, N. Horiike, K. Hino, K. Okita, T. Okanoue, S. Iino, E. Tanaka, K. Suzuki, H. Watanabe, S. Hige, and M. Mizokami.** 2001. Geographic distribution of hepatitis B virus (HBV) genotype in patients with chronic HBV infection in Japan. Hepatology **34:**590–594.

33. **Orito, E., M. Mizokami, H. Sakugawa, K. Michitaka, K. Ishikawa, T. Ichida, T. Okanoue, H. Yotsuyanagi, and S. Iino.** 2001. A case-control study for clinical and molecular biological differences between hepatitis B viruses of genotypes B and C. Hepatology **33:**218–223.
34. **Palumbo, E.** 2007. Hepatitis B genotypes and response to antiviral therapy: a review. Am. J. Ther. **14:**306–309.
35. **Saitou, N., and M. Nei.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:**406–425.
36. **Sakamoto, T., Y. Tanaka, E. Orito, J. Co, J. Clavio, F. Sugauchi, K. Ito, A. Ozasa, A. Quino, R. Ueda, J. Sollano, and M. Mizokami.** 2006. Novel subtypes (subgenotypes) of hepatitis B virus genotypes B and C among chronic liver disease patients in the Philippines. J. Gen. Virol. **87:**1873–1882.
37. **Sall, A. A., S. Starkman, J. M. Reynes, S. Lay, T. Nhim, M. Hunt, N. Marx, and P. Simmonds.** 2005. Frequent infection of *Hylobates pileatus* (pileated gibbon) with species-associated variants of hepatitis B virus in Cambodia. J. Gen. Virol. **86:**333–337.
38. **Sa-nguanmoo, P., C. Thongmee, P. Ratanakorn, R. Pattanarangsan, R. Boonyarittichaikij, S. Chodapisitkul, A. Theamboonlers, P. Tangkijvanich, and Y. Poovorawan.** 2008. Prevalence, whole genome characterization and phylogenetic analysis of hepatitis B virus in captive orangutan and gibbon. J. Med. Primatol. **37:**277–289.
39. **Shin-I, T., Y. Tanaka, Y. Tateno, and M. Mizokami.** 2008. Development and public release of a comprehensive hepatitis virus database. Hepatol. Res. **38:**234–243.
40. **Simmonds, P., and S. Midgley.** 2005. Recombination in the genesis and evolution of hepatitis B virus genotypes. J. Virol. **79:**15467–15476.
41. **Starkman, S. E., D. M. MacDonald, J. C. Lewis, E. C. Holmes, and P. Simmonds.** 2003. Geographic and species association of hepatitis B virus genotypes in non-human primates. Virology **314:**381–393.
42. **Stuyver, L., S. De Gendt, C. Van Geyt, F. Zoulim, M. Fried, R. F. Schinazi, and R. Rossau.** 2000. A new genotype of hepatitis B virus: complete genome and phylogenetic relatedness. J. Gen. Virol. **81:**67–74.
43. **Sugauchi, F., M. Mizokami, E. Orito, T. Ohno, H. Kato, S. Suzuki, Y. Kimura, R. Ueda, L. A. Butterworth, and W. G. Cooksley.** 2001. A novel variant genotype C of hepatitis B virus identified in isolates from Australian Aborigines: complete genome sequence and phylogenetic relatedness. J. Gen. Virol. **82:**883–892.
44. **Sugauchi, F., E. Orito, T. Ichida, H. Kato, H. Sakugawa, S. Kakumu, T. Ishida, A. Chutaputti, C. L. Lai, R. Ueda, Y. Miyakawa, and M. Mizokami.** 2002. Hepatitis B virus of genotype B with or without recombination with genotype C over the precore region plus the core gene. J. Virol. **76:**5985–5992.
45. **Takahashi, K., K. Aoyama, N. Ohno, K. Iwata, Y. Akahane, K. Baba, H. Yoshizawa, and S. Mishiro.** 1995. The precore/core promoter mutant (T1762A1764) of hepatitis B virus: clinical significance and an easy method for detection. J. Gen. Virol. **76**(Pt. 12)**:**3159–3164.
46. **Takahashi, K., B. Brotman, S. Usuda, S. Mishiro, and A. M. Prince.** 2000. Full-genome sequence analyses of hepatitis B virus (HBV) strains recovered from chimpanzees infected in the wild: implications for an origin of HBV. Virology **267:**58–64.
47. **Tanaka, Y., and M. Mizokami.** 2007. Genetic diversity of hepatitis B virus as an important factor associated with differences in clinical outcomes. J. Infect. Dis. **195:**1–4.
48. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:**4673–4680.
49. **Tiollais, P., P. Charnay, and G. N. Vyas.** 1981. Biology of hepatitis B virus. Science **213:**406–411.
50. **Wang, Z., Z. Liu, G. Zeng, S. Wen, Y. Qi, S. Ma, N. V. Naoumov, and J. Hou.** 2005. A new intertype recombinant between genotypes C and D of hepatitis B virus identified in China. J. Gen. Virol. **86:**985–990.
51. **Wiegand, J., D. Hasenclever, and H. L. Tillmann.** 2008. Should treatment of hepatitis B depend on hepatitis B virus genotypes? A hypothesis generated from an explorative analysis of published evidence. Antivir. Ther. **13:**211–220.
52. **Wong, D. K., Y. Tanaka, C. L. Lai, M. Mizokami, J. Fung, and M. F. Yuen.** 2007. Hepatitis B virus core-related antigens as markers for monitoring chronic hepatitis B infection. J. Clin. Microbiol. **45:**3942–3947.

Journal of Human Genetics (2009) 54, 543–546

www.nature.com/jhg

SHORT COMMUNICATION

Genome-wide association database developed in the Japanese Integrated Database Project

Asako Koike[1], Nao Nishida[2], Ituro Inoue[3], Shoji Tsuji[4] and Katsushi Tokunaga[2]

The establishment of high-throughput single-nucleotide polymorphism (SNP)-typing technologies has enabled astonishing progress to be made in genome-wide association studies (GWAS), and various novel genetic factors associated with complex diseases have been discovered. Our organization has created a public repository database (DB) to achieve a continuous and intensive management of GWAS data and to facilitate data sharing among researchers. In the GWAS DB, information on study design, quality control protocols, allele frequencies, genotype frequencies and statistical genetic analysis results are stored as publicly available data and can be accessed freely, whereas individual genotyping data and raw data are stored as restricted data and can only be accessed with authorization. All data are presented by a graphic viewer, which is designed to be user friendly for researchers who are not familiar with GWAS to accelerate disease-related studies. Furthermore, the DB allows users to compare various study results obtained by different institutions and on different platforms. The same data are also managed as a distributed annotation system to call up useful data from other DBs and to superimpose them on the GWAS data for help in interpretation. The DB is accessible at https://gwas.lifesciencedb.jp/.
Journal of Human Genetics (2009) **54,** 543–546; doi:10.1038/jhg.2009.68; published online 24 July 2009

Keywords: database; genome-wide association; SNP

INTRODUCTION

The accomplishment of sequencing of the entire human genome[1,2] and the HapMap project,[3] coupled with the development of cost-effective high-throughput dense single-nucleotide polymorphism (SNP)-typing techniques, have enabled a genome-wide exploration of various complex disease-associated variants. Currently, the high-throughput SNP-typing methods are expected to cover about 80% of the human genome in linkage disequilibrium.[4] A number of large-scale genome-wide cohort studies and case–control studies, such as seven common disease GWAS by the Wellcome Trust Case Control Consortium (WTCCC, 2007), have been planned, and some of them are underway. So far, more than 100 loci of disease-related/causing candidates for about 40 common diseases and traits have been identified,[5] and some loci have led to new insights into pathophysiology and etiological pathways. Because GWAS yields large amounts of raw data and analysis results, the management of GWAS data has become a matter of serious concern. Furthermore, more and more grant-funding agencies, journal editors and research communities are beginning to require the disclosure of GWAS data. Disclosure and data sharing of GWAS data will primarily lead to the following three possibilities: (1) meta-analysis using data sets produced in multiple studies to find novel disease-related SNP candidates; (2) re-use of GWAS data combined with other experimental data, including pathway data and expression data, to deepen the exploration of each disease; and (3) development of methods to analyze and compute genetic statistics. In the case of meta-analysis in particular, the use of raw data is indispensable for quality control and for consideration of population structures. Some studies have successfully found additional disease-related SNP candidates on the basis of meta-analysis.[6,7]

The National Center for Biotechnology Information launched the database (DB) of Genotype and Phenotype in the fall of 2006 as a centralized GWAS system to archive and distribute GWAS data. Currently, results funded by the Genetic Association Information Network and voluntarily submitted data have been accumulated. The European Genotype Archive was created in the spring of 2008 as a repository system for phenotype–genotype relationships, and results primarily from WTCCC have been accumulated and redistributed. To achieve a continuous and intensive management of GWAS data and data sharing among researchers, we established a new DB that is publicly available. This DB is expected to have an essential role in providing easily accessible GWAS data to researchers in various biomedical fields. Some disease-related SNPs are assumed to be buried because of their insufficient *P*-values caused by an insufficient number of case–control samples. It is possible that these SNPs will be revealed by combining the GWAS analysis results with other data possessed by users.

In this paper, we introduce the GWAS DB.

[1]Central Research Laboratory, Hitachi Ltd, Tokyo, Japan; [2]Department of Human Genetics, Graduate School of Medicine, University of Tokyo, Tokyo, Japan; [3]Department of Molecular Life Science and Molecular Medicine, Tokai University School of Medicine, Tokyo, Japan and [4]Department of Neurology, Graduate School of Medicine, University of Tokyo, Tokyo, Japan
Correspondence: Dr A Koike, Central Research Laboratory, Hitachi Ltd, 1-280 Higashi-koigakubo Kokubunji, Tokyo, Japan.
E-mail: asako.koike.ea@hitachi.com
Received 3 June 2009; accepted 27 June 2009; published online 24 July 2009

MATERIALS AND METHODS

Database structure

The DB system consists of an internal GWAS DB and a public GWAS DB. For a maximum of 1 year, or until the acceptance of publication, submitted data are stored in the internal GWAS DB and can be accessed only by the research team that submitted the data for greater convenience in data sharing among research team members living in various locations. Currently, the DB systems are implemented using mysql version 5.0 (http://dev.mysql.com/downloads/mysql/5.0.html), and some of the statistical analysis results are also accumulated in a distributed annotation system (DAS) server. A schematic drawing of the GWAS DB is shown in Figure 1.

In this DB, three types of data access, namely, (1) public access, (2) authorized access accompanied by a data use application, and (3) authorized access accompanied by a data use application and its review by a data access committee, are possible. Principally, frequency data of genotypes and alleles and statistical analysis results can be accessed freely. However, automatic access and frequent access are restricted to prevent the release of frequency data of genome-wide genotypes and alleles, as such a large volume of genotype/allele data leads to the specification of whether the given genome is contained in the case or in the control group, as reported previously.[8] These genome-wide frequency data can be obtained by submitting a data use application to the data access committee. For the use of genotype or raw data, an application that

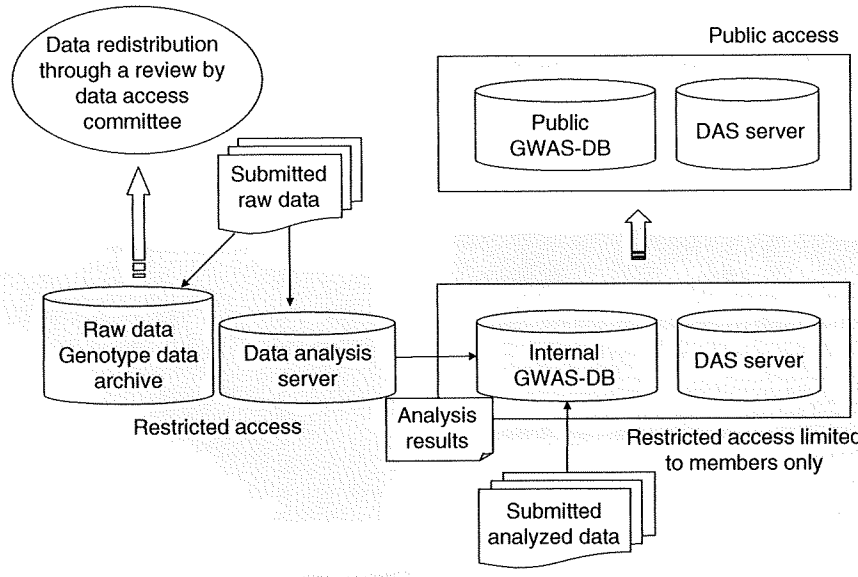

Figure 1 Schematic drawing of genome-wide association study (GWAS) database (DB) systems.

Table 1 Summary of database contents

Contents	*Data sources*
Statistics	
Frequencies of genotypes, alleles and haplotypes	
Statistical genetic analysis	
P-values and odds ratios on genotypic model and allelic model	
P-values and odds ratios on trend model, additive model and recessive model	
Permutation test results	
Bonferroni's corrections and false discovery rate for multiple testing using	
Akaike information criterion	
Hardy–Weinberg equilibrium test	
Haplotype-based χ^2-test	
Epistasis	
Linkage disequilibrium parameters (r^2, D', Lod)	
Other data	
mRNA, amino-acid sequence of each gene	NCBI (http://www.ncbi.nlm.nih.gov/)
mRNA, genome-mapped position	UCSC Hg. 18 (http://hgdownload.cse.ucsc.edu/)
SNP position and SNP kind (cSNP, sSNP, rSNP and so on)	NCBI (http://www.ncbi.nlm.nih.gov/)
OMIM	NCBI (http://www.ncbi.nlm.nih.gov/)
Copy number variation	DGV (http://projects.tcag.ca/variation/)
Gene function	Gene ontology (http://www.geneontology.org/)
Microsatellite polymorphism	UCSC (http://hgdownload.cse.ucsc.edu/)
Manually curated disease-related mutation information	

describes the research purpose and lists the research team members must be submitted to the data access committee. The data access committee deliberates on whether the applicant's research purpose meets the content of the consent form. Only applicants approved by the review committee can use individual genotype data and raw data in accordance with the data handling security rules required by the data access committee and following data use restrictions on the basis of informed consent.

Individual data and raw data are accumulated in the server in a secured computer environment that is different from the public DB server. Only authorized persons can access this server.

Data submission

In principal, both analysis results and unanalyzed data can be submitted. When data have already been analyzed, the analyzed data are accumulated in this DB, along with a detailed description of the analysis protocols. When data have not been analyzed yet, they are analyzed in our site, and the results are accumulated in this DB. When raw data are redistributable under certain conditions, they are also submitted with the contents of the consent form. All data must be submitted with documents explaining the design of the study, as well as ethical consideration.

Data cleaning for quality control

When data are submitted as individual data without analysis results, they are analyzed as follows: (1) SNPs with a call rate $<95\%$ and samples with a call rate $<95\%$ are removed. (2) SNPs, the Hardy–Weinberg equilibrium test result of which in a control group is less than 0.001 or the minor allele frequency of which is less than 0.05, are removed. (3) The principle component analysis (PCA) of these case–control data, along with HapMap data, is carried out using EIGENSTRAT[9] or other programs so that sample outliers and samples with a possible ethnic mixture or a different ethnicity are removed on the basis of the PCA result. Sample outliers in the plot of heterozygosity versus call rate are also removed. The quantile–quantile plot based on the allelic model is calculated and checked. When only genotype frequency data are submitted, PCA and heterozygosity checks are skipped,

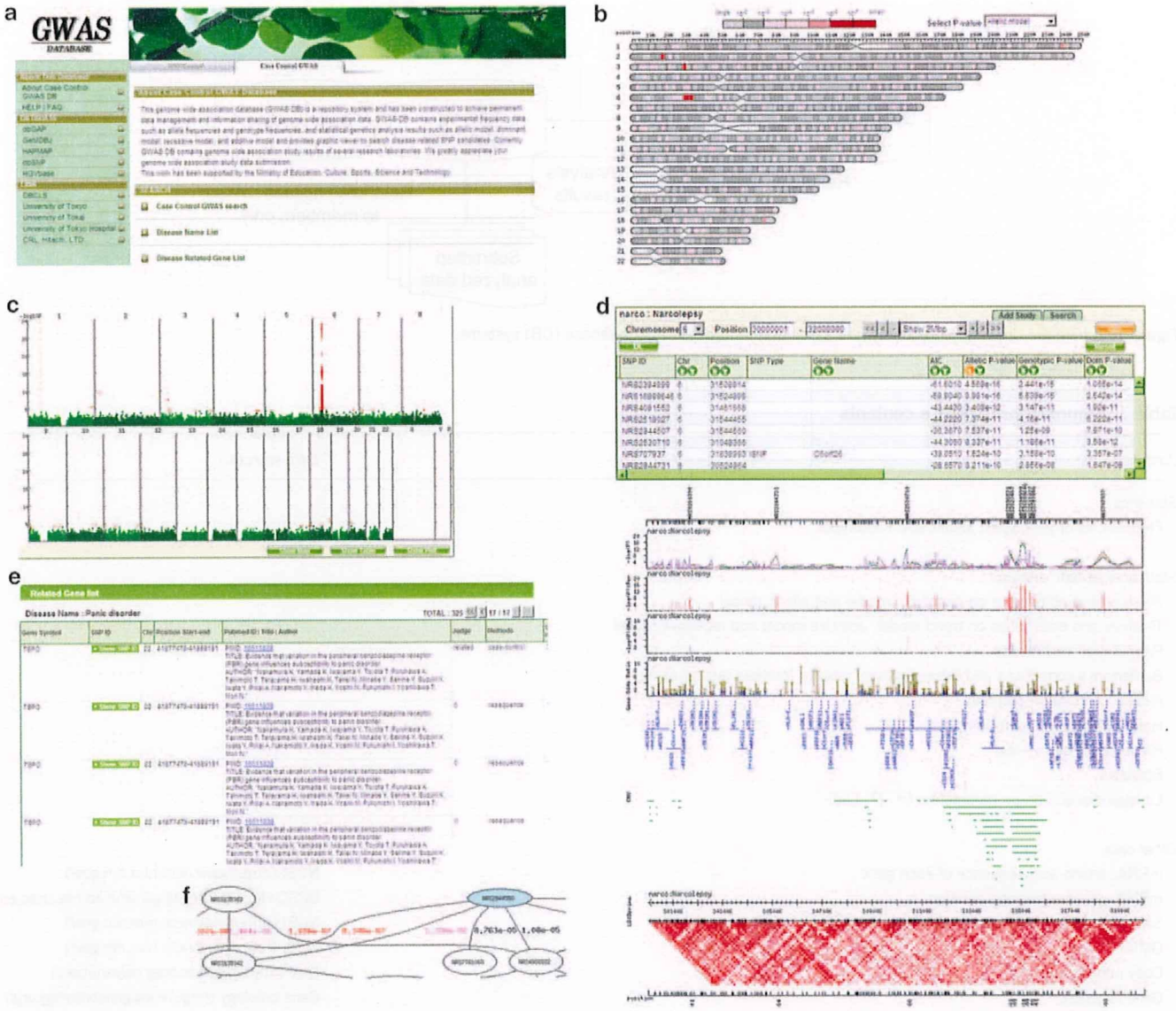

Figure 2 Snapshots of the genome-wide association study (GWAS) database. (**a**) Top page, (**b**) bird's-eye view, (**c**) Manhattan plot, (**d**) region table and graph, (**e**) disease-related gene/single-nucleotide polymorphism (SNP) lists (public data) and (**f**) SNP network based on epistasis.

as they require individual data. The cleaning results are linked from 'study details' on the web.

Data analysis

Standard statistical genetic analyses are performed by plink[10] and Haploview.[11] Additional analyses such as the Akaike information criterion, epistasis and more complicated ones (for example, genetic analysis considering potential case samples existing in the control samples, which sometimes becomes a concern for diseases that develop in old age) are calculated by internally developed programs. The major statistics include *P*-values based on an allelic model, genotypic model, trend model, dominant model, recessive model and permutation test results of these models, and Bonferroni's correction and false discovery rate for multiple testing. These methods are also shown in 'study details.' When submitted data consist of only genotype frequency data, the genome-wide permutation test is skipped.

Database contents and utility

The DB contents (as of April 2009) are summarized in Table 1.

User data other than GWAS data, such as expression data and epigenetic data, are also accumulated and can be displayed on the graph. Although clinical data are not currently accumulated in the DB, they can be added if submitted. Major tables are summarized in Supplementary Table 1.

A snapshot of the GWAS DB is shown in Figure 2. Figure 2a shows the top page of the GWAS DB. When the 'SNP control' tab is selected, the interface jumps to the SNP control DB, which is affiliated to the GWAS DB and contains allelic frequencies, genotypic frequencies, Hardy–Weinberg equilibrium tests and estimated haplotype frequencies of Japanese control samples. Bird's-eye view (Figure 2b) and Manhattan plot (Figure 2c) are provided to draw *P*-values of each model. A genome region can be selected from both (Figures 2b and c), and the results of statistical genetic analysis along with other information such as exon–intron information and copy number variations (CNVs) can be displayed in tables and graphs to facilitate the identification of disease-related SNPs, as shown in Figure 2d. Furthermore, comparisons among various study results obtained by different institutions and/or different platforms can be carried out easily by plotting their graphs on the web (using the 'add study' function in Figure 2d). When the published disease-related gene or SNP is registered as shown in Figure 2e, data are plotted as a known disease-related gene/SNP in the graph (Figure 2d). Epistasis data are also accumulated and drawn as a network graph using Graphviz (http://www.graphviz.org/), as shown in (Figure 2f). Data can be searched by SNP ID (dbSNP ID #rs, affymetrix SNP ID and so on), gene name, disease name and so on. The study design and analysis protocols can also be browsed.

Statistical results are also accumulated on a DAS server, and they can be browsed using the Gmod Gbrowse (http://gmod.org/wiki/Main_Page)-based browser (http://gwas.lifesciencedb.jp/cgi-bin/gbrowse/snpdb/). Furthermore, as a function of the DAS server, data on other DAS servers such as Ensemble can be called up. This function is useful to superimpose data from other DBs onto GWAS data. The GWAS DB is designed to be user friendly for researchers unfamiliar with GWAS to promote disease-related studies.

Further development

A recent topic of interest is genome-wide association analysis coupled with other data such as pathway data[12] to compensate for the low statistical power in disease-associated candidate SNPs. The function to browse or calculate SNP/SNP pair *P*-values on the basis of the GWAS result, along with other data, will be added to this DB to facilitate the generation and understanding of user hypotheses.

The relationships between CNVs and diseases have begun to emerge in recent studies.[13] Although concerns remain about the quality of detected CNVs, genomic locations and frequencies of CNV regions and their case–control association study results will be incorporated into this DB. Furthermore, in the near feature, new high-throughput techniques such as short-read sequencing will be applied for GWAS, and this DB will be improved to suit the new experimental techniques.

ACKNOWLEDGEMENTS

This work was supported by the contract research fund 'Integrated Database Project' from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

1 Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
2 Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
3 The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
4 Barrett, J. C. & Cardon, L. R. Evaluating coverage of genome-wide association studies. *Nat. Genet.* **38,** 659–662 (2006).
5 Manolio, T. A., Brooks, L. D. & Collins, F. S. A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.* **118,** 1590–1605 (2008).
6 Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.* **40,** 638–645 (2008).
7 Houlston, R. S., Webb, E., Broderick, P., Pittman, A. M., Di Bernardo, M. C., Lubbe, S. *et al.* Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat. Genet.* **40,** 1426–1435 (2008).
8 Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J. *et al.* Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* **4,** e000167 (2008).
9 Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).
10 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
11 Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21,** 263–265 (2005).
12 Baranzini, S. E., Galwey, N. W., Wang, J., Khankhanian, P., Lindberg, R., Pelletier, D. *et al.* Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum. Mol. Genet.* **18,** 2078–2090 (2009).
13 McCarroll, S. A. Extending genome-wide association studies to copy-number variation. *Hum. Mol. Genet.* **17** (R2), R135–R142 (2008).

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

特集 SNPが変える肝胆膵診療

総論

疾患関連遺伝子を探し出すためのSNP解析

西田奈央* 徳永勝士*

索引用語：単一塩基多型（SNP），SNPタイピング，ゲノムワイド関連解析，絞り込み，多因子疾患

1 はじめに

近年，ゲノムワイド関連分析法(Genome-wide association study，GWAS)により，ヒトのさまざまな多因子疾患について，その遺伝的な要因を探索する研究が日本をはじめとして世界中で行われている．このGWASは日本の研究者によって先駆的に行われ，これまでにいくつかのヒトの多因子疾患の感受性遺伝子を特定することに成功している[1~4]．また，2003年からオーダーメイド医療実現基盤を構築することを目標とした「オーダーメイド医療実現化プロジェクト」が開始され，30万人の日本人を対象とした疾患関連研究(大規模ケース・コントロール関連解析)が行われている[5]．2008年には，日本における2大プロジェクトである「オーダーメイド医療実現化プロジェクト」と「ミレニアムゲノムプロジェクト」からそれぞれ独立に2型糖尿病に関連する遺伝子であるKCNQ1を発見したという報告がなされた[6,7]．また，われわれの研究室においても，CPT1B遺伝子とCHKB遺伝子の間に存在するSNPが睡眠障害の一つであるナルコレプシーと関連していることを発見し，2008年に報告をした[8]．

ゲノムワイド関連分析法は，ゲノム全域に分布する数十万種以上のSNPについて，非血縁の患者集団と健常者集団を対象として疾患遺伝子と連鎖不平衡(linkage disequilibrium，LD)にある多型マーカーを検出する手法である[9]．ゲノムワイド関連分析によりさまざまな多因子疾患を対象とした疾患感受性遺伝子の探索が行われるようになった背景には，本稿で紹介する大規模なSNP解析技術の進展が非常に大きな役割を果たしている．従来の多くのSNPタイピング法は，個々の多型部位を含むゲノム断片を特異的にPCRで増幅した後でアリルを識別する方法であった[10~14]．これらの方法では，1,000種程度のSNPを対象としたタイピングであれば，PCRプライマーをはじめとする各種試薬にかかるコストを考えても実用可能であるといえる

Nao NISHIDA et al: Identification of susceptibility genes for multifactorial diseases by analyzing single nucleotide polymorphisms

*東京大学大学院医学系研究科人類遺伝学分野 [〒113-0033 東京都文京区本郷7-3-1]

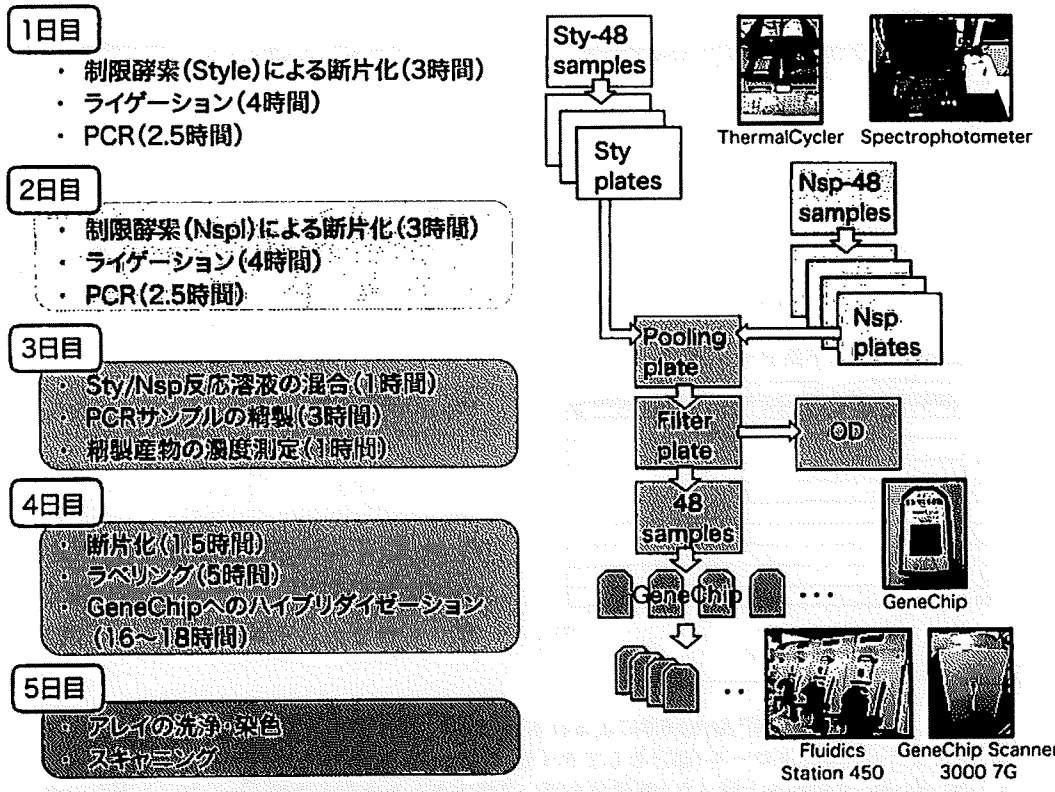

図1　SNP Array 6.0によるSNPタイピングの流れ
制限酵素(Sty I, Nsp I)による断片化反応からスキャニングまで全5日の工程でSNPタイピングが行われる．1検体につき500 ngのゲノムDNAを用いて全909,622種のSNPをタイピングすることができる．

が，数千から数万種を超える数のSNPをタイピングすることは困難となる．一方，近年になって多型部位特異的なPCRを行わずに大規模なSNPタイピングを行う方法が実用化された[15,16]．その一つであるAffymetrix社によって確立された方法では，まず制限酵素反応でゲノムDNAの断片化を行い，続いてそれら断片の両端にアダプター配列を付加し，まとめて増幅した後にマイクロアレイを用いたアリル特異的なハイブリダイゼーションを行う[15]．現在では，この手法を用いて90万種を超えるSNPを同時にタイピングするキットが市販されている(Affymetrix Genome-Wide Human SNP Array 6.0，以下，SNP Array 6.0)．

われわれは，ゲノムワイド関連分析で検出されたヒトの多因子疾患の疾患感受性候補領域の中から真の疾患感受性遺伝子を効率よく特定するためのSNP解析技術として，DigiTag2法を確立した[13]．DigiTag2法は，96カ所(もしくは32カ所)のSNPを同時に解析することができ，また，解析対象によらず同一のマイクロアレイを用いることができるため，専用マイクロアレイを準備する必要のない低コストのSNP解析技術である．われわれの教室に設置したヒトSNPタイピングセ

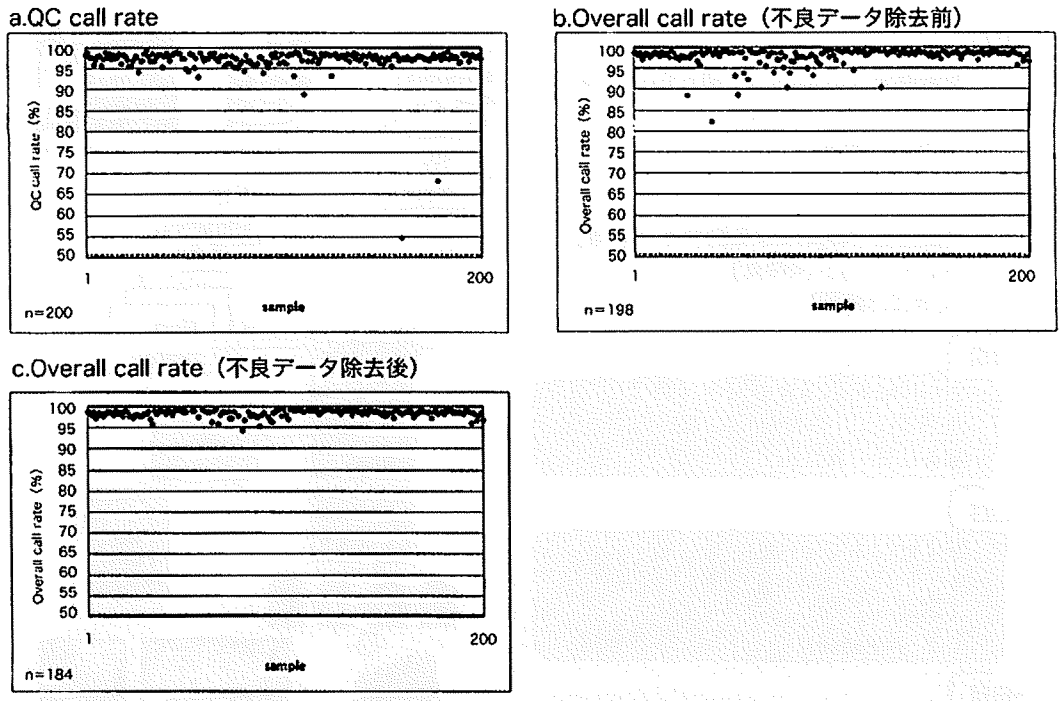

図2　SNP Array 6.0による日本人健常者200名のタイピング結果
a. クオリティーコントロール(QC)としてタイピングされた3,022SNPsのコール率を示す.
b. QC call rateが86%を上回った198検体を用いて決定された全909,622SNPsのコール率を示す.
c. QC call rateを指標として不良データを除去した後の184検体を用いて遺伝子型を決定した際の全909,622 SNPsのコール率を示す.

ンターでは，SNP Array 6.0によるSNPタイピングによりいくつかの多因子疾患についてゲノムワイド関連分析を実施し，さらにDigiTag2法を用いて疾患感受性遺伝子の特定を目的としたSNP解析を実施している.

2 ゲノムワイドSNP解析：SNP Array 6.0の技術原理

SNP Array 6.0によるSNPタイピングは，ゲノムの複雑さを低減しマイクロアレイへのハイブリダイゼーション効率を上げるための酵素反応ステップと，洗浄・染色装置(Fluidics Station 450)およびマイクロアレイ用スキャナー（GeneChip Scanner 3000 7G）を用いた検出ステップで構成される(図1)．1検体につき合計500 ngのゲノムDNAを使用して，2種類の制限酵素(Sty I，Nsp I)によるゲノムDNAの断片化を行った後，断片化したゲノムDNAの両末端にアダプター配列をライゲーション反応により付加する．アダプター配列は，続くPCRで使用されるプライマーと相同な配列を持ち，また制限酵素認識配列を突出端として持つ2本鎖DNAである．PCRでは，目的の長さを持ったDNA断片(250-1100 bp)だけが選択的に増幅される．続いて，Sty IおよびNsp Iそれぞれの PCR産物を混合した後，混合産物を精製し，DNase I制限酵素による断片化を行う．ここで，断

片化されたPCR産物は平均180 bp以下となる．最後にterminal deoxynucleotidyl transferase酵素反応により，断片化したPCR産物の末端にビオチンを導入する．

続いて，専用のマイクロアレイ(GeneChipアレイ)を用いてハイブリダイゼーションを行う．マイクロアレイに固定されるプローブは25塩基長のオリゴDNAで，SNP部位を含む塩基配列を持っている．2種類のアリルを正確に識別するために，SNP部位を25塩基長のプローブの中心に置いたプローブを基本として，SNP部位を中心から4塩基上流(+4)にずらしたプローブから4塩基下流(-4)にずらしたプローブまで7種類のプローブ(-4, -2, -1, 0, +1, +3, +4)を用意し，その中から最適な1種類のプローブを選択する．また，同一のプローブをマイクロアレイ上に3スポット用意することで，SNPタイピングデータの欠損を防ぐ工夫がなされている．

マイクロアレイへのハイブリダイゼーションが終了した後，洗浄・染色装置を用いてマイクロアレイの洗浄および蛍光染色を行う．蛍光染色は，蛍光分子で標識されたストレプトアビジンを，上述のビオチン導入されたPCR断片に結合することにより行われる．また，洗浄・染色装置内ではビオチン修飾された抗ストレプトアビジン抗体を用いてシグナルの増強が行われる．最後に蛍光染色されたマイクロアレイを専用のスキャナーで画像データとして読み取り，続いて専用のソフトウェア(Genotyping Console ver3.0 software)を用いて各SNPの遺伝子型を決定する．

複数の施設で行われたSNP Array 6.0によるSNP解析の結果から，Overall call rate（全909,622種のSNPのうち遺伝子型が決定されたSNPの割合）は平均99％以上となり，また，HapMapデータベースに登録されたタイピングデータとの遺伝子型一致率は99.7％を超えることがAffymetrix社から報告されている．また，タイピング結果が悪いことが明らかとなっている3,022種のSNPをクオリティーコントロール(QC)として用いて，QC call rate（3,022種のSNPのうち遺伝子型が決定されたSNPの割合）が86％を下回る検体を除外したうえで，全909,622種のSNPの遺伝子型は決定される．

3 ゲノムワイドSNP解析：日本人健常者200検体の解析結果

SNP Array 6.0によるSNPタイピングでは，制限酵素(Sty IおよびNsp I)による断片化反応に用いるゲノムDNA量がそれぞれ250 ngとなるように調整することがSNPタイピングの精度に大きな影響を与えることがこれまでの実験結果から明らかとなっている[17]．われわれが行った日本人健常者200検体を対象としたSNP解析を例にあげると，200検体のうち195検体のゲノムDNA濃度は規定濃度である50 ng/μlを満たしており，平均54.8 ng/μlであったが，5検体は規定濃度を下回り平均41.1 ng/μlであった．そこで，規定濃度を下回った5検体は制限酵素断片化反応に6 μlを持ち込み，ゲノムDNAの総量が約250 ngとなるように調整してタイピングを行った．日本人健常者200検体のタイピング結果から，QC call rateは平均97.37％となり，また，QC call rateが86％を下回る検体は200検体のうち2検体となった(図2a)．続いて，QC call rateが86％を上回った198検体を用いて全909,622SNPsのコール率(Overall call rate)を決定したところ，平均99.58％となった(図2b)．

膨大なSNPデータを取り扱うゲノムワイド関連分析において，タイピングエラーが原

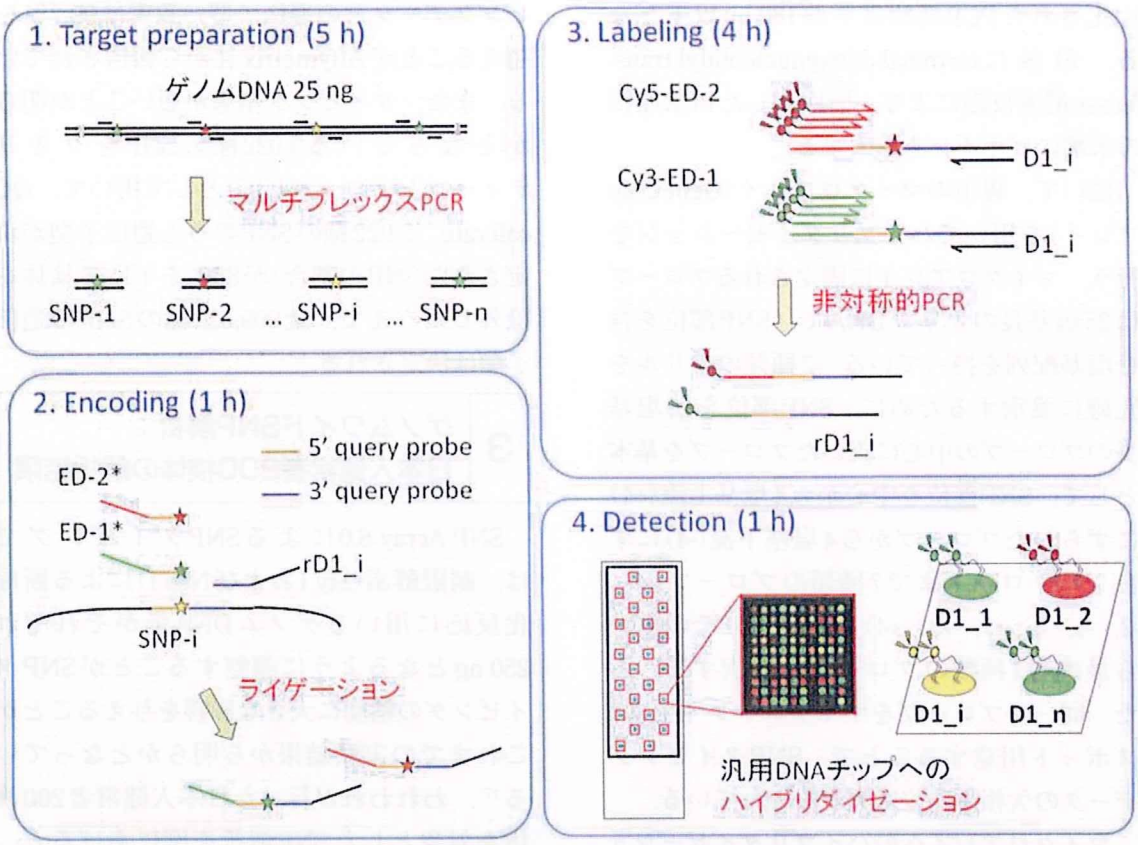

図3　DigiTag2法の概要

DigiTag2法は，ターゲット分子調製，エンコード，ラベリング，検出の4つの工程で構成される．5' query probeにはアリルに対応して2種類のED（ED-1，ED-2）を付加し，また3' query probeにはSNPに応じてD1（D1_i）を付加する．EDおよびD1は物理的，化学的性質が一様となるように設計した23塩基長のオリゴDNAである．配列が相補鎖である場合には配列名称の前に"r"を付けた．

因で生じる偽陽性関連は解析を進める上で大きな障害となる．そこで，SNP Array 6.0に搭載された全90万種のSNPについて，できるだけ多くのSNPの遺伝子型を正確に決定する必要がある．Affymetrix社が提供する遺伝子型決定ソフトウェアは，Birdseedアルゴリズムにより遺伝子型の決定を行う．このBirdseedアルゴリズムはサンプルサイズが小さくても遺伝子型を決定できるものの，高いタイピング精度を得るためには48検体以上を用いて遺伝子型を決定する必要があることが，これまでの解析結果から明らかとなっている[17]．また，QC call rate > 86％ではタイピング不良データを十分に取り除くことができず，タイピング不良データが混在することでOverall call rateを低下させることが明らかとなったため，われわれはQC call rate > 95％を閾値としてタイピング不良データの除去を行うこととした（図2c）．

遺伝子型が決定されたSNPの中にはタイピング精度の悪いSNPが一部含まれており，それらのSNPは偽陽性関連の原因のひとつになると考えられる．これについては，マイナーアリル頻度（MAF），ハーディー・ワイ

ンバーグ平衡検定(Hardy-Weinberg equilibrium test, HWE)およびSNP call rate（各SNPについて，タイピングした全検体のうち遺伝子型を決定できた検体の割合)を指標として，タイピング精度の悪いSNPの大部分を排除することができる[18]．われわれの解析では，MAF > 5%，HWE p値 > 0.001，SNP call rate > 95%を満たすSNPは，590,248 SNPsとなり，この約59万種のSNPによりヒトゲノムの約75%をカバーできることから，日本人においてもSNP Array 6.0を用いたゲノムワイド関連解析が有用であることが期待される．

4 疾患感受性遺伝子特定のためのSNP解析：DigiTag2法の原理と特徴

DigiTag2法は，SNPの遺伝子型をオリゴDNAタグに変換してマルチプレックスSNPタイピング(96-plexもしくは32-plex)を行う(図3)[13]．オリゴDNAタグ(図3中，EDおよびD1)は物理的，化学的に性質が一様となるように設計した23塩基長のオリゴDNAで，オリゴDNAタグを使用することにより正確なDNA分子反応を行うことが可能となる．SNPタイピングで使用するプライマー/プローブは共通の設計水準でデザインするため，解析対象に依存しない共通の実験条件でのSNPタイピングが可能である．また，実験条件の検討が不要であることに加えて，オリゴDNAタグは解析対象となるSNPに対して自由に割り当てることができるため，結果表示に用いるDNAチップを汎用的に使用できるという特徴を持っている．

DigiTag2法はランニングコストが安いうえに，複数カ所のSNPをまとめて解析することができるため，複数のSNPを多検体で解析するのに適した技術である．本技術が他の解析技術と比較して特に優位性が高い点として，タイピング成功率(遺伝子型が決定できたSNP数/解析対象としたSNP総数)が90.72%（929/1,024SNPs)と非常に高いことが挙げられる．また，これまでにDigiTag2法を用いて，合計26,665検体以上を対象として1,100カ所以上のSNP解析を行った実績があり，DigiTag2法は高い成功率でSNPタイピングを実施できるだけでなく，96-plexまたは32-plexのいずれでも，非常に高いCall rate（平均99.53%)でSNPタイピングを行うことのできる技術である．

DigiTag2法は高い成功率でSNPタイピングを行えることから，ゲノムワイド連鎖分析あるいはゲノムワイド関連分析によって検出された候補領域における絞り込み解析を効率的に行う技術として利用されることが期待される．

5 GWASデータのデポジット

文部科学省の「統合データベースプロジェクト」において，われわれはSNPタイピングデータの半永続的な集約管理と研究者間の情報共有を目指して，日本人健常者のデータを登録した標準SNPデータベース，日本人健常者のコピー数多型(CNV)を登録したCNVデータベースおよびゲノムワイド関連解析のデータベース(GWAS-DB)を構築し，なるべく多くの研究グループのデータ登録を広くお願いしている[19]．GWAS-DBは，研究概要，品質基準などの情報と共に，遺伝子型頻度やアリル頻度および遺伝統計解析の結果を登録している．また，GWAS-DBはSNPだけでなくマイクロサテライトやCNVの疾患関連解析の結果も登録・閲覧することができ，エクソン情報やCNVなどの情報と遺伝統計解析

の結果を重ね合わせて可視化する機能を備えている．疾患関連SNPの候補を多面的に選択できるよう，複数の機関が産出した同一疾患のデータおよび異なるプラットフォームの解析結果を比較したり，メタ解析を行ったりする機能を搭載し，専門家以外にも利用しやすいデータベースの構築を目指している．

6 結 語

現在販売されているSNP Array 6.0は，欧米人で頻度の高いSNPが優先的に搭載されているため，日本人試料では約20％のSNPについて多型性が見られなかった．疾患感受性候補領域を最大限に検出するためにも，今後，アジア系集団に適したSNPセットを搭載したプラットフォームが作製されることを期待したい．また，SNP Array 6.0にはCNVを検出するためのプローブが搭載されているものの，CNVを解析するためのソフトウェアの解析精度にはまだ多くの問題が残っており，今後のCNV解析精度の向上が強く望まれる．いずれにせよ，ゲノムワイド多型解析情報は従来にない膨大なデータを産生することから，バイオインフォマティクスに関わる様々な研究者にとって挑戦に値する多くの課題を提供してくれるとともに，その達成によって従来にない実り豊かな成果をわれわれにもたらしてくれるに違いない．

文 献

1) Ozaki K, Ohnishi Y, Iida A et al : Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet 32 : 650–654, 2002
2) Tamiya G, Shinya M, Imanishi T et al : Whole genome association study of rheumatoid arthritis using 27039 microsatellites. Hum Mol Genet 14 : 2305–2321, 2005
3) The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447 : 661–678, 2007
4) Cupples LA, Arruda HT, Benjamin EJ et al : The Framingham Heart Study 100K SNP genome-wide association study resource: overview of 17 phenotype working group reports. BMC Med Genet 8 : s1, 2007
5) 文部科学省リーディングプロジェクト「オーダーメイド医療実現化プロジェクト」[http://www.biobankjp.org/]
6) Unoki H, Takahashi A, Kawaguchi T et al : SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. Nat Genet 40: 1098–1102, 2008
7) Yasuda K, Miyake K, Horikawa Y, et al : Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. Nat Genet 40: 1092–1097, 2008
8) Miyagawa T, Kawashima M, Nishida N et al : Variant between CPT1B and CHKB associated with susceptibility to narcolepsy. Nat Genet 40 : 1324–1328, 2008
9) Ohashi J, Tokunaga K : The power of genome-wide association studies of complex disease genes: statistical limitations of indirect approaches using SNP markers. J Hum Genet 46 : 478–482, 2001
10) Holland PM, Abramson RD, Watson R et al : Detection of specific polymerase chain reaction product by utilizing the 5′ → 3′ exonuclease activity of Thermus aquaticus DNA polymerase. Proc Natl Acad Sci USA 88 : 7276–7280, 1991
11) Pastinen T, Kurg A, Metspalu A et al : Minisequencing: A specific tool for DNA analysis and diagnostics on oligonucleotide arrays. Genome Res 7: 606–614, 1997
12) Bannai M, Higuchi K, Akesaka T et al : Single-nucleotide-polymorphism genotyping for whole-genome-amplified samples using automated fluorescence correlation spectroscopy. Anal Biochem 327: 215–221, 2004
13) Nishida N, Tanabe T, Takasu M et al : Further development of multiplex single nucleotide polymorphism typing method, the DigiTag2 assay. Anal Biochem 364: 78–85, 2007
14) Krjutškov K, Andreson R, Mägi R et al : Development of a single tube 640-plex genotyping method for detection of nucleic acid variations on mi-

croarrays. Nucleic Acids Res 36 : e75, 2008
15) Affymetrix, Inc. [http://www.affymetrix.com/index.affx].
16) Illumina, Inc. [http://www.illumina.com/].
17) Nishida N, Koike A, Tajima A et al : Evaluating the performance of Affymetrix SNP Array 6.0 platform with 400 Japanese individuals. BMC Genomics 9 : 431, 2008
18) Miyagawa T, Nishida N, Ohashi J et al : Appropriate data cleaning methods for genome-wide association study. J Hum Genet 53 : 886-893, 2008
19) 文部科学省「統合データベースプロジェクト」[https://gwas.lifesciencedb.jp/].

* * *

Hepatology Research 2009 doi: 10.1111/j.1872-034X.2009.00607.x

Original Article

A predictive model of response to peginterferon ribavirin in chronic hepatitis C using classification and regression tree analysis

Masayuki Kurosaki,[1] Kotaro Matsunaga,[2] Itsuko Hirayama,[1] Tomohiro Tanaka,[1] Mitsuaki Sato,[1] Yutaka Yasui,[1] Nobuharu Tamaki,[1] Takanori Hosokawa,[1] Ken Ueda,[1] Kaoru Tsuchiya,[1] Hiroyuki Nakanishi,[1] Hiroki Ikeda,[1] Jun Itakura,[1] Yuka Takahashi,[1] Yasuhiro Asahina,[1] Megumu Higaki,[4] Nobuyuki Enomoto[3] and Namiki Izumi[1]

[1]*Division of Gastroenterology and Hepatology and* [2]*Division of Pathology, Musashino Red Cross Hospital, Tokyo,* [3]*First Department of Internal Medicine, University of Yamanashi, Yamanashi, and* [4]*Department of Medical Science, Jikei Medical University, Tokyo, Japan*

Aim: Early disappearance of serum hepatitis C virus (HCV) RNA is the prerequisite for achieving sustained virological response (SVR) in peg-interferon (PEG-IFN) plus ribavirin (RBV) therapy for chronic hepatitis C. This study aimed to develop a decision tree model for the pre-treatment prediction of response.

Methods: Genotype 1b chronic hepatitis C treated with PEG-IFN alpha-2b and RBV were studied. Predictive factors of rapid or complete early virological response (RVR/cEVR) were explored in 400 consecutive patients using a recursive partitioning analysis, referred to as classification and regression tree (CART) and validated.

Results: CART analysis identified hepatic steatosis (<30%) as the first predictor of response followed by low-density-lipoprotein cholesterol (LDL-C) (≥100 mg/dL), age (<50 and <60 years), blood sugar (<120 mg/dL), and gamma-glutamyltransferase (GGT) (<40 IU/L) and built decision tree model. The model consisted of seven groups with variable response rates from low (15%) to high (77%). The reproducibility of the model was confirmed by the independent validation group ($r^2 = 0.987$). When reconstructed into three groups, the rate of RVR/cEVR was 16% for low probability group, 46% for intermediate probability group and 75% for high probability group.

Conclusions: A decision tree model that includes hepatic steatosis, LDL-C, age, blood sugar, and GGT may be useful for the prediction of response before PEG-IFN plus RBV therapy, and has the potential to support clinical decisions in selecting patients for therapy and may provide a rationale for treating metabolic factors to improve the efficacy of antiviral therapy.

Key words: data mining, decision tree, HCV, low-density-lipoprotein-cholesterol, steatosis

INTRODUCTION

COMBINATION THERAPY WITH pegylated interferon (PEG-IFN) and ribavirin (RBV) is now recognized as a standard treatment for patients with chronic hepatitis C.[1] However, the rate of sustained virological response (SVR) to 48 weeks of PEG-IFN RBV combination therapy is only 50% in patients with hepatitis C virus (HCV) genotype 1b and high HCV RNA titer, so called difficult to treat chronic hepatitis C patients.[2,3] Within this difficult to treat group, the response to treatment sometimes can be highly heterogeneous for cases which are apparently equivalent in HCV RNA titer, making the prediction of response before treatment a difficult task. It has been suggested that early virological response (EVR), defined as either undetectable HCV RNA or a 2 log drop in HCV RNA at week 12, is a reliable means to predict SVR.[2,4] More recently, it has been suggested that patients with a rapid virological response (RVR: undetectable HCV RNA at week 4) and a complete EVR (cEVR: undetectable HCV RNA at week 12)

Correspondence: Dr Namiki Izumi, Division of Gastroenterology and Hepatology, Musashino Red Cross Hospital, 1-26-1 Kyonan-cho, Musashino-shi, Tokyo 180-8610, Japan. Email: nizumi@musashino.jrc.or.jp

Received 26 May 2009; revision 25 August 2009; accepted 26 August 2009.

achieve high SVR rates, while patients with a partial EVR (pEVR: 2 log drop in HCV RNA but still detectable at week 12) have lower rates of SVR.[5] Since PEG-IFN RBV combination therapy is costly and accompanied by potential adverse effects, the ability to predict the possibility of RVR or cEVR before therapy and identifying curable patients may significantly influence the selection of patients for therapy. Moreover, identification of baseline predictors of poor response is particularly important to establish a rationale for identifying therapeutic targets to improve the efficacy of antiviral therapy.

Data mining is a method of predictive analysis which explores tremendous volumes of data to discover hidden patterns and relationships in highly complex datasets and enables the development of predictive models. The classification and regression tree (CART) analysis is a core component of the decision tree tool for data mining and predictive modeling,[6] is deployed to decision makers in various fields of business, and currently is being used in the area of biomedicine.[7–13] The results of CART analysis are presented as a decision tree, which is intuitive and facilitates the allocation of patients into subgroups by following the flow-chart form.[14] CART has been shown to be competitive with other traditional statistical techniques such as logistic regression analysis.[15]

In the present study, we used the CART analysis to explore baseline predictors of response to PEG-IFN plus RBV therapy among clinical, biochemical, virological and histological pretreatment variables and to define a pre-treatment algorithm to discriminate chronic hepatitis C patients who are likely to respond to PEG-IFN plus RBV therapy.

MATERIALS AND METHODS

Patients

A TOTAL OF 419 chronic hepatitis C patients were treated with PEG-IFN alpha-2b and RBV at Musashino Red Cross Hospital between December 2001 and December 2007. Among them, 400 patients who fulfilled the following inclusion criteria were enrolled in the present study. (i) infection by genotype 1b (ii) HCV RNA higher than 100 KIU/mL by quantitative PCR (Cobas Amplicor HCV Monitor, Roche Diagnostic systems, CA) which is usually used for the definition of high viral load in Japan (iii) lack of co-infection with hepatitis B virus or human immunodeficiency virus (iv) lack of other causes of liver disease such as autoimmune hepatitis, primary biliary cirrhosis, or alcohol intake of more than 20 g per day, and (v) having completed at least 12 weeks of therapy with an early virological response that could be evaluated. Patients received PEG-IFN alpha-2b (1.5 microgram/kg) subcutaneously every week and were administered a weight adjusted dose of RBV (600 mg for <60 kg, 800 mg for 60–80 kg, and 1000 mg for >80 kg) which is the recommended dosage in Japan. Data from two third of patients (269 patients) were used for the model building set and the remaining one third of patients (131 patients) were used as a validation set. Consent in writing was obtained from each patient and the study protocol conformed to the ethical guidelines of the 1975 Declaration of Helsinki and was approved by the institutional review committee.

Laboratory tests

Blood samples were obtained before therapy, and at least once every month during therapy and analyzed for hematologic tests, blood chemistries, and HCV RNA. In the present study, RVR and cEVR was defined as undetectable HCV RNA by qualitative PCR with a lower detection limit of 50 IU/mL (Amplicor, Roche Diagnostic systems, CA) at week 4 and 12, respectively. SVR was defined as undetectable HCV RNA at week 24 after the completion of therapy.

Histological examination

For all patients, liver biopsy specimens were obtained before therapy and were evaluated independently by three pathologists who were blinded to the clinical details. If there was a disagreement, the scores assigned by the majority of pathologists were used for the analysis. Fibrosis and activity were scored according to the METAVIR scoring system.[16] Fibrosis was staged on a scale of 0–4: F0 (no fibrosis), F1 (mild fibrosis: portal fibrosis without septa), F2 (moderate fibrosis: few septa), F3 (severe fibrosis: numerous septa without cirrhosis) and F4 (cirrhosis). Activity of necroinflammation was graded on a scale of 0–3: A0 (no activity), A1 (mild activity), A2 (moderate activity) and A3 (severe activity). Percentage of steatosis was quantified by determining the average proportion of hepatocytes affected by steatosis and graded on a scale of 0–3: grade 0 (no steatosis), grade 1 (0–9%), grade 2 (10–29%), and grade 3 (over 30%) as we reported previously.[17]

Database for analysis

A pretreatment database of 72 variables was created containing histological findings (grade of fibrosis, activity, and steatosis), laboratory tests including the quantity of HCV RNA by Cobas Amplicor, and clinical information (age, gender, body weight, and body mass index).