doi:10.1111/j.1440-1746.2008.05321.x

HEPATOLOGY

Virological and clinical implication of core promoter C1752/V1753 and T1764/G1766 mutations in hepatitis B virus genotype D infection in Mongolia

Abeer Elkady,* Yasuhito Tanaka,* Fuat Kurbanov,* Tsendsuren Oynsuren† and Masashi Mizokami*

*Department of Clinical Molecular Informative Medicine, Nagoya City University Graduate School of Medical Sciences, Nagoya, Japan; and †Laboratory of Molecular Biology, The Institute of Biology, Mongolian Academy of Sciences, Ulaanbaatar, Mongolia

See *J. Gastroenterol. Hepatol.* 2008; 23: 347–350 for Editorial Comment on this article.

Key words
genotype, hepatitis B virus, mutation.

Accepted for publication 23 October 2007.

Correspondence
Professor Masashi Mizokami, Department of Clinical Molecular Informative Medicine, Nagoya City University Graduate School of Medical Sciences, Kawasumi 1, Mizhuo, Nagoya 467-8601, Japan. Email: mizokami@med.nagoya-cu.ac.jp

Abstract

Background and Aim: The aim of the present study was to reveal virological and clinical features of hepatitis B virus (HBV) genotype D infection.

Methods: One hundred and twenty-two Mongolian chronic liver disease (CLD) patients infected with HBV were subjected for serological HBV-markers screening and HBV-enzyme immunoassay (EIA) genotyping. Nucleotide sequences were analyzed for 48 HBV/D strains (23 isolated from hepatocellular carcinoma (HCC) and 25 from CLD patients).

Results: Prevalence of hepatitis B e antigen (HBeAg) positivity was low (25.9%) in young patients (≤30 years old) indicating early HBeAg seroclearance in HBV/D carriers. The T1764/G1766 double mutation was the most common basal core promoter (BCP) mutation (29.2%) and was frequent in HBeAg-negative patients (39.3%). Patients harboring T1764/G1766 mutants exhibited lower HBV-DNA and HBV core antigen (HBcAg) levels than those with wild-type BCP strains ($P = 0.024$, 0.049, respectively). C1752 and/or V (not T) 1753 mutation was significantly prevalent in HCC patients (HCC *vs* CLD; 52.2% *vs* 20%, $P = 0.033$). T1762/A1764 mutation was detected in 75.0% of HCC patients with high viral load (≥5 log copies/mL). Precore stop codon mutation A1896 was detected in (70.8%) of HBV/D-infected patients.

Conclusions: In Mongolians infected with HBV/D, C1752 and/or V1753 mutation was associated with HCC.

Introduction

Hepatitis B virus (HBV) infection concerns more than 350 million infected people in the world and is a global public health problem. The clinical outcome of HBV infection varies greatly from acute self-limiting disease and inactive carriers up to chronic liver disease (CLD) including liver cirrhosis and hepatocellular carcinoma (HCC).[1] Accumulated evidence indicated that the most important viral factors contributing to development of HCC include genotype,[2] specific genetic mutations[3,4] and coinfections with other hepatitis viruses.[5–7]

Eight genotypes of HBV (A–H) are identified based on the comparison of complete genomes, and most of the genotypes have a distinct geographic distribution.[8–10] Controversial data concerning the clinical characteristics of HBV genotype D infection; HBV/D was associated with severe liver disease, and a higher prevalence of genotype D than genotype A was observed in HCC patients in comparison to asymptomatic carriers.[11] In contrast, other reports indicated the lack of association between this genotype and distinct clinical phenotype.[12–14]

The HBV genomic mutations occur due to a spontaneous error rate of viral reverse transcriptase and evolving of HBV genome under the antiviral pressure of host immune response.[15] Specific mutations may affect the translation of hepatitis B e antigen (HBeAg), as well as the replication of HBV and thus may modify the clinical outcome of HBV infection and contribute to HCC development.[16–18]

Previous studies clearly demonstrated that the coinfection with HBV and hepatitis D virus (HDV) and/or HCV is common in Mongolia, and is significantly associated with development of HCC in Mongolian HBV carriers.[5–7] The age-adjusted incidence rate of HCC in this country was estimated to be 61.8–98.93 per 100 000 men, representing one of the highest age-adjusted incidence rates in the world[19] and HBV/D infection was established in 50–80% of HCC patients according to the recent reports.[5,6] In the present study, we investigated potential association of the basal core promoter (BCP) and precore genomic region characteristics of HBV/D with virological and clinical features of the infection, particularly with development of HCC.

Journal of Gastroenterology and Hepatology 23 (2008) 474–481

Methods

Patients

A group of 122 hepatitis B surface antigen (HBsAg)-positive Mongolian patients with chronic liver disease was analyzed in this study. Coinfection (HBV + HDV) and (HBV + HCV) was found in 51/122 (41.8%) and 14/122 (11.5%), respectively, and triple infection (HBV + HCV + HDV) was detected in 41/122 (33.6%).[5] All patients had their biochemical liver profile examined; alanine aminotransferase (ALT) and aspartate aminotransferase (AST) were measured using kits from HUMAN International (Wiesbaden, Germany) by the Raitmana Frenkle kinetic method on a Humalyzer 3000 (HUMAN International). The clinical diagnosis of HCC was confirmed by biochemical laboratory liver function tests, liver tumor marker, Alpha fetoprotein (AFP) and protein induced by vitamin K antagonist-II (PIVKA-II), ultrasound, computer tomography, and/or liver biopsy with histopathological examination.[5] The patients with and without HCC were classified into two clinical groups, HCC and CLD, respectively.

Serological methods

Hepatitis B virus surface and e antigens were examined by chemiluminescent enzyme immunoassay (Lumipulse; Fujirebio, Tokyo, Japan). HBV genotypes were analyzed using two methods: EIA with pre-S2 epitopes specific monoclonal antibodies[20] (HBV genotype EIA; Institute of Immunology, Tokyo, Japan) and, where possible, by direct sequencing of enhancer II/core promoter and precore/core genomic regions with further phylogenetic analysis. Serum concentration of hepatitis B core antigen (HBcAg) and hepatitis B core-related antigen (HBcrAg) (HBcrAg is a precore/core gene product, comprising HBcAg and HBeAg) were measured using the chemiluminescence enzyme immunoassay (CLEIA) method as described previously.[21,22] Cut-off value for HBcAg positivity was set at 10 pg/mL and for HBcrAg positivity was set at 10 pg/mL.

HBV-DNA quantification

HBV-DNA was quantified using real-time detection polymerase chain reaction (RTD-PCR) as previously reported.[23] The method was applied with slight modification as described previously.[18] The detection limit of this assay was 100 copies/mL.

HBV genome PCR-amplification and sequencing

HBV-DNA was extracted from 0.1 mL serum using QIAamp DNA Blood Mini Kit (QIAGEN Inc., Hilden, Germany). The enhancer II/core promoter and precore regions of the HBV genome were amplified by PCR with a forward primer (IS2-2: 5′-CAT GGA GAC CAC CGT GAA CGC-3′ [nt 1607–1627]) and reverse primer (HBV1917R: 5′-CTC CAC AGA AGC TCC AAA TTC TTT A-3′ [nt 1942–1918]). PCR was initiated by the hot-start technique. The PCR reaction was undertaken for 45 cycles (94°C for 1 min, 60°C for 1 min and 72°C for 1 min) followed by an extension reaction for 7 min. Complete genomes were amplified by primer sets described previously.[24] PCR products were directly sequenced with Prism Big Dye (Applied Biosystems, Foster City, CA, USA) in an ABI 3100 DNA automated sequencer (Applied Biosystems).

Sequence analysis

Sequences were aligned using the CLUSTALW software program (Thompson *et al.* 1997).[25] Phylogenetic trees were constructed using neighbor-joining (NJ) analysis with the six-parameter distance correction method[26] with bootstrap test confirmation performed on 1000 resamplings on the online Hepatitis Virus Database (http://s2as02.genes.nig.ac.jp/).

Complete genome sequences were examined for the presence of intergenotypic recombination using the SIMPLOT software program as described previously[27] and were reconfirmed manually by visual inspection of the alignments.

Nucleotide sequences obtained and analyzed in this study were submitted to DDBJ with consecutive accession numbers AB270534–AB270584.

Statistical evaluation

Statistical analysis was performed with Fisher's exact test and the independent *t*-test for continuous variables using SPSS version 8.0 software packages (SPSS, Chicago, IL, USA). *P*-values (two-tailed) less than 0.05 were considered statistically significant.

Ethical considerations

This study was conducted in accordance with the guidelines of the Declaration of Helsinki and its subsequent amendments. Informed consent was obtained from all patients.

Results

Characteristics of the patients and HBV genotypes

Genotyping by EIA determined HBV/D and HBV/A in 109/122 (89.3%) and 4/122 (3.3%), respectively, and 9/122 (7.4%) cases were untypable by EIA. No significant difference was observed in the genotype distribution between the HCC and CLD groups.

In order to confirm genotyping results and to investigate genetic characteristics of the strains, all the cases were subjected to DNA extraction, PCR using specific primers designed to amplify a part of the HBV genome including enhancer II, BCP, epsilon loop, and a part of the precore/core coding genes followed by direct sequencing. However, only 52/122 (42.6%) cases were amplified by PCR, including 51 of genotype D and one of genotype A as determined previously by EIA.

A phylogenetic analysis of the successfully sequenced strains confirmed one case with genotype A, and 48 cases with genotype D (tree not shown). However, the remaining three cases had discrepancy with the genotyping results by EIA (HBV/D) exhibiting phylogenetic clustering with HBV/C reference sequences (tree not shown). In order to investigate whether this discrepancy was due to coinfection with different genotypes or an intergenotypic recombination event, these cases were subjected for complete genome sequencing. Two cases were successfully amplified, sequenced and were subjected to Bootscan analyses using SIMPLOT software. The

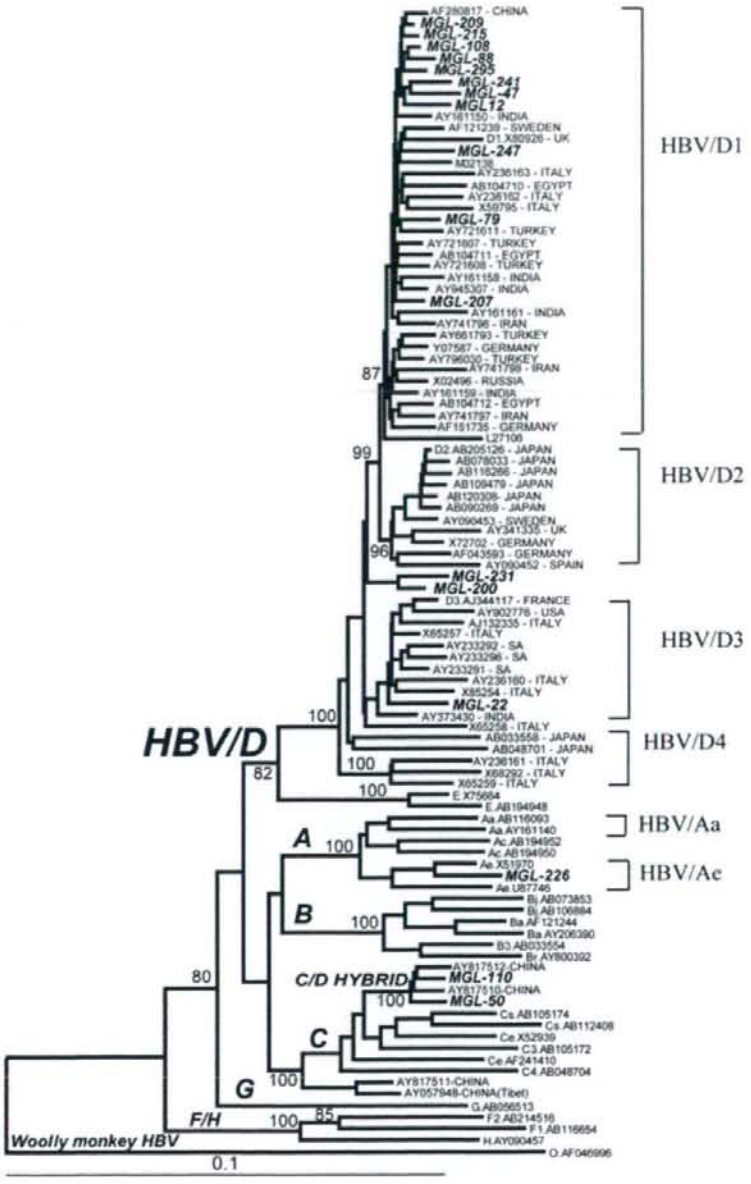

Figure 1 Phylogenetic NJ (Neighbour-joining) tree constructed using 100 complete genome sequences of hepatitis B virus (HBV). Seventeen strains isolated from Mongolia in this study are indicated in bold. Reference sequences retrieved from GenBank/EMBL/DDBJ are indicated with their accession numbers. The origins (in parentheses) are indicated for genotype D and C/D hybrid strains. Shown in the tree roots, bootstrap values higher than 80% were considered as significant. Genotypes (A to H and C/D hybrid) are indicated on the cluster roots. The outgroup consisted of HBV strain isolated from Woolly Monkey.

result indicated that both represented a hybrid HBV variant with a preS2/S region homologous with genotype D and the other regions homologous with genotype C. The breakpoints of the recombination were estimated around the positions of nucleotides 10 and 799.

The phylogenetic tree constructed using the complete genome sequences indicated that the two Mongolian strains were related to the previously reported Northern China C/D hybrid strains having the same recombination breakpoints[28] (Fig. 1). One of the two confirmed C/D hybrid cases had HCC.

The complete genome was also successfully amplified from the one HBV/A case in this study and, phylogenetically, it was related to the previously reported HBV/Ae subgenotype. In addition, 14 HBV/D strains were selected at random from CLD and HCC patients and subjected to complete genome sequencing. As shown in Fig. 1, 11 strains were grouped together with HBV/D1 references, one strain grouped with the HBV/D3 cluster, and the remaining two strains were not joined to any subgenotype groups while still corresponding to HBV/D.

Table 1 Clinical characteristics and viral mutations of patients with HBV/D

	Overall ($n = 48$)	CLD ($n = 25$)	HCC ($n = 23$)	*P*-value*
Age (years)†	46.1 ± 15	37.5 ± 13.3	55.3 ± 10.8	<0.0001
Gender (male)‡	22 (45.8%)	12 (48%)	10 (43.5%)	NS
HBeAg (positive)‡	20 (41.7%)	12 (48%)	8 (34.8%)	NS
ALT (IU/L)†	105.8 ± 62.9	87.7 ± 59.0	125.5 ± 62.2	0.036
AST (IU/L)†	93.1 ± 61.7	70.6 ± 53.2	117.7 ± 62.0	0.007
HBV-DNA (log copies/mL)†	4.4 ± 1.8	4.9 ± 1.9	3.9 ± 1.5	0.060
HBcAg (log pg/mL)†	1.5 ± 1.7	2.0 ± 1.9	0.8 ± 1.2	0.014
HBcrAg (log pg/mL)†	3.1 ± 1.7	3.6 ± 1.8	2.5 ± 1.3	0.020
Mutation prevalence‡				
T1653	7 (14.6%)	4 (16%)	3 (13%)	NS
G1727	7 (14.6%)	2 (8%)	5 (21.7%)	NS
C1752/ or V (not T)1753	17 (35.4%)	5 (20%)	12 (52.2%)	0.033
A1757	42 (87.5%)	23 (92%)	19 (82.6%)	NS
T1762/A1764	6 (12.5%)	3 (12%)	3 (13%)	NS
T1764/G1766	14 (29.2%)	6 (24%)	8 (34.8%)	NS
A1896	34 (70.8%)	16 (64%)	18 (78.3%)	NS
A1915	29 (60.4%)	17 (68%)	12 (52.2%)	NS

*Chronic liver disease group (CLD) *vs* hepatocellular carcinoma group (HCC).
†Mean ± SD.
‡*n* (%).
ALT, alanine aminotransferase; AST, aspartate aminotransferase; HBcAg, hepatitis B core antigen; HBcrAg, hepatitis B core-related antigen; HBeAg, hepatitis B e antigen; HBV, hepatitis B virus; NS, not significant.

Characteristics of HBV/D

To reveal specific mutations in association with HCC among patients with HBV/D, 48 DNA-positive genotype D cases were analyzed. Table 1 summarizes clinical characteristics and prevalence of the featured viral mutations among the 48 patients. The patients in the HCC group were significantly older than in the CLD group ($P < 0.0001$). Aminotransferase (ALT, AST) levels were significantly higher in the HCC group compared to the CLD group ($P = 0.036$, 0.0.007), respectively. HBcAg and HBcrAg levels were significantly higher in the CLD group compared to the HCC group ($P = 0.014$, 0.020), respectively. The level of HBV-DNA was relatively higher in the CLD group than in the HCC group (4.9 *vs* 3.9 log copies/mL, $P = 0.060$).

Comparing several mutations in enhancer II, core promoter, and precore regions, the most frequent BCP mutation was G1764T/C1766G (29.2%), followed by T1753V (not T) (23%), and A1762T/G1764A (12.5%), then A1752C (10%) (Table 1). Interestingly, G1757A substitution was prevalent in the overall population (87.5%) with no significant difference between HCC (92%) and CLD (82.6%) groups.

The V1753 mutation was frequent in HCC patients (34.7%) compared to CLD patients (12%) but did not reach a level of statistical significance ($P = 0.088$), probably due to the small number of samples. However, either or both mutations C1752/V1753 were detected in 17/48 (35.4%). The frequency of either or both C1752/V1753 mutation was significantly higher in the HCC group (52.2%) than in the CLD group (20%, $P = 0.033$). The same trend was also observed in HBeAg-negative patients (HCC *vs* CLD = 53.3% *vs* 15.3%, $P = 0.054$). The frequency of the T1764/G1766 double mutation was comparable between HCC (34.8%) and CLD (24%) groups. However, the T1762/A1764 double mutation was less frequent in both HCC (13%) and CLD (12%) groups. Precore stop codon mutation, G1896A, was detected in 70.8% (34/48) of HBV/D-infected patients including 64% (16/25) in the CLD group and 78.3% (18/23) in the HCC group.

Among the HBeAg-negative patients, the frequency of the precore stop codon mutation A1896 (75%) and BCP mutation T1764/G1766 (39.2%) was higher compared to those positive for HBeAg (65% and 15%), respectively. Studying the electropherogram of HBV strains in HBeAg-positive patients harboring A1896 mutants showed ambiguous patterns of G together with A at nt1896, indicating the presence of both wild-type and mutant strains in those patients.

Coinfection and mutations in association with HCC

It is important to note that only 13.1% of the serum samples tested were from HBV mono-infected patients whereas the remaining serum samples were from those coinfected with HBV + HDV (42%), HBV + HCV + HDV (34%) and HBV + HCV (12%). Therefore, there was a possibility that the observed association of a particular mutation with HCC was the result of coinfection-related bias. To elucidate this, we compared the prevalence of the mutations between HCC and CLD patients carrying any one particular pattern of infection; that is mono-infection with HBV or double-infection with HBV + HDV, or HBV + HCV, or triple-infection with HBV + HDV + HCV.

Among patients carrying HBV + HDV double-infection, we were able to observe that C1752/V1753 (either or both) mutations were significantly associated with HCC (HCC: 75.0% *vs* CLD: 18.2%, $P = 0.023$). The association of C1752/V1753 (either or

Table 2 Virological characteristics and prevalence of viral mutations between CLD and HCC patients in relation to HBV DNA

HBV-DNA (log copies/mL)	HCC		CLD	
	≥5 $n = 4$ (17.4%)	<5 $n = 19$ (82.6%)	≥5 $n = 10$ (40%)	<5 $n = 15$ (60%)
Age (years)	42.3 ± 14.2	58.1 ± 7.9	31.9 ± 13.1	41.3 ± 12.4
HBeAg (positive)	3 (75%)	5 (26.3%)	8 (80%)*	4 (26.7%)
HBcAg (positive)	3 (75%)**	2 (10.5%)	10 (100%)*	4 (26.7%)
Nucleotide substitutions				
T1653	0	3 (15.8%)	2 (20%)	2 (13.3%)
G1727	1 (25%)	4 (21.1%)	0	2 (13.3%)
C1752 and/or V1753	3 (75%)	9 (47.4%)	2 (20%)	3 (20%)
A1757	3 (75%)	16 (84.2%)	10 (100%)	13 (86.7%)
T1762/A1764	3 (75%)**,***	0	2 (20%)	1 (6.7%)
T1764/G1766	0	8 (42.1%)	1 (10%)	5 (33.3%)
A1896	3 (75%)	15 (78.9%)	4 (40%)	12 (80%)
A1915	2 (50%)	10 (52.6%)	7 (70%)	10 (66.7%)

*$P < 0.05$, in CLD group (patients with HBV-DNA ≥5 log copies/mL) *vs* (patients with HBV-DNA <5 log copies/mL).

**$P < 0.05$, in HCC group (patients with HBV-DNA ≥5 log copies/mL) *vs* (patients with HBV-DNA <5 log copies/mL).

***$P < 0.05$ (HCC patients with HBV-DNA ≥5 log copies/mL) *vs* (CLD patients with DNA <5 log copies/mL).

CLD, chronic liver disease group; HBcAg, hepatitis B core antigen; HBeAg, hepatitis B e antigen; HBV, hepatitis B virus; HCC, hepatocellular carcinoma.

both) mutations with HCC was also significant when patients with HBV + HDV dual-infection were combined with those with HBV + HDV + HCV triple-infection (HCC: 57.9% *vs* CLD: 18.8%, $P = 0.036$). However, in patients with HBV mono-infection, as well as those with HBV + HCV dual-infection or HBV + HDV + HCV triple-infection, we were unable to observe a significant association of this mutation with HCC, due to the small number of cases. Furthermore, no mutation or mutation pattern was observed in association with HBV mono-infection or coinfection, when we analyzed HBV BCP nucleotide sequence alignment (data not shown).

Virological characteristics of HBV infection in patients harboring BCP variants

The virological characteristics and prevalence of viral mutations were compared between patients with different HBV viral load in CLD and HCC groups (Table 2). The prevalence of HBeAg and HBcAg positivity was higher in both clinical groups with high HBV-DNA levels (≥5 log copies/mL). The frequency of double-mutation T1762/A1764 was significantly higher in the HCC group with HBV-DNA ≥5 log copies/mL (75%) compared to the HCC group with <5 log copies/mL (0%, $P = 0.002$) and CLD with <5 log copies/mL (6.7%, $P = 0.016$) (Table 2). Of note, the HBV-DNA level was significantly higher in HCC patients with T1762/A1764 mutants (6.1 ± 0.7 log copies/mL) compared to HCC with wild-type BCP strains (3.8 ± 2.0 log copies/mL) ($P = 0.021$), whereas there was no significant difference in the HBV-DNA level between CLD patients with the T1762/A1764 mutation (5.3 ± 0.6 log copies/mL) and those with wild-type BCP strains (5.5 ± 2.1 log copies/mL). The frequency of C1752 and/or V1753 tended to be higher in HCC with HBV-DNA ≥5 log copies/mL (75%) than in CLD with ≥5 log copies/mL (20%) and CLD with <5 log copies/mL (20%) ($P = 0.095$, 0.071, respectively).

Virological characteristics of BCP mutants are shown in Fig. 2. The mean level of HBV-DNA was higher in patients with the T1762/A1764 double mutation (5.7 ± 0.7 log copies/mL) than in those with wild-type BCP strains (4.9 ± 2.1 log copies/mL) with no significant difference (Fig. 2a), whereas in patients with T1764/G1766 mutants, both HBV-DNA levels (3.6 ± 1.1 log copies/mL) and HBcAg levels (0.9 ± 1.1 log pg/mL) were significantly lower than in those with wild-type BCP strains ($P = 0.024$, 0.049, respectively) (Fig. 2).

Discussion

Previous studies have demonstrated that coinfection with HBV and HDV and/or HCV is significantly associated with HCC development in Mongolian HBV carriers.[5–7] The main objective of the present study was to investigate the presence of a particular HBV mutation (or mutation pattern) in association with HCC in HBV genotype D-infected patients in this country. Using direct sequencing, we demonstrated that the C1752/V1753 (either or both) mutations were significantly associated with HCC in the studied cohort. This was also observed when only patients with coinfection were analyzed, suggesting that this mutation associated with HCC independently from coinfection.

The V1753 mutation was recently reported as a predictive factor for HCC among HBeAg-positive HBV/C1 carriers.[29] An *in vitro* study reported a higher replication capacity of HBV with C1753/T1762/A1764 than with T1762/A1764 mutations.[30]

Regarding the molecular aspect, C1753 is considered one of the 'hot spot' mutations of the HBx-encoding gene.[3,31] One of the functions of the HBx protein is transactivation of HBV-DNA transcription, as well as a number of cellular genes.[32] It was reported that C1753 enhanced the transactivation and antiproliferation activity of HBx protein in HBV/B[31] and, thereby, may contribute to carcinogenesis via the induction of a late G1 cell-cycle block,[33]

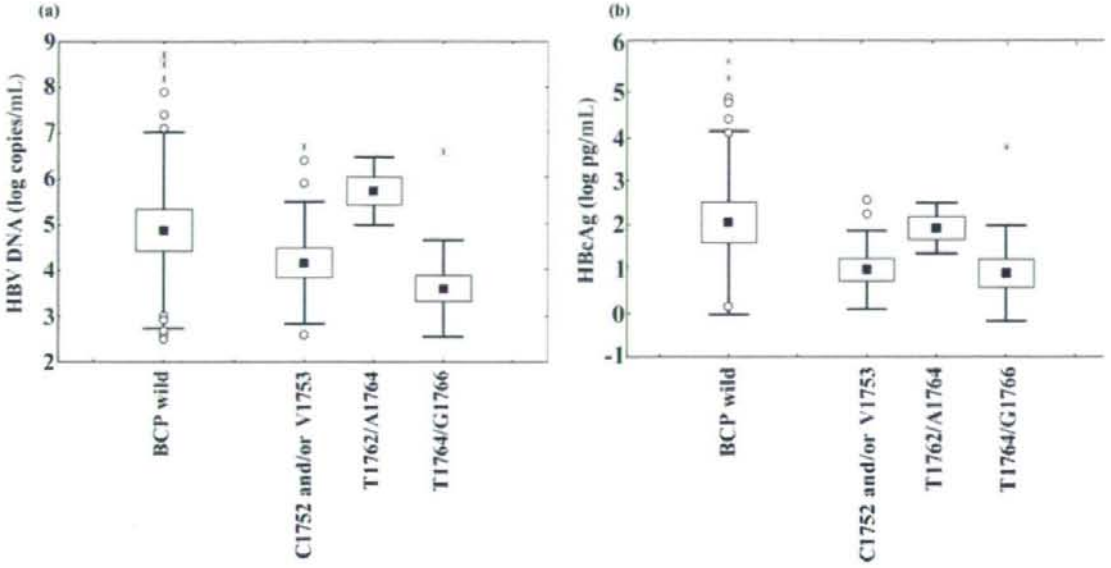

Figure 2 Comparison of (a) hepatitis B virus (HBV)-DNA levels (indicated in log copies/mL) and (b) HBV core antigen (HBcAg) levels (indicated in log pg/mL) among patients harboring basal core promoter wild-type strains (BCP wild); patients with C1752 and/or V1753 mutants (C1752 and/or V1753); patients with double-mutation T1762 and A1764 (T1762/A1764); and patients with double-mutation T1764 and G1766 (T1764/G1766). Box-whiskers, boxes and bars represent mean value, standard error and standard deviation, respectively.

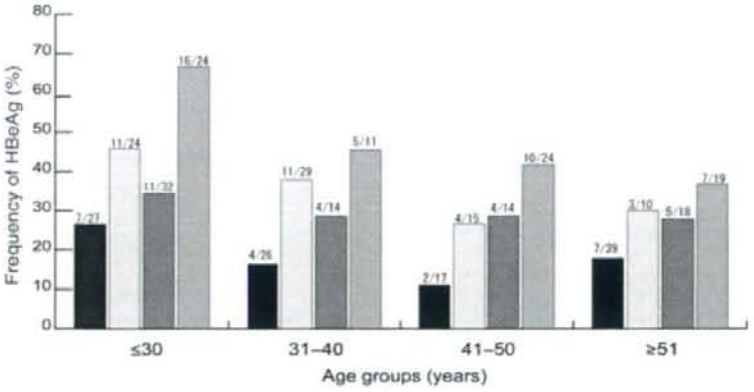

Figure 3 Age-specific seroprevalence of hepatitis B e antigen (HBeAg) in Mongolian patients with hepatitis B virus (HBV)/D enrolled in the present study compared with that observed previously in HBV/D, HBV/Aa (African/Asian type) and HBV/Ae (Europe/USA type).[18] (■), HBV/D (Mongolia); (■), HBV/D (India); (□), HBV/Aa (Africa/Asia); (■), HBV/Ae (Europe).

enhancement of HBV replication[34] leading to higher frequency of HBV integration into the human host genome.[35] However, further prospective study is needed to clarify the pathogenic role of C1752 and/or V1753 in HCC development among HBV/D-infected patients. The current study was limited by the low prevalence of patients with detectable HBV/DNA, probably due to the high frequency of coinfection and triple-infection in the studied cohort. Both HCV and HDV could inhibit HBV replication as reported previously by *in vitro*[36] and *in vivo* studies.[37,38]

Interestingly, our cohort study showed early seroclearance of HBeAg among HBV/D-infected patients in Mongolia compared to previous data as shown in Fig. 3.[18] Previous reports described different viral mutations in association with early clearance of HBeAg in serum among different genotypes; in particular, the mutation in the so-called 'Kozak sequence' immediately upstream of the core initiation codon may interfere with the translation of the HBeAg precursor in HBV/Aa,[39] and the A1896 precore stop codon mutation in HBV/D.[18] Among studied HBV/D-infected patients, the frequency of precore stop codon mutation A1896 was higher among HBeAg-negative patients than those positive for HBeAg, which is in agreement with previous studies,[18,40] suggesting that this mutation is one of the most important factors of

HBeAg early seroconversion in HBV/D-infected individuals. Another mutation found in association with HBeAg-negative status in the studied cohort was the T1764/G1766 double mutation of the BCP region.

A recent report has indicated that the T1764/G1766 double mutation is characteristic of the HBV/D strain and it was associated with a higher viral load.[41] The present study is in agreement with the previous studies regarding the observation that the T1764/G1766 double mutation is a feature of genotype D strains. However, in the present study, this mutation was associated with lower HBV-DNA and HBcAg levels compared to wild-type BCP strains ($P = 0.024$, 0.049, respectively). The discrepancy between the previous and present studies might be due to the difference in the clinical characteristics of studied patients because HBeAg-negative patients were enrolled in the previous study. As has been shown, HBeAg-negative patients with wild strains of core promoter had lower HBV-DNA levels, as the virus replication is sufficiently suppressed by the immune system, whereas HBeAg-negative patients with core promoter variants will escape host immune response and thus maintain higher HBV replication.[42] Finally, the present study is the first to indicate the lack of association of the T1764/G1766 double mutation with HCC in HBV/D-infected patients by comparative analysis of the clinical groups.

The high incidence of T1762/A1764 in HBV/Ba and HBV/C is usually complementary to its presence as a predictive factor for HCC in patients infected with either of these genotypes.[4,43–45] The featured double-mutation T1762/A1764 was infrequently found in our study; overall, 13% of the patients were harboring this mutant, with a similar frequency in both clinical groups (CLD and HCC). The low frequency of double-mutation T1762/A1764 in HBV/D was also reported previously in Iran,[41] USA and India.[18]

Different pathogenic mechanisms have been suggested contributing T1762/A1764 mutants to HCC. Previous reports indicated HBeAg reduction, but high viral replication by T1762/A1764 mutants *in vitro*,[46] in contrast to clinical studies that argued against the enhancement effect of T1762/A1764 on viral replication in regard to T1762/A1764-related hepatocarcinogensis.[39,47] Our study revealed this mutation in 75% of HCC patients with high viral load (≥5 log copies/mL) which might indicate the association of this mutation with enhanced viral replication in HCC patients. However, the number of cases was small and further study is required to confirm this trend.

In conclusion, the association of either or both C1752/V1753 with HCC was indicated in HBV/D-infected patients. High prevalence of BCP T1764/G1766 mutation and precore A1896 may together contribute to early seroclearance of HBeAg in patients with HBV/D. Further large-scale studies are needed to investigate these trends.

Acknowledgments

We thank Mr T. Kimura of Advanced Life Science Institute, Saitama, Japan for examining HBV core protein. This study was supported by a grant-in-aid from the Ministry of Health, Labor and Welfare of Japan (H16-kanen-3) and Toyoaki Foundation.

References

1 Chen DS. From hepatitis to hepatoma: lessons from type B viral hepatitis. *Science* 1993; **262**: 369–70.
2 Miyakawa Y, Mizokami M. Classifying hepatitis B virus genotypes. *Intervirology* 2003; **46**: 329–38.
3 Takahashi K, Ohta Y, Kanai K *et al.* Clinical implications of mutations C-to-T1653 and T-to-C/A/G1753 of hepatitis B virus genotype C genome in chronic liver disease. *Arch. Virol.* 1999; **144**: 1299–308.
4 Kao JH, Chen PJ, Lai MY, Chen DS. Basal core promoter mutations of hepatitis B virus increase the risk of hepatocellular carcinoma in hepatitis B carriers. *Gastroenterology* 2003; **124**: 327–34.
5 Oyunsuren T, Kurbanov F, Tanaka Y *et al.* High frequency of hepatocellular carcinoma in Mongolia; association with mono-, or co-infection with hepatitis C, B, and delta viruses. *J. Med. Virol.* 2006; **78**: 1688–95.
6 Tsatsralt-Od B, Takahashi M, Nishizawa T, Endo K, Inoue J, Okamoto H. High prevalence of dual or triple infection of hepatitis B, C, and delta viruses among patients with chronic liver disease in Mongolia. *J. Med. Virol.* 2005; **77**: 491–9.
7 Kurbanov F, Tanaka Y, Elkady A, Oyunsuren T, Mizokami M. Tracing hepatitis C and Delta viruses to estimate their contribution in HCC rates in Mongolia. *J. Viral. Hepat.* 2008; **14**: 327–34. doi:101111/j1365-2893200700864x.
8 Norder H, Courouce AM, Magnius LO. Complete genomes, phylogenetic relatedness, and structural proteins of six strains of the hepatitis B virus, four of which represent two new genotypes. *Virology* 1994; **198**: 489–503.
9 Okamoto H, Tsuda F, Sakugawa H *et al.* Typing hepatitis B virus by homology in nucleotide sequence: comparison of surface antigen subtypes. *J. Gen. Virol.* 1988; **69** (Pt 10): 2575–83.
10 Stuyver L, De Gendt S, Van Geyt C *et al.* A new genotype of hepatitis B virus: complete genome and phylogenetic relatedness. *J. Gen. Virol.* 2000; **81**: 67–74.
11 Thakur V, Guptan RC, Kazim SN, Malhotra V, Sarin SK. Profile, spectrum and significance of HBV genotypes in chronic liver disease patients in the Indian subcontinent. *J. Gastroenterol. Hepatol.* 2002; **17**: 165–70.
12 Gandhe SS, Chadha MS, Arankalle VA. Hepatitis B virus genotypes and serotypes in western India: lack of clinical significance. *J. Med. Virol.* 2003; **69**: 324–30.
13 Yalcin K, Degertekin H, Bahcecioglu IH *et al.* Hepatitis B virus genotype D prevails in patients with persistently elevated or normal ALT levels in Turkey. *Infection* 2004; **32**: 24–9.
14 Bahri O, Cheikh I, Hajji N *et al.* Hepatitis B genotypes, precore and core promoter mutants circulating in Tunisia. *J. Med. Virol.* 2006; **78**: 353–7.
15 Gunther S, Fischer L, Pult I, Sterneck M, Will H. Naturally occurring variants of hepatitis B virus. *Adv. Virus Res.* 1999; **52**: 25–137.
16 Baumert TF, Rogers SA, Hasegawa K, Liang TJ. Two core promotor mutations identified in a hepatitis B virus strain associated with fulminant hepatitis result in enhanced viral replication. *J. Clin. Invest.* 1996; **98**: 2268–76.
17 Ahn SH, Kramvis A, Kawai S *et al.* Sequence variation upstream of precore translation initiation codon reduces hepatitis B virus e antigen production. *Gastroenterology* 2003; **125**: 1370–8.
18 Tanaka Y, Hasegawa I, Kato T *et al.* A case-control study for differences among hepatitis B virus infections of genotypes A (subtypes Aa and Ae) and D. *Hepatology* 2004; **40**: 747–55.
19 Bosch FX, Ribes J, Diaz M, Cleries R. Primary liver cancer: worldwide incidence and trends. *Gastroenterology* 2004; **127**: S5–16.
20 Usuda S, Okamoto H, Iwanari H *et al.* Serological detection of hepatitis B virus genotypes by ELISA with monoclonal antibodies to

type-specific epitopes in the preS2-region product. *J. Virol. Methods* 1999; **80**: 97–112.
21 Kimura T, Rokuhara A, Matsumoto A *et al.* New enzyme immunoassay for detection of hepatitis B virus core antigen (HBcAg) and relation between levels of HBcAg and HBV DNA. *J. Clin. Microbiol.* 2003; **41**: 1901–6.
22 Kimura T, Rokuhara A, Sakamoto Y *et al.* Sensitive enzyme immunoassay for hepatitis B virus core-related antigens and their correlation to virus load. *J. Clin. Microbiol.* 2002; **40**: 439–45.
23 Abe A, Inoue K, Tanaka T *et al.* Quantitation of hepatitis B virus genomic DNA by real-time detection PCR. *J. Clin. Microbiol.* 1999; **37**: 2899–903.
24 Sugauchi F, Mizokami M, Orito E *et al.* A novel variant genotype C of hepatitis B virus identified in isolates from Australian Aborigines: complete genome sequence and phylogenetic relatedness. *J. Gen. Virol.* 2001; **82**: 883–92.
25 Thompson JD, Higgins DG, Gibson TJ. Clustal W improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 1994; **22**: 4673–80.
26 Gojobori T, Ishii K, Nei M. Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. *J. Mol. Evol.* 1982; **18**: 414–23.
27 Lole KS, Bollinger RC, Paranjape RS *et al.* Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. Virol.* 1999; **73**: 152–60.
28 Wang Z, Liu Z, Zeng G *et al.* A new intertype recombinant between genotypes C and D of hepatitis B virus identified in China. *J. Gen. Virol.* 2005; **86**: 985–90.
29 Tanaka Y, Mukaide M, Orito E *et al.* Specific mutations in enhancer II/core promoter of hepatitis B virus subgenotypes C1/C2 increase the risk of hepatocellular carcinoma. *J. Hepatol.* 2006; **45**: 646–53.
30 Parekh S, Zoulim F, Ahn SH *et al.* Genome replication, virion secretion, and e antigen expression of naturally occurring hepatitis B virus core promoter mutants. *J. Virol.* 2003; **77**: 6601–12.
31 Lin X, Xu X, Huang QL *et al.* Biological impacts of 'hot-spot' mutations of hepatitis B virus X proteins are genotype B and C differentiated. *World J. Gastroenterol.* 2005; **11**: 4703–8.
32 Murakami S. Hepatitis B virus X protein: a multifunctional viral regulator. *J. Gastroenterol.* 2001; **36**: 651–60.
33 Sirma H, Giannini C, Poussin K, Paterlini P, Kremsdorf D, Brechot C. Hepatitis B virus X mutants, present in hepatocellular carcinoma tissue abrogate both the antiproliferative and transactivation effects of HBx. *Oncogene* 1999; **18**: 4848–59.
34 Ozer A, Khaoustov VI, Mearns M *et al.* Effect of hepatocyte proliferation and cellular DNA synthesis on hepatitis B virus replication. *Gastroenterology* 1996; **110**: 1519–28.
35 Bonilla Guerrero R, Roberts LR. The role of hepatitis B virus integrations in the pathogenesis of human hepatocellular carcinoma. *J. Hepatol.* 2005; **42**: 760–77.
36 Wu JC, Chen PJ, Kuo MY, Lee SD, Chen DS, Ting LP. Production of hepatitis delta virus and suppression of helper hepatitis B virus in a human hepatoma cell line. *J. Virol.* 1991; **65**: 1099–104.
37 Sagnelli E, Coppola N, Scolastico C *et al.* Virologic and clinical expressions of reciprocal inhibitory effect of hepatitis B, C, and delta viruses in patients with chronic hepatitis. *Hepatology* 2000; **32**: 1106–10.
38 Jardi R, Rodriguez F, Buti M *et al.* Role of hepatitis B, C, and D viruses in dual and triple infection: influence of viral genotypes and hepatitis B precore and basal core promoter mutations on viral replicative interference. *Hepatology* 2001; **34**: 404–10.
39 Baptista M, Kramvis A, Kew MC. High prevalence of 1762(T) 1764 (A) mutations in the basic core promoter of hepatitis B virus isolated from black Africans with hepatocellular carcinoma compared with asymptomatic carriers. *Hepatology* 1999; **29**: 946–53.
40 Bozdayi AM, Bozkaya H, Turkyilmaz AR *et al.* Nucleotide divergences in the core promoter and precore region of genotype D hepatitis B virus in patients with persistently elevated or normal ALT levels. *J. Clin. Virol.* 2001; **21**: 91–101.
41 Sendi H, Mehrab-Mohseni M, Zali MR, Norder H, Magnius LO. T1764G1766 core promoter double mutants are restricted to Hepatitis B virus strains with an A1757 and are common in genotype D. *J. Gen. Virol.* 2005; **86**: 2451–8.
42 Chu CJ, Keeffe EB, Han SH *et al.* Prevalence of HBV precore/core promoter variants in the United States. *Hepatology* 2003; **38**: 619–28.
43 Orito E, Mizokami M, Sakugawa H *et al.* A case-control study for clinical and molecular biological differences between hepatitis B viruses of genotypes B and C. Japan HBV Genotype Research Group. *Hepatology* 2001; **33**: 218–23.
44 Sugauchi F, Orito E, Ichida T *et al.* Epidemiologic and virologic characteristics of hepatitis B virus genotype B having the recombination with genotype C. *Gastroenterology* 2003; **124**: 925–32.
45 Orito E, Sugauchi F, Tanaka Y *et al.* Differences of hepatocellular carcinoma patients with hepatitis B virus genotypes of Ba, Bj or C in Japan. *Intervirology* 2005; **48**: 239–45.
46 Buckwold VE, Xu Z, Chen M, Yen TS, Ou JH. Effects of a naturally occurring mutation in the hepatitis B virus basal core promoter on precore gene expression and viral replication. *J. Virol.* 1996; **70**: 5845–51.
47 Liu CJ, Chen BF, Chen PJ *et al.* Role of hepatitis B viral load and basal core promoter mutation in hepatocellular carcinoma in hepatitis B carriers. *J. Infect. Dis.* 2006; **193**: 1258–65.

BMC Genomics

BioMed Central

Research article

Open Access

Evaluating the performance of Affymetrix SNP Array 6.0 platform with 400 Japanese individuals

Nao Nishida*[1], Asako Koike[2], Atsushi Tajima[3], Yuko Ogasawara[1], Yoshimi Ishibashi[1], Yasuka Uehara[1], Ituro Inoue[3] and Katsushi Tokunaga[1]

Address: [1]Department of Human Genetics, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan, [2]Central Research Laboratory, Hitachi Ltd, Tokyo, Japan and [3]Division of Molecular Life Science, School of Medicine, Tokai University, Isehara, Japan

Email: Nao Nishida* - nishida-75@umin.ac.jp; Asako Koike - asako.koike.ea@hitachi.com; Atsushi Tajima - atajima@is.icc.u-tokai.ac.jp; Yuko Ogasawara - you-o@m.u-tokyo.ac.jp; Yoshimi Ishibashi - ishi-y@m.u-tokyo.ac.jp; Yasuka Uehara - ysk-u@m.u-tokyo.ac.jp; Ituro Inoue - ituro@is.icc.u-tokai.ac.jp; Katsushi Tokunaga - tokunaga@m.u-tokyo.ac.jp

* Corresponding author

Published: 22 September 2008

BMC Genomics 2008, **9**:431 doi:10.1186/1471-2164-9-431

Received: 18 June 2008
Accepted: 22 September 2008

This article is available from: http://www.biomedcentral.com/1471-2164/9/431

Abstract

Background: With improvements in genotyping technologies, genome-wide association studies with hundreds of thousands of SNPs allow the identification of candidate genetic loci for multifactorial diseases in different populations. However, genotyping errors caused by genotyping platforms or genotype calling algorithms may lead to inflation of false associations between markers and phenotypes. In addition, the number of SNPs available for genome-wide association studies in the Japanese population has been investigated using only 45 samples in the HapMap project, which could lead to an inaccurate estimation of the number of SNPs with low minor allele frequencies. We genotyped 400 Japanese samples in order to estimate the number of SNPs available for genome-wide association studies in the Japanese population and to examine the performance of the current SNP Array 6.0 platform and the genotype calling algorithm "Birdseed".

Results: About 20% of the 909,622 SNP markers on the array were revealed to be monomorphic in the Japanese population. Consequently, 661,599 SNPs were available for genome-wide association studies in the Japanese population, after excluding the poorly behaving SNPs. The Birdseed algorithm accurately determined the genotype calls of each sample with a high overall call rate of over 99.5% and a high concordance rate of over 99.8% using more than 48 samples after removing low-quality samples by adjusting QC criteria.

Conclusion: Our results confirmed that the SNP Array 6.0 platform reached the level reported by the manufacturer, and thus genome-wide association studies using the SNP Array 6.0 platform have considerable potential to identify candidate susceptibility or resistance genetic factors for multifactorial diseases in the Japanese population, as well as in other populations.

Background

Together with technology developments on large-scale single nucleotide polymorphism (SNP) genotyping [1,2], there have been a number of genome-wide association studies (GWAS) to identify candidate susceptibility or resistance genetic factors for multifactorial diseases [3-7]. It is estimated that eleven million SNPs with a greater than 1% minor allele frequency (MAF) are located in the

human genome [8]. Over six million SNPs have been uploaded on public SNP databases through the Human Genome Project and international SNP discovery projects. Among these SNPs, over 900 K SNPs across the human genome are selected with an average MAF of 19.6%, 18.2% and 20.6% in the HapMap Caucasians, Asians and Africans, respectively, and can be simultaneously genotyped using Affymetrix Genome-Wide Human SNP Array 6.0 platform [9]. Several studies have evaluated the coverage of commercial platforms using HapMap population data and genotype data of non-reference Caucasian populations [10-12]. Results from these studies indicated that in a non-reference Caucasian population, as well as the HapMap populations, commercial SNP typing platforms offered similar levels of genome coverage. However, the number of genotyped Japanese individuals in the HapMap project was only 45 samples, which may lead to inaccurate estimation of the number of SNPs with low MAF in the Japanese population.

The SNP Array 6.0 platform offers the genotype calling algorithm "Birdseed" to determine the genotypes of 909,622 SNPs [9]. The Birdseed algorithm performs a multiple-chip analysis to estimate signal intensity for each allele of each SNP, fitting probe-specific effects to increase precision, and then makes genotype calls by fitting a Gaussian mixture model in the two-dimensional A-signal vs. B-signal space, using SNP-specific models to improve accuracy. There was a report that 45% of SNPs observed to be significantly associated with the disease did not agree with Hardy-Weinberg equilibrium (HWE) using the previous version of Mapping 500 K Array set [13]. Some of the miss-called SNPs would be induced by genotype calling algorithms and are likely to be ranked as significantly associated with the disease (false-positive). Therefore, there are strong demands for accurate genotype calls using the Birdseed algorithm.

The SNP Array 6.0 platform has three check points prior to hybridization on GeneChip arrays in order to exclude experimental errors; PCR amplicon size check by electropherograms, DNase I digested fragment size check by electropherograms and quantity check of the purified PCR products. The platform also includes Quality Control (QC) probes for 3,022 SNPs to assess the overall quality for a sample based on the Dynamic Model (DM) algorithm. There are assay criteria to exclude experimental errors and low-quality samples; however, we empirically know that some samples, which pass these criteria, have low-quality genotyping results.

In this study, we genotyped 400 non-HapMap Japanese samples using the SNP Array 6.0 platform in order to evaluate the number of SNPs available for GWAS in the Japanese population, to examine an appropriate approach for acquiring accurate genotype calls using the Birdseed genotype calling algorithm, and to evaluate the assay criteria for preventing low-quality genotyping data.

Results

Genotyping 400 Japanese samples using SNP Array 6.0 platform

We collected 2 sets of 200 Japanese samples for genotyping using the SNP Array 6.0 platform. The average concentration of genomic DNA for the 1st set of 200 samples was 54.8 ng/μl and that for the 2nd set of 200 samples was 52.7 ng/μl. One of the critical points for the SNP Array 6.0 platform to acquire high quality genotyping data is to prepare a uniform quantity of 250 ng genomic DNA for Nsp I and Sty I digestion steps. When an almost 10-fold excess amount of genomic DNA was used, the average overall call rate drastically decreased to about 80% for both Nsp I and Sty I digestion steps with the Mapping 500 K Array (data not shown).

The average concentration of purified PCR products for the two sets of 200 samples was 524.4 ng/μl (range 412.8 to 718.0 ng/μl) and 497.3 ng/μl (range 256.6 to 939.8 ng/μl), respectively (Figure 1a and Figure 2a). In total, 11 samples (2 samples for the 1st set and 9 samples for the 2nd set) showed low QC call rates below the default 86% QC criteria (Figure 1b and Figure 2b). The genotype calls of 909, 622 SNPs for each individual were determined using the Birdseed genotype calling algorithm, embedded in the Affymetrix Genotyping Console 2.0 software (Affymetrix). The 198 samples of the 1st set that were over 86% QC criteria were used to assign genotypes and had an average overall call rate of 99.58%, ranging from 96.42 to 99.90% (Figure 1c). For the 2nd set, 191 samples were over 86% QC criteria and the average overall call rate was 97.54%, ranging from 89.52 to 99.27% (Figure 2c). When genotype calls were determined for every 48 samples analyzed simultaneously in the same batch, the average overall call rate was improved to 99.71% (range, 98.37 to 99.94%) for the 1st set, and 98.66% (range, 94.86 to 99.76%) for the 2nd set (Figure 1d and Figure 2d).

Assay criteria for experimental errors occurring on running batches

The SNP Array 6.0 platform has three check points prior to hybridization on GeneChip arrays in order to remove samples with experimental errors. However, some samples that pass these check points still have relatively low-quality genotyping results with lower overall call rates than 97%; 1 sample for the 1st set of 200 samples and 59 samples for the 2nd set of 200 samples. When genotype calls were determined for every 48 samples simultaneously analyzed in the same batch, the average overall call rate of 48 samples for batch #1 from the 2nd set was 97.21%, which was almost 2% lower than other batches

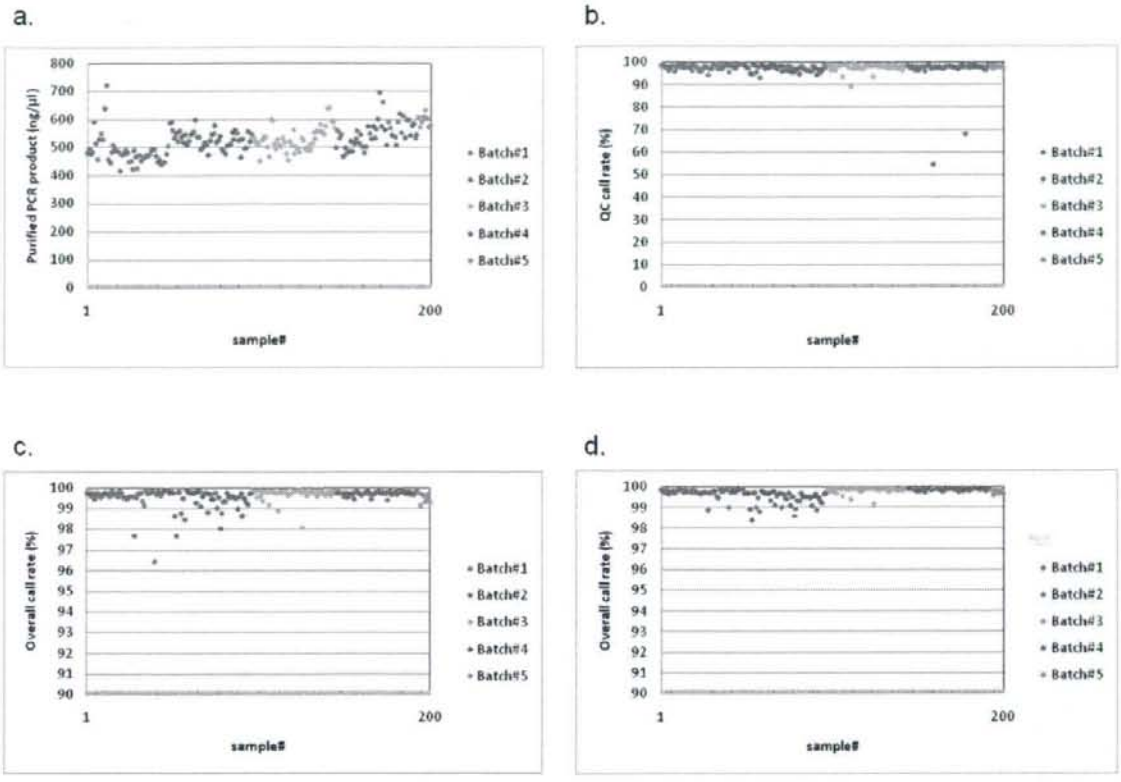

Figure 1
Genotyping results of the 1st set of 200 samples using the SNP Array 6.0 platform. Colours are based on every 48 samples analyzed simultaneously as a batch. a. Concentration of purified PCR products for each sample. b. QC call rate for each sample. c. Overall call rate for each sample, as determined by the Birdseed algorithm using total 198 samples that passed the default 86% QC criteria. d. Overall call rate for each sample, as determined by the Birdseed algorithm using samples in the same batch.

(Figure 2d). The concentration of purified PCR products from batch #1 drastically fluctuated among the 48 samples (Figure 2a). The CV (standard deviation/average) of the purified PCR product concentration for batch #1 was much higher than that for any other batches from the two sets of 200 samples (Figure 3). The CV of the purified PCR product concentration is a new indicator to assess experimental quality for each of the running batches, and may remove the experimental errors occurring on the running batches prior to hybridization on the GeneChip arrays.

For the 48 samples from batch #1 of the 2nd set, the intact genomic DNA could not be detected clearly when the samples were electrophoresed on 1.0% agarose gels (Figure 4). Therefore, these genomic DNAs for batch #1 of the 2nd set may have degraded due to repetitive freezing and thawing, which led to low-quality genotyping results. Preparation of the exact amount of intact genomic DNA is considered to be one of the crucial points for the SNP array 6.0 platform.

In order to assess the performance of the SNP Array 6.0 platform and the Birdseed algorithm, we mainly used genotyping data obtained from the 1st set of 200 samples because the 2nd set contained samples in poor condition.

Genotype calling accuracy with "Birdseed" algorithm

The genotype calling accuracy of the Birdseed algorithm was considered to be improved as the sample number for determining genotype calls increased. We determined 909,622 genotype calls for 12 samples among 198 samples with over 86% QC criteria, and used these genotype calls as a reference. We also determined the genotype calls of the same 12 samples under 6 different sample sizes,

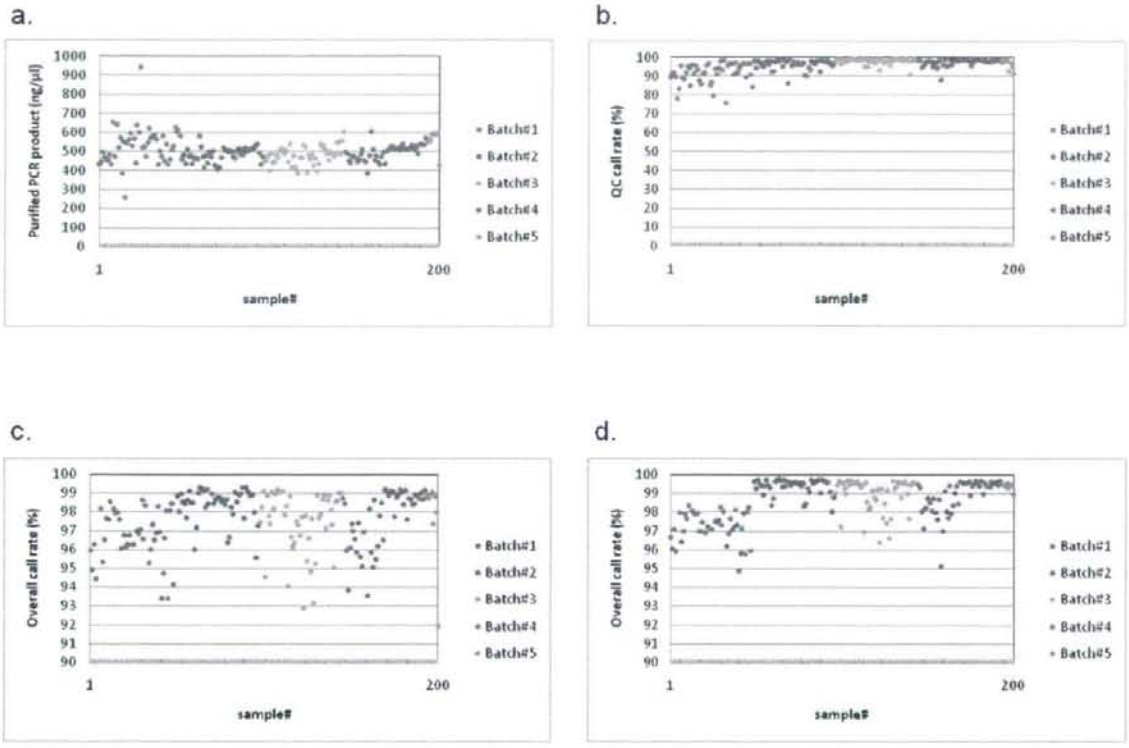

Figure 2

Genotyping results of 2nd set of 200 samples using the SNP Array 6.0 platform. a. Concentration of purified PCR products for each sample. b. QC call rate for each sample. c. Overall call rate for each sample, as determined by the Birdseed algorithm using a total of 191 samples that passed the default 86% QC criteria. d. Overall call rate for each sample, as determined by the Birdseed algorithm using samples in the same batch.

using 12 samples, 24 samples, 36 samples, 48 samples, 72 samples and 96 samples. To investigate the genotype calling accuracy of the Birdseed algorithm, we compared the genotype calls determined under 6 different sample sizes to the reference genotype calls for each of the 12 samples. We prepared 4 sets of 12 samples from a batch of 48 samples (Batch #3) and performed the genotype call comparison for each set of 12 samples. Figure 5 shows the average overall call rate and the average concordance rate for each set of the 12 samples. The average overall call rate for 4 sets of the 12 samples, which were determined with 12 samples, 24 samples, 36 samples, 48 samples, 72 samples, 96 samples and 198 samples, were 99.84%, 99.86%, 99.84%, 99.83%, 99.79%, 99.75% and 99.71%, respectively. The average concordance rate for the 4 sets of the 12 samples under 6 different sample sizes were 99.47%, 99.75%, 99.80%, 99.84%, 99.86% and 99.87%, respectively. Here, "No Calls" was excluded from the concordance calculation.

Our results showed that the average overall call rate of the 12 samples was almost constant when the genotype calls were determined with fewer than 48 samples; however, it gradually decreased as the sample number increased from 48 to 198, which showed a negative correlation with a P value of 0.0053. In contrast, the concordance rate gradually increased as the sample number increased, which showed a positive correlation with a P value of 0.0115.

Removing low-quality samples by adjusting QC criteria

Our results showed that the average overall call rate gradually decreased as the sample number increased, presumably due to low-quality samples included in the genotype calling with the Birdseed algorithm. Indeed, there was one sample which had an overall call rate lower than 97% among the 198 samples with over 86% QC call rate. Therefore, we applied more stringent QC criteria to remove the low-quality samples, because a linear relationship was observed between QC call rate and overall call

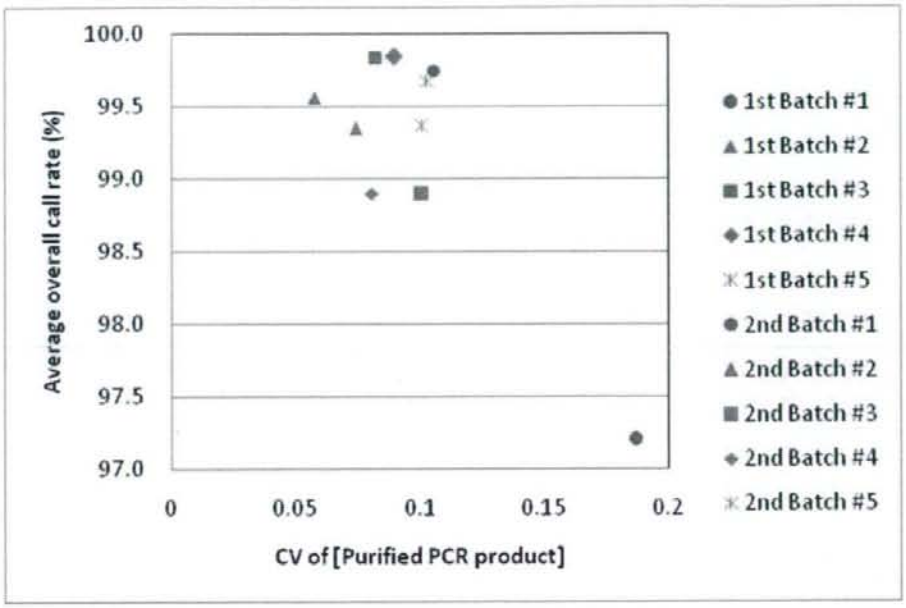

Figure 3
Assay criteria for experimental errors occurring on running batches. The CV of purified PCR product concentration is determined for each running batch. Overall call rate for each sample was determined by the Birdseed algorithm using samples in the same batch.

rate (Figure 6a). When we applied 95% QC criteria, 189 samples passed the QC criteria and the average overall call rate improved from 99.58 to 99.65%. By comparing the overall call rate determined under the 95% QC criteria with that under the default criteria, 187 of 189 samples improved by an average of 0.018% in overall call rate; however, the remaining two samples showed decreased overall call rate (by 0.76% and 0.12%) (Figure 6b). These two samples were considered as outliers on the genotype calling with the Birdseed algorithm and had to be removed. We repeated the removal of samples until none had a lower overall call rate than that determined under the default criteria. A total of 184 samples had an overall call rate that improved over the one determined under the default criteria, with an average change of 0.035%. The average overall call rate for the 184 samples was 99.71%, which was 0.13% higher than the default QC criteria (Figure 6c).

Number of SNPs available for GWAS in the Japanese population

The genotype calls of 909,622 SNPs were determined with 184 samples after sample filtering with adjusted QC criteria. However, these genotype calls still included inaccurate SNPs, which could lead to inflation of false positives, presumably due to systematically miss-called SNPs. Therefore, SNP filtering was considered to be important for a reliable and accurate set of genotype calls that avoid false association signals and false negative signals, allowing rapid identification of disease susceptibility genetic factors. We reported that the poorly behaving SNPs were effectively eliminated with the SNP filtering parameters; MAF > 5% or 1%, HWE p-value > 0.001 and SNP call rate > 95% [14]. Here, SNP call rate was defined for each SNP as the number of successfully genotyped samples divided by the number of total samples genotyped.

Among a total of 909,622 SNPs genotyped using 184 samples, 590,248 SNPs passed the three SNP filtering criteria with MAF > 5%, HWE p-value > 0.001 and SNP call rate > 95%, while 661,559 SNPs passed with MAF > 1%, HWE p-value > 0.001, and SNP call rate > 95%. A total of 180,859 SNPs were observed to be monomorphic in the Japanese population.

Discussion

The emerging SNP typing technologies have enabled genome-wide association studies to be conducted with hundreds of thousands of genotyped SNPs. According to Affymetrix, the SNP Array 6.0 platform can genotype over

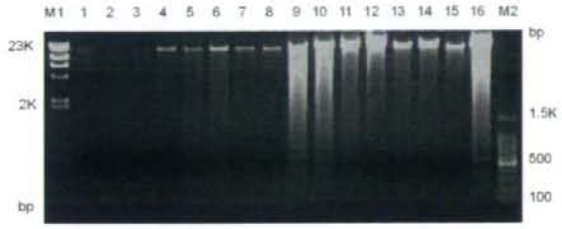

Figure 4
Agarose gel electrophoresis pattern showing genomic DNA from batch #1 of the 2nd set (lanes 1–8) and batch #2 of the 2nd set (lanes 9–16). Fifty nanograms of genomic DNA for each of the sample was electrophoresed on 1.0% agarose gels. M1 and M2 indicate lambda DNA digested with Hind III and 100-bp DNA ladder marker, respectively.

900 K SNP markers across the human genome with an overall call rate of at least 97%, over 99.7% concordant with the HapMap genotypes, and the Mendelian inheritance consistency for 10 Trios of greater than 99.9% when performing analysis under the default 86% QC criteria. To evaluate the SNP 6.0 Array platform and the Birdseed genotype calling algorithm, we genotyped two sets of 200 non-HapMap Japanese samples using the SNP Array 6.0 platform.

When we applied the default 86% QC criteria, 2 samples out of the 1st set of 200 samples were excluded and the average overall call rate was 99.58%. There was one sample with an overall call rate of lower than 97% among the 198 samples. Here, we found a linear relationship between QC call rate and overall call rate. Therefore, we applied stringent QC criteria of over 95% in order to remove the low-quality samples and found that the average overall call rate for 189 samples passing the stringent QC criteria improved to 99.65%. Among the 189 samples, 187 samples had higher overall call rates than those determined under the default QC criteria; however, the remaining two samples showed lower overall call rates (by 0.76% and 0.12%). When we repeated the removal of samples until none had a lower overall call rate than the one determined under the default criteria, none of the remaining 184 samples with an overall call rate lower than 97%. The average overall call rate of 184 samples was thus improved to 99.71%. The decay of average overall call rate may be caused by some samples that pass the QC criteria, but still have a low overall call rate. We can thus improve overall call rate by removing these samples and adjusting the QC criteria.

One of the crucial points for the SNP array 6.0 platform is to prepare the exact amount of intact genomic DNA. A 10-fold excess amount of genomic DNA decreased the overall call rate of each sample to by about 80% and another study revealed that samples with less than 50 ng/μl genomic DNA show low overall call rates [15]. Therefore, we checked the concentration and condition of genomic DNA with the NanoDrop quatitation and agarose gel electrophoresis. The SNP array 6.0 platform has three check points to assess experimental errors prior to hybridization on GeneChip arrays. Here, we found that the CV of the purified PCR product concentration was another critical indicator prior to hybridization in assessing the performance of each running batches. We suggest that samples with a CV value over 0.15 are excluded from the remainder of the assay.

The genotype calling accuracy of the Birdseed algorithm was assessed by comparing the 909,622 genotype calls of 12 samples from among198 samples with over 86% QC criteria, to those of 12 samples determined with six different sample sizes; 12 samples, 24 samples, 36 samples, 48 samples, 72 samples and 96 samples. The concordance rate gradually increased as the number of samples increased. The average concordance rate was almost constant over 99.8%, when the genotype calls were determined with over 48 samples using the Birdseed algorithm. However, the average overall call rate of the 12 samples gradually decreased as the sample number increased from 48 to 198. We could explain the reasons why the overall call rate decreases, and why the concordance rate increases for these 12 samples in a grouping of samples greater than 48 by means of characteristic properties of the Birdseed algorithm and minor allele frequency of each SNP. When the sample number was smaller than 48, all of three clusters designating AA, AB and BB genotypes were rarely observed for the SNPs with low MAF. In such cases, the Birdseed algorithm would determine the genotype as a single cluster, however, would ambiguously genotype as AA, BB and AB (tend to miss-genotype). Therefore, high call rate and low concordance were observed with the sample number smaller than 48. In contrast, when the sample number was greater than 48, two or three clusters would be observed for many SNPs. For these SNPs, the Birdseed algorithm could determine the outlying samples from each cluster as "No Calls", leading to low call rate and high concordance.

We can accurately determine the genotype calls with high overall call rates by determining the genotype calls with more than 48 samples, after removing low-quality samples by adjusting the QC criteria. Our results showed that the SNP Array 6.0 platform reached the expected level reported by the manufacturer, with an average overall call rate of over 99.5% and an average concordance rate of

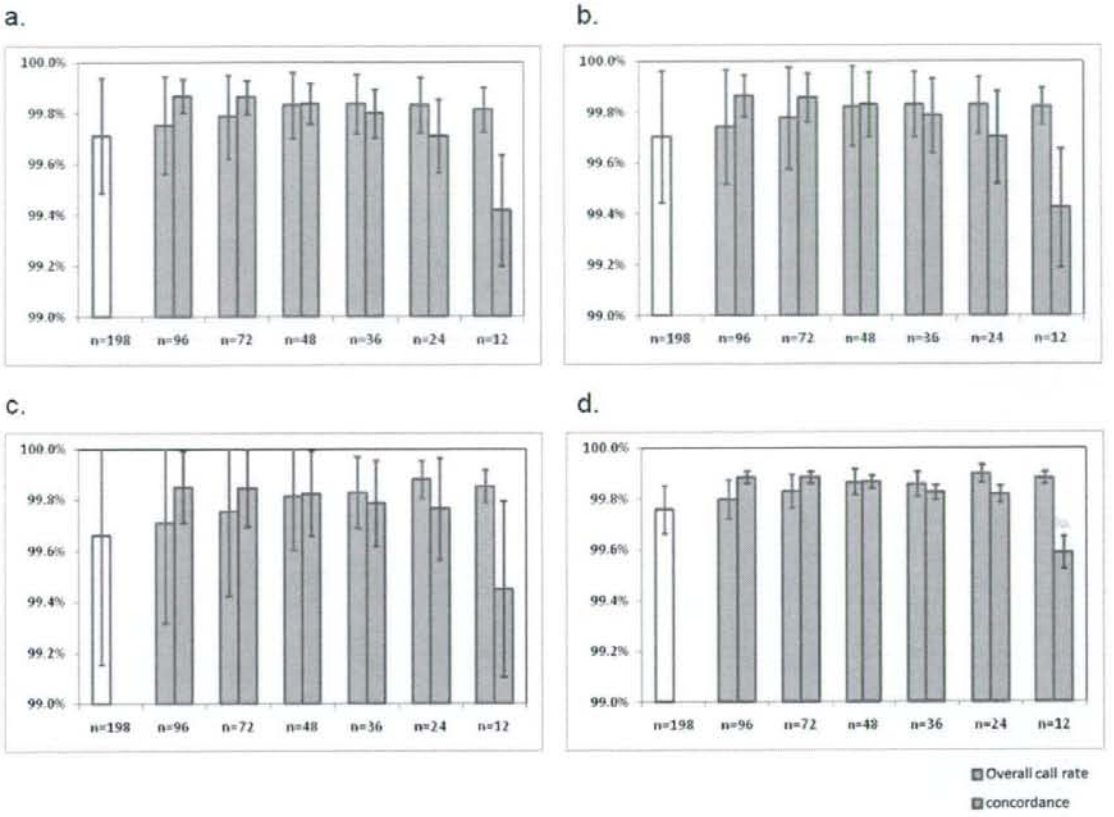

Figure 5
Genotype calling accuracy with Birdseed algorithm. a-d. Genotype calls determined using 198 samples with over 86% QC criteria were used as a reference. The average overall call rate for the 4 sets of the 12 samples were determined with 7 different sample sizes; 12 samples, 24 samples, 36 samples, 48 samples, 72 samples, 96 samples and 198 samples. The average concordance rates for the 4 sets of 12 samples were determined by comparison with the reference genotype calls. A negative correlation with a P value of 0.0053 and a positive correlation with a P value of 0.0115 were shown for overall call rate and concordance rate by fitting the power-law distribution to the data with least-squares approximation.

over 99.8%. However, about 20% of a total of 909,622 SNPs were found to be monomorphic in the Japanese population, which is due to SNP selection methods. The SNPs assayed on the SNP Array 6.0 platform were mainly selected as observed with high MAF in the Caucasian population. Among a total of 909,622 SNPs genotyped using the SNP Array 6.0 platform with 184 Japanese samples, 590,248 SNPs passed three SNP filtering criteria; MAF > 5%, HWE p-value > 0.001 and SNP call rate > 95%. Although the exact number of SNPs within the human genome remains under discussion, it has been reported that the genome coverage of the JPT + CHB population in the Phase II HapMap data was 66% using the Mapping 500 K Array set [10]. The genome coverage of the SNP array 6.0 platform was estimated using the same calculation and was revealed to be 75% with the 590,248 SNPs in the Japanese population.

Conclusion

The current Affymetrix SNP Array 6.0 platform enables the genotyping of over 900 K SNPs with high overall call rate (over 99.5%) and high concordance rate (over 99.8%). The number of SNPs available for GWAS in the Japanese population was revealed to be over 660 K SNPs, all of which passed the three SNP filtering criteria; MAF > 1%, HWE p-value > 0.001 and SNP call rate > 95%. GWAS

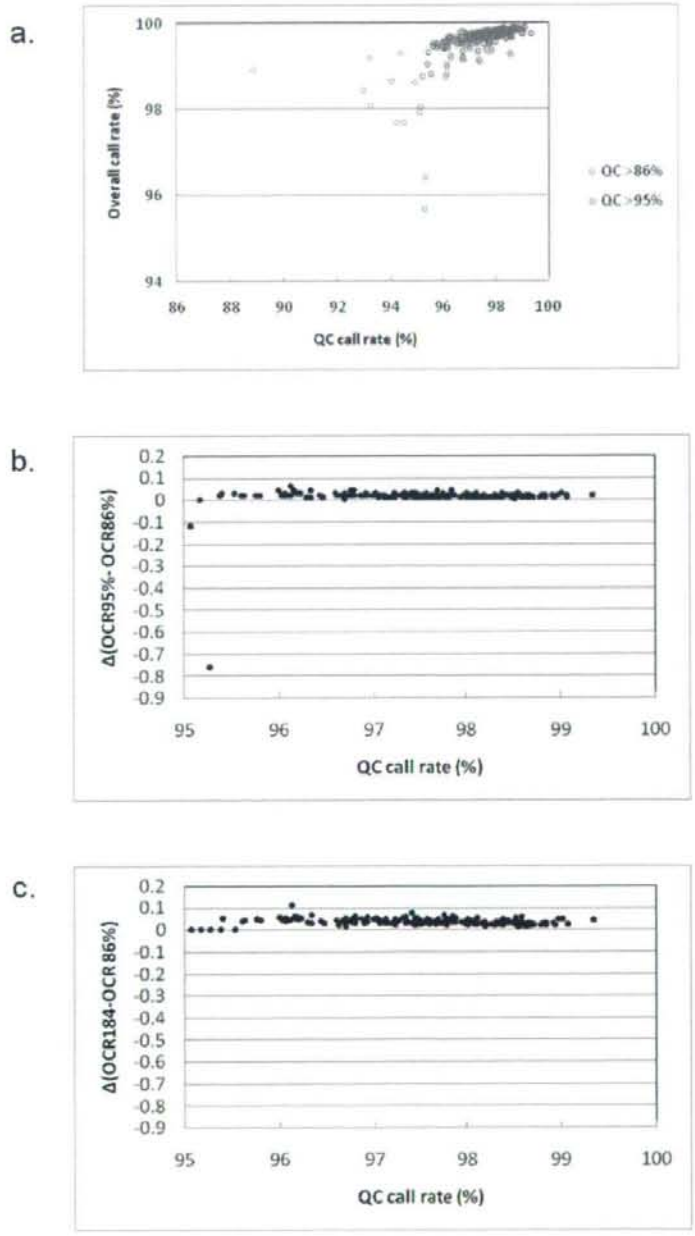

Figure 6
Removal of low-quality samples by adjusting QC criteria. Overall call rate for each sample was determined using total samples that passed the QC criteria. a. Overall call rate and QC call rate for each sample plotted with QC criteria > 86% and > 95%. b. Overall call rate (OCR) determined with 86% QC criteria compared with that determined with 95% QC criteria. c. Overall call rate (OCR) determined with 86% QC criteria compared with that determined using 184 samples.

using the SNP Array 6.0 platform has considerable potential in identifying candidate susceptibility or resistance genetic loci for multifactorial diseases in the Japanese population, as well as in other populations.

Finally, the genotyping data of 400 Japanese samples using the SNP array 6.0 platform will be deposited in a public database to share with the research community [16].

Methods

Study sample

Blood samples were obtained from two sets of 200 Japanese individuals in two institutes. Genomic DNA was extracted from peripheral blood leukocytes using the QIAamp Blood Mini Kit (Qiagen) according to the manufacturer's instructions. All genomic DNA was resuspended with Reduced EDTA TE Buffer (TEKnova) at 50 ng/μl. This study was approved by the Research Ethics Committee of the Faculty of Medicine, The University of Tokyo and Tokai University. Informed consent was obtained from all participants.

Genotyping 400 Japanese samples with SNP Array 6.0 platform

The concentration of genomic DNA for all individuals was measured using a spectrophotometer (NanoDrop ND-1000, NanoDrop Technologies). For the 1st set of 200 samples, five samples had low genomic DNA concentrations with an average of 41.1 ng/μl ranging from 38.2 to 44.5 ng/μl, and the remaining 195 samples had an average of 54.8 ng/μl, ranging from 45.0 to 57.8 ng/μl. For the 2nd set of 200 samples, one sample had 39.1 ng/μl and the remaining 199 samples had an average of 52.7 ng/μl, ranging from 45.0 to 55.9 ng/μl. For each individual assayed, 250 ng of genomic DNA was digested with Sty I and Nsp I (New England BioLabs) by adding 6 μl for the 6 samples with low concentration (five samples for 1st set and one sample for 2nd set) and 5 μl for the remaining samples. For two sets of 200 samples, every 48 samples were simultaneously processed in a single 96-well plate. After the reaction with restriction enzymes, we followed the manufacturer's instructions for the Affymetrix Genome-wide Human SNP array 6.0. The concentration of PCR products after purification with magnetic beads (Agencourt Magnetic Beads, Beckman) was measured using a spectrophotometer (NanoDrop ND-1000). Purified PCR products were diluted 10-fold with TE buffer (pH 8.0) (WAKO) in order to have a suitable concentration for the spectrophotometer to measure. The genotype calls of each individual were determined by the Birdseed version 1 genotype calling algorithm, embedded in the software Affymetrix Genotyping Console 2.0 (Affymetrix). The number of samples used to determine the genotype calls varied depending on the examination.

Acknowledgements

This work was supported by KAKENHI (Grant-in-Aid for Scientific Research) on Priority Areas "Comprehensive Genomics" from the Ministry of Education, Culture, Sports, Science and Technology of Japan, by a grant from the CREST program of the Japan Science and Technology Agency, Japan and by the contract research fund "Integrated Database Project" from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

References

1. Matsuzaki H, Loi H, Dong S, Tsai Y-Y, Fang J, Law J, Di X, Liu W-M, Yang G, Liu G, Huang J, Kennedy GC, Ryder TB, Marcus GA, Walsh PS, Shriver MD, Puck JM, Jones KW, Mei R: **Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array.** *Genome Res* 2004, **14**:414-425.
2. Steemers FJ, Gunderson KL: **Whole genome genotyping technologies on the BeadArray™ platform.** *Biotechnol J* 2007, **2**:41-49.
3. Ohnishi Y, Tanaka T, Ozaki K, Yamada R, Suzuki H, Nakamura Y: **A high-throughput SNP typing system for genome-wide association studies.** *J Hum Genet* 2001, **46**:471-477.
4. Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Sato H, Hori M, Nakamura Y, Tanaka T: **Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction.** *Nat Genet* 2002, **32**:650-654.
5. The Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447**:661-678.
6. Cupples LA, Arruda HT, Benjamin EJ, D'Agostino RB Sr, Demissie S, DeStefano AL, Dupuis J, Falls KM, Fox GS, Gottlib DJ, Govindaraju DR, Guo C-Y, Heard-Costa NL, Hwang S-J, Kathiresan S, Kiel DP, Laramie JM, Larson MG, Levy D, Liu C-Y, Lunetta KL, Mailman MD, Manning AK, Meigs JB, Murabito JM, Newton-Cheh C, O'Connor GT, O'Donnell CJ, Pandey M, Seshadri S, Vasan RS, Wang ZY, Wilk JB, Wolf PA, Yang Q, Atwood LD: **The Framingham Heart Study 100 K SNP genome-wide association study resource: overview of 17 phenotype working group reports.** *BMC Med Genet* 2007, **8**:s1.
7. Eeles RA, Kote-Jarai Z, Giles GG, Al Olama AA, Guy M, Jugurnauth SK, Mulholland S, Leongamornlert DA, Edwards SM, Morrison J, Field HI, Southey MC, Severi G, Donovan JL, Hamdy FC, Dearnaly DP, Muir KR, Smith C, Bagnato M, Ardern-Jones AT, Hall AL, O'Brien LT, Gehr-Swain BN, Wilkinson RA, Cox A, Lewis S, Brown PM, Jhavar SG, Tymrakiewicz M, Lophatananon A, Bryant SL, The UK Genetic Prostate Cancer Study Collaborators, British Association of Urological Surgeons' Section of Oncology, The UK ProtecT Study Collaborators, Horwich A, Huddart RA, Khoo VS, Parker CC, Woodhouse CJ, Thompson A, Christmas T, Ogden C, Fisher C, Jamieson C, Cooper CS, English DR, Hopper JL, Neal DE, Easton DF: **Multiplex newly identified loci associated with prostate cancer susceptibility.** *Nat Genet* 2008, **40**:316-321.
8. Kruglyak L, Nickerson DA: **Variation is the spice of life.** *Nat Genet* 2001, **27**:234-236.
9. **Affymetrix, Inc** [http://www.affymetrix.com/index.affx]
10. Barrett JC, Cardon LR: **Evaluating coverage of genome-wide association studies.** *Nat Genet* 2006, **38**:659-662.
11. Wollstein A, Herrmann A, Wittig M, Mothnagel M, Franke A, Nürnberg P, Schreiber S, Krawczak M, Hampe J: **Efficacy assessment of SNP sets for genome-wide disease association studies.** *Nucleic Acids Res* 2007, **35**:e113.
12. Mägi R, Pfeufer A, Nelis M, Montpetit A, Metspalu A, Remm M: **Evaluating the performance of commercial whole-genome marker sets for capturing common genetic variation.** *BMC Genomics* 2007, **8**:159-166.
13. Hua J, Craig DW, Brun M, Webster J, Zismann V, Tembe W, Joshipura K, Huentelman MJ, Dougherty ER, Stephan DA: **SNiPer-HD: improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays.** *Bioinformatics* 2007, **23**:57-63.
14. Miyagawa T, Nishida N, Ohashi J, Kimura R, Fujimoto A, Kawashima M, Koike A, Sasaki T, Tanii H, Otowa T, Momose Y, Nakahara Y, Gotoh J, Okazaki Y, Tsuji S, Tokunaga K: **Appropriate data cleaning methods for genome-wide association study.** *J Hum Genet* 2008 in press.

15. Woo JG, Sun G, Haverbusch M, Indugula S, Martin L, Broderick JP, Deka R, Woo D: **Quality assessment of buccal versus blood genomic DNA using the Affymetrix 500 K GeneChip.** *BMC Genet* 2007, **8**:79-83.
16. **Ministry of Education, Culture, Sports, Science, and Technology (MEXT) Integrated Database Project** [http://lifesciencedb.mext.go.jp/en/]

J Hum Genet (2008) 53:886–893
DOI 10.1007/s10038-008-0322-y

ORIGINAL ARTICLE

Appropriate data cleaning methods for genome-wide association study

Taku Miyagawa · Nao Nishida · Jun Ohashi · Ryosuke Kimura · Akihiro Fujimoto · Minae Kawashima · Asako Koike · Tsukasa Sasaki · Hisashi Tanii · Takeshi Otowa · Yoshio Momose · Yasuo Nakahara · Jun Gotoh · Yuji Okazaki · Shoji Tsuji · Katsushi Tokunaga

Received: 26 December 2007 / Accepted: 1 July 2008 / Published online: 12 August 2008

Abstract Genome-wide association studies (GWAS) using a large number of single nucleotide polymorphisms (SNPs) have successfully been applied to identify genetic variants of common diseases. However, genotyping using the new array technologies is often associated with spurious results that could unfavorably affect analyses of GWAS. Consequently, data cleaning is of paramount importance in excluding spurious genotyping results. In this study, we investigated the criteria required for the appropriate cleaning of 389 unrelated healthy Japanese samples analyzed using the GeneChip Human Mapping 500K Array Set for GWAS. The samples were randomly subdivided into two groups, and the allele frequencies in the groups were compared for individual SNPs as a quasi-case-control study. Then, observed results were filtered by four parameters (SNP call rate, confidence score obtained using the Bayesian Robust Linear Model with Mahalanobis genotype-calling algorithm, Hardy–Weinberg equilibrium, and minor allele frequency) and assessed for deviation from the null hypothesis. We found that appropriate data cleaning could be achieved using these four parameters. Our findings offer an avenue for obtaining appropriate data from GWAS.

Electronic supplementary material The online version of this article (doi:10.1007/s10038-008-0322-y) contains supplementary material, which is available to authorized users.

T. Miyagawa · N. Nishida · J. Ohashi · R. Kimura · A. Fujimoto · M. Kawashima · K. Tokunaga (✉)
Department of Human Genetics,
Graduate School of Medicine,
The University of Tokyo, 7-3-1 Hongo,
Bunkyo-ku, Tokyo 113-0033, Japan
e-mail: tokunaga@m.u-tokyo.ac.jp

R. Kimura
Department of Forensic Medicine,
Tokai University School of Medicine,
Isehara, Japan

M. Kawashima
Department of Sleep Disorder Research,
Graduate School of Medicine,
The University of Tokyo, Tokyo, Japan

A. Koike
Hitachi Ltd, Central Research Laboratory,
Kokubunji, Japan

T. Sasaki
Health Service Center, The University of Tokyo, Tokyo, Japan

H. Tanii
Department of Psychiatry, Mie University Graduate
School of Medicine, Tsu, Japan

T. Otowa
Department of Neuropsychiatry, Graduate School of Medicine,
The University of Tokyo, Tokyo, Japan

Y. Momose · Y. Nakahara · J. Gotoh · S. Tsuji
Department of Neurology, Graduate School of Medicine,
The University of Tokyo, Tokyo, Japan

Y. Momose · S. Tsuji
The 21st Century COE Program,
Center for Integrated Brain Medical Science,
Graduate School of Medicine,
The University of Tokyo, Tokyo, Japan

Y. Okazaki
Tokyo Metropolitan Matsuzawa Hospital, Tokyo, Japan

Keywords Genome-wide association study · Data cleaning

Introduction

One of the goals underlying the study of common diseases is to identify susceptibility and/or resistance genes associated with them. Previous studies of common diseases include two broad categories: family-based linkage studies across the entire genome, and population-based case-control association studies of candidate genes. Although there have been notable successes, progress has been slow. Linkage studies usually have low power except when a single locus explains a substantial fraction of the disease. In case-control association studies of candidate genes, researchers can target only those genes that have been functionally described.

In contrast, new high-throughput single nucleotide polymorphism (SNP) typing technologies for genome-wide association studies (GWAS) have recently been launched (Matsuzaki et al. 2004; Oliphant et al. 2002). GWAS provide opportunities to identify novel susceptibility and/or resistance loci without any prior information about gene functions. GWAS have successfully identified genetic variants associated with common diseases, including macular degeneration, QT interval prolongation, Crohn's disease, type 2 diabetes, and cerebral infarction (Arking et al. 2006; Dewan et al. 2006; Klein et al. 2005; Kubo et al. 2007; Rioux et al. 2007; Sladek et al. 2007; The Wellcome Trust Case Control Consortium 2007).

However, one fundamental problem is that the results obtained by the new array technologies are not always accurate. In such large data sets, small systematic differences can produce effects that are capable of obscuring the true associations being sought (Clayton et al. 2005; Zondervan and Cardon 2004). The major causes of inaccurate results are systematic errors of array reaction, incomplete genotype-calling algorithms, and SNPs in regions that exhibit copy number variations. Less accurate genotyping data unfavorably affects GWAS analyses. Therefore, data cleaning is of paramount importance, and data should be checked thoroughly (Balding 2006). At present, a consensus for data-cleaning criteria has not been established. In this study, we assessed the following parameters for data cleaning: (1) SNP call rate, (2) confidence score in the Bayesian Robust Linear Model with Mahalanobis (BRLMM) genotype-calling algorithm, (3) fitness to Hardy–Weinberg equilibrium (HWE), and (4) minor allele frequency (MAF). We typed 389 unrelated healthy Japanese samples by the GeneChip Human Mapping 500K Array Set and analyzed different thresholds for these four parameters.

Materials and methods

Subjects and genotyping

The subjects included 389 unrelated, healthy Japanese individuals living in Japan. This study was approved by the research ethics review committees of the University of Tokyo.

Genotyping of 500,568 SNPs was performed using the GeneChip Human Mapping 500K Array Set (Affymetrix). This array set consisted of two chips (Sty I and Nsp I) with approximately 250,000 SNPs each that were used for each individual. Approximately 250 ng of genomic DNA was digested with two restriction enzymes (*Sty*I and *Nsp*I) and processed according to the manufacturer's protocol. The genotyping calls were analyzed using the GCOS 1.4 and GTYPE 4.1 software packages, which adopted the BRLMM genotype-calling algorithm (Rabbee and Speed 2006). BRLMM performs a multiple-chip analysis, enabling simultaneous estimation of probe effects and allele signals for each SNP. The confidence score in BRLMM is assigned for each observation according to the normalized distance from the center of the genotype cluster. The confidence score is d1/d2, where d1 is the smallest distance of the three and d2 is the second-smallest distance. The confidence score threshold is the maximum score at which the algorithm will make a genotype call. All lower-quality confidence calls with scores greater than the threshold result in a no call.

Assessment of data cleaning

To assess the appropriate data-cleaning methods, the 389 healthy control samples were divided into two temporary groups (group A and group B). Each group was separately analyzed by the BRLMM algorithm of GTYPE 4.1 because the analytical capacity of this software was 250 samples (maximum). Then, to remove the bias associated with BRLMM analysis, group A and B samples were equally subdivided into two new groups (group 1 and group 2); group 1 consisted half of group A's and half of group B's samples, and group 2 consisted the remaining samples (Fig. 1). As a quasi-case-control study, group 1 was compared with group 2 using the chi-square test for the difference between allele frequencies of each SNP. SNPs on chromosome X were omitted because these groups were not matched by gender. Next, for reshuffling analysis, 100 new combination sets were prepared using the same 389 healthy control samples. To confirm that the appropriate data cleaning was reproducible in sets having the bias associated with BRLMM analysis, each set was formed from two groups separately analyzed by the BRLMM algorithm, such as the set of groups A and B (Fig. 1). Each set was also analyzed by the chi-square test for the difference in allele frequencies of each SNP.