40. Lengauer C, Kinzler KW, Vogelstein B (1997) Genetic instability in colorectal cancers. Nature 386: 623–627.
41. Kim GP, Colangelo LH, Paik S, O'Connell MJ, Kirsch IR, et al. (2007) Predictive value of microsatellite instability-high remains controversial. J Clin Oncol 25: 4857; author reply 4857–4858.
42. Elsaleh H, Joseph D, Grieu F, Zeps N, Spry N, et al. (2000) Association of tumour site and sex with survival benefit from adjuvant chemotherapy in colorectal cancer. Lancet 355: 1745–1750.
43. Gryfe R, Kim H, Hsieh ET, Aronson MD, Holowaty EJ, et al. (2000) Tumor microsatellite instability and clinical outcome in young patients with colorectal cancer. N Engl J Med 342: 69–77.
44. Ribic CM, Sargent DJ, Moore MJ, Thibodeau SN, French AJ, et al. (2003) Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer. N Engl J Med 349: 247–257.
45. Popat S, Hubner R, Houlston RS (2005) Systematic review of microsatellite instability and colorectal cancer prognosis. J Clin Oncol 23: 609–618.
46. Sinicrope FA, Rego RL, Halling KC, Foster N, Sargent DJ, et al. (2006) Prognostic impact of microsatellite instability and DNA ploidy in human colon carcinoma patients. Gastroenterology 131: 729–737.

OPEN ACCESS Freely available online

PLoS one

Highly Parallel and Short-Acting Amplification with Locus-Specific Primers to Detect Single Nucleotide Polymorphisms by the DigiTag2 Assay

Nao Nishida[1,2]*, Yoriko Mawatari[1,2], Megumi Sageshima[1], Katsushi Tokunaga[1]

1 Department of Human Genetics, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan, **2** Research Center for Hepatitis and Immunology, National Center for Global Health and Medicine, Ichikawa, Japan

Abstract

The DigiTag2 assay enables analysis of a set of 96 SNPs using Kapa 2GFast HotStart DNA polymerase with a new protocol that has a total running time of about 7 hours, which is 6 hours shorter than the previous protocol. Quality parameters (conversion rate, call rate, reproducibility and concordance) were at the same levels as when genotype calls were acquired using the previous protocol. Multiplex PCR with 192 pairs of locus-specific primers was available for target preparation in the DigiTag2 assay without the optimization of reaction conditions, and quality parameters had the same levels as those acquired with 96-plex PCR. The locus-specific primers were able to achieve sufficient (concentration of target amplicon $\geq$5 nM) and specific (concentration of unexpected amplicons $<$2 nM) amplification within 2 hours, were also able to achieve detectable amplifications even when working in a 96-plex or 192-plex form. The improved DigiTag2 assay will be an efficient platform for screening an intermediate number of SNPs (tens to hundreds of sites) in the replication analysis after genome-wide association study. Moreover, highly parallel and short-acting amplification with locus-specific primers may thus facilitate widespread application to other PCR-based assays.

Citation: Nishida N, Mawatari Y, Sageshima M, Tokunaga K (2012) Highly Parallel and Short-Acting Amplification with Locus-Specific Primers to Detect Single Nucleotide Polymorphisms by the DigiTag2 Assay. PLoS ONE 7(1): e29967. doi:10.1371/journal.pone.0029967

Editor: Javier S. Castresana, University of Navarra, Spain

Received September 26, 2011; **Accepted** December 9, 2011; **Published** January 13, 2012

Funding: This work was supported by a KAKENHI [grant number 22710191] Grant-in-Aid for Young Scientists (B) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan, and the Miyakawa Memorial Research Foundation. Partial support by the SENTAN program, Japan Science and Technology Agency, is also acknowledged. The funders had no direct role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: nishida-75@umin.ac.jp

Introduction

Polymerase chain reaction (PCR) is a commonly used technique in molecular biology. Several previously developed methods have employed multiplexed PCR in order to analyze genomic variations such as microsatellites or short tandem repeats (STRs), single nucleotide polymorphisms (SNPs) and insertions/deletions [1–3]. Multiplexed preparation of DNA templates in a single reaction is cost-effective, saving starting materials and run-time, while requiring careful optimization of assay conditions. The optimization process is highly empirical and time consuming, and depending on the combinations of markers, may or may not lead to successful assay development. For the conventional design of multiplex PCR, optimization of reaction conditions and careful pre-selection of targets are required in order to prevent excessive off-target priming by the numerous primers in the reaction. Moreover, the risk of generating errors in multiplex PCR, such as insufficient amplification, biased amplification and considerable primer-dimer formation within primers, tends to increase roughly as the square of the number of added primer pairs [4].

There are several approaches to resolving these drawbacks, including solid-phase assay formats (glass slide arrays, microbeads), oligonucleotides containing locked nucleic acid (LNA) residues and circularized amplification. Primers immobilized on the surface of the solid phase appear to markedly increase product yield on solid supports and may avoid the need for target pre-selection with a modification to enrich the input genomic DNA via a crude solution-phase multiplex PCR [5,6]. LNA pentamers showed high priming efficiency to achieve small biased priming in multiplex PCR [7]. Circularized amplification avoids generating artifacts associated with conventional multiplex PCR where two primers are used for each target [8]. This procedure was shown to perform a 96-plex amplification of an arbitrary set of specific DNA sequences. The arrayed primer extension-based genotyping method (APEX-2) allows efficient homogeneous 640-plex DNA amplification with locus-specific primers [9]. These approaches show effective consequences for multiplex amplification, however, a small number of approaches are practically used in the field of molecular genetics, presumably due to its cost and time consuming steps in preparation.

We developed the DigiTag2 assay for multiplex SNP typing as a simple and cost effective approach by combining multiplex PCR to enrich genetic regions including the target SNPs and an oligonucleotide ligation assay to encode all of the SNP genotypes into well-designed oligonucleotides designated DNA coded numbers (DCNs) [10]. For an effective primer design for multiplex PCR, there are several important physical properties for primer sequences, including melting temperature, Gibbs energy of duplex between primer and template, and interactions between primers and PCR amplicons. The DNA polymerase enzyme used in a multiplex PCR is one of the important factors for a successful unbiased amplification.

The DigiTag2 assay is a suitable approach to analyze an intermediate number of SNPs (tens to hundreds of locus) in the replication study after genome wide association study [11–12]. However, the most time consuming step for the DigiTag2 assay in a total running time of 13 hours is multiplex PCR for target preparation (5.5 hours). Here, we report an improved protocol for the DigiTag2 assay with a short-acting multiplex PCR through the use of Kapa 2GFast HotStart DNA polymerase, which reduces total running time and increases assay throughput. In this study, we also validate the applicability of the 192-plex PCR with locus specific primers to amplify the target regions from genomic DNA, which leads to save genomic DNA samples.

Methods

DNA samples

Genomic DNA samples from 96 unrelated healthy donors were obtained from the Japan Health Science Research Resources Bank (Osaka, Japan). All donors provided written informed consent and samples were anonymized. One microgram of purified genomic DNA was dissolved in 100 µl of TE buffer (pH 8.0) (Wako, Osaka, Japan), followed by storage at −20°C until use.

Primer design

A total of 192 pairs of primer were designed using the Visual OMP software version 7.1.0.0 (DNA software, Ann Arbor, MI, USA) with relatively long length (35–45-mer; average, 39.5-mer) to give amplicon sizes between 312 bp and 995 bp (average, 589 bp), each of which had an SNP site (Table S1). Prediction of DNA melting temperature was calculated using nearest-neighbor thermodynamic models. To avoid spurious amplification products, we employed a two-step protocol (denature and extension steps) using specifically designed primer pairs with an extension temperature at 68°C. The specificity of primer sequences was verified by Blat search in order to predict its location(s) on the human genome (GRCh37), and to confirm no unexpected SNP(s) within the primer sequence. The specificity of primer pairs was verified using MFE primer software, which can predict potential amplicon(s) generated from the human genome (GRCh37, up to 5 kb in amplicon size) [13]. All oligonucleotides (de-salted, 100 pmol/µl in TE (10 mM Tris-HCl, pH 8.0, 1 mM EDTA)) were purchased from Life Technologies (Carlsbad, CA, USA), and were stored at −20°C.

Multiplex PCR with Kapa 2GFast HotStart DNA polymerase

Multiplex PCR mix had a final volume of 10 µl, including 10 ng of genomic DNA, 25 nM each primer, 1.5× KAPA2G Buffer (including 2.25 mM Mg^{2+}), an additional 2.25 mM Mg^{2+} (final concentration of Mg^{2+}: 4.5 mM), 0.2 mM dNTPs and 0.4 U of Kapa 2GFast HotStart DNA polymerase (Kapa Biosystems, Woburn, MA, USA). PCR amplification was conducted using a TGradient (Biometra, Göttingen, Germany) or PTC-225 (MJ Research, Waltham, MA, USA) as follows: 95°C for 3 min, followed by 40 cycles of 95°C for 15 s and 68°C for 2 min. When necessary, the fragment length of PCR products was confirmed by capillary electrophoresis (Agilent 2100 Bioanalyzer, Agilent, Santa Clara, CA, USA) in order to evaluate PCR efficiency. The total running times for multiplex PCR with Kapa 2GFast HotStart DNA polymerase using TGradient and PTC-225 were 1 h 48 min 55 s and 2 h 6 min 59 s, respectively.

Multiplex PCR with QIAGEN Multiplex PCR Kit

Multiplex PCR mix had a final volume of 10 µl, including 10 ng of genomic DNA, 25 nM each primer, 1× Multiplex PCR Buffer (including 3.0 mM Mg^{2+}), 0.2 mM dNTPs and HotStar-Taq DNA polymerase (QIAGEN Multiplex PCR Kit; QIAGEN, Valencia, CA, USA). PCR amplification was conducted using a TGradient or PTC-225 as follows: 95°C for 15 min, followed by 40 cycles of 95°C for 30 s and 68°C for 6 min. The total running times for multiplex PCR with QIAGEN Multiplex PCR Kit using TGradient and PTC-225 were 5 h 27 min 53 s and 5 h 46 min 39 s, respectively.

96-plex genotyping by the DigiTag2 assay

The DigiTag2 assay performs multiplex SNP typing by encoding all of the SNP genotypes into well-designed oligonucleotides, designated DNA coded numbers (Figure 1, DCNs: D1_i, ED-1 and ED-2) [10]. The DCNs are assigned to the target SNPs in an unconstrained manner; therefore, the DNA chips prepared to read out the types of DCNs are universally available for any type of SNP without optimization of assay conditions. The DigiTag2 assay proceeds in four steps; target preparation, encoding, labeling and detection.

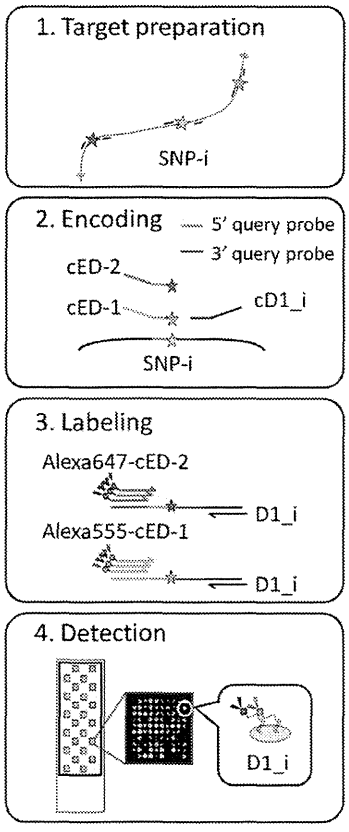


Figure 1. Schematic representation of the DigiTag2 assay. The assay has four steps: target preparation, encoding, labeling and detection. SNP genotypes are encoded into well-designed oligonucleotides, designated DNA coded numbers (DCNs: D1_i, ED-1 and ED-2). D1_i is a variable sequence assigned to each SNP. Reverse complement sequences are written by attaching the character 'c' before the sequence name.
doi:10.1371/journal.pone.0029967.g001

The encoding reactions had a final volume of 15 µl, including 0.5 µl of multiplex PCR products, 20 mM Tris-HCl, pH 7.6, 25 mM potassium acetate, 10 mM magnesium acetate, 10 mM DTT, 1 mM NAD, 0.1% Triton X-100 (1× Taq DNA ligase buffer) with 0.33 nM of each probe and 5 U Taq DNA ligase (New England BioLabs, Ipswich, MA, USA). Encoding reactions were conducted using a TGradient or PTC-225 under the following conditions: 95°C for 5 min, followed by 58°C for 15 min. The reaction was stopped by holding the temperature at 10°C.

The labeling reactions had a final volume of 12 µl, including 6 µl of ligation products, 0.5 µM each labeled primer (Alexa555-cED-1 and Alexa647-cED-2), 2.5 nM each D1 primer (D1_i), 50 mM KCl, 2 mM Mg^{2+}, 0.1 mM DTT, 0.2 mM each dNTP (N = A, G, C), 0.1 mM [^{3}H]-dTTP, 0.25 mg/ml activated salmon sperm DNA (1× *Ex Taq* Buffer) and 0.05 U of *Ex Taq*™ polymerase (TaKaRa, Shiga, Japan). Labeling reactions were conducted using a TGradient or PTC-225 under the following conditions: first held at 95°C for 1 min, followed by 30 cycles of 95°C for 30 s, 55°C for 6 min and 72°C for 30 s. The reaction was stopped by holding the temperature at 10°C. Total running times for labeling using TGradient and PTC-225 were 3 h 49 min 48 s and 4 h 8 min 48 s, respectively.

In the detection step, a hybridization mixture was prepared by mixing 6.25 µl of labeling products with 8.75 µl of hybridization buffer containing 0.5× SSC, 0.1% SDS, 15% formamide, 1 mM EDTA and 3.125 fmol of hybridization control (Alexa555-labeled D1_100 and Alexa647-labeled D1_100). The hybridization control was prepared for ensuring the hybridization step. Ten microliters of hybridization mixture was applied to each block on the universal DNA chip. Hybridization was carried out for 30 min at 37°C in a hybridization oven (ThermoStat plus; Eppendorf, Ham, Germany). After hybridization, glass slides were washed in washing buffer (0.1× SSC, 0.1% SDS) by shaking at 60 rpm for 3 min. Glass slides were consecutively washed in distilled water by shaking at 60 rpm for 1 min and then dried up by centrifugation at 500× g for 1 min. Hybridization images were scanned at photomultiplier voltages of 400 V for Alexa555 and 480 V for Alexa647 using a commercially available DNA chip scanner and fluorescence image analysis was performed using commercially available software (GenePix 4000B unit and GenePix Pro 4.1 software package; Molecular Devices, Sunnyvale, CA, USA).

Labeling with Kapa 2GFast HotStart DNA polymerase

The labeling reactions with Kapa 2GFast HotStart DNA polymerase had a final volume of 12 µl, including 6 µl of ligation products, 0.5 µM each labeled primer (Alexa555-cED-1 and Alexa647-cED-2), 2.5 nM each D1 primer (D1_i), 1.5× KAPA2G Buffer (including 2.25 mM Mg^{2+}), an additional 2.25 mM Mg^{2+} (final concentration of Mg^{2+}: 4.5 mM), 0.2 mM dNTPs and 0.4 U of Kapa 2GFast HotStart DNA polymerase. Labeling reactions were conducted using a TGradient or PTC-225 under the following conditions: first held at 95°C for 1 min, followed by 30 cycles of 95°C for 15 s, 55°C for 120 s and 72°C for 5 s. The reaction was stopped by holding the temperature at 10°C. The total running times for labeling using TGradient and PTC-225 were 1 h 29 min 48 s and 1 h 48 min 34 s, respectively.

Results

Singleplex PCR using 192 pairs of locus-specific primers

Singleplex PCR was conducted under the same reaction condition with multiplex PCR using 25 ng of genomic DNA to ensure target amplicon detection and to confirm the emergence of extra bands (unexpected amplicons). Singleplex PCR with 192 pairs of locus-specific primers revealed that most of the primer pairs are able to achieve sensitive detection (concentration of target amplicon ≥5 nM) and specific amplification without extra bands (concentration of unexpected amplicons <2 nM) except for 14 pairs of primers; low sensitivity (<5 nM) for 5 pairs of primers (61, 99, 102, 189 and 191) and low specificity with extra bands (≥2 nM) for 9 pairs of primers (40, 56, 62, 70, 91, 106, 149, 173 and 174) (Figure 2 and Table S2). Five pairs among the 9 low-specific primer pairs with extra bands (62, 70, 149, 173 and 174) resulted from heteroduplex formation of target amplicons during polyacrylamide gel electrophoresis. Despite the presence of extra bands, the remaining 4 pairs of low-specific primers had a target amplicon with a detectable concentration ≥5 nM.

Validation of efficacy of 192-plex PCR by 96-plex genotyping with the DigiTag2 assay

The DigiTag2 assay enables the simultaneous analysis of 96 target SNPs in: (1) multiplex PCR with locus-specific primers to amplify target genomic regions including target SNPs; (2) multiple oligonucleotide ligation assay with locus-specific probes to determine the genotype of each SNP; and (3) hybridization to the universal DNA chip tethered with probe sequences identical to D1_i (23-mer) (Figure 1) [10]. The validity of 192-plex PCR was assessed with 96 individual DNAs (population control samples) by comparing two sets of 96-plex genotype calls acquired from 96-plex PCR with those from 192-plex PCR (Table 1).

Conversion rate shows the proportion of successfully genotyped SNPs with fewer than 3 undetected samples after excluding low-quality genotyping data, which had more than 5 undetected SNPs in a total of 96 SNPs. However, the composition of failed SNPs in genotyping was not identical, and the conversion rate showed no differences between 192-plex PCR and 96-plex PCR. For the 1st set of 96 SNPs, 7 SNPs among 10 failed SNPs were matched between 192-plex PCR and 96-plex PCR, and for the 2nd set, 8 SNPs among the 9 failed SNPs were matched. The average call rate for successfully genotyped SNPs was over 99.79% for both sets of 96-plex genotyping, even if 192-plex PCR products were adopted for target preparation. Reproducibility was determined by independent genotyping with 96 individuals twice. As a consequence, four discordant genotype calls were observed in the duplicated genotyping data. Concordance of genotype calls between 192-plex PCR and 96-plex PCR was determined using 6,290 genotype calls for the 1st set and 7,884 genotype calls for the 2nd set. Consequently, 14,171 out of 14,174 genotype calls were matched by comparison with 83 SNPs for the 1st set and 86 SNPs for the 2nd set. In total, 3 discordant genotype calls were observed (Figure 3).

Short-acting multiplex PCR by use of Kapa 2GFast HotStart DNA polymerase

Kapa 2GFast HotStart DNA polymerase was employed to perform multiplex PCR with the locus-specific primers for target preparation in genotyping with the DigiTag2 assay. To optimize reaction conditions with Kapa 2GFast HotStart DNA polymerase, singleplex PCR was conducted using 25 ng of genomic DNA with three randomly chosen pairs of locus-specific primers. The designed amplicon sizes for the three pairs of primers were 501 bp, 671 bp and 492 bp. We performed singleplex PCR using a two-step protocol (denature and extension steps) with varied extension periods (15 s, 30 s, 60 s and 120 s) and with varied Mg^{2+} concentrations (3.0 mM and 4.5 mM) (Figure 4). The most sensitive detection and highest levels of amplification for the three

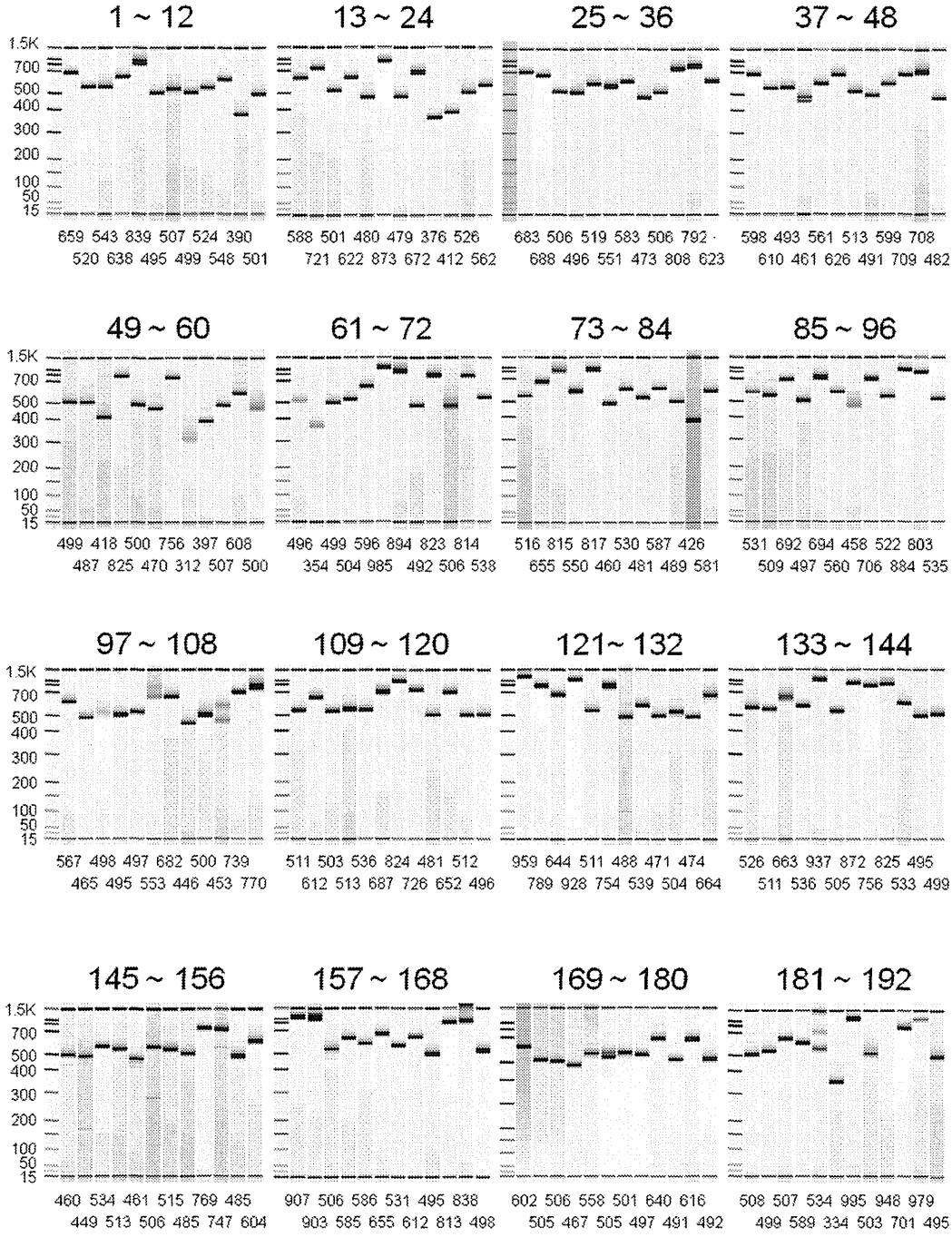


Figure 2. Electropherogram of singleplex PCR products with 192 pairs of locus-specific primers. The designed amplicon size is depicted below each lane.
doi:10.1371/journal.pone.0029967.g002

pairs of primers were observed with 120 s for the extension period and 4.5 mM for the Mg^{2+} concentration. The total running time for multiplex PCR with locus-specific primers was less than 2 hours, which is about 3 h 30 min shorter than the previous protocol (see MATERIALS AND METHODS).

The total running time of the DigiTag2 assay was markedly reduced when the labeling step was also conducted using Kapa 2GFast HotStart DNA polymerase instead of *Ex Taq* polymerase. When the DigiTag2 assay was conducted with Kapa 2GFast HotStart DNA polymerase for multiplex PCR and labeling step, the total running time of the assay was about 7 hours, which is about 6 hours shorter than the previously used protocol in combination with QIAGEN Multiplex PCR Kit for multiplex PCR and *Ex Taq* polymerase for the labeling step.

Table 1. Validation of efficacy of 192-plex PCR by 96-plex genotyping.

		192-plex PCR	96-plex PCR
1st set	Conversion rate	86/96 SNP	86/96 SNP
	Call rate	99.84% (7,728/7,740 genotype)	99.81% (6,695/6,708 genotype)
	reproducibility	99.99% (7,288/7,289 genotype)	100% (6,121/6,121 genotype)
	concordance	99.98% (6,289/6,290 genotype)	
2nd set	Conversion rate	87/96 SNP	87/96 SNP
	Call rate	99.79% (8,074/8,091 genotype)	99.79% (8,161/8,178 genotype)
	reproducibility	99.97% (7,792/7,794 genotype)	99.99% (7,712/7,713 genotype)
	concordance	99.97% (7,882/7,884 genotype)	

doi:10.1371/journal.pone.0029967.t001

Table 2 summarizes the quality parameters (conversion rate, call rate, reproducibility and concordance) when genotyping was conducted with 192-plex PCR or 96-plex PCR by use of Kapa 2GFast HotStart DNA polymerase. The conversion rate was slightly decreased when multiplex PCR was conducted in 192-plex form. However, the conversion rates were better than those observed when multiplex PCR was conducted with the QIAGEN Multiplex PCR Kit. The composition of failed SNPs in genotyping was not consistent for the 1st set of 96 SNPs, in which 4 SNPs were matched between 192-plex PCR and 96-plex PCR. For the 2nd set, a total of 8 failed SNPs in the 96-plex PCR were completely matched to those in the 192-plex PCR. When the composition of failed SNPs were compared between Kapa 2GFast HotStart DNA polymerase and QIAGEN Multiplex PCR Kit, the 1st set had 5 matched SNPs in a total of 8 failed SNPs for 192-plex PCR, and 4 matched SNPs in 5 failed SNPs for 96-plex PCR. From the 2nd set, 5 SNPs in a total of 9 failed SNPs were matched when 192-plex PCR was conducted and 4 SNPs in a total of 8 failed SNPs were matched when 96-plex PCR was conducted. The average call rate for successfully genotyped SNPs was over 99.76% for both sets of 96-plex genotyping, even if 192-plex PCR products were adopted for target preparation. The reproducibility was 100% for the 2nd set; however, three discordant genotype calls were observed for the 1st set. With regard to the concordance of genotype calls between 96-plex PCR and 192-plex PCR, only one discordant genotype call was observed in the comparison for the 1st set, and no discordant genotype calls were observed in the 2nd set.

Table 3 shows the concordance rate in comparison with the genotype calls by the use of Kapa 2GFast HotStart DNA polymerase or QIAGEN Multiplex PCR Kit for multiplex PCR. For the 1st set, there were 4 discordant genotype calls with 96-plex PCR and 8 discordant genotype calls with 192-plex PCR. For the 2nd set of 96 SNPs, there was one discordant genotype call in genotyping with 96-plex PCR and 192-plex PCR.

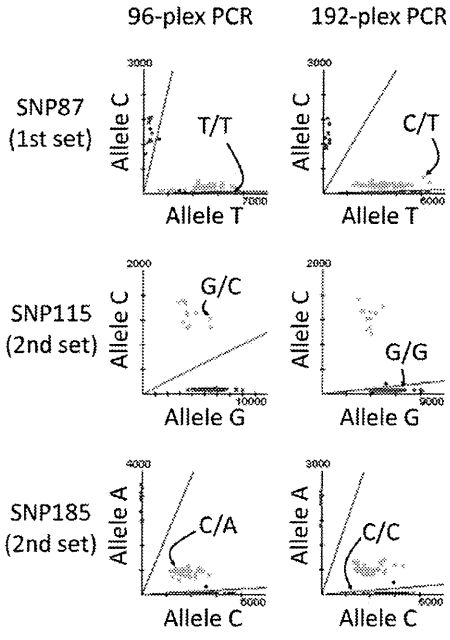


Figure 3. Scatter plots for three SNPs with 3 discordant genotypes. Scatter plots in genotyping with 192-plex PCR and 96-plex PCR are depicted side-by-side. The genotypes of discordant samples are indicated in the scatter plots by arrows.
doi:10.1371/journal.pone.0029967.g003

Discussion

The locus specific primers sufficiently worked in a multiplex form under the same reaction conditions without any optimization processes, either 96-plex PCR or 192-plex PCR. We also found that either 96-plex PCR or 192-plex PCR could be accomplished within two hours through the use of Kapa 2GFast HotStart DNA polymerase. The total running time of the DigiTag2 assay was shortened by 6 hours over the original 13-hour long protocol using Kapa 2GFast HotStart DNA polymerase for both multiplex PCR and the labeling step. The quality parameters (conversion rate, call rate, reproducibility and concordance) observed in genotyping with the new protocol were the same as those observed in the original protocol using QIAGEN Multiplex PCR Kit for multiplex PCR and *Ex Taq* polymerase for the labeling step. The DigiTag2 assay worked with a conversion rate of over 93.2% (179 / 192 SNPs), average call rate of over 99.80% (16,789/16,823 genotypes) and reproducibility of over 99.99% (16,135/16,136 genotypes) using 96-plex PCR under the new protocol. The composition of successfully genotyped SNPs was different when the genotype calls were acquired using the different polymerases (Kapa 2GFast HotStart DNA polymerase and QIAGEN Multiplex PCR Kit), which would result from a varying amplification bias in multiplex PCR. We also found that 192-plex PCR with locus-specific primers worked in 96-plex genotyping with the DigiTag2 assay, giving the same quality parameter data as those observed in genotyping with 96-plex PCR. However, the

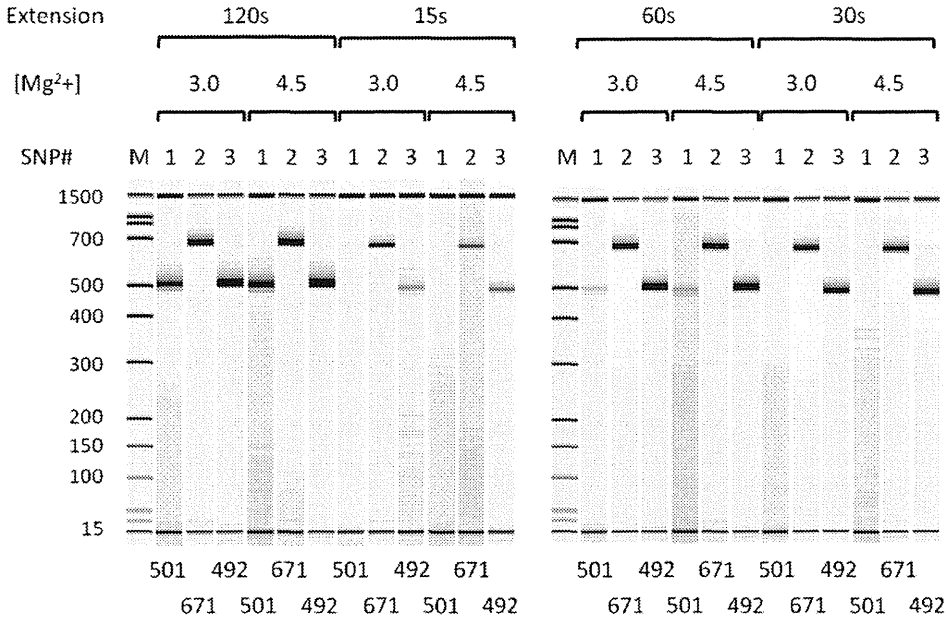


Figure 4. Electropherogram of singleplex PCR products using Kapa 2GFast HotStart DNA polymerase. Singleplex PCR was performed with varied extension periods (15 s, 30 s, 60 s and 120 s) and with varied Mg^{2+} concentrations (3.0 mM and 4.5 mM) using three pairs of locus-specific primers. The designed amplicon size is depicted below each lane.
doi:10.1371/journal.pone.0029967.g004

composition of successfully genotyped SNPs was not consistent between 192-plex PCR and 96-plex PCR, which may be explained by changing the interactions between primer pairs in 192-plex PCR and in 96-plex PCR. The composition of successful SNPs was not consistent when using different polymerases or multiplex systems in the multiplex PCR, which casts some shadows on the reliability of the assay. Regardless of the existing shadows, indeed, 96-plex and 192-plex PCR work with a high conversion rate in genotyping over 93.2%. To clear the existing shadows, it is necessary to continuously accumulate genotyping data.

In this study, fifteen discordant genotype calls were in total observed in the comparison of genotype calls with: i) duplicated genotyping data; ii) genotyping data by use of 192-plex PCR and 96-plex PCR; and iii) genotyping data with different types of polymerases (Table S3). Table S3 shows the genotype calls acquired 8 times under different conditions. All fifteen discordant genotype calls were analyzed with direct sequencing, of which 13 genotype calls were determined. In 8 of 15 discordant genotype calls, the genotype calls were completely different depending on the type of polymerase. The genotype calls acquired using Kapa 2GFast HotStart DNA polymerase were 100% concordant (6 of 6) with those acquired by direct sequencing. This suggests that SNP allelic bias in PCR amplification readily occurred with the QIAGEN Multiplex PCR Kit; however, the error rate in genotyping was only 0.04% (6 out of 14,886 genotypes). The remaining 7 discordant genotype calls were randomly observed in 1 out of 8 different conditions. This shows that the random error rates were almost equal in the genotype data acquired with both types of polymerases (4 out of 62,227 genotypes for QIAGEN Multiplex PCR Kit and 3 out of 66,008 genotypes for Kapa 2GFast HotStart DNA polymerase).

Among the five low-sensitivity primer pairs found on singleplex PCR (61, 99, 102, 189 and 191), no amplicons were detected by primer pair 189 and low concentrations (<5 nM) of amplicon were detected by the 4 other primer pairs (Table S2). Therefore, the SNP189 failed in genotyping, independently of the type of

Table 2. Validation of efficacy of 192-plex and 96-plex PCR with Kapa 2GFast HotStart DNA polymerase.

		192-plex PCR	96-plex PCR
1st set	Conversion rate	88/96 SNP	91/96 SNP
	Call rate	99.84% (8,259/8,272 genotype)	99.76% (8,443/8,463 genotype)
	reproducibility	99.97% (8,069/8,071 genotype)	99.99% (8,339/8,340 genotype)
	concordance	99.99% (7,982/7,983 genotype)	
2nd set	Conversion rate	87/96 SNP	88/96 SNP
	Call rate	99.91% (8,171/8,178 genotype)	99.83% (8,346/8,360 genotype)
	reproducibility	100% (7,705/7,705 genotype)	100% (7,796/7,796 genotype)
	concordance	100% (8,161/8,161 genotype)	

doi:10.1371/journal.pone.0029967.t002

Table 3. Concordance of genotype calls between Kapa 2GFast HotStart DNA polymerase and QIAGEN Multiplex PCR Kit.

		Kapa 2G	QIAGEN
1st set	96-plex PCR	99.94% (6,513/6,517 genotype)	
	192-plex PCR	99.89% (7,441/7,449 genotype)	
2nd set	96-plex PCR	99.99% (7,778/7,779 genotype)	
	192-plex PCR	99.99% (7,700/7,701 genotype)	

doi:10.1371/journal.pone.0029967.t003

polymerase and multiplicity in multiplex PCR (192-plex or 96-plex). However, the SNP191, which was amplified by primer pair 191, was successfully genotyped only when the QIAGEN Multiplex PCR Kit was used for the multiplex PCR. The concentration of amplicon amplified by primer pair 99 was the same as the 2.8 nM observed with the amplicon amplified by primer pair 191. SNP99, which was amplified by primer pair 99, was successfully genotyped independently of polymerase type and multiplicity in multiplex PCR (192-plex or 96-plex). These results suggest that the sensitivity in genotyping with Kapa 2GFast HotStart DNA polymerase was lower than the previously used protocol with QIAGEN Multiplex PCR Kit. These results would be explained by a biased amplification with the shortened protocol using Kapa 2GFast HotStart DNA polymerase, which tends to lead to a consequent biased genotyping. However, the investigated number of primer pairs would not be sufficient to decide the sensitivity in genotyping; therefore, it is necessary to continuously accumulate genotyping data. As the investigated number of primer pairs was only 192 (384 primers) in this study, melting temperature of each primer and the number of potential amplicons predicted by the MFE primer software were strongly associated with low sensitivity and low specificity in an amplification, respectively (multiple regression analysis, $P = 1.26 \times 10^{-37}$ and $P = 1.52 \times 10^{-21}$, respectively).

Through the use of Kapa 2GFast HotStart DNA polymerase, the genotype calls for 96 SNPs can be acquired in about 7 hours by the DigiTag2 assay. The genotyping platform with high conversion rate plays an important role for the replication studies to identify the disease associated genes from candidate loci found in the GWAS (genome-wide association study). The DigiTag2 assay with an improved protocol will be an efficient platform for screening an intermediate number of SNPs (tens to hundreds of sites) in the replication studies. Because of limitations in the variation of DNA coded numbers (DCNs), 192-plex genotyping is not available for the current DigiTag2 assay. However, 192-plex PCR can save genomic DNA samples and time for target preparation. Moreover, 192-plex PCR is also available for direct-sequencing and other PCR-based assays to amplify the target regions from genomic DNA.

Supporting Information

Table S1 Sequence information of 192 pairs of locus specific primer.
(XLSX)

Table S2 Results of singleplex PCR with 192 pairs of locus specific primer.
(XLSX)

Table S3 The 15 discordant genotype calls in 8 different conditions.
(XLSX)

Acknowledgments

We would like to thank M. Takasu for technical support, and H. Adachi, N. Tabei and J. Fujimiya (Dynacom Co., Ltd.) for assistance with primer and probe design.

Author Contributions

Conceived and designed the experiments: NN KT. Performed the experiments: YM MS. Analyzed the data: NN YM MS. Contributed reagents/materials/analysis tools: NN YM MS. Wrote the paper: NN KT.

References

1. Deter J, Gala M, Charbonnel N, Cosson JF (2009) Characterization and PCR multiplexing of polymorphic microsatellite loci in the whipworm *Trichuris arvicolae*, parasite of arvicoline rodents and their cross-species utilization in *T. muris*, parasite of murines. Mol Biochem Parasitol 167: 144–146.
2. Hosseini-Maaf B, Hellberg Å, Chester MA, Olsson ML (2007) An extensive polymerase chain reaction–allele-specific polymorphism strategy for clinical ABO blood group genotyping that avoids potential errors caused by null, subgroup, and hybrid alleles. Transfusion 47: 2110–2125.
3. Goguet de la Salmonière YO, Kim CC, Tsolaki AG, Pym AS, Siegrist MS, et al. (2004) High-throughput method for detecting genomic-deletion polymorphisms. J Clin Microbiol 42: 2913–2918.
4. Landegren U, Nilsson M (1997) Locked on target: strategies for future gene diagnostics. Ann Med 29: 585–590.
5. Pemov A, Modi H, Chandler DP, Bavykin S (2005) DNA analysis with multiplex microarray-enhanced PCR. Nucleic Acids Res 33: e11.
6. Meuzelaar LS, Lancaster O, Pasche JP, Kopal G, Brookes AJ (2007) MegaPlex PCR: a strategy for multiplex amplification. Nat Methods 4: 835–837.
7. Sun Z, Chen Z, Hou X, Li S, Zhu H, et al. (2008) Locked nucleic acid pentamers as universal PCR primers for genomic DNA amplification. Plos One 3: e3701.
8. Dahl F, Gullberg M, Stenberg J, Landegren U, Nilsson M (2005) Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. Nucleic Acids Res 33: e71.
9. Krjutškov K, Andreson R, Mägi R, Nikopensius T, Khrunin A, et al. (2008) Development of a single tube 640-plex genotyping method for detection of nucleic acid variations on microarrays. Nucleic Acids Res 36: e75.
10. Nishida N, Tanabe T, Takasu M, Suyama A, Tokunaga K (2007) Further development of multiplex single nucleotide polymorphism typing method, the DigiTag2 assay. Anal Biochem 364: 78–85.
11. Tanaka Y, Nishida N, Sugiyama M, Kurosaki M, Matsuura K, et al. (2009) Genome-wide association of IL28B with response to pegylated interferon-alpha and ribavirin therapy for chronic hepatitis C. Nat Genet 41: 1105–1109.
12. Miyagawa T, Kawashima M, Nishida N, Ohashi J, Kimura R, et al. (2009) Variant between CPT1B and CHKB associated with susceptibility to narcolepsy. Nat Genet 40: 1324–1328.
13. Qu W, Shen Z, Zhao D, Yang Y, Zhang C (2009) MFEprimer: multiple factor evaluation of the specificity of PCR primers. Bioinformatics 25: 276–278.

J Gastroenterol
DOI 10.1007/s00535-010-0322-5

Pretreatment prediction of response to peginterferon plus ribavirin therapy in genotype 1 chronic hepatitis C using data mining analysis

Masayuki Kurosaki · Naoya Sakamoto · Manabu Iwasaki · Minoru Sakamoto · Yoshiyuki Suzuki · Naoki Hiramatsu · Fuminaka Sugauchi · Hiroshi Yatsuhashi · Namiki Izumi

Received: 22 June 2010 / Accepted: 21 August 2010

Abstract
Background This study aimed to develop a model for the pre-treatment prediction of sustained virological response (SVR) to peg-interferon plus ribavirin therapy in chronic hepatitis C.
Methods Data from 800 genotype 1b chronic hepatitis C patients with high viral load (>100,000 IU/ml) treated by peg-interferon plus ribavirin at 6 hospitals in Japan were randomly assigned to a model building ($n = 506$) or an internal validation ($n = 294$). Data from 524 patients treated at 29 hospitals in Japan were used for an external validation. Factors predictive of SVR were explored using data mining analysis.
Results Age (<50 years), alpha-fetoprotein (AFP) (<8 ng/mL), platelet count ($\geq 120 \times 10^9$/l), gamma-glutamyl-transferase (GGT) (<40 IU/l), and male gender were used to build the decision tree model, which divided patients into 7 subgroups with variable rates of SVR ranging from 22 to 77%. The reproducibility of the model was confirmed by the internal and external validation ($r^2 = 0.92$ and 0.93, respectively). When reconstructed into 3 groups, the rate of SVR was 75% for the high probability group, 44% for the intermediate probability group and 23% for the low probability group. Poor adherence to drugs lowered the rate of SVR in the low probability group, but not in the high probability group.

M. Kurosaki · N. Izumi (✉)
Division of Gastroenterology and Hepatology,
Musashino Red Cross Hospital, 1-26-1 Kyonan-cho,
Musashino, Tokyo 180-8610, Japan
e-mail: nizumi@musashino.jrc.or.jp

M. Kurosaki
e-mail: kurosaki@musashino.jrc.or.jp

N. Sakamoto
Department of Gastroenterology and Hepatology,
Tokyo Medical and Dental University, Tokyo, Japan
e-mail: nsakamoto.gast@tmd.ac.jp

M. Iwasaki
Department of Computer and Information Science,
Seikei University, Tokyo, Japan
e-mail: iwasaki@st.seikei.ac.jp

M. Sakamoto
First Department of Internal Medicine, University of Yamanashi,
Yamanashi, Japan
e-mail: msakamoto@yamanashi.ac.jp

Y. Suzuki
Department of Hepatology, Toranomon Hospital, Tokyo, Japan
e-mail: suzunari@interlink.or.jp

N. Hiramatsu
Department of Gastroenterology and Hepatology,
Osaka University Graduate School of Medicine,
Osaka, Japan
e-mail: hiramatsu@gh.med.osaka-u.ac.jp

F. Sugauchi
Department of Gastroenterology and Metabolism,
Nagoya City University Graduate School of Medical Sciences,
Nagoya, Japan
e-mail: fsugauch@med.nagoya-cu.ac.jp

H. Yatsuhashi
Clinical Research Center, National Nagasaki Medical Center,
Nagasaki, Japan
e-mail: yatsuhashi@nmc.hosp.go.jp

Published online: 10 September 2010

Conclusions A decision tree model that includes age, gender, AFP, platelet counts, and GGT is useful for predicting the probability of response to therapy with peg-interferon plus ribavirin and has the potential to support clinical decisions regarding the selection of patients for therapy.

Keywords Data mining · Decision tree · Alpha-fetoprotein · HCV · Peg-interferon

Introduction

The current standard therapy for genotype 1 chronic hepatitis C is 48 weeks of pegylated interferon (PEG-IFN) plus ribavirin (RBV) [1]. Sustained virological response (SVR), defined as undetectable HCVRNA post-treatment is regarded as a cure of chronic hepatitis C. However, the rate of SVR to this regimen is only 50% in patients with HCV genotype 1b and a high HCVRNA titer [2, 3]. Since PEG-IFN and RBV combination therapy is costly and accompanied by potential adverse effects, the ability to predict the possibility of SVR before therapy may significantly influence the selection of patients for therapy. A recent report revealed that single nucleotide polymorphisms located in the *IL28B* are strongly associated with a response to PEG-IFN plus RBV therapy [4–6]. Besides, the amino acid substitutions in the NS5A [7–9] or core region of HCV were also associated with response to therapy [10, 11]. Unfortunately, these host genetic and viral factors are not yet readily available for general application in actual clinical practice. Fibrosis of the liver is also an important predictor of response, but resources may be limited in some countries. Clinical and non-invasive parameters may be better suited for general practice, but there is no established means by which the likelihood of a response can be predicted prior to therapy.

Data mining is a method of predictive analysis that explores data, without setting the hypothesis, to discover hidden patterns and relationships in highly complex datasets and enables the development of predictive models. Decision tree analysis is a core component of data mining and predictive modeling [12], and it is utilized by decision makers in various fields of business. Recent publications on decision tree analysis indicate its usefulness for defining prognostic factors in various diseases such as prostate cancer [13], diabetes [14], melanoma [15, 16], colorectal carcinoma [17, 18], and liver failure [19]. The results of the analysis are presented as a tree structure, which is intuitive and facilitates the allocation of patients into subgroups by following the flow chart form [20]. We have recently reported the usefulness of decision tree analysis for the prediction of early virological response (undetectable HCVRNA within 12 weeks of therapy) to PEG-IFN and RBV combination therapy in chronic hepatitis C [21].

In the present study, we used decision tree analysis to explore baseline predictors of response to PEG-IFN/RBV therapy so that a pre-treatment algorithm could be created to discriminate chronic hepatitis C patients who are likely to respond to PEG-IFN/RBV therapy from those who are not. For the purpose of use in general practice, only clinical and non-invasive parameters were included in the analysis.

Materials and methods

Patients

This was a multicenter retrospective cohort study supported by the Japanese Ministry of Health, Labor and Welfare. Data were collected from a total of 800 chronic hepatitis C patients who received therapy for 48 weeks with PEG-IFN alpha-2b and RBV at Musashino Red Cross Hospital, Toranomon Hospital, Tokyo Medical and Dental University, Osaka University, Nagoya City University Graduate School of Medical Sciences, Yamanashi University, and their related hospitals. The inclusion criteria to be enrolled in this study were as follows (1) infection by genotype 1b, (2) HCVRNA higher than 100,000 IU/ml by quantitative PCR (Cobas Amplicor HCV Monitor v 2.0, Roche Diagnostic systems, CA), which is typically used for the definition of high viral load in Japan, (3) lack of co-infection with hepatitis B virus or human immunodeficiency virus, (4) lack of other causes of liver disease such as autoimmune hepatitis and primary biliary cirrhosis and (5) completion of at least 12 weeks of therapy. Patients received PEG-IFN alpha-2b (1.5 μg/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 1,000 mg for >80 kg), which is the recommended dosage in Japan. Patients who were treated for more than 49 weeks were not included in the study. For the analysis, patients were randomly assigned to either the model building ($n = 506$) or the internal validation ($n = 295$) group. Consent was obtained from each patient. The study protocol conformed to the ethical guidelines of the Declaration of Helsinki and was approved by the institutional review committee. The baseline characteristics and representative laboratory test results are listed in Table 1. The overall rate of SVR was 47% in the model building set and 49% in the validation set. There were no significant differences in the clinical backgrounds between these 2 groups.

For external validation of the model, we collaborated with another study group supported by the Japanese Ministry of Health, Labor and Welfare. This multicenter study group consisted of 29 medical centers and hospitals

Table 1 Comparison of pre-treatment factors between model building and internal validation patients

	Model (n = 506)	Validation (n = 295)
Age (years)	56 (14–75)	55 (18–74)
Male gender[a]	261/506 (52%)	160/295 (54%)
Body mass index (kg/m^2)	22.9 (14.3–34.0)	23.2 (16.1–33.8)
Albumin (g/dl)	4 (2.7–5.0)	4 (2.8–4.9)
Creatinine (mg/dl)	0.7 (0.4–1.5)	0.7 (0.4–1.1)
AST (IU/l)	60 (11–370)	62 (11–240)
ALT (IU/l)	73 (11–413)	73 (14–390)
GGT (IU/l)	56 (10–328)	55 (7–409)
Total cholesterol (mg/dl)	173 (73–297)	171 (29–273)
Triglyceride (mg/dl)	105 (33–474)	109 (32–372)
White blood cell count (/μl)	4,745 (1,800–10,900)	4,823 (1,200–9,700)
Neutrophil count (/μl)	2,563 (667–7,870)	2,484 (508–7,579)
Red blood cell count (/μl)	448 (313–577)	451 (313–574)
Hemoglobin (g/dl)	14.1 (9.4–18.3)	14.1 (10.0–18.0)
Hematocrit (%)	41.7 (13.3–53.7)	41.9 (15.5–52.7)
Platelets (10^9/l)	164 (52–380)	158 (43–312)
AFP (ng/ml)	14.7 (0.9–680)	13 (0.8–323)
HCVRNA (10^3 IU/ml)	1,852 (100–5,100)	1,870 (100–5,100)
Fibrosis stage: F3–4	73/417 (18%)	48/247 (19%)

Data expressed as median (range) unless otherwise indicated

AST aspartate aminotransferase, *ALT* alanine aminotransferase, *GGT* gamma-glutamyltransferase, *AFP* alpha-fetoprotein

[a] Data expressed as number/available data (percentage)

belonging to the National Hospital Organization. A dataset collected from 524 patients who were treated with PEG-IFN alpha-2b/RBV was used as an external validation dataset, i.e., completely independent from the dataset that was used for model building.

Laboratory tests

Blood samples were obtained before therapy and at least once every month during therapy, and were used for hematologic tests, blood chemistry analysis and determination of HCV RNA. Pretreatment levels of HCV RNA were quantified by Cobas Amplicor (Roche Diagnostic Systems, Pleasanton, CA). SVR was defined as undetectable HCV RNA at week 24 after completion of therapy, as determined by qualitative PCR with a lower end detection limit of 50 IU/ml (Amplicor, Roche Diagnostic Systems). Liver biopsy was available in 664 patients. Fibrosis and activity were scored according to the METAVIR scoring system [22]. 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).

Statistical analysis

A database of pretreatment variables was created containing 6 variables from hematological tests (red blood cells, hemoglobin, hematocrit, white blood cells, neutrocytes and platelets), 8 variables from the blood chemistry test [creatinine, albumin, aspartate aminotransferase, alanine aminotransferase, gamma-glutamyltransferase (GGT), total cholesterol, triglyceride and alpha-fetoprotein (AFP)], serum level of HCV RNA and 3 variables for patient characteristics (age, gender and body mass index). Based on this database, the recursive partitioning analysis algorithm referred to as decision tree analysis was implemented to define meaningful subgroups of patients with respect to the possibility of achieving SVR.

Decision tree analysis is a family of nonparametric regression methods. Software is used to automatically explore the data to search for optimal split variables and to build a decision tree structure [23]. For the analysis, the entire study population was evaluated to determine which variables and cutoff points yielded the most significant division into 2 prognostic subgroups that were as homogeneous as possible for the probability of SVR. Thereafter, the same analytic process was applied to all newly defined subgroups. A restriction was imposed on the tree construction such that the procedure stopped when either no additional significant variable was detected or when the sample size was below 20. For this analysis, the data mining software IBM SPSS Modeler 13 (IBM SPSS Inc., Chicago, IL) was utilized. SPSS software v.15.0 (SPSS Inc., Chicago, IL) was used for multivariate logistic regression analysis.

Results

Decision tree analysis

Decision tree analysis was carried out on the model building dataset from 506 patients using 18 variables. Figure 1 shows the results. The analysis automatically selected 5 predictive variables to produce a total of 7 subgroups of patients. Age was selected as the variable of initial split with an optimal cutoff of 50 years. The possibility of achieving SVR was 41% for patients older than 50 compared to 70% for patients

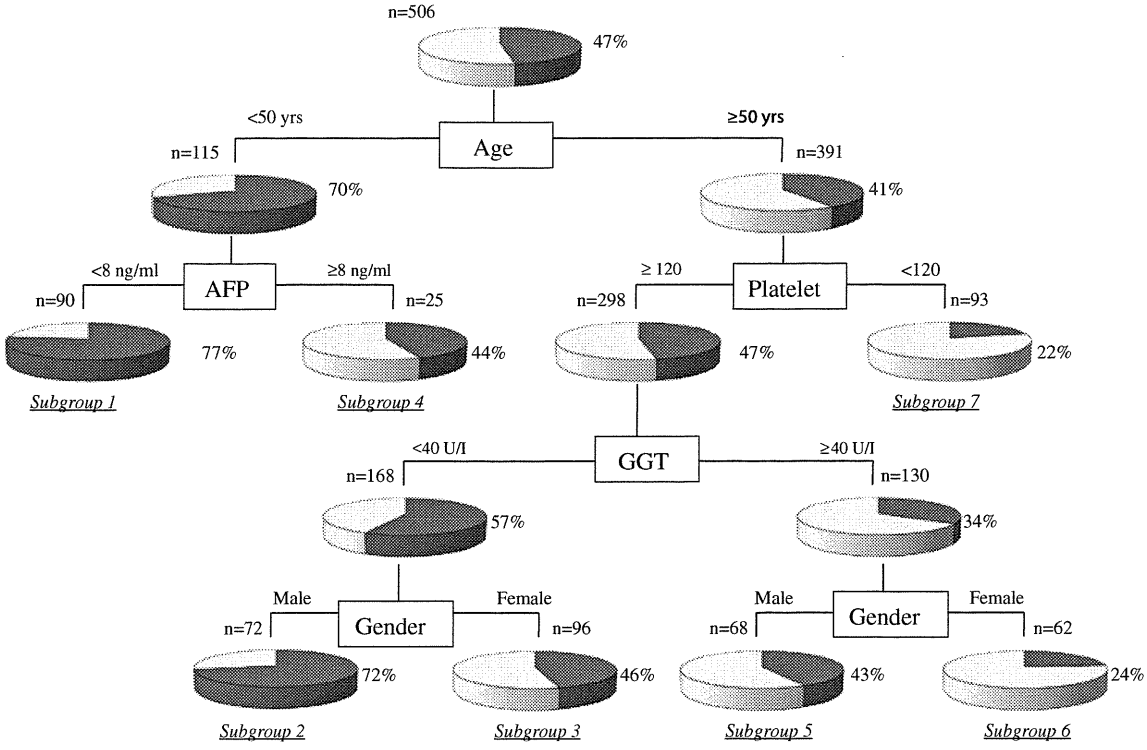


Fig. 1 Decision tree analysis. *Boxes* indicate the factors for splitting and the cutoff value for the split. *Pie charts* indicate the rate of SVR for each group. Terminal subgroups of patients discriminated by the analysis are numbered from 1 to 7. *AFP* alpha-fetoprotein, *GGT* gamma-glutamyltransferase

younger than 50. Among patients younger than 50, the level of serum AFP, with an optimal cutoff of 8 ng/ml, was selected as the variable of second split. Patients with lower AFP levels had a higher probability of SVR (77 vs. 44%). Among older patients, platelet count was selected as the second variable of split, with an optimal cutoff of 120×10^9/l. Patients with higher platelet counts had a higher probability of SVR (47 vs. 22%). Among patients with platelet counts higher than 120×10^9/l, GGT was selected as the third variable of split with an optimal cutoff of 40 IU/l. Patients with a lower GGT level had a higher probability of SVR (57 vs. 34%). Gender was selected as the fourth variable of split, with male gender being a predictor of a higher SVR probability (72 vs. 46% in patients with GGT levels <40 IU/l and 43 vs. 24% in those with GGT ≥40 IU/l). HCVRNA load was included in the analysis but was not selected as a significant variable.

The probabilities of SVR for the 7 subgroups derived by this process were highly variable. The subgroup of young patients (<50 years) with low serum AFP (<8 ng/ml) (subgroup 1) or the subgroup of older (≥50 years) male patients with high platelet counts ($\geq 120 \times 10^9$/l) and low serum GGT (<40 IU/l) (subgroup 2) showed the highest probability of SVR (72 and 77%), while the subgroup of older (≥50 years) patients with low platelet counts ($<120 \times 10^9$/l) (subgroup 7) and older (≥50 years) female patients with high serum GGT (subgroup 6) showed the lowest probability of SVR (22 and 24%).

Validation of the decision tree

The results of the decision tree analysis were validated with an internal validation dataset of 295 cases, which was independent of the model building dataset. Each patient in the validation set was allocated to subgroups 1–7 using the flow-chart form of the decision tree. The rates of SVR were 77% for subgroup 1, 71% for subgroup 2, 55% for subgroup 3, 44% for subgroup 4, 41% for subgroup 5, 17% for subgroup 6, and 30% for subgroup 7. The rates of SVR for each subgroup of patients were closely correlated between the model building dataset and the internal validation dataset ($r^2 = 0.925$) (Fig. 2a).

To further confirm the universality of the results, data collected from 524 patients by a collaborating study group were used for external validation. Thus, the dataset used for external validation was completely independent of the

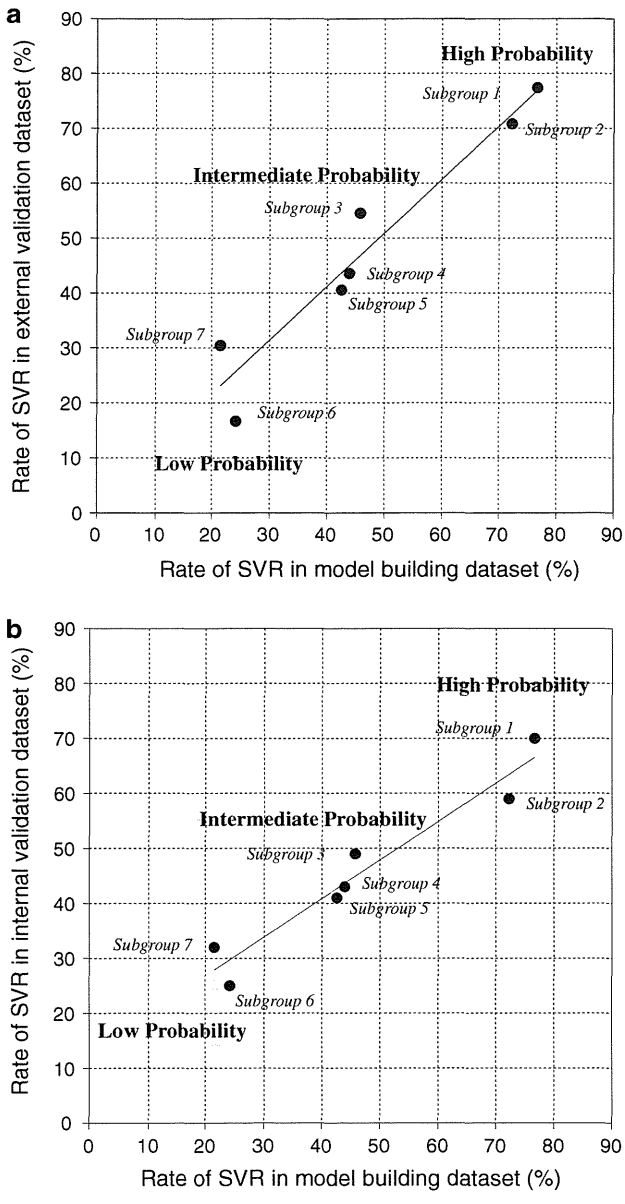


Fig. 2 Validation of the decision tree analysis by an internal and external validation dataset: subgroup-stratified comparison of the SVR rate. The rate of SVR in each subgroup was plotted. The *X* axis represents the model building, and the *Y* axis represents the validation datasets. **a** Internal validation and **b** external validation. There was a close correlation between the model building and the internal validation dataset (correlation coefficient $r^2 = 0.925$) and between the model building and the external validation dataset (correlation coefficient $r^2 = 0.936$)

original dataset used for model building. Each patient in the external validation set was allocated to subgroups 1–7 using the flow-chart form of the tree. The rates of SVR were 70% for subgroup 1, 59% for subgroup 2, 49% for subgroup 3, 43% for subgroup 4, 41% for subgroup 5, 25% for subgroup 6, and 32% for subgroup 7. The rates of SVR for each subgroup of patients were closely correlated

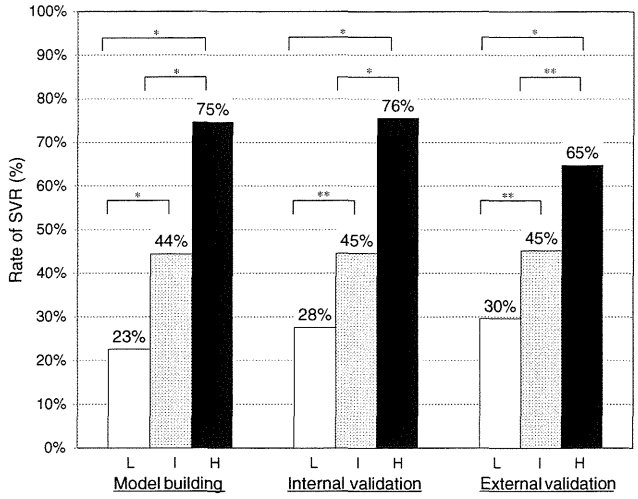


Fig. 3 Comparison of SVR rates between groups divided by the decision tree. The rate of SVR was compared among the 3 groups of patients divided by the decision tree analysis (*white*, *gray* and *black boxes*, indicating a low (*L*), intermediate (*I*) and high (*H*) probability group, respectively). The rate of SVR was significantly different among the 3 groups. $*p < 0.0001$, $**p < 0.001$

between the model-building dataset and the validation dataset ($r^2 = 0.936$) (Fig. 2b).

Construction of 3 groups according to the probability of SVR

Seven subgroups were reconstructed into 3 groups according to their predicted rates of SVR: the high probability group consisted of subgroups 1 and 2, the intermediate probability group consisted of subgroups 3, 4 and 5, and the low probability group consisted of subgroups 6 and 7. The rate of SVR was significantly different among the 3 groups (Fig. 3). The rate of SVR in the high probability group was consistently high: 75% for model building patients, 76% for internal validation patients and 65% for external validation patients. Conversely, the rate of SVR in the low probability group was consistently low: 23% for model building patients, 28% for internal validation patients and 30% for external validation patients. The rate of SVR in the intermediate probability group was 44% for model building patients, 45% for internal validation patients and 45% for external validation patients. Since 28–32% of patients were classified as high probability and 30–32% were classified as low probability, roughly 60% of patients were classified as having either a high or low probability of achieving SVR.

Effect of dose reductions of PEG-IFN and RBV on SVR

The cumulative dose of PEG-IFN and RBV was not included as a variable of analysis since the present study

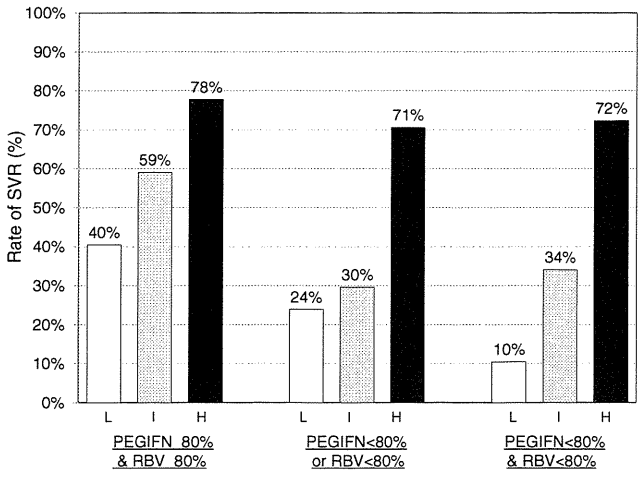


Fig. 4 Comparison of SVR rates among groups stratified by drug adherence. The 3 groups of patients divided by the decision tree analysis (*white*, *gray* and *black boxes* indicating a low (*L*), intermediate (*I*) and high (*H*) probability group, respectively) were further stratified according to the cumulative drug exposure of PEG-IFN and RBV. The good adherence group (≥80% planned dose of both PEG-IFN and RBV) had a higher rate of SVR compared with the poor adherence group (<80% planned dose of both PEG-IFN and RBV) in the low ($p = 0.0003$) and intermediate ($p = 0.007$) but not in the high probability group ($p = 0.53$)

aimed to develop a pre-treatment model for the prediction of response. To analyze the possible effect of drug reductions on the result of the decision tree analysis, 3 groups of patients divided by the decision tree analysis (low, intermediate and high probability group) were further stratified according to the cumulative drug exposure of PEG-IFN and RBV (Fig. 4). Even after adjustment for adherence, 3 groups of patients still had low, intermediate and high probability of achieving SVR, respectively. Of note, the good adherence group (≥80% planned dose of both PEG-IFN and RBV) had higher rates of SVR compared with the poor adherence group (<80% planned dose of both PEG-IFN and RBV) in the low ($p = 0.0003$) and intermediate ($p = 0.007$) probability group, but not in the high probability group ($p = 0.53$).

Factors associated with SVR by multivariate logistic regression analysis

We also explored the factors associated with SVR using a standard statistical analysis. By univariate analysis, age, gender, serum albumin, creatinine, alanine aminotransferase, GGT, red blood cell count, hemoglobin, hematocrit, platelet count and AFP were found to be associated with SVR (Table 2). HCVRNA load was not associated with SVR. By multivariate analysis, age, gender, GGT and platelet count were found to be independently associated with SVR (Table 3). Of note, AFP, which was selected as a significant predictor of response in the decision tree analysis, was not found to be an independent response predictor in the standard multivariate analysis. This indicates a unique feature of the decision tree analysis; i.e., it could identify significant predictors that specifically apply to selected patients, in this case patients younger than 50 years old.

Relationships between decision tree model and stage of fibrosis or HCV RNA load

Liver biopsy was performed in 664 patients. The distribution of fibrosis in three probability groups differed significantly. Advanced fibrosis (F3 or F4) was higher in the low probability group (39%) compared to the intermediate probability group (13%) ($p < 0.0001$) and to the high probability group (6%) ($p < 0.0001$). Advanced fibrosis was also higher in the intermediate group compared to the high probability group ($p = 0.01$). AFP was significantly associated with liver fibrosis stage: medians of AFP levels were 4.9, 5.9, 13.0 and 18.6 for F1, F2, F3 and F4, respectively ($p < 0.0001$, Spearman's rank correlations). Lower platelet counts correlated with advanced fibrosis stages (data not shown). The SVR rate was higher in the high probability group compared to the intermediate or low probability group after stratification by HCV RNA load. Among patients with low HCVRNA load (<400,000 IU/ml), the rate of SVR was 93, 59 and 50% for the high, intermediate and low probability group, respectively ($p = 0.002$ for high vs. intermediate and $p < 0.001$ for high vs. low probability groups). Among patients with a high HCVRNA load (≥400,000 IU/ml), the rate of SVR was 73, 42 and 21% for the high, intermediate and low probability group, respectively ($p < 0.001$ for high vs. low, high vs. intermediate and intermediate vs. low probability groups).

Discussion

Currently, the combination of PEG-IFN and RBV is the recommended therapy for chronic HCV infection. The rate of SVR with 48 weeks of therapy is around 50% in patients with HCV genotype 1b and a high HCV RNA titer [2, 3]. To date, the virological response during therapy is the most reliable means for predicting the likelihood of SVR [2, 24, 25]. More potent therapy, such as a triple combination of protease inhibitor, PEG-IFN and RBV, is being evaluated in clinical trials but is not readily available [26, 27]. Under the circumstances, pre-treatment prediction of the likelihood of SVR may be useful for both patients and physicians to support clinical decisions as to whether to start PEG-IFN/RBV therapy or delay treatment until a new more effective therapy becomes available.

Table 2 Comparison of pre-treatment factors between patients with and without sustained virological response (SVR) among the model building dataset ($n = 506$)

	SVR ($n = 240$)	Non-SVR ($n = 266$)	p
Age (years)	54 (25–75)	60 (36–73)	<0.0001
Male gender[a]	151/240 (63%)	171/266 (41%)	<0.0001
Body mass index (kg/m^2)	22.5 (16.8–32.0)	22.6 (15.5–33.3)	0.244
Albumin (g/dl)	4.1 (3.2–5.0)	4 (2.7–4.9)	0.004
Creatinine (mg/dl)	0.7 (0.44–1.14)	0.69 (0.39–1.47)	<0.0001
AST (IU/l)	59 (11–370)	61 (17–261)	0.457
ALT (IU/l)	58 (11–413)	53 (11–316)	0.031
GGT (IU/l)	31 (10–322)	43 (12–328)	0.005
Total cholesterol (mg/dl)	175 (87–297)	171 (73–274)	0.184
Triglyceride (mg/dl)	105 (36–474)	105 (33–294)	0.992
White blood cell count (/μl)	4,600 (2,200–10,900)	4,425 (1,800–10,810)	0.479
Neutrophils (/μl)	2,507 (667–7,870)	2,423 (900–7,281)	0.321
Red blood cell count (/μl)	455 (336–577)	441 (313–564)	0.001
Hemoglobin (g/dl)	14.3 (10.2–17.6)	13.9 (9.4–17.9)	0.004
Hematocrit (%)	42.1 (13.3–53.7)	41.2 (30.7–52.0)	0.031
Platelets (10^9/l)	178 (81–380)	142 (60–320)	<0.0001
AFP (ng/ml)	4.3 (0.9–680)	6.4 (1.9–468)	0.041
HCVRNA (10^3 IU/ml)	1,400 (100–5,100)	1,700 (100–5,100)	0.659
Fibrosis stage: F3–4[a]	21/198 (11%)	52/219 (24%)	<0.0001

Data expressed as median (range) unless otherwise indicated

AST aspartate aminotransferase, *ALT* alanine aminotransferase, *GGT* gamma-glutamyltransferase, *AFP* alpha-fetoprotein

[a] Data expressed as number/available data (percentage)

Table 3 Multivariate logistic regression analysis for factors associated with sustained virological response (SVR)

	Odds	95% CI	p value
Age (years)	0.96	0.94–0.98	0.001
Platelets (10^9/l)	1.09	1.04–1.14	<0.0001
ALT (IU/l)	1.01	1.00–1.01	0.001
GGT (IU/l)	0.99	0.98–0.99	<0.0001
Male gender	2.92	1.87–4.55	<0.0001

GGT gamma-glutamyltransferase

Using the data mining analysis, we constructed a simple decision tree model for the pre-treatment prediction of response to PEG-IFN/RBV. The analysis highlighted 5 variables relevant to response: age, gender, platelet count, AFP and GGT. Classification based on these variables identified subgroups of patients with high probabilities of achieving SVR among difficult to treat genotype 1b chronic hepatitis C patients. The reproducibility of the model was confirmed by the independent internal and external validation datasets. An advantage of the decision tree analysis over traditional regression models is that the decision tree model is user-intuitive and can be readily interpreted by medical professionals without any specific knowledge of statistics. Patients can be allocated to specific subgroups with a defined rate of response simply by following the flow-chart form. Using this model, an estimate of the response before treatment can be rapidly obtained, which may facilitate clinical decision making. Thus, this model could be readily applicable to clinical practice.

According to the results of the decision tree analysis, patients were categorized into 3 groups: the rate of SVR was 23–30% for the low probability group, 44–45% for the intermediate probability group and 65–76% for the high probability group. About 30% of patients were each categorized in the high and low probability group and the remaining 40% of patients in the intermediate probability group. These results support the evidence-based approach for selecting an optimum treatment strategy for individual patients. For example, patients in the high probability group may be the most suitable candidates for PEG-IFN/RBV therapy, while patients in the low probability group may be advised to wait for a future therapy, such as the combination of protease inhibitor, PEG-IFN and RBV. However, the estimation of low probability should not be used to preclude patients from therapy, and the final decision should be made on a case-by-case basis, taking into consideration the acceptance by the patient of a low likelihood of response and the potential risk of disease progression while waiting for a future therapy.

Another important finding was that poor adherence to drugs lowered the rate of SVR in the low and intermediate probability groups, which implies that effort should be made to maintain ≥80% of the planned dose of PEG-IFN and RBV in those patients. On the other hand, the rate of SVR was high irrespective of drug adherence in the high probability group. Whether shorter duration of therapy is sufficient in this group of patients should be confirmed in future study.

The variables used in the decision tree have been previously reported to associate with the efficacy of IFN therapy. Younger age and male gender are associated with a favorable response [28]. Lower platelet count is a hallmark of advanced fibrosis in chronic hepatitis C and is reported to be associated with poor response to IFN [29]. AFP is usually used for the screening or the diagnosis of hepatocellular carcinoma, but recent studies suggest an association between higher AFP levels and poor response to IFN therapy [30–33]. Previous report speculated that higher expression of AFP by hepatic progenitor cells may be associated with non-response to therapy [30]. Another report speculated that AFP levels predict poor response to therapy through the underlining link to advanced liver fibrosis [31]. Our data support the latter speculation since advanced fibrosis was associated with elevation of AFP levels. Fibrosis of the liver is an important predictor of response, but we did not include this factor in the decision tree analysis since liver biopsy may not always be available in general practice. As a result, two predictive factors that correlate with fibrosis stage (platelet counts and AFP) were selected in the model, and three probability groups reflected the different distribution of fibrosis stage. GGT is reported to be associated with insulin resistance and hepatic steatosis [34–37], a factor that confers resistance to IFN therapy [38–44]. What is unique to the present study is the visualization of response probability by combining these factors and its high reproducibility revealed by a high-quality validation of the model by internal and external validation datasets that were completely independent of the model building dataset. Since factors used in the model were clinical parameters that are readily available by the usual workup of patients, this model could be immediately applicable to clinical practice without imposing costs for additional examinations.

A potential limitation of this study is that data mining analysis has an intrinsic risk of showing relationships that fit to the original dataset but are not reproducible in different populations. Although internal and external validations showed that our model had high reproducibility, we recognize that further validation on a larger external validation cohort, especially in populations other than Japanese, may be necessary to further verify the reliability of our model.

In conclusion, we built a pre-treatment model for the prediction of virological response to PEG-IFN/RBV. Because this decision tree model was made up of simple variables, it can be easily applied to clinical practice. This model may have the potential to support decisions about patient selection for PEG-IFN/RBV based on a possibility of response weighed against the potential risk of adverse events or costs.

Acknowledgments This study was supported by a grant-in-aid from the Ministry of Health, Labor and Welfare, Japan.

References

1. Strader DB, Wright T, Thomas DL, Seeff LB. Diagnosis, management, and treatment of hepatitis C. Hepatology. 2004;39:1147–71.
2. Fried MW, Shiffman ML, Reddy KR, Smith C, Marinos G, Goncales FL Jr, et al. Peginterferon alfa-2a plus ribavirin for chronic hepatitis C virus infection. N Engl J Med. 2002;347:975–82.
3. Manns MP, McHutchison JG, Gordon SC, Rustgi VK, Shiffman M, Reindollar R, et al. Peginterferon alfa-2b plus ribavirin compared with interferon alfa-2b plus ribavirin for initial treatment of chronic hepatitis C: a randomised trial. Lancet. 2001;358:958–65.
4. Tanaka Y, Nishida N, Sugiyama M, Kurosaki M, Matsuura K, Sakamoto N, et al. Genome-wide association of IL28B with response to pegylated interferon-alpha and ribavirin therapy for chronic hepatitis C. Nat Genet. 2009;41:1105–9.
5. Suppiah V, Moldovan M, Ahlenstiel G, Berg T, Weltman M, Abate ML, et al. IL28B is associated with response to chronic hepatitis C interferon-alpha and ribavirin therapy. Nat Genet. 2009;41:1100–4.
6. Ge D, Fellay J, Thompson AJ, Simon JS, Shianna KV, Urban TJ, et al. Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. Nature. 2009;461:399–401.
7. Enomoto N, Sakuma I, Asahina Y, Kurosaki M, Murakami T, Yamamoto C, et al. Comparison of full-length sequences of interferon-sensitive and resistant hepatitis C virus 1b. Sensitivity to interferon is conferred by amino acid substitutions in the NS5A region. J Clin Investig. 1995;96:224–30.
8. Enomoto N, Sakuma I, Asahina Y, Kurosaki M, Murakami T, Yamamoto C, et al. Mutations in the nonstructural protein 5A gene and response to interferon in patients with chronic hepatitis C virus 1b infection. N Engl J Med. 1996;334:77–81.
9. Kurosaki M, Enomoto N, Murakami T, Sakuma I, Asahina Y, Yamamoto C, et al. Analysis of genotypes and amino acid residues 2209 to 2248 of the NS5A region of hepatitis C virus in relation to the response to interferon-beta therapy. Hepatology. 1997;25:750–3.
10. Akuta N, Suzuki F, Sezaki H, Suzuki Y, Hosaka T, Someya T, et al. Association of amino acid substitution pattern in core protein of hepatitis C virus genotype 1b high viral load and non-virological response to interferon-ribavirin combination therapy. Intervirology. 2005;48:372–80.
11. Okanoue T, Itoh Y, Hashimoto H, Yasui K, Minami M, Takehara T, et al. Predictive values of amino acid sequences of the core and NS5A regions in antiviral therapy for hepatitis C: a Japanese multi-center study. J Gastroenterol. 2009;44:952–63.
12. Breiman LJH, Friedman RA, Olshen CJ, Stone CM. Classification and regression trees. Calif: Wadsworth; 1980.
13. Garzotto M, Beer TM, Hudson RG, Peters L, Hsieh YC, Barrera E, et al. Improved detection of prostate cancer using classification and regression tree analysis. J Clin Oncol. 2005;23:4322–9.
14. Miyaki K, Takei I, Watanabe K, Nakashima H, Omae K. Novel statistical classification model of type 2 diabetes mellitus patients for tailor-made prevention using data mining algorithm. J Epidemiol. 2002;12:243–8.
15. Averbook BJ, Fu P, Rao JS, Mansour EG. A long-term analysis of 1018 patients with melanoma by classic Cox regression and tree-structured survival analysis at a major referral center:

implications on the future of cancer staging. Surgery. 2002;132: 589–602.
16. Leiter U, Buettner PG, Eigentler TK, Garbe C. Prognostic factors of thin cutaneous melanoma: an analysis of the central malignant melanoma registry of the german dermatological society. J Clin Oncol. 2004;22:3660–7.
17. Valera VA, Walter BA, Yokoyama N, Koyama Y, Iiai T, Okamoto H, et al. Prognostic groups in colorectal carcinoma patients based on tumor cell proliferation and classification and regression tree (CART) survival analysis. Ann Surg Oncol. 2007;14:34–40.
18. Zlobec I, Steele R, Nigam N, Compton CC. A predictive model of rectal tumor response to preoperative radiotherapy using classification and regression tree methods. Clin Cancer Res. 2005;11:5440–3.
19. Baquerizo A, Anselmo D, Shackleton C, Chen TW, Cao C, Weaver M, et al. Phosphorus ans an early predictive factor in patients with acute liver failure. Transplantation. 2003;75: 2007–14.
20. LeBlanc M, Crowley J. A review of tree-based prognostic models. Cancer Treat Res. 1995;75:113–24.
21. Kurosaki M, Matsunaga K, Hirayama I, Tanaka T, Sato M, Yasui Y, et al. A predictive model of response to peginterferon ribavirin in chronic hepatitis C using classification and regression tree analysis. Hepatol Res. 2010;40:251–60.
22. Bedossa P, Poynard T. An algorithm for the grading of activity in chronic hepatitis C. The METAVIR cooperative study group. Hepatology. 1996;24:289–93.
23. Segal MR, Bloch DA. A comparison of estimated proportional hazards models and regression trees. Stat Med. 1989;8:539–50.
24. Davis GL, Wong JB, McHutchison JG, Manns MP, Harvey J, Albrecht J. Early virologic response to treatment with peginterferon alfa-2b plus ribavirin in patients with chronic hepatitis C. Hepatology. 2003;38:645–52.
25. Lee SS, Ferenci P. Optimizing outcomes in patients with hepatitis C virus genotype 1 or 4. Antivir Ther. 2008;13(Suppl 1):9–16.
26. Hezode C, Forestier N, Dusheiko G, Ferenci P, Pol S, Goeser T, et al. Telaprevir and peginterferon with or without ribavirin for chronic HCV infection. N Engl J Med. 2009;360:1839–50.
27. McHutchison JG, Everson GT, Gordon SC, Jacobson IM, Sulkowski M, Kauffman R, et al. Telaprevir with peginterferon and ribavirin for chronic HCV genotype 1 infection. N Engl J Med. 2009;360:1827–38.
28. Sezaki H, Suzuki F, Kawamura Y, Yatsuji H, Hosaka T, Akuta N, et al. Poor response to pegylated interferon and ribavirin in older women infected with hepatitis C virus of genotype 1b in high viral loads. Dig Dis Sci. 2009;54:1317–24.
29. Shiratori Y, Omata M. Predictors of the efficacy of interferon therapy for patients with chronic hepatitis C before and during therapy: how does this modify the treatment course? J Gastroenterol Hepatol. 2000;15(Suppl):E141–51.
30. Gad RR, Males S, El Makhzangy H, Shouman S, Hasan A, Attala M, et al. Predictors of a sustained virological response in patients with genotype 4 chronic hepatitis C. Liver Int. 2008;28: 1112–9.
31. Akuta N, Suzuki F, Kawamura Y, Yatsuji H, Sezaki H, Suzuki Y, et al. Predictors of viral kinetics to peginterferon plus ribavirin combination therapy in Japanese patients infected with hepatitis C virus genotype 1b. J Med Virol. 2007;79:1686–95.
32. Males S, Gad RR, Esmat G, Abobakr H, Anwar M, Rekacewicz C, et al. Serum alpha-foetoprotein level predicts treatment outcome in chronic hepatitis C. Antivir Ther. 2007;12:797–803.
33. Bayati N, Silverman AL, Gordon SC. Serum alpha-fetoprotein levels and liver histology in patients with chronic hepatitis C. Am J Gastroenterol. 1998;93:2452–6.
34. Fraser A, Ebrahim S, Smith GD, Lawlor DA. A comparison of associations of alanine aminotransferase and gamma-glutamyltransferase with fasting glucose, fasting insulin, and glycated hemoglobin in women with and without diabetes. Hepatology. 2007;46:158–65.
35. Marchesini G, Avagnina S, Barantani EG, Ciccarone AM, Corica F, Dall'Aglio E, et al. Aminotransferase and gamma-glutamyltranspeptidase levels in obesity are associated with insulin resistance and the metabolic syndrome. J Endocrinol Investig. 2005;28:333–9.
36. Soresi M, Tripi S, Franco V, Giannitrapani L, Alessandri A, Rappa F, et al. Impact of liver steatosis on the antiviral response in the hepatitis C virus-associated chronic hepatitis. Liver Int. 2006;26:1119–25.
37. Yaginuma R, Ikejima K, Okumura K, Kon K, Suzuki S, Takei Y, et al. Hepatic steatosis is a predictor of poor response to interferon alpha-2b and ribavirin combination therapy in Japanese patients with chronic hepatitis C. Hepatol Res. 2006;35:19–25.
38. Adinolfi LE, Gambardella M, Andreana A, Tripodi MF, Utili R, Ruggiero G. Steatosis accelerates the progression of liver damage of chronic hepatitis C patients and correlates with specific HCV genotype and visceral obesity. Hepatology. 2001;33:1358–64.
39. Akuta N, Suzuki F, Kawamura Y, Yatsuji H, Sezaki H, Suzuki Y, et al. Predictive factors of early and sustained responses to peginterferon plus ribavirin combination therapy in Japanese patients infected with hepatitis C virus genotype 1b: amino acid substitutions in the core region and low-density lipoprotein cholesterol levels. J Hepatol. 2007;46:403–10.
40. Berg T, Sarrazin C, Herrmann E, Hinrichsen H, Gerlach T, Zachoval R, et al. Prediction of treatment outcome in patients with chronic hepatitis C: significance of baseline parameters and viral dynamics during therapy. Hepatology. 2003;37:600–9.
41. Mazzella G, Salzetta A, Casanova S, Morelli MC, Villanova N, Miniero R, et al. Treatment of chronic sporadic-type non-A, non-B hepatitis with lymphoblastoid interferon: gamma GT levels predictive for response. Dig Dis Sci. 1994;39:866–70.
42. Thomopoulos KC, Theocharis GJ, Tsamantas AC, Siagris D, Dimitropoulou D, Gogos CA, et al. Liver steatosis is an independent risk factor for treatment failure in patients with chronic hepatitis C. Eur J Gastroenterol Hepatol. 2005;17:149–53.
43. Villela-Nogueira CA, Perez RM, de Segadas Soares JA, Coelho HS. Gamma-glutamyl transferase (GGT) as an independent predictive factor of sustained virologic response in patients with hepatitis C treated with interferon-alpha and ribavirin. J Clin Gastroenterol. 2005;39:728–30.
44. Camps J, Crisostomo S, Garcia-Granero M, Riezu-Boj JI, Civeira MP, Prieto J. Prediction of the response of chronic hepatitis C to interferon alfa: a statistical analysis of pretreatment variables. Gut. 1993;34:1714–7.

Genome-Wide Association Study Confirming Association of HLA-DP with Protection against Chronic Hepatitis B and Viral Clearance in Japanese and Korean

Nao Nishida[1,2]*, Hiromi Sawai[2], Kentaro Matsuura[3], Masaya Sugiyama[1], Sang Hoon Ahn[4], Jun Yong Park[4], Shuhei Hige[5], Jong-Hon Kang[6], Kazuyuki Suzuki[7], Masayuki Kurosaki[8], Yasuhiro Asahina[8], Satoshi Mochida[9], Masaaki Watanabe[10], Eiji Tanaka[11], Masao Honda[12], Shuichi Kaneko[12], Etsuro Orito[13], Yoshito Itoh[14], Eiji Mita[15], Akihiro Tamori[16], Yoshikazu Murawaki[17], Yoichi Hiasa[18], Isao Sakaida[19], Masaaki Korenaga[20], Keisuke Hino[20], Tatsuya Ide[21], Minae Kawashima[2], Yoriko Mawatari[1,2], Megumi Sageshima[2], Yuko Ogasawara[2], Asako Koike[22], Namiki Izumi[8], Kwang-Hyub Han[4], Yasuhito Tanaka[3], Katsushi Tokunaga[2], Masashi Mizokami[1]

1 Research Center for Hepatitis and Immunology, National Center for Global Health and Medicine, Ichikawa, Chiba, Japan, **2** Department of Human Genetics, The University of Tokyo, Bunkyo-ku, Tokyo, Japan, **3** Department of Virology and Liver Unit, Nagoya City University Graduate School of Medical Sciences, Nagoya, Aichi, Japan, **4** Department of Internal Medicine, Yonsei University College of Medicine, Seoul, South Korea, **5** Department of Internal Medicine, Hokkaido University Graduate School of Medicine, Sapporo, Japan, **6** Department of Internal Medicine, Teine Keijinkai Hospital, Sapporo, Japan, **7** Department of Gastroenterology and Hepatology, Iwate Medical University, Morioka, Japan, **8** Division of Gastroenterology and Hepatology, Musashino Red Cross Hospital, Tokyo, Japan, **9** Division of Gastroenterology and Hepatology, Saitama Medical University, Saitama, Japan, **10** Department of Gastroenterology, Kitasato University School of Medicine, Sagamihara, Kanagawa, Japan, **11** Department of Medicine, Shinshu University School of Medicine, Matsumoto, Japan, **12** Department of Gastroenterology, Kanazawa University Graduate School of Medicine, Kanazawa, Japan, **13** Department of Gastroenterology, Nagoya Daini Red Cross Hospital, Nagoya, Japan, **14** Molecular Gastroenterology and Hepatology, Kyoto Prefectural University of Medicine, Kyoto, Japan, **15** Department of Gastroenterology and Hepatology, National Hospital Organization Osaka National Hospital, Osaka, Japan, **16** Department of Hepatology, Osaka City University Graduate School of Medicine, Osaka, Japan, **17** Second Department of Internal Medicine, Faculty of Medicine, Tottori University, Yonago, Japan, **18** Department of Gastroenterology and Metabology, Ehime University Graduate School of Medicine, Ehime, Japan, **19** Gastroenterology and Hepatology, Yamaguchi University Graduate School of Medicine, Yamaguchi, Japan, **20** Division of Hepatology and Pancreatology, Kawasaki Medical College, Kurashiki, Japan, **21** Division of Gastroenterology, Department of Medicine, Kurume University School of Medicine, Fukuoka, Japan, **22** Central Research Laboratory, Hitachi Ltd., Kokubunji, Tokyo, Japan

Abstract

Hepatitis B virus (HBV) infection can lead to serious liver diseases, including liver cirrhosis (LC) and hepatocellular carcinoma (HCC); however, about 85–90% of infected individuals become inactive carriers with sustained biochemical remission and very low risk of LC or HCC. To identify host genetic factors contributing to HBV clearance, we conducted genome-wide association studies (GWAS) and replication analysis using samples from HBV carriers and spontaneously HBV-resolved Japanese and Korean individuals. Association analysis in the Japanese and Korean data identified the *HLA-DPA1* and *HLA-DPB1* genes with $P_{meta} = 1.89 \times 10^{-12}$ for rs3077 and $P_{meta} = 9.69 \times 10^{-10}$ for rs9277542. We also found that the *HLA-DPA1* and *HLA-DPB1* genes were significantly associated with protective effects against chronic hepatitis B (CHB) in Japanese, Korean and other Asian populations, including Chinese and Thai individuals ($P_{meta} = 4.40 \times 10^{-19}$ for rs3077 and $P_{meta} = 1.28 \times 10^{-15}$ for rs9277542). These results suggest that the associations between the *HLA-DP* locus and the protective effects against persistent HBV infection and with clearance of HBV were replicated widely in East Asian populations; however, there are no reports of GWAS in Caucasian or African populations. Based on the GWAS in this study, there were no significant SNPs associated with HCC development. To clarify the pathogenesis of CHB and the mechanisms of HBV clearance, further studies are necessary, including functional analyses of the HLA-DP molecule.

Citation: Nishida N, Sawai H, Matsuura K, Sugiyama M, Ahn SH, et al. (2012) Genome-Wide Association Study Confirming Association of HLA-DP with Protection against Chronic Hepatitis B and Viral Clearance in Japanese and Korean. PLoS ONE 7(6): e39175. doi:10.1371/journal.pone.0039175

Editor: Anand S. Mehta, Drexel University College of Medicine, United States of America

Received February 1, 2012; **Accepted** May 16, 2012; **Published** June 21, 2012

Funding: This work was supported by Grants-in-Aid from the Ministry of Health, Labour, and Welfare of Japan (H22-kanen-005, H23-kanen-005), the Japan Science and Technology Agency (09038024), and the Miyakawa Memorial Research Foundation. Partial support by Grant-in-Aid for Young Scientists (B) (22710191) from the Ministry of Education, Culture, Sports, Science, and Technology is also acknowledged. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: AK is an employee of the Central Research Laboratory, Hitachi Ltd. There are no patents, products in development or marketed products to declare. This does not alter the authors' adherence to all the PLoS ONE policies on sharing data and materials, as detailed online in the guide for authors.

* E-mail: nishida-75@umin.ac.jp

Introduction

Overall, one-third of the world's population (2.2 billion) is infected with hepatitis B virus (HBV), and about 15% of these are chronic carriers. About 75% of the chronic carriers live in the east-south Asia and east pacific area, and there are 1.3–1.5 million chronic carriers living in Japan [1]. Of chronic carriers, 10–15% develop liver cirrhosis (LC), liver failure and hepatocellular carcinoma (HCC), and the remaining individuals eventually achieve a state of nonreplicative infection, resulting in hepatitis B surface antigen (HBsAg) negative and hepatitis B core antibody (anti-HBc) positive, i.e. HBV-resolved individuals [2–3]. In Japan, although the major route of HBV transmission was perinatal transmission and horizontal transmission in early childhood, infant HBV carriers have successfully been reduced since 1986 through a selective vaccination policy by the Japanese government [4–7]. However, the prevalence of HBV genotype A in acute HBV (AHB) infection has increased markedly since 2000, reaching approximately 52% in 2008 due to the lack of a universal HB vaccination, and around 10% of AHB cases could be persistent infection [8–9]. Viral factors, as well as host factors, are thought to be associated with persistent HB infection.

In 2009, significant associations between chronic hepatitis B (CHB) and a region including *HLA-DPA1* and *HLA-DPB1* were identified using 786 Japanese individuals having CHB and 2,201 control individuals through a two-stage genome-wide association study (GWAS) [10]. The same group was also subjected to a second GWAS using a total of 2,667 Japanese persistent HBV infection cases and 6,496 controls, which confirmed significant associations between the *HLA-DP* locus and CHB, in addition to associations with another two SNPs located in the genetic region including the *HLA-DQ* gene [11]. The associations between *HLA-DP* variants with HBV infection were replicated in other Asian populations, including Thai and Han Chinese individuals [10,12–13]. With regard to HBV clearance, the association between the human leukocyte antigen (HLA) class II allele and clearance of HBV was confirmed by the candidate gene approach in African, Caucasian and Asian populations [14–18]. However, in a previous GWAS using samples of Japanese CHB and control individuals, the clinical data on HBV exposure in the control individuals were unknown, and this may have led to bias. Moreover, there have been no reports of GWAS using samples from HBV carriers and HBV-resolved individuals to identify host genetic factors associated with HBV clearance other than HLA class II molecules.

Here, we performed a GWAS using samples from Japanese HBV carriers, healthy controls and spontaneously HBV-resolved individuals in order to confirm or identify the host genetic factors related to CHB and viral clearance. In the subsequent replication analysis, we validated the associated SNPs in the GWAS using two independent sets of Japanese and Korean individuals. In our study, healthy controls were randomly selected with clinically no evidence of HBV exposure, therefore, HBV-resolved individuals were prepared to clearly identify the host genetic factors related with CHB or HBV clearance.

Results

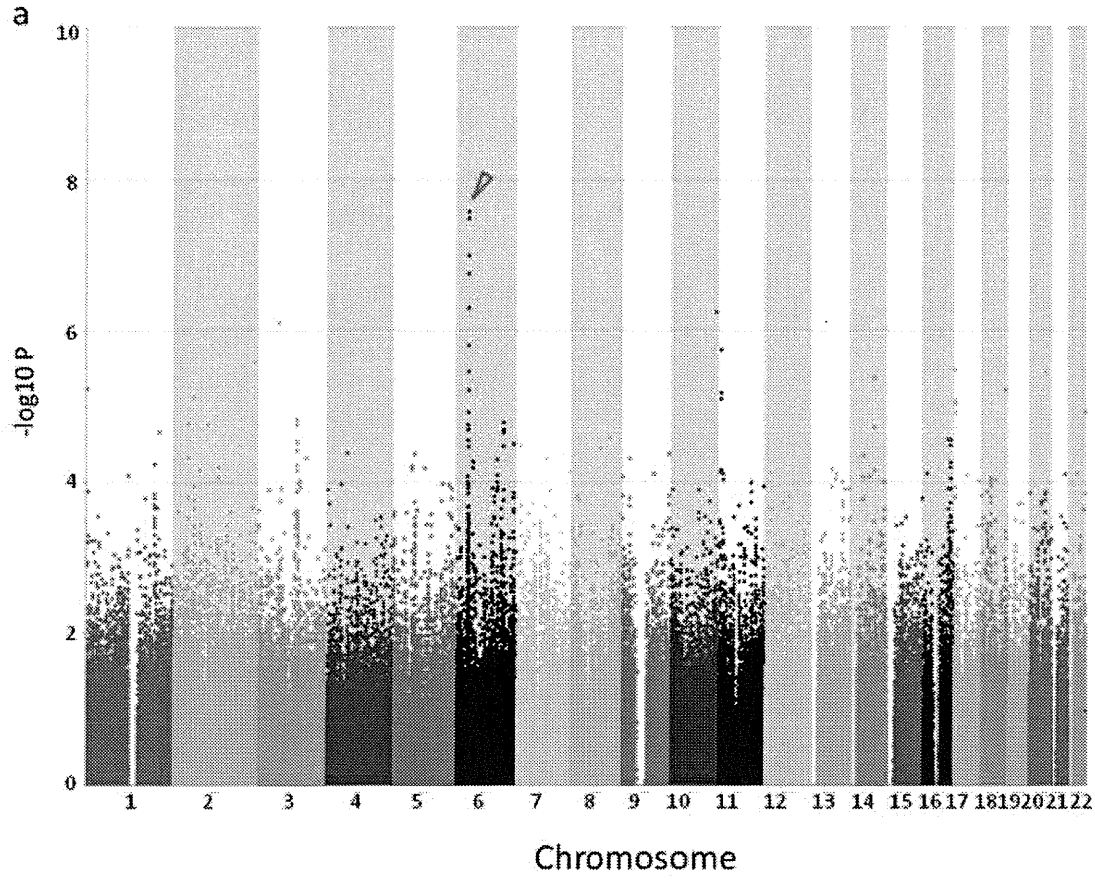
Protective Effects Against Chronic Hepatitis B in Japanese and Korean Individuals

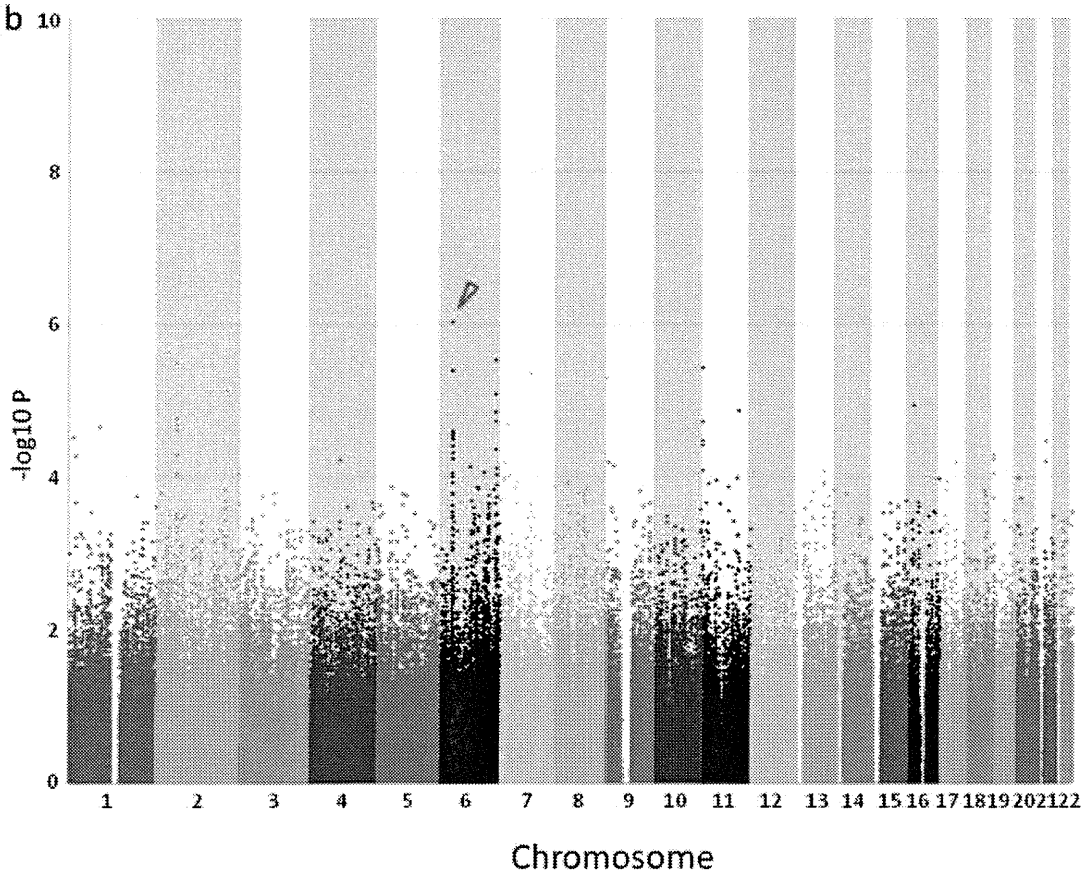
In this study, we conducted a GWAS using samples from 181 Japanese HBV carriers (including asymptomatic carriers (ASC), CHB cases, LC cases and HCC cases, based on the criteria described in Materials and Methods) and 184 healthy controls in order to identify the host genetic factors related to progression of CHB. All samples were genotyped using a genome-wide SNP typing array (Affymetrix Genome-Wide Human SNP Array 6.0 for 900 K SNPs). Figure 1a shows a genome-wide view of the single point association data based on allele frequencies using the SNPs that met the following filtering criteria: (i) SNP call rate ≥95%; (ii) minor allele frequency (MAF) ≥1% for HBV carriers and healthy controls; and (iii) no deviation from Hardy-Weinberg equilibrium (HWE) $P \geq 0.001$ in healthy controls. We identified significant associations of protective effects against CHB with two SNPs (rs3077 and rs9277542) using the allele frequency model, both of which are located in the 3′ UTR of *HLA-DPA1* and in the sixth exon of *HLA-DPB1*, respectively (rs3077, $P = 1.14 \times 10^{-7}$, and rs9277542, $P = 5.32 \times 10^{-8}$, respectively). The association for rs9277542 reached a genome-wide level of significance in the GWAS panel (Bonferroni criterion $P < 8.36 \times 10^{-8}$ (0.05/597,789)).

In order to validate the results of GWAS, a total of 32 SNPs, including the associated two SNPs (rs3077 and rs9277542), were selected for replication in two independent sets of HBV carriers and healthy controls (replication-1:256 Japanese HBV carriers and 236 Japanese healthy controls; and replication-2:344 Korean HBV carriers and 151 Korean healthy controls; Table 1). The associations for the original significant SNP (rs9277542) and marginal SNP (rs3077) on GWAS were replicated in both replication sets [replication-1 (Japanese); rs3077, $P = 2.70 \times 10^{-8}$, OR = 0.48 and rs9277542, $P = 3.33 \times 10^{-6}$, OR = 0.54; replication-2 (Korean); rs3077, $P = 2.08 \times 10^{-6}$, OR = 0.47 and rs9277542, $P = 8.29 \times 10^{-5}$, OR = 0.54, Table 2]. We conducted meta-analysis to combine these studies using the DerSimonian Laird method (random effects model) to incorporate variation among studies. As shown in Table 2, the odds ratios were quite similar across the three studies (GWAS and two replication studies) and no heterogeneity was observed ($P_{het} = 0.80$ for rs3077 and 0.40 for rs9277542). P_{meta} values were 4.40×10^{-19} for rs3077 (OR = 0.46, 95% confidence interval (CI) = 0.39–0.54), and 1.28×10^{-15} for rs9277542 (OR = 0.50, 95% CI = 0.43–0.60). Among the remaining 30 SNPs in the replication study, 27 SNPs were successfully genotyped by the DigiTag2 assay with SNP call rate ≥ 95% and HWE p-value ≥ 0.01. Two SNPs (rs9276431 and rs7768538), located in the genetic region including the *HLA-DQ* gene, were marginally replicated in the two sets of HBV carriers and healthy controls with Mantel-Haenszel P values of 2.80×10^{-7} (OR = 0.56, 95% CI = 0.45–0.70) and 1.09×10^{-7} (OR = 0.53, 95% CI = 0.42–0.67), respectively, when using additive, two-tailed Cochran Mantel-Haenszel (CMH) fixed-effects model with no evidence of heterogeneity ($P_{het} = 0.67$ for rs9276431 and 0.70 for rs7768538) (Table S1).

Meta-analysis using the random effects model across 6 independent studies, including 5 additional published data, showed $P_{meta} = 3.94 \times 10^{-45}$, OR = 0.55 for rs3077, $P_{meta} = 1.74 \times 10^{-21}$, OR = 0.61 for rs9277535 and $P_{meta} = 1.69 \times 10^{-15}$, OR = 0.51 for rs9277542, with the SNP rs9277535 being located about 4-kb upstream from rs9277542 and showing strong linkage disequilibrium of $r^2 = 0.955$ on the HapMap JPT (Table S2). As shown in Table S2, the odds ratio was very similar among the 6 studies, and heterogeneity was negligible with $P_{het} > 0.01$.

Moreover, based on GWAS using samples from 94 chronic HBV carriers with LC or HCC and 87 chronic HBV carriers without LC and HCC, we found no significant SNPs associated with CHB progression (Figure S1).


a
-log10 P
0
2
4
6
8
10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome


b
-log10 P
0
2
4
6
8
10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosome