shows simple steatosis with a good prognosis, but approximately 10–30% of NAFLD histologically diagnosed as nonalcoholic steatohepatitis (NASH) shows hepatocyte degeneration (ballooning hepatocyte), necrosis, inflammation and fibrosis, with a higher frequency of liver-related death both in Japanese and European populations [6,7]. Insulin resistance and oxidative stress are considered to be key players in the progression of NASH [8,9]. However, the progression of NASH has been considered to be complex involving multiple genetic factors interacting with the environment and lifestyle, because only a portion of NAFLD patients develops NASH.

The first Genome-wide association (GWA) study searching for such genetic factors identified the *PNPLA3* gene as a major genetic determinant for the predisposition to NAFLD in Hispanic, African American and European American populations according to liver fat contents [10], which was subsequently confirmed in Europeans and Asians according to liver biopsy. Association of *PNPLA3* with not only fatty liver and TG content, but also inflammation and fibrosis were shown in the subsequent studies, so *PNPLA3* may be widely associated with the development of NAFLD [11–13]. More recently, another GWA study reported the association of four additional genes with NAFLD in Europeans [14]. Also, a candidate gene-based approach revealed the association between NAFLD and the apolipoprotein C3 gene in Indians [15]. However, the precise role of such genes in the development of NASH still remains to be elucidated. In addition, no GWA study has been reported for Asian populations to date although the genetic components and their relative contribution may be different between ethnicities.

The Japan NASH Study Group was founded in 2008 aiming at the identification of genetic determinants predisposing to NASH in the Japanese population. Here we report the first GWA study of NAFLD in the Japanese using DNA samples of patients with liver histology-based diagnoses recruited through this multi-institutional research network.

Results

Genome-wide Association Analysis of NAFLD in Japanese

We conducted a GWA study using DNA samples of 543 patients with NAFLD and 942 controls. After quality controls of genotyping results (see materials and methods for details), a total of 529 patients consisting of four NAFLD subgroups according to Matteoni's classification [2] (type1; 100, type2; 73, type3; 29, type4; 327) and 932 controls were subjected to statistical analyses (Table 1). This index pathologically classifies NAFLD according to the degree of inflammation, hepatocyte degeneration, and the existence of fibrosis and Mallory-Denk body in the liver. Genome scan results of 932 DNA samples collected for other genetic studies were used as general Japanese population controls [16]. After standard quality control procedure as described in materials and methods, genotype distributions of 484,751 autosomal SNP markers were compared between the NAFLD cases and control subjects by exact trend test. A slight inflation of p-values was observed by genomic control method ($\lambda = 1.04$) (Figure S1).

We identified six SNP markers located at chromosome 22q13 showing genome-wide significance ($p<1.04\times10^{-7}$) (Figure 1). Among them, four SNPs, namely, rs2896019, rs926633, rs2076211 and rs1010023, located in the *PNPLA3* gene and in strong linkage disequilibrium (LD) ($r^2>0.93$), returned p-values smaller than 1×10^{-9} ($p=1.5\times10^{-10}$, 7.5×10^{-10}, 1.4×10^{-9} and 1.5×10^{-9}, respectively) (Table 2). Rs738407 and rs3810662 also located in *PNPLA3* showed significant but weaker associations ($p=1.0\times10^{-7}$ and 1.0×10^{-7}, respectively) than the above four SNP markers. Rs738491, rs2073082, rs3761472, rs2235776, rs2143571 and rs6006473 were in the neighboring *SAMM50* gene which is outside of the linkage disequilibrium (LD) block where the top SNP markers were distributed (Figure 2). These markers were in moderate LD with each other ($r^2>0.42$) and showed p-values between 3.9×10^{-6} and 6.4×10^{-7} but did not reach genome-wide significance (Table S1). Rs738409, the SNP which showed the strongest association with NAFLD in the first GWA study [10], was not included in the SNP array used in our study. This SNP was therefore genotyped using Taqman technology in the same case and control samples that were used for genome scan. Rs738409 showed the strongest association with the disease ($p=1.4\times10^{-10}$, OR = 1.66, 95%CI: 1.43–1.94) among all the SNP markers examined in this study. The association remained after the correction for population stratification with EIGENSTRAT [17] ($p=2.3\times10^{-11}$). Although a peak consisting of a cluster of SNPs was observed at the *HLA* locus on chromosome 6 (minimal p-value of 4.10×10^{-7} for rs9262639 located at the 3′ of *C6orf15* gene), the association disappeared when EIGENSTRAT was applied ($p>1.6\times10^{-3}$). We consider this as a result of population stratification between the cases and controls.

Impact of *PNPLA3* Polymorphisms to the Pathogenicity of NAFLD

We next examined whether or not the seven SNPs in the *PNPLA3* gene were associated with the pathogenic status of NAFLD. The genotype distributions of these SNPs were compared by Jonckheere-Terpstra test among the four subgroups of NAFLD patients categorized by Matteoni's classification (type1 to type4). There was a significant increase in the frequency of the risk allele from Matteoni type1 to type4 for all of the seven SNPs (p-values ranging from 3.6×10^{-6} to 0.0017) (Table 2). Among them, rs738409 again showed the strongest association ($p=3.6\times10^{-6}$) as seen in the simple case/control analysis. On the other hand, there was no significant association between control and Matteoni type1 ($p=0.76$).

In order to clarify how rs738409 influences the pathogenicity of NAFLD, we performed pairwise comparisons of genotype distributions in the four subgroups of NAFLD patients. There were marked differences in genotype distributions between type4 subgroup and the other three subgroups by multivariable logistic regression adjusted for age, sex and body mass index (BMI) ($p=2.0\times10^{-5}$, OR = 2.18, 95%CI: 1.52–3.18 between type1 and type4; $p=1.4\times10^{-3}$, OR = 1.81, 95%CI: 1.26–2.62 between type2 and type4; $p=0.027$, OR = 1.85, 95%CI: 1.07–3.19 between type3 and type4) (Figure 3). On the other hand, no significant associations were obtained for type1 to type3 in any combinations. When we performed the same analysis between type4 and the pooled genotypes of type1 to type3, we again obtained a significant difference ($p=4.8\times10^{-6}$, OR = 1.96, 95%CI: 1.47–2.62).

We further examined the specific association of rs738409 with type4 subgroup by using the case/control association results of the initial genome scan. 529 NAFLD patients were divided into 202 patients with type1 to type3 and 327 patients with type4, and genotype distributions of rs738409 in each subgroup were compared with those of 932 control subjects. Exact trend test returned an extremely strong association of rs738409 with type4 subgroup ($p=1.7\times10^{-16}$, OR = 2.18, 95%CI: 1.81–2.63) whereas no association was obtained for type1 to type3 subgroups ($p=0.41$).

Table 1. Clinical characteristics according to the histological classification.

Phenotype	Matteoni classification of NAFLD				Control	*p*-value
	Type 1	Type 2	Type 3	Type 4		
Number of samples	100	73	29	327	932	
Sex (Male/Female)	59/41	47/26	13/16	130/197	471/461	0.0023‡
Age (year)	49.7±15.3	51.5±15.3	49.4±14.0	57.6±14.8	48.8±16.3	<0.001
Physical measurement						
BMI	26.2±4.3	27.7±4.8	27.6±3.5	27.7±5.2	–	0.054
Amount of visceral fat (cm^2)	146.8±65.3	154.3±47.7	136.8±53.8	151.7±57.4	–	0.46
Abdominal circumscript (cm)	90.9±9.9	94.1±10.0	88.5±10.2	94.1±11.8	–	0.10
Biochemical trait						
AST (IU/L)	31.1±14.6	36.4±18.5	52.4±35.1	57.7±48.4	–	<0.001
ALT (IU/L)	48.6±30.8	62.8±47.6	81.5±46.9	74.9±48.4	–	<0.001
GGT (IU/L)	71.0±62.5	67.1±66.9	96.1±91.3	76.6±73.9	–	0.25
Albumin (g/dL)	4.5±0.4	4.4±0.3	4.5±0.3	4.3±0.4	–	<0.001
Total bilirubin (mg/dL)	0.9±0.5	0.9±0.5	0.9±0.6	0.8±0.4	–	0.063
Cholinesterase (unit)	389.1±97.0	354.3±97.2	371.1±109.9	348.9±93.2	–	<0.001
Type IV collagen 7S (ng/dL)	3.8±0.7	3.9±0.9	3.9±0.8	5.1±1.7	–	<0.001
Hyaluronic acid (ng/dL)	25.6±22.5	33.6±29.5	31.5±24.0	80.9±84.3	–	<0.001
Triglycerides (mg/dL)	151.9±73.8	154.0±92.1	166.1±86.5	161.2±85.7	–	0.23
Total cholesterol (mg/dL)	209.1±32.8	194.0±38.0	203.0±39.9	200.3±39.0	–	0.093
HbA1c (%)	6.1±1.1	5.9±1.2	6.5±1.8	6.2±1.3	–	0.13
IRI (μg/dL)	9.1±5.4	11.4±9.0	10.4±6.3	14.9±9.9	–	<0.001
FPG (mg/dL)	112.9±33.7	107.3±27.4	109.9±27.7	114.8±33.8	–	0.14
HOMA-IR	2.4±1.5	2.9±2.4	3.0±2.1	4.2±3.0	–	<0.001
hs-CRP (mg/dL)	1078.9±1407	1048.3±1185.0	865.8±658.4	1579.2±2377.9	–	0.027
Adiponectin (μg/mL)	7.4±4.4	8.5±6.6	6.6±2.6	6.9±4.3	–	0.24
Leptin (ng/mL)	9.9±7.4	9.1±6.2	11.3±9.4	12.4±7.9	–	<0.001
Ferritin (ng/mL)	145.8±101.1	176.5±134.0	271.2±307.0	208.3±180.3	–	0.027
Uric acid (mg/dL)	5.9±1.5	5.7±1.2	5.4±1.9	5.7±1.6	–	0.77
PLT ($\times 10^4/\mu L$)	23.0±5.9	22.9±4.9	21.9±6.7	20.2±6.4	–	<0.001
ANA (0/1/2/3/4)	42/17/4/0/0	31/8/4/1/2	15/6/2/0/0	147/76/31/8/12	–	0.015
Clinical history						
Diabetes (NGT/IGT/DM)	36/11/34	24/7/27	12/8/7	103/35/119	–	0.45*
Hyperlipidemia (+/−)	31/68	31/42	9/20	120/206	–	0.60‡
Hypertension (+/−)	64/35	33/40	19/10	155/172	–	0.013‡
Liver biopsy feature						
Brunt grade (1/2/3)	–	–	19/3/2	149/133/44	–	<0.001‡
Brunt stage (1/2/3/4)	–	–	–	123/74/105/24	–	–
Fat droplet (1/2/3/4)	38/32/19/11	14/29/18/7	7/3/10/4	51/99/104/52	–	<0.001
Iron deposition (0/1/2/3/4)	30/14/21/10/1	24/9/12/2/1	10/5/2/2/0	132/56/29/29/11	–	0.16

Measurements are shown as mean ± standard deviation. Categorical values are shown by the count number. *P*-values are calculated by Jonckheere-Terpstra test unless otherwise stated;
‡Chochran-Armitage trend test,
*Kruskal-Wallis test. Abbreviations used for each trait are summarized in materials and methods.
doi:10.1371/journal.pone.0038322.t001

Association of rs738409 Genotypes with Clinical Traits

The quantitative effects of rs738409 genotypes to clinical traits were examined by multivariable regression adjusted for age, sex and BMI (statistical calculation 1, Table 3). Five categorical ordinals, namely, anti-nuclear antibody (ANA), Brunt grade, Brunt stage, fat deposition and iron deposition, were also tested by an ordinal logistic regression analysis. Potential associations ($p<0.05$) were obtained for 11 traits, namely, aspartate transaminase (AST), alanine aminotransferase (ALT), type IV collagen 7S, hyaluronic acid, hemoglobin A1c (HbA1c), fasting immunoreactive insulin (IRI), fasting plasma glucose (FPG), platelet count (PLT), Brunt grade, fat deposition and iron deposition (Table 3). When the results were further adjusted for Matteoni type (statistical calculation 2), AST, hyaluronic acid, HbA1c, FPG,

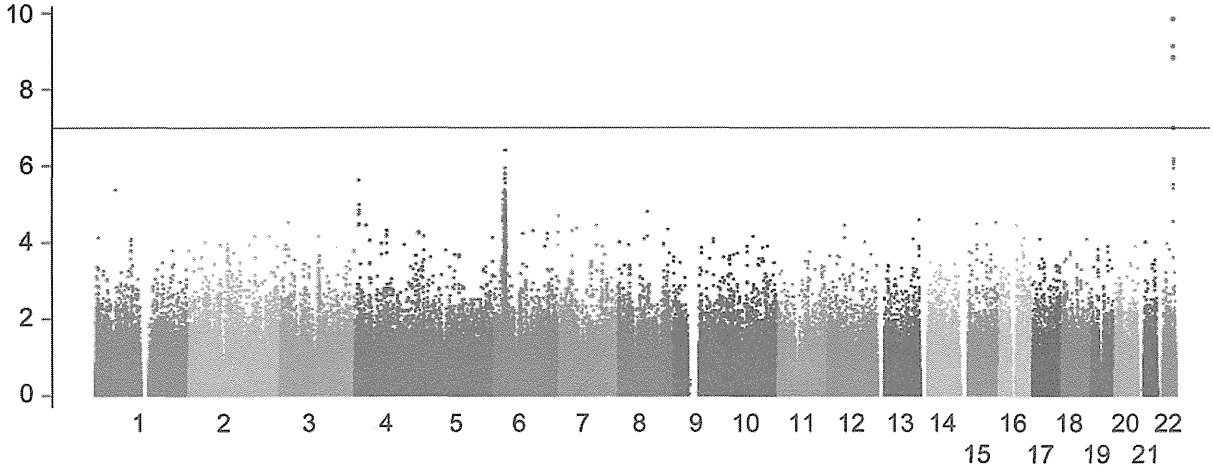

Figure 1. Manhattan plot of the GWA study. Association *p*-values are calculated by exact trend test and plotted along the chromosome in $-\log_{10}$ scale. The horizontal line indicates Bonferroni-adjusted significance threshold ($p = 1.03\times10^{-7}$).
doi:10.1371/journal.pone.0038322.g001

PLT, Brunt grade and iron deposition showed *p*-values smaller than 0.05. The level of serum triglyceride was not significant in the initial analysis, but became significant after being adjusted for Matteoni's type ($p = 0.013$). Among them, only three traits, namely, hyaluronic acid, HbA1c and iron deposition, remained significant ($p<0.0021$) after Bonferroni's correction for multiple testing (Table 3).

Associations of Previously Reported SNPs with NAFLD

Previous genetic studies identified four chromosomal loci, namely, *LYPLAL1* at 1q41, *GCKR* at 2p23, *NCAN* at 19p12 and *PPP1R3B* at 8p23.1, associated with NAFLD in populations of European descent [14]. We examined whether or not the associations were reproduced in the Japanese population by extracting genotype information of SNP markers corresponding to these four loci. As shown in Table 4, the association of rs780094 in *GCKR* with NAFLD was at the border of significance ($p = 0.011$, OR = 0.82, 95%CI: 0.70–0.91) in the case/control analysis. However, the association was lost when examined between rs780094 genotypes and Matteoni types. There were no associations of rs2228603 in *NCAN* and rs12137855 in *LYPLAL1* with either NAFLD or Matteoni types. Rs4240624 in *PPP1R3B* was not in the SNP array used for this study, and this marker was not polymorphic or at a very low frequency in the Japanese (0 in 90

Table 2. List of the SNP markers in the *PNPLA3* locus at chromosome 22q showing genome wide significance.

		Genotyping Result and Allele Frequency of A2						Statistics		
			NAFLD					NAFLD vs. Control		Matteoni
dbSNPID	A1/A2	Control	Total	Type 1	Type 2	Type 3	Type 4	*p*-value†	OR (95%CI)	*p*-value‡
rs738407	T/C	124/447/361	46/200/283	12/51/37	10/28/35	4/14/11	20/107/200	1.0×10^{-7}	1.56(1.32–1.83)	3.4×10^{-5}
		(0.627)	(0.724)	(0.625)	(0.671)	(0.621)	(0.775)			
rs738409	C/G*	247/468/217	88/236/203	20/59/21	21/30/22	8/11/9	39/136/151	1.4×10^{-10}	1.66(1.43–1.94)	3.6×10^{-6}
		(0.484)	(0.609)	(0.505)	(0.507)	(0.518)	(0.672)			
rs2076211	C/T*	248/473/211	92/242/195	21/58/21	21/30/22	8/11/10	42/143/142	1.4×10^{-9}	1.61(1.38–1.87)	3.2×10^{-5}
		(0.480)	(0.597)	(0.500)	(0.507)	(0.534)	(0.653)			
rs2896019	T/G*	246/473/213	91/234/204	20/57/23	22/29/22	7/12/10	42/136/149	1.5×10^{-10}	1.66(1.42–1.93)	2.6×10^{-5}
		(0.482)	(0.607)	(0.515)	(0.500)	(0.552)	(0.664)			
rs1010023	T/C*	249/473/210	94/239/196	21/57/22	22/29/22	7/12/10	44/141/142	1.5×10^{-9}	1.61(1.38–1.87)	6.5×10^{-5}
		(0.479)	(0.596)	(0.505)	(0.500)	(0.552)	(0.650)			
rs926633	G/A*	247/474/211	93/237/199	21/56/23	22/29/22	7/12/10	43/140/144	7.5×10^{-10}	1.62(1.39–1.89)	5.8×10^{-5}
		(0.481)	(0.600)	(0.510)	(0.500)	(0.552)	(0.654)			
rs3810622	T*/C	330/445/157	263/208/58	40/48/12	28/29/16	14/12/3	181/119/27	1.0×10^{-7}	0.64(0.55–0.75)	0.0017
		(0.407)	(0.306)	(0.360)	(0.418)	(0.310)	(0.265)			

Reference (A1) and non-reference (A2) alleles refer to NCBI Reference Sequence Build 36.3 with the effective allele marked by an asterisk. Genotyping results are shown by genotype count of A1A1/A1A2/A2A2 with allele frequency of A2 in parenthesis.
†*P*-values are calculated by exact trend test with odds ratios (OR) calculated for A2 with 95% confidence interval (CI).
‡*P*-values are calculated by Jonckheere-Terpstra test in NAFLD patients for Matteoni type and additive model of genotype. SNPs are ordered by chromosomal location.
doi:10.1371/journal.pone.0038322.t002

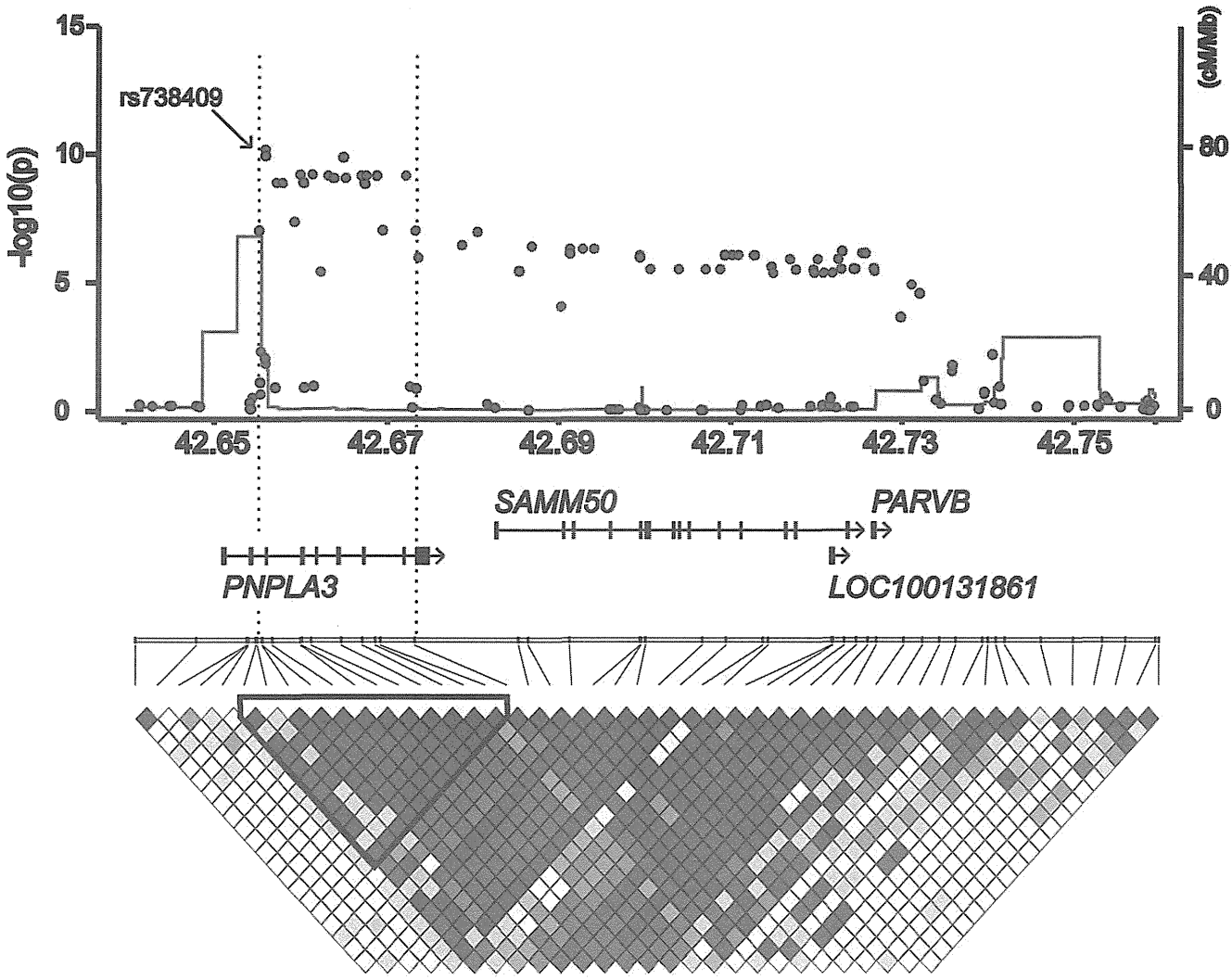

Figure 2. A schematic organization of the human *PNPLA3* locus at 22q13.31 with the genome scan results. *P*-values calculated by the exact trend test were plotted in $-\log_{10}$ scale. Red and blue dots indicate the *p*-values of genotyped and imputed SNPs, respectively. Local recombination rate obtained from HAPMAP release 22 is indicated by a red line plotted in cm/Mb scale. The structure and orientation of four genes in the region are shown below the plots with their transcriptional orientations according to NCBI Reference Sequence Build 36.3. LD blocks were generated according to pairwise LD estimates of the SNPs located within the region using the genome scan results. The LD block showing the strongest association is highlighted with the triangle, and the corresponding chromosomal region is represented by the dotted lines.
doi:10.1371/journal.pone.0038322.g002

chromosomes in the Japanese result of the International HapMap Project).

Discussion

NASH is a type of hepatic steatosis in NAFLD with poor prognosis accompanying liver fibrosis, and subsequent liver cirrhosis and hepatocellular carcinoma [18]. Despite the extensive biochemical and histological investigation of NAFLD, whether or not NASH forms a distinct disease entity in NAFLD still remains unclear. The principle aim of this study was to identify the genetic factors related to the pathogenic status of NAFLD by collecting DNA samples of Japanese NAFLD patients with critically diagnosed disease status by liver biopsy. To our knowledge, this is the first GWA study of NAFLD using patients with known histology-based Matteoni type. In the initial association study using pooled genotyping results of all the cases, we found a significant association of the *PNPLA3* gene at chromosome 22q13.31 with NAFLD in the Japanese. Rs738409 which showed the strongest association with NAFLD in the GWA study of Caucasians was also genotyped and its strongest association with NAFLD was confirmed. These results were in agreement with the former GWA analyses in populations of European descent and in Hispanics, giving strong evidence of the involvement of *PNPLA3* in NAFLD beyond ethnicities. Rs738409 is located in exon3 of the *PNPLA3* gene which is expressed in the liver and adipose tissue. This SNP introduces an amino acid substitution from isoleucine to methionine (I148M), and biological studies demonstrated that its risk allele (G) abolishes the triglyceride hydrolysis activity of *PNPLA3* [19]. These observations strongly suggest rs738409 to be a causative genetic variation for NAFLD. However, future genomic analyses by fine mapping or extensive sequencing may identify additional genetic determinants within the *PNPLA3* locus.

In the current study we did not find other genetic loci showing genome-wide significance ($p<1.0\times10^{-7}$). However, two additional chromosomal loci with *p*-values being smaller than 1×10^{-5} were

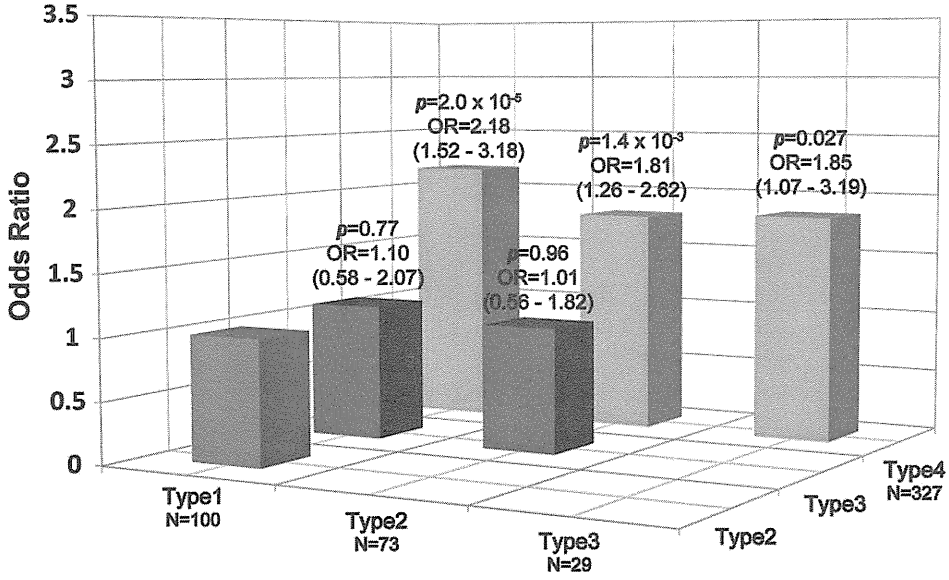

Figure 3. Histogram of odds ratios for genotype distribution of rs738409 between Matteoni types. Each box denotes the odds ratio (OR) comparing the corresponding Matteoni types on the horizontal axes. N represents the number of samples. Odds ratios and *p*-values are calculated for the higher Matteoni type per risk allele (G) on additive model by multivariable logistic regression adjusted for age, sex and BMI, and are shown with 95% CI above each box.
doi:10.1371/journal.pone.0038322.g003

identified on chromosome 1p (rs11206226) and chromosome 4p (rs1390096) neither of which has been reported as being associated with NAFLD in Caucasians (Table S1). Statistical calculation by taking their allele frequencies and effect sizes into account showed that approximately three times as many case and control samples are required to obtain sufficient statistical power (>0.8) for genome wide significance. Hence, further confirmation is required using a larger collection of patients and controls although they may be potential candidates of low-penetrance genes for susceptibility to NAFLD in Japanese.

Subsequent analyses through comparison of genotype distribution among four subgroups of NAFLD (type1 to type4) categorized by Matteoni's classification revealed that the seven NAFLD-associated SNPs in the *PNPLA3* gene were also significantly associated with the pathogenic status of NAFLD. There were also marked differences in genotype distribution of rs738409 between type4 subgroup and the other three groups ($p = 4.8 \times 10^{-6}$, OR = 1.96, 95%CI: 1.47–2.62 between type4 and pooled genotypes of type1 to type3). Moreover, a case/control analysis of rs738409 between Matteoni type4 and controls returned a surprisingly strong association ($p = 1.7 \times 10^{-16}$) which was much stronger than the initial analysis using all NAFLD cases ($p = 1.4 \times 10^{-10}$), whereas the analysis using Matteoni type1 to type3 as cases didn't show significance ($p = 0.41$). There were differences in the score of HOMA-IR and hs-CRP, indicators of insulin resistance and inflammation, respectively, between Matteoni type1 to type3 and type4 subgroups (Table 1). Our results provide compelling evidence that NASH corresponding to Matteoni type4 is both a clinically and genetically different disease subset from other spectrums of NAFLD. Previous studies showed association between *PNPLA3* and fatty liver, inflammation, fibrosis grade and NASH [13]. In our result, strong association between rs738409 and fatty liver was not observed by comparing control and Matteoni type1. In addition, strong association between rs738409 and lobular inflammation was not observed by comparing Matteoni Type1 and Type2. In contrast, a strong association between rs738409 and NASH was observed. Although we could not observe the strong association between rs738409 and fibrosis stage, strong association between rs738409 and Hyaluronic acid suggests that an association exists between *PNPLA3* and fibrosis.

We have also undertaken association analyses of rs738409 and clinical traits in the patients. The multivariable regression analysis adjusted for age, sex, BMI and Matteoni type followed by the correction for multiple testing revealed hyaluronic acid and HbA1c as being significantly associated with rs738409. Hyaluronic acid is one of the principle components of the extracellular matrix and its involvement in fibrosis has been previously suggested [20]. This may indicate another possible functional involvement of *PNPLA3* in the progression of liver fibrosis by influencing the circulating hyaluronic acid levels. A weak association of rs738409 and HbA1c levels was observed in our study population. However, there are no reports to date indicating such an association, and confirmation with different sample sets is needed for definitive conclusion. Also, the association between rs738409 and iron deposition was demonstrated by an ordinal logistic regression analysis. Since the association still remained after the results were adjusted with Matteoni type, rs738409 may play a functional role in the oxidative stress through iron absorption in the liver.

Recently, a genetic analysis of Japanese NAFLD patients was reported demonstrating a significant association in the increase of AST, ALT, ferritin levels and fibrosis stage (Brunt stage) and in the decrease of serum triglyceride with the risk allele (G) of rs738409 [12]. In our study, the association of rs738409 with AST ($p = 1.2 \times 10^{-4}$) and ALT ($p = 0.0016$) was reproduced and that of AST still remained after the results were adjusted for Matteoni type ($p = 0.038$). No association was observed for ferritin level. Brunt stage was available for Matteoni type4 patients only in our study. Although the odds ratio was slightly high (OR = 1.28, 95%CI: 0.95–1.72), it was not possible to examine the association. In addition, the inverse association of the risk allele of rs738409 with decrease of serum triglyceride was confirmed in our study ($p = 0.013$ after being adjusted for Matteoni type). For all of these

Table 3. Association of rs738409 with clinical traits.

Biochemical traits	Statistical calculation1		Statistical calculation 2	
Phenotype	**Coef. (S.E.)**	***p*-value**	**Coef. (S.E.)**	***p*-value**
Biological traits				
AST (IU/L)	0.22 (0.056)	**1.2×10^{-4}**	0.11 (0.052)	0.038
ALT (IU/L)	0.19 (0.058)	**0.0016**	0.093 (0.056)	0.098
GGT (IU/L)	−0.056 (0.061)	0.37	−0.088 (0.062)	0.16
Albumin (g/dL) *	0.015 (0.051)	0.77	−0.012 (0.052)	0.81
Total bilirubin (mg/dL)	−0.011 (0.063)	0.86	0.0059 (0.064)	0.93
Cholinesterase (unit) *	0.062 (0.040)	0.12	0.069 (0.041)	0.092
Type IV collagen 7S (ng/dL) *	−0.19 (0.064)	0.0025	−0.11 (0.062)	0.069
Hyaluronic acid (ng/dL)	0.30 (0.065)	**4.9×10^{-6}**	0.22 (0.063)	**4.6×10^{-4}**
Triglycerides (mg/dL)	−0.10 (0.058)	0.072	−0.15 (0.059)	0.013
Total cholesterol (mg/dL)	−0.066 (0.060)	0.27	−0.057 (0.061)	0.34
HbA1c (%)	−0.17 (0.053)	**0.0012**	−0.18 (0.054)	**0.0011**
IRI (μg/dL)	0.16 (0.063)	0.012	0.086 (0.061)	0.16
FPG (mg/dL)	−0.14 (0.049)	0.0047	−0.15 (0.05)	0.0035
HOMA-IR	0.084 (0.064)	0.19	0.0092 (0.062)	0.88
Hs-CRP (mg/dL)	−0.013 (0.048)	0.79	−0.031 (0.049)	0.52
Adiponectin (μg/mL)	0.048 (0.066)	0.47	0.12 (0.066)	0.072
Leptin (ng/mL)	0.11 (0.068)	0.11	0.10 (0.069)	0.15
Ferritin (ng/mL)	0.031 (0.047)	0.51	−0.0042 (0.048)	0.93
Uric acid (mg/dL)	−0.097 (0.061)	0.11	−0.11 (0.062)	0.067
PLT ($x10^4$/μL)	−0.056 (0.020)	0.0052	−0.045 (0.020)	0.028
Immunological/histological traits				
ANA (0/1/2/3/4)	0.92 (0.70–1.21)	0.56	0.86 (0.65–1.15)	0.31
Brunt grade (1/2/3)	1.42 (1.06–1.92)	0.021	1.38 (1.02–1.87)	0.036
Brunt stage (1/2/3/4)	1.28 (0.95–1.72)	0.11		
Fat deposition (1/2/3/4)	1.44 (1.15–1.81)	0.0019	1.24 (0.98–1.56)	0.76
Iron deposition (0/1/2/3/4)	0.61 (0.47–0.80)	**3.0×10^{-4}**	0.62 (0.47–0.81)	**5.6×10^{-4}**

Associations between distribution of rs738409 genotypes and clinical traits are calculated by multivariable regression. Statistical calculation1 is adjusted for age, sex and BMI, while the Matteoni types are additionally included as covariate in statistical calculation 2. Statistics are calculated by multivariable linear regression for biochemical traits and by multivariable ordinal logistic regression for immunological and histological traits.

Coefficients and odds ratios are calculated for the increase of each trait per risk allele (G). The *p*-values showing significance after Bonferroni's correction for multiple testing ($p = 0.0021$) was shown in bold.

*Reciprocal numbers are used for normalization and a negative coefficient implicates an increase in value according to the increase of the risk allele.

doi:10.1371/journal.pone.0038322.t003

biomarkers, however, the significance was lost after the correction for multiple testing.

A replication analysis of other genetic loci that had been reported for their association with NAFLD in East coast white Americans [14] was performed in our sample collection. We confirmed the association of rs780094 in *GCKR* with NAFLD in a case/control analysis but at a much weaker level ($p = 0.011$, OR = 0.82, 95%CI: 0.70–0.95) than that shown for the populations of European-descent. No associations were found for *LYPLAL1* and *NCAN* loci in our study. There are several reasons to explain such differences, such as the insufficient statistical power with a limited number of study subjects in our study due to the difficulty in the collection of a larger number of histologically diagnosed NAFLD patients. The difference in genetic background between the Japanese and Europeans is also conceivable. Indeed, the risk allele frequency of rs12137855 in *LYPLAL1* was 0.944 in our control subjects but approximately 0.79 in the European populations [14]. Similarly, there was a difference in the risk allele frequency of rs2228603 in *NCAN* (0.049 in Japanese and 0.08 in Europeans). Rs4240624 in *PPP1R3B* was not polymorphic in the Japanese while its risk allele frequency was 0.91 in Europeans.

Materials and Methods

Ethics Statement

In compliance with the Declaration of Helsinki, ethical approval for this study was given by the respective Institutional Review Board and subject written informed consent were obtained for all subjects (Ethical committee of Nara City Hospital; Ethical committee of Saiseikai Suita Hospital; Medical Ethics Committee of Kanazawa University; Ethics committee of Kyoto Prefectural University of Medicine; Ethical Committee of Aichi Cancer Center; Ethical Committee of Kochi Medical School, Kochi University; Ethics Committee of Tokyo Women's Medical University; Ethical Committee on Kawasaki Medical School and Kawasaki Medical School Hospital; Ethical Committee of

Juntendo University; Ethics Committee of Yamagata University School of Medicine; Ethical Committee of the Ikeda Municipal Hospital; Institutional Review Board and Ethics Committee of Kyoto University School of Medicine).

Study Population

A total of 543 patients histologically diagnosed for NAFLD in 2007–2009 were recruited through the Japan study of Nonalcoholic Fatty Liver Disease. Biopsy specimens were stained with H&E and Masson's trichrome for morphological review and assessment of fibrosis. Perl's Prussian blue was performed to evaluate iron load. Biopsy specimens were reviewed by a hepatopathologist (T.O). NAFLD patients were classified into four categories by liver histology according to the classification by Matteoni *et al* [2] as follows; type1: fatty liver alone, type2: fat accumulation and lobular inflammation, type3: fat accumulation and ballooning degeneration, type4: fat accumulation, ballooning degeneration, and either Mallory-Denk body or fibrosis. With these criteria, the 543 patients were classified as type1; 102, type2; 75, type3; 31 and type4; 335. The histological grade and fibrosis stage were also evaluated by the classification of Brunt *et al* [21] for advanced NAFLD cases (type3 and type4) as follows; grade 1: steatosis involving up to 66% of biopsy, occasional ballooned zone 3 hepatocytes and absence or mild portal chronic inflammation, grade2: steatosis, ballooning hepatocytes mild to moderate chronic inflammation, grade3: panacinar steatosis, ballooning and disarray obvious and mild or portal mild to moderate inflammation, stage1: perivenular and/or perisinusoidal fibrosis in zone3, stage2: combined pericellular portal fibrosis, stage3: septal/bridging fibrosis, stage4: cirrhosis. The degree of fat deposition was evaluated by amount of fat droplets as observed under the microscope as follows; 0: <5%, 1: 5–<10%, 2: 10–<34%, 3: 34–<67%, 4: >67%. The degree of iron deposition was categorized by the presence of granules of free iron observed under the microscope as follows; 0: absence by x400, 1: easily identifiable by x400 and rarely identifiable by x250, 2: identifiable by x100, 3: identifiable by x25, 4: identifiable at lower than x25.

Inclusion criteria for NAFLD patients were as follows; (i) no history of alcoholism, (ii) no history for HBV/HCV/HIV infection, (iii) diagnosed by liver biopsy, (iv) information regarding age and BMI available. The sex of two samples was unknown, and was imputed from the results of the genome scan. As general Japanese population controls, the genome scan results of 942 healthy Japanese volunteers from Aichi Cancer Center Hospital and Research Institute were used [22].

Anthropometric and Laboratory Evaluation

We employed conventional methods for the measurement of anthropometry (height, weight, amount of visceral fat and abdominal circumscript). BMI was calculated from the measurements. The following biochemical/hematological/immunological traits were also measured by conventional methods; aspartate aminotransferase (AST), alanine aminotransferase (ALT), γ-glutamyl transpeptidase (GGT), albumin, total bilirubin, cholinesterase, type IV collagen 7S, hyaluronic acid, triglyceride, total cholesterol, hemoglobin A1c (HbA1c), fasting immunoreactive insulin (IRI), fasting plasma glucose (FBS), high sensitive CRP (hs-CRP), adiponectin, leptin, ferritin, uric acid, and platelet (PLT) count. Anti nuclear antibody (ANA) was measured by ELISA and categorized by the detection limit in a serial dilution as follows; 0: <40x, 1: 40–80x, 2: 81–160x, 3: 160x, 4: >320x. Homeostasis model assessment-insulin resistance (HOMA-IR) was calculated from the measurements. Patients were assigned a diagnosis of diabetes mellitus (DM) when they had documented use of oral hypoglycemic medication, a random glucose level >200 mg/dl, or FPG >126 mg/dl. Hyperlipidemia was diagnosed with the cholesterol level being >200 mg/dl and/or triglyceride level being >160 mg/dl. Hypertension was diagnosed when the patient was taking antihypertensive medication and/or had a resting recumbent blood pressure ≥140/90 mmHg on at least two occasions.

DNA Preparation

Genomic DNA was extracted from peripheral blood mononuclear cells by standard phenol-chloroform extraction and resuspended in TE buffer. DNA concentration and purity were measured with Nanodrop 1000 spectrophotometer (Thermo Scientific, Waltham, MA, USA). The samples were stored at −20°C until use.

Genome-wide Genotyping and Quality Control

Genome scan was conducted for 543 patients with NAFLD and 942 healthy subjects using Illumina Human 610-Quad Bead Chip on a Bead Station 500G Genotyping System (Illumina, Inc., San Diego, CA, USA) and subjected to the following quality controls. Initially, ten patients and six control subjects were removed due to low call rates (<0.99). Regarding the SNP markers, 85,472 SNPs with minor allele frequency of smaller than 0.01 in either case or control group, 6,479 SNPs with lower success rates (<0.98) and 35 SNPs with distorted Hardy-Weinberg equilibrium ($p<10^{-7}$) were removed, resulting in 484,751 SNP markers being used for analysis. Principal component analysis by EIGENSOFT [17] including phase II HapMap (http://hapmap.ncbi.nlm.nih.gov/) samples identified no samples that were deviated from the Japanese population. Subsequently, the degree of kinship between individuals was examined by pi-hat in PLINK 1.07 (http://pngu.mgh.harvard.edu/purcell/plink/) [23]. Of the eight pairs of samples (four case pairs and four control pairs) showing high degrees of kinship (PI-HAT>0.4), the sample with the lower call rate in each pair was removed. After these steps, 529 case and 932 controls were used for the analysis.

Statistical Analysis

A case/control association analysis was performed by exact trend test between NAFLD patients and control subjects [24]. The correction of obtained *p*-values for population stratification was performed using EIGENSTRAT [17]. In addition, an association between Matteoni classification (type1 to type4) and additive model of genotype for each SNP was examined using Jonckheere-Terpstra test for NAFLD patients. Assessment of population stratification of inflation of *p*-value was carried out by the genomic control method for asymptotic trend test [25]. Association between each quantitative trait and the genotype of significant SNPs in NAFLD patients were calculated by multivariable linear regression or multivariable ordinal regression adjusted for age, sex and BMI. Each quantitative trait was transformed as follows; natural log for ALT, AST, HOMA-IR, HbA1c, IRI, triglyceride, total bilirubin, adiponectin, hs-CRP, hyaluronic acid, leptin, reciprocal number for albumin, cholinesterase, type IV collagen 7S and square root for uric acid and ferritin. The values of FPG, PLT, total cholesterol, amount of visceral fat, and abdominal circumscript were not transformed. For each trait, values that were within only 4 S.D. were included for analysis. LD indices were calculated by default setting of Haploview [26] and the LD block was defined manually.

Table 4. Replication study of previously reported SNPs.

			Genotyping Result and Allele Frequency of A2					Statistics		
			NAFLD					NAFLD vs. Control		Matteoni
dbSNPID	A1/A2	Gene	Control	Type 1	Type 2	Type 3	Type 4	*p*-value†	OR (95%CI)	*p*-value‡
rs12137855	C*/T	*LYPLAL1*	828/102/2	90/10/0	67/6/0	24/5/0	294/33/0	0.55	0.89	0.98
			(0.056)	(0.050)	(0.041)	(0.086)	(0.050)		(0.64–1.25)	
rs780094	T*/C	*GCKR*	321/433/178	34/54/12	28/34/11	17/11/1	133/139/55	0.011	0.82	0.92
			(0.423)	(0.390)	(0.383)	(0.224)	(0.381)		(0.70–0.95)	
rs4240624	G/A	*PPP1R3B*	–	–	–	–	–	–	–	–
rs2228603	C/T*	*NCAN*	842/88/2	93/7/0	65/8/0	28/1/0	292/31/4	0.80	1.05	0.58
			(0.049)	(0.035)	(0.054)	(0.017)	(0.059)		(0.75–1.48)	

Reference (A1) and non-reference (A2) alleles refer to NCBI Reference Sequence Build 36.3 with the effective allele marked by an asterisk. Genotyping results are shown by genotype count of A1A1/A1A2/A2A2 with allele frequency of A2 in parenthesis. †*P*-values are calculated by exact trend test with odds ratios (OR) calculated for A2 with 95% confidence interval (CI). ‡*P*-values are calculated by Jonckheere-Terpstra test in NAFLD patients for Matteoni type and additive model of genotype.
doi:10.1371/journal.pone.0038322.t004

Supporting Information

Figure S1 QQ plot of the GWA study comparing distribution of the observed and expected *p*-values. Upper box is expressed in antilog scale and the lower box is expressed in $-\log_{10}$ scale. The X- and Y-axis correspond to expected and observed *p*-values. Blue and red dots denote before and after correction by genomic control method ($\lambda = 1.04$), respectively.
(DOC)

Table S1 List of the SNPs showing $p<1.0\times10^{-5}$ in the GWA study. Reference (A1) and non-reference (A2) alleles refer to NCBI Reference Sequence Build 36.3 with the effective allele marked by an asterisk. Genotyping results are shown by genotype count of A1A1/A1A2/A2A2 with allele frequency of A2 in parenthesis. †*P*-values are calculated by exact trend test with odds ratios (OR) calculated for A2 with 95% confidence interval (CI). ‡*P*-values are calculated by Jonckheere-Terpstra test in NAFLD patients for Matteoni type and additive model of genotype. SNPs are ordered by chromosomal location.
(DOC)

Acknowledgments

The authors would like to thank Yutaka Kohgo, Hirofumi Uto and Tetsuo Takehara for sample collection and Hisako Imamura and Hiroyuki Uneme for data management.

Author Contributions

Conceived and designed the experiments: FM TO. Performed the experiments: MT M. Kokubo. Analyzed the data: TK RY FM. Contributed reagents/materials/analysis tools: TK YS AU KM MT TT KY T. Saibara EH M. Kokubo SW SK YI M. Kawanaka T. Shima HP HT KT RY. Wrote the paper: TK MT RY FM TO.

References

1. Ludwig J, Viggiano TR, McGill DB, Oh BJ (1980) Nonalcoholic steatohepatitis: Mayo Clinic experiences with a hitherto unnamed disease. Mayo Clin Proc 55: 434–438.
2. Matteoni CA, Younossi ZM, Gramlich T, Boparai N, Liu YC, et al. (1999) Nonalcoholic fatty liver disease: a spectrum of clinical and pathological severity. Gastroenterology 116: 1413–1419.
3. Cohen JC, Horton JD, Hobbs HH (2011) Human fatty liver disease: old questions and new insights. Science 332: 1519–1523. doi:10.1126/science.1204265.
4. Vernon G, Baranova A, Younossi ZM (2011) Systematic review: the epidemiology and natural history of non-alcoholic fatty liver disease and non-alcoholic steatohepatitis in adults. Aliment Pharmacol Ther 34: 274–285. doi:10.1111/j.1365–2036.2011.04724.x.
5. Okanoue T, Umemura A, Yasui K, Itoh Y (2011) Nonalcoholic fatty liver disease and nonalcoholic steatohepatitis in Japan. J Gastroenterol Hepatol 26 Suppl 1: 153–162. doi:10.1111/j.1440–1746.2010.06547.x.
6. Williams CD, Stengel J, Asike MI, Torres DM, Shaw J, et al. (2011) Prevalence of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis among a largely middle-aged population utilizing ultrasound and liver biopsy: a prospective study. Gastroenterology 140: 124–131. doi:10.1053/j.gastro.2010.09.038.
7. Okanoue T (2011) Recent progress in the research of NASH/NAFLD in Japan. Nihon Shokakibyo Gakkai Zasshi 108: 1161–1169.
8. Berson A, De Beco V, Lettéron P, Robin MA, Moreau C, et al. (1998) Steatohepatitis-inducing drugs cause mitochondrial dysfunction and lipid peroxidation in rat hepatocytes. Gastroenterology 114: 764–774.
9. Day CP (2006) From fat to inflammation. Gastroenterology 130: 207–210. doi:10.1053/j.gastro.2005.11.017.
10. Romeo S, Kozlitina J, Xing C, Pertsemlidis A, Cox D, et al. (2008) Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. Nat Genet 40: 1461–1465. doi:10.1038/ng.257.
11. Sookoian S, Castaño GO, Burgueño AL, Gianotti TF, Rosselli MS, et al. (2009) A nonsynonymous gene variant in the adiponutrin gene is associated with nonalcoholic fatty liver disease severity. J Lipid Res 50: 2111–2116. doi:10.1194/jlr.P900013-JLR200.
12. Hotta K, Yoneda M, Hyogo H, Ochi H, Mizusawa S, et al. (2010) Association of the rs738409 polymorphism in PNPLA3 with liver damage and the development of nonalcoholic fatty liver disease. BMC Med Genet 11: 172. doi:10.1186/1471–2350–11–172.
13. Sookoian S, Pirola CJ (2011) Meta-analysis of the influence of I148M variant of patatin-like phospholipase domain containing 3 gene (PNPLA3) on the susceptibility and histological severity of nonalcoholic fatty liver disease. Hepatology 53: 1883–1894. doi:10.1002/hep.24283.
14. Speliotes EK, Yerges-Armstrong LM, Wu J, Hernaez R, Kim LJ, et al. (2011) Genome-wide association analysis identifies variants associated with nonalcoholic Fatty liver disease that have distinct effects on metabolic traits. PLoS Genet 7: e1001324. doi:10.1371/journal.pgen.1001324.
15. Petersen KF, Dufour S, Hariri A, Nelson-Williams C, Foo JN, et al. (2010) Apolipoprotein C3 gene variants in nonalcoholic fatty liver disease. N Engl J Med 362: 1082–1089. doi:10.1056/NEJMoa0907295.
16. Terao C, Yamada R, Ohmura K, Takahashi M, Kawaguchi T, et al. (2011) The human AIRE gene at chromosome 21q22 is a genetic determinant for the predisposition to rheumatoid arthritis in Japanese population. Human Molecular Genetics 20: 2680–2685. doi:10.1093/hmg/ddr161.
17. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904–909. doi:10.1038/ng1847.
18. Yasui K, Hashimoto E, Komorizono Y, Koike K, Arii S, et al. (2011) Characteristics of patients with nonalcoholic steatohepatitis who develop hepatocellular carcinoma. Clin Gastroenterol Hepatol 9: 428–433; quiz e50. doi:10.1016/j.cgh.2011.01.023.

19. He S, McPhaul C, Li JZ, Garuti R, Kinch L, et al. (2010) A Sequence Variation (I148M) in PNPLA3 Associated with Nonalcoholic Fatty Liver Disease Disrupts Triglyceride Hydrolysis. J Biol Chem 285: 6706–6715. doi:10.1074/jbc.M109.064501.
20. Ueno T, Inuzuka S, Torimura T, Tamaki S, Koh H, et al. (1993) Serum hyaluronate reflects hepatic sinusoidal capillarization. Gastroenterology 105: 475–481.
21. Brunt EM, Janney CG, Di Bisceglie AM, Neuschwander-Tetri BA, Bacon BR (1999) Nonalcoholic steatohepatitis: a proposal for grading and staging the histological lesions. Am J Gastroenterol 94: 2467–2474. doi:10.1111/j.1572–0241.1999.01377.x.
22. Suzuki T, Matsuo K, Sawaki A, Mizuno N, Hiraki A, et al. (2008) Alcohol Drinking and One-Carbon Metabolism-Related Gene Polymorphisms on Pancreatic Cancer Risk. Cancer Epidemiology Biomarkers & Prevention 17: 2742–2747. doi:10.1158/1055–9965.EPI-08–0470.
23. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet 81: 559–575.
24. Yamada R, Okada Y (2009) An optimal dose-effect mode trend test for SNP genotype tables. Genet Epidemiol 33: 114–127. doi:10.1002/gepi.20362.
25. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997–1004.
26. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263–265. doi:10.1093/bioinformatics/bth457.

nature genetics

Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population

Yukinori Okada[1–3,40], Chikashi Terao[4,5,40], Katsunori Ikari[6,40], Yuta Kochi[1,2,40], Koichiro Ohmura[5], Akari Suzuki[1], Takahisa Kawaguchi[4], Eli A Stahl[7,8], Fina A S Kurreeman[7–9], Nao Nishida[10], Hiroko Ohmiya[3], Keiko Myouzen[1], Meiko Takahashi[4], Tetsuji Sawada[11], Yuichi Nishioka[12], Masao Yukioka[13], Tsukasa Matsubara[14], Shigeyuki Wakitani[15], Ryota Teshima[16], Shigeto Tohma[17], Kiyoshi Takasugi[18], Kota Shimada[17], Akira Murasawa[19], Shigeru Honjo[20], Keitaro Matsuo[21], Hideo Tanaka[21], Kazuo Tajima[22], Taku Suzuki[6,23], Takuji Iwamoto[6,23], Yoshiya Kawamura[24], Hisashi Tanii[25], Yuji Okazaki[26], Tsukasa Sasaki[27], Peter K Gregersen[28], Leonid Padyukov[29], Jane Worthington[30], Katherine A Siminovitch[31], Mark Lathrop[32,33], Atsuo Taniguchi[6], Atsushi Takahashi[3], Katsushi Tokunaga[10], Michiaki Kubo[34], Yusuke Nakamura[35], Naoyuki Kamatani[36], Tsuneyo Mimori[5], Robert M Plenge[7,8], Hisashi Yamanaka[6], Shigeki Momohara[6,41], Ryo Yamada[37,41], Fumihiko Matsuda[4,38,39,41] & Kazuhiko Yamamoto[1,2,41]

Rheumatoid arthritis is a common autoimmune disease characterized by chronic inflammation. We report a meta-analysis of genome-wide association studies (GWAS) in a Japanese population including 4,074 individuals with rheumatoid arthritis (cases) and 16,891 controls, followed by a replication in 5,277 rheumatoid arthritis cases and 21,684 controls. Our study identified nine loci newly associated with rheumatoid arthritis at a threshold of $P < 5.0 \times 10^{-8}$, including *B3GNT2, ANXA3, CSF2, CD83, NFKBIE, ARID5B, PDE2A-ARAP1, PLD4* and *PTPN2*. *ANXA3* was also associated with susceptibility to systemic lupus erythematosus ($P = 0.0040$), and *B3GNT2* and *ARID5B* were associated with Graves' disease ($P = 3.5 \times 10^{-4}$ and 2.9×10^{-4}, respectively). We conducted a multi-ancestry comparative analysis with a previous meta-analysis in individuals of European descent (5,539 rheumatoid arthritis cases and 20,169 controls). This provided evidence of shared genetic risks of rheumatoid arthritis between the populations.

Rheumatoid arthritis is a complex autoimmune disease characterized by inflammation and the destruction of synovial joints and affects up to 1% of the population worldwide. To date, more than 35 rheumatoid arthritis susceptibility loci, including *HLA-DRB1*, *PTPN22*, *PADI4*, *STAT4*, *TNFAIP3* and *CCR6*, among others, have been identified by GWAS in multiple populations[1–12] and by several meta-analyses of the original GWAS[13–16]. In particular, each meta-analysis of these GWAS uncovered a number of loci that were not identified in the single GWAS, leading to recognition of the enormous power of the meta-analysis approach for detecting causal genes in disease. However, these previous meta-analyses have been performed solely in European populations[13–16] and not in Asian ones. As multi-ancestry studies on validated rheumatoid arthritis susceptibility loci showed the existence of both population-specific and shared genetic components of rheumatoid arthritis[10,17], additional studies in Asian populations might provide useful insight into the underlying genetic architecture of rheumatoid arthritis, which would otherwise be difficult to capture using the studies in a single population. Here, we report a meta-analysis of GWAS and a replication study for rheumatoid arthritis in a Japanese population that was conducted by the Genetics and Allied research in Rheumatic diseases NETworking (GARNET) consortium[10,12]. We subsequently performed a multi-ancestry comparative analysis that incorporated results from a previously conducted meta-analysis of individuals of European ancestry[15].

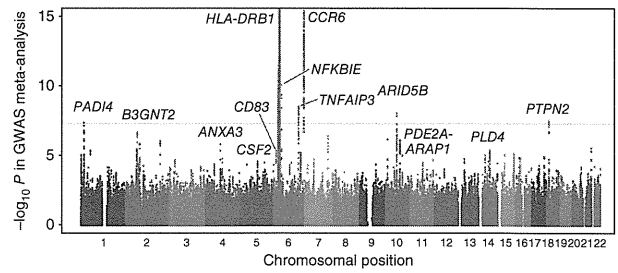

Figure 1 Manhattan plots of the GWAS meta-analysis for rheumatoid arthritis in the Japanese population. The genetic loci that satisfied the genome-wide significance threshold of $P < 5.0 \times 10^{-8}$ (gray line) in the meta-analysis or in the combined study of the meta-analysis and the replication study are presented. The *y* axis shows the $-\log_{10} P$ values of the SNPs in the meta-analysis. The SNPs for which the *P* values were smaller than 1.0×10^{-15} are indicated at the upper limit of the plot.

A full list of author affiliations appears at the end of the paper.

Received 24 October 2011; accepted 1 March 2012; published online 25 March 2012; doi:10.1038/ng.2231

Table 1 Results of the GWAS meta-analysis and the replication studies for rheumatoid arthritis

					Associations in Japanese								Associations in Europeans[c]			
					GWAS meta-analysis				Replication study		Combined study		GWAS meta-analysis			
						Allele 1 freq.							Allele 1 Freq.			
rsID[a]	Chr.	Position (bp)	Cytoband	Gene(s)	Allele 1/2	RA	Control	OR (95% CI)[b]	*P*	OR (95% CI)[b]	*P*	OR (95% CI)[b]	*P*	RA	Control	OR (95% CI)[b]	*P*
SNPs with significant associations ($P < 5.0 \times 10^{-8}$ in the combined study)																	
rs11900673	2	62306165	2p15	*B3GNT2*	T/C	0.31	0.28	1.15 (1.08–1.21)	3.5×10^{-6}	1.09 (1.04–1.14)	6.0×10^{-4}	1.11 (1.07–1.15)	1.1×10^{-8}	0.13	0.13	1.05 (0.98–1.13)	0.17
rs2867461	4	79732239	4q21	*ANXA3*	A/G	0.46	0.44	1.13 (1.08–1.19)	4.7×10^{-6}	1.12 (1.08–1.17)	1.2×10^{-7}	1.13 (1.09–1.17)	1.2×10^{-12}	0.37	0.37	0.98 (0.92–1.04)	0.52
rs657075	5	131458017	5q31	*CSF2*	A/G	0.38	0.36	1.12 (1.06–1.18)	3.2×10^{-5}	1.11 (1.06–1.16)	3.8×10^{-6}	1.12 (1.08–1.15)	2.8×10^{-10}	0.10	0.10	1.04 (0.95–1.13)	0.37
rs12529514	6	14204637	6p23	*CD83*	C/T	0.16	0.14	1.19 (1.10–1.27)	6.8×10^{-6}	1.11 (1.05–1.18)	6.0×10^{-4}	1.14 (1.09–1.19)	2.0×10^{-8}	0.055	0.053	1.11 (0.99–1.24)	0.074
rs2233434	6	44340898	6p21.1	*NFKBIE*	G/A	0.24	0.21	1.23 (1.16–1.31)	9.2×10^{-11}	1.17 (1.11–1.23)	2.2×10^{-9}	1.19 (1.15–1.24)	5.8×10^{-19}	0.059	0.040	1.57 (1.11–2.21)	0.0099
rs10821944	10	63455095	10q21	*ARID5B*	G/T	0.39	0.36	1.17 (1.11–1.23)	1.0×10^{-8}	1.15 (1.10–1.20)	3.0×10^{-10}	1.16 (1.12–1.20)	5.5×10^{-18}	0.29	0.26	1.11 (1.05–1.17)	1.9×10^{-4}
rs3781913	11	72051144	11q13	*PDE2A-ARAP1*	T/G	0.71	0.69	1.11 (1.05–1.17)	3.2×10^{-4}	1.13 (1.08–1.18)	6.7×10^{-7}	1.12 (1.08–1.16)	5.8×10^{-10}	0.45	0.43	1.04 (0.99–1.09)	0.13
rs2841277	14	104462050	14q32	*PLD4*	T/C	0.72	0.69	1.11 (1.05–1.18)	2.8×10^{-4}	1.18 (1.13–1.24)	7.0×10^{-12}	1.15 (1.11–1.19)	1.9×10^{-14}	0.47	0.46	1.02 (0.96–1.09)	0.54
rs2847297	18	12787694	18p11	*PTPN2*	G/A	0.37	0.33	1.16 (1.11–1.23)	3.5×10^{-8}	1.06 (1.01–1.11)	0.013	1.10 (1.07–1.14)	2.2×10^{-8}	0.36	0.34	1.10 (1.05–1.15)	9.2×10^{-5}
SNPs with suggestive associations ($5.0 \times 10^{-8} \leq P < 5.0 \times 10^{-6}$ in the combined study)																	
rs4937362	11	127997949	11q24	*ETS1-FLI1*	T/C	0.71	0.68	1.13 (1.07–1.19)	2.0×10^{-5}	1.07 (1.02–1.12)	0.0061	1.09 (1.06–1.13)	7.5×10^{-7}	0.46	0.44	1.06 (1.01–1.11)	0.015
rs3783637	14	54417868	14q22	*GCH1*	C/T	0.76	0.74	1.13 (1.07–1.20)	6.5×10^{-5}	1.07 (1.02–1.13)	0.0062	1.10 (1.06–1.14)	2.0×10^{-6}	0.88	0.88	0.99 (0.88–1.11)	0.87
rs1957895	14	60978085	14q23	*PRKCH*	G/T	0.40	0.39	1.12 (1.06–1.18)	4.1×10^{-5}	1.07 (1.02–1.12)	0.0022	1.09 (1.05–1.13)	3.6×10^{-7}	0.093	0.089	1.01 (0.95–1.07)	0.73
rs6496667	15	88694672	15q26	*ZNF774*	A/C	0.38	0.35	1.13 (1.07–1.19)	4.7×10^{-5}	1.07 (1.02–1.11)	0.0050	1.09 (1.05–1.13)	1.4×10^{-6}	0.21	0.20	1.07 (1.01–1.13)	0.031
rs7404928	16	23796341	16p12	*PRKCB1*	T/C	0.65	0.62	1.13 (1.07–1.19)	1.5×10^{-5}	1.05 (1.01–1.10)	0.026	1.08 (1.05–1.12)	4.0×10^{-6}	0.75	0.75	1.01 (0.94–1.09)	0.79
rs2280381	16	84576134	16q24	*IRF8*	T/C	0.86	0.84	1.16 (1.08–1.25)	1.0×10^{-4}	1.09 (1.03–1.15)	0.0049	1.12 (1.07–1.17)	2.4×10^{-6}	0.62	0.60	1.05 (0.99–1.11)	0.081
SNPs in previously reported rheumatoid arthritis susceptibility loci ($P < 5.0 \times 10^{-8}$ in the GWAS)																	
rs766449	1	17547439	1p36	*PADI4*	T/C	0.44	0.40	1.17 (1.11–1.24)	4.6×10^{-8}	–	–	–	–	0.38	0.37	1.09 (1.03–1.05)	0.0022
rs2157337	6	32609122	6p21.3	*HLA-DRB1*	C/T	0.59	0.44	1.99 (1.88–2.11)	2.6×10^{-118}	–	–	–	–	0.69	0.46	2.50 (2.39–2.62)	$< 1.0 \times 10^{-300}$
rs6932056	6	138284130	6q23	*TNFAIP3*	C/T	0.092	0.073	1.35 (1.23–1.49)	3.2×10^{-9}	–	–	–	–	0.044	0.034	1.41 (1.24–1.60)	1.3×10^{-7}
rs1571878	6	167460832	6q27	*CCR6*	C/T	0.54	0.48	1.31 (1.24–1.39)	3.2×10^{-19}	–	–	–	–	0.47	0.43	1.13 (1.08–1.19)	5.9×10^{-7}

Chr., chromosome; Freq., frequency; RA, rheumatoid arthritis; OR, odds ratio; CI, confidence interval.
[a]SNPs with $P < 5.0 \times 10^{-6}$ in the combined study of the GWAS meta-analysis and the replication study or SNPs with $P < 5.0 \times 10^{-8}$ in the GWAS meta-analysis are annotated according to forward strand and NCBI Build 36.3. Full results of the replication study are provided in **Supplementary Table 3**. [b]Odds ratio of allele 1. [c]Associations in the previous meta-analysis in European populations[15].

The meta-analysis included 4,074 rheumatoid arthritis cases (with 81.4% and 80.4% of the subjects being positive for antibody to cyclic citrullinated peptide (anti-CCP) and rheumatoid factor, respectively) and 16,891 controls from three GWAS of Japanese subjects (from the BioBank Japan Project[10,18], Kyoto University[12] and the Institute of Rheumatology Rheumatoid Arthritis (IORRA)[19]; **Supplementary Table 1**). After the application of stringent quality control criteria, including principal-component analysis (PCA; **Supplementary Fig. 1**) for each GWAS, the meta-analysis was conducted by evaluating ~2.0 million autosomal SNPs with minor allele frequencies (MAFs) ≥0.01, which were obtained through whole-genome imputation of genotypes on the basis of the HapMap Phase 2 East Asian panels (Japanese in Tokyo (JPT) and Han Chinese in Beijing (CHB)). The inflation factor of the test statistics in the meta-analysis λ_{GC} was as low as 1.036, suggesting no substantial effects of population structure (**Supplementary Table 2**). The quantile-quantile plot of *P* values showed a marked discrepancy in the values in its tail from those anticipated under the null hypothesis that there is no association—even after removal of the SNPs located in the human leukocyte antigen (HLA) region, the major rheumatoid arthritis susceptibility locus—thereby showing the presence of significant associations in the meta-analysis (**Supplementary Fig. 2**).

We identified seven loci in the current meta-analysis that satisfied the genome-wide significance threshold of $P < 5.0 \times 10^{-8}$. These included previously known rheumatoid arthritis susceptibility loci, such as *PADI4* at 1p36, *HLA-DRB1* at 6p21.3, *TNFAIP3* at 6q23 and *CCR6* at 6q27 (refs. 1,3,6,10,15) (the smallest $P = 2.6 \times 10^{-118}$ was found at the *HLA-DRB1* locus; **Fig. 1** and **Table 1**). To our knowledge, the other three loci identified, *NFKBIE* at 6p21.1, *ARID5B* at 10q21 and *PTPN2* at 18p11, are newly associated ($P = 9.2 \times 10^{-11}$, 1.0×10^{-8} and 3.5×10^{-8}, respectively).

To validate the associations identified in the meta-analysis, we conducted a replication study of two independent Japanese rheumatoid arthritis case-control cohorts (cohort 1: 3,830 rheumatoid arthritis cases and 17,920 controls, cohort 2: 1,447 rheumatoid arthritis cases and 3,764 controls; **Supplementary Table 1**). To increase the number of subjects and enhance statistical power, genotype data obtained from other GWAS projects conducted for non-autoimmune diseases in Japanese using Illumina platforms were used for the replication control panels. For each of the 46 loci that exhibited $P < 5.0 \times 10^{-4}$ in

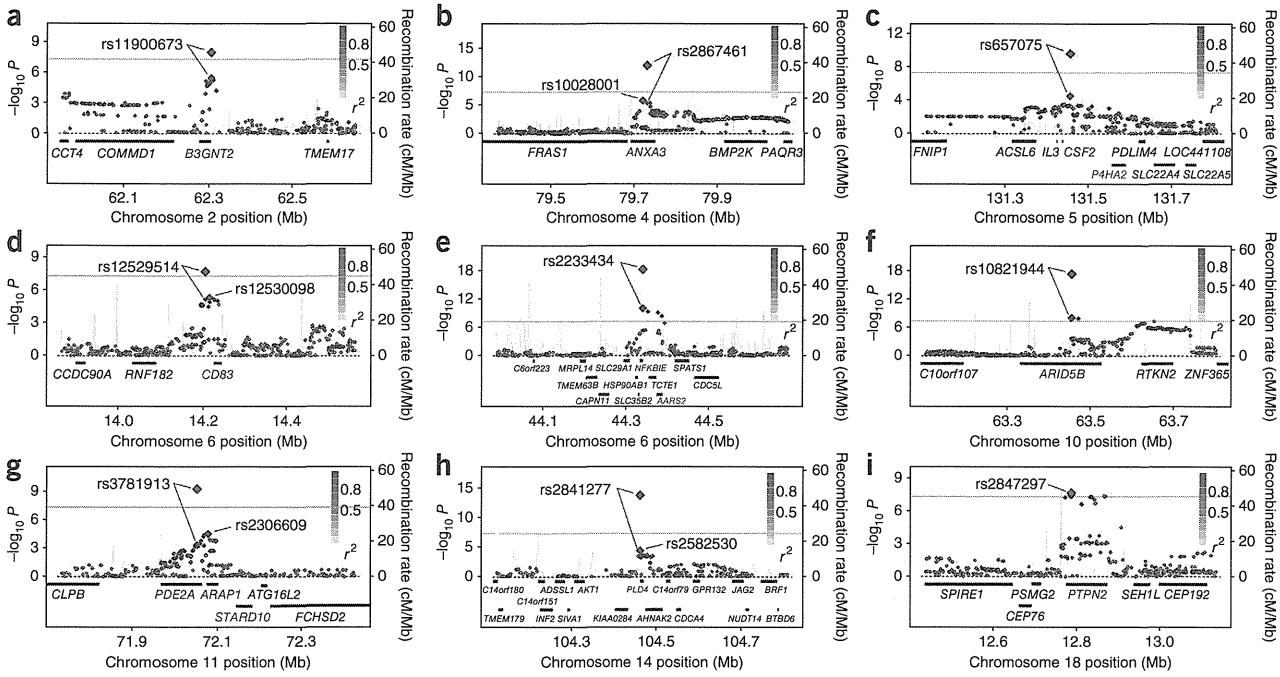

Figure 2 Regional plots of the loci newly associated with rheumatoid arthritis at the genome-wide significance threshold of $P < 5.0 \times 10^{-8}$ in the combined study of the meta-analysis and the replication study. (**a–i**) Regional plots are shown at *B3GNT2* (**a**), *ANXA3* (**b**), *CSF2* (**c**), *CD83* (**d**), *NFKBIE* (**e**), *ARID5B* (**f**), *PDE2A-ARAP1* (**g**), *PLD4* (**h**) and *PTPN2* (**i**). Diamonds represent the $-\log_{10} P$ values of the SNPs, and the red diamonds represent the $-\log_{10} P$ values of the SNPs in the meta-analysis. Red color for the smaller circles represents the r^2 value with the most significantly associated SNP (larger red circle). The purple circle represents the *P* value in the combined study. The blue line shows the recombination rates given by the HapMap Phase 2 east Asian populations (release 22). RefSeq genes at the loci are indicated below. Genes nearest to the marker SNPs at the loci are colored blue (**Supplementary Note**), and genes implicated in eQTL analysis are colored red (**Supplementary Table 4**). At 11q13, two genes (*PDE2A* and *ARAP1*) that are nearest to the SNP selected for the replication study and the most significant SNP in the meta-analysis are highlighted. The plots were drawn using SNP Annotation and Proxy Search (SNAP) version 2.2.

the meta-analysis and had not been reported as rheumatoid arthritis susceptibility loci[1–16], we selected a marker SNP for the replication study (Online Methods and **Supplementary Table 3**).

In the combined analyses of the meta-analysis and the replication study, including a total of 9,351 rheumatoid arthritis cases and 38,575 controls, we identified six newly associated loci, in addition to the *NFKBIE*, *ARID5B* and *PTPN2* loci, that satisfied the significance threshold of $P < 5.0 \times 10^{-8}$, including *B3GNT2* at 2p15, *ANXA3* at 4q21, *CSF2* at 5q31, *CD83* at 6p23, *PDE2A-ARAP1* at 11q13 and *PLD4* at 14q32 (**Figs. 1** and **2** and **Table 1**). Of these loci, *NFKBIE* had the smallest *P* value (5.8×10^{-19}). Although association with rheumatoid arthritis has been described for the *CSF2* and *PTPN2* loci[11,15,16,20,21], ours is the first report to our knowledge validating these associations with a threshold of $P < 5.0 \times 10^{-8}$. Suggestive associations were also observed in *ETS1-FLI1* at 11q24, *GCH1* at 14q22, *PRKCH* at 14q23, *ZNF774* at 15q26, *PRKCB1* at 16p12 and *IRF8* at 16q24 ($5.0 \times 10^{-8} \leq P < 5.0 \times 10^{-6}$). A summary of the genes in the newly associated loci and the results of *cis* expression quantitative trait locus (*cis* eQTL) analysis of the marker SNPs are provided (**Supplementary Table 4** and **Supplementary Note**).

Previous studies have reported associations of rheumatoid arthritis susceptibility loci with other autoimmune diseases[4,10,15,16]. Therefore, we assessed the association of these newly identified susceptibility loci with systemic lupus erythematosus (SLE) by examining the results of an SLE GWAS in the Japanese population (891 cases and 3,384 controls)[22] and in Graves' disease by genotyping 1,783 cases[10] (the controls from the SLE analysis were used for testing for Graves' disease). We observed significant associations of the *ANXA3* locus with SLE and of the *B3GNT2* and *ARID5B* loci with Graves' disease, which showed the same directional effects of the alleles as in rheumatoid arthritis ($P < 0.05/9 = 0.0056$, Bonferroni correction of the number of loci; **Supplementary Table 5**). It should be noted that relatively small sample sizes in the SLE and Graves' disease cohorts might yield limited statistical power, and further evaluations enrolling larger numbers of subjects would be desirable.

To highlight genetic backgrounds of rheumatoid arthritis that are common and divergent in different ancestry groups, we conducted a multi-ancestry comparative analysis of the present study in Japanese and a previous GWAS meta-analysis in Europeans that included 5,539 rheumatoid arthritis cases and 20,169 controls[15] (**Fig. 3a–c**). First, we compared associations in the reported[1–16] or newly identified rheumatoid arthritis susceptibility loci (**Fig. 3a** and **Supplementary Table 6**). Of the 46 rheumatoid arthritis risk variants evaluated, 6 were monomorphic in Japanese, and all were polymorphic in Europeans. We observed significant associations at 22 loci in Japanese and at 36 loci in Europeans (false discovery rate (FDR) < 0.05, $P < 0.0030$), with 14 loci being shared between the populations. Of the newly associated rheumatoid arthritis susceptibility loci identified in our Japanese meta-analysis, significant associations were also observed in the European meta-analysis at the *ARID5B* and *PTPN2* loci ($P = 1.9 \times 10^{-4}$ and 9.2×10^{-5}, respectively; **Table 1**). Significant positive correlation of odds ratios was observed between the studies ($r = 0.822$, $P = 8.1 \times 10^{-11}$; **Fig. 3b**), suggesting that a substantial proportion of genetic factors are shared between

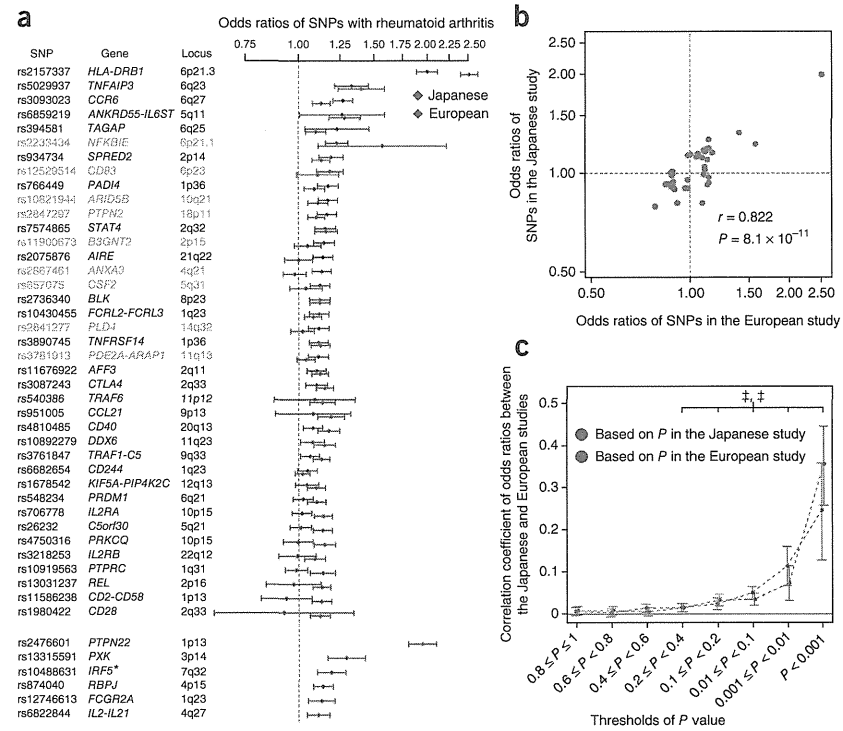

Figure 3 Overlap of the associations with rheumatoid arthritis between Japanese and European populations. (**a**) Forest plots of SNPs in the rheumatoid arthritis susceptibility loci (**Supplementary Table 6**). We selected the genetic loci that have been validated to be associated with rheumatoid arthritis susceptibility by showing associations in the reports of multiple cohorts or satisfying the genome-wide significant threshold ($P < 5.0 \times 10^{-8}$) in previous studies, including in the meta-analysis and replication phases[1–16]. For each of the loci, the most significant SNP among those reported in the previous or present study were selected[1–16]. SNPs in the newly identified rheumatoid arthritis susceptibility loci are colored green. Odds ratios and 95% confidence interval (CI) values are based on rheumatoid arthritis risk alleles, and the SNPs are ordered according to the odds ratios in the Japanese study. Several SNPs were monomorphic in the Japanese population. The odds ratios of these SNPs in the European study are presented below. The asterisk indicates that an association of another variant at the *IRF5* locus was reported in the Japanese population[24]. (**b**) Correlation of the odds ratios of the SNPs in the validated rheumatoid arthritis susceptibility loci between the two populations. SNPs that were polymorphic in both populations were used; odds ratios were based on the minor allele in the Japanese population. (**c**) Correlation of the odds ratios of the genome-wide SNPs, excluding the rheumatoid arthritis susceptibility loci. Correlations were evaluated for sets of SNPs stratified by the thresholds based on the meta-analysis *P* values in each population after pruning of the SNPs by LD ($r^2 < 0.3$). Correlation coefficient and 95% CI are indicated on the *y* axis. Significant correlation of the odds ratios was observed (‡, $P < 0.005$), even for the SNPs that showed moderate associations with rheumatoid arthritis (meta-analysis $P < 0.4$ in each population).

the two ancestry groups[17]. When the rheumatoid arthritis cases of the Japanese GWAS meta-analysis were stratified into anti-CCP–positive or rheumatoid factor–positive cases ($n = 3{,}209$) and controls ($n = 16{,}891$), similar results were observed (data not shown). Nevertheless, most of the SNPs assessed here are not necessarily causal variants, and further fine mapping of the loci is warranted to precisely evaluate the shared genetic predisposition between the populations.

Next, we compared regional associations within each of the loci and identified unique patterns in the *ARID5B* locus at 10q21 (**Supplementary Fig. 3**). In Japanese, three peaks of association were observed ($P = 1.0 \times 10^{-8}$ at rs10821944, $P = 5.7 \times 10^{-8}$ at rs10740069 and $P = 8.5 \times 10^{-6}$ at rs224311). These three variants were in weak linkage disequilibrium (LD) in Japanese ($r^2 < 0.10$), indicating independent associations with each of the other SNPs that satisfied a region-wide significance threshold of $P < 3.5 \times 10^{-5}$ (conditional $P = 4.3 \times 10^{-6}$, 1.7×10^{-5} and 1.8×10^{-5}, respectively) (**Supplementary Fig. 3**). In contrast, there was only one peak of association in Europeans ($P = 1.2 \times 10^{-6}$ at rs12764378; $r^2 = 0.59$ with rs10821944 in Europeans), and no additional association was observed in conditional analysis with rs12764378 (the smallest conditional $P = 2.2 \times 10^{-4}$), suggesting that the number of independent associations may be different at this locus in the two populations.

Finally, we conducted polygenic assessment for common variants showing modest associations to rheumatoid arthritis (those not meeting the genome-wide association threshold). This approach has been recognized to be a means to explain a substantial proportion of genetic risk[23]. For the SNPs that were shared between the two meta-analyses but not included in the validated rheumatoid arthritis susceptibility loci, we adopted LD pruning of the SNPs ($r^2 < 0.3$). We then evaluated the correlation of odds ratios of the SNPs between the two meta-analyses and observed a significant positive correlation ($r = 0.023$, $P < 1.0 \times 10^{-300}$). When the SNPs were stratified according to the *P* values in each meta-analysis, significant positive correlations of odds ratios were observed for the SNPs, even for those showing modest association ($P < 0.4$ in the meta-analysis of Japanese or Europeans; $r = 0.014$–0.36 for each *P* value range, $P < 0.005$ for each correlation test) (**Fig. 3c**). Correlations (r) of odds ratios observed herein suggest substantial overlap of the genetic risk of rheumatoid arthritis between the two populations, not only in the validated rheumatoid arthritis susceptibility loci but also at the loci showing nonsignificant associations. This suggests the usefulness of a meta-analysis approach involving multiple ancestry groups in identifying additional susceptibility loci.

In summary, we identified multiple new loci associated with rheumatoid arthritis through a large-scale meta-analysis of GWAS in Japanese. Multi-ancestry comparative analysis provided evidence of significant overlap in the genetic risks of rheumatoid arthritis between Japanese and Europeans. Thus, findings from the present study should contribute to the further understanding of the etiology of rheumatoid arthritis.

URLs. GARNET consortium, http://www.twmu.ac.jp/IOR/garnet/home.html; The BioBank Japan Project (in Japanese), http://biobankjp.org/; International HapMap Project, http://www.hapmap.org/; PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/; EIGENSTRAT, http://genepath.med.harvard.edu/~reich/Software.htm; MACH and mach2dat, http://www.sph.umich.edu/csg/abecasis/MACH/index.

html; R statistical software, http://cran.r-project.org/; SNAP, http://www.broadinstitute.org/mpg/snap/index.php; NCBI GEO database, http://www.ncbi.nlm.nih.gov/geo/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

The authors acknowledge the essential role of the GARNET consortium in developing the study. In this study, the following GARNET members are included: CGM of RIKEN, University of Tokyo, the BioBank Japan Project, Kyoto University and IORRA. We would like to thank all the doctors and staff who participated in sample collection for the RIKEN cohort and the BioBank Japan Project. We thank K. Kobayashi and M. Kitazato for their technical assistance. We thank T. Raj for calculation of composite of multiple signals (CMS). We thank M. Kokubo for DNA extraction, GWAS genotyping and secretarial assistance. We would also like to thank H. Yoshifuji, N. Yukawa, D. Kawabata, T. Nojima, T. Usui and T. Fujii for collecting DNA samples. We thank Y. Katagiri for her technical efforts. We also appreciate the contribution of E. Inoue and other members of the Institute of Rheumatology, Tokyo Women's Medical University, for their efforts on the IORRA cohort. This study was supported in part by grants-in-aid from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) in Japan, the Ministry of Health, Labour and Welfare (MHLW) in Japan, the Japan Society for the Promotion of Science (JSPS), Core Research for Evolutional Science and Technology (CREST), Solution-Oriented Research for Science and Technology (SORST), INSERM and the Okawa Foundation for Information and Telecommunications.

AUTHOR CONTRIBUTIONS

Y. Okada, C.T., K.I., Y. Kochi and K.O. designed the study and drafted the manuscript. Y. Okada, C.T., K.I., T.K., H.O., N.N., M.T., M.L., K. Tokunaga and M.K. managed genotyping and manipulation of GWAS data. Y. Okada, Y. Kochi, C.T. and K.I. managed genotyping of replication cohorts. Y. Okada, T.K., H.O., E.A.S., A. Takahashi and R.Y. performed statistical analysis. Y. Kochi, A.S., K. Myouzen, T. Sawada, Y. Nishoka, M.Y., T. Matsubara, S.W., R.T. and S.T. collected samples and managed phenotype data for the rheumatoid arthritis cohorts from the BioBank Japan Project and CGM, RIKEN. C.T., K.O., T.K., M.T., K. Takasugi, K.S., A.M., S.H., K. Matsuo, H. Tanaka, K. Tajima and M.L. collected samples and managed phenotype data for the rheumatoid arthritis cohorts from Kyoto University. K.I., T. Suzuki, T.I., Y. Kawamura, H. Tanii, Y. Okazaki and T. Sakaki collected samples and managed phenotype data for the rheumatoid arthritis cohorts from IORRA. Y. Kochi managed the data for the SLE and Graves' disease cohorts. A.S., C.T. and K.I. analyzed the sera of subjects with rheumatoid arthritis. E.A.S., F.A.S.K., P.K.G., J.W., K.A.S., L.P. and R.M.P. managed the data for the rheumatoid arthritis cohorts in European populations. A. Taniguchi, A. Takahashi, K. Tokunaga, M.K., Y. Nakamura, N.K., T. Minori, R.M.P., H.Y., S.M., R.Y., F.M. and K.Y. supervised the overall study.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Suzuki, A. *et al.* Functional haplotypes of *PADI4*, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. *Nat. Genet.* **34**, 395–402 (2003).
2. Kochi, Y. *et al.* A functional variant in *FCRL3*, encoding Fc receptor–like 3, is associated with rheumatoid arthritis and several autoimmunities. *Nat. Genet.* **37**, 478–485 (2005).
3. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
4. Remmers, E.F. *et al. STAT4* and the risk of rheumatoid arthritis and systemic lupus erythematosus. *N. Engl. J. Med.* **357**, 977–986 (2007).
5. Plenge, R.M. *et al. TRAF1–C5* as a risk locus for rheumatoid arthritis—a genomewide study. *N. Engl. J. Med.* **357**, 1199–1209 (2007).
6. Plenge, R.M. *et al.* Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nat. Genet.* **39**, 1477–1482 (2007).
7. Barton, A. *et al.* Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13. *Nat. Genet.* **40**, 1156–1159 (2008).
8. Suzuki, A. *et al.* Functional SNPs in *CD244* increase the risk of rheumatoid arthritis in a Japanese population. *Nat. Genet.* **40**, 1224–1229 (2008).
9. Gregersen, P.K. *et al. REL*, encoding a member of the NF-κB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. *Nat. Genet.* **41**, 820–823 (2009).
10. Kochi, Y. *et al.* A regulatory variant in *CCR6* is associated with rheumatoid arthritis susceptibility. *Nat. Genet.* **42**, 515–519 (2010).
11. Freudenberg, J. *et al.* Genome-wide association study of rheumatoid arthritis in Koreans: population-specific loci as well as overlap with European susceptibility loci. *Arthritis Rheum.* **63**, 884–893 (2011).
12. Terao, C. *et al.* The human *AIRE* gene at chromosome 21q22 is a genetic determinant for the predisposition to rheumatoid arthritis in Japanese population. *Hum. Mol. Genet.* **20**, 2680–2685 (2011).
13. Raychaudhuri, S. *et al.* Common variants at *CD40* and other loci confer risk of rheumatoid arthritis. *Nat. Genet.* **40**, 1216–1223 (2008).
14. Raychaudhuri, S. *et al.* Genetic variants at *CD28*, *PRDM1* and *CD2/CD58* are associated with rheumatoid arthritis risk. *Nat. Genet.* **41**, 1313–1318 (2009).
15. Stahl, E.A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.* **42**, 508–514 (2010).
16. Zhernakova, A. *et al.* Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS Genet.* **7**, e1002004 (2011).
17. Kurreeman, F. *et al.* Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. *Am. J. Hum. Genet.* **88**, 57–69 (2011).
18. Nakamura, Y. The BioBank Japan Project. *Clin. Adv. Hematol. Oncol.* **5**, 696–697 (2007).
19. Yamanaka, H. *et al.* Influence of methotrexate dose on its efficacy and safety in rheumatoid arthritis patients: evidence based on the variety of prescribing approaches among practicing Japanese rheumatologists in a single institute-based large observational cohort (IORRA). *Mod. Rheumatol.* **17**, 98–105 (2007).
20. Yamada, R. *et al.* Association between a single-nucleotide polymorphism in the promoter of the human interleukin-3 gene and rheumatoid arthritis in Japanese patients, and maximum-likelihood estimation of combinatorial effect that two genetic loci have on susceptibility to the disease. *Am. J. Hum. Genet.* **68**, 674–685 (2001).
21. Tokuhiro, S. *et al.* An intronic SNP in a RUNX1 binding site of *SLC22A4*, encoding an organic cation transporter, is associated with rheumatoid arthritis. *Nat. Genet.* **35**, 341–348 (2003).
22. Okada, Y. *et al.* A genome-wide association study identified *AFF1* as a susceptibility locus for systemic lupus eyrthematosus in Japanese. *PLoS Genet.* **8**, e1002455 (2012).
23. Stranger, B.E., Stahl, E.A. & Raj, T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* **187**, 367–383 (2011).
24. Shimane, K. *et al.* A single nucleotide polymorphism in the *IRF5* promoter region is associated with susceptibility to rheumatoid arthritis in the Japanese patients. *Ann. Rheum. Dis.* **68**, 377–383 (2009).

[1]Laboratory for Autoimmune Diseases, Center for Genomic Medicine (CGM), RIKEN, Yokohama, Japan. [2]Department of Allergy and Rheumatology, Graduate School of Medicine, University of Tokyo, Tokyo, Japan. [3]Laboratory for Statistical Analysis, CGM, RIKEN, Yokohama, Japan. [4]Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan. [5]Department of Rheumatology and Clinical Immunology, Graduate School of Medicine, Kyoto University, Kyoto, Japan. [6]Institute of Rheumatology, Tokyo Women's Medical University, Tokyo, Japan. [7]Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [8]Broad Institute, Cambridge, Massachusetts, USA. [9]Department of Rheumatology, Leiden University Medical Center, Leiden, The Netherlands. [10]Department of Human Genetics, Graduate School of Medicine, University of Tokyo, Tokyo, Japan. [11]Department of Rheumatology, Tokyo Medical University Hospital, Tokyo, Japan. [12]Yamanashi Prefectural Central Hospital, Yamanashi, Japan. [13]Department of Orthopaedic Surgery, Yukioka Hospital, Osaka, Japan. [14]Matsubara Mayflower Hospital, Hyogo, Japan. [15]Osaka Minami National Hospital, Osaka, Japan. [16]Department of Orthopedic Surgery, Tottori University, Tottori, Japan. [17]Department of Rheumatology, National Hospital Organization, Sagamihara Hospital, Kanagawa, Japan. [18]Center for Rheumatic Diseases, Dohgo Spa Hospital, Ehime, Japan. [19]Department of Rheumatology, Niigata Rheumatic Center, Niigata, Japan. [20]Saiseikai Takaoka Hospital, Toyama, Japan. [21]Division of Epidemiology and Prevention, Aichi Cancer Center Research Institute, Aichi, Japan. [22]Aichi Cancer Center Hospital and Research Institute, Aichi, Japan. [23]Department of Orthopaedic Surgery, Keio University, Tokyo, Japan. [24]Yokohama Clinic, Warakukai Medical Corporation, Yokohama, Japan. [25]Department of Psychiatry, Mie University School of Medicine, Mie, Japan. [26]Metropolitan Matsuzawa Hospital, Tokyo, Japan. [27]Graduate School of Education, University of Tokyo, Tokyo, Japan. [28]The Feinstein Institute for Medical Research, North Shore–Long Island Jewish Health System, Manhasset, New York, USA. [29]Rheumatology Unit,

Department of Medicine in Solna, Karolinska Institutet and Karolinska University Hospital, Stockholm, Sweden. [30]Arthritis Research Campaign–Epidemiology Unit, The University of Manchester, Manchester Academic Health Science Centre, Manchester, UK. [31]Division of Medicine, University of Toronto, Mount Sinai Hospital and University Health Network, Toronto, Ontario, Canada. [32]Commisariat a l'Energie Atomique (CEA), Institut Genomique, Centre National de Genotypage, Evry, France. [33]Fondation Jean Dausset, Centre d'Etude du Polymorphisme Humain, Paris, France. [34]Laboratory for Genotyping Development, CGM, RIKEN, Yokohama, Japan. [35]Laboratory of Molecular Medicine, Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo, Japan. [36]Laboratory for International Alliance, CGM, RIKEN, Yokohama, Japan. [37]Unit of Statistical Genetics, Center for Genomic Medicine Graduate School of Medicine Kyoto University, Kyoto, Japan. [38]Core Research for Evolutional Science and Technology (CREST) Program, Japan Science and Technology Agency, Kawaguchi, Japan. [39]Institut National de la Santé et de la Recherche Médicale (INSERM), Unité U852, Kyoto University Graduate School of Medicine, Kyoto, Japan. [40]These authors contributed equally to this work. [41]These authors jointly directed this work. Correspondence should be addressed to Y.K. (ykochi@src.riken.jp) or K.O. (ohmurako@kuhp.kyoto-u.ac.jp).

ONLINE METHODS

Subjects. The Japanese participants in the meta-analysis (4,074 rheumatoid arthritis cases and 16,891 controls) and the replication study (5,277 rheumatoid arthritis cases and 21,684 controls) were obtained through the collaborations of the GARNET consortium (**Supplementary Table 1**)[10,12]. The meta-analysis was conducted on three independent GWAS (from the BioBank Japan Project[18] with 2,414 rheumatoid arthritis cases and 14,245 controls[10], Kyoto University with 1,237 rheumatoid arthritis cases and 2,087 controls[12] and IORRA[19] with 423 rheumatoid arthritis cases and 559 controls). The replication study consisted of two independent cohorts (cohort 1 included 3,830 rheumatoid arthritis cases and 17,920 controls, and cohort 2 included 1,447 rheumatoid arthritis cases and 3,764 controls). We employed a case-control cohort of SLE (891 cases and 3,384 controls)[22] and 1,783 cases with Graves' disease[10]. Details of 5,539 rheumatoid arthritis cases and 20,169 controls included in the meta-analysis in European populations were described elsewhere[15]. All participants provided written informed consent for participation in the study, as approved by the ethical committees of the institutional review boards. Detailed descriptions of the participating subjects are provided (**Supplementary Note**).

Genotyping and quality control in the GWAS. Genotyping platforms and quality control criteria for the GWAS, including cutoff values for sample call rates, SNP call rates, MAF and Hardy-Weinberg *P* values, are given (**Supplementary Table 2**). For the subjects enrolled in each of three GWAS, we excluded closely related subjects with first- or second-degree kinship, which was estimated using PLINK version 1.06 (see URLs). We also excluded the subjects determined to be ancestry outliers from East Asian populations using PCA performed by EIGENSTRAT version 2.0 (see URLs) along with HapMap Phase 2 panels (release 24; **Supplementary Fig. 1**). Genotype imputation was performed on the basis of the HapMap Phase 2 East Asian populations, using MACH version 1.0.16 (see URLs) in a two-step procedure as described elsewhere[25]. We excluded imputed SNPs with MAF < 0.01 or *Rsq* < 0.5 from each of the GWAS. Associations of the SNPs with rheumatoid arthritis were assessed by logistic regression models assuming additive effects of the allele dosages of the SNPs using mach2dat software (see URLs).

Meta-analysis. We included 1,948,139 autosomal SNPs that satisfied quality control criteria in all three GWAS (**Supplementary Table 2**). SNP information was based on a forward strand of the NCBI build 36.3 reference sequence. The meta-analysis was performed using an inverse variance method assuming a fixed-effects model from the study-specific effect sizes (logarithm of odds ratio) and the standard errors of the coded alleles of the SNPs determined with the Java source code implemented by the authors[25]. Genomic control corrections[26] were carried out on test statistics of the GWAS using the study-specific inflation factor (λ_{GC}) and was applied or reapplied to the results of our current meta-analysis (**Supplementary Fig. 2**).

Replication study. We selected a SNP for the replication study from each of the loci that exhibited $P < 5.0 \times 10^{-4}$ in the meta-analysis that had not previously been reported as rheumatoid arthritis susceptibility loci[1–16] (**Supplementary Table 3**). For control subjects, we used genotype data obtained from additional GWAS for non-autoimmune diseases or healthy controls, genotyped using Illumina HumanHap550 BeadChips or HumanHap610-Quad BeadChips, and the cases for rheumatoid arthritis and Graves' disease were genotyped with the TaqMan genotyping system (Applied Biosystems; **Supplementary Table 1**). Selection of the SNP was conducted according to the following criteria: if the SNP with the most significant association in the locus was genotyped in the replication control panel, then that SNP was selected; otherwise, a tag SNP in the replication control panel with the strongest LD was selected (mean $r^2 = 0.89$). For the three SNPs that yielded low call rates (<90%), we alternatively selected proxy SNPs with the second strongest LD. As a result, average genotyping call rates of the SNPs were 99.9% and 99.0% for the controls and cases, respectively. We then evaluated concordance rates between the assayed genotypes by applying these two different methods to samples from 376 subjects who were randomly selected. This procedure yielded high concordance rates of ≥99.9%. Associations of the SNPs were evaluated using logistic regression assuming an additive-effects model of genotypes in R statistical software version 2.11.0 (see URLs). The combined study of the meta-analysis and replication study was performed using an inverse variance method assuming a fixed-effects model[25].

***Cis* eQTL analysis.** For each marker SNP of the newly identified rheumatoid arthritis susceptibility locus, correlations between SNP genotypes and expression levels of genes located 300 kb upstream or downstream of the SNP measured in B-lymphoblastoid cell lines (GSE6536) were evaluated using data from the HapMap Phase 2 east Asian populations[27].

Multi-ancestry analysis of the meta-analyses in Japanese and Europeans. We evaluated the associations of the variants in the validated rheumatoid arthritis susceptibility loci by comparing the results from the current meta-analysis in Japanese with those from a previous meta-analysis in Europeans[15]. We assessed two variants in the *IRF5* locus, where different causal variants were identified in the two populations[24]. For the conditional analysis of the regional associations in the *ARID5B* locus (**Supplementary Fig. 3**), we repeated the meta-analysis at that locus by incorporating genotypes of the referenced SNP(s) as additional covariate(s). For comparison of the odds ratios of the SNPs, we first selected SNPs that were shared between the meta-analyses in Japanese and Europeans. Next, we removed the SNPs located more than 1 Mb away from each of the marker SNPs in the validated rheumatoid arthritis susceptibility loci, except for in the HLA region, where we removed the SNPs located between 24,000,000 bp to 36,000,000 bp on chromosome 6 because of the existence of long-range haplotypes with rheumatoid arthritis susceptibility in this region[28]. LD pruning of the SNPs was conducted for the SNP pairs that were in LD ($r^2 \geq 0.3$) in both HapMap Phase 2 East Asian and Utah residents of Northern and Western European ancestry (CEU) populations (release 24). Correlations of the odds ratios were evaluated using R statistical software version 2.11.0.

25. Okada, Y. *et al.* Identification of nine novel loci associated with white blood cell subtypes in a Japanese population. *PLoS Genet.* **7**, e1002067 (2011).
26. de Bakker, P.I. *et al.* Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **17**, R122–R128 (2008).
27. Stranger, B.E. *et al.* Population genomics of human gene expression. *Nat. Genet.* **39**, 1217–1224 (2007).
28. Okada, Y. *et al.* Contribution of a haplotype in the HLA region to anti–cyclic citrullinated peptide antibody positivity in rheumatoid arthritis, independently of HLA-DRB1. *Arthritis Rheum.* **60**, 3582–3590 (2009).

Nonimmunoglobulin target loci of activation-induced cytidine deaminase (AID) share unique features with immunoglobulin genes

Lucia Kato[a], Nasim A. Begum[a], A. Maxwell Burroughs[b], Tomomitsu Doi[a,1], Jun Kawai[b], Carsten O. Daub[b], Takahisa Kawaguchi[c], Fumihiko Matsuda[c], Yoshihide Hayashizaki[b], and Tasuku Honjo[a,2]

[a]Department of Immunology and Genomic Medicine and [c]The Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto 606-8501, Japan; and [b]RIKEN Omics Science Center (OSC), RIKEN Yokohama Institute, Yokohama, Kanagawa 230-0045, Japan

Contributed by Tasuku Honjo, December 28, 2011 (sent for review December 5, 2011)

Activation-induced cytidine deaminase (AID) is required for both somatic hypermutation and class-switch recombination in activated B cells. AID is also known to target nonimmunoglobulin genes and introduce mutations or chromosomal translocations, eventually causing tumors. To identify as-yet-unknown AID targets, we screened early AID-induced DNA breaks by using two independent genome-wide approaches. Along with known AID targets, this screen identified a set of unique genes (*SNHG3, MALAT1, BCL7A,* and *CUX1*) and confirmed that these loci accumulated mutations as frequently as *Ig* locus after AID activation. Moreover, these genes share three important characteristics with the Ig gene: translocations in tumors, repetitive sequences, and the epigenetic modification of chromatin by H3K4 trimethylation in the vicinity of cleavage sites.

deep sequencing | end labeling by biotin oligonucleotide | microarray

Activation-induced cytidine deaminase (AID) is expressed in germinal center (GC) B cells upon antigen stimulation and is essential for two types of genetic alteration in the Ig gene: class switch recombination (CSR) and somatic hypermutation (SHM), which provide the genetic basis for antibody memory (1, 2). CSR produces antibodies with different effector functions by recombination at Ig heavy chain (H) switch (S) regions, so that the μ-chain constant (Cμ) region is replaced by a downstream C_H region. SHM introduces nontemplated point mutations in the rearranged variable (V) region genes, resulting in incremented antigen receptor affinity after clonal selection (3, 4).

Functional studies on AID mutants have shown that distinct AID domains are required for SHM and CSR, although AID has a single catalytic center (cytidine deaminase motif) in the middle of the molecule. Deletions and alterations in the N-terminal region affect both the CSR and SHM activities (5). However, AID C-terminal mutants almost completely lose CSR activity but retain or even increase SHM activity (6, 7). Although C-terminally truncated AID mutants cleave both V and S regions and induce enhanced c-myc-IgH translocations, they cannot mediate CSR, suggesting that the C-terminal domain is not required for DNA cleavage but is required to correctly pair cleaved ends (8).

The DNA cleavage of targets in CSR and SHM (the S region and V region, respectively) requires their transcription (9–12). Indeed, AID-induced mutations (SHM) are generally detected in a region within 2 kb downstream of the transcription start site (TSS) (13, 14). Transcription appears to play two roles in the targeting of cleavage sites. First, transcription is associated with the epigenetic marking of the target locus, particularly by H3K4 trimethylation (H3K4me3). The histone chaperone complex FACT is required to regulate H3K4me3 in the target S region, and FACT knockdown abolishes H3K4me3 and DNA cleavage in this region (15). Second, transcription is probably required to induce non-B structures in highly repetitive sequences such as S regions (16–18), due to excessive negative supercoiling induced immediately downstream of transcription. V regions have also been shown to form stem-loop structures under these conditions (19, 20). Non-B structure involvement has recently been reported in transcription-associated mutations in repetitive sequences such as the dinucleotide repeat hot spots or triplet repeat expansion/contractions causing Huntington's disease (17, 21, 22).

AID-dependent DNA cleavage is, in general, specific to the Ig locus. However, a number of reports have shown that AID can induce DNA cleavage in non-Ig loci. AID non-Ig targets were first demonstrated by studies on AID transgenic mice that produce numerous T lymphomas, in which vast numbers of mutations accumulate in the genes encoding the T-cell receptor, CD4, CD5, c-myc, and PIM1 (23, 24). This finding was followed by the observations that AID deficiency abolishes c-myc-Ig translocation and reduces the incidence of plasmacytoma (25, 26). AID expression is specific to activated B cells under normal conditions. However, AID expression has also been found in non-B cells, especially in cells stimulated by infection with pathogens such as human T-cell leukemia virus type 1 (HTLV1), hepatitis C virus (HCV), Epstein–Barr (EB) virus, and *Helicobacter pylori* (27–30). Based on these observations, AID is postulated to induce tumorigenesis, especially in B lymphomas and leukemias—and AID is expressed in many GC-derived human B-cell lymphomas (31–33). The prognosis of acute lymphocytic leukemia (ALL) and chronic myeloid leukemia (CML) is linked with AID expression (34, 35). It is therefore important to determine which non-Ig genes can be targeted by AID, and what features, if any, they share with Ig genes.

Several approaches have been used to explore AID non-Ig target genes in B cells. Candidate approaches involving the direct sequencing of proto-oncogenes, genes involved in translocations, or genes transcribed in normal GC B cells have shown that AID mutates several non-Ig genes, including *BCL6*, *MYC*, *PIM1*, and *PAX5* (24, 32, 36, 37). More recently, several efforts have been made to identify AID targets in a whole genome. These approaches have used chromatin immunoprecipitation (ChIP) of CSR-related proteins in combination with genome-wide tiling microarrays (ChIP-chip) or deep sequencing (ChIP-seq) on the assumption that proteins involved in CSR bind to AID targets. RPA, Nbs1, AID itself, and Spt5 have been used as marking proteins in this type of study (38–40). However, these approaches did not necessarily show that all of the protein-bound targets are cleaved or mutated by AID. There are indications that some genes identified by such approaches are not tran-

Author contributions: L.K., T.D., and T.H. designed research; L.K., N.A.B., and A.M.B. performed research; T.K. and F.M. contributed new reagents/analytic tools; L.K., N.A.B., A.M.B., J.K., C.O.D., and Y.H. analyzed data; and L.K. and T.H. wrote the paper.

The authors declare no conflict of interest.

[1]Present address: Laboratory Animal Research Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan.

[2]To whom correspondence should be addressed. E-mail: honjo@mfour.med.kyoto-u.ac.jp.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1120791109/-/DCSupplemental.

scribed (39). Therefore, it is important to reexamine non-Ig AID target genes by using a different strategy.

Here, we report four AID targets, identified by a combination of unique techniques. After directly labeling the DNA breakage ends from AID-induced cleavage with a biotinylated linker, we isolated the labeled fragments with streptavidin beads and analyzed them by a combination of promoter arrays and genome-wide sequencing. The candidates identified were then confirmed by quantitative PCR (qPCR) and the actual demonstration of mutations. With these methods, we identified at least four previously unknown AID targets—*SNHG3, MALAT1, BCL7A,* and *CUX1*. We found that these targets share important characteristics with Ig genes, namely, repetitive sequences that can form non-B structures upon efficient transcription, and the accumulation of H3K4me3 histone modifications on the chromatin.

Results

AID-Induced DNA Cleavage Detected by Labeling DNA Break Ends with a Biotinylated Linker. To detect genome-wide AID-induced DNA breaks, we used a modified in situ DNA end-labeling technique as described (8, 41) in BL2 cells, a Burkitt's lymphoma cell line that serves as an in vitro model for studying the SHM mechanism (31, 42, 43). We used the BL2 clone BL2-ΔC-AIDER, which expresses JP8Bdel, an AID mutant lacking the C-terminal 16 residues, fused with the hormone-binding domain of the estrogen receptor (ER) (JP8Bdel-ER). Tamoxifen (4-OHT) treatment induces DNA breakage in the Sμ and Sα regions but not in the Sγ region of JP8Bdel-ER-expressing CH12 cells, which switch almost exclusively from IgM to IgA (8).

BL2-ΔC-AIDER cells were treated with 4-OHT only for 3 h to minimize cell death and DNA break ends were labeled with a biotinylated linker, and the break-enriched biotinylated DNA was used as a PCR template (Fig. 1*A*). In agreement with previous reports (8, 42), we detected DNA breakage in the 5′ Sμ region of the IgH locus only in 4-OHT–treated cells. No breakage was detected in the *B2M* gene, which is expressed in BL2 cells but was shown not to accumulate mutations in activated B cells (Fig. 1*B*).

AID Targets Identified by Promoter Array and Whole Genome Sequencing. Because SHM is normally detected close to the TSS (13, 14), biotin linker-enriched DNA fragments were analyzed by a promoter array to identify unknown AID targets. Table S1 lists the genes whose signals increased after 3 h of 4-OHT treatment, compared with untreated samples with false discovery rate (FDR) values <0.3. We also looked for genes with increased signals after 4-OHT treatment that are known to be targets of chromosomal translocation or genes that had multiple breakage peaks, and we identified >50 genes, among which we found that *BCL7A* and *CUX1* are enriched in the original breakage-enriched library by qPCR (see below). We confirmed by RT-PCR and expression array that *SNHG3*, *MALAT1*, *NIN*, *C9orf72*, *CFLAR*, *SNX25*, *BCL7A,* and *CUX1* were transcribed in BL2 cells (Table S1). Fig. S1 shows the peak signals in a 10-kb segment surrounding the breakage area of *SNHG3, MALAT1, BCL7A,* and *CUX1*. We could not map the breakage in the Ig locus because of the absence of array probes in this region.

Because the promoter array does not detect DNA fragments outside of regions containing probes, we further analyzed the breakage-enriched DNA by direct sequencing of the biotin linker-enriched library. DNA breakage sites in both control and 4-OHT–treated libraries were identified by aligning sequenced tags to the genome, and significantly enriched regions were identified by comparing the local breakage density (*SI Materials and Methods*). Regions were identified in the genes listed in Table S2. Interestingly, *SNHG3* and *MALAT1*, which were identified by the promoter array, appear at the top of the list in the genome-wide sequencing as well.

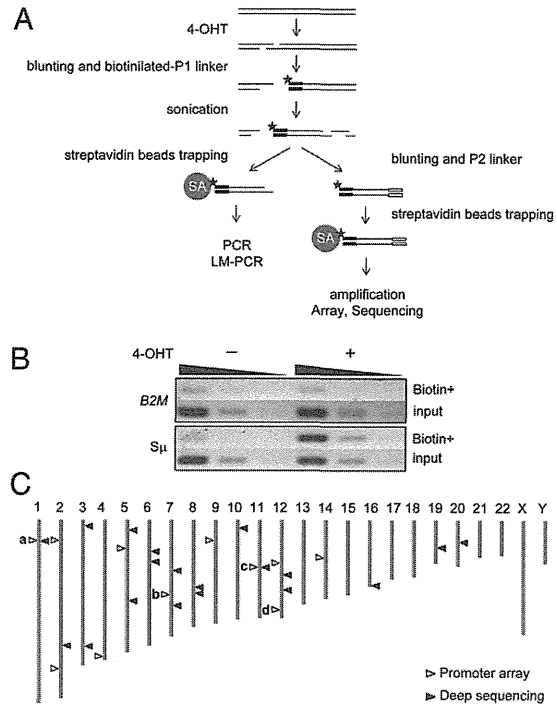

Fig. 1. (*A*) Schematic of the labeling technique. 4-OHT is added to activate AID, and DNA break ends are labeled in situ by biotinylated linker ligation. After genomic DNA is extracted and sonicated, biotinylated fragments are captured by streptavidin beads and used for PCR, array, or sequencing. (*B*) Detection of DNA breaks by PCR. BL2-ΔC-AIDER cells were treated with or without 4-OHT for 3 h, and the break ends were labeled. PCR of Sμ and *B2M* was performed with biotin-labeled DNA or input DNA by using fivefold serially diluted templates. (*C*) Chromosomal distribution of AID targets. a, *SNHG3*; b, *CUX1*; c, *MALAT1*; d, *BCL7A*. White arrowhead, promoter array (FDR < 0.3 plus *BCL7A* and *CUX1*); black arrowhead, whole genome sequencing (FDR < 0.01 and/or remarkable numbers of *P* value clusters).

Fig. 1*C* shows the chromosomal distribution of AID target candidates identified by promoter array or whole-genome sequencing. Breakage seemed to be distributed through the genome without any apparent bias. Surprisingly, of the 29 candidates identified by whole-genome sequencing with strict statistical parameters, only two matched candidates obtained from the promoter array. This discrepancy might be explained in part because most of the breakage-rich regions detected by whole genome sequencing are located in regions that do not contain promoter array probes.

Results may also be limited because of possible bias by PCR amplification of the primary library for microarray and whole-genome sequencing, which could affect the relative genome coverage. To avoid this bias, we relied on the original library and confirmed all candidates by qPCR.

qPCR Analyses of Linker Libraries. To confirm the AID-induced breakage candidates detected by the promoter array and whole-genome sequencing, we used qPCR assays with gene-specific primers to amplify the vicinity of the identified breakage regions in biotin linker-enriched DNA from cells treated with 4-OHT for 3 h (Fig. 2). We examined whether candidate genes were enriched in the 4-OHT–treated DNA library compared with the nontreated library. Among the 29 candidates identified by whole-genome sequencing, only *SNHG3* and *MALAT1* were strongly enriched ($P < 0.0001$ and $P < 0.001$, respectively). Besides these, *BCL7A*, *CUX1,* and *CFLAR*, which were picked up only by the

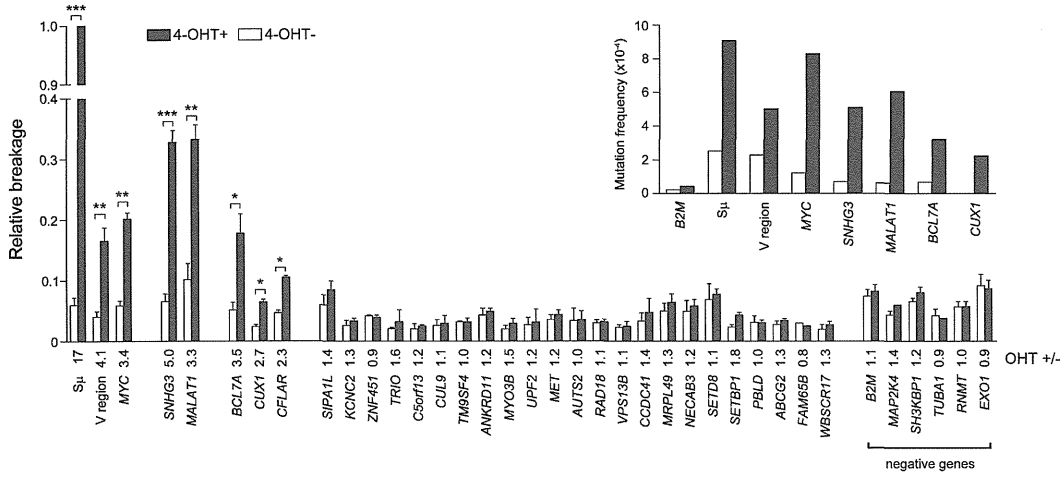

Fig. 2. qPCR measurement of DNA breaks. Break signals are presented relative to Sμ. SD values were derived from at least three independent experiments, and *P* values were calculated by a two-tailed *t* test. $*P < 0.01$, $**P < 0.001$, $***P < 0.0001$. Numbers below the *x* axis indicate the ratio between samples treated and not treated with 4-OHT. (*Inset*) Mutation analysis of genes with significantly increased break signals after AID activation. Cells were treated with or without 4-OHT for 24 h. Only unique mutations were counted. Detailed mutation profiles can be found in Fig. S2 and Table S3.

promoter array, also showed significant enrichment ($P < 0.01$) in the 4-OHT–treated library.

We also confirmed that the Sμ and V regions in BL2 cells were cleaved, because they were enriched in the 4-OHT–treated library. Although *MYC*, which is translocated in an AID-dependent manner in human Burkitt's lymphoma (44), was not identified by either promoter array or whole-genome sequencing, qPCR of the 4-OHT–treated samples clearly revealed *MYC* gene enrichment (Fig. 2). The difference in cleavage detection between the direct candidate qPCR and genome-wide arrays and sequencing suggests that the amplification step required for microarray and whole-genome sequencing methods may introduce bias, either for or against many genes. In the case of sequencing, this bias can lead to low mapping coverage of certain regions, hampering efforts to identify significant enrichment. Therefore, we cannot exclude genes that were not identified by the present methods from being AID targets.

AID Targets Accumulate Somatic Mutations near Cleavage Sites. To test whether the newly identified target genes are mutated upon AID activation, we treated BL2-ΔC-AIDER cells with 4-OHT for 24 h and sequenced regions of ≈600 bp around each area with abundant breakage (Fig. S2 and Table S3). Mutations increased in all of the qPCR-confirmed AID target genes after 4-OHT treatment (Fig. 2, *Inset*), with mutation frequencies ranging from 6.1×10^{-4} for *MALAT1* to 2.2×10^{-4} for *CUX*. These frequencies are comparable to those of the V region (5.0×10^{-4}), the Sμ region (9.1×10^{-4}), and the *MYC* gene (8.3×10^{-4}), and are far higher than that of the control *B2M* gene (4.3×10^{-5}). We also detected mutations in the *CFLAR* gene; however, the mutation frequency (9.2×10^{-5}) was not as high as other AID target genes, although mutations increased significantly in 4-OHT–treated sample ($P = 0.004$) (Table S3).

To compare the distribution profiles of mutated bases and AID-induced DNA breaks in the biotin linker-enriched DNA, we mapped the linker positions by performing ligation-mediated (LM)-PCR with the linker primer and gene-specific primers. These PCR fragments were subsequently cloned and sequenced. Break ends identified by the linker were plotted, together with mutation positions (Fig. 3 and Fig. S2). The results clearly showed that the DNA cleavage marks (biotin linker) were closely associated with mutations, indicating that the DNA cleavage sites identified are functionally relevant to SHM by AID. We used RT-PCR and expression arrays to confirm that the regions

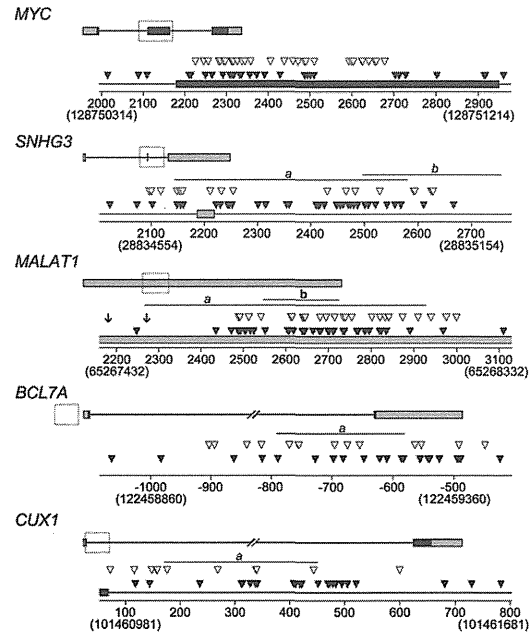

Fig. 3. Somatic mutations and breakpoint distribution in AID target loci. Mutations (open triangles) and breakpoints (filled triangles) detected by LM-PCR (Fig. S2) were plotted on the respective genomic sequences. The top scheme represents exons (rectangles) and introns (bars). Genomic loci are shown in untranslated and translated sequences (gray and black boxes, respectively). The horizontal lines *a* and *b* represent breakage regions identified by promoter array and sequencing, respectively. Regions outlined by dotted boxes are shown in more detail below each genomic locus. For the *MALAT1* locus, the translocation breakpoints reported by Davis et al. (45) are represented by arrows. *x* axis numbers indicate base positions according to RefSeq: NM_002467 (*MYC*), NR_002909 (*SNHG3*), NR_002819 (*MALAT1*), NM_020993 (*BCL7A*), and NM_181552 (*CUX1*). Numbers in parentheses indicate the corresponding base position according to hg19 assembly.

where DNA cleavage and mutations were identified are transcribed (Tables S1 and S2).

Repetitive Sequences Surround the Breakage Regions of Unique Targets. We next examined common features among the AID targets. Although SHM has been reported to prefer the RGYW-WRCY motif (46), we could not find any enrichment of this motif among the break sites in the newly identified targets. It was recently reported that mutations are introduced in regions with sequences prone to forming non-B DNA structure, including tandem repeats, palindromes, and inverted repeats (17, 18). The S region, *MYC,* and V region genes contain sequences prone to forming non-B structure (19, 20, 47, 48). We used REPFIND, a program that identifies clustered, nonrandom short repeats in a given nucleotide sequence, to search the vicinity of identified breakage regions for sequences prone to forming non-B structure. For each repeat cluster, a *P* value is calculated indicating the probability of finding such a repeat cluster randomly (a *P* value of 1×10^{-5} means that such a concentration of that particular repeat occurs an average of once in 100,000 bp by chance) (49). Curiously, we found that various types of repeat sequences cluster in the vicinity of cleaved sites in the newly identified AID target genes. In the *MALAT1* locus, the region within 2 kb surrounding the breakage peaks was rich in clustered short repeat motifs such as GAAG, GCC, GAA, CCG, AAG, GAAGA, and TTAA (Fig. 4). Repeat clusters were also found near the cleavage sites of the *SNHG3*, *BCL7A,* and *CUX1* loci. (Fig. S3). In all cases, the probability of the appearance of these repeats was far below random ($P < 1 \times 10^{-8}$).

H3K4me3 at Cleavage Sites. It was recently shown that S region transcription alone is not sufficient for CSR; specific histone posttranslational modification marks, especially H3K4me3, are required. H3K4me3 depletion strongly inhibits CSR and DNA cleavage in the Sμ and Sα regions (15). We thus asked whether the V region and the newly identified AID targets also carry H3K4me3 marks around the cleavage regions. ChIP analysis showed that both the V region and *MALAT1* locus were abundantly marked by H3K4me3 (Fig. 5). Furthermore, the H3K4me3 distribution profiles corresponded well to the somatic mutation distribution in the rearranged V region and to the breakage signal

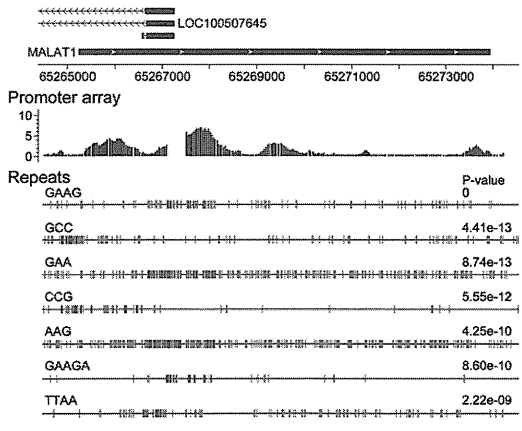

Fig. 4. Repeat sequences surrounding the breakage region in the *MALAT1* gene. (*Top*) Representation of a 10-kb segment surrounding the *MALAT1* locus. *x* axis numbers represent base positions according to hg19 assembly. (*Middle*) Breakage signal distribution detected by promoter array. Regions without bars do not have array probes. (*Bottom*) REPFIND analysis showing significant repeat clusters in the *MALAT1* locus. Motifs depicted as small, colored, vertical bars indicate the cluster with the most significant *P* value; individual repeats are separated by different colors.

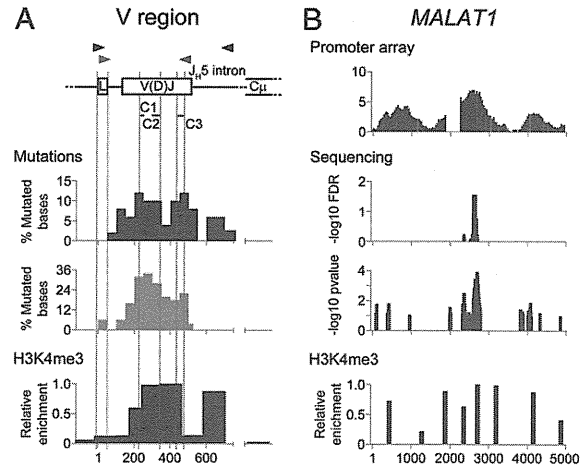

Fig. 5. H3K4me3 distribution in the IgH V region and in the *MALAT1* gene. (*A Top*: Representation of the rearranged IgH V region of BL2 cells. Black and gray arrowheads represent the position of primers used for the mutation analysis shown in *Bottom* (graphs in black and gray, respectively). L, leader; C1, CDR1; C2, CDR2; C3, CDR3. (*A Middle*) Somatic mutation distribution, represented as the percentage of mutated bases per 50 bp sequenced. Graph in black: mutations from Fig. 2, *Inset.* Graph in gray: mutations reported by Denepoux et al. (50). (*Bottom*) ChIP assay using an anti-H3K4me3 antibody. *x* axis numbers indicate the nucleotide position relative to the first V-gene ATG. (*B*) *MALAT1* locus. From top to bottom: Breakage signal distribution detected by promoter array (regions without bars do not have array probes); FDR regions by sequencing; *P* value peaks by sequencing; ChIP assay using an H3K4me3 antibody. *x* axis numbers indicate base positions according to RefSeq NR_002819.

distribution observed by both the promoter array and whole genome sequencing in *MALAT1* (Fig. 5 *A* and *B*). Mutations identified in *MALAT1* overlapped with DNA cleavage signals and H3K4me3 marks (Figs. 3 and 5*B*). We examined the H3K4me3 pattern of other AID targets by using publicly available ENCODE ChIP-seq data for the B-lymphoblastoid cell line GM12878 (51). As expected, all of them, except for *BCL7A*, were highly abundant in H3K4me3 marks overlapping nicely with cleavage sites (Fig. S4). H3K4me3 might be absent at the *BCL7A* locus in GM12878 cells because it is an inducible gene expressed in BL2 cells, but not in the GM12878 cell line (52). We thus conclude that the newly identified AID targets share both *cis* and *trans* marks for AID targeting—non-B structure and H3K4me3, respectively (15, 16).

Discussion

Identified AID Targets Accumulate High-Frequency Mutations. We explored AID targets by combining three different strategies: promoter array, whole genome sequencing, and candidate qPCR in a library containing biotinylated linker-labeled cleaved ends. With these strong criteria, we were able to identify four unique AID targets: *SNHG3, MALAT1, BCL7A,* and *CUX1*. All of these candidates were further confirmed to accumulate mutations. These candidates are thus strong AID cleavage targets; however, these genes represent only very efficient AID targets. The use of the biotinylated linker, which efficiently identifies double-strand breakage with close, staggered nicks on opposite strands, may not detect scattered nicks efficiently, and this may limit identification to targets that are efficiently and specifically cleaved within 3 h of AID activation.

Some well-described SHM target genes, including *MYC, BCL6, PAX5, RHOH,* and *PIM1,* were not detected by either the promoter array or whole genome sequencing. We used qPCR to test whether these genes were enriched in the biotin-labeled