Journal of Human Genetics (2015) 60, 319–326

www.nature.com/jhg

ORIGINAL ARTICLE

# Japan PGx Data Science Consortium Database: SNPs and HLA genotype data from 2994 Japanese healthy individuals for pharmacogenomics studies

Shigeo Kamitsuji[1], Takashi Matsuda[2], Koichi Nishimura[2], Seiko Endo[3], Chisa Wada[3], Kenji Watanabe[3], Koichi Hasegawa[4], Haretsugu Hishigaki[4], Masatoshi Masuda[4], Yusuke Kuwahara[5], Katsuki Tsuritani[5], Kenkichi Sugiura[6], Tomoko Kubota[7], Shinji Miyoshi[7], Kinya Okada[7], Kazuyuki Nakazono[1], Yuki Sugaya[1], Woosung Yang[1], Taiji Sawamoto[2], Wataru Uchida[2], Akira Shinagawa[3], Tsutomu Fujiwara[4], Hisaharu Yamada[5], Koji Suematsu[5], Naohisa Tsutsui[7], Naoyuki Kamatani[1], Shyh-Yuh Liou[6] and the Japan PGx Data Science Consortium (JPDSC)

**Japan Pharmacogenomics Data Science Consortium (JPDSC) has assembled a database for conducting pharmacogenomics (PGx) studies in Japanese subjects. The database contains the genotypes of 2.5 million single-nucleotide polymorphisms (SNPs) and 5 human leukocyte antigen loci from 2994 Japanese healthy volunteers, as well as 121 kinds of clinical information, including self-reports, physiological data, hematological data and biochemical data. In this article, the reliability of our data was evaluated by principal component analysis (PCA) and association analysis for hematological and biochemical traits by using genome-wide SNP data. PCA of the SNPs showed that all the samples were collected from the Japanese population and that the samples were separated into two major clusters by birthplace, Okinawa and other than Okinawa, as had been previously reported. Among 87 SNPs that have been reported to be associated with 18 hematological and biochemical traits in genome-wide association studies (GWAS), the associations of 56 SNPs were replicated using our data base. Statistical power simulations showed that the sample size of the JPDSC control database is large enough to detect genetic markers having a relatively strong association even when the case sample size is small. The JPDSC database will be useful as control data for conducting PGx studies to explore genetic markers to improve the safety and efficacy of drugs either during clinical development or in post-marketing.**
*Journal of Human Genetics* (2015) **60,** 319–326; doi:10.1038/jhg.2015.23; published online 9 April 2015

## INTRODUCTION

Serious adverse drug reaction (ADR) is one of the major problems in clinical medicine. The occurrence of serious ADRs such as serious skin rash (Stevens–Johnson syndrome and toxic epidermal necrolysis) and drug-induced liver injury (DILI) is quite rare (that is, $1\times10^{-3}$ to $1\times10^{-6}$);[1,2] therefore, they are not often recognized during the development of new drugs. In fact, some promising new drugs have been withdrawn from the market because of serious ADRs that occurred in the post marketing phase.[3]

Pharmacogenomics (PGx) has been recognized as an important tool for exploring the genetic factors associated with the efficacy and safety of drugs. In particular, many relevant genetic factors for serious ADRs have been discovered through the PGx approach. They include HLA-B*57:01 for abacavir-induced hypersensitivity,[4] HLA-B*15:02/A*31:01 for carbamazepine-induced Stevens–Johnson syndrome/toxic epidermal necrolysis and HLA-B*58:01 for allopurinol-induced Stevens–Johnson syndrome/toxic epidermal necrolysis.[5–8] These associations were initially identified by the candidate gene approach; however, genome-wide association studies (GWAS) have been widely applied for the screening of genetic factors for serious ADRs.

One of the major practical problems in exploring genetic factors for a serious ADR is the difficulty in collecting DNA samples due to the rarity of the serious ADR. To overcome this problem, Drug-Induced Liver Injury Network (DILIN),[9] EUDRAGENE[10] and DILIGEN[11] were established in the United States, Europe and United Kingdom, respectively. In this framework, International Serious Adverse Events Consortium (iSAEC)[12] identified the genetic marker HLA B*57:01 for flucloxacillin-induced DILI in collaboration with DILIGEN.

From the statistical viewpoint, sample size is an important factor in GWAS analysis to achieve a strong power; however, genetic markers associated with serious ADRs can be often detected with a small sample size because of a rather large effect size of the marker. Nelson *et al.*[13] showed by simulation that the *HLA-B* gene region could be

[1]StaGen, Tokyo, Japan; [2]Astellas Pharma, Tokyo, Japan; [3]Daiichi Sankyo, Tokyo, Japan; [4]Otsuka Pharmaceutical, Tokyo, Japan; [5]Taisho Pharmaceutical, Tokyo, Japan; [6]Takeda Pharmaceutical Company Limited, Osaka, Japan and [7]Mitsubishi Tanabe Pharma Corporation, Osaka, Japan
Correspondence: K Suematsu, PGx, Clinical Research, Taisho Pharmaceutical, 3-24-1 Takada, Toshima-ku, Tokyo 170–8633, Japan.
E-mail: koji.suematsu@po.rd.taisho.co.jp
Received 30 September 2013; revised 12 December 2014; accepted 27 January 2015; published online 9 April 2015

identified with as few as 15 cases and 200 population controls in a sequential analysis.

Although there has been a progress in the development of databases including HapMap Project[14] and the 1000 Genomes project,[15] those databases are not appropriate as control data for Japanese PGx studies because the number of Japanese in these databases are small. To construct a qualified control database for Japanese PGx studies, the Japan Pharmacogenomics Data Science Consortium (JPDSC) was established in 2009 by six pharmaceutical companies, including Astellas, Daiichi Sankyo, Mitsubishi Tanabe, Otsuka, Taisho and Takeda. As a result of the project of the consortium, JPDSC database was established, which contains genotype data obtained from Illumina HumanOmni 2.5–8 and HLA typing kit from 2994 Japanese healthy individuals. In addition, the database contains 121 kinds of clinical information. As a result of the initial analysis, we report a more precise structure of the Japanese population than before and the size of the problem of performing PGx studies without the knowledge of the population structure of the Japanese.

## MATERIALS AND METHODS

### Samples

This study was approved by the ethical committee of JPDSC. Genomic DNA and the corresponding clinical information from 3006 healthy subjects were used in this study after obtaining written informed consent from all participants.

DNA and clinical information from 997 healthy volunteers were the same as those previously reported.[16] The collection of those samples, which was called as the first phase, was performed in the Tokyo Women's Medical University as part of a project of the Pharma SNP Consortium (http://www.jpma.or.jp/information/research/psc/).

The rest of the samples were obtained from 2009 healthy volunteers as a project of JPDSC (the second phase). In this project, the samples were collected from 10 geographic regions in Japan as shown in Figure 1. Sample collection was performed after adjusting the sample size to the population ratio to represent the Japanese population structure. The study plan of the collection in the second phase was shown in Table 1. In particular, the sample size from Okinawa was higher than the expected population ratio[17] because it was well known that the genetic variations in Okinawa are different from those in the other regions in Japan.[18] Moreover, the sample collection was performed such that the male-to-female ratio was 1:1, and three age groups (20–29 years, 30–39 years and ⩾40 years) were 2:2:1, respectively. The 121 variables of the clinical data collected from all the observed subjects are available in Supplementary Table 1.

During the second phase, the samples were collected based on the place of residence of the subjects in each region. As the birthplace is very important in genetic studies and the genetic features of the residence place and birthplace are not necessarily the same,[18] the birthplace data obtained from the questionnaire were used instead of those of the residence place in our present study. In Table 1, the collected size and adjusted size are corresponding to the number of persons deduced from residence place and birthplace based on the questionnaire. Note that the birthplace data for the 997 samples of the first phase were not obtained, so the birthplace of those samples, which were collected in Kanto region, was assumed as Kanto.

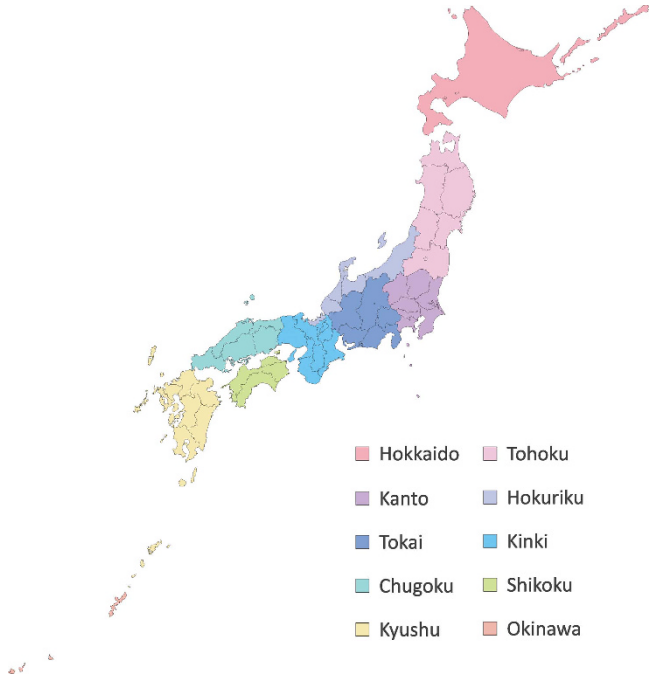


**Figure 1** Geographical regions of Japan. The 2994 Japanese individuals were divided into 10 regions according to their hometown: Hokkaido (119), Tohoku (220), Kanto (1183), Hokuriku (119), Tokai (308), Kinki (336), Chugoku (183), Shikoku (112), Kyushu (298) and Okinawa (116).

### Genotyping

The extraction of genomic DNA from peripheral blood was performed by SRL Co., Ltd. (Shinjuku, Tokyo, Japan). The quality of the genomic DNA was assessed by electrophoresis on agarose gel and optical density measurements. The genotype data were obtained using HumanOmni 2.5-8 BeadChip Kits (Illumina, San Diego, CA, USA) according to the manufacturer's protocol. Five HLA loci (HLA-A, HLA-B, HLA-C, HLA-DRB1 and HLA-DPB1) were genotyped using the Luminex assay system and HLA typing kits (WAKFlow HLA typing kits, Wakunaga, Osaka, Japan; or LABType SSO, One Lambda, Canoga Park, CA, USA). The details of HLA genotyping and its application were previously described by Nakaoka *et al.*[19]

### Sample quality checks

For all the samples, the quality check of the genotype data was performed according to the following criteria. The sample call rate was defined as the

**Table 1 Summary of the numbers and proportions of the subjects included in the study and the data of the Japanese population released by the Japanese Ministry of International Affairs and Communication**

| | *Collected samples 1st+2nd* | *Frequency* | |
|---|---|---|---|
| *Region* | *phase* | *JPDSC[a] (male/female)* | *Japan[b] (male/female)* |
| Hokkaido | 120 | 4.00 (48.70%/51.30%) | 4.30 (47.30%/52.80%) |
| Tohoku | 200 | 7.30 (52.30%/47.70%) | 7.30 (47.90%/52.10%) |
| Kanto | 1197 | 39.50 (41.20%/58.80%) | 33.30 (50.00%/50.00%) |
| Hokuriku | 115 | 4.00 (50.40%/49.60%) | 4.30 (48.30%/51.70%) |
| Tokai | 310 | 10.30 (52.30%/47.70%) | 12.70 (49.50%/50.50%) |
| Kinki | 430 | 11.20 (48.20%/51.80%) | 17.80 (48.20%/51.80%) |
| Chugoku | 160 | 6.10 (54.60%/45.40%) | 5.90 (47.90%/52.10%) |
| Shikoku | 85 | 3.70 (44.60%/55.40%) | 3.10 (47.30%/52.70%) |
| Kyushu | 280 | 10.00 (49.70%/50.30%) | 10.30 (47.00%/53.00%) |
| Okinawa | 109 | 3.90 (50.90%/49.10%) | 1.10 (49.00%/51.00%) |
| Total | 3006 | 100% | 100% |

[a]Percentage of subjects in each region in the adjusted samples used in this study.
[b]Percentage of the subpopulation in each region from the Japanese population data.

proportion of the called genotypes among all the genotypes present in a DNA chip. The samples with sample call rates $<0.99$ were excluded. The frequency of heterozygous SNPs in the probe sets on the × chromosomes was calculated in each self-reported man. A self-reported man with the heterozygous genotype rate $>0.1$ and with the genotype of SNPs on the Y chromosome was excluded from the analysis. In contrast, the frequency of heterozygous in the X chromosomes was also calculated in each self-reported woman. A self-reported woman with the heterozygous genotype rate $<0.1$, and no genotype of SNPs on Y chromosome was excluded from the analysis. For checking the mixture of several identical individuals, the identity by descent between each pair from all the samples was also evaluated as the proportional number of alleles shared identity by descent. Samples with a score $>0.8$ were excluded.

### Population structure

The population structure and the homogeneity of the samples were also evaluated by principal component analysis (PCA) by using 174 Utah residents with ancestry from Northern and Western Europe (CEU), 176 Yoruba residents in Ibadan (YRI), 86 Han-Chinese (CHB) residents and 89 Japanese residents in Tokyo (JPT) genotyping data found in the HapMap database (version HapMap3 Public Release #3, released by the international HapMap consortium, 2003) as previously described, and the samples that did not belong to the Japanese population were excluded.

The differences in the genotype distribution for 10 geographic regions were also evaluated by the results of PCA, by using JPDSC, JPT and CHB samples.

The tag SNPs in each block were selected under the linkage dis-equilibrium coefficient $r^2 \geqslant 0.8$. The SNPs that were in common between our data and HapMap were also selected.

### Re-genotyping

Using the samples that passed the quality check, the genotypes of all SNPs were re-called by the software GenomeStudio (Illumina).

After re-calling the genotype, the quality check process was repeated until no more samples were removed.

### SNP quality checks

SNPs used in the population structure analysis and the GWAS were selected according to the following criteria: SNP call rate, not $<0.95$; minor allele frequency, not $<0.01$; and *P*-value for Hardy–Weinberg equilibrium test, not $<0.001$.

### Replication study of previous reports

Replication study was performed for 10 biochemical and 8 hematological traits, whose associations were previously reported in Kamatani *et al.*[20] To evaluate the association between a trait and the genotypes of SNPs, a multivariable linear regression model was used by incorporating age, sex, weight and height as covariates as in Kamatani *et al.*[20] The significance of the association between a trait and genotypes was evaluated using the Wald test.

### Software

Quality check of genotype data and samples, the estimation of IBD and the association test by linear regression were performed with PLINK software version 1.0.7 (PLINK, Cambridge, MA, USA).[21] For the evaluation of the population structure, EIGENSOFT software (EIGENSOFT, Cambridge, MA, USA)[22] was used. The Manhattan plot was drawn using the Haploview software version 4.2 (Haploview, Cambridge, MA, USA).[23]

## RESULTS

### Sample and SNP quality checks

The 2994 samples out of 3006 were acceptable. The details of the results for sample quality checks are described in the Supplementary Information. The PCA plots for population structure were illustrated in Figure 2 and Supplementary Figure 1.

We also checked the genotype quality using SNP call rates, minor allele frequency (MAF) and goodness-of-fit for the Hardy–Weinberg equilibrium (HWE). The process and results are summarized in Table 2, Supplementary Table 2, Supplementary Figures 2–4. After all these quality check, 1 333 978 SNPs and 2994 samples were used for GWAS analysis.

The JPDSC data (the allele frequency data from about 1.4M SNP of all 2994 Japanese healthy volunteers and each birth place region) are now available in the National Bioscience Database Center, http://humandbs.biosciencedbc.jp/.

### Population structure

We investigated the population structure of the JPDSC samples by PCA and compared the results with the self-reported birthplaces (Figure 2). By the PCA, the JPDSC samples were separated into two major clusters, the Ryukyu and the Hondo clusters as denoted by Yamaguchi-Kabata *et al.*[18] The present study, unlike the previous study, included the residents in the Chugoku and the Shikoku regions (Figure 2). In addition, unlike the previous report, the residents of Tokai and Hokuriku regions were separately analyzed in the present study. The distributions of the plots for the previously reported regions, that is, Okinawa, Kyushu, Kinki, Kanto, Tohoku and Hokkaido were essentially the same as the previous report.[18] We found that the plots for the Shikoku residents showed a similar pattern as those for the Kinki residents that was deviated from the center of the Hondo cluster toward the Han-Chinese cluster from Figure 2. The distribution of the plots for the Hokuriku residents was different from the Tokai residents but was similar to the Kinki residents. The plots for the Chugoku residents were distributed almost in the center of the Hondo cluster, and deviated toward neither the Han-Chinese cluster nor the Ryukyu cluster. It is of interest that the plots for Tohoku, Tokai and Kyushu residents were deviated from the center of the Hondo cluster toward the Ryukyu cluster. From these results, we may be able to speculate that some people came from the Asian continent to the Japanese islands where older ancestors of the Japanese had resided. The people currently living in Kinki, Tokuriku and Shikoku regions may reflect more intensely the genetic factors of the new comers, and the people living in Chugoku region may be a mixture between the two populations.

### Replication study

Supplementary Table 3 shows the results for 56 SNPs with $P<0.05$ obtained from the replication study for the associations reported by Kamatani *et al.*,[20] including the results for eight biochemical traits and six hematological traits. The signs of the coefficient that represents the tendency of association between our study and Kamatani's report were identical. In Supplementary Table 3, 56 of 87 SNPs showed statistical significance ($P<0.05$). Moreover, the signs of the coefficient for 54 of those SNPs were also identical in the Riken and JPDSC control groups. The remaining 31 SNPs were non-replicated SNPs. These SNPs were not replicated in our study possibly because of the difference in the sample size. The sample sizes in our study and Kamatani's study were 2994 and 14 700, respectively.

## DISCUSSION

### Background of collected samples

JPDSC collected DNA samples so that the entire set reflected, as much as possible, the Japanese population, based on the size of the population of each region, sex and age. Table 1 shows the summary of the numbers and the proportions of the subjects included in our study for each region and each sex as well as the proportions from the Japanese population data. The data in this table indicate that our sampling is a good reflection of the Japanese population from the viewpoints of both regional distribution and sex. However, the

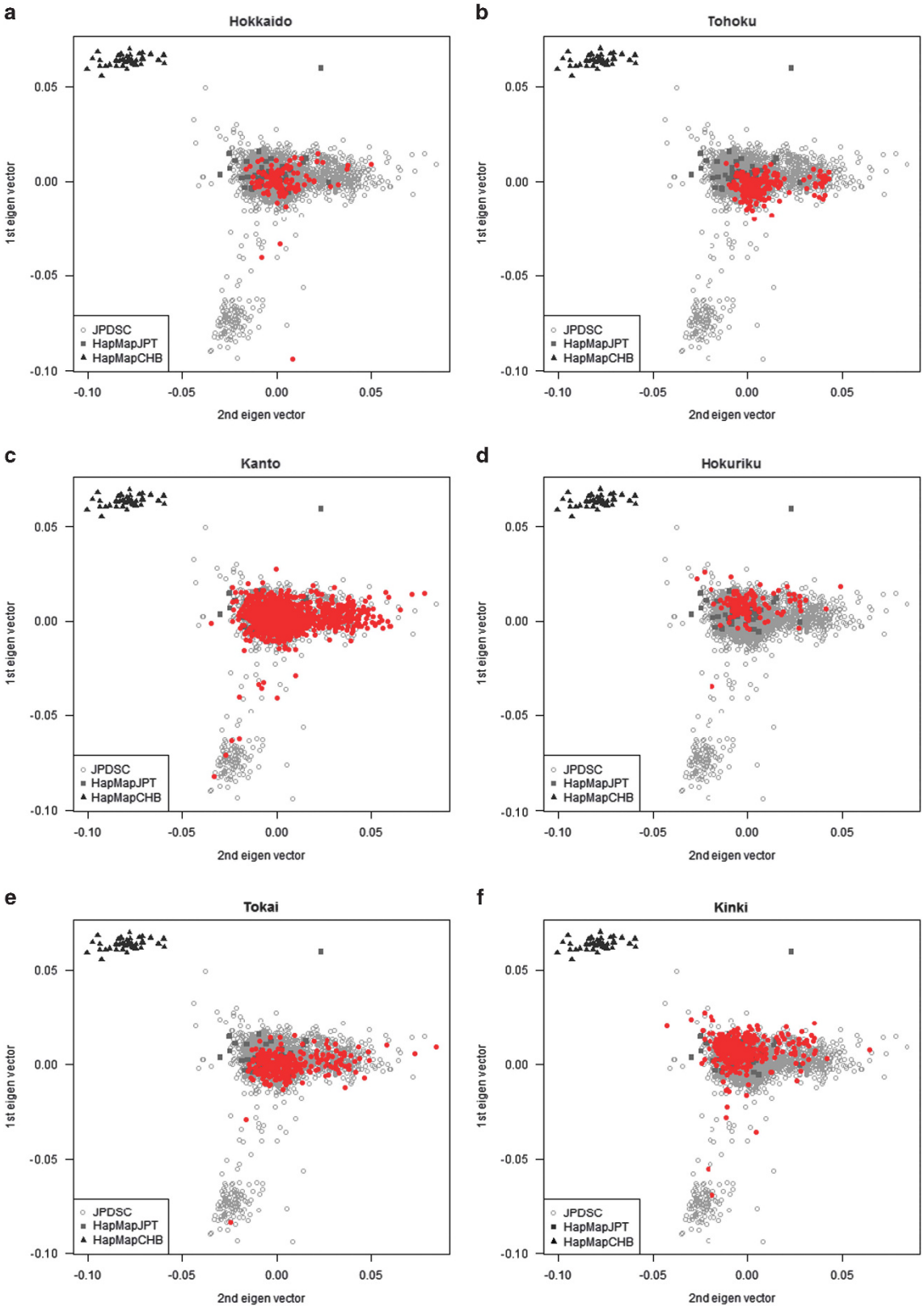


**Figure 2** PCA plots of the Japanese individuals for each geographical region. The 10 PCA plots from (**a**) to (**j**) correspond to the 10 geographic regions (see the details in Figure 1 legend) and the individuals from each region were highlighted by different colors. The gray points correspond to the subjects from HapMap JPT (44) and CHB (45). The regions are indicated as follows: (**a**) Hokkaido, (**b**) Tohoku, (**c**) Kanto, (**d**) Hokuriku, (**e**) Tokai, (**f**) Kinki, (**g**) Chugoku, (**h**) Shikoku, (**i**) Kyushu and (**j**) Okinawa. The number of the subjects in each region was as follows: Hokkaido (119), Tohoku (220), Kanto (1183), Hokuriku (119), Tokai (308), Kinki (336), Chugoku (183), Shikoku (112), Kyushu (298) and Okinawa (116).

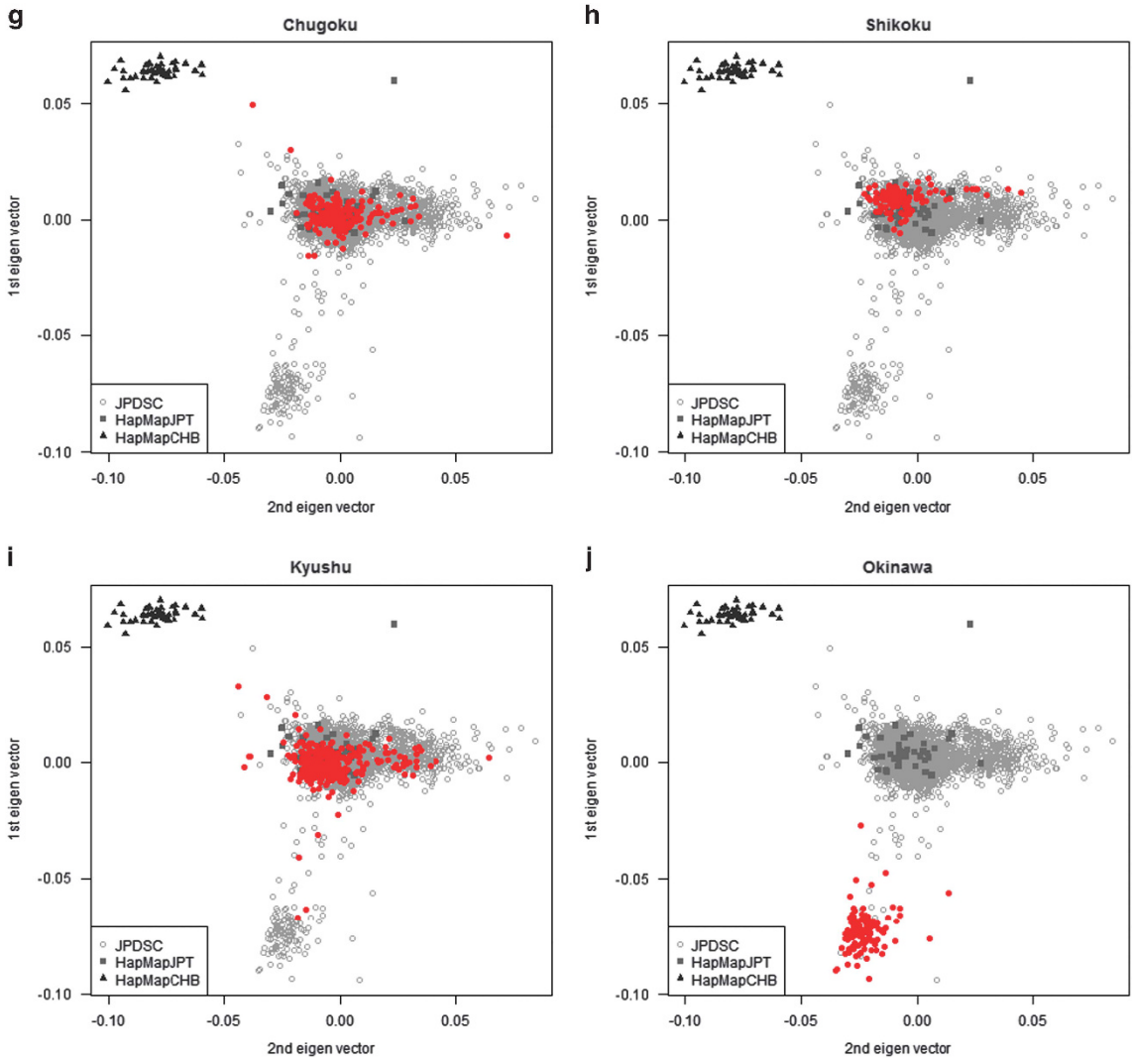


**Figure 2** Continued.

proportion of the subjects from the Okinawa region in our sample is higher than that in the Japanese population data. As the Okinawa region is unique and important to understand the genetic variations in Japan,[18] JPDSC collected the Okinawa sample more intensively than the other regions. The proportion of each sex in each region is also identical to the real population of Japan.

In Section Sample and SNP quality checks, five individuals (three self-reported men and two self-reported women) were excluded from the analysis. The frequencies of heterozygous SNPs on the X chromosome for the two self-reported women were $<0.1$ and almost 0. One of the samples may be derived from a subject with monosomy X, while the other one may have a partial loss of one of the two X chromosomes. Moreover, three self-reported men had high frequencies of heterozygous SNPs on the X chromosome as well as high call rates of SNPs on the Y chromosome. These men may have two or more X chromosomes and one Y chromosome.

### Validity of sample size

It is common that PGx studies often fail to collect sufficient cases to achieve appropriate statistical power. When the sample size is limited, a strategy to enhance the statistical power of PGx is to provide a sufficient sample size of the control group. The control group should have the same genetic background as the case group.

To overcome the problem of small sample size in the case of Japanese PGx studies, JPDSC collected the genotypes and clinical data from about 3000 healthy volunteers. The goal of a sample size of 3000 was decided by the statistical power simulation.

In case of PGx studies with very small sample sizes, the rare risk alleles in the population may be detected through association studies. Figure 3 shows the relationship between MAF and statistical power. Figure 3 indicates that the peak of the power is achieved around MAF values between 0.05 and 0.1. Moreover, this result means that it is hard to detect associated SNPs when MAFs are over 0.3. We observed the increase in the power as the control size increased (Figure 4). Figure 4 shows that the power peaked at the control sample size of 1500 for each odds ratio (OR). In PGx studies, the association with sex may be different, and therefore JPDSC planned to collect 3000 healthy volunteers, including 1500 men and 1500 women.

By using 14 cases and 991 JPDSC controls, Tohkin *et al.*[24] detected the SNPs associated with Stevens–Johnson syndrome and toxic epidermal necrolysis induced by allopurinol. The MAF of the detected SNPs in JPDSC control sample was about 0.0055, and the OR for SNP

**Table 2 Reduction in the number of the SNPs by the quality check. The initial number of available SNPs for analysis was 1 333 978 under the selection criteria CR ⩾ 0.95, MAF ⩾ 0.01 and HWE $P \geqslant 0.001$**

| *Assessment* | *Criteria* | *Number of SNPs* | *Number of SNPs Excluded* |
|---|---|---|---|
| Overall | | 2 379 855 | — |
| SNP call rate | ⩾0.95 | 2 376 936 | 2919 |
| MAF | ⩾0.01 | 2 191 329 | 185 607 |
| MAF (monomorphic SNP) | 0 | 1 337 510 | 853 819 |
| SNP recorded in HapMap project (710 411 SNPs on chip) | | | 84 177 |
| SNP recorded in 1000 Genomes SNP (1 665 231 SNPs on chip) | | | 768 593 |
| HWE *P*-value | ⩾0.001 | 1 333 978 | 3532 |
| Number of selected SNPs | | 1 333 978 | |

Abbreviations: HWE, Hardy–Weinberg equilibrium; MAF, minor allele frequency; SNP, single-nuclotide protein.

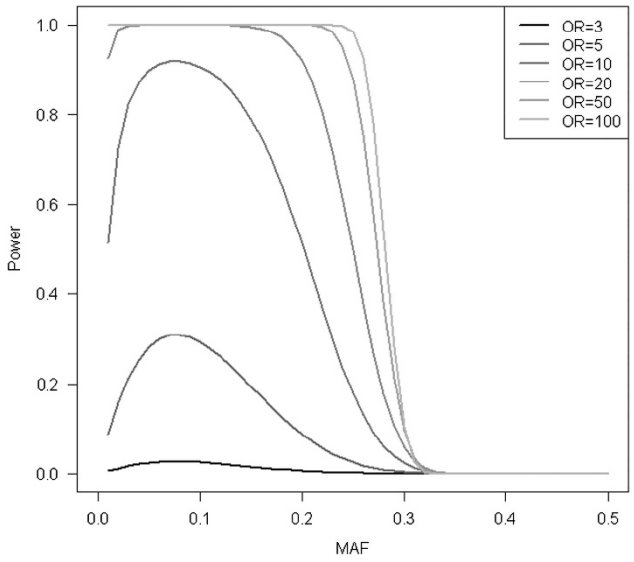


**Figure 3** Power simulation by odds ratio when sample size of case group is very small. Horizontal axis means the minor allele frequency, and the vertical axis means statistical power. In this simulation, sample size of case and control group are fixed as 30 and 2994, respectively. The significant level 3.54E−8 was calculated as 0.05/1411375 under the selection of SNP from quality check with the criteria CR ⩾ 0.95, MAF ⩾ 0.001 and HWE ⩾ 0.001. A full color version of this figure is available at the *Journal of Human Genetics* journal online.

with the strongest association was 66.8 under the dominant mode for the minor allele. Considering the MAF in the control group (0.0055) and the OR (66.8), the statistical power of Tohkin's study is estimated as 95.7% under the assumption that a SNP with the same strength and frequency is statistically significant.

### Available SNPs for GWAS

As indicated in the results of the quality check for SNPs in Section Population Structure, about 1.4 million SNPs were available for

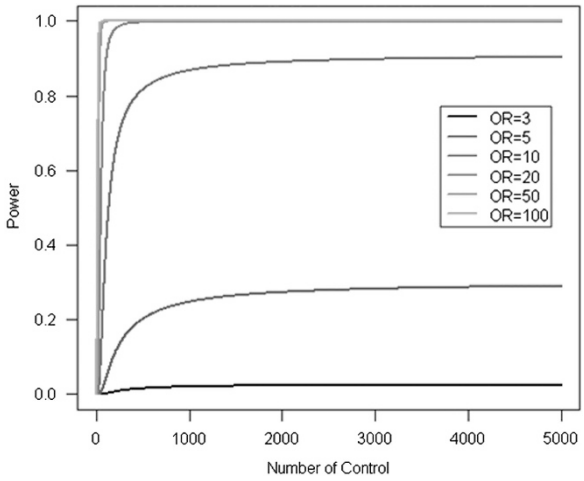


**Figure 4** Power simulation by odds ratio as the size of the control group increases. The horizontal axis represents the control size, and the vertical axis represents the statistical power. In this simulation, the sample size of the case group was fixed as 30, MAF was set as 0.05 and the dominant mode of minor allele was assumed. The significance level ($3.54 \times 10^{-8}$) was calculated as 0.05/1 411 375 under the selection of SNPs from the quality check with the criteria CR ⩾ 0.95, MAF ⩾ 0.001 and HWE ⩾ 0.001. A full color version of this figure is available at the *Journal of Human Genetics* journal online.

statistical analysis. In other words, about 1 million SNPs were excluded based on the general criteria of the quality check. Table 2 shows the reduction in the number of SNPs after each quality control (QC) step, and the classification of the monomorphic SNPs into two categories based on whether they were from the HapMap project or the 1000 Genomes project. The data in this table indicate that most of the monomorphic SNPs in our data were selected from the 1000 Genomes project, although about one-third of the SNPs in the chip were from that project.

SNPs with low frequencies may not be included in a sample when the sample size is limited. The Supplementary Table 4 shows the number of samples that, under the assumption of HWE, are expected to carry at least one rare allele. Here, MAF in SNP was assumed as 0.001–0.1. The Supplementary Table 4 indicates that, when the population MAF is >0.001, at least two samples are expected to be observed in the JPDSC sample. As the number of the Japanese subjects included in 1000 Genomes projects is much smaller than JPDSC, some of our SNPs are unlikely to be included in the data from 1000 Genomes.

### Genetic character of each regions

As described in Section Population Structure, Figure 2 shows the results of PCA indicating the population structure of the JPDSC control. Our results on the Japanese population structure are comparable to those obtained by Yamaguchi *et al.*[18] by using GWAS data.

The distribution of the subjects from Chugoku and Shikoku regions, which had not been included in Yamaguchi's work, was observed. As shown in Figure 2g, the subjects from the Chugoku region seemed to be located almost at the center of the Hondo region and separated from the other regions, similar to the data from the Kyushu. It is interesting to note that the distribution of the subjects from Shikoku was close to that of the subjects from Hokuriku and Kinki regions but far from the subjects of Chugoku region. These

results indicate that the subjects from Shikoku and Chugoku differ in ancestry in spite of being geographically closed. Therefore, it is possible that the inhabitants of the Shikoku and Hokuriku regions may have originated from Kinki region.

Consistent with Yamaguchi's study, Figure 2 also indicates that the distribution of the subjects from Okinawa region differs from that of the population from Hondo region.

To examine the difference between Okinawa and Hondo regions in more detail, we performed a case–control analysis by using 116 individuals from Okinawa and 2878 individuals from Hondo.

*EDAR gene*. Yamaguchi (2008) reported that SNPs on the *EDAR* gene were significantly different between the populations of Okinawa and Hondo. In our results, the kgp3482957 on the *EDAR* gene was also significantly different (OR = 2.638; $P = 4.41 \times 10^{-9}$), and the results were replicated. MAF of the minor allele A for Okinawa and Hondo regions were 0.23 and 0.087, respectively. For other SNPs on *EDAR*, MAFs for Okinawa region were higher than those for Hondo region. EDAR has been reported to be highly differentiated in the Asian subpopulations.[25] EDAR and EDA2R, both involved in the development of hair follicles[25–27] and teeth,[28,29] have probably undergone positive selection in Asia.[25] While the wild-type V is almost fixed in the European and African populations, the nonsynonymous V370A mutation is more frequent among Asians and the associated frequency of the allele A in the Okinawa and Hondo populations is 0.233 and 0.111, respectively.

There were some genome-wide significant SNPs between the Okinawa and Hondo group in genes of drug-metabolizing enzymes and transporters defined as 'ADME gene list' in PharmaADME database (http://pharmaadme.org/) in Supplementary Table 5. But, in 32 important ADME genes defined as 'core ADME gene list' in PharmaADME database, such as CYP2D6, CYP2C9 and CYP2C19, no SNP was significantly different in frequency between the Okinawa and Hondo group.

*Other interesting SNPs*. We also found that the frequencies of the SNPs in the genes described below were significantly different between Okinawa and Hondo regions.

The SNP that showed the lowest *P*-value in the trend test while comparing the genotype frequencies among the PGx-related genes of the Okinawa and Hondo residents was kgp8366488 in ABCB5. ABCB5 is an ABC transporter and P-glycoprotein family member expressed in skin and in malignant melanoma. ABCB5 has been suggested to regulate skin progenitor cell fusion and mediate chemotherapeutic drug resistance in stem-like tumor cell subpopulations in human malignant melanoma. According to our analysis, kgp8366488 SNP, located between the genes *ITGB8* and *ABCB5*, was significantly associated with the difference between Okinawa and Hondo groups (OR = 7.42 and $P = 3.23 \times 10^{-33}$). In addition, the MAF of kgp8366488 for Okinawa was 0.168, while that for Hondo was 0.0265.

The SNP that showed the second lowest *P*-value was rs9729287, located between the genes *PL2G4A* and *FAM5C*. The *PLA2G4A* gene encodes a member of the cytosolic phospholipase A2 group IVA family. This enzyme catalyzes the hydrolysis of membrane phospholipids to release arachidonic acid that is subsequently metabolized into eicosanoids. This gene is also related to the onset of cardiovascular diseases. In our analysis, rs9729287 SNP showed a low *P*-value according to the Cochran-Armitage trend test on Okinawa and Hondo subjects (OR = 6.55 and $P = 2.95 \times 10^{-27}$). The MAF of rs9729287 for Okinawa was 0.168, while that for Hondo was 0.0273.

As JPDSC control sample was originally collected for the purpose of PGx studies, it is important to know the differences in the allele frequencies of ADME and other PGx-related genes between Okinawa and Hondo regions. Therefore, when PGx studies are conducted comparing the genes listed in Supplementary Table 5, it is necessary not to combine the Hondo and Okinawa groups.

### Allele frequency of HLA in JPDSC

To confirm the consistency between JPDSC population and actual Japanese population, the allele frequencies of the HLA loci (HLA-A, HLA-B, HLA-C, HLA-DRB1 and HLA-DPB1) were compared with the database of the HLA laboratory and Central Bone Marrow Data Center. Supplementary Table 6 shows the frequencies of the alleles in five HLA loci obtained from JPDSC and other two databases. This table indicates that the JPDSC control samples represent the general healthy Japanese population.

JPDSC control samples, collected to reflect the general healthy Japanese population, contain >1 300 000 genotypes as well as demographic and clinical laboratory data. The five loci of HLA are also available. Although the main purpose of the JPDSC control data was to use for PGx studies, the data may be useful for various genetic association studies focusing on rare diseases. As the size of the database is quite large, it may be useful for association studies in which the sample size of the case group is rather small but the strength of the association (effect size) is relatively large. As the database also contains regional information, it may also be suitable for genetic studies involving Japanese subjects when the regional differences in the allele frequencies may interfere with the results. In particular, the individuals from Okinawa have been sampled in more detail than the subjects from other regions. Therefore, it is expected that this study will yield a more detailed understanding of the genetic variations in the Japanese population in terms of PGx.

## CONFLICT OF INTEREST

Shigeo Kamitsuji, Kazuyuki Nakazono, Woosung Yang, Yuki Sugaya, and Naoyuki Kamatani are employed by StaGen, Tokyo, Japan.

## ACKNOWLEDGEMENTS

We thank all the study participants for making this study possible. We thank for the data kindly provided by the Japan Pharmacogenomics Data Science Consortium (JPDSC), which is composed of Astellas Pharma, Otsuka Pharmaceutical, Daiichi-Sankyo, Taisho Pharmaceutical, Takeda Pharmaceutical and Mitsubishi Tanabe Pharma Corporation, and is chaired by Dr Ichiro Nakaoka (Takeda Pharmaceutical).

1 Phillips, E. J., Chung, W.-H., Mockenhaupt, M., Roujeau, J.-C. & Mallal, S. A. Drug hypersensitivity: pharmacogenetics and clinical syndromes. *J. Allergy Clin. Immunol.* **127**, 60–66 (2011).
2 Farrell, G. C. & Liddle, C. Drugs and the liver updated. *Semin. Liver Dis.* **22**, 109–113 (2002).
3 Kaplowitz, N. Idiosyncratic drug hepatotoxicity. *Nat. Rev. Drug Discov.* **4**, 489–499 (2005).
4 Mallal, S., Nolan, D., Witt, C., Masel, G., Martin, A. M., Moore, C. *et al.* Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *Lancet* **359**, 727–732 (2002).
5 Hung, S.-I., Chung, W.-H., Liou, L.-B., Chu, C.-C., Lin, M., Huang, H.-P. *et al.* HLA-B*5801 allele as a genetic marker for severe cutaneous adverse reactions caused by allopurinol. *Proc. Natl Acad. Sci. USA* **102**, 4134–4139 (2005).
6 Tassaneeyakul, W., Jantararoungtong, T., Chen, P., Lin, P.-Y., Tiamkao, S., Khunarkornsiri, U. *et al.* Strong association between HLA-B*5801 and allopurinol-induced Stevens-Johnson syndrome and toxic epidermal necrolysis in a Thai population. *Pharmacogenet. Genomics* **19**, 704–709 (2009).
7 Ozeki, T., Mushiroda, T., Yowang, A., Takahashi, A., Kubo, M., Shirakata, Y. *et al.* Genome-wide association study identifies HLA-A*3101 allele as a genetic risk factor for

carbamazepine-induced cutaneous adverse drug reactions in Japanese population. *Hum. Mol. Genet.* **20**, 1034–1041 (2011).
8 McCormack, M., Alfirevic, A., Bourgeois, S., Farrell, J. J., Kasperavičiūtė, D., Carrington, M. *et al.* HLA-A*3101 and carbamazepine-induced hypersensitivity reactions in Europeans. *N. Engl. J. Med.* **364**, 1134–1143 (2011).
9 Fontana, R. J., Watkins, P. B., Bonkovsky, H. L., Chalasani, N., Davern, T., Serrano, J. *et al.* Drug-Induced Liver Injury Network (DILIN) prospective study: rationale, design and conduct. *Drug Saf.* **32**, 55–68 (2009).
10 Molokhia, M. & McKeigue, P. EUDRAGENE: European collaboration to establish a case-control DNA collection for studying the genetic basis of adverse drug reactions. *Pharmacogenomics* **7**, 633–638 (2006).
11 Daly, A. K., Donaldson, P. T., Bhatnagar, P., Shen, Y., Pe'er, I., Floratos, A. *et al.* HLA-B*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin. *Nat. Genet.* **41**, 816–819 (2009).
12 The international Serious Adverse Event Consortium (iSAEC). Available from http://www.saeconsortium.org/.
13 Nelson, M. R., Bacanu, S.-A., Mosteller, M., Li, L., Bowman, C. E., Roses, A. D. *et al.* Genome-wide approaches to identify pharmacogenetic contributions to adverse drug reactions. *Pharmacogenomics J.* **9**, 23–33 (2009).
14 Tanaka, T. The International HapMap Project. *Nature* **426**, 789–796 (2003).
15 Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
16 Kamatani, N., Sekine, A., Kitamoto, T., Iida, A., Saito, S., Kogame, A. *et al.* Large-scale single-nucleotide polymorphism (SNP) and haplotype analyses, using dense SNP Maps, of 199 drug-related genes in 752 subjects: the analysis of the association between uncommon SNPs within haplotype blocks and the haplotypes constructed with hap. *Am. J. Hum. Genet.* **75**, 190–203 (2004).
17 Ministry of International Affairs and Communication. Population by Age (5-Year Age Group) and Sex for Prefectures.
18 Yamaguchi-Kabata, Y., Nakazono, K., Takahashi, A., Saito, S., Hosono, N., Kubo, M. *et al.* Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. *Am. J. Hum. Genet.* **83**, 445–456 (2008).
19 Nakaoka, H., Mitsunaga, S., Hosomichi, K., Shyh-Yuh, L., Sawamoto, T., Fujiwara, T. *et al.* Detection of ancestry informative hla alleles confirms the admixed origins of Japanese population. *PLoS ONE* **8**, e60793 (2013).
20 Kamatani, Y., Matsuda, K., Okada, Y., Kubo, M., Hosono, N., Daigo, Y. *et al.* Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nat. Genet.* **42**, 210–215 (2010).
21 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. a. R., Bender, D. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
22 Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
23 Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
24 Tohkin, M., Kaniwa, N., Saito, Y., Sugiyama, E., Kurose, K., Nishikawa, J. *et al.* A whole-genome association study of major determinants for allopurinol-related Stevens-Johnson syndrome and toxic epidermal necrolysis in Japanese patients. *Pharmacogenomics J.* **13**, 60–69 (2011).
25 Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).
26 Fujimoto, A., Ohashi, J., Nishida, N., Miyagawa, T., Morishita, Y., Tsunoda, T. *et al.* A replication study confirmed the EDAR gene to be a major contributor to population differentiation regarding head hair thickness in Asia. *Hum. Genet.* **124**, 179–185 (2008).
27 Fujimoto, A., Kimura, R., Ohashi, J., Omi, K., Yuliwulandari, R., Batubara, L. *et al.* A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Hum. Mol. Genet.* **17**, 835–843 (2008).
28 Park, J.-H., Yamaguchi, T., Watanabe, C., Kawaguchi, A., Haneji, K., Takeda, M. *et al.* Effects of an Asian-specific nonsynonymous EDAR variant on multiple dental traits. *J. Hum. Genet.* **57**, 508–514 (2012).
29 Kimura, R., Yamaguchi, T., Takeda, M., Kondo, O., Toma, T., Haneji, K. *et al.* A common variation in EDAR is a genetic determinant of shovel-shaped incisors. *Am. J. Hum. Genet.* **85**, 528–535 (2009).

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