OPEN

Journal of Human Genetics (2015) 60, 581–587

www.nature.com/jhg

ORIGINAL ARTICLE

Japonica array: improved genotype imputation by designing a population-specific SNP array with 1070 Japanese individuals

Yosuke Kawai[1,2], Takahiro Mimori[1], Kaname Kojima[1,2,3], Naoki Nariai[1,2], Inaho Danjoh[1], Rumiko Saito[1], Jun Yasuda[1,2], Masayuki Yamamoto[1,2] and Masao Nagasaki[1,2,3,4]

The Tohoku Medical Megabank Organization constructed the reference panel (referred to as the 1KJPN panel), which contains >20 million single nucleotide polymorphisms (SNPs), from whole-genome sequence data from 1070 Japanese individuals. The 1KJPN panel contains the largest number of haplotypes of Japanese ancestry to date. Here, from the 1KJPN panel, we designed a novel custom-made SNP array, named the Japonica array, which is suitable for whole-genome imputation of Japanese individuals. The array contains 659 253 SNPs, including tag SNPs for imputation, SNPs of Y chromosome and mitochondria, and SNPs related to previously reported genome-wide association studies and pharmacogenomics. The Japonica array provides better imputation performance for Japanese individuals than the existing commercially available SNP arrays with both the 1KJPN panel and the International 1000 genomes project panel. For common SNPs (minor allele frequency (MAF) > 5%), the genomic coverage of the Japonica array ($r^2 > 0.8$) was 96.9%, that is, almost all common SNPs were covered by this array. Nonetheless, the coverage of low-frequency SNPs (0.5% < MAF ⩽ 5%) of the Japonica array reached 67.2%, which is higher than those of the existing arrays. In addition, we confirmed the high quality genotyping performance of the Japonica array using the 288 samples in 1KJPN; the average call rate 99.7% and the average concordance rate 99.7% to the genotypes obtained from high-throughput sequencer. As demonstrated in this study, the creation of custom-made SNP arrays based on a population-specific reference panel is a practical way to facilitate further association studies through genome-wide genotype imputations.
Journal of Human Genetics (2015) **60,** 581–587; doi:10.1038/jhg.2015.68; published online 25 June 2015

INTRODUCTION

High-throughput genotyping is now a prerequisite for genome-wide association studies (GWAS). Single nucleotide polymorphism (SNP) genotyping by DNA microarray (SNP array) has been a central part of massive genotyping tools for GWAS. Although whole-genome sequencing (WGS) (or whole exome sequencing) by high-throughput sequencers enables researchers to identify a massive amount of genetic variations, the cost of WGS is still expensive for GWAS that require genotyping of thousands of individuals. Genotype imputation bridges a gap between the cost-effectiveness of SNP arrays and the comprehensiveness of WGS.[1,2] If the collection of haplotypes in reference panel is created from WGS data, the genotypes of whole genomes can be inferred by genotype imputation with appropriate tag SNPs that are usually genotyped by a SNP array. Indeed, many GWAS successfully identified associations of complex diseases and/or quantitative traits with genetic variants that were imputed from whole-genome reference panels,[3,4] such as the International 1000 genomes project (1KGP) panel.[5]

Generally, genotype imputation is less accurate for low-frequency SNPs (0.5% < minor allele frequency (MAF) ⩽ 5%) than common SNPs (MAF > 5%). However, in GWAS, it is desirable that genotypes of variants can be inferred from genotype imputation with a broad MAF range in cases where low-frequency variants are associated with complex diseases.[6] The size and quality of the reference panel are major determinants of the accuracy of genotype imputation.[7] Because a low-frequency allele rarely lies in a certain haplotype in a reference panel (especially when the size of the reference panel is small), larger reference panels that contain diverse haplotypes and precise haplotyping (phasing) can improve imputation accuracy. In addition, the genotype imputation identifies regions in a chromosome shared between a sample and a haplotype in the reference panel, and thus, the optimal configuration of tag SNPs consisting of many alleles that efficiently capture haplotypes in the reference panel also results in accurate genotype imputation.[8] Given the situation, a higher density SNP array is suitable for whole-genome imputation although an increase in the number of SNPs on an array vitiates the

[1]Department of Integrative Genomics, Tohoku Medical Megabank Organization, Tohoku University, Sendai, Japan; [2]Graduate School of Medicine, Tohoku University, Sendai, Japan; [3]Department of Cohort Genome Information Analysis, Tohoku Medical Megabank Organization, Tohoku University, Sendai, Japan and [4]Graduate School of Information Sciences, Tohoku University, Sendai, Japan
Correspondence: Professor M Nagasaki, Department of Integrative Genomics, Tohoku Medical Megabank Organization, Tohoku University, 2-1 Seiryo-machi, Aoba-ku, Sendai 980-8573, Miyagi, Japan.
E-mail: nagasaki@megabank.tohoku.ac.jp
Received 13 January 2015; revised 13 May 2015; accepted 18 May 2015; published online 25 June 2015

cost-effectiveness. Since low-frequency SNPs tend to be population specific, it is expected that a selection of tag SNPs in which the linkage disequilibrium structure of a particular population are taken into account will increase the accuracy of low-frequency SNP imputation.

We are conducting a genome cohort study as part of the Tohoku Medical Megabank Project and constructed a collection of haplotypes from 1070 healthy individuals in Japan (1KJPN).[9] We demonstrated that the haplotype collection from 1KJPN offers practical accuracy and coverage for genotype imputation on a whole-genome scale using commercially available SNP microarrays. However, because the existing arrays were designed for SNPs discovered in HapMap[10] or 1KGP[5] in which only a part of the samples are derived from individuals with Japanese ancestry, there is room for improvement in genotype imputation in the Japanese population. Thus, we designed a new SNP array, which is suitable for individuals with Japanese ancestry by choosing an optimal set of tag SNPs, for conducting GWAS and human genetic studies. Herein, we describe the method and the quality assessment of genotype imputation with the tailored SNP array.

MATERIALS AND METHODS

Summary of the reference panel

We have constructed the reference panel of Japanese individuals based on the deep WGS.[9] Here, we summarize the construction of the reference panel used in this study. The study has been performed as part of the prospective cohort study at the Tohoku Medical Megabank Organization (ToMMo) with the approval of the ethical committee of the Tohoku University School of Medicine. All cohort participants are residents of Miyagi Prefecture, Japan and provided their written consent. The WGS was done for 1201 cohort participants, selected after the sample quality control such as the DNA sample quality check and the removal of outlier samples based on SNP array genotyping. Then, high coverage (32.4 on average) whole-genome sequences were obtained by using HiSeq2500 (Illumina, San Diego, CA, USA) with in-house PCR-free protocol.[11] After quality check of sequenced reads with SUGAR,[12] the read mapping and genotype calling were performed by using Bowtie2[13] (version 2.1.0) and Bcftools[14] (version 0.1.17-dev) programs, respectively. We then phased the genotypes obtained from the WGS using HapMonster[15] and ShapeIT2[16] (version 2.r644) programs. In this study, 1070 whole-genome sequences have been used to construct a reference panel (1KJPN) and the remaining samples were used to evaluate the imputation quality. The summary of age and sex of the 1KJPN panel are shown in Supplementary Table 1. We confirmed that 1070 samples of reference panel and 131 samples of imputation subject (ToMMo131) belong to the same cluster of Japanese in Tokyo (JPT sample of the 1KGP) and are within the genetic diversity of JPT samples (Supplementary Figure 1).

Selection of tag SNPs

Our aim was to select the tag SNPs so that the maximum imputation performance will be achieved for target SNPs that are SNPs of MAF $\geqslant$ 0.5% in the 1KJPN panel. It is generally difficult to call rare SNPs since the cluster of low-frequency genotype may not be well separated. Thus, we excluded SNPs where MAF in the 1KJPN panel $<$0.5% from tag SNPs to avoid miscall due to poor cluster separation. Figure 1 represents the summary of tag SNP selection. In our design, a candidate set of tag SNPs (shortly candidate tag SNPs) is an intersection of target SNPs and the SNPs experimentally validated on the genotyping platform where the array is made (Axiom Genotyping Array, Affymetrix, Santa Clara, CA, USA), which ensure to achieve high conversion and call rates to the designed probes. Tag SNPs were selected from the candidate tag SNPs until the candidate tag SNPs became empty. The only female samples were used for tag SNPs selection of X chromosome. For each tag SNP selection step, the scores of the current candidate tag SNPs were newly re-evaluated based on the already selected tag SNPs, and then the tag SNP with the highest score was selected and in parallel the selected SNP was also removed from the candidate tag SNPs. By repeating the step, all tag SNPs are ranked by scores that reflect their contribution in inferring genotypes of target SNPs in the reference panel. The score of i-th tag SNP S_i is defined as follows:

$$S_i = \sum_{j \in C_i} T_{ij}, \quad (1)$$

where T_{ij} is a score for pair of the i-th tag SNP and j-th target SNP; C_i is an index set of target SNPs that are subjects for the score calculation. T_{ij} is calculated by considering whether the j-th target is tagged by already selected tag SNPs:

$$T_{ij} = \max\left(0, I_{ij} - \max_{k \in U} I_{kj}\right)/n_i,$$

where I_{ij} represents the mutual information (MI) of genotypes at i-th tag SNP and j-th target SNP; U is an index set of selected tag SNPs; and n_i is the number of required probes to select i-th tag, which equals four for SNPs with A/T or C/G alleles and two for other SNPs in Axiom Genotyping platform.

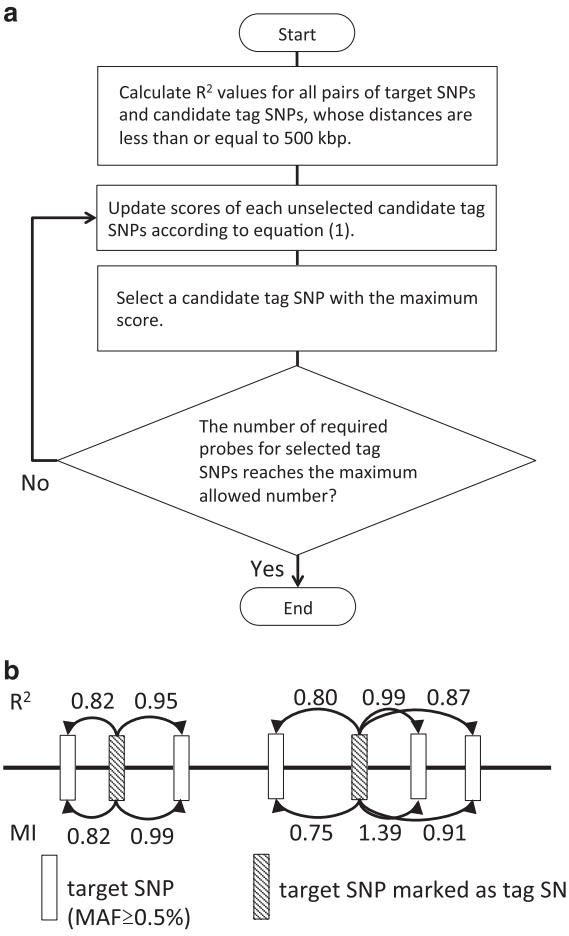

Figure 1 Schematic illustration of tag SNP selection. (**a**) The flowchart represents the algorithm of tag SNP selection. Target SNPs were selected from SNPs of the 1KJPN panel so that the MAF of each target SNP was $\geqslant$0.5%. The tag SNPs were progressively selected from the target SNPs according to the algorithm. (**b**) Schematic illustration of target SNPs and tag SNPs along with a chromosomal region. R^2 is the LD measure calculated as the squared correlation coefficient between genotype frequencies of a pair of SNPs. Note that the R^2 described here is distinct from the measure of imputation accuracy, r^2. The MI is calculated between a pair of SNPs and reflects MAFs and the LD strength of the pair. LD, linkage disequilibrium; MAF, minor allele frequency; MI, mutual information; SNP, single nucleotide polymorphism.

In this study, we set C_i to indicate all the target SNPs located within ± 500 kb from i-th tag SNP and with high linkage disequilibrium ($R^2 \geqslant 0.8$) from i-th tag SNP in the reference panel. In the calculation of R^2 value, genotype is encoded as one of 0, 1 and 2, which corresponds to minor homozygous, heterozygous and major homozygous, respectively. MI of i-th and j-th SNPs (I_{ij}) is defined using entropy of i-th SNP, that of j-th SNP and that of joint distribution of i-th and j-th SNPs as follows:

$$I_{ij} = H_i + H_j - H_{ij},$$

$$H_i = -\sum_{g_i=0}^{2} \frac{n(g_i)}{N} \log_2 \frac{n(g_i)}{N},$$

$$H_{ij} = -\sum_{g_i=0}^{2}\sum_{g_j=0}^{2} \frac{n(g_i, g_j)}{N} \log_2 \frac{n(g_i, g_j)}{N},$$

where $n(g_i)$, $n(g_i, g_j)$ and N are the number of samples with genotype g_i, that with genotypes g_i and g_j, and total number of samples, respectively. The score S_i is based on the MI value instead of the conventional R^2 value. The MI tends to take a larger value for SNPs with higher MAF unlike R^2 value. An example of comparison between MI and R^2 values is shown in Supplementary Figure 2. While MI calculated between SNPs with high MAFs (0.4 and 0.5; example 1 in Supplementary Figure 2) is higher (MI = 0.82) than that between SNPs with low MAFs (0.10 and 0.12; example 2) (MI = 0.47), R^2 values are almost same ($R^2 = 0.82$) despite considerable difference in MAFs between the examples 1 and 2.

Design of the Japonica array

To maximize the imputation performance in low and common frequencies in Japanese population, the probes on the array should be selected from the ranked tag SNPs in their order in the former section. In parallel, we also cared and included SNPs of special interest or purpose (prioritized SNPs) to probes on the SNP array prior to tag SNPs. The prioritized SNPs include those which are listed in the NHGRI GWAS catalog,[17] pharmacogenomics-related SNPs, high impact SNPs (stop gain and splice site changes) that have been difficult to impute in preliminary analyses, and SNPs of Y chromosome and mitochondria. These SNPs are expected to be useful for replication studies or to complement SNPs with low imputation accuracy. The tag SNPs not listed as prioritized SNPs were then added to the list of probes until the number of probes reached the maximum number that an array product allows (Table 1). The full list of SNPs on the Japonica array is publicly available from our website (http://nagasakilab.csml.org/en/japonica).

Genotyping with the Japonica array

We genotyped 288 individuals arbitrarily selected from the 1KJPN panels with the Japonica array to validate the genotyping performance. The Japonica arrays were produced through Axiom myDesign service (Affymetrix). Two hundred nanograms of genomic DNAs were amplified, fragmented and labeled as per manufacturer's instruction with Nimbus automated system (Hamilton, Reno, NV, USA) controlled by Hamilton Run Control-Axiom (v1.1.0 med, Affymetrix) and Gene Titan Multi-channel instrument operated by AGCC Gene Titan Instrument Control (ver 4.1.0.1567, Affymetrix). The genotype calling was conducted using the Affymetrix Power Tools (version 1.16.1, Affymetrix). The genotype concordance rates were calculated by comparing these genotypes with those obtained from the whole-genome sequence of same individuals.

Table 1 Category of SNPs on the Japonica array

Category	*Number of SNPs*[a]	*Array occupancy rate*
Tag SNPs (including X chromosome)	638 269	96.8%
Pharmacogenomics markers	2028	0.31%
Y chromosome	275	0.04%
Mitochondria	70	0.01%
NHGRI GWAS catalog	10 798	1.64%
HLA	3906	0.59%
Untaggable functional SNPs	3990	0.61%
Total	659 253	—

Abbreviations: GWAS, genome-wide association studies; SNP, single nucleotide polymorphism.
[a]Some SNPs are overlapped among categories.

Imputation

The genotypes of 131 Japanese individuals (independent from the 1070 individuals of the 1KJPN panel) were obtained from WGS with the same sequencing protocol and the same variant-calling pipeline as for constructing the reference panel to assess the imputation performance. The genotypes of the same position on each SNP array were used for imputation and all SNPs were used for the evaluation of imputation performance. We also evaluated the imputation performance using 89 samples of JPT panel, in which the whole-genome sequence have been determined on the 1KGP. The imputations were performed using IMPUTE2[18] (version 2.2.2). For IMPUTE2 options, N_e and k_{hap} were set to 20 000 and 1000, respectively. In addition to the 1KJPN panel, we considered the following reference panels for imputation to evaluate their performance: the reference panel from the 1KGP released in December 2013 containing 1092 cosmopolitans (1KGP); a reference panel of 89 JPT individuals from 1KGP (1KGP_JPT); and a reference panel combining data from the 1KGP and 1KJPN (1KJPN+1KGP). Since 89 JPT samples are part of 1KGP panel, we did not conduct imputation of these samples with 1KGP, 1KGP_JPT or 1KJPN+1KGP panels. To assess the agreement between the imputed genotypes and genotype calls of WGS (HiSeq2500), we calculated the squared Pearson correlation r^2 and the discordant rate for each SNP. The r^2 values are calculated between the genotypes of WGS taking the integer values 0, 1 and 2 and the allele dosages of the imputed genotypes valued from 0–2 as in the study by Howie *et al.*[19] The discordance rate is the fraction of genotypes not matched between the genotypes of NGS and the imputed genotypes with the highest genotype probability. The values of SNP position in which probe is designed was set to be 1.0 and 0.0 for r^2 value and discordant rate, respectively. The MAF for each SNP was calculated for each reference panel independently.

RESULTS

We designed a SNP array consisting of 659,253 SNPs, which is almost the maximum number of SNPs of a single array on the Axiom 96-layout plate. The category of prioritized SNPs and their number are presented in Table 1. Probes in the Japonica array were validated by experimental genotyping of 288 samples from the 1KJPN panel. The average call rate across samples was 99.7% (min. 97.5% and max. 99.8%), and 98.4% of SNPs on the array exceeded the call rate above 97.0%. The average genotype concordance rate between the Japonica array and HiSeq2500 was 99.7% (min. 98.4% and max. 99.8%) across samples, and 99.0% of SNPs on the array exceeded the concordance rate above 97.0%. The genotypes that failed to call or are discordant with NGS call are not apparently shared among samples (Supplementary Figure 3). We also compare the genotype calls between the Japonica Array and Illumina HumanOmni2.5 (Omni2.5) on 289 372 overwrapping sites. The genotyping results of HumanOmni2.5 have been obtained in our previous study.[9] The genotype call was carried out using the Genotyping Module in the GenomeStudio software (ver. 2011.1, Illumina) and the default set cluster file was used. The average concordance rate across samples between the Japonica Array and Omni2.5 was 99.8% (min. 98.7% and max. 99.9%) and 99.2% of SNPs exceeded the concordance rare > 97%. These results demonstrated that the genotype quality of the Japonica array was comparable to the existing SNP arrays not only within same platform[8] but also among platforms.

We compared the imputation performance of the Japonica array to the commercially available SNP arrays (Omni2.5, Illumina

HumanOmniExpressExome (OmniExpressExome) and Axiom Genome-wide ASI1 (AxiomASI)) using 1070 samples of 1KJPN as reference panel. These commercial SNP arrays differ by the number of designed positions and the fraction of polymorphic markers compared with the 1KJPN (Table 2). Nearly all the markers on the Japonica array are polymorphic among the 1KJPN panel as we intended (99.7%), meanwhile a substantial fraction of markers on the other SNP arrays is not polymorphic (that is, it is less informative for imputation as tag SNPs). For example, 31.4% of SNPs on OmniExpressExome was not polymorphic. The imputation performance was evaluated by the average r^2 values stratified by the MAF of a reference panel (Figures 2a and c), the genome-wide coverage of the imputed genotype for different r^2 thresholds (Figures 2b and d), and the average discordance rates between imputed genotype with highest genotype probability and genotypes of WGS (Supplementary Figures 4c–e).

For common SNPs, the imputation quality of the Japonica array using 131 samples of our project (ToMMo131) was higher than OmniExpressExome and AxiomASI in terms of the average r^2 value (Figure 2a). In addition, the r^2 value of the Japonica array is almost comparable to that of Omni2.5 that contains 3.6 times as many markers (Table 2). For instance, the average r^2 values of the SNPs

Table 2 Comparison of the Japonica array with the existing SNP arrays

SNP array	*No. of SNP positions*	*No. of polymorphic positions in 1KJPN*	*Genomic coverage with pairwise $r^2 > 0.8$*
Japonica array	659 253	657 152 (99.7%)	72.4%
HumanOmni2.5S	2 391 739	1 422 455 (59.5%)	71.4%
HumanOmniExpressExome	930 717	638 494 (68.6%)	61.2%
Axiom Genome-wide ASI1	627 781	527 859 (88.9%)	60.0%

Abbreviation: SNP, single nucleotide polymorphism.

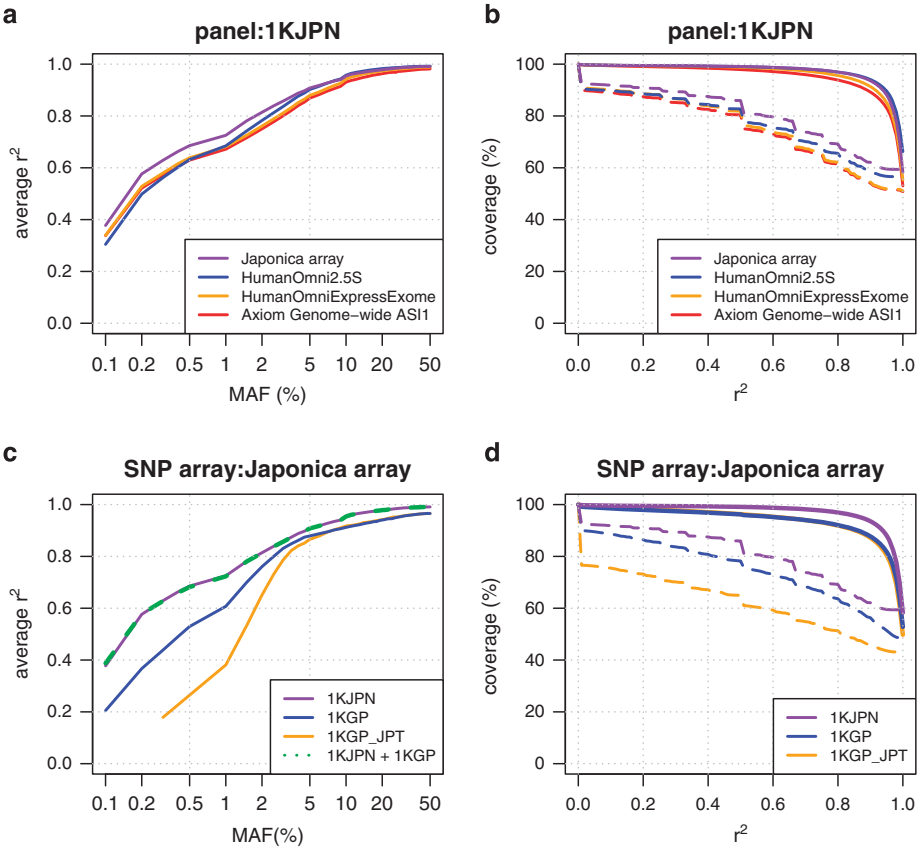

Figure 2 Improvement in imputation accuracy with the Japonica array. Comparison of the imputation accuracy of different SNP arrays using the 1KJPN panel (**a** and **b**) and the imputation accuracy of the Japonica array using different reference panels (**c** and **d**). The imputation was conducted to the 131 individuals (ToMMo131, independent from the 1070 individuals in the 1KJPN panel) using the 1KJPN panel. The average r^2 values are plotted against the MAF (**a** and **c**). The fraction of SNPs in which the genotype was imputed with a given r^2 threshold (*x*-axis) over the total SNPs in the reference panel (genomic coverage) is plotted (**b** and **d**) with solid and dashed lines for common and low-frequency SNPs, respectively. The r^2 value is the squared correlation coefficient between the imputed genotype and the genotype obtained by whole-genome sequencing. MAF, minor allele frequency; SNP, single nucleotide polymorphism.

with MAF >5% were 0.972, 0.975, 0.965 and 0.955 for the Japonica Array, Omni2.5, OmniExpressExome and AxiomASI, respectively. In contrast, for low-frequency SNPs, the imputation quality of the Japonica array were superior to other SNP arrays even when compared with the Omni2.5. The average r^2 values of low-frequency SNPs were 0.802, 0.772, 0.756 and 0.746 for the Japonica Array, Omni2.5, OmniExpressExome and AxiomASI, respectively. Contrary to the r^2 values, the average discordance rate between genotypes of NGS and imputation was higher in Japonica array than Omni2.5 as MAF becomes higher (Supplementary Figure 4c). For example, the average discordance rates were 0.012 and 0.010 for Japonica array and Omni2.5, respectively. This can be explained by the difference in the number of probe-designed SNPs whose discordance rate of SNP was set to be 0.0. The number of such SNPs is larger for Omni2.5 than Japonica Array. Indeed, the discordance rate of common SNP was almost equal (0.013) between Japonica array and Omni2.5 when the probe-designed SNPs were excluded from calculation. The genomic coverage of the Japonica array was higher than the other existing arrays in a broad r^2 threshold especially for low-frequency SNPs (Figure 2b). For common SNPs, the genomic coverage of SNPs with an $r^2>0.8$ was 96.9% for the Japonica array, whereas the coverage of Omni2.5, OmniExpressExome and AxiomASI were 97.0%, 95.6% and 93.9%, respectively. The genomic coverage of low-frequency SNP by the Japonica array (67.2%) was higher than other arrays (63.8% for Omni2.5, 60.0% for OmniExpressExome and 59.4% for AxiomASI). The difference in the genomic coverage by imputation has substantial impact on the absolute number of genotypes, which can be used for downstream analyses, especially for rare and low-frequency SNPs (Table 3). For example, 1 214 767 and 2 077 383 genotypes were imputed from ToMMo131 by the Japonica array for rare and low-frequency SNPs, respectively. This is about 11% larger than those obtained from OmniExpressExome, for example, in which 1 104 194 and 1 854 752 genotypes were imputed for rare and low-frequency SNPs, respectively. Note that these numbers were obtained from 131 samples and the number will increase with the sample size.

It is possible that the imputation performance presented above might be overestimated because individuals of both reference panel (1070 samples) and imputation subject (131 samples) have been recruited at the same region (Miyagi Prefecture, Japan). Thus, we conducted the imputation of 89 samples of HapMap JPT panel (Japanese people in Tokyo) and compared this with those obtained from 131 samples of our project (ToMMo131). The imputation performance was very similar between both samples. For instance, the average r^2 values of 0.976 and 0.810 for common and low-frequency SNPs, respectively, were obtained from the imputation of JPT samples with Japonica array, which is comparable with the average r^2 values (0.972 and 0.802 for common and low-frequency SNPs, respectively) of ToMMo131 samples. This tendency was confirmed with other SNP arrays except for Omni2.5 (Supplementary Figure 4b). The average r^2 of the Japonica array was lower in JPT samples than ToMMo131 samples for low-frequency and rare SNPs, resulting in similar imputation performance with Omni2.5. This is presumably because the tag SNPs of Omni2.5 has been selected from 1KGP panel, which includes the imputation target samples themselves, that is JPT samples.

We next considered the influence of panel selection on the imputation performance. Figures 2c and 2d show the imputation performance of the Japonica array using different reference panels. The 1KJPN panel exhibited better imputation performance compared with the 1KGP and 1KGP_JPT panels, which is consistent with the better imputation efficiency using a closely related reference panel.[20,21] Indeed, the average r^2 values of common SNPs were 0.972, 0.941 and 0.940 for the 1KJPN, 1KGP and 1KGP_JPT, respectively. Difference in the imputation performance by panel selection was more prominent for the low-frequency SNPs. The average r^2 values of low-frequency SNPs were 0.802 for the 1KJPN panel, whereas those for 1KGP and 1KGP_JPT panels were 0.745 and 0.618, respectively. Although the 1KGP_JPT panel consists of haplotypes derived from individuals with Japanese ancestry only, the performance especially for low-frequency SNPs was much worse than the cosmopolitan 1KGP panel, which suggested that the haplotypes in the 1KGP panel (other than those from the JPT) contributed to the genotype imputation. An addition of haplotypes to the 1KJPN panel (that is, 1KJPN+1KGP panel) slightly increased the number of imputed SNPs. For example, 8 278 163 SNPs with $r^2>0.8$ were imputed with 1KJPN+1KGP panel while 8 236 760 SNPs were imputed with the 1KJPN panel. However, the combined panel approach did not substantially affect the imputation performance in terms of r^2 value even though a larger number of haplotypes contained in the panel. The average r^2 of the imputed genotypes of SNPs with MAF>0.5% was almost identical (0.908) between the 1KJPN panel and a combined panel (1KJPN+1KGP) (Figure 2c). In addition, the average discordance rates were also similar between the 1KJPN (0.92%) and 1KJPN+1KGP (0.93%). This is likely due to the huge collection of haplotypes in the 1KJPN panel that includes the haplotypes in the 1KGP panel as a subset.

Table 3 The number of imputed genotype

SNP array	*Rare SNP*[a]	*Low-frequency SNP*[a]	*Common SNP*[a]
Japonica array	1 214 767	2 077 383	4 944 610
HumanOmni2.5S	1 051 158	1 969 616	4 946 935
HumanOmniExpressExome	1 104 194	1 854 752	4 876 863
Axiom Genome-wide ASI1	1 092 543	1 836 323	4 787 601

Abbreviation: SNP, single nucleotide polymorphism.
[a]The number of SNPs with $r^2>0.8$

DISCUSSION

The reference panel 1KJPN is currently comprised of 2140 haplotypes derived from the whole-genome sequences of 1070 Japanese individuals. This is the largest Japanese reference panel to date and contains a large amount of haplotypes that are presumably shared among individuals with Japanese ancestry.

We designed a SNP array suitable for genotype imputation using the 1KJPN panel, termed the 'Japonica array.' The genotype quality of the Japonica array was experimentally validated to be as high as the existing commercial SNP arrays. Nonetheless, we demonstrated that the imputation quality of the Japonica array outperformed the commercially available SNP arrays when applied to Japanese samples. There are two reasons for improvement in imputation quality. First, we selected the SNPs on the Japonica array so that the vast majority of them are polymorphic in the Japanese population by referring to the allele frequencies of SNPs on the 1KJPN reference panel. Indeed, 99.6 % of the SNPs on the Japonica array are polymorphic, which is comparable to 59.5% on the HumanOmni2.5, 68.6% on the OmniExpressExome and 88.9% on the AxiomASI. More importantly, our strategy for tag SNP selection enabled us to capture the highest number of SNPs on the 1KJPN panel as possible. Indeed, the genomic coverage of the tag SNPs (pairwise linkage disequilibrium $R^2>0.8$) was also larger compared with other SNP arrays (Table 1).

We excluded SNPs with MAF < 0.5% from the tag SNP selection to avoid poor cluster separation in genotyping process. In this study, we defined a new score *S* (equation (1)) for tag SNP selection on the basis of the MI, which has been used as a linkage disequilibrium measure instead of conventional R^2 value in the previous study.[22] The MI tends to yield lower value when calculating between low-frequency SNPs in comparison to R^2 value (Supplementary Figure 2). This property would allow us to select higher frequency SNPs, which are expected to improve genotype calls by good cluster separation. Indeed, the relative frequency of rare (MAF < 0.5%) SNPs on the Japonica array was considerably lower than other SNPs (Supplementary Figure 5a). However, the relative frequency of imputed genotype is higher when MAF becomes lower (Supplementary Figure 5b). This implies that the tag SNP selection strategy in this study is effective for the imputation of rare SNPs despite the array containing few probes that directly interrogate rare SNPs.

We evaluated the quality of imputation by comparing the imputed genotypes (or allele dosage) and the genotypes obtained from high coverage (32.4 on average) whole-genome sequences for 131 individuals, which were different from the 1070 individuals in the 1KJPN reference panel. We also conducted the imputation of 89 JPT samples. We then found that the imputation quality was very close to that of 131 samples of our project. These imputations enabled us to assess the accuracy of the imputed genotypes in a whole-genome scale, which is a close situation as actual GWAS. We showed that the Japonica array exhibited better imputation performance from other existing commercial SNP arrays when the haplotypes of the 1KJPN were used as the reference panel. Intriguingly, the imputation quality of the Japonica array also outperformed the other existing commercial SNP arrays even when the 1KGP reference panel was used (Supplementary Figure 4f), indicating that the tag SNPs on the Japonica array effectively captured the haplotypes in the Japanese population irrespective of reference panel in compared with the existing arrays.

Our study showed that the 1KJPN panel is better than the 1KGP panel for the genotype imputation of Japanese samples. This is consistent with previous reports where a population-specific reference panel improved the accuracy of genotype imputation especially for low-frequency and rare variants.[20,21] Almost no improvement was observed in imputation performance with a combined reference panel of 1KJPN and 1KGP (1KJPN+1KGP) compared with the 1KJPN panel in terms of the average r^2 value and the discordance rate. This result is consistent with the Genome of Netherland study,[21,23] which reported that adding haplotypes of the 1KGP panel to a population-specific reference panel (GoNL) had small effects on the imputation quality when Dutch samples were imputed. This result is likely because the larger reference panel (that is, 1KJPN or GoNL) contains the majority of haplotypes in the smaller reference panel (1KGP_JPT or European ancestry panel of 1KGP). This tendency would be prominent for SNPs with lower allele frequencies because such SNPs are population specific.[19]

The development of population-specific SNP arrays will facilitate genome-wide studies inquiring into the genetic basis of complex diseases and traits. In this study, we demonstrated that whole-genome imputation using the Japonica array in combination with the 1KJPN panel was an efficient method to fully utilize the genetic resources of a genome cohort study for downstream studies, such as GWAS. Finally, this approach, a combination of WGS and population-specific SNP arrays, will be applicable to other studies in diverse ethnic groups.

CONFLICT OF INTEREST

YK, TM, KK, NN and MN have a patent pending based on the work reported in this paper. Genotyping service of the Japonica array is provided by Toshiba Corporation under the license from Tohoku University. RS is currently employed by Toshiba Corporation. MN and KK hold the concurrent post at Department of Cohort Genome Information Analysis endowed by Toshiba Corporation. MN received research funding from Toshiba Corporation. The remaining authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported (in part) by the Tohoku Medical Megabank Project (Special Account for reconstruction from the Great East Japan Earthquake). This work was supported by The Center of Innovation Program from Japan Science and Technology Agency, JST. All computational resources were provided by the ToMMo supercomputer system. We are grateful to Takanori Hasegawa for helpful discussion. We are indebted to all volunteers who participated in this ToMMo project. We also thank all other members of ToMMo Japanese Reference Panel Project.

1 Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10**, 387–406 (2009).
2 Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
3 Holm, H., Gudbjartsson, D. F., Sulem, P., Masson, G., Helgadottir, H. T., Zanon, C. *et al.* A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat. Genet.* **43**, 316–320 (2011).
4 Steinthorsdottir, V., Thorleifsson, G., Sulem, P., Helgason, H., Grarup, N., Sigurdsson, A. *et al.* Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat. Genet.* **46**, 294–298 (2014).
5 The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
6 Jonsson, T., Atwal, J. K., Steinberg, S., Snaedal, J., Jonsson, P. V., Bjornsson, S. *et al.* A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature* **488**, 96–99 (2012).
7 Liu, E. Y., Buyske, S., Aragaki, A. K., Peters, U., Boerwinkle, E., Carlson, C. *et al.* Genotype imputation of Metabochip SNPs using a study-specific reference panel of ~4,000 haplotypes in African Americans from the Women's Health Initiative. *Genet. Epidemiol.* **36**, 107–117 (2012).
8 Hoffmann, T. J., Kvale, M. N., Hesselson, S. E., Zhan, Y., Aquino, C., Cao, Y. *et al.* Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. *Genomics* **98**, 79–89 (2011).
9 Nagasaki, M., Yasuda, J., Katsuoka, F., Nariai, N., Kojima, K., Kawai, Y. *et al.* Rare variant discovery by deep whole-genome sequencing of 1070 Japanese individuals. *Nat. Commun.* (in press).
10 Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
11 Katsuoka, F., Yokozawa, J., Tsuda, K., Ito, S., Pan, X., Nagasaki, M. *et al.* An efficient quantitation method of next-generation sequencing libraries by using MiSeq sequencer. *Anal. Biochem.* **466**, 27–29 (2014).
12 Sato, Y., Kojima, K., Nariai, N., Yamaguchi-Kabata, Y., Kawai, Y., Takahashi, M. *et al.* SUGAR: graphical user interface-based data refiner for high-throughput DNA sequencing. *BMC Genomics* **15**, 664 (2014).
13 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
14 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
15 Kojima, K., Nariai, N., Mimori, T., Yamaguchi-Kabata, Y., Sato, Y., Kawai, Y. *et al.* HapMonster: a statistically unified approach for variant calling and haplotyping based on phase-informative reads. *Lect. Notes Comput. Sci.* **8542**, 107–118 (2014).
16 Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
17 Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
18 Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529–e1000529 (2009).
19 Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1**, 457–470 (2011).
20 Deelen, P., Menelaou, A., van Leeuwen, E. M., Kanterakis, A., van Dijk, F., Medina-Gomez, C. *et al.* Improved imputation quality of low-frequency and rare variants

in European samples using the 'Genome of The Netherlands'. *Eur. J. Hum. Genet.* **22**, 1321–1326 (2014).
21 Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K. *et al.* The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**, 327–332 (2014).
22 Liu, Z. & Lin, S. Multilocus LD measure and tagging SNP selection with generalized mutual information. *Genet. Epidemiol.* **29**, 353–364 (2005).
23 The Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46**, 818–825 (2014).

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