


RESEARCH ARTICLE

Open Access



# Regional genetic differences among Japanese populations and performance of genotype imputation using whole-genome reference panel of the Tohoku Medical Megabank Project

Jun Yasuda<sup>1\*</sup> , Fumiki Katsuoka<sup>1</sup>, Inaho Danjoh<sup>1</sup>, Yosuke Kawai<sup>1,9</sup>, Kaname Kojima<sup>1</sup>, Masao Nagasaki<sup>1</sup>, Sakae Saito<sup>1</sup>, Yumi Yamaguchi-Kabata<sup>1</sup>, Shu Tadaka<sup>1</sup>, Ikuko N. Motoike<sup>1</sup>, Kazuki Kumada<sup>1</sup>, Mika Sakurai-Yageta<sup>1</sup>, Osamu Tanabe<sup>1</sup>, Nobuo Fuse<sup>1</sup>, Gen Tamiya<sup>1</sup>, Koichiro Higasa<sup>2</sup>, Fumihiko Matsuda<sup>2</sup>, Nobufumi Yasuda<sup>3</sup>, Motoki Iwasaki<sup>4</sup>, Makoto Sasaki<sup>5,6</sup>, Atsushi Shimizu<sup>6</sup>, Kengo Kinoshita<sup>1,7</sup> and Masayuki Yamamoto<sup>1,8\*</sup>

## Abstract

**Background:** Genotype imputation from single-nucleotide polymorphism (SNP) genotype data using a haplotype reference panel consisting of thousands of unrelated individuals from populations of interest can help to identify strongly associated variants in genome-wide association studies. The Tohoku Medical Megabank (TMM) project was established to support the development of precision medicine, together with the whole-genome sequencing of 1070 human genomes from individuals in the Miyagi region (Northeast Japan) and the construction of the 1070 Japanese genome reference panel (1KJPN). Here, we investigated the performance of 1KJPN for genotype imputation of Japanese samples not included in the TMM project and compared it with other population reference panels.

**Results:** We found that the 1KJPN population was more similar to other Japanese populations, Nagahama (south-central Japan) and Aki (Shikoku Island), than to East Asian populations in the 1000 Genomes Project other than JPT, suggesting that the large-scale collection (more than 1000) of Japanese genomes from the Miyagi region covered many of the genetic variations of Japanese in mainland Japan. Moreover, 1KJPN outperformed the phase 3 reference panel of the 1000 Genomes Project (1KGp3) for Japanese samples, and 1KJPN showed similar imputation rates for the TMM and other Japanese samples for SNPs with minor allele frequencies (MAFs) higher than 1%.

**Conclusions:** 1KJPN covered most of the variants found in the samples from areas of the Japanese mainland outside the Miyagi region, implying 1KJPN is representative of the Japanese population's genomes. 1KJPN and successive reference panels are useful genome reference panels for the mainland Japanese population. Importantly, the addition of whole genome sequences not included in the 1KJPN panel improved imputation efficiencies for SNPs with MAFs under 1% for samples from most regions of the Japanese archipelago.

**Keywords:** Genome reference panel, Genotype imputation, Population genetics, Japan

\* Correspondence: [jyasuda@megabank.tohoku.ac.jp](mailto:jyasuda@megabank.tohoku.ac.jp);  
[masiyamamoto@med.tohoku.ac.jp](mailto:masiyamamoto@med.tohoku.ac.jp)

<sup>1</sup>Sendai, Tohoku Medical Megabank Organization, 2-1, Seiryō-machi, Aoba-ku, Tohoku Medical Megabank, Tohoku University, Sendai 980-8573, Miyagi, Japan

Full list of author information is available at the end of the article



## Background

Genotype imputation is an important step in current genome-wide association studies. Imputation accuracy, as well as genomic coverage of highly accurate imputed genotypes, confers elevated statistical power in association tests. [1] The choice of a haplotype reference panel to maximize imputation performance has often been debated. [2–4] Haplotype reference panels are used to identify haplotypes of individual genomes genotyped by single-nucleotide polymorphism (SNP) arrays, and then to estimate the genotypes missing in the SNP array data. Thus, to enable high-density genotype imputation for SNPs with minor allele frequencies (MAFs) >1% in a population, reference panels are constructed preferably based on the whole-genome sequencing (WGS) of large samples. The influence of panel selection on imputation accuracy in terms of panel size and ancestry matching between the panel and study samples has been assessed by cross-validation. [5] The results showed that better imputation performances were achieved when more samples from various populations were included in the reference panel. Thus, along with improved algorithms for genotype imputation using large panels, great efforts are being made to construct large reference panels for highly accurate genotype imputation, such as that by the Haplotype Reference Consortium. [4] In addition, several cohort studies have been conducted using WGS to construct better, more detailed haplotype reference panels. [2, 6] These studies suggest that increasing the sample sizes of population-specific haplotype reference panels is more effective for improving genotype imputation accuracy than aggregating the haplotype collection from worldwide resources, because the focus is then on specific populations. Although recent studies in human population genetics have revealed clear regional variation in haplotype diversity, even within a single population, [7] the influence of such variation on imputation performance has not yet been assessed.

The Tohoku Medical Megabank (TMM) project was launched in 2011 to investigate effects in the aftermath of the Great East Japan Earthquake in the Miyagi and Iwate prefectures (Northeast Japan). The TMM project developed prospective cohorts in these two prefectures [8] with the aim of aiding in the establishment of precision medicine in this region. To contribute to this, the WGS of 1000 human genomes from individuals in the Miyagi region was undertaken (beginning in 2015) and a 1070 Japanese genome reference panel containing haplotype information was constructed using this data (the 1KJPN panel). [9] In a previous study, we reported the imputation performance using the 1KJPN panel was better than the performance using the 1000 Genomes Project phase 1 panel [10]. We also designed a custom SNP array for a Japanese population (the Japonica array). [11]

The 1KJPN panel consists of haplotypes derived from a cohort of participants in the Miyagi Prefecture, which has approximately 5% of Japan's total population. However, it is not known how much region-specific sampling for a reference panel affects the performance of genotype imputation for samples collected nationwide.

In this study, we evaluated the imputation performance of the 1KJPN panel for participants from three areas other than Miyagi, namely, Iwate, Nagahama (Kinki region, south-central Japan), and Aki (Shikoku Island) (Fig. 1a). We also extended the 1KJPN panel by adding haplotype information from the samples from these three other areas to assess imputation performance given the haplotype variations across the different areas.

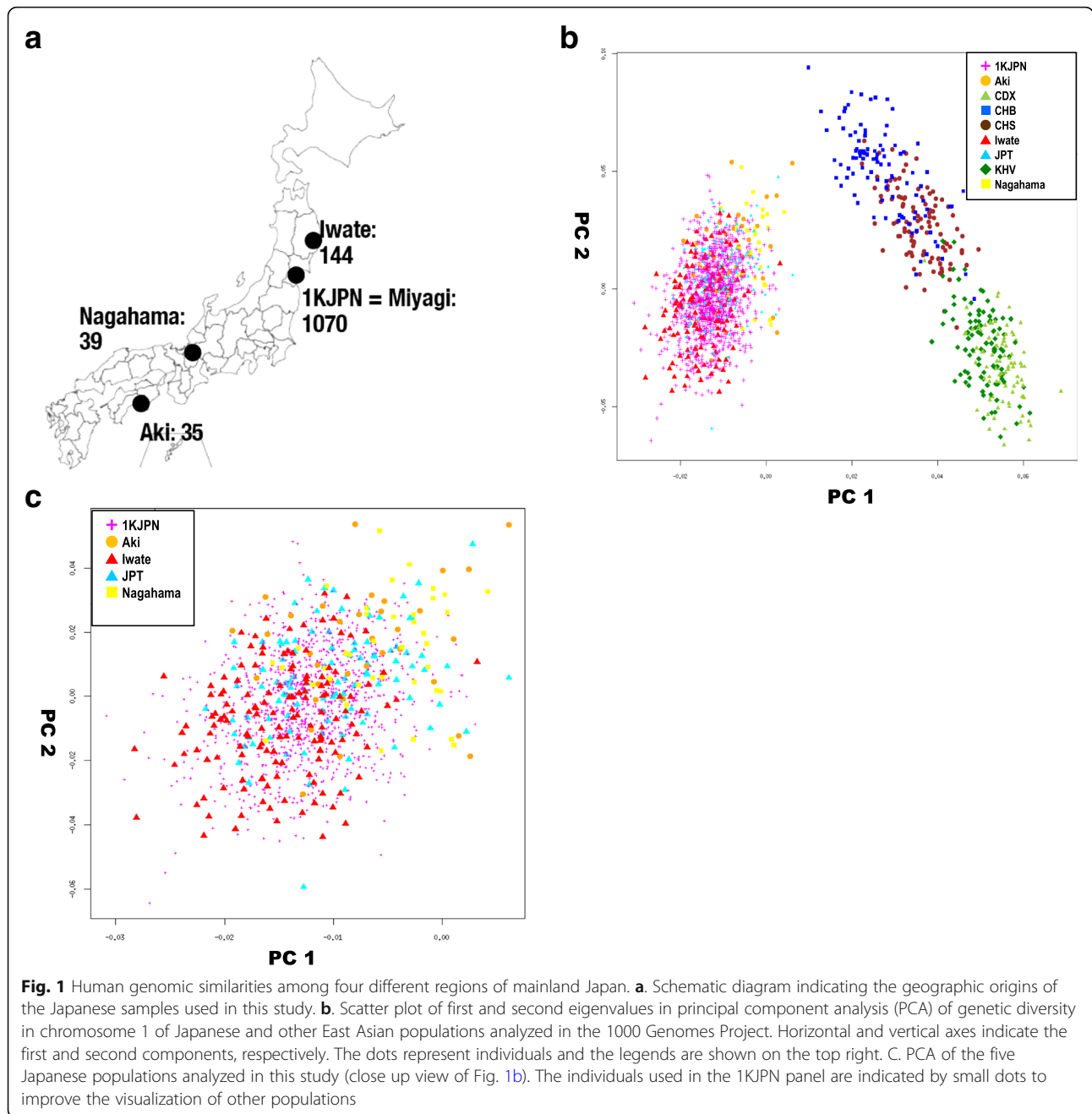
## Results

### Customized reference panel construction

We performed WGS for all the available samples from Iwate, Nagahama, and Aki to construct an extended haplotype reference panel (1KJPN+ panel; Fig. 1a). After removing one Aki sample because of the cryptic relatedness of another member of the cohort, the 1KJPN+ panel consisted of 2560 haplotypes from 1280 samples (1070, 136, 39, and 35 samples from Miyagi, Iwate, Nagahama, and Aki, respectively). We also compared our panels with the phase 3 panel of the 1000 Genomes Project [12] (1KGPp3 panel).

### Genetic diversity of Japanese and other east Asian populations

We compared the diversity of Japanese populations (namely, the Miyagi cohort used to construct 1KJPN and other Japanese populations) with the diversity of populations from elsewhere in East Asia to determine how 1KJPN might reflect these populations. Principal component analysis (PCA) plots with 35,596 SNP genotypes (not indels) on chromosome 1 (see methods) are shown in Fig. 1b. The proportion of variance explained by the first and second principal components was 15.4 and 3.53%, respectively. Japanese individuals from Aki, Iwate, and Nagahama who were newly added to the dataset were clustered with the Miyagi population (= 1KJPN) but separated from other East Asian populations analyzed in the 1000 Genomes Project. Indeed, the Miyagi samples overlapped with most of the other Japanese populations, as shown in Fig. 1c, which is a magnified view of the Japanese populations in Fig. 1b. This indicates that the 1KJPN population is sufficiently similar to populations elsewhere in Japan and that 1KJPN can be used as genomic data representative of the whole Japanese population in mainland Japan.



### Genetic diversity in Miyagi and other parts of Japan

The genetic differentiation of samples in four parts of Japan was investigated in terms of fixation index ( $F_{ST}$ ) and haplotype sharing. Although the PCA did not provide sufficient resolution to separate the samples into the four discrete areas (Fig. 1c), the  $F_{ST}$  values were in good agreement with the geographic locations (Table 1). For instance, the  $F_{ST}$  value between Miyagi and Iwate, which are adjacent prefectures, was the smallest (0.000345) among all pairs of areas examined, which corresponds to the findings in a recent study. [13] No pair

**Table 1** Fixation index ( $F_{ST}$ ) estimation\* of genetic differentiation of samples in four parts of Japan

	Iwate	Nagahama	Aki
Miyagi	0.000345	0.000380	0.00154
Iwate	–	0.000876	0.00192
Nagahama		–	0.000920

\*The  $F_{ST}$  values are based on the SNPs in chromosome 1

of  $F_{ST}$  values among the four regions exceeded the  $F_{ST}$  between the CHB (Han Chinese in Beijing) population in the 1000 Genome Project and our Japanese samples ( $F_{ST} = 0.00777$ ; Additional file 1: Table S1), indicating that 1KJPN may be an appropriate reference panel for the population of mainland Japan.

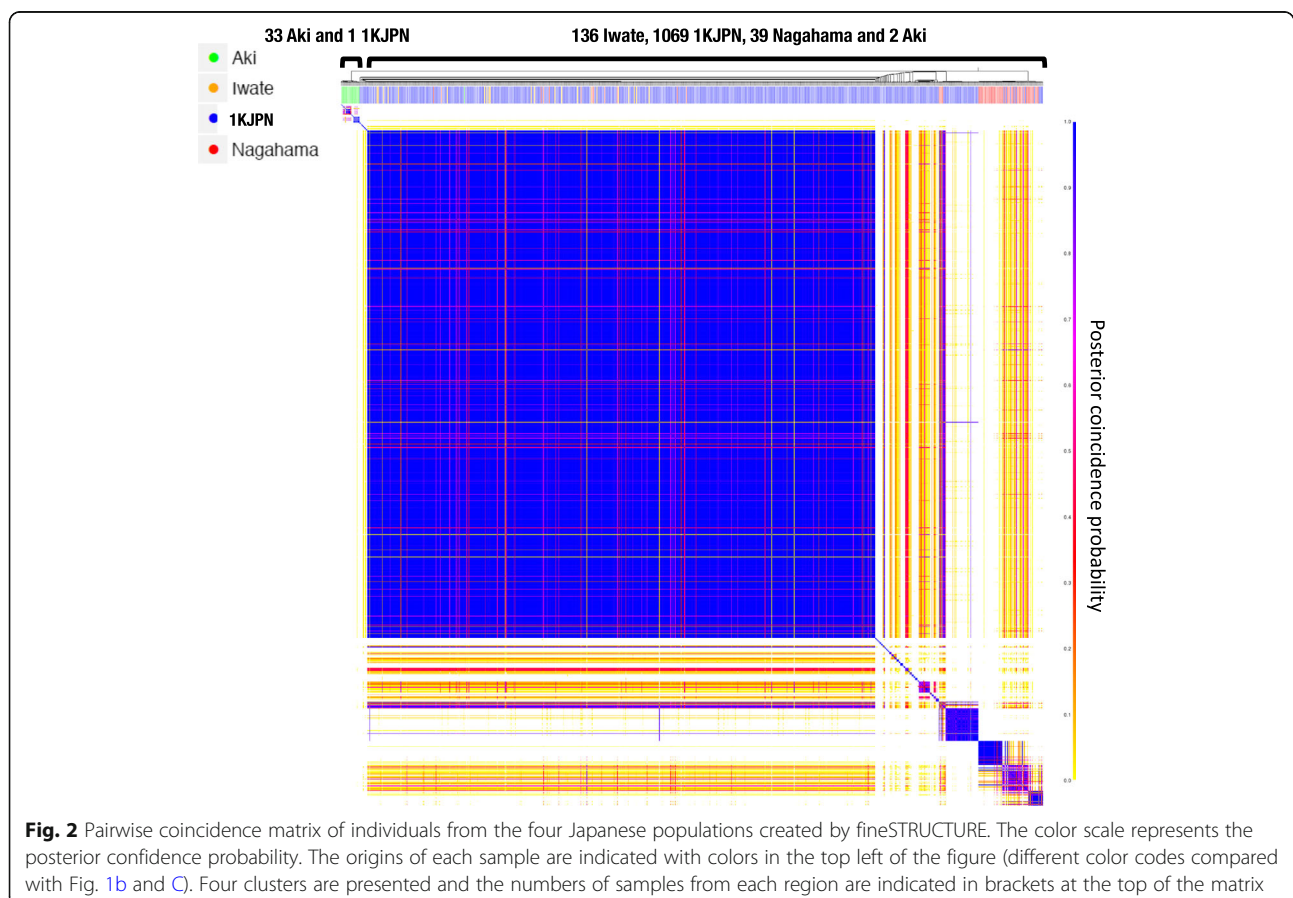
When we analyzed all the 1070 Miyagi samples using fineSTRUCTURE [14], the samples from Miyagi and Nagahama were assigned to the same cluster, whereas most of the samples from Aki formed a distinct cluster (Fig. 2). In this analysis the majority of Aki samples formed a distinct cluster from the other populations, indicating the Aki population from Shikoku Island may be somewhat distinct from the Honshu populations (Iwate, Miyagi, and Nagahama; Fig. 1a). These results show that, in Japan, a large collection of genomic DNA samples from a region the size of a prefectural (around 2% of the Japanese population) may have genetic diversity similar to that of the Japanese population as a whole.

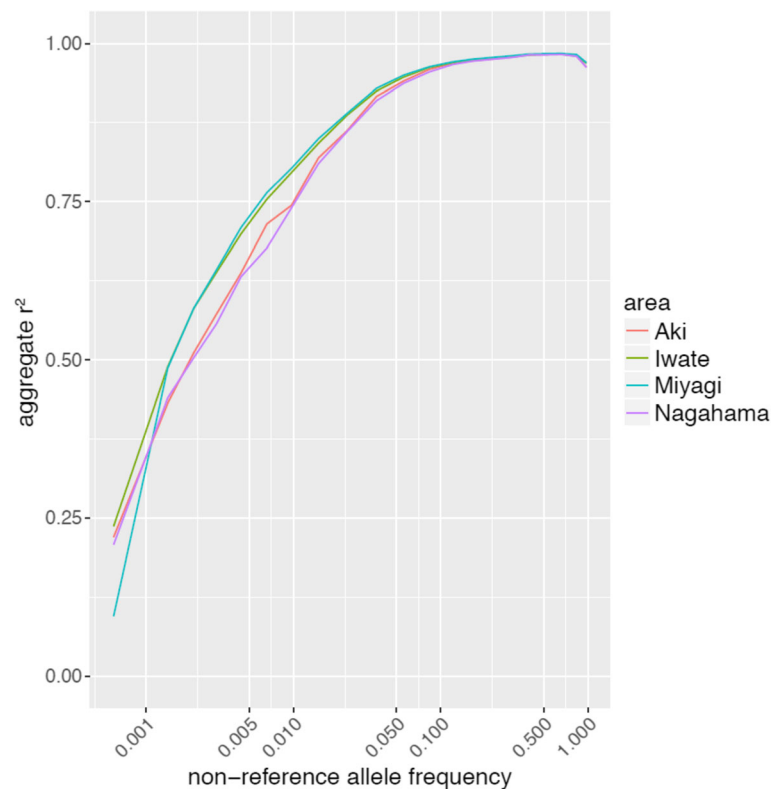
Moreover, we chose 288 of the 1070 samples from Miyagi residents whose maternal grandmother was also born in Miyagi Prefecture to analyze haplotype sharing among the four Japanese regions using fineSTRUCTURE [14] (Additional file 2: Figure S1). We identified four

clusters and found that the cluster positions corresponded to the geographic relationship among the regions. For example, cluster A consisted of samples from Iwate and Miyagi, which are adjacent prefectures, whereas cluster B was dominated by Miyagi samples with small numbers of Iwate and Nagahama samples. However, further investigation is needed to clarify whether the cluster separation among the regions, as shown in Additional file 2: Figure S1, comes from the simple reduction of analyzed individuals, the increase of the Miyagi-specific population, or both.

### 1KJPN genotype imputation efficiencies and effects of genetic differences

To evaluate the difference in imputation performance of the samples from different regions, the aggregate of Pearson's correlation coefficient ( $r^2$ ) values of imputed variants was compared among samples from different areas (Fig. 3). Imputation accuracies among samples from the four areas were comparable ( $r^2 = 0.9891$ , 0.9881, 0.9875, and 0.9878 for Miyagi, Iwate, Nagahama, and Aki, respectively) for common SNPs (non-reference allele frequency > 5%), whereas the accuracies differed among samples from these areas for lower frequency





**Fig. 3** Imputation accuracies using 1KJPN data for Japanese populations not included in the 1KJPN panel. Plot of the imputation accuracy (vertical axis, aggregate  $r^2$  value) against the non-reference allele frequency of reference panel (horizontal axis) when the 1KJPN panel was used as the haplotype reference. Each population is indicated by a different color. Each point on the curves is the average of the corresponding allele frequency bin

variants. For example, for rare variants in 1KJPN (non-reference allele frequency  $\leq 1\%$ ), the  $r^2$  values were 0.7123, 0.7067, 0.6312, and 0.6486 for Miyagi, Iwate, Nagahama, and Aki, respectively.

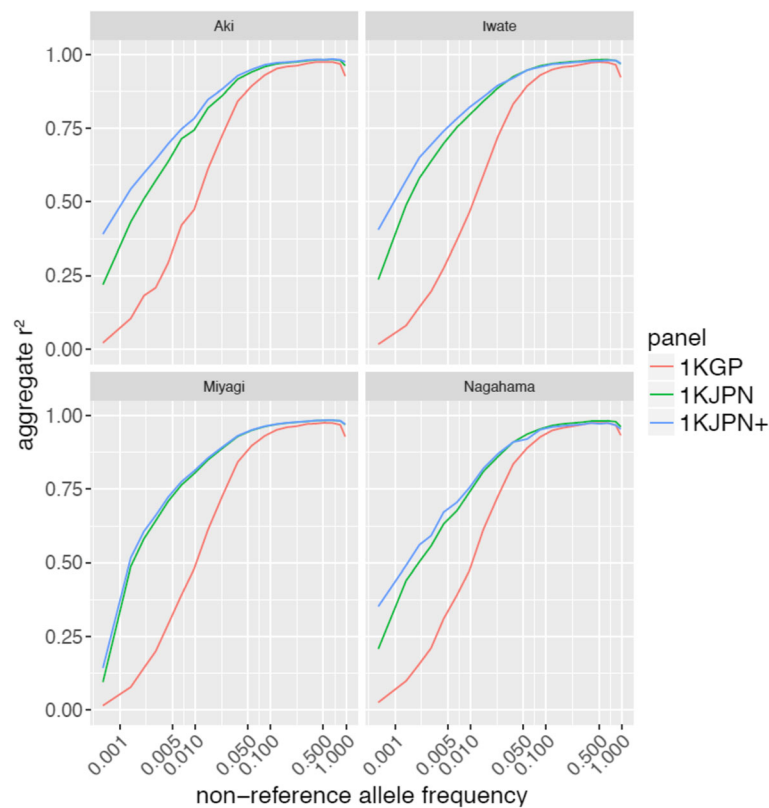
Genotype imputation is difficult for rare variants, so this decrease in aggregate  $r^2$  values was expected. However, SNPs with about 1% MAFs were efficiently imputed with 1KJPN for the Nagahama and Aki samples (Fig. 3;  $r^2$  values around 0.75). These results indicate that 1KJPN is adequate for use as a reference panel for populations from mainland Japan.

To assess the improvement in imputation accuracy gained by adding haplotypes from a different population, genotype imputation accuracies were compared among different reference panels as follows: 1KJPN consisting of 1070 Miyagi samples; 1KJPN+ consisting of 1KJPN and the three other Japanese populations analyzed in this study; and 1KGp3, the phase 3 panel of the 1000 Genome Project (Fig. 4). Imputation accuracy with the 1KJPN+ panel was better than that with the 1KJPN panel and the improvement was prominent for rare variants. The  $r^2$  of rare variants with 1KJPN+ improved to 0.7269, 0.7485, 0.6639, and 0.7037 for Miyagi, Iwate,

Nagahama, and Aki, respectively, compared with the  $r^2$  obtained with 1KJPN. For the three reference panels, the imputation coverage (proportion of variants with confidence: imputation  $r^2 \geq 0.8$ ) across different non-reference allele frequency bins as well as the number of variants that were efficiently imputed are shown in Additional file 3: Table S2. Our results suggest that the addition of samples from other parts of Japan is necessary to further improve 1KJPN as a reference panel for the entire Japanese population. On careful examination, we found several flip-flops occurred between 1KJPN and 1KJPN+ for alleles with high frequencies; for example, Nagahama samples with MAFs around 0.050 (Fig. 4). This may be caused by the design of 1KJPN, which targets the imputation of SNPs with relatively low frequencies (MAF  $\geq 0.5\%$ ). [11].

#### SNPs that differentiate between Miyagi and Nagahama or Aki

The efficiency of 1KJPN as a reference panel for genotype imputation for most Japanese has been shown, [9] and is considered to sufficiently cover SNPs with about 1% MAFs in the Japanese population. However, two medically relevant questions remain unanswered, namely, how many



**Fig. 4** Differences in imputation accuracies using reference panels for Japanese populations. Vertical axis indicates the  $r^2$  values and horizontal axis indicates the minor allele frequencies of the SNPs. Sample regions analyzed are indicated at the top of each panel

biologically relevant SNPs exist among populations in the various regions of Japan, and how different are the MAFs of critical SNPs among Japanese from different regions where the populations might have different genetic disease susceptibilities? To answer these questions, we listed the SNPs that segregate in 1KJPN and exhibit statistically significant differences ( $p < 10^{-7}$ ) in allele frequencies between Miyagi and Nagahama or Aki (Table 2). We detected fewer than 10 SNPs that were present in the Nagahama or Aki population but absent in the Miyagi population (1KJPN), and these SNPs had no exonic variants (Table 2).

Deep-coverage WGS (approximately 30×) of an individual genome can detect very rare variants that were not identified previously. We collected the SNPs not found in 1KJPN (Miyagi population) but found multiple times among the Iwate, Nagahama, and Aki populations (Table 3). The average numbers of such non-1KJPN SNPs were higher in the Aki and Nagahama populations than in the Iwate population, which corresponds to the  $F_{ST}$  data (Table 1, Fig. 1c). This indicates that the addition of samples from distant areas was more effective for collecting rare variants than the addition of samples from neighboring areas. Among the rare variants,

functionally important SNPs that caused amino acid changes or premature stop codons were very rare (0.05 to 0.26 SNPs per genome). The data indicate that there are few deleterious mutations in Japanese from other parts of the country that are not found in 1KJPN. In other words, 1KJPN may be sufficiently comprehensive for genotype imputation for most genome-wide association studies on the population of mainland Japan.

## Discussion

We evaluated the influence of genetic diversity on the accuracy of genotype imputation among populations from different parts of Japan. Previous studies reported clear genetic differentiation between individuals from Okinawa (Ryukyu area) and individuals from the rest of Japan (Hondo), but genetic differentiation among local regions in the Hondo area has been reported to be very low and not to show distinct clusters. [15, 16] In this study, however, we found genetic clusters separated in accordance with geographic location (Miyagi, Iwate, Nagahama, and Aki) using a haplotype-based statistical method [14] (Fig. 1b). Among these areas, genetic diversity was shown to be correlated with geographic distance; for example, the Miyagi and Iwate populations

**Table 2** SNPs that differentiate between Miyagi and Nagahama of Aki based on MAF differences

test	rsid (dbSNP 138)	Chr	Pos (hg19)	Miyagi	Nagahama	Aki	p-value	Annotation
Nagahama-Miyagi	rs1899621	3	157280172	0.03178	0.1795	0.0625	4.44.E-07	PQLC2L: Intronic
Nagahama-Miyagi	rs9501875	6	2685970	0.09331	0.2949	0.04688	8.41.E-07	MYLK4: Intronic
Nagahama-Miyagi	rs148081741	6	135758259	0.02944	0.2051	0.01562	4.32.E-09	AHI1: Intronic
Nagahama-Miyagi	rs141380643	6	135812869	0.02897	0.2051	0.01562	3.54.E-09	AHI1: Intronic
Nagahama-Miyagi	chr7:144127237:C:T	7	144127237	0.0323	0.1795	0.1406	5.31.E-07	intergenic
Nagahama-Miyagi	rs55897843	10	28739479	0.01731	0.1538	0	4.98.E-08	LOC105376468: Intronic
Nagahama-Miyagi	chr11:32314153:G:A	11	32314153	0.03087	0.1795	0.09375	3.26.E-07	intergenic
Nagahama-Miyagi	chr16:47678044:A:T	16	47678044	0.01457	0.1282	0.0625	7.88.E-07	PHKB: Intronic
Nagahama-Miyagi	rs8094961	18	14267309	0.009615	0.1154	0	3.69.E-07	intergenic
Nagahama-Miyagi	rs57064200	21	46286788	0.3396	0.6282	0.371	3.86.E-07	PTTG1IP: Intronic
Aki-Miyagi	rs117933761	1	100267335	0.01402	0.03846	0.1562	8.90.E-08	intergenic
Aki-Miyagi	rs4922078	8	19512537	0.3312	0.359	0.6406	7.15.E-07	CSGALNACT1: Intronic
Aki-Miyagi	rs9423657	10	5607370	0.02682	0.03846	0.1875	3.33.E-07	LOC105376381: Intronic
Aki-Miyagi	rs10899501	11	78131408	0.4565	0.4359	0.1452	3.96.E-07	intergenic
Aki-Miyagi	chr13:103019948:C:T	13	103019948	0.0565	0.1923	0.25	7.78.E-07	FGF14: Intronic
Aki-Miyagi	rs118020607	14	99987537	0.06232	0.1026	0.2656	5.24.E-07	CCDC85C: Intronic
Aki-Miyagi	rs59993898	15	80826474	0.009813	0.01282	0.125	8.60.E-07	ARNT2: Intronic
Aki-Miyagi	rs150711498	18	21327345	0.008879	0.01282	0.125	4.66.E-07	LAMA3: Intronic
Aki-Miyagi	rs11669387	19	33999497	0.1076	0.141	0.3438	7.66.E-07	PEPD: Intronic

were genetically closer than any other pair of areas. The differences in imputation accuracy with the 1KJPN panel among samples from these regions (Fig. 3) were also consistent with this diversity. Because the 1KJPN panel contains samples only from the Miyagi area, more haplotype segments were shared with this area than with other regions. Notably, the imputation accuracy of the Iwate samples was very close to that of the Miyagi samples, even though the Iwate samples are not included in the 1KJPN panel. This is consistent with another report showing that genetic similarities among subpopulations were correlated with geography on the Japanese archipelago. [13] These observations support the idea that ancestry matching between the subjects of genotype imputation and the donors of genomic data for a

whole-genome reference panel is effective for improving imputation accuracy, especially for low-frequency and rare variants.

We demonstrated that haplotypes specific to samples in a local area had substantial impact on imputation performance, especially for low-frequency and rare variants. As mentioned above, genetic differentiation within the Hondo population was small in terms of SNP frequency, but apparent when haplotype sharing between samples was considered. Because imputation algorithms essentially rely on haplotype sharing between the reference panel and the study samples, the genetic differentiation between 1KJPN and other Japanese regions might have been substantial. Our results show that the imputation accuracy for common variants was only marginally affected by the combination of the panel and the area (Fig. 3), and by the addition of region-specific haplotypes to the panel (Fig. 4), suggesting that the common haplotypes contained in the 1KJPN panel cover the haplotype diversity of the Hondo area. However, the imputation accuracy of low frequency and rare variants improved when area-specific haplotypes were added to the 1KJPN panel. This means that long-persisting yet rare haplotypes may exist in each area, and that the imputation accuracy can be improved when matching the haplotype panel and samples in that area. These results provide important information for future extension of haplotype reference panels in population cohorts.

**Table 3** Numbers of SNPs found in three populations but not found in 1KJPN (Miyagi population)

Type	Iwate (per person)	Nagahama (per person)	Aki (per person)
Total SNPs (AC >=2)*	113,541 (834.86)	33,067 (847.87)	44,634 (1275.26)
Exonic	1596 (11.74)	411 (10.54)	593 (16.94)
Mis sense	1032 (7.59)	282 (7.23)	366 (10.46)
Stop gain	14 (0.10)	2 (0.05)	9 (0.26)

AC allele counts in the three regions

## Conclusions

Our data suggest that 1KJPN can cover most of the variants found in the samples from other areas in the Japanese mainland outside of Miyagi and that 1KJPN can be used as a representative of the Japanese population's genomes, making it is useful genome reference panel for other parts of the Japanese mainland. We also showed that the addition of samples not included in 1KJPN improved imputation efficiencies for SNPs with MAFs under 1% from most of the Japanese archipelago.

## Methods

### Sample preparation

The haplotype reference panel (1KJPN panel) was constructed from the whole-genome sequences of 1070 participants from the prospective cohort study of the TMM project. [9] All samples in this panel were obtained from individuals recruited in the Miyagi Prefecture. In the present study, we added samples from individuals recruited in the Iwate Prefecture to our cohort, as well as samples from age- and sex-stratified random samples from external cohorts for comparison, namely, the Nagahama study [17, 18] and the Aki area from the JPHC-NEXT study [19]. WGS and SNP array genotyping were conducted for 136, 39, and 36 samples from the Iwate, Nagahama, and Aki cohorts, respectively. Participants from these cohorts provided written informed consent to undergo WGS in the collaboration studies. For the WGS, 2.5  $\mu\text{g}$  of DNA was dissolved in TE buffer (Tris pH 8.0 10 mM, EDTA 1 mM) or distilled water (100 ng/ $\mu\text{L}$ ). Aliquots of 500 ng were prepared for the SNP array analyses. Detailed WGS methods followed previous studies. [9, 20] SNP array genotyping was performed using Japonica arrays (Toshiba Corporation, Tokyo, Japan).

### Construction of haplotype reference panel

The haplotype reference panel was constructed from WGS data. [9] We constructed an extended haplotype reference panel (1KJPN+ panel) consisting of newly sequenced samples from Iwate, Nagahama, and Aki cohorts using the method described previously, [9] in addition to the 1070 samples originally included in the 1KJPN panel. Variant calling with filtering and haplotype phasing were according to the methods used to construct the 1KJPN panel. [9] Briefly, read mapping and genotype calling were performed using Bowtie2 (version 2.1.0) [21] and Bcftools (version 0.1.17-dev) [22], respectively. Sequence depth criteria for filtering unreliable genotypes were determined on an individual bases to realize genotype concordance between next-generation sequencing variant call data and SNP array call data as 99.8%. We then phased the genotypes obtained from WGS using the SHAPEIT2 program (version 2.r644).

[23] Cryptic relatives inferred through an identity by descent estimate (PI\_HAT value  $> 0.125$ ) were removed from the reference panel. PI\_HAT values were calculated using the PLINK (version 1.9) program. [24]

### SNP array genotyping

Genotype calling was conducted using the apt-probeset-genotype program in the Affymetrix Power Tools suite (version 1.18.2; Thermo Fisher Scientific Inc., Waltham, MA). Quality control (QC) criteria were set in accordance with the manufacturer's recommendations (dish QC  $\geq 0.82$ ; sample call rate  $\geq 97\%$ ) and were met by all the samples. SNP-based QC was conducted using the Ps classification function in the SNPfisher package (version 1.5.2; Thermo Fisher Scientific Inc.). SNPs that were categorized as "recommended" by the Ps classification were retained. SNPs with call rate  $< 97.0\%$ , Hardy-Weinberg equilibrium of  $p < 10^{-6}$ , or MAF  $< 0.5\%$  were excluded from the downstream analysis.

### Population structure analysis

The SNP genotype data of the TMM samples were obtained by whole-genome sequencing or using the Japonica array. We obtained the corresponding SNP genotype data from next-generation sequencing analysis for the cases that were not analyzed with the Japonica array. To analyze the population genetics structure compared with that of East Asian populations, we downloaded the SNP data of chromosome 1 of unrelated East Asian populations (Dai Chinese, CDX; Han Chinese in Beijing, CHB; Han Chinese South, CHS; Tokyo Japanese, JPT; and Kinh Vietnamese, KVH) from the 1000 Genomes Project [25]. We selected the SNPs for which probes were included on the Japonica array [11] and used the VCFtools package to further filter the variants and individuals and the PLINK software package to calculate  $r^2$  scores. Indels and SNPs with maximum detection fraction  $> 0.1$ , smallest MAF 0.05, and maximum  $r^2$  0.8 were filtered out. The calculation of principal components for the SNP genotype was performed using the PLINK package.

Weir and Cockerham's  $F_{ST}$  value estimators [26] were calculated between all pairs of populations using PLINK. Based on the resultant  $F_{ST}$  matrix, a network was inferred among populations using the neighbor-net method [27] in the SplitTree program [28]. Sample clustering by haplotype sharing was performed with the fineSTRUCTURE program [14]. Haplotype phasing for this analysis was carried out using the SHAPEIT2 program (version 2.r644) with the default settings [23].

### Evaluation of imputation performance

We performed genotype imputation using the IMPUTE2 program. [29] Variants in the reference panel that had the same position in the Japonica array [11] (32,913



SNPs) were extracted for use in genotype imputation. The remaining variants in the panel (1,012,074 SNPs) were used to evaluate the accuracy of imputation of the true genotypes. Because the Miyagi samples used to test the reference panels are included in the reference panel (i.e., 1KJPN), we conducted a leave-one-out cross-validation experiment. Namely, each sample in the panel was extracted from the panel one after the other, and then genotype imputation of that sample was conducted against the entire panel without that sample. Because this procedure was repeated for all samples in the panel it required intensive computational resources, so the evaluation of imputation performance was conducted for SNPs only on chromosome 10 (1,044,987 SNPs). Through this process, we obtained imputed genotypes for every sample in the panel. Genotype imputation for the samples that were not included in the reference panel (i.e., Iwate, Nagahama, and Aki samples) was done with the IMPUTE2 program. Imputation accuracy was measured using Pearson's correlation coefficient ( $r^2$ ) between true genotypes, taking a value of 0, 1, or 2, and imputed genotype dosages with values between 0 and 2. The  $r^2$  values were estimated upon aggregating the variants in the reference panel that were stratified by non-reference allele frequency to visualize the imputation accuracies for rare SNPs. These evaluations were conducted on the SNPs that were identified in all the examined reference panels.

## Additional files

**Additional file 1: Table S1.** The imputation coverage (proportion of variants with an imputation  $r^2 \geq 0.8$ ), with each reference panel across different MAF bins. (XLSX 8 kb)

**Additional file 2: Figure S1.** Pairwise coincidence matrix of individuals from the four Japanese populations created by fineSTRUCTURE. Samples from Miyagi residents whose maternal grandmother was also born in Miyagi Prefecture (288 of the 1070 samples) were used to analyze haplotype sharing among the four Japanese populations. Color scale (left panel) represents the posterior confidence probability. Areas of each sample are shown by a colored box at the top of the figure. (PDF 312 kb)

**Additional file 3: Table S2.** The imputation coverage (proportion of variants with confidence: imputation  $r^2 \geq 0.8$ ) across different non-reference allele frequency bins and the number of variants that were efficiently imputed. (CSV 20 kb)

## Acknowledgments

We thank Nozomi Hatanaka, Noriko Takahashi, Masae Kimura, Keiko Tateno, and Chizuru Abe for their technical assistance. We thank Margaret Biswas, PhD, from Edanz Group ([www.edanzediting.com/ac](http://www.edanzediting.com/ac)) for editing a draft of this manuscript.

## Funding

This work was supported in part by the Tohoku Medical Megabank Project from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) and the Reconstruction Agency; the Ministry of Education, Culture, Sports, Science and Technology (MEXT); the Japan Agency for Medical Research and Development (AMED; Grant Numbers JP17km0105001 and JP17km0105002) for Tohoku University and (Grant Numbers 17km0105003j0006 and 17km0105004j0006) for Iwate Medical University; and the Center of Innovation Program from the Japan Science and Technology Agency (JST) for Tohoku University. All computational resources were

provided by the Tohoku University Tohoku Medical Megabank Organization (ToMMo) supercomputer system (<http://sc.megabank.tohoku.ac.jp/en>), which is supported by Facilitation of R&D Platform for AMED Genome Medicine Support conducted by AMED (Grant Number JP17km0405001). The Japan Public Health Center-Based Prospective Study for the Next Generation (JPHC-NEXT) was supported by the National Cancer Center Research and Development Fund and the Japan Science and Technology Agency (JST). The Nagahama Prospective Genome Cohort for Comprehensive Human Bioscience (the Nagahama Study) was supported by MEXT and the Takeda Science Foundation.

## Availability of data and materials

The datasets used and/or analyzed in the current study are available from the representative authors of each of the cohorts on reasonable request. 1KJPN can be obtained from MY; Nagahama data can be obtained from FM; Aki data can be obtained from MI; and Iwate data can be obtained from MS. The genotype data of the other East Asian populations were obtained from the 1000 Genome Project website [25]. The data are described in the 1000 Genomes Project Consortium paper: A global reference for human genetic variation. *Nature* 2015, 526:68–74.

## Authors' contributions

JY and MY planned and organized the study. NM, FM, KH, NY, MI, and MS contributed to sample collection and DNA preparation. FK, ID, SS, and AS performed the whole-genome sequencing and other experimental procedures. JY, YK, Kko, and MN performed the bioinformatics and statistical analyses. ST confirmed sample ID concordances. YYK, INM, Kku, GT, and Kki supervised the bioinformatics and statistical analyses. MSY, OT, and NF supervised the whole genome-sequencing and other experimental procedures. JY and YK contributed to writing the manuscript. YYK, KH, FM, NY, MI, NF, and MY amended the manuscript. All authors read and approved the final manuscript.

## Ethics approval and consent to participate

The ethics approvals were obtained from the ethics committees of the Tohoku University Tohoku Medical Megabank Organization (ID = 2017-4-054) for 1KJPN; the Graduate School of Medicine and Faculty of Medicine, Kyoto University for Nagahama samples (IDs = G0278-11 and G0455-8); the National Cancer Center (IDs = 2011-186 and 2016-305) and Kochi University Medical School (ID = 24-94) for the Aki samples; and Iwate Medical University for the Iwate samples (ID = HGH25-19). Written informed consent was obtained from the participants in all cohorts whose samples were provided for this study.

## Consent for publication

Not applicable.

## Competing interests

The authors declare that they have no competing interests.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Author details

<sup>1</sup>Sendai, Tohoku Medical Megabank Organization, 2-1, Seiryomachi, Aoba-ku, Tohoku Medical Megabank, Tohoku University, Sendai 980-8573, Miyagi, Japan. <sup>2</sup>Center for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto 606-8501, Sakyo-ku, Japan. <sup>3</sup>Department of Public Health, Kochi University Medical School, Nankoku-shi 783-8505, Kochi, Japan. <sup>4</sup>Division of Epidemiology, Center for Public Health Sciences, National Cancer Center, Tokyo 104-0045, Chuo-ku, Japan. <sup>5</sup>Division of Ultrahigh Field MRI, Institute for Biomedical Sciences, Iwate Medical University, 2-1-1 Nishitokuta, Yahaba, Shiwa, Iwate 028-3694, Japan. <sup>6</sup>Iwate Tohoku Medical Megabank Organization, Disaster Reconstruction Center, Iwate Medical University, 2-1-1 Nishitokuta, Yahaba, Shiwa, Iwate 028-3694, Japan. <sup>7</sup>Graduate School of Information Sciences, Tohoku University, Aoba-ku, Sendai 980-8579, Miyagi, Japan. <sup>8</sup>Department of Medical Biochemistry, Graduate School of Medicine, Tohoku University, Aoba-ku, Sendai 980-8575, Miyagi, Japan. <sup>9</sup>Present address: Department of Human Genetics, Graduate School of Medicine, The University of Tokyo, Bunkyo-ku 113-0033, Tokyo, Japan.

Received: 4 April 2018 Accepted: 16 July 2018

Published online: 24 July 2018

## References

- Nelson SC, Doheny KF, Pugh EW, Romm JM, Ling H, Laurie CA, Browning SR, Weir BS, Laurie CC. Imputation-based genomic coverage assessments of current human genotyping arrays. *G3 (Bethesda)*. 2013;3:1795–807.
- Deelen P, Menelaou A, van Leeuwen EM, Kanterakis A, van Dijk F, Medina-Gomez C, Francioli LC, Hottenga JJ, Karssen LC, Estrada K, et al. Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of the Netherlands'. *Eur J Hum Genet*. 2014;22:1321–6.
- Huang J, Howie B, McCarthy S, Memari Y, Walter K, Min JL, Danecek P, Malerba G, Trabetti E, Zheng H-F, et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat Commun*. 2015;6:8111.
- McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*. 2016;48:1279–83.
- Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. *G3 (Bethesda)*. 2011;1:457–70.
- Consortium TUK. The UK10K project identifies rare variants in health and disease. *Nature*. 2015;526:82–90.
- Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, Hutnik K, Royrvik EC, Cunliffe B, et al. The fine-scale genetic structure of the British population. *Nature*. 2015;519:309–14.
- Kuriyama S, Yaegashi N, Nagami F, Arai T, Kawaguchi Y, Osumi N, Sakaida M, Suzuki Y, Nakayama K, Hashizume H, et al. The Tohoku medical megabank project: design and mission. *J Epidemiol*. 2016;26:493–511.
- Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, Kawai Y, Yamaguchi-Kabata Y, Yokozawa J, Danjoh I, Saito S, et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat Commun*. 2015;6:8018.
- Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56–65.
- Kawai Y, Mimori T, Kojima K, Nariai N, Danjoh I, Saito R, Yasuda J, Yamamoto M, Nagasaki M. Japonica array: improved genotype imputation by designing a population-specific SNP array with 1070 Japanese individuals. *J Hum Genet*. 2015;60(10):581–7.
- Consortium TGP. A global reference for human genetic variation. *Nature*. 2015;526:68–74.
- Takeuchi F, Katsuya T, Kimura R, Nabika T, Isomura M, Ohkubo T, Tabara Y, Yamamoto K, Yokota M, Liu X, et al. The fine-scale genetic structure and evolution of the Japanese population. *PLoS One*. 2017;12(11):e0185487.
- Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. *PLoS Genet*. 2012;8:e1002453.
- Yamaguchi-Kabata Y, Nakazono K, Takahashi A, Saito S, Hosono N, Kubo M, Nakamura Y, Kamatani N. 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*. 2008;83:445–56.
- Jinam T, Nishida N, Hirai M, Kawamura S, Oota H, Umetsu K, Kimura R, Ohashi J, Tajima A, Yamamoto T, et al. The history of human populations in the Japanese archipelago inferred from genome-wide SNP data with a special reference to the Ainu and the Ryukyuan populations. *J Hum Genet*. 2012;57:787–95.
- Higasa K, Miyake N, Yoshimura J, Okamura K, Niihori T, Saitsu H, Doi K, Shimizu M, Nakabayashi K, Aoki Y, et al. human genetic variation database, a reference database of genetic variations in the Japanese population. *J Hum Genet*. 2016;61(6):547–53.
- Terao C, Bayoumi N, McKenzie CA, Zelenika D, Muro S, Mishima M, Connell JM, Vickers MA, Lathrop GM, Farrall M et al: Quantitative variation in plasma angiotensin-I converting enzyme activity shows allelic heterogeneity in the ABO blood group locus. *Ann Hum Genet* 2013, 77(6):465–471.
- JPHC-NEXT [<http://epi.ncc.go.jp/jphcnxt/index.html>].
- Motoike IN, Matsumoto M, Danjoh I, Katsuoka F, Kojima K, Nariai N, Sato Y, Yamaguchi-Kabata Y, Ito S, Kudo H, et al. Validation of multiple single nucleotide variation calls by additional exome analysis with a semiconductor sequencer to supplement data of whole-genome sequencing of a human population. *BMC Genomics*. 2014;15:673.
- Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods*. 2012;9:357–9.
- Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. *Bioinformatics*. 2011;27:2987–93.
- Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods*. 2013;10:5–6.
- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*. 2015;4:7.
- IGSR: The International Genome Sample Resource. <http://www.internationalgenome.org>.
- Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. *Evolution*. 1984;38:1358.
- Bryant D, Moulton V. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol*. 2004;21(2):255–65.
- Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol*. 2006;23(2):254–67.
- Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*. 2009;5(6):e1000529.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

