Human Genetics (2021) 140:1097–1108
https://doi.org/10.1007/s00439-021-02281-4

ORIGINAL INVESTIGATION

Geographic variation in the polygenic score of height in Japan

Mariko Isshiki[1] · Yusuke Watanabe[1,2] · Jun Ohashi[1]

Received: 15 September 2020 / Accepted: 12 April 2021 / Published online: 26 April 2021

Abstract

A geographical gradient of height has existed in Japan for approximately 100 years. People in northern Japan tend to be taller than those in southern Japan. The differences in annual temperature and day length between the northern and southern prefectures of Japan have been suggested as possible causes of the height gradient. Although height is well known to be a polygenic trait with high heritability, the genetic contributions to the gradient have not yet been explored. Polygenic score (PS) is calculated by aggregating the effects of genetic variants identified by genome-wide association studies (GWASs) to predict the traits of individual subjects. Here, we calculated the PS of height for 10,840 Japanese individuals from all 47 prefectures in Japan. The median height PS for each prefecture was significantly correlated with the mean height of females and males obtained from another independent Japanese nation-wide height dataset, suggesting genetic contribution to the observed height gradient. We also found that individuals and prefectures genetically closer to continental East Asian ancestry tended to have a higher PS; modern Japanese people are considered to have originated as result of admixture between indigenous Jomon people and immigrants from continental East Asia. Another PS analysis based on the GWAS using only the mainland Japanese was conducted to evaluate the effect of population stratification on PS. The result also supported genetic contribution to height, and indicated that the PS might be affected by a bias due to population stratification even in a relatively homogenous population like Japanese.

Introduction

Over the past decade, a vast number of genome-wide association studies (GWASs) have been performed, in which hundreds of thousands to millions of genetic variants across genomes are tested in various human populations to identify genetic variants associated with diseases and traits (Visscher et al. 2012, 2017; Buniello et al. 2019). GWASs have successfully identified many loci associated with complex diseases, such as type 2 diabetes mellitus (T2DM) (Morris et al. 2012; Xue et al. 2018), and complex traits, such as height (Wood et al. 2014; He et al. 2015; Zoledziewska et al. 2015; Marouli et al. 2017; Tachmazidou et al. 2017; Yengo et al. 2018; Akiyama et al. 2019).

Polygenic scores (PSs) are calculated by aggregating the effects of genetic variants identified by GWAS across the genome to predict the risk of complex diseases or the phenotype of complex traits based on genetic profiling. In conventional GWAS, although a stringent genome-wide significance level of $\alpha = 5 \times 10^{-8}$ has been used to avoid false positives, no attention has been paid to false negatives. In contrast, the PS approach allows the use of single nucleotide polymorphisms (SNPs) that do not reach the genome-wide significance level, even if the set of SNPs includes ones that are not truly associated with the risk of complex diseases or the phenotype of complex traits. The growing number of large-scale GWAS generated from regional and national biobank projects, such as the UK Biobank (UKBB) (Bycroft et al. 2018), has enabled the calculation of more powerful and precise PSs. For example, PSs calculated using the UKBB data successfully identified individuals at high risk for several diseases (Khera et al. 2018). The PS for European height has also been estimated from effect estimates of two large-scale height GWAS, GIANT (Wood et al. 2014) and the UKBB, in several modern and ancient populations (Robinson et al. 2015; Zoledziewska et al. 2015; Martiniano et al. 2017; Berg et al. 2019; Sohail et al. 2019; Uricchio

✉ Jun Ohashi
juno-tky@umin.ac.jp

[1] Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo 113-0033, Japan

[2] Genome Medical Science Project Toyama Project, National Center for Global Health and Medicine, Tokyo 162-8655, Japan

et al. 2019; Chen et al. 2020). They revealed a north–south gradient of polygenic adaptation, although the signal may be overestimated due to population stratification (Berg et al. 2019; Sohail et al. 2019; Uricchio et al. 2019; Chen et al. 2020).

A previous study (Yamaguchi-Kabata et al. 2008), based on the genome-wide SNP data of 7003 Japanese patients treated at hospitals in seven geographic regions, suggested that Japanese were genetically differentiated. Recently, we clarified the genetic heterogeneity of the Japanese population at the prefecture level by analyzing genome-wide SNP data of approximately 11,000 Japanese individuals from all 47 prefectures in Japan (Watanabe et al. 2020). The genetic structure of Japanese people is caused primarily by the extent of admixture between the Jomon people and immigrants from continental East Asia in each prefecture. This result strongly supports the dual structure model (Hanihara 1991), which suggests that the modern mainland Japanese population originated as a result of admixture between indigenous Jomon people and immigrants from continental East Asia. In addition, geographical location was found to largely contribute to the genetic heterogeneity among Japanese prefectures (Watanabe et al. 2020). Therefore, regional differences in the phenotypes of complex traits in Japanese individuals may be caused by the genetic differences among regions.

Several ecological analyses of nation-wide height data in Japan have indicated the existence of a geographical height gradient (Endo et al. 1993; Yokoya 2010; Yokoya et al. 2012). The height of Japanese youth tends to be greater in Northern Japan than in Southern Japan. The gradient is thought to have existed for over 100 years, and differences in the annual temperature and day length among prefectures have been suggested as possible causes of the gradient (Endo et al. 1993; Yokoya 2010; Yokoya et al. 2012). Moreover, height is well known as a polygenic phenotype with high heritability; about 80% of human height variation can be explained by genetic factors (Silventoinen et al. 2003; Visscher et al. 2006), and approximately 4% of human allelic variation has causal effects on height (Boyle et al. 2017). Thus, in this study, we examined whether genetic heterogeneity in the Japanese population contributes to the height gradient in Japan.

Materials and methods

Subjects and data quality control (QC)

An overview of this study is shown in Fig. 1. All subjects in this study were customers of the Japanese Direct to Consumer (DTC) genetic testing service, HealthData Lab (Yahoo! Japan Corporation, Tokyo, Japan). Genomic DNA was extracted from saliva samples and genotyped using the Illumina HumanCore-12 Custom BeadChip and HumanCore-24 Custom BeadChip (Illumina, San Diego, CA), as described in our previous study (Watanabe et al. 2020). The SNP and sample filtering was performed using PLINK version 1.9 (Chang et al. 2015). SNPs with Hardy–Weinberg equilibrium (HWE) P value < 0.01 or missing call

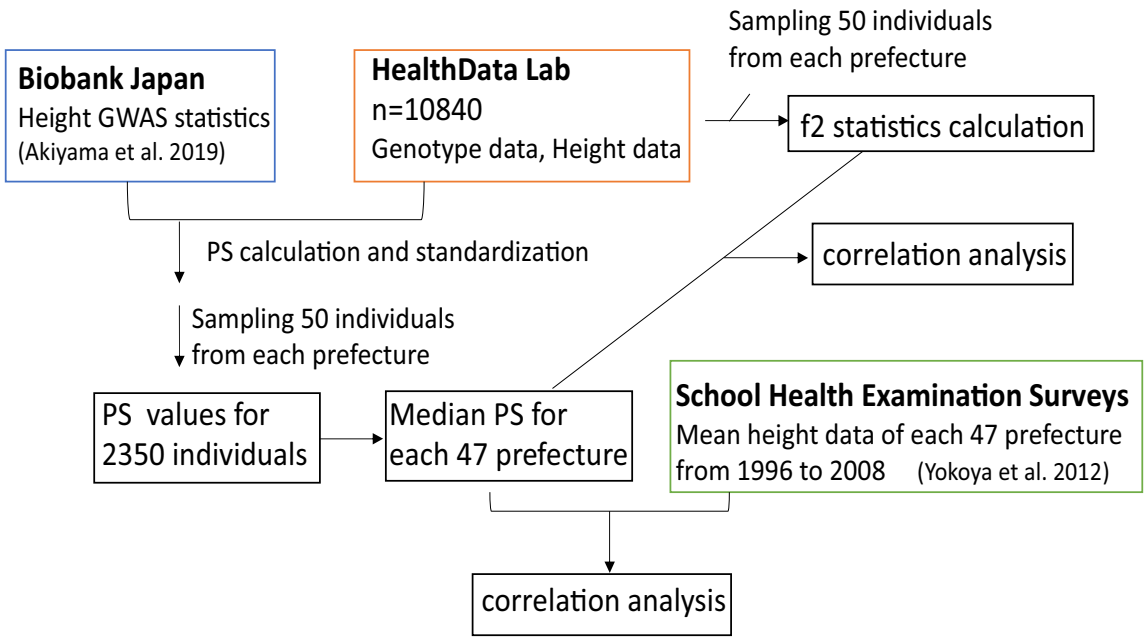

Fig. 1 Overview of this study

rate > 0.01, and samples with missing call rate > 0.1 were excluded. 116 individuals who were close to Han Chinese (CHB) in the 1000 Genomes Project Phase 3 (1KG) (1000 Genomes Project Consortium et al. 2015) in the PCA plot of our previous study (Watanabe et al. 2020) and 111 individuals with IBD values > 0.125 with one or more subjects were excluded; the proportion of IBD for all pairs of subjects within our dataset of the Japanese was calculated using PLINK version 1.912 (Chang et al. 2015) after LD-pruning with the settings of window size = 50 kb, step size = 5 kb and the r^2 threshold = 0.5. We also excluded two individuals born after 1999. The final dataset contained 10,840 individuals from all prefectures in Japan and 183,708 autosomal SNPs. Principal component analysis (PCA) was performed with PLINK version 1.9 (Chang et al. 2015) to adjust for population stratification in the PS calculation (Figure S1). All phenotype data, such as sex, year of birth, and height, were obtained through an online questionnaire. The details of samples were listed in Supplementary Table 1.

Polygenic score

We calculated the PS of height for the dataset of 10,840 Japanese individuals using PRSice-2 (Choi and O'Reilly 2019). Data from a GWAS of height of 159,095 Japanese participants in the Biobank Japan (BBJ) project with 27,896,057 imputed variants (Akiyama et al. 2019) were used as the base dataset since target sample ethnicities should be matched with the ethnicity of GWAS samples (Martin et al. 2017; Duncan et al. 2019). Variants with a minor allele frequency (MAF) < 0.01 and low imputation quality (R square < 0.7) were filtered out from the base dataset. Ambiguous SNPs (e.g., A/T and G/C SNPs) and duplicated SNPs were also excluded from the base dataset. After QC, 175,257 SNPs overlapped between the target dataset and the base dataset. Clumping parameters were set as --clump-kb 250kb --clump-p 1 --clump-r2 0.1. A total of 51,210 SNPs remained after clumping. Sex, year of birth, principal component (PC) 1, and PC2, were used as covariates in a regression model to select the best-fit threshold of PS that explained the highest phenotypic variance in the target dataset. The best-fit threshold was defined such that the difference in R^2 between the full model and the null model (i.e., the R^2 value added by the PSs) was greatest. The Z-Score Normalization was applied to the PS for each individual at the best-fit threshold and the standardized best-fit PS was used in the following analyses.

To evaluate the accuracy of our PS, we fit a linear regression model of height in which standardized PS, sex, year of birth, PC1, and PC2 were used as covariates in the 10,840 target samples. Then, the residuals, the difference between observed height and predicted height, were estimated for each subject. The difference in the residual distribution of each prefecture was examined by Student's t test implemented in R 3.5.3.

Comparison with the UKBB

To evaluate the effect of population stratification, we conducted correlation analysis between estimated effect sizes in the BBJ GWAS and those in the UKBB GWAS using the SNPs that are used to calculate Japanese height PS above. PS based on these effect sizes were also calculated for 1000 Genomes Japanese in Tokyo (JPT) individuals ($n = 104$) (1000 Genomes Project Consortium et al. 2015) using the same SNPs and the correlation between PSs based on the BBJ GWAS and those based on the UKBB GWAS was examined.

Map of height PSs

Japan is divided into 9 regions, which are further divided into 47 prefectures (Fig. 2). We visualized the geographic distribution of height PSs in Japan at prefecture level. Since the sample size is larger in metropolitan areas (Supplementary Table 1), more accurate height PSs would be obtained only in metropolitan areas if the sample size is not adjusted. Therefore, to ensure the same level of certainty in the PS value for each prefecture, random sampling with equal sample size for each prefecture is necessary. In this study, 50 individuals were randomly sampled from each prefecture of the target dataset without replacement (Supplementary Table 2). The median value of the best-fit height PS was calculated for each prefecture and mapped using the R packages choroplethr (https://CRAN.R-project.org/package=choroplethr) and choroplethrAdmin1 (https://CRAN.R-project.org/package=choroplethrAdmin1) in R version 3.5.3.

Independent height data

An independent dataset of standardized mean height at the age of 17 for each prefecture was obtained from Yokoya et al. (2012). The data were generated from reports of the School Health Examination Surveys conducted from 1996 to 2008 by the Ministry of Education. The mean height of each sex in each prefecture was mapped as described above. The association of PS with mean height of each prefecture was examined using R version 3.5.3.

Effect of population genetic ancestry on PS

To reveal the effect of genetic background on height PS, we first estimated the genetic distance between all subjects ($n = 10,840$) and 103 CHB individuals analyzed in the 1000 Genomes Project Phase 3 (1000 Genomes Project Consortium et al. 2015). Here, CHB was used to represent

Fig. 2 Map of Japanese prefectures. The Japanese prefectures were divided into nine regions in this study

continental populations whose ancestors migrated to the Japanese Archipelago. 138,688 autosomal SNPs were shared between our dataset and the CHB individuals. 68,573 SNPs, which was genotyped in all samples and remained after LD-pruning with the settings of window size = 50 kb, step size = 5 kb and the r^2 threshold = 0.5, were extracted using PLINK version 1.9 (Chang et al. 2015). The number of different alleles between each Japanese subject and 103 CHB individuals was counted using PLINK version 1.9 (Chang et al. 2015). The mean value of the number of different alleles from the 103 CHB individuals for each subject was divided by the total SNP number, which was defined as the genetic distance to CHB for each subject. The association of the genetic distance to CHB with PS was examined by correlation analysis using R version 3.5.3.

Then, the relationship between PS and genetic distance to CHB was examined at the prefecture level. From each prefecture, 50 individuals were randomly selected without replacement. The median PS was obtained for each prefecture. The *f*2 statistic (Patterson et al. 2012) between each prefecture and CHB was calculated based on the allele frequencies of 138,638 SNPs. The association between the *f*2 statistic and median PS was examined by correlation analysis using R version 3.5.3. In this study, a *P* value of < 0.05 was considered statistically significant.

Evaluation of Okinawa-mainland population stratification in GWAS

A height GWAS was carried out for 8385 Japanese individuals which excluded individuals who belong to Okinawa cluster in the PCA (Figure S1) to calculate PS without the effect of the Okinawa-mainland population stratification. The GWAS was performed by linear regression model implemented in PLINK 1.912. Sex, year of birth, PC1 and PC2 were included in the model as covariates. The summary statistics were used for PS calculation for 2350 Japanese individuals, 50 individuals each from 47 prefectures. The PS calculation was conducted as described above. The flow of this analysis was described in Figure S2.

Results

The model fit of PSs, which is defined as the R^2 of the full model minus the R^2 of the null model, was calculated to determine the optimal P value threshold (Figure S3). The best threshold, in which the model fit of PSs (i.e., the R^2 value added by the PSs) became greatest, was P value = 0.020. The R^2 value of the full model was $R^2 = 0.64$, and the model fit of PSs was 0.076 at the best P threshold. The 7521 SNPs below the threshold were used for PS calculation (Supplementary Table 3). The standardized best-fit PS was significantly correlated with the observed height both in male (Correlation coefficient $[R] = 0.38$ and P value $< 2.2 \times 10^{-16}$; Fig. 3a) and female ($R = 0.37$ and P value $< 2.2 \times 10^{-16}$; Fig. 3b). Figure S4a shows the relationship between the observed height and the height predicted by linear regression model using the standardized best-fit PS, sex, year of birth, PC1 and PC2 as covariates. The correlation coefficient between the observed and predicted height was $R = 0.80$ (P value $< 2.2 \times 10^{-16}$). The residuals spanned from −40 to 20 cm. The residuals of approximately 95% of subjects ($n = 10{,}262$) fell between −10 and 10 cm (Figure S4b). The distribution of residuals was not different among the prefectures (Figure S4c).

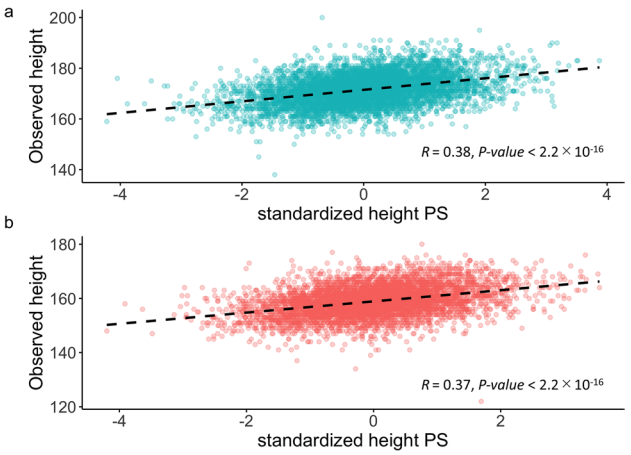

Fig. 3 Correlation between standardized height PS and observed height. **a** Distribution of standardized PS and observed height of 5761 males. **b** Distribution of standardized PS and observed height of 5079 females

The distribution of height PSs varies among prefectures (Fig. 4). Then, we examined whether the distribution of height PSs can explain the geographic gradient of height. To do this, we used another independent Japanese nation-wide height dataset based on the School Health Examination Surveys (Yokoya et al. 2012), in which the height of Japanese youth was measured throughout Japan (Fig. 1). We noted that the height data of our target samples were not used for correlation analysis because these data were used for model fitting in PS calculation. The median value of the height PS was plotted on the map (Figs. 4a and S6a). Northern

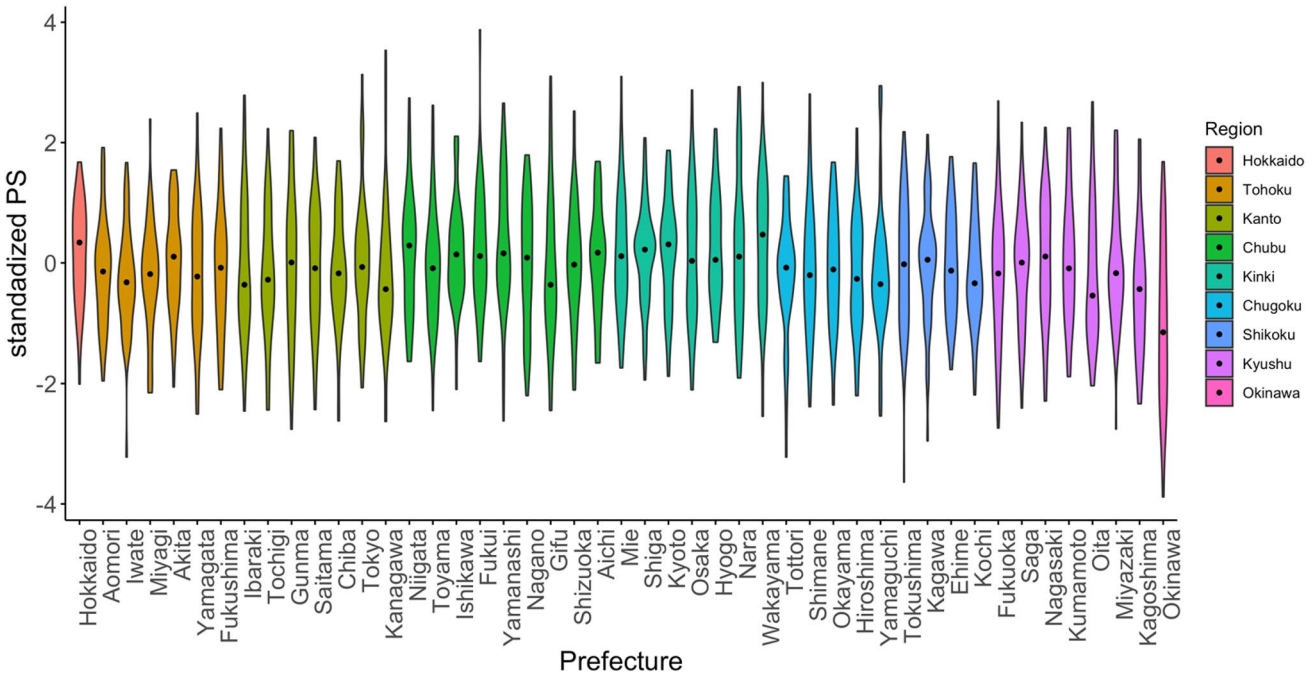

Fig. 4 Violin plot of standardized PS in each prefecture. The black dot indicates the median PS. The color of each region corresponds to that of Fig. 2

regions of Japan tended to show larger PSs. The distribution was similar to the geographic distribution of female height (Figs. 4b and S6b) and male height (Figure S6a) based on the School Health Examination Surveys (Yokoya et al. 2012). The median height PS of each prefecture was significantly correlated with female height ($R = 0.58$ and P value $= 1.6 \times 10^{-5}$; Figure S5c) and male height ($R = 0.53$ and P value $= 0.00012$; Figure S5d). Since the Okinawa Prefecture was an outlier (Figures S6), correlation analysis was also performed for all prefectures except Okinawa. Significant correlations were found in females ($R = 0.42$ and P value $= 0.0033$; Fig. 5c) and males ($R = 0.36$ and P value $= 0.014$; Figure S6b).

To assess if the observed correlations were merely due to population stratification in the BBJ GWAS, the effect sizes in the BBJ of the SNPs used for the PS calculation were compared to those in the UKBB GWAS. These values correlated significantly ($R = 0.51$ and P value $< 2.2 \times 10^{-16}$; Figure S7a). PSs were then calculated for 1000 Genomes JPT individuals using the same SNPs and the effect sizes from the BBJ and UKBB GWASs. The PSs based on the BBJ effect sizes were significantly correlated with those based on the UKBB effect sizes ($R = 0.28$ and P value $= 0.0046$; Figure S7b).

To uncover the effect of genetic ancestry on height PS, the relationship between PS and genetic distance to a continental population, CHB, was examined at both the individual and prefectural level. The genetic distance to CHB was negatively correlated with the PS ($R = -0.094$ and P value $< 2.2 \times 10^{-16}$; Fig. 6a). The $f2$ statistic, corresponding to the genetic distance between each prefecture and CHB, was negatively correlated with the median value of the PS ($R = -0.61$ and P value $= 4.0 \times 10^{-6}$; Fig. 6b). PS and genetic distance to CHB were correlated with each other at both the individual and prefectural level even when individuals belonged to Okinawa cluster in Figure S1 were or Okinawa Prefecture was excluded, respectively (Figure S8).

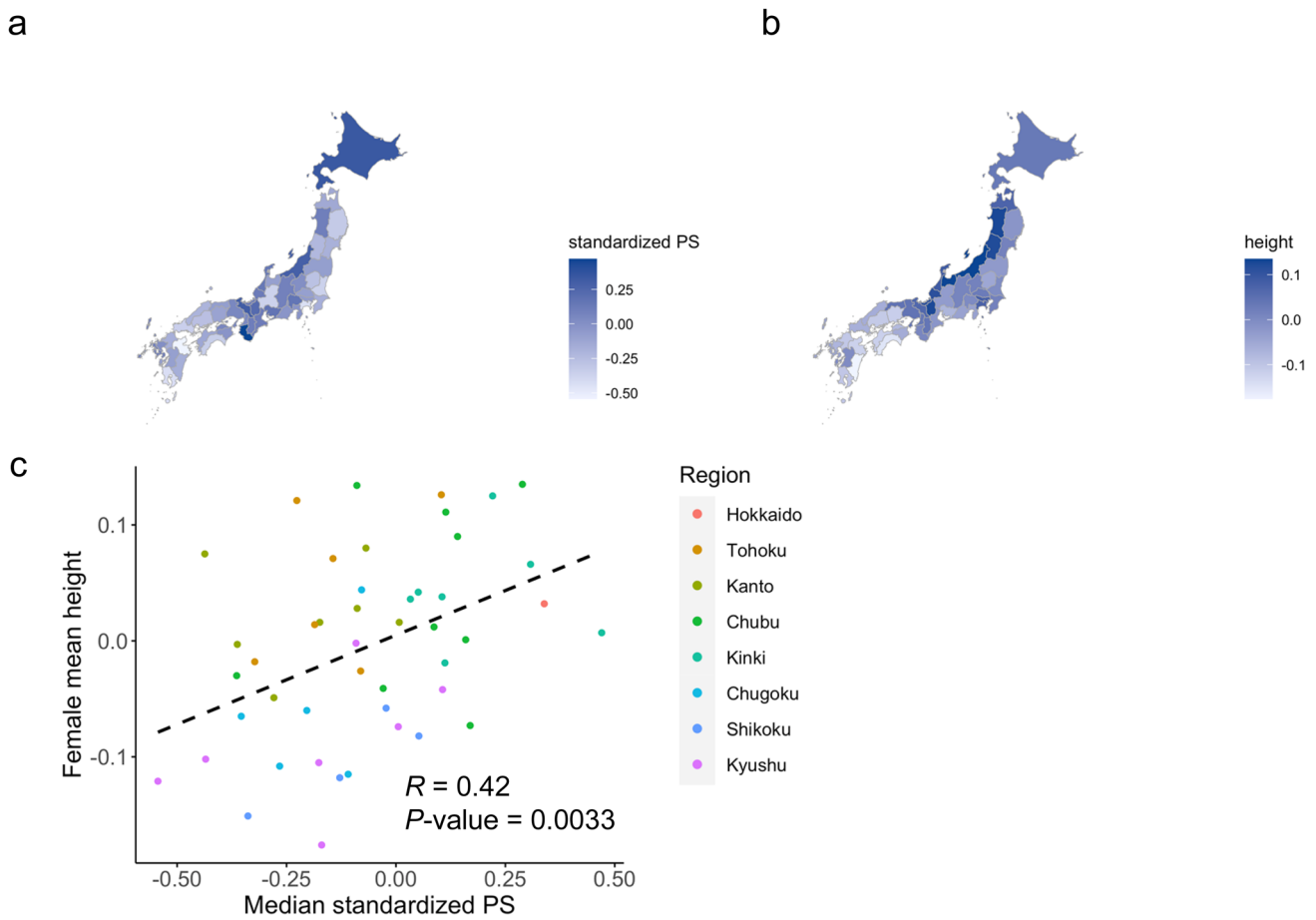

Fig. 5 Geographic distribution of height PS and female height for 46 prefectures. **a** Geographic distribution of the median standardized PS of height of each prefecture except Okinawa. **b** Geographic distribution of the mean height of 17-year-old females in each prefecture except Okinawa. Darker blue indicates larger values. **c** Correlation of the median standardized PS with mean height of 17-year-old females in each prefecture. The color of each region corresponds to that of Fig. 2

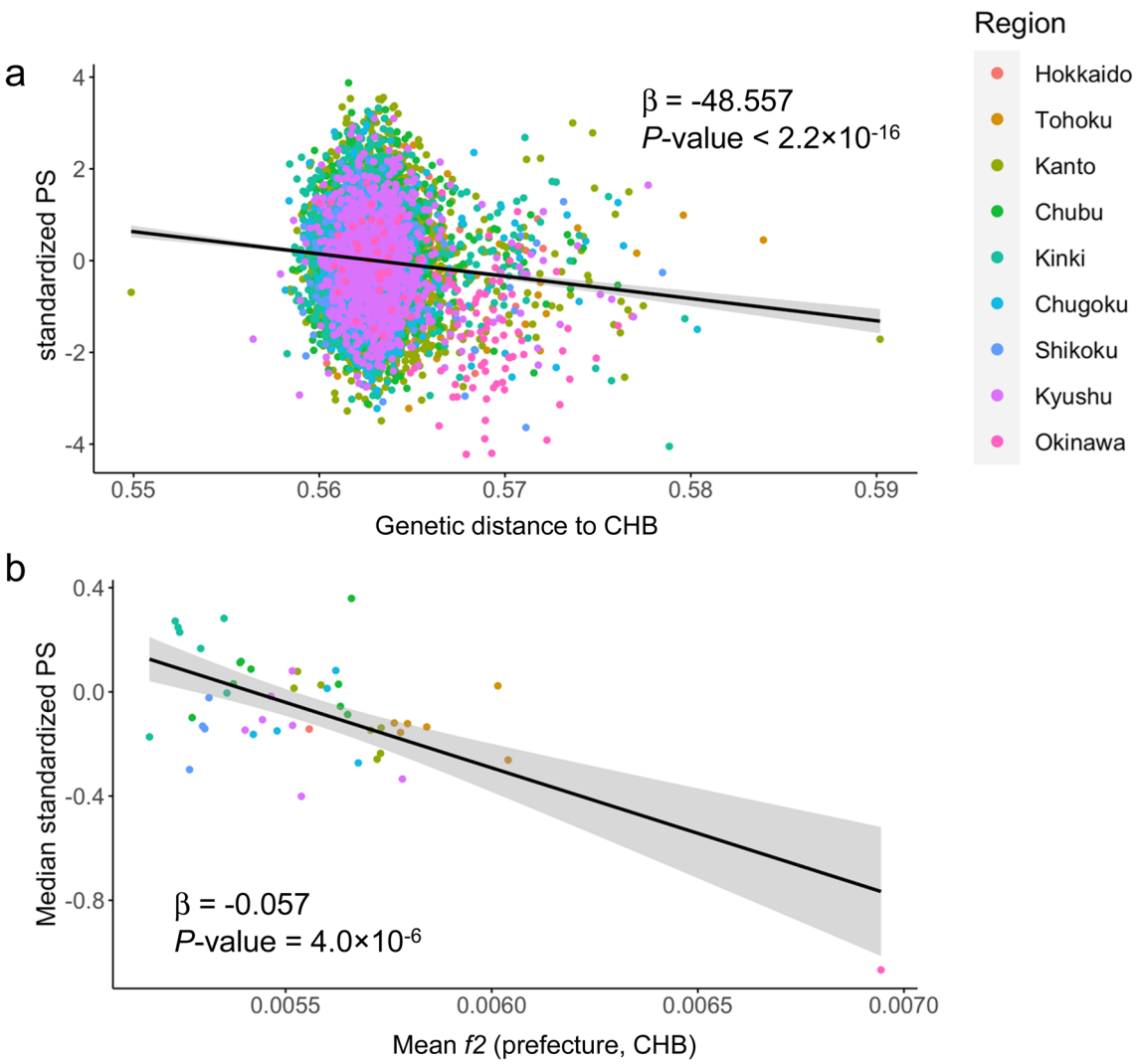

Fig. 6 Correlation of genetic distance to CHB with PS. **a** Correlation of genetic distance to CHB with PS at the individual level. **b** Correlation of the *f*2 statistics, corresponding to the genetic distance between each prefecture and CHB, with PS at the prefecture level. The color of each region corresponds to that of Fig. 2

To evaluate the effect of population stratification between Okinawa and the mainland Japan in GWAS statistics, we calculated PS for 2350 Japanese individuals from 47 prefectures based on the GWAS statistics which were not affected by Okinawa-mainland population stratification (Figure S2). The R^2 value of the full model was $R^2 = 0.57$, and the model fit of PSs (i.e., the R^2 value added by the PSs) was 0.0099 at the best *P* threshold (*P* value = 0.19). The 18,183 SNPs below the threshold were used for PS calculation. The geographic distribution of the newly calculated height PSs is shown in Fig. 7a. The median height PS of each prefecture was significantly correlated with female height ($R = 0.45$ and *P* value = 0.0017) and male height ($R = 0.43$ and *P* value = 0.0024) although the correlation was weak compared to the PS based on the BBJ GWAS statistics. The correlation was still observed even when Okinawa was excluded from the analysis for female ($R = 0.33$ and *P* value = 0.027; Fig. 7b) and male ($R = 0.31$ and *P* value = 0.034). The *f*2 statistic, corresponding to the genetic distance between each prefecture and CHB, was negatively correlated with the median value of the PS when including Okinawa ($R = -0.37$ and *P* value = 0.0097; Fig. 7c) but not when excluding Okinawa ($R = -0.17$ and *P* value = 0.27; Fig. 7d). This is probably because Tohoku region, which was genetically close to Okinawa, exhibited higher PSs. The significant correlation between the *f*2 statistics and the median values of the PS was observed when excluding Tohoku region from the analysis ($R = -0.42$ and *P* value = 0.0063).

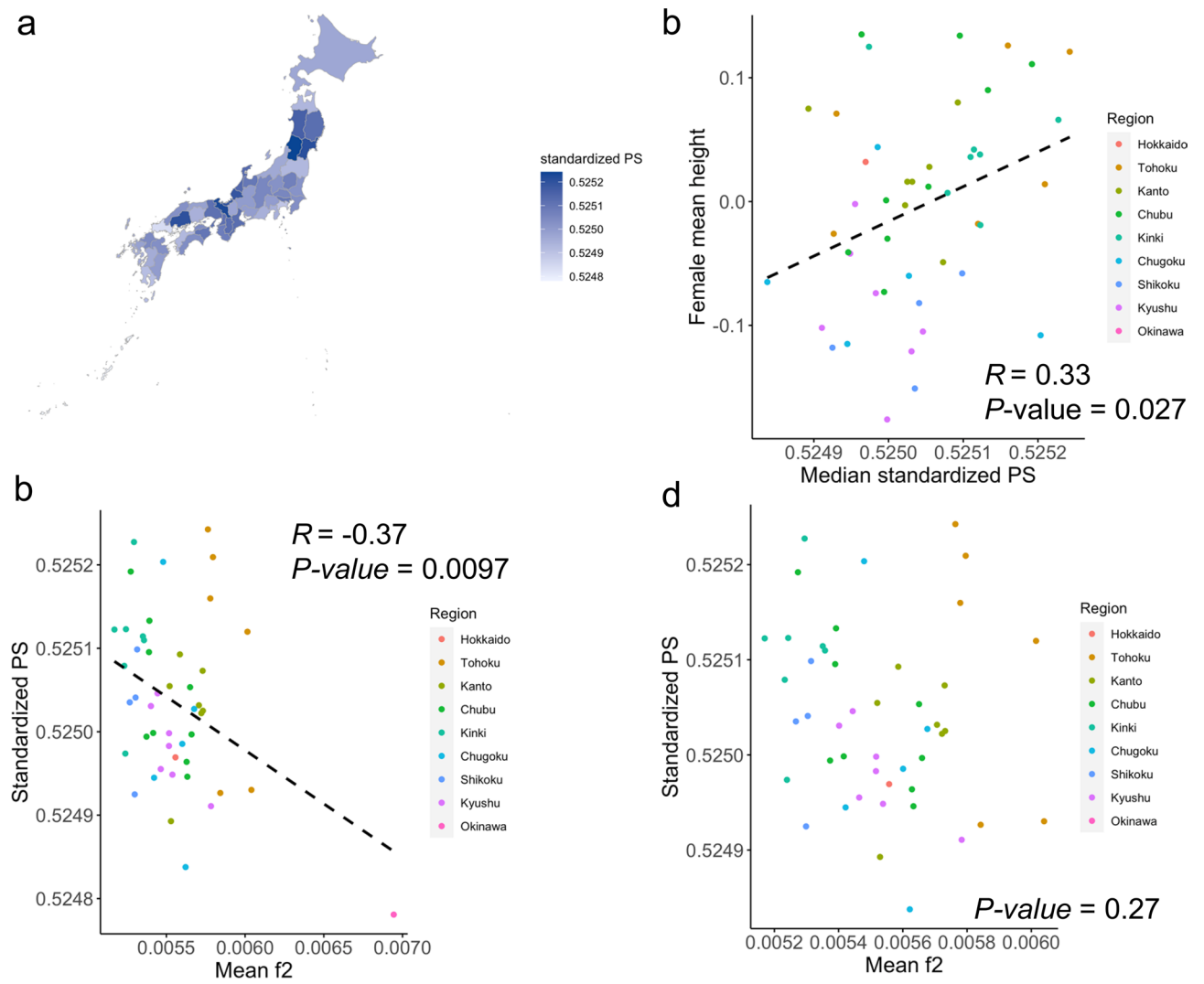

Fig. 7 Height PS based on GWAS without Okinawa-mainland population stratification. Height PS was calculated based on the newly conducted GWAS statistics which were not affected by Okinawa-mainland population stratification. **a** Geographic distribution of the median standardized PS of height of each prefecture. **b** Correlation of the median standardized PS with mean height of 17-year-old females in each prefecture. **c**, **d** Correlation of the *f*2 statistics, corresponding to the genetic distance between each prefecture and CHB, with PS at the prefecture level (**c**) with or (**d**) without Okinawa Prefecture. **b**–**d** The color of each region corresponds to that of Fig. 2

Discussion

In this study, we clarified the geographical distribution of height PS in Japan (Figs. 4a and S6a). To the best of our knowledge, this is the first attempt to elucidate the effect of genetic background on the phenotype of a complex trait at the prefecture level in Japan. As shown in Figure S4b, the predictive accuracy of the height PS was not high. Given that the aim of this study was to ascertain the genetic contribution to the height gradient observed in Japan, rather than to create an accurate genetic prediction of height, the predictive accuracy of the PS was considered to be enough for this study. The PS residuals did not vary among prefectures (Figure S4c), indicating that predictive power of PS was not different among prefectures, which allowed us to discuss geographical distribution of genetic background of Japanese height.

Japanese youth tend to be taller in areas north and east of Japan, especially in areas along the Sea of Japan coast (Figs. 4b and S7a). Since this gradient has existed for approximately 100 years, it was considered to be attributable to climatic factors that have not changed during this period, such as temperature (Endo et al. 1993; Yokoya 2010) and day length (Yokoya et al. 2012). The geographical distribution of Japanese height was negatively correlated with temperature and day length. Temperature has been suggested

to affect food intake, and day length may affect the production of melatonin, a hormone involved in sleep. Although the contribution of genetic factors have not been evaluated because the mainland Japanese population was considered to be genetically homogenous due to the increased migration from rural to urban areas during this period, it has recently been reported that genetic heterogeneity still exists in the Japanese population at the prefecture level (Watanabe et al. 2020). Significant correlations between the median PS and mean height of each prefecture were observed in both sexes when Okinawa was excluded (Figs. 4c and S7b) and when included (Figures S6c and S6d). In addition, compared with the UKBB GWAS (Figure S7), the calculated PSs were likely to reflect genetic difference in height rather than merely due to population stratification in the BBJ GWAS. Taken together, geographic differences in genetic backgrounds are likely to contribute to the observed gradient of Japanese height, along with other environmental factors.

We further examined the effect of the genetic ancestry on the height PS using genetic distance to CHB. Given that the modern Japanese population is considered to have originated as a result of admixture between the indigenous Jomon people and immigrants from continental East Asia, according to the dual structure model (Hanihara 1991), genetic distance to CHB can be used as an index of the extent of continental ancestry. As shown in Fig. 6, individuals and prefectures with a longer genetic distance to CHB tended to have higher PSs, suggesting the continental immigrants were genetically taller than the Jomon people. Our results are compatible with archaeological evidence. Several studies indicated the Jomon people were shorter than the people of the period after the migration from continental East Asia, such as the people of the Yayoi and the Kofun period (Hiramoto 1972; Kaifu 1992; Wada and Motomura 2000). Although the ancient height change may be largely attributed to environmental factors, such as the transmission of rice cultivation, that improved the nutritional status of ancient Japanese people, our results suggest that genetic factors obtained through admixture with continental immigrants, who had a genetic background leading to taller height, also contributed to the height change.

Several studies reported that population stratification can cause a bias in PS prediction (Berg et al. 2019; Sohail et al. 2019; Uricchio et al. 2019; Chen et al. 2020; Sakaue et al. 2020). Previously Japanese people were suggested to be genetically divided into the two subpopulations, the mainland Japanese and Okinawa people (Hammer and Horai 1995; Horai et al. 1996; Hatta et al. 1999; Yamaguchi-Kabata et al. 2008; Jinam et al. 2012; Watanabe et al. 2020). The height GWAS conducted by the BBJ contained participants from Okinawa (Nagai et al. 2017; Akiyama et al. 2019). Therefore, our PS can be affected by a bias caused by the Okinawa-mainland population stratification. To examine this, another GWAS was conducted on 8385 Japanese people, who belonged to the mainland Japanese cluster in the PCA (Figure S1), and the PS was calculated using our HelathData Lab dataset (Figure S2). Since the sample size of the GWAS was much smaller than the BBJ GWAS, the GWAS statistics and the newly calculated PS values were less reliable; the R^2 value added by the PSs was 0.076 for the PS based on the BBJ GWAS statistics (hereinafter called "PS_{BBJ}") while that was 0.0099 for the PS based on our HealthData Lab GWAS (hereinafter called "PS_{HDL}"), and the correlation between the median PS and the mean height of each prefecture was weaker in the PS_{HDL} than the PS_{BBJ} (Figures S6 and 6). The results, however, provided us some insights about the bias in the PS_{BBJ} because the PS_{HDL} were not supposed to be affected by the Okinawa-mainland population stratification although it might still be affected by population structure within the mainland Japanese population. The mean height of prefectures in Tohoku region are high compared to other regions (Figs. 4b and S7a) but the PS_{BBJ} for Tohoku was intermediate (Figs. 4c and S7b) while the PS_{HDL} for Tohoku was relatively high (Fig. 7a). Considering that Tohoku region was genetically close to Okinawa (Figure S1; Yamaguchi-Kabata et al. 2008; Watanabe et al. 2020), PS_{BBJ} for Tohoku was probably underestimated due to the bias caused by the Okinawa-mainland population stratification. Corroborative of Kerminen et al. (2019), our results suggested that geographic distribution of PS was sensitive to the bias caused by population stratification even in a relatively homogenous population like Japanese.

The PS_{HDL} also indicated the possibility that the correlation observed between the PS_{BBJ} and genetic distance to CHB was overestimated. The significant correlation between the $f2$ statistic and the PS_{HDL} was observed when Okinawa was included but not when excluded. The correlation, however, was stronger when prefectures in Tohoku region were excluded, indicating the correlation observed in the PS_{BBJ} was not caused only by the bias due to the population stratification. Thus, the effect of the extent of continental East Asian ancestry on Japanese height might exist in the mainland Japanese to some extent although overestimated in the PS_{BBJ}. The taller height of modern Tohoku people in spite of the lack of genetic components from the continental immigrants is possibly attributable to some genetic factor other than continental East Asian ancestry. For example, Tohoku people may have the Jomon-derived genetic factors associated with taller height. The north–south geographical gradient of the body size was observed in the Jomon people, and Jomon people living in Tohoku region have been reported to be taller than those in other regions (Fukase et al. 2012). Future research is needed to examine whether Jomon people living in Tohoku region were genetically taller than those in other regions.

There are several limitations in this study. First, we cannot eliminate the possibility that the observed association between PS and height of each prefecture was still affected by population stratification within Japanese people. Because of the small number of SNPs in the dataset, we only did a minimal correction of population stratification to avoid reducing the amount of information. Second, the approach of PS calculation in this study, which uses clumping and selects the most predictive PS, can cause overfitting to the target data and produce inflated results and false conclusions (Choi et al. 2020). Third, the prefecture of each individual in this study was where one now lives, not where one originates. Although we examined that there is some correspondence between where one lives and where one's ancestors lived in our previous study which analyzed the same dataset (Watanabe et al. 2020), discrepancies between the residence and the origin of the individuals may affect our results to some extent.

Consequently, we first illustrated the geographic distribution of the genetic background of height in Japan. Geographical differences in genetic background appear to partially explain the observed height gradient in Japan. Our study suggests that even in a relatively homogenous population like the Japanese, geographic differences in the phenotype of complex traits can be explained by geographic differences in genetic background and population stratification can cause a bias on PS calculation. The genetic background of Japanese height was partially attributed to the extent of continental East Asian ancestry. To obtain more precise geographic distribution of genetic background of complex phenotypes, further understanding of the effect of biases arising from population stratification on PSs is needed.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00439-021-02281-4.

Acknowledgements We are grateful to the individuals who participated in the study. We would like to express our deepest gratitude to Masahiro Inoue, Shota Arichi, and Akito Tabira, who obtained the genotype data and provided the technical environment for analyzing them. This study was supported in part by a Grant-in-Aid for Scientific Research (B) (18H02514) and Grant-in-Aid for Scientific Research on Innovative Areas (19H05341, 21H00336) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan, and was supported by AMED under Grant Numbers JP19fk0310115 and JP20km0405211.

Availability of data and materials The SNP datasets analyzed in the current study are not available to avoid personal identification. The statistics per prefectures generated in the current study are available in the Supplementary materials.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethics approval This study was performed in accordance with the principles of the Declaration of Helsinki. Approval was obtained from the Ethics Committee of the Yahoo! Japan Corporation.

Consent to participate (including appropriate statements) Informed consent to participate in the study was obtained from all participants.

Consent for publication (including appropriate statements) Informed consent for publication in the study was obtained from all participants.

References

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA et al (2015) A global reference for human genetic variation. Nature 526:68–74. https://doi.org/10.1038/nature15393

Akiyama M, Ishigaki K, Sakaue S, Momozawa Y, Horikoshi M, Hirata M, Matsuda K, Ikegawa S, Takahashi A, Kanai M et al (2019) Characterizing rare and low-frequency height-associated variants in the Japanese population. Nat Commun 10:4393. https://doi.org/10.1038/s41467-019-12276-5

Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, Boyle EA, Zhang X, Racimo F, Pritchard JK et al (2019) Reduced signal for polygenic adaptation of height in UK biobank. Elife 8:1–47. https://doi.org/10.7554/eLife.39725

Boyle EA, Li YI, Pritchard JK (2017) An expanded view of complex traits: from polygenic to Omnigenic. Cell 169:1177–1186. https://doi.org/10.1016/j.cell.2017.05.038

Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E et al (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 47:D1005–D1012. https://doi.org/10.1093/nar/gky1120

Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J et al (2018) The UK Biobank resource with deep phenotyping and genomic data. Nature 562:203–209. https://doi.org/10.1038/s41586-018-0579-z

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4:7. https://doi.org/10.1186/s13742-015-0047-8

Chen M, Sidore C, Akiyama M, Ishigaki K, Kamatani Y, Schlessinger D, Cucca F, Okada Y, Chiang CWK (2020) Evidence of polygenic adaptation in Sardinia at height-associated loci ascertained from the Biobank Japan. Am J Hum Genet 107:60–71. https://doi.org/10.1016/j.ajhg.2020.05.014

Choi SW, O'Reilly PF (2019) PRSice-2: Polygenic Risk Score software for biobank-scale data. Gigascience 8:1–6. https://doi.org/10.1093/gigascience/giz082

Choi SW, Mak TSH, O'Reilly PF (2020) Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc 15:2759–2772. https://doi.org/10.1038/s41596-020-0353-1

Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, Peterson R, Domingue B (2019) Analysis of polygenic risk score usage and performance in diverse human populations. Nat Commun 10:3328. https://doi.org/10.1038/s41467-019-11112-0

Endo A, Omoe K, Ishikawa H (1993) Ecological factors affecting body size of Japanese adolescents. Am J Phys Anthropol. https://doi.org/10.1002/ajpa.1330910305

Fukase H, Wakebe T, Tsurumoto T, Saiki K, Fujita M, Ishida H (2012) Geographic variation in body form of prehistoric Jomon males in the Japanese archipelago: its ecogeographic implications. Am J Phys Anthropol 149:125–135. https://doi.org/10.1002/ajpa.22112

Hammer MF, Horai S (1995) Y chromosomal DNA variation and the peopling of Japan. Am J Hum Genet 56:951–962

Hanihara K (1991) Dual structure model for the population history of the Japanese. Anthropol Sci 2:1–33. https://doi.org/10.1537/ase.102.455

Hatta Y, Ohashi J, Imanishi T, Kamiyama H (1999) HLA genes and haplotypes in Ryukyuans suggest recent gene flow to the Okinawa Islands. Hum Biol 71:353–365

He M, Xu M, Zhang B, Liang J, Chen P, Lee JY, Johnson TA, Li H, Yang X, Dai J et al (2015) Meta-analysis of genome-wide association studies of adult height in East Asians identifies 17 novel loci. Hum Mol Genet 24:1791–1800. https://doi.org/10.1093/hmg/ddu583

Hiramoto Y (1972) Secular change of estimated stature of Japanese in Kanto district from the prehistoric age to the present day. J Anthropol Soc Nippon 80:221–236. https://doi.org/10.1537/ase1911.80.221

Horai S, Murayama K, Hayasaka K, Matsubayashi S, Hattori Y, Fucharoen G, Harihara S, Park KS, Omoto K, Pan IH (1996) mtDNA polymorphism in East Asian Populations, with special reference to the peopling of Japan. Am J Hum Genet 59:579–590

Jinam T, Nishida N, Hirai M, Kawamura S, Oota H, Umetsu K, Kimura R, Ohashi J, Tajima A, Yamamoto T et al (2012) 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. https://doi.org/10.1038/jhg.2012.114

Kaifu Y (1992) Human skeletal remains of the Yayoi period from the Iwatusbo Cave Site in Gunma Prefecture, Kanto District. J Anthropol Soc Nippon 100:449–483. https://doi.org/10.1537/ase1911.100.449

Kerminen S, Martin AR, Koskela J, Ruotsalainen SE, Havulinna AS, Surakka I, Palotie A, Perola M, Salomaa V, Daly MJ et al (2019) Geographic variation and bias in the polygenic scores of complex diseases and traits in Finland. Am J Hum Genet 104:1169–1181. https://doi.org/10.1016/j.ajhg.2019.05.001

Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT et al (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet 50:1219–1224. https://doi.org/10.1038/s41588-018-0183-z

Marouli E, Graff M, Medina-Gomez C, Sin Lo K, Wood AR, Kjaer TR, Fine RS, Lu Y, Schurmann C, Highland HM et al (2017) Rare and low-frequency coding variants alter human adult height coding variants associated with height. Nat Publ Gr. https://doi.org/10.1038/nature21039

Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, Daly MJ, Bustamante CD, Kenny EE (2017) Human demographic history impacts genetic risk prediction across diverse populations. Am J Hum Genet 100:635–649. https://doi.org/10.1016/j.ajhg.2017.03.004

Martiniano R, Cassidy LM, Ó'Maoldúin R, McLaughlin R, Silva NM, Manco L, Fidalgo D, Pereira T, Coelho MJ, Serra M et al (2017) The population genomics of archaeological transition in west Iberia: investigation of ancient substructure using imputation and haplotype-based methods. PLoS Genet 13:1–24. https://doi.org/10.1371/journal.pgen.1006852

Morris AP, Voight BF, Teslovich TM, Ferreira T, Segrè AV, Steinthorsdottir V, Strawbridge RJ, Khan H, Grallert H, Mahajan A et al (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet. https://doi.org/10.1038/ng.2383

Nagai A, Hirata M, Kamatani Y, Muto K, Matsuda K, Kiyohara Y, Ninomiya T, Tamakoshi A, Yamagata Z, Mushiroda T et al (2017) Overview of the BioBank Japan project: study design and profile. J Epidemiol 27:S2–S8. https://doi.org/10.1016/j.je.2016.12.005

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D (2012) Ancient admixture in human history. Genetics 192:1065–1093. https://doi.org/10.1534/genetics.112.145037

Robinson MR, Hemani G, Medina-Gomez C, Mezzavilla M, Esko T, Shakhbazov K, Powell JE, Vinkhuyzen A, Berndt SI, Gustafsson S et al (2015) Population genetic differentiation of height and body mass index across Europe. Nat Genet 47:1357–1361. https://doi.org/10.1038/ng.3401

Sakaue S, Hirata J, Kanai M, Suzuki K, Akiyama M, Lai Too C, Arayssi T, Hammoudeh M, Al Emadi S, Masri BK et al (2020) Dimensionality reduction reveals fine-scale structure in the Japanese population with consequences for polygenic risk prediction. Nat Commun 11:1569. https://doi.org/10.1038/s41467-020-15194-z

Silventoinen K, Sammalisto S, Perola M, Boomsma DI, Cornes BK, Davis C, Dunkel L, de Lange M, Harris JR, Hjelmborg JVB et al (2003) Heritability of adult body height: a comparative study of twin cohorts in eight countries. Twin Res 6:399–408. https://doi.org/10.1375/136905203770326402

Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, Chiang CW, Hirschhorn J, Daly MJ, Patterson N et al (2019) Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. Elife 8:1–17. https://doi.org/10.7554/elife.39702

Tachmazidou I, Süveges D, Min JL, Ritchie GRS, Steinberg J, Walter K, Iotchkova V, Schwartzentruber J, Huang J, Memari Y et al (2017) Whole-genome sequencing coupled to imputation discovers genetic signals for anthropometric traits. Am J Hum Genet 100:865–884. https://doi.org/10.1016/j.ajhg.2017.04.014

Uricchio LH, Kitano HC, Gusev A, Zaitlen NA (2019) An evolutionary compass for detecting signals of polygenic selection and mutational bias. Evol Lett. https://doi.org/10.1002/evl3.97

Visscher PM, Medland SE, Ferreira MAR, Morley KI, Zhu G, Cornes BK, Montgomery GW, Martin NG (2006) Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet. https://doi.org/10.1371/journal.pgen.0020041

Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet. https://doi.org/10.1016/j.ajhg.2011.11.029

Visscher PM, Wray NR, Zhang Q, Sklar P, Mccarthy MI, Brown MA, Yang J (2017) 10 Years of GWAS discovery: biology, function, and translation. Am J Hum Genet. https://doi.org/10.1016/j.ajhg.2017.06.005

Wada Y, Motomura H (2000) Temporal changes in stature of Western Japanese based on limb characteristics. Anthropol Sci 108:147–168. https://doi.org/10.1537/ase.108.147

Watanabe Y, Isshiki M, Ohashi J (2020) Prefecture-level population structure of the Japanese based on SNP genotypes of 11,069 individuals. J Hum Genet. https://doi.org/10.1038/s10038-020-00847-0

Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, Chu AY, Estrada K, Luan J, Kutalik Z et al (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet 46:1173–1186. https://doi.org/10.1038/ng.3097

Xue A, Yang Wu, Zhu Z, Zhang F, Kemper KE, Zheng Z, Yengo L, Lloyd-Jones LR, Sidorenko J, Yeda Wu et al (2018) Genome-wide association analyses identify 143 risk variants and putative

regulatory mechanisms for type 2 diabetes. Nat Commun. https://doi.org/10.1038/s41467-018-04951-w

Yamaguchi-Kabata Y, Nakazono K, Takahashi A, Saito S, Hosono N, Kubo M, Nakamura Y, Kamatani N (2008) 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. https://doi.org/10.1016/j.ajhg.2008.08.019

Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, Frayling TM, Hirschhorn J, Yang J, Visscher PM (2018) Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. Hum Mol Genet 27:3641–3649. https://doi.org/10.1093/hmg/ddy271

Yokoya M (2010) Geographic variation in the body size of japanese students and its analysis by mesh climate data. Japanese J Nutr Diet 68:263–269. https://doi.org/10.5264/eiyogakuzashi.68.263

Yokoya M, Shimizu H, Higuchi Y (2012) Geographical distribution of adolescent body height with respect to effective day length in japan: an ecological analysis. PLoS ONE 7:5–8. https://doi.org/10.1371/journal.pone.0050994

Zoledziewska M, Sidore C, Chiang CWK, Sanna S, Mulas A, Steri M, Busonero F, Marcus JH, Marongiu M, Maschio A et al (2015) Height-reducing variants and selection for short stature in Sardinia. Nat Genet. https://doi.org/10.1038/ng.3403