nature communications

Article

https://doi.org/10.1038/s41467-024-55147-4

SEAD reference panel with 22,134 haplotypes boosts rare variant imputation and genome-wide association analysis in Asian populations

Received: 6 December 2023

Accepted: 2 December 2024

Published online: 30 December 2024

Check for updates

A list of authors and their affiliations appears at the end of the paper

Limited whole genome sequencing (WGS) studies in Asian populations result in a lack of representative reference panels, thus hindering the discovery of ancestry-specific variants. Here, we present the South and East Asian reference Database (SEAD) panel (https://imputationserver.westlake.edu.cn/), which integrates WGS data for 11,067 individuals from various sources across 17 Asian countries. The SEAD panel, comprising 22,134 haplotypes and 88,294,957 variants, demonstrates improved imputation accuracy for South Asian populations compared to 1000 Genomes Project, TOPMed, and ChinaMAP panels, with a higher proportion of well-imputed rare variants. For East Asian populations, SEAD shows concordance comparable to ChinaMAP, but outperforming TOPMed. Additionally, we apply the SEAD panel to conduct a genome-wide association study for total hip (Hip) and femoral neck (FN) bone mineral density (BMD) traits in 5369 genotyped Chinese samples. The single-variant test suggests that rare variants near *SNTG1* are associated with Hip BMD (rs60103302, MAF = 0.0092, $P = 1.67 \times 10^{-7}$), and variant-set analysis further supports the association ($P_{slide\_window} = 9.08 \times 10^{-9}$, $P_{gene\_centric} = 5.27 \times 10^{-8}$). This association was not reported previously and can only be detected by using Asian reference panels. Preliminary in vitro experiments for one of the rare variants identified provide evidence that it upregulates *SNTG1* expression, which could in turn inhibit the proliferation and differentiation of preosteoblasts.

Genotype imputation is an effective approach to reduce sequencing costs and has become a critical step in genome-wide association studies (GWAS). It can increase the likelihood of identifying likely causal variants[1], facilitate meta-analysis[2,3], and aid in finding pleiotropic effects of risk variants[4,5]. Factors such as haplotype size, ancestry diversity and sequencing depth can affect the imputation accuracy[6]. Generally, a diverse reference panel can improve accuracy in genetically diverse populations[7], while an ancestry-specific reference panel can benefit the corresponding population[8,9]. In our previous study, we found that the imputation accuracy for European population benefited from increased haplotype size and population diversity of the reference panel, while the accuracy for the Han Chinese population reached its peak with a modest addition of diverse samples (8-21%) in the reference panel[10]. Given that imputation accuracy directly impacts the credibility of subsequent analysis, it is crucial to select an appropriate reference panel before imputation.

The 1000 Genomes Project (1kGP) is one of the most renowned and widely used reference panels for genotype imputation, comprising 2504

e-mail: houf.zheng@suda.edu.cn

individuals from 26 populations[11]. The latest update to 1kGP phase 3 sequenced 3202 samples (including an additional 698 samples related to the previous participants) and identified 70,594,286 single nucleotide polymorphisms (SNPs) with 30× depth. Since many rare and low-frequency variants tend to be population-specific[11], an increasing number of human genome projects focus on specific populations to provide population-specific reference panels. However, most of the whole-genome sequencing (WGS) efforts were carried out in individuals of European descent, such as the GoNL project (the Genome of the Netherlands project, $n = 769$)[12], UK10K project (~4000 WGS and ~6000 whole exome sequencing samples in the UK)[13,14] and the TOPMed program (Trans-Omics for Precision Medicine, n =133,597)[15]. These efforts have enabled the development of large-scale combined reference panels for the European population, such as the HRC panel (Haplotype Reference Consortium, n = 32,470)[16]. The update of analytical methods and the emergence of various imputation reference panels have improved the imputation quality for European population. However, one of the complex legacies of the Human Genome Project, after 20 years of development, is that a lack of diversity might hinder the promise of genome science[17]. Recent evaluations of global populations by TOPMed revealed that even with a sample size of over 130,000 individuals, including 14,000 Asian samples, the imputation accuracy for some Asian populations remains insufficient[18]. As the largest continent, Asia accounts for 59.5% of the worldwide population (https://www.worldometers.info/world-population/). Fortunately, the Asian-ancestry populations have recently been sequenced and analyzed to understand the genetic basis of Asian populations., such as in the Japanese[19], Korean[20], Singaporean[21], Indian[22] and Chinese[23–26] populations. Among these efforts, our team initiated the Westlake BioBank for Chinese (WBBC) project[25,27] in 2017. Up to now, we have 4480 Chinese samples whole-genome sequenced (WBBC-seq) and 6080 samples whole-genome genotyped (WBBC-chip), covering 29 of 34 administrative divisions of China[25,28].

Rare variants, which may have low level of pairwise linkage disequilibrium with common variants, can potentially result in significant functional consequences[29]. Next-generation sequencing, with adequate coverage, enables the accurate detection of rare variants[30]. Rare variants are harder to impute than common variants because rare variants only appear a few times in the reference panel. Further, the evaluation of rare variants imputation in Asian populations is hindered by the limited sample size of currently published studies, which typically comprise only a few thousand individuals.

In this study, we integrate WGS data from multiple sources, including the Singapore SG10K pilot project (13.7×, 4563 samples), the GenomeAsia (GAsP) pilot project (36×, 1031 samples), WBBC pilot project (13.9×, 4480 samples), and the high-coverage 1kGP-Asian phase 3 (30×, 993 East and South Asian). These data are combined to create the South and East Asian reference Database (SEAD) panel, encompassing a total of 11,067 individuals, making it a large-scale reference panel for Asian populations. We then focus on evaluating the imputation quality of this panel to impute rare variants in Asian populations. Finally, we apply this augmented panel to the WBBC-chip data to explore its role in imputing likely causal rare variants for bone-related traits.

Results

Integration of the SEAD reference panel

First, we conducted SNP calling and joint calling to merge WBBC-seq (4,480 samples) and 1kGP-Asian (993 samples) datasets by using DeepVariant and GLnexus. The number of variants identified by joint calling for WBBC-seq individuals ranged from 3,220,000 to 3,400,000 with a transition/transversion (ts/tv) ratio of 2.136-2.152. For the WBBC and 1kGP (WBKG) joint-calling, the variant count ranged from 3,000,000 to 3,250,000 with a ts/tv ratio of 2.142-2.162, indicating that both the number and quality of variants were within reasonable limits (Supplementary Fig. 1). Additionally, we evaluated the sequencing quality by comparing the concordance rate in the 179 samples that had both WBBC-seq and WBBC-chip data, considering the chip data as the "true" values. The majority of samples exhibited very high concordance rate around 0.99 for the non-reference allele, heterozygote and homozygote genotypes, validating the use of these variants for further evaluation (Supplementary Fig. 2). After excluding singletons, the WBKG panel contained 5473 samples and 43,542,610 variants.

Because of limited access to the SG10K and GAsP data, we applied a reciprocal imputation approach with the similar protocol as the UK10K did[31]: First, we merged the reference panel of SG10K (4563 samples and 95,597,234 variants) and WBKG by imputing them with each other, we then combined the haplotypes from GAsP (1031 samples and 63,925,145 variants) to the WBKG-SG10K panel. Finally, we obtained the SEAD (South and East Asian reference Database) panel by combining all the panels together (Fig. 1 Step1).

Since the SEAD panel was derived from 4 different datasets, we evaluated for batch effects by principal component analysis (PCA). The results showed that the WBBC-seq matched very well with the 1kGP-EAS (Supplementary Fig. 3A). The SG10K (comprising three populations: Indonesian, Chinese, and Indian) and the GAsP (including a more diverse range of Asian populations), clustered with EAS and SAS, and populations in between (Supplementary Fig. 3B, C). Further extraction of sites with genotype quality (GQ) >40 and depth of coverage (DP) >10 from the GAsP population yielded almost identical results (Supplementary Fig. 3D), confirming the absence of batch effects in the datasets used for constructing the haplotype reference panel.

After removing the singletons, we finally obtained 88,294,957 variants and 22,134 haplotypes for the SEAD panel. The SEAD reference panel is now integrated into an imputation server with a user-friendly web interface for public use (https://imputationserver.westlake.edu.cn/).

SEAD panel improves precision in detecting rare variants in South Asian population

To evaluate the imputation performance of the SEAD panel in South Asian population, we generated three datasets from UK biobank samples with different proportions (50–70%, 70–90% and >90%) of SAS ancestry composition calculated by Rye software[32] (Fig. 1 Step 2). We revealed that the SEAD panel consistently exhibited the highest proportion of well-imputed low-frequency sites (Rsq > 0.8 and MAF < 5%) across all ancestry gradients, particularly in the >90% group (Fig. 2A). In contrast, the proportion of well-imputed sites in TOPMed imputation decreased when the SAS ancestry components increased, and fell to less than half from the 50–70% group to the >90% group (Fig. 2A). Similarly, the imputation performance of the ChinaMAP panel declined with increasing South Asian ancestry. The proportion of high quality sites showed little variation across the gradients in 1kGP imputation (Fig. 2A and Supplementary Data 1). These analyses suggested that despite its large sample size (>130,000), the TOPMed panel performed less effectively in South Asian populations, whereas SEAD panel demonstrated superior performance.

To further assess the imputation accuracy across Asian populations, we segmented 197 Central and South Asian samples from the HGDP, extracted ASA array sites on chromosome 2 to create a pseudo array, and then calculated the concordance rate between the imputed and WGS data (Fig. 2B, C, Supplementary Data 2). Compared to the other three panels, the SEAD panel demonstrated a higher concordance rate for heterozygote and homozygote genotypes, while the ChinaMAP panel exhibited the lowest concordance (Fig. 2B), with 1kGP and TOPMed panels in the middle. Similar trends were observed for specificity and precision (Fig. 2C). In order to further refine the advantage of each panel, we examined the performance in each Central and South Asian population within the HGDP samples. The results showed that the SEAD panel consistently demonstrated the highest imputation precision across most populations, often exceeding 0.94

https://doi.org/10.1038/s41467-024-55147-4

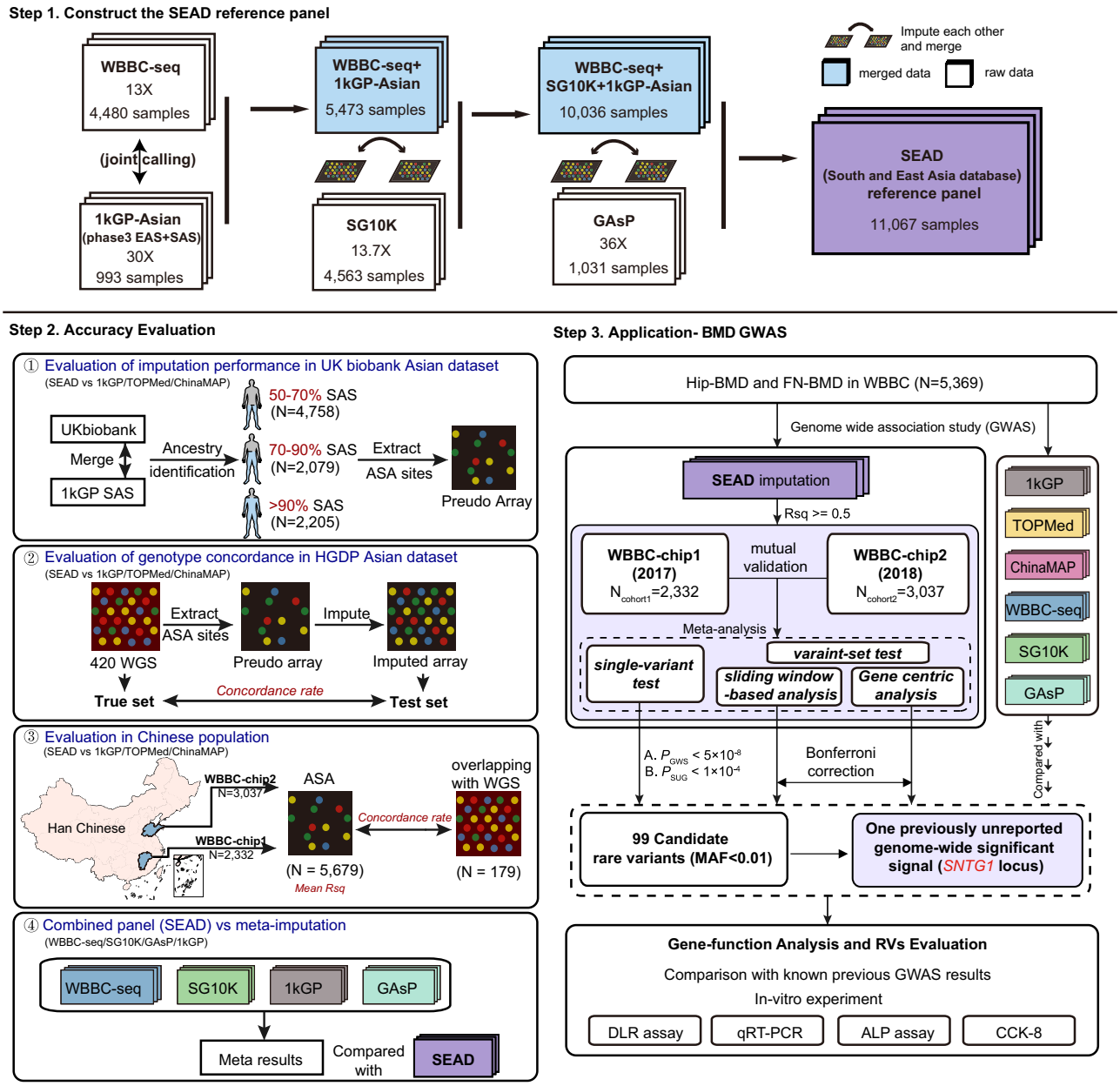


Fig. 1 | Study design. Step 1, the construction process of the SEAD reference panel, we conducted SNP calling and joint calling to merge WBBC-seq (4480 samples) and 1kGP-Asian (993 samples, EAS + SAS) datasets by using DeepVariants and GLnexus, then imputed and merged the obtained panel with SG10K and GAsP panels successively by reciprocal imputation. Finally, the SEAD panel contained 11,067 samples with 88,294,957 variants. Step 2, the imputation performance was compared between SEAD panel and other panels (1kGP, TOPMed, ChinaMAP) in South Asian and East Asian populations, and between meta imputation and combined panel. The human characters in the plot are designed by Freepik (https://www.freepik.com/). Step 3, the application of SEAD reference panel in FN and Hip BMD GWAS analyses.

(Supplementary Fig. 4). For MAF < 0.05 variants, the SEAD panel showed a distinct advantage over the TOPMed and ChinaMAP panels (Fig. 2D and Supplementary Data 3). The ChinaMAP panel consistently exhibited the lowest accuracy, with a median accuracy ranging from 0.7 to 0.8 for low-frequency and rare variants across most populations, reflecting its poorer imputation quality (Fig. 2D and Supplementary Data 3). All these results suggested that the SEAD panel demonstrated optimal performance in South Asian populations.

SEAD yields superior imputation performance in East Asia populations

We further evaluated the imputation accuracy of the 1kGP, TOPMed, ChinaMAP, and SEAD panels in East Asian populations (Fig. 1 Step 2). Initially, we extracted WGS data from 223 East Asian samples in the HGDP dataset and created a pseudo array by extracting ASA chip SNPs. The results showed that the peak of SEAD in the density plots was higher than that of ChinaMAP, indicating that the distribution of concordance rate was more concentrated (Fig. 3A and Supplementary Data 4), while the ChinaMAP panel had the highest heterozygote and homozygote genotype concordance rates. Concordance rates for the TOPMed and 1kGP panels were lower than those of the two Asian panels (Fig. 3A and Supplementary Data 4). For each East Asia population, the SEAD panel outperformed both TOPMed and ChinaMAP panels in Cambodian (which is a Southeast population) and had similar accuracy to the ChinaMAP panel in the Japanese population (Supplementary Fig. 5). In other East Asian populations, the ChinaMAP panel

https://doi.org/10.1038/s41467-024-55147-4

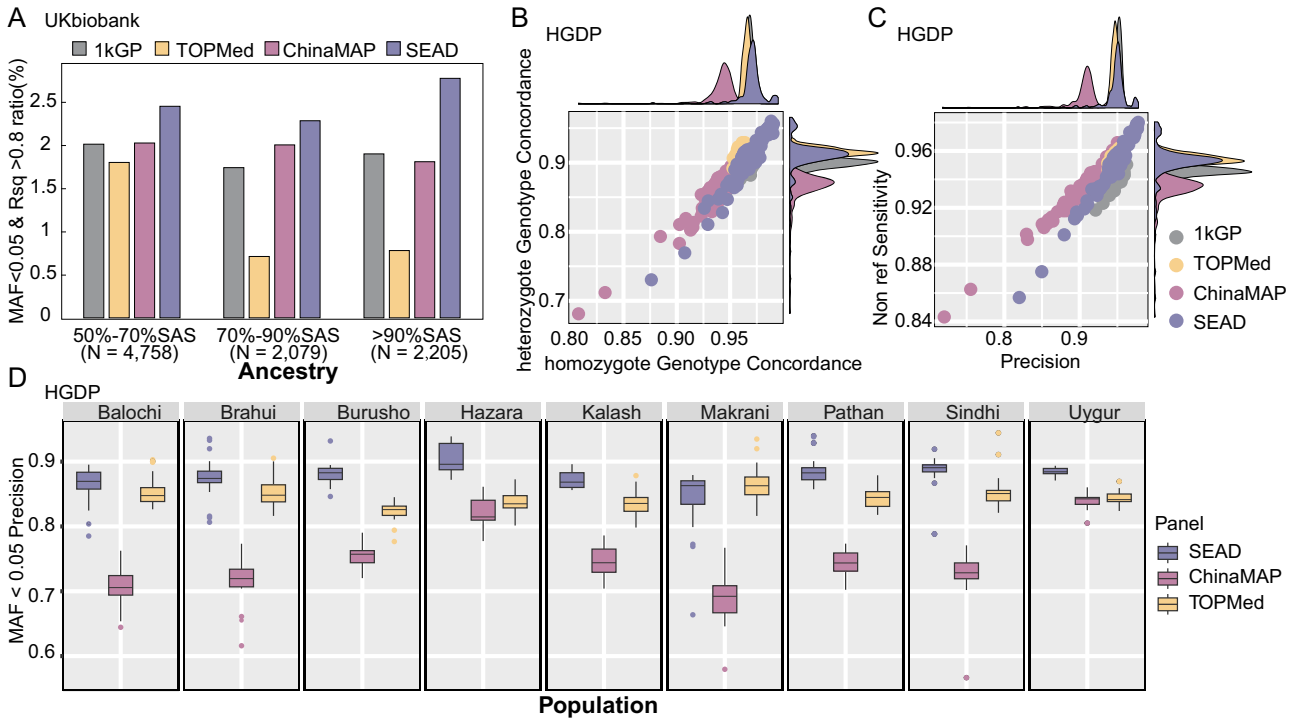


Fig. 2 | The imputation performance in imputing Central and South Asian populations. A the ratio of well imputed low-frequency (MAF < 0.05) variants with SEAD, 1kGP, TOPMed, ChinaMAP panels in samples with 50–70%, 70–90% and >90% SAS ancestry composition from UKbiobank. **B** Non-reference allele (NR-allele) concordance rate distribution (imputed variants vs. WGS variants) in 197 Central and South Asian samples from HGDP. Each dot represents an individual. The plots on the top and right are the corresponding density distributions. **C** No-ref specificity and precision (imputed variants vs. WGS variants) in 197 Central and South Asian samples from HGDP. The plots on the top and right are the corresponding density distributions. **D** Precision of low-frequency (MAF < 0.05) variants in 9 Central and South populations from HGDP. The sample sizes for each population group are as follows: Balochi ($n = 24$), Brahui ($n = 25$), Burusho ($n = 24$), Hazara ($n = 19$), Kalash ($n = 22$), Makrani ($n = 25$), Pathan ($n = 24$), Sindhi ($n = 24$), and Uygur ($n = 10$). Box plots indicate median (middle line), 25th, 75th percentile (box) and 1.5 times the inter-quartile range from the first and third quartiles (whiskers) as well as outliers (single points). All calculation performed on chromosome 2.

consistently showed higher accuracy than the SEAD panel, while TOPMed consistently had the lowest accuracy (Supplementary Fig. 5).

Due to the limited sample size in HGDP East Asian samples (only 223 individuals), it is unable to reflect the imputation performance in MAF (minor allele frequency) bins. We then imputed WBBC-chip data for 5679 Han Chinese samples using the four panels. The results showed that the ChinaMAP panel consistently demonstrated the highest accuracy (mean Rsq) across all MAF bins, while the SEAD panel outperformed TOPMed and 1kGP panels (Fig. 3B and Supplementary Data 5). In terms of the number of well-imputed sites (Rsq > 0.8), the count of the sites for SEAD grew closer to that of the ChinaMAP panel as the MAF increased (Fig. 3B). Additionally, we evaluated the imputation quality by comparing the concordance rate in the 179 samples who had both WBBC-seq and WBBC-chip data, considering the sequencing data as true. The results showed that the SEAD and ChinaMAP panels achieved similar and higher homozygote and heterozygote concordance rates, with the lowest concordance rate for the TOPMed panel (Fig. 3C and Supplementary Data 6). Similar results were observed for non-reference allele sensitivity and precision (Fig. 3D and Supplementary Data 6). These results indicated that the concordance rates of the SEAD and ChinaMAP panels were comparable, whereas the imputation performance of the TOPMed panel was inferior to that of the Asian panels.

Imputation with combined panel outperforms that from meta-imputation

We compared the performance of meta-imputation of four panels (WBBC-seq, 1kGP, SG10K, GAsP) with the combined SEAD panel (Fig. 1 Step 2). Across the seven MAF bins, the number of well-imputed variants obtained through SEAD imputation consistently exceeded those of meta-imputation, particularly for rare/low-frequency variants (Fig. 3E and Supplementary Data 7). Furthermore, the Rsq values achieved by SEAD imputation were higher in each MAF bin than those obtained through meta-imputation, with an increase of ~0.2 in the 0.1–1% MAF bin (Fig. 3E and Supplementary Data 7). Additionally, we calculated the proportion of well-imputed variants relative to the total number of variants within each MAF bin, where SEAD also demonstrated a clear advantage (Supplementary Fig. 6). These results suggested that the combined SEAD panel consistently showed significantly higher well-imputed counts and accuracy across the seven MAF bins compared to the meta-imputation, indicating that the imputation with the combined panel is superior to meta-imputation.

Employment of the SEAD panel in bone mineral density GWAS analysis

After imputing WBBC-chip data (5369 samples) with the SEAD panel, we conducted GWAS on two BMD traits: total hip (Hip) and femoral neck (FN) BMD (Fig. 1 Step3). These two BMD traits were highly correlated both in Pearson correlation ($r = 0.918$) and genetic correlation ($r = 0.932$, SD = 0.025). With a genome-wide significance threshold ($P < 5 \times 10^{-8}$), two common variants (MAF > 0.01) that have been previously reported in GWAS were identified to be associated with both BMD traits, with the top signals at chr1:rs9659023 (an intronic variant of *FMN2*, this locus was reported to be associated with heel[33] and total body BMD[34]) and at chr6:21392233:GA:G (near *SOX4*, this locus was reported to be associated with heel[33] and FN BMD[13]) (Supplementary Fig. 7).

https://doi.org/10.1038/s41467-024-55147-4

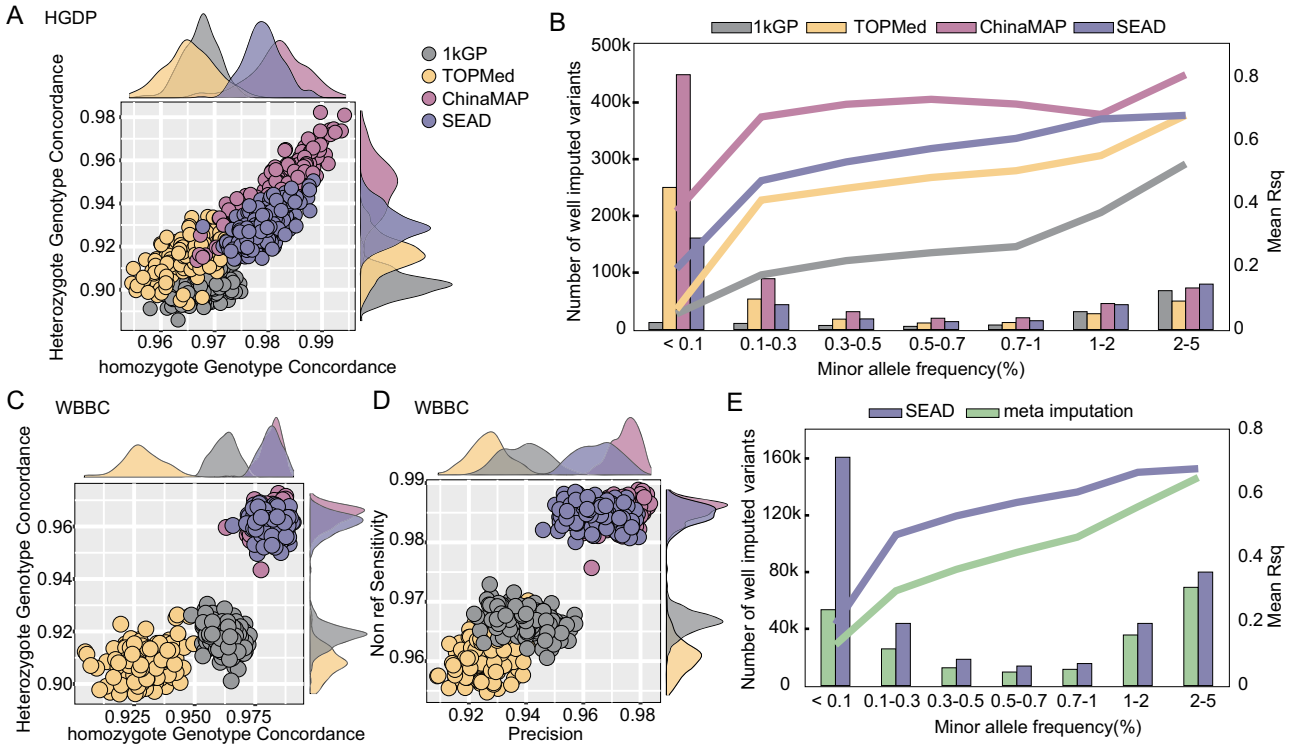


Fig. 3 | The imputation performance in imputing East Asian populations. **A** Non-reference allele (NR-allele) concordance rate distribution (imputed variants vs. WGS variants) in 223 East Asian samples from HGDP. Each dot represents an individual. The plots on the top and right are the corresponding density distributions. **B** The average imputed r-square (Rsq) (line plot) and number of well-imputed (Rsq > 0.8) variants (bar plot) of four reference panels among 7 MAF bins including <0.1%, 0.1%–0.3%, 0.3%–0.5%, 0. 5%–0.7%, 0.7%–1%, 1%-2%, and 2%–5%. **C** Non-reference allele (NR-allele) concordance rate distribution (imputed variants vs. WGS variants) in 179 overlapping samples (both sequenced in WBBC-seq and genotyped in WBBC-chip). Each dot represents an individual. The plots on the top and right are the corresponding density distributions. **D** Non-reference specificity and precision (imputed variants vs. WGS variants) in 179 overlapping samples. The plots on the top and right are the corresponding density distributions. **E** The average Rsq (line plot) and number of well-imputed (Rsq > 0.8) variants (bar plot) of SEAD reference panel and meta imputation with 1kGP, WBBC-seq, SG10K and GAsP among 7 MAF bins including <0.1%, 0.1%–0.3%, 0.3%–0.5%, 0.5%–0.7%, 0.7%–1%, 1%–2%, and 2%–5%. All evaluations conducted on chromosome 2.

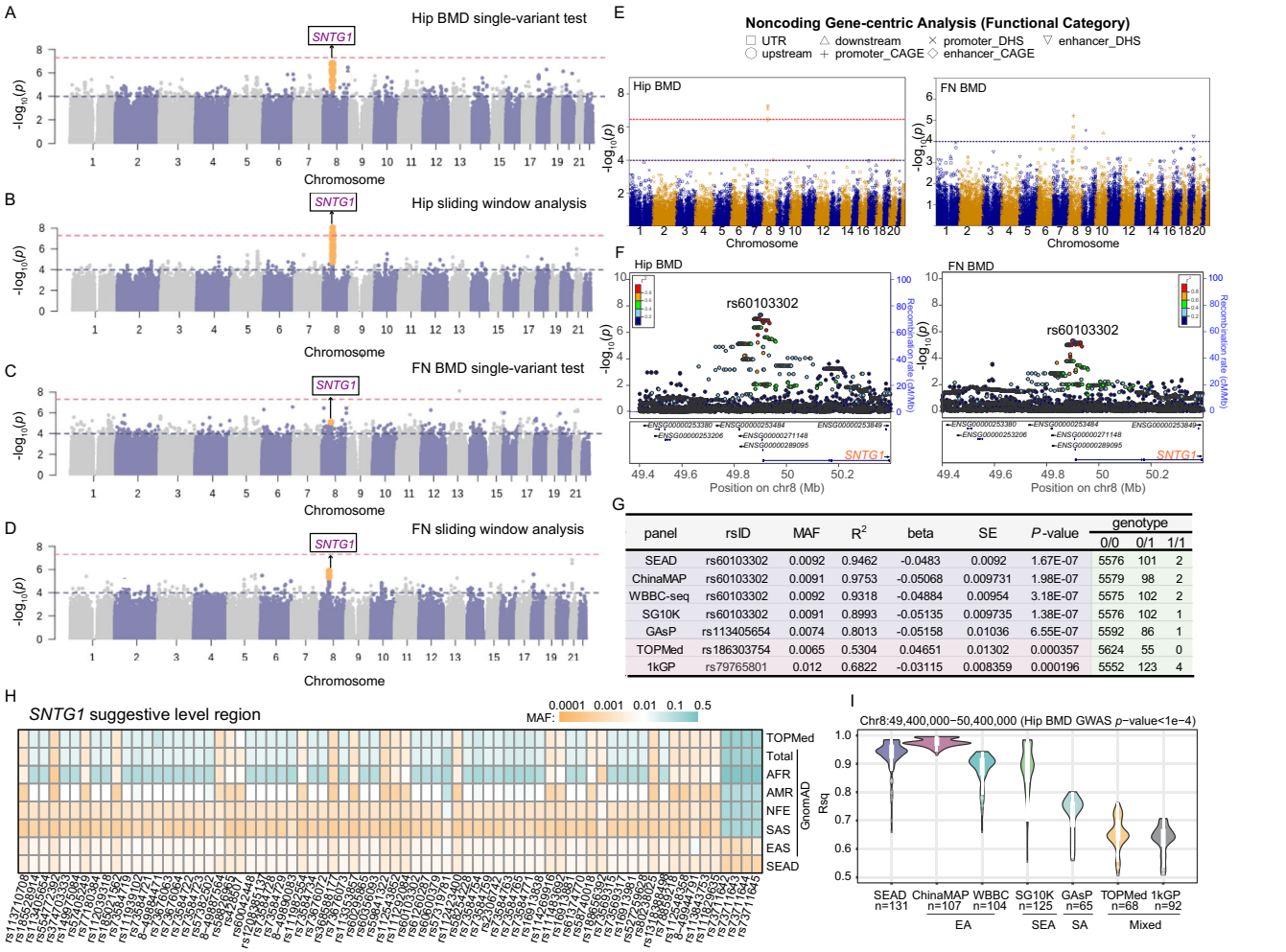
For the rare variants (0.001 < MAF < 0.01), we utilized two analytical strategies: single-variant test with PLINK and variant-set analysis via the STAARpipeline (incorporating both sliding-window and gene-centric methods). In the single-variant test analysis, we prioritized the variants with small *P*-value and replication (see Methods), and identified 106 suggestive variants for Hip BMD (Supplementary Data 8). Among these variants, 68 (64.1%) were annotated to *SNTG1* gene by ANNOVAR (Supplementary Data 8). Four rare variants (rs60103302, rs61260287, rs60600379 and rs57319781), in complete linkage disequilibrium (LD $r^2 = 1.00$), were identified for Hip BMD at near genome-wide significance level ($P = 1.6$ '$\times 10^{-7}$, MAF = 0.0092), located in the intergenic region near *SNTG1* on chromosome 8 (Fig. 4A and F and Supplementary Data 8). As for FN BMD, 1 out of the 35 suggestive variants were annotated to *SNTG1* gene (Fig. 4C and F, and Supplementary Data 8). In the variant-set analysis, we first implemented a sliding window of 2 kb in the STAAR-O test adhering to the previous "discovery and replication" strategy. The results showed that 5 clustering regions were identified as the suggestive signals for Hip BMD (Fig. 4B, Supplementary Fig. 8 and Supplementary Data 9). In the cluster on chromosome 8 (chr8:49873459-49958458), a total of 30 windows were suggested, of which, 3 windows annotated near *SNTG1* were discovered at bonferroni-corrected level $P = 3.59 \times 10^{-8}$ (0.05 divided by 1,392,296 windows), with the top window at chr8:49903244-49905243 ($P = 9.08 \times 10^{-9}$) (Fig. 4B, Supplementary Fig. 8 and Supplementary Data 9). Moreover, in the annotation-based gene-centric analysis, of the 7 non-coding categories of *SNTG1* gene, 8 categories showed suggestive signal for Hip BMD, with the top signal at promoter region ($P_{promoter\_CAGE} = 5.72 \times 10^{-8}$) (Fig. 4E, Supplementary Fig. 8 and Supplementary Data 10), showcasing at least one annotation reaching the genome-wide Bonferroni-corrected level $P = 3.57 \times 10^{-7}$ (0.05 divided by 20,000 genes then divided by 7 categories). In summary, *SNTG1* locus was highlighted by both analytical strategies, using the strictest threshold for Hip BMD, and also achieved a suggestive level in FN BMD ($P_{single\_variant} = 7.79 \times 10^{-6}$, $P_{slide\_window} = 1.34 \times 10^{-6}$, $P_{gene\_centric} = 5.71 \times 10^{-6}$) (Fig. 4C–E, Supplementary Fig. 8, Supplementary Data 9 and 10).

We compared the frequency distribution of the suggestive rare signals identified in the *SNTG1* locus (72 SNPs) with the gnomAD and TOPMed databases (as they both have very big sample size) across various populations (Fig. 4H, Supplementary Data 11). The *SNTG1* variants were predominantly rare in the Non-Finnish European (NFE: 34,029 samples) and South Asian (SAS: 2419 samples), while they were more common in the African/African American (AFR: 20,744 samples) and TOPMed (dominantly European population), highlighting the population-specific patterns of this region. The MAF of *SNTG1* variants imputed by the SEAD panel most closely aligned with the MAF value in the East Asian (EAS) population (2604 samples) within gnomAD v3 (Fig. 4H), validating the appropriateness of SEAD panel to identify Asian-specific associated variants.

SEAD panel compared with other panels in detecting unreported Asian-specific *SNTG1* locus

To test if other panels are also capable of identifying the *SNTG1* locus, we imputed WBBC-chip data using the TOPMed, 1kGP, SG10K, GAsP, ChinaMAP, and WBBC-seq panels and performed association analyses on Hip BMD. We assessed the imputation quality (Rsq) of associated variants (meta *P*-value < 1×10^{-4}) for the WBBC-chip dataset imputed with each panel within the *SNTG1* region on

https://doi.org/10.1038/s41467-024-55147-4



panel	rsID	MAF	R^2	beta	SE	*P*-value	genotype 0/0	0/1	1/1
SEAD	rs60103302	0.0092	0.9462	-0.0483	0.0092	1.67E-07	5576	101	2
ChinaMAP	rs60103302	0.0091	0.9753	-0.05068	0.009731	1.98E-07	5579	98	2
WBBC-seq	rs60103302	0.0092	0.9318	-0.04884	0.00954	3.18E-07	5575	102	2
SG10K	rs60103302	0.0091	0.8993	-0.05135	0.009735	1.38E-07	5576	102	1
GAsP	rs113405654	0.0074	0.8013	-0.05158	0.01036	6.55E-07	5592	86	1
TOPMed	rs186303754	0.0065	0.5304	0.04651	0.01302	0.000357	5624	55	0
1kGP	rs79765801	0.012	0.6822	-0.03115	0.008359	0.000196	5552	123	4

Fig. 4 | Detecting rare locus of Hip and FN BMD in GWAS. A, **C** Meta results of the single-variant test of Hip and FN BMD. The *P*-value of significance threshold of 5×10^{-8} delineated by red lines, while the *P*-value of significance threshold of 1×10^{-4} delineated by blue lines. **B**, **D** The sliding window analysis of Hip and FN BMD. The *P*-value of significance threshold of 5×10^{-8} delineated by red lines, while the *P*-value of significance threshold of 3.55×10^{-8} (0.05/1,400,000) delineated by blue lines. The *SNTG1* gene can be detected by four approaches and labeled with a rectangle. **E** Noncoding Gene-centric analysis of Hip and FN BMD. **F** Locuszoom plots of *SNTG1* gene in Hip and FN BMD single-variant test imputed by SEAD panel. The most significant SNP is rs60103302 at the *SNTG1* locus. SNPs were plotted based on their GWAS −log10 (*P*-values) and genomic position. The color scale of r^2 values is used to label SNPs based on their degree of linkage disequilibrium with the most significant SNP. Recombination rates calculated from SEAD reference data are also displayed in a blue line corresponding to the right vertical axis. **G** Information of top associated SNPs with Hip BMD in single-variant test imputed by seven panels. **H** SNP frequency comparison in *SNTG1* gene regions. The databases contained TOPMed and GnomAD. **I** Violin plot of Rsq distribution in all GWAS *P*-value $< 1 \times 10^{-4}$ variants imputed by seven reference panels. Box plots indicate median (middle line), 25th, 75th percentile (box) and 1.5 times the inter-quartile range from the first and third quartiles (whiskers) as well as outliers (single points).

chromosome 8 (chr8:49,400,000-50,400,000) (Fig. 4I). The results showed that the mean imputation Rsq for the SEAD, ChinaMAP, WBBC-seq, and SG10K panels exceeded 0.9, while for the TOPMed and 1kGP panels, it ranged between 0.6 and 0.7 (Fig. 4I and Supplementary Data 12). After applying a less strict threshold of Rsq > 0.5, the association signals at the *SNTG1* locus were detectable in the WBBC-chip dataset imputed with all Asian panels (Supplementary Fig. 9), but not with the TOPMed and 1kGP panels (Supplementary Fig. 10). As for the most significant SNPs, the SEAD, ChinaMAP, WBBC, and SG10K panels demonstrated the same top SNP (rs60103302), with comparable effect size, MAF, and genotype counts, except for the GAsP panel (Fig. 4G). For the TOPMed and 1kGP panels, the most significant SNPs showed *P*-values $>1 \times 10^{-4}$ with low Rsq values (Fig. 4G and Supplementary Fig. 10). These results demonstrated that, although TOPMed possesses the largest-scale reference haplotypes and contains a subset of Asian samples, it is not efficient to impute data in Asian populations, emphasizing the significance of a population-specific reference panel.

Preliminary in vitro analysis of *SNTG1* gene on osteogenesis

To determine whether the associated SNP would affect the transcriptional activity of the *SNTG1* gene, we selected the intronic SNP rs111829635 (instead of the intergenic top SNP rs60103302) to perform the dual-luciferase assays in two cell lines, 293 T and MC3T3-E1 cell lines. 483-bp fragments containing the rs111829635 SNP were inserted into a luciferase reporter plasmid with respect to a minimal promoter. It was observed that the intronic SNP rs111829635 C-to-T mutation significantly increased luciferase activity in both 293 T cells and MC3T3-E1 cells (Fig. 5A), suggesting that the identified locus would alter the *SNTG1* gene expression.

Moreover, we utilized the CCK-8 proliferation assay to examine the effects of SNTG1 overexpression in the mouse preosteoblast MC3T3-E1. The results showed that the absorbance at 450 nm OD was decreased after *SNTG1* overexpression, and cell density was also decreased ($P < 0.05$, Fig. 5B and C), revealing that overexpression of *SNTG1* led to a significant reduction in cell proliferation. At 72 h, a noticeable decrease in the expression of osteogenic marker genes (*RUNX2*, *COL1A1*, and OCN) confirmed that *SNTG1* inhibited the

https://doi.org/10.1038/s41467-024-55147-4

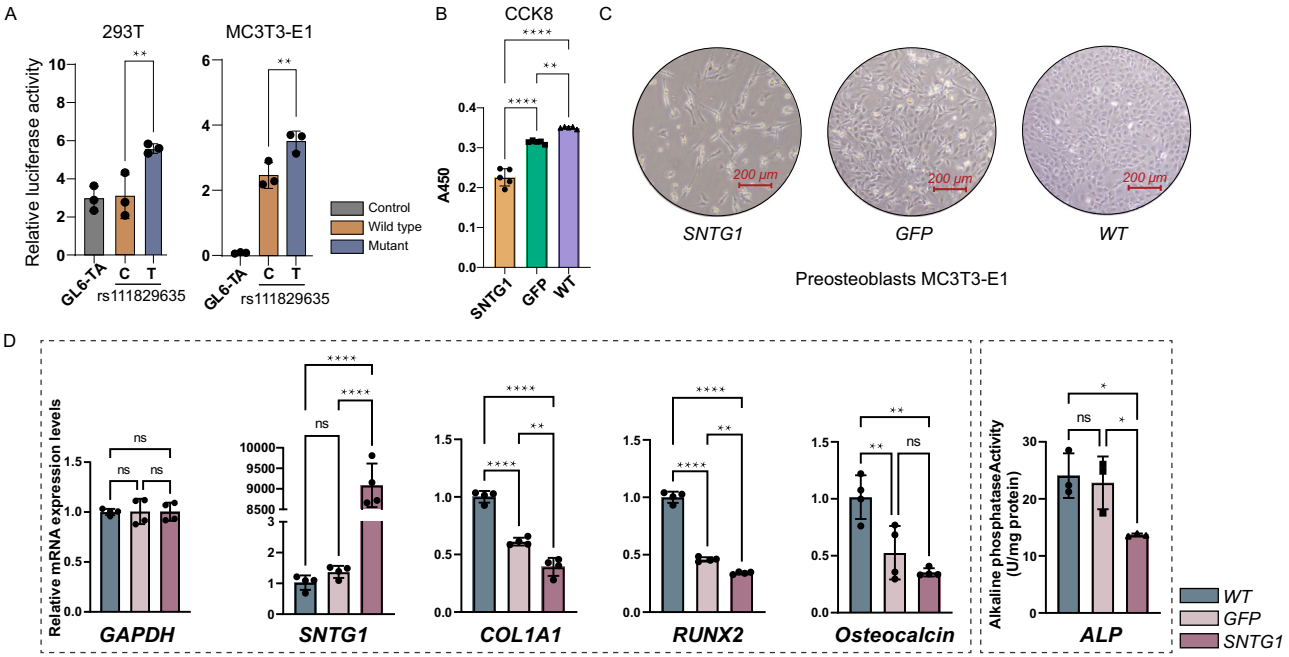


Fig. 5 | Preliminary in vitro analysis of *SNTG1* gene on osteogenesis. A The impact of rs111829635 alleles C and T on the expression of *SNTG1* in 293 T and MC3T3-E1 cells. $n = 3$. The statistic test is *T*-test. **B**, **C** The overexpression of *SNTG1* alone inhibits cell proliferation. $n = 5$. The statistic test is one-way anova in (**B**). **D** The overexpression of *SNTG1* inhibits cell differentiation. with *COL1A1*, *RUNX2*, and Osteocalcin serving as indicators of cell differentiation. $n = 4$. ALP refers to alkaline phosphatase. $n = 3$. The statistic test is one-way anova in **D**. Data are presented as mean + standard deviation (SD). *P*-values are two-sided and adjustments were not made for multiple comparisons. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; $^{****}P < 0.0001$. Source data are provided as a Source Data file.

differentiation of preosteoblast MC3T3-E1 cells ($P < 0.05$, Fig. 5D). Furthermore, we measured alkaline phosphatase (ALP) level, a marker of osteoblast activity, which showed a significant reduction after 6 days of *SNTG1* overexpression (Fig. 5D). Overall, these results supported the notion that *SNTG1* overexpression might inhibit human osteogenic cell proliferation and differentiation, highlighting its potential role in osteogenesis.

Discussion

In this study, we integrated whole-genome sequencing data from SG10K, GenomeAsia, WBBC, and 1kGP-Asian to create a combined reference panel for genotype imputation, the South and East Asian reference Database (SEAD). It comprised a diverse range of populations across Asia, including 11 populations from GenomeAsia, 1 population from WBBC, 3 populations from SG10K, and 8 populations from 1kGP-Asian. With a sample size of 22,134 haplotypes, the SEAD panel stands as one of the most comprehensive panels in terms of coverage across Asia. We compared the concordance rate of the genotypes imputed from the SEAD panel across Asian populations with the TOPMed, ChinaMAP and 1kGP panels, and suggested that the SEAD panel performed the best for the rare variants imputation in South Asian populations. Meanwhile, the concordance rates of the SEAD and ChinaMAP panels were comparable for imputing East Asian populations, whereas the imputation performance of the TOPMed panel was inferior to that of the Asian panels. Finally, we applied the SEAD panel to the bone mineral density GWAS analyses in WBBC-chip data, and identified an Asian-specific rare locus, *SNTG1*, that was not reported even in the large biobank-scale GWAS.

Asia, being the largest and most populous continent worldwide, boasts a wealth of human genetic resources. However, most of the whole-genome sequencing (WGS) efforts were carried out in Caucasian populations in the last decade[17]. In our previous study, we conducted a thorough evaluation and discussion on the imputation of rare variants, highlighting the necessity of constructing a haplotype reference panel for Asian populations[10,35–37]. In recent years, there has been a notable surge in WGS data across Asia, particularly in regions such as Japan[19], Singapore[21], Korea[38] and China[23,25,39]. Despite the abundance of WGS projects, a large-scale reference panel with broad geographical coverage across Asia is still needed. The accumulation of such datasets from diverse Asian populations presents a unique opportunity for amalgamating multiple WGS datasets into a singular, more comprehensive, and expansive reference panel, which would encapsulate a broader spectrum of genetic diversity, thereby enhancing its utility for human genetic study in Asian populations. The recently published Northeast Asian Reference Database, with over half of the samples deriving from Japan and Korea, represented populations from Northeast Asia[40]. In our study, we conducted joint-calling between WBBC-seq and 1kGP-Asian, followed by mutual imputation with SG10K and GAsP to construct the South and East Asian reference Database (SEAD) panel, covering the widest geographical area in Asia, with the majority of its population originating from South and East Asia.

The state-of-the-art imputation reference panel released by the TOPMed includes a diverse range of populations, such as African Americans, Hispanic/Latino, Asian and Caucasian populations. Despite the large and diverse samples in the TOPMed panel, imputation for some populations, notably those from Asia, Oceania and the Pacific, would not fully benefit from it[18]. In our study, we observed that the TOPMed panel performed less effectively than the SEAD panel in both South Asian and East Asian populations, but better than the 1kGP panel in East Asian populations, particularly for the low-frequency variants. The ChinaMAP panel contains the largest number of Chinese samples[24], but it performed the worst in all the Central & South Asian populations, in which the SEAD panel performed the best. Benefiting from the large sample size, the ChinaMAP panel indicated its superior performance in the Chinese population. Prospectively, we will update the SEAD panel in the future with additional East Asian samples to improve its imputation efficiency in the corresponding population. All in all, the SEAD reference panel demonstrated significant advantages in genotype imputation for South Asian populations and exhibited high accuracy when imputing East Asian populations.

In the comparison between TOPMed and ChinaMAP, the genetic ancestry fraction of the UK biobank samples in Fig. 2A was calculated based on the 1kGP (5 super populations) super population South Asian (SAS) (GIH, Gujarati Indian from Houston, Texas; PJL, Punjabi from Lahore, Pakistan; BEB, Bengali from Bangladesh; STU, Sri Lankan Tamil from the UK; ITU, Indian Telugu from the UK). While Central & South Asia populations in Fig. 2D were drawn based on HGDP (7 super populations). The SAS was not classified in HGDP super populations. A previous study including both HGDP and 1kGP data suggested that Central & South Asian populations in HGDP were positioned between 1kGP Asian and European populations[41], that might explain why TOPMed panel yielded a smaller proportion of well-imputed rare sites in South Asian populations (Fig. 2A), but achieved greater accuracy and precision in Central & South Asian populations, compared to ChinaMAP panel (Fig. 2D). Additionally, some populations (such as the Uygur, Burusho and Hazara) in Central & South Asia, exhibiting a higher proportion of EAS ancestry[42], have closer imputation accuracy between TOPMed and ChinaMAP in our study (Fig. 2D).

On the other hand, meta-imputation could integrate the imputation results generated using separate reference panels into a consensus dataset[43]. And Yu et al. demonstrated that this method might generate comparable accuracy to the imputation with a combined reference panel[43]. However, based on our assessment, the combined SEAD panel consistently showed significantly higher well-imputed counts and accuracy across each MAF bin compared to meta-imputation with separate reference panels. In addition, meta-imputation might not improve imputation accuracy for under-represented populations[18].

A selected group of phenotypes were collected within the WBBC project, and the bone-related traits were the main focus[27]. As we reviewed before, despite the fruitful GWAS discoveries in the bone field, few large-scale GWAS studies on the BMD trait have been carried out in the Chinese population[44], noting that the overall prevalence of osteoporosis in the Chinese population exceeded the global average[45]. Therefore, after systematic and comprehensive imputation evaluation for the panels, we applied the SEAD panel to a GWAS analysis for BMD traits at Hip and femoral neck (FN). These two traits were highly correlated in either phenotypic or genetic correlation. Therefore, the association signals reported in both traits would reduce the random error of the results. Through both single-variant association test and variant-set analysis, we detected that rare variants near *SNTG1* gene were associated with Hip BMD. The *SNTG1* gene also reached the suggestive threshold in FN BMD analysis using both methods. The gene and variants were not reported even in large-scale biobank GWAS for BMD[33], and the *SNTG1* variants were predominantly rare in the Non-Finnish European and more common in the African/African American and Latino/Admixed American. The *SNTG1* signal for Hip BMD was identified in GWAS imputed with all Asian panels, but not the TOPMed and 1kGP panels. Specifically, the rare variant rs60103302 (MAF = 0.0092) was identified to be the top signal in imputation with the SEAD panel, as well as the SG10K, WBBC-seq and ChinaMAP panels. However, this SNP had poor Rsq in TOPMed imputation (Rsq = 0.4172), and no SNPs showed SNPs showed *P*-values less than 1×10^{-4}. This is because the low-frequency and rare variant associations are more likely to be population-specific, which would suffer a relative loss in statistical power in GWAS[18]. Therefore, the incapability of TOPMed and 1kGP to impute Asian-specific rare variants further underscores the importance of building Asian-specific panels for the discovery of rare variants unique to Asian populations.

The *SNTG1* (Syntrophin Gamma 1) is a neuronal cell-specific protein prominently expressed in Purkinje neurons of the cerebellum and pyramidal neurons of the hippocampus and cortex[46], suggesting its potential role in neuronal stability and signal transmission within the nervous system[46]. The nervous system, particularly through the sympathetic nervous system (SNS), plays a crucial role in regulating bone metabolism. The SNS provides innervation to bone tissue, with the sympathetic fibers directly interacting with osteoblasts and osteoclasts[47]. Clinical evidence supported the involvement of the SNS in bone metabolism, linking conditions associated with increased sympathetic activity (such as reflex sympathetic dystrophy) to decreased bone mineral density[48,49]. The *SNTG1* was identified as a candidate gene for idiopathic scoliosis[50–52]. A southern Chinese cohort of patients with congenital scoliosis also identified copy number variants of *SNTG1*[53]. The underlying pathological mechanisms of idiopathic scoliosis remain unclear, a leading theory posits that the primary abnormality is a neurological defect. This defect causes abnormal processing in the central nervous system (CNS), leading to developmental anomalies in the spine[50]. Animal models of scoliosis indicated damage to the cerebellum in bipedal rats induced scoliosis[54]. Additionally, the SNTG1 protein was involved in the Dystrophin-associated glycoprotein complex (DGC) pathway. The DGC is known to play a significant role in maintaining the structural integrity of muscle fibers[55]. Reduction in muscle strength diminished mechanical stimuli to bones, adversely affecting bone remodeling and maintenance. Sarcopenia, characterized by decreased muscle mass and strength, can increase the risk of osteoporosis, particularly in the elderly[56,57]. Further, muscle atrophy could influence bone density through hormonal and inflammatory pathways[58,59]. In our study, we suggested that overexpression of *SNTG1* inhibited the proliferation and differentiation of preosteoblasts. Given that osteoblasts are pivotal in orchestrating bone formation and osteogenesis, a reduction in their proliferation and differentiation capacity translates into a diminished osteoblast pool. This, in turn, leads to a decrease in bone formation and the overall process of osteogenesis. This observation aligned with the direction of effect on BMD in our GWAS results.

In summary, we constructed a haplotype reference panel for Asian populations (SEAD panel) with the largest number of samples and the most abundant population diversity in Asia. The reference panel demonstrated excellent performance in imputing East Asian and Central and South Asian populations, especially in detecting rare variants. We provided this optimal imputation service online for free (https://imputationserver.westlake.edu.cn/) for genetic studies in Asian populations. By applying the SEAD panel to impute the genotyping array data in the Chinese population, we, for the first time, successfully identified rare variants near the *SNTG1* gene showing association with bone mineral density.

Methods

Variants calling and quality control for the Asian panels

We collected Asian whole genome sequencing data from the Westlake Biobank for Chinese (WBBC) pilot project[25], the newly released high-coverage 1000 Genomes project (1kGP)[60], SG10K pilot project[21] and the GenomeAsia pilot project (GAsP)[22] (Fig. 1 Step 1).

The WBBC dataset, which was generated previously[25], contains both whole genome sequencing (WBBC-seq) data and high-density Asian Screening Array (ASA) chip data (WBBC-chip) data[25]. The WBBC-seq dataset included 4480 Chinese samples with an average sequencing depth of 13.9×. The CRAM files of WBBC-seq were used to construct the SEAD panel and were stored locally. The CRAM files for the 993 EAS and SAS ancestry samples (1kGP-Asian) were downloaded here (ftp://ftp.sra.ebi.ac.uk/vol1/run/ERR323/). We performed variant calling on the 4480 samples of WBBC-seq and 993 samples of 1kGP-Asian using DeepVariant[61]. We used bcftools filter to segment each chromosome of each sample into 2.5 Mb subsets, followed by joint calling of the same regions across all samples using GLnexus[62]. This resulted in genotype files for 5473 WBKG (WBBC-seq and 1kGP-Asian) samples. Following our previous methods[25], we conducted quality control on the obtained genotype files. We used bcftools filter to remove SNVs and INDELs close to INDELs with SnpGap 3 and IndelGap 5, and filtered out INDELs

>50 bp. Individual genotypes with a genotype quality score (GQ) < 20 were set to missing. Missing and low-confidence genotypes were refined using BEAGLE 5.173[63]. The chromosomes were divided into chunks of 1 Mb with 0.1 Mb overlapping. Lastly, we excluded variants with a Hardy-Weinberg equilibrium (HWE) P-value $< 1 \times 10^{-6}$ using VCFtools (v 0.1.13)[64]. The high-coverage 1kGP dataset (2504 samples) with an average depth of 30×, based on the GRCh38 assembly, was downloaded from the https://www.internationalgenome.org/data-portal/data-collection/30x-grch38. This dataset, not 1kGP-Asian, was used in the comparison of imputation performance of different reference panels (TOPMed, ChinaMAP, 1kGP and SEAD).

The SG10K dataset comprises 4810 whole-genome sequencings (WGS) of Singaporean Chinese, Malays, and Indians, with an average depth of 13.7×. Since the SG10K dataset was based on Genome Reference Consortium Human build 37 (GRCh37), we employed Picard LiftoverVcf (http://broadinstitute.github.io/picard/) to update the genome assembly version from GRCh37 to GRCh38. We merged the 22 autosomes and set the following parameters: java -Xmx100g -XX:ParallelGCThreads = 8 -jar picard.jar LiftoverVcf I = inpute.vcf.gz O = SG10K_all.hg38.vcf.gz CHAIN = hg19ToHg38.over.chain.gz REJECT = rejected_variants.vcf R =Homo_sapiens_assembly38.fasta. The GAsP dataset includes 1163 WGS samples with an average depth of 36×, mainly covering Indians, Korean, Pakistanis, etc. The downloaded VCF files in the GAsP database are categorized into three types based on variant types: INDELs, multiallelic, and SNVs. Initially, each VCF file underwent a coordinate transformation from GRCh37 to GRCh38 using Picard LiftoverVcf. Subsequently, genotype phasing was performed using Shapeit v2 with parameters including windows of size 0.5, 200 states, and an effective size of 14,269, followed by normalization with BCFtools[65]. Multiallelic variants were split into biallelic forms, and positions marked with asterisks were removed. Finally, the three different types of VCF files were merged using BCFtools concat.

We identified relative pairs using KING[66] and excluded sample pairs with estimated kinship coefficients restricted to the range of 0.177–0.354. Ultimately, we retained 4480 samples from WBBC-seq, 993 samples from 1kGP-Asian, 4563 samples from SG10K, and 1031 samples from GAsP (Fig. 1 Step 1 and Supplementary Data 13).

Batch effects evaluated by principal component analysis

Given the differences in sequencing, batch processing, and quality control across the four datasets (WBBC-seq, 1kGP-Asian, SG10K, and GAsP), we evaluated the batch effect using principal component analysis (PCA). We initially conducted a PCA on the WBKG dataset obtained through joint calling mentioned above. Subsequently, we merged the WBKG dataset with the SG10K dataset. We separately converted the processed VCF files into PLINK formats, following the merging guidelines available at https://github.com/baharian/merge_PLINK. The pipeline initially examined each dataset individually, inspecting each chromosome for SNPs identified by multiple or different names (e.g., rsID) at the same position and removing any such occurrences. Subsequently, variants with missing calls were identified and removed from each dataset. The intersection of SNPs present in all datasets was determined, and SNPs not in this intersection were excluded. Next, the datasets were cross-examined for SNPs that might pose issues due to strand misassignment, reference/alternate allele misassignment, ambiguous strand or reference/alternate allele misassignment, and non-biallelic variants. Appropriate actions, such as flipping or removing SNPs, were taken for each category. After merging these datasets, an in-depth LD-based strand assignment cross-check was conducted on the merged data to identify and correct potential strand flips. We performed another PCA on the combined dataset (WBKG and SG10K). Finally, following the same methodology, we merged this dataset with the GAsP dataset and conducted another PCA analysis. The PLINK2.0 was used to compute the 20 principal components (PCs) with the parameters --pca approx 20 --maf 0.01 on the merged data. To further validate the PCA results, we extracted SNPs from the GAsP dataset with GQ > 40 and DP > 20, and computed the PCs using the same parameters.

Quality control of the WBBC-chip data

Besides the whole genome sequencing data, the WBBC study also genotyped 6080 individuals with the high-density Illumina Asian Screening Array (ASA, based on GRCh37), resulting in the identification of a total of 659,184 SNPs[25,27]. Among them, 184 samples underwent both whole-genome sequencing and array genotyping.

As part of quality control, we employed GCTA[67] to calculate the pairwise genetic relationship matrix using common variants and remove samples with a coefficient > 0.025 (Supplementary Data 14). We then excluded samples with missing call rates ≥ 5% and excluded SNPs with missing call rates ≥ 5%, MAF < 1% and Hardy-Weinberg equilibrium at $P < 1 \times 10^{-6}$ by using PLINK1.9[68]. The genotype assembly version of the WBBC-chip data was updated from GRCh37 to GRCh38. Finally, we retained a total of 470,242 variants and 5679 samples in the WBBC array dataset. The data were phased using SHAPEIT v2 with a window size of 0.5, 200 states, and an effective size of 14,269.

Evaluation of imputation performance in UK biobank Asian dataset

To evaluate the imputation performance of the SEAD panel in South Asian (SAS) populations, we used Rye software[32] to infer the genetic ancestry composition of each sample in UK Biobank, and selected 4758 samples with 50%–70% SAS ancestry composition, 2079 samples with 70%–90% SAS composition, and 2205 samples with >90% SAS composition (Fig. 1 Step 2). In brief, the UK Biobank data was directly downloaded upon request (Application 41376) by our team[69–75], and the 1kGP phase 3 data including 2504 samples from five ancestral groups, Africa (AFR), East Asia (EAS), Europe (EUR), South Asia (SAS) and the Americas (AMR)[11], served as the ancestry reference populations. Rye infers genetic ancestry composition based on principal component (PC) analysis of samples from ancestral reference populations, and compares them to the person being tested. To obtain PCs of each individual, we merged the UK Biobank data with the 1kGP phase 3 data using PLINK2 and extracted 565,631 slightly LD-pruned HapMap3 variants to calculate the first 20 principal components (PCs). The LD-pruning parameters used in PLINK2 were: window size = 1000 kb, step size = 100, $r^2 = 0.9$ and MAF ≥ 0.01[76,77].

For the selected SAS ancestry samples in the UK biobank, we extracted Asian Screening Array variants (659,184 variants) and converted the genotype coordinates to the GRCh38 reference to produce the pseudo ASA arrays. We applied parameters --geno 0.05, --hwe 1e-6, --maf 0.01, --make-bed, and --mind 0.05 to do the quality control in PLINK2. Finally, the samples with 50–70% South Asian (SAS) ancestry retained 610,279 variants, those with 70–90% SAS ancestry retained 583,566 variants, and those with > 90% SAS ancestry retained 576,772 variants. We employed SHAPEIT v2[78] for phasing and conducted imputation locally using the SEAD and 1kGP panels respectively. For the imputation with TOPMed (https://imputation.biodatacatalyst.nhlbi.nih.gov/) and ChinaMAP (http://www.mbiobank.com/imputation/help/) panels, we submitted the phased genotypes to the online websites for processing. All these analyses were performed on chromosome 2.

Evaluation of genotype concordance in HGDP Asian dataset

The 929 genomes from 55 diverse human populations in the Human Genome Diversity Project (HGDP)[79] dataset were utilized to estimate the genotype concordance of imputation across Asian populations (Fig. 1 Step 2). The HGDP dataset includes 104 samples from African populations, 61 samples from American populations, 197 samples from Central and South Asia populations, 223 samples from East Asia

populations, 155 samples from European populations, 161 samples from Middle East populations, and 28 samples from Oceania populations (Supplementary Data 15). We obtained the HGDP data from the following source: ftp://ngs.sanger.ac.uk/production/hgdp. The phased autosome VCF file was split into 22 chromosomes, which were considered as the true set. To compare the imputation concordance of the reference panels (TOPMed, ChinaMAP, 1kGP, and SEAD), We extracted Asian Screening Array sites (659,184 variants) in Central and South Asia populations (197 samples) and East Asia populations (223 samples) as the Asian array for comparison purposes. Before imputation, variants in the pseudo arrays with MAF <1% and samples with calling rate below 95% were excluded. We calculated the non-reference heterozygote concordance rate, non-reference homozygote concordance rate, precision, and non-reference sensitivity between true sequencing data and imputed genotypes pseudo arrays for each individual (Supplementary Fig. 11), as we did previously[25]. All these analyses were performed on chromosome 2.

Evaluation in Chinese population

The procedure to assess the imputation performance in the Chinese population using the reference panels (TOPMed, ChinaMAP, 1kGP and SEAD) was similar to the evaluation in HGDP dataset (Fig. 1 Step 2). As we had thousands of WBBC-chip samples, we were able to assess imputation performance for low-frequency and rare variants. Genotype imputation was performed using Minimac4[80,81] with a chunk length of 20 Mb and a chunk overlap of 4 Mb. We used R-square as the estimated value, defined as the squared correlation between imputed genotypes and observed genotypes, produced by Minimac4. Sites with an imputed r-square (Rsq) value >0.8 were considered well-imputed variants. We grouped the variants into seven Minor Allele Frequency (MAF) bins: <0.1%, 0.1–0.3%, 0.3–0.5%, 0.5–0.7%, 0.7–1%, 1–2% and 2–5%. We counted the number of well-imputed variants and calculated the average Rsq within each MAF bin. This analysis was performed on chromosome 2.

In the WBBC-chip data, 179 samples underwent both whole-genome sequencing and array genotyping. We evaluated imputation quality by comparing the concordance rate in the 179 samples, with the genotypes obtained from WGS serving as the true set. Employing the methods from the previous section, we calculated the non-reference heterozygote concordance rate, and non-reference homozygote concordance rate, precision and non-reference sensitivity for each individual.

Meta-imputation

We initially generated four independent non-singleton reference panels for the WBBC, 1kGP, SG10K, and GAsP datasets using Minimac3 (Fig. 1 Step 2). Subsequently, we imputed the 5679 WBBC-chip samples with these four reference panels, utilizing Minimac4 with the parameter set to -meta to produce the empiricalDose.vcf.gz file format for subsequent meta-analysis. Finally, we merged the imputation results from the four panels using MetaMinimac2 (https://github.com/yukt/MetaMinimac2)[43].

Genome wide association study of BMD traits

As mentioned above, the WBBC pilot study genotyped 6080 individuals with the ASA array (WBBC-chip), and a selected group of bone-related phenotypes was collected within WBBC[27]. After quality control, a total of 470,242 variants and 5679 samples were retained. Here, we took two correlated bone mineral density (BMD) traits (Hip and FN BMD) as examples, removed individuals with missing phenotype data, and excluded outliers using the mean ± 4 standard deviations (n = 5,369 remaining) (Fig. 1 Step 3). We then grouped the ASA samples according to the assessment place of collection: 2332 samples recruited from Jiangxi province (samples were mainly from Southern China) were taken as discovery cohort (WBBC-chip1), and 3037 samples recruited from Shandong province (samples were mainly from Northern China) were taken as replication cohort (WBBC-chip2), and vice versa. The baseline statistics of the study samples were shown in Supplementary Data 16. Comprehensive GWAS analyses were conducted with the imputed genotypes from the augmented SEAD reference panel. We kept the imputed variants with Rsq > 0.5 in the GWAS analysis. The imputation boosted the analyzed genetic variants from 470,242 to 19,235,129, with 3,195,758 variants with MAF between 0.1% and 1%. The BMD phenotypes (Hip and FN BMD) in this study were analyzed as continuous outcomes, adjusting for 'sex', 'age', 'BMI', 'geographical region', and '10 principal components (PCs)'. The parameters applied in PLINK were '--geno' of 0.05, '--mind' of 0.05, '--hwe' of 1×10^{-6} and '--maf' of 0.001. Furthermore, we performed the meta-analysis of GWAS summary statistics in cohort 1 and cohort 2 using METAL software[82]. Rare variants (0.001 < MAF < 0.01) were considered as candidates if they met the following criteria: a P-value $<1 \times 10^{-4}$ in either cohort, a P-value <0.05 in the other cohort, and a P-value less than 1×10^{-4} in the meta-analysis. Concurrently, we performed GWAS analysis on Hip BMD using the WBBC-chip imputed with ChinaMAP, SG10K, WBBC-seq, TOPMed, 1kGP and GAsP panel. All processing steps were consistent with those used for the data imputed with SEAD panel.

Variant-set analysis for association using annotation information

We performed the STAAR (variant-set test for association using annotation information)[83] framework to identify rare variants that would be associated with BMD traits (Hip and FN BMD) (Fig. 1 Step 3). The STAAR pipeline facilitates rare variant variant-set analyses, including sliding window-based analysis and gene-centric analysis[83]. In each test, we included variants with a minor allele frequency (MAF) of 0.1% to 1%. We employed fixed-size sliding window analysis, allowing for the systematic examination of distinct genomic regions by moving a 2 kb window every 1 kb across the genome, a total of 1,392,296 windows were generated. The gene-centric analysis method is the variant-set test for association using annotation information for a gene (STAAR-O), enhancing the power of rare variant association test by incorporating multiple variant functional annotations. For the gene-centric non-coding variants, aggregation was performed based on 7 categories: downstream, enhancer variants overlaid with Cap Analysis of Gene Expression (CAGE) sites, promoter CAGE, enhancer variants overlaid with DNase hypersensitivity (DHS), promoter DHS, upstream, and UTR2, and clustering regions with rare variations fewer than 2 were excluded. Both the sliding window and gene-centric methods were performed separately on WBBC-chip1 and WBBC-chip2.

Cell culture

HEK293T cells were cultured at 37 °C in a humidified atmosphere with 5% CO2 using DMEM basal medium supplemented with 10% fetal bovine serum. The MC3T3-E1 subclone 14 cells were cultured under similar condition with ascorbate-free αMEM basal medium supplemented with 10% fetal bovine serum. The medium was refreshed every 3–4 days, and cells were passaged when they reached 90% confluence. To initiate differentiation, the media were replaced with a calcification-inducing medium containing 50 ng/L vitamin C (VC), 10 nM dexamethasone, and 10 mM beta-glycerophosphate.

Luciferase reporter assays

Dual-Luciferase reporter assays were conducted following previously described methods[84]. In summary, the wild-type luciferase plasmid was constructed by inserting the sequence of GRCh38 chr8: 49957722-49958204 into the pGL6-TA firefly luciferase reporter vector (Beyotime). The risk luciferase plasmid was designed based on the *SNTG1* intronic SNP rs111829635 (GRCh38, chr8:49958116 C-T). The 293 T or MC3T3-E1 cells were seeded in a 24-well plate for 16 h and

subsequently transfected with luciferase reporter plasmids along with the control Renilla plasmid (Beyotime), using lipofectamine 3000 (Thermo Fisher Scientific). After 48 h of transfection, cells were collected and lysed using the lysis buffer provided in the Dual-Luciferase Reporter Assay System kit (Beyotime). Luciferase activity was analyzed in accordance with the guidelines outlined in the technique manual (Beyotime).

SNTG1 overexpression and qRT-PCR analysis

The cDNA of mouse *SNTG1* was amplified via polymerase chain reaction (PCR) and subsequently cloned into the pEF-GFP vector (a gift from Connie Cepko, Addgene plasmid # 11154[85]), by replacing GFP to create vector pEF-SNTG1. To initiate the overexpression of *SNTG1*, MC3T3-E1 cells were plated in 6-well and 24-well plates at a density of 6000 cells per square centimeter. Following an incubation period of 16–20 h, pEF-SNTG1 was transfected using Lipofectamine 3000 transfection reagent (Thermo Fisher Scientific). As a negative control, pEF-GFP was utilized. Additionally, a set of duplicate wells without any transfection was included and labeled as "WT". After a further 4–6 h, the cell culture medium was replaced with differentiation-induction medium.

For quantitative real-time polymerase chain reaction (qRT-PCR) analysis, cells were harvested 72 h after transfection. Total RNA was extracted from the target cells using TRIZOL (Invitrogen) according to the manufacturer's protocol. The isolated RNA was then reverse transcribed into complementary DNA (cDNA) using a reverse transcription kit (TransGen Biotech). For qRT-PCR analysis, a 2xSYBR Green Mix (TransGen Biotech) was utilized. The expression level of the target gene was normalized to that of GAPDH, which served as an endogenous control, enabling the comparison of samples. Cells intended for alkaline phosphatase (ALP) activity measurements were harvested 6 days after transfection.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The sequencing and vcf data of WGS from WBBC have been deposited in the Genome Sequence Archive (GSA)[86] in National Genomics Data Center[87], China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences, under accession number HRA001385 (https://ngdc.cncb.ac.cn/gsa-human/)[25]. The Fastq data are available under restricted access for privacy protection and access can be obtained by application on the website. The user can register and login to the GSA database website (https://ngdc.cncb.ac.cn/gsa-human/) and follow the guidance of "Request Data" to request the data step by step. The access authority can be obtained for Research Use Only. The SG10K (dataset ID:EGAD00001005337 on https://ega-archive.org/)[21] and GAsP (dataset ID: EGAS00001002921 on https://ega-archive.org/)[22] datasets were obtained by applying to the consortium. The high-coverage 1kGP dataset with an average depth of 30×, based on the GRCh38 assembly, were downloaded from the https://www.internationalgenome.org/data-portal/data-collection/30x-grch38. The CRAM files for Asian ancestry samples (EAS and SAS) from 1kGP were downloaded from here (ftp://ftp.sra.ebi.ac.uk/vol1/run/ERR323/). The SEAD reference panel integrates data from above sources, each governed by specific access restrictions. Researchers interested in accessing the underlying data must submit requests to the consortium in accordance with their respective policies. Detailed access conditions, including contact information and application processes, can be found in the availability statements of the original studies. For imputation purposes, the SEAD panel is available via the imputation server at https://imputationserver.westlake.edu.cn/. This integration allows researchers to utilize the panel without requiring direct access to the raw data. Users can register and create imputation jobs freely by uploading their bgzipped array data (VCF-formatted) to the server under a strict policy of data security. The UK Biobank data was directly downloaded upon request (Application 41376) (https://www.ukbiobank.ac.uk/). The GWAS summary statistics for the Hip and FN BMD of the WBBC-chip data are fully available at https://wbbc.westlake.edu.cn/downloads.html. Source data are provided with this paper.

References

1. Mahajan, A. et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet* **50**, 1505–1513 (2018).
2. Zheng, H. F. et al. Meta-analysis of genome-wide studies identifies MEF2C SNPs associated with bone mineral density at forearm. *J. Med. Genet.* **50**, 473–478 (2013).
3. Zhu, X. W. et al. Comprehensive assessment of the association between FCGRs polymorphisms and the risk of systemic lupus erythematosus: evidence from a meta-analysis. *Sci. Rep.* **6**, 31617 (2016).
4. Hoffmann, T. J. et al. Imputation of the rare HOXB13 G84E mutation and cancer risk in a large population-based cohort. *PLoS Genet* **11**, e1004930 (2015).
5. Handsaker, R. E. et al. Large multiallelic copy number variations in humans. *Nat. Genet* **47**, 296–303 (2015).
6. Das, S., Abecasis, G. R. & Browning, B. L. Genotype imputation from large reference panels. *Annu. Rev. Genomic Hum. Genet.* **19**, 73–96 (2018).
7. Nelson, S. C. et al. Improved imputation accuracy in hispanic/latino populations with larger and more diverse reference panels: applications in the hispanic community health study/study of latinos (HCHS/SOL). *Hum. Mol. Genet.* **25**, 3245–3254 (2016).
8. Lert-Itthiporn, W. et al. Validation of genotype imputation in Southeast Asian populations and the effect of single nucleotide polymorphism annotation on imputation outcome. *BMC Med. Genet.* **19**, 23 (2018).
9. Vergara, C. et al. Genotype imputation performance of three reference panels using African ancestry individuals. *Hum. Genet* **137**, 281–292 (2018).
10. Bai, W. Y. et al. Genotype imputation and reference panel: a systematic evaluation on haplotype size and diversity. *Brief. Bioinform.* **6**, bbz108 (2019).
11. Genomes Project C., Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
12. Genome of the Netherlands C. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet* **46**, 818–825 (2014).
13. Zheng, H. F. et al. Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. *Nature* **526**, 112–117 (2015).
14. Consortium U. K. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
15. Jun, G. et al. Structural variation across 138,134 samples in the TOPMed consortium. *bioRxiv* **25**, 2023.01.25.525428 (2023).
16. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet* **48**, 1279–1283 (2016).
17. Jones, K. M. & Cook-Deegan, R. Complicated legacies: the human genome at 20. *Science* **371**, 564–569 (2021).
18. Cahoon, J. L. et al. Imputation accuracy across global human populations. *Am. J. Hum. Genet* **111**, 979–989 (2024).
19. Nagasaki, M. et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. *Nat. Commun.* **6**, 8018 (2015).
20. Jeon, S. et al. Korean genome project: 1094 Korean personal genomes with clinical information. *Sci. Adv.* **6**, eaaz7835 (2020).

21. Wu, D. et al. Large-scale whole-genome sequencing of three diverse Asian populations in Singapore. *Cell* **179**, 736–49 e15 (2019).
22. GenomeAsia KC. The GenomeAsia 100 K project enables genetic discoveries across Asia. *Nature* **576**, 106–111 (2019).
23. Zhang, P. et al. NyuWa genome resource: a deep whole-genome sequencing-based variation profile and reference panel for the Chinese population. *Cell Rep.* **37**, 110017 (2021).
24. Li, L. et al. The ChinaMAP reference panel for the accurate genotype imputation in Chinese populations. *Cell Res.* **31**, 1308–1310 (2021).
25. Cong, P. K. et al. Genomic analyses of 10,376 individuals in the westlake BioBank for Chinese (WBBC) pilot project. *Nat. Commun.* **13**, 2939 (2022).
26. Wang, C. et al. Analyses of rare predisposing variants of lung cancer in 6,004 whole genomes in Chinese. *Cancer Cell* **40**, 1223–39 e6 (2022).
27. Zhu, X. W. et al. Cohort profile: the westlake BioBank for Chinese (WBBC) pilot project. *BMJ Open* **11**, e045564 (2021).
28. Cong, P. K. et al. Identification of clinically actionable secondary genetic variants from whole-genome sequencing in a large-scale Chinese population. *Clin. Transl. Med.* **12**, e866 (2022).
29. Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet* **13**, 135–145 (2012).
30. Alioto, T. S. et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.* **6**, 10001 (2015).
31. Huang, J. et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**, 8111 (2015).
32. Conley, A. B. et al. Rye: genetic ancestry inference at biobank scale. *Nucleic Acids Res.* **51**, e44 (2023).
33. Morris, J. A. et al. An atlas of genetic influences on osteoporosis in humans and mice. *Nat. Genet* **51**, 258–266 (2019).
34. Medina-Gomez, C. et al. Life-course genome-wide association study meta-analysis of total body BMD and assessment of age-specific effects. *Am. J. Hum. Genet* **102**, 88–102 (2018).
35. Zheng, H. F., Ladouceur, M., Greenwood, C. M. & Richards, J. B. Effect of genome-wide genotyping and reference panels on rare variants imputation. *J. Genet Genomics* **39**, 545–550 (2012).
36. Zheng, H. F. et al. Performance of genotype imputation for low frequency and rare variants from the 1000 genomes. *PLoS One* **10**, e0116487 (2015).
37. Chou, W. C. et al. A combined reference panel from the 1000 genomes and UK10K projects improved rare variant imputation in European and Chinese samples. *Sci. Rep.* **6**, 39313 (2016).
38. Yoo, S. K. et al. NARD: whole-genome reference panel of 1779 Northeast Asians improves imputation accuracy of rare and low-frequency variants. *Genome Med.* **11**, 64 (2019).
39. Cao, Y. et al. The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell Res.* **30**, 717–731 (2020).
40. Choi, J. et al. A whole-genome reference panel of 14,393 individuals for East Asian populations accelerates discovery of rare functional variants. *Sci. Adv.* **9**, eadg6319 (2023).
41. Lu, D. & Xu, S. Principal component analysis reveals the 1000 genomes project does not sufficiently cover the human genetic diversity in Asia. *Front Genet* **4**, 127 (2013).
42. Li, J. Z. et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).
43. Yu, K. et al. Meta-imputation: an efficient method to combine genotype data after imputation with multiple reference panels. *Am. J. Hum. Genet* **109**, 1007–1015 (2022).
44. Zhu, X., Bai, W. & Zheng, H. Twelve years of GWAS discoveries for osteoporosis and related traits: advances, challenges and applications. *Bone Res.* **9**, 23 (2021).
45. Salari, N. et al. The global prevalence of osteoporosis in the world: a comprehensive systematic review and meta-analysis. *J. Orthop. Surg. Res.* **16**, 609 (2021).
46. Hogan, A. et al. Interaction of gamma 1-syntrophin with diacylglycerol kinase-zeta. Regulation of nuclear localization by PDZ interactions. *J. Biol. Chem.* **276**, 26526–26533 (2001).
47. Shi, H. & Chen, M. The brain-bone axis: unraveling the complex interplay between the central nervous system and skeletal metabolism. *Eur. J. Med. Res.* **29**, 317 (2024).
48. Schwartzman, R. J. New treatments for reflex sympathetic dystrophy. *N. Engl. J. Med.* **343**, 654–656 (2000).
49. Rosen, C. J. & Bouxsein, M. L. Mechanisms of disease: is osteoporosis the obesity of bone? *Nat. Clin. Pr. Rheumatol.* **2**, 35–43 (2006).
50. Bashiardes, S. et al. SNTG1, the gene encoding gamma1-syntrophin: a candidate gene for idiopathic scoliosis. *Hum. Genet* **115**, 81–89 (2004).
51. Tassano, E. et al. Scoliosis with cognitive impairment in a girl with 8q11.21q11.23 microdeletion and SNTG1 disruption. *Bone* **150**, 116022 (2021).
52. Surface, L. E. et al. ATRAID regulates the action of nitrogen-containing bisphosphonates on bone. *Sci. Transl. Med.* **12**, eaav9166 (2020).
53. Lai, W. et al. Identification of copy number variants in a Southern Chinese cohort of patients with congenital scoliosis. *Genes* **12**, 1213 (2021).
54. O'Kelly, C. et al. The production of scoliosis after pinealectomy in young chickens, rats, and hamsters. *Spine* **24**, 35–43 (1999).
55. Omairi, S. et al. Regulation of the dystrophin-associated glycoprotein complex composition by the metabolic properties of muscle fibres. *Sci. Rep.* **9**, 2770 (2019).
56. Cheng, L. & Wang, S. Correlation between bone mineral density and sarcopenia in US adults: a population-based study. *J. Orthop. Surg. Res.* **18**, 588 (2023).
57. Chou, Y. Y. et al. The associations of osteoporosis and possible sarcopenia with disability, nutrition, and cognition in community-dwelling older adults. *BMC Geriatr.* **23**, 730 (2023).
58. Thapa, S., Nandy, A. & Rendina-Ruedy, E. Endocrinal metabolic regulation on the skeletal system in post-menopausal women. *Front Physiol.* **13**, 1052429 (2022).
59. Terkawi, M. A. et al. Interplay between inflammation and pathological bone resorption: insights into recent mechanisms and pathways in related diseases for future perspectives. *Int J. Mol. Sci.* **23**, 1786 (2022).
60. Byrska-Bishop, M. et al. High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *Cell* **185**, 3426–40 e19 (2022).
61. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
62. Yun, T. et al. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* **36**, 5582–5589 (2021).
63. Browning, B. L., Zhou, Y. & Browning, S. R. A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet* **103**, 338–348 (2018).
64. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
65. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
66. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
67. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet* **88**, 76–82 (2011).
68. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
69. Zhao, P. et al. Shared genetic architecture highlights the bidirectional association between major depressive disorder and fracture risk. *Gen. Psychiatr.* **37**, e101418 (2024).

70. Zhao, P. et al. Deciphering the complex relationship between type 2 diabetes and fracture risk with both genetic and observational evidence. *Elife* **12**, RP89281 (2024).
71. Ge, Q. et al. Ambient PM2.5 exposure and bone homeostasis: analysis of UK biobank data and experimental studies in mice and in vitro. *Environ. Health Perspect.* **131**, 107002 (2023).
72. Zheng, C. et al. Targeting sulfation-dependent mechanoreciprocity between matrix and osteoblasts to mitigate bone loss. *Sci. Transl. Med.* **15**, eadg3983 (2023).
73. Xia, J. W. et al. Both indirect maternal and direct fetal genetic effects reflect the observational relationship between higher birth weight and lower adult bone mass. *BMC Med.* **20**, 361 (2022).
74. Zhu, X. W. et al. General and abdominal obesity operate differently as influencing factors of fracture risk in old adults. *iScience* **25**, 104466 (2022).
75. Xia, J. et al. Systemic evaluation of the relationship between psoriasis, psoriatic arthritis and osteoporosis: observational and mendelian randomisation study. *Ann. Rheum. Dis.* **79**, 1460–1467 (2020).
76. Jiang, L., Zheng, Z., Fang, H. & Yang, J. A generalized linear mixed model association tool for biobank-scale data. *Nat. Genet* **53**, 1616–1621 (2021).
77. Altshuler, D. M. et al. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
78. Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
79. Bergstrom, A. et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, eaay5012 (2020).
80. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet* **48**, 1284–1287 (2016).
81. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation. *Bioinformatics* **31**, 782–784 (2015).
82. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
83. Li, Z. et al. A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nat. Methods* **19**, 1599–1611 (2022).
84. Bai, W. Y. et al. Identification of PIEZO1 polymorphisms for human bone mineral density. *Bone* **133**, 115247 (2020).
85. Matsuda, T. & Cepko, C. L. Electroporation and RNA interference in the rodent retina in vivo and in vitro. *Proc. Natl. Acad. Sci. USA* **101**, 16–22 (2004).
86. Chen, T. et al. The genome sequence archive family: toward explosive data growth and diverse data types. *Genomics Proteom. Bioinforma.* **19**, 578–583 (2021).
87. Partners C-NMa. Database resources of the national genomics data center, China national center for bioinformation in 2021. *Nucleic Acids Res.* **49**, D18–d28 (2021).

Acknowledgements

We thank the "SG10K_Pilot Investigators" for providing the SG10K_Pilot data (EGAD00001005337). The data from the "SG10K_Pilot Study" reported here were obtained from EGA. This manuscript was not prepared in collaboration with the "SG10K_Pilot Study" and does not necessarily reflect the opinions or views of the "SG10K_Pilot Study". We also thank the "Genome Asia 100 K consortium" for providing the "GenomeAsia pilot project" (EGAS00001002921). We thankfully acknowledge the High-performance Computing Center at Westlake University. This work was supported by the National Natural Science Foundation of China (#82370887), the Chinese National Key Technology R&D Program, Ministry of Science and Technology (#2021YFC2501702). And the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (#2023C03164).

Author contributions

H.-F.Z. conceptualized and designed the study. M.-Y.Y., J.-D.Z., G.T., and W.-Y.B. conducted the data analysis. X.L. conducted in vitro experiments. S.-H.Y., W.-W.Z., J.-Q.L. and Y. S. conducted the whole sequencing experiments. C.-D.Y., M.-C.Q., Y.-H.F., C.-F.Y., P.-K.C., K.S., S.-R.G., P.-P.Z., P.-L.G., Y.Q., J.-G.T., X.-J.Y., J.-X.G., X.C., M.-M.M., L.-X.L., G.T., S.-Y.X., L.X., F.H., J.-C.L., J.-F.G., B.-S.T., L.Y. and D.K. contributed to the sample collection, processing and preliminary data analysis. J.-J.Y., Y.-H.L. and N.L. designed the online website resource. M.-Y.Y. and J.-D.Z. drafted the manuscript, H.-F.Z. reviewed and edited manuscript. All authors contributed, discussed and approved manuscript.

Competing interests

S.-H.Y., W.-W.Z. and J.-Q.L. Y.S. are employee of KingMed Diagnostics Co., Ltd. The other authors have no competing interests to declare.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-55147-4.

Correspondence and requests for materials should be addressed to Hou-Feng Zheng.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Meng-Yuan Yang[1,2,3,16], Jia-Dong Zhong[2,3,16], Xin Li[1,2,3,16], Geng Tian[4,16], Wei-Yang Bai[3,16], Yi-Hu Fang[5], Mo-Chang Qiu[5], Cheng-Da Yuan[6], Chun-Fu Yu[7], Nan Li[8], Ji-Jian Yang[8], Yu-Heng Liu[8], Shi-Hui Yu[9], Wei-Wei Zhao[9], Jun-Quan Liu[9], Yi Sun[9], Pei-Kuan Cong[3], Saber Khederzadeh[3], Pian-Pian Zhao[3], Yu Qian[2,3], Peng-Lin Guan[1,2,3], Jia-Xuan Gu[1,2,3], Si-Rui Gai[1,2,3], Xiang-Jiao Yi[3], Jian-Guo Tao[1,2,3], Xiang Chen[2,3], Mao-Mao Miao[3], Lan-Xin Lei[10], Lin Xu[4], Shu-Yang Xie[4], Jin-Chen Li[11,12,13], Ji-Feng Guo[11,12,13], David Karasik[14], Liu Yang[15], Bei-Sha Tang[11,12,13], Fei Huang[4] & Hou-Feng Zheng[2,3] ✉

[1]School of Life Sciences, Zhejiang University, Hangzhou, Zhejiang, China. [2]Center for Health and Data Science (CHDS), the Second Affiliated Hospital of Soochow University, Suzhou, Jiangsu, China. [3]Diseases & Population (DaP) Geninfo Lab, School of Life Sciences, Westlake University, Hangzhou, Zhejiang, China. [4]WBBC Shandong Center, Binzhou Medical University, Yantai, Shandong, China. [5]WBBC Jiangxi Center, Jiangxi Medical College, Shangrao, Jiangxi, China. [6]Department of Dermatology, Hangzhou Hospital of Traditional Chinese Medicine, Hangzhou, Zhejiang, China. [7]Department of Orthopedic Surgery, Shangrao Municipal Hospital, Shangrao, Jiangxi, China. [8]The High-Performance Computing Center, Westlake University, Hangzhou, Zhejiang, China. [9]Clinical Genome Center, KingMed Diagnostics, Co., Ltd, Guangzhou, Guangdong, China. [10]Medical Biosciences, Imperial College London, London, United Kingdom. [11]National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University, Changsha, Hunan, China. [12]Department of Neurology, Xiangya Hospital, Central South University, Changsha, Hunan, China. [13]Center for Medical Genetics & Hunan Key Laboratory, School of Life Sciences, Central South University, Changsha, Hunan, China. [14]Azrieli Faculty of Medicine, Bar-Ilan University, Safed, Israel. [15]Institute of Orthopedic Surgery, Xijing Hospital, Fourth Military Medical University, Xi'an 710032, China. [16]These authors contributed equally: Meng-Yuan Yang, Jia-Dong Zhong, Xin Li, Geng Tian, and Wei-Yang Bai. ✉e-mail: houf.zheng@suda.edu.cn