medRxiv preprint doi: https://doi.org/10.1101/2023.12.23.23300480; this version posted December 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

SEAD: an augmented reference panel with 22,134 haplotypes boosts the rare variants imputation and GWAS analysis in Asian population

Meng-Yuan Yang[1,2,3#], Jia-Dong Zhong[1,2,3#], Xin Li[1,2,3#], Wei-Yang Bai[2,3#], Cheng-Da Yuan[4], Mo-Chang Qiu[5], Ke-Qi Liu[5], Chun-Fu Yu[6], Nan Li[7], Ji-Jian Yang[7], Yu-Heng Liu[7], Shi-Hui Yu[8], Wei-Wei Zhao[8], Jun-Quan Liu[8], Yi Sun[8], Pei-Kuan Cong[2,3], Saber Khederzadeh[2,3], Pian-Pian Zhao[1,2,3], Yu Qian[1,2,3], Peng-Lin Guan[1,2,3], Jia-Xuan Gu[1,2,3], Si-Rui Gai[1,2,3], Xiang-Jiao Yi[1,2,3], Jian-Guo Tao[1,2,3], Xiang Chen[1,2,3], Mao-Mao Liao[1,2,3], Lan-Xin Lei[9], Lin Xu[10], Shu-Yang Xie[10], Geng Tian[10], 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[10], Hou-Feng Zheng[1,2,3*]

[1]The affiliated Hangzhou first people's hospital, School of Medicine, Westlake University, Hangzhou, Zhejiang, China

[2]Diseases & Population (DaP) Geninfo Lab, School of Life Sciences, Westlake University, Hangzhou, Zhejiang, China

[3]Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang, China

[4]Department of Dermatology, Hangzhou Hospital of Traditional Chinese Medicine, Hangzhou, Zhejiang, China

[5]WBBC Jiangxi Center, Jiangxi Medical College, Shangrao, Jiangxi, China

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.12.23.23300480; this version posted December 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

[6]Department of Orthopedic Surgery, Shangrao Municipal Hospital, Shangrao, Jiangxi, China

[7]The High-Performance Computing Center, Westlake University, Hangzhou, Zhejiang, China

[8]Clinical Genome Center, KingMed Diagnostics, Co., Ltd. Guangzhou, Guangdong, China

[9]Imperial College London, London, The United Kingdom of Great Britain and Northern Ireland

[10]WBBC Shandong Center, Binzhou Medical University, Yantai, Shandong, China

[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

#These authors contributed equally: Meng-Yuan Yang, Jia-Dong Zhong, Xin Li, Wei-Yang Bai

medRxiv preprint doi: https://doi.org/10.1101/2023.12.23.23300480; this version posted December 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

*Correspondence: zhenghoufeng@westlake.edu.cn (H.-F.Z.)

medRxiv preprint doi: https://doi.org/10.1101/2023.12.23.23300480; this version posted December 27, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Abstract

Here, we present the South and East Asian Reference Database (SEAD) reference panel (https://imputationserver.westlake.edu.cn/), which comprises whole genome sequencing data from 11,067 individuals across 17 countries in Asia. The SEAD panel, which excludes singleton variants, consists of 22,134 haplotypes and 80,367,720 variants. Firstly, we assessed the concordance rate in global populations using HGDP datasets, notably, the SEAD panel showed advantage in East Asia, Central and South Asia, and Oceania populations. When imputing the disease-associated variants of Asian population, the SEAD panel displayed a distinct preponderance in imputing low-frequency and rare variants. In imputation of Chinese population, the SEAD panel imputed a larger number of well-imputed sites across all minor allele frequency (MAF) bins. Additionally, the SEAD panel exhibited higher imputation accuracy for shared sites in all MAF bins. Finally, we applied the augmented SEAD panel to conduct a discovery and replication genome-wide association study (GWAS) for hip and femoral neck (FN) bone mineral density (BMD) traits within the 5,369 Westlake BioBank for Chinese (WBBC) samples. The single-variant test suggests that rare variants near *SNTG1* gene are associated with hip BMD (rs60103302, MAF=0.0091, $P=4.79\times10^{-8}$). The spatial clustering analysis also suggests the association of this gene ($P_{slide_window}=1.08\times10^{-8}$, $P_{gene_centric}=4.72\times10^{-8}$). The gene and variants achieved a suggestive level for FN BMD. This gene was not reported previously, and the preliminary experiment demonstrated that the identified rare variant can upregulate the *SNTG1* expression, which in turn inhibits the proliferation and differentiation of preosteoblast.

Key words: reference panel, GWAS, rare variants, hip BMD, FN BMD

Introduction

Genotype imputation is an effective way to save the cost of sequencing and has become a key step in genome-wide association studies (GWAS). It can increase the chance 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 the 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 pervious study, we found that the imputation accuracy for European population benefit from an increasing in haplotype size and population diversity of the reference panel, while the accuracy for the Han Chinese population reached its peak with a fraction of an extra diverse sample (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 panel for genotype imputation, comprising 2,504 individuals from 26 populations (11). The latest update to 1kGP phase 3 sequenced 3,202 samples (including an additional 698 samples related to the previous participants) and identified 70,594,286 single nucleotide polymorphisms (SNPs) with 30× depth. As many rare and low-frequency variants associated with diseases tend to be population-specific (11), an increasing number of human genome projects are focusing on specific populations to provide population-specific reference panels. However, most of the whole-genome sequencing (WGS) efforts were carried out in European-descent individuals, such as the GoNL project (the Genome of the Netherlands project, n=769) (12), UK10K project (~4,000 WGS and ~6,000 whole exome sequencing samples in UK)(13, 14) and the TOPMed program (Trans-Omics for Precision Medicine, n=~138,000) (15). These efforts have enabled the provision of a large-scale combined reference panel 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 complicated legacies of human genome project after 20 years' development is that lacking of diversity might hinder the promise of genome science (17). As the largest continent, Asia accounts for 59.5 % of the worldwide population

102 (https://www.worldometers.info/world-population/). Fortunately, the Asian ancestry-dominant
103 populations have been recently sequenced and analyzed to understand the genetic basis of the
104 Asian, such as the Japanese (18), Korean (19), Singaporean (20), Indian (21) and Chinese (22-
105 25). Among these efforts, the Westlake BioBank for Chinese (WBBC) project (26) was initiated
106 in 2017 by our team. Up to now, we have whole-genome sequenced 4480 Chinese samples,
107 covering 29 out of 34 administrative divisions of China (24, 27).
108
109 Rare variants, which may have low level of pairwise linkage disequilibrium with common
110 variants, can potentially result in significant functional consequences (28). Next-generation
111 sequencing, with adequate coverage, enables the accurate detection of rare variants (29). Rare
112 variants are harder to impute than common variants because rare variants only appear a few
113 times in the reference panel. Further, the evaluation of rare variant imputation in Asian
114 populations is hindered by the limited sample size of currently published studies, which typically
115 comprise only a few thousand individuals.
116
117 In this study, we integrated the WGS data from multiple sources, including Singapore SG10K
118 pilot project (13.7×, 4563 samples), GenomeAsia pilot project (36×, 1031 samples), WBBC pilot
119 project (13.9×, 4480 samples), and the high-coverage 1kGP-Asian phase 3 (30×, 993 East and
120 South Asian from 3202 samples). These data were combined to create the South and East Asian
121 Reference Database (SEAD), encompassing a total of 11,067 individuals, making it a large-scale
122 reference panel for Asian populations. We compared the concordance rate of the genotypes
123 imputed from the SEAD panel across populations around the world. We then focused on
124 evaluating the imputation quality of this panel to impute rare variants in Asian populations.
125 Finally, we applied this augmented panel to the WBBC genome-wide genotyping data to explore
126 its role in imputing possible likely causal rare variants for bone-related traits.
127

128 # Methods

129 ## Quality control of the four Asian reference panels

130 We collected Asian haplotypes from the SG10K pilot project (20), the GenomeAsia pilot project
131 (GAsP) (21), the newly released high-coverage 1000 Genomes project (1kGP) (30) and Westlake

132 Biobank for Chinese pilot project (WBBC) (24). The SG10K dataset comprises 4,810 whole-
133 genome sequencings (WGS) of Singaporean Chinese, Malays, and Indians, with an average
134 depth of 13.7×. The GAsP dataset includes 1,163 WGS samples with an average depth of 36×,
135 mainly covering Indians, Korean, Pakistanis, etc. These two datasets were obtained by applying
136 to the consortium. Since the SG10K and GAsP datasets were based on Genome Reference
137 Consortium Human build 37 (GRCh37), we employed liftover (31) to update the genome
138 assembly version from GRCh37 to GRCh38. For the GAsP dataset, we transformed multi-allelic
139 variants into multiple bi-allelic variants using Bcftools (32) and performed haplotype re-phasing
140 using SHAPEIT v2 with parameters including windows of size 0.5, 200 states, and an effective
141 size of 14,269. The high-coverage 1kGP dataset with an average depth of 30×, based on the
142 GRCh38 assembly, were downloaded from the https://www.internationalgenome.org/data-
143 portal/data-collection/30x-grch38. Then we extracted East Asian (EAS) and South Asian (SAS)
144 dataset from 1kGP and removed sites with minor allele count (MAC) equals zero. The WBBC
145 dataset, which were generated by our own, included 4,480 Chinese samples with an average
146 sequencing depth of 13.9×, supporting both GRCh37 and GRCh38. We identified relative pairs
147 using KING (33) and excluded sample pairs with estimated kinship coefficients restricted to the
148 range of 0.177-0.354. Ultimately, we retained 4,480 samples from WBBC 4,563 samples from
149 SG10K, 1,031 samples from GAsP, and 993 samples from the EAS/SAS subset of the 1kGP
150 dataset (1kGP-Asian) (Supplementary Table 1).
151

152 Evaluation of genotype concordance in HGDP dataset.

153 The 929 genomes from 55 diverse human populations from Human Genome Diversity Project
154 (HGDP) (34) dataset was utilized to estimate the genotype concordance of imputation across
155 global populations. The HGDP dataset includes 104 samples from African populations, 61
156 samples from American populations, 197 samples from Central and South Asia populations, 223
157 samples from East Asia populations, 155 samples from European populations, 161 samples from
158 Middle East populations, and 28 samples from Oceania populations (Supplementary Table 2).
159 We obtained the HGDP data from the following source: ftp://ngs.sanger.ac.uk/production/hgdp.
160 The phased autosome VCF file was split into 22 chromosomes, which were considered as the
161 true set. To compare the imputation concordance of the reference panels (WBBC, SG10K, 1kGP,
162 GAsP and SEAD), we extracted the genetic variants from the 929 HGDP genome data with

163 Axiom array loci (820,967 variants) to create pseudo array, which was then used as the test set
164 for imputation. We downloaded Axiom array SNP list from
165 https://www.ukbiobank.ac.uk/enable-your-research/about-our-data/genetic-data. We imputed the
166 genotypes from the 5 panels, then compared the concordance between the imputed genotypes
167 and true set. We also extracted Asian Screening Array sites (659,184 variants) in Central and
168 South Asia populations and East Asia populations (420 samples in total) as Asian array for
169 comparison purpose. Before imputation, variants in the pseudo arrays with MAF less than 1%
170 and samples with calling rate below 95% were excluded. Here, the singletons were excluded in
171 the 5 reference panels, consists of 35,616,674 variants for WBBC, 50,246,865 variants for
172 SG10K, 70,594,286 variants for 1kGP (not just Asian), 20,203,158 variants for GAsP,
173 80,367,720 variants for SEAD. We calculated the non-reference genotype concordance rate, the
174 non-reference heterozygote concordance rate and non-reference homozygote concordance rate
175 between true genotypes and imputed genotypes pseudo arrays for each individual
176 (Supplementary Figure 1), as what we did before (24). The average concordance rate in single
177 population was calculated as the mean of all the individuals' concordance rate in this population.
178 We quantified the ratio of NR concordance rate of the SEAD panel to other four panels by using
179 the following formula, respectively.

180 $$\mathrm{logFC} = \log\left(\frac{\text{NR concordance rate1}}{\text{NR concordance rate2}}\right)$$

181 All these analyses were performed on chromosome 2.
182

183 Quality control of the Chinese genotyping array data

184 Besides the whole genome sequencing data, the WBBC study also genotyped 6,080 individuals
185 with the high-density Illumina Asian Screening Array (ASA, based on GRCh37), resulting in the
186 identification of a total of 659,184 SNPs (24, 26). Of which, 184 samples underwent both whole-
187 genome sequencing and array genotyping.
188

189 As part of quality control, we employed GCTA (35) to calculate the pairwise genetic relationship
190 matrix using common variants and remove samples with a coefficient > 0.025 (Supplementary
191 Table 3). We then excluded samples with missing call rates ≥ 2% and excluded SNPs with
192 missing call rates ≥ 5%, MAF<1% and Hardy-Weinberg equilibrium at $P < 1 \times 10^{-6}$ by using

193 PLINK1.9 (36). The genotype assembly version of the WBBC array data was updated from
194 GRCh37 to GRCh38. Finally, we retained a total of 470,242 variants and 5,679 samples in the
195 WBBC array dataset. The data were phased using SHAPEIT with a window size of 0.5, 200
196 states, and an effective size of 14,269.
197

198 Evaluation of reference panel for imputation in the Chinese 199 population

200 The procedure to assess the imputation performance in the Chinese population using the five
201 reference panels (WBBC, SG10K, 1kGP, GAsP, and SEAD) was similar to the evaluation in
202 HGDP dataset. As we have thousands of ASA array samples, we could assess the imputation
203 performance for low-frequency and rare variants. We grouped the variants into seven Minor
204 Allele Frequency (MAF) bins: < 0.1%, 0.1%-0.3%, 0.3%-0.5%, 0.5%-0.7%, 0.7%-1%, 1%-2%,
205 and 2%-5%. Genotype imputation was performed using Minimac4 (37, 38), with a chunk length
206 of 20 Mb and a chunk overlap of 4 Mb. We used R-square as the estimated value, defined as the
207 squared correlation between imputed genotypes and observed genotypes, produced by
208 Minimac4. Sites with an R-square value greater than 0.8 were considered well-imputed variants.
209 We counted the number of well-imputed variants and calculated the average R^2 within each
210 MAF bin. To compare the imputation accuracy among different reference panels, we extracted
211 8,525,136 variants that were shared by all five panels and evaluated the imputation performance
212 again.
213
214 In the WBBC ASA array data, 184 samples underwent both whole-genome sequencing and array
215 genotyping. We then utilized these samples to evaluate the imputation concordance rate. We
216 randomly masked 2,600 variants on chromosome 2 from the array data and subsequently
217 imputed these SNPs using the five Asian panels. We extracted the imputed 2,600 loci as the test
218 set, while the original genotype of the 2,600 loci from the array data served as the true set.
219 Employing the methods from the previous section, we calculated both the non-reference
220 homozygous and heterozygous concordance rates.
221

222 # Genome wide association study of BMD traits

223 As we mentioned above, the WBBC pilot study genotyped 6,080 individuals with ASA array,
224 and a bunch of bone-related phenotypes were collected within WBBC (26). After quality control,
225 a total of 470,242 variants and 5,679 samples were retained. Here, we took two correlated bone
226 mineral density (BMD) traits (total hip and femoral neck BMD) as example, we removed
227 individuals with missing phenotype data, and excluded outliers using the mean ± 4 standard
228 deviations (n=5,369 left). We then grouped the ASA samples according to the assessment place
229 of collection: 2,332 samples recruited from Jiangxi province (samples were mainly from
230 Southern China) were taken as discovery cohort (cohort 1), and 3,037 samples recruited from
231 Shandong province (samples were mainly from Northern China) were taken as replication cohort
232 (cohort 2), and ***vice versa***. The baseline statistics of the study samples were shown in
233 Supplementary Table 4. Comprehensive GWAS analyses were conducted with the imputed
234 genotypes from the augmented SEAD reference panel. We kept the imputed variants with
235 Rsq >0.5 in the GWAS analysis. The imputation boosted the analyzed genetic variants from
236 470,242 to 19,235,129, with 3,195,758 variants with MAF between 0.1% and 1%. The BMD
237 phenotypes (total hip and femoral neck BMD) in this study was analyzed as continuous
238 outcomes, adjusting for 'sex', 'age', 'BMI', 'geographical region', and '10 principal components
239 (PCs)'. The parameters applied in PLINK were '--geno' of 0.05, '--mind' of 0.05, '--hwe' of
240 1×10^{-6} and '--maf' of 0.001. Furthermore, we performed the meta-analysis of GWAS summary
241 statistics in cohort 1 and cohort 2 using METAL software (39). Rare variants
242 (0.001<MAF<0.01) were considered as candidates if they met the following criteria: a *P*-value
243 less than 1e-04 in either cohort, a *P*-value less than 0.05 in the other cohort, and a *P*-value less
244 than 1e-04 in the meta-analysis.
245

246 # Spatial clustering analysis

247 We also performed the STAAR (variant-set test for association using annotation information)
248 (40) framework to identify rare variants that would be associated with BMD traits (Hip and FN
249 BMD). The STAAR pipeline facilitates rare variants spatial clustering analyses, including sliding
250 window-based analysis and gene-centric analysis (40). In each test, we included variants with a
251 minor allele frequency (MAF) of 0.1% to 1%. We employed fixed-size sliding window analysis,

252 allowing for the systematic examination of distinct genomic regions by moving a 2kb window
253 every 1kb across the genome, a total of 1,392,296 windows was divided. The gene-centric
254 analysis method is the variant-set test for association using annotation information for a gene
255 (STAAR-O), enhancing the power of rare variant association test by incorporating multiple
256 variant functional annotations. For the gene-centric non-coding variants, aggregation was
257 performed based on 7 categories: downstream, enhancer variants overlaid with Cap Analysis of
258 Gene Expression (CAGE) sites, promoter CAGE, enhancer variants overlaid with DNase
259 hypersensitivity (DHS), promoter DHS, upstream, and UTR2, and clustering regions with rare
260 variations less than 2 were excluded. Both the sliding window and gene-centric methods were
261 performed separately on the cohort 1 and cohort 2.
262

263 Cell culture

264 The HEK293T cells were cultured at 37 °C in a humidified atmosphere with 5% CO2 using
265 DMEM basal media supplemented with 10% fetal bovine serum. The MC3T3-E1 subclone 14
266 cells were cultured with similar condition with ascorbate-free αMEM basal media supplemented
267 with 10% fetal bovine serum. The media were refreshed every 3-4 days, and cells were passaged
268 when they reached 90% confluence. To initiate differentiation, the media were replaced with a
269 calcification-inducing medium containing 50 ng/L vitamin C (VC), 10 nM dexamethasone, and
270 10 mM beta-glycerophosphate.
271

272 Luciferase Reporter assays

273 Dual-Luciferase reporter assays were conducted following previously described methods (41). In
274 summary, the wild type luciferase plasmid was constructed by inserting the sequence of hg38
275 chr8: 49957722-49958204 into pGL6-TA firefly luciferase reporter vector (Beyotime). The risk
276 luciferase plasmid was designed according to the *SNTG1* intronic SNP rs111829635 (hg38,
277 chr8:49958116 C-T). The 293T or MC3T3-E1 cells were seeded in a 24-well plate for 16 hours
278 and subsequently transfected with luciferase reporter plasmids along with the control Renilla
279 plasmid (Beyotime), using lipofectamine 3000 (Thermo Fisher Scientific). After 48 hours of
280 transfection, cells were collected and lysed using the lysis buffer provided in the Dual-Luciferase

281 Reporter Assay System kit (Beyotime). Luciferase activity was analyzed in accordance with the
282 guidelines outlined in the technique manual (Beyotime).
283

284 *SNTG1* overexpression and qRT-PCR analysis

285 The cDNA of mouse *SNTG1* was amplified via polymerase chain reaction (PCR) and
286 subsequently cloned into the pEF-GFP vector (a gift from Connie Cepko, Addgene plasmid #
287 11154 (42)), by replacing GFP to create vector pEF-SNTG1. To initiate the overexpression of
288 *SNTG1*, MC3T3-E1 cells were plated in 6-well and 24-well plates at a density of 6,000 cells per
289 square centimeter. Following an incubation period of 16-20 hours, transfection of pEF-SNTG1
290 was performed using Lipofectamine 3000 transfection reagent (Thermo Fisher Scientific). As a
291 negative control, pEF-GFP was utilized. Additionally, a set of duplicate wells without any
292 transfection was included and labeled as "WT". After a further 4-6 hours, the cell culture
293 medium was replaced with differentiation-induction medium.
294
295 For quantitative real-time polymerase chain reaction (qRT-PCR) analysis, cells were harvested
296 72 hours after transfection. Total RNA was extracted from the target cells using TRIZOL
297 (Invitrogen) according to the manufacturer's protocol. The isolated RNA was then reverse
298 transcribed into complementary DNA (cDNA) using a reverse transcription kit (TransGen
299 Biotech). For qRT-PCR analysis, a 2xSYBR Green Mix (TransGen Biotech) was utilized. The
300 expression level of the target gene was normalized to that of GAPDH, which served as an
301 endogenous control, enabling the comparison of samples. Cells intended for alkaline phosphatase
302 (ALP) activity measurements were harvested after 6 days following transfection.
303

304 Protein-protein interaction network analysis

305 To uncover potential pathways and functions of candidate genes, we explored their direct
306 relationships with previously reported genes for bone mineral density (BMD) using a protein
307 network. The known genes were derived from 1,103 independent genome-wide significant SNPs
308 annotated to 863 genes, as reported in the study by Morris et al. (43). We utilized the STRING
309 database (44) online resource to predict the relationships between these 863 known genes and the
310 10 suggestive candidate genes (including three lincRNAs), ultimately identifying 19 known

genes directly associated with the candidate genes. The interconnections between these genes were visualized using Cytoscape (45). We mainly considered two databases, the Gene Ontology (GO) database (for biological processes and cellular components) and the KEGG database, to perform pathway enrichment and functional annotation analyses on the selected 29 proteins. We manually selected pathways related to bone formation and development with a false discovery rate (FDR) of less than 0.05 for annotation.

Results

Integration of the SEAD reference panel

After quality control of each reference panel (see Methods), we retained 4,480 samples and 78,429,408 variants from WBBC, 4,563 samples and 95,597,234 variants from SG10K, 1,031 samples and 63,925,145 variants from GAsP, and 993 samples and 35,157,155 variants from 1kGP-Asian. To maximize the number of variants as we did before (46), we applied reciprocal imputation approach step by step with minimac4 (38): first, we merged the reference panel of SG10K and WBBC by imputing with each other, we then combined haplotypes from 1kGP-Asian to the WBBC-SG10K panel, next, we obtained the SEAD (South and East Asian Reference Database) panel by combining all the 4 panels together (Figure 1A). After removing the singletons, we finally got 80,367,720 variants and 22,134 haplotypes for SEAD panel. The SEAD reference panel is now available online (https://imputationserver.westlake.edu.cn/).

Imputation performance in global populations

The imputation concordance rate of NR-alleles were assessed across five reference panels using data from 929 global samples representing 55 populations. The samples were then categorized into regions, including Africa, America, Europe, Middle East, Oceania, Central and South Asia, and East Asia (Supplementary Table 2). Two types of pseudo arrays were generated: the Axiom array loci (820,967 variants) for all populations, and the Asian Screening array loci (659,184 variants) for East Asian and Central and South Asian population (Figure 2A).

Among the non-Asian populations, the NR allele concordance rates from the 1kGP panel (this panel contains all 1kGP populations, not just Asian) were higher than that of the other reference

panels in most of the populations, particularly in African populations (Supplementary Table 5 and Supplementary Figure 2). To visualize these differences among the panels, we calculated the fold change of the NR allele concordance rate between the panels. When compared to the 1kGP panel, the SEAD reference panel demonstrated a lower level of NR-allele concordance for most of the non-Asian populations, however, it is worth noting that the SEAD reference panel demonstrated a slight advantage in imputing Oceanian populations (Figure 2A). Although SEAD might outperform the other three Asian reference panels (Supplementary Figure 3), we did not suggest to use SEAD to impute the populations outside Asian.

As for East Asian populations, the distributions of the non-reference homozygous allele concordance rate and the non-reference heterozygous allele concordance rate imputed by the SEAD reference panel were concentrated on higher values. The higher peak of the SEAD reference panel in the density indicated that the NR-allele concordance rate of this panel was more concentrated (Supplementary Figure 4A). The SEAD reference panel consistently demonstrated higher NR-allele concordance when compared to the 1kGP panel (Figure 2A and Supplementary Figure 2). In comparison to three other Asian reference panels, WBBC and SG10K might be slightly better in some of the populations (Supplementary Figure 3). For the Central and South Asian populations, the NR-allele concordance rate corresponding to the peak of the SEAD reference panel also showed slightly higher than that of other reference panels (Supplementary Figure 4B). The SEAD panel outperformed GAsP, WBBC and SG10K in all populations (Supplementary Figure 3). Compared with the 1kGP panel, the SEAD panel also demonstrated its advantage in imputing most Central and South Asian populations (Figure 2A). These results indicated that SEAD panel was a suitable imputation reference panel for Asian populations.

Imputation of disease-associated variants for Asian population

We downloaded the complete dataset of genome-wide association data from the GWAS catalog, (47) which comprised 331,513 association sites. As these association loci were predominantly identified in European population data, we manually examined the population ancestry for each catalog ID, resulting in the identification of 9,810 association loci specific to Asian population (Supplementary Table 6). We assessed the imputation performance for the reference panels of

372 the 9,810 loci extracted from WBBC array imputation (Figure 1B). The SEAD panel exhibited
373 the highest number of well imputed variants, amounting to 4,424 (45.10%), followed by SG10K
374 (Figure 2B and Supplementary Table 7). Only 25.97% of the disease-associated variants were
375 imputed in GAsP panel (Figure 2B and Supplementary Table 7). The main advantage of SEAD
376 panel lied in the imputation of low-frequency and rare variants, encompassing a total of 356 low-
377 frequency variants and 133 rare variants (Supplementary Table 7). When categorizing the
378 variants into seven Minor Allele Frequency (MAF) bins: < 0.1%, 0.1%-0.3%, 0.3%-0.5%, 0.
379 5%-0.7%, 0.7%-1%, 1%-2%, and 2%-5%, the SEAD panel demonstrated high number of well
380 imputed sites compared to other panels (Figure 2C and Supplementary table 7). The WBBC and
381 SG10K also performed well in some of the MAF bins (Figure 2C and Supplementary table 7).
382 These results suggested SEAD panel performed well to impute the disease-associated variants in
383 Asian population.
384

385 # Improvement in detecting rare variants in Chinese population

386 To further assess the imputation performance of the reference panels in Chinese population
387 (Figure 1B), we utilized GAsP, 1kGP, WBBC, SG10K, and the SEAD reference panels to
388 impute WBBC array data. The SEAD reference panel imputed a total of 7,976,019 well-imputed
389 sites, surpassing all the other panels, with SG10K ranking second with 6,725,695 sites (Figure
390 3A and Supplementary Table 8). GAsP panel imputed fewest well-imputed variants, because
391 only a thousand samples were included in this panel (Figure 3A and Supplementary Table 8).
392

393 Given that common variants have been extensively studied, our primary focus was on the
394 imputation effect of reference panels for low-frequency (1% < MAF < 5%) and rare variants
395 (MAF < 1%). The SEAD reference panel had obtained 3,209,490 well-imputed low/rare sites,
396 and as many as twice rare mutations compared to the 1kGP panel (Figure 3A and Supplementary
397 Table 8). Across the 7 MAF bins for rare and low-frequency variants, as MAF decreased, the
398 SEAD reference panel exhibited a higher number of well-imputed sites compared to other panels
399 (Figure 3B and Supplementary Table 8). Especially in the intervals of MAF < 0.1% and 0.1% <
400 MAF < 0.3%, the imputation advantage of SEAD was most pronounced. (Figure 3B and
401 Supplementary Table 8). To make the results comparable across the five reference panels, we
402 also extracted a total of 8,525,136 shared SNPs for analysis. Within the 7 MAF bins, the SEAD

403 reference panel always exhibited the highest Rsq value and largest number of well-quality sites
404 (Figure 3C and Supplementary Table 9-10).
405
406 As 184 samples underwent both whole-genome sequencing and array genotyping, we further
407 assessed the imputation concordance of the five panels. The imputation accuracy of the SEAD,
408 WBBC, and SG10K reference panels was considerably superior to that of the GASP and 1KGP
409 panels. The concordance rate values corresponding to the peak of the SEAD imputation panel
410 were slightly higher than those of WBBC and SG10K, but the differences were minimal (Figure
411 3D).
412

413 Employment of the SEAD reference panel in bone mineral 414 density GWAS analysis

415 After imputation with SEAD panel for WBBC array data, we conducted GWAS analysis on two
416 BMD traits (total hip and femoral neck BMD) (Figure 1C). These two BMD traits were highly
417 correlated in either Pearson correlation (r = 0.918) or genetic correlation (r = 0.932, SD = 0.025).
418 With genome-wide significance threshold ($P < 5\times10^{-8}$), two common loci (MAF > 0.01) that
419 have been previously reported in GWAS were identified to be associated with both BMD traits,
420 with the top signals at chr1: rs9659023 (an intronic variant of *FMN2*, this locus was reported to
421 be associated with heel (43) and total body BMD (48)) and at chr6: rs138367190 (near *SOX4*,
422 this locus was reported to be associated with heel (43) and FN BMD (13)) (Supplementary
423 Figure 5).
424
425 For the relative rare variants (0.001 < MAF < 0.01), we utilized two analytical strategies: single-
426 variant test with PLINK and spatial clustering analysis via the STAARpipeline (incorporating
427 both sliding-window and gene-centric methods). In the single-variant test analysis, we prioritized
428 the variants with small *P*-value and with replication (see Methods), and identified 95 suggestive
429 variants for Hip BMD (Supplementary Table 11). Among these variants, 70 (73.7%) were
430 annotated to *SNTG1* gene by ANNOVAR (Supplementary Table 11). Of which, four rare
431 variants (rs60103302, rs61260287, rs60600379, and rs57319781), in complete linkage
432 disequilibrium (LD r^2 = 1.00), were identified for Hip BMD at genome-wide significance level

433 ($P = 4.79\times10^{-8}$, MAF = 0.0091), these variants located in the intergenic region near *SNTG1* on
434 chromosome 8 (Figure 4A and 4F, and Supplementary Table 11). As for FN BMD, 18 (46.2%)
435 out of the 39 suggestive variants were annotated to *SNTG1* gene (Figure 4C and 4F, and
436 Supplementary Table 11). In the spatial clustering analysis, we first implemented a sliding
437 window of 2kb in STAAR-O test and adhered to the previous "discovery and replication"
438 strategy. The results showed that 5 clustering regions were identified as the suggestive signals
439 for Hip BMD (Figure 4B and Supplementary Table 12). In the cluster on chromosome 8
440 (chr8:49873459-49958458), a total of 22 windows were suggested, of which, 3 windows
441 annotated near *SNTG1* were discovered at bonferroni-corrected level $P = 3.59\times10^{-8}$ (0.05 divided
442 by 1,392,296 windows), with top window at chr8:49898459-49900458 ($P = 1.08\times10^{-8}$) (Figure
443 4B and Supplementary Table 12). Moreover, in the annotation-based gene-centric analysis, of
444 the 7 non-coding categories of *SNTG1* gene, 5 categories showed suggestive signal for Hip
445 BMD, with the top signal at promoter region ($P_{promoter_CAGE} = 4.72\times10^{-8}$) (Figure 4E and
446 Supplementary Table 13), showcasing at least one annotation reaching the genome-wide
447 bonferroni-corrected level $P = 3.57\times10^{-7}$ (0.05 divided by 20,000 genes then divided by 7
448 categories). In summary, *SNTG1* locus was highlighted in both analytical strategies with the
449 strictest threshold applied to Hip BMD, and also achieved a suggestive level in FN BMD
450 ($P_{single_variant} = 4.85\times10^{-6}$, $P_{slide_window} = 2.36\times10^{-6}$, $P_{gene_centric} = 5.43\times10^{-6}$) (Figure 4C, 4D, 4E,
451 and Supplementary Table 12 and 13),
452
453 We further exploited the preliminary function of *SNTG1* gene. Dual-luciferase assays were then
454 performed in two cell lines, 293T and MC3T3-E1. It was observed that the intronic SNP
455 rs111829635 C-T mutation significantly increased luciferase activity in both 293T cells and
456 MC3T3-E1 cells (Figure 4G), suggesting that the identified locus would alter the *SNTG1* gene
457 expression. Moreover, the results from the CCK-8 proliferation assay showed that the
458 absorbance at OD 450 nm was decreased after *SNTG1* overexpression, and cell density was also
459 decreased (P <0.05, Fig 4H and Fig 4I), revealing that overexpression of *SNTG1* led to a
460 significant reduction in cell proliferation. At the 72 hours, a noticeable decrease in the expression
461 of osteogenic marker genes (*RUNX2*, *COL1A1* and OCN) confirmed that *SNTG1* inhibited the
462 differentiation of preosteoblast MC3T3-E1 cells (P <0.05, Fig 4J). Furthermore, there was also a
463 significant reduction in alkaline phosphatase (ALP) activity observed after six days. Overall,

464 these results supported the notion that *SNTG1* overexpression might inhibit human osteogenic
465 proliferation and differentiation.
466

467 Other suggestive candidate genes and loci

468 Besides the *SNTG1* locus, other suggestive loci were ranked in a credible priority order: 3 loci
469 (*GRM7*, *CPT2*, *LINC00407*) were identified by both single-variant test and spatial clustering
470 analyses (level A), and 6 loci (*SH3TC2*, *C4BPA*, *LINC02572*, *LINC00461*, *MORC2*, *ADIG*) were
471 identified in both BMD traits by either method (level B) (Figure 4A, 4B, 4C, 4D, and
472 Supplementary Table 14). Of these, 7 were annotated near genes, three candidates were
473 lincRNAs. We then employed the protein-protein interaction analysis to include the known
474 BMD-associated genes reported in previously GWAS (43) and suggested that 19 known genes
475 directly associated with the 7 candidate genes (Supplementary Table 15). Functional enrichment
476 of directly associated proteins revealed 24 potential categories with a false discovery rate (FDR)
477 <= 0.05 (Supplementary Table 16). Among these enrichment analysis, we prioritized four
478 functional annotations closely related to bone biology (Figure 5A), including "abnormal skeletal
479 muscle morphology" ('Monarch Human Phenotype', FDR = 0.0246), "Dystrophin-associated
480 glycoprotein complex" ('GO Cellular Component', FDR = 0.0047), "Fatty acid metabolic
481 process" ('GO Biological Process', FDR = 0.0493) and "Complement and coagulation cascades"
482 ('KEGG pathway', FDR = 0.0011).
483

484 We compared the frequency distribution of all the rare variants identified in level-A and -B (84
485 SNPs) with the gnomAD and TOPMed databases across various populations (Figure 5B,
486 Supplementary Table 17). The MAF of these rare variants most closely aligned with the MAF
487 value in the East Asian (EAS) population (2,604 genomes) within gnomAD, validating the
488 appropriateness of our panel for imputing genotype data in East Asian populations. Furthermore,
489 the *SNTG1* variants were predominantly rare in the Non-Finnish European (NFE: 34,029
490 genomes) and South Asian (SAS: 2,419 genomes) populations, while they were more common in
491 the African/African American (AFR: 20,744 genomes) and Latino/Admixed American (AMR:
492 7,647 genomes) populations, highlighting population-specific patterns. Additionally, the other
493 variants were present in either the SAS or EAS population but were largely absent in other

494 populations, further illustrating the panel's ability to detect variations specific to East Asian and
495 South Asian populations.
496

497 Discussion

498 In this study, we integrated whole-genome sequencing data from SG10K, GenomeAsia, WBBC,
499 and 1kGP-Asian to create a combined reference panel for genotype imputation, the South and
500 East Asian Reference Database (SEAD). It comprised a diverse range of populations across Asia,
501 including 11 populations from GenomeAsia, one population from WBBC, three populations
502 from SG10K, and 8 populations from 1kGP-Asian. With a sample size of 22,134 haplotypes, the
503 SEAD panel stands as one of the most comprehensive panels in terms of coverage across Asia.
504 We compared the concordance rate of the genotypes imputed from the SEAD panel across
505 populations around the world, and suggested that the SEAD panel performed the best for the rare
506 variants imputation in Asian populations. The SEAD panel also had advantage in detecting
507 disease-associated low/rare variants. Subsequently, we applied the SEAD panel to the bone
508 mineral density GWAS analyses in WBBC array cohorts, and identified a rare locus, *SNTG1*,
509 that was not reported even in the large biobank-scale GWAS.
510

511 Asia, being the largest and most populous continent worldwide, boasts a wealth of human
512 genetic resources. However, most of the whole genome-sequencing (WGS) efforts were carried
513 out in Caucasian populations in the last decade (17). In our previous study, we conducted a
514 thorough evaluation and discussion on the imputation of rare variants, highlighting the necessity
515 of constructing a haplotype reference panel for Asian populations (10, 49-51). In recent years,
516 there has been a notable surge WGS data across Asia, particularly in regions such as Japan,
517 Singapore, Korea and China. Despite the abundance of WGS projects, large-scale reference
518 panel with broad geographical coverage across Asia is still needed. The accumulation of such
519 vast datasets from diverse populations presents a unique opportunity: the combination of
520 multiple WGS datasets to form a more comprehensive and expansive reference panel. The
521 recently published Northeast Asian Reference Database, with over half of the samples deriving
522 from Japan and Korea, represented populations from Northeast Asia (52). In our study, we
523 integrated the data from WBBC with other Asian populations to construct the South and East

524 Asian Reference Database (SEAD) reference panel, covering the widest geographical area in
525 Asia, with the majority of its population originating from South and East Asia.
526
527 With the SEAD panel, we first assessed the imputation performance across populations at the
528 global scale by using HGDP database. We extracted the UK Biobank Axiom array sites as
529 pseudo array from all HGDP populations, and extracted the Illumina Asian Screening Array sites
530 from HGDP East Asian and Central and South Asian populations as pseudo array. Then the non-
531 reference (NR) allele concordance rates were compared between true and imputed genotypes. It
532 was not surprising that the SEAD reference panel outperformed other reference panels in
533 imputing all East Asian populations and some of the Central and South Asian populations.
534 Interestingly, the relative high concordance rate was achieved by the SEAD reference panel in
535 the Oceania population, this might be attributable to shared haplotypes between the Oceania
536 population and Singaporean Malays (20). Among all the Asian populations imputed by the
537 SEAD panel, the NR concordance rate obtained from the pseudo array based on the Axiom array
538 was consistently higher compared to that of the pseudo array extracted from Asian Screening
539 array sites. These higher concordance rates could be attributed to the fact that the Axiom array
540 contains approximately 150,000 more variants compared to the Asian Screening array.
541
542 As WBBC cohort had samples that were both sequenced and genotyped, we then assessed the
543 performance of the panels on Chinese population imputation. We found that the concordance rate
544 values of the SEAD panel were slightly higher than those of WBBC and SG10K, but the
545 differences were minimal. This suggested that the accuracy did not significantly improve after
546 the combination, this is consistent with our pervious study that the imputation accuracy would
547 decrease if the diverse sample exceed a set amount (10). However, the SEAD panel
548 outperformed the other four panels in terms of well-imputed variants, with almost twice the
549 number of variants at MAF bin less than 0.1% compared to SG10K and WBBC imputation.
550 Finally, if we focused on the variants that had been identified to be associated with diseases and
551 traits in Asian population in GWAS catalog, still, the SEAD panel performed the best.
552
553 Besides the systematical and comprehensive imputation evaluation for the panels, we
554 successfully applied the SEAD panel to a GWAS analysis for BMD traits at hip and femoral

555 neck (FN). These two traits were highly correlated in either phenotypic or genetic correlation.
556 Therefore, the association signals reported in both traits would reduce the random error of the
557 results. Through both single-variant association test and spatial clustering analysis, we detected
558 that rare variants near *SNTG1* gene were associated with hip BMD. The *SNTG1* gene also
559 reached the suggestive threshold in FN BMD analysis using both methods. The gene and variants
560 were not reported even in large-scale biobank GWAS for BMD (43), and the *SNTG1* variants
561 were predominantly rare in the Non-Finnish European and more common in the African/African
562 American and Latino/Admixed American. The *SNTG1* (Syntrophin Gamma 1) gene is known to
563 encode a cytoplasmic peripheral membrane protein, which a candidate gene for idiopathic
564 scoliosis (53). A southern Chinese cohort of patients with congenital scoliosis also identified
565 copy number variants of *SNTG1* (54). It was also noteworthy that *SNTG1* genetic variations had
566 been associated with neurodevelopmental disorders, suggesting a potential role in developmental
567 processes (55). The SNTG1 protein was involved in the Dystrophin-associated glycoprotein
568 complex (DGC) as shown in our pathway enrichment analysis. The DGC was known to play a
569 significant role in maintaining the structural integrity of muscle fibers (56). In our study, we
570 investigated the preliminary function of this gene, and suggested that overexpression of *SNTG1*
571 inhibited the proliferation and differentiation of preosteoblasts. This observation aligned with the
572 direction of effect on BMD in our GWAS results.
573
574 In summary, we constructed the haplotype reference panel for Asian population (SEAD panel)
575 with the largest number of samples and the most abundant population diversity in Asia. The
576 reference panel demonstrated excellent performance on imputing East Asian and Central and
577 South Asian populations, especially in detecting rare variants. We provided this optimal
578 imputation service online for free (https://imputationserver.westlake.edu.cn/) for genetic studies
579 in Asian populations. By applying the SEAD panel to impute the genotyping array data in
580 Chinese population, we, for the first time, successfully identified rare variants near *SNTG1* gene
581 showing association with bone mineral density.
582

583 Data availability

584 GWAS summary statistics for the Hip and FN BMD from our analysis of the WBBC data are
585 fully available at https://wbbc.westlake.edu.cn/downloads.html

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 100K consortium” for providing the “GenomeAsia pilot project” (EGAS00001002921). We thankfully acknowledge the High-performance Computing Center at Westlake University.

Funds

This work was supported by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (#2023C03164), the National Natural Science Foundation of China (#82370887), the Chinese National Key Technology R&D Program, Ministry of Science and Technology (#2021YFC2501702), and the funds from the Westlake Laboratory of Life Sciences and Biomedicine (#202208014).

Author contributions

H.-F.Z. conceptualized and designed the study. M.-Y.Y., J.-D.Z., and W.-Y.B. conducted the data analysis. X.L. conducted experiments. S.-H.Y., W.-W.Z., J.-Q.L. and Y. S. conducted the whole sequencing experiments. C.-D.Y., M.-C.Q., K.-Q.L., C.-F.W., 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.L., 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

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

615

616 # Reference

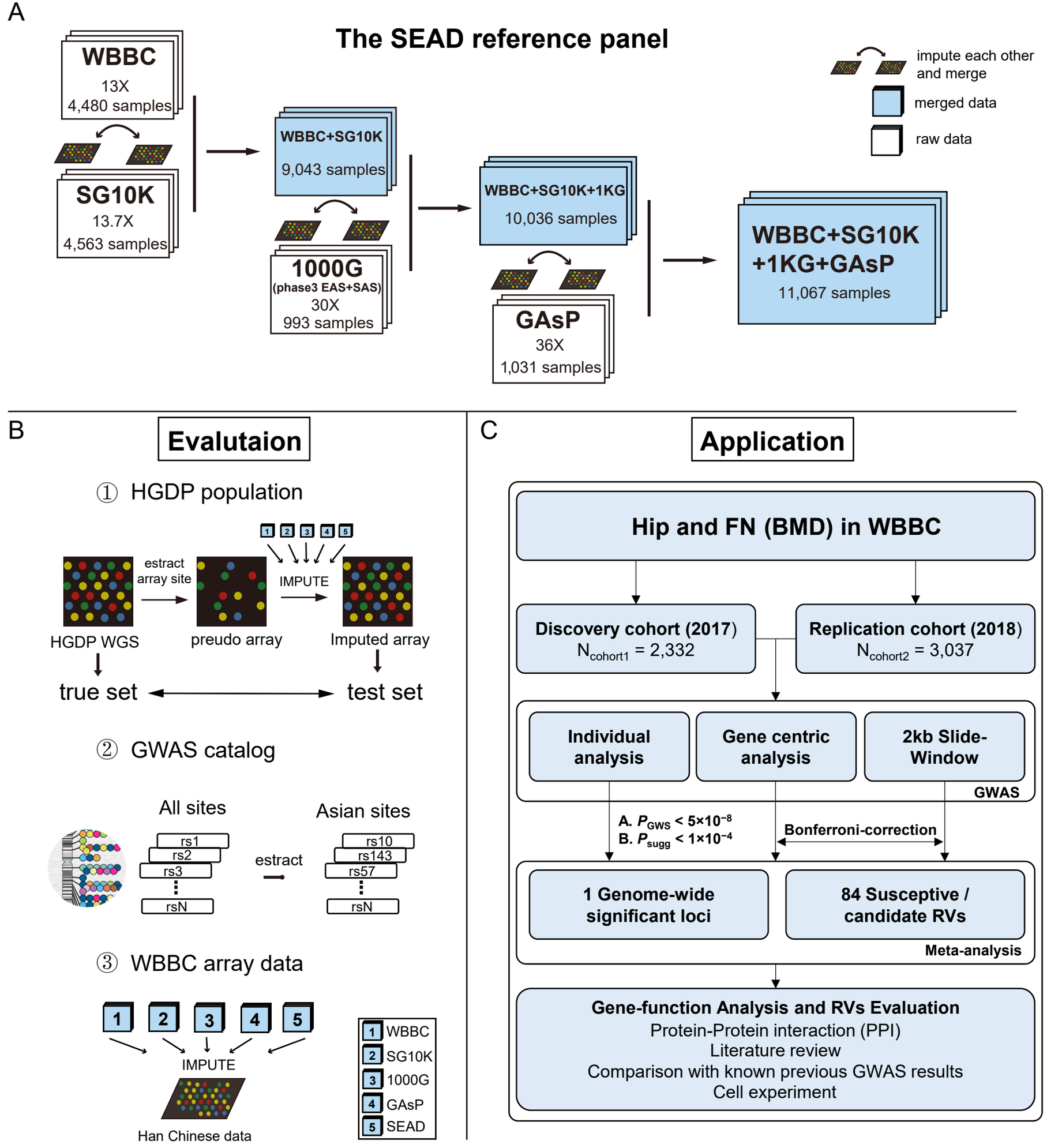
617 1. Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, et al. Fine-
618 mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-
619 specific epigenome maps. Nat Genet. 2018;50(11):1505-13.
620 2. Zheng HF, Duncan EL, Yerges-Armstrong LM, Eriksson J, Bergstrom U, Leo PJ, et al.
621 Meta-analysis of genome-wide studies identifies MEF2C SNPs associated with bone mineral
622 density at forearm. J Med Genet. 2013;50(7):473-8.
623 3. Zhu XW, Wang Y, Wei YH, Zhao PP, Wang XB, Rong JJ, et al. Comprehensive
624 Assessment of the Association between FCGRs polymorphisms and the risk of systemic lupus
625 erythematosus: Evidence from a Meta-Analysis. Sci Rep. 2016;6:31617.
626 4. Hoffmann TJ, Sakoda LC, Shen L, Jorgenson E, Habel LA, Liu J, et al. Imputation of the
627 rare HOXB13 G84E mutation and cancer risk in a large population-based cohort. PLoS Genet.
628 2015;11(1):e1004930.
629 5. Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, et al. Large
630 multiallelic copy number variations in humans. Nat Genet. 2015;47(3):296-303.
631 6. Das S, Abecasis GR, Browning BL. Genotype Imputation from Large Reference Panels.
632 Annu Rev Genomics Hum Genet. 2018;19:73-96.
633 7. Nelson SC, Stilp AM, Papanicolaou GJ, Taylor KD, Rotter JI, Thornton TA, et al. Improved
634 imputation accuracy in Hispanic/Latino populations with larger and more diverse reference
635 panels: applications in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL).
636 Hum Mol Genet. 2016;25(15):3245-54.
637 8. Lert-Itthiporn W, Suktitipat B, Grove H, Sakuntabhai A, Malasit P, Tangthawornchaikul N,
638 et al. Validation of genotype imputation in Southeast Asian populations and the effect of single
639 nucleotide polymorphism annotation on imputation outcome. BMC Med Genet. 2018;19(1):23.
640 9. Vergara C, Parker MM, Franco L, Cho MH, Valencia-Duarte AV, Beaty TH, et al. Genotype
641 imputation performance of three reference panels using African ancestry individuals. Hum
642 Genet. 2018;137(4):281-92.

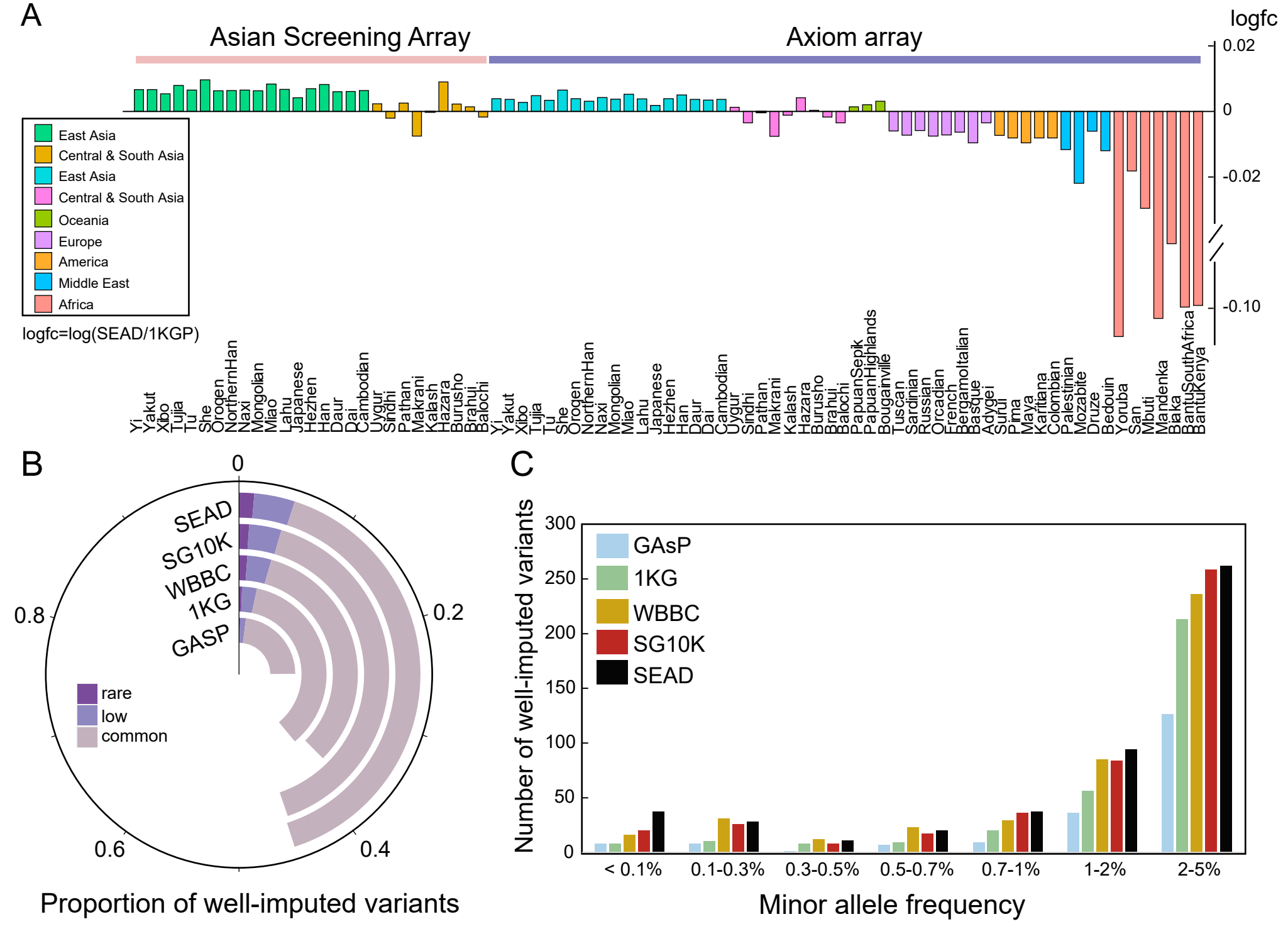
643 10. Bai WY, Zhu XW, Cong PK, Zhang XJ, Richards JB, Zheng HF. Genotype imputation and
644 reference panel: a systematic evaluation on haplotype size and diversity. Brief Bioinform. 2019.
645 11. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A
646 global reference for human genetic variation. Nature. 2015;526(7571):68-74.
647 12. Genome of the Netherlands C. Whole-genome sequence variation, population structure and
648 demographic history of the Dutch population. Nat Genet. 2014;46(8):818-25.
649 13. Zheng HF, Forgetta V, Hsu YH, Estrada K, Rosello-Diez A, Leo PJ, et al. Whole-genome
650 sequencing identifies EN1 as a determinant of bone density and fracture. Nature.
651 2015;526(7571):112-7.
652 14. Consortium UK, Walter K, Min JL, Huang J, Crooks L, Memari Y, et al. The UK10K
653 project identifies rare variants in health and disease. Nature. 2015;526(7571):82-90.
654 15. Jun G, English AC, Metcalf GA, Yang J, Chaisson MJ, Pankratz N, et al. Structural
655 variation across 138,134 samples in the TOPMed consortium. bioRxiv. 2023.
656 16. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference
657 panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016;48(10):1279-83.
658 17. Jones KM, Cook-Deegan R. Complicated legacies: The human genome at 20. Science.
659 2021;371(6529):564-9.
660 18. Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, Kawai Y, et al. Rare variant
661 discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Nat Commun.
662 2015;6:8018.
663 19. Jeon S, Bhak Y, Choi Y, Jeon Y, Kim S, Jang J, et al. Korean Genome Project: 1094 Korean
664 personal genomes with clinical information. Science Advances. 2020;6(eaaz7835).
665 20. Wu D, Dou J, Chai X, Bellis C, Wilm A, Shih CC, et al. Large-Scale Whole-Genome
666 Sequencing of Three Diverse Asian Populations in Singapore. Cell. 2019;179(3):736-49 e15.
667 21. GenomeAsia KC. The GenomeAsia 100K Project enables genetic discoveries across Asia.
668 Nature. 2019;576(7785):106-11.
669 22. Zhang P, Luo H, Li Y, Wang Y, Wang J, Zheng Y, et al. NyuWa Genome resource: A deep
670 whole-genome sequencing-based variation profile and reference panel for the Chinese
671 population. Cell Rep. 2021;37(7):110017.
672 23. Li L, Huang P, Sun X, Wang S, Xu M, Liu S, et al. The ChinaMAP reference panel for the
673 accurate genotype imputation in Chinese populations. Cell Res. 2021;31(12):1308-10.

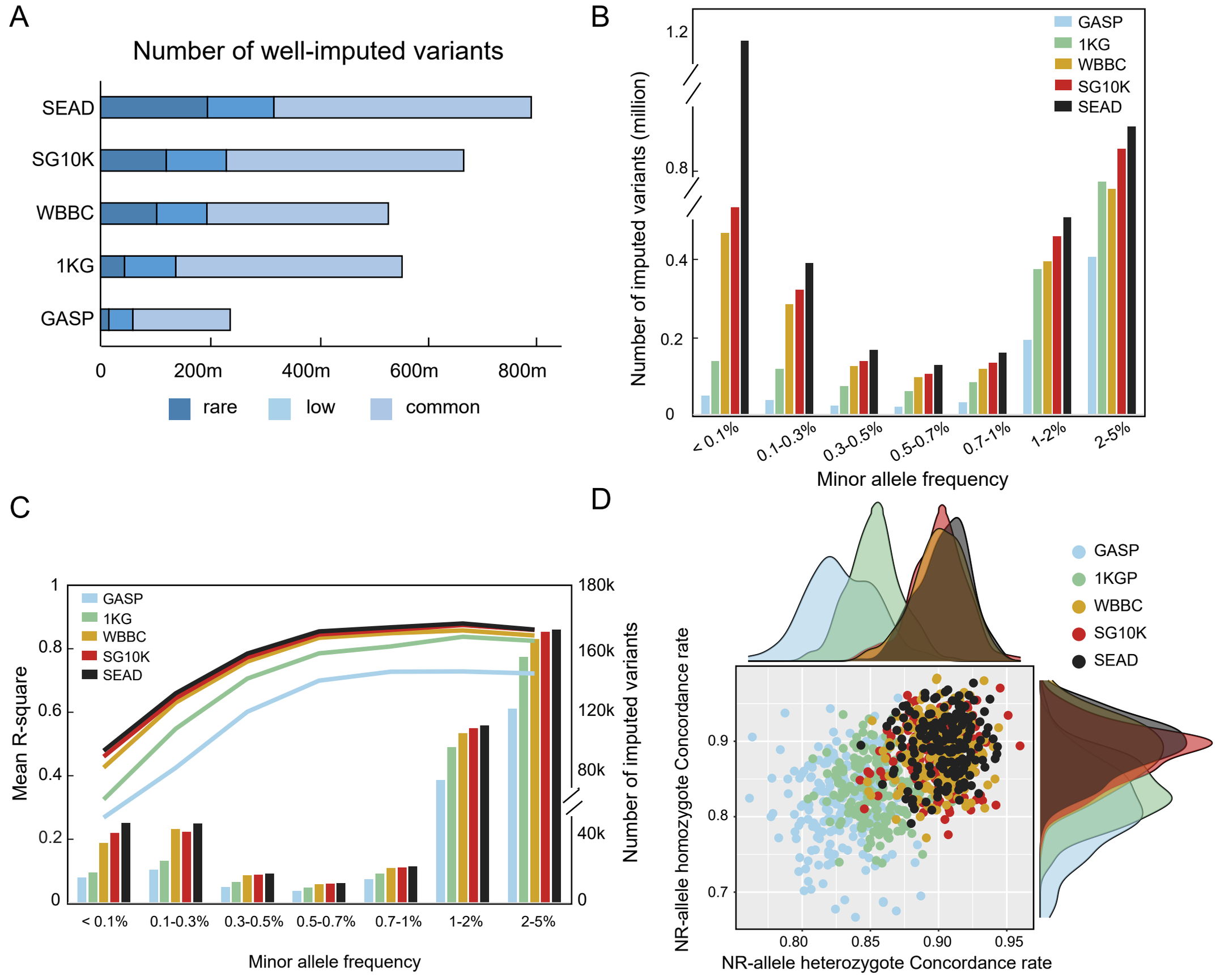
674 24. Cong PK, Bai WY, Li JC, Yang MY, Khederzadeh S, Gai SR, et al. Genomic analyses of
675 10,376 individuals in the Westlake BioBank for Chinese (WBBC) pilot project. Nat Commun.
676 2022;13(1):2939.
677 25. Wang C, Dai J, Qin N, Fan J, Ma H, Chen C, et al. Analyses of rare predisposing variants of
678 lung cancer in 6,004 whole genomes in Chinese. Cancer Cell. 2022;40(10):1223-39 e6.
679 26. Zhu XW, Liu KQ, Wang PY, Liu JQ, Chen JY, Xu XJ, et al. Cohort profile: the Westlake
680 BioBank for Chinese (WBBC) pilot project. BMJ Open. 2021;11(6):e045564.
681 27. Cong PK, Khederzadeh S, Yuan CD, Ma RJ, Zhang YY, Liu JQ, et al. Identification of
682 clinically actionable secondary genetic variants from whole-genome sequencing in a large-scale
683 Chinese population. Clin Transl Med. 2022;12(5):e866.
684 28. Gibson G. Rare and common variants: twenty arguments. Nat Rev Genet. 2012;13(2):135-
685 45.
686 29. Alioto TS, Buchhalter I, Derdak S, Hutter B, Eldridge MD, Hovig E, et al. A comprehensive
687 assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat
688 Commun. 2015;6:10001.
689 30. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, et al. High-coverage
690 whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios.
691 Cell. 2022;185(18):3426-40 e19.
692 31. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, et al. The
693 UCSC Genome Browser Database: update 2006. Nucleic Acids Res. 2006;34(Database
694 issue):D590-8.
695 32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
696 Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
697 33. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship
698 inference in genome-wide association studies. Bioinformatics. 2010;26(22):2867-73.
699 34. Bergstrom A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into
700 human genetic variation and population history from 929 diverse genomes. Science.
701 2020;367(6484).
702 35. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait
703 analysis. Am J Hum Genet. 2011;88(1):76-82.

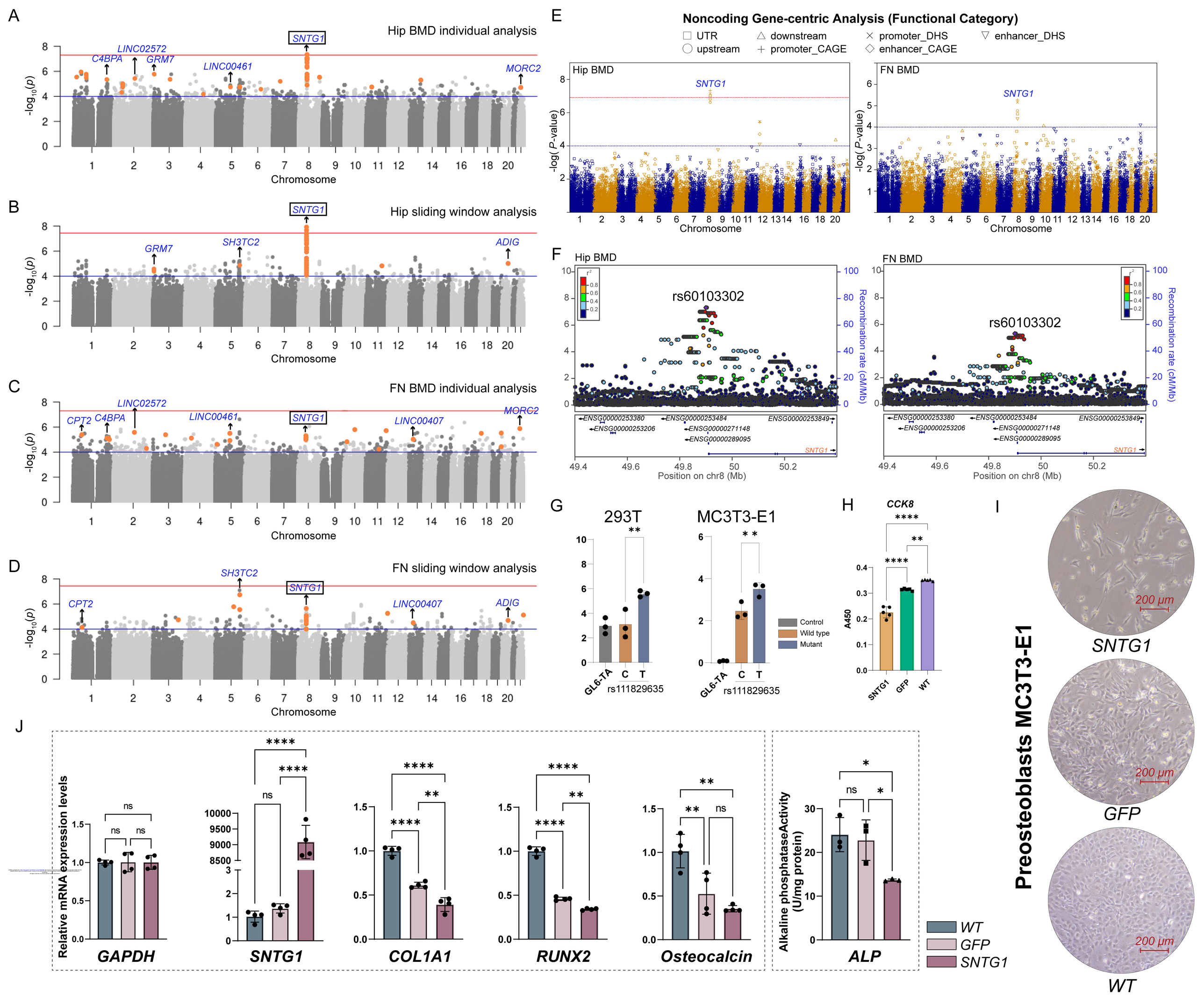
704 36. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation
705 PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
706 37. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation
707 genotype imputation service and methods. Nat Genet. 2016;48(10):1284-7.
708 38. Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation.
709 Bioinformatics. 2015;31(5):782-4.
710 39. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide
711 association scans. Bioinformatics. 2010;26(17):2190-1.
712 40. Li Z, Li X, Zhou H, Gaynor SM, Selvaraj MS, Arapoglou T, et al. A framework for
713 detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies.
714 Nat Methods. 2022;19(12):1599-611.
715 41. Bai WY, Wang L, Ying ZM, Hu B, Xu L, Zhang GQ, et al. Identification of PIEZO1
716 polymorphisms for human bone mineral density. Bone. 2020;133:115247.
717 42. Matsuda T, Cepko CL. Electroporation and RNA interference in the rodent retina in vivo
718 and in vitro. Proceedings of the National Academy of Sciences of the United States of America.
719 2004;101(1):16-22.
720 43. Morris JA, Kemp JP, Youlten SE, Laurent L, Logan JG, Chai RC, et al. An atlas of genetic
721 influences on osteoporosis in humans and mice. Nat Genet. 2019;51(2):258-66.
722 44. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, et al. The
723 STRING database in 2023: protein-protein association networks and functional enrichment
724 analyses for any sequenced genome of interest. Nucleic Acids Res. 2023;51(D1):D638-D46.
725 45. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a
726 software environment for integrated models of biomolecular interaction networks. Genome Res.
727 2003;13(11):2498-504.
728 46. Huang J, Howie B, McCarthy S, Memari Y, Walter K, Min JL, et al. Improved imputation of
729 low-frequency and rare variants using the UK10K haplotype reference panel. Nat Commun.
730 2015;6:8111.
731 47. Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, et al. The NHGRI-EBI GWAS
732 Catalog: knowledgebase and deposition resource. Nucleic Acids Res. 2023;51(D1):D977-D85.

733 48. Medina-Gomez C, Kemp JP, Trajanoska K, Luan J, Chesi A, Ahluwalia TS, et al. Life-
734 Course Genome-wide Association Study Meta-analysis of Total Body BMD and Assessment of
735 Age-Specific Effects. Am J Hum Genet. 2018;102(1):88-102.
736 49. Zheng HF, Ladouceur M, Greenwood CM, Richards JB. Effect of genome-wide genotyping
737 and reference panels on rare variants imputation. J Genet Genomics. 2012;39(10):545-50.
738 50. Zheng HF, Rong JJ, Liu M, Han F, Zhang XW, Richards JB, et al. Performance of genotype
739 imputation for low frequency and rare variants from the 1000 genomes. PLoS One.
740 2015;10(1):e0116487.
741 51. Chou WC, Zheng HF, Cheng CH, Yan H, Wang L, Han F, et al. A combined reference panel
742 from the 1000 Genomes and UK10K projects improved rare variant imputation in European and
743 Chinese samples. Sci Rep. 2016;6:39313.
744 52. Choi J, Kim S, Kim J, Son H-Y, Yoo S-K, Kim C-U, et al. A whole-genome reference panel
745 of 14,393 individuals for East Asian populations accelerates discovery of rare functional
746 variants. SCIENCE ADVANCES. 2023;9(eadg6319).
747 53. Bashiardes S, Veile R, Allen M, Wise CA, Dobbs M, Morcuende JA, et al. SNTG1, the gene
748 encoding gamma1-syntrophin: a candidate gene for idiopathic scoliosis. Hum Genet.
749 2004;115(1):81-9.
750 54. Lai W, Feng X, Yue M, Cheung PWH, Choi VNT, Song YQ, et al. Identification of Copy
751 Number Variants in a Southern Chinese Cohort of Patients with Congenital Scoliosis. Genes
752 (Basel). 2021;12(8).
753 55. Lemos RR, Oliveira DF, Zatz M, Oliveira JR. Population and computational analysis of the
754 MGEA6 P521A variation as a risk factor for familial idiopathic basal ganglia calcification
755 (Fahr's disease). J Mol Neurosci. 2011;43(3):333-6.
756 56. Omairi S, Hau KL, Collins-Hooper H, Scott C, Vaiyapuri S, Torelli S, et al. Regulation of
757 the dystrophin-associated glycoprotein complex composition by the metabolic properties of
758 muscle fibres. Sci Rep. 2019;9(1):2770.

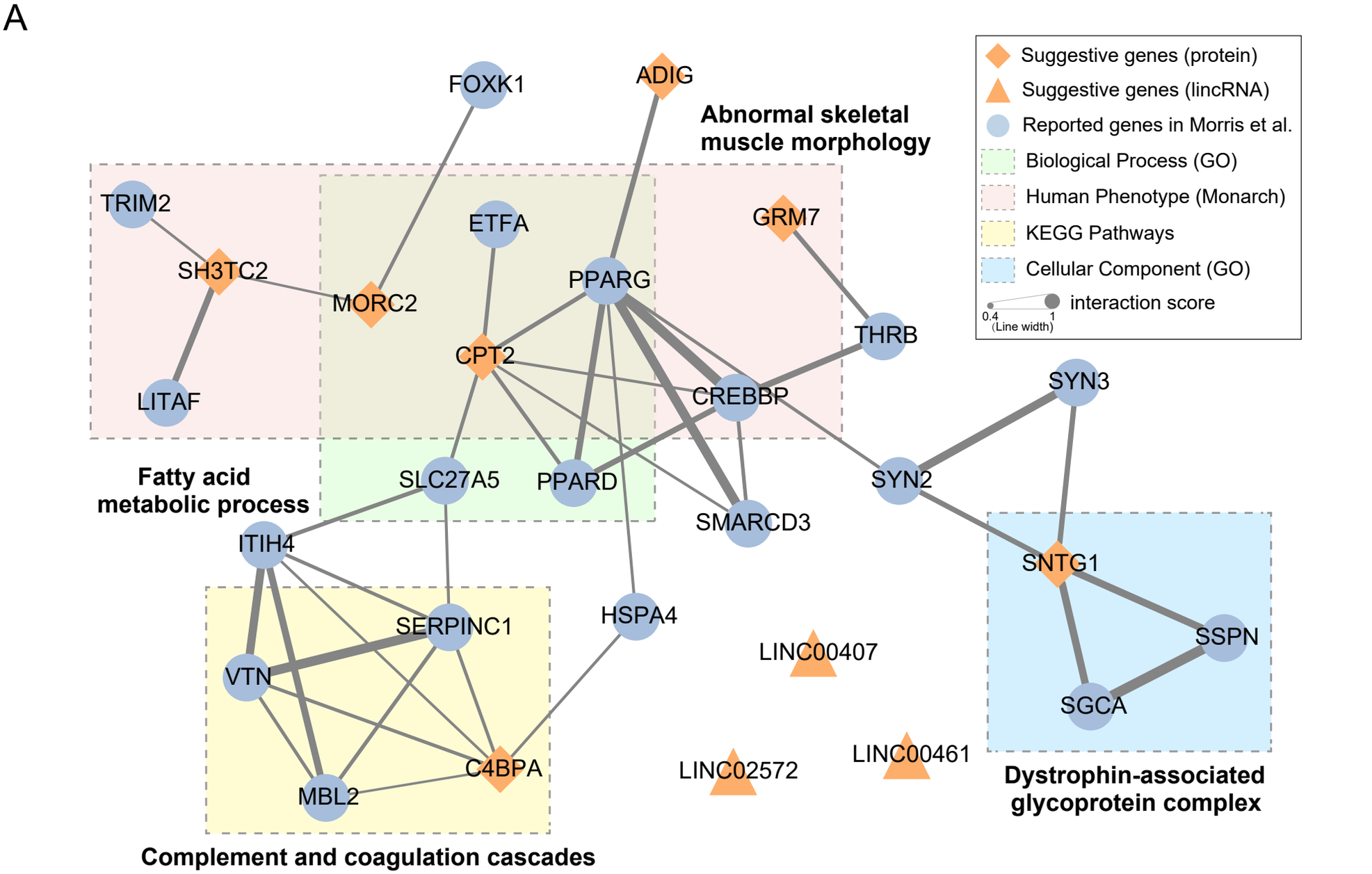

A
The SEAD reference panel
WBBC
13X
4,480 samples
SG10K
13.7X
4,563 samples
WBBC+SG10K
9,043 samples
1000G
(phase3 EAS+SAS)
30X
993 samples
WBBC+SG10K+1KG
10,036 samples
GAsP
36X
1,031 samples
WBBC+SG10K
+1KG+GAsP
11,067 samples
impute each other
and merge
merged data
raw data
B
Evalutaion
① HGDP population
estract
array site
IMPUTE
HGDP WGS
preudo array
Imputed array
true set
test set
② GWAS catalog
All sites
Asian sites
rs1
rs2
rs3
rsN
estract
rs10
rs143
rs57
rsN
③ WBBC array data
IMPUTE
Han Chinese data
1 WBBC
2 SG10K
3 1000G
4 GAsP
5 SEAD
C
Application
Hip and FN (BMD) in WBBC
Discovery cohort (2017)
N cohort1 = 2,332
Replication cohort (2018)
N cohort2 = 3,037
Individual analysis
Gene centric analysis
2kb Slide-Window
GWAS
A. P GWS < 5×10−8
B. P sugg < 1×10−4
Bonferroni-correction
1 Genome-wide significant loci
84 Susceptive / candidate RVs
Meta-analysis
Gene-function Analysis and RVs Evaluation
Protein-Protein interaction (PPI)
Literature review
Comparison with known previous GWAS results
Cell experiment


A
Asian Screening Array
Axiom array
logfc
0.02
0
-0.02
-0.10
East Asia
Central & South Asia
East Asia
Central & South Asia
Oceania
Europe
America
Middle East
Africa
logfc=log(SEAD/1KGP)
Yi
Yakut
Xibo
Tujia
Tu
She
Orogen
NorthernHan
Naxi
Mongolian
Miao
Lahu
Japanese
Hezhen
Han
Daur
Dai
Cambodian
Uygur
Sindhi
Pathan
Makrani
Kalash
Hazara
Burusho
Brahui
Balochi
PapuanSepik
PapuanHighlands
Bougainville
Tuscan
Sardinian
Russian
Orcadian
French
Bergamoltalian
Basque
Adygei
Surui
Pima
Maya
Karitiana
Colombian
Palestinian
Mozabite
Druze
Bedouin
Yoruba
San
Mbuti
Mandenka
Biaka
BantuSouthAfrica
BantuKenya
B
0
0.2
0.4
0.6
0.8
SEAD
SG10K
WBBC
1KG
GASP
rare
low
common
Proportion of well-imputed variants
C
Number of well-imputed variants
300
250
200
150
100
50
0
GAsP
1KG
WBBC
SG10K
SEAD
< 0.1%
0.1-0.3%
0.3-0.5%
0.5-0.7%
0.7-1%
1-2%
2-5%
Minor allele frequency

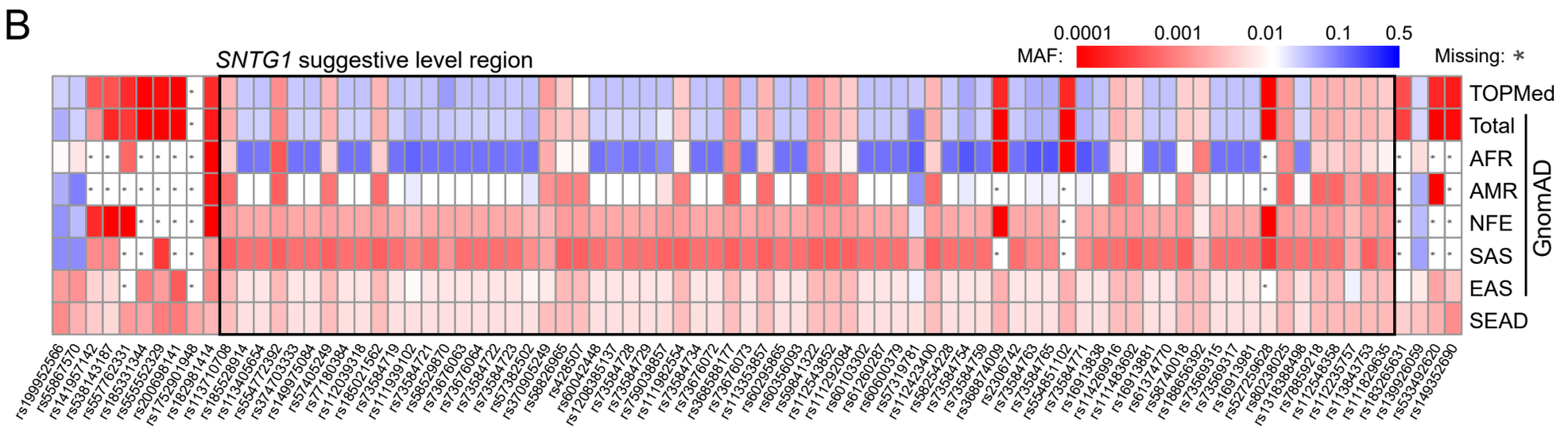

A
Number of well-imputed variants
SEAD
SG10K
WBBC
1KG
GASP
0
200m
400m
600m
800m
rare
low
common
B
GASP
1KG
WBBC
SG10K
SEAD
Number of imputed variants (million)
1.2
0.8
0.4
0.2
0
< 0.1%
0.1-0.3%
0.3-0.5%
0.5-0.7%
0.7-1%
1-2%
2-5%
Minor allele frequency
C
GASP
1KG
WBBC
SG10K
SEAD
Mean R-square
1
0.8
0.6
0.4
0.2
0
Number of imputed variants
180k
160k
120k
80k
40k
0
< 0.1%
0.1-0.3%
0.3-0.5%
0.5-0.7%
0.7-1%
1-2%
2-5%
Minor allele frequency
D
GASP
1KGP
WBBC
SG10K
SEAD
NR-allele homozygote Concordance rate
0.9
0.8
0.7
0.80
0.85
0.90
0.95
NR-allele heterozygote Concordance rate


A
Hip BMD individual analysis
SNTG1
LINC02572
C4BPA
GRM7
LINC00461
MORC2
-log10(p)
Chromosome
B
Hip sliding window analysis
SNTG1
GRM7
SH3TC2
ADIG
C
FN BMD individual analysis
LINC02572
CPT2
C4BPA
LINC00461
SNTG1
LINC00407
MORC2
D
FN sliding window analysis
SH3TC2
SNTG1
CPT2
LINC00407
ADIG
E
Noncoding Gene-centric Analysis (Functional Category)
UTR
upstream
downstream
promoter_CAGE
promoter_DHS
enhancer_CAGE
enhancer_DHS
Hip BMD
FN BMD
SNTG1
-log(P-value)
Chromosome
F
Hip BMD
FN BMD
rs60103302
Recombination rate (cM/Mb)
Position on chr8 (Mb)
ENSG00000253380
ENSG00000253206
ENSG00000253484
ENSG00000271148
ENSG00000289095
ENSG00000253849
SNTG1
G
293T
MC3T3-E1
GL6-TA
C
T
rs111829635
Control
Wild type
Mutant
H
CCK8
A450
SNTG1
GFP
WT
I
Preosteoblasts MC3T3-E1
200 μm
SNTG1
GFP
WT
J
Relative mRNA expression levels
GAPDH
SNTG1
COL1A1
RUNX2
Osteocalcin
Alkaline phosphataseActivity (U/mg protein)
ALP
WT
GFP
SNTG1

A


Suggestive genes (protein)
Suggestive genes (lincRNA)
Reported genes in Morris et al.
Biological Process (GO)
Human Phenotype (Monarch)
KEGG Pathways
Cellular Component (GO)
interaction score
0.4
1
(Line width)
Abnormal skeletal muscle morphology
Fatty acid metabolic process
Complement and coagulation cascades
Dystrophin-associated glycoprotein complex
FOXK1
ADIG
TRIM2
SH3TC2
MORC2
ETFA
GRM7
PPARG
THRB
CPT2
CREBBP
LITAF
SYN3
SLC27A5
PPARD
SYN2
SMARCD3
ITIH4
SNTG1
SERPINC1
HSPA4
SSPN
LINC00407
VTN
SGCA
C4BPA
LINC00461
LINC02572
MBL2

B


SNTG1 suggestive level region
MAF:
0.0001
0.001
0.01
0.1
0.5
Missing: *
TOPMed
Total
AFR
AMR
NFE
SAS
EAS
SEAD
GnomAD