nature genetics

Article

https://doi.org/10.1038/s41588-026-02612-z

Genome-wide associations of structural variants with human traits through imputation from long-read assemblies

Received: 20 October 2025

Accepted: 24 April 2026

Published online: 20 May 2026

Check for updates

Wei-Yang Bai[1,2], Shuli Liu[1,2], Zhongqu Duan[1,2], Ji-Jian Yang[3], Jie Chen[2,4], Junren Hou[1,2], Lianfeng Wu[2,4], Nan Li[3], Ting Qi[1,2,5] & Jian Yang[1,2]

Structural variants (SVs) are a major type of genetic variation, yet their role in human traits remains largely uncharacterized, primarily due to challenges in genotyping them on a genome-wide scale in large cohorts. Here we identified 171,233 high-quality, genome-wide SVs from 482 haplotype-resolved genome assemblies derived from PacBio HiFi long-read sequencing of 241 individuals. We developed a reference panel and a web application (ImputeSV) to impute these SVs from single-nucleotide polymorphism (SNP) data and demonstrated high imputation accuracy at both the individual and cohort levels. Using this tool, we imputed 54,578 common SVs (minor allele frequencies (MAFs) ≥1%) in 456,643 UK Biobank (UKB) participants of European ancestry. Through analysis of UKB data and simulations, we estimated that SVs contributed to at least 4.7% of the common genetic variation for complex traits. Genome-wide association analyses of SVs for 2,624 UKB traits identified 17,335 SV–trait associations, including 958 unlikely to be driven by small genetic variants. Our study demonstrates the power of using long-read assemblies for imputing SVs from SNPs, unveils the role of SVs in complex trait variation and provides a catalog of SV associations in the UKB.

Structural variants (SVs), typically defined as genomic alterations of at least 50 base pairs (bp), are a major source of genetic variation in humans[1]. These variants, which are present throughout the human genome, manifest in multiple forms such as insertions, deletions, inversions and other complex events[2]. A considerable number of SVs are found within repetitive genomic regions, often composed of tandemly repeated sequences of a specific motif, such as variable number tandem repeats (VNTRs)[3,4]. Despite an extensive body of research highlighting the impact of SVs on human traits and diseases[5–8], genome-wide association studies (GWASs) for SVs lag far behind those for small genetic variants (SGVs)[9], primarily due to the challenges of detecting and genotyping SVs at a genome-wide scale in large cohorts[10,11].

Detecting and genotyping SVs using single-nucleotide polymorphism (SNP) array data is feasible, but it is generally limited to a few large SVs[12]. Identifying SVs from short-read whole-genome sequencing (srWGS) data, such as those from the UK Biobank (UKB)[13] and All of Us[14] studies, has been markedly more successful than using SNP arrays[2]. However, pinpointing SVs in complex genomic regions (for example, repetitive elements) using srWGS data remains a formidable challenge, restricting the number of reliably identified SVs to between 9,000 and 13,000 per individual[13–17]. The current gold standard for calling

[1]New Cornerstone Science Laboratory, School of Life Sciences, Westlake University, Hangzhou, China. [2]Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, China. [3]High-Performance Computing Center, Westlake University, Hangzhou, China. [4]School of Life Sciences, Westlake University, Hangzhou, China. [5]Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China. e-mail: jian.yang@westlake.edu.cn

SVs is high-accuracy long-read WGS (for example, the PacBio HiFi technology)[18], which allows for the identification of ~25,000 SVs per individual[19,20]. Despite its effectiveness, the cost of these technologies presents a considerable obstacle to their application in large cohorts[21]. One alternative strategy is to build a graph-based pangenome from highly accurate de novo assemblies in a relatively small cohort and to map short-read data from large cohorts to the pangenome for SV genotyping[22]. However, this strategy requires the availability of srWGS data in large cohorts, leaving vast and valuable SNP array datasets underused and highlighting the critical need for a comprehensive SV reference panel to enable accurate imputation from SNP data.

Genotype imputation has proven successful in inferring untyped SGVs, especially those that are common in the population[23–25]. Yet, the task of SV imputation remains challenging. While prior efforts have used srWGS data or early-stage moderate-coverage Oxford Nanopore Technologies to create SV imputation panels, these approaches often under-represent complex SVs or face challenges due to high error rates (Supplementary Note 1).

Assembly-based SV calling is currently considered one of the most powerful and reliable approaches for SV detection[26]. In this study, we aimed to leverage de novo genome assemblies recently constructed from high-accuracy, long-read WGS data to identify a comprehensive set of high-quality SVs and to develop a tool (ImputeSV) to impute these SVs in individuals with either SNP array or srWGS data. We identified 171,233 high-quality SVs from 482 haplotype-resolved long-read assemblies derived from HiFi long-read data for 241 individuals of diverse ancestries. When benchmarked against 9,228 well-established SVs, our SV imputation exhibited high recall, precision and genotype concordance, especially when the input SNP coverage was high, even if most SNPs were imputed. Using this approach, we imputed 54,578 common SVs (minor allele frequencies (MAFs) ≥1%) in 456,643 UKB individuals of European ancestry, quantified the variance in various complex traits explained by the SVs and conducted genome-wide SV association analyses for 2,624 traits in the UKB.

Results

Reference panel and online tool for genome-wide SV imputation

We obtained 482 haplotype-resolved genome assemblies derived from high-coverage HiFi (hcHiFi) long-read data for 241 individuals of diverse ancestries, including 53 Africans (AFR), 24 Admixed Americans (AMR), 62 East Asians (EAS), 38 Europeans (EUR), 53 West Asians (WAS) and 11 South Asians (SAS; Extended Data Fig. 1 and Supplementary Tables 1–3). We evaluated three assembly-based variant calling methods and benchmarked them against Genome in a Bottle (GIAB) standards, ultimately selecting PAV-called SVs supported by at least one additional caller for highest accuracy (Methods; Extended Data Fig. 2, Supplementary Note 2 and Supplementary Figs. 1 and 2). Applying this strategy, we identified an average of 15,765 high-confidence, noncentromeric autosomal SVs per assembly (Fig. 1a).

Considering that the same SVs can be detected with different representations in different assemblies[27], we merged SVs using Jasmine[28] (Methods; Supplementary Fig. 3), yielding a total of 171,233 unique SVs across all assemblies (Fig. 1b). The length of SVs varied with a median of 160 bp and a mean of 1,589 bp (Fig. 1c), comparable to those in prior work[29]. We annotated all SVs using AnnotSV[30] and found that, while more than 98% SVs were in intergenic and intronic regions, 3,070 SVs were located entirely within exons, and 911 overlapped splicing sites (Fig. 1d). Furthermore, we developed a pipeline to discover VNTRs by analyzing the sequence of each SV allele in every assembly (Methods; Supplementary Fig. 4). We identified 18,360 autosomal VNTR loci (Supplementary Table 4), of which 292 were located within exons (Supplementary Fig. 5). Each VNTR consisted of at least 2 alternative alleles, with an average of 21.4 alleles per locus, and 779 loci contained 100 or more alleles (Fig. 1e). Approximately 66.0% SVs (113,055/171,233) identified in this study were in tandem repeat regions (TRRs), and, as expected, the size of these SVs differed from that of other SVs (Supplementary Fig. 6).

To facilitate public use, we adapted a standard imputation pipeline using our assembly-derived variants to construct optimized reference panels for general SVs and specific VNTRs (Supplementary Note 3 and Supplementary Fig. 7). This pipeline is publicly accessible at https://yanglab.westlake.edu.cn/ImputeSV.

Evaluating the performance of SV imputation

We evaluated our SV imputation panel using the well-characterized HG002 genome and a leave-one-out imputation strategy across our 241 reference individuals. Using various simulated SNP arrays and pre-imputed SNPs as inputs, our pipeline demonstrated high recall, precision and genotype concordance against established benchmark datasets (GIAB, HGSVC3 and 1KGP3), including within complex and medically relevant genomic regions (Fig. 2a,b, Extended Data Figs. 3 and 4, Supplementary Fig. 8 and Supplementary Note 4; Methods). Pre-imputing SNPs from array data substantially improved SV imputation (achieving F1 scores of 0.92 and genotype concordance rates up to 0.97 against GIAB benchmarks), yielding performance nearly comparable to using WGS-derived SNPs. At the population level, common SVs (MAF $\geq$ 0.01) were imputed with high accuracy (median, $r^2 = 0.78$) across diverse ancestries (Fig. 2c and Supplementary Note 4).

We further compared our hcHiFi panel against existing long-read (moderate-coverage Oxford Nanopore Technologies) and srWGS imputation panels. Our panel provided more comprehensive genome-wide SV coverage, achieved a substantially higher F1 score and yielded over twice as many accurately imputed SVs (Supplementary Note 4, Extended Data Fig. 5 and Supplementary Fig. 9). Finally, our VNTR-specific panel successfully captured multi-allelic VNTR length variations across ancestries and exhibited strong genotype concordance with srWGS calls in the UKB (Fig. 2d, Supplementary Fig. 10 and Supplementary Note 4).

Estimation of variance explained by SVs for complex traits in the UKB

We applied our pipeline to impute SVs in all the UKB-EUR individuals based on TOPMed-imputed SNPs with imputation info scores of ≥0.9, using SHAPEIT4 (ref. 31) for SNP phasing and Minimac4 (ref. 32) for SV imputation (Methods; Extended Data Fig. 6). Across all samples, an average of 19,578 SVs were imputed per individual. Imputed SVs with MAF of ≥0.01 were retained, resulting in 54,578 SVs. Of these, 93.7% were found among the SVs called in the EUR assemblies of our reference panel (Extended Data Fig. 7), and large SVs (>1 kbp) in high-mappability regions had high MAF concordance with those called using short-read data in gnomAD[33] ($r = 0.83$) and UKB-EUR ($r = 0.93$; Extended Data Fig. 7). We also included in the analysis ~8.8 million TOPMed-imputed SGVs with MAF of ≥0.01 that were available from the UKB. We estimated the proportion of phenotypic variation explained by the imputed common SVs and SGVs, denoted by $h^2_{cSV(imp)}$ and $h^2_{cSGV(imp)}$, respectively, for 636 quantitative traits in ~350,000 unrelated UKB-EUR individuals using GCTA-GREML[34] based on two models, a joint model fitting SVs and SGVs simultaneously as two random-effect components and a marginal model fitting each component separately. In the separate analysis, the estimates of $h^2_{cSV(imp)}$ and $h^2_{cSGV(imp)}$ (denoted by $\hat{h}^2_{cSV(imp)}$ and $\hat{h}^2_{cSGV(imp)}$) averaged across the 636 traits were 6.7% and 20.9%, respectively (Fig. 3a and Supplementary Table 5). When SVs and SGVs were analyzed jointly, the mean $\hat{h}^2_{cSV(imp)}$ decreased substantially to 1.0% (s.e.m. = 0.32%), whereas $\hat{h}^2_{cSGV(imp)}$ only decreased slightly to 20.2% (s.e.m. = 1.1%), indicating that the imputed SVs accounted for at least 4.7% (that is, $\frac{1.0\%}{1.0\% + 20.2\%}$) of the common genetic variance on average across traits (Fig. 3b). The estimated variance explained by the imputed SVs and SGVs was correlated (correlation of 0.40 between the

https://doi.org/10.1038/s41588-026-02612-z

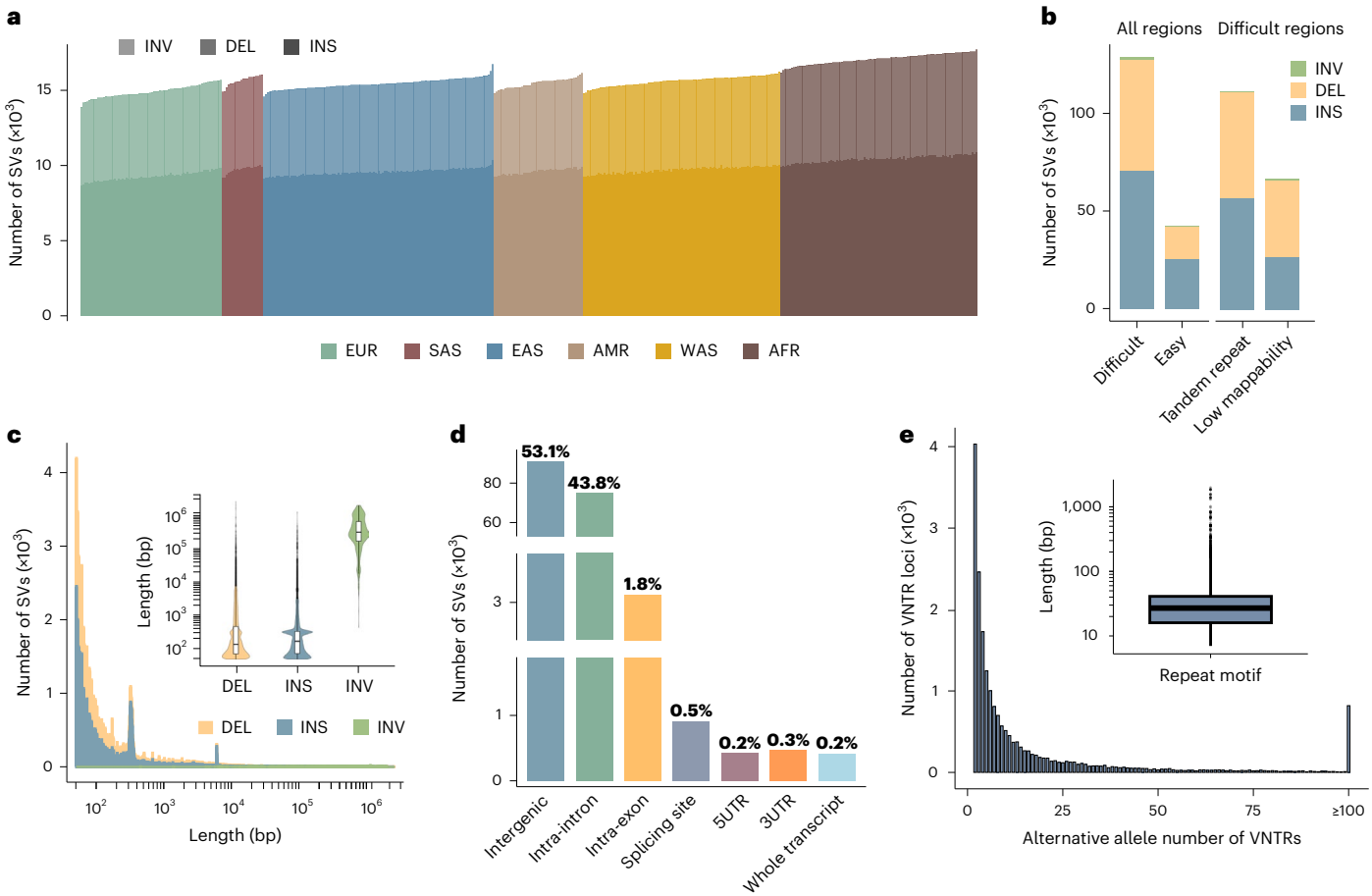


Fig. 1 | Descriptive summary of SVs and VNTRs detected from 482 long-read assemblies. a, The number of SVs stratified by INS, DEL and INV per haplotype. The assemblies were collected from six super-populations, namely, AFR, AMR, EAS, EUR, WAS and SAS. **b**, The number of SVs across all samples in stratified genomic regions defined by GIAB (v3.5). **c**, The length distribution of SVs ($m_{DEL} = 74,176$; $m_{INS} = 95,942$; $m_{INV} = 1,115$). The Alu, SVA and LINE peaks are observed at approximately 300 bp, 1–3 kb and 6 kb, respectively. **d**, The number of SVs located in different genomic regions. The *y* axis is truncated to better illustrate results. **e**, The number of alternative alleles of each VNTR locus identified from 482 haplotypes. The length distribution of motifs ($m = 18,360$) is shown, with the median, IQR and whiskers extending to 1.5× IQR; data beyond this range are displayed as outliers. The box plots in **c** and **e** display the medians, upper quartiles and lower quartiles; whiskers extend to 1.5× IQR; outliers are defined as ±1.5× IQR. INS, insertion; DEL, deletion; INV, inversion; IQR, interquartile range.

joint $\hat{h}^2_{cSV(imp)}$ and $\hat{h}^2_{cSGV(imp)}$ and of 0.96 between the marginal $\hat{h}^2_{cSV(imp)}$ and $\hat{h}^2_{cSGV(imp)}$; Fig. 3c) due to shared effector genes and the tagging of causal SVs by SGVs and vice versa. There were 132 traits for which the joint $\hat{h}^2_{cSV(imp)}$ was significant (false discovery rate < 0.05), suggesting a contribution from SV effects even after accounting for the effects of SGVs, with the joint $\hat{h}^2_{cSV(imp)}$ for 48 widely studied traits illustrated in Fig. 3d. We also performed the imputation and GREML analyses in the UKB individuals of African (UKB-AFR) and American (UKB-AMR) ancestries for 111 quantitative traits with sample sizes of >3,000. For UKB-AFR, the mean $\hat{h}^2_{cSV(imp)}$ and $\hat{h}^2_{cSGV(imp)}$ from the joint analysis were 3.4% (s.e.m. = 1.1%) and 28.9% (s.e.m. = 2.0%), respectively, and, for UKB-AMR, they were 3.6% (s.e.m. = 1.2%) and 19.4% (s.e.m. = 2.1%), respectively (Extended Data Fig. 8 and Supplementary Table 6). The estimates for non-European ancestries were more variable due to smaller sample sizes.

To investigate the large difference in estimated heritability between the joint and marginal models, we performed a simulation study based on 1KGP3-EUR WGS data (Supplementary Note 5). The simulations revealed that the marginal model overestimates h^2_{cSV} due to the tagging of causal SGVs, whereas the joint model underestimates it due to competition from imputed SGVs. By combining our empirical GREML results with these simulations, we extrapolate that common SVs may account for ~8% the common genetic variance for complex traits on average (Supplementary Note 6 and Supplementary Figs. 11–14).

Genome-wide associations of SVs with 2,624 traits in the UKB

We conducted genome-wide association analyses of the imputed SVs (referred to as SV-GWAS hereafter) for 2,624 UKB traits in up to 456,643 individuals of European ancestry using GCTA-fastGWA[35,36], controlling for genetic relatedness among the individuals. We identified a total of 17,335 SV–trait associations with $P_{SV-GWAS} < 5 \times 10^{-8}$, involving 4,397 SVs and 656 traits (Supplementary Table 7). All association results can be queried, visualized and downloaded through our online data portal (https://yanglab.westlake.edu.cn/data/ukb_sv_gwas). More than half of the trait-associated SVs were associated with multiple traits (Fig. 4a), with the top 20 most pleiotropic SVs listed in Supplementary Table 8. The three most pleiotropic SVs, located in the 4q24, 9q34.2 and 7p22.2 genomic loci, were associated with more than 50 traits, mainly related to brain gray matter, blood biochemistry and fornix microstructural integrity. The trait-associated SVs were significantly longer (Supplementary Fig. 15) and located closer to transcriptional start sites (TSS) and splice sites (SS; Fig. 4b,c), compared with other SVs. Comparing our SV associations with a large-scale short-read study[37] highlighted the complementary strengths of our hcHiFi-based panel in resolving complex genomic regions versus the statistical power of larger short-read panels (Supplementary Note 7 and Extended Data Fig. 9).

Due to extensive linkage disequilibrium (LD), it is challenging to determine whether an observed association is driven by SVs or nearby

https://doi.org/10.1038/s41588-026-02612-z

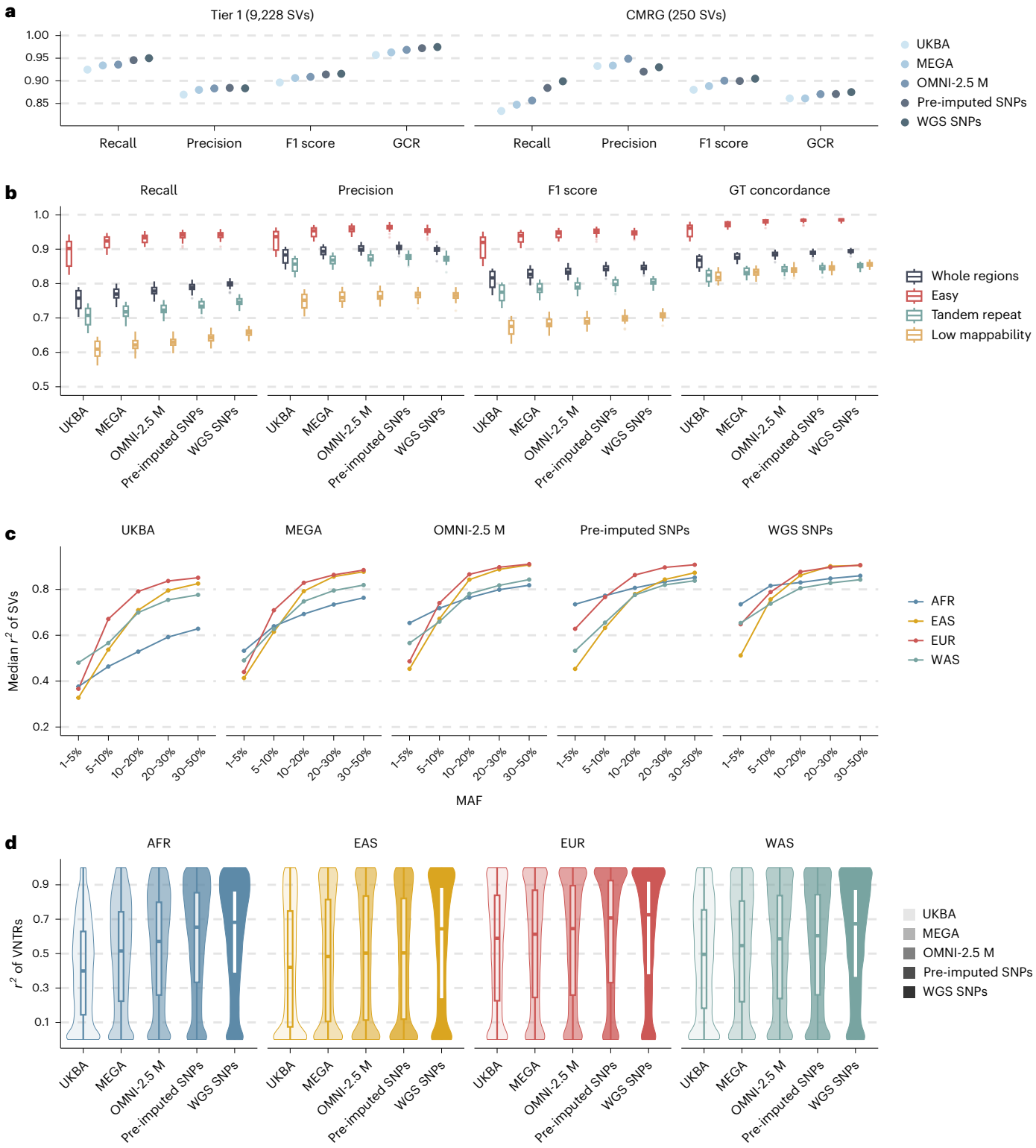


Fig. 2 | Evaluation of the performance of the SV imputation pipeline. Five input SNP sets were tested, including UKBA (the UKB Axiom array), MEGA (Multi-Ethnic Global array), OMNI-2.5M (Infinium Omni-2.5M array), pre-imputed array (imputed SNPs based on UKBA using the TOPMed reference panel) and WGS SNPs (SNPs called from assemblies). **a**, Imputed SVs of HG002 benchmarked against the GIAB Tier 1 (v0.6) SV set (left) and the CMRG SV set (right). GCR, genotype concordance rate. **b**, Leave-one-out analysis. HGSVC3 samples ($n = 65$) were analyzed. The SVs were stratified by different genomic regions defined in the GIAB genome stratification BED files. **c**, Squared correlation (r^2) between the genotype of SV called from genome assemblies and that imputed using the SV panel for each of the five input sets with diverse ancestry. Leave-one-out strategy was used for the imputation. Please note that AMR and SAS ancestries were excluded due to insufficient sample size for reliable r^2 calculation. **d**, The r^2 between the length of VNTRs ($m = 18,360$) called from genome assemblies and that imputed using the VNTR panel with each of the five input sets. The box plots in **b** and **d** display the medians, upper quartiles and lower quartiles; whiskers extending to 1.5× IQR; outliers are defined as ±1.5× IQR. 'Whole region' refers to 'genome-wide regions', 'easy' refers to 'not in any difficult-to-locate regions', 'tandem repeat' refers to 'TRRs' and 'low mappability' refers to 'low mappability regions'.

https://doi.org/10.1038/s41588-026-02612-z

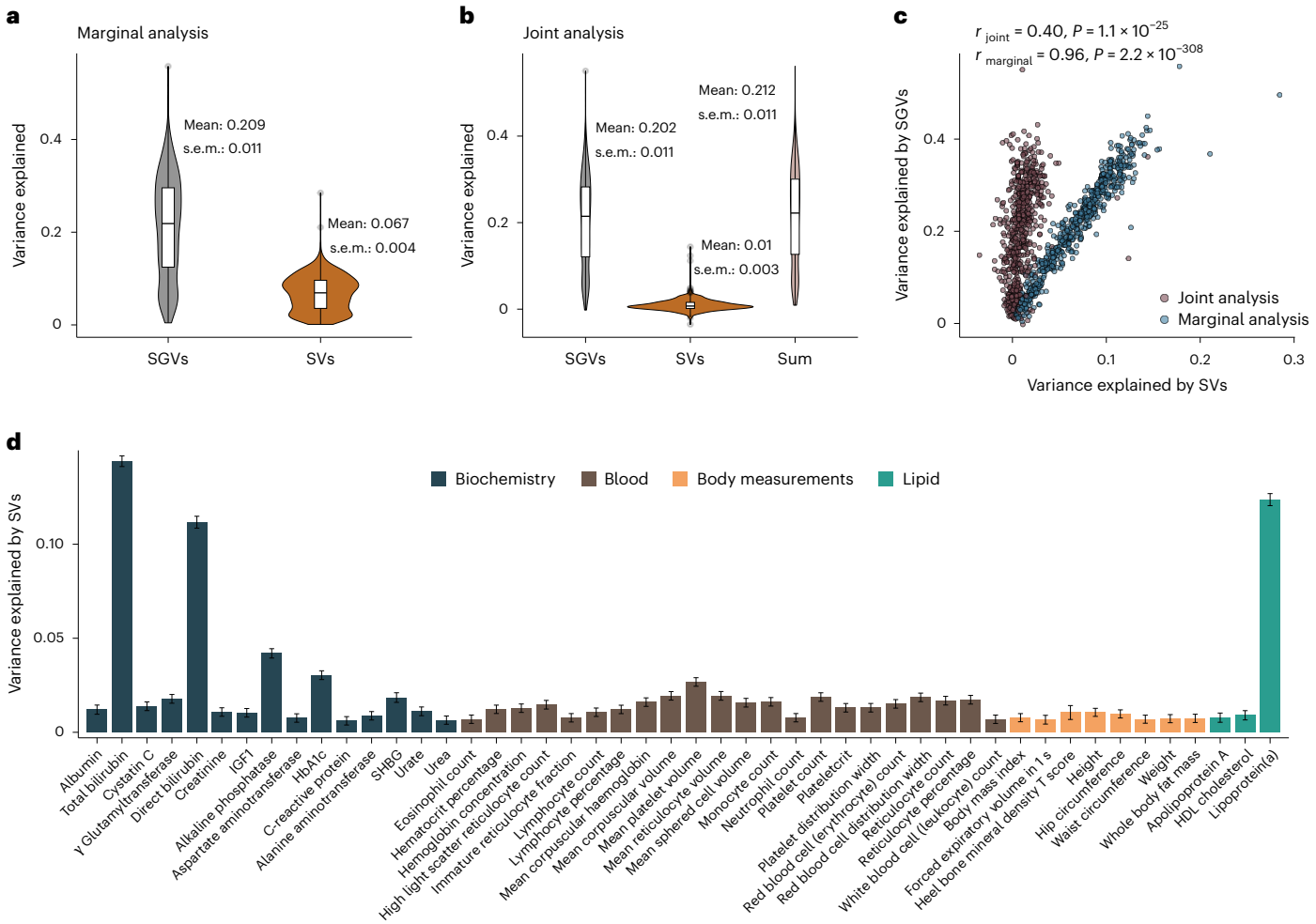


Fig. 3 | Estimated SV-based and SGV-based heritability for 636 quantitative traits in the UKB. a,**b**, The estimates from a marginal analysis where the SVs or SGVs were fitted separately in a single component model (**a**) and a joint analysis where the SVs and SGVs were fitted together in a two-component model ($m_{\text{trait}} = 636$) (**b**) are shown. The medians, upper quartiles, lower quartiles of the estimates and whiskers extending to 1.5× IQR are plotted, with outliers defined as ±1.5× IQR. **c**, Comparison of estimates of variance explained by SVs and SGVs. Each point represents a single trait, plotting the estimated proportion of phenotypic variance explained by SVs against that explained by SGVs. The Pearson correlation coefficients between the heritability estimates from the joint and marginal models are shown. *P* values were calculated using a two-sided test. **d**, Estimated variance explained by SVs for 48 widely studied quantitative traits. The plot shows the estimated proportion of phenotypic variance explained by SVs in a joint model that fits both SVs and SGVs for the 44 UKB traits, for which this estimate was statistically significant. Traits are grouped into four categories. Sample sizes for each trait are provided in Supplementary Table 5. Each bar represents the heritability estimate for a single trait, and error bars indicate ±1 s.e. SHBG, sex hormone-binding globulin.

SGVs. Through simulations, we established stringent criteria, requiring the SV to be the lead variant, remain significant after conditioning on SGVs, or possess the highest fine-mapping posterior inclusion probability (PIP), to identify SV associations highly unlikely to be driven by SGVs (Supplementary Note 8). Applying these criteria to our UKB results yielded 958 distinct, high-confidence SV-derived association loci (Supplementary Tables 9–11). Although SVs are larger, their individual effect sizes were comparable to SGVs (Supplementary Fig. 16).

Functional implications of the trait-associated SVs

To further explore the functional relevance of trait-associated SVs, we assessed their enrichment across functional genomic categories in multiple tissues and cell types. Compared to other SVs, the trait-associated SVs were significantly enriched in transcriptionally active regions and functional regulatory elements, such as gene bodies, TSS and enhancers, across most cell types (Fig. 4d,e) and depleted in constitutive heterochromatin and regions with stably repressed transcriptional activity. Besides directly disrupting the coding regions of genes by sequence insertion or deletion, another plausible hypothesis, given the considerable sizes of the SVs, is that they influence traits through effects on chromatin contacts or protein binding to DNA or RNA[38]. This hypothesis is supported by our observation that the trait-associated SVs were enriched at the boundaries of topologically associating domains (TAD) and RNA-binding protein (RBP) binding sites across various cell types and tissues (Fig. 4f,g). Additionally, to determine whether the SVs affect the traits through effects on gene expression, we applied our SV imputation pipeline to the Genotype-Tissue Expression (GTEx; v8) genotype data[39] and conducted an SV-based expression quantitative trait locus (referred to as SV-eQTL) analysis using OSCA[40]. This analysis revealed that 38.8% trait-associated SVs (1,804 of 4,655) were SV-eQTLs in at least one of the GTEx tissues (Supplementary Table 12), including 208 *trans*-SV-eQTLs that exerted long-range effects (>5 Mbp) on remote genes. While the mechanisms through which SVs affect complex traits are likely diverse and require in-depth investigations in the future, we illustrate in the examples below potential mechanisms through which SVs affect complex traits by disrupting functional elements (for example, enhancer and HNRNPL binding sites) that regulate gene expression.

https://doi.org/10.1038/s41588-026-02612-z

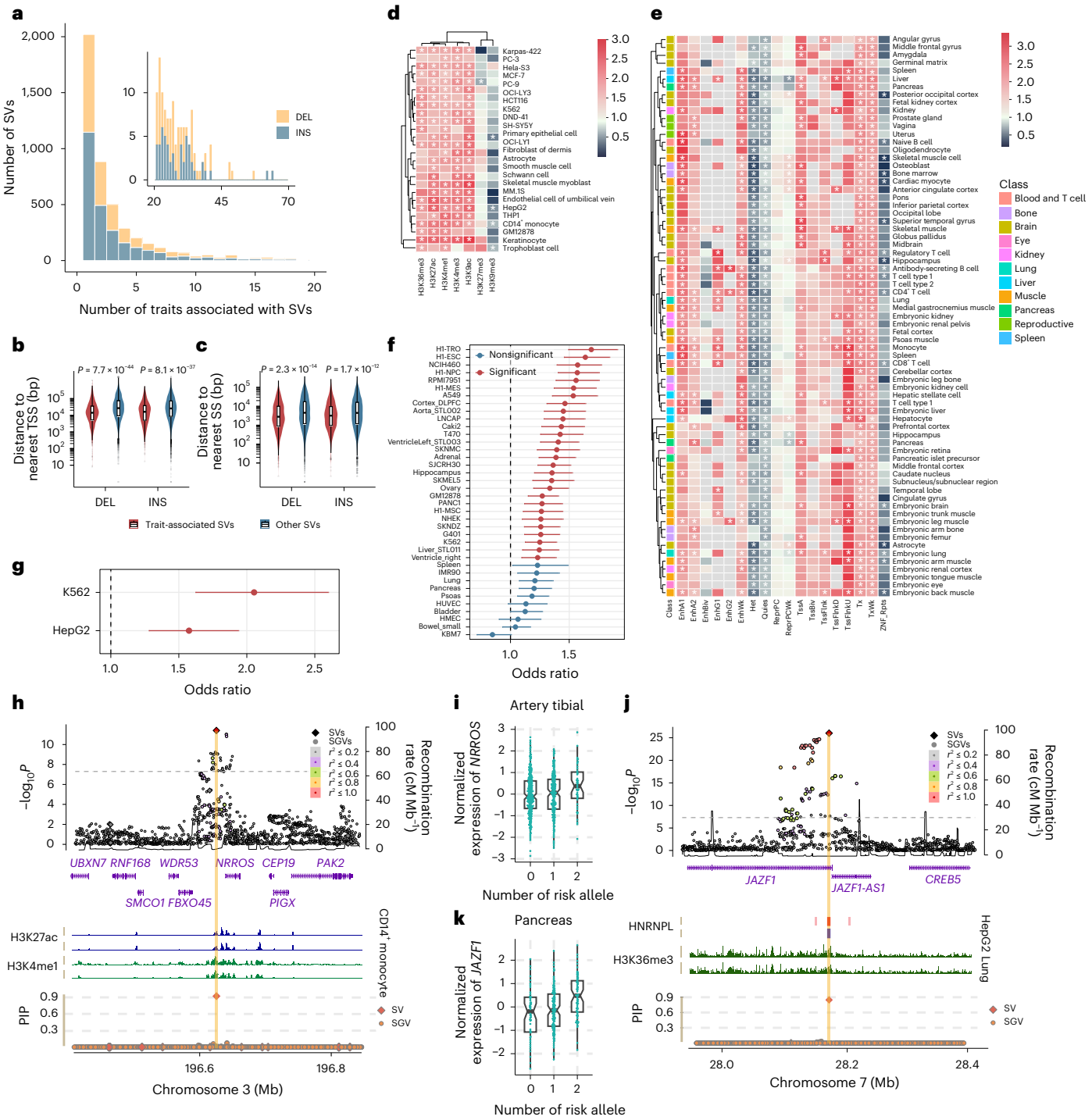


Fig. 4 | Trait-associated SVs. a, Distribution of the number of SVs associated with multiple traits. **b**,**c**, Comparison between trait-associated SVs ($m = 4,397$) and other SVs ($m = 32,372$) in terms of distance to the nearest TSS (**b**) and SS (**c**). The two-sided Wilcoxon rank-sum test was used to determine whether there is a significant difference across groups. **d**–**g**, Enrichment of trait-associated SVs ($m = 4,397$) in histone modifications (**d**), chromatin states (**e**), TAD boundaries (**f**) and RBP binding sites (**g**). The *y* axis represents various cell types and tissues. In the heatmap, the odds ratio is indicated by a color bar in the legend, with red indicating enrichment and blue indicating depletion. Asterisks indicate statistically significant results after correction for multiple testing. Points indicate odds ratios, and horizontal error bars represent 95% CIs. **h**, The association plot shows the SV locus (3q29) associated with asthma (top), H3K27ac and H3K4me1 modification peaks in CD14$^+$ monocytes (middle) and fine-mapping PIP (bottom). The dashed line in the association plot represents a significance level of 5×10^{-8}. **i**, Expression level of *NRROS* in the tibial artery according to the GTEx data ($n = 476$). **j**, The association plot displays the SV locus (7p15.1) associated with T2D (top), eCLIP signal for HNRNPL in HepG2 cells, H3K36me3 ChIP–seq peaks in lung (middle) and fine-mapping PIP (bottom). **k**, The expression level of *JAZF1* in the pancreas according to the GTEx data ($n = 243$). The box plots in **b**,**c**,**i**,**k** show the medians, upper quartiles and lower quartiles; whiskers extend to 1.5× the IQR; outliers are defined as ±1.5× the IQR. CIs, confidence intervals.

These examples also demonstrate how the identification of a putative causal SV can facilitate inferring the underlying mechanism.

An intergenic deletion (denoted by 3q29-DEL-60bp; chr3: 196,625,684–196,625,743), situated within the regions of H3K27ac (chr3:196,622,786–196,626,104) and H3K4me1 (chr3:196,625,100–196,626,481) modifications in $CD14^+$ monocytes, exhibited the strongest association with asthma among all genetic variants in this locus (Fig. 4h). Thus, 3q29-DEL-60bp was accurately imputed with an info score of 0.98 and is located within a putative *cis*-regulatory element (GeneHancer ID−GH03J196623) that interacts with *NRROS*[41], which has been previously reported as a negative regulator of reactive oxygen species during host defense and autoimmunity[42]. A significant association between 3q29-DEL-60bp and the expression level of *NRROS* in the tibial artery ($P_{SV-eQTL} = 4.0 \times 10^{-12}$) was observed in our SV-eQTL analysis (Fig. 4i). Fine-mapping analysis of the SV and SGV GWAS data in this locus revealed a PIP of 0.93 for 3q29-DEL-60bp. Additionally, it was also associated with eosinophil percentage with PIPs of 0.84 (Supplementary Table 11), which is implicated in eosinophilic asthma, a subtype of severe asthma[43].

Another well-imputed deletion SV (info score = 0.94), located in the gene body of *JAZF1* (denoted as 7p15.1-DEL-364bp; chr7: 28,174,681–28,175,044), was associated with type 2 diabetes (T2D; Fig. 4j). Previous studies have shown that *JAZF1* expression is reduced in the pancreatic islets of type 2 diabetes patients, and that increased expression of type *JAZF1* is associated with higher insulin secretion[44–46]. In this study, 7p15.1-DEL-364bp overlaps with an RBP binding site (chr7: 28,174,808–28,174,851) of HNRNPL in HepG2 cells (ENCODE−ENCSR724RDN), which is a component of the heterogeneous nuclear ribonucleoprotein complexes and is hypothesized to work in conjunction with H3K36me3 to have a major role in the formation, processing and function of mRNA[47,48]. In our study, 7p15.1-DEL-364bp was associated with *JAZF1* expression in multiple tissues (Supplementary Table 12), with the strongest SV-eQTL signal observed in the pancreas ($P_{SV-eQTL} = 2.2 \times 10^{-17}$; Fig. 4k). Although several variants, including both SVs and SGVs in this locus, displayed a high LD with one another, 7p15.1-DEL-364bp had the highest PIP of 0.85 in fine mapping (Fig. 4j), suggesting a distinct association signal.

Genome-wide associations of VNTRs with the UKB traits

SVs are often present in the genome as tandem repeats. To investigate whether the effect of such an SV changes in proportion to its number of repeat motifs, we applied our VNTR imputation pipeline to the UKB-EUR data and imputed 16,448 VNTRs, 84.1% of which were present in the EUR assemblies of our reference panel (Extended Data Fig. 7). We quantified the repeat number for each VNTR and associated them with each of the 2,624 UKB traits (Methods). The VNTR association analysis was performed in 349,662 unrelated individuals, as no GWAS tool is currently available for VNTR association analysis while controlling for relatedness in biobank-scale data. In total, we identified 7,295 VNTR associations at $P_{VNTR-GWAS} < 5 \times 10^{-8}$, involving 1,861 VNTRs and 831 traits (Supplementary Table 13). Of the 1,861 trait-associated VNTRs, 1,040 were associated with at least two traits (Supplementary Fig. 17a), and 1,017 (54.6%) were VNTR-eQTLs with $P_{VNTR-eQTL} < 5 \times 10^{-8}$ in at least one tissue in the GTEx data (Supplementary Table 14). The trait-associated VNTRs were significantly closer to TSS and SS than other VNTRs (Supplementary Fig. 17b). Most of the trait-associated VNTRs were in noncoding regions, with motif lengths ranging from 7 bp to 472 bp. A comparison with a short-read VNTR study[8] underscored the distinct capability of our HiFi-based panel for robust VNTR discovery, complementing the extensive allelic diversity captured by larger short-read panels (Supplementary Note 9 and Extended Data Fig. 10). In the following examples, we demonstrate plausible mechanisms by which a noncoding VNTR motif serves as a specific binding site for regulatory elements. Consequently, a larger number of motifs corresponds to an increased likelihood of binding, thereby leading to amplified functional consequences.

An intronic VNTR (denoted by 22q11.23-VNTR-7, chr22: 24,605,024–24,606,149), located within *GGT1*, was strongly associated with γ-glutamyltransferase level ($P_{VNTR-GWAS} = 5.02 \times 10^{-241}$). The level of this enzyme showed an ascending trend with an increased number of motifs (Fig. 5a). The eQTL analysis results indicate that 22q11.23-VNTR-7 was also a VNTR-eQTL, positively associated with *GGT1* expression in multiple tissues in the GTEx data (Supplementary Table 14), with the strongest association in the thyroid ($P_{VNTR-eQTL} = 2.5 \times 10^{-24}$; Fig. 5b). We compared the sequence of the repeat motif, $(AT)_3A$, to known motifs using MEME Suite[49] and found a significant match to an RBP binding site (Supplementary Fig. 18a). This site is targeted by multiple RBPs that contain RNA recognition motifs, including RBMS1, RBMS2 and RBMS3 (ref. 50). According to the GTEx data, the expression levels of *GGT1* and the *RBMS* genes were significantly correlated in multiple tissues, with the strongest correlation in the small intestine (Fig. 5c and Supplementary Table 15). These results indicate that 22q11.23-VNTR-7 may affect γ-glutamyltransferase level through its effect on RBP-regulated expression of *GGT1*.

Another VNTR (denoted by 1q44-VNTR-75, chr1: 247,870,261–247,871,078), located in the intron of *TRIM58*, was associated with both reticulocyte count ($P_{VNTR-GWAS} = 6.6 \times 10^{-88}$) and high light scatter reticulocyte count ($P_{VNTR-GWAS} = 3.5 \times 10^{-62}$), and the effect sizes increased with an increased number of motifs (Fig. 5d). Interestingly, this VNTR did not show a significant association with *TRIM58* expression in any tissue according to the GTEx data but was associated with a neighboring gene, *OR2T8* ($P_{VNTR-eQTL} = 3.2 \times 10^{-8}$), residing ~50 kbp downstream of 1q44-VNTR-75 (Fig. 5e) in the thyroid. The motif sequence of 1q44-VNTR-75 corresponded to a transcription factor binding site of POU1F1 (Supplementary Fig. 18b), which is known to regulate the expression of several genes involved in pituitary development and hormone expression[51]. These findings suggest that the repeat polymorphisms of 1q44-VNTR-75 may influence both the reticulocyte count and high light scatter reticulocyte count through the modulation of gene expression mediated by POU1F1.

Discussion

SVs are a major type of genomic variation, but their role in complex traits, including diseases, remains largely unknown, primarily due to the challenges involved in detecting and genotyping them on a genome-wide scale in large cohorts. In this study, we demonstrated the feasibility of leveraging long-read genome assemblies for imputing genome-wide SVs from SNPs. Genotype imputation relies on the presence of haplotype-resolved variants in both the reference panel and the target samples to accurately identify shared haplotype segments. Using an assembly-derived reference panel not only improves imputation accuracy by reducing haplotype phasing errors in the reference panel[52] but also yields sequences of SVs with high accuracy. Consequently, our reference panel and online SV imputation tool (ImputeSV) offer a distinctive advantage, facilitating the imputation of sequence-resolved SVs, even within complex genomic regions, for large cohorts. This unlocks the potential of vast, existing SNP datasets for SV-based discovery, providing a powerful and accessible resource for the genetics community. We have made the imputation pipeline, along with the imputation reference panel, publicly accessible through an online server. This allows re-analysis of all existing and forthcoming GWAS data to assess associations of SVs, including VNTRs, with a broad range of complex traits and diseases, particularly those underrepresented in the UKB data. It also opens a new avenue for genome-wide associations of SVs for molecular phenotypes, such as gene expression, epigenomic modifications and protein abundances. It further provides an opportunity to incorporate SVs into genetic risk prediction. We are committed to continuously expanding our reference panel, leveraging new high-quality assemblies, especially from diverse and underrepresented populations, and up-to-date methods. Additionally, graph-based pangenome methods such as PanGenie[22] genotype

https://doi.org/10.1038/s41588-026-02612-z

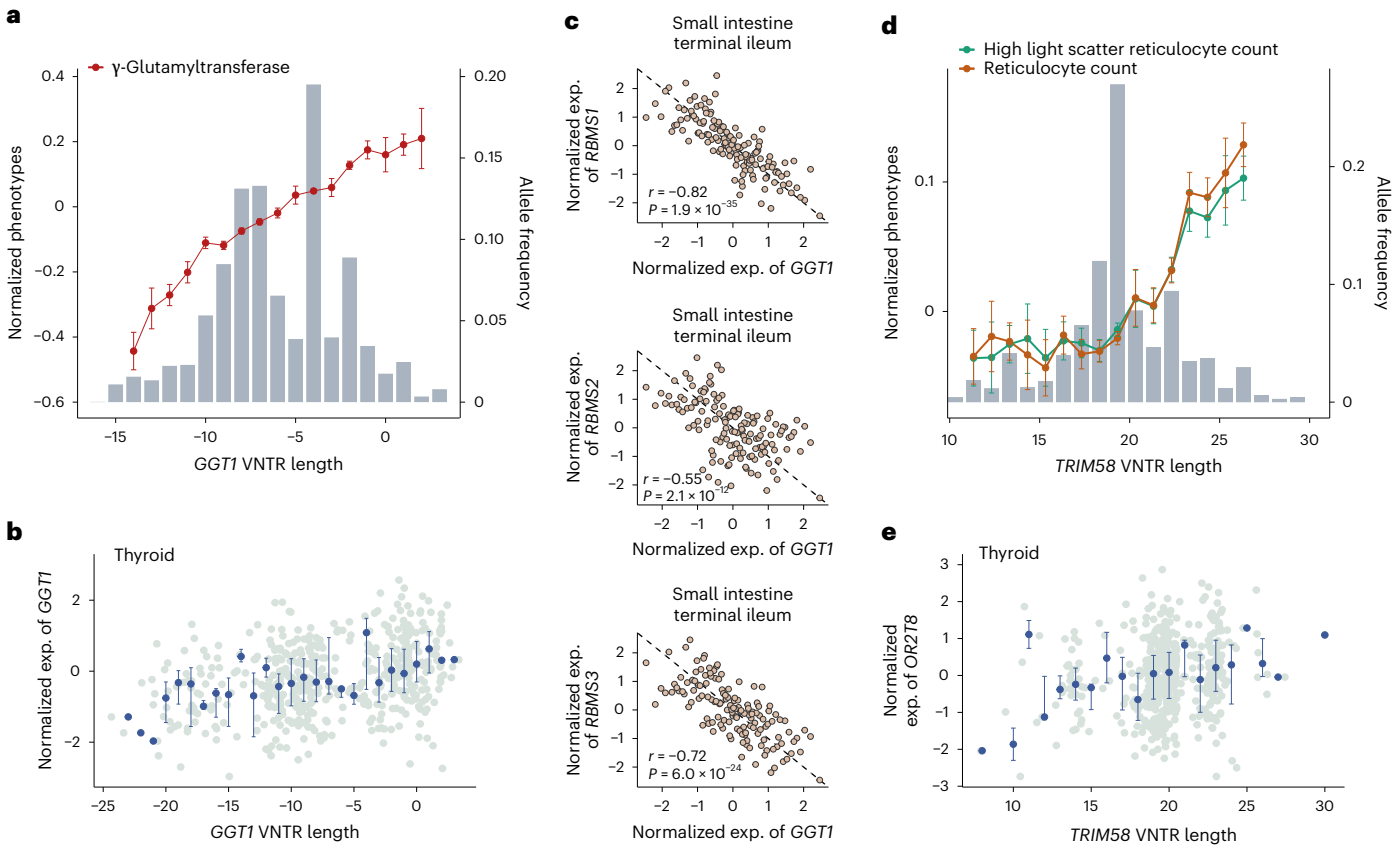


Fig. 5 | Associations of *GGT1* and *TRIM58* repeat polymorphisms with gene expression and complex traits. a, Normalized γ-glutamyltransferase levels (dotted line, left *y* axis) and frequency distribution of the VNTR alleles binned by length (histogram, right *y* axis) in UKB-EUR (n = 333,225). Data are presented as mean ± 95% CIs. **b**, Expression of *GGT1* in GTEx carriers of VNTR alleles (binned by length) in the thyroid (n = 482). The bars indicate the IQR (Q1–Q3), with the median shown. **c**, Scatterplots of the expression of *GGT1* against those of the three *RBMS* genes in the small intestine terminal ileum according to the GTEx data, with each dot representing an individual. Pearson's correlation (r) between the expression levels is shown for each pair of genes. *P* values were calculated using a two-sided test. **d**, Normalized phenotypes of high light scatter reticulocyte count (n = 333,535) and reticulocyte count (n = 333,535 line, left *y* axis), and frequency distribution of the VNTR alleles binned by length (histogram, right *y* axis). Data are presented as mean ± 95% CIs. **e**, Expression level of *OR2T8* in carriers of VNTR alleles (binned by length) in the thyroid according to the GTEx data (n = 482). The bars indicate the IQR (Q1–Q3), with the median shown. exp., expression.

SVs by mapping short reads to a pangenome graph, an approach that effectively resolves nested rearrangements and reduces reference bias. However, this approach requires srWGS data and is computationally intensive for large-scale studies. In contrast, SV imputation leverages widely available SNP array data, providing a highly scalable and cost-effective solution for genotyping. This positions pangenome-based genotyping as a high-resolution tool for complex rearrangements, while establishing imputation as a practical method for massive-scale studies. Future work should consider these trade-offs and leverage their respective strengths.

We conducted SV-GWAS for 2,624 traits in the UKB, leading to the discovery of 17,335 SV–trait associations and 7,295 VNTR–trait associations, all of which can be browsed, visualized and downloaded through our online data portal (https://yanglab.westlake.edu.cn/data/ukb_sv_gwas). This analysis not only supplements UKB srWGS-based SV associations but also constitutes a large-scale proof of concept that powerfully demonstrates our panel's utility. Regarding the SV associations discovered in our analysis of the UKB data, although a subset of the associated SVs were located in coding regions, the majority were found in noncoding regions, exhibiting significant enrichment in various genomic regions of functional importance, such as enhancers and RBP binding sites, suggesting potential mechanisms through which SVs may influence gene expression by affecting these elements. SVs are considerably larger than SGVs, which increases their likelihood of affecting functional genomic elements and, consequently, their potential to be causal in trait associations.

Nevertheless, due to the extensive LD across the genome, it is challenging to determine whether SVs, SGVs or both contribute to the genetic effects at loci with associations to both types of variants. Through simulation, we have demonstrated that, at loci with co-occurring SV and SGV associations, it is highly improbable for the genetic association signals to be driven exclusively by causal SGVs if an SV meets any of the following conditions: it is the lead variant identified by GWAS, remains genome-wide significant after conditioning on the associated SGVs or is the primary variant after fine mapping. We have identified 958 loci in the UKB data that meet at least one of these conditions. However, these criteria are stringent, providing high confidence about a distinct SV association signal at only a subset of loci but leaving a substantial portion unresolved. Future work is needed to develop methods with improved sensitivity to detect loci with distinct SGV and SV signals, using test statistics or other quantitative metrics.

Despite certain limitations (Supplementary Note 10), our study demonstrates the great potential of using long-read assemblies for genome-wide SV imputation. By leveraging 482 haplotype-resolved long-read assemblies from 241 samples, we have developed an SV imputation online server and made it publicly accessible at https://yanglab.westlake.edu.cn/ImputeSV. This enables all currently available and forthcoming GWAS and molecular quantitative trait locus data to be used to assess the roles of SVs and VNTRs in complex traits and molecular phenotypes. By applying this method to the UKB data, we were able to quantify the heritability attributable to SVs for common variants and identified a

large number of SV associations. These associations offer new insights into the mechanisms underlying genetic variation in complex traits.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-026-02612-z.

References

1. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
2. Ho, S. S., Urban, A. E. & Mills, R. E. Structural variation in the sequencing era. *Nat. Rev. Genet.* **21**, 171–189 (2020).
3. Audano, P. A. et al. Characterizing the major structural variant alleles of the human genome. *Cell* **176**, 663–675 (2019).
4. Sulovari, A. et al. Human-specific tandem repeat expansion and differential gene expression during primate evolution. *Proc. Natl Acad. Sci. USA* **116**, 23243–23253 (2019).
5. Zhang, F. et al. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat. Genet.* **41**, 849–853 (2009).
6. Carvalho, C. M. et al. Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nat. Genet.* **43**, 1074–1081 (2011).
7. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013).
8. Mukamel, R. E. et al. Repeat polymorphisms underlie top genetic risk loci for glaucoma and colorectal cancer. *Cell* **186**, 3659–3673 (2023).
9. Wheeler, M. M. et al. Whole genome sequencing identifies structural variants contributing to hematologic traits in the NHLBI TOPMed program. *Nat. Commun.* **13**, 7592 (2022).
10. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–46 (2011).
11. Kosugi, S. et al. Detection of trait-associated structural variations using short-read sequencing. *Cell Genom.* **3**, 100328 (2023).
12. Aguirre, M., Rivas, M. A. & Priest, J. Phenome-wide burden of copy-number variation in the UK Biobank. *Am. J. Hum. Genet.* **105**, 373–383 (2019).
13. UK Biobank Whole-Genome Sequencing Consortium. Whole-genome sequencing of 490,640 UK Biobank participants. *Nature* **645**, 692–701 (2025).
14. All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024).
15. Collins, R. L. et al. A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
16. Byrska-Bishop, M. et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440 (2022).
17. Collins, R. L. & Talkowski, M. E. Diversity and consequences of structural variation in the human genome. *Nat. Rev. Genet.* **26**, 443–462 (2025).
18. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
19. Liao, W. W. et al. A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
20. Logsdon, G. A. et al. Complex genetic variation in nearly complete human genomes. *Nature* **644**, 430–441 (2025).
21. Method of the year 2022: long-read sequencing. *Nat. Methods* 20, 1 (2023).
22. Ebler, J. et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat. Genet.* **54**, 518–525 (2022).
23. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).
24. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
25. Sun, B. B. et al. Genetic associations of protein-coding variants in human disease. *Nature* **603**, 95–102 (2022).
26. Pendleton, M. et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015).
27. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
28. Kirsche, M. et al. Jasmine and Iris: population-scale structural variant comparison and analysis. *Nat. Methods* **20**, 408–417 (2023).
29. Beyter, D. et al. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet.* **53**, 779–786 (2021).
30. Geoffroy, V. et al. AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics* **34**, 3572–3574 (2018).
31. Delaneau, O., Zagury, J. F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019).
32. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
33. Chen, S. et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).
34. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
35. Jiang, L. et al. A resource-efficient tool for mixed model association analysis of large-scale data. *Nat. Genet.* **51**, 1749–1755 (2019).
36. 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).
37. Halldorsson, B. V. et al. The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).
38. Merla, G. et al. Submicroscopic deletion in patients with Williams-Beuren syndrome influences expression levels of the nonhemizygous flanking genes. *Am. J. Hum. Genet.* **79**, 332–341 (2006).
39. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
40. Qi, T. et al. Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat. Commun.* **9**, 2282 (2018).
41. Fishilevich, S. et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)* **2017**, bax028 (2017).
42. Noubade, R. et al. NRROS negatively regulates reactive oxygen species during host defence and autoimmunity. *Nature* **509**, 235–239 (2014).
43. Van Hulst, G. et al. Eosinophil diversity in asthma. *Biochem. Pharmacol.* **179**, 113963 (2020).
44. Taneera, J. et al. A systems genetics approach identifies genes and pathways for type 2 diabetes in human islets. *Cell Metab.* **16**, 122–134 (2012).
45. Kobiita, A. et al. The diabetes gene JAZF1 is essential for the homeostatic control of ribosome biogenesis and function in metabolic stress. *Cell Rep.* **32**, 107846 (2020).

46. Zeggini, E. et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.* **40**, 638–645 (2008).
47. Pinol-Roma, S., Swanson, M. S., Gall, J. G. & Dreyfuss, G. A novel heterogeneous nuclear RNP protein with a unique distribution on nascent transcripts. *J. Cell Biol.* **109**, 2575–2587 (1989).
48. Kerschbamer, E. et al. CHD8 suppression impacts on histone H3 lysine 36 trimethylation and alters RNA alternative splicing. *Nucleic Acids Res.* **50**, 12809–12828 (2022).
49. Bailey, T. L. et al. MEME Suite: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
50. Ray, D. et al. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172–177 (2013).
51. Martinez-Ordonez, A. et al. POU1F1 transcription factor induces metabolic reprogramming and breast cancer progression via LDHA regulation. *Oncogene* **40**, 2725–2740 (2021).
52. Das, S., Abecasis, G. R. & Browning, B. L. Genotype imputation from large reference panels. *Annu. Rev. Genomics Hum. Genet.* **19**, 73–96 (2018).

Methods

Ethical approval

This study is approved by the Ethics Committee of Westlake University (approval 20200722YJ001).

Haplotype-resolved genome assemblies

A total of 482 high-quality, haplotype-resolved genome assemblies from 241 individuals representing six ancestry groups, namely, 53 AFR, 24 AMR, 62 EAS, 38 EUR, 53 WAS and 11 SAS, were included in our imputation reference panel (Supplementary Table 1). Of these assemblies, 424 were obtained from public resources, including the HPRC[19], HGSVC[20], Arab pangenome reference (APR)[53], Chinese Pangenome Consortium (CPC)[54], CN1 (ref. 55) and YAO[56]. These assemblies were derived from high-depth PacBio HiFi reads, complemented by multiple sequencing technologies, such as Illumina srWGS, strand-seq and Hi–C, where available, to improve the assembly process, enhance error correction and boost data completeness. The detailed methodologies and procedures for the sequencing and assembly processes are described elsewhere[19,20,53–56]. Furthermore, 58 high-quality, haplotype-resolved assemblies from 29 1KGP3 individuals were newly generated in this study using PacBio HiFi sequencing; detailed protocols for DNA extraction, assembly construction and quality assessment are provided in Supplementary Note 11 (Supplementary Tables 1–3 and Extended Data Fig. 1).

Assembly-based variant calling

We used three commonly used tools, namely, Dipcall (v0.3)[57], PAV (v2.2.4.1)[27], and SVIM-asm (v1.0.2)[58], for variant calling and genotyping. The 482 assemblies were first aligned to the human reference genome, GRCh38, excluding alternate and decoy sequences[27]. We used minimap2-unimap (v2.22)[59] for mapping, with the parameter '-x asm5' consistently used for the assembly-to-reference alignment. Please note that Dipcall and PAV have built-in pipelines that can call minimap2 for sequence mapping, while SVIM-asm uses a standalone mapping process. SAMtools (v1.14) was used to sort and index the alignments[60]. Variant calling was performed by each of the three tools using its respective default settings. All detected variants subsequently underwent stringent quality control, including regional masking, cross-caller consensus and haplotype consistency checks, to yield a highly reliable set of autosomal SVs and SGVs (Supplementary Note 12).

Benchmarking of assembly-based SVs against the GIAB HG002 SVs

We created three SV sets for HG002 using each of the three callers, along with a combined set. We obtained the GIAB HG002 Tier 1 (v0.6) SV benchmark set (9,357 SVs) and lifted it over to GRCh38 using the UCSC liftOver tool[61,62]. After excluding SVs that failed the liftover, the benchmark ultimately included 9,228 SVs. We compared our call sets against the benchmark using Truvari (v4.2.2). Based on the performance comparison among the callers, we opted to use the PAV-based SVs supported by at least one additional caller for subsequent analyses.

Mendelian transmission error test

We conducted Mendelian transmission error tests on trio samples using BCFtools with the mendelian2 option. Four trios representing different ancestries, namely, AFR (NA19238, NA19239 and NA19240), AMR (HG00731, HG00732 and HG00733), EAS (HG00512, HG00513 and HG00514) and WAS (APRTrioF, APRTrioM and APRTrioS), were included in the analysis. SVs exhibiting genotypes that violated Mendelian transmission rules across the trio were classified as discordant SVs.

Benchmarking of assembly-based SVs against the HGSVC3 and HPRC Year 1 SVs

Two external SV datasets, namely, the HGSVC3 SVs[20] and the HPRC Year 1 pangenome-derived SVs (decomposed VCF; from Minigraph-Cactus (v1.1) and PGGB (v1.0))[19], were leveraged to benchmark our called SVs. Truvari (v4.2.2) was used for the benchmarking analysis. The analysis was restricted to Dipcall-confident alignment SVs and subsequently stratified into GIAB Easy and Difficult regions (genome stratification (v3.5)). Please note that all comparisons were performed at the haplotype-resolved assembly level for individuals shared across datasets.

VNTR discovery

We identified VNTRs for each assembly. Contrary to the read-depth-based approach typically applied to short-read data[63], the assemblies allowed to use a more straightforward method to identify tandem repeats from long continuous sequences. First, we identified TRRs of GRCh38 using Tandem Repeat Finder (v4.10.0) with recommended parameters (trf 2 5 7 80 10 50 2000 -l 6)[63,64]. Second, we retrieved the FASTA sequences of SVs and INDELs and searched for tandem repeats within these sequences. Subsequently, all tandem repeats were projected onto the reference genome by their physical positions, and only tandem repeats that met the following three criteria were retained: (1) they were located within the reference TRR ± the length of the repeat unit; (2) the repeat period size was the same as that of the TRR; (3) the difference in A, G, C and T content between the tandem repeat and the corresponding TRR was less than 5%. Finally, all passing tandem repeats with a unit length of ≥7 were considered a VNTR allele, and a VNTR locus was defined as a TRR containing at least two alternative alleles.

Merging redundant representations of SV alleles across assemblies

We tested three SV merging methods, namely, Jasmine[28], SVanalyzer[62] and SURVIVOR[65], to merge redundant SVs. We created pseudo-arrays for panel samples using SNP sites from the UKB array and imputed them using a leave-one-out strategy with reference panels constructed by each of the three methods. The imputed SVs were benchmarked against HGSVC3 (Freeze4) SVs to compare the accuracy and reliability of imputation. Based on the performance comparison among the mergers, we opted to use Jasmine for the merging process. For VNTRs, we used SVanalyzer with the options --reldist 0.95 --relsizediff 0 to eliminate highly similar VNTR alleles while preserving allele length diversity across assemblies.

Haplotype reference panels for SV and VNTR imputation

We consolidated all samples into a population-level variant set using BCFtools (merge -m none)[66]. As the individual-level variant call set did not include nonvariant sites (that is, sites where both maternal and paternal alleles matched the reference genome), we inferred such sites by using assembly alignments and Dipcall-confident regions[57]. Variant sites not covered by confident alignments in over half of the haplotypes were excluded, and the remaining missing sites were assumed to align with the reference using BCFtools (--missing-to-ref). We constructed two haplotype reference panels for imputation, one for general SVs (SV panel) and another for VNTRs (VNTR panel). Given the abundance of SNPs for identifying shared haplotype segments between the reference panel and a target sample, we excluded INDELs (length < 50 bp) from the SV panel to prevent a potential decrease in the performance of SV imputation. In the VNTR panel, we retained tandem repeat INDELs of a length of ≥7 bp, as these are considered VNTR alleles. Finally, we converted the panels to MSAV format using Minimac4 (v4.1.6) to expedite subsequent analysis[32].

Benchmarking of the imputed SVs against the GIAB HG002 SVs

We generated pseudo-arrays of HG002 by extracting SNPs from our call set that were present on the UKBA, MEGA and OMNI-2.5M arrays, consisting of 0.8, 1.8 and 2.4 million SNPs, respectively.

Moreover, we created an input set of ~3.2 million SNPs by imputing the UKBA SNP set with the TOPMed reference panel. These SNPs with

their imputed genotypes were then used as another input set (referred to as pre-imputed SNPs) for SV imputation. We also included an input set that contained ~3.4 million genome-wide SNPs to represent those typically called from WGS data (referred to as WGS SNPs). We excluded HG002 haplotypes from the reference panel and imputed the five input sets with the haplotypes of the remaining individuals. The imputation was conducted using Minimac4 (v4.1.6) with default settings[32]. We compared the imputed SVs against two widely used SV benchmarks of HG002, including the GIAB Tier 1 (v0.6)[62] and GIAB CMRG (v1.0)[67] that comprised 9,228 and 250 SVs, respectively. We used Truvari (v4.2.2)[68] for the benchmarking tests. We further conducted a leave-one-out analysis to evaluate the robustness of imputation for SVs not represented in the HG002 benchmark sets (Supplementary Note 13).

Assessing the performance of SV and VNTR imputation at population level

We evaluated the performance of SV imputation at population level using the assembly-based SVs used in this study, UKB srWGS-based SVs[37] and 1KGP moderate-coverage ONT WGS-based SVs[69]. We conducted leave-one-out imputation as described above to exclude individuals or their relative(s) overlapping with the reference panel. We assessed imputation accuracy by calculating the squared correlation between imputed and observed genotypes for each SV locus and VNTR locus. Additionally, we compared the allele frequencies of imputed common SVs (>1 kbp and located in high-mappability regions) in the UKB with those from gnomAD[33] (v4.1; non-Finnish European population) and UKB short-read-based SV datasets[37]. Truvari (v4.2.2)[68] was used to align SVs across datasets.

SV and VNTR imputation in the UKB data

Having seen the performance gain of SV imputation from using imputed SNPs over array SNPs, we performed SV and VNTR imputation in the UKB (application ID 66982) on RAP, using the TOPMed pre-imputed SNP data generated by the UKB team[70]. Variants with an imputation info score of <0.9 or an MAF of $<1\times10^{-4}$ were filtered out, and the remaining variants were lifted over to GRCh38 using the UCSC liftOver tool[61]. The ancestries of UKB individuals were inferred by projecting them onto the 1KGP3 data through principal component projection analysis implemented in GCTA[35,71]. Ultimately, 477,856 individuals were retained and grouped into the following five super-populations: EUR ($n = 456,643$), AFR ($n = 8,362$), AMR ($n = 7,322$), SAS ($n = 3,753$) and EAS ($n = 1,776$). Participants who had withdrawn from the UKB were excluded. We split all variants into chunks of 15,000 markers to save computing memory, with an overlap of 2,500 markers among adjacent chunks. SHAPEIT4 (v4.2) was used for phasing with a high number of iterations to improve accuracy (--mcmc-iterations 10b, 1p, 1b, 1p, 1b, 1p, 1b, 1p, 10 m --pbwt-depth 8)[31]. All phased variant chunks were finally ligated back by matching the phase state at overlapping haplotypes using BCFtools (concat --ligate)[66]. We conducted imputation for each ancestry group in the UKB. The pre-imputed array data of the EUR population were split into 46 subsets (10,000 individuals per subset) to lessen computational load. The imputations were carried out in chunks using Minimac4 (v4.1.6)[32]. All imputed subsets were then merged using BCFtools (merge -m none). The poorly imputed VNTR loci with a squared correlation below 0.1 in our benchmarking analysis, as well as imputed SVs with an MAF of <0.01, an info score of <0.3 or those that violated the Hardy–Weinberg equilibrium (HWE) test ($P < 1\times10^{-6}$), were excluded from the subsequent analysis using PLINK2 (ref. 72). This process finally resulted in 54,578, 81,858, 62,957, 50,347 and 56,545 common SVs in the EUR, AFR, AMR, EAS and SAS populations, respectively.

Estimating variance explained by all common SVs and SGVs

We used GCTA-GREML[34] to estimate the variance in a phenotype explained by all common SVs and SGVs for 652 quantitative traits in up to 349,230 unrelated UKB individuals of European ancestry. Traits with sample sizes smaller than 10,000 were excluded, leaving a total of 636 traits for analysis. Due to the computational constraints of GCTA-GREML, which intensify as the sample size exceeds 50,000, we divided all the unrelated UKB individuals into up to seven subsets. The division depended on the sample size of the trait being examined, with each subset containing approximately 50,000 individuals. Like SVs, SGVs with an MAF of <0.01, an info score of <0.3 or those that failed the HWE test ($P < 1\times10^{-6}$) were filtered out using PLINK2 (ref. 72). In each subset of individuals, we used GCTA to compute an SV-based GRM with 54,578 imputed SVs and an SNP-based GRM with 8,801,428 imputed SGVs. We conducted both a marginal analysis, where the SVs and SGVs were fitted separately in a one-component model, and a joint analysis, where the SVs and SGVs were fitted jointly in a two-component model. In both analyses, we fitted age, age squared, sex, age × sex, age squared × sex and the first 20 principal components provided by the UKB team as fixed-effect covariates. To avoid selection biases, we allowed variance component estimates to be negative, using the flag --reml-no-constrain. We conducted an inverse-variance meta-analysis to combine the estimates of variance explained across all subsets using the formula $-\hat{h}^2_{\text{meta}} = \sum \frac{\hat{h}^2_i}{\text{s.e.}^2_i} / \sum \frac{1}{\text{s.e.}^2_i}$ with $\text{s.e.} = \sqrt{1/\sum \frac{1}{\text{s.e.}^2_i}}$, where $\hat{h}^2_i$ denotes the estimated variance explained in subset i, and s.e.^2_i denotes the corresponding standard error[71]. The standard error of the mean heritability estimates across all traits was calculated following a previous study[73]. We further conducted simulations to quantify variation at causal SVs or SGVs captured by imputed SVs and SGVs (Supplementary Note 14 and Supplementary Table 16).

LD analysis for SVs and SGVs

We selected 2,000 random variants, comprising both imputed SVs and SGVs, from chromosome 20 as target variants for an LD analysis in 10,000 unrelated UKB-EUR individuals. For each target SV or SGV, we identified its LD friends (that is, variants significantly correlated with the target) and calculated the LD score (that is, the sum of LD r^2 between the target variant and all variants in a region) with a window size of 10 Mbp using GCTA-LDF[74] and GCTA-LDS[71,75]. The analyses were performed in four scenarios, depending on the type of variants—a target SV in LD with SVs, a target SV in LD with SGVs, a target SGV in LD with SVs and a target SGV in LD with SGVs.

Genome-wide association analyses of the imputed SVs in the UKB

We performed GWAS for the 54,578 imputed common SVs for 2,624 traits in up to 456,643 UKB individuals of European ancestry. The phenotype processing and GWAS for the imputed SGVs were described elsewhere[35,36]. In brief, we generated 835 Phecode traits using International Classification of Diseases, Tenth Revision (ICD-10) records (UKB fields 41202 and 41204) and the ICD-10 to Phecode map (v.1.2). The remaining traits were derived following the quality control pipeline from the Neale Lab with the PHESANT package[76]. A total of 652 quantitative traits were converted into z scores using the rank-based inverse normal transformation. Multicategorical traits were transformed into binary traits, and any binary traits with the number of cases <100 or a total sample size <5,000 were excluded. GCTA-fastGWA[35] and GCTA-fastGWA-GLMM[36] were used for the association analysis of quantitative and binary traits, respectively, with a sparse GRM derived from SGVs used to control for relatedness among individuals. The covariates fitted in these analyses were the same as those used in the GREML analyses.

COJO and FINEMAP analyses of the SVs and SGVs

We used GCTA-COJO[77] to perform a multiple-regression-based-model-selection analysis of a pooled set of SV and SGV associations to test whether an SV association signal remained significant after accounting for the SGV associations. We conducted this analysis on a chromosome-by-chromosome basis, with a significance threshold of $P < 5\times10^{-8}$ and a window size of 10 Mbp. For ease of computation for GCTA-COJO, only a random subset of 50,000 unrelated UKB-EUR individuals was used as the LD reference. Moreover, we used FINEMAP

(v1.4.1)[78] to perform a fine-mapping analysis of a pooled set of SV and SGV associations. We set a window size of 1 Mbp for each locus centered around a lead SV, with the LD matrix for each locus computed from all unrelated UKB-EUR individuals.

eQTLs analysis for SVs in the GTEx

We performed SV imputation using SNPs called from srWGS data of 694 unrelated individuals of European ancestry in the GTEx (v8) project[39] and associated the SVs with gene expression levels (referred to as SV-eQTL hereafter) in 49 tissues (n = up to 588). Similar to the analysis in the UKB, imputed SVs with an MAF of <0.01, an info score <0.3 or those that failed the HWE test ($P < 1 \times 10^{-6}$) were excluded. The normalized gene expression data, along with the covariate matrices, were obtained from the GTEx portal (https://www.gtexportal.org/home/datasets)[39]. The SV-eQTL mapping was performed using the OSCA tool with the default settings[40]. The GENCODE[79] Gene Set (v43) was used for annotations.

Functional enrichment analyses of trait-associated SVs

We tested the enrichment of trait-associated SVs in various functional regions, including histone modifications, chromatin states, RBP binding sites and TAD boundaries, using logistic regression. Comprehensive details regarding the acquisition, filtering and processing of these functional annotation datasets are provided in Supplementary Note 15. Here we defined the outcome variable as the overlap states (coded as 0 or 1) between SVs and each type of genomic region above. A total of 54,578 SVs were classified into 'trait-associated' and 'other SVs' groups based on a threshold of P value of $<5 \times 10^{-8}$. The MAF, length, type, info score and distance to the nearest TSS and SS of each SV were included as covariates in the logistic regression model. We pruned SVs for LD with a window size of 500 kbp and an LD r^2 threshold of 0.1 using PLINK2, to avoid the inclusion of SVs in substantial LD in the analysis. All datasets used in the analysis were based on GRCh38 coordinates. The P-value threshold after correcting for multiple tests was $0.05/(26 \times 7)$ for histone modifications, $0.05/(18 \times 71)$ for chromatin states, 0.05/38 for TAD boundaries and 0.05/2 for RBP binding sites.

VNTR association analyses in the UKB and GTEx data

SV imputations in the UKB and GTEx datasets are described above. The same input data were used for VNTR imputation, using the VNTR panel. A VNTR locus was defined as having at least two nonreference alleles. After prior work[63], we computed the length of a VNTR for each individual based on the imputed genotype dosage, which reflects imputation uncertainty, using the following formula: $G_{\text{vntr}} = \sum v_i n_i g_i$, where v_i is an indicator variable to indicate whether allele i is an insertion or deletion (taking a value of 1 or −1, respectively), as we normalized the VNTR lengths according to the reference genome GRCh38, n_i represents the copy number of the repeat unit of allele i and g_i is the imputed dosage score of allele i. We performed association analyses for 2,624 complex traits in the UKB (referred to as VNTR-GWAS analyses) and gene expression traits across multiple tissues in the GTEx data (referred to as VNTR-eQTL analyses), using a linear regression model for quantitative or categorical traits and a logistic regression model for binary traits in R, with the same covariates as used in the SV association analyses above. Only the unrelated individuals of European ancestry (n = up to 349,660 in the UKB; n = up to 588 in the GTEx) were included in the analysis. A trait-associated VNTR or VNTR-eQTL with a P value $<5 \times 10^{-8}$ was considered significant.

Reporting summary

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

Data availability

The assembly-level SVs and reference panels constructed in this study can be accessed at https://yanglab.westlake.edu.cn/ImputeSV/resource. The individual-level genotype and phenotype data are available upon formal application to the UKB (http://www.ukbiobank.ac.uk). The SV-GWAS and VNTR-GWAS summary statistics generated in this study are available at https://yanglab.westlake.edu.cn/data/ukb_sv_gwas. The 29 1KGP3-EUR assemblies generated in this study are available at https://yanglab.westlake.edu.cn/ImputeSV/resource. The HPRCY1 assemblies are available at https://github.com/human-pangenomics/HPP_Year1_Data_Freeze_v1.0. The HGSVC3 assemblies are available at https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/working/. The CPC assemblies are available at https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA011422. The APR assemblies are available at https://www.mbru.ac.ae/the-arab-pangenome-reference/. The CN1 assemblies are available at https://genome.zju.edu.cn/Downloads. The YAO assemblies are available at https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA017932. The GIAB Tier 1 SV benchmark set for HG002 (v0.6) is available at dbVar (accession nstd175) and at https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/NIST_SV_v0.6/. The GIAB genome stratification (v3.5) is available at https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/. The population-level HGSVC3 Freeze4 SV sets are available at https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/working/20240415_Freeze4. The assembly-level HGSVC3 Freeze4 SV sets are available at https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/working/20240307_PAV_VCF/. The HPRCY1 pangenome-derived SV sets are available at https://github.com/human-pangenomics/hpp_pangenome_resources. SVs called from the moderate-coverage ONT data of 888 1KGP3 individuals are available at http://1kgp-sv-imputation.s3-website-eu-west-1.amazonaws.com/?prefix=panel/. SVs called from high-coverage ONT data of 100 1KGP3 individuals are available at https://s3.amazonaws.com/1000g-ont/index.html?prefix=Gustafson_etal_2024_preprint_SUPPLEMENTAL/. The gnomAD SV set (v4.1) is available at https://gnomad.broadinstitute.org/downloads. The 1000 Genomes Project Phase 3 genotypes are available at https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20201028_3202_phased/. The GTEx (v8) data are available at https://www.gtexportal.org/home/datasets. The histone ChIP-seq peak and RBP eCLIP peak data are available through the ENCODE portal (https://www.encodeproject.org). The 18-state Roadmap model annotation data are available at EpiMap (https://compbio.mit.edu/epimap/). The TAD coordinate data are available at the 3D Genome Browser (https://3dgenome.fsm.northwestern.edu/datasets). The deCODE SV-GWAS summary data are available at https://www.decode.com/summarydata/. The genome partition of LD-independent blocks is available at https://bitbucket.org/nygcresearch/ldetect-data/.

Code availability

The online SV Imputation Server developed in this study can be accessed at https://yanglab.westlake.edu.cn/ImputeSV. The scripts used for reference panel construction and data analysis can be accessed at the Zenodo repository (https://doi.org/10.5281/zenodo.15825718).

References

53. Nassir, N. et al. A draft UAE-based Arab pangenome reference. *Nat. Commun.* **16**, 6747 (2025).
54. Gao, Y. et al. A pangenome reference of 36 Chinese populations. *Nature* **619**, 112–121 (2023).
55. Yang, C. et al. The complete and fully-phased diploid genome of a male Han Chinese. *Cell Res.* **33**, 745–761 (2023).
56. He, Y. et al. T2T-YAO: a telomere-to-telomere assembled diploid reference genome for Han Chinese. *Genomics Proteomics Bioinformatics* **21**, 1085–1100 (2023).
57. Li, H. et al. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat. Methods* **15**, 595–597 (2018).

58. Heller, D. & Vingron, M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* **36**, 5519–5521 (2021).
59. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
60. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
61. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
62. Zook, J. M. et al. A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* **38**, 1347–1355 (2020).
63. Mukamel, R. E. et al. Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. *Science* **373**, 1499–1505 (2021).
64. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
65. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
66. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
67. Olson, N. D. et al. PrecisionFDA Truth Challenge V2: calling variants from short and long reads in difficult-to-map regions. *Cell Genom.* **2**, 100129 (2022).
68. English, A. C., Menon, V. K., Gibbs, R. A., Metcalf, G. A. & Sedlazeck, F. J. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biol.* **23**, 271 (2022).
69. Noyvert, B. et al. Imputation of structural variants using a multi-ancestry long-read sequencing panel enables identification of disease associations. Preprint at *medRxiv* https://doi.org/10.1101/2023.12.20.23300308 (2025).
70. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
71. 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).
72. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
73. Zhu, Z. et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* **9**, 224 (2018).
74. Yang, J. et al. Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807–812 (2011).
75. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
76. Millard, L. A. C., Davies, N. M., Gaunt, T. R., Davey Smith, G. & Tilling, K. Software application profile: PHESANT: a tool for performing automated phenome scans in UK Biobank. *Int. J. Epidemiol.* **47**, 29–35 (2018).
77. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).
78. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
79. Frankish, A. et al. GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).

Acknowledgements

The authors thank Y. Wang and X. Sun for helpful discussions. This research was partly supported by the National Natural Science Foundation of China (U23A20165 to J.Y.), the National Key R&D Program of China (2024YFC3405800 and 2024YFC3405802 to J.Y.), the 'Pioneer' and 'Leading Goose' R&D Program of Zhejiang (2024SSYS0032 to J.Y.) and the New Cornerstone Science Foundation (to J.Y.). The authors thank the Westlake University High-Performance Computing Center, the UK Biobank Research Analysis Platform and the TOPMed imputation server for their computational support. This research has been conducted using the UK Biobank Resource under application 66982. The authors also gratefully acknowledge the use of publicly available assemblies and data from the HPRC, HGSVC, CPC and APR consortia; the T2T-CN1 and T2T-YAO projects; and the 1000 Genomes Project. The authors thank all reviewers for their valuable comments and constructive suggestions, which have substantially improved the quality and rigor of this study.

Author contributions

J.Y. conceived and directed the study. W.-Y.B. and J.Y. designed the experiments and developed the analytical pipelines with input from S.L., Z.D. and T.Q. W.-Y.B., J.-J.Y. and J.H. developed the online tool and the data portal under the guidance of N.L. and J.Y. W.-Y.B. performed the data analyses and simulations with assistance and guidance from S.L., T.Q. and J.Y. W.-Y.B., Z.D., J.C., L.W. and J.Y. contributed to the generation of the HiFi data for the 29 1KGP3 individuals. W.-Y.B. and J.Y. wrote the paper with contributions from T.Q. All authors reviewed and approved the final paper.

Competing interests

All authors declare no competing interests.

Additional information

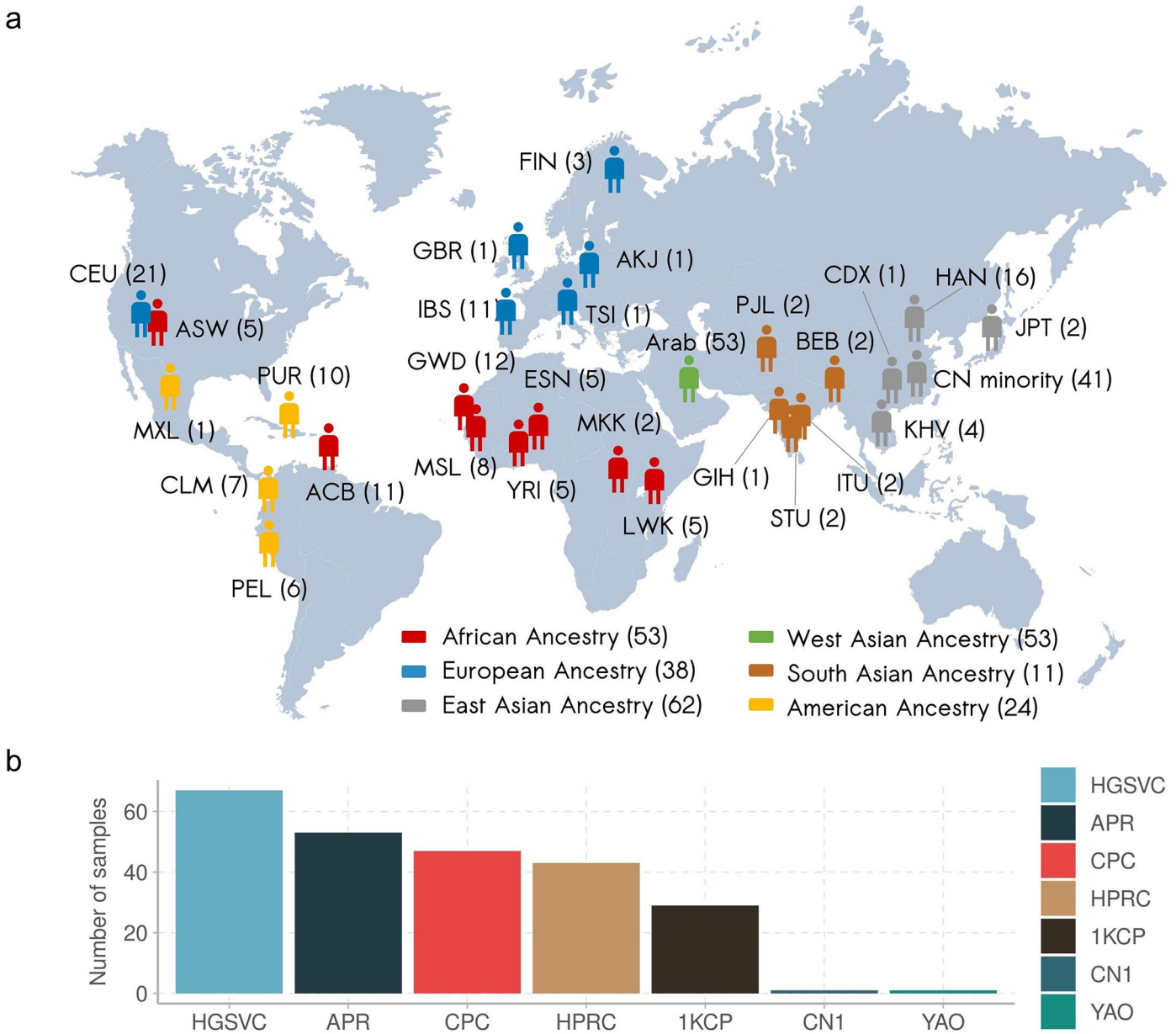
Extended data is available for this paper at https://doi.org/10.1038/s41588-026-02612-z.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-026-02612-z.

Correspondence and requests for materials should be addressed to Jian Yang.

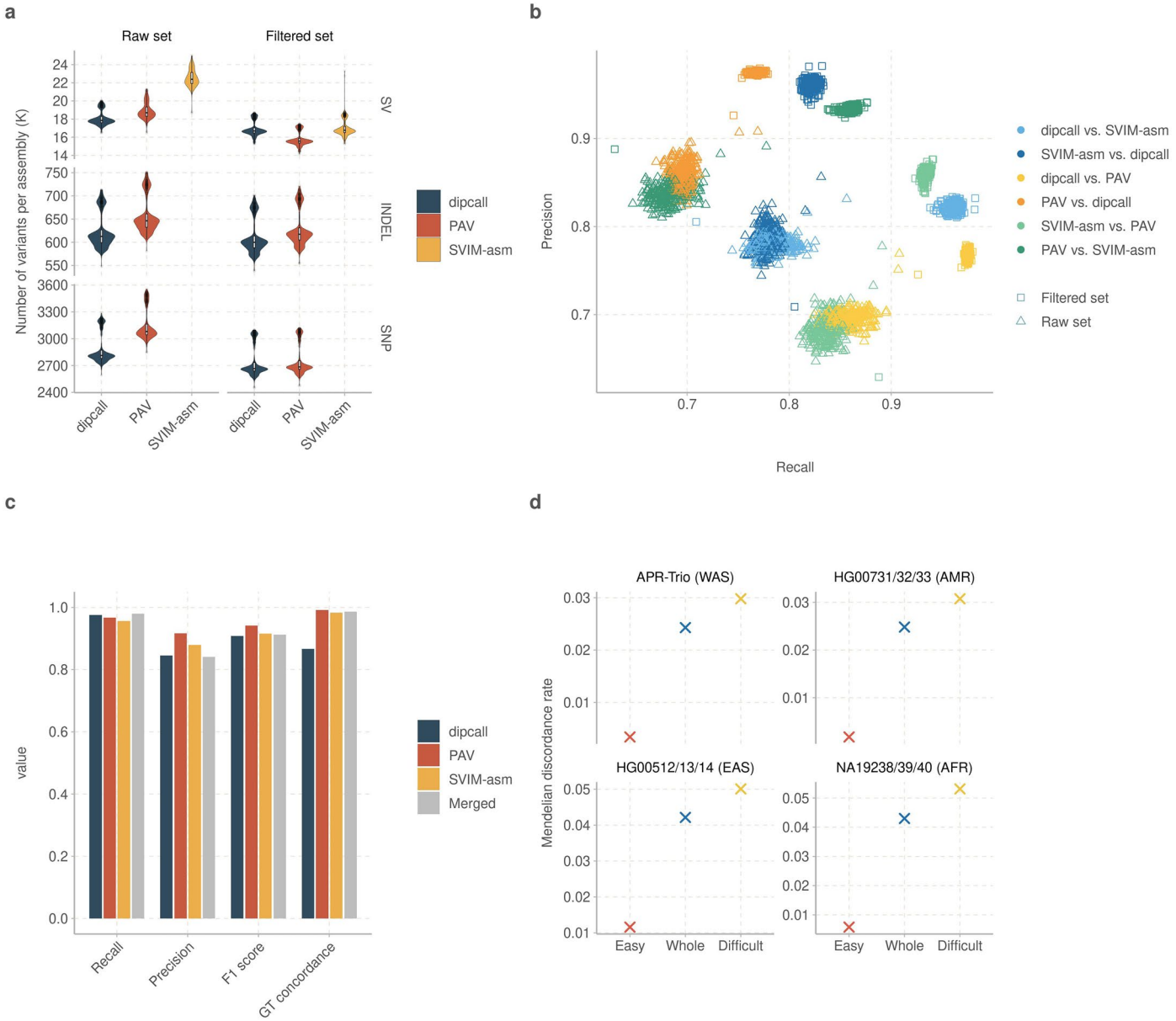
Peer review information *Nature Genetics* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41588-026-02612-z



Extended Data Fig. 1 | Geographical distribution of the individuals with high-coverage HiFi data included in our SV imputation panel. a,**b**, A total of 482 high-quality assemblies from 241 individuals are represented, with their diverse ancestries detailed in **a** and their source cohorts shown in **b**. In **a**, the colors correspond to different ancestry groups, and the sample size within each group is indicated in parentheses. In **b**, the colors denote the source cohort. The population codes are as follows: ACB: African Caribbean; AKJ: Ashkenazi Jewish; ASW: African Ancestry in Southwest USA; BEB: Bengali; CDX: Chinese Dai; CEU: CEPH; CN minority: Chinese minority; CLM: Colombian; ESN: Esan; FIN: Finnish; GBR: British; GIH: Gujarati; GWD: Gambian Mandinka; HAN: Han Chinese; IBS: Iberian; ITU: Indian Telugu; JPT: Japanese; KHV: Kinh Vietnamese; LWK: Luhya; MSL: Mende; MXL: Mexican Ancestry; PEL: Peruvian; PJL: Punjabi; PUR: Puerto Rican; STU: Sri Lankan Tamil; TSI: Toscani; YRI: Yoruba. Image provided by Servier Medical Art (https://smart.servier.com/), licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.1038/s41588-026-02612-z

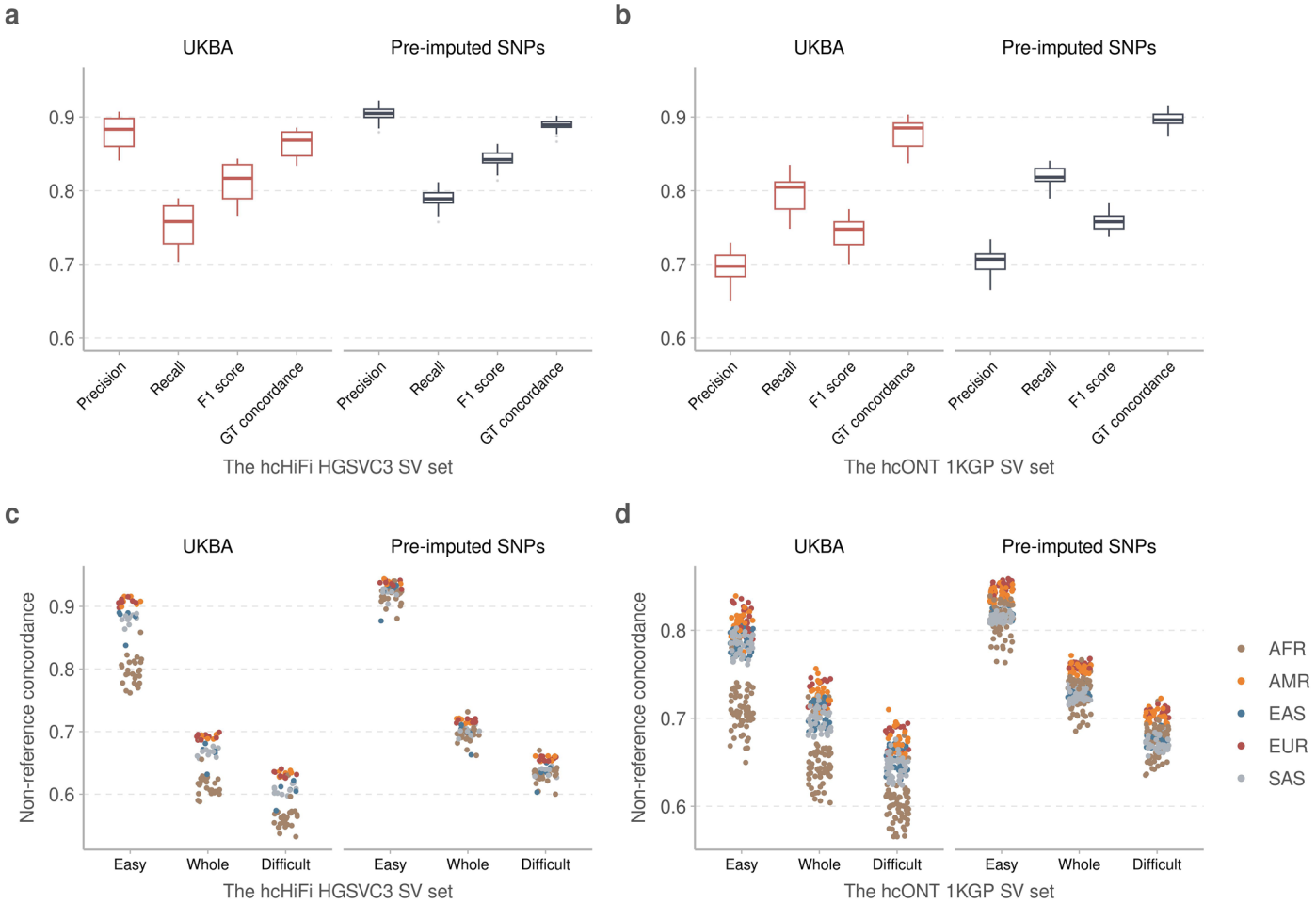


Extended Data Fig. 2 | SVs called from 482 genome assemblies. a, Distribution of the number of variants called by Dipcall, SVIM-asm, and PAV. The numbers of called SNPs, INDELs, and SVs (rows) before and after filtering (columns) are shown. In each violin plot, the centerline indicates the median, the box represents the interquartile range (IQR; Q1–Q3), and whiskers extend to 1.5× the IQR; values beyond this range are shown as outliers. The sample size is 482. SNPs and INDELs called by SVIM-asm are not shown, as this caller is currently not applicable for SGVs. **b**, Reciprocal benchmarking tests across the three callers. The benchmarking tests were performed for each combination of the three SV callers using Truvari (v4.2.2) with options --dup-to-ins -C 1000 -O 0.0 -p 0.0 -P 0.3 -s 50 -S 15. The triangle and square represent the SV sets before and after filtering, respectively, for each sample. **c**, Benchmarking test for called SVs against the HG002 (v0.6) Tier 1 SVs from the GIAB. The benchmark included 9,228 well-established SVs. **d**, Mendelian transmission analysis of SVs in family trios. Mendelian discordance rates were calculated for SV calls in trios from four different ancestry groups using BCFtools +mendelian2. The analysis was restricted to SVs in Dipcall-confident regions and subsequently stratified into 'easy' and 'difficult' regions using the GIAB (v3.5) genome stratification.

https://doi.org/10.1038/s41588-026-02612-z

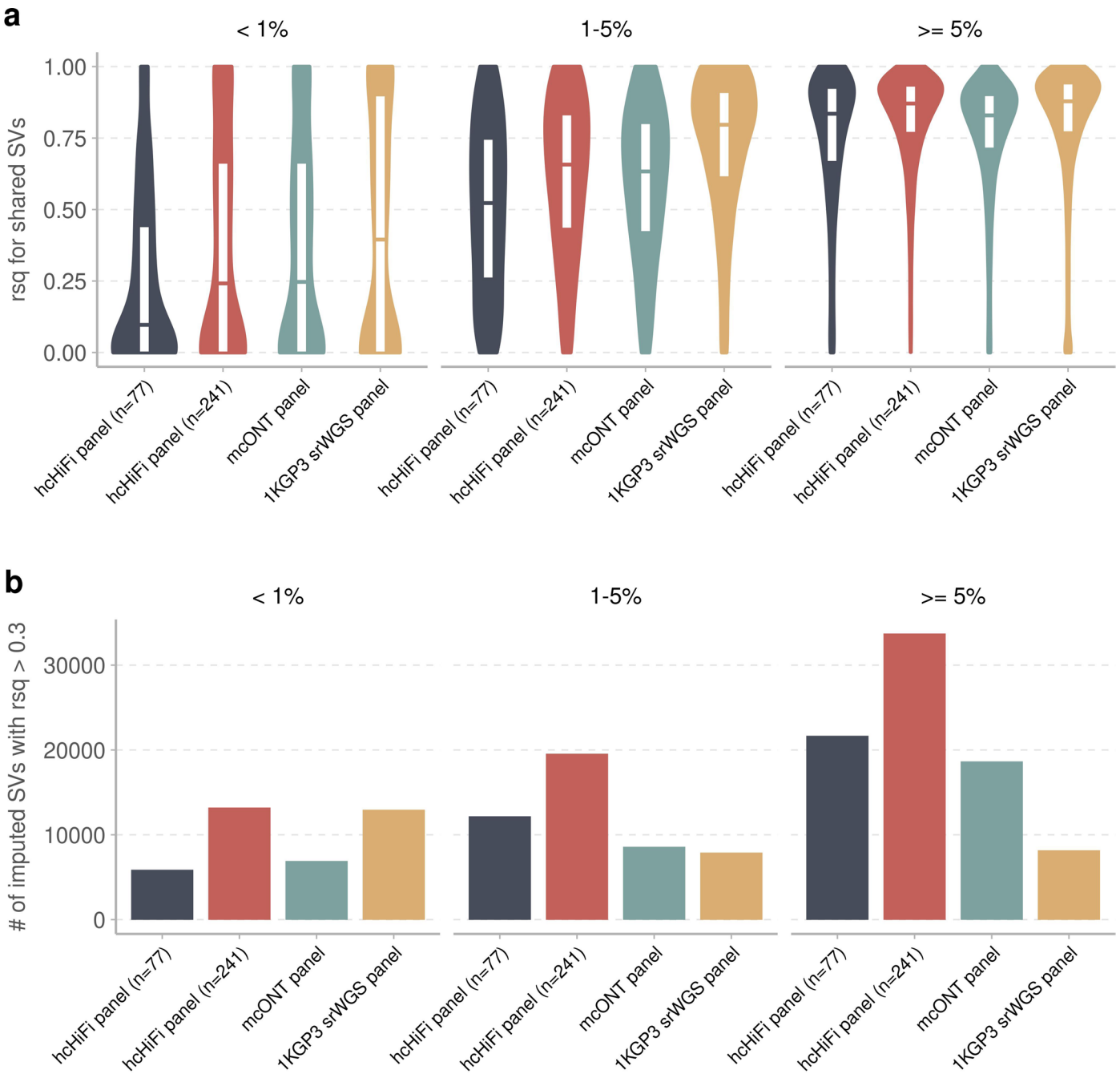
Extended Data Fig. 3 | Benchmarking of imputed SVs under various scenarios. **a**, Impact of reference panel size and ancestry composition on SV imputation accuracy. Benchmarking was performed for 65 individuals against the HGSVC3 SV truth set. Leave-one-out imputation was performed for HGSVC3 samples using the TOPMed pre-imputed array as the input set. The three reference panels tested were: (1) a matched panel, which comprised samples of matched ancestry to the target samples (38 EUR, 62 EAS, 53 AFR, 53 MEA, 24 AMR, and 11 SAS); (2) a half-mixed panel, which included samples of matched ancestry to the target samples and an equal sample size of other ancestries; and (3) a maximum panel, which consisted of all available samples with various ancestries. Truvari (v4.2.2) was used for benchmarking with the parameters: --dup-to-ins -C 1000 -O 0.0 -p 0.0 -P 0.3 -s 50 -S 15. The box plots display the medians, upper quartiles, and lower quartiles; whiskers extending to 1.5× IQR; outliers are defined as ±1.5× IQR. **b**, Benchmarking the performance of our hcHiFi SV imputation reference panels using lenient (--dup-to-ins -C 1000 -O 0.0 -p 0.0 -P 0.3 -s 50 -S 15) and stringent (--dup-to-ins -C 1000 -O 0.0 -p 0.5 -P 0.5 -s 50 -S 15) parameters. Each dot represents an individual.

https://doi.org/10.1038/s41588-026-02612-z



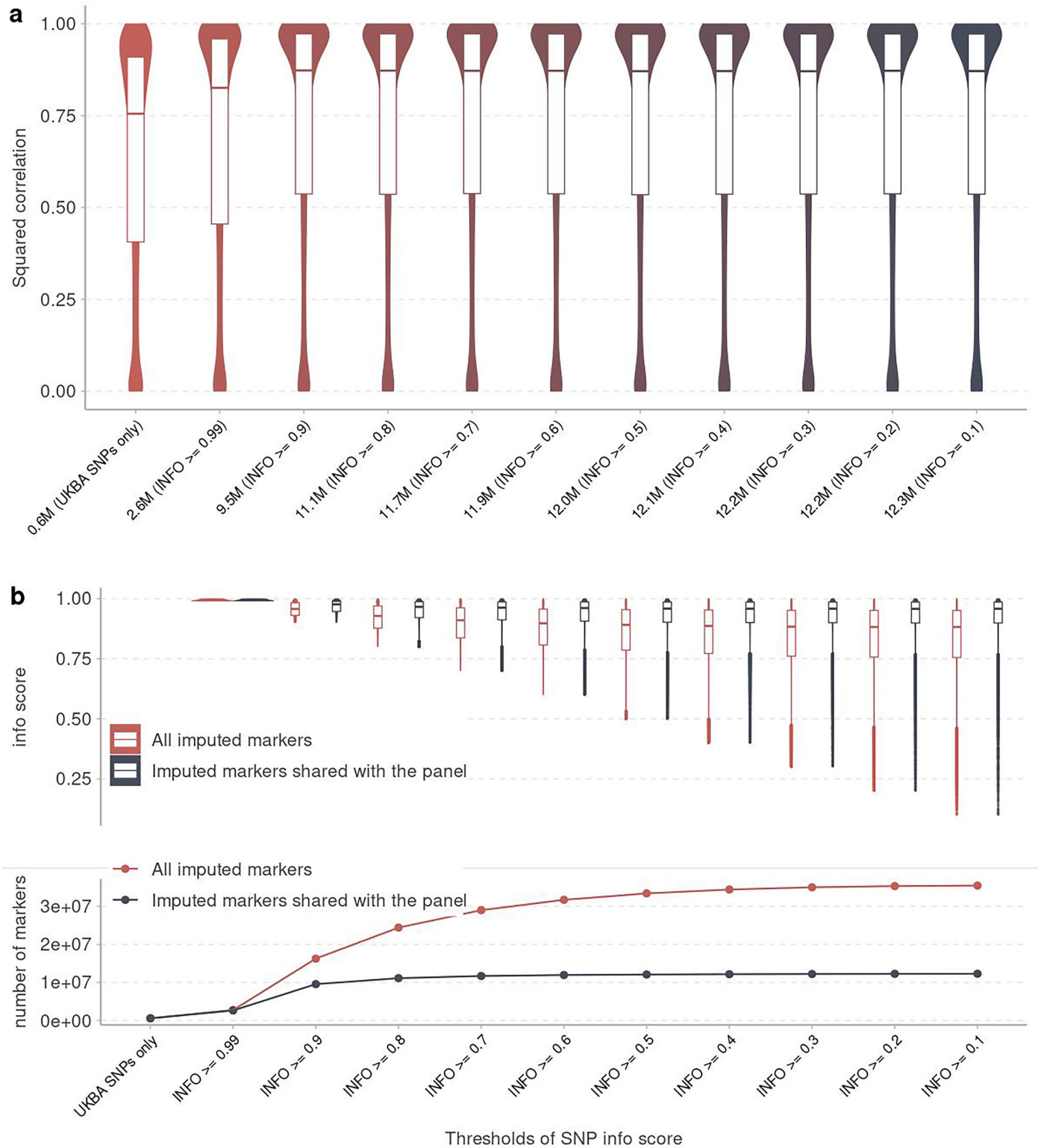
Extended Data Fig. 4 | Benchmarking of imputed SVs against external SV sets. a–**d**, Leave-one-out SV imputation was performed for samples in our reference panel that also appear in the HGSVC3 high-coverage HiFi (hcHiFi; **a** and **c**) or 1KGP3 high-coverage ONT (hcONT; **b** and **d**) cohort, using either the UKB-array SNPs or the TOPMed pre-imputed SNPs as input. The corresponding called SV sets (HGSVC3 hcHiFi and 1KGP3 hcONT) were used as the ground truth for their respective cohorts. Truvari (v4.2.2) was used for the analysis with the following options: --dup-to-ins -C 1000 -O 0.0 -p 0.0 -P 0.3 -s 50 -S 15. In **a** and **b**, the genotype concordance rate was computed for shared SVs between the imputed set and the called set for each benchmarking individual. In **c** and **d**, the nonreference concordance rate was calculated for SVs that appeared in all the benchmarking individuals. In **a** (n = 65) and **b** (n = 100), the box plots represent the median (centerline), IQR (box from Q1 to Q3), and whiskers extending to 1.5× IQR; values outside this range are plotted as outliers.

https://doi.org/10.1038/s41588-026-02612-z



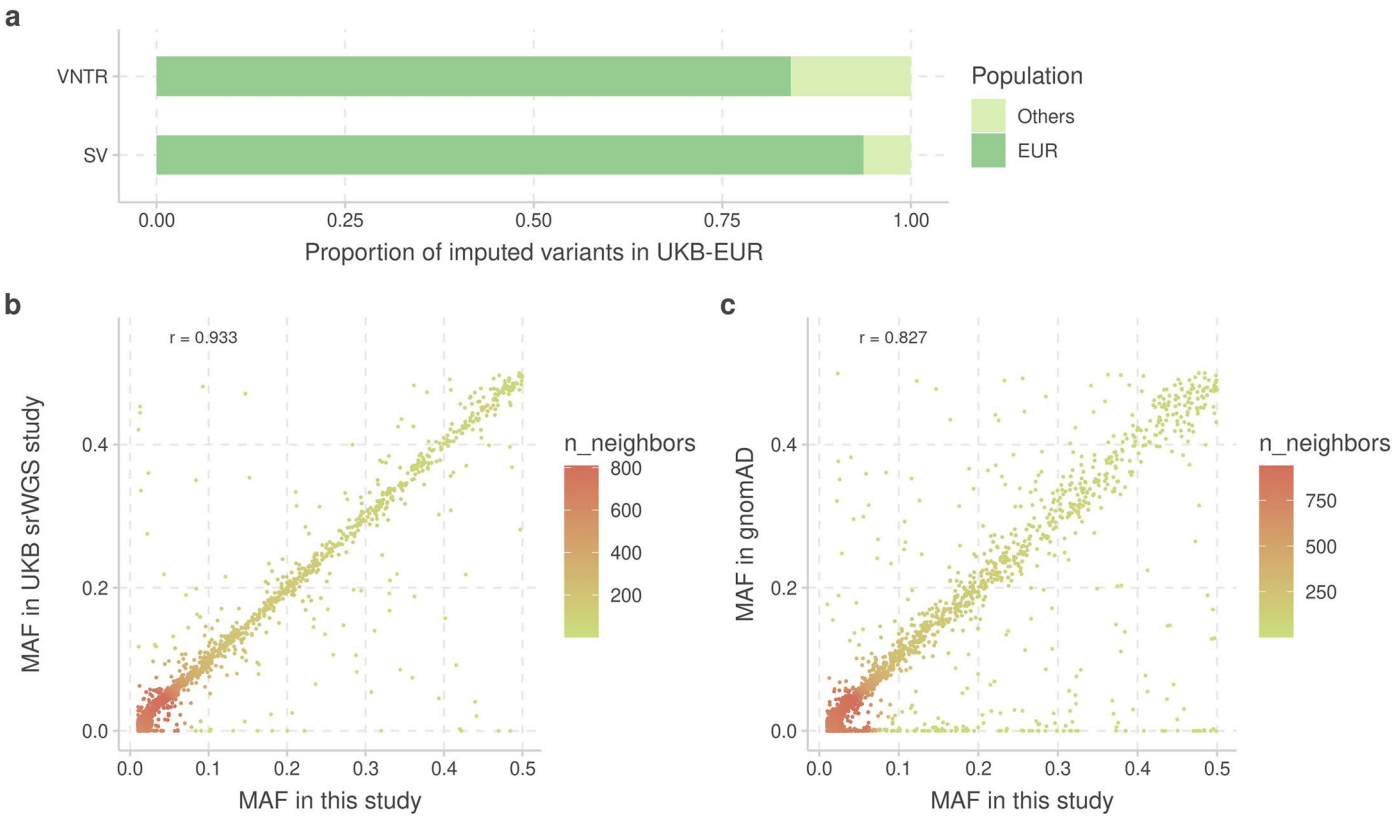
Extended Data Fig. 5 | Squared correlation between imputed and called genotypes of SVs using different reference panels. SVs called from the HiFi assemblies of 241 samples were used as the ground truth. We employed a leave-one-out imputation strategy, excluding the individual to be imputed (and their close relatives, if any) from the reference panel to avoid potential bias. The hcHiFi panels include two versions: one built from 77 samples (denoted as 'n = 77'), and another built from the full set of 241 samples (denoted as 'n = 241'). **a**, Squared correlation for SVs shared across all panels. SVs were grouped into three minor allele frequency (MAF) bins: <1% (m = 4,149), 1–5% (m = 6,075), and ≥5% (m = 7,876). Truvari (v4.2.2) with the options --dup-to-ins -C 1000 -O 0.0 -p 0.0 -P 0.3 -s 50 -S 15 was used to align SVs across datasets. **b**, The total number of imputed SVs with a squared correlation greater than 0.3 using different reference panels. All imputed SVs were considered for each panel, regardless of their presence in other panels. The box plots display the medians, upper quartiles, and lower quartiles; outliers are defined as ±1.5× interquartile range (IQR).

https://doi.org/10.1038/s41588-026-02612-z



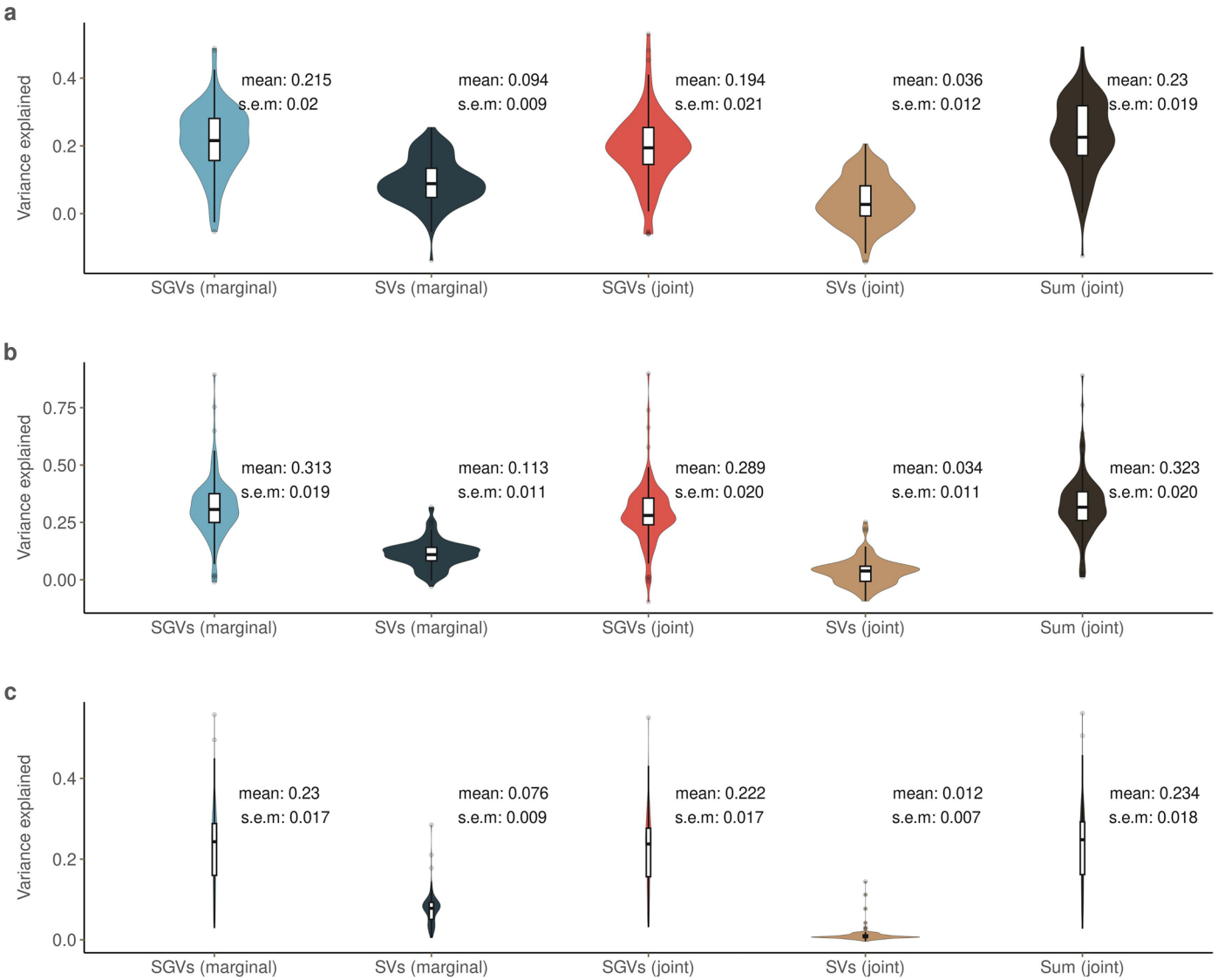
Extended Data Fig. 6 | The impact of the quality and density of input SNPs on SV imputation. a, Squared correlation (r^2) between imputed and srWGS-called SV genotypes in the UKB. SV imputation was performed in 5,000 unrelated UKB participants of European ancestry using TOPMed-imputed SNPs filtered at various imputation info-score thresholds. The x-axis corresponds to the number of input SNP markers remaining after each filtering step (M: million), and the y-axis shows the resulting median r^2 for common SVs (MAF ≥ 0.01). SVs called from srWGS data were used as the ground truth. Truvari (v4.2.2) with the options --dup-to-ins -C 1000 -O 0.0 -p 0.0 -P 0.3 -s 50 -S 15 was used to match imputed and called SVs. **b**, Quality and quantity of input SNPs used for imputation. The top panel displays the distribution of imputation info scores (box plots) for the TOPMed-imputed SNPs that overlap with the SV reference panel at each filtering threshold. The bottom panel shows the total number of these overlapping SNPs for each corresponding threshold. The box plots display the medians, upper quartiles, and lower quartiles; whiskers extending to 1.5× IQR; outliers are defined as ±1.5× IQR.

https://doi.org/10.1038/s41588-026-02612-z



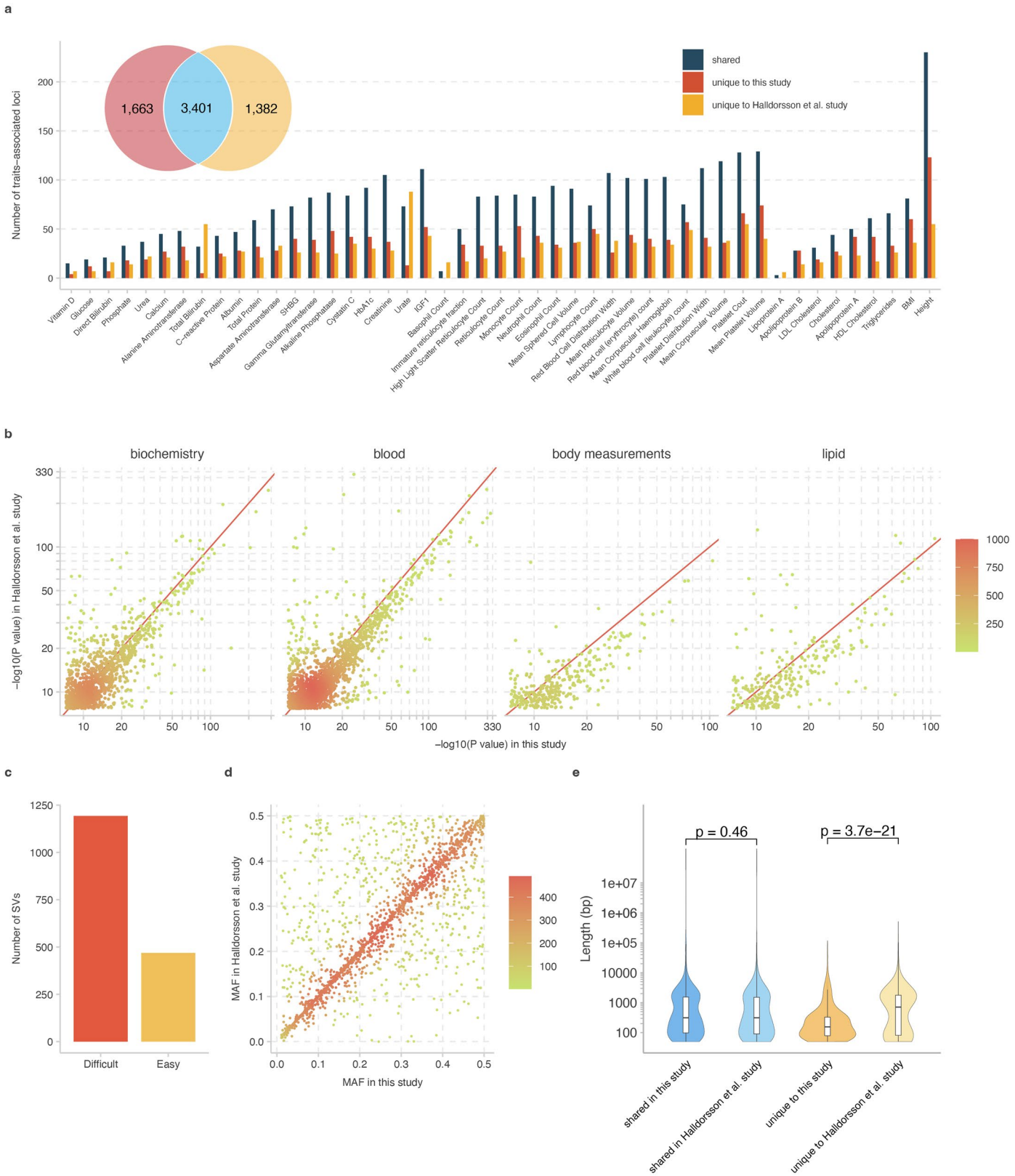
Extended Data Fig. 7 | Common SVs imputed in UK Biobank. a, Proportion of imputed common SVs and VNTRs in UKB-EUR shared with assembly-based called SVs from the EUR population. **b**,**c**, Comparison of minor allele frequency between imputed and short-read-based called SVs. Two external SV sets, UKB short-read WGS SVs (**b**) and gnomAD SVs (**c**), were used for comparison. Each dot represents a single imputed SV with a length of >1 kbp and in high mappability regions (GIAB genome stratification v3.5). The Pearson correlation (r) of minor allele frequencies is labeled in each plot.

https://doi.org/10.1038/s41588-026-02612-z



Extended Data Fig. 8 | Distribution of the estimates of variance explained by SVs and SGVs in three ancestry groups. a–**c**, The heritability estimation analyses were conducted in the AMR (**a**), AFR (**b**), and EUR (**c**) ancestry groups for 111 shared quantitative traits with a sample size of over 3,000 using GCTA-GREML. The x-axis indicates the variance component (SGVs or SVs), fitted in a marginal or joint model. The y-axis indicates the estimated variance explained. The median, upper quartiles, lower quartiles and whiskers (extending to 1.5× the IQR) of the estimates are shown in each violin plot (n = 111), with outliers defined as ±1.5× IQR.

https://doi.org/10.1038/s41588-026-02612-z

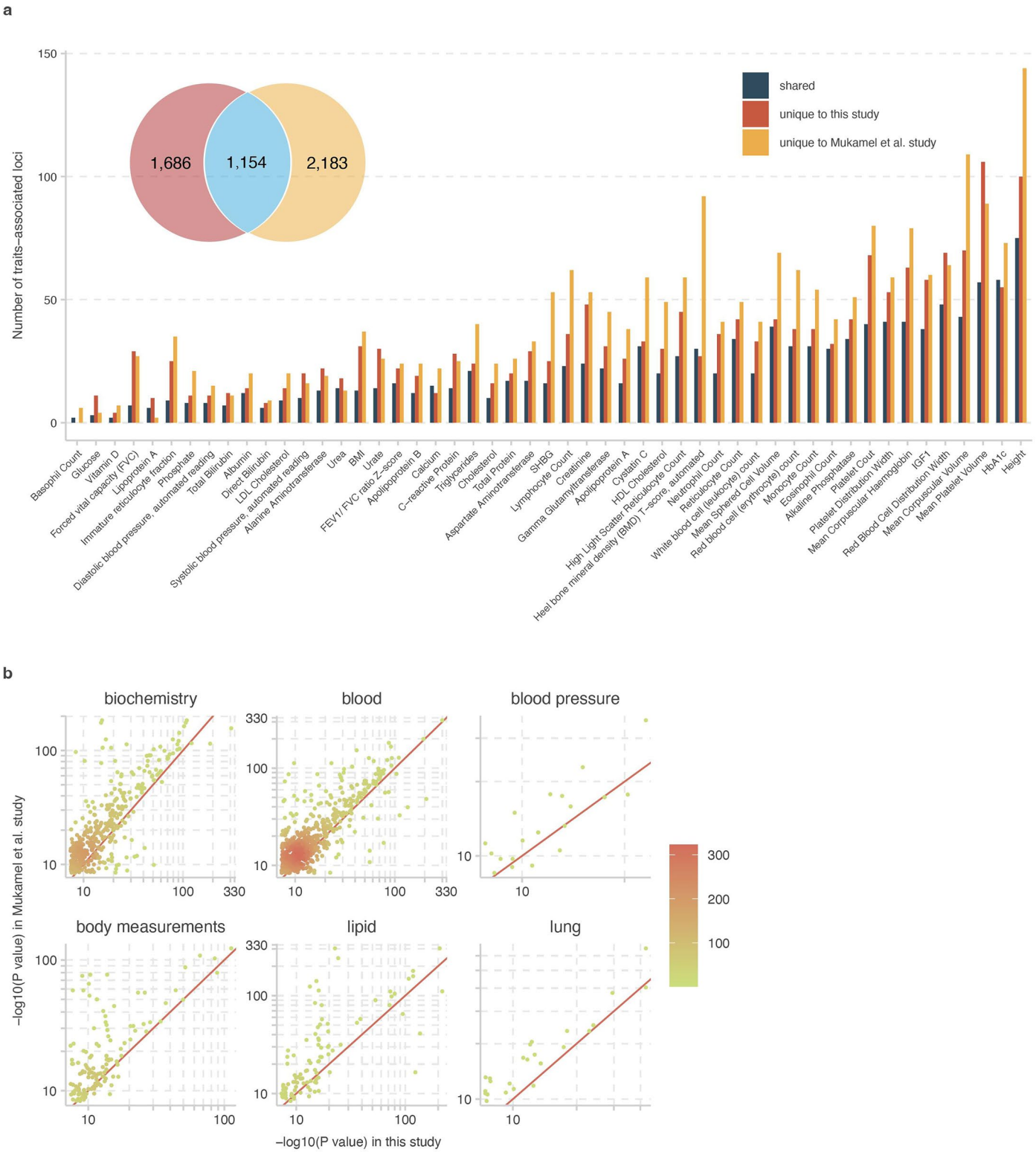


Extended Data Fig. 9 | See next page for caption.

https://doi.org/10.1038/s41588-026-02612-z

Extended Data Fig. 9 | Comparison of SV associations between ref. 37 and this study. To facilitate the comparison between SVs called from short-read (ref. 37) and long-read (this study) WGS data, we used the genome partition of quasi-independent LD blocks to define independent SV associations. **a**, Numbers of study-specific and shared SV associations for 47 quantitative traits analyzed in both studies. **b**, Scatterplots comparing the P values of SV associations shared between the two studies. The traits were categorized into four groups. Each dot in the plots represents an SV association. The x-axis represents our study (sample size up to 456,643), while the y-axis represents ref. 37 (sample size up to 430,136). The color bar in the legend represents dot density. The red straight line in each plot is the diagonal line. **c**, Numbers of SV associations unique to this study in confident and complex genomic regions. These regions were defined using the GIAB genome stratification (v3.5). **d**,**e**, Scatter plots (**d**) and violin plots (**e**) comparing the minor allele frequency and length of trait-associated SVs between this study and ref. 37. The numbers of SVs shared, unique to this study, and unique to ref. 37 are 1,588, 1,045, and 768, respectively. P-values from a two-sided Wilcoxon rank-sum test are labeled. Each box plot displays the median, upper quartile, and lower quartile; whiskers extending to 1.5× the IQR; outliers are defined as ±1.5× IQR.

https://doi.org/10.1038/s41588-026-02612-z



Extended Data Fig. 10 | Comparison of VNTR associations between ref. 8 and this study. a, Number of study-specific and shared VNTR associations for 51 quantitative traits analyzed in both in ref. 8 and this study. **b**, Scatterplot for P values of VNTR associations shared between the two studies. The traits were classified into six groups. Each dot represents a VNTR association. The x-axis and y-axis represent this study and ref. 8, respectively. The color bar in the legend represents dot density, and a diagonal line is shown for each plot.

nature portfolio

Corresponding author(s): Jian Yang

Last updated by author(s): Mar 25, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection beyond standard accession from the public repositories listed in the Data Availability Statement.
Data analysis	The online SV imputation tool developed in this study is available at https://yanglab.westlake.edu.cn/impute_sv. The scripts utilized for reference panel construction and data analysis can be accessed at the Zenodo repository: https://zenodo.org/records/15825718. Other software or packages used in this study are summarized as follows. HiFiAdapterFilt (v2.0.0): https://github.com/sheinasim-USDA/HiFiAdapterFilt Hifiasm (v0.19.8): https://github.com/chhylp123/hifiasm QUAST (v5.0.2): https://github.com/ablab/quast Merqury (v1.3): https://github.com/marbl/merqury BUSCO (v5.2.2): https://busco.ezlab.org dipcall (v0.3): https://github.com/lh3/dipcall SVIM-asm (v1.0.2): https://github.com/eldariont/svim-asm PAV (v2.2.4.1): https://github.com/EichlerLab/pav minimap2 (v2.22): https://github.com/lh3/minimap2 unimap: https://github.com/lh3/unimap SAMtools (v1.14): https://github.com/samtools/samtools BCFtools (v1.14): https://github.com/samtools/bcftools SVanalyzer: https://github.com/nhansen/SVanalyzer truvari (v3.4.0): https://github.com/ACEnglish/truvari AnnotSV (v3.4): https://github.com/lgmgeo/AnnotSV

UCSC liftOver tool: https://genome.ucsc.edu/cgi-bin/hgLiftOver
Tandem Repeat Finder (v4.10.0): https://github.com/Benson-Genomics-Lab/TRF
WhatsHap (v1.4): https://github.com/whatshap/whatshap
BEDtools (v2.29.1): https://github.com/arq5x/bedtools2
GCTA (v1.93.3): https://yanglab.westlake.edu.cn/software/gcta/#Overview
MINIMAC3 (v2.0.1): https://github.com/Santy-8128/Minimac3
SHAPEIT4 (v4.2): https://odelaneau.github.io/shapeit4/
MINIMAC4 (v1.0.2): https://github.com/statgen/Minimac4
GCTA-fastGWA: https://yanglab.westlake.edu.cn/software/gcta/index.html#fastGWA-GLMM
GCTA-fastGWA-GLMM: https://yanglab.westlake.edu.cn/software/gcta/index.html#fastGWA-GLMM
GCTA-COJO: https://yanglab.westlake.edu.cn/software/gcta/index.html#COJO
PLINK (v2.0): https://www.cog-genomics.org/plink
GCTA-GREML: https://yanglab.westlake.edu.cn/software/gcta/index.html#GREML
FINEMAP (v1.4.1): http://www.christianbenner.com/
WashU Epigenome Browser: https://epigenomegateway.wustl.edu/
region-plot: https://github.com/pgxcentre/region-plot
Cytoscape: https://cytoscape.org/
OSCA: https://yanglab.westlake.edu.cn/software/osca/#Overview
TOPMed Imputation Server: https://imputation.biodatacatalyst.nhlbi.nih.gov/#!pages/home
R (v3.6.3): https://www.r-project.org
ggplot2 (v3.5.1): https://github.com/tidyverse/ggplot2

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data used and generated in this study are summarized as follows:

The assembly-level SVs and reference panels constructed in this study can be accessed at https://yanglab.westlake.edu.cn/impute_sv/resource.
The individual-level genotype and phenotype data are available upon formal application to the UKB (http://www.ukbiobank.ac.uk).
The SV- and VNTR-GWAS summary statistics generated in this study are available at https://yanglab.westlake.edu.cn/data/ukb_sv_gwas.
The HPRCY1 assemblies are available at https://github.com/human-pangenomics/HPP_Year1_Data_Freeze_v1.0.
The HGSVC3 assemblies are available at https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/working/.
The CPC assemblies are available at https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA011422.
The APR assemblies are available at https://www.mbru.ac.ae/the-arab-pangenome-reference/.
The CN1 assemblies are available at https://genome.zju.edu.cn/Downloads.
The YAO assemblies are available at https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA017932.
The GIAB Tier 1 SV benchmark set for HG002 (v0.6) is available at dbVar (accession: nstd175) and at ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/analysis/NIST_SVs_Integration_v0.6/.
The GIAB genome stratification (v3.5) is available at https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/.
The population-level HGSVC3 Freeze4 SV sets are available at https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/working/20240415_Freeze4.
The assembly-level HGSVC3 Freeze4 SV sets are available at https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/working/20240307_PAV_VCF/.
The HPRCY1 pangenome-derived SV sets are available at https://github.com/human-pangenomics/hpp_pangenome_resources.
SVs called from the moderate-coverage ONT data of 888 1KGP3 individuals are available at http://1kgp-sv-imputation.s3-website-eu-west-1.amazonaws.com/?prefix=panel/.
SVs called from high-coverage ONT data of 100 1KGP3 individuals are available at https://s3.amazonaws.com/1000g-ont/index.html?prefix=Gustafson_etal_2024_preprint_SUPPLEMENTAL/.
The gnomAD SV set (v4.1) is available at https://gnomad.broadinstitute.org/downloads.
The 1000 Genomes Project Phase 3 genotypes are available at https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20201028_3202_phased/.
The GTEx (v8) data are available at https://www.gtexportal.org/home/datasets.
The histone ChIP-seq peak and RBP eCLIP peak data are available through the ENCODE portal (https://www.encodeproject.org).
The 18-state Roadmap model annotation data are available at EpiMap (https://compbio.mit.edu/epimap/).
The TAD coordinate data are available at the 3D Genome Browser (http://3dgenome.fsm.northwestern.edu/publications.html).
The deCODE SV-GWAS summary data are available at https://www.decode.com/summarydata/.
The genome partition of LD-independent blocks is available at https://bitbucket.org/nygcresearch/ldetect-data/.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Sex was included as a biological variable and used as a covariate in our analyses. The data was obtained from the UK Biobank

Reporting on sex and gender	(UKB), which records sex at recruitment.
Population characteristics	The association analyses were performed on 456,643 UKB individuals of European ancestry.
Recruitment	We analyzed existing UKB data sets. Thus, no recruitment was performed.
Ethics oversight	This study was approved by the Ethics Committee of Westlake University (No. 20200722YJ001).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For the SV imputation reference panels, 241 individuals were included, and for the SV association analyses, 456,643 UKB individuals of European ancestry were included.
Data exclusions	Participants who withdraw from UK Biobank were excluded.
Replication	The robustness of our results was confirmed through a series of sensitivity analyses with various statistical settings and filtering criteria.
Randomization	The SV reference panel was constructed using all available samples; therefore, randomization was not applicable. Covariates such as sex, age, and genetic principal components were controlled for in the SV association analyses.
Blinding	The SV reference panel was constructed using all available samples, and analyses were performed on existing datasets; therefore, no blinding measures were implemented in this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging