

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

In the format provided by the
authors and unedited

Contents

Supplementary Notes 1-15

Supplementary Figures 1-18

Supplementary References

Supplementary Note 1: Limitations of previous structural variant imputation panels

Prior work has utilized SVs identified from short-read data of the 1000 Genomes Project Phase 3 (1KGP3) as a reference for SV imputation¹. Unfortunately, most insertions, tandem repeats, and segmental duplications are underrepresented in this reference panel due to the complexities in identifying these types of SVs using short-read data. More recent efforts have used Oxford Nanopore Technologies (ONT) long-read sequencing to create more comprehensive SV datasets²⁻⁴. Two studies, for instance, used moderate-coverage ONT data to build imputation panels from 3,622 individuals (~134K SVs)⁴ and 888 individuals (~108K SVs)³. However, the numbers of SVs called were relatively small considering the sample sizes, likely due to the high error rate in these ONT data⁵. Another study identified ~16K VNTRs from 64 long-read assemblies from the Human Genome Structural Variation Consortium (HGSVC)⁶ and genotyped the VNTRs in 4,688 individuals with short-read data⁷, but the majority of other SVs were not examined.

Supplementary Note 2: Benchmarking and selection of assembly-based variant callers

Three commonly used, assembly-based variant calling methods, Dipcall⁸, PAV⁶, and SVIM-asm⁹, were used for variant calling (Methods and Supplementary Fig. 1). The numbers of SVs and SGVs detected per assembly aligned well across callers after quality control (Extended Data Fig. 2a) and with those reported previously¹⁰. Comparing SVs using Truvari (v4.2.2)¹¹ between pairwise callers revealed strong concordance among SVs identified by the three callers (Extended Data Fig. 2b). To enhance the robustness of our results, we discarded SVs detected by only one caller¹². We then benchmarked the SVs called from the individual HG002 by the three callers against the Genome in a Bottle (GIAB) v0.6 Tier 1 SVs¹³, a meticulously curated benchmark set from the GIAB consortium. This benchmark analysis showed that while all three callers achieved a high F1 score of over 0.9, PAV marginally outperformed the others, including the combined set of SVs from the three callers, with a recall of 0.97, a precision of 0.92, and a genotype concordance rate (GCR) of 0.99 (Extended Data Fig. 2c). Thus, we chose SVs called by PAV and supported by at least one additional caller for subsequent analyses. Applying this strategy to our 482 assemblies, we identified an average of 15,765 high-confidence, non-centromeric autosomal SVs (including 9,722 insertions, 6,017 deletions, and 26 inversions) per assembly (22,295 autosomal non-centromeric SVs per sample) (Fig. 1a), with a mean Mendelian discordance rate of 3.4% as quantified in four trios (Extended Data Fig. 2d). Benchmarking our SVs against those from HGSVC¹⁴ and the Human Pangenome Reference Consortium Year 1 (HPRCY1)¹⁰ yielded high F1 scores of 0.95 and 0.81, respectively (Supplementary Fig. 2). This discrepancy likely reflects inherent differences between graph-based (HPRCY1) and linear-reference SV calls¹⁰. Consistent with previous work^{1,10}, assemblies of African ancestry exhibited a greater number of SVs than those of other ancestries, indicating higher genetic diversity in this population (Fig. 1a).

Supplementary Note 3: Development of the SV and VNTR imputation reference panels and online server

Utilizing the variants derived from the haplotype-resolved assemblies, we constructed two reference panels for imputation: a panel for SV imputation in general (SV-panel) and a panel for VNTR imputation specifically (VNTR-panel). The SV-panel included only SNPs and SVs, as our benchmark analysis suggested that the inclusion of small insertion/deletion variants (INDELs) degraded the imputation accuracy of SVs (Supplementary Fig. 7), likely due to interference from the ambiguous breakpoints of uncertain INDELs. The VNTR-panel consisted of SNPs, VNTR SVs, and small tandem repeat INDELs with a repeat unit length greater than 6 bp, as these INDELs could also act as VNTR alleles. Complex VNTRs, which typically require a larger sample size to adequately represent motif type diversity, were not included in this panel. We utilized an imputation pipeline previously developed for imputing SGVs¹⁵ and modified it for SV imputation by treating each SV or VNTR allele, along with the allele in the reference genome, as a biallelic genetic marker. We have packaged this pipeline into an online tool for public use (https://yanglab.westlake.edu.cn/impute_sv).

Supplementary Note 4: Detailed evaluation of SV and VNTR imputation performance

We evaluated the performance of our SV imputation panel on the well-characterized HG002 genome, taking advantage of its rigorously validated SVs (Methods). We obtained the SNP genotype data of HG002 and extracted the SNPs that overlapped with those on three widely used SNP arrays, namely UKB Axiom (denoted by UKBA, with 0.8M SNPs), Multi-Ethnic Global (denoted by MEGA, with 1.7M SNPs), and the Infinium Omni 2.5M-8 SNP array (denoted by OMNI 2.5M, with 2.5M SNPs), to form three pseudo-arrays as the input for SV imputation. Additionally, we created an input set of 3.2M SNPs imputed based on UKBA using the TOPMed reference panel¹⁶ (denoted by pre-imputed SNPs), and another input set of 3.4M assembly-called SNPs to represent those called from WGS (denoted by WGS SNPs). For the pre-imputed SNPs, the imputed genotypes were used for SV imputation. The phasing information of these five input sets was obtained from the two assemblies of HG002. To avoid overlap between the input and the imputation reference panel, the haplotypes of HG002 were excluded from the reference panel. SV imputation was performed using these five input sets, and the imputed SVs were benchmarked against both the GIAB v0.6 Tier 1 SVs (9,228 SVs)¹³ and the GIAB Challenging Medically Relevant Genes (CMRG) v1.0 SVs (250 SVs)¹⁷. Overall, our imputation pipeline achieved a high level of performance for all the input sets, with a slight increase in recall and GCR, and a minor decrease in precision as the number of input SNPs increased (Fig. 2a). There was a noticeable improvement in F1 score and GCR from UKBA to the pre-imputed SNPs, and this improvement was even more pronounced in imputation accuracy

quantified at population level (see below). For instance, in the two benchmark SV sets, the pre-imputed SNPs showed F1 scores of 0.92 each and GCRs of 0.97 and 0.87, respectively, a performance nearly comparable to that of the WGS SNPs (Fig. 2a). These findings suggest that our proposed strategy can impute most SVs in both benchmark sets with reasonably high accuracy, including in regions harboring medically relevant genes that are difficult to accurately analyze due to their repetitive nature and complexity.

To assess the robustness of imputing SVs that are underrepresented in the GIAB benchmark set and to avoid potential bias toward a single individual, we conducted a leave-one-out imputation analysis. That is, we performed SV imputation for each of the 241 reference individuals using a reference panel from which the individual being imputed and their relatives, if any, were excluded. Using the HGSC3 SVs (Freeze4)¹⁴ as the ground truth, we benchmarked the imputed SVs in the entire genome or specific regions stratified by their complexity (Methods)¹⁸. We found that imputation using the full reference panel showed higher accuracy than using only the ancestry-matched subset (Extended Data Fig. 3a). For the genome-wide SVs, all the input sets achieved a median F1 score of ≥ 0.82 and GCR of ≥ 0.87 (e.g., F1 score = 0.84 and GCR = 0.89 for the pre-imputed SNPs; Fig. 2b), which is slightly lower than those of GIAB Tier 1, partly because the genome-wide SV set contains more SVs in complex regions than the GIAB Tier 1 set. Sensitivity analysis under more stringent benchmarking conditions confirmed the robustness of the results (Extended Data Fig. 3b). Not surprisingly, the imputation performance depended heavily on the complexity of the genomic regions (Fig. 2b), as the resolution offered by using SNPs for accurate haplotype selection may be limited in complex regions. For confident regions where reads are easy to map, all the input sets achieved a median F1 score of >0.93 and GCR of >0.96 (e.g., F1 score = 0.95 and GCR = 0.98 for the pre-imputed SNPs). For regions of low mappability or with a high level of tandem repeats, where most SVs are difficult to reliably call by short-read WGS, the imputation still yielded a median F1 score of >0.68 and GCR of >0.82 . In addition, when using SVs called from the 1KGP3 high-coverage ONT data (hcONT)² as the ground truth, the imputed SVs achieved an F1 score of 0.76 for genome-wide SVs using pre-imputed SNPs (Extended Data Fig. 4), slightly lower than that of HGSC3 hcHiFi SVs, likely due to the incomplete coverage of our imputed SVs by the ONT call set. Together, these results demonstrate the feasibility of leveraging haplotype-resolved assemblies, although currently available in a relatively small number of individuals, to impute most SVs throughout the entire genome with high accuracy.

We next quantified imputation accuracy at population level by calculating the squared correlation (r^2) between the imputed and actual genotypes of SVs in the leave-one-out imputation. In each of the four ancestry groups (i.e., AFR, EAS, EUR, and WAS), we grouped the imputed SVs into different

MAF bins and observed an increasing trend of r^2 with higher MAF or density of input SNPs (Fig. 2c). The performance of the pre-imputed SNPs was once again comparable to that of the WGS SNPs. For common SVs (MAF ≥ 0.01), the median r^2 was 0.78 averaged across ancestry groups when using our assembly-called SVs as ground truth (Fig. 2c). For the EUR ancestry, the median r^2 was 0.80 and 0.83 when using SVs called from the moderate-coverage ONT data (mcONT)³ of 1KGP3-EUR individuals and the high-coverage short-read WGS data (srWGS)¹⁹ of a subset of UKB-EUR individuals as ground truth, respectively (Supplementary Fig. 8). Compared to common SVs, the performance for rare SVs (MAF < 0.01) was dramatically lower (Supplementary Fig. 8) and, therefore, was not considered in the real data applications discussed below.

We compared our SV panel with two existing panels, the 1KGP3 mcONT panel (N=888)³ and the srWGS panel (N=3,202)¹. Our hcHiFi panel provided more comprehensive genome-wide SV coverage than the mcONT and srWGS panels (Supplementary Fig. 9a). We conducted leave-one-out imputation using the three panels with the UKBA as input. Benchmarked against HGSVC3 Freeze4 SVs¹⁴, our panel achieved the highest F1 score (0.82 vs. 0.58 for the mcONT panel and 0.27 for the srWGS panel) (Supplementary Fig. 9b). Under more stringent benchmarking conditions for long-read-based SVs, the F1 score of the hcHiFi panel remained robust, while that of the mcONT panel decreased slightly, likely due to differences in SV representation between the two long-read technologies (Supplementary Fig. 9c). For shared SVs among the three imputed sets, all three sets showed similar genotype concordance rates; however, it should be noted that these shared SVs were biased towards those in high mappability regions due to their intersection with SVs called from srWGS. Moreover, owing to its higher genome-wide SV coverage, our hcHiFi panel yielded 2.0 and 2.3 times more SVs with $r^2 > 0.3$ than the mcONT and srWGS panels, respectively (Extended Data Fig. 5). These findings suggest an important consideration for future efforts: substantially less sequencing may be necessary to build an accurate imputation panel if it is based on hcHiFi data, although the choice of a sequencing strategy for any given project will require a customized cost-benefit analysis that considers factors such as the genetic diversity of the target population. In addition, the performance difference between two versions of our panel highlights the improvement gained from increasing the sample size of an hcHiFi panel (Extended Data Fig. 5).

We also evaluated the performance of the VNTR-panel in imputing VNTR length (i.e., copy number of repeat units), which is more challenging than imputing other non-repeat SVs. For each multi-allelic VNTR locus, we calculated the r^2 between the imputed VNTR length and the VNTR length called from assemblies in each ancestry group, using the leave-one-out imputation strategy. Across all imputed multi-allelic VNTRs, those imputed based on the pre-imputed SNPs captured a substantial proportion of the length variation with a median r^2 of 0.70, 0.66, 0.60 and 0.50 in EUR,

AFR, WAS and EAS, respectively (Fig. 2d). We further benchmarked the imputed VNTRs against those called from srWGS data in the UKB-EUR¹⁹. Due to the challenge of accurately identifying tandem repeats within each VNTR locus using short-read data, we focused on the imputed alleles that were also identified by the UKB srWGS and calculated the genotype concordance; the concordance rate between the imputed and called genotypes was 0.78 (Supplementary Fig. 10).

Supplementary Note 5: Simulation study assessing joint and marginal variant-based heritability estimation models

The large difference in $\hat{h}_{\text{cSV}(\text{imp})}^2$ between the two models led us to perform a simulation study to investigate the difference between the imputed SVs and SGVs in capturing causal genetic variations. To emulate the real-world scenarios, we simulated a phenotype based on called SVs and SGVs, from short-read WGS data of 505 unrelated individuals of European ancestry from the 1KGP3 (1KGP3-EUR), and performed the GREML analyses using imputed SVs and SGVs (Methods). Imputed variants with MAF < 0.01 were excluded. To ensure the GREML analysis was applicable to the small sample size of 1KGP3-EUR²⁰, we confined the simulation to a single chromosome: chromosome 2, 10, or 20. The phenotype was generated by randomly sampling certain proportions of the called common and rare SVs and SGVs as causal variants (Methods). On average across simulations, when all the causal variants were SVs, 46.1% (67.2%) and 58.1% (88.2%) of the variance explained by all common causal SVs (denoted by h_{cSV}^2) were captured by the imputed common SVs and SGVs, respectively, in the joint (marginal) analysis (Supplementary Fig. 11). Conversely, when all the causal variants were SGVs, the imputed common SGVs captured 98.1% (98.5%) of the variance attributed to all common causal SGVs (denoted by h_{cSGV}^2), whereas the imputed common SVs only captured 1.1% (32.9%) of h_{cSGV}^2 , in the joint (marginal) analysis (Supplementary Fig. 11). These results are consistent with the observations that in most cases, any SV of interest was well tagged by SGVs, as measured by the sum of linkage disequilibrium (LD) r^2 between the target SV and the SGVs nearby; in contrast, an SGV of interest was often not well tagged by SVs, given the sparse distribution of SVs across the genome (Supplementary Fig. 12). These results suggest that $\hat{h}_{\text{cSV}(\text{imp})}^2$ from the marginal model overestimated h_{cSV}^2 because of the tagging of causal SGVs, whereas $\hat{h}_{\text{cSV}(\text{imp})}^2$ from the joint model underestimated h_{cSV}^2 primarily because of the competition from the imputed SGVs in tagging the causal SVs. By combining our real-data heritability with the simulation results, we extrapolated that all common SVs may account for ~8% of the common genetic variance for complex traits on average (Supplementary Figs. 13-14). However, this value should be interpreted with caution, as the extrapolation depends heavily on the assumed genetic architecture of the simulated traits.

Supplementary Note 6: Extrapolating of the proportion of variance explained by all common causal SVs for complex traits

We utilized GCTA-GREML to estimate the overall contributions of the imputed common SVs and SGVs to phenotypic variance, denoted as $h_{\text{cSV}(\text{imp})}^2$ and $h_{\text{cSGV}(\text{imp})}^2$, respectively. The estimated mean $\hat{h}_{\text{cSV}(\text{imp})}^2$ and $\hat{h}_{\text{cSGV}(\text{imp})}^2$ across 636 quantitative traits in the UKB were 6.7% and 20.9%, respectively, when estimated separately and were 1.0% and 20.2%, respectively, when estimated jointly. We showed by simulation that $\hat{h}_{\text{cSV}(\text{imp})}^2$ based on the marginal model overestimated h_{cSV}^2 (i.e., the proportion of phenotypic variance explained by all common causal SVs) while that based on the joint model underestimated h_{cSV}^2 . Given $\hat{h}_{\text{cSV}(\text{imp})}^2$ and $\hat{h}_{\text{cSGV}(\text{imp})}^2$ quantified from real data and the proportions of causal genetic variation captured by imputed variants (i.e., the π values), we sought to extrapolate h_{cSGV}^2 (i.e., the proportion of phenotypic variance explained by all common causal SGVs), and more importantly h_{cSV}^2 , for real traits under certain assumptions.

Assuming negligible impacts of rare variants on the variance explained by imputed common variants, the joint $h_{\text{cSV}(\text{imp})}^2$ and $h_{\text{cSGV}(\text{imp})}^2$ can be approximated as:

$$h_{\text{cSV}(\text{imp})}^2 \approx \pi_{\text{cSV} \leftarrow \text{cSV}(\text{imp})} h_{\text{cSV}}^2 + \pi_{\text{cSGV} \leftarrow \text{cSV}(\text{imp})} h_{\text{cSGV}}^2 \quad (\text{Eq. 1})$$

$$h_{\text{cSGV}(\text{imp})}^2 \approx \pi_{\text{cSV} \leftarrow \text{cSGV}(\text{imp})} h_{\text{cSV}}^2 + \pi_{\text{cSGV} \leftarrow \text{cSGV}(\text{imp})} h_{\text{cSGV}}^2 \quad (\text{Eq. 2})$$

where $\pi_{\text{cSV} \leftarrow \text{cSV}(\text{imp})}$ and $\pi_{\text{cSGV} \leftarrow \text{cSV}(\text{imp})}$ ($\pi_{\text{cSV} \leftarrow \text{cSGV}(\text{imp})}$ and $\pi_{\text{cSGV} \leftarrow \text{cSGV}(\text{imp})}$) denote the proportions of variation at common causal SVs or SGVs, respectively, captured by imputed common SVs (SGVs). These π values were quantified by simulations (see Methods for details about the simulation strategy and settings) (Supplementary Fig. 13). Recalling that the mean joint $\hat{h}_{\text{SV}(\text{imp})}^2$ and $\hat{h}_{\text{SGV}(\text{imp})}^2$ across 636 UKB traits were 1.0% and 20.2%, respectively, we may predict h_{SV}^2 and h_{SGV}^2 for those traits to be approximately 1.7% and 19.6%, respectively, by solving the two equations above. This suggests that on average, SVs could account for ~8% of the overall common genetic variance for complex traits (Supplementary Figs. 13 and 14). This percentage remained relatively robust (ranging from 6% to 11%) when the π values were quantified from simulations in which we oversampled causal SVs from complex regions or genomic regions harbouring candidate cis-regulatory elements, varied the numbers of or the proportions of variance explained by causal SVs and SGVs, or altered the variance attributed to rare causal variants.

Our extrapolated estimates of h_{cSV}^2 and h_{cSGV}^2 at 1.7% and 19.6%, respectively, imply that around 8% of common genetic variation could be attributed to SVs on average across traits. However, this extrapolation needs to be interpreted with caution, as the estimates of $\pi_{\text{cSV} \leftarrow \text{cSV}(\text{imp})}$ and

$\pi_{\text{cSV} \leftarrow \text{cSGV}(\text{imp})}$ depend on the assumed genetic architecture of the simulated traits, particularly the complexity of the genomic regions where the causal SVs are located. In our simulation study, we quantified $\pi_{\text{cSV} \leftarrow \text{cSV}(\text{imp})}$ and $\pi_{\text{cSV} \leftarrow \text{cSGV}(\text{imp})}$ by treating the SVs called from short-read data as the ground truth, which might be biased towards SVs in less complex genomic regions. Despite this potential bias, the effect appeared to be small, as demonstrated by the consistent estimates of both $\pi_{\text{cSV} \leftarrow \text{cSV}(\text{imp})}$ and $\pi_{\text{cSV} \leftarrow \text{cSGV}(\text{imp})}$ even after we strongly enriched the causal SVs in complex regions (Supplementary Fig. 13). Furthermore, the extrapolation assumed a negligible impact of rare variants on $\hat{h}_{\text{cSV}(\text{imp})}^2$ and $\hat{h}_{\text{cSGV}(\text{imp})}^2$, which was supported by further simulations showing that the estimated π values remained largely unchanged when we varied the proportion of variance attributed to rare causal variants (Supplementary Fig. 13).

Ultimately, a more accurate and unbiased estimation of the variance explained by all SVs, including rare ones, will require direct SV calling from long-read sequencing technologies in tens of thousands of individuals in the future.

Supplementary Note 7: Comparison of SV associations with a prior short-read-based imputation study

Recently, Halldorsson et al. established an SV imputation panel using the short-read WGS data of ~150K UKB individuals, imputed SVs in the remaining individuals of the UKB, and conducted association analyses of the imputed SVs¹⁹. We compared our SV associations with those from Halldorsson et al. across 47 quantitative traits analyzed in both studies. A substantial fraction of SV-trait associations (3,401 out of a union set of 6,442) was shared between the two studies, with comparable P-values (Extended Data Fig. 9). While 1,382 of the 6,442 associations were unique to the Halldorsson et al. study, potentially benefiting from the larger sample size of the reference panel, 1,663 associations were unique to our study. The associations unique to our study likely stem from the superior accuracy of our HiFi assembly-based panel, especially in complex genomic regions (e.g., areas with low mappability or repetitive sequences) where short-read WGS is less effective. Indeed, most of our unique findings were located in these challenging regions. Additionally, the higher precision of HiFi-derived SV breakpoints can improve variant merging and enhance statistical power. Conversely, the associations unique to Halldorsson et al. likely reflect the statistical power afforded by their large reference panel. Ultimately, the substantial overlap and complementary nature of the results suggest that our HiFi-based panel performs comparably to a short-read-based panel over 600 times its size. This demonstrates the robustness of our approach and the significant potential of our online tool for SV imputation and association discovery.

Supplementary Note 8: Establishing criteria for distinct SV association signals

SVs are considerably longer than SGVs and are thus often presumed to have a higher functional impact than SGVs. However, due to the extensive genome coverage of SGVs and their effective tagging of SVs, determining whether SVs or SGVs are the underlying causal variants driving the observed associations in the presence of both SV and SGV association signals in a locus remains a challenge. Nevertheless, we reason that there are certain circumstances in which the SV associations are very unlikely to be driven by SGVs, as indicated by our simulations. In these simulations, phenotypes were generated based on the causal variants all sampled entirely from called SGVs and the association analyses were performed with our imputed SVs and TOPMed-imputed SGVs (see Methods and also the text above for details about the simulation settings). Under this simulation scenario, an imputed SV was the lead variant in only 0.6% of cases, remained significant after accounting for the SGVs in 0.6% of cases, or was the variant with the highest posterior inclusion probability (PIP) in fine-mapping in 0.7% of the cases; it fulfilled any of the above conditions in only 0.6% of cases. This means that in real data analysis, when we observe an SV association fulfilling any of these conditions, the SV association is highly unlikely to be driven by SGVs, thereby likely representing a distinct association signal although the observed SV may not itself be causal.

Supplementary Note 9: Comparison of our VNTR associations with a prior short-read-based imputation study

We compared our results to those from Mukamel et al.⁷, who identified 15,653 VNTRs from the 64 HGSC assemblies, created an imputation panel by genotyping 9,376 of these VNTRs in 4,688 individuals using short-read WGS data, and subsequently imputed these into the UKB. Across 51 traits analyzed in both studies, 1,154 of a combined 5,023 trait-associated VNTRs were shared (Extended Data Fig. 10). While 2,183 VNTR associations were unique to the Mukamel et al. study, 1,686 were exclusive to our study. This relatively small overlap highlights the complementary strengths of the two approaches. VNTRs are particularly challenging due to difficulties in their accurate calling and imputation, and because their extensive allelic diversity requires larger reference samples. The associations unique to our study likely stem from the superior ability of our purely HiFi assembly-based panel to identify a more comprehensive set of VNTR loci. Conversely, the associations unique to Mukamel et al. likely reflect the statistical power of their much larger reference panel, which, despite being based on short-read data, is better suited for capturing the extensive diversity of allele lengths at any given locus. Therefore, our study not only demonstrates the significant potential of using HiFi assemblies for robust VNTR discovery but also makes clear that expanding our assembly-based panel in the future will be critical to achieving the statistical power needed for even more comprehensive VNTR association analysis.

Supplementary Note 10: Limitations of this study

Our study has several limitations. First, although we have demonstrated the feasibility and accuracy of imputing common SVs from SNP data, imputation of rare SVs remains a challenge, partly because the limited sample size of the genome assemblies that are currently available. With the reduction in long-read sequencing costs and the collective efforts of the international community, we anticipate a dramatic increase in the sample size of long-read assemblies in the future, which will likely lead to improved accuracy for both common and rare SV imputations, as has been observed with SGV imputations. Second, our association analyses for 2,624 traits, while extensive, do not encompass all phenotypes available in the UKB. Specifically, to ensure adequate statistical power and the robustness of our findings, we excluded binary disease phenotypes with low case numbers (e.g., < 100 cases). Consequently, while our results provide novel insights for a wide range of traits, the current association data portal should not be viewed as an exhaustive resource for all diseases within the UKB, particularly for those with lower prevalence. This highlights the necessity for future, targeted association studies for these specific traits. A key contribution of our work is that the public imputation server we provide directly enables such follow-up analyses, empowering researchers to investigate the role of SVs in a broader spectrum of human diseases. Third, the SV association analyses were performed in the UKB-EUR, as the sample sizes of other ancestries were too small to provide sufficient power to detect individual SV associations. Nevertheless, our imputation reference panel includes genome assemblies of diverse ancestries, including six continental populations, which, in principle, could be applied to GWAS data of most non-European populations. Finally, in our VNTR analysis, we did not specifically include complex VNTRs (i.e. a VNTR locus containing more than one type of motif with significant difference), which could be better represented using repeat-pangenome²¹.

Supplementary Note 11: Generation of 58 1KGP3 European assemblies

58 assemblies from 29 1KGP3 individuals were generated in this study. High-quality DNA was extracted from the lymphoblastoid cell lines (Coriell Institute) of these individuals and sequenced on the PacBio Revio platform (25M SMRT Cells), yielding ~36.4x HiFi reads per individual (Supplementary Table 2). Adapters were removed using HiFiAdapterFilt (v2.0.0)²² with default parameters, and haplotype-resolved assemblies were constructed with Hifiasm (v0.19.8)²³, leveraging 1KGP3 trio-based high-coverage srWGS data for phasing. Quality assessment with QUAST (v5.0.2)²⁴, Merquy (v1.3)²⁵, and BUSCO (v5.2.2; mammalia_odb10)²⁶ using default parameters confirmed the high quality of the assemblies, revealing an average genome length of 2.98 Gb, an N50 of 45.43 Mb, and an auN of 52.03 Mb. The assemblies also showed a high quality value (QV) of 58.39 and a low percentage of missing genes at 4.31% (Supplementary Table 3).

More details about the assemblies can be found in Supplementary Tables 1-3 and Extended Data Figure 1.

Supplementary Note 12: Quality control of the called variants

We applied stringent quality control to the variants detected by the three callers. Firstly, we normalized all variants and excluded low-quality variants, as indicated by the FILTER flag in VCF files. We also discarded variants located in centromeres and heterochromatic regions, as annotated by the Genome Reference Consortium²⁷. Additionally, we filtered out variants detected in poorly aligned regions using a BED file generated by Dipcall. For each individual, the BED file defined the high-confidence regions as those covered by alignments with a length ≥ 50 Kbp and a mapping quality ≥ 5 , while not being covered by other alignments with a length ≥ 10 Kbp⁸. To assess consistency and enhance confidence in the SVs detected by Dipcall, PAV, and SVIM-asm, we conducted a reciprocal reproducibility test among the three callers using Truvari (v4.2.2)¹¹ with the options¹⁰ `--dup-to-ins -C 1000 -O 0.0 -p 0.0 -P 0.3 -s 50 -S 15`. We selected PAV-based SVs that were supported by at least one additional caller and created a combined SV set. For SGVs, we only retained those detected by both Dipcall and PAV with a consistent genotype, as SVIM-asm cannot detect small variants. Additionally, we calculated the consistency between genotypes derived from assembly-based haplotypes and those from 1KGP3 haplotypes for the same sample using WhatsHap (v1.4)²⁸, and excluded SNPs that displayed inconsistent genotypes between the two datasets. All quality control processes of the called SVs and SGVs were conducted using BCFtools (v1.14)²⁹, and this study focused only on autosomes.

Supplementary Note 13: Leave-one-out imputation experiment

We created pseudo-arrays for the 241 individuals in the reference panel, using the same pipeline described for benchmarking with HG002. Each individual and their relatives (if any) were held out separately, and the remaining individuals were used to build the reference panel. We constructed reference panels representing three distinct ancestry compositions: 1) the ancestry-matched panel, which included only individuals with ancestry matching that of the target individuals; 2) the half-mixed panel, which contained an equal number of samples from the matched ancestry and other ancestries; and 3) the maximum panel, which incorporated all available samples, representing diverse ancestries. The imputation was performed using the five input sets described above (i.e., UKBA, MEGA, OMNI 2.5M, pre-imputed SNPs, and WGS SNPs) for each individual with the corresponding leave-one-out reference panel. We calculated the precision, recall, and genotype concordance rates of imputed SVs against external HGSVC3 high-coverage HiFi-based SVs (sample N=65)¹⁴ and 1KGP3 high-coverage ONT-based SVs (sample N=100)² using Truvari (v4.2.2). We partitioned the genomic regions into "easy", "low mappability", "tandem repeat", and

"whole" regions and evaluated the performance for these regions respectively, leveraging a series of BED files provided by the GIAB genome stratification (v3.5)³⁰. Easy regions refer to genomic regions with relatively high mappability where variants can be confidently detected. Conversely, low mappability and tandem repeat regions often contain consecutive repetitions of short elements or exhibit high similarity to other genomic regions, making it difficult to align them uniquely to a reference genome. More details about these regions can be found at <https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.5/GRCh38/>.

Supplementary Note 14: Simulations to quantify variation at causal SVs or SGVs captured by imputed SVs and SGVs

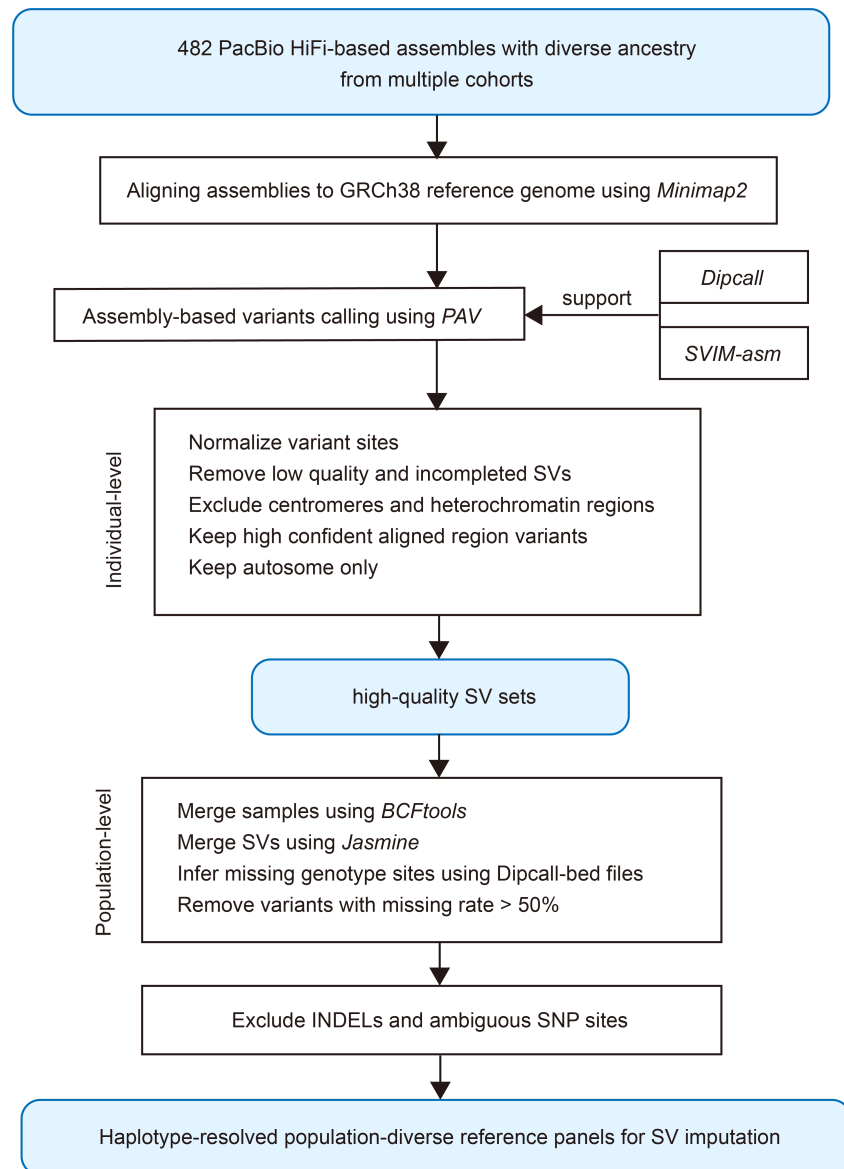
We used the genotypes of SVs and SGVs called from high-coverage short-read WGS data of 505 unrelated 1KGP3-EUR individuals¹ as the true genotype set for simulations. We extracted the UKBA SNPs from the true set for SGV imputation using the TOPMed reference panel¹⁶, and then utilized the imputed SNPs for SV imputation using our SV imputation panel, excluding overlapping individuals with the 1KGP3-EUR. As the sample size of 1KGP3-EUR was small, to make GCTA-GREML²⁰ applicable to the simulated data, we confined the simulation to a single chromosome, i.e., chromosome 2, 10, or 20. The numbers of called and imputed variants on the three chromosomes are presented in Supplementary Table 16. We performed the first set of simulations where the causal variants were all SVs or SGVs. The number of causal variants was set to 60, and the variance explained by all causal variants (h^2) was set to 0.3, with 20% of h^2 attributed to rare variants as indicated in prior work^{31,32}. The phenotype was simulated based on the genotypes of the causal variants from the real set using GCTA (--simu-qt)³³. We then conducted the second set of simulations using both SVs and SGVs as causal variants, with the number of causal SVs (m_c) set to 60 and the number of SGVs varying from m_c to $3m_c$, $6m_c$, and $9m_c$. All the causal variants were randomly selected across an entire chromosome, with potential enrichment in specific genomic regions specified below.

For a robustness test, we varied the simulation settings as follows: 1) sampling 90%, 80%, or 70% of causal SVs from complex regions, which occupy 21.3% of the genome, defined using GIAB genome stratification (v3.5)³⁰, to compensate for the underrepresentation of short-read-derived SVs in these regions; 2) sampling 90%, 80%, or 70% of causal variants from +/- 500 bp regions of candidate cis-regulatory elements (cCREs), which occupy 38.1% of the genome, defined by ENCODE annotations³⁴; 3) varying m_c from 60 to 30 and 90; 4) varying h^2 from 0.3 to 0.2 and 0.4; 5) varying the proportion of h^2 attributed to rare variants from 20% to 10% and 30%. Each simulation scenario was repeated 7,500 times.

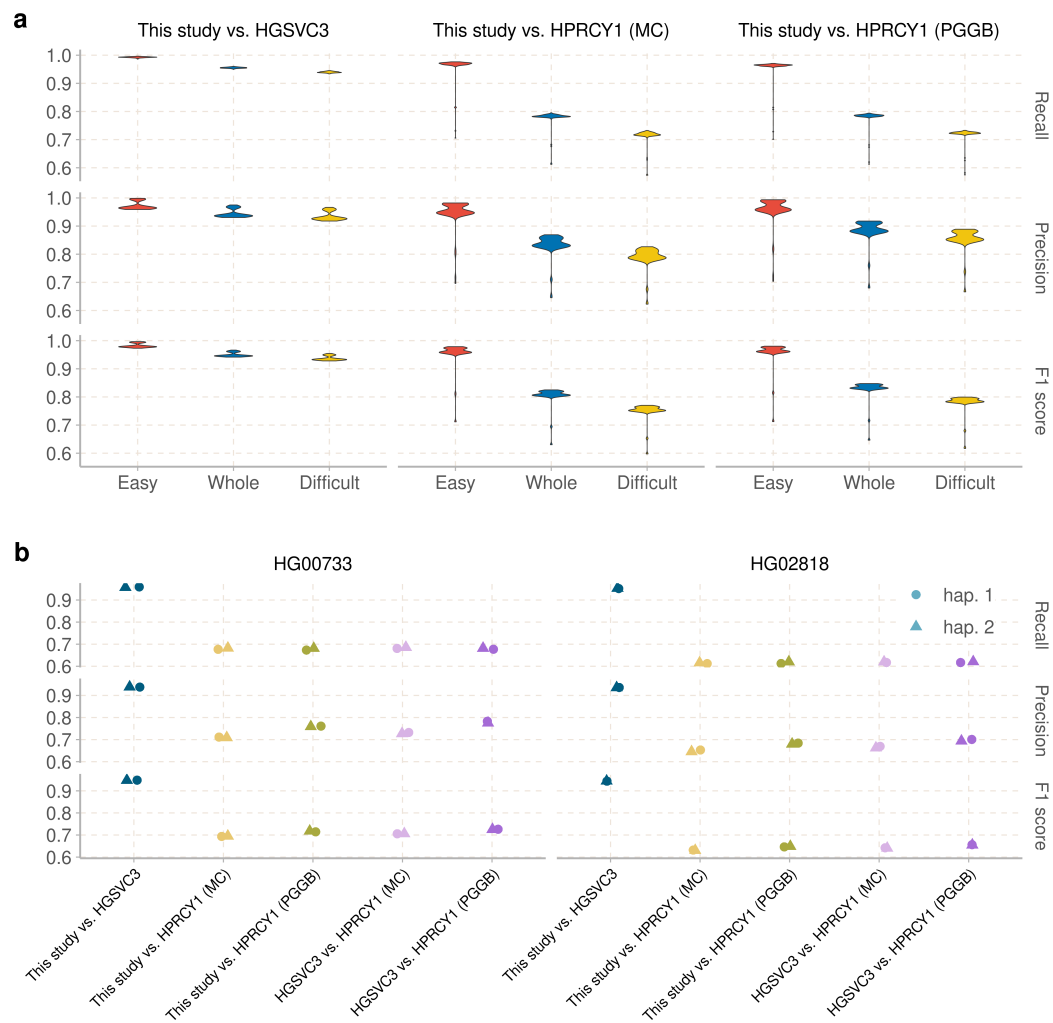
In each simulation replicate, we estimated the proportions of variance in the simulated phenotype explained by all imputed common SVs and SGVs using both the one-component marginal and two-component joint models. We also performed association analyses using the imputed SVs and SGVs to quantify the proportion of cases in which an imputed SV remains significant after accounting for the SGVs, is the lead association signal, or is the lead fine-mapped variant when the underlying causal variant is an SGV. Note that for the association analysis, to ensure sufficient power to detect SV associations at $P < 5e-8$, we reduced the number of causal variants to 6 and increased the phenotypic variance explained by all causal variants to 60%.

Supplementary Note 15: Functional enrichment analyses of trait-associated SVs

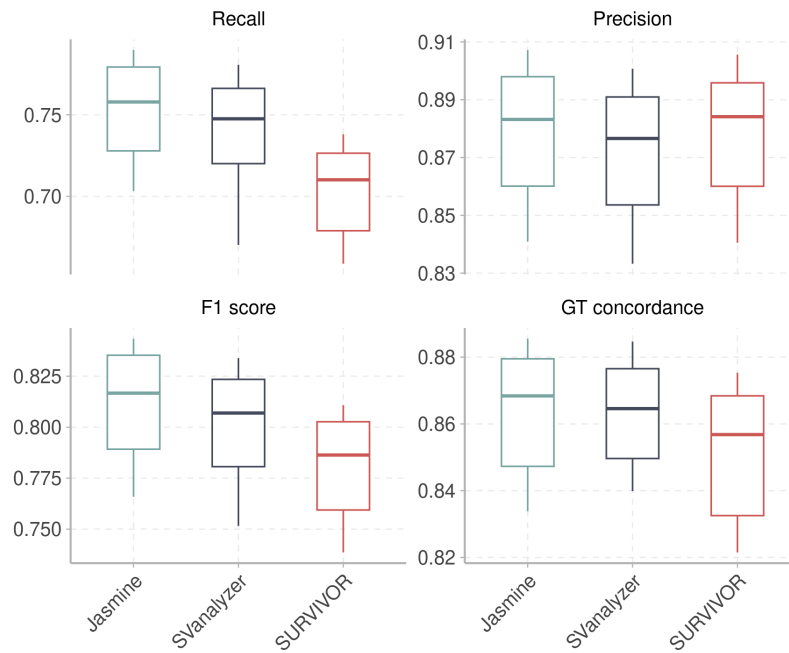
For histone modifications, we obtained histone ChIP-seq peak data for the seven most commonly used marks, namely H3K4me1, H3K4me3, H3K36me3, H3K27me3, H3K9me3, H3K27ac, and H3K9ac, from the ENCODE data portal³⁴. The ChIP-seq peak data were assayed for 26 cell types. We filtered out low-confidence peaks with a fold enrichment of peak calling < 2 or a P-value $> 1e-5$. The remaining peaks from the same histone marker and cell type were sorted and merged using BEDtools (v2.29.1)²⁹. For chromatin states, we obtained the 18-state Roadmap model annotation data³⁵ from Epigenome Integration across Multiple Annotation Projects (EpiMap)³⁶. We included 11 organ/tissue groups: blood, bone, brain, eye, kidney, liver, lung, muscle, pancreas, reproductive, and spleen. For RBP binding sites, we downloaded eCLIP peaks from ENCODE, including 140 target RBPs in K562 and HepG2 cell lines assayed by the Gene Yeo Lab³⁷. For chromatin interactions, we downloaded TAD coordinate datasets of 38 tissue/cell types from the 3D Genome Browser³⁸. We defined TAD boundary regions as the flanking regions of a TAD, spanning $\pm 50\%$ of its length³⁹. The modeled regions of heterochromatin and centromeres were excluded²⁷. We employed a logistic regression model in R to perform the enrichment analysis for trait-associated SVs compared to other SVs.



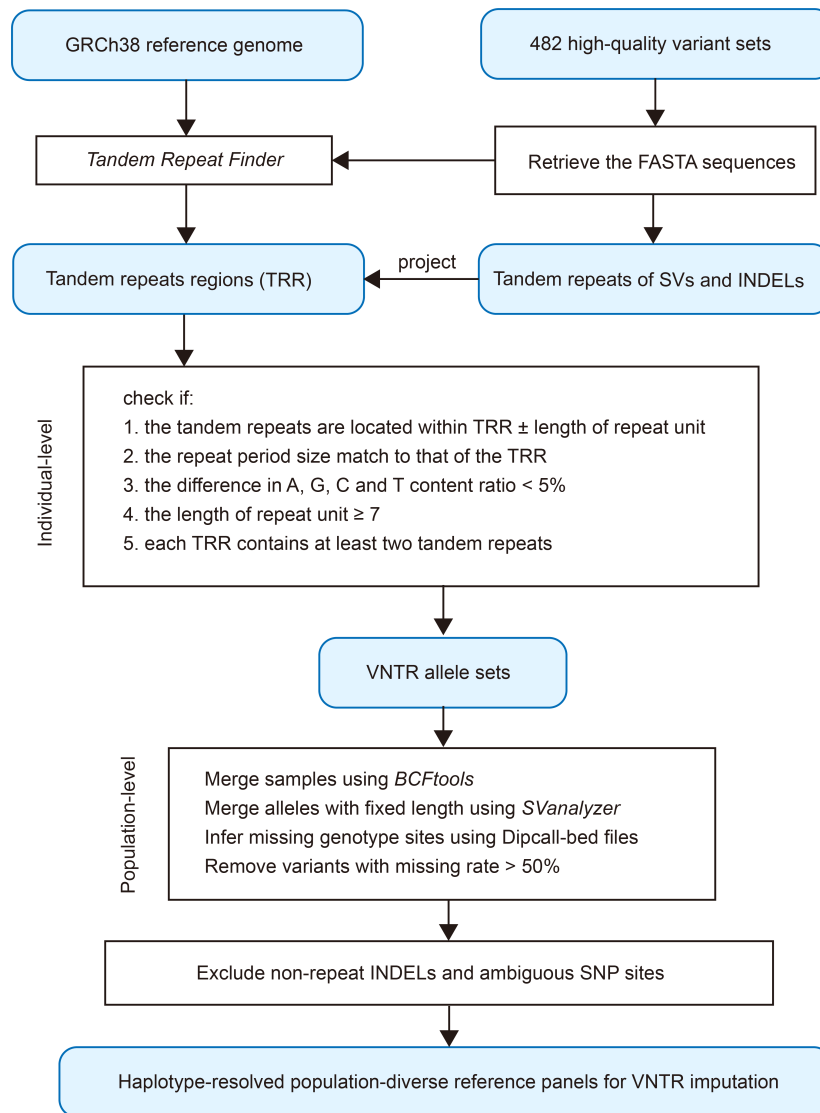
Supplementary Figure 1. Workflow of SV calling and SV imputation panel construction. Three widely used assembly-based SV calling methods were utilized in this study. Each SV was required to obtain support from PAV⁶ and at least one another caller. Stringent quality control was conducted for SVs at both the individual and population levels. The haplotype-resolved SVs and SNPs were combined to form the imputation panel, excluding SNPs that were inconsistent with those called using the high-coverage short-read WGS data from the 1KGP3¹.



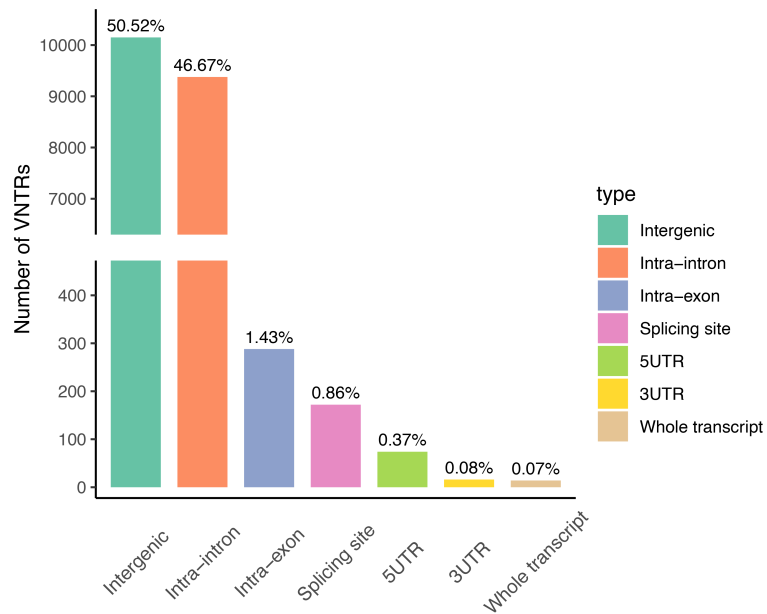
Supplementary Figure 2. Benchmarking of our assembly-level SVs against the HGSVC3 and HPRCY1 call sets. Truvari (v4.2.2)¹¹ with the options¹⁰ --dup-to-ins -C 1000 -O 0.0 -p 0.0 -P 0.3 -s 50 -S 15 --sizemax 100000 was used for the benchmarking analysis ($n_{\text{HGSVC3}} = 65$; $n_{\text{HPRCY1}} = 44$). a) The analysis was restricted to SVs in Dipcall-confident regions and subsequently stratified into “easy” and “difficult” regions using the GIAB v3.5 genome stratification³⁰. Outliers are defined as ± 1.5 times the IQR. b) Benchmarking analysis for two outlier samples (HG00733 and HG02818).



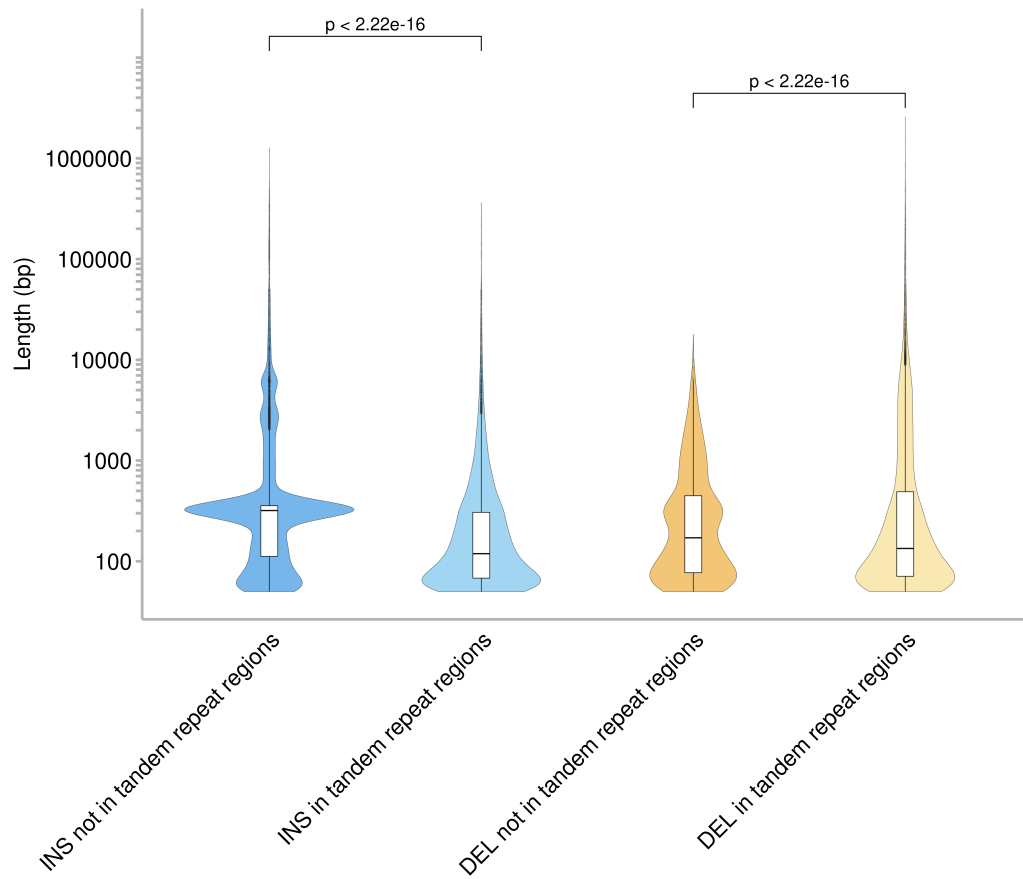
Supplementary Figure 3. Comparison of three SV merging methods for imputation accuracy. The performance of three SV merging methods (Jasmine⁴⁰, SURVIVOR⁴¹, and SVanalyzer¹³) was benchmarked based on the accuracy of imputed SVs. For each method, a reference panel was created and used for leave-one-out imputation of HGSVC3 individuals using UKB array SNPs. The accuracy of the imputed SVs was then evaluated against the HGSVC3¹⁴ (Freeze 4, $n = 65$) ground-truth set using Truvari¹¹ (v4.2.2; options¹⁰: --dup-to-ins -C 1000 -O 0.0 -p 0.0 -P 0.3 -s 50 -S 15). The boxplots display the medians, upper quartiles, and lower quartiles; whiskers extending to $1.5 \times$ the IQR; outliers are defined as $\pm 1.5 \times$ the IQR.



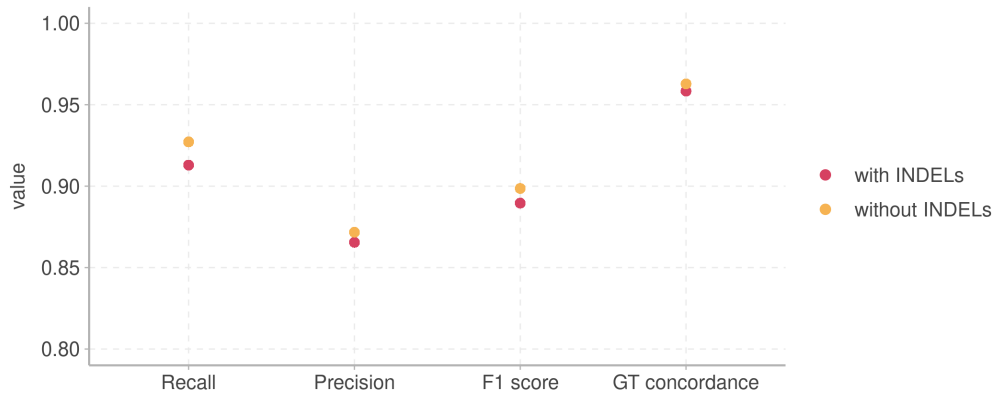
Supplementary Figure 4. Workflow of VNTR imputation panel construction. The tandem repeats of each assembly were discovered from the SV and INDEL sequences using Tandem Repeat Finder⁴². The individual-level tandem repeats were projected onto the GRCh38 tandem repeat regions. The haplotype-resolved VNTR alleles and SNPs were combined to form the VNTR imputation panel, excluding SNPs showing discrepancies with those called from the 1KGP3 short-read WGS data¹.



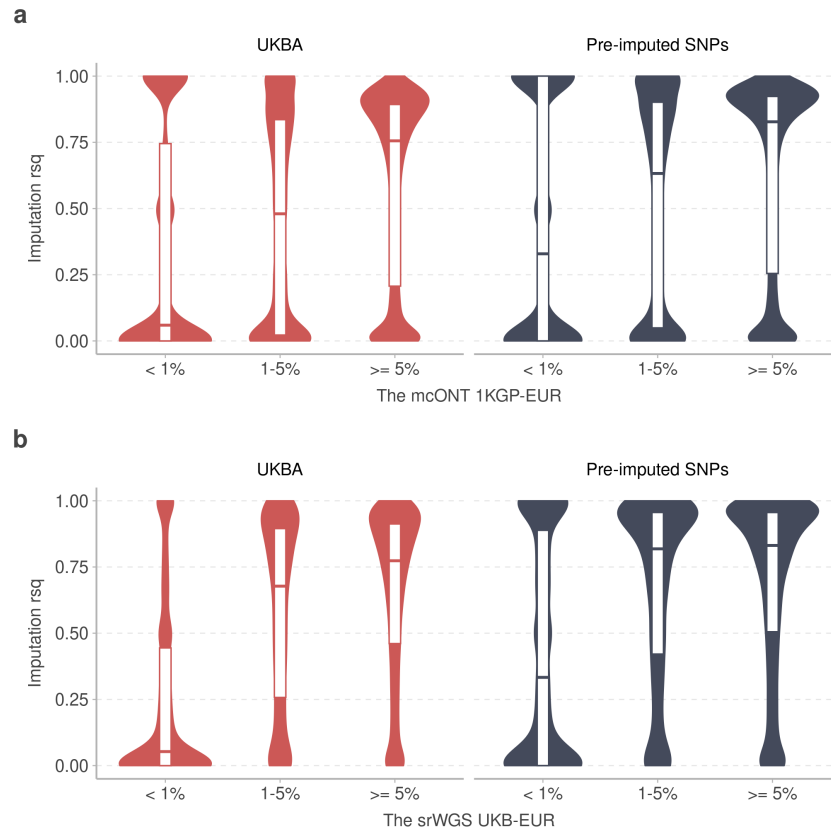
Supplementary Figure 5. Distribution of VNTRs across the genome. The x-axis displays different categories of genomic regions. The y-axis represents the number of VNTR loci, truncated for better visualization of the results. Each bar is labeled with the proportion of VNTR loci for the corresponding region, out of all VNTR loci.



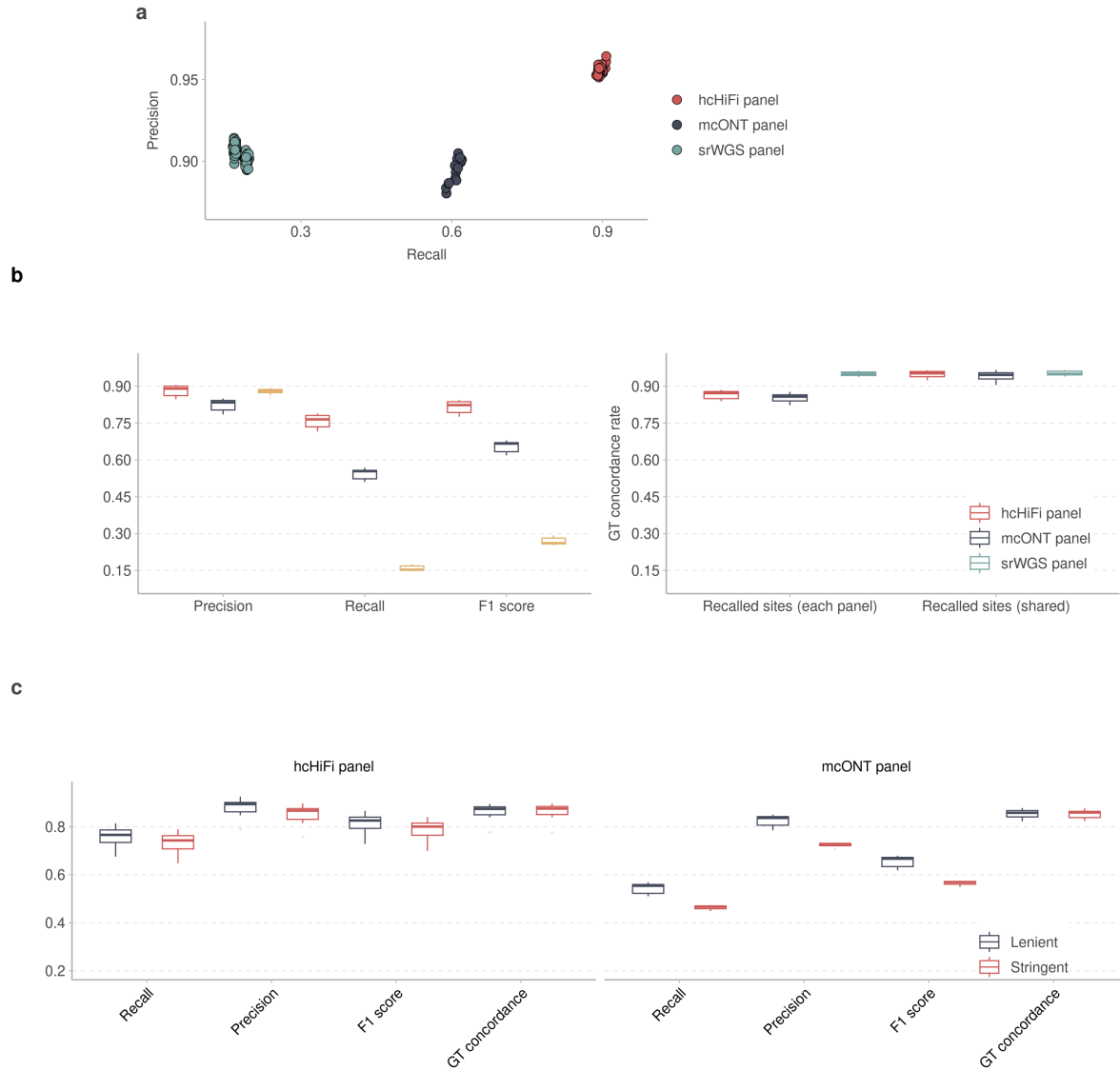
Supplementary Figure 6. Comparison of SV lengths between those located within and outside tandem repeat regions. Four categories are shown: DEL in tandem repeat regions ($m = 55,131$); DEL not in tandem repeat regions ($m = 19,045$); INS in tandem repeat regions ($m = 57,615$); INS not in tandem repeat regions ($m = 37,212$) P-values from a two-sided Wilcoxon rank-sum test are labeled. The boxplots display the medians, upper quartiles, and lower quartiles; whiskers extending to $1.5 \times$ the IQR; outliers are defined as $\pm 1.5 \times$ the IQR.



Supplementary Figure 7. Accuracy of imputing SVs using a reference panel with or without INDELs. The imputed SVs of HG002 were benchmarked against the GIAB HG002 Tier1 v0.6 SV set using 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`. The SNPs on the UKB array were used as the input for SV imputation. GT concordance is the genotype concordance of SVs shared between the imputed and HG002 Tier1 v0.6 SVs.

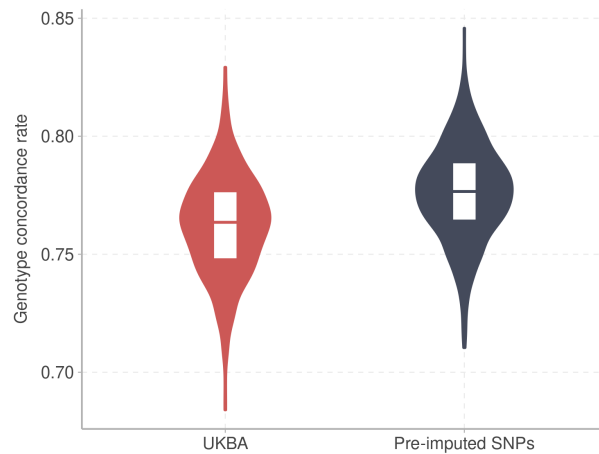


Supplementary Figure 8. Squared correlation between imputed and called SV genotypes in the 1KGP moderate-coverage ONT (mcONT) and UKB short-read WGS (srWGS) datasets. The SV imputation was conducted with our updated panel using the UKB array SNPs or the TOPMed pre-imputed SNPs. The 1KGP3 mcONT SV set³ (a; $n = 156$) and the UKB srWGS SV set¹⁹ (b; $n = 5,000$) served as ground truth. 1KGP3 samples overlapping with our reference panel were excluded from the analysis. Truvari (v4.2.2)¹¹ was utilized to align SVs across datasets with the options¹⁰: --dup-to-ins -C 1000 -O 0.0 -p 0.0 -P 0.3 -s 50 -S 15. The boxplots display the medians, upper quartiles, and lower quartiles; outliers are defined as $\pm 1.5 \times$ the IQR.

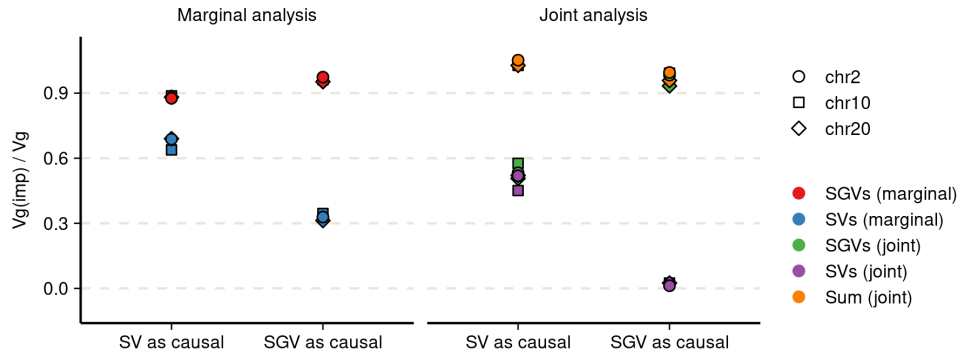


Supplementary Figure 9. Benchmarking analysis of SVs from three reference panels with different sequencing technologies against the HGSC3 Freeze4 SV set. The panels include: the high-coverage HiFi panel from this study (hcHiFi), the 1KGP moderate-coverage ONT panel³ (mcONT), and the 1KGP high-coverage short-read WGS panel¹ (srWGS). a) Benchmarking of SVs in each panel against HGSC3 SVs¹⁴. Truvari (v4.2.2)¹¹ was used to compare SVs in shared individuals between each panel and the HGSC3 ground truth, with parameters¹⁰ --dup-to-ins -C 1000 -O 0.0 -p 0.0 -P 0.3 -s 50 -S 15. Only SVs within HGSC3 Dipcall-confident alignment regions were considered. Shared sample sizes were: hcHiFi ($n = 65$), mcONT ($n = 15$), and srWGS ($n = 63$). b) Performance evaluation of SV imputation reference panels. The three panels (hcHiFi, mcONT, and srWGS) were assessed for SV imputation accuracy against the HGSC3 truth set. The leave-one-out imputation strategy was applied to 15 individuals shared across all panels, using UKB array SNPs as input. Genotype concordance rates are shown for both panel-specific imputed

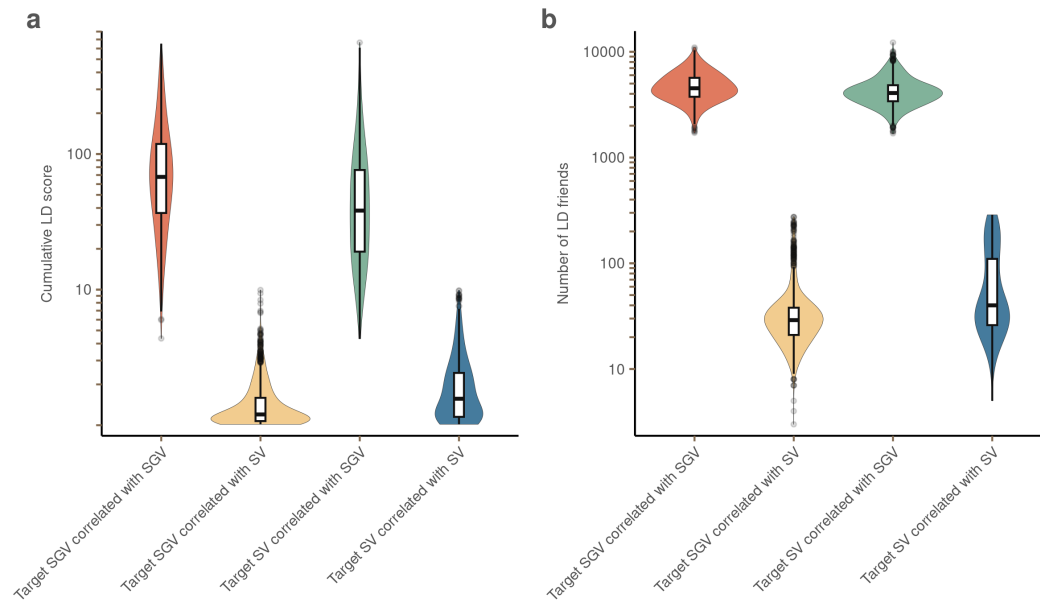
SVs and those shared across all three panels. The 15 shared individuals were: HG00096, HG00171, HG00268, HG00513, HG00731, HG01596, HG01890, HG02554, HG02953, HG03009, NA18989, NA19238, NA19331, NA19347, and NA20847. c) Two sets of benchmarking parameters were tested for the 15 samples: lenient (--dup-to-ins -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. The boxplots display the medians, upper quartiles, and lower quartiles; whiskers extending to $1.5 \times$ the IQR; outliers are defined as $\pm 1.5 \times$ the interquartile range (IQR).



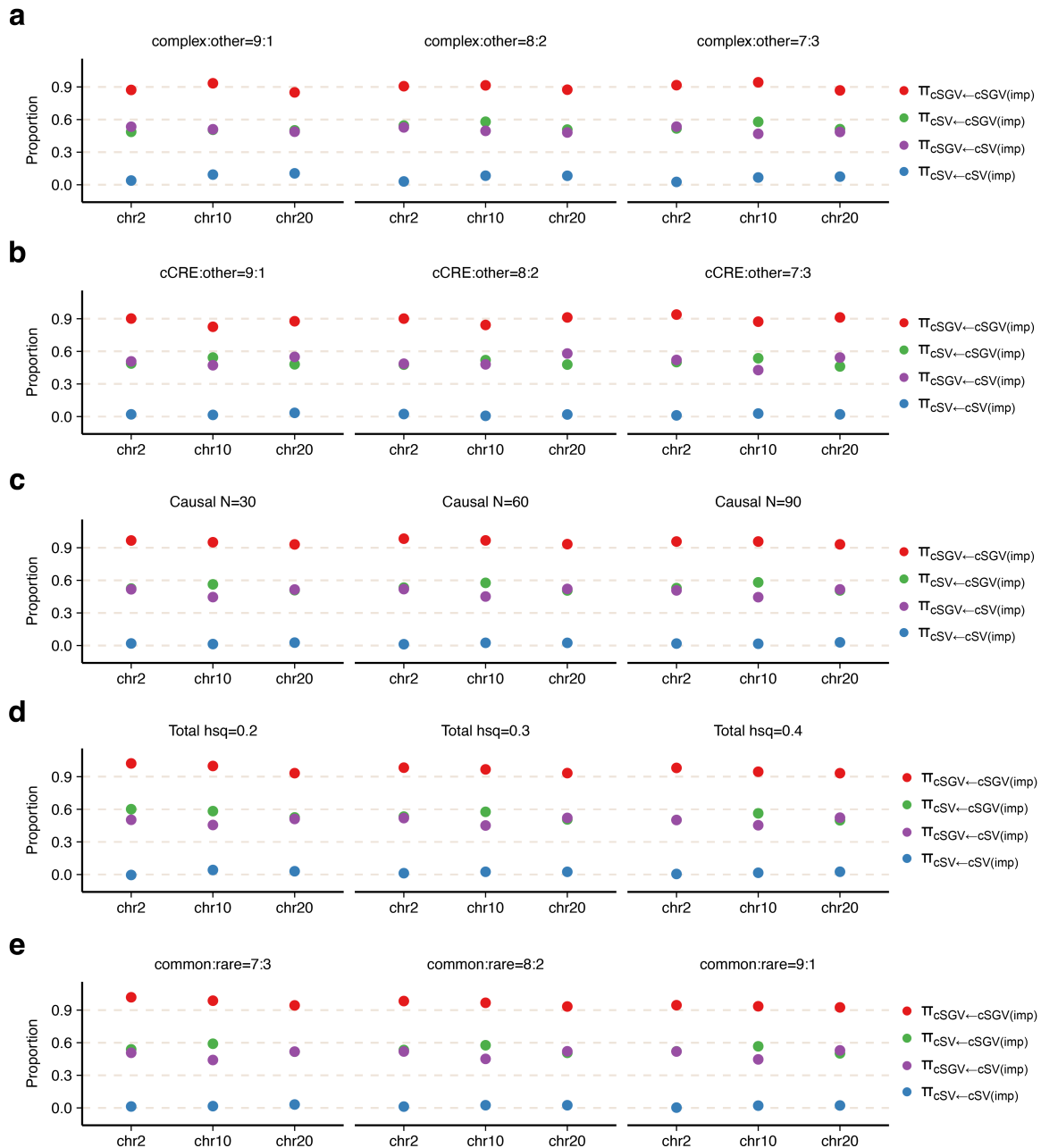
Supplementary Figure 10. Genotype concordance rates for VNTR alleles shared between the imputed set and the srWGS-called set in the UKB. Imputation was performed using our VNTR reference panel, with either the UKB array SNPs or the TOPMed pre-imputed SNPs as input. A total of 5,000 randomly sampled, unrelated EUR individuals were tested. Truvari¹¹ (v4.2.2) was used to compare genotypes of matching VNTR alleles between the imputed and called sets, with the following parameters¹⁰: --dup-to-ins -C 1000 -O 0.0 -p 1 -P 0.3 -s 50 -S 15. The boxplots display the medians, upper quartiles, and lower quartiles; outliers are defined as $\pm 1.5 \times$ the interquartile range (IQR).



Supplementary Figure 11. Proportion of common genetic variance explained by imputed common SVs or SGVs quantified by simulations. The estimates of variance explained were obtained using both the marginal (left column) and joint (right column) models. The y-axis represents the estimated proportion of genetic variance attributed to common causal SVs or SGVs explained by the common imputed SVs or SGVs, denoted by $V_{g(\text{imp})}/V_g$. In the simulated data, the phenotype was generated by 60 causal SVs or SGVs randomly sampled from the entire chromosome 2 (chr2), 10 (chr10), or 20 (chr20). The total heritability (h^2) was set to 0.3, with 20% of h^2 attributed to rare variants.

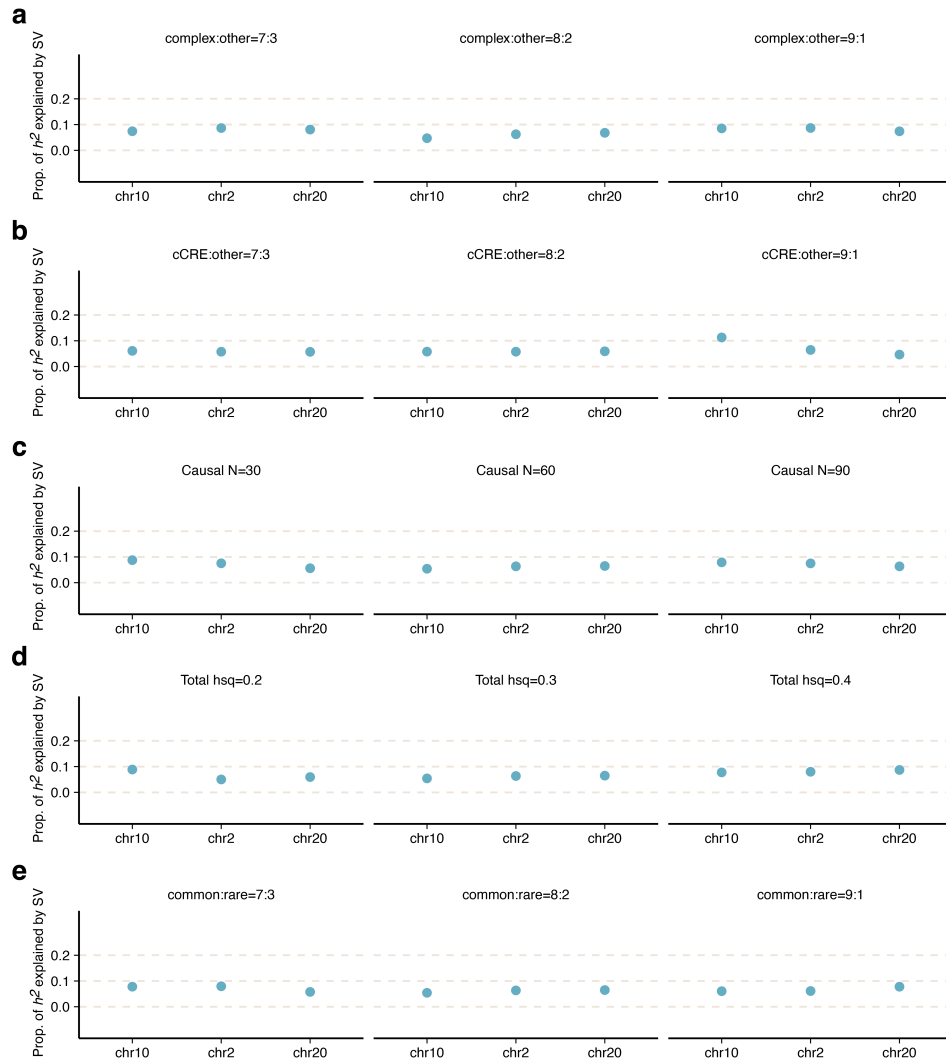


Supplementary Figure 12. LD scores and LD friends of the imputed SVs and SGVs computed from the UKB-EUR data. The calculation of LD scores (a) and the search for LD friends (b) were performed using genotype data on chromosome 20 of 10,000 unrelated UKB-EUR individuals. The x-axis displays four groups representing 2,000 target SVs or SGVs in LD with other SVs or SGVs. The y-axis is log-scaled to better visualize the results. The median, upper quartiles, lower quartiles, whiskers (extending to $1.5 \times$ the IQR) and outliers ($\pm 1.5 \times$ the IQR) are displayed in each violin plot.

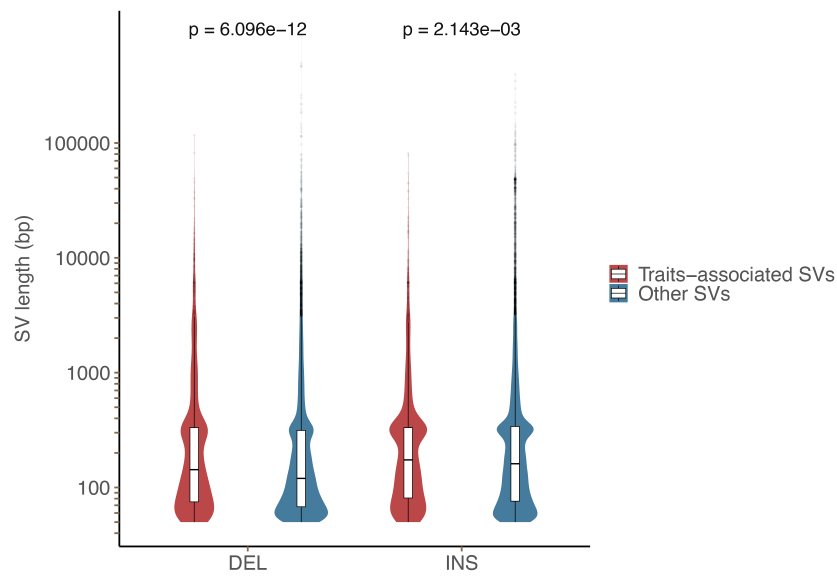


Supplementary Figure 13. The π values quantified using variance components estimated from the joint model across various simulation scenarios. The x-axis indicates the chromosome. Four types of π values - specifically, the proportions of variation at common causal SVs or SGVs captured by imputed common SVs and SGVs - are denoted by different colors. Apart from the simulations presented in Supplementary Fig. 11, we performed additional simulations for sensitivity analysis under the following scenarios: a) changing from the strategy of sampling causal variants at random from an entire chromosome to that of sampling 90% (left column), 80% (middle column) and 70% (right column) of causal SVs from complex regions that occupy 21.3% of the genome; b) changing from the strategy of sampling causal variants at random from an entire chromosome to that of sampling 90% (left column), 80% (middle column) and 70% (right column)

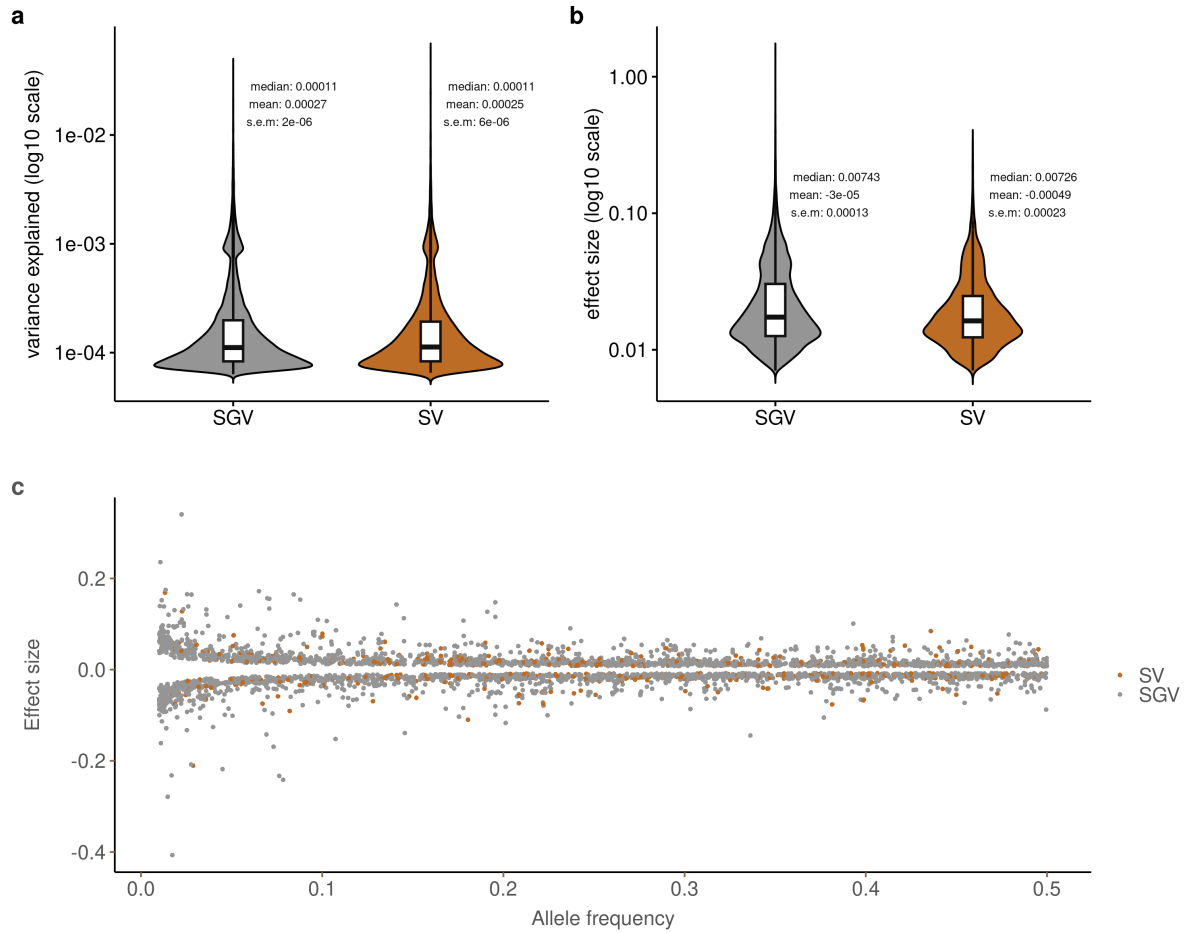
of causal variants from cCRE regions that occupy 38.1% of the genome; c) varying the total heritability (h^2) from 0.3 (middle column) to 0.2 (left column) and 0.4 (right column); d) varying the number of causal SVs from 60 (middle column) to 30 (left column) and 90 (right column) (note that the number of SGVs were changed accordingly as described in the Methods section); e) varying the proportion of h^2 attributed to rare variants from 20% (common:rare=8:2, middle column) to 10% (common:rare=9:1, left column) and 30% (common:rare=7:3, right column).



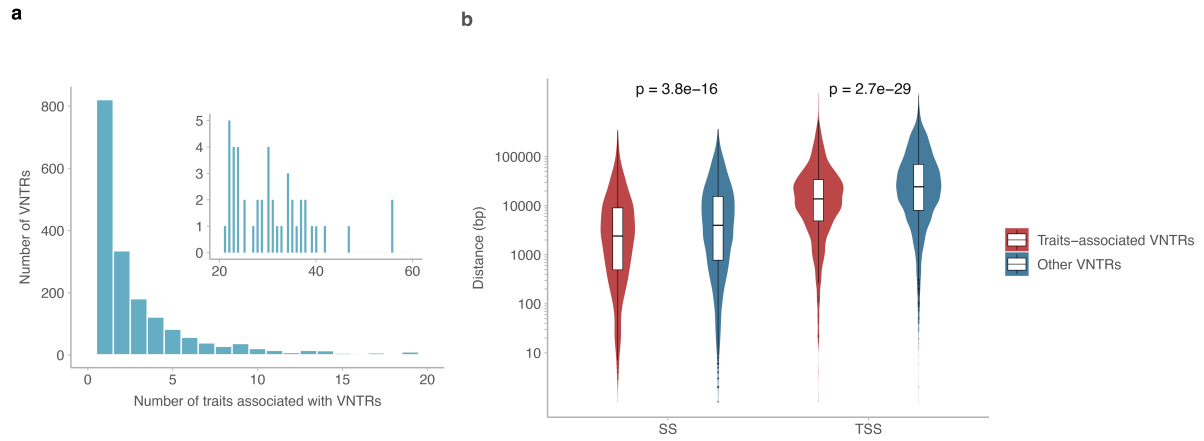
Supplementary Figure 14. Extrapolated proportion of the common genetic variance explained by all common causal SVs on average across traits. This proportion was quantified based on the four π values quantified on different chromosomes in various simulation scenarios as presented in Supplementary Figure 13.



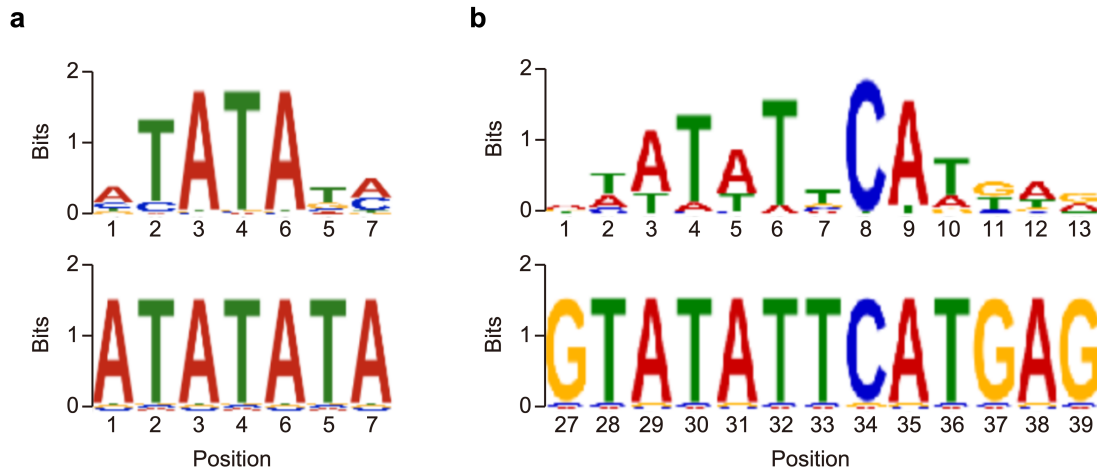
Supplementary Figure 15. Length distribution of the trait-associated SVs ($m = 4,397$) vs. other SVs. SV associations ($m = 32,372$) with a $P_{SV-GWAS} < 5e-8$ were considered significant. The y-axis is log-scaled to better show SV lengths. P-values from a two-sided Wilcoxon rank-sum test are labeled. The median, upper quartiles, lower quartiles, whiskers (extending to $1.5\times$ the IQR) and outliers ($\pm 1.5\times$ the IQR) are displayed in each violin plot.



Supplementary Figure 16. Effect size distributions of the trait-associated SGVs and SVs. a) Distribution of the proportions of variance in a phenotype explained by individual SVs and SGVs. We compared the variance explained between SVs and SGVs. For ease of interpreting the results, we focused only on the quantitative traits from the UKB. Variance explained by an LD-clumped SV was as computed based on the equation: $\frac{2p(1-p)b^2}{var(y)}$, where p is the MAF and b is the effect of a genetic variant on the phenotype, and $var(y)$ is the variance of the phenotype. b) Distribution of the absolute values of the effects of SVs and SGVs. c) The scatter plot of the effect of a genetic variant against its allele frequency. A total of 420 quantitative traits with significant SV associations were included. The variants were clumped with a P -value threshold of $5e-8$ ($m_{SV} = 16,544$; $m_{SGV} = 85,912$). The boxplots display the median, upper quartile, and lower quartile; whiskers extending to $1.5 \times$ the IQR; outliers are defined as $\pm 1.5 \times$ the IQR.



Supplementary Figure 17. Distribution of the number of traits associated with VNTRs and the distance to the splice site (SS) or transcriptional start site (TSS) of the trait-associated VNTRs. a) Distribution of the number of traits associated with VNTRs. b) Distribution of the distance of a VNTR to the nearest splice site (SS) and transcriptional start site (TSS). The VNTRs were grouped into trait-associated VNTRs ($m = 1,861$) and other VNTRs ($m = 14,587$) based on a cutoff $P_{\text{VNTR-GWAS}}$ of $5e-8$. P-values from a two-sided Wilcoxon rank-sum test are labeled. Each boxplot displays the median, upper quartile, and lower quartile; whiskers extending to $1.5\times$ the IQR; outliers are defined as $\pm 1.5\times$ the IQR.



Supplementary Figure 18. Matching the *GGT1* and *TRIM58* VNTR motifs with known motifs. a) The *GGT1* VNTR motif, (AT)₃A, was aligned against a compendium of RNA-binding motifs⁴³ (https://hugheslab.ccb.utoronto.ca/supplementary-data/RNAcompete_eukarya/) using MEME-TOMTOM with the default settings⁴⁴. The RBMS binding site motif (ID: RNCMPT00057) was identified with an alignment P-value of 1.97e-04 and an E-value of 0.02. The E-value is from a built-in compensation process of TOMTOM after correcting for multiple testing by multiplying the P-value by twice the number of target motifs in the database. b) The *TRIM58* VNTR motif was aligned against the motifs in the HOCOMOCO core database⁴⁵. The POU1F1 binding site motif, GTATATTTCATGAG (ID: PIT1_HUMAN.H11MO.0.C), was found to be embedded entirely in the *TRIM58* VNTR motif, with an alignment P-value of 1.18e-05 and E-value of 0.0042.

Supplementary References

- 1 Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426-3440 e3419 (2022).
- 2 Gustafson, J. A. *et al.* High-coverage nanopore sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of human genetic variation. *Genome Res* **34**, 2061-2073 (2024).
- 3 Noyvert, B. *et al.* Imputation of structural variants using a multi-ancestry long-read sequencing panel enables identification of disease associations. *medRxiv*, 2023.2012.2020.23300308 (2025).
- 4 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).
- 5 Workman, R. E. *et al.* Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat Methods* **16**, 1297-1305 (2019).
- 6 Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372** (2021).
- 7 Mukamel, R. E. *et al.* Repeat polymorphisms underlie top genetic risk loci for glaucoma and colorectal cancer. *Cell* **186**, 3659-3673 e3623 (2023).
- 8 Li, H. *et al.* A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat Methods* **15**, 595-597 (2018).
- 9 Heller, D. & Vingron, M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* **36**, 5519-5521 (2021).
- 10 Liao, W. W. *et al.* A draft human pangenome reference. *Nature* **617**, 312-324 (2023).
- 11 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).
- 12 Wang, T. *et al.* The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437-446 (2022).
- 13 Zook, J. M. *et al.* A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol* **38**, 1347-1355 (2020).
- 14 Logsdon, G. A. *et al.* Complex genetic variation in nearly complete human genomes. *Nature* **644**, 430-441 (2025).
- 15 Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284-1287 (2016).
- 16 Kowalski, M. H. *et al.* Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet* **15**, e1008500 (2019).

- 17 Wagner, J. *et al.* Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat Biotechnol* **40**, 672-680 (2022).
- 18 Olson, N. D. *et al.* PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions. *Cell Genom* **2** (2022).
- 19 Halldorsson, B. V. *et al.* The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732-740 (2022).
- 20 Visscher, P. M. *et al.* Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. *PLoS Genet* **10**, e1004269 (2014).
- 21 Lu, T. Y., Smaruj, P. N., Fudenberg, G., Mancuso, N. & Chaisson, M. J. P. The motif composition of variable number tandem repeats impacts gene expression. *Genome Res* **33**, 511-524 (2023).
- 22 Sim, S. B., Corpuz, R. L., Simmonds, T. J. & Geib, S. M. HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. *BMC Genomics* **23**, 157 (2022).
- 23 Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170-175 (2021).
- 24 Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. & Gurevich, A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**, i142-i150 (2018).
- 25 Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 245 (2020).
- 26 Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212 (2015).
- 27 Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* **27**, 849-864 (2017).
- 28 Martin, M., Ebert, P. & Marschall, T. Read-Based Phasing and Analysis of Phased Variants with WhatsHap. *Methods Mol Biol* **2590**, 127-138 (2023).
- 29 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).
- 30 Krusche, P. *et al.* Best practices for benchmarking germline small-variant calls in human genomes. *Nat Biotechnol* **37**, 555-560 (2019).
- 31 Yang, J., Zeng, J., Goddard, M. E., Wray, N. R. & Visscher, P. M. Concepts, estimation and interpretation of SNP-based heritability. *Nat Genet* **49**, 1304-1310 (2017).

- 32 Wainschtein, P. *et al.* Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat Genet* **54**, 263-273 (2022).
- 33 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).
- 34 Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res* **48**, D882-D889 (2020).
- 35 Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015).
- 36 Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**, 300-307 (2021).
- 37 Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* **13**, 508-514 (2016).
- 38 Wang, Y. *et al.* The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol* **19**, 151 (2018).
- 39 McArthur, E. & Capra, J. A. Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. *Am J Hum Genet* **108**, 269-283 (2021).
- 40 Kirsche, M. *et al.* Jasmine and Iris: population-scale structural variant comparison and analysis. *Nat Methods* **20**, 408-417 (2023).
- 41 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).
- 42 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-580 (1999).
- 43 Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172-177 (2013).
- 44 Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**, W202-208 (2009).
- 45 Kulakovskiy, I. V. *et al.* HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res* **46**, D252-D259 (2018).