Imputation of structural variants using a multi-ancestry long-read sequencing panel enables identification of disease associations

Boris Noyvert[1*], A Mesut Erzurumluoglu[1*], Dmitriy Drichel[2*], Steffen Omland[2*], Till F M Andlauer[1,3*], Stefanie Mueller[1], Lau Sennels[2], Christian Becker[2], Aleksandr Kantorovich[2], Boris A Bartholdy[1], Ingrid Brænne[1], Julio Cesar Bolivar-Lopez[1], Costas Mistrellides[1], Gillian M Belbin[4], Jeremiah H Li[4], Joseph K Pickrell[4], Johann de Jong[1], Jatin Arora[1], Yao Hu[1], **Boehringer Ingelheim – Global Computational Biology and Digital Sciences**, Clive R Wood[5], Jan M Kriegl[1], Nikhil Podduturi[2], Jan N Jensen[1], Jan Stutzki[2+], Zhihao Ding[1+]

1. Global Computational Biology and Digital Sciences (gCBDS), Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany
2. BI X GmbH, Ingelheim am Rhein, Germany
3. Department of Neurology, School of Medicine, Technical University of Munich, Munich, Germany
4. Gencove, New York, USA
5. Discovery Research, Boehringer Ingelheim Pharma GmbH & Co. KG, Ingelheim am Rhein, Germany

*: These authors contributed equally to this work
+: Joint senior authors

*Full list of **Boehringer Ingelheim – Global Computational Biology and Digital Sciences** authors listed in **Supplementary Information***

Abstract

Advancements in long-read sequencing technology have accelerated the study of large structural variants (SVs). We created a curated, publicly available, multi-ancestry SV imputation panel by long-read sequencing 888 samples from the 1000 Genomes Project. This high-quality panel was used to impute SVs in approximately 500,000 UK Biobank participants. We demonstrated the feasibility of conducting genome-wide SV association studies at biobank scale using 32 disease-relevant phenotypes related to respiratory, cardiometabolic and liver diseases, in addition to 1,463 protein levels. This analysis identified thousands of genome-wide significant SV associations, including hundreds of conditionally independent signals, thereby enabling novel biological insights. Focusing on genetic association studies of lung function as an example, we demonstrate the added value of SVs for prioritising causal genes at gene-rich loci

compared to traditional GWAS using only short variants. We envision that future post-GWAS gene-prioritisation workflows will incorporate SV analyses using this SV imputation panel and framework.

Main Text

Human disease associations of single nucleotide variants (SNVs) and short insertions and deletions are routinely identified in genome-wide association studies (GWASs)[1,2]. By contrast, large structural variants (SVs) of >50 base pairs (bp) are typically neglected, despite functional roles in the context of disease. Each human carries 23,000-31,000 SVs[1], often overlapping protein-coding genes or regulatory regions, thus enabling fine-mapping causal genetic variants[2,3]. Studies on single populations have already demonstrated the value of sequencing SVs for identifying causal variants underlying disease associations[4,5].

Robust SV calling from traditional short-read sequencing is challenging because SVs are often longer than the average short-read length (**Figure 1a**)[3,6]. Long-read sequencing captures SVs reliably, but the high cost impairs its application to large-scale datasets. Reference panels constructed for imputation of SVs from genotyped samples enable biobank-scale genome-wide analyses of SVs. For example, imputing SVs in UK Biobank (UKB) can accelerate research on the genetic underpinnings of diverse diseases and facilitate the identification of novel therapeutic targets. To this end, we generated a publicly available multi-ancestry long-read sequencing-based SV imputation panel (**Figure 1b**). Simultaneously, an effort is ongoing to develop new methods for improved SV calling that utilises this dataset (Schloissnig *et al*., in preparation).

Characterization of SV diversity

We performed long-read whole-genome sequencing of 906 individuals sampled from the 1000 Genomes Project, with median read length of ~6.2 kbp (**Supplementary Table 1**) and 15x median coverage (alignment with minimap2; **Supplementary Table 2, Supplementary Figure 9**). The 888 samples passing quality control (QC) included 164 Europeans, 144 (Admixed) Americans, 168 East Asians, 171 South Asians, and 241 Africans (**Figures 2a** and **2b; Supplementary Table 3**).

Joint (multi-sample) calling was performed using Sniffles2 v2.0.7[7]. We identified 107,445 SVs passing quality control (**Supplementary Table 4**). The Genome in a Bottle (GIAB) consortium described which genomic regions confidently produce reliable genotypes[8], and 41.9% of our SVs mapped to high-confidence ('confident' henceforth) regions. The most common SV types were insertions (55.8%), followed by deletions (35.8%), inversions (5.3%), breakends (typically unresolved SVs; 2.5%), and duplications (<0.06%) **(Figure 2c)**. Most SVs were rare (minor allele frequency (MAF) <0.005: 41.8%), the remainder being low-frequency (MAF 0.005-0.05: 28.7%) and common (MAF >0.05: 29.6%) variants, with the latter being enriched for insertions and deletions.

Sizes of most SV deletions and insertions are between 50 and 1000bps (**Figure 2d**). Duplications and inversions predominantly range from 1 to 30 kbp. A considerable number of SVs extend up to 1 Mbp and beyond. However, SVs that are longer than 1 Mbp are likely to be artifacts of the SV calling process (see **Extended Methods** section).

In a comparison of our long-read-based SV calls to short-read-based SV calls previously generated by a high-depth resequencing effort of the 1000 Genomes Project[3,9], we observed high recall rates for common SVs. Namely, we observed an overall recall rate of 74.8% for variants with MAF>5% within the same 888 samples in the 1000 Genomes dataset, with higher recall rates for insertions/duplications (75.4%) compared to deletions (74.7%) and inversions (21.1%). Additionally, we observed higher recall rates for variants within the GIAB-defined high-confidence regions compared to those outside those regions (see

Online Methods; **Supplementary Figures 1** and **2**). Overall, we identified 79,377 SVs that were absent in the short-read-based 1000 Genomes dataset.

Of the 107,445 SVs, 4,406 SVs were predicted using SnpEff[10] to have high functional impact. Of these, 1,198 are predicted to cause frameshifts, and 581 the complete loss of exons. Based on GWAS Catalog[11], 2,465 SVs overlapped positions of variants with known GWAS associations (**Supplementary Table 5**).

On average, individuals carried 16,065 SVs, with individuals of African ancestry exhibiting the highest (mean: 18,822 SVs), and East Asian-ancestry individuals exhibiting the lowest (14,729) SV diversity (**Figure 2e, Supplementary Table 6**)[3,12]. This observation was most pronounced for insertions and deletions, which are more frequent and can be called with high confidence. Only 36.6% (39,273) of SVs were shared across all superpopulations with individuals of African ancestry exhibiting the most SVs not shared with other populations (13,153) and (Admixed) Americans the least (841) (**Supplementary Table 7**).

Generation of an SV imputation panel and application to UKB

To enable imputation of SVs in other genotyped studies, we constructed a haplotype reference panel by integrating the SVs generated in the present study with the ~45M variants from the 1000 Genomes Project Phase 3 release[13] present in the 888 individuals (see **Online Methods**).

We assessed the accuracy of SV imputation using this resource by performing leave-one-out imputation for all individuals in the panel. Here, we specifically assessed the quality of SV imputation in UKB by using all genotyped UKB SNVs present in the reference panel (~702K SNVs) as the basis for imputation. To facilitate benchmarking, we not only imputed SVs but also ~57K randomly selected SNVs from the panel. We calculated the per-variant non-reference genotype concordance[14] and the imputation quality metric r^2_{imp} (see **Figure 3b**; **Online Methods**; **Supplementary Table 9**). Both metrics varied based on MAF, GIAB region type, and variant type. Generally, rare SVs showed poorer imputation quality than common insertions and deletions, which produced high-quality imputation metrics. The mean non-reference concordance for common insertions and deletions was 0.718 and 0.721, respectively, with mean r^2_{imp}=0.921 and 0.924 (which was even higher for SVs in confident regions, see **Figure 3b**). These metrics demonstrate sufficient imputation quality for conducting GWAS of common variants. As anticipated, the imputation quality of SVs was lower than that of imputed SNVs across GIAB region types and MAF classes, due to the greater difficulty in reliably calling SVs; however, this difference was not substantial (**Figure 3b**).

We also computed the per-individual non-reference genotype concordance for each superpopulation, based on common SVs and SNVs stratified by GIAB region type (**Figure 3a**; **Supplementary Table 8**). African-ancestry individuals, being the most diverse, had a mean non-reference concordance of 0.69 based on all common SVs, which was the lowest amongst all superpopulations. By comparison, individuals of European ancestry showed a mean concordance of 0.760 for common SVs. Overall, we did not observe significant outliers in the per-individual concordance values, indicating that there were no issues with sequencing and data processing of the samples. In conclusion, the SV reference panel offers a robust foundation for imputing SVs, particularly common insertions and deletions, into UKB and for performing subsequent GWAS.

Error! Reference source not found.We used our imputation panel to impute SVs into the genomes of 488,130 UKB participants. To this end, genotyped UKB SNVs were lifted to the GRCh38 human genome assembly. Of these SNVs, 92.19% were present in the 1000 Genomes reference panel. Imputation quality in UKB mirrored our observations from the leave-one-out validation in the 1000 Genomes study: r^2_{imp} was higher for common variants and in high-confidence regions. For example, common deletions and insertions in high-confidence regions showed mean r^2_{imp}=0.915, compared to 0.961 for common SNVs in the same regions (**Supplementary Tables 10** and **11**; **Figure 3c**). Notably, UKB predominantly consists of European-ancestry individuals, for which we observed higher average imputation qualities in the 1000 Genomes leave-one-out analyses.

Proof-of-principle SV imputation and genome-wide association studies in UK Biobank

To demonstrate the added value of the imputation panel, we selected 19 exemplary continuous traits and 13 binary traits available in UKB relevant to respiratory (n=6 traits), metabolic (n=16), and liver (n=10) diseases (**Supplementary Tables 12** and **13**). On these phenotypes, we performed GWASs of the imputed SVs in up to 453,754 European UKB participants[15].

Filtering variants by imputation metric INFO>0.7 and MAF>0.01, 3,858 SV associations (in 1,898 unique SVs) passed the established genome-wide significance threshold of $p<5\times10^{-8}$ in any phenotype (**Supplementary Table 14**). For an in-depth assessment of potential causal genes, we selected all significantly associated SVs from these GWASs that overlapped functional protein-coding genes (some SVs spanned several genes), thereby prioritizing 689 unique genes (**Figure 4a**; **Supplementary Table 14**; **Online Methods**). We also analysed how SVs influence the levels of 1,463 proteins in blood plasma. Here, we identified 10,518 significant ($p<5\times10^{-8}/1463=3.4\times10^{-11}$, INFO>0.7, MAF>0.01) SV-based pQTLs (3,723 unique SVs) for 1,101 proteins (**Supplementary Table 15**), including 84 (excluding the major histocompatibility complex region i.e. 6p21) that were conditionally independent of nearby SNVs (**Online Methods**; **Supplementary Table 16**).

Systematic analysis of SV information in identifying novel associations and causal genes

To follow up on the significant SV associations identified in the UKB GWASs, we examined whether these SV associations provide added value compared to analysing SNVs alone. To this end, we carried out traditional GWASs on imputed UKB SNVs derived from genotype data, using the same sample inclusion criteria and phenotype transformations as in the SV analyses. To identify whether SVs were the causal variants at the SNV-based GWAS and pQTL-associated loci, we combined the SNV and SV association results (using the filtering criteria above). At 55 genome-wide significant GWAS loci, an SV showed the lowest *p*-value (**Supplementary Table 17**). Conditional analyses revealed that SVs constituted a secondary, independent signal at 23 further loci; 38 of these 78 SVs overlapped with protein-coding genes.

We then systematically assessed the contribution of SV associations to the prioritization of causal genes by comparing genes implicated by SVs to the putatively causal genes reported by the latest GWAS of lung function measures published by Shrine *et al.*[16]. The authors used several post-GWAS approaches (e.g., nearest gene, e/pQTLs, polygenic priority score (PoPS, a tool for gene prioritisation[17]), rare respiratory disease causal genes) to map associated loci to genes, thereby prioritizing on average ~3 genes per locus. Of the reported autosomal loci associated with ≥1 of the three lung function phenotypes (*i.e.*, FEV_1, FVC, or FEV_1/FVC – the latter a clinically important lung function measure utilised in diagnosing chronic obstructive pulmonary disease), 70 harboured SVs (*i.e.*, SV mapped within ±500kb of the top SNV) significantly associated with the same primary lung function trait in our UKB-based GWASs

(**Supplementary Table 18**). At 55 of these 70 loci, the gene implicated by our SV analyses was also implicated by ≥1 of the post-GWAS methods utilised by Shrine *et al.* – strengthening the evidence for causality for the respective genes. In the remaining 15 loci where SV-implicated genes had not been prioritised yet, SVs pointed to genes very likely involved in lung health via influencing smoking behaviour (e.g., *SLC1A2* - highly expressed in the brain; **Supplementary Figure 3**) or other lung-disease causal mechanisms such as abnormal respiratory ciliary function[18] (e.g., *DNAH12* and *DYNLRB1*) and promotion of inflammation[19] (e.g., *PRDX1*).

Added value of SV information to gene prioritisation: examples from GWASs of lung function measures

In addition to the identification of the abovementioned genes, the added value of SV information for identifying the causal gene *i.e.*, improving locus-to-gene (L2G) prioritisation pipelines at the loci identified by Shrine *et al.*[16] can be demonstrated using four representative examples:

In our GWASs of quantitative lung function measures in UKB, SVs constituted the conditionally independent variant in primary or secondary signals of 14 loci, thereby facilitating identification of the respective causal genes (**Supplementary Table 17**). One of these FEV_1/FVC-associated loci contained an 841-base SV deletion in an intron of *CFDP1* (Sniffles2.DEL.3639MF; $p=1.1x10^{-65}$; MAF=0.413; **Figure 4b**). Shrine *et al.* prioritized three potentially causal genes at this locus (including *CTRB1* and *BCAR1*), with *CFDP1* only being implicated by its proximity to the top SNV (rs11864587) without any functional or regulatory support. Follow-up analyses on this gene, including phenome-wide cis-eQTL-based Mendelian Randomisation (MR) and colocalization analyses (not carried out by Shrine *et al.*) provided strong evidence (min MR $p=1.1\times10^{-22}$; colocalisation posterior probability (PP) >90%) for putatively causally linking *CFDP1* expression in various (bulk GTEx) tissues including fibroblasts – a cell type relevant for respiratory disease – to a lower FEV_1/FVC measure (**Supplementary Figure 4**; **Supplementary Table 19**). *CFDP1* and *BCAR1* protein expression levels were not assayed by the UKB-PPP[20], thus information from SV-based pQTLs could not be utilised to distinguish between the genes.

At a second FEV_1/FVC-associated locus (top-associated SNV: rs947350), Shrine *et al.* prioritized ten genes, including *MEGF6* with suggestive evidence from rare coding variants. In our study, an FEV_1/FVC-associated SV deletion (Sniffles2.DEL.6E8M0; $p=2.1x10^{-16}$; MAF=0.069) mapped only to *MEGF6*. Similar to the *CFDP1* example, phenome-wide MR and colocalization analyses provided strong evidence (min MR $p=6.3\times10^{-7}$; colocalization PP>97%) putatively causally linking *MEGF6* expression in various (bulk GTEx) tissues including skeletal muscle and heart tissue (a likely respiratory disease-relevant cell type as a smooth muscle-containing tissue) to a lower FEV_1/FVC measure (**Supplementary Figure 5**; **Supplementary Table 20**).

Third, *AAGAB* was solely implicated by two different FVC-associated SV deletions (Sniffles2.DEL.268AME; $p=1.6x10^{-12}$; MAF=0.231; and Sniffles2.DEL.2689ME; $p=2.6x10^{-12}$; MAF=0.233) but not by any of the post-GWAS methods utilised by Shrine *et al*. Similar to the above examples, phenome-wide MR and colocalization analyses provided strong evidence (min MR $p=1.7\times10^{-15}$; colocalization PP>90%) putatively causally linking *AAGAB* expression in various (bulk GTEx) tissues, including lung tissue, directly to lung function measures (**Supplementary Figure 6**; **Supplementary Table 21**).

As a final example, Shrine *et al.* identified rs7108992 to be associated with FVC and FEV_1. They prioritized two genes, *ETS1* (by proximity) and *FLI1* (via PoPS). In our analysis, we identified an FVC-associated SV

deletion only mapping to *FLI1* (Sniffles2.DEL.5C9EMA; p=2.0x10^{-30}; MAF=0.322). Phenome-wide MR and colocalization analyses strongly linked *FLI1* expression in lung and smooth muscle-containing tissues to lung function measures (MR p=3.9×10^{-4}; colocalization PP=96%), pulmonary heart disease risk (MR p=8.8×10^{-7}; colocalization PP=80%), *FGF10*[21] (MR p=7.5×10^{-5}; colocalization PP=89%), and *LRP1* protein levels[22] (MR p=2.8×10^{-5}; colocalization PP=87%) – known factors contributing to respiratory diseases[21] (**Supplementary Figure 7**).

Taken together, these examples illustrate the potential added value of SV information for the identification of novel gene-disease associations, and for improving gene prioritisation pipelines applied to GWAS summary statistics (e.g., the composite locus-to-gene (L2G) score calculated in Open Targets Genetics[23]).

Discussion

Long-read sequencing holds the promise of conducting reliable association studies of SVs in large cohorts, but its widespread adoption is impeded by its significant cost. For example, comprehensive long-read sequencing of all UKB participants would cost approximately 0.5 billion USD, based on an estimate of 1,000 USD per whole genome. It was recently suggested to reduce costs by sequencing only a limited number of SVs[24]. By contrast, our SV imputation approach allows for analyses of a comprehensive, genome-wide SV panel without additional sequencing costs. Therefore, use of an SV imputation panel constitutes a practical and cost-effective solution for the robust analysis of common SVs. The multi-ancestry imputation panel applied in the present study to Europeans from UKB can also be used to impute SVs in diverse ancestries, e.g., from BioBank Japan[25], Qatar Biobank[26], China Kadoorie Biobank[27], or the Singapore Precision Medicine Programme[28].

We observed that the quality of SV calling, and imputation is strongly stratified by variant type, frequency, and the complexity of genomic regions. Using genome stratification files provided by the GIAB consortium, we identified robust SV calls and showed that deletions and insertions with a MAF>0.05 were most reliable. These insights motivate future refinement of variant stratification methods.

We demonstrated the utility of our SV imputation panel by imputing SVs in UKB and conducting GWASs on 32 disease-relevant traits. Here, 907 SVs were significantly associated, mapping to 689 functional protein-coding genes. We also ran GWAS for expression levels of ~1.5k proteins and identified 1720 significant SVs overlapping with 1197 genes. Using selected examples, we highlighted the relevance of the identified potentially causal genes to respiratory diseases.

Our imputation panel facilitates analyses of SVs overlapping exons, splice-sites, promoters, or enhancers of protein-coding genes at GWAS-associated loci. Therefore, it has the potential to become a routine component of post-GWAS gene prioritization workflows. We also envisage that our SV imputation panel will enable diverse applications of integrating SVs with other *omics* data, e.g., in machine learning-based frameworks. For example, genome-wide SV-based features could be used in models for high-throughput post-GWAS gene prioritization[27], disease subtyping/precision medicine[29,30], and drug response/adverse event prediction[30].

As next steps in applying our SV panel resource, we suggest, first, to impute SVs in other international biobanks, making more SV data available to the research community. Second, to carry out GWASs utilizing SVs to identify novel associations and disease-causal genes. Finally, to carry out comprehensive analyses of the effects of SV inversions (e.g., spanning transcription factor binding sites), translocations, and

duplications (e.g., to analyse dosage-dependent effects of SNVs affected by the duplications) for which more research is needed. Thereby, this resource will strongly advance our knowledge of the genetic underpinnings of disease.

Figures

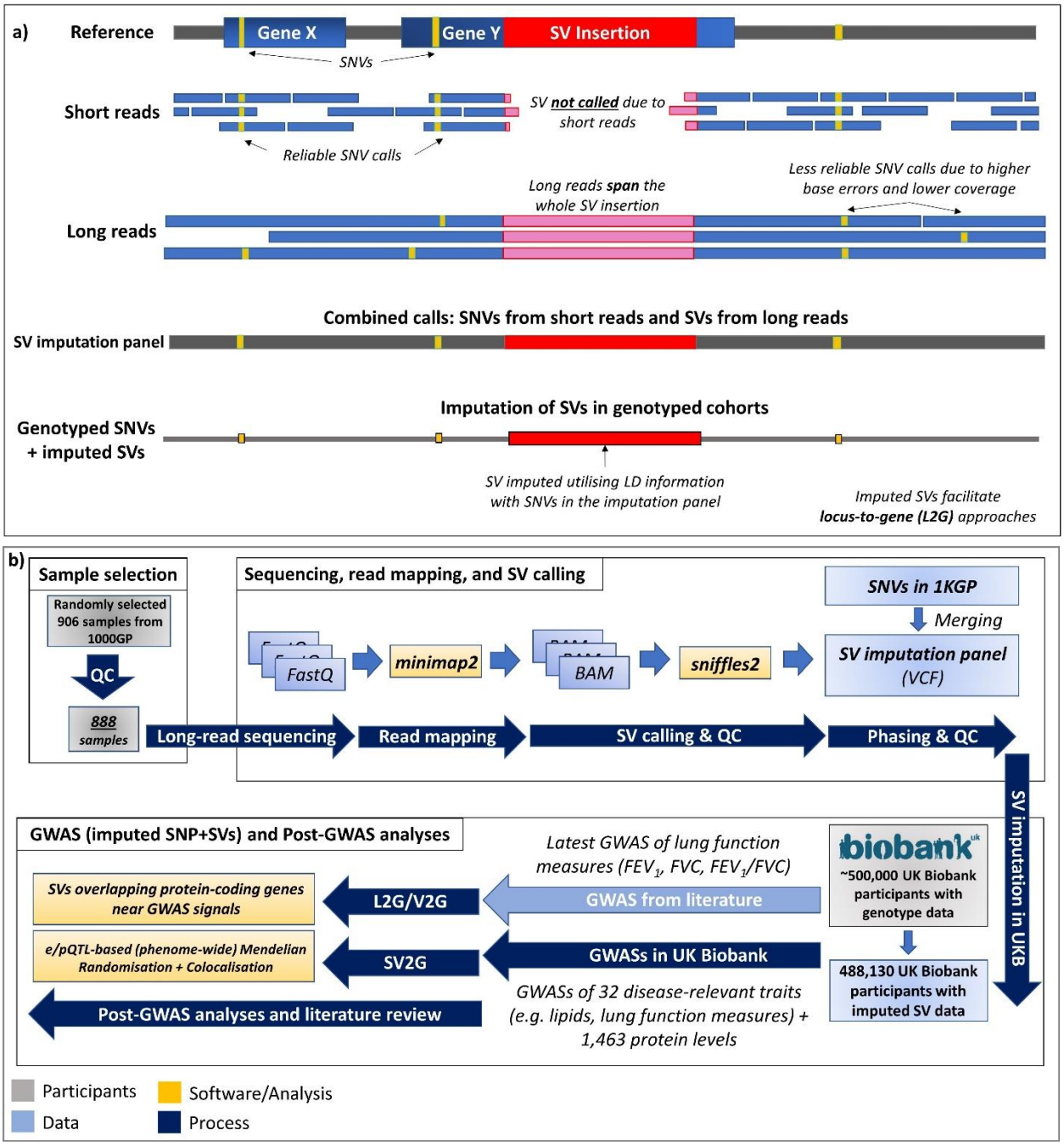

Figure 1. ***Study design and visual abstract of purpose and benefits of long-read sequencing and a multi-ancestry imputation panel.*** *a) Long-reads that span large SVs enable identification of novel SVs and*

improve calling of known SVs, which better inform locus-to-gene (L2G) approaches. b) Study design including the sample selection, SV calling, SV imputation in UK Biobank, GWAS, and Post-GWAS stages (not exhaustive). For the GWASs in UK Biobank, separate analyses using the same phenotype definitions were carried out using (i) the imputed SNVs, and (ii) imputed SVs. The two GWAS summary statistics were then combined for the relevant post-GWAS analyses. L2G/V2G: locus-to-gene/variant-to-gene (carried out on top SNVs identified by Shrine et al, which have an SV deletion within 500kb); SV2G: SV-to-gene (carried out ***only*** *on SV deletions that were the most significant variants in the associated loci in the GWASs we carried out in UK Biobank, which also overlap with protein-coding genes); 1KGP: 1000 Genomes Project (Phase 1); UKB: UK Biobank.*

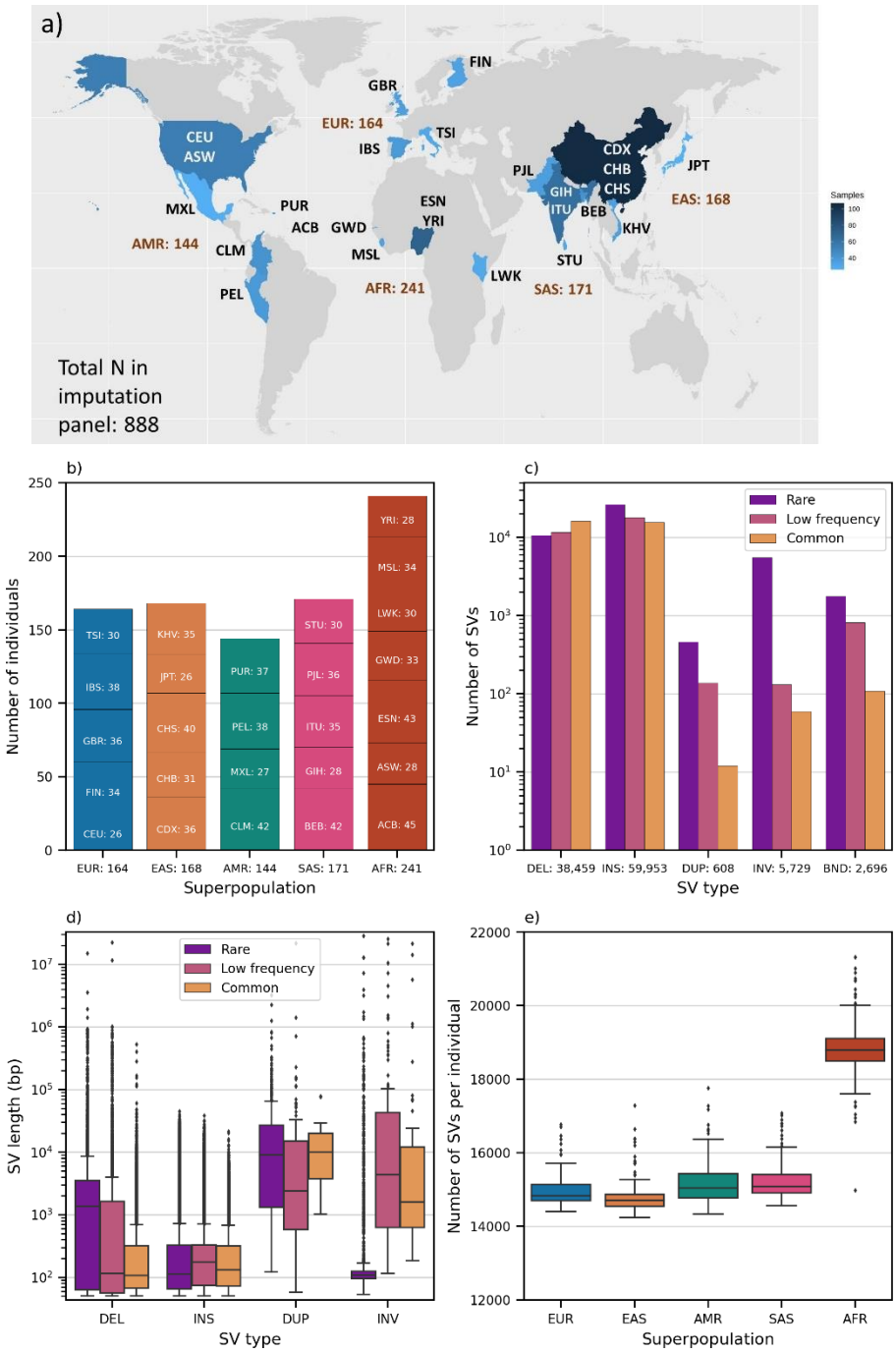

Figure 2. **Characterization of the quality-controlled imputation panel** (N=888 samples and n=107,445 SVs). **a)** A multi-ancestry, long-read sequencing-based imputation panel enables robust SV imputation in all biobanks, including UK Biobank – which we utilised for proof-of-concept. AMR: Admixed American ancestry; AFR: African ancestry; EUR: White European ancestry; EAS: East Asian ancestry; SAS: South Asian ancestry; ACB: African Caribbean in Barbados; MSL: Mende in Sierra Leone; GWD: Gambian in Western Division – Mandinka; PUR: Puerto Ricans in Puerto Rico. **b)** Sample counts by superpopulation and population codes (population code description in **Supplementary Table 1**), using abbreviations from 1000 Genomes Project. **c)** Number of SVs by variant type (DEL: deletion, INS: insertion, DUP: duplication, INV:

inversion, BND: breakend) and frequency class (common: 0.05<MAF≤0.5, low frequency: 0.005<MAF≤0.05, rare: 0<MAF≤0.005), with total SV counts by SV type. **d)** SV size distributions by SV type and frequency class, excluding breakends, which do not have a size. **e)** Number of minor SV alleles per individual by superpopulation.

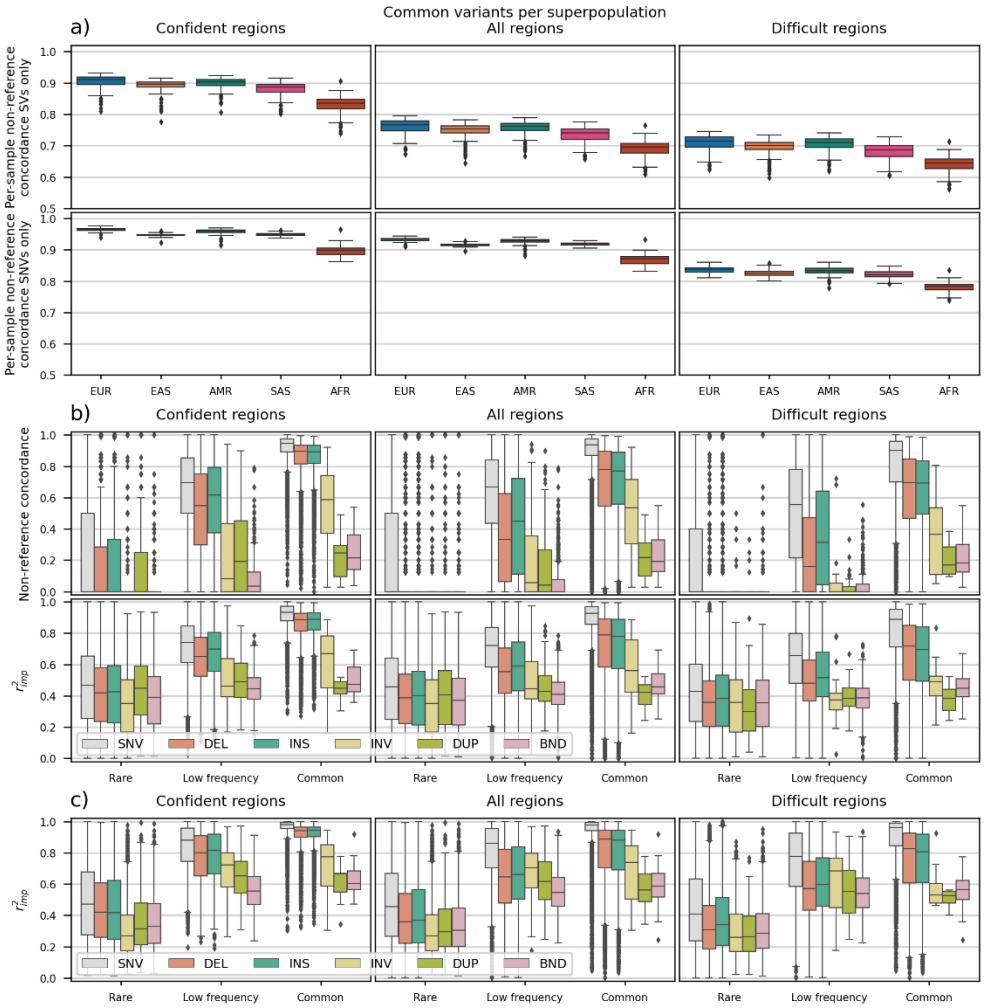

Figure 3. **Evaluation of imputation quality: a)** and **b)** Imputation metrics from leave-one-out cross-validation by MAF frequency class, for high-confidence (confident) regions, all regions, and difficult regions. **a)** Sample-wise non-reference concordance by superpopulation, using only imputed common SVs and only imputed common SNVs. Note the range of the y-axis. **b)** Variant-wise non-reference concordance by variant type. **c)** Assessment of imputation quality in imputed UK Biobank data. Quality is assessed using r^2_{imp} score in UK Biobank for imputed SNVs and SVs, by variant type and MAF class, for confident regions, all regions, and difficult regions.

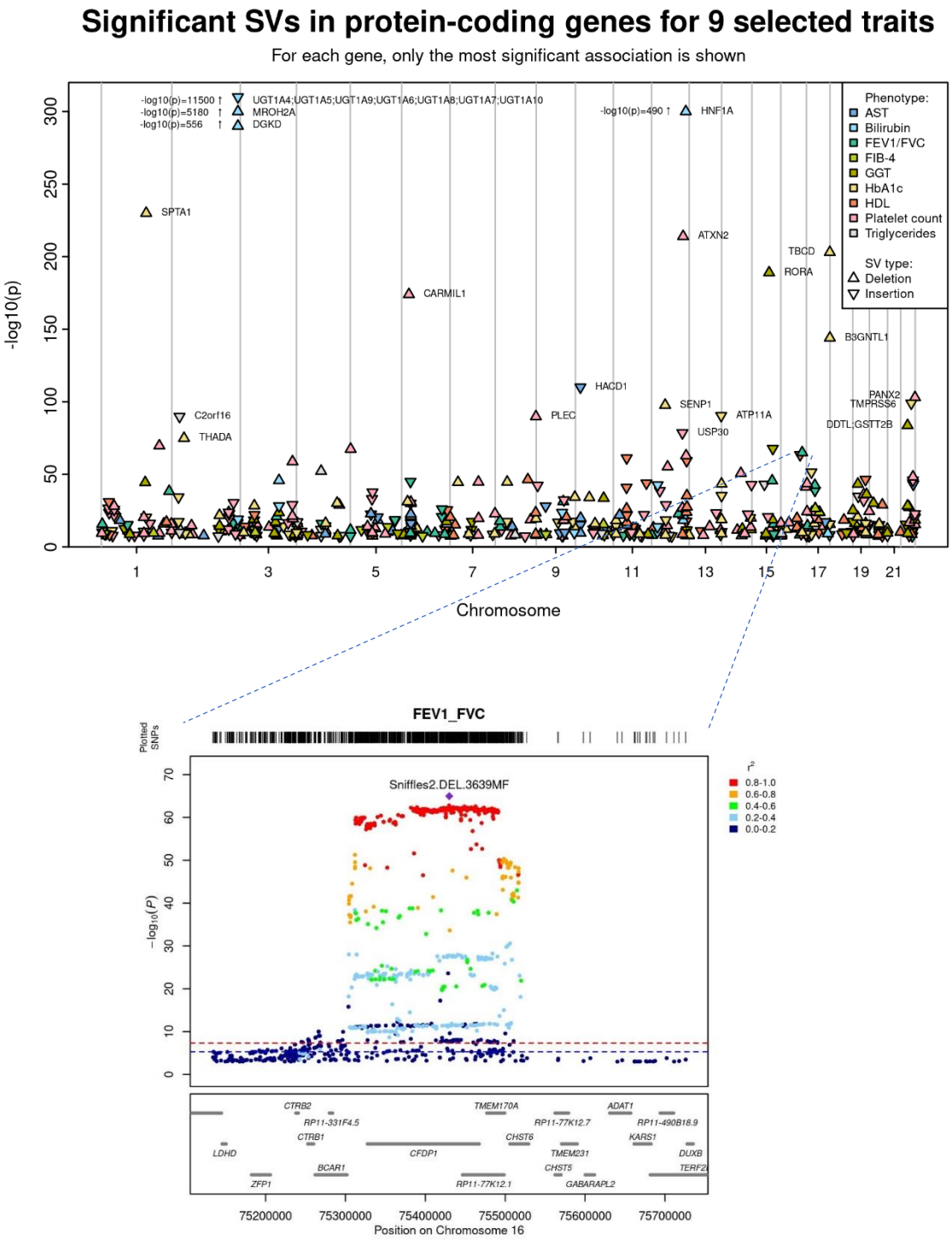

Figure 4. **a)** Manhattan plot of 9 selected traits, with only GW-significant SVs overlapping with protein-coding genes shown. Phenotype definitions and full list of GW-significant SVs are available in **Supplementary Tables 12-13** and **14**, respectively. **b)** Region plot showing the FEV_1/FVC association of Sniffles2.DEL.3639MF (841-base SV deletion in an intron of *CFDP1*) and nearby SNVs in a GWAS of UK

Biobank participants. LD information between the variants was calculated using the same population utilised in the UKB GWASs. Additional details on the SV can be found in **Supplementary Table 5**.

Online Methods

Building the SV imputation panel

Oxford nanopore long-read sequencing

Sample sequencing was executed using Oxford Nanopore Technologies' (ONT) PromethION P48. After DNA extraction, libraries were prepared as per ONT ligation sequencing kit SQK-LSK110 protocols with r9.4.1 flowcells (further details can be found in the Qiagen Gentra Puregene Handbook).

For basecalling, we used guppy v6.0.1 with the super high accuracy model (SUP). Sequencing statistics for the individual fastq files were generated with the NanoStat[31] software (**Supplementary Table 1; Supplementary Figures 8** and **9**).

Alignment and calling

The alignment was performed against the GRCh38 assembly using minimap2 v2.24[32] and recommended default parameters for Oxford Nanopore long-reads ("-ax -map-ont").

Alignment metrics were computed with Picard's CollectWgsMetrics tool, with modified parameters to adapt to Oxford Nanopore reads and our analysis setup. Specifically, Picard was instructed to count unpaired reads ("--COUNT_UNPAIRED true"), all base qualities ("--MINIMUM_BASE_QUALITY 0"), and only coverages at sites in the reference on autosomal chromosomes (interval file via "--INTERVALS"). For descriptive statistics see **Supplementary Table 2**.

Joint variant calling was conducted with Sniffles2 v2.0.7, supplemented by tandem repeat annotations to improve variant calls in these regions, using the default annotation file provided by the tool's authors ("--tandem-repeats human_GRCh38_no_alt_analysis_set.trf.bed").

Imputation panel QC

After calling, raw variants were normalized with bcftools v1.15.1[33] ("bcftools norm") and SVs smaller than <50bp were removed. After extensive sample QC, we also excluded private SVs (singletons and private doubletons), SVs with genotype missing rates ≥0.2, and with very large apparent sizes >30Mbp (see **Supplementary Table 22** and **Supplementary Figures 10-12** for statistics on raw and post-QC SV calls). In total, 107,445 SVs remained in the imputation panel after QC (see **Supplementary Information** for extended methods).

We performed extensive sample identity verification: For each sample, we counted nanopore reads supporting each allele at the ~12,000 common (MAF 0.45-0.55) SNV sites across the genome. To verify sample identity, we compared read ratios to the publicly available SNV genotypes of the 1000 Genomes Project individuals. We identified two samples to be swapped during sequencing and seven samples to be contaminated with DNA from another individual. The swapped samples were corrected in the panel and the contaminated samples were removed. In addition, seven descendants of other individuals present in the panel and two samples with high read duplication levels were excluded from the imputation panel. Finally, one sample was found to be not from the 1000 Genomes Project and one sample was removed due to very low leave-one-out metrics indicating sequencing problems for this sample. In total, we

removed 18 samples resulting in the final panel of 888 individuals (see **Supplementary Information** for further details).

Panel imputation and phasing

Beagle was used for phasing (Beagle5) and imputation (Beagle4) of sporadically missing genotypes in the panel, which allowed us to generate the final reference panel calls. Default parameters were used except for the "--gtgl" flag instead of "--gl" during imputation because genotype likelihoods were absent for SVs. Beagle was run on chunks of 55,000 variants each, flanked by 3,000 variants up- and downstream.

Liftover of UK Biobank genotyped SNVs

The genomic positions of 805,426 directly genotyped UK Biobank variants were lifted over from GRCh37 to GRCh38 human genome assembly using the DNAnexus pipeline[34] based on Picard's[35] LiftoverVcf tool. 803,700 (99.8%) markers were successfully lifted over. 702,480 of 764,685 (91.87%) of the genotyped variants could be matched to variants in the 1000 Genomes Project, serving as the basis for SV imputation.

Preprocessing and imputation of UKB data

Imputation of SVs was performed in UK Biobank[15] under the approved research proposal 57952.

After liftover, genotype files were converted to VCF format with PLINK2 using the "--ref-from-fa" setting to ensure the order of REF and ALT alleles remained compatible with the GRCh38 genome assembly. These genotyped variants for 488,130 UKB individuals were phased using SHAPEIT4 v4.2.2[36].

Imputation was performed for each chromosome separately in batches of 10,000 individuals with Beagle5.4[37]. The r^2_{imp} score was re-computed based on the data from all individuals[38] (further details in **Supplementary Information**).

Genome-wide association study utilizing imputed structural variants in UKB

Sample selection

For the selection of samples with "White European" ancestry, we adhered to a strategy previously established in the literature[39,40]. In summary, we applied k-means clustering (with k=6) on the first two genetic principal components from the principal component analysis (PCA) results provided by UKB ("ukb_sqc_v2_pca.txt"). We excluded individuals who had withdrawn consent or did not map to the European cluster (**Supplementary Figure 8**). We identified 171 duplicate pairs among the selected samples using KING v2.3.0 (--related). For each pair, we excluded the sample with the most missing phenotypes of interest. In cases where no significant differences were found, we randomly selected a sample from the pair to exclude. This resulted in a dataset of 455,589 samples for association analyses.

Genome-wide association studies in UKB

GWASs and pQTLs were calculated in UKB using imputed SV dosages and a linear mixed model, incorporating a leave-one-chromosome-out whole-genome regression model to account for population stratification as implemented in REGENIE v3[41]. Covariates included in the analysis were sex, age, age^2, genotype array (UKB data field 22000), and 20 genetic principal components as determined by PCA. Protein levels were treated as quantitative phenotypes. For lung function-related traits (FEV_1, FVC, FEV_1/FVC) we added 'ever smoked' (data field 20160) as a binary covariate. PCA was computed for the set of individuals included in each GWAS or pQTL analysis with PLINK2 ("--pca approx 20") on pruned genotypes after exclusion of rare variants ("–maf 0.005 --indep-pairwise 200kb 0.5"). For all 19 quantitative phenotypes, the measurement at the first examination was used to maximise sample size.

Rank-based inverse normal transformation was applied to quantitative phenotype and protein level values. GWASs were checked for unaccounted genomic inflation via quantile-quantile plots and no extreme values were observed. Thus, P-values were not corrected for genomic inflation. We used Manhattan plots to visualize and manually check the GWAS results.

Signal selection and conditional analyses

For each phenotype and protein level, the significant SVs (GWAS p-value<$5x10^{-8}$; pQTL p-value<$5x10^{-8}$/1463) were split into sets with a genomic distance ≥100kb. For each set, we performed conditional analyses: First, SV coordinates were lifted from GRCh38 to GRCh37, then the SV imputed data was merged with the standard UKB short-variant imputed genotype data (data field 22828) in the region of the SV set ±500kb flanking regions on each side. Next, we ran an iterative conditional analysis on the merged dataset to select independent association signals, conditioning on the set of top variants and adding the identified independent variants in a stepwise manner, stopping when the p-value of the top variant reached >$5x10^{-6}$ (> $5x10^{-6}$/1463 for pQTLs) or after five iterations. We considered the SV as an independent and novel signal if it appeared as a top variant in any iteration of the conditional analysis.

Utilizing SV information to identify novel disease-relevant genes (SV2G) and improve post-GWAS L2G prioritisation

From identified SV associations to causal genes (SV2G)

We conducted the following steps to identify the putatively causal gene(s) in the associated regions where an SV overlapping a protein-coding gene was the top variant (see 'SV2G' in **Figure 1b**): a) Examination of the literature, b) cis-e/pQTL-based phenome-wide MR analyses (using the TwoSampleMR package[42]), c) finemapping and colocalization (coloc v5), for all genes implicated by the SV(s) at each locus. GCTA-COJO and LD pruning (r^2<0.01) were used to select cis-e/pQTLs (within ±1Mbp of the gene's transcription start site) as instrumental variables (IVs) for each exposure (list of exposures in **Supplementary Table 23a**). Where we identified a strong MR association between gene expression and a trait (e.g., between *MEGF6* expression and a lung-disease-relevant trait), we manually checked the regional plots to ensure that there was strong evidence for local colocalization between the cis-e/pQTL signals used as the MR instrument(s) and the outcome/trait GWAS (examples in **Supplementary Figures 3-7**). The outcome GWASs (n=7,429; **Supplementary Table 23b**) consisted mostly of a manually curated list from IEU OpenGWAS[43] but also from internally conducted GWASs on 44 endophenotypes and outcomes related to respiratory and/or fibrotic diseases[44,45]. A Bayesian method was used to finemap each associated locus to a set of variants that – assuming the causal variant was also included in the analysis – contains the underlying causal variant with 95% probability[46]. We set the parameter *W* (i.e., the variance of the prior distribution of effect sizes) to 0.04 (≈0.21^2) in the approximate Bayes factor formula – which equates to a 95% belief that the absolute relative risk is $<\frac{3}{2}$. To estimate the probability that a single variant explains both the cis-e/pQTL signal and the signal in the trait/outcome GWAS, we manually inspected the region plots in addition to using the "coloc.abf" function[47].

From SNV-based GWAS signals to causal genes utilizing SV information (L2G)

To systematically analyse the contribution of SV information to post-GWAS locus-to-gene (see 'L2G' in **Figure 1b**) prioritization approaches, we utilized the list of independent associations identified by the most recent GWAS of lung function carried out by Shrine *et al*[16] and the table of implicated genes (consisting of genes implicated by various post-GWAS methods such as e/pQTL association, PoPS[17], and rare respiratory disease causal genes) for each associated locus (Column AB in **Supplementary Table 18**). First, we lifted

over the SNV coordinates reported by Shrine *et al* (in hg19) to GRCh38 positions. We then looked for SVs that were (i) within 500kb of the independent SNPs reported by Shrine *et al*, (ii) significantly associated ($p<5x10^{-8}$ in our UKB-based SV-WAS) with the primary lung function measure reported by Shrine *et al*[16] (e.g. FEV_1/FVC for rs2355210), and (iii) which spanned protein-coding genes (as defined by GENCODE v43). As a pragmatic approach to focus on the SVs likely to be real and functional, we restricted the analyses to a set of well-imputed SVs (INFO metric >0.7) not mapping to 'difficult' regions.

Author contributions

Study conception and design: ZD, JS, SO, AME, BN, JNJ, NP, JK. Data governance/infrastructure, statistical analysis, and/or interpretation: BN, JS, SO, DD, AME, CM, JCBL, BAB, CB, LS, SM, AK, JHL, GMB, IB. Preparation of the manuscript: ZD, AME, BN, TFMA, DD, SO, JS, JKP, JHL, GMB, JdJ, JA, IB. All authors (incl. all under the 'gCBDS' banner) have critically reviewed and approved the final version of this paper, including the authorship statement.

Data availability

Long-read sequencing imputation panel will be made available via the **OpnMe** initiative of Boehringer Ingelheim GmbH (details: https://opnme.com/genomiclens). Imputed SVs of UK Biobank participants will be made available via **UKB RAP**. Full summary statistics for the (SV- and SNV-based) GWASs carried out in UK Biobank are available upon request.

Conflicts of interest statement

Boehringer Ingelheim, a privately-owned pharmaceutical company, funded this initiative. DD and LS are independent contractors and declared no conflicts of interest. GMB, JHL, and JKP are employees of Gencove and declared no conflicts of interest.

Ethics declarations

This research has been conducted using the UK Biobank Resource under Application Number 57952.

Acknowledgements

This research has been conducted using the UK Biobank, a major biomedical database (www.ukbiobank.ac.uk). We thank all UK Biobank and 1000 Genomes Project participants, without whom this project would not have been possible.

We express our gratitude to the MARVL Initiative – a collaboration between the Research Institute of Molecular Pathology (IMP), BI X and gCBDS (Boehringer Ingelheim). In particular: Siegfried Schloissnig, Klaus Ehrlinger, Julien Charest, Mila Asparuhova and Patrick Hüther for sequencing the samples.

We thank Gilean McVean and Jan Korbel for guidance during the project and providing critical feedback on the manuscript.

References

1. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* **15**, 461–468 (2018).

2. Jakubosky, D. *et al.* Properties of structural variants and short tandem repeats associated with gene expression and complex traits. *Nat Commun* **11**, 2927 (2020).

3. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).

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. Wu, Z. *et al.* Structural variants in the Chinese population and their impact on phenotypes, diseases and population adaptation. *Nat Commun* **12**, 6501 (2021).

6. Abel, H. J. *et al.* Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583**, 83–89 (2020).

7. Smolka, M. *et al.* Comprehensive Structural Variant Detection: From Mosaic to Population-Level. *Biorxiv* 2022.04.04.487055 (2022) doi:10.1101/2022.04.04.487055.

8. GIAB consortium, genome stratification files. https://github.com/genome-in-a-bottle/genome-stratifications.

9. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426-3440.e19 (2022).

10. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**, 80–92 (2012).

11. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).

12. Abel, H. J. *et al.* Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583**, 83–89 (2020).

13. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

14. Linderman, M. D. *et al.* Analytical validation of whole exome and whole genome sequencing for clinical applications. *Bmc Med Genomics* **7**, 20 (2014).

15. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

16. Shrine, N. *et al.* Multi-ancestry genome-wide association analyses improve resolution of genes and pathways influencing lung function and chronic obstructive pulmonary disease risk. *Nat Genet* **55**, 410–422 (2023).

17. Weeks, E. M. *et al.* Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *Nat. Genet.* **55**, 1267–1276 (2023).

18. Leigh, M. W. *et al.* Clinical and genetic aspects of primary ciliary dyskinesia/Kartagener syndrome. *Genet. Med.* **11**, 473–487 (2009).

19. Liu, D. *et al.* Proteomic Analysis of Lung Tissue in a Rat Acute Lung Injury Model: Identification of PRDX1 as a Promoter of Inflammation. *Mediat. Inflamm.* **2014**, 469358 (2014).

20. Sun, B. B. *et al.* Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).

21. Jiang, T. *et al.* Fibroblast growth factor 10 attenuates chronic obstructive pulmonary disease by protecting against glycocalyx impairment and endothelial apoptosis. *Respir Res* **23**, 269 (2022).

22. Nichols, C. E. *et al.* Lrp1 Regulation of Pulmonary Function. Follow-Up of Human GWAS in Mice. *Am J Resp Cell Mol* **64**, 368–378 (2020).

23. Mountjoy, E. *et al.* An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat Genet* **53**, 1527–1533 (2021).

24. Auwerx, C. *et al.* The individual and global impact of copy-number variants on complex human traits. *The Am. J. Hum. Genet.* **109**, 647–668 (2022).

25. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol* **27**, S2–S8 (2017).

26. Thani, A. A. *et al.* Qatar Biobank Cohort Study: Study Design and First Results. *Am J Epidemiol* **188**, 1420–1433 (2019).

27. Walters, R. G. *et al.* Genotyping and population structure of the China Kadoorie Biobank. *Medrxiv* 2022.05.02.22274487 (2022) doi:10.1101/2022.05.02.22274487.

28. Wong, E. *et al.* The Singapore National Precision Medicine Strategy. *Nat Genet* 1–9 (2023) doi:10.1038/s41588-022-01274-x.

29. Chandak, P., Huang, K. & Zitnik, M. Building a knowledge graph to enable precision medicine. *Sci Data* **10**, 67 (2023).

30. Jong, J. de *et al.* Towards realizing the vision of precision medicine: AI based prediction of clinical drug response. *Brain* **144**, 1738–1750 (2021).

31. Coster, W. D., D'Hert, S., Schultz, D. T., Cruts, M. & Broeckhoven, C. V. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).

32. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

33. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).

34. DNAnexus liftover_plink_beds. https://github.com/dnanexus-rnd/liftover_plink_beds.

35. Picard. https://broadinstitute.github.io/picard/.

36. 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).

37. Browning, B. L., Zhou, Y. & Browning, S. R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am J Hum Genetics* **103**, 338–348 (2018).

38. Das, S., Abecasis, G. R. & Browning, B. L. Genotype Imputation from Large Reference Panels. *Annu Rev Genom Hum G* **19**, 1–24 (2018).

39. Wain, L. V. *et al.* Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. *Nat Genet* **49**, 416–425 (2017).

40. Shrine, N. *et al.* New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat Genet* **51**, 481–493 (2019).

41. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet* **53**, 1097–1103 (2021).

42. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**, e34408 (2018).

43. Elsworth, B. *et al.* The MRC IEU OpenGWAS data infrastructure. *Biorxiv* 2020.08.10.244293 (2020) doi:10.1101/2020.08.10.244293.

44. Hemani, G., Tilling, K. & Smith, G. D. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet.* **13**, e1007081 (2017).

45. Bowden, J., Smith, G. D. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiology* **44**, 512–525 (2015).

46. Wakefield, J. A Bayesian Measure of the Probability of False Discovery in Genetic Epidemiology Studies. *Am J Hum Genetics* **81**, 208–227 (2007).

47. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, e1004383 (2014).