HGG Advances

ARTICLE

Landscape of genomic structural variations in Indian population-based cohorts: Deeper insights into their prevalence and clinical relevance

Krithika Subramanian,[1,2] Mehak Chopra,[1] and Bratati Kahali[1,3,*]

Summary

Structural variations (SV) are large (>50 base pairs) genomic rearrangements comprising deletions, duplications, insertions, inversions, and translocations. Studying SVs is important because they play active and critical roles in regulating gene expression, determining disease predispositions, and identifying population-specific differences among individuals of diverse ancestries. However, SV discoveries in the Indian population using whole-genome sequencing (WGS) have been limited. In this study, using short-read WGS having an average 42X depth of coverage, we identify and characterize 36,210 SVs from 529 individuals enrolled in population-based cohorts in India. These SVs include 24,574 deletions, 2,913 duplications, 8,710 insertions, and 13 inversions; 1.26% (456 out of 36,210) of the identified SVs can potentially impact the coding regions of genes. Furthermore, 56 of these SVs are highly intolerant to loss-of-function changes to the mapped genes, and five SVs impacting *ADAMTS17, CCDC40,* and *RHCE* are common in our study individuals. Seven rare SVs significantly impact dosage sensitivity of genes known to be associated with various clinical phenotypes. Most of the SVs in our study are rare and heterozygous. This fine-scale SV discovery in the underrepresented Indian population provides valuable insights that extend beyond Eurocentric human genetic studies.

Introduction

Structural variations (SVs) in the human genome are a diverse set of large regions of rearrangements in the DNA sequence spanning for more than 50 base pairs (bp), comprising unbalanced insertions, deletions, and duplications, and balanced classes of inversions and translocations.[1] SVs are widespread in the human genome, and the sequence length of SVs can extend well beyond several megabases, therefore being responsible for more nucleotide changes than other classes of sequence variations, for example, single nucleotide polymorphisms (SNPs), and short insertions and deletions (InDels) (length <50 bp) in the human genome.[2–4] Structural variations are known to be associated with human phenotypes and disease traits, for example, obesity, certain types of cancer, autism, schizophrenia, and cognitive dysfunction, among others, as well as play crucial roles in molecular and cellular processes and gene expression,[5–7] thus establishing that SV discovery and characterization are crucial in human health studies.

Globally, short-read whole-genome sequencing (WGS) is an established technology used in large-scale human cohort studies for identifying genetic variants and understanding their contributions to disease traits. Although initial discoveries of SVs were done by array-based genomic hybridization techniques[8,9] and single molecule or long reads sequencing technologies,[10,11] current SV detection approaches from high-coverage short-read WGS has facilitated large-scale population-level discovery of SVs in recent times with better size and breakpoint resolutions, albeit with some limitations.[12–18] The Indian population is extremely diverse, yet the population-level genetic makeup has been inferred primarily from SNP[19–21] and limited structural variation discovery studies.[18] Our study will help address this lacuna. In this study, we present a detailed analysis of structural variations identified through WGS of individuals enrolled in the Center for Brain Research TATA Longitudinal Study of Aging (CBR-TLSA) and Center for Brain Research Srinivaspura Aging Neurosenescence and Cognition (CBR-SANSCOG) study.[22,23] Both the CBR-TLSA and CBR-SANSCOG cohorts consist of adults 45 years and older, recruited from community settings in Bangalore, and the villages of Srinivaspura taluk (sub-district) located in urban Bangalore and Kolar districts, respectively, in the state of Karnataka, India. The CBR-TLSA study individuals forming the majority set of this work belong to various communities and population subgroups from across the country, although currently residing in metropolitan Bangalore. In this study, we identify and characterize SVs in 529 deeply sequenced (average depth of coverage 42X) human genomes from the Indian population. These 529 individuals, belonging to more than 30 distinct population subgroups represent a modest proportion of the rich genetic diversity in India.

We discover a set of 36,210 SVs, comprising 24,574 deletions, 2,913 duplications, 8,710 insertions, and 13 inversions. The SVs we discover are predominantly rare (≤1%

[1]Centre for Brain Research, Indian Institute of Science, Bangalore 560012, India; [2]Manipal Academy of Higher Education, Manipal, Karnataka 576104, India
[3]Lead contact
*Correspondence: bratati@iisc.ac.in
https://doi.org/10.1016/j.xhgg.2024.100285.

frequency in our dataset) and have limited overlap with results from individuals of European ancestry, thus reflecting the population specificity of these SVs for Indians. Deletions and duplications are mostly between 100 bp and 1 kb in length, with average lengths of 310 bp and 908 bp, respectively. Insertions are relatively small, with an average length of 164 bp; 1.26% (456 of 36,210) of the resultant SVs have the potential to impact coding regions of mapped genes. Combined with loss-of-function (LoF) propensities for the respective annotated genes, we observe that only five of these SVs are commonly present in our study individuals and mostly in a heterozygous state, and corresponds to three genes (*ADAMTS17, CCDC40,*and *RHCE*). Seven SVs, all being rare and present in heterozygote carriers, significantly impact dosage sensitivity of genes that are known to be associated with various clinical phenotypes. We thus observe a greater preponderance of individuals carrying a single copy of the identified SVs. Most of the SVs in our study are rare and longer SVs are rarer. The low proportion of SVs capable of causing genomic alterations in coding regions or known to be disease relevant, as well as the limited number of long SVs (length >100 kb constitutes about 5% of the total SVs identified) may stem from evolutionary selection against large regions of genetic rearrangements. This might be attributed to the fact that our study individuals are derived from population-based cohort studies rather than being specifically recruited in specific disease cohorts. Our work alleviates the Eurocentric bias of genomic studies, by contributing to the discovery and characterization of SVs from short-read WGS in the much-understudied Indian population. We expect that with more such discoveries, regular SV detection in population-based and disease cohort studies will help to utilize and appreciate their potential for downstream applications and research.

Material and methods

Sociodemographic details of samples

The CBR TATA Longitudinal Study of Aging (CBR-TLSA)[23] and CBR Srinivaspura Aging Neurosenescence and Cognition (CBR-SANSCOG)[22] longitudinal cohort studies are approved by the Institutional Human Ethics Committees of the Indian Institute of Science and Center for Brain Research. We recruited the participants for the CBR-TLSA study from community settings in Bangalore through flyers and oral communication. The Field Data Collector team conducted awareness campaigns and recruited CBR-SANSCOG study participants by contacting individuals through phone or in-person home visits in the village of Srinivaspura. All participants have signed the written informed consent form. The study participants in CBR-TLSA and CBR-SANSCOG underwent comprehensive clinical examination, neuropsychological tests, blood biochemical, and neuroimaging assessments. Genomic studies consist of genome-array wide genotyping and WGS to better understand the factors contributing to aging in this population. In the past decade, there have been significant advances in worldwide genomics research, highlighted by pioneering studies like the 1000 Genomes Project[14] and the Genome Aggregation Database (gnomAD),[17] which contained data from a vast group of over 195,000 individuals. Essentially, genetics and genomic studies in CBR-TLSA and CBR-SANSCOG study individuals, who are not ascertained for particular diseases during recruitment, aid in estimating the genome-wide variant allele frequencies in the population that will facilitate further clinical and disease-based interpretations. Our current work consists of 529 individuals recruited in these two studies, with overall 47% females and 53% males, and a mean age of 63.6 years (SD = 10.57).

Distribution of samples across states and ethnicities

Our study samples constitute over 30 well-defined population ethnic groups and their distributions across various states in India. The representation of these ethnic groups correlates with the different states in our study dataset. However, due to confidentiality considerations related to this cohort, we are unable to disclose ethnic backgrounds of the samples. Instead, we depict the sample distribution by state on a map of India in Figure S1A. The majority of our sampled population is primarily reflective of South India, exhibiting a notably lower representation from northern India. While our study is designed with the intention of achieving diversity and inclusivity, it is important to note that the scope of our sampled demographics might not fully encapsulate the extensive diversity prevalent across the entirety of India. Furthermore, we conducted principal component analysis to reflect the genetic ancestry of the study individuals (Figure S1B). We utilize the SmartPCA[24] package to generate a PCA plot, revealing associations between genetic variations and geographical locations in our samples. Figure S1 (A and B) shows the distribution of individuals mirroring their genetic structure in relation to the geographic location, with a high proportion of the individuals belonging to the southern states of India.

WGS

We conducted genomic DNA extraction from peripheral blood samples obtained from the study participants using a QIAamp DNA Blood Midi Kit (100); Cat No./ID: 51185 (NucleoSpin Blood L Midi kit, cat number 740954.20). We assessed the quality and quantification of genomic DNA using NanoDrop and Qubit 4 fluorometer (Thermo Fisher Scientific, Cat #Q33228). We used the TruSeq DNA PCR-Free kit (Illumina, Cat # 20015962) for library preparation as per the manufacturer's instructions. TruSeq DNA PCR-Free offers superior coverage of areas that are traditionally difficult to sequence, such as GC-rich regions, promoters, and repetitive content. We estimated the quantity of the DNA libraries using the Qubit 1X dsDNA HS Assay kit (Thermo, Cat. number: Q33231). We checked the quality of the DNA libraries using Agilent Fragment Analyzer (Agilent Technologies, Inc.) with the high-sensitivity next generation sequencing (NGS) fragment analysis kit (Cat. No. DNF-474-1000). We quantified the libraries by an absolute quantification method using QuantStudio 6 pro-real-time PCR system (Thermo Fisher) and KAPA Library Quantification Kit – Illumina/Universal (Roche Sequencing solutions, Cat # KK4828). We performed WGS on the Illumina NovaSeq 6000 platform (Illumina Inc., San Diego, CA, USA) using the NovaSeq 6000 S4 reagent kit (Illumina, Cat. No. 20028312) following library preparation and quality check. We generated a total of 529 paired-end WGS data across 23 different runs on the NovaSeq 6000 sequencer and further processed the WGS data analysis pipeline to identify structural variations (Figure S2).

Workflow for identifying structural variations

The entire workflow for detecting structural variations by split-read, read-pair, and assembly-based methods from raw WGS short reads is depicted in Figure S2. Our stringent SV identification approach utilizes an ensemble of three callers, described below.

We processed the raw.bcl files obtained from NovaSeq 6000 sequencer and converted these files into the analyzable.fastq sequence format to conduct quality check and unaligned bam (ubam) files for processing the WGS pipeline.[25] Following this, we sorted the ubam files and marked the Illumina adapter sequences. We then reconverted these reads back to.fastq format and mapped the paired-end.fastq sequences of 529 individuals against the GRCh38.p13 build human reference assembly using the BWA-MEM algorithm.[26] After mapping, we identified and marked duplicate reads while applying base quality recalibration scores using the Genome Analysis Toolkit (GATK) (Figure S2). Throughout the crucial steps of whole-genome data analysis, we conducted quality checks to ensure the accuracy of our results. The average coverage of these 529 genomes was 42X (Figure S3A) and 151 bp read length. Additionally, we observed a mean Phred score of 36 for the sequenced bases, with 97% of reads successfully mapping to the reference genome. Only 8% of the reads were mapped in multiple regions, known as duplicate reads, and 0.3% of reads were mapped to the reference genomes in either the forward or reverse reads, termed singleton reads. These duplicate and singleton reads were marked and not considered for variant calling. Consistently, these parameters were observed across the majority of samples, as illustrated in Figures S3B–S3D. All the genome sequences successfully passed the mentioned quality checks. For SVs discovery and genotyping, we utilize three callers: LUMPY (v0.2.13),[27] DELLY (v0.8.5),[28] and MANTA (v1.6.0),[29] which employ split-read, read-pair, and assembly-based methods. We use three callers to ensure concurrent, robust, and accurate SV calling for our study dataset. Comparable to large-scale studies such as the 1000 Genomes[14] and gnomAD,[16] our methodology distinguishes itself by strategically employing these methodologies with stringent filtering and merging approaches (section below).

Comparison of results from SV discovery workflow with standard GIAB call sets

We downloaded the HG002/NA24385 benchmark set of SV call from Genome in a Bottle Consortium (GIAB: https://www.nist.gov/programs-projects/genome-bottle) in VCF file format (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/analysis/NIST_HG002_DraftBenchmark_defrabbV0.011-20230725/), comprising ~70,000 SVs.[30] Independently, we procured the GM24385 cell line sample from Corriell (https://www.coriell.org/1/NIGMS/Collections/NIST-Reference-Materials) and sequenced it using the S4 flow cell kit in NovaSeq 6000, following Illumina's recommended protocol at our in-house sequencing facility. The coverage of GM24385 from our sequencing is 38X. We then processed the WGS data analysis pipeline of GM24385 (Figure S2) to generate a binary alignment mapping (.bam) file and identified structural variations using three SV caller methods. To check the concordance of our SV calling pipeline, we converted the downloaded HG002/NA24385 high-confidence truth set, and our sequenced and analyzed VCF files of HG002/GM24385 to BED format using *"svtk vcf2bed"* (https://github.com/talkowski-lab/svtk). We compare the resulting BED files of NA24385 and GM24385 (truth set and CBR sequenced for the same sample, respectively) using BEDTools intersect[31] with 50% reciprocal overlap (flag of "-f 0.5"). This comparison results in a precision of deletions at 96.57%, duplications at 53.13%, and insertions at 86% for the identified SVs in our study dataset, which were supported by corresponding SVs in the same GIAB samples. Nevertheless, the high precision value ensures that there are minimal false positives in our robust SV discovery workflow for identifying the SVs.

Genotyping and merging at the cohort level

After the initial calling of SVs by three methods (Lumpy, DELLY, and MANTA), we conducted genotyping, after filtering at the sample level for each caller, and merging the variations at the cohort level. The LUMPY outputs of all 529 individuals are re-genotyped using SVTyper,[32] taking the VCF file of SV sites generated by LUMPY and the aligned BAM file from the WGS pipeline as input. SVTyper provides genotyped SV sites for each sample in VCF format. We discarded all SVs with split-read support of ≤ 3 and break end (BND) SV type for each individual to avoid false discovery of SVs. Additionally, DELLY is utilized to sensitively and accurately detect structural variations by integrating the read-pair (discordant paired-end) and split-reads algorithms. The obtained genotyped SV sites are stored in binary call format (bcf), which is later converted to variant call format (vcf). From the resulting DELLY SV output, SVs with split-read support of ≤ 3 and BND SV type of the chromosomal coordinate of END value not equal to the coordinate of START are filtered out. Using MANTA, we simultaneously carry out an assembly-based method, which involves constructing a graph of all break-end relationships in the genome and processing the components for variant hypothesis creation, assembly, scoring, and VCF reporting. We excluded split-read support of ≤ 3 and break end with lengths less than 50 in the MANTA SV vcf file for each sample (Figure S2).

We merged the identified structural variations (SVs) at two levels using "SURVIVOR merge": sample level and cohort level. At the sample level, we performed the merging of SVs called DELLY, LUMPY, and MANTA only when they have been discovered by all three methods (Figure S2), and the distance of 1000 bp between the relative breakpoints at each end, taking SVs type into consideration where the minimum length of the SV is 50 bp. Then we merged the SVs detected in individual samples using all three methods across all 529 samples. We applied a 1,000-bp distance between the relative breakpoints at each end, while also considering the SV types and minimum length of 50 bp. This allowed us to generate a comprehensive call set of SVs for our study individuals at the population level (Figure S2).

Next, we calculated the discovery allele frequencies for the final SV site using "svtools afreq" (http://www.lib4dev.in/info/hall-lab/svtools/INSTALL.md). This allows us to determine the frequency distribution of the SVs within our dataset of 529 individuals. Based on the discovery of allele frequencies, we categorized the SVs into different frequency groups. Ultra-rare variants are considered as any variant present in less than or equal to three (0.5%) individuals. SVs with frequencies greater than four individuals but less than or equal to six individuals ($>0.5\%$ to $\leq 1\%$) are classified as rare. Similarly, SVs with frequencies more than seven individuals but fewer than 26 individuals in our study dataset ($>1\%$ to $\leq 5\%$) are categorized as low frequency. Finally, SVs observed in more than 5%, corresponding to over 27 individuals, are categorized as common variants. This categorization allows us to get insights into the distribution and prevalence of SVs within our dataset, providing a comprehensive understanding of their frequency in the population under study.

Filtering SVs

In order to retain good quality and reliable SVs in our final call set, we conducted thorough manual checks after the discovery workflow and discarded SVs identified to be present in repeat regions of the human genome such as human telomeric, centromeric, and repeat regions from the following link: https://github.com/dellytools/delly/blob/main/excludeTemplates/human.hg19.excl.tsv. Next, we performed a liftover of these regions from hg19 to hg38 genomic coordinates using the UCSC liftover. Afterward, we excluded SVs located within these specific regions to narrowing our focus on variants with known functional significance. Accurately identifying repeat regions from short reads sequencing data can be challenging due to the complex and repetitive nature. Moreover, the short reads may not span the entire length of the repeat region, leading to difficulties in accurately aligning the reads to the reference genome or identifying unique mapping locations. Within our study dataset, it was observed that 16 translocations present within the final call set contained either a single nucleotide repeat or a dinucleotide repeat. Therefore, we did not include these 16 translocations in our subsequent analysis. Additionally, we carried out manual assessments of SVs located in the X and Y chromosomes. In our final call set, we included pseudoautosomal regions (PAR1 and PAR2), that behave similarly to autosomes. Since longer SVs (>1 MB) are harder to be robustly detected from short reads, we again performed manual assessment and removed those that overlapped with other types of SVs and could be ambiguous. As a result, we excluded 22 SVs in our final call set.

We perused the mobile element data sourced from the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTables) to check for presence of Alu and LINE regions within our SV callset. Using BEDTools intersect with a 50% reciprocal overlap, we observed one Alu and three LINE1 elements overlapping with our SV insertions, which were present in a limited number (n = 15, nonoverlapping) of individuals in our study dataset. Furthermore, we delved deeper for these elements at the individual level in our study dataset. This analysis revealed that the length of these mobile elements was unknown, and only left (named as LEFT_SVINSSEQ) and right (RIGHT_SVINSSEQ) partial insertion sequences were provided in the VCF INFO field of the individual call set for those samples. Usually, such denominations are provided for large insertions of unresolved length. We could also see that the insertion detected in these regions in our study samples are repetitive in nature, thus alluding to the incomplete detection of mobile elements from short-read sequencing methods. As a result, we excluded four such structural variations from our study dataset, since they cannot be confidently resolved from the short-read sequences.

To summarize the SV detection workflow, we have considered the potential pitfalls given the nature of WGS and inferences being made, and introduced stringent checks and filtering criteria for repeat and complex genomic regions to ensure the quality and reliability of our findings, while looking for evidence from all the methods for population-level calling of SV. Employing identical merging parameters at both the individual level and population level, with consensus from three supporting callers, matching strands as a distinctive feature provides robust evidence for the identified SVs.

Annotation of structural variations

We annotated the identified structural variations using AnnotSV (v3.0.7) with GRCh38.p13 build of the human genome. The resulting output file includes gene-based annotations, information on repeats, genic intolerance, and overlapping features from various databases such as the Database of Genomic variants,[33] 1000 Genomes,[14] gnomAD,[17] ExAC[34] databases, disease-based annotation from OMIM,[35] pathogenic features from dbVar,[36] and American College of Medical Genetics and Genomics (ACMG) categories.[37] By annotating the structural variations, we are able to determine the number of genes affected in our SV call set. We are extending our filtration process for impacting clinically relevant SVs by emphasizing the overlap of coding regions with a minimum overlap of 8 bps. Furthermore, to understand their relevance to disease mapping, we used the eDGAR database[38] to map the genes that were annotated by our identified structural variations. The eDGAR database contained curated information on gene-disease mapping derived from sources like OMIM,[35] humsavar,[39] and ClinVar.[40] This comprehensive analysis allows us to gain insights into the potential relationships between the identified genes and diseases.

Our study identified 13 inversions that are located within the intronic regions of six genes: *KIF17, CCDC3, DLG2, SLC8A1, SLC8A1-AS1*, and *TTLL11*. Inversions refer to the balanced rearrangement of DNA segments, involving the reversal of a section of DNA orientation, and potentially disrupting the binding of transcription factors, which are proteins regulating gene expression.[41,42] Therefore, to understand if these inversions have any impact on the transcription of these genes, we manually examined them using the Transcription Factor Binding Site Prediction (TFBSPred) tool.[43] Our analysis determined that none of the inversions affect the transcription sites and they are not likely to affect the binding of transcription factors at the start site of the gene. This was also supported by the observation that the individuals in our dataset carrying the corresponding inversions did not receive any diagnoses related to the abnormalities causing rearrangements in these genes.

Genome-wide estimation of deleterious variants

To understand intolerance to changes inflicted by our identified structural variations in the genomic regions, we extracted LOEUF and ExAC pLI score metrics from the results obtained from AnnotSV results. The LOEUF metrics represent the LoF observed/expected upper bound fraction, where low LOEUF scores (e.g., 0 to 1) indicate strong selection against predicted loss-of-function (pLoF) variation in a gene, while high LOEUF scores (e.g., 9) suggest a relatively higher tolerance to inactivation. Conversely, the pLI is a score that indicates the probability of a gene being intolerant to an LoF variation. A gene with a pLI value of 0.9 or higher is considered an extremely LoF intolerant gene. To identify SVs mapping to genes that exhibit high intolerance to LoF, as determined by these metrics, we further focused on SVs that have the potential to cause frameshift changes, which can have a significant impact on the function of the affected gene. According to the LOEUF metrics, we identified 32, 40, and two distinct genes impacted by respectively the same number of deletions, duplications, and insertions as highly intolerant. Similarly, based on pLI scores, we identify 27 deletions, 34 duplications, and one insertion as highly intolerant (Table S1).

Gene-based estimation of dosage sensitivity

We conducted an estimation of the dosage sensitivity of the genomic regions mapping to our called SVs using annotation from the Clinical Genome Resource (ClinGen) consortium.[37] This estimation involves two independent rating systems: a haploinsufficiency (HI) score for the LoFs and a triplosensitivity (TS) score for the gain of functions. We consider the score of 3 for both HI and TS parameters that suggests evidence of dosage pathogenicity for the

genic region and are associated with clinical phenotype. In addition, we refined our analysis by applying additional filters to focus only on genes that are affected by loss or gain of functions and frameshift changes and performed an in-depth analysis of the genes mapping with diseases using the eDGAR database.[38]

Inspecting correlated neighboring loci associated with traits enlisted in EBI-GWAS catalog

We performed an analysis to examine correlated neighboring loci associated with traits listed in the EBI-GWAS catalog. For this study, we utilized the GWAS catalog version 1.0.2, which contains all associations. The catalog was downloaded from the provided link on July 26, 2023 (https://www.ebi.ac.uk/gwas/api/search/downloads/alternative/gwas_catalog_v1.0.2-associations_e110_r2023-07-20.tsv). Our aim was to determine if the SVs we identified could be implicated, functional, or responsible for traits for which GWAS data are available in the EBI catalog. We ensured consistency by using the exact same genomic coordinates in GRCh38. We extracted SNV-phenotype results from the downloaded EBI catalog and then those SNVs are mapped with our study cohort SNP dataset. We employed Plink2[44] tools to calculate LD (linkage disequilibrium) correlations between these SNVs that mapped with the EBI catalog and our identified SVs, treating them as part of the same set. The correlation calculations were constrained to ±500-kb windows, ensuring that we only considered correlations between an SNV and SV if they were located within 500 kb of each other[45] and thoroughly examined the trait-specific findings reported in EBI to characterize the SVs for disease. The population genomic details of SNVs from our study individuals are provided in the supplemental information ("Single nucleotide variants and small insertions and deletions" section).

Overlap of SVs in Indians with the worldwide dataset

We retrieved the South Asian (SAS) population-specific structural variations from the 1000 Genomes latest release,[46] nstd152 dataset,[47] IndiGen-SV dataset,[18] and gnomAD[17] SVs in VCF format from the following links: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20210124.SV_Illumina_Integration/, https://ftp.ncbi.nlm.nih.gov/pub/dbVar/data/Homo_sapiens/by_study/vcf/, https://clingen.igib.res.in/indigen/download, and https://gnomad.broadinstitute.org/downloads#v2-structural-variants. Then, we utilized BEDTools intersect with a 50% reciprocal overlap threshold (flagged as "-f 0.5") to identify the overlaps based on genomic coordinates between our SV call set and the other four SV datasets for each SV type independently. In this process, we aimed to determine the shared regions between the datasets. In addition, we calculated the Pearson correlation coefficient (r^2) to assess the direction and strength of the linear relationship between allele frequencies of SVs overlapping between our dataset and the global datasets of 1000 Genomes, gnomAD, and IndiGen SVs. We additionally, tested the results for the overlapping SVs after categorizing them into two groups based on our study frequency distribution: rare variants-present at less than 1%, and common variants-present at greater than or equal to 1%. This analysis allows us to understand the similarities and differences between the global dataset and the call set we identified.

Comparison for array-based and WGS-based deletion and duplication SVs

We identified the deleted and duplicated chromosomal regions using array-based genotyping data from the high-throughput output of raw.cel files using Affymetrix Axiom Precision Medicine Research Array in GeneTitan Multi-Channel Microarray Instrument for 529 samples from our study dataset. We included only the samples with a call rate above 98.5%, resulting in 515 samples for further analysis. Furthermore, we generated the summaries, calls, reports, and confidence files using the Axiom Analysis Suite (v5.1.1.1). Then, we utilized the Affymetrix Axiom CNV Summary tool (v1.1.0.85) to generate files in PennCNV output format. These files contain log R ratio and B allele frequency values for all the samples associated with the CEL files. We retrieved the annotation file from ThermoFisher (https://www.thermofisher.com/order/catalog/product/sec/assets?url=TFS-Assets/LSG/Support-Files/Axiom_PMRA.na35.r3.a1.annot.db.zip) to detect CNVs. Furthermore, we utilized pfb (Population frequency of B allele) and gcmodel (GC content) files, generated using compile_pfb.pl and cal_gc_snp.pl scripts, respectively. We employed the default affygw6.hmm files in PennCNV[9] for this step. Then, we converted the genomic coordinates from the GRCh37 build to the GRCh38 build using the Crossmap.py tool. In conclusion, a comprehensive comparison was performed between the deletions and duplications identified by PennCNV method and final call set of structural variations of deletions and duplications independently. The analysis was focused on intervals of ±125 kb, and an in-house python script was utilized to facilitate the evaluation.[48]

Results

Structural variations identified and characterized

We uncovered a total of 36,210 structural variations, comprising 24,574 deletions, 2,913 duplications, 8,710 insertions, and 13 inversions, in the genomes of 529 individuals using split-read, read-pair, and assembly-based methods (Figure 1A). Reflective of the total number, we found that the deletions (67%) are the more predominant types of structural variations followed by insertions (21%) and duplications (9%) (Figure 1B), over each chromosome as well. We observed on average 3,010 high-confidence deletions, 3,248 insertions, and 193 duplications per genome (Figure 1C). The majority of deletions (67%) occurred in fewer than 10 individuals, while more than 50% of insertions were present in more than 10 individuals, indicating a higher insertion frequency per genome. This was interesting because overall deletions outnumbered insertions by 3-fold, yet the per-genome estimates showed more insertions in our study dataset. Duplications detected per chromosome were less common, likely due to challenges in detecting them using short-read sequencing. The average number of insertions per genome aligned closely with South Asian results from 1000 Genomes (3,378 insertions on average per genome).[46] However, average deletions (4,066) and duplications (1,168) per genome in our dataset appeared to differ from the 1000 Genomes discovery due to population characteristics or pipeline variations. Taking all SV types together, our study dataset revealed 3,251 high-confidence SVs per genome. Notably, a substantial proportion of deletions (77%), duplications (84%), and insertions (73%) were attributed to a cohort of fewer than 50 individuals, as illustrated in the initial two bars of Figure 1D within

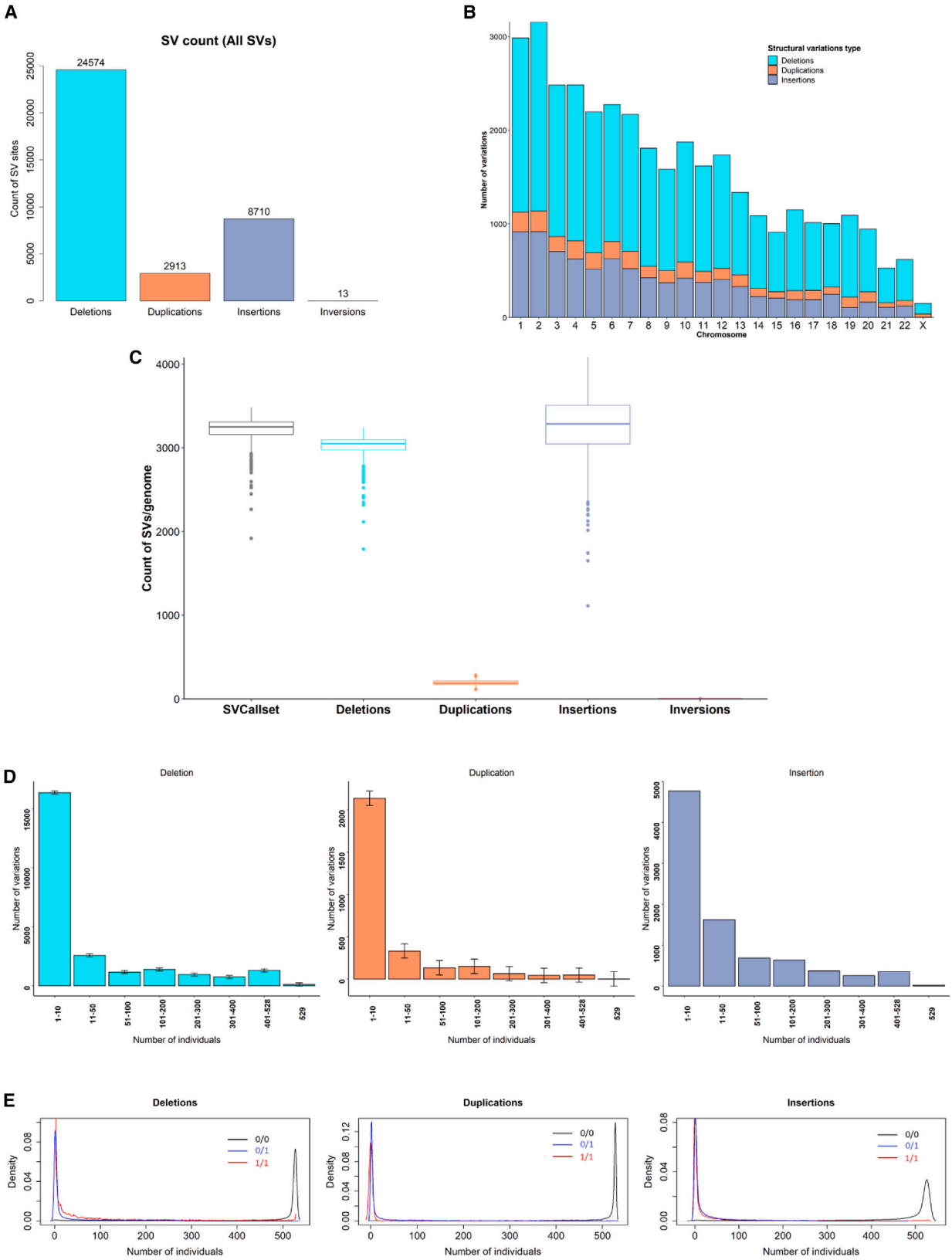

Figure 1. SV discovery in the high-coverage WGS data of Indian samples

(A) SV count: This bar plot illustrates the number of identified SVs categorized by variant type. The dataset comprises a total of 24,574 deletions, 2,913 duplications, 8,710 insertions, and 13 inversions.

(B) Distribution of SVs per chromosome: This stacked bar plot depicts the distribution of structural variations identified in each chromosome. Deletions are the most prevalent types of SVs, followed by insertions and duplications.

(legend continued on next page)

our study dataset, highlighting that limited subset of individuals carried the majority of structural variations. This distribution pattern of structural variations is consistent with previous studies,[16] indicating that these large genomic rearrangements are not common in the population. Our results emphasized the heterogeneity in the occurrence and distribution of structural variations at the individual level (Figures 1D and S4), which played a crucial role in comprehending the genomic landscape of these variations.

Genotype distributions

We analyzed the genotype distributions of identified SVs across all samples. For this we tallied the number of homozygous reference (0/0) individuals having the two copies of reference alleles, homozygous alternate genotypes (1/1) representing individuals with the same alternate alleles, and heterozygous (0/1) genotypes representing individuals having one reference allele and one alternate allele at the locus, indicating heterozygosity. Figure 1E, depicted a higher frequency of variants in the small number of heterozygous and homozygous alternate individuals (as represented by the peaks of the blue and red lines). This is also supported by variant-specific details in Table S2. These patterns were observed consistently across all three types of structural variations. Sixty-six deletions, 16 insertions, and two duplications were the only SVs present across all 529 individuals in our study dataset. These deletions and insertions were for the homozygous alternate variant, and duplications were present in the heterozygous state. Notably, all of these deletions, duplications, and insertions present in all individuals, were located outside the coding regions. This implies the potential impact on protein-coding sequences is minimal, suggesting that these structural variations may not directly influence the function of the encoded proteins, even though some SVs are present in all individuals. Moreover, from our analysis results, we observed that most rare alleles, present in heterozygous and homozygous states, were found in only a few individuals as expected since those SVs are large rearrangements of the genome (Table S2). The individuals carrying those rare SVs can be further studied to gain more insights into their conditions. We also observed that for the common SVs (present in more than 26 individuals out of 529), there was a higher average proportion of heterozygote carriers (130 individuals, 7,581 SVs, color-coded in green in Table S2) compared with those with a homozygous alternate status (average of 65 individuals, 2,732 SVs, and color-coded in blue Table S2) (Figure 1E).

Length of SVs

Figure 2A illustrates the length distribution of deletions, duplications, and insertions within the genome. Across all three types of SVs in our study dataset, a predominant occurrence was observed within the range of 50 bp to 500 bp—deletions (65%), duplications (43%), and insertions (98%). Notably, duplications are mostly longer (median length = 908 bp) compared with deletions (median length = 310 bp). Moreover, a substantial portion of insertions consisted of relatively small lengths, with a median length of 164 bp. We also conducted a detailed analysis of the length of structural variations across all the chromosomes (Table S3). Less than half of the deletions and duplications (13.23% and 15.12%, respectively) in chromosome 1 exceeded average lengths of 3.4 kb and 18 kb, while 36.95% of insertions in chromosome 1 were longer than the average length of 166 bp. A similar trend was observed in all other chromosomes for deletions, duplications, and insertions, as shown in Table S3. This was reflected in the fact that over all chromosomes, only 13.52% of deletions, 15.48% of duplications, and 36.06% of insertions have longer stretches than their respective average lengths. Therefore, longer variations were less common in our study, indicating that the structural variations in the analyzed chromosomes were generally smaller in size compared with their average lengths.

Allele frequencies

We classified the structural variations into four categories based on their discovery frequency among the individuals in our dataset: common (>5%), low frequency (>1% but ≤5%), rare (>0.5% but ≤1%), and ultra-rare (≤0.5%). In our analysis, we observed that 26.91% of deletions, 19.40% of duplications, and 34.50% of insertions were common in our dataset, amounting to 10,185 variations (Figure 2B and Table S4). Notably, a substantial number of variations (50.38% of deletions, 54.34% of duplications, and 35.10% of insertions) occurred in fewer than three individuals, falling into the ultra-rare category, totaling to 17,028 such variations out of overall 36,210 (Table S4).

(C) Mean SV count per sample by variant type: In this box and whisker plot, by a horizontal line the mean count of SVs per genome is shown for the overall SVCallset, deletions, duplications, insertions, and inversions. The SVCallset category represents all SVs in our study dataset. The boxplot illustrates the median with a horizontal line and displays the first and third interquartile ranges.
(D) Distribution of SVs across samples: The plot illustrates the distribution of each SV type across all individuals over eight bins based on the number of individuals in each bin (1–10, >11–50, >50–100, >101–200, >201–300, >301–400, >401–528, and all 529 samples). The initial bar, representing 1 to 10 individuals carrying the majority of structural variations, suggests that SVs are infrequently observed in population-based studies.
(E) Genotype distribution of SVs: The genotype distributions of identified SVs across all samples are depicted by this line plot. The x axis represents the number of individuals, while the y axis represents the estimated probability density of genotypes for each individual. A higher count of SVs in a limited number of individuals is emphasized as observed in the peaks closer to the origin of the plots. These variations are more likely to be found in either a heterozygous state (0/1, as indicated by the blue line) or a homozygous alternate state (1/1, as indicated by the red line). The presence of merely 66 deletions, two duplications, and 16 insertions among all depicted individuals is demonstrated by the corresponding plot's tail in these respective categories.

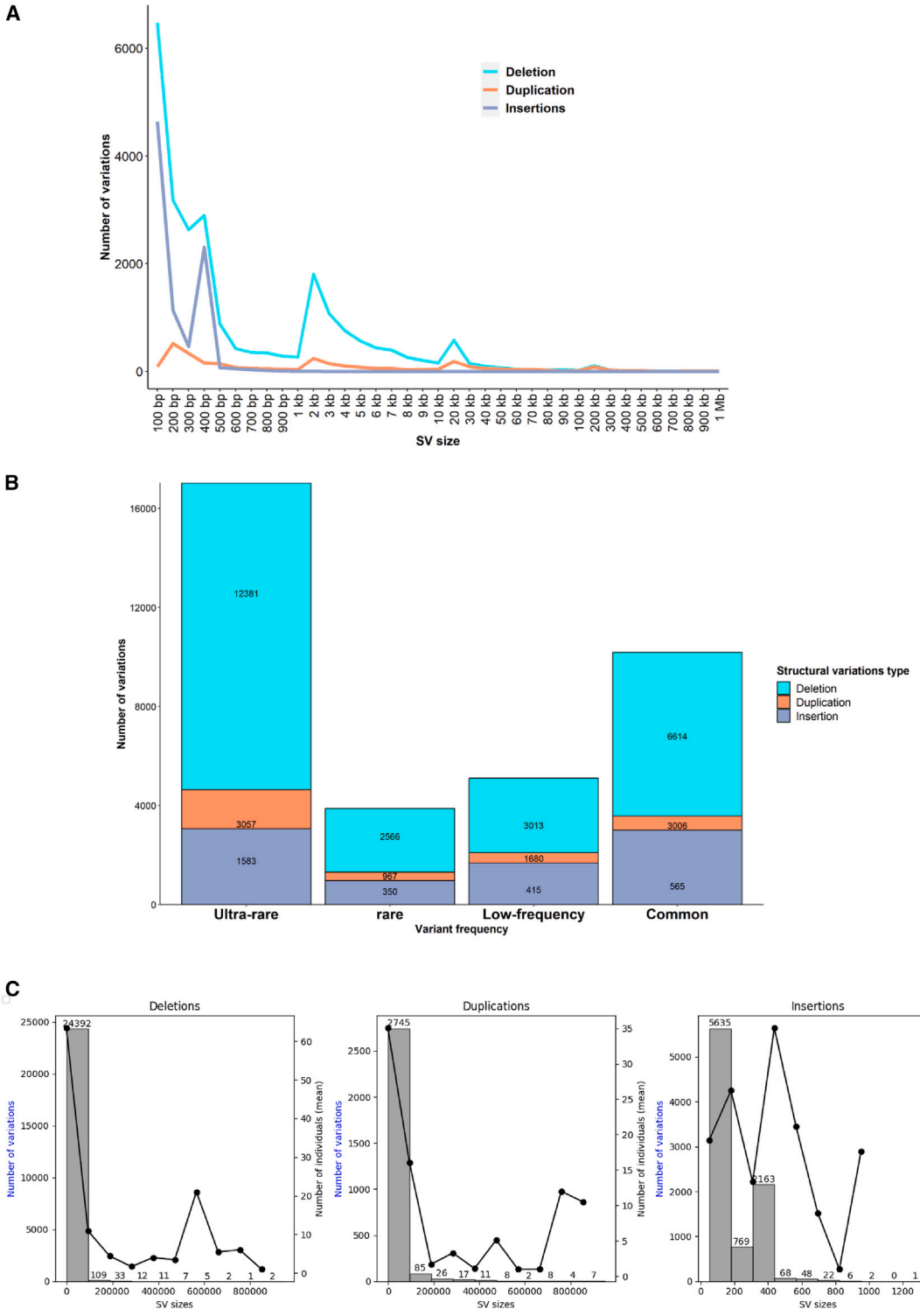

Figure 2. Length and frequency distributions of SVs

(A) Distribution of SV Lengths: The line plot illustrates the length distribution of identified SVs within our study dataset. Predominantly, the majority of SVs exhibit lengths ranging from 50 bp to 500 bp across all SV types, followed by the next most frequent range observed between 1 kb and 3 kb for deletions and duplications. Overall, the lesser number of variations for deletions, duplications, and insertions depicts that longer SVs are less commonly observed in our study samples.

(B) Frequency Distribution of SVs: This stacked bar plot illustrates the frequency distribution of SVs based on four categories: common (>5%), low frequency (>1% but ≤ 5%), rare (>0.5% but ≤ 1%), and ultra-rare (≤0.5%). Our study dataset reveals a prevalence of rare and ultra-rare variations (involving at most five individuals), followed by common SVs (involving over 26 individuals), and low-frequency SVs (involving six to 26 individuals) is observed within our study dataset.

(legend continued on next page)

Additionally, 2,566 (10.44%), 350 (12.02%), and 967 (11.11%) were rare variations occurring in three to five individuals (Figure 2B and Table S4). The remaining SVs, 12.26% (3,013 out of 24,574) deletions, 14.25% (415 out of 2,913) duplications, and 19.29% (1,680 out of 8,710) insertions were low frequency, being present in more than five but fewer than 27 individuals in our dataset (Table S4). The majority of deletions and duplications were either in rare or ultra-rare categories, being present in fewer than five individuals (Figures 2B, S5, and Table S4). We found that 40% of deletions, 43% of duplications, and 26% of insertions appeared as "singletons" (that is, only one allele observed across all samples). In our analysis encompassing all SV classes, we observed that the majority of SVs were small, with a median size of 274 bp, and rare (allele frequency $\leq$ 1%), as depicted in Figure 2C.

In our study, we also presented a graph that illustrates three parameters intended to support the relationship between the number of structural variations, their presence in study individuals, and variety in their length (Figure 2C). We observed that a major proportion of deletions and duplications (the first bars in Figure 2C, first two panels) whose lengths were less than 200 kb, are present on an average in about 100 individuals in the dataset. Similarly, for insertions, although they could go up to 1.4 kb in length, about 180 individuals on average had insertions that spanned fewer than 400 bases. This observation highlights that the extent of DNA rearrangement was most likely influenced by evolutionary processes and longer stretches of the genome were to be encountered less often unless they are favored by natural selection or chance events that remain neutral to biological functions. This phenomenon had also been noted in other population-scale studies focused on uncovering structural variations within populations.[16,17]

Annotation of SVs based on genomic regions

Using Ensembl and RefSeq,[49] we annotated our identified SVs for genic regions and reported schematic representations of multiple instances in Figure 3 to provide a comprehensive understanding of our annotation workflow. In our study dataset, a total of 18,581 structural variations had an impact on 11,405 genes. These genes encompass various regions such as 5′ and 3′ UTR, exons, and introns. Notably, the remaining structural variations are located outside genic regions, thus having no direct influence on the genes. Interestingly, we observed that a few SVs affect specific parts of the genic region. For instance, one SV impacts the starting region of certain genes, while other regions of the same genes, such as exonic or intronic regions, are affected by different SVs in the vicinity. This phenomenon is highlighted in instance 1 within Figure 3. In other instances (instance 2 in Figure 3), some SVs impact multiple adjacent genes due to their larger size. On average, in our study we observed that about 21 (=11,405/529) genes per genome could be affected by SVs. Furthermore, our specific focus was on the SVs found in the coding regions of the genomic regions. In instance 3, as depicted in Figure 3, we observed 6,832 SVs located in those coding regions. To gain a deeper understanding, we conducted a thorough analysis of those coding region SVs to identify their impact on functional changes caused by frameshift SVs. We observed 456 SVs that had an impact on frameshifts, with 298 deletions, 155 duplications, and three insertions affecting the start and stop sites in genes. Out of the total of 456 frameshift changes, 387 were rare SVs (frequency <1%), while the total number of rare SVs that we identified was 20,907 (Figure 2B). Our study dataset shows that a total of 1.26% (=456/36,210) of all SVs can cause changes in the protein-coding regions of the genome. Thus, on average there can be approximately one frameshift alteration per genome (=456/529). The rarity of protein-altering SVs in our results might be reflective of the fact that our cohorts were not specifically ascertained for any particular disease, thus decreasing the propensity of finding disease-related SVs in the discovery analysis.

Implication to diseases for identified SVs in coding regions

We analyzed the potential disease-related impact of identified common (present in >5% of the study individuals) SVs on the coding region for causing frameshift changes, since those are the variants that can alter protein amino acid sequences, resulting in non-functional or truncated proteins. We have noted in previous sections that most structural variations are present in <1% of individuals, with heterozygous genotypes. We uncovered 21 genes in 21 deletions, 12 genes in 12 duplications, and two genes in two insertions that are present in at least 5% of individuals of our dataset and causing frameshift changes. For the 21 deletions, two genes (*ADAMTS17* and*CCDC40*) mapped to specific genetic disorders. The rest of the genes had no reported evidence for disease mapping. Deletions in *ADAMTS17* (ADAM metallopeptidase with thrombospondin type 1 motif 17) are known to occur in Weill-Marchesani Syndrome (https://omim.org/entry/607511?search=ADAMTS17&highlight=adamts17); however, the inheritance pattern is mostly autosomal recessive, and 155 individuals in our study sample carry a 6.2-kb deletion; however, all of them are heterozygous, which explains the absence of the syndrome diagnosis in these 155 individuals. Homozygous mutations or compound heterozygote mutations in *CCDC40* (coiled-coil domain containing 40) present in 156 individuals in

(C) Relationship between Length and Frequency: The dual-y-axis plot illustrates the relationship between SV size (lengths) and two key metrics: the number of structural variations and the mean number of individuals in each SV size bin. The x axis represents SV size, while the primary y axis corresponds to the count of structural variations, and the secondary y axis corresponds to the mean number of individuals. The graph highlights a notable pattern wherein smaller SV sizes are associated with a greater number of variations. Additionally, it reveals that these smaller SVs tend to be present in fewer individuals.

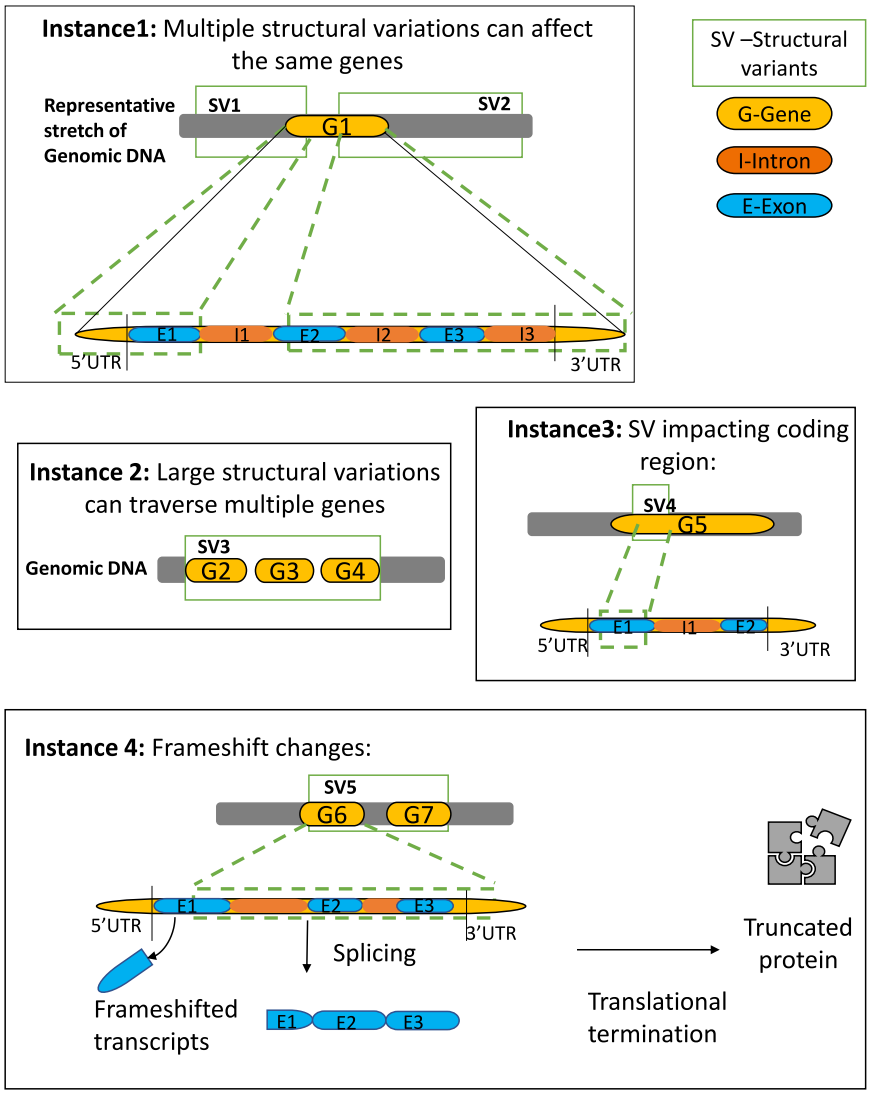

Figure 3. Workflow of genomic annotation of SVs

This figure provides a comprehensive understanding of our SVs annotation workflow, presenting multiple instances of how SVs plausibly affect genic regions in the genome and thereby their function, based on the extent of overlap of the SV with the transcript. In instance 1, multiple structural variations affecting the same gene. In instance 2, one large structural variation can traverse multiple genes. Instance 3 depicts how SVs can impact 5′ or 3′ UTRs or exons in the genic coding regions. Instance 4: Frameshift changes, which can have significant implications for protein-coding genes and their associated functions are shown.

the heterozygous state, showing no symptoms of the diseases.[50] Additionally, 12 duplications present in 268, and 156 individuals respectively map to two genes *ADAMTS17*and *RHCE*. Notably, the deletions and duplications corresponding to the *ADAMTS17* gene affect distinct samples in our study dataset, meaning the same individuals do not contain the deletion and duplication. The amorph type Rh-null disease corresponding to *RHCE* gene (https://omim.org/entry/111700?search=rhce&highlight=rhce) results from a homozygous genotype; however, all our study individuals are heterozygous for the duplication. The Rhesus system polypeptide is a specific protein that is encoded by the *RHCE* gene, which plays a crucial role in the Rh system by determining the presence or absence of the E antigen on the surface of red blood cells. No genetic disorders are identified mapping to the genes and frameshift changes affected by the insertions in our study. In our study, these SVs exhibited a heterozygous genotype state indicating that individuals carrying these SVs may have had one copy of the variant allele. This crucial information from a population-based SV detection study enhances comprehension of variant effects on the human genome and could aid in diagnostic and therapeutic development. Our in-depth analysis revealed that although automated gene and disease mapping programs are excellent resources for screening and annotation purposes, detailed knowledge about the genomic coordinates and zygosity of the corresponding variants in the study samples, combined with the knowledge of inheritance, is vital to draw conclusions regarding disease prevalence attributed to SVs in a population.

Intolerance to LoF

We assessed the probability of intolerance to LoF caused by SVs in genes using two standard methods. The first method relies on a probability score (pLI) of 0.9 or higher, indicating a high likelihood of intolerance to LoF variations (nSV = 60), and the second one designates LOEUF values ranging from 0 to 1, with lower values indicating a greater likelihood of intolerance to LoF variations (nSV = 69). When we consider the genes that are intolerant to LoF due to being impacted by the deletions, duplications, and insertions, we observed that most of them are frameshift variations. We focused on these SVs that affected frameshifts and observed deleterious effects by both methods (pLI and LOEUF) in our study dataset, resulting in a total of 56 highly intolerant SVs comprising 24 deletions, 31 duplications, and one insertion. Because these measures of genic intolerance were indicative of evolutionary conservation, observing them for frameshift causing SVs in our study samples seems contrary to expectations. However, this is mitigated by the fact that the majority of these highly intolerant SVs (50 out of 56) were rare occurrences within our dataset. Additionally, it is noteworthy that the remaining six SVs causing frameshift in genic regions exhibiting

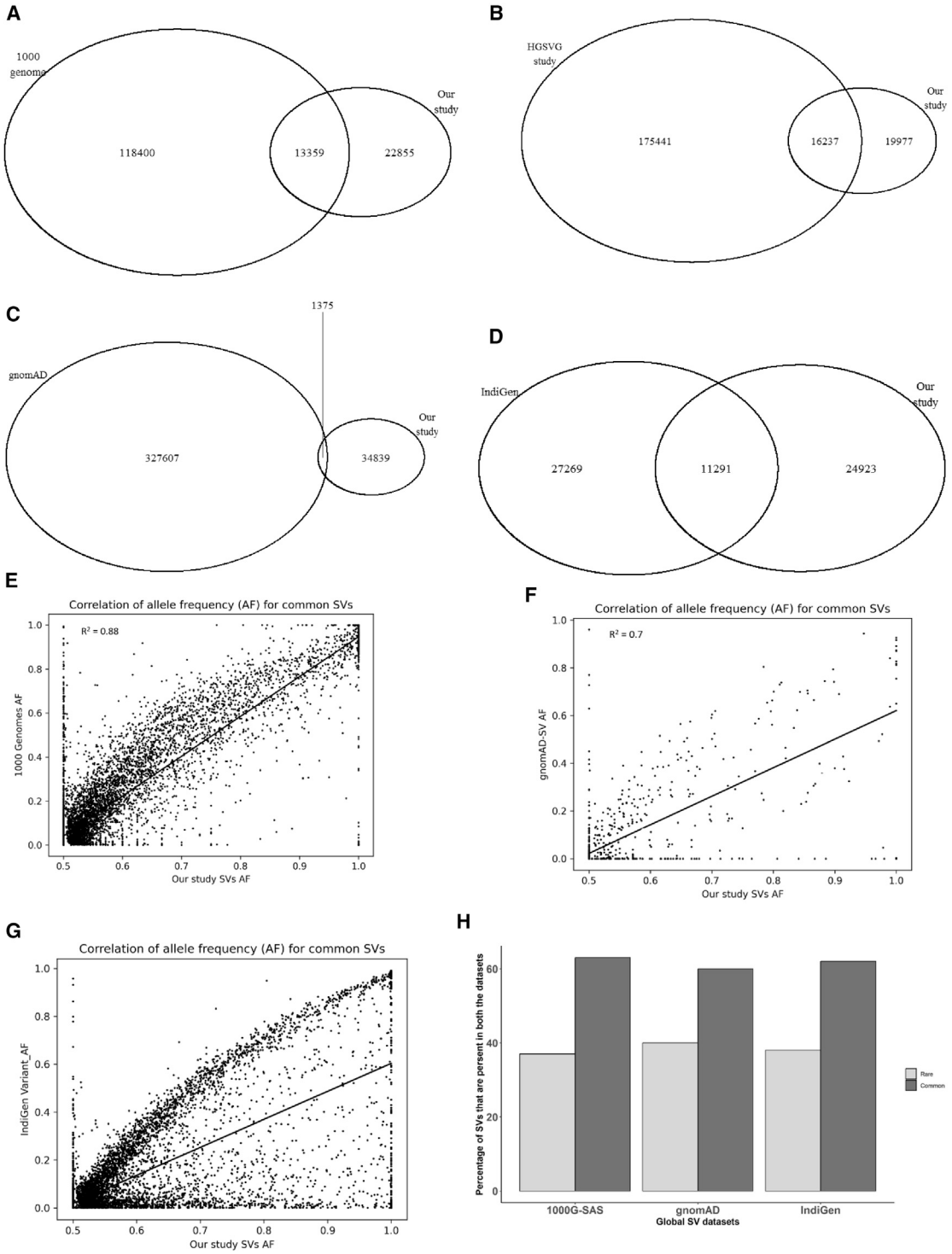

Figure 4. Comparisons with global datasets

These comparisons and frequency correlation analyses provide valuable insights into the concordance and differences between our study dataset and various global datasets. The Venn diagrams visually illustrate the extent of shared SVs among the datasets, while the frequency correlations shed light on the consistency of SV frequencies across different datasets.

(A–D) Venn diagram comparing multiple global datasets: (A) Venn diagram comparing the overlap between the 1000 Genomes dataset and our study dataset. (B) Venn diagram comparing the overlap between the HGSVG trios dataset (dstd152) and our study dataset. (C) Venn diagram comparing the overlap between the IndiGen-SV dataset and our study dataset. (D) Venn diagram comparing the overlap between the gnomAD-SV dataset and our study dataset.

(E–G) Correlations between frequencies of overlapping SVs: (E) Frequency correlation analysis of SVs overlapping between our study dataset and the 1000 Genomes dataset. (F) Correlation of frequencies for SVs overlapping between our study dataset and the HGSVG trios (nstd152) dataset. (G) Correlation of frequencies for SVs overlapping between our study dataset and the IndiGen-SV dataset.

(legend continued on next page)

high intolerance to LoF are present in at most 87 individuals, averaging at only 26 individuals in the entire set of study samples. We observed that five SVs (one deletion in one individual, three duplications in three different individuals, one insertion in four individuals) present in the homozygous alternate state affected genes with high propensity for LoF, respectively being *CARD11* for deletion; *ANO8, USP37, PDZD2* for duplication; and *FLT4* for insertion. Although *ANO8, USP37,* and *PDZD2* are not immediately known for specific diseases, *CARD11* and *FLT4* are known to be implicated in B cell-induced immunodeficiency and lymphedema, respectively; however, clinical symptoms indicative of such conditions were not present in our study individuals. This is not surprising though, given that a spectrum for disease conditions could exist in a population, and thus having genetic knowledge relating to such conditions could empower clinical practices for precise diagnosis.

Dosage sensitivity of SVs

We annotated our identified structural variations (SVs) using ClinGen[37] recommended parameters for dosage sensitivity-haploinsufficiency and TS. Haploinsufficiency occurs when a single functional copy of a gene is inadequate to maintain normal function, while TS describes abnormal phenotypic effects resulting from an additional copy of a specific gene. Through our analysis, we discovered 100 deletions, 17 duplications, and 42 insertions that impact dosage sensitivity and are associated with a clinical phenotype. Among these SVs, frameshift changes that affected the start and stop sites in six genes (*TNRC6B, FLG, COL1A1, BRIP1, SHANK3,* and *ARID1B*) were observed for deletions, and one gene (*AUTS2*) for duplications (Tables S4A and S4B). However, on a closer observation, we noticed that the SVs identified are most of the time not in a genomic region of reported mutations for these associated diseases. For example, mutations, including copy number variations, and large translocations in *SHANK3* are well-known for causing inherited Schizophrenia 15, Phelan-McDermid syndrome. However, only two individuals in our study sample carry the 336-bp deletion with no manifestations of the above clinical features, thus possibly indicating that not all genetic variations in the same gene will result in identical extreme clinical phenotypes. The annotation results depicted here are obtained from ClinGen datasets that possibly places more significance toward reported inherited disorders, and less on the actual sequence regions. Table S5 provides additional insights into these genes and their potential disease associations, including the number of individuals carrying these genes. It is important to note that the number of individuals carrying these genes in our dataset is relatively low, being maximum of two, suggesting that the significance of these genes being associated with the reported diseases may be limited.[51,52]

Overlap of SVs in Indians with the worldwide datasets

We compared the genomic coordinates of the SVs identified in our dataset with multiple global SV datasets, namely 1000 Genome SAS latest release,[46] HGSVG trios of nstd152 SV dataset,[47] IndiGen-SV,[18] and gnomAD[17] SV datasets. Upon comparing with the 1000 Genomes project of 3,202 samples that have a sizable number of individuals belonging to South Asian ancestry, we observed a substantial overlap with our discovery set. Specifically, we found an overlap of 11,022 (45%) deletions, 1,056 (37%) duplications, and 1,275 (15%) insertions within our dataset (Figure 4A). Furthermore, we conducted a comparison of our SV results against SVs obtained in the HGSVG trios (nstd152) and found that 8,695 (35%) deletions, 1,099 (37%) duplications, and 6,436 (74%) insertions overlapped with the trios' dataset (Figure 4B). It can be noted here that the nstd152 dataset is derived from PacBio long-read sequencing, Illumina 3.5 kbp and 7.5 kbp jumping libraries, and optical mapping, with subsequent long-range phasing and haplotype structure determination. Comparing our study dataset with IndiGen-SV reveals overlaps of 8,697 (35%) deletions, 690 (24%) duplications, and 1,904 (22%) insertions (Figure 4C). Notably, taking all the SV types together, we observed 31% overlap between IndiGen and our study samples, possibly because the sampled individuals in these two datasets are from different ethnic groups across the country with IndiGen more biased toward northern India-based sample collection, thus highlighting the significant diversity among Indian populations, and the need to do more such SV discovery studies for comprehensive results. Then, we compared our dataset with gnomAD-SVs and observed 1,187 deletions overlapping, although the latter reports more than 169,000 deletions. Additionally, 157 duplications and 37 insertions from our dataset overlap with gnomAD (Figure 4D). This low level of match was expected because gnomAD contains about 329,000 SVs from individuals representing diverse world populations with an overrepresentation of European ancestry and an underrepresentation of Indian individuals. Nevertheless, the strongest positive correlation $r^2 = 0.88$ between allele frequencies of overlapping SVs are observed for the gnomAD dataset (Figure 4E). We also found a strong positive correlation of 0.7 (Figure 4F), 0.71 (Figure 4G) between allele frequencies of overlapping SVs discovered in 1000 Genomes-SAS SV and IndiGen-SV, respectively. Nonetheless, when we considered for the entire set of previously

(H) Distribution of allele frequencies among the SVs that overlap with a global dataset of SVs: The overlapping SVs were categorized into two groups based on our study frequency distribution: rare variants, present at less than 1%, and common variants, present at greater than 1%. Notably, a substantial proportion was found to be common within our study dataset—60% from gnomAD, 63% from 1000 Genomes-SAS, and 62% from IndiGen-SV—highlighted in dark gray. The remaining SVs that overlapped between datasets were considered rare within our study and are represented in light gray.

discovered structural variations from 1000 Genomes, and IndiGen-SV, that comprises individuals of Indian descent or South Asian ancestry, we observed that about 54% of the SVs overlap with our results, thereby rendering about 46% of our identified SVs to be currently specific to the Indian population. This result is particularly notable as these datasets encompass samples from the South Asian population, thus enhancing the generalizability and applicability of our study's findings to a broader genetic context.

We conducted further analysis on the allele frequencies of SVs that overlapped between our study dataset and a global dataset. These overlapping SVs were categorized into two groups based on our study frequency distribution: rare variants, present at less than 1%, and common variants, present at greater than 1%. Notably, a substantial proportion was found to be common within our study dataset—60% from gnomAD, 63% from 1000 Genomes-SAS, and 62% from IndiGen-SV—highlighted in dark gray. The remaining SVs that overlapped between datasets were considered rare within our study and are represented in light gray (Figure 4H). This conforms to the expectation that common variants will be ubiquitous in several populations, and rare variants will exhibit population-specific signatures.

Comparisons with EBI-GWAS catalog

We examined the impact of SVs on observable traits (phenotypes) by analyzing their genetic linkage with known trait-associated variants. From the entire downloaded EBI-GWAS catalog, our analysis identified 148 variants therein that are in strong genetic linkage (LD $r^2 \geq 0.7$) with 145 of our identified SVs. This suggests that these SVs may play a role in explaining the observed trait associations (Table S5). Among them, 44 SVs located in coding regions demonstrate strong LD with 42 variants from the GWAS catalog. Notably, within our study dataset, we found one deletion (chr4:11024191-11028005), reported for "Alzheimer disease or family history of Alzheimer disease"-associated SNP (chr4:11024404, genetic linkage with $r^2 = 1$), in two individuals and another insertion in four individuals (chr17:71643353-71643415), which is reported for "cognitive impairment (MoCA score) in Parkinson disease" (genetic linkage with $r^2 = 0.75$) associated SNP (chr17:71680028). To gain more insight into their cognitive status, we examined the *APOE* SNPs (rs429358 and rs7412, responsible for the *APOE* isoforms implicated in cognitive impairment) from our in-house SNPs dataset for these six individuals. We observed that only one individual (age = 55, female) exhibits the ε3/ε4 genotype known to increase the risk of developing Alzheimer disease and has a high score of 30 in the cognitive screening test HMSE (Hindi Mental State Examination). In the future, our goal is to closely follow up this individual and investigate the cognitive phenotypes to better understand the implications of these findings. Additionally, we observed a deletion in *KCNAB1* in strong LD with SNP associated previously to aging traits and another deletion in *LAMA1* in strong LD to SNP known to be associated previously to type 2 diabetes (Table S5). Moreover, we identified a deletion involving *FNTB* and *MAX* genes, which is in significant LD and associated with white blood cell count (Table S5). These results highlight the relevance of SVs in explaining various phenotypic associations and provide ground for inspecting these specific SVs as potential causal variants for the reported trait associations.

Comparison of SVs from array-based and WGS methods

In our study dataset, we identified a total of 4,803 duplications and 7,860 deletions from the genome-wide array based experimentally derived genotypes using the PennCNV method. To explore the relationship between array data deletions and duplications and our identified SVs, which consist of 24,574 deletions and 2,910 duplications, we conducted a comprehensive analysis. Our analysis revealed that 2,572 deletions and 454 duplications overlapped between the array data and WGS-based SVs. Additionally, we performed an in-depth analysis of the lengths of these overlapping SVs to understand which length categories exhibited more overlap. Interestingly, we observed significant overlap with array data for larger SVs. This especially is meaningful considering that array-based technologies for large variant detection is more sensitive toward uncovering longer (tens of kilobases) deletions and duplications,[53] thus corroborating the robust identification of longer SVs (length >10 kb–1 Mb) in our discovery set (Figure S6).

Discussion

The depiction of genetic variations in their entirety remains unfulfilled if we exclude SVs from population genetic sequencing studies. Moreover, SV plays crucial roles in genomic and cellular processes as well as disease associations. The Indian population has been severely underrepresented in genomic studies. Here, we report mapping and characterization of SVs in Indians from population-based cohorts identified by short-read high-coverage (average depth 42X) WGS. The robust analysis using multiple discovery tools upon deep coverage WGS data has empowered us to identify 36,210 SVs from the consensus of these methods, including mapping of rare SVs at high genomic resolution from 529 individuals across more than 30 distinct population subgroups (Figures S1A and S1B) representing a modest proportion of the rich genetic diversity of India. Previously, discovery of SVs in individuals of South Asian descent have been made in the 1000 Genomes study, focusing on a small number of individuals with ancestral connections to Gujaratis in west India and Telugus in southern India, in addition to Bengalis and Pakistanis. Another recent study has also contributed to SV discovery in the Indian population.[18] Upon cross-referencing our identified SVs with the SVs uncovered from these two studies taken together, we note that 54% of

our identified SVs overlap with previously known ones, and therefore 46% are unique to our study dataset, most of them being rare in prevalence. This is also perceivable from our resultant SVs where low-frequency or rare variants make up about half of the total 36,210 SVs discovered. Considering the diversity of India's population, the relatively moderate level of overlap that we observed emphasizes the distinctiveness of our results obtained from a different set of study individuals sampled from different geographic regions and probably belonging to different population subgroups compared with the IndiGen or 1000 Genomes. This phenomenon is also highlighted in the IndiGen study where they found that 55% of the uncovered SVs are unique to Indians.[18] It also underlines the importance of conducting exhaustive studies across various regions and including diverse ethnic groups to facilitate SV discovery in less studied non-European populations. We would like to note that the sample size is a limitation in this work, and 529 samples are not sufficient to capture the huge population size and rich genetic diversity of India for characterizing structural variation for the entire population. Also, the extant Indian population groups have evidence of their ancestry derived from the genetically distinct and divergent ancestral north Indians (ANIs) and the ancestral south Indians (ASIs),[54,55] and our samples mostly belong to the south Indian states with lesser representation from north Indian states. Nevertheless, it may be acknowledged that recruiting individuals from the community, data generation costs including high-coverage WGS and computational analysis burden as well as data storage are substantial barriers to feasibly achieving a large sample size for such studies. We expect to add more samples in the future in order to have a larger and improved map of SVs in the Indian population. We have also shown results from the single nucleotide variations analysis in our supplemental information, that more than 6% of them are unique when compared with the latest 1000 Genomes-SAS results. Our dataset most likely captures more of the ASI-specific genetic variations, and this group was earlier underrepresented in studies.

We report an average of 3,248 high-confident insertions, 3,010 deletions, and 193 duplications per genome in Indian individuals. At the individual level though, there is substantial heterogeneity in the distribution of SVs, for example, much fewer number of SVs are ubiquitously present in all the genomes, in contrast to a much greater number of variations carried by a fraction (less than 50 out of 529) of the individuals (77% of deletions, 84% of duplications, and 73% of insertions), which has been shown in other populations as well.[17] At the chromosomal level on average, deletions (69%) are the more predominant type of SVs, followed by insertions (21%), and duplications (9%). We note that the larger the variation, the less frequent it is in our study. SV duplications are observed to be mostly longer compared with other variations in our dataset. The median length of duplications (908 bp) is longer than that of deletions (310 bp) and insertions (164 bp). Most of the SVs are enriched in rare or ultra-rare variants (fewer than five individuals in our dataset), whereas only 26.91%, 19.4%, and 34.50% of deletions, duplications, and insertions, respectively, are commonly present in our study samples. It is encouraging to note that there is significant overlap between the SVs identified through WGS and array-based methods for the longer ones (>10 kb length, those might be spuriously identified by short-read WGS), which reiterates that multiple discovery methods are essential for consistent and reliable identification of a larger gamut of SVs at population scale.

We observed a lesser proportion 7.5% (=2,732/36,210) of SVs being carried in the homozygous state compared with the heterozygous state (21% from 7,582/36,210), and given that our study individuals have not been ascertained for particular diseases, this observation is aligned with the hypothesis that in autosomal genes in a randomly mating population, evolutionary selection acts in favor of heterozygotes so that there is one well-functioning copy of the gene, and homozygous carriers of the harmful variant are less prone to be observed in a non-diseased population.[56] On a deeper analysis of the homozygous SVs present in high LoF genes, we observe that the individuals carrying these variants do not show obvious manifestations of relevant disease phenotypes. This is not surprising given that potentially harmful genetic variants can be present in individuals in a general population, and the phenotypes could also represent a spectrum that makes them undetectable. Also, many such disease traits are complex and polygenic, and thus apparently healthy individuals in a general population will carry many such variants. Our detailed analysis to unravel the medical and clinical significance of identified SVs corroborates these ideas and highlights the importance of population-scale genomic sequencing all the more in order to delineate population-specific attributable genetic risk for several complex and rare disease conditions.

Of all the resultant SVs (36,210) in our study samples, 456 (1.263%) have the potential to cause frameshift changes in the genome. The majority of these frameshifts are inflicted by rare (387) SVs in our study dataset. Even though we identified 20,904 rare and ultra-rare SVs, only a minuscule fraction (387 out of 20,904 = 1.85%) can cause coding region disruptions in 529 individuals. Furthermore, we conducted an in-depth analysis of these frameshift SVs, specifically focused on SVs that were common, that is, present in at least 27 individuals, and map to disease database. We identified three such genes (*ADAMTS17, CCDC40,* and *RHCE*) known to be implicated in inherited disorders. However, the SVs we have identified mapping to these genes are present in heterozygous carriers, whereas the implicated diseases are mostly known to have autosomal recessive inheritance. Thus, manual inspection of zygosity, knowledge of disease diagnosis in study samples, and mode of inheritance for known disorders are crucial factors that ought to be considered in addition to automated disease annotation while characterizing genetic variants in population-based studies.

We found that 50 SVs are rare and highly intolerant to LoF. Also, we observed that six deletions and one duplication are rare and associated with dosage sensitivity for a loss of phenotype, exhibiting frameshift changes affecting the start and stop sites, indicating potential alterations in the protein-coding sequences. We also found that 44 SVs located in coding regions are in strong LD ($r^2 > 0.7$) with 42 variants reported for disease associations in the EBI-GWAS catalog. This could form the basis for future studies in the Indian and other world populations to help pinpoint causal variants and elucidate disease mechanisms better with larger sample sizes and targeted cohorts for the specific disease conditions.

Our results take a step toward realizing the goal of having an SV map for the Indian population. We uncovered rare as well as common unique SVs in Indians with robust sensitivity and specificity; this could facilitate elucidating the role of SVs in not only rare diseases but also complex disorders that affect the general population, akin to propositions in earlier studies.[57] While our study encompassed a considerable cohort of 529 individuals and utilized robust methodologies for data analysis, it is crucial to acknowledge that our sample might not entirely encompass the diverse spectrum of the diverse Indian population. Consequently, our results emphasize the necessity for future endeavors involving multicentric studies that encompass various regions and population subsets across India. Leveraging this previously undocumented genetic diversity from this genomic resource based on the Indian population will help contribute to the global genomic landscape as well. Differential causality and low penetrance of disease traits in different populations with considerable contributions from the environment and gene-environment interaction factors emphasizes the need for more such studies to proactively use genomic insights in medicine. Our in-depth characterization of identified SVs signifies that a comprehensive evaluation of the contribution of SVs to disease association studies will improve their power, thus furthering the medical genetics scenario for India and worldwide.

Data and code availability

All structural variations (deletion, duplication, insertions, and inversions) of our study samples (n = 529) are available in VCF formats via the CBR FTP site: https://ftp.cbr.res.in/.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2024.100285.

Acknowledgments

We are grateful to the volunteers who participated in the CBR-TLSA and CBR-SANSCOG studies. We acknowledge the valuable advice and academic input of CBR's International Advisory Board members. We are grateful to B.K. lab member Mr. Vishak Madhwaraj Kadambalithaya for his help in preparing Figure S1A, and Ms. Anupriya Sadhasivam for initial work on a few samples in this project, which has been repeated with more vigorous workflow for all the study samples. The TATA Longitudinal Study of Aging is supported by the Tata Trusts, India. TATA Trusts also funded the next generation sequencing (NGS) facility at CBR, IISc. The SANSCOG study is funded by CBR, IISc. Computational infrastructure used to perform analysis for this paper is supported by grants to B.K.: BT/RLF/Re-entry/29/2016, DST/NSM/R&D_HPC_Applications/2021/03.12, ECR/2018/001429. B.K. is also currently supported by DBT-Wellcome Trust India Alliance Intermediate Fellowship (Ref: IA/I/23/1/506750).

Declaration of interests

The authors declare no competing interests.

Received: September 14, 2023
Accepted: March 20, 2024

References

1. Freeman, J.L., Perry, G.H., Feuk, L., Redon, R., McCarroll, S.A., Altshuler, D.M., Aburatani, H., Jones, K.W., Tyler-Smith, C., Hurles, M.E., et al. (2006). Copy number variation: new insights in genome diversity. Genome Res. *16*, 949–961. https://doi.org/10.1101/gr.3677206.
2. Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., et al. (2010). Origins and functional impact of copy number variation in the human genome The WTCCC collaborated on array design. Validation experiments were performed by Europe PMC Funders Group. Nature *464*, 704–712. https://doi.org/10.1038/nature08516.
3. Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., Abyzov, A., Chris Yoon, S., Ye, K., Keira Cheetham, R., et al. (2011). Mapping Copy Number Variation by Population-Scale Genome Sequencing. https://doi.org/10.1038/nature09708.
4. Sudmant, P.H., Mallick, S., Nelson, B.J., Hormozdiari, F., Krumm, N., Huddleston, J., Coe, B.P., Baker, C., Nordenfelt, S., Bamshad, M., et al. (2015). Global diversity, population stratification, and selection of human copy-number variation. Science *349*, aab3761. https://doi.org/10.1126/science.aab3761.
5. Wheeler, E., Huang, N., Bochukova, E.G., Keogh, J.M., Lindsay, S., Garg, S., Henning, E., Blackburn, H., Loos, R.J.F., Wareham, N.J., et al. (2013). Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. Nat. Genet. *45*, 513–517. https://doi.org/10.1038/ng.2607.
6. Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J.O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. Nat. Rev. Genet. *14*, 125–138. https://doi.org/10.1038/nrg3373.
7. Chiang, C., Scott, A.J., Davis, J.R., Tsang, E.K., Li, X., Kim, Y., Hadzic, T., Damani, F.N., Ganel, L., Montgomery, S.B., et al. (2017). The Impact of Structural Variation on Human Gene Expression. https://doi.org/10.1038/ng.3834.
8. de Smith, A.J., Tsalenko, A., Sampas, N., Scheffer, A., Yamada, N.A., Tsang, P., Ben-Dor, A., Yakhini, Z., Ellis, R.J., Bruhn, L.,

et al. (2007). Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. Hum. Mol. Genet. *16*, 2783–2794. https://doi.org/10.1093/hmg/ddm208.

9. Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F.A., Hakonarson, H., and Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. *17*, 1665–1674. https://doi.org/10.1101/gr.6861907.
10. Pendleton, M., Sebra, R., Wing, A., Pang, C., Ummat, A., Franzen, O., Rausch, T., Stütz, A.M., Stedman, W., Anantharaman, T., et al. (2015). Assembly and Diploid Architecture of an Individual Human Genome via Single-Molecule Technologies, *780* (Articles). https://doi.org/10.1038/nmeth.3454.
11. Chaisson, M.J.P., Huddleston, J., Dennis, M.Y., Sudmant, P.H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., et al. (2014). Resolving the Complexity of the Human Genome Using Single-Molecule Sequencing. https://doi.org/10.1038/nature13907.
12. Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome Structural Variation Discovery and Genotyping (Nature Publishing Group). https://doi.org/10.1038/nrg2958.
13. MacDonald, J.R., Ziman, R., Yuen, R.K.C., Feuk, L., and Scherer, S.W. (2014). The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res. *42*, D986–D992. https://doi.org/10.1093/nar/gkt958.
14. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H.-Y., et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75–81. https://doi.org/10.1038/nature15394.
15. Almarri, M.A., Bergström, A., Prado-Martinez, J., Yang, F., Fu, B., Dunham, A.S., Chen, Y., Hurles, M.E., Tyler-Smith, C., and Xue, Y. (2020). Population Structure, Stratification, and Introgression of Human Structural Variation. Cell *182*, 189–199.e15. https://doi.org/10.1016/j.cell.2020.05.024.
16. Abel, H.J., Larson, D.E., Regier, A.A., Chiang, C., Das, I., Kanchi, K.L., Layer, R.M., Neale, B.M., Salerno, W.J., Reeves, C., et al. (2020). Mapping and characterization of structural variation in 17,795 human genomes A population-scale map of SVs. Nature *583*, 83–89. https://doi.org/10.1038/s41586-020-2371-0.
17. Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., et al. (2020). A structural variation reference for medical and population genetics Aggregation Database Production Team. Genome Aggregation Database Consortium*. *581*, 444–451. https://doi.org/10.1038/s41586-020-2287-8.
18. Divakar, M.K., Jain, A., Bhoyar, R.C., Senthivel, V., Jolly, B., Imran, M., Sharma, D., Bajaj, A., Gupta, V., Scaria, V., and Sivasubbu, S. (2023). Whole-genome sequencing of 1029 Indian individuals reveals unique and rare structural variants. J. Hum. Genet. *68*, 409–417. https://doi.org/10.1038/s10038-023-01131-7.
19. Wall, J.D., Stawiski, E.W., Ratan, A., Kim, H.L., Kim, C., Gupta, R., Suryamohan, K., Gusareva, E.S., Purbojati, R.W., Bhangale, T., et al. (2019). The GenomeAsia 100K Project enables genetic discoveries across Asia. Nature *576*, 106–111. https://doi.org/10.1038/s41586-019-1793-z.
20. Jain, A., Bhoyar, R.C., Pandhare, K., Mishra, A., Sharma, D., Imran, M., Senthivel, V., Divakar, M.K., Rophina, M., Jolly, B., et al. (2021). IndiGenomes: a comprehensive resource of genetic variants from over 1000 Indian genomes. Nucleic Acids Res. *49*, D1225–D1232. https://doi.org/10.1093/nar/gkaa923.
21. Nakatsuka, N., Moorjani, P., Rai, N., Sarkar, B., Tandon, A., Patterson, N., SriLakshmi Bhavani, G., Mohan Girisha, K., Mustak, M.S., Srinivasan, S., et al. (2017). The Promise of Discovering Population-specific Disease-Associated Genes in South Asia. https://doi.org/10.1038/ng.3917.
22. Ravindranath, V.; and SANSCOG Study Team (2023). Srinivaspura Aging, Neuro Senescence and COGnition (SANSCOG) study: Study protocol. Alzheimers Dement. *19*, 2450–2459. https://doi.org/10.1002/alz.12722.
23. Sundarakumar, J., Chauhan, G., Rao, G.N., Sivakumar, P.T., Rao, N.P., Ravindranath, V., and Investigators, S. and T. (2020). Srinivaspura Aging, Neuro Senescence and COGnition (SANSCOG) study and Tata Longitudinal Study on Aging (TLSA): Study protocols. Alzheimer's Dementia *16*, e045681. https://doi.org/10.1002/alz.045681.
24. Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. *2*, 190. https://doi.org/10.1371/journal.pgen.0020190.
25. Panda, A., Subramanian, K., and Kahali, B. (2021). Implementation of human whole genome sequencing data analysis: A containerized framework for sustained and enhanced throughput. Inform. Med. Unlocked *25*, 100684. https://doi.org/10.1016/j.imu.2021.100684.
26. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.
27. Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014). LUMPY: A Probabilistic Framework for Structural Variant Discovery. https://doi.org/10.1186/gb-2014-15-6-r84.
28. Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: Structural Variant Discovery by Integrated Paired-End and Split-Read Analysis, *28*, pp. 333–339. https://doi.org/10.1093/bioinformatics/bts378.
29. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220–1222. https://doi.org/10.1093/bioinformatics/btv710.
30. Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J.-P., and Wang, L. (2014). Genome Analysis CrossMap: A Versatile Tool for Coordinate Conversion between Genome Assemblies, *30*, pp. 1006–1007. https://doi.org/10.1093/bioinformatics/btt730.
31. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. BIOINFORMATICS APPLICATIONS NOTE *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.
32. Chiang, C., Layer, R.M., Faust, G.G., Lindberg, M.R., Rose, D.B., Garrison, E.P., Marth, G.T., Quinlan, A.R., and Hall, I.M. (2015). SpeedSeq: ultra-fast personal genome analysis and interpretation. Nat. Methods *12*, 966–968. https://doi.org/10.1038/nmeth.3505.
33. Macdonald, J.R., Ziman, R., Yuen, R.K.C., Feuk, L., and Scherer, S.W. The Database of Genomic Variants: A Curated Collection of Structural Variation in the Human Genome. 10.1093/nar/gkt958.
34. Karczewski, K.J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D.M., Kavanagh, D., Hamamsy, T., Lek, M., Samocha, K.E., Cummings, B.B., et al. (2017). The ExAC browser:

displaying reference data information from over 60 000 exomes. Nucleic Acids Res. *45*, D840–D845. https://doi.org/10.1093/nar/gkw971.
35. Amberger, J.S., and Hamosh, A. (2017). Searching Online Mendelian Inheritance in Man (OMIM): A Knowledgebase of Human Genes and Genetic Phenotypes. Curr. Protoc. Bioinformatics *58* (1), 1.2.1–1.2.12. https://doi.org/10.1002/cpbi.27.
36. Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J.D., Garner, J., Chen, C., Maguire, M., Corbett, M., Zhou, G., et al. (2013). DbVar and DGVa: public archives for genomic structural variation. Nucleic Acids Res. *41*, D936–D941. https://doi.org/10.1093/nar/gks1213.
37. Rivera-Muñoz, E.A., Milko, L.V., Harrison, S.M., Azzariti, D.R., Kurtz, C.L., Lee, K., Mester, J.L., Weaver, M.A., Currey, E., Craigen, W., et al. (2018). ClinGen Variant Curation Expert Panel experiences and standardized processes for disease and gene-level specification of the ACMG/AMP guidelines for sequence variant interpretation. Hum. Mutat. *39*, 1614–1622. https://doi.org/10.1002/humu.23645.
38. Babbi, G., Martelli, P.L., Profiti, G., Bovo, S., Savojardo, C., and Casadio, R. (2017). eDGAR: a database of Disease-Gene Associations with annotated Relationships among genes. BMC Genom. *18*, 554. https://doi.org/10.1186/s12864-017-3911-3.
39. UniProt Consortium (2015). UniProt: a hub for protein information. Nucleic Acids Res. *43*, D204–D212. https://doi.org/10.1093/nar/gku989.
40. Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. *44*, D862–D868. https://doi.org/10.1093/nar/gkv1222.
41. Redin, C., Brand, H., Collins, R.L., Kammin, T., Mitchell, E., Hodge, J.C., Hanscom, C., Pillalamarri, V., Seabra, C.M., Abbott, M.-A., et al. (2017). The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. Nat. Genet. *49*, 36–45. https://doi.org/10.1038/ng.3720.
42. Spielmann, M., Lupiáñez, D.G., and Mundlos, S. (2018). Structural variation in the 3D genome. Nat. Rev. Genet. *19*, 453–467. https://doi.org/10.1038/s41576-018-0007-0.
43. Zogopoulos, V., Spaho, K., Ntouka, C., Lappas, G., Kyranis, I., Bagos, P., Spandidos, D., and Michalopoulos, I. (2021). TFBSPred: A functional transcription factor binding site prediction webtool for humans and mice. Int. J. Epigen. *1*, 9. https://doi.org/10.3892/ije.2021.9.
44. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*. https://doi.org/10.1186/s13742-015-0047-8.
45. Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H.P., Bjornsson, E., Jonsson, H., Atlason, B.A., Kristmundsdottir, S., Mehringer, S., Hardarson, M.T., et al. (2021). 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. https://doi.org/10.1038/s41588-021-00865-4.
46. Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K., et al. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell *185*, 3426–3440.e19. https://doi.org/10.1016/j.cell.2022.08.004.
47. Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O.L., Guo, L., Collins, R.L., et al. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat. Commun. *10*, 1784. https://doi.org/10.1038/S41467-018-08148-Z.
48. Zhou, B., Arthur, J.G., Guo, H., Hughes, C.R., Kim, T., Huang, Y., Pattni, R., Lee, H., Ji, H.P., and Song, G. (2017). Automatic detection of complex structural genome variation across world populations. Preprint at bioRxiv. https://doi.org/10.1101/200170.
49. Firth, H.V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R.M., and Carter, N.P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Am. J. Hum. Genet. *84*, 524–533. https://doi.org/10.1016/j.ajhg.2009.03.010.
50. Alsamri, M.T., Alabdouli, A., Iram, D., Alkalbani, A.M., Almarzooqi, A.S., Souid, A.-K., and Vijayan, R. (2021). A Study on the Genetics of Primary Ciliary Dyskinesia. J. Clin. Med. *10*, 5102. https://doi.org/10.3390/jcm10215102.
51. Nmezi, B., Giorgio, E., Raininko, R., Lehman, A., Spielmann, M., Koenig, M.K., Adejumo, R., Knight, M., Gavrilova, R., Alturkustani, M., et al. (2019). Genomic deletions upstream of lamin B1 lead to atypical autosomal dominant leukodystrophy. Neurol. Genet. *5*, e305. https://doi.org/10.1212/NXG.0000000000000305.
52. Giorgio, E., Robyr, D., Spielmann, M., Ferrero, E., Di Gregorio, E., Imperiale, D., Vaula, G., Stamoulis, G., Santoni, F., Atzori, C., et al. (2015). A large genomic deletion leads to enhancer adoption by the lamin B1 gene: a second path to autosomal dominant adult-onset demyelinating leukodystrophy (ADLD). Hum. Mol. Genet. *24*, 3143–3154. https://doi.org/10.1093/hmg/ddv065.
53. Zhang, X., Du, R., Li, S., Zhang, F., Jin, L., and Wang, H. (2014). Evaluation of copy number variation detection for a SNP array platform. BMC Bioinf. *15*, 50. https://doi.org/10.1186/1471-2105-15-50.
54. Reich, D., Thangaraj, K., Patterson, N., Price, A.L., and Singh, L. (2009). ARTICLES Reconstructing Indian population history. Nature *461*, 489–494. https://doi.org/10.1038/nature08365.
55. Moorjani, P., Thangaraj, K., Patterson, N., Lipson, M., Loh, P.-R., Govindaraj, P., Berger, B., Reich, D., and Singh, L. (2013). Genetic Evidence for Recent Population Mixture in India. Am. J. Hum. Genet. *93*, 422–438. https://doi.org/10.1016/j.ajhg.2013.07.006.
56. Jiggins, C. (2010). Elements of Evolutionary Genetics. B. Charlesworth & D. Charlesworth. Roberts & Company. 2010. 768 pages. Price $80 (hardback). Genet. Res. (Camb) *92*, 323. https://doi.org/10.1017/S001667231000042X.
57. Stankiewicz, P., and Lupski, J.R. (2010). Structural Variation in the Human Genome and its Role in Disease. Annu. Rev. Med. *61*, 437–455. https://doi.org/10.1146/annurev-med-100708-204735.