A Catalogue of Structural Variation across Ancestrally Diverse Asian Genomes

Nicolas Bertin (Nicolas_Bertin@gis.a-star.edu.sg)
Genome Institute of Singapore, A*STAR

Joanna Tan
Genome Institute of Singapore, A*STAR

Zhihui Li
Genome Institute of Singapore, Agency for Science, Technology and Research

Mar Gonzalez-Porta
Genome Institute of Singapore, Agency for Science, Technology and Research

Ramesh Rajaby
https://orcid.org/0000-0001-9980-1913

Rodrigo Jimenez
Genome Institute of Singapore, A*STAR

Weng Khong Lim
SingHealth Duke-NUS Institute of Precision Medicine https://orcid.org/0000-0003-4391-1130

Ye An Tan
National University of Singapore and National University Health System, Singapore

Renyi Teo
Genome Institute of Singapore, A*STAR

Maxime Hebrard
Genome Institute of Singapore (GIS) https://orcid.org/0000-0003-2346-4513

Jack Ling Ow
Genome Institute of Singapore, A*STAR

Shimin Ang
Genome Institute of Singapore, A*STAR

Justin Jeyakani
Genome Institute of Singapore

Yap Seng Chong
National University of Singapore

Tock Han Lim
National Healthcare Group Eye Institute

Liuh Ling Goh
Tan Tock Seng Hospital https://orcid.org/0000-0002-1387-3682

Yih-Chung Tham
Singapore National Eye Centre

Khai Pang Leong
Tan Tock Seng Hospital

Calvin Chin
National Heart Centre Singapore

Sonia Davila
SingHealth Duke-NUS

neer Gilmore
UNC

Ching-Yu Cheng
Singapore National Eye Centre
John Chambers
Nanyang Technological University, Lee Kong Chian School of Medicine
E Shyong Tai
Saw Swee Hock School of Public Health, National University of Singapore and National University Health System https://orcid.org/0000-0003-2929-8966
Jianjun Liu
Genome Institute of Singapore https://orcid.org/0000-0002-3255-3019
Xueling Sim
National University of Singapore https://orcid.org/0000-0002-1233-7642
Wing-Kin Sung
Genome Institute of Singapore
Shyam Prabhakar
Genome Institute of Singapore https://orcid.org/0000-0002-6409-0661
Patrick Tan
Genome Institute of Singapore, Agency for Science, Technology and Research (A*STAR)
Jin-Fang Chai
Saw Swee Hock School of Public Health, National University of Singapore
Jimmy Lee
Institute of Mental Health https://orcid.org/0000-0002-7724-7445
Eng Sing Lee
National Healthcare Group https://orcid.org/0000-0003-4963-535X
Joanne Ngeow
Lee Kong Chian School of Medicine, Nanyang Technological University https://orcid.org/0000-0003-1558-3627
Paul Elliott
Imperial College London https://orcid.org/0000-0002-7511-5684
Elio Riboli
Imperial College London https://orcid.org/0000-0001-6795-6080
Hong Kiat Ng
Nanyang Technological University, Lee Kong Chian School of Medicine
Theresia Mina
Nanyang Technological University, Lee Kong Chian School of Medicine
Darwin Tay
Nanyang Technological University, Lee Kong Chian School of Medicine
Nilanjana Sadhu
Nanyang Technological University, Lee Kong Chian School of Medicine
Pritesh Rajesh Jain
Nanyang Technological University, Lee Kong Chian School of Medicine
Dorrain Low
Nanyang Technological University, Lee Kong Chian School of Medicine https://orcid.org/0000-0002-7742-3189
Xiaoyan Wang
Nanyang Technological University, Lee Kong Chian School of Medicine
Khung Keong Yeo
National Heart Centre Singapore https://orcid.org/0000-0002-5457-4881
Stuart Alexander Cook
National Heart Centre Singapore
Chee Jian Pua
National Heart Center https://orcid.org/0000-0003-4683-3043
Chengxi Yang

National Heart Centre Singapore
Tien Wong
Singapore National Eye Center
Charumathi Sabanayagam
Singapore Eye Research Institute https://orcid.org/0000-0002-4042-4719
Lavanya Raghavan
Singapore National Eye Centre
Tin Aung
Singapore Eye Research Institute
Miao Ling Chee
Singapore Eye Research Institute
Miao Chee
12. Singapore Eye Research Institute, Singapore National Eye Centre
Hengtong Li
Singapore National Eye Centre
Rob van Dam
National University of Singapore and National University Health System
Yik-Ying Teo
National University of Singapore
Chia Wei Lim
Tan Tock Seng Hospital
Pi Kuang Tsai
Tan Tock Seng Hospital
Wen Jie Chew
Tan Tock Seng Hospital
Wey Ching Sim
Tan Tock Seng Hospital
Li-xian Grace Toh
Tan Tock Seng Hospital
Johan Eriksson
Singapore Institute for Clinical Sciences, Agency for Science, Technology and Research
Peter Gluckman
University of Auckland https://orcid.org/0000-0002-5711-1655
Yung Seng Lee
Yong Loo Lin School of Medicine, National University of Singapore, Singapore https://orcid.org/0000-0002-1253-0557
Fabian Yap
KK Women's and Children's Hospital https://orcid.org/0000-0003-1083-7958
Kok Hian Tan
https://orcid.org/0000-0003-1945-0266

Article

Keywords:

Posted Date: October 3rd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3376868/v1

Abstract

Structural variants (SVs) are significant contributors to inter-individual genetic variation associated with traits and diseases. Current SV studies using whole-genome sequencing (WGS) have a largely Eurocentric composition, with little known about SV diversity in other ancestries particularly from Asia. Here, we present a WGS catalogue of 152,655 SVs from 8,392 Singaporeans of East Asian, Southeast Asian and South Asian ancestries, of which ~75% (113,446 SVs) are novel. We show that Asian populations can be stratified by their global SV patterns and identified 82,003 novel SVs that are specific to Asian populations. 38% of these novel SVs are restricted to one of the three major ancestry groups studied (Indian, Chinese or Malay). We uncovered SVs affecting ACMG-defined clinically actionable loci. Lastly, by identifying SVs in linkage disequilibrium with single-nucleotide variants, we demonstrate the utility of our SV catalogue in the fine-mapping of Asian GWAS variants and identification potential causative variants. These results augment our knowledge of structural variation across human populations, thereby reducing current ancestry biases in global references of genetic variation afflicting equity, diversity and inclusion in genetic research.

Introduction

Human genomic variation plays a critical role in health and disease, making its study a vital area of biological and medical research[1,2]. To improve our understanding of genetic variation across diverse human genomes and populations, international consortia such as the 1000 Genomes Project[3] (1KGP), Genome Aggregation Database (gnomAD)[4], and national efforts such as the U.K. 100,000 Genomes Project[5] and NIH's All of Us program[6] have reported large-scale population-based sequencing efforts to comprehensively delineate common and rare genetic mutations across different geographies and ancestry groups. Currently, most of these studies have focused primarily on base-pair level variations such as single nucleotide polymorphisms (SNPs) and short insertions/deletions (indels)[3,4,7]. Recently, structural variants (SVs) have emerged as another important source of variation[8,9]. SVs are genome rearrangements >50bp and can be classified into different classes such as deletions, duplications, insertions (including mobile element insertions), translocations and inversions[10]. Different classes of SVs have been proposed to arise through a variety of mechanisms, including non-allelic homologous recombination or mobile element insertion events[11].

With the availability of whole-genome sequencing (WGS) and development of SV calling algorithms, researchers are increasingly leveraging short-read WGS data to characterise the spectra of human SVs. In 2015, the 1000 Genome Project[12] analyzed 2,504 low-pass genomes (~7x coverage) to discover 68,818 SVs affecting 2.5x more base pairs in the genome compared to SNPs. The gnomAD-SV project (gnomAD)[10] identified 335,470 SVs from 14,891 WGS samples, clarifying the impact of SVs in different portions of the genome and generating SV catalogues to facilitate identification of SVs associated with medical and phenotypic traits. Some phenotypically/medically relevant SVs include Chr17p11.2 duplications leading to *PMP22* gene overexpression and Charcot-Marie-Tooth disease (an inherited neurological disorder)[13], and Chr7 deletions affecting the *ELN* (Elastin) gene associated with Williams neurodevelopment syndrome[14]. Some SVs may be pleiotropic, such as the aforementioned Chr7 deletions which are associated with autism[15], schizophrenia[16] and cancer[17]. Knowledge of SVs can also improve our understanding of human evolution, as some SVs display population and ancestry-specific patterns[10,12]. For instance, amylase, a key enzyme involved in the digestion of starch has a higher copy number in Asian populations where rice (starch) is a staple food[18]. These studies highlight the importance of characterising the diversity of SV landscapes on a global scale.

Asia accounts for 60% of the world population, however, many of the current large-scale SV profiling projects have focused on individuals of European ancestry, resulting in an under-representation of SVs reflective of Asian populations (gnomAD: 1,304 Asian genomes; 8%, 1000 Genomes Project: 993 Asian genomes). Moreover, despite recent efforts to close this gap, current SV studies of Asian populations are still of limited sample size and have focused on single ancestry groups[19,20].

Singapore is a multi-ancestry country populated by individuals of Indian, Chinese and Malay ethnicity due to its immigration history. Majority of the residents (~74%)[21] in Singapore are Chinese, who are mainly descendants of Han Chinese from the southern provinces of China[22]. Malays represent 13.6%[21] of the population forms the second largest ethnic group in Singapore. The Malay community in Singapore are mainly descendants of Austronesian people in Southeast Asia, particularly from Malaysia and Indonesia. Lastly, Indians form the third largest ethnic group in Singapore. Majority of the Indians in Singapore are descendants of Indian migrants from south-eastern part of India[22]. Given the genetic diversity of the population, Singapore can serve in the first

approximation as a snapshot of East Asian, South-East Asia, and South Asia populations, and is uniquely suited for cataloguing Asian SV landscapes and genomic variation.

The Singapore Genome Variation Project (SGVP)[22], the SG10K_Health[23] and the SG10K_med[24] projects, which focussed on small variants (SNP and lesser than 50bp long indels) have previously demonstrated the value of Singaporean genomes for precision medicine. Here, we describe one of the first and to our knowledge the largest multi-ancestry study of SVs in Asians. Using WGS data from 8,392 individuals (SG10K_Health) along with specialized SV-calling tools, we identified and characterized SVs in these three Asian populations and related these SVs to regulatory and biological effects. Our results contribute to the growing body of research on SVs and fill a critical gap in deciphering the genomic variation landscape across Asian populations.

Results

SV Catalogues of Three Major Ancestry Groups

We analysed Illumina short-read WGS data of 9,770 samples from the SG10K_Health study[23], comprising participants of Chinese (58%), Indians (24%) and Malays (18%) ethnicities. After CRAM-level quality control (QC) and removing samples failing at least 1 of 9 QC metrics (Methods), 8,392 samples were retained. This data set is subsequently referred to as SG10K Structural Variant release 1.3 ("SG10K-SV-r1.3"). Besides Chinese and Indians which other groups have studied[19,20], SG10K-SV-r1.3 contains 1,620 individuals of Malay ancestry (**Supplementary Table 1**), a population which have to date not been included in previous large population-based SV studies[10,12].

The SG10K-SV-r1.3 dataset comprises multiple sub-cohorts sequenced at heterogeneous depths and using different library construction methods, which can influence SV detection accuracy (**Supplementary Table 1**). To ensure robust SV analysis and to reduce technical confounding factors, we first focused on a discovery cohort of 5,487 individuals' derived WGS libraries both constructed (PCR plus) and sequenced at an average depth of 15x in a consistent fashion. Other datasets, which included 1,523 individuals sequenced at a depth of 15x using a PCR-less WGS library construction method (referred to as "15x_validation") and 1,922 individuals sequenced at a depth of 30x using a PCR-employing WGS library construction method (referred to as "30x_validation") were used as validation datasets to ensure that results observed in the discovery dataset are reproducible. Even when confined to the discovery cohort alone, this study represents one of the largest Asian SV studies to date, covering over 4.21 times as many individuals of Asian ancestries as previous studies (**Figure 1A**).

We focused on the three most common SV types: deletions, insertions, and duplications (**Figure 1B**, **Supplementary Figure 1**, Methods). Due to their distinct genomic properties, it is challenging to accurately identify SVs using a single analytic tool[25], and most previous SV cataloguing efforts have employed a combined suite of SV class-specialized algorithms[10,12]. In this study, we employed Manta[26] to identify deletions and insertions separately in single samples, followed by SVimmer[27] to obtain a putative cohort-wide consensus set. Individual-level genotype calls within this uniformly-defined discovery SV set were then refined using Graphtyper2[28]. In addition, to address previously-reported limitations of these tools[29], we also used MELT[30], an algorithm specially designed for identifying mobile element insertion (MEIs) events.

As the Manta-SVimmer-Graphtyper SV pipeline relies solely on discordant read pairs and split-read alignments, it has inherent limitations to accurately detect duplication events created by the presence of tandem repeat sequences (e.g., microsatellites and minisatellites)[31,32]. We thus complemented the above algorithms with SurVIndel2[33], an in-house developed algorithm that can detect duplication events at high sensitivity (**Figure 1C**). To demonstrate the robustness of SurVindel2, we assessed false discovery rate (FDR) and true positive (TP) statistics for duplications relative to Manta-SVimmer-Graphtyper, against a truth set of high quality SVs obtained by haplotype-resolved long-read sequencing of a selected subset of 1000 Genomes Project analyzed samples[34]. We measured an average per-sample duplication identification FDR of 12% and 36% for SurVindel2 and Manta-SVimmer-Graphtyper, respectively. SurVIndel2 yielded a better sensitivity than Manta-SVimmer-Graphtyper (**Figure 1D, Supplementary Table 2**). Furthermore, the gains in sensitivity were more pronounced for tandem repeats (**Supplementary Figure 2**).

Using this pipeline, we identified 152,655 SVs comprising 35,584 insertions (including MEIs), 84,607 deletions, and 32,464 duplications. Approximately 75% of these events were novel (**Figure 1E**) with respect to gnomAD-SV[10], reflecting the potential for new discoveries by analysing underrepresented Asian genomes (a more detailed comparison of SG10K-SV-r1.3 to gnomAD-SV is reported

in later sections). We also detected 84,336 and 103,183 SG10K-SV-r1.3 SVs in the 15x PCR minus and 30x PCR plus validation dataset, respectively.

SG10K_Health SV Landscape

On average, each SG10K_Health individual harboured SVs covering 0.41% of the genome, with 1,905 insertions (0.017%), 2,486 deletions (0.367%), and 1,103 duplications (0.030%). These figures were consistent across all three ancestries (**Figure 2A**). Compared to gnomAD-SV, we detected fewer insertions and deletions per individual (insertions: 1,905 in SG10K_Health vs 2,612 in gnomAD-SV; Deletions: 2,486 vs 3,505), likely due to the higher sequencing depth of gnomAD samples (32x [10]). Confirming this hypothesis, we detected comparable insertion/deletion counts per individual in our 30x_validation dataset (2,751/5,692; **Supplementary Figure 3**). However, despite lower sequencing depth in our discovery cohort, we detected comparable numbers of duplications compared to gnomAD-SV (1,103 vs 1,346), likely reflecting the improved sensitivity of the SurVIndel2 duplication-detection pipeline. Similar to previous studies[10], the majority (>70%; 107,548) of deletions, insertions and duplications were rare events with allele frequencies (AF) less than or equal to 1% (**Figure 2B, Supplementary Figure 4**). Nevertheless, we identified 700 SVs with allele frequency greater than 0.95 in our discovery cohort; in these cases, the reference genome bears the minor allele.

While most detected SVs were small (10kbp; **Figure 2C**), we identified 6,444 deletions and 2,065 duplications longer than 10kbp. There was a striking abundance of SVs at 300bp, 2kb and 6kb (**Figure 2C**). The 300bp and 6kb insertions corresponded to Alu and LINE1 elements respectively, the two most abundant classes of transposable elements in the human genome (~11%[35] and ~17%[36] of the genome). The 2 kb SVs represent composite SVA (SINE, Variable Number Tandem Repeat, and Alu) transposons. These results highlight the pervasive contribution of repeat elements (Alu, LINE1, SVAs) in sculpting human genomic variation, and high-level similarities between our SV catalogue and other studies[12].

SVs have been reported to cluster at specific genomic regions (“hotspots”). Several factors have been proposed to influence the location of SV hotspots, such as segmental duplications and the local presence of transposable elements[37]. These factors may contribute to SV formation due to their higher propensity for DNA breakage and repair, with local transposable elements increasing the likelihood of non-allelic homologous recombination (NAHR)[38]. To identify SV hotspots, we employed hotspotter[39] (bandwidth:200,000, num.trial=10,000, pval=5 X 10^{-3}) and identified 331 regions containing higher-than-expected SV densities (**Supplementary Table 3**). Together, these 331 regions affected ~303Mbp, in line with previous findings[34]. Notably, 28.1% (93 out of 331) of the hotspot regions are located within 5Mbp of the ends of the chromosomes as well as near the centromeric regions. Excluding these sub-telomeric and centromeric hotspots, 122 hotspots were unique to SG10K-SV compared to gnomAD-SV. For example, we identified a 725,560bp (chr12:124034930-124760490) hotspot region containing 89 SVs. This hotspot overlaps the *NCOR2* (Nuclear receptor corepressor 2) gene, a corepressor that is frequently altered in prostate cancer[40].

Impact of SVs on Regulatory Elements and Gene Bodies

To assess the impact of SG10K-SV-r1.3 on different categories of functional genomic regions, we overlapped the SVs with gene regulatory elements identified by ENCODE and the Epigenomics Roadmap project[41]. Regulatory elements surveyed included 926,535 putative regulatory elements annotated as distal enhancers (667,599), proximal enhancers (141,830), insulators (CTCF sites, 56,766), promoters (34,803), poised elements (exhibiting DNase I hypersensitivity but are likely functionally gated by additional trans-acting signals), and non-promoter K4me3 regions (25,537)[41].

Common deletions (AF≥1%) were significantly depleted at putative enhancers and insulators, consistent with a model of negative selection acting on alterations affecting gene expression (**Figure 3A**). In contrast, rare (4592; 1%> AF >=0.1%) and ultra-rare (12,705; AF <0.1%) deletions did not exhibit similar depletion signals - it is possible that these latter SVs may have arisen later in human evolution with insufficient time for purifying selection. Promoter regions exhibited a trend (albeit not significant) for enrichment in common deletions, duplications and insertions, perhaps reflecting the higher GC content of promoters and susceptibility to errors in DNA replication[42]. Common duplications were also significantly depleted at distal and proximal enhancers (**Figure 3A**) again suggesting the action of purifying selection, and ultrarare duplications were also depleted at enhancers, though not as strongly as common duplications. Following a similar qualitative trend, common insertions (MEIs) were more strongly depleted at enhancers, insulators and K4me3 regions than ultrarare insertions.

Unexpectedly, we observed common duplications being enriched at annotated non-promoter H3K4me3 regions. To deepen this observation, we examined the intersect of 81 non-promoter H3K4me3 regions overlapping common duplications, and found that they were highly and significantly enriched for tandem repeats relative to all 25,537 H3K4me3 regions (fold enrichment: 4.6 : hypergeometric p-value: 2.45×10^{-23}). We speculate that since read mapping artifacts are common at tandem repeats, it is possible that these genome duplications may have contributed, at least in part, a degree of artifactual H3K4me3 ChIP-seq peaks. These results highlight how more refined genomic annotations taking SVs into account can improve the accuracy of other orthogonal data sets such as regular ChIP-seq maps.

We then analyzed gene bodies (UTRs, CDS, exons or introns). SVs of all three categories were strongly depleted at gene bodies, including 3'UTRs, 5'UTRs, CDS, exons, and introns (**Figure 3B**). For example, common insertions were depleted 11-fold at coding exons, against reflecting high selection pressure on coding sequences. Similar to enhancers, rare and ultrarare SVs showed weaker depletion patterns in exons of all types. Interestingly, intronic regions showed no deviations from background, except for a modest elevation in rare and ultrarare insertions. This may reflect the increased propensity of certain MEIs families to insert into the gene bodies of actively transcribed genes or GC-rich regions[43,44].

SVs deleting gene regions may cause complete or partial loss of function (LOF) effects. Conversely, duplications may lead to gene copy gain, augmenting gene dosage. Employing SVTK[45] to assess the potential impact of the SG10K_Health SVs on protein-coding regions, we identified 5,438 SVs (3.5% of 152,655) with direct predicted impact on protein coding integrity (**Figure 3C**). Of these, 4,143 SVs resulted in likely gene LOF. LOF-associated SVs tended to occur at low allele frequencies (AF<1%). We identified 1,023 duplications predicted to cause copy number gain of one or several consecutive protein-coding genes. Copy number gain events were typically larger compared to LOF events (median size 96kb vs 4.6kb). These patterns are in line with gnomAD where the majority of protein coding affecting SVs resulted in LOF, and copy gain events exhibited larger sizes.

We assessed the potential impact of SVs on clinically actionable genes, focusing 78 actionable genes (ACMG v3.1[46]) associated with highly penetrant and actionable genetic conditions. AnnotSV[47] was used to identify SVs potentially affecting at least one ACMG v3.1 gene. We found 35 SVs affecting coding sequence integrity in 21 clinically actionable ACMG genes. For example, we identified two separate 2.3kb and 9.4kb deletions in 5 and 3 Chinese individuals, affecting *TRDN* (**Figure 3D**), encoding triadin and a key component of the calcium release complex[48]. We also found, in 48 individuals (27 Chinese, 12 Indian, 9 Malay) a 134bp deletion affecting *DSG2* (**Supplementary Figure 5**), an essential component of desmosomes that provides mechanical strength and stability to heart and skin tissues[49]. Both *TRDN* and *DSG2* genes have been associated with severe cardiac dysfunction (catecholaminergic polymorphic ventricular tachycardia (CPVT), and arrhythmogenic right ventricular cardiomyopathy (ARVC) respectively).

SV Patterns Between International Cohorts

Reflecting the novelty of the SG10K-SV-r1.3 catalog, 74.3% (113,446/152,655) of the catalog were not previously reported in gnomAD-SV (**Figure 2A**). To compare the SG10K-SV-r1.3 cohort with gnomad-SV more stringently, we then applied a 50% call rate cut-off across each population within SG10K_SV, resulting in 85,162 SVs not exhibiting any overlap with gnomAD-SV. We hereby refer to these SVs as novel “Asian-specific" SVs. This included 47,064 deletions, 20,462 duplications, and 17,636 insertions. The majority of novel Asian-specific SVs were detected at lower allele frequencies compared to SVs commonly found both in SG10K_SV and gnomAD-SV (**Supplementary Figure 6**). Additionally, we identified 39,209 SV events in SG10K-SV which overlaps gnomAD-SV events. We further focused on this subset to identify events with a higher prevalence in Asian populations, employing Fst[50] analysis on the gnomAD-SV dataset as described in the Methods section. Using this approach, we further detected 32,085 events in gnomAD-SV displaying such differences which overlaps with 14,198 SV in SG10K-SV dataset

Notable examples of Asian-specific events include a previously reported 2,903 bp deletion in intron 2 of the *BIM* gene, which is associated with resistance to tyrosine kinase inhibitors[51]. This SV is present in gnomAD-SV at a higher AF in East-Asians compared to other ethnicities (AF EAS: 7.37×10^{-2}, AF others: 1.04×10^{-4}). Another example comprises a rare 19.3kbp deletion spanning the *HBA1* and *HBA2* genes, associated with α-thalassemia and detected more frequently in Asian populations (AF EAS: 9.93×10^{-3}, AF others: 1.04×10^{-4})[20]. It is worth noting that without the availability of SG10K-SV as a tool to enrich for Asian events, identifying these SV AF differences solely through an internal comparison of the gnomAD-SV database would have been challenging due to the large number of events in gnomAD-SV. Specifically, conducting an equivalent analysis on the entire gnomAD-SV database would

have yielded 1,647 events with significant AF differences between Asian and non-Asian populations. These 1,647 events represent 0.542% of the queried events, in contrast to the 21% (8,285/39,209) obtained when utilizing a comparison between SG10K-SV and gnomAD-SV. This approach serves two important purposes. Firstly, it identifies new events that had not been previously discovered, which constitute the majority of our findings. Secondly, it provides an enhanced framework to facilitate the detection of Asian-specific events within existing published resources.

SVs Between Asian Ancestry Groups

We then investigated SV patterns distinctive to the three major Asian ancestries. Principal components analysis (PCA) on the full set of SG10K_SV demonstrated ancestry-specific population clustering (**Figure 4A**), which was further replicated using either insertions, deletions, or duplication events (**Supplementary Figure 7**). These results support pervasive differences across the three SV classes contributing to population differentiation. 38% of SVs were seen in only one ancestry, 15% were shared across two ancestries, and half of the SVs (47%) were in all three populations (**Figure 4B**). However, as the numbers of events detected as unique in a population correlated with cohort size (**Supplementary Table 1**) and were enriched for low-frequency SVs (**Supplementary Figure 8**), it remains possible that some of these SVs may be present in other populations, but remain undetected due to low allele frequency.

To gain a more granular understanding of ancestry-specific SV patterns, we calculated fixation indexes (Fst)[50] for each of the detected events and assigned a significance score to each observation using permutation analysis (see **Methods**). By examining the resulting Fst trends, we found that SVs with extreme Fst values (0.7 and above) were mostly detected in small numbers of individuals (call rate < 2%) not reaching significance thresholds (**Figure 4C**). Of SVs exhibiting statistically significant Fst values, we identified 18,076 SVs displaying ancestry-specific frequency patterns, comprising 8,346 deletions, 5,660 insertions, and 4,070 duplications (see **Methods, refer to VCF in SG10K_Health Chorus Browser**).

The set of 18,076 ancestry-specific SVs were further filtered to those annotated to harbour functional consequences (**Figure 3C**). This analysis yielded a subset of 143 ancestry-specific SVs, comprising 91 deletions, 47 duplications, and 5 insertions (**Figure 4C, Supplementary Table 4**). By plotting AF trends for the top 50 events with highest Fst (**Figure 4D**), we observed ancestry-specific events across a range of allele frequencies. SVs associated with Indian ancestry drove differences for the majority of the cases (N=45/50). We also identified 10 SVs for which the GRCh38 reference genome contained the minor allele (MAF>50% in at least one population), underscoring the importance of moving beyond a single human genome reference[52] to establish reference genomes better reflecting the genetic diversity of global human populations.

Curation of the 143 events confirmed previously reported ancestry-specific SVs. For example, we observed a 27.6kbp deletion in the *ACOT1* gene, involved in fatty acid metabolism. This ancient deletion is marked by significant AF differences between continents, almost reaching fixation in Asian populations[53]. The *ACOT1*-associated deletion exhibited a lower AF in Indians (AF SG-Chinese: 0.873, AF SG-Indian: 0.532, AF SG-Malay: 0.769). Another example was a 32kbp deletion in *CYP2A6*, a member of the cytochrome P450 (CYP-450) superfamily involved in drug metabolism[54]. This SV also exhibited significantly lower AFs in Indians (AF SG-Chinese: 0.139, AF SG-Indian: 0.045, AF SG-Malay: 0.171). A third example was a 21.5kbp duplication overlapping *MPV17L2* and *RAB3A*, present in gnomAD-SV East Asians but rare in other ancestries[55]. We observed a similar frequency for Chinese, with lower AFs in Malay and Indians (AF SG-Chinese: 0.030, AF SG-Indian: 0.002, AF SG-Malay: 0.011). There was also a 204.4kbp duplication overlapping multiple genes in chromosome 4, observed only in SG-Indians (AF SG-Chinese: 0, AF SG-Indian: 0.015, AF SG-Malay: 0.002). While not identified in gnomAD-SV, we confirmed detection of this SV in 12 individuals from the 1KGP dataset, all of South-Asian ancestry.

Importantly, we also discovered previously unreported SVs. One such event was a 3kbp deletion overlapping *AHNAK2*, encoding a nucleoprotein involved in calcium signalling. The AF of this SV was higher in Indians compared to Chinese and Malay (AF SG-Chinese: 0.018, AF SG-Indian: 0.117, AF SG-Malay: 0.014). We also detected a 59bp deletion in *TNNT3*, which encodes Troponin T3, a protein involved in muscle contraction and distal Arthrogryposis[56,57]. This event was detected with the highest AF in Malays (AF SG-Chinese: 0.005, AF SG-Indian: 0.032, AF SG-Malay: 0.056). Other Malay-specific deletions involved *OR2B2* (1.3kbp; AF SG-Chinese: 0.005, AF SG-Indian: 0.003, AF SG-Malay: 0.029) and *FAM3B* (1.6kbp; both with AF SG-Chinese: 0.002, AF SG-Indian: 0.002, AF SG-Malay: 0.028). *OR2B2* encodes an olfactory receptor, a gene family known for population stratification, whilst *FAM3B* encodes a secreted cytokine-like protein involved in glucose metabolism and linked to type 2 diabetes. One more Malay specific insertion included a 209 bp event overlapping *CEACAM3* (AF SG-Chinese: 0.05, AF SG-Indian: 0.07, AF SG-Malay: 0.11), a cell adhesion

molecule that plays a crucial role in the innate immune response to bacterial infections. Finally, we identified a 12.2kbp duplication overlapping *CLPSL1* and *CLPS*, which encode enzymes involved in the digestion of dietary proteins. The AF of this duplication was lower in Chinese compared to Malay and Indian individuals (AF SG-Chinese: 0.52, AF SG-Indian: 0.60, AF SG-Malay: 0.62). The observed minor AF was greater than 50% in all three populations, indicating that this duplication is a common event. Collectively, our analyses demonstrate that numerous population-specific SVs among Asians can be detected using SG10K-SV.

SVs Exhibit *cis*-linkage to Disease GWAS Loci

Finally, SVs are gaining prominence as potential genetic drivers of disease susceptibility, drug response and other phenotypes[58]. To explore potential associations between SVs and biological phenotypes, we hypothesized that certain trait-associated lead SNPs identified by GWAS (GWAS-lead SNPs) might not (and indeed often do not) represent the actual causative variant. Conventional GWAS analysis thus often requires pinpointing underlying causal variants using fine-scale genetic mapping to assess variants showing high linkage disequilibrium (LD) with lead SNPs. Since SVs are large variants in terms of genomic span, it is possible that certain SVs in strong LD with GWAS lead SNPs might also be causative[59].

To explore this possibility, we performed LD analysis between SG10K-SVs and previously reported SG10K_Health SNPs/short indels inferred from WGS[23]. LD was computed for high-confidence (call rate > 0.8) common (MAF>1%) SVs (n=6,453) and small variants (n=9,206,351) located within a 1Mbp distance (**Figure 5A**). 14% of SVs were not in LD with any SG10K_Health small variants (R^2<0.2), suggesting that a substantial proportion of SVs represent genetic variability that might be overlooked in conventional genetic association analyses. 4,047 of the 6,453 high-confidence common SVs were in strong LD with 172,698 SG10K_Health SNPs (R^2>0.8). Of these, 748 SVs were in strong LD with 1,814 SG10K_Health SNPs matching lead SNPs from the EBI GWAS catalogue (genomic location and allelic alteration), with 75 SVs (35 deletions, 4 duplications, and 36 insertions) in strong LD with 174 lead SNPs from GWAS focused on Asian cohorts. **Supplementary table 5** provides these 75 SG10K-SVs and their associated lead SNPs.

From the 75 SVs, we focused on the subset that overlapped exons, since they could most directly be assigned a functional consequence. These included a predicted LOF deletion (chr1: 89,010,225-89,012,941) of exons 7 and 8 of *GBP3* that was in strong LD (R^2=0.968) with a missense variant (C->R) in the same gene (rs17433780; **Figure 5B**). Notably, rs17433780 is associated with markers of subclinical atherosclerosis in Chinese individuals ($P=2x10^{-6}$ [60]). While this missense SNP is certainly a candidate, our analysis suggests that the linked LOF SV should also be considered a potential causal variant for subclinical atherosclerosis in this locus.

GWAS-lead SNPs are often found in non-coding regions of the genome. Our analysis highlighted one exonic-associated SVs in high LD with these non-exonic SNPs, where the former may represent underlying causal variants. For example, a predicted LOF SV (chr11:55,264,123-55,271,064) deleting exons 2 to 6 of *TRIM48* exhibited strong LD (R^2=0.903) with an intergenic GWAS-lead SNP (chr11:54,697,371; rs11532186) associated with altered glomerular filtration rate. Notably, an integrative analysis of genetic association and gene expression in a cohort of patients with reduced kidney function identified *TRIM48* among the top causal candidates for urine metabolite variation[61]. These examples support the value of including SG10K-SVs in analyses of genetic drivers of phenotypic variation in Asian cohorts. The full list of SVs in high LD with GWAS lead SNPs is reported in **Supplementary Table 5**.

Discussion

We generated a comprehensive catalogue of SVs in 8,392 Singaporeans containing 152,655 SVs. Compared to previous studies analysing primarily populations of European ancestry, our samples enabled us to assess patterns of SV genetic diversity across Asia, leveraging on Singapore as a diverse multi-ethnic community. In particular, little is known about the SV landscape in Malay individuals. Malay is the third largest ethnic group in Asia. Individuals of Malay ethnicity are geographically distributed across several countries in Southeast Asia, including Malaysia, Singapore, Java and Sri Lanka[62]. There are approximately 220 million individuals of Malay ethnicity in the world, with Indonesia and Malaysia accounting for the majority. Previously, Wu *et al.* investigated the population structure of the three Singaporean populations with the 1000G project populations using small variants and reported an ancestral component that is largely specific to the Malays in Singapore[7]. This result indicates the importance of including individuals of Malay ethnicity in large-scale population-based SV studies so as to uncover SVs unique to the Malay community. Our study is the first population-scale SV study to include individuals of Malay ethnicity. Notably, we identified several

SVs enriched in individuals of Malay ethnicity. For example, we observed a 1.6kbp deletion in *FAM3B*, a gene involved in glucose metabolism and linked to type 2 diabetes, with a higher allele frequency in Malays compared to Chinese and Indians. Previous studies have characterized SVs in the Chinese (East Asian) and Indian (South Asian) populations[10,12]. However, the current study provided a much more comprehensive catalogue of SVs for these populations by analysing a much large number of samples than previous efforts. Overall, our findings reiterate the importance of creating a comprehensive population-specific database of SVs to fill the gap of our understanding of genetic diversity in Asian populations.

While clearly a first-generation catalogue, the SG10K-SV database allows us to identify Asian-specific variants. Given the large number of Asian samples in our dataset, we demonstrate the ability to pinpoint novel variants that occur in higher frequency in the Asian population but were missed in other large-scale population SV studies such as gnomAD-SV. In addition, the use of SG10K-SV also enables us to identify variants that are highly prevalent in Asians within the existing public database. By calculating the Fst between Asians and non-Asians for variants in gnomAD-SV alone, we detected 1,647 SVs, which shows significant differences in their allele frequencies between Asians and non-Asians. However, by incorporating data from SG10K-SV and identifying variants that were shared by the two datasets, we detected 8,285 SVs in gnomAD-SV that showed significant differences in their allele frequencies between Asians and non-Asians. As such, we demonstrate that SG10K-SV can be used to complement existing SV catalogues to identify Asian-specific variants.

In addition, due to the genetic diversity within the Singapore population, the SG10K-SV dataset enabled the detection of variants that are unique to each of the three Asian populations. We identified 18,076 SVs displaying ancestry-specific frequency patterns. We were able to detect many SVs that were reported previously in Asian population. For instance, we observed a 21.5kbp duplication overlapping *MPV17L2* and *RAB3A* in gnomAD-SV East Asians. This duplication has a higher allele frequency in the Chinese within SG10K-SV than in Malays and Indians. More importantly, we discovered previously unreported SVs that were more prevalent within one of the ethnic groups. We found a 3kbp deletion overlapping *AHNAK2*, encoding a nucleoprotein involved in calcium signalling, that occurs at a higher frequency in Indians than Malays and Chinese. We also identified a 1.3kb deletion overlapping the *FAM3B* gene, which encodes a secreted cytokine-like protein involved in glucose metabolism and linked to type 2 diabetes. This deletion has a higher occurrence in Malay individuals compared to Chinese and Indians. We also identified a 12.2kbp duplication overlapping *CLPSL1* and *CLPS*, which encode enzymes involved in the digestion of dietary proteins. This duplication has a high allele frequency in the Indians and Malays compared to Chinese. Interestingly, the observed minor AF was greater than 50% in all three populations, indicating that this duplication common within the Asian population. Our findings reiterate the importance of population-specific reference data to reduce biases in genetic discovery.

Apart from identifying ancestry-specific variants, the SG10K-SV catalogue has also enabled us to identify potential SVs associated with human phenotypes. Beyond SVs affecting gene function, integrating SG10K-SVs with SNPs enabled us to identify LD patterns between these two categories of genomic variation (SVs and SNPs). Specifically, we found 75 SVs in strong LD with GWAS lead SNPs from Asian cohorts. Of these 75 SVs, we identified a *GBP3* deletion in strong LD with a GWAS lead SNP that is associated with subclinical atherosclerosis, demonstrating the value of the SG10K-SV database, which allows the identification of potential causative SVs that are in strong LD with GWAS lead SNPs associated with disease phenotypes in the Asian cohort.

Our study has several limitations. SV discovery is challenging, and the full spectrum of SVs in the human genome remains poorly understood. The findings presented here are primarily derived from 15x short-read WGS and are clearly underpowered both in terms of sequencing read length and sequence coverage to capture all possible SVs present in the Asian population. Existing algorithms rely on sequencing coverage and split-reads from the short-read WGS data to detect SVs, and hence, it is impossible to identify the exact coordinates and length of tandem duplicates and large insertions using short-read data. In addition, at present, our SV callers captured only the three most commonly analysed SVs (deletions, insertions and duplications), but did not consider other SV classes (inversions, translocation) that are also present in human genomes and are likely to have biological consequences. Using long-read sequencing in the near future, either as a single modality or coupled with high-coverage short-read sequencing, will allow us to identify substantially more SVs, clarify SVs in repetitive regions, and define new classes of SVs. Notwithstanding these shortcomings, the SG10K-SV dataset is, by far, the largest Asian SV database. This resource will be valuable to understanding the genetic diversity of the Asian population and how these variations underpin health and disease in the Asian population.

Materials and Methods

WGS data quality control

We processed WGS data collected from the SG10K_Health[23] study. SG10K Health comprises alignments and variant calls for SNVs and INDELs from 9 local cohorts, including 9,770 healthy individuals. Data generation involved WGS of blood DNA samples (Illumina short-reads) and subsequent analysis following GATK best practices (GATK4 GRCh38)[63] to generate individual sample level CRAM files. It also included QC checks intended to discard samples with poor sequencing quality (e.g. hard filters for error rate and contamination), unusual numbers of calls (e.g. MAD-based filters on het/hom ratio), chromosome aneuploidies, and/or samples with related individuals in the same cohort (see methods in Wong *et al.*, 2023 for additional details).

Using an in-house developed pipeline, we calculated the coverage, alignment and GC bias metrics from the SG10K Health CRAMs. In total, nine metrics were considered for downstream filtering, chosen to represent the type of evidence used by SV calling algorithms:

- Median_autosome_coverage: The median coverage in autosomes, excluding (i) bases in reads with low mapping quality (mapq < 20); (ii) bases in reads marked as duplicates, and (iii) overlapping bases in read pairs; calculated with mosdepth[64].
- Mad_autosome_coverage: The median absolute deviation of coverage in autosomes after coverage filters are applied (see median_autosome_coverage); calculated with mosdepth[64].
- Pct_autosomes_1x: The percentage of bases that attained at least 1X sequence coverage in autosomes, after coverage filters are applied (see median_autosome_coverage); calculated with mosdepth[64].
- Pct_reads_aligned: The percentage of PF reads that align to the reference; calculated with picard AlignmentSummaryMetrics[65].
- Pct_reads_properly_paired: The percentage of reads that align as proper pairs; calculated with samtools stats[66].
- Median_insert_size: The median insert size of aligned reads; calculated with picard InsertSizeMetrics[65].
- Mad_insert_size: The median absolute deviation of insert sizes; calculated with picard InsertSizeMetrics[65].
- gc_dropout: Illumina-style GC dropout metric; calculated with picard GcBiasSummaryMetrics[65].
- at_dropout: Illumina-style AT dropout metric; calculated with picard GcBiasSummaryMetrics[65].

In each cohort, we discarded samples outside 8 MAD from the median for at least one of the nine metrics considered. Such filters led to the exclusion of 1,378 samples, thus leaving 8,392 samples for downstream analysis.

Deletions and insertions detection

Manta v1.6 was executed in the single sample mode to identify deletions and insertions in the discovery dataset. We used the default parameters and further filtered the single-sample VCF to retain (i) calls that pass filters, (ii) with a length of 50bp or more and (iii) of the selected variant types (deletions and insertions).

SV discovery from short read data is notedly a challenging task[67]. Moreover, since the majority of our dataset consists of 15X genomes (**Supplementary Table 1**), we expect lower sensitivity compared to what has been reported in higher-depth studies[10]. In order to overcome these limitations, we have incorporated additional clustering and re-genotyping steps, which are known to improve detection power in short-read-based studies. In brief, the goal is to aggregate all SV candidates identified when evaluating each sample individually (SV clustering), and then re-assess the original data for the presence/absence of these calls (SV re-genotyping).

Prior to clustering, we sought to discard any samples that displayed an unusual number of calls for any of the SV types considered, by applying an 8-MAD filter on a per-cohort basis, analogous to the strategy previously used during sample QC. For Manta, no samples were discarded after applying such a filter, suggesting that the upstream sample QC is already adequate to flag unusual samples. We then clustered SV candidates in each of the call sets obtained during the discovery step using svimmer[27], which we ran with default parameters to aggregate events across all samples in the discovery dataset. Lastly, we performed re-genotyping for each sample using Graphtyper[28] v2.5.1 with default parameters. We then retained calls made under the aggregated genotyping model for downstream analysis.

We further enriched the annotations from Graphtyper2 with depth-based information by running Duphold v0.2.3[68] on each individual sample. Subsequently, we filtered calls using the following criteria: (i) For homozygous reference genotypes, we retain calls with FORMAT/FT=="PASS".(ii) For heterozygous and homozygous alternate genotypes, we retain calls as follows: FILTER=="PASS" &

FORMAT/FT=="PASS" & INFO/PASS_AC>0 & INFO/QD>12 & (INFO/ABHet>0.30 | INFO/ABHet<0) & (INFO/AC / INFO/NUM_MERGED_SVS)<25 & INFO/MaxAAS>4 & ((INFO/SB>0.1 & INFO/SB<0.9) | INFO/SB<0). (iii) For deletions, we retain calls with INFO/DHFFC < 0.7. Any genotype calls that did not pass the filters mentioned above was set to null using Hail[69].

Mobile Element Insertions (MEIs) detection

MELT v2.2.2[30] was executed using MELT-Split with default parameters in a four-step process to identify different classes of mobile element insertions (Alu, SVA, LINE1) in the SG10K-SV discovery set. First, indivAnalysis was used to identify MEI in each sample. Second, GroupAnalysis was used to aggregate MEIs across all samples in the discovery dataset. Third, we performed re-genotyping for each sample using the merged MEI information obtained from step 2 using Genotype feature in MELT. Lastly, MELT-Split uses the MakeVCF function to filter and merge MEIs information across all samples into a single VCF. The four-step MEI discovery was run separately for each MEI class. We extract only variants that PASS the filters indicated by MELT for downstream analysis.

For the two validation datasets, we used the output file from GroupAnalysis, which contains aggregated MEIs across all samples in the discovery dataset, to re-genotype MEIs in each sample in the two validation datasets. Lastly, we used the MakeVCF function to filter and merge MEIs across all samples into a single VCF. We extract only variants that PASS the filters indicated by MELT for downstream analysis.

Duplications detection

We ran SurVIndel2[33] with default parameters on each sample in the discovery set and only retained tandem duplications. Duplications were left-aligned using the normalised utility in SurVIndel2. Then, we clustered the duplications as recommended in the manuscript of SurVIndel2, in order to obtain a set of duplications in the studied population.

Next, we used the companion re-genotyper of SurVIndel2, SurVTyper, to genotype each duplication in each sample. The genotyped duplications for each sample were merged using bcftools merge. Finally, we set calls such that FORMAT/FT != PASS as not genotyped.

Callset refinement and merging of individual variant callset into SG10K-SV Release 1.3 discovery and validation datasets

For the last step of the SV pipeline, we used a combination of regional, call and event-specific filters to further refine the outputs of the re-genotyping step, aiming to reduce the number of false positives in our dataset. Region-specific filters were applied consistently across all samples before generating the final SG10K-SV release 1.3 to (i) retain events in autosomal contigs (chr1-22) and (ii) exclude those that occur in centromeres, telomeres, heterochromatin region[70] and regions in the primary assembly that overlap with ALT contigs.

Benchmarking of tools

Benchmarking structural variations (SVs) generated by short-read methods is often done using long-read based ground truth catalogues. The Human Genome SV Consortium (HGSVC) released HGSVC2, a comprehensive set of SVs detected in 35 samples in the 1000 Genome Project using PacBio HiFi reads[34]. Additionally, CRAM files at 30x coverage are available for all the samples[71]. We used 10 samples for our benchmarking effort. We downsampled these 10 samples to a sequencing depth of 15x using samtools to mimic our discovery set. Next, we ran our pipeline on a dataset comprising 5,487 discovery samples plus the 10 benchmarking samples. Finally, we obtained a call set for each sample by retaining SVs with an allele count of at least 1 and an FS value of PASS. We used an in-house tool (https://github.com/Mesh89/SVComparator) to compare, for each sample, the predicted SVs with the set of SVs reported in HGSVC2. Our pipeline reports tandem duplications and insertions separately, while HGSVC2 only reports deletions and insertions; tandem duplications are considered insertions. For this reason, we could not measure the sensitivity of our duplications and insertions separately.

One of the significant challenges when generating a dataset of SVs for a large population is maintaining a low level of noise. Our benchmarking efforts show that our call set is precise (average precision is 0.73 for deletions, 0.88 for duplications and 0.79 for insertions) (**Supplementary Table 6**). Unsurprisingly, PacBio HiFi reads can discover far more SVs compared to 15x Illumina paired-end reads. However, the number of deletions, duplications and insertions we discover is comparable to recent studies such as

gnomAD[8] while we use lower sequencing depth. Coupled with the good precision, we conclude that our pipeline is in line with the state of the art in the field.

Principal component analysis (PCA)

To investigate the relationship between the different ethnic groups in Singapore, we performed principal component analysis (PCA) using all variants (deletions, insertions, duplications and MEIs) genotypes using Hail[69]. Briefly, monomorphic sites are removed prior to Hard-Weinberg Equilibrium (HWE) normalization. Genotypes were normalized using HWE[72]. Lastly, we used the normalized genotypes matrix for principal component estimation. We performed PCA on all samples in the discovery dataset. The results indicate that PC1 and PC2 can segregate the individuals by their ethnic groups. We also performed PCA on all samples in the discovery dataset for each variant type separately. The results obtained per variant type recapitulated the population structure when all variants were analysed together.

Comparison of the number of Asian samples across different population-based SV studies

We obtain the ancestry composition of 3 major studies with SV, namely 1) gnomAD-SV[10] 2) 1000 Genomes Project (1KG)[12] 3) Centers for Common Disease Genomics (CCDG)[8]. Samples in gnomAD-SV were grouped into “EAS” (gnomAD-SV East Asian (EAS) sample) and “Other” (all other non-EAS sample), while 1KG was grouped into “EAS” (1KG’s sample found in superpopulation of East Asian Ancestry (EAS)), “SAS” (1KG’s sample found in superpopulation of South Asian Ancestry (SAS)) and “Other” (1KG’s superpopulation which are not EAS and SAS) and CCDG was grouped into “EAS” (CCDG’s sample of EAS ancestry), “SAS” (CCDG’s sample of SAS ancestry) and “Other” (CCDG’s sample of non-EAS or non-SAS ancestry). SG10K-SV's sample were grouped into SG-CHI (individuals of self-reported “Chinese” ancestry), SG-MAL (individuals of self-reported “Malay” ancestry) and SG-IND (individuals of self-reported “Indian” ancestry). Sample count of each group was plotted in a stacked barplot for each project.

Comparison to SVs from gnomAD-SV

We obtained the hg38 lift-over gnomAD-SV callset from NCBI’s dbvar study “nstd166”.
We considered any SG10K-SV to be novel if no overlapping gnomAD-SV could be identified using a approach similar to our svimmer-based clustering of individual sample derived SV candidates, aggregating events across gnomAD-SV and SG10K-SV with svimmer[27] default parameters.

Enrichment analysis

To calculate the relative enrichment for genic and non-coding regions of the genome, we downloaded the ENCODE cCRE track[73] and gencode v40[74] annotation from UCSC table browser.

First, we partitioned the SG10K-SV dataset into three groups (ultra-rare, rare and common) based on the allele frequency of the variants using bcftools (version 1.16) filter function. Ultra-rare variants are variants with AF < 0.001; rare variants are variants with AF >= 0.001 and AF < 0.01 and lastly, common variants are variants with AF >= 0.01. The partitioned VCF files were transformed into bed files with bcftools query and a custom script. To calculate the relative enrichment of SVs in non-coding cCRE regions, we retain only variants that do not overlap any exons using bedtools (v2.30.0) intersect. Next, we count the number of variants which overlaps cCRE regions and genic regions using bedtools intersect. Lastly, we performed permutation tests for the different cCRE regulatory elements or genic regions that overlap SVs. For the permutation tests, the null distribution is calculated by the number of overlaps between cCRE regulatory elements or genic regions and randomly shuffled SV locations. We generated 10,000 random SV sets. For this analysis, we required the coordinates of the shuffled SVs to be within the same chromosome and non-overlapping. The enrichment of a specific cCRE regulatory elements or gene region and SV overlap is expressed as the log2 fold change of the number of actual SVs that overlap the specific regulatory or gene regions divided by the average of the null distribution. A positive log2 fold change indicates an enrichment of SVs in the specific regulatory or gene region compared to a random null distribution, whereas a negative log2 fold change indicates a depletion of SVs in the specific regulatory or gene region when compared against the null distribution. Lastly, the p-value was calculated as follows:

$$p - value = \frac{[Number\ of\ times\ abs(log2\ simulated\ fold\ change) >= abs(log2\ fold\ change\ actual)]}{10{,}000}$$

SV annotations

We annotated the SV VCF using SVTK[45] v0.27.1-beta with default parameters to associated SVs with gencode release 40 genes and transcripts. We focused on SVs that were annotated as loss of function (LOF), copy gain, duplications LOF (DUP_LOF). A deletion is predicted as LOF when it overlap at least one exon of a gene. A duplication is predicted as LOF when both the start and end of the duplication are contained within the exon of a gene. On the other hand, a duplication is annotated as DUP_LOF if a duplication overlaps at least one exon of a gene. A duplication is annotated as copy gain if it spans the entire gene. Lastly, an insertion is predicted as a LOF if a sequence is inserted into an exon.

To identify SVs affecting medically relevant genes, we annotated the SG10K-SV VCF using AnnotSV v3.2.3[47] with default parameters to identify SVs overlapping with the genes listed in ACMG version 3.1[46].

Identifying hotspots in SG10K-SV Release 1.3

To identify SV hotspot in the SG10K-SV dataset and gnomAD-SV dataset, we employed hotspotter from the primatR package[39] with the following parameters: (bandwidth:200,000, num.trial=10,000, pval=5 $x10^{-3}$). To identify hotspots unique to our dataset, we used bedtools[75] intersect with the "-v" function to find hotspot regions that are absent in gnomAD.

Linkage disequilibrium (LD) analysis between SNPs and SVs

To explore the relationship between SVs and SNPs, we conducted pairwise linkage disequilibrium (LD) analysis between each SV and small variants identified in SG10K_Health[23] . We used PLINK[76] to calculate the R^2 value between each SV and all SNPs co-localized within a 1Mbp window.

Known GWAS lead SNPs were retrieved from the NHGRI-EBI GWAS catalogue v1.0.2, only studies involving Asian individuals containing cohorts were retained. Finally, we found SNPs in common between the filtered NHGRI-EBI GWAS catalogue and SG10K-SNP that were in high LD ($R^2 \geq 0.8$) with an SV in SG10K-SV.

Fixation index (Fst) calculation

We computed Fst values using the "hudson_fst" function from the "scikit allel" Python package. The calculation involved comparing allele frequencies (AF) between pairs of populations. Specifically, we compared overall AFs in SG10K-SV-r1.3 to East Asian-specific AFs within gnomAD when comparing events in SG10K-SV vs. gnomAD-SV. For SG10K-SV, we performed 3 pairwise comparisons , 1) Chinese vs. Indian, 2) Chinese vs. Malay, and 3) Indian vs. Malay populations. The resulting Fst values were obtained for each pair and the maximum Fst value was kept for each SG10K-SV event along with the annotation of which pair-wise comparison generated the Fst value

Next, to assign p-values to each Fst value, we conducted permutation analysis. This involved preserving the original genotype matrix while randomly shuffling the ancestry labels.

For gnomAD-SV Fst calculation, we compared EAS versus the non-EAS ancestry group using the VCF downloaded from

https://ftp.ncbi.nlm.nih.gov/pub/dbVar/data/Homo_sapiens/by_study/genotype/nstd166/gnomad_v2.1_sv.sites.accessioned.vcf.gz

which contains the necessary tags of the ancestry group's allele call type, for example the "EAS_N_HOMREF", "EAS_N_HET" and "EAS_N_HOMALT" tags representing East Asian's number of sample called homozygous reference (hom_ref), heterozygous (het) and homozygous alternate (hom_alt) allele respectively. With the count for each allele call type, we generated a "GenotypeArray" in "allel" package with each element in the "GenotypeArray" being genotype status, [0,0] for hom_ref, [0,1] for het and [1,1] for hom_alt, based on the count of EAS ancestry group allele call type and a similar "GenotypeArray" was produced for the non-EAS ancestry group. The EAS "GenotypeArray" and non-EAS "GenotypeArray" was used to calculate the allele count for the 2 group with "count_alleles" function and the generated allele count used to calculate Fst with "hudson_fst" function.

To generate the p-value for gnomAD-SV, we combine the EAS "GenotypeArray" and non-EAS "GenotypeArray" and noted the length, n, of EAS "GenotypeArray", we then shuffle the combined "GenotypeArray", then split the shuffled "GenotypeArray" into shuffled EAS

“GenotypeArray” with the first n genotype in the shuffled "GenotypeArray” and the rest being shuffled non-EAS “GenotypeArray”. We calculated the Fst between this 2 shuffled "GenotypeArray” and note down the Fst (Fst-shuffled).

For both SG10K-SV and gnomAD-SV, we repeated the shuffling process 1,000 times and determined the number of instances where the observed Fst exceeded that of the real dataset, thus obtaining the p-value. We applied false discovery rate (FDR) correction across the entire dataset to account for multiple comparisons.

Subsequently, we applied additional filtering on the obtained FDR values to identify SVs with significant Fst. Specifically, we focused on events with an FDR threshold of less than 1% and an Fst value greater than the mean of the entire dataset.

Declarations

Author contributions

T.H.J.J., L.Z.H., R.R., M.G.P., L.J.J., S.W.K.K., N.B., S.P. and P.T. contributed to the writing of the manuscript. T.H.J.J., L.Z.H., R.R., M.G.P., T.R.Y., N.B., R.T.J., S.X.L., T.Y.A. and L.W.K. contributed to the generation of figures and analysis of the data. T.H.J.J., L.Z.H., R.R. and M.G.P. contributed to the production and quality control of the SG10K-SV dataset. M.H., J.L.O., S.A., J.J., Y.S.C., T.H.L., L.L.G., Y.C.T., K.P.L., C.W.L.C., S.D., N.K., C.Y.C., J.C.C., E.S.T., and the SG10K_Health Consortium contribute to the production and collection of the data. All authors reviewed the manuscript.

Ethics declaration

All authors declare no competing interests. This project is approved by the NPM Data Access Committee with project ID: NPM00002.

Data Availability

Aggregated data can be downloaded via the National Precision Medicine programme’s SG10K_Health Chorus variant browser accessible through the SG10K_Health web portal (https://npm.a-star.edu.sg/help/NPM).

Data including WGS and intermediate files for all analyses and regeneration of all display items contain individual-level data including genotypes. The data is governed by the NPM Data Access Committee (DAC). The data generated in this study is made available to researchers registered through the SG10K_Health Data Access Portal. Users are required to submit a Data Access Request to the NPM DAC for approval. The forms and data access policy can be downloaded via the SG10K_Health web portal (https://npm.a-star.edu.sg/help/NPM). For more information, users can contact the National Precision Medicine Programme Coordinating Office, A*STAR (contact_npco@gis.a-star.edu.sg).

Code Availability

Acknowledgement

This study made use of data generated as part of the Singapore National Precision Medicine program funded by the Industry Alignment Fund (Pre-Positioning) (IAF-PP: H17/01/a0/007).

This study made use of data / samples collected in the following cohorts in Singapore:

The Health for Life in Singapore (HELIOS) study at the Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore (supported by grants from a Strategic Initiative at Lee Kong Chian School of Medicine, the Singapore Ministry of Health (MOH) under its Singapore Translational Research Investigator Award (NMRC/STaR/0028/2017) and the IAF-PP: H18/01/a0/016);

The Growing up in Singapore Towards Healthy Outcomes (GUSTO) study, which is jointly hosted by the National University Hospital (NUH), KK Women’s and Children’s Hospital (KKH), the National University of Singapore (NUS) and the Singapore Institute for Clinical Sciences (SICS), Agency for Science Technology and Research (A*STAR)(supported by the Singapore National Research Foundation under its Translational and Clinical Research (TCR) Flagship Programme and administered by the Singapore Ministry of Health’s National Medical Research Council (NMRC), Singapore - NMRC/TCR/004-NUS/2008; NMRC/TCR/012-NUHS/2014. Additional funding is provided by SICS and IAF-PP H17/01/a0/005);

The Singapore Epidemiology of Eye Diseases (SEED) cohort at Singapore Eye Research Institute (SERI)(supported by NMRC/CIRG/1417/2015; NMRC/CIRG/1488/2018; NMRC/OFLCG/004/2018);

The Multi-Ethnic Cohort (MEC) cohort (supported by NMRC grant 0838/2004; BMRC grant 03/1/27/18/216; 05/1/21/19/425; 11/1/21/19/678, Ministry of Health, Singapore, National University of Singapore and National University Health System, Singapore);

The SingHealth Duke-NUS Institute of Precision Medicine (PRISM) cohort (supported by NMRC/CG/M006/2017_NHCS; NMRC/STaR/0011/2012, NMRC/STaR/ 0026/2015, Lee Foundation and Tanoto Foundation);

The TTSH Personalised Medicine Normal Controls (TTSH) cohort funded (supported by NMRC/CG12AUG17 and CGAug16M012).

This research is also supported by the National Research Foundation (NRF) Singapore under its National Precision Medicine Program (NPM) Phase II Funding (MOH- 000588) and administered by the Singapore MOH's National Medical Research Council (NMRC).

The views expressed are those of the author(s) are not necessarily those of the National Precision Medicine investigators, or institutional partners. We thank all investigators, staff members and study participants who made the National Precision Medicine Project possible.

References

1. Eichler, E.E. Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. *N Engl J Med* **381**, 64-74 (2019).
2. Frazer, K.A., Murray, S.S., Schork, N.J. & Topol, E.J. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics* **10**, 241-251 (2009).
3. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
4. Karczewski, K.J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020).
5. Smedley, D. *et al.* 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. *N Engl J Med* **385**, 1868-1880 (2021).
6. The “All of Us” Research Program. *New England Journal of Medicine* **381**, 668-676 (2019).
7. Wu, D. *et al.* Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore. *Cell* **179**, 736-749.e15 (2019).
8. Abel, H.J. *et al.* Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583**, 83-89 (2020).
9. Almarri, M.A. *et al.* Population Structure, Stratification, and Introgression of Human Structural Variation. *Cell* **182**, 189-199.e15 (2020).
10. Collins, R.L. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444-451 (2020).
11. Carvalho, C.M.B. & Lupski, J.R. Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews Genetics* **17**, 224-238 (2016).
12. Sudmant, P.H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81 (2015).
13. Lupski, J.R. Charcot-Marie-Tooth Polyneuropathy: Duplication, Gene Dosage, and Genetic Heterogeneity. *Pediatric Research* **45**, 159-165 (1999).
14. Pérez Jurado, L.A., Peoples, R., Kaplan, P., Hamel, B.C. & Francke, U. Molecular definition of the chromosome 7 deletion in Williams syndrome and parent-of-origin effects on growth. *Am J Hum Genet* **59**, 781-92 (1996).
15. Marshall, C.R. *et al.* Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet* **82**, 477-88 (2008).
16. Walsh, T. *et al.* Rare Structural Variants Disrupt Multiple Genes in Neurodevelopmental Pathways in Schizophrenia. *Science* **320**, 539-543 (2008).
17. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112-121 (2020).
18. Perry, G.H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nat Genet* **39**, 1256-60 (2007).

19. Divakar, M.K. *et al.* Whole-genome sequencing of 1029 Indian individuals reveals unique and rare structural variants. *J Hum Genet* **68**, 409-417 (2023).
20. Wu, Z. *et al.* Structural variants in the Chinese population and their impact on phenotypes, diseases and population adaptation. *Nature Communications* **12**, 6501 (2021).
21. https://www.singstat.gov.sg/find-data/search-by-theme/population/population-and-population-structure/visualising-data/population-dashboard, 2023).
22. Teo, Y.Y. *et al.* Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. *Genome Res* **19**, 2154-62 (2009).
23. Wong, E. *et al.* The Singapore National Precision Medicine Strategy. *Nature Genetics* **55**, 178-186 (2023).
24. Chan, S.H. *et al.* Analysis of clinically relevant variants from ancestrally diverse Asian genomes. *Nature Communications* **13**, 6694 (2022).
25. Kosugi, S. *et al.* Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology* **20**, 117 (2019).
26. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220-2 (2016).
27. Github:DecodeGenetics/svimmer. Structural Variant Merging Tool. (https://github.com/DecodeGenetics/svimmer, 2021).
28. Eggertsson, H.P. *et al.* GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature Communications* **10**, 5402 (2019).
29. Delage, W.J., Thevenon, J. & Lemaitre, C. Towards a better understanding of the low recall of insertion variants with short-read based variant callers. *BMC Genomics* **21**, 762 (2020).
30. Gardner, E.J. *et al.* The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res* **27**, 1916-1929 (2017).
31. Rajaby, R. & Sung, W.K. SurVIndel: improving CNV calling from high-throughput sequencing data through statistical testing. *Bioinformatics* **37**, 1497-1505 (2021).
32. Dashnow, H. *et al.* STRetch: detecting and discovering pathogenic short tandem repeat expansions. *Genome Biology* **19**, 121 (2018).
33. Rajaby, R. & Sung, W.-K. SurVIndel2: improving local CNVs calling from next-generation sequencing using novel hidden information. *bioRxiv*, 2023.04.23.538018 (2023).
34. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**(2021).
35. Price, A.L., Eskin, E. & Pevzner, P.A. Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. *Genome Res* **14**, 2245-52 (2004).
36. Beck, C.R. *et al.* LINE-1 retrotransposition activity in human genomes. *Cell* **141**, 1159-70 (2010).
37. Lin, Y.L. & Gokcumen, O. Fine-Scale Characterization of Genomic Structural Variation in the Human Genome Reveals Adaptive and Biomedically Relevant Hotspots. *Genome Biol Evol* **11**, 1136-1151 (2019).
38. Perry, G.H. *et al.* Hotspots for copy number variation in chimpanzees and humans. *Proceedings of the National Academy of Sciences* **103**, 8006-8011 (2006).
39. Github:daewoooo/primatR. PrimatR. (https://github.com/daewoooo/primatR, 2022).
40. Long, M.D. *et al.* Reduced NCOR2 expression accelerates androgen deprivation therapy failure in prostate cancer. *Cell Rep* **37**, 110109 (2021).
41. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015).
42. Kiktev, D.A., Sheng, Z., Lobachev, K.S. & Petes, T.D. GC content elevates mutation and recombination rates in the yeast Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences* **115**, E7109-E7118 (2018).
43. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
44. Niu, Y. *et al.* Characterizing mobile element insertions in 5675 genomes. *Nucleic Acids Res* **50**, 2493-2508 (2022).
45. Github:talkowski-lab/svtk. SVTK. (https://github.com/talkowski-lab/svtk, 2021).

46. Miller, D.T. *et al.* ACMG SF v3.1 list for reporting of secondary findings in clinical exome and genome sequencing: A policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet Med* **24**, 1407-1414 (2022).
47. Geoffroy, V. *et al.* AnnotSV: an integrated tool for structural variations annotation. *Bioinformatics* **34**, 3572-3574 (2018).
48. Chopra, N. & Knollmann, B.C. Triadin regulates cardiac muscle couplon structure and microdomain Ca(2+) signalling: a path towards ventricular arrhythmias. *Cardiovasc Res* **98**, 187-91 (2013).
49. Schinner, C. *et al.* Stabilization of desmoglein-2 binding rescues arrhythmia in arrhythmogenic cardiomyopathy. *JCI Insight* **5**(2020).
50. Hudson, R.R., Slatkin, M. & Maddison, W.P. Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**, 583-9 (1992).
51. Ng, K.P. *et al.* A common BIM deletion polymorphism mediates intrinsic resistance and inferior responses to tyrosine kinase inhibitors in cancer. *Nat Med* **18**, 521-8 (2012).
52. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312-324 (2023).
53. Lin, Y.L., Pavlidis, P., Karakoc, E., Ajay, J. & Gokcumen, O. The evolution and functional impact of human deletion variants shared with archaic hominin genomes. *Mol Biol Evol* **32**, 1008-19 (2015).
54. PharmGKG. CYP2A6. (https://www.pharmgkb.org/vip/PA166169430).
55. Browser, g. (https://gnomad.broadinstitute.org/variant/DUP_19_47180?dataset=gnomad_sv_r2_1).
56. Sung, S.S. *et al.* Mutations in TNNT3 cause multiple congenital contractures: a second locus for distal arthrogryposis type 2B. *Am J Hum Genet* **73**, 212-4 (2003).
57. Wei, B. & Jin, J.P. TNNT1, TNNT2, and TNNT3: Isoform genes, regulation, and structure–function relationships. *Gene* **582**, 1-13 (2016).
58. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J.O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nature Reviews Genetics* **14**, 125-138 (2013).
59. Stranger, B.E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848-53 (2007).
60. Xie, G. *et al.* Genome-wide association study on progression of carotid artery intima media thickness over 10 years in a Chinese cohort. *Atherosclerosis* **243**, 30-7 (2015).
61. Feofanova, E.V. *et al.* A Genome-wide Association Study Discovers 46 Loci of the Human Metabolome in the Hispanic Community Health Study/Study of Latinos. *Am J Hum Genet* **107**, 849-863 (2020).
62. Deng, L. *et al.* Dissecting the genetic structure and admixture of four geographical Malay populations. *Scientific Reports* **5**, 14375 (2015).
63. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**, 491-498 (2011).
64. Pedersen, B.S. & Quinlan, A.R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867-868 (2018).
65. Broad Institute of, M.I.T. Picard Tools. (https://broadinstitute.github.io/picard/, 2018).
66. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**(2021).
67. Amarasinghe, S.L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biology* **21**, 30 (2020).
68. Pedersen, B.S. & Quinlan, A.R. Duphold: scalable, depth-based annotation and curation of high-confidence structural variant calls. *GigaScience* **8**(2019).
69. HailTeam. Hail 0.2. (2021).
70. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333-i339 (2012).
71. 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).
72. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-59 (2000).

73. Abascal, F. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699-710 (2020).
74. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Research* **49**, D916-D923 (2020).
75. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
76. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).

Figures

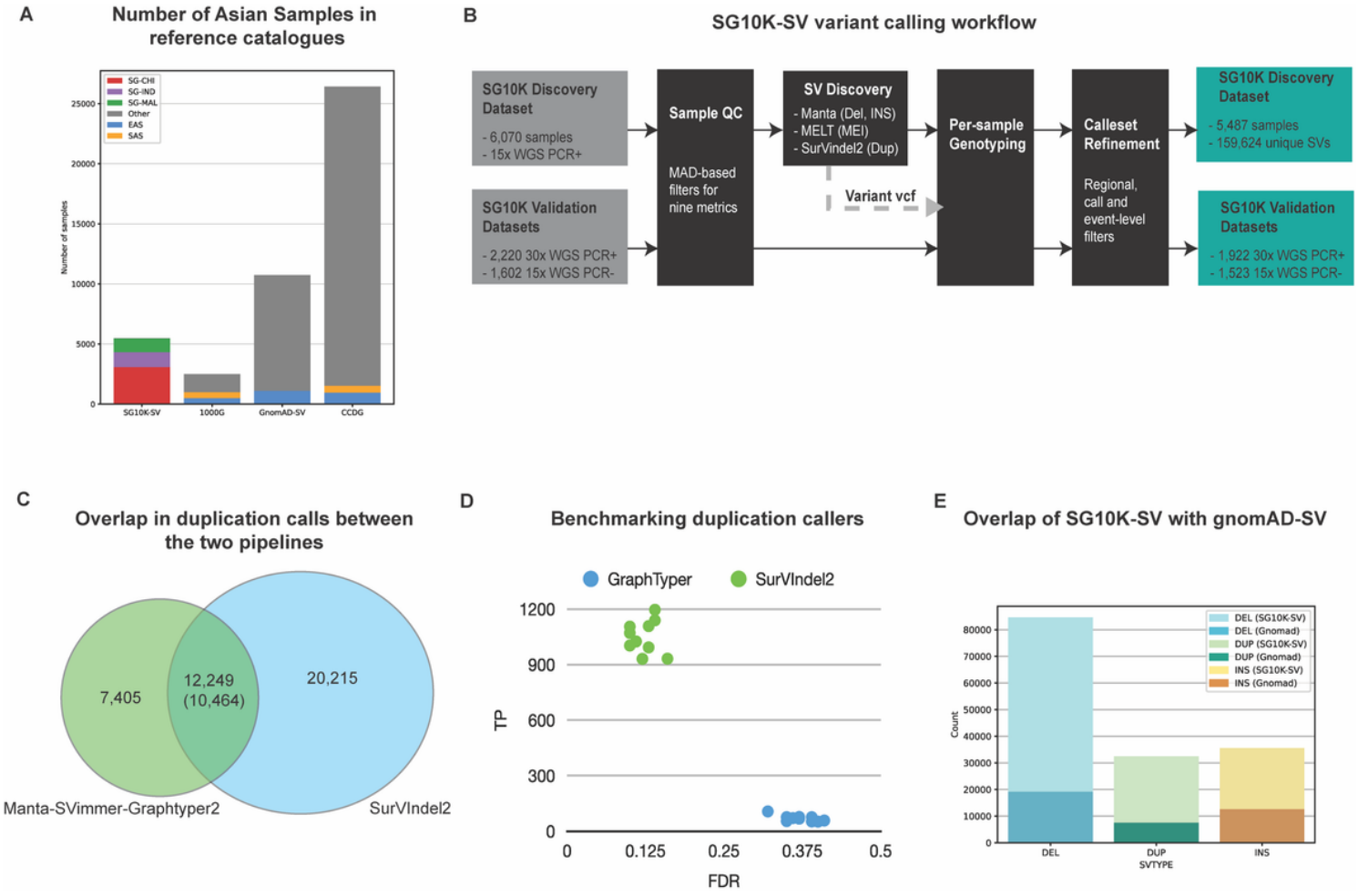

Figure 1

SG10K-SV-r1.3 Structural Variant landscape

A: Number of Asian samples in SG10K-SV-r1.3 compared to (short-read derived) 1000 genomes SV, gnomAD-SV and CCDG reference studies.

B: SG10K-SV-r1.3 analysis pipeline diagram

C: Comparison of the number of duplications detected by Manta-Graphtyper and SurVIndel2

D: Scatterplot comparing the number of true positives detected duplication and FDR achieved with Manta-SVimmer-Graphtyper2 and SurVindel2 for a truth set of high quality SVs obtained by haplotype-resolved long-read sequencing of a selected subset of 1000 Genomes Project analyzed samples[34].

E: Number of SG10K-SV-r1.3 variants that overlap with gnomAD-SV.

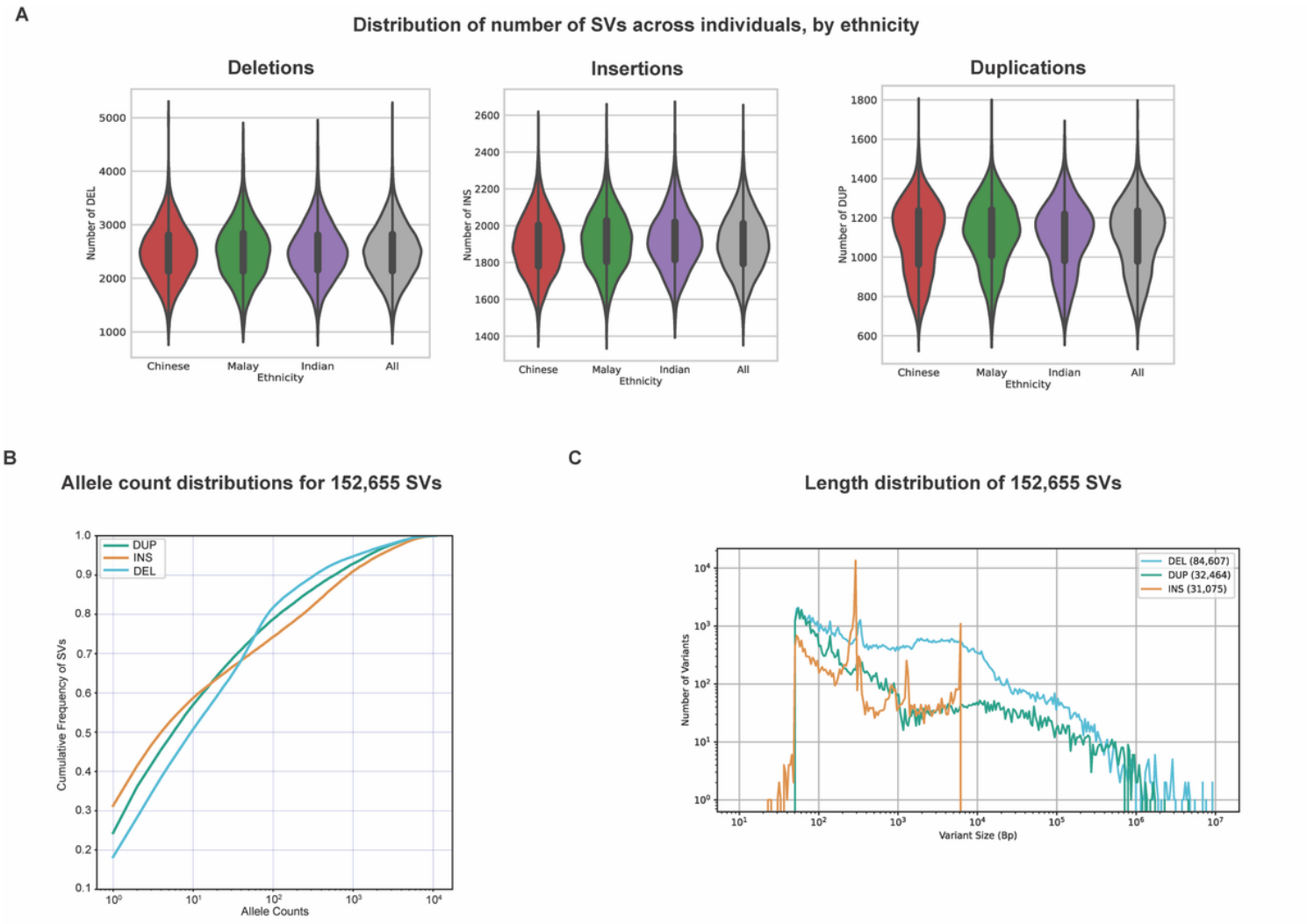

Figure 2

SG10K-SV-r1.3 Structural Variant catalogue properties

A: Violin plot showing the number of events per genome. DEL, deletions; DUP, duplications; INS, insertions (including MEIs).

B: Distribution of allele frequencies for different classes of SVs in the SG10K dataset. The majority of the SVs are rare variants (AF < 1%).

C: Size distribution of SVs detected from SG10K cohort. DEL, deletions; DUP, duplications; INS, insertions (including MEIs). Expected Alu, SVA and LINE1 MEIs peaks at around 300bp, 2100bp and 6000bp, respectively.

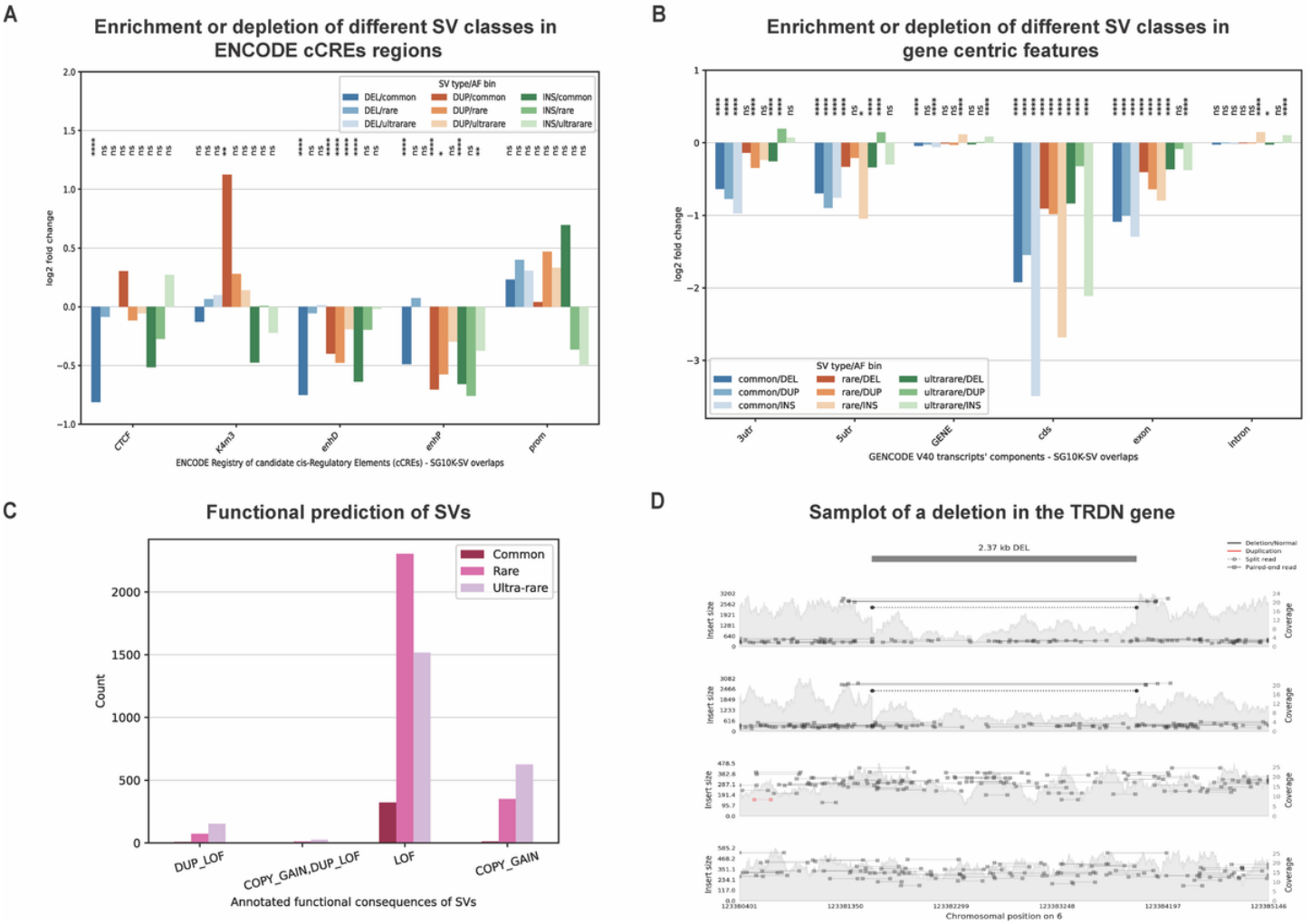

Figure 3

Functional impact of structural variations in the SG10K-SV-r1.3

A. Enrichment or depletion of different classes of non-exonic SV in regulatory elements across allele frequency bins. Common indicates variants with allele frequency ≥ 0.01; rare indicates variants with allele frequency ≥ 0.001 and allele frequency < 0.01; ultra-rare variants refers to variants with allele frequency < 0.001.

B. Distribution of SVs (Deletions, Insertions, Duplications) disrupting gene centric features across allele frequency bins. Ns indicates not significant p-value, * indicates p-value < 0.05, ** indicates p-value < 0.01, *** indicates p-value <0.001, **** indicates p-value < 0.0001.

C. *in silico* prediction of functional consequences of SVs segregated by allele frequencies.

D. Samplot of a 2,373 bp deletion event overlapping the TRDN gene region.

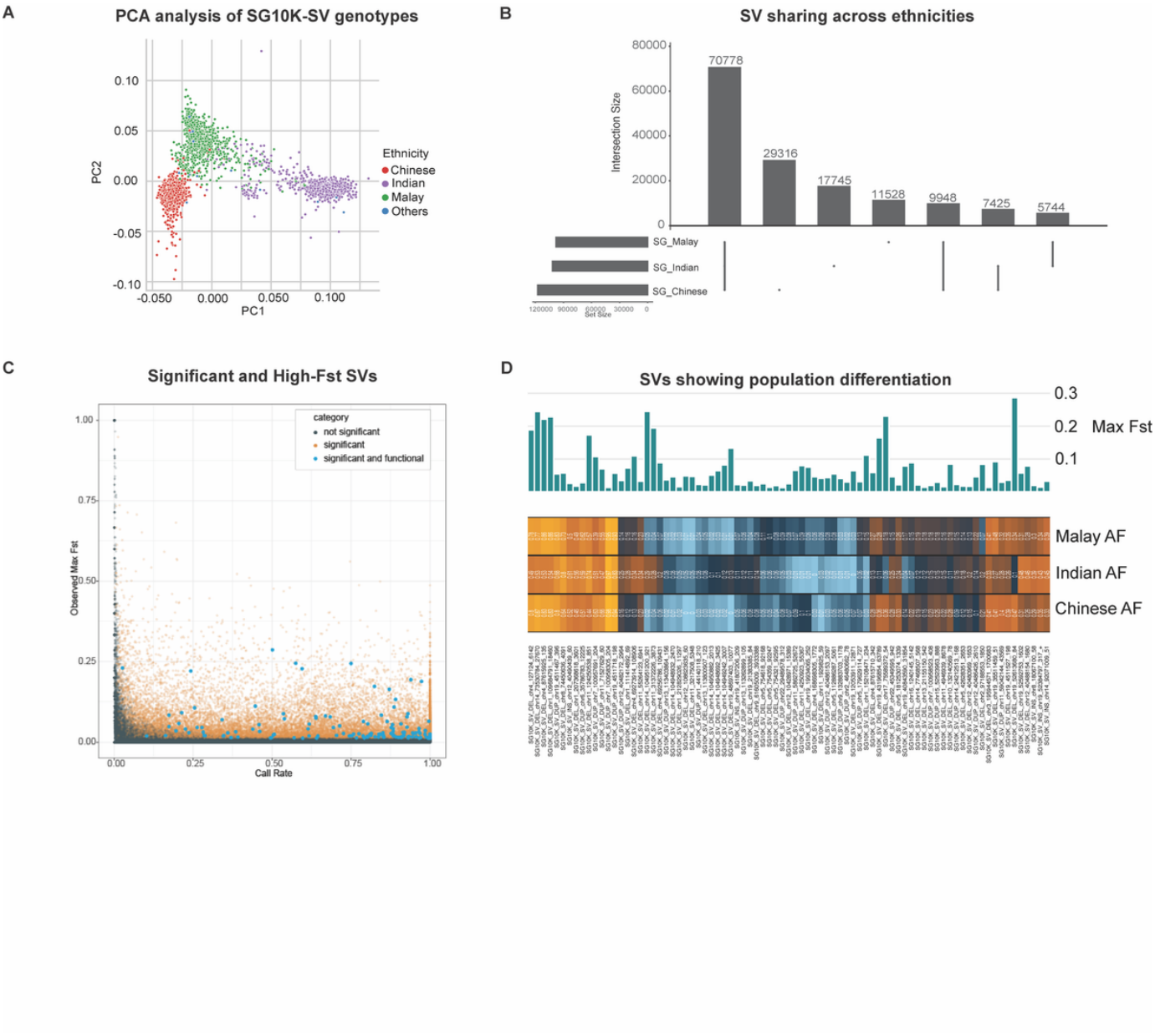

Figure 4

Population specificity of SVs

A. Population structure revealed by PCA analysis of SG10K-SV-r1.3 genotype values. Each point corresponds to an individual, coloured according to its ethnicity, x and y axis represents the first two principal component respectively.

B. Proportion of SVs found in all, two or one populations.

C. Scatter plot of SV's fixation index (Fst) as a function of their call rate.

D. Allele frequencies in Chinese, Indian and Malay for selected SV events with elevated fixation index (Fst).

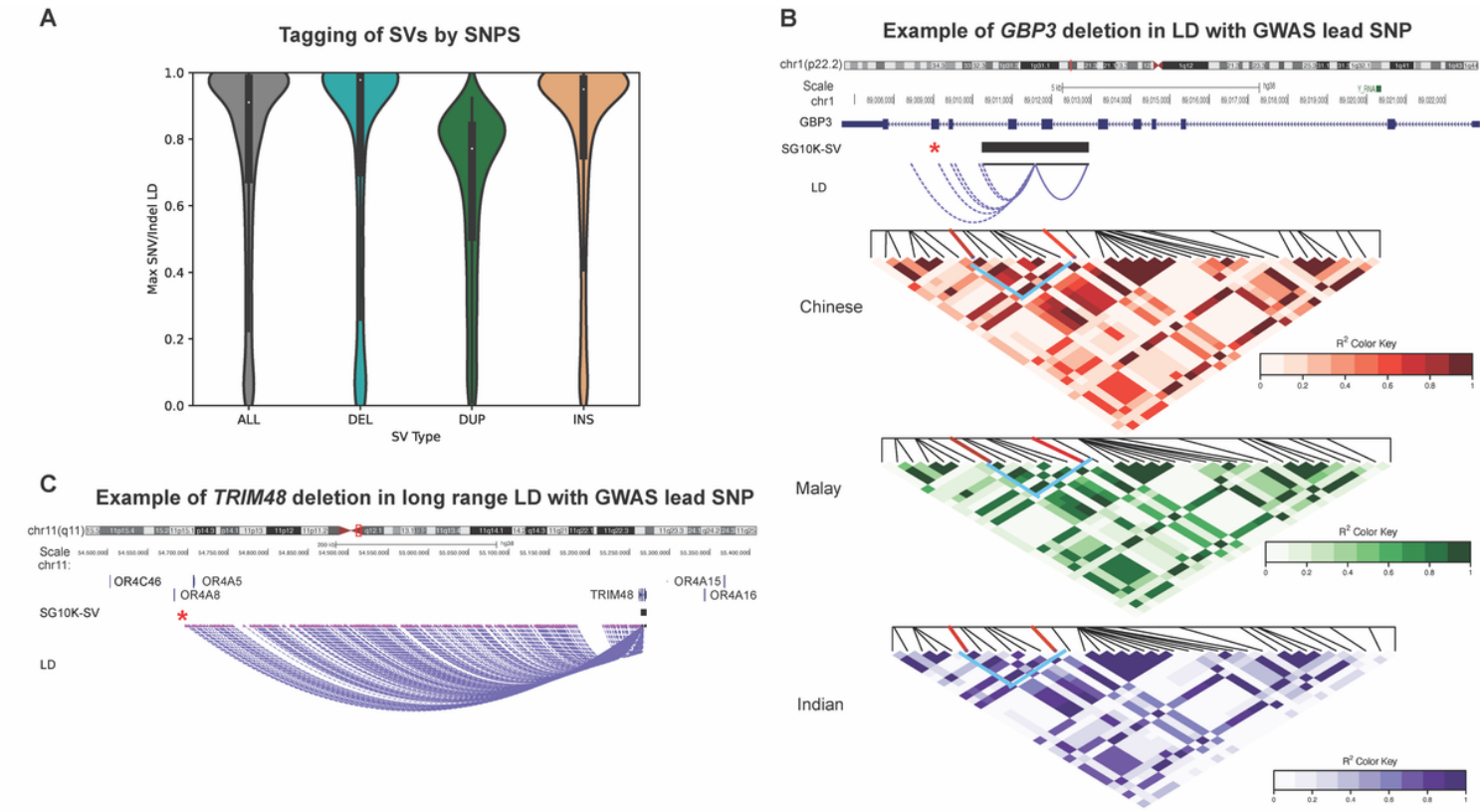

Figure 5

Linkage disequilibrium between SVs and SNPs

A. Tagging of SVs by SNPs: Distribution of the maximum R^2 value to SNPs for each SV.

B. Candidate causal SV: Example of a deletion affecting exons 7 and 8 of the *GBP3* gene, in high LD with a carotid artery intima media thickness GWAS SNP in exon 10 of *GBP3*. The SNP is significantly associated with carotid artery intima media thickness. LD structure plots are shown for the three ethnicities. The star indicates the GWAS lead SNP and the black bar indicates the SV.

C. Candidate causal SV: Example of a deletion in *TRIM48* gene, in high LD with an intergenic GWAS-lead SNP associated with altered glomerular filtration rate. The lines indicate LD between GWAS-lead SNP and deletion with r2 >= 0.8. The star indicates the GWAS lead SNP and the black bar indicates the SV.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TanetalSG10KSVSupplementaryTables.xlsx
- TanetalSG10KSVNatCommSupplementaryInformation.docx