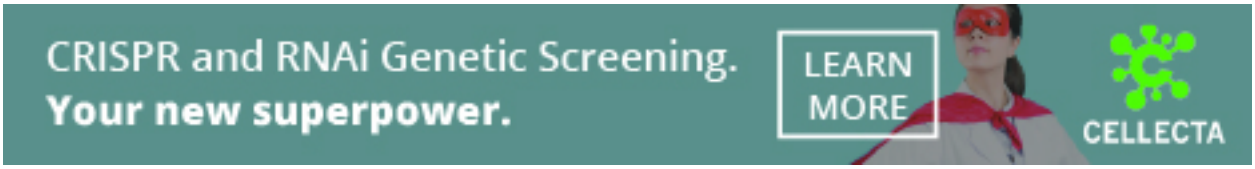
Nanopore sequencing identifies high-frequency somatic structural variations in laryngeal squamous cell carcinoma genomes

Xuyan Liu, Lin Xia, Yixin Qiao, et al.

Genome Res. published online April 8, 2026
Access the most recent version at doi:10.1101/gr.281046.125



Nanopore sequencing identifies high-frequency somatic structural variations in laryngeal squamous cell carcinoma genomes

Xuyan Liu[1#], Lin Xia[1#], Yixin Qiao[2#], Yang Li[1], Yan Huang[1], Bingyan Yue[2], Xi Liang[2], Xin Yang[2], Honghui Zhang[2], Jiaxun Zhang[1], Xiao Chen[1], Dan Xie[1]*, Jifeng Liu[2,3]*

1. Laboratory of Omics Technology and Bioinformatics, Frontiers Science Center for Disease-related Molecular Network, State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu, Sichuan, 610041, China
2. Department of Otolaryngology-Head & Neck Surgery, West China Hospital, Sichuan University, Chengdu, Sichuan, 610041, China
3. Deep Underground Space Medical Center, West China Hospital, Sichuan University, Chengdu, Chengdu, Sichuan, 610041, China

# These authors contributed equally

* Corresponding author

E-mail: danxie@scu.edu.cn (DX), liujifeng777@wchscu.cn (JFL)

Running title

Somatic variations in LSCC by long-read sequencing

Abstract

Laryngeal squamous cell carcinoma (LSCC) is an aggressive cancer with poor quality of life. Understanding the somatic mutations in its genome is crucial for elucidating its occurrence and progression. Although somatic structural variations (SVs) have been documented in LSCC, conventional short-read sequencing lacks the sensitivity to effectively detect high-frequency SVs shared across multiple samples—variants that play a crucial role in tumorigenesis. Here, we present SomaGauss-SV, a somatic SVs detection workflow leveraging nanopore long-read sequencing data. Benchmarking against five paired tumor cell line datasets shows SomaGauss-SV consistently achieves a balanced high precision and recall. SomaGauss-SV applied to 15 paired LSCC tumor-blood samples uncovers a comprehensive SVs landscape and reveals a significant positive correlation between somatic deletion burden and smoking intensity. Furthermore, a high-frequency somatic simple repeat expansion is identified in 28/39 (71.79%) of LSCC patients, which upregulates the expression of genes *TP53BP2* and *FBXO28* through spatial proximity. These findings underscore the potential of long-read sequencing and SomaGauss-SV for uncovering recurrent somatic SVs in LSCC, providing valuable resources for biomarker discovery.

Introduction

Laryngeal squamous cell carcinoma (LSCC) is the second most common subtype of head and neck squamous cell carcinoma (HNSCC), characterized by marked gender differences and regional variability in both incidence and mortality rates. The incidence rate in males is approximately eight times that of females(Han et al. 2024) LSCC pathogenesis is closely related to modifiable risk factors, such as tobacco exposure and alcohol consumption(Bray et al. 2024; Sung et al. 2021; Chen et al. 2024; Johnson et al. 2020). However, the molecular mechanisms by which these exogenous factors induce genomic instability remain unclear. Early symptoms of LSCC are often overlooked, resulting in more than 50% of patients being diagnosed at advanced stages, particularly those with supraglottic tumors(Steuer et al. 2017; Nocini et al. 2020). This delay in diagnosis typically requires aggressive treatments that severely impair patients' respiratory and vocal functions, thereby significantly dimishing their quality of life(R et al. 2019). Despite advances in treatment regimens, combining surgery with radiotherapy and chemotherapy, the five-year overall survival rate has shown limit

improvement(Cavaliere et al. 2021; Wang et al. 2020), highlighting the urgent need for molecular biomarkers specific to LSCC. Compared to other HNSCC subtypes, LSCC has the unique biological characteristics and epidemiological features(Cavaliere et al. 2021; Pan et al. 2025), making specialized research essential to address its unique diagnostic challenges and the critical need for organ function preservation.

In recent years, pan-cancer genomic studies have highlighted the pivotal role of structural variants in the occurrence and progression of cancer(Cosenza et al. 2022). Structural variants (SVs), which encompass large genomic rearrangements exceeding 50 bp, include deletions (DEL), insertions (INS), duplications (DUP), inversions (INV), and translocations (TRA) (van Belzen et al. 2021). These variants drive oncogenesis through mechanisms such as gene disruption and fusion, copy number alterations, and three-dimensional genome reorganization(Beroukhim et al. 2016). Large-scale cancer cohort studies have shown that somatic SVs are the most prevalent class of driver mutations in cancer, vastly outnumbering single nucleotide variants (SNVs) and small insertions/deletions (indels)(Cosenza et al. 2022). Given their abundance and diverse roles in cancer genomes, a comprehensive understanding of somatic SV's scope and mechanisms is essential for uncovering key cancer mutations in patient tumors.

However, despite the widespread clinical application of somatic SV detection methods based on second-generation short-read sequencing, these approaches have inherent limitations in sensitivity for smaller structural variants (≤10 kb), and in resolving repetitive sequence variants (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium 2020; Choo et al. 2023). The advent of nanopore long-read sequencing technology offers a promising solution to overcome this bottleneck. With read length exceeding 10 kb, nanopore sequencing can span SVs breakpoint regions in their entirety, enabling comprehensive mapping of somatic SV landscapes. Recently, several somatic SVs algorithms, such as nanomonsv(Shiraishi et al. 2023), Severus(Keskus et al. 2025), and SVision-pro(Sahlin et al. 2023), have been developed to harness this potential. However, existing tools still face challenge in correcting alignment biases in repetitive regions, distinguishing mixed signals of multi-scale INSs and DELs, and their practical application in clinical samples of sufficient scale remains limited(Sahlin et al. 2023). Therefore, developing detection workflows suitable for different research scenarios is a

key step for advancing the clinical interpretation of SVs.

Here, we present SomaGauss-SV, a nanopore-based somatic SV caller developed to address this limitation. We rigorously benchmark its performance in paired tumour–normal cell lines using complementary, orthogonal validation strategies to establish robustness before clinical deployment. We then apply this framework to delineate the somatic SV landscape of primary LSCC and interrogate potential biological and etiological associations. Ultimately, this study aims to deliver a high-performance workflow for somatic SV discovery, elucidate recurrent SV-linked mechanisms in LSCC, and generate a comprehensive, clinically annotated SV resource to enable future mechanistic and translational investigations.

Results

SomaGauss-SV demonstrates excellent performance in detecting somatic structural variations

Current long-read sequencing algorithms for detecting SVs often exhibit reduced accuracy in repetitive genomic regions. This limitation arises from inherent biases in traditional seed-chain alignment methods(Sahlin et al. 2023) (Supplemental Fig. S1). The lack of systematic alignment error correction in existing SVs detection tools further constrains their utility for somatic SVs calling. To address these challenges, we present SomaGauss-SV, a novel workflow (Fig. 1A) that integrates a mixed Gaussian distribution model with alignment error correction to improve the accuracy of somatic SVs detection. Its core innovation lies in applying tailored Gaussian modeling for different SV types coupled with stringent population-based filtering to distinguish somatic events (see Methods for full details).

To systematically assess the performance of SomaGauss-SV in somatic SVs detection, we utilized a public resource that provides five tumor-normal cell line pairs (HCC1395/HCC1395BL, HCC1937/HCC1937BL, HCC1954/HCC1954BL, H1437/H1437BL, H2009/H2009BL,Supplementary Table 1) with matched sequencing data and a curated SV ensemble as the ground truth (Keskus et al. 2025). We employed two independent evaluation strategies, Truvari (English et al. 2022) and minda (Keskus et al. 2025), to compare SomaGauss-SV against established somatic SV callers: nanomonsv (Shiraishi et al. 2023),

SVision-pro (Wang et al. 2025), and Severus (Keskus et al. 2025). For a fair comparison in this controlled setting, we omitted the panel-of-normal (PON) filtration step.

SomaGauss-SV consistently achieved the highest F1 scores across all five cell line pairs using the Truvari evaluation strategy, achieving a mean F1-score of 62.8% (±19.2% SD), significantly outperforming Severus (51.5% ±18.0%), SVision-pro (42.0% ±16.3%), and nanomonsv (28.9% ±11.1%) (Fig. 1B). Notably, SomaGauss-SV achieved peak performance (84.9% F1-score) in the H2009 cell line, which is characterized by insertion-dominant SVs profiles (mean comparator performance: 57.3%, Supplemental Fig. S2A), highlighting its particular strength in resolving this challenging variant class. In contrast, the HCC1395 cell line, with its high SVs burden, showed narrower performance differentials between SomaGauss-SV (74.9%) and Severus (72.0%), both substantially outperforming the other tools (mean: 42.5%). The genome-doubled HCC1954 cell line exhibited universally lower F1-scores across all platforms (mean: 28.2% vs. 50.8% in non-polyploid lines), suggesting that polyploidization introduces systematic challenges in SVs detection. A critical analysis revealed the precision-recall trade-offs of each tool (Fig. 1C): nanomonsv exhibited higher precision (59.1% vs. competitors' mean precision 50.7%), while Severus and SVision-pro prioritized recall (64.3% vs. competitors' mean recall 40.6%). SomaGauss-SV uniquely balanced both metrics (mean precision: 64.0%, mean recall: 61.9%). Considering practical application, we also evaluated the computational space and time complexity. SomaGauss-SV exhibited the smallest Maximum Resident Set Size, and its Elapsed Real Time was significantly less than that of Severus, despite their similar F1-scores (Supplemental Fig. S2B).

The inter-cell line performance variation strongly correlated with the composition of somatic SVs types. To explore this relationship further, we conducted stratified analyses by SVs types using Truvari with type-specific ensemble settings (Supplemental Fig. S2C). SomaGauss-SV consistently outperformed other tools across all SVs types. Notably, in INS detection, SomaGauss-SV achieved an F1-score of 59.3%, outperforming comparator tools by 13.9–93.1 % (mean F1-score of comparators: 25.5%). For DUP, SomaGauss-SV exhibited performance comparable to

Severus (67.7% vs. 69.1%). While all tools showed relatively higher efficacy in DEL detection (mean F1-scores: 45.4–70.2%), both breakend (BND) and inversion (INV) detection proved challenging across platforms (mean F1 < 5.7%). This limitation was most pronounced in the BND-rich HCC1954 dataset, where BNDs constituted 64.2% of SVs (Supplemental Fig. S2C). To ensure a robust and unbiased evaluation, we conducted a complementary assessment using minda (see Supplementary Table 2 for full details), which matches variants based on breakpoint coordinates and length, independent of SV type annotation. Re-evaluation with minda yielded two key insights. First, under this more lenient and type-agnostic matching scheme, F1-scores for all tools increased, with the most substantial improvements (15.56%–48.47%) observed in the BND-rich HCC1954 dataset (Supplemental Fig. S2D), highlighting the inherent difficulty in computationally distinguishing BND. Second, and more importantly, SomaGauss-SV's advantage in precision became even more pronounced. It achieved the highest average precision (79.20%) across four cell lines (Supplemental Fig. S2E), demonstrating its ability to generate a highly reliable, low false-positive call set—a critical feature for downstream biological discovery. In light of the inherent difficulty in computationally detecting complex SVs like INV and BND, our final analysis pipeline incorporates a mandatory manual curation step for these variant types in tumor–normal pairs to ensure accuracy (see Methods). Beyond type-specific performance, we analyzed the concordance between different SV detection tools using the HCC1395/HCC1395BL cell line pair (Fig. 1D). SomaGauss-SV demonstrated superior consensus validity (92.2% overlap) and the highest true positive rate (23.94%) among its unique calls (Fig. 1E). Although consensus calling improved true positive rates, it incurred substantial variant loss (52.3%) and systematic class biases (e.g., the proportion of DEL detection increased by 29.0%; Supplemental Fig. S2F), underscoring its limitations for comprehensive SV profiling. Collectively, these findings position SomaGauss-SV as the optimal solution for our LSCC study, balancing precision and recall while preserving both high-confidence consensus variants and biologically relevant unique calls - critical advantages for downstream clinical interpretation.

Pathological information and sequencing quality data display of ONT WGS samples

To interrogate the Somatic SVs landscape of LSCC, we included tumor samples, comprising with matched blood samples from 15 treatment-naive with primary, HPV-negative LSCC (Supplementary Table 3, Fig. 2A). These tumor samples represented a broad spectrum of clinical stages and anatomical subtypes (Fig. 2B and 2C). All patients were male, with a mean age of 63 years (median: 60). Smoking exposure was assessed using the smoking index, calculated as the number of cigarettes per day multiplied by years of smoking(Zhou et al. 2023). 14/15 patients had a history of smoking, with an average smoking index of 648 (median: 610) (Supplementary Table 3, Fig. 2D). These cohort characteristics align with previous studies, demonstrating that LSCC predominantly affects middle-aged and elderly males and is strongly associated with smoking(Johnson et al. 2020; Lechien et al. 2022), reflecting the typical demographic profile of LSCC patients.

Whole-genome sequencing of paired tumor and blood samples was performed using the PromethION platform (R9: 9 samples, R10: 6 samples; Supplementary Table 3), generating 2.10 Tb of data. After quality control (sequencing quality > 7, read length > 1 kb), tumor samples yielded an average N50 of 28 kb (range: 20 kb to 36 kb) and an average of 94 billion bases per sample, equating to a mean sequencing depth of 30× (range: 28× to 33×, Fig. 2E). Matched blood samples generated an average N50 of 28 kb (range: 19 kb to 40 kb) and an average of 49 billion bases per sample, resulting in a mean sequencing depth of 15× (range: 12× to 18×, Fig. 2E). Alignment of the clean long-read sequencing data to the hg38 reference genome using minimap2 yeilded average alignment rate of 99.95% for tumor samples and 99.90% for blood samples (Supplementary Table 4, Fig. 2F).

Discovery of the positive relation between somatic DEL burden and tobacco exposure

Using SomaGauss-SV, 8128 somatic SVs were identified across the 15 paired samples, with an average of 541 somatic SVs per tumor samples. Somatic insertions (INS) were the most common variant, comprising 54.7% of the total, followed by deletions (DEL, 19.7%), rearrangements (BND, 18.0%), inversions (INV, 4.0%), and duplications

(DUP, 3.6%) (Supplemental Fig. S3A). The distribution of SV types in individual samples was consistent with this overall pattern (Fig. 3A). Notably, samples with higher smoking indices displayed a significantly greater somatic SV burden (Pearson correlation coefficient, $p<0.05$, $r=0.54$; Supplemental Fig. S3B). DEL contributed most prominently to this trend, exhibiting the strongest correlation with smoking index among all SV types (Pearson correlation coefficient, $p < 0.05$, $r = 0.67$; Fig. 3B). To dissect this relationship more rigorously, we performed negative binomial regression, treating the burden of specific SV types as the dependent variable while controlling for TNM stage, age, and sequencing depth. This multivariate analysis identified smoking index as a highly significant, independent positive predictor for DEL burden specifically ($p = 0.000596$; Supplemental Fig. S3C). To strengthen the generalizability of this finding, we validated the association in an independent head and neck squamous cell carcinoma cohort from the ICGC (n=44, 19 of whom had no smoking history). After controlling for TNM stage, age, and sex via negative binomial regression, smoking index remained significantly associated with DEL burden ($p = 0.0498$; Supplemental Fig. S3D).This finding similar with previous study indicating that smoking increases the somatic mutation burden in bronchial epithelial cells and LSCC(Yoshida et al. 2020; Degawa et al. 1994), suggested that tobacco exposure may promote somatic DEL by introducing DNA double-strand breaks(Yoshida et al. 2020; Degawa et al. 1994; Aw et al. 2021).

To investigate the functional impact of somatic deletions in smoking-related LSCC, we focused on the 42 somatic DEL regions (after merging adjacent windows, as described in Methods: Sample Enrichment Analysis) that were shared by three or more samples from smokers. Using GREAT (McLean et al. 2010) , these loci were mapped to 49 putative distal regulatory target genes. Pathway enrichment analysis revealed that six of the top ten enriched pathways ($q < 0.05$) were immune-related, including Dectin-2 family(Graham and Brown 2009), Translocation of ZAP-70 to Immunological Synapse(Blanchard et al. 2002), and Complement Cascade(Afshar-Kharghan 2017)(Fig. 3C). GO analysis further supported this immune-focused pattern, with nine of the top ten biological processes ($q < 0.05$) related to lymphocyte activation and cytokine responses (Supplemental Fig. S3E).

These findings supported the role of smoking in tumorigenesis through immune-inflammatory responses(Aw et al. 2021), suggested that somatic DEL frequently carried by high-smoking LSCC patients may regulate the smoking-associated immune microenvironment.

Among the immune-related somatic DEL regions, a particular a somatic DEL, shared by four smoking-related LSCC samples, may affect the expression of *HLA-DRA* and *HLA-DRB5*, which are key immune modulators involved in tumor immune evasion. This high-frequency DEL is located near a CTCF binding site, identified in the MCF-7 and A673 cell lines, 265 bp apart (Fig. 3D). The CTCF site is predicted to regulate *HLA-DRA* and *HLA-DRB5* (The ENCODE Project Consortium 2012). We propose that the somatic DEL disrupts CTCF binding, leading to altered chromatin architecture and dysregulated expression of *HLA-DRA* and *HLA-DRB5*. Additionally, we identified a somatic DEL in the *MUC3A* gene, which encodes mucin and has been associated with smoking-induced changes in chronic obstructive pulmonary disease (COPD)(Merikallio et al. 2023). *MUC3A* expression is reported to be higher in smokers than in non-smokers due to tobacco exposure(Gendler and Spicer 1995). The somatic DEL, located at exon 2 of *MUC3A* gene (DEL length: 1120 bp to 1123 bp), contains a simple tandem repeat sequence (motif: ACC; Fig. 3E). PCR analysis of cDNA from clinical samples demonstrated that this genomic deletion results in a deletion within the *MUC3A* transcript, resulting in a shorter band (~346-349 bp) compared to the expected full-length product (1469 bp) (Supplemental Fig. S3F). This exon encodes a peptide rich in threonine (Thr) and serine (Ser), which are essential for glycosylation and for mucin 3A's role in protecting the epithelial barrier from environmental damage (Bhatia et al. 2019; Kufe 2009). Therefore, the somatic DEL is likely to cause truncation of this critical peptide domain, potentially impairing mucin 3A function.

Charateristics of somatic SVs

We characterized somatic SVs in LSCC by analyzing SV size, sequence context, and breakpoint distribution. Our results highlight distinct length preferences across SV types (Fig. 4A). Somatic DEL and INS were characterized by shorter lengths and

narrower distributions. DEL (mean: 1493 bp, range: 50 bp to 30 kb), had a peak at size of 70 bp, while INS (mean 1107 bp, range: 50 bp to 24 kb), exhibiting a tri-modal distribution with peaks at 70 bp, 170 bp, and 300 bp. In contrast, DUP and INV displayed larger, more dispersed length distributions (DUP mean: 39 kb, INV mean: 5.3 Mb). These patterns are similar with prior study on somatic SVs in HCC(Zeng et al. 2025).

To investigate the sequence determinants of somatic SVs, we analyzed the sequence feature of SVs at characteristic size peaks. Given the relatively uniform length distributions of DUP and INV (Fig. 4A and Supplemental Fig. S4A), we focused on DEL and INS (Fig. 4B and 4C). At the 70 bp peak, both somatic DEL and INS were significant enriched in simple repeat sequences. Notably, AAAG motifs predominated in somatic DEL (30%, Supplemental Fig. S4B), whereas somatic INS exhibited a strong AT-richness (58.6% AT, Supplemental Fig. S4C). The 170-bp INS peak showed a notable association with centromeric regions. Specifically, 45.4% of these 170-bp INS harbored Satellite/CEN (monomers: ~170 bp (Aldrup-MacDonald et al. 2016), Pearson's Chi-squared test, p-value = 1.1×10^{-6}). Among these satellite/CEN-containing INS, 55.6% were directly located within centromeres, a significant enrichment compared to the genome-wide background of 12.3% (Pearson's Chi-squared test, p-value < 0.001; Supplemental Fig. S4D). Even those not residing inside centromeres clustered in close proximity, with 62.7% found within 1 Mb of centromeric regions (Wilcoxon test, p-value < 0.001, Supplemental Fig. S4D). This pericentromeric enrichment pattern was further confirmed by re-alignment of Chromosome 20 reads to the complete T2T-CHM13 assembly (Wilcoxon test, p-value < 0.001, Supplemental Fig. S4E). At the 300-bp peak, SINE/*Alu* elements, primarily the *AluY* subclass, were predominant in INS, comprising 43.9% of the total (Pearson's Chi-squared test, p-value = 3×10^{-10}) (Supplemental Fig. S4F). This observation may be attributed to the *AluY* subfamily, the youngest and most active member of the *Alu* family(Hormozdiari et al. 2011; Witherspoon et al. 2010), whose high activity has been identified in both cancer and healthy populations(Wu et al. 2021; Zeng et al. 2025; Ma and Pl 2002).

Somatic SVs breakpoints were non-uniformly distributed across the genome, with significant enriched in repetitive elements (TE enrichment score: 0.85; repeat enrichment score: 10.20; non-repetitive regions enrichment score: 0.82). Hyper-repetitive regions (including satellite, simple repeats, low complexity regions and SVA elements) showed substantial enrichment (mean enrichment score > 6.56, Fisher's exact test, *p*-value < 0.05, Fig. 4D). Notably, somatic DEL and INS were more frequently observed in centromeric satellite sequences (enrichment score > 6.4, Fisher's exact test, *p*-value < 0.05), whereas larger SVs, such as INV and DUP, were underrepresented in these regions (enrichment score < 0.54). This pattern may be due to satellite sequences' susceptibility to replication fork slippage or mispairing, combined with the compact chromatin structure of centromeric regions that limits DNA accessibility and favors conserved repair mechanisms, reducing the occurrence of large-scale SVs (Kashi and King 2006). Sequence annotation revealed that over 60% of INS inserted in satellite/center, satellite, and simple repeat regions harbored inserted sequences identical to flanking repeats (Supplemental Fig. S4G). A representative case in simple repeats demonstrated a 12-fold amplification of the reference TTCT motif in tumor samples (Supplemental Fig. S4H), indicating that somatic INS preferentially propagate existing repetitive architectures through replication- or repair-associated templated insertion.

To elucidate the regulatory consequences of somatic SVs, we performed an enrichment analysis of SV breakpoints across functional genomic regions. Somatic INS, DUP, DEL, and INV exhibited significant enrichment within proximal transcriptional regulatory domains (Fig. 4E), particularly in promoter-proximal (2.5kb) and termination regulatory regions (downstream 2.5 kb), suggested that these SVs may alter gene expression through *cis*-regulatory interference(Cosenza et al. 2022). Specifically, INS and BND breakpoints were broadly localized to *cis*-regulatory elements, with the strongest enrichment observed at CTCF-binding sites (enrichment score > 3.19, Fisher's exact test, *p*-value < 0.05, Supplemental Fig. S4I), suggested potential disruption of topologically associating domain boundary. In contrast, DUP and DEL breakpoints preferentially disrupted enhancers and promoters near transcription/termination sites, suggesting they preferred modify the regulatory

relationships between elements and genes (enrichment score >1.40, Fisher's exact test, *p*-value < 0.05, Supplemental Fig. S4I).

Identify extensive high-frequency somatic SVs regions in LSCC

To detect recurrent somatic SVs across genome, we performed a non-overlapping sliding window approach for sample carrier rate quantification. We identified 226 high-frequency SV regions (sample carrier rate > 26%) exhibiting non-random chromosomal distribution (Fig. 5, Circos plot (Krzywinski et al. 2009)), and significant co-localization with known tumor-related genes (Tate et al. 2019) (Chi-squared test, *p*-value < 0.05, Supplementary Table S5). To assess robustness, we performed orthogonal validation using an independent SV caller (Severus) on the same dataset. Notably, two INS hotspots identified by SomaGauss-SV were concordantly confirmed by Severus, underscoring their biological relevance beyond algorithmic artifacts (Supplemental Fig. S5).

We further characterized two high-frequency somatic INS regions, Chr7:121,603,263–121,603,506 (shared by 9 LSCC samples, Supplemental Fig. S6A) and Chr15:29,945,811–29,950,000 (shared by 7 LSCC samples, Supplemental Fig. S6B), which overlapped with recurrent repeat expansions (rREs) previously identified in the TCGA pan-cancer cohort (Erwin et al. 2023). The former INS aligned with a rRE specifically detected in lung squamous cell carcinoma, while the latter corresponds to a rRE characteristic of prostate cancers. The inserted sequences of these somatic INS matched the motifs described in previous study, suggested that these somatic INS have the potential to serve as pan-cancer biomarker across multiple malignancies. Moreover, we identified several high-frequency DEL that may influence nearby genes by disrupting regulatory elements. For instance, the region Chr11:117,260,000–117,262,041 harbored somatic DEL (235-353 bp) in six tumors (Supplemental Fig. S6C). This region is located within an intron of *RNF214* and overlapped with an enhancer element predicted by ENCODE(Je et al. 2020). This enhancer is predicted to regulate *RNF214* expression in PC-3 prostate cancer cells(Je et al. 2020). RNF214 is an E3 ubiquitin ligase. Its overexpression promotes tumor cell

proliferation, migration, and invasion(Lin et al. 2024). Additionally, in Chr12:54,206,733–54,210,000, somatic DEL (163-178 bp) in five samples overlapped with an enhancer predicted to regulate *SMUG1* expression in Panc1 pancreatic cancer cells(Je et al. 2020) (Supplemental Fig. S6D). *SMUG1* is involved in base excision repair and genomic stability maintenance(Iakovlev et al. 2019). High *SMUG1* expression correlates with poorer survival in the TCGA HNSCC cohort (Log-rank test, *p*-value = 0.003) (Supplemental Fig. S6E).

We also identified eight somatic INV shared by more than four samples (Supplemental Fig. S6F). The most frequent event occurred in region Chr7:77,175,000–102,599,637 (9/15 samples). This interval encompasses 197 protein-coding genes, including 11 annotated proto-oncogenes. Pathway enrichment analysis revealed significant enrichment for the Cytochrome P450 metabolic pathway ($q < 0.05$), which includes genes within this region such as *CYP3A5* and *CYP3A7*. Notably, CYP family enzymes can be induced by polycyclic aromatic hydrocarbons in tobacco smoke, enhancing the metabolic activation of procarcinogens and thereby promoting tumorigenesis(Zevin and Benowitz 1999).Analysis of INV breakpoints showed clustering at similar genomic positions in 7 of the 9 positive samples, resulting in INV fragments of 18–30 Mb. These clustered into two distinct groups: LSCC1, 3, 6, 14 and LSCC5, 7, 11 (Supplemental Fig. S7). Of particular note, the breakpoints in the first group (LSCC1, 3, 6, 14) localized within an intron of *CCDC146* and exon 7 of *RASA4* (Supplemental Fig. S7). *RASA4* encodes a calcium-dependent GTPase-activating protein that inactivates RAS and suppresses the MAPK pathway, functioning as a tumor suppressor. Its loss of function is implicated in malignant progression across various cancers(Chen et al. 2021; Prior et al. 2020). Since large inversions can disrupt gene architecture (Xu et al. 2023), this recurrent event may contribute to LSCC pathogenesis by altering *RASA4* transcription or function.

Analysis and validation of a somatic SRE in 71.8% LSCC patients

A somatic INS, observed in 66% (10/15) of LSCC tumors, was the most frequent

recurrent SV in our study (Fig. 6A). It was located at Chr1: 224,011,691-224,016,553, a simple repeat region annotated by RepeatMasker(Tarailo-Graovac and Chen 2009). The region initially consisted of a short tandem repeat with the (AATGG) repeat motif in normal samples, while the somatic alteration in 9/10 LSCC samples involved an expansion of this pattern, specifically a simple repeat expansion (SRE) (Fig. 6B).The LSCC11 sample differed, harboring Human Satellite II as its repeat motif (HSATII , a satellite DNA derived from the (CATTC)n repeat (Altemose et al. 2022)). The somatic SRE exhibited low supporting read ratios in six samples (10%-30%), while four exhibited stronger support (>50%) (Supplemental Fig. S8A). SRE lengths varied among tumors (median: 581 bp, Fig. 6C and S8A). LSCC15 had a 3,610 bp INS, while other samples showed peak SRE lengths around 450 bp (mean 517 bp, SD 382 bp). Additionally, two patients harbored multiple INS of different lengths. For example, the LSCC1 sample harbored 390 bp, 650 bp, and 1,030 bp INS (Supplemental Fig. S8A and Fig. 6A), indicating more tumor heterogeneity.

To validated the presence of somatic SRE identified through LRS WGS, target PCR amplification of the SRE region was performed on DNA from avialable 14 paired tumor and blood samples (including 9 tumors with INS identified by LRS WGS). Gel electrophoresis indicated that 6/9 tumors displayed longer fragments compared to blood samples (Fig. 6D and 6E). The remaining 3 samples presented challenges due to low SRE supporting ratio (20% and 30% for LSCC10 and LSCC11) and a long insertion length in LSCC15 (3,610 bp)(Supplemental Fig. S8A), complicating the gel electrophoresis validation. Despite these challenges, our results confirm the authenticity and reliability of these high-frequency somatic SRE.

To assess the prevalence of the high-frequency somatic SRE in LSCC tumors, we further collected 24 paired LSCC samples for PCR validation and ONT Sequence (Supplemental Fig. S8B-S8G, S9-S19). Gel electrophoresis and LRS data of the target PCR products revealed that 18/24 LSCC samples carried the somatic SRE with an (AATGG) repeat motif, with insertion lengths ranging from 130 bp to 3,693 bp (Supplemental Fig. S9-S19, Supplementary Table S6). Furthermore, this SRE was detected in the peritumoral tissues of 15/21(71.42%) LSCC tumors (Supplemental Fig.

S8G, Supplementary Table S6). Together, the high prevalence of this somatic SRE in peritumoral tissue suggested it is an early genomic event, while its significantly elevated allele frequency in tumor compare with peritumoral tissue (*t*-test, *p*-value = 0.0046; Supplemental Fig. S8H) indicates positive selection of the SRE-harboring clone during tumor evolution.

To further explore the functional role of the high-frequency somatic SRE, we analysis the chromatin environment surrounding the SRE. In the A673, A549, and SK-N-MC cell lines, the SRE region overlapped with DNase-H3K4me3 signals, which are known to mark transcriptional start sites(Je et al. 2020)(Fig. 6C). While the target genes regulated by this DNase-H3K4me3 mark have not been extensively characterized, Hi-C data from the hypopharyngeal cancer FaDu cell line suggested that the region resides within the same topological domain (TAD) as the nearby coding genes *TP53BP2* and *FBXO28* (Supplemental Fig. S20A). Furthermore, data from the Four-Dimensional (4D) Nucleosome Consortium(Zhu et al. 2022) supported the presence of three-dimensional chromatin interactions between this DNase-H3K4me3 region and the genes *TP53BP2* and *FBXO28* (Fig. 6C). Critically, analysis of matched TCGA LSCC samples confirmed that, among genes near the SRE, the expression of *FBXO28* and *TP53BP2* exhibited the most significant tumor-specific upregulation (*t*-test, *p*-value < 0.001; Supplemental Fig. S20B). These convergent lines of evidence raised the possibility that this somatic SRE may alter the regulation of these two genes. Consistent with this hypothesis, qPCR analysis of our clinical LSCC samples revealed that tumors harboring the SRE exhibited significantly higher expression levels of both *TP53BP2* and *FBXO28* compared to SRE-negative tumors (*t*-test, *p*-value = 0.00039 and *p*-value = 8.1×10^{-6}, Fig. 6F), directly linking the presence of the SRE to upregulated expression of its putative target genes. *TP53BP2* encodes the apoptosis-stimulating protein, which interacts with P53 family members to modulate cell apoptosis and growth(Huo et al. 2023). *FBXO28*, which encodes a ubiquitin ligase, has been reported to be aberrantly overexpressed in human epithelial cancer cell lines(Song et al. 2024).

We next examine the regulatory effect of the high-frequency somatic SRE on gene expression using cell line models. With gel electrophoresis, we identified longer fragments in two LSCC cell lines (SNU46 and SNU899) compared to a normal human bronchial epithelial cell line (BEAS-2B) (Supplemental Fig. S8B). Nanopore sequencing of the PCR products further confirmed that both LSCC cell lines harbored an expected INS with AATGG repeat motif. Specifically, SNU46 and SNU899 carried homozygous INS with lengths of 165 bp and 240 bp, respectively (Supplemental Fig. S20C and S20D). In contrast, no significant sequence length variation was observed in BEAS-2B cells compared to the reference genome. We then selected the SNU899 cell line, which has a homozygous SRE with longer insertion, and the BEAS-2B cell line for quantitative real-time PCR (QPCR) to assess whether the SRE affects the expression levels of *TP53BP2* and *FBXO28*. Consistent with findings in primary tumors, We found that *TP53BP2* and *FBXO28* expression levels were significantly higher in the SNU899 cell line than in the BEAS-2B cell line (Fig. 6G, *t*-test, *TP53BP2*: p-value < 0.01; *FBXO28*: p-value < 0.001). To rule out tissue-specific baseline differences as the cause, we analyzed normal tissue data from TCGA (LSCC vs. LUSC) and cell line data from DepMap. This confirmed that expression of these genes in laryngeal contexts was not inherently higher than in pulmonary contexts, supporting the SRE's specific role in upregulation (Supplemental Fig. S20E, F). To directly test whether the SRE alters local chromatin architecture, we performed 3C-PCR with an anchor primer (Anchor-S) placed downstream of the simple repeat region where the SRE occurs. In SNU899 (SRE-positive), the Anchor-S locus exhibited strong, specific spatial proximity to the F1 regulatory site upstream of *FBXO28*. In BEAS-2B (SRE-negative), the same Anchor-S locus showed a markedly weaker interaction with *FBXO28*-F1. Instead, its predominant interactions were with sites near *TP53BP2* (T3/T8) (Supplemental Fig. S20G). This reciprocal pattern demonstrates that the presence of the SRE is associated with a distinct three-dimensional chromatin conformation, specifically promoting spatial proximity to the regulatory region of *FBXO28*. Together, these findings suggest that the somatic SRE may upregulate *TP53BP2* and *FBXO28* expression in LSCC by altering the local three-dimensional chromatin architecture.

Discussion

Somatic structural variations (SVs) are critical hallmarks of tumorigenesis (Cosenza et al. 2022). However, technical limitations of SRS in sensitively detecting somatic SVs, compounded by the tissue heterogeneity characteristic of HNSCC, have left the landscape of high-frequency somatic SVs in LSCC underexplored. This gap in knowledge has hindered the identification of potential LSCC-specific molecular biomarkers for effective cancer management. To address these challenges, we developed SomaGauss-SV, a LRS-optimized computational framework for somatic SVs detection. Benchmarking against established tools revealed that SomaGauss-SV offered a superior balance of sensitivity and specificity, with an F1-score improvement of 11.3-33.9%. SomaGauss-SV excels in two key areas: (1) it reduces alignment errors in repetitive genomic regions through a multi-step filtering approach, and (2) it accurately distinguishes true somatic INS, DUP and DEL in the presence of size-variable INS/DUP/DEL using a Gaussian mixture model. Despite these advances, significant technological challenges persist. Current LRS platforms still struggle with resolving complex breakpoint architectures, particularly those involving nested rearrangements and ultra-long SVs (>100 kb)(Keskus et al. 2025), limitations shared across existing detection tools, including SomaGauss-SV. Furthermore, the current benchmarking landscape relies heavily on consensus-based “gold standard” sets derived from multiple sequencing platforms and computational tools rather than orthogonal experimental validation. This approach may introduce methodological bias, especially for complex SVs, which remain difficult and costly to validate experimentally, resulting in limited and potentially incomplete ground-truth datasets (e.g., the COLO829 benchmark provides only 68 gold-standard SVs (Espejo Valle-Inclan et al. 2022)). We also observed that the choice of benchmarking strategy (minda vs Truvari) itself significantly impacts performance evaluation. There is currently no consensus in the field on whether SV type should be incorporated as a matching criterion during benchmarking. While classifying complex SVs can be ambiguous, applying a consistent typing standard is necessary for the robust evaluation of simple somatic variants (e.g., INS, DEL, DUP), as type-agnostic matching may misclassify genuine events. However, recent innovations in high-throughput ultra-long-read sequencing technologies (Wang et al. 2021) and

graph-based or assembly-based long-read alignment algorithms (Li and Durbin 2024) offer promise for overcoming these obstacles.

We systematically analyzed the LRS WGS data of 15 paired LSCC samples using SomaGauss-SV, revealing the distribution and sequence features of somatic SVs. We identified a dose-dependent positive correlation between smoking intensity and somatic DEL burden in LSCC. Previous studies indicated tobacco exposure is the primary driver of HNC, with tobacco users exhibiting higher frequencies of somatic SNVs and indels(Alexandrov et al. 2016; Torrens et al. 2025). Notably, tobacco-related mutational signatures show higher burdens and frequencies in laryngeal cancers compared to other HNC subsites(Alexandrov et al. 2016; Torrens et al. 2025). However, the link between tobacco and somatic SVs remains less clearly defined(Yoshida et al. 2020; Alexandrov et al. 2016; Torrens et al. 2025; Jethwa and Khariwala 2017). Our study fills this gap by demonstrating the role of chemical carcinogens in chromosomal instability. Specifically, we suggested that the accumulation of somatic DEL may result from the synergistic effects of multiple carcinogens in tobacco smoke. DNA adducts from polycyclic aromatic hydrocarbons and nitrosamines can block replication forks and cause DNA double-strand breaks(Yoshida et al. 2020). These breaks, when repaired poorly, may lead to chromosomal deletions. Also, free radicals and reactive oxygen species from tobacco can cause more DNA damage through oxidative stress, promoting deletions during faulty repair(Yoshida et al. 2020). Tobacco components might also disrupt cell division or cycle regulation(Aw et al. 2021), increasing chromosomal instability. Although high expression of *CYP1A1*, the key enzyme involved in the metabolism of benzopyrene and nitrosamines, has been proposed as a potential cause of the elevated tobacco-related mutation signatures in laryngeal tumors(Degawa et al. 1994; Hecht and Hatsukami 2022; Yamazaki et al. 1992), the specific tobacco carcinogens responsible for the observed DEL burden remain to be identified in vitro. Clinically, DEL burden could serve as a potential biomarker for risk stratification and therapeutic target screening in LSCC. Further investigation into high-frequency deletions associated with smoking may provide evidence for preventive interventions targeting high-dimensional genomic regions. For example, somatic DEL might disrupt CTCF binding in 3D chromatin architecture, influencing the expression of immune genes like *HLA-DRB5* and *HLA-DRA*. However, the limited sample size in this study restricts the generalizability of the observed association between somatic SVs burden and smoking.

Other potential confounders, such as alcohol consumption and dietary factors, may also influence the results. Future studies with larger cohorts and the incorporation of multivariable models or in vivo/vitro experiments are needed to confirm these findings.

In addition, 226 high-frequency SV loci were identified, exhibiting selective advantages during tumor clonal evolution. Cross-cancer comparison analysis revealed that high-frequency SVs at Chr7:121,603,263-121,603,506 and Chr15:29,945,811-29,950,000 overlap with hotspot somatic repeat expansions reported in pan-cancer studies(Erwin et al. 2023), suggesting their involvement in conserved carcinogenic pathways across various cancer types. Significant CNV regions in laryngeal cancer from the TCGA database also overlap with the high-frequency large DELs, DUPs and INSs, further validating the unique strength of LRS technology in resolving multi-scale genomic rearrangements. A key issue to explore further is how to define the boundary between CNVs identified through sequencing depth and DEL/INS/DUP detected by long reads. Furthermore, we investigated co-occurring SV patterns within individual reads, which may represent complex structural rearrangements. For instance, in sample LSCC1, a single read simultaneously contained an insertion of a simple repeat (AT), a deletion within a LINE/L1 element, and an insertion of a SINE/*AluY* sequence (Supplemental Fig. S21). In LSCC3, we identified an event structurally resembling an "inverted duplication" as previously reported(Schloissnig et al. 2025), involving both INV and DUP on the same allele (Supplemental Fig. S22). These examples illustrate the capacity of long-read sequencing to resolve complex, multi-variant rearrangements.

Notably, analysis of both WGS and an independent validation cohort identified a somatic high-frequency SRE in 71.79%(28/39) of LSCC patients. Among the 21 patients with available peritumoral tissue, the SRE was detected in both tumor and matched peritumoral samples in 15 individuals. The allele frequency of the SRE was significantly higher in tumor tissue than in peritumoral tissue, suggested that this alteration may arise early during tumorigenesis and undergo subsequent positive selection. We also observed heterogeneity in SRE lengths within the same tumor, indicative of intratumoral diversity and potentially distinct selective advantages associated with different expansion sizes. This dynamic evolution was particularly

evident in a patient with lymph node metastasis, where the SRE was detected in the primary tumor (400 bp), peritumoral tissue (125 bp), and metastatic lymph node (200 bp). This suggested that the SRE evolves dynamically throughout tumor progression and metastasis, conferring varying selective advantages in different microenvironments. However, the mechanisms linking SRE length variation to phenotypic diversity in tumor cells require further investigation.

The recurrent somatic SRE identified in this study holds significant potential as a biomarker for LSCC. Integrating 3C-PCR and qPCR results from our cell line models, we propose a hypothesis in which the SRE may regulate the expression of neighboring genes by altering local three-dimensional chromatin architecture. Specifically, the SRE-mediated sequence elongation could increase spatial proximity between a distal regulatory element and the promoters of *TP53BP2* and *FBXO28*, thereby enhancing their transcription. Although external data suggested that baseline expression of these genes in laryngeal contexts is not inherently higher than in pulmonary tissues, the inherent tissue-of-origin heterogeneity between the lung epithelial cell line BEAS-2B and the LSCC-derived SNU899 line warrants caution in interpreting cross-tissue comparisons. Nevertheless, the functional relevance of these genes in cancer is well-supported: elevated *FBXO28* expression is linked to poor prognosis in ovarian cancer, where it activates the TGF-β1/Smad2/3 pathway to promote proliferation and invasion (Song et al. 2024). *TP53BP2*, a member of the ASPP family, regulates apoptosis, proliferation, and autophagy, and its upregulation correlates with poor prognosis in colorectal cancer and other malignancies (Huo et al. 2023). To definitively test our mechanistic hypothesis, future studies employing CRISPR-Cas9 to delete the SRE in relevant models would be essential to directly assess its impact on *TP53BP2* and *FBXO28* expression and downstream phenotypes, thereby excluding the influence of tissue-of-origin heterogeneity.

In conclusion, the bioinformatics pipeline developed in this study achieves an optimal balance between sensitivity and specificity, offering a robust tool for future research on tumor somatic SVs. This study represents the first comprehensive genome-wide survey of somatic SVs in LSCC using LRS, with the generated whole-genome sequencing data and identified SVs offering a valuable resource for clinical research in the field.

Method

Sample and patient information collection

This study involved 39 patients with LSCC who underwent surgical resection at West China Hospital, Sichuan University. Tumor samples and matched blood samples were collected from these patients during surgery and venipuncture. Throughout the tissue collection and utilization process, this study strictly adhered to the Declaration of Helsinki and recieved formal approval from the Ethics Review Committee of West China Hospital, Sichuan University. All patients participating in this study have signed written informed consent forms. In addition, key clinical information, including clinical staging, smoking history, age, and gender, was throroughly documented for each patient.

Cell culture

Human cell lines, obtained from Procell (Wuhan Pricella Biotechnology Co., Ltd.) were utilized in this study. The cells were grown in DMEM medium (Gibco, C11995500BT) or RPMI 1640 medium (Gibco, C11875500BT), supplemented with 10% fetal bovine serum (FBS) and 100 U/mL Penicillin-Streptomycin (P/S). All cell lines were maintained at 37□°C and 5% CO_2. Periodic mycoplasma testing was conducted using the Mycoplasma Detection Kit (Vazyme, D101). Cells were passaged at 70–90% confluency with 0.25% Trypsin-EDTA (Gibco, 25200-072).

Gene expression, chromatin conformation, and genomic validation assays

For gene expression analysis, total RNA was extracted, reverse transcribed, and quantified by qRT-PCR using gene-specific primers. To investigate chromatin interactions between the SRE region and target genes (*FBXO28* and *TP53BP2*), Chromatin Conformation Capture (3C) libraries were generated from cross-linked SNU899 cells via EcoRI digestion and intra-molecular ligation, and specific interactions were quantified by 3C-qPCR. Additionally, genomic DNA was used to validate the presence of the SRE in clinical samples and cell lines by PCR amplification followed by Sanger sequencing. The primer sequences are listed in the oligonucleotide table (Supplementary Table S7). Detailed protocols, including all primer sequences and reaction conditions, are provided in the Supplemental Methods.

Nanopore sequencing library preparation

For whole-genome sequencing, high-molecular-weight genomic DNA was extracted and used for library preparation with Oxford Nanopore ligation kits (SQK-LSK109 or SQK-LSK114), followed by sequencing on PromethION flow cells. For targeted sequencing of the SRE locus, PCR amplicons were prepared into libraries using either the Oxford Nanopore or Qitan Tech nanopore sequencing platforms and sequenced on GridION X5 or QNome-3841 instruments, respectively. Detailed protocols for library preparation, including specific reagents, reaction conditions, and purification steps, are provided in the Supplemental Methods.

Preprocessing of long-read sequencing data

For data preprocessing, raw long-read sequencing data were quality-filtered and assessed using NanoFilt (version 2.8.0)(Lee et al. 2021) and NanoPlot (version 1.39.0)(De Coster and Rademakers 2023) , respectively. Filtered reads were aligned to the hg38 reference genome using minimap2 (version 2.26)(Li 2018). Resulting SAM files were converted, sorted, and read-tagged using SAMtools (version 1.9)(Danecek et al. 2021) to generate final BAM files. For the analysis of SRE-targeted PCR products , a similar preprocessing pipeline was applied, with specific length filtering to ensure complete spanning of the repeat region. The processed alignments were visualized using the Integrative Genomics Viewer (IGV) (Robinson et al. 2011). Detailed parameters and steps are provided in the Supplemental Methods.

Acquisition and preprocessing of nanopore sequencing data and gold standard somatic structural variations data from cell lines

According to the reference(Keskus et al. 2025), we downloaded the nanopore sequencing data for five paired cell lines: HCC1395/HCC1395BL, H1437/H1437BL, H2009/H2009BL, HCC1937/HCC1937BL, and HCC1954/HCC1954BL, as well as the corresponding gold standard sets of somatic structural variations for these cell lines. To ensure the accuracy of the analysis, we filtered the gold standard data, retaining only those somatic structural variations that could be detected by the nanopore sequencing platform, totaling 4058, including DEL 1493, INS 906, DUP 385, INV 48, and BND 1226. Subsequently, following the standard workflow for nanopore sequencing data preprocessing, we aligned and processed these data, ultimately obtaining BAM files with "blood" and "tumor" tags, respectively.The LRS WGS data and benchmark SV sets of five cell line pairs were downloaded from NCBI SRA BioProject

PRJNA1086849.

SomaGauss-SV workflow

The SomaGauss-SV workflow for somatic SV detection involved four key stages. First, merged tumor-normal BAM files were analyzed with Sniffles2(version 2.0.6)(Smolka et al. 2024) to call candidate variants. Second, a type-specific screening strategy was applied: candidate DEL were validated using a Gaussian mixture model on locally extracted reads, while candidate INS and DUP were re-analyzed using Straglr(Chiu et al. 2021), a length-based Gaussian modeling tool optimized for repetitive regions. Third, for clinical samples, variants were filtered against a Panel of Normals derived from Chinese population SV data (Wu et al. 2021). using Jasmine (version 1.1.5)(Kirsche et al. 2023) to remove common germline polymorphisms. Finally, complex variants (INV and BND) were manually validated through visualization using the Integrative Genomics Viewer (IGV) (Robinson et al. 2011). Detailed parameters and scripts are provided in the Supplemental Methods.

To validate our foundational tool selection and demonstrate the incremental value of each step in the SomaGauss-SV pipeline, we conducted systematic benchmarking. First, Sniffles2 was empirically selected over cuteSV(Jiang et al. 2020) as the initial variant caller due to its superior performance (mean F1 score: 68.7% vs. 53.4%, Supplemental Fig. S23 A and B). Subsequently, a stepwise comparison of three workflows: (1) Original Sniffles: The merged VCFs obtained by running Sniffles2 alone, (2) Filter Reads: The SV set after further filtering out variants with any supporting reads in the matched normal (blood) sample, (3) SomaGauss (Full Pipeline): The final output after applying our Gaussian Mixture Model (GMM)-based filtering step to the "Filter Reads" set. It significantly enhanced precision (from 14.5% to 71.51%) and F1 scores (overall +39.63%, with a more pronounced improvement of +47.86% in challenging VNTR regions) by effectively removing subtle false positives, particularly those arising from alignment errors in repetitive sequences. This confirms that each component of SomaGauss-SV adds substantial value, culminating in a highly accurate and robust somatic SV detection workflow (Supplemental Fig. S23 C-F).

Benchmarking and performance evaluation of SV detection tools

For benchmarking comparisons, somatic structural variants were independently detected using three established calls: nanomonsv(Shiraishi et al. 2023), Severus (Keskus et al. 2025), and SVision-pro (Wang et al. 2025), following each call's recommended parameters and workflow. The performance of all callers was then evaluated against the gold-standard variant set using two evaluation tools: Truvari(English et al. 2022) and minda (Keskus et al. 2025). Detailed command-line parameters are provided in the Supplemental Methods.

Evaluating software runtime efficiency

To evaluate the runtime efficiency of each software tool, we used the command "/usr/bin/time -v" followed by the respective software command line to measure both execution time and maximum memory usage. Given that most tools involve multiple distinct steps in their workflows—for example, Severus includes separate phases for SNP calling and phasing of the normal BAM, haplotagging of normal and tumor BAMs, and somatic SV calling, while SomaGauss-SV comprises steps such as merging BAM files, calling SVs from the merged BAM, and running Straglr. We adopted the following practical approach to summarize resource consumption: Total Elapsed Real Time: We summed the Elapsed Real Time from each step of a given tool to represent its overall runtime. Maximum resident set size: We recorded the maximum Resident Set Size observed across all steps as the tool's peak memory footprint. This method provides a consolidated and realistic view of computational resource requirements for each somatic SV detection pipeline under evaluation.

Variant sequence annotation and genomic context analysis

For sequence annotation and refinement, the consensus sequences of somatic INS were first corrected using Iris (version 1.0.4)(Kirsche et al. 2023) and then annotated for repetitive elements using RepeatMasker (version 4.1.2)(Tarailo-Graovac and Chen 2009). To characterize the genomic context of somatic DEL, their breakpoint regions were intersected with a RepeatMasker-annotated genome using BEDTools (version 2.30.0)(Quinlan and Hall 2010) to identify overlaps with repetitive sequences. Detailed parameters for each tool are provided in the Supplemental Methods.

Genomic enrichment and recurrence analysis of somatic SV

To assess genomic enrichment of somatic SVs, we performed two complementary analyses. First, SV breakpoints were intersected with functional genomic elements (genes and candidate *cis*-regulatory elements from UCSC (Cm et al. 2020)) using BEDTools. Enrichment fold and statistical significance were calculated using Fisher's exact test based on a contingency table of variant counts within and outside these elements relative to genomic background. Second, to address alignment ambiguity in repetitive regions (Supplemental Fig. S1, (Sahlin et al. 2023; Ahsan et al. 2023)), we constructed a non-uniform window BED file by combining annotated simple repeat regions with 500 bp/5 kb windows in non-repetitive regions. Somatic SV loci were then intersected with these windows using BEDTools to calculate sample carrier rates, and adjacent regions with shared variants were merged prior to downstream analysis. Detailed procedures are provided in the Supplemental Methods.

Data access

The BAM files generated in this study have been submitted to the Genome Sequence Archive (Genomics, Proteomics & Bioinformatics 2025) in the National Genomics Data Center (Nucleic Acids Res 2025), China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences (GSA-Human; https://ngdc.cncb.ac.cn/gsa-human/) under accession number HRA017212.

Code availability

SomaGauss-SV is available at https://github.com/XuYan000131/SomaGauss-SV and as Supplemental Code.

Acknowledgements

We thank Ranlei Wei for managing the computational resources required for this study. This work was supported by grants from the National Natural Science Foundation of China (82371883 and 12441517 to JFL, 82173383 to DX, 32200508 to LX), the Noncommunicable Chronic Diseases-National Science and Technology Major Project (2026ZD0553400, 2026ZD0553404 to DX), the National Key Science and Technology Special Project for Deep Earth Research (2024ZD1000600 and 2024ZD1000606 to JFL), the National Key Research and Development Program of China (2024YFF1207002,4 to JFL and YXQ) , the Foundation

of Sichuan Provincial Science and Technology Program (2025ZNSFSC0728 to JFL), the 1·3·5 project for disciplines of excellence, West China Hospital, Sichuan University (ZYYC23024 to DX), the Project Supported by Scientific Reserch Fund of SiChuan Provincial Education Department (2024NSFC1187 to YL).

Authors' contributions

LX and XYL conceived and developed the methodology, performed bioinformatic analyses. JFL and YXQ collected the clinical samples and clinical data. BYY, XL, XY, XC and HHZ collected the clinical samples. YH and YL designed and performed experiments. LX, XYL, YXQ, YH wrote the manuscript. DX and JFL responsible for supervision of research, data interpretation, and manuscript preparation. All authors read and approved the final manuscript.

Competing interests

Dan Xie is the co-founder of Qitan Technology. Other authors declare no competing interests.

References

Afshar-Kharghan V. 2017. The role of the complement system in cancer. J Clin Invest 127: 780–789.

Ahsan MU, Liu Q, Perdomo JE, Fang L, Wang K. 2023. A survey of algorithms for the detection of genomic structural variants from long-read sequencing data. Nat Methods 20: 1143–1158.

Aldrup-MacDonald ME, Kuo ME, Sullivan LL, Chew K, Sullivan BA. 2016. Genomic variation within alpha satellite DNA influences centromere location on human chromosomes with metastable epialleles. Genome Res 26: 1301–1311.

Alexandrov LB, Ju YS, Haase K, Van Loo P, Martincorena I, Nik-Zainal S, Totoki Y, Fujimoto A, Nakagawa H, Shibata T, et al. 2016. Mutational signatures associated with tobacco smoking in human cancer. Science 354: 618–622.

Altemose N, Glennis A, Bzikadze AV, Sidhwani P, Langley SA, Caldas GV, Hoyt SJ, Uralsky L, Ryabov FD, Shew CJ, et al. 2022. Complete genomic and epigenetic maps of human centromeres. Science 376: eabl4178.

Aw C, S T, A B. 2021. Relationships among smoking, oxidative stress, inflammation, macromolecular damage, and cancer. Mutation research Reviews in mutation research 787. https://pubmed.ncbi.nlm.nih.gov/34083039/ (Accessed March 25, 2025).

Beroukhim R, Zhang X, Meyerson M. 2016. Copy number alterations unmasked as enhancer hijackers. Nat Genet 49: 5–6.

Bhatia R, Gautam SK, Cannon A, Thompson C, Hall BR, Aithal A, Banerjee K, Jain M, Solheim JC, Kumar S, et al. 2019. Cancer-associated mucins: role in immune modulation and metastasis. Cancer Metastasis Rev 38: 223–236.

Blanchard N, Di Bartolo V, Hivroz C. 2002. In the immune synapse, ZAP-70 controls T cell polarization and recruitment of signaling proteins but not formation of the synaptic pattern. Immunity 17: 389–399.

Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A. 2024. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 74: 229–263.

Cavaliere M, Bisogno A, Scarpa A, D'Urso A, Marra P, Colacurcio V, De Luca P, Ralli M, Cassandro E, Cassandro C. 2021. Biomarkers of laryngeal squamous cell carcinoma: a review. Ann Diagn Pathol 54: 151787.

Chen J, Huang J, Huang Q, Li J, Chen E, Xu W. 2021. RASA4 inhibits the HIFα signaling pathway to suppress proliferation of cervical cancer cells. Bioengineered 12: 10723–10733.

Chen JP, Diekmann C, Wu H, Chen C, Della Chiara G, Berrino E, Georgiadis KL, Bouwman BAM, Virdi M, Harbers L, et al. 2024. scCircle-seq unveils the diversity and complexity of extrachromosomal circular DNAs in single cells. Nat Commun 15: 1768.

Chiu R, Rajan-Babu I-S, Friedman JM, Birol I. 2021. Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. Genome Biology 22: 224.

Choo Z-N, Behr JM, Deshpande A, Hadi K, Yao X, Tian H, Takai K, Zakusilo G, Rosiene J, Da Cruz Paula A, et al. 2023. Most large structural variants in cancer genomes can be detected without long reads. Nat Genet 55: 2139–2148.

Cm L, Gp B, J C, H C, M D, Jn G, As H, Bt L, Lr N, Cc P, et al. 2020. UCSC Genome Browser enters 20th year. Nucleic acids research 48. https://pubmed.ncbi.nlm.nih.gov/31691824/ (Accessed March 30, 2025).

Cosenza MR, Rodriguez-Martin B, Korbel JO. 2022. Structural Variation in Cancer: Role, Prevalence, and Mechanisms. Annu Rev Genomics Hum Genet 23: 123–152.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. Gigascience 10: giab008.

De Coster W, Rademakers R. 2023. NanoPack2: population-scale evaluation of long-read sequencing data. Bioinformatics 39: btad311.

Degawa M, Stern SJ, Martin MV, Guengerich FP, Fu PP, Ilett KF, Kaderlik RK, Kadlubar FF. 1994. Metabolic activation and carcinogen-DNA adduct detection in human larynx. Cancer Res 54: 4915–4919.

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74.

English AC, Menon VK, Gibbs RA, Metcalf GA, Sedlazeck FJ. 2022. Truvari: refined structural variant comparison preserves allelic diversity. Genome Biol 23: 271.

Erwin GS, Gürsoy G, Al-Abri R, Suriyaprakash A, Dolzhenko E, Zhu K, Hoerner CR, White SM, Ramirez L, Vadlakonda A, et al. 2023. Recurrent repeat expansions in human cancer genomes. Nature 613: 96–102.

Espejo Valle-Inclan J, Besselink NJM, de Bruijn E, Cameron DL, Ebler J, Kutzera J, van Lieshout S, Marschall T, Nelen M, Priestley P, et al. 2022. A multi-platform reference for somatic structural variation detection. Cell Genom 2: 100139.

Gendler SJ, Spicer AP. 1995. Epithelial mucin genes. Annu Rev Physiol 57: 607–634.

Graham LM, Brown GD. 2009. The Dectin-2 family of C-type lectins in immunity and homeostasis. Cytokine 48: 148–155.

Han B, Zheng R, Zeng H, Wang S, Sun K, Chen R, Li L, Wei W, He J. 2024. Cancer incidence and mortality in China, 2022. Journal of the National Cancer Center 4: 47–53.

Hecht SS, Hatsukami DK. 2022. Smokeless tobacco and cigarette smoking: chemical mechanisms and cancer prevention. Nat Rev Cancer 22: 143–155.

Hormozdiari F, Alkan C, Ventura M, Hajirasouliha I, Malig M, Hach F, Yorukoglu D, Dao P, Bakhshi M, Sahinalp SC, et al. 2011. Alu repeat discovery and characterization within human genomes. Genome Res 21: 840–849.

Huo Y, Cao K, Kou B, Chai M, Dou S, Chen D, Shi Y, Liu X. 2023. TP53BP2: Roles in suppressing tumorigenesis and therapeutic opportunities. Genes Dis 10: 1982–1993.

Iakovlev DA, Alekseeva IV, Vorobjev YN, Kuznetsov NA, Fedorova OS. 2019. The Role of Active-Site Residues Phe98, His239, and Arg243 in DNA Binding and in the Catalysis of Human Uracil-DNA Glycosylase SMUG1. Molecules 24: 3133.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. 2020. Pan-cancer analysis of whole genomes. Nature 578: 82–93.

Je M, Mj P, He P, Cb E, N S, J A, T K, Ca D, A D, R K, et al. 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature 583. https://pubmed.ncbi.nlm.nih.gov/32728249/ (Accessed April 10, 2025).

Jethwa AR, Khariwala SS. 2017. Tobacco-related carcinogenesis in head and neck cancer. Cancer Metastasis Rev 36: 411–423.

Johnson DE, Burtness B, Leemans CR, Lui VWY, Bauman JE, Grandis JR. 2020. Head and neck squamous cell carcinoma. Nat Rev Dis Primers 6: 92.

Kashi Y, King DG. 2006. Simple sequence repeats as advantageous mutators in evolution. Trends Genet 22: 253–259.

Keskus AG, Bryant A, Ahmad T, Yoo B, Aganezov S, Goretsky A, Donmez A, Lansdon LA, Rodriguez I, Park J, et al. 2025. Severus detects somatic structural variation and complex rearrangements in cancer genomes using long-read sequencing. Nat Biotechnol. https://www.nature.com/articles/s41587-025-02618-8 (Accessed April 22, 2025).

Kirsche M, Prabhu G, Sherman R, Ni B, Battle A, Aganezov S, Schatz MC. 2023. Jasmine and Iris: population-scale structural variant comparison and analysis. Nat Methods 20: 408–417.

Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: An information aesthetic for comparative genomics. Genome Res 19: 1639–1645.

Kufe DW. 2009. Mucins in cancer: function, prognosis and therapy. Nat Rev Cancer 9: 874–885.

Lechien JR, Sadoughi B, Hans S. 2022. Laryngeal cancers in paediatric and young adult patients: epidemiology, biology and treatment. Curr Opin Otolaryngol Head Neck Surg 30: 145–153.

Lee S, Nguyen LT, Hayes BJ, Ross EM. 2021. Prowler: a novel trimming algorithm for Oxford Nanopore sequence data. Bioinformatics 37: 3936–3937.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094–3100.

Li H, Durbin R. 2024. Genome assembly in the telomere-to-telomere era. Nat Rev Genet 25: 658–670.

Lin M, Zheng X, Yan J, Huang F, Chen Y, Ding R, Wan J, Zhang L, Wang C, Pan J, et al. 2024. The RNF214-TEAD-YAP signaling axis promotes hepatocellular carcinoma progression via TEAD ubiquitylation. Nat Commun 15: 4995.

Ma B, Pl D. 2002. Alu repeats and human genomic diversity. Nature reviews Genetics 3. https://pubmed.ncbi.nlm.nih.gov/11988762/ (Accessed March 22, 2025).

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol 28: 495–501.

Merikallio H, Pincikova T, Kotortsi I, Karimi R, Li C-X, Forsslund H, Mikko M, Nyrén S, Lappi-Blanco E, Wheelock ÅM, et al. 2023. Mucins 3A and 3B Are Expressed in the Epithelium of Human Large Airway. Int J Mol Sci 24: 13546.

Nocini R, Molteni G, Mattiuzzi C, Lippi G. 2020. Updates on larynx cancer epidemiology. Chin J Cancer Res 32: 18–25.

Pan X, Che Q, Liu D, Xie Y, Li B, Zhang S, Li T, Li G, Li X, Zheng Q, et al. 2025. Development and validation of a novel endoplasmic reticulum stress-related lncRNA signature in laryngeal squamous cell carcinoma. Sci Rep 15: 12497.

Prior IA, Hood FE, Hartley JL. 2020. The Frequency of Ras Mutations in Cancer. Cancer Res 80: 2969–2974.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

R O, M R, C T. 2019. The Treatment of Laryngeal Cancer. Oral and maxillofacial surgery clinics of North America 31. https://pubmed.ncbi.nlm.nih.gov/30449522/ (Accessed March 24, 2025).

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. Nat Biotechnol 29: 24–26.

Sahlin K, Baudeau T, Cazaux B, Marchet C. 2023. A survey of mapping algorithms in the long-reads era. Genome Biology 24: 133.

Schloissnig S, Pani S, Ebler J, Hain C, Tsapalou V, Söylev A, Hüther P, Ashraf H, Prodanov T, Asparuhova M, et al. 2025. Structural variation in 1,019 diverse humans based on long-read sequencing. Nature. https://www.nature.com/articles/s41586-025-09290-7 (Accessed August 4, 2025).

Shiraishi Y, Koya J, Chiba K, Okada A, Arai Y, Saito Y, Shibata T, Kataoka K. 2023. Precise characterization of somatic complex structural variations from tumor/control paired long-read sequencing data with nanomonsv. Nucleic Acids Res 51: e74.

Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, Kalef-Ezra E, Gandhi M, Hong K, Pehlivan D, et al. 2024. Detection of mosaic and population-level structural variants with Sniffles2. Nat Biotechnol 42: 1571–1580.

Song G, Sun Z, Chu M, Zhang Z, Chen J, Wang Z, Zhu X. 2024. FBXO28 promotes cell proliferation, migration and invasion via upregulation of the TGF-beta1/SMAD2/3 signaling pathway in ovarian cancer. BMC Cancer 24: 122.

Steuer CE, El-Deiry M, Parks JR, Higgins KA, Saba NF. 2017. An update on larynx cancer. CA Cancer J Clin 67: 31–50.

Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. 2021. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 71: 209–249.

Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics Chapter 4: 4.10.1-4.10.14.

Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, et al. 2019. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res 47: D941–D947.

Torrens L, Moody S, de Carvalho AC, Kazachkova M, Abedi-Ardekani B, Cheema S, Senkin S, Cattiaux T, Cortez Cardoso Penha R, Atkins JR, et al. 2025. The complexity of tobacco smoke-induced mutagenesis in head and neck cancer. Nat Genet 57: 884–896.

van Belzen IAEM, Schönhuth A, Kemmeren P, Hehir-Kwa JY. 2021. Structural variant detection in cancer genomes: computational challenges and perspectives for precision oncology. NPJ Precis Oncol 5: 15.

Wang N, Yan H, Wu D, Zhao Z, Chen X, Long Q, Zhang C, Wang X, Deng W, Liu X. 2020. PRMT5/Wnt4 axis promotes lymph-node metastasis and proliferation of laryngeal carcinoma. Cell Death Dis 11: 864.

Wang S, Lin J, Jia P, Xu T, Li X, Liu Y, Xu D, Bush SJ, Meng D, Ye K. 2025. De novo and somatic structural variant discovery with SVision-pro. Nat Biotechnol 43: 181–185.

Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. 2021. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol 39: 1348–1365.

Witherspoon DJ, Xing J, Zhang Y, Watkins WS, Batzer MA, Jorde LB. 2010. Mobile element scanning (ME-Scan) by targeted high-throughput sequencing. BMC Genomics 11: 410.

Wu Z, Jiang Z, Li T, Xie C, Zhao L, Yang J, Ouyang S, Liu Y, Li T, Xie Z. 2021. Structural variants in the Chinese population and their impact on phenotypes, diseases and population adaptation. Nat Commun 12: 6501.

Xu L, Wang X, Lu X, Liang F, Liu Z, Zhang H, Li X, Tian S, Wang L, Wang Z. 2023. Long-read sequencing identifies novel structural variations in colorectal cancer. PLoS Genet 19: e1010514.

Yamazaki H, Inui Y, Yun CH, Guengerich FP, Shimada T. 1992. Cytochrome P450 2E1 and 2A6 enzymes as major catalysts for metabolic activation of N-nitrosodialkylamines and tobacco-related nitrosamines in human liver microsomes. Carcinogenesis 13: 1789–1794.

Yoshida K, Gowers KHC, Lee-Six H, Chandrasekharan DP, Coorens T, Maughan EF, Beal K, Menzies A, Millar FR, Anderson E, et al. 2020. Tobacco smoking and somatic mutations in human bronchial epithelium. Nature 578: 266–272.

Zeng T, Liao H, Xia L, You S, Huang Y, Zhang J, Liu Y, Liu X, Xie D. 2025. Multisite long-read sequencing reveals the early contributions of somatic structural variations to HBV-related hepatocellular carcinoma tumorigenesis. Genome Res.

Zevin S, Benowitz NL. 1999. Drug interactions with tobacco smoking. An update. Clin Pharmacokinet 36: 425–438.

Zhou Y, Hu Y, Yan X, Zheng Y, Liu S, Yao H. 2023. Smoking index and COPD duration as potential risk factors for development of osteoporosis in patients with non-small cell lung cancer – A retrospective case control study evaluated by CT Hounsfield unit. Heliyon 9: e20885.

Zhu X, Zhang Y, Wang Y, Tian D, Belmont AS, Swedlow JR, Ma J. 2022. Nucleome Browser: an integrative and multimodal data navigation platform for 4D Nucleome. Nat Methods 19: 911–913.

Fig. 1. Overview of somatic SV detection softwares evaluation. (A). Schematic of the SomaGauss-SV workflow, which consists of three main steps: (1) detection of candidate somatic SVs, screening of somatic SVs through a merge method, (2)filtering of insertions (INS) and deletions (DEL) using a Gaussian Mixture Model (GMM), and (3) filtering of SVs using data from the Chinese normal population. **(B).**Bar chart shwoing the F1 scores for different software. The height of the bars represents the average F1 score for each software, and different shapes of points indicate the F1 scores for different cell lines. **(C).**Contour plot of

software evaluation. The X-axis represents precision, the Y-axis represents recall, the contours in the plot represent F1 scores, and software and cell lines are distinguished by the color and shape of the points, respectively. **(D).**Venn diagram of software detection results in the HCC1395 cell line. Numbers indicate the number of SVs in the intersections between software, and red lines indicate the intersections involving SomaGauss-SV with other software. **(E).**UPSet plot of true positive (TP) proportions for different software in the HCC1395 cell line. Yellow indicates TP, gray indicates false positives (FP), red boxes indicate the proportion of TP unique to each software, and black circles below indicate intersections between software.

Fig. 2. Overview of pathological information and sequencing data. (A). Display of sampling sites for LSCC. There are 3 cases of supraglottic type, 11 cases of glottic type, and 1 case of subglottic type. **(B).**Pie chart of the proportion of sampling sites for LSCC. **(C).**Pie chart of the proportion of TNM staging for LSCC patients. **(D).**Pathological scatter plot for LSCC patients. The X-axis represents the smoking index of patients (calculated as: number of cigarettes per day × smoking years), the Y-axis represents patient age, and the shape of the points indicates TNM staging. **(E).**Reads length distribution plot of nanopore sequencing. The upper part shows the distribution of average read lengths (bp) for tumor (pink) and blood (blue), and the lower part shows the N50 distribution for tumor and blood. **(F).**Box plot of read Mapping rates. Blue represents blood, and pink represents tumor.

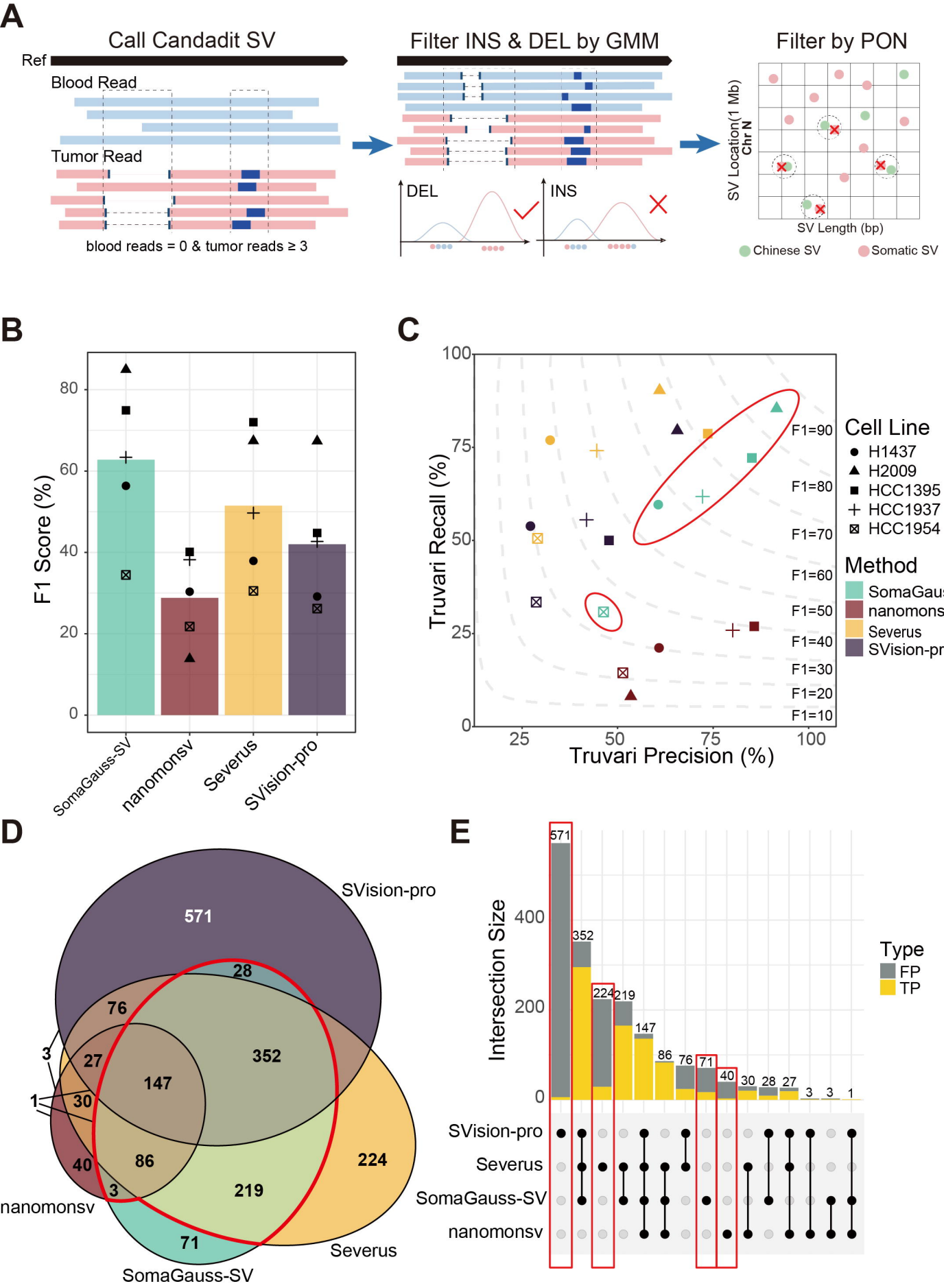
Fig. 3. Relationship between somatic DEL and tobacco exposure. (A). Stacked bar chart of somatic structural variation counts and smoking index distribution. The upper stacked bar chart shows the number of different SVs for each sample, and the lower bubble chart shows the smoking index for corresponding samples, where higher indices are represented by redder colors. (**B).**Correlation plot of somatic DEL counts and smoking index. The X-axis represents the smoking index, and the Y-axis represents the number of DEL, with degrees of freedom, *p*-value, and Pearson correlation coefficient marked on the plot. **(C).**Reactome pathway enrichment plot for genes affected by common DEL in multiple samples from smokers (q-value < 0.05). The X-axis represents $-\log_{10}(q\text{-value})$, and the Y-axis represents the names of enriched pathways, with immune-related pathways highlighted in red. The size of the circles indicates the number of enriched genes. **(D).**Browser plot of somatic DEL (Chr6:32,477,000-32,477,500) illustrating the following epigenomic signals: sample DEL location, ENCODE-annotated gene transcript, and Candidate *cis*-regulatory Elements (cCREs) predicted by ENCODE. **(E).**Browser plot of somatic DEL (Chr7:100,953,500-100,954,000) illustrating the following epigenomic signals: sample DEL location, ENCODE-annotated gene transcript, and Simple Tandem Repeats identified by TRF.

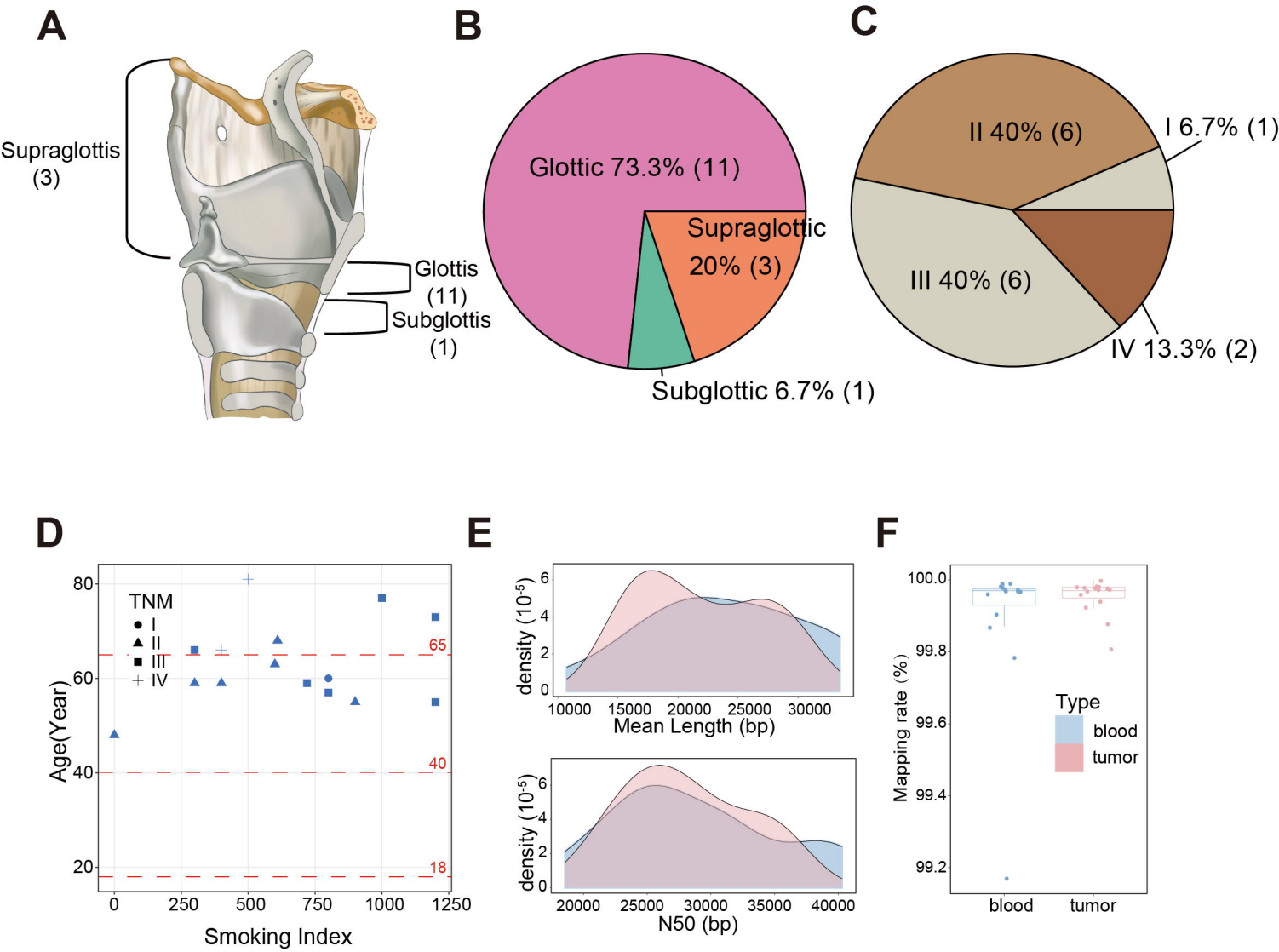
Fig. 4. Overview of SV length and site enrichment. (A). Length distribution plot for different SV types. From top to bottom are DEL, INS, DUP, and INV, with red dashed lines marking the peaks for each SV type. **(B).**Stacked bar chart of sequence annotation proportions at DEL peaks.**(C).**Stacked bar chart of sequence annotation proportions at INS peaks. **(D).**Heatmap of somatic SVs site enrichment in repeat regions. Each row represents a different SV type, and

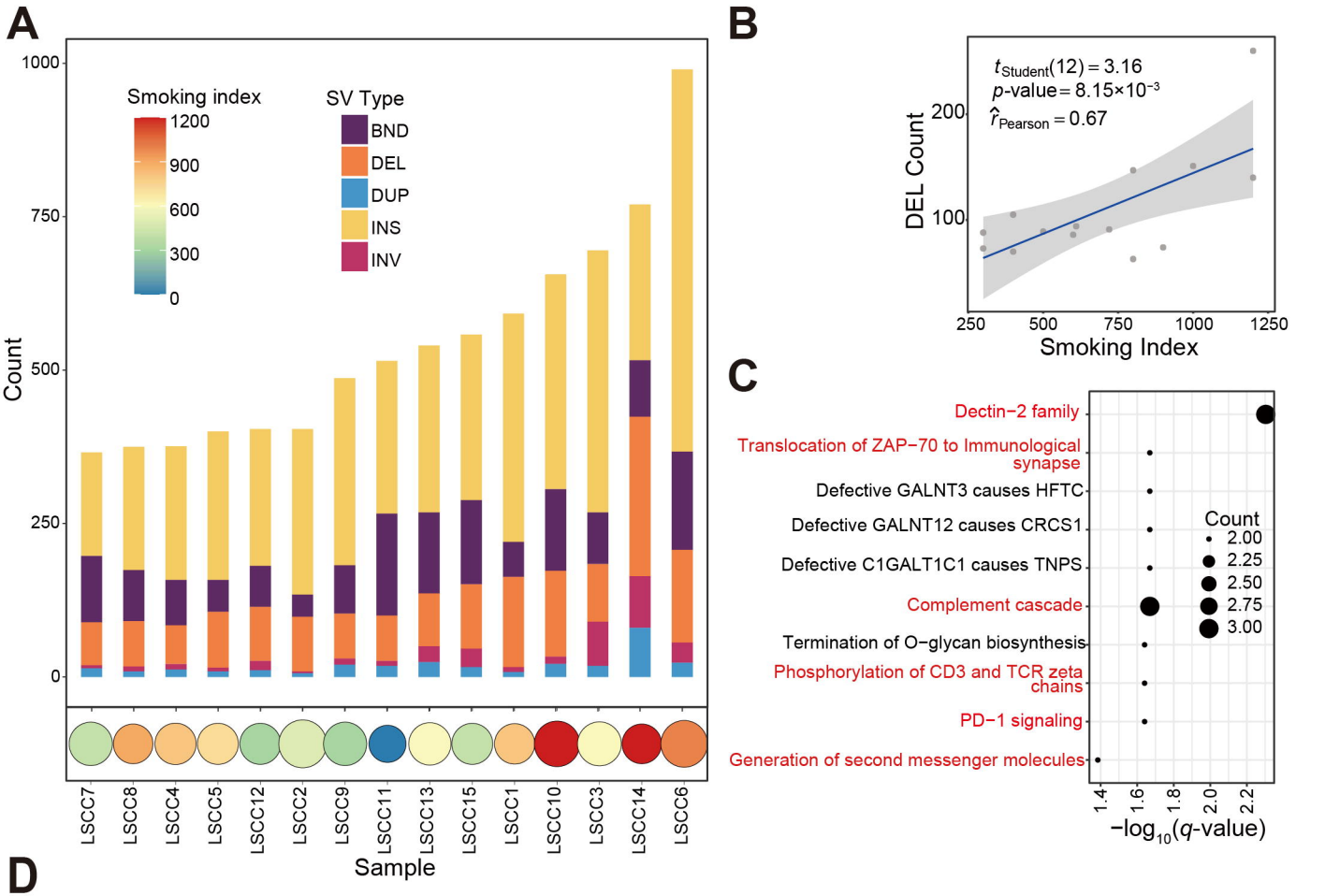
each column represents a different repeat region type, with color intensity indicating log_2(Fold Enrichment). Repeat regions are categorized into TE (transposable elements region), Repeat (repeat region), and de novo (non repititive regions). **(E).**Bar chart of somatic SVs site enrichment for gene elements. The X-axis represents log_2(Fold Enrichment), the Y-axis represents gene element names, and facets represent different SV types (Fisher's exact test, **p*-value < 0.05, ***p*-value < 0.01, ****p*-value < 0.001).

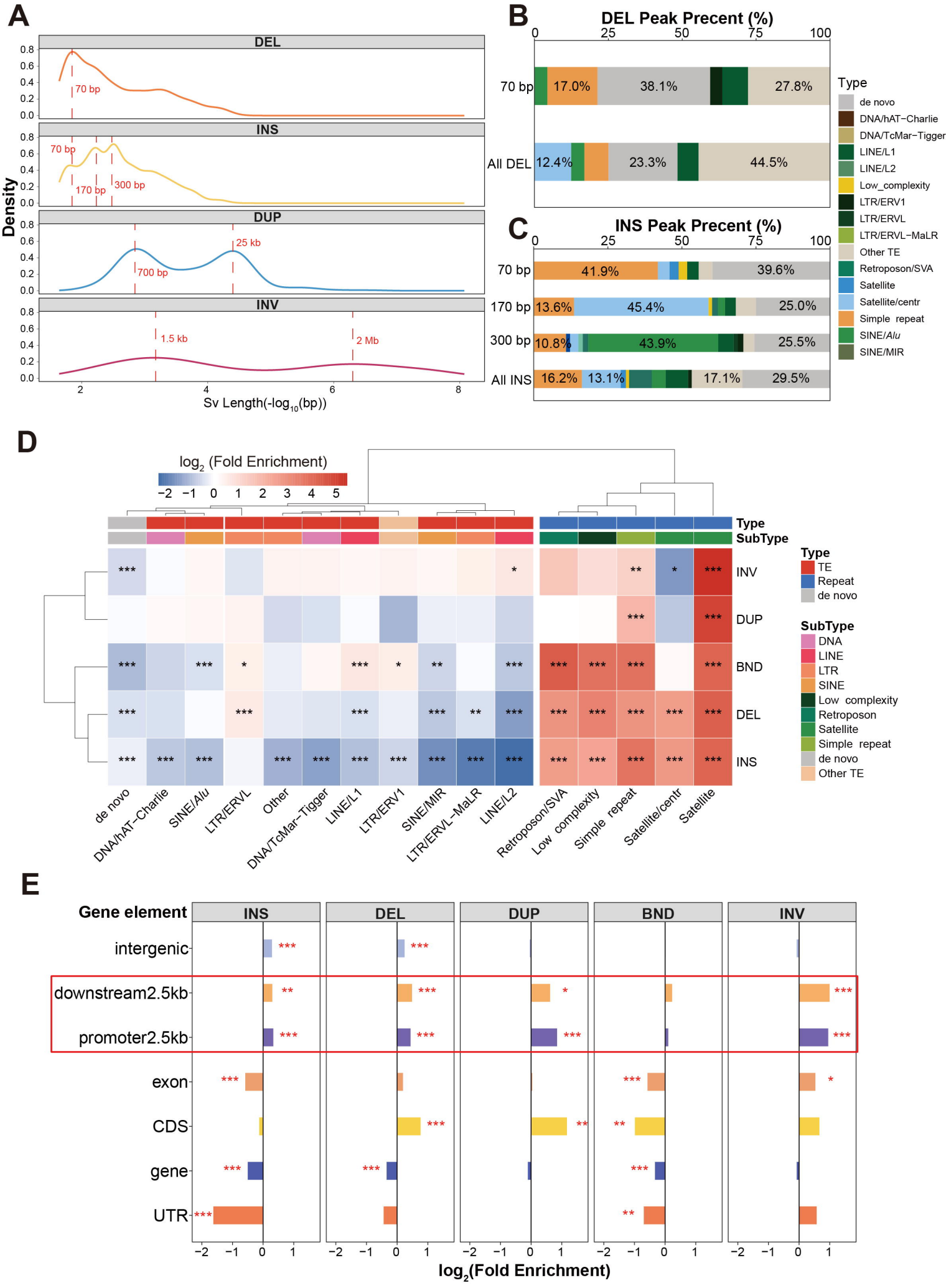
Fig. 5. Circos plot of genome-wide sample enrichment. The tracks in the figure represent the enrichment of different SVs types in samples, from the innermost to the outermost being INV, BND, DUP, INS, and DEL. The plot highlights regions with the top 2 sample counts for INS and DEL that may affect genes, the top 3 regions for INV sample counts, and regions intersecting with those reported in previous studies.

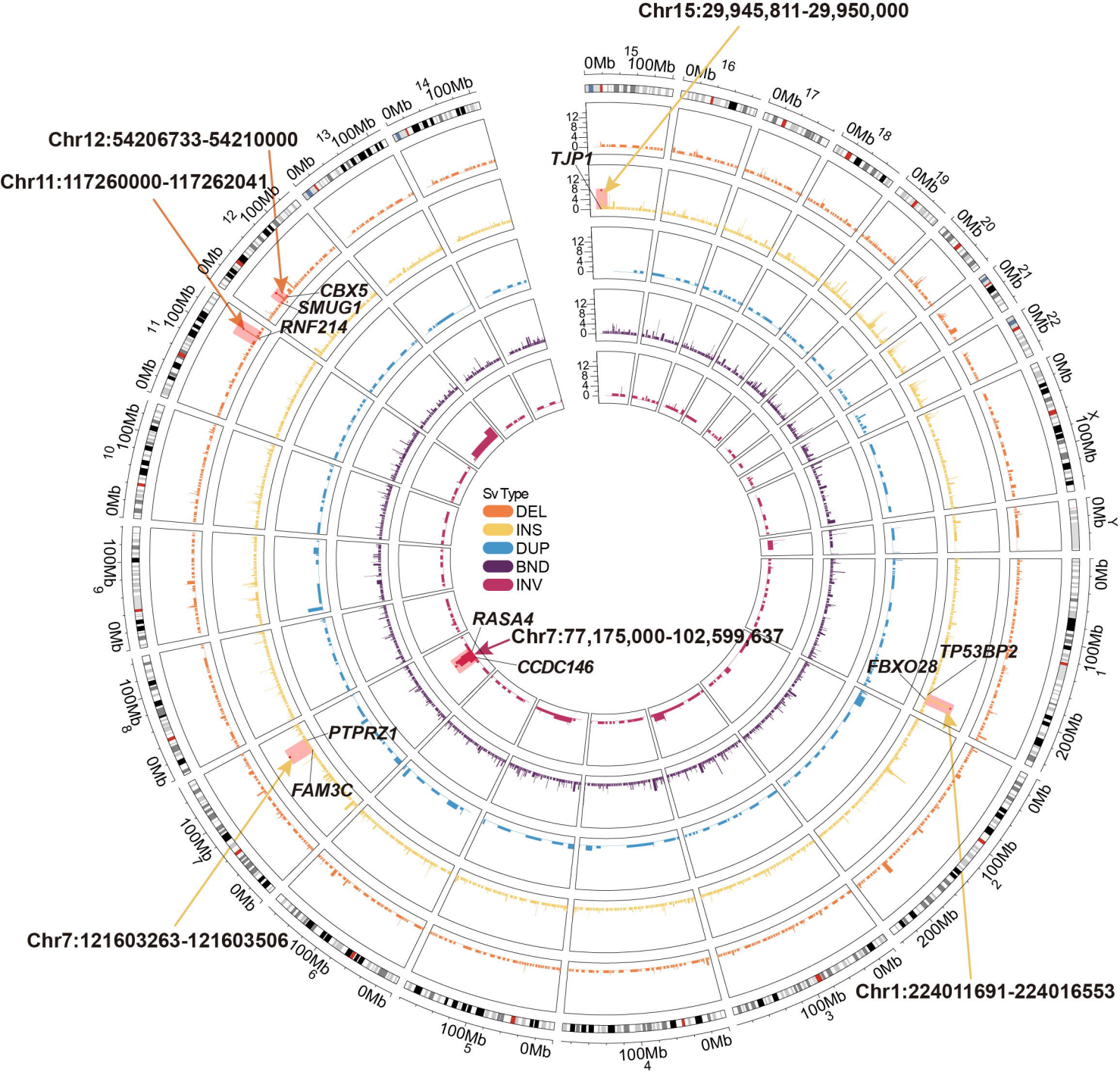
Fig. 6. Overview of somatic SRE shared by 66% of samples. (A). Oncoplot plot of INS sample enrichment regions. Rows represent samples, and columns represent regions. **(B).**IGV plot of somatic SRE (LSCC 13, Chr1:224,009,042-224,018,679). Light blue represents reads from blood samples, pink represents reads from tumor samples, and dark blue indicates the Simple repeat region with the AATGG motif. **(C).**Browser plot of somatic SRE (Chr1:224,009,042-224,018,679) illustrating the following epigenomic signals: a stacked bar chart of SRE positions and inserted sequences, with different colors representing different bases; the plot concludes with the AATGG motif of the repeat, the ENCODE-annotated gene transcript, and Candidate *cis*-regulatory Elements (cCREs) predicted by ENCODE. Also included are H3K4me1, H3K4me3, and H3K27ac marks predicted by ENCODE. In the upper left corner, the predicted 4D nucleosome structure around the LSCC-enriched SRE, *TP53BP2*, and *FBXO28* is displayed. **(D).**Horizontal gel electrophoresis images 1 showing DNA fragments from blood and tumor tissues of LSCC samples sequenced using nanopore technology. Each sample's blood and tumor are labeled separately, and green dots and red lines mark the longer bands in the tumor relative to the blood. **(E).**Horizontal gel electrophoresis images 2 showing DNA fragments from blood and tumor tissues of LSCC samples sequenced using nanopore technology. Each sample's blood and tumor are labeled separately, and green dots and red lines mark the longer bands in the tumor relative to the blood. **(F).**Relative expression levels of *TP53BP2* and *FBXO28* determined by QCPCR in SRE and No SRE Clinical Samples (*t*-test). **(G).**Relative expression levels of TP53BP2 and FBXO28 determined by QCPCR in BEAS 2B and SNU899 cell lines (*t*-test, **p*-value < 0.05, ***p*-value < 0.01, ****p*-value < 0.001).

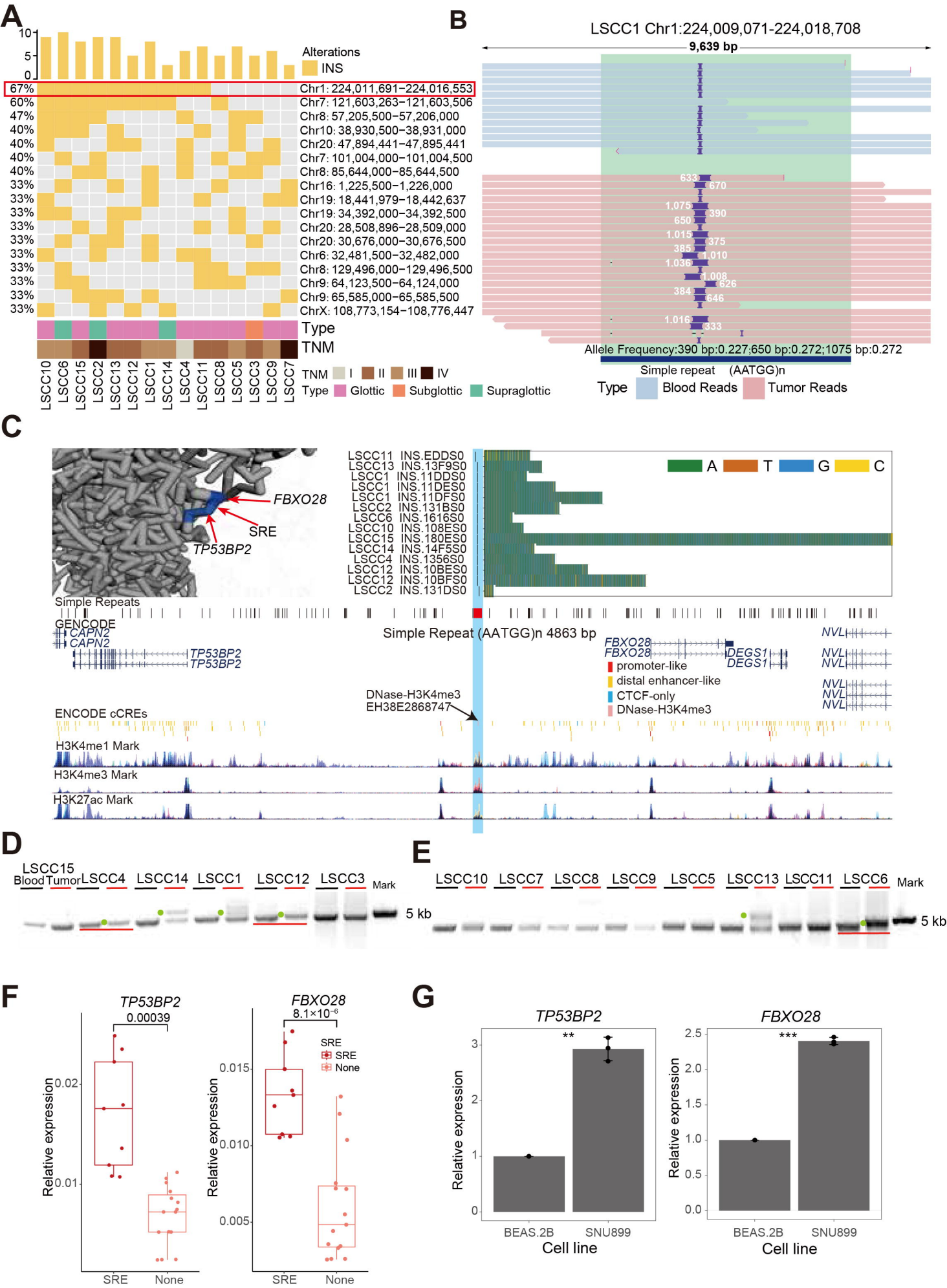

A
Call Candadit SV
Ref
Blood Read
Tumor Read
blood reads = 0 & tumor reads ≥ 3
Filter INS & DEL by GMM
DEL
INS
Filter by PON
SV Location(1 Mb)
Chr N
SV Length (bp)
Chinese SV
Somatic SV
B
F1 Score (%)
SomaGauss-SV
nanomonsv
Severus
SVision-pro
C
Truvari Recall (%)
Truvari Precision (%)
F1=90
F1=80
F1=70
F1=60
F1=50
F1=40
F1=30
F1=20
F1=10
Cell Line
H1437
H2009
HCC1395
HCC1937
HCC1954
Method
SomaGauss-SV
nanomonsv
Severus
SVision-pro
D
SVision-pro
571
28
76
3
27
352
147
1
30
40
86
224
3
219
nanomonsv
71
Severus
SomaGauss-SV
E
Intersection Size
571
352
224
219
147
86
76
71
40
30
28
27
3
3
1
SVision-pro
Severus
SomaGauss-SV
nanomonsv
Type
FP
TP


A
Supraglottis
(3)
Glottis
(11)
Subglottis
(1)
B
Glottic 73.3% (11)
Supraglottic
20% (3)
Subglottic 6.7% (1)
C
II 40% (6)
I 6.7% (1)
III 40% (6)
IV 13.3% (2)
D
TNM
I
II
III
IV
65
40
18
Age(Year)
Smoking Index
E
density (10^{-5})
Mean Length (bp)
density (10^{-5})
N50 (bp)
F
Mapping rate (%)
Type
blood
tumor
blood
tumor


A
Smoking index
1200
900
600
300
0
SV Type
BND
DEL
DUP
INS
INV
Count
1000
750
500
250
0
LSCC7
LSCC8
LSCC4
LSCC5
LSCC12
LSCC2
LSCC9
LSCC11
LSCC13
LSCC15
LSCC1
LSCC10
LSCC3
LSCC14
LSCC6
Sample
B
$t_{Student}(12) = 3.16$
$p\text{-value} = 8.15\times10^{-3}$
$\hat{r}_{Pearson} = 0.67$
DEL Count
200
100
250
500
750
1000
1250
Smoking Index
C
Dectin−2 family
Translocation of ZAP−70 to Immunological synapse
Defective GALNT3 causes HFTC
Defective GALNT12 causes CRCS1
Defective C1GALT1C1 causes TNPS
Complement cascade
Termination of O−glycan biosynthesis
Phosphorylation of CD3 and TCR zeta chains
PD−1 signaling
Generation of second messenger molecules
Count
2.00
2.25
2.50
2.75
3.00
1.4
1.6
1.8
2.0
2.2
$-\log_{10}(q\text{-value})$
D
DEL Chr6:32,477,000-32,477,500
LSCC1 Sniffles2.DEL.1B8DS5
LSCC11 Sniffles2.DEL.FAAS5
LSCC2 Sniffles2.DEL.1E54S5
LSCC4 Sniffles2.DEL.1C36S5
LSCC1 Sniffles2.DEL.1B95S5
LSCC8 Sniffles2.DEL.2E59S5
LSCC5 Sniffles2.DEL.2BB3S5
Basic Gene Annotation Set from GENCODE Version 42
HLA-DRA
HLA-DRA
HLA-DRA
ENCODE cCREs
CTCF
HLA-DRB5
HLA-DRB5
E
DEL Chr7:100,953,500-100,954,000
LSCC4 Sniffles2.DEL.3570
LSCC5 Sniffles2.DEL.5C30
LSCC11 Sniffles2.DEL.17E2
LSCC8 Sniffles2.DEL.5D89
Basic Gene Annotation Set from GENCODE Version 42
MUC3A
MUC3A
MUC3A
ENSG00000274993
MUC12
MUC12
MUC12
Simple Tandem Repeats by TRF
Simple Tandem Repeat


A
DEL
INS
DUP
INV
70 bp
170 bp
300 bp
700 bp
25 kb
1.5 kb
2 Mb
Density
Sv Length(-log10(bp))
B
DEL Peak Precent (%)
70 bp
All DEL
17.0%
38.1%
27.8%
12.4%
23.3%
44.5%
C
INS Peak Precent (%)
70 bp
170 bp
300 bp
All INS
41.9%
39.6%
13.6%
45.4%
25.0%
10.8%
43.9%
25.5%
16.2%
13.1%
17.1%
29.5%
Type
de novo
DNA/hAT−Charlie
DNA/TcMar−Tigger
LINE/L1
LINE/L2
Low_complexity
LTR/ERV1
LTR/ERVL
LTR/ERVL−MaLR
Other TE
Retroposon/SVA
Satellite
Satellite/centr
Simple repeat
SINE/Alu
SINE/MIR
D
log2 (Fold Enrichment)
Type
SubType
INV
DUP
BND
DEL
INS
de novo
DNA/hAT−Charlie
SINE/Alu
LTR/ERVL
Other
DNA/TcMar−Tigger
LINE/L1
LTR/ERV1
SINE/MIR
LTR/ERVL−MaLR
LINE/L2
Retroposon/SVA
Low complexity
Simple repeat
Satellite/centr
Satellite
TE
Repeat
DNA
LINE
LTR
SINE
Retroposon
E
Gene element
INS
DEL
DUP
BND
INV
intergenic
downstream2.5kb
promoter2.5kb
exon
CDS
gene
UTR
log2(Fold Enrichment)


Chr15:29,945,811-29,950,000
Chr12:54206733-54210000
Chr11:117260000-117262041
TJP1
CBX5
SMUG1
RNF214
Sv Type
DEL
INS
DUP
BND
INV
RASA4
Chr7:77,175,000-102,599,637
CCDC146
FBXO28
TP53BP2
PTPRZ1
FAM3C
Chr7:121603263-121603506
Chr1:224011691-224016553


A
Alterations
INS
67% Chr1: 224,011,691−224,016,553
60% Chr7: 121,603,263−121,603,506
47% Chr8: 57,205,500−57,206,000
40% Chr10: 38,930,500−38,931,000
40% Chr20: 47,894,441−47,895,441
40% Chr7: 101,004,000−101,004,500
40% Chr8: 85,644,000−85,644,500
33% Chr16: 1,225,500−1,226,000
33% Chr19: 18,441,979−18,442,637
33% Chr19: 34,392,000−34,392,500
33% Chr20: 28,508,896−28,509,000
33% Chr20: 30,676,000−30,676,500
33% Chr6: 32,481,500−32,482,000
33% Chr8: 129,496,000−129,496,500
33% Chr9: 64,123,500−64,124,000
33% Chr9: 65,585,000−65,585,500
33% ChrX: 108,773,154−108,776,447
Type
TNM
LSCC10 LSCC6 LSCC15 LSCC2 LSCC13 LSCC12 LSCC1 LSCC14 LSCC4 LSCC11 LSCC8 LSCC5 LSCC3 LSCC9 LSCC7
TNM I II III IV
Type Glottic Subglottic Supraglottic
B
LSCC1 Chr1:224,009,071-224,018,708
9,639 bp
Allele Frequency;390 bp:0.227;650 bp:0.272;1075 bp:0.272
Simple repeat (AATGG)n
Type Blood Reads Tumor Reads
C
FBXO28
SRE
TP53BP2
LSCC11 INS.EDDS0
LSCC13 INS.13F9S0
LSCC1 INS.11DDS0
LSCC1 INS.11DES0
LSCC1 INS.11DFS0
LSCC2 INS.131BS0
LSCC6 INS.1616S0
LSCC10 INS.108ES0
LSCC15 INS.180ES0
LSCC14 INS.14F5S0
LSCC4 INS.1356S0
LSCC12 INS.10BES0
LSCC12 INS.10BFS0
LSCC2 INS.131DS0
A T G C
Simple Repeats
GENCODE
CAPN2
TP53BP2
Simple Repeat (AATGG)n 4863 bp
FBXO28
DEGS1
NVL
promoter-like
distal enhancer-like
CTCF-only
DNase-H3K4me3
DNase-H3K4me3
EH38E2868747
ENCODE cCREs
H3K4me1 Mark
H3K4me3 Mark
H3K27ac Mark
D
LSCC15
Blood Tumor LSCC4 LSCC14 LSCC1 LSCC12 LSCC3 Mark
5 kb
E
LSCC10 LSCC7 LSCC8 LSCC9 LSCC5 LSCC13 LSCC11 LSCC6 Mark
5 kb
F
TP53BP2
0.00039
FBXO28
8.1×10⁻⁶
SRE
SRE
None
Relative expression
SRE None
G
TP53BP2
FBXO28
**
***
Relative expression
BEAS.2B SNU899
Cell line