J Genet Resour2026; 12(1): 64-73 *Homepage: http://sc.journals.umz.ac.ir/*

RESEARCH ARTICLE *DOI: 10.22080/jgr.2026.30992.1455*

High-confidence Mapping and Clustered Patterns of LncRNA-associated SNP Reveal Recurrent Non-coding Variants in Colorectal Cancer Transcriptomes

Maryam Esmaeili and Malek Hossein Asadi*

Department of Biotechnology, Institute of Science and High Technology and Environmental Sciences, Graduate University of Advanced Technology, Kerman, Iran

ARTICLEINFO

Article history:
Received 26 November 2025
Accepted 28 December 2025
Available 21 January 2026

Keywords:
Colorectal cancer
Long non-coding RNA
Non-coding variants
RNA sequencing

Corresponding authors:
✉ MH. Asadi
mh.asadi@kgut.ac.ir

p-ISSN 2423-4257
e-ISSN 2588-2589

ABSTRACT

Long non-coding RNAs (lncRNAs) are important regulators of gene expression and cellular homeostasis, yet sequence-level variation within these transcripts remains poorly characterized in colorectal cancer (CRC). Although evidence suggests that lncRNA-embedded variants influence transcriptional regulation and disease progression, the exact mechanistic function of these variants is not well understood. Publicly available CRC transcriptome sequencing libraries were systematically obtained and subjected to rigorous quality control to ensure high analytical reliability. Filtered libraries were aligned to a curated reference set of lncRNA transcripts derived from the RefSeq human genome assembly. Variant detection was carried out using RNA-seq-based variant-calling approaches, followed by stringent filtering based on alternate read depth and genotype confidence to minimize false-positive calls. Recurrence-based filtering, combined with hierarchical clustering of transcriptomes using shared variant profiles, was applied to identify reproducible patterns of sequence variation across independent datasets. Initial variant detection identified more than one hundred thousand candidate single nucleotide variants within lncRNA transcripts. Application of depth-, confidence-, and recurrence-based filtering substantially reduced this set to a focused catalogue of highly recurrent variants shared across multiple transcriptome clusters. In total, 549 high-confidence single nucleotide variants distributed across 99 lncRNAs were identified. Several of the transcript-embedded variants have been previously implicated in CRC biology, supporting the biological relevance of the detected variants. These findings demonstrated that expressed lncRNA-associated SNPs are non-randomly distributed in CRC transcriptomes and potentially accumulate in non-coding variant hotspots. This study provides a scalable framework for prioritizing recurrent lncRNA-associated sequence variants and offers a resource for future functional validation, biomarker-oriented analyses, and precision oncology studies in CRC.

Please cite this paper as: Esmaeili, M., & Asadi, M.H. (2026). High-confidence mapping and clustered patterns of LncRNA-associated SNPs reveal recurrent non-coding variants in colorectal cancer transcriptomes. *Journal of Genetic Resources,* 12(1), 64-73. doi: 10.22080/jgr.2026.30992.1455

Introduction

Genetic variation in non-coding loci, particularly single nucleotide polymorphisms (SNPs) within lncRNA loci, is increasingly recognized key source of cancer heterogeneity through its strong impact on transcriptional regulation, chromatin dynamics, and oncogenic signaling pathways (Du *et al.,* 2020; Zhou *et al.,* 2023). LncRNAs are transcripts longer than 200 nucleotides that lack protein-coding capacity but play essential regulatory roles at transcriptional, post-transcriptional, and epigenetic levels.

Dysregulation or sequence variation within lncRNAs can modify RNA structure, RNA stability, and regulatory interactions, thereby affecting gene regulatory networks contributing to tumor initiation, progression, metastasis, and therapy resistance (Bian *et al.,* 2018; Liu *et al.,* 2019; Hennig *et al.,* 2021). CRC was selected for this analysis because of its high global incidence, status as the third most commonly diagnosed cancer worldwide, pronounced molecular heterogeneity, and the availability of large-scale public RNA-seq datasets that enable systematic interrogation of non-coding genetic variation. Although protein-coding alterations in CRC have been greatly characterized, the distribution and frequency of lncRNA-associated SNPs across CRC transcriptomes remain poorly understood (Yan *et al.,* 2021). Most previous studies have focused on DNA-based variation or differential lncRNA expression, leaving RNA-seq–based discovery of expressed lncRNA-associated variants relatively underexplored (Kerachian *et al.,* 2022; Andrabi *et al.,* 2024). Recent developments in bioinformatics have made it feasible to perform variant calling directly from RNA-seq data, enabling the detection of expressed SNPs in transcribed regions, including lncRNAs (Wu *et al.,* 2023; Chen *et al.,* 2025). This approach provides a functional layer of evidence by identifying variants that are not only present at the sequence level but also expressed in cancer transcriptomes (Lulli *et al.,* 2022). Nevertheless, distinguishing high-confidence lncRNA-associated SNPs from sequencing artefacts, RNA-editing events, and sample redundancy remains a major challenge, particularly when integrating large public datasets from heterogeneous experimental sources (Liu *et al.,* 2025). Therefore, stringent filtering based on read depth, genotype likelihood, recurrence, and clustering is required to improve the reliability of RNA-seq-derived variant catalogues. The biological relevance of lncRNA-associated polymorphisms is increasingly supported by studies showing that variants in non-coding RNA loci can alter chromatin accessibility, RNA secondary structure, RNA-protein interactions, and transcriptional output (Minotti *et al.,* 2018; He *et al.,* 2019; Zhang *et al.,* 2021). In CRC, lncRNAs contribute to key tumor-related processes, including Wnt/β-catenin signaling, epithelial-mesenchymal transition, immune regulation, chemoresistance, and metastatic progression (Tang *et al.,* 2021; Tufail, 2023). However, recurrent sequence-level variation within CRC-associated lncRNAs has not been systematically characterized at the transcriptome level.

We hypothesized that recurrent, expression-supported SNPs in lncRNAs are non-randomly distributed across CRC transcriptomes and preferentially accumulate in a subset of transcripts representing potential non-coding variant hotspots. To test this hypothesis, we developed a lncRNA-focused RNA-seq variant-calling pipeline using public CRC datasets, stringent depth- and likelihood-based filtering, transcript-to-genome coordinate mapping, recurrence analysis, and hierarchical clustering. This strategy identified 549 high-confidence recurrent SNPs distributed across 99 lncRNAs, providing a reproducible framework for prioritizing lncRNA-associated sequence variants in CRC.

Materials and Methods

Data acquisition and reference preparation

Human lncRNA sequences were obtained from the NCBI RefSeq assembly GRCh38.p14 (accession GCF_000001405.40). Files were downloaded from the genome FTP server and checked using the provided MD5 checksum. The rna.fna FASTA file, containing all RefSeq RNA transcripts (NR_ and XR_), was used as a reference for alignment, and the genomic annotation file GCF_000001405.40_GRCh38.p14_genomic.gff provided exon/transcript coordinates for later transcript-to-genome conversion. To build a lncRNA-specific reference, only NR_/XR_ transcripts were kept; for multi-isoform genes, the longest isoform was retained. The final reference (final_unique_longest_lncRNA.fna) comprised ~21,300 unique lncRNA transcripts and was used for all downstream analyses.

RNA-seq processing and alignment

RNA-seq data from CRC were obtained from the NCBI SRA (335 accessions; tumor tissues or CRC cell lines) using the SRA Toolkit (v3.0.3) and converted to FASTQ. Read quality (base scores, GC content, duplication, adapters) was assessed with FastQC (v0.12.1).

A Bowtie2 index was built from final_unique_longest_lncRNA.fna, and reads were aligned in sensitive local mode using Bowtie2 (v2.5.1). Sorted, indexed BAM (Binary Alignment/Map) files, were generated with SAMtools (v1.19). Libraries with <80% alignment were discarded; the remaining libraries showed a mean mapping rate of 86.7%. Software versions and key parameters were recorded, and temporary files were removed after BAM generation.

Table 1. Components of the RefSeq GRCh38.p14 ncRNA dataset used in this study.

File name	Description	Purpose in this study
RNA.fna	FASTA-formatted sequences of all annotated RefSeq RNA transcripts (including *NR_* and *XR_* prefixes).	Used as reference for Bowtie2 indexing and alignment.
data_report.jsonl	Metadata file listing transcript accessions, species, annotation source, and release version.	Used for verifying transcript origin and completeness.
dataset_catalog.json	File list and content summary with file types and sizes.	Quality control and reference tracking.
README.md	Dataset documentation describing annotation origin and usage.	Reference for citation and reproducibility.
md5sum.txt	MD5 hash codes for all downloaded files.	Verification of data integrity.
GCF_000001405.40_GRCh38.p14_genomic.gff	GFF3 file containing exon and transcript coordinates mapped to the genome.	Used for transcript-to-genome SNP mapping in later steps.

ncRNA = non-coding RNA; GFF = General Feature Format; MD5 = Message-Digest Algorithm 5; NR_ = RefSeq non-coding transcript prefix; XR_ = predicted non-coding transcript prefix.

Variant calling and filtering of lncRNA-associated SNPs

Expressed SNPs within lncRNA transcripts were called from each BAM using SAMtools/BCFtools (v1.19), restricting variants to SNPs and excluding indels. Individual VCFs (Variant Call Format files) were merged into a single dataset including sample ID, lncRNA transcript, position, alleles, quality metrics, and depth. Library-level QC (raw SNP counts, depth, genotype quality, Ts/Tv ratio) was used to remove outlier libraries.

High-confidence SNP filtering: Filtering was performed in R (v4.3.1) with vcfR and dplyr. From genotype fields, we extracted alternate allele depth (Alt_DP) and Phred-scaled genotype likelihoods (PL, converted to genotype probabilities). SNPs were retained only if Alt_DP > 50 and the probability of the best non-reference genotype was $< 1\times10^{-5}$. Remaining high-confidence variants were stored in Very_Strong_SNPs_All.csv with full positional and sample metadata.

Descriptive summaries: Distributions of Alt_DP, genotype likelihoods (probabilities supporting each possible genotype), and Alt_DP–probability relationships were summarised in R using ggplot2 (v3.4.2) to evaluate the performance of the filtering strategy.

Quantification of high-confidence SNPs per RNA-seq library

Per-sample SNP burden was calculated from Very_Strong_SNPs_All.csv using dplyr by grouping variants by library ID and counting retained SNPs. The resulting table (SNP_Count_Per_Sample.csv) was used to compare mutation load across CRC libraries and as input for downstream clustering.

Sample metadata mining and group classification

For all SRA accessions, metadata were retrieved from NCBI SRA using rentrez (v1.2.3) in R. XML records were searched for tumor-related terms (“tumor”, “cancer”, “carcinoma”, “adenocarcinoma” and CRC cell-line names such as HCT116, Caco-2, SW480, LoVo, HT29, DLD-1) and for normal/control terms (“normal”, “healthy”, “control”, “non-tumor”, “adjacent”). Libraries were classified as tumor, normal or unknown based on keyword matches. The classification table (sample_groups.txt) was used to interpret clustering results and lncRNA-level SNP patterns.

Transcript-to-genome coordinate conversion

Transcript-relative SNP positions were converted to genomic coordinates using the RefSeq GRCh38.p14 GFF3 annotation in R with rtracklayer, GenomicRanges, and GenomicFeatures. Exon models were built per transcript and used to map each SNP from transcript coordinate to genomic chromosome, 1-based position, and strand. These genomic

coordinates were merged with existing SNP metadata and exported as all_mapped_snps_to_genome.csv for genome-based analyses.

SNP presence–absence matrix construction and recurrence filtering

A binary SNP–library matrix was built in R, with rows representing unique SNPs and columns representing libraries (1 = presence, 0 = absence). Library counts per SNP were obtained from row sums. To focus on recurrent variants, only SNPs present in ≥ 10 libraries were retained, generating the filtered matrix and summary statistics. Recurrent SNPs were annotated using a curated lncRNA metadata object, producing an annotated SNP dataset. Global sharing patterns were visualized as heatmaps using pheatmap.

Mapping SNPs to lncRNA genes and lncRNA-level summarization

Genome-mapped SNPs (SNPs_mapped_to_genes.csv) were linked to lncRNA gene symbols by merging transcript IDs with annotations extracted from final_unique_longest_lncRNA.fna headers. The combined table (SNPs_with_lncRNA_names.csv) included transcript ID, lncRNA gene name, genomic coordinate, alleles, and sample ID. SNP counts per lncRNA were computed and ranked in R, yielding lncRNA_SNP_counts.csv, which was used to describe the distribution of recurrent SNPs across lncRNAs and to generate bar plots (via ggplot2) of the most strongly mutated genes.

Results

Characteristics of the lncRNA Reference Dataset

The reference used in this study was constructed from the NCBI RefSeq genome assembly GRCh38.p14 (accession GCF 000001405.40). The downloaded archive contained 54,932 RNA records, including both coding and non-coding transcripts. Filtering for entries annotated with the NR_ or XR_ prefixes yielded 23,488 lncRNA transcripts. To minimize isoform redundancy and provide a stable alignment target, only the longest transcript per lncRNA gene was retained, resulting in a non-redundant reference set comprising approximately 21,300 unique lncRNA sequences. In this final dataset, the mean transcript length was 2,156 nucleotides, ranging from 212 nt to 12,870 nt. The reference included both intergenic and intronic lncRNAs distributed across all autosomes and sex chromosomes, consistent with the GRCh38.p14 RefSeq annotation. Integrity checks based on md5sum verification showed a 100% checksum match, confirming that no data corruption occurred during data transfer. Together with the associated metadata and genomic GFF3 annotation, this curated reference provided a robust framework for RNA-seq alignment and downstream SNP analysis.

RNA-seq data processing and alignment performance

A total of 335 CRC RNA-seq libraries were retrieved from the NCBI SRA and processed successfully. Quality assessment indicated high read quality and minimal technical artefacts. Reads were aligned to the curated set of 21,300 non-redundant lncRNA transcripts using Bowtie2, achieving a mean mapping rate of 86.7%. Libraries with alignment rates below 80% were excluded. The remaining datasets were converted into sorted and indexed BAM files for downstream variant calling and allele-depth analysis.

High-confidence Dataset of lncRNA-associated SNPs

Application of the lncRNA-focused variant-calling pipeline to CRC RNA-seq data initially identified a large set of candidate SNPs across RefSeq lncRNA transcripts. Following stringent filtering based on alternate read depth (Alt_DP > 50) and genotype likelihood (PL_Prob < 1×10^{-5}), a high-confidence catalogue of lncRNA-associated SNPs was generated (Very_Strong_SNPs_All). These variants were supported by strong read depth and high genotype confidence, providing a reliable basis for recurrence-based analyses.

Distribution and quality assessment of high-confidence lncRNA-associated SNPs across CRC libraries

Application of stringent depth and genotype-likelihood filters (Alt_DP > 50; PL_Prob < 1×10^{-5}) yielded a catalogue of 159,410 high-confidence lncRNA-associated SNPs across 137

CRC RNA-seq libraries (Very_Strong_SNPs_All). These variants formed the basis for subsequent recurrence and clustering analyses. Sequential recurrence filtering further reduced this set to 1,477 SNPs detected in at least 10 independent libraries, thereby enriching for consistently observed variants and minimizing the contribution of low-frequency or sample-specific events. Substantial heterogeneity in SNP burden was observed across libraries, reflecting both biological variability and dataset heterogeneity. The retained variants exhibited strong alternate read support, with most SNPs supported by approximately 60–120 alternate reads (Figure 1B). In addition, genotype likelihood analysis confirmed high confidence, as the majority of variants displayed PL_Prob values well below the defined threshold (Figure 1A). Analysis of nucleotide substitution patterns revealed a predominance of transition mutations, particularly A→G, G→A, and C→T changes (Figure 1C), consistent with established mutational signatures in CRC. Collectively, these results demonstrate that the applied filtering strategy generates a statistically robust and biologically consistent catalogue of lncRNA-associated SNPs, suitable for downstream recurrence-based clustering and hotspot identification.

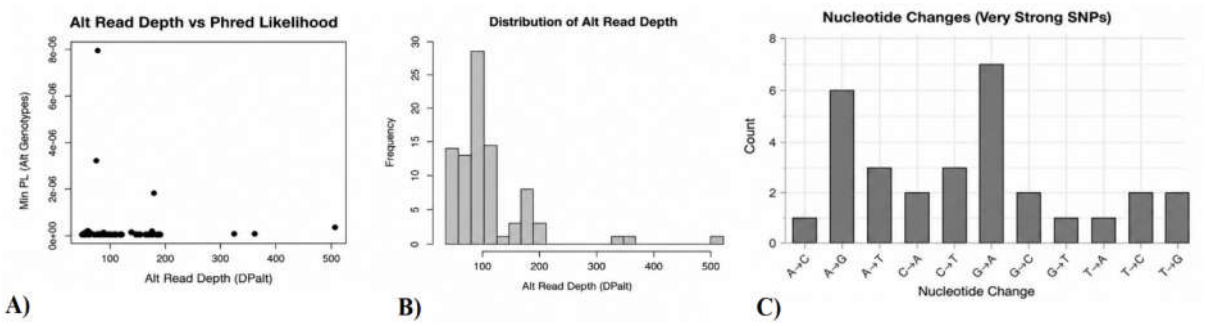

Fig. 1. Quality assessment of high-confidence lncRNA-associated SNPs in CRC RNA-seq data: A) Relationship between Alt_DP and genotype likelihood probability (PL_Prob; minimum across alternate genotypes), demonstrating that SNPs with higher alternate read depth consistently display very low PL_Prob values, confirming the stringency and robustness of the filtering strategy.
B) Distribution of alternate read depth (Alt_DP) for all retained variants, showing that most SNPs are supported by >50–100 alternate reads, with a small tail of very highly covered sites.
C) Spectrum of nucleotide substitutions among "high-confidence" SNPs, highlighting the predominance of transition events (A→G, G→A and C→T) over transversions.

Genomic position of lncRNA-associated SNPs

All high-confidence lncRNA-associated SNPs were projected from transcript-relative to genomic coordinates using the RefSeq GRCh38.p14 GFF3 annotation. For the majority of variants, a unique genomic locus was successfully assigned based on exon structure, while SNPs lacking valid transcript models or exhibiting ambiguous mappings were excluded. The resulting genome-based catalogue (*all_mapped_snps_to_genome.csv)* provides standardized chromosomal coordinates, strand orientation, and transcript identifiers for each variant. This unified coordinate framework enables integration with external genomic resources and facilitates downstream analyses, including recurrence assessment, clustering, and potential overlap with regulatory elements. Importantly, the ability to map transcript-derived variants to genomic loci provides a critical bridge between RNA-level variation and genome-level analysis.

Recurrent lncRNA-associated SNP patterns throughout colorectal cancer libraries

Construction of the SNP–library presence/absence matrix identified 159,410 lncRNA-associated SNPs across 137 CRC RNA-seq libraries. Application of a recurrence threshold (≥10 libraries) reduced this set to 1,477 recurrent SNPs, enriching for consistently observed variants across independent transcriptomes. Hierarchical clustering based on

Jaccard distance grouped libraries into 36 clusters characterized by distinct SNP-sharing profiles. Notably, clusters 6, 4, and 30 exhibited the highest enrichment of recurrent variants, whereas several clusters showed minimal or no SNP accumulation, indicating substantial heterogeneity in lncRNA-associated mutation patterns.

To further improve strong candidates, an additional cluster-level filter retaining SNPs present in at least five clusters yielded a compact set of 549 highly recurrent variants. This multi-level filtering strategy (library-level recurrence and cluster-level consistency) substantially reduces the likelihood of random or sample-specific artefacts and highlights variants reproducibly observed across independent biological contexts. Heatmap and bar plot visualizations (Figure 2) revealed a highly non-uniform distribution of SNP sharing, with a limited subset of clusters contributing disproportionately to the overall mutation burden. This pattern suggests the presence of structured, non-random variation in lncRNA-associated SNP profiles across CRC transcriptomes. Collectively, these findings indicate that recurrent lncRNA-associated SNPs are not randomly distributed but instead concentrate within specific transcriptomic contexts, supporting the existence of non-coding mutational hotspots and providing strong candidates for downstream functional studies and biomarker-oriented analyses.

Clustering and hotspot distribution of recurrent lncRNA-associated SNPs

To further explore the distribution patterns of recurrent variants, we examined their localization across lncRNA transcripts. SNPs were not uniformly distributed, but instead showed clear clustering within specific lncRNAs. Several transcripts, including *LINC00342, LOC100996740*, and *WAC-AS1*, harbored multiple SNPs within relatively short transcript regions, suggesting the presence of localized sequence variation hotspots.

This non-random distribution supports the hypothesis that recurrent lncRNA-associated SNPs are structured and biologically relevant rather than randomly dispersed across the transcriptome.

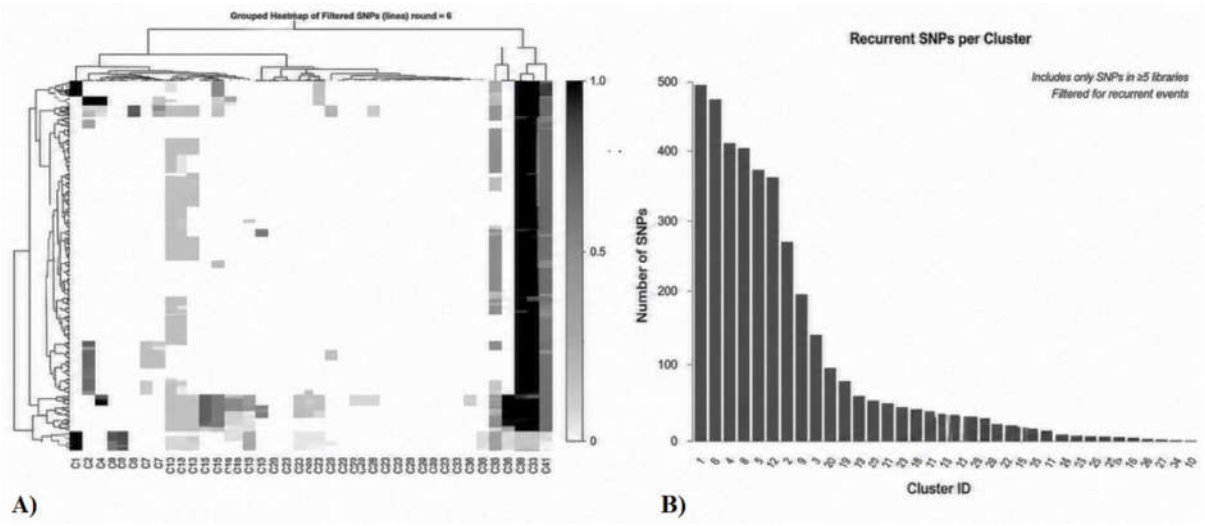

Fig. 2. Recurrent lncRNA-associated SNP patterns across clustered CRC libraries: A) Cluster-level heatmap showing the distribution of 549 highly recurrent lncRNA-associated SNPs across 36 clusters derived from hierarchical clustering of SNP presence/absence profiles using Jaccard distance (cut height = 0.3). Rows represent SNPs and columns represent clusters. Color intensity reflects the number of libraries within each cluster in which a SNP is detected (capped at 6 for visualization). Clusters 4, 6, and 30 show the highest SNP density, whereas several clusters exhibit sparse variant representation; B) Bar plot illustrating the number of recurrent SNPs per cluster following cluster-level filtering (presence in ≥5 clusters). A small subset of clusters accounts for the majority of recurrent SNPs, highlighting non-uniform distribution of lncRNA-associated variants.

Table 2. Cluster-level summary of recurrent lncRNA-associated SNPs across colorectal cancer RNA-seq libraries.

Clusters	Num_Libraries	Num_SNPs	Clusters	Num_Libraries	Num_SNPs
Cluster_1	1	503	Cluster_19	1	51
Cluster_2	3	196	Cluster_20	2	73
Cluster_3	1	87	Cluster_21	2	31
Cluster_4	11	416	Cluster_22	3	25
Cluster_5	3	373	Cluster_23	2	37
Cluster_6	15	485	Cluster_24	1	39
Cluster_7	1	382	Cluster_25	1	5
Cluster_8	2	413	Cluster_26	1	6
Cluster_9	2	139	Cluster_27	1	2
Cluster_10	5	0	Cluster_28	1	27
Cluster_11	1	7	Cluster_29	4	27
Cluster_12	1	273	Cluster_30	13	45
Cluster_13	1	5	Cluster_31	4	15
Cluster_14	1	7	Cluster_32	1	30
Cluster_15	1	17	Cluster_33	1	5
Cluster_16	1	3	Cluster_34	14	1
Cluster_17	5	47	Cluster_35	7	14
Cluster_18	3	34	Cluster_36	3	3

Clusters were defined by hierarchical clustering of RNA-seq libraries using Jaccard distance (cut height = 0.3). Num_Libraries indicates the number of RNA-seq libraries assigned to each cluster, and *Num_SNPs* represents the number of high-confidence lncRNA-associated SNPs detected in at least one library within that cluster.

lncRNA-level distribution of recurrent SNPs

Mapping the 549 highly recurrent SNPs to lncRNA gene symbols revealed a markedly non-uniform distribution throughout 99 distinct lncRNAs. The SNP burden per gene was highly skewed, with the majority of lncRNAs harboring only a small number of recurrent variants, while a limited subset accumulated a disproportionately high number of SNPs.

Among all genes, *MIR17HG* exhibited the highest SNP burden (~120 recurrent variants), followed by *ZEB2-AS1* (~70 variants). Several additional, largely uncharacterized lncRNAs, including LOC124907404, LOC124890606, and LOC105374140, contained intermediate SNP loads (approximately 20-60 variants), placing them among the most variant-enriched transcripts (Figure 3). Importantly, the enrichment of recurrent SNPs in a small subset of lncRNAs suggests the presence of non-coding mutational hotspots rather than random distribution of sequence variation. The identification of both well-characterized cancer-associated lncRNAs (e.g., *MIR17HG*, *ZEB2-AS1*) and poorly annotated transcripts within this high-burden group indicates that recurrent sequence variation may highlight functionally relevant non-coding loci as well as novel candidate regulators in colorectal cancer. These findings suggest that lncRNA-associated SNP accumulation may reflect selective pressures or structural constraints within specific transcripts, potentially influencing RNA stability, regulatory interactions, or chromatin-associated functions. Consequently, these high-burden lncRNAs represent strong candidates for future functional and biomarker-oriented investigations.

Bar plot illustrating the 20 lncRNA genes with the highest numbers of recurrent SNPs identified in the final filtered dataset (549 SNPs distributed across 99 lncRNAs). *MIR17HG* and *ZEB2-AS1* display the greatest SNP burdens, whereas several other lncRNAs (including *LOC124907404*, *LOC124890606*, *LOC105374140*, *LINC00881* and *MIR9-3HG*) exhibit intermediate variant counts (approximately 20–60 SNPs). The observed distribution indicates a markedly non-uniform pattern of lncRNA-associated SNP accumulation across CRC transcriptomes.

Discussion

In this study, we developed a lncRNA-focused RNA-seq variant-calling pipeline and applied it to a large collection of CRC transcriptomes. By aligning reads to a curated RefSeq lncRNA reference, applying stringent depth- and likelihood-based filters, and accounting for non-independence among libraries through clustering, we reduced an initially large and noisy variant set to 549 high-confidence recurrent SNPs mapped to 99 lncRNA genes. The substantial reduction from

raw variant calls highlights the high level of noise intrinsic to RNA-seq–based SNP discovery, where many initial variants are supported by limited read depth or confined to single libraries, suggesting technical artefacts or low-confidence events.

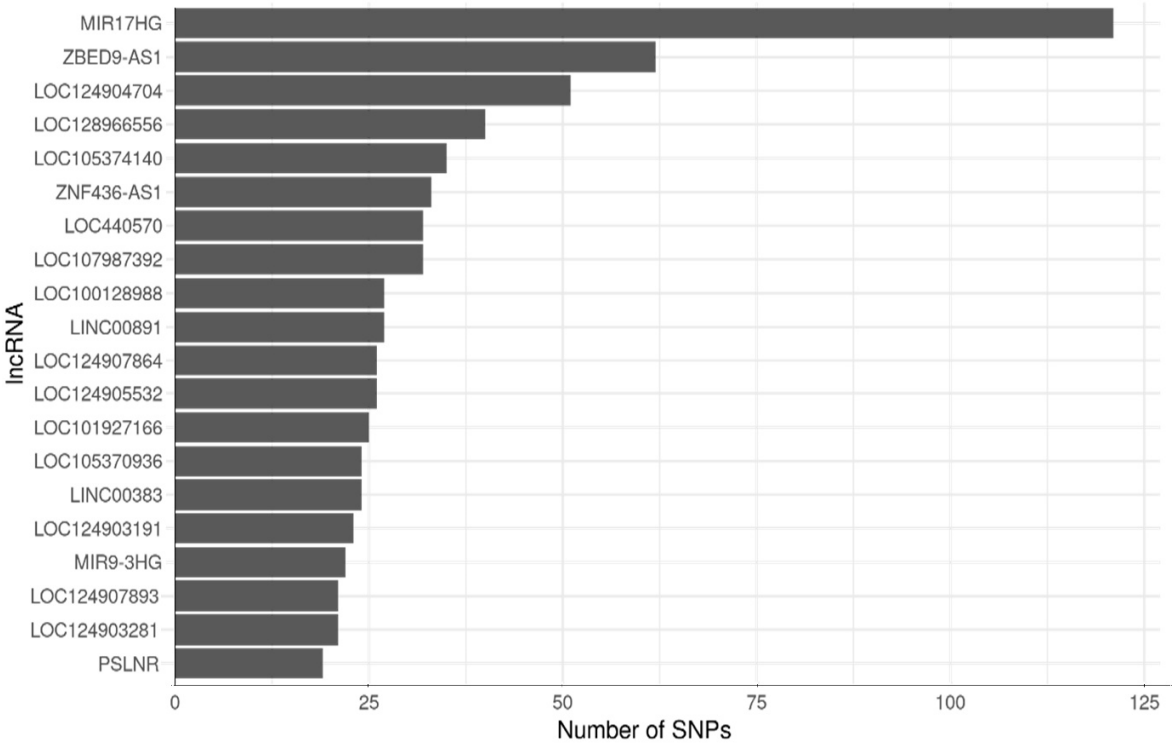

Fig. 3. Top lncRNAs ranked by recurrent SNP burden.

Prioritizing specificity through strict filtering and recurrence criteria substantially increased confidence in the retained SNPs, albeit at the cost of excluding rare or sample-specific variants. A key strength of this study is the explicit handling of library non-independence. By clustering RNA-seq libraries based on shared SNP profiles and collecting variant recurrence at the cluster level, we minimized inflation arising from technical and biological replicates and prioritized variants consistently detected across independent experimental contexts. Compared with previous studies, which have largely focused on differential expression of lncRNAs, our approach specifically targets sequence-level variation and integrates recurrence-based filtering with clustering, thereby improving the robustness and reproducibility of variant detection across heterogeneous RNA-seq datasets. Importantly, recurrent SNPs were not uniformly distributed across lncRNAs, but instead exhibited clear clustering within a limited subset of transcripts. This non-random distribution pattern suggests the presence of transcript-level mutational hotspots that may reflect underlying structural or regulatory constraints within specific lncRNA loci rather than randomly background variation. Notably, *MIR17HG* and *ZEB2-AS1* ranked among the top lncRNAs in terms of recurrent SNP burden, supporting their potential relevance as structurally or functionally sensitive loci. In contrast, several well-established lncRNAs in CRC, including *H19*, *MALAT1*, and *NEAT1*, did not exhibit a comparably high burden of recurrent sequence variants after stringent filtering. This discrepancy likely reflects fundamental differences between expression-based and sequence-variant–focused analyses. Importantly, the absence of high recurrent SNP burden in these lncRNAs does not contradict their biological relevance but instead suggests that their contribution to tumorigenesis may arise through

mechanisms other than recurrent sequence variation, such as transcriptional dysregulation or epigenetic modulation. Although comprehensive functional annotation of all identified variants was beyond the scope of this study, the observed clustering of SNPs within specific lncRNA transcripts provides indirect functional evidence. The accumulation of multiple variants within restricted transcript regions suggests the presence of localized mutational hotspots that may influence RNA secondary structure, stability, or interactions with RNA-binding proteins. Such non-random SNP distribution is unlikely to arise solely from technical artefacts and instead indicates underlying biological constraints or selective pressures acting on specific lncRNA loci. From a biological perspective, these findings support the concept that non-coding regions, particularly lncRNAs, can harbor structured patterns of recurrent variation that may contribute to regulatory dysregulation in CRC . In addition, several poorly characterized lncRNAs with elevated SNP burden were identified, highlighting novel candidate loci for future functional investigation. From a translational perspective, the recurrent lncRNA-associated SNPs identified in this study represent stable and reproducible molecular features that may serve as candidate biomarkers for CRC. Their recurrence across independent transcriptomic datasets suggests potential utility in patient stratification, molecular subtype classification, and integration into multi-omics biomarker panels. However, further experimental validation and clinical studies are required to confirm their diagnostic and prognostic value. A key limitation of this study is the inability to distinguish somatic mutations from germline polymorphisms due to the lack of matched normal samples. Therefore, a proportion of the identified variants may represent inherited variation rather than tumor-specific alterations. In addition, RNA-seq-based variant detection is inherently dependent on gene expression levels, which may bias variant discovery toward highly expressed transcripts. Nevertheless, the recurrence of these SNPs across independent datasets, combined with stringent filtering and cluster-level validation, reduces the likelihood of random or technical artefacts and supports their potential biological relevance. Overall, this study provides a scalable and robust framework for identifying recurrent lncRNA-associated sequence variants in CRC . Beyond its methodological contributions, the resulting SNP catalogue constitutes a valuable resource for future integrative and experimental studies aimed at elucidating the functional impact of lncRNA sequence variation in tumorigenesis. From an application-oriented perspective, these recurrent variants represent promising candidates for incorporation into biomarker-driven and precision oncology analyses, although their direct biological and clinical relevance requires further validation.

Acknowledgments

The authors express their sincere appreciation to the Institute of Science and High Technology and Environmental Sciences, Graduate University of Advanced Technology, Kerman, Iran, for their support and assistance.

Ethical Statement

This study did not involve human participants or animal experiments; therefore, ethical approval and informed consent were not required. All analyses were performed in accordance with academic integrity principles, including originality of work and avoidance of data manipulation or duplicate publication.

Data Availability

The processed variant catalogue, genomic coordinates, and clustering summary data are available as supplementary files and from the corresponding author upon reasonable request.

Conflict of interests

The authors declare that there are no conflicts of interest.

References

Andrabi, M. Q., Kesavan, Y., & Ramalingam, S. (2024). Non-coding RNAs as biomarkers for survival in CRC patients. *Current Aging Science,* 17(1), 5-15. https://doi.org/10.2174/1874609816666230202101054

Bian, Z., Zhang, J., Li, M., Feng, Y., Wang, X., Zhang, J., ... & Huang, Z. (2018). LncRNA-FEZF1-AS1 promotes tumor proliferation and metastasis in colorectal cancer by regulating

PKM2 signaling. *Clinical Cancer Research,* 24(19), 4808-4819. https://doi.org/10.1158/1078-0432.CCR-17-2967

Chen, W., Liu, X., Wu, Z., Tan, H., Yu, F., Wang, D., ... & Chen, Z. (2025). Unveiling the diagnostic power of lncRNAs in colorectal cancer: a meta-analysis. *BioMedical Engineering OnLine,* 24(1), 103. https://doi.org/10.1186/s12938-025-01431-3

Du, D., Shen, X., Zhang, Y., Yin, L., Pu, Y., & Liang, G. (2020). Expression of long non-coding RNA SFTA1P and its function in non-small cell lung cancer. *Pathology-Research and Practice,* 216(9), 153049. https://doi.org/10.1016/j.prp.2020.153049

He, F., Wei, R., Zhou, Z., Huang, L., Wang, Y., Tang, J., ... & Su, Z. (2019). Integrative analysis of somatic mutations in non-coding regions altering RNA secondary structures in cancer genomes. *Scientific Reports,* 9(1), 8205. https://doi.org/10.1038/s41598-019-44489-5

Hennig, E. E., Kluska, A., Piątkowska, M., Kulecka, M., Bałabas, A., Zeber-Lubecka, N., ... & Ostrowski, J. (2021). GWAS links new variant in long non-coding RNA LINC02006 with colorectal cancer susceptibility. *Biology,* 10(6), 465. https://doi.org/10.3390/biology10060465

Kerachian, M. A., & Azghandi, M. (2022). Identification of long non-coding RNA using single nucleotide epimutation analysis: A novel gene discovery approach. *Cancer Cell International,* 22(1), 337. https://doi.org/10.1186/s12935-022-02752-2

Liu, H., Li, Y., Karsidag, M., Tu, T., & Wang, P. (2025). Technical and biological biases in bulk transcriptomic data mining for cancer research. *Journal of Cancer,* 16(1), 34-43. https://doi.org/10.7150/jca.100922

Liu, Y., Liu, B., Jin, G., Zhang, J., Wang, X., Feng, Y., ... & Huang, Z. (2019). An integrated three-long non-coding RNA signature predicts prognosis in colorectal cancer patients. *Frontiers in Oncology,* 9, 1269. https://doi.org/10.3389/fonc.2019.01269

Lulli, M., Napoli, C., Landini, I., Mini, E., & Lapucci, A. (2022). Role of non-coding RNAs in colorectal cancer: Focus on long non-coding RNAs. *International Journal of Molecular Sciences,* 23(21), 13431. https://doi.org/10.3390/ijms232113431

Minotti, L., Agnoletto, C., Baldassari, F., Corra, F., & Volinia, S. (2018). SNPs and somatic mutations in long non-coding RNAs: a new frontier in cancer studies? *High-Throughput,* 7(4), 34. https://doi.org/10.3390/ht7040034

Tang, C., Liu, J., Hu, Q., Zeng, S., & Yu, L. (2021). Metastatic colorectal cancer: Perspectives on long non-coding RNAs and promising therapeutics. *European Journal of Pharmacology,* 908, 174367. https://doi.org/10.1016/j.ejphar.2021.174367

Tufail, M. (2023). HOTAIR in colorectal cancer: Structure, function, and therapeutic potential. *Medical Oncology,* 40(9), 259. https://doi.org/10.1007/s12032-023-02131-5

Wu, Y., & Xu, X. (2023). Long non-coding RNA signature in colorectal cancer: Research progression and clinical application. *Cancer Cell International,* 23(1), 28. https://doi.org/10.1186/s12935-023-02867-0

Yan, T., Shen, C., Jiang, P., Yu, C., Guo, F., Tian, X., ... & Fang, J. Y. (2021). Risk SNP-induced lncRNA-SLCC1 drives colorectal cancer through activating glycolysis signaling. *Signal Transduction and Targeted Therapy*, 6(1), 70. https://doi.org/10.1038/s41392-020-00446-7

Zhang, Z., Gu, M., Gu, Z., & Lou, Y. R. (2021). Role of long non-coding RNA polymorphisms in cancer chemotherapeutic response. *Journal of Personalized Medicine,* 11(6), 513. https://doi.org/10.3390/jpm11060513

Zhou, H., Hao, X., Zhang, P., & He, S. (2023). Noncoding RNA mutations in cancer. *Wiley Interdisciplinary Reviews: RNA,* 14(6), e1812. https://doi.org/10.1002/wrna.1812