Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Transcriptomic Profiling of Colorectal Cancer Cell Lines Reveals Novel Divergent Splicing Modalities in Long Non-Coding RNAs GAS5 and SNHG1

**John Mathew Manipadam**

jm_manipadam@yahoo.com

VPS LAKESHORE https://orcid.org/0000-0003-3771-4480

**Research Article**

**Keywords:** transcriptomic profiling, lnc rna, GAS5, SNHG1 , Colorectal cancer

**Posted Date:** July 14th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10321302/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** The authors declare no competing interests.

# Abstract

Esophageal and colorectal carcinomas exhibit widespread transcriptional dysregulation, yet the contribution of non-coding RNA splicing remains poorly characterized. In a preceding transcriptomic study utilizing primary esophageal adenocarcinoma (EAC) cell lines, long non-coding RNAs GAS5 and PVT1 were identified as major pan-cancer splicing hotspots via an integrated exon- and intron-centric computational pipeline. To determine if these architectural alterations represent ubiquitous gastrointestinal mechanisms or tissue-specific modalities, the current study deployed an identical reproducible pipeline—incorporating STAR alignment, rMATS, and MntJULiP—to evaluate deep-sequencing RNA-seq datasets of 36 colorectal cancer (CRC) cell lines against 18 normal colorectal mucosa samples. Following stringent concordance filtering, GAS5 and SNHG1 emerged as the most significantly spliced and upregulated lncRNAs in CRC.

A direct coordinate-level comparison revealed striking tissue-specific divergence: while the GAS5 locus at chr1:173.865 Mb previously demonstrated a compact Retained Intron (RI) companion cluster in EAC, the CRC cohort exhibited exclusively exon-centric variations (Mutually Exclusive Exons and Alternative 5' Splice Sites) at the identical genomic coordinate without significant intron retention. Crucially, while these specific GAS5 and SNHG1 isoforms possess exact intron matches in GENCODE v40 annotations, they displayed negligible junction support in GTEx normal colonic mucosa and were entirely absent from the TCGA PanCancer Atlas splicing catalogs. We propose that their absence in public databases is an artifact of the 'tumor dilution effect' inherent to bulk tumor sequencing, combined with standard Poly-A selection biases. This study highlights the growing significance of GAS5 splicing and upregulation in gastrointestinal malignancies and establishes SNHG1 as a potent, previously under-reported structural target in CRC progression.


# Introduction

Colorectal cancer (CRC) is a leading cause of cancer-related mortality worldwide. While large-scale transcriptomic efforts by consortia such as The Cancer Genome Atlas (TCGA) have established pervasive transcriptional activation across specific pathways, the contribution of alternative RNA splicing—particularly within long non-coding RNAs (lncRNAs)—has remained an underexplored continent in tumor pathophysiology. LncRNAs govern crucial cellular processes including chromatin architecture, transcriptional output, and RNA stability. Several lncRNAs are known to exhibit highly cancer-specific splice patterns that regulate these functions.

In our recent integrated analysis of esophageal adenocarcinoma (EAC), we identified the lncRNAs GAS5 and PVT1 as profound non-coding splicing hotspots [1]. Furthermore, that esophageal study demonstrated significant upregulation of GAS5, as evidenced by the gene_exp.diff profiling available in the EAC Zenodo repository. However, GAS5 was originally identified in the 1990s as a tumor suppressor gene that is characteristically downregulated in various tumors [2]. Utilizing orthogonal exon-centric and intron-centric detection algorithms, that prior study demonstrated that tumor-specific lncRNA isoforms

are frequently masked in standard workflows due to algorithmic constraints and the 'tumor dilution effect' inherent to bulk-tissue sequencing [1, 3]. To determine if these architectural alterations found in my esophageal study represent ubiquitous gastrointestinal mechanisms or tissue-specific modalities, the current study deployed an identical reproducible pipeline—incorporating STAR alignment, rMATS [4], and MntJULiP [5]—to evaluate deep-sequencing RNA-seq datasets of 36 colorectal cancer (CRC) cell lines against 18 normal colorectal mucosa samples—offering high signal-to-noise ratios devoid of stromal interference—against fresh-frozen normal colorectal mucosa.

The primary objectives of this study were to:

- Perform differential gene expression and splicing analysis on CRC cell lines versus normal mucosa using Cuffdiff (v2.2.1) [6], rMATS (exon-centric) [4], and MntJULiP (intron-centric) [5].
- Identify the most frequently spliced lncRNAs in CRC and compare their splicing architecture to previously established EAC models.
- Structurally validate candidate events against GENCODE annotations and assess their biological novelty using the GTEx and TCGA databases.

## Materials and Methods

# Data Acquisition and Cohort Standardization

Raw paired-end RNA-sequencing files representing the malignant cohort were acquired from the expanded Cancer Cell Line Encyclopedia (CCLE, BioProject PRJNA523380), consisting of 36 distinct CRC cell lines. To eliminate technical batch effects and preserve exon-intron boundary integrity, a meticulously matched normal cohort comprising 18 samples of fresh-frozen human normal colorectal mucosa (GEO Accession GSE50760) was utilized. Both cohorts were subjected to oligo-dT magnetic bead Poly-A selection. To mitigate sequencing depth bias, a strict library size threshold (< 10 GB) was enforced, thereby preventing ultra-deep outliers from artificially inflating the false discovery rate for novel splice junctions [7] (see Supplementary Code File (SCF): Code Cell (CC)- 1).

## Alignment and Splicing Quantification

Reads were aligned to the GRCh38 primary assembly using STAR (v2.7.10a) with GENCODE v40 annotations, utilizing restrictive multimapping thresholds to maximize novel junction discovery stringency (SCF, CC-1).

Differential gene expression was calculated using Cuffdiff (v2.2.1) [6]. Cuffdiff-derived gene_exp.diff values were used to compute raw log2 fold changes (SCF: CC-2). Isoform-level splicing burden was derived from isoform_exp.diff. A robust linear model adjusted gene-level log2 fold changes for splicing burden, and a second model adjusted for sample/batch effects (SCF: CC-3). The final bias-corrected list appears in TopUpregulated_BiasCorrected_Combined.csv (Table 1).

Exon-centric splicing variations were quantified using rMATS-Turbo (v4.1.2) [4] (SCF, CC-4). Intron-centric differential regulation was captured utilizing MntJULiP (v1.5.2) [5] with a q-value < = 0.05 constraint (SCF, CC-5). Splicing outputs were deterministically paired using an anchor-based matching algorithm utilizing a strict threshold of FDR < = 0.05 and |ΔPSI| >=0.20 (SCF: CC-6, 7) and structurally queried against the GENCODE annotated database (SCF, CC-8) and GTEx and TCGA databases via the Snaptron API utilizing an OS-level Web Application Firewall bypass strategy (see SCF CC-9).

## Results

Integration of rMATS and MntJULiP revealed extensive alternative splicing across the colorectal transcriptome (available in https://doi.org/10.5281/zenodo.21270412) [8]. Upon restricting the analysis to non-coding RNA biotypes and applying our stringent dual-algorithm concordance filter, GAS5 and SNHG1 emerged as the topmost significant splicing variants in the CRC cohort, concurrent with notable overall gene upregulation (Table 2).

It is found after Cuffdiff analysis that GAS5 is significantly upregulated (log2 fold change = 1.32, q-value = 0.00025). Cuffdiff results showed that out of 9,539 significantly differentially expressed genes, there is a massive shift with 4,683 genes significantly upregulated and 4,856 downregulated. This confirms a broad, system-wide transcriptomic dysregulation. GAS5 ranks 2,997th overall in terms of its log2 fold change magnitude with 78 distinct transcripts detected, 22 of which are significantly upregulated (available in the Zenodo repository Cuffdiff results).

## Divergent Splicing Modalities at the GAS5 Locus

A striking structural divergence was identified when comparing the GAS5 splicing architecture in the colorectal cohort against our previously reported esophageal findings. In EAC, the chr1:173.865 Mb locus presented as a compact Retained Intron (RI) cluster flanked by MXE and A5SS events—necessitating simultaneous exon- and intron-centric algorithmic validation. Conversely, the CRC cell lines analyzed herein exhibited pervasive and exclusively exon-centric variations at the identical genomic coordinate.

CRC models demonstrated massive multi-exon switching (Mutually Exclusive Exons) and Alternative 5' Splice Sites, without the significant intron retention companions seen in the esophageal cohort. This structural divergence was visualized by extracting exact junction bounds to render overlapping events on a shifted axis (SCF: CC-10, Fig. 1) and visually validated via group-averaged Sashimi plots (SCF, CC-11, Fig. 2). This contrast highlights a tissue-specific modality of transcriptomic dysregulation, suggesting that while the genomic locus itself represents a pan-gastrointestinal anchor of structural instability, its downstream processing is profoundly dictated by the local tumor microenvironment. Similar plotting was done for SNHG1 (Figs. 3, 4).

## Discussion

This study deployed a unified exon-intron splicing workflow to dissect the non-coding transcriptome of colorectal cancer. Building upon our initial discovery of GAS5 dysregulation in esophageal adenocarcinoma [1], this analysis confirms that the GAS5 locus is a major structural hotspot across gastrointestinal malignancies, albeit with highly tissue-specific splicing configurations and unexpected upregulation. This upregulation was also found in my previous study on esophageal adenocarcinoma cell lines (Cuffdiff results of this study available at Zenodo DOI: 10.5281/zenodo.17644639). This stands in contrast to the historical belief that GAS5 is strictly a tumor suppressor gene that is globally downregulated in tumors [2].

The transition from an intron-retention-dominant signature in EAC to the complex exon-switching networks (MXE, A5SS) observed in CRC suggests that distinct gastrointestinal tumors exploit entirely different RNA processing pathways to modulate the tumor related functions of GAS5 [9].

Furthermore, this study highlights critical variations in SNHG1, an lncRNA increasingly recognized for its role in tumor progression and invasion [11]. SNHG1 has been heavily implicated in tumor progression through the modulation of competitive endogenous RNA (ceRNA) networks, promoting cellular proliferation and invasion in CRC [10, 11].

The specific exon-centric SNHG1 and GAS5 splice junctions identified are structurally annotated as exact matches within the GENCODE v40 catalog. However, cross-referencing with the TCGA PanCancer Atlas (TCGA-COAD) and GTEx normal colonic mucosa via Snaptron yielded negligible junction support. We posit that the systematic absence of these events in major public databases represents a technical artifact rather than biological irrelevance. Standard bulk-tissue sequencing frameworks suffer heavily from the 'tumor dilution effect,' where rare or highly cancer-specific lncRNA splice events are obscured by transcriptomic noise from surrounding stromal and immune infiltration [3, 7]. Our utilization of pure epithelial cell line models bypassed this dilution, providing the signal-to-noise clarity required to resolve these 'invisible' targets. This growing body of evidence—spanning both esophageal and colorectal domains—underscores the critical, under-reported role of alternative splicing in lncRNA-mediated oncogenesis, offering precise, coordinate-resolved candidates for future RNA-directed interventions.

## Declarations

# Data Availability:

Code and processed datasets, including Cuffdiff, rMATS, and MntJULiP results, are available at Zenodo (https://doi.org/10.5281/zenodo.21270412) [8].

## References

1. Manipadam JM Integrated Exon and Intron Splicing Analysis Identifies Novel GAS5 and PVT1 Regulatory Junctions in Esophageal Adenocarcinoma. JCO Oncology Advances. (Submitted)

2. Coccia EM et al (1992) Isolation and characterization of a novel gene downregulated in growth-arrested mammalian cells. Mol Cell Biol 12(8):3514–3521. 10.1128/mcb.12.8.3514
3. Ryan M, Cleland J, Kim R, Wong WC, Weinstein JN (2016) TCGA SpliceSeq: a compendium of alternative mRNA splicing in cancer. Nucleic Acids Res 44(D1):D1018–D1022. 10.1093/nar/gkv1288
4. Shen S et al (2014) rMATS: Robust detection of splicing. Proc Natl Acad Sci USA 111:E5593–E5601. 10.1073/pnas.1419161111
5. Yang G, Sabunciyan S, Florea L (2022) Comprehensive and scalable quantification of splicing differences with MntJULiP. Genome Biol 23(1):82. 10.1186/s13059-022-02767-y
6. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31(1):46–53. 10.1038/nbt.2450
7. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP (2014) Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet 15(2):121–132. 10.1038/nrg3642
8. Zenodo Data Repository rMATS and MntJULiP Results. https://doi.org/10.5281/zenodo.21270412
9. Kino T et al (2010) Reciprocal regulation of GAS5 and glucocorticoid receptor. Mol Cell Biol 30(13):3242–3254. 10.1128/MCB.01516-09
10. Pickard MR, Williams GT (2015) Molecular and Cellular Mechanisms of Action of Tumour Suppressor GAS5 LncRNA. Genes (Basel) 6(3):484–499. 10.3390/genes6030484
11. Wang M et al (2019) LncRNA SNHG1 promotes cell proliferation, migration, and invasion in colorectal cancer. Biomed Pharmacother 117:109165. 10.1016/j.biopha.2019.109165

## Tables

Tables 1 and 2 are available in the Supplementary Files section.

## Figures

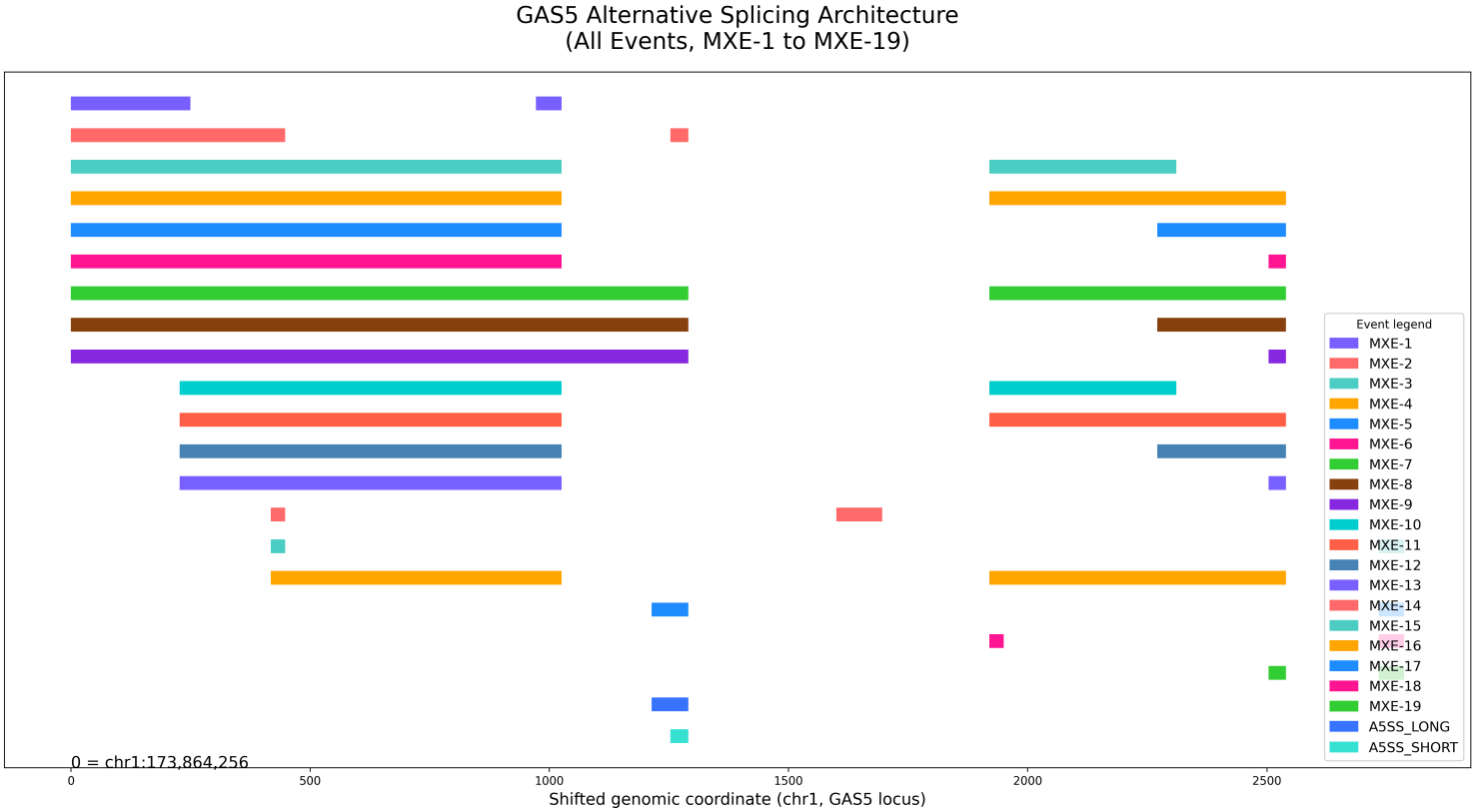



**Figure 1**

GAS5 splicing architecture map




## Figure 2

GAS5 sashimi plots

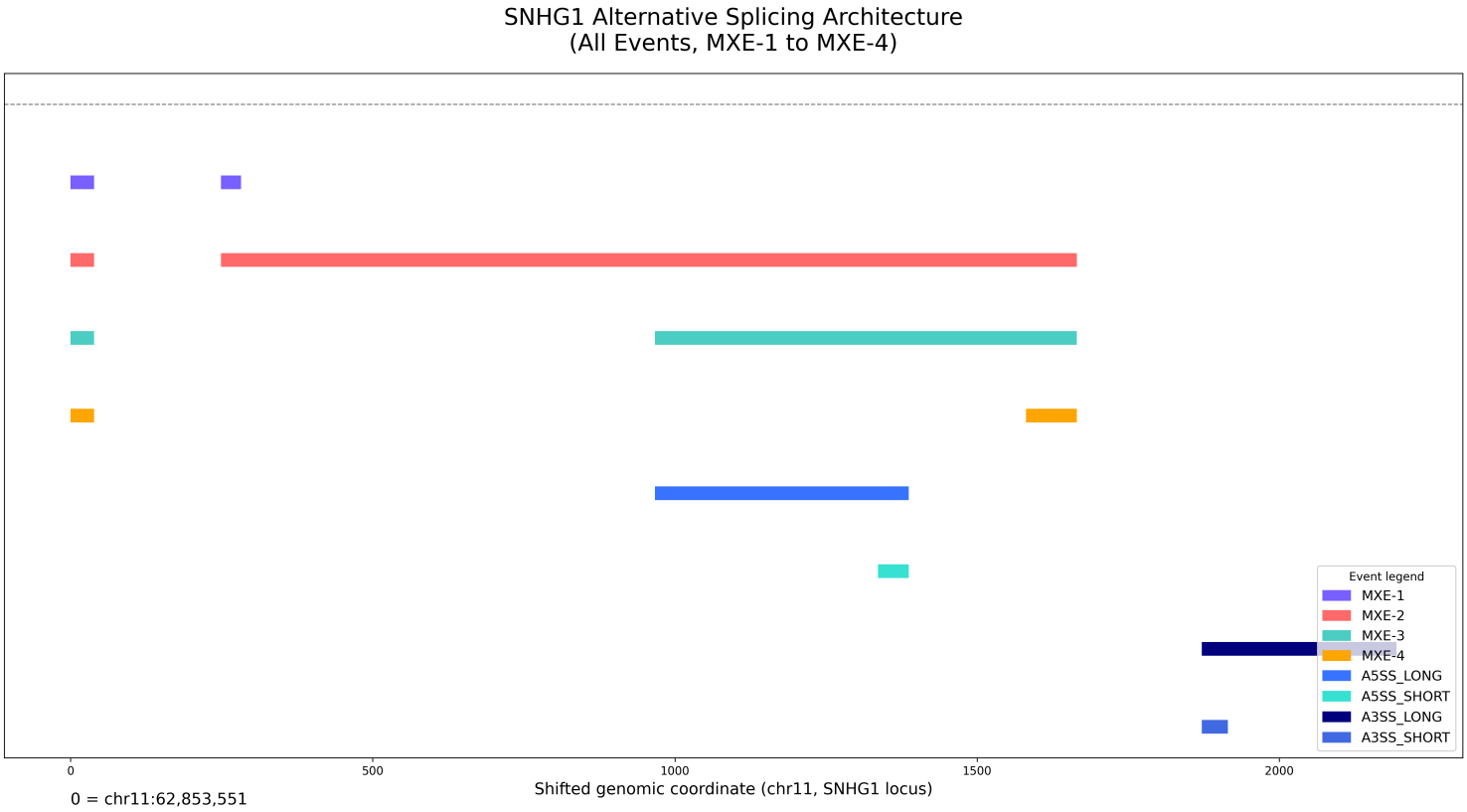



**Figure 3**

SNHG1 splicing architecture map

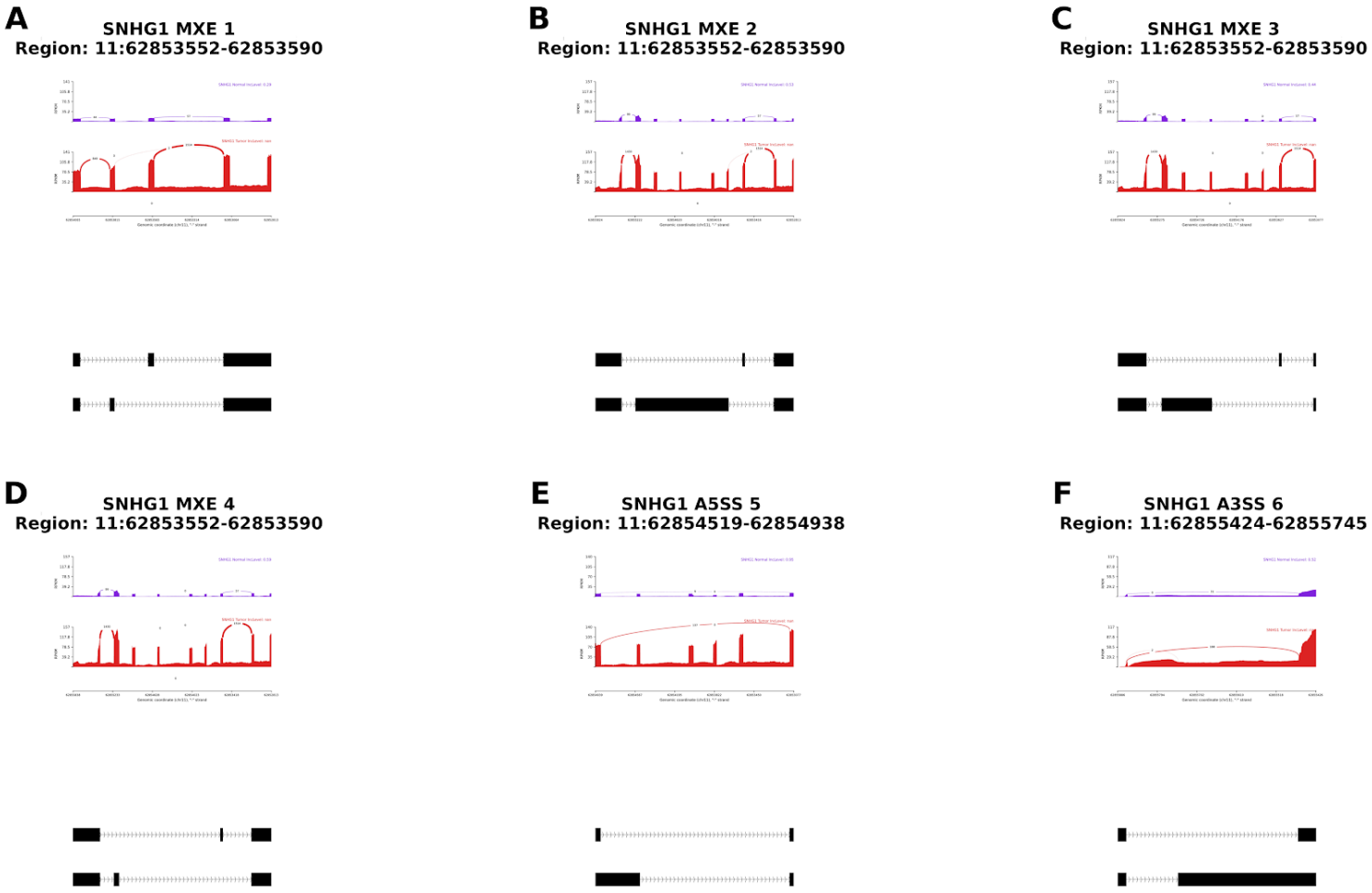



**Figure 4**

SNHG1 sashimi plots

# Supplementary Files

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

- CopyofUntitled163.ipynb
- TopUpregulatedBiasCorrectedCombined.csv
- FinalValidatedCRCTargetsDualQuery.csv