Briefings in Bioinformatics, 2025, **26(3)**, bbaf219
https://doi.org/10.1093/bib/bbaf219
Case Study

Multicancer analyses of short tandem repeat variations reveal shared gene regulatory mechanisms

Feifei Xia [1,2,3], Max Adriaan Verbiest [1,2,3], Oxana Lundström [4], Tugce Bilgin Sonay [1,3], Michael Baudis [2,3], Maria Anisimova [1,3,*]

[1] Institute of Computational Life Sciences, Zurich University of Applied Sciences, Schloss 4, 8820 Wädenswil, Switzerland
[2] Department of Molecular Life Sciences, University of Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland
[3] Swiss Institute of Bioinformatics, Amphipôle, Quartier UNIL-Sorge, 1015 Lausanne, Switzerland
[4] Department of Computer Science and Media Technology, Linnaeus University, Universitetsplatsen 1, 352 52 Växjö, Sweden

[*] Corresponding author. E-mail: anis@zhaw.ch

Abstract

Short tandem repeats (STRs) have been reported to influence gene expression across various human tissues. While STR variations are enriched in colorectal, stomach, and endometrial cancers, particularly in microsatellite instable tumors, their functional effects and regulatory mechanisms on gene expression remain poorly understood across these cancer types. Here, we leverage whole-exome sequencing and gene expression data to identify STRs for which repeat lengths are associated with the expression of nearby genes (eSTRs) in colorectal, stomach, and endometrial tumors. While most eSTRs are cancer-specific, shared eSTRs across multiple cancers exhibit consistent effects on gene expression. Notably, coding-region eSTRs identified in all three cancer types show positive correlations with nearby gene expression. We further validate the functional effects of eSTRs by demonstrating associations between somatic eSTR mutations and gene expression changes during the transition from normal to tumor tissues, suggesting their potential roles in tumorigenesis. Combined with DNA methylation data, we perform the first quantitative analysis of the interplay between STR variations and DNA methylation in tumors. We identify eSTRs where repeat lengths are associated with methylation levels of nearby CpG sites (meSTRs) and show that >70% of eSTRs are significantly linked to local DNA methylation. Importantly, the effects of meSTRs on DNA methylation remain consistent across cancer types. Overall, our findings enhance the understanding of how functional STR variations influence gene expression and DNA methylation. Our study highlights shared regulatory mechanisms of STRs across multiple cancers, offering a foundation for future research into their broader implications in tumor biology.

Keywords: short tandem repeats; gene expression; DNA methylation; quantitative trait locus analysis; multi-omics.

Introduction

Short tandem repeats (STRs), also known as microsatellites, are repetitive sequences of 1 to 6 base pairs that are highly polymorphic and widely distributed throughout the human genome [1, 2]. They account for ~3% of the human genome [3]. STR mutation rates are orders of magnitude higher than nucleotide substitution rates, thus representing a significant source of genetic variations [4–6]. Pathogenic expansions of STRs have been known to cause Mendelian disorders and neurological diseases, such as Fragile X syndrome and Huntington’s disease [7]. The underlying pathological mechanisms are diverse and depend on their genomic location, sequence composition, and length [8, 9]. While coding-region STR expansions can directly lead to misfolding of the protein, the majority of studied STR expansions are in noncoding regions, potentially affecting gene expression through a variety of mechanisms, such as epigenetic silencing of the gene *FMR1* [10], acting as nucleosome positioning signals [11], altering the affinity of nearby DNA-binding sites [12]. Beyond Mendelian disorder, STRs have been proposed to contribute to complex traits in diverse organisms. Recently, multiple genome-wide association studies have revealed that STR length variations play a significant role in molecular and cellular processes, including gene expression [13–17], DNA methylation [18, 19], and alternative splicing [20].

STRs are closely linked to cancer, particularly through microsatellite instability (MSI), which arises from DNA mismatch repair (MMR) deficiency [21]. MSI is a critical phenotype in colorectal, stomach, and endometrial cancers, affecting >15% of patients and associated with favorable responses to immunotherapy [22–24]. STR variations in coding regions can result in frameshift mutations that may alter the function of tumor driver genes and also provide a substantial source of tumor-specific neoantigens [25, 26]. Consequently, MSI detection and characterization with next-generation sequencing have been extensively studied [27–31]. However, the regulatory functions of STR variations on gene expression in cancer remain largely unexplored [32].

DNA methylation has emerged as another important diagnostic and prognostic marker for many cancers, including colorectal, stomach, and endometrial cancer [33–35]. It has been demonstrated that DNA methylation plays a critical role in regulating gene expression by modifying transcription factor binding sites,

Received: February 7, 2025. **Revised:** April 22, 2025. **Accepted:** April 25, 2025

acting either as a repressive or activating mark depending on the methylation region [36]. One of the key mechanisms leading to MSI is the promoter methylation of the MMR gene MLH1, which is an essential component of the mismatch repair system [37]. Furthermore, MSI tumors often overlap with the CpG island methylator phenotype (CIMP), a distinct epigenetic subtype characterized by widespread CpG island hypermethylation [38]. Despite these well-established interactions, the interplay between STR variations and local DNA methylation levels has not been quantitatively investigated in cancer.

Building on our previous identification of expression short tandem repeats (eSTRs) in colorectal cancer (CRC) [32], we extended our analyses to stomach and endometrial cancers and incorporated DNA methylation data to explore the dual role of STRs functioning as eSTRs and methylation and expression short tandem repeats (meSTRs) across multiple cancers. Using tumor-derived STR genotypes, gene expression, and DNA methylation data, we performed quantitative trait locus (QTL) analyses to identify eSTRs and subsequently meSTRs where eSTR repeat lengths are associated with nearby DNA methylation levels. Our results revealed a strong concordance between eSTRs and meSTRs, with consistent regulatory effects across cancers. Overall, these findings highlight the multifaceted regulatory roles of STRs and suggest shared regulatory mechanisms on nearby gene expression and DNA methylation.

Methods

Dataset collection

Whole-exome sequencing. Whole-exome sequencing (WES) data from colon adenocarcinoma (COAD), rectum adenocarcinoma (READ), stomach adenocarcinoma (STAD), and uterine corpus endometrial carcinoma (UCEC) from the TCGA database were accessed through dbGap under the study phs000178.v11.p8. The COAD and READ datasets were merged into a single CRC dataset. The specific sample selection used for the analyses is detailed in Supplementary Note 1.

Gene expression and DNA methylation. Gene expression (TPM normalized) and DNA methylation (Human methylation 450k) for CRC, STAD, and UCEC cohorts were downloaded using the GDC data transfer tool from the Genomics Data Commons Data Portal (https://gdc-portal.nci.nih.gov/).

MSI phenotype. The MSI phenotypes (MSI-H, MSI-L/MSS) of CRC, STAD, and UCEC tumors were obtained from the GDAC website (https://gdac.broadinstitute.org). Previous studies have suggested that MSI-L and MSS samples do not show distinct molecular and clinical features [39, 40]. Therefore, for simplicity, we merged MSI-L and MSS samples and collectively referred to them as MSS samples, while MSI-H samples were referred to as MSI.

Principal component analysis of STR profiles

STR genotyping was performed using GangSTR [41] on WES data for colorectal (CRC), stomach (STAD), and endometrial (UCEC) normal and tumor samples. Further details on STR genotyping and preprocessing are provided in Supplementary Note 2. To ensure the data quality of STR genotypes from tumors, we compared the STR profiles from paired normal and tumor samples using principal component analysis (PCA). For this analysis, we constructed an STR allele length matrix $\mathbf{S}_{n,m}$ where n represents the number of STR loci and m represents the number of normal or tumor samples for each cancer type. Each element $c_{i,j}$ in the matrix corresponds to the mean STR allele length for a diploid call at a given locus in a specific sample. $c_{i,j}$ is set to NaN when the ith STR call of the jth sample is missing. Before further analysis, STR loci missing in >50% of the total m samples were discarded. Additionally, STR loci with length variation <0.1 across m samples were also excluded. The filtered matrices were then imputed using k-nearest neighbors with the parameters `n_neighbors=5, weights=distance`, implemented in the KNNImputer function from the Python scikit-learn library [42]. The imputed matrices were subsequently used as input for PCA, also using the Python scikit-learn library.

eSTRs identification

For each cancer type, we filtered out STR genotypes called in <10% of total tumor samples and then restricted our analyses to variable STRs with a standard deviation of repeat lengths > 0.1 within the cohort. We mapped STRs to a 5kb upstream window of gene regions to define STR–gene pairs. We excluded genes with median TPM values of zero and the remaining expression values were quantile normalized separately for each cancer type to a standard normal distribution. Analyses were further restricted to protein-coding genes based on GENCODE v.36 annotation. Altogether, 17 039, 11 097, and 22 147 STR–gene pairs remained in 451 CRC, 409 STAD, and 466 UCEC tumors, respectively. For each STR–gene pair, we fitted a linear regression model between STR mean allele length and gene expression levels:

$$\mathrm{Y} = \beta_0 + \beta\,\mathrm{X} + \boldsymbol{\beta_{1..k}}\,\mathbf{PC_{(1..k)}} + \boldsymbol{\beta_n}\,\mathbf{N} + \epsilon, \tag{1}$$

where X denotes STR genotypes, specifically mean allele length at the STR locus, Y denotes gene expression levels, β denotes the effect size, and ϵ is the error term. To control for technical variations in expression, we applied PCA correction for gene expression [43]. To ascertain the number of principal components (PCs) that capture technical variability, we carried out multiple rounds of association analyses for each cancer type separately, each time incorporating 0,1,2,3,4,5 PCs as covariates. This approach identified 2, 0, and 2 PCs for CRC, STAD, and UCEC, respectively, as optimal covariates for maximizing eSTR discovery (Supplementary Fig. S4). **N** represents a vector of additional covariates, including ***population structure, gender***, and ***year***. To correct for ***population structure***, we used inferred population stratification from an aggregated model for patients without self-reported population [44]. ***Year*** represents technical variation identified in STR genotypes (Supplementary Fig. S2). A separate regression analysis was performed for each STR–gene pair in each tumor type. Linear regression was fitted using the ordinary least squares function from the Python `statsmodels.api` module [45]. Samples with missing STR genotypes or expression values were excluded from each regression analysis. To reduce the effect of outlier STR genotypes, we removed samples with genotypes observed in fewer than three samples. To control the false discovery rate (FDR), we applied Bonferroni multiple testing correction for the number of evaluated STR–gene pairs tested within each cancer type. eSTRs were defined as STR loci where the mean allele length was significantly associated with the expression level of a nearby gene ($FDR < 0.05$).

Cross-cancer eSTR analysis

To compare the effects of eSTRs across different cancer types, we first identified overlapping eGenes that showed significant associations with eSTRs in multiple cancers. We then applied regression analysis to calculate residual gene expression by accounting for the aforementioned covariates. Using these residuals, we estimated the posterior effect sizes of each eSTR in each cancer type. Subsequently, we performed pairwise correlation analyses of the adjusted effect sizes for the overlapping eGenes across cancer types. For eGenes associated with multiple eSTRs, we averaged the effect sizes to ensure a consistent and comparable evaluation.

GO enrichment analysis

To gain insight into the biological functions and pathways associated with the identified eGenes, we performed Gene Ontology (GO) enrichment analysis using DAVID (Database for Annotation, Visualization, and Integrated Discovery) [46]. The analysis was conducted separately for each cancer type. We focused on GO categories for biological process, molecular function, and cellular component to identify overrepresented functional terms. GO terms with FDR corrected P-values $< .05$ were considered statistically significant in CRC. However, due to the smaller number of eGenes from STAD and UCEC tumors, no significantly enriched GO terms were identified for these cancer types.

eSTR mutability analysis

We compared the mutability of eSTRs and non-eSTRs in MSI and MSS tumors, grouping STRs by unit size and reference allele length [32]. For each repeat type, we retrieved somatic mutations for eSTRs and non-eSTRs, excluding repeat types observed <50 times per cancer type. We then calculated the fraction of mutated eSTRs and non-eSTRs within each repeat type, defining eSTRs as more mutable when their mutation fraction exceeded non-eSTRs by ≥0.05. To assess statistical significance, we performed 10 000 permutations, randomly shuffling eSTR labels and recording the fraction of repeat types with increased eSTR mutability. This generated null distributions serving as baseline models. The observed fraction of repeat types with elevated eSTR mutability was then compared with these null distributions using permutation tests to determine statistical significance.

eSTR somatic mutation and gene expression change

Matched gene expression data were available in 33, 33, and 19 paired tumor and normal samples for CRC, STAD, and UCEC cohorts. Among these paired samples, we detected 3285, 302, and 94 somatic mutations in putative eSTRs using the STR mutation calling method described above. For each eSTR and its corresponding eGene, gene expression changes were determined by calculating the differences between paired tumor and normal samples after performing quantile normalization separately in the normal and tumor samples. Using the eSTR mutation lengths Δx and the corresponding expression change Δy of the associated eGenes, we assessed whether our linear models could predict gene expression changes in response to somatic eSTR mutations both directionally and quantitatively. To evaluate directional predictions, we determined whether the actual direction of gene expression change Δy matched the predicted change based on the eSTR mutation length Δx and the adjust coefficient β'. We recalculated the coefficients β' for each eSTR based on the residuals of adjusted gene expression values Y'. A directional prediction was considered correct if both the gene expression change (Δy) and the predicted change ($\beta' * \Delta x$) showed the same sign (both positive or negative). For eSTR mutations that correctly predicted the direction, we further calculated the predictive score to assess the accuracy of quantitatively predicting gene expression changes, as defined by the following formula:

$$predictive\_score = 1 - \left| \frac{\beta' * \Delta x - \Delta y}{\Delta y} \right| \qquad (2)$$

Negative predictive scores were observed for 11.5%, 10.6%, and 11.7% of eSTR mutations in CRC, STAD, and UCEC, respectively. These negative values likely reflect noise or other factors that influence gene expression. We retained only the eSTR mutations with positive predictive scores for visualization.

Conditional analysis

Given that the MSI phenotype can affect both STR length variation and gene expression, we incorporated MSI status as a covariate in our regression models to account for its confounding effect. Specifically, we conducted a conditional analysis to determine whether the association between STR length variation and gene expression is independent of MSI phenotype. For each eSTR, we performed the following analysis:

$$Y = \beta_0 + \beta X + \beta_m MSI\_status + \boldsymbol{\beta_{1..k}}\,\mathbf{PC_{(1..k)}} + \boldsymbol{\beta_n}\,\mathbf{N} + \epsilon \qquad (3)$$

MSI_Status is a binary variable indicating the MSI phenotype. eSTRs were considered as regulators of gene expression independent of MSI phenotype when the resulting association upon conditioning was still significant (nominal $P < .05$) and had the same directionality as in the unconditioned analysis.

Overlap of eGenes and SNP-associated genes

SNP-associated genes in COAD, READ, STAD, and UCEC were downloaded from the database PancanQTL [47]. The database conducted expression quantitative trait locus (eQTL) analysis using gene expression data and SNP genotypes, and comprehensively provided the results of eQTLs from the TCGA. We filtered the SNP-gene association with a threshold of $FRD < 0.05$.

meSTRs identification and analysis

Among the tumors analyzed, DNA methylation data were available for 393 (CRC), 395 (STAD), and 431 (UCEC) samples, with methylation levels represented by β-values (0 = unmethylated, 1 = fully methylated). To ensure measurement reliability, we excluded probes mapping to multiple genomic locations [48]. CpG site annotations from Illumina's 450k methylation arrays were obtained using R [49] and minfi [50], then lifted to hg38 using bedtools [51]. CpG sites with a call rate below 80% were filtered out. We then selected CpG sites located within 5 kb upstream or gene regions of each eSTR. Association testing was performed for 33 910 (CRC), 10 672 (STAD), and 2598 (UCEC) eSTR:CpG pairs, controlling for ***population structure, gender***, and ***year***, followed by multiple testing correction using the Benjamini–Hochberg method.

To evaluate whether eSTRs are more frequently linked to DNA methylation, we generated background sets from 11 879 (CRC), 8223 (STAD), and 16 832 (UCEC) non-eSTRs. For each cancer type, we randomly sampled 1000 non-eSTR subsets (with replacement), ensuring subset sizes matched eSTR counts. Non-eSTR:CpG pairs were mapped using the same procedure as eSTRs, and identical association analyses were performed. We then quantified the proportion of methylation-associated STRs (mSTRs) within each subset, establishing a null expectation distribution for each cancer type.

To compare the effects of meSTRs across different cancer types, we first selected the eSTR:CpG pairs showing significant associations in multiple cancer types. We then applied linear regression analysis to calculate the residual methylation levels, accounting for relevant covariates. Using these residuals, we estimated the posterior effect sizes for each overlapped pair. Subsequently, we performed pairwise correlation analyses between the effect-size estimates for the overlapping eSTR:CpG pairs across cancer types to examine consistency in meSTR effects.

Results

STR profiles capture MSI phenotype and population structure in tumors

To examine STR length variation patterns in CRC, STAD, and UCEC, we analyzed WES data from 433, 439, and 464 matched normal and tumor samples from TCGA. STR genotyping was performed using GangSTR [41], followed by PCA on variable STRs separately for tumor and matched normal samples in each cancer type (Fig. 1a).

Across all three cancers, the distinction between MSI and MSS tumors was well captured by the first PC derived from tumor samples (tumor PC1). Tumor PC1 explained a higher proportion of variance (20%–30%) compared with that of matched normal samples (<=10%), largely due to extensive somatic STR mutations in MSI tumors, particularly deletions (Fig. 1d, Supplementary Note 3). Additionally, the second PC from tumor samples (tumor PC2) was strongly correlated with the first PC from the matched normal samples (normal PC1) (Fig. 1b), confirming the consistency of STR genotyping and its reflection of true biological variation. Population structure patterns were evident in tumor PC2 and normal PC1 (Fig. 1c, Supplementary Fig. S1), further reinforcing the reliability of STR genotyping. Finally, technical artifacts were identified (Supplementary Fig. S2) and accounted for as covariates in subsequent analyses.

eSTRs identification and characterization

We performed an exome-wide analysis to identify associations between repeat length at each STR locus and the expression of nearby genes in CRC, STAD, and UCEC tumors. Our analysis focused on 1326 tumor samples with high-quality WES and RNA-sequencing data available from the TCGA database (Fig. 2a). After filtering low-quality STR calls (see Methods), we fitted linear models to test for associations between gene expression and the mean length of each STR located within the gene bodies or up to 5 kilobases (kb) upstream. The models were adjusted for gender, population structure, and technical covariates (see Methods). In total, 50 283 STR–gene pair tests were performed across the three cancer types. After applying Benjamini–Hochberg multiple testing correction with a significance threshold of $FDR < 0.05$, we identified 1411, 382, and 108 significant STR–gene associations in CRC, STAD, and UCEC, respectively (Fig. 2b, Supplementary Table 1). These STR loci were considered as eSTRs with putative regulatory roles. Over 60% of eSTRs were mononucleotide repeats located within intronic regions, with a small proportion located in coding or noncoding exons and untranslated regions (UTRs) (Fig. 2c). These findings align with previous research showing that mononucleotide repeats are the most abundant STRs in the human genome [2] and represent a significant source of STR variability in MSI tumors [28, 52]. We next examined effect-size biases among STR–gene associations. Overall, we did not observe any significant bias in the direction of STR effects on gene expression. Among the eSTRs detected in CRC, we observed interesting associations. For instance, the repeat length of an intronic eSTR positively correlated with the expression level of UBR5 (Fig. 2d), a gene that has been studied to be of clinical and biological significance in the progression of CRC [53, 54]. Similarly, we identified a negative association between the repeat length of an intronic eSTR and the expression of *JAK2* (Fig. 2e), which is involved in the pathogenesis of CRC [55, 56].

While most eGenes (genes associated with at least one eSTR) were linked to a single eSTR, there were 223, 26, 8 eGenes in CRC, STAD, and UCEC associated with multiple eSTRs (Fig. 2f). In CRC, some eGenes were associated with up to five eSTRs. Notably, these multiple eSTRs generally showed the same direction of correlation on gene expression. For example, 26 eGenes associated with multiple eSTRs showed concordant effect directions in STAD (Fig. 2g). This pattern of consistent effects suggests that eSTRs are not simply tagging other causal variants, as indicated by previous studies [57]. However, exceptions in CRC included *ERAP2*, *KLKB1*, *TRAPPC4*, and *POLQ*, which were associated with eSTRs of different repeat unit lengths. This could be related to the hypothesis that distinct repeat classes may affect gene expression through different regulatory mechanisms [14]. Additionally, for eGenes *ST7*, *FAHD1*, and *WFDC3*, the associated eSTRs were assigned to multiple genes due to overlapping transcripts.

Compared with our previous study [32], which identified 1295 significant STR–gene pairs, we confirmed 476 eSTRs here with our larger sample size and more robust confounder adjustments. While the previous study did not specifically investigate the consistent directional effects of multiple eSTRs on a single gene, this pattern of multiple eSTRs was also observed in the previous study except for gene *ATP6V1D*, *SYNE2*, and *ERAP2*, reinforcing the reliability and biological relevance of this finding.

Consistent effects of eSTRs on gene expression across cancers

Among the significant STR–gene associations, there were 1113, 351, and 100 unique eGenes in CRC, STAD, and UCEC, respectively. To explore the regulatory roles of eSTRs on gene expression across cancers, we visualized the overlap of eGenes identified in each cancer type (Fig. 3a). For each pair of cancers, we selected shared eGenes and computed the Pearson correlation between their effect sizes (Fig. 3b). When eGenes were associated with multiple eSTRs, we averaged their effect sizes. Remarkably, 82.7%, 90%, and 93.3% of overlapped eGenes were linked to at least one common eSTR. The effect sizes of these shared eSTRs and eGenes exhibited strong correlations across cancer types: CRC and STAD ($R = 0.69, P = 2.26e^{-20}$), CRC and UCEC ($R = 0.63, P = 1.8e^{-4}$), STAD and UCEC ($R = 0.90, P = 5.01e^{-6}$). Our results showcase that certain eSTRs may regulate gene expression through shared regulatory mechanisms across different cancer types. This supports previous findings that most eSTRs act by global mechanisms across diverse human tissues [14].

Next, we examined the 10 eGenes shared across all three cancers (Fig. 3c). Interestingly, the expression of *ABCF1*, *LTN1*, *SEC31A*, and *TNKS2* positively correlated with eSTRs located in coding regions (Supplementary Fig. S5). STR variations in coding regions have been known to cause frameshift mutations, which can provide a substantial source for tumor-specific neoantigens [26, 58]. These findings suggest that eSTRs in coding regions may not only regulate gene expression but also contribute to the immunogenic landscape of cancer by facilitating the generation of tumor-specific neoantigens.

To further explore the biological processes and functions influenced by eSTRs, we performed GO enrichment analysis on the eGenes identified in each cancer type (Supplementary Table 2). In CRC, the enrichment analysis revealed significant enrichment in the molecular function category *protein binding*

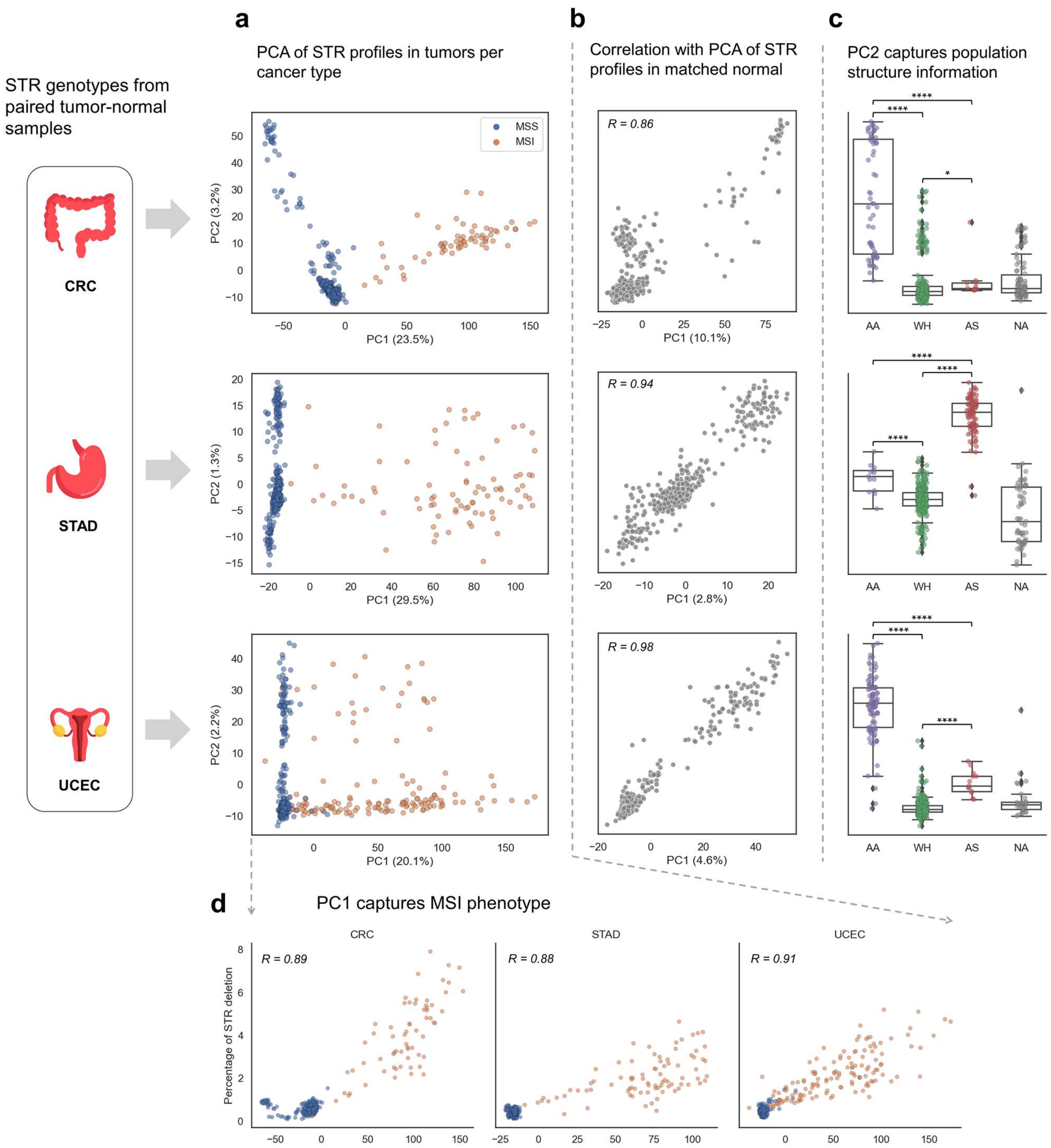

Figure 1. Tumor STR profiles capture MSI phenotype and population structure. (**a**) PCA plot of the first two PCs based on STR profiles from tumor samples per cancer type. The MSI status of each sample is indicated by colors. The numbers in brackets show the proportion of variance explained by each PC. (**b**) Pairwise comparisons of tumor PC2 and normal PC1 in three cancers, as shown by Pearson correlation coefficients ($P < 1e-10$). (**c**) Boxplots showing the distribution of tumor PC2 across population groups in each cancer type (AA: Black or African American; WH: white; AS: asian; NA: unknown). Two-sided Mann–Whitney–Wilcoxon tests were used to assess differences between population groups AA, WH, and AS. Significant differences are indicated by asterisks ($*$: $P < .05$, **** :$P< 1e-4$). (**a,b, and c**) share the same y-axes as shown in (**a**). The variance explained by normal PC2 is not shown in the correlation plot. See Supplementary Fig. S1 for detailed variance contributions of both PC1 and PC2 in normal samples. (**d**) Pairwise comparisons of tumor PC1 and the percentage of STR somatic deletion in each cancer type, as shown by Pearson correlation coefficients ($P < 1e-10$). Colon, stomach and uterus icons are from Freepik and licensed under the Freepik Free License (www.freepik.com).

(Fig. 3d). In particular, 74%, 71%, and 77% of the eGenes detected in CRC, STAD, and UCEC respectively, were enriched under the *protein binding* category. These findings align with the previous observation [59, 60], which reported that genes containing variable repeats were frequently involved in protein binding. Moreover, eGenes from CRC and STAD showed significant enrichment in the *Phosphoprotein* and *Acetylation* categories, whereas eGenes from UCEC were exclusively enriched in the *Acetylation* category. These distinct enrichment patterns suggested a potential regulatory mechanism where eSTRs may influence gene expression through posttranslational modifications (PTMs).

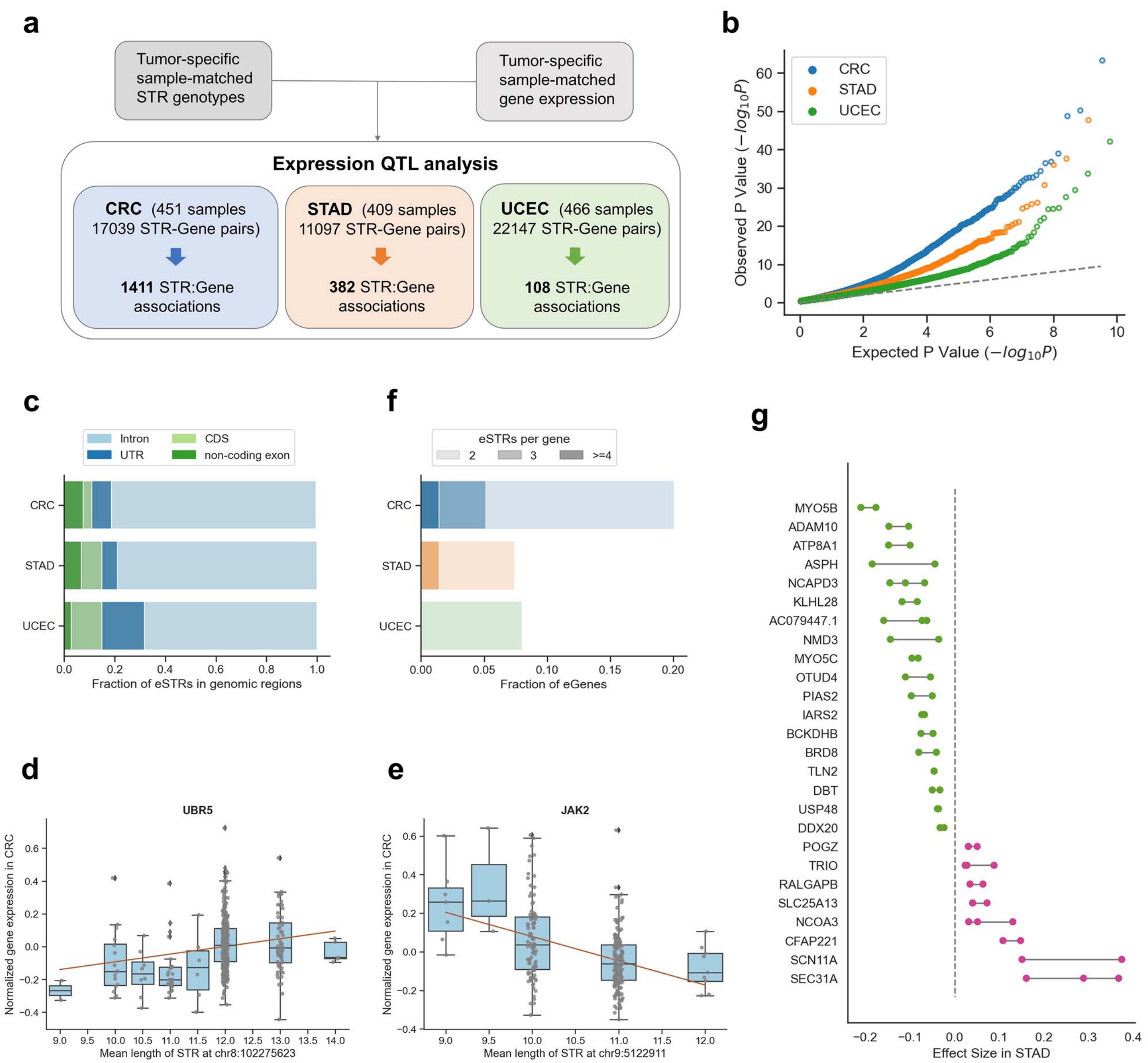

Figure 2. eSTR identification and characterization. (**a**) Schematic design of expression QTL analysis to identify eSTRs. We analyzed eSTRs using sample-matched gene expression and STR genotypes from WES data from TCGA CRC, STAD, and UCEC. (**b**) eSTR association results. The quantile–quantile plot compares observed *P*-values for each STR versus the expected uniform distribution for each tumor. (**c**) Barplots indicating fractions of eSTRs located within genomic regions (CDS, coding sequence; UTR, untranslated exon region) in CRC, STAD, and UCEC tumors, respectively. (**d**) Scatterplot showing an example of positive correlation between repeat length of eSTR in intronic region and UBR5 gene expression. (**e**) Scatterplot showing an example of negative correlation between repeat length of eSTR in intronic region and JAK2 gene expression. (**f**) Distribution of the number of eSTRs per gene, stratified by cancer type. Only genes with at least two identified eSTRs are shown in the plot. (**g**) Forest plot showing the consistent effects of multiple eSTRs on each gene. A total of 26 genes with multiple eSTRs in STAD tumors are shown in the plot. Each dot represents an eSTR, with the direction of effects shown in pink (positive correlations) or green (negative correlations).

eSTRs show higher mutability and regulate gene expression during tumorigenesis

We next sought to compare the somatic mutability of eSTRs and non-eSTRs using STR genotypes from paired samples. Due to the significant differences in STR mutability between MSI and MSS tumors, we assessed the mutability of eSTRs separately for MSI and MSS tumors. Similarly to our previous study [32], we categorized the STRs based on their repeat unit size and allele length and then compared their mutability in each category. In MSI CRC tumors, eSTRs exhibited higher mutability than non-eSTRs in 25 of the total 31 categories, significantly higher than expected under a null distribution generated through random permutations of eSTR labels (Fig. 4a, Supplementary Fig. S6). The trend of elevated mutability for eSTRs was consistent across nearly all MSI and MSS tumors, with the exception of MSS tumors in UCEC patients. Additionally, in both MSI and MSS patients, the mutability of eSTRs followed a descending trend from CRC, to STAD, and then to UCEC tumors.

To further explore the functional consequences of eSTR mutability on gene expression, we focused on a subset of patients with matched gene expression data from both tumor and normal tissues. For each mutated eSTR, we calculated the eSTR mutation length (Δx) and quantified the expression shift of the associated gene (Δy) between the matched normal and tumor samples.

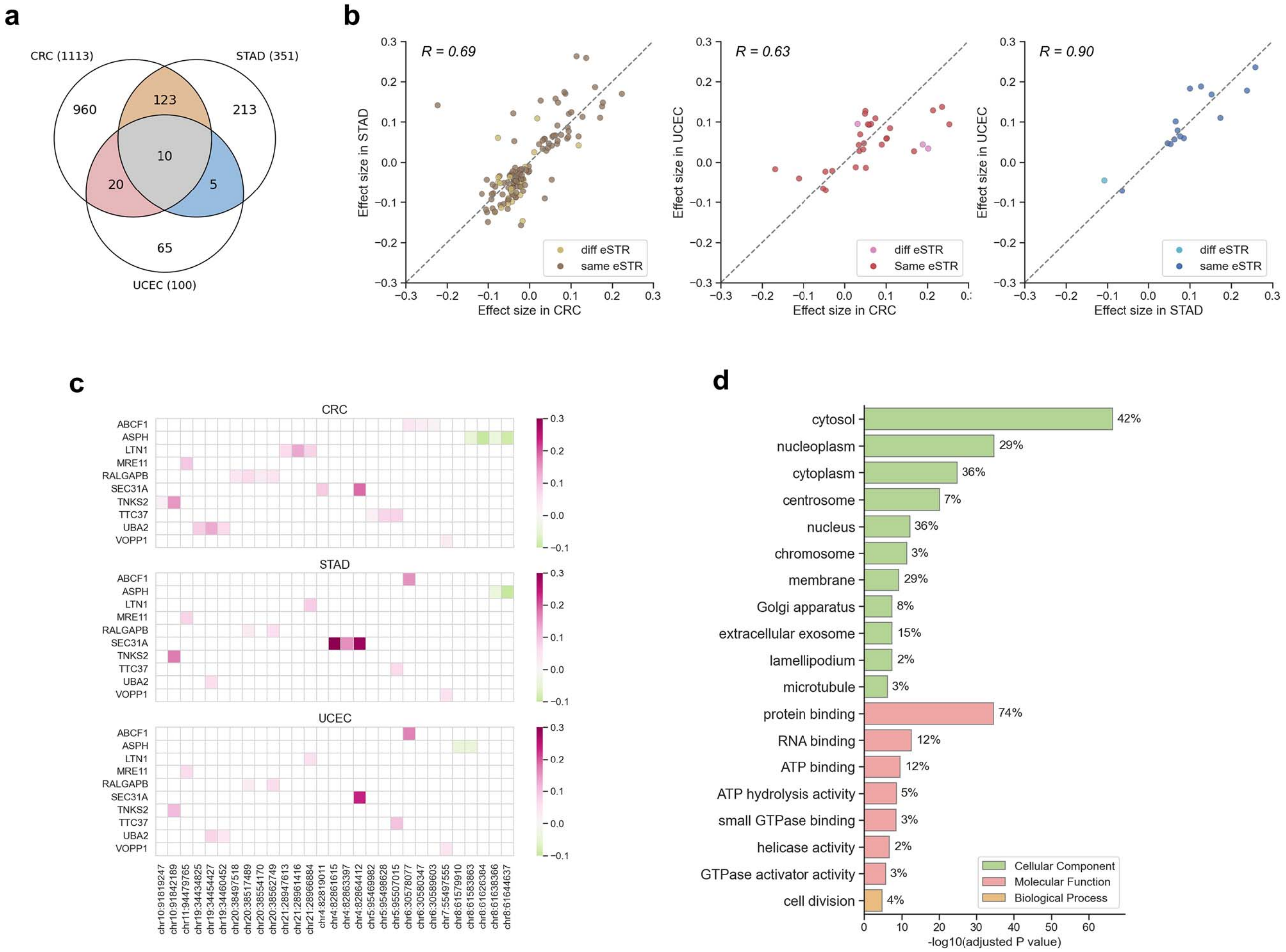

Figure 3. The effects of eSTRs on gene expression across cancers. (**a**) Venn diagram represents the overlap between eGenes in each cancer type. There are 1113, 351, and 100 unique eGenes detected in CRC, STAD, and UCEC, respectively. (**b**) Satterplots for comparison of effect sizes of overlapped eGenes between CRC and STAD, CRC and UCEC, and STAD and UCEC (from left to right). The lighter dots (diff eSTR) denote the eGenes are not associated with common eSTR. The darker dots (same eSTR) denote the eGenes are associated with at least one common eSTR. (**c**) Heatmap showing the effect sizes of 10 eGenes that are detected in all three cancer types. The purple cells represent positive effects, and the green cells represent negative effects. (**d**) Enriched GO of 1113 eGenes identified in CRC from GO enrichment analysis (*FDR* < 0.01).

We aimed to evaluate whether eSTR mutations could predict the direction and magnitude of corresponding gene expression changes during the transition from normal to tumor samples. The expected impact on gene expression changes was estimated by multiplying the eSTR mutation length (Δx) with the adjusted effect sizes (β') derived from our regression models. In general, the observed accuracy of predicting the direction of gene expression change was 55.6%, 57.5%, and 53.3% for CRC, STAD, and UCEC, respectively, exceeding the random expectation of 50%. To assess the significance of these accuracies, we performed permutation tests for each cancer type by generating a null distribution from permutations of the eSTR effect sizes (β'). The resulting *P*-values (Supplementary Fig. S7) indicated that the observed accuracies were significantly higher than expected under the null hypothesis for CRC and STAD. Furthermore, we grouped the eSTR mutations by the absolute predicted impacts ($|\beta' * \Delta x|$) and calculated the average mutation impact for each group. We observed that eSTR mutations with higher average impacts resulted in higher accuracy in predicting the direction of gene expression changes (Supplementary Fig. S8a). For eSTR mutations with accurate direction predictions, we next evaluated their ability to predict the magnitude of corresponding gene expression changes. We calculated the predictive score (see Methods) for each eSTR mutation and demonstrated that the overall predictive scores were significantly higher than those of the baseline models (see Methods; Supplementary Fig. S9). When we again grouped them based on their absolute expected impacts, we observed a significantly increasing trend between the average mutation impacts and predictive scores across all cancer types (Supplementary Fig. S8b).

Finally, we wondered whether eSTR mutation length might affect their predictions about gene expression changes. We grouped the eSTR mutations based on their mutation lengths (Δx) and tested for associations between eSTR mutation lengths and two metrics: the accuracy of predicting the direction of gene expression changes and the predictive score. Notably, we observed a significant positive correlation between eSTR mutation length and predictive score, but not with directional accuracy (Fig. 4b; Supplementary Fig. S8c). Overall, our results further validated the linear relationship between eSTR length and gene expression, highlighting the functional role of eSTRs in mediating gene regulatory changes during tumorigenesis. For smaller eSTR mutation lengths, the effect on gene expression may be minimal or masked by other regulatory mechanisms. However, as the eSTR mutation length increased, the influence of eSTR mutation on gene expression became more pronounced.

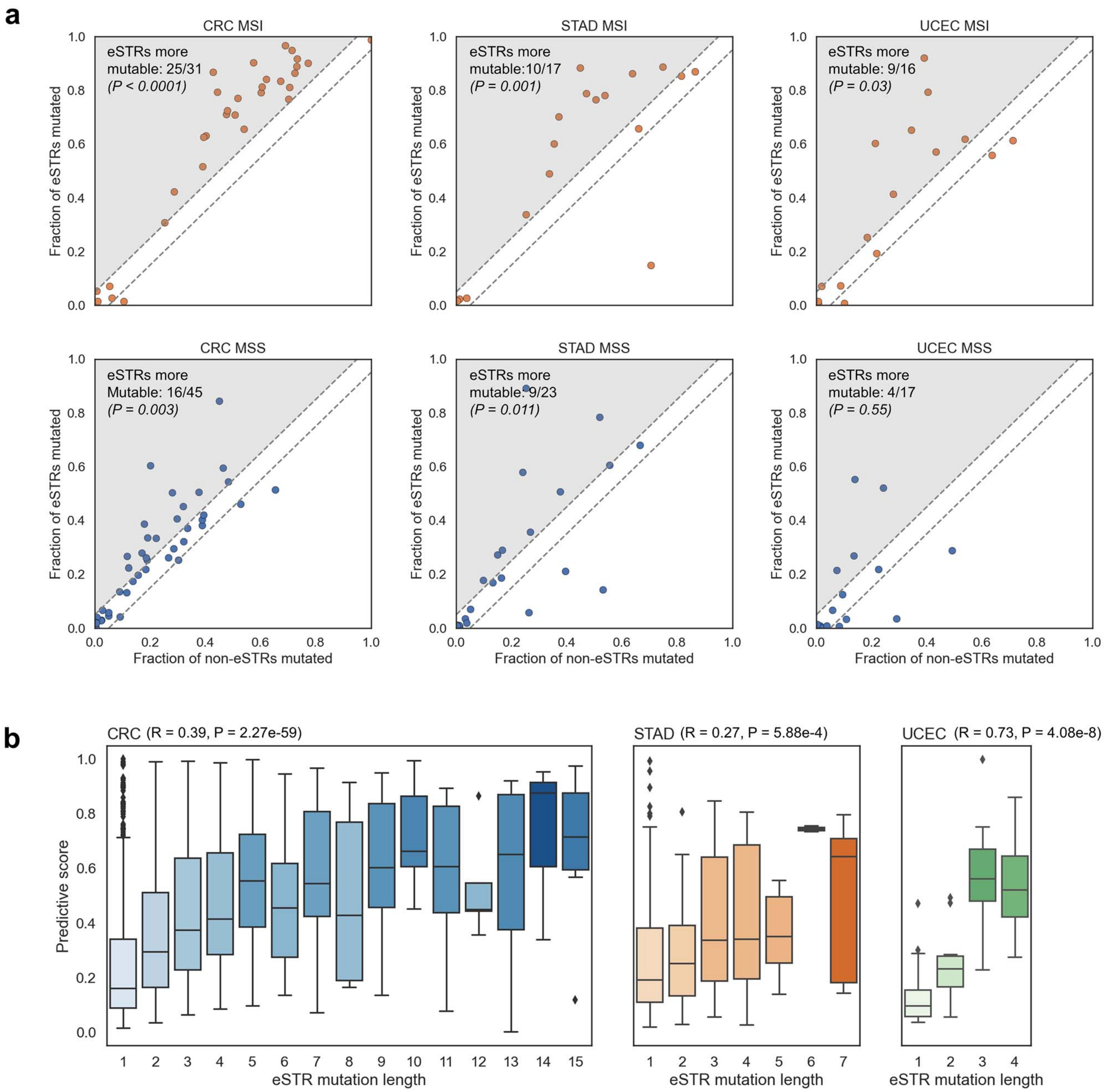

Figure 4. eSTRs have higher mutability and are predictive of gene expression change (**a**) Comparison of the mutability between eSTRs and non-eSTRs. The top row shows results for MSI patients and the bottom row shows results for MSS patients. Each dot in the scatterplot represents a STR category characterized by the STR unit size and allele length. The fraction of mutated non-eSTRs is shown on the x-axis, and the fraction of mutated eSTRs on the y-axis. The dots that fall between the dashed lines represent repeat types for which no difference in mutability between eSTRs and non-eSTRs was observed. For repeat types that fall in the shaded region, eSTRs were more mutable than non-eSTRs (their numbers are noted in the top left). *P*-values obtained from comparing the observed values with their respective null distributions using permutation tests are shown in the top left. (**b**) Boxplots showing the relationship between the eSTR mutation length and the predictive score of eSTR mutation prediction for gene expression change from normal to tumor samples. The x-axis represents different eSTR mutation lengths. Boxes are colored based on the prediction of gene expression change prediction. Larger mutation lengths lead to higher predictive score in predicting gene expression changes.

Further evidence for regulatory roles of eSTRs

MSI and MSS tumors exhibit distinct molecular features. MSI tumors are characterized by frequent point mutations and hypermethylation, whereas MSS tumors typically display high chromosomal instability [21, 23, 24, 61]. In addition to the variability of STRs, numerous other genetic and epigenetic factors may contribute to gene expression differences between MSS and MSI tumors. To provide further evidence of the regulatory role of eSTRs, we analyzed their effects in the context of the MSI phenotype. Specifically, we performed association analyses while conditioning the MSI phenotype of tumors for each cancer type. If the observed effects of eSTRs were solely attributable to the confounding influence of MSI, which can simultaneously affect both STR allele length and gene expression, then the conditioned effect of eSTRs should be randomly distributed compared with their unconditioned effects (Fig. 5b). Alternatively, if the effects of eSTRs were independent of the MSI phenotype, the direction of conditioned effects should align with the unconditioned effects (Fig. 5c). We considered eSTRs to be more likely causal regulators of gene expression if after conditioning, the direction of the associations stayed consistent and the associations remained statistically significant ($P < .05$). Using this approach, we found that 24.5%, 30.4%, and 63.9% of eSTR signals were independent of MSI phenotype in CRC, STAD, and UCEC, respectively (Fig. 5a, Supplementary Table S3).

Previous studies have indicated that single nucleotide polymorphisms (SNPs) adjacent to mononucleotide repeats often show increased variability [62], thus the signals attributed to eSTRs might be explained by nearby SNPs in linkage disequilibrium [13, 14]. To evaluate this, we leveraged the database PancanQTL [47], which identified SNPs functioning as eQTLs across diverse cancer types. We examined the overlap between STR-associated eGenes and SNP-associated genes from the PancanQTL database in each cancer type. Our analysis revealed that the majority of eGenes, 67.7%, 73.5%, and 90.0% in CRC, STAD, and UCEC, respectively, did not overlap with SNP-associated genes (Fig. 5d).

After filtering through both conditional and overlap analyses, 19.1%, 24.2%, and 60.0% of the original eGenes in CRC, STAD, and UCEC, respectively, were retained. Notably, the genes *ABCF1*, *SEC31A*, *TNKS2*, and *MRE11* still appeared in the overlap. Among these, *ABCF1*, *SEC31A*, and *TNKS2* were associated with coding-region eSTRs. Collectively, these results suggest that eSTRs linked to these eGenes were more likely to as as independent casual regulators of gene expression.

eSTRs are more likely to link to nearby DNA methylation

Functional STRs have previously been reported to influence nearby DNA methylation in the human genome [18, 19]. Additionally, the interactions between MSI and tumor DNA methylation have been studied in the pathogenesis of CRC [63]. Therefore, we sought to determine whether our eSTRs were enriched for associations with nearby DNA methylation. Using DNA methylation data from the TCGA database, we first mapped each STR locus to nearby CpG sites and then performed eSTR:CpG association analyses for each cancer type (see Methods). We identified that 84.9%, 81.3%, and 72.6% of eSTRs were also meSTRs (eSTRs that are associated with nearby DNA methylation) in CRC, STAD, and UCEC, respectively ($FDR < 0.05$) (Supplementary Table 4), suggesting a strong concordance between eSTRs and meSTRs. To assess the enrichment of eSTRs for DNA methylation associations, we performed a comparative analysis with non-eSTRs. From the larger collection of non-eSTRs, we generated 1000 background sets through random subsampling, ensuring each set matched the corresponding eSTR set in size. We then determined the percentage of mSTRs in which STR lengths were associated with DNA methylation levels of nearby CpG sites in these background sets, which served as null distributions. Notably, the proportion of meSTRs among eSTRs was two to three times higher than that observed in the null distributions (Fig. 6a), indicating a significant enrichment of DNA methylation associations for eSTRs.

Next, we investigated the functional characteristics of meSTRs to gain insight into their biological roles across the three cancer types. First, we examined effect-size biases in meSTR associations. Overall, meSTRs were more likely to show positive correlations between repeat length and DNA methylation in CRC and STAD (binomial one-sided $P < 1.0e^{-4}$), but not in UCEC (binomial one-sided $P = .92$). Additionally, we observed that 77.7%, 82.4%, and 59.7% of meSTRs in CRC, STAD, and UCEC, respectively, displayed opposing correlation directions between meSTR:gene and meSTR:CpG associations. In most cases, increased repeat length of a meSTR was associated with lower gene expression and higher DNA methylation levels at one or more nearby CpG sites. These inverse correlations suggest that these meSTRs may regulate gene expression by modulating local DNA methylation. For example, the length of an intronic eSTR was positively correlated with the expression of gene *ABCF1* and negatively correlated with DNA methylation levels at a nearby CpG site (Fig. 6c). To test this hypothesis more directly, we performed single-variant colocalization analysis between meSTR–CpG and meSTR–gene associations (Supplementary Note 4) [64]. Notably, ∼20% of meSTRs in each cancer type exhibited high posterior probabilities for a shared causal variant (PP.H4 > 0.8), supporting a model in which meSTR variations influence both DNA methylation and gene expression (Supplementary Fig. S10).

Finally, we examined the consistency of meSTR associations across cancer types. The effects of overlapping meSTR associations were highly correlated between CRC and STAD ($R = 0.87, P = 1.50e^{-54}$), CRC and UCEC ($R = 0.95, P = 1.33e^{-23}$), and STAD and UCEC ($R = 0.59, P = 2.68e^{-2}$) (Fig. 6b), similar to the relationship between overlapping eSTRs and gene expression. These consistent patterns further emphasized the robust effects of eSTRs on gene expression and highlighted potential shared regulatory mechanisms involving DNA methylation in these cancers.

Discussion

This study provides a comprehensive analysis of STR variations and their functional roles in CRC, STAD, and UCEC tumors. To our knowledge, this is the first study to analyze STRs across three distinct cancer types while integrating gene expression and DNA methylation data. This multi-cancer, multi-omics approach allowed us to uncover not only the functional impact of eSTRs on gene expression and DNA methylation, but also highlighted shared regulatory mechanisms through eSTRs and meSTRs conserved across cancers. These overlapping eSTRs likely represent a core set of key regulatory elements across cancers.

Our integrated analysis provided valuable insights into the regulatory roles of STRs. We identified 1383, 370, and 107 unique eSTRs in CRC, STAD, and UCEC tumors, respectively. We then used conditional analyses accounting for MSI phenotype and overlapping analyses with SNP-associated genes to investigate the causality of our putative eSTRs, which determined that 19.1%, 24.2%, and 60.0% of them are more likely to act as independent regulators of nearby gene expression. We further demonstrated that eSTRs show higher mutability and their mutation lengths are strongly associated with gene expression changes during the transition from normal to tumor tissues, particularly those with larger mutation lengths. These findings provided statistical evidence for the regulatory role of eSTRs and indicated that eSTR mutations may play a crucial role in driving gene expression alterations during tumorigenesis.

Our findings have several important implications for understanding the regulatory mechanisms of STRs on gene expression. First, by integrating DNA methylation data, we demonstrated that over 70% eSTRs were significantly associated with methylation levels of nearby CpG sites, forming what we termed meSTRs. Notably, the regulatory effects of overlapping eSTRs on gene expression were consistent across cancers. Similarly, overlapping meSTRs exhibited strong concordance in their influence on DNA methylation. These consistent patterns across cancers suggest that eSTRs may influence gene expression through epigenetic modulation in tumors, specifically by altering the methylation levels of adjacent CpG sites. Further colocalization analysis provided statistical evidence that ∼20% of meSTRs are likely

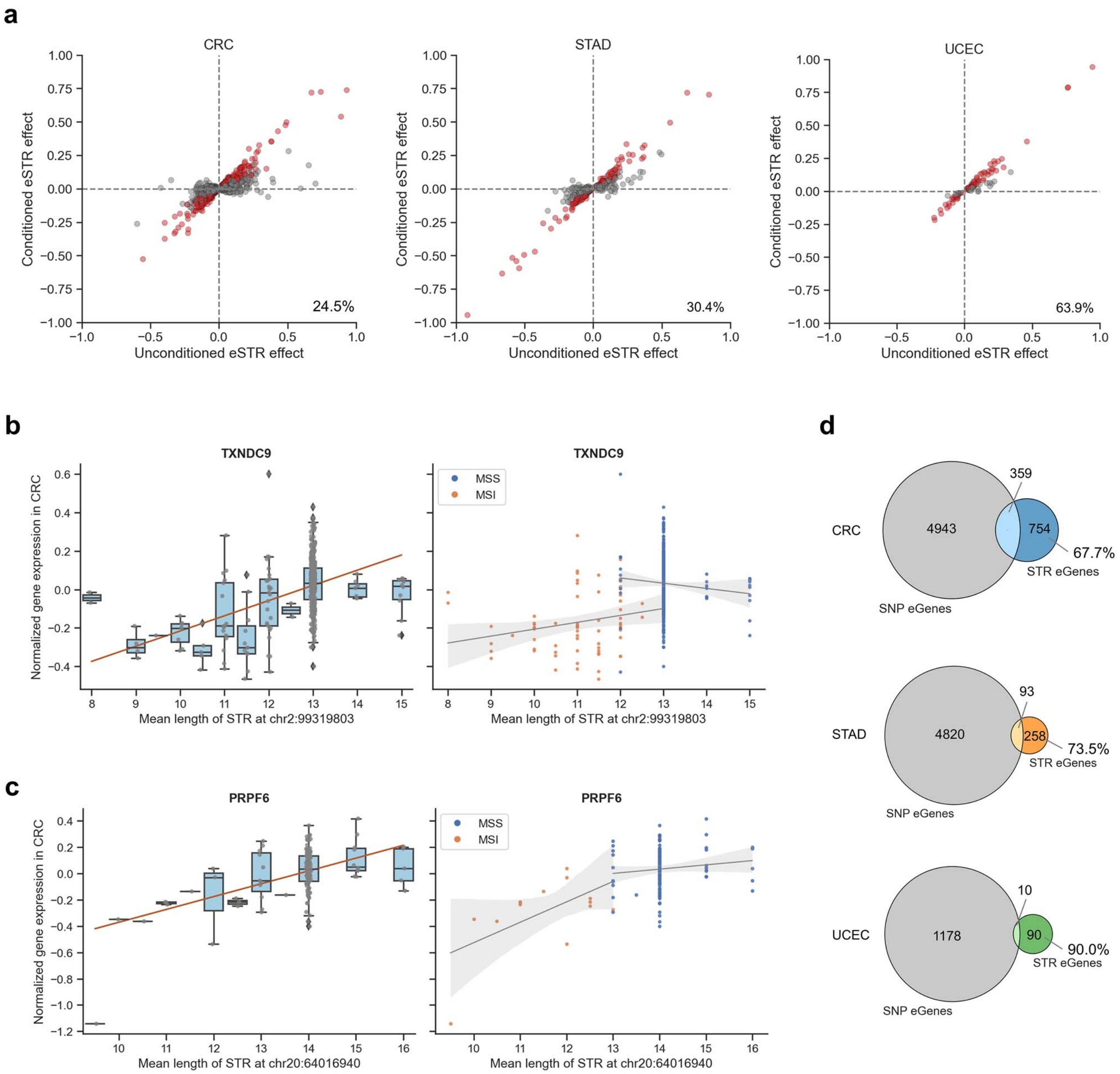

Figure 5. Further evidence for a regulatory role of eSTR. (**a**) Orignial eSTR effect versus conditioned eSTR effect. The red dots represent eSTRs whose directions of effect were consistent in both MSI tumors and MSS and whose associations remained significant (*P* < .05) upon conditioning for the MSI phenotype. The gray dots represent eSTRs that either became nonsignificant or showed opposite direction of effects after conditioning. (**b**) An example of STR:gene associations (TXNDC9), after conditioning on MSI phenotype, the direction of associations are different in MSS and MSI samples. (**c**) An example of STR:gene associations (PRPF6), after conditioning on MSI phenotype, the direction of associations remain the same and have statistical significance (*P* < .05). (**d**) Overlap of STR-associated eGenes and SNP-associated eGenes.

responsible for both methylation and expression changes, reinforcing their role in coordinated epigenetic regulation. Second, functional enrichment analyses of eGenes showed significant enrichment in PTMs. They are critical for modulating protein–protein interactions, protein stability, and localization in cancers [65]. The observed enrichment suggests that eSTRs may regulate gene expression by affecting PTM-related processes, such as transcriptional activation or repression, chromatin accessibility, and protein–protein interactions. While these mechanisms are likely primary drivers, additional multi-omics studies are needed to uncover potential interacting regulatory pathways.

Our study also revealed a dual role for coding-region eSTRs identified in all the three cancers. We observed that the expression of genes *ABCF1*, *LTN1*, *SEC31A*, and *TNKS2* was positively correlated with the allele length of eSTRs located in the coding regions. Among these genes, *SEC31A* has been identified as one of five genes encoding shared immunogenic frameshift mutations in colorectal, stomach, and endometrial MSI tumors [66]. Additionally, *LTN1* was among the six candidate genes that were predicted and validated to generate shared frameshift peptides with functionally confirmed immunogenicity in both colorectal and stomach MSI tumors [67]. Furthermore, *TGFBR2* and *SLC35F5*, which showed positive associations with eSTRs located in coding regions in both colorectal and stomach cancer, were also among these six candidate genes. Collectively, we hypothesize that eSTRs in coding regions may have a dual

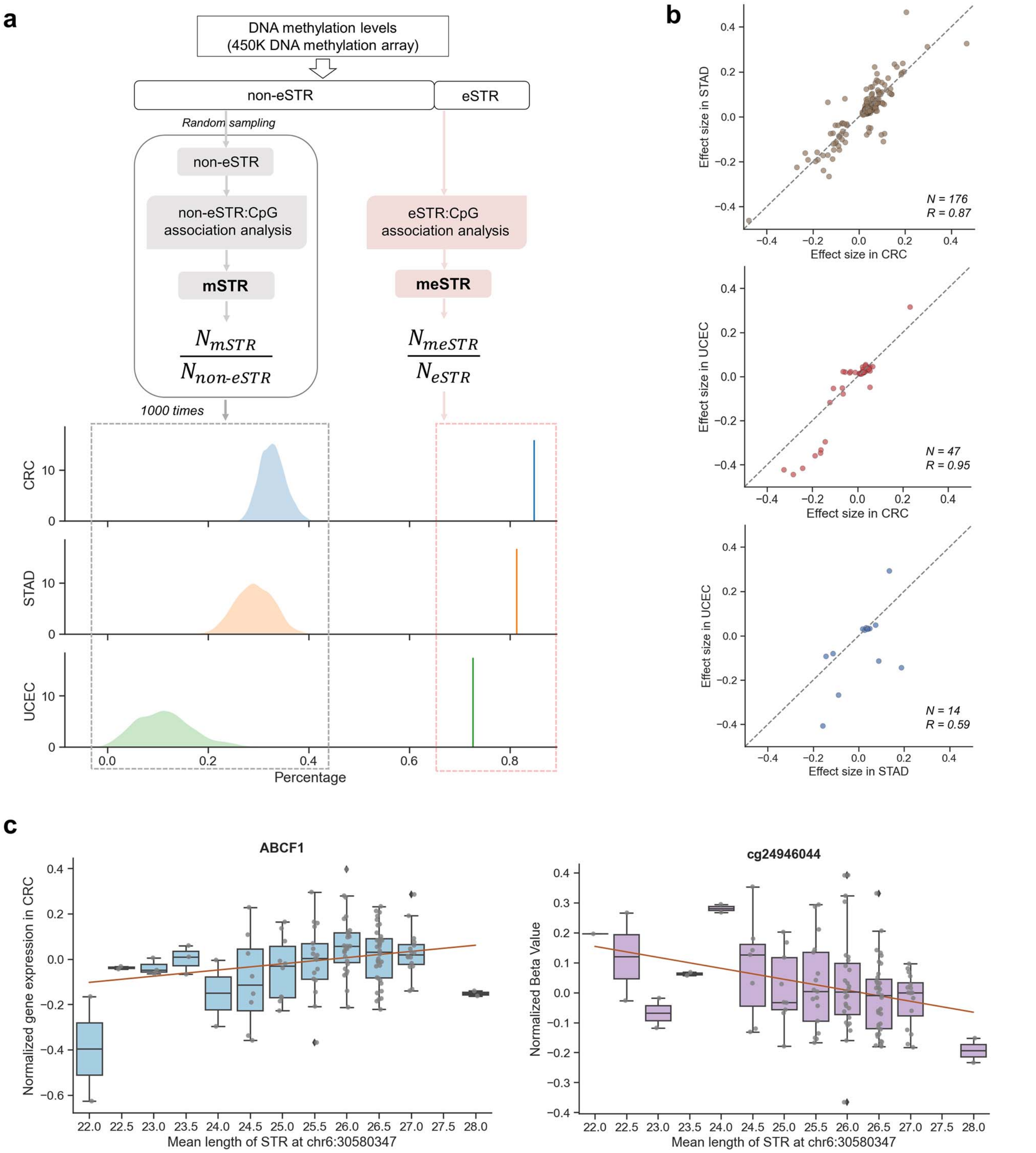

Figure 6. eSTRs are more likely to link to DNA methylation. (**a**) Illustration and comparison of mSTR/meSTR identification from non-eSTRs and eSTRs. To generate a null distribution, we randomly selected the amount of non-eSTRs as the same number of eSTRs per cancer type. The resulting distributions are shown in the left panel of the density plots. The proportions of meSTRs identified from eSTRs are shown in the right panel. (**b**) Comparison of effects of shared STR-CpG pairs between CRC and STAD, CRC and UCEC, and STAD and UCEC (from left to right). The dots denote the same meSTR-CpG pairs between two cancers. (**c**) Scatterplots showing an example of the opposite correlation between the mean length of an eSTR with the gene ABCF1 expression (left plot) and nearby DNA methylation at CpG site cg24946044 (right plot).

impact: (1) affecting gene expression and (2) generating tumor-specific neoantigens through frameshift mutations, particularly in MSI tumors. This dual role could have further diagnostic and therapeutic implications in multiple cancers.

Our study has several limitations. (1) STR data were genotyped using WES, which limited our analysis to a relatively small subset of STR loci and may have excluded potentially distal functional STRs. Future studies utilizing large-scale whole-genome or long-read sequencing data could enable a more comprehensive and extensive investigation of functional STR variations. (2) Our models assumed linear relationships between STR length and gene expression, but nonlinear effects may exist. Future analyses exploring alternative patterns could uncover additional regulatory complexity. (3) We established links between eSTRs and DNA methylation, but the mechanistic interactions between STR variation, DNA methylation, and gene expression remain unclear. Further studies incorporating multi-omics data and advanced models are needed to unravel the mechanisms underlying their interactions.

Conclusions

Overall, our integrated analyses of STR variations, gene expression, and DNA methylation provide valuable insights into the functional role of STRs among colorectal, stomach, and endometrial cancers. Our findings suggest that STRs may influence gene expression and DNA methylation through shared regulatory mechanisms in these cancers, highlighting their potential significance as key regulators in tumorigenesis. These results lay the groundwork for future studies to explore the clinical relevance of STRs as biomarkers or therapeutic targets.

Key points

- We conducted the first comprehensive study of exome-wide STR variations in colorectal, stomach, and endometrial cancers integrating gene expression and DNA methylation data, which revealed shared regulatory mechanisms of STRs across cancers.
- The expression QTL analysis identified eSTRs in each cancer type. Further analyses demonstrated that they have higher mutability and their mutation length strongly correlates with gene expression changes, particularly during normal-to-tumor transitions, suggesting a role in tumorigenesis.
- Coding region eSTRs exhibited positive correlations with nearby gene expression in all three cancers, including genes *ABCF1*, *LTN1*, *SEC31A*, and *TNKS2*.
- Through methylation QTL analysis, we found that over 70% of eSTRs were significantly associated with methylation levels of nearby CpG sites, forming meSTRs. The consistent regulatory effects of eSTRs and meSTRs across cancers suggest that STRs influence gene expression through epigenetic modulation.

Acknowledgments

We express our gratitude to the TCGA research network for generously sharing their invaluable data.

Author contributions

Feifei Xia (Formal analysis, Writing original draft, Investigation), Oxana Lundstrm (Data curation), Max Adriaan Verbiest (Formal analysis, Writing review & editing), Tugce Bilgin Sonay (Formal analysis, Writing review & editing), Michael Baudis (Formal analysis, Writing review & editing), and Maria Anisimova (Conceptualization, Formal analysis, Funding acquisition, Writing review & editing). All authors reviewed and approved the final version of the manuscript.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Competing interests

No competing interest is declared.

Funding

F.X., M.A.V., and O.L. were supported by SNSF Sinergia grant CRSII5_193832 to MA. This work was also supported by the EU Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 823886.

Data and code availability

The analyses were performed using Python 3.10.4 and R 4.1.3. Scripts for analyses and reproducing figures are available here: https://github.com/acg-team/multicancer_STR. Data from the COAD, READ, STAD, and UCEC cohorts were downloaded from the GDC knowledge base (data release version 38.0). Summary statistics of STRs in the study are available in the database WebSTR [68]. Access to restricted TCGA data was granted under dbGaP study phs000178.v11.p8.

References

1. Ellegren H. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* 2004;**5**:435–45. https://doi.org/10.1038/nrg1348
2. Tóth G, Gáspári Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res* 2000;**10**:967–81. https://doi.org/10.1101/gr.10.7.967
3. Fan H, Chu JY. A brief review of short tandem repeat mutation. *Genomics Proteomics Bioinformatics* 2007;**5**:7–14. https://doi.org/10.1016/S1672-0229(07)60009-6
4. Sun JX, Helgason A, Masson G. *et al.* A direct characterization of human mutation based on microsatellites. *Nat Genet* 2012;**44**:1161–5. https://doi.org/10.1038/ng.2398
5. Willems T, Gymrek M, Highnam G. *et al.* The landscape of human STR variation. *Genome Res* 2014;**24**:1894–904. https://doi.org/10.1101/gr.177774.114
6. Redelings BD, Holmes I, Lunter G. *et al.* Insertions and deletions: computational methods, evolutionary dynamics, and biological applications. *Mol Biol Evol* 2024;**41**:msae177. https://doi.org/10.1093/molbev/msae177
7. Mirkin SM. Expandable DNA repeats and human disease. *Nature* 2007;**447**:932–40. https://doi.org/10.1038/nature05977
8. Gymrek M. A genomic view of short tandem repeats. *Curr Opin Genet Dev* 2017;**44**:9–16. https://doi.org/10.1016/j.gde.2017.01.012
9. Uguen K, Michaud JL, Génin E. Short tandem repeats in the era of next-generation sequencing: from historical loci to population databases. *Eur J Hum Genet* 2024;**32**:1037–44. https://doi.org/10.1038/s41431-024-01666-z
10. Liu XS, Wu H, Krzisch M. *et al.* Rescue of fragile X syndrome neurons by DNA methylation editing of the FMR1 gene. *Cell* 2018;**172**:979–992.e6. https://doi.org/10.1016/j.cell.2018.01.012
11. Raveh-Sadka T, Levo M, Shabi U. *et al.* Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat Genet* 2012;**44**:743–50. https://doi.org/10.1038/ng.2305
12. Afek A, Schipper JL, Horton J. *et al.* From the cover: protein-DNA binding in the absence of specific base-pair recognition. *Proc Natl Acad Sci USA* 2014;**111**:17140–5. https://doi.org/10.1073/pnas.1410569111

13. Gymrek M, Willems T, Guilmatre A. *et al.* Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat Genet* 2016;**48**:22–9. https://doi.org/10.1038/ng.3461
14. Fotsing SF, Margoliash J, Wang C. *et al.* The impact of short tandem repeat variation on gene expression. *Nat Genet* 2019;**51**: 1652–9. https://doi.org/10.1038/s41588-019-0521-9
15. Jakubosky D, D'Antonio M, Bonder MJ. *et al.* Properties of structural variants and short tandem repeats associated with gene expression and complex traits. *Nat Commun* 2020;**11**:2927. https://doi.org/10.1038/s41467-020-16482-4
16. Shi Y, Niu Y, Zhang P. *et al.* Characterization of genome-wide STR variation in 6487 human genomes. *Nat Commun* 2023;**14**:1–18. https://doi.org/10.1038/s41467-023-37690-8
17. Horton CA, Alexandari AM, Hayes MGB. *et al.* Short tandem repeats bind transcription factors to tune eukaryotic gene expression. *Science* 2023;**381**:eadd1250. https://doi.org/10.1126/science.add1250
18. Quilez J, Guilmatre A, Garg P. *et al.* Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Res* 2016;**44**:3750–62. https://doi.org/10.1093/nar/gkw219
19. Martin-Trujillo A, Garg P, Patel N. *et al.* Genome-wide evaluation of the effect of short tandem repeat variation on local DNA methylation. *Genome Res* 2023;**33**:184–96. https://doi.org/10.1101/gr.277057.122
20. Hamanaka K, Yamauchi D, Koshimizu E. *et al.* Genome-wide identification of tandem repeats associated with splicing variation across 49 tissues in humans. *Genome Res* 2023;**33**:435–47. https://doi.org/10.1101/gr.277335.122
21. Li K, Luo H, Huang L. *et al.* Microsatellite instability: a review of what the oncologist should know. *Cancer Cell Int* 2020;**20**:16. https://doi.org/10.1186/s12935-019-1091-8
22. Boland CR, Goel A. Microsatellite instability in colorectal cancer. *Gastroenterology* 2010;**138**:2073–2087.e3. https://doi.org/10.1053/j.gastro.2009.12.064
23. Puliga E, Corso S, Pietrantonio F. *et al.* Microsatellite instability in gastric cancer: between lights and shadows. *Cancer Treat Rev* 2021;**95**:102175. https://doi.org/10.1016/j.ctrv.2021.102175
24. Kanopiene D, Vidugiriene J, Valuckas KP. *et al.* Endometrial cancer and microsatellite instability status. *Open Medicine* 2014;**10**: 70–6. https://doi.org/10.1515/med-2015-0005
25. Woerner SM, Yuan YP, Benner A. *et al.* SelTar base, a database of human mononucleotide-microsatellite mutations and their potential impact to tumorigenesis and immunology. *Nucleic Acids Res* 2010;**38**:D682–9. https://doi.org/10.1093/nar/gkp839
26. Turajlic S, Litchfield K, Xu H. *et al.* Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol* 2017;**18**:1009–21. https://doi.org/10.1016/S1470-2045(17)30516-8
27. Kim TM, Laird PW, Park PJ. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. *Cell* 2013;**155**:858–68. https://doi.org/10.1016/j.cell.2013.10.015
28. Salipante SJ, Scroggins SM, Hampel HL. *et al.* Microsatellite instability detection by next generation sequencing. *Clin Chem* 2014;**60**:1192–9. https://doi.org/10.1373/clinchem.2014.223677
29. Bonneville R, Krook MA, Kautto EA. *et al.* Landscape of microsatellite instability across 39 cancer types. *JCO Precis Oncol* 2017;**1**:1–15. https://doi.org/10.1200/PO.17.00073
30. Cortes-Ciriano I, Lee S, Park WY. *et al.* A molecular portrait of microsatellite instability across multiple cancers. *Nat Commun* 2017;**8**:15180. https://doi.org/10.1038/ncomms15180
31. Fujimoto A, Fujita M, Hasegawa T. *et al.* Comprehensive analysis of indels in whole-genome microsatellite regions and microsatellite instability across 21 cancer types. *Genome Res* 2020;**30**:334–46. https://doi.org/10.1101/gr.255026.119
32. Verbiest MA, Lundström O, Xia F. *et al.* Short tandem repeat mutations regulate gene expression in colorectal cancer. *Sci Rep* 2024;**14**:1–11. https://doi.org/10.1038/s41598-024-53739-0
33. Lakshminarasimhan R, Liang G. The role of DNA methylation in cancer. *Adv Exp Med Biol* 2016;**945**:151. https://doi.org/10.1007/978-3-319-43624-1_7
34. Ashktorab H, Brim H. DNA methylation and colorectal cancer. *Current colorectal cancer reports* 2014;**10**:425–30. https://doi.org/10.1007/s11888-014-0245-2
35. Zeng Y, Rong H, Xu J. *et al.* DNA methylation: an important biomarker and therapeutic target for gastric cancer. *Front Genet* 2022;**13**. https://doi.org/10.3389/fgene.2022.823905
36. Dhar GA, Saha S, Mitra P. *et al.* DNA methylation and regulation of gene expression: Guardian of our health. *The Nucleus* 2021;**64**: 259–70. https://doi.org/10.1007/s13237-021-00367-y
37. Cunningham JM, Christensen ER, Tester DJ. *et al.* Hypermethylation of the hMLH1 promoter in colon cancer with microsatellite instability. *Cancer Res* 1998;**58**:3455–60.
38. Issa JP. CpG Island methylator phenotype in cancer. *Nat Rev Cancer* 2004;**4**:988–93. https://doi.org/10.1038/nrc1507
39. Tomlinson I, Halford S, Aaltonen L. *et al.* Does MSI-low exist? *J Pathol* 2002;**197**:6–13. https://doi.org/10.1002/path.1071
40. Ogino S, Goel A. Molecular classification and correlates in colorectal cancer. *J Mol Diagn* 2008;**10**:13–27. https://doi.org/10.2353/jmoldx.2008.070082
41. Mousavi N, Shleizer-Burko S, Yanicky R. *et al.* Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res* 2019;**47**:e90. https://doi.org/10.1093/nar/gkz501
42. Pedregosa F, Varoquaux G, Gramfort A. *et al.* Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;**12**:2825–30.
43. Zhou HJ, Li L, Li Y. *et al.* PCA outperforms popular hidden variable inference methods for molecular QTL mapping. *Genome Biol* 2022;**23**:1–17. https://doi.org/10.1186/s13059-022-02761-4
44. Carrot-Zhang J, Chambwe N, Damrauer JS. *et al.* Comprehensive analysis of genetic ancestry and its molecular correlates in cancer. *Cancer Cell* 2020;**37**:639–654.e6. https://doi.org/10.1016/j.ccell.2020.04.012
45. Seabold S, Perktold J. Statsmodels: econometric and statistical Modeling with python. In: van der Walt S, Millman J, (eds), *Proceedings of the 9th Python in Science Conference; 2010 Jun 28–Jul 3; Austin, TX*. Austin (TX): SciPy; 2010. p. 92–6.
46. Sherman BT, Hao M, Qiu J. *et al.* DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res* 2022;**50**:W216–21. https://doi.org/10.1093/nar/gkac194
47. Gong J, Mei S, Liu C. *et al.* PancanQTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types. *Nucleic Acids Res* 2018;**46**:D971–6. https://doi.org/10.1093/nar/gkx861
48. Ya C, Lemire M, Choufani S. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* 2013;**8**:203–9. https://doi.org/10.4161/epi.23470
49. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2022.
50. Aryee MJ, Jaffe AE, Corrada-Bravo H. *et al.* Minfi: a flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* 2014;**30**: 1363–9. https://doi.org/10.1093/bioinformatics/btu049
51. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**:841–2. https://doi.org/10.1093/bioinformatics/btq033

52. Haghighi MM, Javadi GR, Parivar K. *et al.* Frequent MSI mononucleotide markers for diagnosis of hereditary nonpolyposis colorectal cancer. *Asian Pac J Cancer Prev* 2010;**11**:1033–5.
53. Wang J, Zhao X, Jin L. *et al.* UBR5 contributes to colorectal cancer progression by destabilizing the tumor suppressor ECRG4. *Dig Dis Sci* 2017;**62**:2781–9. https://doi.org/10.1007/s10620-017-4732-6
54. Xie Z, Liang H, Wang J. *et al.* Significance of the E3 ubiquitin protein UBR5 as an oncogene and a prognostic biomarker in colorectal cancer. *Oncotarget* 2017;**8**:108079–108092. https://doi.org/10.18632/oncotarget.22531
55. Du W, Hong J, Wang YC. *et al.* Inhibition of JAK2/STAT3 signalling induces colorectal cancer cell apoptosis via mitochondrial pathway. *J Cell Mol Med* 2012;**16**:1878–88. https://doi.org/10.1111/j.1582-4934.2011.01483.x
56. Park SY, Lee CJ, Choi JH. *et al.* The JAK2/STAT3/CCND2 Axis promotes colorectal cancer stem cell persistence and radioresistance. *J Exp Clin Cancer Res* 2019;**38**:1–18. https://doi.org/10.1186/s13046-019-1405-7
57. Lamkin M, Gymrek M. The emerging role of tandem repeats in complex traits. *Nat Rev Genet* 2024;**25**:452–3. https://doi.org/10.1038/s41576-024-00736-8
58. Mandal R, Samstein RM, Lee KW. *et al.* Genetic diversity of tumors with mismatch repair deficiency influences anti–PD-1 immunotherapy response. *Science* 2019;**364**:485–91. https://doi.org/10.1126/science.aau0447
59. O'Dushlaine CT, Edwards RJ, Park SD. *et al.* Tandem repeat copy-number variation in protein-coding regions of human genes. *Genome Biol* 2005;**6**:1–12. https://doi.org/10.1186/gb-2005-6-8-r69
60. Legendre M, Pochet N, Pak T. *et al.* Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Res* 2007;**17**:1787–96. https://doi.org/10.1101/gr.6554007
61. Vilar E, Gruber SB. Microsatellite instability in colorectal cancer—the stable evidence. *Nat Rev Clin Oncol* 2010;**7**:153–62. https://doi.org/10.1038/nrclinonc.2009.237
62. Siddle KJ, Goodship JA, Keavney B. *et al.* Bases adjacent to mononucleotide repeats show an increased single nucleotide polymorphism frequency in the human genome. *Bioinformatics* 2011;**27**:895–8. https://doi.org/10.1093/bioinformatics/btr067
63. Jasmine F, Haq Z, Kamal M. *et al.* Interaction between microsatellite instability (MSI) and tumor DNA methylation in the pathogenesis of colorectal carcinoma. *Cancer* 2021;**13**:4956. https://doi.org/10.3390/cancers13194956
64. Giambartolomei C, Vukcevic D, Schadt EE. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* 2014;**10**:e1004383. https://doi.org/10.1371/journal.pgen.1004383
65. Geffen Y, Anand S, Akiyama Y. *et al.* Pan-cancer analysis of post-translational modifications reveals shared patterns of protein regulation. *Cell* 2023;**186**:3945–3967.e26. https://doi.org/10.1016/j.cell.2023.07.013
66. Roudko V, Bozkus CC, Orfanelli T. *et al.* Shared immunogenic poly-epitope frameshift mutations in microsatellite unstable tumors. *Cell* 2020;**183**:1634–1649.e17. https://doi.org/10.1016/j.cell.2020.11.004
67. Ballhausen A, Przybilla MJ, Jendrusch M. *et al.* The shared frameshift mutation landscape of microsatellite-unstable cancers suggests immunoediting during tumor evolution. *Nat Commun* 2020;**11**:4740. https://doi.org/10.1038/s41467-020-18514-5
68. Lundström OS, Verbiest MA, Xia F. *et al.* WebSTR: a population-wide database of short tandem repeat variation in humans. *J Mol Biol* 2023;**435**:168260. https://doi.org/10.1016/j.jmb.2023.168260

Briefings in Bioinformatics, 2025, **26(3)**, bbaf219
https://doi.org/10.1093/bib/bbaf219
Case Study