nature communications

Article | Article in Press

https://doi.org/10.1038/s41467-026-76741-8

Alternative tandem transcription initiation links noncoding variants to human disease through translational control

Received: 16 November 2025

Accepted: 5 August 2026

Published online: 20 August 2026

Check for updates

Xudong Zou[1,7], Wei Wang[2,7], Yinuo Wang[1,7], Hui Chen[1,3,7], Xing Li[1], Shuxin Chen[1], Feihong Weng[4], Yangcan Li[5], Qin Li[6], Chen Yu[2] ✉ & Lei Li[1] ✉

Alternative tandem transcription initiation is a pervasive mechanism of gene regulation, yet its genetic impact on human disease remains largely unknown. Here, we systematically quantify the genetic regulation of alternative tandem transcription initiation across 25,859 samples from 49 human normal tissues and 33 tumor tissues. We identify approximately 0.4 million genetic variants associated with alternative transcription initiation in 5295 genes, with 32% operating independently of gene expression. Moreover, we discover 2238 multi-tissue alternative tandem transcription initiation outliers enriched for rare deleterious promoter and 5′ UTR variants, demonstrating that both common and rare variants modulate transcription initiation. Strikingly, 74% of disease variants that colocalize with genetic variants regulating alternative transcription initiation cannot be identified through expression quantitative trait loci. Transcriptome-wide association studies identify 614 disease susceptibility genes associated with alternative transcription initiation, including known cancer drivers such as *MAFF* and *MLLT10*. Functional validation uncovers *OSGEP* as a breast cancer risk gene, where the alternative allele lengthens the 5′ UTR and reduces protein abundance through upstream open reading frame-mediated translation repression, and suppresses breast cancer cell proliferation. Our findings establish alternative transcription initiation as a major, underappreciated mechanism associating noncoding variation to disease, providing a critical resource for interpreting disease risk loci.

Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with human traits and complex diseases[1,2]. However, elucidating the underlying molecular mechanisms remains a significant challenge as most disease-associated variants lie within noncoding regions[3]. One prevailing hypothesis is that many noncoding variants contribute to traits by modulating gene expression or alternative splicing events[4,5]. Therefore, molecular phenotypes, such as expression quantitative trait loci (eQTLs), have provided mechanistic insights into disease-associated loci. Nevertheless, a substantial proportion of GWAS signals remain without molecular interpretation, suggesting that additional layers of regulatory variation have yet to be explored.

Alternative transcription initiation represents a highly regulated and critical step in gene expression, constituting a significant source of

[1]Institute of Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen, China. [2]Institute of Cancer Research, Shenzhen Bay Laboratory, Shenzhen, China. [3]Department of Physiological Science, School of Medicine and Health Sciences, University of Barcelona, Barcelona, Spain. [4]Liver Transplant Center, Organ Transplant Center, West China Hospital, Sichuan University, Chengdu, China. [5]Department of Pathology, the Affiliated Changsha Hospital of Xiangya School of Medicine, Central South University, Changsha, China. [6]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA. [7]These authors contributed equally: Xudong Zou, Wei Wang, Yinuo Wang, Hui Chen. ✉e-mail: yu@szbl.ac.cn; lei.li@szbl.ac.cn

transcriptomic and proteomic diversity independent of alternative splicing and polyadenylation[6–9]. More than 50% of human protein-coding genes harbor multiple transcription start sites[10,11]. Systematic 5′ -end mapping efforts, such as cap analysis of gene expression (CAGE), have revealed that 76% of detected TSSs reside within 5′ UTRs[12]. Transcription initiation from alternative promoters typically exhibits two distinct patterns: alternative first exons (AFE), in which the initiations occur from distant transcription TSSs, and alternative tandem TSSs (ATTSS), in which initiation sites differ at their 5′ end within the same 5′ UTR. These alternative 5′ UTRs can alter key regulatory elements, including upstream open reading frames (uORFs) and CpG islands, thereby influencing mRNA stability and translational efficiency[13]. In contrast, usage of AFEs often alters coding sequences or triggers nonsense-mediated decay, leading to truncated or unstable proteins. Both mechanisms are crucial for fine-tuning gene expression during development and are frequently dysregulated in disease[14,15]. For example, CAGE sequencing analysis of mRNA promoter dynamics across embryonic developmental stages has revealed numerous instances of TSS alterations, including the shifting of TSS positions within the promoter region of cyclin D1 (*CCND1*) gene from the maternal to the zygotic stage[16]. Another example is the dynamic regulation of alternative promoters in brain-derived neurotrophic factor (*BDNF*), which harbors nine functional promoters that exhibit differential activities across various cell types[17]. In addition, widespread promoter alterations have been observed across diverse tumor types, suggesting that alternative TSS usage may contribute to oncogenic processes[14,18]. Despite accumulating evidence for dynamic alternative transcription initiation in various biological and pathological contexts, the genetic basis governing alternative tandem transcription initiation (ATTSS) in human tissues remains largely uncharacterized.

Here, we present a comprehensive atlas of genetic regulation of alternative transcription initiation across diverse human tissues. Using our DATTSS algorithm[19], we quantified the distal TSS usage of detected ATTSS events from 25,859 RNA-seq datasets with matched genotype data spanning 49 normal and 33 tumor tissues. We identified approximately 0.40 million common *cis*-acting genetic variants associated with distal TSS usage across 6829 ATTSS events in 5295 genes. Notably, 32.1% of these 5′aQTLs showed no overlap with eQTLs. Further integration with GWAS summary statistics revealed significant enrichment of 5′aQTLs in several disease risk loci, with ~7.6% of GWAS variants colocalizing with 5′aQTLs. Furthermore, 5′aQTL-based transcriptome-wide association analyses (5′aTWAS) identified 614 susceptibility genes, including known oncogenic drivers and gene candidates, such as *OSGEP*. We experimentally validated that alternative 5′aQTL alleles lengthen the 5′ UTR in *OSGEP*, reducing OSGEP protein levels independent of mRNA level and thereby impacting breast cancer cell proliferation. Overall, our findings reveal a regulatory mechanism by which noncoding genetic variation is linked to human disease through alternative transcription initiation.

Results

A multi-tissue atlas of genetic regulation of alternative tandem transcription initiation

To systematically quantify alternative tandem transcription initiation (ATTSS) across diverse human tissues, we first curated a comprehensive reference set of transcription start sites by aggregating from GENCODE (v44), RefSeq (v200), and CAGE data from FANTOM5 and the ENCODE[20] (see "Methods"), yielding 27,792 transcripts from 17,078 genes annotated with multiple TSSs within their 5′ UTRs. We then applied our DATTSS[19] algorithm, which incorporates these reference TSSs to quantify the relative usage of the distal TSS for each transcript, expressed as the distal TSS usage index (DTUI), which represents the relative expression contribution of upstream (distal) TSS compared to the remaining (proximal) TSSs within each transcript. Biologically, this index reflects alternative promoter activation, which results in RNA isoforms with different 5′ UTR lengths. Given the enrichment of regulatory elements such as upstream open reading frames (uORFs) and RNA structural components, change of 5′ UTR length may impact gene expression regulation by affecting transcript stability or translational efficiency. We analyzed 25,859 RNA-seq datasets from 49 normal tissues and 33 tumor types (Fig. 1a). Across normal tissues, we identified 26,345 ATTSS events in 13,550 genes, representing approximately 79% of all multi-TSS genes in our reference set. In tumor samples, 25,355 ATTSS events were detected in 13,457 genes. On average, each normal tissue and each tumor type exhibited 9249 and 7799 alternative TSS events, respectively (Supplementary Fig. 1a, b). Hierarchical clustering analysis based on DTUI values demonstrated that samples grouped tightly by tissue and cancer type (Supplementary Fig. 1c, d), confirming that distal TSS usage patterns are biologically informative.

To identify genetic variants associated with alternative tandem transcription initiation, we performed 5′ UTR alternative TSS quantitative trait locus (5′aQTL) analysis. After quantile normalizing DTUI values and correcting for known covariates, including sex, sequencing platform, PCR (Polymerase Chain Reaction), and hidden technical covariates inferred using probabilistic estimation of expression residual (PEER)[21] factors (Supplementary Fig. 2a, b; detailed correlation between PEER factors and known technical covariates for all 49 tissues were listed in Supplementary Data 1), we used QTLTools (v1.3.1)[22] to identify common genetic variants (MAF > 1%) associated with differential usage of normalized DTUI values in each tissue or cancer type. Using a false-discovery rate (FDR) threshold of 5%, we identified in total approximately 0.27 million variants associated with DTUI of 4679 genes across 49 normal tissues, representing approximately 47.2% of all expressed genes with multiple TSSs (Fig. 1b and Supplementary Fig. 3a). In addition, we identified 0.13 million genetic variants associated with DTUI of 3818 genes in 28 tumor types (Supplementary Fig. 3a, b). On average, 36.3% of these genes across tumor types were not identified in matched GTEx tissues (Supplementary Fig. 3c). Notably, the proportion of 5′aGenes identified only in TCGA was moderately correlated with sample size (Supplementary Fig. 3d), suggesting that differences in statistical power may partially contribute to the observed TCGA-specific fraction. This finding suggests that incorporating TCGA samples substantially expands the discovery of *cis*-regulatory variants controlling alternative TSS selection. Significant variant-DTUI associations (FDR < 5%) were defined as 5′aQTLs, with the corresponding genes termed 5′aGenes. Many of these 5′ aGenes were associated with important disease-related genes. For example, we identified significant 5′aQTLs for *STAT6* (signal transducer and activator of transcription 6), a key transcription factor required for the IL-4-mediated biological response[23], and *FES* (tyrosine-protein kinase Fes/Fps), a known proto-oncogene[24] (Supplementary Fig. 4a, b). Moreover, the 5′ UTR length of translocase of inner mitochondrial membrane (*TIMM13*), previously shown to use a shorter 5′ UTR in lung cancer samples compared with normal tissues[19], was significantly modulated by a 5′aQTL located in its 5′ UTR (rs4806847, A/G; $P = 3.0 \times 10^{-4}$) (Fig. 1c).

We next explored the genetic architecture of ATTSS regulation by estimating heritability and performing genetic fine-mapping. Using a linear mixed model implemented in the GCTA[25], we estimated the heritability of the ATTSS variation attributable to 5′aQTLs within a 1-megabase (Mb) *cis* region. We observed that 5′aQTLs can explain, on average, 10.1% of ATTSS variations (Fig. 1d), ranging from 8.6% to 31.9% of ATTSS variations in individual tissues, and specifically, over 40% of ATTSS variations in 24 genes. For example, 43.5% of the variation in alternative tandem TSS usage of *TMEM45A* can be attributed to 5′ aQTLs. Although the *cis*-heritability of ATTSS is modest on average, it is consistent with estimates reported for other molecular phenotypes. For reference, cis-heritability of gene expression averages ~14–26% per tissue[26,27], with considerable heterogeneity across tissues and genes. To account for linkage disequilibrium (LD) among the identified 5′

https://doi.org/10.1038/s41467-026-76741-8

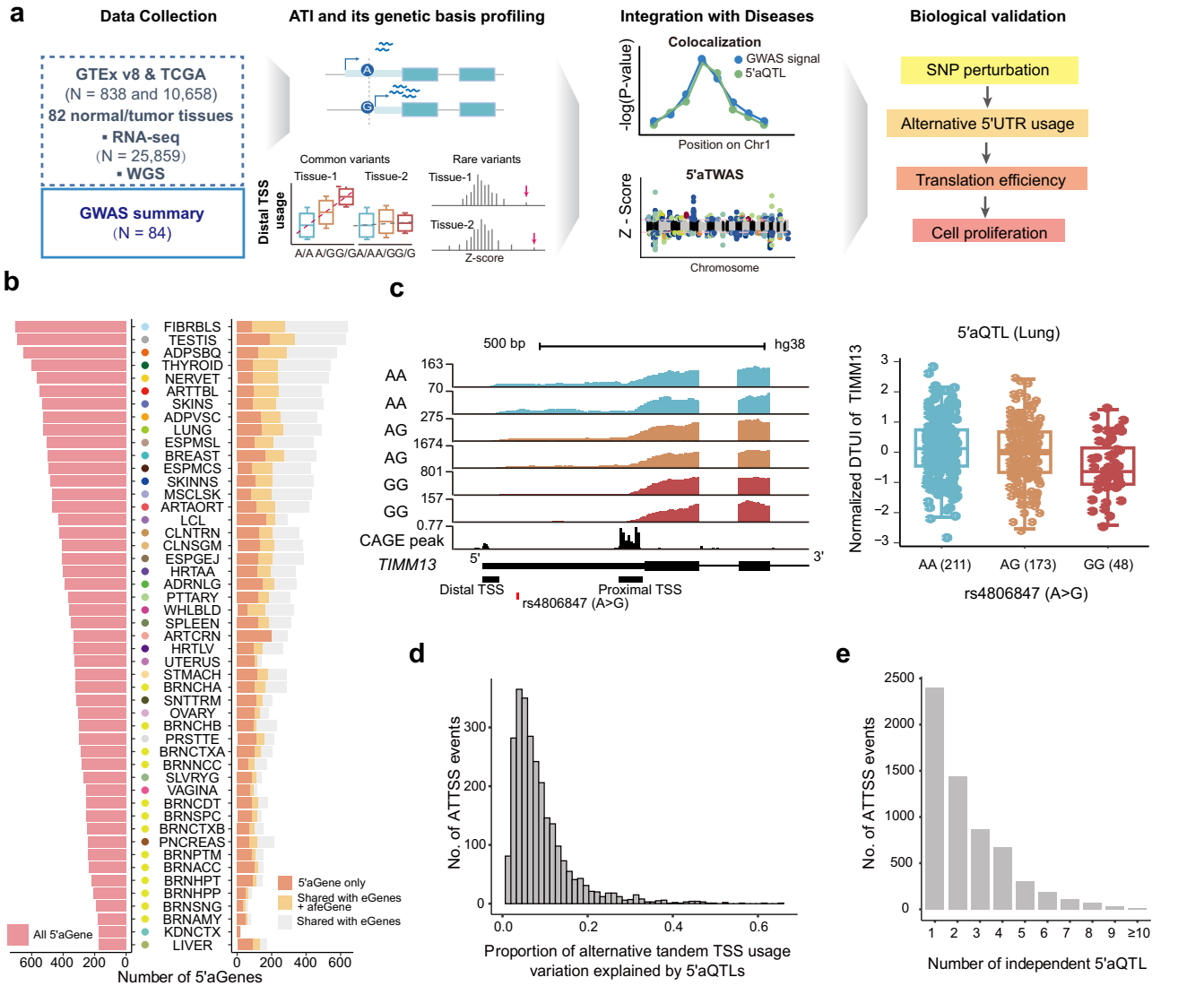


Fig. 1 | A multi-tissue atlas of genetic regulation of alternative transcription initiation. a Overview of the data and analyses in this study. **b** Distribution of the number of 5′aGene for each normal tissue, sorted by the 5′aGene number and number of 5′aGenes shared by eQTL or afeQTL. 5′aGene only (dark orange) means 5′aGene not shared by eGenes or afeGenes; light orange indicates the number of 5′ aGene shared by both eGenes and afeGenes, gray indicates 5′aGenes shared by only eGenes. **c** Example of a 5′aVariant (rs4806847) that is strongly associated with the *TIMM13* 5′ UTR usage in lung. Left: RNA-seq coverage track for the *TIMM13* 5′ UTR. The bottom four tracks display the CAGE peak, RefSeq gene structure, TSS site location as predicted by DATTSS, and 5′ avariant location. Right: Distribution of the quantile normalized DTUI for each genotype. Each dot represents the normalized DTUI value from one lung RNA-seq sample ($n = 432$ total samples), and the numbers in parentheses indicate the sample size for each genotype group. Box plots show the median values, with boxes spanning the 25th to 75th percentiles and whiskers extending to the most extreme values within 1.5 × IQR from the box. **d** Average fraction of tandem TSS usage variations that can be explained by 5′aQTLs for each transcript. The *y* axis represents the total transcripts across all human tissues studied. **e** The distribution of independent 5′aQTLs across tissues. Source data are provided as a Source Data file.

aQTLs, we used the sum of single effects (SuSiE)[28] to fine-map independent 5′aQTLs (summarized as 95% single-effect credible sets) for each ATTSS event in each tissue. We observed multiple independent 5′ aQTLs for 61.3% of tissue-transcript pairs, revealing widespread allelic heterogeneity in the genetic regulation of transcription initiation (Fig. 1e).

In summary, these results establish a foundational atlas of alternative tandem transcription initiation regulation across human tissues and cancers. By integrating nearly 26,000 RNA-seq datasets with comprehensive TSS annotations, our analysis reveals that alternative tandem transcription initiation is widespread, genetically controlled, and highly tissue-dependent. The discovery of thousands of 5′aQTLs provides a valuable resource for dissecting how common genetic variation shapes transcription initiation, thereby bridging a critical gap between noncoding variants and transcriptional initiation regulation.

Genetic regulation of alternative transcription initiation reveals tissue-restricted patterns

We next examined the cross-tissue sharing of the genetic regulation of alternative tandem transcription initiation. Using a metric adapted from Storey's π_1 metrics[29–31] (weighted π_1), we quantified the degree of 5′aQTL sharing between tissues. Specifically, weighted π_1 metric (Methods) is a composite measure designed to quantify the extent of cross-tissue sharing while accounting for differences in the number of evaluable associations between tissue pairs. As shown in Fig. 2a and Supplementary Fig. 5a and b, tissues with similar physiological functions, such as various brain regions, cluster together. This pattern

https://doi.org/10.1038/s41467-026-76741-8

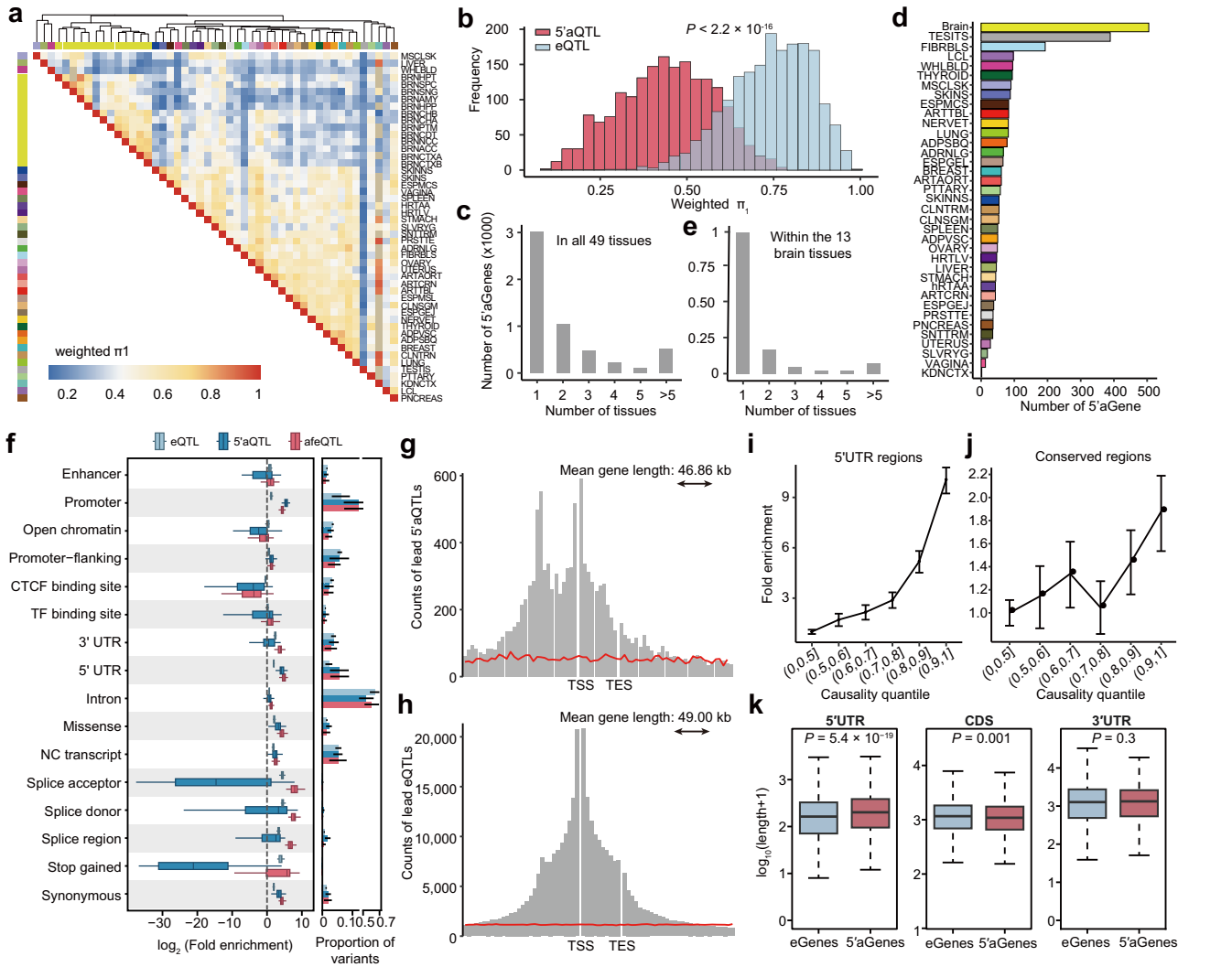


Fig. 2 | Tissue-specific 5′aQTLs and distinct genetic variation beyond eQTLs. **a** Heatmap shows the sharing relationships between tissues as estimated by the weighted π1 metric. The rows of the heatmap present the GTEx tissues (discovery tissues) and tissues order was determined by the method hclust with default settings in R basic package, and the tissues order in columns (replicate tissues) were the same as rows. Missing values were filled with gray. The diagonal presents self-comparison, and the value was assigned as 1 by definition. The color of each cell in the heatmap depend on the weighted π1 value. **b** The frequency of weighted π_1 for pairwise 5′aQTL and eQTL sharing across tissues. *P*-value was obtained from a two-sided Wilcoxon rank-sum test. **c** Number of shared tissues for all 5′aGene across 49 tissues. **d** Barplot shows the number of tissue-specific 5′aGene in each of the 49 tissues. **e** Number of shared tissues for all 5′aGene across the 13 brain tissues. **f** The proportion and enrichment of 5′aQTLs, eQTLs and afeQTLs under different annotations. Box plots show the distribution of log2-transformed fold enrichment values for each QTL type across 49 tissues. The center lines represent the median values, boxes span the 25th to 75th percentiles, and whiskers extend to the most extreme values within 1.5 × IQR from the box. The error bars in the right panel indicate the standard deviation of the proportion of variants across 49 tissues. **g**, **h** Distances between lead 5′aQTLs (**g**) and eQTLs (**h**) relative to the TSS of associated genes. The red line represents randomly selected positions within the ± 1 Mb window for each gene. TSS, transcription start site; TES, transcription end site. **i** Fold enrichment and 95% confidence intervals (CIs) for 5′aQTLs in each causality bin for the intersection with 5′ UTR regions. Points represent estimated fold enrichment values, and error bars indicate 95% CIs. **j** Fold enrichment and 95% CIs of 5′aQTLs that intersect with conserved regions, which were defined as regions with UCSC phastCons conservation scores > 0.9. Points represent estimated fold enrichment values, and error bars indicate 95% CIs. **k** Comparison of 5′ UTR, CDS length and 3′ UTR length between 5′aGenes and eGenes. A total of 3929 matched gene pairs were analyzed, including 5′aGenes with non-zero annotated 5′ UTR, CDS and 3′ UTR lengths identified in our tissues and corresponding eGenes with comparable expression levels and exon numbers. *P*- values were calculated using a two-sided *t* test. Box plots show the median values, with boxes spanning the 25th to 75th percentiles and whiskers extending to the most extreme values within 1.5 × IQR from the box. Source data are provided as a Source Data file.

extends to other systems. For example, within the circulatory system, similar tissues (e.g., the three artery tissues and two heart tissues) show higher weighted π_1 (two-sided Wilcoxon rank-sum test, $P = 0.00087$). We also used R package mashr v0.2.73[32], a frequently used method which borrows statistical strength across tissues to boost power and accurately identify shared and tissue-specific QTL effects, to evaluate the cross-tissue sharing of 5′aQTLs. The tissue-sharing patterns derived from mashr were similar to those estimated from weighted π_1 (Supplementary Fig. 5c). Collectively, these results indicate that biologically similar tissues share higher reproducibility of 5′aQTLs, supporting that the observed patterns reflect genuine biological relationships rather than technical artifacts. We then compared the weighted π_1 of each tissue pair calculated from 5′aQTL with that from eQTL. Strikingly, 5′aQTLs had significantly smaller weighted π_1 values than eQTLs (median weighted π_1 of 0.44 *vs.* 0.75, Wilcoxon rank−sum test $P < 2.2 \times 10^{-16}$; Fig. 2b), indicating substantially higher tissue specificity of 5′aQTLs compared to eQTLs.

To further characterize the presence of 5′aGenes across tissues, we identified 3012 5′aGenes (55.9%) present in only one tissue (Fig. 2c). These tissue-restricted 5′aGenes were predominantly enriched in brain tissues and testis, consistent with the known transcriptional complexity of these tissues and their pervasive use of alternative promoters. Specifically, 505 and 389 5′aGenes were uniquely detected in the brain and testis, respectively (Fig. 2d). For example, anoctamin 8 (*ANO8*), which encodes a transmembrane protein ubiquitously expressed across all human tissues, was identified as a significant 5′ aGene only in the brain tissues (Supplementary Fig. 5d–f). Similarly, potassium channel tetramerization domain 7 (*KCTD7*) exhibited a testis-specific 5′aQTL absent from all other tissues (Supplementary Fig. 5g–i). Given the variation in sample size across tissues, which raises false negatives in 5′aQTL effects, we further examined tissue specificity of 5′aQTL effects using multivariate adaptive shrinkage (MASH)[32], which jointly estimates effect sizes across tissues and borrows information from multiple tissues to improve detection of underpowered shared signals. Following the approach implemented in mashr and adopted by previous studies, we estimate tissue-sharing of 5′aQTL effect by sign (effects have the same sign across tissue pairs) and by magnitude (effects have similar magnitude, here defined as within a factor of 2). Across all 49 tissues, we observed similar tissue-restricted patterns (57.34%) by magnitude compared to the above significance-threshold method, while the majority of them show shared signs across tissues (Supplementary Fig. 5j, k). Similarly, in brain tissues, although a majority of 5′aQTLs were found to be restricted to a single brain region by considering effect magnitude in mashr (Fig. 2e and Supplementary Fig. 5l), most of them show concordant effect sign across brain regions (Supplementary Fig. 5m).

In summary, the genetic regulation of alternative transcription initiation is widespread and is characterized by broadly shared regulatory directions across tissues but substantial tissue-dependent differences in effect magnitude. These results reveal a distinct and context-dependent layer of transcriptional regulation shaped by genetic variation.

5′aQTLs represent a distinct regulatory layer beyond eQTLs

To determine whether 5′aQTLs capture distinct genetic variation beyond eQTLs and other molecular QTLs, including splicing QTL (sQTL) and alternative first exon QTL (afeQTL), we systematically compared 5′aQTLs with these QTL types across 49 tissues. We found that, on average, 36.6% of 5′aGenes across tissues were exclusively identified by 5′aQTL rather than any of the other QTL types (Supplementary Fig. 6a). Specifically, 48.9% of 5′aGenes per tissue were not identified as eGenes, 79.4% and 66.2% of 5′aGenes were not identified as afeGenes and sGenes, respectively (Fig. 1b and Supplementary Fig. 6a). Moreover, even among the shared 5′aGenes and eGenes (average π_1 value of 0.62 *vs.* 0.04, $P = 3.5 \times 10^{-15}$; Supplementary Fig. 6b), 71.3% had different causal variants as exemplified by 5′aGene *MYEOV* (myeloma overexpressed) (Supplementary Fig. 6c, d), the lead SNP (rs72930287) was significantly ($P = 8.8 \times 10^{-9}$) associated with ATTSS rather than mRNA abundance of this gene, suggesting independent *cis*-regulation of 5′ UTRs relative to mRNA abundance. To further characterize the molecular features distinguishing 5′aQTLs from eQTLs, we performed functional annotation enrichment analysis. We classified 5′aQTL variants based on functional annotations and used Torus[33] to assess enrichment across various genomic features. Consistent with their role in transcription initiation, 5′aQTL variants exhibited greater enrichment in promoter and 5′ UTR regions compared to eQTL variants (Fig. 2f and Supplementary Fig. 7). Both 5′aQTL variants and eQTL variants were highly enriched in regions near TSSs (Fig. 2g, h). We further applied fine-mapping using CAVIAR[34] and performed enrichment analysis across six causality bins based on posterior probabilities. We found that 5′aQTLs in the most causal bin (larger than 90th quantile) exhibited 10-fold enrichment in 5′ UTR regions compared with 5′aQTLs in the least causal bin (less than 50th quantile) (Fig. 2i). Moreover, 5′aQTLs showed a tendency to be enriched in conserved regions (phastCons conservation score > 0.9) (Fig. 2j), suggesting potential selective constraint on variants affecting transcriptional initiation. We also observed distinct genomic architectures between 5′aGenes and eGenes. Compared to eGenes—which were randomly shuffled to match the exon numbers and expression levels of 5′aGenes—the 5′aGenes exhibited significantly longer 5′ UTRs ($P < 5.4 \times 10^{-19}$). In contrast, only marginal differences were found in CDS lengths ($P = 0.001$), and no significant difference was observed in 3′ UTR lengths ($P = 0.3$; Fig. 2k). In summary, 5′aQTLs represent a distinct class of genetic regulatory variants, preferentially localized to promoter and 5′ UTR regions and frequently operating independently of eQTL effects, probably correlated with protein abundance (Supplementary Fig. 8). These findings reveal that genetic control of transcription initiation provides unique and functionally important mechanisms beyond regulation of mRNA abundance.

Transcription factor binding underlies 5′aQTL regulation

Alterations in core promoter elements can explain only a small percentage of 5′aQTLs (Fig. 2f and Supplementary Fig. 7), suggesting that most 5′aQTLs affect 5′ UTR length via other mechanisms. We hypothesized that many 5′aQTLs modulate transcription initiation by altering transcription factor binding sites. To test this hypothesis, we analyzed 407 ChIP-seq datasets of 126 transcription factors, obtained from the Roadmap Epigenomics and ENCODE projects[20]. We systematically examined 5′aQTL enrichment within ChIP-seq peaks of each transcription factor and compared them with matched random genomic sequence regions as controls. We applied the computational strategy mentioned earlier[35] to predict the *trans*-regulators and epigenetic regulators of transcription initiation and identified 123 transcription factors whose binding regions show enrichment for 5′aQTLs in at least one tissue. Among these, 49 transcription factors were enriched in 5′aQTLs across more than 30 tissues, including the well-known transcription initiation factors such as SP1, MAZ, NFY, ETS-family, and YY1(Fig. 3a). We conducted de novo TF binding motif enrichment analysis for 5′aQTL lead variants and identified 23 TF motifs that were significantly enriched in 5′aQTLs, including known transcription initiation factors like members of the ETS family and SP1 (Fig. 3b).

To characterize how 5′aQTLs modulate transcription initiation by altering transcription factor bindings, we intersected fine-mapped (PIP > 0.5) 5′aQTLs with transcription factor binding sites and identified 168 fine-mapped 5′aQTLs that were located in known transcription factor binding motifs (Fig. 3c and Supplementary Data 2). For instance, we identified a fine-mapped 5′aQTL (rs7786826, A > G) located in the 5′ UTR of RNA-binding protein 33 (RBM33), which alters the binding motif of the transcription initiation factor MAZ (Supplementary Fig. 9a, b). To test whether MAZ indeed regulates *RBM33* transcription initiation, we analyzed RNA-seq data from the K562 cell line with MAZ knockdown. We revealed altered distal TSS usage relative to control samples (Supplementary Fig. 9c), confirming that MAZ binding at this site functionally regulates *RBM33* transcription initiation. In addition to validating known transcription initiation factors, our analysis also uncovered regulatory relationships involving factors not previously associated with transcription initiation control. We identified a fine-mapped 5′aQTL (rs3120064) that impacts transcription initiation regulation of O-sialoglycoprotein endopeptidase (*OSGEP*) by altering the binding motif of transcription factor E2F4 (Fig. 3d). To determine whether this 5′aQTL indeed modulates *OSGEP* transcription by disrupting E2F4 binding, we performed E2F4 knockdown in MCF-7 and MCF-10A cell lines, followed by RNA sequencing (Fig. 3e, f), and observed consistent increase of distal TSS usage of *OSGEP* upon E2F4 knockdown in both cell lines (Fig. 3g, h and Supplementary Fig. 9d–f), while no consistent *OSGEP* expression changes were observed. These

https://doi.org/10.1038/s41467-026-76741-8

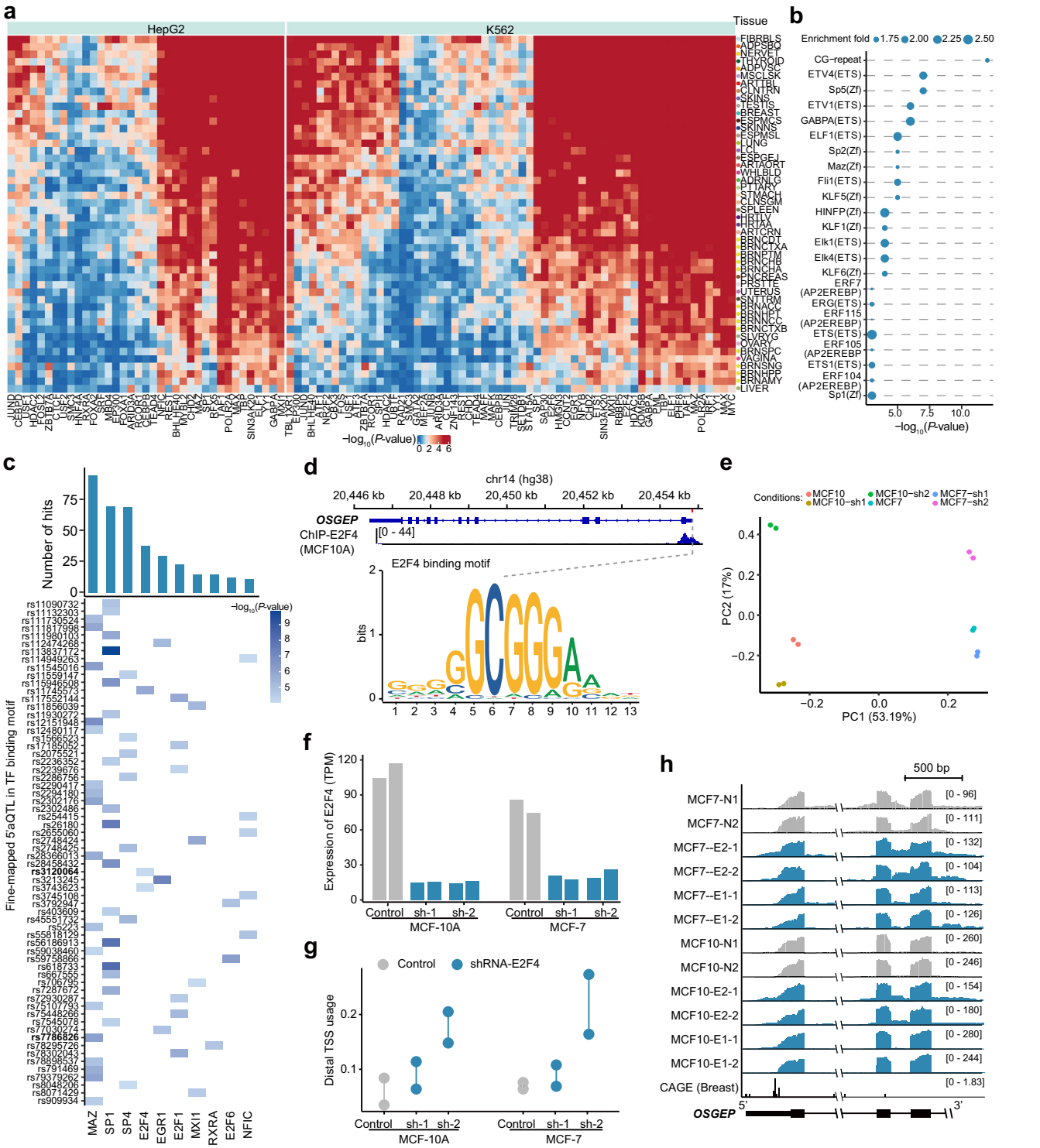


Fig. 3 | Transcription factor binding and chromatin context underlie 5′aQTL regulation. a Heatmap showing the 5′aVariant significance for TFs identified by ENCODE in each tissue (row). The right bar shows the color code for each tissue; the top color bar represents the HepG2 and K562 cell lines, separately. Values in the heatmap represent the degree of enrichment for 5′aQTLs in TF binding sites compared with the control. The tissue of each row of the heatmap was indicated by a dot with a color consistent with that used throughout the manuscript. **b** The TF motif enrichment using de novo motif scans. Fold enrichment is shown as mean (dot). *P*-values were calculated using Fisher's exact test. **c** Top 10 transcription factors (TFs) most frequently perturbed by 5′aQTLs. The bar plot shows the number of significant hits for each TF, and the heatmap represents the corresponding $-\log_{10}$(*P*-value). *P*-values were calculated using Fisher's exact test. **d** An example of a causal 5′aQTL SNP disrupting E2F4 binding and modulating ATTSS in *OSGEP*. Genomic view of the *OSGEP* locus on chromosome 4. The track shows E2F4 ChIP-seq signal in MCF-10A cells, revealing a binding peak at the location of the fine-mapped causal SNP rs3120064 (A > G). The SNP is located within the E2F4 binding motif. **e** PCA plot shows the relationships of samples with/without E2F4 knockdown in MCF-7 and MCF-10A cell lines. **f** The expression level of E2F4 after knockdown in MCF-10A and MCF-7 cell lines. Two shRNAs were used. **g** The effects of knockdown E2F4 on TSS usage of *OSGEP*. **h** RNA-seq coverage tracks for the *OSGEP* 5′ UTR in samples with/without E2F4 knockdown in MCF-7 and MCF-10A cell lines. Source data are provided as a Source Data file.

https://doi.org/10.1038/s41467-026-76741-8

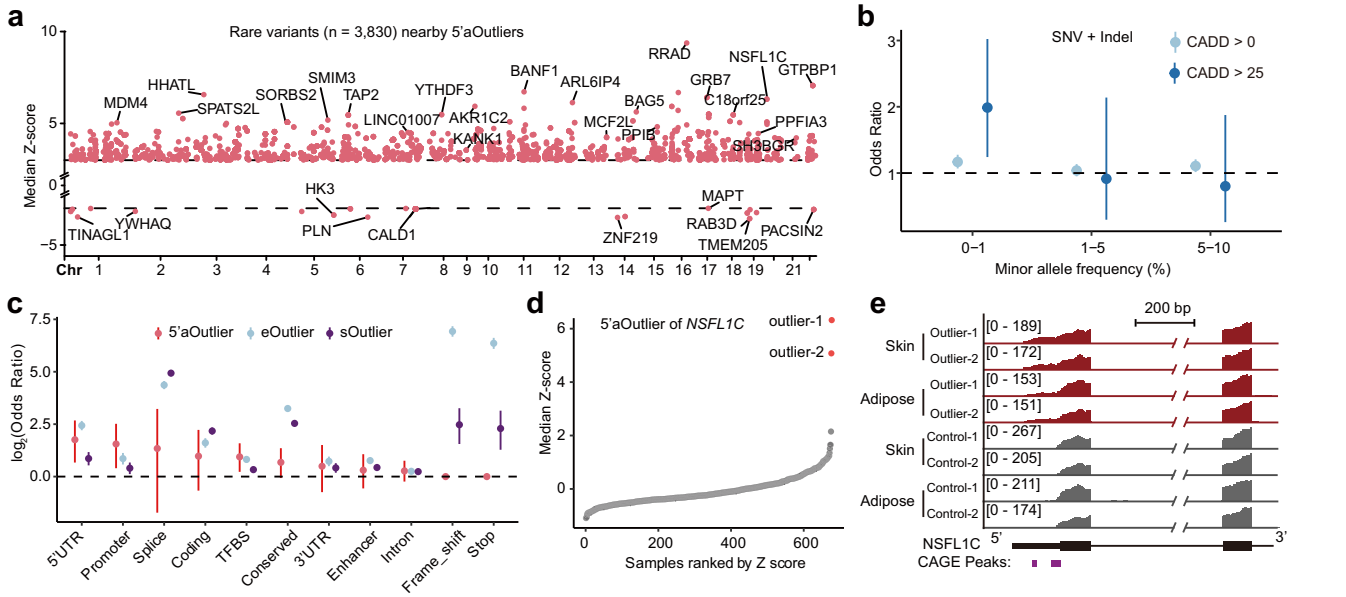


Fig. 4 | Rare variants influence tandem TSS usage across human tissues. **a** Distribution of 5′aOutliers across the human genome. Genes with the highest (for positive median Z-scores) or lowest (for negative median Z-scores) Z-score at each chromosome region were labeled. **b** Enrichment of deleterious single-nucleotide variants (SNVs) and small INDELs in 5′aOutliers; variants within 1 kb of 5′aOutlier genes were counted. Points represent estimated ORs, and error bars indicate 95% CIs. **c** Enrichment of rare variants of different categories in 5′aOutlier (red), eOutlier (light blue), and sOutlier (purple). Points represent log2-transformed estimated ORs, and error bars indicate 95% CIs. **d** Median Z-score distribution of the *NSFL1C* gene across individuals. Outliers are highlighted with red dots. **e** RNA-seq coverage track of the *NSFL1C* gene 5′ UTR in outlier individuals (red) and nonoutlier individuals (gray) in the Skin and Adipose tissues. Source data are provided as a Source Data file.

observations suggest a role for E2F4 in modulating the alternative transcription initiation of *OSGEP*.

In summary, 5′aQTLs are preferentially located within active promoter regions enriched for transcription initiation factor binding. Through 5′aQTL analysis, we identified additional transcription initiation regulators.

Rare variants influence tandem TSS usage across human tissues

While common variants explain a substantial portion of alternative transcription initiation variations, thousands of rare genetic variants are present in individual genomes[36], and many have been found to contribute to both rare and common diseases[37,38]. To determine whether rare genetic variants (minor allele frequency < 1%) affect complex disease risk by modulating tandem TSS usage, we explored characteristics of rare variants influencing tandem TSS usage using an outlier enrichment approach[39,40]. We identified 2238 multi-tissue 5′ alternative tandem transcription initiation outlier events (5′aOutliers). We use the term 5′aOutlier, defined as the individual-gene combination that is an outlier in the majority of an individual's sampled tissues (Fig. 4a). We required individual-gene pairs with ATTSS events quantified in at least two tissues and absolute median Z-scores > 3 to be classified as multi-tissue 5′ aOutliers. These 5′aOutliers comprised 1735 ATTSS events (1532 genes) across 593 individuals, resulting in 3.8 outlier events for each individual. Unlike previously reported gene expression outliers (eOutliers), splicing outliers (sOutliers), and alternative polyadenylation outliers (aOutliers)[41], the majority of 5′aOutliers exhibited positive Z scores, indicating that these genes have longer 5′ UTRs in outlier individual(s) compared with the majority of individuals (Supplementary Fig. 10).

To test whether nearby rare variants in outlier individuals contributed to 5′aOutlier events, we systematically analyzed the genomic context of these outliers and identified 3830 rare variants within 10 kb of the first exon region associated with multi-tissue 5′aOutliers (Fig. 4a and Supplementary Data 3). As expected, we observed significant enrichment of rare genetic variants, including single-nucleotide variants (SNVs) and small insertions and deletions (indels) (Fig. 4b). Importantly, enrichment was greater when restricted to deleterious variants (Combined Annotation-Dependent Depletion, CADD score > 25; Fig. 4b). In addition to CADD annotations, we examined rare variants across different genomic annotation categories in 5′aOutliers relative to eOutliers and sOutliers. Notably, 5′aOutliers are not enriched for most classes of rare variants that showed significant enrichment in eOutliers and sOutliers. Instead, 5′aOutliers were significantly enriched only for 5′ UTR and promoter rare variants (Fig. 4c). Among these 5′aOutliers, several significant genes were identified, including *NSFL1C*, a cofactor for the VCP/p97 ATPase and involved in proteasome-mediated protein degradation. Two 5′aOutlier individuals were identified for this gene. *NSFL1C* with two 5′aOutlier individuals was found to utilize the longer 5′ UTR in outlier individuals (Fig. 4d, e).

In summary, rare variants with substantial functional impact can perturb tandem TSS usage, predominantly through promoter and 5′ UTR alterations. These findings extend the contribution of rare variation to transcription initiation regulation and reveal an additional mechanism by which rare noncoding mutations can influence complex disease risk.

Genetic regulation of transcription initiation contributes to complex trait heritability

To investigate the contribution of 5′aQTLs to the heritability of complex traits and disease risk, we collected GWAS summary statistics for 84 human traits and diseases, including 52 cancer GWASs (Supplementary Data 4). Using functional genome-wide association analysis (fgwas)[42], we evaluated enrichment of 5′aQTL variants among trait-associated GWAS SNPs across tissues. We observed significant enrichment of 5′aQTL variants in 13.2% of tissue-trait pairs, particularly in disease-relevant tissues. For instance, we detected strong enrichment (OR = 3.17, 95% CI: 1.56 to 4.62) in skin tissue for basal cell carcinoma (Fig. 5a). Compared with eQTLs, 5′aQTLs exhibited more substantial enrichment in 92 tissue–trait pairs, suggesting that transcription initiation provides additional context-specific regulatory insights into disease risk.

https://doi.org/10.1038/s41467-026-76741-8

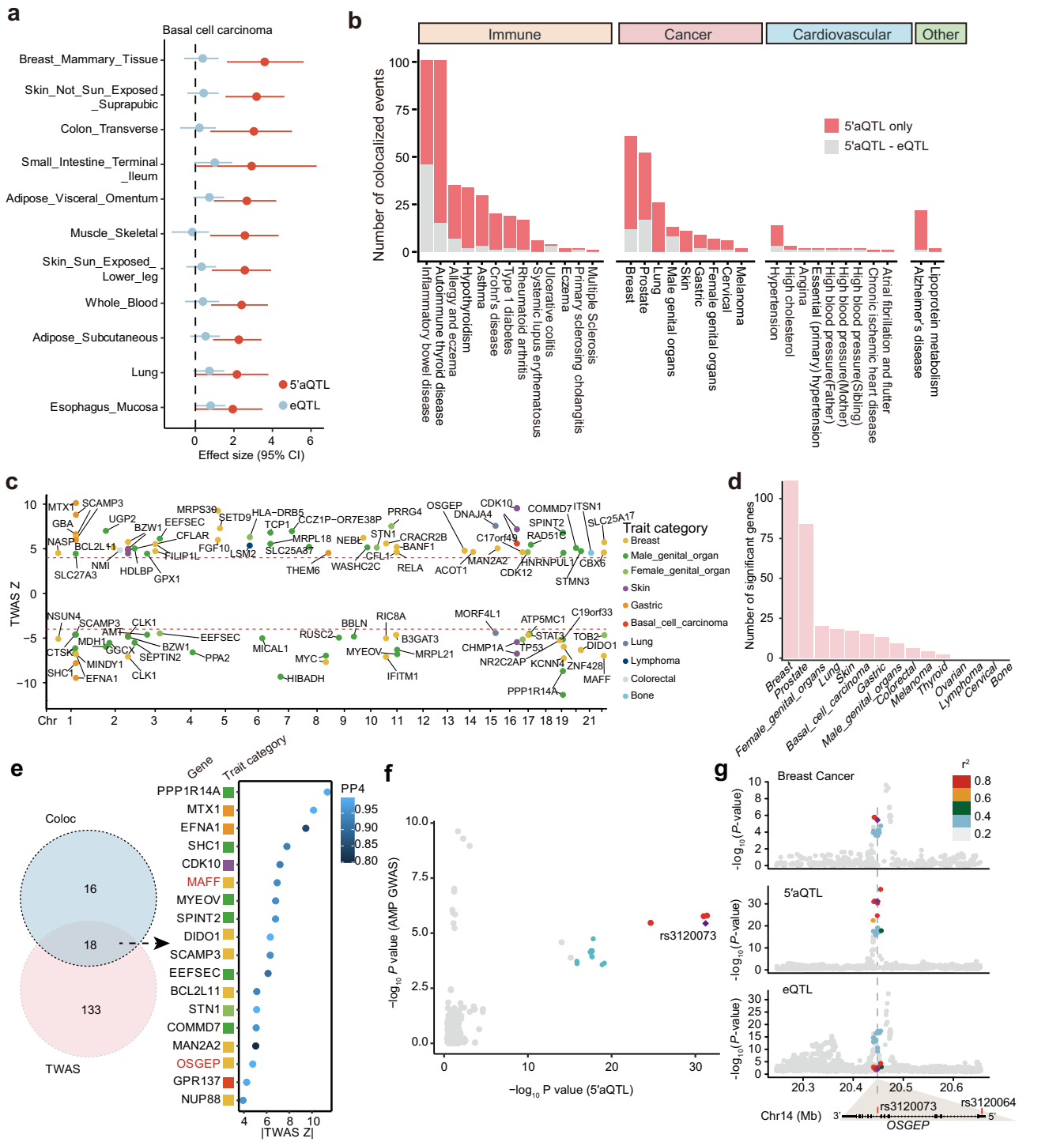


Fig. 5 | Genetic regulation of transcription initiation contributes to complex trait heritability. a Tissues with 5′aVariant enrichment but no eQTL enrichment for basal cell carcinoma traits. The enrichment values (effect size) were calculated using functional genome-wide association analysis, which quantifies the relationships between trait-associated variants and 5′aQTLs/eQTLs. Points represent estimated effect sizes, and error bars indicate 95% CIs. **b** Number of 5′aQTL-colocalized genes, colored based on whether the gene also colocalized with eQTLs. Gray represents 5′aQTL-colocalized genes that are also eQTL colocalized genes, and red represents 5′aQTL-colocalized genes are not eQTL colocalized genes. **c** Manhattan plots of 5′aTWAS results in 10 cancers using prediction models from 49 human tissues. Each point represents the *Z*-score of a single 5′aTWAS association. Colored points represent significant associations with cancer types at FDR < 0.05, with each of the 10 colors representing 1 of 10 different cancer types. **d** Bar plots show the number of 5′aTWAS significant genes (FDR < 0.05) for 15 cancer traits in 49 human tissues. **e** The number of disease susceptibility genes identified by both TWAS and Coloc. The left Venn diagram shows the gene overlapping situation identified by the two methods, while the right scatter plot displays the PP4 values of 18 genes identified by both methods. The color of the squares represents different types of traits (same as the trait category shown in c). **f** Correlation of the *P*- values of 5′aQTL associations for *OSGEP* and matched *P*-values in breast cancer GWAS. *P*-values of 5′aQTL were obtained from linear regression models used for 5′aQTL mapping. **g** Aligned Manhattan plots of Breast cancer GWAS (top), 5′aQTLs (middle), and eQTLs (bottom) at the *OSGEP* locus. SNPs are colored by LD (r^2). *P*-values of QTL were obtained from linear regression models used for QTL mapping. Source data are provided as a Source Data file.

To systematically characterize 5′aQTLs that share genetic signals with GWAS risk variants, we conducted colocalization analysis to identify ATTSS events that contribute to disease risk through genetic effects. By integrating 5′aQTL data from 49 human tissues with GWAS summary statistics from 32 non-cancer and 52 cancer traits (Supplementary Data 4) using Coloc, we identified 116 loci (7.6%) from 27 non-cancer GWASs and 37 loci (7.1%) from 15 cancer GWASs that exhibited colocalized signals with 5′aQTLs (Supplementary Data 5). In total, we identified 425 significant colocalized signals spanning 76 of the 5′aQTL genes for the non-cancer GWAS loci and 176 colocalized signals spanning 32 of the 5′aQTL genes across the cancer GWAS loci (Fig. 5b, Supplementary Fig. 11 and Supplementary Data 6) using Coloc PP4 > 0.75 and PP4/(PP4 + PP3) > 0.9. Among these colocalized 5′ aGenes, we identified several known cancer driver genes. Notably, *MAFF* (MAF BZIP Transcription Factor F), previously implicated in promoting tumor invasion and metastasis[43], exhibited colocalization with breast cancer GWAS signals (Supplementary Data 6). In addition, we identified multiple immune-related genes as immune disease susceptibility loci, including *IFITM1* and *IFITM2* from the interferon-induced transmembrane protein family, as well as *STAT6* (signal transducer and activator of transcription 6) (Supplementary Data 6). For comparison, we also performed colocalization analyses for eQTLs using the same datasets. Notably, 74.0% of trait-colocalized 5′aQTL signals were independent of eQTLs (Fig. 5b), indicating a distinct role of 5′aQTLs in disease risks. These findings establish transcription initiation regulation as a key molecular intermediary associating genetic variation in noncoding regions to tissue-specific disease mechanisms.

To systematically identify disease susceptibility genes associated with ATTSS events, we adapted FUSION[44] to examine associations between GWAS summary statistics and ATTSS events. TWAS enables the identification of additional trait-associated loci compared to traditional GWAS. Briefly, for each dataset, we used a mixed-linear model to estimate the heritability of ATTSS events (normalized DTUI values after covariates correlation) explained by *cis*-SNPs proximal to the TSS of each gene in a reference panel. Only genes with significant heritability estimates were included in subsequent analyses. For each FUSION-trained model, including BLUP, LASSO, and Elastic Net, we used cross-validation to select the model with the optimal 5′aTWAS prediction accuracy for each ATTSS event. We obtained 5025 tissue-specific 5′aTWAS prediction models from GTEx reference panels, spanning 2580 ATTSS events. The number of 5′aTWAS prediction models was highly correlated with the sample sizes of the reference panels (Supplementary Fig. 12a). The average in-sample prediction accuracy (measured by heritability-normalized R^2, $R^2/cis\text{-}h^2$) of 5′ aTWAS models was 0.67 for GTEx (Supplementary Fig. 12b), which is similar to that of expression and splicing TWAS models. These results indicate that most cis-regulated transcription initiation variations are effectively captured by *cis*-SNPs.

We applied these 5′aTWAS models to investigate the interplay between genetic variation and ATTSS events in disease risk. The 5′ aTWAS models identified 151 genes (Fig. 5c, d) associated with cancer traits (FDR < 0.05) and 536 genes associated with non-cancer traits (Supplementary Fig. 13). To enhance the detection of disease susceptibility genes associated with alternative transcription initiation, we integrated all genes identified by both Coloc and 5′aTWAS. Consequently, we identified 18 5′aGenes in seven cancer types, including *MAFF* in breast cancer (Fig. 5e). Nine of these genes were found to be associated with breast cancer risk, with most having well-characterized roles in breast cancer development. These findings highlight the role of ATTSS events in cancer development.

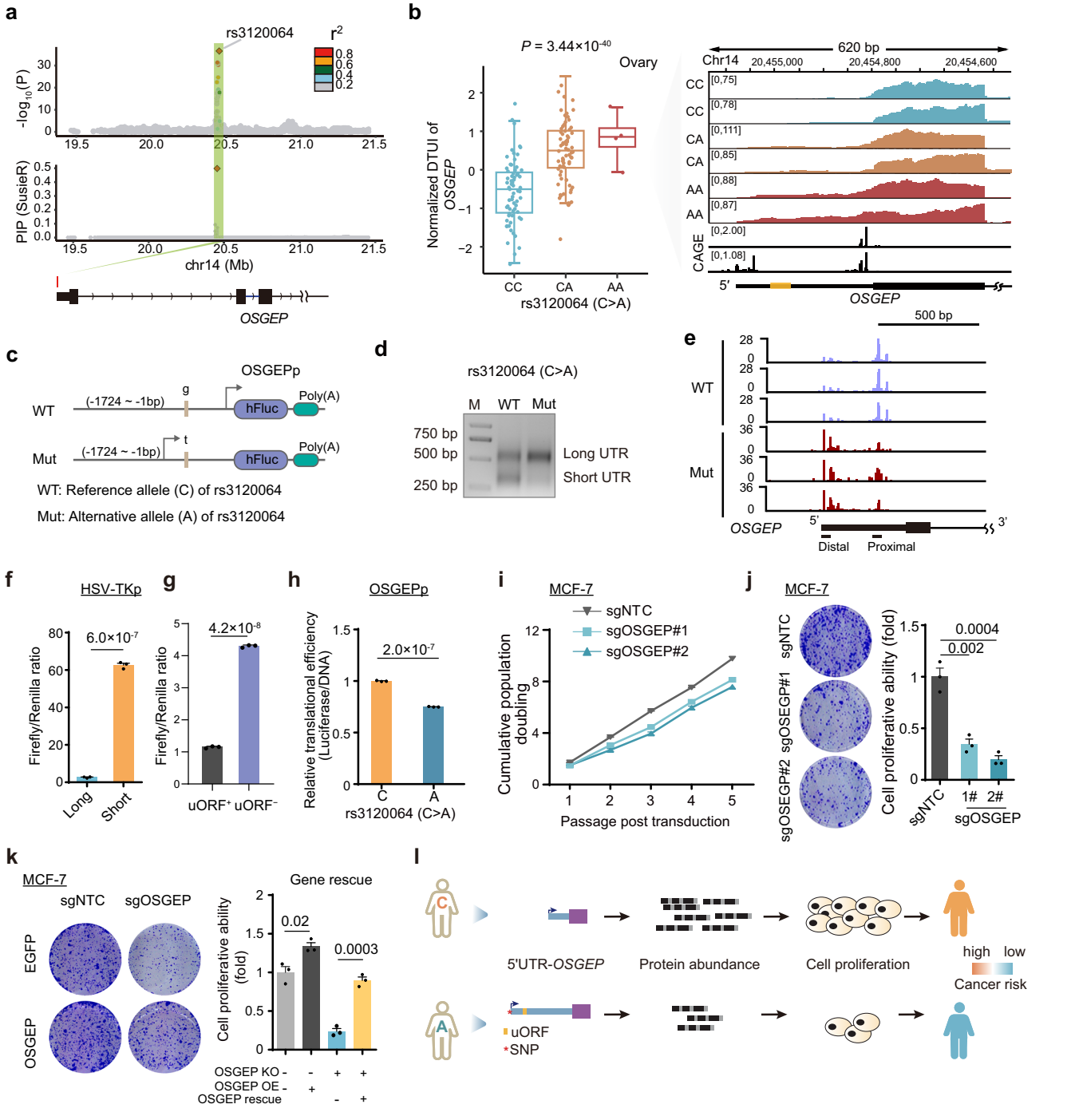
In summary, 5′aQTLs substantially contribute to the heritability of complex traits, and 5′aTWAS extends conventional transcriptome-wide association approaches by incorporating variation in transcription initiation. This analysis identifies known and additional candidate susceptibility genes, demonstrating that alternative TSS regulation provides a complementary and mechanistically distinct path associating genetic variation to cancer risk.

Alternative TSS regulation of OSGEP is associated with breast cancer risk

Through both colocalization and TWAS analysis, we identified O-sialoglycoprotein endopeptidase (*OSGEP*) as a breast cancer risk gene (Fig. 5f, g). *OSGEP* is involved in tRNA threonylcarbamoyladenosine modification and has been implicated in rare neurological disorders (OMIM:610107) but has not been previously linked to breast cancer. Motif analysis at this locus revealed 5′aQTL rs3120064 that disrupts transcription factor E2F4 binding (Figs. 3d, 6a, b) and exhibits strong linkage disequilibrium (LD $R^2 = 1.0$ in EUR) with the breast cancer risk variant rs3120073 (Supplementary Fig. 14a). To functionally assess the role of alternative alleles at the causal SNP in tumor cellular phenotypes, we observed higher distal TSS usage associated with the alternative allele (rs3120064 C > A) (Fig. 6b and Supplementary Fig. 14b). To directly validate this observation, we performed 5′rapid amplification of cDNA ends (RACE) in MCF-7 breast cancer cells to carry either the reference or alternative allele, followed by sequencing (Fig. 6c). Briefly, the promoter of *OSGEP* (-1 ~ 1724 bp) with reference allele (rs3120064 C) or alternative allele (rs3120064 A) was cloned into a luciferase lentiviral vector, three days post-infection, cells were collected for 5′ RACE. We observed that the alternative allele of rs3120064 lead to significantly higher distal TSS usage (Fig. 6d, e), demonstrating the causality of 5′aQTL in modulating *OSGEP* transcription initiation. To investigate how alternative TSSs regulate OSGEP protein abundance, we next constructed vectors containing the long 5′ UTR or the short 5′ UTR of *OSGEP* for use in dual-luciferase reporter assays, where long 5′ UTR isoform produced by the distal TSS was 296 nt longer than the short 5′ UTR isoform produced by the proximal TSS (Supplementary Fig. 14b, c). After two days transfection, Renilla and Firefly luciferase activities were measured and we found that the short 5′ UTR exhibited significantly higher luciferase activity than the construct containing the long 5′ UTR (normalized to Renilla luciferase activities) ($P = 6.0e\text{-}7$; Fig. 6f). Inspection of the 5′ UTR sequence revealed that the long 5′ UTR contains an upstream open reading frame (uORF) absent from the short isoform, disrupting this uORF (with mutation in start codon sequence; see details and uOFR sequences in Methods) significantly increased luciferase activity (Fig. 6g). To assess the effect of genetic variation and 5′ UTR length on protein levels, we generated reporter constructs containing the *OSGEP* promoter with either the reference allele or the alternative allele driving luciferase expression. The *OSGEP* promoter with the alternative allele decreased protein abundance (normalized to DNA copies) (Fig. 6h and Supplementary Fig. 14d) but had minimal effect on transcript levels (Supplementary Fig. 14e, f). These findings suggest that alternative 5′aQTL alleles that alter 5′ UTR length contribute to reduced OSGEP protein levels through translational control involving uORF-mediated translational repression.

To determine whether OSGEP protein levels functionally impact breast cancer-relevant cellular phenotypes, we further investigated the phenotypic consequences of *OSGEP* dysregulation to mimic the alternative tandem transcription initiation-mediated protein changes. *OSGEP* deficiency induced by CRISPR/Cas9 significantly inhibited cell proliferation and decreased colony formation compared with non-targeting control (Fig. 6i, j). Significantly, overexpressing *OSGEP* (modified with a synonymous mutation in the sgRNA-targeted genomic region) in OSGEP-depletion MCF-7 cells rescued the cell proliferative phenotype ($P = 0.003$; Fig. 6k). These results establish that OSGEP protein levels directly regulate breast cancer cell proliferation. To further establish more direct evidence supporting the association between 5′aQTL and breast cancer cell phenotypes, we introduced the A/A or C/C genotype at the rs3120064 locus in MCF-7 cells through

https://doi.org/10.1038/s41467-026-76741-8



allele-specific base editing (C > A or A > C). We observed that the introduction of the alternative allele (C > A) significantly decreased the cell proliferative ability (Supplementary Fig. 14g, h). These results suggested that the causal genetic variants modulate *OSGEP* 5′ UTR length, potentially by modulating E2F4 binding, thereby regulating cell proliferation and breast cancer risk (Fig. 6l).

Collectively, these findings identify several cancer susceptibility genes associated with ATTSS and demonstrate that alternative 5′aQTL alleles influence *OSGEP* distal TSS usage, resulting in longer 5′ UTR usage and reduced protein levels.

Discussion

In this study, we establish a comprehensive atlas of ATTSS regulation across human tissues and cancers. By integrating nearly 26,000 RNA-seq and matched genotype datasets, along with a unified analytical framework, we systematically quantified tandem TSS usage and identified ~0.40 million 5′aQTLs across 49 normal and 33 tumor tissues. These results reveal that transcription initiation represents a distinct regulatory layer of gene regulation that complements traditional eQTL mechanisms. We observed that the overall median *cis*-heritability of common SNPs on ATTSS variation is 10.1%. Although modest, it falls in a similar range of known values for other molecular phenotypes[45,46]. For comparison, *cis*-heritability of gene expression averages ~14–26% per tissue, and this value is heterogeneous among tissues and genes[26]. Our analyses suggest that genetic regulation of alternative TSS usage is characterized by broadly shared regulatory directions across tissues but substantial tissue-dependent differences in effect magnitude, highlighting the context-dependent nature of transcription initiation

Fig. 6 | Alternative TSS regulation of OSGEP is associated with breast cancer risk. a Regional association plots of the *OSGEP* gene showing the nominal *P*-values (top) and fine-mapping posterior probabilities (bottom) for the *cis*-SNPs. **b** 5′aVariant (rs3120064) that is strongly associated with the *OSGEP* 5′ UTR usage. Left: Distribution of the normalized DTUI for each genotype. *P*-value was obtained from linear regression models used for QTL mapping. Each dot in the box plot represents the normalized DTUI value for one particular sample. Box plots show the median values, with boxes spanning the 25th to 75th percentiles and whiskers extending to the most extreme values within 1.5 × IQR from the box. Right: RNA-seq coverage track for the *OSGEP* 5′ UTR. **c** Schematic diagram for the 5′RACE strategy in MCF-7 cells. **d** Changes of 5′ UTR usage in MCF-7 breast cells detected by 5′ RACE. The experiment was repeated independently for $n = 3$ times. **e** Sequencing revealed distal TSS of OSGEP changes between WT and Mut. **f** Luciferase activity from a reporter system containing the short and long 5′ UTR of *OSGEP* in HEK 293 T cells. Data are presented as mean Firefly/Renilla ratio from $n = 3$ biological replications, and error bars represent the standard error of the mean (SEM). **g** Luciferase activity from a reporter system containing the uORF and the one without the uORF 5′ UTR of *OSGEP* in HEK 293 T cells. Data are presented as the mean ratio of luciferase levels to input DNA copies from $n = 3$ biological replicates, normalized to the value of the C allele, and error bars represent the SEM. **h** Luciferase activity from a reporter system containing reference allele (rs3120064 C) or alternative allele (rs3120064 A) of *OSGEP* promoter in HEK 293 T cells, normalized to DNA copies. Data are presented as mean normalized luciferase activity from n = 3 biological replicates, and error bars represent the SEM. **i** Cell proliferation of sgRNA-mediated knockdown cells was analyzed during passage. **j** Colony formation assay for MCF-7 breast cells treated with the indicated sgRNAs. Left panel: Images are representative results from three independent experiments. Right panel: Quantification of colony numbers from the left panel. Data are presented as mean values from $n = 3$ biological replicates, and error bars represent the SEM. **k** Colony formation assay for MCF-7 breast cells upon ectopic expression of *OSGEP* in sg- *OSGEP* #2-transduced cells. Left panel: Images are representative results from three independent experiments. Right panel: Quantification of colony numbers from the left panel. Data are presented as mean values from $n = 3$ biological replicates, and error bars represent the SEM. **l** Schematic depicts alternative alleles of 5′ UTR variants, mediated 5′ UTR lengthening promotes cancer progression. Source data are provided as a Source Data file.

regulation. Over half of 5′aQTL-regulated genes are restricted to a single tissue, with strong enrichment in brain and testis, two tissues often characterized by their dynamic promoter switching and complex transcript diversity. The tissue specificity of 5′aQTL exceeds that observed for eQTLs, indicating that alternative tandem transcription initiation usage is a finely tuned and context-dependent regulatory mechanism. By comparing 5′aQTLs with eQTLs and integrating other datasets, we demonstrate that common variants affecting transcription initiation are preferentially located within promoters and 5′ UTRs. These findings suggest that subtle alterations in transcription initiation rather than mRNA abundance can mediate the effects of noncoding variants on cellular function. Extending to rare variation, we found that deleterious promoter and 5′ UTR mutations are strongly enriched near outlier individuals with aberrant TSS usage, revealing that rare noncoding variants can perturb transcription initiation across multiple tissues.

Based on our findings on the E2F4-dependent 5′aQTL associated with *OSGEP*, we propose a potential regulatory model in which transcriptional output is shaped by initiation events occurring across multiple alternative TSSs. In this framework, E2F4 binding near the proximal promoter of *OSGEP* may contribute to both preferential proximal TSS usage and local transcription initiation efficiency, potentially through modulation of promoter occupancy and the assembly of the transcriptional machinery. Disruption of E2F4 binding, either through transcription factor depletion or sequence variation affecting TF binding motif, therefore may shift the balance between proximal and distal TSS usage while altering transcriptional output. The broader regulatory circuitry underlying these processes remains to be established. Our functional characterization of *OSGEP* exemplifies how TSS regulation mediates genetic risk through translation control rather than transcription. Breast cancer–associated variants within the *OSGEP* promoter promote distal TSS usage, producing a longer 5′ UTR that introduces upstream open reading frames and reduces translation efficiency. This mechanism directly links noncoding variation in promoter regions to reduced protein abundance and tumor growth inhibition, underscoring the translational consequences of transcription initiation control.

Collectively, our work establishes alternative transcription initiation as a key mediator of the relationship between genetic variation and phenotypic diversity. By integrating the effects of common and rare variants, this atlas expands the functional interpretation of noncoding variants and provides a foundational resource for future studies of transcriptional regulation and disease. Extending this framework to single-cell and long-read transcriptomics will further refine our understanding of how promoter choice shapes cell-type–specific gene expression and human pathophysiology.

Methods

Ethics statement

The study was designed and conducted in compliance with all relevant ethical regulations regarding the use of human study participants and in accordance with the Declaration of Helsinki. Collection and use of the human genotype and RNA-seq data from GTEx and TCGA in this study were conducted in accordance with the ethical guidelines of their respective consortia. The ethical procedures for data collection, including informed consent and ethics approval, have been described in the original publications of both the GTEx project[47] and TCGA. As this study exclusively utilized de-identified, publicly available data from these repositories, no further institutional ethics review was required.

Statistics and reproducibility

No statistical method was used to predetermine sample size, which was primarily based on the availability of the samples in the GTEx cohort. The experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment, as the study is not a randomized controlled trial. We conducted most statistical tests using R (v4.1.3), including Fisher's exact test, Wilcoxon rank-sum test, and Student's *t* test, utilizing the basic function fisher.test, wilcox.test and t.test, respectively. Please refer to the corresponding sections in Methods for details on statistical tests.

Collection and processing of GTEx and TCGA RNA-seq data and genotype data

We downloaded all 17,832 RNA-seq BAM files from 49 tissues, obtained from 838 donors, and corresponding genotype data from the GTEx v8 release in dbGaP (accession no. phs000424.v7.p2). Publicly available RNA-seq and genotype data from The Cancer Genome Atlas (TCGA) were obtained from the Genomic Data Commons (GDC) Data Portal (https://portal.gdc.cancer.gov/). All BAM files were converted to BigWig files using bamCoverage (v3.3.2) with the "binSize" parameter set to 1, which means the coverage was counted at single-base resolution. Genotype files (in VCF format) were filtered and processed with PLINK (v1.9) to remove low-quality and low-frequency (minor allele frequency less than 5%) variants. This generated a bed format of the genotype file. Genotype PCA analysis was conducted using QTLtools to obtain the PCA components of the genotype data.

Construction of the reference tandem TSSs

We constructed a reference tandem TSS dataset from three publicly available resources, including transcription start sites (TSSs) annotated in GENCODE (version 44), RefSeq (version 200) annotations, and

TSSs annotated by Cap Analysis of Gene Expression (CAGE) from FANTOM5. For TSS annotations in GENCODE and RefSeq, we merged transcripts whose first exon ends were identical based on GENCODE and RefSeq annotations. We chose the 5′ most TSS as the start position of the first exon region. Then we obtained all tandem TSSs for non-redundant first exon regions of all transcripts. Suppose a given annotated TSS is located between the 500nt upstream start position and the end position of the first exon region of a specific transcript on the same strand. In that case, the TSS is considered a tandem TSS of the transcript. Transcripts with multiple TSSs in the first exon region were collected. This provided us with a comprehensive map of annotated tandem TSSs to the first exon regions of 27,792 transcripts in 17,078 genes.

Generation of genome-wide tandem TSS quantitative data

BigWig files transformed from BAM files in each of the 49 tissues and 33 cancer types were analyzed using our DATTSS algorithm to quantify 5′ UTR tandem TSS usage. Briefly, the coverage of each base in the annotated first exon region in the reference TSSs dataset for each sample was extracted from the corresponding bigwig file with pyBigWig (v0.3.22). DATTSS algorithm inferred the relative usage of the distal (5′ most) TSS in each first exon region through the coverage information, and the relative usage of distal TSS was defined as the distal TSS usage index (DTUI) representing the relative expression contribution of upstream (most distal) TSS compared to the predominant or most proximal TSS within each first exon region of each gene. Biologically, the distal TSS usage reflects alternative promoter activation that can influence transcript architecture by altering 5′ UTR length, therefore may further impact upstream translation regulatory elements (e.g., upstream open reading frame, uORF) and RNA secondary structure, and in some cases the N-terminal coding sequence. To obtain a solid estimate of distal TSS usage, we set the minimum required mean read coverage across the 5′ UTR to be 10; otherwise, the usage was assigned as NA for the corresponding samples.

Covariate correction

We first corrected for population structure to account for hidden batch effects and unobserved covariates in each tissue. Sites marked as 'wasSplit' from the GTEx analysis freeze variant call format (VCF) were removed using BCFtools v.1.3. The variants were further filtered with a call rate of > 99% and MAF > 1%; LD pruning was performed using PLINK v.2.0 with a sliding window of 200 bp, a step size of 100 bp, and the r^2 threshold of 0.1. The top three principal components from the principal component analysis were consistent with the known three main subpopulations, including White, Black or African American, and Asian, in the GTEx samples. We used PEER with sex, PCR, platform, and the top 5 genotype principal components as the known covariates to estimate a set of latent covariates for the DTUI values in each tissue. The top 5 genotype PCs were included to account for population structure, as they have been shown to capture subpopulation differences among GTEx WGS samples[48]. The number of PEER factors was optimized based on suggestions from the GTEx Consortium; for tissue sample sizes < 150, 15 PEER factors were chosen. Thirty PEER factors were selected if the sample size ranged from 150 to 250, and 35 PEER factors were selected for > 250 samples.

5′aQTL mapping

We applied a linear regression model in QTLtools (version 1.3.1) to test the association between the quantile-normalized DTUI values and the genotype of SNPs within 1 Mb around the 5′ UTR region, adjusting for known covariates (including sex, PCR, platform, and the top five genotype PCs) and hidden covariates inferred by PEER. The number of PEER covariates for each tissue was determined as suggested by the GTEx Consortium. We performed 1000 rounds of permutation (with the parameter "--permutation 1000") to obtain adjusted *P*-values for the association between the quantile-normalized DTUI and the top variants in *cis*.

5′aQTL mapping for TCGA datasets

The QTL process for TCGA samples was also conducted by using QTLtools. Genotype data of samples in each tumor type were obtained from GDC, and SNPs were filtered by call rate > 99%, minor allele frequency (MAF) > 1%. SNPs within the 1 Mb cis window of tested TSSs were used for 5′aQTL mapping. DTUI values in samples of each tumor type were quantile-normalized. Genes with > 50% missing DTUI values or samples with > 80% missing DTUI values were removed, and the remaining missing DTUI values were imputed using the median DTUI value across samples, following the same procedure as that used in the GTEx dataset. Known factors, such as sample age and sex, the top five genotype PCs, and PEER factors inferred from DTUI, were used as covariates for QTL mapping. The number of PEER factors for each tumor type was determined based on sample size, following an approach similar to that used in GTEx. We also performed 1000 rounds of permutation to adjust *P*-values.

The SNP heritability estimation for individual ATTSS events

The genome-wide complex trait analysis (GCTA, v1.94.1) toolkit[25] was applied to estimate the heritability of individual ATTSS events. Specifically, the restricted maximum-likelihood (REML) model implemented in GCTA was applied to estimate the proportion of variance in ATTSS usage explained by common SNPs (MAF > 0.01). For each ATTSS event, normalized distal TSS usage index (DTUI) values across samples within each tissue were used as the phenotype, and common SNPs located within 1 Mb of the ATTSS event were included to construct the genetic relationship matrix for the analysis. Heritability was defined as the proportion of phenotypic variance in DTUI attributable to the additive effects of the assessed genetic variants.

Fine-mapping of causal variants to 5′aQTL

Fine-mapping of causal variants for 5′aQTL was conducted using SuSiE software[28] and CAVIAR[34]. In brief, the SuSiE can operate on individual-level data (genotypes and ATTSS phenotypes) and efficiently analyze loci containing many independent effect variables. We used default parameters in SuSiE, which allowed a maximum of 10 independent effects for each gene in the current study. Linkage disequilibrium (LD) matrices used in SuSiE were calculated directly from individual-level genotype data within GTEx. In parallel, CAVIAR was applied using marginal association statistics and LD matrices derived from the same GTEx genotype data to estimate the posterior probability of each variant being causal. The maximum number of causal variants per locus was set to 2, and 95% credible sets were constructed to prioritize candidate causal variants.

QTL enrichment in genomic annotations

We downloaded a reference annotation (WGS_Feature_overlap_collapsed_VEP_short_4torus.MAF01.txt.gz) from the GTEx data portal. This reference annotation file contains 18 genomic features for each variant, including enhancer, promoter, open_chromatin_region, CTCF_binding_site, TF_binding_site, 3_prime_UTR_variant, 5_prime_UTR_variant, frameshift_variant, intron_variant, missense_variant, non_coding_transcript_exon_variant, splice_acceptor_variant, splice_donor_variant, splice_region_variant, stop_gained, and synonymous_variant. Then we use TORUS to perform enrichment analysis of 5′aQTL on each genomic feature.

Estimation of QTL sharing between ATTSS and expression

To estimate the sharing between 5′aQTL and eQTL in matched tissues, we first identified the best SNP-gene pair of 5′aQTL (the most significant one). We then used Storey's π_1 method[31] (qvalue) to estimate

the sharing between 5′aQTL and eQTL by considering the association *P*-value of the best SNP-gene pair from ATI in expression.

Colocalization between 5′aQTL and eQTL

To examine whether 5′aQTLs share genetic signals with eQTLs in matched tissues, we performed colocalization analysis using the Coloc R package (v5.2.3). For each tissue, genes identified as both 5′aGenes and eGenes were included in the analysis. Coloc computes five posterior probabilities corresponding to five alternative hypotheses, as described in the original colocalization framework applied to GWAS. Based on the posterior probabilities for hypothesis 3 (PP3) and hypothesis 4 (PP4), genes were classified into two categories: (1) PP4 > 0.8, indicating that the 5′aQTL and eQTL signals likely share a common causal variant; (2) PP3 > 0.8, indicating that the 5′aQTL and eQTL signals are likely driven by independent causal variants.

Tissue sharing and specificity analysis of 5′aQTL

To estimate the tissue specificity of 5′QTL between tissues, we designed a weighted π_1 metric by considering the proportion of the best SNP-gene pairs from tissue A shared in tissue B. In brief, the weighted π_1 between tissue A and tissue B is calculated as: weighted $\pi_1(A,B) = p(A,B) \times \pi_1(A \mid B)$, where p(A,B) is the proportion of significant 5′aQTL SNP-gene pairs from tissue A that are also testable in tissue B, and $\pi_1(A \mid B)$ is the Storey's π_1 statistic estimating the proportion of true positives among these shared associations in tissue B. Notably, weighted π_1 metric is not equivalent to the within-tissue true positive rate of the original QTL discovery analysis. Rather, it is a composite measure designed to quantify the extent of cross-tissue sharing while accounting for differences in the number of evaluable associations between tissue pairs.

The cross-tissue sharing of 5′QTL was additionally evaluated using multivariate adaptive shrinkage (mash) as implemented in R package mashr (v0.2.73)[32]. Briefly, mash jointly models effect size estimates and their corresponding standard errors across multiple tissues, allowing information to be shared across multiple tissues, allowing information to be shared across tissues to improve estimation accuracy and statistical power. For each 5′aQTL locus, the nominal *p*-value and effect size estimated in all 49 tissues were used as input to the mash model. Posterior effect estimates and local false sign rate (LFSR) were calculated, and a 5′aQTL effect was considered significant in a tissue if the corresponding LFSR < 0.05. Following the approach implemented in mashr and adopted by previous GTEx studies, two tissues were considered shared by sign if significant effects (LFSR < 0.05) had the same direction (Eq. 1), and shared by magnitude (Eq. 2) if the effects had the same direction and their posterior effect sizes were within a factor of two of each other.

Sharing by sign:

$$\mathrm{sign}(\hat{\beta}_i) = \mathrm{sign}(\hat{\beta}_j) \tag{1}$$

Sharing by magnitude:

$$0.5 \le \frac{\hat{\beta}_i}{\hat{\beta}_j} \le 2 \tag{2}$$

Enrichment analysis of 5′aQTL in transcription binding regions

We obtained 407 ChIP-seq data for 126 unique transcription factors (TF) and 162 CLIP-seq data for 114 RNA-binding proteins (RBPs) from the Encyclopedia of DNA Elements (ENCODE) data portal for HepG2, K562 cells and other cell lines. To assess the enrichment of 5′aQTLs in TF/RBP binding peaks, we selected 5′aQTL variants within promoter regions representing the region of 3 kb upstream and downstream of the transcription start site (TSS). We created a control variant set by randomly shuffling the genomic locations of a selected 5′QTL variant set within the same promoter regions. We then counted the 5′aQTL variants in binding peaks of each TF/RBP using bedtools. Finally, we compared the number of 5′aQTL variants to that of random variant sets in TF/RBP binding regions using Fisher's exact test to determine the enrichment.

Multi-tissue 5′aOutlier calling and rare variants enrichment analysis

We called 5′aOutliers following the previous methods used for calling outliers of alternative polyadenylation (3′aOutliers)[41]. Briefly, we firstly removed covariates for ATTSS quantifications by calculating DTUI residuals, in which known covariates, including population structure, age, sex, and sequencing platforms, and inferred hidden covariates from raw DTUI metrics were regressed out using the function "PEER_getResiduals" in the PEER package. DTUI residuals were then scaled and centered to get the Z score of each ATTSS event in each tissue. We defined 5′aOutlier as "ATTSS gene-individual" pairs with an absolute median Z score across at least two different tissues greater than 3. Notably, the 5′aOutlier analysis was called only for European-ancestry individuals ($N = 715$), which accounts for 85% of GTEx donors, to avoid confounding from population bias.

To examine the enrichment of rare variants (RVs, minor allele frequency < 1%) near 5′aOutlier genes, we counted the RVs present within 10 kb of the outlier genes in both outlier and control individuals and built a 2 × 2 contingency table for each outlier event, containing the number of 5′aOutliers with RVs, the number of nonoutlier controls with RVs, the number of 5′aOutliers without RVs, and the number of controls without RVs. We then calculated Odds Ratios (ORs), *P*-value, and 95% confidence interval (CI) using Fisher's exact test in the R base package. Similar enrichment analysis was also conducted for RVs stratified by functional annotations and constraint scores.

SNP heritability estimation within GWAS risk loci and genetic correlation

We applied stratified LD score regression (v1.0.1) to cancer and non-cancer GWAS results to assess the enrichment of heritability attributable to 5′aQTLs and eQTLs within GWAS risk loci. Briefly, genome-wide association study (GWAS) summary statistics for each trait were first converted into the sumstats format using the script munge_sumstats.py. To ensure allele alignment between our summary statistics and the reference data used for LD score estimation, we restricted analyses to HapMap3 SNPs with the option --merge-alleles w_hm3.snplist, where *w_hm3.snplist* contains the list of SNPs and corresponding alleles. SNP heritability was then calculated from the formatted sumstats files using pre-computed ancestry-matched LD scores, obtained from the LDSC website.

Enrichment of molecular QTLs within GWAS risk loci

We used *fgwas* (v.0.3.6) to assess the enrichment of molecular QTLs within GWAS risk loci. Briefly, GWAS loci were annotated as 5′aQTLs (or eQTLs) in a binary manner. We considered all molecular QTLs that were significant (FDR < 5%). *fgwas* then constructed a hierarchical Bayesian model to estimate the enrichment effects of different molecular annotations within GWAS loci.

Colocalization of 5′aQTL and eQTL with GWAS signal

We collected well-curated GWAS summary statistics for 84 traits/diseases (Supplementary Data 4) from the UK Biobank and the literature. Colocalization analysis of 5′aQTL and GWAS signals was conducted with Coloc (v5.2.3) software[49]. In detail, Coloc uses the approximate Bayes factor test approach and calculates five posterior probabilities (PP0-PP4) representing five hypotheses: PP0, the posterior probabilities of the null model of no association; PP1, the posterior probabilities that causal SNPs are associated with GWAS signals only; PP2, the posterior probabilities that causal SNPs are associated with 5′

aQTLs only; PP3, the posterior probabilities that causal SNPs of GWAS signals and 5′aQTLs are independent; PP4, the posterior probabilities of GWAS signals and 5′aQTLs share causal SNPs. In the current study, we consider a significant colocalization event if PP4 > 0.75 and PP4/(PP4 + PP3) > 0.9, as also used in our previous study[35].

Building 5′aTWAS model

We used FUSION[50] to construct a 5′aTWAS model between ATTSS phenotypes and individual-matched WGS genotype data. Known covariates, the same as used in 5′aQTL mapping, including sex, PCR, platform, and hidden covariates calculated from PEER, were used to residualize DTUI values. Residuals of DTUI values were then used to train the cross-tissue 5′aTWAS models, along with the corresponding genotype data. Prediction models with a heritability of Bonferroni-corrected $P < 0.05$ were used for 5′aTWAS analysis. For model construction and downstream TWAS analysis, LD was obtained from the 1000 Genomes Project European ancestry reference panel. To ensure ancestry matching, only European-ancestry individuals in GTEx (identified by genotype principal component analysis) were included for model training.

Applying 5′aTWAS prediction models to GWAS summary statistics

We applied our 5′aTWAS models to the 84 sufficiently powered GWAS summary statistics (Supplementary Data 7) used for colocalization analysis in the current study. Variants in GWAS summary statistics were filtered to remove variants with minor allele frequency less than 0.01, indels, and all variants with ambiguous ref/alt alleles using the LDSC software. We then applied 5′aTWAS models to all the cleaned GWAS summary statistics data to associate the ATTSS with disease risks.

Cell culture

HEK 293 T (GNHu44) cells were purchased from the Cell Bank of the Type Culture Collection at the Shanghai Institute of Biochemistry & Cell Biology, Chinese Academy of Science and HEK 293 T and MCF-7 cells were maintained in high-glucose DMEM (Thermo Fisher Scientific) containing 10% FBS (VISTECH) and 1% penicillin/streptomycin (Thermo Fisher Scientific). MCF-10A (ATCC-originated) were maintained in Mammary Epithelial Basal Medium (Thermo Fisher Scientific) with 10% FBS (VISTECH), 20 ng/ml Epidermal Growth Factor (Thermo Fisher Scientific), 100 ng/ml cholera toxin (Sigma-Aldrich), and 1% Penicillin/Streptomycin (Thermo Fisher Scientific). No mycoplasma contamination was detected during cell culture.

Rapid amplification of cDNA ends

To investigate the impact of alternative alleles on tandem TSS usage, the promoter of *OSGEP* (-1 ~ 1724 bp) with reference alleles (rs3120064 C) or alternative alleles (rs3120064 A) was inserted into hPGK-luciferase (Addgene 19360) by replacing the hPGK promoter. To minimize the impact of other promoters, we deleted the PGK-puromycin sequence. All constructs were validated by Sanger sequencing. Viruses were harvested from HEK293T cells transfected with the indicated plasmid and the packaging plasmids pMD2G and psPAX2 using PEI MAX transfection reagents (Polysciences), concentrated, and titrated. MCF-7 and HEK 293T cells were transfected at a calculated MOI of 0.8 with eight μg/ml polybrene (Thermo Fisher Scientific) for 16 hr at 37 °C. Three days post-infection, cells were collected for RNA and DNA extraction. The full length of 5′ UTR of *OSGEP* was identified and amplified from the total RNA of MCF-7 cells, which were infected with the indicated plasmid, by 5′-RACE using Template Switching RT Enzyme Mix (NEB, M0466S) following the manufacturer's protocol. Partial RACE-PCR products were separated on a 2% agarose gel for gel detection, and others were extracted by M5 Hiper Gel Extraction Kit (Mei5bio) for Next-generation sequencing (NGS). The primers used for 5′-RACE are listed in Supplementary Data 8.

Luciferase reporter assay

To investigate the impact of alternative 5′ UTR sequence on translation, longer and shorter versions of 5′ UTRs were inserted into psiCHECK-2 Vector (Promega, catalog no. C8021) for Dual-Luciferase assay. Briefly, the 5′ UTR of interest was cloned downstream of an HSV-TK promoter and upstream of the Firefly luciferase ORF. All constructs were validated by Sanger sequencing. HEK293T cells were seeded in 24-well plates at a density of 30 ~ 40%, and 500 ng of the indicated plasmid was transfected using PEI MAX transfection reagents. Renilla and Firefly luciferase activities were measured two days post-transfection using the Dual-Luciferase Reporter 1000 Assay System (Promega) on a BioTek Synergy H1 plate reader with whole waveband.

To investigate the impact of an alternative allele on translation, the promoter of OSGEP (-1 ~ 1724 bp) with reference allele (rs3120064 C) or alternative allele (rs3120064 A) was inserted into the psiCHECK-2 Vector by replacing the SV40 promoter, Renilla luciferase, and HSV-TK promoter, upstream of the Firefly luciferase ORF. HEK293T cells were seeded in 24-well plates at a density of 30 ~ 40%, and 500 ng of the indicated plasmid was transfected using PEI MAX transfection reagents. Two days post-transfection, Firefly luciferase activity was measured using the Dual-Luciferase Reporter 1000 Assay System (Promega) on a BioTek Synergy H1 plate reader with a whole-wavelength range. Genomic DNA was extracted from the same lysate, and firefly luciferase DNA levels were measured by qRT-PCR analysis to standardize firefly luciferase activity. All the primer sequences are listed in Supplementary Data 8.

To evaluate the impact of uORF residing in the long 5′ UTR, on the translation of OSGEP, we constructed two vectors containing the wild-type uORF ($uORF^+$) (atgtgcaagagggaaaaaaacagtggagaaggctgggctggacc acgggaaagagataggaaaacggatacgaatgtaaggcacagagttggctcggcgtccctgg cagttaggaagacaggtccctggaggggccttggcgggtctccgctcgcgggagttccgcg gtgtctttcccagtgctatctgcaggctggccagcttcctctgcgctccggaaagctgcggcc-cagcgcggactag) and the mutant uORF ($uORF^-$), respectively. $uORF^+$ vector was constructed with the long UTR sequence and luciferase; $uORF^-$ vector was constructed with the long UTR sequence with a mutation (to disrupt the start codon sequence of the uORF) and was inserted into psiCHECK-2 Vector. Both constructs were validated by Sanger sequencing. HEK293T cells were seeded in 24-well plates at a density of 30 ~ 40%, and 500 ng of indicated plasmid was transfected by using PEI MAX transfection reagents. Two days post-transfection, Renilla and Firefly luciferase activities were measured using the Dual-Luciferase Reporter 1000 Assay System (Promega) on a BioTek Synergy H1 plate reader with full waveband.

DNA and RNA preparation

Genomic DNA was extracted with the TIANamp Genomic DNA Kit (DP304, TIANGEN) for DNA analysis. For RNA analysis, total RNA was extracted with Quick-RNA™ Miniprep Kit (cat#: R1055; Zymo Research), followed by cDNA synthesis with Template Switching RT Enzyme Mix (YNEB, M0466S). Quantitative RT-qPCR was performed in a Real-Time PCR system (Bio-Rad) using SYBR Green Supermix. The primer sequences used are listed in Supplementary Data 8.

Cell colony formation assay

Cell colony formation assay was performed as previously described. 2000 MCF-7 cells were seeded into 12-well plates for long-term culture for approximately 2 weeks. The cultured cells were then washed with PBS and fixed with 4% paraformaldehyde for 20 min. Subsequently, they were stained with 0.1% crystal violet (Sangon Biotech, cat #: A600331-0025) for 20 min at room temperature. Colony counting was performed using ImageJ.

Lentivirus preparation and titration

To construct lentiviral vectors expressing sgRNA targeting *OSGEP*, sgNTC, or other sgRNA, corresponding sgRNA oligonucleotides

(Supplementary Data 8) were inserted into the cloning site of lentiCRISPRv2 (Addgene, #52961) following the manufacturer's instructions. To construct lentiviral vectors expressing shRNA targeting E2F4, the corresponding shRNA oligonucleotides (Supplementary Data 8) were inserted into the cloning site of pLKO.1 (Addgene, # 10878) following the manufacturer's instructions. To construct lentiviral vectors expressing OSGEP, the corresponding cDNAs were amplified by PCR and cloned into hPGK-luciferase (Addgene 19360) by replacing the hPGK promoter with the EF1α promoter. Lentivirus was packaged as previously described. For virus titration, viruses were tested by counting the cell clones after puromycin selection.

DNA and RNA preparation

Genomic DNA was extracted with the TIANamp Genomic DNA Kit (DP304, TIANGEN) for DNA analysis. For RNA analysis, total RNA was extracted with Quick-RNA™ Miniprep Kit (cat#: R1055; Zymo Research), followed by cDNA synthesis with Template Switching RT Enzyme Mix (YNEB, M0466S). Quantitative RT-qPCR was performed in a Real-Time PCR system (Bio-Rad) using SYBR Green Supermix. The primer sequences used are listed in Supplementary Data 8.

RNA sequencing library construction

To investigate the impact of E2F Transcription Factor 4 (E2F4) on tandem TSS usage of the *OSGEP* promoter, MCF-7 cells were transfected with shE2F4 and the corresponding control shRNA. After puromycin selection, total RNA was extracted using the Quick-RNA™ Miniprep Kit (cat#: R1055; Zymo Research) according to the manufacturer's protocol. After total RNA was quantified using a Fragment Analyzer (Advanced Analytical), 1–2 μg of total RNA was used to prepare a sequencing library with a TruSeq RNA Sample Preparation Kit (Illumina) according to the manufacturer's protocol.

RNA sequencing data processing

The raw RNA sequencing reads were subjected to quality control using FastQC to assess sequencing quality metrics. Adapter sequences and low-quality bases were trimmed using TrimGalore (v0.6.6), with default parameters. Cleaned reads were aligned to the human reference genome (GRCh38) using STAR (v2.7.3a). Only uniquely mapped reads were retained for downstream analyses. Gene-level expression quantification was performed using RNA-SeQC (v2.4.3), generating read count and normalized expression estimates based on GENCODE gene annotations (v38). Standard quality control metrics, including mapping rate, rRNA content, and gene body coverage, were evaluated to ensure data integrity. In parallel, the RNA-seq data in K562 cell lines with MAZ knockdown (ENCSR129WCZ) and three controls (ENCSR126QVF, ENCSR109YDL, and ENCSR287SHU) aligned to the human reference genome hg38 were downloaded from the ENCODE database, and further gene-level quantification was performed with RNA-SeQC.

Prime editing of rs3120064 in the promoter of *OSGEP*

To investigate the impact of alternative alleles on tandem TSS usage of the *OSGEP* promoter in vivo, we firstly cloned the lentiV2-EFS-nCas9/RT (PEmax) plasmid. We first excised the U6-sgRNA cassette from the lentiCRISPR v2 plasmid (Addgene, #52961) and replaced the Cas9 cassette with an nCas9/MMLV-RT cassette (sequence from the pCMV-PEmax plasmid (Addgene, #174820)). Subsequently, we cloned prime-editing guide RNAs (pegRNAs) into the lentiV2-pegRNA vector backbone (Supplementary Data 8). MCF-7 cells were seeded in 24-well plates at a density of 12,500 cells per well and co-infected with the PEmax and pegRNA lentiviral vectors. Following infection, transduced cells were enriched by fluorescence-activated cell sorting (FACS) based on dual positivity for mCherry and EGFP. Cells were then cultured for six weeks post-infection before harvest.

To assess prime editing efficiency, genomic DNA was extracted from the harvested cells, and the target genomic loci were amplified by PCR from the purified DNA for next-generation sequencing (NGS) analysis. Briefly, sequencing libraries were prepared using a two-step PCR protocol. The NGS primers are listed in Supplementary Data 8.

Construction of the 5′aQTL-atlas website

We developed a web-accessible database, the 5′aQTL Atlas (https://bioinfo.szbl.ac.cn/5aQTLatlas), as a comprehensive resource for alternative tandem transcription start site (TSS) events and their associated genetic variants. All 5′aQTL data spanning 49 normal human tissues and 33 tumor types are stored in a MySQL database.

The web interface was implemented using Vue 3 with TypeScript, built with Vite. UI components were provided by PrimeVue and Ant Design Vue. IGV.js was used for genome browser visualization. The backend API was implemented in Python using Flask and deployed on an Ubuntu Linux server with Apache HTTP Server.

The 5′aQTL Atlas provides an interactive environment for exploring tissue-specific 5′aQTLs and their potential disease associations. Users can search and browse data by gene name or genetic variant and visualize results through an integrated genome browser. The platform also supports querying colocalization results between 5′ aQTLs and GWAS signals, with visualization via locus-level plots. In addition, users can download DTUI profiles and summary statistics for 5′aQTLs across tissues in batch mode. The database is freely accessible without registration or login requirements.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The raw data from the GTEx project V8 used in this study are available at the database of Genotypes and Phenotypes (dbGaP) under accession number phs000424.v7.p2 [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v7.p2]. The cancer transcriptome and genotype data of The Cancer Genome Atlas (TCGA) are available from the Genomic Data Commons (GDC) Data Portal (https://portal.gdc.cancer.gov/). GWAS summary statistics are from NHGRI–EBI GWAS catalog (https://www.ebi.ac.uk/gwas/), UK Biobank GWAS (http://www.nealelab.is/uk-biobank/), FinnGen (https://www.finngen.fi/en), and JENGER (http://jenger.riken.jp), with a complete list of GWAS summary statistics in Supplementary Data 4. RNA-seq data in K562 cell lines were downloaded from ENCODE (https://www.encodeproject.org/) under accession numbers: ENCSR129WCZ, ENCSR126QVF, ENCSR109YDL, and ENCSR287SHU. The RNA-seq data generated in this study have been deposited in the GEO database under accession code GSE311085. The data described in this study are freely available for querying, visualizing, and downloading at https://bioinfo.szbl.ac.cn/5aQTLatlas/, a website portal dedicated to 5′aQTL. Source data are provided in this paper.

Code availability

Code for detecting and quantifying alternative tandem TSS events is available on GitHub (https://github.com/ZhaozzReal/DATTSS). Codes used for the analyses presented in this paper are available on GitHub (https://github.com/Xu-Dong/5aQTL)[51] and Zenodo (https://doi.org/10.5281/zenodo.21287300).

References

1. Loos, R. J. F. 15 years of genome-wide association studies and no signs of slowing down. *Nat. Commun.* **11**, 5900 (2020).
2. Sud, A., Kinnersley, B. & Houlston, R. S. Genome-wide association studies of cancer: current insights and future perspectives. *Nat. Rev. Cancer*. **17**, 692–704 (2017).

3. Hindorff, L. A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
4. Hormozdiari, F. et al. Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat. Genet.* **50**, 1041–1047 (2018).
5. Liu, Y. et al. Noncoding RNAs regulate alternative splicing in Cancer. *J. Exp. Clin. Cancer Res.* **40**, 11 (2021).
6. Pal, S. et al. Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. *Genome Res.* **21**, 1260–1272 (2011).
7. Alfonso-Gonzalez, C. & Hilgers, V. Alternative) transcription start sites as regulators of RNA processing. *Trends Cell Biol.* **34**, 1018–1028 (2024).
8. Shabalina, S. A., Ogurtsov, A. Y., Spiridonov, N. A. & Koonin, E. V. Evolution at protein ends: major contribution of alternative transcription initiation and termination to the transcriptome and proteome diversity in mammals. *Nucleic Acids Res.* **42**, 7132–7144 (2014).
9. Rojas-Duran, M. F. & Gilbert, W. V. Alternative transcription start site selection leads to large differences in translation activity in yeast. *RNA* **18**, 2299–2305 (2012).
10. Carninci, P. et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38**, 626–635 (2006).
11. Kimura, K. et al. Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.* **16**, 55–65 (2006).
12. Wang, X., Hou, J., Quedenau, C. & Chen, W. Pervasive isoform-specific translational regulation via alternative transcription start sites in mammals. *Mol. Syst. Biol.* **12**, https://doi.org/10.15252/msb.20166941 (2016).
13. Leppek, K., Das, R. & Barna, M. Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nat. Rev. Mol. Cell Biol.* **19**, 158–174 (2018).
14. Demircioglu, D. et al. A pan-cancer transcriptome analysis reveals pervasive regulation through alternative promoters. *Cell* **178**, 1465–1477 (2019).
15. Schuster, S. L. & Hsieh, A. C. The Untranslated Regions of mRNAs in Cancer. *Trends Cancer.* **5**, 245–262 (2019).
16. Haberle, V. et al. Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature* **507**, 381–385 (2014).
17. Pruunsild, P., Kazantseva, A., Aid, T., Palm, K. & Timmusk, T. Dissecting the human BDNF locus: bidirectional transcription, complex splicing, and multiple promoters. *Genomics* **90**, 397–406 (2007).
18. Hashimoto, K. et al. CAGE profiling of ncRNAs in hepatocellular carcinoma reveals widespread activation of retroviral LTR promoters in virus-induced tumors. *Genome Res.* **25**, 1812–1824 (2015).
19. Zhao, Z. et al. Pan-cancer transcriptome analysis reveals widespread regulation through alternative tandem transcription initiation. *Sci. Adv.* **10**, eadl5606 (2024).
20. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
21. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
22. Delaneau, O. et al. A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* **8**, 15452 (2017).
23. Hebenstreit, D., Wirnsberger, G., Horejs-Hoeck, J. & Duschl, A. Signaling mechanisms, interaction partners, and target genes of STAT6. *Cytokine Growth Factor Rev.* **17**, 173–188 (2006).
24. Greer, P. Closing in on the biological functions of Fps/Fes and Fer. *Nat. Rev. Mol. Cell Biol.* **3**, 278–289 (2002).
25. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
26. Wang, J. et al. Imputing gene expression in uncollected tissues within and beyond GTEx. *Am. J. Hum. Genet.* **98**, 697–708 (2016).
27. Grundberg, E. et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat. Genet.* **44**, 1084–1089 (2012).
28. Gao Wang, A. S., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).
29. Zhang, Z. et al. Genetic analyses support the contribution of mRNA N(6)-methyladenosine (m(6)A) modification to human disease heritability. *Nat. Genet.* **52**, 939–949 (2020).
30. Li, Y. I. et al. RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600–604 (2016).
31. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
32. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019).
33. Wen, X. Molecular QTL discovery incorporating genomic annotations using Bayesian false discovery rate control. *Annal. Appl. Stat.* **10**, 1619–1638(2016).
34. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).
35. Li, L. et al. An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. *Nat. Genet.* **53**, 994–1005 (2021).
36. Keinan, A. & Clark, A. G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**, 740–743 (2012).
37. Consortium, U. K. et al. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
38. Nelson, M. R. et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**, 100–104 (2012).
39. Ferraro, N. M. et al. Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science* **369**, https://doi.org/10.1126/science.aaz5900 (2020).
40. Li, X. et al. The impact of rare variation on gene expression across tissues. *Nature* **550**, 239–243 (2017).
41. Zou, X. et al. Impact of rare non-coding variants on human diseases through alternative polyadenylation outliers. *Nat. Commun.* **16**, 682 (2025).
42. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).
43. Moon, E. J. et al. The HIF target MAFF promotes tumor invasion and metastasis through IL11 and STAT3 signaling. *Nat. Commun.* **12**, 4308 (2021).
44. Mancuso, N. et al. Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. *Nat. Commun.* **9**, 4079 (2018).
45. Gamazon, E. R. et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
46. Wright, F. A. et al. Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.* **46**, 430–437 (2014).
47. Carithers, L. J. et al. A novel approach to high-quality postmortem tissue procurement: the GTEx project. *Biopreserv. Biobank.* **13**, 311–319 (2015).
48. Consortium, G. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

49. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
50. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
51. Zou, X. et al. Alternative tandem transcription initiation links non-coding variants to human disease through translational control. *Zenodo* https://doi.org/10.5281/zenodo.21287301 (2026).

Acknowledgements

We thank members of the Li laboratory for helpful discussions and the Shenzhen Bay Laboratory High-Performance Computing and Informatics Core Facility for high-computing support.

Author contributions

L.L. conceived and designed the study. L.L. and X.Z. acquired funding. L.L. was responsible for project administration. X.Z. curated the data and performed bioinformatic analyses with help from Y.W. and H.C. W.W., F.W., S.C., and Y.L. conducted the experiments. X.Z. and W.W. draw all figures with help from Y.W., H.C. and X.L. X.Z. and W.W. wrote the original draft of the manuscript; Q.L., L.L., and Y.C. reviewed and comment on the manuscript, L.L., X.Z., W.W., Y.W., and H.C. revised the manuscript; all authors contributed to review and editing.

Funding

This work was funded by grants from the Shenzhen Medical Research Fund (grant no. C2503001 to L.L., grant no. B2401002 to C.Y), the Major Program of Shenzhen Bay Laboratory (grant no. C1012524001 to L.L.), and the National Natural Science Foundation of China (grant no. 32370721, 32100533 to L.L., grant no. 32570730 to X.Z.), the Guangdong Basic and Applied Basic Research Foundation (grant no. 2024A1515012982 to W.W.), the Shenzhen Clinical Research Center for Oral Diseases (grant no. 20210617170745001 to X.Z.).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-76741-8.

Correspondence and requests for materials should be addressed to Chen Yu or Lei Li.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.