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

Xudong Zou[1#], Wei Wang[2#], Xing Li[1], Yinuo Wang[1], Hui Chen[1], Shuxin Chen[1], Feihong Weng[3], Qin Li[4], Chen Yu[2*], Lei Li[1]*

[1] Institute of Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen 518055, China.

[2] Institute of Cancer Research, Shenzhen Bay Laboratory, Shenzhen 518055, China.

[3] Liver Transplant Center, Organ Transplant Center, West China Hospital of Sichuan University, Chengdu 610000, China

[4] Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

# The authors have contributed equally

*Corresponding author: yu@szbl.ac.cn, lei.li@szbl.ac.cn

Abstract

Alternative tandem transcription initiation (ATI) is a pervasive mechanism of gene regulation, yet its genetic impact on human disease remains largely unknown. Here, we systematically quantified genetic regulation of ATI across 25,859 samples from 49 human normal tissues and 33 tumor tissues. We identified ~0.4 million 5′ UTR ATI quantitative trait loci (5′aQTLs) in 5,295 genes, with 32% operating independently of gene expression. Moreover, we discovered 2,238 multi-tissue ATI 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 5′aQTLs cannot be identified by eQTLs. Transcriptome-wide association studies identified 614 ATI-mediated disease susceptibility genes, including known cancer drivers such as *MAFF* and *MLLT10*. Functional validation uncovered *OSGEP* as a novel breast cancer risk gene, where the alternative allele lengthens the 5′ UTR and reduces protein abundance through upstream open reading frame (uORF)-mediated translation repression, and suppresses breast cancer cell proliferation. Our findings establish ATI as a major, underappreciated mechanism linking noncoding variation to disease, providing a critical resource for interpreting disease risk loci.

Introduction

Genome-wide association studies (GWAS) have identified thousands of genetic variants to 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 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 6,829 ATTSS events in 5,295 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 approximately 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 novel 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 previously underappreciated molecular 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). 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 9,249 and 7,799 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, and hidden technical covariates inferred using probabilistic estimation of expression residual (PEER)[21] factors (Supplementary Fig. 2a, b), 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 4,679 genes across 49 normal tissues, representing approximately 47.2% of all expressed genes with multiple TSSs (Fig. 1b). Additionally, we identified 0.13 million genetic variants associated with DTUI of 3,818 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). 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 5′aQTL located in its 5′ UTR (rs4806847, A/G; *P*=3.0e-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. To account for linkage disequilibrium (LD) among the identified 5′aQTLs, we used the sum of single effects (SuSiE)[26] 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 is highly tissue-specific

We next examined the tissue specificity of the genetic regulation of alternative tandem transcription initiation. Using a metric adapted from Story's π_1 metrics[27-29] (weighted π_1), we quantified the degree of 5′aQTL sharing between tissues. Hierarchical clustering of the weighted π_1 values revealed that biologically similar tissues clustered together (Fig. 2a). For example, all brain tissues exhibited high correlation with one another. This pattern indicates that our weighted π_1 metrics effectively captured the biological tissue differences underlying alternative transcription initiation regulation. 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 $P < 2.2\times10^{-16}$; Fig. 2b), indicating substantially higher tissue specificity of 5′aQTLs compared to eQTLs.

To further characterize the tissue specificity, we identified 3,012 5′aGenes (55.9%) detected in only one tissue (Fig. 2c). These tissue-specific 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 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. 5a-c). Similarly, potassium channel tetramerization domain 7 (*KCTD7*) exhibited testis-specific 5′aQTL absent from all other tissues (Supplementary Fig. 5d-f). We further examined these brain-specific 5′aGenes across the 13 brain tissues and found that 84.2%

were only present in one brain tissue (Fig. 2e), indicating that most brain-specific 5′aGenes are also region-specific.

In summary, the genetic regulation of alternative transcription initiation is widespread yet highly tissue-specific. Compared with eQTLs, 5′aQTLs show substantially lower cross-tissue sharing, with over half of regulated genes displaying tissue-restricted effects, particularly in the brain and testis. 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, we systematically compared 5′aQTLs with eQTLs across 49 tissues. We found that, on average, 48.9% of 5′aGenes were not identified as eGenes. Moreover, even among the shared 5′aGenes and eGenes (average π_1 value of 0.62 *vs.* 0.04, P=3.5×10^{-15}; Supplementary Fig. 6a), 71.3% had different causal variants as exemplified by 5′aGene *MYEOV* (myeloma overexpressed) (Supplementary Fig. 6b,c), the lead SNP (rs72930287) was significantly (P=8.8e-9) associated with ATTSS rather than gene expression of this gene, suggesting independent *cis*-regulation of 5′ UTRs relative to gene expression. To further characterize the molecular features distinguishing 5′aQTLs from eQTLs, we performed functional annotation enrichment analysis. We classified 5′QTL variants based on functional annotations and used Torus[30] 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. 2e). Both 5′QTL variants and eQTL variants were highly enriched in regions near TSSs (Fig. 2f-g). We further applied fine-mapping using CAVIAR[31] 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. 2h). Moreover, 5′aQTLs were highly enriched in conserved regions (phastCons conservation score > 0.9) (Fig. 2i), suggesting strong selective constraint on variants affecting transcriptional initiation. We also observed that 5′aGenes and eGenes differ considerably in their genomic architecture. Compared with eGenes, 5′aGenes possessed shorter coding sequences ($P<2.2\times10^{-16}$) and 3′UTR (P=0.04) but significantly longer 5′ UTRs ($P<2.2\times10^{-16}$) (Fig. 2j).

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. 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. 2e), 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[32] 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 in ETS family and SP1 (Fig. 3b).

To characterize how 5′aQTLs modulate transcription initiation by altering transcription factor bindings, we intersect fine-mapped 5′aQTLs with transcription factor binding sites and identified 169 5′aQTLs that were located in known transcription factor binding sites (Fig. 3c; Supplementary Table 1). 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. 7a, 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. 7c), confirming that MAZ binding at this site functionally regulates *RBM33* transcription initiation. In addition to validating known transcription initiation factors, our analysis also uncovered novel 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 that E2F4 knockdown resulted in increased distal TSS usage of *OSGEP* in both cell lines (Fig. 3g,h), suggesting a new 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 novel 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[33], and many have been found to contribute to both rare and common diseases[34,35]. 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[36,37]. We identified 2,238 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 1,735 ATTSS events (1,532 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)[38], the majority of 5′aOutliers exhibited positive Z scores, indicating that these genes have longer 5′ UTR in outlier individual(s) compared with the majority of individuals (Supplementary Figure 8).

To test whether nearby rare variants in outlier individuals contributed to 5′aOutlier events. We systematically analyzed the genomic context of these outliers. We identified 3,830 rare variants within 10kb of the first exon region associated with multi-tissue 5′aOutliers (Fig. 4a; Supplementary Table 2). 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 were 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 Table 3). Using functional genome-wide association analysis (fgwas)[39], 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). Comparison 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.

To systematically characterize 5′aQTLs that share causal variants 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 Table 3) 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 Table 4). In total, we identified 714 significant colocalized signals spanning 76 of the 5′aQTL genes for the non-cancer GWAS loci and 385 colocalized signals spanning 34 of the 5′aQTL genes across the cancer GWAS loci (Fig. 5b, Supplementary Fig. 9; Supplementary Table 5) 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[40], exhibited colocalization with breast cancer GWAS signals (Supplementary Table 5). 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 Table 5). 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 linking genetic variation in noncoding regions to tissue-specific disease mechanisms.

To systematically identify disease susceptibility genes associated with ATTSS events, we adapted FUSION[41] 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 5,025 tissue-specific 5′aTWAS prediction models from GTEx reference panels, spanning 2,580 ATTSS events. The number of 5′aTWAS prediction models was highly correlated with the sample sizes of the reference panels (Supplementary Fig. 10a). The average in-sample prediction accuracy (measured by heritability-normalized R^2, R^2/cis-h^2) of 5′aTWAS models was 0.67 for GTEx (Supplementary Fig. 10b), 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. 11). To enhance the detection of disease susceptibility genes mediated by 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 novel susceptibility genes, demonstrating that alternative TSS regulation provides a complementary and mechanistically distinct path linking genetic variation to cancer risk.

Alternative TSS regulation of OSGEP links genetic variation to breast cancer risk

Through both colocalization and TWAS analysis, we identified O-sialoglycoprotein endopeptidase (*OSGEP*) as a novel 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. Fine-mapping at this locus revealed 5′aQTL rs3120064 that disrupts transcription factor E2F4 binding (Fig. 3d; Fig. 6a-b) and exhibits strong linkage disequilibrium (LD R^2=1.0 in EUR) with the breast cancer risk variant rs3120073 (Supplementary Fig. 12a). 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 (rs120064 C>A) (Fig. 6b; Supplementary Fig. 11b). 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). We observed that 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 (Supplementary Fig. 12b,c). We found that the short 5′ UTR exhibited significantly higher luciferase activity than the construct containing the long 5′ UTR (*P*=6.0e-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 (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 but had minimal effect on transcript levels (Supplementary Fig. 12d,e). 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 nontargeting 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. Finally, we analyzed survival in breast cancer patients and found that it was strongly associated with *OSGEP* 5′ UTR length (HR=4.13; 95% CI:1.41-12.11) but not expression level (HR=1.35; 95% CI:0.48-3.82; Fig. 6l). Specifically, patients whose tumors exhibited preferential usage of the long 5′UTR (corresponding to lower OSGEP protein levels) showed significantly worse survival outcomes, consistent with the tumor-suppressive role of *OSGEP* suggested by our functional studies. Importantly, *OSGEP* mRNA expression levels alone did not predict survival, reinforcing that 5′ UTR-mediated translational control, rather than transcriptional regulation, represents the key disease-relevant mechanism. 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. 6m).

Collectively, these findings identify several cancer susceptibility genes linked to 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 the first 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. Our analyses demonstrate that genetic control of alternative TSS usage is highly tissue-specific. 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.

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

This study complies with all relevant ethical regulations. Collection and use of the human genotype and RNA-seq data from GTEx in this study were approved by the GTEx consortium. The ethical procedures for data collection, including informed consent and ethics approval, have been described in the original GTEx project documentation and publications[42]. No additional ethics review was required for this study.

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 RNA-seq and genotype data from The Cancer Genome Atlas (TCGA) 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. They 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 in the same strand. In that case, the TSS is considered a tandem TSS of the transcript. Transcripts with multiple TSS 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). 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 will be assigned as “NA” for the corresponding samples.

Covariate correction

We first corrected the sample genotype 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. 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 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 normalized DTUI values and the genotype of SNPs within an interval of 1 Mbp from 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 1,000 rounds of permutation (with the parameter "--permutation 1000") to obtain adjusted *P* values for the association between the normalized DTUI and the top variants in *cis*.

Fine-mapping of causal variants to 5′aQTL

Fine-map of causal variants for 5′aQTL was conducted using SuSiE software[26]. In brief, SuSie can operate 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.

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 Story’s π_1 method[29] (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. We used a bootstrap approach to compute confidence intervals for our π_1 estimation. Briefly, we resampled *P*-values with replacement *N* times (where *N* is the initial number of samples) and recomputed the π_1 estimate.

Tissue 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 in tissue B. In brief, tissue sharing was measured by multiplying the proportion of shared SNP-gene associations between tissues by Storey's p-value.

SNP heritability estimation 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 Table 3) from the UK Biobank and the literature. Colocalization analysis of 5′aQTL and GWAS signals was conducted with Coloc (v5.2.3) software[43]. 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[32].

Building 5′aTWAS model

We used FUSION[44] 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.

Applying 5′aTWAS prediction models to GWAS summary statistics

We applied our 5′aTWAS models to the 84 sufficiently powered GWAS summary statistics (Supplementary Table 6) 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 link the

ATTSS to disease risks.

Cell culture

HEK 293T and MCF-7 (ATCC-originated) 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 (rs120064 C) or alternative alleles (rs120064 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 Table 6.

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 (rs120064 C) or alternative allele (rs120064 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 Table 6.

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 Table 6.

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 Table 6) 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 Table 6) 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 Table 6.

RNA sequencing library construction

To investigate the impact of E2F Transcription Factor 4 (E2F4) on tandem TSS usage of *OSGEP* promoter, MCF-7 cells were transfected with shE2F4 and 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.

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/), Finn Gen (https://www.finngen.fi/en), and JENGER(http://jenger.riken.jp), with a complete list of GWAS summary statistics in Supplementary Table 3.

Code availability

Code for detecting and quantifying alternative tandem TSS events is available on GitHub (https://github.com/ZhaozzReal/DATTSS).

Acknowledgments

We thank members of the Li laboratory for helpful discussions. This work was funded by grants from the Major Program of Shenzhen Bay Laboratory (grant no. C1012524001 to L.L.), the Shenzhen Medical Research Fund (grant no. C2503001 to L.L.), and the National Natural Science Foundation of China (grant no. 32370721, 32100533 to L.L., grant no. 32570730 to X.Z.). We also thank the Shenzhen Bay Laboratory High-Performance Computing and Informatics Core Facility for high-computing support.

Author contributions

L. L. and Y.C. conceived and designed the project. X.Z., L.X., Y.N., and C.H. collected the data and performed the analyses. W. W., W. F., and C.S. conducted the experiments. X.Z., W.W., and C.S. wrote the manuscript; L.L. and Y.C. reviewed and revised the manuscript.

Declaration of interests

The authors declare no competing interests.

Reference

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 U S A* **106**, 9362-7 (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**, 1260-72 (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**, 7132-44 (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**, 2299-305 (2012).

10. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. **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. **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. *Molecular Systems Biology* **12**(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 e17 (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**, 1812-24 (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**, 500-7 (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**, 173-88 (2006).

24. Greer, P. Closing in on the biological functions of Fps/Fes and Fer. *Nat Rev Mol Cell Biol* **3**, 278-89 (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. Gao Wang, A.S., Peter Carbonetto and Matthew Stephens. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82**, 1273-1300 (2020).

27. 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).

28. Li, Y.I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600-4 (2016).

29. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440-5 (2003).

30. Wen, X. Molecular QTL discovery incorporating genomic annotations using Bayesian false discovery rate control. *The Annals of Applied Statistics* **10**(2016).

31. 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).

32. 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).

33. Keinan, A. & Clark, A.G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336**, 740-3 (2012).

34. Consortium, U.K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90 (2015).

35. Nelson, M.R. *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**, 100-4 (2012).

36. Ferraro, N.M. *et al.* Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science* **369**(2020).

37. Li, X. *et al.* The impact of rare variation on gene expression across tissues. *Nature* **550**, 239-243 (2017).

38. Zou, X. *et al.* Impact of rare non-coding variants on human diseases through alternative polyadenylation outliers. *Nat Commun* **16**, 682 (2025).

39. Pickrell, J.K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am J Hum Genet* **94**, 559-73 (2014).

40. Moon, E.J. *et al.* The HIF target MAFF promotes tumor invasion and metastasis through IL11 and STAT3 signaling. *Nat Commun* **12**, 4308 (2021).

41. Mancuso, N. *et al.* Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. *Nat Commun* **9**, 4079 (2018).

42. Carithers, L.J. *et al.* A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreserv Biobank* **13**, 311-9 (2015).

43. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).

44. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* **48**, 245-52 (2016).

Figures and legends

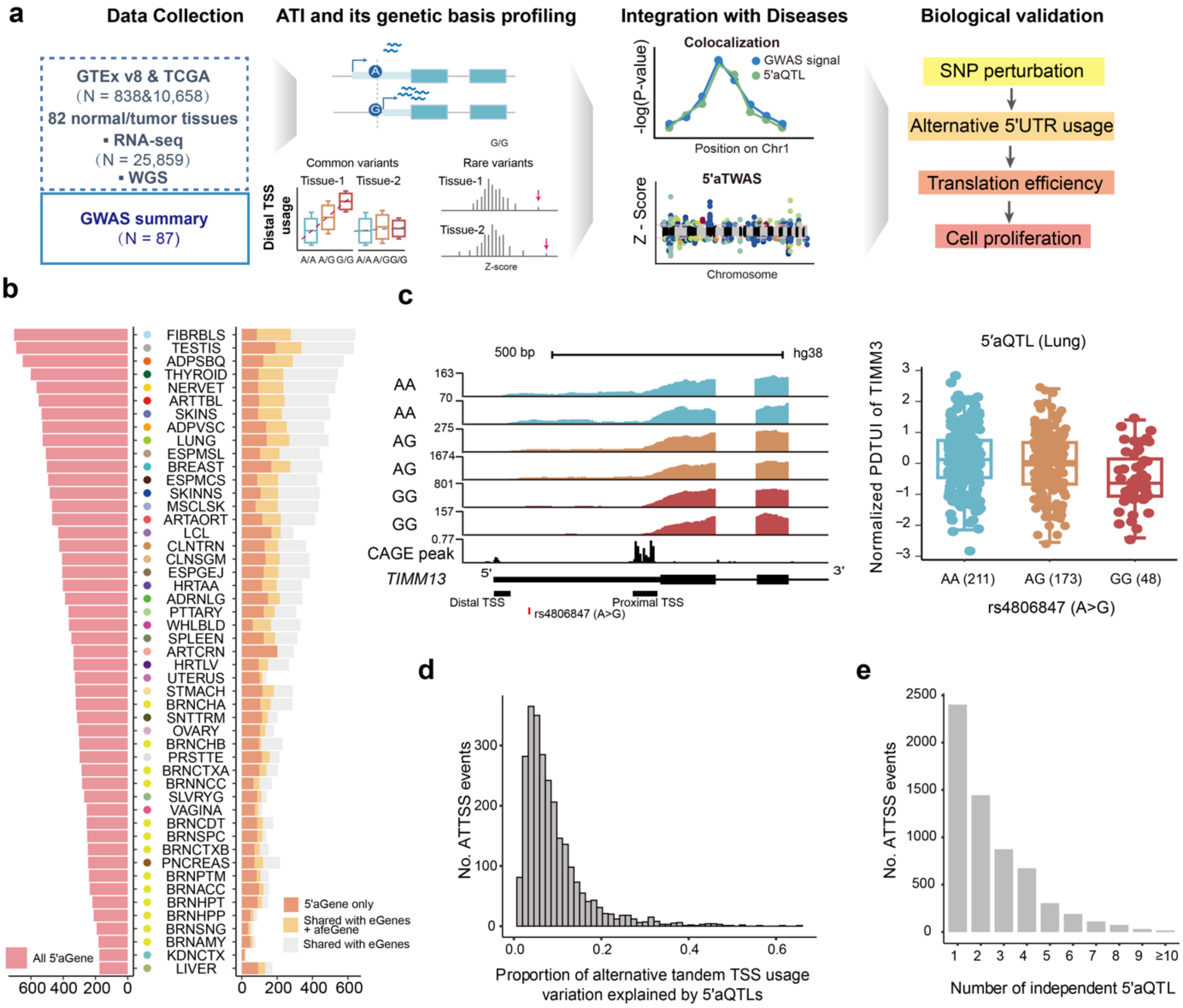

Fig. 1. A multi-tissue atlas of genetic regulation of alternative transcription initiation. a, Overview of the data and analyses in this study. **b,** Distribution of the number of 5′aGene for each normal tissue, sorted by the 5′aGene number and number of 5′aGenes shared by eQTL or afeQTL. 5′aGene only (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, grey 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. Top: 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. Bottom: Distribution of the normalized DTUI for each genotype. Each dot in the box plot represents the normalized DTUI value for one particular sample ($n = 432$). The center horizontal lines within the plot represent the median values, and the boxes span from the 25th to the 75th percentile. **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.

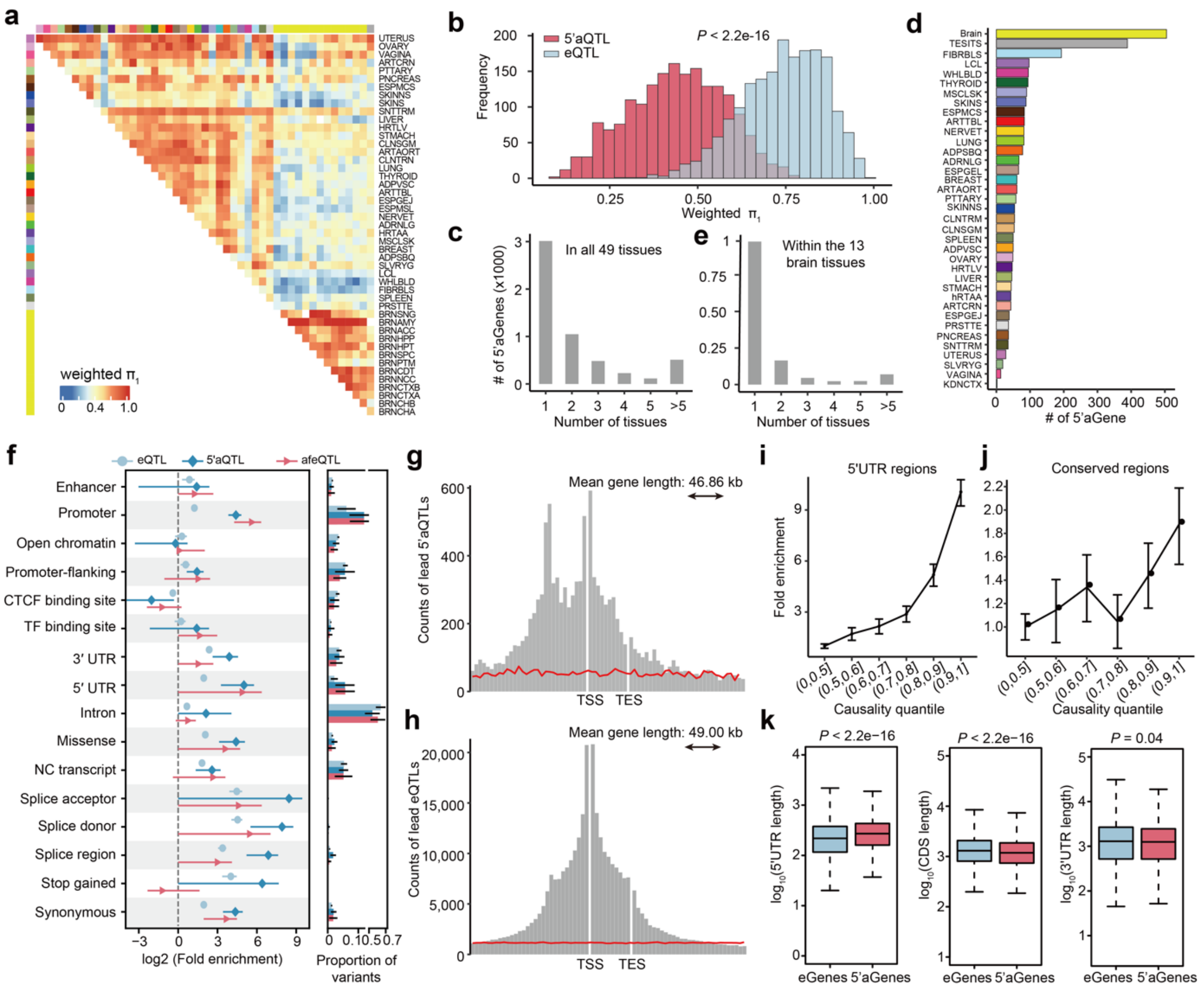

Fig. 2. Tissue-specific 5'aQTLs and distinct genetic variation beyond eQTLs. a, Pairwise 5′aQTL sharing among tissues. The degree of sharing for significant 5′aQTL signals between tissues was calculated by weighted π_1. **b**, The frequency of weighted π_1 for pairwise 5′aQTL and eQTL sharing across tissues. *P* value was obtained from 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. Fold enrichment was shown as mean (dot) ± standard deviation ($\log_2$ scaled, error bar) across 49 tissues. The error bars in the right panel indicate standard deviation of the proportion of variants across 49 tissues. **g-h**, Relative distances between lead 5′aQTLs (**g**) and eQTLs (**h**) and their associated genes. TSS, transcription start site; TES, transcription end site. The red line represents randomly selected positions within the ±1Mb window for each gene. **i**, Fold enrichment and 95% confidence intervals (CIs) for 5′aQTLs in each causality bin for the intersection with 5′ UTR regions. **j**, Fold enrichment and 95% CIs of 5′aQTLs that intersect with conserved regions, which were defined as regions with UCSC

phastCons conservation scores >0.8. **k**, Comparison of 5′ UTR, CDS length and 5′ UTR length between 5′aGenes and eGenes. *P* values were calculated using a two-sided *t*-test. The center horizontal lines of the box plot show the median values and the boxes span from the 25th to the 75th percentile.

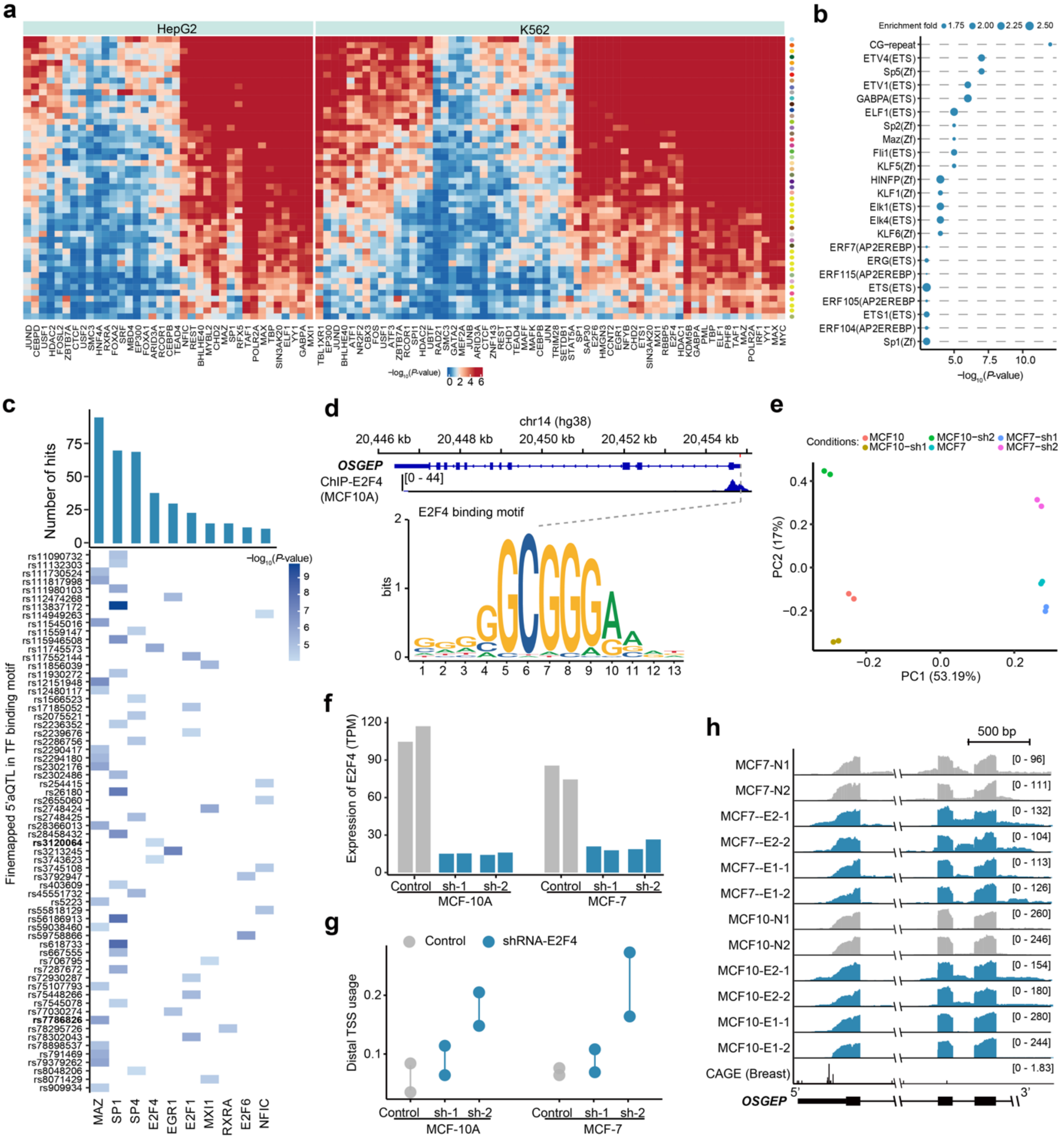

Fig. 3. Transcription factor binding and chromatin context underlie 5′aQTL regulation. a, Heatmap showing the 5′aVariant significance for TFs identified by ENCODE in each tissue. The right bar shows the color code for each tissue; the top color bar represents the HepG2 and K562 cell lines, separately. Values in the heatmap represent the degree of enrichment for 5′aQTLs in TF binding sites compared with the control. **b**, The TF motif enrichment using de novo motif scans. Fold enrichment is shown as mean (dot). **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 represent the corresponding -log$_{10}$(P-value). **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 MAZ 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.

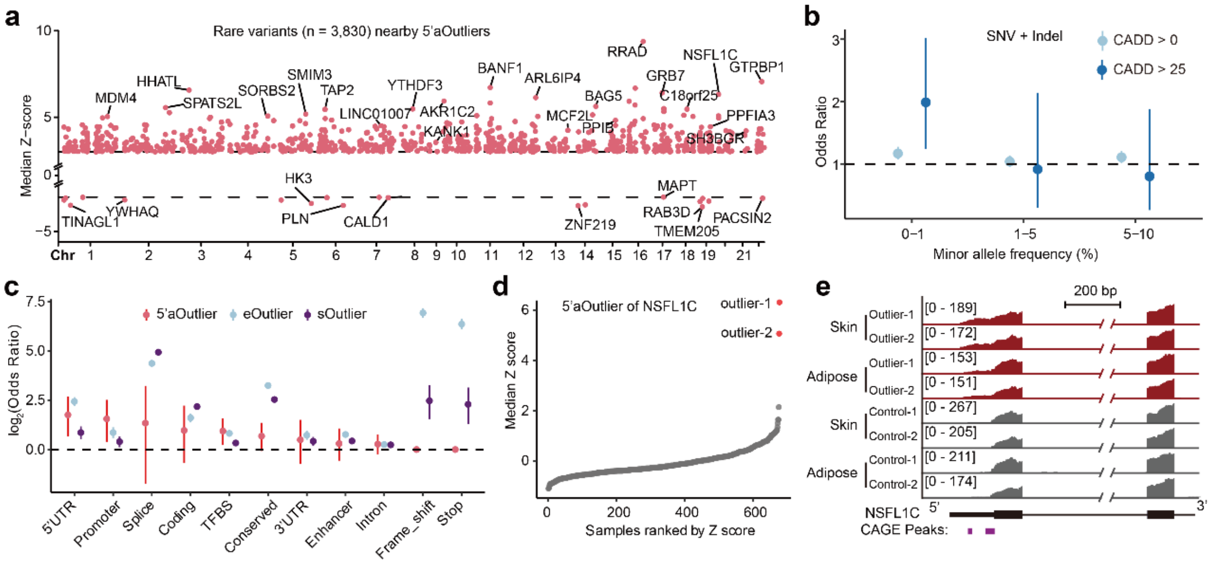

Fig. 4. Rare variants influence tandem TSS usage across human tissues. a, Distribution of 5′aOutliers across the human genome. Genes with the highest (for positive median Z-scores) or lowest (for negative median Z-scores) Z-score at each chromosome region were labeled. **b,** Enrichment of deleterious single-nucleotide variants (SNVs) and small INDELs in 5′aOutliers; variants within 1 kb of 5′aOutlier genes were counted. Data are presented as ORs and 95% CIs. **c,** Enrichment of rare variants of different categories in 5′aOutlier (red), eOutlier (light blue), and sOutlier (purple). **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 (green) in the Skin and Adipose tissues.

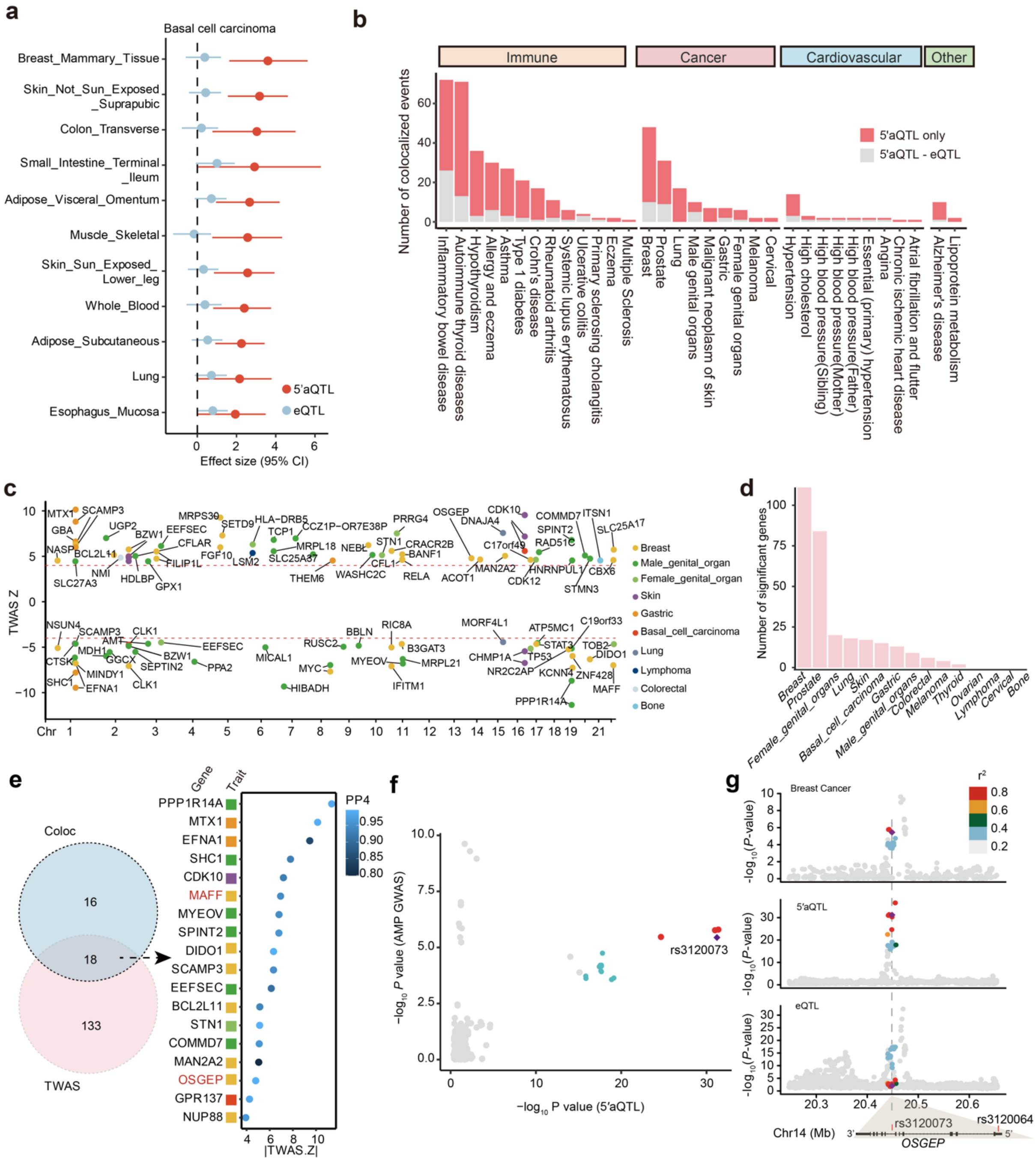

Fig. 5. Genetic regulation of transcription initiation contributes to complex trait heritability. a, Tissues with 5′aVariant enrichment but no eQTL enrichment for basal cell carcinoma traits. The enrichment values (effect size) were calculated using functional genome-wide association analysis, which quantifies the relationships between trait-associated variants and 5′aQTLs/eQTLs. The estimated lower and upper bound 95% CIs for the enrichment value are also shown. **b**, Number of 5′aQTL-colocalized genes, colored based on whether the gene also colocalized with eQTLs. Grey 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. **f,** Correlation of the *P* values of 5′aQTL associations for *OSGEP* and matched *P* values in breast cancer GWAS. **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).

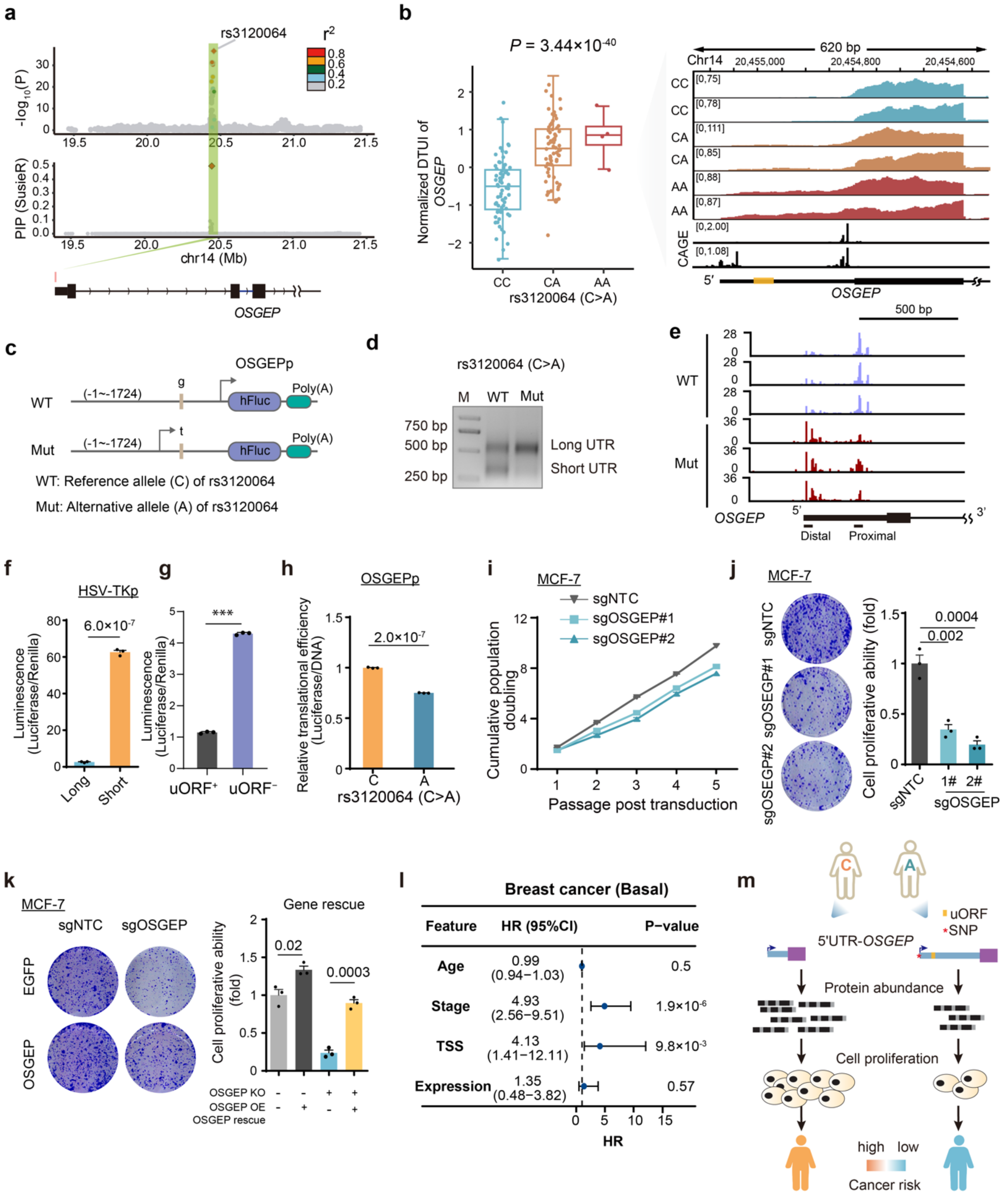

Fig. 6. Alternative TSS regulation of OSGEP links genetic variation to breast cancer risk. a, Regional association plots of the *OSGEP* gene showing the nominal *P* values (top) and fine-mapping posterior probabilities (bottom) for the *cis*-SNPs. **b,** 5′aVariant (rs3120064) that is strongly associated with the *OSGEP* 5′ UTR usage. Left: Distribution of the normalized DTUI for each genotype. Each dot in the box plot represents the normalized DTUI value for one particular sample. The center horizontal lines within the plot represent the median values and the boxes span from the 25th to the 75th percentile. 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. **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 293T cells. **g,** Luciferase activity from a reporter system containing the uORF and the one without the uORF 5′ UTR of *OSGEP* in HEK 293T cells. **h,** Luciferase activity from a reporter system containing reference allele (rs120064 C) or alternative allele (rs120064 A) of *OSGEP* promoter in HEK 293T cells, normalized to DNA copies. **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. **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. **l,** Forest plot displaying the results of multivariable cox regression analysis for distinct features in basal-type breast cancer. Each feature is shown with its hazard ratio (HR), 95% confidence interval (95% CI), and associated p-value. The vertical line at HR=1 indicates no effect. Features with a confidence interval that does not cross this line are considered statistically significant. **m,** Schematic depicts alternative alleles of 5′ UTR variants mediated 5′ UTR lengthening promotes cancer progression.