JJCO Japanese Journal of Clinical Oncology

Japanese Journal of Clinical Oncology, 2023, 53(2)97–104
https://doi.org/10.1093/jjco/hyac175
Advance Access Publication Date: 3 December 2022
Original Article

OXFORD

Original Article

Classification and characterization of alternative promoters in 26 lung adenocarcinoma cell lines

Yamato Hamaya[1,2], Ayako Suzuki[3], Yutaka Suzuki[3], Katsuya Tsuchihara[1,2,*] and Riu Yamashita[2,3,*]

[1]Department of Integrated Biosciences, Graduate School of Frontier Sciences, The University of Tokyo, Chiba, Japan, [2]Division of Translational Informatics, National Cancer Center, Exploratory Oncology Research and Clinical Trial Center, Chiba, Japan and [3]Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Chiba, Japan

*For reprints and all correspondence: Riu Yamashita, Division of Translational Informatics, National Cancer Center, Exploratory Oncology Research and Clinical Trial Center, 6-5-1 Kashiwanoha, Kashiwa, Chiba 277-8577, Japan. E-mail: riuyamas@east.ncc.go.jp; Katsuya Tsuchihara, Division of Translational Informatics, National Cancer Center, Exploratory Oncology Research and Clinical Trial Center 6-5-1 Kashiwanoha, Kashiwa, Chiba 277-8577, Japan. E-mail: ktsuchih@east.ncc.go.jp

Received 16 August 2022; Editorial Decision 7 October 2022; Accepted 21 October 2022

Abstract

Background: Genome-wide landscape of alternative promoter use remains unknown. We determined expression profiles of promoters in 26 lung adenocarcinoma cell lines using the transcriptional start site-sequencing data and proposed an index 'canonical promoter usage' to quantify the diversity of alternative promoter usage.

Methods: Transcriptional start site-sequencing and other datasets were obtained from the DataBase of Transcriptional Start Sites. Transcriptional start site-sequencing read clusters were mapped onto RefGene to determine the promoters. Commonly used promoters were designated as canonical promoters. The sequence logos, CpG islands, DNA methylation and histone modifications of canonical and non-canonical promoters were examined. Canonical promoter usage was calculated by dividing 'read counts of a canonical promoter' by 'read counts of all the units of promoters' on each gene. The expressed genes were subjected to hierarchical clustering according to their canonical promoter usage.

Results: Among 104 455 promoters for 14 297 genes, 8659 canonical and 68 197 non-canonical promoters were identified. Corresponding to higher expression, canonical promoters showed core promoter sequences, higher CpG island positivity, less DNA methylation and higher transcription-promoting histone modifications. Gene ontology enrichment analysis revealed that the clusters with lower canonical promoter usage were related to signalling pathways, whereas clusters of tightly regulated genes with higher canonical promoter usage were related to housekeeping genes.

Conclusion: Canonical promoters were regulated by conventional transcriptional machinery, while non-canonical promoters would be targets of 'leaky' expression. Further investigation is warranted to analyse the correlation between alternative promoter usage and biological characteristics contributing to carcinogenesis.

Key words: alternative promoter, lung adenocarcinoma, cell line, transcriptional start site, canonical promoter usage

Introduction

Transcription is an essential step in reading genomic information. Transcriptional initiation occurs at multiple regions called 'alternative promoters', and 58% of 19 142 genes in the human genome have alternative promoters (1–3). Alternative promoters may cause diversity in transcriptional profiles and contribute to biological phenomena. For example, *TP73*, a member of the *TP53* tumour suppressor gene family, selectively uses alternative promoters. Upstream promoters code for transcriptionally active isoforms and serve as tumour suppressors. Conversely, downstream promoters code isoforms lacking an N-terminal transactivation domain and antagonize the active form to induce tumorigenesis (4,5). Alternative promoter usage has been reported to be associated with the clinical subtypes of breast cancer (6), suggesting critical roles for alternative promoters in tumour biology. However, the genome-wide landscape of alternative promoters in cancer genomes remains unclear.

Mammalian promoters contain closely separated transcription start sites (TSSs) (7). These promoters are regulated by various factors, including distal enhancers, DNA methylation and histone modifications (8–10). Multi-omics datasets of various cancer types are useful for studying transcriptional regulation (11–13). We have obtained datasets of TSS sequencing (TSS-seq) of 26 lung adenocarcinoma (LUAD) cell lines, along with data of whole genome sequencing, RNA sequencing, bisulfite-sequencing (BS-seq), chromatin immunoprecipitation sequencing (ChIP-seq) and assays for transposase-accessible chromatin sequencing (ATAC-seq), and they are publicly available via the DataBase of Transcriptional Start Sites (DBTSS) (11,14).

In this study, we determined the position and expression levels of all promoters in the genomes of 26 LUAD cell lines using TSS-seq data. BS-seq and ChIP-seq data were then overlaid on the TSS-seq data to characterize the regulatory features of alternative promoters. Further, we propose an index to quantify the diversity of alternative promoter usage to explore its biological significance.

Materials and methods

Dataset for multi-omics analysis

TSS-seq (DDBJ accession number, DRA005903), BS-seq (DRA001841), RNA-seq (DRA001846) and ChIP-seq (DRA001860) data of 26 LUAD cell lines were obtained from DBTSS ver.9.0 (Supplementary Table S1) (15). Cap Analysis of Gene Expression-sequence (CAGE-seq) read count data of two LUAD cell lines (A549 and PC14) were obtained from the Functional ANnoTation Of the Mammalian Genome website (http://fantom.gsc.riken.jp/5/datafiles/reprocessed/hg38_latest) (7). To align the CAGE-seq reads with the promoter regions defined from the TSS-seq reads, the regions were expanded 10 bp upstream and downstream to prevent miscounting of reads owing to a few base pair gaps between TSS-seq and CAGE-seq reads. All data were in tsv, csv, and bed format with coordinate information on the hg38 genomes.

Determination of promoters based on TSS-seq reads

Promoter regions based on TSS-seq reads were defined according to previous studies with some modifications (16). Briefly, all TSS-seq reads were mapped into the hg38 human genome by BWA (17). The mapped TSS-seq reads were clustered to construct TSS-seq read clusters (TSCs) within 500 bp intervals in the genomes for each 26 cell line. If a TSC overlapped with other TSCs, the TSC region was extended to include all neighbouring TSCs. All extended TSCs from 1000 bp upstream of the annotated transcriptional start sites to the transcriptional end site of each gene were mapped on the UCSC reference gene coordinate information (RefGene, version 5.6.26). The extended TSCs spanning more than 5000 bps length were omitted since visual inspection of several arbitrarily selected extremely long TSCs suggested to contain multiple promoters. When stratified by the length of the promoter, more than 95% of number of genes and TSS read counts were included below 5000 bp, thus this was used as the threshold. A TSS that corresponded to the highest TSS-seq reads in the promoter was defined as the representative TSS. Sequence logo analysis was performed around the representative TSS using the Python package Logomaker (https://github.com/jbkinney/logomaker) (18). We also used our original scripts written by Python to produce alternative promoters from TSS-seq reads.

Definition of canonical and non-canonical promoters

The promoters with the highest expression for each gene in at least 13 of the 26 LUAD cell lines were designated as 'canonical promoters' and other promoters of that gene were defined as non-canonical promoters. For the genes in which canonical promoters were not identified, canonical-like promoters were defined as the promoters with the highest expression for each gene in more than half of the cell lines in which the gene was expressed.

Extraction of CpG islands

The GC contents around representative TSSs with 250 bp upstream and downstream regions were calculated, and the ratio of observed CpG and expected CpG was estimated using the reference genome of GRCh38/hg38 (downloaded from the UCSC Genome Data). Promoters with more than 55% GC content and more than 0.65 observed CpG/expected CpG were identified as CpG island positive promoters (19).

Epigenetic analysis of promoter regions

TSS, BS, and ChIP-seq reads mapped information to the human genome 19 (hg19) in DBTSS were obtained, and the genome coordinates of hg19 were converted to hg38 using the UCSC LiftOver tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver) (20). The mapped TSS-seq and ChIP-seq read counts were normalized to Reads Per Million (RPM). The DNA methylation ratio was calculated using the methylated-C read count/(methylated-C read count + non-methylated-C read count) from the BS-seq. To estimate the DNA methylation ratio of the promoters, regions extended by 250 bp upstream and downstream from each representative TSSs were used in the calculation. ChIP-seq reads of RNA polymerase II and seven histone modifications (H3K4me1, H3K4me3, H3K27me3, H3K36me3, H3K9me3, H3K27ac and H3K9/14 ac) were mapped on an area of 1.5 kb upstream and downstream of determined promoters of representative TSSs. To score and visualize, the mapped ChIP-seq reads were normalized to $\log_2$ scale by 'bamCompare' and applied to 'computeMatrix' and 'plotHeatmap' of deeptools (version 3.5.0) (https://github.com/deeptools/deepTools) (21).

Definition of canonical promoter usage as the density of the expression of alternative promoters

For genome-wide analysis of alternative promoter usage, a unique quantitative value 'canonical promoter usage' (CPU) for each gene

was defined as below:

$$\mathrm{CPU} = \frac{P_{\mathrm{canonical}}}{\sum_{k=1}^{n} P_k},$$

where P_k is the total TSS-seq reads (RPM) belonging to promoter k in a gene and $P_{\mathrm{canonical}}$ is the total TSS-seq reads (RPM) on the canonical promoter in the same gene.

Clustering and enrichment analysis of the genes and cell lines

To segregate genes and cell lines based on the characteristics of CPU values, hierarchical clustering (Euclidean distance, Ward's method) was applied. Five gene clusters were classified using the *k*-means clustering method. Reads Per Kilobase of exon per Million mapped reads (RPKM) values based on RNA-seq for each of the 26 cell lines were obtained to estimate the RNA expression. Based on the UCSC refGene.txt file, the longest transcription length for each splicing variant was defined as the gene length. The *P* values of the statistical test were obtained with the CPU, normalized read count of the CPs and NCPs, promoter number, gene length and RPKM from RNA-seq on the five gene clusters by 'Steel-Dwass test' with scikit-posthocs (version 0.7.0) of Python package. The cut-off criterion was *P* value <0.01. Enrichment analysis of Gene Ontology (GO) was performed with the Metascape (22).

Results

Determination of promoters in genomes of 26 LUAD cell lines

We defined promoters based on TSS-seq reads, according to a previous study (Fig. 1a) (16). We used hg38 lift-overed TSS-seq reads of 33 164 208 (min: 5347033–max: 80455398) on average for genomes of 26 LUAD cell lines and obtained 767 560 (min: 288801–max: 2918312) transcriptional start sites. TSSs were clustered into a total of 187 584 TSCs and defined as promoters. Each cell line had an average of 159 087 (min: 107582–max: 357818) promoters in its genome. These promoters were annotated using RefSeq gene regions. However, 7010 promoters of 951 genes with extremely long regions were excluded as potential mixed promoters (Supplementary Fig. S1a and b). Finally, we obtained 104 455 promoters for 14 297 genes from the genomes of 26 cell lines (Fig. 1b and Supplementary Table S1). The defined promoters were compared with and validated by CAGE-seq data (23). CAGE-seq reads from A549 and PC14 cells were aligned to the promoter regions defined above. The results showed that 95.0% and 96.7% of CAGE-seq reads overlapped with TSS-seq-based promoters in A549 and PC14 cells, respectively. The correlation coefficients of CAGE-seq and TSS-seq expression levels on these promoters were 0.655 and 0.491 and statistically significant ($P < 0.001$) in A549 and PC14 cells, respectively (Supplementary Fig. S1c).

Among 19 014 RefSeq genes, an average of 9659 (min: 8865–max: 10465) genes were expressed in each cell line. Of 19 014 genes, 17.5–30.1% had a single promoter and 18.3–35.2% had multiple promoters (Fig. 1c). The highest expressed promoters tended to be detected at the 5′-end (first promoters) in the RefGene region (Supplementary Fig. S1d). For example, in the *STAU2* gene, which reportedly has multiple alternative promoters (6), a total of 36 promoters have been identified from 26 LUAD cell lines. The number of promoters in PC3 cells was 16, while those in the other LUAD cell lines ranged from one to six. Among the 36 promoters from 25 cell lines, the 5′-end promoters (p1: chr8, 73746374–73748017) showed the highest expression, but in PC3 cells, the p24 promoter (chr8: 73585207–73586432) showed the highest expression (Fig. 1d, Supplementary Fig. S1e). This suggested that even in the same cancer, the expression patterns of the alternative promoters differed in different cell lines.

Classification of canonical and non-canonical promoters in genomes of 26 LUAD cell lines

As described for the *STAU2* gene, the most highly expressed promoters were largely common among the cell lines. We designated these common promoters as canonical promoters (CPs). In this study, CPs were the promoters with the highest expression for each gene in at least 13 of the 26 LUAD cell lines, and other promoters of that gene were defined as non-canonical promoters (NCPs) (Fig. 2a). CPs were identified in 8659 genes. Meantime, 68 197 NCPs were identified in these genes. According to this definition, CPs were not able to be identified in 5638 genes since the promoters with highest expression were detected in less than 13 cell lines, mainly due to the lack of the expression of the genes in some cell lines. To further evaluate characters of the promoters in these genes, canonical-like promoters were defined as the promoters with the highest expression for each gene in more than half of the cell lines in which the gene was expressed. As a result, canonical-like promoters were identified in 4382 genes with 11 752 non-canonical-like promoters. Finally, 11 465 promoters of the 1256 genes remained as unclassified (Supplementary Fig. S2). Of identified CPs, 56.6% (4905 CPs) were commonly expressed in all 26 cell lines (Fig. 2b). In each cell line, an average of 7990 (min: 7083–max: 8345) out of 8659 CPs (92.3%) were expressed. In contrast, among the 68 197 NCPs, 52.0% (35 475 NCPs) were uniquely expressed in a single cell line (Fig. 2b). In each cell-line, an average of 10 146 (min: 5332–max: 17937) NCPs (14.9%) were expressed (Supplementary Table S2).

Expression levels and genomic contexts of CPs and NCPs

The expression levels of each CPs and NCPs had an average of 74.3 [standard deviation (SD) ± 3.5] and 0.609 (SD ± 0.173) RPM, respectively (Fig. 2c). To address the genomic background of the expression of CPs and NCPs, CpG islands in these promoters were estimated. We detected 79.7% (6908/8659 promoters) of CPs and 1.5% (1002/68197 promoters) of NCPs were located inside CpG islands (Fig. 2d). We further evaluated the DNA sequence patterns of the core promoter region and drew a sequence logo for a short region around the representative TSSs (±5 bp). In the CPs, pyrimidines and purines were dominant at the position one and position minus one of transcriptional start site, respectively. In contrast, no characteristic motifs were detected in NCPs (Fig. 2e). These results suggest that CPs are under the control of conventional and efficient transcriptional regulation, while NCPs may be expressed in a rather abnormal manner, resulting in less efficient transcription.

Epigenetic regulation of CPs and NCPs

Next, we evaluated the epigenetic modifications of CPs and NCPs. The DNA methylation ratios of the CPs and NCPs were evaluated. Among the expressed CPs, 6927 (SD ± 355, 80.0%) promoters were hypomethylated (methylation ratio < 0.25) and 188 (SD ± 25,

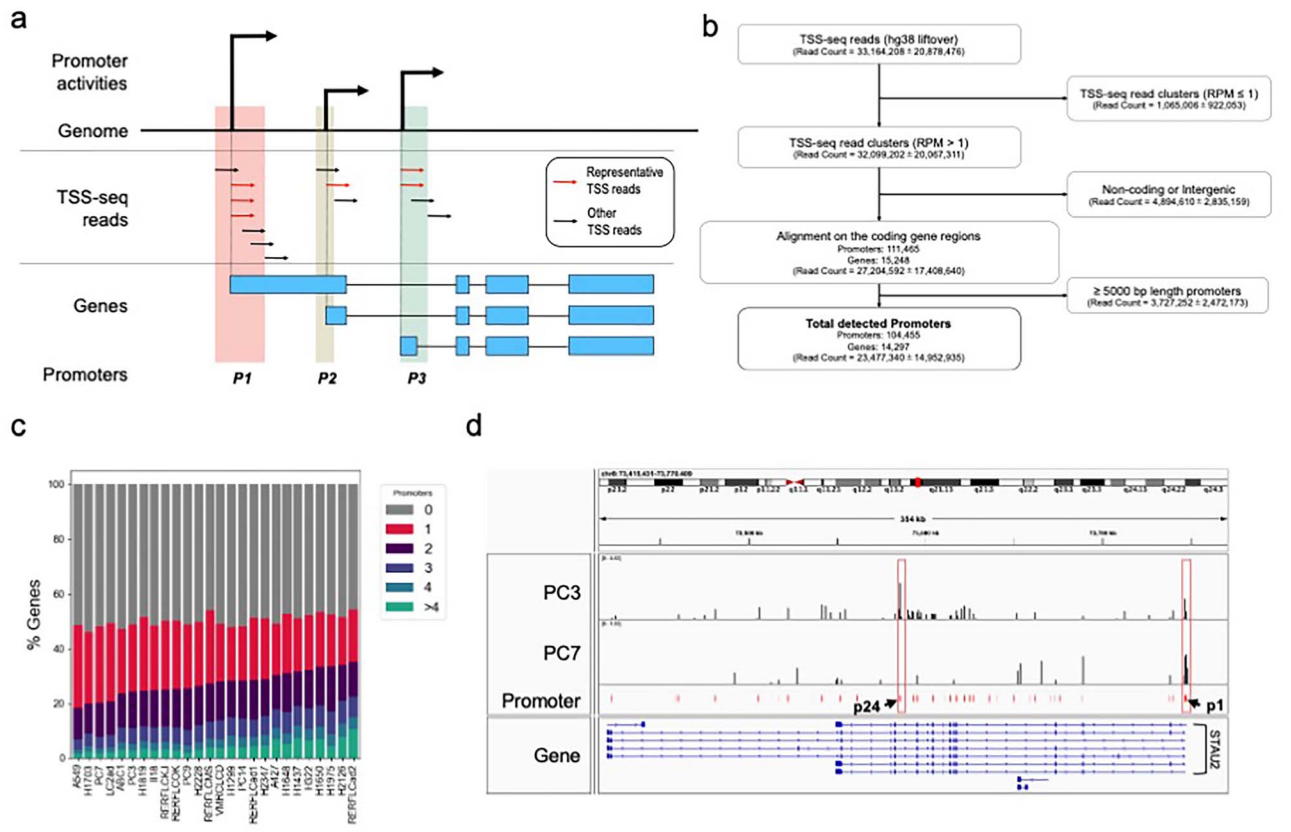

Figure 1. Overview of alternative promoters. (a) Definition of the promoters from transcriptional start site-sequencing (TSS-seq) data. Most frequent TSS reads in each promoter were defined as representative TSS reads (red allows). Other TSS (black arrows) reads within 500 bp bins were recognized as the reads within the same promoter. Promoter activity is calculated from the total number of TSS reads within each promoter region. (b) Flow of the extraction of the promoters in 26 lung adenocarcinoma (LUAD) cell lines based on the TSS-seq read data. (c) Percentage of the RefSeq genes according to the number of the determined promoters (0, 1, 2, 3, 4 and over 4 promoters) in the 26 LUAD cell lines. (d) Representative alternative promoter usage in *STAU2* gene (right to left) in PC3 and PC7 cells. The middle box shows the positions of stacked TSS reads and identified promoters. Red boxes indicate the positions of p1 and p24.

2.2%) promoters were hypermethylated (methylation ratio > 0.75), whereas among expressed NCPs, 627 (SD ± 88, 0.91%) were hypomethylated and 2406 (SD ± 713, 3.5%) were hypermethylated (Fig. 3a and Supplementary Table S3).

Examples of chromatin modifications in PC14 and RERF-LC-Ad2 cells are shown in Fig. 3b and Supplementary Fig. S3a, respectively. Around the representative TSSs of expressed CPs, ChIP-seq scores of Pol II and active histone marks, such as H3K4me3, H3K27ac and H3K9/14 ac, were elevated, whereas repressive marks (H3K27me3 and H3K9me3), markers of gene bodies (H3K36me3) and enhancers (H3K4me1) were low. Using a cut-off value of 0.5 with each normalized ChIP-seq reads in the region near the promoter as calculated by computeMatrix in deeptools, 94.87% of expressed CPs in PC14 cells had transcription-promoting histone modifications, and these marks highly overlapped with each other (Fig. 3c). This trend was observed across all cell lines, with 94.58% of the expressed CPs of 26 cell lines having transcription-promoting histone modifications (Supplementary Fig. S3b). These findings are consistent with the hypothesis that CPs are subjected to conventional transcriptional regulation. However, even in the expressed NCPs, gene body marks were predominant. Active histone marks were detected in 6.95% of the expressed NCPs of PC14 and 15.58% of 26 cell lines, and the overlap of each marker was much less than that in the CPs (Fig. 3c and Supplementary Fig. S3b). Similar results were obtained for canonical-like promoters. Active histone marks were more frequently detected in the expressed canonical-like promoters (73.53%) than in the expressed non-canonical-like promoters (30.19%) (Supplementary Fig. S3c).

Diversities of histone modifications among NCPs

Although canonical promoters were suggested to play important roles, it was anticipated that some of the non-canonical promoters would also be functionally significant. We focused on the cell lines in which those canonical promoters were not expressed and selected the most highly expressed non-canonical promoters in those genes in those cell lines. The proportion of promoters with active histone marks was increased in these non-canonical promoters with the highest expression (29.88%) compared with that of total non-canonical promoters (15.58%) (Supplementary Fig. S4a).

Next, we compared the histone modifications of non-canonical promoters commonly expressed in 26 cell lines (indicated by the open arrowhead in Fig. 2b) to those of non-canonical promoters uniquely expressed in 1–3 cell lines (indicated by the closed arrowhead in Fig. 2b). The proportion of active histone marks in commonly expressed non-canonical promoters (27.45%) were higher than those in uniquely expressed non-canonical promoters (14.87%) (Supplementary Fig. S4b). Though the proportion of active histone marks of those specific non-canonical promoters was still lower than those of canonical promoters, these findings suggest that some

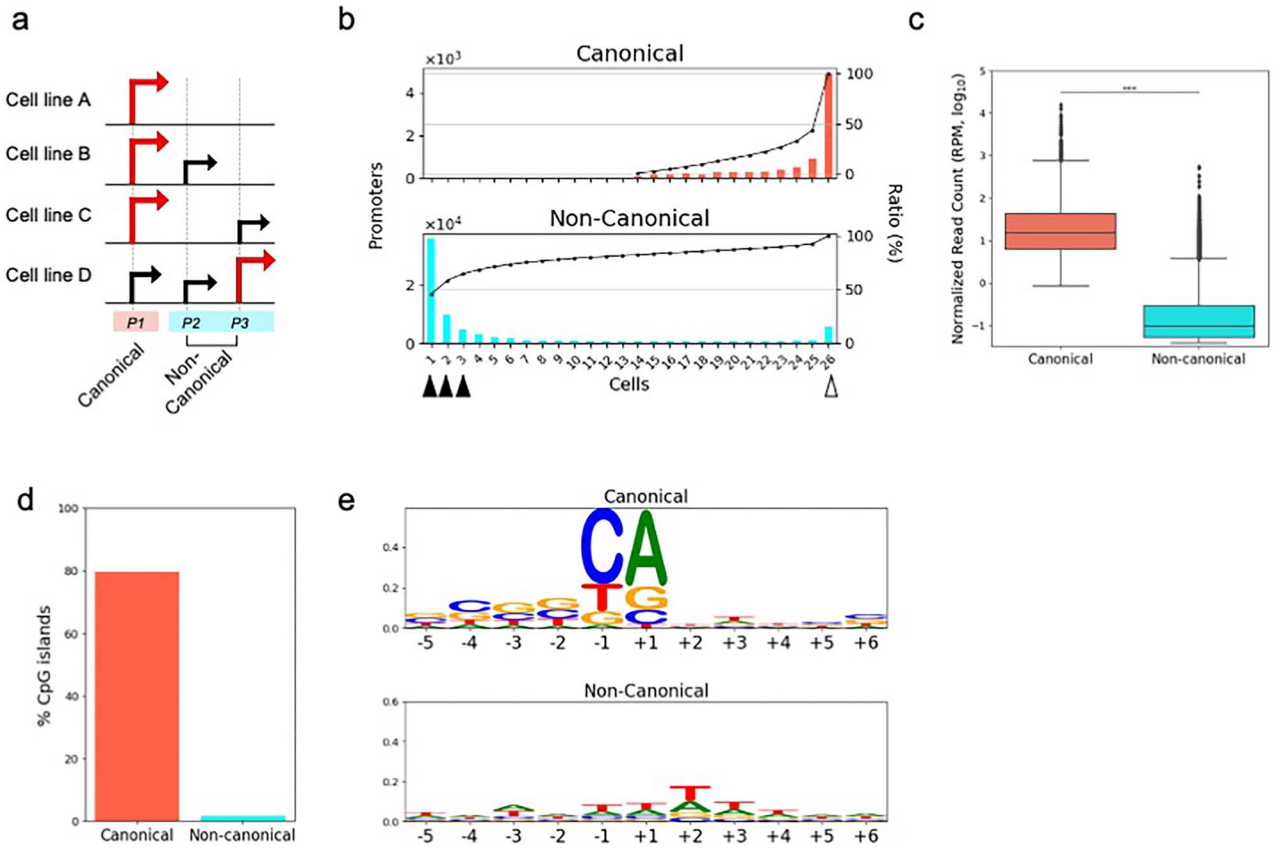

Figure 2. Canonical promoters (CPs) and non-canonical promoters (NCPs) in genomes of 26 lung adenocarcinoma (LUAD) cell lines. (a) Definition of CP and NCPs. CPs are the promoters with the highest expression (red arrows) for each gene in the majority of the cell lines. (b) Bar plots showing the number of CPs (top box) and NCPs (bottom box) stratified by the number of cell lines in which the promoters were expressed. The point plot in each box shows the cumulative number of CPs and NCPs. An open arrowhead at the bottom box means commonly expressed NCPs and closed arrowheads mean uniquely expressed NCPs. (c) Expression of CPs and NCPs shown as mean $\log_{10}$ Reads Per Million value. (d) Percentage of the CpG islands positive CPs and NCPs. (e) The Sequence logo in representative TSS ± 5 bp of CPs and NCPs.

non-canonical promoters were regulated by a canonical promoter-like manner.

Diversity of alternative promoter usage among the genes in 26 LUAD cell lines

A genome-wide evaluation of the usage of CPs and NCPs would allow us to analyse the biological significance of alternative promoters. We defined CPU as an index to evaluate the bias of CP in alternative promoters in gene expression. CPU is calculated by dividing 'read counts of the TSS-seq on a CP' by 'read counts of the TSS-seq on all the units of promoters (CP and NCPs)' on each gene. For example, the CPU value for the *STAU2* gene was 0.18 (29.0/161.0 RPM) in PC3 cells and 0.96 (26.3/27.5 RPM) in PC7 cells. CPUs were calculated for all genes in each cell line. Figure 4a illustrates the distribution of CPUs per gene in the PC14 and PC7 cell lines with mean CPU values of 0.81 and 0.91, respectively.

We performed a cluster analysis based on the CPU values. There were 14 297 genes for which promoters were identified in any of the cell lines. Among them, 5301 genes that were expressed in all 26 cell lines were used for analysis. Genes were divided into five clusters (Clusters 1–5, Supplementary Table S4), and cell lines were sorted by mean CPU value (Fig. 4b). Supplementary Figure S5 shows the cluster characteristics. Genes in clusters 1 and 2 had lower CPU, suggesting relatively high or frequent expression of NCPs. Conversely, cluster 3 had the highest CPU, suggesting a relatively high expression in CPs. Although the total number of promoters included in the genes tended to be higher in clusters 1 and 2 and lower in cluster 3, there was no significant difference in the length of the genes included in those clusters. No differences were observed in gene expression levels in clusters 1–4 evaluated by RNA-seq data.

Next, GO enrichment analysis was performed on the clusters (Fig. 4c and Supplementary Table S5). In clusters 1 and 2, terms related to signalling pathways, such as 'enzyme-linked receptor protein signalling pathway' and 'regulation of small GTPase mediated signal transduction', were found. On the other hand, cluster 3 had an enrichment of terms related to translation, such as 'ribosome assembly'.

Alternative promoter usage and characteristics of LUAD cell lines

The average CPU of non-cancerous small airway epithelial cell was relatively high, between the top 4th and 5th LUAD cell lines (Supplementary Fig. S6a). The relationship between the average CPU of each cell line and the growth speed and proliferation rate was examined, but no significant correlation was observed (Supplementary Fig. S6b). Next, an association between the average

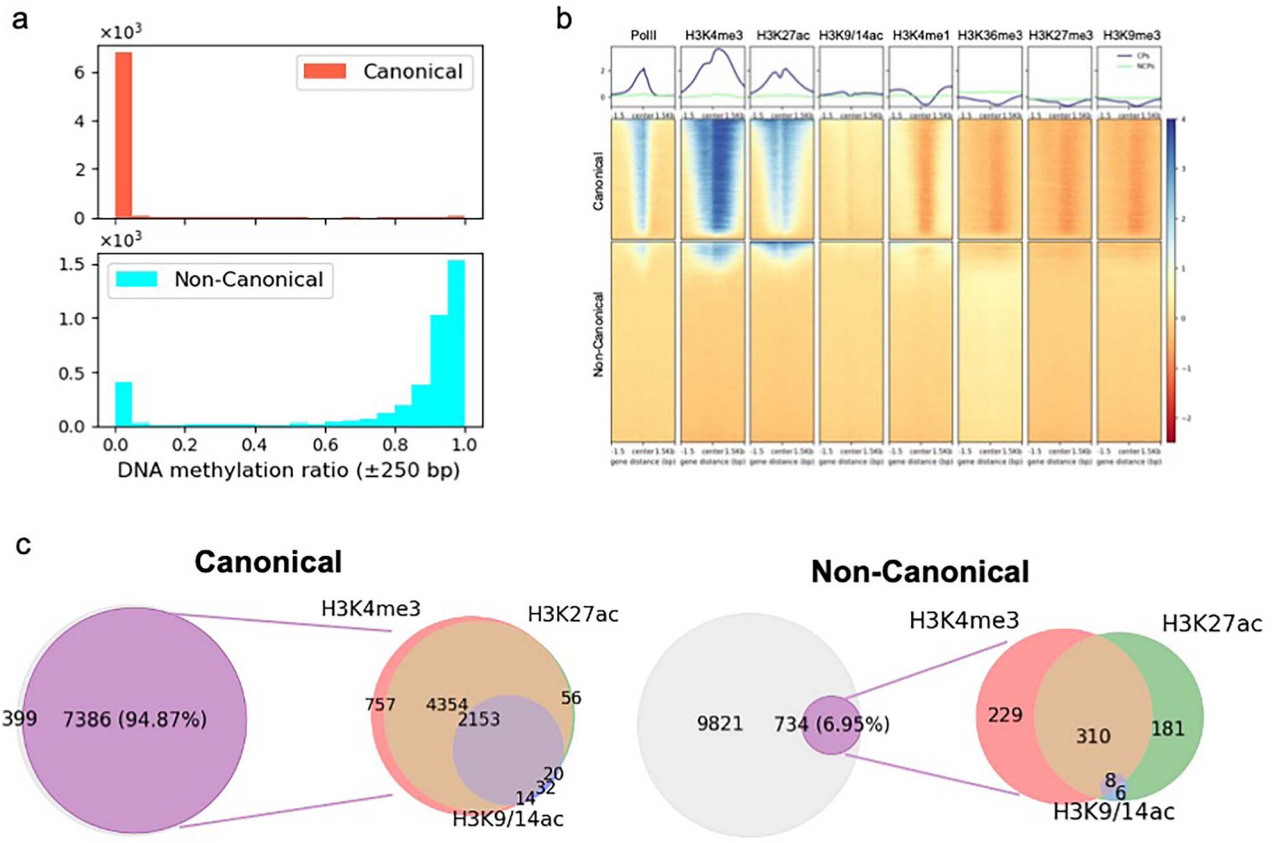

Figure 3. Epigenetic regulation of canonical promoters (CPs) and non-canonical promoters (NCPs). (a) Number of CPs (top box) and NCPs (bottom box) stratified by DNA methylation ratio. (b) Density heatmap showing the peaks of normalized read count of chromatin immunoprecipitation sequencing of PolII and histone markers on CPs and NCPs in PC14 cells. (c) Venn diagram showing the number of CPs and NCPs with histone modifications in PC14 cells. Right side purple circles represent promoters with any active histone marks. The circles on the left side represent the overlap of the active histone marks.

CPU and the presence of LUAD-associated driver mutations was examined (Supplementary Fig. S6c). For the overall distribution of all the associated genes, no clear correlation was observed. However, the cell lines with previously reported mitogenic driver mutations such as *EGFR*, *KRAS* and *NRAS* tended to have lower average CPUs.

Discussion

The diversity in the regulation of gene expression is considered an important event in oncogenesis. In addition to epigenetic regulation, the regulation of transcripts by non-coding RNAs or an increase in splice variants has also attracted attention from diagnostic and therapeutic perspectives. Dysregulation of alternative promoters diversifies translated protein isoforms and functional non-coding RNAs, leading to the disruption of cellular homeostasis. A previous study reported that the alternative promoter of *ERBB2* is associated with a worse prognosis and is related to patient survival in breast cancer (6). However, the genome-wide diversity of alternative promoter usage in cancer biology has not yet been well studied. In this study, we catalogued alternative promoter usage using multi-omics data from LUAD-derived cell lines that were considered histologically close.

We classified all the observed promoters in each gene into CPs, which are commonly and frequently used across cell lines and other NCPs. Analysis of CpG islands, core promoter sequences and histone modifications revealed conventional and efficient transcriptional regulation in CPs. In contrast, NCPs were likely involved in gene bodies and had fewer transcription initiation motifs and CpG islands. In addition, NCPs had fewer active histone marks, and the overlap among these marks was poor. These findings suggested that these regions were located within genes and would be targets of 'leaky' expression that were not strictly transcriptionally regulated. The relationship between changes in chromatin structure and alternative promoter usage needs to be investigated, such as overlaying with high-resolution ATAC-seq data and whether CPU is altered by inhibitors of chromatin-modifying enzymes.

Although leaky promoters were likely to increase with longer gene length and higher overall gene expression, there were no differences between clusters of genes with high or low CPU. Interestingly, GO enrichment analysis suggested that the high-CPU gene clusters were related to housekeeping genes, indicating that genes maintaining fundamental cellular functions would tend to use strictly regulated single promoters. These findings were consistent with a previous study that analysed human full-length cDNAs derived from oligo-cap cDNA libraries to identify differences in alternative promoter usage between housekeeping and tissue-specific genes (24). In contrast, the low-CPU gene cluster was enriched for genes related to signalling pathways. The possibility that fluctuations in alternative promoter usage may be associated with aberrant signalling in cancer cells is intriguing.

Following our definition, canonical promoters were not determined in a considerable number of genes. So, we redefined canonical-like promoters as the promoters with the highest expression for

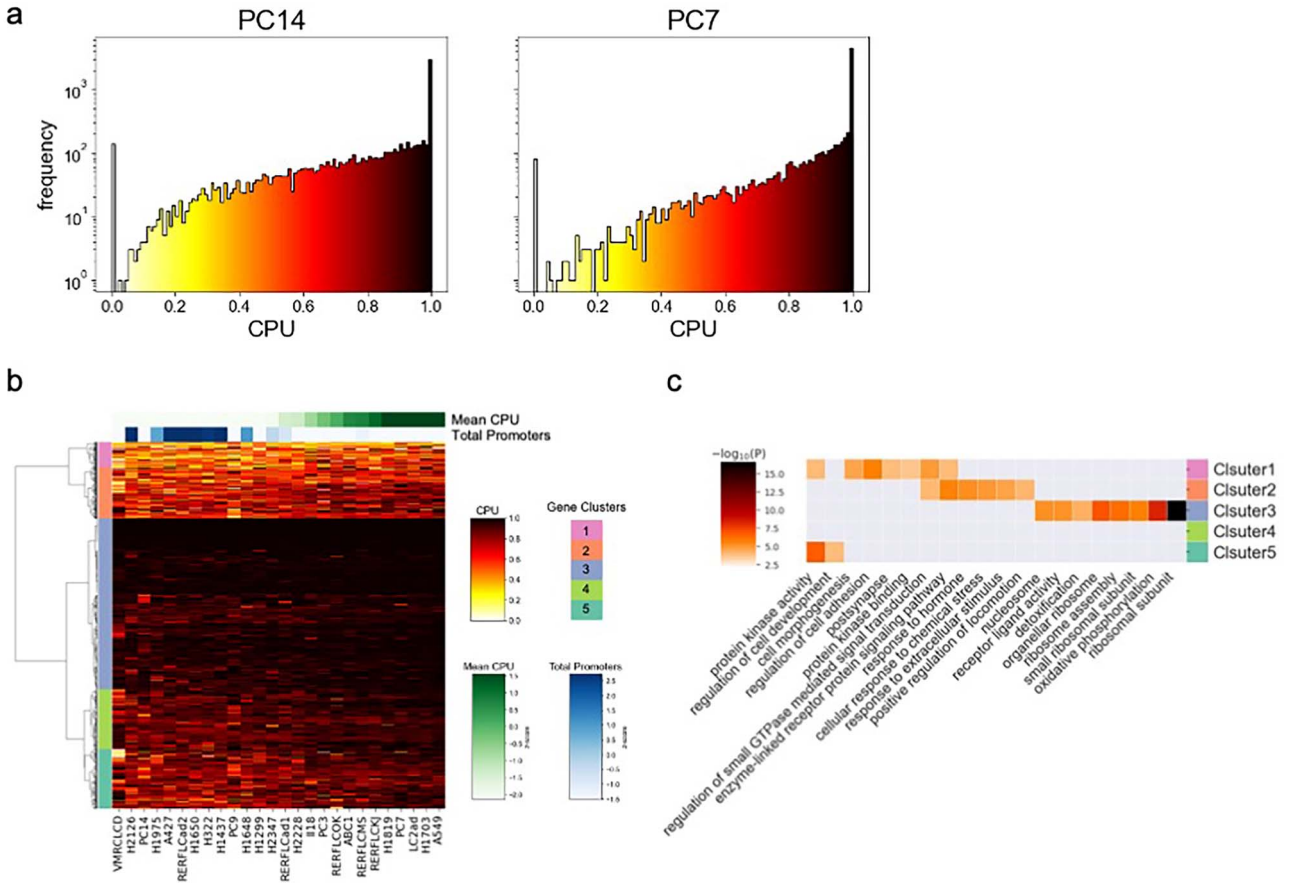

Figure 4. Alternative promoter usage indicated by canonical promoter usage (CPU). (a) Distribution of the CPU values in PC14 and PC7 cells. Mean CPU value of PC7 cells was higher than that of PC14 cells. The CPU = 0 peak was the genes which had the expressed NCPs and did not have the expressed CP. (b) Hierarchical clustering with the CPU values in 5301 genes among 26 lung adenocarcinoma cell lines. Genes were divided into five clusters. Cell lines were sorted by mean CPU value and the number of total promoters was represented in a heat map on the top of the diagram. (c) A heatmap showing the Gene Ontology enrichment of each gene clusters by CPU values.

each gene in more than half of the cell lines in which the gene was expressed. As well, the possibility that NCPs contain functionally important promoters is also considered and we selected NCPs with the highest expression and commonly expressed NCPs. The positivity of active histone marks for these redefined promoters was intermediate between CPs and NCPs. These findings suggest that some non-canonical promoters were regulated by a canonical promoter-like manner. More detailed classification is needed and will be challenged in future investigation.

Since the distribution of CPUs per gene differed among cell lines. The CPU value of non-cancerous lung epithelial cells was relatively high. In other words, a higher proportion of promoters under conventional regulation was expected in non-cancerous cells. Though, no clear correlation between the average CPU of each cell line and the growth speed and proliferation rate was observed, interestingly, *EGFR*, *KRAS* and *NRAS* mutations, which are considered strong mitogenic drivers, were more prevalent in the low CPU group. Though the number of cell lines used in this study is limited and analysis using a larger number of clinical specimens is needed, these results may suggest that regulation of alternative promoters has some link to the carcinogenic process of LUAD.

In this study, we classified alternative promoters in LUAD cell lines in a genome-wide manner and demonstrated their distinct transcriptional regulatory mechanisms. By designating a novel indicator, CPU, we suggested different profiles of alternative promoter usage in different cell lines. Further investigation is warranted to elucidate the detailed molecular mechanisms using this indicator to analyse the correlation between alternative promoter usage and biological characteristics of individual cancers. Thus, it is expected to be applicable to the analysis of the diversity of carcinogenesis and treatment response.

Supplementary Material

Supplementary material can be found at *Japanese Journal of Clinical Oncology* online.

Acknowledgements

The authors thank Shun-Ichiro Kageyama, Sung-Gi Chi, Junyan Du, Masato Aoshima and Shunsuke Sakai for their assistance. The super-computing resource was provided by Human Genome Center (the Univ. of Tokyo). We thank Editage (www.editage.com) for English language editing.

Funding

National Cancer Center Research and Development Fund [31-A-10], [2021-A-10] and JSPS KAKENHI (22H03084).

References

1. Macleod D, Ali RR, Bird A. An alternative promoter in the mouse major histocompatibility complex class II I-Aβ Gene: implications for the origin of CpG islands. *Mol Cell Biol* 1998;18:4433–43.
2. Ushijima T, Hanada K, Gotoh E, et al. Light controls protein localization through phytochrome-mediated alternative promoter selection. *Cell* 2017;171:1316–1325.e12.
3. Carninci P, Sandelin A, Lenhard B, et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* 2006;38:626–35.
4. Pozniak CD, Radinovic S, Yang A, et al. An anti-apoptotic role for the p53 family member, p73, during developmental neuron death. *Science* 2000;289:304–6.
5. Zaika AI, Slade N, Erster SH, et al. δNp73, a dominant-negative inhibitor of wild-type p53 and TAp73, is up-regulated in human tumors. *J Exp Med* 2002;196:765–80.
6. Demircioğlu D, Cukuroglu E, Kindermans M, et al. A pan-cancer transcriptome analysis reveals pervasive regulation through alternative promoters. *Cell* 2019;178:1465–1477.e17.
7. Forrest ARR, Kawaji H, Rehli M, et al. A promoter-level mammalian expression atlas. *Nature* 2014;507:462–70.
8. Bulger M, Groudine M. Functional and mechanistic diversity of distal transcription enhancers. *Cell* 2011;144:327–39.
9. Weber M, Hellmann I, Stadler MB, et al. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet* 2007;39:457–66.
10. Heintzman ND, Stuart RK, Hon G, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* 2007;39:311–8.
11. Suzuki A, Makinoshima H, Wakaguri H, et al. Aberrant transcriptional regulations in cancers: Genome, transcriptome and epigenome analysis of lung adenocarcinoma cell lines. *Nucleic Acids Res* 2014;42:13557–72.
12. Berg KCG, Eide PW, Eilertsen IA, et al. Multi-omics of 34 colorectal cancer cell lines - a resource for biomedical studies. *Mol Cancer* 2017;16:1–16.
13. Vasaikar SV, Straub P, Wang J, Zhang B. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res* 2018;46:D956–63.
14. Suzuki A, Onodera K, Matsui K, et al. Characterization of cancer omics and drug perturbations in panels of lung cancer cells. *Sci Rep* 2019;9:1–15.
15. Suzuki A, Wakaguri H, Yamashita R, et al. DBTSS as an integrative platform for transcriptome, epigenome and genome sequence variation data. *Nucleic Acids Res* 2015;43:D87–91.
16. Yamashita R, Sathira NP, Kanai A, et al. Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. *Genome Res* 2011;21:775–89.
17. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754–60.
18. Tareen A, Kinney JB. Logomaker: Beautiful sequence logos in Python. *Bioinformatics* 2020;36:2272–2274.
19. Takai D, Jones PA. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A* 2002;99:3740–5.
20. Hinrichs AS, Karolchik D, Baertsch R, et al. The UCSC genome browser database: update 2006. *Nucleic Acids Res* 2006;34:590–8.
21. Ramírez F, Ryan DP, Grüning B, et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* 2016;44:W160–5.
22. Zhou Y, Zhou B, Pache L, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* 2019;10:1523.
23. Shiraki T, Kondo S, Katayama S, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci USA* 2003;100:15776–81.
24. Kimura K, Wakamatsu A, Suzuki Y, et al. Diversification of transcriptional modulation: Large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res* 2006;16:55–65.