A Pan-Cancer Transcriptome Analysis Reveals Pervasive Regulation through Alternative Promoters

Deniz Demircioğlu[1,2], Engin Cukuroglu[1,] Martin Kindermans[1], Tannistha Nandi[1], Claudia Calabrese[3,4,20], Nuno A. Fonseca[3,5,20], André Kahles[6,7,8,9,10,20], Kjong-Van Lehmann[6,8,9,10,20], Oliver Stegle[3,4,11], Alvis Brazma[3,21], Angela N. Brooks[12,21], Gunnar Rätsch[6,7,8,9,10,13,21], Patrick Tan[14,15,16,17,18,19,22], Jonathan Göke[1,18,22,23]

[1]Computational and Systems Biology, Genome Institute of Singapore, Singapore 138672, Singapore

[2]School of Computing, National University of Singapore, Singapore 117417, Singapore

[3]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK

[4]Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, 69117, Germany

[5]CIBIO/InBIO - Research Center in Biodiversity and Genetic Resources, Universidade do Porto, Vairão 4485-601, Portugal

[6]Department of Computer Science, ETH Zurich, Zurich, 8092 Switzerland

[7]Department of Biology, ETH Zurich, Zurich 8093, Switzerland

[8]Computational Biology Center, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA

[9]SIB Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland

[10]Biomedical Informatics Research, University Hospital Zurich, Zurich 8091, Switzerland

[11]Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany

[12]Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA

[13]Weill Cornell Medical College, New York, NY 10065, USA

[14]Program in Cancer and Stem Cell Biology, Duke-NUS Medical School, Singapore 169857, Singapore

[15]Cancer Science Institute of Singapore, National University of Singapore, Singapore 117599, Singapore

[16]Cancer Therapeutics and Stratified Oncology, Genome Institute of Singapore, Singapore 138672, Singapore

[17]SingHealth/Duke-NUS Institute of Precision Medicine, National Heart Centre Singapore, Singapore 169856, Singapore

[18]Cellular and Molecular Research, National Cancer Centre, Singapore 169610, Singapore

[19]Singapore Gastric Cancer Consortium, Singapore 119074, Singapore

[20,21]Alphabetical order

[22]Senior authors

[23]Lead contact.

*Correspondence: gokej@gis.a-star.edu.sg (J.G.)

Summary

Most human protein-coding genes are regulated by multiple, distinct promoters, suggesting that the choice of promoter is as important as its level of transcriptional activity. However, while a global change in transcription is recognized as a defining feature of cancer, the contribution of alternative promoters still remains largely unexplored. Here we infer active promoters using RNA-Seq data from 18,468 cancer and normal samples, demonstrating that alternative promoters are a major contributor to context-specific regulation of transcription. We find that promoters are deregulated across tissues, cancer types, and patients, affecting known cancer-genes and novel candidates. For genes with independently regulated promoters we demonstrate that promoter activity provides a more accurate predictor of patient survival than gene expression. Together, our study suggests that a dynamic landscape of active promoters shapes the cancer transcriptome, opening new diagnostic avenues and opportunities to further explore the interplay of regulatory mechanisms with transcriptional aberrations in cancer.

Keywords: transcriptomics, RNA-Sequencing, transcriptional regulation, alternative promoters, cancer, genomics, computational biology

INTRODUCTION

The key element in regulation of transcription is the region upstream of the transcription start sites (TSS), the promoter. Promoters contain the elements required to initiate transcription, and they integrate signals from distal regulatory elements and epigenetic modifications that together determine the level of transcription. In the human genome, the majority of protein coding genes are regulated by multiple promoters that initiate transcription for different gene isoforms (Carninci et al., 2006; Sandelin et al., 2007). In contrast to alternative splicing which regulates gene isoform expression post-transcriptionally, the usage of alternative transcription start sites provides a way to regulate gene isoform expression pre-transcriptionally (Ayoubi and Van De Ven, 1996). Therefore, promoters not only determine when a gene is active and how active it is, they also regulate which gene isoforms will be expressed.

In cancer, somatic mutations, genomic re-arrangements, and changes in the regulatory or epigenetic landscape have been found to affect the promoter of several oncogenes, and it has been suggested that promoters contribute to the malignant transformation of the cells (Khurana et al., 2016; Sharma et al., 2010; Vogelstein et al., 2013). Genome-wide studies of promoters using the H3K4me3 histone modification, an epigenetic mark at active promoters, or CAGE tag sequencing of the 5’ end of transcripts have found that transcription start sites frequently are differentially used in cancer (Bernstein et al., 2002; Chi et al., 2010; Gherardi et al., 2012; Hashimoto et al., 2015; Kaczkowski et al., 2016; Kodzius et al., 2006; Muratani et al., 2014; Santos-Rosa et al., 2002; Takahashi et al., 2012). However, as data such as H3K4me3 profiles or CAGE tag is not available for most cancer studies, the role of alternative promoters in cancer remains largely unexplored.

Because any change in a cell's identity and function will be reflected in a change in gene expression, transcriptome profiling by RNA-Sequencing is one of the most widely studied large-scale molecular phenotypes in cancer. Analysis of gene expression in cancer has uncovered fundamental insights of tumor biology (Hoadley et al., 2018), enabled stratification of cancer types (Cancer Genome Atlas Research, 2012), predicted clinical outcome (Gerstung et al., 2015), and guided treatment decisions (Cancer Genome Atlas Research, 2011), forming a cornerstone of data driven precision oncology. RNA-Seq data measures the transcriptome largely unbiased, and as promoters regulate expression of isoforms with distinct 5’ start sites, it could potentially be used to identify active promoters (Feng et al., 2016; Pal et al., 2011; Reyes and Huber, 2018).

In this manuscript, we infer active promoters from RNA-Seq data, enabling the analysis of promoter activity in thousands of samples using publicly available expression data, thereby generating the largest currently available catalog of active promoters in human tissues and cancers. We apply this approach to comprehensively analyze alternative promoters in 18,468 samples covering 42 different cancer types, including the Pan-Cancer-Analysis of Whole Genomes (PCAWG) cohort of 1,188 patient samples with matched whole genome sequencing data, the PanCanAtlas cohort of 11,251 samples with exome sequencing data, and 6,674 normal samples from the GTEx project (Calabrese et al., 2018; Consortium et al., 2017; Hoadley et al., 2018). We find that alternative promoters are frequently used to increase isoform diversity and that a number of known cancer genes and novel candidates show deregulation of promoters in cancer. Our data suggests that the landscape of active promoter is highly dynamic and associated with patterns of somatic mutations, and that patient-to-patient variation in promoter activity is associated with survival. We propose that the precise knowledge of which promoter is active in each patient helps understanding the genetic, transcriptional, and pathological profile of individual tumors.

RESULTS

Identification of active promoters in 18,468 RNA-Seq samples

The promoter is defined as the regulatory region upstream of the transcription start site. Using Gencode (release 19) annotations, we compiled a set of 113,076 possible promoters, assuming that isoforms which have identical or very close TSSs are regulated by the same promoter (Table S1, 2) (Frith et al., 2008; Harrow et al., 2012). As the number of promoters is much smaller than the number of isoforms per gene, the problem of promoter activity estimation is heavily reduced in complexity, resulting in more robust inference (Figure S1A). To further reduce the number of false positives, we restricted our analysis to promoters that can be uniquely identified (Figure S1B, see Methods).We then defined promoter activity as the total amount of transcription initiated at each promoter. By quantifying the expression that is initiated at each promoter using the set of unique junction reads, we can then infer levels of promoter activity from RNA-Seq data (Figure 1A, see Methods). Following this approach, we quantified promoter activity in 18,468 samples from the PCAWG, TCGA, and GTEx cohorts, covering 42 cancer types (Table S3). Across all samples we identified the most active promoter (major promoter) for 17,182 (30%) genes, we identified 5,115 (9%) additional promoters that are active at lower levels (minor promoters), and we found 61% (35,127) of promoters to be inactive (Figures 1B and S1C). In the absence of regulatory genomics data, the first promoter of a gene is often assumed to be active. However, our data shows that the dominating major promoters can occur at any position within a gene. We find that 1 out of 3 major promoters are located downstream of the first TSS (Figure 1C), demonstrating how RNA-Seq data adds information and context to genome annotations.

To evaluate the accuracy of expression-based estimation of promoter activity, we compared the former to publicly available H3K4me3 ChIP-Seq and CAGE tag data from a variety of different cell lines and tissues (Consortium, 2012; Consortium et al., 2014; Davis et al., 2018; Lizio et al., 2015). We observed highest levels of H3K4me3 support for major promoters whereas inactive promoters show the lowest H3K4me3 levels (Figures 1D, S1D, and S1E; Kruskal-Wallis $p < 2.2e-16$), demonstrating that expression and epigenetic based estimates display a remarkable level of consistency. Our analysis of CAGE tag data confirmed these findings (Figures 1E and S1F), demonstrating that promoters which are uniquely identified have significantly higher CAGE tag support compared to non-unique (“internal”) promoters (Figures S1G, S1H and S1I). Furthermore, RNA-Seq based promoter activity estimates were most similar to ChIP-Seq profiles from matching cell lines for the majority of the tissues (44 out of 61; Figure 1F; Table S4). However, while promoter activity estimates from patients were generally highly consistent, cell lines showed a much higher variance (Figure S1J, Table S4). It has been observed before that cancer cell lines differ from the primary tissue (Consortium et al., 2014), suggesting that RNA based estimates from patient samples more accurately reflect the promoter landscape of the tumor than cell line based estimates.

To compare our framework with other methods, we estimated promoter activity using established RNA-Seq quantification methods (Salmon and Kallisto) and first exon read counts (see Methods) (Bray et al., 2016; Patro et al., 2017). Our promoter activity estimates were more similar to Salmon and Kallisto estimates compared to the first exon read count approach, with consistently high levels of correlation (Pearson’s correlation coefficient > 0.85; Figures S1K and S1L). However, we observed that our approach shows higher levels of agreement with ChIP-Seq data compared to all other methods ($p < 2.2e-16$), an observation that is possibly caused by over-estimation of inactive transcripts by transcript-based quantification methods (Figures S1M and S1N) (Soneson et al., 2019). Overall, this analysis demonstrates that our approach enables the quantitative, robust, and reproducible estimation of promoter activity from RNA-Seq data.

Alternative promoters are a major contributor to isoform diversity

Genome-wide, we find that promoter activity is dominated by the tissue and cell of origin for each cancer type (Figure 2A). This closely resembles the observation from gene expression, despite using only the minimal set of discriminative reads indicative of promoter activity (Figure S2A). In contrast to gene-level expression estimates, promoter activity enables us to investigate the contribution of each promoter to the overall expression pattern. Among all expressed protein-coding genes, 23% have at least 2 active promoters that contribute to more than 10% of the overall gene expression (Figures 2B and 2C). In principle, these promoters are independent regulatory units which can be used in a different context to control changes in isoform expression. The usage of such *alternative promoters* - promoters whose activity depends on the context but not on the activity of the gene's remaining promoters - will not be detectable with gene level based expressions analysis. Therefore, even though globally promoter activity reflects gene expression, there is additional information in promoter activity that cannot be detected at the gene expression level.

To approximate the prevalence of alternative promoters as context-specific regulators of transcription, we searched for promoters that show significantly changed activity across tissues at genes that do not show an overall change in expression (FDR adjusted p-values < 0.05; Figures 2D, S2B and S2C; Table S5; see Methods for details). Strikingly, our data demonstrates that even genes that do not show any tissue-specificity at the gene expression level can be under control of 2 independent, highly tissue-specific alternative promoters which regulate distinct gene isoforms (Figures 2E, 2F and 2G). The majority of tissue-specific alternative promoters activate single isoforms, providing a direct link between transcriptional regulation and isoform expression (Figure 2H). Alternative promoters often correspond to minor promoters that are expressed at lower levels compared to the constitutively active major promoter (Figure 2I). However, for 15% of genes we observe that the major promoter is switched (Figure 2I). Interestingly, on a global level, 58% of all isoform switching events involve a switch in promoters (Figures 2J and S2D), demonstrating that alternative promoters are a major contributor to tissue-specific transcriptional diversity.

To understand the consequence of alternative promoter usage on the gene product, we examined how the functional regions (5' UTR, CDS, 3'UTR) differ compared to the major promoter (Figure 2K). As expected, use of an alternative promoter is almost always associated with a change in the 5' UTR region (Figure S2E), with on average less than 20% of the 5'UTR sequence being shared between alternative promoters (Figures 2L and S2F). A change in promoters also dramatically effects the coding part of RNAs, often involving a change of almost 50% of the protein coding sequence (Figures 2L and S2G). We further find that almost 90% of alternative promoters encode for isoforms that potentially use a different 3' UTR sequence based on annotations (Figures S2E and S2H). This suggests that promoters not only regulate transcription initiation, but that they specifically regulate alternative isoforms that are marked by distinct sequences, possibly influencing post transcriptional regulation, translation, and protein structure in a context-specific manner.

Cancer-associated promoters regulate isoform switching of oncogenes and tumor suppressors

Many cancer-associated genes and pathways have been discovered by comparing the expression profile of cancer with the expression profile of normal tissues (Fay et al., 2003; Gross et al., 2015; Hippo et al., 2002; Rapin et al., 2014). The large number of context-specific alternative promoters found in this study suggests that promoters might be among the unknown driving forces behind the transcriptional changes in cancer. To investigate this hypothesis, we searched for promoters that show a change in activity in cancer compared to normal tissue using adjacent samples from the PCAWG and TCGA data sets and additionally 5,260 samples from GTEx (Figure 3A) (Consortium et al., 2017). For the majority of tumor types the most similar tissue is indeed the tumor tissue (Figures 3B and S3A). Interestingly, lung squamous cell carcinomas are most similar to normal skin tissue, reflecting the cell of origin for these tumors (Cancer Genome Atlas Research, 2014). Using these matched tissue groups, we then identified cancer-associated alternative promoters. For each tissue we find between 73 and 633 promoters that are significantly differentially regulated in cancer compared to normal (Figures 3C and S3B; Table S5; see Methods for details). An analysis using the subset of paired cancer and normal samples from the same individuals, and an analysis using a smaller subset of these data confirms our results, suggesting that alternative promoters are consistently found across patients (Figures S3D, S3F, S3G and S3H). The change in expression due to cancer-associated promoters is largely independent from the other promoters for each gene, confirming that alternative promoters indeed act as independent regulatory units which can specifically be deregulated in cancer (Figures 3D, S3C and S3E). Again, we find that the choice of promoters changes the 5’UTR, CDS, and 3’UTR sequences, indicating that transcriptional changes in cancer are translated into functional differences in the gene product (Figures S3I and S3J). Among the genes that show alternative promoter activation in cancer are known cancer biomarkers such as *SEPT9* (deVos et al., 2009) or *TNFRSF19* (*TROY*) (Paulino et al., 2010), the well described proto-oncogene *CTNNB1* (β-catenin) (Lazar et al., 2008), *BID,* a pro-apoptotic target gene of p53 (Lee et al., 2004), or *MLLT1*, which has been associated with childhood kidney cancer (Perlman et al., 2015) (Figure S3K), *CDK4* (Figures 3E and 3F) (Lapenna and Giordano, 2009) and *PRKACA* (Moody et al., 2015). To understand the underlying regulatory changes leading to the use of cancer associated promoters, we performed a de-novo motif analysis. We searched for enrichment of transcription factor motifs in alternative promoters compared to the set of active promoters (see Methods for details). Across all cancer types we find several enriched transcription factor motifs (Figures S3M, S3N, S3O, S3P and S3Q; Table S6, see Methods), suggesting that changes in the activity of promoters are partially driven through a change in upstream regulatory networks.

Interestingly, alternative promoters also differ between closely related tumor types from the same tissue. For the 3 different kidney tumor types we find a large number of genes that only show minor changes in overall gene expression, but where alternative promoter usage causes a significant tumor-type-specific change in isoform expression (Figures 3G, 3H and S3L; FDR adjusted p-value < 0.05; Table S5; see Methods for details). Similarly, we identify a number of genes that use distinct alternative promoters across the clinical subtypes of breast cancer (Figures 3I, 3J, 3K and 3L; Table S5; see Methods for details), confirming that alternative promoter activation is indeed associated with the molecular characteristic of tumors.

Pan-Cancer deregulation of alternative promoters

While some promoters were specifically deregulated in single tumor types, we hypothesized that other alternative promoters might be deregulated across multiple tumor types from different tissues compared to their matched normal counterpart. Indeed, overall, we find 184 such promoters, several of which belong to known oncogenes and tumor suppressors such as *TES* or *SPOP* (Figures 4A, 4B, 4C, S4A and S4B; Table S5; FDR adjusted p-values < 0.05, see Methods for details) (Futreal et al., 2004). While these genes have been implied in cancer, the usage of an alternative promoter in cancer has not been described (Barbieri et al., 2012; Tobias et al., 2001). Tissue-specific promoter switching is more frequent, yet these events further demonstrate the prevalence of alternative promoter regulation.

Patterns of noncoding promoter mutations in cancer

Accumulation of somatic mutations plays a central role in cancer not only by affecting protein coding genes, also by disrupting noncoding gene regulatory elements (Kandoth et al., 2013; Rheinbay et al., 2017b; Weinhold et al., 2014). The accumulation patterns of somatic mutations in cancer is known to be highly heterogeneous (Lawrence et al., 2013b; Maruvka et al., 2017). To better understand which properties of promoters are associated with accumulation of somatic (single nucleotide) mutations in cancer, we investigated the whole genome sequencing data for all patients with matched RNA-Seq data in the PCAWG cohort. We find that promoters of genes with a less complex promoter architecture show higher number of mutations (Figure 4D). These genes are more often non-coding (Figure 4E) and within regions associated with later replication timing (Figure 4F), confirming that distinct groups of promoters are exposed to different mutational patterns. Most of these mutational patterns are dominated by passenger mutations, and accordingly only a small set of driver promoter mutations has been identified in the PCAWG cohort (*TERT*, *PAX5*, *WDR74*, *HES1*, *IFI44L*, *RFTN1*, and *POLR3E*) (Rheinbay et al., 2017a). Possibly due to the limited number of samples for each cancer type, no significant associations between mutation burden and alternative promoter activity was found, however, it is expected that the number increases with higher sample number (Calabrese et al., 2018; Rheinbay et al., 2017a). As RNA-Seq data is among the most widely generated data, our approach will provide a powerful tool to better understand the relation between somatic mutations, promoter activity and regulatory drivers in cancer.

Alternative promoter usage is associated with patient survival

Gene expression varies from patient to patient, a property that has enabled the discovery of gene expression biomarkers to predict cancer patient survival (Director's Challenge Consortium for the Molecular Classification of Lung et al., 2008; Finak et al., 2008; Salazar et al., 2011). As our data suggests that alternative promoters are often independently regulated, we hypothesized that patient-to-patient variation in promoter activity might provide a more accurate predictor for genes that use multiple promoters. To test this hypothesis, we first identified candidate genes that show signs of promoter switching within a cancer type (2 distinct major promoters in at least 10% of samples). We then investigated the association of promoter activity with survival estimates using 9,459 TCGA samples with matched clinical data (Figure 5A) (Liu et al., 2018). Indeed, we find a number of genes that show a significant association with survival only for a specific promoter, but not for overall gene expression (Figure 5B). This particularly affects genes which use independently regulated alternative promoters that show a low correlation in promoter activity as gene expression is unable to capture such promoter switching (Figures 5B, 5C, S5A and S5B). Among the genes that are predictive of patient survival through alternative promoter usage are several unknown genes such as *EML2*, but also known cancer genes that have not been reported to rely on promoter switching such as *CDKN2A* in kidney cancer and *ERBB2* (also known as *HER2*) in lower grade glioma (Figures 5D, 5E, S5C and S5D; Table S7).

High gene expression levels of *ERBB2* have been associated with aggressive tumor types (Slamon et al., 1987) and there is a targeted therapy available that makes *ERBB2* a marker of precision oncology (Slamon et al., 2001). *ERBB2* uses 2 promoters in lower grade gliomas, both of which are independently regulated (Pearson correlation 0.053, Figure 5F). Using our catalog of promoter activity, we find that only the second promoter (P2) of *ERBB2* in lower grade glioma patients is predictive of poor outcome (p = 2.01e-19), whereas the major promoter (P1) shows no significant association with patient survival (p = 0.8520, Figures 5G, 5H, 5I and 5J). As we did not find any underlying regulatory promoter mutation, we searched for co-occurrence of exonic somatic mutations with high *ERBB2* promoter activity for 510 samples with matched exome sequencing data. We find that *ERBB2* P2 activation co-occurs significantly with *EGFR* and *PTEN* missense mutations, but is mutually exclusive with p53 and *IDH1* missense mutations (Figure 5K). In contrast, levels of *ERBB2* P1 are not associated with somatic mutations (Figures S5E, S5F and S5G), only the relative levels compared to *ERBB2* P2 is informative. In samples with *IDH1* mutations *ERBB2* P1 is more active compared to *ERBB2* P2, whereas *ERBB2* P2 acts as dominating, major promoter in samples without *IDH1* mutations, providing a striking example of an association between alternative promoter activation and somatic mutations in cancer (Figure 5L). *IDH1*, *EGFR*, and *PTEN* mutations are themselves associated with survival (Figures S5H and S5I). Remarkably, even in the absence of *IDH1*, *EGFR*, or *PTEN* mutations (n = 80 patients), *ERBB2* promoter usage predicts patient survival (Figures S5J and S5K), demonstrating that alternative promoter usage can potentially provide a highly robust, predictive biomarker.

Our findings indicate that survival is either associated with the underlying regulatory changes which could be used as a diagnostic marker (Hegi et al., 2005), or with the differential usage of gene isoforms determined through the choice of promoters which could be explored for novel therapeutic targets. As none of these novel promoter biomarkers has been explored in detail, there is still a huge potential to utilize them in diagnostic or therapeutic applications in cancer.

DISCUSSION

Promoters are the key elements that link gene regulation with expression. Studies using ChIP-Seq and CAGE tag data have demonstrated a role of alternative promoters in cancer (Kaczkowski et al., 2016; Muratani et al., 2014; Qamra et al., 2017), yet due to limitations in sample numbers for such technologies, the landscape of active promoters and their variation across cancers and patients has not been described. By analyzing 18,468 RNA-Seq samples, we provide the largest survey of promoter activity in human tissues and cancers, confirming known examples, and identifying many alternative promoters that have not been associated with cancer. The scale of these data allows us to describe for the first time patient-to-patient variation in promoter usage. Our analysis suggests that the choice of promoter is tightly regulated, has a significant influence on the cancer transcriptome, and indicates that promoters possibly contribute to the cellular transformation of cancer.

By using RNA-Seq data, our approach enables the analysis of promoter activity in the PCAWG, PanCanAtlas and GTEx cohorts without the need for additional experiments. Similar approaches have been applied to normal tissue expression (Reyes and Huber, 2018), and embryonic stem cells (Feng et al., 2016). Overall these estimates are highly accurate although we observe an increased uncertainty for some promoters due to use of short read sequencing data. In particular, we find that transcription start sites that lie within internal exons or that overlap with splice acceptor sites are difficult to accurately identify. Information from the 3’ end of transcripts can be used to predict their activity, however this approach heavily depends on accurate annotations and high quality isoform abundance estimates, and a high level of uncertainty remains (Teng et al., 2016). Both CAGE tag data and ChIP-Seq data suggest that these “internal” TSSs are less used compared to the remaining TSSs, therefore our analysis still captures an accurate and comprehensive view of the promoter landscape in cancer, enabling the analysis of patient-specific promoter activity on a much larger scale compared to other genomic assays.

It is known that alternative promoters contribute to isoform diversity (Consortium et al., 2014; Reyes and Huber, 2018), yet only few such events have been described in a disease context. In cancer, genes such as *MET* (Muratani et al., 2014) , *TP73* (Deyoung and Ellisen, 2007), or *ALK* have been reported to use alternative promoters (Wiesner et al., 2015). By analyzing the role of alternative promoters in this large scale cohort we demonstrate that many more cancer-associated genes use alternative promoters, and that their activity systematically alters the cancer transcriptome across all major cancer types. Our results suggest that transcriptional regulation, possibly involving sequence specific transcription factors and epigenetic modifiers, provides a robust way to pre-transcriptionally determine isoform expression in tumors. The choice of promoter often has an impact on the coding sequence, suggesting that a switch in promoters will alter protein isoforms or result in noncoding transcription. Interestingly, we also observe a change in the 3’ UTR sequence that contains regulatory elements such as miRNA binding sites (Lai, 2002), indicating a possible relation between pre- and post-transcriptional regulation. Alternative promoters often show lower levels of activity, and the functional consequence of such transcripts remains to be validated. However, we also find a number of promoter switching events that dramatically change the gene product. Such alternative promoters are frequently found in cancer, most of which are unknown, demonstrating that this aspect has a large potential to be further explored.

In summary, our study demonstrates the pervasive role of alternative promoters in context-specific isoform expression, regulation of isoform diversity, and highlights how patient-to-patient variation in promoter activation is linked to pathological properties of cancer. As RNA-Seq data is among the most widely generated data types, our approach has many applications beyond cancer. Here, we provide a comprehensive catalog of active promoters and their expression pattern across 42 cancer types and tissues that will be a highly useful resource to understanding the roles of gene regulation and noncoding mutations in cancer. Tissue and cancer-specific promoters could also become highly relevant as sensors and tumor-restricted activators for immunotherapy and the development of novel cancer drugs (Nissim et al., 2017), diagnostic approaches such as liquid biopsy (Ulz et al., 2016) and they will enable accurate designs of genome wide functional screens (Klann et al., 2017; Marx, 2017). As the vast majority of alternative promoters in cancer has not been described before, our study opens numerous possibilities to explore their contribution to tumor formation, diagnosis, or treatment.

ACKNOWLEDGEMENTS

This work is funded by the Agency for Science, Technology and Research (A∗STAR), Singapore and supported by the Singapore Ministry of Health's National Medical Research Council under its Individual Research Grant funding scheme (OFIRG16nov019). A portion of this research is supported by the National Research Foundation Singapore and the Singapore Ministry of Education under its Research Centres of Excellence initiative. P.T. is supported by National Medical Research Council grants TCR/009-NUHS/2013 and NMRC/STaR/0026/2015. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The GTEx data used for the analyses described in this manuscript were obtained from the dbGaP accession number phs000424.v6.p1 on 04/09/2015. The TCGA data used in this manuscript were obtained from the dbGap accession number phs000178.v10.p8. The results published here are in whole or part based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov. We acknowledge the ENCODE Consortium and the ENCODE production laboratories for generating the datasets used in this study.

AUTHOR CONTRIBUTIONS

Conceptualization: P.T. and J.G.

Methodology: D.D and J.G.

Software: D.D.

Formal Analysis: D.D. and J.G.

Investigation: D.D., M.K., T.N., P.T., J.G.

Data Curation: D.D., E.C., C.C., N.A.F., A.K., K.L., O.S., A.B., A.N.B., G.R., J.G.

Writing – Original Draft: D.D. and J.G.

Writing – Review & Editing: D.D., J.G., P.T.,

Supervision: P.T. and J.G.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Main Figure Titles and Legends

Figure 1: Promoter activity estimation using RNA-Seq data and comparison with ChIP-Seq and CAGE data.

(A) Schematic representation of promoter activity quantification using RNA-Seq data. Transcripts which are regulated by the same promoter are grouped and promoter activity is estimated using the set of unique junction reads spanning the first intron of each transcript.

(B) Categorization of annotated promoters based on the average promoter activity estimates across all samples (pan-cancer). Promoters are separated into 3 groups: major promoters, the most active promoter of the gene; minor promoters, other active promoters of the gene; inactive, promoters with estimated activity < 0.25.

(C) Major/minor promoter proportions across TSSs ranked by position (5' to 3'), based on multi-promoter genes with at least one active promoter.

(D) Mean H3K4me3 (ChIP-Seq) coverage across 59 ENCODE cell lines for the pan-cancer major (green), minor (orange), and inactive promoters (purple) within ± 2000 bps of the TSSs. Overlapping promoter regions (± 1500 bps) are excluded from this analysis.

(E) Percentage of samples with CAGE tag support (for FANTOM samples) for inactive (< 0.25), low, and high (≥ 1.5) activity promoters. Promoters with more than 0 CAGE tag reads within a sample are considered supported for the corresponding sample.

(F) Mean promoter activity and mean H3K4me3 read counts for the same promoters show high levels of correlation for the matching RNA-Seq and ChIP-Seq samples. The promoter activity and log H3K4me3 counts are averaged across replicates for each cell line (biosample term) within ENCODE where both data were available. The top 3 cell lines with the highest Spearman correlation are highlighted for each column.

See also Figure S1, Tables S1, S2, S3, S4.

Figure 2: Alternative promoters are a major contributor to isoform diversity.

(A) T-SNE plot using the top 1,500 promoters with the highest variance in promoter activity.

(B) Comparison of major promoter activity and gene expression (sum of all promoters). A single promoter often does not fully explain gene expression, minor promoters contain additional information.

(C) Most active promoters are observed at genes with multiple promoters.

(D) Heatmap showing the standardized (per tissue) mean relative promoter activity for tissue-specific alternative promoters of genes which do not change in overall expression.

(E) Shown is the mean read count at the *GJB1* gene locus for samples from the central nervous system (CNS) and from all other tissues. The light blue and red regions highlight the two alternative promoters.

(F) The second *GJB1* promoter (prmtr.50930) is more active in CNS compared to all other tissues, whereas the first promoter (prmtr.50929) is inactive in CNS samples. The significance level is reported for the tissue specific alternative promoter, prmtr.50390: **** p ≤ 0.0001.

(G) Comparison of gene expression levels for *GJB1*.

(H) Number of isoforms that can be transcribed from tissue-specific alternative promoters.

(I) Alternative promoters are most often minor promoters.

(J) Shown is the fraction of isoform switching events across tissues that can be attributed to a change in promoter.

(K) Schematic representation for the analysis of 5'UTR, CDS, and 3' UTR regions and their association with alternative promoter usage. Regions unique to the major and alternative promoter, and regions shared among them are quantified for each tissue-specific alternative promoter.

(L) Shown is the percentage of the 5'UTR, 3'UTR, and CDS sequence that is unique to the tissue specific promoter gene isoforms (green), unique to the major promoter gene isoforms (orange), and that is shared (gray).

See also Figure S2 and Table S5.

Figure 3: Identification of cancer-associated alternative promoters.

(A) Overview of cancer and normal data obtained by combining PCAWG, TCGA and GTEx samples.

(B) Cancer samples are most similar (highest average Pearson correlation of promoter activity) to the normal samples from the same tissue type. Tumor types with less than 15 normal and cancer samples are excluded from this analysis, normal samples are batch corrected to adjust for different data sources (PCAWG, TCGA and GTEx).

(C) Heatmap showing the promoter activity estimates for KIRC cancer and kidney normal samples, ranked by mean difference. The promoter activity estimates are capped ± 3 sd.

(D) Difference in cancer-associated promoter activities (upper panel) and gene expression excluding alternative promoters (lower panel) for KIRC cancer and kidney normal samples, ranking is similar to heatmap in (C).

(E) Shown is the mean read count at the *CDK4* gene locus for KIRC cancer and kidney normal samples (left panel). The light blue and red regions highlight the cancer-associated promoters. The first promoter (prmtr.29918) is inactive in cancer samples and active in normal samples whereas the second promoter (prmtr.29917) is the major promoter and displays high activity in both cancer and normal samples.

(F) The promoter activity for the first (red) and second (light blue) promoters of *CDK4* reflect the switch of promoters between cancer and normal samples. The significance level is shown for cancer associated alternative promoter, prmtr.29918: **** p ≤ 0.0001.

(G) UMAP plot using the top 1,500 promoters with the highest variance in promoter activity in kidney samples.

(H) Shown is the *JAZF1* locus (top) and mean relative promoter activity across different kidney cancer and normal samples (bottom). The 3’ most promoter of *PIK3R1* (prmtr.40310, light blue) displays KIRP tumor subtype specific activation.

(I) The 3’ most promoter (prmtr.2834) of the *STAU2* gene displays subtype specific activity for basal breast cancer samples. (Top) Enrichment of the distinct breast cancer subtypes in the samples with high activity of prmtr.2834. (Bottom) Promoter activity for the 3 active promoters of *STAU2*, sorted by prmtr.2834.

(J) Shown is the mean read count at the *STAU2* gene locus for samples from the basal subtype of breast cancer and from all other subtypes of breast cancer. The light red, light blue and blue rectangle regions highlight the top 3 most active promoters.

(K) The most 3’ *STAU2* promoter (prmtr.2834) is more active in the basal subtype compared to all other subtypes, whereas the other 2 active promoters (prmtr.2837 and prmtr.2839) show comparable or higher activity levels in all other subtypes. The significance level is shown for molecular subtype specific alternative promoter, prmtr.2834: **** $p \leq 0.0001$.

(l) Boxplot showing the gene expression levels for *STAU2*.

See also Figure S3 and Table S5, Table S6.

Figure 4: Pan-cancer-associated promoters and heterogeneity of promoter mutations.

(A) Heatmap showing the mean promoter activity for promoters whose activity significantly differ between cancer and normal samples across multiple cancer types (pan-cancer associated promoters). Promoters of known cancer-associated genes are highlighted (Futreal et al., 2004). Promoter activities are capped at ±3 sd.

(B, C) Relative activity profile of pan-cancer-associated promoters. Shown is the *TES* (B) and the *SPOP* (C) gene locus which show lower activity of an alternative promoter across multiple cancer types.

(D) Proportions of single and multi TSS genes with mutated promoters across different numbers of mutated samples. Single TSS genes are more frequently mutated than genes with multiple promoters (multiple TSS genes).

(E) lncRNAs are more frequently mutated at the promoters than protein coding genes.

(F) Boxplots comparing the replication timing across promoters with different number of mutated samples, higher numbers of mutated samples are associated with later replication timing.

See also Figure S4 and Table S5.

Figure 5: Alternative promoter usage predicts patient survival.

(A) Schematic overview: Firstly, genes which use 2 or more promoters in a cancer type were selected, and for each promoter the top 10% of the samples with the highest activity were selected. We then estimated significance of difference in survival for each promoter.

(B) Mean correlation (Pearson) of alternative promoter activity with the activity of all other promoters of a gene; green: gene expression and promoter activity is associated with survival; orange: only promoter activity is associated with survival; grey: neither gene expression or promoter activity are associated with survival, for different significance thresholds.

(C) Boxplots showing the correlation of the alternative promoter activity with the promoter activity of all other promoters of a gene. Promoters of genes with predictive promoters are less correlated than promoters of genes with predictive gene expression, indicating that gene expression is limited as a biomarker for genes with independently regulated promoters.

(D) Scatterplot of adjusted p-values for genes with alternative promoters, x-axis showing the p-value for the most predictive promoter, y-axis showing the p-value for the other promoters. Orange indicates genes with predictive alternative promoters.

(E) The *ERBB2* gene in Lower Grade Glioma uses 2 alternative promoters (P1 and P2).

(F) Promoter activity of *ERBB2* promoter 1 and promoter 2. Both promoters are independently regulated.

(G) Promoter activity for *ERRB2* P1 high and low activity groups. The significance levels are reported for both promoters, P1: **** $p \leq 0.0001$, P2: not significant $p > 0.05$.

(H) Promoter activity of *ERBB2* promoter 1 is not associated with patient survival.

(I) Promoter activity for *ERRB2* P2 high and low activity groups. The significance levels are reported for both promoters, P1: not significant $p > 0.05$, P2: **** $p \leq 0.0001$.

(J) Promoter activity of *ERBB2* promoter 2 is significantly associated with patient survival.

(K) Enrichment of somatic missense mutations in patients with *ERBB2* P2 activation.

(L) Gene expression of *IDH1* and *ERBB2* P2 promoter usage. *IDH1* mutations (orange) are mutually exclusive with *ERBB2* P2 activation as major promoter (>0.5, grey line).

See also Figure S5 and Table S7.

Supplemental Figure Titles and Legends

Figure S1: RNA-Seq data can be used to identify active promoters. Related to Figure 1, Tables S1, S2, S3, S4.

(A) Correlation of expression estimates across samples of the same tumor type for genes (dark blue), active promoters (blue), and isoforms of multi isoform genes (light blue). The correlations are shown for the tumor types with more than 100 samples in the PCAWG cohort. A higher correlation of promoter activity estimates suggest a higher level of robustness compared to isoform estimates.

(B) Correlation of activity for promoters that cannot be uniquely identified (light blue), and promoters that can be uniquely identified (blue) across the sample pairs of the same tumor type. Similar to (A), the same set of samples from PCAWG is used for this analysis. The promoter activity estimates is obtained by using the split reads ratios approach (see Methods).

(C) Number of major, minor and inactive promoters per tumor type (left) for all tumor types and per sample (right) for GTEx muscle samples. 100 samples are available for GTEx muscle tissue.

(D) Pan-cancer major promoters have a greater number of H3K3me3 (ChIP-Seq) reads, a sign of active transcription, overlapping with the promoter region (±2000 bps from TSS) compared to minor and inactive promoters for 59 ENCODE cell lines. Outliers are not shown.

(E) Mean H3K4me3 ChIP-Seq read coverage across 59 ENCODE cell lines for the pan-cancer major, minor, and inactive promoters at the 1st, 2nd and 3rd TSS positions (from left to right) respectively.

(F) High activity promoters have greater number of CAGE tag reads overlapping with the promoter region (±100 bps from TSS) compared to low activity and inactive promoters.

(G) Non-internal promoters (promoters that can be uniquely identified) have higher CAGE tag support for high ($p = 0.02$) and low activity ($p = 0.00013$) promoters whereas less support for inactive ($p < 2.2e-16$) promoters compared to internal promoters (promoters that cannot be uniquely identified) indicating non-internal promoters are more accurate. Similar to (F), promoter region is identified as ±100 bps from TSS.

(H - I) Percentage of samples with CAGE tag support (for FANTOM samples) for inactive, low, and high activity promoters for non-internal (H) and internal promoters (I). Promoters with more than 0 CAGE tag reads within a sample are considered supported for the corresponding sample.

(J) Correlation of promoter activity estimates and H3K4me3 ChIP-Seq signal for matching blood, blood vessel, brain, cervix, colorectal, heart, kidney liver, muscle, prostate and skin ENCODE cell lines and RNA-Seq samples. RNA-Seq samples show higher correlation with ChIP-Seq data from the matching tissue.

(K) Correlation matrix of mean promoter activity for BRCA samples using junction read counts, Kallisto, Kallisto with bias correction, Salmon, Salmon with bias correction and first exon read counts methods. Pearson correlation is reported.

(L) Median of per sample promoter activity correlation (Pearson) for each sample group across different promoter activity estimation methods.

(M) Mean H3K4me3 ChIP-Seq read coverage across 59 ENCODE cell lines for the pan-cancer major promoters identified using junction reads, Salmon bias corrected isoform estimates and first exon read counts.

(N) Major promoters identified using junction reads counts have significantly higher H3K4me3 read count support compared to major promoters identified by Salmon bias corrected isoform estimates and first read exon counts estimates.

Figure S2: Alternative promoters display context specific regulation independent from gene expression. Related to Figure 2 and Table S5.

(A) t-SNE plot using the top 1,500 genes with the highest variance in gene expression.

(B) Difference in alternative promoter activities (upper panel) and gene expression excluding alternative promoters (lower panel) across pan-cancer. Alternative promoters' contribution to tissue specificity is independent from gene expression.

(C) Number of alternative promoters for each tumor type.

(D) Proportion of isoform switching events that can be explained by alternative splicing and alternative promoters per tissue type.

(E) Shown is the percentage of times a change has occurred in the 5'UTR, 3'UTR, and CDS sequence that is unique to the tissue specific promoter (green), unique to the major promoter (orange), and that is shared (gray).

(F - H) Shown is the percentage of the 5'UTR (F), CDS (G), and 3' UTR (H) sequence that is unique to the tissue specific promoter gene isoforms (top panel), unique to the major promoter gene isoforms (middle panel), and that is shared (bottom panel).

Figure S3: Identification of cancer-associated alternative promoters for tumor types and subtypes. Related to Figure 3 and Tables S5, and S6.

(A) Cancer samples from PCAWG and TCGA match to normal samples from the same tissue (blue) regardless of data source (PCAWG, TCGA or GTEx). Squamous cell carcinoma of the lung (LUSC), is assigned to skin reflecting the origin of cancer cell not the tissue.

(B) Heatmap showing the promoter activity estimates for BRCA cancer and breast normal samples, ranked by mean difference. Promoter activity estimates are capped at ± 3 sd.

(C) Difference in cancer-associated promoter activities (upper panel) and gene expression excluding alternative promoters (lower panel) for BRCA cancer and breast normal samples, ranking is similar to heatmap in (B).

(D) Heatmap showing promoter activity for BRCA alternative promoters using only paired cancer and normal samples. The promoter are ordered according to the paired sample activity difference.

(E) Paired promoter activity differences in cancer and normal samples for the BRCA associated alternative promoters using paired BRCA samples only (top panel). The remaining gene expression difference after excluding the contribution of alternative promoters (bottom panel). The promoters identified to be up- and down- regulated in cancer using all BRCA samples are shown in red and blue respectively. Mean ± standard error is shown for each promoter across the 117 paired samples.

(F) The paired relative promoter activity (top) and remaining gene expression (bottom) for BRCA associated alternative promoter prmtr.21925 of DNMT3A cancer census gene.

(G) The enrichment of down/up regulated alternative promoters identified using paired samples only relative to the set of alternative promoters identified by using all samples. Significances are estimated using Fisher's test.

(H) Randomized alternative promoter analysis of varying subsamples of BRCA sample cohort re-identifies BRCA associated alternative promoters. (Top) Enrichment of cancer associated alternative promoters for 10 random subsamples of size 50 (25 for each condition). (Bottom) Enrichment of cancer associated alternative promoters for 10 random subsamples of size 100 (50 for each condition). Significances are estimated using Fisher's test.

(I, J) Shown is the percentage change in occurrence (I) and length (J) of the 5'UTR, 3'UTR, and CDS sequences that are unique to the cancer associated promoters (green), unique to the major promoters (orange), and that are shared (gray).

(K) Shown is the mean read count at the *MLLT1* gene locus for KICH tumor samples and kidney normal samples (bottom-left). The red regions highlight the cancer-associated promoter in KICH cancer samples. The cancer associated deactivation of prmtr.26482 can be seen in relative (bottom-middle) and absolute (bottom right) promoter activities across normal and cancer samples. The significance level is shown for the cancer associated promoter, prmtr.26482: **** p ≤ 0.0001.

(L) Shown is the relative promoter activity for the *JAZF1*. The 3' most promoter (prmtr.40310) is active in KIRP samples and inactive in all other kidney cancer types, displaying KIRP cancer type specific regulation. The significance levels (compared to normal) are shown for the cancer type specific alternative promoter, prmtr.40310: **** p ≤ 0.0001.

(M) Sample motifs identified by RSAT de-novo motif discovery webserver to be enriched for different cancer associated promoter sets.

(N - O) The transcription factor *KLF9* binds to a motif similar to the identified de-novo motif for cancer associated down-regulated alternative promoters for LUAD samples. (N) The JASPAR motif (MA1107.1) for *KLF9* transcription factor. (P) The expression of *KLF9* gene in LUAD cancer and lung normal samples.

(P - Q) The transcription factor *SMAD4* binds to a motif similar to the identified de-novo motif for cancer associated down-regulated alternative promoters for LUAD samples. (P) The JASPAR motif (MA1153.1) for *SMAD4* transcription factor. (Q) The expression of *SMAD4* gene in LUAD cancer and lung normal samples.

Figure S4: Overview of relative promoter activity for pan-cancer associated alternative promoters. Related to Figure 4 and Table S5.

(A) The relative promoter activities for the first (light blue, prmtr.29726) and second (red, prmtr.29727) promoters of *TES* show the deactivation of the second promoter in cancer samples compared to normal. The significance level is shown for the pan-cancer associated alternative promoter, prmtr.29727: **** $p \leq 0.0001$.

(B) The relative promoter activities for the first (light blue, prmtr.22517) and second (red, prmtr.22516) promoters of *SPOP* in cancer and normal samples show the deactivation of the second promoter in cancer samples compared to normal. The significance level is shown for the pan-cancer associated alternative promoter, prmtr.22516: **** $p \leq 0.0001$.

Figure S5: Alternative promoter usage predicts patient survival independently from somatic mutations. Related to Figure 5 and Table S7.

(A) Boxplots showing the correlation of alternative promoter activity with the activity of all other promoters of a gene for different sets of promoters and different significance thresholds.

(B) Mean correlation of alternative promoter activity with the activity of all other promoters of a gene; color represents different sets of promoters for different significance thresholds.

(C) Survival data for *CDKN2A* in KIRC samples. Among the 2 active promoters of *CDKN2A*, only the second promoter (prmtr.37826) is predictive of patient survival.

(D) Survival data for *CDKN2A* KICH samples. Among the 2 active promoters of *CDKN2A*, only the second promoter (prmtr.37826) is predictive of patient survival.

(E) Enrichment of somatic missense mutations in patients with *ERBB2* P1 activation. Highlighted in blue are genes where missense mutations are significantly associated with *ERBB2* P2 activation.

(F) *ERBB2* P2 activity and gene expression for *IDH1*, *EGFR*, *PTEN*, *TP53* (from left to right). Orange indicates a missense mutation in the respective gene.

(G) *ERBB2* P1 activity and gene expression for *IDH1*, *EGFR*, *PTEN*, *TP53* (from left to right). Orange indicates a missense mutation in the respective gene. *ERBB2* P1 activity is largely independent from mutation status of these genes.

(H) Survival for patients with missense mutations in *IDH1*, *EGFR*, *PTEN* (from left to right).

(I) Survival for patients with missense mutations in *IDH1*, *EGFR*, *PTEN* and *ERBB2* P2 activation (from left to right).

(J) Survival for patients with *ERBB2* P2 activation, only patients without mutations in *EGFR*, *PTEN*, and IDH1 are shown. *ERBB2* P2 activation predicts patient survival in the absence of mutations.

(K) Survival for patients with *IDH1*, *EGFR*, *PTEN* mutations and *ERBB2* P2 activation, shown are all combinations.

STAR METHODS

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Jonathan Göke (gokej@gis.a-star.edu.sg). The resources generated in this study are provided as supplemental tables and listed in the Key Resources Table.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Tumor and normal RNA-Seq samples were obtained from GTEx, TCGA and PCAWG projects. The GTEx (phs000424.v6.p1) data contains 6,674 normal RNA-Seq samples and their sample metadata. The TCGA (phs000178.v10.p8) transcriptomics data consists of 11,251 cancer and normal RNA-Seq samples and sample metadata. Additionally, we downloaded the clinical data and the SNV calls for donors when available for the mutation and survival analysis respectively. The PCAWG data includes 543 cancer and normal RNA-Seq samples in addition to the samples within the TCGA dataset. Furthermore, we obtained the SNV calls for 1,188 patients in the PCAWG cohort with whole genome sequencing data for the noncoding mutation analysis. For the evaluation of promoter activity estimates, we retrieved data from the ENCODE and FANTOM5 projects. From ENCODE, we obtained 1,158 (361 ChIP-Seq and 797 RNA-Seq) data sets for the analysis of matched RNA-Seq and ChIP-Seq data. Additionally, we downloaded ChIP-Seq data for 59 cell lines with H3K4me3 data available. We used CAGE tag read counts for 1,829 samples from primary human cell types and tissues obtained from the FANTOM5 project. The sample identifiers are listed in their corresponding supplementary tables, and the respective data repositories are listed in the Key Resources Table.

METHOD DETAILS

Identification of promoters from annotations

Here, we used Gencode annotations (release 19) to identify TSSs (defined as the start of the first exon) for all annotated transcripts. Since transcripts with identical or very close TSSs are regulated by the same promoter, we used the TSSs from overlapping first exons to determine the set of transcripts regulated by each promoter. This provided us with a mapping of promoter to transcripts and promoters to genes (Table S1). We used this mapping information in the downstream analysis to quantify promoter activities per gene per sample (see below). Since a single promoter can be composed of multiple TSSs, we chose the 5' most TSS as the TSS for each promoter (Table S2).

Promoter activity estimation

Here, we used the concept of weighted splicing graphs to estimate promoter activity. A splicing graph is a directed acyclic graph that captures all the splice variants of a single gene in one data structure (Heber et al., 2002).

Let G, P and T denote the set of all genes, promoters and transcripts and S be the set of all samples. Then $G_g = (V_g, E_g)$denotes the weighted splicing graph with edges E_g and vertices V_g for gene$g \in G$ with promoters$P_g$and transcriptsT_g. The nodes of the splicing graph V_g represent the set of splicing sites for all the transcripts for gene g. The edges E_g denote the set of introns and exons connecting these splice sites. A transcript t is said to support an edge e if the intron or exon region identified by edge e is part of the transcript t.

Each edge e has the following properties: $type(e)_t$ denotes whether edge e is an intron or an exon for transcript t; $rank(e)_t$ is the intron or exon rank of the edge e for transcript t depending on $type(e)_t$; and $weight(e)_s$ is the count of reads uniquely mapping to the edge e in sample s.

Uniquely identifiable promoters

We calculated promoter activity estimates only for uniquely identified promoters. A uniquely identifiable promoter can be defined as follows. Let E_p be the set of first intron edges for promoter p with transcripts T_p.

$$E_p = ¿ t \in T_p \{ e : type(e)_t = intron \wedge rank(e)_t = 1 \}$$

Then then the set of uniquely identifiable promoters P' is defined as

$$P' = \{ p \in P_g : \max_{e \in E_p, t \in T_e} rank(e)_t = 1 \}$$

whereT_e is the set of transcripts supporting edge e. Promoters that can't be identified uniquely are either single exon promoters ($E_p = \emptyset$) or promoters that uses an internal intron ($rank(e)_t > 1$, "internal promoters", see below for details).

Junction read counts method

In this study, we quantify both absolute and relative promoter activities for uniquely identifiable promoters in each sample using the junction read counts method. For this method, the absolute promoter activity $A_{p,s}$ of promoter $p \in P'$ in sample s is proportional to the log_2 of total count of the junction reads aligning to the set of first introns belonging to the transcripts T_p of promoter p. Hence absolute promoter activity $A_{p,s}$ is

$$A_{p,s} = \log_2 \frac{\sum_{e \in E_p} weight(e)_s}{n_s}$$

where n_s is the sample specific normalization factor for sample s. Here, we normalized the total junction read counts using DESeq2 (v1.20.0) size factors to obtain normalized counts for the combined data set (PCAWG, GTEx and TCGA together) (Love et al., 2014). We then used the log2 transformed normalized read counts as promoter activity in the downstream analyses.

The gene expression $Z_{g,s}$ of gene g in sample s is calculated as the total absolute promoter activity for all the promoters in P', hence $Z_{g,s} = \sum_{p \in P'} A_{p,s}$. We normalized each promoter's activity by the gene expression to obtain relative promoter activities, $R_{p,s}$:

$$R_{p,s} = \frac{A_{p,s}}{\sum_{p \in P'} A_{p,s}} = \frac{A_{p,s}}{Z_{g,s}}$$

Split read ratios method

An internal promoter is identified as a promoter p which has an internal intron, i.e. $\exists e \in E_p$, such that $rank(e)_t > 1$ for some transcript $t \in T_g$. We excluded internal promoters from our analysis with the junction read counts methods since the junction reads mapping to an internal intron cannot be unambiguously assigned to the promoters. However, to be able to quantify and compare internal promoters with uniquely identifiable promoters, we developed a “split read ratio” method that accounts for this ambiguity. We quantified promoter activity for these internal promoters by normalizing the read count for splice donor sites by the read count for splice acceptor sites for the first exons of the transcripts belonging to a promoter. The split read ratio method can be described as follows:

Let the set of splice acceptor sites V_a for all the first exons of promoter p be

$$V_a = \text{¿}\, t \in T_p \{ v_k : \exists e(v_k, v_l) \wedge type(e)_t = exon \wedge rank(e)_t = 1 \}$$

Then the set of intron edges E_a with a splice acceptor site at the first exons of promoter p can be defined as

$$E_a = \text{¿}\, t \in T_g \{ e(v_i, v_j) : type(e)_t = intron \wedge v_j \in V_a \}$$

Similarly, we can define the set of splice donor sites V_d for all the first exons of promoter p and the set of intron edges $E_d$with a splice donor site at the first exons of promoter p as follows:

$$V_d = \text{¿}\, t \in T_p \{ v_l : \exists e(v_k, v_l) \wedge type(e)_t = exon \wedge rank(e)_t = 1 \}$$

$$E_d = \text{¿}\, t \in T_g \{ e(v_i, v_j) : type(e)_t = intron \wedge v_i \in V_d \}$$

Using these definitions, the absolute promoter activity $A^{SR}_{p,s}$ for promoter p in sample s is

$$A^{SR}_{p,s} = \frac{\log_2 \frac{\sum_{e \in E_d} weight(e)_s + 1}{\sum_{e \in E_a} weight(e)_s + 1}}{n_s}$$

where n_s is the sample specific normalization factor. Here, each sample is normalized by the average promoter activity across all promoters in the sample, hence $n_s = \frac{\sum_{p \in P} A^{SR}_{p,s}}{|P|}$ where $|P|$ is the number of promoters. The gene expression and relative promoter activity estimates are calculated similarly to the junction read counts method. We used the internal promoter activity estimates only for the robustness analysis and excluded it from any further downstream analysis.

Definition of Major, Minor, Inactive Promoters

We divided the promoter set into 3 different categories depending on their absolute promoter activity, namely, major, minor and inactive promoters. We mark the promoters with the highest average activity for each gene across the sample cohort as major promoters. Promoters with average activities smaller or less than 0.25 constitute inactive promoters whereas the other promoters of the gene constitute minor promoters.

Alternate Promoter Activity Estimation Methods

To demonstrate that promoter activity estimates obtained by junction read counts approach are reproducible, we used alternative approaches of promoter activity estimation for comparison. Since we define the promoter activity as the total transcription initiated at each promoter, we used transcript expression based methods and first exon read counts to estimate promoter activity.

Promoter quantification using transcript expression

We used Salmon (v0.10.0) and Kallisto (v0.44.0), with and without bias correction enabled, to estimate the isoform expression of the entire data cohort (18,468 samples in total). Transcript read counts were normalized to fragment per kilobase of million mapped with upper quartile normalization (FPKM-UQ) where total read counts in the FPKM definition has been replaced by the upper quartile of the read count distribution multiplied by total number of protein-coding

transcripts (Calabrese et al., 2018). We then summed the expression of transcripts belonging to a single promoter to obtain promoter activity estimates. We also calculated the mean promoter activity per tumor type and for the pan-cancer cohort. We identified the major, minor and inactive promoters. Here, we consider promoters with activity ≤ 0.5 FPKM-UQ as inactive.

Promoter quantification using first exon read counts

To calculate the promoter activity using first exon reads we used featureCounts (Rsubread v1.30.9) to count the total number of reads overlapping with the combined first exon ranges belonging to transcripts regulated by the same promoter (Liao et al., 2019). Similar to isoform based methods, we normalized exon read counts by using FPKM-UQ normalization where we used the total length of the combined first exons belonging to a promoter for length normalization. We used the first exon read counts approach only for the two largest tissue cohorts namely breast (BRCA: 1227 samples, GTEx Breast: 218 samples) and kidney (KICH: 91, KIRC: 610, KIRP: 323, GTEx Kidney: 36 samples) tissues. Promoters with activity ≤ 0.25 FPKM-UQ are labeled as inactive promoters.

RNA-Seq Data Sources and Alignment

We downloaded GTEx (phs000424.v6.p1) and TCGA (phs000178.v10.p8) raw data in fastq format from dbGap. We aligned reads to the human reference genome (GRCh37.p13) using TopHat2 (v2.0.12) using the “--b2-very-sensitive” option (Kim et al., 2013). Gencode (release 19) annotations were used as the reference annotation. For the ICGC/PCAWG cohort, we downloaded aligned bam files and STAR junction files processed according to the PCAWG alignment pipeline (Calabrese et al., 2018; Dobin et al., 2013). TopHat2 and STAR junction files are used to estimate absolute and relative promoter activity for each sample.

Identification of tissue-specific alternative promoters

To examine tissue specific alternative promoters for each cancer type, we used all the tumor samples within TCGA cohort only. We identified the up-regulated tissue specific alternative promoters using the following linear model for each tissue i:

$$A_{p,i}=\beta_{0,i,p}+\beta_{1,i,p}x_i+\varepsilon_{p,i}$$

$$R_{p,i}=\beta_{0,i,p}^R+\beta_{1,i,p}^R x_i+\varepsilon_{p,i}^R$$

where $x_i=x_{i,1},x_{i,2},\ldots,x_{i,|S|}$ indicates whether sample s is a cancer sample of tissue i. The nominal p-value for each promoter p for tissue i is calculated using the t-statistics of the β_1 and β_1^R coefficients of the linear regression for both the absolute and relative promoter activity, respectively. These nominal p-values are subsequently corrected for multiple testing across all promoters by using the Benjamini-Hochberg (BH) method. A promoter p is required to have BH adjusted p-value ≤ 0.05 for relative promoter activity estimate to be considered as a candidate tissue specific alternative promoter. We only consider the top 5,000 promoters with the lowest p-values for relative promoter activity as candidate tissue specific promoters.

Let S_i denote the set of samples of tissue i and $S_i^{'}$ be the rest of the samples. Then, $\overline{A}_{p,S_i}$ represent the mean absolute promoter activity of promoter p for samples in the tissue i and $\overline{A}_{p,S_i^{'}}$ represent the mean absolute promoter activity of promoter p for the samples from all other tissues. Similarly, $\overline{R}_{p,S_i}$, $\overline{R}_{p,S_i^{'}}$ and $\overline{Z}_{p,S_i}$, $\overline{Z}_{p,S_i^{'}}$ denote the mean relative promoter activities and mean gene expressions for same sample sets respectively. To identify tissue-specific alternative promoters, we filtered out inactive genes by enforcing $\overline{Z}_{p,S_i}\geq 1$ and $\overline{Z}_{p,S_i^{'}}\geq 1$. Additionally, we required the absolute and relative promoter activity to be above a certain threshold in both conditions, namely $\overline{A}_{p,S_i}\geq 0.25$ and $\overline{A}_{p,S_i^{'}}\geq 0.25$, $\overline{R}_{p,S_i}\geq 0.25$ and $\overline{R}_{p,S_i^{'}}\geq 0.25$. To remove candidates with inconsistent relative promoter activity, we imposed $\sum_{p\in P^{'}}\overline{R}_{p,S_i}\geq 0.9$ and $\sum_{p\in P^{'}}\overline{R}_{p,S_i^{'}}\geq 0.45$. Finally, to identify differential promoter activity without differential gene expression, we required tissue specific alternative promoters to have at least 2 fold change in mean absolute promoter activity and less than 1.5 fold change in mean gene expression across different conditions. We note that this is a conservative threshold, as many alternative promoters will change the gene expression as well.

Identification of cancer-associated promoters

In order to identify cancer associated promoters, we compared normal samples from the GTEx project in addition to the normal samples from PCAWG and TCGA with cancer samples from PCAWG and TCGA. We matched tissue between GTEx and TCGA/PCAWG by clustering of samples based on their mean promoter activity profile. We first removed the batch effect that might originate from using 3 different data sets by using the “removeBatchEffect” function from the ‘limma’ (v3.36.5) R package (Table S3) (Ritchie et al., 2015). We clustered the combined normal samples by hierarchical clustering where the distance of two tissues is defined as $d(i,j)=1-correlation(\acute{A}_i,\acute{A}_j)$. For downstream analysis, we used tumor types with at least 15 normal and 15 tumor samples.

Cancer associated alternative promoters

We identified cancer associated alternative promoters for each cancer type i using the following linear model:

$$A_{p,i}=\beta_{0,i,p}+\beta_{1,i,p}c_i+(\sum_{n=1}^{q-1}\beta_{n+1,i,p}d_{n,i})+\varepsilon_{p,i}$$

$$R_{p,i}=\beta^R_{0,i,p}+\beta^R_{1,i,p}c_i+(\sum_{n=1}^{q-1}\beta^R_{n+1,i,p}d_{n,i})+\varepsilon^R_{p,i}$$

where $c_i=c_{i,1},c_{i,2},\ldots,c_{i,\vee S_i\vee\text{¿¿}}$ indicates whether sample s is a cancer sample or not for samples S_i of cancer type i. Similarly, $d_{n,i}=d_{n,i,1},d_{n,i,2},\ldots,d_{n,i,\vee S_i\vee\text{¿¿}}$ indicates whether sample s is coming from study n where GTEx is used as reference group for study (here $q=3$). By including the study as a factor, we are adjusting for the possible batch effects that might confound the results. Similar to tissue specific alternative promoters, p-values for each promoter p are calculated using the t-statistics of the β_1 and β_1^R coefficients of the linear regression for both the absolute and relative promoter activity, respectively and corrected for multiple testing using the BH method. We used the same significance and expression thresholds as in the tissue-specific alternative promoter analysis, except we required $\sum_{p\in P'}\overline{R}_{p,S_i}\geq 0.9$ and $\sum_{p\in P'}\overline{R}_{p,S'_i}\geq 0.9$ for both cancer and normal to identify promoters with consistent relative promoter activity profiles across both conditions.

Multi-type associated alternative promoters

To identify multi-type associated alternative promoters for individual tumor types of a single tissue i, we estimated the following linear model:

$$A_{p,i}=\beta_{0,i,p}+(\sum_{k=1}^{j}\beta_{k,i,p}m_{k,i})+(\sum_{n=1}^{q-1}\beta_{n+1,i,p}d_{n,i})+\varepsilon_{p,i}$$

$$R_{p,i}=\beta^R_{0,i,p}+(\sum_{k=1}^{j}\beta^R_{k,i,p}m_{k,i})+(\sum_{n=1}^{q-1}\beta^R_{n+1,i,p}d_{n,i})+\varepsilon^R_{p,i}$$

where $m_{k,i}=m_{k,i,1},m_{k,i,2},\ldots,m_{k,i,\vee S_i\vee\text{¿¿}}$ indicates whether the sample s is of tumor type k, and normal sample type was used as the reference group for a tissue with j tumor types. To adjust for possible batch effects that might originate from using data from multiple studies, study is included as a factor ($d_{n,i}$) similar to cancer associated promoter analysis. For each tumor type k, p-values are calculated using the t-statistics of β_k and β_k^R of linear regression and multiple test corrected using BH method. For candidate multi-type associated alternative promoters, we required BH adjusted p-values $\leq$1e-5 for absolute and relative promoter activities. Additionally, we require $\overline{A}_{p,S_{i,k}}\geq 0.25$ and $\overline{A}_{p,S'_{ik}}\geq 0.25$ where $S_{i,k}$ is the set of tumor samples for cancer type k of tissue i and $S'_{i,k}$ is the set of normal samples for tissue i. Finally we filtered candidate alternative promoters with ¿ 2 fold change in mean absolute promoter activity and ¿ 2 fold change in mean gene expression across different conditions.

Subtype specific alternative promoters

Breast cancer molecular subtype specific alternative promoters are identified using the same approach as the tissue specific alternative promoters. The only difference in the linear model used is the indicative variable x_i, which is now defined as $x_i = x_{i,1}, x_{i,2}, \ldots, x_{i,|S_{BRCA}|}$ where $x_{i,s}$ indicates whether the cancer sample s is of molecular subtype i or not. S_{BRCA} contains 1068 BRCA tumor samples where the molecular subtype information is available. The same set of statistical significance and expression change criteria has been used in subtype analysis as with the tissue specific alternative promoter analysis.

Pan-cancer associated alternative promoters

To identify pan-cancer associated alternative promoters for all samples, we applied linear regression with adjustment for tissue type i and study n:

$$A_p = \beta_{0,p} + \beta_{1,p} c + \left(\sum_{i=1}^{l-1} \beta_{i+1,p} y_i\right) + \left(\sum_{n=1}^{q-1} \beta_{n+1,p} d_n\right) + \varepsilon_p$$

$$R_p = \beta_{0,p}^R + \beta_{1,p}^R c + \left(\sum_{i=1}^{l-1} \beta_{i+1,p}^R y_i\right) + \left(\sum_{n=1}^{q-1} \beta_{n+1,p}^R d_n\right) + \varepsilon_p^R$$

where $c = c_1, c_2, \ldots, c_{¿S \vee ¿¿}$ indicates whether samples s is a cancer sample or not. To adjust for tissue type i, we use $y_i = y_{i,1}, y_{i,2}, \ldots, y_{i,|S|}$ which indicates whether the sample s is from tissue type i (adrenal gland tissue type was used as a reference group for tissue type among $l-1$ different tissue types in the pan-cancer cohort). Also d_n is an indicator variable for study that adjusts for potential batch effects similar to methods above. P-values for each promoter p is calculated using the t-statistics of the β_1 and β_1^R coefficients of linear regression and corrected using the BH multiple test correction. We used the same significance and expression thresholds as cancer associated alternative promoter analysis with the following exceptions. Instead of using the pan-cancer mean cancer and normal activity for expression filters, we used the mean of mean per tumor and normal promoter activity, i.e. mean of $\overline{A}_{p,S_h}$ and $\overline{A}_{p,S_h'}$ for all cancer types h. Finally, we required $\sum_{p \in P'} \overline{R}_{p,S_C} \geq 0.1$ and $\sum_{p \in P'} \overline{R}_{p,S_C'} \geq 0.1$ for both cancer and normal sets respectively. The threshold is lowered to accommodate the variance in absolute promoter activities across multiple tissue and tumor types.

ChIP-Seq analysis

To evaluate the junction read count approach for promoter activity quantification, we compared RNA-Seq based promoter activity estimates with H3K4me3 histone marks levels, a mark of active transcription at promoters. We downloaded matching RNA-Seq data and H3K4me3 histone mark ChIP-Seq data for all the 68 ENCODE cell lines where both data are available (Table S4). We mapped RNA-Seq reads using TopHat2 (v2.0.12) and ChIP-Seq reads using BWA (v0.7.10-r789) to Gencode (release 19) annotation (Li and Durbin, 2009). Including replicates and cell lines which are treated with chemicals, we processed 1,158 (361 ChIP-Seq and 797 RNA-Seq) data sets. We estimated the promoter activity for all ENCODE RNA-Seq samples using the junction read count approach. We examined the coverage of promoter regions (± 2,000 bps from TSS) for coverage plots, and we used the $\log_2$ of the read counts overlapping these regions for the correlation analysis. The reads overlapping promoter regions are identified by using *featureCounts()* function of Rsubread (v.1.30.9) package in R (Liao et al., 2019). We calculated the Spearman correlation between mean promoter activity estimates from RNA-Seq samples with the mean ChIP-Seq signal for each cell line (biosample term).

To examine the level of ChIP-Seq support for the major, minor and inactive promoters identified by using tumor and normal samples in the pan-cancer cohort, we compared estimated promoter activities from RNA-Seq data with ChIP-seq data obtained from ENCODE project cell lines. For this analysis we used 59 cell ENCODE cell lines that have H3K4me3 data available (Table S4). To identify the matched cell lines, we examined the correlation between mean tissue promoter activity and the ChIP-Seq read counts for each cell line. Only the cell lines that are from same tissue (according to ENCODE metadata) and showing the highest correlation (among top 10) are considered matched.

CAGE Tag Analysis

CAGE tag read count data for 1,829 samples are downloaded from FANTOM5 (http://fantom.gsc.riken.jp/5/datafiles/latest/extra/CAGE_peaks/hg19.cage_peak_phase1and2combined_counts_ann.osc.txt.gz) (Lizio et al., 2017; Lizio et al., 2015). The number of CAGE tag reads overlapping the promoter regions, 100 bps upstream and 50 bps downstream of TSSs, are used as CAGE tag support for each promoter.

Identification of isoform switch events

We find the major transcript of each gene in each tumor type using the mean activity across all TCGA tumor samples (tumor-specific major transcript). Additionally, we find the major transcript based on the pan-cancer mean activity (pan-cancer major transcript). For each tumor type, we then identify the changes in major transcript by comparing tumor specific and pan-cancer major transcripts. A change in major transcript can occur via 2 different mechanisms: either the new tumor specific major transcript is regulated by a different promoter than the pan-cancer major transcript (i.e. a promoter switching event), or the promoter is still the same as the pan-cancer major transcript’s promoter but only the major transcript of this promoter is changed (i.e. a splicing event). For each tumor type, we count the number of major transcript changes for both of these mechanisms. We report the proportions of isoform switching events that can be explained by alternative promoters and splicing.

5’UTR, CDS, and 3’UTR analysis

To understand the functional effect of alternative promoters, we compared the major and alternative promoters for the samples of each tumor type. We determined the major promoters by the mean promoter activity across the samples of the corresponding tumor type. Then, we identify the regions unique to the major promoter, alternative promoter and the regions that are common in both. For each of these regions, we looked at the Gencode (release 19) annotations to determine the functional composition, i.e. 5’ untranslated region (5’UTR), exon, coding sequence (CDS) and 3’ untranslated region (3’UTR). We determined for each region whether we observe these functional regions, and also the fraction of the total region that is observed. To obtain the pan-cancer overview, we considered all the promoter changes occurring across all the tumor types.

Paired Sample Analysis

To inspect the consistency of alternative promoter events across patients, we selected all individuals that have paired cancer and normal samples. In total, we have 766 individuals in 18 cancer types for which we can obtain per-patient estimates of promoter switching. We performed this analysis for all the tumor types where paired samples are available. Using the paired t-test for paired samples, we identified consistently down- and up- regulated alternative promoters for both absolute and relative promoter activity (Benjamini–Hochberg adjusted paired t-test p-values < 0.05). We then tested if this analysis identifies the same alternative promoters as we reported in the manuscript using Fisher’s exact test (Figure S3G).

Randomized Analysis

To investigate the impact of sample sizes on alternative promoter identification, we randomly selected a smaller set of cancer and normal samples across the entire BRCA cohort and compared the resulting set of alternative promoters with the set obtained from the complete data (1,227 BRCA samples and 218 GTEx Breast samples). We used a sample size of 50 (25 cancer and 25 normal) and 100 (50 cancer and 50 normal samples), and repeated this analysis 10 times. For each of these randomly selected data sets (“shuffles”), we calculated the significance of their overlap with the results from the entire data sets using Fisher’s exact test (Figure S3H).

Motif analysis

To uncover the putative transcription factors that might be involved in driving the alternative promoter events, we performed de-novo motif analysis using the RSAT – Metazoa webserver (Thomas-Chollier et al., 2012a; Thomas-Chollier et al., 2012b). We provided the sequences 200bps upstream region of the transcription start sites of of promoters for this enrichment analysis. We estimate enrichment of motifs in cancer-associated alternative promoters compared to the background set of all promoters from active genes in each tissue, thereby specifically searching for motifs that contribute more frequently to alternative promoters than to the generally active promoters in that tissue. We identified the enriched transcription factor motifs by using RSAT webserver’s comparison tool with the 2018 JASPAR transcription factor DNA-binding preferences database (Khan et al., 2018). The list of transcription factor motifs that are enriched for each cancer type can be found in Table S6.

Mutation burden analysis

We examined the noncoding mutation burden at each promoter to study mutation patterns. We used the samples with whole genome sequencing and RNA-Seq data available within the PCAWG cohort for this analysis. The somatic single nucleotide variants (SNV) are identified by the PCAWG consortium using a uniform processing pipeline (Rheinbay et al., 2017a). Here we used the somatic mutation calls provided by the PCAWG consortium to estimate mutation burden. Noncoding mutation burden is calculated as the total number of SNVs within the promoter regions (200bps upstream to TSS) excluding the SNVs overlapping with an exon.

Survival Analysis

We used tumor types with more than 40 samples to identify promoters predictive of survival. We examined the association of absolute promoter activity with patient survival for the patients where clinical data is available. Promoters of genes that have at least 2 promoters with >10 junction reads per sample and a promoter with relative activity > 0.5 in at least %10 of the tumor samples are chosen for this analysis. Samples are stratified into high and low expression groups according to the absolute promoter activity where samples with promoter activity in the top 10% are considered as high expression for each promoter. Survival analysis is performed using the survival package (v.2.42.3) in R and the promoters is considered predictive of survival if BH adjusted p-value < 0.01 (Table S7) (Therneau, 2015; Therneau and Grambsch, 2000). Somatic SNV data used for survival analysis is downloaded from https://gdc.cancer.gov/about-data/publications/pancanatlas (Hoadley et al., 2018).

QUANTIFICATION AND STATISTICAL ANALYSIS

Quantitative and statistical methods are noted above according to their respective technologies and analytic approaches.

R version 3.5.1 was used for all statistical analysis and visualization (R Core Team, 2018). We used the GenomicFeatures (v1.32.3) and GenomicAlignments (v1.16.0) R packages to estimate promoter activity (Lawrence et al., 2013a).

To generate the T-SNE plots we selected the 1,500 promoters and genes with the largest variance across the whole data set. The tsne package (v0.1.3) in R was used to generate T-SNE plots (Donaldson, 2016). All boxplots show median and the inter quartile range (IQR) of the underlying data with whiskers extending to ± 1.5 IQR from boxes. All outliers are shown unless stated otherwise. All figures are generated using ggplot2 (v3.1.0), ggpubr (v0.2) and gplots (v3.0.1) (Kassambara, 2018; Warnes et al., 2019; Wickham, 2009). We used the Wilcoxon test for statistical testing unless stated otherwise. Significance levels are defined as follows: n.s: p ¿ 0.05, *: p ≤ 0.05, **: p ≤ 0.01, ***: p ≤ 0.001 and ****: p ≤ 0.0001.

DATA AND CODE AVAILABILITY

The implementation of the junction read count approach is available online (https://github.com/GoekeLab/proActiv). The datasets generated during this study are provided as supplemental tables and listed in the Key Resources Table.

Supplemental Table Titles

Table S1: The transcript ids with corresponding transcription start site ids, promoter ids and gene ids according to Gencode (release 19) annotations. Related to Figures 1 and S1.

Table S2: The transcription start site coordinates for the compiled promoters. Related to Figures 1 and S1.

Table S3: The sample metadata for the combined RNA-Seq data set including PCAWG, GTEx and TCGA data cohorts. Related to Figures 1 and S1.

Table S4: The sample metadata for ENCODE cell lines which have both RNA-Seq and H3K4me3 ChIP-Seq data available, and matching tissue information for comparison with tumor samples. Related to Figures 1 and S1.

Table S5: The complete list of alternative promoters including tissue specific, cancer associated, multi-cancer associated, BRCA molecular subtype associated and pan-cancer associated alternative promoters. Related to Figures 2, 3, 4, S2, S3 and S4.

Table S6: The transcription factor motifs enriched for cancer associated alternative promoters. Related to Figure 3, S3 and STAR Methods.

Table S7: Promoters which are significantly associated with patient survival. Related to Figures 5 and S5.

REFERENCES

Ayoubi, T.A., and Van De Ven, W.J. (1996). Regulation of gene expression by alternative promoters. FASEB J *10*, 453-460.

Barbieri, C.E., Baca, S.C., Lawrence, M.S., Demichelis, F., Blattner, M., Theurillat, J.P., White, T.A., Stojanov, P., Van Allen, E., Stransky, N., *et al.* (2012). Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet *44*, 685-689.

Bernstein, B.E., Humphrey, E.L., Erlich, R.L., Schneider, R., Bouman, P., Liu, J.S., Kouzarides, T., and Schreiber, S.L. (2002). Methylation of histone H3 Lys 4 in coding regions of active genes. Proc Natl Acad Sci U S A *99*, 8695-8700.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol *34*, 525-527.

Calabrese, C., Davidson, N.R., Fonseca, N.A., He, Y., Kahles, A., Lehmann, K.-V., Liu, F., Shiraishi, Y., Soulette, C.M., Urban, L., *et al.* (2018). Genomic basis for RNA alterations revealed by whole-genome analyses of 27 cancer types. bioRxiv, 183889.

Cancer Genome Atlas Research, N. (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609-615.

Cancer Genome Atlas Research, N. (2012). Comprehensive genomic characterization of squamous cell lung cancers. Nature *489*, 519-525.

Cancer Genome Atlas Research, N. (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. Nature *507*, 315-322.

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A., Taylor, M.S., Engstrom, P.G., Frith, M.C., *et al.* (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet *38*, 626-635.

Chi, P., Allis, C.D., and Wang, G.G. (2010). Covalent histone modifications--miswritten, misinterpreted and mis-erased in human cancers. Nat Rev Cancer *10*, 457-469.

Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Consortium, F., the, R.P., Clst, Forrest, A.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J., Haberle, V., Lassmann, T., *et al.* (2014). A promoter-level mammalian expression atlas. Nature *507*, 462-470.

Consortium, G.T., Laboratory, D.A., Coordinating Center -Analysis Working, G., Statistical Methods groups-Analysis Working, G., Enhancing, G.g., Fund, N.I.H.C., Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida, *et al.* (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204-213.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., *et al.* (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res *46*, D794-D801.

deVos, T., Tetzner, R., Model, F., Weiss, G., Schuster, M., Distler, J., Steiger, K.V., Grutzmann, R., Pilarsky, C., Habermann, J.K., *et al.* (2009). Circulating methylated SEPT9

DNA in plasma is a biomarker for colorectal cancer. Clin Chem *55*, 1337-1346.

Deyoung, M.P., and Ellisen, L.W. (2007). p63 and p73 in human cancer: defining the network. Oncogene *26*, 5169-5183.

Director's Challenge Consortium for the Molecular Classification of Lung, A., Shedden, K., Taylor, J.M., Enkemann, S.A., Tsao, M.S., Yeatman, T.J., Gerald, W.L., Eschrich, S., Jurisica, I., Giordano, T.J., *et al.* (2008). Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat Med *14*, 822-827.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Donaldson, J. (2016). tsne: T-Distributed Stochastic Neighbor Embedding for R (t-SNE).

Fay, M.J., Longo, K.A., Karathanasis, G.A., Shope, D.M., Mandernach, C.J., Leong, J.R., Hicks, A., Pherson, K., and Husain, A. (2003). Analysis of CUL-5 expression in breast epithelial cells, breast cancer cell lines, normal tissues and tumor tissues. Mol Cancer *2*, 40.

Feng, G., Tong, M., Xia, B., Luo, G.Z., Wang, M., Xie, D., Wan, H., Zhang, Y., Zhou, Q., and Wang, X.J. (2016). Ubiquitously expressed genes participate in cell-specific functions via alternative promoter usage. EMBO Rep *17*, 1304-1313.

Finak, G., Bertos, N., Pepin, F., Sadekova, S., Souleimanova, M., Zhao, H., Chen, H., Omeroglu, G., Meterissian, S., Omeroglu, A., *et al.* (2008). Stromal gene expression predicts clinical outcome in breast cancer. Nat Med *14*, 518-527.

Frith, M.C., Valen, E., Krogh, A., Hayashizaki, Y., Carninci, P., and Sandelin, A. (2008). A code for transcription initiation in mammalian genomes. Genome Res *18*, 1-12.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. Nat Rev Cancer *4*, 177-183.

Gerstung, M., Pellagatti, A., Malcovati, L., Giagounidis, A., Porta, M.G., Jadersten, M., Dolatshad, H., Verma, A., Cross, N.C., Vyas, P., *et al.* (2015). Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. Nat Commun *6*, 5901.

Gherardi, E., Birchmeier, W., Birchmeier, C., and Vande Woude, G. (2012). Targeting MET in cancer: rationale and progress. Nat Rev Cancer *12*, 89-103.

Gross, A.M., Kreisberg, J.F., and Ideker, T. (2015). Analysis of Matched Tumor and Normal Profiles Reveals Common Transcriptional and Epigenetic Signals Shared across Cancer Types. PLoS One *10*, e0142618.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., *et al.* (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res *22*, 1760-1774.

Hashimoto, K., Suzuki, A.M., Dos Santos, A., Desterke, C., Collino, A., Ghisletti, S., Braun, E., Bonetti, A., Fort, A., Qin, X.Y., *et al.* (2015). CAGE profiling of ncRNAs in hepatocellular carcinoma reveals widespread activation of retroviral LTR promoters in virus-induced tumors. Genome Res *25*, 1812-1824.

Heber, S., Alekseyev, M., Sze, S.H., Tang, H., and Pevzner, P.A. (2002). Splicing graphs and EST assembly problem. Bioinformatics *18 Suppl 1*, S181-188.

Hegi, M.E., Diserens, A.C., Gorlia, T., Hamou, M.F., de Tribolet, N., Weller, M., Kros, J.M., Hainfellner, J.A., Mason, W., Mariani, L., *et al.* (2005). MGMT gene silencing and benefit from temozolomide in glioblastoma. N Engl J Med *352*, 997-1003.

Hippo, Y., Taniguchi, H., Tsutsumi, S., Machida, N., Chong, J.M., Fukayama, M., Kodama, T., and Aburatani, H. (2002). Global gene expression analysis of gastric cancer by oligonucleotide microarrays. Cancer Res *62*, 233-240.

Hoadley, K.A., Yau, C., Hinoue, T., Wolf, D.M., Lazar, A.J., Drill, E., Shen, R., Taylor, A.M., Cherniack, A.D., Thorsson, V., *et al.* (2018). Cell-of-Origin Patterns Dominate the

Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell *173*, 291-304 e296.

Kaczkowski, B., Tanaka, Y., Kawaji, H., Sandelin, A., Andersson, R., Itoh, M., Lassmann, T., Hayashizaki, Y., Carninci, P., Forrest, A.R., *et al.* (2016). Transcriptome Analysis of Recurrently Deregulated Genes across Multiple Cancers Identifies New Pan-Cancer Biomarkers. Cancer Res *76*, 216-226.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., *et al.* (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333-339.

Kassambara, A. (2018). ggpubr: 'ggplot2' Based Publication Ready Plots.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., van der Lee, R., Bessy, A., Cheneby, J., Kulkarni, S.R., Tan, G., *et al.* (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res *46*, D260-D266.

Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M.A., and Gerstein, M. (2016). Role of non-coding sequence variants in cancer. Nat Rev Genet *17*, 93-108.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol *14*, R36.

Klann, T.S., Black, J.B., Chellappan, M., Safi, A., Song, L., Hilton, I.B., Crawford, G.E., Reddy, T.E., and Gersbach, C.A. (2017). CRISPR-Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. Nat Biotechnol *35*, 561-568.

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., *et al.* (2006). CAGE: cap analysis of gene expression. Nat Methods *3*, 211-222.

Lai, E.C. (2002). Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. Nat Genet *30*, 363-364.

Lapenna, S., and Giordano, A. (2009). Cell cycle kinases as therapeutic targets for cancer. Nat Rev Drug Discov *8*, 547-566.

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013a). Software for computing and annotating genomic ranges. PLoS Comput Biol *9*, e1003118.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., *et al.* (2013b). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214-218.

Lazar, A.J., Tuvin, D., Hajibashi, S., Habeeb, S., Bolshakov, S., Mayordomo-Aranda, E., Warneke, C.L., Lopez-Terrada, D., Pollock, R.E., and Lev, D. (2008). Specific mutations in the beta-catenin gene (CTNNB1) correlate with local recurrence in sporadic desmoid tumors. Am J Pathol *173*, 1518-1527.

Lee, J.H., Soung, Y.H., Lee, J.W., Park, W.S., Kim, S.Y., Cho, Y.G., Kim, C.J., Seo, S.H., Kim, H.S., Nam, S.W., *et al.* (2004). Inactivating mutation of the pro-apoptotic gene BID in gastric cancer. J Pathol *202*, 439-445.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Liao, Y., Smyth, G.K., and Shi, W. (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res.

Liu, J., Lichtenberg, T., Hoadley, K.A., Poisson, L.M., Lazar, A.J., Cherniack, A.D., Kovatich, A.J., Benz, C.C., Levine, D.A., Lee, A.V., *et al.* (2018). An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell *173*, 400-416 e411.

Lizio, M., Harshbarger, J., Abugessaisa, I., Noguchi, S., Kondo, A., Severin, J., Mungall, C., Arenillas, D., Mathelier, A., Medvedeva, Y.A., *et al.* (2017). Update of the FANTOM web resource: high resolution transcriptome of diverse cell types in mammals. Nucleic Acids Res *45*, D737-D743.

Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., Abugessaisa, I., Fukuda, S., Hori, F., Ishikawa-Kato, S., *et al.* (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol *16*, 22.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550.

Maruvka, Y.E., Mouw, K.W., Karlic, R., Parasuraman, P., Kamburov, A., Polak, P., Haradhvala, N.J., Hess, J.M., Rheinbay, E., Brody, Y., *et al.* (2017). Analysis of somatic microsatellite indels identifies driver events in human tumors. Nat Biotechnol *35*, 951-959.

Marx, V. (2017). Choosing CRISPR-based screens in cancer. Nat Methods *14*, 343-346.

Moody, S.E., Schinzel, A.C., Singh, S., Izzo, F., Strickland, M.R., Luo, L., Thomas, S.R., Boehm, J.S., Kim, S.Y., Wang, Z.C., *et al.* (2015). PRKACA mediates resistance to HER2-targeted therapy in breast cancer cells and restores anti-apoptotic signaling. Oncogene *34*, 2061-2071.

Muratani, M., Deng, N., Ooi, W.F., Lin, S.J., Xing, M., Xu, C., Qamra, A., Tay, S.T., Malik, S., Wu, J., *et al.* (2014). Nanoscale chromatin profiling of gastric adenocarcinoma reveals cancer-associated cryptic promoters and somatically acquired regulatory elements. Nat Commun *5*, 4361.

Nissim, L., Wu, M.R., Pery, E., Binder-Nissim, A., Suzuki, H.I., Stupp, D., Wehrspaun, C., Tabach, Y., Sharp, P.A., and Lu, T.K. (2017). Synthetic RNA-Based Immunomodulatory Gene Circuits for Cancer Immunotherapy. Cell *171*, 1138-1150 e1115.

Pal, S., Gupta, R., Kim, H., Wickramasinghe, P., Baubet, V., Showe, L.C., Dahmane, N., and Davuluri, R.V. (2011). Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. Genome Res *21*, 1260-1272.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods *14*, 417-419.

Paulino, V.M., Yang, Z., Kloss, J., Ennis, M.J., Armstrong, B.A., Loftus, J.C., and Tran, N.L. (2010). TROY (TNFRSF19) is overexpressed in advanced glial tumors and promotes glioblastoma cell invasion via Pyk2-Rac1 signaling. Mol Cancer Res *8*, 1558-1567.

Perlman, E.J., Gadd, S., Arold, S.T., Radhakrishnan, A., Gerhard, D.S., Jennings, L., Huff, V., Guidry Auvil, J.M., Davidsen, T.M., Dome, J.S., *et al.* (2015). MLLT1 YEATS domain mutations in clinically distinctive Favourable Histology Wilms tumours. Nat Commun *6*, 10013.

Qamra, A., Xing, M., Padmanabhan, N., Kwok, J.J.T., Zhang, S., Xu, C., Leong, Y.S., Lee Lim, A.P., Tang, Q., Ooi, W.F., *et al.* (2017). Epigenomic Promoter Alterations Amplify Gene Isoform and Immunogenic Diversity in Gastric Adenocarcinoma. Cancer Discov *7*, 630-651.

R Core Team (2018). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).

Rapin, N., Bagger, F.O., Jendholm, J., Mora-Jensen, H., Krogh, A., Kohlmann, A., Thiede, C., Borregaard, N., Bullinger, L., Winther, O., *et al.* (2014). Comparing cancer vs normal gene expression profiles identifies new disease entities and common transcriptional programs in AML patients. Blood *123*, 894-904.

Reyes, A., and Huber, W. (2018). Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. Nucleic Acids Res *46*, 582-592.

Rheinbay, E., Nielsen, M.M., Abascal, F., Tiao, G., Hornshøj, H., Hess, J.M., Pedersen, R.I., Feuerbach, L., Sabarinathan, R., Madsen, T., *et al.* (2017a). Discovery and characterization of coding and non-coding driver mutations in more than 2,500 whole cancer genomes. bioRxiv, 237313.

Rheinbay, E., Parasuraman, P., Grimsby, J., Tiao, G., Engreitz, J.M., Kim, J., Lawrence, M.S., Taylor-Weiner, A., Rodriguez-Cuevas, S., Rosenberg, M., *et al.* (2017b). Recurrent and functional regulatory mutations in breast cancer. Nature *547*, 55-60.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res *43*, e47.

Salazar, R., Roepman, P., Capella, G., Moreno, V., Simon, I., Dreezen, C., Lopez-Doriga, A., Santos, C., Marijnen, C., Westerga, J., *et al.* (2011). Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. J Clin Oncol *29*, 17-24.

Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y., and Hume, D.A. (2007). Mammalian RNA polymerase II core promoters: insights from genome-wide studies. Nat Rev Genet *8*, 424-436.

Santos-Rosa, H., Schneider, R., Bannister, A.J., Sherriff, J., Bernstein, B.E., Emre, N.C., Schreiber, S.L., Mellor, J., and Kouzarides, T. (2002). Active genes are tri-methylated at K4 of histone H3. Nature *419*, 407-411.

Sharma, S., Kelly, T.K., and Jones, P.A. (2010). Epigenetics in cancer. Carcinogenesis *31*, 27-36.

Slamon, D.J., Clark, G.M., Wong, S.G., Levin, W.J., Ullrich, A., and McGuire, W.L. (1987). Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science *235*, 177-182.

Slamon, D.J., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., Bajamonde, A., Fleming, T., Eiermann, W., Wolter, J., Pegram, M., *et al.* (2001). Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. N Engl J Med *344*, 783-792.

Soneson, C., Love, M.I., Patro, R., Hussain, S., Malhotra, D., and Robinson, M.D. (2019). A junction coverage compatibility score to quantify the reliability of transcript abundance estimates and annotation catalogs. Life Sci Alliance *2*.

Takahashi, H., Kato, S., Murata, M., and Carninci, P. (2012). CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks. Methods Mol Biol *786*, 181-200.

Teng, M., Love, M.I., Davis, C.A., Djebali, S., Dobin, A., Graveley, B.R., Li, S., Mason, C.E., Olson, S., Pervouchine, D., *et al.* (2016). A benchmark for RNA-seq quantification pipelines. Genome Biol *17*, 74.

Therneau, T.M. (2015). A Package for Survival Analysis in S.

Therneau, T.M., and Grambsch, P.M. (2000). Modeling Survival Data: Extending the Cox Model (Springer-Verlag New York).

Thomas-Chollier, M., Darbo, E., Herrmann, C., Defrance, M., Thieffry, D., and van Helden, J. (2012a). A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. Nat Protoc *7*, 1551-1568.

Thomas-Chollier, M., Herrmann, C., Defrance, M., Sand, O., Thieffry, D., and van Helden, J. (2012b). RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. Nucleic Acids Res *40*, e31.

Tobias, E.S., Hurlstone, A.F., MacKenzie, E., McFarlane, R., and Black, D.M. (2001). The TES gene at 7q31.1 is methylated in tumours and encodes a novel growth-suppressing LIM domain protein. Oncogene *20*, 2844-2853.

Ulz, P., Thallinger, G.G., Auer, M., Graf, R., Kashofer, K., Jahn, S.W., Abete, L., Pristauz, G., Petru, E., Geigl, J.B., *et al.* (2016). Inferring expressed genes by whole-genome sequencing of plasma DNA. Nat Genet *48*, 1273-1278.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546-1558.

Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., *et al.* (2019). gplots: Various R Programming Tools for Plotting Data.

Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., and Lee, W. (2014). Genome-wide analysis of noncoding regulatory mutations in cancer. Nat Genet *46*, 1160-1165.

Wickham, H. (2009). ggplot2 : Elegant Graphics for Data Analysis (Springer-Verlag New York).

Wiesner, T., Lee, W., Obenauf, A.C., Ran, L., Murali, R., Zhang, Q.F., Wong, E.W., Hu, W., Scott, S.N., Shah, R.H., *et al.* (2015). Alternative transcription initiation leads to expression of a novel ALK isoform in cancer. Nature *526*, 453-457.