Methods

**TCGA samples.** In brief, each TCGA code corresponds to a specific patient, from which both Tumor and matching “Normal” material from the same tissue are available. The tumor types tested and the respective TCGA nomenclature are: BRCA (Breast invasive carcinoma), KIRC (Kidney renal clear cell carcinoma), LIHC (Liver hepatocellular carcinoma), COAD (Colon adenocarcinoma), HNSC (Head and Neck squamous cell carcinoma), THCA (Thyroid carcinoma), UCEC (Uterine corpus endometrioid carcinoma), LUAD (Lung adenocarcinoma), PRAD (Prostate adenocarcinoma) and KIRP (Kidney renal papillary cell carcinoma). RNA for small RNA library construction was recovered as follows: Total RNA (two micrograms) that met the quality standards on the BCR submission documentation was mixed with oligo(dT) MicroBeads and eluted through a 96-well MACS column in a MultiMACS separator (Miltenyi Biotec, Germany). The flow-through was recovered and RNA including small RNAs was extracted by ethanol precipitation. miRNA-Seq libraries were constructed using a plate-based protocol developed at the BCGSC described here [39] (page 21 of Supplementary Material). In brief, after adapter ligations, first-strand cDNA was synthesized using Superscript II Reverse Transcriptase (Invitrogen) and the respective adapter related RT primer. Indexes were inserted during PCR and the respective products were pooled, then size selected to remove large cDNA fragments and adapter contaminants. Libraries were size separated with a cutoff corresponding to small RNAs with length up to 30-31 bases using an automated custom robotic system. RNA quality was checked using an Agilent Bioanalyzer RNA Nano chip.

**Library construction.** We employed an in house developed plate-based protocol including ligation of a 3’ adapter single strand DNA (T4 RNA ligase2, NEB Canada, cat. M0242L) and an RNA 5’ adapter (T4 RNA ligase, Ambion USA, cat. AM2141), followed by 1st strand cDNA synthesis (Superscript II Reverse Transcriptase, Invitrogen, cat.18064 014). During the PCR for the library, sequences were introduced to enable library indexing during sequencing. Negative controls were added during RNA load onto the plate, before the 3’ adapter ligation, and before PCR. PCR conditions were at 98°C (30 sec), 15 cycles of 98°C(15 sec)/ 62°C(30 sec) / 72°C
(15 sec), and 72°C (5min). Quality was checked using a Caliper LabChipGX DNA chip. Subsequently, PCR products were selected based on their size using a 96-channel automated size selection robot developed in house. Then, they were evaluated with an Agilent Bioanalyzer DNA1000 chip and a Qubit fluorometer (Invitrogen, cat. Q32854) regarding their quality and quantity, respectively, before cluster generation and load into the flow cell for sequencing. The way RNAseq libraries and data was produced for this study (selection of poly A transcripts) eliminates from the final sequenced product small or short RNAs that may have been produced during transcription but did not lead to a final mRNA product (for example after Pol pausing).

**smRNA analysis.** BED files with reads ranging from 19nt to 30nt were uploaded to Seqmonk 0.16.0 and quantified as follows: Probes were defined based on the Feature Probe Generator for coordinates of first exon of protein coding genes in Ensemble (via Biomart, http://www.ensembl.org, as of Dec 2011). For each cancer type only probes containing at least one read were used in subsequent analysis, while exact probe duplicates were removed. Two quantifications for each feature were made based on count of reads either in sense or antisense direction of the feature/probe, corrected for total read count (per million reads) and for the probe length and log2 transformed. The term smRNA feature used in the text represents the genomic coordinates of the above mentioned exon 1 probes. For identification of positions of reads present in at least three samples the SeqMonk read probe generation approach was applied. For estimation of the number of samples in which the identified read locations are present, samples were quantified only regarding read counts at these probes and in house perl scripts were used to select these probes per sample with at least one count. For visualization of the reads in the genome, the SeqMonk 0.16.0 genome browser was used.

**General bioinformatic analysis.** Models of distribution of read locations around TSS (and other hotspots) in the genome were estimated using in house python scripts constructing an accumulation model around a hypothetical common TSS for all genes, in which the numbers of read locations around each different TSS (based on the 5' end) were calculated and attributed to defined points around the model TSS. These points were in regular distances within a specific
radius to both directions around TSS (or the base 1 of other hotspots). Cumulative gene models of read density within specific regions such as exon 1 were constructed based on a similar approach as above but using genomic intervals that were superimposed to construct the model instead of genomic points. For each sample reads in each region of interest were summed in windows equal to the 5% of region length and divided to the total length of each section to produce a read density value along the aligned genomic intervals representing a set of regions. This density was normalized based on the total read counts for each sample, and the mean value of all normal samples was divided to that of tumor and then was log2 transformed to give the log fold change of read densities between normal and tumor at the regions of interest.

Annotation datasets used for these models were acquired via Biomart from Ensembl regarding exon 1 of protein coding genes and Ensembl Regulation 67 regarding Transcription factors and PolII (http://www.ensembl.org). For DNAse hypersensitive regions, and DNA methylation data in Fig. 4 (except 4e), data from the Human Epigenome Atlas project (http://www.genboree.org/epigenomeatlas/multiGridViewerPublic.rhtml) in breast (vHMEC) and Breast Luminal Epithelial Cells, respectively, were used. DNA methylation sequencing tracks for Breast of Fig 4e with the respective DNA methylation percentages were acquired from ENCODE [38] data coordination center at UCSC and correspond to DNA methylation by reduced representation bisulfite seq data. The respective DNA methylation and smRNA relative distribution plots were constructed using the respective Seqmonk trend plot functions.

**TCGA DNA methylation:** Patient DNA methylation data used corresponded to level 3 DNA methylation data TCGA Analysis of DNA Methylation for BRCA using Illumina Infinium HumanMethylation450 platform. The patients for which this data was available and has been used are marked with an asterisk in Supplementary Table 4.

**Clustering.** Non-negative matrix-factorization consensus clustering was done as described before [33] using the Gene Pattern Platform from Broad Institute [34]. NMF clustering’s potential to distinguish Cancer from Normal tissues when applied to biologically meaningful data has already been successfully validated in the respective work mentioned above [33].
smRNA values were antiloged and NMF consensus clustering was performed for the following parameters (k 2-5/or up to 10 for all Normal and Tumor tissues, num clusterings 20, max num iterations 2000, error function divergence, random seed 123456789, stop convergence 40, stop frequency 10). K-means consensus clustering for samples of Fig.3d was done on the Gene Pattern Platform (kmax 5, resampling iterations 20, seed value 12345, clustering algorithm KMEANS, distance measure EUCLIDEAN, merge type average, descent iterations 2000). Subsequently, the clustering with the best and more clear separation was finally selected (k=3).

**Differential expression.** For identification of exon 1 features with differential smRNA expression between Normal and Tumor samples the Comparative Marker Selection module of Gene Pattern was used. Each sample in the two datasets (Normal or Tumor) was treated as a different biological replicate in the subsequent statistical evaluation (corrected for multiple testing) and features were selected based on FDR values less than 0.05. smRNA profiles for the selected features were visualized in a heatmap produced by the respective module in Gene pattern applying row normalization. Gene ontology analysis was performed on the David functional annotation platform [35] for an EASE score 0.1 and adjusted values based on Benjamini correction. We tested enrichment for terms in three different databases, the Gene ontology one, Reactome pathways and CGAP_SAGE_Quartile.

**Status prediction.** For the construction of the predictor model, the KNN algorithm available in Gene Pattern was used (num of neighbors 3, euklidean distance). The feature list used was the same identified by Comparative Marker Selection and the input smRNA values were the anti-logged ones.

**TCGA Clinical data, mRNA data, survival and statistical analysis.** Clinical and mRNA data for the patients are available from the TCGA data portal (see below). RNAseq RPKM mRNA gene/level III data were used. The TCGA codes of patients belonging in each group (1,2 or 3) of Fig. 3D are listed in Supplementary Table 5. Survival analysis was done using the R package “survival” (Kaplan- Meier estimator and long-rank test). The statistical test applied
was as described in the package (Terry Therneau (2012). A Package for Survival Analysis in S. R package version 2.36-14). mRNA data for the patients of interest of Suppl. Fig. S7 were retrieved from and tested using the cBIO platform [36] which performed also the respective combinatorial with mRNA values survival analysis. mRNA data tested were as follows (mRNA Expression z-Scores (RNA Seq RPKM), threshold +/- 2). For statistical analysis regarding differential expression and Gene Ontology analysis see previous section and refer to GenePattern module description (http://www.broadinstitute.org/cancer/software/genepattern/modules). Confidence in predictor model corresponds to the proportion of votes to the winning class by the K-nearest-neighbor algorithm and the respective values are listed in Supplementary Table 2. The statistical test regarding enrichment of smRNA locations in exon 1 was done as described previously [37].

**Accession of primary data.** Primary data can be downloaded according to the respective TCGA instructions and data release policy (https://tcga-data.nci.nih.gov/tcga/tcgaAbout.jsp). For mRNA data, and DNA methylation (450-array) data the respective level III files have been used. For bam files deposited in TCGA sections of CGHub (https://cghub.ucsc.edu/), instructions on how to access the files are available here: (https://cghub.ucsc.edu/docs/quickstart.html) and the exact name codes of the bam files in CGHub are available in the Galaxy data sharing platform 's area for this manuscript (https://main.g2.bx.psu.edu/u/publicdata/h/bcgscsmrnaszovoilisetal2013) and can be used to retrieve each bam file from the repository together with the exact smRNA locations of figs 1 and 3.

**Additional Comments**

The “hot-spot” distribution pattern of smRNAs can influence short read RNA sequencing studies that do not employ size or any other form of mRNA selection. For example, when combining all overlapping RNAseq reads or during microarray analysis with probes located in
different exons - strong density/hybridization in a certain portion of a transcript may not indicate alternative exon usage, but rather presence of independently-regulated small RNA species.

In addition, a question tested in this study was whether smRNA values are directly connected with the underlying mRNA values, implying mature mRNA degradation (for example cut by endonucleases) or whether the transcription process producing them is rather indirect and independent of successful mRNA transcription (for example RNA pol II pausing).

In the first scenario, increase in mRNA values should be accompanied by a direct increase in exon1 smRNA values (and correlation values comparable for example to those of Fig S16, in which mature mRNA products concordance post-transcriptionally in different samples was far above r=0.75). The observed correlation of r=0.26, which is very near to what is usually conventionally regarded as the border between no and very weak correlation (r=0.25) is rather explained by Fig. 3f that shows that smRNAs are preferentially (but not exclusively) located at loci of active transcription. Intuitively, in order for these small RNAs to be produced a level of transcription is required. This in turn may favor open chromatin states, in which at a given time point a percentage of genes will be always transcribed introducing a level of transcriptional noise. However, as shown in figure 3f and the supplementary figure S14 a significant number of genes, with no final mRNA product (no gene expression) present expression of small RNAs. Figure 3h shows that the r=0.26 correlation refers to all protein coding genes. The vast majority of these genes do not differ regarding their mRNA / smRNA expression, (figure 3i). Thus, the above mentioned transcriptional noise is expected to be higher in such cases. A more representative approach would be to test the correlation only for genes that statistically differ from each other regarding their smRNA/mRNA profiles as in figures 3g,j and k. In these cases essentially no correlation exists and for the vast number of genes actually the mode of change of smRNAs is the opposite of what it would be normally expected by mRNAs in a degradation scenario. To exclude any bias where a potential reverse correlation might exist, up and down regulated genes were also tested separately.
Of course, any possible regulatory potential of these small RNAs is expected to be realized by affecting production of the final mRNA product either post-transcriptionally or by changing the chromatin state/DNA methylation without necessarily affecting mRNA transcription. To this extent, it is reasonable to assume that smRNA underlying genes should be cancer related, which as shown in Fig. 3b it is true. However, if the discriminatory effect of small RNAs was due to transcriptional noise from cancer related mRNAs we would expect that these potential should be also present if not higher in small RNAs produced by the rest of the exons, which as shown above are more related with mRNA transcription. However, this was not the case.

In addition, small RNAs in the antisense direction upstream TSS have been described to be produced by bidirectional Pol II transcription that also produces the smRNAs at exon1. Any increase in the transcription rate of mRNAs would normally affect also the antisense RNAs and create a reverse association with transcribed cancer related mRNAs. However, we did not found any prediction potential in these RNAs. We neither found such a potential to small RNAs coming from introns, which theoretically correspond to RNAs closer to the initial hnRNA that was produced and thus lack the prediction bias that would have been introduced by post-transcriptional processing.

These results suggest that among all TSS associated or pervasive transcription associated small RNAs, something special about only those in exon1 confers them a prediction potential independent of that of the respective mature mRNA. As discussed in the text, DNA methylation could be a reasonable alternative to explain this prediction potential.
SUPPLEMENT REFERENCES


