The expression level of small non-coding RNAs derived from the first exon of protein-coding genes is predictive of cancer status

Athanasios Zovoilis^{1,2}, Andrew J Mungall¹, Richard Moore¹, Richard Varhol¹, Andy Chu¹, Tina Wong¹, Marco Marra^{1,3} & Steven JM Jones^{1,3,4*}

Abstract

Small non-coding RNAs (smRNAs) are known to be significantly enriched near the transcriptional start sites of genes. However, the functional relevance of these smRNAs remains unclear, and they have not been associated with human disease. Within the cancer genome atlas project (TCGA), we have generated small RNA datasets for many tumor types. In prior cancer studies, these RNAs have been regarded as transcriptional “noise,” due to their apparent chaotic distribution. In contrast, we demonstrate their striking potential to distinguish efficiently between cancer and normal tissues and classify patients with cancer to subgroups of distinct survival outcomes. This potential to predict cancer status is restricted to a subset of these smRNAs, which is encoded within the first exon of genes, highly enriched within CpG islands and negatively correlated with DNA methylation levels. Thus, our data show that genome-wide changes in the expression levels of small non-coding RNAs within first exons are associated with cancer.

Keywords small non-coding RNAs; cancer; TCGA

Subject Categories RNA Biology; Chromatin, Epigenetics, Genomics & Functional Genomics; Molecular Biology of Disease

DOI 10.1002/embr.201337950 | Received 4 September 2013 | Revised 13 January 2014 | Accepted 14 January 2014 | Published online 17 February 2014

EMBO Reports (2014) 15, 402–410

Introduction

A number of recent studies have revealed significant generation of short and small ncRNAs near key positions in the genome and especially in proximity to transcription initiation sites (TSS) (PASRs, TSSa-RNAs, tiRNAs, PROMPT's, other promoter associated RNAs) [1–11] and have challenged the idea that many small non-coding

RNAs observed throughout the genome may be just transcriptional noise or RNA degradation by-products [12–14]. However, their transcription *per se* may not be enough to imply a functional role within the cell or in disease [15], and any relation of these small non-coding RNAs to human disease remains unknown [11]. The Cancer Genome Atlas Project (TCGA) [16] now provides the depth of sampling in both diseased and normal tissues to address this question. Within the initial aims of this project, we have determined miRNA sequences using thousands of samples from different cancer tissues which also includes other small non-coding RNAs (smRNAs) up to 30 nucleotides in length, providing one of the largest resources of smRNAs ever derived. TCGA also generates a significant amount of associated data (genomic, clinical, DNA methylation, mRNA expression). Therefore, this resource provides the unprecedented possibility to test whether these RNAs can be associated with human disease.

Results and Discussion

smRNA locations are conserved between different individuals

Firstly, we mapped small RNA reads from 47 TCGA patients with breast invasive carcinoma (BRCA) (47 pairs of tumor and matched “normal” samples) (347M reads) (Fig 1A). We filtered out all known small non-coding RNAs resulting in 26M putative smRNAs. Fig 1B,C depict the mapping locations of these smRNAs for a randomly selected genome segment.

As shown in Fig 1B, the distribution of smRNAs seems rather chaotic when smRNA data from all samples are superimposed and pooled. However, comparing separately smRNAs among samples reveals a striking consistency at most of the transcribed locations across different individuals and samples (Fig 1C). Comparing all the samples with each other indicates that the vast majority of smRNAs (87%) are transcribed from regions that produce an smRNA in at least 3 different samples, while 43% of them in more than 10

¹ BC Cancer Agency Genome Sciences Centre, Vancouver, BC, Canada

² Department of Molecular Biology, Harvard Medical School, Boston, MA, USA

³ Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada

⁴ Department of Molecular Biology and Biochemistry, Simon Fraser University, Vancouver, BC, Canada

*Corresponding author. Tel: +1 604 877 6083; Fax: +1 604 876 3561; E-mail: sjones@bcgsc.ca

Figure 1. smRNA genomic locations are consistent across different individuals and samples.

- A** Analysis overview.
- B** smRNA locations at a randomly selected region. The top two lines of rectangles represent genes and mRNAs. Each line represents smRNA locations for pooled smRNA reads from either all samples, all normal or all tumor samples.
- C** Each line represents locations of smRNAs reads sequenced from a single sample for all 94 BRCA normal and tumor samples. Red and blue colors represent transcription in the sense or antisense direction, respectively.
- D** Proportion of smRNA locations producing smRNAs in multiple samples. The pie chart depicts the classification to 10 groups of these smRNA locations for samples of (A).
- E** NMF consensus clustering [31] of all samples (normal and tumor) based on smRNA sense expression profiles of each sample at (i) exon 1 (normalized per number of reads/per feature length) (left), (ii) at regions 500 bp upstream (middle) or (iii) downstream (right panel) of TSS. The lower panels depict the resulting two clusters of samples (eclipses). The number represents misclassification rate.
- F** NMF consensus clustering as in (E) based on smRNA antisense expression profiles.

samples (Fig 1D). As shown previously, smRNA coverage of the human genome is less than 1% [2]. Thus, this degree of specificity and conservation of smRNA locations across a large number of

different samples, including cancer tissues, suggests common underlying mechanisms that have produced them at exactly these genomic locations.

Previous studies revealed two peaks of smRNAs around TSS: sense and antisense and downstream and upstream, respectively [1–11]. Consistent with what has been shown for reads [5,8], conserved smRNA locations are enriched around TSS (Supplementary Fig S1A), transcription factor and PolII binding sites and DNase hypersensitive locations (Supplementary Fig S2). These results suggest that smRNA read distribution models described previously around TSS in other systems correspond to transcripts produced from locations conserved among individuals.

smRNAs in exon 1 show distinct expression profiles that can predict cancer status

The abbreviation smRNAs used throughout the study stands for small non-coding RNAs and, in case of RNAs near TSS, may correspond to classes of small RNAs near TSS mentioned in the introduction. These RNA classes include smRNAs that are upstream or downstream of TSS as well as in exon or intron regions. We questioned whether any of these classes may have a biological relevance and whether this is correlated with their exact location with respect to TSS.

Sixty-one percent of our smRNA locations were mapped within gene boundaries, and the overlap with exon 1 was disproportionately higher than it would be expected by chance (7.4% compared with the expected 0.7%, $P = 0.004975$, $FDR < 10\%$). As shown in Supplementary Fig S1B, first exon smRNAs are a distinct subclass of smRNAs regarding expression pattern compared with other regions. In addition, a previous study found that a significant portion (11–18%) of genes have short and small RNAs only in the first exon [2]. For these reasons, we focused on smRNAs at exon 1 and questioned how the 94 BRCA samples are classified based on their smRNA profiles.

Surprisingly, using an unsupervised clustering approach based on sense smRNA profiles of 82,633 exon 1 probes, the resulting separation of our samples in clusters corresponded to a separation of the samples based on their origin from tumor or normal breast tissues (Fig 1D). smRNAs did not cluster together samples coming from the same individuals, but instead clustered the vast majority of samples to either normal or cancer clusters. Thus, the distinct expression profile of smRNAs at exon 1 is correlated with either the cancer or normal tissue origin of a sample.

As shown in Supplementary Fig S3, among the other known small non-coding RNAs that were tested, the only ones that showed a similar classification potential such as smRNAs were miRNAs, which are already well-established cancer prediction markers compared with tRNAs, snRNAs and snoRNAs that have more general cell housekeeping functions [17–24].

smRNA cancer status prediction is a widely observed characteristic

That smRNA profiles were able to distinguish efficiently between normal and breast cancer tissues raised the question how applicable

this ability is to other tumor types. Interestingly, in 8 of 10 TCGA tumor types tested (for which the respective normal tissues of the same patients were available, including BRCA), smRNAs shared this characteristic (Fig 2). Moreover, misclassification rates were comparable to those of the respective miRNAs in most tissues (Supplementary Fig S4). These findings also suggest that future small ncRNA cancer prediction assays could take into account also exon 1 smRNAs.

Moreover, smRNA profiles were able to classify samples coming from different normal tissues, suggesting a tissue-specific smRNA profile for the tissues tested (Supplementary Fig S5).

Differentially expressed smRNAs provide discriminating features between cancerous and normal tissues

Next, we investigated how the distinct expression profile of smRNAs at exon 1 was connected with the cancer or normal tissue origin of BRCA samples. To this end, we identified 3,660 exon 1 features of protein-coding genes for which smRNAs are differentially expressed between normal and tumor BRCA samples tested above ($FDR < 0.05$, $\log_2$ fold change threshold ± 1) (Fig 3A). Consistent with the ability of smRNAs to distinguish between normal and tumor tissues, differentially expressed locations correspond to genes strongly related with oncogenic processes (programmed cell death, telomere maintenance and apoptosis) which have been previously identified to be altered in breast cancer (Fig 3B). The vast majority of features are located in genes with high CpG content promoters, which have been previously associated with cell cycle and transcription regulation [25] (Supplementary Fig S6). Moreover, as shown in Supplementary Fig S7, when the identified locations based on smRNA expression profiles are combined with TCGA mRNA data for these locations and TCGA survival data for the same patients, changes in mRNA expression of these genes are strongly connected with a highly tumorigenic phenotype. These findings reveal that changes in smRNA profiles occur at hotspots of the genome with high tumorigenic and regulatory potential.

To test whether these changes in smRNA profiles can be used to predict cancer status of a sample, the identified differentially expressed features were used to train a prediction model on the dataset of the 94 BRCA samples tested above and subsequently to predict cancer status of a new independent dataset of 114 BRCA samples (57 normal and 57 tumor). smRNA profiles predicted efficiently with an error rate of 0.08, the correct status of samples in the new set (Fig 3C) with a confidence presented in Supplementary Table 2. Interestingly, predictive power of smRNAs (91%) was almost equal with that of an already published and widely cited solid tumor miRNA signature [18] (92%) (Supplementary Fig S8A). In contrast, use of randomly selected smRNA signatures provided much lower classification efficiencies (78%) (Supplementary Fig S8B). Similar results were observed in KIRC and LIHC samples for which number of samples allowed this type of analysis (Supplementary Figs S9 and S10). The ability of differentially expressed smRNAs to predict cancer status of an unknown sample is of high

Figure 2. NMF consensus clustering for nine additional TCGA sets of normal and tumor samples based on their smRNA profiles as described in Fig 1. The number under the clusters represents the number of misclassifications to the total number of clustered samples. For the last two tumor types, separation of samples was not possible.

Figure 3. smRNA profiles can predict cancer status. Differential expression analysis for smRNAs between normal and tumor BRCA samples (set A) identified 3,660 exon 1 features.

- A The heatmap depicts smRNA expression for the 3,660 features (rows) for all 94 samples (columns) (normalized per number of reads and feature length, $\log_2$ -transformed).
- B Function annotation for genes at the 3,660 feature locations (red: direct connection with tumorigenesis).
- C Prediction of cancer status of a new set of BRCA normal and tumor samples (Set B) based on smRNA profiles at the 3,660 features.
- D Consensus k-means clustering based on smRNA profiles at exon 1 of the 104 BRCA tumor samples. The survival plot depicts outcomes for the resulting three groups of patients.
- E smRNA expression profile for the 143 exon 1 features with smRNA differential expression between patients with worst (Group 1) or better outcome (Groups 2 and 3) (FDR < 0.05, log fold threshold 1). Lower panel: estrogen and progesterone receptor status for each tumor sample (red lines: negative receptor status).
- F Mean mRNA values of regions with/without exon 1 smRNA expression in BRCA samples. The colograde diagram represents exon 1 smRNA expression values of all protein-coding genes, sorted from lower to higher ($\log_2$ scale).
- G Circos diagram representing heatmap of exon 1 smRNA values (in normal vs. tumor) for the smRNA differentially expressed regions of Fig 3A (right) and the respective mRNA ratios (left) ($\log_2$ scale) (range of red: upregulated in normal, range of blue: upregulated in tumor). Heatmaps correspond to the same features ordered based on either smRNA or mRNA values. The lines connect smRNA and mRNA values that belong to the same feature and show reverse mode of change between normal and tumor. The numbers correspond to numbers of features upregulated in normal (red) or tumor (blue). mRNA values are RPKM values and smRNAs as in (A).
- H Relationship between exon 1 smRNA and mRNA values for all protein-coding genes ($P < 2.2\text{e-}16$).
- I Relationship between normal/tumor ratios of smRNAs and mRNAs for all genes with smRNA expression.
- J–K Same as in (I) but for smRNA differentially expressed regions (j, upregulated in normal; k, upregulated in tumor).

Data information: In (H–K), data are depicted in $\log_2$ scale, $R < 0.25$ was regarded as no correlation and no P -values are depicted.

importance, because it may be applicable and transferable to routine, diagnostic procedures, a possibility that has yet to be extensively addressed.

Based on the ability of smRNA profiles to define cancer status of the samples, we questioned whether these profiles could be also used to classify patients in groups of different survival outcomes. Indeed, as shown in Fig 3D, applying k-means consensus clustering, global smRNA profiles identified three groups of patients, one of which had a worse clinical outcome and was characterized by an increased portion of negative estrogen and progesterone receptor status (Fig 3E). smRNA features that contribute to this difference and are differentially expressed between the different clinical outcome groups are depicted in Fig 3E and listed in Supplementary Table 3. Consistent with the observed different survival outcomes, these features are involved in p53 signaling and apoptosis, suggesting that separation based on smRNAs has a biologically meaningful basis (Supplementary Fig S11).

smRNAs near TSS are a heterogeneous group

We questioned whether the above potential of exon 1 smRNAs is also shared by the other smRNAs in the region. In contrast to exon 1 smRNAs, smRNAs in the antisense direction (Fig 1F) as well as smRNAs at other exons, introns and promoter region (1 kb upstream TSS) (Supplementary Figs S12 and S13) did not show the same potential. The fact that a portion of exons that are annotated as internal could also be first due to yet unannotated alternative TSS may have contributed to the slightly better classification potential of this group compared with the other smRNAs not in exon 1. This group contains also other previously described smRNAs associated with 3' exons end and termination of transcription, for which however no better discrimination potential was observed. In addition, the proximity of smRNAs to the TSS itself was not sufficient to classify efficiently normal and tumor samples as smRNAs 500 bp upstream or downstream of the TSS in both sense and antisense direction did not show comparable results with the smRNAs of exon 1 (Fig 1E,F). These find-

ings suggest that the smRNAs near TSS and promoters are a rather heterogeneous group in regard to their function potential and that there is something specific in exon 1 that contributes to their prediction potential. Thus, we further focused on two possible aspects that could explain this potential: mRNA production and DNA methylation in these loci.

Exon 1 smRNAs are not directly associated with mature mRNA levels

We questioned whether differences in smRNA profiles between normal and tumor samples are directly associated with differences in the mRNA expression of the respective genes, which would imply an mRNA degradation scenario and confound their prediction potential. As shown in Fig 3F, mRNAs can be divided into two classes concerning whether they express first exon smRNAs (52%) or not (48%). Consistent with previous reports [2,7], mRNA expression was found to be significantly higher at regions of smRNA production compared with exon 1 regions with no smRNA production. However, as shown in Supplementary Fig S14D, first exons with negligible or no mature mRNA expression can also express smRNAs. In addition, within the class of mRNAs at regions with smRNA expression, the levels of first exons smRNAs are very weakly correlated with mRNA levels ($r = 0.26$, $P < 2.2\text{e-}16$) (Fig 3 H) compared with a stronger, but still weak correlation observed in smRNAs from other exons ($r = 0.33$, $P < 2.2\text{e-}16$) (Supplementary Fig S15). This is in accordance with previous studies suggesting the production of these RNAs through events such as PolII pausing and other transcriptional events [7,8,12]. The very weak correlation ($r = 0.26$) could be attributed to the fact that these small RNAs in order to be produced at least a basal level of transcription is required within open chromatin sites, where at any given time point, a percentage of genes will be always transcribed introducing a level of noise when testing across all genes in Fig 3 H. Importantly, we observed no correlation between mRNA and smRNA levels from genes differentially expressed between normal and cancer tissues ($r = 0.05$, $P < 0.00001$) (Supplementary Fig S16). In addition, while smRNA

and mRNA levels in normal samples are highly correlated with smRNA and mRNA levels in tumor samples, respectively, ($r > 0.75$) (Supplementary Fig S16), smRNA differential expression between normal and tumor does not correlate with the respective mRNA differential expression (Fig 3I,J,K). As shown in Fig 3G, upregulation or downregulation of smRNAs between normal and tumor samples is not necessarily followed by the same mode of change regarding the mRNA levels of the respective genes.

These results suggest that the potential of exon 1 smRNAs to predict cancer is more associated with the active state of transcription at a locus rather than simply correlated with the production of the underlying mature mRNA itself. However, the possibility that these smRNAs may still reflect mRNA transcription influencing their cancer prediction potential cannot be absolutely excluded. Thus, we applied the same cancer prediction models for the same loci but using the mature mRNA levels instead of those of smRNAs

and questioned whether they would result in similar results. As shown in Supplementary Fig S17, although the connection of these mRNAs with cancer is strong and mRNA levels have a cancer prediction potential themselves, the observed smRNA predictive power cannot be explained by these mRNA levels across all patients studied. Moreover, a simple scenario of smRNA/mature mRNA dependence cannot explain why smRNAs derived from other exons or other possible mRNA transcription by-products that come from the same mRNA final product do not have such a discriminatory potential as those from the first exon (Fig 1, Supplementary Figs S12 and S13).

CpG islands are the center of exon 1 smRNA locations

We questioned whether there is another key regulatory element preferentially located in exon 1 that could be connected with exon 1 smRNAs and their cancer prediction potential. Aberrant methylation of CpG island promoters has been described in the past in various cancer types [26]. As in case of smRNAs, CpG islands are preferentially located within exon 1 of genes (more than 56%) and as shown in Supplementary Fig S6 the vast majority of smRNAs that are differentially expressed derive from high CpG content promoters. Thus, we questioned whether there is any connection between smRNAs and CpG island locations.

As shown in Fig 4B, smRNA locations are mainly found in first exons containing CpG islands. In fact, CpG islands are found to be the center of these locations within exon 1 (Fig 4D,F,H). As shown in Fig 4D, there is a sharp increase in the spatial distribution of smRNAs at the beginning of CpG islands followed by an equal decrease at their end. This enrichment is observed only in CpG islands within first exons (Fig 1F). Consistent with this finding, the correlation of smRNAs with TSS proximity within first exons (Fig 4A,C) is eliminated in first exons lacking CpG islands (Fig 4C). Therefore, the small RNAs (smRNAs) described in this study appear to represent a distinct subclass of previously described TSS-associated RNAs, the locations of which are highly enriched in CpG islands.

smRNAs and DNA methylation status

Since transcriptional activity has been connected with the protection of CpG islands from DNA methylation [27], we asked whether smRNA levels are connected with DNA methylation. Indeed, as shown in Fig 4E and H, smRNA expression is directly associated with a low DNA methylation status. This association is stronger in the first exons containing CpG islands (Fig 4G), and, in particular, within the actual CpG islands of first exons, high levels of smRNAs are connected with low DNA methylation levels and vice versa (Fig 4 H, Supplementary Fig S18). To this end, differences between normal and tumor samples in smRNA levels at specific features were associated with aberrant methylation of CpGs at these locations (Supplementary Fig S19). There is also a direct correlation between the spatial distribution of smRNAs within exon 1 and DNA methylation (Fig 4C). Since the correlation of smRNAs with CpG islands and low methylation could be confounded by the prevalent higher level of general transcription at those loci, we performed the same analysis also for loci with no or negligible gene expression which however do express

smRNAs, and we determined the same reverse correlation (Supplementary Fig S14).

Thus, these findings raise the intriguing possibility of exon 1 smRNAs participating in regulation of DNA methylation at CpG islands. Future studies could test whether such a mechanism underlies the remarkable association of differences in their expression with cancer despite their relative independence from the expression levels of their associated mature mRNAs.

Our study reveals that genome-wide changes in expression levels of small non-coding RNAs within first exons are associated with cancer. In contrast to what may have been previously considered as transcriptional “noise” in cancer, we show that smRNA expression profiles can potentially discriminate between cancerous and normal tissue. This is the first time that small ncRNAs near TSS have been associated with human disease. The current work provides a basis to address whether the alteration of transcription at these hotspots is contributory to different cancer phenotypes, especially regarding the aberrant methylation of CpG islands. Based on this data, we acknowledge the potential role of these smRNAs as contributing factors in oncogenesis, a potential that may extend to the molecular biology of other diseases.

Materials and Methods

Small RNA deep sequencing and mapping

Details on TCGA are provided in the following links: <https://wiki.nci.nih.gov/display/TCGA/TCGA+User's+Guides> and <http://cancergenome.nih.gov/>. A list of the public TCGA codes regarding material from cancer patients used in this study is available in Supplementary Table 4. miRNA-Seq library construction, sequencing at an Illumina GAIIx or HiSeq 2000 and mapping of reads were done as previously described [28]. Reads were attributed to different samples based on their index read sequences. For reads from libraries passing the quality control (regarding abundance of reads from each indexed sample in the pool, adapter dimers, sequencing errors), adapter sequence was trimmed off as described previously [28]. Trimmed reads for each sample were aligned either to NCBI36 (hg18) or later to NCBI GRCh37 (hg19) reference genome using BWA (bwa-0.5.7, default mis-matches on single end mode) [29] and stored in BAM format files. hg18 aligned samples were also converted to hg19 reference genome.

Small RNA filtering

BEDTools [30] were used to convert sequence alignments of each sample separately from BAM to BED format and subsequently to filter out those within genomic coordinates of known non-coding RNAs. Before filtering, for samples with read coordinates in hg18, conversion into hg19 coordinates using liftOver (UCSC Genome Browser) was applied. Coordinates of known non-coding RNAs were acquired for microRNAs, snoRNAs, snRNAs, rRNAs, tRNAs via SeqMonk v.0.16.0 (<http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/>) from GRCh37 human genome datasets which are annotated by Ensembl as of Dec 2011 and from the fRNAdb database version 3.0 for all the above as well as piRNAs (<http://www.ncrna.org/frnadb/download>). Filtering was done by intersect-

Figure 4. CpG islands are a center of smRNA locations.

- A Spatial distribution of exon 1 smRNAs within exon 1.
 B Number of smRNA locations in first exons with and without CpG islands.
 C Distribution of smRNA and MeDIP seq reads in first exons that either contain (upper panel) or do not contain CpG islands (lower panel). Densities are weighted per feature in a model representing exons divided in 100 bins.
 D Enrichment of CpG islands in smRNA reads. Left and right panel correspond to the left and right border, respectively, of CpG islands. The distribution corresponds to read densities around these borders, each of which is located in bin number 100.
 E Aggregate distribution models of smRNAs around possible DNA methylation sites in hypomethylated or hypermethylated regions.
 F Distribution of smRNA reads in CpG islands within (left) or outside (right) exon 1.
 G DNA methylation levels in CpG islands with low (smRNAs bottom 10%) or high (top 10%) smRNA content.
 H The same as in (G), but with a representation depicting the spatial distribution of smRNAs and DNA methylation around CpG islands.

ing the sample reads with the known ncRNA coordinates and reporting only those reads that have no overlap with them.

Accession of primary data

Primary data can be downloaded from the TCGA data portal, <http://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp>, using the TCGA codes mentioned in Supplementary Table 4. Lower levels of sequencing

data such as alignments in form of bam files are deposited in TCGA sections of CGHub (<http://cghub.ucsc.edu/>).

Supplementary information for this article is available online: <http://embo.embopress.org>

Acknowledgements

The TCGA Research Network. Supported by Grant Number U24CA143866 from the National Cancer Institute. The content is solely the responsibility of the

authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health. Funded also by the Canadian Cancer Society (grant No. 2010-700329 and grant No. 018080) and the BC Cancer Foundation. SJMJ is a Scholar of the Michael Smith Foundation for Health Research. AZ is supported by the European Molecular Biology Organization. We thank our colleagues at the Michael Smith Genome Sciences Centre and the Technology Development groups for data production and pipeline optimization.

Author contributions

AZ and SJMJ contributed to conception and design; AJM, RM, RV, AC, TW, MM and SJMJ were involved in generation of mapped sequenced reads; AZ performed data analysis; AZ and SJMJ carried out data interpretation and wrote the manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447: 799–816
2. Kapranov P, et al (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316: 1484–1488
3. Clark MB, et al (2011) The reality of pervasive transcription. *PLoS Biol* 9: e1000625
4. Preker P, et al (2008) RNA exosome depletion reveals transcription upstream of active human promoters. *Science* 322: 1851–1854
5. Seila AC, et al (2008) Divergent transcription from active promoters. *Science* 322: 1849–1851
6. Jacquier A (2009) The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat Rev Genet* 10: 833–844
7. Taft R, et al (2008) Tiny RNAs associated with transcription start sites in animals. *Nat Genet* 41: 572–578
8. Taft R, et al (2010) Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans. *Nat Struct Mol Biol* 17: 1030–1034
9. Mercer TR, et al (2011) Expression of distinct RNAs from 3 untranslated regions. *Nucleic Acids Res* 39: 2393–2403
10. Preker P, et al (2011) PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. *Nucleic Acids Res* 39: 7179–7193
11. Gingeras T (2012) RNA discrimination: Patience is a virtue. *Nature* 482: 310–311
12. FejesToth K, et al (2009) Post-transcriptional processing generates a diversity of 5-modified long and short RNAs. *Nature* 457: 1028–1032
13. Kanhere A, et al (2010) Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. *Mol Cell* 38: 675–688
14. Guang S, et al (2010) Small regulatory RNAs inhibit RNA polymerase II during the elongation phase of transcription. *Nature* 465: 1097–1101
15. Kowalczyk M, Higgs D (2012) RNA discrimination: Quantity or quality? *Nature* 482: 310
16. The Cancer Genome Atlas Research Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455: 1061–1068
17. Lu J, et al (2005) MicroRNA expression profiles classify human cancers. *Nature* 435: 834–838
18. Volinia S, et al (2006) A microRNA expression signature of human solid tumors defines cancer gene targets. *PNAS* 103: 2257–2261
19. Murakami Y, et al (2006) Comprehensive analysis of microRNA expression patterns in hepatocellular carcinoma and non-tumorous tissues. *Oncogene* 25: 2537–2545
20. Mitra R, et al (2010) SFSSClass: an integrated approach for miRNA based tumor classification. *BMC Bioinformatics* 11(Suppl. 1): S22
21. Bandyopadhyay S, et al (2010) Development of the human cancer microRNA network. *Silence* 1: 6
22. Zheng Y, Kwok C (2006) Informative microRNA expression patterns for cancer classification. *Data Mining Biomed Appl* 3916: 143–154
23. Caramuta S, et al (2010) MicroRNA expression profiles associated with mutational status and survival in malignant melanoma. *J Invest Dermatol* 130: 2062–2070
24. Li X, et al (2010) Survival prediction of gastric cancer by a seven-microRNA signature. *Gut* 59: 579–585
25. Saxonov S, et al (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *PNAS* 103: 1412–1417
26. Jones P, Baylin S (2002) The fundamental role of epigenetic events in cancer. *Nat Rev Genet* 3: 415–428
27. Ginno P, et al (2012) R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Mol Cell* 45: 814–825
28. The Cancer Genome Atlas Network (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487: 330–337.
29. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 26: 589–595
30. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26: 841–842
31. Brunet JP, et al (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci USA* 101: 4164–4169