SCIENCE CHINA
Life Sciences

• RESEARCH PAPER •

June 2013 Vol.56 No.6: 503–512
doi: 10.1007/s11427-013-4485-1

Genome-wide identification of cancer-related polyadenylated and non-polyadenylated RNAs in human breast and lung cell lines

ZHAO GuoGuang[1,2†], JIAO Fei[3†], LIAO Qi[4], LUO HaiTao[1], LI Hui[1], SUN Liang[1,5], BU DeChao[1], YU KunTao[1], ZHAO Yi[1*] & CHEN RunSheng[1,6*]

[1]*Bioinformatic Research Group, Key Laboratory of Intelligent Information Processing, Advanced Computing Research Laboratory, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;*
[2]*University of Chinese Academy of Sciences, Beijing 100190, China;*
[3]*Department of Biochemistry and Molecular Biology, Binzhou Medical College, Yantai 264003, China;*
[4]*Department of Preventive Medicine, School of Medicine, Ningbo University, Ningbo 315211, China;*
[5]*College of Computer Science and Technology, Jilin University, Changchun 130012, China;*
[6]*Bioinformatics Laboratory and National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China*

Received January 14, 2013; accepted April 15, 2013; published online May 10, 2013

Eukaryotic mRNAs consist of two forms of transcripts: poly(A)+ and poly(A)–, based on the presence or absence of poly(A) tails at the 3′ end. Poly(A)+ mRNAs are mainly protein coding mRNAs, whereas the functions of poly(A)– mRNA are largely unknown. Previous studies have shown that a significant proportion of gene transcripts are poly(A)– or bimorphic (containing both poly(A)+ and poly(A)– transcripts). We compared the expression levels of poly(A)– and poly(A)+ RNA mRNAs in normal and cancer cell lines. We also investigated the potential functions of these RNA transcripts using an integrative workflow to explore poly(A)+ and poly(A)– transcriptome sequences between a normal human mammary gland cell line (HMEC) and a breast cancer cell line (MCF-7), as well as between a normal human lung cell line (NHLF) and a lung cancer cell line (A549). The data showed that normal and cancer cell lines differentially express these two forms of mRNA. Gene ontology (GO) annotation analyses hinted at the functions of these two groups of transcripts and grouped the differentially expressed genes according to the form of their transcript. The data showed that cell cycle-, apoptosis-, and cell death-related functions corresponded to most of the differentially expressed genes in these two forms of transcripts, which were also associated with the cancers. Furthermore, translational elongation and translation functions were also found for the poly(A)– protein-coding genes in cancer cell lines. We demonstrate that poly(A)– transcripts play an important role in cancer development.

polyadenylation, non-polyadenylation, function annotation, high throughput sequencing, cancer bioinformatics

Citation: Zhao G G, Jiao F, Liao Q, et al. Genome-wide identification of cancer-related polyadenylated and non-polyadenylated RNAs in human breast and lung cell lines. Sci China Life Sci, 2013, 56: 503–512, doi: 10.1007/s11427-013-4485-1

Eukaryotic mRNAs include two main types: poly(A)+ and poly(A)– transcripts, based on the presence or absence of a poly(A) tail at the 3′ end [1,2]. However, some mRNAs can have both poly(A)+ and poly(A)– forms, and are 'bimorphic' [3]. Polyadenylation at the 3′ ends of mRNA transcripts helps to stabilize mRNA and enable smooth nucleocytoplasmic exportation and protein translation. Thus, the majority of known mRNA transcripts are poly(A)+ [4], whereas the poly(A)– forms usually encode ribosomal RNAS [5], histone RNAs [6], tRNAs, and certain small [7] and long non-coding RNAs [8]. It was previously consid-

†Contributed equally to this work
*Corresponding author (email: crs@sun5.ibp.ac.cn; biozy@ict.ac.cn)

ered that all functional transcripts were polyadenylated; thus, numerous studies have identified and investigated the functions only of poly(A)+ transcripts. In contrast, studies on the poly(A)− form of mRNA have been limited [9,10]. However, since the development of better sequencing technologies, increasing numbers of poly(A)− transcripts have been identified and reported [1,11]. For example, Cheng et al. [12] used tilling array analyses to show that 43.7% of all transcribed gene sequences from 10 human chromosomes in eight cell lines are the poly(A)− form. In addition, using 454 sequencer technology, Wu et al. found that 64% of 13467 distinct 3′ expressed sequence tags of cytoplasmic transcripts from HeLa cells are poly(A)− or bimorphic [1]. Furthermore, 1911 (15.8%) and 2828 (25.8%) poly(A)− and bimorphic transcripts in H9 human embryonic stem cells (hESCs) and HeLa cells, respectively, were identified using RNA-seq analyses [11]. Taken together, the poly(A)− and bimorphic RNA transcripts comprise a significant proportion of the human transcriptome. Nevertheless, only a few studies have investigated their functions, especially those differentially expressed in normal and cancerous cells. A recent study showed that ZNF, a member of the zinc finger factor protein family, has the poly(A)− non-histone mRNAs znf460 and sesn3 [11]. Thus, there is a need to determine the levels of poly(A)− or the bimorphic form of RNA expression and their functions in normal and cancerous cells to better understand the cell biology and mechanism of cancer development and progression.

In this study, we identified the poly(A)+ and poly(A)− transcripts in HMEC, MCF-7, NHLF, and A549 cell lines and explored their potential functions in human cancer. We analyzed eight RNA transcript datasets [13] representing poly(A)+ and poly(A)− transcripts in the four human normal, adenocarcinoma breast and lung tissues cell lines.

1 Materials and methods

1.1 Data set

We used eight strand-specific RNA-seq datasets, which contained transcriptomic information of ~200 bp poly(A)+ and poly(A)− RNA fragments from the normal human mammary epithelial cell line HMEC, the breast cancer cell line MCF7, the normal human lung fibroblast cell line NHLF and the lung cancer cell line A549 (Table S1). The ENCODE/Cold Spring Harbor Laboratory using the Illumina GAIIx platform generated the eight datasets, following the protocol described in http://www.ncbi.nlm.nih.gov/pubmed/19620212. The datasets are part of the Encyclopedia of DNA Elements (ENCODE) Project [13] and contain 808.7 million 76 bp paired-end sequences in terms of total reads and approximately 101.1 million reads per sample. All of the long RNA-seq data used in this study can be downloaded from the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under the accession number GSE30567.

The human reference genome, Hg19, annotation resources, including the RefSeq data, the UCSC data, GENCODE data, Ensembl data, and the pseudogene annotation were downloaded from UCSC in September 2011 (http://genome.ucsc.edu/) (Table S2).

1.2 Identification of poly(A)+, poly(A)−, and bimorphic transcripts

To comprehensively construct transcripts from RNA-seq data and to identify the poly(A)+, poly(A)−, and bimorphic transcripts, we developed a computational workflow comprising three steps (Figure 1).

1.2.1 Integration of transcripts

We constructed the transcriptome from each sample and integrated these transcripts with annotated transcripts as a unified transcript set (Figure 1A).

(1) RNA-seq read mapping. All the reads of the RNA-seq data were aligned to the human genome (Hg19) using the fast splice junction mapper, TopHat version V1.4.1 [14]. Briefly, it aligned the reads to the human genome using the ultra high-throughput short read aligner Bowtie [15] and then analyzed the mapping results to identify splice junctions between exons based on canonical and non-canonical splice sites flanking the aligned reads. In this analysis workflow, two iterations of TopHat alignments were used to acquire the splice site information derived from these eight samples of four human cell lines in the ENCODE project [13] (GEO accession No. GSE30567). Before analyzing these eight samples from the four cell lines, we constructed a splice site pool from the eight samples using TopHat. We then re-aligned each sample using the pool (using 'raw-juncs' and 'no-novel-juncs' parameters).

(2) RNA-seq transcriptome assembly. To construct the transcriptome of each sample, Cufflinks [16] was used to assemble the mapped reads aligned to the genome. Briefly, Cufflinks assembled the transcripts by accepting the aligned RNA-seq reads and assembling the alignments into a parsimonious set of transcripts. It was based on a probabilistic model that provided a maximum likelihood explanation of the expression data at a given locus [17].

(3) Integration of the transcripts. We then integrated constructed transcripts using Cuffcompare [16] with all annotation resources, including RefSeq, the UCSC annotation for humans, and the Ensembl database and lncRNA transcripts from GENCODE Version 11. The integration helped us to determine a unique set of isoforms for each transcript locus and reduced the redundant calculation of FPKM (fragments per kilobase of transcript per million mapped reads).

1.2.2 Filtering of background transcripts

This module calculated expression values (i.e., FPKM) for

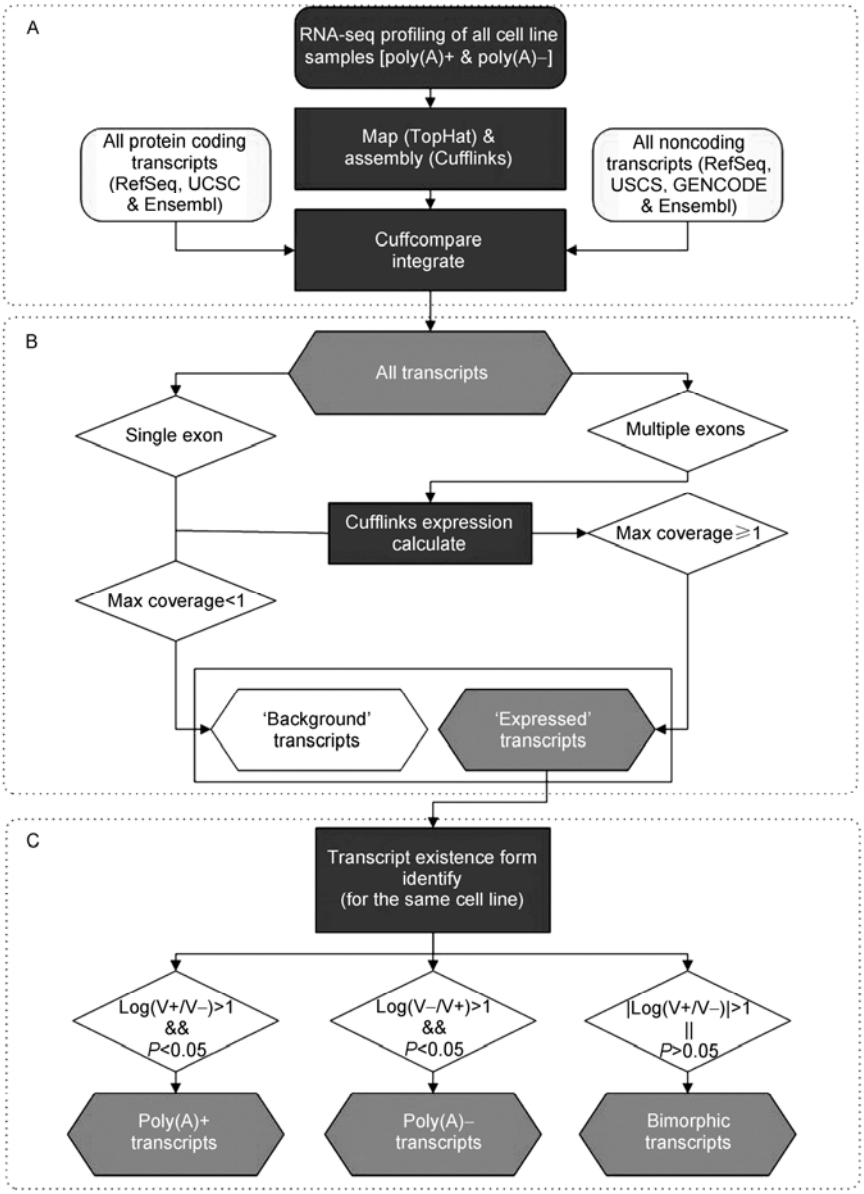


Figure 1 Computational workflow for the identification of different forms of gene transcripts in cells. A, Graphical representation of the generation and integration of transcripts via bioinformatics analyses. Transcripts were reconstructed from Cufflinks, and all coding genes and non-coding RNAs were determined from RefSeq, UCSC, Ensembl and GENCODE to integrate into 'all transcripts' using Cuffcompare. B, Screening of the expressed transcripts from the integrated transcripts. Lower coverage with one read and single exon transcripts were filtered as 'background transcripts'; transcripts on the left are considered 'expressed transcripts' in these samples. C, Screening of different forms of transcripts. V+ means FPKM value in poly(A)+ sample, V– means FPKM value in poly(A)– sample.

the integrated transcript set in a given sample and filtered out low expression transcripts (Figure 1B).

(1) Calculation of transcript expression levels. We calculated expression values (FPKM) for the integrated transcript set in each sample using the same computational process, because Cufflinks [16] uses a random algorithm in the program and offers different FPKM values for different transcript sets and computational processes.

(2) Filtering of low expression transcripts. One challenge of this workflow was how to distinguish lncRNAs expressed at a low level from unreliable transcripts assembled from RNA-seq data [18]. Transcripts with a maximal coverage in all eight samples below 1 were eliminated. The outputs from this step were termed 'expressed transcripts' (Figure 1B).

1.2.3 Identification of poly(A)+, poly(A)–, and bimorphic transcripts

This module classified all expressed transcripts into three subgroups, i.e., poly(A)+, poly(A)–, or bimorphic [11], according to their relative abundance using FPKM values for each gene in the poly(A)+ and poly(A)– samples from the

given cell line (Figure 1C).

We then recalculated the merged transcripts' relative abundance (i.e., FPKM), their differential expression in two corresponding samples, and Poly(A)+ and Poly(A)− levels in the same cell line using Cuffdiff [16]. Poly(A)+ predominant transcripts ('poly(A)+ transcripts' for short) were defined as those expressed transcripts with $P<0.05$ and at least two-fold greater expression enrichment from the poly(A)+ library compared with the poly(A)− library. In contrast, poly(A)− predominant transcripts ('poly(A)− transcripts') were defined as those expressed transcripts with $P<0.05$ and at least two-fold greater enrichment from the poly(A)− library compared with the poly(A)+ library. Bimorphic predominant transcripts ('bimorphic transcripts') were defined as those expressed transcripts with $P>0.05$ or less than two-fold relative expression between the poly(A)+ and poly(A)− libraries.

2 Results

2.1 Overview of transcriptome assembly

The RNA-seq data contained 808.7 million reads, and approximately 80% were mapped to the human genome. Using Cufflinks [16], 929754 transcripts were *ab initio* constructed, 82.1% of which (762941) were from poly(A)− samples and 17.9% (166813) from poly(A)+ samples. Among the poly(A)− transcripts, 78.3% (597441) were from cancer cell lines (i.e., MCF7 and A549). Most of the assembled transcripts (64.2%) were from poly(A)− samples of cancer cell lines (Table 1). Therefore, based on the numbers of exons, we divided the transcripts into two types: multiple exon transcripts and single exon transcripts. For the assembled transcripts with multiple exons, 51.9% (96072) of them were assembled from poly(A)+ samples and 48.1% (89192) were from poly(A)− samples. Hence, poly(A)+ samples represented more multiple exon transcripts (an average of 60.6%), especially for the two normal cell lines (an average of 73.0%), compared with that of poly(A)− samples (average of 16.0%), particularly for samples from cancer cell lines (average of 9.2%). For the single exon transcripts, poly(A)− samples represented significantly more one exon transcripts (an average of 84.1%), especially for the cancer cell line samples (an average of 90.8%), in comparison with poly(A)+ samples (an average of 39.4%). In addition, the average length of transcripts reconstructed from poly(A)+ samples was longer than that of transcripts reconstructed from poly(A)− samples. The length distribution of transcripts reconstructed from the eight samples of the four cell lines can be seen in Figure S1.

Subsequently, we compared the multiple exon transcripts from each sample to the 33243 protein coding transcripts from the RefSeq data using Cuffcompare [16]. Approximately 66.6% of the transcripts overlapped with exons of known coding transcripts. The remaining transcripts belonged to known non-coding RNAs or novel non-coding RNAs. Thus, these sequencing data represented the expression of all genes in the cells well.

2.2 Classification of poly(A)+, poly(A)− and bimorphic predominant transcripts in four cell lines

We designed a pipeline (Figure 1) to classify the transcripts into poly(A)+, poly(A)− and bimorphic transcripts, according to their different expression patterns in poly(A)+ and poly(A)− samples from cancer cells. For the coding transcripts filtered from all transcripts that were integrated using annotation of RefSeq, UCSC, Ensembl and GENCODE Version 11 databases and including all protein coding transcripts and long non-coding transcripts, most (44.7%) were bimorphic transcripts, i.e., 37.5% were poly(A)+ transcripts and 17.8% were poly(A)− transcripts (Figure S2). However, taking RefSeq protein coding transcripts as an example, there were different compositions of poly(A)+, poly(A)− and bimorphic transcripts, as shown in Figure 2. For example, 15841 and 16777 expressed RefSeq protein coding transcripts were screened from four human breast samples and four lung cell lines samples, respectively (File S1). The protein coding poly(A)+ transcripts were the most abundant in HMEC, MCF7, and NHLF cell lines, with an average of 52.3%; however, bimorphic transcripts accounted for the greatest ratio in the HMEC cell line with 44.7%. As ex-

Table 1 Assembly data of the eight samples and comparison with the RefSeq annotated gene set

Sample name	Assembly results	Multiple exons	Multiple exon ratio (%)	Exonic overlap with RefSeq	Annotation ratio (%) of multiple exons
HMEC-Poly(A)+	29326	22274	76	16078	72.1
HMEC-Poly(A)−	82360	16727	20.3	10885	65.1
MCF-7-Poly(A)+	55854	24971	44.7	16716	66.9
MCF-7-Poly(A)−	379206	26829	7.1	16276	60.1
NHLF-Poly(A)+	35416	24455	69.1	17807	72.8
NHLF-Poly(A)−	83140	20886	25.1	12960	62.1
A549-Poly(A)+	46217	24372	52.7	16713	68.6
A549-Poly(A)−	218235	24750	11.3	16146	65.2

pected, poly(A)− transcripts were the least abundant, with an average of 10% in these four cell lines, especially in the HMEC cell line, where they represented only 3.5% (Figure 2).

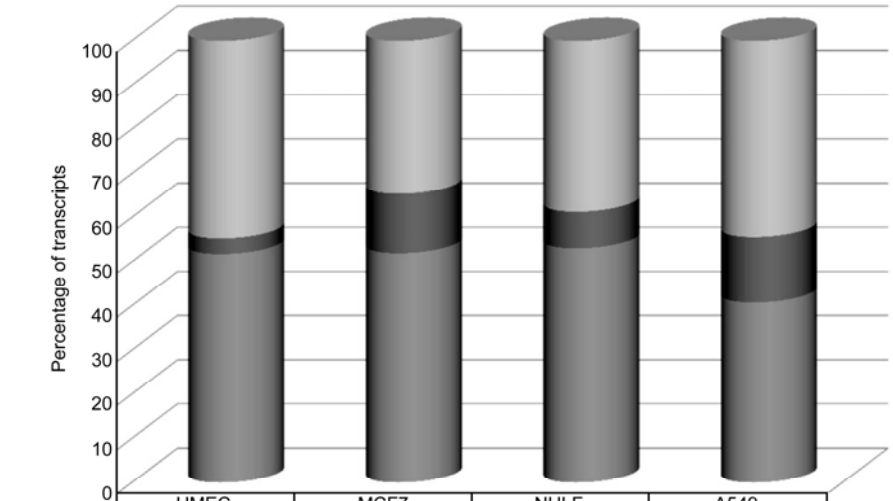
To further characterize the basic features of the different forms of protein coding transcripts, we compared the sequence characteristics and expression patterns among the three types of transcripts (Figure 2). The three types had the same average numbers of exons (11 exons) but different average lengths, i.e., 3.8 kb for poly(A)+ and 4.7 kb for poly(A)−, whereas bimorphic transcripts averaged 4.2 kb (Figure 3). In terms of their expression patterns, the expression levels of poly(A)+ protein-coding transcripts were greater than those of poly(A)− and bimorphic transcripts in these four cell lines (Figure 3).

2.3 Differentially expressed poly(A)+, poly(A)−, and bimorphic genes among these four cell lines

To obtain differentially expressed transcripts (DETs), we first obtained the poly(A)+ and poly(A)− transcripts from the interactions of tumor and normal samples (Figure S3, File S2). We then analyzed the differential expressions of these poly(A)+ or poly(A)− transcripts between tumor and normal cell lines using Cuffdiff [16] (gene list in File S3). Using two-fold greater expression enrichment and a *P*-value of less than 0.05 as a cutoff point, we found that the number of upregulated and downregulated transcripts (tumor vs. normal) for poly(A)+ DETs of lung tissue were 1043 and 1172, respectively. In poly(A)+ samples of breast cells, we found 1422 upregulated and 1188 downregulated poly(A)+ DETs (Table 2). Compared with poly(A)+ DETs, there were fewer poly(A)− DETs in tumor vs. normal cell lines, i.e., there were 45 upregulated and 192 downregulated poly(A)− DETs between lung cancer and normal lung cell lines. In addition, there were 24 upregulated and 37 downregulated poly(A)− DETs observed in poly(A)− samples of breast cancer vs. normal cells (Table 2).

2.4 Potential functions of these differentially expressed poly(A)+ and poly(A)− genes between tumor and normal cell lines

We then used the David bioinformatics tool with gene ontology (GO) functional terms to predict the functions of these different coding transcript sets and presented the data using GO-FAT biological process terms. We found that the functions of upregulated poly(A)− coding genes in lung cancer cells compared with normal cells were mainly associated with gene translation elongation, protein amino acid dephosphorylation, translation, the enzyme-linked receptor protein signaling pathway, the transmembrane receptor protein serine/threonine kinase signaling pathway, and the BMP (bone morphogenetic protein) signaling pathway (P<0.01, Figure 4). For upregulated poly(A)− coding genes in breast cancer cells compared with normal cells, the data showed the same functions, e.g., gene translation elongation and translation. Some of the downregulated poly(A)-coding



	HMEC RefSeq Total: 15841 (%)	MCF7 RefSeq Total: 15841 (%)	NHLF RefSeq Total: 16777 (%)	A549 RefSeq Total: 16777 (%)
Bimorphic	44.7	34.6	38.7	44.4
Poly(A)−	3.5	13.4	8.2	14.8
Poly(A)+	51.8	52	53.1	40.8

Figure 2 Three different forms of the expressed RefSeq transcripts in the four cell lines. Expression levels of the expressed transcripts were analyzed in four cell lines using FPKM values. Three forms of gene transcripts for each cell line were obtained according to their relative FPKM values from each gene compared with the poly(A)+ vs. poly(A)− samples of the same cell line.

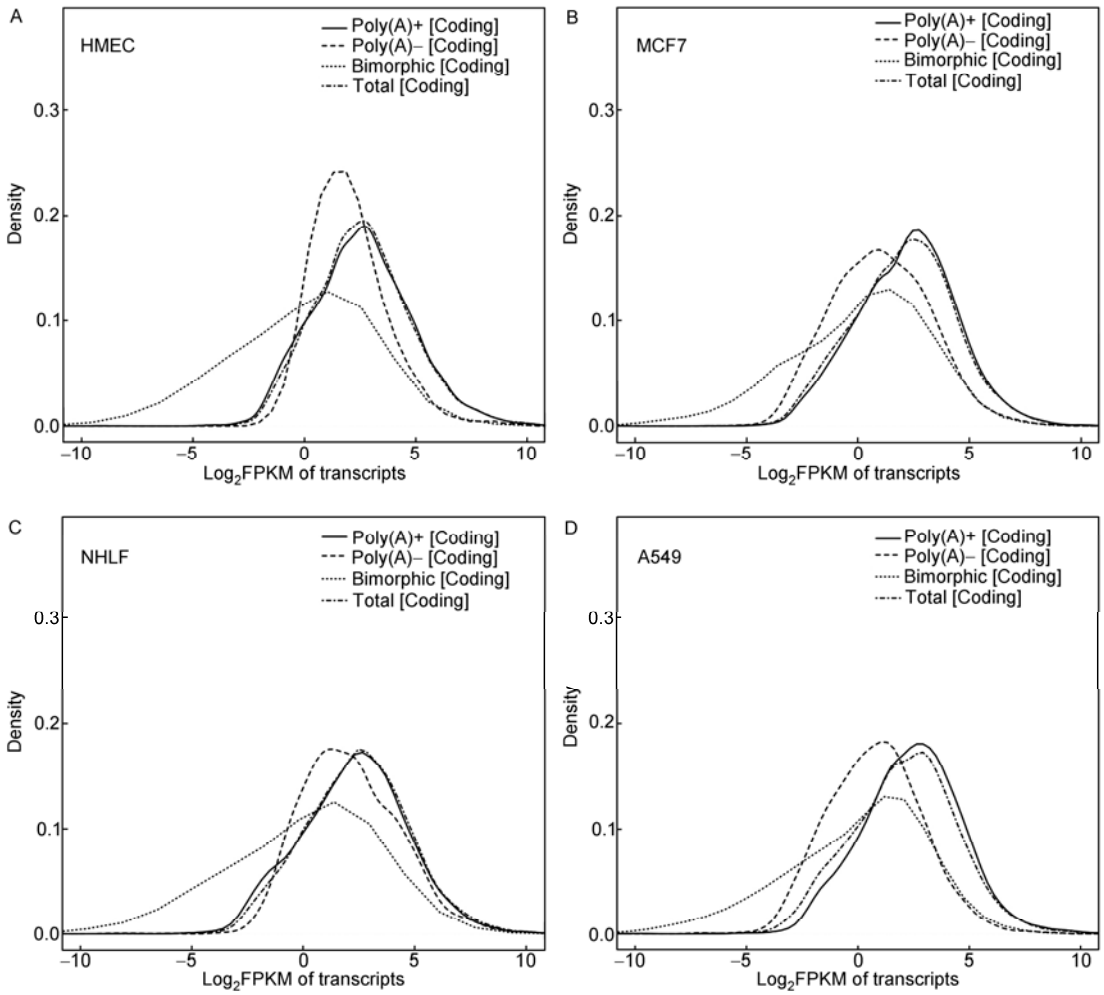


Figure 3 Expression patterns of the three forms of protein-coding transcripts in four cell lines. The expression levels of the three forms of the expressed RefSeq transcripts were compared in four cell lines using $\log_2$FPKM values (Figure 2).

Table 2 Differentially expressed poly(A)+ and poly(A)– transcripts between normal and cancer cell lines[a)]

	Lung	Breast
Poly(A)+ (↑)	1043	1422
Poly(A)+ (↓)	1172	1188
Poly(A)– (↑)	45	24
Poly(A)– (↓)	192	37

a) ↑, upregulated transcripts of tumor vs. normal cells; ↓, downregulated transcripts of tumor vs. normal cells. Details of these transcripts are listed in File S3 in Supporting Information.

genes in cancer cells appear to have functions related to cell death and the cell cycle in these two pairs of cell lines (Figure S4).

Furthermore, the functions of the poly(A)+ coding genes were mostly associated with the cell cycle, apoptosis, and cell death in cancer cell lines. These genes are reported to be highly expressed in cancer cells [19] (Figures S5 and S6).

In addition, the poly(A) state of the expressed transcripts was altered between cancer and normal cells. For example, poly(A)+ transcripts were converted into poly(A)– during carcinogenesis. The number of transcripts converted was approximately 460 and 476 in lung and breast cancer cells, respectively (gene list in File S4). Similarly, some poly(A)– transcripts in normal cells displayed a poly(A)+ characteristic in cancer cells. We found 153 and 223 genes with such an alteration in lung and breast cancer cells, respectively (gene list in File S4). Furthermore, functional enrichment analysis demonstrated that the functions of genes with mutual conversion of poly(A)+ to poly(A)– in normal and cancer cells were mainly associated with cell cycle progression and apoptosis (Figure 5), whereas the functions of genes with mutual conversion of poly(A)– to poly(A)+ were negative regulators of molecular function and RNA localization in normal and cancer cells (Figure S7).

3 Discussion

We developed an integrative method to reconstruct transcripts from eight lung and breast cell lines samples. Our

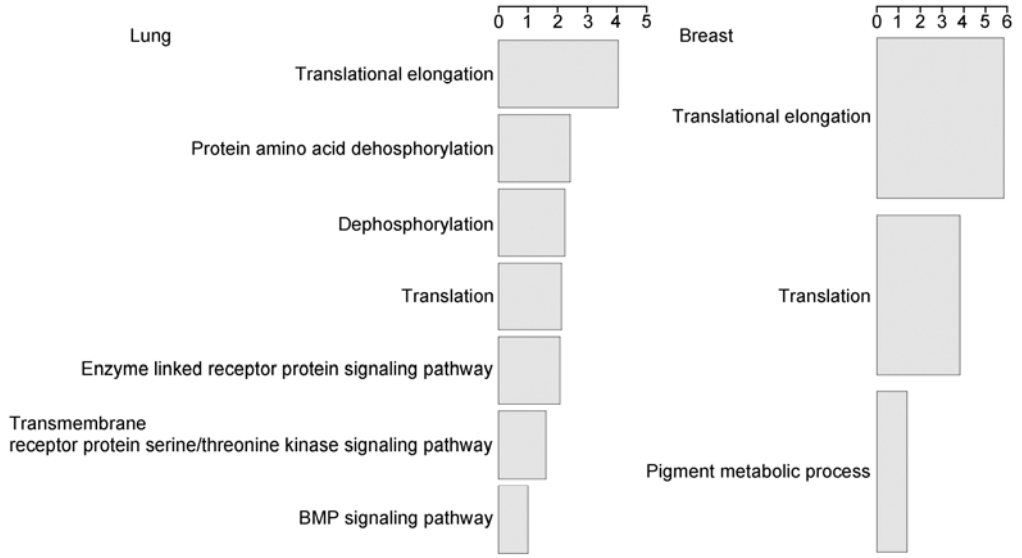


Figure 4 GO analysis of the upregulated poly(A)– coding genes between normal lung and lung cancer cell lines (left) and between normal breast and breast cancer cell lines (right).

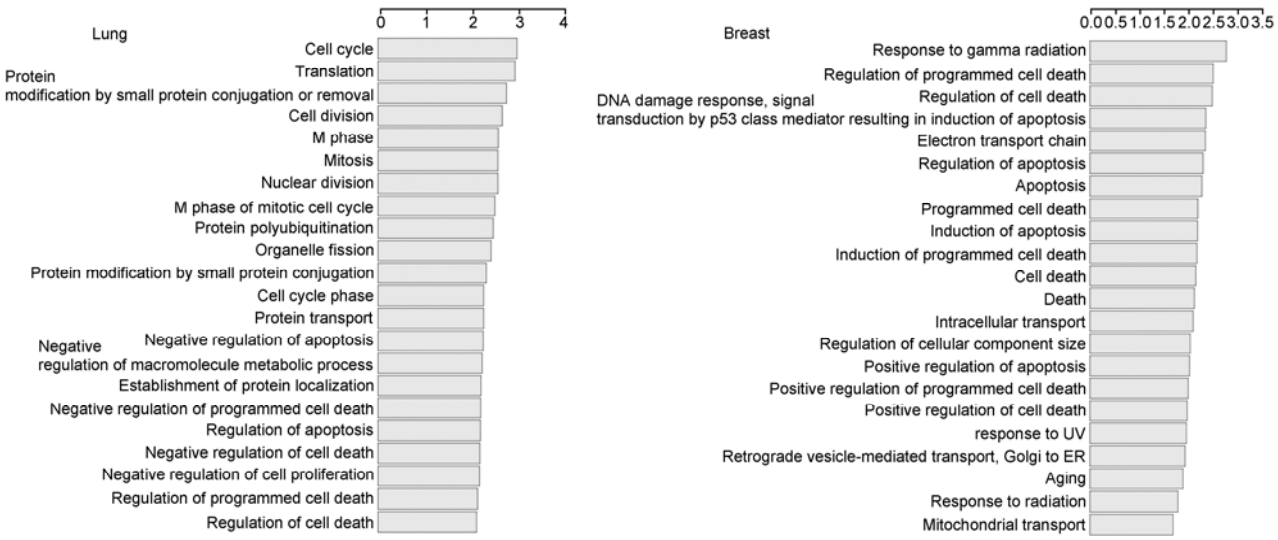


Figure 5 GO analysis of mutual conversion genes. Poly(A)+ in normal and poly(A)– forms of mRNA between normal and cancer cell lines. The lung cell line data are shown on the left, while the breast cell line data are shown on the right.

study provided several important observations: (i) the transcripts of some annotated coding genes were in the poly(A)– form; (ii) according to their relative abundance (FPKM values), all expressed transcripts could be divided into three subgroups: poly(A)+, poly(A)– and bimorphic; (iii) half (50.0% on average in these cell lines) of these expressed protein-coding transcripts were poly(A)+, while the other half were bimorphic (40%) or poly(A)– (10%); and (iv) GO enrichment analyses showed that poly(A)+ and some poly(A)– DETs were associated with cell cycle progression, DNA damage, and apoptosis-related genes, whereas poly(A)– DETs were also associated with gene translational elongation and translation functional categories, suggesting that poly(A)+ and poly(A)– transcripts play different roles in tumor development and progression. Thus, we show that poly(A)– transcripts may play an important role in cancer progression.

To the best of our knowledge, this is the first study to report differential expression and functions of poly(A)+ and poly(A)– genes between normal and tumor cell lines using high-throughput sequencing data. We found that the protein-coding genes with upregulated poly(A)+ transcripts were associated with DNA damage and the cell cycle in lung cancer (e.g., genes that respond to DNA damage stimuli, DNA repair, cell mitosis, and the M-phase of cell cycle). These genes are upregulated in cancer and many other dis-

eases [20,21], and the stimuli of gene upregulation causing DNA damage are usually from the environment, including ultraviolet or ionizing radiation, and various chemicals, or from inside cells [22]. Moreover, the regulation of cell cycle progression is altered during cancer development and progression [23,24]. However, for poly(A)− DETs, the functions of these genes are associated with protein translation elongation; for example, ribosomal protein SA (RPSA) causes protein translation elongation and is upregulated in lung cancer [25,26]. EIF4EBP1, a repressor of translation with a short poly(A) tail, is overexpressed in patients with lung cancer [27,28]. EIF4EBP1 directly interacts with eukaryotic translation initiation factor 4E (eIF4E) to inhibit protein translation complex assembly and repress translation. EIF4EBP1 is activated through phosphorylation by various signals (such as UV irradiation and insulin signaling), resulting in its dissociation from eIF4E and activation of mRNA translation [29]. EIF4EBP1 is also a key effector for the oncogenic activation of the AKT and ERK signaling pathways, which have integrated functions in different cancers [30].

The downregulated poly(A)− coding genes (Figure S3) are associated with cell death and cell cycle regulation, which are associated with tumorigenesis and tumor progression [19,31]. In addition, these mutual conversion genes mainly function to regulate the cell cycle, cell death, and apoptosis. Alterations of these genes frequently occur during tumorigenesis and cancer progression [19,32]. Poly(A) status can influence mRNA stability and translation efficiency; therefore, it is plausible that the transformation between poly(A)+ and poly(A)− may change the expression of a transcript and thus play an important role in cancer development and progression. The transcripts affected are involved in cell cycle regulation; therefore, it is conceivable that cell proliferation would be further influenced. Ultimately, tumorigenesis is triggered because of inappropriate cell proliferation [33]. Similarly, inappropriate apoptosis may cause tumorigenesis because apoptosis can serve as a natural barrier to cancer development [19]. Collectively, poly(A)− genes or an alterable poly(A) tails may influence the emergence and development of a tumor. Taken together, these results suggest that poly(A)− genes play an important role in cancer.

In addition, increasing numbers of non-coding RNAs have been identified that play a role in human cancers [34–36]. In the current study, we observed that certain non-coding RNAs were able to regulate gene expression. These non-coding RNAs were also present in three forms, i.e., poly(A)+, poly(A)− and bimorphic, and may play a role in cancer development. Indeed, previous studies have shown that p15AS (an antisense lncRNA) could silence expression of the p15 tumor suppressor gene in human leukemia [37], while lincRNA-p21, a repressor of p53-dependent transcriptional responses [38], and HOX antisense intergenic RNA (HOTAIR) play an active role in modulating the cancer epigenome and mediating cell transformation [39]. Furthermore, PCAT-1, a transcriptional repressor, has been implicated in a subset of prostate cancer patients [40]. Thus, the different forms of non-coding RNAs participate and influence tumorigenesis and tumor progression. Use of the coding-non-coding co-expression network to annotate the functions of the different forms of non-coding RNAs represents a novel tool for future research [41–44].

However, the methodology used in our current study presented some challenges. For example, we mainly used gene annotation after transcript alignment and assembly in the integrating transcripts step using Cuffcompare. To increase the precision of splice junction detection and transcript assembly, the gene annotation file for TopHat and Cufflinks is highly recommended by others. In our study, we performed *de novo* assembly without use of TopHat and Cufflink databases. Moreover, we did not use any of the annotation files from RefSeq, UCSC, Ensembl, or GENCODE as the reference annotation method for Cuffcompare to assist the process of transcript comparison and merging from different samples. Nonetheless, we did use Cuffcompare to remove the duplicate transcripts that overlapped in different databases. The principle is that Cuffcompare examines the structure of each transcript and matches transcripts that agree with the coordination and orders of all of their introns as well as the strand. Matched transcripts were allowed to differ on the length of the first and last exons, since these lengths will naturally vary from database to database of different sources. Thus, this approach could be useful to produce informative data. Another challenge is that it is impossible to assess the accuracy and sensitivity of the method because of the lack of "golden sets" for the three forms of genes to date. In a previous study [23], different proportions of the three forms of genes for different cell lines were observed, i.e., the form of a gene varies under different conditions. These results indicate that the analysis of accuracy and sensitivity is not practical for this kind of study. Overall, our current data demonstrated proof-of-principle; however, more studies are needed to investigate the functions of these three types of mRNAs in cancer and other human diseases.

This work was supported in part by the National Natural Science Foundation of China (31000564, 31071137, 91229120), the Beijing Natural Science Foundation (5122029), and the Knowledge Innovation Program of the Chinese Academy of Sciences (KSCX2-EW-R-01).

1 Wu Q, Kim Y C, Lu J, et al. Poly A-transcripts expressed in HeLa cells. PLoS ONE, 2008, 3: e2803
2 Cheng J, Kapranov P, Drenkow J, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science, 2005, 308: 1149–1154
3 Katinakis P, Slater A, Burdon R. Non-polyadenylated mRNAs from eukaryotes. FEBS Lett, 1980, 116: 1–7
4 Moore M J, Proudfoot N J. Pre-mRNA processing reaches back to transcription and ahead to translation. Cell, 2009, 136: 688–700
5 Grummt I. Regulation of mammalian ribosomal gene transcription by

RNA polymerase I. Prog Nucleic Acid Res Mol Biol, 1998, 62: 109–154
6 Detke S, Stein J L, Stein G S. Synthesis of histone messenger RNAs by RNA polymerase II in nuclei from S phase HeLa S3 cells. Nucleic Acids Res, 1978, 5: 1515–1528
7 Willis I M. RNA polymerase III. Genes, factors and transcriptional specificity. Eur J Biochem, 1993, 212: 1–11
8 Sunwoo H, Dinger M E, Wilusz J E, et al. MEN epsilon/beta nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. Genome Res, 2009, 19: 347–359
9 Wang E T, Sandberg R, Luo S, et al. Alternative isoform regulation in human tissue transcriptomes. Nature, 2008, 456: 470–476
10 Li J B, Levanon E Y, Yoon J K, et al. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science, 2009, 324: 1210–1213
11 Yang L, Duff M O, Graveley B R, et al. Genomewide characterization of non-polyadenylated RNAs. Genome Biol, 2011, 12: R16
12 Cheng J, Kapranov P, Drenkow J, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science, 2005, 308: 1149–1154
13 Rosenbloom K R, Dreszer T R, Long J C, et al. ENCODE whole-genome data in the UCSC Genome Browser: update 2012. Nucleic Acids Res, 2012, 40: D912–D917
14 Trapnell C, Pachter L, Salzberg S L. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 2009, 25: 1105–1111
15 Langmead B, Trapnell C, Pop M, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol, 2009, 10: R25
16 Trapnell C, Williams B A, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnol, 2010, 28: 511–515
17 Garber M, Grabherr M G, Guttman M, et al. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Meth, 2011, 8: 469–477
18 Cabili M N, Trapnell C, Goff L, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev, 2011, 25: 1915–1927
19 Hanahan D, Weinberg R A. Hallmarks of cancer: the next generation. Cell, 2011, 144: 646–674
20 Hoeijmakers J HJ. DNA damage, aging, and cancer. N Engl J Med, 2009, 361: 1475–1485
21 Gong X, Wu R H, Wang H W, et al. Evaluating the consistency of differential expression of microRNA detected in human cancers. Mol Cancer Ther, 2011, 10: 752–760
22 Jackson S P, Bartek J. The DNA-damage response in human biology and disease. Nature, 2009, 461: 1071–1078
23 Massagué J. G1 cell-cycle control and cancer. Nature, 2004, 432: 298–306
24 Evan G I, Vousden K H. Proliferation, cell cycle and apoptosis in cancer. Nature, 2001, 411: 342–348
25 Wang L S, Xiong Y Y, Sun Y H, et al. HLungDB: an integrated database of human lung cancer research. Nucleic Acids Res, 2010, 38: D665–D669
26 Landi M T, Dracheva T, Rotunno M, et al. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. PLoS ONE, 2008, 3: e1651
27 Rohrbeck A, Neukirchen J, Rosskopf M, et al. Gene expression profiling for molecular distinction and characterization of laser captured primary lung cancers. J Transl Med, 2008, 6: 69
28 Wrage M, Ruosaari S, Eijk P P, et al. Genomic profiles associated with early micrometastasis in lung cancer: relevance of 4q deletion. Clin Cancer Res, 2009, 15: 1566–1574
29 Barnhart B C, Lam J C, Young R M, et al. Effects of 4E-BP1 expression on hypoxic cell cycle inhibition and tumor cell proliferation and survival. Cancer Biol Ther, 2008, 7: 1441–1449
30 She Q B, Halilovic E, Ye Q, et al. 4E-BP1 is a key effector of the oncogenic activation of the AKT and ERK signaling pathways that integrates their function in tumors. Cancer Cell, 2010, 18: 39–51
31 Elmore S. Apoptosis: a review of programmed cell death. Toxicol Pathol, 2007, 35: 495–516
32 Zamai L, Ponti C, Mirandola P, et al. NK cells and cancer. J Immunol, 2007, 178: 4011–4016
33 Park M T, Lee S J. Cell cycle and cancer. J Biochem Mol Biol, 2003, 36: 60–65
34 Schadt E E, Linderman M D, Sorenson J, et al. Computational solutions to large-scale data management and analysis. Nat Rev Genet, 2010, 11: 647–657
35 Yang J, Yang F, Ren L, et al. Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. J Clin Microbiol, 2011, 49: 3463–3469
36 Sun L, Luo H, Liao Q, et al. Systematic study of human long intergenic non-coding RNAs and their impact on cancer. Sci China Life Sci, 2013, 56: 324–334
37 Yu W, Gius D, Onyango P, et al. Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. Nature, 2008, 451: 202–206
38 Huarte M, Guttman M, Feldser D, et al. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell, 2010, 142: 409–419
39 Esteller M. Non-coding RNAs in human disease. Nat Rev Genet, 2011, 12: 861–874
40 Prensner J R, Iyer M K, Balbin O A, et al. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat Biotechnol, 2011, 29: 742–749
41 Bu D C, Yu K T, Sun S, et al. NONCODE v3.0: integrative annotation of long noncoding RNAs. Nucleic Acids Res, 2012, 40: D210–D215
42 Guo X L, Gao L, Liao Q, et al. Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks. Nucleic Acids Res, 2013, 41: e35
43 Liao Q, Xiao H, Bu D C, et al. ncFANs: a web server for functional annotation of long non-coding RNAs. Nucleic Acids Res, 2011, 39: W118–W124
44 Liao Q, Liu C N, Yuan X Y, et al. Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. Nucleic Acids Res, 2011, 39: 3864–3878

Supporting Information

Figure S1 The length distribution of transcripts reconstructed from eight samples of these four cell lines.

Figure S2 The three forms of the expressed transcripts in these four cell lines.

Figure S3 Intersection of protein-coding transcripts between normal and cancer cell lines for poly(A)+ (top), poly(A)– (middle), and bimorphic (bottom) transcripts. HMEC, normal lung cell line; MCF-7, breast cancer cell line; NHLF, normal breast cell line; A549, lung cancer cell line.

Figure S4 GO analysis of the down-regulated poly(A)– coding genes between normal and lung cancer cell lines (left) as well as between normal and breast cancer cell lines (right).

Figure S5 GO analysis of the up-regulated poly(A)+ coding genes between normal and lung cancer cell lines (left) as well as between normal and breast cancer cell lines (right).

Figure S6 GO analysis of the down-regulated poly(A)+ coding genes between normal and lung cancer cell lines (left) as well as between normal and breast cancer cell lines (right).

Figure S7 GO analysis of the mutually converted genes as poly(A)– in normal and poly(A)+ in cancer of different forms between normal and cancer cell lines. Left, lung; right, breast. The "mutual conversion genes" indicate that they were originally poly(A)– genes in the normal cell line, whereas poly(A)+ genes were in the tumor cell line. After conversion, the opposite was true.

Table S1 Human paired-end RNA-seq data sets from ENCODE Cold Spring Harbor Labs

Table S2 Human reference genome sequences of hg19 from annotation resources

File S1 three_form_of_refseq_genes.xlsx

File S2 three_forms_intersection_between_normal_cancer.xlsx

File S3 lung_mammary.three_forms.xlsx

File S4 transformed_genes.xlsx

The supporting information is available online at life.scichina.com and www.springerlink.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.