High-throughput long noncoding RNA profiling for diagnostic and prognostic markers in cancer: opportunities and challenges

“The increasing discoveries will help to better understand the deadly disease and find more promising diagnostic and prognostic markers.”

Keywords: cancer • challenges • diagnostic marker • long noncoding RNA • outcome prediction • prognostic marker • treatment target

Long noncoding RNAs (lncRNAs) are a large class of gene transcripts discovered or better characterized in recent years, mostly as a result of high throughput technologies like RNA sequencing [1–3]. They account for roughly 68% of transcribed genes from about 70% of human genome with transcription activity [3,4]. The largely unexplored desert has attracted great interest in the research community and publications in this topic have grown dramatically for the past few years, particularly for cancer. The increasing discoveries will help to better understand the deadly disease and find more promising diagnostic and prognostic markers. On the other hand, as one of the most poorly understood and annotated classes of gene transcripts, we are facing many challenges that need to be adequately addressed so that novel and clinical translational discoveries can be made.

The unique features of lncRNAs make them more favorable for potential diagnostic and prognostic markers (or treatment targets); however, they are more complex and difficult to study than protein coding genes. lncRNAs do not code proteins but play regulatory roles. This regulatory function is often executed through sequence match to its targets, which can be blocked as a treatment target. lncRNAs are usually shorter than protein coding genes with mostly two to three exons [2]. Although most lncRNAs are polyadenylated so that they can be profiled through the common polyA enrichment-based RNA sequencing

methods, some without the polyadenylation would need non-polyA methods to be measured. Unlike mRNAs, which are transported into cytoplasm for protein translation, most lncRNAs are localized to the nucleus for their actions [2]. RNA extraction methods or RNAs isolated from which compartment of cells (cytoplasm vs nucleus) can make a significant difference in lncRNAs that are quantified and studied [5]. Finally, lncRNA expressions seem more tissue specific at lower expression levels [2], which makes them excellent differential diagnostic markers. However, many at very low expression are difficult to distinguish from detecting noise in gene expression microarray or need higher depth sequencing and may not be a good candidate for clinical testing.

According to their genomic position relative to adjacent protein-coding genes, lncRNAs are classified into: antisense RNAs, long intergenic noncoding RNAs (lincRNAs), sense overlapping intronic, sense nonoverlapping intronic and processed transcripts without ORF [2]. Except lincRNAs that do not overlap with protein-coding genes, others are intertwined with protein-coding genes and are difficult to distinguish from the overlapping genes without strand-specific RNA sequencing. In this case, lncRNA studies mostly focus on lincRNAs. The latest ENSEMBL (Version 80) and GENCODE (Version 22) annotations have 14,863 and 15,900 lncRNAs defined; however, vast majority are still putative with-

Zhifu Sun

Division of Biomedical Statistics & Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA
Tel.: +1 507-266-1894
Fax: +1 507-284-0360
sun.zhifu@mayo.edu

out a standard name approved by the HUGO Gene Nomenclature Committee (only 191 lncRNAs have). With defined guidelines [6,7] and more functional annotations, this will be expected to improve.

The genome-wise search for cancer lncRNA biomarkers is mainly carried out through gene-expression microarray, either lncRNA-specific array [8] or remapping of existing microarrays to limited noncoding genes [9,10], and RNA sequencing. While the former profiles known lncRNAs, the latter identifies many novel ones. The largest study so far mined over 7000 RNA sequencing libraries consisting of various tissue types and cancers and discovered over 7000 lineage- or cancer-associated lncRNAs, with *PCa3* and *SClLAPI*, for example, as specific markers for prostate cancer [4]. Furthermore, *SClLAPI* was demonstrated as a consistent predictor for metastasis or recurrence [11]. The detection of this lncRNA in urine can potentially make it a noninvasive biomarker for clinical practice. The promising findings in other cancer types have also been reported. Analysis of 567 lung adeno- and squamous-cell carcinomas identified a 100 of novel lincRNAs that were differentially regulated between tumors and normal lung tissues and '*LCAL1*' was found to be associated with tumor growth with prognostic value [12]. In profiling lncRNAs in over 200 patients with acute myeloid leukemia, lncRNAs were found to be associated with gene mutation status and to predict treatment response and patient survival [13]. A six-lncRNA panel in glioblastoma multiforme was reported to provide an independent survival prediction benefit in addition to age and *MGMT* promoter methylation status [10]. Several studies in colorectal cancer through lncRNA-specific microarray or mining the existing data for a small number of lncRNAs present on the array platforms found that a subset of lncRNAs may not only classify colorectal cancer subtypes with clinical relevance [14] but also have an improved prediction value for survival [9].

The impressive advances within a short time period are a big step forward for potential lncRNA biomarkers. However, we need to realize the significant challenges in this evolving and fast growing field. The tissue-specific expression and incomplete annotation of lncRNAs often justify novel prediction in a study. The process involves many complex bioinformatics steps and there is no standard way to discover and nominate novel lncRNAs. The transcriptome assembly is required and various programs like Cufflinks, Scripture or StringTie can be used but their performances can vary [15]. The defining feature of lncRNAs is lack of protein coding capability and the current prediction is mostly through machine learning or sequence feature examination using programs such as Coding

Potential Assessment Tool [16], iSeeRNA [17], Coding Potential Calculator [18] or checking the presence of a known Pfam domain. Even for well-characterized protein-coding RNAs and lncRNAs, there is a small percentage (up to 5%) of transcripts which often gets misclassified [16,17]. Whether to take lncRNA expression level or filter out low confidence or false mapping into consideration in nomination of a novel lncRNA also makes differences [19]. Furthermore, most RNA sequencing data are nonstrand specific and polyA selected. lncRNAs that overlap with protein coding are not easily investigated; lncRNAs that are not polyadenylated need to be captured by total RNA rRNA removal protocols. RNA extraction by the TRIzol method tends to capture more nuclear unprocessed transcripts and lncRNAs than the Qiagen RNeasy Mini Kit method [5]. These variations of sequencing and lncRNA prediction algorithms need to be factored in comparing and judging the results of significance across different studies. Independent validation of results becomes increasingly important in such a heterogeneous environment.

The functions of most lncRNAs are largely unknown and result interpretation from lncRNA studies is very challenging. lncRNAs may interact with mRNAs and modulate their translation either in a positive or negative way [20]. lncRNAs appear more positively correlated with the expression of antisense protein coding genes [2]. lincRNAs may act on their nearby protein-coding genes or remote genes. Such interactions may be predicted digitally [21,22] or from experimental data of binding or correlation between lncRNAs and proteins [23]. When interaction relationships lack, a common approach is to conduct correlative analysis with protein-coding genes genome-wide and then use the correlated protein-coding genes for pathway or network-enrichment analysis where canonical pathways or protein-protein interaction networks are well known. A caveat is with so many combinations between lncRNAs and protein-coding genes, it is inevitable to incorporate random correlations and true interactions need to be further validated.

A good alternative biomarker needs to pass several tests: it demonstrates a better performance than existing ones; it can be easily tested clinically; and it can be reliably tested. lncRNAs are essentially no different from protein-coding mRNAs as clinical biomarkers. Experience from gene-expression microarray or the markers derived from the technology for the past decade should provide us good guidance to find more promising markers [24,25]. There is no surprise that lncRNA expression patterns in tumors are distinct from their normal counterpart tissues as tumors change gene-expression programs dramatically. The fact that

the patterns may define molecular subtypes of tumors is commonly seen in the mRNA expression. The harder questions are whether the new classification is better than protein-coding expression patterns or existing classification methods. mRNAs and lncRNAs are likely correlated, which may be further correlated with histologic features or subtypes. Evaluation of molecular signatures needs to take those known prognostic factors into consideration so that the molecular predictors can provide an added clinical value [26,27]. Not many RNA-based biomarkers or panels have moved to clinical practice. Oncotype DX®, PAM50 and MammaPrint are the most promising gene signatures to predict breast cancer recurrence and guide adjuvant chemotherapy selection [28]. However, their clear clinical benefits are still being evaluated through large clinical trials.

The financial constraint limits most studies to a small scale that lacks a sufficient power for an out-

come association study. Incomplete clinical phenotypic information for some large public data repositories may hinder the identification of a good prognostic marker or treatment target. Rigorous evaluation of such markers is needed before these candidates can be moved to further development as clinical markers. With more attention and resources invested, promising findings from lncRNA research are only a matter of time.

Financial & competing interests disclosure

This work is partially supported by the independent research funds from the Mayo Clinic Center for Individualized Medicine. The author has no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

References

- Jia H, Osak M, Bogu GK, Stanton LW, Johnson R, Lipovich L. Genome-wide computational identification and manual annotation of human long noncoding RNA genes. *RNA* 16(8), 1478–1487 (2010).
- Derrien T, Johnson R, Bussotti G *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* 22(9), 1775–1789 (2012).
- Djebali S, Davis CA, Merkel A *et al.* Landscape of transcription in human cells. *Nature* 489(7414), 101–108 (2012).
- Iyer MK, Niknafs YS, Malik R *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* 47(3), 199–208 (2015).
- Sultan M, Amstislavskiy V, Risch T *et al.* Influence of RNA extraction methods and library selection schemes on RNA-seq data. *BMC Genomics* 15, 675 (2014).
- Wright MW. A short guide to long non-coding RNA gene nomenclature. *Hum. Genomics* 8, 7 (2014).
- Mattick JS, Rinn JL. Discovery and annotation of long noncoding RNAs. *Nat. Struct. Mol. Biol.* 22(1), 5–7 (2015).
- Xue Y, Ma G, Gu D *et al.* Genome-wide analysis of long noncoding RNA signature in human colorectal cancer. *Gene* 556(2), 227–234 (2015).
- Hu Y, Chen HY, Yu CY *et al.* A long non-coding RNA signature to improve prognosis prediction of colorectal cancer. *Oncotarget* 5(8), 2230–2242 (2014).
- Zhang XQ, Sun S, Lam KF *et al.* A long non-coding RNA signature in glioblastoma multiforme predicts survival. *Neurobiol. Dis.* 58, 123–131 (2013).
- Prensner JR, Zhao S, Erho N *et al.* RNA biomarkers associated with metastatic progression in prostate cancer: a multi-institutional high-throughput analysis of SchLAP1. *Lancet Oncol.* 15(13), 1469–1480 (2014).
- White NM, Cabanski CR, Silva-Fisher JM, Dang HX, Govindan R, Maher CA. Transcriptome sequencing reveals altered long intergenic non-coding RNAs in lung cancer. *Genome Biol.* 15(8), 429 (2014).
- Garzon R, Volinia S, Papaioannou D *et al.* Expression and prognostic impact of lncRNAs in acute myeloid leukemia. *Proc. Natl Acad. Sci. USA* 111(52), 18679–18684 (2014).
- Chen H, Xu J, Hong J, Tang R, Zhang X, Fang JY. Long noncoding RNA profiles identify five distinct molecular subtypes of colorectal cancer with clinical relevance. *Mol. Oncol.* 8(8), 1393–1403 (2014).
- Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33(3), 290–295 (2015).
- Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* 41(6), e74 (2013).
- Sun K, Chen X, Jiang P, Song X, Wang H, Sun H. iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC Genomics* 14(Suppl. 2), S7 (2013).
- Kong L, Zhang Y, Ye ZQ *et al.* CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* 35, W345–W349 (2007).
- Sun K, Zhao Y, Wang H, Sun H. Sebnif: an integrated bioinformatics pipeline for the identification of novel large intergenic noncoding RNAs (lincRNAs) – application in human skeletal muscle cells. *PLoS ONE* 9(1), e84500 (2014).
- Gong C, Maquat LE. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature* 470(7333), 284–288 (2011).
- Lu Q, Ren S, Lu M *et al.* Computational prediction of associations between long non-coding RNAs and proteins. *BMC Genomics* 14, 651 (2013).

- 22 Bellucci M, Agostini F, Masin M, Tartaglia GG. Predicting protein associations with long noncoding RNAs. *Nat. Methods* 8(6), 444–445 (2011).
- 23 Li JH, Liu S, Zheng LL *et al.* Discovery of protein-lncRNA interactions by integrating large-scale CLIP-Seq and RNA-Seq datasets. *Front. Bioeng. Biotechnol.* 2, 88 (2014).
- 24 Simon R. Analysis of DNA microarray expression data. *Best Pract. Res. Clin. Haematol.* 22(2), 271–282 (2009).
- 25 Cheng J, Greshock J, Shi L, Zheng S, Menius A, Lee K. Good practice guidelines for biomarker discovery from array data: a case study for breast cancer prognosis. *BMC Syst. Biol.* 7(Suppl. 4), S2 (2013).
- 26 Sun Z, Yang P. Gene expression profiling on lung cancer outcome prediction: present clinical value and future premise. *Cancer Epidemiol. Biomarkers Prev.* 15(11), 2063–2068 (2006).
- 27 Yang P, Sun Z. Gene-expression profiling in lung cancer: still early days. *Pharmacogenomics* 8(2), 129–132 (2007).
- 28 Goncalves R, Bose R. Using multigene tests to select treatment for early-stage breast cancer. *J. Natl Compr. Canc. Netw.* 11(2), 174–182 (2013).