EBioMedicine 35 (2018) 369–380

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

EBioMedicine
Published by THE LANCET

Research paper

Capturing functional long non-coding RNAs through integrating large-scale causal relations from gene perturbation experiments

Jinyuan Xu [a,1], Aiai Shi [a,1], Zhilin Long [a,1], Liwen Xu [a,1], Gaoming Liao [a], Chunyu Deng [a], Min Yan [a], Aiming Xie [a], Tao Luo [a], Jian Huang [c], Yun Xiao [a,b], Xia Li [a,b,*]

[a] *College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, Heilongjiang 150081, China*
[b] *Key Laboratory of Cardiovascular Medicine Research, Harbin Medical University, Harbin, Heilongjiang 150086, China*
[c] *Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 611731, China*

ARTICLE INFO

Article history:
Received 6 July 2018
Received in revised form 15 August 2018
Accepted 22 August 2018
Available online 1 September 2018

Keywords:
lncRNA
Function
Gene perturbation
Cancer
Immunotherapy

ABSTRACT

Characterizing functions of long noncoding RNAs (lncRNAs) remains a major challenge, mostly due to the lack of lncRNA-involved regulatory relationships. A wide array of genome-wide expression profiles generated by gene perturbation have been widely used to capture causal links between perturbed genes and response genes. Through annotating >600 gene perturbation profiles, over 354,000 causal relationships between perturbed genes and lncRNAs were identified. This large-scale resource of causal relations inspired us to develop a novel computational approach LnCAR for inferring lncRNAs' functions, which showed a higher accuracy than the co-expression based approach. By application of LnCAR to the cancer hallmark processes, we identified 38 lncRNAs involved in distinct carcinogenic processes. The "activating invasion & metastasis" related lncRNAs were strongly associated with metastatic progression in various cancer types and could act as a predictor of cancer metastasis. Meanwhile, the "evading immune destruction" related lncRNAs showed significant associations with immune infiltration of various immune cells and, importantly, can predict response to anti-PD-1 immunotherapy, suggesting their potential roles as biomarkers for immune therapy. Taken together, our approach provides a novel way to systematically reveal functions of lncRNAs, which will be helpful for further experimental exploration and clinical translational research of lncRNAs.

1. Introduction

Over the past decade, advances in sequencing technologies have detected large amounts of long non-coding RNAs (lncRNAs) in the human genome. Some lncRNAs have been proved to play key roles in transcriptional regulation, chromatin modification, cell differentiation and immune responses [25,28,37,50]. In human diseases, particularly cancer, a large number of genetic variations in non-coding regions, including lncRNAs, are discovered and some are highly correlated with pathogenesis of the disease [18,81]. Although lncRNAs have been proved to exert important effects in numerous diseases, however, little is known about the functions of most lncRNAs, which seriously obstructs our further understanding of their dysfunctional mechanisms underlying complex disease [81]. Therefore, the functional characterization of lncRNAs is a key fundamental challenge in the field of lncRNA biology.

Although perturbation experiments, such as knockdown or overexpression, have been developed or adapted to properly yield biological insights into lncRNAs, they are not suited to study such extensive pool of candidates [5,67]. Even though large-scale RNAi screens have been successful in investigating hundreds of functional lncRNAs in specific biological processes, they are frequently not efficient to reduce lncRNA levels and may suffer substantial off-target effects [10,25,40]. Recently, CRISPR/Cas9-based screening strategies, which are highly dependent on specific genome editing and can effectively reduce off-target effects [8], are developed to identify lncRNAs required for cellular growth [45,90]. However, the Cas9-mediated approaches are shown to be unsuitable to target lncRNAs that are positioned in close proximity to other genes, making it difficult to obtain biologically significant results for a majority of lncRNAs [21].

Considering the fact that experimental methods can be time-consuming and laborious, the computational methodologies are proposed to systematically infer lncRNA functions based on various types of biological data, such as sequence conservation [54] and RNA

[*] Corresponding author at: College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, Heilongjiang 150081, China.

E-mail addresses: hj@uestc.edu.cn (J. Huang), xiaoyun@ems.hrbmu.edu.cn (Y. Xiao), lixia@hrbmu.edu.cn (X. Li).

[1] These authors contributed equally to this work as first author

https://doi.org/10.1016/j.ebiom.2018.08.050
2352-3964/

Research in context

Evidence before this study

In the past decade, a large number of perturbation-based expression data were produced, which capture a wide range of causal relations between perturbed genes and response genes for revealing the biological functions. We collected gene expression profiles of single-gene perturbation experiments from the Gene Expression Omnibus (GEO), the ENCODE project and the gene perturbation atlas (GPA). Perturbation experiments from the GEO were curated before June 2016 by using the keywords 'knock out', 'knock down', 'RNAi', 'knock in', 'overexpression', 'high expression', 'low expression', 'siRNA' or 'shRNA'. By assessing the reannotation performance for every microarray platforms, we observed that the majority of platforms had a few re-annotated lncRNAs and only two platforms including Affymetrix Human Genome U133 Plus 2.0 Array (GPL570) and Affymetrix Human Exon 1.0 ST Array (GPL5175) had enough lncRNAs, which were thus selected for the following analysis. Moreover, perturbations with less than two experimental samples were filtered. In addition, considering the influence of other biochemical factors, experiments with additional treatments, such as dexamethasone, hypoxia and insulin, were also filtered.

Added value of this study

We constructed a comprehensive resource of causal relations of lncRNAs and protein-coding genes and proposed a novel approach, named LnCAR, to capture functions of lncRNAs based on a resource of causal relations from a large scale gene perturbation profiles. LnCAR is a robust and flexible approach for identifying lncRNAs related to any function of interest. To facilitate the convenient use of our approach to infer lncRNAs' functions, we also developed an online tool at the following URL: http://biocc.hrbmu.edu.cn/LnCAR/.

Implications of all the available evidence

In this study, we not only inferred cancer-related lncRNAs but also identified lncRNAs involved in each cancer hallmark, allowing to look into the underlying mechanisms of lncRNAs in cancer progression. Our method can effectively identify lncRNAs of tumor metastasis and demonstrate their potential as biomarkers for diagnosis and prognosis of cancer progression. For lncRNAs involved in "evading immune destruction", we not only found their significant association with different immune cell infiltration but also with some chemokines/receptors in various cancer types, highlighting their important roles in immune response. Furthermore, two lncRNAs, RP11-705C15.3 and SNHG5, were found to be highly correlated with response to anti-PD-1 immunotherapy, and be correlated with patient survival and better stratification, showing their strong potential as novel clinical predictors for immunotherapy response.

secondary structure predicted from DNA sequences [68] as well as transcriptomic [24] and epigenomic data [88]. Undoubtedly, the most commonly used data is gene expression profile, by means of which the guilt-by-association approach that assumes genes with similar expression patterns should share common biological functions or pathways, is widely used to predict lncRNA functions [43,67]. Following the identification of co-expression relationships based on the guilt-by-association principle, genome-wide clustering methods and network-based approaches can be intuitively used to organize transcriptome data and reveal lncRNA functions [9,85,86]. However, it should be noted that these guilt-by-association methods are suffering many false positive functional associations predicted from co-expression relations, thus affecting the performance of predicting lncRNA functions [63].

In the past decade, a large number of perturbation-based expression data were produced, which capture a wide range of causal relations between perturbed genes and response genes for revealing the biological functions of the perturbed genes [36,56,83]. Importantly, these causal relationships that reflect the influence of perturbed genes on downstream genes, when compared with co-expression-based relationships, provide a more reliable and more direct way to build functional links. Moreover, it is intuitively reasonable that these causal relationships allow bridging known functional genes to many functionally uncharacterized genes, especially largely uncharacterized lncRNAs. With the wide application of RNA sequencing technology and the development of re-annotation approaches for microarray data, lncRNAs can be properly detected in a large number of gene expression datasets that are publicly available in gene expression repositories, thus making it possible to detect lncRNA-associated causal relations.

In this study, we presented a novel approach, called LnCAR, to infer functional lncRNAs based on the causal relations inherent to gene perturbation experiments. By using LnCAR, we successfully captured lncRNAs involved in cell cycle and tumorigenesis processes and proved the reliability and accuracy of the approach. We further showed that lncRNAs that were inferred to be associated with tumor metastasis and immune response can be served as potential markers for clinical diagnosis of metastatic cancer and response to immunotherapy, respectively.

2. Material and methods

2.1. Gene perturbation resource

We collected gene expression profiles of single-gene perturbation experiments from the Gene Expression Omnibus (GEO), the ENCODE project and the gene perturbation atlas (GPA). Perturbation experiments from the GEO were curated before June 2016 by using the keywords 'knock out', 'knock down', 'RNAi', 'knock in', 'overexpression', 'high expression', 'low expression', 'siRNA' or 'shRNA'. We downloaded the original microarray datasets from GEO, and processed the data by re-annotating lncRNAs using a custom pipeline [89]. By assessing the reannotation performance for every microarray platforms, we observed that the majority of platforms had a few re-annotated lncRNAs and only two platforms including Affymetrix Human Genome U133 Plus 2.0 Array (GPL570) and Affymetrix Human Exon 1.0 ST Array (GPL5175) had enough lncRNAs [17], which were thus selected for the following analysis. Moreover, perturbations with less than two experimental samples were filtered. In addition, considering the influence of other biochemical factors, experiments with additional treatments, such as dexamethasone, hypoxia and insulin, were also filtered. RNA-seq datasets treated with shRNA knockdown and generated using strand-specific transcriptome sequencing were obtained from the ENCODE project and the gene quantifications generated by the ENCODE processing pipeline were used to construct transcriptional profiles.

2.2. Data preparation

The protein-coding gene and lncRNA annotations were obtained from the UCSC Gene track and GENCODE (v19), respectively. For microarray data, we used a custom pipeline to re-annotate the probe sequences provided by Affymetrix (http://www.affymetrix.com) to thousands of lncRNAs according to our previous study [89]. This results in 15,692 protein-coding genes and 2673 lncRNAs for GPL570 and 18,376 and 10,092 for GPL5175, respectively. The raw data were normalized using the RMA normalization method and gene expression variations between perturbation and control were obtained by Student's *t*-test analysis. For RNA-seq data, we first unified gene identities of protein-coding genes and lncRNAs to our annotations and filtered genes without or with multiple IDs. According to the official instructions of the GENCODE, only genes annotated with "3prime_overlapping_ncrna", "antisense", "lincRNA", "processed_transcript", "sense_intronic" and "sense_overlapping" are regarded as lncRNAs. Considering the low consistency of less expressed genes between RNA-seq and microarray, genes expressed at least 2 read counts in 75% samples were retained. Raw count data with paired samples (knockdown and control) were analyzed using DESeq R package.

2.3. Generation of gene ranking lists

We first obtained expression fold change (FC) and statistical significance (P-value) from differential expression analysis of perturbation profiles. A significance score which combined FC and P-value was calculated to rank causal relations [84].

$$\pi_i = |\log_2(FC_i)|.(-\log_{10}p_i)$$

π-value is a non-negative index, and the larger it is, the more significant gene expression changes. For each perturbation experiment, we can get the ranked list of genes where expression is most affected by the perturbed gene.

2.4. LnCAR

The LnCAR approach provides a set of optimal lncRNAs as well as a prioritized list of genes associated with the function of interest. For a given function, we obtained the causal relations from our gene perturbation resource whose perturbed genes were in the function, and calculated π-value for each causal relation. Then, the ranked gene lists affected by the perturbed genes were generated and the genes annotated in all of the perturbation profiles (without perturbed genes) were chosen for the following analysis. The LnCAR procedure consists of two major steps: (1) rank aggregation and (2) selection of optimal lncRNAs.

We first employed a modified hybrid Bayesian-rank integration method (BIRRA) for aggregation of our gene rankings [4]. The rank-based method is proper to solve the problem of the heterogeneity among different high-throughput genomic experiments and the use of Bayesian framework allows us to weight the reliability of individual datasets. To address the requirement of a prior probability before aggregation, we integrated another method RRA to produce a positive class for the prior probability calculation (see Supplementary Information) [39]. According to the modified prior probability and the Bayes factor calculation, a confidence score for each gene would be obtained to represent how likely it influenced by the function. Then, a high-confidence ranking of genes was constructed according to their scores.

Next, we proposed a sliding window method to extract optimal lncRNAs that most associated with the function. We used 6 different bin sizes (20, 25, 30, 50, 75 and 100) to slide the ranking constructed above separately to determine the optimal position. For each bin, the association between the genes in the bin and the given function was identified by the network analysis based on the random walk with restart (RWR) algorithm (see Supplementary Information) [74]. We used genes in the bin and genes in that function as seeds to prioritize genes in the network (from the STRING database) separately and calculated an association score to reflect the association strength between the two gene sets. If the gene sets are in the top ranking area of each other, it will obtain a high association score. The association score can be defined as follows:

$$\text{median}\left(1-\frac{\text{Rank}(\text{GeneOf}(\text{Step}_i))}{\text{Totalgenes}}\right) \times \text{median}\left(1-\frac{\text{Rank}(\text{GeneOf}(\text{Function}))}{\text{Totalgenes}}\right)$$

We further applied a permutation test to assess the significance of each score. We simulated the ranking 1000 times, used the same bin sizes to calculate the association scores and evaluated the significant score for each bin. The significant results of the bins from all six scale types were used to generate a multiscale view and help precisely define the position (on the basis that the bins near the top should be continuously significant, see Supplementary Information). Finally, lncRNAs above the position were regarded as optimal lncRNAs.

3. Method comparison

The functional catalogs assigned to the lncRNAs and the lncRNAs recorded in each function were retrieved from Huarte [30]. To expand the repertoire of known functional lncRNAs, we also added lncRNAs recorded in the lncRNAdb database by searching each function in the database [61]. After filtering lncRNAs that were not adopted in the two approaches, six functional catalogs were used for comparative analysis, which included eight known lncRNAs. The co-expression based approach used Spearman correlation coefficients to calculate the correlations for each lncRNA–mRNA pair based on the gene expression data from the GTEx Portal (including 7497 samples from 36 different tissues). For each lncRNA, protein-coding genes were ranked according to their correlation coefficients and gene set enrichment analysis (GSEA) was used to identify significantly enriched functions based on the GO terms with more than fifteen gene members and FDR < 0.25 [24]. The lncRNAs significantly enriched in the six functions were assigned to each function and compared to the known lncRNAs and the lncRNAs captured by LnCAR in the corresponding function.

4. Results

4.1. A comprehensive resource of causal relations of lncRNAs and protein-coding genes

Transcriptional profiles after gene perturbations, such as knock out, RNAi, overexpression, and CRISPRi, have been widely used to discover causal relations between perturbed genes and downstream factors. We totally collected >2700 high-throughput protein-coding gene perturbation experiments (based on microarray and RNA-seq techniques) which contained ~8800 perturbed samples and ~6200 controls. After filtering experiments with additional treatments and annotating lncRNAs in both microarray and RNA-seq data, we constructed a resource of causal relations between perturbed genes and lncRNAs/protein-coding genes from 672 independent perturbation experiments (see Methods, Supplementary Fig. S1, Supplementary Table S1). A total of over 1,180,000 causal relations were included, containing 13,870 lncRNAs and 18,734 protein-coding genes (Table S2). Almost all (99.99%) of the lncRNAs were experimentally validated (n = 2005) or manually annotated (n = 11,863), only two lncRNAs were derived from computational analysis. Among them, over 354,000 causal relationships were identified between 419 perturbed genes and these lncRNAs. Note that these perturbed protein-coding genes are involved in many different functional categories, such as cell cycle and developmental process, highlighting the high functional coverage and diversity implicated in the causal relations (Supplementary Fig. S1).

4.2. LnCAR: a novel computational approach to infer lncRNA functions using causal relations

It has been shown that the causal relations produced by an individual gene perturbation could implicate tight functional links [56]. We reasoned that when multiple genes from a common function were perturbed separately, their co-influenced factors should be also involved in that function. This hypothesis allowed us to predict the functions of abundant uncharacterized molecules, including various non-coding RNAs. LncRNAs, as a major class of regulatory molecules,

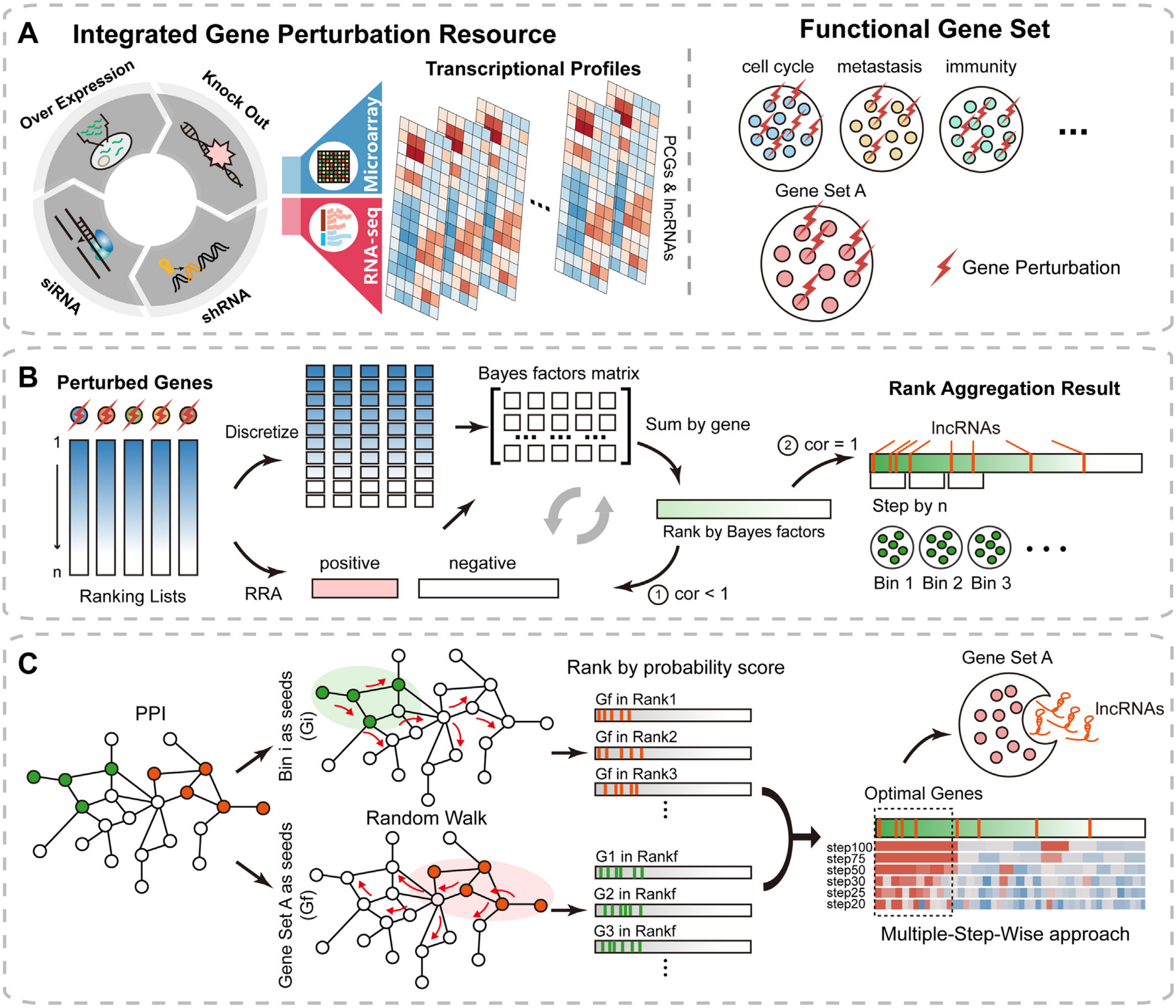


Fig. 1. Overview of the LnCAR approach. (A) General view of the gene perturbation resource and a functional gene set used in the approach. (B) Schematic of the rank aggregation method in LnCAR. (C) Schematic of the network-based method to capture lncRNAs involved in the function.

are still facing challenges in functional identification. The causal relations between perturbed genes and downstream lncRNAs could provide us a new way to reversely infer the functions of lncRNAs. Thus, we designed an approach termed LnCAR that used the resource of causal relations to capture lncRNA functions (Fig. 1).

Briefly, for a given biological function, we extracted all causal relations from our causal resource in which the perturbed genes were involved in that function. The causal relations for each perturbed gene were ranked according to their varying degree after gene perturbation and a gene ranking list was generated. We aggregated the ranking lists derived from the given function using a modified hybrid Bayesian-rank integration method which could adaptively compute a prior probability for the 'positive class' and iteratively fit Bayes factors to recompute the integration result (see Methods). Then, to capture lncRNAs most relevant to the given function, we adopted a sliding window method to analyze the functional connections of genes in each bin with that function based on protein interaction network. By using different bin sizes, a multiscale view of functional connections was investigated and an optimal threshold was determined. Finally, lncRNAs ranked above the optimal position were considered to be involved in that function (see Methods).

4.3. Using LnCAR for identifying lncRNAs involved in cell cycle

To test the performance of LnCAR, we extracted causal relations of perturbed 46 cell cycle genes (such as CDK1, CCNB1 and TP53) to capture lncRNAs involved in cell cycle. By applying LnCAR to causal relations derived from these 46 genes, an aggregated gene ranking list reflecting the relevance to cell cycle was produced and then the top 205 genes (excluded the perturbed cell cycle genes) were captured as cell cycle-related molecules, which included 201 protein-coding genes and 4 lncRNAs (Fig. 2A, Supplementary Fig. S2, Table S3). Many of the protein-coding genes have been reported to be related to cell cycle. For example, the top one protein-coding gene GDF15 has been widely investigated as a regulator in cell cycle by targeting cell cycle-regulated protein kinases [13,48]. Two of the three D-type cyclins, CCND1 and CCND2, that control cell cycle progression through the G1 phase [62], were also identified (the other cyclin CCND3 was included in the perturbed gene list). Statistically, we observed a significant enrichment of the other cell cycle members (not included in the perturbed list) at the top of the aggregated ranking list (P-value = 2.80e-03, chi-square test, Fig. 2B). Further analysis using the gene set enrichment analysis (GSEA) also confirmed the result (nominal P-value < 0.001).

When compared to the ranking lists of individual perturbed genes, our approach based on the integrative strategy showed obviously more superior performance (Fig. 2D). In parallel, functional enrichment analysis showed that the 201 protein-coding genes were significantly involved in cell cycle (FDR < 0.05, hypergeometric test, Fig. 2E and F), and significantly over-represented in S, G2/M and M phases (P-values < 0.01, Fisher's exact test, Fig. 2C) [49].

Importantly, we identified four lncRNAs, from the top 205 genes, including MALAT1, NEAT1, H19 and CCDC18-AS1. All of them have been validated to contribute to cell cycle. MALAT1, which ranked first in the aggregated ranking list, can modulate the expression of cell cycle genes and is required for G1/S and mitotic progression [51]. MALAT1-depletion can prevent the activation of genes involved in G1/S transition and S-phase progression [76]. Two studies also demonstrated the essential roles of NEAT1 and H19 in cell cycle [7,14]. CCDC18-AS1 is transcribed divergently from the promoter region of down-regulator of transcription 1 (DR1), a global transcriptional repressor related to RNA synthesis during mitosis [79].

Taken together, these results together demonstrated the effective performance of LnCAR to identify new molecules (including lncRNAs and protein-coding genes) involved in a particular function.

4.4. Comparing LnCAR to co-expression based approach

The guilt-by-association strategies have been widely used to predict the putative functions of lncRNAs based on the co-expressed protein-

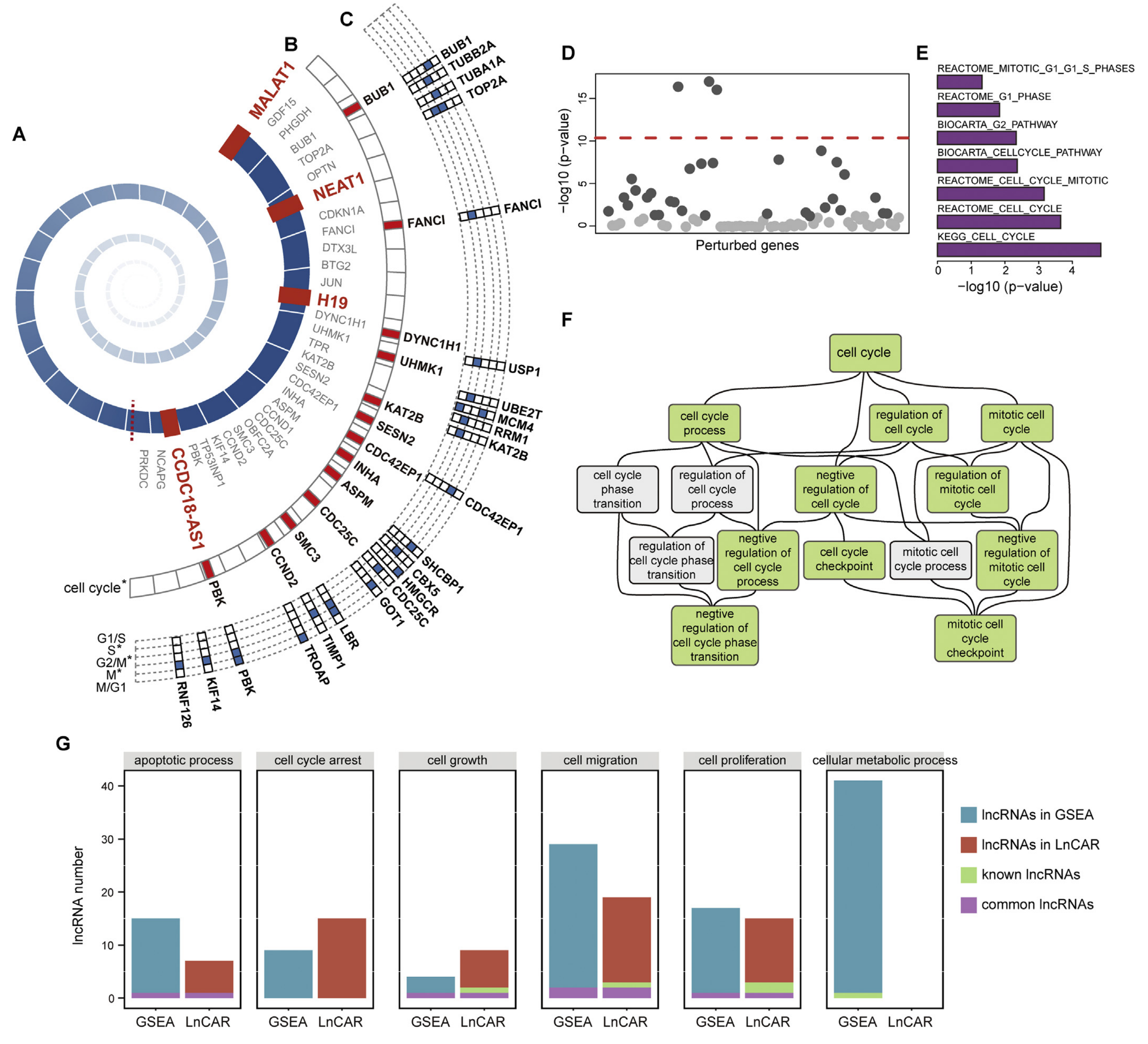


Fig. 2. Systematic assessment of LnCAR's performance. (A) Spiral diagram illustrating the rank of genes associated with cell cycle. The lncRNAs involved in cell cycle are labeled in red. (B) Enrichment of cell cycle members. The members in the optimal result are labeled in red. (C) Representation of signature gene sets for cell-cycle phases. The genes in each of the five phases (G1/S, S, G2/M, M and M/G1) are labeled in blue. (D) Predicting genes to cell cycle based on the perturbations of individual genes. Each bubble represents the Wilcoxon test P-value, showing the difference between cell cycle members and other genes in the ranking list. The horizontal line shows the P-value for the aggregated list and the colour shows if the P-value is significant (dark grey: significant; light grey: not significant). (E-F) Functional enrichment of the optimal protein-coding genes in (E) Gene Ontology and (F) three pathway databases (KEGG, REACTOME and BIOCATA). (G) Bar plots of identified lncRNAs in six functions by LnCAR and GSEA. The known lncRNAs in each function (green) and the lncRNAs identified by both methods (purple) were also showed.

coding genes [71]. To compare with our approach, we applied both LnCAR and co-expression based approach to identify lncRNAs involved in the functions assigned to the well-known lncRNAs [30], including apoptotic process, cell migration, cell proliferation, etc., and evaluated the performance of the two methods (see Methods). For the co-expression based approach, we used the GTEx expression data in 36 tissues [15] to calculate the correlations between lncRNAs and protein-coding genes. Spearman correlation coefficients were used to rank protein-coding genes for each lncRNA and then GSEA was used to identify significantly enriched functions (FDR < 0.25) [24]. We found that LnCAR could correctly capture proved lncRNAs in some functions, while lncRNAs identified by the co-expression based approach were rarely characterized (Fig. 2G). Although co-expression based approach could identify more lncRNAs than LnCAR, they do not contain some well-known lncRNAs identified by LnCAR, such as H19 and MALAT1. It should be noted that other well-studied lncRNAs were also identified by LnCAR, such as NEAT1 and FTX in cell proliferation [12,44]; H19, MALAT1 and NEAT1 in cell growth and cell cycle arrest [20,35,73,76,80].

Recently, a CRISPRi-based genome-scale screening system has been developed to reveal functional lncRNAs and identified a set of lncRNAs required for robust cellular growth [45]. Based on the result, we further compared these lncRNAs to the lncRNAs identified in the "cell growth" function by the co-expression based approach and LnCAR, respectively. As a result, three lncRNAs identified by our approach were shown to be essential for cellular growth (P-value = 8.52e-03, hypergeometric test), while no one was found by the co-expression based approach. Furthermore, considering the correlations between lncRNAs and protein-coding genes may be driven by the tissue-specific function, we obtained three temporal expression profiles on cell growth (GSE10979, n = 9, three time points; GSE18913, n = 12, three time points; GSE21912, n = 12, four time points) to calculate the correlations and then identified significantly enriched functions for each lncRNA. Unfortunately, none of these data sets could figure out lncRNAs involved in cell growth. These results showed the good performance of our method when compared with the co-expression based approach.

4.5. Discovering cancer lncRNAs by application of LnCAR to known cancer driver genes

Accumulating evidence have indicated that lncRNAs play important roles in cancer biology, and recent data suggest that they may serve as master drivers of carcinogenesis [59]. The interactions between cancer genes and lncRNAs constitute a complicated regulation circuit to cooperatively drive carcinogenesis and progression. Therefore, we believed that using the perturbation data of these cancer genes can help us to identify lncRNAs associated with cancer. To identify cancer lncRNAs, we applied LnCAR to known cancer genes listed in Cancer Gene Census (CGC) that involved 61 perturbed genes. As a result, the optimal genes showing significant functional association with the cancer genes were identified, including 7 lncRNAs and 394 protein-coding genes (Fig. 3C, Supplementary Supplementary Fig. S2, Table S3).

Among the 394 protein-coding genes, we found some well-known cancer genes that were not included in our perturbation resource, such as BUB1B and COL1A1, which play a tumor suppressor role in rhabdomyosarcoma and oncogene roles in both dermatofibrosarcoma protuberans and aneurysmal bone cyst, respectively [19]. Furthermore, these genes were significantly enriched for the cancer-associated genes obtained from the genetic association database (GAD) (hypergeometric test, P-value < 0.0001) [6]. To further assess their functional significance in oncogenesis, functional enrichment analysis was performed against known cancer-associated gene sets from 3 collections ("hallmark gene sets", "curated gene sets" and "oncogenic signatures") of the Molecular Signature Database (MSigDB) [72]. We found that the optimal protein-coding genes were highly associated with many cancer gene sets (hypergeometric test, FDR < 0.05, Fig. 3A and B), such as "INTERFERON_GAMMA_RESPONSE" and "EPITHELIAL_MESENCHYMAL_TRANSITION" in the "hallmark gene sets". These results implied the importance of these genes in carcinogenesis.

Due to the strong association of the optimal protein-coding genes with carcinogenesis, it is reasonable to believe that the 7 lncRNAs also contribute to the tumorigenesis. We discovered that three well-studied lncRNAs (MALAT1, H19 and NEAT1) can all act as oncogenes, and have been shown to regulate tumor metastasis, growth and proliferation in bladder, ovarian and breast cancer, respectively [55]. And the 7 lncRNAs were significantly enriched for cancer-associated lncRNAs from Lnc2Cancer (hypergeometric test, P-value = .005). Recently, one study has used the CRISPR–Cas9 strategy targeting lncRNAs on a genome-scale to identify lncRNAs that can disrupt or stimulate cancer cell proliferation [90]. By GSEA, the 7 lncRNAs were shown to both significantly disrupt HeLa cell proliferation of cervical cancer and stimulate Huh7.5 cell proliferation of liver cancer (FDR < 0.25). To assess their potential biological functionality, we then investigated the alteration of these lncRNA expression levels in various tumor types. We conducted a pan-cancer analysis of publicly available expression data from 4225 tumors and 532 matched adjacent normal samples across 11 cancers from TANRIC. All the 7 lncRNAs were altered in at least 3 cancer types, with consistent dysregulation in at least two tumor types (Fig. 3C). Especially, MIR22HG was consistently downregulated in all 11 cancer types, in line with previous reports in the miTranscriptome study [32], indicating its tumor suppressor role in cancer. Consistently upregulated in 5 cancer types, LINC00263 is located near protein-coding gene SCD, whose increased expression promotes the proliferation of androgen receptor (AR)-positive LNCaP prostate cancer cells [38]. Furthermore, the targeted investigation of LINC00263 for cell growth by CRISPR interference (CRISPRi) has shown that it could decrease the growth of U87 cell line [45]. These results together demonstrated the crucial roles of these 7 lncRNAs to drive carcinogenesis.

To further advance our understanding of the functional mechanisms for these 7 cancer-related lncRNAs in cancer, we explored which cancer hallmarks they might be involved in, thereby contributing to the tumorigenesis and progression. By applying LnCAR to the genes associated with 10 hallmarks collected from our manual curation [29,58], respectively, a total of 38 lncRNAs were found to be involved in the hallmarks, with obvious enrichment for known cancer-associated lncRNAs (hypergeometric test, P-value < 0.001). The 7 lncRNAs identified using CGC were included in these hallmark-related lncRNAs. Notably, they were all involved in the hallmark "Activating Invasion and Metastasis" (Fig. 3D). Among them, 3 known cancer lncRNAs (MALAT1, H19 and NEAT1) have been reported to act as pivotal players in the metastasis of cancer [22,66]. A recent study has demonstrated that MIR22HG repression impaired migration, invasion and viability of ovarian cancer cells [42]. Among the 38 hallmark-related lncRNAs, 8 and 4 lncRNAs were respectively associated with "Sustaining proliferative signaling" and "Evading growth suppressors". Based on two previous CRISPR experiments, the lncRNAs involved in these 2 hallmarks were found to disrupt cell proliferation (GSEA, FDR < 0.25) and modify the growth of cancer cells (hypergeometric test, P-value = 0.018), respectively. These results showed that these 38 lncRNAs, especially the 7 lncRNAs, may exert their tumor suppressive and/or oncogenic functions by regulating the malignant phenotype of cancer cells. Furthermore, we checked whether there exists wet lab experiment evidence supporting our results. One predicted apoptosis-associated lncRNA TNRC6C-AS1 was recently reported to powerfully modulate cell apoptosis. After down-regulation of TNRC6C-AS1 by transfecting siRNA in papillary thyroid cancer derived cell lines TPC1, apoptosis of TPC1 cells were significantly increased when compared to the control in an apoptosis assay by flow cytometry [52]. LncRNAs MIR17HG and LINC00263 were predicted to be involved in cell proliferation and growth (hallmark "Sustaining proliferative signaling") and LINC00263 also contributes to hallmark "Evading growth suppressors". Consistently, CRISPR interference (CRISPRi) of MIR17HG and LINC00263 showed that they played key roles in modulating cell growth by affecting cell growth phenotype

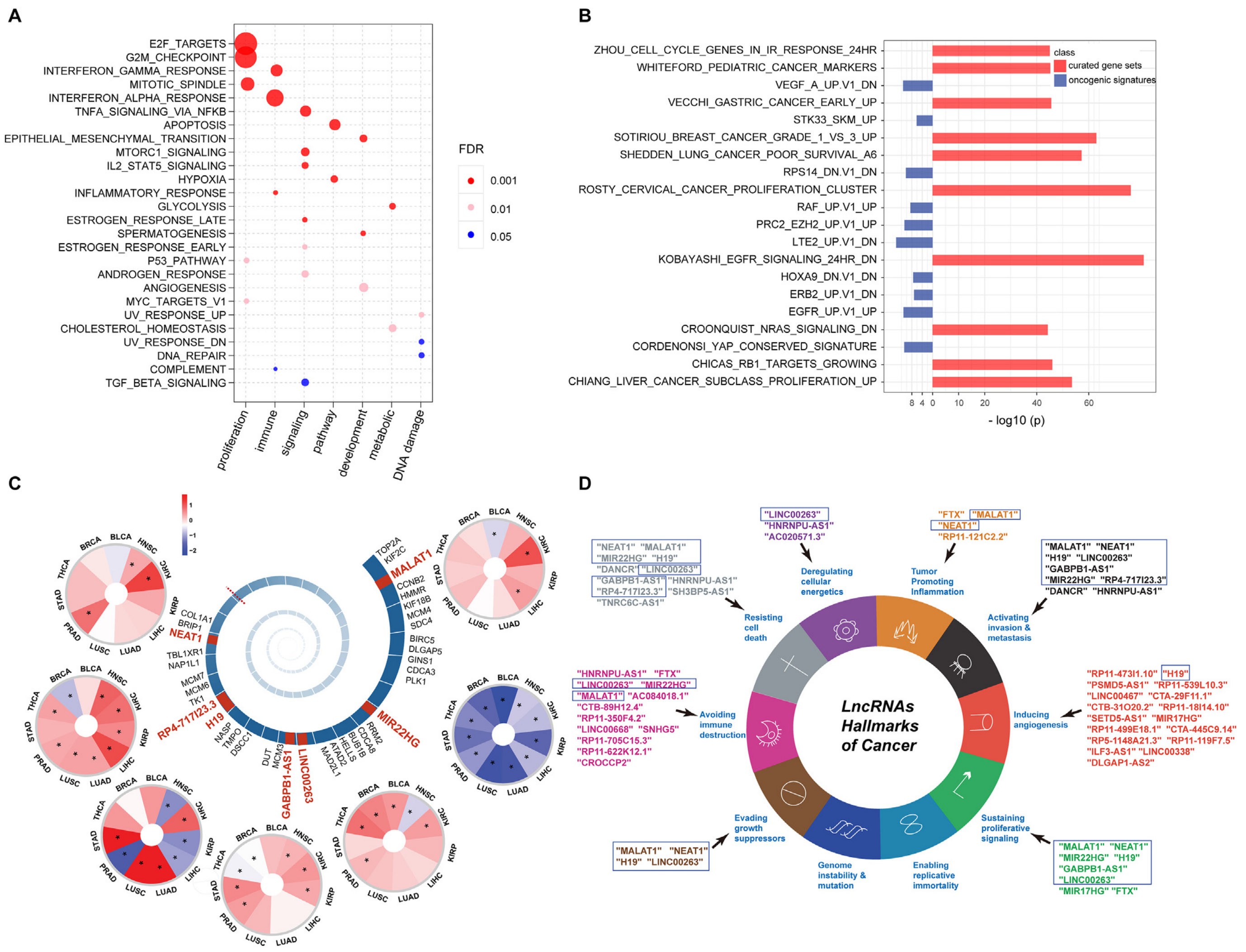


Fig. 3. Inferring lncRNAs associated with cancer. (A) Scatter chart of the significantly enriched hallmark gene sets from MSigDB database. Different colors and sizes indicate the significance levels and the percentage of overlapped genes upon hallmark gene sets, respectively. (B) Bar chart of the top 10 significantly enriched cancer-associated gene sets from MSigDB curated gene sets (red) and oncogenic signatures (blue), respectively. (C) Spiral diagram illustrating the rank of genes associated with cancer and circular heatmaps showing the differential expression patterns of the 7 captured lncRNAs in 11 tumor types from TCGA. The TCGA tumor type abbreviations are BRCA, breast invasive carcinoma; BLCA, bladder urothelial carcinoma; HNSC, head and neck squamous cell carcinoma; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; PRAD, prostate adenocarcinoma; STAD, stomach adenocarcinoma; THCA, thyroid carcinoma. (D) LncRNAs involved in the 10 hallmarks of cancer. The lncRNAs in blue rectangles are the 7 lncRNAs identified before.

[45]. We also obtained expression profiles after lncRNA perturbation for RP5-1148A21.3 (GSE85011), MIR17HG (GSE85011), LINC00467 (GSE52985). All of these lncRNAs were predicted to be involved in hallmark "Inducing angiogenesis". Based on the GSEA, we found that these lncRNAs were significantly associated with several angiogenesis-associated biological processes (Table S4) (MIR17HG was also significantly associated with cell growth and proliferation based on GSEA, data not shown). Together, these lncRNA perturbation experiments further substantiate our predictions, highlighting the efficiency of our method to capture lncRNA function.

4.6. Identifying metastatic lncRNAs and their clinical significance

Using perturbation experiments associated with the “Activating Invasion and Metastasis” hallmark in our approach, we identified nine lncRNAs (MALAT1, NEAT1, H19, LINC00263, GABPB1-AS1, MIR22HG, RP4-717I23.3, DANCR and HNRNPU-AS1) related to cancer metastasis (Fig. 4A). Research evidence supports that some identified lncRNAs are found to be involved in tumor invasion and metastasis [11,46,87]. For instance, MALAT1, metastasis-associated lung adenocarcinoma transcript 1, has already been widely accepted as having an important role in lung cancer metastasis [23,33]. High expression of lncRNA DANCR was shown to promote osteosarcoma cells proliferation, migration and invasion by upregulating AXL through competitively binding to miR-33a-5p [34]. Silencing DANCR repressed the β-catenin signaling and then inhibited hepatocellular carcinoma cell proliferation and invasion in vitro and in vivo [47]. To further confirm their roles in cancer metastasis, we collected publicly available expression data from GEO, involving 434 metastasis samples and 1177 non-metastasis tumors across five different cancer types (Table S5). As a result, seven out of nine lncRNAs presented significantly aberrant expression in at least one cancer type (Fig. 4B). For example, in prostate cancer cohort, the expression of NEAT1 and HNRNPU-AS1 showed strong association with tumor metastasis (NEAT1, P-value = 0.0001; HNRNPU-AS1, P-value = 0.0019, Fig. S3).

In clinical research, these metastatic molecules may be of great potential significance for diagnosis and prognosis of cancer metastasis. Hence we took prostate cancer as an example and evaluated the clinical benefit of NEAT1 and HNRNPU-AS1 that showed highly associations with metastasis in the previous result. We found that their expression

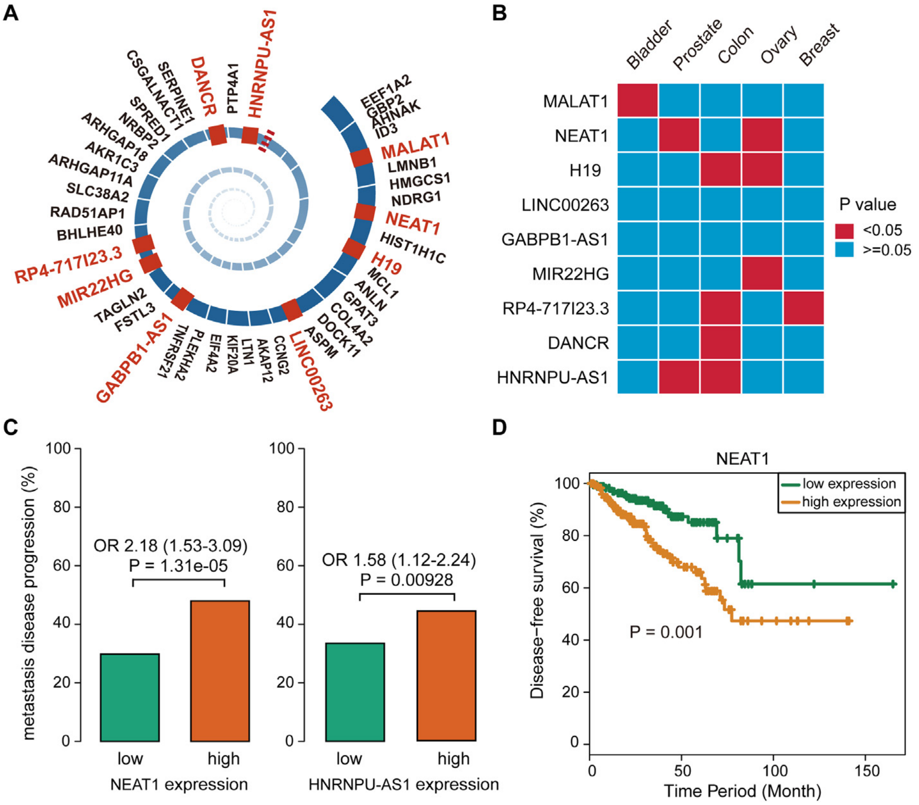


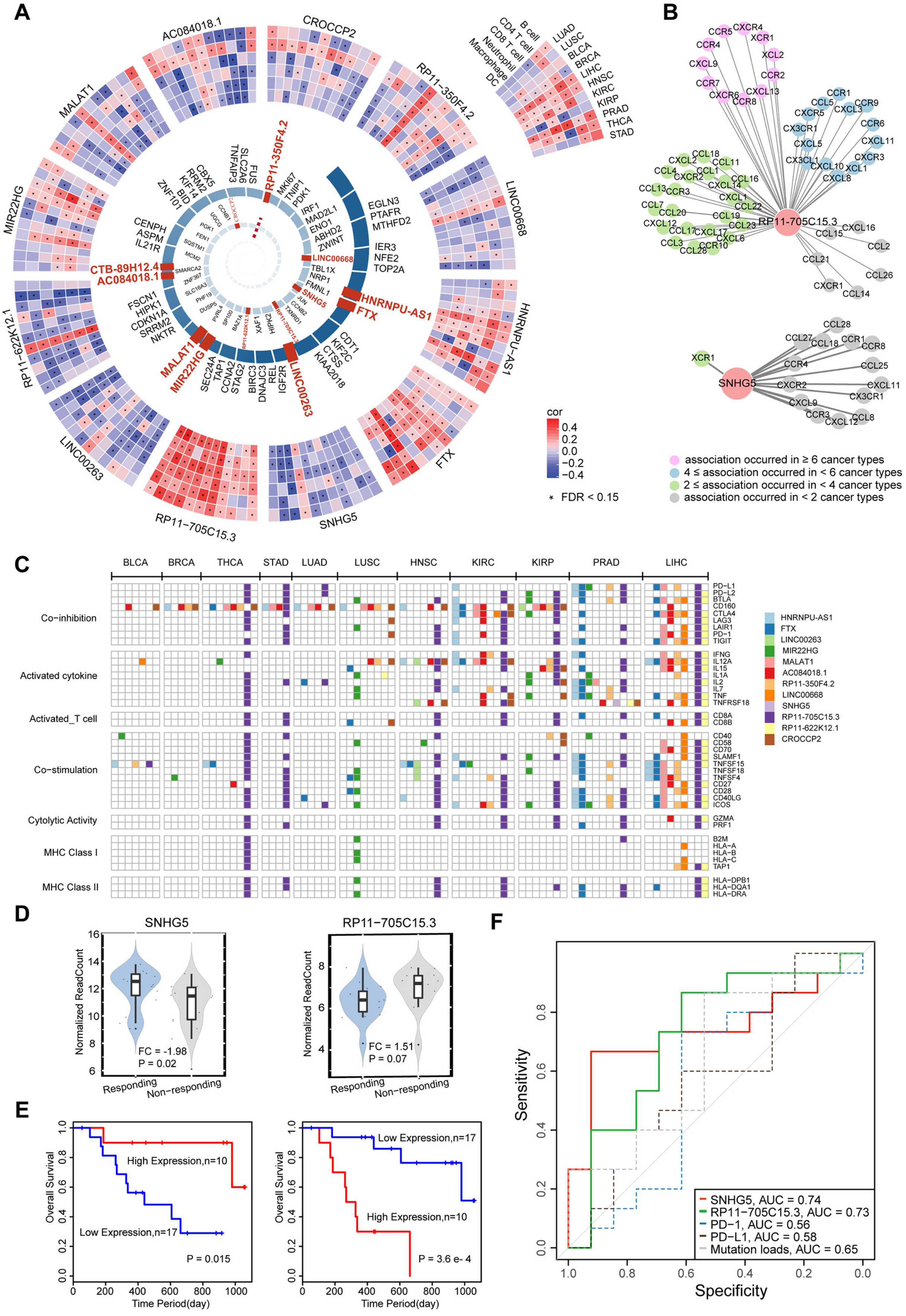
Fig. 4. Clinical analysis of metastatic lncRNAs. (A) Spiral diagram illustrating the rank of genes associated with hallmark "Activating Invasion and Metastasis". Red parts represent the optimal lncRNAs we identified. (B) Differential expression of nine metastasis-related lncRNAs across five different cancer types. The P-values were calculated by two-tailed Student's *t*-test. (C) Bar plots of patient outcomes of metastasis in the Mayo Clinic I cohort, stratified by NEAT1 or HNRNPU-AS1 expression. P-values were calculated by chi-square test. Odds ratio (OR) was presented with 95% CI. (D) Kaplan-Meier analysis of prostate cancer outcome in the TCGA cohort. The P-value was calculated by log-rank test.

levels were significantly positive correlated with Gleason score (6–9) in the Mayo Clinic I cohort (NEAT1, Cor = 1.93, P-value = 7.34e-06; HNRNPU-AS1, Cor = 0.23, P-value = 1.52–07, Supplementary Fig. S4), which is an important clinical pathologic indicator for risk stratification and therapeutic decision making in prostate cancer [26,57]. Moreover, high NEAT1 and HNRNPU-AS1 expression were associated with higher risks for metastasis (P-value = 1.31e-5, OR = 2.18 [1.53–3.09]; P-value = 0.00928, OR = 1.58 [1.12–2.24], respectively, Fig. 4C). Notably, NEAT1 was still a significant predictor of the prostate cancer metastasis even after patient stratification by Gleason score (P-value = 0.0002744, OR = 2.00 [1.37–2.92], Supplementary Fig. S5). Multivariate logistic regression analysis further confirmed that high expression of NEAT1 emerged as an independent risk factor for metastasis (P-value = 7.25e-4, OR = 1.95 [1.33–2.87], Table S6). A recent research has shown that the AUC performance of prostate-specific antigen (PSA) for metastatic progression was 0.56 [60], which approved by the FDA to monitor and predict the progression of prostate cancer in men [69]. We then evaluated the predictive power of NEAT1 and found its performance was superior to PSA (AUC = 0.62, Supplementary Fig. S6). Additionally, we utilized prostate data from TCGA and assessed the prognostic value of the NEAT1 expression. The analysis result showed that high expression of NEAT1 was significantly associated with higher pathological N stage, higher Gleason score (Table S7) and a lower rate for disease-free (P-value <0.001, Fig. 4D). Furthermore, multivariable Cox analysis revealed that NEAT1 expression was an independent prognostic variable for disease-free survival after adjusting for age, PSA, Gleason score, Clinical T stage and pathological N stage (DFS HR = 2.00 [1.03–3.91], P-value = 0.0400, Table S8). Taken together, our results demonstrated that our method can effectively identify lncRNAs of tumor metastasis, and revealed their potential as biomarkers for diagnosis and prognosis of cancer progression.

4.7. LnCAR discovered lncRNAs participating in immune response

Cancer cells exploit multiple mechanisms in order to avoid the immune attack, fortunately, immunotherapy strategy with checkpoint blockade has been raised as a promising weapon against immune escape. However, it remains a challenge to elucidate the molecular biomarkers of immune response in tumor biology. As shown above, we found 13 lncRNAs referred to the cancer hallmark "Evading Immune Destruction" (Fig. 5A). To validate their roles in the immune system, we characterized their associations with immune cell state in the tumor microenvironment (TME). First, using the estimated abundance of six main tumor-infiltrating immune cells (CD8 T cells, CD4 T cells, B cells,

Fig. 5. Immune-linked lncRNAs predict response to anti-PD-1 immunotherapy. (A) Spiral diagram indicating the rank of genes associated with "Evading Immune Destruction" and Circular heatmaps showing partial Spearman's correlations of the expression levels of lncRNAs with different immune cell infiltration levels in 11 cancers. Significant correlations at a false discovery rate (FDR) of 0.15 were indicated by asterisks. (B) Associations of RP11-705C15.3 and SNHG5 with chemokines and its receptors across cancers (Spearman's correlation >0.2 and P-value <.05), the colors reflected the frequencies they occurred across cancer types. (C) Associations of lncRNAs and immune response markers in 11 cancer types. The immune response markers covering seven different classes, including co-inhibition, activated cytokine, activated T cell, co-simulation, cytolytic activity, MHC Class I and MHC Class II. Significant associations were displayed by colored squares (Spearman's correlation >0.2 with P-value <.05). (D) Distribution of the expression levels of SNHG5 (left) and RP11-705C15.3 (right) between the responders and non-responders following anti-PD-1 immunotherapy. (E) Kaplan-Meier analysis of overall survival in 27 melanoma patients. (F) Receiver operating characteristic (ROC) curves showing the predictive powers of SNHG5 and RP11-705C15.3 (solid line) to stratify patients into responders and non-responders, compared with PD-1 and PD-L1 expressions and mutation load (dash line).


A
AC084018.1
CROCCP2
RP11-350F4.2
LINC00668
HNRNPU-AS1
FTX
SNHG5
RP11−705C15.3
LINC00263
RP11-622K12.1
MIR22HG
MALAT1
B cell
CD4 T cell
CD8 T cell
Neutrophil
Macrophage
DC
LUAD
LUSC
BLCA
BRCA
LIHC
HNSC
KIRC
KIRP
PRAD
THCA
STAD
CTB-89H12.4
AC084018.1
MALAT1
MIR22HG
LINC00263
HNRNPU-AS1
FTX
LINC00668
SNHG5
cor
0.4
0.2
0.0
−0.2
−0.4
* FDR < 0.15
B
RP11-705C15.3
SNHG5
association occurred in ≥ 6 cancer types
4 ≤ association occurred in < 6 cancer types
2 ≤ association occurred in < 4 cancer types
association occurred in < 2 cancer types
C
BLCA
BRCA
THCA
STAD
LUAD
LUSC
HNSC
KIRC
KIRP
PRAD
LIHC
Co−inhibition
Activated cytokine
Activated_T cell
Co−stimulation
Cytolytic Activity
MHC Class I
MHC Class II
PD−L1
PD−L2
BTLA
CD160
CTLA4
LAG3
LAIR1
PD−1
TIGIT
IFNG
IL12A
IL15
IL1A
IL2
IL7
TNF
TNFRSF18
CD8A
CD8B
CD40
CD58
CD70
SLAMF1
TNFSF15
TNFSF18
TNFSF4
CD27
CD28
CD40LG
ICOS
GZMA
PRF1
B2M
HLA−A
HLA−B
HLA−C
TAP1
HLA−DPB1
HLA−DQA1
HLA−DRA
HNRNPU-AS1
FTX
LINC00263
MIR22HG
MALAT1
AC084018.1
RP11-350F4.2
LINC00668
SNHG5
RP11-705C15.3
RP11-622K12.1
CROCCP2
D
SNHG5
Normalized ReadCount
FC = -1.98
P = 0.02
Responding
Non−responding
RP11−705C15.3
FC = 1.51
P = 0.07
E
Overall Survival
High Expression,n=10
Low Expression,n=17
P = 0.015
Time Period(day)
Low Expression,n=17
High Expression,n=10
P = 3.6 e- 4
F
Sensitivity
Specificity
SNHG5, AUC = 0.74
RP11−705C15.3, AUC = 0.73
PD−1, AUC = 0.56
PD−L1, AUC = 0.58
Mutation loads, AUC = 0.65

neutrophils, macrophages and dendritic cells), we investigated the connection between these lncRNAs and the immune cell infiltration [41]. We found that the expression of 12 lncRNAs (expression level of CTB-89H12.4 is unavailable) reflected the extent of different immune cell infiltration in various cancer types after correcting for tumor purity (Fig. 5A). For example, HNRNPU-AS1, FTX and RP11-705C15.3 were significantly associated with the CD8 T cell infiltration in HNSC and KIRC that has been reported previously [41,65]. Remarkably, lncRNA RP11-705C15.3 and SNHG5 were associated with infiltration of all six immune cells in most cancer types, suggesting their outstanding roles in immune infiltration regulation. Since chemokines drive the recruitment of immune cells via interactions with chemokine receptors, we thus wondered whether these two lncRNAs were involved in chemokine-induced immune cell infiltration. We found significant associations of RP11-705C15.3 with some chemokines/receptors in different cancer types, such as CXCL9 and CXCL10 in a substantial types of cancer (7 of 11 cancers and 5 of 11 respectively; both Cor > 0.2, P-value < 0.05), two chemokines recruiting effector CD8 T cells and TH1 cells into tumors (Fig. 5B) [53].

Moreover, we noticed that RP11705C-15.3 was significantly associated with dysregulation of signatures linked to immune response, including T-cell activation, cytolytic activity, antigen presentation and T-cell inhibition as well in a pan-cancer analysis (Fig. 5C). These findings highlighted the importance of these lncRNAs in affecting different tumor microenvironments and tumor-induced immune response.

4.8. Immune-linked lncRNAs predict response to anti-PD-1 immunotherapy

Although cancer immunotherapy based on the checkpoint blockade strategy has been approved for use in several advanced malignancies [27,82], there are still plenty of patients failed to respond, making the identification of predictive markers an urgent problem. Currently, PD-1 and PD-L1 expressions and mutation load have been proposed as predictors of clinical response [64,75,77], but they provided limited power to effectively optimize the patients [70]. Here, the great relation of the identified lncRNAs with T cell infiltration and activity led us to explore their potential as novel clinical predictors for checkpoint blockade efficacy. Indeed, we noticed many significant associations between these lncRNAs and inhibitory checkpoints, such as widely used targets PD-1, PD-L1 and CTLA4 across the 11 cancers (Fig. 5C, Supplementary Fig. S7). Also, RP11-705C15.3 and SNHG5 were more significantly changed in those cancer types approved for the use of checkpoint inhibitors (LUSC, KIRC, BLCA, LUAD, HNSC; Supplementary Fig. S8) [78]. Furthermore, we obtained the RNA-seq data of 28 pre-treatment melanomas following anti-PD-1 immunotherapy, in which 15 samples had clinical response while 13 had no response [31]. Expressions of RP11-705C15.3 and SNHG5 had significant associations with response to treatment (FC > 1.5, Fig. 5D). Patients with up one-third of RP11-705C15.3 expression had a 20% (2 of 10) response to anti-PD-1 therapy, compared with 72.2% (13 of 18 cases) for the low two-third (odds ratio 10.4 [95%CI 1.617–66.898]; P-value = 0.009). For SNHG5, patients with up one-third of expression had a 90% (9 of 10) response, compared to 33.3% (6 of 18) for the lower two-thirds (odds ratio 0.056 [95%CI 0.006–0.547]; P-value = 0.005). The overall survival was significantly improved in patients with lower RP11-705C15.3 expression (HR 0.117, 95% CI 0.029–0.462; log-rank P-value = 3.6e-4) and patients with higher SNHG5 expression (HR 0.119, 95% CI 0.015–0.930; log-rank P-value = 0.015). Multivariate Cox regression analysis showed they were both independent of age and gender (P-value = 0.001 for RP11-705C15.3 and P-value = 0.03 for SNHG5; Fig. 5E, Table S9). Importantly, RP11-705C15.3 and SNHG5 expressions enabled better stratification of patients into responders and non-responders with AUC of 0.73 and 0.74 respectively. By contrast, the AUC of PD-1 and PD-L1 expressions and mutation load were only 0.56, 0.58 and 0.65, respectively (Fig. 5F). Taken together, the immune-associated lncRNAs showed strong potential to be novel predictors for immunotherapy response.

4.9. The LnCAR analysis tool

To facilitate the convenient use of our approach to infer lncRNAs' functions, we developed an online tool at the following URL: http://biocc.hrbmu.edu.cn/LnCAR/. It allows users to explore lncRNAs associated with any interested biological function based on the incorporated gene perturbation resource. For a functional gene set inputted by users (a protein-coding gene set), the tool will screen relevant perturbed data and only when there are 5 or more perturbed genes in the input set, the LnCAR method will be applied to generate a rank list characterizing links with the inputted function. The lncRNAs ranked at the top of the list were predicted to be involved in the function. Moreover, users can quickly search the perturbation information of each perturbed gene, such as platform and perturbation technology, and can use a customized perturbation dataset to perform the calculation.

5. Discussion

While a large amount of lncRNAs has been exploded, only a small proportion of lncRNAs have functionally characterized roles. The limited characterization of lncRNAs will impede our further understanding of biological mechanisms and processes regulated by lncRNAs. In this paper, we proposed a novel approach, named LnCAR, to capture functions of lncRNAs based on a resource of causal relations from a large scale gene perturbation profiles. LnCAR is a robust and flexible approach for identifying lncRNAs related to any function of interest. In our study, we first captured lncRNAs involved in cell cycle process and systematically validated the performance of our approach. Then we applied LnCAR to infer cancer-associated lncRNAs and lncRNAs contributed to each cancer hallmark. We showed that the lncRNAs identified in the "activating invasion & metastasis" hallmark were strongly associated with metastatic progression in various cancer types. And lncRNAs inferred from "evading immune destruction" were proved to be important in immune cell infiltration and activity, which could also be served as indicators of the response of immunotherapy.

The use of transcriptome profiles after gene perturbations can provide us adequate causal relations between perturbed genes and downstream affected factors. Comparing to widely adopted co-expression relationships between protein-coding and non-coding genes, the causal relation would provide a more reliable measure to analyze functional mechanisms of lncRNAs. In our approach, we used a rank-based method to integrate causal relations produced by gene perturbation experiments, which can solve the problem of the heterogeneity among different platforms and improve the data availability dramatically [39]. It should be noted that LnCAR was able to capture lncRNAs from any gene set with similar functional properties and enough perturbation profiles in our resource. In this study, we not only inferred cancer-related lncRNAs but also identified lncRNAs involved in each cancer hallmark, allowing to look into the underlying mechanisms of lncRNAs in cancer progression. For lncRNAs involved in "evading immune destruction", we not only found their significant association with different immune cell infiltration but also with some chemokines/receptors in various cancer types, highlighting their important roles in immune response. Furthermore, two lncRNAs, RP11-705C15.3 and SNHG5, were found to be highly correlated with response to anti-PD-1 immunotherapy, and be correlated with patient survival and better stratification, suggesting their potential as novel clinical predictors for immunotherapy response.

In our resource, the perturbation experiments were generated from multiple technical platforms, which can lead to different numbers of re-annotated lncRNAs and thus different numbers of causal relationships captured by perturbation experiment. To reduce the impact of platform heterogeneity, one of the major advantages of LnCAR applies a weight-based rank aggregation method in which the reliability of individual datasets is estimated, which can, to some extent, mitigate the impact of the heterogeneity among different technical platforms [2–4,39]. Even so, integration of perturbation data derived from different platform can still block LnCAR to capture more exhaustive causal relationships, resulting in the loss of some valuable lncRNAs. Besides, our method is designed to capture the co-influenced factors across multiple perturbed genes from a common function, with limited performance for some lncRNAs highly dependent on specific components in the function of interest. With the wider of applications of sequencing technology in perturbation experiment, the performance of LnCAR will become more robust and effective. Moreover, when more RNA-seq perturbation data derived from RNA-seq experiments on total RNA are generated in the future, non-polyadenylated lncRNAs can be also analyzed. It is conceivable that other types of perturbation experiments following genome-wide expression profiles and continuously updated lncRNA information can further facilitate revealing lncRNAs' functions. It is known that there exist some lncRNAs playing completely different roles in different contexts (some lncRNAs may function as oncogenes in certain types of cancer, but may function as tumor suppressors in other types of cancer). The cellular context is an important aspect to be considered. Recently, combination of CRISPR/Cas9-based gene editing technology and single cell RNA-seq technology has been used to generate perturbation-based expression data in specific cells [1,16]. We expect that more and more perturbation-based expression data will be produced, which can help us to infer lncRNAs' functions in the specific cellular context. Considering that lncRNAs often exert their function through interaction with various components such as DNA elements, RNAs or proteins, integrating more types of data (e.g., ChIP-seq, RIP-seq and ChIRP-seq data) will further strength our method, which can make our prediction more biologically relevant. In the future, we anticipate that the applicability of LnCAR would be continually growing as we incorporate new datasets into our existing resource of causal relations. The way we delivered for leveraging causal relations shed new light on the characterization of lncRNAs, which would be helpful for further experimental research and clinical applications of lncRNA class.

Funding

This work was supported in part by the National High Technology Research and Development Program of China [863 Program, Grant Nos. 2014AA021102], the National Program on Key Basic Research Project [973 Program, Grant Nos. 2014CB910504], the National Natural Science Foundation of China [Grant Nos. 91439117, 61473106, 61573122], the China Postdoctoral Science Foundation (2016M600260), Wu lienteh youth science fund project of Harbin medical university [Grant Nos. WLD-QN1407], Special funds for the construction of higher education in Heilongjiang Province [Grant Nos. UNPYSCT-2016049], and the Heilongjiang Postdoctoral Foundation (LBH-Z16098).

Declaration of Interests

The authors have no financial conflicts to declare.

Author Contributions

XL, YX and JH conceived, designed, and supervised the study. JX, AS, ZL, LX, GL, CD, MY, AX, and TL collected data. Analysis and data interpretation were done by JX, AS, ZL and LX. JX, AS, ZL and LX wrote the drafts of the manuscript. JX, AS, ZL, LX, GL, CD, MY, AX, and TL commented on and revised the drafts of the manuscript. All authors read and approved the final report.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ebiom.2018.08.050.

References

[1] Adamson B, Norman TM, Jost M, Cho MY, Nunez JK, Chen Y, et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. Cell 2016;167:1867–82 (e1821).

[2] Adler P, Kolde R, Kull M, Tkachenko A, Peterson H, Reimand J, et al. Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. Genome Biol 2009;10:R139.

[3] Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, et al. Gene prioritization through genomic data fusion. Nat Biotechnol 2006;24:537–44.

[4] Badgeley MA, Sealfon SC, Chikina MD. Hybrid Bayesian-rank integration approach improves the predictive power of genomic dataset aggregation. Bioinformatics 2015;31:209–15.

[5] Bassett AR, Akhtar A, Barlow DP, Bird AP, Brockdorff N, Duboule D, et al. Considerations when investigating lncRNA function in vivo. Elife 2014;3:e03058.

[6] Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. Nat Genet 2004;36:431–2.

[7] Berteaux N, Lottin S, Monte D, Pinte S, Quatannens B, Coll J, et al. H19 mRNA-like noncoding RNA promotes breast cancer cell proliferation through positive control by E2F1. J Biol Chem 2005;280:29625–36.

[8] Boettcher M, McManus MT. Choosing the right Tool for the Job: RNAi, TALEN, or CRISPR. Mol Cell 2015;58:575–85.

[9] Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 2011;25:1915–27.

[10] Chakraborty D, Kappei D, Theis M, Nitzsche A, Ding L, Paszkowski-Rogacz M, et al. Combined RNAi and localization for functionally dissecting long noncoding RNAs. Nat Methods 2012;9:360–2.

[11] Chakravarty D, Sboner A, Nair SS, Giannopoulou E, Li R, Hennig S, et al. The oestrogen receptor alpha-regulated lncRNA NEAT1 is a critical modulator of prostate cancer. Nat Commun 2014;5:5383.

[12] Chen X, Kong J, Ma Z, Gao S, Feng X. Up regulation of the long non-coding RNA NEAT1 promotes esophageal squamous cell carcinoma cell progression and correlates with poor prognosis. Am J Cancer Res 2015;5:2808–15.

[13] Cheung PK, Woolcock B, Adomat H, Sutcliffe M, Bainbridge TC, Jones EC, et al. Protein profiling of microdissected prostate tissue links growth differentiation factor 15 to prostate carcinogenesis. Cancer Res 2004;64:5929–33.

[14] Clemson CM, Hutchinson JN, Sara SA, Ensminger AW, Fox AH, Chess A, et al. An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. Mol Cell 2009;33:717–26.

[15] Consortium GT. Human genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 2015;348:648–60.

[16] Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell 2016;167:1853–66 (e1817).

[17] Du Z, Fei T, Verhaak RG, Su Z, Zhang Y, Brown M, et al. Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. Nat Struct Mol Biol 2013;20:908–13.

[18] Esteller M. Non-coding RNAs in human disease. Nat Rev Genet 2011;12:861–74.

[19] Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer 2004;4:177–83.

[20] Gabory A, Jammes H, Dandolo L. The H19 locus: role of an imprinted non-coding RNA in growth and development. Bioessays 2010;32:473–80.

[21] Goyal A, Myacheva K, Gross M, Klingenberg M, Duran Arque B, Diederichs S. Challenges of CRISPR/Cas9 applications for long non-coding RNA genes. Nucleic Acids Res 2017;45:e12.

[22] Guo S, Chen W, Luo Y, Ren F, Zhong T, Rong M, et al. Clinical implication of long non-coding RNA NEAT1 expression in hepatocellular carcinoma patients. Int J Clin Exp Pathol 2015;8:5395–402.

[23] Gutschner T, Hammerle M, Eissmann M, Hsu J, Kim Y, Hung G, et al. The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. Cancer Res 2013;73:1180–9.

[24] Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 2009;458:223–7.

[25] Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, et al. lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 2011;477:295–300.

[26] Ham WS, Chalfin HJ, Feng Z, Trock BJ, Epstein JI, Cheung C, et al. New prostate cancer grading system predicts long-term survival following surgery for gleason score 8-10 prostate cancer. Eur Urol 2017;71:907–12.

[27] Hamid O, Robert C, Daud A, Hodi FS, Hwu WJ, Kefford R, et al. Safety and tumor responses with lambrolizumab (anti-PD-1) in melanoma. N Engl J Med 2013;369:134–44.

[28] Heward JA, Lindsay MA. Long non-coding RNAs in the regulation of the immune response. Trends Immunol 2014;35:408–19.

[29] Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA, et al. Super-enhancers in the control of cell identity and disease. Cell 2013;155:934–47.

[30] Huarte M. The emerging role of lncRNAs in cancer. Nat Med 2015;21:1253–61.

[31] Hugo W, Zaretsky JM, Sun L, Song C, Moreno BH, Hu-Lieskovan S, et al. Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. Cell 2017;168:542.

[32] Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, et al. The landscape of long noncoding RNAs in the human transcriptome. Nat Genet 2015;47:199–208.

[33] Ji P, Diederichs S, Wang W, Boing S, Metzger R, Schneider PM, et al. MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene 2003;22:8031–41.

[34] Jiang N, Wang X, Xie X, Liao Y, Liu N, Liu J, et al. lncRNA DANCR promotes tumor progression and cancer stemness features in osteosarcoma by upregulating AXL via miR-33a-5p inhibition. Cancer Lett 2017;405:46–55.

[35] Jiao F, Hu H, Yuan C, Wang L, Jiang W, Jin Z, et al. Elevated expression level of long noncoding RNA MALAT-1 facilitates cell growth, migration and invasion in pancreatic cancer. Oncol Rep 2014;32:2485–92.

[36] Kemmeren P, Sameith K, van de Pasch LA, Benschop JJ, Lenstra TL, Margaritis T, et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. Cell 2014;157:740–52.

[37] Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci U S A 2009;106:11667–72.

[38] Kim SJ, Choi H, Park SS, Chang C, Kim E. Stearoyl CoA desaturase (SCD) facilitates proliferation of prostate cancer cells through enhancement of androgen receptor transactivation. Mol Cells 2011;31:371–7.

[39] Kolde R, Laur S, Adler P, Vilo J. Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics 2012;28:573–80.

[40] Lennox KA, Behlke MA. Cellular localization of long non-coding RNAs affects silencing by RNAi more than by antisense oligonucleotides. Nucleic Acids Res 2016;44:863–77.

[41] Li B, Severson E, Pignon JC, Zhao H, Li T, Novak J, et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. Genome Biol 2016;17:174.

[42] Li J, Yu H, Xi M, Lu X. Long noncoding RNA C17orf91 is a potential prognostic marker and functions as an oncogene in ovarian cancer. J Ovarian Res 2016;9:49.

[43] Liao Q, Liu C, Yuan X, Kang S, Miao R, Xiao H, et al. Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. Nucleic Acids Res 2011;39:3864–78.

[44] Liu F, Yuan JH, Huang JF, Yang F, Wang TT, Ma JZ, et al. Long noncoding RNA FTX inhibits hepatocellular carcinoma proliferation and metastasis by binding MCM2 and miR-374a. Oncogene 2016;35:5422–34.

[45] Liu SJ, Horlbeck MA, Cho SW, Birk HS, Malatesta M, He D, et al. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. Science 2017;355.

[46] Luo M, Li Z, Wang W, Zeng Y, Liu Z, Qiu J. Long non-coding RNA H19 increases bladder cancer metastasis by associating with EZH2 and inhibiting E-cadherin expression. Cancer Lett 2013;333:213–21.

[47] Ma X, Wang X, Yang C, Wang Z, Han B, Wu L, et al. DANCR Acts as a diagnostic biomarker and promotes tumor growth and metastasis in hepatocellular carcinoma. Anticancer Res 2016;36:6389–98.

[48] Maaser K, Borlak J. A genome-wide expression analysis identifies a network of EpCAM-induced cell cycle regulators. Br J Cancer 2008;99:1635–43.

[49] Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 2015;161:1202–14.

[50] Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. Nat Rev Genet 2009;10:155–9.

[51] Michalik KM, You X, Manavski Y, Doddaballapur A, Zornig M, Braun T, et al. Long noncoding RNA MALAT1 regulates endothelial cell function and vessel growth. Circ Res 2014;114:1389–97.

[52] Muhanhali D, Zhai T, Jiang J, Ai Z, Zhu W, Ling Y. Long non-coding antisense RNA TNRC6C-AS1 is activated in papillary thyroid cancer and promotes cancer progression by suppressing TNRC6C expression. Front Endocrinol 2018;9:360.

[53] Nagarsheth N, Wicha MS, Zou W. Chemokines in the cancer microenvironment and their relevance in cancer immunotherapy. Nat Rev Immunol 2017;17:559–72.

[54] Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature 2014;505:635–40.

[55] Ning S, Zhang J, Wang P, Zhi H, Wang J, Liu Y, et al. Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. Nucleic Acids Res 2016;44:D980–5.

[56] Parnas O, Jovanovic M, Eisenhaure TM, Herbst RH, Dixit A, Ye CJ, et al. A Genome-wide CRISPR Screen in primary Immune Cells to Dissect Regulatory Networks. Cell 2015;162:675–86.

[57] Pierorazio PM, Walsh PC, Partin AW, Epstein JI. Prognostic Gleason grade grouping: data based on the modified Gleason scoring system. BJU Int 2013;111:753–60.

[58] Plaisier CL, Pan M, Baliga NS. A miRNA-regulatory network explains how dysregulated miRNAs perturb oncogenic processes across diverse cancers. Genome Res 2012;22:2302–14.

[59] Prensner JR, Chinnaiyan AM. The emergence of lncRNAs in cancer biology. Cancer Discov 2011;1:391–407.

[60] Prensner JR, Zhao S, Erho N, Schipper M, Iyer MK, Dhanasekaran SM, et al. RNA biomarkers associated with metastatic progression in prostate cancer: a multi-institutional high-throughput analysis of SChLAP1. Lancet Oncol 2014;15:1469–80.

[61] Quek XC, Thomson DW, Maag JL, Bartonicek N, Signal B, Clark MB, et al. lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. Nucleic Acids Res 2015;43:D168–73.

[62] Quelle DE, Ashmun RA, Shurtleff SA, Kato JY, Bar-Sagi D, Roussel MF, et al. Overexpression of mouse D-type cyclins accelerates G1 phase in rodent fibroblasts. Genes Dev 1993;7:1559–71.

[63] Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, et al. A large-scale evaluation of computational protein function prediction. Nat Methods 2013;10:221–7.

[64] Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science 2015;348:124–8.

[65] Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell 2015;160:48–61.

[66] Shen XH, Qi P, Du X. Long non-coding RNAs in cancer invasion and metastasis. Mod Pathol 2015;28:4–13.

[67] Signal B, Gloss BS, Dinger ME. Computational Approaches for Functional Prediction and Characterisation of Long Noncoding RNAs. Trends Genet 2016;32:620–37.

[68] Smith MA, Gesell T, Stadler PF, Mattick JS. Widespread purifying selection on RNA structure in mammals. Nucleic Acids Res 2013;41:8220–36.

[69] Sohn E. Screening: diagnostic dilemma. Nature 2015;528:S120–2.

[70] Spencer KR, Wang J, Silk AW, Ganesan S, Kaufman HL, Mehnert JM. Biomarkers for Immunotherapy: Current Developments and Challenges. , vol. 35American Society of Clinical Oncology Educational Book American Society of Clinical Oncology Meeting; 2016; e493–503.

[71] Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science 2003;302:249–55.

[72] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50.

[73] Sun C, Li S, Zhang F, Xi Y, Wang L, Bi Y, et al. Long non-coding RNA NEAT1 promotes non-small cell lung cancer progression through regulation of miR-377-3p-E2F3 pathway. Oncotarget 2016;7:51784–814.

[74] Suratanee A, Plaimas K. DDA: a Novel Network-based Scoring Method to Identify Disease-Disease Associations. Bioinforma Biol Insights 2015;9:175–86.

[75] Topalian SL, Hodi FS, Brahmer JR, Gettinger SN, Smith DC, McDermott DF, et al. Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. N Engl J Med 2012;366:2443–54.

[76] Tripathi V, Shen Z, Chakraborty A, Giri S, Freier SM, Wu X, et al. Long noncoding RNA MALAT1 controls cell cycle progression by regulating the expression of oncogenic transcription factor B-MYB. PLoS Genet 2013;9:e1003368.

[77] Tumeh PC, Harview CL, Yearley JH, Shintaku IP, Taylor EJ, Robert L, et al. PD-1 blockade induces responses by inhibiting adaptive immune resistance. Nature 2014;515:568–71.

[78] Turajlic S, Litchfield K, Xu H, Rosenthal R, McGranahan N, Reading JL, et al. Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. Lancet Oncol 2017;18:1009–21.

[79] van der Meijden CM, Lapointe DS, Luong MX, Peric-Hupkes D, Cho B, Stein JL, et al. Gene profiling of cell cycle progression through S-phase reveals sequential expression of genes required for DNA replication and nucleosome assembly. Cancer Res 2002;62:3233–43.

[80] Wang P, Wu T, Zhou H, Jin Q, He G, Yu H, et al. Long noncoding RNA NEAT1 promotes laryngeal squamous cell cancer through regulating miR-107/CDK6 pathway. J Exp Clin Cancer Res 2016;35:22.

[81] Wapinski O, Chang HY. Long noncoding RNAs and human disease. Trends Cell Biol 2011;21:354–61.

[82] Wolchok JD. PD-1 Blockers. Cell 2015;162:937.

[83] Xiao Y, Gong Y, Lv Y, Lan Y, Hu J, Li F, et al. Gene Perturbation Atlas (GPA): a single-gene perturbation repository for characterizing functional mechanisms of coding and non-coding genes. Sci Rep 2015;5:10889.

[84] Xiao Y, Hsiao TH, Suresh U, Chen HI, Wu X, Wolf SE, et al. A novel significance score for gene selection and ranking. Bioinformatics 2014;30:801–7.

[85] Xiao Y, Lv Y, Zhao H, Gong Y, Hu J, Li F, et al. Predicting the functions of long noncoding RNAs using RNA-seq based on Bayesian network. Biomed Res Int 2015;2015:839590.

[86] Yao P, Lin P, Gokoolparsadh A, Assareh A, Thang MW, Voineagu I. Coexpression networks identify brain region-specific enhancer RNAs in the human brain. Nat Neurosci 2015;18:1168–74.

[87] Zhang L, Yang F, Yuan JH, Yuan SX, Zhou WP, Huo XS, et al. Epigenetic activation of the MiR-200 family contributes to H19-mediated metastasis suppression in hepatocellular carcinoma. Carcinogenesis 2013;34:577–86.

[88] Zhao T, Xu J, Liu L, Bai J, Wang L, Xiao Y, et al. Computational identification of epigenetically regulated lncRNAs and their associated genes based on integrating genomic data. FEBS Lett 2015;589:521–31.

[89] Zhao T, Xu J, Liu L, Bai J, Xu C, Xiao Y, et al. Identification of cancer-related lncRNAs through integrating genome, regulome and transcriptome features. Mol Biosyst 2015;11:126–36.

[90] Zhu S, Li W, Liu J, Chen CH, Liao Q, Xu P, et al. Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR-Cas9 library. Nat Biotechnol 2016;34:1279–86.