Abstract Number: 3353

Discovery and validation of orphan noncoding RNA profiles across multiple cancers in TCGA and two independent cohorts

Exai Bio

Jeffrey Wang[1,2], Helen Li[1], Lisa Fish[1,2], Kimberly H. Chau[1], Patrick Arensdorf[1], Hani Goodarzi[2], Babak Alipanahi[1]

[1]Exai Bio Inc., Palo Alto, CA; [2]UCSF School of Medicine, University of California, San Francisco, CA

UCSF School of Medicine

Background

- Small non-coding RNAs (sncRNAs) have established roles as post-transcriptional regulators of cancer pathogenesis.
- We recently reported a novel and previously unannotated class of sncRNAs that were found in breast cancer tissue but not in normal tissue adjacent to the tumor (hereinafter normal), which we termed orphan non-coding RNA (oncRNA).[1]
- We showed that one of these oncRNAs is exploited by breast cancer cells to promote cancer metastasis,[1] which suggests that other oncRNAs may also have pathological roles in cancer.
- We now hypothesize that oncRNAs are present in many types of cancer and that oncRNAs may enable a new, RNA-based liquid biopsy strategy for early detection and monitoring in a wide variety of cancers.

Goals

- Identify and validate novel oncRNAs in six different cancer sites in 10,963 samples (7,942 cancer and 3,021 normal samples) across three large independent cohorts to create the largest known library of oncRNAs.
- Develop and validate an artificial intelligence-based (AI) approach to predict cancer tissue-of-origin by leveraging oncRNA expression profiles.

Samples

- Two sources provided small RNA (smRNA) data for 10 types of cancer from 6 tissue sites and for their corresponding normal tissues.
  1. **The Cancer Genome Atlas (TCGA)**. A joint NCI and NHGRI program that provides open-source tumor/normal sequencing data. TCGA smRNA data were used here to discover new oncRNAs and to train a predictive AI model.
  2. **IndivuType**. A comprehensive multi-omics dataset provided by Indivumed GmbH that offers sequencing data from a global patient sample collection. IndivuType provided two cohorts, A and B, which were used here to validate oncRNAs identified in TCGA and to validate the AI tissue-of-origin model.
- The 6 cancer sites studied here (breast, colorectal, gastric, kidney, liver, lung) represent 46% of new cancer diagnoses and 51% of cancer deaths worldwide, per GLOBOCAN 2020.[2]
- Sample collection/preparation and RNA sequencing for the TCGA and IndivuType cohorts were performed prior to and independently of this study, using standard methods. Patients had provided informed consent and contributing centers had obtained IRB approval.

Methods

- Fisher's Exact Test followed by Benjamini-Hochberg correction was used to identify cancer-specific oncRNAs in TCGA data. OncRNAs were considered validated in IndivuType data if statistical test *P* values ≤ 0.05.
- An eXtreme Gradient Boosting (XGB) model for prediction of cancer tissue-of-origin from oncRNA profiles was trained with TCGA data and validated in the two IndivuType cohorts.

Results 1: *Three Large Cohorts*

Table 1. TCGA Samples used for Identifying oncRNAs

Numbers of independent samples evaluated from TCGA cohort, grouped by cancer site and tissue type (cancer or normal tissue).

Cancer site	TCGA code	TCGA Tissue Type	
		Cancer (n)	Normal (n)
Breast	BRCA	1,103	104
Colorectal	COAD, READ	619	11
Gastric	STAD	446	45
Kidney	KICH, KIRC, KIRP	903	130
Liver	LIHC	375	50
Lung	LUAD, LUSC	999	91
TOTAL	*All 10 codes above*	4,445	431

Table 2. IndivuType Samples used for Validating oncRNAs

Samples evaluated from non-overlapping IndivuType cohorts, A and B, are grouped by cancer site and tissue type (cancer or normal tissue).

Cancer site	Cohort A		Cohort B	
	Cancer (n)	Normal (n)	Cancer (n)	Normal (n)
Breast	117	59	359	0*
Colorectal	906	682	230	149
Gastric	164	163	46	53
Kidney	111	96	201	186
Liver	145	142	62	66
Lung	802	637	354	357
TOTAL	2,245	1,779	1,252	811

*Cohort B lacked normal tissues for breast cancer, which excluded it from validation.

Results 2: *Heat Map of oncRNAs in TCGA Cohort*

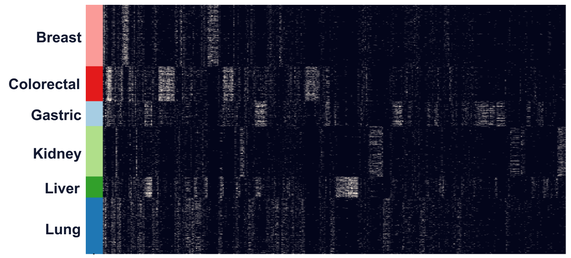


Figure 1. Discovery of cancer-specific orphan non-coding RNAs across 6 cancer sites.
749 representative oncRNAs discovered in 4,445 cancer samples across 6 cancer sites are shown.

Results 3: *Identification and Validation of oncRNAs*

Table 3. Identification of new oncRNAs in TGCA Cohort and Validation in IndivuType cohorts.

- In total, 144,695 distinct oncRNAs were identified among the six cancers.
- For each cancer site category, the majority of oncRNAs identified in TCGA were validated in one IndivuType cohort (Union column).
- A total of 51,208 oncRNAs were validated in both independent IndivuType cohorts.

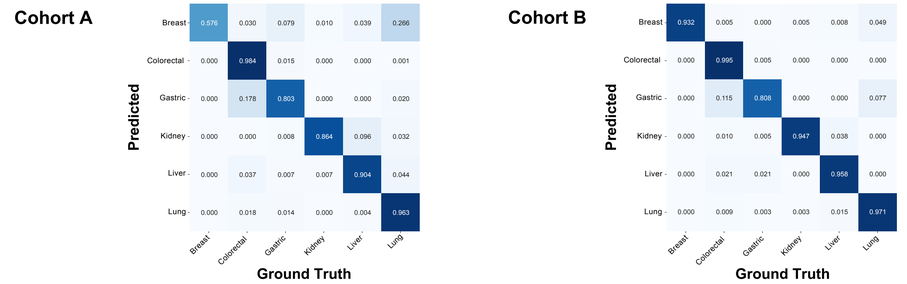
Cancer site*	Numbers of oncRNAs			oncRNAs Validated in Cohorts A and B		
	Discovery	Validation				
	TCGA cohort	Cohort A	Cohort B	Union (*A or B*)	Intersection (*A and B*)	Combined *P*‡ (*A and B*)
Breast	15,869	7,538	NA**	NA**	NA**	NA**
Colorectal	57,663	35,013	31,657	38,589 (66.9%)	28,081 (48.7%)	35,816 (62.1%)
Gastric	34,895	13,104	11,685	16,621 (47.6%)	8,168 (23.4%)	14,094 (40.4%)
Kidney	42,183	6,271	10,580	12,553 (29.8%)	4,298 (10.1%)	9,217 (21.9%)
Liver	36,929	9,749	12,703	15,910 (43.1%)	6,542 (17.7%)	12,809 (34.7%)
Lung	86,008	41,018	34,996	47,829 (55.6%)	28,185 (32.8%)	42,793 (49.8%)
TOTAL†	144,695	68,310	64,580	82,046 (56.7%)	51,208 (35.4%)	70,575 (48.8%)

* TCGA codes for each cancer site appear in Table 1. **Normal tissue samples were not available from Cohort B breast cancer patients. †Totals are less than the sum of rows above it because some oncRNAs were found in >1 cancer site. ‡Stouffer's method was used to combine *P* values from Cohort A and Cohort B for each oncRNA.

Results 4: *AI Analysis of oncRNA Profiles to Predict Cancer Tissue-of-Origin*

Figure 2. Validation in 2 IndivuType Cohorts of an XGB prediction model that was Trained on TCGA oncRNA Data.

- Accuracies were: 91.5% (95% CI: 90.3%–92.7%) for IndivuType Cohort A, and 96% (94.7%–97.0%) for IndivuType Cohort B.



Numbers across a row indicate the frequencies of the ground truth cancer among samples for which the model predicted the cancer site named at left.

Conclusions

- We have identified 144,695 distinct oncRNAs across 6 types of cancer, of which 51,208 were validated in each of two independent cohorts.
- We developed an artificial intelligence model that uses oncRNA profiles to predict cancer tissue-of-origin with high accuracy.
- These results suggest that oncRNAs are a unique, generalizable feature of cancers with potential for use in cancer detection and monitoring.

Acknowledgments: Mathias Saver and Margarita Krawczyk from Indivumed are thanked for performing data processing. Indivumed combines the world's most comprehensive multi-omics cancer data with extensive medical and bioinformatics expertise. Samples are collected within a global clinical network using a standardized approach to ensure biospecimen quality.

Disclosure: JW, HL, LF, KC are full-time employees of Exai Bio. BA and PA are co-founders, stockholders, and full-time employees of Exai Bio. HG is co-founder, stockholder, and advisor of Exai Bio.

References

1. Fish L. *et al., Nature Med.* 2018;24:1743-51.
2. Sung H. *et al., CA Cancer J Clin. 2021*: 71: 209- 249.