Lantern: a semi-automated pipeline and repository for annotating lncRNAs with ontologies

INDIANA UNIVERSITY
SCHOOL OF INFORMATICS AND COMPUTING
IUPUI

Swapna Vidhur Daulatabad[1] , Rajneesh Srivastava[1] , Sarath Chandra Janga[1, 2, 3]

[1] Department of BioHealth Informatics, Indiana University School of Informatics and Computing, IUPUI. [2] Centre for Computational Biology and Bioinformatics, Health Information and Translational Sciences, Indianapolis. [3] Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis.

ABSTRACT

With advancements in in omics technologies, the range of biological processes long non-coding RNAs (lncRNAs) are involved in is expanding extensively [1, 2]. The accelerating rate of evidence discovery of lncRNAs' role in various critical biochemical, cellular, and physiological processes is necessitating a robust platform of lncRNA annotation resources. Available resources with lncRNA ontology annotations are rare: despite a plethora of resources for annotating genes, and an extensive body of lncRNA literature. Here, we present a lncRNA annotation extractor and repository (Lantern), was developed using PubMed's abstract retrieval engine and NCBO's recommender annotation system [3]. Between 1-150 abstracts were extracted per lncRNA , which were subsequently used for extracting annotations with respect to each ontology by querying NCBO's recommender system via Application Programming Interface (API). To evaluate the quality of annotations in Lantern, benchmarking analysis was performed by deploying Lantern's pipeline over 182 lncRNAs from lncRNAdb [4] and compared the extracted annotations against annotations mapped onto the lncRNAdb's manually curated free text. Benchmarking analysis suggested that Lantern has a recall of 0.62 against lncRNAdb for 182 lncRNAs and precision of 0.8 based on manual verification of ontology annotations for 50 lncRNAs. Additionally, lncRNAs were also annotated with multiple omics information, like: RBP-interaction, tissue specific expression, protein c-expression, coding potential, sub-cellular localization and SNPs for around 11000 lncRNAs; retrieved and analyzed by robust NGS tools and pipelines. The extracted annotations for 11000 lncRNAs are available at http://www.iupui.edu/~sysbio/lantern/.

OBJECTIVES

- Develop an automated pipeline for annotating lncRNAs with multiple ontologies.
- Extract up to date annotations for multiple ontologies using latest literature.
- Develop a lncRNA annotation resource with a web interface.
- Annotate lncRNAs with high quality controlled multiple omics annotations for improving the annotation of the noncoding transcriptome.

MATERIALS & METHODS

- LncRNA relevant literature extracted using PubMed API.
- Ontology annotations from the literature were extracted using NCBO ontology recommender.
- Extracted ontologies were benchmarked against annotations extracted from lncRNAdb.
- The following formulae were used to calculate precision and recall:

$$Recall = \frac{\text{Number of annotations common across extracted annotations and gold standard annotations}}{\text{Total number of annotations extracted}}$$

Precision was calculated based on manual verification of GO annotations for 50 randomly selected lncRNAs. The formula used to calculate precision is as follows:

$$Precision = \frac{\text{Number of true annotations extracted}}{\text{Total number of annotations extracted}}$$

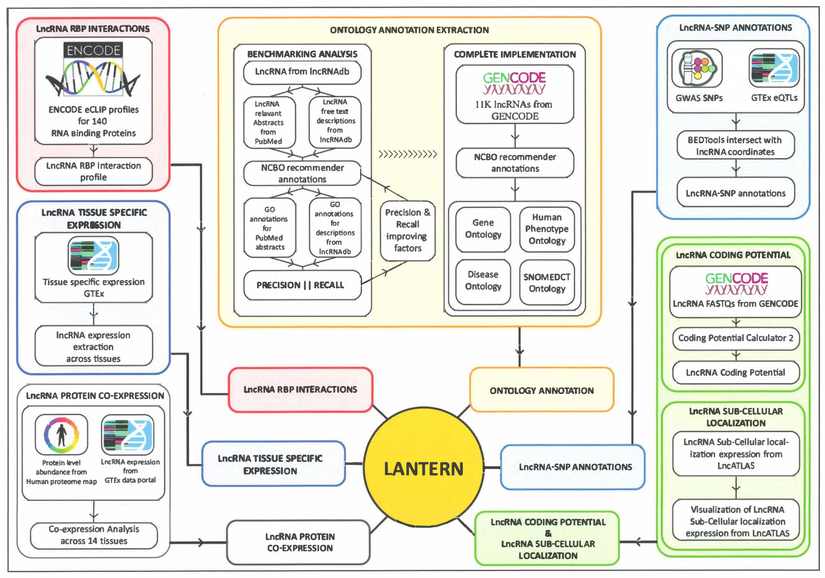

Figure 1: Lantern workflow overview. Flow chart show various NGS and ontology extraction pipelines merging to annotate lncRNAs with multi-level omics and ontology information. All the annotation pipelines were integrated, and the extracted annotations are hosted on the online resource. The ontology annotation extraction pipeline showing two part work flow, first part benchmarks the pipeline developed, and the annotations extracted, second part shows the implementation of the developed pipeline over all the human lncRNAs from GENCODE. Respective flowcharts on either side showing data resources and steps involved in extracting diverse set of omics data for annotating lncRNAs.

RESULTS

- Lantern currently hosts annotation information for 11,290 lncRNAs.
- Around 26% of all the lncRNAs with gene ontology annotations extracted using our novel pipeline were benchmarked.
- Users can search and visualize various levels of lncRNA annotations, information like:
  - where a lncRNA localizes in the cell,
  - which tissue is more representative of which lncRNA,
  - which RBPs interact with which lncRNAs,
  - SNPs on lncRNA specific to a tissue and phenotype and
  - Controlled ontology annotations (gene ontology, disease ontology, human phenotype ontology and SNOMEDCT ontology annotations).

Annotation	Number of lncRNAs with this annotation	Features
Ontology	769	Gene ontology, human phenotype ontology, disease ontology and SNOMEDCT ontology extracted from 6942 abstracts.
RBP-interactions	7942	Across 18 cell lines.
Tissue specific expression	9982	Across 53 tissues.
Protein co-expression	5331	Across 14 tissues.
Coding Potential	9898	For 27907 transcripts.
Sub-cellular localization	11290	For 10 cellular compartments across 15 cell lines.
GTEx eQTL SNPs	6714	Across 48 tissues.
GWAS SNPs	2569	1421 phenotypic characteristics.

Table 1: Various components on Lantern with number of lncRNAs with the respective annotations, along with the domain feature.

Benchmarking of predicted lncRNA annotations from Lantern pipeline reveals high recall and precision

- Recall was computed for the extracted gene ontology annotations of 182 lncRNAs with annotations both sources (PubMed and lncRNAdb)
- Recall was calculated to be 0.62 for gene ontology annotations across 182 lncRNAs.
- Precision was calculated in a non-traditional manner, by randomly selecting 50 lncRNAs, and manually curating annotations. From which the precision was calculated with equation to be 0.8.

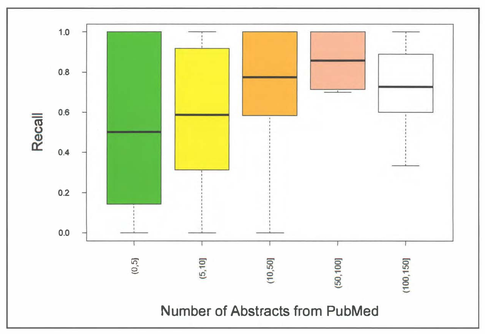

Figure 2: Boxplots showing the distribution of recall values across lncRNAs categorized based on the number of abstracts retrieved per lncRNA from PubMed. The analysis suggests that if more abstracts are available for a lncRNA: annotations extracted by Lantern will have a higher recall. The data shows an optimal peak of recall at 100 to 150 abstracts extracted per lncRNA.

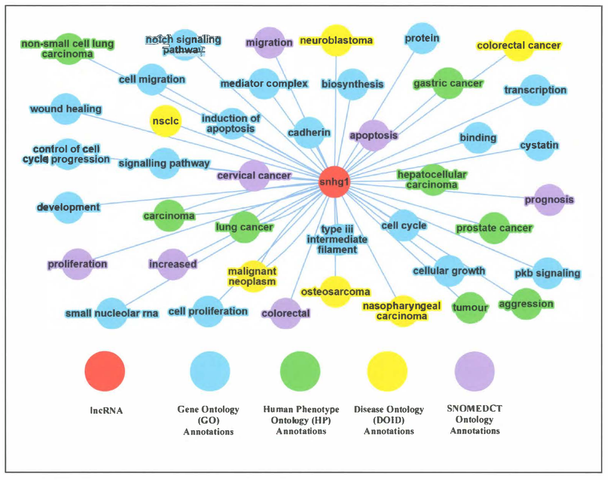

Figure 3: lncRNA ontology annotation network on Lantern. LncRNA and ontology annotation interaction network showing robust ontology annotations of lncRNA SNHG1, extracted from recent literature.

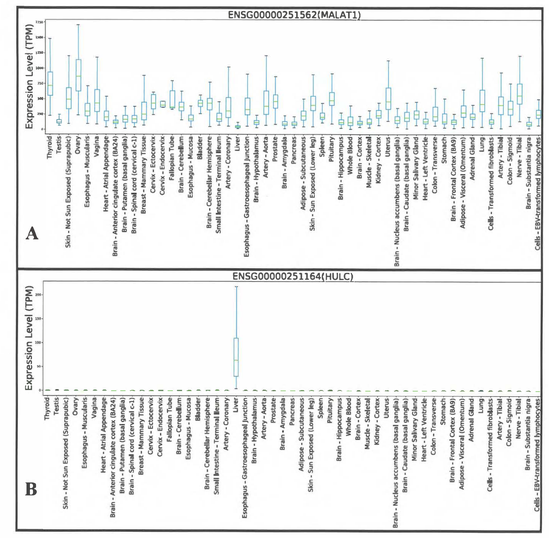

Figure 4: Tissue specific expression of lncRNAs. A: boxplots showing global expression of putative lncRNA MALAT1 across tissues. B: boxplots showing liver specific expression of lncRNA HULC, observed to highly up-regulated in liver tissue.

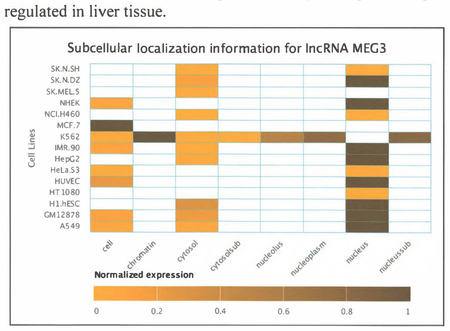

Figure 5: Sub-cellular localization of lncRNAs across cell lines. Figure showing enriched abundance of lncRNA MEG3 in nuclear and chromatin compartments as observed in previous literature.

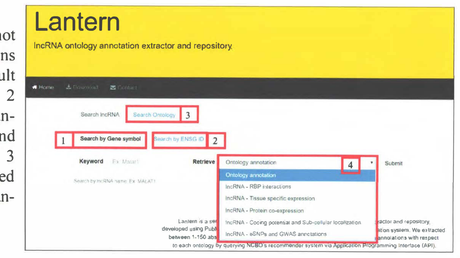

Figure 6: Lantern web interface screenshot showing various lncRNA and annotations searchable options. Box no. 1 showing default search lncRNA by lncRNA name, Box no. 2 showing lncRNA search by ENSG options. Lantern also enable user to browse ontologies and respectively annotated lncRNAs, Box no. 3 highlighting the button for ontology based search. Box no. 4 highlights all the available annotations dropdown.

CONCLUSION

- A semi-automated pipeline for annotating lncRNA with ontology information using literature mining was developed.
- The extracted ontology annotations were benchmarked against a manually curated lncRNAdb.
- All the extracted ontology information is reposited onto a web interface to easily navigate and retrieve ontology annotations for lncRNA.
- A molecular level interaction profile of lncRNAs was put together, along with tissue specific expression, eQTLs and SNP information utilizing contemporary NGS data analysis pipelines.
- Subsequently, a public resource with high quality-controlled ontology annotations and comprehensive omics annotations was developed for improving the annotation of the noncoding transcriptome.

REFERENCES

1. Mohanty, V., et al., Role of lncRNAs in health and disease—size and shape matter. Briefings in functional genomics, 2014. 14(2): p. 115-129.
2. Quinn, J.J. and H.Y. Chang, Unique features of long non-coding RNA biogenesis and function. Nature Reviews. Genetics, 2016. 17(1): p. 47.
3. Martínez-Romero, M., et al., NCBO Ontology Recommender 2.0: an enhanced approach for biomedical ontology recommendation. Journal of biomedical semantics, 2017. 8(1): p. 21.
4. Quek, X.C., et al., lncRNAdb v2. 0: expanding the reference database for functional long noncoding RNAs. Nucleic acids research, 2014. 43(D1): p. D168-D173.

Acknowledgment: The authors would like to thank members of the Janga lab for helpful discussions in the course of this project and support from National Institute of General Medical Sciences (NIGMS) under Award Number R01GM123314 to SCJ.