Systematic tissue annotations of genomic samples by modeling unstructured metadata

Nathaniel T. Hawkins[1], Marc Maldaver[1], Lindsay A. Guare[1,2,3], Arjun Krishnan[1,2]

[1]Dept. of Computational Mathematics, Science and Engineering. Michigan State University. East Lansing, MI, United States.
[2]Dept. of Biochemistry and Molecular Biology. Michigan State University. East Lansing, MI, United States.
[3]Dept. of Microbiology and Molecular Genetics. Michigan State University. East Lansing, MI, United States.

hawki235@msu.edu
www.nathawkins.info
thekirshnanlab.org

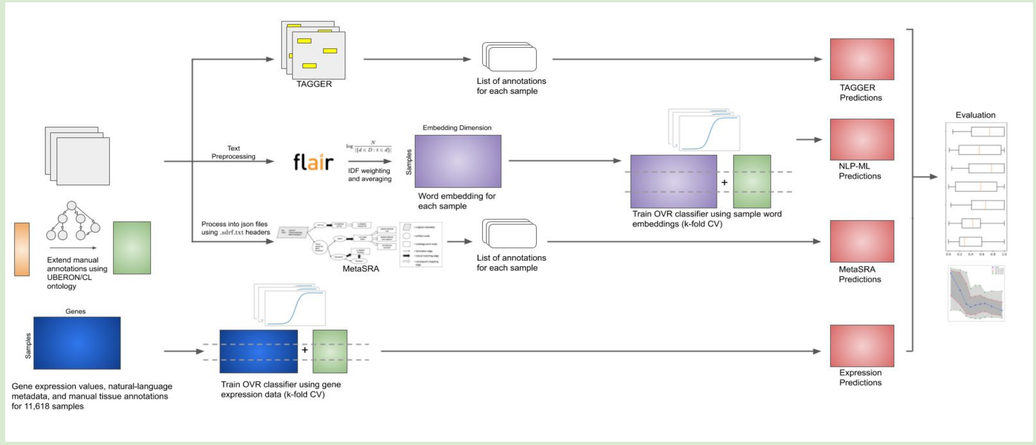
CMSE

MICHIGAN STATE UNIVERSITY

OVERVIEW

The reuse of publicly available omics data is limited by the lack of systematic annotations for tissues or cell types of origin. Previous work has used the underlying omics data to predict annotations for a given sample, but these tools are not widely usable due to the need to retrain for different data types. We propose a novel approach that leverages free-text, natural-language metadata to identify the tissue or cell-type of origin of omics samples. Using sample metadata files from ArrayExpress, we create a numerical representation for each sample using flair, a python NLP library. These representations are used train OVR logistic regression models, one for each tissue or cell type we have labels for, to predict a sample's annotations given its metadata. We compare our models' predictions to annotations predicted by models trained on gene expression data and those from two existing tools: TAGGER, an NER tool, and MetaSRA, a method that uses prior knowledge and ontologies.

WORKFLOW



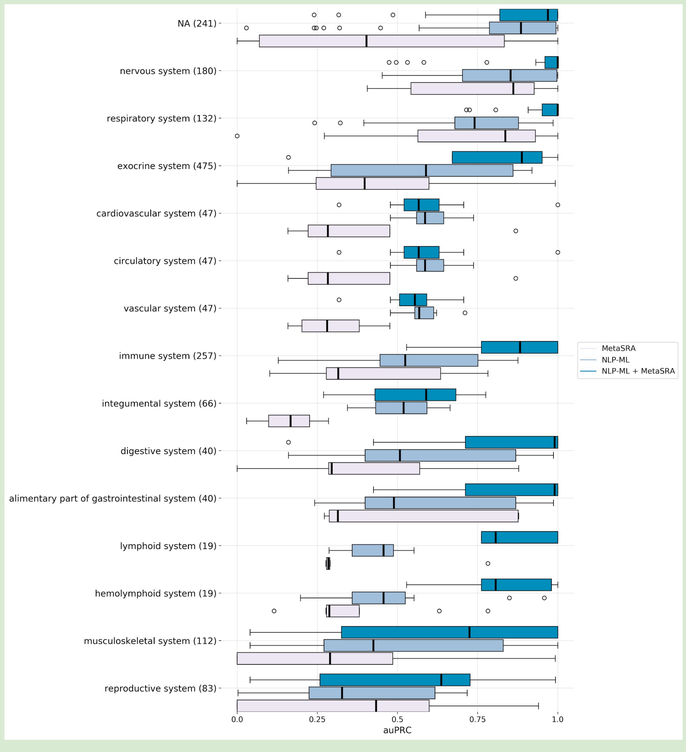
MODEL PERFORMANCE ON NOVEL DATA TYPES

	RNA-seq of Coding RNA	ChIP-seq	Methylation Profiling by Array	Comparative Genomic Hybridization by Array	Transcription Profiling by Array	Sum
Adipose Tissue	7	8	10	3	10	38
Brain	10	10	10	9	10	49
Colon	6	10	10	5	9	40
Neural Tube	10	10	10	7	9	46
Muscle Tissue	9	0	10	0	10	29

We randomly selected 10 samples from 5 data types and evaluated the predictions from 5 of our top performing model. The average precision of our models on this task is 80%. The number of correct predictions for each model is shown above. Our models can make accurate predictions on new data types without needing to retrain.

COMPARISON OF TEXT-BASED METHODS

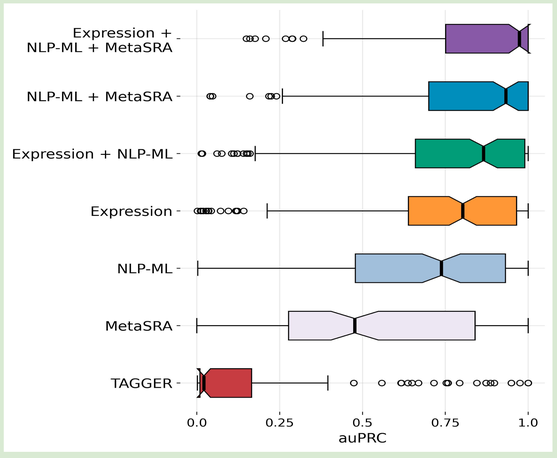
Our method outperforms both TAGGER and MetaSRA. The combined predictions of both MetaSRA and our NLP-based ML outperforms all other text-based methods. This is consistent across anatomical systems as well.



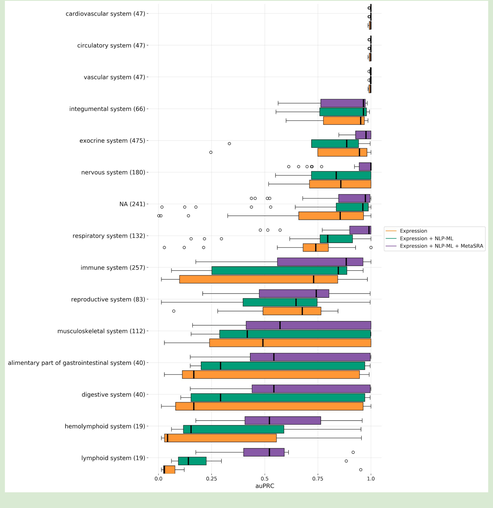
1. Bernstein et al., 2017, *Bioinformatics*, 33, 18, 2914-2923.
2. Lee et al., 2013, *Bioinformatics*, 29, 23, 3036–3044.
3. Jensen, 2016, *bioRxiv*, doi.org/10.1101/067132.
4. Akbik et al., 2019, *NAACL Proceedings,* 724–728.
5. Basha et al., 2020, *Bioinformatics*, 36, 9, 2821-2828.

This work was primarily supported by US National Institutes of Health (NIH) grants R35 GM128765 to AK and in part by MSU start-up funds to AK and MSU Rasmussen Doctoral Recruitment Award and Engineering Distinguished Fellowship to NTH.

COMPARISON TO AND COMBINATION WITH EXPRESSION-BASED METHODS

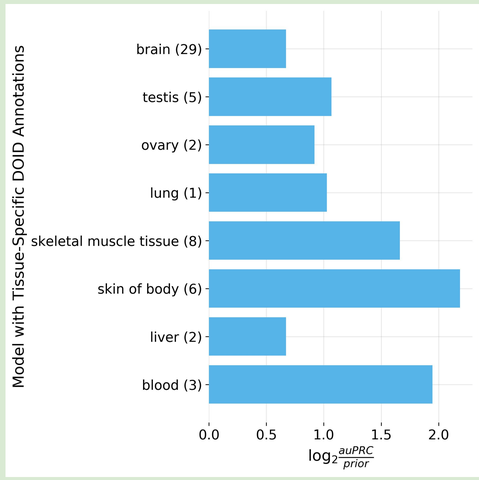


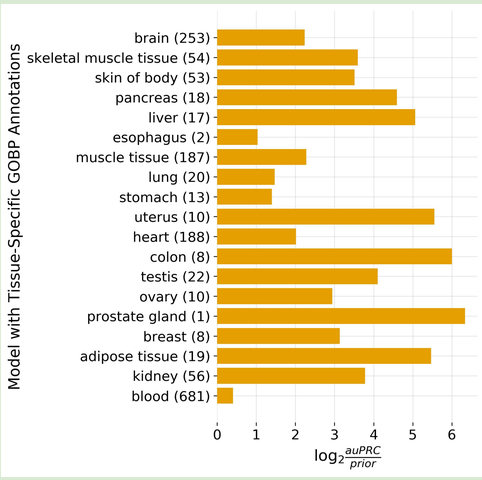
Expression-based ML outperforms all text based methods. Adding the predictions from our NLP-ML improves performance over expression alone, and is further improved by incorporating MetaSRA. Text-based methods therefore capture a unique and meaningful signal that expression does not. This is consistent across anatomical systems.



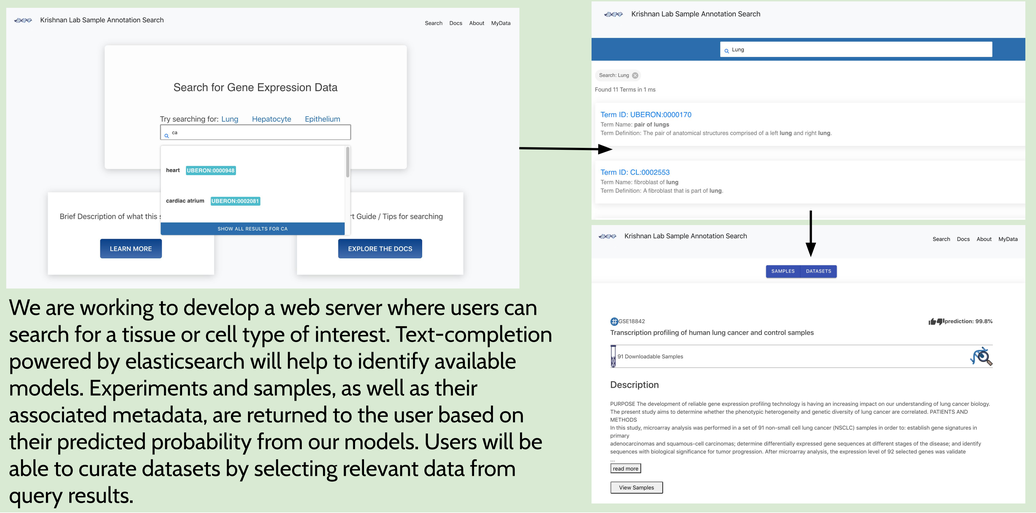
TISSUE-SPECIFIC GO BIOLOGICAL PROCESS AND DISEASE ONTOLOGY TERMS

Model performances on list of embeddings created from plain-text definitions for curated tissue-specific GOBP and DOID terms. Our models capture biologically-relevant signals from text alone.





WEB SERVER DEVELOPMENT



We are working to develop a web server where users can search for a tissue or cell type of interest. Text-completion powered by elasticsearch will help to identify available models. Experiments and samples, as well as their associated metadata, are returned to the user based on their predicted probability from our models. Users will be able to curate datasets by selecting relevant data from query results.