Inferring disease-associated lncRNAs using expression data and disease-associated protein coding genes

Xiaoyong Pan

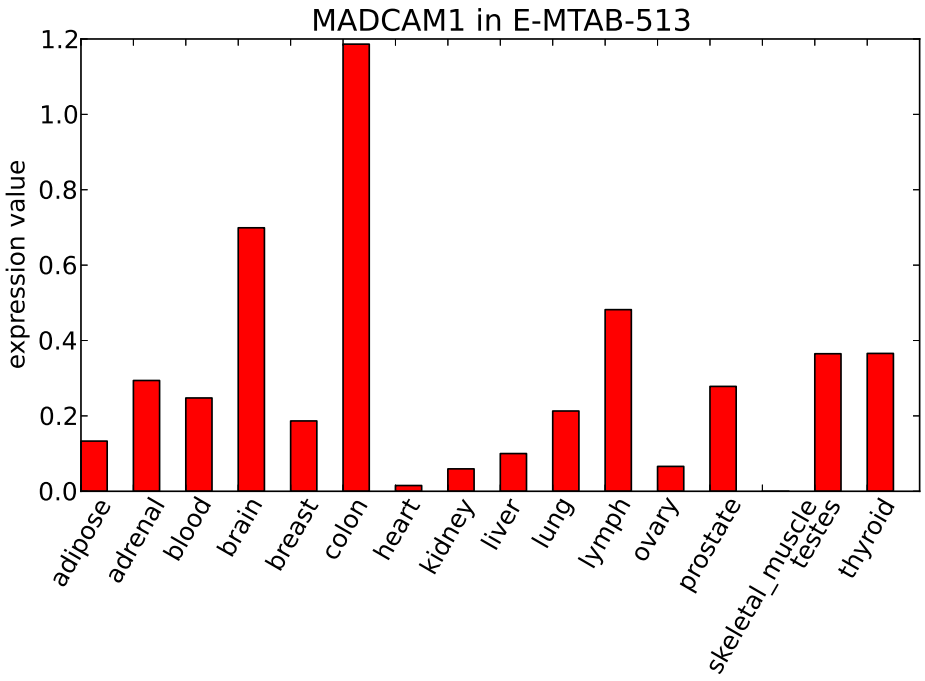
Center for non-coding RNA in Technology and Health,
Department of Clinical Veterinary and Animal Science,
University of Copenhagen
Department of Disease Systems Biology,
Novo Nordisk FoundationCenter for Protein Research,
University of Copenhagen, Denmark.

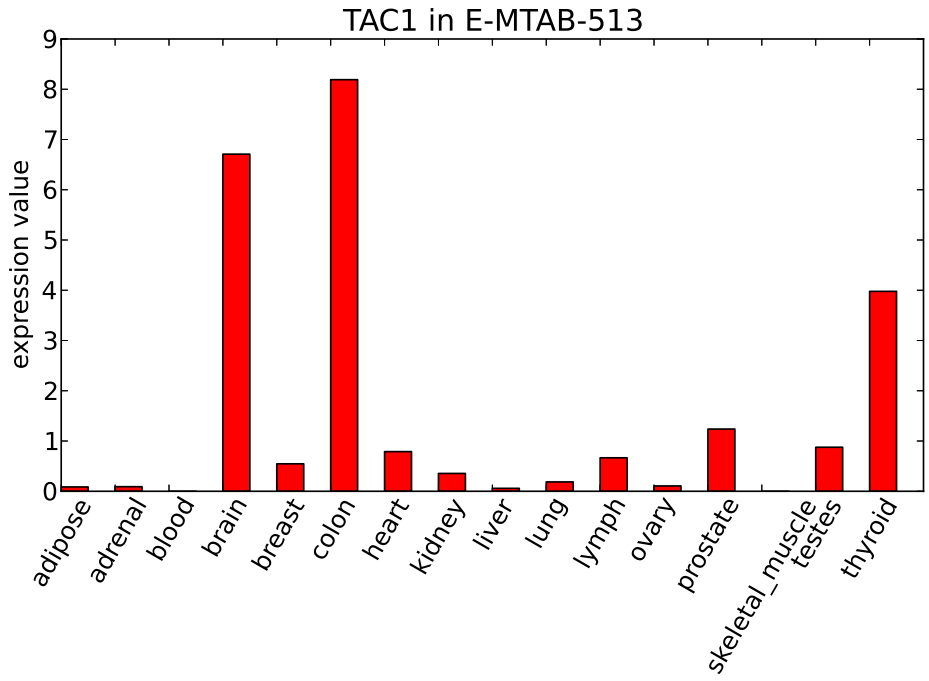
February 20, 2016

Introduction

1. lncRNAs are emerging as important regulator in different diseases.
2. Gene is more tissue-specific than individual-specific [Melé 2015, Science]
3. Different diseases relevant to specific tissues[Lage 2008, PNAS]
4. Disease-associated genes have similar expression pattern.

Two genes associated with inflammatory bowel disease





Tissue

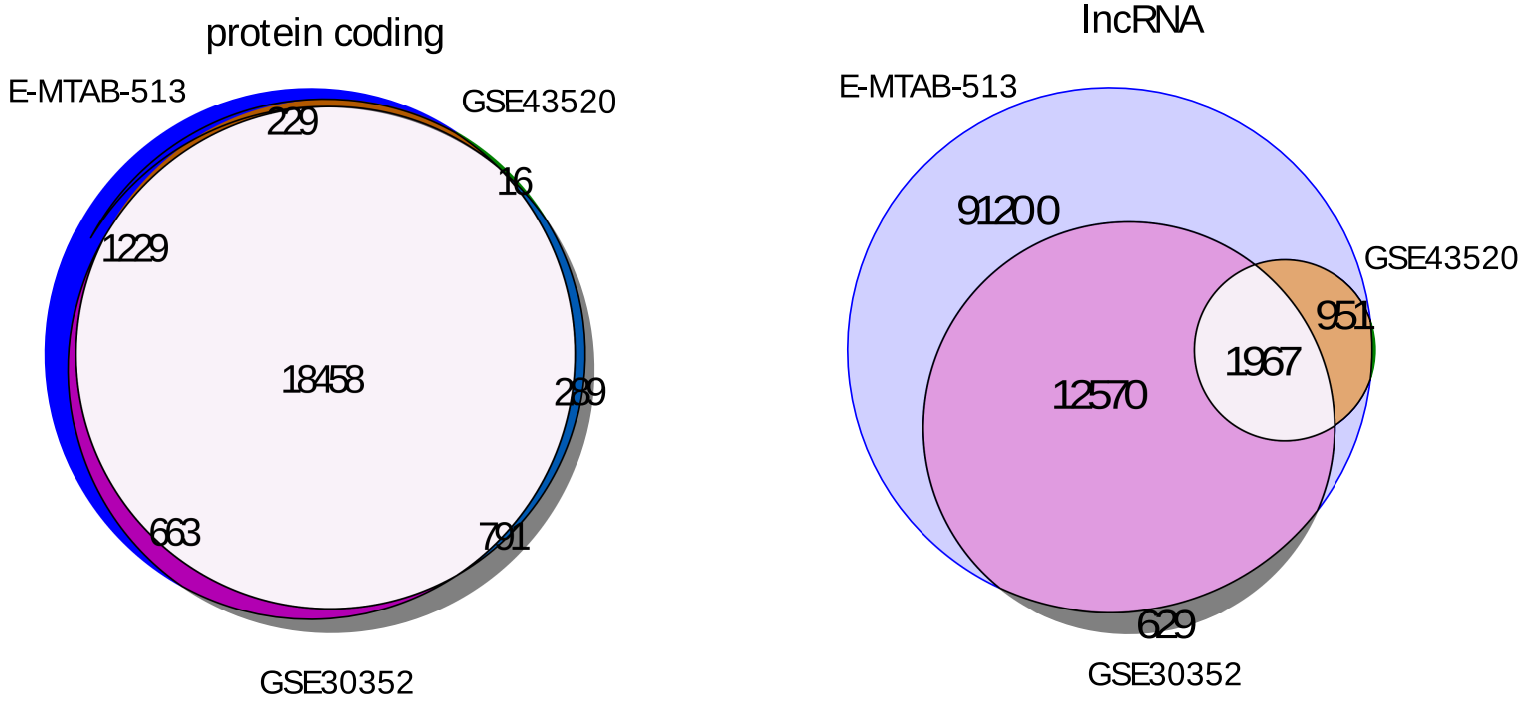
Project goal

1. Infer disease-associated lncRNAs from protein coding and lncRNA co-expression dataset.

2. **Method**: Train random forest model on disease-associated protein coding gene expression profiles, then predict for lncRNA.

Dataset

1. 3 RNAseq datasets:
   - E-MTAB-513, human body map, 16 tissues, GENCODE v7.
   - GSE43520, evolution of lncRNA in tetrapods, 15 tissues, highly conserved lncRNAs.
   - GSE30352, evolution of genes in mammalians, 6 tissues, Ensembl based annotation.



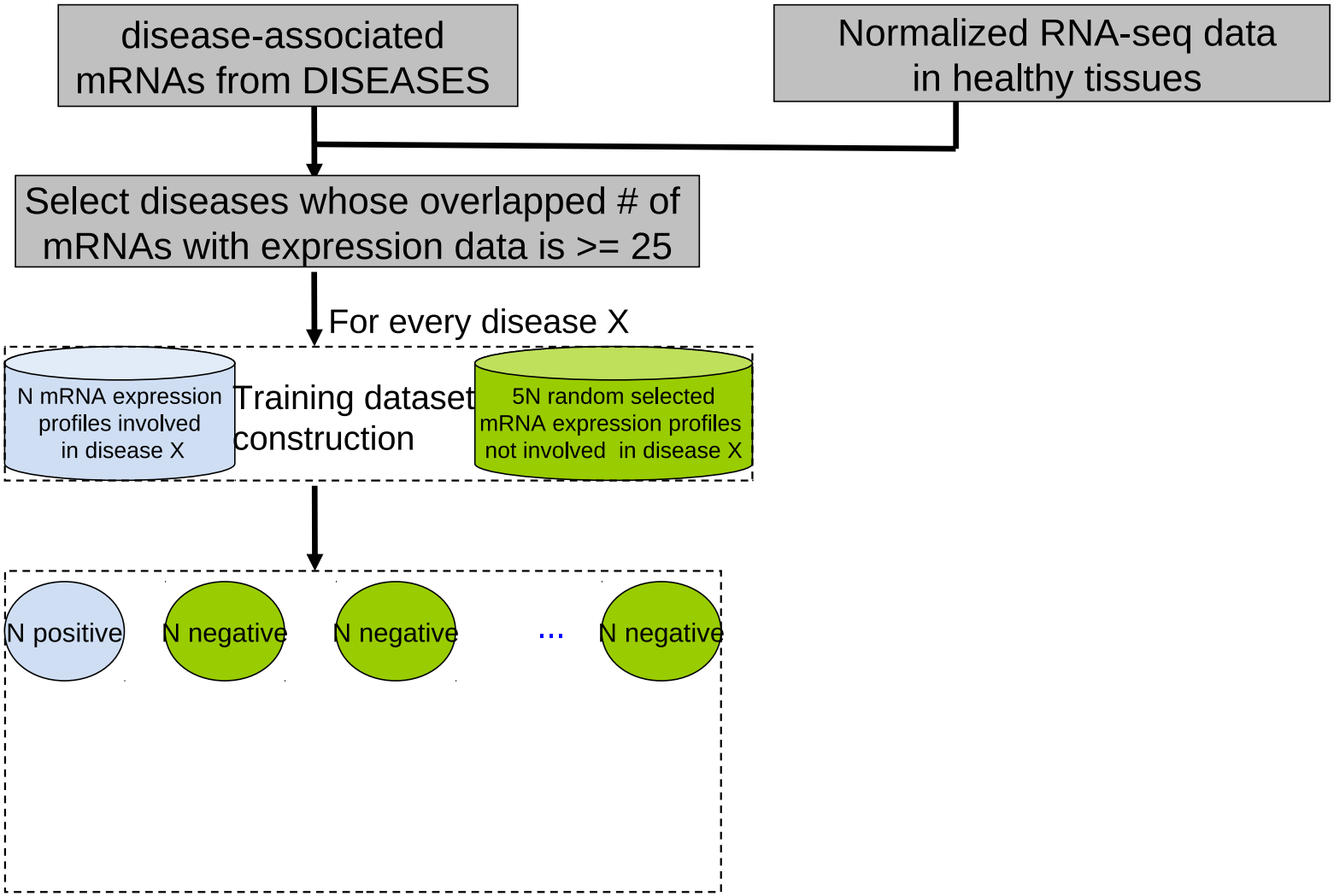
[Derrien 2012, Genome Res], [Necsulea 2014, Nature], [Brawand 2011, Nature]

Dataset

1. DISEASES [Pletscher-Frankild 2015, Methods] database for disease-associated protein coding genes.
   - 543,405 associations between 17,606 genes and 4,610 diseases.

2. LncRNADisease [Chen 2013, Nucleic Acids Res] for verified disease-associated lncRNAs.
   - More than 1,000 association between 321 lncRNAs and 221 diseases.

Infer disease-associated lncRNA from co-expression profile



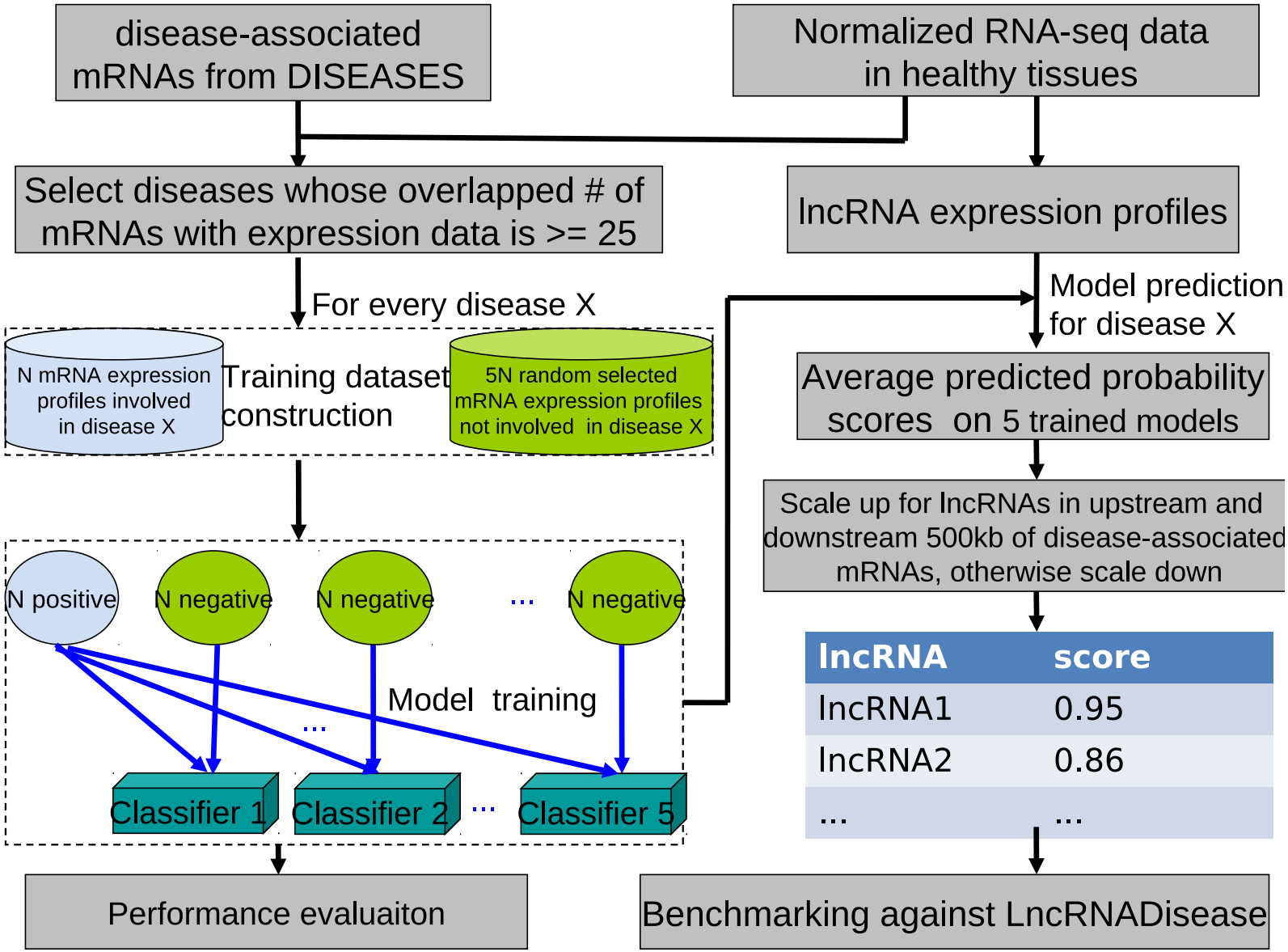
Constructing training data

1. For each disease X:
   - Randomly sampling 5N mRNAs subsets not involved in disease X with 5 times the number of mRNAs associated with disease X.
   - assign label 1 to disease-associated mRNAs, otherwise 0.

Label		adipose	adrenal	blood	brain	breast	colon	heart	kidney	liver
1	PFKP	0.95426	0.64932	0.04886	0.65482	0.36988	0.340933	0.08663	0.131352	0.06026
1	GALT	0	0.04175	0.01892	0.1727	0.01518	5.43E-09	0	0.031027	0
1	ACAA1	49.50935	188.05339	23.36094	44.89562	101.19716	33.72076	31.20994	57.01054	19.12016
1	7SK	105.66857	128.62117	4.06798	113.55834	124.79909	48.27836	13.8542	42.67313	4.27873
...	...	...	...	...	...	...	...	...	...	...
0	A-575C2.4	1.38353	4.154284	0.84911	2.68101	2.292396	0.92442	0.80375	1.66876	0.86089
0	A1BG	3.74192	6.41188	5.59976	2.97937	1.36397	4.04595	0.55942	1.77137	392.052
0	A1BG-AS1	0.87483	2.75723	2.50365	1.19136	0.52167	1.03712	0.62707	1.83091	15.2022
0	A1CF	0	0.03268	0	0.02983	0	0.536022	0	1.672	17.6027

2. NOTICE: when randomly select negative mRNAs, this mRNAs should be associated with other diseases (well studied genes instead of rare genes), have no any evidence for disease X in DISEASES database.

Infer disease-associated lncRNA from co-expression profile



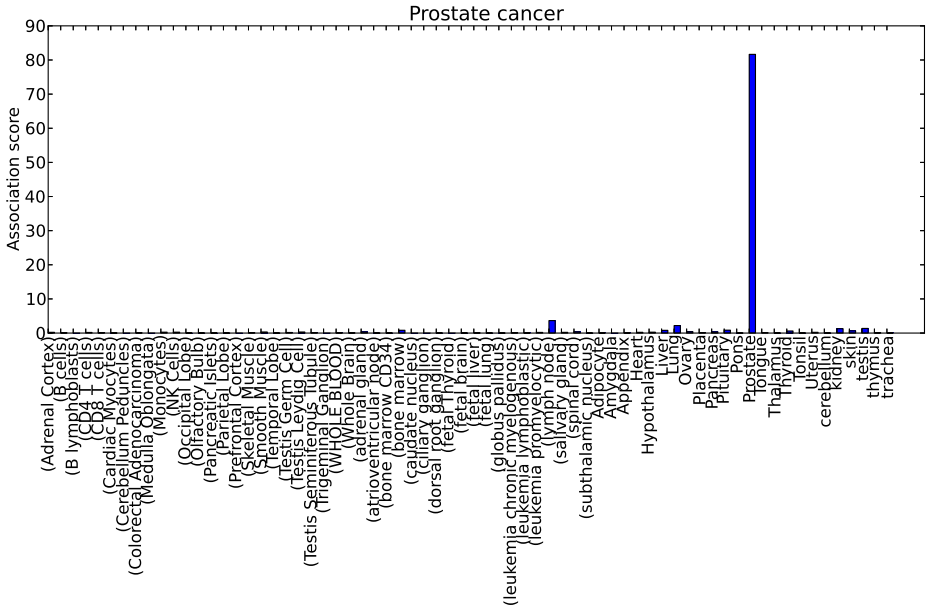
Performance using mRNA expression profile

Table: Average performance for diseases, whose # of associated mRNAs overlapping with mRNAs in expression data is greater than 25 in DISEASES database, using mRNA expression profiles.

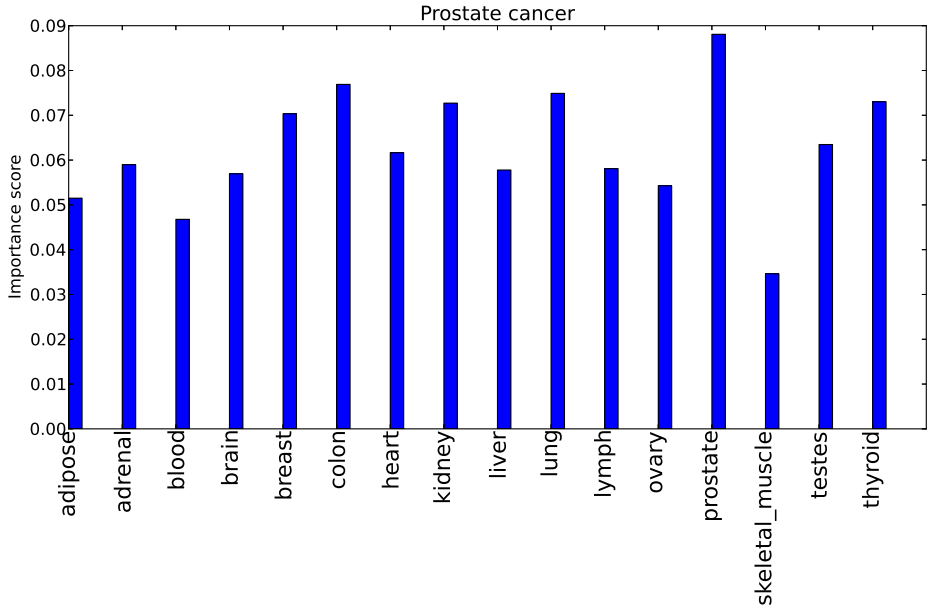
Dataset	# diseases	Accuracy	Sensitivity	Specificity	Precision	MCC
E-MTAB-513	114	0.851	0.568	0.895	0.842	0.631
GSE43520	114	0.864	0.589	0.917	0.854	0.661
GSE30352	120	0.831	0.536	0.905	0.817	0.607

Tissue importance for disease-associated gene classification

- disease-tissue association score[Lage 2008, PNAS].
- Random forest feature importance analysis.
- Expression value in tissues are features, tissue important score is ranked by random forest.



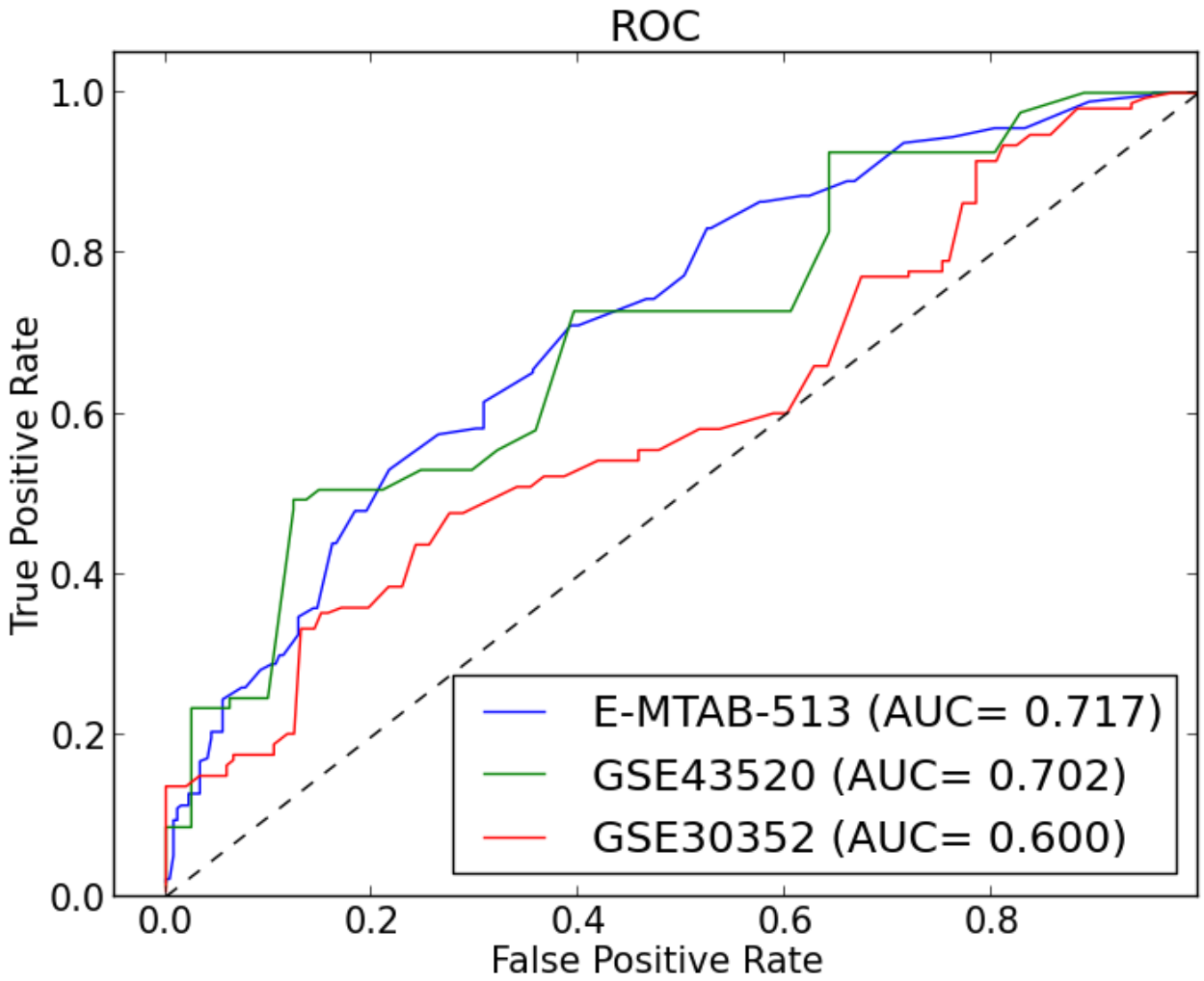
Lage 2008, PNAS



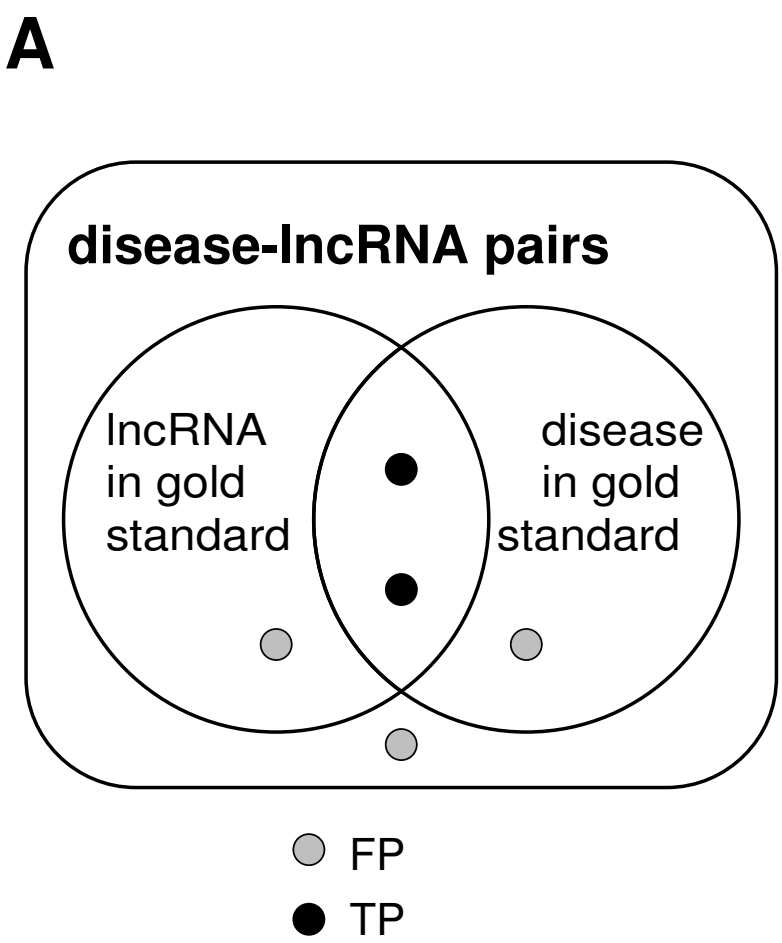
Tissue

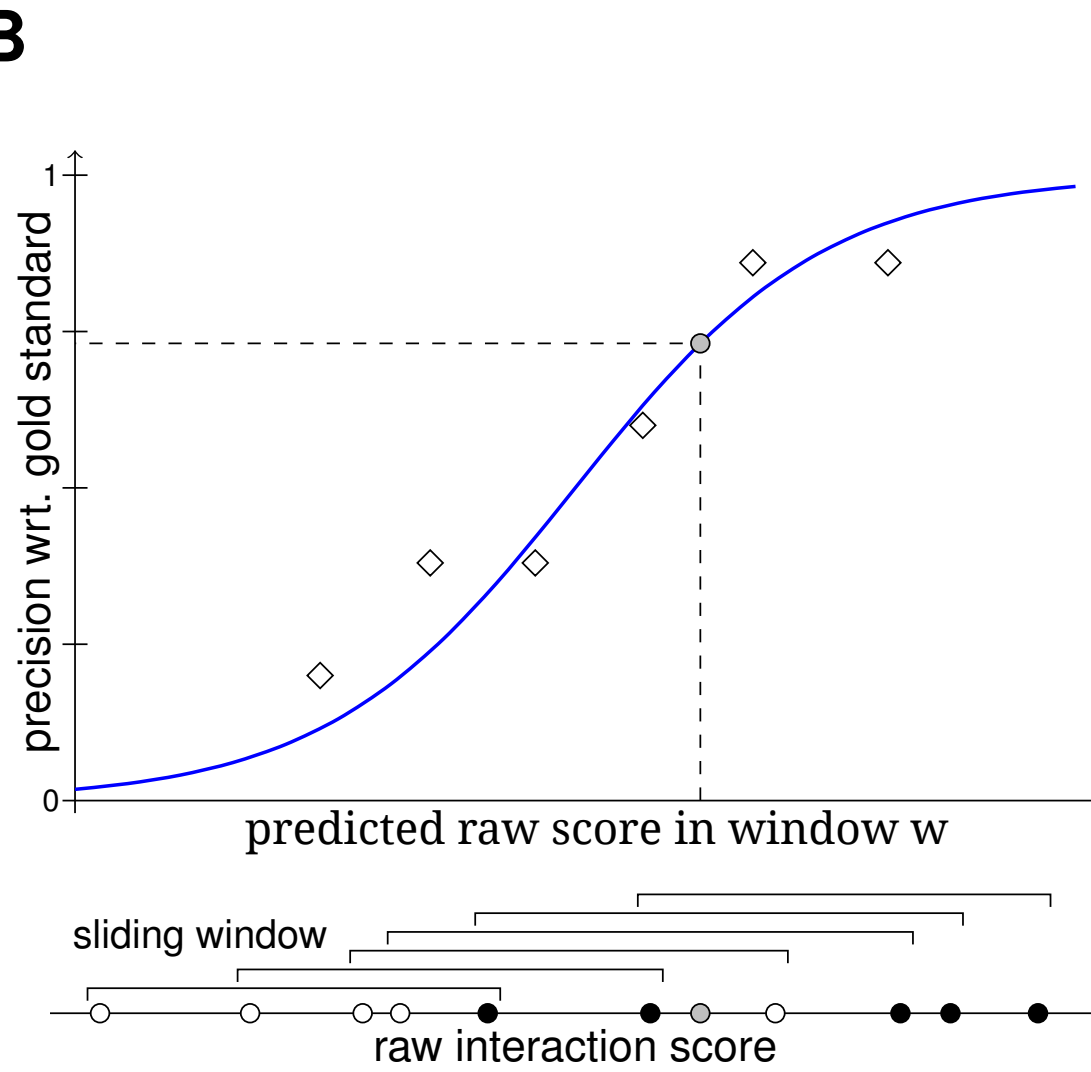
Infer disease-associated lncRNA

1. For each association in LncRNADisease, randomly select another lncRNA for this disease as negative pair.

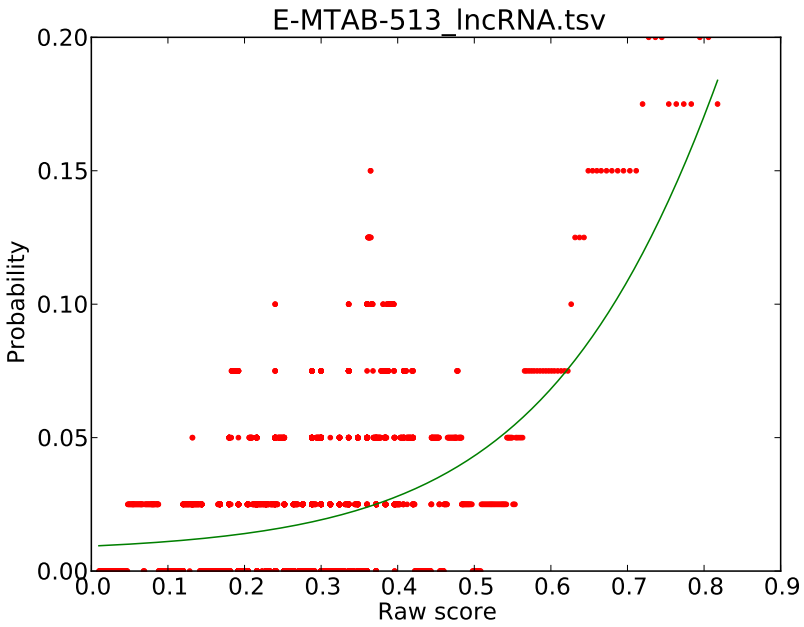


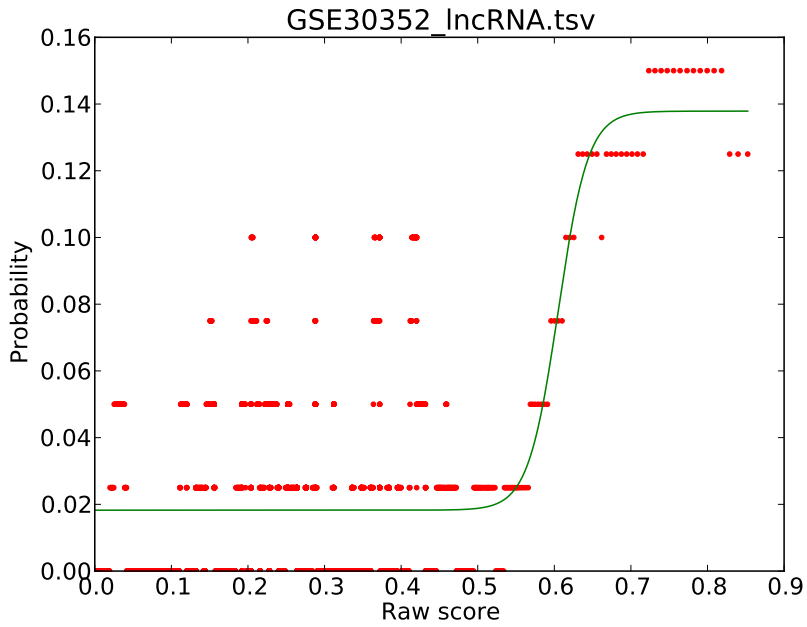
Benchmarking predicted disease-association lncRNAs

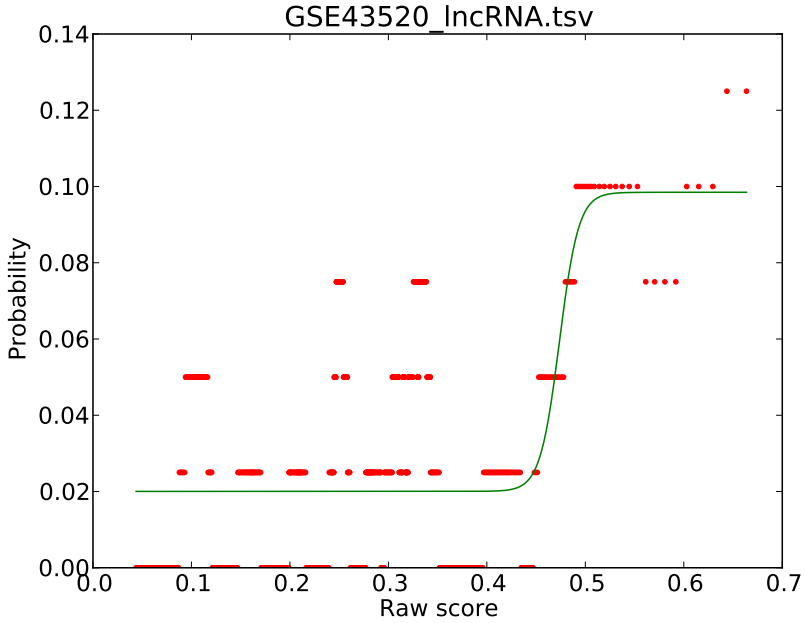




Benchmark against disease-lncRNA from LncRNADisease







Conclusion and outlook

1. In this study, we infer disease-associated lncRNA from expression data and disease-associated protein coding genes.
2. Integrate GWAS SNP data with predicted score to prioritize disease-associated lncRNAs.
3. Text mining disease-lncRNA associations and compared our prediction to it.
4. How to better select negative genes for model training.

Acknowledge!

Lars Juhl Jensen

Jan Gorodkin

RTH and DSB group

Funding: Innovation fund Denmark
PhD scholarship from from Faculty of Health and Medical Science, University of Copenhagen

Thanks for your attention!