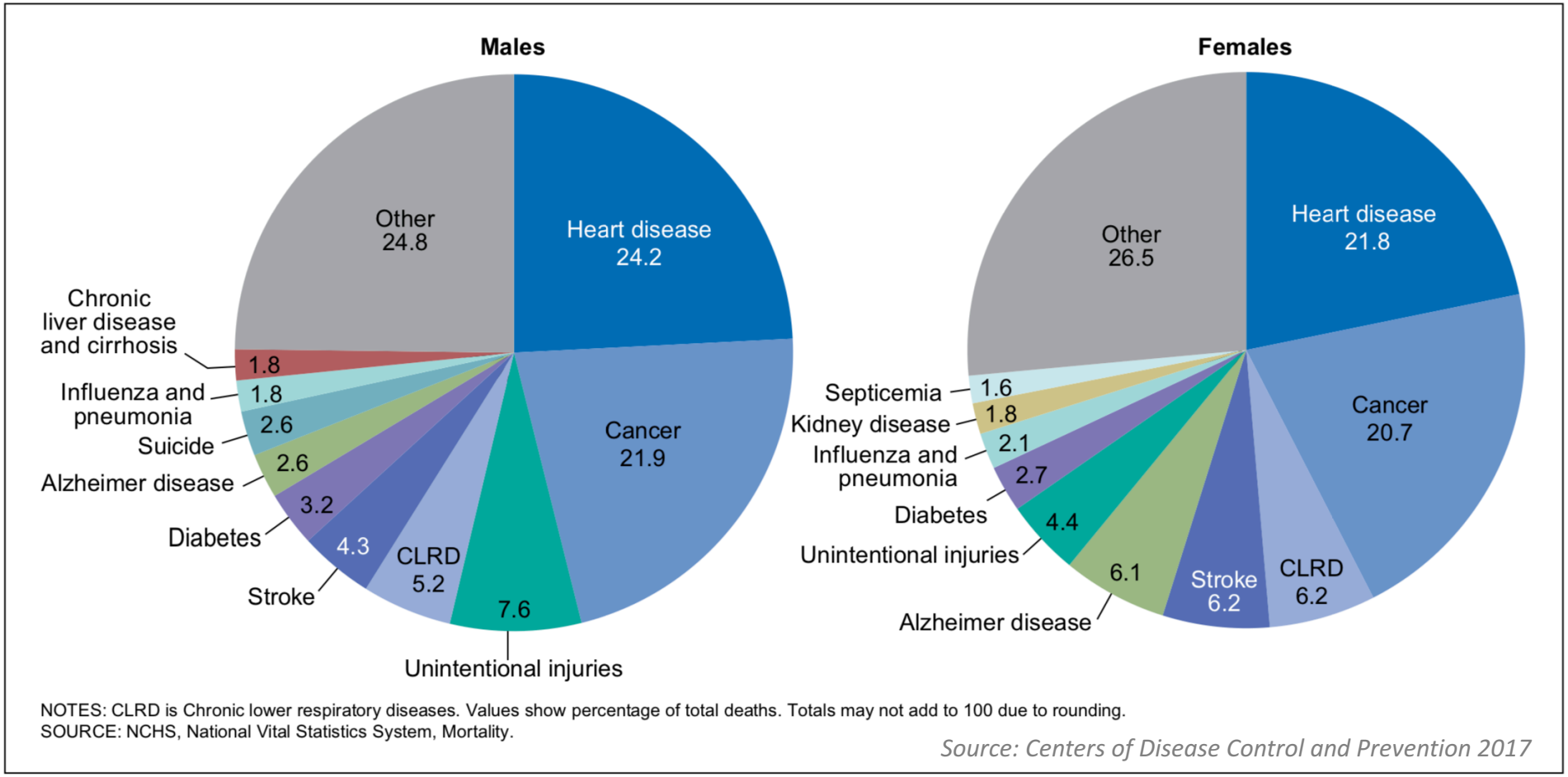
FIU Engineering & Computing
School of Computing & Information Sciences

# Feature Selection and Classification Reveal key lncRNAs for Multiple Cancers

**Abdullah Al Mamun**

Machine Learning and Data Analytics Group (MLDAG)
Florida International University, Miami, FL, USA

# Leading Cause of Death in US




NOTES: CLRD is Chronic lower respiratory diseases. Values show percentage of total deaths. Totals may not add to 100 due to rounding.
SOURCE: NCHS, National Vital Statistics System, Mortality.

*Source: Centers of Disease Control and Prevention 2017*

# Expected Statistics in US

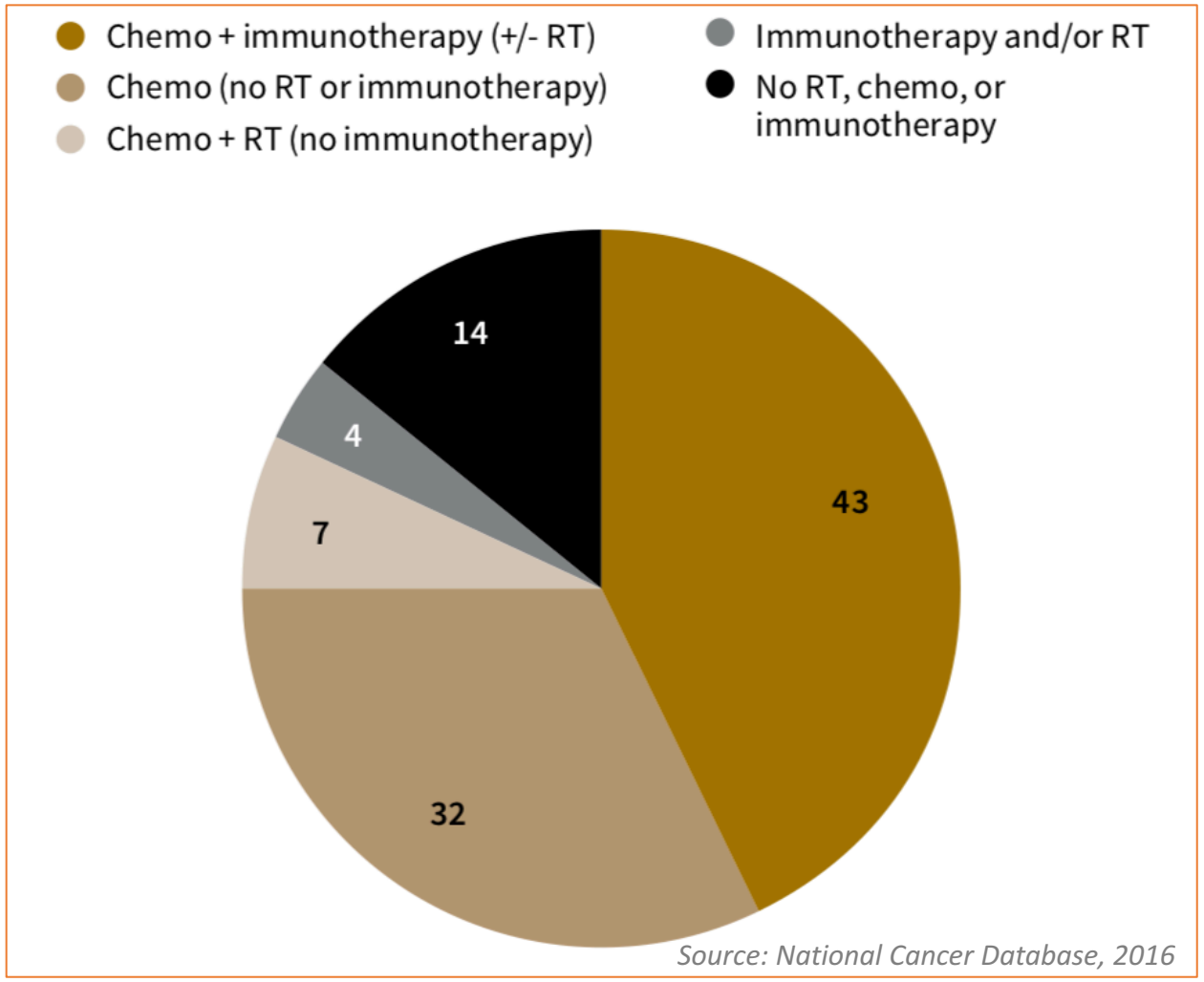
| | |
|---|---|
| Estimated New Cases in 2019 | 1,762,450 |
| % of All New Cancer Cases | 100.0% |

| | |
|---|---|
| Estimated Deaths in 2019 | 606,880 |
| % of All Cancer Deaths | 100.0% |

| Percent Surviving 5 Years |
|---|
| **67.1%** |
| 2009-2015 |

*Source: National Cancer Institute 2019*

# Targeted Therapy Treatment




**Breast cancer:**
Human epidermal growth factor receptor 2 (***HER2***)

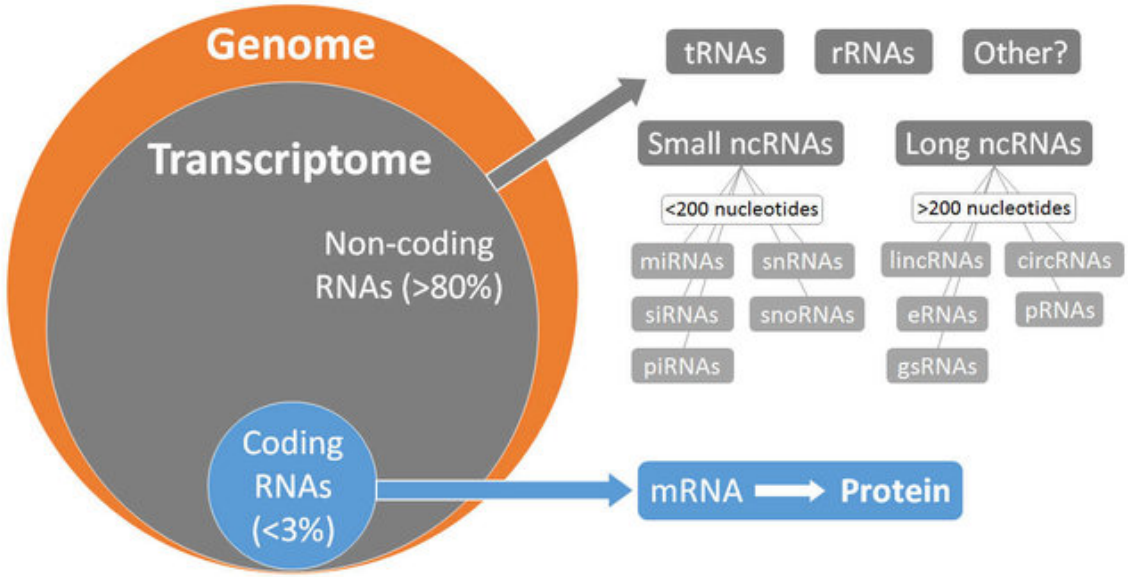
**Colorectal cancer:**
Epidermal growth factor receptor (***EGFR***)

**Lung cancer:** *EGFR*

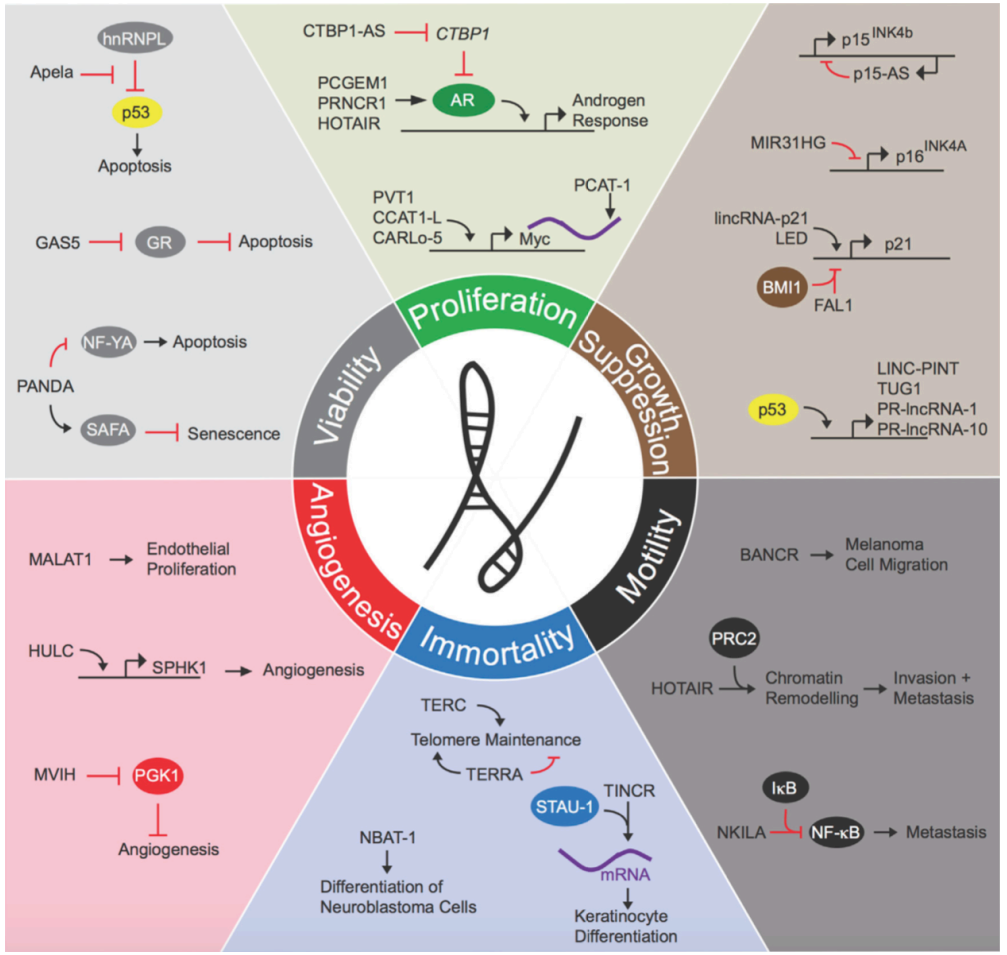
**Melanoma:** *BRAF*

***What about other RNAs?***

# lncRNAs in Cancer Pathways



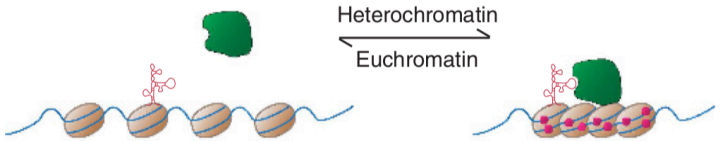

*Source: Hubé, Florent et al., 2018*




*Schmitt, Adam M. et al., 2016*

# lncRNAs in Cancer (*Cheetham, S. W., et al, 2013*)

Chromatin Remodeling

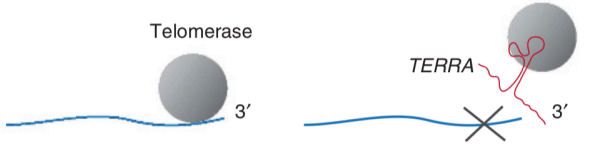



***ANRIL*** ***XIST*** ***HOTAIR***

Transcriptional co-activation

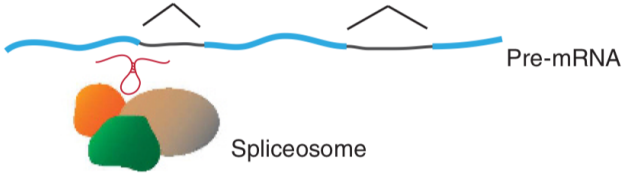

***H19*** ***SRA***

Protein Inhibition




***TERRA***

Post-transcriptional modifications




***MALAT1***

# Aim

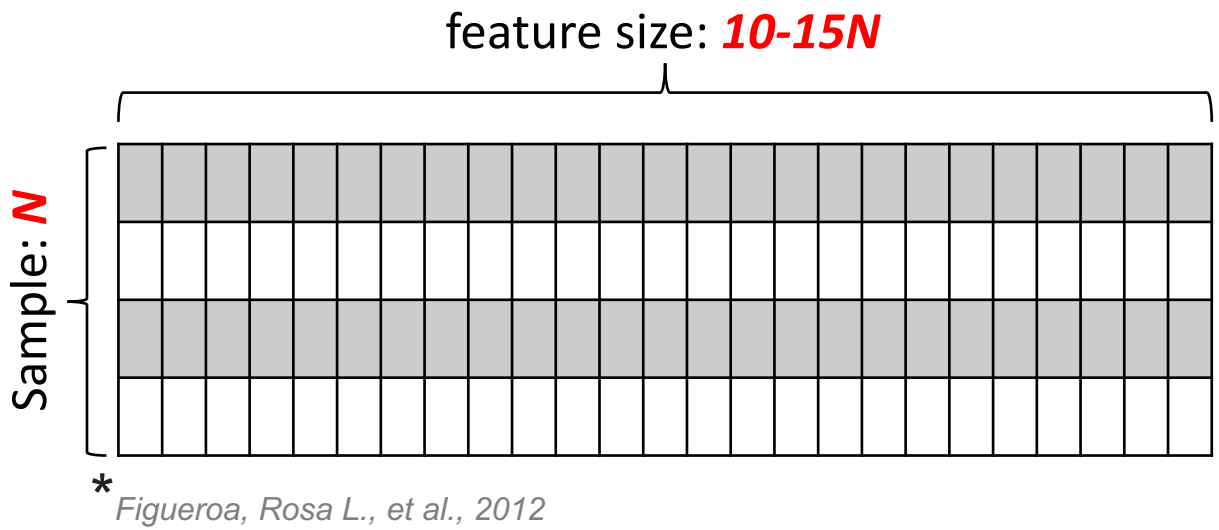
- ❑ Identifying cancer specific key lncRNAs
- ❑ Cancer classification based on those important lncRNAs expression

# Limitation and Challenges

**1. DEG**

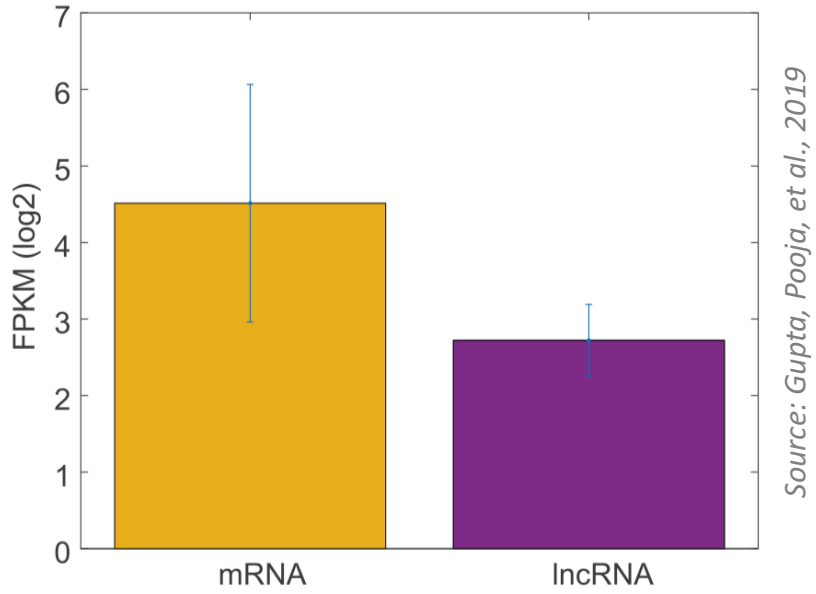
- Need case and control

**2. Optimal feature* size should be $\sqrt{N}$ *to* $N$**




*Figueroa, Rosa L., et al., 2012*

**3. Expression level: *Low***



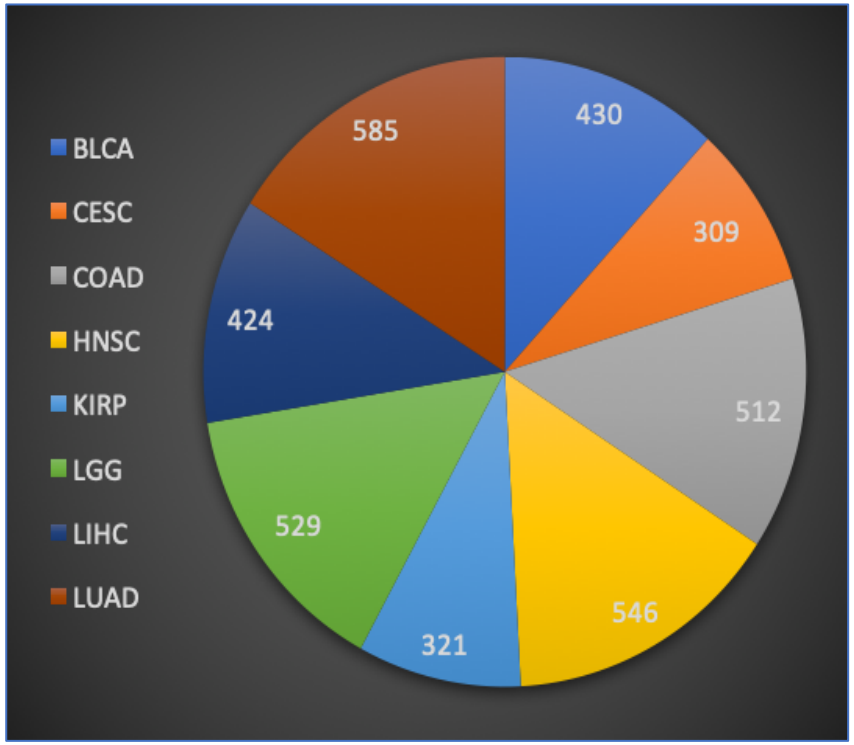

**4. Cancer specific lncRNA**

# Solution

- ***Feature Selection*** → **Identify key lncRNAs**
- ***Machine Learning*** → **Cancer Classification**

# Samples and its Distribution

| Tumor Types | Short Name | # Datasets |
|---|---|---|
| Bladder Cancer | BLCA | 12 |
| Cervical Cancer | CESC | 12 |
| Colon Cancer | COAD | 13 |
| Head and Neck Cancer | HNSC | 12 |
| Kidney Papillary Cell Carcinoma | KIRP | 13 |
| Lower Grade Glioma | LGG | 12 |
| Liver Cancer | LIHC | 12 |
| Lung Adenocarcinoma | LUAD | 13 |




Data source: TCGA
8 cancer types with dimension : *3656 * 12309*

# LASSO: Least Absolute Selection and Shrinkage Operator

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon.$$

- Imposes a constraint on the sum of the absolute values of the model parameters
- This constraint causes regression coefficients for some variables to shrink towards zero
- Picks the variables those are not zero
- Identifies the variables most strongly associated with the target

# Why LASSO?

- It can provide greater prediction accuracy
- Good for small number of observations
- It can increase model interpretability
  - Overfitting causes difficulty for model interpret
  - With Lasso, the coefficients for unimportant variables are reduced to zero that selects only the most important predictors.
- Tuning parameter: lamda
- Bias increases and variance decreases as lambda increases.

# RFE

- Goal of recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features.
- First, the estimator is trained on the initial set of features and the importance of each feature is obtained either through from coefficient or through a feature importance.
- Then, the least important features are pruned from current set of features.
- That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.

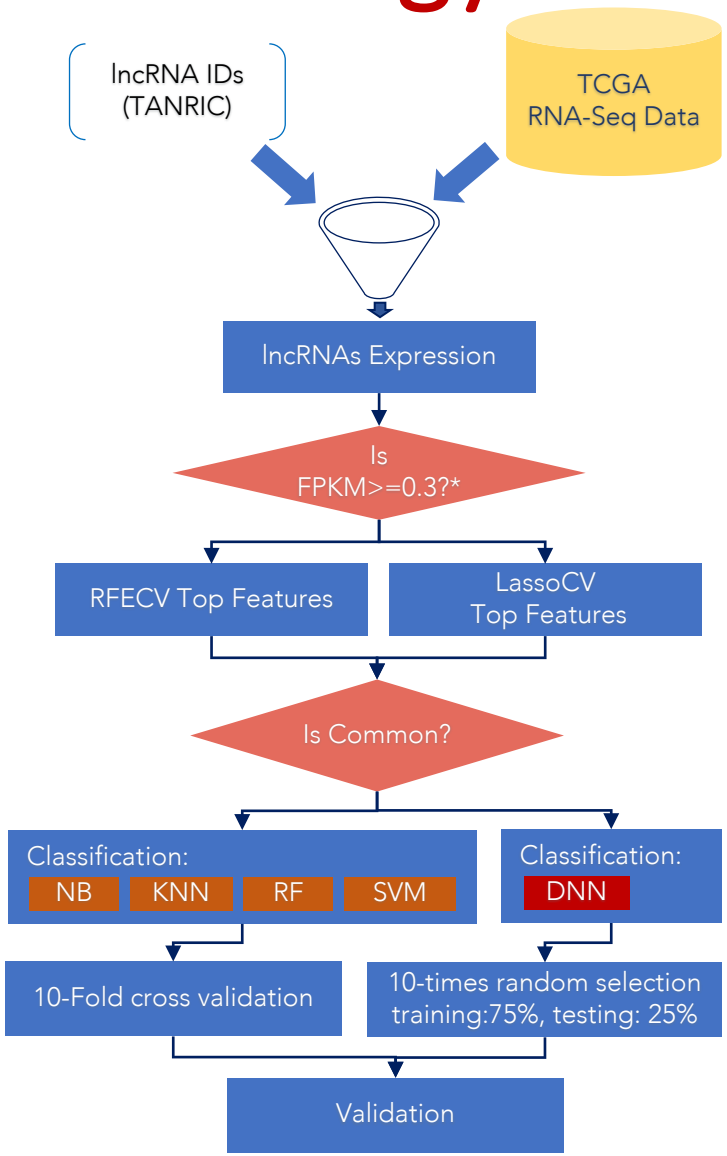
# Methodology

## Feature Selection
- *LASSO*
- *RFE*

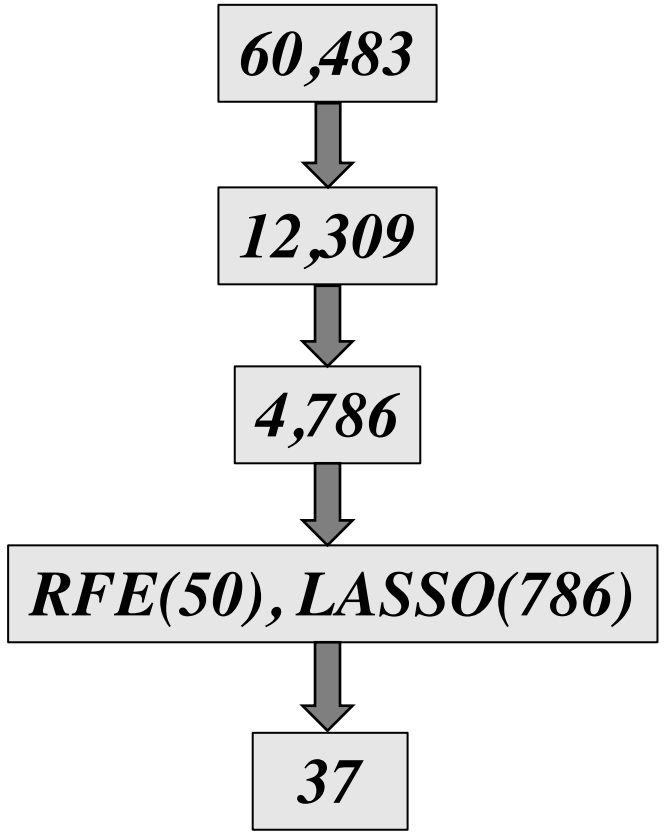
## Classification
- *NB, KNN, RF, SVM, DNN*

## Parameter Tuning
- *Grid Search*

## Validation
- *tSNE*
- *Survival Analysis*




**Feature Size**



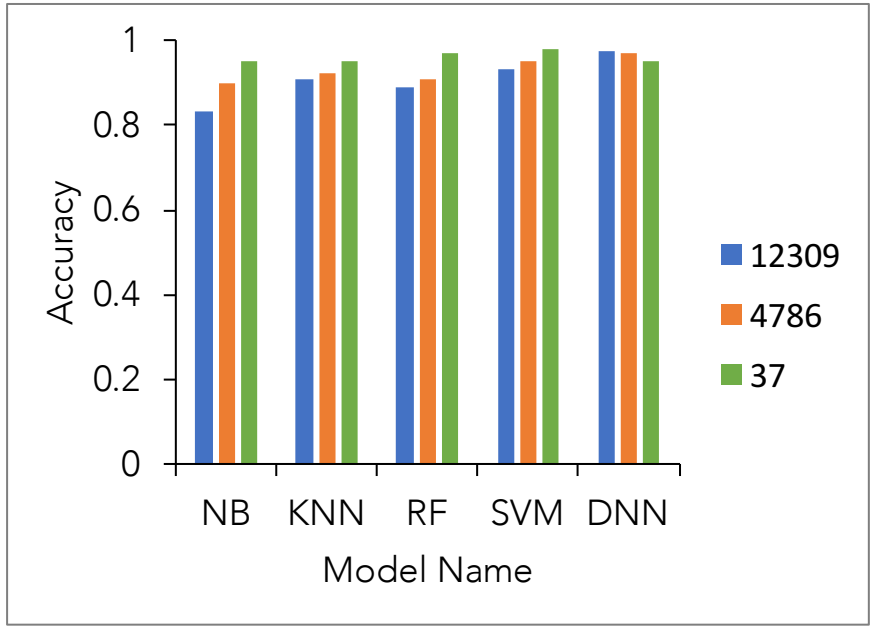

*Han, Leng et al., 2014*

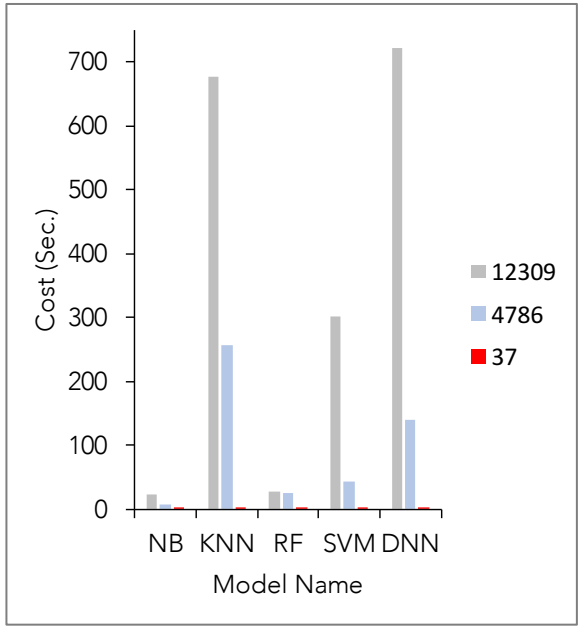
# Key lncRNAs

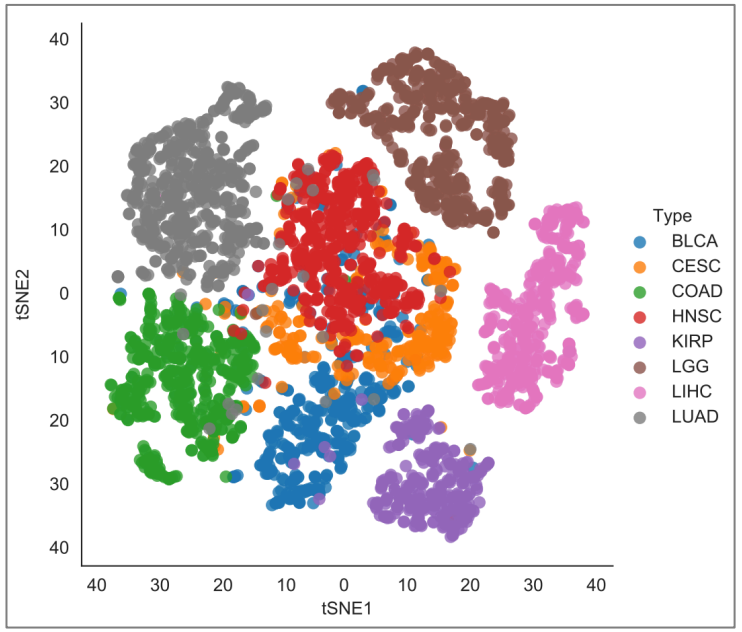
| Key lncRNA (n=37) |
| --- |
| AC000111.6, AC005082.12, AC005355.2, AC009299.3, AL450992.2, AP001626.1, BBOX1-AS1, CTA-384D8.31, EMX2OS, FAM182A, FENDRR, GATA3-AS1, ***H19***, HAGLR, HOXA10-AS, HOXA11-AS, HOXD-AS2, KIZ, LINC00857, LINC00958, LINC01082, LINC01158, MIR205HG, NKX2-1-AS1, RP11-157J24.2, RP11-30K9.5, RP11-373D23.2, RP11-435O5.6, RP11-445O3.2, RP11-535M15.1, RP11-76C10.5, SFTA1P, TBX5-AS1, TMEM51-AS1, TP53TG1, UCA1, ***XIST*** |

# Performance and Visualization

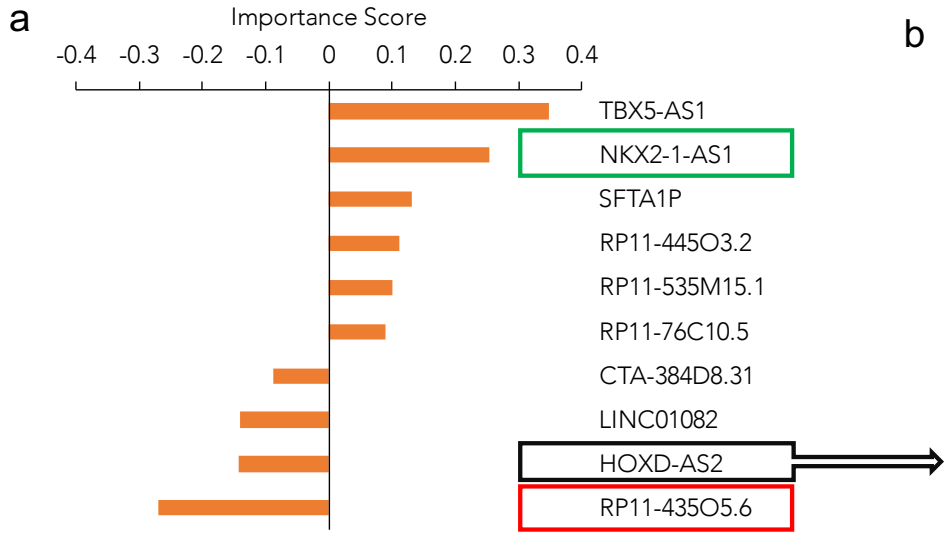
| #Total Features | Model Name | Recall | Precision | Accuracy | Cost (sec.) |
|---|---|---|---|---|---|
| 37 | NB | 0.95 (+/- 0.02) | 0.94 (+/- 0.02) | 0.95 (+/- 0.01) | 0.09 |
| | KNN | 0.94 (+/- 0.01) | 0.95 (+/- 0.01) | 0.95 (+/- 0.01) | 1.44 |
| | RF | 0.97 (+/- 0.01) | 0.97 (+/- 0.01) | 0.97 (+/- 0.01) | 2.28 |
| | **SVM** | **0.97 (+/- 0.01)** | **0.97 (+/- 0.01)** | **0.98 (+/- 0.01)** | **0.82** |
| | DNN | 0.95(+/- 0.01) | 0.95(+/- 0.01) | 0.95(+/- 0.01) | 4.07 |

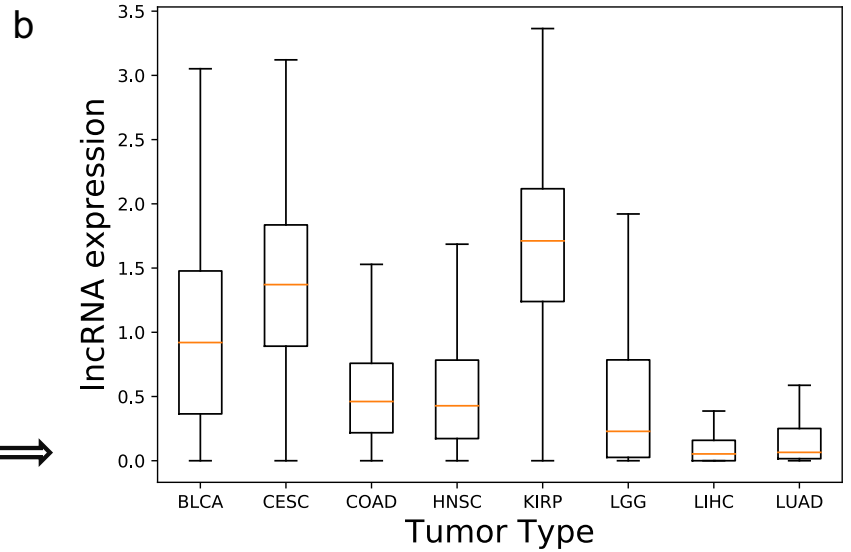


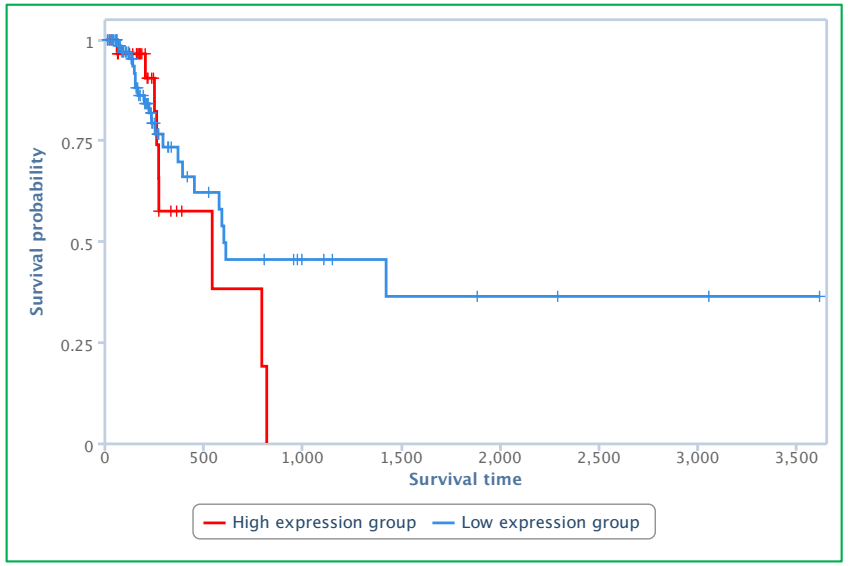



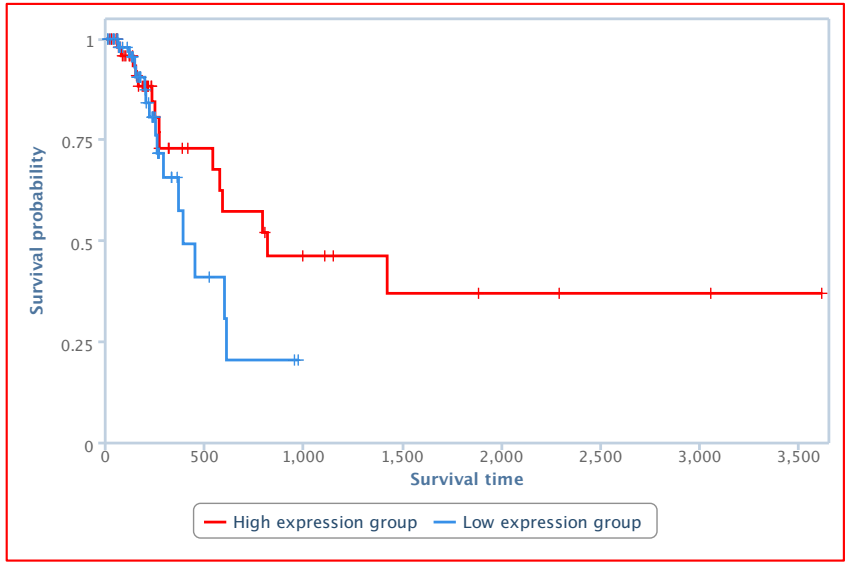



# Validation


a
Importance Score
-0.4 -0.3 -0.2 -0.1 0 0.1 0.2 0.3 0.4
TBX5-AS1
NKX2-1-AS1
SFTA1P
RP11-445O3.2
RP11-535M15.1
RP11-76C10.5
CTA-384D8.31
LINC01082
HOXD-AS2
RP11-435O5.6


b
lncRNA expression
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
BLCA CESC COAD HNSC KIRP LGG LIHC LUAD
Tumor Type


c
Survival probability
1
0.75
0.5
0.25
0
0 500 1,000 1,500 2,000 2,500 3,000 3,500
Survival time
High expression group
Low expression group


d
Survival probability
1
0.75
0.5
0.25
0
0 500 1,000 1,500 2,000 2,500 3,000 3,500
Survival time
High expression group
Low expression group

# Conclusion and Future Directions

## ❑Contribution

Computational framework for key lncRNA identification and cancers classification.

## ❑Results

37 lncRNAs can be used as diagnostic and prognostic features for 8 cancers

## ❑Future works

- Framework can be used for cancer subtyping
- More validations for key lncRNAs
  - Cross check with Literatures
  - Functional/pathway/enrichment analysis

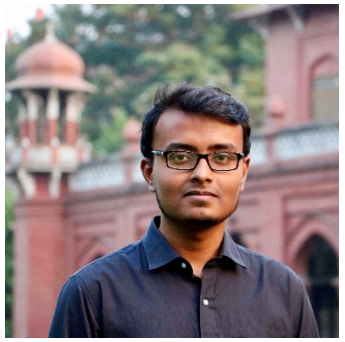
# Acknowledgement

This study is supported by ***NSF CAREER*** award

## Thanks to

Tasmia Aqila

Mona Maharjan



Raihanul Bari Tanvir

## Collaborators

- Professor Giri Narasimhan, Bioinformatics Research Group (BioRG), SCIS, FIU
- Professor Wenrui Duan, Department of Human & Molecular Genetics, FIU
- Professor Lidia Kos, Biomolecular Sciences Institute, FIU

# Some Important terms

**Bias**

Bias are the simplifying assumptions made by a model to make the target function easier to learn.

Generally, linear algorithms have a high bias making them fast to learn and easier to understand but generally less flexible. In turn, they have lower predictive performance on complex problems that fail to meet the simplifying assumptions of the algorithms bias.

**Low Bias**: Suggests less assumptions about the form of the target function.

**High-Bias**: Suggests more assumptions about the form of the target function.

- Examples of **low-bias** machine learning algorithms include: Decision Trees, k-Nearest Neighbors and SVM
- Examples of **high-bias** machine learning algorithms include: Linear Regression, Linear Discriminant Analysis and Logistic Regression.

**Variance**

Variance is the amount that the estimate of the target function will change if different training data was used.

**Low Variance**: Suggests small changes to the estimate of the target function with changes to the training dataset.

**High Variance**: Suggests large changes to the estimate of the target function with changes to the training dataset.

Generally, nonlinear machine learning algorithms that have a lot of flexibility have a high variance. For example, decision trees have a high variance, that is even higher if the trees are not pruned before use.

**Fitting**

# Methods for Feature Selection

- ***Filter based:*** We specify some metric and based on that filter features. An example of such a metric could be correlation/chi-square.
- ***Wrapper-based:*** Wrapper methods consider the selection of a set of features as a search problem. Example: **RFE** Recursive Feature Elimination
- ***Embedded:*** Embedded methods use algorithms that have built-in feature selection methods. For instance, **Lasso** and RF have their own feature selection methods.

# Acknowledgement

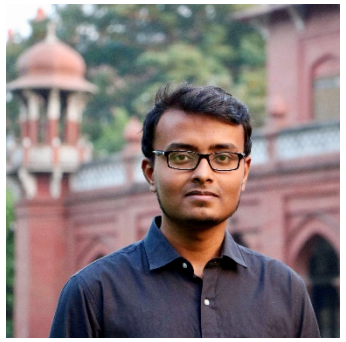
This study is supported by ***NSF CAREER*** award

**Thanks to**

**Dr. Ananda M. Mondal**

Tasmia Aqila

Mona Maharjan



Raihanul Bari Tanvir

## Collaborators

- Professor Giri Narasimhan, Bioinformatics Research Group (BioRG), SCIS, FIU
- Professor Wenrui Duan, Department of Human & Molecular Genetics, FIU
- Professor Lidia Kos, Biomolecular Sciences Institute, FIU