Survival Prediction for Brain Cancer Using Machine Learning Models

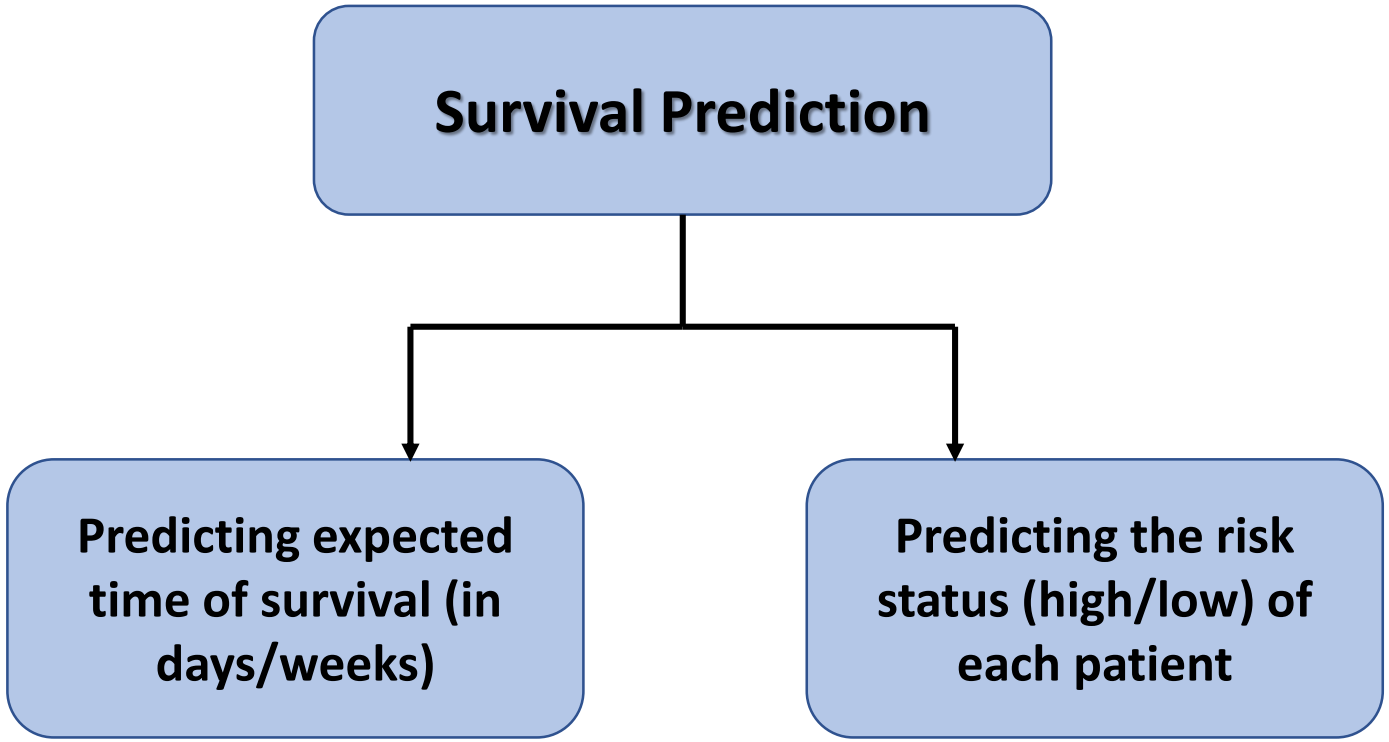
Presented by -

Shreya Johri (2016BB10039)

Supervisors -

Prof. D. Sundar and Prof. Sumeet Agarwal

Introduction



Introduction

- Need for Survival Prediction
  - Important in clinical setting – effective decision making
  - Helps advance scientific knowledge about disease biomarkers
- Low Grade Gliomas
  - Grade II and III brain cancers
  - Develop in supporting glial cells of the brain
  - Have the potential to transform into serious glioblastomas - survival prediction important

Introduction

- Current models for survival prediction
  - Random Forest
  - Cox Proportional Hazard
  - Kaplan Meier
  - Deep Neural Networks

Why Multi Omics?

- Model accuracy with single omics not at par with best reported
- Multi Omics -
  - Can reduce the effect of experimental and biological noise in the data
  - Different omics can reveal different cellular aspects, such as effects manifest at the genomic and epigenomic levels
  - Even within the same molecular aspect, each omic can contain data that is not present in other omics (e.g. mutation and copy number)

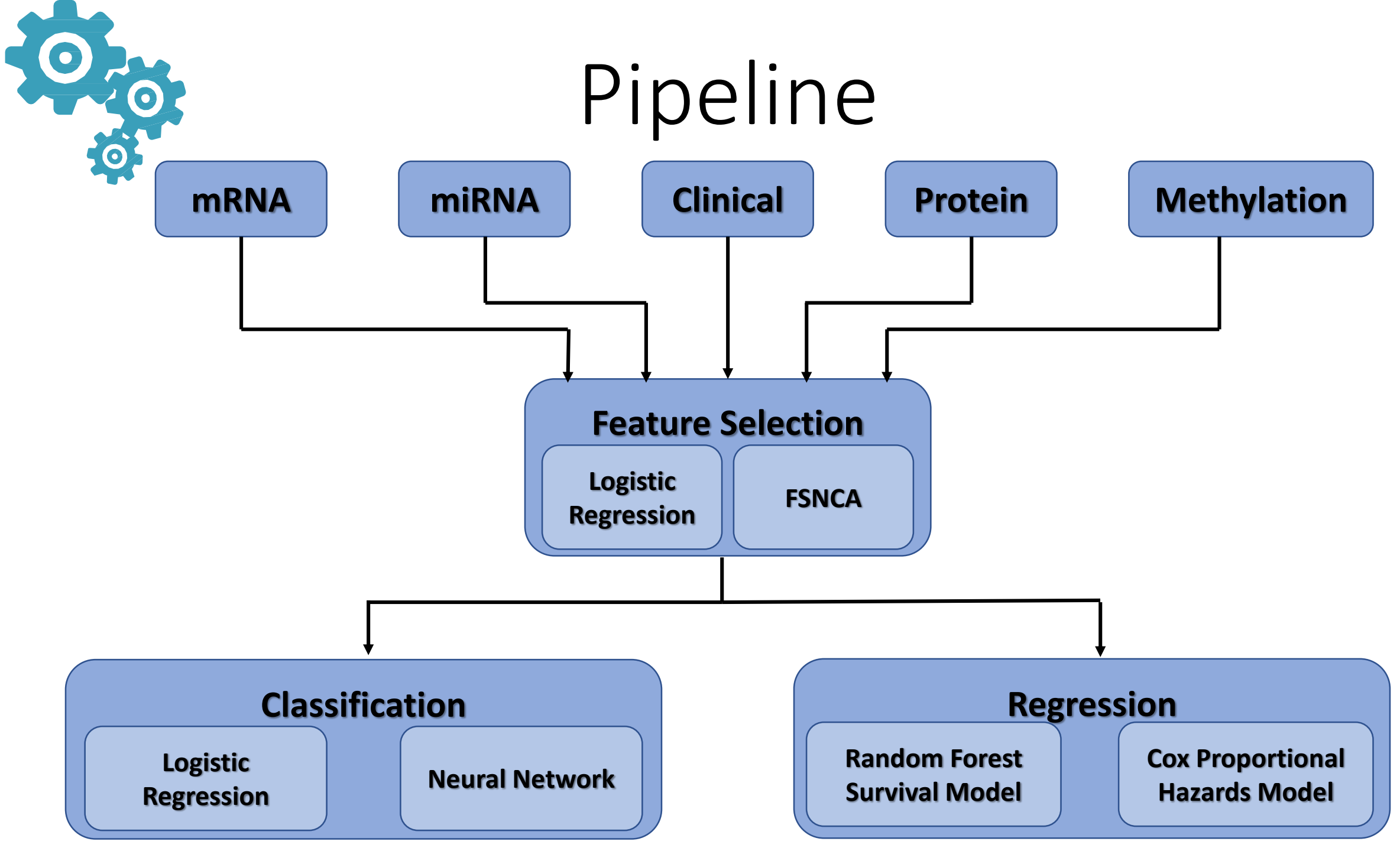
Objectives

- Improving the accuracy of machine learning models
- Comparing the contribution of different omics to the accuracy
- Interpreting the biological significance of models

Data Preparation

- Multi-omics data used from The Cancer Genome Atlas (TCGA)
  - mRNA data – sequencing data
  - methylation – standardised beta values
  - miRNA data – standardised data
  - Protein data – reverse phase protein assay
  - Clinical data – survival status, survival time (days)
- Data Pre-processing
  - Standard normalisation done (all values normalised to -1 to 1)
  - Subjects with missing data <=10% - imputation by mean value
  - Subjects with missing data >=10% - removed from study

Pipeline


mRNA
miRNA
Clinical
Protein
Methylation
Feature Selection
Logistic Regression
FSNCA
Classification
Logistic Regression
Neural Network
Regression
Random Forest Survival Model
Cox Proportional Hazards Model

Feature Selection

Supervised learning done to ensure selection of best features

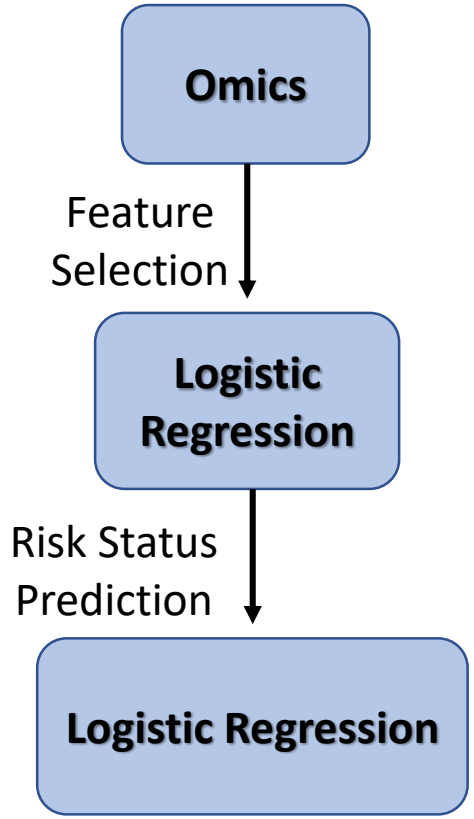
- Logistic Regression with L1 (lasso) penalty
  - Reduces the weights of features with low importance to zero
  - Only features with non-zero weights selected
- Feature selection using Neighbourhood Component Analysis (FSNCA)
  - Optimises the selected features according to best classification for KNN
  - Features with weights less than a set threshold selected

Results

- Prediction of survival time (days) –
  - CoxNet performs better than random forest survival models
  - Issue of overfitting resolved
- Prediction of Risk Status (high/low) –
  - Logistic Regression performs better over neural networks
  - Overfitting yet to be resolved

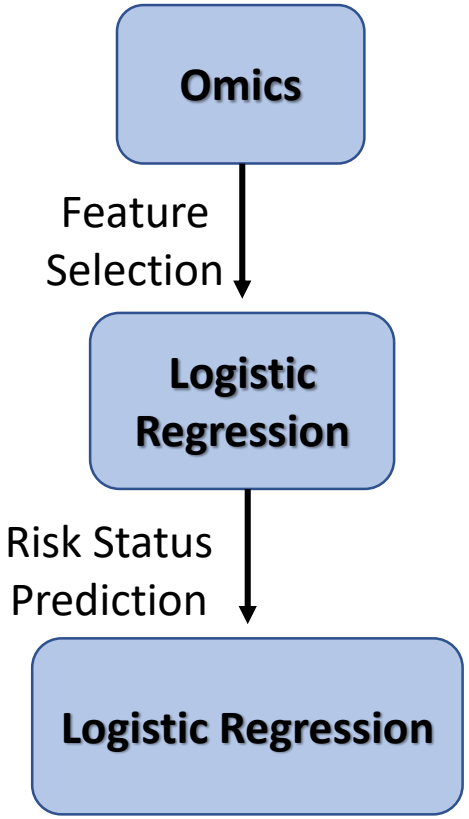
Prediction of Risk Status (High/Low)

Omics	Cross Validation Concordance	Train Concordance
miRNA + methylation + mRNA + protein	0.812	0.924
miRNA + mRNA	0.834	0.928



Prediction of Risk Status (High/Low)

Omics	Cross Validation Concordance	Train Concordance
miRNA + methylation + mRNA + protein	0.812	0.924
miRNA + mRNA	0.834	0.928

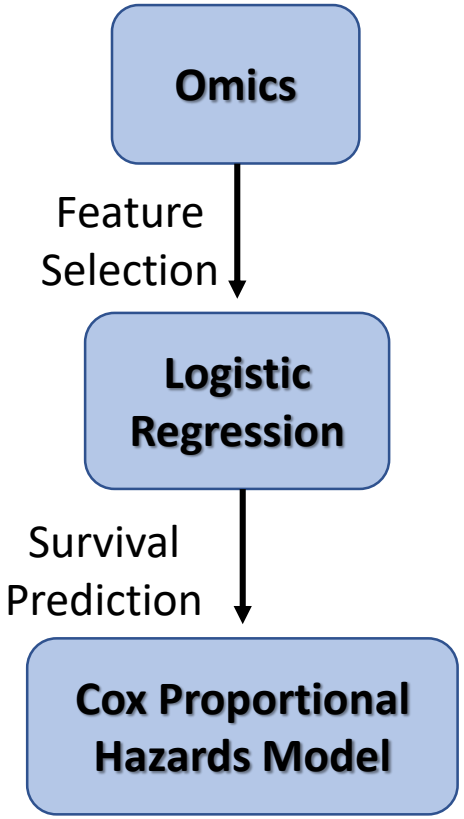


Key conclusions:

- Not all omics combined together give the best results
- The increase in accuracy of miRNA + mRNA vs all omics combined is not huge => redundancy in features still present

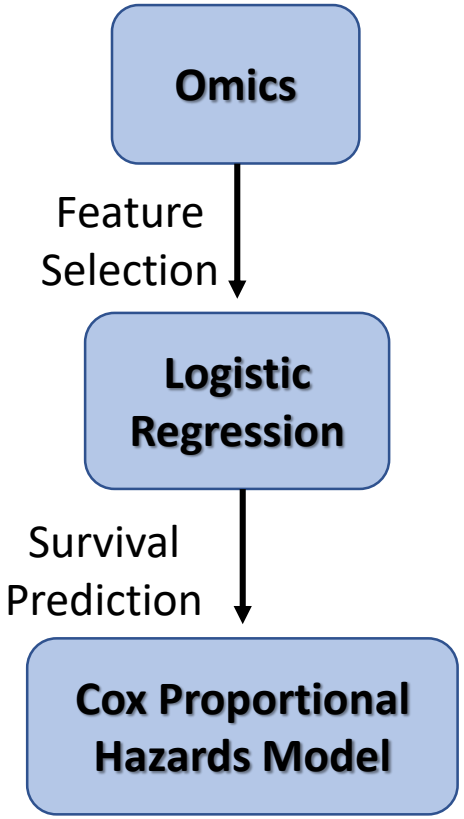
Prediction of Survival Time (days)

Omics	Cross Validation Concordance	Train Concordance
miRNA + methylation + mRNA + protein	0.890	0.963
methylation + mRNA	0.892	0.934
methylation	0.863	0.870
mRNA	0.882	0.916



Prediction of Survival Time (days)

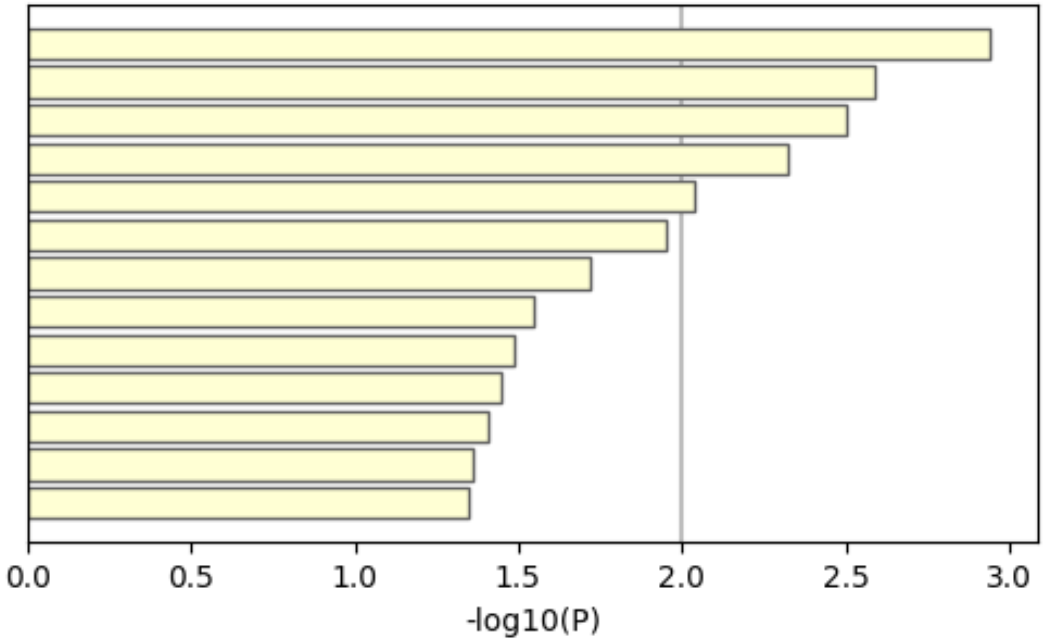
Omics	Cross Validation Concordance	Train Concordance
miRNA + methylation + mRNA + protein	0.890	0.963
methylation + mRNA	0.892	0.934
methylation	0.863	0.870
mRNA	0.882	0.916



Key conclusions:

- methylation + mRNA give the best results
- Redundancy in features from miRNA and protein
- Improvement in mRNA + methylation vs accuracy of individual omics is not huge => features selected are overlapping

Biological Interpretation

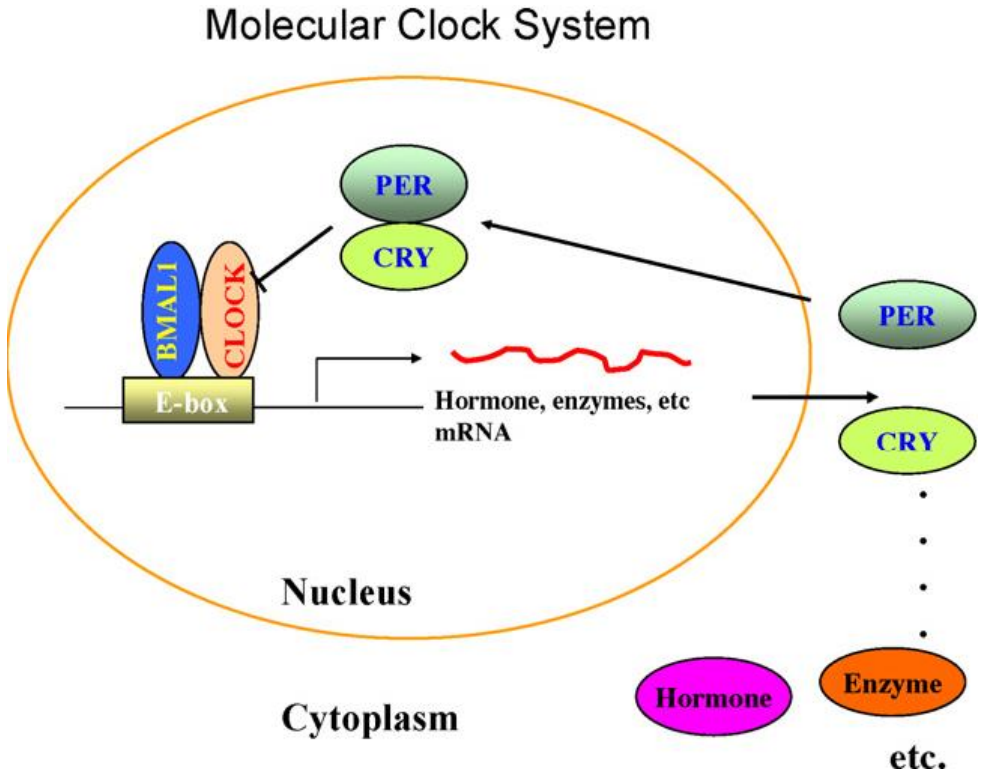


GO:0032922: circadian regulation of gene expression
GO:0034976: response to endoplasmic reticulum stress
GO:0008217: regulation of blood pressure
GO:0007631: feeding behavior
GO:0071407: cellular response to organic cyclic compound
GO:0045580: regulation of T cell differentiation
GO:0007156: homophilic cell adhesion via plasma membrane adhesion molecules
GO:0007276: gamete generation
R-HSA-418555: G alpha (s) signalling events
GO:0006888: ER to Golgi vesicle-mediated transport
GO:0006631: fatty acid metabolic process
GO:0009755: hormone-mediated signaling pathway
GO:0048705: skeletal system morphogenesis

Pathway enrichment analysis using *Metascape*

- Significant pathways shown (p<=0.05)
- Circadian rhythm pathway most affected

Biological Interpretation

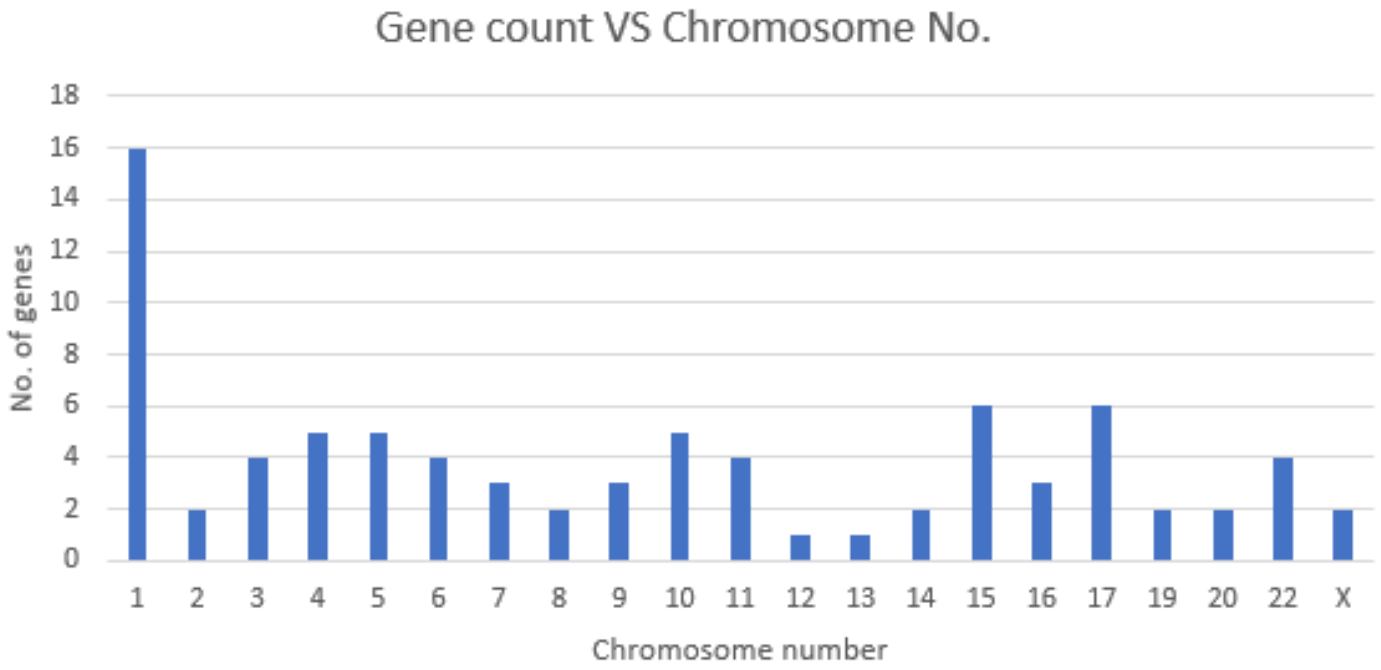


Mutations of *Per2* accelerates tumor growth due to an increase in proliferation rate upon circadian rhythm disruption, because tumor suppressor and key cell cycle genes are under clock control.

Genes found by the model : ARNTL (clock gene), PER3, CIART, IFNB1, EIF2AK4

[1] Kiessling, Silke, et al. "Enhancing circadian clock function in cancer cells inhibits tumor growth." BMC biology 15.1 (2017): 13.
[2] Shimba, Shigeki, and Yuichi Watabe. "Crosstalk between the AHR signaling pathway and circadian rhythm." Biochemical pharmacology 77.4 (2009): 560-565 Figure1

Biological Interpretation



- Most of the genes found in chromosome 1
- Anomalies in chromosome 1 detected in breast cancer, no such study in brain cancer

Orsetti, Batrice, et al. "Genetic proling of chromosome 1 in breast cancer: mapping of regions of gains and losses and identication of candidate genes on 1q." British journal of cancer 95.10 (2006): 1439.

Conclusion

- The feature selection pipeline is selecting features which are of importance in cancer
  - Many features have been validated in previous studies
  - Many features are new, paving way for future experimental studies
- CoxNet for prediction of survival time is performing better than state-of-the-art models
- Neural Network for prediction of risk status needs more optimisation

Future work

- Extending risk status prediction to unsupervised settings like k-nearest neighbours
- Explore Deep Learning architectures for survival prediction time
- Include radiology data as input to the model and evaluate its contribution

Thank you!

Prior Work

- Conclusions of single omics study :
  - High accuracy not obtained even after multi layer feature selection and positive feedback loops
  - Pathways involved only in brain cancer not found identified explicitly by the model

	Test Accuracy	Train Accuracy
Selected from Grade Data (2500 genes)	0.354	0.967
Selection after feedback loop (1000 genes)	0.531	0.906