Bladder Cancer - Survival Analysis

Arthur Lui
Christine Ma

Department of Statistics
Brigham Young University

April 22, 2014

Bladder Cancer

- USA 2014: 74690 new cases, 15580 deaths
- Interested in relationship between gene expression and bladder cancer
- Want to compare different statistical methods

Description of Data

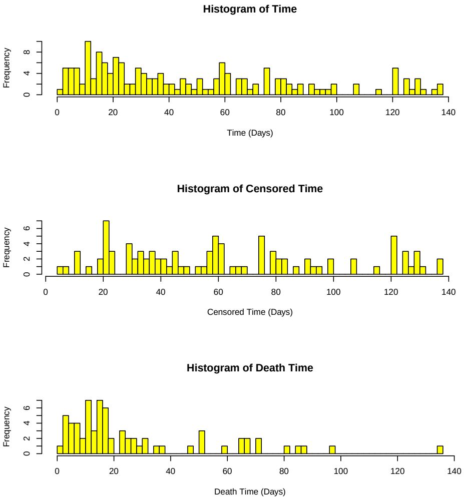
- Biomarkers (43149)
- Survival Times
- Censoring Indicators
- Censoring Rate = 58%
- Number of Observations (Patients) = 165
- No dichotomization was done
- Removed the column of NA's in data set

Summary Statistics

Summary Table of Survival Times

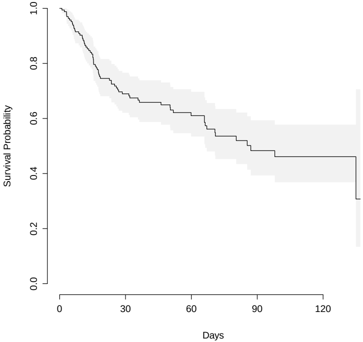
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Censored	5.30	32.33	58.25	63.00	87.17	137.00
Died	1.03	10.40	16.67	28.05	35.70	135.00
Overall	1.03	17.13	36.57	48.38	74.17	137.00

Histogram of Survival Times



Note: We have more data on lower survival times. And more deaths occurred at lower survival times.

KM Curve



Median = 87.07 (33.97, 140.16)

FDR

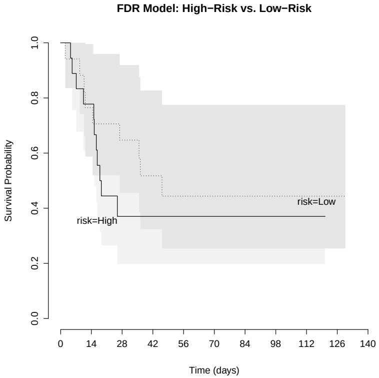
+ Appropriate for large size of independent and dependent coefficients
- Average fraction of false rejections has to be made or obtained using cross validation

Interaction terms were not included

Cox Model Using Variables with FDR $< .025$

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
geneILMN_1666893	0.28	1.32	0.26	1.04	0.30
geneILMN_1689037	0.82	2.26	0.22	3.70	0.00
geneILMN_1690017	0.32	1.37	0.28	1.15	0.25
geneILMN_1702933	0.45	1.57	0.32	1.42	0.15
geneILMN_1714118	0.48	1.61	0.51	0.94	0.35
geneILMN_1714592	-0.15	0.86	0.37	-0.40	0.69
geneILMN_1718866	0.18	1.20	0.40	0.45	0.65
geneILMN_1745238	0.23	1.26	0.33	0.69	0.49
geneILMN_1757351	0.00	1.00	0.16	0.03	0.98
geneILMN_1767685	-0.26	0.77	0.48	-0.55	0.59
geneILMN_1807525	-0.01	0.99	0.56	-0.03	0.98
geneILMN_1809336	0.40	1.49	0.28	1.42	0.15
geneILMN_1889811	0.69	2.00	0.42	1.64	0.10

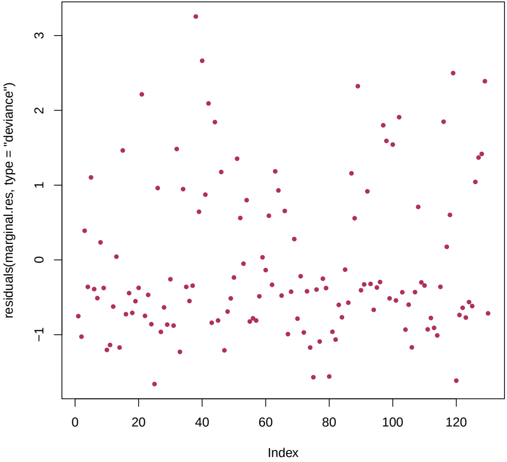
FDR KM Plots



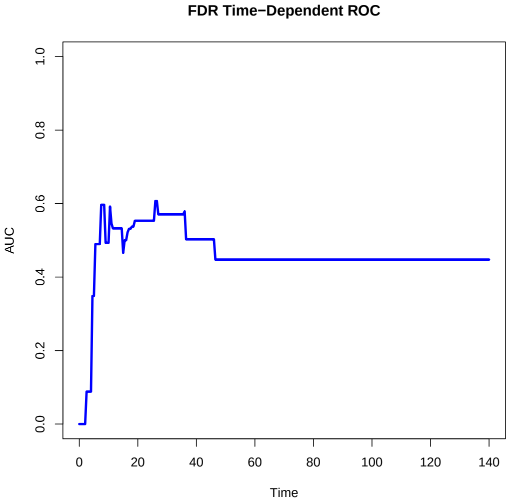
Low Median = 18.2. High Median = 46.2

Likelihood ratio test=63 on 13 df, p=1.49e-08 n= 130.

Residuals Plot



FDR AUC

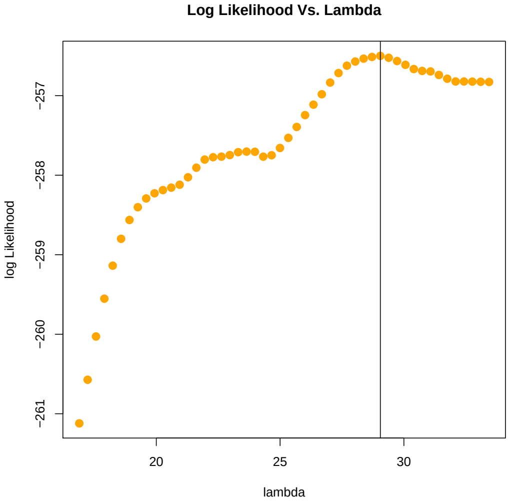

FDR Time−Dependent ROC
AUC
0.0
0.2
0.4
0.6
0.8
1.0
Time
0
20
40
60
80
100
120
140

Lasso

+ Performs model selection
- Tuning parameter needs to be estimated

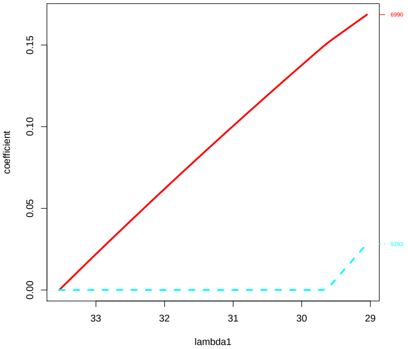
Interaction terms were not included

Selecting Tuning Parameter λ



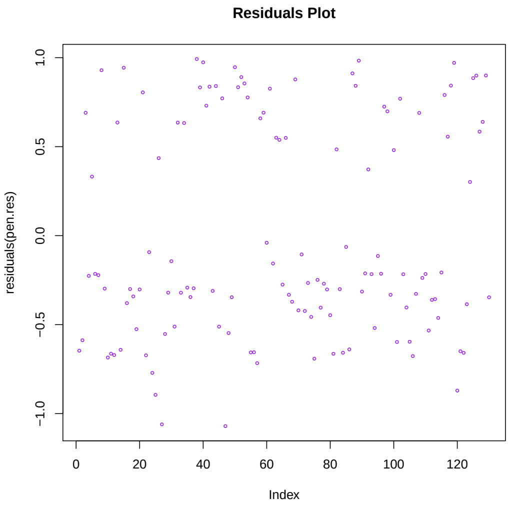
$\lambda = 29.95$

Selected Variables

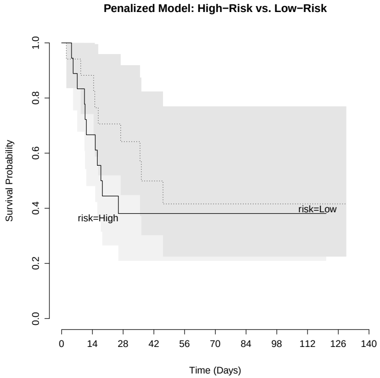


	ILMN_1689037	ILMN_1702933
1	0.17	0.03

Residuals Plot



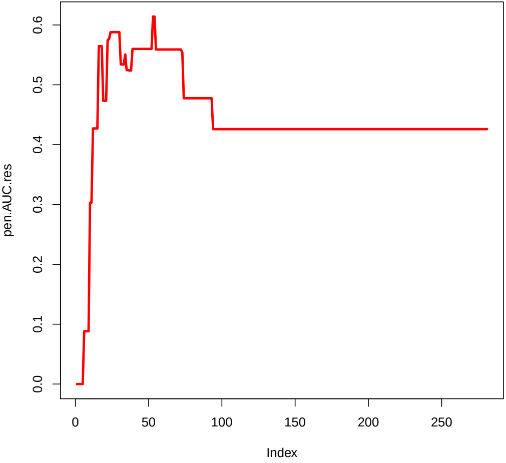
Lasso KM Plots



Low Median = 36.3
High Median = 18.2

Likelihood ratio test=41.9 on 2 df, p=7.83e-10

Lasso AUC



Random Forests Model

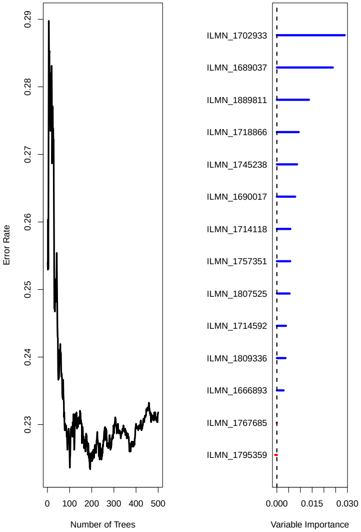
1. A regression tree is a model that predicts the response of an input based on a sequence of decisions
2. A Random Forest is created from many trees
3. The predicted response of the random forest is the mean of the predictions of the individual trees

Random Forest

+ Good for modelling non-linear data (data assumed to be nonlinear)
- Lower prediction accuracy

Interaction terms were not included

Variable Importance



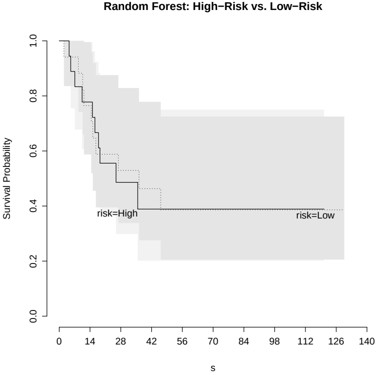
Cox Model Using Important Variables from Random Forest

	coef	exp(coef)	se(coef)	z	p
genesILMN_1689037	0.68	1.98	0.18	3.70	0.00
genesILMN_1702933	1.03	2.80	0.25	4.08	0.00
genesILMN_1704154	0.32	1.37	0.19	1.69	0.09
genesILMN_1749989	-2.32	0.10	1.33	-1.74	0.08

Likelihood ratio test=49.5 on 4 df, p=4.58e-10 n= 130, number of events= 49

Random Forest KM Plots

Important Markers: ILMN_1689037, ILMN_1702933, ILMN_1704154, ILMN_1749989



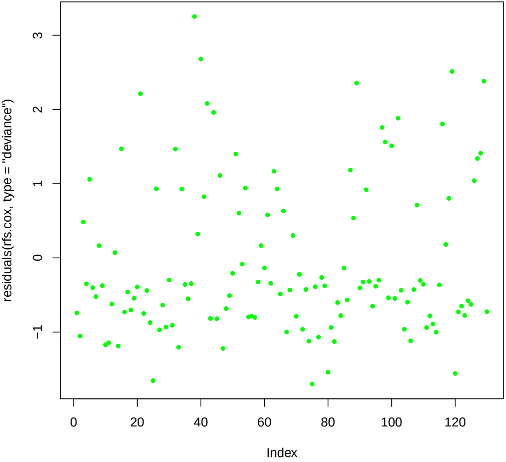
Random Forest: High−Risk vs. Low−Risk

Median

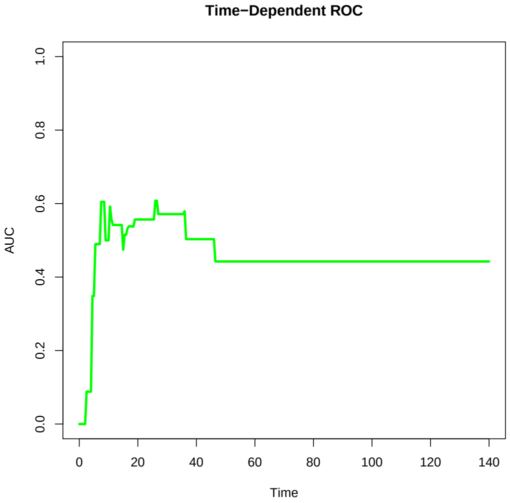
	Estimate	CI.Lower	CI.Upper
Low-Risk Group	36.30	-39.43	112.03
High-Risk Group	25.83	-9.78	61.45

Likelihood Ratio Test = 0.02 on 1 df, p=0.8958 (Curves similar)

Residuals Plot



Random Forest AUC



Hierarchical Clustering Model

1. Identify hyperplane that provides maximum separation between clusters

Hierarchical Clustering

+ Good result visualization

+ Will obtain a hierarchy of clusters

+ Fast computation

+ Helpful for identifying gene expression data patterns in time and space

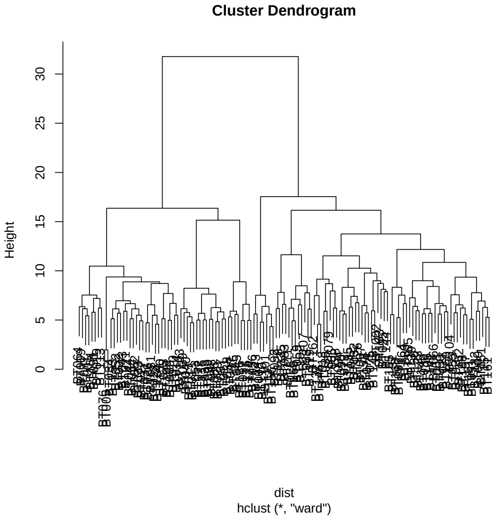
- Doesn't identify best clusters

- Sensitive to noise and outliers

- Might break for large clusters

Interaction terms were not included

H-Clust model

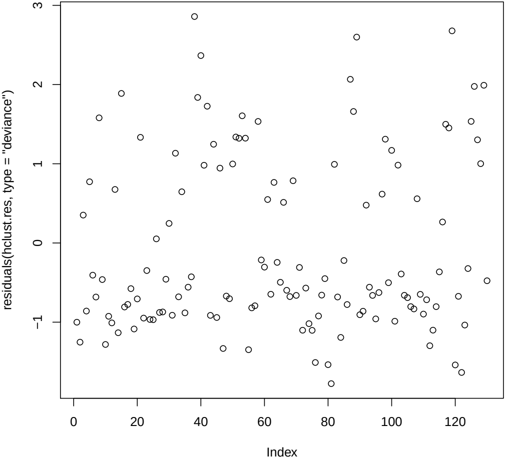


Cox Model Using Important Variables from H-Clust

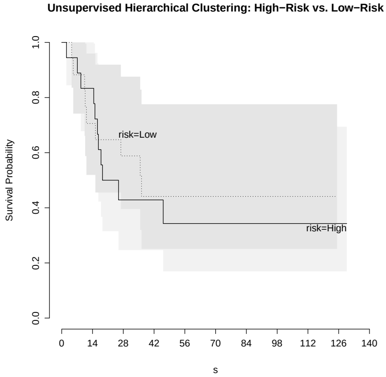
	coef	exp(coef)	se(coef)	z	p
genesILMN_1651236	-0.19	0.83	0.41	-0.45	0.65
genesILMN_1651260	0.44	1.55	0.38	1.16	0.25
genesILMN_1651429	0.26	1.29	0.14	1.78	0.08
genesILMN_1651433	-0.02	0.98	0.35	-0.05	0.96
genesILMN_1651438	0.47	1.60	0.29	1.65	0.10
genesILMN_1651557	-0.23	0.80	0.25	-0.92	0.36
genesILMN_1651574	-0.24	0.78	0.11	-2.21	0.03
genesILMN_1651611	0.17	1.19	0.16	1.04	0.30
genesILMN_1651652	-0.42	0.66	0.32	-1.31	0.19
genesILMN_1651694	0.32	1.38	0.25	1.27	0.20
genesILMN_1651799	0.24	1.27	0.20	1.18	0.24

(p-value $< 10^{-5}$)

Plot of Deviance Residuals

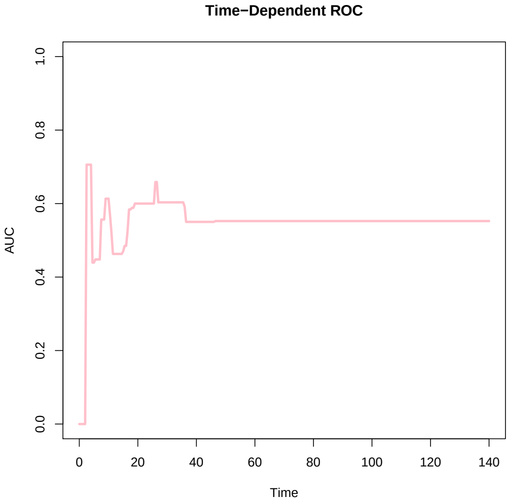


H-Clust KM Plots



Low Risk Median = 36.3 (23.1, 49.5)
High Risk Median = 25.8
Likelihood ratio test= 22.25 on 11 df. p-value=0.0225

H-Clust AUC


Time−Dependent ROC
AUC
0.0
0.2
0.4
0.6
0.8
1.0
Time
0
20
40
60
80
100
120
140

Principal Component Analysis (PCA)

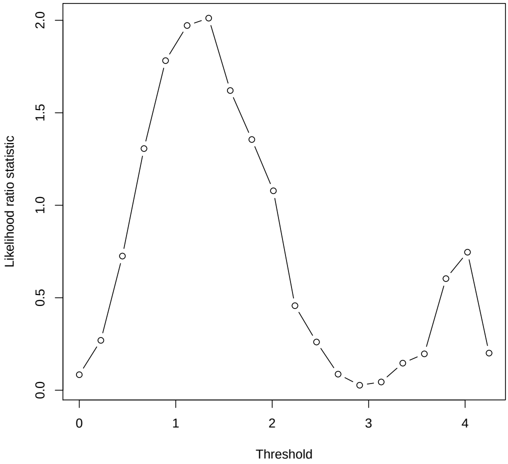
1. Orthogonal Transformation
2. Convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables

PCA

+ Lack of redundancy of data
+ Reduced complexity
+ Smaller database representation
+ Reduced noise b/c the maximum variation basis is chosen (small variations are ignored)
- The covariance matrix is hard to evaluate
- Ability to capture variance depends on the training data

Interaction terms were not included

PCA LRT Threshold



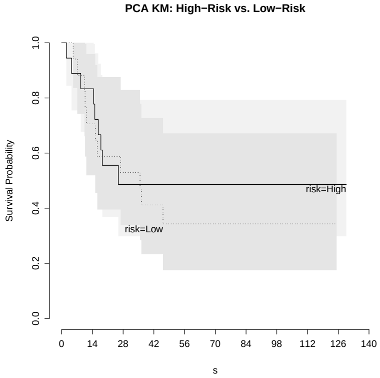
threshold ≈ 1.34

Cox Model Using Principal Components

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
geneILMN_1651574	-0.17	0.84	0.13	-1.37	0.17
geneILMN_1651429	0.23	1.25	0.21	1.08	0.28
geneILMN_1651237	-0.27	0.77	0.20	-1.32	0.19
geneILMN_1651611	-0.07	0.93	0.16	-0.46	0.65
geneILMN_1651832	-0.19	0.82	0.35	-0.55	0.58
geneILMN_1651428	0.16	1.18	0.27	0.60	0.55
geneILMN_1651496	0.02	1.02	0.23	0.10	0.92
geneILMN_1651776	0.23	1.26	0.32	0.70	0.48
geneILMN_1651745	-0.51	0.60	0.23	-2.24	0.02
geneILMN_1651364	0.39	1.48	0.38	1.03	0.30
geneILMN_1651789	0.46	1.58	0.22	2.07	0.04
geneILMN_1651538	0.02	1.02	0.36	0.05	0.96
geneILMN_1651872	0.74	2.09	0.39	1.90	0.06
geneILMN_1651254	-1.15	0.32	0.44	-2.61	0.01
geneILMN_1651336	-0.43	0.65	0.52	-0.83	0.41
geneILMN_1651544	-0.97	0.38	0.40	-2.42	0.02
geneILMN_1651375	0.30	1.35	0.27	1.09	0.28
geneILMN_1651517	-1.59	0.20	0.88	-1.82	0.07

Likelihood ratio test=3.76 on 3 df, p=0.288 n= 35

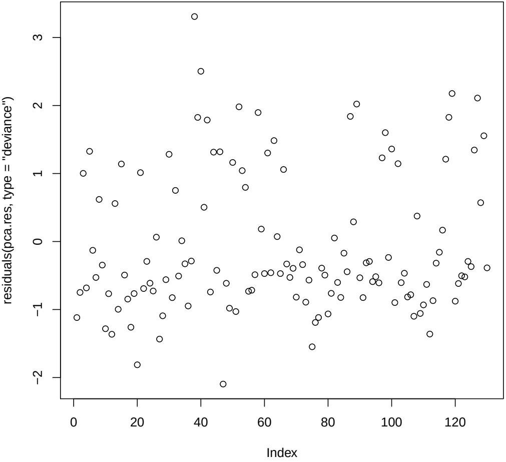
PCA KM Plots



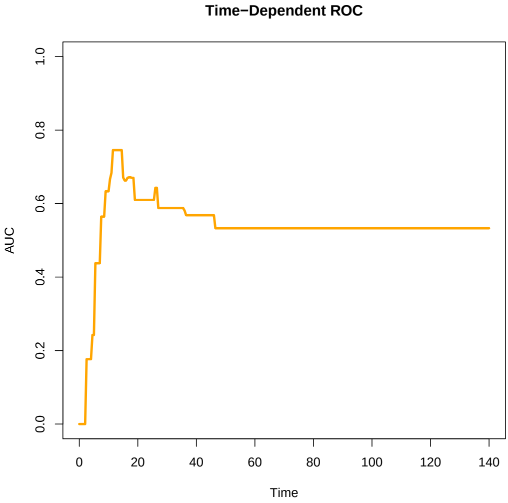
Low Risk Group Median = 35.7. High Risk Group Median = 25.8.

Likelihood ratio test=59.8 on 18 df, p=2.21e-06

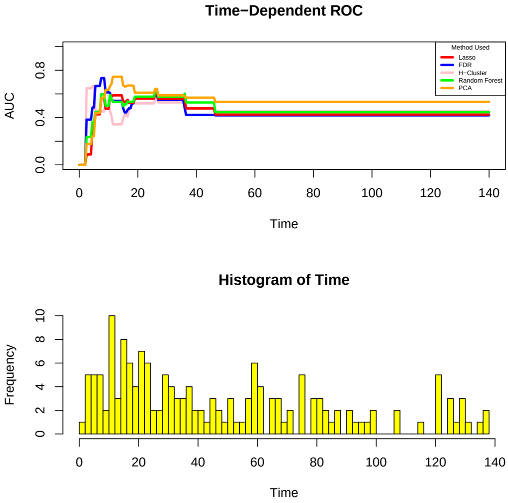
Plot of Deviance Residuals



PCA AUC


Time−Dependent ROC
AUC
0.0
0.2
0.4
0.6
0.8
1.0
Time
0
20
40
60
80
100
120
140

Comparison of Methods



Covariates that appeared most frequently

	FDR	Lasso	RF	H-Clust	PCA
ILMN_1702933	*	*	*	.	.
ILMN_1689037	*	*	*	.	.
ILMN_1651611	.	.	.	*	*
ILMN_1651574	.	.	.	*	*
ILMN_1651429	.	.	.	*	*

Future

Include other covariates