Statistical Techniques to Analyze Expression Level Difference and Survival Conditions in mRNA/miRNA

Ken Lau

Statistical Methods:

1. Group Comparison Techniques
   - Kolmogorov-Smirnov
   - T-Test
   - Box Plot
   - Quantile-Quantile Plot

2. Survival Analysis
   - Kaplan Meier Survival Curves
   - Cox Proportional Hazards Regression

Input Data

Gene: hsa-mir-181c.MIMAT0000258

hsa-mir-181a.MIMAT0000256

Groups: Good/Bad Survival

Variable: Expression Level

Samples: 165 Patients

	Library	hsa-mir-181c#MIMAT0000258_0	ID
1	TCGA-AB-2802	70.695277	GOOD
2	TCGA-AB-2803	213.342981	GOOD
3	TCGA-AB-2805	184.114325	GOOD
4	TCGA-AB-2806	269.754716	GOOD
5	TCGA-AB-2807	491.639123	BAD
6	TCGA-AB-2808	176.775151	GOOD
7	TCGA-AB-2810	262.917753	BAD

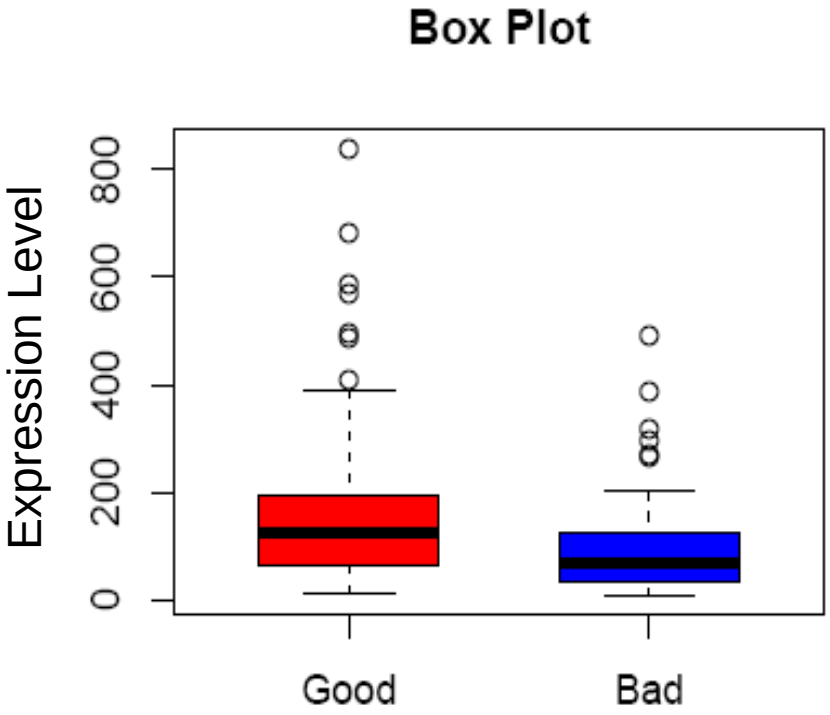
Kolmogorov-Smirnov	T-Test
Ho: The samples are drawn from the same distribution	Ho: The true mean parameters between two populations are the equal

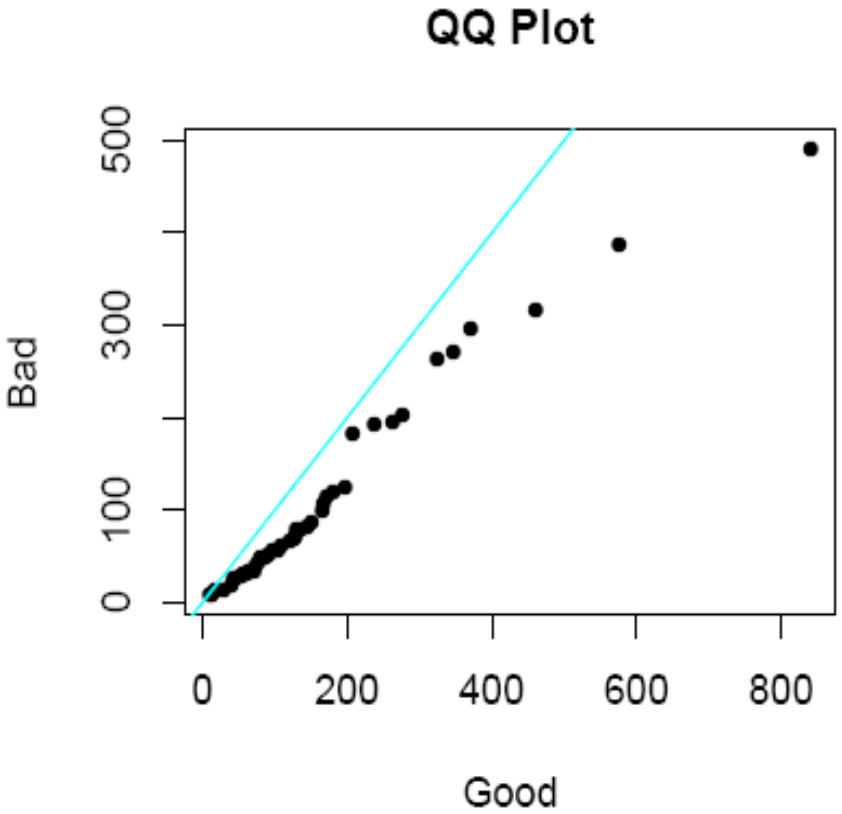
Box Plot: Graphical Display of Interquartile Range, median, min, max, and outliers

QQ Plot: Quantile comparison between two distributions

hsa-mir-181c.MIMAT0000258

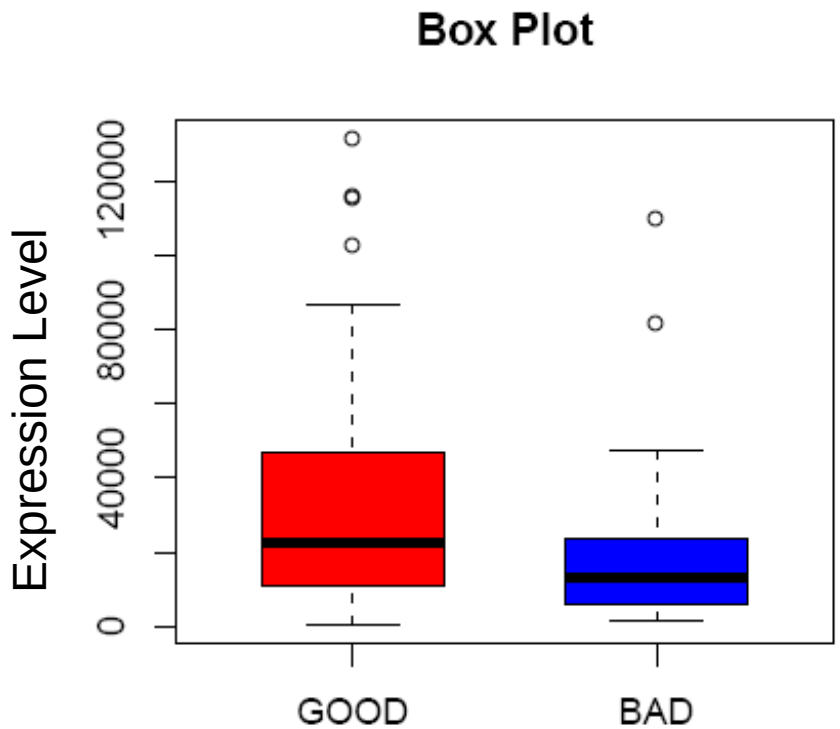
Statistical Test	P-Value
Kolmogorov-Smirnov	0.018643
T-Test	0.023208

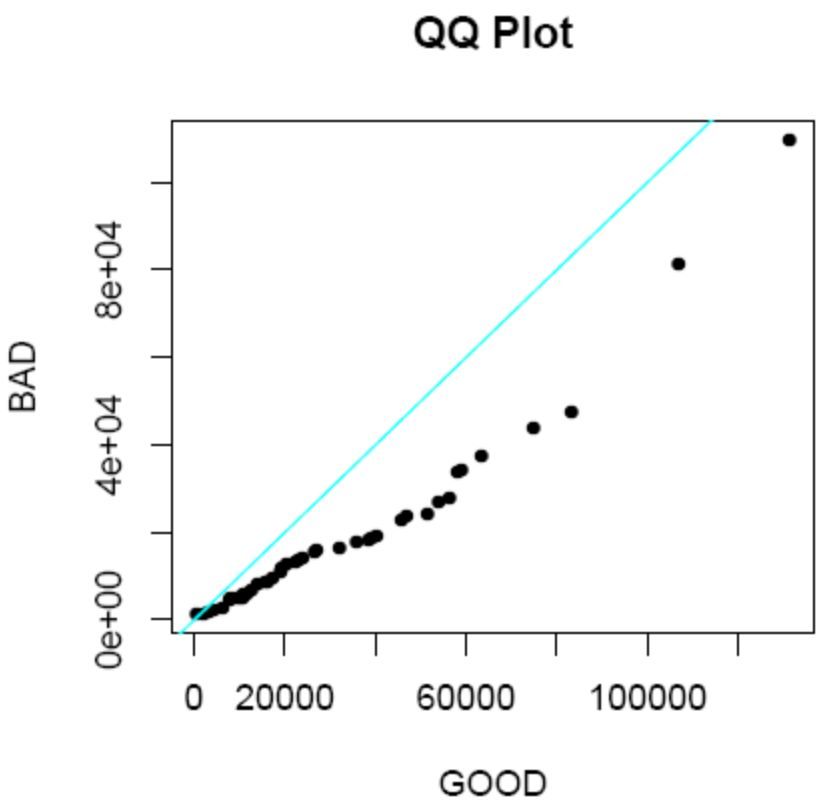




hsa-mir-181a.MIMAT0000256

Statistical Test	P-Value
Kolmogorov-Smirnov	0.007332688
T-Test	0.003161189





Statistical Methods:

1. Group Comparison Techniques
   - Kolmogorov-Smirnov
   - T-Test
   - Box Plot
   - Quantile-Quantile Plot

2. Survival Analysis
   - Kaplan Meier Survival Curves
   - Cox Proportional Hazards Regression

Input Data

Gene: hsa-mir-181c.MIMAT0000258

hsa-mir-181a.MIMAT0000256

Explanatory Variable: Expression Level

High	Med	Low

Survival Variables: Survival Time

Months

Status

1 = Death	0 = Alive

Samples: 187 Patients

	ID	status	months	q.id
1	TCGA-AB-2865-03A-01T	1	2.3	med
2	TCGA-AB-2949-03B-01T	1	32.6	high
3	TCGA-AB-2956-03A-01T	1	5.7	low
4	TCGA-AB-2857-03A-01T	1	10.0	low
5	TCGA-AB-2878-03A-01T	1	12.2	high
6	TCGA-AB-2996-03A-01T	0	73.0	med

Kaplan Meier Survival Curves

- *T = Time of Survival*

- *Survival Function:*

  S(t) = P(T > t)

Hazard Function:

$$h(t) = -\frac{d}{dt}\log S(t)$$

- *Kaplan Meier Survival Estimation:*

$$\hat{S}(t_i) = \prod_{t_i \le t}\left(1 - \frac{d_i}{n_i}\right)$$

n_i = number of subjects at beginning of time period t_i

d_i = number of subjects who die during time period t_i

Cox Proportional Hazards Regression

Y = Survival Time/Status

X = Expression (H/M/L)

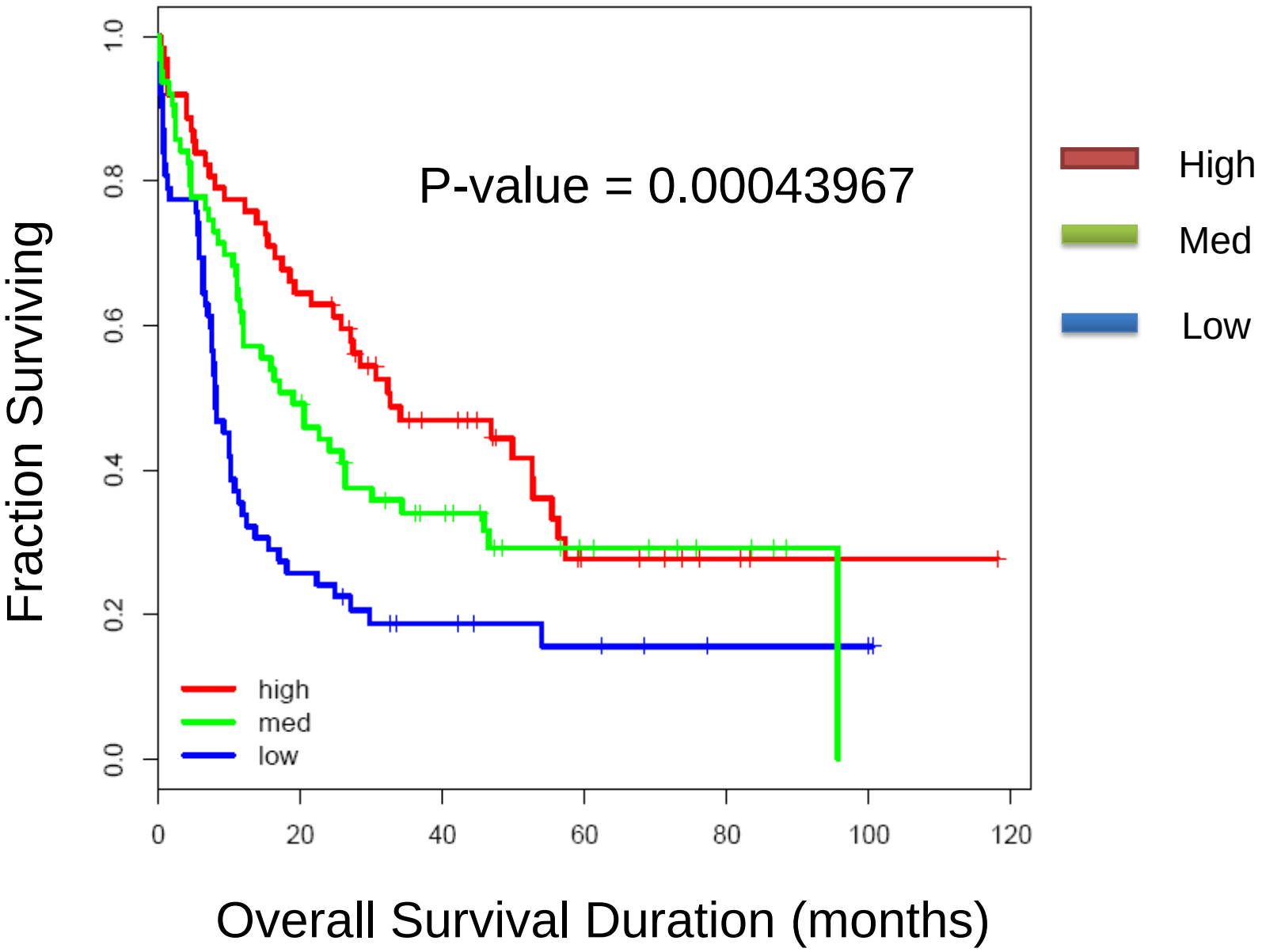
$$h_i(t) = h_0(t)\exp(\beta_1 x_{i1})$$

$\hat{h}(t)$

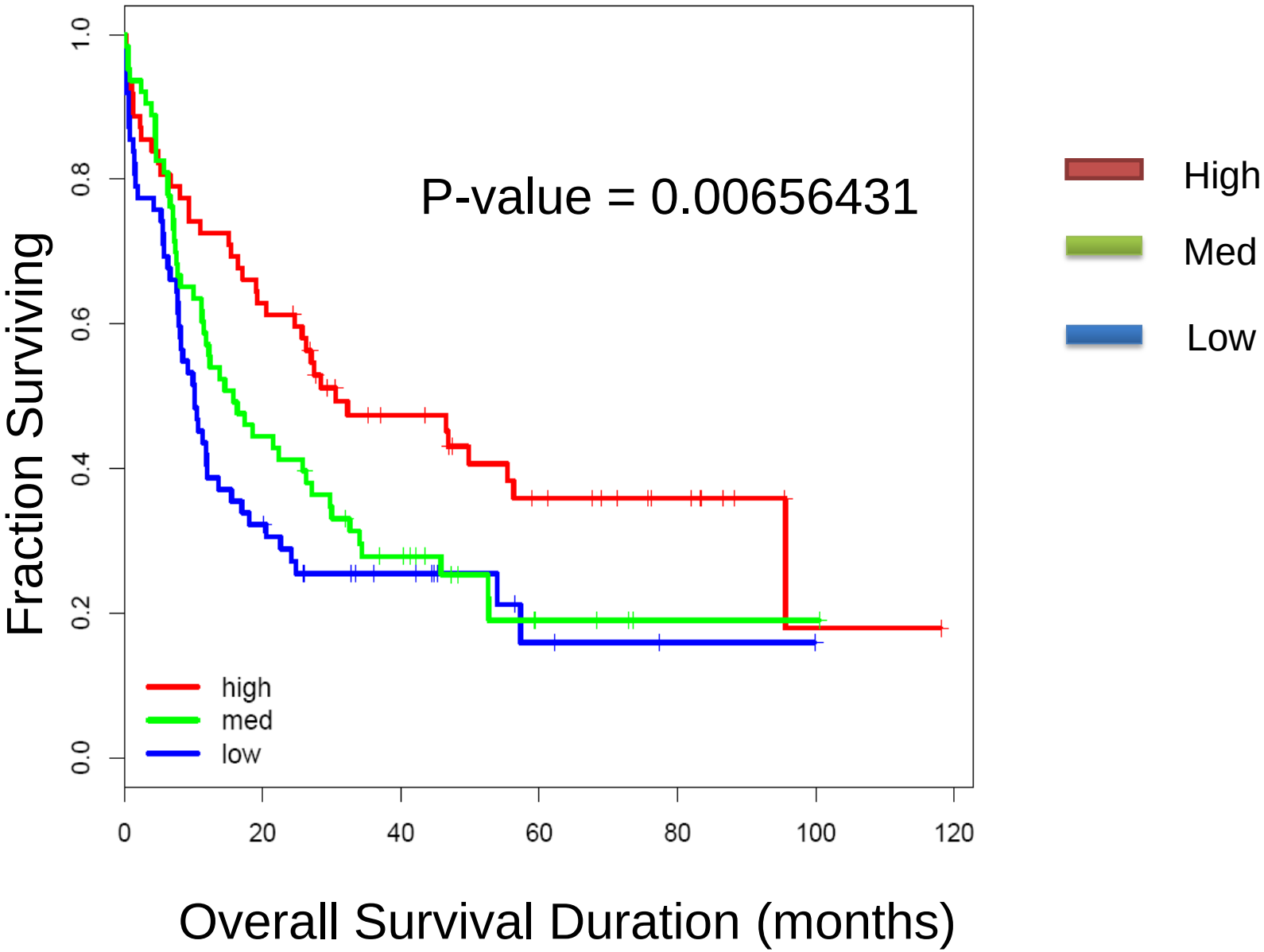
$\hat{S}(t)$

Log-Rank P-Value Test: There is no difference between population curves

hsa-mir-181c.MIMAT0000258



hsa-mir-181a.MIMAT0000256



Conclusion + Additional/Future Work

- Statistical techniques to compare expression level groups with good/bad survival
- Survival analysis to compare high/med/low expression
- Multivariate/Boosted/Penalized Cox Regression Models
- Multiple Testing Correction on p-values

Questions?