



Making Sense of Two Human Lung Carcinomas Gene Expression Data: Integration and Analysis of Two Affymetrix Platform Experiments

**Xiwu Lin, Daniel C. Park, Sergio Eslava,
Kwan R. Lee, Raymond L. Lam, and Lei A. Zhu**

**BIostatISTICS AND DATA SCIENCES
GlaxoSmithKline
November 13, 2003**



OUTLINE

- CEL files vs Processed Data?
- Data Integration and normalization
- “Validation” of the integrated data
- Survival Analysis
- Discussion



OUTLINE

- **Processed data vs Raw data (CEL files) Data?**
- Data Integration and normalization
- “Validation” of the integrated data
- Survival Analysis
- Discussion

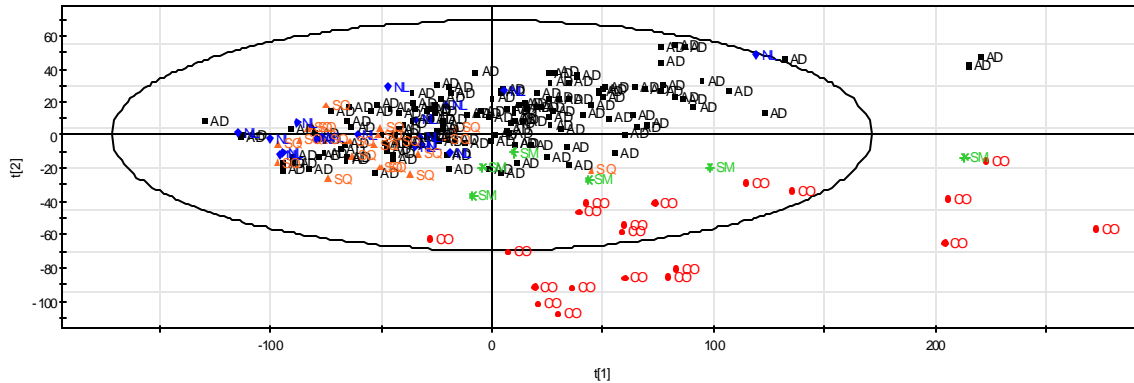


PROCESSED DATA VS RAW DATA?

- Processed Data
 - What kind of expression measure is this?
- MAS 5.0 data generated from CEL files had better separation of disease groups compared to existing processed data



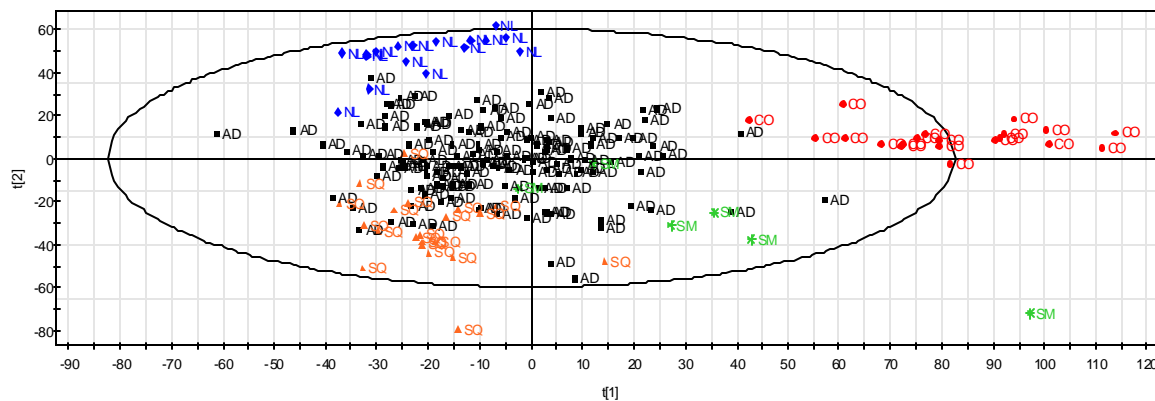
PROCESSED DATA VS. RAW DATA



Harvard processed data

AD = Lung and Other Adenocarcinomas
CO = Pulmonary Carcinoids
SM = Small Cell Carcinomas

SQ = Squamous Cell Carcinomas
NL = Normal Lung



Harvard data generated by MAS 5.0

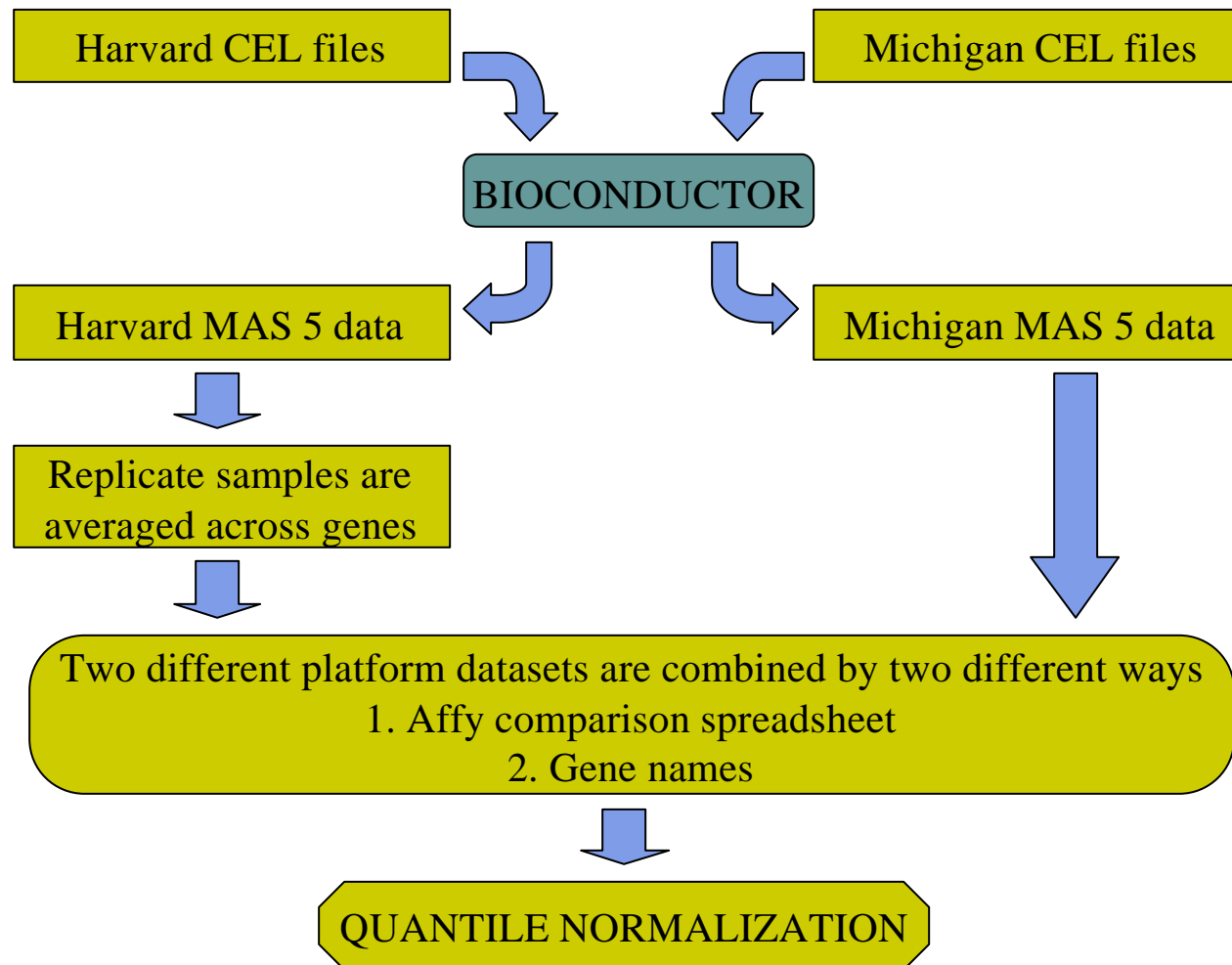


OUTLINE

- CEL files vs Processed Data?
- **Data Integration and normalization**
- “Validation” of the integrated data
- Survival Analysis
- Discussion



DATA INTEGRATION PROCESS



Flow chart for preprocessing of the data



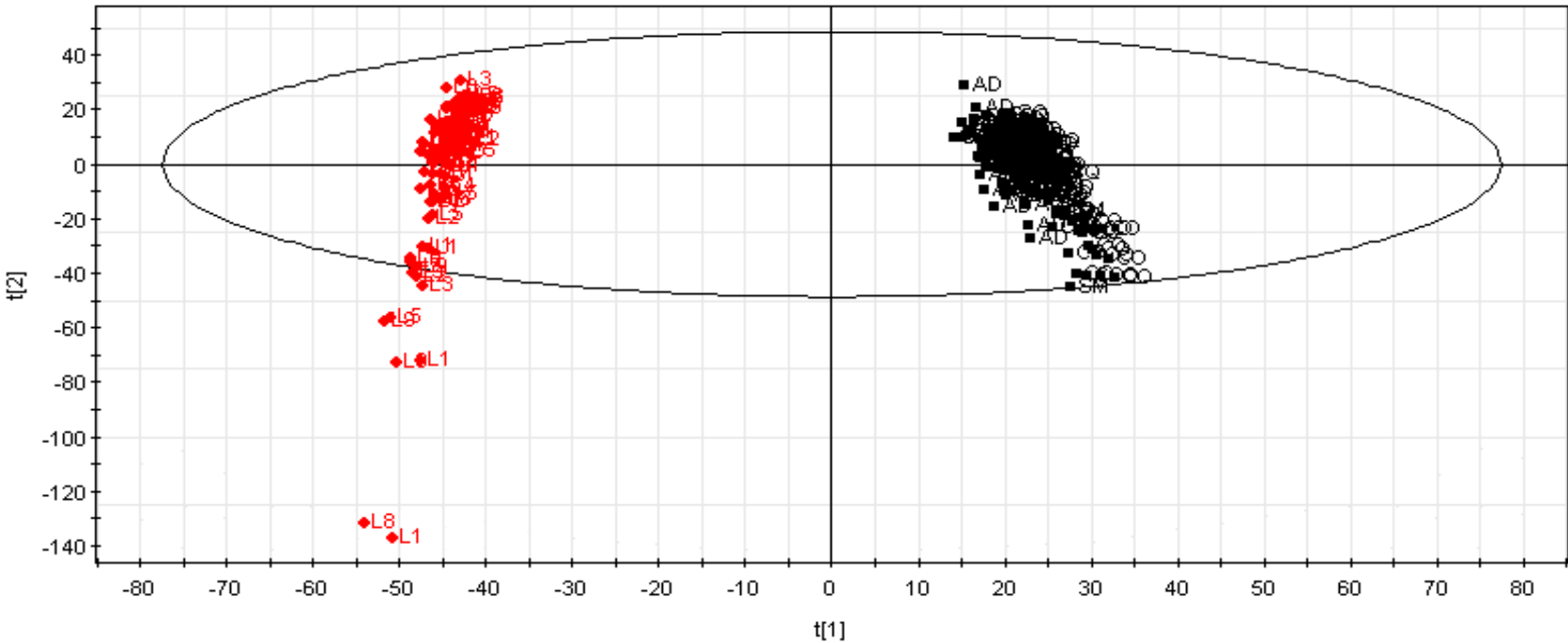
INTEGRATED DATA

- Harvard data set.
 - 203 samples and 12600 probe sets.
- Michigan data
 - 96 samples and 7129 probe sets.
- The integrated dataset
 - Using ACS: 6041 probe sets
 - Using gene name: 4837 genes
 - Either method of combining the datasets resulted in approximately the same information.



BEFORE NORMALIZATION

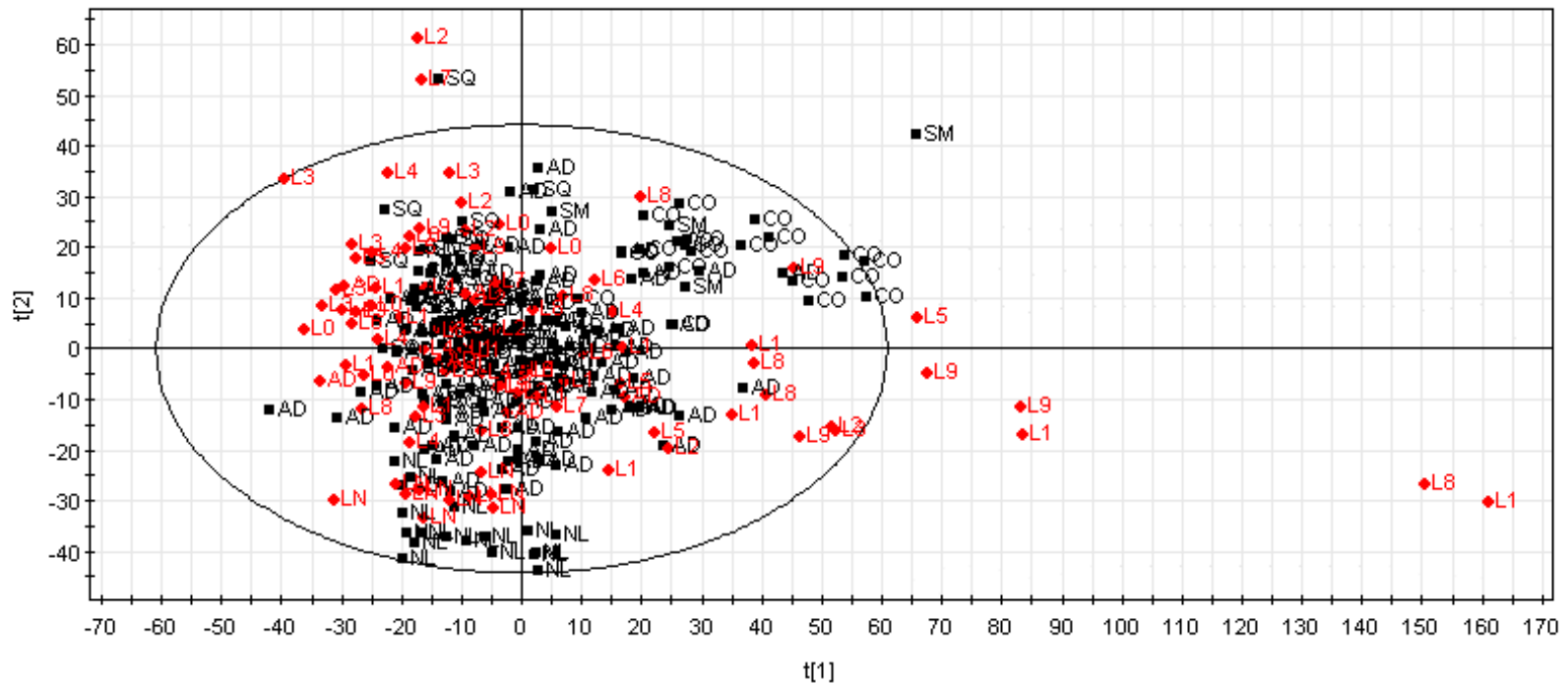
PCA plot of combined data before Q-normalization
(RED : Harvard data BLACK : Michigan data)





AFTER NORMALIZATION

PCA plot of combined data after Q-normalization
(RED : Harvard data BLACK : Michigan data)



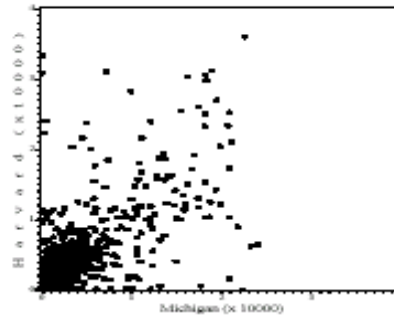
SCATTER PLOTS:HARVARD VS MICHIGAN



**Before
Normalization**

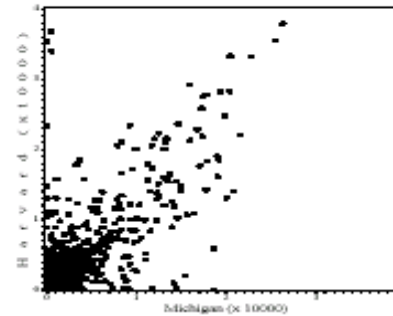
Normal

Normal (Before Q-N)



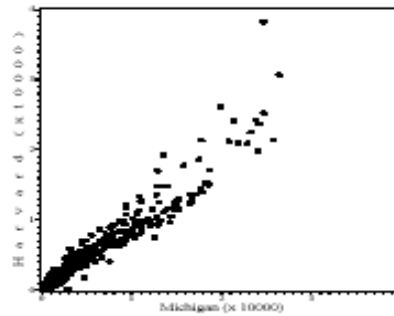
Carcinomas

Patient (Before Q-N)

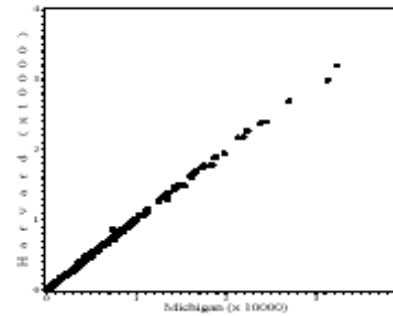


**After
Normalization**

Normal (After Q-N)



Patient (After Q-N)





QUANTILE NORMALIZATION OF COMBINED DATA FROM DIFFERENT PLATFORMS

Algorithm:

- a. Denote $X=(X^1, \dots, X^k)$ of dimension $p \times N$ where X^1, \dots, X^k are data from platform 1 to platform k respectively and $N=N_1+\dots+N_k$;
- b. Rank each row of X^1, \dots, X^k to give $X^1_{\text{rank}}, \dots, X^k_{\text{rank}}$
- c. Calculate sample CDF $p^m(I,J)=(X^m_{\text{rank}}(I,J)-1)/(N_m-1)$ where $I=1,\dots,p$ and $J=1,\dots,N_m$ for each platform $m=1,\dots,k$
- d. The normalized value of $X^m(I,J)$ is the average of $p^m(I,J)$ - quantile of the i th row of $X^1_{\text{rank}}, \dots, X^k_{\text{rank}}$

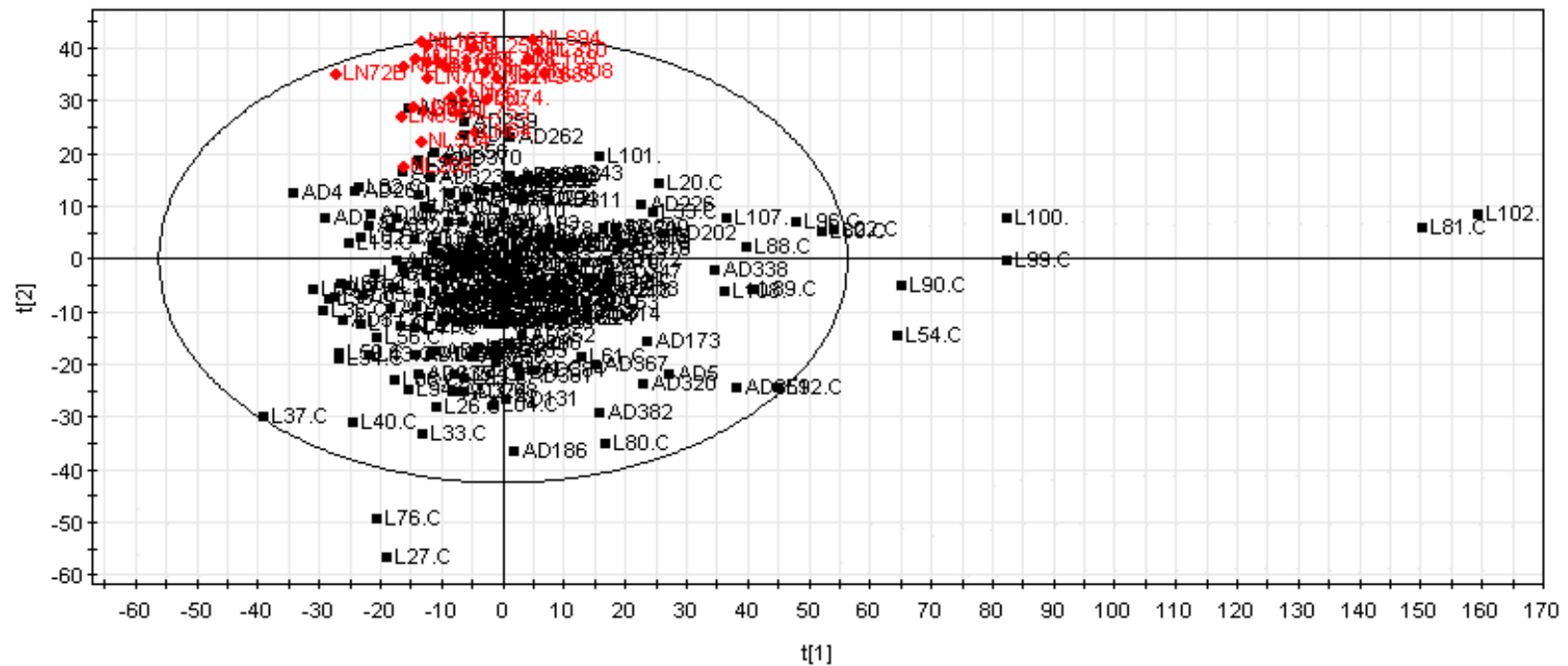


OUTLINE

- CEL files vs Processed Data?
- Data Integration and normalization
- **“Validation” of the integrated data**
 - **Discrimination of Normal lungs from Adenocarcinomas**
 - **Predict one platform from another**
- Survival Analysis
- Discussion



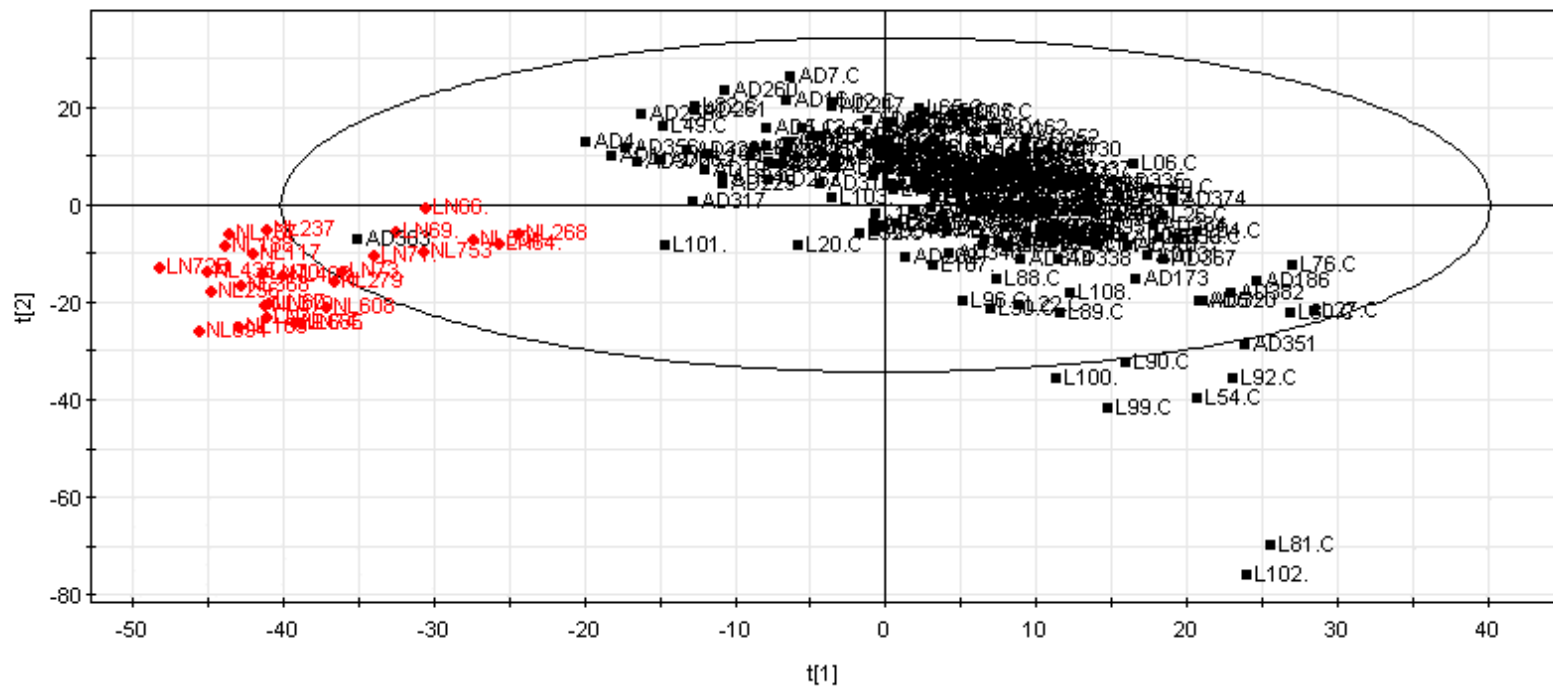
DISCRIMINATION OF NORMAL LUNGS FROM THE ADENOCARCINOMAS



PCA plot after Q-normalization (**Class 1 = normal**, **Class 2 = carcinomas**)



DISCRIMINATION OF NORMAL LUNGS FROM THE ADENOCARCINOMAS



PLS plot after Q-normalization (**Class 1 = normal**, Class 2 = carcinomas)



PREDICTION OF ONE PLATFORM FROM ANOTHER

- Predictive Classification Models were built using the Harvard dataset and then validated by classifying new cases from the Michigan dataset
- Target variable: AD (adenocarcinoma) vs Normal
- Feature selection algorithm: CHAID
- Classification algorithms: C5, CART and NNs

	C5	CART	NN
Sensitivity	98.84%	98.84%	100.00%
Specificity	80.00%	80.00%	90.00%
PPV	97.70%	97.70%	98.85%
NPV	88.89%	88.89%	100.00%
Accuracy	96.88%	96.88%	98.96%

Summary of performance statistics for the 3 predictive classification models on the test data



OUTLINE

- CEL files vs Processed Data?
- Data Integration and normalization
- “Validation” of the integrated data
- **Survival Analysis**
- Discussion



SURVIVAL ANALYSIS

- Identify those genes associated with high risk of mortality.
- 211 patients: 125 from Harvard data and 86 from Michigan data
- Frailty (mixed effects) Cox proportional hazard model
 - Gene expression as fixed effect and
 - Study effect (Harvard vs. Michigan) as random
- Recursive Partitioning Method (Tree method) for Survival Data
 - Exponential scaling is performed to make the observed time follow exponential distribution
 - Recursive partitioning for Poisson data is used



SURVIVAL ANALYSIS

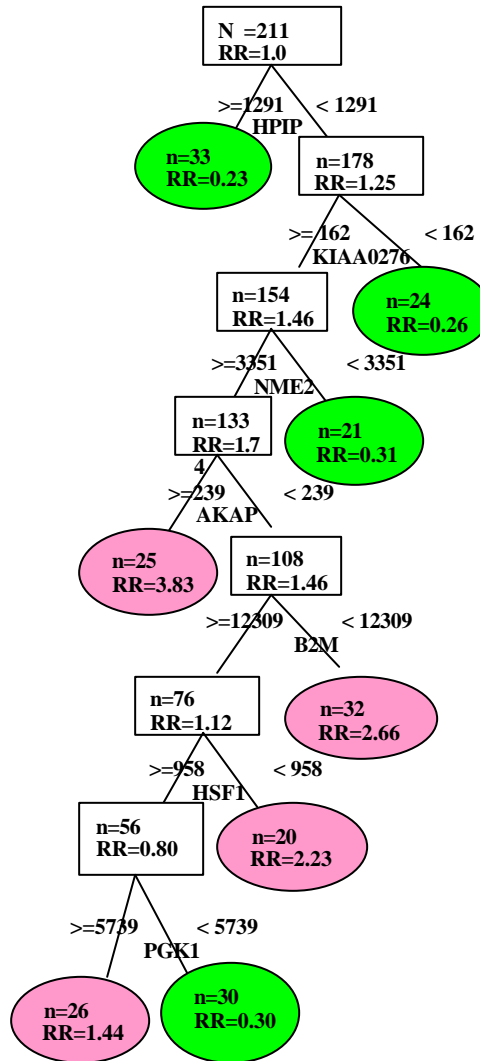
GENE	Coefficient	RAW P-Value	FDR adjusted P-Value
KIAA0211	-0.0025069	0.0000011	0.0054398
CTSL*	0.0002727	0.0000368	0.0312787
KRT18	0.0001400	0.0000490	0.0312787
LHX1	0.0019983	0.0000362	0.0312787
PGK1	0.0001655	0.0000426	0.0312787
PRKCBP1	0.0034964	0.0000275	0.0312787
STX1A*	0.0009447	0.0000517	0.0312787
VEGFC	0.0026009	0.0000309	0.0312787
P4HA1	0.0010053	0.0000646	0.0347284
INHA	0.0009011	0.0001035	0.0484484
RALA	0.0025610	0.0001102	0.0484484

Significant mortality gene list from frailty (mixed effect) Cox Model

*** these two genes appeared in the list in Beer et al (2002)**



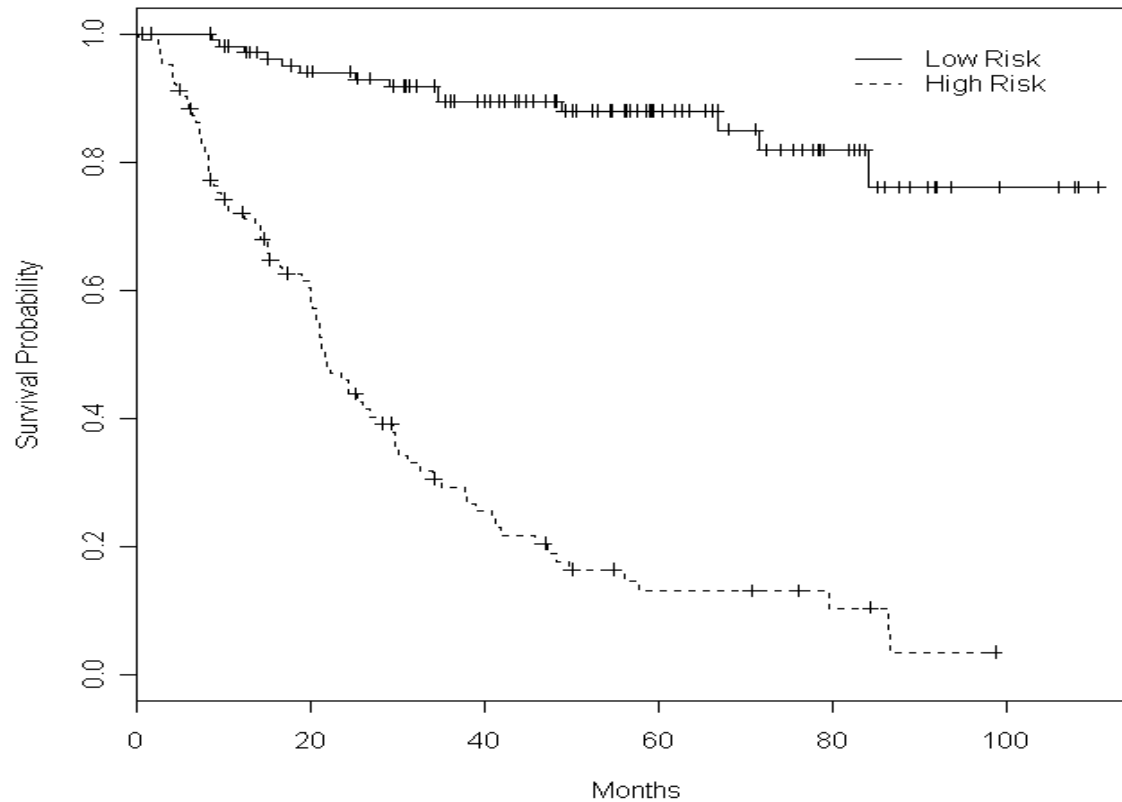
SURVIVAL ANALYSIS



Survival tree diagram



SURVIVAL ANALYSIS



K-M plot by high risk and low risk group defined



OUTLINE

- CEL files vs Processed Data?
- Data Integration and normalization
- “Validation” of the integrated data
- Survival Analysis
- **Discussion**



DISCUSSION

- Summary
 - Integrated data vs individual platform
 - Normalization
 - Survival analysis showed the possibility of using integrated gene expressions data to cluster the adenocarcinoma samples into different mortality risk groups.
- Further work:
 - Covariate information (age, stage, smoking history etc.) might be incorporated in the analysis
 - Apply other statistical and data mining methods
 - Compare results from integrated data with previous published results
 - Further investigation of the results obtained from this survival analysis should be of biological interest.