

SUPPLEMENTARY MATERIALS

Table of Content

1. IDENTIFICATION OF 35-GENE SIGNATURE	2
2. USING THE 35-GENE SIGNATURE TO PREDICT LUNG CANCER OUTCOMES	3
2.1. DESCRIPTION OF EACH PATIENT COHORT	4
2.2. LIST OF OVERLAPPING GENES IN OUR SIGNATURE FOR EACH DATASET	6
2.3. COEFFICIENT-CORRELATION STUDY ON VARIOUS COHORTS	7
2.3.1. <i>Normalization</i>	7
2.3.2. <i>Correlation Coefficient Study</i>	7
3. ASSOCIATION OF GENE-EXPRESSION DEFINED RISK GROUPS AND CLINICAL PARAMETERS.....	9
4. DIFFERENTIAL GENE EXPRESSION IN HIGH-RISK GROUPS	10
5. COMPARISON WITH OTHER GENE SIGNATURES.....	12
5.1. 5-GENE SIGNATURE ANALYSIS	12
5.2. 133-GENE SIGNATURE ANALYSIS	16
6. VALIDATION OF THE GENE SIGNATURE USING RT-PCR ANALYSIS OF LUNG CANCER TUMOR SPECIMENS.....	18

1. Identification of 35-Gene Signature

Table1: The identified 35-gene prognostic signature for non-small cell lung cancer

Genes	Probe set	Function (Unigene comment)	Classification
FCN2	D63160_at	Innate immunity	Immunity
HFL3 (CFHL2)	X64877_s_at	Complement factor H-related	Immunity
IGL@	X57809_s_at	Immunoglobulin lambda locus	Immunity
ATP5A1	D14710_at	ATP synthesis	Metabolism
ATP8A2	U82313_at	ATPase, aminophospholipid transporter-like	Metabolism
FUT7	X78031_at	Glycosylation	Metabolism
GUCA2B	Z70295_at	Endogenous activator of intestinal guanylate cyclase	Metabolism
HRMT1L2	Y10807_s_at	Histone methyltransferase	Metabolism
OGT	U77413_at	Glycosylation	Metabolism
ARHGAP19	U79256_at	Rho GTPase activating protein	Oncogene
EGF	X04571_at	Growth factor	Oncogene
GHRHR	L01406_at	Growth factor receptor, cancer development	Oncogene
INSR	X02160_at	Growth factor receptor: insulin receptor	Oncogene
MT3	M93311_at	Bind to heavy metals	Oncogene
TAL2	HG4068.HT4338_at	T cell leukemogenesis, brain development	Oncogene
TAX1BP2 (VAC14)	U25801_at	Cellular transformation, gene activation	Oncogene
TNFSF9	U03398_at	Tumor necrosis factor family	Oncogene
UBE1	M58028_at	Ubiquitin-activating protein	Protein degradation
UBE2I	U45328_s_at	Ubiquitin-activating protein	Protein degradation
AHNAK	HG180.HT180_at	AHNAK nucleoprotein (AHNAK), transcript variant 2	Signaling transduction
ARHGDIG	U82532_at	Cell signaling protein	Signaling transduction
EMK1 (MARK2)	X97630_a_t	Protein kinase	Signaling transduction
GNB1	X04526_at	Cell signaling transduction	Signaling transduction
LBC (AKAP13)	HG2167.HT2237_at	Scaffolding protein for rho and PKA signaling	Signaling transduction
MSX2	HG3729.HT3999_f_at	Transformation suppressor genes	Signaling transduction
RER1	AJ001421_at	Endoplasmic reticulum membrane proteins	Structure
TUBA3	X01703_at	Encode microtubules	Structure

ATRX	U09820_s_at U72935_cds3_s_at	Transcriptional regulator	Transcription
CHD4	X86691_at	Transcription regulator	Transcription
CREB3	AF009368_at	Transcriptional factor	Transcription
EZFIT (ZNF71)	HG3565.HT3768_r_at	Regulate transcriptional control	Transcription
FBRNP (HNRPA3)	HG1078.HT1078_at	heterogeneous nuclear ribonucleoprotein A3	Transcription
ILF3	U10324_at	Transcriptional factor	Transcription
NP220 (ZNF638)	D83032_at	DNA binding protein pack aging, transferring, or processing transcripts	Transcription
E2F4	U15641_s_at	Transcriptional factor, cell cycle apoptosis	Transcription, Oncogene

2. Using the 35-gene signature to predict lung cancer outcomes

We sought to investigate the predictive power of these 35 genes in assessing lung cancer outcomes. Examined clinical outcomes include overall survival (OS). Kaplan-Meier analyses were carried out with software R.

Beer et al.	PMID: 12118244
Bild et al.	PMID: 16273092
Garber et al.	PMID: 11707590
Raponi et al.	PMID: 16885343

Table 2: Datasets and sub-datasets with their sample size, number of measured signature genes in the datasets and clinical endpoints.

Lung Cancer Datasets	Sample Size	Number of Overlapping Genes	Survival Type
Beer	86	34	OS
Bild	111	34	OS
Garber	24	21	OS
Raponi	129	29	OS

2.1. Description of each patient cohort

BEER: This cohort contains 86 lung cancer patients of which, 67 were of stage 1 and 19 of stage 3. 16 of them had p53 nucleus accumulation positive and 69 had p53 nucleus accumulation negative. The cohort contains 39 K-ras mutation positive and 46 K-ras mutation negative. The histopathology of all 86 patients is Adenocarcinoma. 23 have well differentiation, 41 have moderate and 21 have poor differentiation. 9 patients never smoked and 74 were smokers.

BILD: This cohort contains 111 patients of which, 67 were of stage 1 (including 1A and 1B), 18 of stage 2 (including 2A and 2B), 26 of stage 3A, 3B, or 4. There were two cell types: Adenocarcinoma (58 patients) and Squamous (43 patients).

GARBER: This cohort contains 24 patients. There was 1 sample with grade 1, 11 samples with grade 2 and 12 samples with grade 3. 5 patients were of stage 1A and 2 of stage 1B, 1 of stage 2B, 6 of stage 3A and 10 of stage 4.

RAPONI: This cohort contains 129 patients with Squamous. 4 patients never smoked. 27 were of stage 1A, 46 of stage 1B, 6 of stage 2A, 27 of stage 2B, 17 of stage 3A, and 6 of stage 3B.

Table 3: The clinical and pathological characteristics of patient cohorts

<i>Cohorts</i>	<i>Beer</i> (n=86)	<i>Bild</i> (n=111)	<i>Garber</i> (n=24)	<i>Raponi</i> (n=130)
Parameters				
Age				
<60	28	31		29
>=60	58	80		100
Grade				
1			1	
2			11	
3			12	
Differentiation				
Poor	21			
Moderate	41			
Well	24			
K-ras mutation				
Positive	39			
Negative	46			
NA	1			
p53 nuclear accumulation				
Positive	16			
Negative	69			
NA	1			
Smoking				
Non-Smoker	9			4
Smoker	74			119
NA	3			6
Stage				
1	67	7	-	-
1A	-	33	5	27
1B	-	27	2	46
2	-	1	-	-
2A	-	5	-	6
2B	-	12	1	27
3	19	-	-	-
3A	-	6	6	17
3B	-	15	-	6
4	-	5	-	-
Histopathology				
Adenocarcinoma	86	58	22	-
Squamous	-	53	1	129
Large Cell	-	-	1	-
Other	-	-	-	-

2.2. List of overlapping genes in our signature for each dataset

Table 4: The list of overlapping genes for each dataset

Gene Symbol	Beer	Bild	Garber	Raponi
AHNAK	X	X	X	X
AKAP13 (LBC)	X	X	X	X
ARHGAP19	X	X		X
ARHGDIG	X	X		X
ATP5A1	X	X		
ATP8A2	X	X	X	
ATRX	X	X	X	X
CFHL2 (HFL3)	X	X		X
CHD4	X	X	X	X
CREB3	X	X		X
E2F4	X	X	X	X
EGF	X	X	X	X
EMK1 (MARK2)	X	X		X
FCN2	X	X		X
FUT7	X	X		X
GHRHR	X	X		X
GNB1	X	X	X	X
GUCA2B	X	X		
HNRPA3 (FBRNP)	X	X	X	X
HRMT1L2	X	X	X	X
IGL@	X		X	
ILF3	X	X	X	X
INSR	X	X	X	X
MSX2	X	X	X	X
MT3	X	X		X
OGT	X	X	X	X
RER1	X	X	X	X
TAL2	X	X		
TAX1BP2 (VAC14)	X	X		X
TNFSF9	X	X	X	X
TUBA3	X	X	X	X
UBE1	X	X	X	X
UBE2I	X	X	X	X
ZNF638 (NP220)	X	X		X
ZNF71 (EZFIT)	X	X	X	

2.3. Coefficient-Correlation study on Various Cohorts

2.3.1. Normalization

In order to adjust the gene expression values across different DNA microarray platforms into comparable scales, following normalization methods were used. For Affymetrix data, dChip (1) was used to normalize the samples at probe levels with perfect match (PM) only. A quantile normalization method was then used without \log_2 transformation. Furthermore, the expression values for each gene were sample-wise normalized in each patient using the following formula:

$$\frac{g(x) - \text{mean}}{\text{standard deviation}}$$

where $g(x)$ is the expression of a specific gene, *mean* is the average of all the gene expression values in this patient sample, and *standard deviation* is the statistical standard deviation of all the gene expression levels in this sample. For cDNA data, only sample-wised normalized was applied using the in the formula as described above.

2.3.2. Correlation Coefficient Study

Firstly, using the training dataset from Beer et al (2), we obtained two groups: patients who survived less than 2.5 years as poor-prognosis group, and patients who survived 5 years or longer as good-prognosis group. Then, average expression values of each gene in both groups were computed. Next, we obtained the correlation coefficient between the good-prognosis centroid of the training set and the gene expression signature in each patient sample from the validation datasets from Garber et al (3), Bild et al (4), and Raponi et al (5).

A correlation coefficient of 0.32 was identified as the cutoff from Garber's cohort. This cutoff was applied on Bild's cohort to obtain significant patient stratification. Nevertheless, this cutoff did not stratify patients from Raponi's cohort into distinct prognostic groups. The Raponi's cohort was randomly partitioned into two sets: training set (with 65 samples) and a testing set (with 64 samples). A correlation coefficient of -0.15 with the good prognosis centroid in Beer's cohort was identified as the cutoff from the training set. The same cutoff was applied on the testing set. This stratification generated distinct prognostic groups in both training and testing sets.

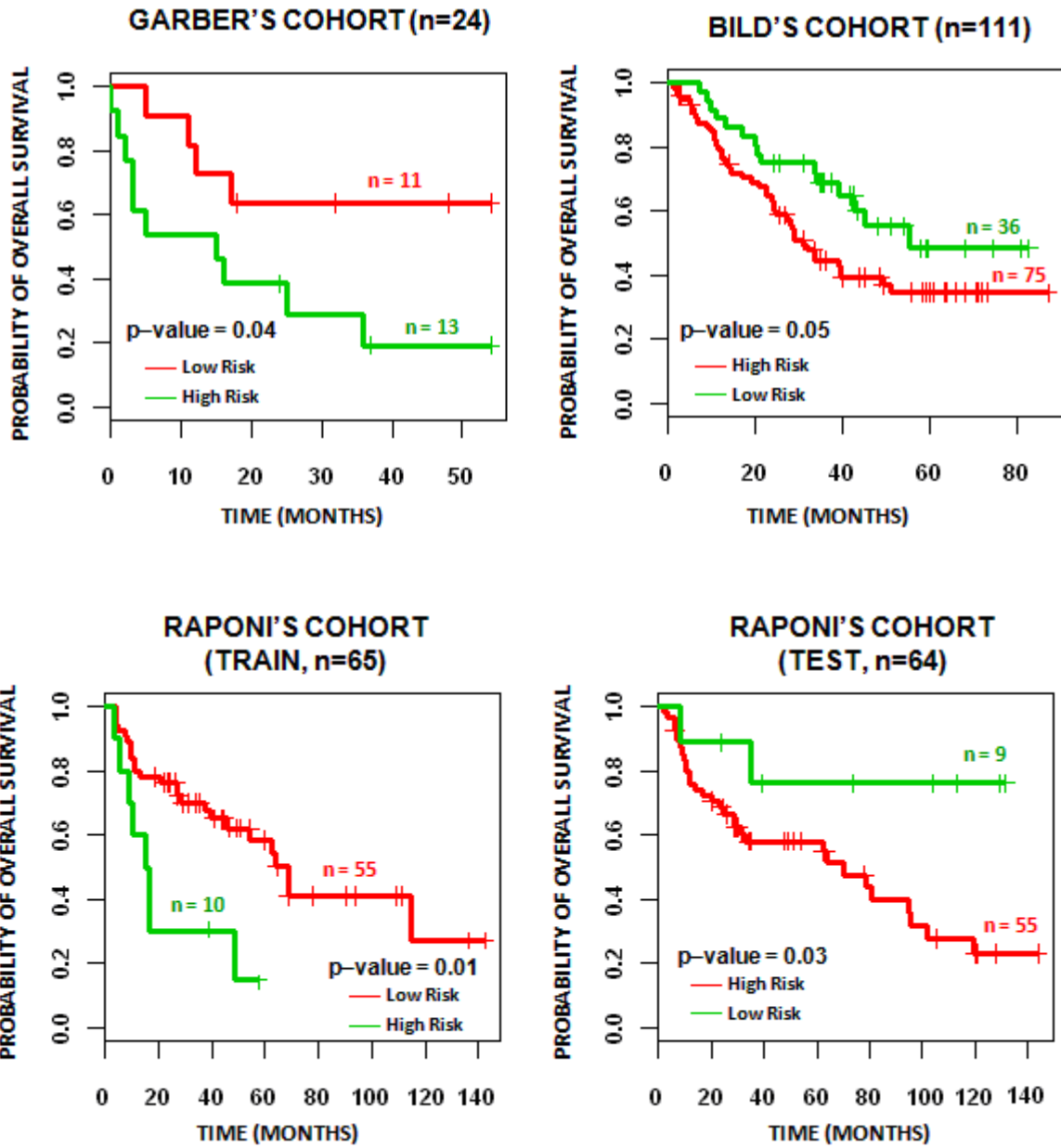


Figure 23: The 35-gene signature stratified patients of the three cohorts into high- and low-risk groups with distinct overall survival based on a nearest centroid method. Patients in these validation sets were classified based on the correlation coefficient of the 35-gene expression signature in the tumor sample and the good prognostic centroid in Beer's cohort.

3. Association of gene-expression defined risk groups and clinical parameters

In this study, we evaluated the association between risk groups and clinical-pathological parameters on the study cohorts by using either Chi-square test or Fisher's exact test (two-sided). Chi square test was used if its assumptions were satisfied. Otherwise, Fisher's test was used. In the datasets from Garber et al, Raponi et al, and Bild et al, the prognostic groups were defined as described above. Table 7 reports the *P* values resulted from the tests. There was no significant association between the prognostic signature and age, smoking status, tumor grade, tumor stage, or differentiation in the studied cohorts. The above section has shown that the 35-gene signature stratifies patients with the same stage into distinct prognostic subgroups, indicating that the prognostic signature is an independent of tumor stage in prognostic prediction.

Table 7: Association between gene expression-defined risks groups (high and low-risk groups) and clinicopathologic parameters in overall survival

<i>P</i>-values	Bild	Garber	Raponi
Age < 60 vs. > 60	1		0.67
Stage	0.16	0.70	0.34
Smoking			
Differentiation			0.44
Grade		0.62	

4. Differential Gene Expression in High-risk Groups

We analyzed the differential gene expression in high-risk vs. low risk groups. Low-risk groups included patients who lived longer than 5 years, and high-risk group included patients who died before 2.5 years from Beer et al. and Bild et al. Low-risk groups from Garber et al. and Raponi et al, included patients who live longer than 3 years.

In the step of comparing the genes, If $\text{mean}(\text{High-risk}) > \text{mean}(\text{Low-risk})$ then we have signed this gene as over-expressed, and marked with red (significant) and light pink (insignificant) in Table 8. If $\text{mean}(\text{High-risk}) < \text{mean}(\text{Low-risk})$ then we have signed this gene as under-expressed, and marked with dark green (significant) and light green (insignificant) in Table 10. P-values are added in the significant boxes. Some genes had multiple probes in one dataset. If multiple probes showed both over-expression and under-expression in high-risk groups, the gene is marked as black.

Table 8: Differential gene expression analysis of the 35-gene signature in the studied patient cohorts.

Gene Symbol	Beer	Bild	Garber	Raponi
AHNAK	0.003			0.019
ARHGAP19	0.0012	0.05		5.87E-05
ARHGDIG	0.00094			
ATP5A1	0.014			
ATP8A2	1.70E-05			
ATRX	5.80E-08			
CHD4	1.30E-06			
CREB3	0.0014			
E2F4	5.20E-06			
EGF	0.0018			
EMK1 (MARK2)	0.00011			
EZFIT (ZNF71)	0.00014			
FBRNP (HNRPA3)	0.00054			
FCN2	7.50E-05			
FUT7	0.00027			
GHRHR	5.40E-05			
GNB1	3.80E-05	0.013		0.0028
GUCA2B	0.016			
HFL3 (CFHL2)	0.00057			
HRMT1L2	0.0085			
IGL@	0.00011		0.011	
ILF3	0.0043		0.0052	
INSR	3.80E-05			
LBC (AKAP13)	0.0033	0.018		
MSX2	1.80E-05			
MT3	3.20E-05			
NP220 (ZNF638)	0.00022			
OGT	0.0023			
RER1	0.026	0.005		0.0048
TAL2	0.0061			
TAX1BP2 (VAC14)	0.046			
TNFSF9	0.0089			0.00062
TUBA3	0.017			
UBE1	3.50E-06			
UBE2I	2.70E-05			

significant under express with p-value	insignificant under express
significant over express with p-value	insignificant over express

5. Comparison with Other Gene Signatures

5.1. 5-Gene Signature Analysis

In order to evaluate the performance of our 35-gene signature, the 5-gene signature model developed in Chen et al. (6) was validated. The five signature genes were extracted from the following datasets:

Beer – 5 overlapping genes:

Gene Symbol	Probe ID
ERBB3	S61953_at
LCK	M26692_s_at
DUSP6	X93920_at
STAT1	M97936_at
MMD	X85750_at

Garber – 10 overlapping probes:

Gene Symbol	SPOT ID	NAME	Clone ID	Cluster ID	Accession No.
ERBB3	2852	14284	IMAGE:486828	Hs.199067	AA042878
ERBB3	3553	8089	IMAGE:267420	Hs.199067	N24966
LCK	6834	2168	IMAGE:730410	Hs.1765	AA420981
LCK	8528	2168	IMAGE:730410	Hs.1765	AA420981
STAT1	1343	14294	IMAGE:545503	Hs.21486	AA079495
STAT1	20994	14265	IMAGE:545242	Hs.21486	AA076085
STAT1	23446	15527	IMAGE:840691	Hs.21486	AA486367
DUSP6	1337	14161	IMAGE:1337808	Hs.180383	AA811335
DUSP6	16905	19954	IMAGE:854899	Hs.180383	AA630374
MMD	3871	16111	IMAGE:841331	Hs.79889	AA487434

Bild – 15 overlapping probes:

Gene Symbol	Probe ID
ERBB3	1563252_at
ERBB3	1563253_s_at
ERBB3	202454_s_at
ERBB3	215638_at
ERBB3	226213_at
LCK	204890_s_at
LCK	204891_s_at
DUSP6	208891_at
DUSP6	208892_s_at

DUSP6	208893_s_at
STAT1	200887_s_at
STAT1	209969_s_at
STAT1	232375_at
MMD	203414_at
MMD	244523_at

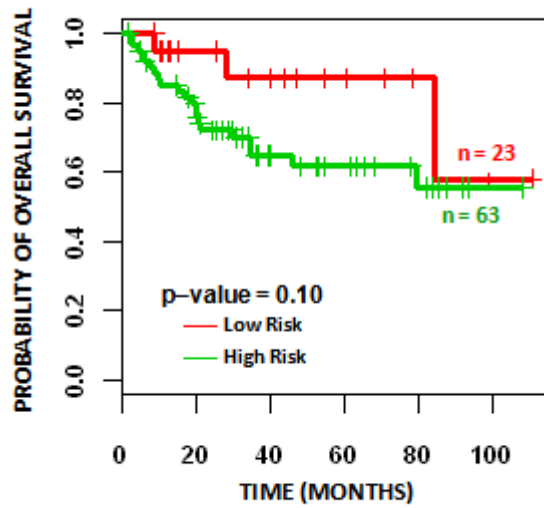
Raponi – 10 overlapping probes:

Gene Symbol	Probe ID
ERBB3	202454_s_at
ERBB3	215638_at
LCK	204890_s_at
LCK	204891_s_at
DUSP6	208891_at
DUSP6	208892_s_at
DUSP6	208893_s_at
STAT1	200887_s_at
STAT1	209969_s_at
MMD	203414_at

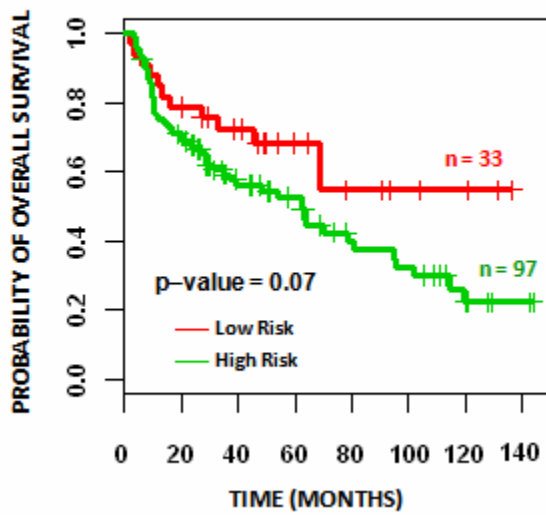
The gene expression values were log-transformed and normalized as described in Chen et al (6). Patients were stratified based on the classification tree model (Supplementary Figure 1 in Chen et al (6)). The classification tree was implemented with nested if-else statements in R code. Based on the patient stratification, Kaplan-Meier analysis and log-rank tests were performed to assess the difference in the survival rates of high- and low-risk groups. For datasets having duplicated probes for one gene symbol, two approaches were performed in the analysis:

- One of the duplicated probes of the same gene was randomly selected.
- Average gene values of duplicated probes were used; results are shown in Raponi et al. (average), Bild et al. (average), and Garber et al. (average).

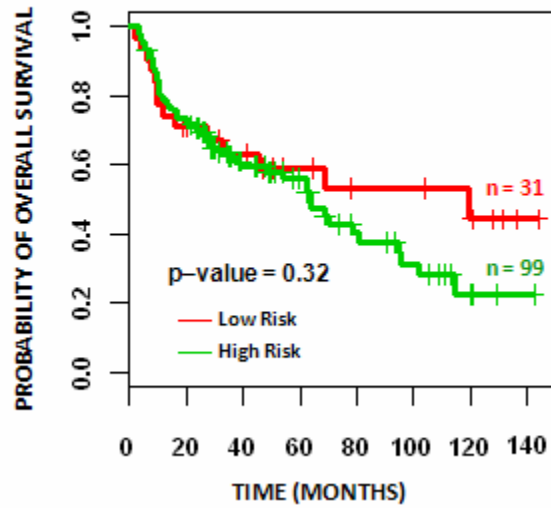
BEER'S COHORT (n=86)



RAPONI'S COHORT (n=130)



RAPONI'S COHORT (average, n=130)



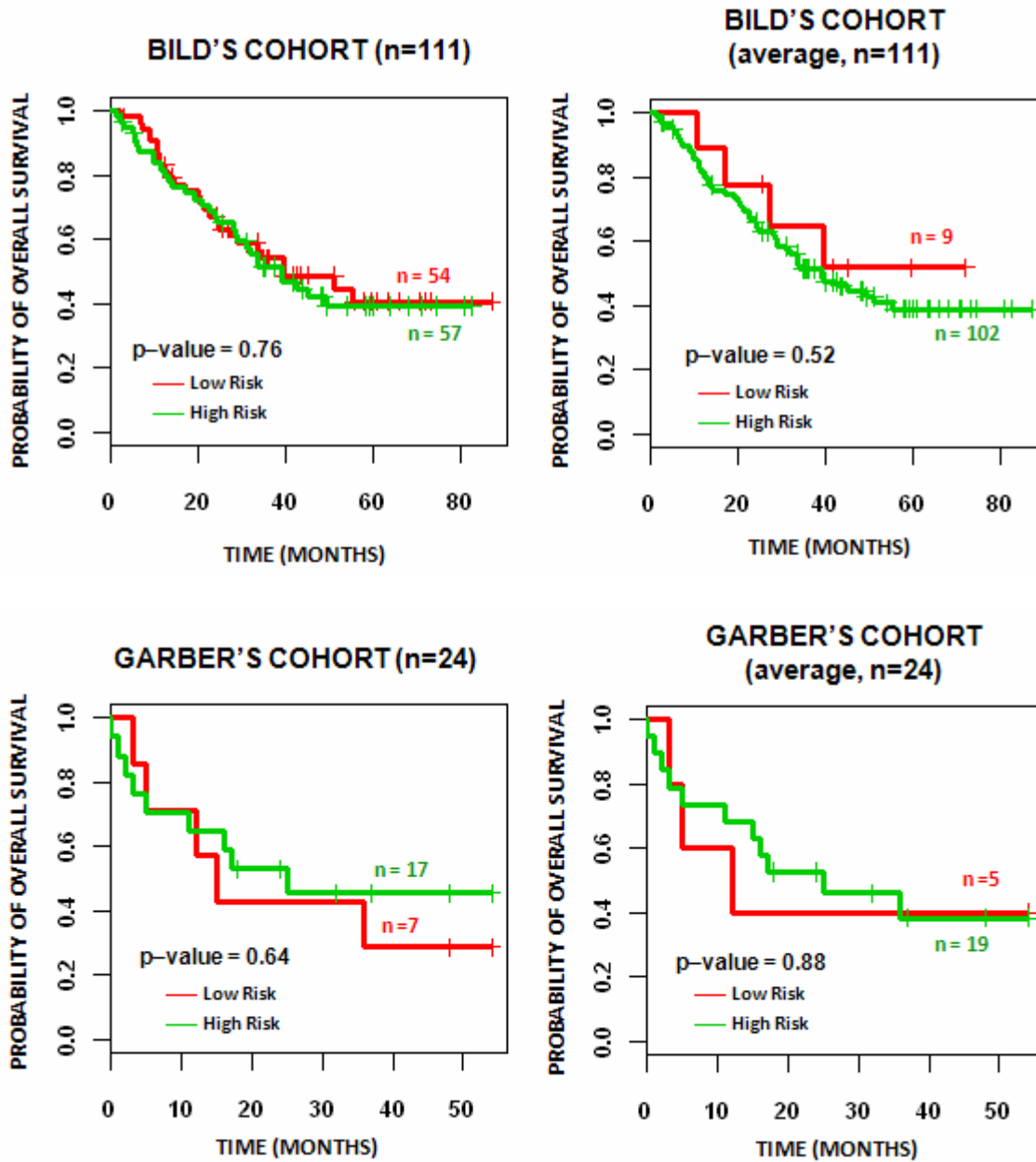


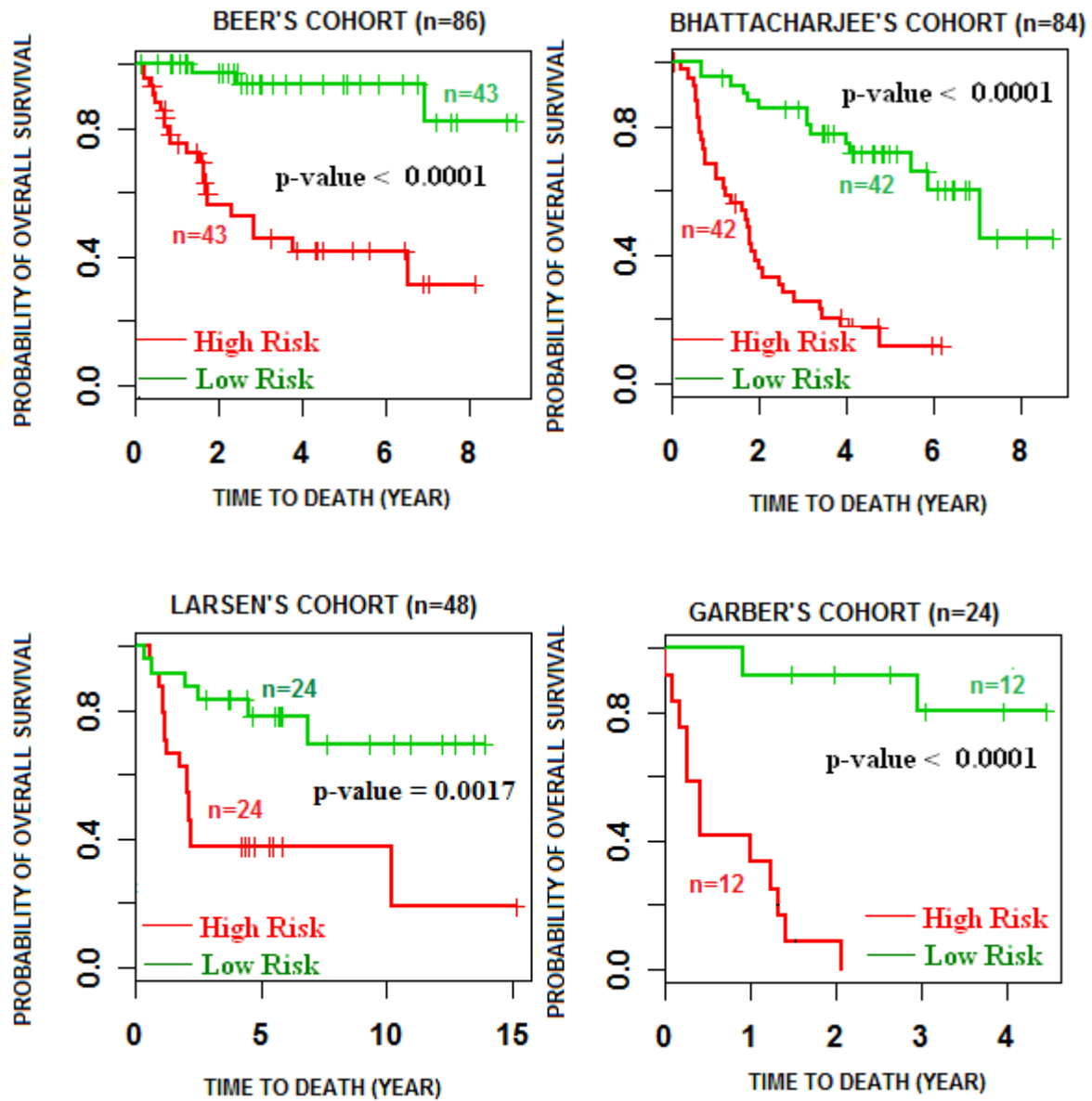
Figure 24: Patient stratification based on the 5-gene model from Chen et al.

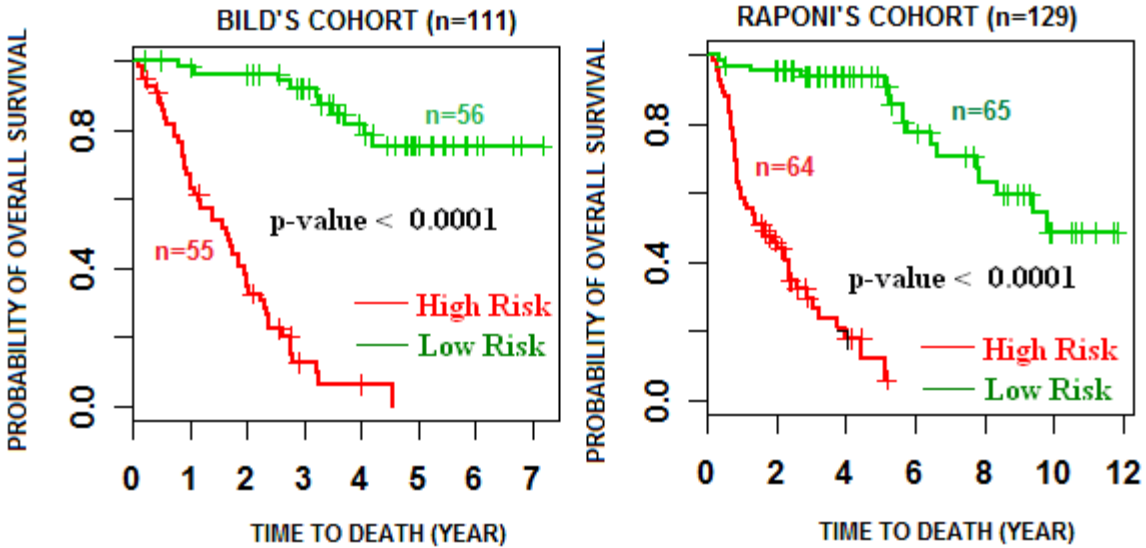
5.2. 133-Gene Signature Analysis

The 133-gene signature from Potti et al. (7) was evaluated in the validation sets. There are 67 overlapping gene in Beer's, 18 in Bhattacharjee's, 13 in Garber's, 14 in Larsen's datasets. Their area under the time dependent ROC curve (AUC) and Kaplan Meire plots are given below.

There was a convergence and fitting problem in Bild's and Raponi's datasets. There are 489 overlapping probe set for Bild's and 302 overlapping probe set for Raponi's datasets. Mean value of multiple probes have been taken to obtain a unique expression for each gene. There are 123 overlapping genes for both datasets. But with using the entire 123 gene we couldn't fit the Cox model in R. Problem occurred in that part. We could fit the model with 71 or less genes for Bild's dataset and 88 or less genes for Raponi's dataset. Correlation coefficient of each gene calculated and used to get these 71 genes ($|\text{corr}| > 0.44$) and 88 genes ($|\text{corr}| > 0.46$). AUC figures are given below.

Kaplan-Meier Analysis for 133 gene signature





6. Validation of the Gene Signature using RT-PCR analysis of lung cancer tumor specimens

RNA extraction

Total RNA was extracted from frozen lung tissue using the RNeasy mini kit according to the manufacturer's protocol (Qiagen, USA). RNA was eluted in 30 μ l of RNase-free water and stored at -80°C . The quality and integrity of the total RNA was evaluated on the 2100 Bioanalyzer (Agilent Technologies, CA).

Reverse transcription

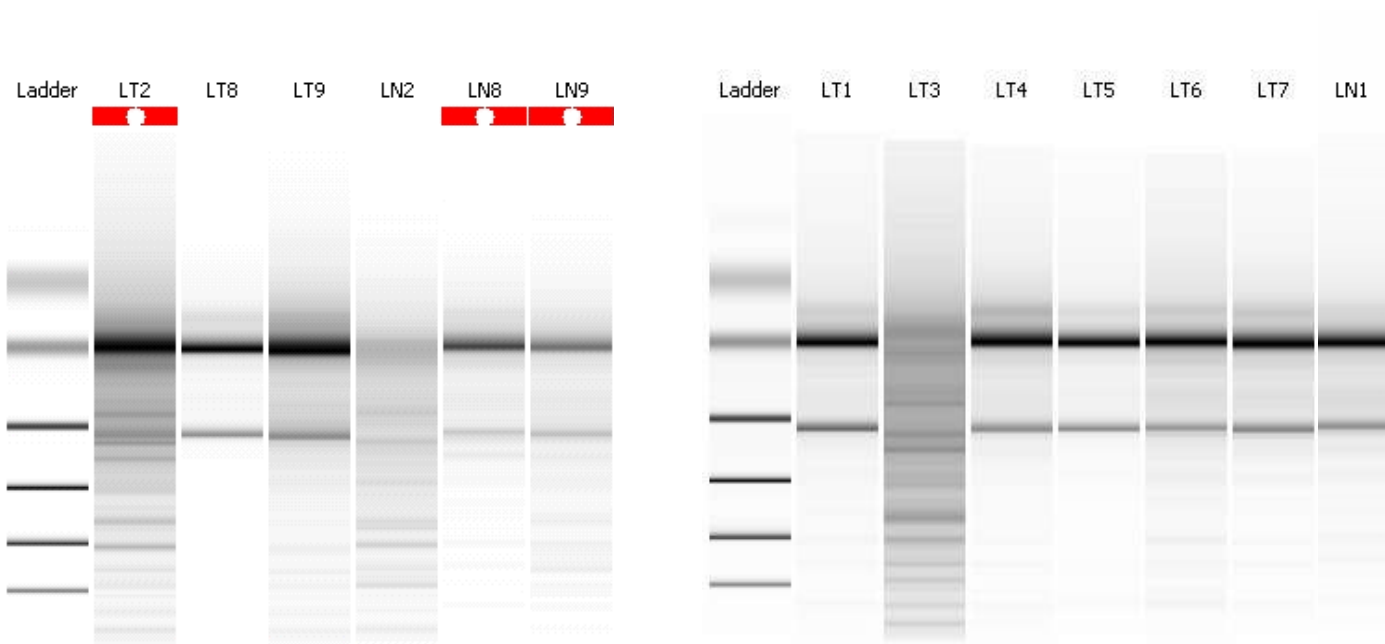
From each sample, 1 μ g of RNA was used to generate cDNA using the High Capacity cDNA kit according to manufacturer's protocol (Applied biosystems, CA).

RT-PCR

An endogenous control gene Taqman Low Density Array card (Applied Biosystems, CA) was run on the ABI PRISM 7900HT Sequence Detection System for samples LT1, LT4, LT6, LT8, LT9, LN1, LN8 and LN9 in order to choose a gene that had the most relatively constant expression in the different tissue samples. Three genes, namely 18S, UBC and POLR2A had constant expression in the different tissue samples. Constant expression of mRNA for 18S and UBC genes was also confirmed for all lung tissue samples using the individual TaqMan[®] Gene Expression Assays.

Expression of mRNA for 35 genes was measured in each of the lung tissues by real-time PCR using TaqMan[®] Gene Expression Assays on ABI PRISM 7500 HT Sequence Detection System (Applied Biosystems, CA). Due to sample limitations, each gene was amplified only once per sample. However, samples LT4, LT7, LT8, LN8 and LN9 were run in duplicates for the following genes: UBC, E2F4, GNB1, ILF3, EGF, MT3 and RER1 in order to confirm consistent pipetting technique. On each plate, one no-template control was also run.

Total RNA samples run on an Agilent 2100 Bioanalyzer RNA 6000 Nano LabChip.



References

- (1) Li C. Automating dChip: toward reproducible sharing of microarray data analysis. *BMC Bioinformatics* 2008;9:231.
- (2) Beer DG, Kardia SL, Huang CC, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* 2002;8:816-24.
- (3) Garber ME, Troyanskaya OG, Schluens K, et al. Diversity of gene expression in adenocarcinoma of the lung. *Proc Natl Acad Sci U S A* 2001;98:13784-9.
- (4) Bild AH, Yao G, Chang JT, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 2006;439:353-7.
- (5) Raponi M, Zhang Y, Yu J, et al. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Res* 2006;66:7466-72.
- (6) Chen HY, Yu SL, Chen CH, et al. A five-gene signature and clinical outcome in non-small-cell lung cancer. *N Engl J Med* 2007;356:11-20.
- (7) Potti A, Mukherjee S, Petersen R, et al. A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. *N Engl J Med* 2006;355:570-80.