Genome wide copy number variation pattern analysis and a classification signature for Non-small cell lung cancer

Supplementary Materials

Jia-hao Bi[1], Zhe-wei Qiu[1], Adi F. Gazdar[2,3] and Kai Song[1,2]

[1] School of Chemical Engineering and Technology, Tianjin University, 300072 Tianjin, P.R. China;

[2]Hamon Center for Therapeutic Oncology, [3]Department of Pathology, University of Texas Southwestern Medical Center, 75390, Dallas, Texas, USA;

Correspondence should be addressed to: Kai Song (ksong@tju.edu.cn) or Adi F. Gazdar (adi.gazdar@utsouthwestern.edu)

Molecular signature identification based on the histologic classification of NSCLC

The big challenge in identifying the molecular signature through the microarray data is that the number of variables (e.g. genes, probes and so on) is usually much bigger than the number of available samples (e.g. LUAD and LUSC patients and so on). Additionally, inherent high noise of microarray data and the complicated multi-relationship among genes make it a much more challenging issue. Therefore, to improve the classification accuracy and to identify the real signature genes, T-test, Elastic Net (EN), Partial least squares (PLS) and Naïve Bayes were used together for the complementary advantages of each other.

- T-test is a simple univariate significant testing method. It cannot consider the multi-relationships among CNVs of all genes but it can select genes with significantly different CNVs between LUAD and LUSC. Hence, it was used to roughly select important genes.

- Elastic Net (EN) is a regularized regression method that linearly combines the L_1 and L_2 penalties of the lasso (least absolute shrinkage and selection operator) and ridge regression methods (Tibshirani 1996). It cannot overcome the multi-relationships among genes but it is capable of selecting groups of correlated genes (Zou and Hastie 2005).

- Partial least squares (PLS) algorithm is an efficient statistical regression or classification technique that is highly suited for the analysis of high-dimensional data. It is a powerfully proven method for analyzing genomic and proteomic data, especially for the problems of classification and dimension reduction in bioinformatics (Nguyen and Rocke 2002; Song, et al. 2012).

- Naïve Bayes (NB) classifier is a basically probabilistic classifier based on Bayes' theorem with strong assumptions that features are independent to each other. Therefore, the main drawback of NB is that it cannot overcome the multi-relationships among genes (Caruana and Niculescu-Mizil 2006; Rish 2001). In this study, the orthogonal LVs extracted by PLS from the CNVs of signature genes

were used as the new input variables to NB.

To overcome the inherent high noise of microarray data and the complicated multi-relationship among genes and to select potential signature genes distinguishing LUADs and LUSCs out of approximately twenty thousand genes, an algorithm integrating Elastic Net (EN) (Hughey and Butte 2015), partial least squares (PLS) (Song 2012), and naive byes (NB) (Langarizadeh and Moghbeli 2016) was applied to complement the advantages of each other. The integrated EN-PLS-NB algorithm is described below with the corresponding flowchart in Figure S1A.

(1) An initial step before applying the algorithm is to select the top 10,000 genes ranked by the *p*-values of two-tailed *t*-test comparing CNV values between LUAD and LUSC samples. This step reduces both random noise and computational burden for EN in step (2) by omitting genes with negligible differences between LUAD and LUSC.

(2) Apply EN method to select the top genes out of the 10,000 genes at a step size of 100. A nested cross-validation (CV) procedure with 10-fold and 3-fold CV for the inner and outer loops, respectively, was employed. This step exploits the advantage of EN in cooperating group genes selection.

(3) Apply the PLS method to a list of genes, which, initially, is the output of step (2). Sort the genes by the absolute coefficient values in descending order and derive the orthogonal LVs. A 5-fold CV procedure was employed. PLS compensates the shortage of EN that cannot overcome the multi-correlation among genes.

(4) Apply NB classifier on the orthogonal LVs derived from the PLS analysis to calculate classification accuracy of NSCLC by a 5-fold CV. The orthogonality of LVs can improve the classification accuracy.

(5) Remove genes at the bottom of the list in step (3) and repeat steps (3)-(4) until only one gene is left.

(6) Select the gene set with the highest prediction accuracy and the best balance between sensitivity and specificity (defined as below) as the signature gene set and the corresponding classification model as the final one.

Each method is briefly summarized as follows:

Elastic Net

The elastic net is based on the lasso, which is a penalized least squares method imposing an L_1-penalty on the regression coefficients. The elastic net is a regularized regression method that linearly combines the L_1 and L_2 penalties of the lasso and ridge regression methods (Zou and Hastie 2005), and for any fixed non-negative λ_1 and λ_2, the elastic net can be formulated as:

$$L(\lambda_1, \lambda_2, \beta)=|y-\mathbf{X}\beta|^2+\lambda_2|\beta|^2+\lambda_1|\beta|_1 \quad (1)$$

where $|\beta|^2 = \sum_{j=1}^{p} \beta_j^2, |\beta|_1 = \sum_{j=1}^{p} |\beta_j|$. The elastic net estimator (1) is seen to be equivalent to the minimizer of: $\hat{\beta} = arg\, min_{\beta} \{L(\lambda_1, \lambda_2, \beta)\}$. On setting $\alpha=\lambda_2/\lambda_1+\lambda_2$, solving $\hat{\beta}$ in equation (1) is equivalent to the optimization problem:

$$\hat{\beta} = arg\, min_{\beta} |y - \boldsymbol{X}\beta|^2, \text{ subject to } (1-\alpha)|\beta|_1+\alpha|\beta|^2 \leq t \text{ for some } t \quad (2)$$

The function $(1-\alpha)|\beta|_1+\alpha|\beta|^2$ is a convex combination of the lasso and ridge penalty. The elastic net simplifies to simple ridge regression when $\alpha=1$ and to the lasso when $\alpha=0$. The L_1 part of the elastic net does automatic variable selection, while the L_2 part encourages grouped selection and stabilizes the solution paths with respect to random sampling, thereby improving prediction.

Partial least squares

Partial least squares (PLS) is an efficient statistical regression technique that is highly suited for the analysis of high-dimensional data, a powerfully proven method for analyzing genomic and proteomic data, especially problems of classification and dimension reduction in bioinformatics and genomics (Nguyen and Rocke 2002; Song, et al. 2012).

Suppose that the data $\mathbf{X}$ is an n×p matrix of n samples and p genes (the raw data set should be scaled to zero mean and unit variance), and let $\mathbf{Y}$ denote the n×q vector of response values, such as the indicator of classification of LUAD and LUSC. When n<p, the usual regression tools such as ordinary least squares (OLS), cannot be applied since the p×p covariance matrix $\mathbf{X}^T\mathbf{X}$ is singular. In contrast, PLS may be applied also to the cases, whose aims is to describe linear relationship between the predictor matrix $\mathbf{X}\in\mathbf{R}_{n\times p}$ and the response $\mathbf{Y}\in\mathbf{R}_{n\times q}$,

$$\mathbf{Y}=\mathbf{XB}+\mathbf{V} \tag{3}$$

where $\mathbf{B}\in\mathbf{R}_{p\times q}$ is the regression coefficient matrix and $\mathbf{V}\in\mathbf{R}_{n\times q}$ is the residual matrix. PLS regression is based on the basic principal component decomposition:

$$\mathbf{Y}=\mathbf{TQ}^{T}+\mathbf{F} \tag{4}$$

$$\mathbf{X}=\mathbf{TP}^{T}+\mathbf{E} \tag{5}$$

where $\mathbf{T}\in\mathbf{R}_{n\times m}$ is the latent variables (LVs) matrix, $\mathbf{P}\in\mathbf{R}_{p\times m}$ and $\mathbf{Q}\in\mathbf{R}_{q\times m}$ are matrices of coefficients, $\mathbf{E}\in\mathbf{R}_{n\times p}$ and $\mathbf{F}\in\mathbf{R}_{n\times q}$ are matrices of random errors, m is the number of LVs.

From equation (3), (4), and (5), the **T** is the key. The objective criterion for constructing components in PLS is to sequentially maximize the covariance between the response variable and a linear combination of the predictors. That is, in PLS, the components are constructed to maximize the objective criterion based on the sample covariance between **Y** and **XW**, thus,

$$w_k = arg\,max\, cov^2_{w^T w=1}(\boldsymbol{XW}, \boldsymbol{Y}) \tag{6}$$

Subject to the orthogonal constraint,

$$w_k^T \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{W}_i = 0 \quad \text{for all} \quad 1\le k<i \tag{7}$$

where $\mathbf{W}\in\mathbf{R}_{p\times m}$ is a matrix of weights.

To derive the T, PLS can all be seen as methods to construct a matrix of latent components T as a linear transformation of **X**,

$$\mathbf{T}=\mathbf{XW} \tag{8}$$

If **T** is constructed, $\mathbf{Q}^{T}$ and is obtained as the least squares solution of Equation (4):

$$\mathbf{Q}^{T}=(\mathbf{T}^{T}\mathbf{T})^{-1}\mathbf{T}^{T}\mathbf{Y} \tag{9}$$

The matrix B regression coefficients matrix is constructed from Equation (3):

$$\mathbf{B}=\mathbf{W}(\mathbf{T}^{T}\mathbf{T})^{-1}\mathbf{T}^{T}\mathbf{Y} \tag{10}$$

The number of LVs is the only parameter of PLS which need to be decided, with the increase of LVs, the information of original data preserved is increasing, until reaching the maximal value, which is the rank of **X**, all the information of original data is contained in LVs.

Naïve Bayes classifier

The Naïve Bayes Classifier (NB) is a simple probabilistic classifier based on Bayes' theorem with strong assumptions that the feature values are conditionally

independent given the class. Despite their naive design and apparently oversimplified assumptions, the Naïve Bayes Classifier is outperformed by other approaches, such as boosted trees or random forests (Caruana and Niculescu-Mizil 2006; Rish 2001). Given a new sample observation, NB estimates the conditional probabilities of classes using the joint probabilities of training sample observations and classes,

$$p(C|F_1, F_2, \dots, F_p) = \frac{p(C)p(F_1,F_2,\dots,F_p|C)}{p(F_1,F_2,\dots,F_p)} \quad (11)$$

Because the denominator does not depend on C and the values of the features F_i are given, the denominator is a constant, and there is interest only in the numerator of that fraction. On the basis of the joint probability model and the conditional probability model:

$$\begin{aligned} p(C)p(F_1, F_2, \dots, F_p|C) &= p(C, F_1, F_2, \dots, F_p) \\ &= p(C)p(F_1|C)p(F_2, \dots, F_p|C, F_1) \\ &= p(C)p(F_1|C)p(F_2|C, F_1) \dots p(F_p|C, F_1, F_2, \dots, F_{p-1}) \end{aligned} \quad (12)$$

the 'naive' conditional independence assumptions come into play: assume that each feature F_i is conditionally independent of every other feature F_j for $i \neq j$ given the category C: $p(F_i|C, F_j) = p(F_i|C)$, $p(F_i|C, F_j, F_k) = p(F_i|C)$ and so on, for $i \neq j,k$. The conditional distribution over the class variable C is:

$$p(C|F_1, F_2, \dots, F_p) = \frac{p(C)}{p(F_1,F_2,\dots,F_p)} \prod_{i=1}^{p} p(F_i|C) \quad (13)$$

Given a sample with f_1, f_2,… f_p, value, we can know the probability of belonging to this class C.

Classification scores

The classification scores of each sample in the training set and validation set were calculated whose magnitude can be viewed as an estimate of the prediction's confidence. All scores were normalized between -100 to +100. We interpret positive scores as predicting ADC histology while negative scores predict SCC histology. The cutoff scores of the grey area were chosen as

±STD [abs(all scores of training samples)] (14)

where STD is the standard deviation and abs is the absolute value.

The cutoff is equal to ±18.28 in the TCGA training set and can be viewed as a prediction threshold: values above 18.28 are predicted to be LUAD while values below -18.28 are predicted to be LUSC. Intermediate values are predicted as poorly differentiated.

Performance measurements

To evaluate the performance of the classification of NSCLC, prediction accuracy, specificity and sensitivity defined as follows:

$$Accurary = \frac{TP + TN}{TP + TN + FP + FN}, \tag{15}$$

$$Sensitivity = \frac{TP}{TP + FN}, \tag{16}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{17}$$

where TP, FP, TN and FN denote true positive, false positive, true negative, and false negative, respectively. For example, when classifying NSCLC, LUADs and LUSCs were designated as the positive and negative samples, respectively. Correspondingly, sensitivity is the proportion of LUADs correctly classified, specificity is the proportion of LUSCs correctly classified, and accuracy is the proportion of both types of samples correctly classified.

Five genes are relevant to lung cancer

The 33 signature genes selected for the classification model included two genes (SOX2 and PIK3CA) which are known to be involved in lung cancer pathogenesis. We explored whether any of the other 31 genes had a similar relevance by searching the PubMed database using the Medical Subject Headings (MeSH) terms lung carcinoma and the gene name. Five of the genes were found to have relevance to lung

cancer.

- Adiponectin (ADIPOQ) gene polymorphisms may play a role in the susceptibility and prognosis of NSCLC and increased plasma levels are found after targeted therapy of EGFR mutant tumors (Cui, et al. 2011; Umekawa, et al. 2013).
- ABCF3 and ABCC5 are members of the ATP-binding cassette (ABC) superfamily of proteins and ABCC5 is associated with cisplatin resistance in lung cancer (Weaver, et al. 2005).
- Serpin Peptidase Inhibitor, Clade I Member 2 (SERPINI2) is a member of a family of proteins that acts as inhibitors of serine proteases and is unregulated in LUSC tumors and in the bronchial epithelium of smokers (Boelens, et al. 2009).
- The hormone somatostatin (SST) is differentially over expressed in LUSCs and has been proposed as a potential target for novel therapies (Kang, et al. 2009).
- Recombination signal-binding protein Jkappa (RBPJ) is a key transcription factor downstream of receptor activation in Notch signaling pathway, and may stimulate lung growth (Lv, et al. 2015).

Table S1. Summary of patient clinical parameters

	TCGA		SPORE		EDRN/Canary
	LUAD	LUSC	LUAD	LUSC	LUAD
Number	496	490	105	46	79
Age	65	67	64	68	-
Gender					
Female	238	85	53	13	55
Male	200	264	52	33	24
Unknown	58	141	-	-	-
Stage					
Stage I	245	175	59	16	-
Stage II	99	101	15	9	-
Stage III	72	65	28	21	-
Stage IV	21	4	3	-	-
Unknown	59	145	-	-	79
Smoking History					
Smoker	356	329	91	45	51
Non-smoker	68	13	14	1	28
unknown	72	148	-	-	-
Vital status					
Dead	106	124	68	19	-
Alive	332	225	37	27	-
Unknown	58	141	-	-	79
Chemotherapy					
with	92	57	11	11	-
without	359	292	94	35	-
Unknown	45	141	-	-	79

-: Not available

*In TCGA dataset, due to the missing of clinical information of some samples, the summary is only come from part of it.

Table S2. Summary of the datasets

Dataset	Tumor type	Sample Number	Usage
TCGA	LUAD	496	Training samples
	LUSC	490	Training samples
	Non-malignant	556+523	Validation samples
SPORE	LUAD	105	Validation samples
	LUSC	46	Validation samples
EDRN/Canary	LUAD	79	Validation samples

Table S3. Tumor associated genes for lung cancer

No	Gene symbols	Location	p-value[1]	Tumor types[2]
1	***PIK3CA***	**3q26.3**	**1.55E-78**	**LUSC**
2	***SOX2***	**3q26.33**	**7.48E-57**	**LUSC**
3	***FGFR1***	**8p11**	**2.57E-10**	**LUSC**
4	***NKX2-1***	**14q33.3**	**6.32E-7**	**LUAD**
5	*TERT*	5p15.33	0.01	Both
6	*EGFR*	7p11.2	0.03	Both
7	*MYCL*	1p34.2	0.42	Neither
8	*MYC*	8q24.21	0.42	Both
9	*ERBB2 (HER2)*	17q12	0.43	Both
10	*MDM2*	12q15	0.54	Both
11	*MET*	7q.31	0.97	Both

1. p-value: two-tailed *t*-test between LUADs and LUSCs, Bonferroni correction cutoff 2.10E-6; Genes whose p-values lower than the cutoff were shown in boldface.

2. Tumor type:

- *LUSC: the median deflections of genes were greater for LUSC tumors;*
- *LUAD: the median deflections of genes were greater for LUAD tumors;*
- *Both: the median deflections of genes for both LUSC and LUAD tumors were big enough and were similar to each other;*
- *Neither: no big median deflections of genes in LUSC or LUAD tumors*

Table S4 Percentage of significantly different genes among LUAD, LUSC and non-malignant samples (TCGA only)

	LUAD vs Non-malignant	LUSC vs Non-malignant	LUAD vs LUSC
Total	65.05%	64.74%	47.3%
Amplification	34.35%	37.01%	18.5%
Deletion	30.70%	27.73%	28.8%

Table S5. The classification results between LUAD and LUSC obtained by 7 CNV biomarkers*

Gene set	Dataset	Sensitivity	Specificity	Accuracy
LUAD vs LUSC	TCGA	0.81	0.94	0.88
	SPORE	0.83	0.74	0.80
	EDRN/Canary	0.97	NA	NA
Tumor vs non-malignant	LUAD (TCGA)	0.83	0.99	0.91
	LUSC (TCGA)	0.99	0.96	0.97

*The 7 genes are the top 7 genes in Table 2. There are only LUAD samples in EDRN/Canary dataset, so only Sensitivity measurement can be calculated. There we used Sensitivity as the accuracy of EDRN/Canary dataset.

Table S6. Genome-wide CNV difference among different tumors

	Tumor types	Overall	Gains	Losses
Adenocarcinomas	LUAD vs. LUSC	47.3%	28.8%	18.5%
	LUAD vs. CRCA	52.3%	22.3%	30.0%
	LUAD vs. BRCA	48.0%	23.2%	14.8%
	LUAD vs. PRAD	48.4%	23.1%	15.3%
	LUAD vs. OV	55.0%	33.4%	21.6%
Squamous cells	LUSC vs. HNSC	42.5%	23.6%	18.9
	LUSC vs. ESSC	2.9%	1.8%	1.1%

Table S7. The percentage of significant genes compared with the whole genome genes

Sample ID	Predicted score	NKX2-1 (TTF-1)	NAPSA	TP63	KRT5
Diagnosed as LUAD samples by TCGA					
TCGA-44-5643-01	-72.32	5.74	9.76	13.52	17.49
TCGA-50-5931-01	-51.75	5.21	7.78	11.64	15.30
TCGA-50-6595-01	16.15	8.21	10.85	9.81	10.52
TCGA-55-7726-01	17.93	5.63	9.84	11.81	15.40
TCGA-55-8204-01	20.86	7.63	9.06	13.05	16.69
TCGA-62-A471-01	30.94	7.19	10.89	9.30	13.01
TCGA-64-1679-01	10.11	10.02	11.07	9.67	10.30
TCGA-75-6214-01	15.40	8.22	10.44	11.94	13.64
Diagnosed as LUSC samples by TCGA					
TCGA-22-1017-01	16.33	11.82	15.06	6.26	4.83
TCGA-43-2581-01	11.94	11.18	14.99	5.98	6.11
TCGA-60-2714-01	4.41	12.01	14.61	4.80	7.40
TCGA-63-6202-01	17.71	11.29	14.70	4.78	6.67
TCGA-85-A513-01	14.90	12.06	15.87	5.23	8.66
TCGA-NC-A5HJ-01	2.52	10.79	14.70	5.74	8.17
TCGA-O2-A52Q-01	13.45	10.71	12.92	6.16	7.08

Note: The cutoff values of four markers are: 10.3 (NKX2-1), 12.5 (NAPSA), 8.2 (TP63), 9.9 (KRT5). All expression values are log2 transformed.

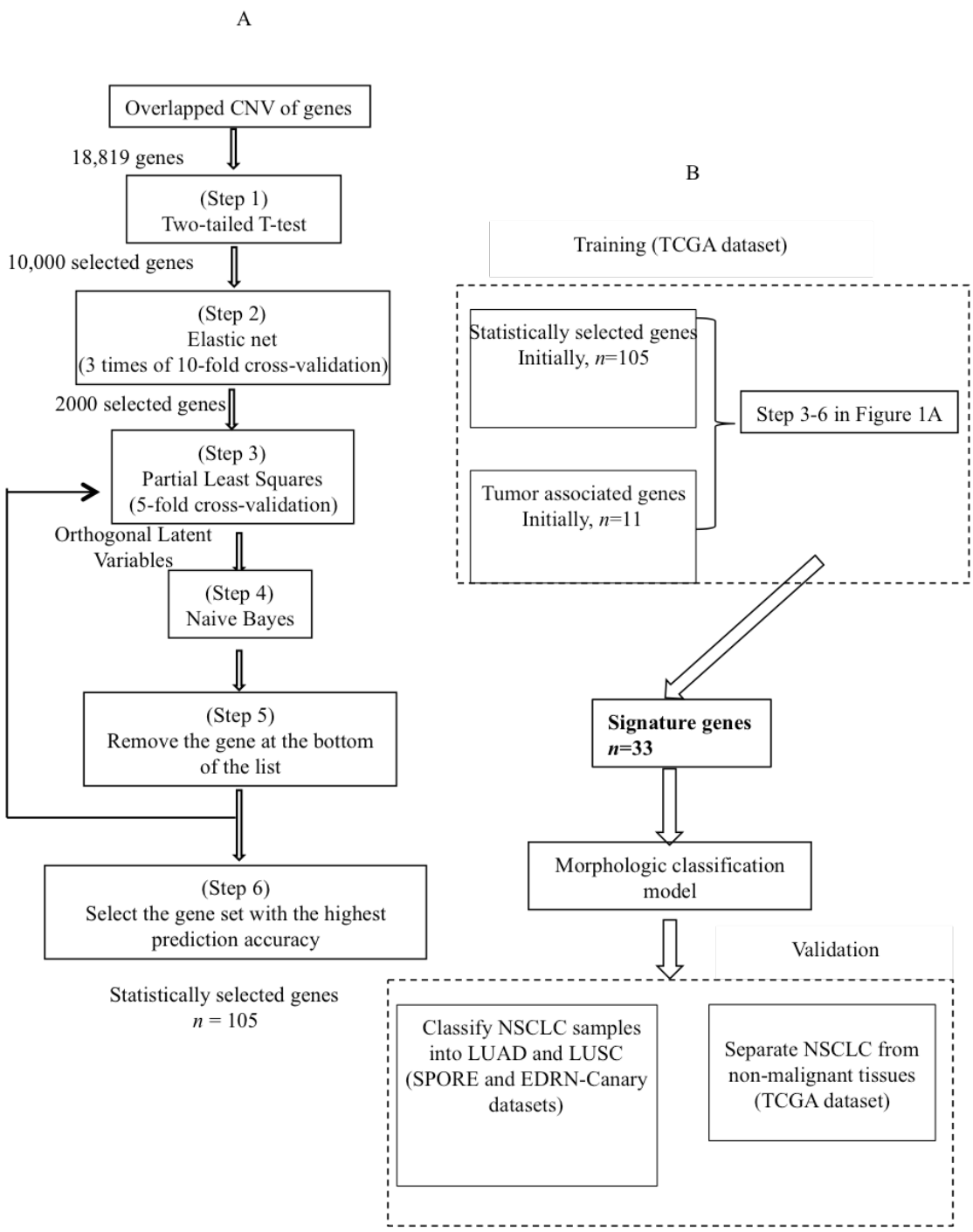

Figure S1 The identification and verification process for the integrated EN-PLS-NB algorithm and the flowchart of the CNV signature genes

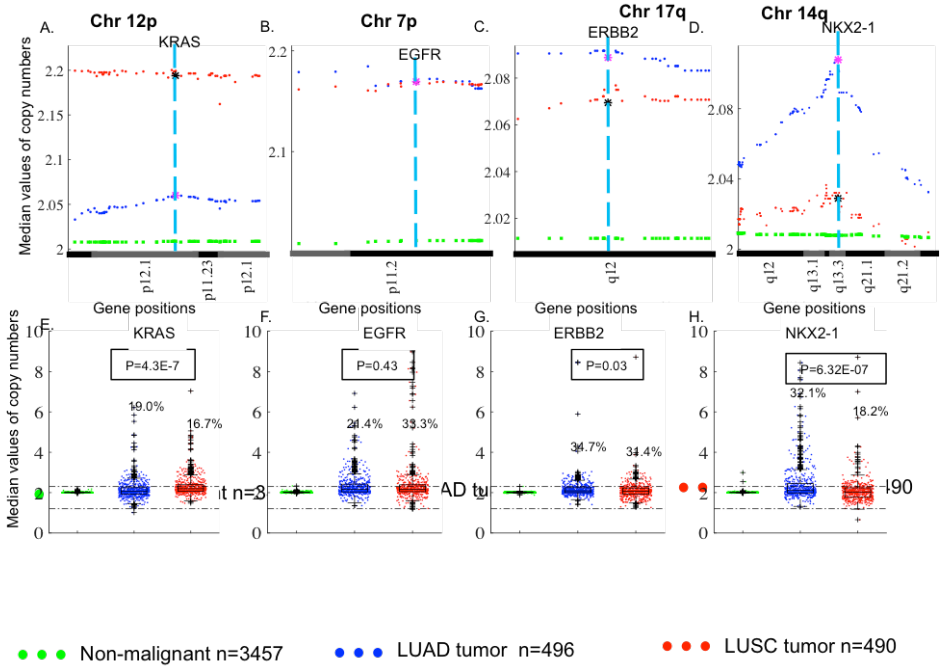

Figure S2 Chromosome location of important genes in EGFR pathyway and their Beeswarm copy number distribution in different groups.

Each spot is the median value of copy numbers of each gene in the corresponding group. The genes are sorted according to their locations. The space between two arms of each chromosome is the location of the corresponding centromere.

P: p values of t-test between LUSC and LUAD tumor samples.

The percentages of samples with copy numbers higher than 2.3 or lower than 1.2 of each gene are shown in the corresponding Beeswarm figures.

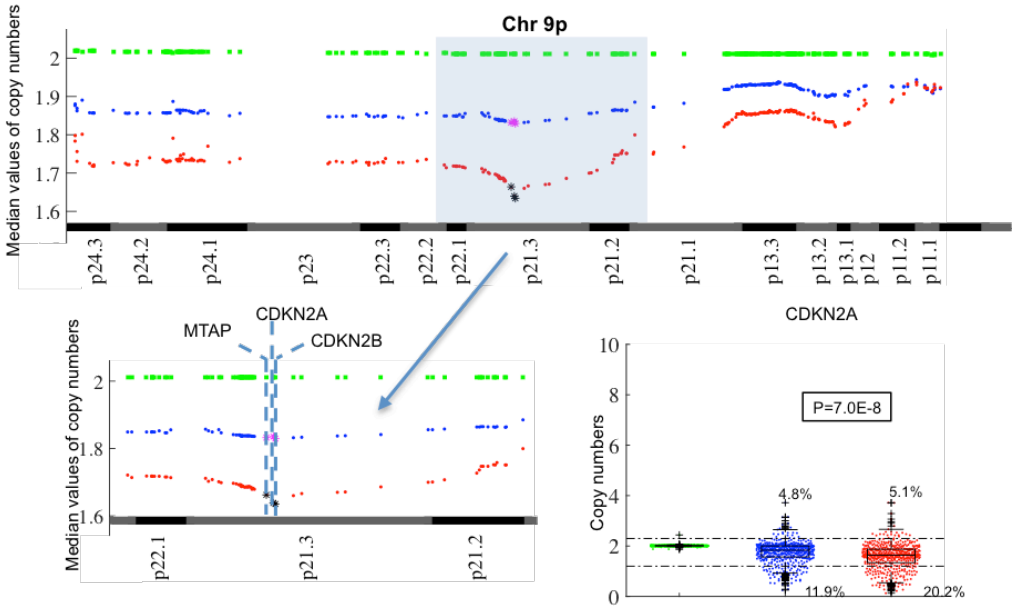

Figure S3 Narrow loss on Chromosome 9 and the Beeswarm copy number distribution of CDKN2A in different groups.

Each spot is the median value of copy numbers of each gene in the corresponding group. The genes are sorted according to their locations. The space between two arms of each chromosome is the location of the corresponding centromere.

P: p values of t-test between LUSC and LUAD tumor samples.

The percentages of samples with copy numbers higher than 2.3 or lower than 1.2 of each gene are shown in the corresponding Beeswarm figures.

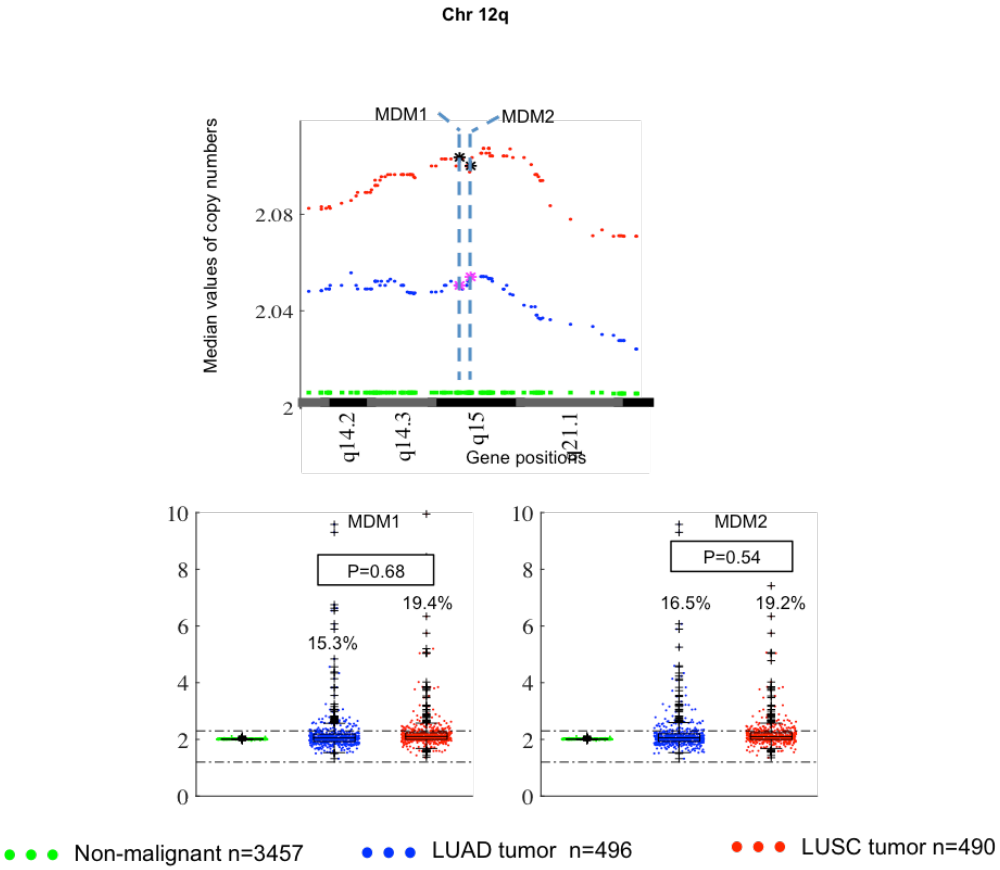

Figure S4 Gains 12q and the Beeswarm copy number distribution of specific genes in different groups.

Each spot is the median value of copy numbers of each gene in the corresponding group. The genes are sorted according to their locations. The space between two arms of each chromosome is the location of the corresponding centromere.

P: p values of t-test between LUSC and LUAD tumor samples.

The percentages of samples with copy numbers higher than 2.3 or lower than 1.2 of each gene are shown in the corresponding Beeswarm figures.

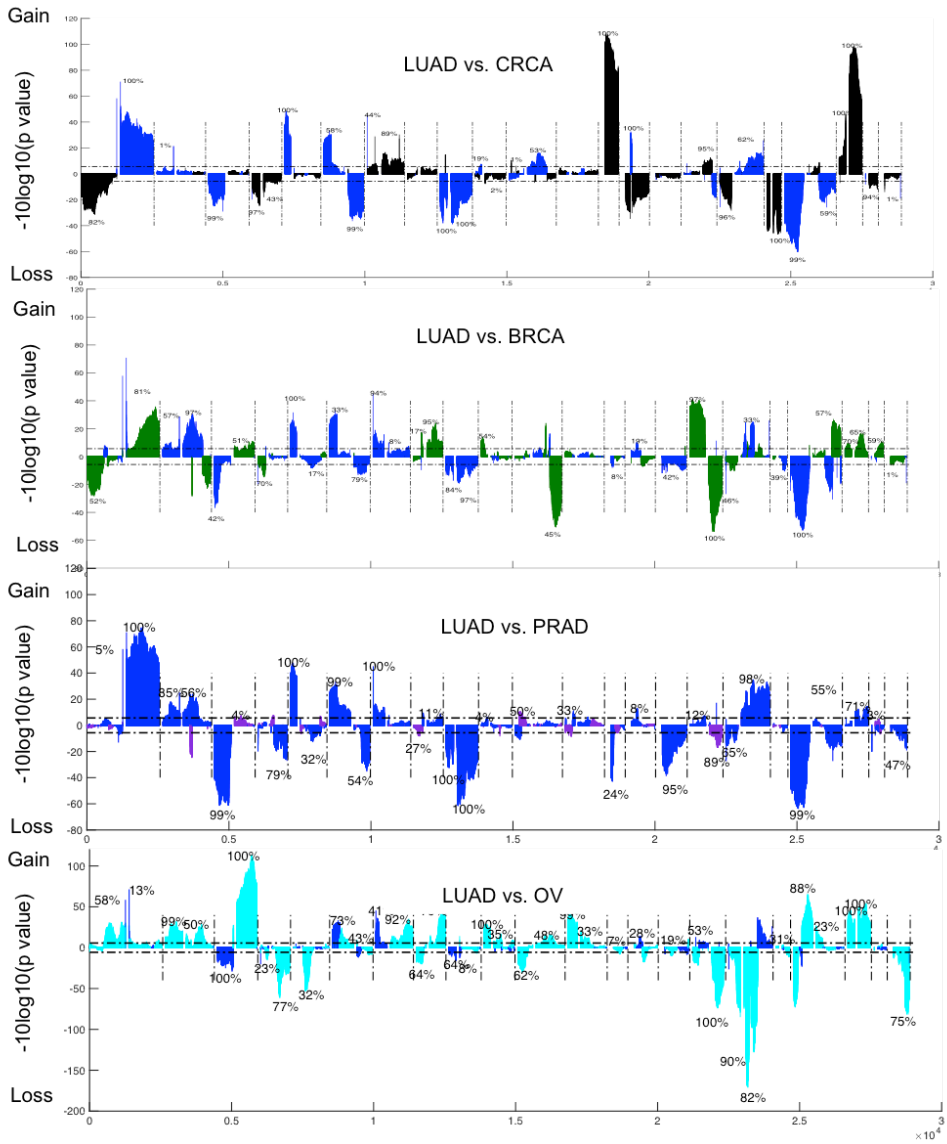

Figure S5 Genome-wide CNVs t-test between LUAD and CRCA/BRCA/PRAD/OV tumor samples in the TCGA dataset

Blue color indicates that the deflection was greater for LUAD, whereas black and green colors indicate that the deflection was greater for CRCA/BRCA/PRAD/OV, respectively. The dashed horizontal lines corresponding to the cutoff p-values of 2.1×10^{-6} (Bonferonni-correction). The vertical dashed lines separate the data from each chromosome. A gap within the individual chromosome data indicates the location of the centrosome. Note that for chromosomes 13, 14, 15, 21, and 22 only genes on the q arm were represented on the microarray.

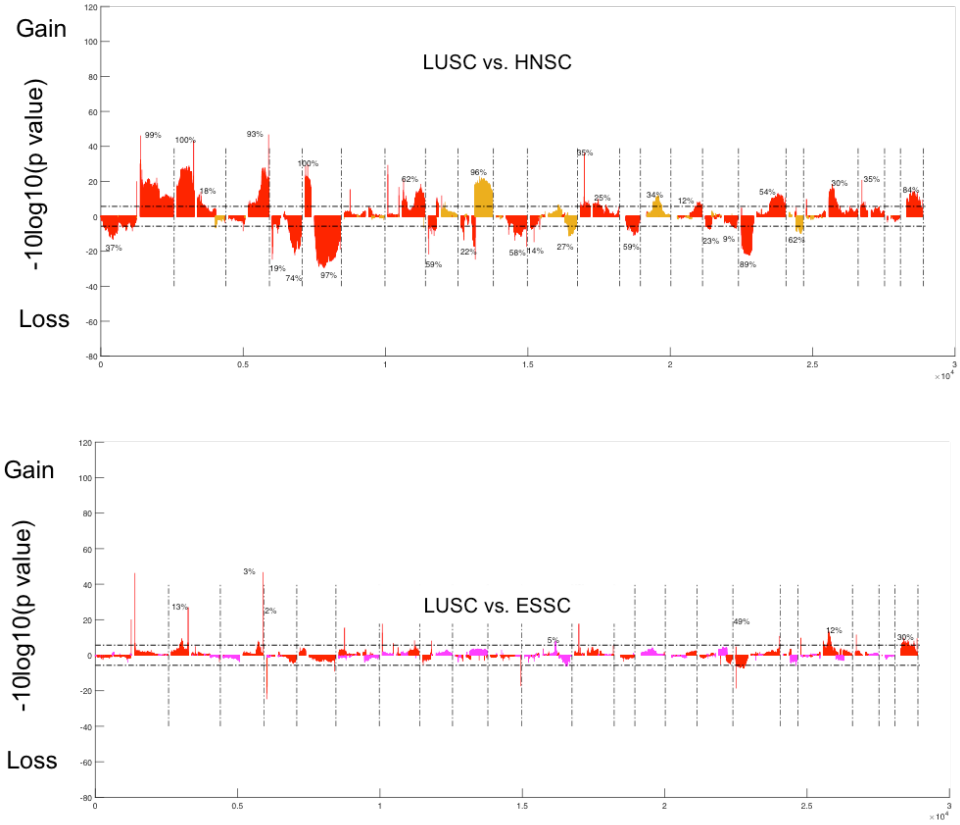

Figure S6 Genome-wide CNVs t-test between LUSC and HNSC/ESSC tumor samples in the TCGA dataset

Red color indicates that the deflection was greater for LUSC, whereas orange and pink colors indicate that the deflection was greater for HNSC or ESSC, respectively. The dashed horizontal lines corresponding to the cutoff p-values of 2.1×10^{-6} (Bonferonni-correction). The vertical dashed lines separate the data from each chromosome. A gap within the individual chromosome data indicates the location of the centrosome. Note that for chromosomes 13, 14, 15, 21, and 22 only genes on the q arm were represented on the microarray.

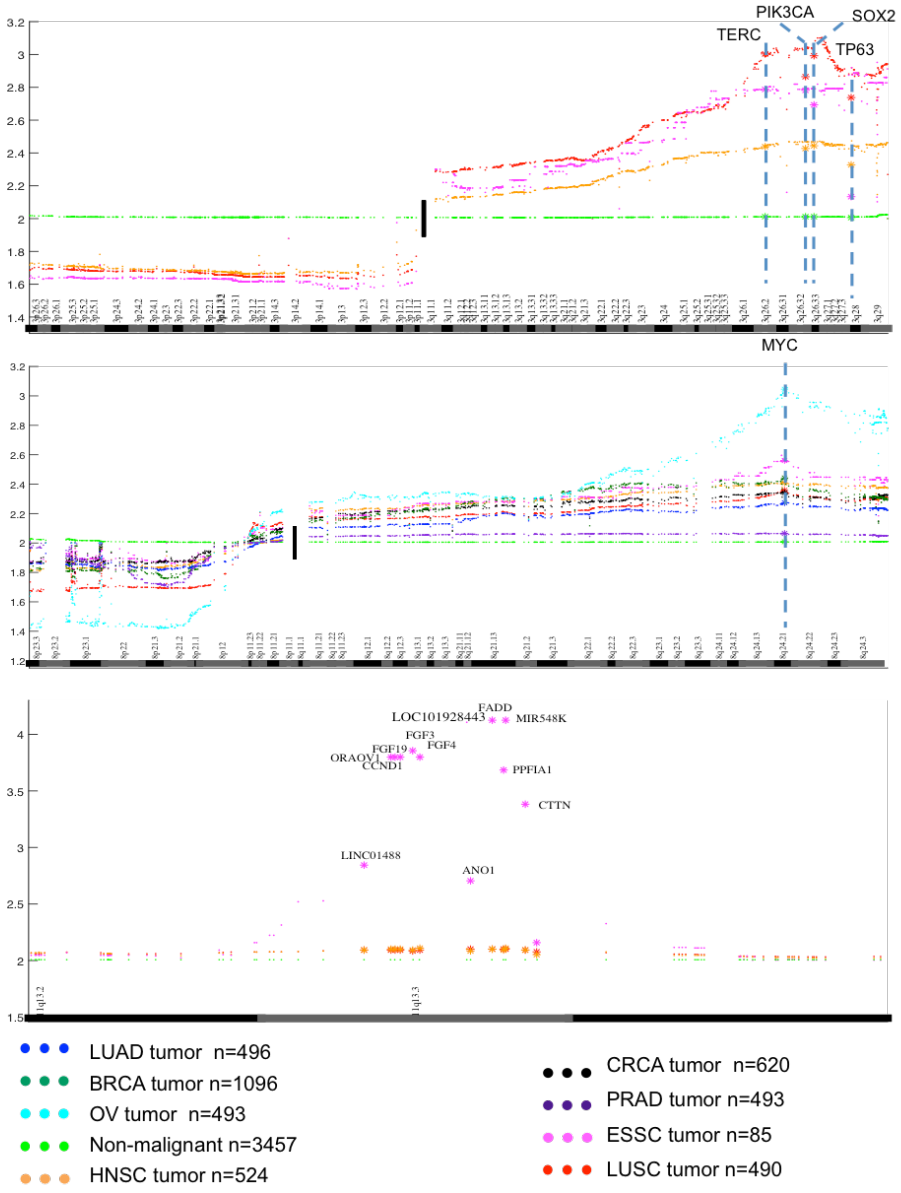

Figure S7 The genome locations of specific genes in different cancer types.

Each spot is the median value of copy numbers of each gene in the corresponding group. The genes are sorted according to their locations. The space between two arms of each chromosome is the location of the corresponding centromere.

Note: Chromsome3 9 and 11 have different axis limits.

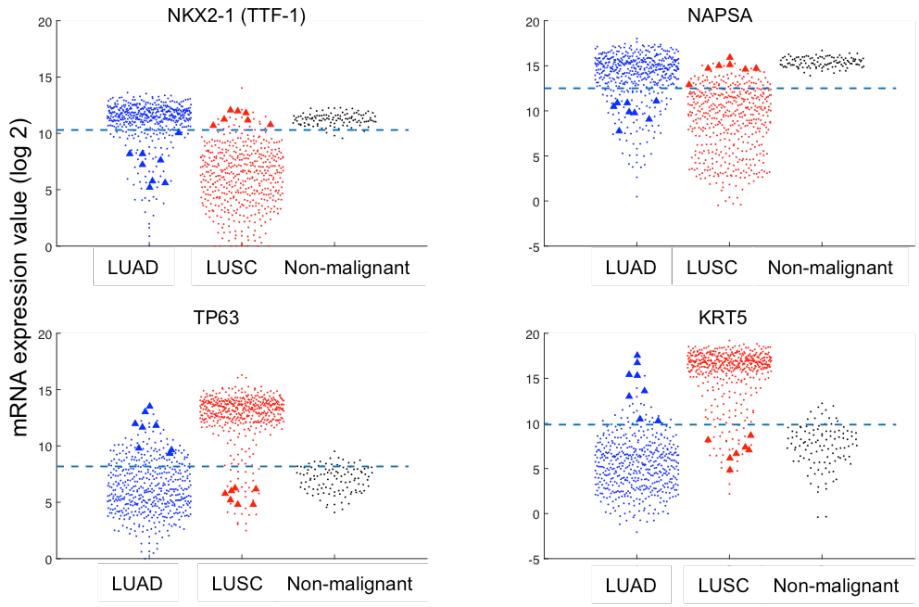

Figure S8 The bee-swarm of the mRNA expression values of four gene markers in LUAD and LUSC in TCGA samples

The lines in each figure are the cutoff values.

▲ Diagnosed as LUAD by TCGA but double negative by LUAD markers and double positive by LUSC markers;

▲ Diagnosed as LUSC by TCGA but double negative by LUSC markers and double positive by LUAD markers;

▪ Other LUAD samples in TCGA dataset;

▪ Other LUSC samples in TCGA dataset;

▪ Other non-maglinant samples in TCGA dataset;

LUAD n=490; LUSC n=487; Non-malignant n=110

Supplementary References

Boelens MC, van den Berg A, Fehrmann RS, Geerlings M, de Jong WK, te Meerman GJ, Sietsma H, Timens W, Postma DS, Groen HJ. 2009. Current smoking-specific gene expression signature in normal bronchial epithelium is enhanced in squamous cell lung cancer. J Pathol 218(2):182-191.

Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms; 2006. ACM. p 161-168.

Cui E, Deng A, Wang X, Wang B, Mao W, Feng X, Hua F. 2011. The role of adiponectin (ADIPOQ) gene polymorphisms in the susceptibility and prognosis of non-small cell lung cancer. Biochem Cell Biol 89(3):308-313.

Kang JU, Koo SH, Kwon KC, Park JW, Kim JM. 2009. Identification of novel candidate target genes, including EPHB3, MASP1 and SST at 3q26.2-q29 in squamous cell carcinoma of the lung. BMC Cancer 9:237.

Lv Q, Shen R, Wang J. 2015. RBPJ inhibition impairs the growth of lung cancer. Tumour Biol.

Nguyen DV, Rocke DM. 2002. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 18(1):39-50.

Rish I. An empirical study of the naive Bayes classifier; 2001. p 41-46.

Song K, Zhang Z, Tong TP, Wu F. 2012. Classifier assessment and feature selection for recognizing short coding sequences of human genes. J Comput Biol 19(3):251-260.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological):267-288.

Umekawa K, Kimura T, Kudoh S, Suzumura T, Nagata M, Mitsuoka S, Matsuura K, Oka T, Yoshimura N, Kira Y, Hirata K. 2013. Reaction of plasma adiponectin level in non-small cell lung cancer patients treated with EGFR-TKIs. Osaka City Med J 59(1):53-60.

Weaver DA, Crawford EL, Warner KA, Elkhairi F, Khuder SA, Willey JC. 2005. ABCC5, ERCC2, XPA and XRCC1 transcript abundance levels correlate with cisplatin chemoresistance in non-small cell lung cancer cell lines. Mol Cancer 4(1):18.

Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67(2):301-320.