Supplementary materials for

JDINAC: Joint density-based non-parametric differential interaction network analysis and classification using high-dimensional sparse omics data

Jiadong Ji[1], Di He[2], Yang Feng[3], Yong He[1], Fuzhong Xue[4], Lei Xie[2,5,*]

[1]School of Statistics, Shandong University of Finance and Economics, Jinan, 250014, China.

[2]The Graduate Center, The City University of New York, New York, NY 10016, USA.

[3]Department of Statistics, Columbia University, New York, NY 10027, USA.

[4]Departmentof Biostatistics, School of Public Health, Shandong University, Jinan, 250012, China.

[5]Department of Computer Science, Hunter College, The City University of New York, New York, NY 10065, USA.

*To whom correspondence should be addressed: lxie@iscb.org

Supplementary Methods

1. Simulation study in the non-linear scenario

It is quite difficult to quantify the non-linear relationship in real world scenario. We randomly selected 10 genes from the BRCA data, and described the pairwise scatterplot matrix (Figure S10). The diagonal panel of the matrix is the kernel density curve of individual gene. The *i*-th row and *j*-th column of the matrix is the pairwise scatter plot and smooth curve fitted by a generalized additive model. Overall, the non-linear relationship among these genes cannot be neglected.

To make our conclusions more **c**onvincing, we have simulated other non-linear relationship pattern. The two variables are independent in one group and have exponential relationship in the other group (see Figure S5). The data are generated as follows: In class 0, generate data $X^{(0)} = (X_1^{(0)}, \cdots, X_p^{(0)})$, where $X_j^{(0)} = \exp(u_j) + v_j$, $j = 1, \cdots, p/2$, $u_j \sim Unif(-2, 2)$ and $v_j \sim N(0, 2)$; $X_j^{(0)} = u_j$, $j = p/2+1, \cdots, p$, $p = 100$. In class 1, generate data $X^{(1)} = (X_1^{(1)}, \cdots, X_p^{(1)})$, where $X_j^{(1)} = \exp(u_{j+p/2}) + v_j$, $j = 1, \cdots, p/2$, $u_j \sim Unif(-2, 2)$ and $v_j \sim N(0, 2)$; $X_j^{(1)} = u_j$, $j = p/2+1, \cdots, p$, $p = 100$. $n_1 = n_2 = 300$. Figure S6 illustrates the ROC curves in this non-linear scenario. JDINAC still performs the best among the 5 methods over 50 simulations. JDINAC still performs the best among the 5 methods.

In the non-linear scenario (scenario 4), the area under the ROC curve (AUROC) is almost 1 due to the joint density of variables are quite different between two groups. We simulated new scenario with reduced difference between class 0 and class 1,through increasing the variance of variables. The data are generated as follows: In class 0, generate data $X^{(0)} = (X_1^{(0)}, \cdots, X_p^{(0)})$, where $X_j^{(0)} = u_j^2 + v_j$, $j = 1, \cdots, p/2$, $u_j \sim Unif(-2, 2)$ and $v_j \sim N(-4/3, e)$, $e = 1, 1.5, 2$; $X_j^{(0)} = u_j$, $j = p/2+1, \cdots, p$, $p = 100$. In class 1, generate data $X^{(1)} = (X_1^{(1)}, \cdots, X_p^{(1)})$, where $X_j^{(1)} = u_{j+p/2}^2 + v_j$, $j = 1, \cdots, p/2$, $u_j \sim Unif(-2, 2)$ and $v_j \sim N(-4/3, e)$, $e = 1, 1.5, 2$; $X_j^{(1)} = u_j$, $j = p/2+1, \cdots, p$, $p = 100$. $n_1 = n_2 = 300$. Figure S7 illustrates the ROC curves of 5

methods for the classification in the non-linear scenario with different variance over 50 simulations. As expected, Fig. S7(a) showing the highest AUROC followed by Fig. S7(b), (c). The proposed method JDINAC still has more advantageous performance than the other four.

2. Simulation study in the cases of multidimensional outliers

To evaluate the model performance in the cases of multidimensional outliers, we generated data with missing in one group. We generated 50 pairs of datasets, each representing the case (class 1) and the control (class 0) conditions. Each dataset contains $n_i (i = 1, 2)$ observations with p variables drawn from the multivariate normal distribution with mean 0 and covariance matrix $\mathbf{\Sigma}$, that is, $\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{\Sigma})$. $\mathbf{\Sigma}$ consistsof 3 blocks along the diagonal. $\mathbf{\Sigma} = diag(\Sigma_1, \Sigma_2, \Sigma_3)$, $\Sigma_1 = (\sigma_{ij})_{m \times m}$, $\sigma_{ij} = \rho^{|i-j|}$ for $i, j = 1, \cdots, m$; $m = 80$; $\Sigma_2 = \Sigma_3 = (\sigma^*_{ij})_{10 \times 10}$. In class 0, $p = 100$, $\rho = 0.5$, $\sigma^*_{ii} = 1$ for $i = 1, \cdots, 10$, $\sigma^*_{ij} = 0$ for $i \neq j$; Five percent variables in class 0 are randomly chosen to be missed, then for each missing variable five percent samples are randomly chosen to be missed. In class 1, $p = 100$, $\rho = 0.5$, $\sigma^*_{ii} = 1$ for $i = 1, \cdots, 10$, $\sigma^*_{ij} = 0.7$ for $i \neq j$. $n_1 = n_2 = 300$.

For all the methods mentioned in our manuscript, we first impute the missing value using the corresponding mean value of observed samples. Then, we ran all the methods to assess their performance. The results of classification are shown in Figure S8, it indicated that the proposed JDINAC has the best performance. Table S1 presents the TPR, TNR and TDR of the JDINAC, DiffCorr, DEDN and cPLR. It shows that JDINAC has the highest TDR and TPR, and acceptable TPR. It indicates that JDINAC performs well in the case with multidimensional outliers.

3. Simulation study in imbalanced case/control setting

We conducted simulation study in imbalanced case/control setting, to evaluate the optimal cut-off of predictions. We generated 50 pairs of datasets, each

representing the case (class 1) and the control (class 0) conditions. Each dataset contains $n_i (i = 1, 2)$ observations with p variables drawn from the multivariate normal distribution with mean 0 and covariance matrix $\mathbf{\Sigma}$, that is, $\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{\Sigma})$. $\mathbf{\Sigma}$ consists of 3 blocks along the diagonal. $\mathbf{\Sigma} = diag(\Sigma_1, \Sigma_2, \Sigma_3)$, $\Sigma_1 = (\sigma_{ij})_{m \times m}$, $\sigma_{ij} = \rho^{|i-j|}$ for $i, j = 1, \cdots, m$; $m = 80$; $\Sigma_2 = \Sigma_3 = (\sigma^*_{ij})_{10 \times 10}$. In class 0, $p = 100$, $\rho = 0.5$, $\sigma^*_{ii} = 1$ for $i = 1, \cdots, 10$, $\sigma^*_{ij} = 0$ for $i \neq j$; in class 1, $p = 100$, $\rho = 0.5$, $\sigma^*_{ii} = 1$ for $i = 1, \cdots, 10$, $\sigma^*_{ij} = 0.7$ for $i \neq j$.

The maximum value of Youden index (also called Youden's J statistic, J = Sensitivity + Specificity -1) was used as a criterion for selecting the optimum cut-off point. Table S2 shows the optimal cut-off point for all methods, as expected, 0.5 is not the optimal cut-off point. Furthermore, Figure S9 described the corresponding ROC curves. Different criteria can lead to different optimal cut off in real word scenario, Youden index puts equal weights to the sensitivity and specificity. In some special diagnostic tests, sensitivity is more important than specificity, more weight should be given to sensitivity.

4. BRCA dataset

The BRCA RNASeq Version 2 expression data and clinical data were obtained through the standard RNASeq data processing pipeline of software TCGA-Assembler. First, download the raw BRCA patients' RNA expression data through TCGA-Assmbler downloading module. Specifically, in our work, we used the RNASeqV2 normalized gene expression data, which uses MapSplice to do the alignment and RSEM to perform the quantization. After getting the raw data downloaded, with the TCGA-Assembler RNASeqV2 normalized gene expression data processing module, we extract the normalized count values as our training data and test data. For detailed data description, please refer to NIH NCI (National Cancer Institute) Wiki. Screening method was used to shrink the candidate gene pair numbers before applying JDINAC. In this study, we only used the data of 114 patients who have both tumor and matched normal samples. The patients' ID are listed as follows,

[1]	TCGA-BH-A0B7	TCGA-BH-A0BW	TCGA-BH-A0B2	TCGA-BH-A0H5	TCGA-BH-A0DL
[6]	TCGA-BH-A18N	TCGA-BH-A18P	TCGA-E2-A158	TCGA-BH-A0BS	TCGA-BH-A0BT
[11]	TCGA-BH-A18J	TCGA-BH-A18S	TCGA-BH-A18K	TCGA-BH-A18L	TCGA-BH-A18M
[16]	TCGA-BH-A18U	TCGA-BH-A18V	TCGA-E2-A153	TCGA-BH-A18Q	TCGA-BH-A18R
[21]	TCGA-E2-A15M	TCGA-BH-A0BZ	TCGA-BH-A0C3	TCGA-BH-A0DD	TCGA-BH-A0AZ
[26]	TCGA-BH-A1EO	TCGA-BH-A1ET	TCGA-BH-A1EU	TCGA-BH-A1EV	TCGA-BH-A1EW
[31]	TCGA-E2-A15I	TCGA-BH-A1EN	TCGA-BH-A1F2	TCGA-BH-A1F6	TCGA-BH-A1FM
[36]	TCGA-BH-A0B5	TCGA-BH-A0DG	TCGA-BH-A0DV	TCGA-E2-A1L7	TCGA-E2-A1LB
[41]	TCGA-E9-A1NA	TCGA-E9-A1ND	TCGA-BH-A1F0	TCGA-E2-A1IG	TCGA-BH-A1FU
[46]	TCGA-BH-A1FN	TCGA-E2-A1LH	TCGA-E9-A1N5	TCGA-E9-A1N6	TCGA-BH-A1F8
[51]	TCGA-BH-A1FC	TCGA-BH-A1FH	TCGA-BH-A0HA	TCGA-BH-A0E0	TCGA-BH-A0E1
[56]	TCGA-BH-A0H7	TCGA-BH-A0HK	TCGA-BH-A1FR	TCGA-A7-A13G	TCGA-E9-A1N9
[61]	TCGA-E9-A1NF	TCGA-A7-A0D9	TCGA-A7-A0DB	TCGA-A7-A0DC	TCGA-BH-A0AY
[66]	TCGA-BH-A0B3	TCGA-BH-A0B8	TCGA-BH-A0BC	TCGA-BH-A0BJ	TCGA-BH-A0BM
[71]	TCGA-BH-A0BV	TCGA-BH-A0C0	TCGA-BH-A0DH	TCGA-BH-A0DK	TCGA-BH-A0DP
[76]	TCGA-BH-A0DQ	TCGA-E9-A1RB	TCGA-E9-A1RC	TCGA-E9-A1RD	TCGA-E9-A1RF
[81]	TCGA-E9-A1N4	TCGA-E9-A1NG	TCGA-AC-A2FB	TCGA-AC-A2FF	TCGA-E9-A1R7
[86]	TCGA-AC-A23H	TCGA-BH-A204	TCGA-BH-A208	TCGA-BH-A203	TCGA-E9-A1RH
[91]	TCGA-BH-A1FD	TCGA-BH-A1FE	TCGA-BH-A1FG	TCGA-BH-A1FJ	TCGA-E9-A1RI
[96]	TCGA-BH-A209	TCGA-E2-A1LS	TCGA-GI-A2C8	TCGA-BH-A0BQ	TCGA-BH-A0DT
[101]	TCGA-BH-A0DO	TCGA-A7-A13E	TCGA-A7-A13F	TCGA-BH-A0AU	TCGA-BH-A1FB
[106]	TCGA-BH-A0BA	TCGA-GI-A2C9	TCGA-AC-A2FM	TCGA-BH-A0DZ	TCGA-E2-A15K
[111]	TCGA-A7-A0CE	TCGA-A7-A0CH	TCGA-E2-A1BC	TCGA-BH-A0H9	

Supplementary Figures

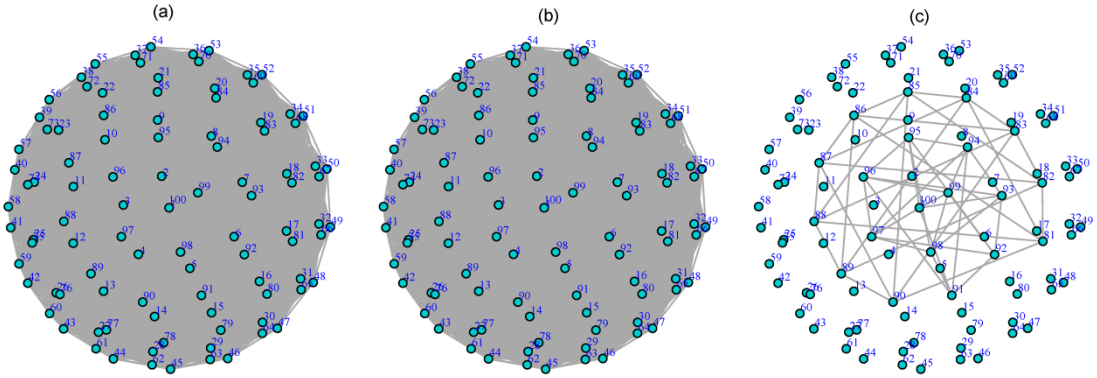

Fig. S1. The networks of simulation scenario 1. (a) The network of class 0; (b) The network of class 1; (c) The differential network. The edges in (a) and (b) indicate the two linked nodes are correlated. An edge in the differential network represent that the correlation between two given nodes are different across the two classes.

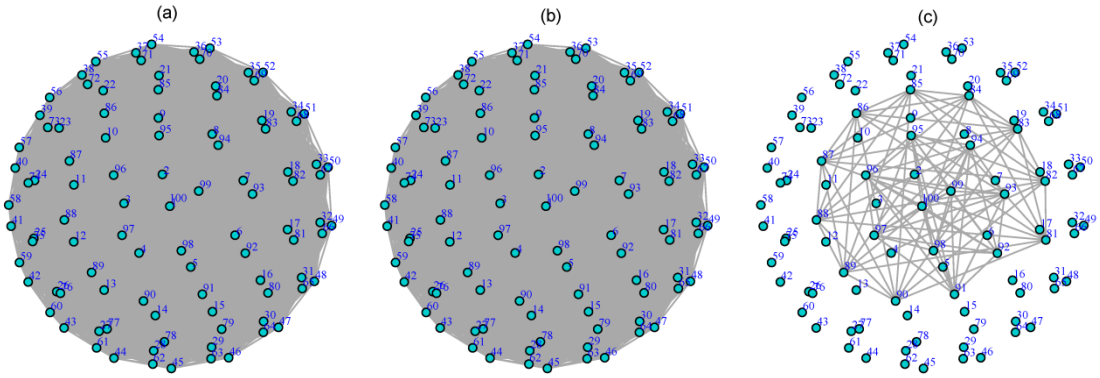

Fig. S2. The networks of simulation scenario 2. (a) The network of class 0; (b) The network of class 1; (c) The differential network. The edges in (a) and (b) indicate the two linked nodes are correlated. An edge in the differential network represent that the correlation between two given nodes are different across the two classes.

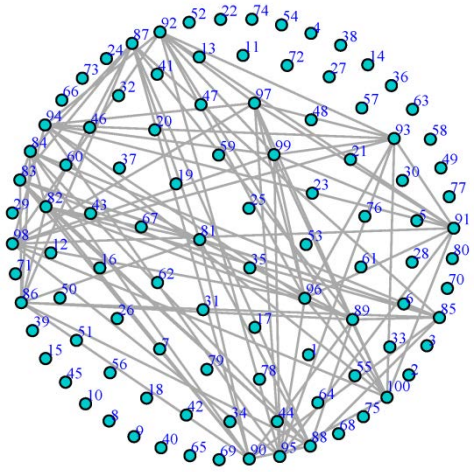

Fig. S3. The differential network of simulation scenario 3. An edge in the network represent that the joint density of two given nodes are different across the two classes.

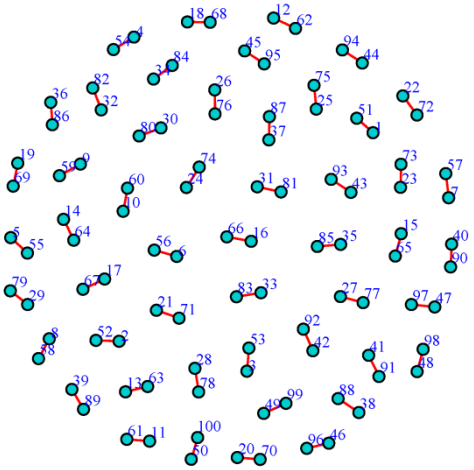

Fig. S4. The differential network of simulation scenario 4. An edge in the network represent that the dependency between two given nodes are different across the two classes.

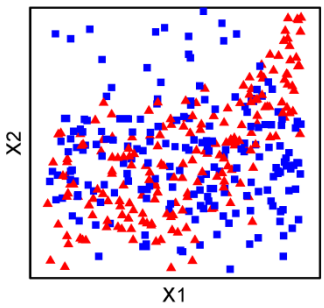

Fig. S5. The non-linear scenario. The blue square and red triangle represents the scatter plots for the two variables in class 0 and class 1 respectively, the two variables are independent in one group and have exponential relationship in the other group. (see Supplementary Methods 1)

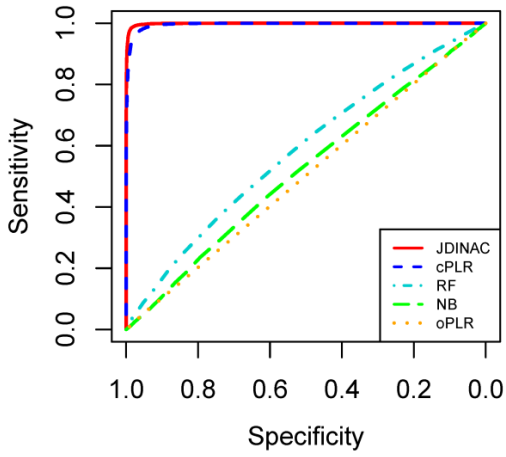

Fig. S6. The ROC curves in the non-linear scenario. (see Supplementary Methods 1)

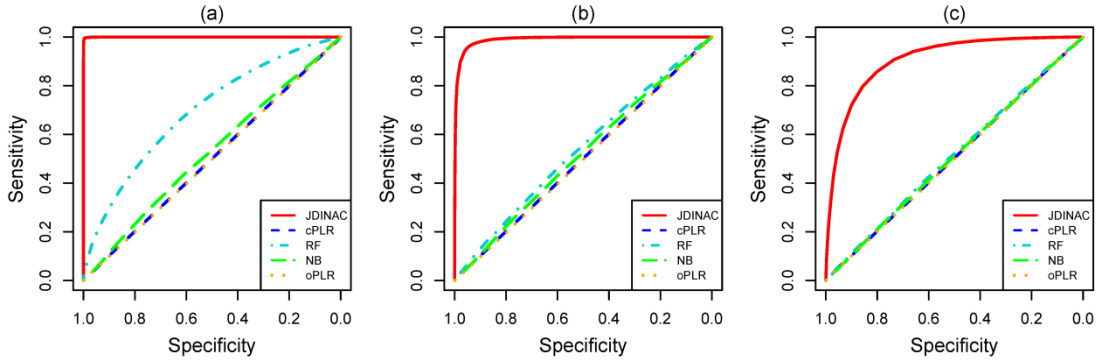

Fig. S7. ROC curves of 5 methods for the classification in the non-linear scenario with different variance. (a) $v_j \sim N(-4/3,1)$; (b) $v_j \sim N(-4/3,1.5)$; (c) $v_j \sim N(-4/3,2)$. (see Supplementary Methods 1)

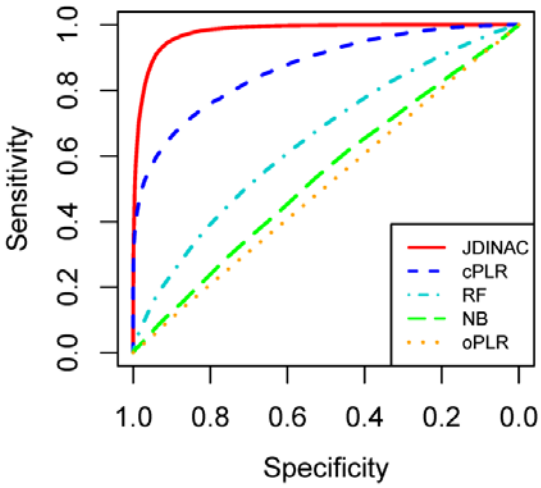

Fig. S8. ROC curves of 5 methods for the classification with multidimensional outliers. (see Supplementary Methods 2)

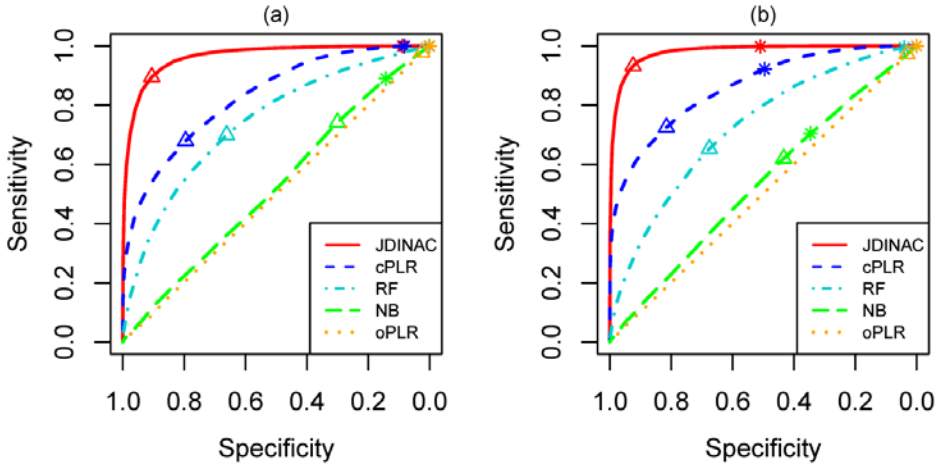

Fig. S9. ROC curves of 5 methods for the classification. (a) sample size n_0=100, n_1=500; (b) sample size n_0=200, n_1=400. The asterisk indicates the location where the cutoff of prediction was set to 0.5. The triangle indicates the optimal cutoff point obtained by maximum Youden index. (see Supplementary Methods 3)

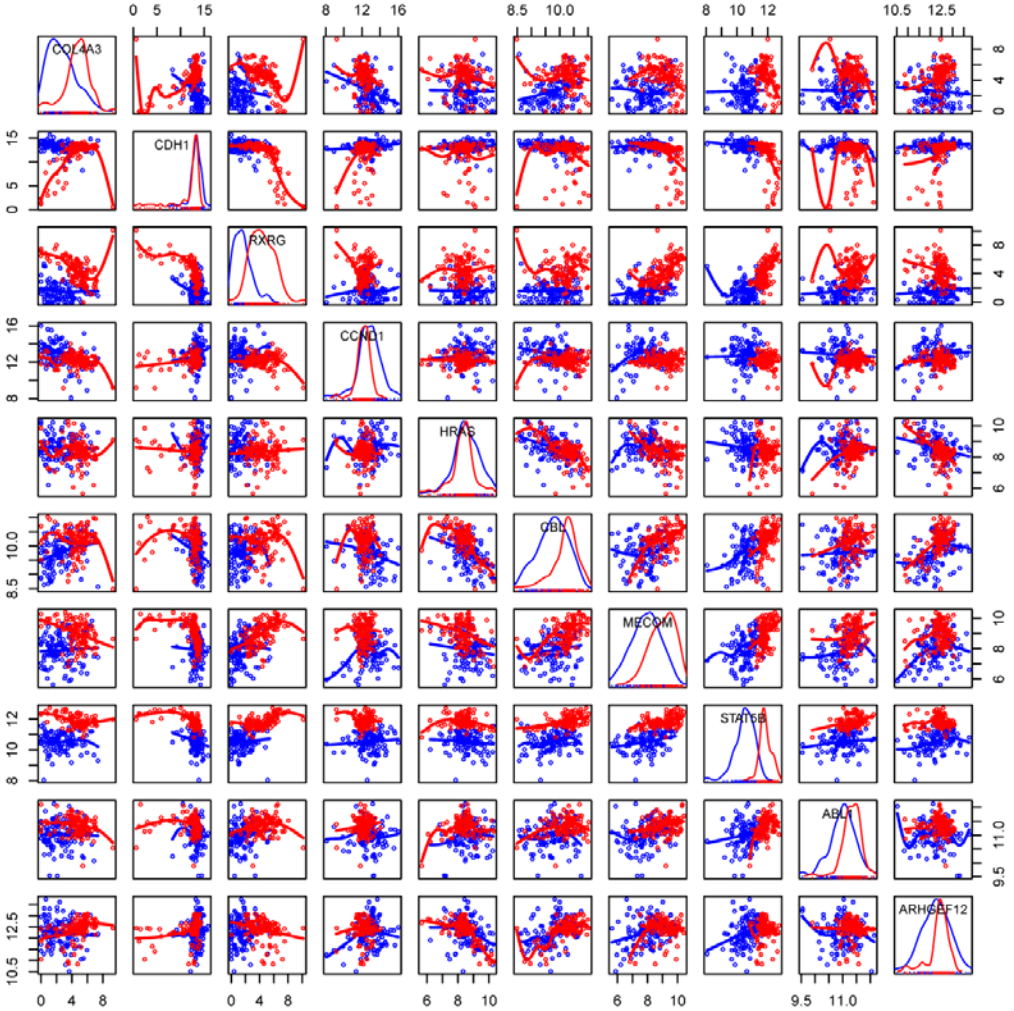

Fig. S10. Scatterplot matrix of 10 genes. The diagonal panel of the matrix is the kernel density curve of individual gene. The *i*-th row and *j*-th column of the matrix is the pairwise scatter plot and smooth curve fitted by a generalized additive model. The blue and red represents tumor group and control group respectively. (see Supplementary Methods 1)

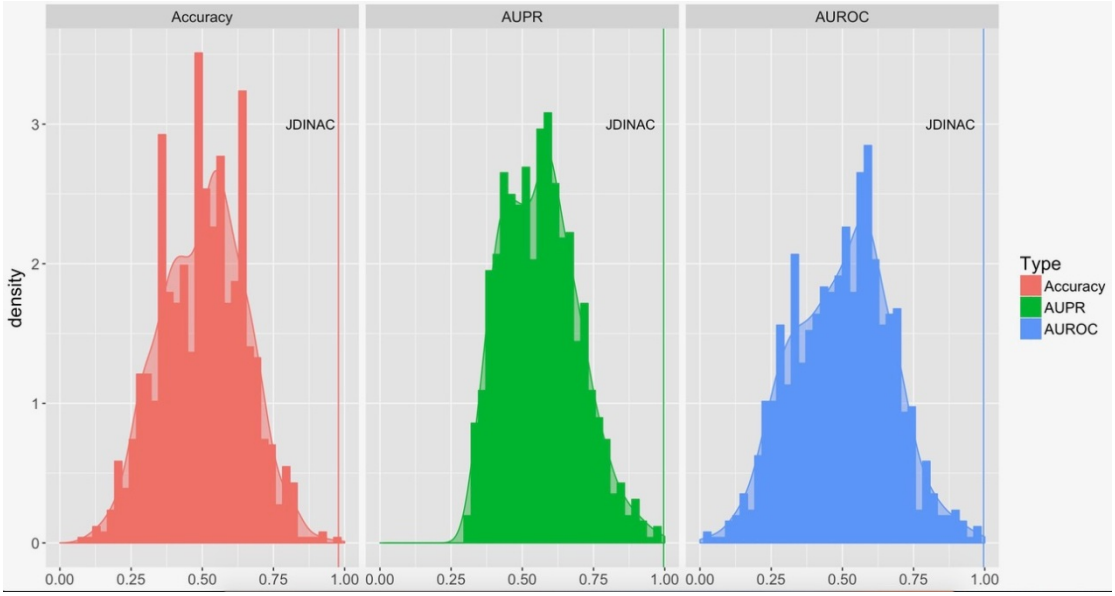

Fig. S11. Y-randomization result of JDINAC classification model of BRCA case and control. The area under the ROC curve (AUROC), the area under the precision-recall curve (AUPR).

Supplementary Tables

Table S1. The TPR, TNR and TDR of different methods. (average of 50 replications, %), The best performance is highlighted in bold. (see Supplementary Methods 2)

Methods	TDR	TPR	TNR
JDINAC[a]	**99.20**	73.38	**99.99**
DiffCorr[b]	89.74	**100.00**	99.79
DEDN	13.84	89.71	93.16
cPLR	40.87	47.82	98.60

a. Pair (G_i, G_j) was taken as differential edge in the network for JDINAC, when the differential dependency weight $w_{ij} \geq 4$.

b. Set to control the false discovery rate equal to 0.1.

Table S2. The optimal cutoff point obtained by maximum Youden index. (see Supplementary Methods 3)

	JDINAC	cPLR	RF	NB	oPLR
n_0=100, n_1=500	0.83	0.83	0.80	0.52	0.79
n_0=200, n_1=400	0.67	0.65	0.65	0.51	0.63

Table S3. The overlapped edges betweendifferent methods.

Methods	Overlap edges
JDINAC & cPLR	*LEF1--FOXO1, TPM3--CKS1B, RXRG--LAMA3, COL4A6--AGTR1, LAMC1--FGF1, TGFBR2--LEF1*
JDINAC & DiffCorr	*EGFR--AR, LAMC2--CBLC, GSTP1--EGFR, WNT7B--LAMC2, ARNT2--AGTR1, TGFB3--RUNX1T1, FZD7--CTBP2, TGFB3--FOXO1, COL4A6--CBLC, LAMC2--CDK2, TGFA--COL4A6, PDGFA--FOXO1, TGFBR2--GNG7, GNB5--FGFR2, GNG7--FZD4, PDGFA--FZD4, PDGFA--GNG2, JUP--GNG7, TGFBR2--PDGFA*
DiffCorr & cPLR	*KIT--CDH1, TCF7--CDH1, RUNX1--CDK2, FZD4--ERBB2, PDGFA--ETS1, LAMA4--FGF1, JUP--FZD7, LAMA4--FZD7, LAMC1--FZD7, LPAR2--FZD7, TP53--FZD7, PDGFA--JUP, LPAR3--LEF1, VEGFB--LEF1*

Table S4. The overlapped GO terms and KEGG pathway among different methods.

ID	Functional term	Category	JDINAC	DiffCorr	cPLR
			***p*-value (count/size)**	***p*-value (count/size)**	***p*-value (count/size)**
GO:0031581	hemidesmosome assembly	BP	0.006212655 (4/74)	0.04962503 (4/128)	0.04805773 (3/70)
GO:0005198	structural molecule activity	MF	0.01694874 (9/74)	0.004009204 (14/128)	0.0026791628 (10/70)
GO:0005201	extracellular matrix structural constituent	MF	0.03082378 (5/74)	0.003969823(8/128)	0.0004478932 (7/70)
GO:0005604	basement membrane	CC	0.002956678 (10/74)	0.02389143 (12/128)	0.001859320 (10/70)
GO:0005605	basal lamina	CC	0.005858540 (7/74)	0.03905771 (8/128)	0.004174482 (7/70)
GO:0044420	extracellular matrix component	CC	0.002956678 (10/74)	0.02389143 (12/128)	0.001859320 (10/70)
GO:0044421	extracellular region part	CC	0.005011179 (37/74)	0.03430913 (55/128)	0.013529227 (34/70)
hsa04512	ECM-receptor interaction	pathway	0.01352785 (10/74)	0.006079814 (15/128)	0.0005460853 (12/70)

Table S5. Hub genes identified by JDINAC

Methods	Hub genes (number of neighbor genes)
JDINAC	*FGF1* (8), *TGFB3* (8), *ARNT2* (7), *LAMA3* (6), *LAMC2*(6),*PDGFA* (5), *EGFR* (4), *FOXO1* (4), *FZD7* (4), *LEF1* (4), *WNT6* (4)
DiffCorr	*CBLC* (27), *TRAF4* (27), *FZD3* (26), *TGFA* (25), *FZD7* (24), *HDAC1* (24), *ERBB2* (23), *MAP2K1* (23), *PDGFA* (23), *CTBP2* (22), *FZD4* (21), *LPAR2* (21), *JUP* (20), *RUNX1* (20), *CDK2* (19), *FZD1* (19), *VEGFB* (19), *KIT* (18), *LAMA4* (18), *CDH1* (17), *COL4A5* (17), *COL4A6* (17), *ITGA2* (16), *CDKN2B* (15), *GNAI1* (15), *GNG7* (15), *TCF7* (15), *GNB5* (14), *ITGA3* (14), *TP53* (14), *WNT10A* (14), *LPAR3* (13), *IGF1R* (12), *WNT7B* (12), *GNG11* (11), *GNG2* (11), *JAK1* (11), *LAMA3* (11)
cPLR	*LEF1* (6), *FZD7* (5), *KIT* (5), *JUP* (4), *LAMA4* (4), *WNT10A* (4)