Genetics and Molecular Research

Correlation-based linear discriminant classification for gene expression data

M. Pan[1,2] and J. Zhang[3]

[1]Department of Optoelectronic Engineering, Jinan University, Guangzhou, China
[2]Guangdong Provincial Key Laboratory of Optical Fiber Sensing and Communications, Jinan University, Guangzhou, China
[3]Department of Physics, Jinan University, Guangzhou, China

Corresponding author: J. Zhang
E-mail: tjiez@jnu.edu.cn

Genet. Mol. Res. 16 (1): gmr16019357
Received September 21, 2016
Accepted September 21, 2016
Published January 23, 2017
DOI http://dx.doi.org/10.4238/gmr16019357

ABSTRACT. Microarray gene expression technology provides a systematic approach to patient classification. However, microarray data pose a great computational challenge owing to their large dimensionality, small sample sizes, and potential correlations among genes. A recent study has shown that gene-gene correlations have a positive effect on the accuracy of classification models, in contrast to some previous results. In this study, a recently developed correlation-based classifier, the ensemble of random subspace (RS) Fisher linear discriminants (FLDs), was utilized. The impact of gene-gene correlations on the performance of this classifier and other classifiers was studied using simulated datasets and real datasets. A cross-validation framework was used to evaluate the performance of each classifier using the simulated datasets or real datasets, and misclassification rates (MRs) were computed. Using the simulated data, the average MRs of the correlation-based

classifiers decreased as the correlations increased when there were more correlated genes. Using real data, the correlation-based classifiers outperformed the non-correlation-based classifiers, especially when the gene-gene correlations were high. The ensemble RS-FLD classifier is a potential state-of-the-art computational method. The correlation-based ensemble RS-FLD classifier was effective and benefited from gene-gene correlations, particularly when the correlations were high.

Key words: Classifier; Gene expression; Correlation-based; Linear discriminant analysis; Random subspace

INTRODUCTION

Microarray gene expression technology provides a systematic approach to cancer classification (Golub et al., 1999; Ross et al., 2000) and continues to be used to clarify the mechanisms of human diseases (Novianti et al., 2014). However, microarray data pose a great challenge for computational techniques owing to their large dimensionality and small sample sizes (Saeys et al., 2007). Correlations among genes present additional challenges and are often not incorporated into analyses (Xu et al., 2015).

The classical Fisher linear discriminant analysis (FLDA) deals with multivariate (multi-gene) correlations when the sample size exceeds the dimensions. When the dimensions are higher than the sample size, the sample covariance matrix is not invertible, and the FLDA is not applicable (McLachlan, 2004). Many methods have been proposed to overcome this problem (see Slawski et al., 2008 and Kalina, 2014 for a summary). Some of these methods do not consider multivariate correlations (i.e., they are not correlation-based), e.g., the diagonal linear discriminant analysis (DLDA), which assumes that the within-class covariance matrices are diagonal (Slawski et al., 2008). Other methods account for correlations (i.e., they are correlation-based), e.g., the ensemble of random subspace (RS) Fisher linear discriminant (FLD) classifier, which averages the decisions of the base FLD classifiers in lower-dimensional subspaces and gene-gene correlations might be utilized (Durrant and Kabán, 2015).

The performance of classifiers varies among datasets. Gene-gene correlations within datasets affect the performance of classifiers (Jong et al., 2014). A previous study (Novianti et al., 2015) has shown that the within-class correlation has a positive effect on the accuracy of classification models. However, an earlier study (Kim and Simon, 2011) showed that the correlation structure has a negative impact on the performance of classifiers, such as the PAM (prediction analysis of microarrays) probabilistic classifier, the Bayesian composite probabilistic classifier (both of which are non-correlation-based), and the penalized logistic regression (PLR) classifier (correlation-based), but the PLR is effective for strong correlations. Another study (Dudoit et al., 2002) showed that the DLDA classifier (non-correlation-based) performed well compared with FLDA (correlation-based).

In this study, the recently developed ensemble RS-FLD classifier was utilized as model correlation-based linear discriminant classifier and was compared with DLDA, PAM, and other classifiers. The impact of gene-gene correlations on classifier performance was studied using simulated datasets and real datasets.

MATERIAL AND METHODS

Classifiers

The ensemble RS-FLD classifier (referred as enRS-FLD) was chosen to account for gene-gene correlations. It uses all variables and correlations among variables in subspaces composed of randomly selected variables. The analysis was conducted in Matlab according to the algorithm described in Durrant and Kabán (2015). The same parameters as in Durrant and Kabán (2015) were also used. The dimension of the subspaces, k, was set to (N - 2) / 2 to optimize the classification, where N was the number of samples in the training dataset. The number of subspaces, M, was set to 1000.

The DLDA classifiers using all genes (referred to as allDLDA) and filtered genes (referred to as filterDLDA) were chosen for comparison. The filtered genes were obtained by deleting unrelated genes (false discovery rate multiple testing adjustment P value > 0.05) and redundant genes (|correlation coefficient| > 0.8) (Valavanis et al., 2015). The classification was conducted in Matlab according to the algorithms described in Dudoit et al. (2002) and Valavanis et al. (2015).

PAM is an enhanced nearest prototype (centroid) classifier that uses "nearest shrunken centroids" to identify subsets of genes that best characterize each class (Tibshirani et al., 2002). Soft thresholding was used to perform variable selection (Slawski et al., 2008), which was implemented using the R package pamr (Hastie et al., 2014). The classifier was trained using training samples and the threshold parameter was tuned at the training stage. The classifier was then applied to the test samples to obtain class predictions.

SVM (support vector machine) classification is a correlation-based binary classification that works with selected variables (features) and fits an optimal hyperplane between two classes by maximizing the margin between the closest points (Statnikov et al., 2005; Novianti et al., 2015). This classification was implemented in Matlab using the functions svmtrain and svmclassify.

Availability of data and materials

The datasets supporting the conclusions of this article are available in (Dataset *Colon*, 2016; Dataset *ALL-AML,* 2016; Dataset *Prostate*, 2016), and the Gene Expression Omnibus website (Gene Expression Omnibus, 2016) under the accession numbers GSE42133, GSE57162, GSE49710, GSE4922, and GSE19159. The R package pamr is available in (Hastie et al., 2014).

Classifier performance

Cross-validation (CV) (Durrant and Kabán, 2015; Tan et al., 2015) was used to evaluate the performance of each classifier using the simulated datasets and real datasets. In the process of CV, the dataset was split into testing set, of which the classes were assumed unknown, and training set, by which the classifier was trained. Some classifiers might also carry out inner-set cross-validation in training to tune the parameters (e.g., pamr). Being trained by the training set, the classifier could predict the classes for the testing set. Misclassification rates (MRs) were computed to evaluate the performance of the classifier. For simulated datasets, the number of samples was small and leave-one-out CV was used. For real datasets, CV was performed according to Durrant and Kabán (2015), in which we ran experiments on 100

independent splits, and in each split we took 12 samples for testing and used the remainder for training. MRs were averaged over the 100 splits.

Simulated datasets

The Monte Carlo setup was described previously (Kim and Simon, 2011). Simulated datasets were generated using multivariate normal distributions with class-specific mean vectors and common covariance matrices. Each dataset contained 30 samples that were randomly assigned to class 0 or 1. The simulations were performed with $p = 1000$ genes (variables), and the first $p_1 = 50$ were differentially expressed between classes with a mean difference of 1. The remaining 950 genes had the same mean for each class. For the common covariance matrix, Structures 1-3 as defined in Kim and Simon (2011) were used. In Structure 1, few genes were correlated; in Structure 3, many genes were correlated. The correlation coefficients for the three structures were 0.25, 0.5, and 0.75. The case of full independence was also studied. In total, 1000 simulated datasets were generated for each structure.

Real datasets

Three gene expression datasets, Colon (Alon et al., 1999; Dataset *Colon*, 2016), ALL-AML (Golub et al., 1999; Dataset *ALL-AML*, 2016), and Prostate (Singh et al., 2002; Dataset *Prostate*, 2016), which have been extensively studied (Dembélé and Kastner, 2014; Durrant and Kabán, 2015), were included in the study. Five Gene Expression Omnibus datasets used in recently published studies, GSE42133 (Pramparo et al., 2015), GSE57162 (Gowrishankar et al., 2015), GSE49710 (Zhang et al., 2015), GSE4922 (Ivshina et al., 2006; Blagus and Lusa, 2015), and GSE19159 (Gravier et al., 2010; Zhang and Pan, 2016), were also included. The datesets are summarized in Table 1. Ethics approval was obtained in previous studies and accordingly was not required for this study.

Table 1. Gene expression datasets.

Source	Dataset	Samples analyzed	Nsample	Ngene
Alon et al. (1999)	Colon	Tumor/Normal	62 (40/22)	2,000
Golub et al. (1999)	ALL-AML	ALL/AML	72 (47/25)	7,129
Singh et al. (2002)	Prostate	Tumor/Normal	102 (52/50)	12,533
Pramparo et al. (2015)	GSE42133 (Autism)	ASD/Control	147 (91/56)	47,231
Gowrishankar et al. (2015)	GSE57162 (Renal cortical neoplasm)	Benign/Malignant	191 (36/155)	43,102
Zhang et al. (2015)	GSE49710 (Neuroblastoma)	HR, Event yes/no	176 (120/56)	43,291
		Class unfavorable/favorable	272 (91/181)	43,291
Ivshina et al. (2006)	GSE4922 (Breast cancer)	ER +/-	245 (211/34)	22,215
		Grade 1&2/3	289 (234/55)	22,215
Gravier et al. (2010)	GSE19159 (Breast carcinoma)	Outcome, Good/Poor	168 (111/57)	2,905

Nsample: number of samples; Ngene: number of genes; ALL: acute lymphoblastic leukemia; AML: acute myeloid leukemia; ASD: autism spectrum disorder; HR: high risk; ER: estrogen receptor status.

RESULTS

Classifier performance using simulated datasets

The MR for each simulated dataset was computed using leave-one-out CV and

averaged over 1000 simulation replications to reduce the standard deviation. The results are shown in Figure 1.

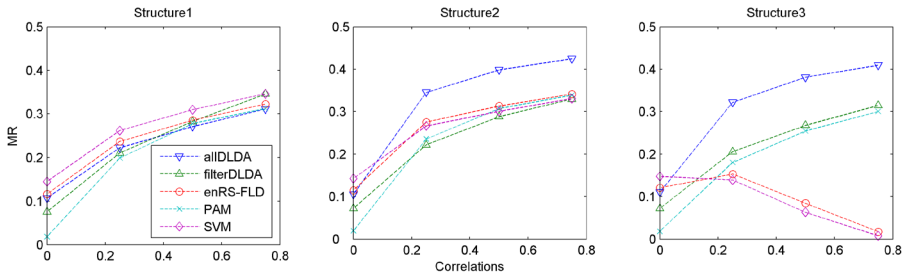

Figure 1. Average MR (1000 replications) for simulated datasets.

Overall, the average MRs for the classifiers allDLDA, filterDLDA, and PAM (all of which are non-correlation-based) increased as the correlation increased. The average MRs of the classifiers enRS-FLD and SVM (both of which are correlation-based) increased as the correlation increased for Structures 1 and 2 and decreased as the correlation increased for Structure 3.

The average MR for filterDLDA was lower than that of allDLDA for Structures 2 and 3. For Structure 1, the average MR of filterDLDA was somewhat lower than that of allDLDA for low correlations and similar to allDLDA for high correlations.

The average MR for PAM was somewhat lower than that of filterDLDA for Structure 3. For Structures 1 and 2, it was somewhat lower than that of filterDLDA for low correlations and somewhat higher than that of filterDLDA for high correlations.

The average MR for enRS-FLD was similar to that of SVM for Structures 2 and 3, and somewhat lower than that of SVM for Structure 1. The average MRs for enRS-FLD and SVM were largely lower than those of the three other classifiers at high correlations for Structure 3. They were similar to or lower than that of allDLDA and similar to or somewhat higher than those of filterDLDA and PAM for Structure 2. They were similar to or somewhat higher than those of the three non-correlation-based classifiers for Structure 1.

Classifier performance with real datasets

The MR for each real dataset was computed using CV of splitting and averaged over 100 splits. The results are shown in Table 2.

Overall, the MRs for filterDLDA were lower than those of allDLDA except for dataset 'GSE49710, HR-Event', the MRs of enRS-FLD were lower than those of filterDLDA, and the MRs of PAM and SVM were similar to those of enRS-FLD. However, the MRs of enRS-FLD and SVM were significantly lower than that of PAM for dataset GSE42133.

Impact of gene-gene correlations on classifier performance

For each real dataset, within-class correlations (averaged over the two classes) were computed for all gene pairs. Distributions of within-class correlations are shown in, **Figures S1-S10**. The average of absolute within-class correlations (denoted *withincor*) was calculated for each dataset and is shown in Table 2 and **Table S1**.

Table 2. Misclassification rate for the classifiers.

		Classifiers				
Dataset	*withincor*	allDLDA	filterDLDA	enRS-FLD	PAM	SVM
Colon	0.474	0.289 ± 0.015*	0.202 ± 0.011	0.180 ± 0.009	0.164 ± 0.011	0.184 ± 0.010
ALL-AML	0.166	0.143 ± 0.011	0.063 ± 0.009	0.034 ± 0.006	0.023 ± 0.004	0.029 ± 0.005
Prostate	0.331	0.404 ± 0.012	0.278 ± 0.014	0.100 ± 0.008	0.103 ± 0.009	0.091 ± 0.008
GSE42133	0.095	0.283 ± 0.012	0.278 ± 0.012	0.239 ± 0.012	0.313 ± 0.012	0.228 ± 0.011
GSE57162	0.182	0.053 ± 0.006	0.044 ± 0.006	0.015 ± 0.003	0.033 ± 0.005	0.015 ± 0.003
GSE49710, HR-Event	0.353	0.312 ± 0.013	0.366 ± 0.013	0.302 ± 0.010	0.333 ± 0.013	0.308 ± 0.011
GSE49710, Class	0.305	0.137 ± 0.009	0.099 ± 0.008	0.038 ± 0.006	0.054 ± 0.007	0.038 ± 0.006
GSE4922, ER	0.110	0.121 ± 0.009	0.117 ± 0.009	0.102 ± 0.008	0.117 ± 0.009	0.115 ± 0.008
GSE4922, Grade	0.108	0.139 ± 0.010	0.143 ± 0.010	0.140 ± 0.011	0.145 ± 0.010	0.153 ± 0.010
GSE19159	0.122	0.224 ± 0.011	0.209 ± 0.011	0.214 ± 0.012	0.246 ± 0.013	0.217 ± 0.012

*Each rate is accompanied with a standard error.

To study the impact of gene-gene correlations on classifier performance, we computed the MR ratio of enRS-FLD to allDLDA and the MR ratio of enRS-FLD to filterDLDA for each real dataset. The relationships between gene-gene correlations and the MR ratios are demonstrated in Figure 2. Overall, most of the MR ratios were less than 1, and they decreased as the correlations increased. The relationships for simulated datasets are also shown for comparison. The relationships between gene-gene correlations and the MR ratios of enRS-FLD to filterDLDA for real datasets were similar to those of simulated datasets for Structure 3.

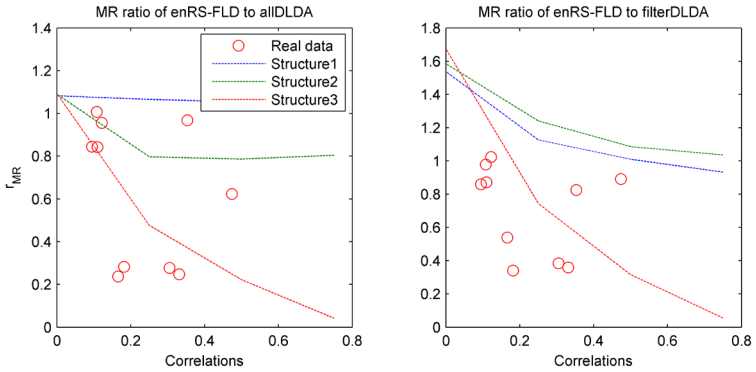

Figure 2. MR ratio of classifiers for real datasets and simulated datasets.

Classifier performance with simulated datasets using modified settings

The fold-changes and the number of differentially expressed genes (DEGs) may also affect classifier performance (Novianti et al., 2015). These parameters were computed for each real dataset and are shown in **Figures S1-S10** and **Table S1**. They differed among datasets; DEGs typically represented greater than 5% of genes; the (normalized) fold changes were normally less than 1. Accordingly, simulations were performed using modified settings in which the first p_1 = 200 genes were differentially expressed between classes with a mean difference of 0.5. The results are summarized in Figure 3.

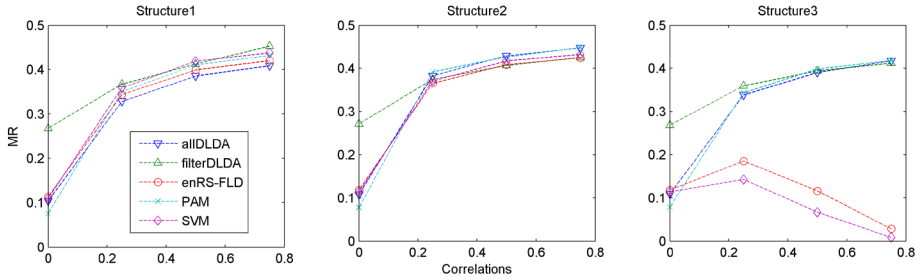

Figure 3. Average MR of simulated datasets using modified settings.

Overall, the average MRs of the three non-correlation-based classifiers were similar to or higher than those of the two correlation-based classifiers in all cases. For Structure 3, the average MRs of the two correlation-based classifiers were largely lower than those of the three non-correlation-based classifiers in the case of non-zero correlations.

DISCUSSION

Classifier performance

In this study, the performances of non-correlation-based classifiers (allDLDA, filterDLDA, and PAM) and correlation-based classifiers (enRS-FLD and SVM) were studied using simulated datasets and real datasets. Using simulated datasets, the correlation-based classifiers benefited from high correlations in the case of Structure 3, in which more genes were correlated. Using the real datasets, the classifier enRS-FLD outperformed the classifiers allDLDA and filterDLDA with respect to prediction accuracy, especially when the correlations were high, and outperformed PAM, especially for dataset GSE42133. Additionally, the enRS-FLD algorithm was simpler and more robust than SVM, and SVM might fail occasionally owing to a lack of convergence.

The relationships between gene-gene correlations and the MR ratios of enRS-FLD to filterDLDA for real datasets were similar to the relationships for simulated datasets for Structure 3, suggesting that Structure 3 was more realistic in this case.

Comparison with previous studies

The MR results for enRS-FLD were consistent with previous results obtained for enRS-FLD (Durrant and Kabán, 2015) using the shared datasets Colon, ALL-AML, and Prostate. The enRS-FLD analysis in this study outperformed previous results using PLR (Kim and Simon, 2011) with simulated datasets and outperformed FLDA (Dudoit et al., 2002) with the shared dataset ALL-AML. They were consistent with results obtained for the module-based classifier in (Pramparo et al., 2015) using the dataset GSE42133 and a test dataset and outperformed the classification results in (Gowrishankar et al., 2015) using the shared dataset GSE57162. They were consistent with the results of the boosting classifiers in (Blagus and Lusa, 2015) using the shared datasets ‘GSE4922, ER’ and ‘GSE4922, Grade’ and were consistent with the results obtained using the classifiers in (Gravier et al., 2010) with the shared dataset GSE19159. Accordingly, the ensemble RS-FLD classifier is an effective new technique.

CONCLUSION

The correlation-based classifier ensemble RS-FLD was effective and benefited from gene-gene correlations, particularly when the correlations were high.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

The authors are grateful to the authors of the datasets and those responsible for the availability of the data.

REFERENCES

Alon U, Barkai N, Notterman DA, Gish K, et al. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* 96: 6745-6750. http://dx.doi.org/10.1073/pnas.96.12.6745

Blagus R and Lusa L (2015). Boosting for high-dimensional two-class prediction. *BMC Bioinformatics* 16: 300. http://dx.doi.org/10.1186/s12859-015-0723-9

Dataset *ALL-AML* (2016). Available at [http://www.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=43]. Accessed January 30, 2016.

Dataset *Colon* (2016). Available at [http://www.weizmann.ac.il/mcb/UriAlon/sites/mcb.UriAlon/files/uploads/DownloadableData/affydata.zip]. Accessed January 30, 2016.

Dataset *Prostate* (2016). Available at [http://www-genome.wi.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=75]. Accessed January 30, 2016.

Dembélé D and Kastner P (2014). Fold change rank ordering statistics: a new method for detecting differentially expressed genes. *BMC Bioinformatics* 15: 14. http://dx.doi.org/10.1186/1471-2105-15-14

Dudoit S, Fridlyand J and Speed TP (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* 97: 77-87. http://dx.doi.org/10.1198/016214502753479248

Durrant R and Kabán A (2015). Random projections as regularizers: learning a linear discriminant from fewer observations than dimensions. *Mach. Learn.* 99: 257-286. http://dx.doi.org/10.1007/s10994-014-5466-8

Gene Expression Omnibus (2016). Available at [http://www.ncbi.nlm.nih.gov/geo/browse/]. Accessed January 30, 2016.

Golub TR, Slonim DK, Tamayo P, Huard C, et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286: 531-537. http://dx.doi.org/10.1126/science.286.5439.531

Gowrishankar B, Przybycin CG, Ma C, Nandula SV, et al. (2015). A genomic algorithm for the molecular classification of common renal cortical neoplasms: development and validation. *J. Urol.* 193: 1479-1485. http://dx.doi.org/10.1016/j.juro.2014.11.099

Gravier E, Pierron G, Vincent-Salomon A, Gruel N, et al. (2010). A prognostic DNA signature for T1T2 node-negative breast cancer patients. *Genes Chromosomes Cancer* 49: 1125-1134. http://dx.doi.org/10.1002/gcc.20820

Hastie T, Tibshirani R, Narasimhan B and Chu G (2014). Package 'pamr': Pam: prediction analysis for microarrays. R Package Version 1.55. Available at [http://cran.r-project.org/package=pamr]. Accessed January 30, 2016.

Ivshina AV, George J, Senko O, Mow B, et al. (2006). Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res.* 66: 10292-10301. http://dx.doi.org/10.1158/0008-5472.CAN-05-4414

Jong VL, Novianti PW, Roes KC and Eijkemans MJ (2014). Exploring homogeneity of correlation structures of gene expression datasets within and between etiological disease categories. *Stat. Appl. Genet. Mol. Biol.* 13: 717-732. http://dx.doi.org/10.1515/sagmb-2014-0003

Kalina J (2014). Classification methods for high-dimensional genetic data. *Biocybern. Biomed. Eng.* 34: 10-18. http://dx.doi.org/10.1016/j.bbe.2013.09.007

Kim KI and Simon R (2011). Probabilistic classifiers with high-dimensional data. *Biostatistics* 12: 399-412. http://dx.doi.org/10.1093/biostatistics/kxq069

McLachlan GJ (2004). *Discriminant analysis and statistical pattern recognition*. Hoboken NJ: Wiley-Interscience.
Novianti PW, Roes KC and Eijkemans MJ (2014). Evaluation of gene expression classification studies: factors associated with classification performance. *PLoS One* 9: e96063. http://dx.doi.org/10.1371/journal.pone.0096063
Novianti PW, Jong VL, Roes KC and Eijkemans MJ (2015). Factors affecting the accuracy of a class prediction model in gene expression data. *BMC Bioinformatics* 16: 199. http://dx.doi.org/10.1186/s12859-015-0610-4
Pramparo T, Pierce K, Lombardo MV, Barnes CC, et al. (2015). Prediction of autism by translation and immune/inflammation coexpressed genes in toddlers from pediatric community practices. *JAMA Psychiatry* 72: 386-394. http://dx.doi.org/10.1001/jamapsychiatry.2014.3008
Ross DT, Scherf U, Eisen MB, Perou CM, et al. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.* 24: 227-235. http://dx.doi.org/10.1038/73432
Saeys Y, Inza I and Larrañaga P (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics* 23: 2507-2517. http://dx.doi.org/10.1093/bioinformatics/btm344
Singh D, Febbo PG, Ross K, Jackson DG, et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1: 203-209. http://dx.doi.org/10.1016/S1535-6108(02)00030-2
Slawski M, Daumer M and Boulesteix AL (2008). CMA: a comprehensive Bioconductor package for supervised classification with high dimensional data. *BMC Bioinformatics* 9: 439. http://dx.doi.org/10.1186/1471-2105-9-439
Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, et al. (2005). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics* 21: 631-643. http://dx.doi.org/10.1093/bioinformatics/bti033
Tan XH, Cheng R, Hu HP and Bai YP (2015). Classification of colon cancer based on the expression of randomly selected genes. *Genet. Mol. Res.* 14: 12628-12635. http://dx.doi.org/10.4238/2015.October.19.6
Tibshirani R, Hastie T, Narasimhan B and Chu G (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA* 99: 6567-6572. http://dx.doi.org/10.1073/pnas.082099299
Valavanis I, Maglogiannis I and Chatziioannou AA (2015). Exploring robust diagnostic signatures for cutaneous melanoma utilizing genetic and imaging data. *IEEE J. Biomed. Health Inform.* 19: 190-198. http://dx.doi.org/10.1109/JBHI.2014.2336617
Xu P, Zhu J, Zhu L and Li Y (2015). Covariance-enhanced discriminant analysis. *Biometrika* 102: 33-45. http://dx.doi.org/10.1093/biomet/asu049
Zhang J and Pan M (2016). A high-dimension two-sample test for the mean using cluster subspaces. *Comput. Stat. Data Anal.* 97: 87-97. http://dx.doi.org/10.1016/j.csda.2015.12.004
Zhang W, Yu Y, Hertwig F, Thierry-Mieg J, et al. (2015). Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biol.* 16: 133. http://dx.doi.org/10.1186/s13059-015-0694-1

Supplementary material

Table S1. Gene expression structure.

Figure S1. Distributions of fold change and within-class correlation for dataset Colon. Solid lines represent the results of real dataset. Dashed lines represent the results of simulation dataset generated according to independent normal distribution.

Figure S2. Distributions of fold change and within-class correlation for dataset ALL-AML.

Figure S3. Distributions of fold change and within-class correlation for dataset Prostate.

Figure S4. Distributions of fold change and within-class correlation for dataset GSE42133.

Figure S5. Distributions of fold change and within-class correlation for dataset GSE57162.

Figure S6. Distributions of fold change and within-class correlation for dataset 'GSE49710, HR-Event'.

Figure S7. Distributions of fold change and within-class correlation for dataset 'GSE49710, Class'.

Figure S8. Distributions of fold change and within-class correlation for dataset 'GSE4922, ER'.

Figure S9. Distributions of fold change and within-class correlation for dataset 'GSE4922, Grade'.

Figure S10. Distributions of fold change and within-class correlation for dataset GSE19159.