# A Comparative Study of Classification Methods for Microarray Data Analysis

**Hong Hu**[1] **Jiuyong Li**[1] **Ashley Plank**[1] **Hua Wang**[1] **Grant Daggard**[2]

Department of Mathematics and Computing[1]
Department of Biological and Physical Sciences[2]
University of Southern Queensland,
Toowoomba, QLD 4350, Australia
Email: huhong@usq.edu.au

**Abstract**

In response to the rapid development of DNA Microarray technology, many classification methods have been used for Microarray classification. SVMs, decision trees, Bagging, Boosting and Random Forest are commonly used methods. In this paper, we conduct experimental comparison of LibSVMs, C4.5, BaggingC4.5, AdaBoostingC4.5, and Random Forest on seven Microarray cancer data sets. The experimental results show that all ensemble methods outperform C4.5. The experimental results also show that all five methods benefit from data preprocessing, including gene selection and discretization, in classification accuracy. In addition to comparing the average accuracies of ten-fold cross validation tests on seven data sets, we use two statistical tests to validate findings. We observe that Wilcoxon signed rank test is better than sign test for such purpose.

**Keywords:** Microarray data, classification.

## 1 Introduction

In recent years, the rapid development of DNA Microarray technology has made it possible for scientists to monitor the expression level of thousands of genes with a single experiment (Schena, Shalon, Davis & Brown 1995, Lockhart, Dong, Byrne & et al. 1996). With DNA expression Microarray technology, researchers will be able to classify different diseases according to different expression levels in normal and tumor cells, to discover the relationship between genes, to identify the critical genes in the development of disease. There are many active research applications of Microarray technology, such as cancer classification (Golub, Slonim, Tamayo & et al. 1999, Veer, Dai, de Vijver & et al. 2002, PetricoinIII, Ardekani, Hitt, Levine & et al. 2002), gene function identification (Lu, Patterson, Wang, Marquez & Atkinson 2004, Santin, Zhan, Bellone & Palmieri 2004), clinical diagnosis (Yeang, Ramaswamy, Tamayo & et al. 2001), and drug discovery studies (Maron & Lozano-Pérez 1998).

This project was partially supported by Australian Research Council Discovery Grant DP0559090.



A main task of Microarray classification is to build a classifier from historical Microarray gene expression data, and then it uses the classifier to classify future coming data. Many methods have been used in Microarray classification, and typical methods are Support Vector Machines (SVMs) (Brown, Grundy, Lin, Cristianini, Sugnet, Furey, Jr & Haussler 2000, Guyon, Weston, Barnhill & Vapnik 2002), k-nearest neighbor classifier (Yeang et al. 2001), C4.5 decision tree (Li & Liu 2003, Li, Liu, Ng & Wong 2003), rule-base classification method (Yeang et al. 2001) and ensemble methods, such as Bagging and boosting (Tan & Gibert 2003, Dietterich 2000).

SVMs, decision trees and ensemble methods are most frequently used methods in Microarray classification. Reading through the literature of Microarray data classification, it is difficult to find consensus conclusions on their relative performance. We are very interested in classifying Microarray data using C4.5 since it provides more interpretable results than other methods do. Therefore, we design an experiment to find out the classification performance of C4.5, AdaBoostingC4.5, BaggingC4.5, Random Forests, Libsvms on seven Microarray cancer data sets.

In the experimental analysis, we use sign test and Wilcoxon signed rank test to compare classification performance of different methods. We find that Wilcoxon signed rank test is better than sign test for such comparison. We also find inconsistent results in accuracy test and Wilcoxon signed rank test, and we interpret the results in a reasonable way.

The rest of this paper is organized as follows. In Section 2, we describe the relevant methods in this comparison study. In Section 3, we introduce our experimental design. In Section 4, we show our experimental results and present discussions. In Section 5, we conclude the paper.

## 2 Algorithm selected for comparison

Numerous Microarray data classification algorithms have been proposed in recent years. Most of them have been adapted from current data mining and machine learning algorithms.

C4.5 (Quinlan 1993, Quinlan 1996) was proposed by Quinlan in 1993 and it is a typical decision tree algorithm. C4.5 partitions a training data into some disjoint subsets simultaneously, based on the values of an attribute. At each step in the construction of the decision tree, C4.5 selects an attribute which separates data with the highest information gain ratio (Quinlan 1993). The same process is repeated on all subsets until each subset contains only one class. To simplify the decision tree, the induced decision tree is pruned using pessimistic error estimation (Quinlan 1993).

SVMs was proposed by Cortes and Vapnik (Cortes

& Vapnik 1995) in 1995 and It has been a most influential classification algorithm in recent years. SVMs are classifiers which transform the input samples into a high dimensional space by a kernel function and use a linear hyperplane to separate two classes mapped to that high dimensional space by support vectors which are selected vectors from training samples. SVMs has been applied to many domains, for example, text categorization (Joachims 1998), cancer classification (Furey, Christianini, Duffy, Bednarski, Schummer & Hauessler 2000, Brown et al. 2000, Brown, Grundy, Lin, Cristianini, Sugnet, Ares & Haussler 1999).

In the past decade, many researchers have devoted their efforts to the study of ensemble decision tree methods for Microarray classification. Ensemble decision tree methods combine decision trees generated from multiple training data sets by re-sampling the training data set. Bagging, Boosting and Random forests are some of the well-known ensemble methods in the machine learning field.

Bagging was proposed by Leo Breiman (Breiman 1996) in 1996. Bagging uses a bootstrap technique to re-sample the training data sets. Some samples may appear more than once in a data set whereas some samples do not appear. A set of alternative classifiers are generated from a set of re-sampled data sets. Each classifier will in turn assign a predicted class to an incoming test sample. The final predicted class for the sample is determined by the majority vote. All classifiers have equal weights in voting.

The boosting method was first developed by Freund and Schapire (Freund & Schapire 1996) in 1996. Boosting uses a re-sampling technique different from Bagging. A new training data set is generated according to its sample distribution. The first classifier is constructed from the original data set where every sample has an equal distribution ratio of 1. In the following training data sets, the distribution ratios are made differently among samples. A sample distribution ratio is reduced if the sample has been correctly classified; otherwise the ratio is kept unchanged. Samples which are misclassified often get duplicates in a re-sampled training data set. In contrast, samples which are correctly classified often may not appear in a re-sampled training data set. A weighted voting method is used in the committee decision. A higher accuracy classifier has larger weight than a lower accuracy classifier. The final verdict goes along with the largest weighted votes.

Based on Bagging, Leo Breiman introduced another ensemble decision tree method called Random Forests (Breiman 1999) in 1999. This method combines Bagging and random feature selection methods to generate multiple classifiers.

## 3 Experimental design methodology

### 3.1 Ten-fold cross-validation

Tenfold cross-validation is used in this experiment. In tenfold cross-validation, a data set is equally divided into 10 folds (partitions) with the same distribution. In each test 9 folds of data are used for training and one fold is for testing (unseen data set). The test procedure is repeated 10 times. The final accuracy of an algorithm will be the average of the 10 trials.

### 3.2 Test data sets

Seven data sets from Kent Ridge Biological Data Set Repository (**?**) are selected. These data sets were collected from very well researched journal papers, namely Breast Cancer (Veer et al. 2002), Lung Cancer (Gordon, Jensen, Hsiao, Gullans & et al. 2002), Lymphoma (Alizadeh, Eishen, Davis, Ma & et al. 2000), ALL-AML Leukemia (Golub et al. 1999), Colon (Alon & et al. 1999), Ovarian (PetricoinIII et al. 2002) and Prostate (Singh & et al. 2002). Table 1 shows the summary of the characteristics of the seven data sets. We conduct our experiments by using tenfold cross-validation on the merged original training and test data sets.

| Data set | Genes | Class | Record |
|---|---|---|---|
| Breast Cancer | 24481 | 2 | 97 |
| Lung Cancer | 12533 | 2 | 181 |
| Lymphoma | 4026 | 2 | 47 |
| Leukemia | 7129 | 2 | 72 |
| Colon | 2000 | 2 | 62 |
| Ovarian | 15154 | 2 | 253 |
| Prostate | 12600 | 2 | 21 |

Table 1: Experimental data set details

### 3.3 Softwares used for comparison

We have done our experiments with C4.5, C4.5AdaBoosting, C4.5Bagging, Random forests, LibSVMs with the Weka-3-5-2 package which is available online (`http://www.cs.waikato.ac.nz/ml/weka/`). Default settings are used for all compared ensemble methods. We were aware that the accuracy of some methods on some data sets can be improved when parameters were changed. However, it was difficult to find another uniform setting good for all data sets. Therefore, we did not change default settings since the default produced high accuracy on average.

### 3.4 Microarray data preprocessing

We used information gain ratio for gene selection and used Fayyad and Irani's MDL discretization method provided by Weka to discretize numerical attributes. Our previous results (Hu, Li, Wang & Daggard 2006) show that with preprocessing, the number of genes selected affects the classification accuracy. The overall performance is better when data sets contain 50 to 100 genes. For our experiment, we set the number of genes as 50. After the data preprocessing, each data set contains 50 genes with discretized values.

### 3.5 Sign test

Sign test (Conover 1980) is used to test whether one random variable in a pair tends to be larger than the other random variable in the pair. Given $n$ pairs of observations. Within each pair, either a plus, tie or minus is assigned. The plus corresponds to that one value is greater than the other, the minus corresponds to that one value is less than the other, and the tie means that both equal to each other. The null hypothesis is that the number of pluses and minuses are equal. If the null hypothesis test is rejected, then one random variable tends to be greater than the other.

### 3.6 Wilcoxon signed rank test

Sign test only makes use of information of whether a value is greater, less than or equal to the other in a pair. Wilcoxon signed rank test (Conover 1980, Daniel 1978) calculates differences of pairs. The absolute differences are ranked after discarding pairs with the difference of zero. The ranks are sorted in ascending order. When several pairs have absolute differences that are equal to each other, each of these

| Data set | C4.5 | Random Forests | AdaBoostC4.5 | BaggingC4.5 | LibSVMs |
|---|---|---|---|---|---|
| Breast Cancer | 84.5 | 88.7 | 90.7 | 85.6 | 72.2 |
| Lung Cancer | 98.3 | 99.5 | 98.3 | 97.8 | 100.0 |
| Lymphoma | 74.5 | 93.6 | 89.4 | 89.4 | 55.3 |
| Leukemia | 88.9 | 98.6 | 95.8 | 95.8 | 100.0 |
| Colon | 88.7 | 83.9 | 90.3 | 90.3 | 90.3 |
| Ovarian | 96.8 | 99.2 | 98.8 | 98.0 | 100.0 |
| Prostate | 95.2 | 100 | 95.2 | 95.2 | 100.0 |
| Average | 89.6 | 94.8 | 94.1 | 93.2 | 88.3 |

Table 2: Average accuracy of seven preprocessed data sets with five classification algorithms based on tenfold cross-validation

| | C4.5 | Random Forests | AdaBoostC4.5 | BaggingC4.5 | LibSVMs |
|---|---|---|---|---|---|
| C4.5 | – | | | | |
| Random Forests | 0.063 | – | | | |
| AdaboostC4.5 | **0.031** | 0.63 | – | | |
| BaggingC4.5 | 0.11 | **0.0088** | **0** | – | |
| LibSVMs | 0.23 | 0.34 | 0.34 | 0.34 | – |

Table 3: Summary of sign test between any two of the compared classification methods. P-values of the test are given, and significant p-values at 95% confidence level are highlighted.

several pairs is assigned as the average of ranks that would have otherwise been assigned. The hypothesis is that the differences have the mean of 0.

## 4 Experimental results and discussions

Table 2 shows the individual and average accuracy results of all the compared methods based on seven preprocessed data sets with the tenfold cross-validation method. Table 5 shows the individual and average accuracy results of the compared methods based on seven original data sets with tenfold cross-validation method.

Based on Table 2, we have the following conclusions: with preprocessed data sets, all ensemble methods on average perform better than C4.5 and LibSVMs. Both C4.5 and LibSVM perform similar to each other.

Those results demonstrate that the ensemble decision tree methods can improve the accuracy over single decision tree method on Microarray data sets. These results are consistent with most machine learning study.

To determine whether the ensemble methods consistently outperform single classification methods, we also conducted a sign test. The results are shown in Table 3. Based on the sign test, we have the following conclusions.

1. AdaBoostC4.5 is the only one among the all compared classification algorithms that outperforms C4.5.

2. Comparing between ensemble methods, Random Forests and AdaBoostC4.5 outperform BaggingC4.5 significantly.

3. No sufficient evidence supports that any ensemble method and C4.5 outperform LibSVMs.

We have the following observations from the sign test. The average difference of 6.5% (between Random Forest and LibSVM) may not be statistically significant, but the average difference of 0.9% (between AdaBoostC4.5 and BaggingC4.5) are statistically significant. This may sounds strange, but is understandable. The average accuracy indicates the average performance of a method on the data sets. However, the sign test indicates if a method is consistently better than another on each test data set. The accuracy difference can be very small. For example, each accuracy value of AdaboostC4.5 is slightly higher than Bagging C4.5, and hence Sign test shows that AdaBoostC4.5 is significantly better than BaggingC4.5. However, the accuracy improvement is marginal.

This also indicates a limitation of the sign test: the difference of 0.01 and 10.0 are considered the same in the sign test since only plus or minus is used. We conducted a Wilcoxon signed rank test based on Table 2. The results of Wilcoxon signed rank test is shown in Table 4

Table 4 shows that all ensemble methods, Random Forest, AdaBoostC4.5 and BaggingC4.5, are significantly more accurate than C4.5. This conclusion is consistent with most research literature. Though AdaBoostC4.5 performs marginally better than BaggingC4.5 on each data set. The Wilcoxon signed rank test does not support that the differences are significant. We tend to believe that the Wilcoxon signed rank test is better than sign test for our purpose.

Based on Table 2 and table 4, we can conclude that all ensemble methods significantly outperform C4.5. We do not have sufficient evidence to show whether LibSVM and another method is better. Though Table 2 give a large average accuracy difference between an ensemble method and LibSVM, we do not know wether LibSVM and an ensemble method will perform better on a data set. This is because that SVM and decision trees are two different types of classification methods. They are suitable for different data sets.

To show that all methods benefit from data preprocessing, we conducted experiments on original data sets, and show their accuracy results in Table 5.

Table 2 and Table 5, clearly indicate that all classification methods on data preprocessed by discretization and gene selection methods achieve higher average accuracy over themselves on data without data preprocessing. After data preprocessing, accuracy performance has been improved significantly for all compared classification algorithms with up to 17.4% improvement.

To show that this improvement is significant, we conducted Sign test and Wilcoxon signed rank test on differences between accuracies on preprocessed data and original data. The test results are shown in Table 6 and Table 7.

Based on a sign test of 95% confidence level, All methods except C4.5 improve the predictive accuracy on the preprocessed Microarray data sets than the original data sets. Not enough evidence supports that C4.5 performs significantly better on the preprocessed

| | C4.5 | Random Forests | AdaBoostC4.5 | BaggingC4.5 | LibSVMs |
|---|---|---|---|---|---|
| C4.5 | – | | | | |
| Random Forests | $\leq$ **0.05** | – | | | |
| AdaboostC4.5 | **0.005** | 0.2-0.3 | – | | |
| BaggingC4.5 | **0.025** | 0.1-0.2 | 0.091 | – | |
| LibSVMs | 0.5 | 0.4-0.5 | 0.4-0.5 | 0.4-0.5 | – |

Table 4: Summary of Wilcoxon signed rank test between any two of the compared classification methods. P values are shown and significant p-values at 95% confidence level are highlighted.

| Data set | C4.5 | Random Forests | AdaBoostC4.5 | BaggingC4.5 | LibSVMs |
|---|---|---|---|---|---|
| Breast Cancer | 62.9 | 61.9 | 61.9 | 66.0 | 52.6 |
| Lung Cancer | 95.0 | 98.3 | 96.1 | 97.2 | 82.9 |
| Lymphoma | 78.7 | 80.9 | 85.1 | 85.1 | 55.3 |
| Leukemia | 79.2 | 86.1 | 87.5 | 86.1 | 65.3 |
| Colon | 82.3 | 75.8 | 77.4 | 82.3 | 64.5 |
| Ovarian | 95.7 | 94.1 | 95.7 | 97.6 | 87.0 |
| Prostate | 33.3 | 52.4 | 33.3 | 42.9 | 61.9 |
| Average | 75.3 | 78.5 | 76.7 | 79.6 | 67.1 |
| Difference | 14.3 | 16.3 | 17.4 | 13.6 | 21.2 |

Table 5: Average accuracy on seven original data sets of five classification methods based on tenfold cross-validation. The last row shows the differences in average accuracy between the average accuracy based on preprocessed data and original data for every compared classification method

Microarray data sets than the original data set.

These results show that the data precessing method improves the predictive accuracy of classification. As we mentioned before, Microarray data contains irrelevant and noisy genes. Those genes do not help classification but reduce the predictive accuracy . Microarray data preprocessing is able to reduce the number of irrelevant genes in Microarray data classification and therefore can generally help to improve the classification accuracy.

Apart from predictive accuracy, the representation of predictive results is another important fact for determining the quality of a classification algorithm. Among the compared algorithms, the classifier of C4.5 is a tree, and the classifier of an ensemble method is formed by a group of trees. Trees are more easier to be evaluated and interpreted by users. By contrast, the outputs of SVMs are numerical values and are less interpretable.

## 5 Conclusion

In this paper, we conducted a comparative study of classification methods for Microarray data analysis. We compared five classification methods, namely LibSVMs, C4.5, BaggingC4.5, AdaBoostingC4.5, and Random Forest, on seven Microarray data sets, with or without gene selection and discretization. The experimental results show that all ensemble methods are significantly more accurate than C4.5. Data preprocessing significantly improves accuracies of all five methods. We conducted both sign test and Wilcoxon signed rank test to evaluate the performance differences of comparative methods. We observed that the Wilcoxon signed rank test is better than the sign test. We also found that there is no sufficient evidence to support the performance difference between the SVM and an ensemble method although the average accuracy of SVM is much lower than that of an ensemble method. A possible explanation is that they are two different classification schemes, and hence one may be able to suits for a data set whereas the other does not.

## References


Alizadeh, A., Eishen, M., Davis, E., Ma, C. & et al. (2000), 'Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling', *Nature* **403**, 503–511.

Alon, U. & et al. (1999), 'Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays', *PNAS* **96**, 6745–6750.

Breiman, L. (1996), 'Bagging predictors', *Machine Learning* **24**(2), 123–140.

Breiman, L. (1999), Random forests–random features, Technical Report 567, University of California, Berkley.

Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey, T., Jr, M. & Haussler, D. (2000), Knowledge-based analysis of microarray gene expression data by using suport vector machines, *in* 'Proc. Natl. Acad. Sci.', Vol. 97, pp. 262–267.

Brown, M., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C., Ares, M. & Haussler, D. (1999), Support vector machine classification of microarray gene expression data, Technical Report UCSC-CRL-99-09, University of California, Santa Cruz, Santa Cruz, CA 95065.

Conover, W. J. (1980), *Practical nonparametric statistics*, Wiley, New York.

Cortes, C. & Vapnik, V. (1995), 'Support-vector networks.', *Machine Learning* **20**(3), 273–297.

Daniel, W. W. (1978), *Applied nonparametric statistics*, Houghton Mifflin, Boston.

Dietterich, T. G. (2000), 'An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization', *Machine learning* **40**, 139–157.

Freund, Y. & Schapire, R. E. (1996), Experiments with a new boosting algorithm, *in* 'International Conference on Machine Learning', pp. 148–156.

Furey, T. S., Christianini, N., Duffy, N., Bednarski, D. W., Schummer, M. & Hauessler, D. (2000), 'Support vector machine classification and validation of cancer tissue samples using microarray expression data.', *Bioinformatics* **16**(10), 906–914.

| | | | with original data | | |
|---|---|---|---|---|---|
| with preprocessed data | C4.5 | Random Forests | AdaBoostC4.5 | BaggingC4.5 | LibSVMs |
| C4.5 | 0.0625 | | | | |
| Random Forests | | **0.0078** | | | |
| AdaboostC4.5 | | | **0.0078** | | |
| BaggingC4.5 | | | | **0.0078** | |
| LibSVMs | | | | | **0.0156** |

Table 6: Summary of sign test between accuracy of the compared classification methods on original and preprocessed data sets. P values at 95% confidence level are highlighted.

| | | | with original data | | |
|---|---|---|---|---|---|
| with preprocessed data | C4.5 | Random Forests | AdaBoostC4.5 | BaggingC4.5 | LibSVMs |
| C4.5 | **0.025** | | | | |
| Random Forests | | **0.005** | | | |
| AdaboostC4.5 | | | **0.005** | | |
| BaggingC4.5 | | | | **0.005** | |
| LibSVMs | | | | | **0.01** |

Table 7: Summary of Wilcoxon signed rank test between accuracy of the compared classification methods based on original and preprocessed data sets. P values at 95% confidence level are highlighted


Golub, T., Slonim, D., Tamayo, P. & et al. (1999), 'Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring', *Science* **286**, 531–537.

Gordon, G., Jensen, R., Hsiao, L.-L., Gullans, S. & et al. (2002), 'Translation of microarray data into clinically relevant cancer diagnostic tests using gege expression ratios in lung cancer and mesothelioma', *Cancer Research* **62**, 4963–4967.

Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. (2002), 'Gene selection for cancer classification using support vector machines', *Machine Learning* **46**(1-3), 389–422.

Hu, H., Li, J., Wang, H. & Daggard, G. (2006), Combined gene selection methods for microarray data analysis, *in* '10th International Conference on KnowledgeBased & Intelligent Information & Engineering Systems. To appear'.

Joachims, T. (1998), Text categorization with support vector machines: learning with many relevant features, *in* 'Proceedings of 10th European Conference on Machine Learning', number 1398, pp. 137–142.

Li, J. & Liu, H. (2003), Ensembles of cascading trees, *in* 'ICDM', pp. 585–588.

Li, J., Liu, H., Ng, S.-K. & Wong, L. (2003), Discovery of significant rules for classifying cancer diagnosis data, *in* 'ECCB', pp. 93–102.

Lockhart, D., Dong, H., Byrne, M. & et al. (1996), 'Expression monitoring by hybridization to high-density oligonucleotide arrays', *Nature Biotechnology* **14**, 1675–1680.

Lu, K., Patterson, A. P., Wang, L., Marquez, R. & Atkinson, E. (2004), 'Selection of potential markers for epithelial ovarian cancer with gene expression arrays and recursive descent partition analysis', *Clin Cancer Res* **10**, 291–300.

Maron, O. & Lozano-Pérez, T. (1998), A framework for multiple-instance learning, *in* M. I. Jordan, M. J. Kearns & S. A. Solla, eds, 'Advances in Neural Information Processing Systems', Vol. 10, The MIT Press, pp. 570–576.

PetricoinIII, E., Ardekani, A., Hitt, B., Levine, P. & et al. (2002), 'Use of proteomic patterns in serum to identify ovarian cancer', *The lancet* **359**, 572–577.

Quinlan, J. (1996), 'Improved use of continuous attributes in C4.5', *Artificial Intelligence Research* **4**, 77–90.

Quinlan, J. R. (1993), *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, California.

Santin, A., Zhan, F., Bellone, S. & Palmieri, M. (2004), 'Gene expression profiles in primary ovarian serous papillary tumors and normal ovarian epithelium: idnetification of candidate molecular markers for ovarian cancer diagnosis and therapy', *International Journal of Cancer* **112**, 14–25.

Schena, M., Shalon, D., Davis, R. & Brown, P. (1995), 'Quantitative monitoring of gene expression patterns with a complementary DNA microarray', *Science* **270**, 467–470.

Singh, D. & et al. (2002), 'Gene expression correlates of clinical prostate cancer behavior', *Cancer Cell* **1**, 203–209.

Tan, A. C. & Gibert, D. (2003), 'Ensemble machine learning on gene expression data for cancer classification', *Applied Bioinformatics* **2**(3), s75–s83.

Veer, L. V., Dai, H., de Vijver, M. V. & et al. (2002), 'Gene expression profiling predicts clinical outcome of breast cancer', *Nature* **415**, 530–536.

Yeang, C., Ramaswamy, S., Tamayo, P. & et al. (2001), 'Molecular classification of multiple tumor types', *Bioinformatics* **17**(Suppl 1), 316–322.