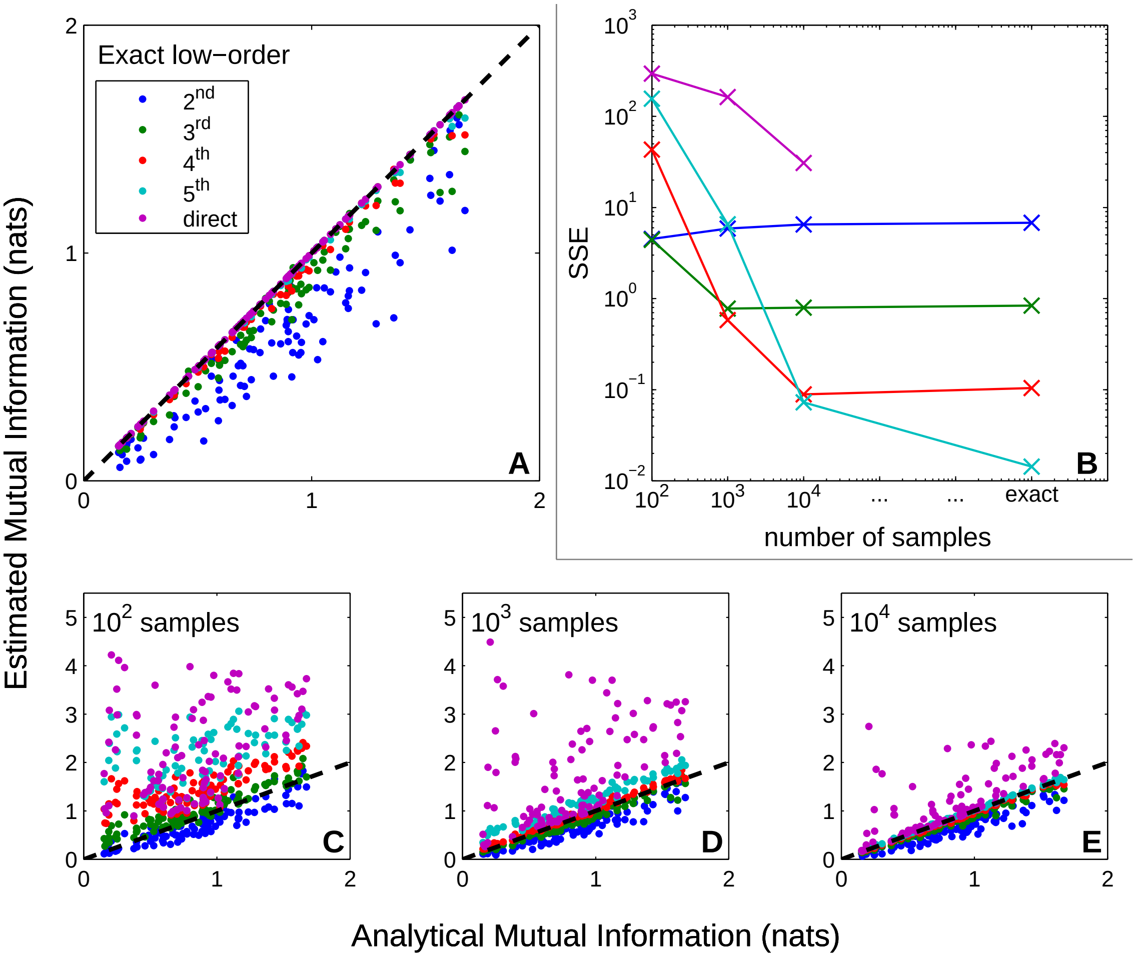

Figure S1: *Direct validation of MIST MI approximation.* To evaluate the MIST framework, we simulated 100 randomly generated networks with analytically computable joint entropies and applied the metrics using a range of sample sizes. Half of each network was randomly chosen and the MI between one half and the other was computed analytically or using the MIST approximation of various orders. When the analytical entropies are known exactly (A), the higher-order approximations performing increasingly well. When the entropies are estimated from a finite sample, however (C–E), the approximations provide the best estimates, with the higher-order approximations performing better as more data become available. This behavior is quantified by computing the sum-of-squared error of each metric as a function of the sampling regime (B). The best approximation to use depends upon the amount of data available, but for all cases examined with finite sample size, the approximations outperform direct estimation and the second-order approximation provides a good estimate.

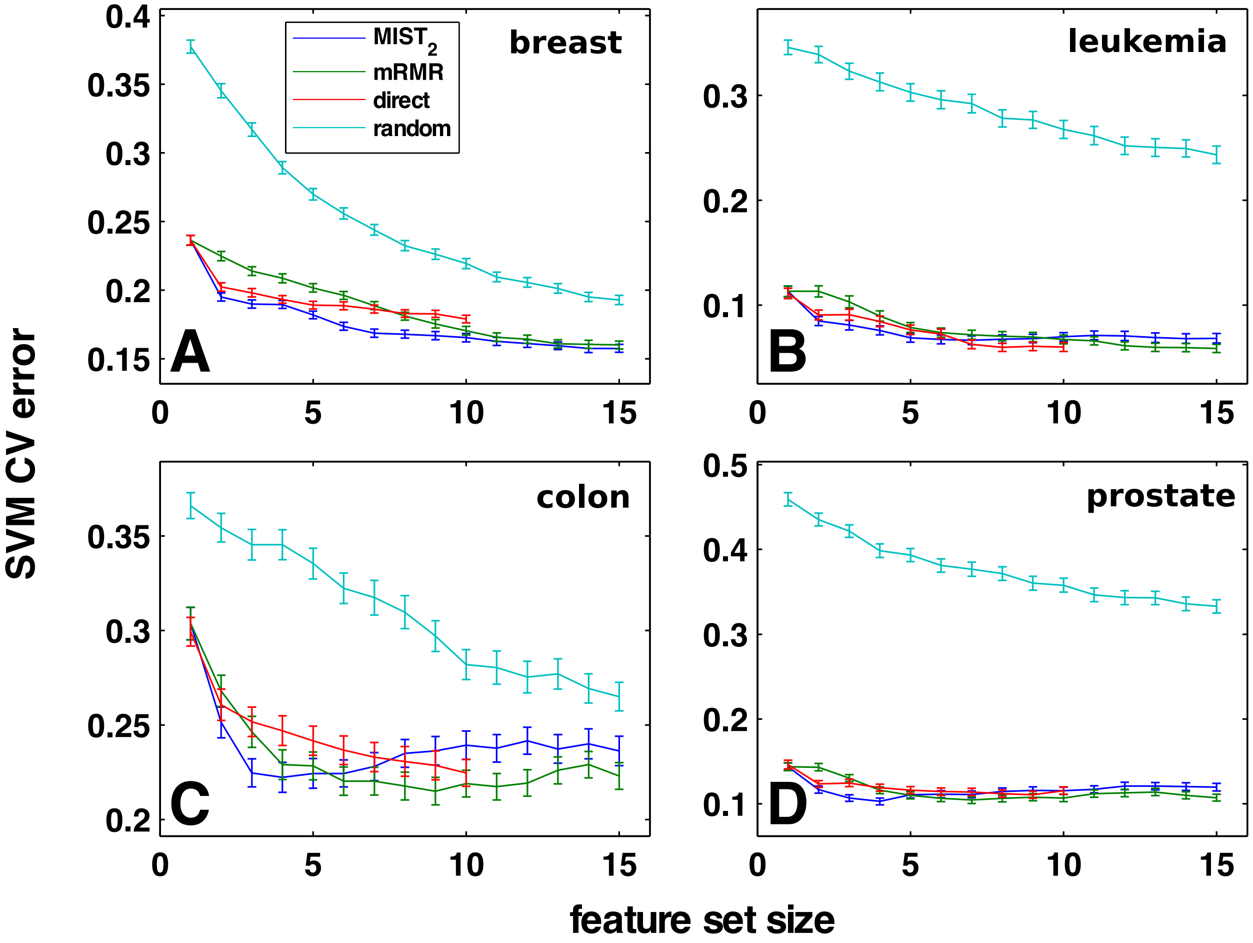

Figure S2: **Gene subset selection for cancer classification.** Subsets of gene expression levels were chosen incrementally to maximize the information content with the cancer class variable according to MIST$_2$, direct estimation of MI, mRMR, or at random and the chosen sets were scored by the cross-validation error of an SVM classifier trained to discriminate the cancer type. For all data sets, 75% of the data was separated and used to select features and train the model; the classifier was then used to classify the remaining 25% of the samples. The mean classification error and standard error of the mean for 200 such training/testing partitioning are reported. Genes were selected for data sets relating to (A) breast, (B) leukemia, (C) colon, and (D) prostate cancer.

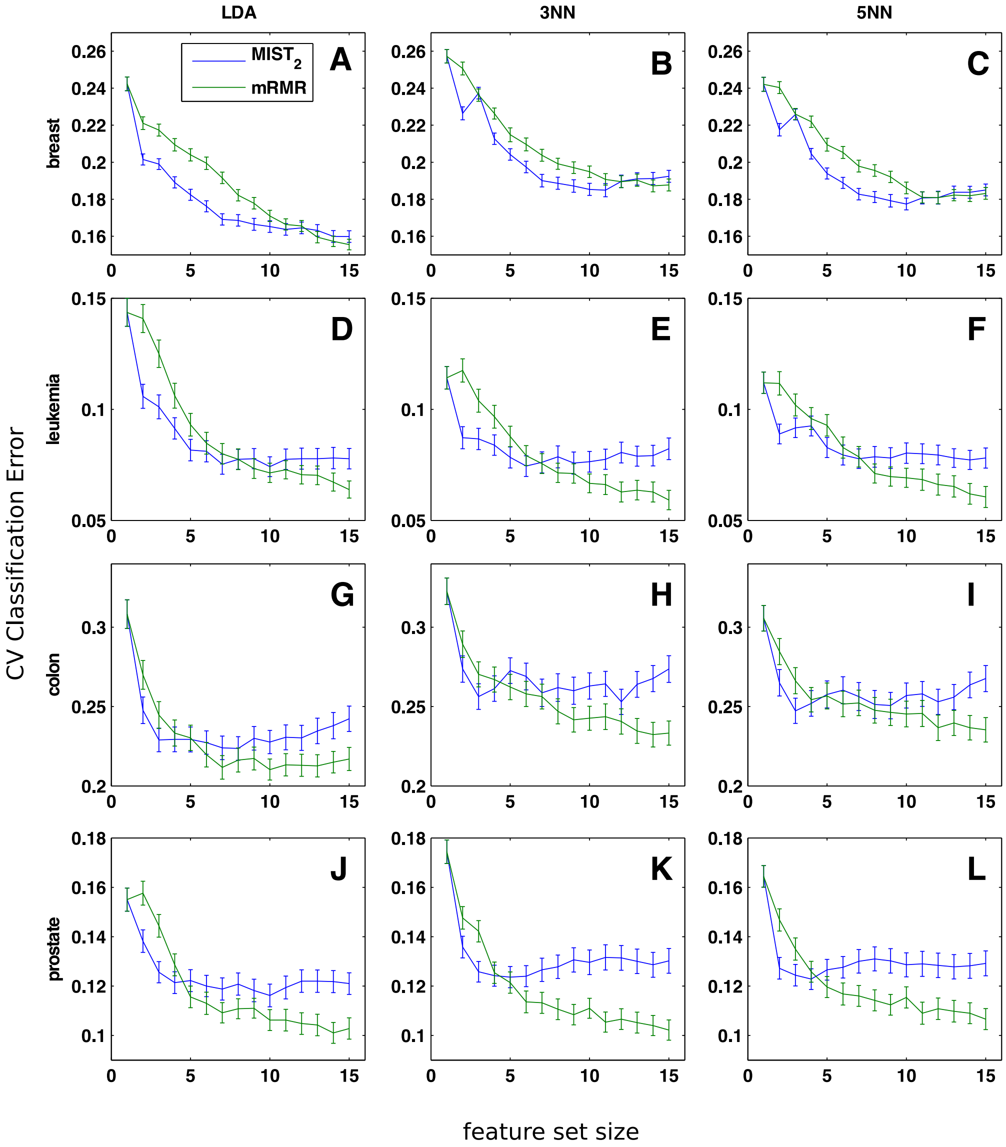

Figure S3: *Gene subset selection for cancer classification.* Subsets of gene expression levels were chosen incrementally to maximize the information content with the cancer class variable according to MIST$_2$ or mRMR and the chosen sets were scored by the cross-validation error of an LDA (A,D,G,J), 3NN (B,E,H,K), or 5NN (C,F,I,L) classifier trained to discriminate the cancer type. For all data sets, 75% of the data was separated and used to select features and train the model; the classifier was then used to classify the remaining 25% of the samples. The mean classification error and standard error of the mean for 200 such training/testing partitioning are reported. Genes were selected for four data sets relating to (A,B,C) breast, (D,E,F) leukemia, (G,H,I) colon, and (J,K,L) prostate cancer. Results using an SVM classifier and including direct estimation-based feature selection are shown in Figure 4.

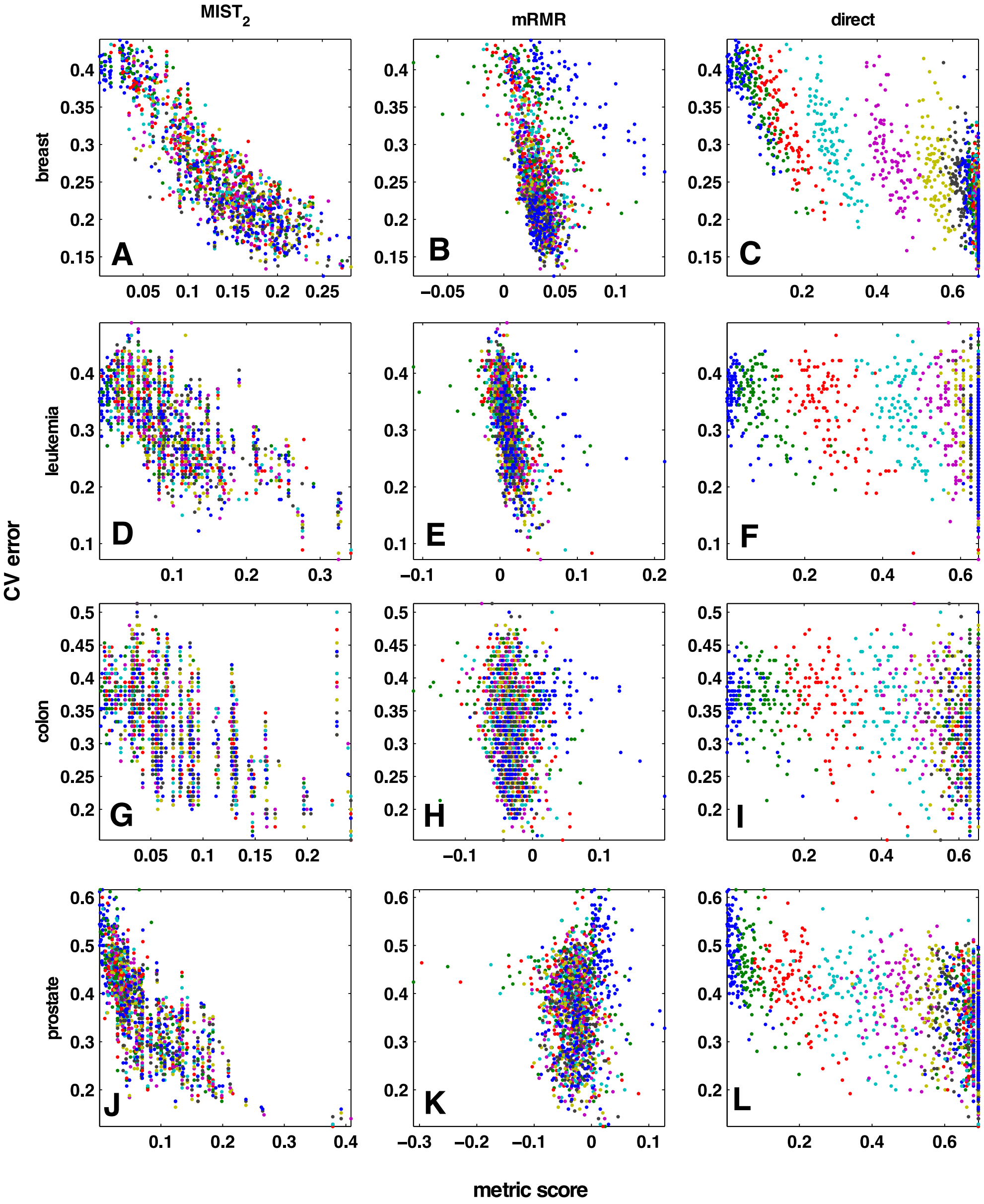

Figure S4: *Correlation of classification error and MI metrics.* The classification error of randomly chosen subsets of 1–15 genes was computed through cross-validation with an SVM based classifier. The same sets were then scored by MIST$_2$ (A,D,G,J), MI computed with direct estimation (B,E,H,K), and mRMR (C,F,I,L) and these metrics are shown plotted against the CV classification error. The color of the points relates to the size of the feature set, cycling through blue, green, red, cyan, magenta, yellow, black for increasing set size. The correlation coefficients between metrics as a function of set size is shown in Figure 3. Notably, MIST$_2$ has strong negative correlation across all feature set sizes.

Table S1: Microarray data sets for cancer classification

| Tissue | # Samples | # Genes | Class Type | Ref |
|---|---|---|---|---|
| breast | 295 | 70 | good/bad prog | [24] |
| leukemia | 72 | 7070 | AML/ALL | [11] |
| colon | 62 | 2000 | normal/tumor | [2] |
| prostate | 102 | 12600 | normal/tumor | [20] |

Table S2: Genes selected by MIST$_2$

| Tissue | # | Gene ID | Reproduce % | Cancer Relevance | Other Studies |
|---|---|---|---|---|---|
| breast | 1 | NM_003981 | 91.0* | [17] | [14, 5, 21] |
| | 2 | AI918032 | 91.0* | | [5] |
| | 3 | NM_003239 | 85.5* | [9] | [5] |
| | 4 | AW024884 | 52.0* | | |
| | 5 | AA404325 | 68.5* | | |
| | 6 | AF055033 | 77.0* | | [5, 21] |
| | 7 | AW014921 | 77.0* | | |
| | 8 | AL080059 | 49.5* | | [26] |
| | 9 | AI738508 | 1.5 | | |
| | 10 | AK000745 | 17.0 | | |
| leukemia | 1 | M27891 | 33.0* | | [10, 3, 6, 26, 4, 8] |
| | 2 | U29175 | 3.5* | | [3, 8] |
| | 3 | U72621 | 19.0* | [1] | [3] |
| | 4 | U88047 | 7.5* | | [8] |
| | 5 | M92287 | 24.0* | [19] | [3, 4, 8] |
| | 6 | M19507 | 2.0 | | [3, 6, 4, 8] |
| | 7 | D84294 | 0.5 | | |
| | 8 | HG3549-HT3751 | 6.5* | | |
| | 9 | M32304 | 6.5* | | [3] |
| | 10 | AF005043 | 1.0 | | |
| colon | 1 | M63391 | 22.0* | [7] | [3, 4, 8] |
| | 2 | U30825 | 3.5 | | [3, 8] |
| | 3 | T57468 | 4.5* | | [8] |
| | 4 | T47377 | 21.5* | | [3, 4, 8] |
| | 5 | M26383 | 19.0* | | [3, 4, 8] |
| | 6 | R39209 | 24.5* | | [8] |
| | 7 | M76378 | 5.5* | | [3, 4, 8] |
| | 8 | M80815 | 3.0 | | [3, 8] |
| | 9 | Y00097 | 4.5* | [18] | |
| | 10 | X90858 | 1.0 | [12] | [3] |
| prostate | 1 | X07732 | 90.0* | [13] | [6, 25, 22] |
| | 2 | U24577 | 33.0* | | |
| | 3 | M62895 | 6.0* | [16] | |
| | 4 | U12472 | 14.0* | | |
| | 5 | D80010 | 17.5* | | |
| | 6 | AB014545 | 15.0* | | |
| | 7 | AB023204 | 27.0* | | |
| | 8 | U67615 | 23.5* | | |
| | 9 | M21536 | 12.5* | [15] | |
| | 10 | AF038451 | 4.0* | [23] | |

*Bonferroni adjusted pval$\leq$0.01 for gene occuring this often in 200 random 10-feature selection runs.

# References

[1] A. Abdollahi, A. K. Godwin, P. D. Miller, L. A. Getts, D. C. Schultz, T. Taguchi, J. R. Testa, and T. C. Hamilton. Identification of a gene containing zinc-finger motifs based on lost expression in malignantly transformed rat ovarian surface epithelial cells. *Cancer Res*, 57(10):2029–2034, 1997.

[2] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A*, 96(12):6745–6750, 1999.

[3] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles. *J Comput Biol*, 7(3-4):559–583, 2000.

[4] T. Bo and I. Jonassen. New feature subset selection procedures for classification of expression profiles. *Genome Biol*, 3(4):RESEARCH0017, 2002.

[5] A. Choudhary, M. Brun, J. Hua, J. Lowey, E. Suh, and E. R. Dougherty. Genetic test bed for feature selection. *Bioinformatics*, 22(7):837–842, 2006.

[6] W. Chu, Z. Ghahramani, F. Falciani, and D. L. Wild. Biomarker discovery in microarray gene expression data with gaussian processes. *Bioinformatics*, 21(16):3385–3393, 2005.

[7] P. Dias, P. Kumar, H. B. Marsden, P. H. Morris-Jones, J. Birch, R. Swindell, and S. Kumar. Evaluation of desmin as a diagnostic and prognostic marker of childhood rhabdomyosarcomas and embryonal sarcomas. *Br J Cancer*, 56(3):361–365, 1987.

[8] C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol*, 3(2):185–205, 2005.

[9] T.-V. Do, L. A. Kubba, H. Du, C. D. Sturgis, and T. K. Woodruff. Transforming growth factor-beta1, transforming growth factor-beta2, and transforming growth factor-beta3 enhance ovarian cancer metastatic potential by inducing a smad3-dependent epithelial-to-mesenchymal transition. *Mol Cancer Res*, 6(5):695–705, 2008.

[10] J. Fand and J. Grzymala-Busse. *Leukemia Prediction from Gene Expression Data—A Rough Set Approach*, volume 4029 of *Lecture Notes in Computer Science*, pages 1611–3349. Springer, Berlin, 2006.

[11] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

[12] A. Kanzaki, Y. Takebayashi, H. Bando, J. F. Eliason, S.-i. Watanabe Si, H. Miyashita, M. Fukumoto, M. Toi, and T. Uchida. Expression of uridine and thymidine phosphorylase genes in human breast carcinoma. *Int J Cancer*, 97(5):631–635, 2002.

[13] K. A. Kelly, S. R. Setlur, R. Ross, R. Anbazhagan, P. Waterman, M. A. Rubin, and R. Weissleder. Detection of early prostate cancer using a hepsin-targeted imaging agent. *Cancer Res*, 68(7):2286–2291, 2008.

[14] H. Liu, J. Li, and L. Wong. Use of extreme patient samples for outcome prediction from gene expression data. *Bioinformatics*, 21(16):3377–3384, 2005.

[15] G. Petrovics, A. Liu, S. Shaheduzzaman, B. Furusato, C. Sun, Y. Chen, M. Nau, L. Ravindranath, Y. Chen, A. Dobi, V. Srikantan, I. A. Sesterhenn, D. G. McLeod, M. Vahey, J. W. Moul, and S. Srivastava. Frequent overexpression of ets-related gene-1 (erg1) in prostate cancer transcriptome. *Oncogene*, 24(23):3847–3852, 2005.

[16] S. A. Reeves, C. Chavez-Kappel, R. Davis, M. Rosenblum, and M. A. Israel. Developmental regulation of annexin ii (lipocortin 2) in human brain and expression in high grade glioma. *Cancer Res*, 52(24):6871–6876, 1992.

[17] A. Shimo, T. Nishidate, T. Ohta, M. Fukuda, Y. Nakamura, and T. Katagiri. Elevated expression of protein regulator of cytokinesis 1, involved in the growth of breast cancer cells. *Cancer Sci*, 98(2):174–181, 2007.

[18] S. Shin, K. L. Rossow, J. P. Grande, and R. Janknecht. Involvement of rna helicases p68 and p72 in colon cancer. *Cancer Res*, 67(16):7572–7578, 2007.

[19] E. Sicinska, I. Aifantis, L. Le Cam, W. Swat, C. Borowski, Q. Yu, A. A. Ferrando, S. D. Levin, Y. Geng, H. von Boehmer, and P. Sicinski. Requirement for cyclin d3 in lymphocyte development and t cell leukemias. *Cancer Cell*, 4(6):451–461, 2003.

[20] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209, 2002.

[21] Z. Su, H. Hong, H. Fang, L. Shi, R. Perkins, and W. Tong. Very important pool (vip) genes–an application for microarray-based molecular signatures. *BMC Bioinformatics*, 9 Suppl 9:S9, 2008.

[22] Y. Tang, Y.-Q. Zhang, Z. Huang, X. Hu, and Y. Zhao. Recursive fuzzy granulation for gene subsets extraction and cancer classification. *IEEE Trans Inf Technol Biomed*, 12(6):723–730, 2008.

[23] D. A. Thompson and R. J. Weigel. hag-2, the human homologue of the xenopus laevis cement gland gene xag-2, is coexpressed with estrogen receptor in breast cancer cell lines. *Biochem Biophys Res Commun*, 251(1):111–116, 1998.

[24] M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. M. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*, 347(25):1999–2009, 2002.

[25] Y. Yap, X. Zhang, M. T. Ling, X. Wang, Y. C. Wong, and A. Danchin. Classification between normal and tumor tissues based on the pair-wise gene expression ratio. *BMC Cancer*, 4:72, 2004.

[26] K. Y. Yeung, R. E. Bumgarner, and A. E. Raftery. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics*, 21(10):2394–2402, 2005.