Published in final edited form as:

Conf Proc IEEE Eng Med Biol Soc. 2008 ; 2008: 5660–5663. doi:10.1109/IEMBS.2008.4650498.

Improving Microarray Sample Size Using Bootstrap Data Combination

John H. Phan, Richard A. Moffitt, Andrea B. Barrett, and May D. Wang

Department of Biomedical Engineering, Georgia Institute of Technology

John H. Phan: jhphan@gatech.edu; Richard A. Moffitt: r.moffitt@gatech.edu; Andrea B. Barrett: andreabarrett@gatech.edu; May D. Wang: maywang@bme.gatech.edu

Abstract

Microarray technology has enabled us to simultaneously measure the expression of thousands of genes. Using this high-throughput technology, we can examine subtle genetic changes between biological samples and build predictive models for clinical applications. Although microarrays have dramatically increased the rate of data collection, sample size is still a major issue when selecting features. Previous methods show that combining multiple microarray datasets improves feature selection using simple methods such as fold change. We propose a wrapper-based gene selection technique that combines bootstrap estimated classification errors for individual genes across multiple datasets and reduces the contribution of datasets with high variance. We use the bootstrap because it is an unbiased estimator of classification error that is also effective for small sample data. Coupled with data combination across multiple datasets, we show that our meta-analytic approach improves the biological relevance of gene selection using prostate and renal cancer microarray data.

Introduction

Microarray gene expression profiling has become a popular tool for identifying biomarkers in diseases such as cancer. They allow us to examine the expression levels of thousands of genes in a single experiment. However, cancer may only alter a tiny portion of the human genome. Therefore, we need a powerful and effective feature selection scheme, in addition to a large sample size, to identify these potential biomarkers. While the number of gene expression datasets available to the scientific community is growing, the sample size of each dataset remains small compared to the number of features. As such, methods for combining multiple datasets have the potential for increasing the power of microarray data analysis by pooling information.

Combining datasets can be difficult when we use different microarray platforms or apply different probe normalization and summarization techniques. Even when we use the same chip hardware and software, the laboratory effect can, in some cases, be more significant than the choice of chip platform when assessing reproducibility [1]. Differences in reproducibility, sensitivity, and specificity between datasets from separate test sites can lead

to different sets of candidate biomarkers [2, 3]. In addition to all of these technical obstacles, the practical limitation of finding datasets which measure the same scientific question further hampers data combination. Thus, most current biomarker identification studies are limited to single, small-sample datasets.

A common goal in microarray analysis is the creation of predictive classifiers. The first step in creating a classifier is often feature selection, which involves systematically excluding a number of weakly-informative genes in order to increase the overall performance of the classifier. Methods for feature selection fall into two categories: filter methods and wrapper methods.

Filter methods are a two step process, beginning with individual scoring of each feature, followed by selection based on this scoring. At the end of the filtering procedure, we build a predictive classifier using a different method from the one used to score and select individual genes. Common filtering methods include fold change and T-test. However, the classification accuracy of biomarkers resulting from such methods is not necessarily high. Because of the inclusion of redundant information, resulting classifiers may become highly complex without significant gain in accuracy [4]. Furthermore, these methods are sensitive to small-sample data and depend on strict assumptions. Calculation of the T-statistic, for example, breaks down when the number of features included is larger than the sample size. Statistics such as mean and variance may be significantly biased when calculated from small sample data, leading to false conclusions of significance. The dependence of the T-test on data normality is also problematic, since this assumption is often not true for gene expression data [5].

For wrapper methods, the final classifier is intrinsic to the feature selection process. Instead of scoring genes independently, a wrapper method will assess groups of genes based on their synergistic performance, usually measured by estimating the error-rate of classification. Using classification error-rate as a selection criterion is appropriate when the aim is to design a discriminant rule [6]. Furthermore, error estimation techniques such as the bootstrap do not depend on assumptions of data normality. Studies have shown that various bootstrap and cross validation resampling methods are accurate estimators of predictive performance for small-sample data [7].

Several studies examine methods for combining multiple microarray datasets in order to improve sample size. These methods include large-scale data-mining, functional integration, ANOVA models, and effect size meta-analysis [8–11]. Of particular interest is the study by Wang et al., which combines the fold change of genes between classes from three microarray datasets. Their computed statistic also takes into account the different variances within datasets [11]. Choi et al. compute a combined z-statistic and find that this combined statistic identifies more potential biomarkers than those of single datasets [12]. None of these methods, however, explore data combination using wrapper-based gene selection methods that estimate classification error. Such methods would have both the benefits of accurate identification of predictive genes from small datasets as well as increased sample size by combining multiple datasets.

The meta-analytic method that we propose combines heterogeneous microarray datasets by combining bootstrap estimated classification errors for each gene. Our method, adapted from Wang et al., weights the combined classification error from each dataset by the inverse of its bootstrap variance [11]. This weighting reduces the overall contribution of datasets with large variance. It is easily extended to any number of datasets and has the potential to improve the biological and statistical relevance of candidate biomarkers.

Methods

Microarray Data

We use two groups of microarray datasets to test the bootstrap meta-analysis method. The first group includes samples from two renal cancer studies. Each study contains samples from clear cell (CC), chromophobe (CHR), and oncocytoma (ONC) renal cancer subtypes. We are interested in identifying genes that are differentially expressed between the CC and the combined group of both CHR and ONC subtypes. The smaller dataset from Schuetz et al. contains 13 CC, 4 CHR, and 3 ONC samples. The larger dataset from Jones et al. contains 32 CC, 6 CHR, and 12 ONC samples. Samples from both datasets are derived from Affymetrix microarrays and contain at least 8793 genes [13, 14]. The Jones dataset contains more genes, but we reduce the number of genes to a subset of 8793 common to both datasets.

The second group includes two prostate cancer datasets with 12625 genes on Affymetrix microarrays. Between the two datasets, there are a total of 113 tumor samples and 113 normal samples [15, 16]. Normal samples are extracted from tissue adjacent to prostate tumors. The smaller Singh dataset contains 52 tumor samples and 50 normal adjacent samples. The Chandran dataset contains 61 tumor samples and 63 normal adjacent samples.

Feature Ranking

We use a wrapper-based approach to rank genes by classification accuracy, estimated using a linear support vector machine (SVM) classifier and bootstrap resampling. The SVM classifier predicts the class of a test sample based on its gene expression by constructing a maximal margin discriminating hyperplane from training samples. We use a dataset-specific SVM cost parameter, C , of approximately $1/n$. Thus, the cost parameter is 0.05 for the Schuetz dataset and 0.02, 0.01, and 0.01 for the Jones, Singh, and Chandran datasets, respectively.

We define the m th estimate $E_{j,k,l,m}$ of the true error of classification: $E_{j,k,l}$, for each data source j , each gene k , and set of class labels l , with $l=0$ defining the correct labeling. m enumerates all possible partitions of the data into training and testing sets. To estimate $E_{j,k,l}$, we sample the space of m with a bootstrapping algorithm as follows: For a dataset with $n=1:N$ samples, the bootstrap algorithm randomly selects N samples with replacement [17]. The unique N^* samples selected by the bootstrap ($N^* < N$) are designated as the training set (Straining) for the SVM classifier. The remaining $N-N^*$ samples are the testing set (Stesting). Comparing the SVM output labels for these $N-N^*$ samples to the actual class labels (defined by l), we calculate the classification error with

$$E_{j,k,l,m} = \frac{\sum_{n \in S_{\text{testing}}} |L_{j,k,SV M,m,n} - L_{j,k,l,m,n}|}{N - N^*} \quad (1)$$

where $L_{j,k,l,m,n}$ is the class label of the n^{th} sample. Both L 's are either 0 or 1 depending on the class. We repeat this bootstrap procedure for 100 iterations and compute the mean (E) and variance ($\sigma^2_{j,k,l,\bullet}$) of the $j,k,l,\bullet$ classification error across the $B=100$ bootstrapped samplings.

$$E_{j,k,l,\bullet} = \frac{1}{B} \sum_{m=1}^B E_{j,k,l,m} \quad (2)$$

$$\sigma^2_{j,k,l,\bullet} = \frac{1}{B-1} \sum_{m=1}^B (E_{j,k,l,\bullet} - E_{j,k,l,m})^2 \quad (3)$$

In a typical wrapper-based analysis, we can use these $E_{j,k,l,\bullet}$ for feature ranking and $\sigma^2_{j,k,l,\bullet}$ as measures of confidence in these errors.

Data Combination

Differences in sample size prevent us from directly comparing wrapper-based feature selection between two or more datasets. The distribution of $E_{j,k,l,\bullet}$ over all genes k often depends on the balance of samples between classes within a dataset. Furthermore, the total number of samples in a dataset affects $\sigma^2_{j,k,l,\bullet}$. Thus, we need to empirically estimate a null distribution for $E_{j,k,l,\bullet}$ for each data source j , and gene k in order to generate a universally comparable (and combinable) significance score. A common method for estimating the null distribution $P(E^0 = x)$ is to randomly permute the j,k dataset class labels for each gene, in order to remove any information that the true class labels hold. We can estimate the cumulative distribution function of this null distribution by sampling space of l , T times:

$$cdf(E_{j,k}^0) \approx \frac{1}{T} \int_{-\infty}^x \sum_l \delta(\tau - E_{j,k,l,\bullet}) d\tau \quad (4)$$

Ideally, we should sample the space of l extensively for each gene k to estimate a gene-specific null distribution. In practice, however, we assume that the null distributions are similar for each gene, and thus we only sample l once for each of the T genes. This results in one aggregate null distribution for a data source, $P(E^0 = x)$, described by:

$$cdf(E_j^0) \approx \frac{1}{T} \int_{-\infty}^x \sum_{l,k} \delta(\tau - E_{j,k,l,\bullet}) d\tau \quad (5)$$

with corresponding sample mean and variance of E_j^0 and $\hat{s}^2$ estimated from the set of T observations from $\{E_{j,k,l,\bullet}\}$. We can now normalize any bootstrap error $E_{j,k,l,\bullet}$ against its null distribution such that its distribution is approximately normal with zero mean and unit variance:

$$Z_{j,k,l,\bullet} = \frac{1}{\hat{s}_j} (E_{j,k,l,\bullet} - \bar{E}_j^0). \quad (6)$$

The variance of a bootstrap measurement $\sigma_{j,k,l,\bullet}^2$ estimates the reliability of that measurement. When combining bootstrap measurements from multiple data sources, we should favor measurements with low variance. Therefore, the combined classification score is a weighted combination of normalized bootstrap errors from individual datasets. We adapt this combination formula from Wang et al. [11]:

$$E_{\text{comb},k,l,\bullet} = \left(\sum_j Z_{j,k,l,\bullet} / \sigma_{j,k,l,\bullet}^2 \right) / \left(\sum_j 1 / \sigma_{j,k,l,\bullet}^2 \right) \quad (7)$$

A null distribution for combined classification score can also be computed similarly to (5):

$$cdf(E_{\text{comb}}^0) \approx \frac{1}{T} \int_{-\infty}^x \sum_{l,k} \delta(\tau - E_{\text{comb},k,l,\bullet}) d\tau \quad (8)$$

We can identify differentially expressed genes using individual datasets by ranking values of $E_{j,k,0,\bullet}$ for all genes k or by estimating p-values using the estimated null values of distribution $P(E_j^0 = x)$. Similarly, we can identify significant genes from combined datasets by ranking values of $E_{\text{comb},k,0,\bullet}$ for all genes k , or by estimating p-values using the estimated null distribution $P(E_{\text{comb}}^0 = x)$. The p-value for a gene k from an individual dataset j is taken as the value of the estimated $cdf(E_{j,k}^0)$ at $E_{j,k,0,\bullet}$.

In addition to the bootstrap error method described above, we combine data sources on the basis of fold change and T-test as described by Wang et al. [11].

Feature Selection

We decide that a gene is significantly differentially expressed based on p-value. For assessing overlap between methods, we use a consistent p-value threshold of 0.01. For both T-test and fold-change methods, we assign p-values to raw scores based on a two-tailed normal distribution with mean and standard deviation statistics estimated from the observed T and fold change distributions, respectively. For the bootstrap error estimator, we estimate p-values based on the empirically-determined null-distribution described in the previous section.

Validating the Relevance of Gene Ranking

We assess the effect of data combination by comparing the result of feature selection in each individual dataset to that of the combined dataset. It is often difficult to compare feature selection between methods or similar datasets due to the large number of features and the subjectivity of feature interpretation. Because of our limited understanding of biological mechanisms and the noise inherent in microarray technology, we can often only verify the validity of a selected gene by independent assays such as qRT-PCR [18]. However, in order to avoid the costly and time-consuming validation of genes selected in this study, we use a method introduced by Mukherjee et al. to compute the probability of successfully selecting differentially expressed genes [19]. This method is based on a simple scenario that compares the ranks of two genes, one of which is known to be differentially expressed. Mukherjee et al. computes the probability that a given ranking algorithm correctly ranks the differentially expressed gene by ranking it more favorably. They extend this method to multiple genes and define the random variables T_0 and T_1 , which represent the ranks of null and differentially expressed genes, respectively. Assuming that a higher rank number corresponds to a more differentially expressed gene, they explicitly compute the probability $P(T_1 < T_0)$ which they call the Binary Selection Accuracy (BSA) [19]. This probability is equivalent to the area under the curve (AUC) of a receiver-operator characteristic (ROC) curve produced by classifying genes into either differentially expressed or null groups using a rank threshold [20].

Instead of comparing different ranking algorithms, we compare the ranks produced from individual datasets to those of the combined dataset. In order to compute the BSA, we need a set of reference genes that are known to be differentially expressed. For both the renal cancer and prostate cancer datasets, we identify genes from literature that have also been validated with qRT-PCR, the most common method for validating differentially expressed genes [21]. When computing the ROC curves, we weight each gene by one minus its p-value. Thus, statistically significant genes have a larger contribution to the resulting AUC.

Results and Discussion

Bootstrap Meta Analysis

Data suggests that the no-information null distribution of the bootstrap is approximately normal, but the actual distribution of errors trained on correct class labels is skewed towards low error (figure 1). This is expected, as the magnitude of skew should be a function of the separability of the classes under investigation and the information content of the genes present on the microarray.

For the prostate cancer datasets, this skew is less prevalent, as the empirical distribution more resembles the null distribution. Small deviations still exist at the low error side of the distribution, corresponding to potential differentially expressed genes (figure 2).

Validating the Relevance of Gene Ranking

We identify several validated genes for both the renal and prostate cancer datasets and use these genes as a reference for assessing the effect of data combination. We restrict these

reference genes to those validated with qRT-PCR. This restriction improves the reliability of the reference genes since qRT-PCR is a common and established method for validating microarray assays [21].

In renal cancer, carbonic anhydrase IX (CA9) and low density lipoprotein-related protein 2 (LRP2) are upregulated in CC compared to the ONC and CHR renal cancer subtypes [13, 22]. Additionally chloride channel Kb (CLCNKB), defensin beta 1, and parvalbumin (PVALB) are upregulated in ONC and CHR compared to CC (table 1) [13, 22]. Using these reference genes, we compute the BSA (AUC) for each of the individual and combined renal datasets. We compare the results of the bootstrap meta-analysis to the original fold change combination method by Wang et al. as well as to a similar T-test method [11]. The ROC curves (figure 3) for the combined data are more similar to the Jones dataset when using T-test or bootstrap ranking. ROC curves for the fold change method are similar between both individual datasets and the combined dataset. BSA values (figure 5) for the combined dataset are higher than both of the individual datasets when ranking genes using fold change and bootstrap methods, but not the T-test method. Fold change tends to perform well for both individual datasets in terms of favorably ranking the reference genes. Thus, we also expect fold change to perform well for the combined data.

Compared to the Jones data, the small sample size of the Schuetz data seems to reduce the efficiency of ranking using the T-test and bootstrap methods. Small sample size generally corresponds to higher measurement variance. Therefore we expect the data combination method to reduce the overall contribution of the Schuetz data. However, the BSA for the T-test combination method is higher than that of the individual Schuetz dataset and lower than that of the individual Jones data. This suggests that, for the T-test method, the contribution of the Schuetz data to the combined data ranking has not been properly weighted to account for its higher variance. The results of bootstrap ranking using individual datasets produces similar BSA values compared to the T-test. However, our bootstrap method properly weights the Schuetz data to reduce its overall contribution in the combined data. Thus, the BSA of the combined data ranking is higher than both individual datasets (figure 5).

For the prostate cancer data, we select six validated genes from literature, four of which are over-expressed in tumor tissue while two are under-expressed (table 2) [23]. We compute the ROC curves and BSA scores for the prostate cancer datasets using these reference genes (figure 4, figure 6). Compared to the renal cancer data, detection of these reference genes via ranking is much less efficient, resulting relatively low BSA scores. This decrease in efficiency may be due to biological heterogeneity among the three datasets from Singh, Chandran, and Ernst. Despite the low BSA scores, the genes identified from the Ernst dataset have been qRT-PCR validated and should serve as a point of reference for assessing the quality of the Singh and Chandran datasets. Examining the ROC curves and the corresponding BSA values, for both the T-test and fold change data combination methods, we see that combining data improves gene detection compared to the individual Chandran dataset. However, for these ranking methods, the combined dataset still performs worse than the individual Singh dataset. The bootstrap method performs equally well on both the individual Singh and Chandran datasets. Furthermore, the combined method improves the overall performance of gene detection compared to both individual datasets.

Overlap Among Methods

Both renal cancer datasets (Schuetz and Jones) exhibit a much larger trend of significantly low bootstrap errors than is found in either of the two prostate datasets. The number of significant genes found via T-test and fold change, however remains moderate in the renal data. This is most likely because the bootstrap error is determined to be significant based on a separately generated null distribution, while it is assumed that, for the fold change and T-test scoring, the null distributions are identical to the observed distribution. We make this assumption to maintain consistency and comparability with prior work. However, this assumption may be faulty and could be improved. Another way to think about this phenomenon is to consider a random variable exhibiting a nearly Gaussian distribution, such as fold change and T-statistic (data not shown). Normalizing the variable with its own population mean and standard deviation should always result in 1% of all observations assayed to be significant at the $p < 0.01$ level.

The difference between the number of significant genes identified via bootstrap in each of the renal or prostate datasets (figure 7) may best be attributed to both the sample size and the intrinsic difficulty of the classification problem, although we lack a biological basis for this conclusion. The prostate dataset may be a more difficult problem to classify since the normal samples are derived from tissue adjacent to the prostate tumor, which are also highly heterogeneous [24].

Figure 8 illustrates a case where one dataset is dominant during combination due to a disparity in variances. In this example, the Schuetz data has lower variance for both the T-test and fold change (top), but not for the bootstrap error (bottom) when compared to the Jones dataset. As a result, almost all of the genes identified from the Schuetz data alone are also identified as significant after dataset combination for T-test and fold change, with the trend reversing and favoring Jones for bootstrap. Because the Schuetz dataset has a smaller pool of samples, one would expect the confidence of classification to be lower for repeated measurements, and this is indeed the case for the bootstrap error, which is computed 100 times. We therefore conclude that estimating the confidence of a T-test or fold change value by calculating the variance of the statistic across all genes is a poor estimate of the true confidence of measurement. For future experiments, we recommend re-computing each statistic using a cross-validation procedure to estimate variance on a per-gene basis. Hopefully, this would lead to a better estimate of variance, and therefore, a more fruitful dataset combination.

Interpretation of Selected Genes

Data combination using bootstrap meta-analysis improves the ranks of two of the prostate cancer reference genes. Both of these genes, tripartite motif containing 29 (TRIM29) and pyrroline-5-carboxylate reductase 1 (PYCR1), were previously validated by Ernst et al [23]. For TRIM29, individual dataset p-values of 0.07 and 0.05 are reduced to 0.02 in the combined data. Likewise, for PYCR1, individual dataset p-values of 0.19 and 0.16 are reduced to 0.10 in the combined data. The ranks for the other four reference genes are not improved in the combined data compared to both individual datasets. However, the individual dataset p-values of these four genes are larger than those of TRIM29 and PYCR1.

Although these genes have been validated with qRT-PCR, their reliability for these particular datasets is questionable and provides further evidence of the analytical difficulties due to prostate tissue heterogeneity.

In addition to the reference genes, we identify several other prostate cancer-related genes using the bootstrap-combined data. Interestingly, brain-derived neurotrophic factor (BDNF), has been linked to prostate cancer [25]. We easily identify BDNF in the Singh dataset (rank=3), but not in the Chandran data (rank=420). The combined data ranks the gene at 17 (lower rank is better). Similarly, the combined data improves the recognition of the metastasis-associated protein 1 (MTA1, combined rank=18, Singh rank=16, Chandran rank=360) and chemokine receptor 2 (CCR2, combined rank=35, Singh rank=5, Chandran rank=580), again by discounting the contribution of the Chandran data [26, 27]. The combined data also favorably ranks some relevant genes that are not ranked favorably in individual datasets. For example, ITIH3 (Singh rank=25, Chandran rank=326, combined rank=6) and CHD5 (Singh rank=165, Chandran rank=95, combined rank=7) have both been linked to human cancer [28, 29].

Data combination also improves the ranks of two of the renal cancer reference genes, CLCNKB and PVALB. The rank of DEFB1 is only improved compared to the individual Schuetz dataset (Schuetz rank=6158, Jones rank=552, combined data=1011). However, the p-value of DEFB1 in the Schuetz dataset is large. Likewise, the combined data rank of CA9 is improved compared to the Jones dataset, but not the Schuetz dataset (Schuetz rank=1900, Jones rank=2808, combined data=2344). Some genes implicated in renal cancer are favorably ranked in the combined data, but not in either of the individual datasets. For example, CXCR4 has a combined data rank of 28, but individual ranks of 113 and 188 in the Schuetz and Jones dataset, respectively [30]. Many other examples of improved biological relevance exist in both prostate and renal cancer datasets. However, a more rigorous biological analysis is beyond the scope of this paper.

Conclusion

Until microarray technology becomes standardized, we must develop statistical methods to handle small sample data. Standardization protocols are currently not well defined. However, the ideal scenario for standardization should allow us to compare quantitative measurements across different platforms or to expect all clinical measurements to be acquired with the same technology. Regardless of standardization, microarrays are still subject to technological variance when experiments are performed at different times or locations. Our proposed method is a possible solution to reduce technical bias by computing differential gene expression scores on distinct microarray groups, then combining these scores across multiple groups of microarrays. The bootstrap accurately estimates classification errors for genes of individual small datasets while the combination method favors genes whose scores have lower individual dataset variances. The results of this method applied to prostate and renal cancer datasets indicate that bootstrap meta-analysis improves the biological relevance of gene selection by increasing data sample size.

Acknowledgments

This research has been supported by grants from National Institutes of Health (R01CA108468, P20GM072069, U54CA119338), National Science Foundation (Fellowship to RAM), Microsoft Research Funding, and Georgia Cancer Coalition (Distinguished Cancer Scholar Award to MDW).

References

1. Irizarry RA, et al. Multiple-laboratory comparison of microarray platforms. *Nature Methods*. 2005; 2(5):345–350. [PubMed: 15846361]
2. Patterson T, et al. Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. *Nature Biotechnology*. 2006; 24(9):1140–1150.
3. Vo, TM., et al. Reproducibility of Differential Gene Detection Across Multiple Microarray Studies; Proc. 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS; 2007.
4. Xiong M, et al. Biomarker Identification by Feature Wrappers. *Genome Research*. 2001; 11:1878–1887. [PubMed: 11691853]
5. Troyanskaya O, et al. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*. 2002; 18(11):1454–1461. [PubMed: 12424116]
6. Xiao Y, et al. Multivariate search for differentially expressed gene combinations. *BMC Bioinformatics*. 2004; 5(164)
7. Braga-Neto U, Dougherty E. Is cross-validation valid for small-sample microarray classification? *Bioinformatics*. 2004; 20:374–380. [PubMed: 14960464]
8. Huttenhower C, et al. A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics*. 2006; 22(23):2890–2897. [PubMed: 17005538]
9. Park T, et al. Combining multiple microarrays in the presence of controlling variables. *Bioinformatics*. 2006; 22(14):1682–1689. [PubMed: 16705015]
10. Yoon S, et al. Large scale data mining approach for gene-specific standardization of microarray gene expression data. *Bioinformatics*. 2006; 22(23):2898–2904. [PubMed: 17032674]
11. Wang J, et al. Differences in gene expression between B- cell chronic lymphocytic leukemia and normal B cells: a meta-analysis of three microarray studies. *Bioinformatics*. 2004; 20(17):3166–3178. [PubMed: 15231529]
12. Choi J, et al. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*. 2003; 19:i84–i90. [PubMed: 12855442]
13. Schuetz AN, et al. Molecular Classification of Renal Tumors by Gene Expression Profiling. *J Mol Diagn*. 2005; 7(2):206–218. [PubMed: 15858144]
14. Jones J, et al. Gene Signatures of Progression and Metastasis in Renal Cell Cancer. *Clin Cancer Res*. 2005; 11(16):5730–5739. [PubMed: 16115910]
15. Chandran U, et al. Differences in gene expression in prostate cancer, normal appearing prostate tissue adjacent to cancer and prostate tissue from cancer free organ donors. *BMC Cancer*. 2005; 5(45)
16. Singh D, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*. 2002; 1(2):203–209. [PubMed: 12086878]
17. Efron B. Estimating the error rate of a prediction rule: some improvements on cross-validation. *J of the Am Stat Assoc*. 1983; 78:316–331.
18. Chuaqui R, et al. Post-analysis follow-up and validation of microarray experiments. *Nature Genetics*. 2002; 32:509–514. [PubMed: 12454646]
19. Mukherjee S, Roberts S. A theoretical analysis of the selection of differentially expressed genes. *J Bioinformatics Comput Biol*. 2005; 3:627–643.
20. Bradley A. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*. 1997; 30(7):1145–1159.
21. Morey J, et al. Microarray validation: factors influencing correlation between oligonucleotide microarrays and real- time PCR. *Biol. Proced. Online*. 2006; 8(1):175–193. [PubMed: 17242735]

22. Chen Y, et al. Messenger RNA Expression Ratios among Four Genes Predict Subtypes of Renal Cell Carcinoma and Distinguish Oncocytoma from Carcinoma. *Clin Cancer Res.* 2005; 11:6558–6566. [PubMed: 16166433]
23. Ernst T, et al. Decrease and Gain of Gene Expression Are Equally Discriminatory Markers for Prostate Carcinoma: A Gene Expression Analysis on Total and Microdissected Prostate Tissue. *American Journal of Pathology.* 2002; 160(6):2169–2180. [PubMed: 12057920]
24. Wang N, et al. Evaluation of tumor heterogeneity of prostate carcinoma by flow- and image DNA cytometry and histopathological grading. *Analytical Cellular Pathology.* 2000; 20(1):49–62. [PubMed: 11007438]
25. Bronzetti E, et al. A possible role of BDNF in prostate cancer detection. *Oncol Rep.* 2008; 19(4): 969–974. [PubMed: 18357383]
26. Hofer M, et al. The Role of Metastasis-Associated Protein 1 in Prostate Cancer Progression. *Cancer Research.* 2004; 64:825–829. [PubMed: 14871807]
27. Lu Y, et al. CCR2 expression correlates with prostate cancer progression. *J Cell Biochem.* 2007; 101(3):676–685. [PubMed: 17216598]
28. Hamm A, et al. Frequent expression loss of Inter-alpha- trypsin inhibitor heavy chain (ITI_H) genes in multiple human solid tumors: A systematic expression analysis. *BMC Cancer.* 2008; 8(25)
29. Baqchi A, et al. CHD5 is a tumor suppressor at human 1p36. *Cell.* 2007; 128(3):459–475. [PubMed: 17289567]
30. Schrader A, et al. CXCR4/CXCL12 expression and signaling in kidney cancer. *Br J Cancer.* 2002; 86(8):1250–1256. [PubMed: 11953881]

Figure 1.

Distributions of normalized classification errors for the renal cancer datasets. Solid lines represent individual or combined distributions while dashed lines represent null distributions. Both the Schuetz (top left) and Jones (top right) datasets deviate significantly from the null distributions, indicating a large number of differentially expressed genes. This deviation is reflected in the combined data (bottom).

Figure 2.

Distributions of normalized classification errors for the prostate cancer datasets (Singh, top left; Chandran, top right; combined, bottom). Compared to the renal cancer datasets, these datasets do not deviate significantly from the null distribution.

Figure 3.

ROC curves for detecting validated reference genes for renal cancer. Red lines are combined data and dashed lines are individual datasets. For all datasets, fold change (top right) tends to detect reference genes efficiently compared to t-test (top left) and bootstrap (bottom). Combining data using fold change and bootstrap slightly improves detection efficiency. This corresponds to an increase in area under the ROC curve.

Figure 4.

ROC curves for detecting validated reference genes for prostate cancer. Red lines are combined data and dashed lines are individual datasets. Combined data does not improve efficiency of reference gene detection when using the t-test (top left) or fold change (top right) methods. The bootstrap method (bottom) slightly increases detection performance of the reference genes for combined data.

Figure 5.

BSA (AUCs) of individual and combined renal cancer datasets for detecting reference genes.

The red bars are AUCs of the combined data. T, FC, and BS correspond to t-test, fold change, and bootstrap, respectively. S, J, and C correspond to Schuetz, Jones, bootstrap (right bars), the relevance of ranking for the combined data is at least as good, if not better, than both individual datasets.

Figure 6.

BSA (AUCs) of individual and combined prostate cancer datasets for detecting reference genes. The red bars are AUCs of the combined data. T, FC, and BS correspond to t-test, fold change, and bootstrap, respectively. S, Ch, and C correspond to Singh, Chandran, and combined data, respectively. The bootstrap combination method (right bars) outperforms both the t-test and fold change methods.

Figure 7.

Venn Diagrams showing overlap of genes selected by T-test, fold change, and bootstrap error ($p < 0.01$). The left panel shows results for data combination from the renal cancer group, and the right panel shows results from the prostate cancer group.

Figure 8.

Venn Diagrams showing overlap of selected genes selected by the individual datasets, and then by the combined meta-analysis for the T-test, Fold Change, and Bootstrap.

Table 1

Validated renal cancer reference genes.

Gene	Regulation	Source
GA9	Up in CC	Chen 2005
CLCNKB	Up in ONC/CHR	Chen 2005
DEFB1	Up in ONC/CHR	Schuetz 2004
LRP2	Up in CC	Schuetz 2004
PVALB	Up in ONC/CHR	Chen 2005

Table 2

Validated prostate cancer reference genes.

Gene	Regulation	Source
AMACR	Up in Tumor	Ernst 2002
PLA2G7	Up in Tumor	Ernst 2002
HPN	Up in Tumor	Ernst 2002
PYCR1	Up in Tumor	Ernst 2002
SCGB1A1	Down in Tumor	Ernst 2002
TRIM29 (ATDC)	Down in Tumor	Ernst 2002