Benefits of Using Paired Controls for Analyzing Gene Expression of Prostate Cancer

Scott Haney, Moshe Kam, and Leonid Hrebien

Abstract— **The genetics of prostate cancer are highly variable and not well understood. No consistent pattern of mutations across prostate cancer samples has yet been found. Due to this inherent heterogeneity it is natural to wonder whether or not using paired controls in gene expression studies might be useful. Although other studies have previously analyzed the use of paired controls for the expression of a handful of genes at a time no study has yet been performed to assess the benefits of using paired controls rather than independent controls on a large scale. By using a prostate cancer microarray data set that consisted of 58 pairs of paired cancer and control samples as well as 18 independent controls we found that searches for differentially expressed genes and for upregulated genes were significantly enhanced by using paired controls instead of independent controls.**

I. INTRODUCTION

In general the molecular genetics of prostate cancer is highly variable and not well understood [1], [2], [3]. Unlike several other common cancers that have been shown to contain common mutations such as the APC mutation in colorectal cancer, the VHL mutation in renal cancer, and the BRCA1 and BRCA2 mutations in breast cancer [4] no consistent pattern of mutations has been found to occur across a large number of prostate cancer patients [5]. This heterogeneity has also been shown to exist in gene expression data. For instance, a microarray study by Luo et al. found that the profiles of highly aggressive tumor samples were noticeably distinct from the profiles of organ-confined tumors [6]. Although the sample sizes were small (only three highly aggressive tumor samples) it is interesting to note that amongst prostate cancer samples different tumors can exhibit remarkably distinct gene expression patterns from one another. Further evidence for this heterogeneous biological nature can be found in a study of 30 rapid autopsy samples from men who died of androgen-independent prostate cancer. In this study a customized cDNA microarray detected large variations in genetic expression across the 30 samples. For instance, androgen receptor (AR) expression levels were less than 10% for 100 of the 265 tumor samples and greater than 50% for 83 of the 265 tumor samples [2].

Due to the heterogeneous nature of prostate cancer genetics it is natural to consider the use of paired controls. Pairing would reduce variability and could possibly simplify the analysis of gene expression in prostate cancer. Previous work using paired controls for the analysis of gene expression in prostate cancer has been run on small subsets of genes by several groups [7], [8]. However, these groups did not quantitatively investigate the importance of using paired controls in their studies. No test was performed to determine the benefit of using paired samples. Although it might seem obvious that paired samples should provide more information in a gene expression study this fact should be verified quantitatively. Also, it is certainly not clear how much improvement is made when paired controls are used instead of independent controls.

In order to assess the impact of paired controls on a large scale we will utilize a data set from the gene expression omnibus (GEO) which contains 58 pairs of paired prostate cancer and control samples as well as 18 independent controls. We chose to focus on how pairing impacts two simple factors which were differential expression (change in mean) and relative expression (serial increases or decreases in expression across patients relative to pre-tumor expression levels). If these two factors are significantly influenced by pairing then it is likely that more complicated methods such as neural networks, random graphs, clustering etc. would also be significantly influenced by pairing. Also, differential expression and relative expression have very simple and obvious biological interpretations. For these reasons analyzing the impact of pairing on differential expression and relative expression is an important task that can quickly reveal significant benefits from pairing.

II. DATA DESCRIPTION

Our data set was taken from the gene expression omnibus (GEO) website (www.ncbi.nlm.nih.gov/geo/). This data set contained 171 CEL files for independent samples and was labeled GDS2545. Of the samples, there were 58 paired samples and 18 independent control samples. Each of the 58 pairs consisted of a sample from the primary tumor and a sample from adjacent normal prostate tissue in the same patient. Each of the 134 samples (58 pairs and 18 independent controls) used the HG_U95A affymetrix microarray platform and contained expression values for 12,625 tags. The original 134 CEL files were background corrected, summarized, and normalized all together using the justRMA() function available for the R programming language through the affy library from bioconductor (www.bioconductor.org).

The biological distribution of the samples is given in Tables 1 and 2. It is important to note that most of the tumors are stage T2b and stage T3a. In stage T2b the tumor invades both lobes of the prostate and in stage T3a the tumor

S. Haney is a Graduate Student, Electrical Engineering, Drexel University, Philadelphia, PA 19104, USA swh23@drexel.edu

M. Kam and L. Hrebien are with the Faculty of Electrical Engineering, Drexel University, Philadelphia, PA 19104, USA kam@minerva.ece.drexel.edu, lhrebien@coe.drexel.edu

TABLE I
TUMOR STAGE DISTRIBUTION

Tumor Stage	Number of Tumors
T2a	2
T2b	19
T3a	24
T3b	11
T4	1
T4a	1

TABLE II
GLEASON SUM DISTRIBUTION

Gleason Sum	Number of Tumors
5	2
6	12
7	25
8	7
9	12

unilaterally extends the prostate capsule. For both of these stages the tumor has not extended beyond the prostate into any neighboring anatomy or organs. The Gleason sum for the tumors falls in the range of 5 to 9. Depending upon the observed structure of the tumor a Gleason score of anywhere between 1 and 5 can be assigned. Since multiple tumor patterns are often seen the Gleason scores for the two most predominant patterns are summed together to give the Gleason sum [9].

III. DATA ANALYSIS

Both a paired t-test and a two sample t-test were run on the 58 pairs of cancer and control samples. It was found that 3,544 genes were selected at the α = .05 level using the paired t-test and 2,971 genes were selected at the α = .05 level using the two sample t-test. The possibility that the difference in the number of genes between these two lists was due solely to chance was evaluated. Under the null hypothesis that the difference is solely due to chance the probability of observing a difference of 3,544 - 2,971 = 573 genes is given by the probability that x false positives were observed in the paired t-test and $\geq$ (x + 573) false positives were observed in the two sample t-test where x ranges from 0 to (12,625 - 573). This probability was approximately 2.8×10^{-77} which shows that the difference is unlikely to be due to chance alone. Of the 3,544 genes selected by the paired t-test 602 of these were not chosen using the two sample t-test.

A further test for differential expression was run using a paired t-test on the first 18 pairs of cancer and normal samples and a two sample t-test on the first 18 cancer samples and the 18 independent normal samples. It was found that the paired t-test selected 864 genes at the α = .05 level and the two sample t-test selected 2,620 genes at the α = .05 level. The probability that the difference in the number of genes between the two lists was due to chance was negligible. Of the 864 genes selected by the paired t-test 449 of these were not chosen by the two sample t-test.

The affect of relative expression on the 58 pairs of cancer and control samples was tested using a paired t-test that was run on 1,000 random reorderings of the pairings. Only upregulation in cancer samples was considered (a pair is upregulated in cancer if the largest expression in the pair is the expression from the cancer tissue). A gene was selected as upregulated if the probability that the number of upregulated genes would be generated by chance no more than 5% of the time. Under the null hypothesis of no effect the probability of a gene being upregulated should be 50%. This leads to a gene being taken as upregulated if at least 36 pairs are upregulated which corresponds to a significance level of α = .0435 (since the distribution is discrete it was impossible to get α = .05 exactly). Using these criteria there were found to be 2,146 upregulated genes in the original pairing and 998 out of the 1,000 reorderings had fewer than 2,146 upregulated genes. The average number of upregulated genes over the 1,000 reorderings was 1,177 which shows that on average the random reorderings selected 969 fewer upregulated genes than the original pairing.

IV. DISCUSSION

Changes in differential expression were evaluated using the t-test. Our first test analyzed the difference between using the 58 paired samples in the obvious manner (paired t-test) and using the 58 pairs of paired cancer and control samples as if they were independent of each other (two sample t-test). It was found that significantly more genes were selected using the paired t-test than the two sample t-test. More importantly, of the genes selected by the paired t-test 602 were unique to that list. Since the difference in the size of the lists was 573 we see that the paired t-test chose all but 29 (1% of the total results) of the same genes as the two sample t-test (plus a few extra). This shows that the paired t-test not only effectively duplicated the results of the two sample t-test but also added extra information (602 genes worth). The use of the pairing was definitely significant in this case.

Although the previous test seems to be fairly conclusive, there is an important issue that was not addressed. We assumed that by treating the 58 control samples from the 58 pairs as independent that they really were independent. However, we know that the 58 control samples were really dependent with the 58 cancer samples so it is possible that this could have affected the results. To test this we need to use the 18 truly independent control samples from the data set. Since there were only 18 independent controls we only used the first 18 pairs of samples from the 58 pairs. This reduction was performed because differences in sample sizes cause differences in statistical power. It would be nearly impossible to accurately determine the effect that a difference in power could have on the number of selected genes since it is unknown how many genes are differentially expressed at each fold change value. The results showed that significantly more genes were chosen by the two sample t-test. This is

in stark contrast to the previous test where the paired t-test chose more genes. Also, it is important to notice that approximately half of the genes chosen by the paired t-test (449 out of 864 genes) were not also found in the list of genes chosen by the two sample t-test. This implies that the two tests are not well correlated. In other words they are choosing strikingly different subsets of genes.

These two tests show that the use of pairing is very important when determining differential expression. When a pairing exists and is not used it was seen that a large number of results (602 genes in the case above) are thrown away. Also, when independent controls were used instead of the paired controls the outcome of the analysis was radically different. From these results it is apparent that for primary prostate cancer the use of paired controls instead of independent controls can make a large difference in the final outcome of the analysis. This shows that heterogeneity in primary prostate cancer is significant enough to warrant the extra effort needed to obtain paired controls when testing differential expression.

For relative expression we must test whether or not the pairing is important for determining whether a gene is upregulated. Our null hypothesis is that the pairing is not important and under this hypothesis the calculation of relative expression can not be affected by changing the pairing of the samples. To test this hypothesis we calculated the paired t-test on 1,000 random reorderings of the pairings for the 58 pairs of paired cancer and control samples. It was found that in 998 of the 1,000 reorderings the number of selected genes was less than the number of genes selected from the original pairing. This shows that by disturbing the pairing we have lost information (on average 969 genes or 45%) and this implies that the pairing was important.

Further tests were not run to test the impact of independent controls on relative expression because simply reordering the pairings provides rather convincing evidence that the pairings can not be done away with. More comparisons could be run to see whether or not possible pairings using the independent controls could be reproduced using the paired controls. This would in effect be testing whether or not the results from the paired data add any new information to any correlations between values of cancer samples and independent controls. The obvious impact of the pairings on detecting upregulated genes in the cancer samples shows that the pairings are highly significant. This implies that the impact of pairing on correlation data will be substantial. This should come as no surprise since paired data would be expected to provide more information on correlations. What is somewhat surprising is the size of the improvement.

V. CONCLUSION

Paired controls provide obvious benefits to studies on both differential expression and relative expression. By treating paired controls as if they were independent it was found that the paired controls led to the discovery of 602 more genes. Using only the independent controls it was found that over half of the genes selected by the paired t-test were not in the list of genes selected using the independent controls. This suggests that the results from using paired controls are radically different than the results using independent controls. With regards to relative expression, it was found that in 998 out of 1,000 reorderings of the pairings information was lost. On average the number of genes lost was 969 which was 45% of the number of genes selected when the original pairing was used. This shows that the pairings are indeed important and significant. The large effect that the pairing produces on the simple t-test shows how much information can be gained by using paired controls rather than independent controls. Since differential expression and relative expression are so affected by the use of paired controls we find it likely that almost any data analysis method will benefit from using paired controls when assessing gene expression data in prostate cancer.

VI. FUTURE WORK

Although the impact of pairing seems obvious the actual quantitative effect should be analyzed on more complicated methods such as neural networks. Also, if possible, more data sets should be found that include both paired controls and independent controls to test for the effects of pairing on other diseases. Due to the fact that pairing has such a significant impact on simple factors it is likely that it also has a large impact on more complicated ones. It seems likely that the use of paired controls may be crucial in many cases.

REFERENCES

[1] A. M. DeMarzo, W. G. Nelson, W. B. Isaacs, and J. I. Epstein, "Pathological and molecular aspects of prostate cancer," *Lancet*, vol. 361, pp. 955 – 964, 2003.

[2] R. B. Shah, R. Mehra, A. M. Chinnaiyan, R. Shen, D. Ghosh, M. Zhou, G. R. MacVicar, S. Varambally, J. Harwood, T. A. Bismar, R. Kim, M. A. Rubin, and K. J. Pienta, "Androgen-independent prostate cancer is a heterogeneous group of diseases: lessons from a rapid autopsy program," *Cancer Res.*, vol. 64, pp. 9209 – 9216, Dec. 2004.

[3] T. Visakorpi, "The molecular genetics of prostate cancer," *Urology*, vol. 62, pp. 3 – 10, 2003.

[4] L. W. K. Chung, W. B. Isaacs, and J. W. Simons, Eds., *Prostate Cancer: Biology, Genetics, and the New Therapeutics*, 2nd ed. Totowa, N.J.: Humana Press, 2007.

[5] A. Meeker, "Telomeres and telomerase in prostatic intraepithelial neoplasia and prostate cancer biology," *Urol. Oncol.*, vol. 24, pp. 122 – 130, 2006.

[6] J. Luo, D. J. Duggan, Y. Chen, J. Sauvageot, C. M. Ewing, M. L. Bittner, J. M. Trent, and W. B. Isaacs, "Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expressoion profiling," *Cancer Res.*, vol. 61, pp. 4683 – 4688, 2001.

[7] A. Latil, I. Bièche, L. Chêne, I. Laurendeau, P. Berthon, O. Cussenot, and M. Vidaud, "Gene expression profiling in clinically localized prostate cancer: a four-gene expression model predicts clinical behavior," *Clin. Cancer Res.*, vol. 9, pp. 5477 – 5485, Nov. 2003.

[8] J. Edwards, N. S. Krishna, K. M. Grigor, and J. M. S. Bartlett, "Androgen receptor gene amplification and protein expression in hormone refractory prostate cancer," *Br. J. Cancer*, vol. 89, pp. 552 – 556, Aug. 2003.

[9] R. S. Kirby, T. J. Christmas, and M. K. Brawer, *Prostate Cancer*. England: Mosby, 2001.