Bioinformatics Advance Access published December 19, 2008

Matching Methods for Observational Microarray Studies

Ruth Heller [1,*], Elisabetta Manduchi [2] and Dylan Small [1*]

[1]Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6340

[2]Computational Biology and Informatics Laboratory, Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA 19104-6021

Associate Editor: Dr. Joaquin Dopazo

ABSTRACT

Motivation: We address the problem of identifying differentially expressed genes between two conditions in the scenario where the data arise from an *observational study*, in which confounding factors are likely to be present.

Results: We suggest to use matching methods to balance two groups of observed cases on measured covariates, and to identify differentially expressed genes using a test suited to matched data. We illustrate this approach on 2 microarray studies: the first study consists of data from patients with two cancer subtypes, and the second study consists of data from AMKL patients with and without Down syndrome.

Availability: R code (www.r-project.org) for implementing our approach is included as supplementary material.

Contact: ruheller@wharton.upenn.edu

1 INTRODUCTION

The goal in microarray analysis is often to identify the genes that are differentially expressed between two conditions. For this purpose, randomized experiments are very useful. The random assignment of cases to conditions practically ensures (in a large enough sample of cases) that the two groups of cases differ only in the condition assignment. However, in microarray studies, random assignment is not always feasible. For example, the "conditions" may be cancer and non-cancer tissues, or different types of cancer, so one cannot control the assignment of cases to conditions. Potter (2003) emphasizes that such a study is not an experiment because exposures are not being assigned randomly.

In an observational study, the groups of cases from two conditions may differ in aspects other than those defining the two conditions. These differences may be mistaken for differential expression between the two conditions. In Section 3 we give an example where two subtypes of cancers are investigated for differential expression but there is an imbalance in the distribution of sex across the groups of cases (where a case is a patient). If a gene is differentially expressed between the sexes, then without adjusting for the imbalance in sex, it is not possible to attribute to the cancer subtype an observed differential expression in that gene between the cancer subtypes, even if the comparison is highly statistically significant. However, the observed difference can be attributed to cancer subtype if the groups are first balanced on confounding factors such as sex. As Potter (2003) remarks, in order to attribute differences in gene expression to differences in disease state, it is necessary to control for potential confounders such as age, sex, genetic profile, histology and treatment of the person.

The group differences, other than those defining the groups, that might affect the gene expression levels can be of two kinds: those that have been accurately measured, called *measured confounders*, and those that have not been measured but are suspected to exist, called *unmeasured confounders*. Accounting for measured confounders through adjustments and addressing uncertainty about unmeasured confounders through a sensitivity analysis are central statistical topics in the analysis of observational studies (see Rosenbaum (2002)).

There are two main strategies for adjusting for measured confounders: 1) including them in a regression model (Smyth (2005), Hummel et al. (2008)), and 2) matching methods. The second strategy has the following advantages over the first strategy in an observational microarray study. First, it is easy to examine the balance between the matched cases to understand the limits of the analysis. In a regression framework, this balance is obscured. Rubin (1979) demonstrates in simulation studies that if there is not sufficient overlap in the covariates between the two groups, then any comparison of the groups has to be based on an extrapolation and is not generally reliable. Second, once the matching is done, it is straightforward to test all genes for differential expression in a nonparametric way. In the regression framework, on the other hand, a different model may be appropriate for each gene. Examining the fit of tens of thousands of regression models is impractical. Thus, typically regression analysis assumes that a linear model holds for each gene. This strong linearity assumption is not needed in matching methods. Third, matching methods prevent explorations of the data in such a way that the inference is ultimately invalidated (Rubin (2007)). Adjustment by matching encourages the analyst to focus on balancing the two groups rather than modeling the relationship between the groups and the expression levels. In the regression modeling framework, the temptation exists to explore different models for each gene and then report the ones that yield the most "interesting discoveries". In large data sets with thousands of genes, the analyst is bound to find models that discover interesting genes even if there are no truly interesting genes. A strategy that performs a large number of analyses and reports only the most promising analysis would invalidate the resulting discoveries in the

*to whom correspondence should be addressed

sense that they may not be reproducible. Matching methods do not have this problem.

Our goal in this paper is to focus on measured confounders and introduce applications of matching methods to microarray analysis. These methods should be used when random assignment is not feasible or ethical, and potential confounders have been measured. In Section 2 we introduce the methods. In Sections 3 and 4 we apply our approach to examples. In Section 5 we give final remarks.

2 ANALYSIS METHOD

The framework is as follows. We have two groups of cases. For each case, we have the expression levels on tens of thousands of genes as well as measured covariates (typical covariates may be patient characteristics such as sex and disease state). In Section 2.1 we introduce a matching method for cases with respect to measured covariates that is appropriate for a microarray study, taking into account the fact that often some of the covariates have missing values and the groups are of unequal size. After matching, the groups can be compared and the expression data are used to test every gene for differential expression. In Section 2.2 we describe tests for differential expression on the matched sets.

2.1 Matching on Measured Covariates

In microarray studies there are often cases that have missing data for some of their covariates. Based on the discussions in Rosenbaum and Rubin (1984) and Hansen (2004) on how to handle missing data, we suggest the following. If a case has missing information for a covariate, then the average covariate value across cases with non-missing information is assigned to this case. Categorical covariates are first transformed into indicator functions. For example, the covariate sex can be coded as 1 if the case is a male, 0 if the case is a female, and the proportion of males out of the cases with non-missing sex information if the case has missing sex information. A dummy variable indicating missing cases for each covariate is created to be considered as an additional covariate when matching. Thus, the matching will group together subjects that are comparable in terms both of observed covariate values and of "covariate missing-ness".

The most common matching method matches pairs. However, this method drastically reduces the sample size if one group is much larger than the other. Therefore, we suggest using a *full matching* method instead. This method was introduced in Rosenbaum (1991) and is compared to common matching methods in Hansen (2004) and Gu and Rosenbaum (1993). Full matching divides the cases into a collection of matched sets, each consisting of one case from one group and a positive number of cases from the other group. In one matched set the individual case matched to many cases may be from the first group, but in another matched set the individual case may come from the other group. To improve power, it may be helpful to put bounds on the number of cases allowed in a matched set (Hansen (2004)).

The full matching method minimizes a measure of average discrepancy within matched sets between the groups. We say that a matching has improved comparability if the average discrepancy within matched sets is smaller. For example for the covariate sex, if a matched set consists of one male case from the first group matched to n cases from the second group, the higher the proportion of males in these n cases the more comparable the groups are in terms of sex within this matched set.

A typical discrepancy measure can be derived by the following steps: (1) define a pairwise distance between cases, based on the measured covariates; (2) for each matched set, the *matched set discrepancy* between the groups is the sum of the pairwise distances between cases, within the matched set, that belong to different groups; then (3) the *average discrepancy* is the average of the matched set discrepancies.

For microarray data, an attractive distance function to use in step (1) is the Mahalanobis distance on the ranks, which is calculated as follows. First, replace each covariate value by its rank among all cases. Let $\mathbf{q_1}, \mathbf{q_2}$ be the vectors of covariate ranks for two cases. The distance between these cases is defined as the Mahalanobis distance $d(\mathbf{q_1}, \mathbf{q_2}) = \sqrt{(\mathbf{q_1} - \mathbf{q_2})^t \hat{\Sigma}^{-1} (\mathbf{q_1} - \mathbf{q_2})}$, where $\hat{\Sigma}$ is an estimated covariance matrix of the ranks, e.g. it could be estimated by pooling within group covariance matrices. The attractiveness of this distance function is that the ranks are robust to outliers, and the Mahalanobis distance takes the correlation among the covariates into account. Full matching using the above Mahalanobis distance function will find an optimal matching in terms of the *average discrepancy*.

Another goal of matching is to achieve *balance*, assessed here by comparing the *standardized difference* for each covariate before and after matching, as suggested in Haviland et al. (2007). Specifically, the standardized difference for a covariate x before matching is the mean of that variable in group 1, $\overline{x}_1$, minus the mean of that variable in group 2, $\overline{x}_2$, divided by $\sqrt{s_1^2/2 + s_2^2/2}$, where s_1 and s_2 are the within group standard deviations of group 1 and group 2 respectively. Formally, $d_{\text{Before}}(x) = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{s_1^2/2 + s_2^2/2}}$. The standardized difference for a covariate x within I matched sets is calculated by the following steps: (1) for each matched set i compute the difference in the mean of the variable in group 1 , $\overline{x}_{1i}$, minus the mean of the variable in group 2, $\overline{x}_{2i}$, divided by $\sqrt{s_1^2/2 + s_2^2/2}$; and (2) compute the average values from (1). Formally, $d_{\text{After}}(x) = \frac{1}{I} \sum_{i=1}^{I} (\frac{\overline{x}_{1i} - \overline{x}_{2i}}{\sqrt{s_1^2/2 + s_2^2/2}})$. A more comprehensive discussion of balance can be found in Gu and Rosenbaum (1993) and Rubin and Thomas (2000).

A good summary of the overall balance is the standardized difference on the log odds propensity scores (Rubin (2007)). The *propensity score* of a case is defined as the conditional probability that the case will be in the first group rather than in the second group given a set of covariates used to predict to which group the case belongs. Propensity scores are not known in practice, but can be estimated from the data by logistic regression. The propensity score for each covariate vector $\mathbf{x}$ (corresponding to a case), $e(\mathbf{x})$, is estimated from the logistic regression model, $\hat{e}(\mathbf{x})$. The log odds estimated propensity score vector, $\log(\frac{1-\hat{e}}{\hat{e}})$, has components $\log(\frac{1-\hat{e}(\mathbf{x})}{\hat{e}(\mathbf{x})})$, where x varies across the covariate vectors. The standardized difference of the log odds propensity scores before and after matching is $d_{\text{Before}}(\log(\frac{1-\hat{e}}{\hat{e}}))$ and $d_{\text{After}}(\log(\frac{1-\hat{e}}{\hat{e}}))$ respectively.

To achieve both good distance and balance, Rosenbaum and Rubin (1985) suggest to add a *propensity score* caliper. The 'propensity score caliper with Mahalanobis metric matching' method minimizes the sum of the matched set discrepancies with a propensity score caliper that guarantees the estimated propensity

scores are not too far apart within matched sets. In the example in Section 3, if the difference in the log odds propensity scores between two cases from different groups is larger than 0.2 standard deviations of the estimated propensity scores, the distance between the cases is set to $d(\mathbf{q_1}, \mathbf{q_2}) + 100000$.

Other methods can be used to achieve both good distance and balance. Gu and Rosenbaum (1993) give some guidance based on simulations. When the number of covariates is small (say 5), they suggest matching using the Mahalanobis distance with a propensity score caliper. When the number of covariates is large (say 20), they suggest matching using a distance function based on propensity scores.

How small should the summary of the overall balance be in order for the matching to be considered successful? Imai et al. (2008) suggest that it should be as small as possible subject to efficiency constraints (such as the number of cases allowed in a matched set). Hansen (2004) suggests comparing full matching with efficiency constraints to the full matching without constraints, and choosing the matching with efficiency constraints that results in close overall balance to that of full matching without constraints. Having a standardized absolute difference of the log odds propensity scores below 0.05 is reassuring, but there is no universal threshold in the literature above which the matching is considered unacceptable.

2.2 The test

Once the cases are matched, each gene is tested for differential expression. In this section we describe how to test each gene.

The expression data for a gene in a group are not necessarily normal and may be contaminated by outliers, so nonparametric tests that operate on the ranks have been suggested for testing whether a gene is differentially expressed between groups of cases (see e.g. Troyanskaya et al. (2002) and Neuhauser and Senske (2004)). For unmatched data, a popular test is the *Wilcoxon rank sum test* (also called the Mann-Whitney test, Ewens and Grant (2005) page 142). For paired data, the *Wilcoxon signed rank test* can be used (Ewens and Grant (2005) page 143).

Since the full matching yields matched sets with more than two cases we need a suitable test. Hodges and Lehmann (1962) introduced *the aligned rank test*. We explain this test by an example. Suppose the expression levels of two groups of sizes 3 and 7 are to be compared, where each case from the first group is matched to either 3 or 2 cases from the second group. Table 1 shows the expression levels of the cases from each group within the matched sets. The first step in the computation of the test statistic is to bring the cases in each matched set into alignment with one another by subtracting from each observation the mean observation in the matched set to which it belongs. In Table 1, the residual data after alignment is shown in the last two columns. Once the observations are aligned they are ordered and ranked without regard to their set assignment, see Table 2.

The test statistic is the sum of the ranks of the observations that are labeled to come from the first group: 1 + 3 + 5 = 9.

Under the null hypothesis, the assignment of group labels is done at random within each matched set. The null distribution can be calculated by computing the test statistic for all possible permutations within matched sets, or a Monte Carlo sample thereof. Specifics follow. Let w_i be the sum of the ranks after alignment of matched set i for group 1. Then the test statistic is $w = w_1 + w_2 + \ldots + w_I$ where I is the number of matched sets. Let m_i be the number of cases in group 1 in set i and N_i be the total number of cases in i. Then there are $\binom{N_i}{m_i}$ possible permutations for computing W_i and $\prod_{i=1}^{I} \binom{N_i}{m_i}$ possible permutations for computing W. The p-value for a one sided test is the proportion of times that the permuted test statistic is larger (or smaller) than the observed test statistic. Note that for I reasonably large, W is approximately normally distributed, since it is the sum of I independent random variables, and thus the p-value computation can be based on the normal approximation as follows. Denote the N_i ranks in matched set i by r_{ij}. The expectation and variance of W_i are $E(W_i) = \frac{m_i}{N_i} \sum_{j=1}^{N_i} r_{ij}$ and $var(W_i) = \frac{m_i(N_i - m_i)}{N_i - 1}[\frac{1}{N_i} \sum r_{ij}^2 - (\frac{1}{N_i} \sum r_{ij})^2]$, so the (one sided) p-value is approximated by $1 - \Phi[(w - \sum_{i=1}^{I} E(W_i))/\sqrt{\sum_{i=1}^{I} var(W_i)}]$.

We conclude with three remarks about this test. First, note that the alignment is necessary to remove the effect of the matched sets: if there is a large set effect, then without alignment, one set can have systematically larger ranks than another set and there will be no power to detect differences between the group labels. Second, note that the method of alignment and the choice of test statistic may be adjusted to the problem (see Podgor (1994) for a comparison of various tests). Third, note that for paired samples, the aligned rank test is not identical to the Wilcoxon signed rank test but they are equivalent for large enough samples (see Hodges and Lehmann (1962) for details).

Given a microarray study, the unadjusted p-values from the aligned rank test for each gene need then to be adjusted by applying a multiple comparisons procedure. For example, the *BH procedure* in Benjamini and Hochberg (1995) can be applied to control the FDR at level q.

We implemented in R the aligned rank test. The code is available as supplementary material. The R code to run the example in the next section 3 (including the multiple comparisons procedure) is also available as supplementary material. The matching was performed using the function `fullmatch` of the R-package `optmatch` (Hansen et al. (2008)).

Table 1. An example of expression data for a gene for testing the differential expression between two groups that are matched into 3 sets.

Matched set ID	Group 1	Group 2	Mean	Resid Group 1	Resid Group 2
1	5.3	5.7, 5.9, 6.3	5.8	-0.5	-0.1, 0.1, 0.5
2	6.5	7.6, 9.6	7.9	-1.4	-0.3, 1.7
3	3.4	3.1, 4.8	3.7	-0.2	-0.6, 0.8

Table 2. The sorted residuals and their ranks.

Resid	-1.4	-0.6	-0.5	-0.3	-0.2	-0.1	0.1	0.5	0.8	1.7
Rank	1	2	3	4	5	6	7	8	9	10

Table 3. Distribution of measured covariates in cases from B-cell and T-cell ALL. For indicator variables the entries are the proportion of 1s.

Covariate	Description	B-cell type	T-cell type
age	age (years)	mean 33, sd 15	mean 30, sd 11
sex	1=male, 0 = female	59/93	24/32
mdr	1= multi-drug resistant, 0 otherwise	18/93	6/32
stage	1=stage of cell differentiation < 3, 0 otherwise	55/90	16/28
kinet	1 = chromosome number > 46, 0 otherwise	20/90	7/31

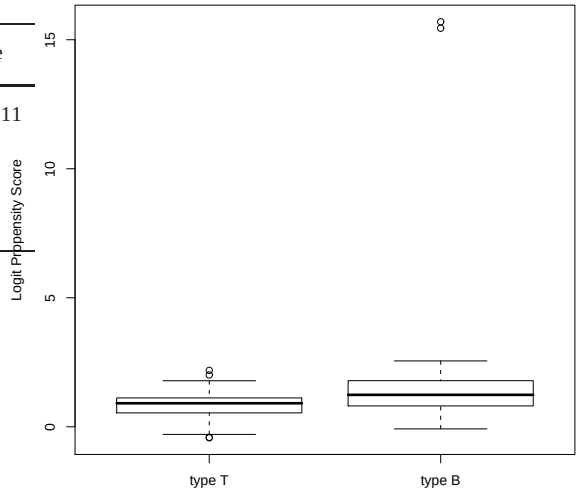

Fig. 1. Boxplots of the estimated log odds propensity score for type T and type B Leukemia patients.

3 A COMPARISON OF TWO CANCER SUBTYPES

We analyzed the data set by Chiaretti et al. (2004), available in the Bioconductor ALL package at `www.bioconductor.org`. The data set was collected to identify genes that distinguish subgroups of leukemia patients. 128 patients are split into 95 with B-cell and 33 with T-cell type acute lymphoblastic leukemia (ALL). The data consist of 12625 expression profiles from the HGU95aV2 Affymetrix chip for each patient.

Information was available for the following additional covariates: age, sex, multi drug resistance (mdr), the stage of cell differentiation (stage), and an indicator of whether the chromosome number was larger than 46 (kinet). Table 3 shows the distribution of these variables in cases from the T-cell type and B-cell type.

As discussed in Section 2.1, we assigned the mean value of a variable in the cases where this variable is missing, and we included a dummy variable indicating missing cases of each variable, so the matching balances both the observed data and the pattern of missing data.

We dealt with missing values as suggested in Section 2.1. Propensity scores were estimated by logistic regression. We used full matching based on the Mahalanobis distance of ranks between each pair using a propensity score caliper of 0.2.

The standardized absolute differences of the log odds propensity scores before and after matching were 0.26 and 0.45 respectively, indicating that the matching failed to balance the propensity scores. The distribution of log odds propensity scores in figure 1 shows that two cases in the B-cell group are not comparable to any cases in the T-cell group. These cases have missing age information. Consequently we removed these two cases and limited our inferences to cases that have age information. Repeating the matching, the standardized absolute difference of the log odds propensity scores before and after matching were then 0.37 and 0.04 respectively, indicating that the matched sets are balanced on the propensity scores. Table 4 compares the standardized absolute difference between the covariates on the subset of 126 cases before and after matching.

Examination of the number of cases per matched set revealed that two of the sets have at least 8 B-cell cases matched to one T-cell case, whereas others have only one B-cell case matched to one T-cell case. To increase power without introducing imbalance, we repeated the matching on the 126 cases with an additional constraint, that the ratio of matches between the two groups in any matched set be at most 1 to 6. The standardized difference of the log odds propensity score was only slightly higher, 0.05. Had we chosen 5 instead of 6 as a bound, the standardized difference of the log odds propensity scores would have been higher, 0.12. The last column in Table 4 shows the standardized absolute difference between the variables on the subset of 126 cases after matching with this constraint.

Table 4. The standardized absolute difference between measured covariates before and after matching the 126 cases, as well as after matching the 126 cases with a constraint on the ratio between the two groups in the matched sets.

Covariate	Before Matching	After Matching	After constraint matching
age	0.18	0.05	0.02
sex	0.19	0.06	0.07
mdr	0.02	0.09	0.00
stage	0.06	0.03	0.03
kinet	0.00	0.07	0.08
log odds propensity score	0.37	0.04	0.05

Once the measured covariates were successfully balanced and the ratio of cases between the groups in matched sets was restricted, we turned to the analysis of the expression data. To calculate an unadjusted p-value for each gene we used the aligned rank test. To control the FDR at the 0.01 level on the 12625 genes tested, the BH procedure was applied at the 0.01 level on the unadjusted p-values. 1684 genes were discovered.

For comparison, we performed a typical analysis that did not control for measured confounders by calculating the p-values from the Wilcoxon rank sum test without matching. Applying the BH procedure at the 0.01 level, 1931 genes were discovered. Out of these discovered genes, 1600 genes were also discovered using our procedure.

Although we discovered fewer genes than the unmatched analysis, we have more confidence that the differential expression of the discovered genes can indeed be attributed to the difference between T-cell and B-cell ALL.

To reinforce this point, we further did the following comparison. An unmatched analysis for the effect of sex on differential expression using the Wilcoxon rank sum p-values and a BH procedure at the 0.01 level discovered 12 genes. Two of the gene discoveries had also been discovered in the matched set analysis. Although there was an imbalance of sex between the groups (75% males in T-cell ALL and 63% males in B-cell ALL), since sex was matched upon in the latter analysis, we can exclude sex as the sole explanation of the differential expression for these two genes.

4 A COMPARISON OF AMKL PATIENTS WITH AND WITHOUT DOWN SYNDROME

We analyzed the data set by Bourquina et al. (103), available in `ftp://ftp.ncbi.nih.gov/pub/geo/DATA/SeriesMatrix/GSE4119/`. The goal was to identify genes that distinguish between acute megakaryoblastic leukemia (AMKL) patients with and without Down syndrome (DS). There were 21 DS and 38 non-DS AMKL samples. Among the available covariates was the age of the patients. The age distribution was extremely imbalanced between the two groups: in the DS group all patients were toddlers between 8 and 36 months old, with a median of 18 months; in the non-DS group 24 patients were toddlers between 0.03 and 38 months old, with a median of 14 months, but the remaining 14 patients range from 4.25 years to 76 years in age with a median of 42.5 years. The expression data consist of 22283 expression profiles from the U133A Affymetrix chip for each subject.

Our goal in this example is to compare a matching based analysis to a regression based analysis. For simplicity, we analyze the data adjusting just for one covariate, age, but in practice the additional measured covariates need to be adjusted for.

Matching based analysis We used pair matching on age. Every DS patient was matched to a non-DS patient that is of similar age. Since we had 21 DS patients and 38 non-DS patients, pair matching resulted in 21 pairs. The distribution of log odds propensity scores in figure 2 shows that two non-DS cases were not comparable to any cases in DS group. These cases have ages 51 and 72 months. Consequently we removed these cases and limited our inferences to toddlers aged 7 months to 38 months. We had 21 DS and 19 non-DS toddlers, so pair matching resulted in 19 pairs. On this subset of toddlers, the standardized absolute difference for age before and after matching was 0.09 and 0.03 respectively, indicating that the matched sets are balanced on age. For each of the 22283 genes, the differential expression between the 19 matched pairs was tested using the Wilcoxon signed rank test, and adjusted for multiplicity using the BH procedure for FDR control at level 0.05. 1628 genes were discovered.

Regression based analysis We used the package *limma* (Smyth (2005)), available in Bioconductor, for the regression based analysis. A linear regression analysis of the log expression levels on the group and on age, on the 21 DS and 38 non-DS patients, yielded 3469 gene discoveries (using the BH procedure for FDR control at level 0.05), 1494 of them discovered also in the matching based analysis. However, these results cannot be trusted due to a large extrapolation of the linear model inferences, since the DS group consists solely of toddlers, yet only 60% of the non-DS group are toddlers and in the remaining 40% most patients are grown-ups.

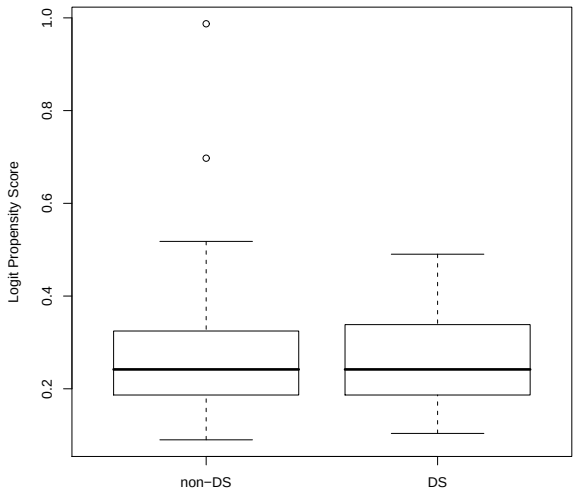

Fig. 2. Boxplots of the estimated log odds propensity score for DS and non-DS AMKL patients.

The matching based analysis guided us to the subset of patients with similar age information: toddlers aged 7 months to 38 months. Restricting the regression to this subset of 21 DS and 19 non-DS patients, 1715 genes were discovered (using the BH procedure for FDR control at level 0.05). 1208 of these genes were also discovered in the matching based analysis. The regression analysis assumes that the relationship of age with the log expression values is linear. Moreover, it assumes there are no outliers. Figure 3 shows the residual plot from the regression, as well as the scatter plot of log expression on age, for 2 genes that were discovered by the regression based analysis but not by the matching based analysis. The linearity assumption is questionable in both these genes, and there are outliers in the data. Similar plots on many other genes, not illustrated, show deviations from linearity and outliers. The matching based analysis, on the other hand, does not assume linearity, is robust to outliers, and indicates how balanced the data are so there is no danger of extrapolation.

5 DISCUSSION

In a microarray experiment a procedure that controls for an error measure such as the FDR is typically applied to identify differential expression between two groups. Thus some of the discoveries may be false, but there is a limit on the proportion of false discoveries among all discoveries. In an observational microarray study, discoveries may in addition be false due to bias. In order to remove bias, the groups need to be balanced for measured confounders. We introduced matching methods to balance the groups before testing for differential expression. We illustrated by

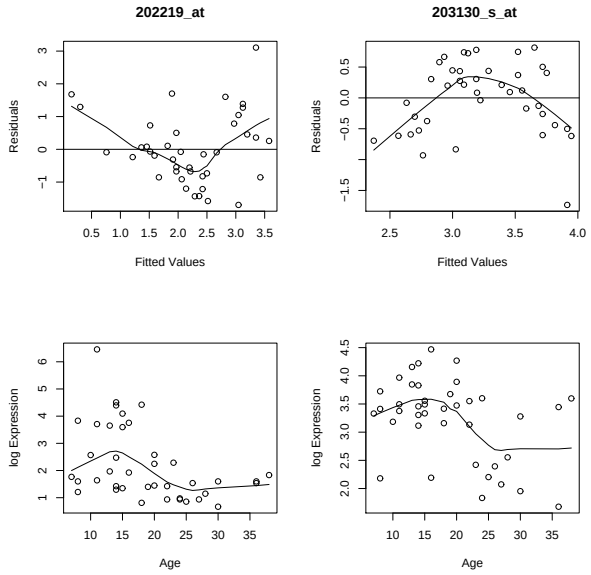

Fig. 3. For two genes that were discovered by the regression based analysis but not by the matching based analysis: the residual plot from the regression of log expression on age and group (top), and a scatter plot of log expression by age (bottom), including the locally-weighted polynomial regression curve.

examples that many discoveries are still made after balancing the covariates. Moreover, the confidence that these discoveries are true discoveries is far greater than in an analysis that does not first attempt to balance the groups. This added confidence is due to the fact that after matching, the groups are comparable in terms of measured covariates and therefore, if we find a differential expression between the groups, it cannot be due to these covariates.

In order to remove bias it may be necessary to restrict the analysis to a subpopulation. In Section 4, the DS group was comprised solely of toddlers but the non-DS group included older children and adults as well, therefore to achieve balance on age between the groups it was necessary to restrict the analysis that compares the DS group to the non-DS group to toddlers. Thus our inference is valid only for toddlers, and the genes that were found to be associated with DS in toddlers may or may not be associated with DS in other age groups.

In this paper we focused on measured confounders. Methods for addressing the possible effects of unmeasured confounders (see e.g. Rosenbaum (2002)) may be considered for observational microarray studies in future work.

REFERENCES

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B Met.*, 57 (1):289–300.

Bourquina, J., Subramanian, A., Langebraked, C., Reinhardte, D., Bernardf, O., Ballerinig, P., Baruchelh, A., H., C., Dastuguej, N., Haslek, H., Kaspersl, G., Lessardn, M., Michauxo, L., Vyasp, P., Weringm, E., Zwaanm, C., Golub, T., and Orkina, S. (2006 (103)). Identification of distinct molecular phenotypes in acute megakaryoblastic leukemia by gene expression profiling. *PNAS*, 9:3339–3344.

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J., and Foa, R. (2004). Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, 103:2771–2778.

Ewens, W. and Grant, G. (2005). *Statistical Methods in Bioinformatics: An Introduction*. Springer, New York.

Gu, X. and Rosenbaum, P. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2(4):405–420.

Hansen, B. (2004). Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association*, 99(467):609–618.

Hansen, B., Bertsekas, D., and Tseng, P. (2008). Functions for optimal matching. r package version 0.4-1. *http://www.stat.lsa.umich.edu/ bbh/optmatch.html*.

Haviland, A., Nagin, D., and Rosenbaum, P. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods*, 12 (3):247–267.

Hodges, J. and Lehmann, E. (1962). Rank methods for combination of independent experiments in analysis of variance. *The Annals of Mathematical Statistics*, 33(2):482–497.

Hummel, M., Meister, R., and Mansmann, U. (2008). Globalancova: Exploration and assessment of gene group effects. *Bioinformatics*, 24(1):78–85.

Imai, K., King, G., and Stuart, E. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, series A*, 171:481–502.

Neuhauser, M. and Senske, R. (2004). The baumgartner-wei ss-schindler test for the detection of differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 20(18):3553–3564.

Podgor, M. (1994). A cautionary note on applying scores in stratified data. *Biometrics*, 50:1215–1218.

Potter, J. (2003). Epidemiology, cancer genetics and microarrays: making correct inferences, using appropriate designs. *Trends in Genetics*, 19 (12):690–695.

Rosenbaum, P. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society*, 53:597–610.

Rosenbaum, P. (2002). *Observational Studies*. Springer, New York.

Rosenbaum, P. and Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524.

Rosenbaum, P. and Rubin, D. (1985). Constructing a control-group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, 39(1):33–38.

Rubin, D. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366):318–328.

Rubin, D. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomization trials. *Statistics in medicine*, 26:20–36.

Rubin, D. and Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, 95 (450):573–585.

Smyth, K. (2005). Limma: linear models for microarray data. In Gentleman, R., Carey, V., Dudoit, S., Irizarry, R., and Huber, W., editors, *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, pages 397–420, New York. Springer.

Troyanskaya, O., Garber, M., Brown, P., Botstein, D., and Altman, R. (2002). Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, 18(11):1454–1461.