Identifying genes involved in the growth of adrenocortical tumors: An approach based on partial linear models

Jonas Mueller
UC Berkeley Department of Statistics

Abstract

Many diseases exist for which some numerical severity indicator can be measured. To identify the differentially expressed genes involved in the progression of such a condition, we propose a method that is more informative than traditional analyses because it accounts for severity measurements in addition to comparing diseased vs. control samples. Our approach incorporates the effects of disease severity as a nonlinear term which is added to a linear model to explain the variation in each gene’s expression level between normal and diseased samples. Applying this method to microarray data from human adrenocortical carcinomas, where expression profiles of each tumor are accompanied by mitotic rate measurements, we report interesting findings on the relationship between gene expression and mitotic rates in adrenocortical cancer. These models have been implemented in the R package plmDE, which facilitates their flexible application to differential gene expression analyses containing information on relevant quantitative phenotypic variables.

Introduction

Based on autopsy studies, up to 9% of the population has adrenal gland tumors [1], but fortunately the vast majority of cases are benign; adrenocortical cancer is very aggressive and because it is so rare, little was known about the condition until recently [2]. The malignancy of a tumor can be quantified through both objective measurements of its size and mitotic rate and also through semi-quantitative subjective scoring systems like Weiss grading, which have proven to be fairly good indicators of mortality [3]. A multitude of other conditions such as type-II diabetes, COPD, and hypertrophic cardiomyopathy also have severity scales which can be estimated by numerical figures (eg. insulin resistance for diabetes [4], FEV1/FVC ratios for COPD [5], and ventricle thickness for hypertrophic cardiomyopathy [6]).

While there has been extensive work done in attempting to predict disease severity from transcriptional profiles [7], there is less literature that deals with the complimentary problem of learning more about gene expression through the relationship between these severity scores and the associated transcriptional activity. Furthermore, even when severity measurements have been incorporated into an analysis of differentially expressed genes, they are often brought in as factors, despite their inherent continuity, because of sample size constraints imposed by the high price tag of each transcriptome. In numerous gene expression studies where real-valued severity assessments such as disease duration in arthritis [7] or extent of coronary artery stenosis [8], were available, investigators opted to bin subjects into ranges of disease affliction such as mild vs. severe and conduct pairwise t or Fisher tests. However, such ordered factoring approaches miss any differential
expression effects which might only be exaggerated between two of these imposed levels, especially if the particular gene in question happens to have sharp fluctuations in its expression over close values of the severity measure.

Although simple two-way comparisons of gene expression in healthy and diseased groups have become widely adopted in biology, such solely factor-based analysis is likely not the most productive method of examining the underlying causes of many disorders. To illustrate why, consider the case of cancer, for which there are at least six hallmarks present in all incidences of the condition [9]. Because these traits are common to all cases, inferences made entirely based on the set of genes identified as differentially expressed in a transcriptome comparison of cancerous and noncancerous tissues are limited to determination of these genes as possibly involved in one of the numerous traits of the disease. Without resorting to other means, we are left with no means to ascertain which specific trait of the disease is associated with a specific gene’s up or down-regulation, and can only conclude that this gene might play a role in some aspect of the disease. In a complex condition like cancer, where the list of known hallmarks continues to grow [10], cures at the RNA level depend on understanding each these characteristics well, something which cannot be accomplished through a simple comparison of diseased vs. control samples.

However if these traits can be quantified, then ascertaining their relationships with levels of gene expression provides a far more insight on the trait-specific effects of up/down-regulation of certain genes. As the increasing affordability of sequencing experiments widens the availability of gene-expression-paired-with-disease-severity data [11], which will result in more expression profiles at differing levels of affliction, we predict approaches based on treating these severities as continuous measures will become far more prevalent than they are today. In this paper, we propose some methods for utilizing quantitative severity measures to identify sets of differentially expressed genes that are potentially of greater interest for the aforementioned reasons.

**Methods**

In a disease for which severity level (or any specific trait of interest) can be numerically expressed by some measure $S$, it is reasonable to suppose the measured expression level of a gene $Y$ in a profiling experiment (where $I_D$ indicates the presence of the disease) can be described by the following additive partially linear model:

$$E[Y|I_D, S] = \beta_0 + \beta_1 I_D + I_D f(S)$$

(1)

where we do not impose the condition of linearity in the relationship between $S$ and $Y$ due to the inherently complex nature of the interactions between genes and their environment. The last term describes the interaction effect of the disease on the severity measure. The expression-level measures $Y$ could be probe intensities in a microarray experiment or gene-level counts of read-mapping in a profiling by high-throughput sequencing, in which case we might want to account for the discrete nature of $Y$ by assuming it is Poisson or negative-binomial distributed and employing a link function to generalize this partially linear model. We also impose the constraint $f(0) = 0$ on the intercept of the function which is needed to ensure the identifiability of this model. If we identified differentially expressed genes for which the hypothesis:

$$\beta_1 = f(S) = 0$$

(2)

is rejected, we would presumably obtain a set of genes whose differential expression is more likely attributed to their effect on $S$ in the course of this disease, than genes that fall the rejection region.
of the test for $\beta_1 = 0$ in the traditional case where the function of severity is disregarded. We now illustrate how this method can be implemented in practice.

In 2006, Giordano et al. [2] performed transcriptome profiling by microarray on a set of 65 independent samples of adrenal cortex tissue from humans. Assayed by the 54,675 probe sets (representing 19,686 different genes) of the Affymetrix HG U133 2.0 Plus Array, these samples stem from different individuals and are divided into 3 categories: 10 are from individuals with a normal adrenal cortex, 22 from individuals with adenomas (benign tumors), and 33 are from individuals with adrenocortical carcinomas. To supplement this sequence data, the investigators also recorded the weight and mitotic rate of each cancerous tumor as well as the weights of the benign tumors. The data and collection methodology is recorded in NCBI’s Gene Expression Omnibus under GEO Series accession number GSE33371.

After the exclusion of 33,581 probes-pairs at which perfect matches minus mismatched was less than 100 in a representative sample, the remaining probe-set intensity data was normalized via a nearly-full quantile procedure that exactly mapped quantiles from 0.01 to 0.99 in increments of 0.01 to the quantiles of the representative sample and used linear interpolation to correspondingly transform the rest of the intensities. Finally, the data were log transformed according to $\log(\max(x, 50) + 50)$. Through a cluster analysis of gene expression, Giordano et al. ascertained that one sample, ACC053, which was classified as a carcinoma due to its large tumor size, was actually closer to the cohort of adenoma patients. Because the cancer diagnoses were made solely based on tumor size and they also report that this individual revealed no evidence of recurrent or metastatic disease, we removed this subject from any further analysis to avoid further ambiguity.

From this transcriptome data, we wish to identify the genes that are not only differentially expressed in adrenocortical carcinomas, but whose differential expression is linked with increased rates of mitosis in the tumors. Mitotic rate has been designated a good prognostic for the condition as well as an accurate predictor of survival [12]. Also interested in this problem, Giordano et al., like others we previously mentioned, chose to categorize the numerical values of the mitotic rates either as low-grade or high-grade and they then compared these two subgroups of the cancerous cohort to identify differentially expressed genes. However, such an analysis is sub-optimal, both for the reasons we mentioned earlier and because it ignores interaction between mitotic rates and the other gene expression effects in cancer.

After the data have been processed and normalized to reflect accurately comparable measures of gene expression, we apply model (1) in the following form:

$$E[Y_{j,k} | I_{cancer}(k), S_k] = \beta_{0,j} + \beta_{1,j}I_{cancer}(k) + I_{cancer}(k)f_j(S_k)$$  \hspace{1cm} (3)

where $Y_{j,k}$ is the probe intensity of probe $j$ (log-transformed) in the array assay of sample $k$, $I_{cancer}(k)$ indicates whether or not sample $k$ was a carcinoma, and $S_k$ is the mitotic rate measured from sample $k$. Although there exist iterative approaches to fit $f_j$ such as local scoring and the back fitting algorithm [13] or marginal integration [14], we choose to adopt a much simpler method, realizing that we must fit this model to over 54,675 probe sets. A linear model can be formed from this one by decomposing $f_j$ into a set of basis functions $\{b_1, \ldots, b_m\}$ so that the problem of estimating $f_j$ has been converted into a problem of estimating the coefficients of this basis. The model now takes the following form:

$$E[Y_{j,k} | I_{cancer}(k), S_k] = \beta_{0,j} + \beta_{1,j}I_{cancer}(k) + I_{cancer}(k)\sum_{i=1}^{m} \gamma_{ij}b_i(S_k)$$  \hspace{1cm} (4)

To fit this model to the data, we select the $b_i$ as the B-spline basis, which possesses nice properties for nonparametric estimation and does not require many parameters to fit smoothly [15]. Because
the limited sample size of the data (27 cancerous samples left after the removal of those with missing mitotic rate measurements), only one knot (at the median of the mitotic rate values) is chosen at first to avoid losing too many degrees of freedom in the estimation of the functions $f_j$. To ensure the identifiability of this model, the intercept of the basis was fixed at 0, leaving me with 21 degrees of freedom since 4 parameters are needed to estimate $f_j$ and 2 to estimate $\beta_0$ and $\beta_1$. Defining $\epsilon_{j,k}$ as the error of the fitted expression level $\hat{Y}_{j,k}$, this model requires the normal assumptions of linear regression to hold for each gene: accurate measurement of $S_k$, linearity ($E[\epsilon_{j,k}] = 0$) and i.i.d. error distribution with constant variance $\epsilon_{j,k} \sim N(0, \sigma^2)$. Although we cannot explicitly check the assumptions for all $j = 1, \ldots, 54675$, we take them to hold as they have been previously deemed reasonable for log-transformed, normalized, microarray data [16].

Once the coefficients of this model are estimated, we use the moderated $F$-static [16] to simultaneously test $\beta_1,j = \gamma_1,j = \ldots \gamma_4,j = 0$ for each gene $j$. The moderated $F$ and $t$ test statistics, which control the estimated variance across genes, have been shown to be more informative for microarray data because they do not become as significant as their standard counterparts in the biologically-less-interesting case where the log-fold changes in expression level are small, but their variance is estimated to be much smaller [16].

**Results**

For the adrenocortical cancer data, testing the combined significance of $\beta_1,j, \gamma_1,j, \ldots, \gamma_4,j$ with the standard false discovery rate method for multiple testing correction returned a list of 5,378 genes (9,376 probe sets) that were expressed at significantly differing levels at the level 0.05, while the moderated $t$-test for $\alpha_1,j = 0$ under the standard model:

$$E[Y_{j,k} \mid I_{cancer}(k), S_k] = \alpha_{0,j} + \alpha_{1,j}I_{cancer}(k) \quad (5)$$

returned a list of 4,641 genes (6,522 probe sets). (Note: a gene was deemed differentially expressed if any of the probe sets which mapped to it according to Affymetrix annotation provided significant evidence of differential expression). These two tests produce very different results as evidenced by the fact that lists of the ten genes with the most evidence for differential expression under each test were entirely disjoint. Figure 1A provides an example of a probe set at which the $p$-value for differential expression under the $t$-test of model (5) fails to be significant (adjusted $p = 0.078$), but is under model (1): adjusted $p = 7.5 \cdot 10^{-7}$. Because the figure illustrates a distinct relationship between the rate of mitosis and notable expression level changes in the carcinoma group, a test to find genes potentially involved with abnormal mitosis in cancer should identify this probe set’s matched gene as significant.

Figure 1B proves any example of a probe set at which the $p$-value for differential expression is significant under model (1) (adjusted $p = 0.006$), but fails to be significant under the methods of Giordano et al. [2], which involved categorizing the mitotic rates as high-grade or low-grade and then comparing the high-grade carcinoma group with the low-grade individuals (adjusted $p$-value = 0.21). Again, the data suggest a fairly evident (nonlinear) relationship between expression level and mitotic rate, and we would therefore want our test to identify the genes represented by such probes as significant.

Because of the single-knot limit imposed on the B-spline basis functions, our model is robust to observations without high leverage, which might nevertheless exert undue high influence (especially under a non-linear fit) stemming from erroneous large deviations in expression level from the rest of the probe measurements. Figure 2 shows that increasing the number of knots in our B-spline approximation ensures that our $f_j$ estimate (for $j$ corresponding to Probe 1552266_at) better fits
Figure 1: Expression levels (log scale of normalized intensity) of two different probe sets. The red curve illustrates the fitted values of our model while the black dotted line represents a loess fit of the actual observed log probe-set intensities on the mitotic rate.

the intensity data from the probe. However, an outlying point of low leverage (e.g., the individual with mitotic rate 70 and high expression at this probe) can then drastically alter the shape of the function. In this case, the coefficients of the B-spline become far more significant and while this probe is not identified as differentially expressed under our model with single knot B-splines (adjusted p-value = 0.11), it is identified as so (adjusted p-value = 0.002) by a similar model which uses 3 knots placed at the quartiles of the mitotic rate distribution. Decisions of how we would like to treat such a probe must be taken into consideration when choosing the spline basis. In the presence of more data, one promising option would be to impose penalties on the splines to reduce over fitting, perhaps turning to the P-spline methods introduced in [15].

Figure 2: Expression levels (log normalized intensity) of probe set 1552266. The fitted values of our model with a single-knot B-spline (red), or a B-spline with 3 knots (blue dotted curve).

We now perform a basic assessment of our model, comparing its performance against a simplified model which assumes that a subset of the genes display expression levels that share a linear relationship with mitotic rate. This model can be written as:

\[ E[Y_{j,k} | I_{cancer}(k), S_k] = \alpha_{0,j} + \alpha_{1,j}I_{cancer}(k) + \alpha_{2,j}I_{cancer}(k) \cdot S_k \]  

and we test \( \alpha_{1,j} = \alpha_{2,j} = 0 \), again using the moderated t-statistic of limma at the fdr-adjusted level of 0.05 and considering a gene differentially expressed if the null hypothesis is rejected for any of
its probes. We can then compare $G_1$, the set of 4,026 genes with significant evidence under model (1), with $G_2$, the set of such genes determined by (6), which contains 3,165 genes (represented by 7,048 probe sets).

We analyze the ontology-annotated subsets of both $G_1$ and $G_2$ with GO::Termfinder [17] looking for terms from the ontology of biological processes which are significantly enriched in each set. Because we are querying so many genes involved in a multitude of processes, we had GO::Termfinder return any process with significant evidence for overrepresentation using a lax cut-off of 0.1 for the unadjusted $p$-value, so that we would obtain numerous “significant” processes in which the genes of each set could potentially be involved. As Whitefield et al. have shown that even vastly different cancers have closely shared proliferation signatures [18], it is reasonable to consider their list of annotated cell proliferation processes present in genes involved in the proliferation of breast cancer (see Supplementary Table 2 of [18]) to be representative of processes that drive the proliferation of adrenocortical cancer. Thus, we can compare the GO::Termfinder $p$-values for the significance of a proliferation-related term in each set to evaluate how the methods for generating $G_1$ and $G_2$ compare in determining the genes whose up/down-regulation is linked with the proliferation of adrenocortical cancer cells.

From Table 1, we find that these the great majority of these terms are far more significantly present in $G_1$ than in $G_2$, providing evidence for the utility of methods that do not assume a linear relationship between gene expression and mitotic rate. Although model (1) identified far more differentially expressed genes, these results suggest that this greater number also contained far more genes involved in determining mitotic rates than the genes identified by the simple linear model. Despite this greater number of genes, we find through permutation testing (randomly distributing the cancerous label and mitotic rate label along with it to samples in all three cohorts and then performing the model fitting again), that in 100 such permutations, the mean number of differentially expressed genes identified by model (1) (using the fdr $p$-value adjustment each time) was 2,291 which implies a false discovery rate of 0.116. This is not much worse than the results of permutation testing done on the simple pairwise test which suggested a false discovery rate of 0.99, providing further evidence validating the use of (1). Finally, an examination of the distribution of $p$-values the model computes for the significance of differential expression in each probe (Figure 3) illustrates a spike in the low $p$-values and reasonable uniformity over the others illustrating the presence of a large subset of genes that is differentially expressed in adrenocortical cancer.

![Figure 3: Histogram of the FDR-adjusted $p$-values returned by our model for the intensities of each probe.](image)

One especially interesting item we noted while looking over plots of the model fit on expression data from various differentially expressed probes was the repeated appearance curves shaped like
Table 1: P-Values (from GO::Termfinder) for the significance of enrichment of various proliferation-related terms in two sets of genes identified as differentially expressed

<table>
<thead>
<tr>
<th>Gene ontology process</th>
<th>$G_1$ P-Value</th>
<th>$G_2$ P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 mitotic cell cycle</td>
<td>3.2E-04</td>
<td>1.1E-03</td>
</tr>
<tr>
<td>2 cell cycle</td>
<td>9.0E-05</td>
<td>2.8E-03</td>
</tr>
<tr>
<td>3 M phase</td>
<td>7.0E-06</td>
<td>3.7E-04</td>
</tr>
<tr>
<td>4 mitosis</td>
<td>3.0E-04</td>
<td>3.5E-04</td>
</tr>
<tr>
<td>5 M phase of mitotic cell cycle</td>
<td>3.2E-03</td>
<td>1.5E-02</td>
</tr>
<tr>
<td>6 cell division</td>
<td>2.9E-07</td>
<td>6.6E-06</td>
</tr>
<tr>
<td>7 DNA metabolism</td>
<td>3.8E-03</td>
<td>8.9E-04</td>
</tr>
<tr>
<td>8 regulation of progression through cell cycle</td>
<td>8.3E-04</td>
<td>5.8E-03</td>
</tr>
<tr>
<td>9 DNA replication</td>
<td>1.9E-03</td>
<td>6.0E-03</td>
</tr>
<tr>
<td>10 interphase</td>
<td>1.0E-03</td>
<td>9.3E-04</td>
</tr>
<tr>
<td>11 interphase of mitotic cell cycle</td>
<td>4.7E-04</td>
<td>3.4E-03</td>
</tr>
<tr>
<td>12 DNA-dependent DNA replication</td>
<td>4.5E-04</td>
<td>7.6E-04</td>
</tr>
<tr>
<td>13 biopolymer metabolism</td>
<td>3.3E-04</td>
<td>1.6E-04</td>
</tr>
<tr>
<td>14 cell proliferation</td>
<td>4.3E-04</td>
<td>1.8E-03</td>
</tr>
<tr>
<td>15 DNA replication initiation</td>
<td>3.4E-04</td>
<td>1.1E-02</td>
</tr>
<tr>
<td>16 response to endogenous stimulus</td>
<td>3.9E-04</td>
<td>9.8E-04</td>
</tr>
<tr>
<td>17 G2/M transition of mitotic cell cycle</td>
<td>4.7E-02</td>
<td>3.8E-03</td>
</tr>
<tr>
<td>18 regulation of mitosis</td>
<td>1.1E-03</td>
<td>7.3E-04</td>
</tr>
<tr>
<td>19 G1 phase of mitotic cell cycle</td>
<td>2.1E-03</td>
<td>4.8E-04</td>
</tr>
<tr>
<td>20 G1 phase</td>
<td>5.0E-07</td>
<td>4.7E-04</td>
</tr>
<tr>
<td>21 traversing start control point of mitotic cell cycle</td>
<td>3.7E-03</td>
<td>2.2E-03</td>
</tr>
<tr>
<td>22 cell cycle checkpoint</td>
<td>8.3E-04</td>
<td>5.7E-04</td>
</tr>
<tr>
<td>23 chromosome segregation</td>
<td>1.6E-03</td>
<td>5.3E-04</td>
</tr>
<tr>
<td>24 mitotic checkpoint</td>
<td>2.4E-06</td>
<td>4.6E-05</td>
</tr>
<tr>
<td>25 regulation of cyclin dependent protein kinase activity</td>
<td>8.9E-04</td>
<td>1.2E-03</td>
</tr>
<tr>
<td>26 DNA replication and chromosome cycle</td>
<td>3.4E-04</td>
<td>1.2E-03</td>
</tr>
<tr>
<td>27 mitotic sister chromatid segregation</td>
<td>8.1E-03</td>
<td>1.7E-03</td>
</tr>
</tbody>
</table>

We are left with 288 genes whose representative probe intensities offer very significant evidence of differential expression. Running this list through GO::Termfinder, we find highly significant enrichment of biological process terms such as “mitotic checkpoint” (Bonferroni-corrected $p$-value = $4.52 \cdot 10^{-18}$), “cell cycle” (Bonferroni-corrected $p$-value = $2.90 \cdot 10^{-14}$), “traversing start control...
Figure 4: Common shape of the estimated function at many differentially expressed probes. The red curve illustrates the fitted values of our model while the black dotted line represents a loess fit of the actual observed log probe-set intensities on the mitotic rate.

point of mitotic cell cycle” (Bonferroni-corrected $p$-value = $2.11 \cdot 10^{-11}$), “regulation of mitosis” (Bonferroni-corrected $p$-value = $1.82 \cdot 10^{-11}$), and “G1 phase of mitotic cell cycle” (Bonferroni-corrected $p$-value = $8.69 \cdot 10^{-11}$). This suggests that these genes, which are increasingly down regulated over increases in low mitotic rates especially warrant further investigation, as they could potentially provide insight on how mitotic rates accelerate over the course of adrenocortical cancer.

Discussion

Having illustrated the utility of the model introduced here, we discuss how it can may be adapted to help answer a number of other interesting biological questions. The data from Giordano et al. also contain information on the tumor weights of from both cancerous and benign samples (adenomas). One interesting topic to pursue is the identification of genes whose expression is involved in differentiating the growth of cancerous tumors from the growth of benign adenoma tumors. To seek an answer, we fit a additive partially linear model to the data of the form:

$$
E[Y_{j,k} \mid \text{sample data}] = 
\beta_0 + \beta_{1,j} I_{\text{adenoma}}(k) + I_{\text{adenoma}}(k)f_j(W_k) + \beta_{2,j} I_{\text{carcinoma}}(k) + I_{\text{carcinoma}}(k)g_j(W_k)
$$

where $Y_{j,k}$ is taken to be the observed expression level of gene/probe $j$ conditional on the observed data, $W_k$ is the tumor weight of the $k$th sample, and we impose the constraint $f_j(0) = g_j(0) = 0$ for all $j$. In this case, the functions $f_j$ and $g_j$ measure the interactions between adenoma and tumor weight and cancer and tumor weight, and to answer the question, we test the expression levels of each probe $j$ against the null hypothesis $\beta_{1,j} = \beta_{2,j}$ and $f_j = g_j$. Using the equivalent linear model which can be approximated by a B-spline basis, there are numerous ways we can choose to test this hypothesis. We can fit one common B-spline basis (with intercept constraint) $\{b_1, \ldots, b_m\}$ to the
measurements of tumor weight, adopting model (9) and testing $\beta_{1,j} - \beta_{2,j} = \psi_{1,j} = \cdots = \psi_{m,j} = 0$.

$$E[Y_{j,k} \mid \text{sample data}] = \beta_{0,j} + \beta_{1,j}I_{\text{adenoma}}(k) + \beta_{2,j}I_{\text{carcinoma}}(k) + \sum_{i=1}^{m} b_i(W_k)[\gamma_{i,j} + \psi_{i,j}I_{\text{carcinoma}}(k)]$$

(9)

Or, we could fit two B-splines (with constraints for identifiability) $\{b_1, \ldots, b_m\}$ and $\{b_1^*, \ldots, b_p^*\}$ where for example, the second basis could be chosen with more knots placed among the measurements of cancerous tumors to reflect the fact that this group has more samples present in the data. In this case, we could adopt a model like (10) and test $\beta_{1,j} - \beta_{2,j} = \psi_{1,j} = \cdots = \psi_{p,j} = 0$.

$$E[Y_{j,k} \mid \text{sample data}] = \beta_{0} + \beta_{1,j}I_{\text{adenoma}}(k) + \beta_{2,j}I_{\text{carcinoma}}(k) + \sum_{i=1}^{m} \gamma_{i,j}b_i(W_k) + I_{\text{carcinoma}}(k) \sum_{i=1}^{p} \psi_{i,j}b_i^*(W_k)$$

(10)

Figure 5: Measured expression levels (log normalized intensity) at probe set 1552319_at. The fit of the full (in red) and reduced null model (in blue) which are tested against each other in (9) (on left) and (10) (on right). One can see that because any sum of B-splines is simply another B-spline, the full models estimate the same function regardless of the number of bases used, as long as their knots are kept in place. However, the estimation differs significantly between the similar null models of the two tests.

Fitting these two models to data from Giordano et al. and testing for differential expression can easily be done through the R package plmDE, which we have implemented along with this analysis. After excluding carcinomas that were much heavier than any of the adenomas to keep the basis functions fitted to the same domain, we used this package to fit both models (9) and (10) to the micro-array data, and the genes they identify as differentially expressed in relation to differences in tumor size between adenomas and carcinomas are fairly different, illustrating the importance of selecting a good model before pursuing any analysis. Figure 5 depicts how these models fit the data.
from specific probes, and such figures are easy to make using the plmDE package. Exploring the
data through such fits can be very informative as we discovered upon finding the earlier observed
trend between mitotic rate and expression level that was exhibited by numerous genes.

We can also generalize these methods to multiple continuous covariates. For example, if we have
a disease $D$ with subgroups $D_1, \ldots D_G$ and control group $N$ as well as other quantitative covariates
(such as various measures of severity) $S_1, \ldots S_C$, then we can use the following model to describe
$Y_j$, the expression level of each gene $j$ in a patient $X$:

$$
E[Y_j \mid \text{severity information about } X] = \beta_{N,j} + \sum_{i=1}^{G} \beta_{i,j} I_{D_i}(X) + \sum_{i=1}^{C} f_{i,j}(S_{i,X})
$$

A multitude of interesting hypotheses can be tested from this model using the B-spline approx-
imation to transform it into an ordinary linear model, and choosing an appropriate contrast matrix
for the desired hypothesis. Because models such as this would allow us to investigate the way traits
like tumor size and mitotic rate are related at the genetic level, we believe these methods will prove
increasingly informative as transcriptome data of this sort accumulates.
References


