Mathematical Statistics
Stockholm University

Statistical methodology in case-control 5'-nuclease assays: Statistical design, modeling and inference for identification of differentially expressed genes

Rolf Sundberg

Mathematical Statistics, Stockholm University, SE-106 91 Stockholm, Sweden

e-mail: rolfs@matematik.su.se; web: www.matematik.su.se/~rolfs

Anja Castensson, Elena Jazin

Evolutionary Biology, Uppsala University, SE-752 36 Uppsala, Sweden

Research Report 2002:9

ISSN 1650-0377

Postal address:
Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:
http://www.matematik.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2002:9**,
http://www.matematik.su.se/matstat

Statistical methodology in case-control 5'-nuclease assays: Statistical design, modeling and inference for identification of differentially expressed genes

Rolf Sundberg

Mathematical Statistics, Stockholm University, SE-106 91 Stockholm, Sweden

e-mail: rolfs@matematik.su.se; web: www.matematik.su.se/~rolfs

Anja Castensson, Elena Jazin

Evolutionary Biology, Uppsala University, SE-752 36 Uppsala, Sweden

October 2002

Abstract

With background in mRNA expression data from a case–control 5'-nuclease assay for the investigation of genes suggested to be schizophrenia-related, aspects on experimental design, statistical modeling, and statistical inference in such studies are discussed. With more individuals than positions per plate, a balanced incomplete design will be required. Based on a multivariate (for several genes jointly) random-effects analysis of covariance (RANCOVA) model, incorporating heterogeneity both between and within individuals, it was found highly efficient to use reference genes to form additional regressors. In this regression on reference gene values it was found important first to average over the intra-individual replicates, because regressions between and within individuals are usually different. This has some consequences for the estimation of plate effects. Topics also discussed in the paper are testing for a significant mean disease effect, multiple genes aspects, and the difficulty to identify genes affected in only a subgroup of cases.

Keywords: ANCOVA, cDNA, mRNA, RANCOVA, Real-time PCR, Schizophrenia, Subgroup, TaqMan.

1 Introduction

For the understanding of the mechanisms of a disease, and the potential development of drugs for its treatment, it is of interest to find genes which are differentially expressed between sick and healthy individuals. The degree of expression is quantified by measurements of mRNA concentrations.

The role of mRNA (messenger-RNA) is to carry the information in the DNA of a gene to the protein synthesis of the ribosomes. If some protein synthesizing gene is expressed at different levels in sick and healthy individuals, in their brains say, the difference might be detected by measuring and comparing the expression of the corresponding mRNA in brain samples from the two groups. Two techniques for mRNA measurements widely used are microarrays and 5'-nuclease assays (TaqMan assays; 5' refers to use of the ribosome binding end region of the mRNA). Microarrays allow the simultaneous study of a large number of genes, but at the expense of lower precision and fewer samples for the same cost, in comparison with the 5'-nuclease assay, that allows many individual samples to be distinguished but not tried on very many genes. Therefore the two study types have dual roles. Microarrays is a natural choice for a screening study on pooled samples, which can suggest candidate genes or gene families to be studied by more sensitive techniques. The 5'-nuclease assay is such a more sensitive technique, with a larger dynamic range and less demand for sample material than the microarray technique. When differential expressions are not large monogenetic, this sensitivity may be required. In this paper we shall discuss statistical models and methods for 5'-nuclease studies. The discussion will largely be based on a concrete study, where 16 proposed susceptibility genes (candidate genes) for schizophrenia were tested on cortical brain samples (autopsies) from 110 deceased individuals (Castensson et al., 2002). Some of these genes were suggested from microarray studies.

In a 5'-nuclease assay the mRNA expression level in a specimen is measured by real-time PCR (see e.g. Gibson et al, 1996, Heid et al, 1996). This means that the increasing fluorescence intensity of accumulating PCR product is recorded during the PCR reaction until a prespecified level is attained. The time (in number of cycles) is measured and used as a measure linear in the log concentration of cDNA (copyDNA). A calibration by a dilution series yields the coefficients of this linear function. The amounts of cDNA and mRNA are proportional, and therefore the estimate of the amount of cDNA can be used as an mRNA expression level for comparisons between samples, for a fixed gene.

The cDNA is generated from the mRNA in a reverse transcription (RT), which yields an RT master plate, with the different specimens in separate positions (wells) on the plate. Replica plates are produced by aliquoting diluted contents of the RT master plate into TaqMan optical plates, one for each gene under consideration. On the TaqMan plates, the cDNA amounts are amplified and measured by real-time PCR technique, as described above.

Many applications of 5'-nuclease assay concern supervised or unsupervised classification of individuals in groups (discrimination or cluster analysis). This was not the aim of the schizophrenia study, which concerned possible differences in gene expression between two groups of individuals (patients and controls). Differences to look for could be of several types. It might be a difference in mean value between the two populations, and/or

a difference in variance. Only the former type of difference would be detectable if pooled samples were compared. Also, it is probably unrealistic to expect a whole population of patients to differ from controls in a particular gene. Hence, another possibility to look for is that only a subgroup of a heterogeneous patient population shows the differential expression. This would affect both mean value and variance for the sample.

Measurements on biological material, such as brain samples, will of course show a substantial natural variation between individuals. As much as possible of this variation must be eliminated by experimental and statistical techniques, in order to increase the power of the study, that is to increase the chances that a possible difference between sick and healthy individuals will be revealed. Large samples of individuals may compensate for variability between individuals. To some extent this is necessary, in order to reduce the risk of false signals caused by particularities in the two samples compared; in fact, many of the announcements in the literature come from studies with quite small sample sizes. However, large samples are expensive and not very efficient for reducing variation, so more sophisticated methods must also be utilized to reduce the need for very large samples. This was so also in the schizophrenia study, despite the fact that the number of individuals in it was higher than in most other studies of similar type, comprising 55 patients and 55 controls. If no covariates are used and the standard deviation between individuals is σ, a difference in mean values must be of a magnitude as high as $2\sigma\sqrt{2/55} \approx 0.4\sigma$ in order to make the chance 50% to detect on the 5% significance level that there is a difference in means. If this difference only affects a subgroup, the chances will be smaller, of course. Fortunately, as we will see, covariates allow the inference to be made much more precise.

Experimental and observational study designs have a role in this context, for example to control for possibly confounding factors such as sex, age and brain bank. A particular aspect to consider is the allocation of specimens to plates when such a large number of brain specimens is available that there is not room for all of them on a single RT plate. This was the case with the brain sample material, with its 110 individuals and replicate samples of tissue for each individual. Design aspects are discussed in Section 2.

The introduction of explanatory covariates is a well-known statistical technique for increasing precision by reducing the influence of natural variation. Besides the factors mentioned above, variation in the time post mortem when the autopsy specimen was taken contributes substantially to the variability between individuals. Measurements on reference genes, known or expected to be unaffected, are other important covariates, see below. In the schizophrenia study, the following factors were available as covariates. Other studies are likely to have covariates of similar type.

- Categorical covariates: Sex and brain bank. Autopsies came from three brain banks of quite different age distributions and presumably also other systematic differences (racial composition, medical diagnostic criteria).

- Quantitative covariates: Age and time post mortem, but also measurements on reference genes (to be further discussed below)

The material originally comprised three specimens of brain tissue from each individual, taken from the same part of the brain. They will be regarded as replicate observations on the same individual. Replicates make it possible to reduce part of the variation, by averaging over them. It will be seen later that a dominating portion of the variation

between specimens from different individuals is due to the variation between individuals, so the replicates do not have a particularly large effect on the precision. However, the replicates are beneficial in other ways. They will be needed in the design, as illustrated in Section 2, and they help to identify gross errors and to distinguish these from values for extreme individuals.

Statistical models are introduced in Sec. 3. Based on these models, some statistical inference aspects are discussed in Sec. 4.

2 Design aspects

The design of the study involves aspects from both observational and experimental studies. If differences in mRNA levels are seen between the two groups of individuals, it is of course essential that this difference can be attributed to the fact that one group has the disease and the other group has not. To allow such conclusions the selection of individuals should preferably be balanced with respect to possible confounders, so that the two groups do not differ substantially in other important respects than the disease. In the schizophrenia study, the two groups of individuals were in fact matched within each combination of brain bank and sex. This means that these two factors will not be disturbing confounders, but only help to make possible conclusions more general than if they had not been varied at all.

Given the two samples of individuals, it remains to allocate the samples on (master) plates. Each plate has 96 wells (=positions), but some wells are reserved for various standards and checks (Negative Amplification Control, Negative Template Control). Therefore the available number of wells per plate will not be much more than 80. Plate is an experimental block factor, to some extent analogous with the array factor in micro-array experiments. It is certainly important to allow differences between plates, meaning that the overall level of the signal could differ between different replica plates. We do not recommend estimation of plate effects from internal controls only; rather the role of the latter will be in checking for anomalies. In principle we could also imagine position effects on the plate, for example in the form of row or column effects. Our experience from previous TaqMan experimental work had not pointed towards any position effects, so in the schizophrenia study such effects were assumed to be absent. However, even if row and coilumn effects are not believed to matter, it would be wise to make the design row/column-balanced.

Since we had 110 individuals in the schizophrenia study, they could not all be allocated to the same plate, but an incomplete design had to be chosen. To some extent this makes the statistical analysis more complicated, but it is a cost worth paying for the advantage of having many individuals represented. The design should preferably be chosen to be (at least approximately) balanced. Since the variation between individuals is crucial, it is essential that all individuals are used, in some way or another. On the other hand, it was found less important that all three samples per individual were used, in the first round. After discussing several balanced incomplete designs, we decided to use three plates and only two samples per individual. This allowed more genes to be tested at the same total cost. The third sample was kept in reserve, waiting for a promising gene, when it could

be used for confirmation purposes, if desired.

More precisely, the following balanced incomplete design was chosen: 54 ($= 3 \times 18$) patients and 54 controls, with two samples per individual, were allocated to three plates such that each plate received 36 ($= 2 \times 18$) of the 54 individuals of each type. This was done in a balanced way, which means that each subgroup of 18 patients was combined once with each of the other two subgroups of 18 patients, and likewise for controls. The allocation to subgroups was performed such that the approximate balance in the other factors (sex and bank) was retained. The odd 55th individual was added to one of the three groups, making it size 19 instead of 18.

Design aspects are important also in microarray studies, see Kerr & Churchill (2001a, 2001b). However, an important difference is that microarrays have typically two varieties but many genes on the same array, where 5'-nuclease plates have many individuals but only one gene per replica plate.

3 Statistical modeling aspects

3.1 A basic RANCOVA model

First, let us consider a single gene for the individuals of the control group, so that by definition its mRNA expression is not affected by the presence of the disease. Partially expressed in terms of the schizophrenia study, we found the following nested linear model reasonable to describe the variation in the data for a single gene, taking covariates and design effects into account and measuring the fluorescence y on a log scale:

$$y_{hij} = \mu + \alpha_h + \beta' u_{hi} + \gamma_{k(hij)} + \delta_{hi} + \epsilon_{hij} \tag{1}$$

with indices
h = categorical covariate index (sex, brain bank);
i = index for numbering individuals within group h;
j = index for numbering samples within individual;
$k = k(h, i, j)$ = index for the plate number to which sample (h, i, j) is allocated.
The variables and effects in (1) are characterized as follows:
μ = constant, the overall mean value (grand mean);
α_h = main effects for the systematic variation between groups;
u_{hi} = vector of individual characteristics such as age and time post mortem, common for all samples from an individual;
β = regression parameter vector for the u_{hi} (age and time post mortem) effects. These effects are believed and assumed not to depend on group (h), and to be reasonably linear in suitable covariates;
γ_k = plate effect, allowing differences between the three plates. These effects could be thought of as randomly generated. However, since these effects are of interest only in an adjustment process, they will here rather be regarded as fixed effects;
δ_{hi} = individual deviation from group mean α_h, reduced for the covariates effect $\beta' u_{hi}$, that is "individual pure random variation" characterized by a variance σ^2_δ;
ϵ_{hij} = contribution from random variation between brain samples within individual and

from pure random experimental and measurement errors, characterized by a variance σ_ϵ^2.

Model (1) may be called a random-effects analysis of covariance model (RANCOVA; Longford, 1993, Ch. 2). The only additional complication so far to a typical RANCOVA model is the plate effect. To model log-values of fluorescence intensities additively is standard, also in microarray data analysis (Kerr & Churchill, 2001b). In the schizophrenia study, the assumption of constant variances σ_δ^2 and σ_ϵ^2 was seen to be largely but not fully correct. The variability in log-values y tends to be higher when the intensity is low. Another problem with low intensities is that there is a detection limit for low intensities. Some individuals have overall relatively low intensities, and some genes have lower intensities than others. Treatment of non-constant variance and measurements below the detection limit will be discussed in Sec. 3.3.

With data from several genes jointly, the model becomes multivariate. We still use expression (1), but now we interpret y_{hij} as a vector of the corresponding dimension. Likewise, α, γ, δ and ϵ are vectors, and β is a matrix of regression coefficients. We must expect and therefore allow (and utilize) correlation between genes in their inter- and intra-individual variation, i.e. correlation between components in the vector δ as well as in ϵ. In other words, for δ and ϵ we have to generalize the variances σ_δ^2 and σ_ϵ^2 to variance–covariance matrices Σ_δ and Σ_ϵ, respectively, of the same dimension as the number of genes considered jointly.

Turning now to the patient group, the same model (1), with the same parameter values, should hold as long as the model is restricted to genes not differentially expressed between patients and controls. A disease effect in a candidate gene might appear in some or (in the extreme case) all individuals of the patient group. A constant disease effect could be represented by a shift in μ, but it is unlikely that an effect would only or even dominantly affect the mean value. A disease effect in a candidate gene might interact with other factors in the study, for example its presence/absence directly with sex, or its magnitude with age. By dependence on the individual overall intensity level, the disease effect might be seen to interact indirectly with reference gene intensities. It does not appear sensible to allow all imaginable such interactions explicitly in the model.

The following might be a more reasonable modeling procedure, at least for an exploratory analysis. Let the systematic parameters μ, α, β and γ be as defined for the controls, and estimated from the controls data. Allow the patients to have a different distribution for δ than the controls, including in particular a possibly different mean value. We shall express this by replacing δ_{hi} by δ_{hi}^* for the patient group and a candidate gene. That is, corresponding to model (1) for the controls, the model for the patients is

$$y_{hij} = \mu + \alpha_h + \beta' u_{hi} + \gamma_{k(hij)} + \delta_{hi}^* + \epsilon_{hij} \,, \tag{2}$$

with a possibly different distribution for δ_{hi}^* than for δ_{hi} when the gene under consideration is a candidate gene, and with interaction allowed between δ_{hi}^* and other factors. There is a possibility to think of δ_{hi}^* as the combined result of two effects, $\delta_{hi}^* = \delta_{hi} + \eta_{hi}$, where δ_{hi} for these individuals is a hypothetical (contrafactual) random effect that would have been realized, had the patient not got the disease, and η_{hi} is an individual disease effect, whose mean value and distribution we are interested in. This interpretation involves conceptual difficulties, however, and will be avoided in the sequel, but is close to the

natural temptation to interpret an apparently large value of an individual δ^*_{hi} as showing an individual disease effect.

For a given candidate gene, it is thus of primary interest to compare the distributions of δ^*_{hi} and δ_{hi} over the populations of patients and controls, respectively. However, δ^*_{hi} and δ_{hi} cannot be separated from the measurement noise represented by ϵ. For simplicity, let us provisionally not bother about the fixed effects in (1) and (2), whether they are known or estimated, but consider μ, α_h, β and γ_k as given. When we adjust the observed y-values, forming residuals by subtracting the fixed effects from $\bar{y}_{hi.}$, we will see the sample distribution of $\delta + \bar{\epsilon}$ for the control group and of $\delta^* + \bar{\epsilon}$ for the patient group. These two distributions are of course equal precisely when the distributions of δ^* and δ are equal. The distributions should be compared, to see if there is any systematic difference or extra variability in $\delta^* + \bar{\epsilon}$, that can be seen above the background level of variability of $\delta + \bar{\epsilon}$. When a disease effect has been established for a candidate gene, we can next look for possible interaction with other factors, in order to understand the effect better.

It is a further complication that the parameters of model (1) must be estimated, and due consideration paid to the degrees of freedom lost in this process. Note that by reduction to average values per individual we retain all information for estimation of the fixed parameters except the plate effects γ. This is seen by observing that the full data of model (1) are equivalent to the averages $\bar{y}_{hi}$ over replicates jointly with the differences $\triangle y_{hi} = y_{hi1} - y_{hi2}$ (or more generally the corresponding deviations from the averages), following the model equations

$$\bar{y}_{hi} = \mu + \alpha_h + \beta' u_{hi} + \bar{\gamma}_{k(hi)} + \delta_{hi} + \bar{\epsilon}_{hi}, \quad (3)$$

$$\triangle y_{hi} = \triangle\gamma_{k(hi)} + \triangle\epsilon_{hi}, \quad (4)$$

with analogous average/difference definitions for $\bar{\gamma}_{k(hi)}$, $\bar{\epsilon}_{hi}$, $\triangle\gamma_{k(hi)}$, and $\triangle\epsilon_{hi}$.

The inter-individual information from model part (3) about the plate effects γ is likely to be much less than the intra-individual information from part (4). This can be quantified as follows. In least squares estimation of plate effects under a balanced incomplete design as proposed above, with three plates and two replicates per individual, the intra-individual differences yield a plate effect estimation variance of size $4\sigma^2_\epsilon/n$, with n as the number of individuals, whereas after averaging, the variance would be $3(2\sigma^2_\delta + \sigma^2_\epsilon)/n$, with a contribution from the component of variance between individuals, σ^2_δ, that is likely to make the latter variance substantially larger than the former one.

3.2 Correlation results

The multivariate version of model (1) was fitted to data from 12 genes, two of which were going to have the roles of reference genes and the others were the ten earliest candidate genes. The covariances and correlations between genes were estimated with respect to variability between and within individuals, Σ_δ and Σ_ϵ, respectively. The correlations were high, indeed, in particular in Σ_δ. A MANOVA analysis primarily gave estimates of Σ_ϵ and of approximately $\Sigma_\epsilon + 2\Sigma_\delta$. An estimate $\widehat{\Sigma}_\delta$ of Σ_δ was obtained by subtraction. A PCA on the corresponding correlation matrix estimate showed that its first principal component explained 94% of the total variability. This PC was quite close to proportional to a vector

of ones. This means that the variability between individuals could be described by a common factor influencing all genes similarly. As a consequence, the technique of using reference genes should be able to increase precision substantially.

3.3 Covariates from reference genes

The natural background variability of $\delta+\bar{\epsilon}$ is likely to be large, mostly because of biological heterogeneity, both of natural type and of more or less artificial character (e.g. different medical treatments, different tissue storage times). By utilization of some (highly) correlated reference gene, or a couple of such genes, the precision can be increased considerably. We reduce the variability in $\delta + \bar{\epsilon}$ for the candidate gene by using the reference gene to adjust for the predictable part of $\delta+\bar{\epsilon}$. This is done by regressing the candidate gene values on reference gene values, and in this way replacing 'original' δ-values by their residuals in this regression. In other words, we use the reference genes to form additional covariates in models (1) and (2). This causes some statistical complications.

A candidate gene and a reference gene can both be represented by model equations of type (3–4), with response variables represented by notations y and x, respectively. Since averages and pairwise differences are mutually uncorrelated, also between genes, we need only consider regression of average on average and of difference on difference,

$$\bar{y}_{hi} = \mu + \alpha_h + \beta' u_{hi} + \bar{\gamma}_{k(hi)} + \theta_b \bar{x}_{hi} + \delta_{hi} + \bar{\epsilon}_{hi}, \quad (5)$$

$$\triangle y_{hi} = \triangle\gamma_{k(hi)} + \theta_w \triangle x_{hi} + \triangle\epsilon_{hi}. \quad (6)$$

With x inserted on the right hand side of the model, μ, α, β and γ change their meanings and values from those in model (3–4), since the model for x has the same explanatory variables as the model for y. This need not be a problem, however, since we are not interested in these parameters per se. Of more importance is to note that the regression coefficients θ_b and θ_w in the x-terms are not likely to be the same 'between individuals' in (5) as 'within individuals' in (6). This has consequences for the plate effect estimation when blocks (plates) are incomplete. In analogy with the discussion ending Section 3.1, the difference data following (6) will provide a relatively precise estimate of the linear combination $\gamma^{(y)} - \theta_w\gamma^{(x)}$ of the plate effects $\gamma^{(x)}$ and $\gamma^{(y)}$ in x and y. On the other hand, analogously, the term involving plate effects in model (5) is based on $\gamma^{(y)} - \theta_b\gamma^{(x)}$. Hence, if $\theta_b \neq \theta_w$, the 'within' data in (6) do not provide an estimate of the plate effect desired in (5).

Our experience from real data is that θ_b and θ_w typically differ substantially. For most genes in the schizophrenia study, θ_b was about 50% higher than θ_w (albeit one exceptional gene showed remarkably little variation between individuals and had a θ_b-value close to θ_w). Since the utilization of reference genes is so important for the precision, we conclude that we must be content with the less efficient 'between individuals' estimate of the average plate effects appearing in model (5). Even though they are not as precise as could be desired, note that if we are interested in the average disease effect on a candidate gene, this average will be unaffected by possible plate effects in x as well as in y, by the balance properties of the design.

Note that the assumption $\theta_b = \theta_w$ should not only be generally distrusted, it would also be more complicated to fit models (5) and (6) jointly, with $\theta_b = \theta_w = \theta$, than model

(5) separately. This is because there would be information about θ to combine from (5) and (6). The estimates of θ_b and θ_w should optimally be weighed together by weights depending on the ratio σ_w/σ_b, which must be jointly estimated from the same data, see Longford (1993, ch.). If θ_b and θ_w are the same, we would gain in precision for their common value. This would be important if, by chance, their estimates differed much, but on the other hand, if the two estimates differ substantially, we should distrust the assumption that the parameters are the same.

The role of the pairwise differences, by model (6), will thus not be for direct inference about the disease effects. Their role in the inference is rather in outlier detection. Gross errors must be reckoned with, for example caused by DNA contamination, but usually such errors are signalled by a too large pairwise difference. Typically in such cases, one of the response values is relatively close to the estimated or predicted value, and then it will be clear that the other one is the outlier.

For low-expressing genes, frequent response values will already initially be missing, because they are too low to be detected, that is a case of left censoring. For the schizophrenia data it was difficult to specify a detection limit in each case, and as a pragmatic way of avoiding the complications with left-censored data, such missing values were replaced by the low value $\log 1 = 0$, corresponding to a single observed copy. Likewise, estimated or predicted values below this limit were replaced in the same way. This reduced some residuals, but this effect was to some extent positive, because the random variability on the log-scale was somewhat larger for very weak responses than for moderate or strong responses. See the upper diagram in Figure 2 for an illustration.

For genes with a wide span in response values, it is perhaps generally advisable to let the variance in model (3) be some negative power or some other function of the expected value. We would then use a recursively weighted regression fitting method, in which the weights are updated in each step. The problem is to determine the form of this variance function. No such generalised variance model was tried in the schizophrenia study.

4 Statistical inference aspects

4.1 Identification of differentially expressed genes

As stated in Section 3.1 above, we want to compare the distribution of δ^* for patients with the reference distribution of δ for controls. We here concentrate on testing for a difference in mean values, because a systematic difference between patients and controls in a particular gene is regarded as being of primary interest. Note that the unvoidable error term $\bar{\epsilon}$ will not affect the mean difference, but only the test power.

Under the null hypothesis of no disease effect, both patients and controls should follow the same covariance analysis (ANCOVA) model, formula (5). A conventional and simple test for a null disease effect would be a standard ANCOVA t-test, testing for equality of intercepts when patients and controls are allowed to follow models with different intercepts, but all other parameters in common. This test can be carried out by use of standard software. However, the test would be optimally efficient only under the assumption of a constant disease effect. With a highly variable disease effect, in particular with only a

subgroup affected, such a test would not be very efficient at all, because it would pool the two groups together in all other parameters and it would use the pooled residual variance. In a simple two-sample situation without covariates, the standard two-sample t-test suffers from the same inefficiency for detecting when only a subgroup is affected. Alternative tests will be discussed below.

When we go from testing a null hypothesis to constructing confidence intervals for the mean disease effect, we must specify the alternative model, and in practice we will then be more or less forced to rely on such a vulnerable assumption of a constant disease effect η. The other fixed effects will be assumed to be the same for the two groups, and the residual variance to be constant (according to the relation $\delta^*_{hi} = \delta_{hi} + \eta$). The magnitude of a possible disease effect is certainly of importance, but a fully meaningful confidence interval for the population difference also requires representative samples from two well-defined populations of interest. A significance test is less demanding.

Now, consider testing the null hypothesis against alternatives allowing not only a mean value shift but also an increased variance, $\mathrm{Var}(\delta^*) > \mathrm{Var}(\delta)$, and allowing possible interactions with other factors in the model. Then it can be more efficient to let the model parameters be estimated from the control group alone, so that they are unaffected by possible variability in disease effects. Two different approaches to the problem will here be seen to yield the same test statistic.

For the sequel it will be convenient to simplify the description of the model, by imagining dummy covariates for the parameters α_h and bringing them into the regression, at the same time omitting the index h and the bar over y and x, and incorporating all regression into the term $\theta' x$, where θ and x are now vectors. Under H_0, the model can then be written, for controls and patients respectively, as follows:

$$E(y_i^{(c)}) = \mu + \theta' x_i^{(c)} \tag{7}$$

$$E(y_i^{(p)}) = \mu + \theta' x_i^{(p)} \tag{8}$$

$$\mathrm{var}(y_i^{(c)}) = \mathrm{var}(y_i^{(p)}) = \sigma^2 \,. \tag{9}$$

Directing the interest towards tests for a difference in μ-values between patients and controls, we are led to a test statistic of the form

$$\frac{\bar{y}^{(p)} - \bar{y}^{(c)} - \hat{\theta}'(\bar{x}^{(p)} - \bar{x}^{(c)})}{\{\hat{\sigma}^2/n_p + \hat{\sigma}^2/n_c + (\bar{x}^{(p)} - \bar{x}^{(c)})'\widehat{\mathrm{var}}(\hat{\theta})(\bar{x}^{(p)} - \bar{x}^{(c)})\}^{1/2}} \,, \tag{10}$$

where $\hat{\theta}$ and $\hat{\sigma}$ are suitable estimates of θ and σ, respectively, and the denominator represents the standard error of the numerator.

If θ and σ are estimated from the joint patients and controls data, we have the standard ANCOVA t-test statistic, mentioned above. If θ and σ are estimated from the controls alone, to be written $\hat{\theta}_c$ etc., we also have a t-test statistic, but with reduced degrees of freedom. The latter test may be more powerful under such alternatives where the variance σ^2 is higher for patients, and under alternatives in which θ differs betwen patients and controls due to the disease. If n_c is large, the resulting loss of degrees of freedom in $\hat{\sigma}^2$ is unimportant, but the higher variance in $\mathrm{var}(\hat{\theta}_c)$ may possibly cost more, unless $\bar{x}^{(p)} - \bar{x}^{(c)}$ is small.

A more predictive approach, leading to the same test statistic as the last one, is to compare the observed patient responses with their predicted values, as predicted under a model fitted to the controls,

$$y_i^{(p)} - \hat{y}_i^{(p)}, \tag{11}$$

where $\hat{y}_i^{(p)} = \hat{\mu}_c + \hat{\theta}_c' x_i^{(p)}$. The average of these prediction errors is

$$\bar{y}^{(p)} - \hat{\mu}_c + \hat{\theta}_c' \bar{x}^{(p)} = \bar{y}^{(p)} - \bar{y}^{(c)} - \hat{\theta}_c'(\bar{x}^{(p)} - \bar{x}^{(c)}), \tag{12}$$

that is the numerator in (10) above, with $\hat{\theta}$ from the controls. Hence, we obtain the same test statistic again.

When $\hat{\theta}_c$ is used in the statistic (10), it can be difficult to compute the denominator from standard package output. A simpler alternative is to approximate the denominator by its expected value over a distribution for x assumed to be the same for patients and controls:

$$\begin{aligned} E\{(\bar{x}^{(p)} - \bar{x}^{(c)})' \widehat{\mathrm{var}}(\hat{\theta}_c)(\bar{x}^{(p)} - \bar{x}^{(c)})\} &= \mathrm{tr}\ E\{\widehat{\mathrm{var}}(\hat{\theta}_c)(\bar{x}^{(p)} - \bar{x}^{(c)})(\bar{x}^{(p)} - \bar{x}^{(c)})'\} \\ &\approx \sigma^2 \left(\frac{1}{n_c} + \frac{1}{n_p}\right) \frac{\dim(x)}{n_c}, \end{aligned}$$

neglecting second order terms in $\dim(x)/n_c$. This yields the approximation for the denominator of formula (10) as the square root of

$$\hat{\sigma}^2 \left(\frac{1}{n_c} + \frac{1}{n_p}\right) \left(1 + \frac{\dim(x)}{n_c}\right). \tag{13}$$

This approximation was used in Castensson et al. (2002). It will tend to be conservative if the x-variables are balanced by design. It is instructive to compare (13) with the corresponding approximation when the full data estimator $\hat{\theta}$ is used. With this estimator, the last factor of (13) is reduced to

$$\left(1 + \frac{\dim(x)}{n_c + n_p}\right).$$

This is the gain in precision from using all data in the estimation of θ, valid under the null hypothesis.

In order to look for a difference between patients and controls in a more exploratory way, we could compare a histogram of the prediction errors (11) for the patients with a histogram of the residuals for the controls. However, a direct such comparison has several defects. Residuals are not i.i.d., nor are prediction errors, since their variances depend on the regressor x and they are all correlated in their dependence on $\hat{\theta}$ (or $\hat{\theta}_c$). Furthermore, residuals have smaller variances than prediction errors, not only for a fixed x but typically also on the average. So if histograms should be compared, they should be histograms for variance-standardized residuals and variance-standardized prediction errors, respectively. This is illustrated in Figure 1 for those two genes, denoted HTR2C and MAOB, which were classified as showing significant differences between controls and

patients in the schizophrenia study. However, the variance standardization did not change the picture much for these genes.

We have also found it informative to display data in the form of a scatter-plot of observed responses against predicted values, $y_i^{(p)}$ against $\hat{y}_i^{(p)}$, over-laid the analogous plot for the controls. See Figure 2 for illustrations, for the same two genes as in Figure 1. The HTR2C gene is seen to be more problematic than the MAOB gene, by showing a large spread in responses and predicted values, with several values below the detection limit, both for patients and controls. This was discussed near the end of Section 3.3.

Since the predicted values are linear in the covariates x, Figure 2 type of plots reveals the extent to which y is explained by x. The reader can judge residuals and prediction errors with respect to the position of the predicted value $\hat{y}$, for example informally putting less weight on very low predictions or response values. Not only do outliers in y given x stand out, vertically from the line $y = \hat{y}$, but also do outliers in x, via both y and $\hat{y}$, along the line $y = \hat{y}$. More diagrams of this type are found in Castensson et al. (2002). With few or a moderate number of genes, the plot for each gene may be looked at, whereas if the number of genes is very high, consideration is likely to be confined to those genes which are to some degree significant.

When regarding Figure 2 it should be kept in mind that residuals are slightly less random than prediction errors. An attractive alternative plot, in which controls and patients are more on the same footing, is formed if, for the controls, we plot y_i against leave-one-out cross-validated $\hat{y}_i$. This will be much more computer intensive, however.

With several genes tested we must pay consideration to the multiple testing effects, of course. Castensson et al. (2002) used simple Bonferroni-corrected significance levels. Since the number of genes studied here is of a much smaller magnitude than in microarray studies, the mass significance problem is not at all as severe as with microarrays, cf. Dudoit et al. (2002) for a recent discussion of p-value adjustment methods in the latter case. Even if the number of genes is not very high, a quantile–quantile-plot could be suggested as an informal way to judge the lower p-values. Also, even if each gene has been selected for testing on its own merits, we are not likely to be helped by an outcome signalling a large number of moderately significant genes, but only the most significant ones will then be of much further interest. An alternative to ranking the significant genes with respect to their p-values is to order and evaluate them according to their average effects, (12), normed by their corresponding residual standard deviation, $\hat{\sigma}$, as an indication of the natural level of variation of y given x. To norm by the total variation in y could be seriously misleading, since there are several irrelevant sources for the apparent variability seen between individuals.

4.2 Multivariate aspects for multiple genes

Multivariate T^2 test extensions of the univariate t-tests from Section 4.1 are immediate. They will be more powerful tests than the univariate ones if most genes contribute a little to the difference between the two groups. However, if the groups differ only in a few genes, those few effects might be so diluted by the other genes that the difference is not seen in the multivariate test. Also, note that the aim of the assay is not to demonstrate a significant difference between patients and controls, but to find those genes which possibly

show a difference. The multivariate aspect concerns whether patients showing relatively high (positive or negative) effects in one gene tend to show corresponding effects in some other gene, to a higher extent than the controls. This would be expressed as a differential co-regulation. To this end, we could look for the direction, that is the linear combination over genes, which maximises the ratio of mean squares in the standardized residuals, with the patients mean square in the numerator and the controls mean square in the denominator. This is the eigenvector with largest eigenvalue of the corresponding ratio of sums of squares and products matrices.

If we neglect the difference between true mean values and estimated mean values, with estimates based on the controls, this would also be an invariant test statistic for testing equality of two covariance matrices, but here used to identify a combination of mean value effects and covariance effects. However, as taken over the whole set of genes, this criterion turned out to be useless. Too much, the criterion appeared to favour directions of low variation by chance in the controls.

When attention was restricted to the relatively few significant genes, it was easier to make relevant comparisons of the correlation structure between patients and controls. In the schizophrenia study only two genes were identified as significant. The variance-standardized residuals turned out to be uncorrelated for the controls ($r = -0.02$), but the corresponding standardized prediction errors for the patient group were substantially negatively correlated ($r = -0.45$). This corresponded to one high eigenvalue in the matrix ratio of sums of squares and products matrices mentioned above. Its eigenvalues were 2.9 and 1.0. The difference in covariance structures is clearly seen in a two-dimensional plot of the variance-standardized residuals, Figure 3. The scientific conclusion is that there is largely a joint reason behind the increased values of MAOB and the reduced values of HTR2C for the patient group. In other words, the two genes are co-regulated for the patients exclusively.

5 Discussion

In this paper we have discussed aspects of design and statistical inference in 5'-nuclease assays. An analogous discussion concerning design and analysis of microarray studies has been provided by Kerr & Churchill (2001b). The design problems are different, as remarked in Section 2. In the analysis phase, Kerr & Churchill also fit linear (ANOVA) type models for experimental factors. However, there are major differences. For example, in microarray analyses heteroscedasticity between genes is a problem, because gene is a factor with all its levels represented on the same plate/array, whereas in 5'-nuclease assays the genes are represented by components of a multivariate vector, for which it is just natural to allow an arbitrary covariance matrix.

Multivariate linear models with a nested error structure (RANCOVA) has a natural role in the first stage of modelling. The high mutual correlation between genes in their random model errors motivated the use of reference genes, to form additional covariates. However, this also required some care, in order to avoid systematic errors due to different regressions between and within individuals.

The statistical inference problem appears to have some nonstandard features, as com-

pared with ordinary ANCOVA type situations. In particular, a differential effect need not be seen in or is not even likely to be seen in the whole patient population, but rather only in a patient subgroup, if it is a genetic effect. For the latter reason reason, a somewhat more powerful modification of the standard ANCOVA t-test was proposed, in which parameters are estimated from controls only.

Acknowledgements

Thanks are due to Lina Emilsson for her part in the experimental work and to Niclas Sjögren for his cooperation in the early parts of the statistical work.

References

Castensson, A., Emilsson, L., Sundberg, R. & Jazin, E. (2002). Decrease of serotonin receptor 2C in schizophrenia brains identified by high-resolution mRNA expression analysis. Submitted.

Dudoit, S., Yang, Y.H., Callow, M.J. & Speed, T.P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111-139.

Gibson, U.E.M., Heid, C.A. & Williams, P.M. (1996). A novel method for real time quantitative competitive RT–PCR. *Genome Research*, **6**, 995-1001.

Heid, C.A., Stevens, J., Livak, K.J. & Williams, P.M. (1996). Real time quantitative PCR. *Genome Research*, **6**, 986-994.

Kerr, M.K. & Churchill, G.A. (2001a) Experimental design for gene expression microarrays. *Biostatistics*, **2**, 183-201.

Kerr, M.K. & Churchill, G.A. (2001b) Statistical design and the analysis of gene expression microarrays. *Genetical Research*, **77**, 123-128.

Longford, N.T. (1993) *Random Coefficient Models.* Oxford: Oxford University Press.

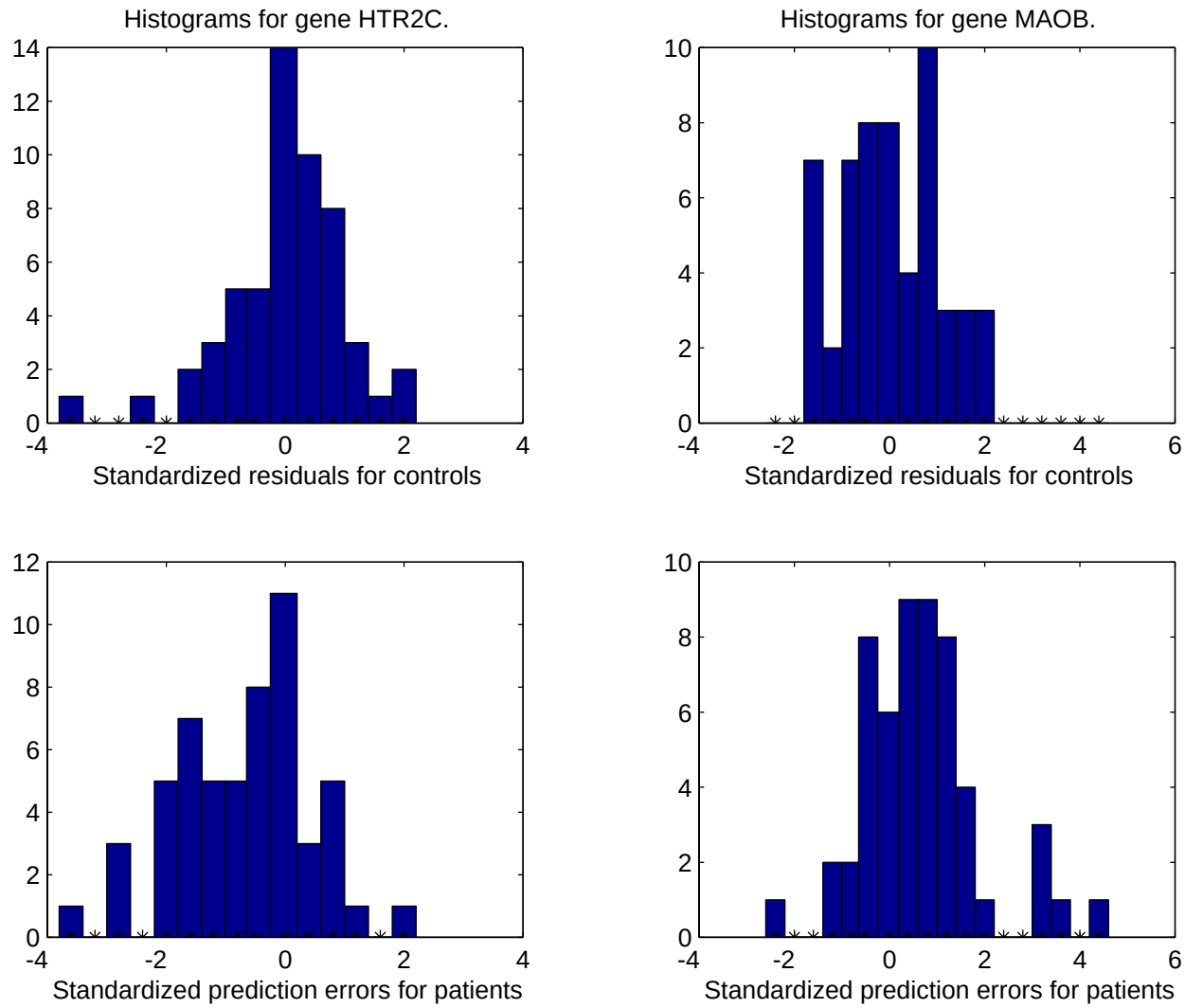

Figure 1: Histograms for genes HTR2C and MAOB of standardized residuals/prediction errors for controls/patients.

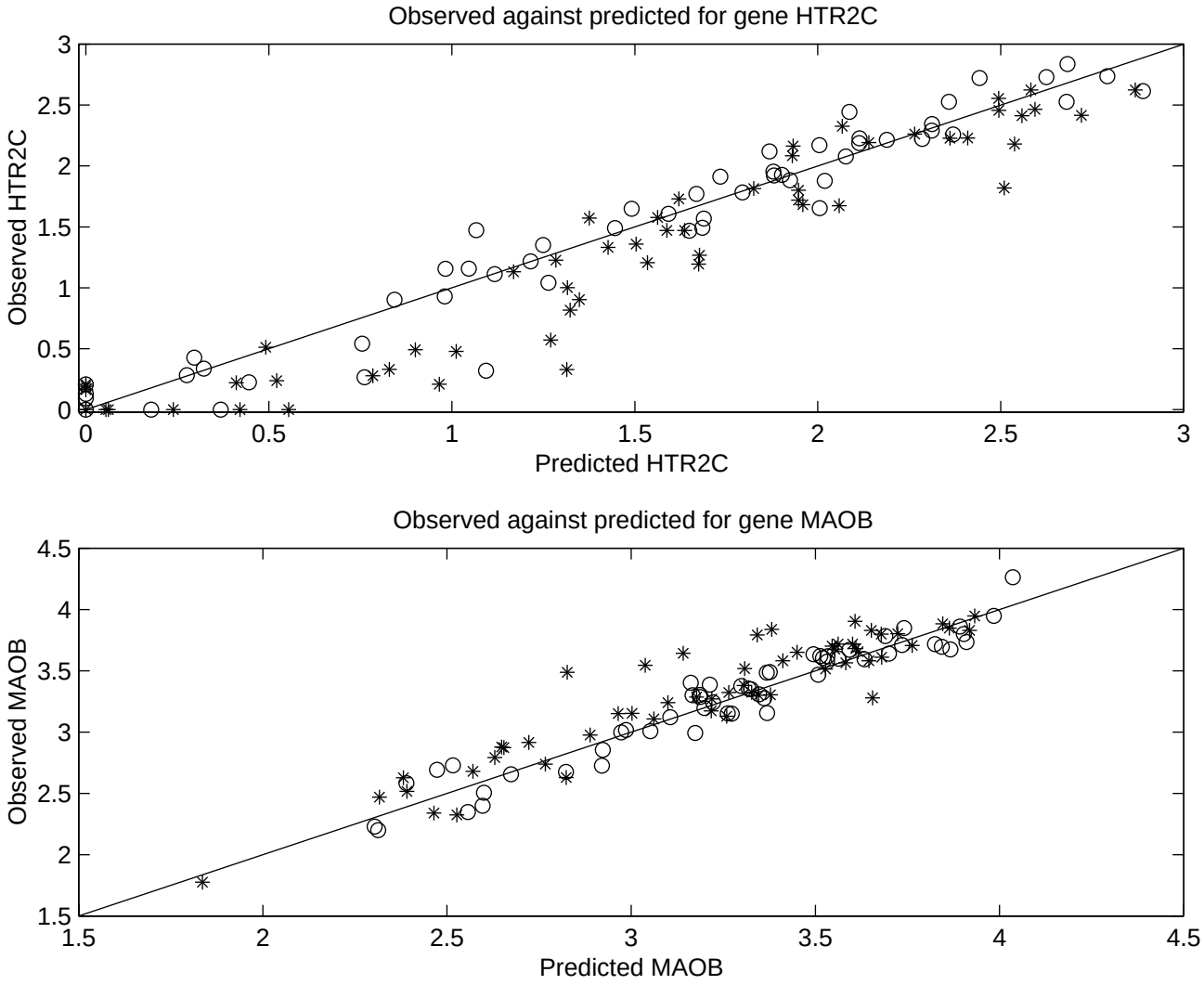

Figure 2: Observed against predicted for genes HTR2C and MAOB for controls/patients. Circles denote residuals for controls, and asterisks denote prediction errors for patients.

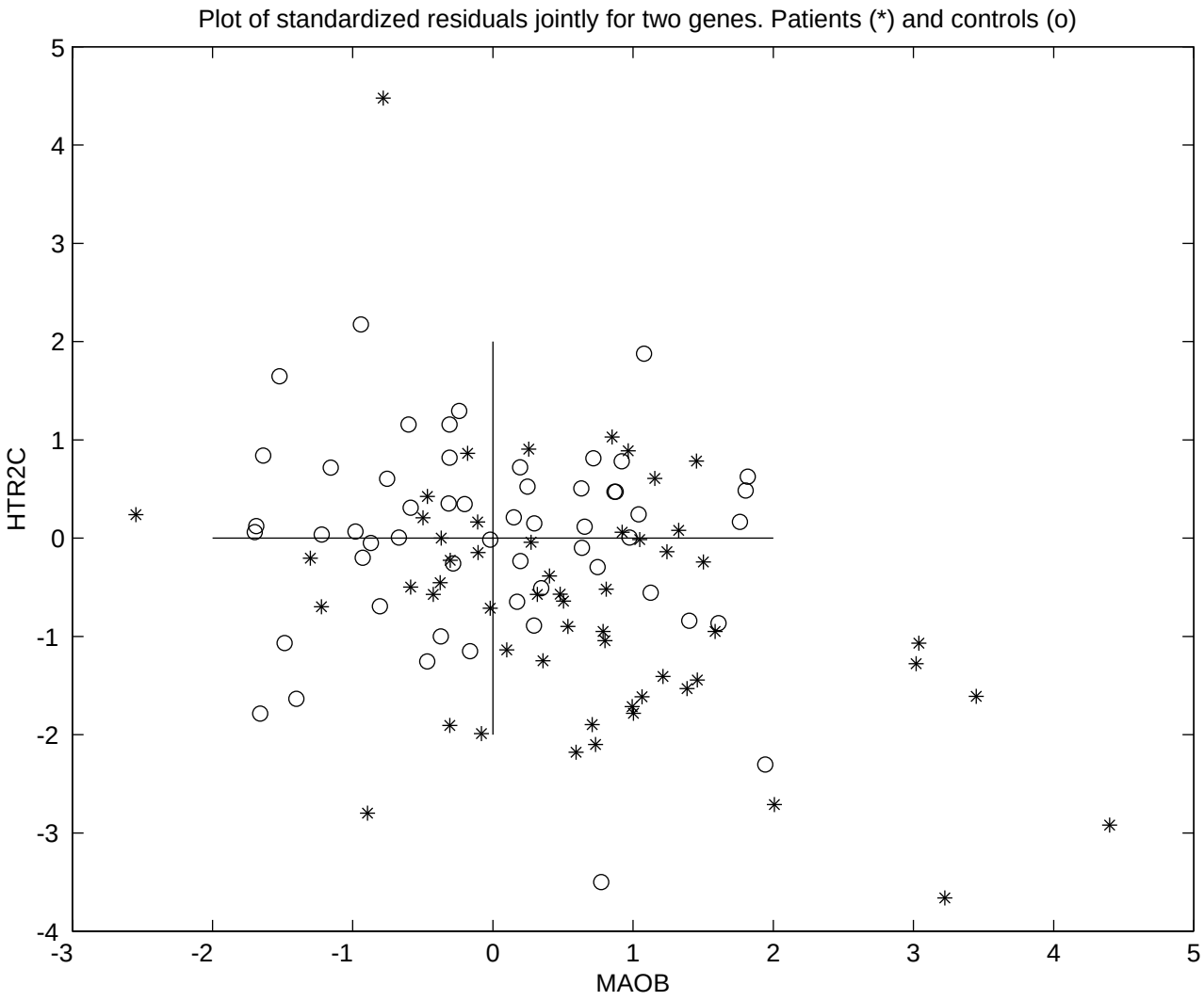

Figure 3: Plot of variance-standardized residuals/prediction errors for genes HTR2C and MAOB against each other. Circles denote residuals for controls, and asterisks denote prediction errors for patients. Correlations are -0.08 for controls and -0.42 for patients.