Bioinformatics Advance Access published July 7, 2010

ROAST: rotation gene set tests for complex microarray experiments

Di Wu, Elgene Lim, François Vaillant, Marie-Liesse Asselin-Labat, Jane E. Visvader and Gordon K. Smyth*

The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia, and The University of Melbourne, Victoria 3010, Australia.

Associate Editor: Dr. Jonathan Wren

ABSTRACT

Motivation: A gene set test is a differential expression analysis in which a p-value is assigned to a set of genes as a unit. Gene set tests are valuable for increasing statistical power, organizing and interpreting results, and for relating expression patterns across different experiments. Existing methods are based on permutation. Methods which rely on permutation of probes unrealistically assume independence of genes, and in any case are suitable only for 2-group comparisons with a good number of replicates in each group.

Results: We present ROAST, a statistically rigorous gene set test which allows for gene-wise correlation while being applicable to almost any experimental design. Instead of permutation, ROAST uses rotation, a Monte-Carlo technology for multivariate regression. Since the number of rotations does not depend on sample size, ROAST gives useful results even for experiments with minimal replication. ROAST allows for any experimental design which can be expressed as a linear model, and can also incorporate array weights and correlated samples. ROAST can be tuned for situations in which only a subset of the genes in the set are actively involved in the molecular pathway. ROAST can test for uni or bi-direction regulation. Probes can also be weighted to allow for prior importance. The power and size of the ROAST procedure is demonstrated in a simulation study, and compared to a representative permutation method. Finally, ROAST is used to test the degree of transcriptional conservation between human and mouse mammary stems.

Availability: ROAST is implemented as a function in the Bioconductor package *limma* available from www.bioconductor.org.

Contact: smyth@wehi.edu.au

1 INTRODUCTION

A gene set test is a differential expression analysis in which a set of putatively co-regulated genes is treated as a unit. A single p-value is evaluated for the set rather than evaluating individual p-values for invidual genes. Typically the gene set is chosen to represent a particular molecular pathway. In this way, gene set testing can simplify and organise differential expression analyses by focusing attention on larger and more biologically meaningful processes than individual genes. Gene set tests are valuable for relating expression patterns across different studies, even across different platforms or species (Manoli *et al.*, 2006). Gene set tests can be statistically more powerful than genewise tests as evidence is accumulated from many genes.

There are a number of gene set testing methodologies with different aims. This article is concerned with what might be called *focused* gene set testing, in which interest focuses on one or more gene sets of special relevant to the experiment at hand (Goeman *et al.*, 2004; Tian *et al.*, 2005; Dinu *et al.*, 2007; Jiang and Gentleman, 2007). This approach contrasts with what might be called *battery* gene set testing, in which a large database of gene sets is evaluated on a microarray data set, to see whether any sets stand out from the others (Saxena *et al.*, 2006; Efron and Tibshirani, 2007; Dørum *et al.*, 2009). Battery gene set testing was made popular by the Gene Set Enrichment Analysis (GSEA) method of Mootha *et al.* (2003) and Subramanian *et al.* (2005). In focused gene set testing, each set is evaluated on its own terms. In battery gene set testing, sets are evaluated relative to the other sets in the database. For this reason, methods designed for battery testing cannot be applied to individual gene sets.

Focused gene set tests can be classified into those which evaluate p-values by permuting samples and those which permute genes. Methods which permute genes are limited by the fact that they treat genes as if statistically independent. This assumption is usually suspect, particularly when the set is specifically chosen to contain co-regulated genes. Moreover, gene permutation p-values are very sensitive to inter-gene correlations, potentially leading to dangerously over-stated statistical significance (Efron and Tibshirani, 2007; Dørum *et al.*, 2009). This has lead some authors to conclude that statistically rigorous testing is only possible by permuting samples (Goeman and Bühlmann, 2007).

The need to permute samples to obtain p-values severely limits the experimental designs which can be analysed. Permutation is basically limited to two-group comparisons, with a moderate to large number of replicates in each group. Some authors have adapted permutation to the needs of one or two-way ANOVA designs. Adewale *et al.* (2008) and Hummel *et al.* (2008) suggested permuting sample labels while holding all covariates fixed except the covariate of interest. Oron *et al.* (2008) permuted sample labels within each level of a blocking factor. However these stratified permutation methods are still limited to special designs and moderate to large numbers of replicates.

*to whom correspondence should be addressed

Another possibility is to try to model inter-gene correlations explicitly using mixed linear models (Wang *et al.*, 2008). As genes operate with complex covariance patterns, this will be a simplification of the truth and, again, the method is restricted to experimental designs with a special structure.

We present a statistically rigorous gene set test which fully allows for gene-wise correlations while being applicable to almost any experimental design. We have in mind the sort of small but complex experimental designs which arise frequently in experimental medicine, which may have many experimental factors but only a few biological replicates. Instead of permutation, we use rotation, a Monte-Carlo simulation technology recently proposed for multivariate regression models (Langsrud, 2005). Rotation has also been used recently for a battery gene set testing (Dørum *et al.*, 2009), but here we use it for focused testing. Rotation can be viewed as analogous to fractional permutation. Since there is no limit on the number of rotations which can be done, the problem of granularity of p-values in small sample sizes is avoided. Rotation can be applied to the residual space of a linear model, and so utilizes all the available degrees of freedom regardless of the experimental design.

The next section of this article describes our statistical model for microarray data. Our implementation of ROAST (rotation gene set testing) allows for any experimental design which can be expressed as a linear model, and also accommodates array quality weights and correlations between samples. To ensure good behaviour in very small samples, ROAST uses empirical Bayes t-statistics for gene-level differential expression (Smyth, 2004). Then our approach to gene set testing is described. Alternative hypotheses are considered for gene sets in which all genes are expected to be regulated in the same direction and for those in which genes may vary in both directions. Scenarios are considered in which only a subset of the genes in the set actively contribute to the overall result. Different gene set summary statistics are developed which are appropriate for detecting different proportions of active genes.

Next, the numerical implementation of the ROAST algorithm is outlined. The power and size of the ROAST procedure is demonstrated in a simulation study, and compared to a representative permutation method. Finally, ROAST is used to test the degree of transcriptional conservation between human and mouse mammary stems

2 THE STATISTICAL MODEL

2.1 Data and gene set

We assume that an experiment or observational study has been conducted resulting in expression data on G probes in each of n target RNA samples. Different treatments, phenotypes or other characteristics are associated with the n samples, and we are interested in finding genes that are differentially expressed (DE) between the samples in some particular way. For example, the samples might be in two or more groups, and we want to find genes DE between two specified groups. Or we might want to find genes which show an interaction between two treatments. Or we might want to find genes which show a trend in a time-course experiments. Our aim is to accommodate quite general and arbitrarily complex experiments, so there is no limit on the number of treatment factors associated with the experiment, provided of course there is enough data available to estimate all the effects.

We assume that a particular set of probes or genes is of prior interest. This *a priori* specified gene set might represent a molecular pathway believed to be relevant to the experiment, or it might be a gene list from a previous microarray experiment hypothesized to be related to the current experiment. We want to test whether the gene set contains any probes which are DE. In some cases, the expected direction and magnitude of change may be specified in advance for individual genes. We are particularly interested in detecting sets which contain a good proportion, say 25–50% or more, of co-regulated DE genes. Sets which contain a small proportion of DE genes are of a lower level of interest.

2.2 The linear model

To be as general as possible, we use a linear model representation for the experiment, as described previously (Smyth, 2004). Write y_{gi} for the $\log_2$-expression value for sample i and probe g. We assume that the expression values have already been background corrected, normalized and, perhaps, filtered in a way appropriate for the microarray or expression platform. Write $\mathbf{y}_g = (y_{g1}, \ldots, y_{gn})^T$ for the vector of expression values for probe g. We assume

$$E(\boldsymbol{y}_g) = X\boldsymbol{\alpha}_g$$

where X is an $n \times p$ design matrix of full column rank and $\boldsymbol{\alpha}_g$ is an unknown coefficient vector of length p. The matrix X represents the design of the experiment, and describes how the different treatment factors are assigned to the RNA samples. The coefficients α_{gj} which make up $\boldsymbol{\alpha}_g$ represent the treatment effects or differences associated with probe g. We also assume

$$\text{var}(\boldsymbol{y}_g) = W^{-1}\sigma_g^2$$

where σ_g^2 is the unknown probewise variance and W is a positive-definite matrix of weights. The weight matrix W provides the ability to incorporate array weights if desired (Ritchie *et al.*, 2006) or to incorporate correlations between samples (Smyth *et al.*, 2005). If W is unknown, then it is estimated from the expression data on all G probes, as described in the papers just cited. After this step, W is treated as known.

2.3 Distributional assumptions across genes

We assume that the y_{gi} are multivariate normal, at least for the probes in our gene set, with a general but unknown correlation structure between the probes. Multivariate normality is theoretically a strong assumption, but we show later that our test procedure is robust against departures from normality.

In order to borrow information across probes when estimating standard errors, we assume a hierarchical model for the probewise variances, as described previously (Lonnstedt and Speed, 2002; Wright and Simon, 2003; Smyth, 2004). We assume that the probewise variances are sampled from an inverse-chisquare distribution,

$$\frac{1}{\sigma_g^2} \sim \frac{1}{s_0^2}\chi^2_{d_0}$$

where s_0^2 is the prior variance and d_0 is the prior degrees of freedom. The prior variance represents typical variability and the prior degrees of freedom control how consistent the variability is across probes.

2.4 Probe level tests

A particular contrast of the coefficients, represented by $\beta_g = \mathbf{c}^T \boldsymbol{\alpha}_g$, is assumed to be of interest. We want to find genes for which β_g is nonzero. The linear model described above is fitted to the expression data for each probe. The t-statistic for testing $\beta_g = 0$ is

$$t_g = \frac{\hat{\beta}_g}{s_g \sqrt{v}}$$

where $\hat{\beta}_g = \mathbf{c}^T \hat{\boldsymbol{\alpha}}_g$ is the least squares estimator of β_g, s_g is the residual standard deviation for probe g, and $v = \mathbf{c}^T (X^T W X)^{-1} \mathbf{c}$ is the unscaled standard deviation of $\hat{\beta}_g$. Under the null hypothesis, t_g follows a t-distribution on $d = n - p$ degrees of freedom.

Following Wright and Simon (2003) and Smyth (2004), an improved test can be obtained by computing the posterior variances

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d\, s_g^2}{d_0 + d} \tag{1}$$

and moderated t-statistics

$$\tilde{t}_g = \frac{\hat{\beta}_g}{\tilde{s}_g \sqrt{v}}$$

Under the null hypothesis, $\tilde{t}_g$ follows a t-distribution on $d_0 + d$ degrees of freedom. The gain in degrees of freedom of the moderated over the ordinary t-statistic reflects the information which is borrowed from other probes when making inferences about an individual probe. The moderated t-test has been shown to be superior to other tests in comparative studies (Kooperberg *et al.*, 2005; Diboun *et al.*, 2006; Murie *et al.*, 2009).

The hyper-parameters s_0 and d_0 in the prior distribution for σ_g^2 are estimated from the expression data on all G probes as described by Smyth (2004). After this step, s_0 and d_0 are treated as known.

For the calculations in this paper, the moderated t-statistics are transformed to equivalent standard normal random variables,

$$z_g = F^{-1}(F_t(\tilde{t}_g)) \tag{2}$$

where F and F_t are the cumulative distribution functions of the standard normal and t_{d_0+d} distributions respectively.

3 APPROACH TO GENE SET TESTING

3.1 Gene set hypotheses

Let $\mathcal{S}$ be the set of indices of the probes in our gene set of interest. We want to test the null hypothesis H_0 that $\beta_g = 0$ for all $g \in \mathcal{S}$. The alternative hypothesis depends on whether the DE genes in $\mathcal{S}$ are expected to change in the same direction or not, and whether this direction is specified in advance. We consider three possible alternative hypotheses. The *up* hypothesis H_u is that $\beta_g > 0$ for at least one $g \in \mathcal{S}$. The *down* hypothesis H_d is that $\beta_g < 0$ for at least one $g \in \mathcal{S}$. The *mixed* hypothesis H_m is that $\beta_g \neq 0$ for at least one $g \in \mathcal{S}$, i.e., the genes can change in mixed (up or down) directions. Clearly, it is possible for more than one of the alternative hypotheses H_m, H_u or H_d to be true for the same gene set.

3.2 Gene weights

In some cases there are prior reasons to give more weight to some genes in the gene set than others, for example genes that are more highly expressed might be of more interest. We therefore allow the possibility of gene weights a_g, which are used to weight the z_g when the gene set summary statistic T is computed, similar to a suggestion of Jiang and Gentleman (2007).

Positive or negative gene weights can be used to reflect the expected direction of change of genes in the gene set. For example, the gene weights might be ± 1 depending on whether the genes are known to be up or down regulated in a particular pathway, or the gene weights might be set equal to the log-fold changes for these genes in a prior experiment.

3.3 Gene set statistics

A test statistic T for the gene set $\mathcal{S}$ is constructed from the moderated t-statistics for probes in the set. Several gene set level summary statistics have been proposed in previous research on gene set testing (Jiang and Gentleman, 2007; Ackermann and Strimmer, 2009). We propose a number of new summary statistics which have a good power for different gene set scenarios. Our statistics are computed in terms of the z-scores z_g.

When all genes in the set $\mathcal{S}$ are differentially expressed by about the same amount, the weighted mean of the genewise statistics is a logical gene set statistic. Write $A = \sum_{g \in \mathcal{S}} |a_g|$ for the sum of absolute gene weights of genes in the set. To test the directional hypotheses H_u or H_d, we define $T_{\text{mean}} = (\sum_{g \in \mathcal{S}} a_g z_g)/A$. To test H_m, $a_g z_g$ is replaced by $|a_g z_g|$.

When only a few genes in the set are differentially expressed, or if some log-fold-changes are much larger than others, the mean of the squared genewise statistics is a more sensitive measure. To test the mixed hypothesis H_m, we define $T_{\text{msq}} = (\sum_{g \in \mathcal{S}} |a_g| z_g^2)/A$. To test H_u, the sum in the numerator of the statistic is taken only over those $a_g z_g$ which are positive. To test H_d, the sum in the numerator of the statistic is taken only over those $a_g z_g$ which are negative.

We define two more gene set statistics, which are designed to be sensitive to gene sets in which about half of the genes are differentially expressed. The first is the mean-50 statistic. Let h be the smallest integer greater than or equal to half the number of genes in the set, i.e., $h = \lceil (m+1)/2 \rceil$ where m is the number of genes in the gene set. The mean50 statistic T_{mean50} is the weighted mean of the top h most significant z-statistics. To test H_m, the mean50 statistic is the mean of the h largest absolute $a_g z_g$ values. To test H_u, the mean50 statistic is the mean of the h largest $a_g z_g$ values. To test H_d, the mean50 statistic is the mean of the h smallest $a_g z_g$ values.

The last statistic is inspired by the max-mean statistic of Efron and Tibshirani (2007). We call it *floor-mean* because it applies a floor value to the z-statistics. To test H_u, we compute the floored genewise statistics $f_g = \max(z_g, 0)$. To test H_d, the floored statistics are $f_g = \min(a_g z_g, 0)$. To test H_m, the floored statistics are $f_g = \max(|z_g|, 0.67)$, where 0.67 is the square-root of the median of the χ_1^2 distribution. In each case, $T_{\text{floormean}}$ is computed as for T_{mean} but with f_g in place of z_g. The floor-mean statistic behaves similarly to the mean50 statistic, but is faster to compute.

3.4 Assigning p-values

We now seek to assign a p-value to the gene set statistic T. The distribution of T is unknown, because the correlation of expression scores between probes is unknown, so a resampling method must be used to assign the p-value. We avoid permuting genes because this

would destroy the inter-probe dependence (Goeman and Bühlmann, 2007). We also avoid permuting samples, for several reasons. First, permutation requires a large number of replicate samples in order to provide a reliable p-value estimate, whereas we wish to analyse experiments with small numbers of replicates. Secondly, permutation does not have the ability to test general linear model hypotheses, such as we have specified. Thirdly, permutation assumes samples to be identically distributed and exchangeable, whereas we wish to accommodate various types of weighting and correlation structures.

We instead adapt the idea of rotation tests from Langsrud (2005). Rotation test use a type of simulation to generate p-values. The first step is to remove the nuisance parameters in the linear model, all the α_{gj} other than the contrast of interest β_g, by projecting the data for each probe onto the $d+1$ dimensional residual space orthogonal to them. This yields a set of $d+1$ independent residuals, such that the t-statistic $\tilde{t}_g$ can be computed from the first residual. This step allows us to test $\beta_g = 0$ without making assumptions about the other coefficients in the linear model. The second step randomly rotates the residuals in $d+1$ dimensional space. For each rotation, the gene set statistic T is re-computed, and compared to the observed value. The final p-value is $p = (b+1)/(B+1)$, where B is the total number of rotations and b is the number which yield a rotation statistic at least as extreme as that observed. This is an exact p-value Barnard (1963).

3.5 Estimating the active proportion

The above procedure attaches a p-value to the gene set. When the p-value is statistically significant, it is of interest to know how many genes in the set are contributing to this result. We consider a gene to be *active* in the result if $z_g > \sqrt{2}$ (for H_u) or $z_g < -\sqrt{2}$ (for H_d) or $|z_g| > \sqrt{2}$ for (for H_m). The threshold of $\sqrt{2}$ is somewhat arbitrary but is motivated by Akaike's information criterion in model selection theory, wherein the addition of one parameter to a model is considered worthwhile if it improves the chisquare goodness of fit by 2 or greater.

4 NUMERICAL IMPLEMENTATION

4.1 Independent residual effects

For the statistical model considered here, we are able to substantially simplify the rotation p-value computations outlined by Langsrud (2005), resulting in a very fast algorithm. All the computations below are for genes g in $\mathcal{S}$ only. Other genes can be ignored.

The first step is to remove the nuisance parameters, i.e., the contrasts not of interest, from the linear model. This is done by projecting each $\boldsymbol{y}_g$ onto the space orthogonal to the columns of X not involved in the null hypothesis.

Let C be an invertible $p \times p$ matrix with last column equal to $\mathbf{c}$, and write $\boldsymbol{\beta}_g = C^T \boldsymbol{\alpha}_g$. Note

$$X\boldsymbol{\alpha}_g = X(C^T)^{-1}\boldsymbol{\beta}_g$$

and that β_g, as defined in Section 2.4, is the last element of $\boldsymbol{\beta}_g$. In this way, the linear model is re-parametrized so that the null hypothesis concerns the last regression coefficient.

The projection is obtained as a byproduct of the usual QR decomposition used in the numerical calculations for fitting the linear model,

$$W^{1/2}X(C^T)^{-1} = QR.$$

We use the full QR-decomposition in which Q is $n \times n$. Let Q_2 be Q with the first $p-1$ columns removed. Then the nuisance parameters are removed by the projection

$$\mathbf{u}_g = Q_2^T W^{1/2}\mathbf{y}_g$$

for each g in $\mathcal{S}$. In the actual numerical computations, the vector $\mathbf{u}_g$ can be obtained from the QR-decomposition without forming Q or Q_2 explicitly.

Under the null hypothesis, the elements of $u_{g1}, \ldots, u_{g,d+1}$ are independent and identically $N(0, \sigma_g^2)$. We have

$$s_g^2 = \frac{1}{d}\sum_{i=2}^{d+1} u_{gi}^2,$$

then the posterior variance is computed from (1), and finally the moderated t-statistics are

$$\tilde{t}_g = u_{g1}/\tilde{s}_g$$

4.2 Rotation

Let $\rho_g^2 = \mathbf{u}_g^T\mathbf{u}_g$. The rotation test method rotates the vector of residuals $\mathbf{u}_g$ to a random point $\mathbf{u}_g^*$ on the $d+1$-sphere of radius ρ_g. For every rotation, the moderated t-statistics and the overall gene set statistic T are recomputed. After a large number of rotations, the Monte Carlo p-value is computed as above.

It is actually only necessary to randomly generate the first element u_1^* of $\mathbf{u}_g^*$, because the residual variances for the rotated data can be computed from $s_g^{*2} = (\rho_g^2 - u_1^{*2})/d$. The computation is extremely efficient. A random rotation vector $\mathbf{r}$ is generated satisfying $\mathbf{r}^T\mathbf{r} = 1$. Then u_1^* is generated for all probes in the gene set by

$$\mathbf{u}_1^* = U\mathbf{r}$$

where U is the $m \times (d+1)$ matrix with rows $\mathbf{u}_g^T$ for g in $\mathcal{S}$. This yields rotated t-statistics and z-statistics for each gene in the gene set, and hence to a null distribution value T^* of the gene set statistic. New rotation vectors $\mathbf{r}$ and T^* are generated a large number of times, typically $B = 10000$.

5 SIMULATION STUDY

5.1 Scenarios

Four multi-group experimental designs were simulated. The first design (D1-3) was a three-group experiment with $n_1 = n_2 = 3$ replicate samples in the first two groups and $n_3 = 20$ replicate samples in the third. Interest is in differential expression between the first two groups. The second design (D1-5) was the same but with $n_1 = n_2 = 5$ replicate samples in the first two groups. The third (D2-3) and fourth (D2-5) designs were two-way factorial designs with two levels per factor, with 3 samples and 5 samples per group respectively. For the factorial designs, interest was in differential expression for the first factor, the other being a blocking factor.

Each simulated dataset had $G = 10,000$ probes per array. Genes were divided either into 250 gene sets of 40 genes each, or into 10 gene sets of 1000 genes each. In each case, only the first gene set was simulated to contain differentially expressed genes. Genewise variances were simulated according to the model described in Section 2 with $d_0 = 4$ and $s_0 = 0.25$.

For each design, datasets were simulated in 10 scenarios with different proportions of up and down regulated genes in the set, and either independent or correlated expression profiles. In each scenario, the differentially expressed genes share the same fold-change. The fold-change was chosen for each scenario to make the highest powers around 80%. Table 1 shows the scenarios for designs D1-3 and D1-5. Smaller fold-changes were simulated with 1000 genes in each set than with 40 so as to keep the power about the same in each case (Table 1).

Table 1. Ten simulated scenarios for designs D1-3 and D1-5. Scenarios differ according to the proportion of up and down regulated genes in the set and the intergene correlation.

Scenario	Proportion up	Proportion down	Correlation	log-fold change	Hypothesis tested
1	1.00	0.00	0.0	0.1(0.02)	up
2	1.00	0.00	0.1	0.2(0.2)	up
3	0.25	0.00	0.0	0.3(0.07)	up
4	0.25	0.00	0.1	0.4(0.2)	up
5	0.50	0.50	0.0	0.2(0.07)	mixed
6	0.50	0.50	0.1	0.2(0.2)	mixed
7	0.20	0.20	0.0	0.3(0.1)	mixed
8	0.20	0.20	0.1	0.4(0.2)	mixed
9	0.00	0.00	0.0	0.0(0.0)	up/mix
10	0.00	0.00	0.1	0.0(0.0)	up/mix

Fold-changes are for 40 genes per set (or 1000 genes per set). Genes regulated in the same direction are positively correlated, those in opposite directions are negatively correlated. In scenario 10, the correlation applies to all genes in the set.

5.2 Size

The most important property of a statistical test is that its size (type I error rate) is controlled correctly. It follows from the theory of rotation tests that ROAST must hold its size correctly if the data is normally distributed, and this was confirmed by simulations (Tables S1, S4).

To explore the robustness of ROAST to non-normal data, we simulated expression data to be exponentially distributed. The exponential distribution is highly right skew, far more non-normal than would be seen in any real microarray experiment. To simulate correlated exponential random variables, data was first simulated as for the normal simulations (Table 1, scenarios 9 and 10), then transformed to be exponentially distributed, via the appropriate normal and exponential cumulative distribution functions. Even in this extreme situation, ROAST continued to hold its power correctly when used with the mean, mean50 or floor-mean gene set statistics (Table S2). This was not quite true with the msq gene summary statistic but, even then, the true test sizes were only slightly higher than nominal sizes. We conclude that ROAST is likely to be robust to non-normality in realistic data situations.

5.3 Power

ROAST was found to have good power to detect sets containing differentially expressed genes with very modest fold-changes, across a range of scenarios (Tables S3 and S5). The mean set statistic was the best of the statistics for detecting scenarios with all genes changing (Tables S3, S5, scenarios 1,2,5,6). The msq set statistic was the best of the statistics for detecting scenarios with only a minority of genes changing Tables S3, S5, scenarios 3,4,7,8). The mean50 and floor-mean statistics was intermediate between the mean and msq statistics in both scenarios. The statistics can be ordered as mean, floor-mean, mean50, then msq, in terms of increasing sensitivity to a subset of differentially expressed genes and decreasing sensitivity to all genes changing by the same amount. In all cases, power was reduced for a given fold-change when the genes in set had correlated expression values. More genes in the gene set increased power, but this increase was not so apparent when the genes were correlated (Tables S3 and S5).

5.4 Comparison with permutation tests

The key advantage of ROAST is the ability to handle situations for which no other gene set test is suitable. It is of interest however to see how ROAST compares with other methods when conditions are suitable for them. We compared ROAST to GSEAlm (Oron *et al.*, 2008). We chose GSEAlm because it is a high quality representative of permutation methods, and because it has more flexible options that most permutation software, being able to handle one-way anova or block designs.

No permutation algorithm can be expected to work well on design D1-3, with only 3 arrays per group. Indeed we found that GSEAlm failed to hold its size when the number of available permutations was not large (data now shown). This can be traced to the way in which the permutation p-values are computed. Whereas ROAST computes an exact p-value, GSEAlm, like all permutation software that we know of, computes an estimate $\hat{p} = b/B$ of the p-value, where B is the number of permutations and b is the number of permuted statistics as extreme as that observed. This estimated p-value can often be zero if the number of permutations is modest, resulting in an over-statement of statistical significance (Ernst, 2004).

GSEAlm was compared with ROAST on designs D1-5, D2-3 and D2-5. GSEAlm is computationally time consuming, so only 100 data sets were simulated for each scenario for each design. On design D1-5, GSEAlm is somewhat less powerful that ROAST, regardless of scenario and regardless of ROAST set statistic (Table S6). This was presumably because ROAST is able to make use of residual degrees of freedom from the 20 arrays in the third group that is not involved in the hypothesis test. GSEAlm was not tested on the mixed scenarios, as it is not able to test bi-directional hypotheses. On the two-factor designs D2-3 and D2-5, for testing directional hypotheses, GSEAlm is similar in power to ROAST with the mean gene set statistic (Tables S7 and S8).

6 MAMMARY STEM CELLS

Mammary epithelial cells can be sorted into a family tree of cell populations, including a mammary stem cell enriched population (MaSC), luminal progenitor cells (LP) and mature luminal cells (ML) (Visvader and Lindeman, 2006). This family tree is of tremendous interest for many reasons, for example, because mammary stem cells or luminal progenitor cells may represent the cell of origin for different types of breast cancer (Lim *et al.*, 2009). Most experimental research on mammary cells is undertaken using mouse as the model organism. Hence it is critically important to establish that results observed for mice will remain valid for humans.

Gene set testing provides a powerful way to relate expression profiles across different platforms or different species. In this example, we use gene set tests to demonstrate that the transcriptional differences between MaSC, LP and ML cells are broadly conserved between mouse and human. We focus particularly on stem cells. We define a gene set associated with stem cells in mouse, then examine the profile of this set in the human data. This allows us to confirm

Table 2. Human mammary cell populations profiled.

Array	Cell population	Patient	BeadChip	Block
1	MaSC enriched	A	4380071023	1
2	Stroma	A	4380071023	1
3	ML	A	4380071023	1
4	LP	A	4380071023	1
5	MaSC enriched	B	4380071023	2
6	Stroma	B	4380071023	2
7	ML	B	4380071027	3
8	LP	B	4380071027	3
9	MaSC enriched	C	4380071027	4
10	Stroma	C	4380071027	4
11	ML	C	4380071027	4
12	LP	C	4380071027	4

Block represents unique patient–BeadChip combinations.

Table 3. ROAST p-values for distinguishing human cell populations using the mouse MaSC signature gene set.

Summary statistic	MaSC−LP		MaSC−ML	
	H_u	H_d	H_u	H_d
mean	0.001	1.000	0.001	1.000
mean50	0.001	0.397	0.001	0.396
floormean	0.001	0.345	0.001	0.318
msq	0.001	0.074	0.001	0.048

p-values based on 999 rotations.

conclusions from Lim *et al.* (2010), but here we use ROAST to formally take account of inter-gene correlation in evaluating statistical significance, which we were not able to do in the earlier publication.

Mammary epithelial cells and stroma cells were sorted from breast tissue samples from three human patients. RNA was profiled on two Illumina HumanWG-6 V3 BeadChips, comprising 12 arrays (Table 2). A similar experiment was undertaken using mouse samples and Illumina mouse WG-6 V2 BeadChips. The microarray data was normalized and annotated as previously described (Lim *et al.*, 2010). The data is available as GEO database series GSE16997 and GSE19446 for human and mouse respectively.

A set of genes was selected to represent the transcriptional signature of MaSC cells in mouse, as previously described (Lim *et al.*, 2010). Briefly, we identified signature probes as those significantly up-regulated or down-regulated in MaSC cells versus both of the other two mammary epithelial populations, LP and ML. This yielded 2616 up and 2305 down-regulated MaSC signature probes in mouse, corresponding to 3410 unique gene symbols, of which 2754 had human orthologs. This defined then a set of 2754 human gene symbols. We set gene weights a_i equal to the log-fold-change observed for that gene in the mouse date, specifically the average of the log-fold-changes observed for MaSC vs LP and MaSC vs ML. Hence genes were weighted according to the strength and direction of their regulation in MaSC cells.

It is important to account for biological and technical correlations when analysing microarray data, as this improves the precision and reliability of the results, even with high quality data such as we analyse here. For the human samples described in Table 2, it is essential to account for correlations between samples taken from the same patient. Also to be expected are more or less subtle batch effects between BeadChips. Our linear modelling approach permits us to adjust for patient and BeadChip effects in a variety of ways. We could include patient as a fixed effect in our linear model, when testing for differential expression between the cell populations. Alternatively, we could include patient as a random effect, with BeadChip as a fixed effect. The effects were relatively subtle, so we combined patient and BeadChip into one blocking factor (see Table 2) which was included as a random effect. In this approach, samples are treated as correlated only if they corresponded to the same patient and the same BeadChip. Using the method of Smyth *et al.* (2005), the intra-block correlation was estimated to be 0.08, a small but detectable positive correlation between samples in the same block. Our weight matrix W therefore was made up of correlations ρ_{ij}, where ρ_{ij} was 1 if $i = j$, 0.08 if arrays i and j are in the same block and zero otherwise.

ROAST was used to test whether the mouse MaSC signature gene set was able to distinguish human MaSC from LP cells, and human MaSC from ML (Table 3). Genes were weighted in the test by their direction of change in mouse, hence the alternative hypothesis H_u corresponds to genes changing in the same direction in human as they did in mouse, whereas H_d represents change in opposite directions. The highly significant results for H_u show that the main body of genes change in the same direction in human as in mouse. The non-significant p-values for H_d for the mean, mean50 and floormean statistics deny any sizeable group of genes changing in the opposite direction. The msq statistic is highly sensitive to even a small number of genes, and gives suggestive p-values for H_d. Close inspection of individual genes shows that a small number of the signature genes do indeed move in the reverse directions in human and mouse, explaining the msq result. Nevertheless, the high degree of conservation across species supports mouse as a model system for the study of mammary gland development.

7 DISCUSSION

ROAST is the first focused gene set test which correctly allows for inter-gene correlation and gives statistically rigorous p-values for small or complex microarray experimental designs. Like Dørum *et al.* (2009), we use rotation to evaluate p-values, but our methodology is for self-contained tests with specially targeted gene sets whereas they considered competitive tests between a large database of gene sets. ROAST is the only existing software that could have been used for our mammary stem cell data example.

The use of rotation offers many advantages over permutation. It is computationally much faster. It yields exact p-values, so that ROAST holds its nominal size correctly even for small samples. The number of rotations can be chosen large enough to avoid any problem with granularity of p-values. Dørum *et al.* (2009) considered one or two-group t-tests, whereas we take full advantage of the possibilities of linear models. ROAST is applicable to any experiment which can be analysed using an linear model, i.e., to arbitrarily complex experiments. The gene set can be tested in any contrast between the coefficients, not necessarily a simple comparison between two groups. Indeed, our development could easily be extended to test gene sets in two or more contrasts simultaneously. In that case

we would replace our moderated t-statistics with F-statistics. Our implementation allows for correlations between RNA samples and for array quality weights, assuming that the correlations or weights can be estimated from the entire ensemble of probes on the arrays. In all cases, the test takes advantage of all available residual degrees of freedom. Our implementation uses empirical Bayes test statistics, which should add additional stability in small samples, especially when the gene set contains only a few genes.

Rotation theoretically depends on multivariate normality, but simulations with grossly non-normal data shows that ROAST is insensitive to departures from normality, agreeing with analogous results of Dørum *et al.* (2009).

We propose a number of new set summary test statistics. Like Jiang and Gentleman (2007), but unlike Ackermann and Strimmer (2009), we find that the choice of summary statistic does affect performance. Our summary statistics are designed to vary in their sensitivity to gene sets that contain a small proportion of differentially expressed genes. No statistic is universally better than the others in all situations. When the mean statistic is used, a gene set can be significant for the up or down hypothesis, but not both. With the other set statistics, a gene set can be judged as significant in both directions, if there is a subset of genes that are up-regulated and a subset which are down-regulated. Therefore, floormean, mean50 and especially msq give higher resolution results, more nuanced but potentially less clear cut. To help make judgements in this respect, ROAST gives an estimate of the proportion of genes which actively contribute to a significant result. The simulations presented in this paper assumed a uniform fold-change for all DE genes. Simulations with random fold-changes tend to improve the performance of msq, and also of mean50 and floormean, relative to that of the mean statistic (data not shown). We suggest mean50 as a good compromise in many biological situations.

Potential applications for ROAST include those where the set might not be made up of genes. We have for example used it in exon-level expression analyses to test whether any exon of a given gene is differentially expressed.

ACKNOWLEDGEMENT

Funding: NHMRC Program Grant 490037.

REFERENCES

Ackermann, M. and Strimmer, K. (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, **10**, 47.

Adewale, A. J., Dinu, I., Potter, J. D., Liu, Q., and Yasui, Y. (2008). Pathway analysis of microarray data via regression. *J Comput Biol*, **15**(3), 269–77.

Barnard, G.A. (1963). Discussion of The spectral analysis of point processes (by MS Bartlett). *Journal of the Royal Statistical Society*, **25**, 294.

Diboun, I., Wernisch, L., Orengo, C. A., and Koltzenburg, M. (2006). Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma. *BMC Genomics*, **7**, 252.

Dinu, I., Potter, J. D., Mueller, T., Liu, Q., Adewale, A. J., Jhangri, G. S., Einecke, G., Famulski, K. S., Halloran, P., and Yasui, Y. (2007). Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*, **8**, 242.

Dørum, G., Snipen, L., Solheim, M., and Saebo, S. (2009). Rotation testing in gene set enrichment analysis for small direct comparison experiments. *Stat Appl Genet Mol Biol*, **8**(1), Article34.

Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *Ann. Appl. Statist.*, **1**, 107–129.

Ernst, M. (2004). Permutation methods: A basis for exact inference. *Stat. Sci.*, **19**, 686–696.

Goeman, J. J. and Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**(8), 980–7.

Goeman, J. J., van de Geer, S. A., de Kort, F., and van Houwelingen, H. C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**(1), 93–9.

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

Jiang, Z. and Gentleman, R. (2007). Extensions to gene set enrichment. *Bioinformatics*, **23**(3), 306–13.

Kooperberg, C., Aragaki, A., Strand, A. D., and Olson, J. M. (2005). Significance testing for small microarray experiments. *Stat Med*, **24**(15), 2281–98.

Langsrud, Ø. (2005). Rotation tests. *Statist. Comput.*, **15**, 53–60.

Lim, E., Vaillant, F., Wu, D., Forrest, N. C., Pal, B., Hart, A. H., Asselin-Labat, M. L., Gyorki, D. E., Ward, T., Partanen, A., Feleppa, F., Huschtscha, L. I., Thorne, H. J., Fox, S. B., Yan, M., French, J. D., Brown, M. A., Smyth, G. K., Visvader, J. E., and Lindeman, G. J. (2009). Aberrant luminal progenitors as the candidate target population for basal tumor development in brca1 mutation carriers. *Nat Med*, **15**(8), 907–13.

Lim, E., Wu, D., Pal, B., Bouras, T., Asselin-Labat, M. L., Vaillant, F., Yagita, H., Lindeman, G. J., Smyth, G. K., and Visvader, J. E. (2010). Transcriptome analyses of mouse and human mammary cell subpopulations reveals multiple conserved genes and pathways. *Breast Cancer Res*, **12**(2), R21.

Lonnstedt, I. and Speed, T. (2002). Replicated microarray data. *Statistica Sinica*, **12**, 31–46.

Manoli, T., Gretz, N., Grone, H. J., Kenzelmann, M., Eils, R., and Brors, B. (2006). Group testing for pathway analysis improves comparability of different microarray datasets. *Bioinformatics*, **22**(20), 2500–6.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, **34**(3), 267–73.

Murie, C., Woody, O., Lee, A. Y., and Nadon, R. (2009). Comparison of small n statistical tests of differential expression applied to microarrays. *BMC Bioinformatics*, **10**, 45.

Oron, A. P., Jiang, Z., and Gentleman, R. (2008). Gene set enrichment analysis using linear models and diagnostics. *Bioinformatics*, **24**(22), 2586–91.

Ritchie, M. E., Diyagama, D., Neilson, J., van Laar, R., Dobrovic, A., Holloway, A., and Smyth, G. K. (2006). Empirical array quality weights in the analysis of microarray data. *BMC Bioinformatics*, **7**, 261.

Saxena, V., Orgill, D., and Kohane, I. (2006). Absolute enrichment: gene set enrichment analysis for homeostatic systems. *Nucleic Acids Res*, **34**(22), e151.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, **3**, Article3.

Smyth, G. K., Michaud, J., and Scott, H. S. (2005). Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, **21**(9), 2067–75.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, **102**(43), 15545–50.

Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., and Park, P. J. (2005). Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci U S A*, **102**(38), 13544–9.

Visvader, J. E. and Lindeman, G. J. (2006). Mammary stem cells and mammopoiesis. *Cancer Res*, **66**(20), 9798–801.

Wang, L., Zhang, B., Wolfinger, R. D., and Chen, X. (2008). An integrated approach for the analysis of biological pathways using mixed models. *PLoS Genet*, **4**(7), e1000115.

Wright, G. W. and Simon, R. M. (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics*, **19**(18), 2448–55.