Rauschenberger *et al. BMC Bioinformatics* (2016) 17:118
DOI 10.1186/s12859-016-0961-5

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Testing for association between RNA-Seq and high-dimensional data

Armin Rauschenberger[1], Marianne A. Jonker[1], Mark A. van de Wiel[1,2] and Renée X. Menezes[1*]

Abstract

Background: Testing for association between RNA-Seq and other genomic data is challenging due to high variability of the former and high dimensionality of the latter.

Results: Using the negative binomial distribution and a random-effects model, we develop an omnibus test that overcomes both difficulties. It may be conceptualised as a test of overall significance in regression analysis, where the response variable is overdispersed and the number of explanatory variables exceeds the sample size.

Conclusions: The proposed test can detect genetic and epigenetic alterations that affect gene expression. It can examine complex regulatory mechanisms of gene expression. The R package *globalSeq* is available from Bioconductor.

Keywords: High-dimensional, Overdispersion, Negative binomial, Global test, Integration, RNA-Seq

Background

Genetic and epigenetic factors contribute to the regulation of gene expression. A better understanding of these regulatory mechanisms is an important step in the fight against cancer. Of interest are genetic alterations such as single nucleotide polymorphisms (SNPs), copy-number variations (CNVs) and loss of heterozygosity (LOH), as well as epigenetic alterations such as DNA methylation, microRNA expression levels and histone modifications.

From a statistical perspective, it makes sense to represent the expression of one gene as a response variable that changes when some covariates are altered. As a starting point, we assume that all covariates come from a single genetic or epigenetic molecular profile. Typically, more covariates are of interest than there are samples.

A plethora of methods for the analysis of gene expression and covariates has emerged in the last years. Many of these methods test each covariate individually, and subsequently correct for multiple testing or rank the covariates by significance. An alternative approach is the global test from Goeman et al. [1]. The global test does not test the individual but the joint significance of covariates. It allows for high dimensionality, reduces the multiple testing burden, and successfully detects small effects that encompass many covariates. Due to its desirable properties, the global test has become a widely used tool in genomics (e.g. [2–4]).

Currently, gene expression microarrays are being supplanted by high-throughput sequencing. The negative binomial distribution seems to be a sensible choice for modelling RNA sequencing data [5, 6]. One of its parameters describes the dispersion of the variable. If this parameter is unknown, the negative binomial distribution is not in the exponential family. As the global test from Goeman et al. [1] is limited in its current form to the exponential family of distributions, a new test is needed for RNA-Seq data. We will provide here such a test.

After proposing a global test for the negative binomial setting, we perform a simulation study, and analyse two publicly available datasets. The first application concentrates on method validation, overdispersion, and individual contributions. The second application concentrates on robustness against multicollinearity, the method of control variables, and the simultaneous analysis of multiple molecular profiles.

Although we focus on RNA-Seq gene expression data, the test developed here is applicable whenever associations between a count variable and large sets of quantitative or binary variables are of interest. In essence, it can be applied to any other type of sequencing data, such as

*Correspondence: r.menezes@vumc.nl
[1] Department of Epidemiology and Biostatistics, VU University Medical Center, 1007 MB, Amsterdam, The Netherlands
Full list of author information is available at the end of the article

ChIP-Seq (chromatin immunoprecipitation), microRNA-Seq or meth-Seq (methylation).

Methods

The random-effects model

The human genome contains several thousand protein-coding genes. In the following, only one gene is considered at a time. Accordingly, the expression of one gene across all samples is our response variable $\boldsymbol{y} = (y_1, \ldots, y_n)^T$. If we were interested whether a given subset of SNPs affected gene expression, these SNPs would be our p covariates. The $n \times p$ covariate matrix $\boldsymbol{X}$ is potentially high-dimensional ($p \gg n$).

We represent the relationship between the response and the covariates using the generalised linear model framework from McCullagh and Nelder [7]:

$$\mathrm{E}[y_i] = h^{-1}\left(\alpha + \sum_{j=1}^{p} X_{ij}\beta_j\right),$$

where h^{-1} is an inverse link function, α is the unknown intercept, X_{ij} is the entry in the i^{th} row and j^{th} column of $\boldsymbol{X}$, and $\beta_1, \ldots, \beta_p$ are the unknown regression coefficients. This model holds for all samples i ($i = 1, \ldots, n$).

We are interested in testing the joint significance of all regression coefficients. This is challenging because the regression coefficients cannot be estimated by classical regression methods if there are more covariates than samples. Goeman et al. [1] took a novel approach for testing $H_0 : \beta_1 = \ldots = \beta_p = 0$ against $H_1 : \beta_1 \neq 0 \cup \ldots \cup \beta_p \neq 0$. The decisive step from Goeman et al. [1] was to assume $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ to be random, with the expected value $\mathrm{E}[\boldsymbol{\beta}] = \boldsymbol{0}$ and the variance-covariance matrix $\mathrm{Var}[\boldsymbol{\beta}] = \tau^2\boldsymbol{I}$, where $\boldsymbol{I}$ is the $p \times p$ identity matrix and $\tau^2 \geq 0$. Then a random-effects model is obtained:

$$\mathrm{E}\left[y_i | r_i\right] = h^{-1}(\alpha + r_i), \qquad r_i = \sum_{j=1}^{p} X_{ij}\beta_j. \tag{1}$$

This random-effects model allows to rephrase the null and the alternative hypotheses. Defining the random vector $\boldsymbol{r} = (r_1, \ldots, r_n)^T$, it can be deduced that $\mathrm{E}[\boldsymbol{r}] = \boldsymbol{0}$ and $\mathrm{Var}[\boldsymbol{r}] = \tau^2\boldsymbol{X}\boldsymbol{X}^T$. Now the null hypothesis of no association between the covariate group and the response is given by $H_0 : \tau^2 = 0$. To construct a score test against the one-sided alternative hypothesis $H_1 : \tau^2 > 0$, we need to assume a distribution for $y_i | r_i$.

The testing procedure

We assume the negative binomial distribution $y_i | r_i \sim \mathrm{NB}(\mu_i, \phi)$, where the mean parameter μ_i depends on the sample, but the dispersion parameter ϕ does not. We parametrise the negative binomial distribution such that $\mathrm{E}[y_i | r_i] = \mu_i$ and $\mathrm{Var}[y_i | r_i] = \mu_i + \phi\mu_i^2$. Its density function is given by

$$f(y_i) = \frac{\Gamma\left(y_i + \frac{1}{\phi}\right)}{\Gamma\left(\frac{1}{\phi}\right)\Gamma(y_i + 1)} \left(\frac{1}{1 + \mu_i\phi}\right)^{\frac{1}{\phi}} \left(\frac{\mu_i}{\frac{1}{\phi} + \mu_i}\right)^{y_i}.$$

Various link functions come into consideration for the negative binomial model. We favour the logarithmic link in order to relate the negative binomial model directly to the Poisson model (see below). As library sizes can be unequal, we include the offset $\log(m_i/\overline{m})$, where m_i denotes the library sizes, and $\overline{m}$ their *geometric* mean. Thus the mean function becomes

$$\mu_i = \exp\left(\alpha + r_i + \log\frac{m_i}{\overline{m}}\right) = \frac{m_i}{\overline{m}}\exp(\alpha + r_i). \tag{2}$$

When τ^2 is close to zero, the score test is the most powerful test of the null hypothesis $H_0 : \tau^2 = 0$ against the alternative hypothesis $H_0 : \tau^2 > 0$ [8]. Here the score function is the first derivative of the logarithmic marginal likelihood with respect to τ^2. Intuitively, if the marginal likelihood reacts sensitively to changes in τ^2 close to 0, there is evidence against $\tau^2 = 0$. Using results from le Cessie and van Houwelingen [9], we show in the Additional file 1 how to calculate the score function. This function contains the unknown parameters α and ϕ, but they can be estimated by maximum likelihood. Replacing the unknown parameters by their estimates leads to the test statistic

$$\begin{aligned} u_{nb} = &\left\{\sum_{i=1}^{n}\sum_{k=1}^{n}\frac{R_{ik}}{2}\,\frac{(y_i - \hat{\mu}_i)(y_k - \hat{\mu}_k)}{(1 + \hat{\phi}\hat{\mu}_i)(1 + \hat{\phi}\hat{\mu}_k)}\right\} \\ &- \sum_{i=1}^{n}\frac{R_{ii}}{2}\,\frac{(\hat{\mu}_i + y_i\hat{\phi}\hat{\mu}_i)}{(1 + \hat{\phi}\hat{\mu}_i)^2}, \end{aligned} \tag{3}$$

where R_{ij} is the entry in the i^{th} row and j^{th} column of the $n \times n$ matrix $\boldsymbol{R} = (1/p)\boldsymbol{X}\boldsymbol{X}^T$, and $\hat{\mu}_{0,i} = (m_i/\overline{m})\exp(\hat{\alpha})$ is the estimated mean under the null hypothesis. For simplicity we always write $\hat{\mu}_i$ instead of $\hat{\mu}_{0,i}$. In the Additional file 1 the test statistic is rewritten in matrix notation.

Statistical hypothesis testing depends on the null distribution of the test statistic u_{nb}, which is unknown. We will obtain p-values by permuting the response $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ together with the mean $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)^T$. Since this is a one-sided test [10], if the observed test statistic is larger than most of the test statistics obtained by permutation, there is evidence against the null hypothesis.

As we are not using a parametric form for the null distribution of the test statistic, no adjustments for the estimation of α and ϕ are necessary. Furthermore, maximum likelihood estimation does not depend on the order of the elements in $\boldsymbol{y} = (y_1, \ldots, y_n)^T$. Because neither $\hat{\alpha}$

nor $\hat{\phi}$ vary under permutation, computational efficiency can be achieved with these parameters as given.

When testing for associations between RNA-Seq data and another molecular profile, numerous genes might be of interest. Because one test is performed per gene, the multiple testing problem reappears. (In the applications from below we omit multiple testing correction when analysing the distribution of p-values.)

Relation to the poisson model

For comparison we also consider the Poisson distribution $y_i|r_i \sim \text{Pois}(\mu_i)$ with $\text{E}[y_i|r_i] = \text{Var}[y_i|r_i] = \mu_i$ and a logarithmic link function. Proceeding as above we obtain the test statistic

$$u_{pois} = \left\{ \sum_{i=1}^{n} \sum_{k=1}^{n} \frac{R_{ik}}{2} (y_i - \hat{\mu}_i)(y_k - \hat{\mu}_k) \right\} - \sum_{i=1}^{n} \frac{R_{ii}}{2} \hat{\mu}_i, \tag{4}$$

where the estimates $\hat{\mu}_i$ are the same as in the negative binomial model.

In the case of $\hat{\phi}\hat{\boldsymbol{\mu}} = \mathbf{0}$ we would have $u_{nb} = u_{pois}$, but in practice only situations with $\hat{\boldsymbol{\mu}} > \mathbf{0}$ are of interest. The fact that $\hat{\phi} = 0$ implies $u_{nb} = u_{pois}$ is convenient since a negative binomial distribution with a dispersion parameter close to zero is practically equivalent to a Poisson distribution.

Individual contributions

Following Goeman et al. [1], the test statistic u_{nb} can be rewritten to reveal the influence of individual samples and covariates.

The contribution of sample i ($i = 1, \ldots, n$) to the test statistic is

$$s_i = \left\{ \sum_{k=1}^{n} \frac{R_{ik}}{2} \frac{(y_i - \hat{\mu}_i)(y_k - \hat{\mu}_k)}{(1 + \hat{\phi}\hat{\mu}_i)(1 + \hat{\phi}\hat{\mu}_k)} \right\} - \frac{R_{ii}}{2} \frac{(\hat{\mu}_i + y_i \hat{\phi} \hat{\mu}_i)}{(1 + \hat{\phi}\hat{\mu}_i)^2}. \tag{5}$$

If s_i is positive, the sample i increases the evidence against the null hypothesis. Though, s_i not only depends on the sample i, but through $\boldsymbol{R}$, $\hat{\boldsymbol{\mu}}$ and $\hat{\phi}$ also on the other samples.

Especially useful is the contribution of covariate j ($j = 1, \ldots, p$) to the test statistic:

$$c_j = \frac{1}{2p} \left\{ \sum_{i=1}^{n} X_{ij} \frac{y_i - \hat{\mu}_i}{1 + \hat{\phi}\hat{\mu}_i} \right\}^2 - \sum_{i=1}^{n} \frac{X_{ij}^2}{2p} \frac{(\hat{\mu}_i + y_i \hat{\phi} \hat{\mu}_i)}{(1 + \hat{\phi}\hat{\mu}_i)^2}. \tag{6}$$

Note that multiplying c_j by p gives the u_{nb} that would have been obtained if only the covariate j had been tested. Similar to Goeman et al. [1], the test statistic for a group of covariates is the average of the individual test statistics. If c_j is positive, the covariate j increases the evidence against the null hypothesis. Conveniently, c_j is independent of all other covariates.

By construction we have $u_{nb} = \sum_{i=1}^{n} s_i$ and $u_{nb} = \sum_{j=1}^{p} c_j$. Even though a single hypothesis is tested on the covariate group, these decompositions allow to determine which samples and which covariates are the most influential on the test result. If samples or covariates can be put into categories, decomposing the test statistic and grouping samples by category could visualise how each category contributes to the test results. Similarly, if samples or covariates can be ordered according to some genomic or phenomic criteria, patterns might be detected.

Method of control variables

One drawback of obtaining p-values via permutation is the computational burden. Here we will make use of the work from Senchaudhuri et al. [11] in order to estimate p-values efficiently.

The proposed test statistic and the test statistic from Goeman el al. [1] have different advantages: whereas the former adequately models overdispersed count data, the latter has a known asymptotic null distribution. Usually we would obtain an unbiased estimate of the p-value using $1/k \sum_{i=1}^{k} \mathbf{1}[u_i \geq u_0]$, where $\mathbf{1}$ is the indicator function and u_i represents the proposed test statistic for a permutation ($i = 1, ..., k$) or for the observed data ($i = 0$). Following Senchaudhuri et al. [11], we could also obtain an unbiased estimate using $1/k \sum_{i=1}^{k} \mathbf{1}[u_i \geq u_0] - \mathbf{1}[q_i \geq q_0] + p^*$, where q_i and p^* are the test statistic and asymptotic value, respectively, from Goeman et al. [1]. If the test statistics u_i and q_i have a strong positive correlation, then this alternative estimate is preciser than the usual estimate [11]. (In the applications from below we only use the method of control variables when explicitly stated.)

Multiple molecular profiles

Not only SNPs but also other molecular mechanisms regulate gene expression. For instance, aberrant DNA methylation levels in promoter regions can activate oncogenes and inactivate tumour suppressor genes. Thus it could be interesting to test for associations between RNA-Seq gene expression data on one hand, and on the other SNP data as well as methylation data.

Let $\boldsymbol{X}$ represent the $n \times p$ SNP data matrix, and let $\boldsymbol{Z}$ represent the $n \times q$ methylation data matrix. The model from Eq. 1 allows to test single covariate sets, leading to the test statistic $u_{nb} = u(\boldsymbol{X})$ for SNP data, and to the test statistic $u_{nb} = u(\boldsymbol{Z})$ for methylation data.

Menezes et al. [12] provided a test for analysing multiple molecular profiles simultaneously, for responses with a distribution in the exponential family. As the negative binomial distribution with an unknown dispersion

parameter is not in the exponential family, we have to adapt this test. Following Menezes et al. [12], we include a second covariate set in the random-effects model from Eq. 1:

$$\mathrm{E}[y_i|r_i] = h^{-1}(\alpha + r_i), \qquad r_i = \sum_{j=1}^{p} X_{ij}\beta_j + \sum_{j=1}^{q} Z_{ij}\gamma_j. \tag{7}$$

Using the ideas and the notation from above: for the random vectors $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^T$ we assume $\mathrm{E}[\boldsymbol{\beta}] = \mathrm{E}[\boldsymbol{\gamma}] = 0$, $\mathrm{Var}[\boldsymbol{\beta}] = \tau^2\boldsymbol{I}$, $\mathrm{Var}[\boldsymbol{\gamma}] = \upsilon^2\boldsymbol{I}$ and $\mathrm{Cov}[\boldsymbol{\beta}, \boldsymbol{\gamma}] = 0$, where $\tau^2 \geq 0$ and $\upsilon^2 \geq 0$. Consequently, the random vector $\boldsymbol{r} = (r_1, \ldots, r_n)^T$ has $\mathrm{E}[\boldsymbol{r}] = \boldsymbol{0}$ and $\mathrm{Var}[\boldsymbol{r}] = \tau^2\boldsymbol{X}\boldsymbol{X}^T + \upsilon^2\boldsymbol{Z}\boldsymbol{Z}^T$. The joint test of both covariate sets is described by

$$H_0 : \tau^2 = \upsilon^2 = 0 \quad \text{versus} \quad H_1 : \tau^2 \neq 0 \cup \upsilon^2 \neq 0.$$

Menezes et al. [12] showed that ignoring the correlation between the individual test statistics entails little loss of power, and proposed to use the sum of the individual test statistic as a joint test statistic. As mean and variance of the individual test statistics should be brought onto the same scales [12], our joint test statistic is

$$u(\boldsymbol{X}, \boldsymbol{Z}) = \frac{u(\boldsymbol{X}) - \hat{\mathrm{E}}[u(\boldsymbol{X})]}{\sqrt{\widehat{\mathrm{Var}}[u(\boldsymbol{X})]}} + \frac{u(\boldsymbol{Z}) - \hat{\mathrm{E}}[u(\boldsymbol{Z})]}{\sqrt{\widehat{\mathrm{Var}}[u(\boldsymbol{Z})]}}. \tag{8}$$

Permuting as above, we estimate the first two central moments of $u(\boldsymbol{X})$ and $u(\boldsymbol{Z})$ under the null hypothesis, and calculate a p-value for the joint test. Note that this framework can be extended to an arbitrary number of covariate sets. Under k covariate sets the joint test statistic is the standardised sum of k individual test statistics.

Results

Simulation study

We perform a simulation in order to study the power of the proposed test in various circumstances. Instead of randomly generating covariates, we extract a $n \times p$ covariate matrix $\boldsymbol{X}$ from the HapMap data (see below) at a random position. This maintains the correlation structure between SNPs, and thereby ensures a realistic noise level. Initially we set all coefficients in $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ equal to zero. Then we randomly select a subset of r consecutive coefficients, and assign with the probabilities 80 % and 20 % the values s and $2s$ to them, where s is the effect size. Using the relation $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, we calculate the mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$, and simulate the response vector $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ under the distributional assumption $y_i \sim \mathrm{NB}(\mu_i, \phi)$. This procedure ensures that $\boldsymbol{y}$ and $\boldsymbol{X}$ are associated. If we wanted to obtain comparable data under the null hypothesis, we would shuffle the elements in $\boldsymbol{\mu}$. In either case it is of interest how much evidence the proposed test finds for an association between $\boldsymbol{y}$ and $\boldsymbol{X}$.

After simulating numerous response vectors independently and identically, we calculate the specificity and sensitivity of the proposed test at various significance levels, and visualise their relation in a ROC curve. All other things being held equal, we either vary the dispersion parameter ϕ, the sample size n, the effect size s, or the number of non-zero coefficients r. In the last case we do not select another subset of coefficients, but shorten or lengthen the original subset. It is reassuring that the area under the curve changes in the expected directions (see Figure A in the Additional file 1) and that the type I error rates are maintained (see Table A in the Additional file 1).

A slight modification of this simulation study allows to compare the statistical power between testing all covariates at once and testing them one by one. For this we extract various covariate matrices $\boldsymbol{X}$ from the HapMap data, and let the coefficient vector $\boldsymbol{\beta}$ exclusively take non-zero values. For each covariate matrix $\boldsymbol{X}$ we simulate one response vector $\boldsymbol{y}$ under the alternative hypothesis. Using the proposed test, we test the joint as well as the individual significance of the p covariates. Subsequently, we compare the joint p-value with the minimum of the FDR-corrected individual p-values (false discovery rate correction). In our setting with many small effects, joint testing is more powerful than individual testing (see Table B in the Additional file 1). Note that this might not hold in situations with fewer or stronger effects.

Application: HapMap

Here we verify that the proposed test finds biologically meaningful signals, examine whether overdispersion is present, and measure the influence of covariates and samples.

We use the datasets from Montgomery et al. [13] and Pickrell et al. [14] that were made available in a preprocessed form by Frazee et al. [15]. They include RNA-Seq gene expression data for 59 individuals from the population "Utah residents with ancestry from northern and western Europe" (CEU) and 69 individuals from the population "Yoruba in Ibadan, Nigeria" (YRI). Excluding genes outside the 22 autosomes, without any variation within the sample, or without annotations, 11 700 genes are left. For each individual, SNP data is obtained from the International HapMap Consortium [16]. Throughout this application we use the term SNP to designate the number of minor alleles per locus (0, 1 or 2), considered quantitatively.

Stratified permutation test

Considering one gene at a time, its expression level over all individuals is used as a response vector, and the SNPs in the neighbouring region are used as a covariate matrix. The aim is to detect regions where causal SNPs might be. To be precise, we test each of the 11 700 gene expression

vectors for associations with the respective SNPs that are within a window of ± 1 000 base pairs around the gene. This window size leads to $p > n$ for approximately 13 % of the genes, with a maximum of $p = 5\,152$. Under the null hypothesis of no association between gene expression and local SNPs, the p-values would follow a uniform distribution.

Each sample either belongs to the population CEU or to the population YRI, and we account for this grouping by restricting permutations to keeping samples within the same population. As the distribution of p-values is weakly positively skewed, the overall evidence against the null hypotheses is small (see Figure B in the Additional file 1). Only 40 genes reach the minimal p-value given by the reciprocal of the number of permutations (see Table C in the Additional file 1). As in Hulse and Cai [17], we find some genes in the major histocompatibility complex family to be associated with nearby SNPs. Our results display good overlap with the examined results from Lappalainen et al. [18] (see Figure C in the Additional file 1), leading us to conclude that the proposed test identifies biologically meaningful signals.

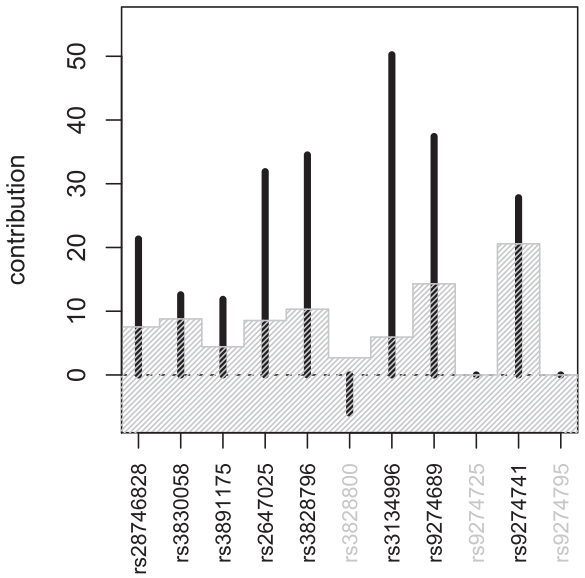

Fig. 1 Contributions of covariates to the test statistic for gene *HLA-DQA2*. The *shaded* area indicates their lower 99 % confidence interval under the null hypothesis

Presence of overdispersion

The reliability of the global test depends on how well the underlying distribution of RNA-Seq gene expression data is approximated. We are interested whether this dataset requires a model with an offset as well as an dispersion parameter, or whether a simpler model would be sufficient.

Fitting under the null hypothesis of no association between gene expression and local SNPs, we observe that the Poisson distribution without an offset has a poor fit, and that including an offset for different library sizes or using the negative binomial distribution improves the fit (see Figure D in the Additional file 1).

In this example the Poisson model with an offset seems to fit almost equally well to the data as the negative binomial with or without an offset. This might be caused by genetic homogeneity within populations or by the absence of diseases. In cancer datasets we expect a much higher variability between individuals (see below).

Individual contributions

For each of the 11 700 tests (one test per gene), the test statistic can be decomposed to show the contribution of individual samples or covariates (Eqs. 5 and 6). By construction these contributions can be positive or negative, but the same holds for their expected values under the null hypothesis. We select two tests (i.e. genes) in order to illustrate these decompositions.

For gene *HLA-DQA2*, most covariates have a larger influence than expected under the null hypothesis (Fig. 1). This suggests that several SNPs might be associated with the expression of the gene. Indeed, if they are tested individually using 10 000 permutations, almost half of them obtain the minimal p-value of 0.0001.

For gene *CIRBP*, the samples from the population CEU tend to contribute positively to the test statistic, whereas those from the population YRI tend to have negative contributions (Fig. 2). Accordingly, the ordinary permutation test would give a much smaller p-value than the stratified permutation test (0.001 versus 0.065). In the case of

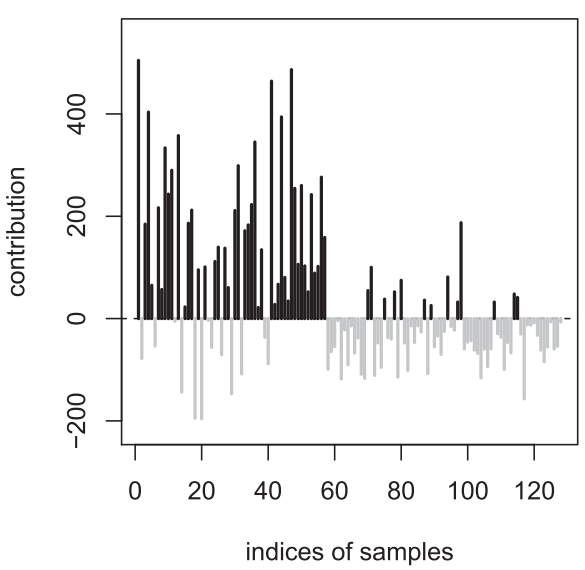

Fig. 2 Contributions of samples to the test statistic for gene *CIRBP*. Samples 1 to 59 are from population CEU, samples 60 to 128 are from population YRI

gene *CIRBP* we cannot detect any sample with an extreme contribution to the test statistic.

Application: TCGA

In this application we illustrate that the proposed test is robust against multicollinearity of the covariates, apply the method of control variables, and test for association with multiple covariate sets simultaneously.

We use a dataset on prostate cancer from TCGA et al. [19]. It includes expression levels of 17 678 genes, DNA methylation levels at 482 486 sites, and DNA copy numbers measured at 30 000 locations for 162 individuals. Section B in the Additional file 1 gives further information about this dataset, including preprocessing. Examining some randomly selected genes, it becomes clear that the Poisson distribution fits badly, but the negative binomial distribution with a free dispersion parameter fits well to the gene expression data (see Figure E in the Additional file 1). Given that the RNA-Seq data has been adjusted for different library sizes, we do not use an offset.

Robustness to multicollinearity

McCarthy, Chen and Smyth [20] developed a test of differential expression between conditions defined by one or more covariates. Taking the design matrix into account when estimating the dispersion parameters, this generalised linear model likelihood-ratio test is powerful for testing small numbers of covariates jointly. However, as in all regression models, multicollinearity may have undesirable consequences.

When testing for associations between gene expression and local genetic or epigenetic variations, high-dimensional situations can occur. Then the likelihood-ratio test breaks down due to singularity, but the global test is still applicable.

But also in low-dimensional situations perfect multicollinearity poses a practical problem. For example, copy number data has a relatively high chance of being perfectly multicollinear, because it correlates highly between locations. If we wanted to apply the likelihood-ratio test nonetheless, we would have to drop some covariates. In contrast, the global test exploits this correlation.

Method of control variables

Here we compare the method of control variables with the crude permutation test, based upon randomly selected genes. Testing the expression of each gene for associations with copy numbers that are within 1 000 000 base pairs around the gene, we estimate the precision of the estimated *p*-values by repeating each permutation test many times. The precision of the estimated *p*-values not only increases with the number of permutations, but according to Table 1 also when switching from the crude permutation test to the method of control variables. For the genes (i.e. tests) in Table 1 the correlation between the two test statistics is sufficiently strong to make this happen, but this is not necessarily true for all genes. However, also in the application HapMap this improvement occurs at all randomly selected genes (see Table D in the Additional file 1). Before deciding between the two methods, we advise to estimate the precision analytically [11].

Table 1 Precision of estimated *p*-values from tests with 100 permutations, estimated from 1,000 repetitions

	EXOSC9	FRMD1	SLC22A6	CNFN	PDHB	U2AF1L4
crude	3.50E+03	4.25E+02	4.13E+02	5.74E+03	1.75E+03	3.38E+03
MCV	2.78E+04	9.17E+04	1.29E+03	1.28E+13	1.89E+04	1.43E+04
	ENTPD6	TMED2	POU6F1	ANP32E	CLDND1	C2orf54
crude	1.91E+03	1.61E+03	1.67E+03	1.38E+05	3.91E+03	5.57E+02
MCV	7.31E+03	8.47E+03	1.50E+04	7.24E+10	3.93E+04	1.12E+03

At all randomly selected genes (*columns*) the crude permutation test (*first row*) is outperformed by the method of control variables (*second row*) in terms of precision

Multiple molecular profiles

Several molecular mechanisms are believed to have an impact on gene expression. In the following, the simultaneous analysis from Eqs. 7 and 8 is applied to chromosome 1. We test for associations between RNA-Seq gene expression data on one hand, and on the other methylation values within $\pm$ 50 000 base pairs, or copy numbers within $\pm$ 2 000 000 base pairs around the start location of the gene. To make the comparison meaningful, the same 1 000 permutations are used for the individual tests and the joint test.

Figure 3 shows: (1) the evidence against null hypotheses is stronger for methylations than for copy numbers; (2) testing methylations and copy numbers jointly leads to an increase in power compared to testing only copy numbers or only methylations; (3) the joint *p*-values are strongly correlated with both sets of individual *p*-values.

Because window sizes are arbitrary, great care is required for biological interpretations of (1). However, (2) and (3) imply that the joint test adds some information to the individual tests. Indeed, in 13 % of the cases the joint test gives smaller *p*-values than both individual tests (Fig. 4). This illustrates the fact that the joint test finds effects that are missed by *both* individual tests. At a false discovery rate of 5 %, Table E in the Additional file 1 lists all genes that are insignificant in both individual tests but significant in the joint test. Extreme examples are the genes *CNKSR1*, *ZNHIT6*, *TMEM56*, *PRPF38B*, and *SLC39A1*, where both individual *p*-values are larger than 0.005, but the joint *p*-values are equal to 0.001. Among these genes, *ZNHIT6* and *SLC39A1* have been linked to prostate or breast cancer [21].

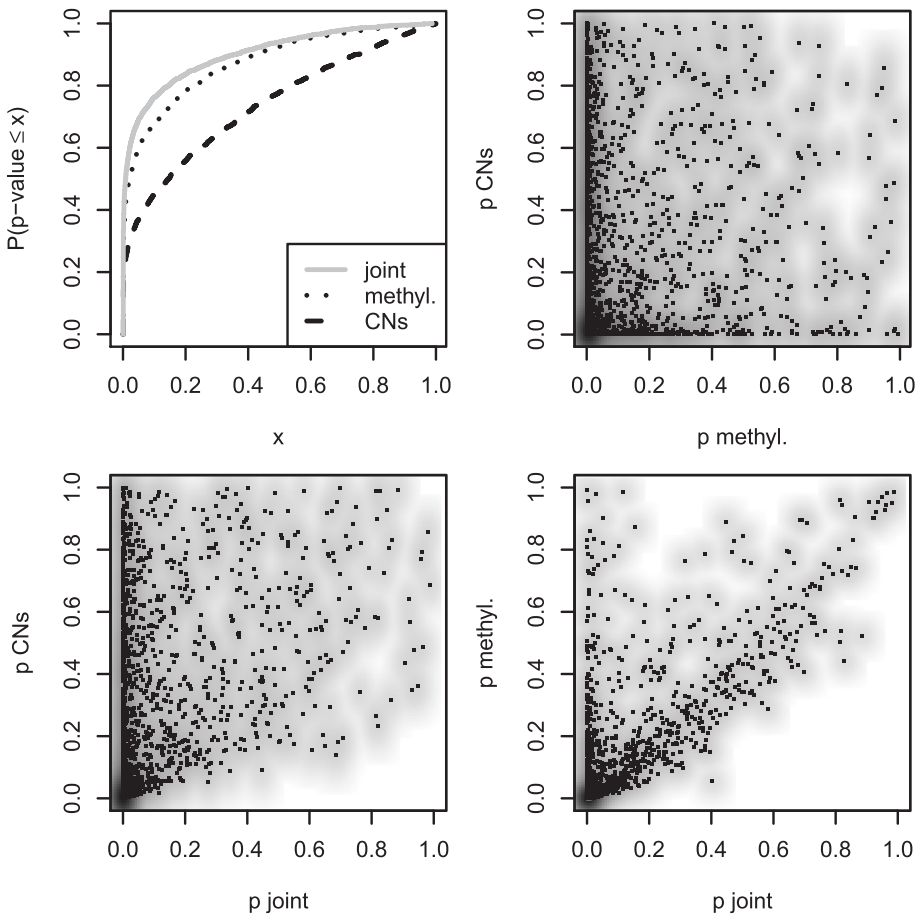

Fig. 3 Empirical cumulative distribution functions and scatterplots of *p*-values. We test for associations between RNA-Seq on one hand, and either copy numbers, methylations or both on the other. The corresponding Spearman correlation coefficients are 0.04 (*top right*), 0.55 (*bottom left*) and 0.72 (*bottom right*)

Discussion

We have proposed a test for association between RNA-Seq data and other molecular profiles. By virtue of the negative binomial distribution, we have accounted for overdispersion in the RNA-Seq data. And owing to a random-effects model, we have allowed for the high dimensionality of the other molecular profiles. Varying library sizes are naturally dealt with by an offset in the model.

We applied the proposed test to detect regulatory mechanisms of gene expression. Thereby we illustrated some of its advantages: (1) stratified permutation allows to account for simple groupings; (2) if overdispersion is absent, the proposed test is equivalent to the one based on the Poisson distribution; (3) the test statistic can be decomposed to show the influence of covariates or samples; (4) the test is applicable in presence of multicollinearity; (5) an extension allows to analyse multiple covariate sets simultaneously.

We use simple offsets and dispersion estimates, but more sophisticated results can easily be integrated into the proposed test. In this regard, sharing information on overdispersion would probably improve the performance of the test under small sample sizes.

The proposed test is based on permutations. Due to the lower multiple testing burden, testing the joint significance of covariates requires much less permutations than testing their individual significance. Even though the computation time for a single test is usually much shorter than one second, genome-wide analyses can be computationally expensive. Running several processes in parallel and interrupting permutation when it becomes impossible to reach a predefined significance level [22] reduces the computation time of a genome-wide analysis to a couple of minutes. If expressions for the mean and the variance of the test statistic were obtained, it would be possible to approximate its null distribution without using permutations. This would allow to obtain significant *p*-values under small sample sizes, and lead to a drastic reduction of computation time. An alternative way of achieving precision as well as speed is the discussed method of control variables.

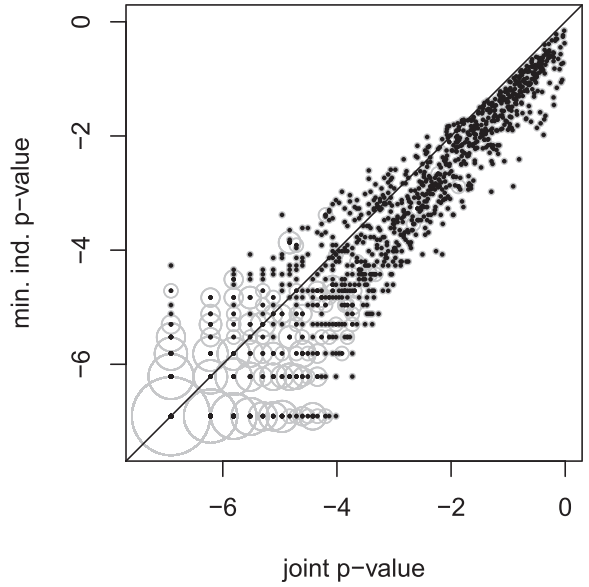

Fig. 4 Scatterplot of logarithmic *p*-values from the simultaneous analysis of multiple covariate sets. *Black points* match the minimal individual *p*-values with the corresponding joint *p*-values. *Grey circles* visualise how often these combinations occur

Conclusions

We have proposed a powerful test for finding eQTL effects based upon RNA-Seq data. It can be computed efficiently and is able to handle sets of highly correlated covariates.

Software

The R package *globalSeq* runs on any operating system equipped with R-3.3.0 or later. It is available from Bioconductor under a free software license: http://bioconductor.org/packages/globalSeq/.

Additional file

Additional file 1: Appendix. Mathematical details, supplementary plots and information on reproducibility. (PDF 487 kb)

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Using ideas from MAJ, MAvdW and RXM, AR developed the method and drafted the manuscript. MAJ, MAvdW and RXM revised the manuscript critically. All authors read and approved the final manuscript.

Acknowledgements

We are grateful to J. J. Goeman for helpful discussions, and to two anonymous reviewers for constructive criticism. We also acknowledge the use of data produced by The Cancer Genome Atlas (TCGA) Research Network to illustrate methods introduced in this work. This research was funded by the Department of Epidemiology and Biostatistics, VU University Medical Center.

Author details

[1]Department of Epidemiology and Biostatistics, VU University Medical Center, 1007 MB, Amsterdam, The Netherlands. [2]Department of Mathematics, VU University, 1081 HV Amsterdam, The Netherlands.

Received: 12 October 2015 Accepted: 18 February 2016
Published online: 08 March 2016

References

1. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics. 2004;20:93–99.
2. Smid M, Wang Y, Zhang Y, Sieuwerts AM, Yu J, Klijn JG, et al. Subtypes of breast cancer show preferential site of relapse. Cancer Res. 2008;68:3108–14.
3. Sanchez-Carbayo M, Socci ND, Lozano J, Saint F, Cordon-Cardo C. Defining molecular profiles of poor outcome in patients with invasive bladder cancer using oligonucleotide microarrays. J Clin Oncol. 2006;24:778–89.
4. Roehle A, Hoefig KP, Repsilber D, Thorns C, Ziepert M, Wesche KO, et al. MicroRNA signatures characterize diffuse large B-cell lymphomas and follicular lymphomas. Br J Haematol. 2008;142:732–44.
5. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11:R106.
6. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.
7. McCullagh P, Nelder JA. Generalized linear models, 2nd ed. London: Chapman and Hall; 1989.
8. Goeman JJ, van de Geer SA, van Houwelingen HC. Testing against a high dimensional alternative. J R Stat Soc Ser B Stat Methodol. 2006;68:477–93.
9. le Cessie S, van Houwelingen HC. Testing the fit of a regression model via score tests in random effects models. Biometrics. 1995;51:600–14.
10. Verbeke G, Molenberghs G. The use of score tests for inference on variance components. Biometrics. 2003;59:254–62.
11. Senchaudhuri P, Mehta CR, Patel NR. Estimating exact p values by the method of control variates or Monte Carlo rescue. J Am Stat Assoc. 1995;90:640–8.
12. Menezes RX, Mohammadi L, Goeman JJ, Boer J. Analysing multiple types of molecular profiles simultaneously: connecting the needles in the haystack. BMC Bioinformatics. 2016;17:77.
13. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010;464:773–7.
14. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010;464:768–72.
15. Frazee AC, Langmead B, Leek JT. ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets. BMC Bioinformatics. 2011;12:449.
16. The International HapMap Consortium. The international HapMap project. Nature. 2003;426:789–96.
17. Hulse AM, Cai JJ. Genetic variants contribute to gene expression variability in humans. Genetics. 2013;193:95–108.
18. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PA, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501:506–11.
19. The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45:1113–20.
20. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012;40:4288–97.
21. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: integrating information about genes, proteins and diseases. Trends Genet. 1997;13:163.
22. van Wieringen WN, van de Wiel MA, van der Vaart AW. A test for partial differential expression. J Am Stat Assoc. 2008;103:1039–49.