Supplementary material for "Exploiting expression patterns across multiple gene isoforms to identify radiation response biomarkers in early-stage breast cancer patients"

Chaitanya R. Acharya, Kouros Owzar, Janet K. Horton and Andrew S. Allen

Our models

Model for differential expression (DE) analysis

$$Y = T\alpha + R\beta + Au + Bv + \xi \tag{1}$$

where T is a $ntg \times t$ dimensional matrix of gene expresion levels in t isoforms of a gene in g groups and n individuals, α is a fixed effect representing the isoform-specific intercepts, R is a $ngt \times 1$-dimensional matrix of radiation dose identifiers such that $R \in \{0, 1\}$, 0 indicates no radiation and 1 indicates radiation, β is a fixed effect indicating the average effect of radiation on gene expression. $u \sim N\left(0, \tau AA^T\right)$ indicates subject-specific random intercept, $v \sim N\left(0, \gamma BB^T\right)$ is random effect that denotes the interaction between gene-isoform and radiation (isoform-specific radiation effect), and $\xi \sim N\left(0, \epsilon I\right)$. I is $ntg \times ntg$ dimensional identity matrix . The matrices J, A and B are design matrices with B being a function of radiation dose. J is $ntg \times t$ dimensional matrix denoting the design matrix for the tissue-specific intercepts. A is $ntg \times n$ design matrix for the subject-specific intercepts. B is a $ntg \times t$ design matrix of stacked radiation dose identifiers.

Parameters of interest are γ, δ, β and ϕ; α, λ, τ, θ and ϵ are nuisance parameters. We test the null hypothesis that $H_0 : \beta = \phi = \gamma = \delta = 0$, i.e. the variant does not affect gene expression across any of the tissues. Our joint score test will test for the effect of genotype on 1) an overall shift in the gene expression, 2) tissue-specific interaction ($G \times T$), 3) overall methylation ($G \times M$), and 4) tissue-specific methylation ($G \times M \times T$). All the random effects are independent to each other.

From the above model, the log-likelihood function conditioned on radiation is –

$$\ell\left(\beta, \theta\right) = c - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\left(Y - T\alpha - R\beta\right)^T \Sigma^{-1}\left(Y - T\alpha - R\beta\right) \tag{2}$$

where θ represents the vector of all the variance components involved in Σ, and β while c is a constant.

From Jiming Jiang's *Linear and Generalized Linear Mixed Models and Their Applications* [4] –

$$\frac{\partial \ell}{\partial \beta} = R^T\Sigma^{-1}Y - R^T\Sigma^{-1}R\beta$$

$$\frac{\partial \ell}{\partial \theta_r} = \frac{1}{2}\left\{\left(Y - T\alpha - R\beta\right)^T \Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_r}\Sigma^{-1}\left(Y - T\alpha - R\beta\right) - \mathrm{Tr}\left(\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_r}\right)\right\}$$

where θ_r is the r^{th} component of θ such that $\theta \in (\tau, \gamma, \epsilon)$.

$$E\left[\frac{\partial^2 \ell_i}{\partial \beta \partial \beta^T}\right] = -R^T\Sigma^{-1}R$$

$$E\left[\frac{\partial^2 \ell_i}{\partial \beta \partial \theta_r}\right] = 0$$

$$E\left[\frac{\partial^2 \ell_i}{\partial\theta_r \partial\theta_s}\right] = -\frac{1}{2}\,\mathrm{Tr}\left(\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_r}\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_s}\right)$$

Let the parameters of interest be $\psi = (\beta, \gamma)^T$ and the nuisance parameters be $\eta = (\alpha, \tau, \epsilon)^T$. The following is constructed under the null (H_0)

$$U_\psi = \begin{pmatrix} \frac{\partial l}{\partial \beta} \\ \frac{\partial l}{\partial \gamma} \end{pmatrix} - \begin{bmatrix} I_{\beta\alpha} & I_{\beta\tau} & I_{\beta\epsilon} \\ I_{\gamma\alpha} & I_{\gamma\tau} & I_{\gamma\epsilon} \end{bmatrix} \begin{bmatrix} I_{\alpha\alpha} & I_{\alpha\tau} & I_{\alpha\epsilon} \\ I_{\tau\alpha} & I_{\tau\tau} & I_{\tau\epsilon} \\ I_{\epsilon\alpha} & I_{\epsilon\tau} & I_{\epsilon\epsilon} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial l}{\partial \alpha} \\ \frac{\partial l}{\partial \tau} \\ \frac{\partial l}{\partial \epsilon} \end{bmatrix}$$

Some algebra will result in the following –

$$U_\beta = \left(R - \bar{R}\right)^T \hat{\Sigma}_n^{-1}\left(Y - T\hat{\alpha}\right) \tag{3}$$

and

$$U_\gamma = \frac{1}{2}\left(Y - T\hat{\alpha}\right)^T \hat{\Sigma}_n^{-1} BB^T \hat{\Sigma}_n^{-1}\left(Y - T\hat{\alpha}\right) \tag{4}$$

$$\begin{aligned} U_\psi &= \left(\boldsymbol{a_\beta} U_\beta^2 + \boldsymbol{a_\gamma} U_\gamma\right) \\ &= \left(Y - T\hat{\alpha}\right)^T \hat{\Sigma}_n^{-1}\left[a_\beta\left(R - \bar{R}\right)\left(R - \bar{R}\right)^T + a_\gamma \frac{1}{2} BB^T\right]\hat{\Sigma}_n^{-1}\left(Y - T\hat{\alpha}\right) \end{aligned} \tag{5}$$

Model for differentially enriched gene-set analysis

$$Y = T\alpha + G\lambda + R\beta + Au + Bv + Cw + \xi \tag{6}$$

where T is a $ntjg \times t$-dimensional matrix of expression levels in t isoforms of a gene, j genes, g groups and n individuals, α is a fixed effect representing t isoform-specific intercepts, λ is a fixed effect representing g gene-specific intercepts, R is a $ntjg \times 1$ dimensional matrix of radiation dose identifiers such that $R \in \{0, 1\}$, 0 indicates no radiation and 1 indicates radiation, β is a fixed effect indicating the average effect of radiation on a pathway or gene-set. $u \sim N\left(0, \tau AA^T\right)$ indicates subject-specific random intercept, $v \sim N\left(0, \gamma BB^T\right)$ is a random effect that denotes the interaction between gene-isoform and radiation (isoform-specific radiation effect), $w \sim N\left(0, \phi CC^T\right)$ is a random effect that denotes the interaction between gene and radiation (gene-specific radiation effect), and $\xi \sim N\left(0, \epsilon I\right)$. I is $ntjg \times ntjg$-dimensional identity matrix . The matrices J, A and B are design matrices with B being a function of radiation dose. J is $ntjg \times t$ dimensional matrix denoting the design matrix for the tissue-specific intercepts. A is $ntjg \times n$ design matrix for the subject-specific intercepts. B is a $ntjg \times t$ design matrix of stacked radiation dose identifiers and C is a $ntjg \times g$ dimensional design matrix of the $R \times G$ effect. All the random effects are independent to each other.

The log-likelihood function conditioned on the radiation is given by –

$$\ell\left(\beta, \theta\right) = c - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\left(Y - T\alpha - G\lambda - R\beta\right)^T \Sigma^{-1}\left(Y - T\alpha - G\lambda - R\beta\right) \tag{7}$$

where θ represents the vector of all the variance components involved in Σ, and β while c is a constant.

Let the parameters of interest be $\psi = (\beta, \gamma, \phi)^T$ and the nuisance parameters be $\eta = (\alpha, \lambda, \tau, \epsilon)^T$. The following can be constructed under the null (H_0) –

$$U_\zeta = \begin{bmatrix} \frac{\partial l}{\partial \beta} \\ \frac{\partial l}{\partial \gamma} \\ \frac{\partial l}{\partial \phi} \end{bmatrix} - \begin{bmatrix} I_{\beta\alpha} & I_{\beta\lambda} & I_{\beta\tau} & I_{\beta\epsilon} \\ I_{\gamma\alpha} & I_{\gamma\lambda} & I_{\gamma\tau} & I_{\gamma\epsilon} \\ I_{\phi\alpha} & I_{\phi\lambda} & I_{\phi\tau} & I_{\phi\epsilon} \end{bmatrix} \begin{bmatrix} I_{\alpha\alpha} & I_{\alpha\lambda} & I_{\alpha\tau} & I_{\alpha\epsilon} \\ I_{\lambda\alpha} & I_{\lambda\lambda} & I_{\lambda\tau} & I_{\lambda\epsilon} \\ I_{\tau\alpha} & I_{\tau\lambda} & I_{\tau\tau} & I_{\tau\epsilon} \\ I_{\epsilon\alpha} & I_{\epsilon\lambda} & I_{\epsilon\tau} & I_{\epsilon\epsilon} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial l}{\partial \alpha} \\ \frac{\partial l}{\partial \lambda} \\ \frac{\partial l}{\partial \tau} \\ \frac{\partial l}{\partial \epsilon} \end{bmatrix}$$

Some algebra will result in the following –

$$U_\beta = \left(R - \bar{R}\right)^T \hat{\Sigma}_n^{-1} \left(Y - T\hat{\alpha} - G\hat{\lambda}\right) \tag{8}$$

$$U_\gamma = \frac{1}{2} \left(Y - T\hat{\alpha} - G\hat{\lambda}\right)^T \hat{\Sigma}_n^{-1} BB^T \hat{\Sigma}_n^{-1} \left(Y - T\hat{\alpha} - G\hat{\lambda}\right) \tag{9}$$

$$U_\phi = \frac{1}{2} \left(Y - T\hat{\alpha} - G\hat{\lambda}\right)^T \hat{\Sigma}_n^{-1} CC^T \hat{\Sigma}_n^{-1} \left(Y - T\hat{\alpha} - G\hat{\lambda}\right) \tag{10}$$

$$\begin{aligned} U_\zeta &= \left(\boldsymbol{a_\beta} U_\beta^2 + \boldsymbol{a_\gamma} U_\gamma + \boldsymbol{a_\phi} U_\phi\right) \\ &= \left(Y - T\hat{\alpha} - G\hat{\lambda}\right)^T \hat{\Sigma}_n^{-1} \left[a_\beta \left(R - \bar{R}\right)\left(R - \bar{R}\right)^T + a_\gamma \frac{1}{2} BB^T + a_\phi \frac{1}{2} CC^T\right] \hat{\Sigma}_n^{-1} \left(Y - T\hat{\alpha} - G\hat{\lambda}\right) \end{aligned} \tag{11}$$

Optimal weights a_β, a_γ and a_ϕ were derived using Lagrangians.

Null simulations

We carried out two simulation studies for each method. While the results from the power simulations were made available in the main manuscript, here are the tables showing the results from 'null' simulations in order to estimate the type I error at $\alpha = 0.05$.

Null simulations for the DE score test statistic

We evaluated our method to detect DE genes using two simulation studies. Here we present results from null simulations from both simulation studies. Briefly, each Monte Carlo simulated dataset from the first simulation study was comprised of data for a single gene, whose expression is measured across 5 or 10 transcripts in 50 paired individuals. Each individual pair's radiation status is either a zero or a one indicating before and after radiotherapy, respectively. Since the transcript-specific effect is modeled as a random effect, a test of whether there is any transcript-specific effect due to radiation is equivalent to testing whether the variance of the random effect (γ) is zero.

Test	Type I error	Lower CI	Upper CI
DE Score Test	0.0481	0.04398826	0.0524774
TBT paired t-test	0.0469	0.04162092	0.05261547
TBT Wilcoxon test	0.0438	0.03869686	0.04934269
Gene-leve paired t-test	0.0449	0.04092429	0.04914339

Table 1: DE of genes - Null Simulation results from our first simulation study at 5% FDR with 95% confidence interval. Our score test is referred to as "DE Score Test".

In the second simulation study, each Monte Carlo dataset, comprised of gene expression data for 50 genes over 50 observations, each gene with unequal number of isoforms, was simulated from a multivariate normal distribution with a known variance-covariance matrix. We varied the mean difference in differential gene expression between the two phenotypes, and the proportion of differentially expressed gene-isoforms. At the transcript level, we applied paired t-test and a non-parametric alternative in Wilcoxon's paired t-test and combined the p values over all the transcripts of a gene using Fisher's method. At the gene-level, we combined the gene expression values by computing either the median or Winsorized mean of all the transcripts within a given gene. Paired t-test was run on this gene-level data.

	DE Score Test	TBT Paired t-test	TBT Wilcoxon's test	Gene-level paired t-test
Type I error	0.0578	0.0528	0.0456	0.0528

Table 2: DE of genes - Null Simulation results at 5% FDR from our second simulation study. Our score test is referred to as "DE Score Test".

Null simulations for the gene-set enrichment score test statistic

We evaluated our method to detect DE gene-sets or pathways using two simulation studies. Here we present results from two simulations studies. Briefly, each Monte Carlo simulated dataset from the first simulation study was comprised of data for a single gene-set comprising of 5 genes, whose expression is measured across 3 transcripts in 50 paired individuals. Each individual pair's radiation status is either a zero or a one indicating before and after radiotherapy, respectively. Since the transcript-specific effect is modeled as a random effect, a test of whether there is any transcript-specific effect on the gene-sets due to radiation is equivalent to testing whether the variances of the random effects (γ and ϕ) are zero.

Test	Type I error	Lower CI	Upper CI
Gene-set Score Test	0.0436	0.03968114	0.04778736
TBT paired t-test	0.0505	0.04489782	0.05655453

Table 3: Gene-set enrichment analysis - Null Simulation results from our first simulation study at 5% FDR with 95% confidence interval. Our score test is referred to as "Gene-set Score Test".

In our second simulation study, each Monte Carlo simulation consisted of 100 genes over 5 observations across the two phenotypes. We generated gene expression data using the same approach as described in the previous section. We simulated 10 gene-sets under both scenarios (with non-overlapping and overlapping genes) and compared the performance of our method with the other gene-set enrichment methods at the gene-level. We varied the sizes of gene-sets between 2 and 10 genes. Gene-level analysis is performed by computing the median gene expression values across all the transcripts within a gene followed by an implementation of gene set variational analysis (GSVA), Pathway Level analysis of Gene Expression (PLAGE), single sample GSEA (ssGSEA) and the combined z-score (ZSCORE). We estimated the empirical type I error rate at 5% FDR both in the presence and absence of any gene overlap among the simulated gene-sets.

	Gene-set overlap	Gene-set Score Test	GSVA	PLAGE	ssGSEA	ZSCORE
Type I error	Yes	0.053	0.057	0.054	0.052	0.044
Type I error	No	0.045	0.057	0.050	0.056	0.049

Table 4: Gene-set enrichment analysis - Null Simulation results from our second simulation study at 5% FDR. We present type I error rates for two cases - gene-sets share genes (overlap) and gene-sets with unique genes (no overlap). Our score test is referred to as "Gene-set Score Test".

Gene-set analysis methods

We compared the performance of our score test method for gene-set analysis with gene-set variational analysis (GSVA) [3], pathway level analysis of gene expression (PLAGE) [5], the combined z-score method (ZSCORE) [2] and single sample GSEA (ssGSEA) [1]. Briefly, PLAGE standardizes expression profiles into z-scores over the samples and then calculates the singular value decomposition $Z = UDV'$ on the z-scores of the genes in the gene-set. The coefficients of the first right-singular vector (first column of V) are taken as the gene-set summaries of expression over the samples. ZSCORE method also standardizes expression profiles into z-scores over the samples, but combines them together for each gene-set at each individual sample in the following way.

Given a gene-set $\gamma = \{1, \cdots, k\}$ with z-scores $Z_1, \cdots, Z_k$ for each gene, the combined z-score Z_γ for the gene-set γ is $\frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}}$. ssGSEA, on the other hand, calculates a gene-set enrichment score per sample as the normalized difference in empirical cumulative distribution functions of gene expression ranks inside and outside the gene-set.

All of these methods do not perform analysis at the transcript-level. In order to apply the aforementioned methods on both simulated and real data, we combined gene expression values over all the transcripts for a given gene by computing the median expression values thus, performing the analysis at the gene-level to obtain a matrix single sample enrichment scores. Paired t-test was then performed on this matrix.

Reproducibility of the analysis

All the scripts and the accompanied documentation for reproducing our analyses are located at `https://github.com/cramanuj/Radiation_Rcodes`.

References

[1] David A. Barbie, Pablo Tamayo, Jesse S. Boehm, So Young Kim, Susan E. Moody, Ian F. Dunn, Anna C. Schinzel, Peter Sandy, Etienne Meylan, Claudia Scholl, Stefan Frohling, Edmond M. Chan, Martin L. Sos, Kathrin Michel, Craig Mermel, Serena J. Silver, Barbara A. Weir, Jan H. Reiling, Qing Sheng, Piyush B. Gupta, Raymond C. Wadlow, Hanh Le, Sebastian Hoersch, Ben S. Wittner, Sridhar Ramaswamy, David M. Livingston, David M. Sabatini, Matthew Meyerson, Roman K. Thomas, Eric S. Lander, Jill P. Mesirov, David E. Root, D. Gary Gilliland, Tyler Jacks, and William C. Hahn. Systematic rna interference reveals that oncogenic kras-driven cancers require tbk1. *Nature*, 462(7269):108–112, 11 2009.

[2] Lee E, Chuang H-Y, Kim J-W, Ideker T, and Lee D. Inferring pathway activity toward precise disease classification. *PLoS Computational Biology*, 4(11):e1000217. doi:10.1371/journal.pcbi.1000217, 2008.

[3] S Hänzelmann, R Castelo, and J Guinney. Gsva: gene set variation analysis for microarray and rna-seq data. *BMC Bioinformatics*, 14(7):DOI: 10.1186/1471–2105–14–7, 2013.

[4] J Jiang. *Linear and Generalized Linear Mixed Models and Their Applications*. Springer Series in Statistics. Springer-Verlag New York, 233 Springer Street, New York, NY 10013, USA, 1 edition, 2007.

[5] J Tomfohr, J Lu, and T.B. Kepler. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 6(225):DOI: 10.1186/1471–2105–6–225, 2005.