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

Abstract

In an effort to understand the underlying biology of radiation response along with whole transcriptome effects of preoperative radiotherapy in early-stage breast tumors, we propose two new statistical methods that exploit gene expression patterns across all available gene transcript isoforms and identify potential biomarkers, which are representative of tumor microenvironment and tumor biology of radiation response. We demonstrate the effectiveness of these methods using extensive simulation studies that show that both of our methods give an improved performance, in terms of statistical power, over the most commonly used methods. By exploiting radiation-induced changes in all available gene transcript isoforms i.e. human transcriptome, we identified several statistically significant differentially genes related to PI3K-AKT and JAK-STAT signaling pathways along with radiation-induced oncogenic signaling pathways and tumor microenvironment gene signatures that could be potential targets to improve response to radiotherapy in breast tumors.

Background

Radiation therapy or radiotherapy is utilized as a curative therapy in many solid tumors including gynecologic, head and neck, gastrointestinal, breast, prostate, lung, central nervous system and pediatric malignancies. Approximately 60% of cancer patients receive radiotherapy as part of their treatment either as a stand alone pre-operative therapy or combined with other modalities such as chemotherapy following surgery in an adjuvant setting. Radiotherapy has played a significant role in treating both invasive and non-invasive breast tumors over the years. However, response to radiation in breast cancer patients has not been uniform across all breast tumor subtypes (for example, basal, luminal, etc.) leading to a significant percentage of patients being either over- or under-treated [1, 2]. This can be attributed to variable transcriptional response to radiation, which is very similar to response to chemotherapy except that the mechanisms underlying radiation response have not been well understood and studied [3, 4].

Constantly evolving high-throughput gene expression profiling technologies, such as RNA-Seq or ultra high-resolution microarrays, have enabled us to interrogate all transcript isoforms in the human transcriptome with probes designed to target coding transcripts, exon-exon splice junctions, and non-coding transcripts. The end goal of using these technologies is to exploit the gene expression patterns across multiple isoforms or transcripts of a gene in order to map biomarkers such as genes and pathways that help illuminate the molecular pathology of complex diseases at the RNA level. Existing analytic tools or methods for biomarker analysis involves combining gene expression over all gene isoforms or transcripts prior to data analysis, resulting in a gene-level interrogation of biological conditions. Such an approach has at least two significant limitations. First, it fails to fully exploit expression patterns across gene isoforms either by combining information across multiple transcripts or by not explicitly identifying effects that differ across transcripts. Second, and more importantly, it fails to account for alternative splicing or poly-adenylation events. We propose two distinct approaches, one to identify treatment-induced (for example, radiation treatment) gene expression biomarkers in an isoform-specific differential expression (DE) analysis and another to perform isoform-specific gene-set or pathway enrichment analysis. We test these methods extensively using simulation studies and then evaluate the effectiveness of these two methods on an array-based gene expression dataset containing 26 paired breast tumor samples. These samples originated from a clinical trial and were treated with preoperative radiotherapy. We specifically investigate the effects of radiation on tumor microenvironment (TM) and some hallmark oncogenic signaling pathways that can make for potential targets in any downstream analysis.

We hypothesize that our methods are effective in identifying more biomarkers than the most commonly used approaches. Investigating the tumor microenvironment and the oncogenic signaling pathways before and after radiation will help us understand any radiation-induced changes in individual patients, which may serve as a surrogate to understand patient response to radiation.

Methods

Microarray analysis of the breast cancer dataset

Raw microarray data for sixty two early-stage breast cancer patients were obtained from NCBI's gene expression omnibus (GEO ID: GSE65505) repository. All the patients are at least 55 years old, clinically node negative, ER-positive and/or PR-positive, HER2-negative (biologically favorable tumors) with T1 invasive carcinomas or low-intermediate grade *in situ* disease $\leq 2cm$. These patients received pre-operative radiotherapy (radiation dose prior to surgical resection of tumor). All the samples were arrayed on Affymetrix Human Transcriptome Array 2.0 [5], which was designed with approximately ten probes per exon and four probes per exon-exon splice

junction. At the top level, each transcript cluster roughly corresponds to a gene. Each transcript cluster is comprised of exon clusters that a) shared splice sites, b) or were derived from overlapping exonic sequences, c) or were single-exon clusters bounded on the genome by spliced content. Each exon cluster is further fragmented into probe selection regions (PSRs), which are non-overlapping contiguous sequences. Gene-level and gene isoform/transcript-level expression data were obtained using R/Bioconductor packages *oligo*, *affyio* and *pd.hta.2.0*, and pre-processed by robust multi-array average (RMA) method, which summarizes the probe level expression data into a probe set level expression value. Principal component analysis was conducted to check for batch effects in both gene-level and transcript-level data, and any batch effects that were identified were corrected using a popular Empirical Bayes approach (ComBat). DE analysis was performed on genes with at least two transcript isoforms. This resulted in a dataset with more than 800,000 transcripts.

Strategy to identify gene expression biomarkers of radiation: Differential Expression (DE) analysis

Given two distinct biological groups (before and after radiation treatment), gene expression for each gene can be modeled in the following way

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

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 $\xi \sim N(0, \epsilon I)$. 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.

We test the null hypothesis that $H_0 : \beta = 0; \gamma = 0$ i.e radiation does not affect gene expression. From our model above, we derive our score test statistic, U_ψ as

$$U_\psi \equiv \hat{Y}^T \hat{\Sigma}^{-1} \left[a_\beta \left(R - \bar{R}\right)\left(R - \bar{R}\right)^T + a_\gamma \left(\frac{1}{2}BB^T\right) \right] \hat{\Sigma}^{-1} \hat{Y}, \tag{2}$$

where a_β and a_γ are scalar constants chosen to minimize the variance of U_ψ (see Supplementary methods). The p values are approximated using Satterthwaite method [6]. The maximum likelihood estimates, obtained from fitting a standard linear mixed model using lme4 [7], are computed only once per gene since under the null, there is no effect due to radiation on the gene expression. The p values obtained from applying our method were adjusted for multiple hypothesis within the false discovery rate (FDR) framework. Genes with FDR adjusted p values (q values) less than 0.05 were selected to be differentially expressed. More information on our method is available in supplementary methods.

Strategy to perform radiation-induced isoform-specific gene-set enrichment analysis

Gene expression data for each pathway is modeled in the following way

$$Y = T\alpha + G\lambda + R\beta + Au + Bv + Cw + \xi \tag{3}$$

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 $\xi \sim N(0, \epsilon I)$. 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.

We test the null hypothesis that $H_0 : \beta = 0; \gamma = 0; \phi = 0$ i.e radiation does not affect gene expression. From our model above, we derive our score test statistic, U_ζ as

$$U_\zeta \equiv \hat{Y}^T \hat{\Sigma}^{-1} \left[a_\beta \left(R - \bar{R}\right)\left(R - \bar{R}\right)^T + a_\gamma \left(\frac{1}{2}BB^T\right) + a_\phi \left(\frac{1}{2}CC^T\right) \right] \hat{\Sigma}^{-1}\hat{Y}, \tag{4}$$

where a_β, a_γ and a_ϕ are scalar constants chosen to minimize the variance of U_ζ. The p values are approximated using Satterthwaite method [6]. Similar to our earlier method, the maximum likelihood estimates, obtained from fitting a standard linear mixed model using lme4 [7], are computed only once per gene-set since under the null,

there is no effect due to radiation on the gene expression. The p values obtained from applying our method were adjusted for multiple hypothesis within the false discovery rate (FDR) framework. Genes with FDR adjusted p values (q values) less than 0.05 were selected to be differentially expressed. More details on our method are available in supplementary methods.

Simulations

Testing our method for DE analysis

We have performed the following two simulation studies in order to verify our approach. In our first study, we simulated one gene at a time from the following linear model and varied the following parameters- β (additive effect due to radiation), the proportion of variation explained by γ or $R \times T$ effect $\left(PVE_\gamma \equiv \left(\frac{\gamma}{\tau + \epsilon} \right) \right)$ and the number of transcripts. For a positive integer tg that represents the combined number of transcripts (t) and groups (g), if $\mathbf{1}$ denotes a column vector of tg ones and $\mathbb{I}$ denotes the corresponding $tg \times tg$ diagonal matrix, following the tg-variate normal law denoted by $N_{tg}[\mu, \Sigma]$ with mean $\mu \in \mathbb{R}^{tg}$ and variance $\Sigma \in \mathbb{R}^{tg \times tg}$, expression levels of a target gene j by using the following vectorized form of the linear mixed model –

$$y_{ijg} = \alpha_j + \beta_j r_g + \mathbf{1} a_i + b_j r_g + \xi_{ijg} \qquad \xi_{ijg} \overset{i.i.d.}{\sim} N(0, \epsilon \mathbb{I}) \tag{5}$$

where y_{ijg} is a $tg \times 1$ vector of gene expression data, α_t is the transcript-specific intercept ($\alpha_t \in \mathbb{R}^t$), β_j describes the main additive effect ($\beta_j \in \mathbb{R}^1$), r_g is a vector of length tg such that $r \in (0_t, 1_t)$. The random effect $b_j \in \mathbb{R}^{tg}$ represents transcript-specific interaction effect of radiation, and $a_i \in \mathbb{R}^1$ is a subject-specific random intercept. We assume that all the random effects are independent and that $a_i \sim N_1(0, \tau)$, $b_j \sim N_{tg}(0, \gamma \mathbb{I})$. A linear mixed effects model was fit using the package *lme4*[7] in the statistical environment R (R Core Team).

We then compared our method with a standard paired t-test and a non-parametric alternative in Wilcoxon's test [8]. The test statistic in case of transcript-by-transcript (TBT) analysis is the minimum p value over the total number of transcripts from either t-test or Wilcoxon's test performed separately in each transcript for each paired sample. A gene-level test was constructed over all the transcripts by taking the median expression value across the transcripts followed by a standard paired t-test. Statistical significance was determined at a nominal p value of 0.05 for all power simulations (in case of TBT analysis, it is $\frac{0.05}{k}$, where k is the number of transcripts). We used 10,000 data replicates to evaluate the type I error and 1,000 data replicates for power calculations.

We have also tested our method on a synthetic dataset simulated from a multivariate normal distribution containing two classes of data. Each gene was simulated to have variable number of transcripts. We used this dataset

with increasing number of genes (by also keeping a small proportion differentially expressed) and tested our approach at both transcript-level (paired t-test and Wilcoxon's test) and gene-level. The most commonly used method to combine p values of all the transcripts of a gene is Fisher's method however, under the assumption that all the p values are independent. This assumption may be frequently violated since different isoforms of a gene may be correlated and the resulting p values are dependent on each other. At the gene-level, paired t-tests were run on gene expression values of a gene that were aggregated over its transcripts by either their median expression values or Winsorized mean [9] expression values.

Testing our method for gene-set enrichment analysis

Similar to the above analyses, we have performed two simulations studies in order to verify our approach. In our first study, we simulated one gene-set at a time from the following linear model and varied the following parameters- β (additive effect due to radiation), the proportion of variation explained by γ or $R \times T$ effect $\left(PVE_\gamma \equiv \left(\frac{\gamma}{\tau+\phi+\epsilon}\right)\right)$, the proportion of variation explained by ϕ or $R \times G$ effect $\left(PVE_\phi \equiv \left(\frac{\phi}{\tau+\phi+\epsilon}\right)\right)$ and the number of transcripts. For a positive integer tjg that represents the combined number of transcripts (t), genes (j) and groups (g), if $\mathbf{1}$ denotes a column vector of tjg ones and $\mathbb{I}$ denotes the corresponding $tjg \times tjg$ diagonal matrix, following the tjg-variate normal law denoted by $N_{tjg}[\mu, \Sigma]$ with mean $\mu \in \mathbb{R}^{tjg}$ and variance $\Sigma \in \mathbb{R}^{tjg \times tjg}$, expression levels of a target geneset k by using the following vectorized form of the linear mixed model –

$$y_{ijkg} = \alpha_j + \beta_k r_g + \lambda_k + \mathbf{1}a_i + b_j r_g + c_k r_g + \xi_{ijkg} \qquad \xi_{ijkg} \overset{i.i.d.}{\sim} N(0, \epsilon\mathbb{I}) \qquad (6)$$

where y_{ijkg} is a $tjg \times 1$ vector of gene expression data, α_t is the transcript-specific intercept ($\alpha_t \in \mathbb{R}^t$), β_k describes the main additive effect ($\beta_k \in \mathbb{R}^1$), r_g is a vector of length tjg such that $r \in (0_{tg}, 1_{tg})$. The random effect $b_t \in \mathbb{R}^{tjg}$ represents transcript-specific interaction effect of radiation, the random effect $c_j \in \mathbb{R}^{tjg}$ represents transcript-specific interaction effect of radiation, and $a_i \in \mathbb{R}^1$ is a subject-specific random intercept. We assume that all the random effects are independent and that $a_i \sim N_1(0, \tau)$, $b_t \sim N_{tg}(0, \gamma\mathbb{I})$ and $c_j \sim N_{jg}(0, \gamma\mathbb{I})$. A linear mixed effects model was fit using the package *lme4*[7] in the statistical environment R (R Core Team).

We then compared our method with a standard paired t-test and a non-parametric alternative in Wilcoxon's test [8]. The test statistic in case of transcript-by-transcript (TBT) within a gene analysis is the minimum p value over the total number of transcripts and genes from either t-test or Wilcoxon's test performed separately in each transcript for each paired sample. A gene-level test was constructed over all the transcripts by taking the median expression value across the transcripts followed by a standard paired t-test. Statistical significance was determined at a nominal p value of 0.05 for all power simulations (in case of TBT analysis, it is $\frac{0.05}{k}$, where k is

the product of the number of transcripts and genes). We used 10,000 data replicates to evaluate the type I error and 1,000 data replicates for power calculations.

A second set of simulations involved generating a synthetic gene expression data from a multivariate normal distribution containing two classes of data. Each gene was simulated to have variable number of transcripts. We defined two types of gene-sets, one with overlapping genes and the other with non-overlapping genes, and randomly assigned some gene-sets to to contain differentially expressed genes. Since most, if not all of the current methods involve gene-set analysis at the gene level, we compared our method with Gene Set Variational Analysis (GSVA) [10], Pathway Level Analysis of Gene Expression (PLAGE) [11], single sample GSEA (ssGSEA) [12] and the combined z-score (ZSCORE) [13] methods. Both, PLAGE and the ZSCORE are parametric and assume that gene expression profiles are jointly normally distributed. More about these methods in the supplementary material.

Defining the gene-sets and gene-set analysis

All the hallmark oncogenic signaling pathways used in our primary data analysis were obtained from the Molecular Signature Database version 3 (MSigDB) collection []. We focussed our attention on 24 specific oncogenic signaling pathways that were most likely associated with radiation response. We defined tumor microenvironment as a collection of proteins produced by cells present in and around the tumor that support the growth of the cancer cells. We included gene-sets representing hypoxia [14], invasiveness/metastases gene signature [15], epigenetic stem cell signature in cancer [16], inflammatory pathway involving tumor necrosis factors [17], angiogenesis [18], immune signatures [19] and a form of genomic instability called chromosomal instability [20], which determines the tumor cell's ability to respond to its microenvironment. In order to visualize sample set enrichment of these gene-sets (enrichment level of a gene-set in a sample), we employed Gene Set Analysis (GSA) software [21], which implements a supervised method (class labels are known before the analysis) that computes a "maxmean" summary statistic for each gene-set. Briefly, GSA computes the average of both positive and negative aspects of gene-scores (for example, fold changes) over each gene in a gene-set, and choose the one that is larger in absolute value [22]. These values were clustered using hierarchical clustering algorithm (with Euclidian distance metric and complete linkage method) [23].

Multiple hypothesis correction

Wherever applicable, we use multiple hypothesis correction based on the Benjamini-Hochberg (BH) approach [24] to obtain corrected p values. In case of gene-set analysis, BH approach may result in a conservative estimate

of the false discovery rate (FDR) because of overlapping gene-sets that have highly correlated genes. We used the BH method only as a demonstration of statistical power.

Results

Evaluating our method to identify differentially expressed (DE) genes using simulated data

We evaluated our method to detect DE genes using two 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. Thus, our model to detect DE genes involves testing two scalar parameters in β and γ. Simulations under the null hypothesis (no effect of radiation on overall gene expression) confirm that our method has the right type I error. More details in the supplementary section.

Additive Effect	$PVE_\gamma(\%)$	DE Score Test	TBT Paired t-test	TBT Wilcoxon's test	Gene-level paired t-test
		Number of transcripts per gene = 5			
NO	0	0.051 [0.038-0.067]	0.052 [0.036-0.073]	0.054 [0.037-0.075]	0.044 [0.032-0.059]
NO	9	0.36 [0.33-0.391]	0.291 [0.255-0.329]	0.263 [0.228-0.3]	0.114 [0.095-0.135]
NO	13	0.629 [0.598-0.659]	0.536 [0.495-0.577]	0.504 [0.463-0.545]	0.205 [0.18-0.231]
YES	0	0.373 [0.343-0.404]	0.259 [0.224-0.296]	0.239 [0.205-0.275]	0.385 [0.355-0.416]
YES	9	0.634 [0.603-0.664]	0.515 [0.474-0.556]	0.493 [0.452-0.534]	0.418 [0.387-0.449]
YES	13	0.759 [0.731-0.785]	0.66 [0.62-0.698]	0.627 [0.587-0.666]	0.447 [0.416-0.478]
		Number of transcripts per gene = 10			
NO	0	0.053 [0.04-0.069]	0.043 [0.027-0.064]	0.039 [0.024-0.059]	0.059 [0.045-0.075]
NO	9	0.534 [0.503-0.565]	0.352 [0.31-0.396]	0.318 [0.277-0.361]	0.135 [0.114-0.158]
NO	13	0.861 [0.838-0.882]	0.682 [0.639-0.723]	0.642 [0.598-0.684]	0.21 [0.185-0.237]
YES	0	0.539 [0.508-0.57]	0.302 [0.262-0.344]	0.264 [0.226-0.305]	0.646 [0.615-0.676]
YES	9	0.831 [0.806-0.854]	0.633 [0.589-0.675]	0.588 [0.543-0.632]	0.63 [0.599-0.66]
YES	13	0.92 [0.901-0.936]	0.832 [0.796-0.864]	0.803 [0.766-0.837]	0.604 [0.573-0.634]

Table 1: *DE of genes - Simulation results at 5% FDR with 95% confidence interval. We varied additive effect i.e. average effect of radiation on the whole transcriptome and proportion of variation explained by γ i.e. radiation $\times$ transcripts interaction effect. Our score test is referred to as "DE Score Test".*

Power simulations were done by varying the following parameters- 1) additive effect of radiation (β), 2) the proportion of variation explained by the interaction effect between radiation and transcript (PVE_γ) and 3) the number of transcripts. Table 1 shows that our method significantly outperforms transcript-by-transcript paired t-test and Wilcoxon test (a non-parametric alternative to t-test) in all simulated situations. However, the gene-level paired t-test seems to work the best when there is an overall shift in gene expression due to radiation but

absence of any transcript-specific effects.

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. We varied the proportion of genes that are differentially expression and the mean difference between the genes that are differentially expressed and those that aren't. Statistical power and empirical type I error rates were estimated at 5% FDR. Figure 1 displays the performance of all the methods, measured both in terms of statistical power and area under the curve (AUC). AUC for all the methods was estimated using R package ROCR [25]. We see that our method does well compared to the rest of the methods based on AUC plot. Given how the gene expression data was generated, every gene may have a fraction of transcripts differentially expressed. Consequently, any method to identify DE genes must account for this transcript-specific variability. By combining gene expression values over all the transcripts of a gene (as evidenced by any gene-level methods), we are not able to fully exploit transcript-specific gene expression patterns. This is evident in Figures 1a and 1b, where the gene-level tests perform poorly compared to the transcript-level tests, including our approach.

Evaluating our method to identify DE gene-sets using simulated data

We evaluated our method to detect DE gene-sets or pathways using two simulation 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. Thus, our model to detect enriched gene-sets involves testing three scalar parameters in β, γ and ϕ. Simulations under the null hypothesis (no effect of radiation on overall gene expression) confirm that our method has the right type I error (see supplementary material).

Power simulations were done by varying the following parameters- 1) additive effect of radiation (β), 2) the
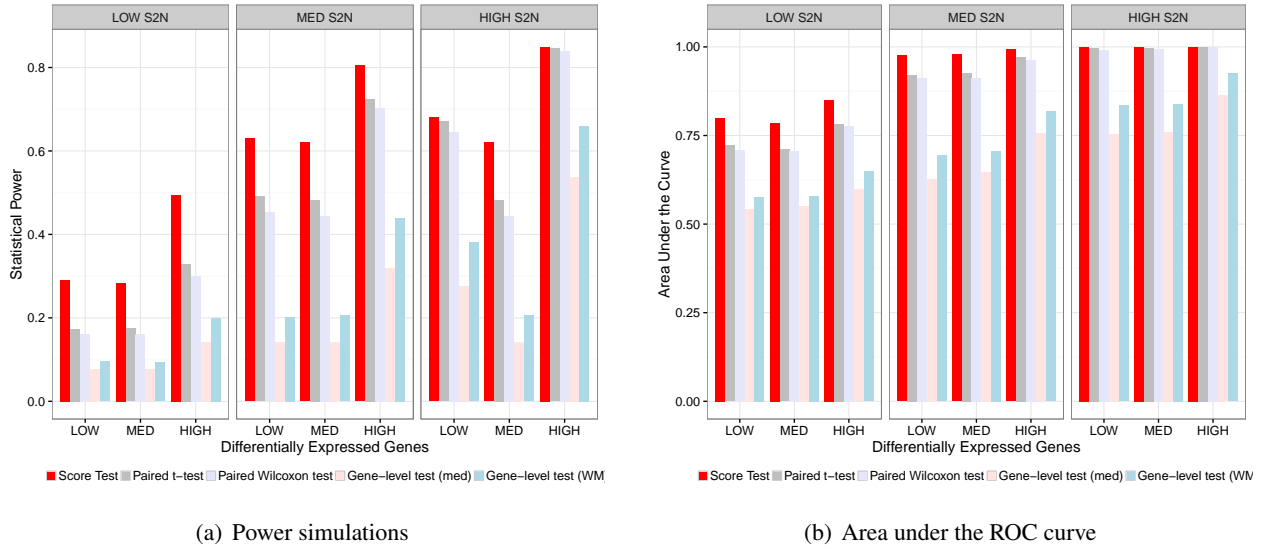
(a) Power simulations

(b) Area under the ROC curve

Figure 1: The performance of all the methods in detecting DE genes. A) Bar plot depicting the statistical power of each method under changing number of differentially expressed genes and the mean difference in gene expression (signal-to-noise ratio) between the two phenotypes (before and after radiation). We compared our method with two transcript-level tests in paired t-test and paired wilcoxon test (p values combined at gene-level by Fisher's method), and with two gene-level tests, where the gene expression values are combined by median and Winsorized mean values followed by a paired t-test. B) Bar plot depicting the area under the curve (AUC) of all the methods under the aforementioned conditions.

Additive Effect	$PVE_\gamma(\%)$	$PVE_\phi(\%)$	Gene-set Score Test	TBT Paired t-test	TBT Wilcoxon's test	Gene-level paired t-test
NO	0	0	0.048 [0.036-0.063]	0.047 [0.03-0.07]	0.044 [0.027-0.066]	0.042 [0.027-0.061]
NO	0	7	0.546 [0.515-0.577]	0.234 [0.196-0.275]	0.198 [0.162-0.237]	0.316 [0.279-0.355]
NO	0	9	0.753 [0.725-0.779]	0.384 [0.339-0.43]	0.313 [0.271-0.358]	0.465 [0.424-0.506]
NO	7	0	0.408 [0.377-0.439]	0.202 [0.166-0.242]	0.17 [0.137-0.207]	0.12 [0.095-0.149]
NO	7	7	0.756 [0.728-0.782]	0.413 [0.367-0.46]	0.386 [0.341-0.432]	0.376 [0.337-0.416]
NO	6	9	0.859 [0.836-0.88]	0.558 [0.511-0.604]	0.515 [0.468-0.562]	0.526 [0.485-0.567]
NO	9	0	0.584 [0.553-0.615]	0.353 [0.309-0.399]	0.294 [0.253-0.338]	0.178 [0.148-0.211]
NO	9	6	0.806 [0.78-0.83]	0.546 [0.499-0.592]	0.481 [0.434-0.528]	0.415 [0.375-0.456]
NO	8	8	0.897 [0.876-0.915]	0.655 [0.61-0.699]	0.601 [0.555-0.646]	0.606 [0.565-0.646]
YES	0	0	0.716 [0.687-0.744]	0.178 [0.144-0.216]	0.167 [0.134-0.204]	0.289 [0.253-0.327]
YES	0	7	0.801 [0.775-0.825]	0.386 [0.341-0.432]	0.334 [0.291-0.379]	0.483 [0.442-0.524]
YES	0	9	0.878 [0.856-0.898]	0.542 [0.495-0.588]	0.483 [0.436-0.53]	0.651 [0.611-0.69]
YES	7	0	0.738 [0.71-0.765]	0.414 [0.368-0.461]	0.365 [0.321-0.411]	0.334 [0.296-0.374]
YES	7	7	0.876 [0.854-0.896]	0.588 [0.541-0.634]	0.538 [0.491-0.584]	0.549 [0.508-0.59]
YES	6	9	0.924 [0.906-0.94]	0.654 [0.609-0.698]	0.607 [0.561-0.652]	0.66 [0.62-0.698]
YES	9	0	0.763 [0.735-0.789]	0.478 [0.431-0.525]	0.438 [0.392-0.485]	0.349 [0.31-0.389]
YES	9	6	0.88 [0.858-0.899]	0.654 [0.609-0.698]	0.598 [0.551-0.643]	0.57 [0.529-0.61]
YES	8	8	0.944 [0.928-0.957]	0.727 [0.684-0.767]	0.682 [0.637-0.725]	0.65 [0.61-0.689]

Table 2: *DE of gene-sets - Gene-set simulation results at 5% FDR with 95% confidence interval. We varied additive effect i.e. average effect of radiation on the whole transcriptome, proportion of variation explained by γ i.e. radiation $\times$ transcripts interaction effect, and the proportion of variation explained by ϕ i.e. radiation $\times$ genes interaction effect. Our score test is referred to as "DE Score Test".*

proportion of variation explained by the interaction effect between radiation and transcript (PVE_γ) and 3) the proportion of variation explained by the interaction effect between radiation and gene (PVE_ϕ) . We kept the number of transcripts and genes constant for all these simulations. Table 2 shows that our method significantly outperforms both transcript-level and gene-level methods. More specifically, our method captures the transcript-specific variability due to radiation within each gene more efficiently than the other tests.

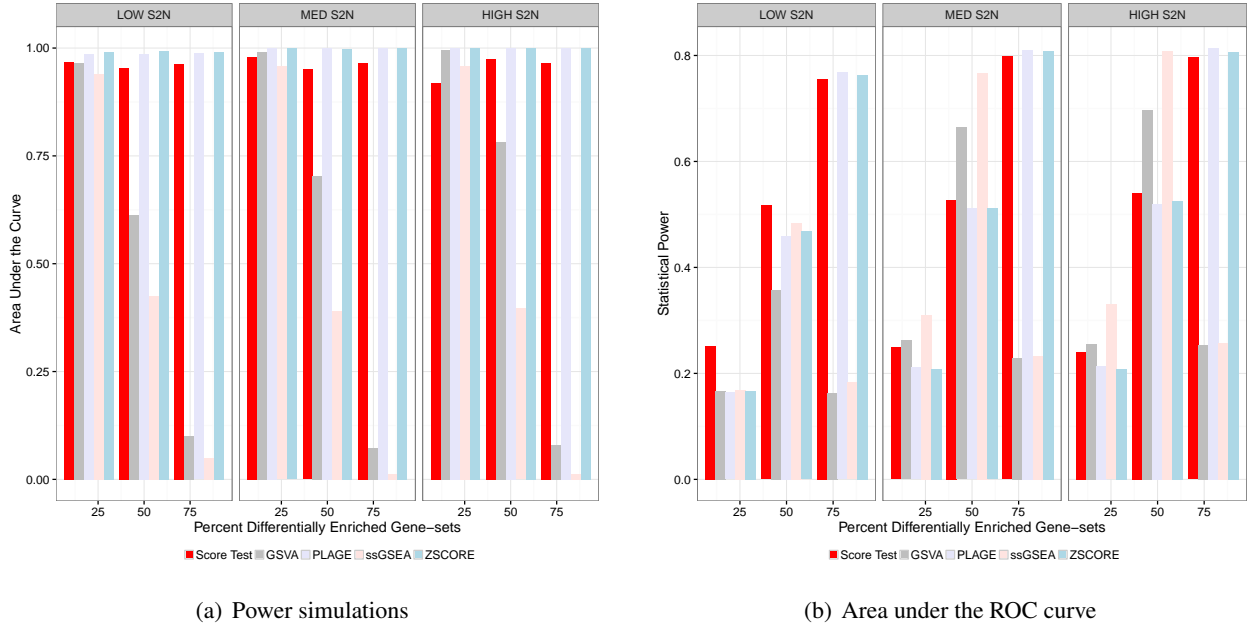(a) Power simulations

(b) Area under the ROC curve

Figure 2: The performance of all the methods in detecting differentially enriched gene-sets when each gene-set is comprised of unique set of genes. A) Bar plot depicting the statistical power of each method under changing number of differentially enriched gene-sets and the mean difference in gene expression (signal-to-noise ratio) between the two phenotypes (before and after radiation). We compared our method with several gene-level tests,by computing the median gene expression values across all the transcripts within a gene. B) Bar plot depicting the area under the curve (AUC) of all the methods under the aforementioned conditions

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. See supplementary methods for more details. In case on no gene overlap, we simulated 10 gene-sets with varying degrees of gene overlap

(20%, 50% and 80%), and varying mean differences between the two phenotypes (low, medium and high). We compared the performance of all the methods by measuring statistical power and area under the curve in case of gene-sets with no overlapping genes. In the case where gene-sets shared genes, we measured only statistical power.
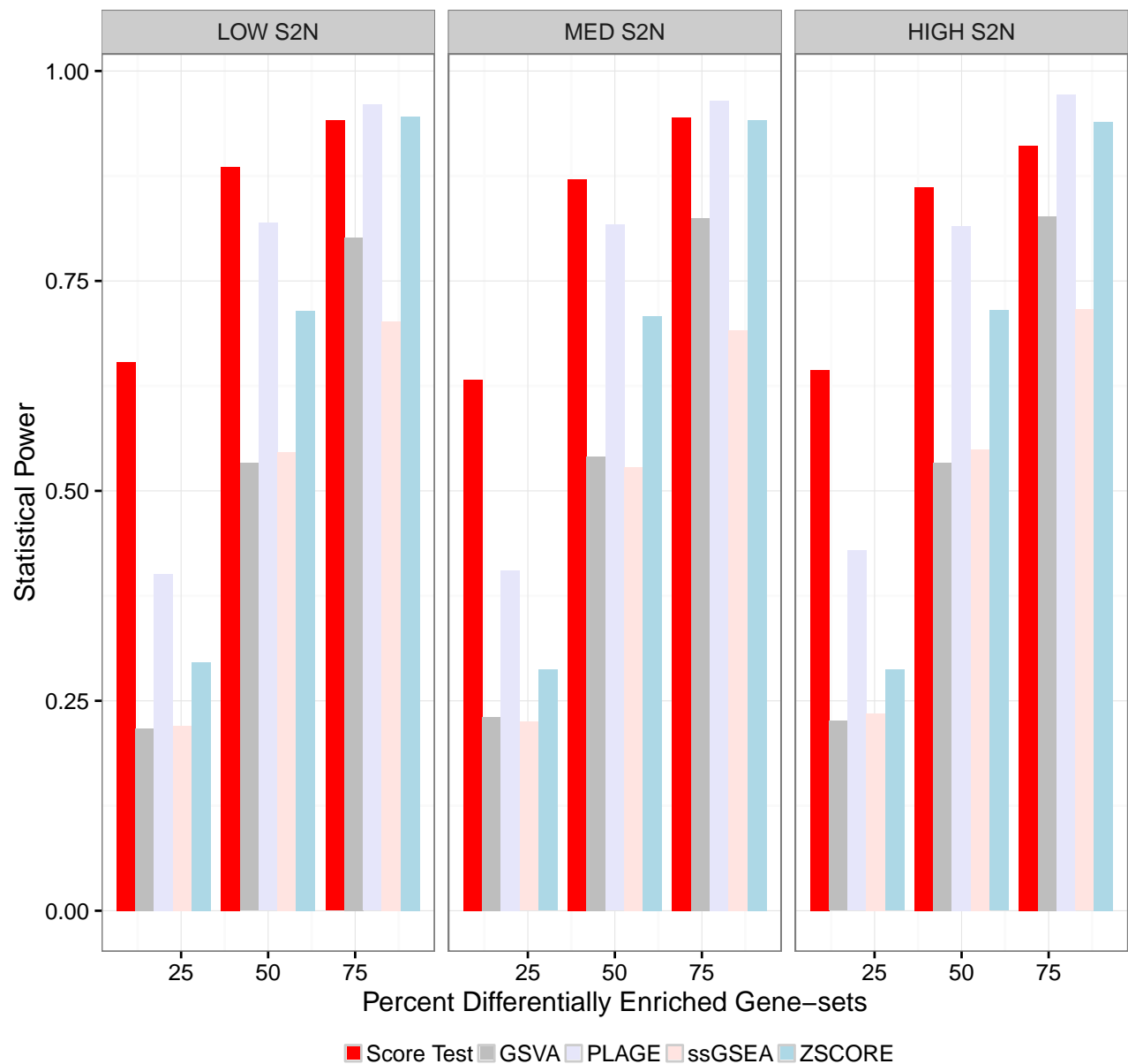

Figure 3: The performance of all the methods in detecting differentially enriched gene-sets when each gene-set is comprised of shared genes. Bar plot depicting the statistical power of each method under changing number of differentially enriched gene-sets and the mean difference in gene expression (signal-to-noise ratio) between the two phenotypes (before and after radiation). We compared our method with several gene-level tests, by computing the median gene expression values across all the transcripts within a gene.

Figues 2a and 2b show the performance of all the methods when the gene-sets do not share any genes. Even though, this is not a general scenario, our method is competitive with the rest of the methods. In situations where the power of our method is low (relative to the other methods), the accuracy of our method is high given the AUC values. Figure 3 displays the performance of all the methods when the gene-sets have over-lapping genes or shared genes. This is the most common scenario and our method performs well, in terms of statistical power, in almost all cases.

Transcriptome-wide response to radiotherapy in breast tumors

Isoform-specific DE analysis

Transcriptome expression profiling of the early-stage breast cancer patients before and after preoperative radiotherapy using our method has revealed many DE genes. Current methods perform DE analysis at the gene-level and not at the transcript-level. One method performs a standard paired t-test at the transcript-level and combines the p values using Fisher's method [26, 27], which tests whether the combined p values are jointly significant. However, Fisher's method assumes that the transcript-level p values for each gene are independent. This method identified 11,944 genes at 5% FDR. Another most commonly used approach is to combine the gene expression values of all transcripts of a gene *a priori* by computing either the median expression values or Winsorized mean expression values (to avoid any outliers). Paired t-tests were then run on the combined data. These two ways of combining the data identified 4,729 and 3,353 genes, respectively at 5% FDR. Our method identified a total of 12,414 DE genes at 5% FDR, which is more than the ones identified by the aforementioned methods. To assess the biological relevance of the DE genes, we performed a KEGG pathway term enrichment analysis [28] for each set of results separately. KEGG pathways were considered overrepresented if a set of at least three genes from different linked regions is observed to be overrepresented with an adjusted significance level of an adjusted p value < 0.05, calculated from a hypergeometric test [29].

Table 4 presents a list of top 10 signaling pathways that were shown be overrepresented in the dataset without any specifics on the directionality (up- or down-regulation) of the pathway deregulation. For example, PI3K-AKT signaling pathway shown in the table, a potential target for radiosensitizing cancer cells, is one the many pro-survival signaling pathways that get activated by radiation that may lead to suppression of apoptosis, initiation of DNA repair mechanisms and induction of cell-cycle arrest [30]. Together with mTOR signaling pathway, PI3K-AKT are activated in many different cancers. Drugs like rapamycin, CCI-779 and RAD-001 target mTOR signaling pathway while perofisine, PX-866 target AKT pathway [31].

KEGG ID	Description	p values	Adjusted p values	q value
hsa05200	Pathways in cancer	4.56E-09	1.34E-06	7.68E-07
hsa04151	PI3K-AKT signaling pathway	3.69E-08	5.45E-06	3.11E-06
hsa01100	Metabolic pathways	1.64E-07	1.61E-05	9.20E-06
hsa04060	Cytokine-cytokine receptor interaction	1.21E-06	8.91E-05	5.09E-05
hsa04510	Focal adhesion	1.59E-06	9.38E-05	5.36E-05
hsa04630	JAK-STAT signaling pathway	3.35E-06	0.000164626	9.40E-05
hsa04144	Endocytosis	9.88E-06	0.000386927	0.000220904
hsa05166	HTLV-I infection	1.13E-05	0.000386927	0.000220904
hsa04360	Axon guidance	1.24E-05	0.000386927	0.000220904
hsa04210	Apoptosis	1.43E-05	0.000386927	0.000220904

Table 3: A list of top 10 over-represented KEGG pathways based on the functional enrichment of our DE gene list.

Isoform-specific gene-set analysis

Instead of focusing on individual genes, we turned our focus on functionally related genes referred to as gene-sets or pathways and assess their behavior before and after treatment with radiation. Gene-set enrichment analysis (GSEA) and other similar methods such as Gene Set Analysis (GSA) make use on the entire gene expression profile before they assess changes of small magnitude in functionally related genes. The aforementioned methods are supervised, which require an *a priori* knowledge on the experimental design. However, methods such as single sample GSEA, GSVA, PLAGE, and ZSCORE are unsupervised and focus on the relative enrichment of pathways across the sample space rather than the absolute enrichment with respect to a phenotype [10]. All of these methods work at a gene-level and require us to combine gene expression values at the transcript level before any analysis. Our method identified differentially expressed gene-sets by leveraging transcript-specific effects without having to aggregate gene expression over all the probes of a gene. On this basis, we interrogated critical radiation-associated oncogenic signaling pathways and tumor microenvironment signatures and compared the performance of our method with the rest of the methods. Many of the radiation-associated oncogenic signaling pathways were obtained from the "halmark" gene-set collection of the Molecular Signatures Database (MSigDB), which were generated by a hybrid approach that combines an automated computational procedure with manual expert curation [32]. All of the investigated 24 oncogenic signaling pathways and 7 tumor microenvironment gene signatures were found to be statistically significant at 5% FDR by our method.

All other methods were applied at the gene-level i.e. aggregated gene expression values over all isoforms using median expression values. GSVA identified 22 gene-sets (70.9%), PLAGE identified 26 gene-sets (83.8%), ssGSEA identified 25 gene-sets (80.6%), and ZSCORE identified 27 gene-sets (87%) at 5% FDR. In order to

Gene-set name	Gene-set type	Score Test	GSVA	PLAGE	ssGSEA	ZSCORE
NFKB	Oncogenic Signaling Pathway	2.00E-16	1.42E-09	1.89E-06	2.99E-09	2.28E-09
WNT/BETA CATENIN	Oncogenic Signaling Pathway	2.00E-16	1.29E-02	1.32E-02	1.01E-03	3.12E-04
TGF BETA	Oncogenic Signaling Pathway	2.00E-16	1.92E-02	1.95E-02	9.77E-03	5.55E-03
JAK/STAT	Oncogenic Signaling Pathway	2.00E-16	2.47E-08	1.61E-04	5.53E-07	2.60E-08
DNA REPAIR	Oncogenic Signaling Pathway	2.00E-16	7.04E-02	2.72E-02	6.95E-03	1.28E-02
G2M CHECKPOINT	Oncogenic Signaling Pathway	2.00E-16	7.32E-01	4.94E-02	4.63E-01	9.63E-01
APOPTOSIS	Oncogenic Signaling Pathway	2.00E-16	2.65E-04	2.49E-03	5.70E-05	2.06E-05
NOTCH SIGNALING	Oncogenic Signaling Pathway	2.00E-16	9.46E-02	2.98E-02	1.98E-02	1.89E-02
ADIPOGENESIS	Oncogenic Signaling Pathway	2.00E-16	4.82E-02	3.71E-02	1.97E-02	1.66E-02
INTERFERON ALPHA	Oncogenic Signaling Pathway	2.00E-16	6.37E-03	1.32E-02	1.23E-02	4.01E-03
INTERFERON GAMMA	Oncogenic Signaling Pathway	2.00E-16	3.37E-05	4.45E-04	3.03E-05	5.36E-06
HEDGEHOG SIGNALING	Oncogenic Signaling Pathway	2.00E-16	3.54E-02	2.72E-02	7.42E-02	9.08E-03
PI3K/AKT/MTOR	Oncogenic Signaling Pathway	2.00E-16	1.24E-02	1.95E-02	6.16E-03	1.71E-03
MTORC1	Oncogenic Signaling Pathway	2.00E-16	5.68E-02	2.72E-02	1.24E-02	1.05E-02
E2F	Oncogenic Signaling Pathway	1.62E-11	4.95E-01	1.22E-01	5.12E-01	5.31E-01
EMT	Oncogenic Signaling Pathway	2.00E-16	3.93E-05	1.61E-04	2.86E-06	9.82E-06
INFLAMMATORY RESPONSE	Oncogenic Signaling Pathway	2.00E-16	1.10E-10	1.61E-04	2.99E-09	1.41E-11
FATTY ACID METABOLISM	Oncogenic Signaling Pathway	2.00E-16	2.70E-01	1.33E-01	1.34E-01	9.63E-02
OXIDATIVE PHOSPHORYLATION	Oncogenic Signaling Pathway	2.00E-16	1.50E-01	6.49E-02	7.42E-02	5.40E-02
GLYCOLYSIS	Oncogenic Signaling Pathway	2.00E-16	7.32E-02	4.17E-02	1.98E-02	1.53E-02
ROS	Oncogenic Signaling Pathway	2.00E-16	2.60E-03	2.98E-02	1.04E-02	1.71E-03
P53	Oncogenic Signaling Pathway	2.00E-16	2.77E-06	1.61E-04	1.11E-08	6.89E-07
IL2/STAT5	Oncogenic Signaling Pathway	2.00E-16	7.98E-06	4.45E-04	1.83E-06	6.52E-07
KRAS	Oncogenic Signaling Pathway	2.00E-16	2.70E-08	4.45E-04	5.53E-07	8.42E-10
CHROMOSOMAL INSTABILITY	Tumor Microenvironment	2.00E-16	9.38E-04	2.16E-02	2.11E-02	3.12E-04
HYPOXIA	Tumor Microenvironment	2.00E-16	2.37E-04	4.45E-04	2.86E-06	2.06E-05
INVASIVENESS POTENTIAL	Tumor Microenvironment	2.00E-16	7.56E-02	5.00E-02	3.22E-02	2.17E-02
EPIGENETIC STEM CELL	Tumor Microenvironment	2.00E-16	1.80E-03	5.56E-02	8.83E-02	1.05E-02
TNF	Tumor Microenvironment	2.00E-16	5.12E-03	1.32E-02	9.77E-03	7.00E-04
ANGIOGENESIS	Tumor Microenvironment	2.00E-16	4.92E-02	1.95E-02	1.24E-02	6.37E-03
IMMUNE RESPONSE	Tumor Microenvironment	2.00E-16	1.69E-04	2.00E-01	2.43E-04	2.42E-04

Table 4: A comparison of gene-set analysis of oncogenic signaling pathways and tumor microenvironment using our score test method, GSVA, PLAGE, ssGSEA and ZSCORE methods. All the p values are adjusted for multiple comparisons using Benjamini-Hochberg method.

visualize the patterns of pathway regulation, we obtained a matrix containing sample set enrichment scores of all the 31 gene-sets over all the samples using the popular GSA method. The enrichment data clearly clusters into two phenotypic classes, before and after radiation. From the heatmap, radiation induces a hypoxic state, enhances tumor necrosis factors and suppresses angiogenesis. Radiation-induced inflammatory pathways and immune response signatures can be targeted by therapeutics that improve the clinical outcome of radiotherapy by enhancing the radiosensitivity and decreasing any putative metabolic effects.

Discussion

Tailoring a patient's treatment to exploit an individual's tumor biology remains an elusive goal in cancer therapy. Similar to cytotoxic therapy, response to radiation in a given population of 'eligible' patients is markedly
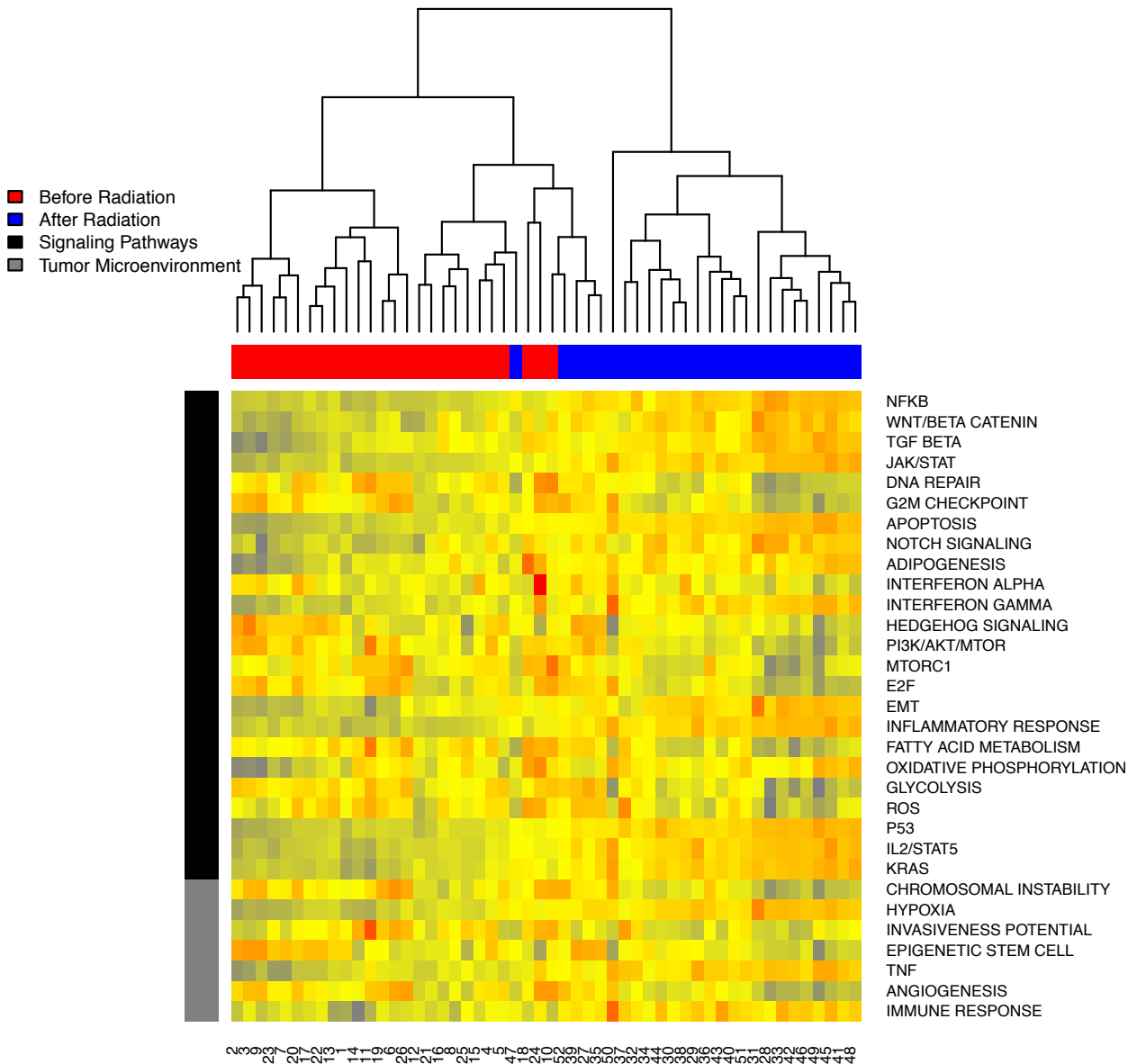
Figure 4: Heatmap showing differentially enriched oncogenic signaling pathways and signatures of tumor microenvironment between patients before and after receiving radiotherapy. The matrix containing sample set enrichment score as computed by the GSA software were clustered by hierarchical clustering method to generate this heatmap. Red indicates a higher collective expression and blue indicates a lower collective expression of genes in that gene-set.

heterogeneous. While chemotherapy serves to address systemic disease, radiation acts as effective local therapy. In many instances, patients resistant to radiation have limited to no options to control local disease [33, 34, 35]; thus, prospectively determining tumor radiosensitivity is important to identify cohorts of patients most likely to respond and to minimize the incidence of radiation-related adverse events in patients who might not otherwise respond. Also, if the molecular underpinnings of radiation response can be elucidated and exploited, the radioresistance of tumors could potentially be abrogated with novel therapeutics. While many mechanisms of

radiation resistance, including alterations in DNA repair mechanisms [36], upregulation of pathways regulating angiogenesis [37], apoptosis [38] and cell cycle [39], have been previously described, a comprehensive evaluation of biological events to identify key oncogenic signaling events regulating radiation response, at a genomic and transcriptomic level, is largely unknown. Recent technological advances in quantifying gene expression (i.e. high-throughput sequencing assays) will allow us to interrogate whole exomes or transcriptomes with a higher precision than mRNA expression microarrays thus, overcoming the limitations in detecting and quantifying coding transcript isoforms. However, current statistical methods allow us to interrogate genes and gene-sets at the gene-level by aggregating gene expression across all possible gene isoforms thus, not taking advantage of alternative splicing mechanisms that result in multiple isoforms of the same gene.

Overall, our efforts are primarily directed to understanding two very specific aspects - 1) the effect of radiation-induced gene isoform-level variability on gene expression, oncogenic signaling pathways involved in radiation response and tumor microenvironment, and 2) the overall effect of radiation on gene expression. Currently, there are no established methods that leverage gene isoform-specific effects in order to quantify gene expression and investigate tumor biology at a higher resolution. Our methods provide an efficient framework to model transcript-specific and gene-specific effects to map biomarkers association with radiation response. The dataset used here used a high-resolution array-based platform that includes an overwhelming number of gene transcripts in the human transcriptome with >6 million probes targeting coding transcripts, exon-exon splice junctions, and non-coding transcripts. However, we predict that our method will be applicable to gene expression data quantified using RNA-Seq analysis since we make distributional assumptions that preclude their direct application to RNA-Seq count data.

Finally, our methods and analyses are helpful in only generating biological or clinical hypotheses that require substantial verification using *in vitro* and *in vivo* models. Eventually, by correctly interpreting these data, we enhance our ability to accurately identify individuals most likely to be resistant to radiation therapy based on the patterns of pathway activation, which further emphasizes the need to identify novel compounds/drugs that could modulate radiation response and function as radiosensitizers.

References

[1] Horton J, Siamakpour-Reihani S, Lee C, Zhou Y, Chen W, Geradts J, Fels D, Hoang P, Ashcraft K, Groth J, Kung H, Dewhirst M, Chi JT: **FAS Death Receptor: A Breast Cancer Subtype-Specific Radiation Response Biomarker and Potential Therapeutic Target**. *Radiation Research* 2015, **184**(5):456–69.

[2] Langlands FE, Horgan K, Dodwell DD, Smith L: **Breast cancer subtypes: response to radiotherapy**

and potential radiosensitisation. *British Journal of Radiology* 2013, **86**(1023):20120601.

[3] Macaeva E, Saeys Y, Tabury K, Janssen A, Michaux A, Benotmane MA, De Vos WH, Baatout S, Quintens R: **Radiation-induced alternative transcription and splicing events and their applicability to practical biodosimetry**. *Scientific Reports* 2016, **6**:19251 EP –, [http://dx.doi.org/10.1038/srep19251].

[4] Wushou A, Jiang YZ, Hou J, Liu YR, Guo XM, Shao ZM: **Development of triple-negative breast cancer radiosensitive gene signature and validation based on transcriptome analysis**. *Breast Cancer Research and Treatment* 2015, **154**:57–62, [http://dx.doi.org/10.1007/s10549-015-3611-0].

[5] Affymetrix: **Affymetrix, GeneChip Human Transcriptome Array 2.0 Data Sheet.** [http://www.affymetrix.com/support/technical/datasheets/hta_array_2_0_datasheet.pdf].

[6] Satterthwaite F: **An approximate distribution of estimates of variance components**. *Biometrics Bulletin* 1946, **2**(6):110–114.

[7] Bates D, Maechler M, Bolker B, Walker S: *lme4: Linear mixed-effects models using Eigen and S4* 2014, [http://CRAN.R-project.org/package=lme4]. [R package version 1.1-7].

[8] Wilcoxon F: **Individual Comparisons by Ranking Methods**. *Biometrics Bulletin* 1945, **1**(6):80–83.

[9] Wilcox R, Keselman H: **Modern robust data analysis methods: Measures of central tendency**. *Psychological Methods* 2003, **8**(3):254–274.

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

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

[12] Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, Schinzel AC, Sandy P, Meylan E, Scholl C, Frohling S, Chan EM, Sos ML, Michel K, Mermel C, Silver SJ, Weir BA, Reiling JH, Sheng Q, Gupta PB, Wadlow RC, Le H, Hoersch S, Wittner BS, Ramaswamy S, Livingston DM, Sabatini DM, Meyerson M, Thomas RK, Lander ES, Mesirov JP, Root DE, Gilliland DG, Jacks T, Hahn WC: **Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1**. *Nature* 2009, **462**(7269):108–112, [http://dx.doi.org/10.1038/nature08460].

[13] E L, H-Y C, J-W K, T I, D L: **Inferring Pathway Activity toward Precise Disease Classification**. *PLoS Computational Biology* 2008, **4**(11):e1000217. doi:10.1371/journal.pcbi.1000217.

[14] Chi J, Wang Z, Nuyten D, Rodriguez E, Schaner M, Salim A, Wang Y, Kristensen G, Helland A, Børresen-Dale A, Giaccia A, Longaker M, Hastie T, Yang Y, van de Vijver M, Brown P: **Gene expression programs in response to hypoxia; cell type specificity and prognostic significance in human cancers**. *PLoS Medicine* 2006, **3**(3):e47. doi:10.1371/journal.pmed.0030047.

[15] Liu R, Wang X, Chen G, Dalerba P, Gurney A, Hoey T, Sherlock G, Lewicki J, Shedden K, Clarke M: **The prognostic role of a gene signature from tumorigenic breast-cancer cells. The prognostic role of a gene signature from tumorigenic breast-cancer cells**. *N Engl J Med* 2007, **356**(3):217–226.

[16] Widschwendter M, Fiegl H, Egle D, Mueller-Holzner E, Spizzo G, Marth C, Weisenberger DJ, Campan M, Young J, Jacobs I, Laird PW: **Epigenetic stem cell signature in cancer**. *Nat Genet* 2007, **39**(2):157–158, [http://dx.doi.org/10.1038/ng1941].

[17] Viemann D, Goebeler M, Schmid S, Nordhues U, Klimmek K, Sorg C, Roth J: **TNF induces distinct gene expression programs in microvascular and macrovascular human endothelial cells.** *Journal of Leukocyte Biology* 2006, **80**:174–185.

[18] Chang H, Sneddon J, Alizadeh A, Sood R, West R, Montgomery K, Chi JT, van de Rijn M, Botstein D, Brown P: **Gene Expression Signature of Fibroblast Serum Response Predicts Human Cancer Progression: Similarities between Tumors and Wounds.** *PLoS Biology* 2004, **2**(2):e7. doi:10.1371/journal.pbio.0020007.

[19] Hsu D, Kim B MKand Balakumaran, Acharya C, Anders C, Clay T, Lyerly H, Drake C, Morse M, Febbo P: **Immune Signatures Predict Prognosis in Localized Cancer**. *Cancer Investigation* 2010, **28**(7):765–773.

[20] Carter S, Eklund A, Kohane I, Harris L, Z S: **A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers.** *Nature Genetics* 2006, **38**(9):1043–1048.

[21] Efron B, Tibshirani R: *GSA: Gene set analysis* 2010, [https://CRAN.R-project.org/package=GSA]. [R package version 1.03].

[22] Efron B, Tibshirani R: **On testing the significance of sets of genes**. Tech. rep., Stanford University 2006.

[23] Eisen M, Spellman P, Brown P, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *PNAS* 1998, **95**(25):14863–8.

[24] Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995, **57**:289–300.

[25] Sing T, Sander O, Beerenwinkel N, T L: **ROCR: visualizing classifier performance in R.** *Bioinformatics* 2005, **21**(20):3940–1.

[26] Fisher R: *Statistical Methods for Research Workers*. Oliver and Boyd, fifth edition 1934.

[27] Birnbaum A: **Combining Independent Tests of Significance**. *Journal of American Statistical Association* 1959, **59**(267):559–574.

[28] Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M: **KEGG as a reference resource for gene and protein annotation**. *Nucleic Acids Research* 2016, **44**(D1):D457–62.

[29] Yu G, Wang LG, Han Y, He QY: **ClusterProfiler: An r package for comparing biological themes among gene clusters.** *OMICS: A Journal of Integrative Biology* 2012, **16**:284–287.

[30] Hein A, Ouellette M, Yan Y: **Radiation-induced signaling pathways that promote cancer cell survival (Review)**. *International Journal of Oncology* 2014, **45**(5):1813–1819.

[31] LoPiccolo J, Blumenthal G, Bernstein W, Dennis P: **Targeting the PI3K/Akt/mTOR pathway: effective combinations and clinical considerations**. *Drug Resistance Update* 2008, **11**(1-2):32–50.

[32] Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov J, Tamayo P: **The Molecular Signatures Database Hallmark Gene Set Collection**. *Cell Systems* 2015, **1**(6):417–425.

[33] Torres-Roca J: **A molecular assay of tumor radiosensitivity: a roadmap towards biology-based personalized radiation therapy**. *Per Med.* 2012, **9**(5):547–557.

[34] Torres-Roca J, Eschrich S, Zhao H, Bloom G, Sung J, McCarthy S, Cantor A, Scuto A, Li C, Zhang S, Jove R, Yeatman T: **Prediction of radiation sensitivity using a gene expression classifier**. *Cancer Research* 2005, **65**(16):7169–76.

[35] Baumann M, Krause M, Overgaard J, Debus J, Bentzen S, Daartz J, Richter C, Zips T D andBortfeld: **Radiation oncology in the era of precision medicine**. *Nature Reviews Cancer* 2016, **16**(4):234–49.

[36] Willers H, Dahm-Daphi J, Powell S: **Repair of radiation damage to DNA**. *British Journal of Cancer* 2004, **90**:1297–1301.

[37] Li J, Huang S, Armstrong E, Fowler J, Harari P: **Angiogenesis and radiation response modulation after vascular endothelial growth factor receptor-2 (VEGFR2) blockade**. *International Journal of Radiation Oncology*Biology*Physics* 2005, **62**(5):1477–1485.

[38] Balcer-Kubiczek E: **Apoptosis in radiation therapy: a double-edged sword**. *Experimental Oncology* 2012, **34**(3):277–285.

[39] Bernhard E, Maity A, Muschel R, McKenna W: **Effects of ionizing radiation on cell cycle progression. A review.** *Radiation and environmental biophysics* 1995, **34**(2):79–83.