ssPATHS: Single Sample PATHway Score (version 0.1.0)

Natalie R. Davidson, Philipp Markolin, Gunnar Rätsch
Biomedical Informatics, ETH Zürich

2019
July

1 Introduction

Precision oncology requires that a single patient can be accurately and meaningfully characterized in order to tailor treatments. Using our method, ssPATHS, we are able to estimate pathway deviations for a single patient, by first learning a discriminative gene weighting from a reference cohort. In this vignette, we use TCGA gene expression data to learn a weighting on hypoxia-related genes then apply it to PDAC cancer cell lines to estimate the level of hypoxia within each sample.

2 Formatting Data

First, we load the appropriate packages we will need.

```
> library("ROCR")
> library("ggplot2")
> library("ssPATHS")
```

Next, we will read in the TCGA data and format it appropriately.

```
> data(tcga_expr_df)
```

Let's look at the format of our data. To learn our gene weights we will need a Y column and a *sample_id* column.

```
> tcga_expr_df[1:6,1:5]
                        tcga_id study is_normal
1 TCGA-CQ-6224-01A-11R-1915-07  HNSC     FALSE
2 TCGA-TQ-A7RP-01A-21R-A34F-07   LGG     FALSE
3 TCGA-13-1510-01A-02R-1565-13    OV     FALSE
4 TCGA-HC-8265-01A-11R-2263-07  PRAD     FALSE
```

```
5 TCGA-HC-7079-01A-11R-1965-07  PRAD     FALSE
6 TCGA-4X-A9FC-01A-11R-A42C-07  THYM     FALSE
  libsize_75percent ENSG00000078369
1         0.5201350       146190.89
2         0.4887488       105250.39
3         0.5159627        58773.63
4         0.3658994       101336.60
5         0.4617693        81393.47
6         0.3735174       101473.18
```

Now we will need to transform the data.frame into a SummarizedExperiment object.

```
> tcga_se <- SummarizedExperiment(t(tcga_expr_df[ , -(1:4)]),
                                  colData=tcga_expr_df[ , 2:4])
> colnames(tcga_se) <- tcga_expr_df$tcga_id
> colData(tcga_se)$sample_id <- tcga_expr_df$tcga_id
>
```

Since we are interested in hypoxia, we want to learn a weighting only on genes associated with hypoxia. In this package we have a helper function to retreive these genes, but other gene sets can be used for different pathways of interest. Gene sets can be easily fetched using the package *msigdbr*.

```
> hypoxia_gene_ids <- get_hypoxia_genes()
> hypoxia_gene_ids <- intersect(hypoxia_gene_ids, rownames(tcga_se))
```

Now we will need to identify how we want to discriminate our samples. Here, we use the assumption that normal samples are less hypoxic than tumor samples. Therefore, we will use the *is_normal* column as our Y column. We set normal samples to 0 and tumor samples to 1. This implies that a higher score indicates a more hypoxic sample.

```
> colData(tcga_se)$Y <- ifelse(colData(tcga_se)$is_normal, 0, 1)
>
```

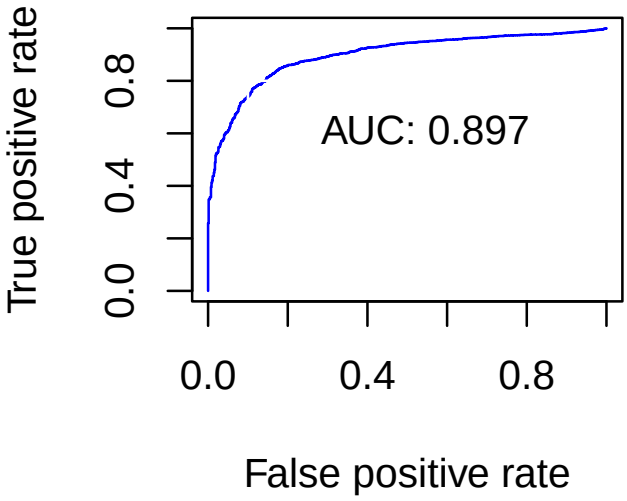
3 Get Reference Gene Weightings

Now that our data is in the appropriate format, we can learn the weightings. Within the method *get_gene_weights*, there is a normalization step where we scale across all the genes available in the SummarizedExperiment assay. For this to be stable and consistent, we recommend that the assay contain at least 500 genes that are consistently expressed across all samples in addition to the genes in the pathway of interest. We also assume the most of the genes in our hypoxia geneset are increasing, meaning that our genes move unidirectionally.

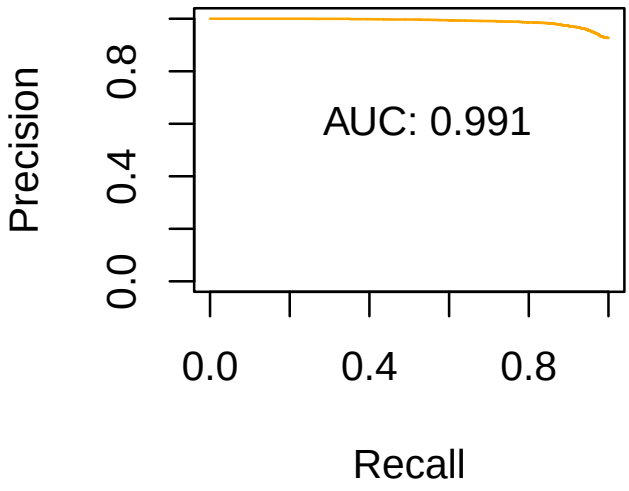
```
> res <- get_gene_weights(tcga_se, hypoxia_gene_ids, unidirectional=TRUE)
> gene_weights <- res[[1]]
> sample_scores <- res[[2]]
```

Now let's see how well we did in seperating the two classes defined by Y.

```
> training_res <- get_classification_accuracy(sample_scores, positive_val=1)
> # plot the ROC curve
> plot(training_res[[4]], col="blue", ylim=c(0, 1))
> roc_text <- paste("AUC:", round(training_res$auc_roc,3))
> legend(0.1,0.8, roc_text,
        border="white",cex=1,box.col = "white")
```




```
> # plot the PR curve
> plot(training_res[[3]], col="orange", ylim=c(0, 1))
> pr_text <- paste("AUC:", round(training_res$auc_pr,3))
> legend(0.1,0.8, pr_text,
        border="white",cex=1,box.col = "white")
```



4 Apply Gene Weightings to New Samples

Now, using the gene weightings learned from the reference set, we can apply it to a new sample.

```
> data(new_samp_df)
> new_samp_se <- SummarizedExperiment(t(new_samp_df[ , -(1)]),
                                      colData=new_samp_df[ , 1, drop=FALSE])
> colnames(colData(new_samp_se)) <- "sample_id"
> new_score_df <- get_new_samp_score(gene_weights, new_samp_se)
> new_score_df
DataFrame with 12 rows and 2 columns
           sample_id pathway_score
         <character>     <numeric>
1    exp_norm_ctrl_C     -0.344500
2    exp_norm_ctrl_A     -0.335998
3    exp_norm_ctrl_B     -0.282953
4    exp_hyp_noHIF_A     -0.256307
5   exp_norm_noHIF_C     -0.225995
...              ...           ...
8    exp_hyp_noHIF_C    -0.0822687
```

```
9    exp_norm_noHIF_B    -0.0717889
10     exp_hyp_ctrl_C     0.1123652
11     exp_hyp_ctrl_A     0.1177899
12     exp_hyp_ctrl_B     0.2467603
```

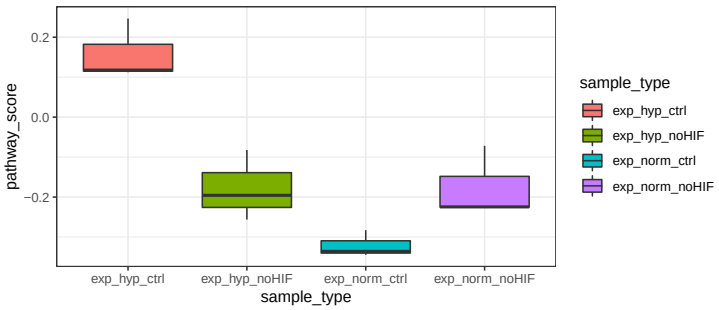
Now lets see if the derived score match our experimental expectations. Samples with **hyp** or **norm** in the sample id are cell lines that were exposed to hypoxic or normoxic conditions respectively. Samples with **ctrl** or **noHIF** were samples that were able to produce a HIF-mediated hypoxic response or not, respectively.

```
> plot_scores <- function(hif_scores){

    # format the sample IDS
    hif_scores$sample_type <- substr(hif_scores$sample_id, 1,
                                     nchar((hif_scores$sample_id))-2)
    colnames(hif_scores)[2] <- "pathway_score"

    gg <- ggplot(hif_scores, aes(x=sample_type, y=pathway_score,
                                 fill=sample_type)) +
        geom_boxplot() +
        theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
        theme_bw()

        return(gg)
 }
> gg <- plot_scores(as.data.frame(new_score_df))
> print(gg)
```



Accordingly, we find the **hyp_ctrl** samples have the highest pathway score. According to our labeling (tumor/hypoxic is 1 and normal/normoxic is 0), this implies that these samples are the most hypoxic. Furthermore, we see that the samples that were not able to produce a hypoxic response, even in the absence of oxygen (**hyp_noHIF**) are found to have similar score to the normoxic samples.