An overview of the TIN package

Bjarne Johannessen

April 27, 2020

Contents

1 Introduction

This document gives an overview and demonstration of the `TIN` package. The package provides a set of tools for transcriptome instability analysis based on exon-level microarray expression profiles. Alternative splicing is an important mechanism for gene expression, and disruption from normal splicing patterns can be harmful for eukaryotic cells. By applying high-throughput technologies, it is possible to identify genes and exons that are subject to splicing discrepancies. The `TIN` package includes a set of tools for aberrant exon usage calculations, and for analyzing correlation between transcriptome instability and splicing factor expression.

1.1 Data input

Input data to the `TIN` package is raw expression data (CEL files) and preprocessed gene-level expression values.

1.2 Data sets

The package includes three data sets:

- `splicingFactors`: A list of 280 splicing factor genes [1].
- `geneSets`: 1,454 Gene Ontology gene sets [2].
- `geneAnnotation`: Matching gene symbols and Affymetrix transcript cluster identifiers.

In addition, two toy data sets are included in the package. See the worked example below for a demonstration.

2 Example

The following example illustrates the analysis pipeline of the `TIN` package. First, load the package in R:

```
> library(TIN)
```

2.1 Data access

We will need access to all three data sets included in the package:

```
> data(splicingFactors)
> data(geneSets)
> data(geneAnnotation)
```

The splicingFactors data set contains gene symbols and Affymetrix transcript cluster identifiers for 280 genes known to be involved in splicing. The geneSets data set is a collection of 1,454 Gene Ontology gene sets included in the package to enable comparisons of associations between aberrant exon usage and expression levels with other general gene sets. The list comprises one major collection of gene sets in the Molecular Signatures Database, MSigDB [2]. The geneAnnotation data set is a list of matching gene symbols and Affymetrix transcript cluster identifiers.

2.2 Sample data

Two sample data sets are included for educational purposes.

```
> data(sampleSetFirmaScores)
> data(sampleSetGeneSummaries)
```

The two sample datasets include small parts of a comprehensive prostate cancer data set (GEO accession number GSE21034) published in [3].

2.3 FIRMA analysis

By issuing the first command,

```
> fs <- firmaAnalysis(useToyData=TRUE)
```

raw CEL files are being read, and background correction, normalization (customized RMA approach), and alternative splicing analysis is performed according to the FIRMA method (http://www.aroma-project.org/vignettes/FIRMA-HumanExonArrayAnalysis). Local path to the aroma.affymetrix root directory and the name of the sample set is sent as parameters to the function. The function returns a data.frame with log2 FIRMA (alternative splicing) scores for each probeset/sample combination.

Next we read preprocessed gene-level expression values by

```
> gs <- readGeneSummaries(useToyData=TRUE)
```

The input parameter should be a table tab-separated file with one row for each gene and one column for each sample. Affymetrix transcript cluster identifiers should be used as row names, whereas sample names for each sample should be used as column names. These values can be generated by using for instance Affymetrix Power Tools or Expression Console.

After reading input data, aberrant exon usage can be calculated by

```
> tra <- aberrantExonUsage(1.0, sampleSetFirmaScores)
```

This function makes use of the data.frame from 'firmaAnalysis' (containing log2 FIRMA scores for all probe sets/exons (rows) in all samples (columns)), and a number (default 1.0) indicating which top percentile value of the global FIRMA scores to be used as threshold for denoting aberrant exon usage. The tra object is a list containing one number for each sample, indicating to what degree each sample possess aberrant exon usage. An object called aberrantExons, consisting of two lists representing how many probe sets for each sample that are nominated as having high or low aberrant exon usage, is also created by calling the aberrantExonUsage function. In addition, the two expression values that are used as threshold for detecting aberrant exon usage are stored in the quantiles object.

Next, we create permutations of the FIRMA scores for each probe set/exon across all samples,

```
> aberrantExonsPerms <- probesetPermutations(sampleSetFirmaScores, quantiles)
```

The perms object contains lists indicating high and low aberrant exon usage for each sample after the data in the initial firmaScores object has been reshuffled at each probe set. To calculate the correlation between sample-wise amounts of aberrant exon usage and splicing factor expression levels, the correlation function is applied in the following way

```
> corr <- correlation(splicingFactors, sampleSetGeneSummaries, tra)
```

Correlation between aberrant exon usage and expression levels for a number of gene sets is calculated by

```
> gsc <- geneSetCorrelation(geneSets, geneAnnotation, sampleSetGeneSummaries,
+     tra, 100)
```

The function calculates Pearson correlation between sample-wise aberrant exon usage amounts and expression levels of all genes for all gene sets defined by the input parameter list geneSets.

2.4 Data plotting

Four different plotting methods are included in the `TIN` package. First, the cluster plot visualizes the hierarchical clustering of the samples based on splicing factor expression levels.

```
> clusterPlot(sampleSetGeneSummaries, tra, "euclidean", "complete",
+     "TIN-cluster.pdf")

Plot was saved in  /tmp/RtmpBmnmtA/Rbuild496651bfaf82/TIN/vignettes/TIN-cluster.pdf
pdf
  2
```

Second, a scatter plot visualizes the relative amounts of aberrant exon usage for each sample

```
> scatterPlot("TIN-scatter.pdf", TRUE, aberrantExons, aberrantExonsPerms)

Plot was saved in  /tmp/RtmpBmnmtA/Rbuild496651bfaf82/TIN/vignettes/TIN-scatter.pdf
pdf
  2
```

Third, the correlationPlot creates a plot that visualizes the number of splicing factor genes with expression levels significantly correlated with the sample-wise total relative amounts of aberrant exon usage.

```
> correlationPlot("TIN-correlation.pdf", tra, sampleSetGeneSummaries,
+     splicingFactors, 1000, 1000)
```

```
Plot was saved in  /tmp/RtmpBmnmtA/Rbuild496651bfaf82/TIN/vignettes/TIN-correlation.pdf
pdf
  2
```

Additionally, the posNegCorrPlot is a scatterPlot that compares the amount of splicing factor genes for which expression levels are significant positively (vertical axis) and negatively (horizontal axis) correlated with the total relative amounts of aberrant exon usage per sample.

```
> posNegCorrPlot("TIN-posNegCorrPlot.pdf", tra, sampleSetGeneSummaries,
+     splicingFactors, 1000, 1000)
```

```
Plot was saved in  /tmp/RtmpBmnmtA/Rbuild496651bfaf82/TIN/vignettes/TIN-posNegCorrPlot.pdf
pdf
  2
```

References

[1] Sveen, A., Agesen, TH., Nesbakken, A., Rognum, TO., Lothe, RA., Skotheim, RI., *Transcriptome instability in colorectal cancer identified by exon microarray analyses: Associations with splicing factor expression levels and patient survival*, Genome Medicine *3*, 32 (2011).

[2] Subramanian, A., Tamayo P., Mootha, VK., Mukherjee S., Ebert, BL., Gillette, MA., Paulovich, A., Pomeroy, SL., Golub, TR., Lander, SL., Mesirov, JP., *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles*, Proceedings of the National Academy of Sciences of the United States of America *102*, 15545-15550 (2005).

[3] Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, Arora VK, Kaushik P, Cerami E, Reva B, Antipin Y, Mitsiades N, Landers T, Dolgalev I, Major JE, Wilson M, Socci ND, Lash AE, Heguy A, Eastham JA, Scher HI, Reuter VE, Scardino PT, Sander C, Sawyers CL, Gerald WL: *Integrative genomic profiling of human prostate cancer.* Cancer Cell *18*:11-22 (2010).