TIN: An R Package for Transcriptome Instability Analysis

Bjarne Johannessen[1,2], Anita Sveen[1,2] and Rolf I. Skotheim[1–3]

[1]Department of Molecular Oncology, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway, [2]Centre for Cancer Biomedicine, Faculty of Medicine, University of Oslo, Oslo, Norway. [3]Department of Informatics, Faculty of Natural Sciences and Mathematics, University of Oslo, Norway.

ABSTRACT: Alternative splicing is a key regulatory mechanism for gene expression, vital for the proper functioning of eukaryotic cells. Disruption of normal pre-mRNA splicing has the potential to cause and reinforce human disease. Owing to rapid advances in high-throughput technologies, it is now possible to identify novel mRNA isoforms and detect aberrant splicing patterns on a genome scale, across large data sets. Analogous to the genomic types of instability describing cancer genomes (eg, chromosomal instability and microsatellite instability), transcriptome instability (TIN) has recently been proposed as a splicing-related genome-wide characteristic of certain solid cancers. We present the R package TIN, available from Bioconductor, which implements a set of methods for TIN analysis based on exon-level microarray expression profiles. TIN provides tools for estimating aberrant exon usage across samples and for analyzing correlation patterns between TIN and splicing factor expression levels.

KEYWORDS: alternative splicing, exon microarray, splicing factor, R software, transcriptome instability

CITATION: Johannessen et al. TIN: An R Package for Transcriptome Instability Analysis. *Cancer Informatics* 2015:14 109–112 doi: 10.4137/CIN.S31363.

TYPE: Software or Database Review

RECEIVED: July 06, 2015. **RESUBMITTED:** August 10, 2015. **ACCEPTED FOR PUBLICATION:** August 12, 2015.

ACADEMIC EDITOR: J.T. Efird, Editor in Chief

PEER REVIEW: Six peer reviewers contributed to the peer review report. Reviewers' reports totaled 1267 words, excluding any confidential comments to the academic editor.

FUNDING: This work was funded by the Norwegian Health Region South-East (project numbers 2011024 and 2012067) and the Norwegian Cancer Society (PR-2007–0166) and partly supported by the Research Council of Norway through its Centres of Excellence funding scheme (project number 179571). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: Rolf.I.Skotheim@rr-research.no

Published by Libertas Academica. Learn more about this journal.

Introduction

Cancers often harbor genomic types of instability, including chromosomal instability and microsatellite instability. However, cancer-associated variation may occur at several levels of gene regulation and, in particular, the processing of pre-mRNA into mature mRNAs is important for proper protein synthesis and cell function. Alternative pre-mRNA splicing is a major source of genetic variation in human beings, and disruption of the splicing process may cause cancer.[1,2] An improved understanding of the mechanisms that cause such structural transcript variation may provide important insights into disease development and progression. Alternative splicing is regulated by splicing factors, proteins that remove certain introns from the pre-mRNA, thereby joining the exons of the mRNA together. We have recently described transcriptome instability (TIN) in cancer, a genome-wide characteristic defined by the amounts of aberrant exon usage per sample, and shown that this is strongly and nonrandomly associated with splicing factor expression levels in several cancer types.[3,4]

High-resolution microarrays allow for genome-wide expression profiling at the exon level, enabling the detection of alternative splicing across a large series of samples. Here, we describe TIN, an R package enabling analysis of TIN from expression data obtained by Affymetrix Human Exon 1.0 ST Arrays. A major challenge in large-scale data analysis is reproducibility. With this aim, the TIN package consists of a set of unambiguous procedures that use raw expression data (cell intensity [CEL] files) as input, which are readily accessible and easy to extend. Information on how to install the package is provided in the Supplementary File.

Methods

The TIN software package is a collection of R modules that make use of the aroma.affymetrix[5] framework to analyze exon-level expression data. Starting from raw CEL files, the TIN tool applies the Finding Isoforms using Robust Multichip Analysis (FIRMA) method[6] for preprocessing and alternative splicing detection. The FIRMA method is an extension of the robust multichip analysis (RMA) approach[7] that not only estimates expression levels but also detects alternative splicing patterns between samples. Using the FIRMA method, the first two preprocessing steps, background correction of perfect match probes and inter-chip quantile normalization, are performed in concordance with standard RMA procedures. For the summarization step, a more general model that includes the relative change for each sample in a particular exon is introduced in the FIRMA approach to allow for alternative splicing or different levels of expression for each exon along the gene. For each exon sample combination, the FIRMA method calculates alternative splicing scores, FIRMA scores, based on whether the probes systematically deviate from the expected gene expression level. Thus, the FIRMA scores are a measure of the relative ratio between exon expression level and corresponding gene expression level. Strong positive and negative scores are indicative of differential exon inclusion and skipping, respectively.

The main idea is to test the association between splicing factor expression levels and the amounts of aberrant exon usage among samples. Sample-wise total relative amounts of aberrant exon usage are recorded from exons with FIRMA scores exceeding user-defined thresholds, and the correlation between aberrant exon usage amounts and splicing factor expression levels is tested across all samples. Two methods are implemented for testing if the correlation between sample-wise aberrant exon usage amounts and splicing factor expression levels is stronger than expected by chance. First, permutations of the FIRMA scores are done for each probe set/exon across all samples, and the sample-wise amounts of aberrant exon usage are recalculated based on the permutations. If the correlation between the aberrant exon usage amounts and splicing factor expression levels is considerably lower when based on permutations compared to the original FIRMA scores, it is an indication of splicing factor expression having impact on the aberrant exon usage in the samples. Second, correlation is tested using a number of miscellaneous gene sets instead of the original set of 280 splicing factor genes. Equivalently, poorer correlation for random gene sets compared to the splicing factor set can be considered an indication that the aberrant exon usage to some extent was attributable to the expression levels of the splicing factor genes. An overview of the pipeline is outlined in Figure 1.

Example

Five R data sets are included in the package. By issuing the following commands:

```
data(splicingFactors)
data(geneSets)
data(geneAnnotation),
```

data.frames with the three sets of data will become available. The first object is a comprehensive list of 280 splicing factor genes created by combining search results from several public annotation databases.[3] Second, one of the major collections of gene sets in the Molecular Signatures Database, MSigDB,[8] comprising 1,454 Gene Ontology gene sets, is included to see if the association between aberrant exon usage and gene expression levels is different in the splicing factor gene set compared to more general gene sets. Third, a list of matching gene symbols and Affymetrix transcript cluster identifiers for the full genome (core set of human genes) are provided in the annotation data set to provide easy access and enable generation of new gene sets. The main purpose behind the TIN package is to facilitate reproducibility through a consistent set of

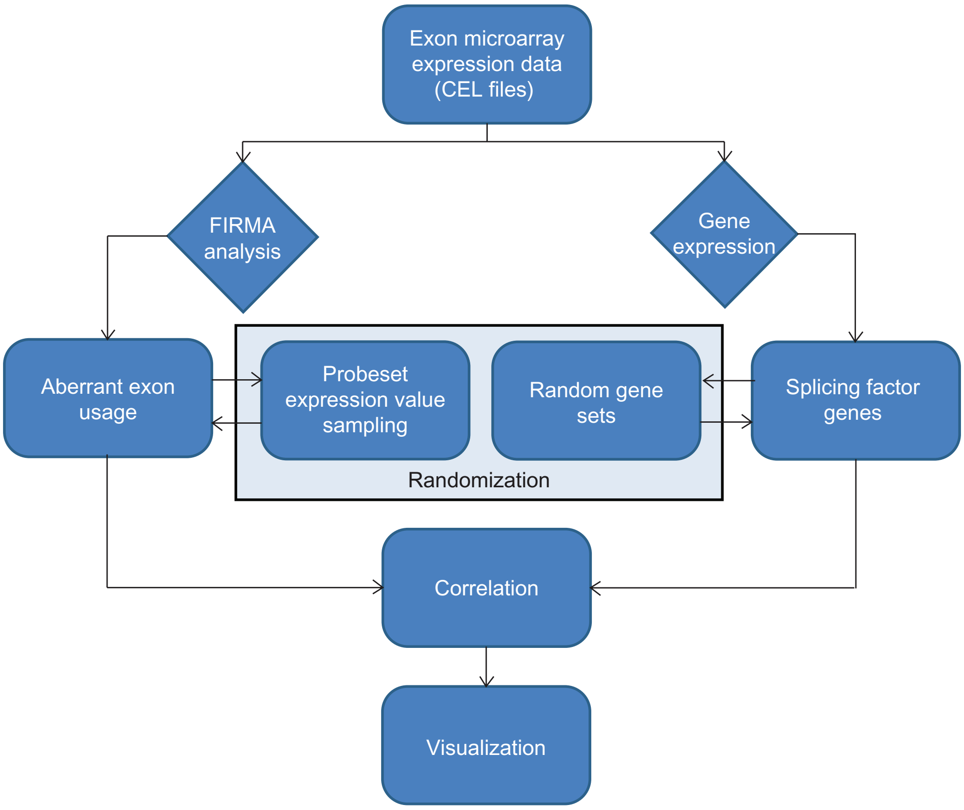

Figure 1. Pipeline to investigate TIN in tumor samples based on exon-level microarray data. CEL files with raw expression data is taken as input, along with gene-level expression data. The FIRMA algorithm is used to identify exon skipping and inclusion events, and user-defined thresholds (such as the upper and lower first percentile) are used for denoting exons as aberrantly spliced. The correlation between aberrant exon usage and splicing factor gene expression is evaluated and tested against random associations in two ways. First, the correlation step is carried out using permutations of the expression data at each probe set. Second, the correlation is calculated using random gene sets instead of known splicing factor genes.

algorithms, which may be applied on real-world data; however, for educational purposes, a small toy data set is embedded in the release. Thus, preprocessed FIRMA scores for 16 samples and 10,000 randomly selected probe sets are included in the sampleSetFirmaScores object. Equivalently, gene-level expression data for the same 16 samples across the core set of human genes are provided through the sampleSetGeneSummaries object. Summary files for real gene-level expression data can be generated by using, for instance, Affymetrix Power Tools or Expression Console prior to applying the TIN package. The analysis pipeline is outlined in the following example, with expression data from 131 prostate cancers.[9] The data set is publicly available from NCBI's Gene Expression Omnibus (GEO; accession number GSE21034).

```
fs <- firmaAnalysis(useToyData = FALSE, aromaPath = "/path/to/aroma.affymetrix", dataSetName = "Prostate")
gs <- readGeneSummaries(useToyData = FALSE, summaryFile = "/path/to/prostate-gene-level-summary.txt")
```

To use the small toy data set supplied with the package instead, load the sample data by issuing the following two commands

```
data(sampleSetFirmaScores)
data(sampleSetGeneSummaries,
```

and copy the two objects into the fs and gs variables, respectively.

```
tra <- aberrantExonUsage(1.0, fs)
perms <- probesetPermutations(fs, quantiles)
corr <- correlation(splicingFactors, gs, tra)
gsc <- geneSetCorrelation(geneSets, geneAnnotation, gs, tra, 100)
```

In the example, the lower and upper 1st percentiles are used as threshold values to score exons with deviating skipping or inclusion (Fig S1). Information on where to find documentation of the different functions is provided in the Supplementary File.

Having performed FIRMA analysis and entered gene-level expression values, sample-wise amounts of aberrant exon usage are calculated. Pearson correlation between relative amounts of aberrant exon usage and splicing factor expression is obtained using tools from the WGCNA package (Bioconductor).[10] To assess the association, correlation is also calculated for random permutations of the FIRMA scores at each probe set and for random sets of genes.

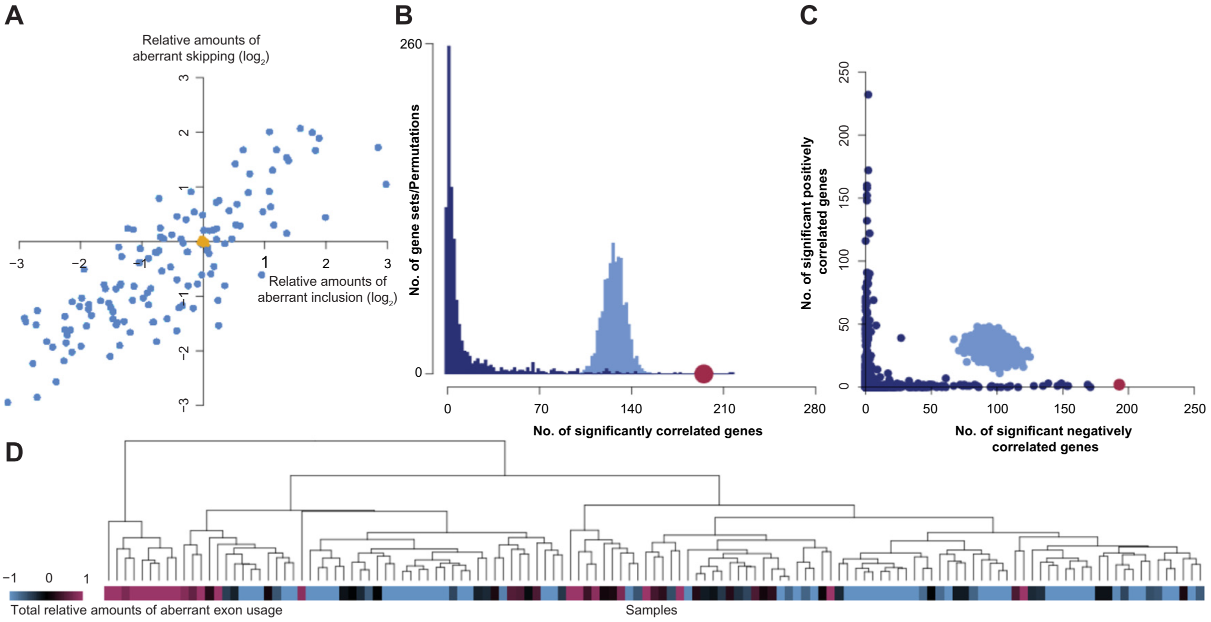

Figure 2. (**A**) Sample-wise relative amounts (blue dots) of aberrant exon inclusion (horizontal axis) and exon skipping (vertical axis) events for the 131 prostate cancers in the worked example, compared to random sample-wise amounts calculated from permuted FIRMA scores (yellow dots). (**B**) Correlation between estimated aberrant exon usage and splicing factor expression compared with random gene sets and permuted TIN-estimates. In the example cancer dataset, 195 of the 280 (70%) splicing factor genes had expression levels that were significantly correlated ($P < 0.05$; Pearson correlation; red dot; horizontal axis). This is more than expected by chance, as compared with first making 1,000 random permutations of the TIN-estimates (bar graphs in dark blue) and second by selecting 1,000 random sets of 280 genes (bar graphs in light blue). (**C**) Negative correlation between TIN-estimates and splicing factor expression in the example prostate cancer dataset. Inverse relationship with strong associations between TIN-estimates and expression levels of splicing factors ($n = 280$), with a much higher percentage of significantly negatively (horizontal axes) than positively (vertical axes) correlated splicing factor genes (red). The shift was higher than expected by chance, as demonstrated by comparing first with each of 1,000 permutations of the TIN-estimates (dark blue) and second with genes in each of 1,000 random sets of 280 genes (light blue). (**D**) Unsupervised hierarchical clustering analysis (Euclidean distance metrics; complete linkage) of all the 131 samples based on the expression levels of all 280 splicing factor genes. The example prostate series is separated into clusters with some samples having predominantly lower (blue) or higher (red) relative amounts of deviating exon usage than the more average sample (black).

Visualization. The TIN package implements four different methods for visualizing the results (Fig. 2). First, a scatter plot visualizes sample-wise relative amounts of aberrant exon inclusion vs exon exclusion and optionally includes amounts calculated from random permutations of the FIRMA scores for comparison (Fig. 2A). Second, the package includes a function for comparing sample-wise correlation between splicing factor gene expression and total relative amounts of aberrant exon usage, with correlations obtained either by making permutations of the sample-wise amounts of aberrant exon usage or by using randomly generated gene sets (Fig. 2B). Third, a scatter plot that compares the amount of splicing factor genes for which expression levels are significantly positively and negatively correlated with the total relative amounts of aberrant exon usage per sample is created (Fig. 2C). This plot may also include results based on permutations of the sample-wise aberrant exon usage amounts and randomly constructed gene sets. In addition, a function for hierarchical clustering of the samples based on splicing factor expression levels is included to test for separation of samples according to aberrant exon usage amounts (Fig. 2D). Example commands for creating visualization plots are outlined below:

```
scatterPlot(“scatter.png”, TRUE, hits, perms)
correlationPlot(“correlation.png”, tra, gs, splicingFactors, 1000, 1000)
posNegCorrPlot(“posNegCorr.png”, tra, gs, splicingFactors, 1000, 1000)
clusterPlot(gs, tra, “euclidean”, “complete”, “cluster.png”)
```

Further instructions on parameter usage and how the methods work are provided in the accompanying vignette and documentation of the package.

Results and Discussion

We have developed the TIN package (Bioconductor) to analyze TIN in cancer or other disease conditions from exon-level microarray data. By using computational tools already available to create algorithms for analyzing TIN, the package offers a framework for calculating and visualizing the correlation between sample-wise aberrant exon usage amounts and expression levels of multiple gene sets, including splicing factors. The R software has been applied to expression data from different cancer types, and we have shown that TIN is a common feature of several types of solid cancer. In most cancer types studied, we found strong and nonrandom ($P < 0.001$) correlations between the estimated aberrant exon usage and the expression levels of splicing factor genes.[4] When analyzing multiple data sets, it is of great importance to be able to repeat and standardize computational methodology. The TIN package facilitates reproducibility through an unambiguous analysis pipeline.

Author Contributions

Conceived and designed the methods: BJ, AS, RIS. Wrote the software code: BJ. Wrote the first draft of the manuscript: BJ. Contributed to the writing of the manuscript: AS, RIS. Agreed with manuscript results and conclusions: BJ, AS, RIS. Developed the structure and arguments for the paper: BJ. Made critical revisions and approved the final version: BJ, AS, RIS. All the authors reviewed and approved the final manuscript.

Supplementary Material

Supplementary File. This file contains installation guidelines, links to documentation for the TIN package, and Figure S1, a distribution plot of all FIRMA scores in the worked example.

REFERENCES

1. David CJ, Manley JL. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes Dev*. 2010;24(21):2343–64.
2. Skotheim RI, Nees M. Alternative splicing in cancer: noise, functional, or systematic? *Int J Biochem Cell Biol*. 2007;39(7–8):1432–49.
3. Sveen A, Agesen T, Nesbakken A, Rognum T, Lothe R, Skotheim R. Transcriptome instability in colorectal cancer identified by exon microarray analyses: associations with splicing factor expression levels and patient survival. *Genome Med*. 2011;3(5):32.
4. Sveen A, Johannessen B, Teixeira M, Lothe R, Skotheim R. Transcriptome instability as a molecular pan-cancer characteristic of carcinomas. *BMC Genomics*. 2014;15(1):672.
5. Bengtsson H, Simpson K, Bullard J, Hansen K. aroma.affymetrix: a generic framework in R for analyzing small to very large Affymetrix data sets in bounded memory. 2008. Tech. rep. 745. Department of Statistics, University of California, Berkeley.
6. Purdom E, Simpson KM, Robinson MD, Conboy JG, Lapuk AV, Speed TP. FIRMA: a method for detection of alternative splicing from exon array data. *Bioinformatics*. 2008;24(15):1707–14.
7. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003;4(2):249–64.
8. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545–50.
9. Taylor BS, Schultz N, Hieronymus H, et al. Integrative genomic profiling of human prostate cancer. *Cancer Cell*. 2010;18(1):11–22.
10. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;2008(9):559.