Chapter 8

Analysis of Isoform Expression from Splicing Array Using Multiple Comparisons

T. Murlidharan Nair

Abstract

There is a high prevalence of alternatively spliced genes (isoforms) in the human genome. Studies toward understanding aberrantly spliced genes and their association with diseases have lead researchers to profile the expression of alternatively spliced products. High-throughput profiling of isoforms has been done using microarray technology. Expression of isoforms reflects regulation both at transcriptional and posttranscriptional levels. This chapter details the methods to perform exhaustive comparison of isoforms using the R statistical framework.

Key words: mRNA isoforms, Multiple comparisons

1. Introduction

Alternative pre-mRNA splicing (AS) responsible for generating multiple transcripts from a single gene plays a central role in generating complex proteomes (1). It is estimated that more than 90% of the human genes have alternatively spliced products. Over the years, studies directed toward understanding alternative splicing using computational approaches have gained increased attention (2–5). Several studies have used microarray technology to quantify isoform expression levels either directly or indirectly (6–9). Quantifying isoform expression levels has the advantage in that it reflects the integrated outcome of the regulations at transcriptional and posttranscriptional levels. There is evidence that points to the functional integration of processes involved in transcription and RNA processing (10).

There are several disparate microarray platforms that have been used for expression analysis (11, 12); however, most platforms are not designed to specifically query isoforms. Multiplex mRNA isoform detection assays known as RASAL or DASL

Junbai Wang et al. (eds.), *Next Generation Microarray Bioinformatics: Methods and Protocols*,
Methods in Molecular Biology, vol. 802, DOI 10.1007/978-1-61779-400-1_8,

(RNA/DNA-mediated annealing, selection, and ligation), coupled with microarray were designed to uniquely profile mRNA isoforms in a high-throughput manner (13, 14). This chapter provides the computational methods for analyzing and extracting biological information from isoform expression data. For the purpose of this chapter, we have used data from Illumina BeadArray technology; however, the method described here can be easily extended to data collected from other high-throughput technologies, with some preprocessing of the data.

2. Materials

2.1. Hardware and Software Requirements

The computational protocol that is described here requires the following:

R is an open-source statistical computing environment available under the GNU Public License for different platforms (Windows/Linux/Unix/Mac) (15). R was developed by Robert Gentleman and Ross Ihaka. It has quickly become the language of choice for most large-scale computational analyses in Biostatistics and Bioinformatics. R has a command line interface where R commands are typed in. R has a rich library of add-on packages that has been developed for specific types of analyses. All the packages are available free to the user.

R can be downloaded from ref. 16. Binary versions are easy and straightforward to install. The analysis described in this chapter makes use of the multcomp package to carry out multiple-hypothesis testing (17). The multcomp may be installed using the R interface. It can be done by clicking on "packages" from the main menu and choosing "Install package(s)." Choose a mirror site closest to you geographically, and then choose the required package, in this case "multcomp", to be installed.

2.2. Dataset

The methods described here use the data generated using Illumina BeadArray (6, 18). For details of how the data was generated, the reader may refer to the original article by Li et al. (6). While there are several technologies that have been used for gathering information on expression of isoforms, the methods described in this chapter are not specific to any particular type of data set. However, some preprocessing of the data may be required so as to map the data obtained using other technologies to the one obtained using the BeadArray. For instance, the Affymetrix approach uses multiple probes to query a transcript; thus, care should be taken to combine the expression values from probes that query the same exon. This can then be used to compare expression levels of different exons within the same transcript using the method described here.

Table 1
Example of normalized isoform expression data: columns represent cell lines and rows represent isoforms/splicing event associated with the ATP-binding cassette, subfamily G, member 1 gene (ABCG1)

Isoform	HCE-7	HCE-7	MDA.MB-468	MDA.MB-468	PC3-E	PC3-E	PC3-E	DU145-E	DU145-E	DU145-E
ABCG1-0489	461.76	488.89	391.04	380.10	1,088.46	999.51	1,153.58	403.81	373.71	394.07
ABCG1-0490	507.13	479.91	541.21	676.18	275.79	272.72	260.40	277.59	258.48	257.83
ABCG1-0491	329.69	316.55	375.43	369.25	272.33	256.72	271.71	269.83	265.45	255.15
ABCG1-0492	337.54	338.47	441.79	456.05	248.00	248.04	260.30	246.47	245.60	252.56
ABCG1-0494	197.15	195.35	210.37	193.22	215.94	215.83	207.31	204.67	194.74	216.36
ABCG1-0495	279.85	326.66	491.95	601.24	1,132.31	1,207.01	1,260.44	429.28	421.67	389.16
ABCG1-1482	257.93	286.40	308.56	378.81	632.61	664.51	657.83	341.80	323.43	325.45
ABCG1-1483	212.34	200.27	203.42	219.46	188.74	214.67	209.19	220.03	196.99	207.44

The numbers following the gene name ABCG1 correspond to the different splicing events and are assigned at the time of experimental design.

2.2.1. Isoform Expression Data

The isoform expression data is read from a comma-separated value file (csv): each column represents a biological sample (cell line/tissue) and each row represents a different isoform or splicing event (see Table 1).

3. Methods

3.1. Experimental Design and Normalization

When profiling expression of isoforms/splicing events from biological samples, it is important to ensure that one takes the necessary steps to process the samples in batches and have biological and technical replicates. Careful attention should be paid when designing probes to minimize interference with hybridization due to secondary structure. Expression data from RASL/DASL assay used here has high specificity and sensitivity in querying isoform expression. The ligation step contributes to the specificity and the PCR step significantly enhances the sensitivity (6) in the assay. When extracting isoform expression information from other technologies like Affymetrix that use multiple probes, appropriate care should be taken to assign expression values to isoforms/splicing events (see Note 1) (19, 20).

Microarray data needs to be normalized before different data sets can be cross compared. Normalization enhances meaningful data characteristics and accounts for systematic differences across data sets. There are several methods that may be used to normalize expression data (21–23). The data used here was normalized against a synthetic average using locally weighted polynomial regression (LOWESS) (24). LOWESS uses a polynomial of degree 1 or 2, thus avoiding over-fitting. The procedure divides the data domain into several windows and uses the polynomial only to approximate over a narrow interval. Since normalization is not a one-size-fits-all solution, the user should decide, based on the data they have, which method is most suitable for their data. It is assumed here that data has been normalized.

3.2. Multiple Comparisons of Isoform Expression

In analyzing isoform expression data, we are confronted with the problem of testing the differences in expression between many means. This can be conveniently tackled using multiple comparisons. Differential analysis of isoform expression involves all possible comparisons and can be conveniently done using the R multcomp package (25). It is noteworthy to mention that such comparisons are compute intensive and it is advisable to use parallel processing (see Note 2). The output is in the form of confidence intervals, significant comparisons are those that do not intersect the zero line. We demonstrate the exhaustive comparisons using the data given in Table 1. R-code given in Table 2 can be used to carry out the analysis.

Table 2
R-code for carrying out the exhaustive comparisons using the multcomp package

```
1  library(mvtnorm)
2  library(multcomp)
3  par(mfrow = c(1,1),cex = 0.7, mai = c(3,2,1,2), ask = T)
4  complete.data < −read.csv("isoformSubset.csv",header = T)
5  lgth < −length(complete.data[1,])-1
6  complete.data.mat < −as.matrix(complete.data[,1:lgth + 1])
7  cell.line < −0
8  complete.data.frame < −as.data.frame(complete.data)
9  filename < −as.vector(complete.data.frame$Isoform)
10 cell.line < −colnames(complete.data.mat)
11 cell.line < −as.factor(substr(cell.line, 1,c(5,5,8,8,5,5,5,5,7,7,7)))
12 number.rows < −nrow(complete.data.mat)
13 i < −0
14 Expression < −0
15 mult.comp < −0
16 for(i in 1:number.rows)
17 {
18  cat("Now computing::- > ", filename[i],"\n")
19  for(j in 1:(lgth)){
20   Expression[j] < −complete.data.mat[i,j]
21 }
22  Expression < −as.numeric(Expression)
23  isoform.expression < −data.frame(cell.line,Expression)
24  isoform.expression$cell.line < −factor(isoform.expression$cell.line)
25  amod < −aov(Expression ~ cell.line, data = isoform.expression)
26  mult.comp < −glht(amod,linfct = mcp(cell.line = "Tukey"))
27  conf.int < −confint(mult.comp,level = 0.99)
28  plot(conf.int, main = filename[i],xlab = "99% Confidence interval")
29  p.value < −summary(mult.comp)$test$pvalues
30  out.data.mat < −data.frame(conf.int$confint[,1:3],p.value)
31  filename.csv < −paste(filename[i], "csv",sep = ".")
32  write.table(out.data.mat, file = filename.csv, sep = ",", qmethod = "double", col.name = NA)
33  rm(amod,mult.comp,conf.int,p.value,out.data.mat,filename.csv)
34 }
```

Table 3
Output of the comparison of isoform ABCG1-0490

	Estimate	lwr	upr	*p*-Value
HCE.7-DU145.E	228.88653	44.68726	413.0858	0.003281
MDA.MB.4-DU145.E	344.06312	159.8638	528.2624	0.000266
PC3.E-DU145.E	4.9996982	−159.753	169.7525	0.998626
MDA.MB.4-HCE.7	115.17659	−86.6036	316.9568	0.10403
PC3.E-HCE.7	−223.8868	−408.086	−39.6876	0.003638
PC3.E-MDA.MB.4	−339.0634	−523.263	−154.864	0.000376

The preceding code may be written using any ASCII editor and saved as an R file. Lines 1 and 2 ensure that the two libraries are loaded. Line 3 sets the parameter for plotting. You may change these according to your requirements. Reading the isoform expression data is achieved in line 4. It is assumed here that the name of the file is "isoformSubset.csv." You should substitute your isoform expression data file name. Line 9 uses the isoform name from the expression data to create a file name to store the results of the analysis for a particular isoform. Line 11 creates a factor, in this case using the cell line names from the expression data. The substr function in line 11 is used to eliminate any additional differentiators that R introduces when the file header contains duplicate names. You may need to make changes to the substr function to reflect the size of the headers you have used. Lines 25 through 27 help achieve the multiple comparison. Confidence level used in computing the confidence intervals is set to 0.99 in line 27 to ensure low probability of type I error. Line 32 writes the output of each comparison to a file that has the isoform name as its filename. Table 3 shows a typical output that is written to the file created in line 32. In the interest of brevity, data contained in only one output file is shown.

3.3. Interpretation and Further Processing of the Output

The plots obtained from execution of line 28 are shown in Fig. 1. These plots are the graphical representation of the confidence intervals for the comparisons. The significant comparisons are those that do not intersect the zero line. Only comparisons for four of the isoforms are shown. The plots clearly show that there is a significant difference in expression of the isoform ABCG1-0490 between HCE.7 and DU145.E, and between MDA.MB.4 and DU145.E. The isoform ABCG1-0495 does not show a significant difference in expression between HCE.7 and DU145.E, and between MDA.MB.4 and DU145.E. Further, the isoform ABCG1-0494 does not show any significant difference in expression in any of the comparisons, as in all cases we see an intersection of the zero line.

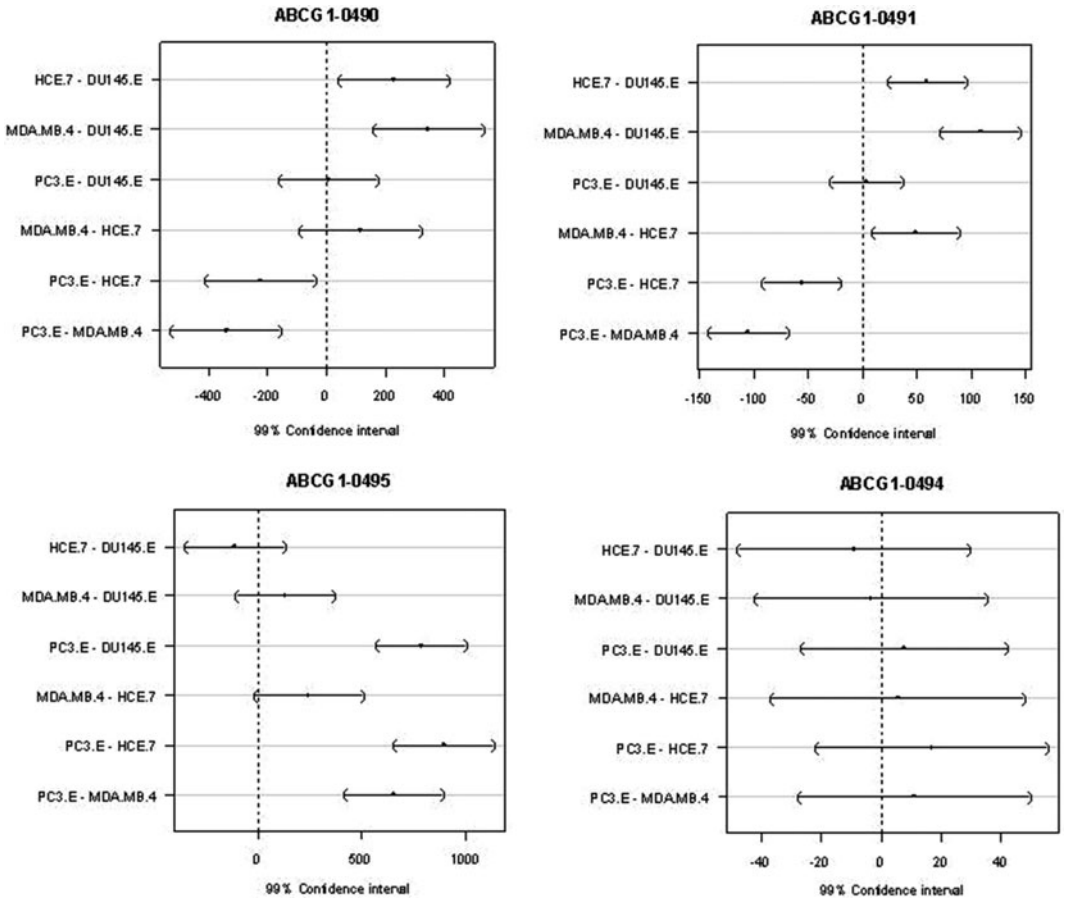

Fig. 1. Multiple comparisons on expression level of four different isoforms of the gene ABCG1. Comparisons that show significant difference in expression level are the ones that do not intersect the zero line.

The subset of data used here was part of a study to identify differential expression of isoforms in prostate cancer cell lines and nonprostate cancer cell lines (6, 18). The data generated as a result of this study consisted of isoform expression from cell lines. The cell lines for which expression data were collected included five prostate cancer cell lines, viz., LNCap, LAPC4, RWPE2, PC3, and DU145, and twelve nonprostate cancer cell lines, viz., colon cancer line (HT29, SW480, HCT116, LS174, Fet), breast cancer line (MCF7, MDA.MB-468), kidney cancer line (Caki-2), lung epidumoid carcinoma line (CALU1), and esophageal cancer lines (HCE-7, EC17 and TE3). Isoforms that exhibit differential expression between two classes of samples can be delineated from the output generated using multiple comparisons. Each isoform is given a unit score for every significant difference it showed in a comparison. The sum of the scores can be used to rank the isoform. In the example that we are using here, the isoforms ABCG1-0490 and ABCG1-0491 each have a sore of 4.

Even though the comparison between HCE.7 and MDA.MB.4 is significant, it is not considered, as both are nonprostate cancer cell lines. Isoform ABCG1-0495 has a score of 3, while ABCG1-0494 has a score of 0. Assignment of scores may be decided depending upon the question you are trying to answer, that is, whether you are doing a within-class comparison or a between-class comparison. Top ranking isoforms may be used as features for class separation or may be further studied to understand their potential to serve as biomarkers. Further, isoform levels may also reflect on the different levels of control that may be teased out in a problem-specific manner (see Note 3).

4. Notes

1. *Processing of expression data from disparate microarrays.* Not all microarrays permit the direct measurement of isoform expression. The data used in this chapter was from specially designed arrays that queried for splicing events. Isoform expression may be derived from Affymetrix that uses multiple probes. However, this would require deducing isoform information based on the probes that query the gene of interest. Care must be taken when such preprocessing is done and would require careful annotation of the probes to reflect the isoform being queried.
2. *Computational capacity issues.* Multiple comparisons are compute intensive, especially when one handles large datasets. It is advisable to use a cluster and process the data in parallel. The R/Parallel package helps to conveniently achieve this (26). In addition to this, computing efficiency may be improved by processing subsets of data and avoiding redundant comparisons.
3. *Deconvoluting controls at levels of transcription and splicing.* Controls of mRNA expression may be regulated at levels of transcription, RNA stability, and splicing. Depending on the type of data collected, it may be possible to tease this information from the data. For instance, multiple isoforms that are similarly elevated or depressed would indicate coordinated changes in transcription and/or RNA stability (6). The transcript change may be computed as the sum of the weighted fold change of the isoforms involved. The splicing change may be computed as the difference in fold change of the two isoforms. Thus, for isoforms that are similarly up- or down-regulated, the splicing change would be close to zero. These computations are data dependent and the reader is referred to an earlier work by the author for details of a specific case (6).

Acknowledgement

TMN would like to thank IUSB for research funding.

References

1. Matlin AJ, Clark F, Smith CW (2005) Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol 6:386–398.
2. Kim N, Lee C (2008) Bioinformatics detection of alternative splicing. Methods Mol Biol 452:179–197.
3. Ferreira EN, Galante PA, Carraro DM et al (2007) Alternative splicing: a bioinformatics perspective. Mol Biosyst 3:473–477.
4. Chacko E, Ranganathan S (2009) Comprehensive splicing graph analysis of alternative splicing patterns in chicken, compared to human and mouse. BMC Genomics 10:S5.
5. Lee C, Wang Q (2005) Bioinformatics analysis of alternative splicing. Brief Bioinform 6:23–33.
6. Li HR, Wang-Rodriguez J, Nair TM et al (2006) Two-dimensional transcriptome profiling: identification of messenger RNA isoform signatures in prostate cancer from archived paraffin-embedded cancer specimens. Cancer Res 66:4079–4088.
7. Blencowe BJ (2006) Alternative splicing: new insights from global analyses. Cell 126:37–47.
8. Johnson JM, Castle J, Garrett-Engele P et al (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science 302:2141–2144.
9. Pando MP, Kotraiah V, McGowan K et al (2006) Alternative isoform discrimination by the next generation of expression profiling microarrays. Expert Opin Ther Targets 10:613–625.
10. Pandit S, Wang D, Fu XD (2008) Functional integration of transcriptional and RNA processing machineries. Curr Opin Cell Biol 20:260–265.
11. Hardiman G (2004) Microarray platforms – comparisons and contrasts. Pharmacogenomics 5:487–502.
12. Lee NH, Saeed AI (2007) Microarrays: an overview. Methods Mol Biol 353:265–300.
13. Yeakley JM, Fan JB, Doucet D et al (2002) Profiling alternative splicing on fiber-optic arrays. Nat Biotechnol 20:353–358.
14. Fan JB, Yeakley JM, Bibikova M et al (2004) A versatile assay for high-throughput gene expression profiling on universal array matrices. Genome Res 14:878–885.
15. http://www.r-project.org.
16. http://cran.r-project.org.
17. Hothorn T, Bretz F, Westfall P (2008) Simultaneous inference in general parametric models. Biom J 50:346–363.
18. Nair TM (2009) On selecting mRNA isoform features for profiling prostate cancer. Comput Biol Chem 33:421–428.
19. Bemmo A, Benovoy D, Kwan T et al (2008) Gene expression and isoform variation analysis using Affymetrix Exon Arrays. BMC Genomics 9:529.
20. Bemmo A, Dias C, Rose AA et al (2010) Exon-level transcriptome profiling in murine breast cancer reveals splicing changes specific to tumors with different metastatic abilities. PLoS ONE 5: e11981.
21. Bolstad BM, Irizarry RA, Astrand M et al (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19:185–193.
22. Zeller G, Henz SR, Laubinger S et al (2008) Transcript normalization and segmentation of tiling array data. Pac Symp Biocomput: 527–538.
23. Haldermans P, Shkedy Z, Van Sanden S et al (2007) Using linear mixed models for normalization of cDNA microarrays. Stat Appl Genet Mol Biol 6:Article 19.
24. Cleveland WS (1979) Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association 74:829–836.
25. Hothorn T, Bretz F, Westfall P et al (2008) Multcomp: Simultaneous Inference for General Linear Hypotheses. URL http://CRAN.R-project.org.
26. Vera G, Jansen RC, Suppi RL (2008) R/parallel – speeding up bioinformatics analysis with R. BMC Bioinformatics 9:390.