XhybCasneuf package

Tineke Casneuf

October 29, 2020

This package contains the data that was generated by and investigated in our study, 'The Effect of Cross-Hybridisation on Expression Correlations Inferred from Microarrays' (manuscript submitted).
In the first part, we investigated the relationship between reporter-to-transcript sequence similarity and correlation of expression signals. We assessed the extent to which off-target reporters in probe sets, i.e. reporters that (partially) align to another transcript than the one intended, influence the expression correlation of the intended and off-target probe set. For a given probe set X, intented to target gene X and a potential off-target gene Y, the variables were calculated as followed:
First X's reporters' sequences were aligned to the transcript of Y. To quantify the potential off-target affinity of probe set X to Y, the 75th percentile Q_{XY}^{75} was then calculated of these alignment scores $\{a_1, \ldots, a_n\}$. We obtained the expression correlation measure by calculating the Pearson correlation coefficient between the intensity values of X and Y in 14 different plant tissues. These data are a subset of the AtgenExpress project data [1, 2].
We illustrated our findings with three detailed examples of cross-hybridisation. We also ran a simulation to demontstrate the effect of individual reporters on summarisation.
Our study shows that numerous probe sets on a widely used commercial array platform contain off-target reporters, and many probe sets show a signal pattern that is highly similar to that of unintended transcripts. In addition, a positive correlation is revealed between off-target alignment score of different reporters and the magnitude of their off-target correlation. We evaluated the conventional probe set design, as defined by the array manufacturer (Affymetrix CDF), and compared it to a custom-made CDF. We demonstrate that careful reporter mapping alleviates cross-hybridisation effects to a substantial extent. This analysis was conducted on the ATH1 Affymetrix GeneChip for *Arabidopsis thaliana*.
This package contains, for both the Affymetrix and the custom-made CDF, the data of probe set pairs with a off-target sensitivity score of ≥ 55.
Load the library:

```
> library("XhybCasneuf")
```

1 Probe set off-target sensitivity and expression correlation

First, we will have a look at the relation between expression correlation and the off-target sensitivity of these all pairs of the Affymetrix and our custom-made CDF. Because a positive trend between (reporter) alignment strength and expression correlation is not unexpected for functionally related genes like paralogous genes or genes that share protein domains, we omitted gene pairs that aligned to each other with BLAST [3] in at least one direction with an E-value smaller than 10^{-10}.

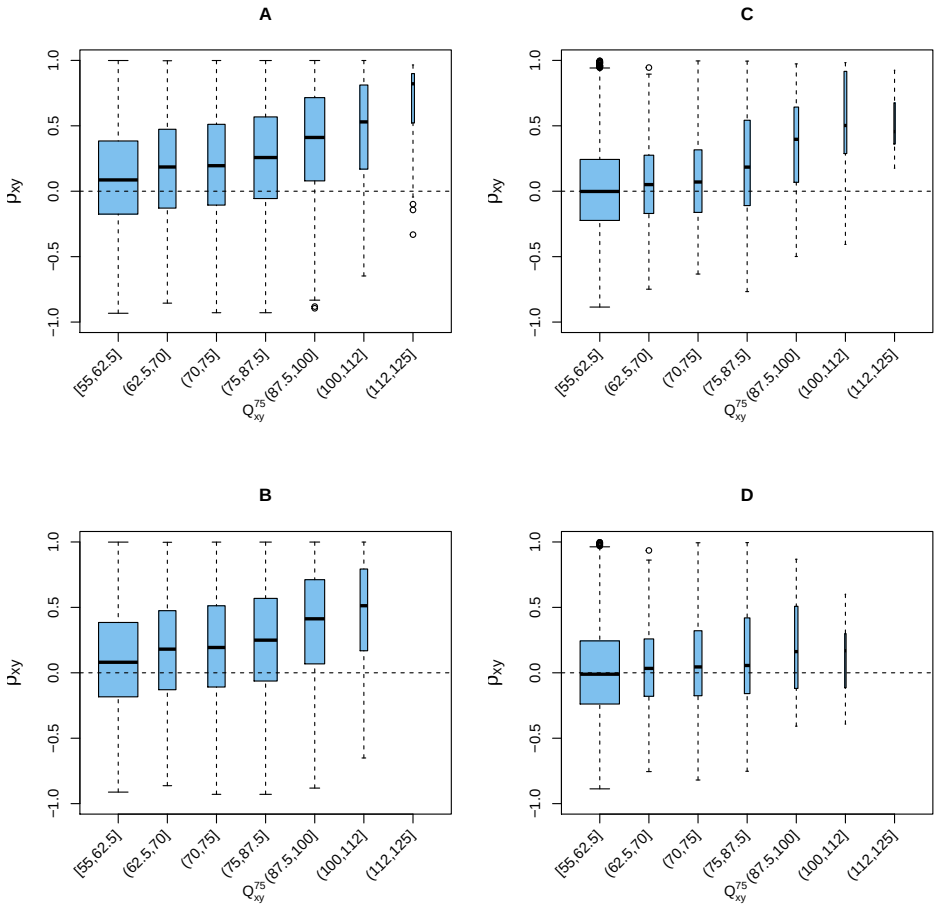
```
> ## all probe set pairs
> data(AffysTissue)
> data(CustomsTissue)
> ## probe set pairs of genes that do not align to each other with BLAST with a E-value smal
> data(AffysTissue.noBl)
> data(CustomsTissue.noBl)
```

We now write function that will construct boxplots of these 4 data sets:

```
>  myXs <- c(seq(55,70, length.out =3),seq(75,125,length=5))
>  tiltedmyboxF <- function(X,Y, main){
+    par(mar = c(7, 4, 4, 2) + 0.1)
+    boxplot(Y~X, col = "skyblue2", ylim=c(-1,1), ylab = expression(rho[xy]), varwidth=T,xax
+    axis(1, labels = FALSE)
+    text(seq_len(nlevels(X)), par("usr")[3] - 0.15, srt = 45, adj = 1, labels = levels(X),
+    text(4, par("usr")[3] - 0.6, adj = 1, labels = expression(Q[xy]^75), xpd = TRUE)
+    abline(0,0, lty=2)
+   }
```

And we plot them:

```
> # X11(height = 10, width = 10)
>  layout(matrix(1:4, nrow = 2, byrow = F))
>   # ALL PAIRS of the custom and affymetrix CDF
>   tiltedmyboxF(cut(AffysTissue$alSum, myXs, include.lowest = TRUE, right = TRUE), AffysTis
>   tiltedmyboxF(cut(CustomsTissue$alSum, myXs, include.lowest = TRUE, right = TRUE), Custom
>   # EXCLUDE gene pairs with BLAST HIT with E-value < 10^-10
>   tiltedmyboxF(cut(AffysTissue.noBl$alSum, myXs, include.lowest = TRUE, right = TRUE), Afi
>   tiltedmyboxF(cut(CustomsTissue.noBl$alSum, myXs, include.lowest = TRUE, right = TRUE), C
```



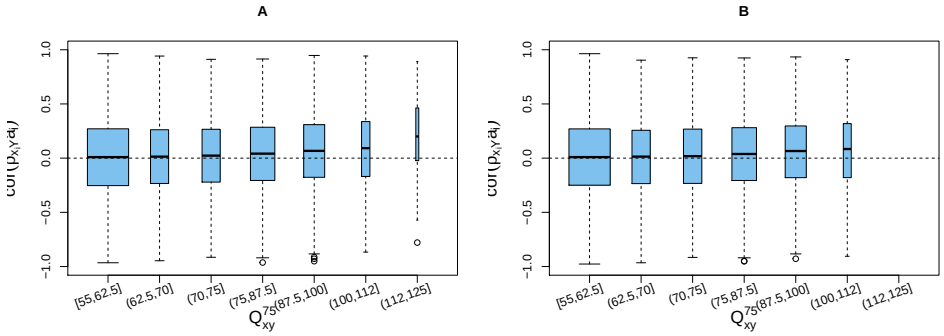
These boxplots reveal a positive relation between the two variables: a gene whose expression is measured by reporters that align well to a different transcript tends to have an expression signal that is correlated with that of the other transcript (plots A and B). This positive trend is present even between gene pairs that do not share longer stretches of sequence similarity and where the reporter to off-target alignment is only based on short near-matches (plots C versus A and D versus B).

2 Reporter off-target sensitivity and expression correlation

We also studied the behavior of off-target sensitivity and signal correlation of different reporters within a probe set. For a probe set X and an off-target gene Y, we calculated the metacorrelation $\mathrm{cor}(\rho_{X_iY}, a_i)$ between the alignment scores a_i of X's reporters to Y's transcript sequence and the Pearson correlation coefficients of the reporters' signal patterns to the expression pattern of Y. We reasoned that if cross-hybridisation occurs, a positive trend between reporter to off-target correlation and the alignment score a_i can be detected. Conversely, lack of such a trend may indicate that cross-hybridisation is negligible.

```
> data(AffysTissueMC)
> data(CustomsTissueMC)
> myboxplot <- function(X, Y, main){
+  boxplot(Y~X, col = "skyblue2", ylim=c(-1,1), ylab = expression(cor(rho[x[i]*Y]*a[i])), va
+  axis(1, labels = FALSE)
```

```
+  text(seq_len(nlevels(X))+0.25, par("usr")[3] - 0.12, srt = 20, adj = 1, labels = levels(X
+  abline(0,0, lty=2)
+ }
>  par(mfrow=c(1,2))
>  # ALL PAIRS of the Affymetrix CDF
>  X <- cut(AffysTissueMC$alSum, myXs, include.lowest = TRUE, right = TRUE)
>  myboxplot(X, AffysTissueMC$Mcor, main = "A")
>  # ALL PAIRS of the custom-made CDF
>  X <- cut(CustomsTissueMC$alSum, myXs, include.lowest = TRUE, right = TRUE)
>  myboxplot(X, CustomsTissueMC$Mcor, main = "B")
>
```

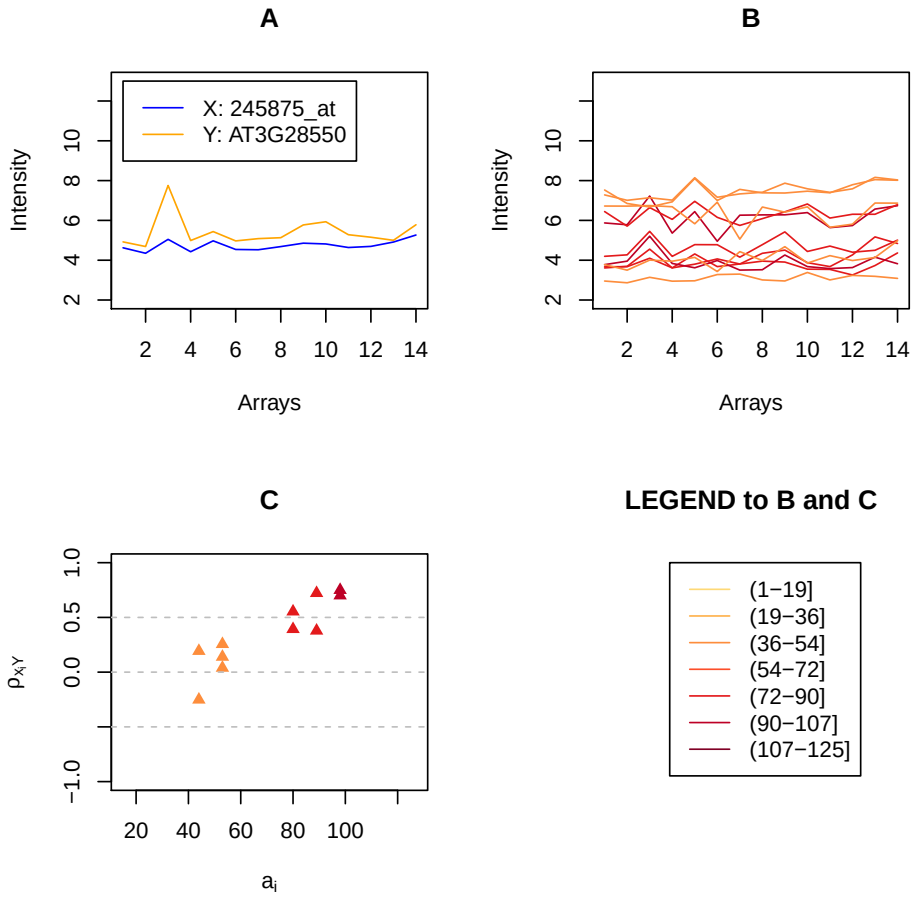


The distribution of the metacorrelations of most probe set pairs corresponds to a random distribution centered around zero. However for those strata with high off-target sensitivity score the distribution is shifted upwards. This means that within these probe sets some reporters do not correlate with the off-target, while others do, depending on their alignments score.

3 Examples

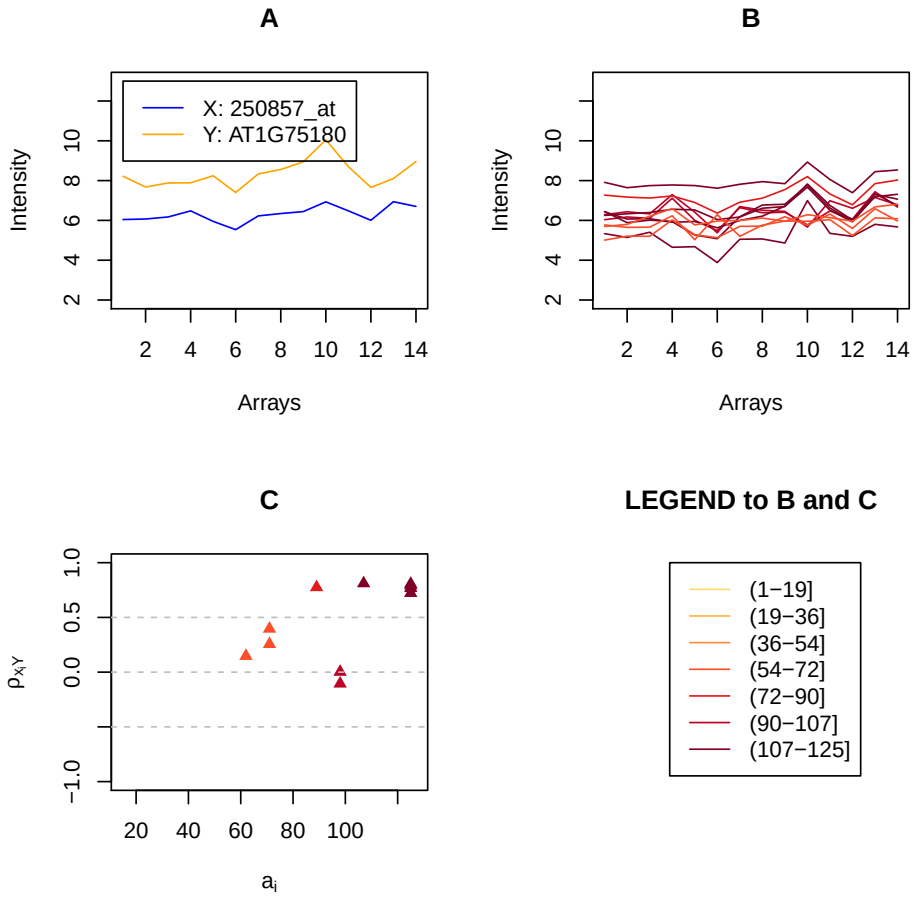
We illustrate our findings with three examples. The data for these are contained in an 'XhybExamples' -S4- class object. 'plotExample' is a method that takes 'XhybExamples' objects and plots them:
Example 1:

```
> data(ex1)
> plotExample(ex1)
```



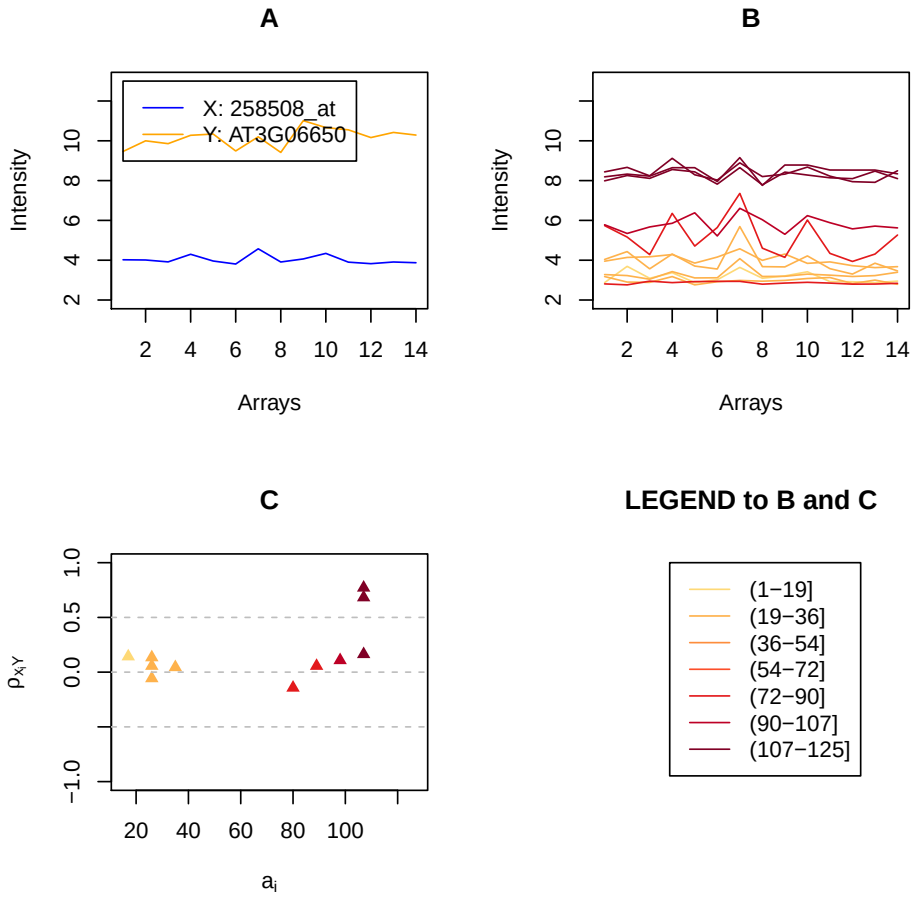
Example 2:

```
> data(ex2)
> plotExample(ex2)
```



Example 3:

```
> data(ex3)
> plotExample(ex3)
```



Plot A contains the summarized expression values of a probe set X (in blue) and an off-target gene Y (in orange) in the tissue dataset. Plot B shows the background corrected, normalized signal profiles of X's reporters. In plot C, for each reporter ρ_{X_iY}, the Pearson correlation coefficient calculated between its signal profile and that of Y (orange in A-D-G) is plotted in function of its alignment score a_{X_iY}. The colors used to plot the profiles in B and the data in C correspond to the alignment score of the particular reporter to Y's transcript and is explained in the legend.

4 Simulation

We also used a simulation to show the effect of individual reporters on summarisation.
A given gene A has a sinusoidal expression pattern over the course of 14 time points in an experiment. Plot A shows the signal profiles of the 11 reporters of this gene's probe set, with data simulated using an established error model for microarray data [4]. The 11 reporters of a probe set B in plot B show random signals without any underlying trend. Nine of the reporters of a probe set C have identical signals as nine reporters of probe set B, while the remaining two reporters cross-hybridize with the transcript of gene A (plot C). The correlation coefficients calculated on the summarised expression values obtained by different methods (median polish [5] (shown in plot D), dChip [6, 7] introduced by Li-Wong and one-step Tukey's Biweight [8] used in Affymetrix' MAS 5 software are printed to the screen.

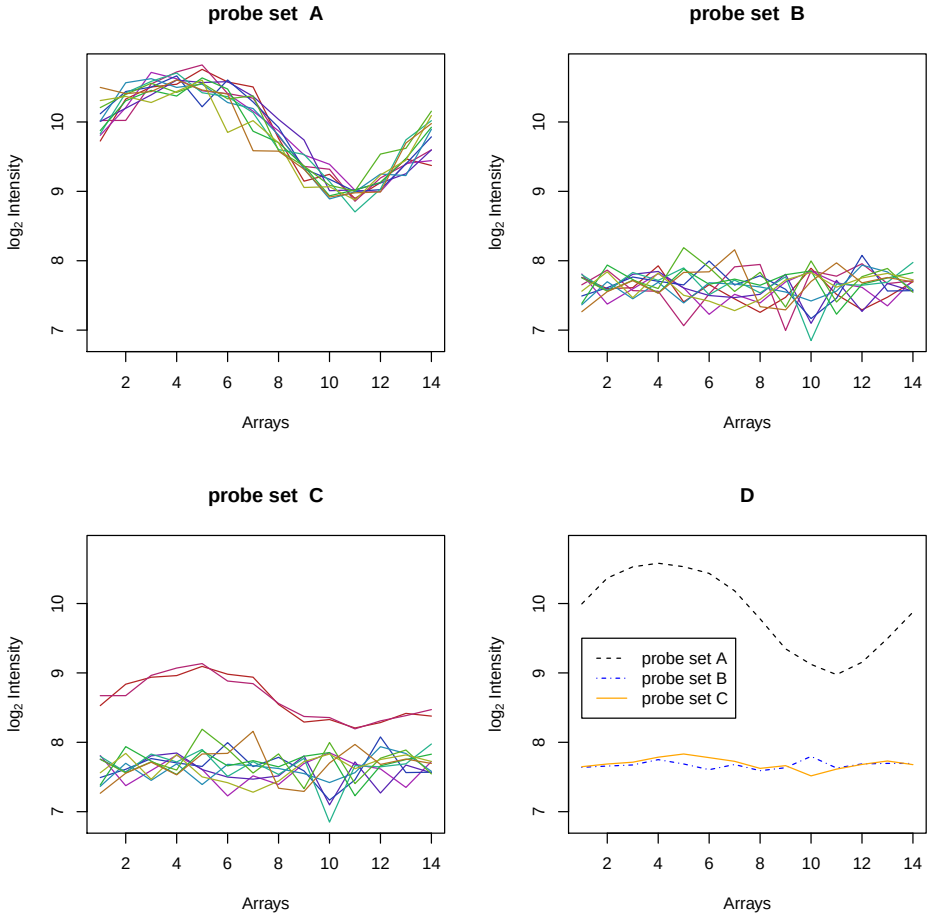
```
> runSimulation()
```

```
Plot background-corrected, normalised expression data of reporters:
   - of probe set  A  -> plot  A
   - of probe set  B  -> plot  B
   - of probe set  C  -> plot  C
Plot median-polish summarised expression data of the three probe sets -> plot D


***  Pearson correlation coefficients between expression values summarised with:
*** median polish ***
corr.: probe sets A and B:  -0.0692
corr.: probe sets A and C:  0.731

*** tuckey's biweight ***
corr.: probe sets A and B:  -0.223
corr.: probe sets A and C:  -0.19

*** dChip ***
corr.: probe sets A and B:  0.296
corr.: probe sets A and C:  0.947
```



References

[1] Schmid, M. and Davison, T.S. and Henz, S.R. and Pape, UJ and Demar, M. and Vingron, M. and Scholkopf, B. and Weigel, D. and Lohmann, JU: **A gene expression map of Arabidopsis thaliana development** *Nature Genetics* 2005, **37**(5): 501-506

[2] AtGenExpress dataset `http://www.weigelworld.org/resources/microarray/AtGenExpress/AtGE_dev_samples.pdf`

[3] Altschul, S.F. and Gish, W. and Miller, W. and Myers, E.W. and Lipman, D.J.: **Basic local alignment search tool** *Journal of Molecular Biology*, 1990, **215**: 403-410

[4] Rocke, D.M. and Durbin B. : **A Model for Measurement Error for Gene Expression Arrays** *Journal of Computational Biology* 2001, **8**(6): 557-569

[5] Irizarry, R.A. and Bolstad, B.M. and Collin, F. and Cope, L.M. and Hobbs, B. and Speed, T.P. **Summaries of Affymetrix GeneChip probe level data** *Nucleic Acids Res* 2003, **31**(4):e15

[6] Li, C. and Wong, W.H. **Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application** *Genome Biology* 2001, **2**(8): 1-11

[7] Li, C. and Wong, W.H. **Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection** *Proceedings of the National Academy of Sciences of the United Sates of America* 2001, **98**(1): 31-36

[8] Statistical Algorithms Description Document (2002) `http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf`