Reciprocal Cross in RNA-seq

Vasyl Zhabotynsky * Wei Sun Fei Zou

July 12, 2013

1 Overview

This vignette describes how to use `R/rxSeq` to perform an analysis on RNA-seq data from F1 reciprocal crosses.

```
> library(rxSeq)
```

2 Introduction

RNA sequencing (RNA-seq) not only measures total gene expression but may also measure allele-specific gene expression in diploid individuals. RNA-seq data collected from F1 reciprocal crosses in mouse can powerfully dissect strain and parent-of-origin effects on allelic imbalance of gene expression. This R package, rxSeq, implements a novel statistical approach for RNA-seq data from F1 and inbred strains. Zou *et al.* (2013) [4]

3 Citing `R/rxSeq`

When using the results from the `R/rxSeq` package, please cite:

> Zou F *et al.* (2013) 'RNA-seq analysis for F1 reciprocal crosses', *submitted*.

The article describes the methodological framework behind the `R/rxSeq` package.

4 rxSeq implementation and output

4.1 Fitting the data

4.1.1 Joint model (TReCASE model) for total read counts (TReC) and allele specific expression (ASE) counts

The TReC and ASE can be produced using the R package `R/asSeq`[1] developed by our group. A detailed pipeline of producing gene-level (or transcript-level) allele specific counts can be found in the asSeq document.

For autosomal genes, the TReCASE model requires the following input data:

*vasyl@unc.edu

- (1) a vector **index**, classifying each mouse into a cross type: for one sex or female mice AB=1,BA=2,AA=3,BB=4, and for male AB=5,BA=6,AA=7,BB=8
- (2) matrix of total counts **y** with columns representing mice, and rows - genes, F1s in first columns
- (3) matrix of allele specific counts for both alleles **n** (note, that the columns for these mice should match columns of F1 mice for total read counts)
- (4) matrix of allele specific counts for allele allele B **n0B**
- (5) a vector of log(total read counts for each mouse) - **kappas**. If not provided, the given set of total read counts **y** will be used to estimate it.
- (6) a vector of gene IDs **geneid**, if not calculated, row names are used, if they are NULL, 1:nrow(**y**) will be substituted.
- (7) **hessian** - a logical value requesting to calculate a Hessian matrix. The default value is FALSE. It is not needed for the basic analysis, however, if a subset of genes of special interest is identified, this option can be switched to TRUE, and in addition to the regular output an extra item will be added: a list of Hessian matrices for each gene.

```
> #fit trecase autosome genes:
> trecase.A.out<-proc.trecase.A(index=data.A$index,kappas=data.A$kappas,
+                 y=data.A$y[1:2,],n=data.A$n[1:2,],n0B=data.A$n0B[1:2,],
+                 geneids=data.A$geneids[1:2])

processing 2 genes
```

The following command runs the TReCASE model for Chromsome X genes, which requires two additional parameters for dealing with X-chromosome inactivations:

- (8) a vector of **tausB** - *Xce* effect for a given cross (can be estimated using overall allele specific counts imbalance for an AB cross, or literature values could be used.)
- (9) a vector **genes.switch** of geneids for which *Xce* should be switched to $1 - \mathbf{tausB}$, for example *Xist*. The default value is an Ensembl ID for a known gene with switched *Xce* effect - *Xist*: ENSMUSG00000086503.

```
> #fit trecase X chromosome genes:
> trecase.X.out<-proc.trecase.X(index=data.X$index,kappas=data.X$kappas,
+                 tausB=data.X$tausB,y=data.X$y[1:2,],n=data.X$n[1:2,],
+                 n0B=data.X$n0B[1:2,],geneids=data.X$geneids[1:2])

found 1 genes to switch Xce effect:
ENSMUSG00000086503
processing 2 genes
```

These functions return the following outputs: parameter estimates from the full models and associated p-values, and all reduced short models, followed by the list of errors:

```
> names(trecase.A.out)

 [1] "pvals"        "coef.full"    "coef.add"     "coef.poo"     "coef.dom"
 [6] "coef.same"    "coef.ase.add" "coef.sex"     "coef.sex.add" "coef.sex.poo"
[11] "coef.sex.dom" "errorlist"

> trecase.A.out$pval[,1:2]

                      pval_add  pval_poo
ENSMUSG00000055725 7.224770e-02 0.5868794
ENSMUSG00000015568 1.515202e-25 0.5460404

> names(trecase.X.out)

 [1] "pvals"        "coef.full"    "coef.add"     "coef.poo"     "coef.dom"
 [6] "coef.same"    "coef.ase.add" "coef.sex"     "coef.sex.add" "coef.dev.dom"
[11] "errorlist"

> trecase.X.out$pval[,1:2]

                      pval_add  pval_poo
ENSMUSG00000086503 1.009569e-02 0.5095113
ENSMUSG00000049775 6.992049e-05 0.9503648
```

4.1.2 TReC model for TReC only

The package also allows for fitting the data with only TReC when for a given gene, there is no enough SNP or indel information for estimating ASE.

The following function fits the TReC model for autosomal genes:

```
> #fit trec autosome genes
> trec.A.out<-proc.trec.A(index=data.A$index,kappas=data.A$kappas,
+             y=data.A$y[1:2,],geneids=data.A$geneids[1:2])

processing 2 genes

> names(trec.A.out)

 [1] "pvals"        "coef.full"    "coef.add"     "coef.poo"     "coef.dom"
 [6] "coef.sex"     "coef.sex.add" "coef.sex.poo" "coef.sex.dom" "errorlist"

> trec.A.out$pval[,1:2]

                      pval_add  pval_poo
ENSMUSG00000055725 2.334687e-02 0.5858691
ENSMUSG00000015568 5.106511e-09 0.8125544
```

The following function fits the TReC model for Chromosome X genes:

```
> #fit trec X chromosome genes
> trec.X.out<-proc.trec.X(index=data.X$index,kappas=data.X$kappas,
+             tausB=data.X$tausB,y=data.X$y[1:2,],
+             geneids=data.X$geneids[1:2])

found 1 genes to switch Xce effect:
ENSMUSG00000086503
processing 2 genes

> names(trec.X.out)

[1] "pvals"        "coef.full"    "coef.add"     "coef.poo"     "coef.dom"
[6] "coef.sex"     "coef.sex.add" "coef.dev.dom" "errorlist"

> trec.X.out$pval[,1:2]

                     pval_add  pval_poo
ENSMUSG00000086503 0.29390171 0.3133116
ENSMUSG00000049775 0.06834156 0.3018140
```

4.2 Estimating *Xce* effect for X chromosome

Both proc.trecase.X and proc.trec.X require an estimate of the *Xce* effect. In the above examples, we used an estimated value from the data in Crowley (2013) [2].

The following function estimates the *Xce* effect for any given data:

```
> get.tausB(n=data.X$n,n0B=data.X$n0B,geneids=data.X$geneids,
+           Xist.ID="ENSMUSG00000086503")

           FG_0125_F_hapG FG_0162_F_hapG FG_0163_F_hapG FG_0164_F_hapG
med.tauB        0.2266945      0.2512354      0.2888816      0.2984825
ave.tauB        0.2338611      0.2511211      0.2864014      0.2961086
all.genes       8.0000000      8.0000000      8.0000000      8.0000000
used.genes      8.0000000      8.0000000      8.0000000      8.0000000
           FG_0167_F_hapG FG_0168_F_hapG GF_0164_F_hapG GF_0165_F_hapG
med.tauB        0.2381954      0.2433292      0.3331889      0.2372911
ave.tauB        0.2354252      0.2525122      0.3477522      0.2317843
all.genes       8.0000000      8.0000000      8.0000000      8.0000000
used.genes      8.0000000      8.0000000      8.0000000      8.0000000
           GF_0166_F_hapG GF_0168_F_hapG GF_0238_F_hapG
med.tauB        0.2395825      0.3396480      0.3413311
ave.tauB        0.2529066      0.3592291      0.3367434
all.genes       8.0000000      8.0000000      8.0000000
used.genes      8.0000000      8.0000000      8.0000000
```

For genes that are known to escape X inactivation or have different *Xce* control effects, adjusted analysis can be done provided their ids are given. A default gene - *Xist* which is known to have an opposite inactivation pattern with the other X chromosome genes, we set its estimate to $1 - Xce$. We may also exclude genes with too low ASE (which is set to 50 by default) and/or with too low proportion of one of the alleles. The default value for the latter is set to 0.05 to avoid fully imprinted genes.

```
> data.X$tausB

FG_0125_F_hapG FG_0162_F_hapG FG_0163_F_hapG FG_0164_F_hapG FG_0167_F_hapG
     0.2346327      0.2520325      0.3043478      0.3000000      0.2555848
FG_0168_F_hapG GF_0164_F_hapG GF_0165_F_hapG GF_0166_F_hapG GF_0168_F_hapG
     0.2645804      0.3488372      0.2486188      0.2712934      0.3781178
GF_0238_F_hapG
     0.3611111
```

The first row of the **get.tausB** output provides a medain estimate of the *Xce* effect and the second row provides an average estimate of *Xce* effect. The two estimates are expected to be close, though median would be more stable.

```
> get.tausB(n=data.X$n,n0B=data.X$n0B,geneids=data.X$geneids,Xist.ID = "")

           FG_0125_F_hapG FG_0162_F_hapG FG_0163_F_hapG FG_0164_F_hapG
med.tauB        0.2303523      0.2534435      0.2897196      0.2986111
ave.tauB        0.3733453      0.3372797      0.3296875      0.3400289
all.genes       9.0000000      9.0000000      9.0000000      9.0000000
used.genes      9.0000000      9.0000000      9.0000000      9.0000000
           FG_0167_F_hapG FG_0168_F_hapG GF_0164_F_hapG GF_0165_F_hapG
med.tauB        0.2466844      0.2501718      0.3350168      0.2475884
ave.tauB        0.3291692      0.3398780      0.4076566      0.3148905
all.genes       9.0000000      9.0000000      9.0000000      9.0000000
used.genes      9.0000000      9.0000000      9.0000000      9.0000000
           GF_0166_F_hapG GF_0168_F_hapG GF_0238_F_hapG
med.tauB        0.2437753      0.3507246      0.3437500
ave.tauB        0.3357056      0.4229374      0.3998674
all.genes       9.0000000      9.0000000      9.0000000
used.genes      9.0000000      9.0000000      9.0000000
```

5 Simulations

RNA-seq data can be simulated using function **simRX** which requires the following input variables:

- (1)**b0f** - a female additive strain effect
- (2)**b0m** - a male additive strain effect
- (3)**b1f** - a female parent of origin effect
- (4)**b1m** - a male parent of origin effect
- (5)**beta_sex** - a sex effect
- (6)**beta_dom** - a dominance effect
- (7)**beta_k** - an effect associated with the library size kappas
- (8)**phi** - a Negative-Binomial overdispersion, default value is 1

- (9)**theta** - a Beta-Binomial overdispersion, default value is 1
- (10)**n** - a vector defining number of mice in each cross, default value is 6
- (11)**mean.base.cnt** - a target expected number of counts for the base group (with no effects), default value is 50
- (12)**range.base.cnt** - a range in which the expected number of counts for the base group will vary, default value is 60
- (13)**perc.ase** - percent of TReC that are allele-specific, default value is 35%
- (14)**n.simu** - a number of simulations, default value is 1E4
- (15)**is.X** - a flag for X chromosome genes (TRUE), default value is FALSE
- (16)**tauB** - a value describing allelic imbalance - *Xce* effect, default value is NULL, i.e. 0.5.
- (17)**seed** - a random seed, not set by default.

It produces three data matrices:

- (1)**y** - TReC
- (2)**n** - total ASE
- (3)**n0B** - allele specific counts associated with allele B

```
> dat.A<-simRX(b0f=.5,b0m=.6,b1f=.3,b1m=.4,beta_sex=.1,beta_dom=.1,n.simu=1E1)
> names(dat.A)

[1] "y"     "n"     "n0B"   "index"

> dat.X<-simRX(b0f=.5,b0m=.6,b1f=.3,b1m=.4,beta_sex=.1,beta_dom=.1,n.simu=1E1,
+ is.X=TRUE,tauB=.3)
> names(dat.X)

[1] "y"     "n"     "n0B"   "index"
```

6 References

References

[1] Wei Sun, Vasyl Zhabotynsky (2013) asSeq: A set of tools for the study of allele-specific RNA-seq data. *http://www.bios.unc.edu/ weisun/software/asSeq.pdf, to be provided on Bioconductor.*

[2] Crowley, J. J., Zhabotynsky, V., Sun, W., Huang, S., Pakatci, I. K., Kim, Y., Wang, J. R., Morgan, A., P., Calaway, J. D., Aylor, D. L., Yun, Z., Bell, T. A., Buus, R. J., Calaway, M. E., Didion, J. P., Gooch, T. J., Hansen, S. D., Robinson, N. N., Shaw, G. D., Spence, J. S., Quackenbush, C. R., Barrick, C. J., Xie, Y., Valdar, W., Lenarcic, A. B., Wang, W., Welsh, C. E., Fu, C. P., Zhang, Z., Holt, J., Guo, Z., Threadgill, D.

W., Tarantino, L. M., Miller, D., R., Zou, F., McMillan, L., Sullivan, P. F., and Pardo-Manuel de Villena, F. (2013), Pervasive allelic imbalance revealed by allele-specific gene expression in highly divergent mouse crosses., *To be submitted.*

[3] Collaborative Cross Consortium (2012), The Genome artichecture of the Collaborative Cross Mouse Genetic Reference Population, *Genetics.* **190**(2):389-401

[4] Zou, F., Sun, W., Crowley, J. J., Zhabotynsky, V., Sullivan, P. F., and Pardo Manuel de Villena, F. (2013) (2013) RNA-seq analysis for F1 reciprocal crosses., *Submitted.*.