Joing estimation of additive and parent of origin effects in trios

Vasyl Zhabotynsky * Wei Sun Kaoru Inoue Terry Magnuson
Mauro Calabrese

August 30, 2018

1 Overview

This vignette describes how to use `R/rcppreqtl` to perform an analysis on RNA-seq data from childrens of genotyped family trios

```
> library(rcppreqtl, lib.loc="/nas02/home/z/h/zhabotyn/research/package9/")
```

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 individuals from genotyped family trios can dissect strain and parent-of-origin effects on allelic imbalance of gene expression. This R package, rcppreqtl, implements a novel statistical approach for RNA-seq data collected using a new experimental design. Zhabotynsky *et al.* (2018) [[**?**]]
The package allows to fit the joint model of the total read counts for subject (assuming Negatvie-Binomial model to allow for an overdispersion) and allele specific counts (Beta-Binomial model). In the provided data example the counts are aggregated on gene level, though as long as counts are not too small, any level of generalization can be used: transcript level, exon level, etc.

3 Citing `R/rcppreqtl`

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

> Zhabotynsky, Vasyl, Sun, Wei, Inoue, Kaoru, Magnuson, Terry, Calabrese, Mauro (2018) TReCASE under family trio design: A Statistical Method for Joint Estimation of *Cis*-eQTLs and Parent-of-Origin Effects under Family Trio Design

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

*vasyl@unc.edu

4 rcppreqtl 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 model aimes to combine the total read counts (TReC) and allele specific expression (ASE) counts, estimate simultaneously additive strain effect, parent of origin effects as well as adjusts for covariates such as individual total level of expression of a subject. At the same time, model allows to reduce type II error by estimating overdispersion of the count data.
First lets create an input data object

```
> percase = 0.1
> dblcnt = 0.2
> mn = 100
> b0 = 0;b1 = 0;th = .5;dv=4;niter = 100;betas = c(3,.2,.05,.5);ss=2
> set.seed(12345)
> library(VGAM)
> library(MASS)
> phiNB = th
> phiBB = th/dv
> dep = makeXmatr(ss)
> dat = simu4(num=niter, Xmatr=dep$Xmatr, haplotype=dep$thp, totmean=mn,
+             percase=percase, dblcnt=dblcnt, phiNB=phiNB, phiBB=phiBB,
+             b0=b0, b1=b1, betas=betas)
```

For autosomal genes, the full TReCASE model can be fitted as:

```
> #fit trecase autosome genes:
> fullest = fit(subset=1:2, data=dat, traceit=FALSE)
```

Note, that it requires both TReC and ASE counts, and assumes that mice and genes match in the data matrices.

5 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.*