EDDA: Experimental Design in Differential Abundance analysis

Luo Huaien*, Li Juntao*,Chia Kuan Hui Burton, Niranjan Nagarajan

1 April 2014

1 Introduction

EDDA aids in the design of a range of common experiments such as RNA-seq, Nanostring assays, RIP-seq and Metagenomic sequencing, and enables researchers to comprehensively evaluate the impact of experimental decisions on the ability to detect differential abundance.

2 Simple example

This is a simple example for EDDA code.

```
> library("EDDA")
> data <- generateData(EntityCount=500)
>test.obj <- testDATs(data,DE.methods=c("DESeq","edgeR"),nor.methods="default")
>auc.obj  <- computeAUC(test.obj)
>plotROC(auc.obj)
>plotPRC(auc.obj)
```

3 Simulation of Count Data in EDDA

EDDA can either generate synthetic data by using the experimental characteristics specified by the user or if input is real experimental data, EDDA will first learn experimental

characterisitcs from this real data and generate synthetic data that mimics the real data accordingly.

This is an example to generate data with experimental characteristics specified by user.

```
>data <- generateData(EntityCount=1000,SampleVar="high",numDataPoints=500,
>                      ControlRep=5,FC="Unif(3,5)",perDiffAbund=0.1,minAbund=0,
>                      modelFile="BP")

>dim(data$count)
>dim(data$DiffAbundList)
>data$dataLabel
```

This is an example to generate synthetic data if real pilot data (count.txt) are available.

```
>x <- matrix(rnbinom(1000*15,size=1,mu=10), nrow=1000, ncol=15);
>x.lable=c(rep(0,10),rep(1,5))
>x[1:50,11:15] <- x[1:50,11:15]*10
>x.name=paste("g",1:1000,sep="");
>write.table(cbind(x.name,x),"count.txt",row.names =FALSE, sep ='\t')
>data <- generateData(inputCount="count.txt",inputLabel=x.lable)
>dim(data$count)
>dim(data$DiffAbundList)
>data$dataLabel
```

This is an example to generate synthetic data that captures the properties of real data but with number of replicates and entity count redefined by the user.

```
>data <- generateData(inputCount="count.txt",inputLabel=x.lable,
>                     ControlRep=10,CaseRep=10,EntityCount=3000,SimulType="auto2")

>dim(data$count)
>dim(data$DiffAbundList)
>data$dataLabel
```

The default model for data generation in EDDA is Full model, which generates synthetic counts from Negative Binomial distribution in conjunction with Multinomial distribution. EDDA also provides options to generate data directly from Negative Binomial distribution or Multinomial distribution. In cases where previously mentioned distributions are not applicable, EDDA can also generate data using Model Free approach by subsampling.

This is an example to generate synthetic data using Model Free approach by setting sampling data the same as real pilot data (count.txt).

```
>data <- generateData(SimulModel="ModelFree", inputCount="count.txt",inputLabel=x.la
>dim(data$count)
>dim(data$DiffAbundList)
>data$dataLabel
```

Another example to generate synthetic data using Model Free approach by subsampling the avaible single cell RNA-seq data.

```
>data <- generateData(SimulModel="ModelFree", inputCount="count.txt",inputLabel=x.la
>dim(data$count)
>dim(data$DiffAbundList)
>data$dataLabel
```

4 Differential abundance testing and performance evaluation

The differential abundance testing (DAT) module will allow users to easily run a panel of DATs (eg. edgeR, DESeq) on any given dataset, so as to assess the variability of results across DATs, and to compute the intersection and union of these results. Therefore, users can make more informed decisions on calling entities of differential abundance for downstream analysis.

Performance evaluation will allow users to evaluate the relative performance of various DATs based on the characteristics of their experimental setting.

This is an example to perform differential abundance testing and performance evaluation.

```
>data <- generateData(EntityCount=500)
>test.obj <- testDATs(data,DE.methods=c("DESeq","edgeR"),nor.methods="default")
>auc.obj  <- computeAUC(test.obj)
>plotROC(auc.obj)
>plotPRC(auc.obj)
```

5 References

Luo Huaien, Li Juntao,Chia Kuan Hui Burton, Paul Robson, Niranjan Nagarajan, The importance of study design for detecting differentially abundant features in high-throughput experiments, under review.