InPAS guide

Jianhong Ou, Sung Mi Park, Michael R. Green and Lihua Julie Zhu

July 28, 2015

Contents

1 Introduction

Alternative polyadenylation (APA) is one of the most important post-transcriptional regulation mechanisms which occurs in most human genes. APA in a gene can result in altered expression of the gene, which can lead pathological effect to the cells, such as uncontrolled cell cycle and growth. However, there are only limited ways to identify and quantify APA in genes, and most of them suffers from complicated process for library construction and requires large amount of RNAs.

RNA-seq has become one of the most commonly used methods for quantifying genome-wide gene expression. There are massive RNA-seq datasets available publicly such as GEO and TCGA. For this reason, we develop the InPAS algorithm for identifying APA from conventional RNA-seq data.

The workflow for InPAS is:

1. Calculate coverage from BEDGraph files
2. Predict cleavage sites
3. Estimate 3UTR usage

2 How to

To use InPAS, BSgenome and TxDb object have to be loaded before run.

```
> library(InPAS)
> library(BSgenome.Mmusculus.UCSC.mm10)
> library(TxDb.Mmusculus.UCSC.mm10.knownGene)
> path <- file.path(find.package("InPAS"), "extdata")
```

Users can prepare annotaiton by utr3Annotation with a TxDb and org annotation. The 3UTR annotation prepared by utr3Annotation includes the gaps to next annotation region.

```
> library(org.Hs.eg.db)
> samplefile <- system.file("extdata", "hg19_knownGene_sample.sqlite",
+             package="GenomicFeatures")
> txdb <- loadDb(samplefile)
> utr3.sample.anno <- utr3Annotation(txdb=txdb,
+                     orgDbSYMBOL="org.Hs.egSYMBOL")
> utr3.sample.anno
```

```
GRanges object with 177 ranges and 6 metadata columns:
                           seqnames                ranges strand   |     feature
                              <Rle>             <IRanges>  <Rle>   | <character>
    uc001bum.2_5|IQCC|utr3     chr1 [ 32673684,  32674288]      +   |     unknown
   uc001fbq.3_3|S100A9|utr3     chr1 [153333315, 153333503]      +   |     unknown
   uc001gde.2_2|LRRC52|utr3     chr1 [165533062, 165533185]      +   |     unknown
  uc001hfg.3_15|PFKFB2|utr3     chr1 [207245717, 207251162]      +   |     unknown
  uc010psc.2_17|PFKFB2|utr3     chr1 [207252365, 207254368]      +   |     unknown
                         ...      ...                   ...    ... ...         ...
    uc004dst.3_22|PHF8|CDS     chrX [ 53965591,  53965679]      -   |     unknown
    uc004dsx.3_15|PHF8|CDS     chrX [ 53969797,  53969835]      -   |     unknown
   uc004ehz.1_5|ARMCX3|CDS     chrX [100879970, 100881109]      +   |     unknown
   uc004elw.3_6|FAM199X|CDS     chrX [103434289, 103434459]      +   |     unknown
    uc004fmj.1_10|GAB3|CDS     chrX [153906455, 153906571]      -   |     unknown
                                    id          exon  transcript        gene      symbol
                           <character>   <character> <character> <character> <character>
    uc001bum.2_5|IQCC|utr3        utr3  uc001bum.2_5  uc001bum.2       55721        IQCC
   uc001fbq.3_3|S100A9|utr3        utr3  uc001fbq.3_3  uc001fbq.3        6280      S100A9
   uc001gde.2_2|LRRC52|utr3        utr3  uc001gde.2_2  uc001gde.2      440699      LRRC52
  uc001hfg.3_15|PFKFB2|utr3        utr3 uc001hfg.3_15  uc001hfg.3        5208      PFKFB2
  uc010psc.2_17|PFKFB2|utr3        utr3 uc010psc.2_17  uc010psc.2        5208      PFKFB2
                       ...         ...           ...         ...         ...         ...
    uc004dst.3_22|PHF8|CDS         CDS uc004dst.3_22  uc004dst.3       23133        PHF8
    uc004dsx.3_15|PHF8|CDS         CDS uc004dsx.3_15  uc004dsx.3       23133        PHF8
   uc004ehz.1_5|ARMCX3|CDS         CDS  uc004ehz.1_5  uc004ehz.1       51566      ARMCX3
   uc004elw.3_6|FAM199X|CDS         CDS  uc004elw.3_6  uc004elw.3      139231     FAM199X
    uc004fmj.1_10|GAB3|CDS         CDS uc004fmj.1_10  uc004fmj.1      139716        GAB3
  -------
  seqinfo: 27 sequences from hg19 genome; no seqlengths
```

Users can load mm10 and hg19 annotation from pre-prepared data. Here we loaded the prepared mm10 3UTR annotation file.

```
> ##step1 annotation
> # load from dataset
> data(utr3.mm10)
```

The coverage is calculated from BEDGraph file. The RNA-seq BAM files could be converted to BED-Graph files by bedtools genomecov tool (parameter: -bg -split).PWM and a classifier of polyA signal can be used for adjusting CP sites prediction.

```
> load(file.path(path, "polyA.rds"))
> library(cleanUpdTSeq)
> data(classifier)
> bedgraphs <- c(file.path(path, "Baf3.extract.bedgraph"),
+                file.path(path, "UM15.extract.bedgraph"))
> hugeData <- FALSE
> ##step1 Calculate coverage
> coverage <- coverageFromBedGraph(bedgraphs,
+                                  tags=c("Baf3", "UM15"),
+                                  genome=BSgenome.Mmusculus.UCSC.mm10,
+                                  hugeData=hugeData)
> ##step2 Predict cleavage sites
> CPsites <- CPsites(coverage=coverage,
+                    gp1="Baf3", gp2="UM15",
+                    genome=BSgenome.Mmusculus.UCSC.mm10,
+                    utr3=utr3.mm10,
+                    search_point_START=200,
+                    cutEnd=.2,
+                    long_coverage_threshold=3,
+                    PolyA_PWM=pwm,
+                    classifier=classifier,
+                    shift_range=100)
> ##step3 Estimate 3UTR usage
> res <- utr3UsageEstimation(CPsites=CPsites,
+                            coverage=coverage,
+                            utr3=utr3.mm10,
+                            genome=BSgenome.Mmusculus.UCSC.mm10,
+                            gp1="Baf3", gp2="UM15",
+                            short_coverage_threshold=15,
+                            long_coverage_threshold=3,
+                            adjusted.P_val.cutoff=0.05,
+                            dPDUI_cutoff=0.3,
+                            PDUI_logFC_cutoff=0.59)
> res[1]
```

```
GRanges object with 1 range and 28 metadata columns:
             seqnames                ranges strand |  transcript        gene      symbol
                <Rle>             <IRanges>  <Rle> | <character> <character> <character>
  uc009eet.1     chr6 [128846245, 128850081]      - |  uc009eet.1      232406    BC035044
             fit_value Predicted_Proximal_APA Predicted_Distal_APA        type utr3start
             <numeric>             <numeric>           <numeric> <character> <numeric>
  uc009eet.1  128.8251             128849128           128846245 novalDistal 128849981
              utr3end total.gp1  long.gp1 short.gp1 total.gp2  long.gp2 short.gp2
             <numeric> <numeric> <numeric> <numeric> <numeric> <numeric> <numeric>
  uc009eet.1 128848044  39.59496  7.874827  31.72014  19.62749  21.95978         0
             total.mean.gp1 long.mean.gp1 short.mean.gp1 total.mean.gp2 long.mean.gp2
                  <numeric>     <numeric>      <numeric>      <numeric>     <numeric>
  uc009eet.1       39.59496      7.874827       31.72014       19.62749      21.95978
             short.mean.gp2 test_status  PDUI.gp1  PDUI.gp2     dPDUI        pval
                  <numeric>   <logical> <numeric> <numeric> <numeric>   <numeric>
  uc009eet.1               0        TRUE 0.1988846         1 0.8011154 2.28143e-10
                 adjPval filterPass
               <numeric>  <logical>
  uc009eet.1 4.56286e-10       TRUE
  -------
  seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

The process described above can be done in one step.

```
> if(interactive()){
+     res <- inPAS(bedgraphs=bedgraphs, tags=c("Baf3", "UM15"),
+                genome=BSgenome.Mmusculus.UCSC.mm10,
+                utr3=utr3.mm10, gp1="Baf3", gp2="UM15",
+                txdb=TxDb.Mmusculus.UCSC.mm10.knownGene,
+                search_point_START=200,
+                short_coverage_threshold=15,
+                long_coverage_threshold=3,
+                cutStart=0, cutEnd=.2,
+                hugeData=FALSE,
+                shift_range=50,
+                PolyA_PWM=pwm, classifier=classifier)
+ }
```

InPAS can handle single group data.

```
> res <- inPAS(bedgraphs=bedgraphs[1], tags=c("Baf3"),
+                genome=BSgenome.Mmusculus.UCSC.mm10,
+                utr3=utr3.mm10, gp1="Baf3", gp2=NULL,
+                txdb=TxDb.Mmusculus.UCSC.mm10.knownGene,
+                search_point_START=200,
+                short_coverage_threshold=15,
+                long_coverage_threshold=3,
```

```
+                cutStart=0, cutEnd=.2,
+                hugeData=FALSE,
+                PolyA_PWM=pwm, classifier=classifier)
> res[1]
```

```
GRanges object with 1 range and 18 metadata columns:
             seqnames               ranges strand |  transcript        gene      symbol
                <Rle>            <IRanges>  <Rle> | <character> <character> <character>
  uc009daz.2     chr6 [98018176, 98021358]      + |  uc009daz.2       17342        Mitf
             fit_value Predicted_Proximal_APA Predicted_Distal_APA        type utr3start
             <numeric>              <numeric>            <numeric> <character> <numeric>
  uc009daz.2  18177.92               98018524             98021358      distal  98018176
              utr3end total.gp1  long.gp1 short.gp1 test_status  PDUI.gp1     dPDUI
             <numeric> <numeric> <numeric> <numeric>   <logical> <numeric> <numeric>
  uc009daz.2  98021358  275.7874  293.1735         0        TRUE         1         1
                    pval      adjPval filterPass
               <numeric>    <numeric>  <logical>
  uc009daz.2 4.678114e-05 9.356227e-05       TRUE
  -------
  seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

3 Session Info

```
> toLatex(sessionInfo())
```

- R version 3.2.1 (2015-06-18), x86_64-unknown-linux-gnu
- Locale: LC_CTYPE=en_US.UTF-8, LC_NUMERIC=C, LC_TIME=en_US.UTF-8, LC_COLLATE=C, LC_MONETARY=en_US.UTF-8, LC_MESSAGES=en_US.UTF-8, LC_PAPER=en_US.UTF-8, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=en_US.UTF-8, LC_IDENTIFICATION=C
- Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, stats4, utils
- Other packages: AnnotationDbi 1.30.1, BSgenome 1.36.3, BSgenome.Drerio.UCSC.danRer7 1.4.0, BSgenome.Mmusculus.UCSC.mm10 1.4.0, Biobase 2.28.0, BiocGenerics 0.14.0, BiocParallel 1.2.14, Biostrings 2.36.1, DBI 0.3.1, GenomeInfoDb 1.4.1, GenomicFeatures 1.20.1, GenomicRanges 1.20.5, IRanges 2.2.5, InPAS 1.0.6, RSQLite 1.0.0, S4Vectors 0.6.2, TxDb.Mmusculus.UCSC.mm10.knownGene 3.1.2, XVector 0.8.0, ade4 1.7-2, cleanUpdTSeq 1.6.1, e1071 1.6-6, org.Hs.eg.db 3.1.2, rtracklayer 1.28.6, seqinr 3.1-3
- Loaded via a namespace (and not attached): BiocStyle 1.6.0, Formula 1.2-1, GenomicAlignments 1.4.1, Gviz 1.12.1, Hmisc 3.16-0, MASS 7.3-43, RColorBrewer 1.1-2, RCurl 1.95-4.7, Rcpp 0.12.0, Rsamtools 1.20.4, VariantAnnotation 1.14.6, XML 3.98-1.3, acepack 1.3-3.3, biomaRt 2.24.0, biovizBase 1.16.0, bitops 1.0-6, class 7.3-13, cluster 2.0.3, colorspace 1.2-6, dichromat 2.0-0, digest 0.6.8, foreign 0.8-65, futile.logger 1.4.1, futile.options 1.0.0, ggplot2 1.0.1, grid 3.2.1, gridExtra 2.0.0, gtable 0.1.2, lambda.r 1.1.7,

lattice 0.20-33, latticeExtra 0.6-26, limma 3.24.14, magrittr 1.5, matrixStats 0.14.2, munsell 0.4.2, nnet 7.3-10, plyr 1.8.3, proto 0.3-10, reshape2 1.4.1, rpart 4.1-10, scales 0.2.5, splines 3.2.1, stringi 0.5-5, stringr 1.0.0, survival 2.38-3, tools 3.2.1, zlibbioc 1.14.0