groHMM Tutorial

Minho Chae Charles G. Danko W. Lee Kraus

March 14, 2014

1 Introduction

Global Nuclear Run On and Sequencing (GRO-seq) was developed for comprehensively map transcriptional activity in cells [1, 4]. GRO-seq, which provides a genome wide 'map' of the position and orientation of all transcriptionally active RNA polymerases, has become increasingly widely used in recent years because it has numerous advantages compared to alternative methods of transcriptome profiling, such as expression microarrays and RNA-seq. Among these, GRO-seq provides information on instanteneous transactional responses because it detects primary transcription, as opposed to mature, processed mRNA. In addition, because it is independent of RNA polyadenylation, processing, and stability, GRO-seq provides extensive information on the non-coding transcriptome, including primary miRNAs, lincRNAs, enhancer RNAs, and potentially additional, yet undiscovered classes of transcription occurring in cells [4, 3, 6]. Thus, GRO-seq data provides a complete and instantaneous picture of transcription, and has extensive applications in deciphering the mechanisms of transcriptional regulation.

We have recently developed several important analytical approaches which make use of GRO-seq data to address new biological questions. Our pipleline has been packaged and documented, resulting in the *groHMM* package for Bioconductor. Among the more advanced features, *groHMM* predicts the boundaries of transcriptional activity across he genome *de novo* using a two-state hidden Markov model (HMM). Our model essentially divides the genome into "transcribed" and "non-transcribed" regions in a strand specific manner [4]. We also use HMMs to identify the leading edge of Pol II at genes activated by a stimulus in GRO-seq time course data. This approach allows the genome-wide interrogation of transcription rates in cells [2].

In addtion to these advanced features, *groHMM* provides wrapper functions for counting raw reads [Carlson et al.], generating wiggle files for visualization [5], and creating metagene (averaging) plots. *groHMM* takes over all aspects of analysis after reads have been aligned to a reference genome with short-read alignment tools. Although *groHMM* is tailored towards GRO-seq data, the same functions and analytical methodologies can, in principal, be applied to a wide variety of other short read data sets since the package includes a number of easily usable and extensible functions for general short read data analysis. This guide focuses on the most common application of the package.

2 Preparation

The *groHMM* package is available in the Bioconductor and can be downloaded as follows:

```
> source("http://bioconductor.org/biocLite.R")
> biocLite("groHMM")
```

The following packages are not required to use *groHMM*, but they are used to download annotations and evaluate transcripts, and should be installed for this tutorial.

```
> biocLite("GenomicFeatures")
> biocLite("org.Hs.eg.db")
> biocLite("edgeR")
> biocLite("TxDb.Hsapiens.UCSC.hg19.knownGene")
```

3 *groHMM* Workflow

3.1 Read GRO-seq Data Files

In this tutorial we will use example data from Hah et al. [2011] (GEO accession GSE27463). This experiment was designed to assess transcriptional changes following treatment of MCF-7 cells with 17β estradiol (E2). The data include a time ccurse of GRO-seq data following treatment with E2 (*i.e.*, 0, 10, 40 and 160 min.). Two biological replicates are available for each time point. Bed files were obtained from GEO and lifted over to hg19 using the UCSC *liftOver* tool. In order to make the package size more manageable, we have included data from chromosome 7 only. *groHMM* supports parallel processing and the number of cores to use can be set using `mc.cores` option.

```
> library(groHMM)
> options(mc.cores=9)
```

groHMM uses the *GRanges* class from the *GenomicRanges* packages to represent a collection of genomic features, allowing synergy with other useful packages in Bioconductor. Most of the functions in the package take at least two arguments: 'reads' and 'features'. Reads represent the genomic coordinates of a set of mapped short reads. Features represent a set of genomic coordinates of interest such as genes, exons, or transcripts.

The example data included in this package can be loaded into R using the following commands.

```
> S0mR1 <- as(readGAlignments(system.file("extdata", "S0mR1.bam",
+                 package="groHMM")), "GRanges")
> S0mR2 <- as(readGAlignments(system.file("extdata", "S0mR2.bam",
+                 package="groHMM")), "GRanges")
> S10mR1 <- as(readGAlignments(system.file("extdata", "S10mR1.bam",
+                 package="groHMM")), "GRanges")
> S10mR2 <- as(readGAlignments(system.file("extdata", "S10mR2.bam",
+                 package="groHMM")), "GRanges")
> S40mR1 <- as(readGAlignments(system.file("extdata", "S40mR1.bam",
+                 package="groHMM")), "GRanges")
```

```
> S40mR2 <- as(readGAlignments(system.file("extdata", "S40mR2.bam",
+                 package="groHMM")), "GRanges")
> S160mR1 <- as(readGAlignments(system.file("extdata", "S160mR1.bam",
+                 package="groHMM")), "GRanges")
> S160mR2 <- as(readGAlignments(system.file("extdata", "S160mR2.bam",
+                 package="groHMM")), "GRanges")
```

3.2 Create a Wiggle File

Wiggle files are created for each strand after replicates are combined in order to visualize GRO-seq data in the UCSC genome browser. Wiggle files can be also normalized by the sequencing depth, *i.e.*, average number of reads in the dataset. `writeWiggle` function is a wrapper of `export` in *rtracklayer* for generation of wiggle/bigWig type of files.

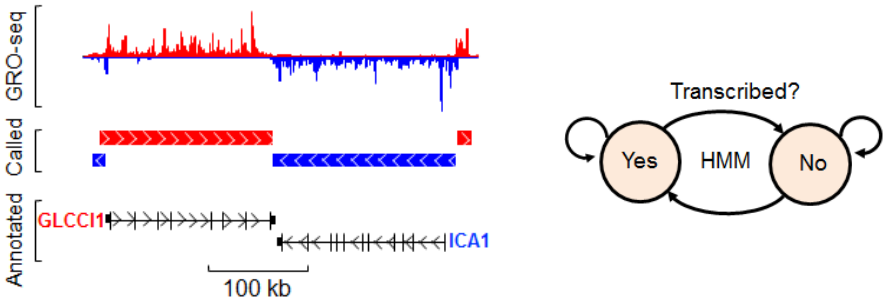
```
> # Combine replicates
> S0m <- sort(c(S0mR1, S0mR2))
> S10m <- sort(c(S10mR1, S10mR2))
> S40m <- sort(c(S40mR1, S40mR2))
> S160m <- sort(c(S160mR1, S160mR2))

> writeWiggle(reads=S0m, file="S0m_Plus.wig", fileType="wig", strand="+",
+                 reverse=FALSE)
> writeWiggle(reads=S0m, file="S0m_Minus.wig", fileType="wig", strand="-",
+                 reverse=TRUE)
> writeWiggle(reads=S0m, file="S0m_Plus.bw", fileType="BigWig", strand="+",
+                 reverse=FALSE)
> # Normalized wiggle files
> expCounts <- mean(c(NROW(S0m), NROW(S10m), NROW(S40m), NROW(S160m)))
> writeWiggle(reads=S0m, file="S0m_Plus_Norm.wig", fileType="wig", strand="+",
+             normCounts=expCounts/NROW(S0m), reverse=FALSE)
```

The resulting wiggle files can be uploaded as 'custom tracks' in the UCSC genome browser, or your visualization software of choice.

3.3 Transcript Calling

In *groHMM*, transcribed regions are detected *de novo* using a two-state hidden Markov model (HMM). The model takes GRO-seq read counts as input across the genome and divides the genome into "transcribed" and "non-transcribed" state as shown in Figure 1. First, a single read set is generated by combining all samples for each time point. This combined approach improves sensitivity for transcripts with low expression levels. Combined reads are used to train the model parameters using the Baum-Welch Expectation Maximization (EM) algorithm. Each strand is modeled separately dividing the genome into non-overlapping 50 bp windows classified as either state.

```
> Sall <- sort(c(S0m, S10m, S40m, S160m))
> hmmResult <- detectTranscripts(Sall, LtProbB=-200, UTS=5,
+                                threshold=1)
> txHMM <- hmmResult$transcripts
```



(a) GRO-seq with called transcripts

(b) 2-state HMM

Figure 1: HMM calling of GRO-seq data

```
> head(txHMM)

GRanges with 6 ranges and 2 metadata columns:
      seqnames              ranges strand |  type            ID
         <Rle>           <IRanges>  <Rle> | <Rle>   <character>
  [1]     chr7 [ 149250,  151349]      + |    tx  chr7_149250+
  [2]     chr7 [ 196150,  231399]      + |    tx  chr7_196150+
  [3]     chr7 [ 558050,  568649]      + |    tx  chr7_558050+
  [4]     chr7 [ 767600,  941699]      + |    tx  chr7_767600+
  [5]     chr7 [1126400, 1130799]      + |    tx chr7_1126400+
  [6]     chr7 [1170250, 1211449]      + |    tx chr7_1170250+
  ---
  seqlengths:
   chr7
     NA
```

The `detectTranscripts` function also uses two hold-out parameters. These parameters, specified by the arguments `LtProbB` and `UTS`, represents the log-transformed transition probability of switching from transcribed state to non-transcribed state and variance of the emission probability for reads in the non-transcribed state, respectively. Holdout parameters are used to optimize the performance of HMM predictions on known genes.

3.4 Evaluation of Transcript Calling

Predicted transcripts are evaluated by comparison to known annotations, making the assumption that GRO-seq transcripts should largely be in agreement with available annotations. Two types of error may occur, as described below. The HMM parameters are evaluated by the sum of the error rates. The procedure involves collecting a set of high-confidence reference transcripts. An annotation dataset can be constructed by downloading from the UCSC database or alteratively, pre-made TranscriptDb objects can be used if they are available in the Bioconductor. We will use the UCSC knownGene track and retrieve transcript annotations with *GenomicFeatures* [Carlson et al.] package.

```
> library(TxDb.Hsapiens.UCSC.hg19.knownGene)
> kgdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
> library(GenomicFeatures)
> # For refseq annotations:
> # rgdb <- makeTranscriptDbFromUCSC(genome="hg19", tablename="refGene")
> # saveDb(hg19RGdb, file="hg19RefGene.sqlite")
> # rgdb <- loadDb("hg19RefGene.sqlite")
> kgChr7 <- transcripts(kgdb, vals=list(tx_chrom = "chr7"),
+                       columns=c("gene_id", "tx_id", "tx_name"))
> seqlevels(kgChr7) <- seqlevelsInUse(kgChr7)
```

Because annotations do not provide precise cell type-specific expression information, overlapping transcripts must be merged into a single set, in which each annotation represents the 5′- and 3′-most boundaries of genes. Different isoforms of each gene are collpsed into one using the ENTREZID. These consensus annotations are used for the evaluation of HMM calling.

```
> # Collapse overlapping annotations
> kgConsensus <- makeConsensusAnnotations(kgChr7, keytype="gene_id",
+                 mc.cores=getOption("mc.cores"))
> library(org.Hs.eg.db)
> map <- select(org.Hs.eg.db,
+         keys=unlist(mcols(kgConsensus)$gene_id),
+         columns=c("SYMBOL"), keytype=c("ENTREZID"))
> mcols(kgConsensus)$symbol <- map$SYMBOL
> mcols(kgConsensus)$type <- "gene"
```

There are two types of error that can be evaluated when comparing predicted transcripts with annotations. (1) The number of transcripts overlapping two or more annotations, (*i.e.*, these transcripts 'run genes together') and (2) the number of annotations that overlap two or more transcripts on the same strand (*i.e.*, we say that these transcript calls 'break up a single annotation') must be determined. The optimal tuning parameters can be found by minimizing the sum of the two errors. This approach allows identification of the model that best fits the existing annotations and should more precisely predict transcripts in non-annotated parts of the genome.

```
> e <- evaluateHMMInAnnotations(txHMM, kgConsensus)
> e$eval

  runGenes brokenUp total errorRate txSize
1       79       40   119 0.0573494   1085
```

3.5 HMM Tuning

Here we demonstrate how the optimal value for each tuning parameters can be obtained by running HMM multiple times over a certain range of the parameters. Among the nine test cases, the sum of errors show minimal at case #4 and #7, as shown below. #4 is chosen over #7 for further analysis due to the lower error rate per called transcript. The variation of errors should be greater if whole chromosomes are used. Also, a larger set of the parameters might be used in practice. This tunning step takes long time, so you may skip it for quick review of the package.

```
> tune <- data.frame(LtProbB=c(rep(-100,3), rep(-200,3), rep(-300,3)),
+            UTS=rep(c(5,10,15), 3))
> Fp <- windowAnalysis(Sall, strand="+", windowSize=50)
> Fm <- windowAnalysis(Sall, strand="-", windowSize=50)
> evals <- mclapply(seq_len(9), function(x) {
+         hmm <- detectTranscripts(Fp=Fp, Fm=Fm, LtProbB=tune$LtProbB[x],
+                                  UTS=tune$UTS[x])
+         e <- evaluateHMMInAnnotations(hmm$transcripts, kgConsensus)
+         e$eval
+         }, mc.cores=getOption("mc.cores"),  mc.silent=TRUE)
> tune <- cbind(tune, do.call(rbind, evals))

> tune

  LtProbB UTS runGenes brokenUp total  errorRate txSize
1    -100   5       72      121   193 0.07082569   1735
2    -100  10       79      147   226 0.06712207   2377
3    -100  15       82      159   241 0.06272775   2852
4    -200   5       79       40   119 0.05734940   1085
5    -200  10       86       68   154 0.06332237   1442
6    -200  15       91       79   170 0.06378987   1675
7    -300   5       91       27   118 0.06233492    903
8    -300  10       96       39   135 0.06332083   1142
9    -300  15      103       45   148 0.06360120   1337

> which.min(tune$total)

[1] 7

> which.min(tune$errorRate)

[1] 4
```

To robustly compare transcripts with known genes, densities representing the frequency of transcripts can be calculated relative to their mapped gene annotations. Conceptually, the plot is divided into three distinct regions as shown in Figure 2, including upstream of known gene annotations, inside genes, and downstream of the annotated polyadenylation site.

Metrics to evaluate the degree of overlap with gene annotations focus on the region upstream of gene annotations, which provides a measure of specificity, and the region inside of genes, which provides a measure of sensitivity. The region downstream of the polyadenylation site is known to contain residual transcription [1] and consequently is not used to define quality.

The metrics are defined relative to an 'ideal' transcript caller, which takes the form of a step function (*i.e.*, red line in the plot). Our quality metrics represent the following (see the plot below for a graphical representation):

1. true transcript density (TTD) = (gene annotation area under the curve) /(max area for matched transcripts),

2. true non-trancript density (TND) = 1 - TTD,

3. false transcript density (FTD) = (5′ overhang area under the curve) /(max area for matched trasncripts),

4. false non-transcript density (FND) = 1- FTD.

Note that these quality metrics are conceptually very similar to true positive, true negative, false positive and false negative, respectively. During the comparison, only expressed annotations are used and genes either tool small or too large in size are excluded. And also size of transcripts and annotations are scaled to a smaller unit, *i.e.*, 1K for a visual representation. Here best overlapped annotations or transcripts are used for either 'run genes together' or 'break up a single annotation' type of error. Final quality metrics are represented as an AUC number.

```
> getExpressedAnnotations <- function(features, reads) {
+       fLimit <- limitToXkb(features)
+       count <- countOverlaps(fLimit, reads)
+       features <- features[count!=0,]
+       return(features[(quantile(width(features), .05) < width(features))
+          & (width(features) < quantile(width(features), .95)),])}
> conExpressed <- getExpressedAnnotations(features=kgConsensus,reads=Sall)

> td <- getTxDensity(txHMM, conExpressed, mc.cores=getOption("mc.cores"))

Run genes together:  65
Broken up a single annotation:  30
Overlaps between transcript and annotation:
Total =  568 Used for density =  446

> u <- par("usr")
> lines(c(u[1], 0, 0, 1000, 1000, u[2]), c(0,0,u[4]-.04,u[4]-.04,0,0),
+       col="red")
> legend("topright", lty=c(1,1), col=c(2,1), c("ideal", "groHMM"))
> text(c(-500,500), c(.05,.5), c("FTD", "TTD"))
> td

$FTD
[1] 0.1566222

$TTD
[1] 0.8322209

$PostTTS
[1] 0.3798408

$AUC
[1] 0.8372885
```

3.6 Repairing Transcript Calling with Annotations

Prediction of transcripts by the HMM is not perfect. Discrepancies with the annotations will occur even after the parameters are optimally tuned. Transcript calls can be 'fixed'

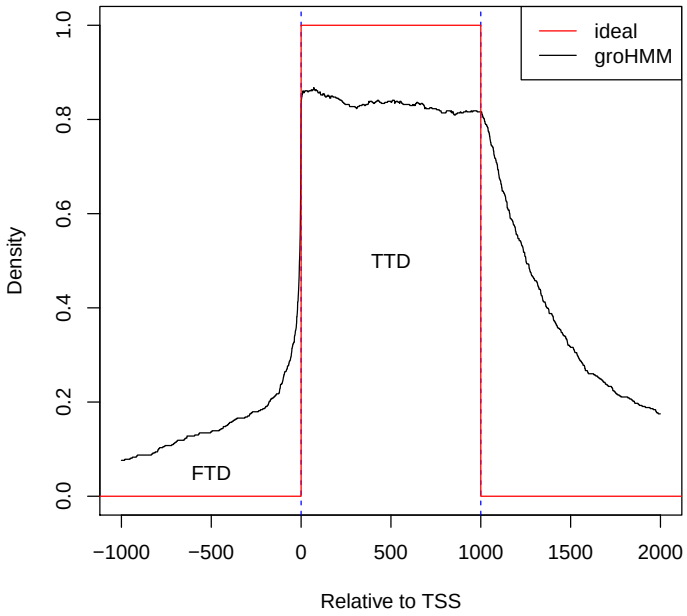


Figure 2: Transcript Density Plot

for known types of error by (1) breaking transcripts that have run together and (2) combining transcripts that have been broken. The following method will generate a final set of transcripts for further analysis.

```
> bPlus <- breakTranscriptsOnGenes(txHMM, kgConsensus, strand="+")
> bMinus <- breakTranscriptsOnGenes(txHMM, kgConsensus, strand="-")
> txBroken <- c(bPlus, bMinus)
> txFinal <- combineTranscripts(txBroken, kgConsensus)

> tdFinal <- getTxDensity(txFinal, conExpressed, mc.cores=getOption("mc.cores")
```

3.7 Differential Analysis with *edgeR*

There are several packages in Bioconductor for differential expression analysis such as *DESeq*, *baySeq*, *DEGSeq*, or *edgeR*. *edgeR* [7] is used for this demonstration. Differential expression analysis can be done using either by called transcripts or known annotations depending on the nature of your research question. The procedures are quite similar except reads are counted using the genomic locations of either called transcript or known annotations. For longer transcripts or annotated genes, we use a window of +1 to +13 kb from the transcription start site (TSS), which was chosen in order to exclude reads from RNA polymerases engaged at the promoter and to allow enough time for the elongation of newly initiated Pol II (See Hah et al., 2011). However, variations can be used, including the entire length of the called or annotated transcripts.

```
> # For called transcripts
> library(edgeR)
> txLimit <- limitToXkb(txFinal)
> ctS0mR1 <- countOverlaps(txLimit, S0mR1)
> ctS0mR2 <- countOverlaps(txLimit, S0mR2)
> ctS40mR1 <- countOverlaps(txLimit, S40mR1)
> ctS40mR2 <- countOverlaps(txLimit, S40mR2)
> pcounts <- as.matrix(data.frame(ctS0mR1, ctS0mR2, ctS40mR1, ctS40mR2))
> group <- factor(c("S0m", "S0m", "S40m", "S40m"))
> lib.size <- c(NROW(S0mR1), NROW(S0mR2), NROW(S40mR1), NROW(S40mR2))
> d <- DGEList(counts=pcounts, lib.size=lib.size, group=group)
> d <- estimateCommonDisp(d)
> et <- exactTest(d)
> de <- decideTestsDGE(et, p=0.001, adjust="fdr")
> detags <- seq_len(NROW(d))[as.logical(de)]
> # Number of transcripts regulated at 40m
> cat("up: ",sum(de==1), " down: ", sum(de==-1), "\n")

up:  95  down:  77

> plotSmear(et, de.tags=detags)
> # 2 fold up or down
> abline(h = c(-1,1), col="blue")


> # For ucsc knownGenes
> kgChr7 <- transcripts(kgdb, vals <- list(tx_chrom = "chr7"),
```

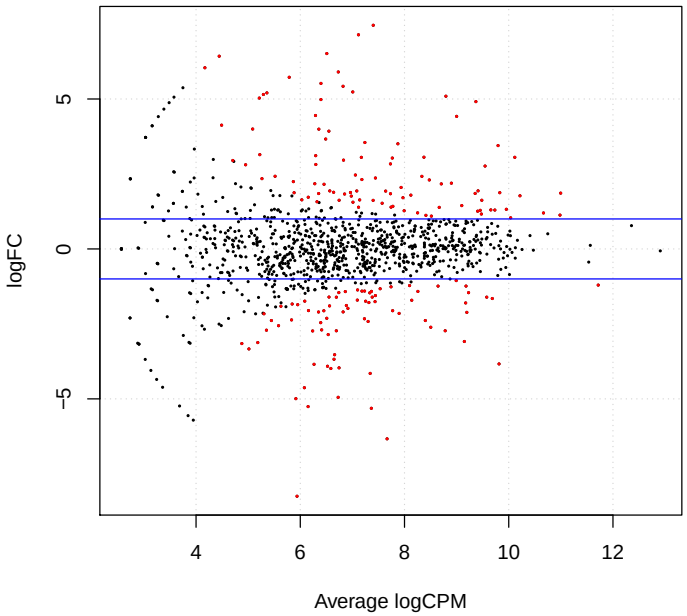


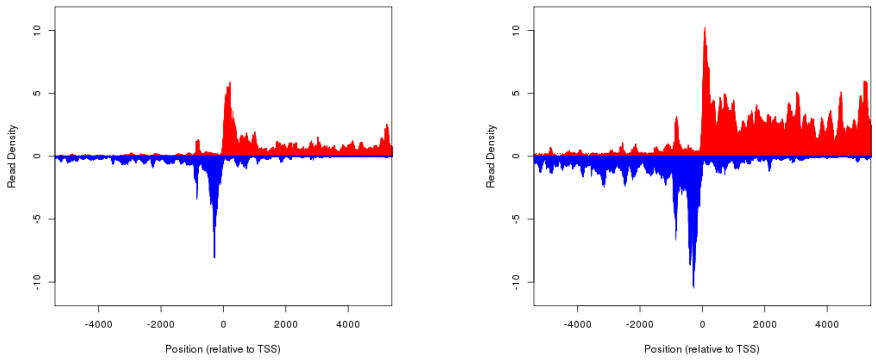
Figure 3: Transcript MAplot

```
+                         columns=c("gene_id", "tx_id", "tx_name"))
> map <- select(org.Hs.eg.db,
+          keys=unique(unlist(mcols(kgChr7)$gene_id)),
+          columns=c("SYMBOL"), keytype=c("ENTREZID"))
> missing <- elementLengths(mcols(kgChr7)[,"gene_id"]) == 0
> kgChr7 <- kgChr7[!missing,]
> inx <- match(unlist(mcols(kgChr7)$gene_id), map$ENTREZID)
> mcols(kgChr7)$symbol <- map[inx,"SYMBOL"]
> kgLimit <- limitToXkb(kgChr7)
> ctS0mR1 <- countOverlaps(kgLimit, S0mR1)
> ctS0mR2 <- countOverlaps(kgLimit, S0mR2)
> ctS40mR1 <- countOverlaps(kgLimit, S40mR1)
> ctS40mR2 <- countOverlaps(kgLimit, S40mR2)
> counts <- as.matrix(data.frame(ctS0mR1, ctS0mR2, ctS40mR1, ctS40mR2))
> group <- factor(c("S0m", "S0m", "S40m", "S40m"))
> lib.size <- c(NROW(S0mR1), NROW(S0mR2), NROW(S40mR1), NROW(S40mR2))
> d <- DGEList(counts=counts, lib.size=lib.size, group=group)
> d <- estimateCommonDisp(d)
> et <- exactTest(d)
> de <- decideTestsDGE(et, p=0.001, adjust="fdr")
> detags <- seq_len(NROW(d))[as.logical(de)]
> symbols <- mcols(kgChr7)$symbol
> # Number of unique genes regulated at 40m
> cat("up: ", NROW(unique(symbols[de==1])), "\n")

up:  58

> cat("down: ", NROW(unique(symbols[de==-1])), "\n")

down:  40

> plotSmear(et, de.tags=detags)
> abline(h = c(-1,1), col="blue")
```

3.8 Metagene Analysis

Metagenes show the distribution of reads near TSS of a set of regulated genes (or some other alignalbe genomic features of interest). It can be thought as a smoothed average of read density weighted by expression over the set of TSS. The `runMetaGene` function has option for sampling. If `TRUE`, 10% of the transcription units are sampled with replacement 1,000 times and median value at each position in the transcription unit over the samples is used for final metagene result. Using subsampling results in an image is more robust to outliers, especially when the size of sample is relatively small.

```
> upGenes <- kgChr7[de==1,]
> expReads <- mean(c(NROW(S0m), NROW(S10m), NROW(S40m), NROW(S160m)))
> # Metagene around TSS
> mg0m <- runMetaGene(features=upGenes, reads=S0m, size=100,
+                         normCounts=expReads/NROW(S0m), sampling=TRUE,
+                         mc.cores=getOption("mc.cores"))
> mg40m <- runMetaGene(features=upGenes, reads=S40m, size=100,
```



(a) 0 min. (b) 40 min.

Figure 4: Metagenes

```
+                          normCounts=expReads/NROW(S40m), sampling=TRUE,
+                                mc.cores=getOption("mc.cores"))

> plotMetaGene <- function(POS=c(-10000:+9999), mg, MIN, MAX){
+     plot(POS, mg$sense, col="red", type="h", xlim=c(-5000, 5000),
+         ylim=c(floor(MIN),ceiling(MAX)), ylab="Read Density",
+         xlab="Position (relative to TSS)")
+      points(POS, (-1*rev(mg$antisense)), col="blue", type="h")
+      abline(mean(mg$sense[5000:8000]), 0, lty="dotted")
+ }
> MAX <- max(c(mg0m$sense, mg40m$sense))
> MIN <- -1*max(c(mg0m$antisense, mg40m$antisense))
> plotMetaGene(mg=mg0m, MIN=MIN, MAX=MAX)
> plotMetaGene(mg=mg40m, MIN=MIN, MAX=MAX)
```

4 Session Info

```
> sessionInfo()

R version 3.0.2 (2013-09-25)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
```

```
attached base packages:
[1] parallel  stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
 [1] edgeR_3.4.2
 [2] limma_3.18.9
 [3] org.Hs.eg.db_2.10.1
 [4] RSQLite_0.11.4
 [5] DBI_0.2-7
 [6] TxDb.Hsapiens.UCSC.hg19.knownGene_2.10.1
 [7] GenomicFeatures_1.14.2
 [8] AnnotationDbi_1.24.0
 [9] Biobase_2.22.0
[10] Rsamtools_1.14.2
[11] Biostrings_2.30.1
[12] groHMM_0.99.0
[13] rtracklayer_1.22.0
[14] GenomicRanges_1.14.4
[15] XVector_0.2.0
[16] IRanges_1.20.6
[17] BiocGenerics_0.8.0
[18] MASS_7.3-29

loaded via a namespace (and not attached):
[1] biomaRt_2.18.0  bitops_1.0-6    BSgenome_1.30.0 RCurl_1.95-4.1
[5] stats4_3.0.2    tools_3.0.2     XML_3.98-1.1    zlibbioc_1.8.0
```

References

[Carlson et al.] M. Carlson, H. Pages, P. Aboyoun, S. Falcon, M. Morgan, D. Sarkar, and M. Lawrence. *GenomicFeatures: Tools for making and manipulating transcript centric annotations*, . R package version 1.14.2.

[Carlson et al.] Marc Carlson, Patrick Aboyoun, and Herve Pages. *An Introduction to GenomicRanges*, . R package version 1.14.4.

[1] Leighton J. Core, Joshua J. Waterfall, and John T. Lis. Nascent rna sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322:1845–1848, 2008.

[2] Charles G. Danko, Nasun Hah, Xin Luo, Andre L. Martins, Leighton Core, John T. Lis, Adam Siepel, and W. Lee Kraus. Signaling pathways differentially affect rna polymerase ii initiation, pausing and elongation rate in cells. *Molecular Cell*, 50: 212–222, 2013.

[3] Nasun Hah, Shino Murakami, Anusha Nagari, Charles G. Danko, and W. Lee Kraus. Enhancer transcripts mark active estrogen receptor binding sites. *Genome Res*, 23:1210–1223, 2009.

[4] Nasun Hah, Charales G. Danko, Leighton Core, Joshua J. Waterfall, Adam Siepel, John T. Lis, and W. Lee Kraus. A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell*, 145:622–634, 2011.

[5] Michael Lawrence, Robert Gentleman, and Vincent Carey. rtracklayer: an r package for interfacing with genome browsers. *Bioinformatics*, 25:1841–1842, 2009.

[6] Xin Luo, Minho Chae, Raga Krishnakumar, Charles G. Danko, and W. Lee Kraus. Dynamic reorganization of the ac16 cardiomyocyte transcriptome in response to tnfÎś signaling revealed by integrated genomic analyses. *BMC Genomics*, 24:155, 2014.

[7] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139–140, 2010.