Single-embryo RNA-sequencing for continuous and sex-specific gene expression analysis on Drosophila

J. Eduardo Pérez-Mojica, Lennart Enders, Kin H. Lau, and Adelheid Lempradl

June 20, 2023

Contents

```
# this chunk is just to keep the _files directory even when we turn off cacheing
```

```
# save start time for script
start_tm <- Sys.time()
start_tm
```

```
## [1] "2023-06-20 10:47:14 EDT"
```

```
outdir <- "./out_files/"

dir.create(outdir, recursive=TRUE)
```

```
## Warning in dir.create(outdir, recursive = TRUE): './out_files' already exists
```

1 Packages loaded

```
library(RaceID)
library(DESeq2)
```

```
## Loading required package: S4Vectors

## Loading required package: stats4

## Loading required package: BiocGenerics

##
## Attaching package: 'BiocGenerics'
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, aperm, append, as.data.frame, basename, cbind,
##     colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
##     get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff, sort,
##     table, tapply, union, unique, unsplit, which.max, which.min
```

```
##
## Attaching package: 'S4Vectors'
```

```
## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname
```

```
## Loading required package: IRanges
```

```
## Loading required package: GenomicRanges
```

```
## Loading required package: GenomeInfoDb
```

```
## Loading required package: SummarizedExperiment
```

```
## Loading required package: MatrixGenerics
```

```
## Loading required package: matrixStats
```

```
##
## Attaching package: 'MatrixGenerics'
```

```
## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars
```

```
## Loading required package: Biobase
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
##
## Attaching package: 'Biobase'

## The following object is masked from 'package:MatrixGenerics':
##
##     rowMedians

## The following objects are masked from 'package:matrixStats':
##
##     anyMissing, rowMedians
```

```
library(splineTimeR)
```

```
## Loading required package: igraph

##
## Attaching package: 'igraph'

## The following object is masked from 'package:GenomicRanges':
##
##     union

## The following object is masked from 'package:IRanges':
##
##     union

## The following object is masked from 'package:S4Vectors':
##
##     union

## The following objects are masked from 'package:BiocGenerics':
##
##     normalize, path, union

## The following objects are masked from 'package:stats':
##
##     decompose, spectrum

## The following object is masked from 'package:base':
##
##     union

## Loading required package: limma

##
## Attaching package: 'limma'

## The following object is masked from 'package:DESeq2':
##
##     plotMA

## The following object is masked from 'package:BiocGenerics':
##
##     plotMA

## Loading required package: GSEABase

## Loading required package: annotate

## Loading required package: AnnotationDbi

## Loading required package: XML

## Loading required package: graph
```

```
##
## Attaching package: 'graph'

## The following object is masked from 'package:XML':
##
##     addNode

## The following objects are masked from 'package:igraph':
##
##     degree, edges, intersection

## Loading required package: gtools

##
## Attaching package: 'gtools'

## The following object is masked from 'package:igraph':
##
##     permute

## Loading required package: splines

## Loading required package: GeneNet

## Loading required package: corpcor

## Loading required package: longitudinal

## Loading required package: fdrtool

## Loading required package: FIs
```

2 Continuous transcriptome analysis part 1; Identification of unfertilized eggs and embryos older than 3 h

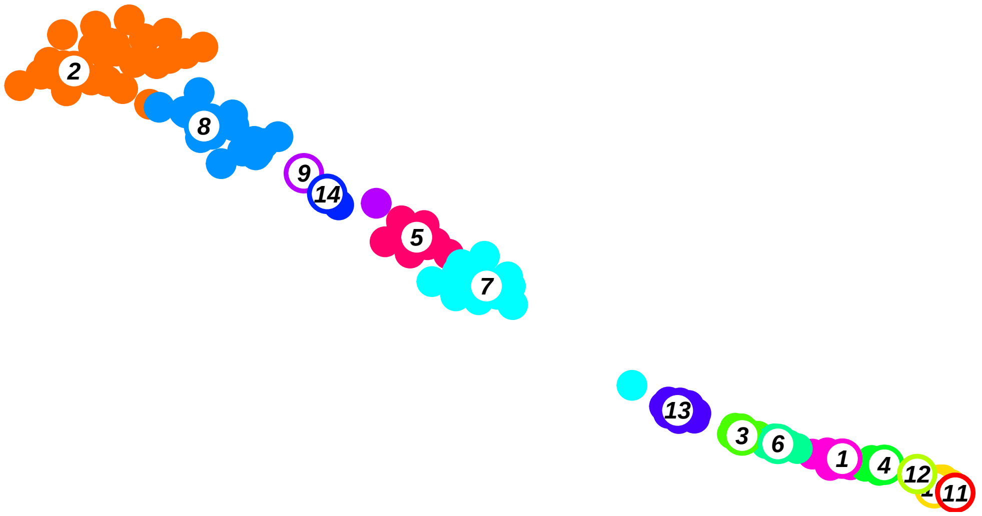
```
library1 <-read.csv("GSM6599295_Sample1.STARsolo_raw.counts.txt", sep="\t",
                    header=TRUE, row.names = 1)
library2 <-read.csv("GSM6599296_Sample2.STARsolo_raw.counts.txt", sep="\t",
                    header=TRUE, row.names = 1)
data <- cbind(library1[,1:96], library2[,97:192])


sc <- SCseq(data)
sc <- filterdata(sc, minexpr = 3, minnumber = 5, LBatch = NULL, mintotal=250000)
sc <- compdist(sc,metric="spearman", FSelect = FALSE,knn = NULL,alpha = 3)
sc <- clustexp(sc, rseed = 12345, samp = 1000 , FUNcluster = "kmedoids")
```

```
## Clustering k = 1,2,..., K.max (= 30): ..
## k = 1 k = 2 k = 3 k = 4 k = 5 k = 6 k = 7 k = 8 k = 9 k = 10 k = 11 k = 12 k = 13 k = 14 k = 15 k =
## done.
## subset 1
## subset 2
## subset 3
## subset 4
## subset 5
## subset 6
## subset 7
## subset 8
## subset 9
## subset 10
## subset 11
## subset 12
## subset 13
## subset 14
## subset 15
## subset 16
## subset 17
## subset 18
## subset 19
## subset 20
## subset 21
## subset 22
## subset 23
## subset 24
## subset 25
## subset 26
## subset 27
## subset 28
## subset 29
## subset 30
## subset 31
## subset 32
## subset 33
## subset 34
## subset 35
## subset 36
## subset 37
## subset 38
## subset 39
## subset 40
## subset 41
## subset 42
## subset 43
## subset 44
## subset 45
## subset 46
## subset 47
## subset 48
## subset 49
## subset 50
```

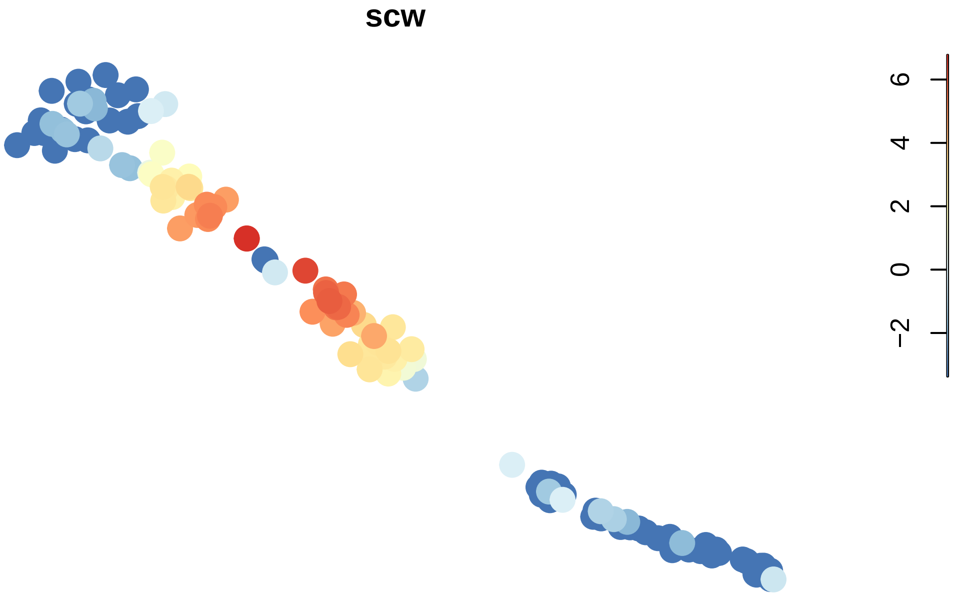
```
sc <- findoutliers(sc, probthr = 0.001, outlg = 3, outminc = 5)
```

```
## find outliers in cluster 1 find outliers in cluster 2 find outliers in cluster 3 find outliers in clu
##
## determine final clustering partition 1 determine final clustering partition 2 determine final cluste
```

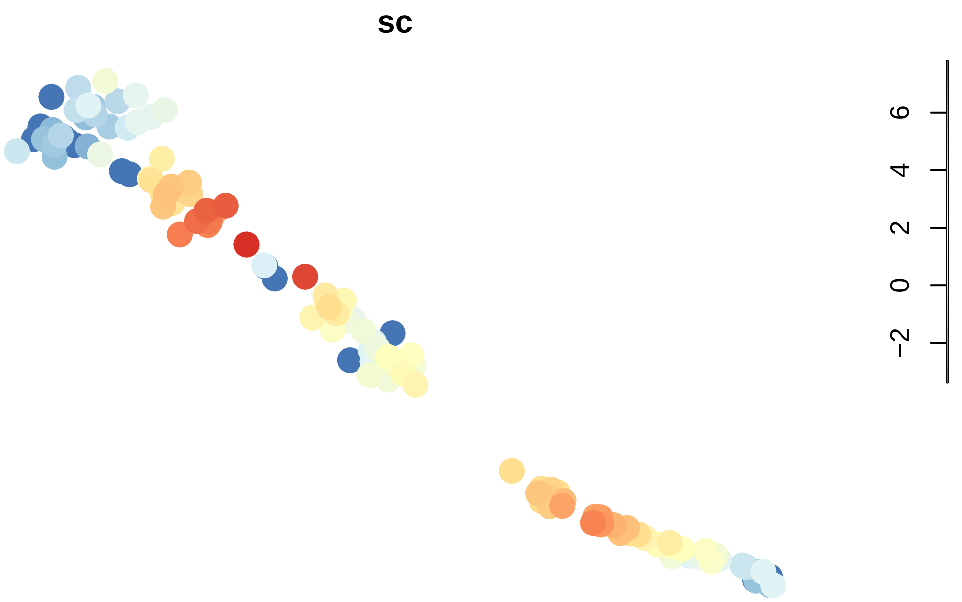
```
#pdf(file = "./out_files/01tsne_maps.pdf",width = 11, height = 7.5)
sc <- comptsne(sc,perplexity = 16, rseed = 420)
plotmap(sc,cex=3)
```



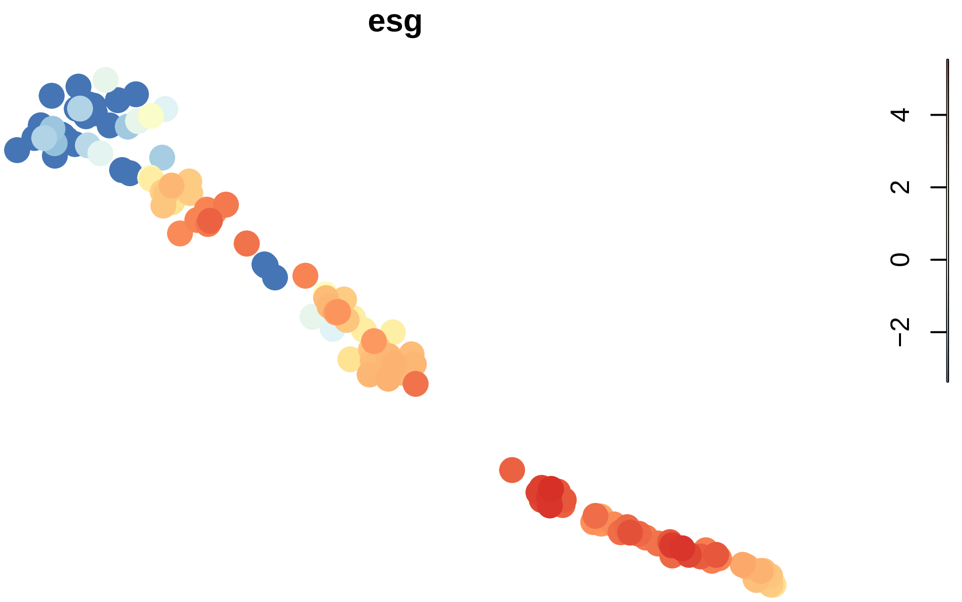
```
plotexpmap(sc, g="scw", n="scw", logsc = TRUE, cex = 3)
```



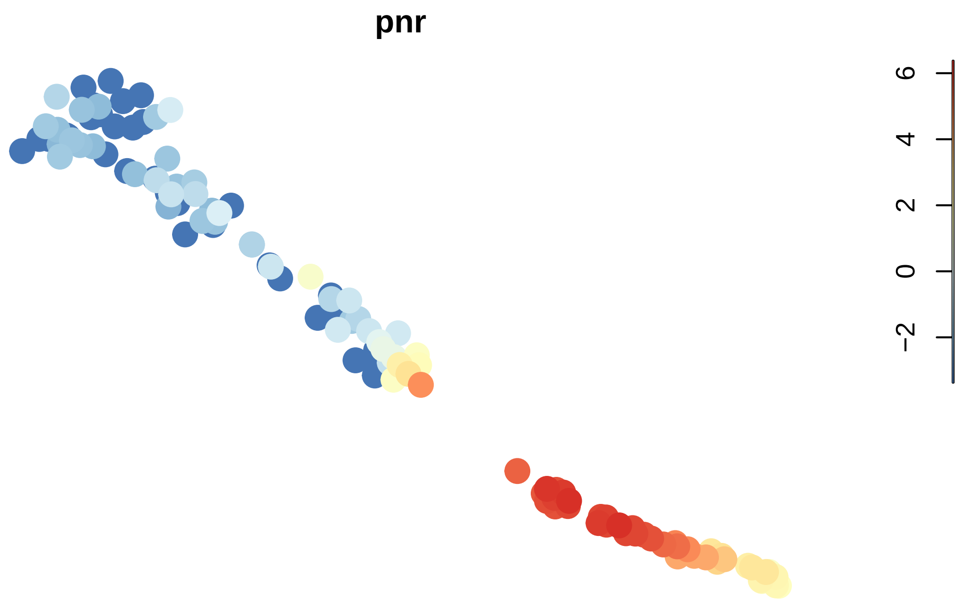
```
plotexpmap(sc, g="sc",  n="sc",  logsc = TRUE, cex = 3)
```



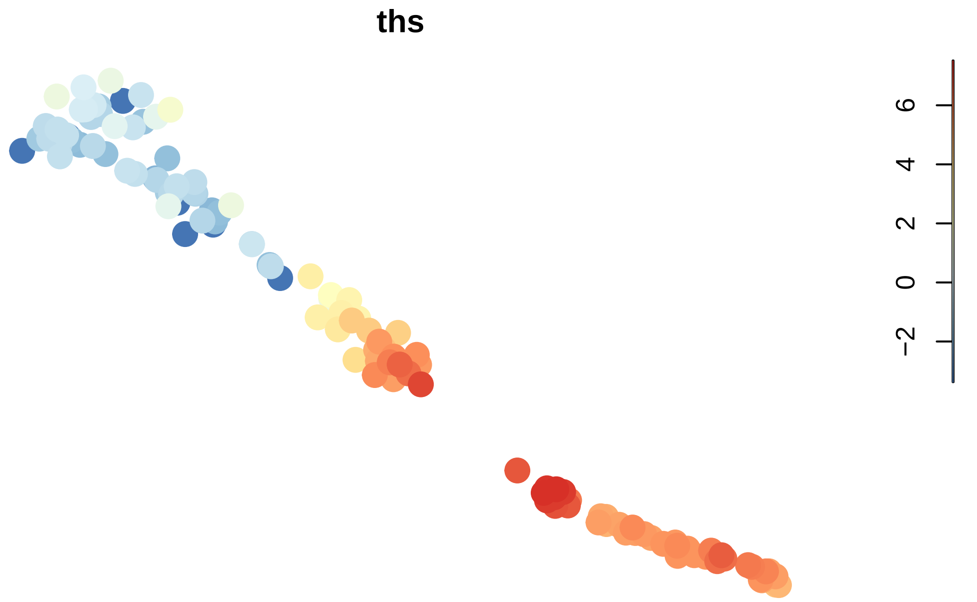
```
plotexpmap(sc, g="esg", n="esg", logsc = TRUE, cex = 3)
```



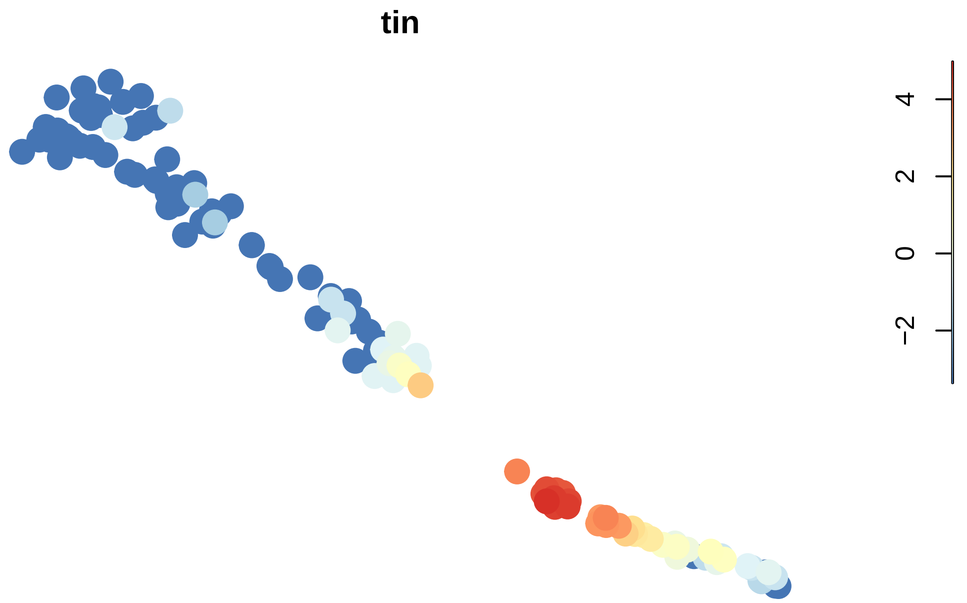
```
plotexpmap(sc, g="pnr", n="pnr", logsc = TRUE, cex = 3)
```



```
plotexpmap(sc, g="ths", n="ths", logsc = TRUE, cex = 3)
```

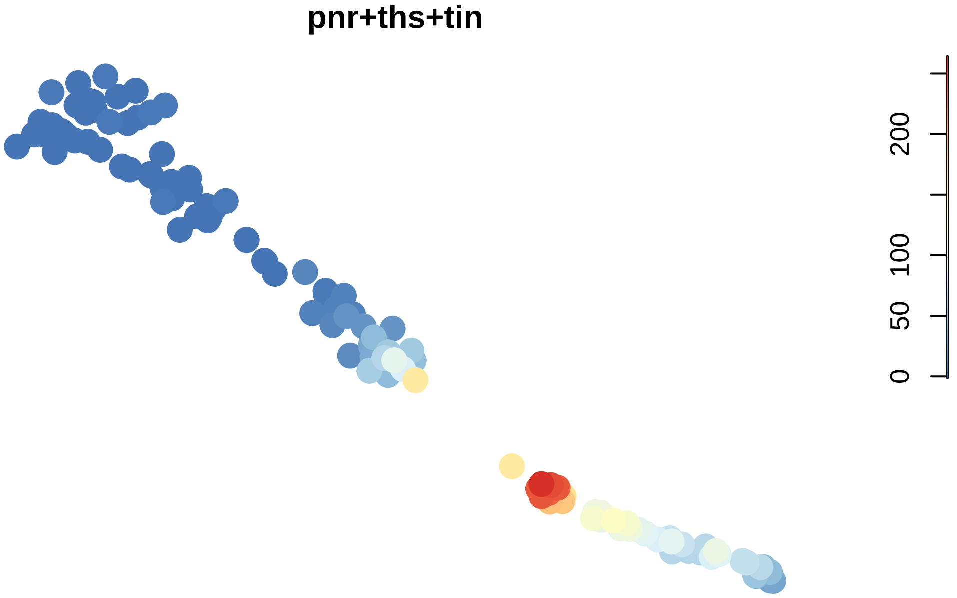


```
plotexpmap(sc, g="tin", n="tin", logsc = TRUE, cex = 3)
```



```
plotlabelsmap(sc, cex = 0.2)
```

```
#dev.off()

#pdf(file = "./out_files/02pnr_ths_tin.pdf",width = 11, height = 7.5)
plotexpmap(sc, g=c("pnr","ths","tin"), n="pnr+ths+tin", logsc = FALSE, cex = 3)
```



```
#dev.off()

write.csv(sc@cpart, file = "./out_files/03sampleid_by_cluster.csv")
```

3 Continuous transcriptome analysis part 2; Generation of a pseudo-time

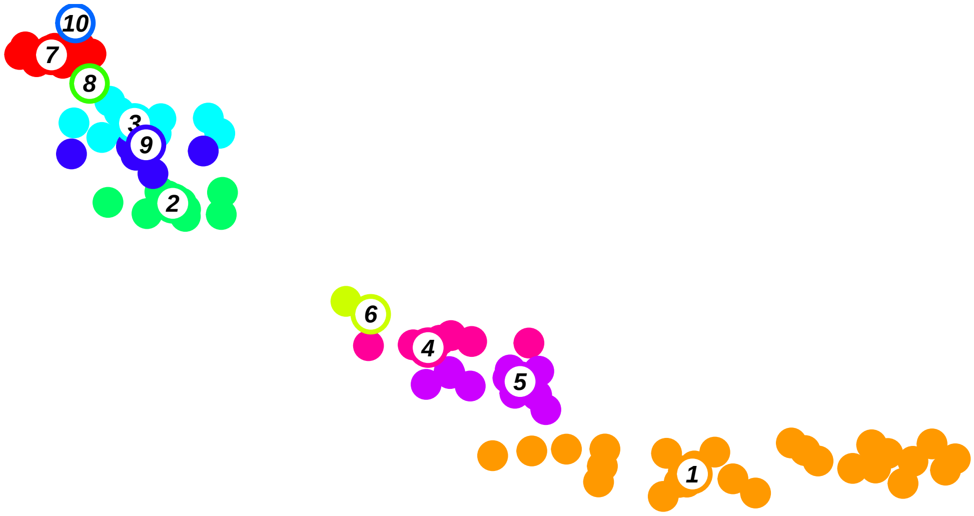
```
# The first 5 samples below are unfertilized eggs and the rest older embryos
exclude <- c("X75","X132","X136","X141","X166","X1",
              "X4","X9","X10","X25","X30","X31","X32",
              "X46","X50","X51","X55","X72","X77","X80",
              "X85","X94","X95","X102","X108","X124",
              "X125","X127","X130","X153","X158","X171",
              "185","X188")
embryos_3h <- data[,!(names(data) %in% exclude)]
sc <- SCseq(embryos_3h)
sc <- filterdata(sc, minexpr = 3, minnumber = 5, LBatch = NULL, mintotal=250000)
sc <- compdist(sc,metric="spearman", FSelect = FALSE,knn = NULL,alpha = 3)
sc <- clustexp(sc, rseed = 12345, samp = 1000 , FUNcluster = "kmedoids")
```

```
## Clustering k = 1,2,..., K.max (= 30): ..
## k = 1 k = 2 k = 3 k = 4 k = 5 k = 6 k = 7 k = 8 k = 9 k = 10 k = 11 k = 12 k = 13 k = 14 k = 15 k =
## done.
## subset 1
## subset 2
## subset 3
## subset 4
## subset 5
## subset 6
## subset 7
## subset 8
## subset 9
## subset 10
## subset 11
## subset 12
## subset 13
## subset 14
## subset 15
## subset 16
## subset 17
## subset 18
## subset 19
## subset 20
## subset 21
## subset 22
## subset 23
## subset 24
## subset 25
## subset 26
## subset 27
## subset 28
## subset 29
## subset 30
## subset 31
## subset 32
## subset 33
## subset 34
## subset 35
## subset 36
## subset 37
## subset 38
## subset 39
## subset 40
## subset 41
## subset 42
## subset 43
## subset 44
## subset 45
## subset 46
## subset 47
## subset 48
## subset 49
## subset 50
```

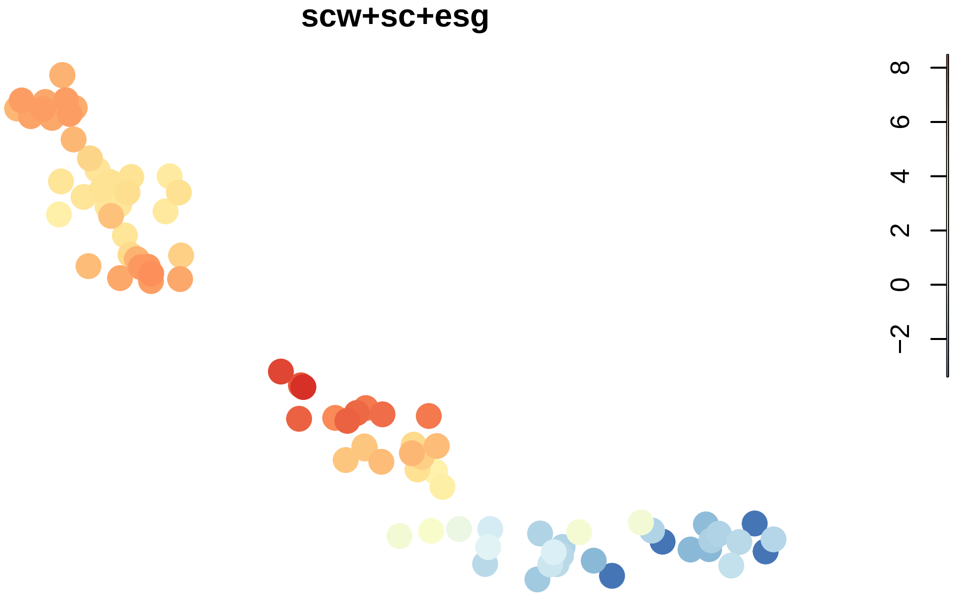
```
sc <- findoutliers(sc, probthr = 0.001, outlg = 3, outminc = 5)
```

```
## find outliers in cluster 1 find outliers in cluster 2 find outliers in cluster 3 find outliers in clu
##
## determine final clustering partition 1 determine final clustering partition 2 determine final cluste
```

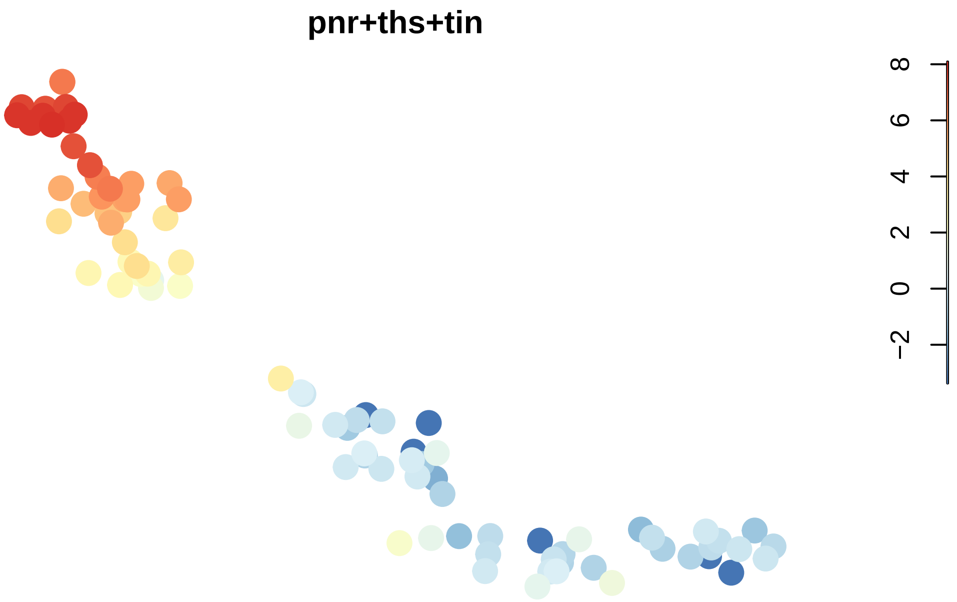
```
#pdf(file = "./out_files/04tsne_maps_3h_embryos.pdf",width = 11, height = 7.5)
sc <- comptsne(sc,perplexity = 10, rseed = 420)
plotmap(sc,cex=3)
```



```
plotexpmap(sc, g=c("scw","sc","esg"), n="scw+sc+esg", logsc = TRUE, cex = 3)
```



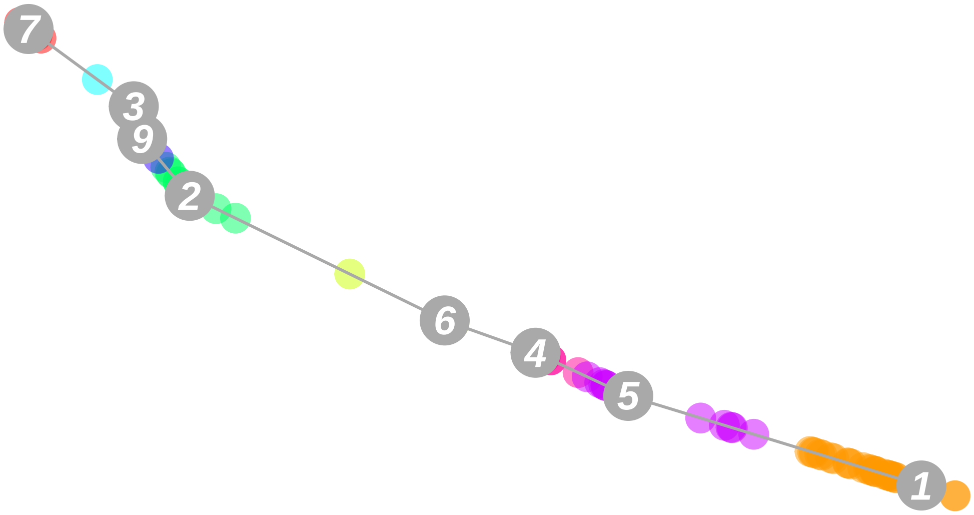
```
plotexpmap(sc, g=c("pnr","ths","tin"), n="pnr+ths+tin", logsc = TRUE, cex = 3)
```



```
#dev.off()

ltr <- Ltree(sc)
ltr <- compentropy(ltr)
ltr <- projcells(ltr,cthr=2,nmode=T,knn=3)
ltr <- lineagegraph(ltr)
```

```
## Building tree:  1 Building tree:  2 Building tree:  3 Building tree:  4 Building tree:  5 Building t
```

```
ltr <- comppvalue(ltr,pthr=0.05, sensitive = T)

#pdf(file = "./out_files/05intercluster_links.pdf", width = 11, height = 7.5)
plotspantree(ltr,cex = 3, projections = T)
```



```
#dev.off()

n <- cellsfromtree(ltr,c(1,5,4,6,2,9,3,7))
list_pseudotime <- row.names(as.data.frame(ltr@sc@cpart[n$f]))
norm_counts <- as.matrix(getfdata(sc))
norm_counts <- norm_counts[, list_pseudotime]
write.csv(norm_counts, file = "./out_files/06normalized_counts_by_pseudotime.csv")
```

4 Differential expression analysis between two clusters (RaceID & DESeq2)

```
A <- names(sc@cpart)[sc@cpart %in% c(1)]
B <- names(sc@cpart)[sc@cpart %in% c(4)]
x <- diffexpnb(sc@expdata,n=c(A,B),DESeq = TRUE, A=A, B=B, method = "per-condition")
```

```
## converting counts to integer mode
```

```
## estimating size factors
```

```
## estimating dispersions
```

```
## gene-wise dispersion estimates
```

```
## mean-dispersion relationship
```

```
## final dispersion estimates
```

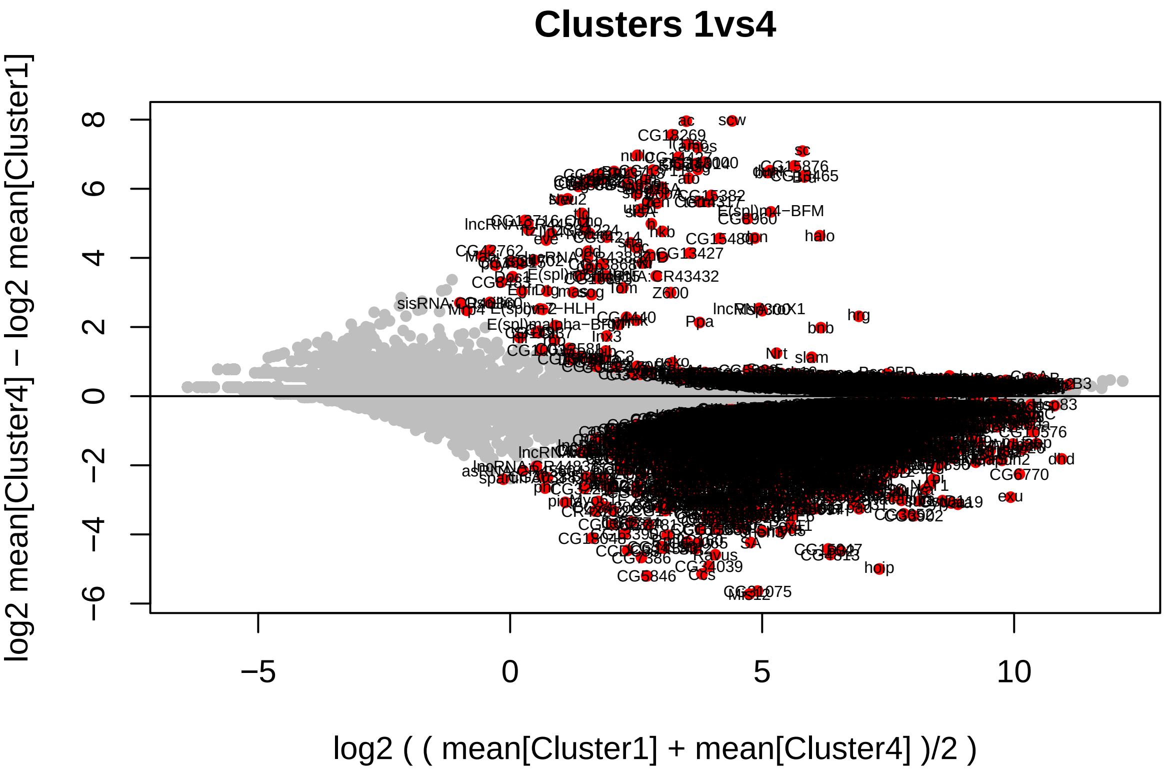
```
## fitting model and testing
```

```
## -- replacing outliers and refitting for 3 genes
## -- DESeq argument 'minReplicatesForReplace' = 7
## -- original counts are preserved in counts(dds)
```

```
## estimating dispersions
```

```
## fitting model and testing
```

```
plotdiffgenesnb(x,pthr=.05,lthr=,mthr=-1, Aname="Cluster1", Bname="Cluster4",
                show_names=TRUE, padj=TRUE, main="Clusters 1vs4")
```



```
write.table(x$res, "./out_files/05results_1vs4.xls", col.names=TRUE, sep="\t",
            quote=FALSE)
```

5 Sex-specific gene expression analysis (SplineTimeR)

```
norm.counts <-read.csv("./out_files/06normalized_counts_by_pseudotime.csv",
                       sep=",", header=TRUE, row.names = 1)
m.data <- read.csv("metadata.csv", sep=",", header=TRUE)
sample_list <- m.data[,1]
norm.counts <- norm.counts[,sample_list]
row.names(m.data) <- m.data[,1]

phenoData <- new("AnnotatedDataFrame", data=m.data)
minimalSet <- ExpressionSet(assayData=as.matrix(norm.counts), phenoData = phenoData)
diffExprs <- splineDiffExprs(eSetObject = minimalSet, df = 7, cutoff.adj.pVal = 0.01,
                             reference = "MALE", intercept = TRUE)
```

```
## --------------------------------------------------
## Differential analysis done for df = 7 and adj.P.Val <= 0.01
## Number of differentially expressed genes:  120
```

```
write.csv(diffExprs, file = "./out_files/07diffExp_males_females.csv")
```

6 SessionInfo

```
sessionInfo()
```

```
## R version 4.2.2 (2022-10-31)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur ... 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] splines   stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] splineTimeR_1.26.0          FIs_1.26.0
##  [3] GeneNet_1.2.16              fdrtool_1.2.17
##  [5] longitudinal_1.1.13         corpcor_1.6.10
##  [7] gtools_3.9.4                GSEABase_1.60.0
##  [9] graph_1.76.0                annotate_1.76.0
## [11] XML_3.99-0.13               AnnotationDbi_1.60.2
## [13] limma_3.54.2                igraph_1.4.1
## [15] DESeq2_1.38.3               SummarizedExperiment_1.28.0
## [17] Biobase_2.58.0              MatrixGenerics_1.10.0
## [19] matrixStats_0.63.0          GenomicRanges_1.50.2
## [21] GenomeInfoDb_1.34.9         IRanges_2.32.0
## [23] S4Vectors_0.36.2            BiocGenerics_0.44.0
## [25] RaceID_0.3.0
##
## loaded via a namespace (and not attached):
##  [1] Rtsne_0.16             colorspace_2.1-0       class_7.3-21
##  [4] modeltools_0.2-23      mclust_6.0.0           som_0.3-5.1
##  [7] XVector_0.38.0         rstudioapi_0.14        leiden_0.4.3
## [10] flexmix_2.3-18         bit64_4.0.5            RSpectra_0.16-1
## [13] fansi_1.0.4            codetools_0.2-19       robustbase_0.95-0
## [16] cachem_1.0.7           geneplotter_1.76.0     knitr_1.42
## [19] jsonlite_1.8.4         umap_0.2.10.0          ica_1.0-3
## [22] kernlab_0.9-32         cluster_2.1.4          png_0.1-8
## [25] pheatmap_1.0.12        compiler_4.2.2         httr_1.4.5
## [28] Matrix_1.5-3           fastmap_1.1.1          cli_3.6.0
## [31] htmltools_0.5.4        tools_4.2.2            gtable_0.3.1
## [34] glue_1.6.2             GenomeInfoDbData_1.2.9 dplyr_1.1.2
## [37] Rcpp_1.0.10            vctrs_0.6.3            Biostrings_2.66.0
## [40] nlme_3.1-162           fpc_2.2-10             xfun_0.39
## [43] lifecycle_1.0.3        irlba_2.3.5.1          princurve_2.1.6
## [46] DEoptimR_1.0-11        zlibbioc_1.44.0        MASS_7.3-58.3
## [49] scales_1.2.1           parallel_4.2.2         RColorBrewer_1.1-3
## [52] FateID_0.2.2           yaml_2.3.7             memoise_2.0.1
## [55] reticulate_1.28        ggplot2_3.4.2          RSQLite_2.3.0
## [58] highr_0.10             randomForest_4.7-1.1   harmony_0.1.1
## [61] permute_0.9-7          BiocParallel_1.32.5    prabclus_2.3-2
## [64] rlang_1.1.0            pkgconfig_2.0.3        bitops_1.0-7
## [67] evaluate_0.20          lattice_0.20-45        cowplot_1.1.1
## [70] bit_4.0.5              tidyselect_1.2.0       magrittr_2.0.3
## [73] R6_2.5.1               generics_0.1.3         DelayedArray_0.24.0
## [76] DBI_1.1.3              pillar_1.9.0           mgcv_1.8-42
## [79] nnet_7.3-18            KEGGREST_1.38.0        RCurl_1.98-1.10
## [82] tibble_3.2.0           crayon_1.5.2           utf8_1.2.3
## [85] runner_0.4.2           rmarkdown_2.20         locfit_1.5-9.7
## [88] grid_4.2.2             blob_1.2.3             FNN_1.1.3.1
## [91] vegan_2.6-4            diptest_0.76-0         digest_0.6.31
## [94] xtable_1.8-4           coop_0.6-3             openssl_2.0.6
## [97] munsell_0.5.0          askpass_1.1            quadprog_1.5-8
```

7 Time

```
# output time taken to run script
end_tm <- Sys.time()
end_tm
```

```
## [1] "2023-06-20 10:48:51 EDT"
```

```
end_tm - start_tm
```

```
## Time difference of 1.619979 mins
```