Tiphaine Martin * Idil Yet † Pei-Chien Tsai † Jordana T. Bell §

Edited: September 2014; Compiled: February 15, 2015

1 Citation

```
citation(package='coMET')
##
## To cite 'coMET' in publications use:
##
     Martin, T., Erte, I, Tsai, P-C, Bell, J.T. coMET: an R plotting package to
##
##
     visualize regional plots of epigenome-wide association scan results QG14, 2014
##
## A BibTeX entry for LaTeX users is
##
##
     @Article{,
##
       title = {coMET: an R plotting package to visualize regional plots of epigenome-wide associ
##
       author = {{Martin} and {T.C.} and {Erte} and {I.} and {Tsai} and {P-C.} and {Bell} and {J.
       journal = {QG14},
##
##
       year = \{2014\},\
       month = \{May\},
##
       url = {http://quantgen.soc.srcf.net/qg14/},
##
##
```

^{*}tiphaine.martin@kcl.ac.uk

[†]idil.yet@kcl.ac.uk

[‡]peichien.tsai@kcl.ac.uk

[§]jordana.bell@kcl.ac.uk

Contents

1	Citation	1
2	Introduction	3
3	Usage	3
4	Files formats 4.1 Format of info file (mandatory)	4 4 5 6 6
5	Creating a plot like the webservice: comet.web	8
6	Creating a plot with the generic function: comet 6.1 coMET plot: pvalue plot, annotation tracks, and correlation matrix 6.2 coMET plot: annotation tracks and correlation matrix	8 9 10
7	Extra information about annotation tracks 7.1 Genes and transcripts from ENSEMBL and UCSC	12 12 13 13
8	coMET: Shiny web-service 8.1 How to use coMET's Shiny web-service	14 14 14
9	SessionInfo	16

2 Introduction

The CoMET package is a web-based plotting tool and R-based package to visualize EWAS (epigenome-wide association scan) results in a genomic region of interest. CoMET provides a plot of the EWAS association signal and visualisation of the methylation correlation between CpG sites (co-methylation). The CoMET package also provides the option to annotate the region using functional genomic information, including both user-defined features and pre-selected features based on the Encode project. The plot can be customized with different parameters, such as plot labels, colours, symbols, heatmap colour scheme, significance thresholds, and including reference CpG sites. Finally, the tool can also be applied to display the correlation patterns of other genomic data in any species, e.g. gene expression array data.

coMET generates a multi-panel plot to visualize EWAS results, co-methylation patterns, and annotation tracks in a genomic region of interest. A coMET figure (cf. Fig. 1) includes three components:

- 1. the upper plot shows the strength and extent of EWAS association signal;
- 2. the middle panel provides customized annotation tracks;
- 3. the lower panel shows the correlation between selected CpG sites in the genomic region.

The structure of the plots builds on snp.plotter (Luna et al., 2007), with extensions to incorporate genomic annotation tracks and customized functions. coMET produces plots in PDF and Encapsulated Postscript (EPS) format.

3 Usage

CoMET requires the installation of R, the statistical computing software, freely available for Linux, Windows, or MacOS. CoMET can be downloaded from bioconductor. Packages can be installed using the install.packages command in R. The coMET R package includes two major functions *comet.web* and *comet*. The function *comet.web* generates output plot with the same settings of genomic annotation tracks as that of the webservice (http://www.epigen.kcl.ac.uk/comet or directly http://comet.epigen.kcl.ac.uk: 3838/coMET/). The function *comet* generates output plots with the customized annotation tracks defined by user.

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

CoMET can be loaded into R using this command:

```
library(coMET)
## Loading required package:
                               qrid
## Loading required package:
                               biomaRt
## Loading required package:
                               Guiz
## Loading required package:
                               BiocGenerics
## Loading required package:
                               parallel
##
## Attaching package: 'BiocGenerics'
##
## The following objects are masked from 'package:parallel':
##
```

```
##
      clusterApply, clusterApplyLB, clusterCall, clusterEvalQ, clusterExport,
##
      clusterMap, parApply, parCapply, parLapply, parLapplyLB, parRapply,
##
     parSapply, parSapplyLB
##
## The following object is masked from 'package:stats':
##
##
      xtabs
##
## The following objects are masked from 'package:base':
##
      anyDuplicated, append, as.data.frame, as.vector, cbind, colnames, do.call,
##
      duplicated, eval, evalq, Filter, Find, get, intersect, is.unsorted, lapply,
##
     Map, mapply, match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##
##
     Position, rank, rbind, Reduce, rep.int, rownames, sapply, setdiff, sort,
##
      table, tapply, union, unique, unlist, unsplit
##
## Loading required package: S4Vectors
## Loading required package:
                              stats4
## Loading required package:
                              IRanges
## Loading required package: GenomeInfoDb
## Loading required package: GenomicRanges
```

The configuration file specifies the options for the coMET plot. Example configuration and input files are also provided on http://www.epigen.kcl.ac.uk/comet. Information about the package can viewed from within R using this command:

```
?comet
?comet.web
```

4 Files formats

There are four types of files that the user should or can give to produce the plot:

- 1. info file is defined in the option DATA.FILE. It is mandatory and has to be a file in tabular format with an header.
- 2. correlation file is defined in the option CORMATRIX.FILE. It is mandatory and has to be a file in tabular format with an header.
- 3. extra info files are defined in the option DATA.FILE.LARGE. It is not mandatory, but if you add it, it has to be a file in tabular format with an header.
- 4. Annotation info file is defined in the option BIOFEAT.USER.FILE.

4.1 Format of info file (mandatory)

It is mandatory and has to be a file in tabular format with an header. Info file can be a list of CpG sites with/without Beta value (DNA methylation level) or direction sign. If it is a site file then it is mandatory to

have the 4 columns as shown below with headers in the same order. Beta can be the 5th column(optional) and it can be either a numeric value (positive or negative values) or only direction sign ("+", "-")

```
extdata <- system.file("extdata", package="coMET",mustWork=TRUE)</pre>
infofile <- file.path(extdata, "cyp1b1_infofile.txt")</pre>
data_info <-read.csv(infofile, header = TRUE,</pre>
                      sep = "\t", quote = "")
head(data_info)
##
       TargetID CHR MAPINFO
                                       Pval
## 1 cg22248750
                   2 38294160 2.749858e-01
## 2 cg11656478
                   2 38297759 7.794549e-01
## 3 cg14407177
                   2 38298023 2.863869e-01
## 4 cg02162897
                   2 38300537 3.148201e-07
## 5 cg20408276
                   2 38300586 1.467739e-06
## 6 cg00565882
                   2 38300707 7.563132e-03
```

Alternatively, the info file can be region-based and if so, the region-based info file must have the 5 columns (see below) with headers in this order. The beta or direction can be included in the 6th column (optional).

```
extdata <- system.file("extdata", package="coMET",mustWork=TRUE)</pre>
infoexp <- file.path(extdata, "cyp1b1_infofile_exprGene_region.txt")</pre>
data_infoexp <-read.csv(infoexp, header = TRUE,</pre>
                         sep = "\t", quote = "")
head(data_infoexp)
##
                                  TargetID CHR MAPINFO.START MAPINFO.STOP
                                                                                     Pval BETA
                                                                   38298453 3.064357e-17
## 1 ENSG00000138061.7_38294652_38298453
                                             2
                                                     38294652
## 2 ENSG00000138061.7_38301489_38302532
                                             2
                                                     38301489
                                                                   38302532 1.145430e-07
## 3 ENSG00000138061.7_38302919_38303323
                                             2
                                                     38302919
                                                                   38303323 1.014050e-08
```

4.2 Format of correlation matrix (mandatory)

It is mandatory and has to be a file in tabular format with an header. The data file used for the correlation matrix is described in the option CORMATRIX.FILE. This tab-delimited file can take 3 formats described in the option CORMATRIX.FORMAT:

- 1. CORMATRIX: pre-computed correlation matrix provided by the user; Dimension of matrix: CpG_number X CpG_number. Need to put the CpG sites/regions in the ascending order of positions and to have a header with the name of CpG sites/regions;
- 2. RAW: Raw data format. Correlations of these can be computed by one of 3 methods Spearman, Pearson, Kendall (option CORMATRIX.METHOD). Dimension of matrix: sample_size X CpG_number. Need to have a header with the name of CpG sites/regions;
- 3. RAW_REV: Raw data format. Correlations of these can be computed by one of 3 methods Spearman, Pearson, Kendall (option CORMATRIX.METHOD). Dimension of matrix: CpG_number X sample_size. Need to have the row names of CpG sites/regions and a header with the name of samples;

4.3 Format of extra info file

It is optional file and if you add one, it has to be a file in tabular format with an header. The extra info files can be described in the option DATA.FILE.LARGE. Different extra info files are separated by a comma.

This can be another type of info file (e.g expression or replication data) and should follow the same rules as the standard info file.

4.4 Format of annotation file

The file is defined in the option BIOFEAT.USER.FILE and the format of file is the format accepted by GViz (BED, GTF, and GFF3).

4.5 Option of config.file

It is a file where each line is one option. The name of option is in capital and is separated to its value by "=". If there are multiple values such as for the option LIST.TRACKS or the options for exta data. If you would like to make your own changes to the plot you can download the configuration file, make changes to it, and upload it into R as shown in the example below.

The important options of a coMET figure include three components:

- 1. The upper plot shows the strength and extent of EWAS association signal.
 - PVAL.THRESHOLD: Significance threshold to be displayed as a red dashed line
 - DISP.ASSOCIATION: This logical option works only if MYDATA.FILE contains the effect direction (MYDATA.FORMAT=SITE_ASSOC or REGION_ASSOC). The value can be TRUE or FALSE: if FALSE (default), for each point of data in the p-value plot, the color of symbol is the color of co-methylation pattern between the point and the reference site; if TRUE, the effect direction is shown. If the association is positive, the color is the one defined with the option COLOR.LIST. On the other hand, if the association is negative, the color is the opposed color.

• DISP.REGION: This logical option works only if MYDATA.FILE contains regions (MYDATA.FORMAT = REGION or REGION_ASSOC). The value can be TRUE or FALSE (default). If TRUE, the genomic element will be shown by a continuous line with the color of the element, in addition to the symbol at the center of the region. If FALSE, only the symbol is shown.

- 2. The middle panel provides customized annotation tracks;
 - LIST.TRACKS (for comet.web function): List of annotation tracks that can be visualised: geneENSEMBL, CGI, ChromHMM, DNAse, RegENSEMBL, SNP, transcriptENSEMBL, SNPstoma, SNPstru, SNPstrustoma, ISCA, COSMIC, GAD, ClinVar, GeneReviews, GWAS, Clin-VarCNV, GCcontent, genesUCSC, xenogenesUCSC.
 - TRACKS.GVIZ, TRACKS.GGBIO, TRACKS.TRACKVIEWER (for comet function): For each option, it is possible to give a list of annotation tracks that is created by the Gviz, GGBio, and TrackViewer bioconductor packages.
- 3. The lower panel shows the correlation between selected CpG sites in the genomic region.
 - CORMATRIX.FORMAT : Format of the input fie CORMATRIX.FILE: either raw data (option RAW if CpG sites are by column and samples by row or option RAW_REV if CpG site are by row and samples by column) or correlation matrix (option CORMATRIX)
 - CORMATRIX.METHOD: If raw data are provided it will be necessary to produce the correlation matrix using one of 3 methods (spearman, pearson and kendall).
 - CORMATRIX.COLOR.SCHEME: There are 5 colors (heat, bluewhitered, cm, topo, gray, bluetored)

```
extdata <- system.file("extdata", package="coMET",mustWork=TRUE)</pre>
configfile <- file.path(extdata, "config_cyp1b1_zoom_4webserver.txt")</pre>
data_config <-read.csv(configfile, quote = "", header=FALSE)</pre>
data_config
##
                                                              V1
## 1
                                               DISP.MYDATA=TRUE
## 2
                                             MYDATA.FORMAT=SITE
                                          MYDATA.REF=cg02162897
## 3
                                   PVAL.THRESHOLD=4.720623e-06
## 4
                                         DISP.ASSOCIATION=FALSE
## 5
## 6
                                              DISP.REGION=FALSE
## 7
                               MYDATA.LARGE.FORMAT=REGION_ASSO
## 8
                                   DISP.ASSOCIATION.LARGE=TRUE
                                         DISP.REGION.LARGE=TRUE
## 9
## 10
                          SAMPLE.LABELS.LARGE=Gene expression
## 11
                                         COLOR.LIST.LARGE=green
                                     SYMBOLS.LARGE=diamond-fill
## 12
## 13
                                                 START=38290160
## 14
                                                   END=38303219
## 15
                                             SAMPLE.LABELS=CpG
                                            SYMBOLS=circle-fill
## 16
## 17
                                                      LAB.Y=log
                                            DISP.COLOR.REF=TRUE
## 18
## 19
                                           CORMATRIX.FORMAT=RAW
## 20
                                         DISP.CORMATRIXMAP=TRUE
## 21
                                      CORMATRIX.METHOD=spearman
## 22
                           CORMATRIX.COLOR.SCHEME=bluewhitered
```

```
## 23
                                           DISP.PHYS.DIST=TRUE
## 24
                                            DISP.COLOR.BAR=TRUE
## 25
                                               DISP.TYPE=symbol
                                               DISP.LEGEND=TRUE
## 26
## 27
                                       LIST.TRACKS=geneENSEMBL
## 28
                                                            CGI
## 29
                                                       ChromHMM
## 30
                                                          DNAse
## 31
                                                     RegENSEMBL
## 32
                                                            SNP
                                          DISP.MULT.LAB.X=FALSE
## 33
                                                 IMAGE.TYPE=pdf
## 34
## 35 IMAGE.TITLE="Example a-DMR in CYP1B1 in Adipose tissue"
## 36
                         IMAGE.NAME=cyp1b1_zoom_plus_name_expr
## 37
                                                 IMAGE.SIZE=3.5
## 38
                                                    GENOME=hg19
## 39
                            DATASET.GENE=hsapiens_gene_ensembl
## 40
                                      DATASET.SNP=hsapiens_snp
## 41
                                          VERSION.DBSNP=snp138
## 42
                            DATASET.SNP.STOMA=hsapiens_snp_som
                       DATASET.REGULATION=hsapiens_feature_set
## 43
## 44
                               DATASET.STRU=hsapiens_structvar
## 45
                    DATASET.STRU.STOMA=hsapiens_structvar_som
## 46
                                    PATTERN.REGULATION=GM12878
## 47
                                           BROWSER.SESSION=UCSC
```

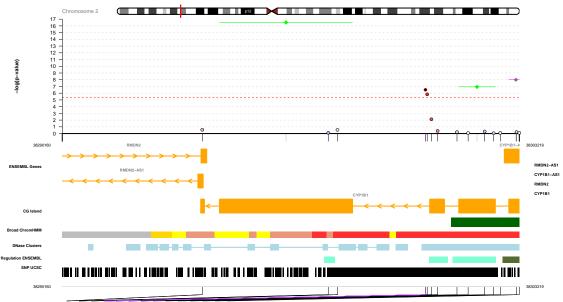
5 Creating a plot like the webservice: comet.web

User can draw coMET via the coMET website (http://epigen.kcl.ac.uk/comet). It is possible to reproduce the web service plotting defaults by using the function comet.web, for example see Figure 1.

6 Creating a plot with the generic function: comet

It is possible to create the annotation tracks by Gviz, trackviewer or ggbio, for example see Figure 2. Currently, the Gviz option for annotation tracks, in combination with the heatmap of correlation values between genomic elements, provides the most informative and easy approach to visualize graphics.

Example a-DMR in CYP1B1 in Adipose tissue



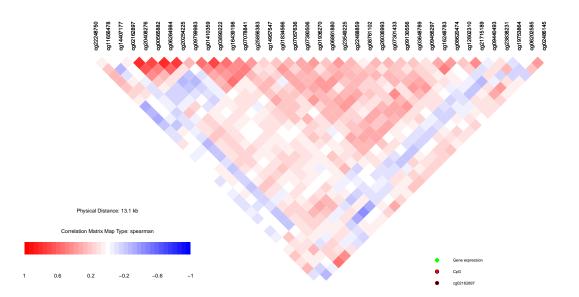


Figure 1: Plot with comet.web function.

6.1 coMET plot: pvalue plot, annotation tracks, and correlation matrix

```
extdata <- system.file("extdata", package="coMET",mustWork=TRUE)
configfile <- file.path(extdata, "config_cyp1b1_zoom_4comet.txt")
myinfofile <- file.path(extdata, "cyp1b1_infofile.txt")
myexpressfile <- file.path(extdata, "cyp1b1_infofile_exprGene_region.txt")
mycorrelation <- file.path(extdata, "cyp1b1_res37_rawMatrix.txt")</pre>
```

6.2 coMET plot: annotation tracks and correlation matrix

It is possible to visualise only annotation tracks and the correlation between genetic elements. In this case, we need to use the option DISP.PVALUEPLOT=FALSE, for example see Figure 3.

```
extdata <- system.file("extdata", package="coMET",mustWork=TRUE)</pre>
configfile <- file.path(extdata, "config_cyp1b1_zoom_4cometnopval.txt")</pre>
myinfofile <- file.path(extdata, "cyp1b1_infofile.txt")</pre>
mycorrelation <- file.path(extdata, "cyp1b1_res37_rawMatrix.txt")</pre>
chrom <- "chr2"
start <- 38290160
end <- 38303219
gen <- "hg19"
strand <- "*"
genetrack <-genesENSEMBL(gen,chrom,start,end,showId=FALSE)</pre>
snptrack <- snpBiomart(chrom, start, end,</pre>
                         dataset="hsapiens_snp_som",showId=FALSE)
strutrack <- structureBiomart(chrom, start, end,</pre>
                                 strand, dataset="hsapiens_structvar_som")
clinVariant<-ClinVarMainTrack(gen, chrom, start, end)</pre>
clinCNV<-ClinVarCnvTrack(gen,chrom,start,end)</pre>
gwastrack <-GWASTrack(gen,chrom,start,end)</pre>
geneRtrack <-GeneReviewsTrack(gen,chrom,start,end)</pre>
```

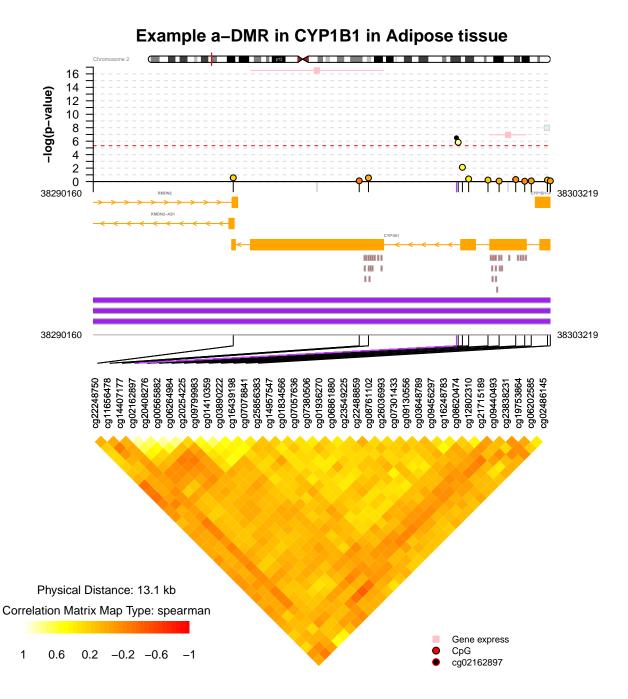


Figure 2: Plot with comet function.

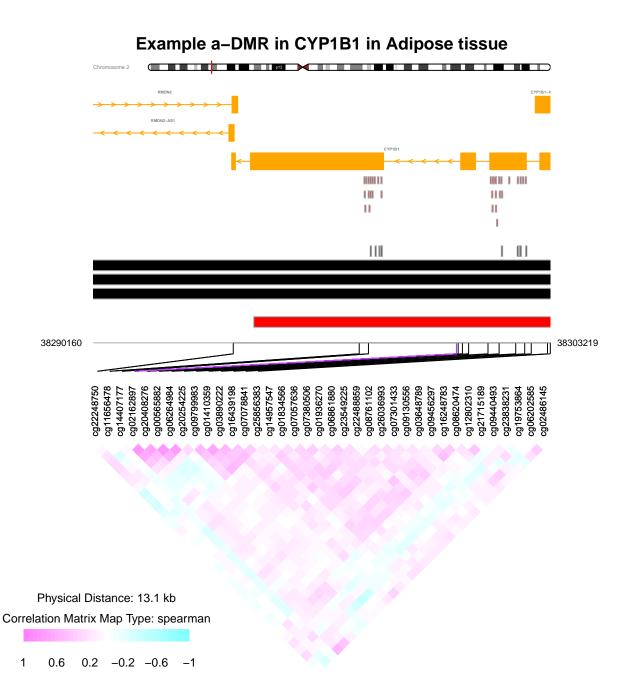


Figure 3: Plot with comet function without pvalue plot.

7 Extra information about annotation tracks

7.1 Genes and transcripts from ENSEMBL and UCSC

The color of genetic elements are defined by the R package Gviz.

7.2 Regulatory elements from ENSEMBL

The color of regulatory elements from ENSEMBL are defined from the same criteria of ENSEMBL done in 2014. The colors and the list of features can be updated in ENSEMBL and not yet in coMET. Please to contact us if you see some difference.

Currently the colors are :

feature color

Promoter Associated darkolivegreen CTCF Binding Site cadetblue1 Gene Associated coral

Non-Gene Associated darkgoldenrod1
Predicted Transcribed Region greenyellow
PolIII Transcription Associated purple
Enhancer gold

Transcription Factor Binding Site darkorchid1
Predicted Weak enhancer/Cis-reg element yellow
Heterochromatin wheat4
Open Chromatin snow3
Promoter Flank tomato
Repressed/Low Activity snow4
Unclassified aquamarine

7.3 ChromHMM from UCSC

The color of regulatory regions from UCSC are defined from the same criteria of UCSC done in 2014. The colors and the list of features can be updated in UCSC and not yet in coMET. Please to contact us if you see some difference.

Currently the colors are :

feature color 1_Active_Promoter firebrick1 2_Weak_Promoter darksalmon 3_Poised_Promoter blueviolet 4_Strong_Enhancer Orange 5_Strong_Enhancer coral 6_Weak_Enhancer yellow 7_Weak_Enhancer gold

8_Insulator cornflowerblue 9_Txn_Transition darkolivegreen 10_Txn_Elongation forestgreen 11_Weak_Txn darkseagreen1 12_Repressed gainsboro 13_Heterochrom/lo gray74 14_Repetitive/CNV gray77 15_Repetitive/CNV gray86

8 coMET: Shiny web-service

8.1 How to use coMET's Shiny web-service

If you want to use coMET via its webservice, you will go to http://www.epigen.kcl.ac.uk/comet and select one of different instances or directly one of instance such as http://comet.epigen.kcl.ac.uk:3838/coMET/. We have created different instances of coMET because we did not have the pro version of Shiny. Nevertheless, it uses the same version of coMET.

If you use coMET from a Shiny webservice, you do not need to install coMET package on you computer and to know the command lines to run on R. It is easy to use it, you have just to load different files and configurate your plot. The creation of coMET's plot can take a time because it makes a connexion to UCSC or/and ENSEMBL for the annotation tracks. First, the plot is created on the webpage, you can save the picture. But if you want to have better quality, you need to push the button download and the plot would be recreated in a file at pdf or eps format.

8.2 How to install coMET's Shiny web-service

These are different steps to install coMET on your Shiny web-service. You need to be root of your server to install it.

- 1. You can install on your server an instance of Shiny http://shiny.rstudio.com/.
- 2. You need also to install R, Bioconductor and, of course, the package coMET.
- 3. In the Shiny's folder (e.g. /var/shiny-server/www), you can create a folder called "COMET".
- 4. Following this, you can install the two scripts of coMET that you find in the folder www of coMET package in the new folder.
- 5. you need change owner and the permission to access this folder. Only the user called Shiny can access.
- 6. You need now to update the configuration file of Shiny (e.g. /etc/shiny-server/shiny-server.conf).
- 7. In this end, you can restart the service Shiny via the command line: sudo restart shiny-server

Your Shiny's configuration file:

```
run_as shiny;
# Define a top-level server which will listen on a port
server {
    # Instruct this server to listen on port 3838
listen 3838;
    # Define the location available at the base URL
location / {
        # Run this location in 'site_dir' mode, which hosts the entire directory
        # tree at '/srv/shiny-server'
        site_dir /var/shiny-server/www;

# Define where we should put the log files for this location
log_dir /var/shiny-server/log;

# Should we list the contents of a (non-Shiny-App) directory when the user
    # visits the corresponding URL?
```

```
directory_index off;

# app_init_timeout 3600;
# app_idle_timeout 3600;
}
```

9 SessionInfo

The following is the session info that generated this vignette:

toLatex(sessionInfo())

- R version 3.1.2 (2014-10-31), x86_64-pc-linux-gnu
- Locale: LC_CTYPE=en_GB.UTF-8, LC_NUMERIC=C, LC_TIME=en_US.UTF-8, LC_COLLATE=en_GB.UTF-8, LC_MONETARY=en_US.UTF-8, LC_MESSAGES=en_GB.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, graphics, grDevices, grid, methods, parallel, stats, stats4, utils
- Other packages: BiocGenerics 0.12.1, biomaRt 2.22.0, coMET 0.99.8, GenomeInfoDb 1.2.4, GenomicRanges 1.18.4, Gviz 1.10.9, IRanges 2.0.1, knitr 1.9, S4Vectors 0.4.0, XVector 0.6.0
- Loaded via a namespace (and not attached): acepack 1.3-3.3, AnnotationDbi 1.28.1, base64enc 0.1-2, BatchJobs 1.5, BBmisc 1.9, Biobase 2.26.0, BiocParallel 1.0.3, BiocStyle 1.4.1, Biostrings 2.34.1, biovizBase 1.14.1, bitops 1.0-6, brew 1.0-6, BSgenome 1.34.1, checkmate 1.5.1, cluster 2.0.1, codetools 0.2-10, colorspace 1.2-4, colortools 0.1.5, DBI 0.3.1, dichromat 2.0-0, digest 0.6.8, evaluate 0.5.5, fail 1.2, foreach 1.4.2, foreign 0.8-62, formatR 1.0, Formula 1.2-0, GenomicAlignments 1.2.1, GenomicFeatures 1.18.3, GGally 0.5.0, ggbio 1.14.0, ggplot2 1.0.0, graph 1.44.1, gridExtra 0.9.1, gtable 0.1.2, gWidgets 0.0-54, gWidgetstcltk 0.0-55, hash 2.2.6, highr 0.4, Hmisc 3.14-6, iterators 1.0.7, lattice 0.20-29, latticeExtra 0.6-26, MASS 7.3-37, matrixStats 0.14.0, munsell 0.4.2, nnet 7.3-9, OrganismDbi 1.8.0, pbapply 1.1-1, plyr 1.8.1, proto 0.3-10, RBGL 1.42.0, RColorBrewer 1.1-2, Rcpp 0.11.4, RCurl 1.95-4.5, reshape 0.8.5, reshape 2 1.4.1, rpart 4.1-9, Rsamtools 1.18.2, RSQLite 1.0.0, rtracklayer 1.26.2, scales 0.2.4, sendmailR 1.2-1, splines 3.1.2, stringr 0.6.2, survival 2.37-7, tcltk 3.1.2, tools 3.1.2, trackViewer 1.2.0, VariantAnnotation 1.12.9, XML 3.98-1.1, zlibbioc 1.12.0