fCCAC: functional Canonical Correlation Analysis to evaluate Covariance between nucleic acid sequencing datasets

Pedro Madrigal

Last Modified: September, 2016. Compiled: October 17, 2016

Wellcome Trust Sanger Institute, Cellular Genetics Programme, Hinxton, Cambridge, UK
University of Cambridge, WT MRC Stem Cell Institute, Department of Surgery, Cambridge, UK

Contents

1 Introduction

Computational evaluation of variability across sequencing datasets is a crucial step in genomic science, as it allows both to evaluate the reproducibility across biological or technical replicates, and to compare different datasets to identify their potential correlations. fCCAC is an application of functional canonical correlation analysis to assess covariance of nucleic acid sequencing datasets such as chromatin immunoprecipitation followed by deep sequencing (ChIP-seq), ChIP-exo, etc. Basic processing of this type of data can be performed using Bioconductor packages such as *NarrowPeaks*, *CSAR*, *CexoR*, *csaw*, *ChIPseeker*, *ChIPQC*, and others. Basic and advanced ChIP-seq workflows are available at the Bioconductor website: https://www.bioconductor.org/help/workflows/sequencing/ and https://www.bioconductor.org/help/workflows/chipseqDB/. Once regions of interest, such as peaks, exons, UTRs, etc. are obtained, bigwig data of the mapped reads are necessesary too for using *fCCAC*, and can be obtained using *rtracklayer*.

Detailed information about the methodology can be found in Madrigal (2016).

2 Example

Data used in the example correspond to H3K4me3 triplicates (wild-type H9-hESCs) downloaded from ArrayExpress (E-ERAD-191) and processed as in (Bertero et al., 2015). Aggregate sets of reproducible peaks from Bertero et al. were considered as genomic regions of interest in the analysis. Region 40000000-48129895 in chromosome 21 was selected for the illustrative example shown in this vignette.

```
R> options(width=60);
R> if (.Platform$OS.type == "windows") { print("...rtracklayer is unable to read bigWig format files in Windows...") }
R> if (.Platform$OS.type == "unix") {
```

```
## hg19. chr21:40000000-48129895 H3K4me3 data from Bertero et al. (2015)

owd <- setwd(tempdir());

library(fCCAC)

bigwig1 <- "chr21_H3K4me3_1.bw"
bigwig2 <- "chr21_H3K4me3_2.bw"
bigwig3 <- "chr21_H3K4me3_3.bw"
peakFile <- "chr21_merged_ACT_K4.bed"
labels <- c( "H3K4me3", "H3K4me3","H3K4me3" )

r1 <- system.file("extdata", bigwig1,  package="fCCAC",mustWork = TRUE)
r2 <- system.file("extdata", bigwig2,  package="fCCAC",mustWork = TRUE)
r3 <- system.file("extdata", bigwig3,  package="fCCAC",mustWork = TRUE)
r4 <- system.file("extdata", peakFile, package="fCCAC",mustWork = TRUE)
ti <- "H3K4me3 peaks (chr21)"

fc <- fccac(bar=NULL, main=ti, peaks=r4, bigwigs=c(r1,r2,r3), labels=labels, splines=15, nbins=100, ncan=15)

head(fc)

setwd(owd)

}
```

```
[1] "Reading peaks..."
[1] "Starting fCCAC..."
[1] "Reading bigWig file...1/3"
[1] "Reading bigWig file...2/3"
[1] "Reading bigWig file...3/3"
[1] "Performing fCCA in pair...1/3"
[1] "H3K4me3_Rep1...vs...H3K4me3_Rep2"
[1] "Performing fCCA in pair...2/3"
[1] "H3K4me3_Rep1...vs...H3K4me3_Rep3"
[1] "Performing fCCA in pair...3/3"
[1] "H3K4me3_Rep2...vs...H3K4me3_Rep3"

R>
```

As we can observe, the triplicates present very high values both in first squared canonical correlation ($>$0.99) and in their F values, which denotes high reproducibility of the experiment.

Finally, if all pairwise comparisons have been computed ('tf=c()' by default), we can plot a heatmap of F values:

```
R> options(width=60)
R> if (.Platform$OS.type == "windows") { print("...rtracklayer is unable to read bigWig format files in Windows...") }
R> if (.Platform$OS.type == "unix" ){ heatmapfCCAC(fc) }
R>
```

Important notes:

- F is an overall measure of shared covariance. It is expected that good replicates will have F values close to 100 (such as in the example above).

- Because heavy smoothing is suggested for functional CCA (Ramsay and Silverman (2005); Ramsay et al. (2009)), a low number of splines (parameter splines) when compared to the total length of the genomic regions is recommended. The parameter 'nbins' can be low for narrow peaks (e.g., 50 for TFs and narrow chromatin marks) and increased for broad domain chromatin marks. The number of canonical correlations to compute (ncan) is limited by the number of splines used and the number of genomics regions to analyse. More information about data approximations used can be found in the Supplementary Information in Madrigal (2016).

3 References

- Madrigal P (2016) fCCAC: functional canonical correlation analysis to evaluate covariance between nucleic acid sequencing datasets. **bioRxiv**, 10.1101/060780. http://biorxiv.org/content/early/2016/06/27/060780.
- Bailey TL, et al. (2013) Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. **PLoS Comput Biol** 9: e1003326.
- Bertero A, et al. (2015) Activin/nodal signaling and NANOG orchestrate human embryonic stem cell fate decisions by controlling the H3K4me3 chromatin mark. **Genes Dev.** 29: 702-17.
- Ramsay JO, Silverman BW (2005) Functional Data Analysis. Springer Verlag, New York.
- Ramsay JO, et al. (2009) Functional Data Analysis with R and MATLAB. SpringerVerlag, New York.

4 Details

This document was written using:

```
R> sessionInfo()

R version 3.3.1 (2016-06-21)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 16.04.1 LTS

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [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] grid      parallel  stats4    stats     graphics
 [6] grDevices utils     datasets  methods   base

other attached packages:
[1] fCCAC_1.0.0          GenomicRanges_1.26.0
[3] GenomeInfoDb_1.10.0  IRanges_2.8.0
[5] S4Vectors_0.12.0     BiocGenerics_0.20.0

loaded via a namespace (and not attached):
 [1] mclust_5.2                Rcpp_0.12.7
 [3] mvtnorm_1.0-5             lattice_0.20-34
 [5] circlize_0.3.9            Rsamtools_1.26.0
 [7] class_7.3-14              Biostrings_2.42.0
 [9] assertthat_0.1            gridBase_0.4-7
[11] plyr_1.8.4                chron_2.3-47
[13] ggplot2_2.1.0             GlobalOptions_0.0.10
[15] zlibbioc_1.20.0           diptest_0.75-7
```

```
[17] data.table_1.9.6           kernlab_0.9-25
[19] whisker_0.3-2              GetoptLong_0.1.5
[21] Matrix_1.2-7.1             labeling_0.3
[23] splines_3.3.1              BiocParallel_1.8.0
[25] readr_1.0.0                stringr_1.1.0
[27] RCurl_1.95-4.8             munsell_0.4.3
[29] rtracklayer_1.34.0         shape_1.4.2
[31] nnet_7.3-12                SummarizedExperiment_1.4.0
[33] tibble_1.2                 matrixStats_0.51.0
[35] dendextend_1.3.0           XML_3.98-1.4
[37] GenomicAlignments_1.10.0   MASS_7.3-45
[39] bitops_1.0-6               gtable_0.2.0
[41] magrittr_1.5               scales_0.4.0
[43] KernSmooth_2.23-15         stringi_1.1.2
[45] impute_1.48.0              XVector_0.14.0
[47] reshape2_1.4.1             flexmix_2.3-13
[49] robustbase_0.92-6          BiocStyle_2.2.0
[51] rjson_0.2.15               RColorBrewer_1.1-2
[53] tools_3.3.1                fpc_2.1-10
[55] BSgenome_1.42.0            Biobase_2.34.0
[57] trimcluster_0.1-2          DEoptimR_1.0-6
[59] seqPattern_1.6.0           plotrix_3.6-3
[61] colorspace_1.2-7           cluster_2.0.5
[63] genomation_1.6.0           prabclus_2.2-6
[65] ComplexHeatmap_1.12.0      fda_2.4.4
[67] modeltools_0.2-21
```