

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

Pedro Madrigal

Last Modified: August, 2021. Compiled: October 24, 2023

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

Contents

1	Introduction	1
2	Example	2
3	References	3
4	Details	3

1 Introduction

Computational evaluation of variability across Next-generation sequencing (NGS) datasets is a crucial step in genomic science. 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 NGS datasets such as chromatin immunoprecipitation followed by deep sequencing (ChIP-seq), ChIP-exo, RNA-seq, ATAC-seq, etc.

Basic processing of NGS data can be performed using Bioconductor packages such as *CSAR*, *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 necessary for using *fCCAC*, and can be obtained using *rtracklayer*.

Detailed information about the statistical 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=50);
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"
```

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=50)
R> if (.Platform$OS.type == "windows") { print("...rtracklayer is unable to read bigWig format files in Windows...") }
```

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

```
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 (2017) fCCAC: functional canonical correlation analysis to evaluate covariance between nucleic acid sequencing datasets. **Bioinformatics**, 33: 746-748. <http://doi.org/10.1093/bioinformatics/btw724>.
- 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 4.3.1 (2023-06-16)  
Platform: x86_64-pc-linux-gnu (64-bit)  
Running under: Ubuntu 22.04.3 LTS  
  
Matrix products: default  
BLAS: /home/biocbuild/bbs-3.18-bioc/R/lib/libRblas.so  
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.10.0  
  
locale:  
[1] LC_CTYPE=en_US.UTF-8  
[2] LC_NUMERIC=C
```

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

```
[3] LC_TIME=en_GB
[4] LC_COLLATE=C
[5] LC_MONETARY=en_US.UTF-8
[6] LC_MESSAGES=en_US.UTF-8
[7] LC_PAPER=en_US.UTF-8
[8] LC_NAME=C
[9] LC_ADDRESS=C
[10] LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8
[12] LC_IDENTIFICATION=C

time zone: America/New_York
tzcode source: system (glibc)

attached base packages:
[1] grid      stats4    stats    graphics
[5] grDevices utils     datasets methods
[9] base

other attached packages:
[1] fCCAC_1.28.0      GenomicRanges_1.54.0
[3] GenomeInfoDb_1.38.0 IRanges_2.36.0
[5] S4Vectors_0.40.0  BiocGenerics_0.48.0

loaded via a namespace (and not attached):
 [1] bitops_1.0-7
 [2] rlang_1.1.1
 [3] magrittr_2.0.3
 [4] clue_0.3-65
 [5] GetoptLong_1.0.5
 [6] gridBase_0.4-7
 [7] matrixStats_1.0.0
 [8] compiler_4.3.1
 [9] png_0.1-8
[10] vctrs_0.6.4
[11] reshape2_1.4.4
[12] stringr_1.5.0
[13] shape_1.4.6
[14] pkgconfig_2.0.3
[15] crayon_1.5.2
[16] fastmap_1.1.1
[17] magick_2.8.1
[18] XVector_0.42.0
[19] labeling_0.4.3
[20] utf8_1.2.4
[21] Rsamtools_2.18.0
[22] deSolve_1.38
[23] rmarkdown_2.25
[24] tzdb_0.4.0
[25] pracma_2.4.2
[26] bit_4.0.5
[27] xfun_0.40
[28] zlibbioc_1.48.0
[29] DelayedArray_0.28.0
[30] BiocParallel_1.36.0
```

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

```
[31] parallel_4.3.1
[32] cluster_2.1.4
[33] R6_2.5.1
[34] stringi_1.7.12
[35] RColorBrewer_1.1-3
[36] rtracklayer_1.62.0
[37] rainbow_3.7
[38] Rcpp_1.0.11
[39] SummarizedExperiment_1.32.0
[40] iterators_1.0.14
[41] knitr_1.44
[42] hdrdce_3.4
[43] readr_2.1.4
[44] Matrix_1.6-1.1
[45] splines_4.3.1
[46] tidyselect_1.2.0
[47] abind_1.4-5
[48] yaml_2.3.7
[49] seqPattern_1.34.0
[50] doParallel_1.0.17
[51] codetools_0.2-19
[52] lattice_0.22-5
[53] tibble_3.2.1
[54] plyr_1.8.9
[55] withr_2.5.1
[56] Biobase_2.62.0
[57] ks_1.14.1
[58] evaluate_0.22
[59] archive_1.1.6
[60] circlize_0.4.15
[61] mclust_6.0.0
[62] Biostrings_2.70.0
[63] pillar_1.9.0
[64] BiocManager_1.30.22
[65] fda_6.1.4
[66] MatrixGenerics_1.14.0
[67] KernSmooth_2.23-22
[68] foreach_1.5.2
[69] pcaPP_2.0-3
[70] generics_0.1.3
[71] vroom_1.6.4
[72] RCurl_1.98-1.12
[73] hms_1.1.3
[74] ggplot2_3.4.4
[75] munsell_0.5.0
[76] scales_1.2.1
[77] BiocStyle_2.30.0
[78] glue_1.6.2
[79] tools_4.3.1
[80] BiocIO_1.12.0
[81] fds_1.8
[82] data.table_1.14.8
[83] BSgenome_1.70.0
[84] GenomicAlignments_1.38.0
[85] mvtnorm_1.2-3
```

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

```
[86] XML_3.99-0.14
[87] Cairo_1.6-1
[88] impute_1.76.0
[89] plotrix_3.8-2
[90] colorspace_2.1-0
[91] GenomeInfoDbData_1.2.11
[92] restfulr_0.0.15
[93] cli_3.6.1
[94] fansi_1.0.5
[95] S4Arrays_1.2.0
[96] ComplexHeatmap_2.18.0
[97] genomation_1.34.0
[98] dplyr_1.1.3
[99] gtable_0.3.4
[100] digest_0.6.33
[101] SparseArray_1.2.0
[102] farver_2.1.1
[103] rjson_0.2.21
[104] htmltools_0.5.6.1
[105] lifecycle_1.0.3
[106] GlobalOptions_0.1.2
[107] bit64_4.0.5
[108] MASS_7.3-60
```