**Infer CNV status in scATAC data**

**General purpose**

This R-script is used for calculating single cell copy number states from scATAC data. It is based on ArchR (<https://www.archrproject.com/>) and particularly, is adapted from codes mentioned in https://github.com/GreenleafLab/ArchR/issues/312 . We have only made a few modifications for the code to run.

**Required input**

An ArchR project. Example of the project could be loaded from ArchR.

**Execution**

In R, use:

|  |
| --- |
| Source(‘1.scCNV.R’)  proj\_arrow <- addCNVMatrix(proj\_arrow) ## add an CNV matrix  cnv\_matrix <- getMatrixFromProject\_mod\_cnv(proj\_arrow) ## get an CNV matrix |

**Algorithm**

The function ` addCNVMatrix ` calculates copy number states of each sliding tile (assumed to be 10 million base pairs, and walks 2 million base pairs per step, across the genome) and count the fragments in it, correct by GC-content (with knn) output the normalized per-base read depth, and smooth across the chromosome.

**Outputs**

The CNV state called by the function could be then extracted from the project using ` getMatrixFromProject\_mod\_cnv ` as a sparse matrix CxW where C = cells and W = windows. The output could be used to remove any peaks that is associated with any CNV in any cell, for downstream analysis. Such application might mask real differences from the cells but nevertheless it removes any significant differences only due to CNV but not native chromatin accessibility.

**Environment**

|  |
| --- |
| > sessionInfo()  R version 3.6.2 (2019-12-12)  Platform: x86\_64-pc-linux-gnu (64-bit)  Running under: CentOS Linux 7 (Core)  Matrix products: default  BLAS: /gpfs/bin/R-3.6.2/lib/libRblas.so  LAPACK: /gpfs/bin/R-3.6.2/lib/libRlapack.so  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 stats4 stats graphics grDevices utils datasets  [8] methods base  other attached packages:  [1] ArchR\_1.0.1 magrittr\_1.5  [3] rhdf5\_2.30.1 Matrix\_1.2-18  [5] data.table\_1.13.0 SummarizedExperiment\_1.16.1  [7] DelayedArray\_0.12.3 BiocParallel\_1.20.1  [9] matrixStats\_0.57.0 Biobase\_2.46.0  [11] GenomicRanges\_1.38.0 GenomeInfoDb\_1.22.1  [13] IRanges\_2.20.2 S4Vectors\_0.24.4  [15] BiocGenerics\_0.32.0 ggplot2\_3.3.5  loaded via a namespace (and not attached):  [1] Rcpp\_1.0.5 pillar\_1.4.6 compiler\_3.6.2  [4] XVector\_0.26.0 bitops\_1.0-6 tools\_3.6.2  [7] zlibbioc\_1.32.0 lifecycle\_0.2.0 tibble\_3.0.4  [10] gtable\_0.3.0 lattice\_0.20-41 pkgconfig\_2.0.3  [13] rlang\_0.4.11 GenomeInfoDbData\_1.2.2 withr\_2.3.0  [16] dplyr\_1.0.2 generics\_0.0.2 vctrs\_0.3.4  [19] grid\_3.6.2 tidyselect\_1.1.0 glue\_1.4.2  [22] R6\_2.4.1 Rhdf5lib\_1.8.0 purrr\_0.3.4  [25] scales\_1.1.1 ellipsis\_0.3.1 colorspace\_1.4-1  [28] RCurl\_1.98-1.2 munsell\_0.5.0 crayon\_1.3.4  [31] Cairo\_1.5-12 |