The Journal of Open Source Software

Generating fragment density plots in R/Bioconductor with VplotR

Jacques Serizay[1] **and Julie Ahringer**[1]

1 The Gurdon Institute and Department of Genetics, University of Cambridge, Cambridge UK

DOI: 10.21105/joss.03009

Software

- Review
- Repository
- Archive

Editor: Charlotte Soneson

Reviewers:

- @henrykironde
- @fgeier

Submitted: 28 January 2021
Published: 01 March 2021

Summary

A multitude of proteins bind to DNA at regulatory regions to control the expression of neighbouring genes. Their binding can be revealed by chromatin accessibility assays such as DNase-seq, MNase-seq or ATAC-seq. The size of the sequencing fragments generated by these assays, as well as their location relative to the center of regulatory regions, can be used to produce two-dimensional fragment density plots called V-plots. Such plots have successfully been used in the past to reveal nucleosome positioning or transcription factor binding sites at regulatory regions in different genomes.

Here, we present VplotR, an R package to easily generate V-plots and one-dimensional footprint profiles over single or aggregated genomic loci of interest. The use of VplotR will improve our understanding of molecular organization at regulatory regions.

Statement of Need

VplotR is an R package facilitating the generation of V-plots, i.e. two-dimensional paired-end sequencing fragment density plots (Henikoff et al., 2011). V-plots have been used in the past to elucidate nucleosome positioning and/or transcription factor binding at regulatory elements (Henikoff et al., 2011; Schep et al., 2015) (e.g. Figure 1). Only a few tools have been developed that can easily generate V-plots (Beati & Chereji, 2020; Schep et al., 2015), and they are provided as scripts to be used with the command-line interface. Thus, they lack the customization and interaction with other datasets typically available in the wealthy R/Bioconductor environment.

VplotR provides functions to import paired-end sequencing bam files from any type of DNA accessibility experiments (e.g. ATAC-seq, DNA-seq, MNase-seq) and can produce V-plots and one-dimensional footprint profiles over single or aggregated genomic loci of interest. The R package is well integrated within the Bioconductor environment and easily fits in standard genomic analysis workflows. Integrating V-plots into existing analytical frameworks has already brought new insights in chromatin organization (Serizay et al., 2020).

The Journal of Open Source Software

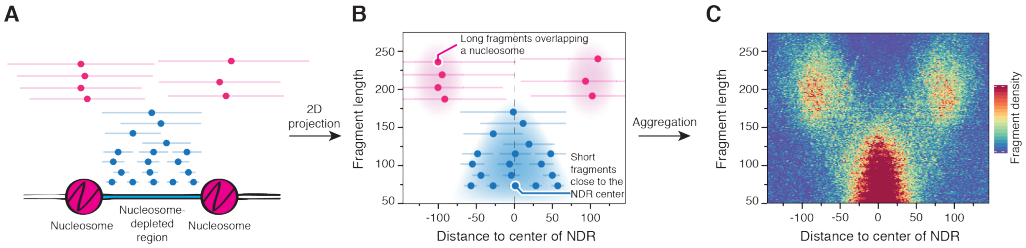

Figure 1: Interpretation of V-plots. A- Sequenced fragments (for instance obtained from ATAC-seq) mapping to a locus of interest can originate from either nucleosomal DNA (in pink) or from nucleosome-free DNA (for instance from nucleosome-depleted regions (NDRs), in blue). B- The fragments can be embedded in a two-dimension graph. The horizontal coordinate represents the distance from the center of a fragment to the center of a locus of interest (for instance the NDR). The vertical coordinate represents the length of the fragment. C- When this projection is done over hundreds of loci, it results in a fragment density plot, i.e. a matrix which can be visualized as a heatmap, with the color gradient representing the density of fragments at each set of coordinates.

Availability

VplotR is available as an R package and can be readily installed from Bioconductor. The development version can be found on GitHub. Package dependencies and system requirements are documented at https://js2264.github.io/VplotR/. VplotR has been tested using R (version 3.5 and later) on macOS (versions 10.11 and later), Ubuntu 18.04.2 and Windows machines.

To ensure stability, VplotR includes unit tests for most functions and supports continuous integration using the Travis CI platform. Code contributions, bug reports, fixes and feature requests are most welcome by opening issues and pull requests at https://js2264.github.io/VplotR/.

Implementation

The main user-level functions of VplotR are `plotVmat()`, `plotProfile()` and `plotFootprint()`. `plotVmat()` is used to generate V-plots (i.e. paired-end fragment density plots aggregated over a set of loci of interest) while `plotProfile()` is used to generate paired-end fragment plots over a single genomic locus. `plotFootprint()` can be used to profile the DNA accessibility footprint measured by a genomic assay over a motif of interest. Additional functions such as `importPEBamFiles()` and `getFragmentsDistribution()` are useful to import and investigate sets of paired-end sequencing fragments. Full examples of how to use the main package functions are described in the package vignette available at https://js2264.github.io/VplotR/articles/VplotR.html.

Workflow

Installation

VplotR and all its dependencies can be installed from Bioconductor:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("VplotR")
library("VplotR")
```

Importing data into R environment

Importing paired-end sequencing fragments (usually from MNase-seq, DNase-seq or ATAC-seq experiments) can be done with VplotR using the `importPEBamFiles()` function. If imported fragments were obtained by ATAC-seq, the argument `shift_ATAC_fragments` should be set to `TRUE` to shift the ATAC-seq sequencing fragments by +4/-5 bp (Adey et al., 2010; Buenrostro et al., 2013).

```
### To import paired-end fragments (e.g. ATAC-seq fragments),
### use `importPEBamFiles()`:
fragments <- importPEBamFiles(
    'path/to/fragments/ATACseq-fragments.bam',
    shift_ATAC_fragments = TRUE
)

### Toy datasets are provided as examples for this study:
data(ABF1_sacCer3)
data(MNase_sacCer3_Henikoff2011)
```

Checking fragment size distribution

The distribution of fragment sizes can be obtained with `getFragmentsDistribution()` and plotted using `ggplot2` (Figure 2):

```
### The one-dimensional distribution of fragment
### sizes is obtained using the
### getFragmentsDistribution() function:
dist <- getFragmentsDistribution(
    MNase_sacCer3_Henikoff2011,
    ABF1_sacCer3
)

### The resulting data frame can be plotted
### with ggplot2 (see Figure 2)
ggplot(dist, aes(x = x, y = y)) +
    geom_line() +
    theme_ggplot2() +
    labs(
        title = 'Distribution of fragment sizes',
        x = 'Fragment size', y = '# of fragments'
    )
)
```

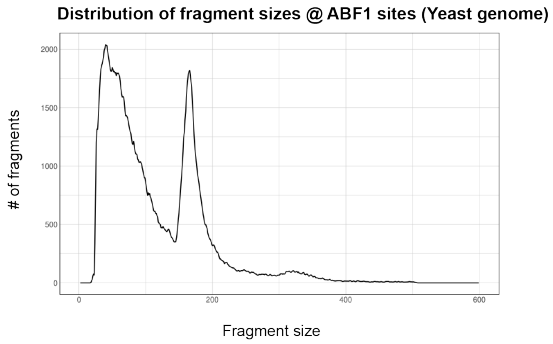

Figure 2: Distribution of fragment sizes for MNase-seq fragments mapping over ABF1 binding sites in Yeast.

Plotting fragment density over a set of genomic loci

`plotVmat()` is used to produce a V-plot (Figure 3):

```
plotVmat(
    MNase_sacCer3_Henikoff2011,
    ABF1_sacCer3
)
```

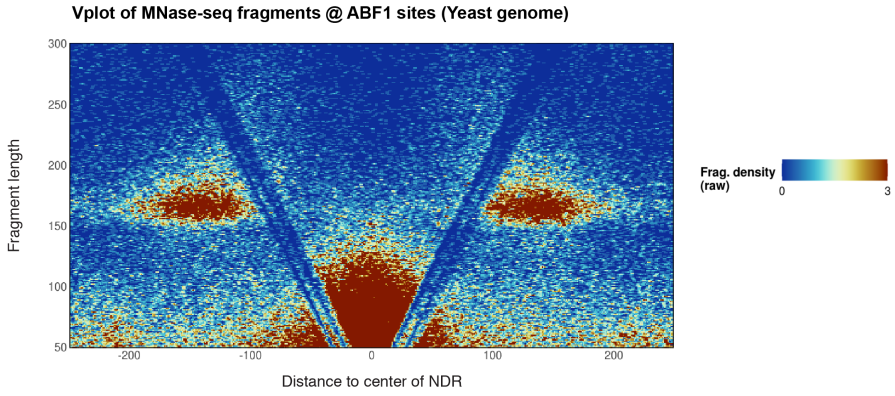

Figure 3: Vplot of MNase-seq fragments over ABF1 binding sites in Yeast. The "V" observed here originates from the protected ABF1 binding site.

Several Vplots can also be produced at once using a list of parameters as input (see https://js2264.github.io/VplotR/articles/VplotR.html#multiple-vplots for more information).

Plotting DNA accessibility footprint

Observations from V-plots can be further investigated by plotting DNA accessibility footprints over these loci. The `plotFootprint()` function can be leveraged to plot these footprint profiles (Figure 4).

```
plotFootprint(
    MNase_sacCer3_Henikoff2011,
    ABF1_sacCer3
)
```

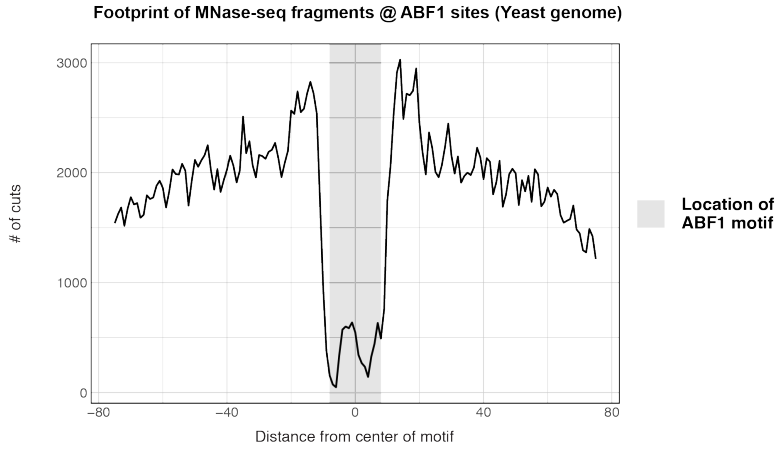

Figure 4: Footprint of MNase-seq fragments over ABF1 binding sites in Yeast.

Plotting fragments over a single genomic locus

Finally, VplotR provides a function to plot the distribution of paired-end fragments over an individual genomic window (Figure 5).

```
genes_sacCer3 <- GenomicFeatures::genes(
    TxDb.Scerevisiae.UCSC.sacCer3.sgdGene::
    TxDb.Scerevisiae.UCSC.sacCer3.sgdGene
)
plotProfile(
    fragments = MNase_sacCer3_Henikoff2011,
    window = "chrXV:186,400-187,400",
    loci = ABF1_sacCer3,
    annots = genes_sacCer3,
    min = 20, max = 200, alpha = 0.1, size = 1.5
)
```

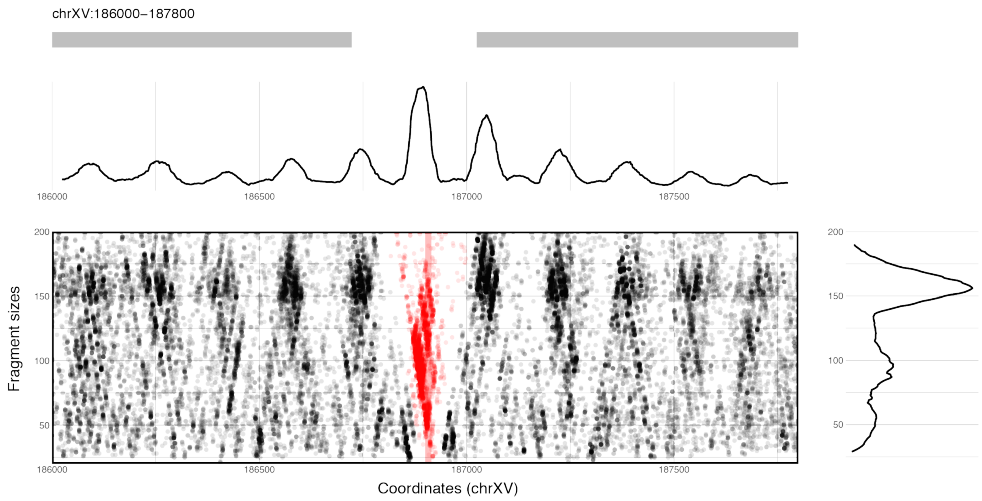

Figure 5: Distribution of MNase-seq fragments over ABF1 binding sites in Yeast at a locus on chrXV (186 kb).

The paired-end fragments overlapping a locus of interest (e.g., binding sites, provided in the `loci` argument) are shown in red while the remaining fragments mapping to the genomic window are displayed in black. Marginal curves are also plotted on the side of the distribution plot; they highlight the smoothed distribution of the position of paired-end fragment midpoints (top) or of the paired-end fragment length (right). Furthermore, genomic features provided in the `annots` arguments are displayed as horizontal bars on top of the plots. Here, the distribution of yeast MNase paired-end fragments over the bi-directional promoter regulating two divergent genes clearly reveals the ABF1 binding site (in red) located in the nucleosome-depleted region (NDR).

Research using VplotR

VplotR was recently leveraged to provide accurate insights into differential organization of nucleosomes flanking ubiquitous or tissue-specific promoters in *C. elegans* (Serizay et al., 2020).

Data availability

Yeast MNase-seq data has been obtained from Henikoff et al. (2011) (SRR3193263). ABF1 and REB1 binding motifs in yeast have been annotated in sacCer3 genome using `TFBStools` (Tan & Lenhard, 2016) and `JASPAR2018` (Khan et al., 2018). The motif occurrences with a relScore >= 0.90 were considered to be real binding sites.

Author contributions

J.S.: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. J.A.: Conceptualization, Supervision, Funding acquisition, Project administration.

Acknowledgments

We would like to thank editors from the Journal of Open Source Software, and notably C. Soneson, for their efforts to ensure a helpful and transparent reviewing procedure. We would also like to thank F. Geier and H. Kironde for their contribution as reviewers. This work was supported by a Wellcome Trust Senior Research Fellowship to J.A. (101863) and a Medical Research Council DTP studentship to J.S..

References

Adey, A., Morrison, H. G., Asan, Xun, X., Kitzman, J. O., Turner, E. H., Stackhouse, B., MacKenzie, A. P., Caruccio, N. C., Zhang, X., & Shendure, J. (2010). Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biology*, *11*(12), 1–17. https://doi.org/10.1186/gb-2010-11-12-r119

Beati, P., & Chereji, R. V. (2020). Creating 2D Occupancy Plots Using plot2DO. In *Stem Cell Transcriptional Networks* (pp. 93–108). Humana, New York, NY. https://doi.org/10.1007/978-1-0716-0301-7_5

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), 1213–1218. https://doi.org/10.1038/nmeth.2688

Henikoff, J. G., Belsky, J. A., Krassovsky, K., MacAlpine, D. M., & Henikoff, S. (2011). Epigenome characterization at single base-pair resolution. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(45), 18318–18323. https://doi.org/10.1073/pnas.1110731108

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., Lee, R. van der, Bessy, A., Cheneby, J., Kulkarni, S. R., Tan, G., Baranasic, D., Arenillas, D. J., Sandelin, A., Vandepoele, K., Lenhard, B., Ballester, B., Wasserman, W. W., Parcy, F., & Mathelier, A. (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research*, *46*(D1), 260–266. https://doi.org/10.1093/nar/gkx1126

Schep, A. N., Buenrostro, J. D., Denny, S. K., Schwartz, K., Sherlock, G., & Greenleaf, W. J. (2015). Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Research*, gr.192294.115. https://doi.org/10.1101/gr.192294.115

Serizay, J., Dong, Y., Jänes, J., Chesney, M., Cerrato, C., & Ahringer, J. (2020). Distinctive regulatory architectures of germline-active and somatic genes in C. elegans. *Genome Research*, *30*, 1752–1765. https://doi.org/10.1101/gr.265934.120

Tan, G., & Lenhard, B. (2016). TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics (Oxford, England)*, *32*(10), 1555–1556. https://doi.org/10.1093/bioinformatics/btw024