Supplementary Text S2 - cageminer: an R/Bioconductor package to prioritize candidate genes by integrating GWAS and gene coexpression networks

Fabricio Almeida-Silva [1] **and Thiago M. Venancio** [1]

[1]Universidade Estadual do Norte Fluminense Darcy Ribeiro

15 July 2022

Contents

Supplementary Text S2 - cageminer: an R/Bioconductor package to prioritize candidate genes by integrating GWAS and gene coexpression networks

1 Introduction

In a previous study, Michno et al. (2020) demonstrated that Camoco (Schaefer et al. 2018) can identify high-confidence candidate genes, but only for short sliding windows relative to each SNP. For larger sliding windows (e.g., >50 kb), there is a steep increase in the background noise, and Camoco is not able to identify any candidate gene (Michno et al. 2020).

Here, we will assess the performance of cageminer for increasing sliding windows relative to each SNP. Our goal here is to check if increasing the sliding windows leads to loss of signal, as happens with the Camoco algorithm. The data set used here was obtained from Almeida-Silva and Venancio (2021), and it contains:

- RNA-seq data of soybean (*Glycine max*) response to fungi infection.
- GWAS-derived SNPs associated with soybean resistance to *Fusarium graminearum*.

2 Assessment

The sliding windows we will consider are:

1. 10 kb
2. 50 kb
3. 100 kb
4. 1 Mb
5. 2 Mb

```r
set.seed(123) # for reproducibility

#----Load packages--------------------------------------------------------------
library(cageminer)
library(here)

#----Load required data----------------------------------------------------------
load(here("data", "cageminer_input", "gene_ranges.rda"))
load(here("data", "cageminer_input", "gmax_se.rda"))
load(here("data", "cageminer_input", "guides.rda"))
load(here("data", "cageminer_input", "snp_positions.rda"))
load(here("data", "cageminer_input", "gcn.rda"))

#----Mine candidates-------------------------------------------------------------
windows <- c(0.01, 0.05, 0.1, 1, 2) # expressed in Mbps

start_time <- Sys.time() # Not required, just to get start time
mined_candidates <- lapply(windows, function(x) {

    message("Sliding window: ", x, " Mbp")

    cand <- mine_candidates(
        gene_ranges = gene_ranges,
        marker_ranges = snp_positions,
        window = x,
        exp = gmax_se,
```

```r
        gcn = gcn,
        guides = guides,
        sample_group = c("Fgr_stress_PI567301B", "Fgr_stress_wyandot")
    )
    return(cand)
})
end_time <- Sys.time() # Not required, just to get end time

# Count number of genes for each interval
candidate_count <- lapply(seq_along(mined_candidates), function(x) {
    count_df <- data.frame(
        window_mb = windows[x],
        ngenes = length(unique(mined_candidates[[x]]$gene))
    )
    return(count_df)
})
candidate_count_cageminer <- Reduce(rbind, candidate_count)
knitr::kable(candidate_count_cageminer, col.names = c("Window (Mb)", "Genes"))
```

Window (Mb)	Genes
0.01	1
0.05	2
0.10	2
1.00	27
2.00	54

As we can see, increasing sliding windows in *cageminer* has the opposite effect of increasing sliding windows in Camoco. The larger the sliding window, the more candidate genes we find. This finding shows that *cageminer* can find high-confidence candidate genes that are not necessarily close to causative SNPs, which includes genes in enhancers and other distal regulatory regions that would have been overlooked by Camoco.

To conclude, let's look at how long it took for *cageminer* to mine candidate genes:

```r
runtime <- end_time - start_time
runtime
## Time difference of 21.31964 secs
```

It took only 21 seconds for *cageminer* to mine candidate genes using 5 different sliding windows.

Session information

This document was created under the following conditions:

```r
sessioninfo::session_info()
## - Session info ---------------------------------------------------------------
##  setting  value
##  version  R version 4.2.1 (2022-06-23)
##  os       Ubuntu 20.04.4 LTS
##  system   x86_64, linux-gnu
```

```
##  ui        X11
##  language  (EN)
##  collate   en_US.UTF-8
##  ctype     en_US.UTF-8
##  tz        Europe/Brussels
##  date      2022-07-15
##  pandoc    2.17.1.1 @ /usr/lib/rstudio/bin/quarto/bin/ (via rmarkdown)
##
## - Packages -----------------------------------------------------------------
##  package            * version   date (UTC) lib source
##  abind                1.4-5     2016-07-21 [1] CRAN (R 4.2.0)
##  annotate             1.74.0    2022-04-26 [1] Bioconductor
##  AnnotationDbi        1.58.0    2022-04-26 [1] Bioconductor
##  AnnotationFilter     1.20.0    2022-04-26 [1] Bioconductor
##  assertthat           0.2.1     2019-03-21 [1] CRAN (R 4.2.0)
##  backports            1.4.1     2021-12-13 [1] CRAN (R 4.2.0)
##  base64enc            0.1-3     2015-07-28 [1] CRAN (R 4.2.0)
##  Biobase              2.56.0    2022-04-26 [1] Bioconductor
##  BiocFileCache        2.4.0     2022-04-26 [1] Bioconductor
##  BiocGenerics         0.42.0    2022-04-26 [1] Bioconductor
##  BiocIO               1.6.0     2022-04-26 [1] Bioconductor
##  BiocManager          1.30.18   2022-05-18 [1] CRAN (R 4.2.0)
##  BiocParallel         1.30.3    2022-06-05 [1] Bioconductor
##  BiocStyle          * 2.25.0    2022-06-15 [1] Github (Bioconductor/BiocStyle@7150c28)
##  biomaRt              2.52.0    2022-04-26 [1] Bioconductor
##  BioNERO              1.4.0     2022-04-26 [1] Bioconductor
##  Biostrings           2.64.0    2022-04-26 [1] Bioconductor
##  biovizBase           1.44.0    2022-04-26 [1] Bioconductor
##  bit                  4.0.4     2020-08-04 [1] CRAN (R 4.2.0)
##  bit64                4.0.5     2020-08-30 [1] CRAN (R 4.2.0)
##  bitops               1.0-7     2021-04-24 [1] CRAN (R 4.2.0)
##  blob                 1.2.3     2022-04-10 [1] CRAN (R 4.2.0)
##  bookdown             0.27      2022-06-14 [1] CRAN (R 4.2.0)
##  broom                0.8.0     2022-04-13 [1] CRAN (R 4.2.0)
##  BSgenome             1.64.0    2022-04-26 [1] Bioconductor
##  cachem               1.0.6     2021-08-19 [1] CRAN (R 4.2.0)
##  cageminer          * 1.2.0     2022-04-26 [1] Bioconductor
##  car                  3.1-0     2022-06-15 [1] CRAN (R 4.2.0)
##  carData              3.0-5     2022-01-06 [1] CRAN (R 4.2.0)
##  checkmate            2.1.0     2022-04-21 [1] CRAN (R 4.2.0)
##  circlize             0.4.15    2022-05-10 [1] CRAN (R 4.2.0)
##  cli                  3.3.0     2022-04-25 [1] CRAN (R 4.2.0)
##  clue                 0.3-61    2022-05-30 [1] CRAN (R 4.2.0)
##  cluster              2.1.3     2022-03-28 [1] CRAN (R 4.2.0)
##  coda                 0.19-4    2020-09-30 [1] CRAN (R 4.2.0)
##  codetools            0.2-18    2020-11-04 [1] CRAN (R 4.2.0)
##  colorspace           2.0-3     2022-02-21 [1] CRAN (R 4.2.0)
##  ComplexHeatmap       2.12.0    2022-04-26 [1] Bioconductor
##  crayon               1.5.1     2022-03-26 [1] CRAN (R 4.2.0)
##  curl                 4.3.2     2021-06-23 [1] CRAN (R 4.2.0)
##  data.table           1.14.2    2021-09-27 [1] CRAN (R 4.2.0)
```

```
##   DBI                1.1.3      2022-06-18 [1] CRAN (R 4.2.0)
##   dbplyr             2.2.1      2022-06-27 [1] CRAN (R 4.2.1)
##   DelayedArray       0.22.0     2022-04-26 [1] Bioconductor
##   DESeq2             1.36.0     2022-04-26 [1] Bioconductor
##   dichromat          2.0-0.1    2022-05-02 [1] CRAN (R 4.2.0)
##   digest             0.6.29     2021-12-01 [1] CRAN (R 4.2.0)
##   doParallel         1.0.17     2022-02-07 [1] CRAN (R 4.2.0)
##   dplyr              1.0.9      2022-04-28 [1] CRAN (R 4.2.0)
##   dynamicTreeCut     1.63-1     2016-03-11 [1] CRAN (R 4.2.0)
##   edgeR              3.38.1     2022-05-15 [1] Bioconductor
##   ellipsis           0.3.2      2021-04-29 [1] CRAN (R 4.2.0)
##   ensembldb          2.20.2     2022-06-16 [1] Bioconductor
##   evaluate           0.15       2022-02-18 [1] CRAN (R 4.2.0)
##   fansi              1.0.3      2022-03-24 [1] CRAN (R 4.2.0)
##   fastcluster        1.2.3      2021-05-24 [1] CRAN (R 4.2.0)
##   fastmap            1.1.0      2021-01-25 [1] CRAN (R 4.2.0)
##   filelock           1.0.2      2018-10-05 [1] CRAN (R 4.2.0)
##   foreach            1.5.2      2022-02-02 [1] CRAN (R 4.2.0)
##   foreign            0.8-82     2022-01-13 [1] CRAN (R 4.2.0)
##   Formula            1.2-4      2020-10-16 [1] CRAN (R 4.2.0)
##   genefilter         1.78.0     2022-04-26 [1] Bioconductor
##   geneplotter        1.74.0     2022-04-26 [1] Bioconductor
##   generics           0.1.2      2022-01-31 [1] CRAN (R 4.2.0)
##   GENIE3             1.18.0     2022-04-26 [1] Bioconductor
##   GenomeInfoDb       1.32.2     2022-05-15 [1] Bioconductor
##   GenomeInfoDbData   1.2.8      2022-05-06 [1] Bioconductor
##   GenomicAlignments  1.32.0     2022-04-26 [1] Bioconductor
##   GenomicFeatures    1.48.3     2022-05-31 [1] Bioconductor
##   GenomicRanges      1.48.0     2022-04-26 [1] Bioconductor
##   GetoptLong         1.0.5      2020-12-15 [1] CRAN (R 4.2.0)
##   GGally             2.1.2      2021-06-21 [1] CRAN (R 4.2.0)
##   ggbio              1.44.1     2022-06-23 [1] Bioconductor
##   ggnetwork          0.5.10     2021-07-06 [1] CRAN (R 4.2.0)
##   ggnewscale         0.4.7      2022-03-25 [1] CRAN (R 4.2.0)
##   ggplot2            3.3.6      2022-05-03 [1] CRAN (R 4.2.0)
##   ggpubr             0.4.0      2020-06-27 [1] CRAN (R 4.2.0)
##   ggsignif           0.6.3      2021-09-09 [1] CRAN (R 4.2.0)
##   ggtext             0.1.1      2020-12-17 [1] CRAN (R 4.2.0)
##   GlobalOptions      0.1.2      2020-06-10 [1] CRAN (R 4.2.0)
##   glue               1.6.2      2022-02-24 [1] CRAN (R 4.2.0)
##   GO.db              3.15.0     2022-05-06 [1] Bioconductor
##   graph              1.74.0     2022-04-26 [1] Bioconductor
##   gridExtra          2.3        2017-09-09 [1] CRAN (R 4.2.0)
##   gridtext           0.1.4      2020-12-10 [1] CRAN (R 4.2.0)
##   gtable             0.3.0      2019-03-25 [1] CRAN (R 4.2.0)
##   here             * 1.0.1      2020-12-13 [1] CRAN (R 4.2.0)
##   Hmisc              4.7-0      2022-04-19 [1] CRAN (R 4.2.0)
##   hms                1.1.1      2021-09-26 [1] CRAN (R 4.2.0)
##   htmlTable          2.4.0      2022-01-04 [1] CRAN (R 4.2.0)
##   htmltools          0.5.2      2021-08-25 [1] CRAN (R 4.2.0)
##   htmlwidgets        1.5.4      2021-09-08 [1] CRAN (R 4.2.0)
```

```
##   httr            1.4.3       2022-05-04 [1] CRAN (R 4.2.0)
##   igraph          1.3.2       2022-06-13 [1] CRAN (R 4.2.0)
##   impute          1.70.0      2022-04-26 [1] Bioconductor
##   intergraph      2.0-2       2016-12-05 [1] CRAN (R 4.2.0)
##   IRanges         2.30.0      2022-04-26 [1] Bioconductor
##   iterators       1.0.14      2022-02-05 [1] CRAN (R 4.2.0)
##   jpeg            0.1-9       2021-07-24 [1] CRAN (R 4.2.0)
##   KEGGREST        1.36.2      2022-06-09 [1] Bioconductor
##   knitr           1.39        2022-04-26 [1] CRAN (R 4.2.0)
##   lattice         0.20-45     2021-09-22 [1] CRAN (R 4.2.0)
##   latticeExtra    0.6-29      2019-12-19 [1] CRAN (R 4.2.0)
##   lazyeval        0.2.2       2019-03-15 [1] CRAN (R 4.2.0)
##   lifecycle       1.0.1       2021-09-24 [1] CRAN (R 4.2.0)
##   limma           3.52.2      2022-06-19 [1] Bioconductor
##   locfit          1.5-9.5     2022-03-03 [1] CRAN (R 4.2.0)
##   magick          2.7.3       2021-08-18 [1] CRAN (R 4.2.0)
##   magrittr        2.0.3       2022-03-30 [1] CRAN (R 4.2.0)
##   Matrix          1.4-1       2022-03-23 [1] CRAN (R 4.2.0)
##   MatrixGenerics  1.8.1       2022-06-26 [1] Bioconductor
##   matrixStats     0.62.0      2022-04-19 [1] CRAN (R 4.2.0)
##   memoise         2.0.1       2021-11-26 [1] CRAN (R 4.2.0)
##   mgcv            1.8-40      2022-03-29 [1] CRAN (R 4.2.0)
##   minet           3.54.0      2022-04-26 [1] Bioconductor
##   munsell         0.5.0       2018-06-12 [1] CRAN (R 4.2.0)
##   NetRep          1.2.4       2020-10-07 [1] CRAN (R 4.2.0)
##   network         1.17.2      2022-05-21 [1] CRAN (R 4.2.0)
##   networkD3       0.4         2017-03-18 [1] CRAN (R 4.2.0)
##   nlme            3.1-158     2022-06-15 [1] CRAN (R 4.2.0)
##   nnet            7.3-17      2022-01-13 [1] CRAN (R 4.2.0)
##   OrganismDbi     1.38.1      2022-06-16 [1] Bioconductor
##   pillar          1.7.0       2022-02-01 [1] CRAN (R 4.2.0)
##   pkgconfig       2.0.3       2019-09-22 [1] CRAN (R 4.2.0)
##   plyr            1.8.7       2022-03-24 [1] CRAN (R 4.2.0)
##   png             0.1-7       2013-12-03 [1] CRAN (R 4.2.0)
##   preprocessCore  1.58.0      2022-04-26 [1] Bioconductor
##   prettyunits     1.1.1       2020-01-24 [1] CRAN (R 4.2.0)
##   progress        1.2.2       2019-05-16 [1] CRAN (R 4.2.0)
##   ProtGenerics    1.28.0      2022-04-26 [1] Bioconductor
##   purrr           0.3.4       2020-04-17 [1] CRAN (R 4.2.0)
##   R6              2.5.1       2021-08-19 [1] CRAN (R 4.2.0)
##   rappdirs        0.3.3       2021-01-31 [1] CRAN (R 4.2.0)
##   RBGL            1.72.0      2022-04-26 [1] Bioconductor
##   RColorBrewer    1.1-3       2022-04-03 [1] CRAN (R 4.2.0)
##   Rcpp            1.0.8.3     2022-03-17 [1] CRAN (R 4.2.0)
##   RCurl           1.98-1.7    2022-06-09 [1] CRAN (R 4.2.0)
##   reshape         0.8.9       2022-04-12 [1] CRAN (R 4.2.0)
##   reshape2        1.4.4       2020-04-09 [1] CRAN (R 4.2.0)
##   restfulr        0.0.15      2022-06-16 [1] CRAN (R 4.2.0)
##   RhpcBLASctl     0.21-247.1  2021-11-05 [1] CRAN (R 4.2.0)
##   rjson           0.2.21      2022-01-09 [1] CRAN (R 4.2.0)
##   rlang           1.0.3       2022-06-27 [1] CRAN (R 4.2.1)
```

```
##   rmarkdown            2.14      2022-04-25 [1] CRAN (R 4.2.0)
##   rpart                4.1.16    2022-01-24 [1] CRAN (R 4.2.0)
##   rprojroot            2.0.3     2022-04-02 [1] CRAN (R 4.2.0)
##   Rsamtools            2.12.0    2022-04-26 [1] Bioconductor
##   RSQLite              2.2.14    2022-05-07 [1] CRAN (R 4.2.0)
##   rstatix              0.7.0     2021-02-13 [1] CRAN (R 4.2.0)
##   rstudioapi           0.13      2020-11-12 [1] CRAN (R 4.2.0)
##   rtracklayer          1.57.0    2022-06-16 [1] Github (lawremi/rtracklayer@2bb0b40)
##   S4Vectors            0.34.0    2022-04-26 [1] Bioconductor
##   scales               1.2.0     2022-04-13 [1] CRAN (R 4.2.0)
##   sessioninfo          1.2.2     2021-12-06 [1] CRAN (R 4.2.0)
##   shape                1.4.6     2021-05-19 [1] CRAN (R 4.2.0)
##   statmod              1.4.36    2021-05-10 [1] CRAN (R 4.2.0)
##   statnet.common       4.6.0     2022-05-02 [1] CRAN (R 4.2.0)
##   stringi              1.7.6     2021-11-29 [1] CRAN (R 4.2.0)
##   stringr              1.4.0     2019-02-10 [1] CRAN (R 4.2.0)
##   SummarizedExperiment 1.26.1    2022-04-29 [1] Bioconductor
##   survival             3.3-1     2022-03-03 [1] CRAN (R 4.2.0)
##   sva                  3.44.0    2022-04-26 [1] Bioconductor
##   tibble               3.1.7     2022-05-03 [1] CRAN (R 4.2.0)
##   tidyr                1.2.0     2022-02-01 [1] CRAN (R 4.2.0)
##   tidyselect           1.1.2     2022-02-21 [1] CRAN (R 4.2.0)
##   utf8                 1.2.2     2021-07-24 [1] CRAN (R 4.2.0)
##   VariantAnnotation    1.42.1    2022-05-15 [1] Bioconductor
##   vctrs                0.4.1     2022-04-13 [1] CRAN (R 4.2.0)
##   WGCNA                1.71      2022-04-22 [1] CRAN (R 4.2.0)
##   xfun                 0.31      2022-05-10 [1] CRAN (R 4.2.0)
##   XML                  3.99-0.10 2022-06-09 [1] CRAN (R 4.2.0)
##   xml2                 1.3.3     2021-11-30 [1] CRAN (R 4.2.0)
##   xtable               1.8-4     2019-04-21 [1] CRAN (R 4.2.0)
##   XVector              0.36.0    2022-04-26 [1] Bioconductor
##   yaml                 2.3.5     2022-02-21 [1] CRAN (R 4.2.0)
##   zlibbioc             1.42.0    2022-04-26 [1] Bioconductor
##
## [1] /home/faalm/R/x86_64-pc-linux-gnu-library/4.2
## [2] /usr/local/lib/R/site-library
## [3] /usr/lib/R/site-library
## [4] /usr/lib/R/library
##
## --------------------------------------------------------------------------
```

References

Almeida-Silva, Fabricio, and Thiago M Venancio. 2021. "Integration of Genome-Wide Association Studies and Gene Coexpression Networks Unveils Promising Soybean Resistance Genes Against Five Common Fungal Pathogens." *Scientific Reports* 11 (1): 1–10.

Michno, Jean-Michel, Junqi Liu, Joseph R Jeffers, Robert M Stupar, and Chad L Myers. 2020. "Identification of Nodulation-Related Genes in Medicago Truncatula Using Genome-Wide Association Studies and Co-Expression Networks." *Plant Direct* 4 (5): e00220.

Supplementary Text S2 - cageminer: an R/Bioconductor package to prioritize candidate genes by integrating GWAS and gene coexpression networks

Schaefer, Robert J, Jean-Michel Michno, Joseph Jeffers, Owen Hoekenga, Brian Dilkes, Ivan Baxter, and Chad L Myers. 2018. "Integrating Coexpression Networks with GWAS to Prioritize Causal Genes in Maize." *The Plant Cell* 30 (12): 2922–42.