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

Fabricio Almeida-Silva and Thiago M. Venancio

Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro, Campos dos Goytacazes, RJ, Brazil.

*FA-S: Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro. Av. Alberto Lamego 2000, P5, sala 217, Campos dos Goytacazes, RJ, Brazil. Email: fabricio_almeidasilva@hotmail.com

*TMV: Laboratório de Química e Função de Proteínas e Peptídeos, Centro de Biociências e Biotecnologia, Universidade Estadual do Norte Fluminense Darcy Ribeiro. Av. Alberto Lamego 2000, P5, sala 217, Campos dos Goytacazes, RJ, Brazil. Email: thiago.venancio@gmail.com

© The Author(s) 2022. Published by Oxford University Press on behalf of the Annals of Botany Company. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
ABSTRACT

Summary
Although genome-wide association studies (GWAS) identify variants associated with traits of interest, they often fail in identifying causative genes underlying a given phenotype. Integrating GWAS and gene coexpression networks can help prioritize high-confidence candidate genes, as the expression profiles of trait-associated genes can be used to mine novel candidates. Here, we present cageminer, an R package to prioritize candidate genes through the integration of GWAS and coexpression networks. Genes are considered high-confidence candidates if they pass all three filtering criteria implemented in cageminer, namely physical proximity to (or linkage disequilibrium with) SNPs, coexpression with known trait-associated genes, and significant changes in expression levels in conditions of interest. Prioritized candidates can also be scored and ranked to select targets for experimental validation. By applying cageminer to a real data set of Capsicum annuum response to Phytophthora infection (RNA-seq and SNPs from an association panel), we demonstrate that it can effectively prioritize candidates, leading to a significant reduction in candidate gene lists.

Availability and implementation
The package is available at Bioconductor (http://bioconductor.org/packages/cageminer).

Keywords: gene discovery, multi-omics data integration, systems biology, bioinformatics, population genomics.
1 Introduction

Over the years, several genome-wide association studies (GWAS) have identified single-nucleotide polymorphisms (SNPs) associated with phenotypes of interest, such as agronomic traits in crops, production traits in livestock, and complex human disorders (Boudhrioua et al., 2020; Maldonado Dos Santos et al., 2019; Wu et al., 2020; Buzanskas et al., 2014). However, finding causative genes from SNPs remains a major bottleneck (Baxter, 2020). First, most GWAS-derived SNPs are located in non-coding portions of the genome, which can be regulatory regions that are very far from a causative gene (Peat et al., 2020). Further, causative variants can be in strong linkage disequilibrium (LD) with non-causative ones, leading to large LD blocks with dozens of putative candidates (Michno et al., 2020).

To address this issue, integrating GWAS with the vast amounts of RNA-seq data in public repositories has become a promising solution, particularly using gene coexpression network (GCN)-based approaches (Michno et al., 2020; Yao et al., 2020; Guo et al., 2020). Currently, the only statistical framework that automates such integration is Camoco, a Python library that identifies sets of densely connected genes for a given sliding window relative to each SNP (Schaefer et al., 2018). However, as sliding windows are expanded (e.g., 50 kb), Camoco loses the ability to discover candidate genes because of background noise (Michno et al., 2020). This is a major limitation, as SNPs can be up to 2 Mb away from the causative genes if they are in distal regions (Brodie et al., 2016).

Here, we present cageminer (candidate gene miner), an R/Bioconductor package that integrates GCNs and GWAS-derived SNPs to prioritize candidate genes associated with traits of interest. cageminer uses a guide gene-based approach to discover novel candidates that are coexpressed with known trait-associated genes and that are significantly induced or repressed in conditions of interest. By relying on researchers’ prior knowledge, cageminer can identify high-confidence candidate genes even in megabase-scale genomic intervals. This package will be instrumental in helping researchers discover genes underlying important quantitative traits.

2 Implementation

Cageminer is implemented as an R package, and all input and output objects belong to base R or common Bioconductor classes to ensure interoperability with other packages. Our algorithm requires three types of input data: i. SNP positions, which must be passed as GRanges or GRangesList objects (for single trait and multiple traits, respectively) (Lawrence et al., 2013); ii. guide genes, either as a character vector or a data frame; and iii. gene coexpression network, which must be passed as a list as returned by the function exp2gcn() from the Bioconductor package BioNERO (Almeida-Silva and Venancio, 2022).
2.1 Algorithm description

cageminer identifies high-confidence candidate genes in three sequential steps (Fig. 1). In the first step, all genes within a sliding window (default: 2 Mb) relative to each SNP are selected as putative candidates. The default 2 Mb sliding window aims to minimize false-negative rates, as SNPs can be located in distal regions (Brodie et al., 2016). If the 2 Mb window returns too many genes to start with, users can simulate different window sizes and visualize the number of genes in a line plot (Supplementary Text). Additionally, users can input a custom interval for each SNP (e.g., based on linkage disequilibrium) by disabling the sliding window expansion.

For the second step of the algorithm, cageminer relies on the module_enrichment() function from the BioNERO package (Almeida-Silva and Venancio, 2022) to perform an enrichment analysis and find candidates from step 1 that co-occur in modules enriched in guide genes. Here, modules are defined as unique clusters of coexpressed genes. Guides are genes known to be associated with the phenotype of interest, which can be passed as a single gene set in a character vector or as a 2-column data frame with gene IDs in the first column and gene classification (e.g., Gene Ontology Terms or KEGG pathways) in the second column. In the latter case, cageminer will look for modules enriched in each class of guide genes rather than guides in general.

In the third step, the gene expression matrix used to infer the GCN is correlated to a binary matrix $m_{ij}$ containing 1 if the sample $m$ corresponds to the condition $j$, and 0 otherwise. This calculation, also known as gene significance, returns a point biserial correlation coefficient ($r_{pb}$) (Langfelder and Horvath, 2008) that indicates if genes have significantly increased or decreased expression levels in a particular condition ($H_0: r_{pb} = 0$). Furthermore, as genes can be negative regulators of the phenotype of interest, negative correlation coefficients are also treated as biologically meaningful. Thus, the absolute value of $r_{pb}$ is considered to define a gene significance threshold, as well as Student asymptotic $P$-values for correlation significance (by default, $r_{pb} \geq 0.2$ and $P < 0.05$).

2.2 Gene scoring

As an optional and additional step, users can score the prioritized candidate genes and further select the top $n$ genes for validation. Genes can be scored with the formula below:

$$S_i = r_{pb} \kappa$$

where

- $\kappa = 2$ if the gene is a transcription factor
- $\kappa = 2$ if the gene is a hub
- $\kappa = 3$ if the gene is a hub and a transcription factor
- $\kappa = 1$ if the gene is neither a hub nor a transcription factor
In BioNERO’s dictionary, hubs are defined as the intersect between the top 10% genes with highest connectivity and genes that have module membership >0.8. The weights given to transcription factors, hubs, and hub transcription factors in the $\kappa$ variable can be modified by users. Guide genes from step 2 of the pipeline can be selected and have their scores used as a reference.

3 Application to real datasets

3.1 Pepper response to Phytophthora root rot

A use case using RNA-seq on pepper (Capsicum annuum) response to Phytophthora root rot (Kim et al., 2018), as well as GWAS-derived SNPs associated with resistance to Phytophthora root rot (Siddique et al., 2019) is available in Supplementary Text S1. Pepper genes encoding transcription factors were downloaded from PlantTFDB 4.0 (Jin et al., 2017), and plant defense-related genes (MapMan annotations) were obtained from PLAZA Dicots 3.0 (Proost et al., 2015). From a list of 1265 putative candidates, cageminer identified 7 high-confidence candidate resistance genes (99.4% reduction).

All prioritized candidates encode proteins related to known plant immunity-related processes (e.g., immune signaling, oxidative stress, and lignan biosynthesis), supporting the effectiveness of the algorithm in finding biologically meaningful genes. For instance, CA10g02780 encodes an acyl-CoA oxidase, a required protein for the biosynthesis of the defense-related hormone jasmonic acid in plant peroxisomes (Schilmiller et al., 2007). Likewise, CA02g16620 encodes a UDP-glycosyltransferase, a stress-responsive class of proteins involved in phytohormone signaling (Rehman et al., 2018; Blanco-Herrera et al., 2015).

3.2 Assessing the impact of increasing sliding windows on candidate gene mining

Michno et al. (2020) demonstrated that Camoco can only identify genes that are close to causative SNPs, as increasing sliding windows relative to each SNP dramatically increases the background noise, hence making Camoco detect no genes. To assess if increasing sliding windows makes cageminer lose power to mine candidates, as it occurs for Camoco, we ran cageminer with increasing sliding windows using data on soybean resistance to Fusarium graminearum obtained from (Almeida-Silva and Venancio, 2021).

We observed that increasing sliding windows makes cageminer detect more genes, the opposite of what happens with Camoco (Supplementary Text S2). The cageminer algorithm identified 1, 2, 2, 27 and 54 high-confidence candidate genes for sliding windows of 10 kb, 50 kb, 100 kb, 1 Mb, and 2 Mb, respectively. As cageminer uses guide genes to narrow down the list of putative candidate genes, larger sliding windows do not lead to high noise. Hence, while Camoco only works if SNPs are inside genes or in promoter regions, cageminer is the only existing tool that also considers SNPs in distal regulatory regions such as enhancers or silencers.
4 Conclusions

cageminer is an R package that integrates GWAS-derived SNPs and gene coexpression networks to prioritize candidate genes involved in phenotypes of interest. This package will likely contribute to the advancement of population genomics and to the identification of genes for biotechnological applications.
Data Availability

All data and code used in this manuscript are available in a GitHub repository (https://github.com/almeidasilvaf/cageminer_benchmark) to ensure full reproducibility.

Figure legends

Figure 1. Summary of the cageminer algorithm. Candidate gene prioritization is performed in three sequential steps that can be run as a pipeline (recommended) or independently. The steps can be interpreted as different sources of evidence that candidates are causative genes. Thus, candidates that pass all three steps are considered high-confidence candidates.

Supporting Information

Supplementary Text S1: Application of cageminer to a real data set on pepper (Capsicum annuum) infected with Phytophthora sp.

Supplementary Text S2: Assessing the impact of increasing sliding windows on candidate gene identification.

Contributions by the authors

Conceived the study: TMV and FA-S. Software development: FA-S. Data analysis: FA-S. Funding, project coordination and infrastructure: TMV. Manuscript writing: FA-S and TMV.

Acknowledgements

This work was supported by Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ; grants E-26/203.309/2016 and E-26/203.014/2018), Coordenação de Aperfeicoamento de Pessoal de Nível Superior - Brasil (CAPES; Finance Code 001), and Conselho Nacional de Desenvolvimento Científico e Tecnológico. The funding agencies had no role in the design of the study and collection, analysis, and interpretation of data and in writing.

Conflicts of interest: none declared.
REFERENCES

Almeida-Silva, F. and Venancio, T.M. (2022). BioNERO: an all-in-one R/Bioconductor package for comprehensive and easy biological network reconstruction. Functional & Integrative Genomics 22: 131–136.

Almeida-Silva, F. and Venancio, T.M. (2021). Integration of genome-wide association studies and gene coexpression networks unveils promising soybean resistance genes against five common fungal pathogens. Scientific Reports 11.

Baxter, I. (2020). We aren’t good at picking candidate genes, and it’s slowing us down. Current Opinion in Plant Biology 54: 57–60.

Blanco-Herrera, F., Moreno, A.A., Tapia, R., Reyes, F., Araya, M., D’Alessio, C., Parodi, A., and Orellana, A. (2015). The UDP-glucose: Glycoprotein glucosyltransferase (UGGT), a key enzyme in ER quality control, plays a significant role in plant growth as well as biotic and abiotic stress in Arabidopsis thaliana. BMC Plant Biology 15.

Boudhrioua, C., Bastien, M., Torkamaneh, D., and Belzile, F. (2020). Genome-wide association mapping of Sclerotinia sclerotiorum resistance in soybean using whole-genome resequencing data. BMC Plant Biol 20: 195.

Brodie, A., Azaria, J.R., and Ofran, Y. (2016). How far from the SNP may the causative genes be? Nucleic Acids Research 44: 6046–6054.

Buzanskas, M.E. et al. (2014). Genome-Wide Association for Growth Traits in Canchim Beef Cattle. PLOS ONE 9: e94802.

Guo, J., Li, C., Zhang, X., Li, Y., Zhang, D., Shi, Y., Song, Y., Li, Y., Yang, D., and Wang, T. (2020). Transcriptome and GWAS analyses reveal candidate gene for seminal root length of maize seedlings under drought stress. Plant Science 292.

Jin, J., Tian, F., Yang, D.C., Meng, Y.Q., Kong, L., Luo, J., and Gao, G. (2017). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res 45: D1040–D1045.

Kim, M.S., Kim, S., Jeon, J., Kim, K.T., Lee, H.A., Lee, H.Y., Park, J., Seo, E., Kim, S.B., Yeom, S.I., Lee, Y.H., and Choi, D. (2018). Global gene expression profiling for fruit organs and pathogen infections in the pepper, Capsicum annuum L. Scientific Data 5: 1–6.

Langfelder, P. and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9: 559.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for Computing and Annotating Genomic Ranges. PLoS Computational Biology 9: 1–10.

Maldonado Dos Santos, J.V., Ferreira, E.G.C., Passianotto, A.L.D.L., Brumer, B.B., Santos, A.B. Dos Soares, R.M., Torkamaneh, D., Arías, C.A.A., Belzile, F., Abdelnoor, R.V., and Marcelino-
Guimarães, F.C. (2019). Association mapping of a locus that confers southern stem canker resistance in soybean and SNP marker development. BMC Genomics 20: 1–13.

Michno, J.M., Liu, J., Jeffers, J.R., Stupar, R.M., and Myers, C.L. (2020). Identification of nodulation-related genes in Medicago truncatula using genome-wide association studies and co-expression networks. Plant Direct 4: 1–10.

Peat, G., Jones, W., Nuhn, M., Marugán, J.C., Newell, W., Dunham, I., and Zerbino, D. (2020). The open targets post-GWAS analysis pipeline. Bioinformatics: 1–2.

Proost, S., Van Bel, M., Vieitezheoutte, D., Van de Peer, Y., Inzé, D., Mueller-Roeber, B., and Vandepoele, K. (2015). PLAZA 3.0: an access point for plant comparative genomics. Nucleic Acids Research 43: D974–D981.

Rehman, H.M., Nawaz, M.A., Shah, Z.H., Ludwig-Müller, J., Chung, G., Ahmad, M.Q., Yang, S.H., and Lee, S.I. (2018). Comparative genomic and transcriptomic analyses of Family-1 UDP glycosyltransferase in three Brassica species and Arabidopsis indicates stress-responsive regulation. Scientific Reports 8.

Schaefer, R.J., Michno, J.-M., Jeffers, J., Hoekenga, O., Dilkes, B., Baxter, I., and Myers, C.L. (2018). Integrating Coexpression Networks with GWAS to Prioritize Causal Genes in Maize. The Plant Cell 30: 2922–2942.

Schilmiller, A.L., Koo, A.J.K., and Howe, G.A. (2007). Functional diversification of acyl-coenzyme A oxidases in jasmonic acid biosynthesis and action. Plant Physiology 143: 812–824.

Siddique, M.I., Lee, H.Y., Ro, N.Y., Han, K., Venkatesh, J., Solomon, A.M., Patil, A.S., Changkwian, A., Kwon, J.K., and Kang, B.C. (2019). Identifying candidate genes for Phytophthora capsici resistance in pepper (Capsicum annuum) via genotyping-by-sequencing-based QTL mapping and genome-wide association study. Scientific Reports 9: 1–15.

Wu, Y. et al. (2020). Multi-trait analysis for genome-wide association study of five psychiatric disorders. Translational Psychiatry 10: 209.

Yao, M. et al. (2020). GWAS and co-expression network combination uncovers multigenes with close linkage effects on the oleic acid content accumulation in Brassica napus. BMC Genomics 21: 1–12.
Figure 1

1. Select all genes within interval $i$

$I = SNP \pm window$

Ref: ATTCGCCTA
Test: ATTCACCCCTA

2. Select candidates from 1 in modules enriched in guide genes

$P < \alpha$

3. Select candidates from 2 correlated with conditions of interest

$|r_{\alpha}| \geq \text{threshold}$