

# Visual integration of GWAS and differential expression results with the `hidecan` R package

Olivia Angelin-Bonnet<sup>1</sup>, Matthieu Vignes<sup>2</sup>, Patrick J. Biggs<sup>3,4</sup>, Samantha Baldwin<sup>5</sup>,  
and Susan Thomson<sup>5</sup>

<sup>1</sup>The New Zealand Institute for Plant and Food Research Limited, Palmerston North, 4442, New Zealand

<sup>2</sup>School of Mathematical and Computational Sciences, Massey University, Palmerston North, 4442, New Zealand

<sup>3</sup>School of Natural Sciences, Massey University, Palmerston North, 4442, New Zealand

<sup>4</sup>School of Veterinary Science, Massey University, Palmerston North, 4442, New Zealand

<sup>5</sup>The New Zealand Institute for Plant and Food Research Limited, Christchurch, 8140, New Zealand

March 2023

## Abstract

**Summary:** We present `hidecan`, an R package for generating visualisations that summarise the results of one or more genome-wide association studies and differential expression analyses, as well as manually curated candidate genes, e.g. extracted from the literature.

**Availability and Implementation:** The `hidecan` package is implemented in R and is publicly available on the CRAN repository (<https://CRAN.R-project.org/package=hidecan>) and on GitHub (<https://github.com/PlantandFoodResearch/hidecan>). A description of the package, as well as a detailed tutorial are available at <https://plantandfoodresearch.github.io/hidecan/>.

**Contact:** [olivia.angelin-bonnet@plantandfood.co.nz](mailto:olivia.angelin-bonnet@plantandfood.co.nz).

**Supplementary information:** Supplementary data are available.

## 1 Introduction

Genome-wide association studies (GWAS) enable researchers to investigate genomic regions associated with a trait of interest [3]. Often, the ultimate goal of such analyses is to highlight causal genes that are involved in biological processes relating to the phenotype under study; for example to detect disease-related genes that can be targeted by new drugs, or facilitate selection or desirable characteristics in breeding programmes. Therefore, association studies can be complemented with the acquisition of transcriptomics data [15, 17]. These can be used to detect genes whose expression is associated with the trait under study through a differential expression (DE) analysis. The results from both GWAS and DE are typically compared with previous results from the literature, in order to assess which findings are substantiated by previous research, and which are novel results that must be confirmed with follow-up studies.

GWAS results are usually displayed through a Manhattan plot, which represents the score of each genomic variant (i.e.  $-\log_{10}$  of its p-value) against its physical position along the genome [14]. DE results are instead typically represented with a volcano plot, which displays the DE score of the genes against their  $\log_2$ -fold change [7]. The drawback of using these plots is that they cannot be easily supplemented

with manual information, for example genes of interest from the literature. In addition, Manhattan plots and volcano plots cannot be combined to represent both GWAS and DE results in a single figure. In recent years, the Circos plot has emerged as an attractive design to display information related to genomics data [6], and thus can be used to visualise the results of both GWAS and DE. It is notably possible to combine several layers of information by stacking several circles on top of each other in a single Circos plot [5, 17]. However, the circular layout of Circos plots makes it difficult to precisely compare the different layers and thus to identify genomic regions of interest in which the information from several layers overlap.

We recently proposed a new visualisation, the HIDECAN plot, that integrates GWAS and DE results as well as candidate genes (from the literature) into one graphic [1]. In this paper, we present the corresponding `hidecan` R package, which can be used to generate HIDECAN plots. The package is publicly available on the CRAN repository (<https://CRAN.R-project.org/package=hidecan>) and on GitHub (<https://github.com/PlantandFoodResearch/hidecan>), with a detailed tutorial available at <https://plantandfoodresearch.github.io/hidecan/>.

## 2 Implementation

The main function of the `hidecan` package is the `hidecan_plot()` function. It takes as input data-frames of GWAS results, DE results, and candidate gene lists. These data-frames should contain information about the genomic position of the genomic variants and genes (i.e. chromosome and physical position or start and end, in base pairs), as well as GWAS or DE scores and log<sub>2</sub>-fold change. The user needs to set a significance threshold on the GWAS scores, DE scores and DE log<sub>2</sub>-fold change in order to select genomics variants and genes that will be displayed in the HIDECAN plot. Several other parameters included in the `hidecan_plot()` function allow the user to control different aspects of the plot, such as which chromosomes should be displayed, or the size of the points in the graphic. The function then returns a HIDECAN plot constructed with the `ggplot2` [16], `ggrepel` [12] and `ggnewscale` [4] packages (Figure 1 A-D). Supplementary Material 1 details the different steps of the `hidecan_plot()` function to generate a HIDECAN plot from the input datasets.

Alternatively, a Shiny app is made available, through which the user can construct a HIDECAN plot via a visual interface rather than with code, by providing as input csv files of GWAS results, DE results and candidate genes lists. The generated HIDECAN plot can be saved as either a PNG or PDF file.

Often, the GWAS and DE analyses are performed using popular R packages such as `edgeR` [10] or `DESeq2` [8]. For example, `GWASpoly` [11] is an R package for performing GWAS for autotetraploid organisms. Through the `GWASpoly` package, it is possible to obtain GWAS scores for several traits at once and with different genetic models. The package then offers several options to compute a significance threshold on the genomic variants score, for each trait and genetic model. In order to facilitate the construction of HIDECAN plots from `GWASpoly` results, we are offering a helper function to extract the necessary information directly from the output of the `GWASpoly` package. The `hidecan_plot_from_gwaspoly()` function reads in the output from the `GWASpoly` package and extracts variants scores and significance thresholds to produce a HIDECAN plot. Such helper functions reduce the amount of data wrangling that users must perform to convert the output of these packages into a suitable format for the `hidecan` package. Due to the modular construction of the package, it is possible to easily implement additional helper functions for other diploid- or polyploid-focused packages if the need arises.

The `hidecan` package has also been included in the visualisation tool `VIEWpoly` [13]. `VIEWpoly` is an R package and Shiny app for visualising and integrating results from polyploid genetic and genomic

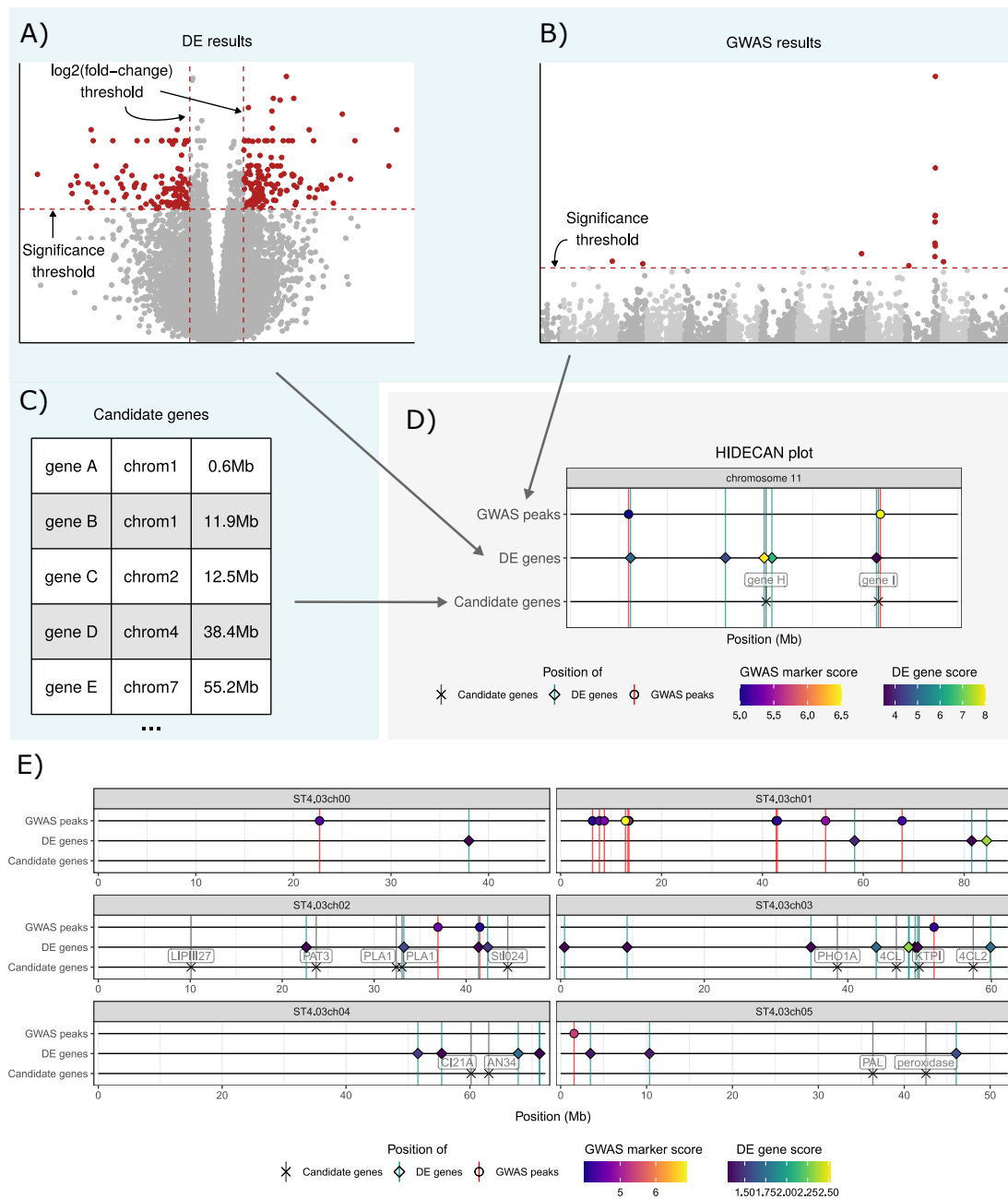


Figure 1: A) to D): Schema of the construction of a HIDECAN plot. The blue box indicates data and plots obtained outside of the hidecan package. A) From a table of differential expression results (represented here with a volcano plot), differentially expressed genes are selected by applying a threshold on the genes' scores and log<sub>2</sub>-fold change. B) Similarly, interesting genomic markers are selected from the GWAS results (here represented with a Manhattan plot) by applying a significance threshold on their association score. C) Lastly, a manually curated table of candidate genes of interest can be specified. D) The HIDECAN plot displays the genomic position (x-axis) of the selected genomic variants, differentially expressed genes and candidate genes along each chromosome. The colour of the points represent the GWAS and DE scores. In this example, only chromosome 11 is represented for better clarity. E) A HIDECAN plot generated with the package, from GWAS and DE results on a potato tuber bruising dataset (only chromosomes 0 to 5 are represented for better clarity, see Supplementary Material for the full figure).

analysis tools. It provides an intuitive interface that facilitates the interactive exploration of linkage mapping and QTL analysis results, obtained from a suite of packages such as MAPpoly [9] or polyqtlR [2]. Users can explore the QTL results and linkage maps in more detail, or use the implemented genome browser to explore the relationships between QTL regions and a corresponding genome annotation. In the newest version of the package (v0.4.0), it is possible to upload GWAS and DE results as well as candidate gene lists as csv files, or GWASpoly results as RData files, into the VIEWpoly Shiny app. A HIDECAN plot is generated from these input files, using the hidecan package. Users can further explore and customise the HIDECAN plot and save the resulting figure as a file with one of several supported extensions.

### 3 Examples

We demonstrate the use of the hidecan package with a dataset from a breeding population of tetraploid potatoes, originally published in [1]. A GWAS analysis was applied to 72,847 genomic variants obtained from 158 progeny plants from a half-sibling breeding population; the results were compared with candidate genes from previous GWAS or QTL mapping studies on potato tuber bruising. In addition, a DE analysis was used to compare the expression of 25,163 transcribed genes between low- and high-bruising tubers two hours after bruising. A subset of these results was used to generate a HIDECAN plot. Figure 1 E showcases the resulting HIDECAN plot for chromosomes 0 to 5. The full HIDECAN plot, as well as the code used to generate the figure and details about the dataset, are presented in Supplementary Material 2. The data used in this example (i.e. result from the GWAS and DE analyses, as well as the list of candidate genes) is available in the hidecan package through the `get_example_data()` function.

In addition, we showcase how the package can be used to visualise the output from the GWASpoly package without any data manipulation. We use the example dataset from the GWASpoly publication [11], in which genomic information and phenotype measurements are obtained for 221 tetraploid potato lines from the SolCAP diversity panel. We focus on the analysis of tuber eye depth, shape and sucrose content, and perform a GWAS analysis with four different genetic models. The code used for this example and resulting HIDECAN plot are presented in Supplementary Material 2.

### 4 Conclusion

We present the hidecan R package for generating HIDECAN plots that display genomic variants and genes highlighted by GWAS and DE analyses, alongside manually curated candidate genes of interest. The HIDECAN plot provides in a single visualisation a genome-wide overview of these three types of information. Critically, it allows users to easily identify genomic regions in which association with a trait of interest is supported at both the genomics and transcriptomics level, and/or supported by previous findings. The hidecan R package provides simple functions that allow the user to rapidly generate such HIDECAN plots. It also offers a Shiny app which allows users to construct a plot through an interactive web-based interface rather than through code. Finally, it provides helper functions that extract results from the output of packages performing the GWAS or DE analysis, to facilitate the use of the package in an R analysis pipeline.

### 5 Acknowledgements

The authors would like to thank Cristiane Hayumi Taniguti, Oscar Riera-Lizarazu, David Byrne, Marcelo Mollinari from the Tools for Polyploids project for their help and for the opportunity to include the

hidecan package into VIEWpoly.

## References

- [1] O. Angelin-Bonnet, S. Thomson, M. Vignes, P. J. Biggs, K. Monaghan, R. Bloomer, K. Wright, and S. Baldwin. Investigating the genetic components of tuber bruising in a breeding population of tetraploid potatoes. Unpublished, 2023.
- [2] P. M. Bourke, R. E. Voorrips, C. A. Hackett, G. van Geest, J. H. Willemsen, P. Arens, M. J. M. Smulders, R. G. F. Visser, and C. Maliepaard. Detecting quantitative trait loci and exploring chromosomal pairing in autopolyploids using polyqtLR. *Bioinformatics*, 37(21):3822–3829, 2021.
- [3] W. S. Bush and J. H. Moore. Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology*, 8(12):e1002822, 2012.
- [4] E. Campitelli. *ggnewscale: Multiple Fill and Colour Scales in 'ggplot2'*, 2022. R package version 0.4.8.
- [5] V. Jha, G. Singh, S. Kumar, A. Sonawane, A. Jere, and K. Anamika. CGDV: a webtool for circular visualization of genomics and transcriptomics data. *BMC Genomics*, 18(1), 2017.
- [6] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. Jones, and M. Marra. Circos: an information aesthetic for comparative genomics. *Genome Res*, 19:1639–45, 2009.
- [7] W. Li. Volcano plots in analyzing differential expressions with mRNA microarrays. *Journal of Bioinformatics and Computational Biology*, 10(06):1231003, 2012.
- [8] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 2014.
- [9] M. Mollinari, B. A. Olukolu, G. da S Pereira, A. Khan, D. Gemenet, G. C. Yench, and Z.-B. Zeng. Unraveling the Hexaploid Sweetpotato Inheritance Using Ultra-Dense Multilocus Mapping. *G3 Genes|Genomes|Genetics*, 10(1):281–292, 2020.
- [10] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2009.
- [11] U. R. Rosyara, W. S. D. Jong, D. S. Douches, and J. B. Endelman. Software for Genome-Wide Association Studies in Autopolyploids and Its Application to Potato. *The Plant Genome*, 9(2), 2016.
- [12] K. Slowikowski. *ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'*, 2022. R package version 0.9.2.
- [13] C. H. Taniguti, G. de Siqueira Gesteira, J. Lau, G. da Silva Pereira, Z.-B. Zeng, D. Byrne, O. Riera-Lizarazu, and M. Mollinari. VIEWpoly: a visualization tool to integrate and explore results of polyploid genetic analysis. *Journal of Open Source Software*, 7(74):4242, 2022.
- [14] S. Turner. qqman: an r package for visualizing gwas results using q-q and manhattan plots. *The Journal of Open Source Software*, 2018.
- [15] L. Wei, H. Jian, K. Lu, F. Filardo, N. Yin, L. Liu, C. Qu, W. Li, H. Du, and J. Li. Genome-wide association analysis and differential expression analysis of resistance to *Sclerotinia* stem rot in *Brassica napus*. *Plant Biotechnology Journal*, 14(6):1368–1380, 2015.

- [16] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer New York, 2016.
- [17] Y. Yuan, H. Xing, W. Zeng, J. Xu, L. Mao, L. Wang, W. Feng, J. Tao, H. Wang, H. Zhang, Q. Wang, G. Zhang, X. Song, and X.-Z. Sun. Genome-wide association and differential expression analysis of salt tolerance in *Gossypium hirsutum* L at the germination stage. *BMC Plant Biology*, 19(1), 2019.