GeneticsMakie.jl: A versatile and scalable toolkit for visualizing locus-level genetic and genomic data

Minsoo Kim[1-3*], Connor T. Jops[1], Michael J. Gandal[1-3*]

[1] Department of Psychiatry, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, 90095, USA.
[2] Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, 90095, USA.
[3] Program in Neurobehavioral Genetics, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, 90095, USA.
* Correspondence to: minsookim@mednet.ucla.edu, mgandal@mednet.ucla.edu

Abstract

With the continued deluge of results from genome-wide association and functional genomic studies, it has become increasingly imperative to quickly combine and visualize different layers of genetic and genomic data within a given locus to facilitate exploratory and integrative data analyses. While several tools have been developed to visualize locus-level genetic results, the limited speed, scalability, and flexibility of current approaches remains a significant bottleneck. Here, we present a Julia package for high-performance genetics and genomics-related data visualization that enables fast, simultaneous plotting of hundreds of association results along with multiple relevant genomic annotations. Leveraging the powerful plotting and layout utilities from Makie.jl facilitates the customization and extensibility of every component of a plot, enabling generation of publication-ready figures.

Introduction

The last decade has seen an exponential increase in the volume of large-scale genetic association results such as those from genome-wide association studies (GWAS) and phenome-wide association studies (PheWAS). The rapid advancements in high-throughput sequencing technologies have further led to an increase in the volume and diversity of molecular genomic readouts, such as 3D genome contacts, ChIP-seq, and ATAC-seq. As these efforts continue to scale, it is becoming increasingly critical to develop efficient ways for simultaneous visualization and integration of multiple such datasets to develop an intuitive understanding of potential underlying biological relationships.

Several tools have been developed to visualize genetic association results within a specific locus along with corresponding genomic annotations, exemplified by the original "LocusZoom" style plots (Pruim et al., 2010). Multiple extensions to these LocusZoom style plots have since been built, spanning a wide array of programming languages, including JavaScript, Python, and R (Jorgenson *et al.*, 2009; Pruim *et al.*, 2010; Boughton *et al.*, 2021; Machiela and Chanock, 2018; Dadaev *et al.*, 2016; Machiela and Chanock, 2015; Juliusdottir *et al.*, 2018; Kramer *et al.*, 2021; Schilder *et al.*, 2021). However, efficient customization and extension is limited with these tools, and in general they are not suitable for parallel visualization of large numbers of data points. For example, only a certain genomic range can be shown or a certain number of genes can be plotted (see **Table 1** for more complete comparisons). Certain packages have a lot of dependencies which can further lead to decreased flexibility.

Ideally, one would be able to visualize GWAS loci across multiple phenotypes and multiple ancestries simultaneously, with layered visualization of molecular QTL results across multiple tissues/cells with ease and speed. The Julia programming language is an optimal platform to achieve this goal (Bezanson *et al.*, 2017) by providing performance of a low-level language while retaining the readability and ease-of-use of a high-level language. Makie.jl (Danisch and Krumbiegel, 2021) is a Julia plotting package that provides powerful plotting utilities and recipes that can be easily extended to visualize most (if not all) genomic data. Using Makie.jl's flexible layout tools, it can be also almost effortless to combine and plot various genetic and genomic data with complex layouts.

Plotting phenome-scale LocusZoom plots

Here, we present the Julia package GeneticsMakie.jl, which builds upon and extends Makie.jl's plotting tools to generate publication-quality figures visualizing multiple genetic and genomic data modalities on different layers, as shown for the *GCKR* locus (**Figure 1**). To further ease this process, we provide functions for munging GWAS (or other association) summary statistics, which can come in various formats (Lyon *et al.*, 2021; Murphy *et al.*, 2021; Bulik-Sullivan *et al.*, 2015) and therefore require harmonization. Linkage disequilibrium (LD) with a SNP of interest is computed on the fly and to compute and visualize LD, any custom reference panel can be loaded using SnpArrays.jl (Zhou *et al.*, 2020). Meanwhile, genes and isoforms with constituent exons and introns are plotted using provided functions with any custom transcriptome annotation file in GTF format. All these functions can be customized or extended easily with minimal loss of performance, which is an inherent strength of the Julia programming language.

To further showcase the power of GeneticsMakie.jl, we share phenome-wide LocusZoom plots for 239 GWAS loci for schizophrenia (**Supplementary Data**; Schizophrenia Working Group of the Psychiatric Genomics Consortium *et al.*, 2020), which are defined as non-overlapping ± 1 Mb windows around the most significantly associated SNPs. We also share such LocusZoom plots for genomic regions known to harbor long-range LD (Anderson *et al.*, 2010), which includes the major histocompatibility complex (MHC) region. Finally, we share LocusZoom plots for high-confidence neuropsychiatric risk genes, which loss-of-function is implicated in increased risk for schizophrenia (Singh *et al.*, 2020), autism spectrum disorder (Satterstrom *et al.*, 2020), and developmental delay disorder (Kaplanis *et al.*, 2020). This permits qualitative assessment of convergence of common variant and rare variant signals for neurodevelopmental and neuropsychiatric disorders.

Table 1. Comparison of functionalities provided by different platforms for creating LocusZoom plots. Shown in parentheses are the names of packages that are suited for a particular function. We note that plotting functions not provided by the current version of GeneticsMakie.jl can be easily added in the future.

	GeneticsMakie.jl	LocusZoom.js	R ecosystem
Munge GWAS summary statistics	0	-	0 (MungeSumstats)
Plot an arbitrary genomic range	0	-	0
Plot an arbitrary number of genes	0	-	-
Color SNPs by LD	0	0	0 (bigsnpr + custom Rcpp code)
Minimal number of dependencies	0	-	-
Plotting various functional genomic data modalities (e.g. Hi-C matrix, enhancer-promoter loops)	-	-	0 (plotgardener)
Can visualize the entire ~10 Mb extended MHC region	0	-	0
Automate drawing many figures	0	-	0
Drawing figures with a very complex layout	0	-	-
Interactivity	-	0	-

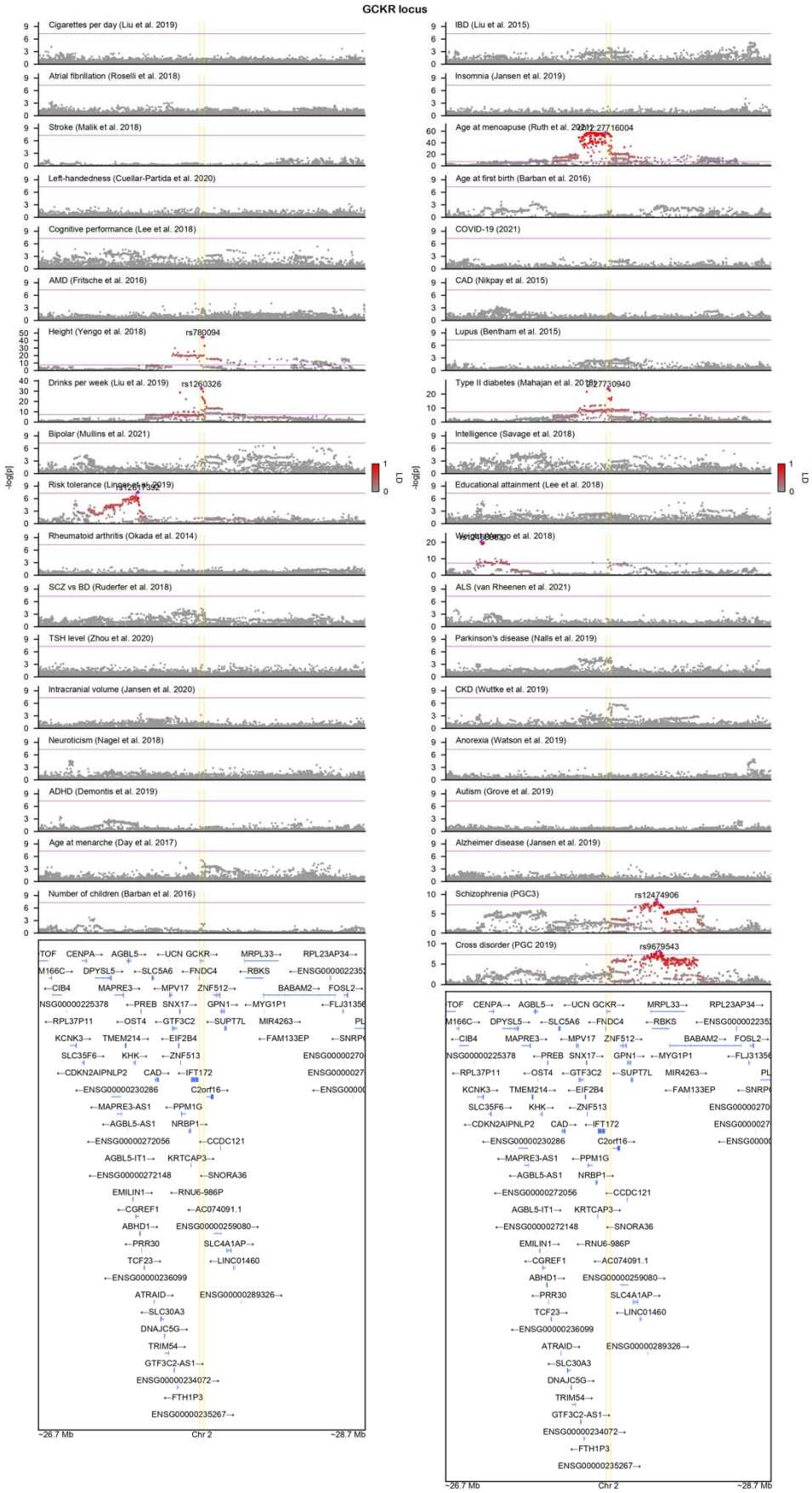

Figure 1. A close look at the known pleiotropic *GCKR* locus using GeneticsMakie.jl. Index SNPs for phenotypes harboring GWAS hits are labeled and corresponding linkage disequilibrium (LD) between other SNPs are displayed with the intensity of red color. Purple line denotes genome-wide significance (*P* = 5e-8), and yellow lines denote gene start and end sites for *GCKR* gene.

Other usage

GeneticsMakie.jl further supports efficient generation of Manhattan plots and corresponding QQ plots for GWAS summary statistics. GeneticsMakie.jl was written with plotting multiple phenotypes in mind such that direct comparison of genetic results across different phenotypes is possible. Visualizing LD blocks as well as plotting correlation results as typically done for reporting genetic correlation (Brainstorm Consortium *et al.*, 2018) is also possible. Other ways of reporting genetic correlation results such as forest plots can be easily drawn using existing plotting functions provided by Makie.jl. Along the same line, colocalization plots (Liu *et al.*, 2019) or circos plots can also be drawn using Makie.jl, which we leave as an exercise to the readers.

Conclusion

In summary, GeneticsMakie.jl allows scalable and flexible visual display of genetic and genomic data within the Julia ecosystem. It produces high-quality, publication-ready figures by default. In the future, we envision other data modalities being plotted on top of what we already have implemented in GeneticsMakie.jl to provide better interpretation of underlying genetic association, hence facilitating the exploratory data analyses (EDA) and generation of novel hypotheses. This includes visualizing PheWAS results (Carroll *et al.*, 2014; Gagliano Taliun *et al.*, 2020) as well as other molecular readouts (Granja *et al.*, 2021; Boix *et al.*, 2021). It would be also worthwhile to implement interactive plots as usually provided by web browser-based tools (Geihs *et al.*, 2015; Kwong *et al.*, 2021).

Acknowledgments

The authors thank members of the Gandal lab and OpenMendel group for testing the package and providing helpful comments.

Code availability

The GeneticsMakie.jl package is open source and distributed under the MIT license via GitHub (https://github.com/mmkim1210/GeneticsMakie.jl). The GitHub repository contains installation instructions as well as examples and documentation for built-in functions.

Funding

This work was supported by the National Institute of Mental Health [R01MH121521 to M.J.G., T32MH073526 and F30MH125523 to M.K.]; and the UCLA Medical Scientist Training Program [T32GM008042 to M.K.].

References

Anderson,C.A. *et al.* (2010) Data quality control in genetic case-control association studies. *Nat. Protoc.*, **5**, 1564–1573.

Bezanson,J. *et al.* (2017) Julia: A Fresh Approach to Numerical Computing. *SIAM Rev.*, **59**, 65–98.

Boix,C.A. *et al.* (2021) Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature*, **590**, 300–307.

Boughton,A.P. *et al.* (2021) LocusZoom.js: Interactive and embeddable visualization of genetic association study results. *Bioinformatics*.

Brainstorm Consortium *et al.* (2018) Analysis of shared heritability in common disorders of the brain. *Science*, **360**.

Bulik-Sullivan,B.K. *et al.* (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, **47**, 291–295.

Carroll,R.J. *et al.* (2014) R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics*, **30**, 2375–2376.

Dadaev,T. *et al.* (2016) LocusExplorer: a user-friendly tool for integrated visualization of human genetic association data and biological annotations. *Bioinformatics*, **32**, 949–951.

Danisch,S. and Krumbiegel,J. (2021) Makie. jl: Flexible high-performance data visualization for Julia. *Journal of Open Source Software*, **6**, 3349.

Gagliano Taliun,S.A. *et al.* (2020) Exploring and visualizing large-scale genetic associations by using PheWeb. *Nat. Genet.*, **52**, 550–552.

Geihs,M. *et al.* (2015) An interactive genome browser of association results from the UK10K cohorts project. *Bioinformatics*, **31**, 4029–4031.

Granja,J.M. *et al.* (2021) ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.*, **53**, 403–411.

Jorgenson,E. *et al.* (2009) VALID: visualization of association study results and linkage disequilibrium. *Genet. Epidemiol.*, **33**, 599–603.

Juliusdottir,T. *et al.* (2018) Toppar: an interactive browser for viewing association study results. *Bioinformatics*, **34**, 1922–1924.

Kaplanis,J. *et al.* (2020) Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature*.

Kramer,N.E. et al. (2021) Plotgardener: Cultivating precise multi-panel figures in R. *bioRxiv*.

Kwong,A. *et al.* (2021) FIVEx: an interactive eQTL browser across public datasets. *Bioinformatics*.

Liu,B. *et al.* (2019) Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.*, **51**, 768–769.

Lyon,M.S. *et al.* (2021) The variant call format provides efficient and robust storage of GWAS summary statistics. *Genome Biol.*, **22**, 32.

Machiela,M.J. and Chanock,S.J. (2018) LDassoc: an online tool for interactively exploring genome-wide association study results and prioritizing variants for functional investigation. *Bioinformatics*, **34**, 887–889.

Machiela,M.J. and Chanock,S.J. (2015) LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*, **31**, 3555–3557.

Murphy,A.E. *et al.* (2021) MungeSumstats: A Bioconductor package for the standardisation and quality control of many GWAS summary statistics. *Bioinformatics*.

Pruim,R.J. *et al.* (2010) LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*, **26**, 2336–2337.

Satterstrom,F.K. *et al.* (2020) Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell*, **180**, 568–584.e23.

Schilder,B.M. *et al.* (2021) echolocatoR: an automated end-to-end statistical and functional genomic fine-mapping pipeline. *Bioinformatics*.

Schizophrenia Working Group of the Psychiatric Genomics Consortium *et al.* (2020) Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. *medRxiv*, 2020.09.12.20192922.

Singh,T. *et al.* (2020) Exome sequencing identifies rare coding variants in 10 genes which confer substantial risk for schizophrenia. *Genetic and Genomic Medicine*.

Zhou,H. *et al.* (2020) OPENMENDEL: a cooperative programming project for statistical genetics. *Hum. Genet.*, **139**, 61–71.