Fauman, Eric B; Keppo, Stella; Cerioli, Nicola; Vyas, Rupesh; Kurki, Mitja; Daly, Mark J.; Reeve, Mary Pat

LAVAA: A lightweight association viewer across ailments

Published in:
Bioinformatics Advances

DOI:
10.1093/bioadv/vbad018

Published: 01/01/2023

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Application Note

LAVAA: A lightweight association viewer across ailments

Eric B. Fauman[1,*], Stella Keppo[2], Nicola Cerioli[2], Rupesh Vyas[2], Mitja Kurki[3-7], Mark Daly[3-7], and Mary Pat Reeve[3]

[1]Internal Medicine Research Unit, Pfizer Worldwide Research, Development and Medical, Cambridge Massachusetts, USA, [2]Visual Communication Design, Aalto University, Finland, [3]Finnish Institute for Molecular Medicine (FIMM), University of Helsinki, Helsinki, Finland,[4]Program for Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA [5]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA [6]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA, [7]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Biobank scale genetic associations results over thousands of traits can be difficult to visualize and navigate.

Results: We have created LAVAA, a visualization web-application to generate genetic volcano plots for simultaneously considering the p-value, effect size, case counts, trait class and fine-mapping posterior probability at a single SNP across a range of traits from a large set of GWAS. We find that user interaction with association results in LAVAA can enrich and enhance the biological interpretation of individual loci.

Availability: LAVAA is available as a stand-alone web service (https://geneviz.aalto.fi/LAVAA/) and will be available in future releases of the finngen.fi website starting with release 10 in late 2023.

Contact: Eric.Fauman@pfizer.com

1 Introduction

The genome-wide association study (GWAS) has been a very successful technique for elucidating the role of inherited variation in human biology and pathophysiology. The importance of a statistical association between a genetic variant and a trait of interest is conveyed by the strength of the association, reflected in the p-value, and the magnitude of the association, reflected in the effect size. The Manhattan plot and the regional association plot are two popular methods to explore the distribution of the p-values across the genome for one trait at a time (Boughton et al., 2021). However, even when interpreting the results of a single GWAS at a single locus it is essential to consider that result in the context of all other genetic studies. The so-called "phenome-wide association scan" or PheWAS, can be presented in a manner analogous to the Manhattan plot, showing the distribution of p-values across a set of traits, but now at a single variant at a time (Denny et al., 2010).

As cohort sizes grow, however, it is becoming increasingly important to consider effect size in addition to the p-value. Indeed, in the limit we might expect that eventually all genes in relevant tissues will show statistical significance with all traits (Boyle et al., 2017). Thus it is likely the case that the most important biological impact of a locus is revealed not in the trait with the strongest p-value but the one with the largest effect.

For continuous traits (e.g., biomarker levels) at a single SNP the trait with the strongest p-value will typically be the trait with the largest effect size (given a fixed sample size, cohort, and standard error across traits) (Fig. 1) (Sham and Purcell, 2014). However this will not be the case for dichotomous traits (e.g., "disease" vs "not-disease"), where the number of "cases" can vary widely across the traits, even if the total number of subjects is constant (Fig. 2).

The genomic context of the association of a single SNP in a GWAS is also important. Because of linkage disequilibrium, the p-value for a trait at a particular SNP can reach genome-wide significance ($p < 5x10\text{-}8$) but may be entirely explained by a much stronger signal 10s or 100s of kilobases away. Statistical fine-mapping techniques (Wang et al., 2020) can pinpoint likely causal variants among correlated variants, but this information is typically absent from visualizations or analyses at a specific SNP.

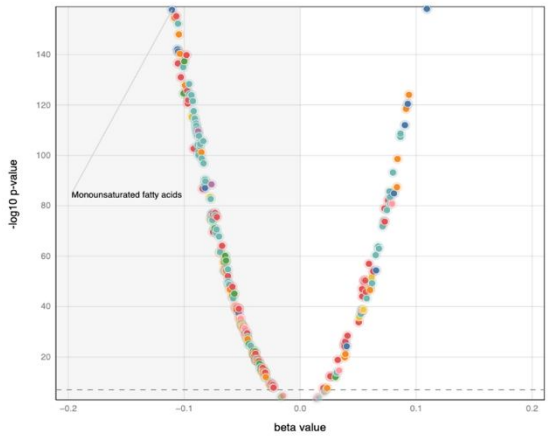


Fig. 1. Sample LAVAA plot for a continuous trait on imported data. This LAVAA plot was created using data imported from OpenGWAS (Hemani et al., 2018) representing metabolomics traits at the GCKR locus (rs1260326) from UK Biobank (Richardson et al., 2022). Each dot represents the -log10(p-value) and beta value for one metabolite trait. The colors represent the classes of measured traits.

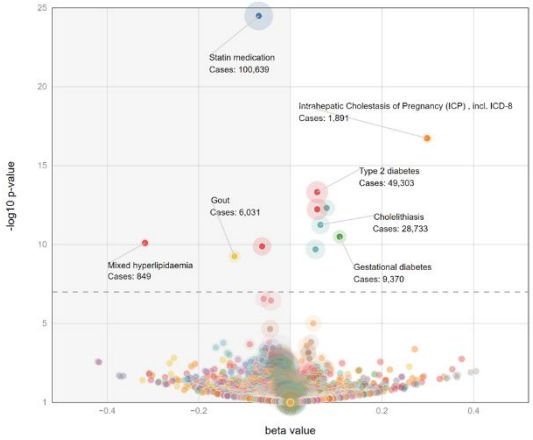
FinnGen is one of the largest nationwide biobank studies, now encompassing 309,154 subjects with genetic results across 3,095 dichotomous traits (as of release 7) (Locke et al., 2019). To assist researchers with extracting valuable genetic, biological, and medical insights from this wealth of data we have designed and implemented LAVAA, a lightweight association viewer across ailments.

2 Methods

LAVAA was developed using the D3 JavaScript library. D3 provides a flexible framework and an active online community to accelerate development of new applications. We also included the following libraries: d3-tip, d3-legend, and d3-format. These libraries further streamline development by adding components created specifically for use by D3-based applications. We also included the simplebar library to support an intuitive user interface design.

The LAVAA user interface is designed to support two aspects of the research process: analysis and display of results. Following "Shneiderman's Mantra" (Schneiderman, 1996) LAVAA first provides an overview of the data with a volcano plot with options to then zoom and retrieve specific details on demand. The initial plot provides the following four dimensions: 1) the magnitude of the association (the beta or the natural log of the odds ratio) on the x-axis; 2) the strength of the association (the -log10 of the p-value from the logistic or linear regression) on the y-axis; 3) the case count displayed as the diameter of the halo around each dot; 4) the category of each trait conveyed by the color of the halo around each dot. In this way every association for a particular variant can be viewed and explored (Fig. 2).

Fig. 2. Sample LAVAA plot for a set of dichotomous traits. Since effective sample size varies drastically from trait to trait the -log10(p-value) and effect size are no longer coupled. This figure represents disease traits from release 7 of FinnGen at the same SNP referenced in Fig. 1. Colors represent groups of related diseases.



LAVAA provides multiple mechanisms for the user to further explore the data. If fine-mapping results are available, traits which have been fine-mapped to the current variant can be highlighted with a dark circle around the central dot. Mousing over any dot displays the association statistics and provides an option to label the dot with those details. A user can zoom in to just the genome-wide significant results ($-\log10(p) > 7.3$) or can sweep and select an arbitrary region of the display. The latter action generates a small table of the association results for all selected variants.

Because the thousands of traits in FinnGen have been mapped to a smaller number of specific categories we implemented a convex hull function which draws the user's eye toward sets of related phenotypes (Fig. 3). A convex hull is defined as the smallest polygon which contains a specific set of points. By changing the focus to categories rather than individual phenotypes the convex hull provides the user with a more global view and reduces the cognitive load (Xu and Chun, 2007).

2.1 Interactive plots of user-provided data

Users can generate their own LAVAA plots by downloading the TSV summary of association results from any variant page in the finngen.fi website (e.g., https://r5.finngen.fi/variant/2-27508073-T-C). This TSV can be directly uploaded to the LAVAA tool here: https://geneviz.aalto.fi/LAVAA/. The LAVAA visualization tool is integrated into the finngen.fi website as of release 10, due to be available to the general public by the end of 2023.

Currently all users can use the standalone LAVAA tool and can also download or fork the LAVAA project on GitHub, here: https://github.com/FINNGEN/volcano_plot

3 Results

We demonstrate the utility of LAVAA plot with an example from the FinnGen. FinnGen employs a rich variety of case-control definitions, refining or expanding the number of subjects included. For example, the broadest definition of "Endocrine, nutritional and metabolic diseases" encompasses over 118,000 cases out of 309,154 subjects. This broad category shows

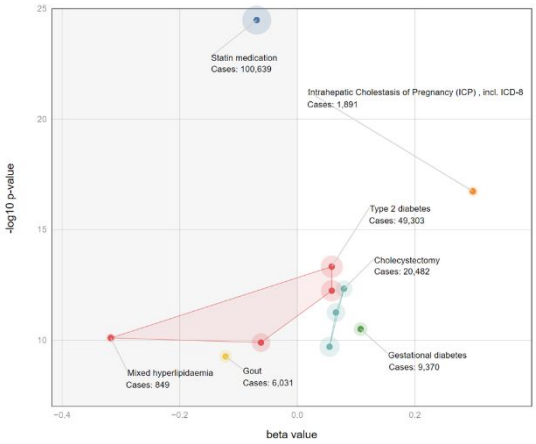


Fig. 3. Example LAVAA plot demonstrating the use of convex hulls to group related traits. This is showing only the significant associations from Fig. 2. The convex hull for "endocrine, nutritional and metabolic diseases" crosses the beta=0 line because the allele at rs1260326 which increases risk for type 2 diabetes also decreases risk for mixed hyperlipidemia.

no significant association at the well-known GCKR non-synonymous variant, rs1260326 (p = 0.074, beta = 0.01 in FinnGen r7). However, as the LAVAA plot reveals, more specific definitions of this disease category can have much stronger associations (Fig. 2). The largest effect size for a genome-wide significant association at rs1260326 is for "mixed hyperlipidemia" with only 849 cases but with an effect size of -0.32 and a p-value of 8x10-11. Similarly, the LAVAA plot illustrates the bifurcation of lipid and diabetic traits with the diabetes risk allele being associated with lower lipid levels and reduced use of statins (Fig. 2 & Fig. 3).

4 Conclusion

LAVAA is a novel web-based visualization tool for assorted traits at a particular locus. By simultaneously representing multiple features per association, LAVAA permits researchers to more deeply explore biobank-scale genetic results such as from FinnGen.

Ethics statement

Patients and control subjects in FinnGen provided informed consent for biobank research, based on the Finnish Biobank Act. Alternatively, separate research cohorts, collected prior the Finnish Biobank Act came into effect (in September 2013) and start of FinnGen (August 2017), were collected based on study-specific consents and later transferred to the Finnish biobanks after approval by Fimea (Finnish Medicines Agency), the National Supervisory Authority for Welfare and Health. Recruitment protocols followed the biobank protocols approved by Fimea. The Coordinating Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS) statement number for the FinnGen study is Nr HUS/990/2017.

The FinnGen study is approved by Finnish Institute for Health and Welfare (permit numbers: THL/2031/6.02.00/2017, THL/1101/5.05.00/2017, THL/341/6.02.00/2018, THL/2222/6.02.00/2018, THL/283/6.02.00/2019, THL/1721/5.05.00/2019, THL/1524/5.05.00/2020, and THL/2364/14.02/2020), Digital and population data service agency (permit numbers: VRK43431/2017-3, VRK/6909/2018-3, VRK/4415/2019-3), the Social Insurance Institution (permit numbers: KELA 58/522/2017, KELA 131/522/2018, KELA 70/522/2019, KELA 98/522/2019, KELA 138/522/2019, KELA 2/522/2020, KELA 16/522/2020, Findata THL/2364/14.02/2020 and Statistics Finland (permit numbers: TK-53-1041-17 and TK/143/07.03.00/2020 (earlier TK-53-90-20).

The Biobank Access Decisions for FinnGen samples and data utilized in FinnGen Data Freeze 7 include: THL Biobank BB2017_55, BB2017_111, BB2018_19, BB_2018_34, BB_2018_67, BB2018_71, BB2019_7, BB2019_8, BB2019_26, BB2020_1, Finnish Red Cross Blood Service Biobank 7.12.2017, Helsinki Biobank HUS/359/2017, Auria Biobank AB17-5154 and amendment #1 (August 17 2020), Biobank Borealis of Northern Finland_2017_1013, Biobank of Eastern Finland 1186/2018 and amendment 22 § /2020, Finnish Clinical Biobank Tampere MH0004 and amendments (21.02.2020 & 06.10.2020), Central Finland Biobank 1-2017, and Terveystalo Biobank STB 2018001..

Funding

The FinnGen project is funded by two grants from Business Finland (HUS 4685/31/2016 and UH 4386/31/2016) and the following industry partners: AbbVie Inc., AstraZeneca UK Ltd, Biogen MA Inc., Bristol Myers Squibb, Genentech Inc., Merck Sharp & Dohme Corp, Pfizer Inc., GlaxoSmithKline Intellectual Property Development Ltd., Sanofi US Services Inc., Maze Therapeutics Inc., Janssen Biotech Inc, and Novartis Pharma AG. Following biobanks are acknowledged for delivering biobank samples to FinnGen: Auria Biobank (www.auria.fi/biopankki), THL Biobank (www.thl.fi/biobank), Helsinki Biobank (www.helsinginbiopankki.fi), Biobank Borealis of Northern Finland (https://www.ppshp.fi/Tutkimus-ja-opetus/Biopankki/Pages/Biobank-Borealis-briefly-in-English.aspx), Finnish Clinical Biobank Tampere (www.tays.fi/en-US/Research_and_development/Finnish_Clinical_Biobank_Tampere), Biobank of Eastern Finland (www.ita-suomenbiopankki.fi/en), Central Finland Biobank (www.ksshp.fi/fi-FI/Potilaalle/Biopankki), Finnish Red Cross Blood Service Biobank (www.veripalvelu.fi/verenluovutus/biopankkitoiminta) and Terveystalo Biobank (www.terveystalo.com/fi/Yritystietoa/Terveystalo-Biopankki/Biopankki/). All Finnish Biobanks are members of BBMRI.fi infrastructure (www.bbmri.fi) and FINBB biobank cooperative (https://finbb.fi/) is the coordinator of the BBMRI-ERIC operations in Finland covering all Finnish biobanks.

Conflict of Interest: EBF is an employee of and shareholder in Pfizer, Inc.

References

Boughton,A.P. et al. (2021) LocusZoom.js: Interactive and embeddable visualization of genetic association study results. Bioinformatics, 37(18):3017-3018.

Boyle,E.A. et al. (2017) An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell, 169, 1177–1186.

Denny,J.C. et al. (2010) PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics, 26, 1205–1210.

FinnGen. https://r7.finngen.fi/ (28 November 2022, date last accessed).

Hemani,G. et al. (2018) The MR-Base platform supports systematic causal inference across the human phenome. eLife, 7:e34408.

Locke,A.E. et al. (2019) Exome sequencing of Finnish isolates enhances rare-variant association power. Nature, 572, 323–328.

Open GWAS. https://gwas.mrcieu.ac.uk/ (28 November 2022, date last accessed).

Richardson,T.G. et al. (2022) Characterising metabolomic signatures of lipid-modifying therapies through drug target mendelian randomisation. PLoS Biol, 20, e3001547.

Schneiderman,B. (1996) The eyes have it: a task by data type taxonomy for information visualizations. Proc. IEEE Symp. Vis. Lang., 336–343.

Sham,P.C. and Purcell,S.M. (2014) Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet. 15(5):335-346.

Sudlow,C. et al. (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med., 12, e1001779.

Wang,G. et al. (2020) A simple new approach to variable selection in regression,

with application to genetic fine mapping. J. R. Stat. Soc. Ser. B Stat. Methodol., 82, 1273–1300.

Xu,Y. and Chun,M.M. (2007) Visual grouping in human parietal cortex. Proc. Natl. Acad. Sci. U. S. A., 104, 18766–18771.