THE PLANT CELL
AMERICAN SOCIETY OF PLANT BIOLOGISTS

Peer Review Report for 10.1105/tpc.16.00551

easyGWAS: A Cloud-based Platform for Comparing the Results of Genome-wide Association Studies

Dominik G. Grimm, Damian Roqueiro, Patrice A. Salomé, Stefan Kleeberger, Bastian Greshake, Wangsheng Zhu, Chang Liu, Christoph Lippert, Oliver Stegle, Bernhard Schölkopf, Detlef Weigel, and Karsten Borgwardt

Plant Cell. Advance Publication December 16, 2016; doi: 10.1105/tpc.16.00551

Corresponding author: Dominik G. Grimm, grimm.dom@gmail.com

Review timeline:

TPC2016-00551-LSB	Submission received:	July 11, 2016
	1st Decision:	Aug. 13, 2016 *revision requested*
TPC2016-00551-LSBR1	1st Revision received:	Oct. 14, 2016
	2nd Decision:	Nov. 18, 2016 *accept with minor revision*
TPC2016-00551-LSBR2	2nd Revision received:	Nov. 20, 2016
	3rd Decision:	Dec. 6, 2016 *acceptance pending, sent to science editor*
	Final acceptance:	Dec. 13, 2016
	Advance publication:	Dec. 16, 2016

REPORT: (The report shows the major requests for revision and author responses. Minor comments for revision and miscellaneous correspondence are not included. The original format may not be reflected in this compilation, but the reviewer comments and author responses are not edited, except to correct minor typographical or spelling errors that could be a source of ambiguity.)

TPC2016-00551-LSB 1st Editorial decision – *revision requested* Aug. 13, 2016

We have received reviews of your manuscript entitled "easyGWAS: A cloud-based platform for comparing the results of genome-wide association studies." Thank you for submitting your best work to *The Plant Cell*. The editorial board agrees that the work you describe is substantive, falls within the scope of the journal (after some discussion), and may become acceptable for publication pending revision, and potential re-review.

We ask you to pay attention to the following points in preparing your revision.

Firstly, please address each point raised by a reviewer separately by phrasing your response followed by listing in detail each of the changes you made in your manuscript accordingly.

In your revision, it is particularly important that you address (i) the effects of differing/only partially overlapping groups of lines used in different GWAS to be compared. (ii) Moreover, a major point to address is the complexity of trait comparisons (caused by different levels of replication, kinship, missing data, etc. between experiments). (iii) Reviewers and editors have observed that increasing the biology by describing/emphasizing novel biological insights revealed by your (meta)analysis would effectively strengthen your manuscript and broaden its interest. Beyond this, the biological context should be clarified and made easier to follow throughout your manuscript. (iv) A number of (but not all of) the editor and reviewers have commented that the manuscript can be interpreted as a narrowly focused tool. Please carefully edit the entire manuscript to enhance the manuscripts accessibility and interest to a broader readership. This publication should enable and encourage the use of GWAS approaches by inexperienced "outsiders" who have not contemplated this in the past. This can be facilitated by including some more background, explaining and defining abbreviations and jargon (for some examples, see below). (v) Finally, maybe you could also implement one additional comment by a reviewer who suggested to mirror all public data between GWAPP and easyGWAS.

In addition to the comments from reviewers, the reviewing editor requests you to address the following issues:

1. Figures: The figures will need extensive revision, taking into account the instructions for authors. In line 283/284, there is a remark that all figures are available online, and maybe the reader should additionally be told in conjunction with the reference to Fig. 1 (and please add information of how precisely to get to the online equivalent of each individual figure where they are mentioned in the text - this is not straightforward at present). In the present versions of the figures, font size ends up too small. Additionally, all panels of a figure must be labeled individually and briefly referred to and explained in legend texts: what is shown, definition of abbreviations used. E.g. Fig. 4 content cannot really be understood in the present way, also because labels are incompletely shown on panel in Fig. 4 (upper left, and also Fig. 7); Figs. 5, 6: explain color code.

2. GWAS tools and methods/approaches should be characterized (linear mixed model against all others) a bit more for the study to receive a broader audience also of non-experts. While this should be far shorter than a review, an overview and characteristics and/or advantages/disadvantages/historical/limits of applicability should be briefly summarized. This should enable a non-expert novel user to gather a first idea which of them can be used in this case and which one(s) may be best. Maybe a Table or Supplemental Table would be suitable to provide this information? And a summary sentence in text, e.g. how you recommend to start.

3. Same for the different alternative methods for multiple hypothesis testing correction, main text lines 177-179.

4. A similar comparison would be useful for meta-analysis methods 1-5 that are partly described in methods.

5. Public are three species. Not clear to me whether users can themselves add additional new species, and - if so - in which steps, within the existing framework, or what the plan is for that in the future. Please make more explicit in main manuscript text.

6. Line 242: please add a little text describing the important results (run-time analyses). Explanations? Consequences? Conclusions? Please add a discussion of this here or later in the manuscript.

7. The possibility to annotate genes and results is mentioned several times, and I feel that a little text on the details and technicalities of this should be included in appropriate positions in the (Results) text.

8. Suppl. Text 3, line 215: An easyGWAS administrator has to first approve the inquiry before a GWAS project is made available. Please list briefly: on which criteria?

9. I feel that it would be best also to briefly outline each of the cases underlying the data shown in Figures 1-5.

10. Concerning the web display: wouldn't it be nice to have, in addition to "phenotypic value", a name for trait, parameter shown/unit, transformation (Fig. 3). And: Why are directions of orfs not shown in Fig. 7, for example?

[*remaining minor comments omitted*]

---------- Reviewer comments:

[Reviewer comments shown below along with author responses]

TPC2016-00551-LSBR1 1st Revision received Oct. 14, 2016

Reviewer comments and author responses:

Thank you for your constructive and insightful reviews. We have thoroughly revised our manuscript based on your comments and recommendations. We have structured the changes in the current submission along the following three axes: i) increased readability of the manuscript and its accompanying supplemental material by adding new guidelines for researchers who may lack extensive experience conducting genome-wide association studies, ii) technical improvements to the easyGWAS framework and added functionality, iii) detailed discussion of the results presented in the *Arabidopsis thaliana* case study. Finally, the technical enhancements to the framework can be summarized as follows:

- Mirroring data with GWAPP: As a first step towards mirroring data with GWAPP (Ü. Seren, et al.) we have started working in close collaboration with the developers of GWAPP. We have created a new database called AraPheno (https://arapheno.1001genomes.org) together with members from the 1001 Genomes Consortium to provide a public database for phenotype data (Note: the AraPheno manuscript is currently under review (Seren, Grimm et al.

2016, under review)). To automatically fetch phenotypes from AraPheno we extended easyGWAS to allow for the integration of public phenotypes from this new platform. This puts us in the path of providing a central platform for all *Arabidopsis thaliana* phenotypes.

- Annotations: We enhanced the annotation functionality of easyGWAS. As an example, the tool now provides more detailed information about variants selected by the user, e.g. if the variant is a missense mutation, frameshift, stop codon or others. This information is automatically fetched from the "Variant Effect Predictor" interface provided by Ensembl.
- Interactive plots: We improved the zooming functionality for Manhattan plots.
- REST-compliant: We added a Representational State Transfer (REST) programming interface. This allows users to obtain information from easyGWAS in various forms, simply by using URLs. An example of a REST query is to get all meta-information (e.g. latitude, longitude, and others) of a particular sample.
- GWAS wizard: We changed the user interface such that Linear Mixed Models are the default algorithm.
- We extended the comparison view to also highlight phenotypes (in red) that are significantly correlated with the genetic kinship matrix to measure whether a phenotype might be highly correlated with population structure

We have received reviews of your manuscript… We ask you to pay attention to the following points in preparing your revision. Firstly, please address each point raised by a reviewer separately by phrasing your response followed by listing in detail each of the changes you made in your manuscript accordingly. In your revision, it is particularly important that you address…

RESPONSE: Thank you for raising the important points (i) and (ii). Before answering them below, it is important to note that we have now clarified and stressed in the manuscript which types of comparative analyses exist in the literature and which of these are offered in easyGWAS: 1) finding joint intersecting associations between different GWAS (which we refer to as 'intersection analysis'), 2) meta-analyses between different studies, 3) multi-trait analyses of several phenotypes on the same individuals. easyGWAS offers 1) intersection analyses and 2) meta-analyses for univariate phenotypes, but not 3) multi-trait analyses, which is currently beyond the scope of easyGWAS, as stated by Reviewer 2 in Comment 2.5 below.

This distinction between 1) and 2) is now stated in the Results section, subsection Comparative GWAS Intersection Analysis View, and that 3) is future work is stated at the end of the Discussion.

(i) The effects of differing/only partially overlapping groups of lines used in different GWAS to be compared.

RESPONSE: Identical, overlapping and differing sets of individuals can lead to different types of biases, spurious associations and/or loss of power, due to:

The violation of the independence assumption: Meta-analyses are commonly based on the assumption that the different datasets are sampled independently. If there is a strong overlap between different datasets, this overlap may lead to a similar association signal in each dataset, thereby artificially confirming this signal and leading to a spurious association found in the meta-analysis.

Difference in sample size: A mere intersection analysis can suffer from differences in sample size between datasets, as weaker–but true–associations may not be discovered in the smaller dataset.

This would lead to false negative findings in the intersection analysis. Meta-analyses tend to correct for this issue by weighting different datasets according to their size.

Interaction effects: An association signal may be present in one dataset, but absent in another, because the individuals in one dataset are exposed to an environmental effect that triggers a gene-environment interaction. Individuals in both datasets may be genetically susceptible to this effect, but it will not be observed in one dataset because the environmental effect is absent there. This would again lead to false negative findings in comparative analyses. A similar phenomenon can occur if a gene-gene interaction affects the phenotype, and the relevant interacting genotype is present in one dataset but not the other. Then the same gene may be significantly associated in only one of the two datasets.

The above text has been incorporated into the manuscript and supplemental material in (a) the Discussion section of the manuscript; (b) Supplemental Material: Suppl. Table 7.

(ii) Moreover, a major point to address is the complexity of trait comparisons (caused by different levels of replication, kinship, missing data, etc. between experiments).

RESPONSE: Thanks for this important comment. We now stress the complexity of trait comparisons in the Discussion section of the manuscript (see Results, subsection Comparative GWAS Intersection Analysis View), and list potential challenges and pitfalls in comparative and meta-analyses in the Supplemental Material (Suppl. Table 7), which—in addition to the points listed in the response to (i)— includes genetic heterogeneity, phenotypic heterogeneity, publication bias, and missing genotypes (see excellent reviews by Rothstein et al., 2005; Bush and Moore, 2012; Evangelou and Ioannidis, 2013).

Genetic heterogeneity: Population structure can cause the finding that certain loci are associated with the phenotype, which are merely correlated to geography and local environmental influences that affect the phenotype. Furthermore, these systematic ancestry differences between different phenotypic classes can lead to spurious associations, just as in GWAS on a single datasets. If two or more phenotypes are significantly correlated with kinship, then they may show shared genetic association signals due to this confounding. easyGWAS offers techniques for association mapping that correct for confounding by population structure in form of Linear Mixed Models. easyGWAS also flags phenotypes in the phenotype correlation matrix that are significantly associated to population structure, to inform the user of this source of potentially spurious joint associations.

Phenotypic heterogeneity: Different protocols to measure phenotypes in different datasets, including different levels of replication of the phenotypes, could lead to artificial differences between associations found in different datasets, despite a common genetic architecture.

Publication bias: It has been observed that studies that find association signals are more likely to be published. This form of publication bias affects the meta-analysis of published studies, because the null hypothesis of no association between genotype and phenotype has already been rejected in each individual dataset (Rothstein et al., 2005). Still, it is not well understood how exactly publication bias affects GWAS. Nevertheless, it is clear that if the studies to be combined in a GWAS meta-analysis focus on results that, a priori, seem more favorable, the bias will then be present (Zeggini and Ioannidis, 2009).

Missing genotypes: The use of different genotyping platforms often results in different sets of genetic features being present in different datasets. As a result, the SNP with the strongest association in one dataset may not even be present in another dataset. This either leads to the need to 1) restrict the analysis to SNPs that are present in all datasets, to 2) analyse hits on the gene level rather than the SNP level, or to 3) impute missing SNPs in each dataset. easyGWAS currently offers option 1) and 2). Special caution needs to be taken with datasets that have been imputed as mentioned in option 3) (Bush and Moore, 2012). If the studies to be combined in a comparative- or meta-analysis have been imputed with different algorithms and/or different haplotype panels drawn from an ethnic population that differs from the target one, this will create additional heterogeneity in the data.

(iii) Reviewers and editors have observed that increasing the biology by describing/emphasizing novel biological insights revealed by your (meta)analysis would effectively strengthen your manuscript and broaden its interest. Beyond this, the biological context should be clarified and made easier to follow throughout your manuscript.

RESPONSE: We have extended the biological interpretation and provided more insights about the findings in the *Arabidopsis thaliana* case study in the Results section. Below we summarize the new additions to the text:

We included the phenotypes that have shared hits of the 87 hits originally reported. These phenotypes are: DTF1: flowering time as days until emergence of visible flowering buds in the center of the rosette from time of sowing. DTF2: flowering time as days until the inflorescence stem elongated to 1 cm. DTF3: flowering time as days until first open flower. RL: rosette leaf number. CL: cauline leaf number.

A thorough biological interpretation of the importance of the three hits associated to RL and in close proximity to *FLOWERING LOCUS T (FT)* was added. In summary, studies show that a polymorphism mapped to *FT* promoter results in delayed flowering time (Schwartz et al., 2009) and that the *FT* promoter length correlates with flowering time (Liu et al., 2014). Most importantly, these results were observed when the coding sequence of *FT* remains unchanged. Similarly, we included a discussion of the hits reported in *FLOWERING LOCUS C* (*FLC*). The same *cis*-regulatory mechanisms mentioned above come into play in this case (Irwin et al., 2016) and of particular importance is the SNP Chr5_3173596 reported by our analysis, which is jointly associated to two different flowering types phenotypes DTF2 and DTF3.

An additional discussion was included in regards to the SNPs associated to the rosette leaf number (RL) phenotype. These SNPs—located in Chr1 at positions 24337820, 24338990 and 24339560—overlap with 3 non-coding RNAs reported by Liu et al., 2012. We discuss the potential role these non-coding RNAs play as enhancer elements modulating the expression of *FT*. We theorize that this behavior may resemble a mechanism described in Ariel et al., 2014 by which the lncRNA *APOLO* regulates the expression of *PINOID* via chromatin looping. Supplemental Figure 14 was added to show the spatial co-location of *FT* with the non-coding RNAs that overlap with the abovementioned SNPs. We have ordered T-DNA insertion lines that may disrupt these lncRNAs to test their role in flowering time.

Finally, all plots and analysis results related to this case study have been made public through the following link in easyGWAS: https://easygwas.ethz.ch/gwas/myhistory/public/14/

In conclusion, the results presented in our case study provide insights into potential pleiotropic effects of certain SNPs. This was possible due to the fact that easyGWAS allows for the systematic comparison of results from GWAS that were conducted under the same controlled and rigorous conditions (i.e. the individual analyses of the nine phenotypes were performed with FaST-LMM which corrects for cryptic relatedness and population structure). The comparison analysis itself provides additional details of the correlation structure of the different phenotypes, as well as reporting the SNPs marked as hits. All of this is done while also correcting for multiple hypothesis testing to limit the number of false positives and to provide the researcher with promising hypotheses that seem, a priori, statistically significant.

All these changes were incorporated in (a) sections/subsections of the manuscript: [Results] Case Study in *Arabidopsis thaliana* (b) Supplemental Material: Suppl. Figure 14.

The last comment raised by the editors is in regards to adding the appropriate biological context to all sections throughout the manuscript. This was achieved by an extensive re-writing of subsections of the manuscript and supplemental material. This topic is addressed in detail in the next comment (iv).

(iv) A number of (but not all of) the editor and reviewers have commented that the manuscript can be interpreted as a narrowly focused tool. Please carefully edit the entire manuscript to enhance the manuscripts accessibility and interest to a broader readership. This publication should enable and encourage the use of GWAS approaches by inexperienced "outsiders" who have not contemplated this in the past. This can be facilitated by including some more background, explaining and defining abbreviations and jargon (for some examples, see below).

RESPONSE: We have followed the recommendations from the editors and reviewers in regards to enhancing the manuscript's accessibility. Both, the main text and supplementary material have been heavily edited to add more background information about the methods implemented by easyGWAS and their goals. The underlying theme in the manuscript is now to: a) highlight the analysis functionality provided by easyGWAS, but also to b) provide guidelines on what association methods are more suitable for each study and c) recommend best practices in order to avoid pitfalls in comparative- and meta-analyses. The main changes that were introduced can be characterized as:

Motivation behind conducting genome-wide association studies (GWAS). The goal is to provide the reader with an overview of what GWAS are and how they differ from other mapping strategies (e.g. linkage mapping).

New section with details on how to conduct comparative analyses of GWAS results. This is one of the main features of easyGWAS that sets it apart from other online tools and the goal of the new text is to guide the reader on how to conduct a comparative analysis. The mathematical foundations behind this type of analyses our now outlined in the supplemental material (Suppl. Text 7). Finally, because these type of analysis has implicit biological assumptions, we also included a list of pitfalls to raise the awareness of the reader, should these assumptions be violated (Suppl. Table 7).

Additional biological insights and extensive discussion of the results presented in the case study of *Arabidopsis thaliana*. A detailed description of the newly added text can be found in the response to the previous comment (iii).

Additional descriptions of the different statistical models that can be used to conduct a genome-wide association study. We also included a series of guidelines to help the reader determine which statistical model may suit their needs best. Because some of these models make assumptions about the distribution of phenotype values (i.e. they assume Gaussianity of the data), a transformation step is normally conducted as pre-processing. We have

included a new section explaining the different transformation methods available in easyGWAS in hopes that this will guide the reader to choose the best transformation for their data (Suppl. Text 8).

Additional recommendations and best practices on how to perform meta-analyses of GWAS and comparison of GWAS. In the case of a meta-analysis, there are also assumptions that should be made about the underlying genotype data (e.g.: do genetic variants have the same effect across all studies?). To assist the reader in deciding what assumptions make more sense in their dataset, we discuss the different types of effects that can be assumed (i.e. fixed- vs. random-effects) and detail the caveats and potential consequences of making the wrong the assumptions (Suppl. Text 9 and Suppl. Table 7).

Detailed step-by-step procedures to help the reader recreate the figures and replicate the analyses presented in the manuscript (Suppl. Text 4).

Technical details of extra functionality added to easyGWAS based on the reviewers' recommendations. Most of the text related to technical aspects of easyGWAS was moved to the Supplemental Material.

Additional descriptions of the methods used to correct for multiple hypothesis testing. The goal is to provide the reader with an intuition of when to apply each method and what the effects of different corrections are when reporting hits (Suppl. Text 10).

All these changes were incorporated in (a) sections/subsections of the manuscript: Introduction, [Results] Comparative GWAS Intersection Analysis View, [Results] Case Study in *Arabidopsis thaliana*, [Methods and Materials] Genome-Wide Association Mapping Methods, [Methods and Materials] Transformation Methods, [Methods and Materials] Meta-Analysis Methods; (b) Supplemental Material: Suppl. Text 7-11, Suppl. Table 4 (list of abbreviations), Suppl. Table 5-7.

In conclusion, we have rephrased existing portions of the manuscript and added considerable amounts of content to enhance the overall accessibility of the manuscript and supplemental material. Our main goal is to provide a sufficient amount of details about how different methods are implemented without getting bogged down in jargon or mathematical details. Moreover, we include guidelines on what assumptions to make for different models and warn the reader of potential pitfalls that can arise when the wrong assumptions are made. We hope now the material is accessible enough to allow the novice reader to start performing GWAS, and comprehensive enough to empower the experienced researcher to conduct sophisticated analyses under well-defined assumptions.

(v) Finally, maybe you could also implement one additional comment by a reviewer who suggested to mirror all public data between GWAPP and easyGWAS.

RESPONSE: We acknowledge that mirroring all public data in GWAPP will be extremely beneficial for easyGWAS and its users. To that effect, we would like to comment that the proposed mirroring is currently an ongoing effort of our group, the Nordborg group and members of the 1001 Genomes Consortium. We are simultaneously developing a central public phenotype-database, called AraPheno (Seren, Grimm et al. 2016, under review), which shall serve as a central repository for public phenotypes for the model organism *Arabidopsis thaliana*. This web-server provides a REST interface to exchange public phenotype data between easyGWAS, GWAPP and the AraPheno portal. AraPheno is currently under review at the NAR database issue. See Supplemental Text 8 for details about the easyGWAS REST interface.

We enhanced easyGWAS such that users can automatically fetch public phenotypes from AraPheno and integrate them into easyGWAS. For this purpose, we added a new function to the phenotype upload manager. This function helps users to fetch data from AraPheno. We updated the text in the main manuscript (see section "Data Repository") and Supplemental Text 1 accordingly.

This is a first step towards mirroring data across different platforms and will be an ongoing effort.

In addition to the comments from reviewers, the reviewing editor requests you to address the following issues:

1. Figures: The figures will need extensive revision, taking into account the instructions for authors. In line 283/284, there is a remark that all figures are available online, and maybe the reader should additionally be told in conjunction with the reference to Fig. 1 (and please add information of how precisely to get to the online equivalent of each individual figure where they are mentioned in the text - this is not straightforward at present). In the present versions of the figures, font size ends up too small. Additionally, all panels of a figure must be labelled individually and briefly referred to and explained in legend texts: what is shown, definition of abbreviations used. E.g. Fig. 4 content cannot really be

understood in the present way, also because labels are incompletely shown on panel in Fig. 4 (upper left, and also Fig. 7); Figs. 5, 6: explain colour code.

RESPONSE: Regarding missing online figures: Thank you for pointing this out. We first apologize for the confusion that all figures are available online. They are not available as graphics (png or pdf) files, our links rather lead to the corresponding webpage in easyGWAS which shows the corresponding results in its original resolution, and the figures in the paper are snapshots of these webpages.

Regarding the description of the panels: To improve presentation, we have followed your advice and colored and named each panel and describe their meaning in the caption of the figures.

Regarding figure resolutions and font sizes: The purpose of Figures 2-4 is to show the overall layout and structure of the easyGWAs website (not a specific result on a specific dataset). We tried to improve their resolution and font sizes, but ran into the following problems:

We experimented a lot with different font-sizes in the web-application. However, larger font-sizes lead to overloaded and sometimes broken websites and thus to less structure.

We also considered providing both zoomed-out overviews of the different views of easyGWAS and a zoomed-in version of every single panel, but this breaks any reasonable page limit.

Hence our current solution to this problem is to continue to show the overview figures in the paper, and to provide the possibility to study their content in detail and in high resolution using the direct weblinks to easyGWAS (as a live website, not as a graphics file).

Regarding reproducibility: Furthermore, to demonstrate the reproducibility of what we show, we provide step-by-step instructions in the Supplement on how to produce the results shown in Figure 2-6. See *Supplemental Text 4: Step-by-Step Procedures to Reproduce the Content of Figures*.

2. GWAS tools and methods/approaches should be characterized (linear mixed model against all others) a bit more for the study to receive a broader audience also of non-experts. While this should be far shorter than a review, an overview and characteristics and/or advantages/disadvantages/historical/limits of applicability should be briefly summarized. This should enable a non-expert novel user to gather a first idea which of them can be used in this case and which one(s) may be best. Maybe a Table or STable would be suitable to provide this information? And a summary sentence in text, e.g. how you recommend to start.

RESPONSE: We thank the editors and reviewers for this suggestion as it greatly increases the readability of the manuscript and the accessibility of the tool. To that effect, we have included detailed descriptions of the statistical models that are used in easyGWAS to determine association. As mentioned in the response to comment (iv), the reader can now assess when each model is applicable to their study. We distinguish the models upon their robustness to deal with confounders and their ability to correct for cryptic relatedness, in order to avoid inflated p-values of association. We stress the importance of the linear mixed models to deal with the previously mentioned relatedness and indicate the assumptions each model makes to better prepare the reader in understanding the hits reported by the model. All these changes were incorporated in the Section *Methods and Materials,* Subsection *Genome-Wide Association Mapping Methods*.

3. Same for the different alternative methods for multiple hypothesis testing correction, main text lines 177-179.

RESPONSE: In a similar fashion to what we mention in the previous comment, we have followed the advice of the editors and reviewers and written new content to better explain the methods for multiple hypothesis testing that are available in easyGWAS. Each method is now described in detail and their assumptions are clearly stated. Moreover, we also mention the benefits—in statistical terms—the reader can reap if the assumptions prove to be true.

The new text is meant to be interpreted by the reader as a journey, in which a novice to GWAS will be satisfied to adopt a simple method like Bonferroni to prevent the occurrence of not even one false positive (with family-wise error rate), whereas a seasoned researcher may be willing to draw more hypotheses by controlling for the false discovery rate using Storey and Tibshirani's q-values.

All these changes were incorporated in the Supplemental Material, Suppl. Text 10.

4. A similar comparison would be useful for meta-analysis methods 1-5 that are partly described in methods.

RESPONSE: Again, we are thankful for this suggestion and we have implemented it by adding a new section in the Supplemental Material (Suppl. Text 9). There we provide additional recommendations and discuss best practices on how to perform meta-analyses of GWAS. We also discuss the assumptions that should be made about the underlying genotype data, such as "Can we safely assume that all genetic variants have the same direction of effects across studies?". To assist the reader in deciding what assumptions are best suited to their dataset, we elaborate on the the different types of effects that can be assumed (i.e. fixed- vs. random-effects) and detail the caveats and potential consequences of making the wrong the assumptions (Suppl. Table 7).

All these changes were incorporated in the Supplemental Material, Suppl. Text 9 and Suppl. Table 7.

5. Public are three species. Not clear to me whether users can themselves add additional new species, and - if so - in which steps, within the existing framework, or what the plan is for that in the future. Please make more explicit in main manuscript text.

RESPONSE: All registered users are able to add additional new species and upload their private data for this species. This can easily be done with the Upload Manager. The user has to select upload "Genotype" button. Here, the user can either select an existing species, or he can create a new species by clicking the "Add new species" button. In addition, the user has to provide a ZIP file with the genotype data in PLINK format (and optionally phenotype, covariate and gene-annotation). After successfully uploading the new genotype data the new species as well as the data will be visible in the "Private Data" repository.

We mention the upload functionality in the main text at "Data Download, Upload and Sharing" and more detailed in Supplemental Text 1.

6. Line 242: please add short text describing the important results (run-time analyses). Explanations? Consequences? Conclusions? Please add discussion of this here or later in the manuscript.

RESPONSE: Thanks, we extended run-time analyses paragraph and added a brief conclusion about its consequences (see section "Runtime Analysis"): "We observe that all algorithms, except for logistic regression, are at least as efficient as the tools compared to. The results show that easyGWAS can compute GWAS with standard models such as linear regression within a few minutes for up to five million SNPs and up to 500 samples. More complex models, such as FaST-LMM or EMMAX, only take minutes for approximately 100 samples and a few hours for up to 500 samples."

7. The possibility to annotate genes and results is mentioned several times, and I feel that short text on the details and technicalities of this should be included in appropriate positions in the (results) text.

RESPONSE: Thanks for mentioning the annotation pipeline. Taking into consideration comments from Reviewer #1 we enhanced the annotation functionality of easyGWAS. As an example, the tool now provides the potential impact of a variant selected by the user, e.g. if the variant is a missense mutation, frameshift, stop codon, amino-acid change or others. This information is automatically fetched from the "Variant Effect Predictor" interface provided by Ensembl. A live demonstration can be accessed via the following link:

https://easygwas.ethz.ch/gwas/results/snp/detailed/cecaaa0d-582a-4358-8251-09cf73a439fa/Chr3/19507447/

8. Suppl. Text 3, line 215: An easyGWAS administrator has to first approve the inquiry before a GWAS project is made available. Please list briefly: on which criteria?

RESPONSE: Good point, thanks. We added a short summary of what will be important to the Suppl. Text 3, e.g. phenotypes, covariates should have meaningful names and descriptions. The GWAS project name and experiment name should be meaningful. The user should provide a short description about the project and the experiments he or she has done and why. We added the following text to Supplemental Text 3:

"Here, administrators check if the user provides meaningful names for the GWAS project, experiments, dataset, phenotypes and samples, but also if a brief description is given about what has been done. This inquiry and approval step should serve as a basic quality check before data and results are made public."

10. Concerning the web display: wouldn't it be nice to have, in addition to "phenotypic value", a name for trait, parameter shown/unit, transformation (Fig. 3). And: Why are directions of orfs not shown in Fig. 7, for example?

RESPONSE: Thank you for this good point. We updated the "Detailed SNP" view to also show the name of the trait and the transformation. We do not show units in this view since it is not necessary to provide this information when phenotypes are integrated into easyGWAS. If units are given, they can be found in the detailed phenotype view. We decided to not show the direction of orfs in Figure 7 to not overload the plots. We provide this information in the zoomed-in Manhattan plots. We have to note that all figures are generated dynamically in the browser using modern HTML5 and JavaScript techniques. We have to make sure that plots can be rendered for as many screens and resolutions as possible. This makes it difficult to provide too much information in plots without running the risk of breaking the visualisation pipeline.

Reviewer #1:

In this work, the authors constructed a web-based platform easyGWAS to facilitate the data analysis (especially for the users without too much bioinformatics background) and enable comparison of GWAS results. The new cloud-based tool should be quite helpful for the users to analyze and display the GWAS results (e.g., Manhattan plots and QQ plots) and carry out some meta-analysis.

I have two major concerns for the work:

1.1 For Arabidopsis, the web-based GWAS platform "GWAPP" has been released and published in *Plant Cell*. There were also several open, web-based platforms (e.g., "Galaxy") for many biological data analyses. Hence, the new features and potential utility of easyGWAS need to be emphasized.

RESPONSE: We agree that there are other web-based tools used to conduct genetic analyses. In addition to the ones mentioned by the reviewer—GWAPP and Galaxy—we also mention in the manuscript EMMA (Kang et al., 2008), DGRP2 (Mackay et al., 2012) and Matapax (Childs et al., 2012). Below we list the novel features provided by easyGWAS and outline the limitations in some of the tools mentioned above:

Interactive visualization of results (for example, not easy to attain with Galaxy)

Support for multiple species (not provided by any other tool)

Upload of custom genotype data (for example, not provided by GWAPP or DGRP2)

Capabilities to conduct GWAS with a wide range of association methods (for example, not easy to attain with Galaxy; and limited in other tools)

Integration and comparison of results obtained from different GWAS (not provided by any other tool)

Publishing and sharing of results between collaborators and registered users (not provided by any other tool)

To address the reviewer's comment, we have rephrased the text in the manuscript highlighting the features that make easyGWAS unique. To that effect, we rephrased the fourth and fifth paragraph in the Introduction making clear what the novel contributions of easyGWAS are. Additionally, we have rephrased the second and third paragraphs of the discussion to emphasize the novel features provided by easyGWAS by contrasting them to what is currently offered by other web-based platforms.

1.2 Currently, the association analysis is not difficult in the studies of plant genetics. The platform needs to pay more attention to the follow-up analysis of GWAS, including the integration of information from expression profiles, coding variants and other literature reports for the candidates in the associated loci. So, the new platform should be improved with multiple functional analysis.

RESPONSE: This is a very interesting point and we thank the reviewer for raising it. As a major revision step, we extended our annotation pipeline to enrich SNPs with additional information. We now provide in our "Detailed SNP View" not only LD plots and gene regions, but also predictions of potential SNP effects from the Ensembl Variant Effect Predictor (McLaren et al., 2016). Users can now see if a SNP is, for example, a missense mutation, a synonymous mutation or stop codon. We also provide information about the transcript, codon and amino-acid change if available. A demo that illustrates these new features is available at: https://easygwas.ethz.ch/gwas/results/snp/detailed/cecaaa0d-582a-4358-8251-09cf73a439fa/Chr3/1950 7447/

Additionally, we extended the SNP Annotation View and now provide a link to the "Detailed SNP View". See example at: https://easygwas.ethz.ch/gwas/results/snpannotations/cecaaa0d-582a-4358-8251-09cf73a439fa/

Reviewer #2:

2.1 This paper outlines an excellent workspace for many plant biologists now screening wild collections (often with published genotype data). It allows them to upload, archive, visualize, transform and share multidimensional phenotype data. Critically it provides easy to run Genome Wide Association Studies, to identify QTL as SNPs. Importantly, SNP effects are estimated individually and the variance explained by the full model together (kinship) is provided.

RESPONSE: Thank you.

2.2 It would be rather trivial to include genomic predictions into easy GWAS. Once the phenotypes for a subset of lines are uploaded and GWAS performed, best linear unbiased estimates of SNP and kinship effects, can be used to estimate a phenotype for other lines that have genotype information in the database but that were not included (or held back) from the initial phenotype screening. This would allow users to select new sets of lines for validation and/or to explore in more detail accessions where there is a large 'residual' between observed and predicted phenotype. It also allows cross validation when different subsets are used for GWAS.

RESPONSE: We agree that this is not mathematically challenging, and that it would be an interesting extension to easyGWAS. We do not think, however, that easyGWAS' core contribution of running and comparing GWAS is incomplete without genomic predictions. The effort of implementing these genomic predictions, which will require a new wizard within easyGWAS with its corresponding visualizations— analogous to the GWAS wizard—would result in several months of full time additional implementation effort. We will therefore pursue this extension in future work.

2.3 The interactive GWAS output is awesome and very useful for exploring candidate genes, which is the rate limiting step. I struggled with the zooming in and out some, perhaps buttons for scroll and zoom can be added or a box to explain the clicks necessary.

RESPONSE: Thank you. There was a small issue with one of the JavaScript libraries. The issue has been resolved and an additional brief description on how to use the zooming features has been included above the Manhattan plots. In a nutshell, to zoom into a region of interest, the user must first select a region in the plot. To do so, the start position of the region is marked with a left click of the mouse. Then, while holding the left mouse button, move the pointer to the right and release the click when the region of interest has been selected. This marks automatically the area to zoom in and the plot is refreshed. An interactive demo of this feature is available at: https://easygwas.ethz.ch/gwas/results/manhattan/view/8557bdde-aa8a-4615-a643-ccce51a4edc0/

2.4 The options for including covariate data is an important advance, this allows sex, other major QTL, *a priori* loci, treatments etc to be included in the GWAS which is sure to alter and improve results of single gene scans. Linear Mixed Models are now the standard for GWAS and should be the default option. The kinship matrix could be re-estimated for each sample set

RESPONSE: Thank you for this suggestion. We updated the wizard such that a linear mixed model is selected by default.

2.5 The phenotype-phenotype correlation plot is a key feature. These days, many correlated phenotypes are measured and it is of interest to know for which loci QTL effects are similar across traits and which are not. A formal multi trait investigation is not easy and would be beyond the scope of this paper. However part of the phenotype-phenotype correlations among traits is due to kinship, rather than specific QTL, e.g. traits correlated with flowering time. It could be possible to show the 'polygenic correlations' and 'nonpolygenic correlations' among traits pairs, by accounting for the trait correlation with kinship. This will become important as we look at pleiotropic QTL. If this cannot be added, it would be worth flagging a high trait ~ kinship correlation that can underly background trait-trait correlations rather than correlations due to pleiotropic QTL.

RESPONSE: We followed your last suggestion and extended the comparison of GWAS to highlight phenotypes that are significantly correlated with the genetic kinship matrix. For this purpose, we use the Hilbert Schmidt Independence Criterion, a kernel-based multivariate measure of statistical dependence (Gretton et al. 2005), which quantifies the dependence between phenotype and kinship. Details on how this measure of statistical dependence can be computed can be found in Supplemental Text 7. Phenotypes that show such a significant dependence to kinship are shown in red in the phenotype-phenotype correlation plot.

2.6 'LD and gene information in close proximity to a (line275) focal SNP' is a fantastic option that is very useful!

RESPONSE: We appreciate your feedback.

Reviewer #3:

3.1 The manuscript 'easyGWAS: A cloud-based platform for comparing the results of genome-wide association studies' by Grimm et al. describes the on-line software tool for GWAS analyses and comparisons between studies. Since the number of GWAS studies is accumulating rapidly this is a timely and very welcome platform. The authors went to great length to make the platform user-friendly and provide a rich set of attractive analysis and visualisation tools. Because there are numerous good packages to perform association analysis the strength of this platform lies in the possibility of comparing and sharing different studies. However, there are a number of issues that require attention. The authors report on a case study of related traits in Arabidopsis, with the convenience that all traits were measured on the same set of lines and mapped using the same genotype data. However, in practice, studies are often performed on different subsets of lines and associated to dissimilar genotype data.

3.2 Likewise, stochasticity in trait values and SNP allele frequency might lead to fluctuations in the tag SNP assigned to an associated locus. As a result, a trait measured in different studies, or two pleiotropically regulated traits, might display a strong association with a single specific locus while this will be represented by different SNPs. Since the comparison tools only seem to detect co-occurrence of significant SNPs this raises the question how many relationships will be missed. It would make much more sense to compare significantly associated loci, e.g. by taking LD around significant SNPs into account or by applying sliding window analyses.

RESPONSE: easyGWAS addresses this issue at least to the degree that it not only finds overlapping SNP hits across several GWAS, but also "shared genes", that is genes in whose neighborhood significant SNP hits were found across several GWAS. https://easygwas.ethz.ch/comparison/results/gene/view/dfaa2551-7b2d-4e3d-9170-6522966b7d2a/

In future work, we will extend this to intergenic regions as well.

3.3 Furthermore, the authors assume that significantly associated SNPs are representative of a single gene. This is often not the case. Especially when a locus is under strong selection in which selective sweeps extent LD over numerous genes. Actually, *FLC* is a good example of this; the most strongly associated SNPs with flowering time are often not located within the gene itself but can be located at considerable distance from *FLC* and often in closer proximity to a neighbouring gene. The authors acknowledge this with the case of *FT* but they do not provide a systematic approach how to deal with this.

RESPONSE: You can change the sequence around a SNP in which easyGWAS looks for neighboring genes through the panel "SNP Annotations > GWAS Annotation Options > Search Window around SNP". For an example, please see: https://easygwas.ethz.ch/gwas/results/snpannotations/4d00706f-ad0f-4f57-9f4e-ac3099b15b94/

3.4 Despite my remarks above I believe this is a very valuable tool that deserves a proper introduction to the research community. I doubt, however, whether *The Plant Cell* is the appropriate journal for dissemination. The manuscript contains very little biological information and the case study does not provide novel insights. I would argue that this manuscript would fit better in a journal with a more methodological scope.

RESPONSE: We thank for your positive comments. We think that this tool is best placed in a journal in which many potential scientific users read and learn about it.

TPC2016-00551-LSBR1 2nd Editorial decision – *accept with minor revision*	Nov. 18, 2016

[Editor and reviewer comments shown below along with author responses]

TPC2016-00551-LSBR2 2nd Revision received	Nov. 20, 2016

Reviewer comments and author responses:

Thank you again for your constructive and insightful reviews. We have thoroughly revised our manuscript based on your comments and recommendations. In addition to your and the reviewer comments we also included a new dataset

into easyGWAS from a recent study about the "Genetic architecture of nonadditive inheritance in *Arabidopsis thaliana* hybrids" (Seymour et. al, 2016, *PNAS*).

Reviewing editor:

[*minor comments omitted*]

Although I do not require this to be implemented during this revision, I would think that the following changes/additions in easyGWAS would be very useful in the future (and maybe a can be implemented - I see it as most important at present):

a) For the user to be able to browse public phenotypes without having to enter (i.e. know) the initial characters

> RESPONSE: This is a very valid point and we are aware that the selection of phenotypes can be improved. In fact, it is not an easy fix to be included in the current revision but we are currently looking into options to make this interface more user-friendly. As mentioned in the first page of the cover letter, we are continuously adding new datasets to easyGWAS and we have realized that simply listing all available options to the user (say, phenotypes or samples) may, at some point, overwhelm the user. Our current implementation – the autocomplete feature – allows the user to choose a phenotype without having to scroll through long lists of values. We currently provide the option of browsing all publicly available phenotypes for a certain species and dataset. The user can navigate to the public data repository: https://easygwas.ethz.ch/data/public/phenotypes/ to get a list of all available phenotypes. Details about the phenotype can be retrieved by clicking on the corresponding phenotype name. Of course, this does not solve the issue raised by the editor but we feel that at least mitigates the problem.

b) To be able to access more information on phenotypes (these are highly abbreviated, and sometimes unclear), i.e. long version of what the phenotype is (description line), additionally an expanded description for example of what was measured in which tissues and how or other important information pertaining to the phenotype data, and finally the ability to access metadata including, for example, growth conditions, age of harvest, tissue harvested, time of day of harvest,...

> RESPONSE: You are right. We updated the phenotype tables to also show more detailed information about the phenotype, e.g. the scoring of the phenotype (please, refer to https://easygwas.ethz.ch/data/public/phenotypes/). More detailed information can then be found in the detailed phenotype view by clicking on the phenotype name.

c) Enable the GWAS user to name/change name of an Experiment/Temporary Experiment.

> RESPONSE: This is already possible for stored experiments. Just click one of the check-boxes next to an experiment and click on "Edit Experiments" in the upper right corner of the experiment history ("My experiments"). In this view users can rename experiments or group them into different projects. Temporary experiments can only be renamed if the users store them in one of their projects in easyGWAS. Otherwise temporary experiments are deleted after 48h.

Reviewer #1:

In this manuscript, Grimm and colleagues introduce 'easyGWAS' an online tool for conducting GWAS. This is an excellent online resource that should greatly facilitate the use of GWAS. Similar online tools have been developed - but as the authors note - these are largely focused on a few model species. Notably, the interface of easyGWAS is intuitive, which should allow it to be quickly adopted by a wide variety of researchers in a wide range of disciplines. The ability to conduct simple (vanilla) GWAS and meta-analyses adds to the value of easyGWAS and will likely help increase the number of researchers comfortable with GWAS. Thank you.

Be that as it may, I have a few (mostly minor) comments for improving the manuscript and resource.

Comments

Page 1, lines 57-9: this sentence is unclear and seems to have suffered from too many edits. It is of course true that GWAS offer higher resolution than traditional linkage mapping approaches, but this is not because "recombination affects the phenotype of interest". It is simply due to the larger number of recombination events that will have occurred in a natural mapping panel.

RESPONSE: Thank you for finding this flawed formulation. We changed the phrase to “Moreover, GWAS offer a higher resolution than linkage mapping because of the larger number of recombination events that will have occurred in natural panels used for association mapping”.

Page 14, lines: 499-503: do these permutation tests take into account population structure? The reader would also benefit from some indication (either in the manuscript or on the website) of the amount of time required for these permutation tests. Even highly 'performant' permutation tests can take several hours (with a reasonable sized dataset), and I expect researchers to repeatedly contact the easyGWAS-staff with questions about their analyses if they wait longer than a few hours. The most impatient researchers will undoubtedly assume that something is awry and resubmit their analyses.

RESPONSE: Very good point. Depending on the algorithm, the permutation tests will also take population structure into account, e.g. if the user selects EMMAX. Here, we compute the p-value using permutations instead of assuming a $\chi 2$ distribution. This leads to the fact that we do not have to assume that the phenotypes have to be normally distributed. We approximate the true null distribution using permutations. You are right regarding the runtime: These tests can take even several days. This is why we print a warning in the web application that permutation based GWAS with more than 100K SNPs can already take several days.

The ability to conduct GWAS online is not novel, but easyGWAS has the potential to become widely used among researchers. The MTMM and MLMM (available through limix, written by two of the co-authors) would facilitate this, and I hope that the authors are able to include this functionality sooner rather than later.

RESPONSE: Thank you for your suggestion and we agree that implementing multi-trait mixed models will add an extra level of functionality to easyGWAS. We are planning to include them in a future release of the web application as they require new visualizations and database models.

[*remaining minor comments omitted*]

TPC2016-00551-LSBR2 3rd Editorial decision – *acceptance pending*	Dec. 6, 2016

Thank you very much for a new round of careful revisions. I apologize for the delay, which was a result of traveling. We are pleased to inform you that your paper entitled "easyGWAS: A cloud-based platform for comparing the results of genome-wide association studies" has been accepted for publication in *The Plant Cell*, pending a final minor editorial review by journal staff.

Final acceptance from Science Editor	Dec. 13, 2016