

Supplemental Material

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

Dominik G. Grimm^{1,2,3,4,*}, **Damian Roqueiro**^{3,4}, **Patrice A. Salomé**^{5,§}, **Stefan Kleeberger**¹, **Bastian Greshake**^{1,§}, **Wangsheng Zhu**⁵, **Chang Liu**^{5,§}, **Christoph Lippert**^{1,§}, **Oliver Stegle**^{1,§}, **Bernhard Schölkopf**⁶, **Detlef Weigel**⁵ and **Karsten M. Borgwardt**^{1,2,3,4,*}

1 Machine Learning and Computational Biology Research Group, Max Planck Institute for Intelligent Systems and Max Planck Institute for Developmental Biology, Tübingen, Germany

2 Zentrum für Bioinformatik (ZBIT), Eberhard Karls Universität Tübingen, Tübingen, Germany

3 Department for Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland

4 Swiss Institute of Bioinformatics, Basel, Switzerland

5 Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany

6 Department of Empirical Inference, Max Planck Institute for Intelligent Systems, Tübingen, Germany

*Correspondence: dominik.grimm@bsse.ethz.ch (Dominik G. Grimm),
karsten.borgwardt@bsse.ethz.ch (Karsten M. Borgwardt)

§ Current addresses: Johann Wolfgang Goethe University of Frankfurt am Main, Germany (Bastian Greshake); Human Longevity, Inc., USA (Christoph Lippert); European Molecular Biology Laboratory, European Bioinformatics Institute, Genome Campus, Hinxton, Cambridge, UK (Oliver Stegle); Department of Chemistry and Biochemistry, UCLA, USA (Patrice A. Salomé); Center for Plant Molecular Biology (ZMBP), University of Tübingen, Germany (Chang Liu)

Supplemental Figures

Species	TaxID	Species Prefix	#Datasets	#Phenotypes	#Covariates
<i>Drosophila melanogaster</i>	7227	DM	1	6	0
<i>Pristionchus pacificus</i>	54126	PP	1	3	4
<i>Arabidopsis thaliana</i>	3702	AT	3	116	0

Supplemental Figure 1. Public and Private Data Repository.

The easyGWAS data repository is divided into a publicly accessible area and one that is only available to registered users. The top panel (A) is the main navigation menu of easyGWAS. A registered user can log in by selecting the “Login” option in the navigation menu. The specific menu options for the public data repository are shown in the left panel (B). The center panel (C) displays the contents of the option selected in the left panel. In this example, it lists all publicly available species. The user can switch between public and private repositories by clicking on “Public Data” or “Private Data” at the top of panel B or through the navigation menu. The “Private Data” section allows the user to upload their own genotype, phenotype, covariate, or gene annotation data. Publicly available genotype, phenotype and covariate data can be downloaded by any user. The view displayed in this figure is obtained by selecting: Public Data [from the main menu]→Public Species [from panel B].

Public Data Private Data Species: *Arabidopsis thaliana*

Public Data

- Public Species
- Public Datasets
- Public Samples
- Public Phenotypes
- Public Covariates

Data Management

- Download Manager
- Upload Manager

Miscellaneous

- Statistics
- Tutorials

General Information

Name: *Arabidopsis thaliana*

NCBI Taxid: 3702

Description: *Arabidopsis thaliana* is a plant. In biological science it is used as a model organism.

Available Datasets

Dataset Name	Build	#Samples	#Chromosomes	#Phenotypes	#Covariates	Shared	Private	Owner	
AiPolyDB (call method 75, Horton et al.)	TAIR9	1307	214051	5	107	0	✘	✘	dggrimm
80 genomes data (Cao et al.)	TAIR9	80	1438752	5	0	0	✘	✘	dggrimm
1001 Genomes Data	TAIR10	1135	6973565	5	9	0	✘	✘	dggrimm


Available Gene-Annotation Sets

Gene Annotation Set Name	Number of Genes
Gene Annotations (TAIR9)	28412
Gene Annotations (TAIR10)	28496

Supplemental Figure 2. Data Repository and Detailed Species View.

The menu on the left contains two sub-panels. The top part (A), in green, allows the user to access detailed information about publicly available species – including those pre-loaded in easyGWAS – datasets, samples, phenotypes, and covariates. The menu at the bottom (B), in sky blue, lists different data management methods. The center panel (C) shows detailed information about the species *Arabidopsis thaliana*. The view displayed in this figure is obtained by selecting: Public Data→Public Species→*Arabidopsis thaliana*.

Sample: TDr-1



General Information A

Name:	TDr-1
ID:	6188
Species:	<i>Arabidopsis thaliana</i>
Dataset:	AtPolyDB (call method 75, Horton et al.)
Country:	SWE
Region:	S Sweden
Latitude:	55.7683
Longitude:	14.1386
Source:	Jakobsson, Mattias
Site:	TDr
Description:	

Additional Meta Information B

Meta Information Field	Meta Information Value
median_intensity	622.0

Publications C

authors	title	pub					
		year	journal	volume	pages	doi	
Matthew Horton et al.	Genome-wide patterns of genetic variation in worldwide <i>Arabidopsis thaliana</i> accessions from the RegMap panel	2012	Nature Genetics	44	212-216	10.1038/ng.1042	<input type="checkbox"/>

Supplemental Figure 3. Detailed Sample View.

The Sample View provides detailed information about a sample. The sample name is displayed at the top. The General Information panel (A) provides details about the sample, in conjunction with a map of its geographic location. The Additional Meta Information panel (B) indicates if there are meta information entries related to the sample. The Publications panel (C) lists the publication(s) associated to the sample. The citation(s) to the publication(s) can be downloaded by marking them (check-mark at the right) and then clicking the download button (top-right corner of panel C). The view displayed in this figure is obtained by selecting: Public Data→Public Samples→*Arabidopsis thaliana*→AtPolyDB→sample name=TDr-1.

Phenotype: K39 PDF

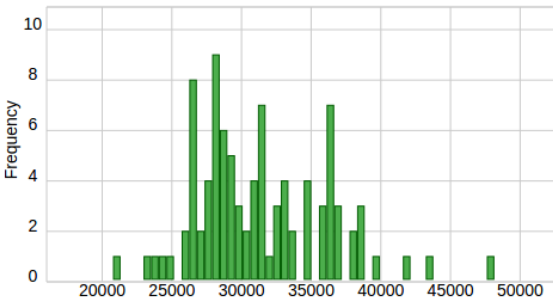
General information A

Phenotype ID:	AT_P_20
Name:	K39
Species:	<i>Arabidopsis thaliana</i>
Dataset:	AtPolyDB (call method 75, Horton et al.)
Variation Type:	continuous
Number of accessions:	93
Category:	ion concentration
Phenotype scoring:	Potassium concentrations in leaves, grown in soil. Elemental analysis was performed with an ICP-MS (PerkinElmer). Sample normalized to calculated weights as described in Baxter et al., 2008
Source:	Ivan Baxter and David E Salt, Purdue University, West Lafayette, Indiana, USA.
Description:	

Additional Meta Information B

Meta Information Field	Meta Information Value
growth_conditions	20°C, 16 hrs daylight

Phenotype Distribution 50 PDF C



Shapiro-Wilk Test Statistic

The Shapiro-Wilk-Test tests the null hypothesis that the data was drawn from a normal distribution.

Test-Statistic:	9.612e-01
P-Value:	7.388e-03







Publications PDF D

authors	title	pub year	journal	volume	pages	doi
Susanna Atwell et al.	Genome-wide association study of 107 phenotypes in <i>Arabidopsis thaliana</i> inbred lines	2010	Nature	465	627-631	10.1038/nature08800

Supplemental Figure 4. Detailed Phenotype View.

The Phenotype View shows detailed information about each phenotype. The phenotype name is displayed at the top with a download button (top-right corner) that allows the user to download the information on screen. The General Information panel (A) provides details about the phenotype. The Additional Meta Information panel (B) indicates if there are meta information entries related to the phenotype. The Phenotype Distribution panel (C) plots a histogram with the distribution of the phenotype values. The number of bins in the histogram can be set with the drop-down list at the top-right corner of panel C. The user can download

the plot as a PDF document by clicking the button to the right of the drop-down list. In the figure, 50 bins were used to create the plot. Results are also shown for a Shapiro-Wilk test of normality (see Suppl. Text 8 for more details). Finally, the Publications panel (**D**) lists the publication(s) associated to the phenotype. The citation(s) to the publication(s) can be downloaded by marking them (check-mark at the right) and then clicking the download button (top-right corner of panel D). The view displayed in this figure is obtained by selecting: Public Data→Public Phenotypes→Arabidopsis thaliana→AtPolyDB→Phenotype ID=AT_P_20.

Download Manager							
Species	Dataset Name	#Samples	#SNPs	#Chromosomes	#Phenotypes	#Covariates	Download
 <i>Drosophila melanogaster</i>	<i>Drosophila Genetic Reference Panel (DGRP, Mackay et al.)</i>	172	2476799	5	6	0	
 <i>Arabidopsis thaliana</i>	<i>80 genomes data (Cao et al.)</i>	80	1438752	5	0	0	
 <i>Arabidopsis thaliana</i>	<i>AtPolyDB (call method 75, Horton et al.)</i>	1307	214051	5	107	0	

Supplemental Figure 5. Download Manager.

The Download Manager allows the user to download publicly available datasets. It can be accessed with the following menu options: Public Data→Download Manager.

Upload a new Genotype

A

Select a publicly available species or create a new one:

Species: + Add new species

Select a publicly available Gene Annotation Set (Optional):

Gene Annotation Set:

B

Add information about your dataset (this is needed to upload the data correctly):

Dataset name:

Dataset version:

Dataset description (optional):

C

Data Upload Options

The data file has to be a ZIP file containing genotype files in PLINK format. Please see the FAQ for more details!

Check if the following files are provided:

Genotype PED file (genotype.ped)	<input checked="" type="checkbox"/>
Genotype MAP file (genotype.map)	<input checked="" type="checkbox"/>
Phenotypes file (phenotypes.pheno)	<input type="checkbox"/>
Covariate file (covariates.cov)	<input type="checkbox"/>
Gene annotation file (geneinfo.gff)	<input type="checkbox"/>

Please select the data upload method of your choice:

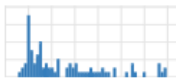
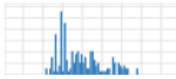
Choose file from your Dropbox:
 Choose from Dropbox

Supplemental Figure 6. Upload Manager.

The Upload Manager is the interface to upload new genotype data, phenotypes, covariates, or gene annotation sets. In order to upload a dataset containing an entire genome-wide association study (including genotype and phenotype information), the data have to be in PLINK format and stored in a single .zip file. The top panel **(A)** indicates to which species and genome version the newly uploaded data will belong. The panel in the middle **(B)** prompts the user to annotate the dataset with a name and version. In the bottom panel **(C)**, the user specifies what types of data files are included in the .zip file. Due to the large size of genetics datasets, the uploads are managed via a personal Dropbox account (Dropbox and the Dropbox logo are trademarks of Dropbox, Inc.). The button “Choose from Dropbox” opens a window that allows the user to select the .zip file to be uploaded. The Upload Manager can be launched by selecting: Private Data→Upload Manager.

1. Species / 2. Phenotypes / 3. Covariates / 4. SNPs / 5. Algorithm / 6. Summary A

Transform/Normalize your Phenotypes (Optional) B

Phenotype Name	Variation Type	Shapiro-Wilk #	Histogram	Transformation ⁺
Secondary Dormancy	continuous	3.2e-08		None
FT10	continuous	2.3e-07 (7.1e-11)		log10 transformation

[#] The p-value of the Shapiro-Wilk-Test, which tests the null hypothesis that the data was drawn from a normal distribution. If the Phenotype is transformed the p-value in parentheses is the original one.
⁺ The number of available transformations depend on the distribution of the Phenotype (automatically-determined).

C

« Back Continue »

Supplemental Figure 7. GWAS Wizard, Step to Normalize Phenotypes.

Once the GWAS Wizard has been launched, the user is guided through a sequence of steps to specify all the parameters needed in an analysis. The top panel (A) details all the steps that are part of the GWAS Wizard. Having selected the species in step 1, the user is currently in the second part of step 2. After selecting the phenotypes, this view allows the user to apply a transformation to them in panel (B). If a regression method will be selected in step 5, it is recommended to apply one type of transformation (see Suppl. Text 8 for details about available transformations). A distribution of values is shown for each phenotype as a histogram. If a transformation is applied, the histogram is recomputed on the fly. Additionally, a Shapiro-Wilk test of normality is computed on the (transformed) phenotype values. The bottom panel (C) are the navigation buttons within the wizard. The user can return to previous steps with the “Back” button and adjust parameters. At every step, when all the information needed has been completed, the “Continue” button allows the user to move on to the next step in the wizard. The view displayed in this figure is obtained by selecting: GWAS Center→New GWAS→Species=Arabidopsis thaliana; Dataset=AtPolyDB; Gene Annotation=TAIR10→Public Phenotype=“Secondary Dormancy” and “FT10”→Transformation for FT10=log10.

Temporary Experiments

Temporary experiments are available for 48h! To store them permanently please save the experiments. 1

Type	Name	Species	Phenotype	Algorithm	date	Running
	Experiment 3	Arabidopsis thaliana	<i>avrRpm1</i>	Logistic Regression	Oct. 14, 2016, 05:20 PM	
	GWAS-Comparison 0	Arabidopsis thaliana	--	Comparison	Oct. 14, 2016, 05:20 PM	Done
	Experiment 1	Arabidopsis thaliana	Secondary Dormancy	FaSTLMM	Oct. 14, 2016, 05:20 PM	Done
	Experiment 2	Arabidopsis thaliana	Fe56	Linear Regression	Oct. 14, 2016, 05:20 PM	Done

Supplemental Figure 8. Temporary History.





The My Temporary History View allows the user to access all recently submitted analyses (also referred to as experiments). Experiments can be accessed through this view for a period of up to 48 hours, after which they will be automatically deleted. The top panel (A) allows for the saving or deletion of the results of the experiment(s) selected by the user. Saving experiments is shown in Suppl. Figure 9. The progress showing the number of unfinished experiments is displayed in the middle panel (B). The bottom panel (C) lists all the experiments in the user's temporary history. The column type shows an icon which, in the figure, is used to differentiate the traditional GWAS (in blue) from the comparative analysis of GWAS (in red). The check-boxes on the right are used in conjunction with the download/delete buttons in panel A.

Save your temporary experiments in your private history

Add your experiments to a Project A

Select a Project (default: General): + Add new project

Update your Experiment Names B

Type	Experiment Name	Phenotype	Covariates	Algorithm	Date
	<input type="text" value="Experiment 3"/>	<i>avrRpm1</i>	None	Logistic Regression	Oct. 14, 2016, 12:16 p.m.
	<input type="text" value="GWAS-Comparison 0"/>	--	None	Comparison	Oct. 13, 2016, 10:18 p.m.
	<input type="text" value="Experiment 1"/>	Secondary Dormancy	None	FaSTLMM	Oct. 14, 2016, 12:16 p.m.
	<input type="text" value="Experiment 2"/>	Fe56	None	Linear Regression	Oct. 14, 2016, 12:16 p.m.

C

Supplemental Figure 9. Save Experiments Permanently into GWAS Projects.

For registered users, the experiments can be saved into the permanent (private) area of the user. This view shows how to save the experiments displayed in Suppl. Figure 8. The top panel (A) allows the user to save the experiments as part of a larger project. Projects can be shared among users. In the middle panel (B), the experiments can be renamed. Finally, in the bottom panel (C) the button “Save experiment” confirms the process. The button “Back” takes the user back to the view shown in Suppl. Figure 8.

Supplemental Figure 10. easyGWAS Data Sharing Dialog.

easyGWAS provides a straightforward way to share GWAS projects with other users and collaborators. To do so, the user has to select the GWAS project and click the sharing button. After typing in the email address of a registered collaborator, the project and its data are shared.

Make your projects and experiments public to the scientific community

Projects have to be approved by an administrator before they will be published!
Please choose meaningful names for your project, experiments, phenotypes and/or covariates.
All private data which is used for this project will be published as well!

Update your Project Information:
Update Project name (please choose an informative name for your project)
Genotype download allowed?

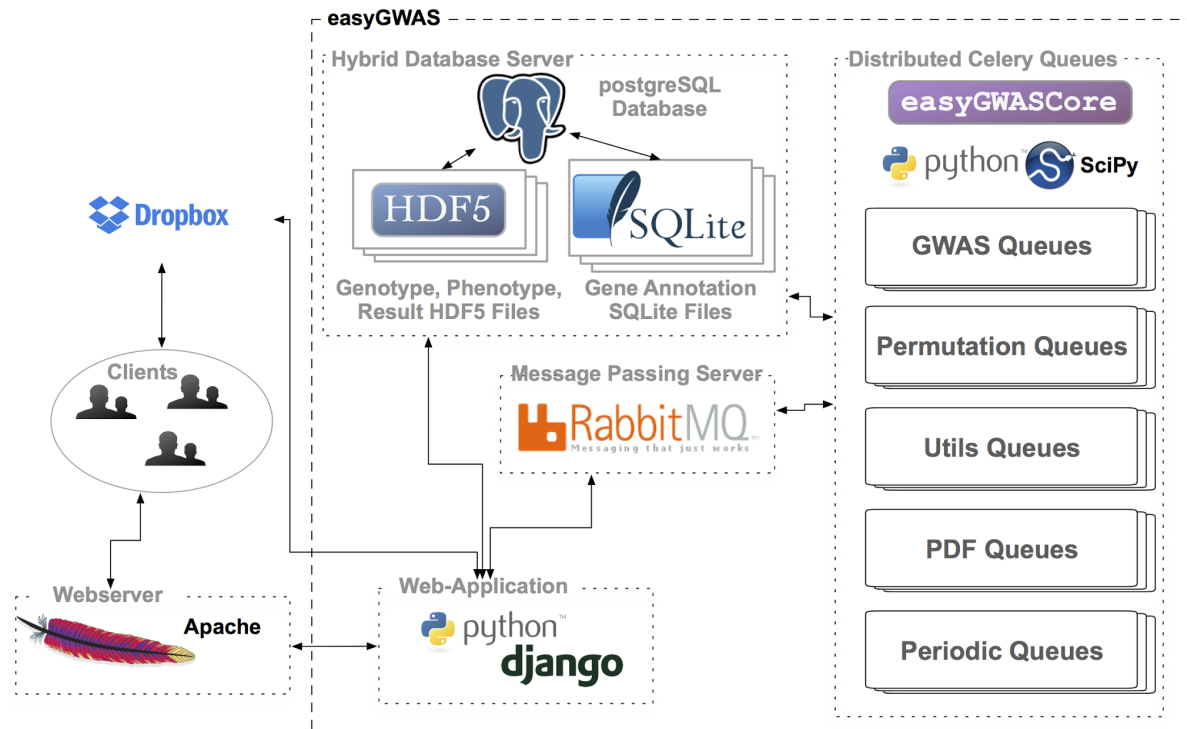
Update your Experiment, Phenotype and Covariate Names if possible

Experiment Name	Phenotype	Covariates	Algorithm	Date
<input type="text" value="Test Experiment"/>	<input type="text" value="Test Phenotype"/>	None	EMMAX	May 20, 2015, 7:26 p.m.

« Back Publish Project »

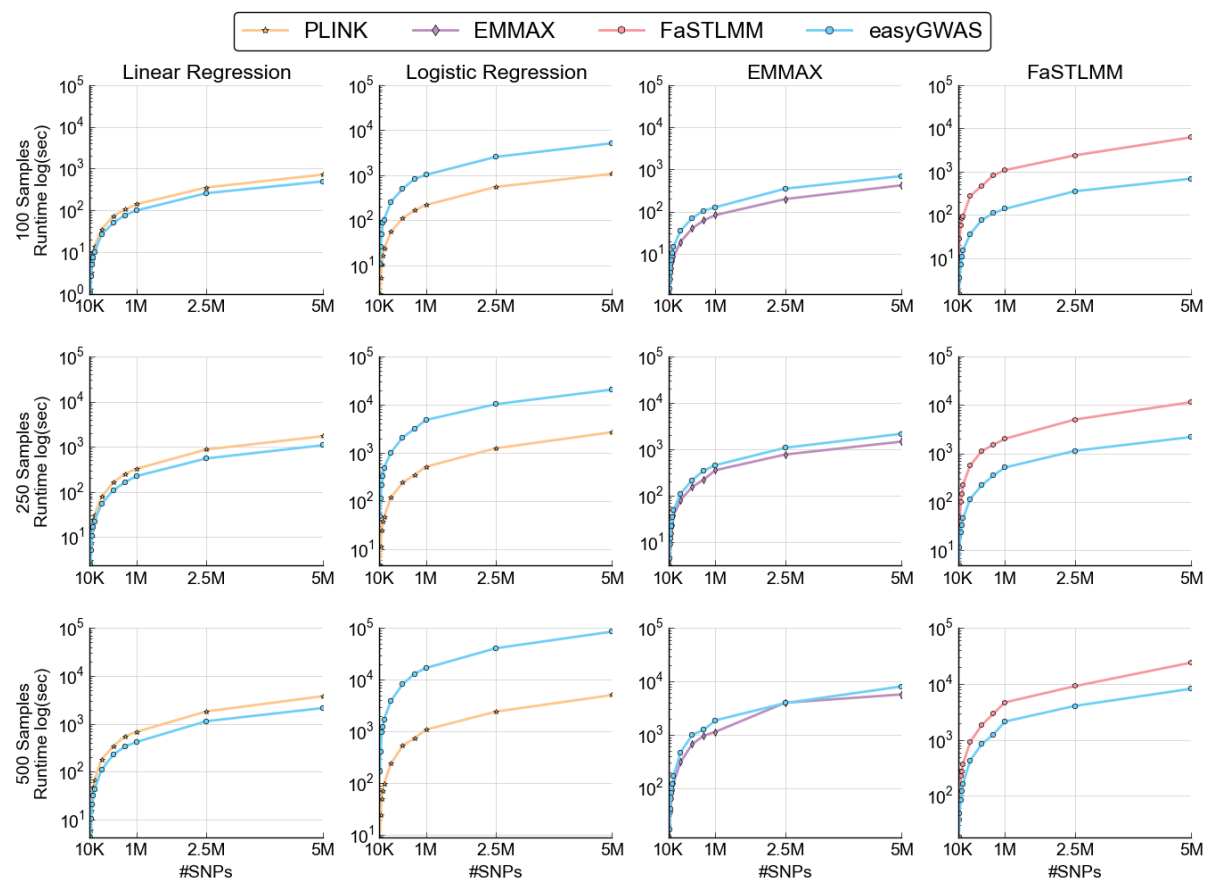
Supplemental Figure 11. GWAS Project Publishing Inquiry Form.

easyGWAS provides a publishing inquiry form. Here, the user can inquire to make their GWAS project, datasets or phenotypes publicly available to the scientific community. Before projects are made public, an easyGWAS administrator has to approve them.



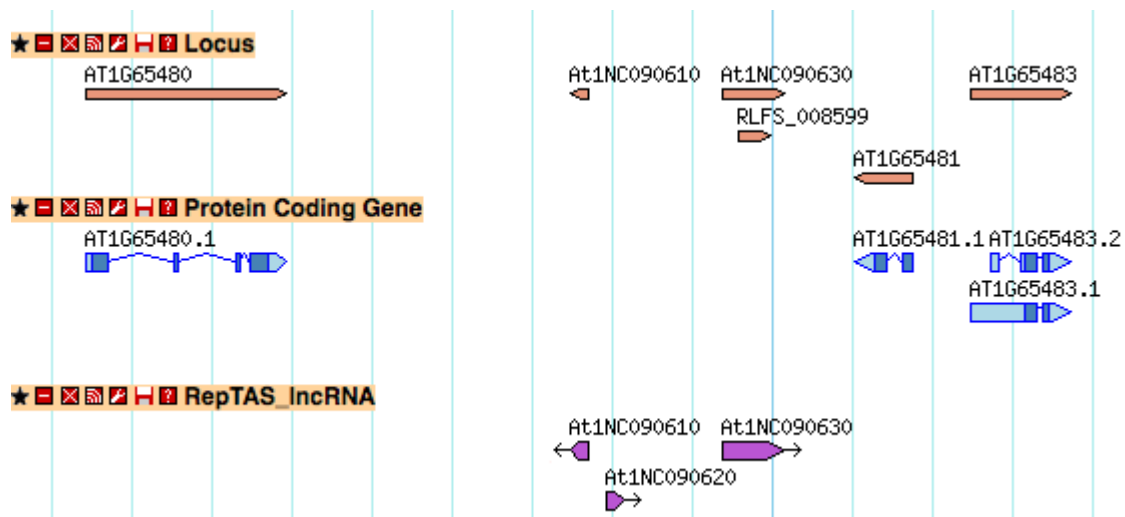
Supplemental Figure 12: Schematics of the easyGWAS Architecture.

Illustration of the internal architecture of easyGWAS including the hybrid database model and different task queues. Communication between the web application and queues is established via a RabbitMQ message passing server. Task queues can be distributed over different computing nodes. The hybrid database can be accessed from the web application, as well as from the different task queues. Users can link their personal Dropbox account to easyGWAS to integrate large genotype datasets (Dropbox and the Dropbox logo are trademarks of Dropbox, Inc.).



Supplemental Figure 13: Runtime Comparison between State-of-the-Art Tools and easyGWAS.

Comparison of four state-of-the-art algorithms implemented in PLINK, EMMAX and FaST-LMM to the easyGWAS implementation. Runtime includes data preprocessing for all tools. Number of SNPs range between 10,000 and five million. Number of samples are varied between 100, 250, and 500.



Supplemental Figure 14: Non-coding RNAs downstream of *FT* (AT1G65480).

The *Arabidopsis thaliana* case study discusses three SNPs that were found to be associated to the rosette leaf number phenotype. The SNPs are located in Chr1 at positions 24337820, 24338990 and 24339560 (see rows 175-177 of **Supplemental Data Set 1**). These SNPs are located downstream of *FT* (id AT1G65480, *FLOWERING LOCUS T*), and they overlap with non-coding RNAs At1NC090610, At1NC090620 and At1NC090630 (marked in purple). The figure is a screenshot of GBrowse (Stein et al., 2002), extracted from the Plant Long noncoding RNA Database (Jin et al., 2015).

Supplemental Tables

Supplemental Table 1. Case Study Results in *Arabidopsis thaliana* using Bonferroni.

Comparison results of nine GWAS performed using a linear mixed model and a Bonferroni ($\alpha=5\%$).

Gene	Brief Gene Description [Other Names]	Chr	Position	P-Value	Phenotype
At5g10100	Haloacid dehalogenase-like hydrolase (HAD) superfamily protein [TPPI, TREHALOSE-6-PHOSPHATE PHOSPHATASE I]	Chr5	3161401	1.932e-08	DTF2
		Chr5	3161477	1.728e-08	DTF2
At4g00752	UBX domain-containing protein	Chr4	317979	2.725e-08	RL
At4g00730	Encodes a homeodomain protein of the HD-GLABRA2 group. Involved in the accumulation of anthocyanin and in root development. Loss of function mutants have increased cell wall polysaccharide content. [AHDP, ANL2, ANTHOCYANINLESS 2, ARABIDOPSIS THALIANA HOMEODOMAIN PROTEIN]	Chr4	299748	1.161e-08	RL
At4g00630	Encodes a K(+)/H(+) antiporter that modulates monovalent cation and pH homeostasis in plant chloroplasts or plastids. [ATKEA2, K+ EFFLUX ANTIPORTER 2, KEA2]	Chr4	262690	2.176e-08	RL
No Gene Found		Chr4	247215	1.868e-08	RL

DTF2: days until the inflorescence stem elongated to 1 cm; RL: rosette leaf number;

Supplemental Table 2. Phenotype Information for the Case Study.

Used phenotypes for the *Arabidopsis thaliana* case study. The table provides information about which phenotype was transformed with which method.

Experiment Name	Link
DTF1	https://easygwas.ethz.ch/gwas/results/manhattan/view/ef55da11-bea8-431d-afef-e0ecce5487fa/
DTF2	https://easygwas.ethz.ch/gwas/results/manhattan/view/bc199ae8-08c0-48d4-9ca7-6f7a8ca8034a/
DTF3	https://easygwas.ethz.ch/gwas/results/manhattan/view/928530a0-74aa-4c13-95dd-aa2c3d26fc44/
RL	https://easygwas.ethz.ch/gwas/results/manhattan/view/57a7cf18-cb0f-408a-8954-49f94d1bfc47/
CL	https://easygwas.ethz.ch/gwas/results/manhattan/view/0c981d39-1c09-42a6-86fe-5c0bfa54dbce/
Diameter	https://easygwas.ethz.ch/gwas/results/manhattan/view/7b30a745-fd39-4510-97db-432ddf557196/
RBN	https://easygwas.ethz.ch/gwas/results/manhattan/view/b8fbd333-4590-4764-9452-337430e6c871/
CBN	https://easygwas.ethz.ch/gwas/results/manhattan/view/074a260c-f476-45ae-b11c-6c8d83b99dbf/
Length	https://easygwas.ethz.ch/gwas/results/manhattan/view/2f66a2a1-4ead-42ca-a93e-4a1b4d423e16/
GWAS-Comparison	https://easygwas.ethz.ch/comparison/results/manhattan/view/2c8da231-96ff-4f28-a17e-fd0e3510d8e1/

Key to experiment abbreviations

DTF1: days until emergence of visible flowering buds in the center of the rosette from time of sowing

DTF2: days until the inflorescence stem elongated to 1 cm

DTF3: days until first open flower

RL: rosette leaf number

CL: cauline leaf number

Diameter: diameter of rosette (end point, after flowering)

RBN: rosette branch number

CBN: cauline leaf number

Length: length of main flowering stem

Supplemental Table 3. Pitfalls when conducting intersection analyses or meta-analyses of GWAS results

The table highlights common pitfalls when conducting an intersection analysis or a meta-analysis with easyGWAS. It stresses the caution the user must observe in regards to the assumptions made and the final interpretation of the results.

Pitfall	Description
Violation of independence	An important assumption in a meta-analysis is that different datasets are sampled independently. An overlap of individuals between different datasets may lead to a similar association signal in each dataset, thereby artificially confirming this signal and leading to a spurious association reported by 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 which 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 intersection 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.
Genetic heterogeneity I	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 a genome-wide association study on a single dataset. If two or more phenotypes are significantly correlated with kinship, then they may show shared genetic association signals due to this confounding effect. 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.
Genetic heterogeneity II	When combining datasets in a meta-analysis, one can assume a fixed-effects model in which the effect of a genetic variant is assumed to be the same in all datasets (see Supplemental Text 9). Although the fixed-effects model is the most popular approach to meta-analysis, if the datasets exhibit large genetic heterogeneity, this will result in a violation of the above mentioned assumption. As a consequence of the latter, it will yield inflated p-values.
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). It is not well understood how exactly does publication bias affect 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 markers 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 leads either to the need of 1) restricting the analysis to SNPs that are present in all datasets, 2) analyzing hits on the gene level rather than the SNP level, or 3) imputing missing SNPs in each dataset. Options 1) and 2) are currently offered by easyGWAS. 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 an intersection- 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.

Supplemental Table 4. Available Genotype Encodings.

Different genotype encodings available in easyGWAS for heterozygous genotypes. The default is the additive encoding

Encoding	Major	Heterozygous	Minor
Additive	0	1	2
Recessive	0	0	1
Dominant	0	1	1
Overdominant	0	1	0

Supplemental Table 5. Transformation Methods

Overview of different methods to transform phenotypes. Different constraints apply to each method. The GWAS wizard determines on-the-fly which transformation method could be applied to which phenotype. Refer to Suppl. Text 8 for more details about each method.

Transformation	Variation Type	Constraint	Description
Zero Mean	continuous, categorical, binary	-	Mean of data is set to 0
Zero Mean & Unit Variance	continuous, categorical, binary	-	Mean of data is set to 0 and variance to 1
Square root (sqrt)	continuous, categorical	-	Square root of data
Logarithmic (log10)	continuous, categorical	No "0" in data allowed	Logarithm base 10 of data
Box-Cox	continuous, categorical	No "0" in data allowed	Box-Cox transformation (Box and Cox, 1964)
Dummy Variable	categorical	data has to be categorical	Encode categorical data into dummy variables

Supplemental Table 6. Phenotype Information for the Case Study.

Used phenotypes for the *Arabidopsis thaliana* case study. The table provides information about which phenotype was transformed with which method.

Phenotype Name	#Samples	Shapiro-Wilk Test (Untransformed)	Transformation	Shapiro-Wilk Test (Transformed)
DTF1	936	4.8e-16	Box-Cox	1.6e-11
DTF2	931	3.3e-17	Box-Cox	1.1e-11
DTF3	923	1.8e-11	Box-Cox	1.8e-11
RL	850	1.7e-07	Box-Cox	3.4e-06
CL	904	1.8e-18	Box-Cox	7.4e-01
Diameter	656	4.9e-04	Box-Cox	1.9e-02
RBN	674	9.0e-11	None	-
CBN	677	1.6e-18	Square root	3.6e-11
Length	680	6.4e-13	Box-Cox	2.3e-04

Key to phenotype abbreviations

DTF1: days until emergence of visible flowering buds in the center of the rosette from time of sowing

DTF2: days until the inflorescence stem elongated to 1 cm

DTF3: days until first open flower

RL: rosette leaf number

CL: cauline leaf number

Diameter: diameter of rosette (end point, after flowering)

RBN: rosette branch number

CBN: cauline leaf number

Length: length of main flowering stem

Supplemental Text

Supplemental Text 1: easyGWAS Data Repository

The easyGWAS data repository contains detailed information about available or integrated species, datasets, samples, phenotypes, and covariates. For each of these data types, different views show additional information about the species or the datasets, such as the total number of available samples and SNPs (see **Supplemental Figure 2** or <https://easygwas.ethz.ch/data/public/species/view/1/>).

The sample view provides meta-information, such as its origin or its source (**Supplemental Figure 3** or <https://easygwas.ethz.ch/samples/public/view/961/>). Meta-information varies for each sample and between datasets and especially species. Therefore, we allow the addition of different types of meta-information and do not limit the user to a predefined set (**Supplemental Figure 3**). Similar views are available for datasets, phenotypes, and covariates. An example for the phenotype view is illustrated in **Supplemental Figure 4**, or can be accessed online via <https://easygwas.ethz.ch/data/public/phenotypes/view/52/>. The detailed view for covariates is similar to those of phenotypes. Phenotypic or covariate measurements are illustrated as histograms. A Shapiro-Wilk test (Shapiro and Wilk, 1965) is provided for the null hypothesis of whether the data could have been drawn from a normal distribution.

The Data Repository, in addition contains a *Download Manager* for publicly available datasets in the widely used PLINK (Purcell et al., 2007) format (**Supplemental Figure 5** or <https://easygwas.ethz.ch/down/1/>).

In addition, easyGWAS provides an *Upload Manager* to support the integration of user-specific genotype, phenotype, covariate or gene annotation data for an arbitrary species (**Supplemental Figure 6**). Upload of private data is only available for registered users. Initially, each user has 5GB of storage available for private data integration. Users can either upload imputed GWAS datasets in PLINK format, or add new phenotypes, covariates or gene annotation sets to existing datasets. Further, users are also allowed to upload summary statistics of already-computed GWAS for further meta-analysis, comparison, or simply for visualization. Tutorials on how to upload data can be found in the online FAQ (<https://easygwas.ethz.ch/faq/>). To upload whole GWAS datasets (including genotype and phenotype data), the data has to be in PLINK format and stored in a single .zip file. Due to the large size of genetics datasets, new ones are uploaded via a personal Dropbox account. easyGWAS can then fetch the data from the personal Dropbox account. (Dropbox and the Dropbox logo are trademarks of Dropbox, Inc.).

In addition, the easyGWAS *Upload Manager* allows the automatic import of public phenotypes from AraPheno¹, a central repository for population scale phenotype data from *Arabidopsis thaliana* (Seren, Grimm et al. 2016).

¹ <https://arapheno.1001genomes.org>

Supplemental Text 2: easyGWAS Wizard

The easyGWAS wizards guide the user through all necessary steps to successfully create a GWAS experiment, meta-analysis, or comparison. In the following, we illustrate the wizard to create a typical GWAS experiment. The other two wizards for meta-analysis and comparisons are similarly straightforward.

To create a new GWAS experiment, the user first has to navigate to the *GWAS Center* and start the wizard by clicking on “New GWAS”. Here, the user has to select an existing species, dataset, and gene annotation set (if available). This can be either a publicly available dataset for an existing species or a privately integrated one. In the second step, up to five different phenotypes can be selected. The wizard will help the user to find the correct phenotype by offering an auto-completion for all available or shared phenotypes for the selected species and dataset. For each selected phenotype, detailed information about the data distribution and a Shapiro-Wilk (Shapiro and Wilk, 1965) are given (see **Supplemental Figure 7**). Here, the user has the opportunity to select different transformation methods to normalize the phenotypic data. The Shapiro-Wilk and histograms are updated dynamically and in real time for interactive exploration. In the next step, the user has the opportunity to add their experiments’ covariates, which can be used to account for various confounding factors, such as environmental effects or simple forms of population stratification. Later, the wizard offers a selection of different algorithms to perform association tests between the selected genotype and phenotypes (see Methods). In addition, different filtering and genotype encoding options are provided. Here, the user can filter SNPs that do not fulfil a certain allele frequency. For heterozygous genotypes, different genotype encodings can be selected. The default encoding, known as the *additive* encoding represents the major allele as 0, the heterozygous allele as 1, and the minor allele as 2. An overview of all available encodings is given in **Supplemental Table 4**. At the end, a summary page is provided such that the user can check all inputs and adjust them if necessary. Eventually, the user can submit his or her experiments to the easyGWAS computation queues. Detailed tutorials with examples on how to apply the wizard can be found in the online FAQ (<https://easygwas.ethz.ch/faq/>).

Supplemental Text 3: easyGWAS GWAS History

All submitted, running, or finished experiments are collected in the user’s temporary history (“*My temporary history*”) for monitoring, as illustrated in **Supplemental Figure 8**. Each user can submit a maximum of five experiments simultaneously. Experiments will be stored for 48-hours in the temporary history and then deleted. Nevertheless, users can save their experiments permanently. For this, experiments can be grouped into GWAS projects and stored in the user’s profile (**Supplemental Figure 9**). All stored experiments are listed in the “*My Experiments*” view. In this view, experiments can be filtered by project, as well as regrouped into new or other projects. An overview of all available projects can be found in the “*My Projects*” view. Projects can be shared with other registered users by entering the e-

mail address in the sharing form (**Supplemental Figure 10**), with the new projects and all associated experiments, datasets, phenotypes, and covariates automatically linked to the others user's profile.

To make projects publicly available to the scientific community, easyGWAS offers a publishing inquiry form (**Supplemental Figure 11**). An easyGWAS administrator has to first approve the inquiry before a GWAS project is made available. 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. Published projects and experiments cannot be changed or deleted subsequently by the user who agrees that the data and experiments can be reused by others.

Supplemental Text 4: Step-by-Step Procedures to Reproduce the Content of Figures

In the easyGWAS manuscript we show many screenshots of different views and visualizations. Here, we show how these different views can be accessed or generated and give a detailed description of the different panels in these views.

How to access the content of Figure 2?

The results of the genome-wide association study illustrated in Figure 2 can be accessed at: <https://easygwas.ethz.ch/gwas/results/manhattan/view/e908dfaf-7f4c-4315-8951-35e8466772a1/>

How to reproduce the contents of Figure 2?

To reproduce the results shown in Figure 2, the user has to follow the step-by-step procedure below:

1. Login to easyGWAS
2. Navigate to "GWAS Center" in the top menu
3. Click on "New GWAS" in the left side menu to start the GWAS wizard
4. Select the species "*Arabidopsis thaliana*"; the dataset "AtPolyDB" and the gene annotation set "TAIR10" and click "Continue"
5. Select the public phenotype "avrRpm1" using the autocomplete search function and click "Continue"
6. The next step in the wizard asks to select a phenotype transformation. Because the phenotype chosen in the previous step is binary, the transformation will not yield a normal distribution of phenotype values. It is therefore not necessary to apply a transformation and this step can be skipped by clicking "Continue".
7. Next, different covariates or principal components can be added. This step can be skipped for this analysis. In general – and whenever available – adding covariates might be necessary to account for confounding factors such as environmental effects.
8. Select all available SNPs by clicking "Continue"
9. Select a 10% minor allele frequency (MAF) filter and the algorithm "EMMAX". Click "Continue"

10. In the final step of the wizard, all inputs can be checked and the genome-wide association study can be submitted to the computation queues.
11. After submitting the experiment the user is redirected to the “Temporary Experiment” view
12. The results can be accessed by clicking on the experiment name after they are finished

How to access the content of Figure 3?

The detailed SNP view of Figure 3 can be accessed via the following link:

<https://easygwas.ethz.ch/gwas/results/snp/detailed/8557bdde-aa8a-4615-a643-ccce51a4edc0/Chr4/429928/>

How to access an arbitrary detailed SNP view?

To access the “Detailed SNP View” for any given SNP, the user can click on a dot – which represents the logarithm base 10 of the p-value of a SNP – in the Manhattan plot. For example click on any SNP in one of the following Manhattan plots:

<https://easygwas.ethz.ch/gwas/results/manhattan/view/e908dfaf-7f4c-4315-8951-35e8466772a1/>

How to access the content of Figure 4?

Figure 4 shows a snapshot of the results of a comparative analysis of GWAS. The results shown in Figure 4 can be accessed via the following link:

<https://easygwas.ethz.ch/comparison/results/manhattan/view/51a98e12-fb0c-4c3e-a0a0-94feb35a4ae6/>

How to reproduce the contents of Figure 4?

To reproduce the comparative analysis of multiple GWAS shown in Figure 4 users have to follow the steps listed below:

1. Login to easyGWAS
2. Navigate to “GWAS Center” in the top menu
3. Click on “New Intersection Analysis” in the left side menu to start the GWAS wizard
4. Select the species “*Arabidopsis thaliana*” and the gene annotation set “TAIR10”. Click “Continue”
5. When having to select the experiments to compare, select the following six public phenotypes using the autocomplete search form: Bs, FLC, FT16, Hiks1, Nickel (Ni60), Storage 28 days. The experiments are named after the phenotypes. Press Enter after selecting one experiment and start typing the name of the next one. Click “Compare GWAS”.
6. The results of the GWAS comparison can be accessed in the “Temporary Experiment” view after the computations are done

How to access the content of Figure 5?

Figure 5 shows the phenotype-phenotype correlation plot created in the comparative analysis of GWAS in our case study. The plot can be found here:

<https://easygwas.ethz.ch/comparison/results/manhattan/view/2c8da231-96ff-4f28-a17e-fd0e3510d8e1/>

How to reproduce the content of Figure 5?

To reproduce the comparative analysis of GWAS from the case study as shown in Figure 5, the user has to follow the steps listed below:

1. Login to easyGWAS

2. Navigate to “GWAS Center” in the top menu
3. Click on “New Intersection Analysis” in the left side menu to start the GWAS wizard
4. Select the species “*Arabidopsis thaliana*” and the gene annotation set “TAIR10”. Click “Continue”
5. In a similar way as we did in Step 5 for Figure 4, select the following nine public phenotypes (experiments) using the autocomplete search form for public phenotypes: DTF1, DTF2, DTF3, RL, CL, CBN, RBN, Diameter, Length. Click “Compare GWAS”.
6. The results of the GWAS comparison can be accessed in the “Temporary Experiment” view after the computations are done

How to access the contents of Figure 6?

The “Detailed SNP View” in Figure 6 can be accessed via the following link:

<https://easygwas.ethz.ch/gwas/results/snp/detailed/57a7cf18-cb0f-408a-8954-49f94d1bfc47/Chr1/24338990/>

Supplemental Text 5: Tips on how to Access Certain Panels

How to access QQ-Plots from GWAS?

To access QQ-Plots of a genome-wide association analysis, click “QQ-Plot” in the “GWAS Result View” submenu:

<https://easygwas.ethz.ch/gwas/results/qplots/e908dfaf-7f4c-4315-8951-35e8466772a1/>

How to access a list of SNP annotations per Chromosome?

To access a list of SNP annotations of a genome-wide association analysis, click “SNP Annotations” in the “GWAS Results View” submenu:

<https://easygwas.ethz.ch/gwas/results/snpannotations/e908dfaf-7f4c-4315-8951-35e8466772a1/>

How to access a detailed summary of a genome-wide association analysis:

To access a detailed summary of genome-wide association analysis, click “Summary/Download”:

<https://easygwas.ethz.ch/gwas/results/summary/e908dfaf-7f4c-4315-8951-35e8466772a1/>

How to access a view showing shared associations between GWAS:

To access a detailed overview about the top x associated SNPs shared between different GWAS, click “Shared Associations” in the “GWAS Comparison View”:

<https://easygwas.ethz.ch/comparison/results/sharedsignal/view/dfaa2551-7b2d-4e3d-9170-6522966b7d2a/>

How to access a view showing genes with shared associations between GWAS:

To access a detailed overview about genes that share associated SNPs between different GWAS, click “Shared Genes” in the “GWAS Comparison View”:

<https://easygwas.ethz.ch/comparison/results/gene/view/dfaa2551-7b2d-4e3d-9170-6522966b7d2a/>

Supplemental Text 6: Variance Explained for Linear Mixed Models

easyGWAS computes, in a 10-fold cross-validation, which parts of the phenotypic variance could be attributed to the genetic contribution (random effect), using the kinship matrix only, and to the covariates (fixed effects). For this purpose, the data is split, to the extent possible, into 10 equal subsets. Then, nine subsets are combined to train a linear mixed model using only the kinship matrix and the covariates. The remaining subset is used to predict the phenotype $\hat{\mathbf{y}}$. This procedure is repeated 10 times. Predictions for $\hat{\mathbf{y}}$ are obtained by summing up the contributions of the random and fixed effects as follows:

$$\hat{\mathbf{y}} = \mathbf{C}_{test}\hat{\boldsymbol{\beta}} + \mathbf{K}_{test} \left(\mathbf{K}_{train} + \hat{\boldsymbol{\delta}}I \right)^{-1} \left(\mathbf{y}_{train} - \mathbf{C}_{train}\hat{\boldsymbol{\beta}} \right),$$

where \mathbf{C} are the included covariates (or a vector of ones if no covariates are included), \mathbf{K} is the kinship matrix, and $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\delta}}$ are the estimated parameters from the training step. The indices $_{train}$ and $_{test}$ indicate which subset the data are coming from. Eventually, the variance explained is computed as follows:

$$v(\mathbf{y}_{test}, \hat{\mathbf{y}}) = 1 - \frac{Var(\mathbf{y}_{test} - \hat{\mathbf{y}})}{Var(\mathbf{y}_{test})}$$

Supplemental Text 7: Procedure to Measure Dependence Between Phenotype and Population Structure

To measure the statistical dependence between a phenotype and population structure, we employ the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005), a kernel-based multivariate measure of statistical dependence. An empirical estimate of HSIC between two random variables – in our case phenotype and kinship – is obtained by computing a kernel matrix on the samples. This kernel matrix on the phenotypic values $\mathbf{L}(i, j)$ is computed by linear kernel, that is the product of the phenotypes y_i and y_j , of individual i and j . The kernel matrix representing kinship is the realized relationship matrix \mathbf{K} , that is, $\mathbf{K}(i, j)$ an scalar product between the genotypes of individuals i and j .

We obtain an empirical estimate of HSIC based on \mathbf{L} and \mathbf{K} , and perform 100,000 random permutations of the phenotypes and recompute HSIC in order to obtain an empirical null distribution of HSIC values under the assumption of no dependence between phenotype and population structure. We compute a p-value for the phenotype y based on this null distribution, and report a statistically significant association if it passes the Bonferroni-corrected significance threshold ($\alpha = 0.05$ divided by the number of phenotypes studied). Phenotypes that show a statistically significant association are shown in red in easyGWAS.

Supplemental Text 8: Transformation Methods

In regression methods, the residuals are assumed to follow a Gaussian distribution. As described in the main document², easyGWAS implements three different regression algorithms: linear regression, logistic regression and two flavors of linear mixed models. In practice, the assumption of residuals being normally distributed does not always hold and a pre-processing of the phenotype is commonly performed to make their values as Gaussian as possible (Fusi et al., 2014). In easyGWAS this pre-processing is referred to as "transformation" and six different transformation methods are implemented:

- Zero mean: subtracts, from each phenotype value, the mean of the phenotype
- Zero mean and unit variance: after subtracting the mean, divide each value by the standard deviation of the phenotype
- Square root: compute the square root of each value
- Logarithmic: compute the logarithm base 10 of each value
- Box-Cox: apply the Box-Cox transformation. As stated in their original paper (Box and Cox, 1964), this transformation requires one parameter λ . When λ tends to zero, the effect is similar to that of a logarithmic transformation. The implementation in easyGWAS does not require the user to specify the value of λ , but optimizes it instead. If unsure of what transformation to apply, the Box-Cox transformation is a safe first guess as it is the most frequently used transformation in regression analyses.
- Dummy variable: encodes categorical data into dummy variables. For example, if phenotype values are 1, 2 and 3, then three dummy variables v_1 , v_2 and v_3 will be created to encode the values as binary combinations of 0 and 1. The phenotype value of 1 will be encoded as $v_1 = 1$, $v_2 = 0$ and $v_3 = 0$; the value of 2 will be encoded as $v_1 = 0$, $v_2 = 1$ and $v_3 = 0$, and the value of 3 as $v_1 = 0$, $v_2 = 0$ and $v_3 = 1$.

Not all transformations can be applied to all phenotypes and easyGWAS controls these options based on the type of data of the phenotype that is being analyzed. **Supplemental Table 5** provides additional details of the methods and of their constraints to different data types.

It is also important to note that there is no principled way to determine what the best transformation for a given dataset is. It has been shown, in the case of linear mixed models, that selecting the wrong transformation for a phenotype can result in significant biases in the heritability estimates (Fusi et al., 2014). To mitigate this problem, easyGWAS allows the user to quickly visualize the distribution of the phenotypes after a transformation method has been selected. Additionally, a Shapiro-Wilk test is performed to test for normality of the transformed data (Shapiro and Wilk, 1965). Nevertheless, the user is cautioned that finding the right transformation method for a genome-wide association study is often times challenging.

² See Methods and Materials, subsection Genome-Wide Association Mapping Methods.

Supplemental Text 9: Meta-Analysis

As mentioned in the main manuscript, independent GWAS that focus on a particular trait tend to report associated genetic variants that have modest effects. The goal of a meta-analysis of GWAS is then to pool results from different studies, thus increasing sample size and power. This has made the meta-analysis a successful tool to discover genetic loci associated to common diseases and phenotypes (Evangelou and Ioannidis, 2013). This success comes at a price, and this section addresses what the main assumptions of a meta-analysis are and how should the user take these assumptions into consideration when conducting a meta-analysis in easyGWAS.

Three of the most common meta-analysis models are implemented in easyGWAS. These are a) the P-value model, b) the fixed-effects model and c) the random-effects model.

P-value model: is based on the combination of p-values obtained from different studies. easyGWAS implements three different methods to combine p-values, namely Fisher's method, Stouffer's Z-score, and the weighted version of Stouffer's Z-score. The general assumption in the p-value model is that the null hypothesis is the absence of true association in the different studies whereas the alternative hypothesis is that there is an association in at least one of them (Evangelou and Ioannidis, 2013). One major disadvantage of the p-value model is that the direction of effects is disregarded or it is simply unknown (i.e. in the case when only p-values are available). The two models discussed below overcome this limitation.

Fixed-effects model: assumes that a given genetic variant has the same effect across all studies. Under this model, any variation in the results of the individual studies is assumed to arise from sampling artifacts. This is the most popular approach when combining GWAS but if its core assumption is violated (i.e. the genetic effects may differ between studies because the populations are very heterogeneous) then there is a risk that the p-values will be inflated.

Random-effects model: assumes the individuals in each study display different magnitudes of genetic effect. When combining studies, the goal is then to estimate the average effect across all populations. Although one may feel inclined to use these model when unsure about the type of effects it is worth noting that if a random-effects model is used when the genetic variants have true fixed-effects, the estimation of p-values will be rather conservative and result in loss of power.

In summary, when the information about the direction and type of effects is available, combining effect sizes is more powerful than combining p-values or Z-scores (Borenstein et al., 2010, 2011). Nevertheless, as discussed before, using the wrong assumptions about the direction of effects will have an impact on the final calibration of the p-values and on their final interpretation. The fixed- and random-effects model require that all GWAS are performed on a distinct set of samples and that the computations are standardized across all studies (e.g. that the measurements had the same scale and had been transformed with the same methods).

As a final remark, currently easyGWAS does not check if there is an overlap of individuals when separate GWAS are combined. The user is cautioned to determine the provenance of all the individuals in a meta-analysis to guarantee there is no overlap as this may result in an inflation of the type I error due to spurious associations (Lin and Sullivan, 2009; Zaykin and Kozbur, 2010).

Supplemental Text 10: Correction methods for multiple hypothesis testing

Performing a form of multiple testing correction is unavoidable when testing thousands of potential associations in GWAS. The correction is done to avoid an abundance of false positive findings. easyGWAS offers four different approaches to multiple testing correction.

Approach 1 is **Bonferroni correction** (Bonferroni, 1936), which controls the family-wise error rate of making at least one false positive finding. Its approach to guarantee this family-wise error rate is to require a stricter significance threshold for each individual association test. Rather than deeming an association significant at level α (typically 0.05 or 0.01), the Bonferroni correction only deems a finding significant at level $\alpha^* = \alpha/n$ where n is the number of association tests performed.

Approaches 2-4 control the false discovery rate rather than the family-wise error rate. Procedures like the Bonferroni correction that control the family-wise error rate tend to be overly conservative because they focus on avoiding any false positive findings. This may lead to a loss of statistical power, that is, true associations may be missed. That is why the less conservative approach of controlling the false discovery rate instead – the expected number of false associations among all associations reported to be significant – gained popularity in many high-dimensional settings. In essence, it has an increased statistical power, i.e. its ability to find true associations. Approach 2 is the original approach to false discovery rate control by **Benjamini and Hochberg** (1995).

Approach 3 is a modified version of false discovery rate control by **Benjamini and Yekutieli** (2001) which takes dependence between tests into account. The original work on false discovery rate by Benjamini and Hochberg makes the assumption that all tests are independent. In GWAS this assumption is often violated, as neighboring genetic markers are highly correlated due to linkage disequilibrium and their corresponding association tests are highly dependent.

Approach 4 controls the false discovery rate using q-values (**Storey and Tibshirani**, 2003). In GWAS, the q-value for a particular SNP is the expected proportion of false positives among all significant findings, when calling this SNP significant. While the p-values of all tests are used in Approach 2 and 3 to define a global p-value threshold that allows to control the false discovery rate, the q-value of each SNP allows to make a statement about the corresponding false discovery rate.

In conclusion, easyGWAS offers a wide range of methods to correct for multiple hypothesis testing. Choosing the appropriate method for a given analysis is subjective and highly dependent on how will the results be followed up. As an example, consider a genome-wide association study in a population of plants where the final goal of identifying SNPs associated to the phenotype is to conduct follow-up validation experiments. If we use a Bonferroni correction with $\alpha = 0.01$, it will be unlikely that our validation will fail on any SNP deemed statistically significant. If, on the other hand, we are willing to accept a percentage of our validation experiments to fail, any of the Approaches 2-4 based on false discovery rate will be more appropriate.

Approach 1, although very basic, is the most common correction method in statistical genetics. On the other hand, Approach 4 is a gold standard in other branches of genomics. As a final recommendation, we suggest the novice user to apply the **Bonferroni correction** in their initial analyses, leaving **Storey and Tibshirani's** q-values for more seasoned or experienced users.

Supplemental Text 11: REST Interface

We provide a Representational State Transfer (REST) web service that allows the user to programmatically query and get results from easyGWAS. The REST interface is language-independent and enables a user to write a script (say, in Python or R) that queries easyGWAS through a web service and returns the results to the invoking program where further post-processing or analyses can take place. Of course, easyGWAS allows the user to download different types of data, including full results of a genome-wide association study computed in the platform. The user can then (offline) process the downloaded files. The REST interface is an alternative medium through which the user can access information in a programmatic way without the need to fully download entire datasets.

The implementation of the REST web service in easyGWAS follows a trend of providing flexible access to genomic data in large biological databases like the Ensembl Genome Browser (Yates et al., 2015) or WormBase ParaSite (Howe et al., 2016).

A full description of all available REST endpoints can be found in the easyGWAS online FAQ (https://easygwas.ethz.ch/faq/#section_7). In the following we will give some examples of how to use the interface:

Example 1: Get a list of all publicly available species

- 1.1 From a browser. The standard format in which the results are returned is JSON. Alternatively, the user can also download the results as a comma-separated file by pasting the following URL:
`https://easygwas.ethz.ch/rest/species/public/list.csv`

1.2 With the curl command

```
#!/bin/sh
host="https://easygwas.ethz.ch"
query="rest/species/public/list/"
curl ${host}/${query}
```

1.3 In Python 2

```
import requests, sys

# Server name and query
host = "https://easygwas.ethz.ch"
query = "/rest/species/public/list"

r = requests.get(host+query, headers={"Content-Type" : "application/json"})

# Determine if GET access was successful
if not r.ok:
    r.raise_for_status()
    sys.exit()

# Get the results
decoded = r.json()
print decoded
```

Example 2: Get all private phenotypes associated to the public species *Arabidopsis thaliana* (species_id = 1) for user john_doe

```
#!/bin/sh
host="https://easygwas.ethz.ch"
species_id=1
query="rest/species/${species_id}/phenotype/private/list/"
curl ${host}/${query} -u john_doe
```

Supplemental References

Bonferroni, C.E. (1936). *Teoria statistica delle classi e calcolo delle probabilita*. Libreria internazionale Seeber.

Box, G.E.P., Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*. 26 (2): 211-252.

Bush, W.S., and Moore, J.H. (2012). Chapter 11: Genome-Wide Association Studies. *PLoS Comput Biol* 8(12): e1002822.

- Fusi, N., Lippert, C., Lawrence, N.D., and Stegle, O.** (2014). Warped linear mixed models for the genetic analysis of transformed phenotypes. *Nature Communications* 5, 4890.
- Gretton, A., Bousquet, O., Smola, A., & Schölkopf, B.** (2005, October). Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory* (pp. 63-77). Springer Berlin Heidelberg.
- Howe, K.L., Bolt, B.J., Cain, S., Chan, J., Chen, W.J., Davis, P., Done, J., Down, T., Gao, S., Grove, C., Harris, T.W., Kishore, R., Lee, R., Lomax, J., Li, Y., Muller, H.-M., Nakamura, C., Nuin, P., Paulini, M., Raciti, D., Schindelman, G., Stanley, E., Tuli, M.A., Van Auken, K., Wang, D., Wang, X., Williams, G., Wright, A., Yook, K., Berriman, M., Kersey, P., Schedl, T., Stein, L., and Sternberg, P.W.** (2016). WormBase 2016: expanding to enable helminth genomic research. *Nucleic Acids Research*, 44(D1) D774-D780
- Jin, J., Liu, J., Wang, H., Wong, L., and Chua, NH.** (2013). PLncDB: Plant Long noncoding RNA Database. *Bioinformatics*, 29(8):1068-71.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., et al.** (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* 81, no. 3:559-75.
- Rothstein, H.R., Sutton, A.J. and Borenstein, M.** (2005) Publication Bias in Meta-Analysis, in *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments* (eds H. R. Rothstein, A. J. Sutton and M. Borenstein), John Wiley & Sons, Ltd, Chichester, UK.
- Seren, Ü., Grimm, D., Fitz, J., Weigel, D., Nordborg, M., Borgwardt, K., Korte, A.** (2016). AraPheno: a public database for Arabidopsis thaliana phenotypes. *Nucleic Acids Research*, gkw986.
- Shapiro, S.S., and Wilk, M.B.** (1965). An Analysis of Variance Test for Normality (complete Samples). *Biometrika* 52, no. 3-4:591-611.
- Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A., and Lewis, S.** (2002). The generic genome browser: a building block for a model organism system database. *Genome Res.* (10):1599-610.
- Yates, A., Beal, K., Keenan, S., McLaren, W., Pignatelli, M., Ritchie, G.R.S., Ruffier, M., Taylor, K., Vullo, A., and Flicek, P.** (2015). The Ensembl REST API: Ensembl Data for Any Language. *Bioinformatics*, 31 (1): 143-145
- Zeggini, E. and Ioannidis, J.P.** (2009). Meta-analysis in genome-wide association studies. *Pharmacogenomics* 10, 191-201.