



**HAL**  
open science

## Full epistatic interaction maps retrieve part of missing heritability and improve phenotypic prediction

Clément Carré, Jean Baptiste Carluer, Christian Chaux, Nicolas Roche,  
André Mas, Gabriel Krouk

### ► To cite this version:

Clément Carré, Jean Baptiste Carluer, Christian Chaux, Nicolas Roche, André Mas, et al.. Full epistatic interaction maps retrieve part of missing heritability and improve phenotypic prediction. 2022. hal-03749709

**HAL Id: hal-03749709**

**<https://hal.science/hal-03749709>**

Preprint submitted on 11 Aug 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# Full epistatic interaction maps retrieve part of missing heritability and improve phenotypic prediction.

## Authors:

Clément Carré<sup>1,2,3\*</sup> ✉, Jean Baptiste Carluer<sup>2,3\*</sup>, Christian Chaux<sup>1</sup>, Nicolas Roche<sup>1</sup>, André Mas<sup>1,2</sup> ✉, Gabriel Krouk<sup>1,3</sup> ✉

## Affiliation:

\*equal-contribution

✉ Correspondence:

[clement.carre@bionomeex.com](mailto:clement.carre@bionomeex.com),

[andre.mas@umontpellier.fr](mailto:andre.mas@umontpellier.fr),

[gkrouk@gmail.com](mailto:gkrouk@gmail.com)

## Contacts:

- 1 BionomeeX, Montpellier, France
- 2 IMAG, Univ. Montpellier, CNRS, France
- 3 IPSiM, Univ. Montpellier, CNRS, INRAE, Montpellier, France

## Abstract

The first Genome Wide Association Studies (GWAS) shed light on the concept of missing heritability. It constitutes a mystery with transcending consequences from plant to human genetics. This mystery lies in the fact that a large proportion of phenotypes are not explained by unique or simple genomic modifications. One has to invoke genetic interactions among different loci, also known as epistasis, to partly account for it. However, current GWAS statistical models are moderately scalable, very sensitive to False Discovery Rate (FDR) corrections and, even combined with High Performance Computing (HPC), they can take years to evaluate for a full combinatorial epistatic space for a single phenotype. Here we propose a modeling approach, named Next-Gen GWAS (NGG) that evaluates, within hours, >60 billions of single nucleotide polymorphism (SNP) combinatorial first-order interactions, on a reasonable computer power. We first benchmark NGG on state of the art GWAS model results, and applied this to *Arabidopsis thaliana* providing 2D epistatic maps at gene resolution. We demonstrate on several phenotypes that a large proportion of the missing heritability can i) be retrieved with this modeling approach, ii) indeed lies in epistatic interactions and iii) can be used to improve phenotype prediction.

During the past decade, genome-wide association studies (GWAS) have allowed the discovery of many genetic variants associated with human (1–3), plant (4), animal (5) phenotypic traits. GWAS success is thus a reality and many discoveries made with this technique led to disruptive insights in biology, impacting basic knowledge as well as translational approaches to agronomy and medicine (6). However, when we observe on the one hand the striking resemblance of human twins, and on the other hand the amount of variation explained by GWAS signals, we are inclined to admit that “mono-dimensional GWAS” (study of genetic variation effect taken one at a time) is somehow limited. The missing heritability (7, 8) can at least partially be explained by variant interactions, ie: epistasis. Approaching epistasis is a difficult problem that lies in the fact that current mathematical models linking genetic variations to phenotypes are extremely sensitive to up-scaling (in particular to the number of individuals in the study) and to False Discovery Rate corrections (9).

To attempt to make full epistatic maps a reality, we decided to use a radically different mathematical formalism combined with solving systems meant to take advantage of Graphic Processing Units (GPUs) being increasingly popular thanks to the rise of gaming and deep learning (10). For this, we established the NGG model that states and define heritability in this framework as done before by Zuk et al. (11) and others: We define  $X$  the matrix with  $n$  rows and  $p$  columns containing the genetic information. Each column displays the coded genetic variants (SNP) for the  $n$  individuals. We also define  $Y$  a vector containing the phenotype. The broad-sense heritability  $H$  may be defined via the following nonparametric “random signal plus noise” model :  $Y = f(X) + \varepsilon$  (NP) where the function  $f$  is unknown and general and  $\varepsilon$  is a random noise, independent from  $X$  that collects all other effects (other than genetic) on the phenotype  $Y$ , such as environmental effects for instance. Thus, the broad-sense heritability is expressed as  $H = \text{var}(f(X))/\text{var}(\varepsilon)$ . The narrow-sense heritability  $h$  also sometimes named additive heritability accounts for part of the variance explained by genetics in the linear model  $Y = X\theta + \varepsilon$  (L). The definition is  $h = \text{var}(X\theta)/\text{var}(\varepsilon)$ . We note that, of course,  $\text{model}(L) \subset \text{model}(NP)$ . Notice for further use that since the slope parameter  $\theta$  is unknown, the narrow sense heritability cannot be computed but only estimated (for example, via a plug-in estimator  $\hat{h} = \text{var}(X\hat{\theta})/\widehat{\text{var}}(\varepsilon)$ ). At last when the estimation method is Ordinary Least Squares (OLS) or some of its (regularized/penalized) variants, the definition above matches the classical  $R^2$  and adjusted  $R^2$ . Below the adjusted  $R^2$  is preferred for reasons related to both the dimensionality of the data (usually  $p$  is much larger than  $n$ ) and the well-known inflation of  $R^2$ .

We further consider two models:

$$\text{Model 1: } Y = X\theta_1 + \varepsilon$$

$$\text{Model 2 : } Y = X\theta_1 + Z\theta_2 + \varepsilon$$

Where  $Z = X \star X$  is the partial face-splitting (or transposed Khatri Rao product) of matrices (12). For self-containedness notice that when :

0

$$X = \begin{bmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{bmatrix} \quad \text{then} \quad X \star X = \begin{bmatrix} a_1 a_2 & a_1 a_3 & a_2 a_3 \\ b_1 b_2 & b_1 b_3 & b_2 b_3 \end{bmatrix}$$

Matrix  $Z$  contains all the pairwise Kronecker products of columns of  $X$ , excluding the products of a column with itself. This matrix  $Z$  will be referred to as the matrix of interactions or shortly 2D-matrix. When  $X$  has  $p$  columns,  $Z$  has  $p(p-1)/2$  columns. The matrix  $Z$  captures all interactions between the SNP's. Although Model 2 remains linear, it is not additive anymore and bridges between Model L (or Model 1) and Model NP. Our resolution method employs acceleration techniques for regularized least square estimation in a sparse linear model (see Methods). These recent techniques from machine learning are coupled with a specific HPC architecture: GPU (See Supplementary Material). The outcome is a sparse estimate  $\hat{\theta}$  collecting the effects of each variant and each SNP interaction, instead of retrieving p-values as a regular GWAS does. As such, NGG can be seen as a sparse signal detection analysis and does not use multiple statistical testing which preclude the use of FDR correction. This classical correction is replaced here by a drastic procedure for variable selection.

To evaluate the performance of our model to retrieve epistatic signals, we first worked on simulated data. The simulation has been performed in two steps. First, we simulated  $X$  and  $Y$  (Fig.1A and 1B), second we simulated  $Y$  for real  $X$  (SNP matrix) retrieved from the 1001 genome project (13) (Fig.1C and 1D, Supplementary Material S1). These simulations are built to control narrow sense heritability ( $h^2$ ) of the trait (Fig.1). Using simulations, we show that the NGG formalism is able to capture simulated epistatic events for a wide range of model parametric values (see Supplementary Material S1). We found that Model 2 is quite resilient to noise but sensitive to the number of individuals used for the analysis (as discussed further, see remarks on Very High Dimension), as it radically improves for larger numbers of individuals (Supplementary Material S1).

We further benchmark our method on state-of-the-art available datasets and modeling approaches (4, 14). For this we first compared unidimensional (i.e. classical) GWAS results using the 107 Arabidopsis phenotypes studied in the landmark paper Atwell et al. [Ref. (4)]. We observed that major signals retrieved with EMMA algorithm (4, 14) are also retrieved by NGG (Fig. 2A and Supplementary Material S2 for the 107 phenotypes). For instance, EMMA and NGG methods both identify a major peak for the phenotype 88: *bacterial disease resistance* (Fig. 2A). This peak directly identifies the resistance gene RESISTANCE TO P. SYRINGAE PV MACULICOLA 1 (RPM1)(15). It is worth noting that for this particular phenotype, some signals emerge in NGG that are not detected by EMMA (Fig. 2) and that for certain phenotypes, NGG and EMMA converge towards a  $x^2$  relationship (Supplemental material S2). The opposite is also true although less frequent (see for the 107 phenotypes Supplementary Material S2). Furthermore, NGG clearly retrieves the effect of FLOWERING LOCUS C (FLC) a major gene in the control of flowering time (16, 17) in the top hits as compared to EMMA. Here we took this gene as an example for which NGG may be good at retrieving such important signals since it is intrinsically built to retrieve  $\hat{\theta}$ , taking into account the other SNP effects (Model 1 and Model 2). For this we compared the capacity of NGG and EMMA to detect signals in the vicinity of the FLC

locus (20 kb window). Interestingly Figure 2B shows that the NGG model indeed retrieves FLC as being the second strongest signal when EMMA reports it as the 30th signal (Fig. 2B).

We then compared the speed of our solution to the fastest GWAS to date, permGWAS (18). We evaluate that the computation time of one particular SNP effect or SNP combination effect (in Model 1 or 2 respectively) by our NGG algorithms takes around  $4.4 \cdot 10^{-8}$  seconds for 1000 samples or individuals. The GPU version of permGWAS, on a comparable (even more powerful) computer setup at ours, runs at a speed of  $2 \cdot 10^{-4}$  seconds per SNP. These four orders of magnitude differences bring the calculation of a full epistatic landscape by NGG below an hour for ~60 billions SNP or combination of SNPs making full epistatic maps possible.

Being confident that NGG has the potential to point to true epistatic effects (Fig. 1) and having in mind that the number of individuals greatly improves the detection capacity of our model (Sup. Info 1), we took advantage of the recent work by Campos et al. 2021 (19) which is, to our knowledge, one of the available phenotype dataset with the greater number of Arabidopsis ecotypes to date. In this work Campos et al. (2021) provide the elementary composition (18 different elements) for >1100 different Arabidopsis ecotypes having been fully sequenced by the *1001 genome project* (13). Figure 3 reports results of unidimensional NGG for Phosphorus content (noted P31) of Arabidopsis leaves, that can be displayed at the same time as i) *support for the model  $x$  |SNP effect|*, or as ii) *pure SNP effect ( $\theta$ )*. The latter provides a new kind of Manhattan plot with negative values that can be interpreted as the SNP having a negative effect on the phenotype as compared to the reference genome (here *Columbia-0* ecotype) (Fig.3). Also, the effect reported in this new kind of Manhattan plot is now expected to be directly proportional to the effect of the genetic variation as compared to Col-0 phenotype, helping to choose for the best variant or gene to study.

We further proceeded with the computation of full epistatic maps or 2D-NGG for phenotypes retrieved from the Campos et al. (2021) and Atwell et al (2010) datasets. We focused on these dataset as they present a relatively high number of ecotypes (>1000), and an important diversity of well known phenotypes respectively. To do this, we prefiltered SNPs having a particular Minor Allele Frequency (MAF) because the probability for the combination of SNPs (that we call MIAF for Minor Interaction Allele Frequency) to be of interest for epistatic measurements directly depends on the MAF as  $Z = X * X$  (above). For this we prefiltered 346094 SNPs, for Campos et al, and between 341067 and 371956 SNPs, for Atwell et al, having a MAF greater than 0.3. The full epistatic landscape is thus 59.890 billion interactions for Campos et al. (2021) and between 58.163 and 69.175 billion interactions for Atwell et al. (2010).

Nowadays, this quantity of data represents a challenge on its own to compute, store, and display the results as it relates to a “Very High-Dimensional” (VHD) framework (20). VHD is mathematically defined in terms of the size of the genotypic matrix  $X$  ( $n$  rows and  $p$  columns) and in terms of sparsity of the unknown parameter to be estimated or tested, here the number of “active” SNPs and interactions for a given phenotype:  $k$ . In this framework (20), we can evaluate the effects of VHD genotypic input matrices on the performance of several popular methodologies (for hypothesis testing, support estimation, prediction) and shows that when  $k \log(p/k)$  is large with respect to  $n$  then statistical estimation and testing errors inflates

dramatically. We believe that this is at least in part the reason for which full epistatic maps (2D-GWAS) were so far out of reach.

Following this line, in our study (Fig. 4)  $n=999$ , and  $p$  is around 60 billions. Following a reasonable estimate for sparsity  $k$ , granting satisfactory and reliable outputs, should be not more than 50. This is why, in our forthcoming study we mainly consider and analyze in a final stage around 30 significant SNPs interactions or composite components.

Figure 4 displays 2D-NGG results for i) Arabidopsis flowering times (Fig. 4A)(4), ii) Arabidopsis phosphorus (P31) leaf content (Fig. 4B) (19). Results are displayed as a square heatmap triangle for which  $\sim 60$  billion signals  $|\hat{\theta}|$  are provided. One 2D-NGG result dataset represents  $\sim 500$  Go of data. To navigate and mind this large dataset a visualization tool has been developed named *Lucio!* that can be understood as a “Google Earth” for full epistatic maps. In short, results are organized in layers as such the max intensity of a genomic region is reported on the higher layers. Here in Figure 4, layer 11 represents our maximum zoom-out condition. A zoom between layer 11 to layer 0 (layer for which a given pixel represents a direct SNP/SNP combination) corresponds to a 4.2 million times zoom. In other words a pixel in layer 11 reports the max intensity of 4.2 million real SNP/SNP interactions underlying layer 0. Observation of full epistatic maps as well as local signal informs on the genetic architecture underlying a given phenotype (Fig. 4).

In the case of the flowering trait (Fig. 4A), around 6 major epistatic signals emerge where 2 of them are close to the diagonal. The proximity of the diagonal refers to potential epistatic interactions of neighboring genes (although few Mb away). To display unambiguous epistasis we thus decided to report here the fourth stronger effect that lies very far from the diagonal. A zoom at the 2D-locus reveals the structure of a 2D-GWAS peak that appears bi-modal (ie: supported by at least 2 distant SNP combinations, 2 local bright spots in the epistatic map, Fig. 4B). This peak points to 2 loci predicted to be epistatic. The first loci is at position CHR4:6524710, and the second one is at CHR1:6243417. Using these coordinates, the matrix  $X$  and phenotype  $Y$  are parsed to plot the phenotypic distribution following the combination of SNPs (a sort of 2D-haplogroup). This is reported by the box plot in Figure 4C. Herein we observe that this epistatic effect involves 2 loci having a moderate effect individually as reported to the  $SNP_i$  and  $SNP_j$  boxplots (left and right panels). However, the combination of the simple effect cannot predict the effect of the combination since the positive effect of  $SNP_j$ , from 0 to 1 modality, seems enhanced by the  $SNP_i$  (0) modality and totally repressed by the  $SNP_i$  (1) modality. This clearly indicates the potential presence of an epistatic effect between these 2 loci.

We also report (Fig. 4D to 4F) the epistatic interactions in the control of plant leaf phosphorus content. This epistatic map reports around 8 strong epistatic signals. Here we zoom into 2 of them being the stronger ones with respect to their predicted value ( $|\hat{\theta}|$ ). The first one is relatively close to the diagonal although both epistatic SNPs lie in chromosome 1 but eleven Mbp away (Fig. 4F top panel). The second one concerns an epistatic effect predicted to involve SNPs on 2 different chromosomes namely CHR5 and CHR3 (Fig. 4F bottom panel). These 2 epistatic signals are built upon the effect of a strong combination of SNP effects as it appears impossible to predict the combinatorial output of these SNPs by solely analyzing the effect of the simple SNP modalities (compare box plots of  $SNP_i$  and  $SNP_j$  to box plot of  $SNP_i:SNP_j$ ). Here,



these effects can totally be missed by previous studies of epistasis that, to date, necessarily implied a selection of genetic variables (21).

We then decided to evaluate the quantity of heritability ( $h^2$ ) retrieved from 2D-GWAS as compared to regular 1D-GWAS (Fig. 5). The differential heritability between 1D and 1D+2D GWAS was estimated by Principal Component Regression (PCR) (22), carried out on a set of selected SNP and SNP-interactions (Fig 5A). The principle of PCR dates back to the late 50's (22). PCR combines Principal Component Analysis (PCA) on the input features of a model followed by linear regression (22). First a PCA of X provides a low number of principal components and a dimension reduction by selecting fewer components associated with the highest eigenvalues modulus of X. Regression is then performed on this reduced set of components (related to the VHD problem that we described above) that play the role of new synthetic inputs. PCA concentrates the information of the large matrix X or Z in a smaller matrix, removing collinearity as well because the components are, by definition, not correlated.

Here PCR is carried out i) on a set of p SNPs then, ii) on a set of the same SNPs as in (i) + q SNP/SNP-interactions (Fig. 5). The plots show the retrieved “missing heritability” (difference between the green line and blue [1D signal] or red lines [1D signal plus 2D\_random]) as measured by the adjusted  $R^2$  when the number of selected components increases (x axis Fig. 5B). For the vast majority (16 of the 18 phenotypes) a good proportion of heritability is retrieved in the 2D signals. Only, Cobalt or Selenium do not display a radical improvement in the explained variance. By applying this method we observe that information in the epistatic landscape indeed contains a good proportion of the missing heritability (Fig. 5B). For the Phosphorus content of Arabidopsis leaves for instance the heritability measures in the 1D GWAS ranges around 22%.  $h^2$  then increases to 33% when the information in the 2D-GWAS is considered.

Having at hand a view of the full epistatic maps can be seen as a new route towards at least two kinds of developments. The first one is obviously experimental validations. These are very labor intensive and may require a very long time to precisely dissect epistatic interactions. These are under investigation but we wished to release our results in the light of the second. The second one is phenotypic prediction. Indeed, one could see the NGG as a very strong variable selection process (Fig 5A) that may benefit precision medicine or several programs of agronomic selection for plants or animals. We thus further evaluate the role of NGG signals for phenotypic predictions through the use of a broad set of proper machine learning algorithms including Deep Neural Networks (DNN), Support Vector Machine (SVM), Gaussian Processes (GP), Gradient boosting (GB), Random Forest (RF), Linear regression, Lasso, Elastic Net. These techniques were used to predict the 18 phenotypes from the Campos et al (2021) work (described above). We also crossed these machine learning techniques with an increasing number of 1D or 2D signals/SNPs (50, 100, 500, 1000, 5000, 10000). To perform a proper control we repeated this *in silico* experiment but instead of providing the models with proper 2D signals, we randomly sampled epistatic signals (named 2D-random) to evaluate our capacity to predict plant mineral content. As classification problems are easier to solve and that the number of individuals is still a bit limited for regression approaches, we also used quantiles to rank phenotypes into 5 or 3 classes. By crossing all these parameters, we ended up with 1728 different models for 1D+2D signals (y axis of the plot Fig. 6A) and the same number of models

for 1D+2D\_random (x axis Fig.6A). Our capacity to predict phenotype is performed on 30% of the dataset (validation set) that were not used to fit or train the models. The quality of the models are evaluated through classical precision/recall curves and F scores. Figure 6 presents the F1 scores of best predicted classes (measure a precision/recall compromise) for 1D+2D random against 1D+2D models. All the models lying above the diagonal ( $x=y$ ) correspond to models for which predictive power is improved by epistatic signals (Fig. 6A).

We observe that 2D epistatic signals largely improve phenotypic classification (Fig. 6A) as 61% of the models are improved. Interestingly, models having a low F1 score and models having a high F1 score seem to benefit the most of the epistatic signals. We wish to highlight some particular points for which we observe a dramatic increase in our capacity to predict phenotypic classification (Fig. 6A). The red arrow Figure 6A points to a model for which 1D signal alone does not allow a good classification (value 0.33), when the same model with epistatic signal reaches a Fscore of 0.65. We also observe some phenotypes such as Na23 (sodium leaf content) for which most models and parametric values of the machine learning procedures globally benefit the epistatic signal showing that retrieved 2D signals are globally bringing new information (purple circle Fig. 6A). Figures 6B and 6C display an example of our capacity to predict phenotype classification for molybdenum leaf concentrations. This level of precision and recall opens avenues for plant selection procedures.

In conclusion, in this work, we provide a first dive into complete epistatic maps with enough SNP to reach a gene resolution and new tools to analyze this VHD problem. We apply it to the model plant *Arabidopsis thaliana* but our technique is fully generic and can easily reach other biological models. We demonstrate that, as hypothesized before (11), a substantial part of the missing heritability indeed lies in epistatic interactions (Fig. 5). We also show that this never observed 2D epistatic signal brings us a bit closer to the prediction of phenotypic values by machine learning procedures in plants, but we hope soon, in other biological models as well.

## METHODS

### Data

Arabidopsis dataset corresponds to data issued from the 1001 genome project (13) and kindly provided by Arthur Korte lab. It consists of a genotype matrix above mentioned as genotype or X matrix containing 9124892 SNPs and 1135 ecotypes. For NGG analysis MAF is controlled ( $0.3 < \text{MAF}$ ) resulting in a MAFed X' matrix containing 346094 SNPs for Campos et al. (2021) and between 341067 and 371956 SNPs, for Atwell et al, (2010).

### Simulations

The simulations (Fig. 1) are performed on R. Code can be found at (<https://github.com/CarluerJB> )

### Computer power

This work has been performed on a PowerEdge T640 DELL Server, RAM 377 Go, 4 NVIDIA Quadro RTX 6000 (24Go).



## Figure Legends

**Fig. 1. Next-Gen GWAS retrieves simulated epistatic interactions.** Var1 (x axis) and Var2 (y axis) are a series of 100 SNPs. The triangle corresponds to SNPs combinations when the diagonal contains simple SNPs effects. z axis reports the estimated  $\hat{\theta}$  of simple SNPs (diagonal) and combinations (rest of the triangle). **A and B)** Genotype and phenotype data are simulated using specific and modulable parameters (see text for details on the simulation). Random noise is added. NGG retrieves the 5 simulated signals (purple) including the pure epistatic effects (outside of the diagonal). We have found that NGG is quite resilient to noise in the data (on phenotypes Sup info 1) and the power of NGG increases quickly with the number of individuals (Sup info 2). **C and D)** Phenotype data only have been simulated while genotypes are from the *Arabidopsis* genome (SNPs are sampled from X matrix). Again epistatic interactions (purple points) are retrieved by NGG.

**Fig. 2. Next-Gen GWAS retrieves 1D-GWAS signal in Arabidopsis comparable to routinely used MML (EMMA (4, 14, 23)) and points to FLC locus for flowering phenotypes.** A) Data from Atwell et al., (2010) have been used to compare efficiency of our algorithm to standards of GWAS in Arabidopsis. NGG and EMMA algorithms largely retrieve similar signals. B) The phenotype 48 (days to flowering trait [8W]) NGG results are displayed. SNPS in the close vicinity of FLC locus (a major component of flowering in plants) is represented by black dots. The scatter plot presents the fact that NGG better detects FLC effect as compared to EMMA.

**Fig. 3. NGG provides direct estimation of SNP effect ( $\theta$ ) on the phenotype (Col-0 being the reference genome).** The upper plot presents the NGG signal combining support (effect or not) x absolute value of the estimated effect of the genetic variation. The lower plot reports the estimated effect of each SNP. Colored data points are emerging from the noise in a bootstrap procedure described in Supplementary Material 1).

**Fig. 4. 2D-NGG results provide an estimation of 61.2 billion SNP combination effect for:** A) Atwell et al, phenotype ID:31, days to flowering time FT10, B) Phosphorus content Campos et al. 2021 measured by ICP-MS. The results are presented as heatmaps and histograms to observe the epistatic interactions between SNPs.

**Fig. 5. Estimation of retrieved missing heritability.** A) Analysis scheme employed to estimate retrieved heritability and phenotypic predictions (in Fig. 6) from 1D signals (Blue Diagonal), and from 2D NGG signals orange triangle representing 61 billion interactions. B) Heritability ( $h^2$  seen as adjusted  $R^2$ ) is measured for an increasing number of PCA components, and for signal retrieved only from 1D-GWAS or 1D-GWAS (V data points) + 2D-NGG (W data points).

**Fig. 6. NGG retrieves genetic markers in epistatic signals improving machine learning procedures.** A) In this dot plot each dot corresponds to a given machine learning model (among: SVM, RF, DNN, Gaussian processes, LASSO, Elastic-Net Classifier) trying to predict a given phenotype (18 elemental concentration of Arabidopsis leaves represented with different colors) combined with different learning data inputs including a different number of classes (3 or 5) and different number of SNPs (50, 100, 500, 1000, 5000, 10 000) (Full dataset provide in Sup Material). The x axis reports max F1 score for the mode provided with SNPs simple 1D signals and randomly picked 2D epistatic SNP combinations (our control). The y axis reports 1D signals and 2D signals retrieved by NGG for the sample model and parameter combinations, respectively. We observe a clear improvement (above the  $y=x$  line) of >65 % of the models. B,C) Example of the good prediction of the Molybdenum (Mo98 phenotype) classified concentrations. B) confusion matrix, C) panel ROC curves for each class of Mo98 phenotype.

## REFERENCES AND NOTES

1. C. Fuchsberger, J. Flannick, T. M. Teslovich, A. Mahajan, V. Agarwala, K. J. Gaulton, C. Ma, P. Fontanillas, L. Moutsianas, D. J. McCarthy, M. A. Rivas, J. R. B. Perry, X. Sim, T. W. Blackwell, N. R. Robertson, N. W. Rayner, P. Cingolani, A. E. Locke, J. F. Tajes, H. M. Highland, J. Dupuis, P. S. Chines, C. M. Lindgren, C. Hartl, A. U. Jackson, H. Chen, J. R. Huyghe, M. van de Bunt, R. D. Pearson, A. Kumar, M. Müller-Nurasyid, N. Grarup, H. M. Stringham, E. R. Gamazon, J. Lee, Y. Chen, R. A. Scott, J. E. Below, P. Chen, J. Huang, M. J. Go, M. L. Stitzel, D. Pasko, S. C. J. Parker, T. V. Varga, T. Green, N. L. Beer, A. G. Day-Williams, T. Ferreira, T. Fingerlin, M. Horikoshi, C. Hu, I. Huh, M. K. Ikrum, B.-J. Kim, Y. Kim, Y. J. Kim, M.-S. Kwon, J. Lee, S. Lee, K.-H. Lin, T. J. Maxwell, Y. Nagai, X. Wang, R. P. Welch, J. Yoon, W. Zhang, N. Barzilai, B. F. Voight, B.-G. Han, C. P. Jenkinson, T. Kuulasmaa, J. Kuusisto, A. Manning, M. C. Y. Ng, N. D. Palmer, B. Balkau, A. Stančáková, H. E. Abboud, H. Boeing, V. Giedraitis, D. Prabhakaran, O. Gottesman, J. Scott, J. Carey, P. Kwan, G. Grant, J. D. Smith, B. M. Neale, S. Purcell, A. S. Butterworth, J. M. M. Howson, H. M. Lee, Y. Lu, S.-H. Kwak, W. Zhao, J. Danesh, V. K. L. Lam, K. S. Park, D. Saleheen, W. Y. So, C. H. T. Tam, U. Afzal, D. Aguilar, R. Arya, T. Aung, E. Chan, C. Navarro, C.-Y. Cheng, D. Palli, A. Correa, J. E. Curran, D. Rybin, V. S. Farooq, S. P. Fowler, B. I. Freedman, M. Griswold, D. E. Hale, P. J. Hicks, C.-C. Khor, S. Kumar, B. Lehne, D. Thuillier, W. Y. Lim, J. Liu, Y. T. van der Schouw, M. Loh, S. K. Musani, S. Puppala, W. R. Scott, L. Yengo, S.-T. Tan, H. A. Taylor Jr, F. Thameem, G. Wilson Sr, T. Y. Wong, P. R. Njølstad, J. C. Levy, M. Mangino, L. L. Bonnycastle, T. Schwarzmayr, J. Fadista, G. L. Surdulescu, C. Herder, C. J. Groves, T. Wieland, J. Bork-Jensen, I. Brandslund, C. Christensen, H. A. Koistinen, A. S. F. Doney, L. Kinnunen, T. Esko, A. J. Farmer, L. Hakaste, D. Hodgkiss, J. Kravic, V. Lyssenko, M. Hollensted, M. E. Jørgensen, T. Jørgensen, C. Ladenvall, J. M. Justesen, A. Käräjämäki, J. Kriebel, W. Rathmann, L. Lannfelt, T. Lauritzen, N. Narisu, A. Linneberg, O. Melander, L. Milani, M. Neville, M. Orho-Melander, L. Qi, Q. Qi, M. Roden, O. Rolandsson, A. Swift, A. H. Rosengren, K. Stirrups, A. R. Wood, E. Mihailov, C. Blancher, M. O. Carneiro, J. Maguire, R. Poplin, K. Shakir, T. Fennell, M. DePristo, M. H. de Angelis, P. Deloukas, A. P. Gjesing, G. Jun, P. Nilsson, J. Murphy, R. Onofrio, B. Thorand, T. Hansen, C. Meisinger, F. B. Hu, B. Isomaa, F. Karpe, L. Liang, A. Peters, C. Huth, S. P. O’Rahilly, C. N. A. Palmer, O. Pedersen, R. Rauramaa, J. Tuomilehto, V. Salomaa, R. M. Watanabe, A.-C. Syvänen, R. N. Bergman, D.

- Bharadwaj, E. P. Bottinger, Y. S. Cho, G. R. Chandak, J. C. N. Chan, K. S. Chia, M. J. Daly, S. B. Ebrahim, C. Langenberg, P. Elliott, K. A. Jablonski, D. M. Lehman, W. Jia, R. C. W. Ma, T. I. Pollin, M. Sandhu, N. Tandon, P. Froguel, I. Barroso, Y. Y. Teo, E. Zeggini, R. J. F. Loos, K. S. Small, J. S. Ried, R. A. DeFronzo, H. Grallert, B. Glaser, A. Metspalu, N. J. Wareham, M. Walker, E. Banks, C. Gieger, E. Ingelsson, H. K. Im, T. Illig, P. W. Franks, G. Buck, J. Trakalo, D. Buck, I. Prokopenko, R. Mägi, L. Lind, Y. Farjoun, K. R. Owen, A. L. Gloyn, K. Strauch, T. Tuomi, J. S. Kooner, J.-Y. Lee, T. Park, P. Donnelly, A. D. Morris, A. T. Hattersley, D. W. Bowden, F. S. Collins, G. Atzmon, J. C. Chambers, T. D. Spector, M. Laakso, T. M. Strom, G. I. Bell, J. Blangero, R. Duggirala, E. S. Tai, G. McVean, C. L. Hanis, J. G. Wilson, M. Seielstad, T. M. Frayling, J. B. Meigs, N. J. Cox, R. Sladek, E. S. Lander, S. Gabriel, N. P. Burt, K. L. Mohlke, T. Meitinger, L. Groop, G. Abecasis, J. C. Florez, L. J. Scott, A. P. Morris, H. M. Kang, M. Boehnke, D. Altshuler, M. I. McCarthy, The genetic architecture of type 2 diabetes. *Nature*. **536**, 41–47 (2016).
2. Wellcome Trust Case Control Consortium, Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. **447**, 661–678 (2007).
  3. J. MacArthur, E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. McMahon, A. Milano, J. Morales, Z. M. Pendlington, D. Welter, T. Burdett, L. Hindorf, P. Flicek, F. Cunningham, H. Parkinson, The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
  4. S. Atwell, Y. S. Huang, B. J. Vilhjálmsson, G. Willems, M. Horton, Y. Li, D. Meng, A. Platt, A. M. Tarone, T. T. Hu, R. Jiang, N. W. Muliyati, X. Zhang, M. A. Amer, I. Baxter, B. Brachi, J. Chory, C. Dean, M. Debieu, J. de Meaux, J. R. Ecker, N. Faure, J. M. Kniskern, J. D. G. Jones, T. Michael, A. Nemri, F. Roux, D. E. Salt, C. Tang, M. Todesco, M. B. Traw, D. Weigel, P. Marjoram, J. O. Borevitz, J. Bergelson, M. Nordborg, Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*. **465**, 627–631 (2010).
  5. D. Tian, P. Wang, B. Tang, X. Teng, C. Li, X. Liu, D. Zou, S. Song, Z. Zhang, GWAS Atlas: a curated resource of genome-wide variant-trait associations in plants and animals. *Nucleic Acids Res.* **48**, D927–D932 (2019).
  6. P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, J. Yang, 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
  7. C. Chatelain, G. Durand, V. Thuillier, F. Augé, Performance of epistasis detection methods in semi-simulated GWAS. *BMC Bioinformatics.* **19**, 231 (2018).
  8. T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorf, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, P. M. Visscher, Finding the missing heritability of complex diseases. *Nature*. **461**, 747–753 (2009).
  9. J. Hind, P. Lisboa, A. J. Hussain, D. Al-Jumeily, A Novel Approach to Detecting Epistasis using Random Sampling Regularisation. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **17**, 1535–1545 (2020).
  10. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature*. **521**, 436–444 (2015).

11. O. Zuk, E. Hechter, S. R. Sunyaev, E. S. Lander, The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 1193–1198 (2012).
12. V. I. Slyusar, A family of face products of matrices and its properties. *Cybern. Syst. Anal.* **35**, 379–384 (1999).
13. 1001 Genomes Consortium. Electronic address: magnus.nordborg@gmi.oeaw.ac.at, 1001 Genomes Consortium, 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*. *Cell.* **166**, 481–491 (2016).
14. A. Korte, B. J. Vilhjálmsón, V. Segura, A. Platt, Q. Long, M. Nordborg, A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat. Genet.* **44**, 1066–1071 (2012).
15. M. R. Grant, L. Godiard, E. Straube, T. Ashfield, J. Lewald, A. Sattler, R. W. Innes, J. L. Dangl, Structure of the *Arabidopsis* RPM1 gene enabling dual specificity disease resistance. *Science.* **269**, 843–846 (1995).
16. S. D. Michaels, R. M. Amasino, FLOWERING LOCUS C encodes a novel MADS domain protein that acts as a repressor of flowering. *Plant Cell* (1999) (available at <http://www.plantcell.org/content/11/5/949.short>).
17. C. C. Sheldon, J. E. Burn, P. P. Perez, J. Metzger, J. A. Edwards, W. J. Peacock, E. S. Dennis, The FLF MADS box gene: a repressor of flowering in *Arabidopsis* regulated by vernalization and methylation. *Plant Cell.* **11**, 445–458 (1999).
18. M. John, M. J. Ankenbrand, C. Artmann, J. A. Freudenthal, A. Korte, D. G. Grimm, Efficient Permutation-based Genome-wide Association Studies for Normal and Skewed Phenotypic Distributions. *bioRxiv* (2022), p. 2022.04.05.487185.
19. A. C. A. L. Campos, W. F. A. van Dijk, P. Ramakrishna, T. Giles, P. Korte, A. Douglas, P. Smith, D. E. Salt, 1,135 ionomes reveals the global pattern of leaf and seed mineral nutrient and trace element diversity in *Arabidopsis thaliana*. *Plant J.* (2021), doi:10.1111/tpj.15177.
20. N. Verzelen, Minimax risks for sparse regressions: Ultra-high dimensional phenomena. *EJSS* . **6**, 38–90 (2012).
21. C. Niel, C. Sinoquet, C. Dina, G. Rocheleau, A survey about methods dedicated to epistasis detection. *Front. Genet.* **6**, 285 (2015).
22. S. H. Park, Collinearity and Optimal Restrictions on Regression Parameters for Estimating Responses. *Technometrics.* **23**, 289–295 (1981).
23. V. Segura, B. J. Vilhjálmsón, A. Platt, A. Korte, Ü. Seren, Q. Long, M. Nordborg, An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* **44**, 825–830 (2012).

## ACKNOWLEDGMENTS

We wish to thank Dr Sandrine Ruffel, Dr Benoit Lacombe and Dr Eric Hosi for their valuable remarks on the manuscript.

**Funding:** Research on phenotype prediction is supported by the CNRS and by Montpellier University (Isite MUSE project AI3P) to G.K. and A.M. This work/project was publicly funded through ANR (the French National Research Agency) under the "Investissements d'avenir" programme with the reference ANR-16-IDEX-0006. The development of a first version of the computer algorithms was supported by SATT-AxLR and by Agropolis Fondation (SATT prematuration project) to G.K. A different version of the NGG resolution algorithm and the visualization algorithm named *Luciol* allowing data exploration has been developed by BionomeeX™.

**Author contributions:** Conceptualization: C.C., G.K. and A.M. Methodology: C.C., J.B.C. (performed prediction, machine learning and simulations), G.K., A.M., N.R., Ch.Ch, Writing – original draft: C.C, J.B.C, A.M, G.K.

### Competing interests:

C.C, G.K, A.M are co-founders of a CNRS and University of Montpellier spin-off company BionomeeX™ (BionomeeX.com) to provide a simple and rapid access to NGG technology and subsequent data analysis.

### Data and materials availability:

Phenotypic and genotypic data have been retrieved online. All code associated with heritability and data modeling can be retrieved at GitHub (<https://github.com/CarluerJB>). Raw data results of 2D-NGG for the phenotypes presented Figure 4 are available upon request to any of the corresponding authors.

## Supplementary Materials

### Supplemental Figure 1. Influence of the number of individuals on the power of epistatic signal detection.

Var1 (x axis) and Var2 (y axis) are a series of 100 SNPs. The triangle corresponds to SNPs combinations when the diagonal contains simple SNPs effects. z axis reports the estimated  $\theta$  of simple SNPs (diagonal) and combinations (rest of the triangle). Genotype and phenotype data are simulated using specific and modulable parameters (see Github for details and code). Random noise has been added. Simulated signals are in purple. Simulations control heritability ( $h^2$ ). The number of individuals ( $n$ ) varies over the figures (see details above panel). The number of individuals has a substantial effect on GWAS signal detection.

### Supplemental Figure 2. Comparison of 100 GWAS results using EMMA or NGG algorithms.

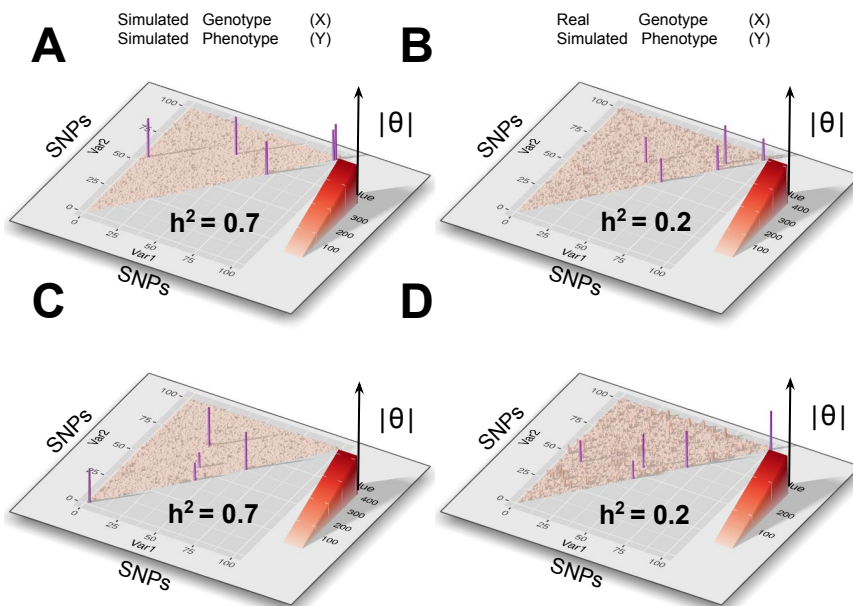
Atwell et al. (2010) dataset has been analyzed using both algorithms. Results are provided as

Manhattan plots and signals between EMMA and NGG are compared by plotting their respective signals against each other.

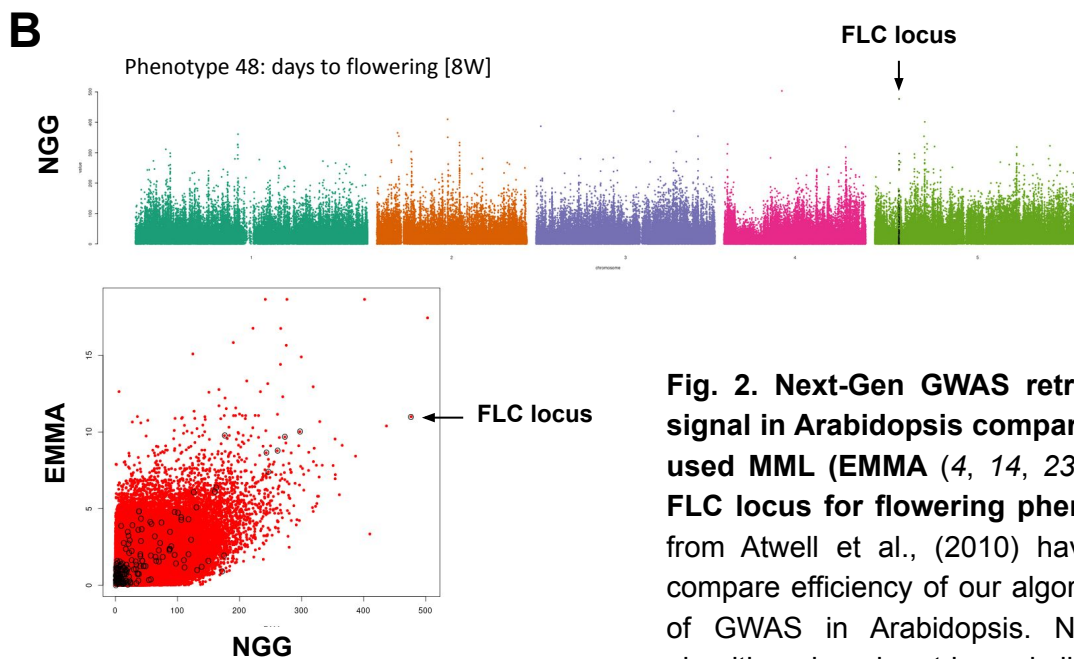
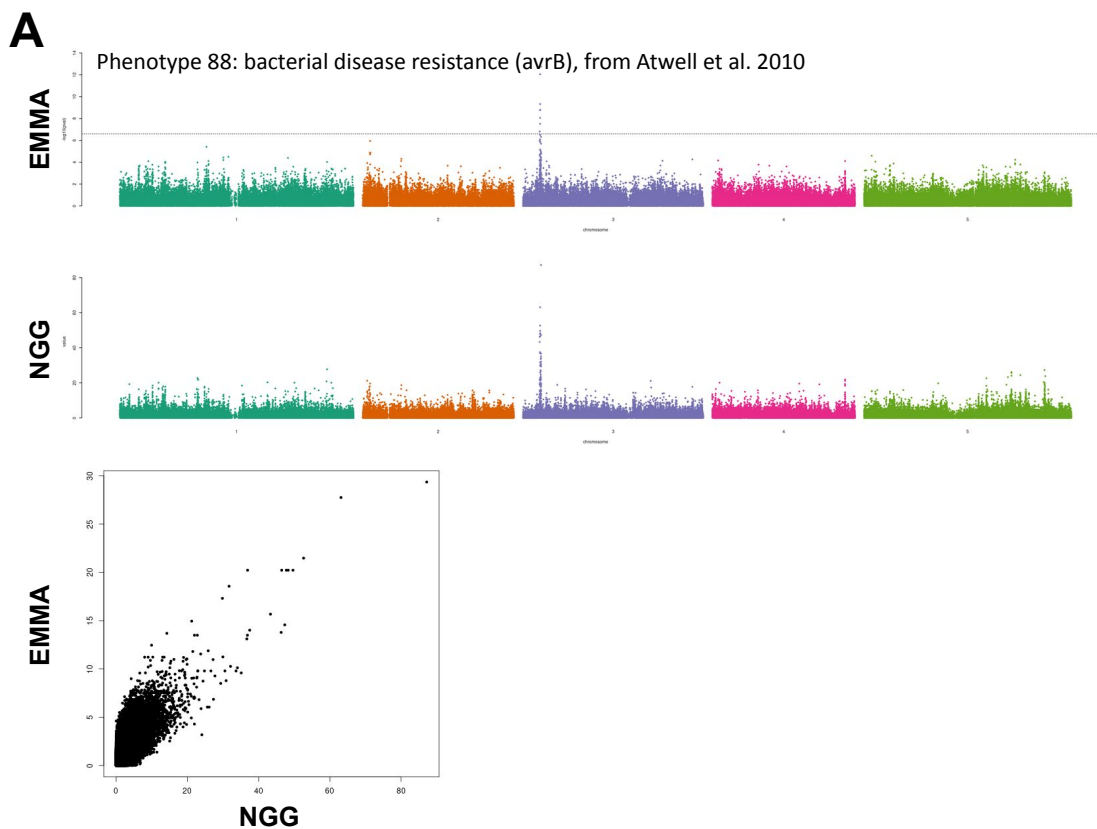
**Supplemental Table 1. Machine Learning Results.**

This table reports scores (precision, recall, Fscores) for the prediction of 18 Arabidopsis Phenotypes (mineral content), and a combination of different parameters of the ML procedures, phenotypic classes, Machine Learning, using 1D signals or 2D signals as input variables. These data led to the making of Figure 6A.

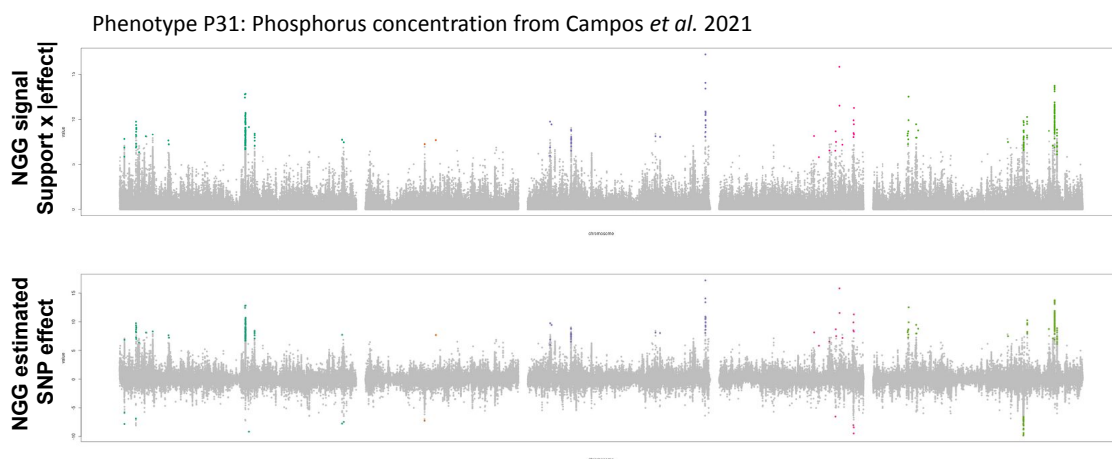




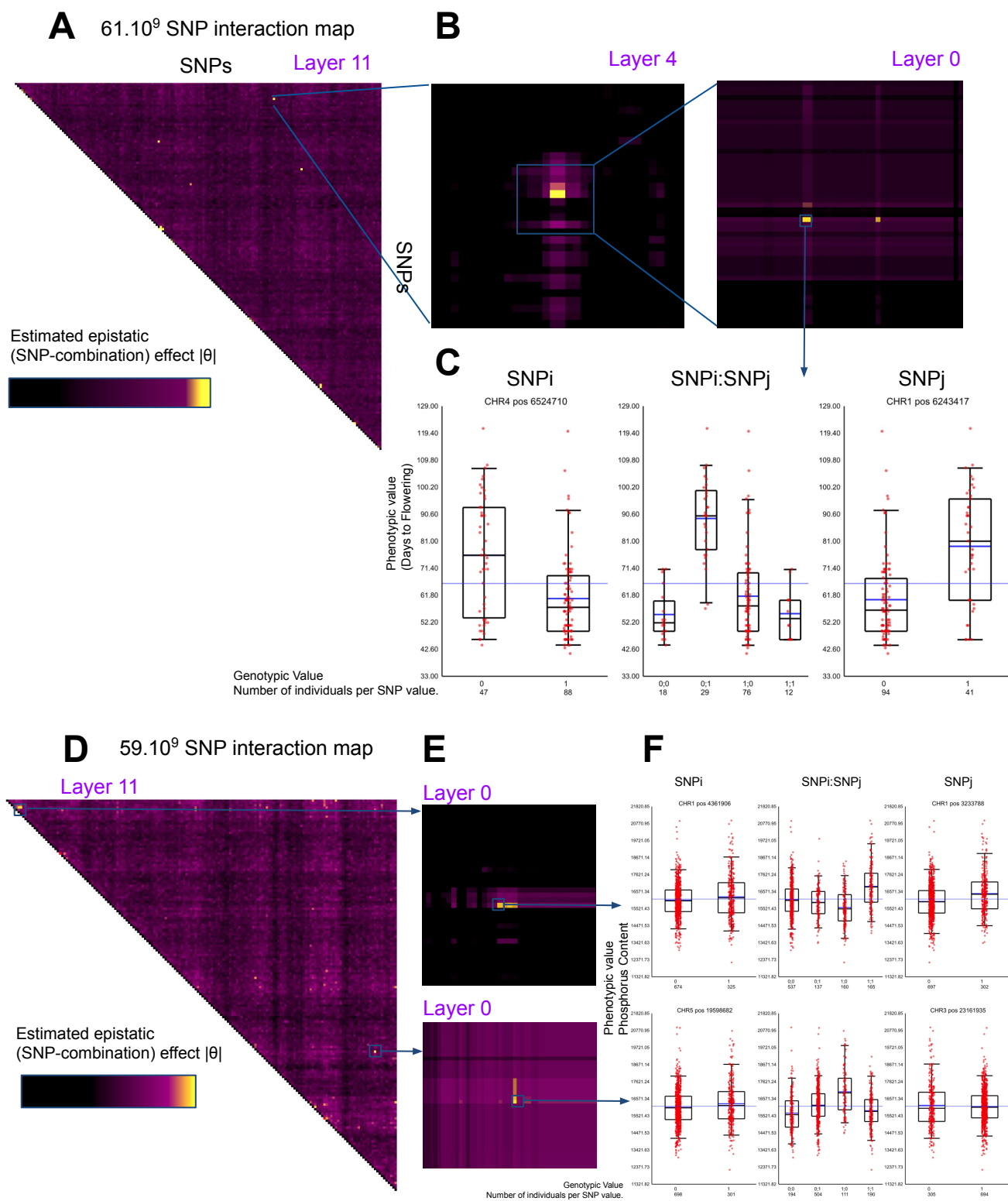
**Fig. 1. Next-Gen GWAS retrieves simulated epistatic interactions.** Var1 (x axis) and Var2 (y axis) are a series of 100 SNPs. The triangle corresponds to SNPs combinations when the diagonal contains simple SNPs effects. z axis reports the estimated  $|\theta|$  of simple SNPs (diagonal) and combinations (rest of the triangle). **A and B**) Genotype and phenotype data are simulated using specific and modulable parameters (see Sup. Material for details on the simulation). Random noise is added. NGG retrieves the 5 simulated signals (purple) including the pure epistatic effects (outside of the diagonal). We have found that NGG is quite resilient to noise in the data (on phenotypes Sup info 1) and the power of NGG increases quickly with the number of individuals (Sup info 2). **C and D**) Phenotype data only have been simulated while genotypes are from the *Arabidopsis* genome (SNPs are sampled from X matrix). Again epistatic interactions (purple points) are retrieved by NGG.



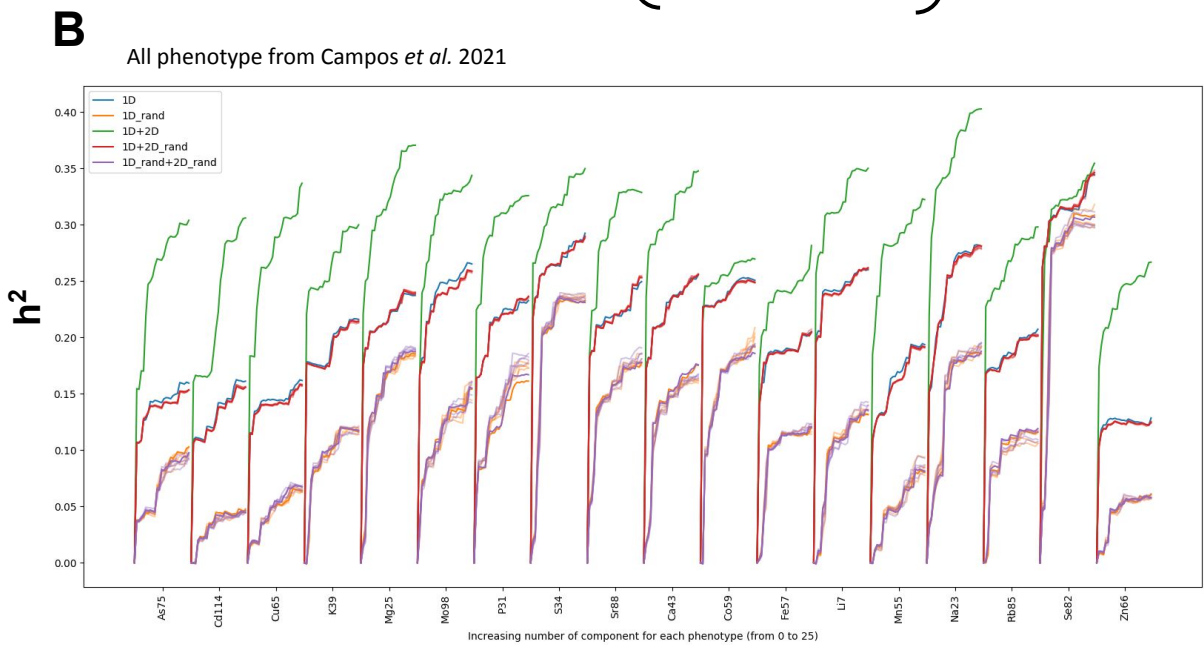
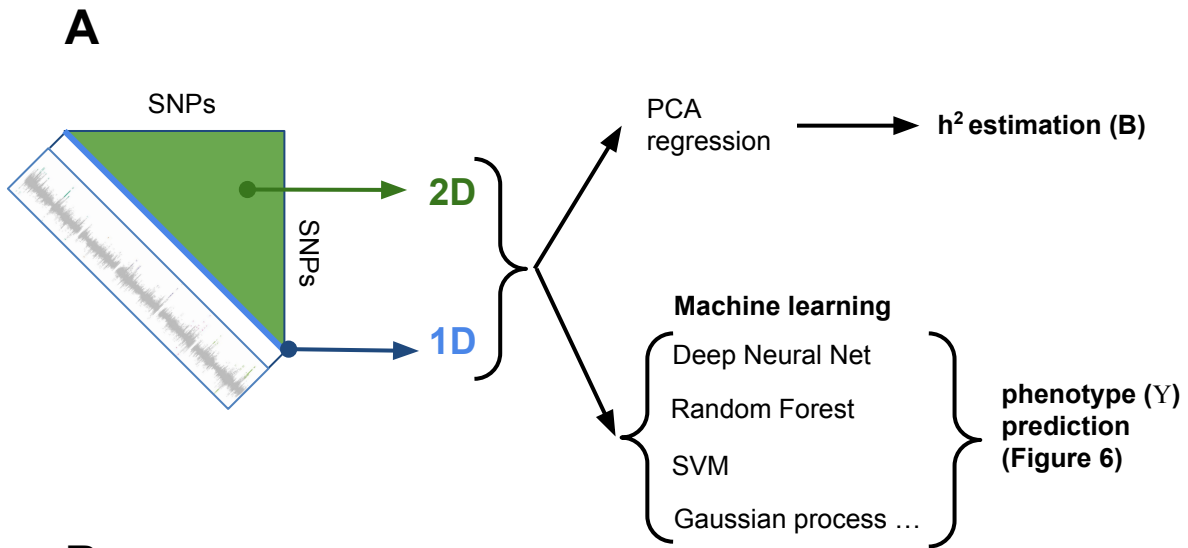
**Fig. 2. Next-Gen GWAS retrieves 1D-GWAS signal in Arabidopsis comparable to routinely used MML (EMMA (4, 14, 23)) and points to FLC locus for flowering phenotypes.** A) Data from Atwell et al., (2010) have been used to compare efficiency of our algorithm to standards of GWAS in Arabidopsis. NGG and EMMA algorithms largely retrieve similar signals. B) The phenotype 48 (days to flowering trait [8W]) NGG results are displayed. SNPs in the close vicinity of FLC locus (a major component of flowering in plants) is represented by black dots. The scatter plot presents the fact that NGG better detects FLC effect as compared to EMMA.



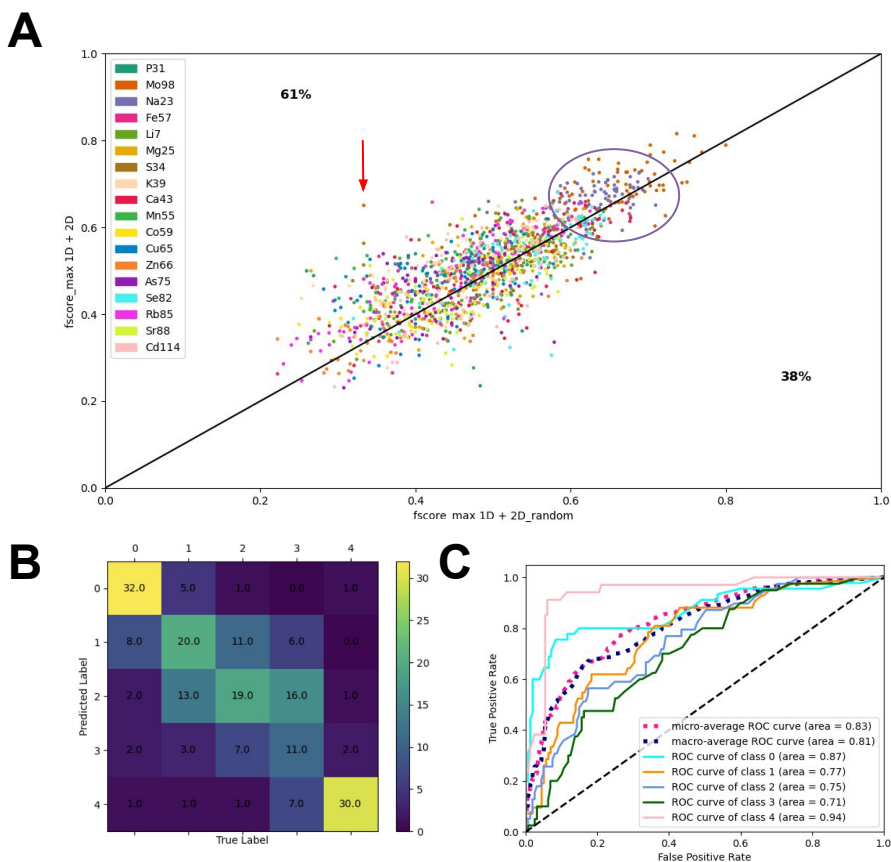
**Fig. 3. NGG provides direct estimation of SNP effect ( $\theta$ ) on the phenotype (Col-0 being the reference genome).** The upper plot presents the NGG signal combining support (effect or not) x absolute value of the estimated effect of the genetic variation. The lower plot reports the estimated effect of each SNP ( $\theta$ ). Colored data points are emerging from the noise in a bootstrap procedure comparable to permGWAS procedure (18).



**Fig. 4.** 2D-NGG results provide an estimation of >60 billion SNP combination effect for: A) Atwell et al, phenotype ID:31, days to flowering time FT10, B) Phosphorus content Campos et al. 2021 measured by ICP-MS. The results are presented as heatmaps and histograms to observe the epistatic interactions between SNPs.



**Fig. 5. Estimation of retrieved missing heritability.** A) Analysis scheme employed to estimate retrieved heritability and phenotypic predictions (in Fig. 6) from 1D signals (Blue Diagonal), and from 2D NGG signals orange triangle representing 61 billion interactions. B) Heritability ( $h^2$  seen as adjusted  $R^2$ ) is measured for an increasing number of PCA components, and for signal retrieved only from 1D-GWAS or 1D-GWAS ( $V$  data points) + 2D-NGG ( $W$  data points).



**Fig. 6. NGG retrieves genetic markers in epistatic signals improving machine learning procedures.** A) In this dot plot each dot corresponds to a given machine learning model (among: SVM, RF, DNN, Gaussian processes, LASSO, Elastic Classifier) trying to predict a given phenotype (18 elemental concentration of Arabidopsis leaves represented with different colors) combined with different learning data inputs including a different number of classes (3 or 5) and different number of SNPs (50, 100, 500, 1000, 5000, 10 000) (Full dataset provide in Sup Material 3). The x axis reports max F1 score for the mode provided with SNPs simple 1D signals and randomly picked 2D epistatic SNP combinations (our control). The y axis reports 1D signals and 2D signals retrieved by NGG for the sample model and parameter combinations, respectively. We observe a clear improvement (above the  $y=x$  line) of  $>61\%$  of the models. B,C) Example of the good prediction of the Molybdenum (Mo98 phenotype) classified concentrations. B) confusion matrix, C) panel ROC curves for each class of Mo98 phenotype.