

Manuscript Number:	GIGA-D-17-00225R2	
Full Title:	Predicting plant biomass accumulation from image-derived parameters	
Article Type:	Research	
Funding Information:	Robert Bosch Stiftung (32.5.8003.0116.0)	Dr. Christian Klukas
	Federal Agency for Agriculture and Food (15/12-13, 530-06.01-BiKo CHN)	Dr. Christian Klukas
	Bundesministerium für Bildung und Forschung (0315958A and 031A053B)	Dr. Christian Klukas
	European Plant Phenotyping Network (284443)	Dr. Christian Klukas
Abstract:	<p>Background: Image-based high-throughput phenotyping technologies have been rapidly developed in plant science recently and they provide a great potential to gain more valuable information than traditionally destructive methods. Predicting plant biomass is regarded as a key purpose for plant breeders and ecologist. However, it is a great challenge to find a predictive biomass model across experiments.</p> <p>Results: In the present study, we constructed four predictive models to examine the quantitative relationship between image-based features and plant biomass accumulation. Our methodology has been applied to three consecutive barley (<i>Hordeum vulgare</i>) experiments with control and stress treatments. The results proved that plant biomass can be accurately predicted from image-based parameters using a random forest model. The high prediction accuracy based on this model will contribute to relieve the phenotyping bottleneck in biomass measurement in breeding applications. The prediction performance is still relatively high cross experiments under similar conditions. The relative contribution of individual features for predicting biomass was further quantified, revealing new insights into the phenotypic determinants of plant biomass outcome. Furthermore, the methods could also be used to determine the most important image-based features related to plant biomass accumulation, which would be promising for subsequent genetic mapping to uncover the genetic basis of biomass.</p> <p>Conclusions: We have developed quantitative models to accurately predict plant biomass accumulation from image data. We anticipate that the analysis results will be useful to advance our views of the phenotypic determinants of plant biomass outcome, and the statistical methods can be broadly used for other plant species.</p>	
Corresponding Author:	Dijun Chen GERMANY	
Corresponding Author Secondary Information:		
Corresponding Author's Institution:		
Corresponding Author's Secondary Institution:		
First Author:	Dijun Chen	
First Author Secondary Information:		
Order of Authors:	Dijun Chen	
	Rongli Shi	
	Jean-Michel Pape	

	Kerstin Neumann
	Daniel Arend
	Andreas Graner
	Ming Chen
	Christian Klukas
Order of Authors Secondary Information:	
Response to Reviewers:	<p>We would like to thank both reviewers again for their further comments/suggestions to improve our manuscript. Our replies start with [Response].</p> <p>Reviewer #1: Chen et al, appear to have addressed each reviewer comment, below are some minor language changes for the revised sections. Minor changes (language changes) 1. Line 47: remove "some other traits" seems unnecessary 2. Line 64: change "they" to Buesmeyer et al. 2013, and change "make it a question" to "question" 3. Line 73 change besides to "Further" 4. Line 75 change to "due to a lack of datasets for assessment" [Response] We appreciate the reviewer for pointing out these. We corrected these in the revised manuscript.</p> <p>Reviewer #2: I thank the authors for the work done on the new manuscript and on Github, that address most of the concerns I raised in my first review. The pipeline published on GitHub now works nicely and allows to reproduces the different analyses. I only had to install manually two packages (earth and e1071). They could be easily added to the list of dependency in the R script to completely automatize the installation. [Response] We thank the reviewer for this suggestion. We fixed this issue in the 'run.R' script.</p> <p>The authors also clarify their analysis of the comparison of models, and the overstatement concerning the RF model has been corrected. I however still think that the abstract should be amended to better match the conclusions of the cross experiment test. The author acknowledged, in their response and in the text (line 226) that one cross experiment test leads to a loss of predictive accuracy. It seems also obvious, from Figure 5, and this should probably be added to the text, that this loss of accuracy is not linked to a greater random dispersion of the points, but to a systematic model bias. I agree with the authors that this may be due to some changes in the experimental conditions. My point is that these changes are not completely captured by the model, even with the inclusion of non structural traits. I therefore still think that there is some overstatement/ambiguity in the abstract, in particular in the sentence 'The high prediction accuracy based on this model, in particular the cross experiment performance, will contribute to relieve the phenotyping bottleneck in biomass measurement in breeding applications' . This may however be easily fixed. [Response] According to the reviewer's comment, we changed the sentence 'The high prediction accuracy based on this model, in particular the cross-experiment performance, will contribute to relieve the phenotyping bottleneck in biomass measurement in breeding applications.' to 'The high prediction accuracy based on this model will contribute to relieve the phenotyping bottleneck in biomass measurement in breeding applications. The prediction performance is still relatively high cross experiments under similar conditions.'</p>
Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics	Yes

<p>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends.</p> <p>Have you included all the information requested in your manuscript?</p>	
<p>Resources</p> <p>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?</p>	<p>Yes</p>
<p>Availability of data and materials</p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.</p> <p>Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?</p>	<p>Yes</p>

Predicting plant biomass accumulation from image-derived parameters

Dijun Chen^{1,§,*}, Rongli Shi¹, Jean-Michel Pape¹, Kerstin Neumann¹, Daniel Arend¹, Andreas Graner¹, Ming Chen², and Christian Klukas^{1,#}

¹*Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstrasse 3, 06466 Gatersleben, Germany.*

²*Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou 310058, China.*

[§]Current address: *Department for Plant Cell and Molecular Biology, Institute for Biology, Humboldt-Universität zu Berlin, 10115 Berlin, Germany.*

[#]Current address: Digitalization in Research & Development (ROM), BASF SE, 67056 Ludwigshafen am Rhein, Germany

*Correspondence should be addressed to D.C. (chendijun2012@gmail.com)

Abstract

Background:

Image-based high-throughput phenotyping technologies have been rapidly developed in plant science recently and they provide a great potential to gain more valuable information than traditionally destructive methods. Predicting plant biomass is regarded as a key purpose for plant breeders and ecologist. However, it is a great challenge to find a predictive biomass model across experiments.

Results:

In the present study, we constructed four predictive models to examine the quantitative relationship between image-based features and plant biomass accumulation. Our methodology has been applied to three consecutive barley (*Hordeum vulgare*) experiments with control and stress treatments. The results proved that plant biomass can be accurately predicted from image-based parameters using a random forest model. The high prediction accuracy based on this model will contribute to relieve the phenotyping bottleneck in biomass measurement in breeding applications. The prediction performance is still relatively high cross experiments under similar conditions. The relative contribution of individual features for predicting biomass was further quantified, revealing new insights into the phenotypic determinants of plant biomass outcome.

1
2
3
4 28 Furthermore, the methods could also be used to determine the most important image-based features related
5
6 29 to plant biomass accumulation, which would be promising for subsequent genetic mapping to uncover the
7
8 30 genetic basis of biomass.
9

10 31 **Conclusions:**

11 32 We have developed quantitative models to accurately predict plant biomass accumulation from image data.

12 33 We anticipate that the analysis results will be useful to advance our views of the phenotypic determinants of
13
14 34 plant biomass outcome, and the statistical methods can be broadly used for other plant species.

15 35 **Keywords:** Barley; High-throughput phenotyping; Phenomics; Biomass; Modeling.
16
17
18
19
20
21 36

22 37 **Introduction**

23
24
25 38 Biomass accumulation is an important indicator of crop final product and plant performance. It is thus
26
27 39 considered as a key trait in plant breeding, agriculture improvement and ecological applications. The
28
29 40 conventional approach of measuring plant biomass is very time consuming and labour intensive since plants
30
31 41 need to be harvested destructively to obtain the fresh or dry weight [1]. Moreover, the destructive method
32
33 42 makes multiple measurements of the same plant over time impossible. With the development of new
34
35 43 technology, digital image analysis has been used more broadly in many fields, as well as in plant research [2-
36
37 44 4]. It allows faster and more accurate plant phenotyping and has been proposed as an alternative way to infer
38
39 45 plant biomass [2, 3, 5].
40
41 46

42
43 47 In recent years, plant biomass has been subject to intensive investigation by using high-throughput
44
45 48 phenotyping (HTP) approaches in both controlled growth chambers [2-3, 6-11] and field environments [5,
46
47 49 12-17], demonstrating that the ability of imaging-based methods to infer plant biomass accumulation. For
48
49 50 example, significant genotypic and environmental effects on plant biomass in *Setaria* were revealed by the
50
51 51 Bellwether Phenotyping Platform under controlled-environmental condition [10]. Yang et al [11] showed that
52
53 52 predicted rice biomass (including shoot fresh and dry weight) based on image-derived morphological and
54
55 53 texture features provided a relatively more complete representation than manual measurements in dissecting
56
57 54 its genetic architecture. In this regard, optimized models plus image-derived features from HTP systems will
58
59 55 improve the power of dissecting genetic architecture of complex traits. Although there are some developed
60
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

56 models for predicting plant biomass, most of them have certain limitations. For example, Golzarian *et al.*
57 (2011) modelled the plant biomass (dry weight) in wheat (*Triticum aestivum* L.) as a linear function of
58 projected area, assuming plant density was constant. However, this method under-estimated dry weight of
59 salt stressed plants and over-estimated that of control plants. Even though the authors argued that the bias
60 was largely related to plant age and the model might be improved by including the factor of plant age [3], the
61 differences in plant density between stressed and control plants may be caused by different physiological
62 properties of plants rather than plant age. In another study, Busemeyer *et al.* (2013) developed a calibrated
63 biomass determination model for triticale (*x Triticosecale* Wittmack L.) under field conditions based on
64 multiple linear regression analysis of a diverse set of parameters, considering both, the volume of the plants
65 and their density. Indeed, this model largely improved the prediction accuracy of the calibration models based
66 on a single type of parameters and can precisely predict biomass accumulation across environments [15].
67 However, Buesmeyer *et al.* (2013) used very limited traits for the model and question whether it could be
68 applied broadly in other cases. As mentioned by Yang *et al.* (2014), noticeable improvement was achieved
69 by adding morphological features or texture features to the biomass-predicting model [11]. This suggests that
70 adding more information/traits could improve the predictive performance of models. Therefore, a more
71 effective and powerful model is needed to overcome these limitations and to allow better utilization of the
72 image-based plant features which are obtained from non-invasive phenotyping approaches.

73
74 Individual studies have recently shown that the prediction accuracy of plant biomass based on image-derived
75 features is relatively high even using the simplest linear regression models [3,10,18]. However, the
76 performance of nonlinear predictive models has not been well evaluated. Further, it is still challenging to
77 apply these models across experiments that are performed in different environmental conditions or with
78 different treatments due to a lack of datasets for assessment. In this study, we present a general framework
79 for investigating the relationships between plant biomass (referred to as shoot biomass hereafter) and image-
80 derived parameters. We applied a multitude of supervised and unsupervised statistical methods to investigate
81 different aspects of biomass determinants by a list of representative phenotypic traits in three consecutive
82 experiments in barley. The results showed that image-based features can accurately predict plant biomass
83 output and collectively reflect large proportions of the variation in biomass accumulation. We elucidated the

1
2
3
4 84 relative importance of different feature categories and of individual features in prediction of biomass
5
6 85 accumulation. The differences in the contribution of the image-based features for prediction of two types of
7
8 86 biomass measurements, fresh weight and dry weight were compared as well. Furthermore, our models were
9
10 87 tested for the possibility of predicting plant biomass in different experiments with different treatments.
11
12 88

15 89 **Results**

16 90 **Development of statistical models for modelling plant biomass accumulation using image-based** 17 18 91 **features**

19
20
21
22 92 In the previous studies [19,20], we have shown that a single phenotypic trait -- the three-dimensional digital
23
24 93 volume, which is a derived feature from projected side and top areas -- can be reasonably predictive to
25
26 94 estimate plant biomass accumulation. We expect that the predictive power could be improved when multiple
27
28 95 phenotypic traits are combined in a prediction model since plant biomass is determined not only by their
29
30 96 structural features but also by their density (physiological properties). To further investigate the relationship
31
32 97 between image-derived parameters and plant biomass accumulation, deep phenotyping data which contain
33
34 98 both structural (e.g., geometric traits) and physiological traits (e.g., plant moisture content as reflected by
35
36 99 near-infrared [NIR]-related traits) were analysed (**Fig. 1, A and B**). Pot weights of the plants were not
37
38 100 included for the analysis although they were weighed regularly. It might reflect the growth tendency of the
39
40 101 whole plants (shoots and roots) where herein we focused mainly on shoots.
41
42 102

43
44 103 Models were constructed to quantify the ability of imaging-based features to statistically predict the biomass
45
46 104 accumulation. The models were developed by using four widely used machine-learning methods (**Fig. 1C**):
47
48 105 multivariate linear regression (MLR), multivariate adaptive regression splines (MARS), random forest (RF)
49
50 106 and support vector regression (SVR), which have extensively been used in accurate prediction of gene
51
52 107 expression [21-25] and DNA methylation levels [26-29]. We combined the biomass measurements (fresh
53
54 108 weight [FW] and/or dry weight [DW]) with image-based features and then divided them into a training data
55
56 109 set and a test data set. A model was trained on the training data set and has then been applied to the test data
57
58 110 set to predict the plant biomass. The relationship between plant biomass accumulation and image-based
59
60
61
62
63
64
65

1
2
3
4 111 features was assessed based on the criterion of the Pearson correlation coefficient (r) between the predicted
5
6 112 values and the actual values, or the coefficient of determination (R^2 ; the percentage of variance of biomass
7
8 113 explained by the model; **Fig. 1D**).

9
10 114
11
12 115 Our methodology was applied to three consecutive experiments (**Fig. 2A**; **Supplemental Table S1** and **Data**
13
14 116 **S1**), which were designed to investigate vegetative biomass accumulation in response to two different
15
16 117 watering regimes under semi-controlled greenhouse conditions in a core set of barley cultivars by non-
17
18 118 invasive phenotyping [20, 30]. There were 312 plants with 18 genotypes for each experiment. Plants were
19
20 119 monitored using three types of sensors (visible, fluorescence [FLUO] and near-infrared [NIR]) in a
21
22 120 LemnaTec-Scanalyzer 3D imaging system. An extensive list of phenotypic traits ranging from geometric
23
24 121 (shape descriptors) to physiological properties (i.e., colour-, FLUO- and NIR-related traits) could be extracted
25
26 122 from the image data (**Supplemental Data S1**) using our image processing pipeline IAP [19]. A representative
27
28 123 list of traits for each plant in the last growth day were selected to test their ability to predict plant biomass.

29
30
31 124

32 125 **Coordinated patterns of plant-image-based profiles and their relation to plant biomass**

33
34 126 We extracted a list of representative and non-redundant phenotypic traits for each plant from image datasets
35
36 127 for each experiment (see **Materials and Methods**; **Fig. 1B**). In common for these experiments, overall thirty-
37
38 128 six high-quality traits which describe plant growth status in the last growth day were obtained. As a result,
39
40 129 each dataset was assigned a matrix whose elements were the signals of different features in different plants
41
42 130 (**Fig. 1C**). Principal component analysis (PCA; **Fig. 2B**) was applied to these datasets. We found that plants
43
44 131 from different experiments with different treatments showed clearly distinct patterns of phenotypic profiles.
45
46 132 For instance, stressed plants and control plants were separated using PCA by their first principal component
47
48 133 (PC1) and also by the top clusters obtained in HCA, while plants from different experiments were
49
50 134 distinguished by PC2 and PC3 in PCA or subordinate clusters in HCA. Accordingly, it could be observed that
51
52 135 biomass (e.g., FW) of plants from different experiments with different treatments was significantly different
53
54 136 (two-way ANOVA, p -value $< 2e-16$; **Fig. 2C**). The relationship was reflected by a dendrogram from cluster
55
56 137 analysis based on the means of FW over genotypes (**Fig. 2D**). Furthermore, the overall phenotypic patterns
57
58 138 of these plants were similar to their biomass output (**Fig. 2, B-D**), revealing that these image-based features

59
60
61
62
63
64
65

1
2
3
4 139 were potential factors reflecting the accumulation of plant biomass. We thus explored the relationship
5
6 140 between the signals of these image-based features and the level of plant biomass output. We calculated the
7
8 141 correlation coefficients for each dataset. The correlation patterns were consistent for different datasets and
9
10 142 more than half of the features revealed high correlation coefficients ($r > 0.5$; **Fig. 2E**). Interestingly, both
11
12 143 structural features (such as digital volume, projected area and the length of the projected plant area border)
13
14 144 and density-related features (such as NIR and FLUO intensities) were involved in the top ranked features.
15
16 145

18 146 **Relating image-based signals to plant biomass output**

20 147 The above analyses suggest that plant biomass can at least be partially inferred from image-based features.
21
22 148 To examine which model has the best performance and to select an appropriate model for biomass prediction,
23
24 149 we then applied our regression models (**Fig. 1C**) to predict plant biomass using image-based features. Our
25
26 150 analyses were focused on the first experiment (i.e., Exp 1), since the phenotypic traits of the corresponding
27
28 151 dataset have been intensively investigated in our previous study [20]. In this experiment, plant biomass was
29
30 152 quantified in two forms: FW and DW. We selected a collection of 45 image-derived parameters from this
31
32 153 dataset that were non-redundant and highly representative.
33
34 154

36 155 We next tried to predict FW and DW based on this set of image-derived features using four different
37
38 156 regression models (MLR, RF, SVR and MARS; **Fig. 3**). The models were respectively tested on control
39
40 157 plants, stressed plants and the whole set of plants (**Fig. 3, A and C**). The prediction accuracy of our models
41
42 158 (the correlation coefficients between the predicted biomass and the actual biomass) was firstly compared
43
44 159 with the ability of individual features to predict biomass. It was found that our models generally showed
45
46 160 better prediction power than the single digital volume-based prediction (**Fig. 3, B and D**), indicating that
47
48 161 additional features improved the predictive power. Then the performance of these models was compared and
49
50 162 evaluated. Overall, the performance of all the tested models showed roughly similar for the prediction of both
51
52 163 FW (**Fig. 3B**) and DW (**Fig. 3D**) under stressed conditions. The prediction accuracy of our models is still
53
54 164 comparable to the results from previous studies [3, 6, 18] based on MLR models, even though much more
55
56 165 features were considered in our study. The RF model slightly outperformed other models in predicting
57
58 166 biomass of control plants, accounting for the most variance ($R^2 = 0.85$ for FW and $R^2 = 0.62$ for DW; **Fig.**

1
2
3
4 167 **3, B and D**, left panels) and showed the best prediction accuracy (Pearson's correlation $r = 0.93$ for FW and
5
6 168 $r = 0.80$ for DW; **Fig. 3, B and D**, middle panels). Of note, RF is the only model showing better performance
7
8 169 than single digital volume-based prediction (**Fig. 3D**). In this study, we focused on the results from the RF
9
10 170 method in the rest of analysis, although results from different methods were highly consistent and led to the
11
12 171 same conclusions.

13
14 172

15 16 173 **Relative importance of different image-based features for predicting plant biomass**

17 174 As mentioned above, the image-based features could be classified broadly into four categories: plant structure
18
19 175 properties, colour-related features, NIR signals, and FLUO-based traits (**Fig. 1B**). The last three types of
20
21 176 features reflect plant physiological properties and can be considered as plant density-related traits and are
22
23 177 thus related to their fresh or dry matter content. For each individual feature or each type of features, we
24
25 178 constructed a degenerate model for biomass prediction using the corresponding feature(s) as the predictor(s).
26
27 179 We compared the capability of each individual or type of feature for predicting biomass accumulation in the
28
29 180 first experiment (i.e., experiment 1). Geometric features showed the most predictive power among the four
30
31 181 categories for prediction of both FW and DW, but were slightly less predictive than all features in a full model
32
33 182 (**Fig. 4, A and B**). Strikingly, the predictability of other types of features (such as colour-related and FLUO-
34
35 183 based traits) was substantial, indicating that these traits may act as unforeseen factors in biomass prediction.
36
37 184 In addition, the NIR-based features showed higher predictive capability for FW than for DW in control and
38
39 185 stressed plants, revealing NIR signals were import factors in determining FW accumulation.
40
41 186

42
43 187

44 187 Next, we investigated the relative importance (RI) of each feature for predicting biomass using a full model
45
46 188 in the whole set of plants (i.e., “control + stressed plants”; **Fig. 4, C and D**, upper panels). In a RF model, the
47
48 189 RI of a feature is calculated as the increase of prediction error (%IncMSE) when phenotypic data for this
49
50 190 feature is permuted [31], and thus indicates the contribution of the feature after considering its intercorrelation
51
52 191 in a model. We found that the top ten most important features in the full model for predicting FW and DW
53
54 192 included both structure and density-related traits. As expected, projected area (from side or top view) and
55
56 193 digital volume were the top ranked features, which have individually been considered as proxies of shoot
57
58 194 biomass in previous studies [3, 20, 18, 32-37]. However, several geometric and colour-related features that

1
2
3
4 195 are top ranked in the prediction have not been used in biomass predictions in previous analysis although they
5
6 196 are widely available among phenotyping platforms.

7
8 197

9
10 198 In principle, we would expect that highly important features in the full model would be related to a high
11
12 199 predictive power in a degenerate model. Surprisingly, there was no clear correlation observed between the
13
14 200 feature importance and their predictive power (**Fig. 4, C and D**). For example, several colour-related and
15
16 201 NIR-based features which were in the top ten list of the most important features revealed insubstantial
17
18 202 predictive power in individual models. This observation implies that the relation of the underlying biomass
19
20 203 determinants is extremely complex and not a linear combinations of the investigated features.

21
22 204

23
24 205 Furthermore, we compared the relative importance of each feature in predicting FW and DW (**Fig. 4E**).
25
26 206 Although a positive correlation ($r = 0.88$) between the feature importance for FW and DW could be observed,
27
28 207 several features showed large differences in their ability to interpret FW or DW, including “nir.intensity”
29
30 208 (derived from side view images), “compactness.01” (top), “hull.pc1” (top), “leaf.count” (side),
31
32 209 “hsv.h.average” (top) and “lab.a.mean” (top). For instance, NIR intensity and plant compactness (top view)
33
34 210 may be important for predicting FW but not for DW. We also performed the above analyses by using only
35
36 211 control (**Supplemental Fig. S1**) or stressed plants (**Supplemental Fig. S2**), respectively. We found that the
37
38 212 patterns of feature importance were distinct between these two groups of plants. For example, NIR intensity
39
40 213 was ranked as the top fifth feature for predicting FW for stressed plants but was not substantially important
41
42 214 for control plants. These findings suggest that there are differences in underlying plant biomass determinants
43
44 215 in these kinds of treatment situations that are also reflected by their image-based phenotypic traits.

45
46
47 216

48
49 217 **Image-based features are predictive of plant biomass across experiments with similar conditions or**
50
51 218 **treatments**

52
53 219 In order to explore whether our models were generalizable across different experiments, we applied our
54
55 220 models trained in one experiment to predict biomass (herein FW) in other experiments using a common set
56
57 221 of features. Examples of such cross-experiment predictions are shown in **Figure 5A**. We tested and illustrated
58
59 222 all possibilities for cross prediction using the whole set of plants in the corresponding experiment. In general,

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

223 the prediction accuracy within individual experiments remained high ($r > 0.97$ and $R^2 > 0.93$ for all three
224 experiments; **Fig. 5B**), revealing that our models were effectively predicting plant biomass based on image-
225 derived feature signals among different experiments. Moreover, the prediction accuracy for cross-experiment
226 prediction, especially between the first two experiments ($r > 0.97$ and $R^2 > 0.94$), was still relatively high,
227 implying that our models generally captured the relationships among the various image-based features.
228 However, the third experiment had relative weaker correlations with the other two experiments for predicting
229 biomass (with $r > 0.81$ and $R^2 > 0.65$; **Fig. 5A**). This might be mainly due to seasonal (temperature and
230 illumination) differences which caused different plants behaviours, namely lower biomass for both control
231 and stressed plants in experiment 3 [30]. This suggests that different plant growth conditions might cause
232 some variation for cross-experiment prediction.

233

234 At the same time, we tested cross predictability of our models using treatment-specific data in the experiments
235 (**Fig. 6**). Similar results were obtained as above using the whole dataset (**Fig. 5B**). The weak predictive power
236 for cross-prediction involving control plants from the third experiment was most clearly observable in the
237 low accuracy in the biomass prediction of this particular subset of plants. Generally, control and stressed
238 plants were found to have very weak predictive power when related to each other (**Fig. 6**), as also supported
239 by the distinct patterns of relative feature importance between these two plant groups (**Supplemental Figs.**
240 **S1 and S2**). For each experiment, the prediction accuracy was higher for stressed plants compared to control
241 plants. This might resulted from the imaging analysis process. Relatively small plants, stressed plants in this
242 case, would gain more clear images due to less overlapping or less area of out range. Therefore, image quality
243 would be an important variation source for our modelling and should be taking into consideration for any
244 application.

245

246 **Discussion**

247 Biomass is a complex but important trait in functional ecology and agronomy for studying plant growth, crop
248 productive potential and plant regeneration capabilities. Many different techniques, either destructive or non-
249 destructive, have been used to estimate biomass [1, 2-3, 5-17]. Compared with the traditional destructive

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

250 methods for measuring biomass, non-destructive imaging methods provide a faster, more accurate approach
251 for plant phenotyping. In recent years, more and more high-throughput plant phenotyping platforms have
252 been set up and applied worldwide. Accordingly, it becomes a current challenge to establish models utilizing
253 the big datasets gained from high-throughput imaging systems. Accurately predicting biomass from image
254 data requires efficient mathematical models as well as representative image-derived features.

255
256 In this study, we have presented a systematic analysis of relationships between plant biomass accumulation
257 and image-derived signals, to confirm the assumption that biomass can be accurately predicted from image-
258 based parameters. We built a random forest model of biomass accumulation using a comprehensive list of
259 representative image-based features. The comparison between a random forest model and alternative
260 regression models indicated that the RF model outperforms other models in terms of (1) better predictive
261 power – especially in comparison with the linear model, confirming the complex phenotypic architecture of
262 biomass, (2) better outperformance than a single-feature-prediction model – arguing the complex phenotypic
263 makeup of biomass, and (3) feasible biological interpretability – the ability to readily extract information
264 about the importance of each feature in prediction. The high prediction accuracy based on this model, in
265 particular the cross-experiment performance, is promising to relieve the phenotyping bottleneck in biomass
266 measurement in breeding applications. For example, based on an established small reference dataset which
267 is used to train a RF model, it is possible to predict biomass in several large plant populations within one
268 experiment or across several experiments using image data by taking advantage of high-throughput
269 phenotyping technologies. Alternatively, the model can be trained from a much larger reference panel of
270 plants that are grown in diverse environmental conditions which is then applied to a diverse set of experiments.
271 The first evidence for this notion is the observation that our model showed more predictive power in plants
272 with two treatments than with a single treatment (**Fig. 3, B and D**). Indeed, when applying our model to the
273 combined dataset from all the three experiments, we found the prediction accuracy remains very high ($R^2 =$
274 0.96 and $r = 0.98$, average values from ten times of ten-fold cross-validation). To keep the high prediction
275 accuracy in other application, there are some points should be take caution. Considering the environmental
276 effects on biomass accumulation, the application of our model will require the testing experiments showing
277 similar conducted conditions with that of the reference experiments. This means the plant cultivation

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

278 conditions should be standardized and any noise which might lower image quality should be avoided.
279 Another approach to improve applicability of models, which could not be tested in this study, would be to
280 improve the data base for the training, by acquiring data from additional environment sensors. Temperature,
281 humidity, and illumination data would certainly help to explain differences in the growth patterns among
282 experiments, performed in different growth seasons. To this end, we expect that our approach is extensible
283 by incorporating such sensor data in the data matrices. Furthermore, our results can provide suggestive hints
284 for biologists to setup phenotyping infrastructures for investigation of plant biomass. For instance, a visible
285 light imaging system would be sufficient to accurately predict fresh weight based on the observation that
286 geometric features alone show high prediction accuracy (**Fig. 4A**). However, to investigate dry weight, it
287 would be helpful to include an additional near-infrared camera system under normal growth conditions and
288 an additional fluorescence camera system under drought stress conditions (**Fig. 4B**).

289
290 In contrast to previous studies [2-3, 6-7, 18, 32-37], in which biomass was investigated using only single
291 image-derived parameter (such as projected area) or several geometric parameters, our analyses extended
292 these studies by incorporating more representative features that cover both structural and physiological-
293 related properties into a more sophisticated model. Although the predictive power of our model is roughly
294 higher than that of single feature-based prediction, such as the digital volume (**Fig. 3**) [20], our model also
295 reveals the relative contribution of individual feature in prediction of biomass. The information regarding the
296 importance of each feature will offer new insights into the phenotypic determinants of plant biomass outcome.
297 Interestingly, we found that several top ranked features, such as digital volume and NIR intensity, showed
298 genetic correlations with biomass of fresh weight (**Fig. 4C**) [20], implying these top ranked features may
299 represent the main “phenotypic components” of biomass outcome and can be further used to dissect genetic
300 components underlying biomass accumulation. As image-based high-throughput phenotyping in plants
301 developed mainly in recent years and therefore few corresponding modelling studies have been performed,
302 we believe that our model could be further improved when new types of cameras and/or newly defined
303 features are available.

304
305 In summary, we have developed a quantitative model for dissecting the phenotypic components of biomass

1
2
3
4 306 accumulation based on image data. Apart from predicting biomass outcome, the methods can be used to
5
6 307 determine the most important image-based features related to plant biomass accumulation, which are
7
8 308 promising for subsequent genetic mapping to uncover the genetic basis of biomass.
9

10 309

11 12 13 310 **Potential Implications**

14
15
16 311 As high-throughput plant phenotyping is a technique which is becoming more and more widely used for
17
18 312 automated phenotype in plant research, especially in plant breeding, we anticipate that the methodologies
19
20 313 proposed in this work will have various potential applications. We anticipate that the analysis results will be
21
22 314 useful to advance our views of the phenotypic determinants of plant biomass outcome, and the statistical
23
24 315 methods can be broadly used for other plant species and therefore assist plant breeding in the context of
25
26 316 phenomics.
27

28 317

29 30 31 318 **Materials and Methods**

32 319 **Germplasm and experiments**

33
34
35
36 320 Barley plant image data were obtained as described previously [20, 30]. Briefly, a core set of 16 two-rowed
37
38 321 spring barley cultivars (*Hordeum vulgare* L.) and two parental cultivars of a double haploid (DH) were
39
40 322 monitored for vegetative biomass accumulation. Three independent experiments with identical setup were
41
42 323 performed in a (semi-) controlled greenhouse at IPK by using the automated phenotyping and imaging
43
44 324 platform LemnaTec-Scanalyzer 3D. Experiments were performed consecutively from May to November
45
46 325 2011 over a period of 58 days each (**Supplemental Table S1**). The greenhouse setup enabled sowing for the
47
48 326 next experiment already 2 days before the old experiment ended. For this, new pots were placed in the middle
49
50 327 of the greenhouse, while the old experiment was still on the conveyer belts.
51

52 328

53
54
55 329 Each experiment consisted of two treatments: well-watered (control treatment) and water limited (drought
56
57 330 stress treatment). In each treatment, nine plants per core set cultivar as well as six plants per DH parent were
58
59 331 tested. This resulted in a total of 312 plants per experiment, corresponding to the maximal capacity of the
60
61
62
63
64
65

1
2
3
4 332 phenotyping platform. Watering and imaging were performed daily. Drought stress was imposed by
5
6 333 intercepting water supply from 27 days after sowing (DAS 27) to DAS 44. Stressed plants were re-watered
7
8 334 at DAS 45. In total, for each of the experiments about 100 GB of raw (image) data was accumulated. At the
9
10 335 end of experiments (DAS 58), plants were harvested to measure above-ground biomass in form of plant fresh
11
12 336 weight (FW; for all experiments) and/or dry weight (DW; for experiment 1).
13

14 337

16 338 **Image analysis**

18 339 Image datasets were processed by the barley analysis pipelines in the IAP software (version v1.1.2) [19].
19
20 340 Analysed results were exported in the csv file format via IAP functionalities, which can be used for further
21
22 341 data inspection. The result table includes columns for different phenotypic traits and rows as plants are
23
24 342 imaged over time. The corresponding metadata is included in the result table as well.
25

26 343

28 344 Each plant was characterized by a set of phenotypic traits also referred to as features, which were grouped
29
30 345 into four categories: geometric features, fluorescence-related (FLUO-related) features, colour-related
31
32 346 features and near-infrared-related (NIR-related) features. These traits were defined by considering image
33
34 347 information from different cameras (visible light, fluorescence and near infrared) and imaging views (side
35
36 348 and top views). See the IAP online documentation (<http://iapg2p.sourceforge.net/documentation.pdf>) for
37
38 349 details about trait definition.
39

40 350

43 351 **Feature selection**

45 352 Feature selection was performed with the same procedure as described in [20]. We applied the feature
46
47 353 selection technique to each dataset. Generally, we captured almost identical subset features from different
48
49 354 datasets. We manually added several representative traits due to removal by variance inflation factors. For
50
51 355 example, the digital volume and projected area are highly correlated with each other but we kept both of
52
53 356 them, because we would investigate the predictive power of both features. Moreover, the regression models
54
55 357 we used are insensitive to collinear features. We thus kept as much representative features as possible. To
56
57 358 apply the prediction models among different datasets, a common set of features supported by all the datasets
58
59 359 was used.
60
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

360

361 **Data transformation**

362 Each plant can be presented by a representative list of phenotypic traits, resulting in a matrix $X_{n \times m}$ for each
363 experiment, where n is the number of plants and m is the number of phenotypic traits. Missing values were
364 filled by mean values of other replicated plants. To make the image-derived parameters from diverse sources
365 comparable, we normalized the columns of X by dividing the values with the maximum value of each column
366 across all plants. Plants with empty values of manual measurements (FW and DW) were discarded for
367 analysis. These transformed data sets were subjected to regression models.

368

369 **Hierarchical clustering analysis and PCA**

370 Hierarchical clustering analysis (HCA) and principle component analysis (PCA) were performed on the
371 transformed data matrix $X_{n \times m}$ in the same way as described in [20]. We also performed HCA using the
372 genotype-level mean value of FW data to check the similarity of overall plant growth patterns in different
373 experiments.

374

375 **Models for predicting plant biomass**

376 To understand the underlying relationship between image-derived parameters and the accumulated biomass
377 (such as FW and DW), we constructed predictive models based on four different machine-learning methods:
378 multivariate linear regression (MLR), multivariate adaptive regression splines (MARS), random forest (RF)
379 and support vector regression (SVR). In these models, the normalized phenotypic profile matrices $X_{n \times m}$ for
380 a representative list of phenotypic traits were used as predictors (explanatory variables) and the measured
381 DW/FW as the response variable Y .

382

383 All these models were implemented in R (<http://www.r-project.org/>; release 2.15.2). To assess the relative
384 contribution of each phenotypic trait to predicting the biomass. We also calculated the relative feature
385 importance for each model. Specifically, for the MLR model, we used the “lm” function in the base
386 installation packages. The relative importance of predictor variables in the MLR model was estimated by a
387 heuristic method [38] which decomposes the proportionate contribution of each predictor variable to R^2 . For

1
2
3
4 388 MARS, we used the “earth” function in the *earth* R package. The “number of subsets (nsubsets)” criterion
5
6 389 (counting the number of model subsets that include the variable) was used to calculate the variables feature
7
8 390 importance, which is implemented in the “evimp” function. For the RF model, we used the *randomForest* R
9
10 391 package which implements Breiman's random forest algorithm [31]. We chose the “%IncMSE” (increase of
11
12 392 mean squared error) to represent the criteria of relative importance measure. For SVR, we utilized the *e1071*
13
14 393 R package which provides functionalities to use the *libsvm* library [39]. The absolute values of the
15
16 394 coefficients of the normal vector to the “optimal” hyperplane can be considered as the relative importance of
17
18 395 each predictor variable contributing to regression [40, 41].
19
20
21

22 397 **Evaluation of the prediction models**

23
24 398 To evaluate the performance of the predictive models, we adopted a 10-fold cross-validation strategy to check
25
26 399 the prediction power of each regression model. Specifically, each dataset was randomly divided into a training
27
28 400 set (90% of plants) and a testing set (10% of plants). We trained a model on the training data and then applied
29
30 401 it to predict biomass for the testing data. Afterwards, the predicted biomass in the testing set was compared
31
32 402 with the manually measured biomass. The predictive accuracy of the model can be measured by
33

- 34 403 1) the Pearson correlation coefficient (PCC; r) between the predicted values and the observed values;
35
36 404 2) the coefficient of determination (R^2) which equals to the fraction of variance of biomass explained
37
38
39 405 by the model, defined as

$$40 406 \quad R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

41 407 where SS_{res} and SS_{tot} are the sum of squares for residuals and the total sum of squares, respectively, \hat{y}_i the
42
43
44 408 predicted and y_i the observed biomass of the i th plant, \bar{y} is the mean value of the observed biomass; and
45
46
47

- 48 409 3) the root mean squared relative error of cross-validation, defined as
49

$$50 410 \quad \text{RMSRE} = \sqrt{\frac{\sum_{i=1}^s \left(\frac{y_i - \hat{y}_i}{y_i} \right)^2}{s}}$$

51
52
53
54
55 411 where s denotes the sample size of the testing dataset.
56

57 412 We repeated the cross-validation procedure ten times. The mean and standard deviation of the resulting R^2
58
59 413 and RMSRE values were calculated across runs.
60
61
62
63
64
65

1
2
3
4 414

5
6 415 To evaluate the applicability of our methods across seasons (thus different growth environments) and
7
8 416 treatments (e.g., control versus drought stress) in the same season, we applied the models in different contexts
9
10 417 with cohort validation. Specifically, we trained the biomass prediction models under one specific context and
11
12 418 predicted biomass in another different context and *vice versa*. The predictive accuracy of the model was
13
14 419 evaluated based on the measures R^2 and RMSRE as described above. Furthermore, the predictive power was
15
16 420 reflected by the bias μ between the predicted and observed values, defined as

17
18
19 421
$$\mu = \frac{1}{n} \cdot \sum_{i=1}^n \frac{\hat{y}_i - y_i}{y_i}$$

20
21 422 where n denotes the sample size of the dataset. This bias indicates over- ($\mu > 0$) or under-estimation ($\mu < 0$)
22
23 423 of biomass.

24
25 424

26
27
28 425 **Availability of source code and requirements**

- 29
30 426
- Project name: Modeling of plant biomass accumulation with HTP data
- 31
32 427
- Project home page: <https://github.com/htpmod/HTPmod>
- 33
34 428
- Operating system(s): Windows, Linux and Mac OS.
- 35
36 429
- Programming language: R
- 37
38 430
- License: open source under GNU GPL v3.0.
- 39

40 431

41
42 432 **Availability of supporting data and materials**

43
44 433 The raw image data sets as well as analysed data supporting the results of this article are available in the PGP
45
46 434 repository [42] under 10.5447/IPK/2017/24, 10.5447/IPK/2017/25 and 10.5447/IPK/2017/26, according to
47
48 435 the ISA-Tab format and the recommendations of the MIAPPE (Minimum Information About a Plant
49
50 436 Phenotyping Experiment) standard [43]. The selected data for modelling are available in the **Supplemental**
51
52 437 **Data S1**. Supporting data, including metadata tables, raw image files and an archival copy of HTPmod are
53
54 438 also available via the *GigaScience* repository GigaDB [44]

55
56 439
57
58
59
60
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

440 **Declarations**

441 **List of abbreviations**

442 DAS: Days After Sowing

443 DW: Dry Weight

444 FLUO: Fluorescence

445 FW: Fresh Weight

446 HCA: Hierarchical Clustering Analysis

447 HTP: High-Throughput Phenotyping

448 MLR: Multivariate Linear Regression

449 MARS: Multivariate Adaptive Regression Splines

450 NIR: Near-Infrared

451 PCA: Principal Component Analysis

452 PCC: Pearson Correlation Coefficient

453 RF: Random Forest

454 RMSRE: Root Mean Squared Relative Error

455 SVR: Support Vector Regression

456

457 **Consent for publication**

458 Not applicable.

459

460 **Funding**

461 This work was supported by the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), the Robert

462 Bosch Stiftung (32.5.8003.0116.0) and the Federal Agency for Agriculture and Food (BEL, 15/12-13, 530-

463 06.01-BiKo CHN) and the Federal Ministry of Education and Research (BMBF, 0315958A and 031A053B).

464 This research was furthermore enabled with support of the European Plant Phenotyping Network (EPPN,

465 grant agreement no. 284443) funded by the FP7 Research Infrastructures Programme of the European Union.

466

467 **Competing interests**

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

468 The authors declare that they have no competing interests.

469

470 **Author contributions**

471 D.C. designed the research. C.K. and M.C. supervised the project. K.N. and G.A. performed the LemnaTec
472 experiments. D.A. created the ISA-Tab formatted description and uploaded data records in the PGP repository.
473 J.M.P. and C.K. analyzed image data. D.C. implemented the methods, analyzed data, interpreted the results,
474 and wrote the manuscript with contribution from R.S.. All authors read and approved the final version of the
475 article.

476

477 **Acknowledgements**

478 We would like to thank Ingo Mücke for his management of the LemnaTec system operations. We thank
479 Michael Ulrich for performing software tests and helping in data analysis. We would like to thank Dr. Malia
480 Gehan and Dr. Christian Fournier for their helpful comments and suggestions.

481

482 **Figure Legends**

483 **Figure1.** Modeling pipeline for predicting plant biomass accumulation based on image-derived parameters.

484 **(A)** Input data, including high-throughput image data and manually measured biomass data. Plants were
485 phenotyped using various cameras such as visible (or color), fluorescence (FLUO) and near-infrared (NIR)
486 sensors. Image analysis was performed with IAP software [10] for feature extraction. The same plants were
487 harvested and measured at the end of growth. Generally, two types of biomass were measured: fresh weight
488 (FW) and dry weight (DW). **(B)** Trait processing. All the phenotypic traits were grouped into four categories:
489 geometric, color-related, FLUO-related and NIR-related traits. Phenotypic data were subjected to quality
490 check to remove low-quality data. **(C)** Each plant was described by a list of traits, resulting in a predictor
491 matrix whose rows represent plants and columns represent image-based traits. This matrix was used to
492 predicted plant biomass accumulation by MLR (multivariate linear regression), MARS (multivariate adaptive
493 regression splines), RF (random forest) and SVR (support vector regression) models. The right panel
494 represents the schema of model validation. In the first schema, a dataset (Dataset 1) was divided into training
495 set and testing set in a ten-fold cross-validation manner. In the second schema, the whole of one dataset
496 (Dataset 1) was used for training and another dataset (Dataset 2) was used for testing. **(D)** Model selection,
497 evaluation and result interpretation. The correlation of the predicted values and measured values was used to
498 assess the overall performance of the model.

500 **Figure 2.** Predictability of image-based traits to plant biomass.

501 **(A)** Schema depicting three consecutive high-throughput phenotyping experiments in barley. Plants in each
502 experiment were harvested for biomass measurements: fresh weight (FW; for all experiments) and dry weight
503 (DW; only for experiment 1). **(B)** Scatter plots showing projections of the top four Principal components
504 (PCs) based on PCA of image-based data. The component scores (shown in points) are colored and shaped
505 according to the experiments (as legend listed in the box). The component loading vectors (represented in
506 lines) of all traits (as colored according to their categories) were superimposed proportionally to their
507 contribution. **(C)** Boxplot showing the distribution of FW across different experiments. **(D)** A dendrogram
508 from cluster analysis based on the means of FW data over genotypes. **(E)** Pearson's correlation (mean values
509 in the three datasets) between image-based traits and FW. Traits with the largest mean correlations values are

1
2
3
4 510 labeled: 1 -- sum of leaf length (side view), 2 -- sum of FLUO intensity (side), 3 -- plant area border length
5
6 511 (side), 4 -- sum of NIR intensity (top), 5 -- sum of FLUO intensity (top), 6 -- projected area (top), 7 --
7
8 512 projected area (side) and 8 -- digital volume.
9

10 513

11
12 514 **Figure 3.** Quantitative relationship between image-based features and plant biomass.

13
14 515 (A) and (C) Scatter plots of manually measured plant biomass (fresh weight [FW] and dry weight [DW])
15
16 516 versus predicted biomass values using four prediction models: multivariate linear regression (MLR),
17
18 517 multivariate adaptive regression splines (MARS), random forest (RF) and support vector regression (SVR).
19
20 518 The red line indicates the expected prediction ($y = x$). The quantitative relationship between image-based
21
22 519 features and biomass was evaluated by Pearson's correlation coefficient (PCC r and its corresponding p -
23
24 520 value), RMSRE (root mean squared relative error) and the percentage of variance explained by the models
25
26 521 (the coefficient of determination R^2). (B) and (D) Summary of the predictive power of each regression model.
27
28 522 The results were based on ten-fold cross-validation with ten trials. Models were evaluated based on control
29
30 523 plants, stressed plants and the whole set of plants. The solid lines represent the predictive performance based
31
32 524 on the single "digital volume" feature.
33

34 525

35
36 526 **Figure 4.** The relative importance of image-based features in prediction of plant biomass.

37
38 527 The capabilities of different types of image-based features to predict plant biomass based on evaluation of
39
40 528 either fresh weight (FW) (A) or dry weight (DW) (B). The overall predictive accuracies of each type of
41
42 529 features are indicated. Grey bar denote the predictive accuracy using all features. The relative importance of
43
44 530 each feature in the Random Forest model (upper panel) and the predictive accuracy of each individual feature
45
46 531 as the single predictor (lower panel) based on investigation of either FW (C) or DW (D). The calculation was
47
48 532 based on the whole set of plants (control and stressed plants). Note that feature labels are shared in the upper
49
50 533 and lower panels. Features are shown in numbers as ordered by their names. The three features highlighted
51
52 534 in the red dash box are digital volume, projected side area and projected top area. (E) Comparison of the
53
54 535 relative importance of features in prediction of FW and DW. The top six most different features are
55
56 536 highlighted and labeled.
57

58
59 537
60
61
62
63
64
65

1
2
3
4 538 **Figure 5.** Comparison of prediction accuracy across different experiments.

539 (A) Biomass prediction across experiments. Models were trained using data from one experiment and were
540 applied to another experiment for prediction. The whole set of plants (i.e., “control + stressed” plant) were
541 used in the analysis. Brown triangles denote stressed plants and green circles control plants. Red box indicates
542 that the prediction accuracy is relatively high between experiments 1 (Exp. 1) and 2. (B) Boxplots of
543 coefficient determination (R^2 , left), Pearson's correlation coefficients (r , middle) and the root mean squared
544 relative error (RMSRE, right) for different comparisons. “Within” denotes a model trained and tested on data
545 from the same dataset with specific treatments (control, stress or both), and “Cross” represents a model
546 trained on one dataset and tested on another dataset. “Control → stress” denotes a model trained on data with
547 control treatment and tested on data with stress treatment, and vice versa for “stress → control”. The number
548 of possible analyses for each category was shown above the boxes.

549
550 **Figure 6.** Comparison of prediction accuracy across different treatments. Refer to **Figure 5A** for legend. The
551 analysis was performed for control and stressed plants separately.

552 553 **Supplemental Data**

554 The following supplemental materials are available.

555 **Supplemental Figure S1.** The relative importance of image-based features in prediction of biomass in
556 control plants. Refer to **Figure 4** for legend. The calculation was based on control plants.

557 **Supplemental Figure S2.** The relative importance of image-based features in prediction of biomass in
558 stressed plants. Refer to **Figure 4** for legend. The calculation was based on stressed plants.

559
560 **Supplemental Table S1.** Overview of three high-throughput phenotyping experiments in barley.

Experiment	#plants/#genotypes ¹	Date of sowing	Date of harvesting	Biomass ²
Exp. 1 (1121KN)	310/18	27.05.2011	24.07.2011	FW & DW
Exp. 2 (1130KN)	310/18	22.07.2011	18.09.2011	FW

Exp. 3 (1137KN)	309/18	16.09.2011	13.11.2011	FW & DW
-----------------	--------	------------	------------	---------

561 ¹Number of plants or genotypes used in analysis (filtered data).

562 ²Types of biomass measurement. FW: fresh weight; DW: dry weight.

563

564 **Supplemental Data S1.** Manual data and image-derived data in the three experiments.

565

566 **References**

- 567 1. Catchpole WR and Wheeler CJ: **Estimating plant biomass: a review of techniques.** *Australian Journal*
568 *of Ecology* 1992, **17**: 121–131.
- 569 2. Tackenberg O: **A new method for non-destructive measurement of biomass, growth rates, vertical**
570 **biomass distribution and dry matter content based on digital image analysis.** *Annals of botany* 2007,
571 **99(4):777-783.**
- 572 3. Golzarian MR, Frick RA, Rajendran K, Berger B, Roy S, Tester M, Lun DS: **Accurate inference of**
573 **shoot biomass from high-throughput images of cereal plants.** *Plant Methods* 2011, **7**:2.
- 574 4. Fahlgren N, Gehan MA and Baxter I: **Lights, camera, action: high-throughput plant phenotyping is**
575 **ready for a close-up.** *Current Opinion in Plant Biology* 2015, **24**:93–99.
- 576 5. Montes JM, Technow F, Dhillon BS, Mauch F, Melchinger AE: **High-throughput non-destructive**
577 **biomass determination during early plant development in maize under field conditions.** *Field Crops*
578 *Research* **121(2)**: 268-273.
- 579 6. Feng H, Jiang N, Huang C, Fang W, Yang W, Chen G, Xiong L, Liu Q: **A hyperspectral imaging system**
580 **for an accurate prediction of the above-ground biomass of individual rice plants.** *Review of*
581 *Scientific Instruments* 2013, **84(9)**:095107-095107.
- 582 7. Neumann K, Zhao Y, Chu J, Keilwagen J, Reif JC, Kilian B, Graner A: **Genetic architecture and**
583 **temporal patterns of biomass accumulation in spring barley revealed by image analysis.** *BMC plant*
584 *biology* 2017, **17(1)**:137.
- 585 8. Zhang X, Huang C, Wu D, Qiao F, Li W, Duan L, Wang K, Xiao Y, Chen G, Liu Q et al: **High-**
586 **Throughput Phenotyping and QTL Mapping Reveals the Genetic Architecture of Maize Plant**

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

587 **Growth.** *Plant Physiology* 2017, **173** (3): 1554-1564.

588 9. Muraya MM, Chu J, Zhao Y, Junker A, Klukas C, Reif JC, Altmann T: **Genetic variation of growth**

589 **dynamics in maize (*Zea mays* L.) revealed through automated non-invasive phenotyping.** *The*

590 *Plant Journal* 2017, **89**(2):366-380.

591 10. Fahlgren N., Feldman M., Gehan M.A., Wilson M.S., Shyu C., Bryant D.W., Hill S.T., McEntee C.J.,

592 Warnasooriya S.N., Kumar I, Ficor T., Turnipseed S., Gilbert K.B., Brutnell T.P., Carrington J.C.,

593 Mockler T.C., and Baxter I: **A Versatile Phenotyping System and Analytics Platform Reveals Diverse**

594 **Temporal Responses to Water Availability in *Setaria*.** *Molecular Plant* 2015, **8**: 1520–1535.

595 11. Yang WN, Zilong Guo ZL, Huang CL, Duan LF, Chen GX, Jiang N, Fang W, Feng H, Xie WB, Lian

596 XM et al: **Combining high-throughput phenotyping and genome-wide association studies to reveal**

597 **natural genetic variation in rice.** *Nature Communications* 2014, **5**:5087.

598 12. Ehlert D, Horn H-J, Adamek R: **Measuring crop biomass density by laser triangulation.** *Computers*

599 *and electronics in agriculture* 2008, **61**(2):117-125.

600 13. Ehlert D, Heisig M, Adamek R: **Suitability of a laser rangefinder to characterize winter wheat.**

601 *Precision Agric* 2010, **11**(6):650-663.

602 14. Erdle K, Mistele B, Schmidhalter U: **Comparison of active and passive spectral sensors in**

603 **discriminating biomass parameters and nitrogen status in wheat cultivars.** *Field Crops Research*

604 2011, **124**(1):74-84.

605 15. Busemeyer L, Ruckelshausen A, Moller K, Melchinger AE, Alheit KV, Maurer HP, Hahn V, Weissmann

606 EA, Reif JC, Wurschum T: **Precision phenotyping of biomass accumulation in triticale reveals**

607 **temporal genetic patterns of regulation.** *Scientific reports* 2013, **3**:2442-2442.

608 16. Cao Q, Miao Y, Wang H, Huang S, Cheng S, Khosla R, Jiang R: **Non-destructive estimation of rice**

609 **plant nitrogen status with Crop Circle multispectral active canopy sensor.** *Field Crops Research*

610 2013, **154**:133-144.

611 17. Fernandez MGS, Bao Y, Tang L, Schnable PS: High-throughput phenotyping for biomass crops. *Plant*

612 *Physiology* 2017, DOI: 10.1104/pp.17.00707.

613 18. Neilson EH, Edwards AM, Blomstedt CK, Berger B, Møller BL, Gleadow RM: **Utilization of a high-**

614 **throughput shoot imaging system to examine the dynamic phenotypic responses of a C4 cereal**

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

615 **crop plant to nitrogen and water deficiency over time.** *Journal of experimental botany* 2015.

616 19. Klukas C, Chen D, Pape JM: **Integrated Analysis Platform: An Open-Source Information System**
617 **for High-Throughput Plant Phenotyping.** *Plant Physiol* 2014, **165**(2):506-518.

618 20. Chen D, Neumann K, Friedel S, Kilian B, Chen M, Altmann T, Klukas C: **Dissecting the phenotypic**
619 **components of crop plant growth and drought responses based on high-throughput image analysis.**
620 *Plant Cell* 2014, **26**:4636-4655.

621 21. Cheng C, Alexander R, Min R, Leng J, Yip KY, Rozowsky J, Yan K-K, Dong X, Djebali S, Ruan Y *et*
622 *al*: **Understanding transcriptional regulation by integrative analysis of transcription factor binding**
623 **data.** *Genome research* 2012, **22**(9):1658-1667.

624 22. Cheng C, Gerstein M: **Modeling the relative relationship of transcription factor binding and histone**
625 **modifications to gene expression levels in mouse embryonic stem cells.** *Nucleic Acids Research* 2012,
626 **40**(2):553-568.

627 23. Cheng C, Yan K-K, Yip KY, Rozowsky J, Alexander R, Shou C, Gerstein M, others: **A statistical**
628 **framework for modeling gene expression using chromatin features and application to**
629 **modENCODE datasets.** *Genome Biol* 2011, **12**(2):R15-R15.

630 24. Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, Gingeras TR, Gerstein M, Guigó R,
631 Birney E *et al*: **Modeling gene expression using chromatin features in various cellular contexts.**
632 *Genome Biol* 2012, **13**(9):R53-R53.

633 25. Karlič R, Chung H-R, Lasserre J, Vlahoviček K, Vingron M: **Histone modification levels are predictive**
634 **for gene expression.** *Proceedings of the National Academy of Sciences* 2010, **107**(7):2926-2931.

635 26. Ma B, Wilker EH, Willis-Owen SAG, Byun H-M, Wong KCC, Motta V, Baccarelli AA, Schwartz J,
636 Cookson WOCM, Khabbaz K *et al*: **Predicting DNA methylation level across human tissues.** *Nucleic*
637 *acids research* 2014, **42**(6):3515-3528.

638 27. Zhang W, Spector T, Deloukas P, Bell J, Engelhardt B: **Predicting genome-wide DNA methylation**
639 **using methylation marks, genomic position, and DNA regulatory elements.** *Genome Biology* 2015,
640 **16**(1):14-14.

641 28. Das R, Dimitrova N, Xuan Z, Rollins RA, Haghighi F, Edwards JR, Ju J, Bestor TH, Zhang MQ:
642 **Computational prediction of methylation status in human genomic sequences.** *Proceedings of the*

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

643 *National Academy of Sciences* 2006, **103**(28):10713-10716.

644 29. Zheng H, Wu H, Li J, Jiang S-W: **CpGIMethPred: computational model for predicting methylation**
645 **status of CpG islands in human genome.** *BMC medical genomics* 2013, **6**(Suppl 1):S13-S13.

646 30. Neumann K, Klukas C, Friedel S, Rischbeck P, Chen D, Entzian A, Stein N, Graner A, Kilian B:
647 **Dissecting spatiotemporal biomass accumulation in barley under different water regimes using**
648 **high-throughput image analysis.** *Plant, cell & environment* 2015.

649 31. Breiman L: **Random forests.** *Machine learning* 2001, **45**(1):5-32.

650 32. Dietz H, Steinlein T: **Determination of plant species cover by means of image analysis.** *Journal of*
651 *Vegetation Science* 1996, **7**(1):131-136.

652 33. Leister D, Varotto C, Pesaresi P, Niwergall A, Salamini F: **Large-scale evaluation of plant growth in**
653 **Arabidopsis thaliana by non-invasive image analysis.** *Plant Physiology and Biochemistry* 1999,
654 **37**(9):671-678.

655 34. Paruelo JM, Lauenroth WK, Roset PA: **Estimating aboveground plant biomass using a photographic**
656 **technique.** *Journal of Range Management* 2000:190-193.

657 35. Walter A, Scharr H, Gilmer F, Zierer R, Nagel KA, Ernst M, Wiese A, Virnich O, Christ MM, Uhlig B
658 *et al*: **Dynamics of seedling growth acclimation towards altered light conditions can be quantified**
659 **via GROWSCREEN: a setup and procedure designed for rapid optical phenotyping of different**
660 **plant species.** *New Phytol* 2007, **174**(2):447-455.

661 36. Arvidsson S, Perez-Rodriguez P, Mueller-Roeber B: **A growth phenotyping pipeline for Arabidopsis**
662 **thaliana integrating image analysis and rosette area modeling for robust quantification of genotype**
663 **effects.** *New Phytol* 2011, **191**(3):895-907.

664 37. Hairmansis A, Berger B, Tester M, Roy SJ: **Image-based phenotyping for non-destructive screening**
665 **of different salinity tolerance traits in rice.** *Rice* 2014, **7**(1):16-16.

666 38. Johnson JW: **A Heuristic Method for Estimating the Relative Weight of Predictor Variables in**
667 **Multiple Regression.** *Multivariate Behavioral Research* 2000, **35**(1):1-19.

668 39. Chang C-C, Lin C-J: **LIBSVM: a library for support vector machines.** *ACM Transactions on*
669 *Intelligent Systems and Technology (TIST)* 2011, **2**(3):27.

670 40. Loo LH, Wu LF, Altschuler SJ: **Image-based multivariate profiling of drug responses from single**

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

671 **cells**. *Nature methods* 2007, **4**(5):445-453.

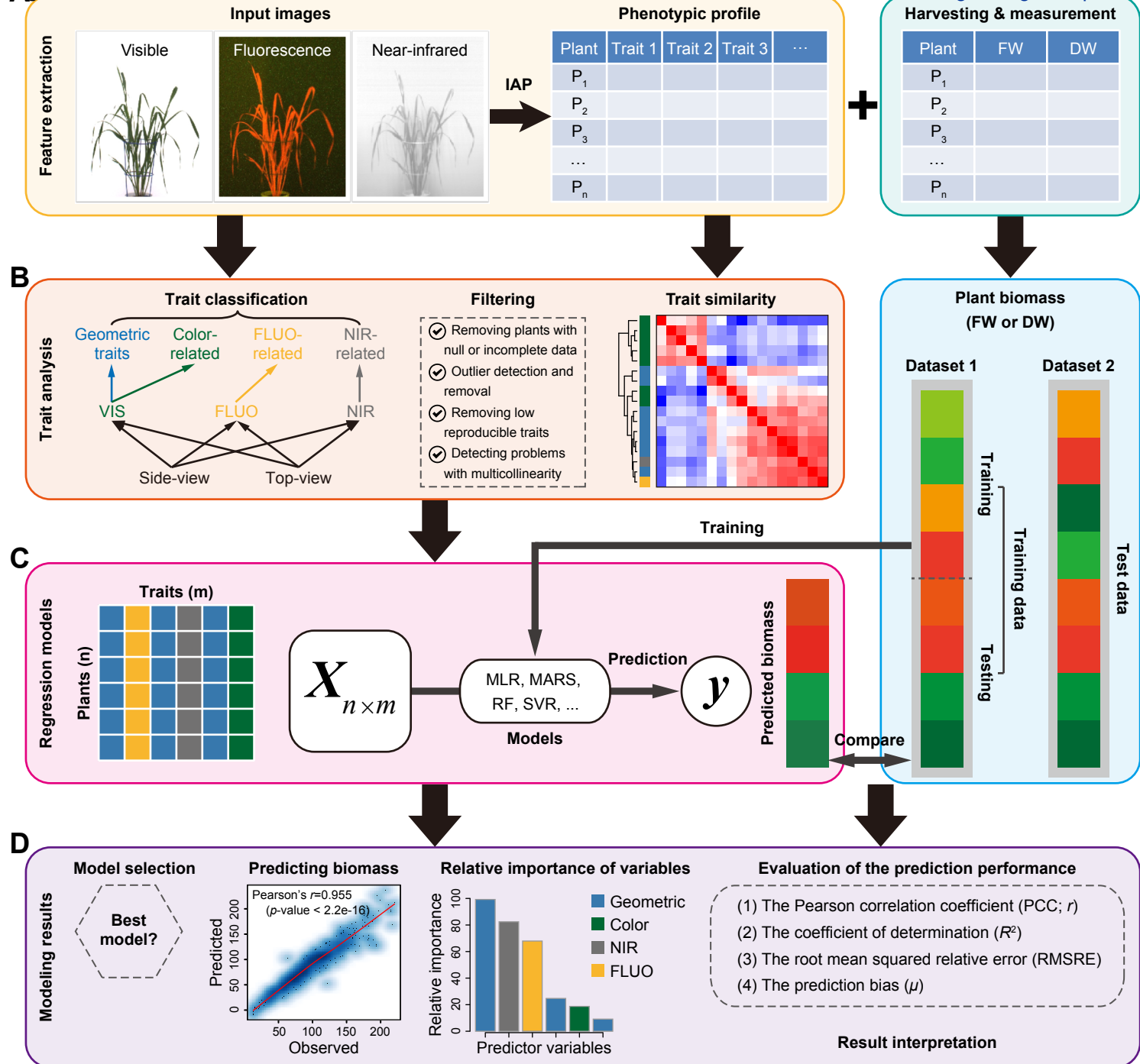
672 41. Iyer-Pascuzzi AS, Symonova O, Mileyko Y, Hao Y, Belcher H, Harer J, Weitz JS, Benfey PN: **Imaging**
673 **and analysis platform for automatic phenotyping and trait ranking of plant root systems**. *Plant*
674 *physiology* 2010, **152**(3):1148-1157.

675 42. Arend D, Junker A, Scholz U, Schuler D, Wylie J, Lange M: **PGP repository: a plant phenomics and**
676 **genomics data publication infrastructure**. *Database : the journal of biological databases and curation*
677 2016, **2016**.

678 43. Cwiek-Kupczynska H, Altmann T, Arend D, Arnaud E, Chen D, Cornut G, Fiorani F, Frohmberg W,
679 Junker A, Klukas C *et al*: **Measures for interoperability of phenotypic data: minimum information**
680 **requirements and formatting**. *Plant methods* 2016, **12**:44.

681 44 Chen D, Shi R, Pape JM, Neumann K, Arend D, Graner A, et al. Supporting data for "Predicting plant
682 biomass accumulation from image-derived parameters". *GigaScience* Database 2018.
683 <http://dx.doi.org/10.5524/100392>

684
685
686



A Figure 2

Experiments performed in 2011

[Click here to download Figure Figure 2.pdf](#)

May 27 July 22/24 September 16/18 November 13

18 genotypes
2 treatments

Exp. 1
312 plants

Exp. 2
312 plants

Exp. 3
312 plants

3 sensors
4 types of traits

High-throughput phenotyping and harvesting

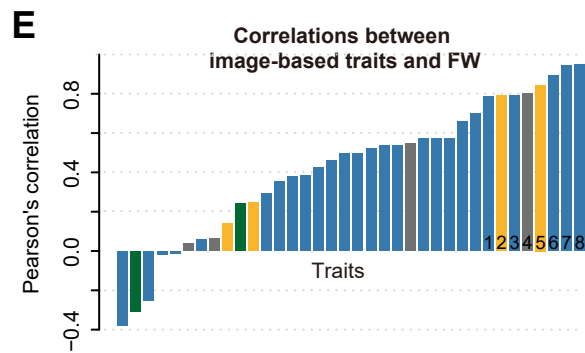
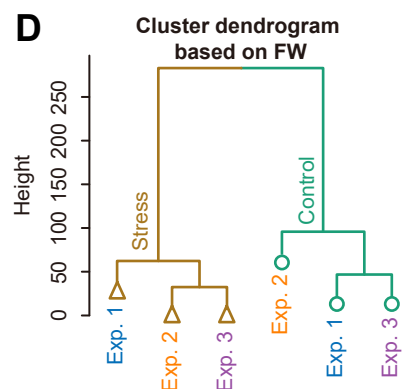
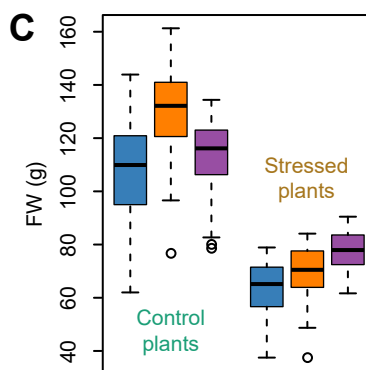
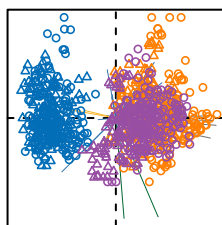
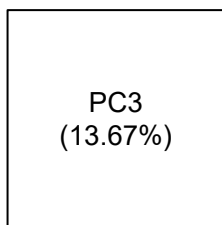
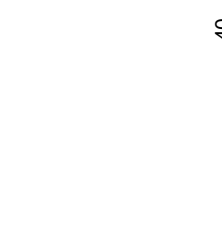
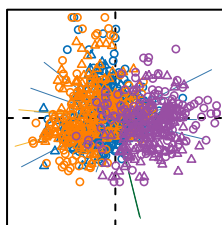
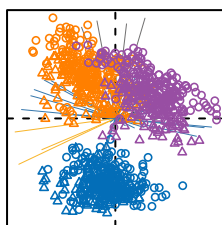
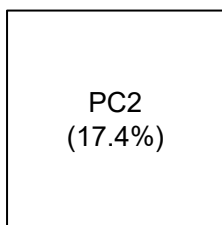
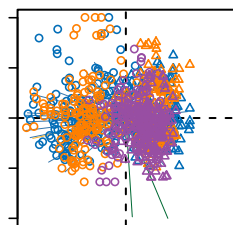
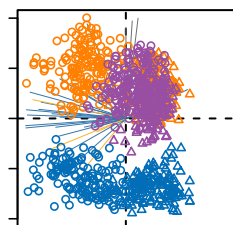
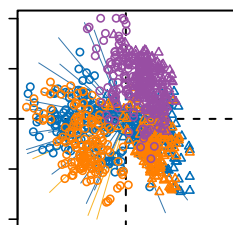
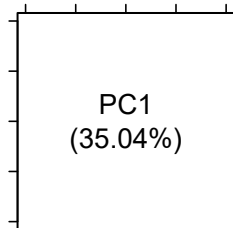
Image data
(features)

Models
Predicting

Biomass
(FW/DW)

B PCA based on image-derived traits

-1.0 -0.5 0.0 0.5 1.0



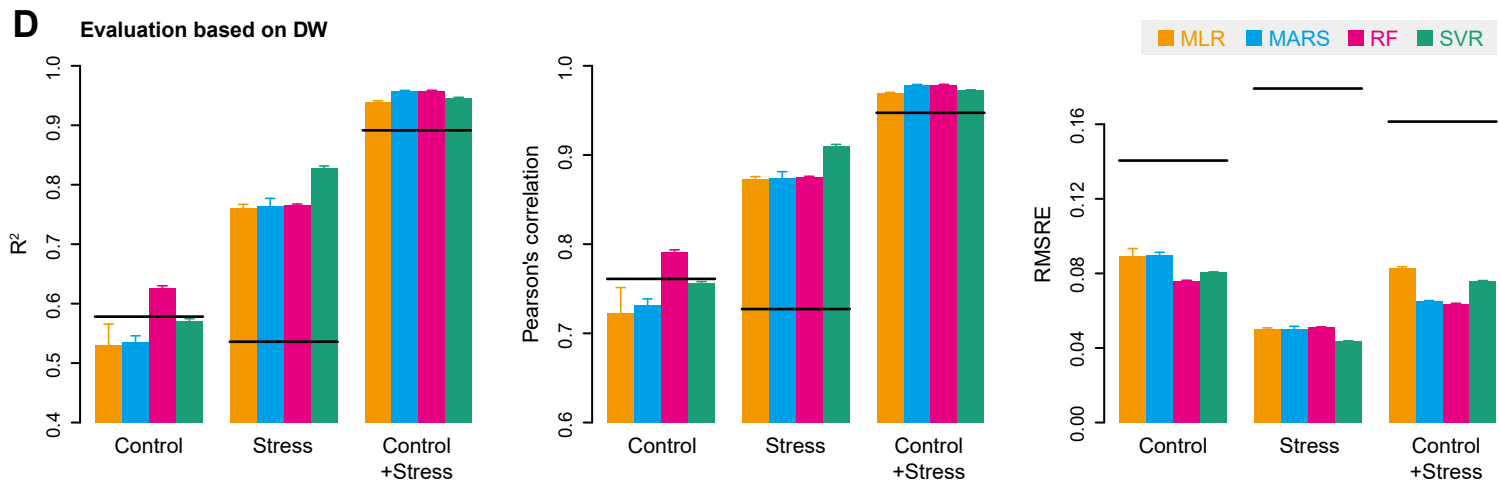
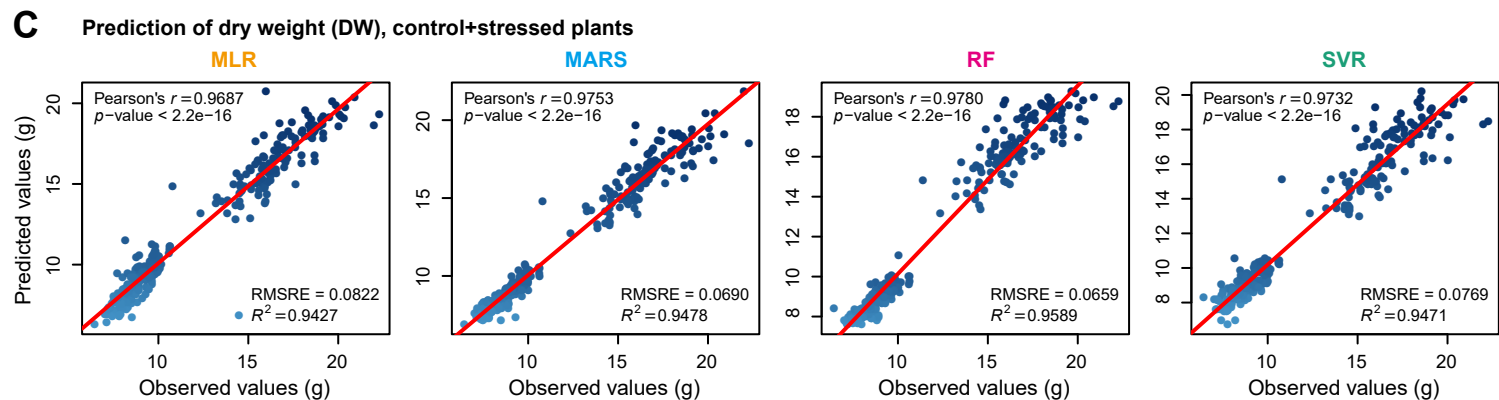
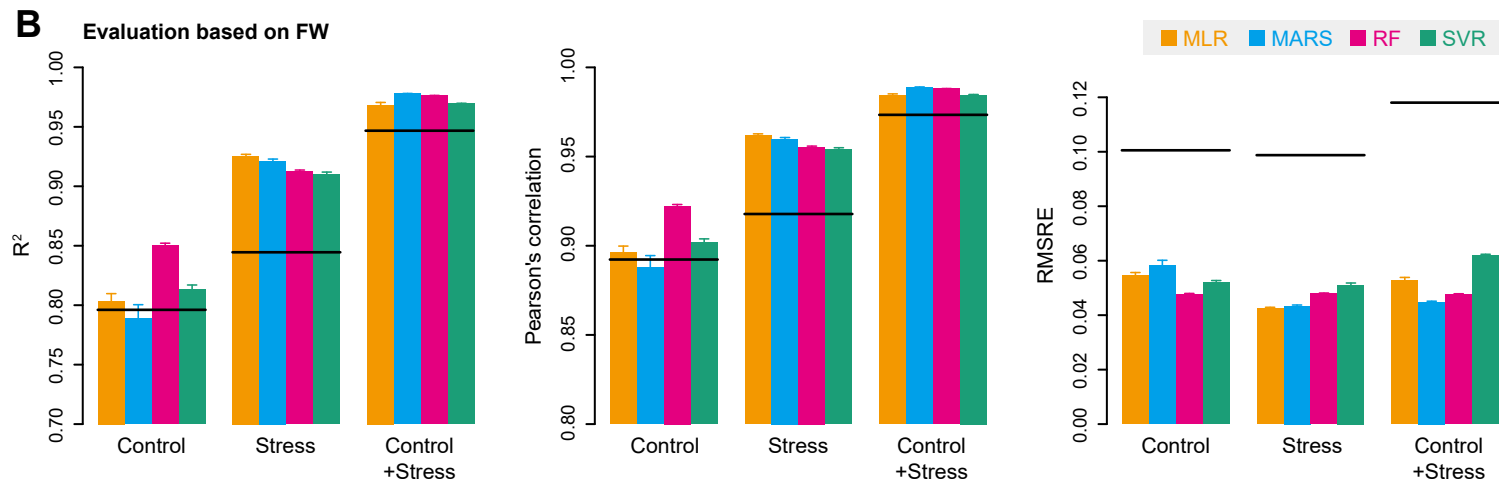
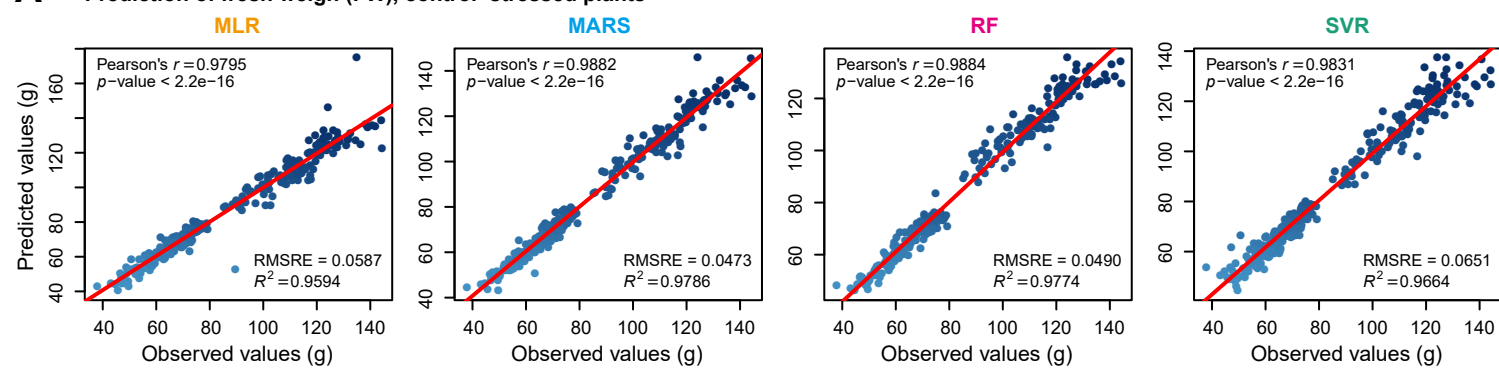
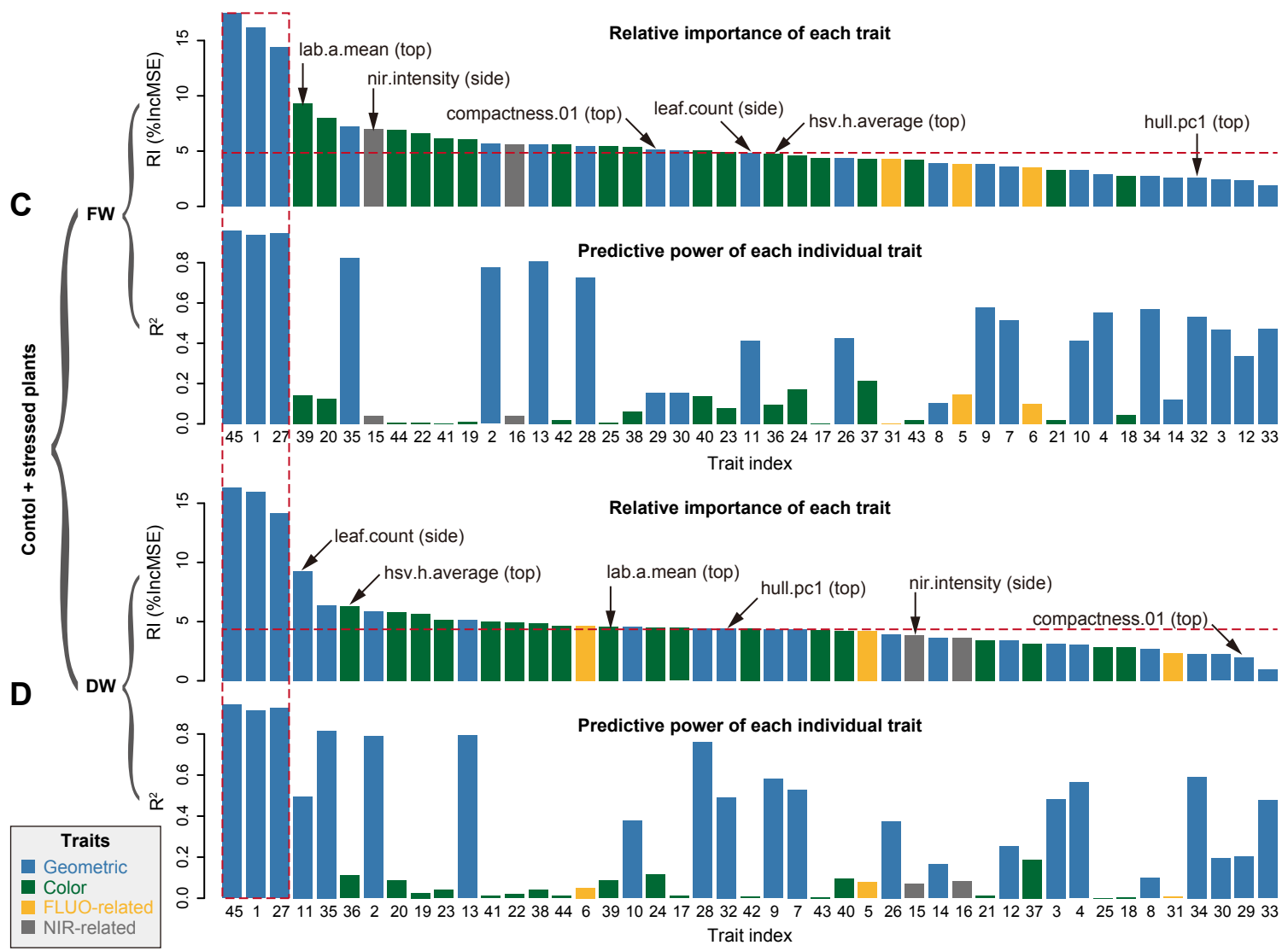
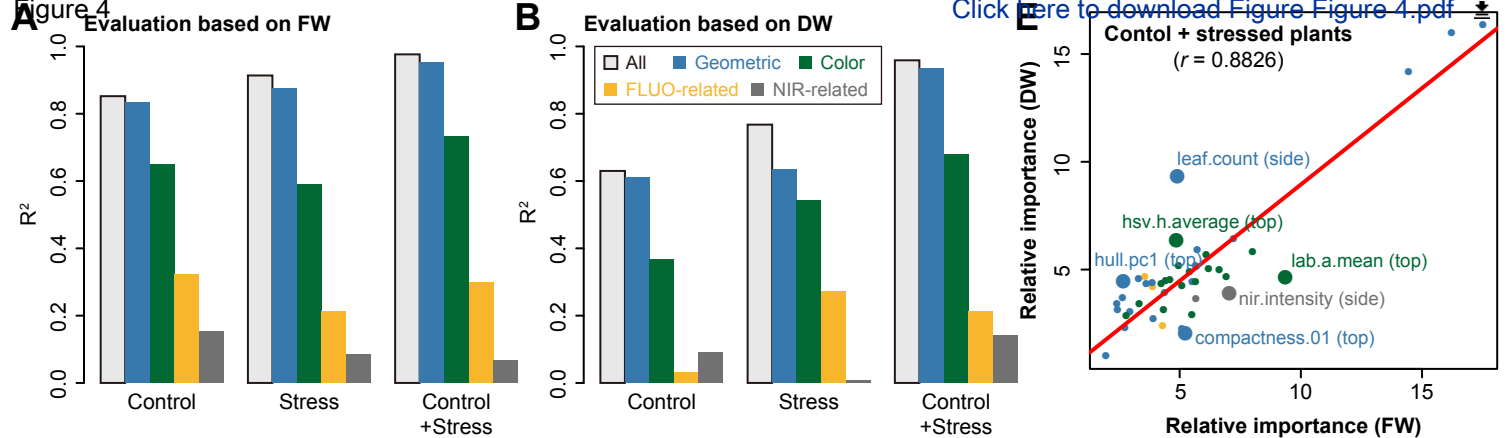
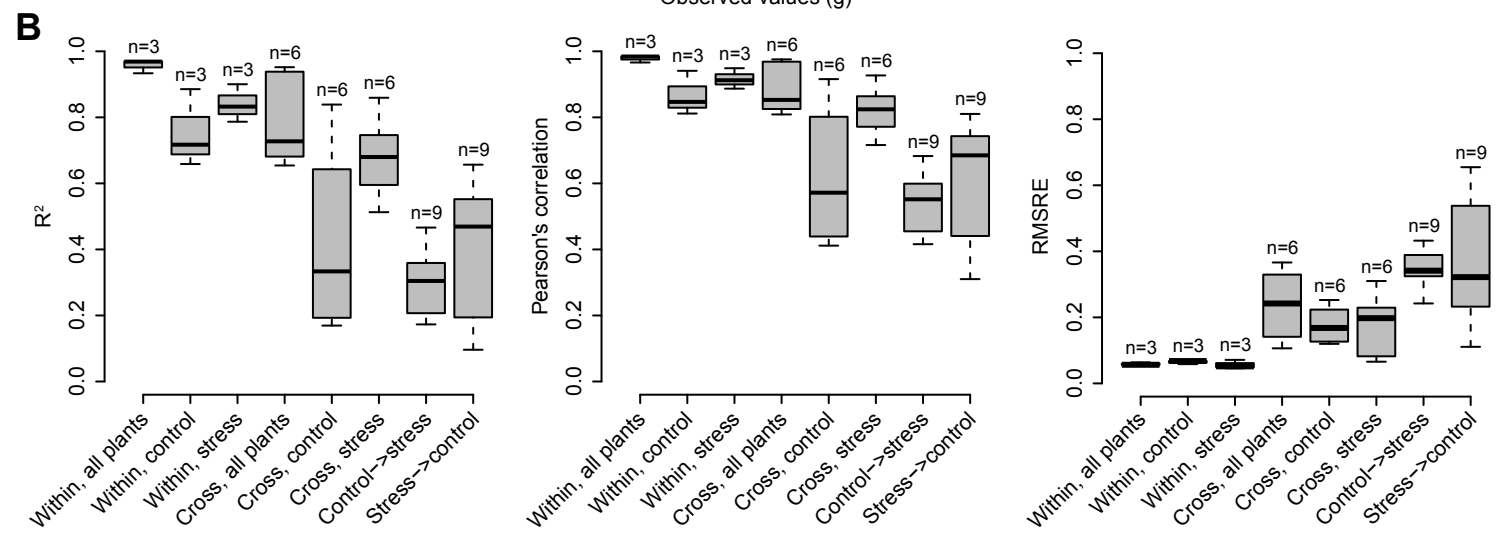
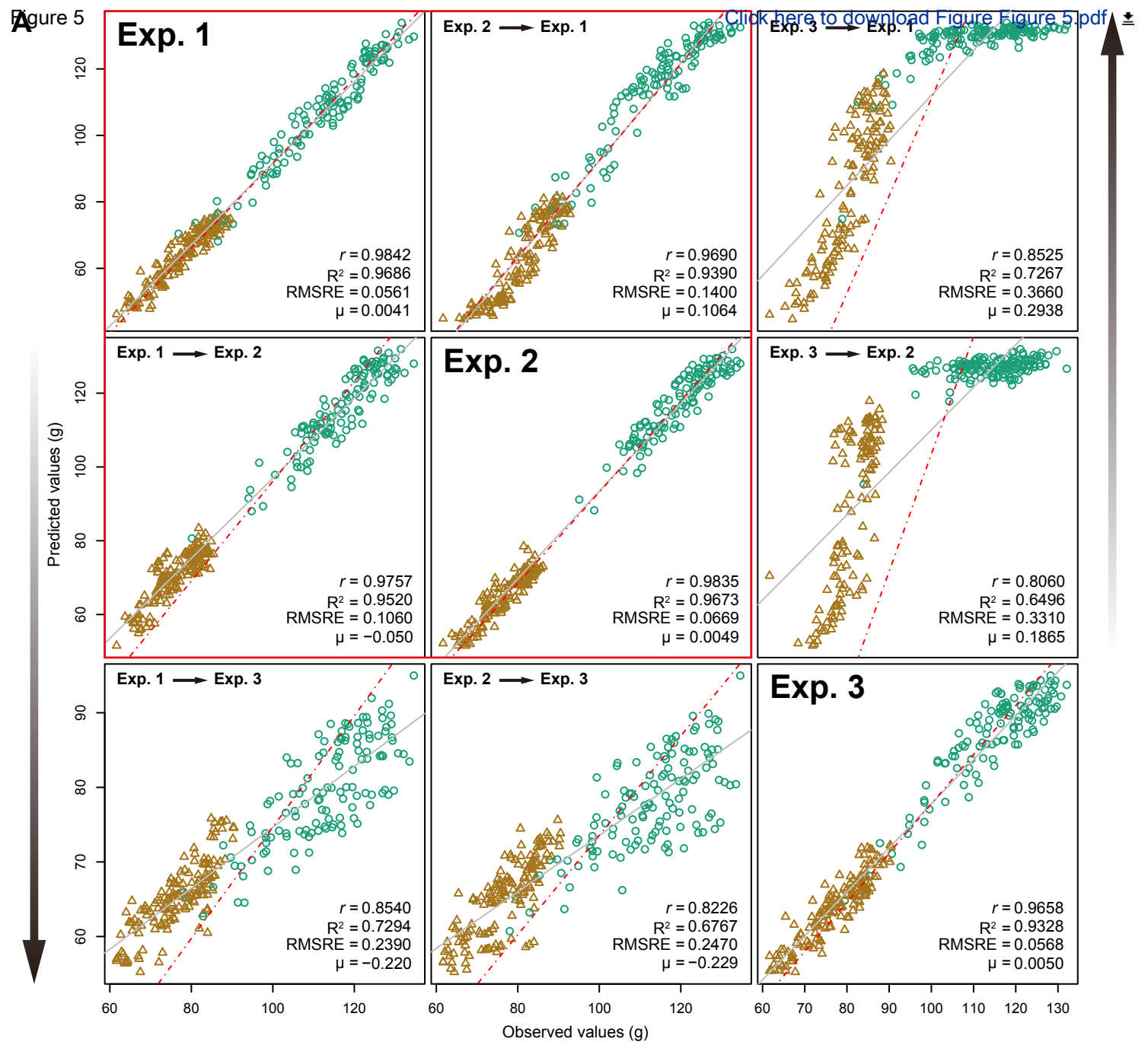
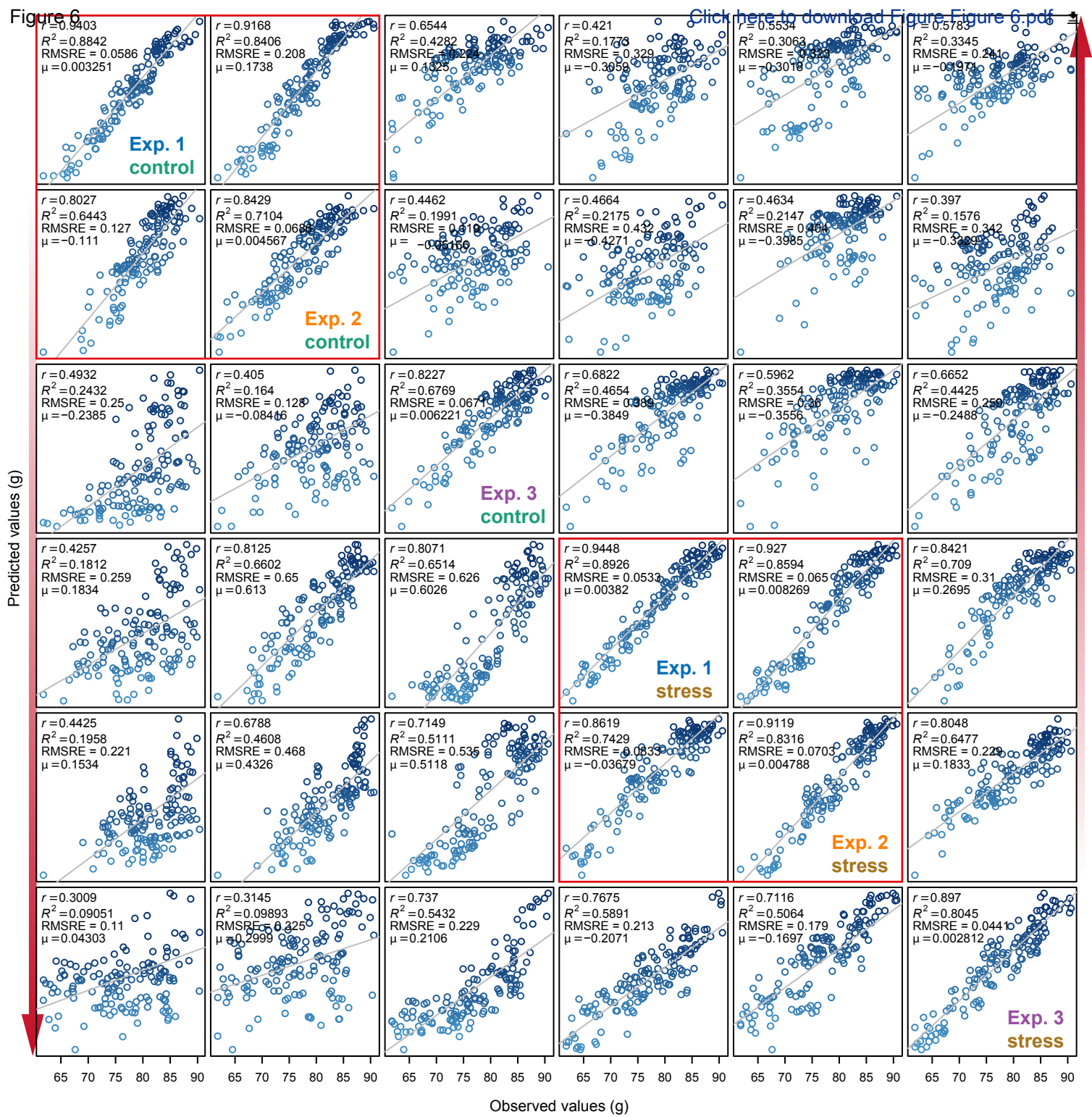


Figure 4




[Click here to download Figure Figure 4.pdf](#)







Click here to access/download
Supplementary Material
Supplemental Figures.pdf





Click here to access/download
Supplementary Material
Supplemental Data S1.xlsx

