

1  
2  
3  
4  
5 1 Combining semi-automated image analysis  
6  
7  
8 2 techniques with machine learning algorithms  
9  
10 3 to accelerate large scale genetic studies.  
11  
12  
13 4  
14 5  
15 6

16 7 Jonathan A. Atkinson<sup>1+</sup>, Guillaume Lobet<sup>2, 3+</sup>, Manuel Noll<sup>4</sup>, Patrick E. Meyer<sup>4</sup>, Marcus Griffiths<sup>1</sup>,  
18 8 Darren M. Wells<sup>1\*</sup>  
19 9  
20 10  
21 11  
22 12  
23 13  
24 14  
25 15  
26 16

- 27 15 1. Centre for Plant Integrative Biology, School of Biosciences, University of Nottingham, United  
28 16 Kingdom  
29 17 2. Agrosphere, IBG3, Forschungszentrum Jülich, Jülich, Germany  
30 18 3. Earth and Life Institute, Université catholique de Louvain, Louvain-la-Neuve, Belgium  
31 19 4. InBios, Université de Liège, Liège, Belgium  
32 20 + These authors contributed equally  
33 21 \* Corresponding author  
34 22  
35 23  
36 24  
37 25  
38 26  
39 27  
40 28  
41 29  
42 30  
43 31  
44 32  
45 33  
46 34  
47 35  
48 36  
49 37  
50 38  
51 39  
52 40  
53 41  
54 42  
55 43  
56 44  
57 45  
58 46  
59 47  
60 48  
61 49  
62 50  
63 51  
64 52  
65 53

48 32 JAA: [jonathan.atkinson@nottingham.ac.uk](mailto:jonathan.atkinson@nottingham.ac.uk), orcid: 0000-0003-2815-0812

49 33 GL: [g.lobet@fz-juelich.de](mailto:g.lobet@fz-juelich.de), orcid: 0000-0002-5883-4572

50 34 MN: [noll@geek-lab.org](mailto:noll@geek-lab.org)

51 35 PEM: [Patrick.Meyer@ulg.ac.be](mailto:Patrick.Meyer@ulg.ac.be)

52 36 MG: [marcus.griffiths@nottingham.ac.uk](mailto:marcus.griffiths@nottingham.ac.uk), orcid: 0000-0003-2349-8967

53 37 DMW: [darren.wells@nottingham.ac.uk](mailto:darren.wells@nottingham.ac.uk), orcid: 0000-0002-4246-4909  
54 38  
55 39  
56  
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

# 1 Abstract

## 2 Background

3 Genetic analyses of plant root systems require large datasets of extracted architectural traits. To  
4 quantify such traits from images of root systems, researchers often have to choose between  
5 automated tools (that are prone to error and extract only a limited number of architectural traits)  
6 or semi-automated ones (that are highly time consuming).

## 7 Findings

8 We trained a Random Forest algorithm to infer architectural traits from automatically extracted  
9 image descriptors. The training was performed on a subset of the dataset, then applied to its  
10 entirety. This strategy allowed us to (i) decrease the image analysis time by 73% and (ii) extract  
11 meaningful architectural traits based on image descriptors. We also show that these traits are  
12 sufficient to identify Quantitative Trait Loci that had previously been discovered using a semi-  
13 automated method.

## 14 Conclusions

15 We have shown that combining semi-automated image analysis with machine learning algorithms  
16 has the power to increase the throughput of large scale root studies. We expect that such an  
17 approach will enable the quantification of more complex root systems for genetic studies. We also  
18 believe that our approach could be extended to other areas of plant phenotyping.  
19

## 20 Keywords

21  
22 Root, plant phenotyping, machine learning, QTL analysis

23

24

# 1 Findings

## 2 Background

3  
4  
5  
6  
7  
8  
9  
10  
11  
12 4 Plant root systems have many physiological roles, including the acquisition of water and nutrients,  
13 5 making them of critical importance for yield establishment in crops. The improvement of root  
14 6 architectural traits will thus be crucial in delivering the yield improvement required to ensure future  
15 7 global food security [1, 2]. Unfortunately, root systems are difficult to analyse and quantify: they  
16 8 are intrinsically complex due to their highly branched tree structure [3], and their growth in an  
17 9 opaque medium (soil) makes them difficult to observe.

10  
11  
12 11 For many years, root researchers have used specific experimental setups to observe and quantify  
13 12 root system architecture. Among these, the "pouch system" is widely used by the community to  
14 13 acquire large number of images of root systems [4–6]. In this approach, plants are grown on the  
15 14 surface of paper allowing the root system to be imaged. The analysis of the resulting root images  
16 15 can be performed either using semi-automated [7, 8] or fully-automated root image analysis  
17 16 software [9, 10]. Semi-automated tools require input and validation by an expert user to faithfully  
18 17 extract the geometry of the root system. However, such user interaction is time consuming, which  
19 18 can strongly hinder the application of these approaches to large datasets (such as those required  
20 19 for quantitative genetic studies). Fully automated software tools are faster, but the extracted  
21 20 descriptors are prone to unexpected errors and the quantified traits are usually less informative  
22 21 [3]. This has led to image analysis being described as a new “bottleneck” in plant phenotyping  
23 22 [11].

24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45 24 Machine learning (a multidisciplinary field of computer science, statistics, artificial intelligence,  
46 25 and information theory) encompasses a range of techniques for the automatic production of  
47 26 analytical models and has been attracting the interest of the plant science community in recent  
48 27 years. Machine learning is breaking new ground in plant science via the automation of procedures  
49 28 and experiments that previously required manual curation. These automated workflows are  
50 29 catalysing the development of new data-driven plant science [12]; including remote sensing [13],  
51 30 species identification [14], and phenotyping [15–18]. Recently, a new approach utilising machine  
52 31 learning algorithms has been proposed for the identification of root system architectural traits; a

1  
2  
3  
4 1 Random Forest model was trained on corresponding ground-truth and image descriptors. The  
5  
6 2 resulting trained model was used to analyse a new set of simulated images and was shown to be  
7  
8 3 much more accurate than the direct image descriptors [3].  
9

10 4  
11 5 Here, we have evaluated this technique using a similar approach with experimental images, and  
12 6 assessed its application to a large scale genetic study. Our rationale was twofold. Firstly, we can  
13 7 reasonably expect a certain level of homogeneity within datasets coming from a single genetic  
14 8 screening as root systems from a given species share common attributes. Secondly, semi-  
15 9 automated root image analysis tools can be used to extract the ground-truth on a subset of  
16 10 images. Such ground truths can be used to train a machine learning algorithm that can then be  
17 11 used to analyse the remaining images in the dataset.  
18  
19 12

20 13 We show that such an approach can (i) yield better results than fully automated software analysis,  
21 14 (ii) is time-efficient compared to performing a semi-automated analysis on the whole dataset and  
22 15 (iii) is able to correctly identify previously found quantitative trait loci (QTL) for root traits.  
23  
24

## 25 16 Overview of the analysis workflow

26 17 The dataset consists of 2614 RGB images of the roots of wheat seedlings obtained from growth  
27 18 pouch experiments as detailed in [5]. The images are of the root systems of the parental lines and  
28 19 94 members of the Savannah x Rialto doubled haploid mapping population, with a minimum of  
29 20 15 replicates per line. All images were processed using the semi-automatic root analysis tool  
30 21 RootNav (RootNav, RRID: SCR\_015584) [7] to extract phenotypic root system architectural traits,  
31 22 representing the “ground-truth” values. A subset of these images has been previously used to  
32 23 identify QTL for root traits [5]. The dataset was divided in two (Fig. 1A): a training dataset,  $D_{train}$ ,  
33 24 of variable size (between 100 and 900 images out of 969) and a test dataset,  $D_{test}$ , of 1645 images,  
34 25 comprising the image set used to identify QTL in the original experimental work [5]. For all images,  
35 26 we first extracted the true values of the traits (ground-truth;  $T_{train}$ ,  $T_{test}$ ) using RootNav [7]. We then  
36 27 extracted a k-dimensional feature vector (“the image descriptor”) of each image using a fully  
37 28 automated analysis pipeline, RIA-J [3]; the corresponding sets are  $I_{train}$ ,  $I_{test}$  (Fig. 1B). We used the  
38 29 extracted data ( $I_{train}$ ,  $T_{train}$ ) to train a Random Forest model  $M: I \rightarrow T$ , to predict the different ground-  
39 30 truths based on the image descriptors [3] (Fig. 1C, F). The trained Random Forest model  $M$  was  
40 31 then applied to the image descriptors  $I_{test}$  from the test dataset  $D_{test}$ , to predict the different ground-  
41 32 truth  $T_{test}$  (named Random Forest estimators, Fig. 1D). The accuracy of both the image descriptors  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 1 and the Random Forest estimators were then compared to the ground-truth acquired with  
5  
6 2 RootNav.

7 3  
8  
9 4 One of the aims of our analysis was to assess the minimal size required for a training dataset.  
10  
11 5 Therefore, we used different numbers of images for training: 100, 200, 300, 400, 500, 600, 700,  
12  
13 6 800 and 900. For each set, we randomly selected the images out of the 969 images that  
14  
15 7 comprised the training dataset, then repeated the training/accuracy procedure described above.  
16  
17 8 To account for the fact that the images were randomly selected, for each test size, we repeated  
18  
19 9 the procedure 10 times.

20 10  
21 11 For each training dataset size, we used the Random Forest estimators to detect QTL regions  
22  
23 12 associated with the different traits quantified (Fig. 1G). The identified QTL regions were then  
24  
25 13 compared to those previously identified using RootNav, as well as those identified using the direct  
26  
27 14 image descriptors.

## 28 29 30 15 **Random Forest estimators have a greater accuracy and greater** 31 32 16 **biological relevance than image descriptors**

33 17  
34  
35 18 It has been previously shown that Random Forest estimators are better at predicting the ground-  
36  
37 19 truth values of various root system metrics compared to direct image descriptors [3]. However,  
38  
39 20 this evaluation used simulated images, rather than a “real” experimental dataset.

40 21  
41 22 Here we show that this approach can also be used with experimental data yielding better results  
42  
43 23 than the direct image descriptors (Fig. 2). We also show that, as expected, increasing the size of  
44  
45 24 the training dataset increases the accuracy of the estimated metrics. For our data, we observe a  
46  
47 25 strong increase in accuracy up to a dataset size of 500 training images, after which the  
48  
49 26 improvement becomes marginal. Our approach also allows for the prediction of new metrics, not  
50  
51 27 obtained using the direct image descriptors. For instance, the direct descriptors do not  
52  
53 28 differentiate between the different root orders, whereas the Random Forest model does.

54 29  
55 30 We observed a decrease in the variability of the predicted values as the number of training images  
56  
57 31 increases. This may be the result of a greater accuracy of the prediction, but may also be due to  
58  
59 32 the fact that the same images are randomly selected for each repetition. As the number of training

1  
2  
3  
4 1 images increases, we expect the number of identical images across repetitions to increase as  
5  
6 2 well (the total number of images in the training set being 969).  
7  
8

### 9 3 Random Forest estimators identify the correct QTLs

10  
11 4  
12 5 Plant phenotyping studies often use mapping populations to dissect the genetic architecture of  
13 6 complex traits by identifying regions of chromosomal DNA that correlate with phenotypic variation  
14 7 termed quantitative trait loci (QTL). The images in our test dataset were used in such a study to  
15 8 identify several QTL for root traits in wheat seedlings [5]. In addition to testing the accuracy of the  
16 9 Random Forest approach in estimating root system parameters, we wanted to know if these  
17 10 parameters could be used reliably for the identification of QTL. Since QTL identification had  
18 11 already been performed on our test dataset, we could directly assess the performance of our new  
19 12 pipeline against the original approach by using the same QTL detection technique on both the  
20 13 direct image descriptors and the traits derived from the random forest models.  
21 14

22  
23 15 The Random Forest models, trained on different numbers of images (100:900), were used on the  
24 16 image descriptors from the test dataset to predict nine estimator datasets (named EST-100 to  
25 17 EST-900) for use in the QTL analysis (see Table 1). This was done to assess the minimum size  
26 18 for the training dataset required for reliable QTL detection, which may be lower than that required  
27 19 to accurately predict the trait values themselves. The R package R/qtl [19] was used for QTL  
28 20 detection on the image descriptor dataset and the nine Random Forest predicted datasets [5].  
29 21 Identified QTL were then directly compared to those found in this paper.  
30 22

31 23 We observed that 12/13 of the expected QTL were correctly identified using the estimators from  
32 24 the Random Forest models trained on 600 or more images (EST-600:EST-900). We also  
33 25 observed that even using the smallest training set of 100 images (EST-100), most of the QTLs  
34 26 were identified (10/12), with 12/13 being identified with the estimators from the model trained with  
35 27 300 images (EST-300). We did not observe an increase in the logarithm of odds (LOD) score with  
36 28 the increase of images (Table 1).  
37 29

38 30 In addition, 4 extra QTL were identified on chromosomes 4D and 6D. Two of these were identified  
39 31 for width and width-depth ratio from EST-300, EST-500 and EST-800 datasets (Table 1). Although  
40 32 in this example, these have been labelled as false positives as they were not detected in the  
41 33 original study, they both have related QTL co-localising in the same positions (the 4D width QTL  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
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

1 co-localises with a W/D QTL and the 6D W/D QTL co-localises with both a width and depth QTL  
2 at the same location). Both QTL were also found using the image descriptors utilised to train the  
3 Random Forest model, possibly explaining their identification. Two additional QTL for seminal  
4 (first order) root number were identified on chromosomes 6D and 7D from the EST-300 to EST-  
5 900 datasets. This is most likely due to the inability of the Random Forest to accurately predict  
6 the seminal number in our dataset ( $r$ -squared < 0.3, Fig. 2).

7  
8 In the majority of cases, the identified QTL had the same confidence intervals and similar peak  
9 marker positions as previously reported for all Random Forest models. Interestingly, the 4D QTL  
10 had a very similar confidence interval (position 0.8-67.6 previously reported vs 0-67.6 here), but  
11 a different peak marker position (position 4.8 previously reported vs position 30-34 here). It was  
12 also noted that lateral root QTL found on 7D had a reduced confidence interval compared to those  
13 previously reported (positions 0-101.8 previously vs 0-62.4 here).

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

1 **Table 1: Results from the QTL comparison for the different estimator datasets:** Green is a  
 2 correct identification compared to results obtained using the RootNav pipeline, Red is a miss,  
 3 yellow is a false positive and grey is not comparable. Numbers represent the significant LOD  
 4 (logarithm of odds) score for each detected QTL generated by R/qrtl [19]. Chr: chromosome, GT:  
 5 ground truth, ID: image descriptors derived from RIA-J, EST-100:900: random forest estimators  
 6 derived from 100 – 900 images.  
 7

	Trait	ID	EST 100	EST 200	EST 300	EST 400	EST5 00	EST 600	EST 700	EST 800	EST 900	GT
4D	Width	2.5			2.7							
	W/D	2.71	2.6	2.2	2.6	2.2	2.2	2.5	2.3	2.9	2.5	2.7
6D	Seminal number				3.6	3.1	4.7	3.3	3.8	3.3	3.1	
	Total root length	17	13.6	15.0	14.4	15.2	14.2	16.0	14.7	16.3	15.3	24
	Mean sem. length		13.4	13.7	13.8	13.8	13.8	14.0	13.9	13.5	15.6	22.2
	Lateral number		12.6	19.0	18.2	17.6	18.5	17.0	17.6	16.7	15.4	9.1
	Tot lateral length		11.2	13.0	14.2	12.0	15.3	12.6	13.3	12.2	11.7	6.4
	Tot seminal length		13.8	13.1	14.1	14.8	13.7	15.2	14.7	14.7	14.4	25.6
	Width	13.5	11.9	13.0	14.8	12.9	12.8	13.1	12.5	12.5	12.5	6.4
	Depth	13.6	14.3	14.3	14.1	14.0	15.6	15.0	15.2	15.8	14.8	22.7
	W/D						2.2			1.9		
7A	Seminal number											2.1
7D	Lateral number		4.3	5.5	5.9	6.6	5.2	5.0	5.3	5.0	4.4	2.4
	Seminal number					3.8		3.4	3.4	4.0	4.5	
	Tot lateral length		4.4	4.0	6.0	4.6	4.9	4.2	4.9	4.2	4.0	2
	Tot root length	4.1		2.7	2.5	2.9		3.1	2.9	4.7	3.3	9
	Tot seminal length				2.9	2.7	2.1	2.8	3.1	2.8	3.2	9.7
	<b>SUM</b>	<b>6</b>	<b>10</b>	<b>11</b>	<b>14</b>	<b>14</b>	<b>13</b>	<b>14</b>	<b>14</b>	<b>15</b>	<b>14</b>	<b>13</b>



# 1 Combining semi-automated analysis and machine learning 2 techniques increase the throughput of our image analysis pipeline

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

4 Extracting meaningful information from images of root systems is a subjective, tedious and often  
5 time-consuming process. As a general rule, automated techniques can only extract a limited  
6 amount of biologically relevant metrics and are often limited to young plants. Semi-automated  
7 tools are able to extract more metrics and with a greater accuracy, but at the expense of user  
8 interaction time (which makes them unsuited for large-scale genetic studies). As a result, large  
9 genetic screens targeting root system traits often focus on a set of simple traits that can be  
10 automatically extracted.

11  
12 Here we have shown that machine learning techniques can be used to automatically extract a  
13 large set of root system metrics. To train the machine learning algorithm on our dataset, we  
14 estimated that 600 root images are needed. Additional images are needed to validate the  
15 accuracy of the machine learning estimators (around 100). These images have to be traced with  
16 a semi-automated tool to extract the parameters in the first place. Thus, instead of tracing all the  
17 images (in our case about 2600), only a subset (700) was needed. It was previously estimated  
18 that tracing one image takes, on average, 2 minutes. In our case, the whole dataset would  
19 represent a workload of 87 hours. With the combined pipeline, the workload decreased to 23  
20 hours (27%).

21  
22 In this example, we used a published dataset, for which the ground-truth data were already  
23 available [5]. In order to easily apply this approach to future studies, we have created the R  
24 application PRIMAL (Pipeline of Root Image analysis using MAchine Learning [20] (Fig. 3). We  
25 recommend the following analysis strategy:

- 26  
27 1. Use a fully automated tool to extract image descriptors for the entire dataset.
- 28  
29 2. Use a semi-automated tool to extract the ground-truth for 200 random images (the  
GROUND-TRUTH DATASET). Remove these images from the global dataset.
- 30  
31 3. Use PRIMAL to train the Random Forest model and analyse the data.
- 32  
33 4. Check the accuracy of the prediction of the Random Forest model. If the prediction is not  
satisfactory, increase the number of images in the ground-truth dataset and repeat the  
procedure. The final number of images in the ground-truth dataset will vary between

1  
2  
3  
4 1 experiments, plants, conditions, etc. In our example, 500 images were needed to reach  
5 a satisfactory accuracy for the Random Forest predictions.  
6 2  
7 3  
8

9 4 A detailed version of this protocol is available from protocols.io [21].  
10 5

11 6 It should be noted that the prediction accuracy of the Random Forest estimation is highly  
12 7 dependent on the homogeneity of the data. For example, a Random Forest model trained on  
13 8 maize root systems will most likely fail when applied to wheat. However, for large scale genetic  
14 9 studies, where only one species is used in the analysis, this should not be an issue. The accuracy  
15 10 of the Random Forest estimators is also function of the variability of the direct descriptors in the  
16 11 dataset. Using a large set of descriptors, that better discriminate the different images, might help  
17 12 increase the accuracy of the Random Forest descriptors.  
18  
19  
20  
21  
22  
23  
24  
25

## 26 13 Conclusions

27 14 Genetic studies on root architecture require large annotated datasets of biologically-relevant  
28 15 traits. Automated analysis tools can be used to extract descriptors from large libraries of root  
29 16 images. Unfortunately, these descriptors are prone to error and their biological relevancy is not  
30 17 always clear. Alternatively, semi-automated tools enable the retrieval of more precise architectural  
31 18 traits but, due to the requirement for skilled user inputs, they are often unsuitable for large  
32 19 datasets.  
33  
34  
35  
36  
37  
38  
39  
40

41 21 Here, we used a Random Forest model to predict architectural traits based on automatically-  
42 22 extracted image descriptors. The model was trained on a subset of the whole dataset that had  
43 23 been previously analysed using a semi-automated tool. This strategy allowed us to (i) decrease  
44 24 the time required for the analysis by 73% (compared to the semi-automated analysis of the whole  
45 25 dataset) and (ii) accurately predict meaningful architectural traits.  
46  
47  
48  
49  
50

51 27 In order to make our pipeline available to the community, we have created an application available  
52 28 at [20].  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

## 1 Methods

2 A detailed version of the protocol described here is available at protocols.io [21].

## 4 Availability of supporting source code and requirements

- 7 - **Project name:** PRIMAL, Pipeline of Root Image analysis using MACHine Learning
- 8 - **Project home page:** <https://plantmodelling.github.io/primal/>
- 9 - **Operating system(s):** Platform independent
- 10 - **Programming language:** R
- 11 - **Other requirements:** none.
- 12 - **License:** GPL

## 15 Availability of supporting data

16 The following supporting data are open and available from the *GigaScience* repository, GigaDB [22]:

- 17 1. Root system image dataset #1. Images of root systems of plants tagged with genotype information. 1665 images from [5].
- 18 2. Root system image dataset #2. Training images without genotype information. 969 images.
- 19 3. Root System Markup Language files for both image datasets.
- 20 4. Full genotype mapping information for the population, from CerealsDB:  
21 [http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/Excel/SavannahXRialto\\_map\\_data\\_May\\_2013.xls](http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/Excel/SavannahXRialto_map_data_May_2013.xls).

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

# 1 Declarations

## 2 List of abbreviations

3 - **QTL**: Quantitative Trait Locus

## 4 Competing interests

5 At the time of publication, GL was a Guest Editor for the *GigaScience* thematic series “Plant  
6 Phenomics”.

## 7 Funding

8 MN is grateful to the F.R.S.-FNRS for a doctoral grant (1.A.320.16F).  
9 DMW, JAA, and MG wish to acknowledge the support of the Biological and Biotechnology Science  
10 Research Council (BBSRC) for responsive mode and CISB awards to the Centre for Plant  
11 Integrative Biology and the European Research Council (ERC) for FUTUREROOTS project  
12 funding (FP7-294729).  
13

## 14 Authors' contributions

- 15 **Conceptualization**: GL, DMW, JAA
- 16 **Formal Analysis**: GL, JAA, MN, PEM
- 17 **Resources**: JAA, MG, MN, PEM
- 18 **Writing – Original Draft**: GL, JAA
- 19 **Writing – Review & Editing**: GL, JAA, MG, DMW, MN
- 20 **Visualization**: GL, JAA

# References

1. Den Herder G, Van Isterdael G, Beeckman T, De Smet I. The roots of a new green revolution. *Trends Plant Sci.* 2010;15:600–7.
2. Lynch JP. Roots of the Second Green Revolution. *Aust. J. Bot.* CSIRO PUBLISHING; 2007;55:493–512.
3. Lobet G, Koevoets IT, Noll M, Tocquin P, Meyer PE, Pagès L, et al. Using a structural root system model to evaluate and improve the accuracy of root image analysis pipelines. *Front. Plant Sci.* [Internet]. *Frontiers*; 2017 [cited 2017 Mar 16];8. Available from: <http://journal.frontiersin.org/article/10.3389/fpls.2017.00447/abstract>
4. Hund A, Trachsel S, Stamp P. Growth of axile and lateral roots of maize: I development of a phenotyping platform. *Plant Soil.* Kluwer Academic Publishers; 2009;325:335–49.
5. Atkinson JA, Wingen LU, Griffiths M, Pound MP, Gaju O, Foulkes MJ, et al. Phenotyping pipeline reveals major seedling root growth QTL in hexaploid wheat. *J. Exp. Bot. Soc Experiment Biol*; 2015;66:2283–92.
6. Gioia T, Galinski A, Lenz H, Müller C, Lentz J, Heinz K, et al. *GrowScreen-PaGe*, a non-invasive, high-throughput phenotyping system based on germination paper to quantify crop phenotypic diversity and plasticity of root traits under varying nutrient supply. *Funct. Plant Biol.* [Internet]. CSIRO PUBLISHING; 2016 [cited 2016 Nov 17]; Available from: <http://www.publish.csiro.au/sci-hub/cc/fp/FP16128>
7. Pound MP, French AP, Atkinson J, Wells DM, Bennett MJ, Pridmore T. RootNav: Navigating images of complex root architectures. 2013;162:1802–14.
8. Lobet G, Pagès L, Draye X. A novel image-analysis toolbox enabling quantitative analysis of root system architecture. *Plant Physiol. American Society of Plant Biologists*; 2011;157:29–39.
9. Leitner D, Felderer B, Vontobel P, Schnepf A. Recovering root system traits using image analysis exemplified by two-dimensional neutron radiography images of lupine. *Plant Physiol.* 2014;164:24–35.
10. Cai J, Zeng Z, Connor JN, Huang CY, Melino V, Kumar P, et al. RootGraph: a graphic optimization tool for automated image analysis of plant roots. *J. Exp. Bot. Soc Experiment Biol*; 2015;66:6551–62.
11. Minervini M, Scharr H, Tsafaris SA. Image Analysis: The New Bottleneck in Plant Phenotyping [Applications Corner]. *IEEE Signal Process. Mag.* 2015;32:126–31.
12. Ma C, Zhang HH, Wang X. Machine learning for Big Data analytics in plants. *Trends Plant Sci.* 2014;19:798–808.
13. Ali I, Greifeneder F, Stamenkovic J, Neumann M, Notarnicola C. Review of Machine Learning Approaches for Biomass and Soil Moisture Retrievals from Remote Sensing Data. *Remote Sensing.* Multidisciplinary Digital Publishing Institute; 2015;7:16398–421.

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

1 14. Babatunde OH, Armstrong L, Leng J, Diepeveen D. A computer-based vision systems for  
2 automatic identification of plant species using kNN and genetic PCA. Jaina [Internet]. 2015;6.  
3 Available from: <http://journal.magisz.org/index.php/jai/article/view/164>

4 15. Minervini M, Abdelsamea MM, Tsafaris SA. Image-based plant phenotyping with incremental  
5 learning and active contours. Ecol. Inform. 2014/9;23:35–48.

6 16. Singh A, Ganapathysubramanian B, Singh AK, Sarkar S. Machine Learning for High-  
7 Throughput Stress Phenotyping in Plants. Trends Plant Sci. Elsevier Ltd; 2015;1–15.

8 17. Weiss U, Biber P, Laible S, Bohlmann K, Zell A. Plant Species Classification Using a 3D  
9 LIDAR Sensor and Machine Learning. 2010 Ninth International Conference on Machine Learning  
10 and Applications. 2010. p. 339–45.

11 18. Pound MP, Atkinson JA, Burgess AJ, Wilson MH, Griffiths M, Jackson AS, Bulat A,  
12 Tzimiropoulos G, Wells DM, Murchie EH, Pridmore TP, French AP. Deep Machine Learning  
13 provides state-of-the-art performance in image-based plant phenotyping. GigaScience (in press).  
14 2017.

15 19. Broman KW, Wu H, Sen S, Churchill GA. R/qtl: QTL mapping in experimental crosses.  
16 Bioinformatics. 2003;19:889–90.

17 20. PRIMAL. Pipeline of Root Image analysis using MACHine Learning.  
18 <https://plantmodelling.github.io/primal/>. Accessed 10 August 2017.

19 21. Guillaume Lobet (2017): PRIMAL: Pipeline of Root Image analysis using MACHine Learning.  
20 [protocols.io](https://protocols.io). <http://dx.doi.org/10.17504/protocols.io.h7bb9in>.

21 22. Atkinson JA, Lobet G, Noll M, Meyer PE, Griffiths M, Wells DM: Supporting data for "Combining semi-  
22 automated image analysis techniques with machine learning algorithms to accelerate large scale genetic  
23 studies." GigaScience Database. 2017. <http://dx.doi.org/10.5524/100346>

# Figure Legends

**Figure 1: Overview of the analysis pipeline used in this study.** **A.** We divided the full dataset (2614 images) into two: a training set (100 to 900 images) and a test set (1645 images). **B.** For each dataset, all the images were analysed using a semi-automated root image analysis tool (RootNav) to extract the ground-truth, as well as with a fully automated root image analysis tools (RIA-J), to extract image descriptors (see text for details). **C.** We trained a Random Forest model on the image descriptors and the ground-truth from the training dataset. **D.** We applied the Random Forest model on the image descriptors from the test dataset. **E.** We compared the image descriptors and the Random Forest estimators from the test dataset with their corresponding ground-truth. **F.** Comparison of biologically-relevant metrics extracted with the automated analysis and the Random Forest analysis. **G.** QTL were identified and compared using both Random Forest estimators and the ground-truth data.

**Figure 2: Accuracy of the Random Forest estimators.** The r-squared values of the linear regression between the Random Forest estimators and the ground-truths were computed for each size and repetition of training datasets. The dotted line represents the r-squared value between the most closely related image descriptors and the ground-truth.

**Figure 3: Screenshot of PRIMAL.** **A.** Variable to evaluate with the Random Forest algorithm. **B.** Random Forest algorithm parameters. **C.** Visualisation of the accuracy of the Random Forest estimators. **D.** Accuracy metrics for the different descriptors.







