A climate selection experiment with 517 *Arabidopsis thaliana* ecotypes

Moises Exposito-Alonso[1*], Rocío Gómez Rodríguez[2], Cristina Barragán[1], Giovanna Capovilla[1], Eunyoung Chae[1], Jane Devos[1], Ezgi Dogan[1], Claudia Friedemann[1], Caspar Gross[1], Patricia Lang[1], Derek Lundberg[1], Belén Méndez-Vigo[3], Vera Middendorf[1], Jorge Kageyama[1], Talia Karasov[1], Sonja Kersten[1], Sebastian Petersen[1], Julian Regalado[1], Lukas Reinelt[1], Beth Rowan[1], Danelle Seymour[1], Efthymia Symeonidi[1], Rebecca Schwab[1], Diep Thi Ngoc Tran[1], Kavita Venkataramani[1], Anna-Lena Van de Weyer[1], François Vasseur[1], George Wang[1], Ronja Wedegärtner[1], Frank Weiss[1], Leily Rabanni[1], Rui Wu[1], Wanyan Xi[1], Maricris Zaidem[1], Wangsheng Zhu[1], Fernando García Arenal[2], Hernán A. Burbano[1], Oliver Bossdorf[3], Detlef Weigel[1].

[1] Max Planck Institute for Developmental Biology, Tübingen, Germany

[2] Centre for Plant Biotechnology and Genomics, Technical University of Madrid, Pozuelo de Alarcón, Spain

[3] University of Tübingen, Tübingen, Germany

Abstract

To quantify phenotypic and genetic natural selection, the gold standard are evolution experiments. However, studies that include whole-genome data are still scarce. Evolution experiments can be longitudinal, as laboratory experiments continued over many generations, or cross-sectional, as field experiments replicated over many geographic locations or environments. For organisms with generations that span months or years such as *Arabidopsis thaliana*, cross-sectional studies are most feasible. Here we present an experiment carried out in a Mediterranean and a Central European field station with rainout shelters which enabled to apply a high and low rainfall treatment in each location. We planted 12 replicates per treatment combination of 517 whole-genome sequenced *A. thaliana* lines covering the global distribution. To make the data and processing code available, we created a companion R package "dryAR" (http://github.com/MoisesExpositoAlonso/dryAR). We believe this could be a useful resource for the evolutionary biology and the Arabidopsis communities.

Field experiment design

The ecotypes from the 1001 Genomes Projects

The 1001 Genomes project (1001 Genomes Consortium 2016) comprises 1135 sequenced accession lines (or ecotypes) (Fig. 1). To select the most informative, less biased lines within the 1001 genomes project, we set some quality criteria. These consisted of several filters: (1) First we aimed to remove those accessions with the lowest genome quality. We discarded those with < 10X genome coverage and $< 90\%$ congruence of SNPs called from Max Planck Institute and Gregor Mendel Institute pipelines (1001 Genomes Consortium 2016), which resulted in 959 accessions. (2) Parallely, we filtered almost-identical individuals. Using Plink (Purcell *et al.* 2007) we computed identity by state genome-wide across the 1135 accessions. For pairs of accessions with < 0.01 differences per SNP, we randomly picked one. This resulted in 889 accessions. After the (1) and (2) criteria 762 accessions remained. (3) Finally, we reduced geographic ascertainment. Sampling for the 1001 Genomes project was not performed in either a random nor regular structured scheme. Some laboratories provided several lines per locations whereas others provided lines that were at least several hundred kilometres apart. Employing latitude and longitude degrees, we computed euclidean distances across the 1135 accessions and

identified pairs that were < 0.0001 distance, that is accessions from the same population (<< 100 meters), and randomly picked one. This resulted in 682 accessions. We merged the resulting lists of accessions after the three quality filtering procedures and obtained a final set of 523 accessions. We propagated accessions in controlled conditions to generate seeds for the field experiment. A total of 517 accessions were finally planted in the field experiments (Fig. 1).

Field settings and watering

We built two 30m x 6m tunnels with similar PVC plastic foils to fully exclude rainfall in Madrid and in Tübingen (Fig. 2). The foil tunnels are different from a greenhouse in that they are completely opened in two sides, thus ambient temperatures vary as much as in any outdoor experiment. In each location, we supplied artificial watering at two contrasting regimes: abundant watering and reduced watering. Inside each tunnel, we created an approximate 10% slope and set up four flooding tables in the ground (1m x 25m). The lower elevation side of the flooding table was used to drain the water provided from the other, higher elevation, side of the table (Fig. 3).

We used trays of 8x5 cells (5.5 x 5.5 x 10cm size). One genotype was planted per cell. We grew a total of 12 replicates per genotype per treatment. Five replicates were planted at a density of 30 counted seeds per cell and were let grow without disturbance ("population replicate"). Seven were planted at low density (ca. 10 seeds) and once germinated one seedling was selected at random ("individual replicate").

We monitored temperature and watering at real time throughout the experiment using multi-purpose sensors (Parrot SA, Paris, France). This enable us to adjust watering depending on the degree of evapotranspiration over the experiment. Overall the soil water content was lower in the blocks of drought treatment compared to the high watering treatments. The sensors outside of the tunnel in Madrid (i.e. only natural rainfall) showed similar soil water content as the drought treatment we artificially imposed inside of the tunnels both in Madrid and in Tübingen. Similarly, the sensors outside of the tunnel in Tübingen showed similar water content as in the high watering treatment in both locations. Temperatures were overall higher in Madrid than in Tübingen, but did not vary between sensors inside and outside the tunnel (see Fig. 4).

Blocking and randomization

We used an incomplete block randomized designed (Fig. 3). A total of 16 blocks were established. For each watering treatment there were two intercalated blocks. Within each flooding table there were four also intercalated blocks, two of individual replicates and two of population replicates. Within each watering x replicate type x replicate number combination block, genotypes were randomly distributed in the trays (Fig. 5). The randomized designed was identical in Madrid and Tübingen.

Removal of errors during sowing and possible contaminations

In a large enterprise such as this field experiment, errors can occur. However, we tried to control such errors by reducing the "degrees of freedom" of the experimenters. We tried to accomplish this by preparing and curating Eppendorf tubes containing the seeds to be sown in cardboard boxes with the same cells as in the target trays and arranged them in their corresponding (randomised) locations. During sowing each experimenter took a box at random and went to the corresponding previously arranged tray in the field (Fig. 3). This reduced the possibility of sowing errors.

We were aware that contamination of neighboring pots was a risk. For that reason the first watering days were gentle. We also chose a day with no wind and we sow the seeds from 1-2 cm height from the soil. During the vegetative growth we identified germination that looked like neighbour contamination and remove those. Although we lost a number of plants, the power of the design was the replication, thus we discarded everything that looked suspicious. During the recording of flowering time, we used the homogeneity of flowering within a pot as a sign of contamination. When a plant had a completely different flowering stage and leaf phenotypes did not coincide with the majority of the pot, this plant was removed.

Recording of flowering time

We visited the field experiment on average every 1-2 days and manually recorded the pots that had flowered. To keep track from previous visits and avoid errors, we stuck blue pins in the pots were flowering had already been

recorded. This removed a source of human error. To calculate flowering time, we counted the number of days from the date of sowing to the assigned flowering date. Fig. 10 shows the raw flowering time data per pot in the original spatial distribution of pots as in Fig. 2. Fig. 11 shows the distribution of flowering time per treatment combination. Note that grey boxes in Fig. 10 are pots with plants that did not survive until flowering. For more visualizations of flowering time see http://github.com/MoisesExpositoAlonso/field/analyses/flowering_exploration.html.

Image production and analysis

Vegetative rosettes

Top-view images were taken every four to five days (median) with a Panasonic DMC-TZ61 digital camera and a customized closed black box (Fig. 5) at a distance of 40 cm from the tray. After trying different camera parameters, we used an exposure of -2/3 and an ISO of 100. White balance was automatically set for flashlight, the only source of illumination to keep the white balance consistent from picture to picture. Photos were saved both in .jpeg and .raw to allow for adjustments a posteriori. In total, we did this at 20 timepoints. Images for analyses are available at http://datadryad.org/updatehere and the software to process and analyse them is available at http://github.com/MoisesExpositoAlonso/hippo. Segmentation was done similarly as in Exposito-Alonso *et al.* (2017). We began by transforming images from RGB to HSV channels. We applied a hard segmentation threshold of HSV values as (H = 30-65, S=65-255, V=20-220). This was followed by several iterations of morphology transformations based on erosion and dilation. Then, for the resulting binary image we counted the number of green pixels. During field monitoring we noticed that some pots did not germinate. Sometimes this was due to lack of seeds or improper soil compaction. In those cases, we left a red mark in those pots, which we could detect in the same way as the existence of green pixels (with threshold H=150-179, S=100-255, V=100-255). Those pots were excluded from survival analyses as they did not contain any plants. An example of transformed images is in Fig. 6.

The resulting raw data consists of green and red area (pixel counts) per pots (Fig. 7A). By error, some trays were photographed two times in the same day. We took the advantage of this to verify whether our camera settings and segmentation pipeline would recover the same area, i.e. to what degree this

pipeline is replicable. In total there were 790 of such observations distributed across 11 timepoints and different trays. Using the R package lme4 1.1-12 (Bates *et al.* 2015), we computed a generalized linear mixed model with Poisson distribution (Bolker *et al.* 2009), and derived the variance proportion of pixel counts explained by the pot identity taken at the same day: $Var_{green}/Var_{tot} = 99.7\%$. As this replicability proportion was high, we summarized the mean number of green pixels for those duplicated pots.

In order to remove pots that did not germinate from the analyses, we performed a moving threshold and analysis of variance to determine what is the number of red pixels above which a pot was highly likely to have a red mark. As expected, the distribution of pixels is bimodal (Fig. 7b), what makes this process straightforward and reliable. This filtering led to a dataset of 434240 pixel observations (from an original of 474100 observations).

Then we aimed to quantify germination timing. We did this by modeling the growth trajectory of green pixels per pot as a sigmoidal curve fitting the function:

$$y = \frac{a}{1 + e^{-(b \times (x-c))}}$$

, starting in the sowing day and until the observed peak of green pixels per pot. The three parameters a, b, and c, would inform about the different shapes of growth curves. We also computed less complex indicators of growth: an analogous linear model that was used to determine the intersection with 1000 pixels, the day that over 1000 green pixels were observed ($\sim$ 1cm^2 as pixels resolution is [need to update]), the day that a fitted spline passed over 1000 green pixels, and a total count of green and red pixels through all timepoints. In Fig. 8, we show a total of 1000 example growth trajectories modeled with the sigmoidal approach and in Fig. 9 we show the estimated number of days from sowing till germination based on the spline approach. The final dataset comprised 22779 observations with germination day from low complex indicators, and 12636 for which a sigmoidal curve could be fitted. A detailed R markdown document of data loading and cleaning can be found at http://github.com/MoisesExpositoAlonso/field/data-cleaning/gen_vegetative.html.

Reproductive plants

Once flowering plants finished the reproductive stage (dry fruits were observed), we harvested them and took a final studio photograph of the rosette

and the inflorescence (Fig. 12). The settings were identical as for the vegetative monitoring but here we included a 18% grey card approximately in the same location in case adjustments a posteriori were needed. The python module to analyse the inflorescence pictures (Fig. 12a) is available at http://github.com/MoisesExpositoAlonso/hitfruit. We first used a cycle of morphological transformations of erode and dilate to produce the segmented image (Fig. 12b). This generated a segmented white/black image without white noise. Then, we used the thin (erode cycles) algorithm from Mahotas module to generate a binary pictures reduced to single-pixel paths — a process called skeletonisation (Fig. 12c). Finally, to detect the branching points in the skeletonised image we used a hit or miss algorithm from Mahotas. We used customized structural elements to maximize the branch (Fig. 12e) and end point detection (Fig. 12d). This resulted in four variables per image: total segmented inflorescence area, total length of the skeleton path, number of branching points and number of ending points (Fig. 12).

Because errors of labeling could have occurred, we build a machine learning model with the labeled organs and the branching points, end points, total area, and skeleton. This produced a concordance of 91.63 % between predicted and labeled. Then, the ones that were badly predicted from the machine learning model were manually re-labeled [need to update result here].

The total number of observations in the inflorescence dataset is 25126. The distribution of total inflorescence area is in Fig. 13. A detailed R markdown document of data loading and cleaning can be found at http://github.com/MoisesExpositoAlonso/field/data-cleaning/gen_harvesting.html.

Searching for labeling errors

Because we also run the same skeletonization software in rosette images, we leverage the different image patterns that rosettes and inflorescences have to identify labeling errors (i.e. mistakes of inflorescences by rosettes). To do this, we first trained a random forest model to predict the manually labeled organ by the four image variables. From this exercise, a total of 92.1% were correctly predicted from image analysis, and about 2,000 images were incorrectly predicted. This could be either because there might be ranges of organ morphology that are relatively similar, for instance we noticed that very small inflorescences and rosettes were confounded, or could be due to real misslabeling. Manually re-labeling about 500 pictures, we discovered that actually only 2.5% were misslabeled. As these were subsequently corrected, the true error

in the dataset should be lower than that value.

Prediction of number of fruits and seeds

Although the study of natural selection is fundamented in studying relative fitness, sometimes it is useful to have at least an approximation of the absolute fitness. In order to provide an approximate number of how many seeds were produced per reproductive plant, we generated two allometric relationship by manual counting of few plants and fruits. The first allometric relationship was built by manually counting the number of fruits of inflorescences of very small, intermediate, and very large size inflorescences (n=11). The correlation between the total inflorescence skeleton size (# pixels) and the carefully counted total number of fruits correlated strongly ($r = 0.97,\ p = 4 \times 10^{-7}$). We believe that the prediction of the number of fruits is appropriate for this type of data, as we had already shown the same linear relationship with 350 plants counted (Vasseur *et al.*). The second relationship was the average number of seeds per fruit. To do this, in the same samples as before, we counted all the seeds inside one fruit (n=11). We tried to sample fruits capturing all the range of fruit size variation. The mean was 28.3 seeds per fruit (standard deviation 11.2). Although this variation is notable, not even fruits at the two size extremes with extreme sizes varied in more than 2.5-fold. The two mentioned allometric relationships were used to predict, first, the number of fruits per inflorescence using the four image analysis variables, and second, the number of seeds corresponding to the number of fruits per inflorescence. Both counts were only done in individual replicates.

Author contributions

MEA conceived and designed the project. MEA carried out the experiment in Tübingen. MEA and RGR carried out the experiment in Madrid. All authors contributed to specific tasks in the experiments (see below). OB provided the field site in Tübingen and FGA provided the site in Madrid. DW funded the project. MEA carried out the analyses and wrote the manuscript. All authors commented and aproved the text. Below is a detailed description of the tasks each author contributed to.

AUTHOR	Conceived_idea	Funding	Advice	Coordination	Materials	Bulking_seeds	Seed_aliquoting	Field_setup	Pictures_plants	Sowing_Madrid	Sowing_Tuebingen	Thinning_seedlings	Field_care	Image_processing	Foil_tunnel_reparation	Fresh_harvesting_Madrid	Fresh_harvesting_Tuebingen	Flowering_monitoring	Image_processing	Data analysis/processing	Writing
Moises Exposito-Alonso	x			x	x	x	x	x	x	x	x	x	x	x	x	x	x	x			x
Rocio Gomez Rodriguez							x	x	x	x			x			x		x			
Detlef Weigel		x	x		x												x				
Xavier Picó			x																		
Hernán A Burbano			x							x											
Oliver Bossdorf			x		x																
Rebecca Schwab			x	x	x												x				
Carlos Alonso Blanco			x							x											
Fernando Garcia Arenal			x		x																
George Wang			x																		
François Vasseur			x								x										
Julian Regalado							x														
Derek Lundberg											x						x				
Ronja Wedegärtner							x	x	x		x		x				x				
Frank Weiss									x												
Belén Méndez-Vigo										x											
Danelle Seymour											x										
Beth Rowan											x				x		x				
Patricia Lang									x		x	x			x	x	x				
Jorge Kagayema											x										
Rui Wu											x				x		x				
Wanyan Xi											x										
Kavita Venkataramani											x				x	x	x				
Giovanna Capovilla												x			x		x				
Efthymia Symeonidi									x			x			x		x				
Vera Middendorf												x						x			
Anna-Lena Van de Weyer												x									
Jane Devos												x									
Diep Thi Ngoc Tran												x									
Sonja Kersten					x						x				x						
Wangsheng Zhu															x						
Maricris Zaidem															x						
Sebastian Petersen																	x				
Ezgi Dogan																	x				
Claudia Friedemann																x	x				
Talia Karasov																x					
Cristina Barragán																x					
Leily Rabbani																	x				
Caspar Gross																	x	x			
Lukas Reinelt													x					x			
Eunyoung Chae																	x				

Acknowledgements

We greatfully thank Leily Rabbani, Belen Mendez-Vigo, and Carlos Alonso-Blanco for field assistance, and Xavi Picó for experimental design advice.

References

1001 Genomes Consortium. (2016). 1,135 genomes reveal the global pattern of polymorphism in arabidopsis thaliana. *Cell*, **166**, 481–491. Retrieved from http://dx.doi.org/10.1016/j.cell.2016.05.063

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear Mixed-Effects models using lme4. *Journal of Statistical Software, Articles*, **67**, 1–48. Retrieved from https://www.jstatsoft.org/v067/i01

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. & White, J.-S.S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends Ecol. Evol.*, **24**, 127–135. Retrieved from http://dx.doi.org/10.1016/j.tree.2008.10.008

Exposito-Alonso, M., Vasseur, F., Ding, W., Wang, G., Burbano, H.A.A. & Weigel, D. (2017). Genomic basis and evolutionary potential for extreme drought adaptation in arabidopsis thaliana. *bioRxiv*, 118067. Retrieved from http://dx.doi.org/10.1101/118067

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., Bakker, P.I.W. de, Daly, M.J. & Sham, P.C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575. Retrieved from http://dx.doi.org/10.1086/519795

Vasseur, F., Exposito-Alonso, M., Ayala-Garay, O., Wang, G., Enquist, B.J., Violle, C., Ville, D. & Weigel, D. Scaling irregularities explained by local adaptation in arabidopsis thaliana. *submitted*.

List of Figures

1 Locations of *Arabidopsis thaliana* accessions used in this experiments (red), accessions included in the 1001 Genomes Project (blue), and all observations of the species in gbif.org 12
2 Aereal picture of foil tunnel setting in Madrid (a) and photo inside the foil tunnel in Tübingen (b). 13
3 Design and spatial distribution of blocks and replicates 14
4 Soil water content and temperature from the 35 sensors monitororing each experimental block and the conditions outside the tunnel. 15
5 Customized black box for image acquisition (a) and example trays (b) of the four watering x replicate type combinations: low watering and population replicate (upper-left), high watering and population replicate (upper-right), low watering and individual replicate (lower-left), high watering and individual replicate (lower-right). 16
6 Example segmentation results from raw image (a) to green (b) and red (c) pixels only. 17
7 Trajectories per pot of number of green pixels for Madrid and Tübingen (a). Distribution of the number of red pixels per pot summed over different time frames (b) and the heuristically chosen threshold to define whether the pot actually had a red label(red vertical line). 18
8 Randomly selected growth trajectories from 1000 pots 19
9 Distribution of germination times 20
10 Days from sowing to flowering in the same spatial distribution as Fig. 3 21
11 Distribution of time since sowing date until flowering 22
12 Example sakeletonisation results from raw image (a) to segmented (b), skeletonised (c), the detected branches (e) and endpoints (d). 23
13 Distribution of total inflorescence size (number of pixels) . . . 24

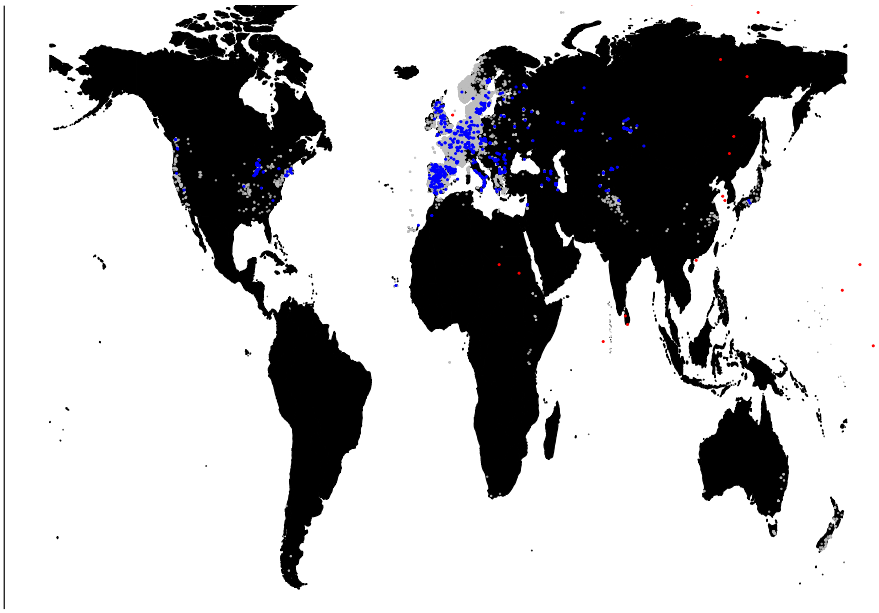


Figure 1: Locations of *Arabidopsis thaliana* accessions used in this experiments (red), accessions included in the 1001 Genomes Project (blue), and all observations of the species in gbif.org

(a) (b)

Figure 2: Aereal picture of foil tunnel setting in Madrid (a) and photo inside the foil tunnel in Tübingen (b).

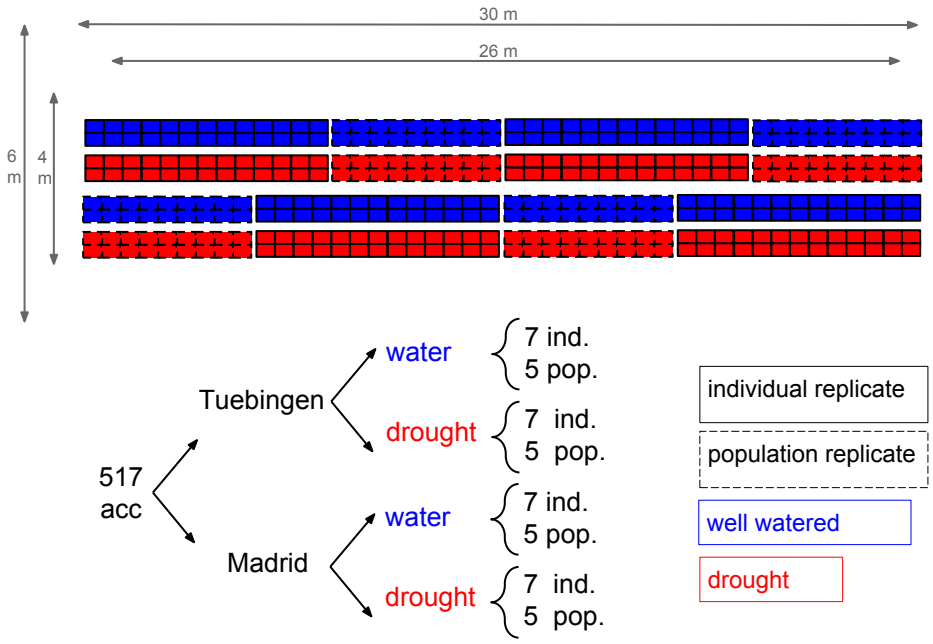


Figure 3: Design and spatial distribution of blocks and replicates

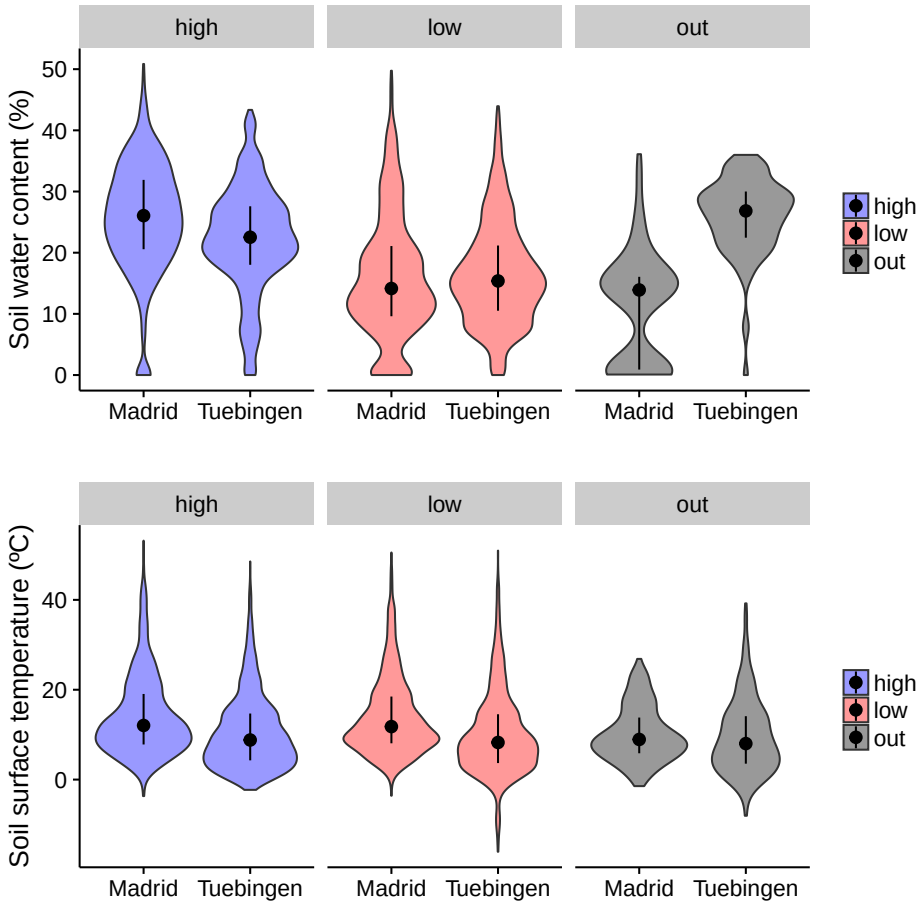
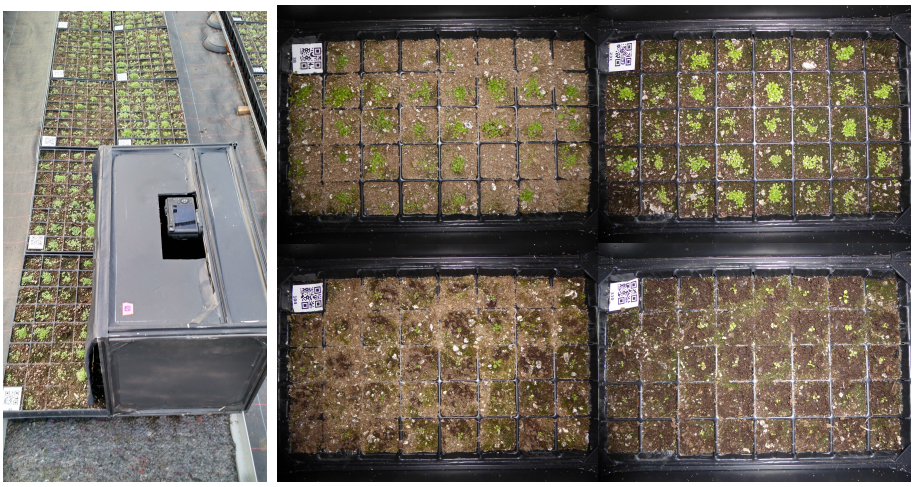
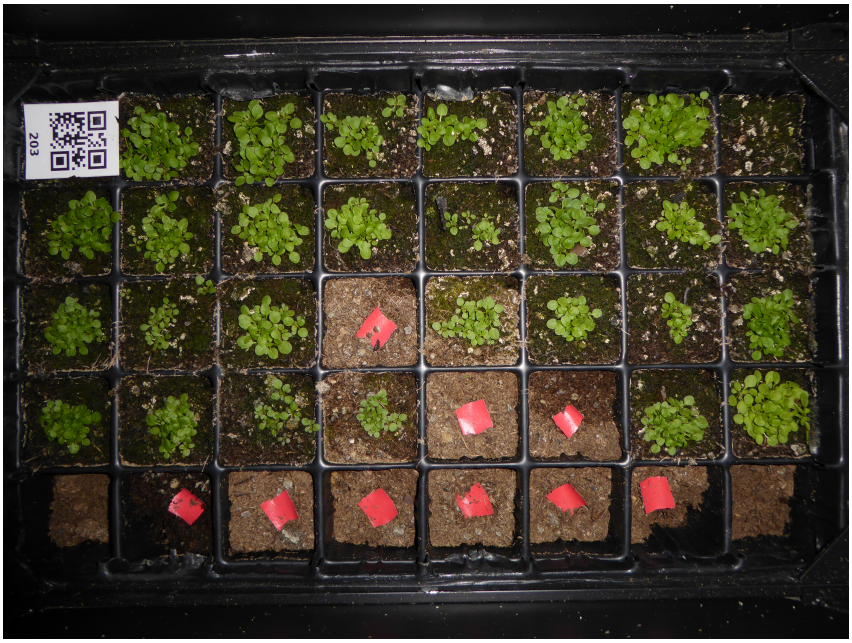


Figure 4: Soil water content and temperature from the 35 sensors monitororing each experimental block and the conditions outside the tunnel.

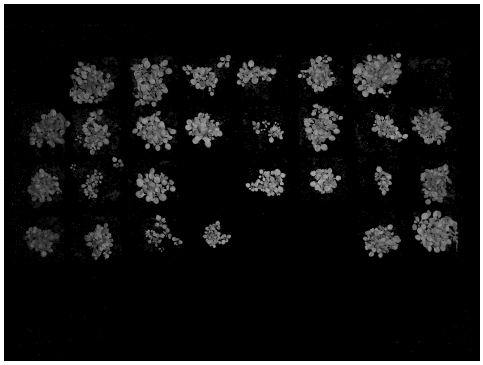


(a) (b)

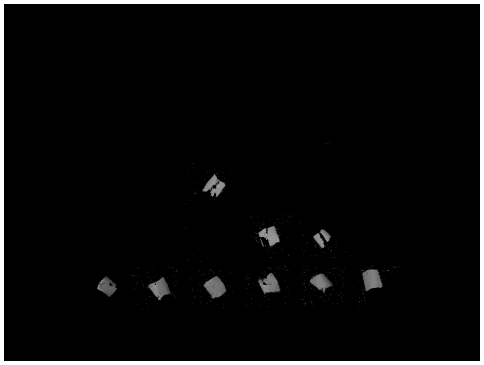
Figure 5: Customized black box for image acquisition (a) and example trays (b) of the four watering x replicate type combinations: low watering and population replicate (upper-left), high watering and population replicate (upper-right), low watering and individual replicate (lower-left), high watering and individual replicate (lower-right).



(a)

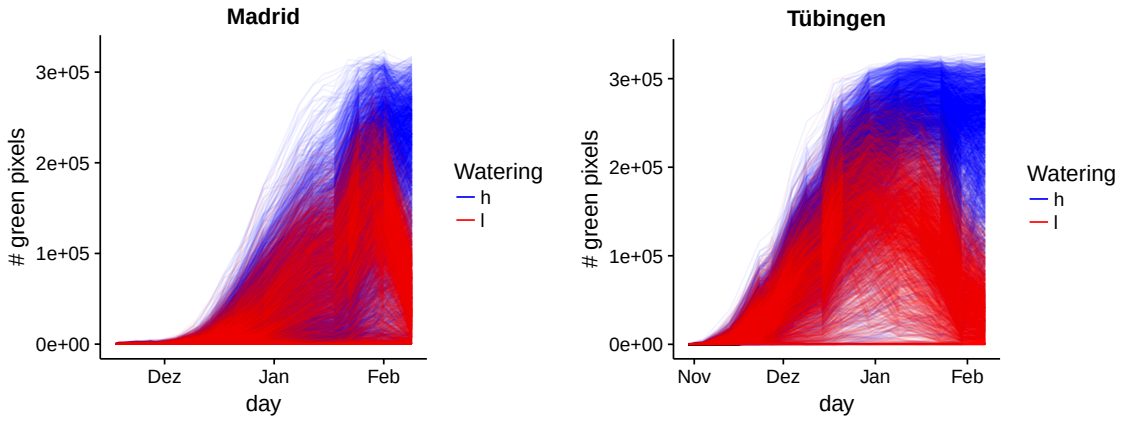


(b)

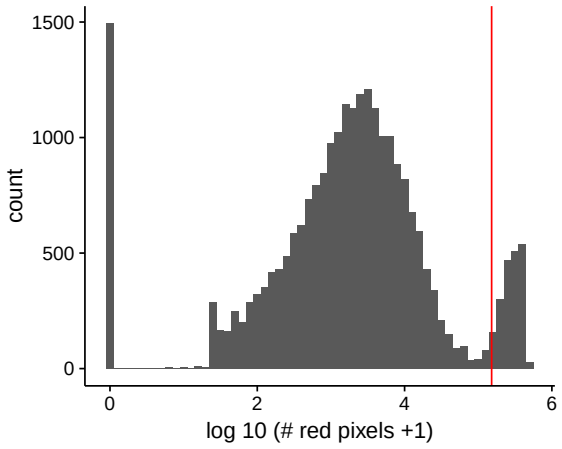


(c)

Figure 6: Example segmentation results from raw image (a) to green (b) and red (c) pixels only.



(a)



(b)

Figure 7: Trajectories per pot of number of green pixels for Madrid and Tübingen (a). Distribution of the number of red pixels per pot summed over different time frames (b) and the heuristically chosen threshold to define whether the pot actually had a red label(red vertical line).

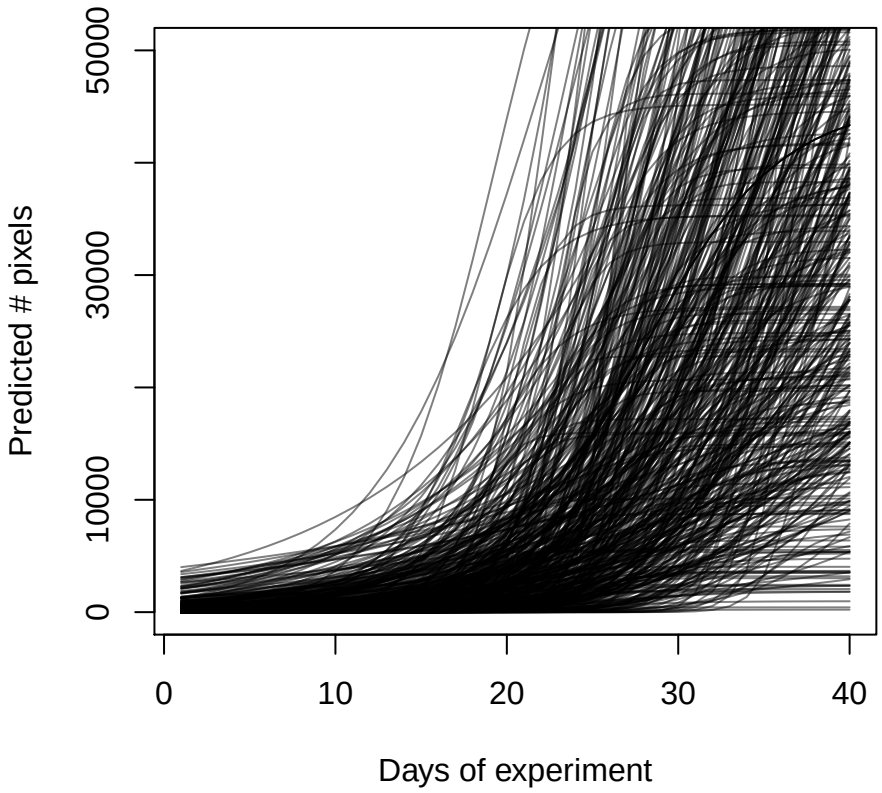


Figure 8: Randomly selected growth trajectories from 1000 pots

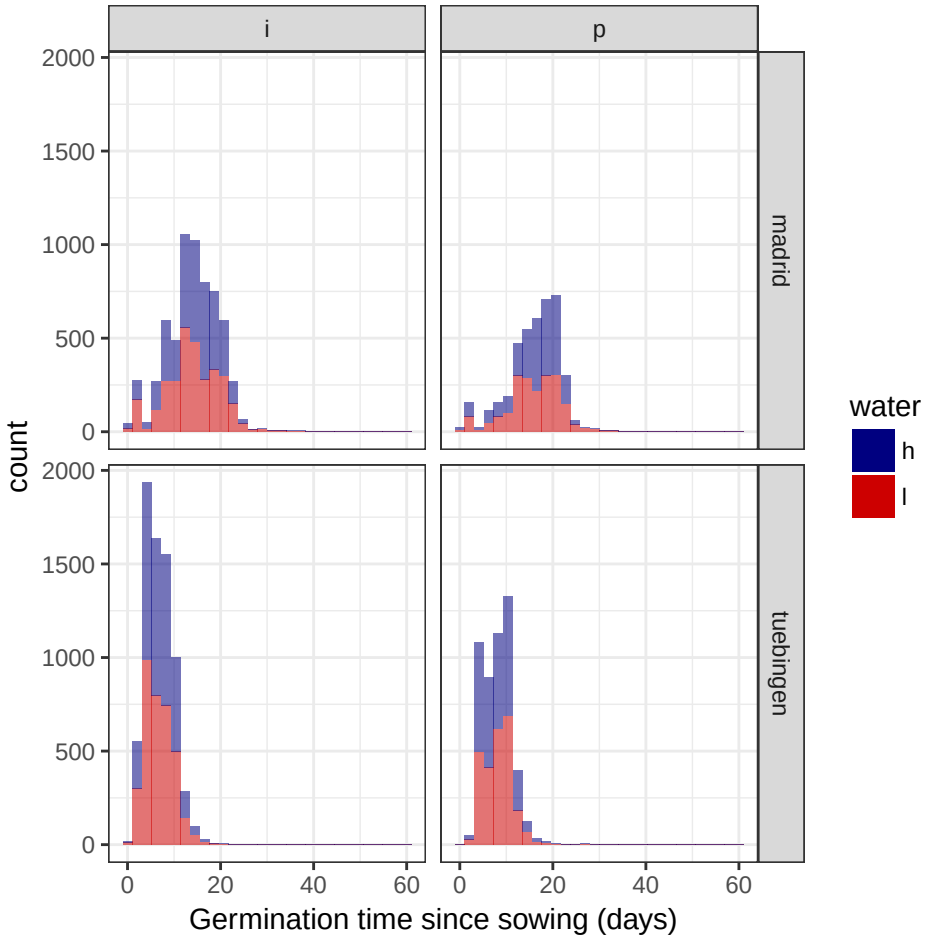


Figure 9: Distribution of germination times

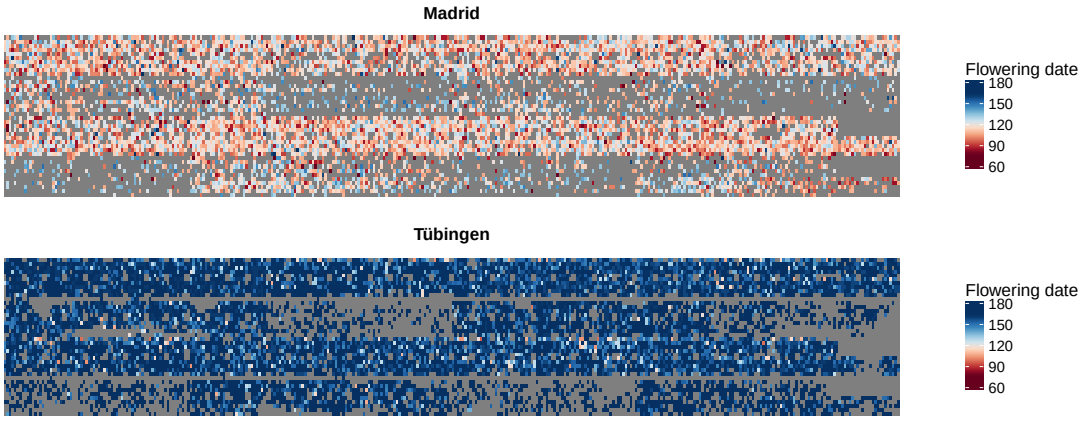


Figure 10: Days from sowing to flowering in the same spatial distribution as Fig. 3

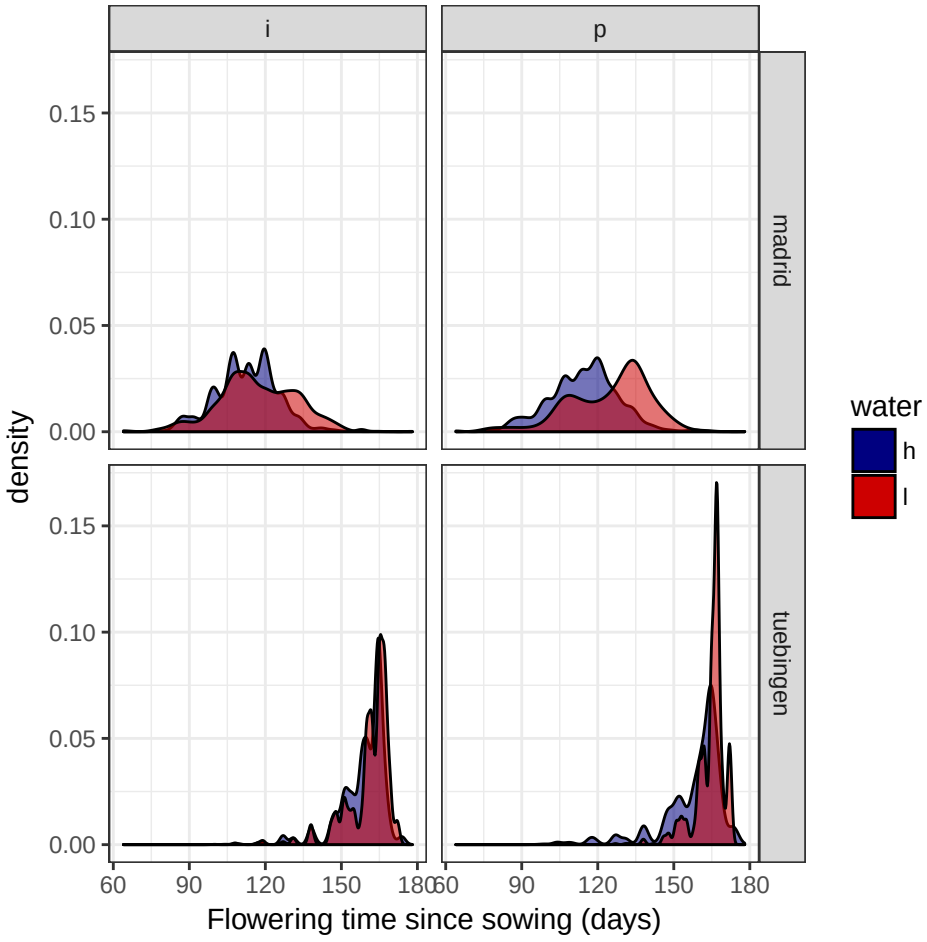
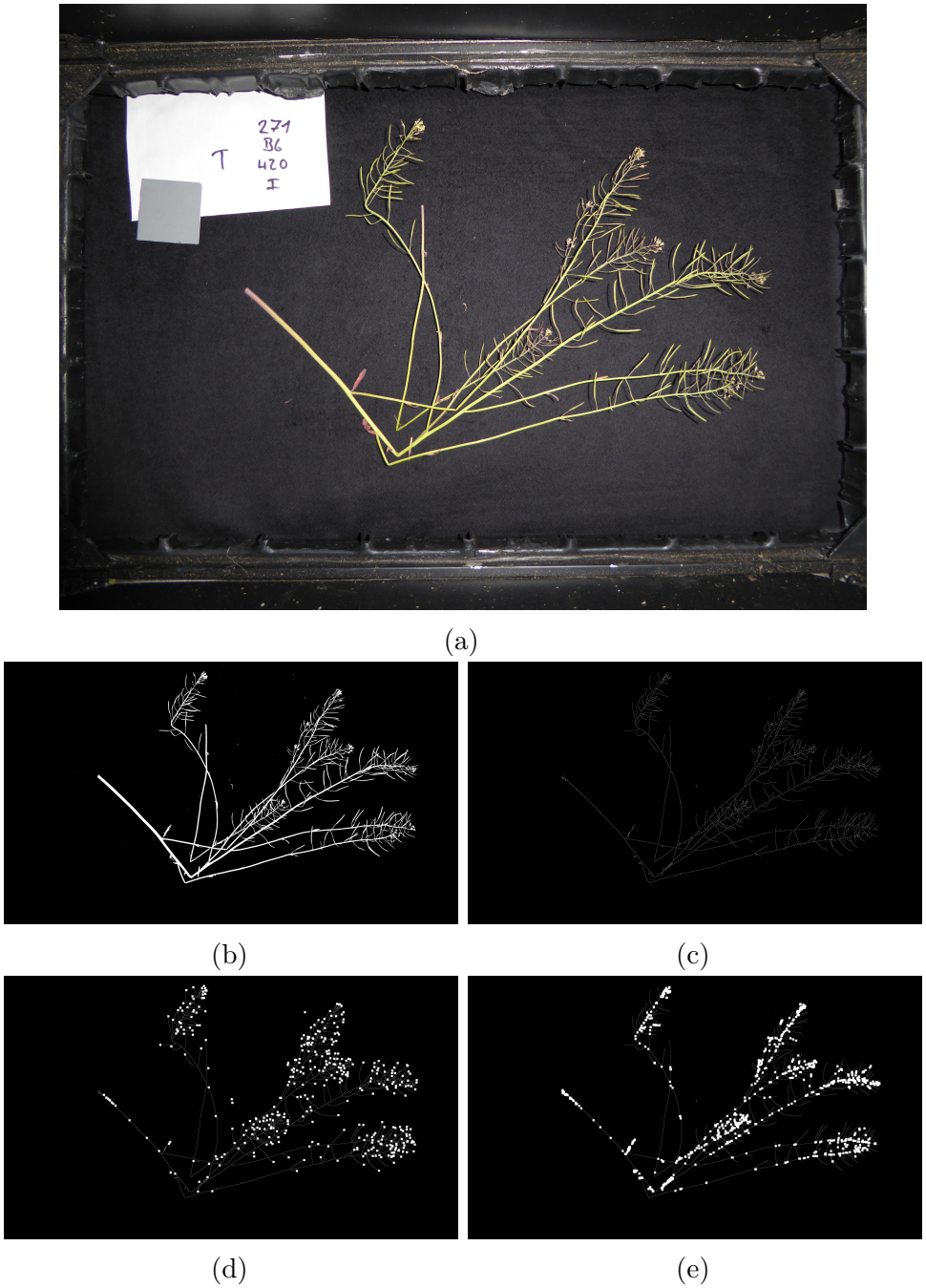


Figure 11: Distribution of time since sowing date until flowering



(a) (b) (c) (d) (e)

Figure 12: Example sakeletonisation results from raw image (a) to segmented (b), skeletonised (c), the detected branches (e) and endpoints (d).

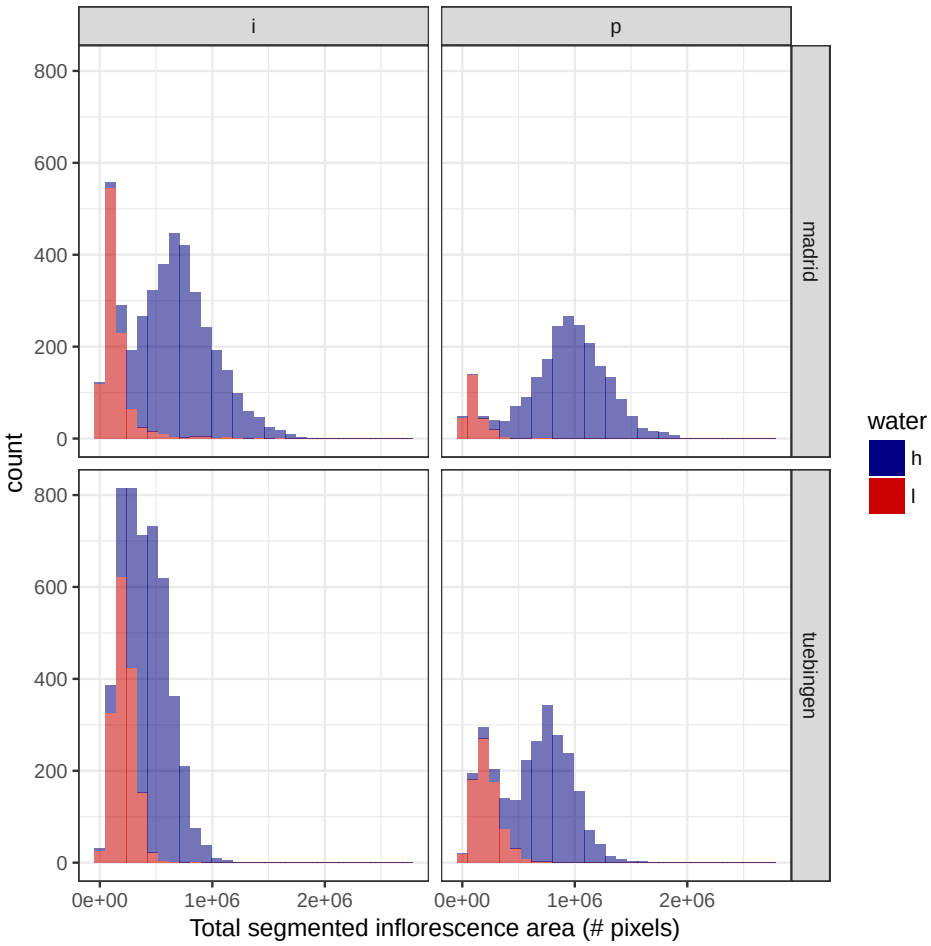


Figure 13: Distribution of total inflorescence size (number of pixels)