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], Leily Rabbani[1], Julian Regalado[1], Beth Rowan[1], Danelle Seymour[1], Efthymia Symeonidi[1], Rebecca Schwab[1], Diep Tran[1], Kavita Venkataramani[1], Anna-Lena Van de Weyer[1], George Wang[1], Ronja Wedegärtner[1], Frank Weiss[1], Rui Wu[1], Wanyan Xi[1], Maricris Zaidem[1], Wangsheng Zhu[1], Fernando García Arenal[2], Carlos Alonso Blanco[3], Xavier Picó[4], Hernán A. Burbano[1], Oliver Bossdorf[5], 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] National Centre of Biotechnology, Cantoblanco, Madrid, Spain

[4] Doñana Biological Station, Sevilla, Spain

[5] 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 stations with rainout shelters so we could apply a high and low rainfall treatment in each location. In those we planted 12 replicates per treatment combination of 517 whole-genome sequenced *A. thaliana* lines covering the global distribution. We included the raw data and processing code in an R package "dryAR" available at 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 project (1001 Genomes Consortium 2016) comprises 1135 accession lines (or ecotypes) sequenced (Fig. 1). Here are the details of the protocol used to select the most informative, less biased sample of the lines within the 1001 genomes project in order to prioritize phenotypic efforts. Filtering consisted in several approaches: (1) First we aimed to remove those accessions with the lowest genome quality. Among 1135, 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). The remaining number were 959 accessions. (2) Parallely, we filtered almost-identical individuals. Using Plink we computed identity by state genome-wide across the 1135 accessions. For those pairs of accessions with < 0.01 differences per SNP, we randomly picked one. This resulted in 889 accessions. The merge between (1) and (2) criteria was 762. (3) Finally, we reduced geographic ascertainment. Sampling for the 1001g 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 reproduced accessions in controlled conditions to generate enough seeds for the field experiment. A total of 517 accessions were finally planted in the field experiments (Fig. 1).

Field settings and watering

We build 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 any outdoor experiment [include environmental information]. 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 2). The soil from trays from the two treatments clearly showed a consistent difference in water content [insert environment].

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").

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 blocks, which were intercalated. Within each flooding table there were four blocks, two of individual replicates and two of population replicates; also intercalated. Within each watering x replicate type x replicate number combination block, genotypes were randomly distributed in the trays (Fig. 4). 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 all eppendorf with 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. Then, during sowing each experimenter took a box at random and went to the corresponding tray in the field previously arranged (Fig. S2). This reduced the possibility of sowing errors, but when an incorrect sowing was done (wrong positions, badly sprayed), we recorded those and exclude them from the analyses.

We were aware that contamination of neighboring pots was a risk. For that reason the first watering days were gentle. Furthermore, we pick a day with no wind and we throw the seeds from 1-2 cm height from the soil. During the vegetative grow 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 or 2 days and recorded manually what pots had flowered. To keep track from previous visits and avoid errors, we sticked blue pins in the pots were flowering was recorded. This removed a source of human error. To calculate flowering time, we counted the number of days from the date of sowing to to the assigned flowering date. Fig. 9 shows the raw flowering time data per pot in the original spatial distribution of pots as in ref{fig:blocks}. Fig. 10 shows the distribution of flowering time per treatment combination. Note that grey boxes in Fig. 9 are pots with plants that did not survive until flowering. For more visualizations of flowering time see http://github.com/MoisesExpositoAlonso/field/data-cleaning/updatehere.

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. 4) at a distance of 40 cm from the tray. Photos were taken on average every x days from the sowing date until flowering plants impede the acquisition – a total of 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 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 contained any plants. An example of transformed images is in Figure 5.

The resulting raw data consists in green and red area (pixel counts) per pots (Fig. 6A). We took the advantage that a tray was photographed the same day several times to verify the replicability of our pipeline. In total there were 790 of such observations distributed in 11 timepoint 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 from the analyses those pots that did not germinate, we performed a moving threshold and analysis of variance to determine what is the number of red pixels above which a pot is highly likely to have a red mark. As expected, the distribution of pixels is bimodal (Fig. 6b), what makes this process straightforward and reliable. This filtering lead to a dataset of 434240 pixel observations (from an original of 474100).

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 next 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 (~ 1cm^2 as pixels resolution is ~xyz), 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. 7, we show a total of 1000 example growth trajectories modeled with the sigmoidal approach and in Fig. 8 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/updatehere.

Reproductive plants

Once flowering plants finished the reproductive stage (dry fruits were observed), we harvested them and took a final photograph of the rosette and the inflorescence (Fig. 11). The python module to analyse the inflorescence pictures (Fig. 11a) 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. 11b). This generated a segmented white/balck 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. 11c). 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. 11e) and end point detection (Fig. 11d). 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. 11).

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 [update here].

The total number of observations in the inflorescence dataset is 30564. The distribution of total inflorescence area is in Fig. 12. A detailed R markdown document of data loading and cleaning can be found at http://github.com/MoisesExpositoAlonso/field/data-cleaning/gen_harvesting.Rmd.

Environmental data

We monitored at real time temperature and watering 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 showed in Madrid similar soil water content as the drought treatment and in Tübingen similar water content as in the high watering treatment. Temperatures were overall higher in Madrid than in Tübingen but did not vary between sensors inside and outside the tunnel 13.

Author contributions

The detailed list of contributed tasks of each authors

AUTHOR	Conceived_idea	Funding	Advice	Coordination	Materials	Bulking_seeds	Seed_aliquoting	Field_setup	Pictures_plants	Sowing_Madrid	Sowing_Tuebingen	Thinning_seedlings	Fiel_care	Image_processing	Foil_tunnel_reparation	Fresh_harvesting_madrid	Fresh_harvesting_Tuebingen	Flowering_monitoring	Image_processing	Data analysis/processing	Data interpretation	Writing
Moises Exposito-Alonso	x			x	x	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					
Xavi Picó			x																			
Hernán A Burbano			x							x												
Oliver Bossdorf			x		x																	
Rebecca Schwab			x	x	x												x					
Carlos Alonso Blanco			x							x												
Fernando García Arenal			x		x																	
George Wang			x																			
François Vasseur			x								x											
Julian Regalado							x															
Derek Lundberg											x						x					
Ronja Wedegärtner							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 Tran												x										
Sonja Kersten					x						x				x							
Wansheng Zhu															x							
Maricris Zaidem															x							
Sebastian Petersen																	x					
Ezgi Dogan																	x					
Claudia Friedemann																x	x					
Talia Karasov																x						
Cristina Barragán																x						
Leily Rabani																	x					
Caspar Gross																	x	x				
Eunyoung Chae																	x					

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

List of Figures

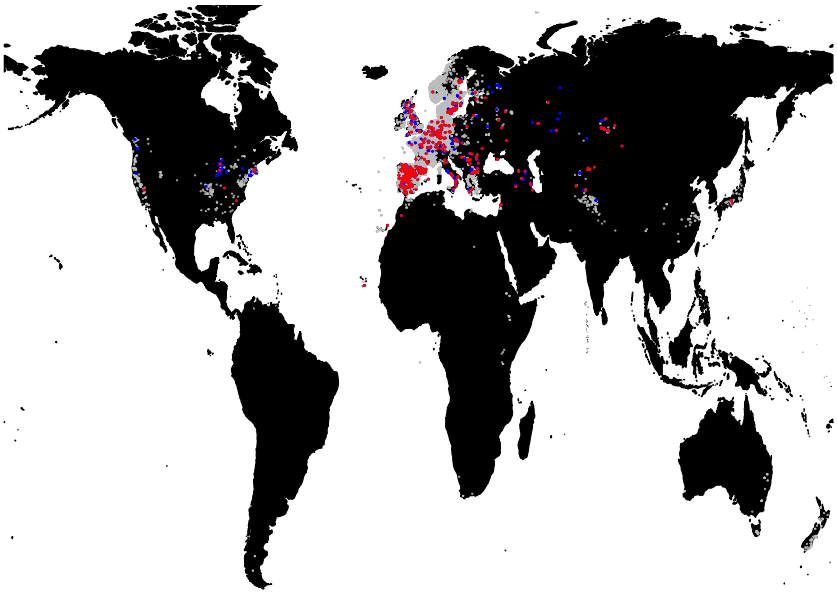


Figure 1: Locations of *Arabidopsis thaliana* accessions used in this experiments (red), accessions included in the 1001 Genomes Project (blue), and all observations of the species in gbif.org

(a) (b)

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

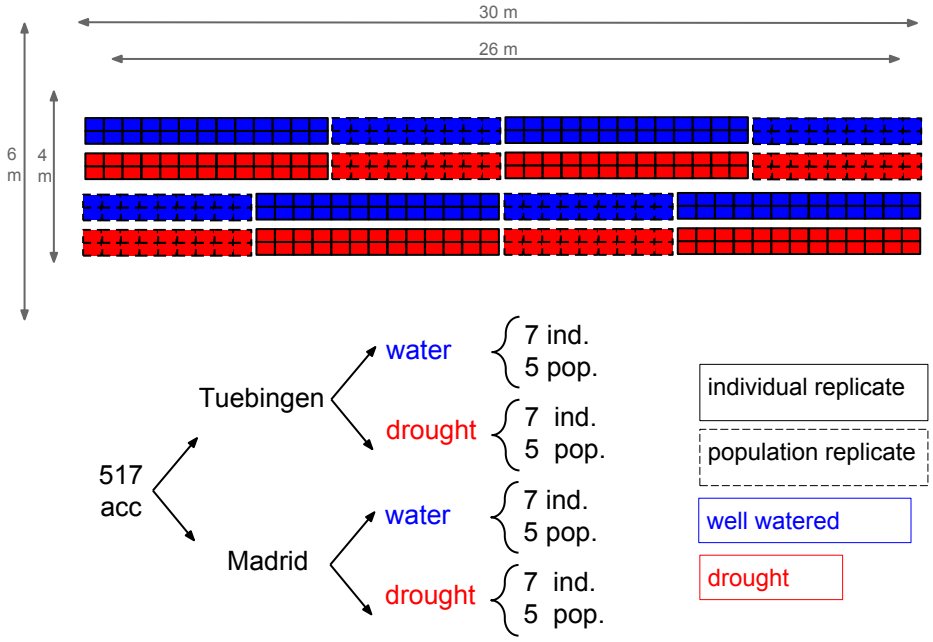
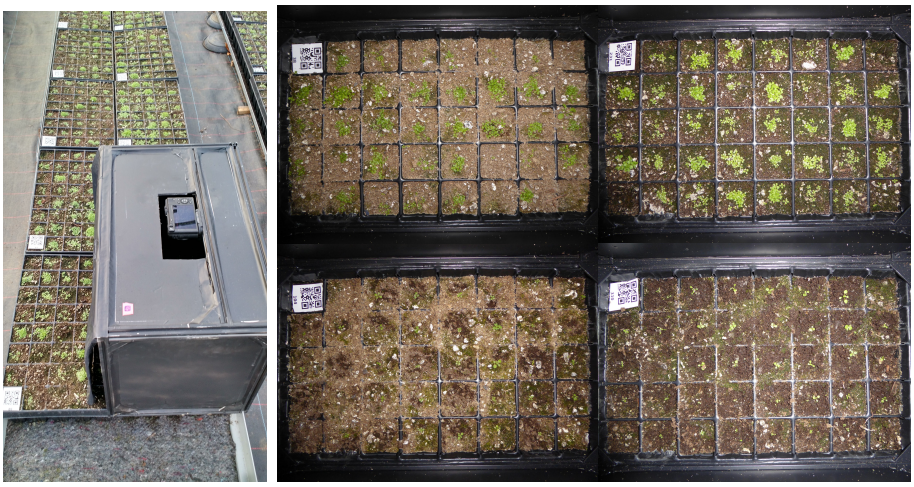
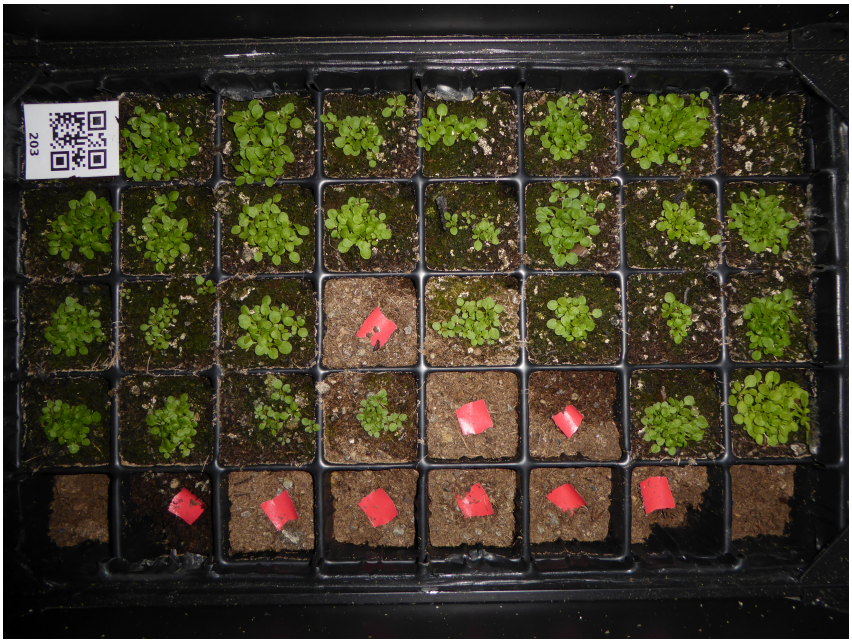


Figure 3: Design and spatial distribution of blocks and replicates

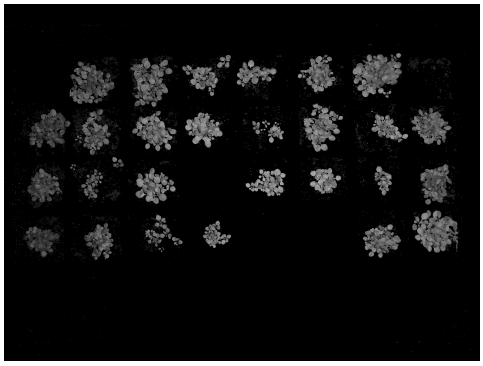


(a) (b)

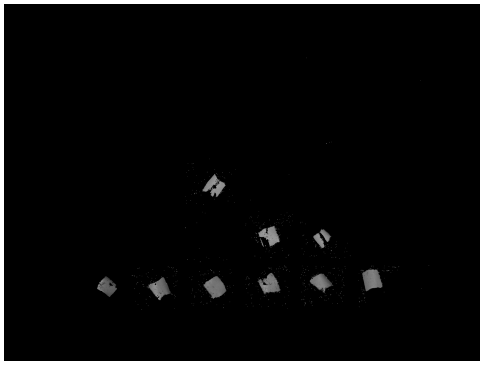
Figure 4: Customized black box for image acquisition (a) and example trays (b) of the four watering x replicate type combinations: low watering and population replicate (upper-left), high watering and population replicate (upper-right), low watering and individual replicate (lower-left), high watering and individual replicate (lower-right).



(a)

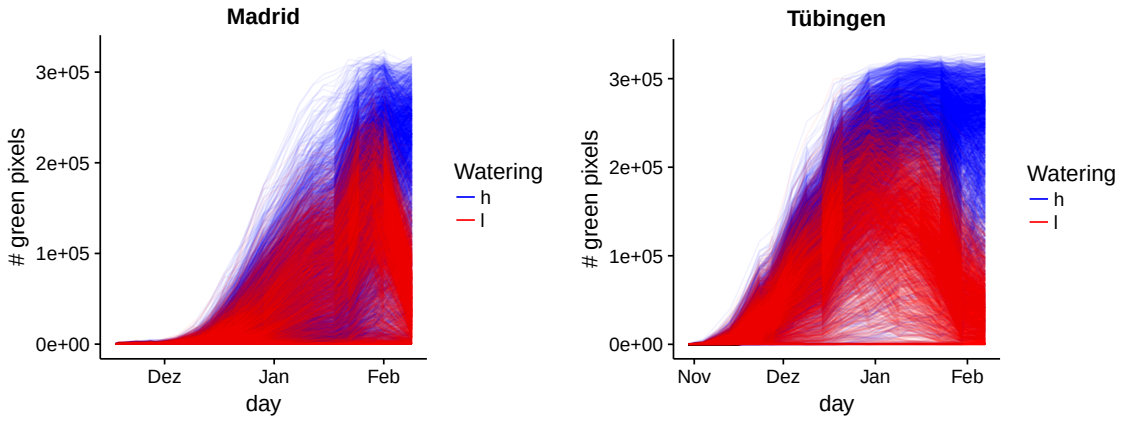


(b)

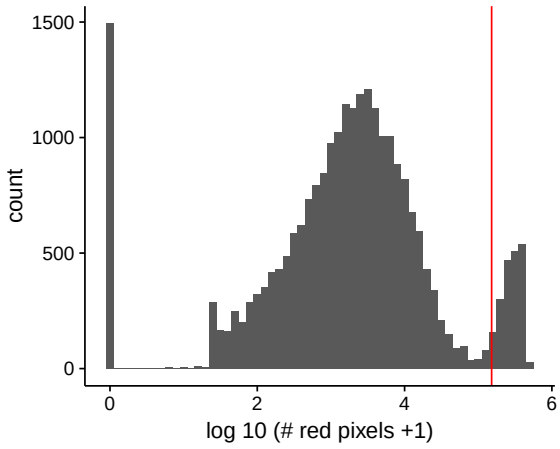


(c)

Figure 5: Example segmentation results from raw image (a) to green (b) and red (c) pixels only.



(a)



(b)

Figure 6: 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).

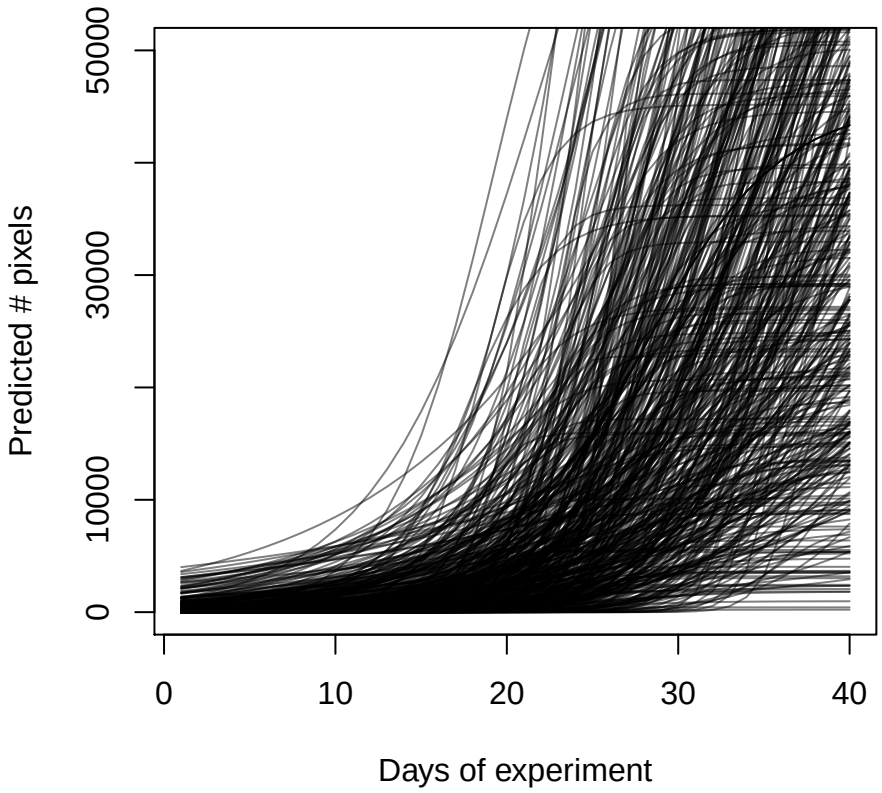


Figure 7: Randomly selected growth trajectories from 1000 pots

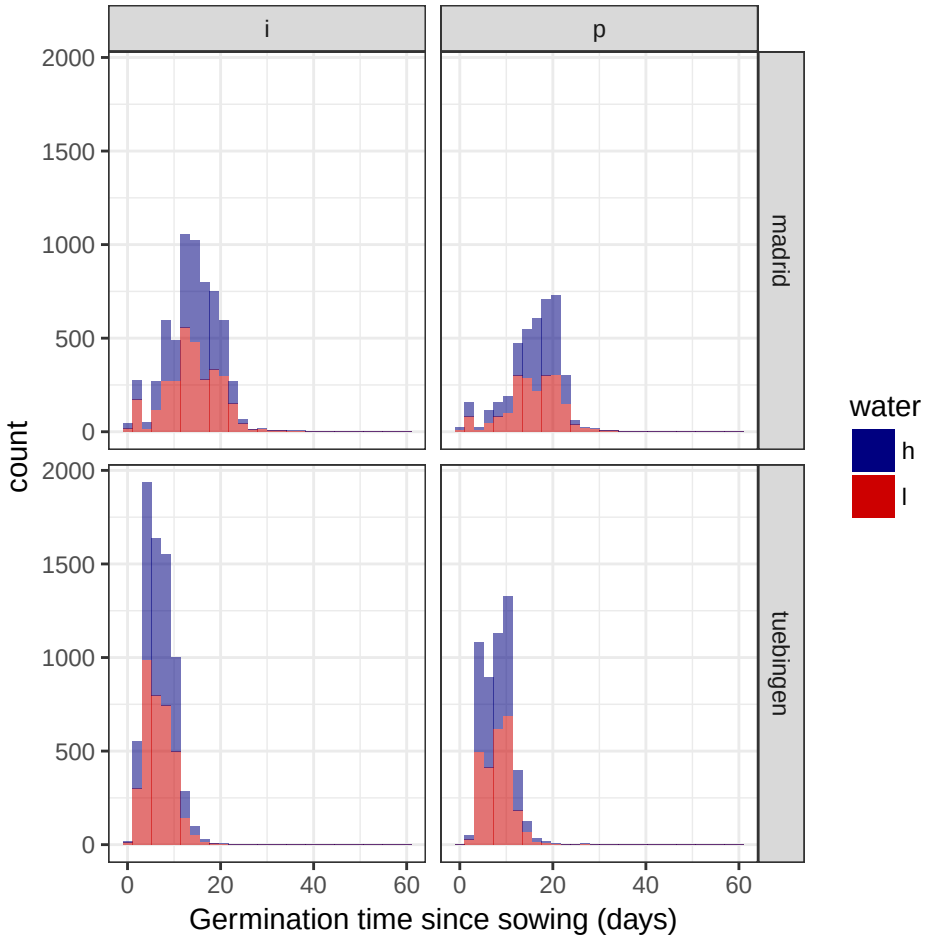


Figure 8: Distribution of germination times

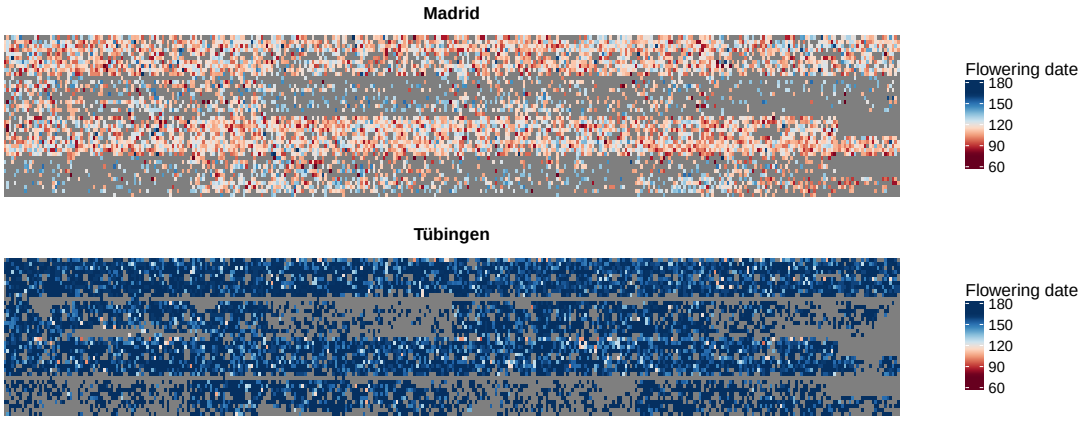


Figure 9: Days from sowing to flowering in the same spatial distribution as Fig. 3

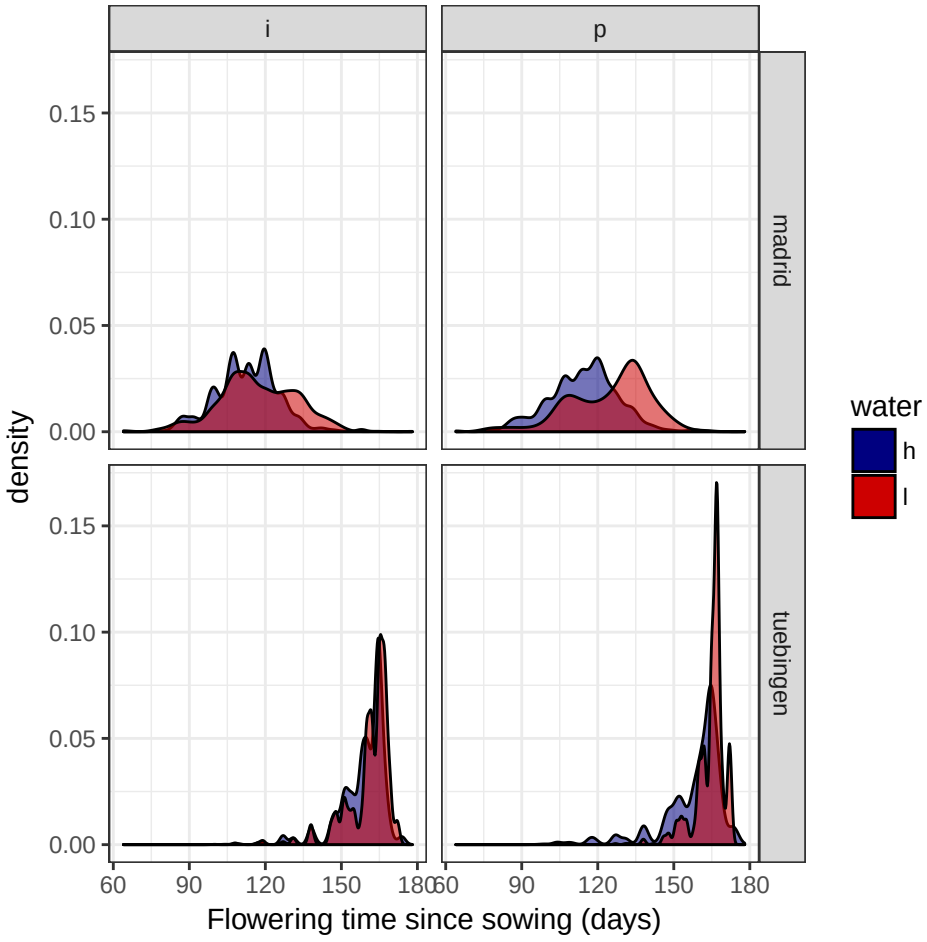
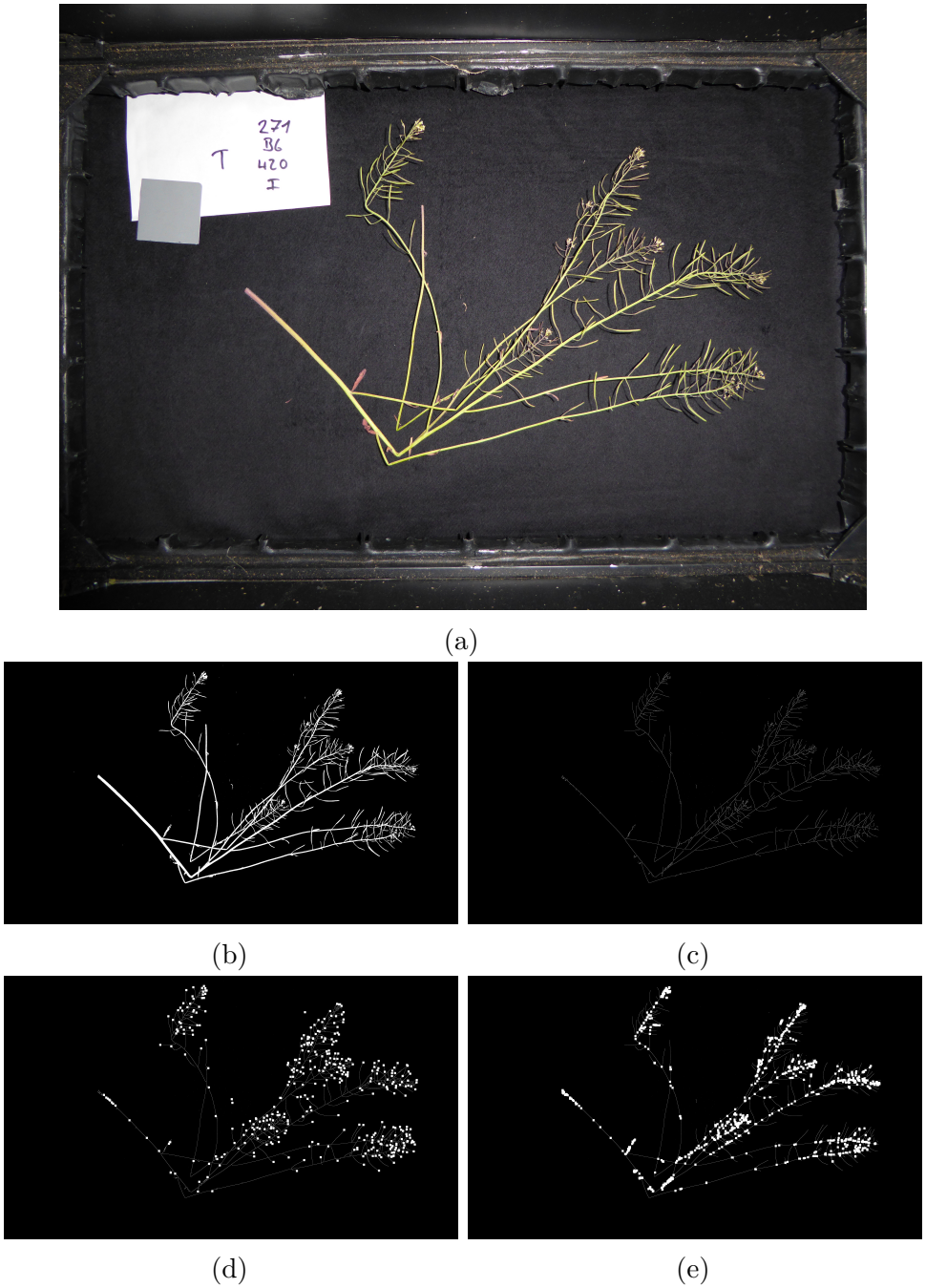


Figure 10: Distribution of time since sowing date until flowering



(a)

(b) (c)

(d) (e)

Figure 11: Example sakeletonisation results from raw image (a) to segmented (b), skeletonised (c), the detected branches (e) and endpoints (d).

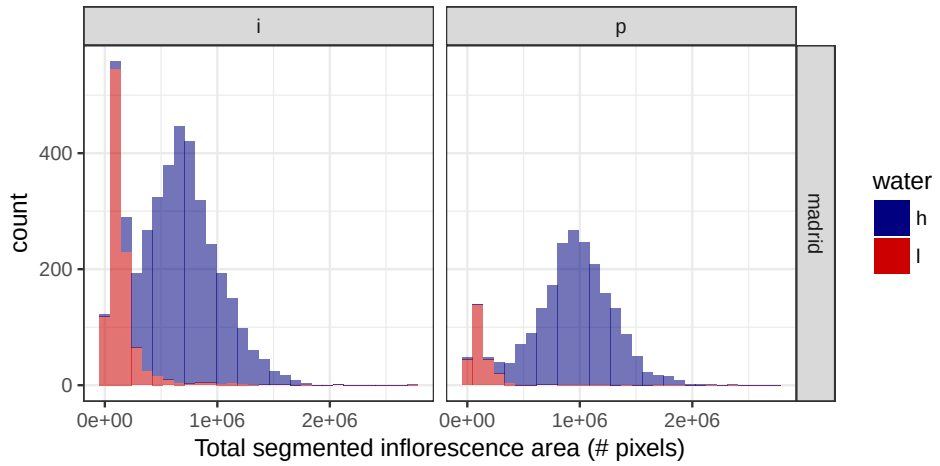


Figure 12: Distribution of total inflorescence size (number of pixels)

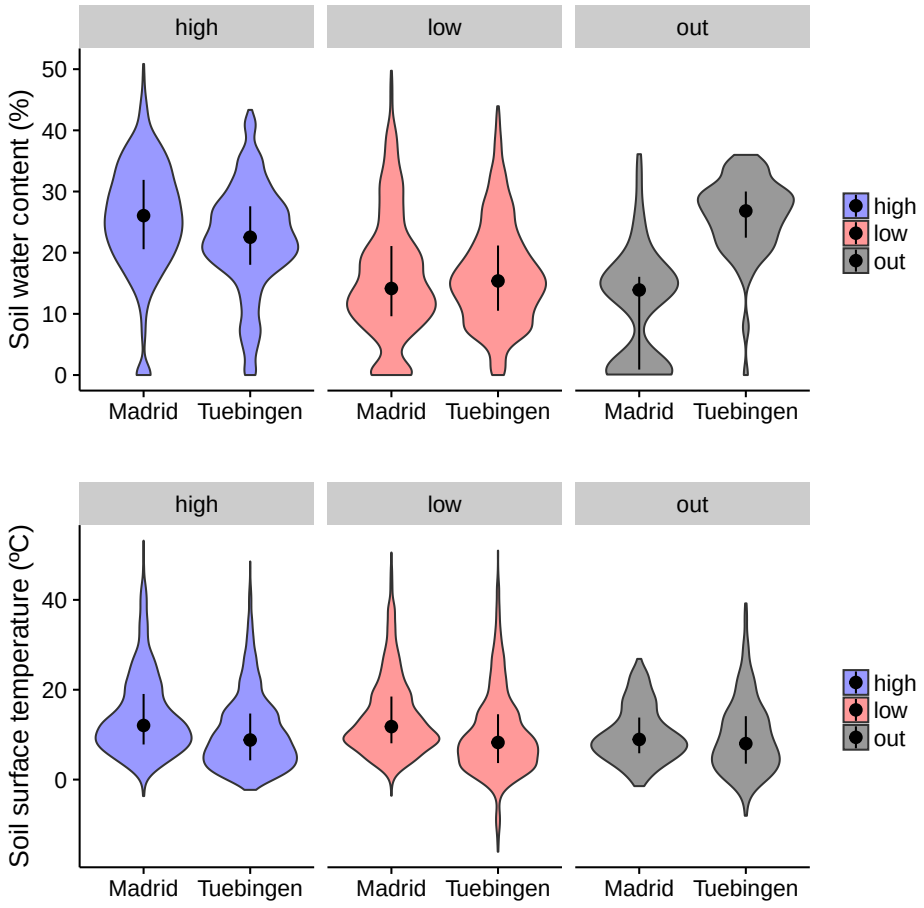


Figure 13: Soil water content and temperature from the 35 sensors monitororing each experimental block and the conditions outside the tunnel.