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], Ronja Wedegärtne[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
[*] correspondence to moisesexpositoalonso@gmail.com

Abstract

To quantify phenotypic and genetic natural selection, the gold standart are evolution experiments. However, studies that include whole-genome data are still unabundant. 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 long-lived organisms (generation time over a 1 year) such as Arabidopsis thaliana, only cross-sectional studies are feasible. Here we present an experiment carried out in a Mediterranean and a Central European field station and in which we additionally manipulated rainfall. We used 517 whole-genome sequenced A. thaliana lines from globally distributed areas. Generally, the experiment suggests that x and x are. We encapsulate the raw data and cleaning and processing code in an R package for easy data sharing. Finally, we complement the dataset with other previously published data of flowering time, root morphology, growth traits, and fitness in four other field stations. 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 neither 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 11135 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 enouch seeds for the field experiment. A total of 517 accessions produced enough seeds and were used in the experiment (Fig. 1).

Field settings and watering

We build two 30m x 6m tunnels with equivalent PVC plastic foils to fully exclude rainfall in Madrid and in Tübingen (Fig 2). The foil tunnels were different from a greenhouse in that they were completely opened in two sides. Then temperature varied 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 setup 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).

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 design was identical in Madrid and Tübingen.

Removal of errors during sowing and possible contaminations

In a large field experiment enterprise errors can occur, but these errors can be reduced 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 quick pots and arranged 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 possible errors, and those positions were detected, were removed from the analyses.

Because we were aware that contamination of neighboring pots was a risk. We were extremely careful during sowing. We pick a day with no wind and we throw the seeds from 1-2 cm height. 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 flowering recording (see section) we observed homogeneity of flowering as a trait that could further indicate contamination

Recording of flowering time

We visited the field experiment on average every 1 or 2 days and recorded what pots had flowered. To keep track from previous visits, we sticked blue pins in the pots were flowering was recorded. This removed a source of human error. To calculate flowering time, to the flowering data we substracted the date of sowing.

Image production and analysis

Vegetative rosettes

Top-view images were taken every x days 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 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 chanels. 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 sucessfuly, which can be due to lack of seeds or improper soil compaction. In those cases, we left a red mark in those pots which we can detect in the same way as the existence of green pixels (with threshold H=150-179, S=100-255, V=100-255). An example of transformed images is in Figure 7.

The resulting data consists in green and red pixel counts per pots (Fig. **??**fig:growth)). 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 a genearlized linear mixed model with Poisson distribution (Bolker *et al.* (2009)), we calculated that the proportion of pixel count variance explained by the pot identity was $Var_{green}/Var_{tot} = 99.6\%$. After that verification, wethe average pixel between the several images was calculated for those pots, ending up with a dataset of the size

In order to remove from the analyses those pots that did not germinater pot. As expected, the distribution of pixels is bimodal (Fig. 6). Therefore, the exclusion of pots is straightforward: a pot is excluded if at any photographed time it was in the second distribution (>1000 pixels). This filtering left a dataset of 30947 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, were stored together with several less complex indicators of growth: an analogous linear model, the day that over 1000 green pixels were observed (~ 1cm^2 as pixels resolution is ~xyz), and a total count of green and red pixels through all timepoints. This data columns comprise 24747 with non-missing data.

[still need to figure out problem here]

Reproductive plants

We used Otsu's adaptive thresholding algorithm from OpenCV module to convert three channel images into a white/black segmented picture. Then we utilized the thin function from Mahotas module to erodes the binary picture to a single-pixel path — called skeletonisation. Finally, to detect the branching points we used

Data visualization

Additional datasets

Vasseur et al. 2017

Fournier-Level et al. 2011

Slovak et al. 2014

1001 Genomes Consortium 2016

Author contributions

The detailed list of contributed tasks of each authors

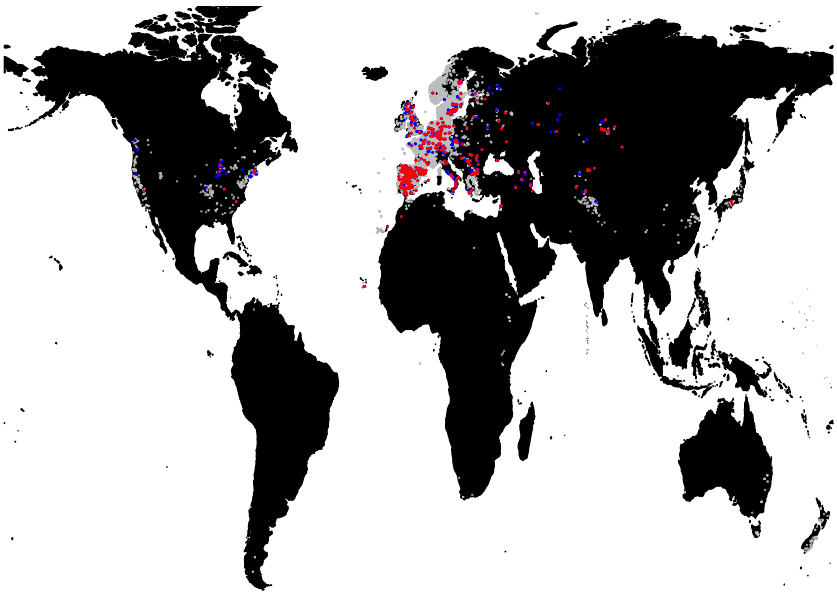


Figure 1: Locations of *Arabidopsis thaliana* accessions used in this experiments (red), accessions included in the 1001 Genomes Project (blue), and all observations of the species in gbif.org

Tables

Figures

(a) (b)

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

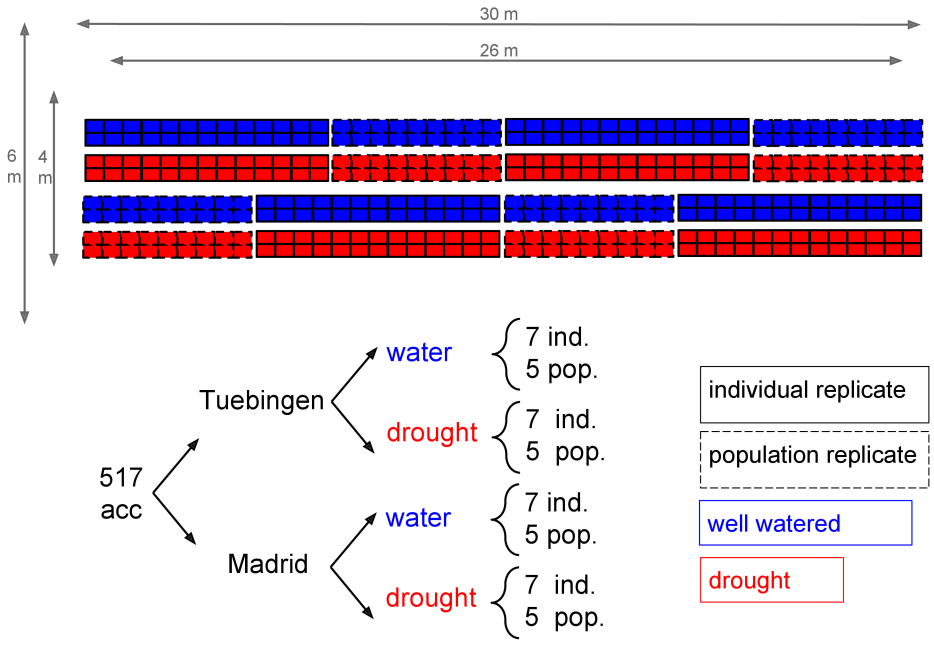
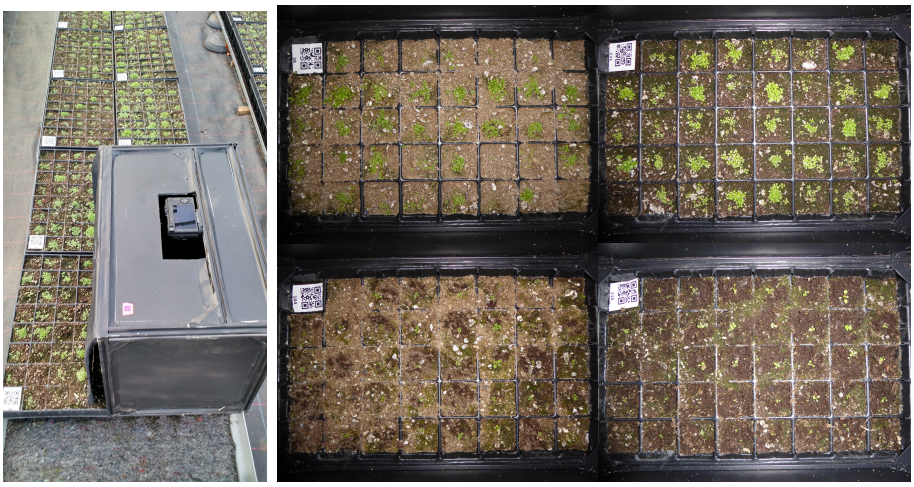


Figure 3: Design and spatial distribution of blocks and replicates



(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).

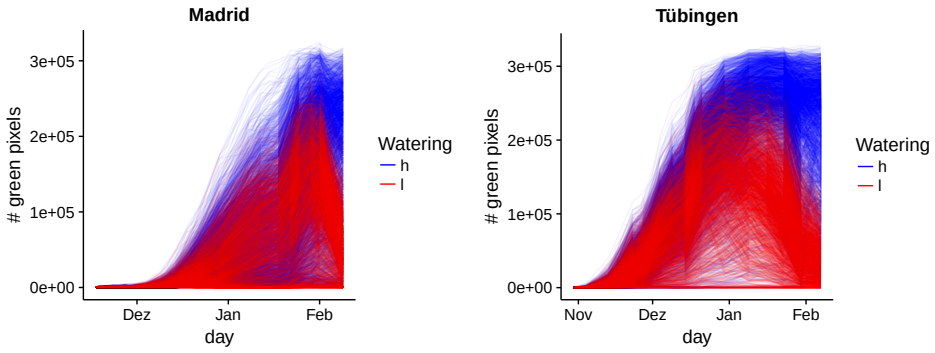


Figure 5: Trajectories per pot of number of green pixels measured from image analysis

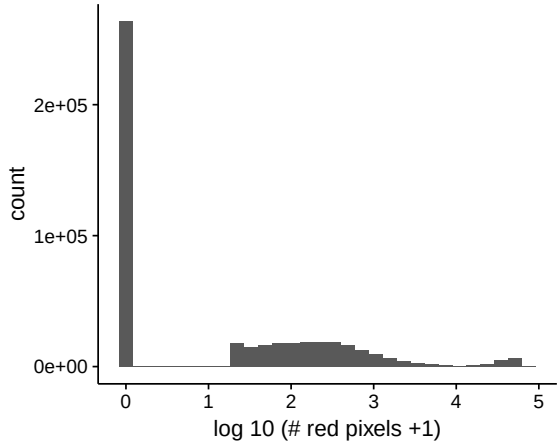


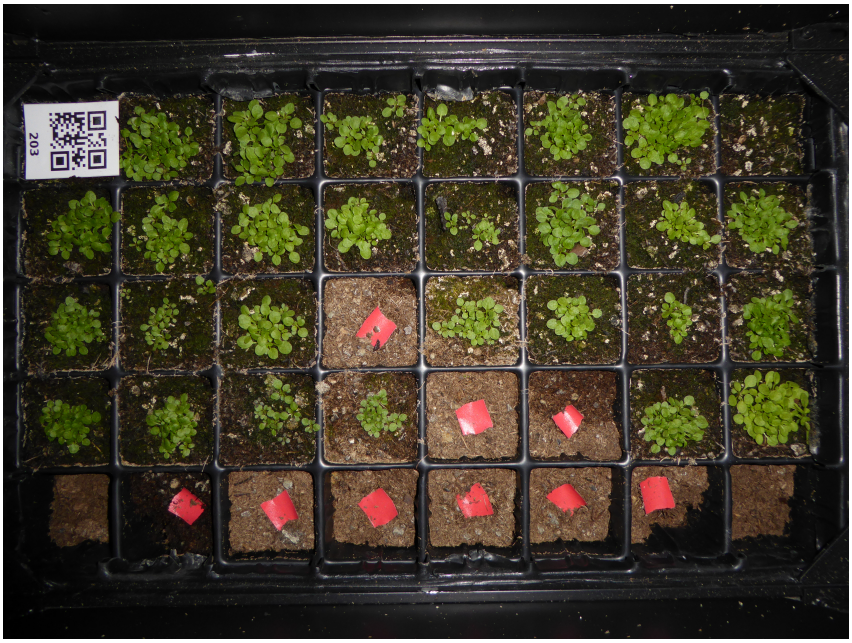
Figure 6: Distribution of total sum of red pixels

References

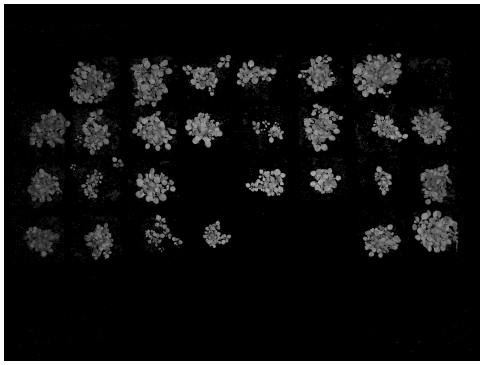
1001 Genomes Consortium. (2016). 1,135 genomes reveal the global pattern of polymorphism in arabidopsis thaliana. *Cell*, **166**, 481–491. Retrieved from http://dx.doi.org/10.1016/j.cell.2016.05.063

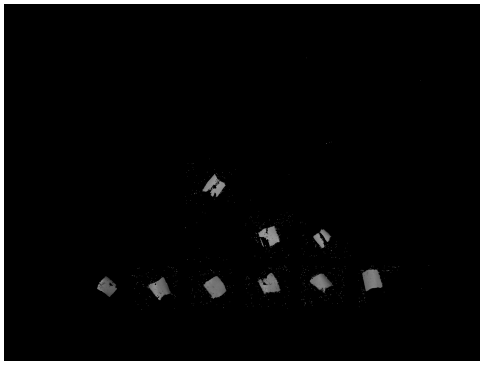
Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. & White, J.-S.S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends Ecol. Evol.*, **24**, 127–135. Retrieved from http://dx.doi.org/10.1016/j.tree.2008.10.008

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



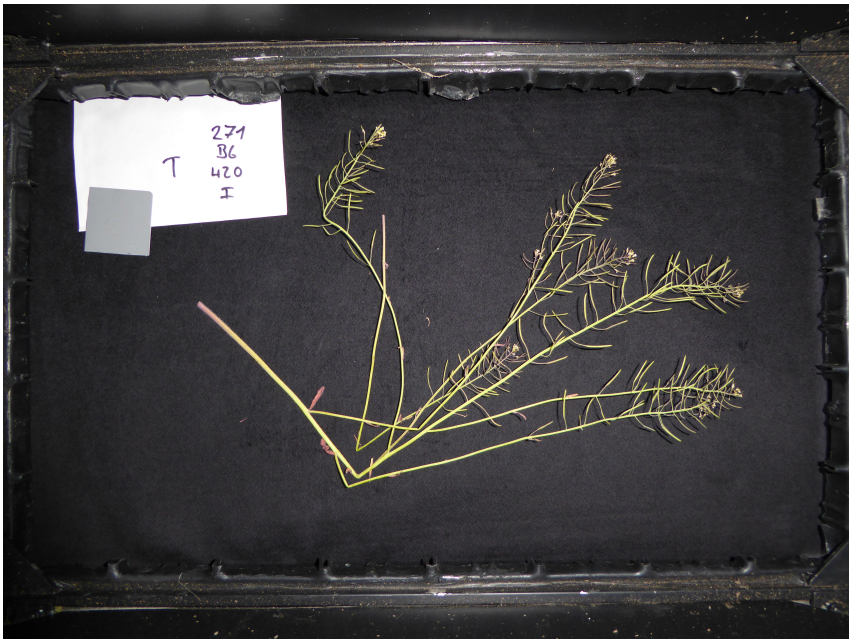
(a)



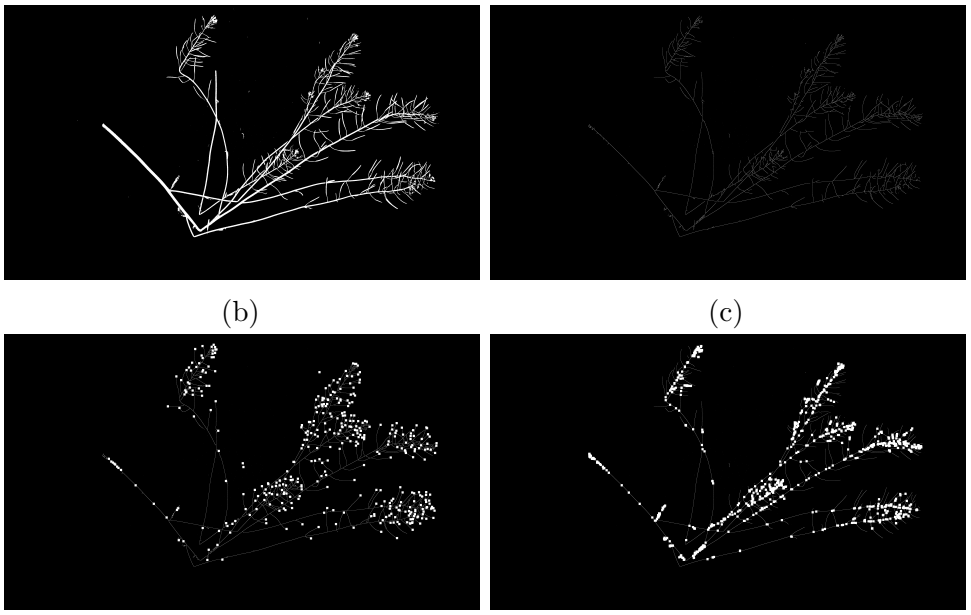


(b) (c)

Figure 7: Example segmentation results from raw image (a) to green (b) and red (c) pixels only.



(a)



(b) (c)

(d) (e)

Figure 8: Example sakeletonisation results from raw image (a) to segmented (b), skeletonised (c), the detected branches (e) and endpoints (d).