A noninvasive, machine learning–based method for monitoring anthocyanin accumulation in plants using digital color imaging

Bryce C. Askey1, Ru Dai1, Won Suk Lee2, and Jeongim Kim1,3

PREMISE: When plants are exposed to stress conditions, irreversible damage can occur, negatively impacting yields. It is therefore important to detect stress symptoms in plants, such as the accumulation of anthocyanin, as early as possible.

METHODS AND RESULTS: Twenty-two regression models in five color spaces were trained to develop a prediction model for plant anthocyanin levels from digital color imaging data. Of these, a quantile random forest regression model trained with standard red, green, blue (sRGB) color space data most accurately predicted the actual anthocyanin levels. This model was then used to noninvasively monitor the spatial and temporal accumulation of anthocyanin in Arabidopsis thaliana leaves.

CONCLUSIONS: The digital imaging–based nature of this protocol makes it a low-cost and noninvasive method for the detection of plant stress. Applying a similar protocol to more economically viable crops could lead to the development of large-scale, cost-effective systems for monitoring plant health.

KEYWORDS: anthocyanin; digital color imaging; early stress detection; machine learning.

The world’s population is projected to reach 9.7 billion by 2050, and agricultural production must increase accordingly to match demand (FAO, 2018). However, the amount of arable land suitable for agricultural expansion is greatly limited by multiple factors, the most significant of which are climate change and land degradation. In addition, accessible freshwater resources (such as rivers and aquifers) in many countries are being depleted at a greater rate than they are regenerated (FAO, 2018). As a result, farmers in the coming decades must produce more food with less land and water inputs by increasing the efficiency of their agricultural systems. Traditional crop growing practices require the application of a treatment (e.g., additional water, fertilizer, or pesticides) to an entire field of crops, based on the assumption that all plants in the field will grow and produce yields relatively homogeneously; however, biotic factors such as pests and disease often affect a field of crops in a heterogeneous manner with respect to time and space (Mahlein et al., 2012). Abiotic stresses, including high temperature and drought, also negatively impact plant productivity, but affect fields of crops in a more homogeneous manner (Zhao et al., 2017).

One of the aims of precision agriculture is to increase agricultural efficiency through the early, site-specific application of treatments. This approach reduces the amount of inputs required while maintaining or increasing total yields (Gebbers and Adamchuk, 2010). A method of monitoring plant stress is an essential part of precision agriculture systems, as it allows for the early detection of abiotic and biotic factors that may impact plant health, giving growers the opportunity to apply specific corrective action before yields are harmed. Imaging-based platforms are the most commonly applied system for monitoring plant health. These platforms record the intensity of light reflected, absorbed, or fluoresced by a plant at various wavelengths, which can change with the health of the plant and severity of any stress factors affecting it (Mutka and Bart, 2015). The popularity of these systems is a result of their capacity for nondestructive and easily automatable high-throughput phenotyping. Imaging-based platforms can also detect and provide a visualization of the heterogeneous development of stress symptoms in a plant, which is helpful for determining the type and severity of stress they are experiencing (Mutka and Bart, 2015).

Hyperspectral, fluorescence, near-infrared, and thermal imaging are several popular imaging techniques used in phenotyping platforms, and have proven to be effective in detecting the photosynthetic status, water content, presence of disease, and other parameters for a variety of plants (Li et al., 2014). However, the cameras and software required for the implementation of these
systems are costly when compared to digital color imaging–based systems. The popularity of digital cameras has made them easily accessible and often included in many modern phones; however, while digital cameras work well for regular consumer use, their applications in spectroscopy are limited by their spectral resolution and range, as they can only detect visible light in three spectral bands: red, green, and blue (RGB). To account for this limitation while taking advantage of their low cost, we investigated a machine learning regression method for processing the color data recorded by a digital color camera. Once trained, a machine learning regression method can be used to predict the state or level of an output variable from the information contained in multiple predictor variables. Compared with traditional regression methods, machine learning regression methods have the potential to describe more complicated relationships between predictor and output variables through the application of flexible, data-driven algorithms (Qin et al., 2011).

Anthocyanins are a group of secondary plant metabolites believed to serve as antioxidants to quench reactive oxygen species (Kovinich et al., 2014). They are produced from the amino acid phenylalanine via the anthocyanin biosynthetic pathway. Many studies have shown that anthocyanin accumulation is positively correlated with various types of stresses (Appendix S1), including high-intensity light, pathogens, wounding, drought, and nutrient deficiency (Shan et al., 2009). Anthocyanin accumulation can therefore be correlated with plant health, and acts as an indicator of the severity of stresses affecting the plant. Traditional chemical analysis methods for anthocyanin quantification can be time consuming, costly, and invasive (Yang et al., 2016); therefore, in this work, we investigated a noninvasive, digital color imaging–based method using a machine learning regression analysis to predict levels of anthocyanin accumulation in detached Arabidopsis thaliana (L.) Heynh. leaves.

METHODS AND RESULTS

A flowchart of the methodology followed to prepare the leaf samples, process image data, and evaluate the accuracy of the developed method is shown in Fig. 1. A detailed benchside protocol for the image background removal and processing, image sorting, regression training and testing, and false-color heatmap generation is provided in Appendix 1.

Leaf sample preparation and imaging

To simulate the varying levels of anthocyanin accumulation caused by stress, four genotypes of Arabidopsis with different anthocyanin accumulation levels were used: Col-0, ref2-1 (ref2), cyp79b2 cyp79b3 (b2b3), and pap1-D (Fig. 2). Col-0 is A. thaliana wild type. pap1-D is an activation tagging mutant with an enhanced expression of MYB75, which is a master regulator of the anthocyanin biosynthesis pathway (Borevitz et al., 2000). Due to the activation of MYB75 in pap1-D, several genes encoding key enzymes in anthocyanin biosynthesis (e.g., chalcone synthase, dihydroflavonol reductase, and anthocyanidin synthase) are upregulated in this mutant, resulting in its overproduction of anthocyanin (Sawano et al., 2017). ref2 and b2b3 are mutants with defects in the enzymes involved in the biosynthesis of defense compounds called glucosinolates (Zhao et al., 2002; Hemm et al., 2003). The altered levels of glucosinolate intermediates in ref2 and b2b3 have been shown to influence the...
stability of phenylalanine ammonia lyase (PAL) functioning at the first step of the phenylpropanoid pathway in these mutants; as a result, ref2 contains a reduced level of phenylpropanoids, including anthocyanin, while b2b3 accumulates more phenylpropanoids than the wild type (Kim et al., 2015, 2019). Because the Col-0, pap1-D, ref2, and b2b3 genotypes accumulate different levels of anthocyanin under nonstressed conditions, we were able to consistently generate Arabidopsis plants with a broad range of anthocyanin contents without the use of additional stress treatments (Fig. 2A).

The Arabidopsis seeds were planted on moist soil and grown for 4–6 weeks in a growth chamber at 23°C, with a light intensity of 140 μE m−2 s−1 in a 16-h light/8-h dark photoperiod. The plants were watered regularly to maintain soil moisture levels. Just before or soon after bolting, the mature rosette leaves were harvested by cutting the petiole using a pair of small scissors, and their fresh weights were determined using an analytical balance.

A digital image of the leaves was taken with a Canon Powershot G15 digital camera (Canon USA Inc., Melville, New York, USA). To standardize the lighting conditions, all photos were taken in the same location, with the only source of lighting being ceiling fluorescent lights. The camera settings were also kept constant for all images, and were set as follows: image size = 12 megapixels (4000 × 3000 pixels), f-stop = f/2.8, exposure time = 1/100 s, ISO speed = ISO-800, exposure bias = −1 step, no flash. To calibrate the camera, its white balance was set using a grey card from the Vello White Balance Card Set (Vello, New York, New York, USA) prior to each imaging session. Eight leaves at a time were arranged on a sheet of nonreflective black paper with the top surface of the leaves facing upward. A tripod was used to position the camera 1–2 feet above the leaves, and the zoom of the camera was adjusted so that the leaves nearly filled the frame of the photograph. A photograph of the leaves was then taken in the standard RGB (sRGB) color space.

Anthocyanin extraction and measurement

The general methodology used for anthocyanin extraction and measurement follows that described by Shan et al. (2009). Immediately after weighing and imaging, each leaf was immersed in 1 mL of extraction buffer (18% 1-propanol, 1% HCl, and 81% water) and boiled for 3 min. The leaves were then incubated in the buffer solution at room temperature in darkness for 24 h. A 1-mL aliquot of the buffer solution from each leaf was transferred to a cuvette, and its absorbance at 535 nm and 650 nm was measured spectrophotometrically. The level of anthocyanin accumulation, designated as the normalized anthocyanin index, was calculated using the following equation:

\[
\text{normalized anthocyanin index (NAI)} = \frac{A_{535} - A_{650}}{\text{Fresh Weight} \times [g]}
\]

Image processing and regression analysis

The background pixels of the leaf images were manually removed with GNU Image Manipulation Program (GIMP, version 2.10.4; https://www.gimp.org/). Each image contained multiple leaves; therefore, each leaf in an image was individually exported as a separate .png file. To quantify the color of the images, the mean color index values of five color spaces were calculated for each image using R (version 3.5.3; R Core Team, 2019). The five color spaces used were sRGB; hue, saturation, value (HSV); luminance, in-phase, quadrature (YIQ); lightness, green to red, blue to yellow (L∗a∗b∗); and luma, blue-difference, red-difference (YCbCr).

For a given component in a color space, the mean color index values were calculated by summing together the color index value of each leaf pixel in an image and dividing by the total number of leaf pixels. For example, in the sRGB color space, the mean red (R) color index value for an image was calculated by summing together each leaf pixel’s R color index value and dividing by the total number of leaf pixels in the image. By repeating this process for the green (G) and blue (B) components of the image, a set of three mean color index values were calculated for each image in the sRGB color space. To calculate the mean color index values in the four other color spaces, the images were first converted into the HSV, YIQ, L∗a∗b∗, and YCbCr color spaces, after which their mean color index values could be calculated using the same method described for the sRGB color space.

A set of 22 regression models were trained and evaluated for their ability to predict the NAI of a leaf from its mean color index values in each color space. The leaves were first randomly sorted into training/validation and test sets. Of the 147 leaves used, 80% were sorted into the training/validation set (n = 118) and the remaining 20% were used as the test set (n = 29). To train and validate the regression models, the “caret” package in R was used (Kuhn, 2008). The mean color index values were set as predictors, and the NAI was set as the response variable. The 22 regression models were then trained in each of the five color spaces previously described, producing a total of 110 trained models.

The accuracy of each trained regression model was evaluated using the test set of leaf data. For each leaf in the test set, the NAI predicted by each trained regression model was compared to the actual NAI measured using the spectrophotometer method. The root mean squared error (RMSE), R², and mean average error (MAE) values were calculated for each regression model. The accuracy data...
of the 10 most accurate regressions, as determined using the RMSE value, are shown in Table 1. A complete list of all regressions and their respective accuracy parameters are provided at https://github.com/bryceaskay/anthocyanin_accumulation. Background pixel removal in GIMP is the only part of the image analysis process completed manually, as all downstream regression training and testing was automated in R. The complete analysis of the 147 leaf images required about 5–6 h of manual work in GIMP and about 8–10 h of computational time in R.

Of all the regression models tested, the quantile random forest model using the sRGB color space data most accurately predicted the actual NAI of the test set, with an RMSE of 9.407 NAI, an $r^2$ of 0.9351, and an MAE of 6.597 NAI. To confirm the regression accuracy, the predicted NAI values were plotted against the actual values (Fig. 3A), and a plot of the normalized residuals was also created (Fig. 3B). The normalized residuals were calculated by dividing the difference between the predicted and actual NAI values using the RMSE of the regression:

$$\text{normalized residual} = \frac{\text{Actual NAI} - \text{Predicted NAI}}{\text{RMSE}}$$

The even spread of residuals around the $x$-axis over the range of predicted NAI values indicates that the training/validation set of images was sufficiently large for training the regression.

**Spatiotemporal monitoring of anthocyanin accumulation**

To demonstrate its potential as an early detection system for plant stress, the trained sRGB quantile random forest model was used to make predictions of the spatiotemporal anthocyanin accumulation in detached leaves. *Arabidopsis* wild-type plants (Col-0) were grown for 4–6 weeks in the same growth chamber conditions described above. Mature leaves were harvested, and a digital color image was taken using the imaging method and camera settings described above. Because melatonin has been shown to enhance stress resistance (Zhang et al., 2014), various concentrations of melatonin were tested, with 1 mM melatonin found to be sufficient for reducing stress symptoms in *Arabidopsis*. The detached leaves were then individually placed into either water containing 0.05% Tween-20, which is considered as a high-stress environment due to mechanical wounding, or water containing 1 mM melatonin and 0.05% Tween-20, in which melatonin can reduce stress symptoms caused by wounding (low-stress environment). All Petri dishes were placed in a growth chamber at 23°C, with a light intensity of 140 μE m$^{-2}$ s$^{-1}$ in a 16-h light/8-h dark photoperiod.

| Color space  | Regression method                  | RMSE (NAI) | $r^2$ | MAE (NAI) |
|--------------|------------------------------------|------------|-------|-----------|
| sRGB         | Quantile random forest             | 9.407      | 0.9351| 6.597     |
| YIQ          | Random forest                      | 10.23      | 0.9266| 8.032     |
| YCbCr        | Quantile random forest             | 10.27      | 0.9257| 7.783     |
| YCbCr        | Random forest                      | 10.32      | 0.9290| 7.653     |
| L*$\ast$b*   | Bayesian neural network            | 10.44      | 0.9323| 7.909     |
| sRGB         | Stochastic gradient boosting       | 10.62      | 0.9219| 7.943     |
| YIQ          | Bayesian neural network            | 10.66      | 0.9296| 7.867     |
| YCbCr        | Bayesian neural network            | 10.68      | 0.9298| 7.759     |
| sRGB         | Random forest                      | 10.71      | 0.9219| 7.205     |
| sRGB         | Bayesian neural network            | 10.86      | 0.9263| 7.763     |

Note: MAE = mean average error; NAI = normalized anthocyanin index; RMSE = root mean squared error.

*Regressions trained using the “caret” package for R with data from the sRGB, HSV, YIQ, L*$\ast$b*$, and YCbCr color spaces.

**FIGURE 3.** Evaluation of the quantile random forest model accuracy using the sRGB values of a test set of leaves ($n = 29$). (A) Normalized anthocyanin index (NAI) values predicted using the quantile random forest model plotted against actual NAI values calculated using the spectrophotometer method. (B) Normalized residuals from the predicted values. Normalized residuals were calculated by dividing the difference between the actual and predicted NAI values for each leaf by the root mean squared error (RMSE) of the regression model.
The leaves were imaged every 24 h for 96 h following the initial immersion. Prior to each imaging session, the leaves were removed from the solution and gently dried with a paper towel, as any liquid on the surface of the leaf could affect the color data recorded by the camera. The imaging was then performed as described above. After imaging, the leaves were re-immersed in the appropriate solution, and the Petri dishes were returned to the growth chamber. The background pixels of the images were removed using GIMP, and the images of each leaf were individually exported as separate .png files.

A false-color heatmap of the anthocyanin accumulation in the leaves was created in R using the trained sRGB quantile random forest model (Fig. 4A, B). This model was then used to predict the NAI at each pixel in the leaf image. To generate the false-color heatmap, a new image was created, and each pixel in the image was assigned a color based on the predicted NAI of the corresponding pixel in the original leaf image. Pixels with a low predicted NAI were colored white in the new image, while those with a high predicted NAI were dark pink/red (Fig. 4C). After 24 h of the immersion treatment, anthocyanin accumulation was barely visible in the RGB images as a slight discoloration in the primary vein; however, this accumulation was much more apparent in the heatmap images, allowing for the easy visualization of the distribution of anthocyanin, which might otherwise go undetected in a visual inspection. A quantitative prediction of anthocyanin accumulation in each leaf image was also performed by averaging the predicted NAI at each pixel over the entire area of the leaf (Fig. 4D).

CONCLUSIONS

This study demonstrated that the application of a machine learning regression to digital color image data enables the accurate prediction of the anthocyanin content of detached *Arabidopsis* leaves. Of the 22 regression models evaluated, a quantile random forest model in the sRGB color space most accurately predicted the actual accumulation of anthocyanin, with an RMSE of 9.407 NAI, an $r^2$ of 0.9351, and an MAE of 6.597 NAI (Table 1). To demonstrate its potential as an early detection system for plant stress, the trained model was then applied to leaf images to monitor the spatiotemporal accumulation of anthocyanin in leaves exposed to different stress conditions (Fig. 4). Anthocyanin accumulates heterogeneously in plants (Anderson et al., 2015), and the generated false-color heatmaps allowed the visualization of its distribution. In comparison with the hyperspectral, thermal, near-infrared, and

**FIGURE 4.** sRGB quantile random forest model used to predict the spatiotemporal accumulation of anthocyanin. (A, B) RGB images (top) and the predicted anthocyanin accumulation (bottom) of detached leaves immersed in water (A), or in water containing 1 mM melatonin (B), over a 96-h period. Scale bar = 1 cm. (C) Color scale used to create false-color images. (D) Average predicted normalized anthocyanin index (NAI) for the leaves with (+) or without (−) melatonin over the 96-h experimental period.
fluorescence imaging–based methods for monitoring plant stress, digital color imaging–based methods are a more cost-effective solution (Mutka and Bart, 2015). Digital cameras are cheap, accessible, and user friendly, and it is likely that most users are already familiar with their operation. In addition, both GIMP and R are open-source software, meaning they are completely free to use.

The application of a machine learning regression model accounts for the complex relationship between each component of a color space and the actual anthocyanin content of the leaf. This multivariate machine learning approach demonstrates an improved accuracy in the estimation of anthocyanin accumulation from a digital color image compared with single-variable regression methods (Murakami et al., 2005; Yang et al., 2016). When compared with hyperspectral imaging–based methods for estimating anthocyanin accumulation, the digital color imaging–based method developed in this study provides a comparable accuracy in controlled conditions, without the need for more expensive hyperspectral imaging equipment (Gitelson et al., 2006; Qin et al., 2011).

As our regression predictions are based on color data, the color accuracy in the digital images used is very important. In our study, images were taken in a laboratory setting under fluorescent ceiling lights, and the camera was calibrated with a grey card before each imaging session. Because the spectral signature of the leaves will change based on the primary source of light in the imaging environment, it is likely that a separate regression would need to be trained to account for the lighting in different environments, such as those lit by natural sunlight or wavelength-controlled LEDs.

Because this study was conducted with detached Arabidopsis leaves, the next step would be to test the applicability of the method at the whole-plant level. Future work should include a comparison of the sensitivity and accuracy of this digital imaging–based method against standard phenotyping methods, as well as further testing with more economically viable crops that also accumulate anthocyanin when exposed to stress, such as cotton (Gossypium hirsutum L.) (Li et al., 2019), maize (Zea mays L.) (Pietrini et al., 2002), and tomato (Solanum lycopersicum L.) (Groher et al., 2018). The automation of the background removal process, the effect of variable lighting conditions on the regression accuracy, and larger-scale field testing should also be a part of future investigations. Finally, the viability of using a cell phone camera as a replacement for the stand-alone camera used here should also be explored. Improvements in these areas would allow this method to be applied as a cost-effective early detection system for plant stress.

ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA) Hatch project (005681), as well as by the University Scholars Program and a startup fund from the Horticultural Sciences Department and Institute of Food and Agricultural Sciences at the University of Florida.

AUTHOR CONTRIBUTIONS

B.C.A. and J.K. conceived and designed the experiments; B.C.A. and R.D. performed the experiments; B.C.A. and W.S.L. developed the methods used for data analysis; and B.C.A. and J.K. wrote the manuscript. All authors read and approved the final manuscript.

DATA AVAILABILITY

All R code used to process and sort the leaf images, train the regressions, and evaluate the accuracy of the regressions can be downloaded from https://github.com/bryceaskey/anthocyanin_accumulation.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

APPENDIX S1. Drought stress induces anthocyanin accumulation. Side view (A) and overhead (B) photos of wild-type Arabidopsis thaliana (Col-0) under either well-watered (control) or water-limited (stress) conditions.

LITERATURE CITED

Anderson, N. A., N. D. Bonawitz, K. Nyffeler, and C. Chapple. 2015. Loss of FERULATE 5-HYDROXYLASE leads to mediator dependent inhibition of soluble phenylpropanoid biosynthesis in Arabidopsis. Plant Physiology 169: 1557–1567.

Borevitz, J. O., Y. Xia, J. Blount, R. A. Dixon, and D. Lamb. 2000. Activation tagging identifies a conserved MYB regulator of phenylpropanoid biosynthesis. Plant Cell 12(12): 2383–2393.

FAO. 2018. The future of food and agriculture: Alternative pathways to 2050. Food and Agriculture Organization of the United Nations, Rome, Italy.

Gebbers, R., and V. I. Adamchuk. 2010. Precision agriculture and food security. Science 327(5967): 828–831.

Gitelson, A. A., G. P. Keydan, and M. N. Merzlyak. 2006. Three-band model for noninvasive estimation of chlorophyll, carotenoids, and anthocyanin contents in higher plant leaves. Geophysical Research Letters 33(11): https://doi.org/10.1029/2006GL026457.

Groher, T., S. Schmittgen, G. Noga, and M. Hunsche. 2018. Limitation of mineral supply as tool for the induction of secondary metabolites accumulation in tomato leaves. Plant Physiology and Biochemistry 130: 105–111.

Hemm, M. R., M. O. Ruegger, and C. Chapple. 2003. The Arabidopsis rfs2 mutant is defective in the gene encoding CYP83A1 and shows both phenylpropanoid and glucosinolate phenotypes. Plant Cell 15(1): 179–194.

Kim, J. I., W. L. Dolan, N. A. Anderson, and C. Chapple. 2015. Indole glucosinolate biosynthesis limits phenylpropanoid accumulation in Arabidopsis thaliana. Plant Cell 27(5): 1529–1546.

Kim, J. I., X. Zhang, P. E. Pascuzzi, C.-J. Liu, and C. Chapple. 2019. Glucosinolate and phenylpropanoid biosynthesis are linked by proteasome-dependent degradation of PAL. New Phytologist https://doi.org/10.1111/nph.16108.

Kovinich, N., G. Kayanja, A. Chanoca, K. Riedl, M. Otegui, and E. Grotewold. 2014. Not all anthocyanins are born equal: Distinct patterns induced by stress in Arabidopsis. Planta 240(5): 931–940.

Kuhn, M. 2008. Building predictive models in R using the caret package. Journal of Statistical Software 28(5): 1–26.

Li, L., Q. Zhang, and D. Huang. 2014. A review of imaging techniques for plant phenotyping. Sensors 14: 20078–2011.

Li, X., X. Ouyang, Z. Zhang, L. He, Y. Wang, Y. Li, J. Zhao, et al. 2019. Overexpression of the red plant gene R1 enhances anthocyanin production and
resistance to bollworm and spider mite in cotton. Molecular Genetics and Genomics 294(2): 469–478.

Mahlein, A., E. Oerke, U. Steiner, and H. Dehne. 2012. Recent advances in sensing plant diseases for precision crop protection. European Journal of Plant Pathology 133(1): 197–209.

Murakami, P. F., M. R. Turner, A. K. van den Berg, and P. G. Scharberg. 2005. An instructional guide for leaf color analysis using digital imaging software. General Technical Report NE-327. USDA Forest Service, Northeastern Research Station, Newtown Square, Pennsylvania, USA. https://doi.org/10.2737/NE-GTR-327.

Mutka, A. M., and R. S. Bart. 2015. Image-based phenotyping of plant disease symptoms. Frontiers in Plant Science 5: https://doi.org/10.3389/fpls.2014.00734.

Pietrini, F., M. A. Ianneli, and A. Massacci. 2002. Anthocyanin accumulation in the illuminated surface of maize leaves enhances protection from photoinhibitory risks at low temperature, without further limitation to photosynthesis. Plant Cell and Environment 25(10): 1251–1259.

Qin, J., D. Rundquist, A. Gitelson, Z. Tan, and M. Steele. 2011. A non-linear model of nondestructive estimation of anthocyanin content in grapevine leaves with visible/red-infrared hyperspectral. Computer and Computing Technologies in Agriculture IV 4: 47–62.

R Core Team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Sawano, H., T. Matsuizaki, T. Usui, M. Tabara, A. Fukudome, A. Kanaya, D. Tanoue, et al. 2017. Double-stranded RNA-binding protein DRB3 negatively regulates anthocyanin biosynthesis by modulating PAP1 expression in Arabidopsis thaliana. Journal of Plant Research 130(1): 45–55.

Shan, X., X. Zhang, W. Peng, Z. Wang, and D. Xie. 2009. Molecular mechanism for jasmonate-induction of anthocyanin accumulation in Arabidopsis. Journal of Experimental Botany 60(13): 3849–3860.

Yang, X., J. Zhang, D. Guo, X. Xiong, L. Chang, Q. Niu, and D. Huang. 2016. Measuring and evaluating anthocyanin in lettuce leaf based on color information. IFAC-PapersOnLine 49(16): 96–99.

Zhao, J. A. K. Hull, N. R. Gupta, K. A. Goss, J. Alonso, J. R. Ecker, J. Normanly, et al. 2002. Trp-dependent auxin biosynthesis in Arabidopsis: Involvement of cytochrome P450s CYP79B2 and CYP79B3. Genes and Development 16(23): 3100–3112.

Zhao, C., B. Liu, S. Piao, X. Wang, D. B. Lobell, Y. Huang, M. Huang, et al. 2017. Temperature increase reduces global yields of major crops in four independent estimates. Proceedings of the National Academy of Sciences USA 114(35): 9326–9331.

APPENDIX 1. Benchside protocol for image background removal and processing, regression training and testing, and false-color heatmap generation.

Equipment list
- Canon Powershot G15 digital camera (Canon USA Inc., Melville, New York, USA; https://shop.usa.canon.com/shop/en/catalog/powershot-g15)
- Vello WB-CMII White Balance Card Set – Medium (Vello, New York, New York, USA; https://www.vellogear.com/product/12273/Vello-WB-CMII-White-Balance-Card-Set-(Medium))

Removal of background pixels from leaf images
1. Download and install the latest version of GNU Image Manipulation Program (GIMP) (https://www.gimp.org/downloads/).

2. Open GIMP. Load an image into the workspace by clicking File > Open and navigating to where the image is saved on the computer. Double click on the image file to select it.

3. Once the image is loaded, save it as a .xcf file by pressing “Ctrl” + “S”, or by navigating to File > Save As. In the “Save Image” popup, use the “Name” box to name the image, and the file explorer to select where it will be saved.

4. Zoom in on a leaf in the image by holding down “Ctrl” and scrolling up with the mouse. To zoom out, hold down “Ctrl” and scroll down. To move the image, hold down the middle mouse button, or the space bar, and move the mouse.

5. Center the leaf in the workspace and use the “Scissors Select Tool” by pressing “I” on the keyboard. A small pair of scissors should appear to the bottom right of the cursor.

6. Begin cutting out the leaf from its background by left clicking on an edge of the leaf. Move the cursor a short distance away, and left click on another edge of the leaf. A line should connect the two points, tracing a section of the edge of the leaf.

7. Continue selecting points on the edge of the leaf in a continuous direction (clockwise or counterclockwise) until the first point is almost reached. The distances separating the selected points should vary based on the contrast between the color of the leaf and its background. In addition, more points will need to be selected at the sharp corners of the leaf (e.g., the end of the stem).

8. Close the selected area by clicking on the first point, then select the leaf area by clicking on it.

9. If there are any errors in the selected area, clean up the edges with the “Free Select Tool”. Press “F” on the keyboard to select the tool. If any part of the leaf is missing from the initial selected area, add it by pressing “Shift”, drawing a shape that encloses the missing area while holding the left mouse button, and pressing “Enter”. If any background is erroneously included in the initial selected area, remove it by pressing “Ctrl”, drawing a shape that encloses the unwanted area while holding left mouse button, and pressing “Enter”.

10. Cut the selected area from the image by pressing “Ctrl” + “X”. With the cursor over the “Layers” tab on the right side of the interface, create a new layer for the selected area by right clicking and selecting “New Layer”. In the popup, use the “Layer Name” box to give a name to the layer, but leave all of the other settings as their default. Press “OK” to create the layer, and paste the selected area into the layer by pressing “Ctrl” + “V”, followed by “Ctrl” + “H”.

11. In the “Layers” tab, select the layer containing the original image.

12. Repeat steps 3–10 for all leaves in the image.

13. To export each leaf-containing layer as a separate .png file, hide all of the layers except the one containing the leaf to be exported. Do this by left clicking the eye-shaped icon that appears to the left of each visible layer in the “Layers” tab.

14. Export the layer by pressing “Ctrl” + “Shift” + “F”. In the “Export Image” popup, use the “Name” box to name the exported image, and ensure that the file extension at the end of the name is .png. Using the file explorer in the popup, navigate to where the image will be saved and click “Export”. An additional popup will appear, titled “Export Image as PNG”. Leave all settings as their default and click “Export”. Note:
To facilitate downstream processing, it is best to name the exported images with their corresponding sample numbers.

15. Repeat steps 13 and 14 for all leaf-containing layers.

**Calculating mean color index values, sorting data into training/validation and test sets, training regressions, evaluating regression accuracy, and making heatmaps**

1. Ensure that all leaf images only contain a single leaf, have had their background removed, and are all saved within a single folder.

2. Prepare a Microsoft Excel (Excel version 16.0.11601.2018 [64-bit]; Microsoft, Redmond, Washington, USA) spreadsheet containing columns with sample names and their associated NAI. Only the first row of the spreadsheet should contain column labels. Sample names should match the name of the corresponding image (a “.png” at the end is not necessary). Save the spreadsheet as a .csv file.

3. Download and install R (https://www.r-project.org/).

4. Download and install RStudio Desktop (https://www.rstudio.com/products/rstudio/#Desktop).

5. Download the R code in the anthocyanin_accumulation repository (https://github.com/bryceaskey/anthocyanin_accumulation) by clicking the green “Clone or download” button and selecting “Download ZIP”. Extract the files from the downloaded .zip.

6. Open the file named “main.R” in RStudio by pressing “Ctrl” + “O” on the keyboard and navigating to the directory into which the files from the .zip folder were extracted.

7. Run the file by pressing “Ctrl” + “Shift” + “S” on the keyboard, and enter the appropriate file directories in the console according to the instructions given. The code will automatically calculate the mean color index values for a set of images, read NAI data from a .csv file, sort data into training/validation and test sets, train regressions, evaluate their accuracy, and use the most accurate regression to generate false-color heatmaps for a set of images.

8. Once the code has finished running, all output data will be displayed under the “Environment” tab in the upper right of RStudio. “allData” contains the mean color index values and actual NAI for each image. “allModels” contains the parameters for each trained regression model. “modelAccuracies” contains a table of accuracy parameters for all trained regression models, sorted from smallest to largest RMSE. “heatmapAvgNAI” contains the average predicted NAI for each heatmap image. If heatmap images were created, they can be found in the same directory that contained the original images. “Heatmap” is added to the end of the original image name to differentiate it from the original image.