Association of leaf spectral variation with functional genetic variants

Cheng Li[1,✉], Ewa A. Czyż[1], Rishav Ray[2], Rayko Halitschke[2], Ian T. Baldwin[2], Michael E. Schaepman[1], and Meredith C. Schuman[1,3]

[1]Department of Geography, Faculty of Science, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
[2]Department of Molecular Ecology, Max Planck Institute for Chemical Ecology, Hans-Knoell-Strasse 8, 07745 Jena, Germany
[3]Department of Chemistry, Faculty of Science, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

The application of in-field and aerial spectroscopy to assess functional and phylogenetic variation in plants has led to novel ecological insights and promises to support global assessments of plant biodiversity. Understanding the influence of plant genetic variation on reflectance spectra will help to harness this potential for biodiversity monitoring and improve our understanding of why plants differ in their functional responses to environmental change. Here, we use an unusually well-resolved genetic mapping population in a wild plant, the coyote tobacco *Nicotiana attenuata*, to associate genetic differences with differences in leaf spectra for plants in a field experiment in their natural environment. We analyzed the leaf reflectance spectra using FieldSpec 4 spectroradiometers on plants from 325 fully genotyped recombinant inbred lines (RILs) of *N. attenuata* grown in a blocked and randomized common garden experiment. We then tested three approaches to conducting Genome-Wide Association Studies (GWAS) on spectral variants. We introduce a new Hierarchical Spectral Clustering with Parallel Analysis (HSC-PA) method which efficiently captured the variation in our high-dimensional dataset and allowed us to discover a novel association, between a locus on Chromosome 1 and the 445-499 nm spectral range, which corresponds to the blue light absorption region of chlorophyll, indicating a genetic basis for variation in photosynthetic efficiency. These associations lie in close proximity to candidate genes known to be expressed in leaves and having annotated functions as methyltransferases, indicating possible underlying mechanisms governing these spectral differences. In contrast, an approach using well-established spectral indices related to photosynthesis, reducing complex spectra to a few dimensionless numbers, was not able to identify any robust associations, while an approach treating single wavelengths as phenotypes identified the same associations as HSC-PA but without the statistical power to pinpoint significant associations. The HSC-PA approach we describe here can support a comprehensive understanding of the genetic determinants of leaf spectral variation which is data-driven but human-interpretable, and lays a robust foundation for future research in plant genetics and remote sensing applications.

Leaf reflectance — Field spectroscopy — Multiparent Advanced Generation Inter-Cross (MAGIC) — Genome-wide association studies (GWAS) — Hierarchical Spectral Clustering with Parallel Analysis — *Nicotiana attenuata*

Correspondence: ***cheng.li@geo.uzh.ch***

1. Introduction

Genetic diversity is essential for the survival and adaptation of species to changing environmental conditions. Radiation reflected, absorbed, and transmitted by plants constitutes the basis for remote sensing of vegetation, commonly in the range from visible to infrared wavelengths corresponding to solar radiation (Knipling, 1970; Thenkabail, Lyon and Huete, 2018). Remote sensing could revolutionize the way we study genetic diversity by offering non-invasive methods with the potential of repeated measurement on large spatial scales (Madritch, Kingdon, Singh, Mock, Lindroth and Townsend, 2014). Remote sensing technologies have not only facilitated the study of Earth's surface features, including vegetation, water bodies, and land use patterns, but are also gaining applications in biodiversity research. Wang and Gamon recently discussed concrete ways in which remote sensing makes unique contributions to monitoring plant biodiversity (Wang and Gamon, 2019). For example, Féret and Asner showcased the capabilities of high-fidelity imaging spectroscopy for mapping the diversity of tropical forest canopies (Asner, Martin, Carranza-Jiménez, Sinca, Tupayachi, Anderson and Martinez, 2014). The Global Ecosystem Dynamics Investigation, as discussed by Dubayah et al., leverages high-resolution laser ranging to monitor forests and topography (Dubayah, Blair, Goetz, Fatoyinbo, Hansen, Healey, Hofton, Hurtt, Kellner, Luthcke et al., 2020). Although they do not capture the structural variation of whole canopies, reflectance spectra from single leaves, which can be captured with controlled lighting and background from plants in field conditions, also offer insights into plant physiology, stress responses, and genetic variation, with recent studies establishing a correlation between genetic diversity and leaf reflectance spectra (Serbin, Singh, McNeil, Kingdon and Townsend, 2014; Cavender-Bares, Meireles, Couture, Kaproth, Kingdon, Singh, Serbin, Center, Zuniga, Pilz et al., 2016; Wang, Gamon, Schweiger, Cavender-Bares, Townsend, Zygielbaum and Kothari, 2018; Asner, Martin, Anderson and Knapp, 2015; Czyż, Schmid, Hueni, Eppinga, Schuman, Schneider, Guillén-Escribà and Schaepman, 2023). We recently showed that genetically variable plant populations are also spectrally more variable in comparison to replicates of an inbred genotype grown under either glasshouse or field conditions, while isogenic plants differing primarily in their gene expression are spectrally similar to replicates of the inbred genotype from which they were derived (Li, Czyz, Halitschke, Baldwin, Schaepman and Schuman, 2023a).

A Genome-Wide Association Study (GWAS) is a powerful tool used to identify genetic variants associated with

specific traits or diseases in populations. By examining the entire genome, researchers can pinpoint specific genetic markers that correlate with phenotypic variants. A primary advantage of GWAS is that it requires no prior knowledge of potential candidate genes and can be used for the discovery of novel genetic associations (Visscher, Wray, Zhang, Sklar, McCarthy, Brown and Yang, 2017). The corresponding disadvantage of GWAS is that it does not provide causal links and thus the mechanisms underlying statistical associations must be dissected and tested. A primary concern of GWAS is the influence of population structure, which, if not addressed, may produce misleading associations (Price, Patterson, Plenge, Weinblatt, Shadick and Reich, 2006). Population structure refers to the presence of subgroups within a population that differ in allele frequencies due to shared ancestry. In GWAS, population structure can be confounding, because genetic variants associated with the subgroup, rather than the trait of interest, may result in false-positive associations. To mitigate this, various methods, such as principal component analysis (PCA) and mixed linear models (Zhang, Ersoz, Lai, Todhunter, Tiwari, Gore, Bradbury, Yu, Arnett, Ordovas et al., 2010), are employed to correct for population structure in GWAS analyses.

Nicotiana attenuata is a model wild plant for molecular ecology native to the Great Basin Desert of the southwestern USA. It is primarily found in large ephemeral populations following fires in sagebrush and pinyon-juniper ecosystems. Its unique germination behavior is stimulated by cues found in wood smoke and the removal of inhibitors from unburned litter, allowing it to thrive in post-fire environments (Bahulikar, Stanculescu, Preston and Baldwin, 2004). Ecologically, *N. attenuata* is a compelling subject for its intricate defense mechanisms against herbivores, its tissue-specific diurnal metabolic rhythms, and its ability to acclimate to varying environmental conditions, including high UVB radiation (Glawe, Zavala, Kessler, Van Dam and Baldwin, 2003; Kim, Yon, Gaquerel, Gulati and Baldwin, 2011; Li, Heiling, Baldwin and Gaquerel, 2016; ĐINH, Galis and Baldwin, 2013). Further, it offers unique opportunities for genetic research due to a well-defined MAGIC (Multiparent Advanced Generation Inter-Cross) population (Ray, Li, Halitschke and Baldwin, 2019; Ray, Halitschke, Gase, Leddy, Schuman, Rodde and Baldwin, 2023). This design comprises 26 phenotypically and genetically differentiated parental lines (PLs) intercrossed to produce 325 genetically mapped recombinant inbred lines (RILs), allowing for the study of phenotype-genotype associations without the significant influence of population structure.

In this study, we harness the genetic diversity of the MAGIC population to discern associations between genetic variants and leaf spectral traits. Our sample comprises two replicates of the 325 *N. attenuata* MAGIC RILs. Our approach advances the state of the art for discerning genetic contributions to variation in complex spectral traits. We test three ways of treating the spectral phenotype data: spectral indices, which are single values derived from ratios of reflectance at specific wavelengths representing spectral features; single wavelengths (SW), i.e., treating each wavelength in the spectrum as a phenotype; and Hierarchical Spectral Clustering with Parallel Analysis (HSC-PA), a new data-driven dimension reduction approach, developed from a method for the analysis of genetic associations with human facial features (Claes, Roosenboom, White, Swigut, Sero, Li, Lee, Zaidi, Mattern, Liebowitz et al., 2018; Sero, Zaidi, Li, White, Zarzar, Marazita, Weinberg, Suetens, Vandermeulen, Wagner et al., 2019), which accounts for correlations within spectral data while retaining interpretable features. We associate the resulting phenotypic data with genotypes from a pool of 183,942 SNPs using Genome-Wide Association Study (GWAS) models based on the Genome Association and Prediction Integrated Tool (GAPIT) version 3 (Wang and Zhang, 2021). We avoid artifacts by incorporating important covariates such as maternal lineage and the number of measured leaves, along with kinship. We employ a set of GWAS models: General Linear Model (GLM), Mixed Linear Model (MLM), Fixed and random model Circulating Probability Unification (FarmCPU), and Bayesian information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK), to assess the robustness of our findings. The resulting Manhattan plots indicate genetic markers and patterns associated with the spectral phenotypes under investigation, which we then filter for significance using an empirically derived threshold to account for multiple testing. Figure 1 provides a summary of the methodologies and tools employed in this study. Our results show that HSC-PA, employed with an appropriate GWAS model, is the most sensitive of the three approaches for revealing genetic associations with spectral variants. We interpret these associations using additional information from the annotated *N. attenuata* genome and transcriptome and hypothesize that variation in specific methylating enzymes explains natural variation in photosynthetic efficiency as revealed by differences in how leaves reflect blue light.

2. Materials and Methods

The methods employed in this study largely replicate those detailed in our previous work. For a comprehensive description, readers are referred to (Li et al., 2023a). A brief summary is provided below for clarity.

2.1. Plant Material and Field Site *Nicotiana attenuata* is a native tobacco species predominantly found in the southwestern United States. The field site for this research was located at the Walnut Creek Center for Education and Research (WCCER) in Prescott, Arizona, within the natural habitat of *N. attenuata*. The MAGIC population of *N. attenuata* was derived from the Utah (UT) accession, a 31st-generation inbred line originally collected from the Desert Inn ranch in Washington County, Utah, USA. Other natural accessions used in this study were detailed in a previous publication (Ray et al., 2023). Germination and cultivation for the Arizona field plantation were previously described in Li et al. (2023a).

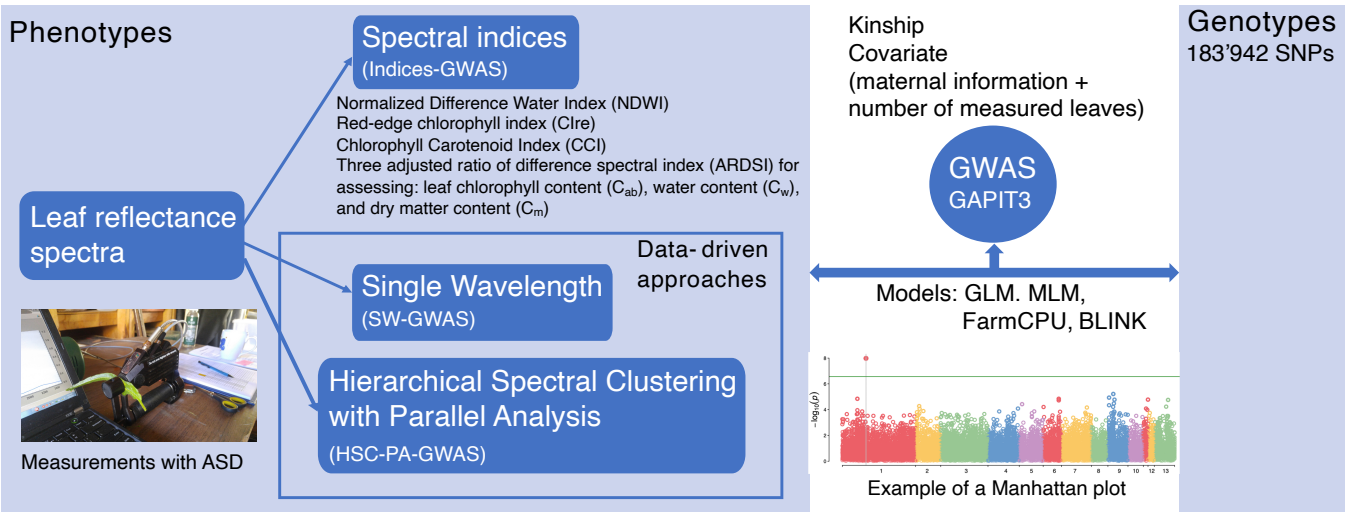

Fig. 1. Approaches and toolbox of this study.

2.2. Generation of MAGIC RIL Population and DNA sequencing To structure the genetic diversity in the MAGIC RIL population, a crossing scheme was developed with 26 parental lines. This entailed five rounds of systematic intercrosses, ensuring that each offspring would have all 26 parental lines as ancestors. The first round involved diallelic crossing of each of the 26 parental lines with each other, resulting in an F1 or A- generation of 325 different crosses. This process was meticulously detailed, with specific procedures to prevent self-pollination and ensure accurate cross-pollination. The subsequent rounds (B- to E-generations) further diversified the genetic makeup of the offspring. By the end of the fifth round, each plant had all 26 MAGIC founders as parents. These systematic crosses were followed by six generations of inbreeding to produce Recombinant Inbred Lines (RILs) with approximately 99% homozygosity, resulting in a MAGIC RIL population consisting of two replicates of 325 RILs. For the sequencing of the 650 MAGIC RILs, seedlings from the L1 and L2 generations were grown in the GH, and the sequencing was conducted at Novogene HK (Ray et al., 2023; Li et al., 2023a).

2.3. Equipment and optical measurement We employed a FieldSpec 4 spectroradiometers (Analytical Spectral Devices, Inc., Malvern Panalytical) 18140 with a plant probe (S/N 445) and leaf clip V2 attachments for measuring leaf optical properties. The FieldSpec encompasses three detectors, covering the visible and near-infrared to the shortwave infrared range of electromagnetic radiation. The devices offer a spectral resolution of 3 nm at 700 nm and 10 nm at 1400 and 2100 nm, and the FieldSpec system is radiometrically calibrated to provide measurements from which values for radiance and subsequently reflectance can be derived for every nanometer between 350 and 2500 nm.

Mature, hydrated, cut leaves harvested from comparable positions of field-grown plants were measured in batches immediately after each harvest, at their widest part, avoiding the midvein, as described in (Li et al., 2023a). Each sample underwent 20 scans under four conditions: white background reference (WR), white background with leaf (WRL), black background reference (BR), and black background with leaf (BRL). All measurement procedures for the Arizona field study were previously described (Li et al., 2023a).

2.4. Data processing Data processing was executed using R (R Core Team, 2023), primarily employing the spectrolab package (version 0.0.10) (Meireles, Schweiger and Cavender-Bares, 2017). We employed a rigorous three-step filtering approach to remove outliers from our dataset, as previously described (Li et al., 2023a). Briefly, we initially conducted a visual inspection of different measurement types, followed by the application of the Local Outlier Factor (LOF) method for each type. A final visual check was performed after calculating reflectance, ensuring the exclusion of outliers that could significantly impact our analysis. The spectral range analyzed spanned from 400-2500 nm, excluding the initial 50 nm (350-400 nm) due to high measurement uncertainty (Petibon, Czyż, Ghielmetti, Hueni, Kneubühler, Schaepman and Schuman, 2021). The calculated reflectance (CR) of a sample was obtained from the mean of scans using the following formula from (Miller, Steven and Demetriades-Shah, 1992):

$$CR = (R_{WR} \cdot R_{BRL} - R_{BR} \cdot R_{WRL})/(R_{WR} - R_{BR}) \quad \textbf{(1)}$$

2.5. Genome-wide association studies (GWAS) With this dataset, we tested three approaches to Genome-Wide Association Study (GWAS) with calculated leaf reflectance spectra.

2.5.1. Indices-GWAS This approach employs spectra to simultaneously assess multiple phenotypes defined by six common spectral indices that are widely recognized for their ability to characterize various vegetation traits. These indices were chosen based on their sensitivity to specific vegetation properties and their proven applicability in previous research, as summarized in Table 1:
The background and significance of these indices are further elaborated below:

Table 1. Six Spectral Indices Used in Indices-GWAS

Index	Formula	Example of Validated Systems	Citations
$NDWI$	$R_{865} - R_{1614}/R_{865} + R_{1614}$	Crops, forests, wetlands, grasslands	(Gao, 1996; Zhang and Zhou, 2019; Karthikeyan et al., 2020; Huete, 2012; Adam et al., 2010; Gu et al., 2007)
CI_{re}	$R_{783}/R_{704} - 1$	Crops, trees and vines, wetlands	(Gitelson et al., 2001; Qian et al., 2021; Gitelson et al., 2005, 2003; DeLancey et al., 2019)
CCI	$R_{560} - R_{664}/R_{560} + R_{664}$	Crops, trees, forests	(Gamon et al., 2016; Dechant et al., 2020; Grabska et al., 2019)
$ARDSIC_{ab}$	$R_{750} - R_{730}/R_{770} + R_{720}$	Crops	(Wan et al., 2021)
$ARDSIC_{w}$	$R_{1360} - R_{1080}/R_{1560} + R_{1240}$	Crops	(Wan et al., 2021)
$ARDSIC_{m}$	$R_{2200} - R_{1640}/R_{2240} + R_{1720}$	Crops	(Wan et al., 2021)

1. Normalized Difference Water Index ($NDWI$): NDWI is used to assess vegetation water content. It is based on the differential absorption of water in the near-infrared and shortwave infrared regions of the spectrum. NDWI has been widely applied in remote sensing to detect water stress in plants (Gao, 1996; Zhang and Zhou, 2019).

2. Red-edge chlorophyll index (CI_{re}): CIre is designed to estimate vegetation chlorophyll content. The red-edge region of the spectrum is sensitive to chlorophyll concentration, making CIre a valuable index for assessing plant vigor and growth (Gitelson et al., 2001; Qian et al., 2021).

3. Chlorophyll Carotenoid Index (CCI): CCI is sensitive to carotenoid/chlorophyll ratio and was initially used to track photosynthetic phenology (Gamon et al., 2016).

We additionally used three adjusted ratio of difference spectral indices ($ARDSI$), which are designed to reduce the differences in estimates obtained for adaxial and abaxial leaf surfaces (Wan et al., 2021), for assessing

4. vegetation chlorophyll content (C_{ab}),

5. water content (C_w), and

6. dry matter content (C_m), which have been validated in other plant species at both leaf and canopy levels.

2.5.2. Data-driven approaches The data-driven approach employed two different methods to identify genetic variants associated with leaf reflectance spectra (ranging from 400 to 2500 nm) in the RILs:

1. Single Wavelength GWAS (SW-GWAS): This approach treats each wavelength as a separate phenotype and runs a GWAS on each wavelength. This allows for the identification of genetic variants associated with specific wavelengths of the leaf reflectance spectra.

2. Hierarchical Spectral Clustering with Parallel Analysis GWAS (HSC-PA-GWAS): This approach, adapted from a method for GWAS on human facial shapes (Claes et al., 2018; Sero et al., 2019), reduces the dimensionality of the data by clustering the spectra into segments based on their similarity. The first principal component (PC) of each segment, which captures the majority of the variation within the segment, was then used as a phenotype for GWAS. The number of segments and the number of PCs to retain for each segment were determined using Parallel Analysis (PA), a statistical method that compares the observed eigenvalues with those obtained from random data (Franklin, Gibson, Robertson, Pohlmann and Fralish, 1995; Hayton, Allen and Scarpello, 2004).

2.5.3. Software and setting A series of GWA studies were conducted using the Genome Association and Prediction Integrated Tool (GAPIT) version 3 (Wang and Zhang, 2021) using the following models:

1. Generalized Linear Model (GLM) (Price et al., 2006): A model that captures the linear relationship between genetic markers and phenotypic traits using a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. GLM does not incorporate kinship.

2. Mixed Linear Model (MLM) (Yu, Pressoir, Briggs, Vroh Bi, Yamasaki, Doebley, McMullen, Gaut, Nielsen, Holland et al., 2006): Extends the GLM by incorporating both fixed effects of population structure and random effects of kinship to control for spurious associations.

3. Fixed and random model Circulating Probability Unification (FarmCPU) (Liu, Huang, Fan, Buckler and Zhang, 2016): Iteratively fits fixed and random effects to improve the power and precision of GWAS.FarmCPU eliminates the confounding effect of kinship by utilizing a fixed-effect model. The kinship derived from the associated markers is then used to select the associated markers using the maximum likelihood method. This method effectively overcomes the

problems of model overfitting that arise with stepwise regression.

4. Bayesian information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK) (Huang, Liu, Zhou, Summers and Zhang, 2019): Iteratively incorporates associated markers as covariates to eliminate their connection to the cryptic relationship among individuals. The associated markers are selected according to linkage disequilibrium, optimized for Bayesian information content, and reexamined across multiple tests to reduce false negatives (Wang and Zhang, 2021).

Testing multiple methods rather than choosing one helps to assess our association procedures. Different models have varying strengths depending on the nature of the data and the underlying genetic architecture of the trait being studied. By employing a suite of models, we aimed to cross-validate results and potentially uncover associations that might be missed by a single model. For the indices-GWAS and HSC-PA-GWAS, where the number of phenotypes was relatively small, all four models were applied. However, for the SW-GWAS, which involves 2101 phenotypes, running all four models would be computationally intensive. Given the promising higher statistical power of the BLINK model and the encouraging results from the HSC-PA-GWAS, we opted to exclusively use the BLINK model for the SW-GWAS.

We included the maternal line that generated each RIL as a covariate in all models to account for potential maternal effects, as suggested and done similarly in a previous study (Joseph, Corwin, Li, Atwell and Kliebenstein, 2013; Ray et al., 2023). We also included the number of measured leaves as a covariate to control for potential effects of the measurement, based on our previous study exploring some non-biological effects where the number of measured leaves was found to affect spectral regions that overlap with genetic affected regions (Li et al., 2023a).

2.5.4. Adjustment for multiple testing and significance thresholds In our analyses, we first employed an exploratory threshold of $p < 1 \times 10^{-5}$ to screen and select a subset of SNPs for further examination. P-values were also adjusted using a false discovery rate procedure (Benjamini and Hochberg, 1995) to account for the multiple testing on 183,942 SNPs and were indicated as the FDR adjusted p-values. Subsequently, to determine the statistically significant associations, we applied the M_{eff} method (Nyholt, 2004) to calculate the number of real independent tests in order to account for phenotypic correlation (testing multiple phenotypes within an approach). We then calculated the significance threshold as $(0.05/M_{eff})$. Any SNPs with FDR-adjusted p-values below this threshold were considered to have significant associations.

2.5.5. Selection of candidate genes To identify candidate genes linked to the significant SNPs, we first extracted all annotated genes within +/- 100 kb of the SNP from the *N. attenuata* reference genome (Ray et al., 2023). We then used the *N. attenuata* Data Hub (http://nadh.ice.mpg.de/NaDH/) to filter out genes for which there was no evidence of transcript accumulation in leaves under a variety of environmental conditions. We used annotation information to identify genes within this filtered list having pertinent functions.

3. Results

3.1. Phenotypic analysis

3.1.1. Spectral Characteristics of Leaves Our study focuses on calculated leaf reflectance spectra spanning from 400 to 2500 nm, including the visible (VIS), near infrared (NIR), and the shortwave infrared (SWIR) regions. This broad spectral range captured variation among the RILs, shown in Figure 2 as an orange shaded region indicating the range of reflectance spectra for all samples. The averaged spectral signature across all lines, evident as a mean profile, exhibited characteristic peaks and troughs, pointing to specific leaf properties such as chlorophyll absorption and water content. When assessing the variability across the samples using the coefficient of variation (CV), we observed fluctuations approximately between 0.05 and 0.2. Interestingly, the near-infrared (NIR) region showed more consistency in reflectance values among the RILs, contrasting with greater variability in the VIS and SWIR regions.

The distribution of six selected spectral indices 2.5.1, encapsulating attributes like water content and chlorophyll concentration, provided more specific information about leaf physiology. The indices aiming to quantify differences in chlorophyll A and B ($ARDSIC_{ab}$) and dry matter content ($ARDSIC_m$) had the narrowest range of values, while the chlorophyll index-red edge index (CI_{re}) had the largest range. The indices for leaf water content ($ARDSIC_w$), chlorophyll and carotenoids (CCI), and normalized differential water index ($NDWI$) showed an intermediate range of values.

Our subsequent application of the Hierarchical Spectral Clustering with Parallel Analysis (HSC-PA) segmented the spectra into 46 distinct patterns. The blank (NA) areas are regions where further clustering was halted due to the retention of a singular Principal Component (PC) from a previous level. This approach allowed for a finer granularity in the spectral data representation while accounting for correlations among wavelengths. The hierarchical clustering process concluded at different levels, ranging from level 3 to level 8, indicating the varying depths of spectral similarities among the segments, and Table 2 shows the detailed spectral region and potential associated features of the segment used in HSC-PA-GWAS. Supplementary table 1 shows the spectral regions of all 46 segments.

By visualizing the 24 segments that each retained only a single PC, we were able to span the entire spectrum from 400 to 2500 nm. These segments, parsed into several interpretable regions, were then used for HSC-PA-GWAS approach, offering invaluable phenotypic data for uncovering associations with underlying genetic variations.

Table 2. Spectral Ranges of Segments Used in GWAS with Potential Associated Features

Segment	Spectral ranges	Potentially associated physiological features
7	708-722	NIR: Leaf structure
8	723-735	NIR: Leaf structure
9	745-945, 1016-1128	NIR: Leaf structure
10	736-744, 946-1015, 1129-1143	NIR: Leaf structure
11	1312-1357	NIR: Water content
12	1144-1311	NIR: Leaf structure
16	1358-1383	NIR: Water content
18	1384-1385, 1592-1752	SWIR: Non-pigment organic composition (e.g. lignin, cellulose, starch, protein, sugar, oil)
19	1525-1564	SWIR: Non-pigment organic composition (e.g. lignin, cellulose, starch, protein, sugar, oil)
20	1386-1398, 1565-1591, 1753-1862	SWIR: Non-pigment organic composition (e.g. lignin, cellulose, starch, protein, sugar, oil)
23	521-604, 696-707	VIS: Chlorophyll a, chlorophyll b, carotenoids, anthocyanins
27	445-499	VIS: Chlorophyll a, chlorophyll b, carotenoids, anthocyanins
28	500-520	VIS: Chlorophyll a, chlorophyll b, carotenoids, anthocyanins
29	648-690	VIS: Chlorophyll a, chlorophyll b, carotenoids, anthocyanins
30	605-647, 691-695	VIS: Chlorophyll a, chlorophyll b, carotenoids, anthocyanins
33	1400-1484	Water content
34	1485-1524	SWIR: Non-pigment organic composition (e.g. lignin, cellulose, starch, protein, sugar, oil)
39	1399, 1863-1880, 2094-2310	NIR-SWIR: Leaf structure, non-pigment organic composition (e.g. lignin, cellulose, starch, protein, sugar, oil)
41	2414-2500	SWIR: Water content
42	2311-2413	SWIR: Non-pigment organic composition (e.g. lignin, cellulose, starch, protein, sugar, oil)
43	1894-2001	SWIR: Water content
44	1881-1893, 2002-2093	SWIR: Non-pigment organic composition (e.g. lignin, cellulose, starch, protein, sugar, oil)
45	423-444	VIS: Chlorophyll a, chlorophyll b, carotenoids, anthocyanins
46	400-422	VIS: Chlorophyll a, chlorophyll b, carotenoids, anthocyanins

Note: The associated features are based on common findings (Jacquemoud and Ustin, 2019) and are intended to indicate physiological properties of leaves known to affect the given spectral ranges. These interpretations are subject to various factors and should not be considered definitive. These do not include all traits demonstrated to be associated with different spectral features; for example, N, P, and phenolic content are commonly associated with several features across the spectrum based on partial least squares models (e.g. (Wang, Chlus, Geygan, Ye, Zheng, Singh, Couture, Cavender-Bares, Kruger and Townsend, 2020)).

3.1.2. Effective Number of Independent Tests (M_{eff}) To account for multiple testing due to the analysis of multiple phenotypes in our GWAS approaches, we determined the effective number of independent tests, denoted as M_{eff}. This analysis aids in adjusting for the correlation among phenotypes, providing a more accurate assessment of the significance of associations. For the Indices-GWAS approach, which employed six distinct spectral indices, the effective number of independent tests, $M_{eff\,ind}$, was determined to be 5. In the Single Wavelength GWAS (SW-GWAS) approach, the effective number, $M_{eff\,sw}$, was notably higher at 1016. Lastly, for the Hierarchical Spectral Clustering with Parallel Analysis GWAS (HSC-PA-GWAS) approach, which clusters the spectra based on similarity, the effective number, $M_{eff\,hsc-pa}$, was found to be 13, reflecting the reduced dimensionality and the focus on major patterns of variation. These calculated M_{eff} values are used to set a threshold for significant associations as described in section 2.5.4.

3.2. SNP markers analysis The kinship relationships among the RILs are visualized in Figure 3. Through a heatmap, the genetic relatedness between individuals is presented. Most values in the histogram gravitate towards zero, indicating a consistent and minimal genetic differentiation among the samples. This uniform distribution of kinship values affirms the thorough genetic mapping of the parental lines within the derived RILs from the MAGIC population and is consistent with the results of no obvious clustering of PCA of the RILs on the same samples (Ray et al., 2023).

Supplementary figure 1 provides additional analyses, including frequency of heterozygosity of individual and markers (a) - (c), and linkage disequilibrium (LD) decay over distance (d) - (i). For individual samples (a), the histogram exhibits a bell-like curve centered around 0.425 to 0.475. The marker heterozygosity distribution (b) reveals a pronounced peak in the initial bin at zero, emphasizing a certain number of markers with no heterozygosity. Beyond this peak, the remaining distribution is left-skewed with a peak around 0.5 heterozygosity. The minor allele frequency (MAF) (c) shows similar pattern, with a peak at around 0.25 - 0.3, and a smaller peak at 0. Continuing with the LD decay over distance, (d) illustrates the relationship between R and markers, where the

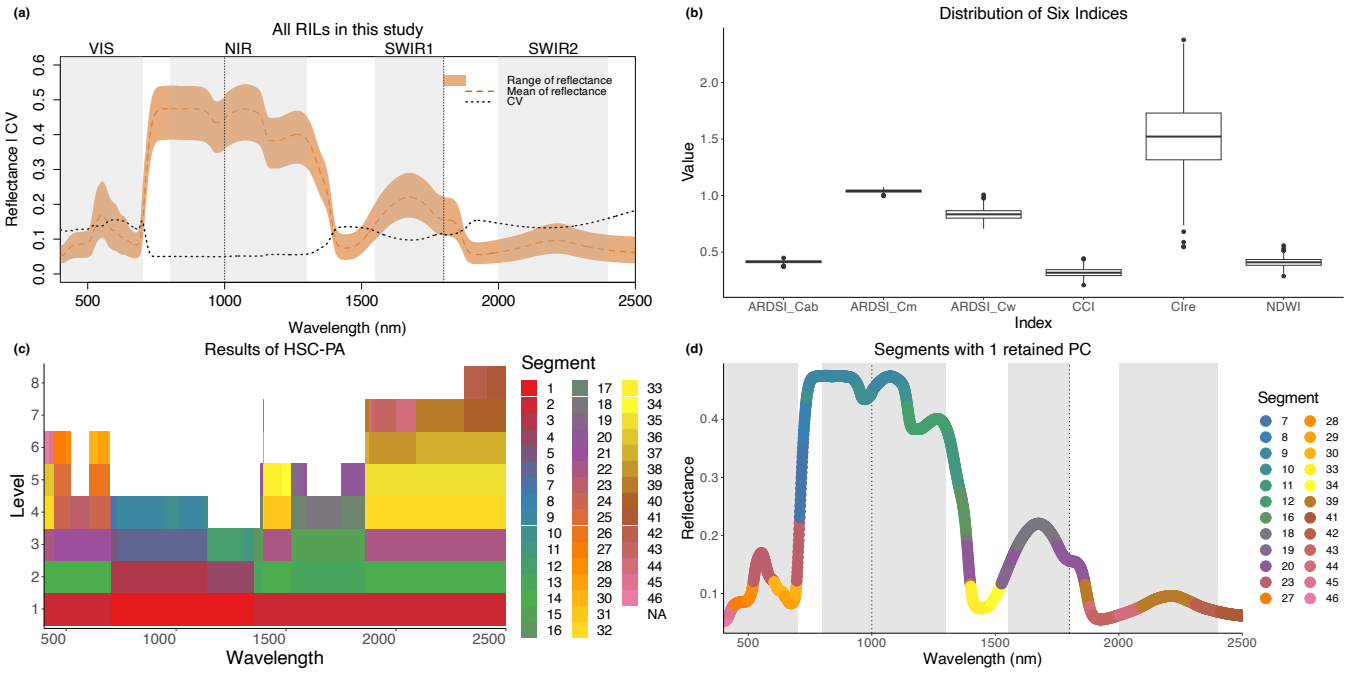

Fig. 2. Spectra as phenotypes. (a) Leaf reflectance spectra for all Recombinant Inbred Lines (RILs) examined in this study. The orange shaded region illustrates the range of spectra, with the dashed orange line depicting the mean reflectance across all samples. The black dotted line represents the coefficient of variation for the entire sample set. (b) Distribution of the six spectral indices utilized in the Indices-GWAS approach. (c) Results from the Hierarchical Spectral Clustering with Parallel Analysis (HSC-PA). A total of 46 segments, differentiated by color, emerged from the clustering. Regions left blank (NA) signify segments that retained one Principal Component (PC) in the prior level, halting further clustering. (d) Presentation of the 24 segments, shaped as example spectra, each maintaining one retained Principal Component (PC). Spanning the entire spectrum from 400 to 2500 nm, these segments served as phenotypes in the HSC-PA-GWAS approach to pinpoint associations with genetic variances.

majority of points are concentrated in the range of -0.25 to 0.8. Figure (e) presents a bell-shaped frequency distribution for R, with a noticeable peak at 0 and a clear higher density on the right side. Figure (f) depicts the relationship between R and distance in kilobases (kb), showing a balanced distribution on both sides of the R=0 line, with most points at a distance of 0 kb. Figure (g) displays the distribution of distances across markers, revealing a dense cluster of points at distances between 0-1 kb that gradually becomes sparser as the distance increases. Figure (h) shows the frequency distribution for distances in kb, with a strong peak in the 0-1 kb bin. Figure (i) outlines the decay of LD as a function of distance in kb, with a smoothed curve indicating a rapid decay till around 0.2 kb.

3.3. Genome-Wide Association Studies (GWAS) Results

3.3.1. Indices-GWAS results A series of GWAS analyses were conducted to uncover potential genetic associations with six spectral indices: $NDWI$, CCI, CI_{re}, $ARDSI\,C_{ab}$, $ARDSI\,C_w$, and $ARDSI\,C_m$. Four models were employed in the analysis: GLM, MLM, FarmCPU, and BLINK. The Manhattan plot, presented in Figure 4a, showcases the associations between the SNPs and the 6 indices using different models. Several genomic regions displayed distinct associations, as signified by the $-\log_{10}$(p-values). An exploratory threshold, defined at 1×10^{-5} (denoted by the green horizontal line), was adopted to shortlist SNPs for a detailed examination.

In Figure 4b, each of the 6 indices is plotted against the SNPs that exceeded the exploratory threshold. The gradient color scale indicates the association's intensity, with deeper hues signifying more significant p-values. This representation furnishes a concise overview of the genomic regions strongly correlated with each spectral index. In total, 19 SNPs exhibited p-values less than 1×10^{-5}: six with CCI, five with $ARDSI\,C_w$, three each with $NDWI$ and $ARDSI\,C_m$, and one each with CI_{re} and $ARDSI\,C_{ab}$. The index CCI displayed the associations with the lowest p-values across the board. Thus, CCI was chosen as a representative example for the QQ plot in Figure 4c. This plot serves as a diagnostic tool for the GWAS results. Each model is represented by a unique color: red for GLM, yellow for MLM, green for FarmCPU, and blue for BLINK. In this example, the BLINK, FarmCPU, and GLM methods all yielded nearly identical results. The alignment of the observed points with the diagonal line indicates that the GWAS results are mostly consistent with the expectations under the null hypothesis. However, deviations can pinpoint regions with stronger genetic signals. The significance level adjusted by M_{eff} is 0.01, and none of the SNPs surpassed this threshold.

3.3.2. SW-GWAS and HSC-PA-GWAS results In the SW-GWAS analysis, the BLINK model was used to investigate the genetic correlations with 2101 spectral wavelengths spanning from 400 to 2500 nm. Figure 5 highlights the genetic associations for the spectral wavelength at 1160 nm, which displayed the strongest association among all phenotypes, serving as a representative example. The diagnostic QQ plot for this wavelength is also presented in Figure 5c. Figure 5b of-

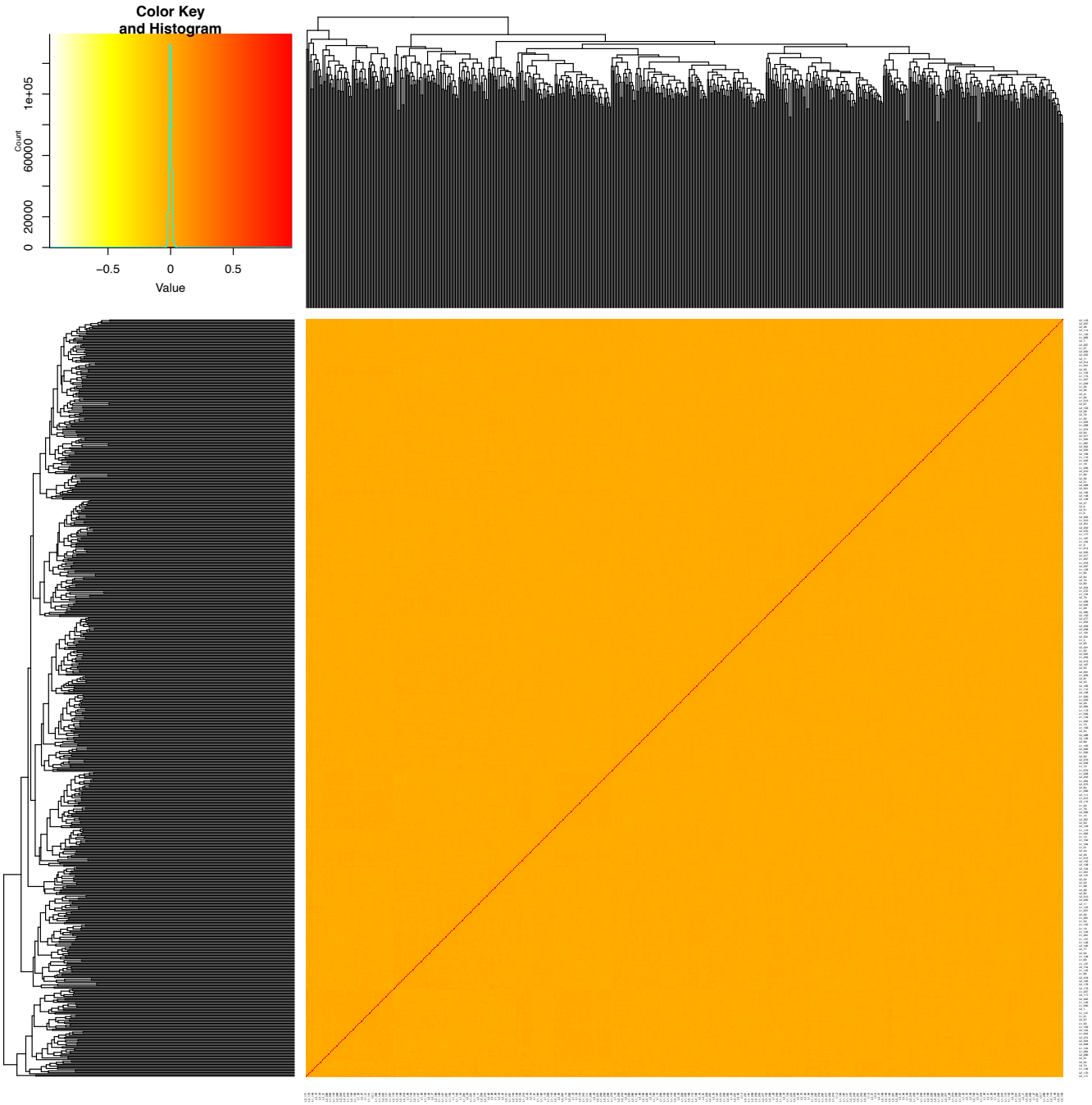

Fig. 3. A heatmap of the kinship matrix indicating the relationship between individuals.

fers an in-depth depiction of associations that exceeded the exploratory threshold. The X-axis signifies the spectral wavelengths, while the Y-axis indicates the associated SNPs, providing a holistic view of the spectral regions correlated with each SNP. A total of 48 SNPs are displayed, with the majority linked to one or multiple continuous spectral regions. This overlap indicates the high correlation inherent in the leaf reflectance spectra data, underscoring the necessity of methods accounting for these correlations. The associations span nearly the entire spectral range and are distributed across almost all chromosomes. The significance level, adjusted by M_{eff}, stands at 4.92×10^{-5}. Notably, no SNP met this stringent threshold.

In the HSC-PA-GWAS, we once again utilized the four models—GLM, MLM, FarmCPU, and BLINK—to probe the genetic associations across 24 delineated segments. The Manhattan plot in Figure 6a summarizes all of these results. Delving deeper, Figure 6b shows the associations that surpassed our exploratory threshold of 1×10^{-5}. A total of 34 SNP associations distributed over 23 segments emerged from the analysis, all of which were also found in the SW-GWAS. Focusing on segment associations obscures spectral information, and so Figure 6c maps these associations back to their corresponding spectral wavelengths for better interpretation. This visualization facilitates the identification of specific spectral regions linked to each SNP, enhancing our

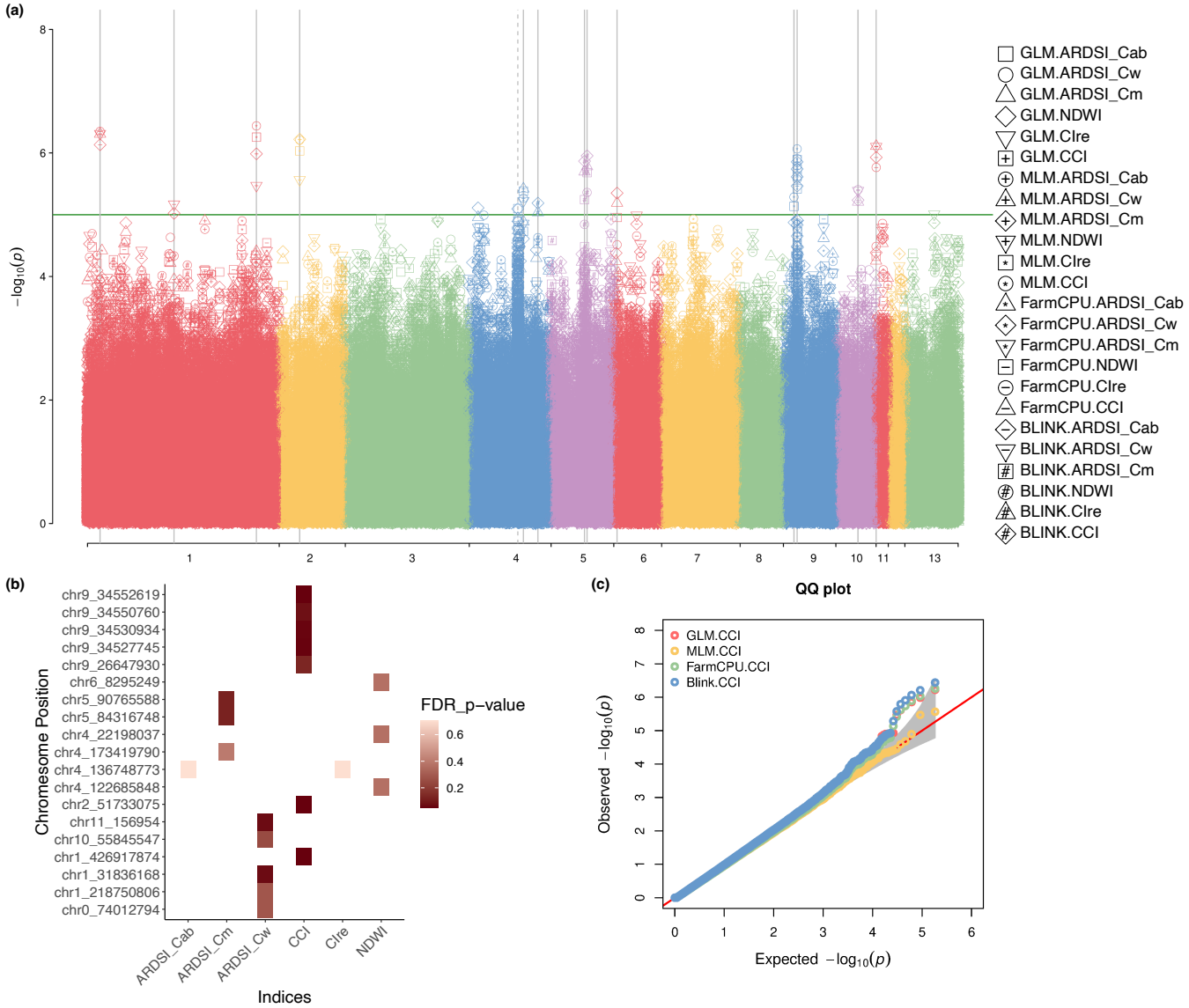

Fig. 4. Indices-GWAS results. (a) Manhattan plot showcasing the $-\log_{10}$(p-values) of SNP associations for the six indices across various models. The horizontal line signifies a significance threshold of 1×10^{-5}, which is used to select associations to generate figure (b). (b) Visualization of associations that have p-values less than 1×10^{-5}. The x-axis enumerates the six indices, while the y-axis displays the corresponding associated SNPs. The gradient red color scale represents the FDR˙adjusted p-values. (c) QQ plot contrasting the observed versus the expected $-\log_{10}$(p-values) for the GWAS analysis on the CCI index. The diagonal line represents the expected distribution under the null hypothesis. The different color schemes depict the models used: red for GLM, yellow for MLM, green for FarmCPU, and blue for BLINK.

comprehension of the associations. It is noteworthy that these associations span a range of chromosomal regions, indicating a complex genetic basis of the studied traits. The significance level, after adjusting for M_{eff}, is set at 3.85×10^{-3}. A singular association, the SNP chr1_156565900 with Segment 27 (445-499 nm), surpassed this threshold. The QQ plot dedicated to Segment 27, as depicted in Figure 6d, serves a dual purpose of validation and diagnostic assessment. Figure 6e delves into the phenotypic distribution of this significant marker (chr1_156565900) affiliated with Segment 27. In this distribution, the numbers 0, 1, and 2 represent homozygous alleles as the reference, heterozygous alleles, and homozygous alternative alleles, respectively. The majority of the RILs display heterozygous alleles at this site, as evidenced by the greater number of points in classes 1 and 2. A clear difference in the mean values of these three classes is evident in the boxplot, warranting further investigation to understand the underlying genetic framework.

To further elucidate the genetic association indicated by HSC-PA-GWAS, we searched for potential candidate genes within a +/- 100 kb window surrounding the SNP chr1_156565900. This led to the identification of two genes of interest: Niat3g_07150 and Niat3g_07151 (more annotation information can be found in supplementary table 2). The former, Niat3g_07150, is annotated as a serine hydroxymethyltransferase (*SHMT*). *SHMT* plays a pivotal role in one-carbon metabolism, catalyzing the reversible conversion of serine and tetrahydrofolate (THF) to glycine and 5,10-methylene-THF (Ravanel, Cherest, Jabrin, Grunwald, Surdin-Kerjan, Douce and Rébeillé, 2001). This enzymatic reaction is integral to various cellular processes, including nucleotide synthesis and amino acid homeostasis (Hanson

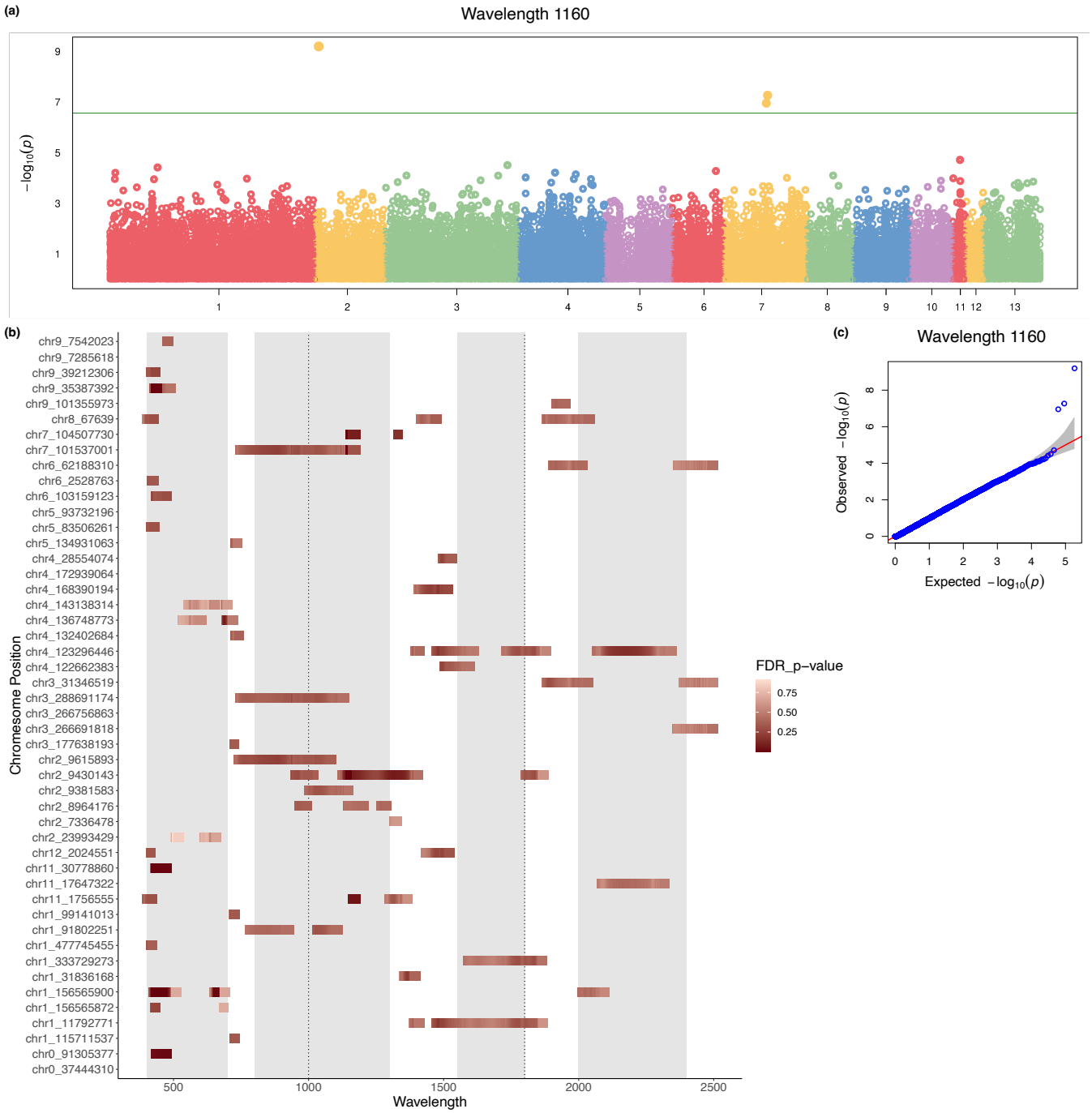

Fig. 5. SW-GWAS results. (a) Manhattan plot showcasing the $-\log_{10}$(p-values) of SNP associations with BLINK for wavelength 1160 as a representative example due to its strongest association. The horizontal line signifies a significance threshold of 1×10^{-5}, which is used to select associations to generate figure (b). (b) Visualization of associations that have p-values less than 1×10^{-5}. The x-axis represents the spectral wavelengths from 400 to 2500 nm, while the y-axis displays the corresponding associated SNPs. The gradient red color scale represents the FDR˙adjusted p-values. (c) QQ plot contrasting the observed versus the expected $-\log_{10}$(p-values) for the GWAS analysis on the wavelength 1160. The diagonal line represents the expected distribution under the null hypothesis. The blue color scheme highlights the BLINK model used for the analysis.

and Roje, 2001). In plants, *SHMT* has been implicated in photorespiration, where it facilitates the recycling of carbon skeletons (Somerville and Ogren, 1981). The latter gene, Niat3g_07151, is characterized by a Methyltransf_11 domain. Proteins with this domain are typically involved in methylation as a post-translational modification that modulates protein function (Bannister and Kouzarides, 2011). Methylation processes are crucial for myriad cellular activities, ranging from gene expression regulation to protein-protein interactions (Law and Jacobsen, 2010). The presence of this domain suggests that Niat3g_07151 might be involved in such regulatory mechanisms, although its specific role in *N. attenuata* remains to be elucidated.

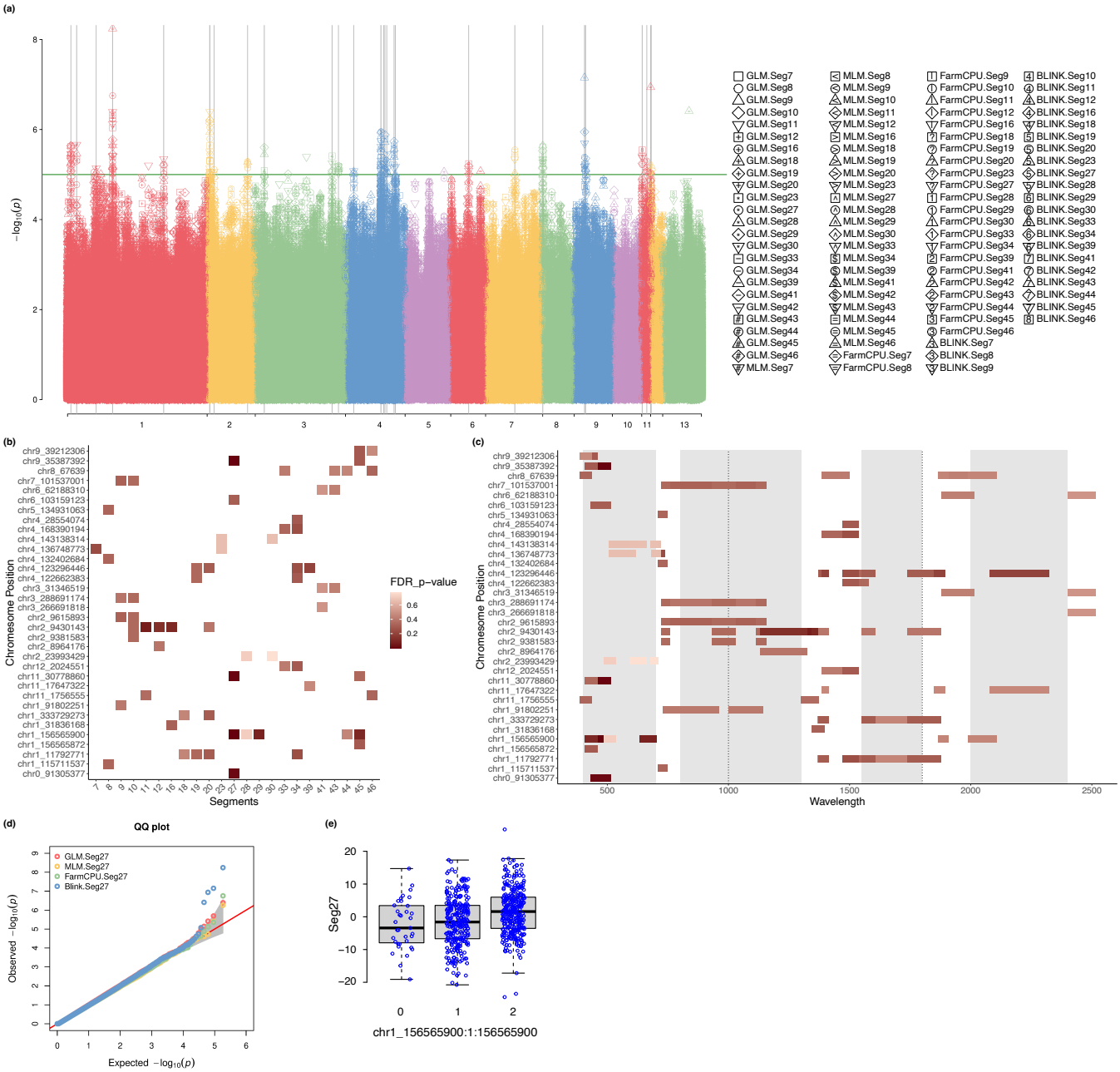

Fig. 6. HSC-PA-GWAS results. (a) Manhattan plot illustrating the genetic associations for the 24 segments using the four models: GLM, MLM, FarmCPU, and BLINK. The horizontal green line represents the exploratory threshold $-\log_{10}$(p-values). (b) Detailed visualization of associations that surpass the exploratory threshold. The X-axis enumerates the 24 segments, while the Y-axis displays the corresponding associated SNPs. (c) An alternative representation of the associations from (b) with the X-axis showcasing the spectral wavelengths, providing insights into the specific spectral regions of the associations. (d) QQ plot highlighting the observed versus expected $-\log_{10}$(p-values) for GWAS analysis on Segment 27, the segment that meets the stringent significance threshold. (e) Phenotypic distribution for the significant marker (chr1˙156565900) in association with Segment 27. 0 represent a homozygous allele from the reference, 1 a heterozygous allele, and 2 a homozygous alternative allele.

4. Discussion

4.1. HSC-PA-GWAS for discovering genetic associations with spectral variation Understanding the genetic foundation of leaf spectral diversity will facilitate advances in biodiversity monitoring and phenotyping, and determining appropriate approaches to handle complex, information-rich spectral phenotypes is a critical step. Although Single-Wavelength GWAS (SW-GWAS) proved beneficial in certain circumstances and was straightforward as a starting point, the high dimensionality of spectral data renders this approach very low-power, and it is not accurate to treat every individual wavelength as an independent variable. Common dimensionality reduction approaches such as PCA, when applied to reflectance spectra, can yield complex components that cannot be simply interpreted in terms of either spectral features or biology (see e.g. PCA loadings in (Li et al., 2023a)). Common approaches from trait-based analyses such as varimax rotation (Weigelt, Mommer, Andraczek, Iversen, Bergmann, Bruelheide, Freschet, Guerrero-Ramírez, Kattge, Kuyper, Laughlin, Meier, van der Plas, Poorter, Roumet, van Ruijven,

Sabatini, Semchenko, Sweeney, Valverde-Barrantes, York and McCormack, 2023) are not straightforward to apply, because spectra are continuous and features commonly have multiple influences, and thus cannot be simply divided into discrete traits on which alignment could be conducted. Spectral indices, a well-established approach to "extract" traits from spectra, retain only a small portion of the total information in leaf reflectance spectra. It is important to note that the interpretation of these indices can differ based on species and other factors, and require validation by comparison with the target traits measured by other methods. (Given that the study of spectral-genetic associations is still in its nascent stages, we included a GWAS on commonly used indices to support initial interpretations, but we do not interpret them directly in terms of biochemical or water content.) Partial least squares regression (PLSR) is commonly used as an alternative to spectral indices for deriving data-driven trait associations with features across entire reflectance spectra, but requires orthogonal measurements of the targeted traits to determine associations (e.g.Wang et al. (2020)). Here, we developed HSC-PA-GWAS as a new approach that retains the information in spectra in a biologically interpretable form without losing meaningful information, and without requiring an *a priori* decision to target specific traits, while maintaining sufficient statistical power to identify potentially meaningful significant associations.

The Hierarchical Spectral Clustering with Parallel Analysis (HSC-PA) GWAS method introduced in this study stands out for its combination of sensitivity and statistical power when handling spectral phenotypes. Associations identified using HSC-PA-GWAS were consistent with the Single-Wavelength GWAS (SW-GWAS) method, as every SNP identified by HSC-PA-GWAS was also detected by SW-GWAS. However, the HSC-PA method surpasses SW-GWAS in several key aspects. One of its strengths is its ability to reduce data dimensionality while accounting for phenotypic correlations within spectra, by transforming multiple correlated wavelengths into a single value, typically represented by the first principal component. This not only mitigates the challenges of multiple testing and correlated phenotypic data but also increases statistical power. By aggregating information across multiple wavelengths, the HSC-PA method offers a more comprehensive and potentially more accurate representation of the underlying traits with its ability to discover the correlation structure among wavelengths from data. The resulting genetic associations are likely more informative than those discovered when parsing spectra into either indices, which cannot encompass all wavelengths influenced by an underlying trait; or single wavelengths, which do not behave independently. Comparing the HSC-PA method with established methods, such as an approach combining partial least squares regression (PLSR) and linear regression (Verrelst, Malenovskỳ, Van der Tol, Camps-Valls, Gastellu-Etchegorry, Lewis, North and Moreno, 2019), indicates that HSC-PA has an advantage in at least two ways: supporting human interpretation of data-derived spectral features and thus genetic associations, and allowing for experimental designs aimed at discovering associations with novel trait variation or where there are limitations on producing appropriate validation datasets required for representing traits via PLSR. Verrelst and colleagues (Verrelst et al., 2019) also emphasized the importance of nonparametric regression and machine learning methods, like the Kernel Ridge Regression (KRR) and Gaussian Processes Regression (GPR), for their ability to capture nonlinear relationships in spectral data. These methods, while powerful, often require careful tuning and can be computationally intensive. In contrast, the HSC-PA method offers a data-driven yet human-interpretable approach. As discussed in studies on human facial shapes, from which the HSC-PA method was adapted (Claes et al., 2018), this methodological shift allows for a transition from a global to a local understanding of spectral variation, and a holistic view of its genetic basis.

4.2. Ecological significance of findings in *N. attenuata*

The discovery of a significant association between the SNP chr1_156565900 and Segment 27 (445-499 nm) in the HSC-PA-GWAS is an important finding from our study. This spectral range, 445-499 nm, encompasses key regions of the light spectrum absorbed by chlorophyll, the primary pigment involved in photosynthesis, and by cryptochrome photoreceptors. Chlorophyll predominantly absorbs light in the blue (430-450 nm) and red (640-680 nm) regions (Katz, 1973; Taiz, Zeiger, Møller, Murphy et al., 2015), with the blue range overlapping with our identified segment. The reported action spectrum for plant cryptochrome is around 365-550 nm with peak activity in response to light at 450 nm (Ahmad, Grancher, Heil, Black, Giovani, Galland and Lardemer, 2002; Li, Wang, Yu, Liu, Yang, Zhao, Liu, Tan, Klejnot, Zhong et al., 2011).

Two interesting candidate genes were identified within 100 kb of this SNP. Niat3g_07150, annotated as a serine hydroxymethyltransferase, catalyzes an important step in photorespiration, a process intricately linked with photosynthesis (Florio, di Salvo, Vivoli and Contestabile, 2011). Photorespiration is essential for plants, especially under conditions where the concentration of carbon dioxide is low (Dusenge, Duarte and Way, 2019). The enzyme serine hydroxymethyltransferase converts serine to glycine, a process crucial for the photorespiration-mediated regeneration of ribulose-1,5-bisphosphate, the substrate for the carbon-fixing enzyme ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO), thereby facilitating photosynthesis (McFadden, 1980; Moreno, Martín and Castresana, 2005). This enzyme has been found to have decreased activity in the progression from C3 to C4 photosynthesis, indicating its significant role in the photorespiratory pathway (Ku, Wu, Dai, Scott, Chu and Edwards, 1991). Moreover, studies have shown that mutations affecting this enzyme can have detrimental effects on photorespiration and photosynthesis, particularly under conditions promoting photorespiration (Somerville and Ogren, 1980). The second candidate gene, Niat3g_07151, annotated as a methyltransf_11 domain-containing protein, may be involved in various plant metabolic processes. While its direct link to photosynthe-

sis or spectral absorption is not immediately evident, methyltransferases function in diverse biological processes including gene expression, protein function, and metabolic reactions (Sarabi and Naghibalhossaini, 2015; Rao, 2020). Interestingly, lines of *N. attenuata* knocked down in RuBPCase activase (RCA) expression were shown to have elevated jasmonate methyltransferase (JMT) activity: these plants accumulated greater amounts of methyl jasmonate than wild-type *N. attenuata* plants, resulting in jasmonoyl-isoleucine (JA-Ile) deficiency and reduced defense responses to herbivory (Mitra and Baldwin, 2014). The RCA knockdown lines also displayed 20% lower C assimilation and ca. 20-40% lower photosynthetic rates, and reduced growth in comparison to the wild-type (Mitra and Baldwin, 2008).

In the context of *N. attenuata*, a wild plant that has evolved myriad strategies to cope with its natural environment, understanding the genetic basis of spectral variation can provide insights into its adaptive mechanisms. The association of the SNP with a spectral range crucial for photosynthesis and blue light sensing might hint at evolutionary pressures that have shaped the genetic makeup of this species, optimizing its photosynthetic efficiency and, by extension, its survival in its native habitat.

4.3. Advantages of the MAGIC design for discovering genetic associations The Multi-parent Advanced Generation Inter-Cross (MAGIC) design has emerged as a powerful tool in the realm of genetic association studies, particularly in Genome-Wide Association Studies (GWAS) (Bandillo, Raghavan, Muyco, Sevilla, Lobina, Dilla-Ermita, Tung, McCouch, Thomson, Mauleon et al., 2013; Mackay, Bansept-Basler, Barber, Bentley, Cockram, Gosman, Greenland, Horsnell, Howells, O'Sullivan et al., 2014; Sannemann, Huang, Mathew and Léon, 2015; Huynh, Ehlers, Huang, Muñoz-Amatriaín, Lonardi, Santos, Ndeve, Batieno, Boukar, Cisse et al., 2018). The generation of these intricate genetic mapping populations represents a substantial investment, but they are an important precision tool to complement ecological studies of population genetic variation. The unique structure of MAGIC populations offers several advantages over traditional biparental genetic mapping populations or the analysis of natural populations.

1. Mitigation of population structure: one of the primary challenges in GWAS is the confounding effect of population structure, which can lead to spurious associations. In the context of our study, the heatmap of kinship, as depicted in Figure 3d, showcases an even pattern without discernible clusters. This uniformity indicates the absence of population structure, a testament to the strength of the MAGIC design. By intercrossing multiple founder lines over several generations, MAGIC populations effectively dilute the population structure, reducing the risk of false positives in association studies.

2. Increased genetic diversity: MAGIC populations amalgamate the genetic diversity from multiple parental lines. This increased diversity ensures a broader representation of alleles, enhancing the resolution of GWAS and increasing the chances of identifying rare or novel alleles associated with traits of interest.

3. More recombination events: due to the multiple rounds of intercrossing, MAGIC populations experience a higher number of recombination events compared to biparental crosses. This leads to finer mapping of quantitative trait loci (QTL), allowing for more precise identification of genomic regions associated with traits.

4. Controlled parental environment: conducting multi-generational breeding under controlled conditions in a MAGIC design allows environmental variables to be kept consistent among all lines throughout the breeding process, allowing for a clearer interpretation of genotype-phenotype associations in light of possible epigenetic effects.

5. Flexibility in analysis: the diverse genetic backgrounds in MAGIC populations allow for both linkage and association analyses, offering flexibility in genetic mapping approaches.

4.4. Limitations and outlook One of the primary limitations of our study lies in the phenotypic interpretation for biological meanings specific to the samples we used. The six spectral indices we selected are derived from existing literature and have not been validated in *N. attenuata* generally, or for this dataset. Without chemical or other analyses to specifically quantify the constituents represented in the indices, such as chlorophyll, our interpretations remain speculative. Moreover, while the HSC-PA-GWAS approach has identified potential candidate genes, any influence of these genes on the traits of interest remains to be verified, e.g. through the generation and testing of knock-out, knock-down, and overexpression lines.

One noteworthy aspect of our study is the unexpectedly high levels of heterozygosity (mean at around 0.5, see supplementary figure 1 (a)) observed in the dataset. While the low-pass sequencing (0.5x coverage) of the RILs (Ray et al., 2023) in our study could be a contributing factor to the underestimation of homozygosity, it is important to consider the role of residual heterozygosity and heterozygote hotspots. Drawing parallels from maize, a study by Liu et al. (Liu, Liu, Li, Pan, Liu, Yang, Yan and Xiao, 2018) identified specific genomic regions, known as residual heterozygosity (RH) hotspots, that are more prone to maintaining heterozygosity. These hotspots are often subject to selection during the development of maize populations, suggesting that they may confer some adaptive or agronomic advantages. The presence of such RH hotspots in our study species, *N. attenuata*, could similarly indicate regions of the genome that are under selective pressure, possibly due to their role in adaptive traits such as photosynthetic efficiency or stress tolerance. In light of these considerations, future studies employing higher sequencing depth could provide a more accurate estimation of homozygosity and heterozygosity in *N. attenuata*. Such

efforts would be invaluable for dissecting the genetic architecture underlying these observations and for validating the robustness of our GWAS findings.

In this study, we have utilized the GAPIT3 tool for GWAS, which is widely recognized as a robust approach. However, we acknowledge that the potential of novel computational GWAS methodologies are not fully captured in this tool. For example, GAPIT3 typically focuses on results for individual SNPs. Approaches allowing for pleiotropy have the potential to increase the power, sensitivity, and meaning of GWAS, similarly as the HSC-PA approach allowed us to obtain more phenotypic information by accounting for correlations in spectral data. As the field of genetics continues to evolve, it is conceivable that advanced techniques, such as machine learning and deep learning (Libbrecht and Noble, 2015; Wu, Karhade, Pillai, Jiang, Huang, Li, Cho, Roach, Li and Divaris, 2021), could offer more meaningful insights into genetic associations. Previous studies have used random forest models to allow for more complex and interactive effects of individual genetic variants on binary and even multivariate phenotypes, although the interpretation of "significance" in associations is not as straightforward with these models (Brieuc, Waters, Drinan and Naish, 2018; Wang, Goh, Wong, Montana and the Alzheimer's Disease Neuroimaging Initiative, 2013). Recently, Saha and colleagues described the "Epi-MEIF" approach using conditional inference forests in place of false discovery rates to allow for locus interactions (epistasis) in the genetic control of complex phenotypes (Saha, Perrin, Röder, Brun and Spinelli, 2022). This may provide a more informed way to delineate significant associations while accounting for testing of multiple loci, and still allowing for continued use of well-established "single-locus" association models such as those implemented in GAPIT.

Looking forward, there is great potential for deepening our understanding of leaf spectral data using the *N. attenuata* system. More than 400 transgenic lines and two genetic mapping populations, including the MAGIC population used here, were developed over decades spent developing this system. As a model system for chemical ecology, metabolomes and transcriptomes from inbred lines of *N. attenuata*, including some of the genotypes used in the MAGIC population, have been extensively characterized under a variety of environmental conditions and in response to different stressors and could support interpretation of the spectral data and help to connect to plausible genetic mechanisms. A recently published transcriptional and metabolomic characterization of MAGIC parental lines before and after different types of experimental herbivory, coupled with investigations of the MAGIC RILs and targeted knock-downs under field and glasshouse conditions, recently identified regulatory mechanisms that may buffer trade-offs and permit plants flexibility in the evolution of growth and defense strategies Ray et al. (2023). This is consistent with the outcome of our analysis, which, perhaps surprisingly, indicates relatively large variance in photosynthetic traits among the wild genotypes contributing to the *N. attenuata* MAGIC population, as discussed previously (Li et al., 2023a). Thus, while photosynthesis is a trait so conserved among plants that it is almost definitive (with the exception of the few parasitic plants that no longer depend on photosynthesis for energy), and is associated with broadly conserved leaf spectral features related to pigment absorbance (Meireles, Cavender-Bares, Townsend, Ustin, Gamon, Schweiger, Schaepman, Asner, Martin, Singh, Schrodt, Chlus and O'Meara, 2020), its expression nevertheless varies substantially even within a single plant species in the context of growth and defense strategies. This variation, likely mediated by variants of regulatory genes embedded in coexpression networks Ray et al. (2023), may lend different genotypes a competitive advantage under changing environmental conditions. The significant associations identified here thus indicate genotype and phenotype targets for novel and interesting in-depth functional studies that could help us to better understand plant survival strategies. A combination of natural variants, genetic modification, and laboratory assays, such as gene expression analyses (Schena, Shalon, Davis and Brown, 1995), protein-protein interactions (Fields and Song, 1989) and chromatin immunoprecipitation assays (Solomon, Larsen and Varshavsky, 1988), comprise the gold standard for functional testing and validation of discovered genetic associations. At this stage, the associations reported here represent hypotheses that should be subjected to these tests.

The broader implications of our research extend beyond its immediate findings. Unraveling the genetic basis of leaf spectral properties has far-reaching implications and could for example revolutionize plant biodiversity monitoring, ecophysiology, and adaptation research, and help to identify genotypes resilient to stresses associated with global change. Our study sets the stage for future research aimed at deciphering these links by presenting a widely applicable new method for data-driven discovery of associations which can be integrated with existing and novel tools to discover genetic associations and move from association to causation or prediction.

5. Conclusions

We tested several spectral data treatments and GWAS approaches to decipher the relationship between genetic variation and leaf spectral properties in a genetic mapping population of the ecological model plant *N. attenuata* in a field experiment. We examined calculated leaf reflectance spectra from 400 to 2500 nm, characterized more and less variable spectral regions, and linked these to features indicative of leaf composition and function. To tackle the high dimensionality and correlation structure inherent in spectral data, we developed the Hierarchical Spectral Clustering with Parallel Analysis (HSC-PA) method, which proved to be more sensitive than working with either spectral indices or individual wavelengths. Resulting associations overlapped with results obtained from a "brute-force" Single-Wavelength GWAS (SW-GWAS) approach and provided greater statistical power while maintaining interpretable spectral features even with reduced dimensionality, and accounting for internal correlations in the spectral data. The HSC-PA approach has advan-

tages for discovering candidate genetic associations in comparison to other data dimensionality reduction methods such as PCA, because it yields more interpretable features, as well as in comparison to other data-driven feature discovery methods such as PLSR, because it does not require independent measurements to be made on all traits of interest to allow feature discovery.

The identification of a significant correlation within the 445-499 nm spectral range, overlapping with the region of blue light absorption by chlorophyll and encompassing the activation spectrum for cryptochrome photoreceptors, indicated a locus which could be related to photosynthetic efficiency and perhaps management of growth and defense strategies, which have been shown to vary in distinct yet complex ways in the pioneer species *N. attenuata* (Ray et al., 2023). The significantly associated SNP is situated near the Niat3g_07150 gene, which is annotated as a serine hydroxymethyltransferase: an enzyme which plays a crucial role in photorespiration, a process intimately linked with photosynthesis. Another nearby gene, Niat3g_07151, which possesses a methyltransf_11 domain, is an additional interesting candidate for explaining the apparent influence of this locus on leaf spectral characteristics. The causal relationship of these candidates to variation in leaf blue light reflectance, if any, requires further testing. It would be interesting if the foliar reflectance of blue light could thus be linked to plant growth and defense strategies through mechanisms not directly related to leaf chlorophyll content.

This study introduces a method and framework that can be broadly applied to attain deeper understanding of the genetic underpinnings of leaf spectral properties. The associations identified here using an advanced genetic toolset built for a native plant, and studied in a field environment within the plant's natural range, have the potential to advance our understanding of the genetic mechanisms underlying photosynthetic efficacy and related growth and defense strategies in plants, and could support new insights in the application of remote sensing to biodiversity assessment.

6. Data and Code Availability

All plant lines are available from the Max Planck Institute for Chemical Ecology. The spectral measurement data underlying the results presented in this paper are available as a published dataset (Li, Czyż, Halitschke, Baldwin, Schaepman and Schuman, 2023b). The MAGIC parents and RIL WGS are available under BioProject id PRJNA907539 (Ray et al., 2022). All processed spectral data, genotypes data, and codes are provided at the GitHub repository: https://github.com/licheng1221/Association-study-of-genetic-variation-and-spectroscopic-imaging-variant

7. Author's contributions

CRediT taxonomy roles are listed with authors in alphabetical order. Conceptualization: I.T.B., M.C.S., M.E.S. Data curation: C.L., R.R. Formal analysis: C.L. Funding acquisition: I.T.B., M.E.S. Investigation: E.A.C., M.C.S. Methodology: C.L., E.A.C., M.C.S. Project administration: M.C.S. Resources: I.T.B., M.E.S., R.H. Supervision: M.C.S. Visualization: C.L. Writing—original draft: C.L. Writing—review and editing: C.L., E.A.C., I.T.B., M.C.S., M.E.S., R.H., R.R.

8. Acknowledgements

The authors acknowledge funding from the NOMIS Foundation and the University of Zurich, including the University Research Priority Program on Global Change and Biodiversity, and the Max Planck Society. We thank our division in the Department of Geography at the University of Zurich, the Remote Sensing Laboratories, for support and for the shared use of equipment. We thank the local managers at WCCER, N. Carlson and R. Carlson, and the 2019 field crews for plant cultivation and plot management. We thank the WCCER for supporting these experiments.

9. Bibliography

Adam, E., Mutanga, O., Rugege, D., 2010. Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation: a review. Wetlands ecology and management 18, 281–296.

Ahmad, M., Grancher, N., Heil, M., Black, R.C., Giovani, B., Galland, P., Lardemer, D., 2002. Action spectrum for cryptochrome-dependent hypocotyl growth inhibition in arabidopsis. Plant Physiology 129, 774–785.

Asner, G.P., Martin, R.E., Anderson, C.B., Knapp, D.E., 2015. Quantifying forest canopy traits: Imaging spectroscopy versus field survey. Remote Sensing of Environment 158, 15–27.

Asner, G.P., Martin, R.E., Carranza-Jiménez, L., Sinca, F., Tupayachi, R., Anderson, C.B., Martinez, P., 2014. Functional and biological diversity of foliar spectra in tree canopies throughout the andes to amazon region. New Phytologist 204, 127–139.

Bahulikar, R.A., Stanculescu, D., Preston, C.A., Baldwin, I.T., 2004. Issr and aflp analysis of the temporal and spatial population structure of the post-fire annual, nicotiana attenuata, in sw utah. BMC ecology 4, 1–13.

Bandillo, N., Raghavan, C., Muyco, P.A., Sevilla, M.A.L., Lobina, I.T., Dilla-Ermita, C.J., Tung, C.W., McCouch, S., Thomson, M., Mauleon, R., et al., 2013. Multi-parent advanced generation inter-cross (magic) populations in rice: progress and potential for genetics research and breeding. Rice 6, 1–15.

Bannister, A.J., Kouzarides, T., 2011. Regulation of chromatin by histone modifications. Cell research 21, 381–395.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) 57, 289–300.

Brieuc, M.S.O., Waters, C.D., Drinan, D.P., Naish, K.A., 2018. A practical introduction to Random Forest for genetic association studies in ecology and evolution. Molecular Ecology Resources 18, 755–766. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12773, doi:10.1111/1755-0998.12773. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.12773.

Cavender-Bares, J., Meireles, J.E., Couture, J.J., Kaproth, M.A., Kingdon, C.C., Singh, A., Serbin, S.P., Center, A., Zuniga, E., Pilz, G., et al., 2016. Associations of leaf spectra with genetic and phylogenetic variation in oaks: prospects for remote detection of biodiversity. Remote Sensing 8, 221.

Claes, P., Roosenboom, J., White, J.D., Swigut, T., Sero, D., Li, J., Lee, M.K., Zaidi, A., Mattern, B.C., Liebowitz, C., et al., 2018. Genome-wide mapping of global-to-local genetic effects on human facial shape. Nature genetics 50, 414–423.

Czyż, E.A., Schmid, B., Hueni, A., Eppinga, M.B., Schuman, M.C., Schneider, F.D., Guillén-Escribà, C., Schaepman, M.E., 2023. Genetic constraints on temporal variation of airborne reflectance spectra and their uncertainties over a temperate forest. Remote Sensing of Environment 284, 113338.

Dechant, B., Ryu, Y., Badgley, G., Zeng, Y., Berry, J.A., Zhang, Y., Goulas, Y., Li, Z., Zhang, Q., Kang, M., et al., 2020. Canopy structure explains the relationship between photosynthesis and sun-induced chlorophyll fluorescence in crops. Remote Sensing of Environment 241, 111733.

DeLancey, E.R., Simms, J.F., Mahdianpari, M., Brisco, B., Mahoney, C., Kariyeva, J., 2019. Comparing deep learning and shallow learning for large-scale wetland classification in alberta, canada. Remote Sensing 12, 2.

Dubayah, R., Blair, J.B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., Hofton, M., Hurtt, G., Kellner, J., Luthcke, S., et al., 2020. The global ecosystem dynamics investigation: High-resolution laser ranging of the earth's forests and topography. Science of remote sensing 1, 100002.

Dusenge, M.E., Duarte, A.G., Way, D.A., 2019. Plant carbon metabolism and climate change: elevated co 2 and temperature impacts on photosynthesis, photorespiration and respiration. New Phytologist 221, 32–49.

Fields, S., Song, O.k., 1989. A novel genetic system to detect protein–protein interactions. Nature 340, 245–246.

Florio, R., di Salvo, M.L., Vivoli, M., Contestabile, R., 2011. Serine hydroxymethyltransferase: a model enzyme for mechanistic, structural, and evolutionary studies. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics 1814, 1489–1496.

Franklin, S.B., Gibson, D.J., Robertson, P.A., Pohlmann, J.T., Fralish, J.S., 1995. Parallel analysis: a method for determining significant principal components. Journal of Vegetation Science 6, 99–106.

Gamon, J.A., Huemmrich, K.F., Wong, C.Y., Ensminger, I., Garrity, S., Hollinger, D.Y., Noormets, A., Peñuelas, J., 2016. A remotely sensed pigment index reveals photosynthetic phenology in evergreen conifers. Proceedings of the National Academy of Sciences 113, 13087–13092.

Gao, B.C., 1996. Ndwi—a normalized difference water index for remote sensing of vegetation liquid water from space. Remote sensing of environment 58, 257–266.

Gitelson, A.A., Gritz, Y., Merzlyak, M.N., 2003. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. Journal of plant physiology 160, 271–282.

Gitelson, A.A., Merzlyak, M.N., Chivkunova, O.B., 2001. Optical properties and nondestructive estimation of anthocyanin content in plant leaves¶. Photochemistry and photobiology 74, 38–45.

Gitelson, A.A., Viña, A., Ciganda, V., Rundquist, D.C., Arkebauer, T.J., 2005. Remote estimation of canopy chlorophyll content in crops. Geophysical research letters 32.

Glawe, G.A., Zavala, J.A., Kessler, A., Van Dam, N.M., Baldwin, I.T., 2003. Ecological costs and benefits correlated with trypsin protease inhibitor production in nicotiana attenuata. Ecology 84, 79–90.

Grabska, E., Hostert, P., Pflugmacher, D., Ostapowicz, K., 2019. Forest stand species mapping using the sentinel-2 time series. Remote Sensing 11, 1197.

Gu, Y., Brown, J.F., Verdin, J.P., Wardlow, B., 2007. A five-year analysis of modis ndvi and ndwi for grassland drought assessment over the central great plains of the united states. Geophysical research letters 34.

Hanson, A.D., Roje, S., 2001. One-carbon metabolism in higher plants. Annual review of plant biology 52, 119–137.

Hayton, J.C., Allen, D.G., Scarpello, V., 2004. Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. Organizational research methods 7, 191–205.

Huang, M., Liu, X., Zhou, Y., Summers, R.M., Zhang, Z., 2019. Blink: a package for the next level of genome-wide association studies with both individuals and markers in the millions. Gigascience 8, giy154.

Huete, A.R., 2012. Vegetation indices, remote sensing and forest monitoring. Geography Compass 6, 513–532.

Huynh, B.L., Ehlers, J.D., Huang, B.E., Muñoz-Amatriaín, M., Lonardi, S., Santos, J.R., Ndeve, A., Batieno, B.J., Boukar, O., Cisse, N., et al., 2018. A multi-parent advanced generation inter-cross (magic) population for genetic analysis and improvement of cowpea (vigna unguiculata l. walp.). The plant journal 93, 1129–1142.

ĐINH, S.T., Galis, I., Baldwin, I.T., 2013. Uvb radiation and 17-hydroxygeranyllinalool diterpene glycosides provide durable resistance against mirid (tupiocoris notatus) attack in field-grown nicotiana attenuata plants. Plant, Cell & Environment 36, 590–606.

Jacquemoud, S., Ustin, S., 2019. Leaf optical properties. Cambridge University Press.

Joseph, B., Corwin, J.A., Li, B., Atwell, S., Kliebenstein, D.J., 2013. Cytoplasmic genetic variation and extensive cytonuclear interactions influence natural variation in the metabolome. Elife 2, e00776.

Karthikeyan, L., Chawla, I., Mishra, A.K., 2020. A review of remote sensing applications in agriculture for food security: Crop growth and yield, irrigation, and crop losses. Journal of Hydrology 586, 124905.

Katz, J.J., 1973. Chlorophyll interactions and light conversion in photosynthesis. Naturwissenschaften 60, 32–39.

Kim, S.G., Yon, F., Gaquerel, E., Gulati, J., Baldwin, I.T., 2011. Tissue specific diurnal rhythms of metabolites and their regulation during herbivore attack in a native tobacco, nicotiana attenuata. PloS one 6, e26214.

Knipling, E.B., 1970. Physical and physiological basis for the reflectance of visible and near-infrared radiation from vegetation. Remote sensing of environment 1, 155–159.

Ku, M.S., Wu, J., Dai, Z., Scott, R.A., Chu, C., Edwards, G.E., 1991. Photosynthetic and photorespiratory characteristics of flaveria species. Plant physiology 96, 518–528.

Law, J.A., Jacobsen, S.E., 2010. Establishing, maintaining and modifying dna methylation patterns in plants and animals. Nature Reviews Genetics 11, 204–220.

Li, C., Czyz, E.A., Halitschke, R., Baldwin, I.T., Schaepman, M.E., Schuman, M.C., 2023a. Evaluating potential of leaf reflectance spectra to monitor plant genetic variation .

Li, C., Czyż, E.A., Halitschke, R., Baldwin, I.T., Schaepman, M.E., Schuman, M.C., 2023b. Uzh sg nicotiana attenuata asd data. SPECCHIO. URL: http://sc22.geo.uzh.ch:8080/SPECCHIO_Web_Interface/search. keyword: UZH_SG_Nicotiana_attenuata_ASD.

Li, D., Heiling, S., Baldwin, I.T., Gaquerel, E., 2016. Illuminating a plant's tissue-specific metabolic diversity using computational metabolomics and information theory. Proceedings of the National Academy of Sciences 113, E7610–E7618.

Li, X., Wang, Q., Yu, X., Liu, H., Yang, H., Zhao, C., Liu, X., Tan, C., Klejnot, J., Zhong, D., et al., 2011. Arabidopsis cryptochrome 2 (cry2) functions by the photoactivation mechanism distinct from the tryptophan (trp) triad-dependent photoreduction. Proceedings of the National Academy of Sciences 108, 20844–20849.

Libbrecht, M.W., Noble, W.S., 2015. Machine learning applications in genetics and genomics. Nature Reviews Genetics 16, 321–332.

Liu, N., Liu, J., Li, W., Pan, Q., Liu, J., Yang, X., Yan, J., Xiao, Y., 2018. Intraspecific variation of residual heterozygosity and its utility for quantitative genetic studies in maize. BMC plant biology 18, 1–15.

Liu, X., Huang, M., Fan, B., Buckler, E.S., Zhang, Z., 2016. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS genetics 12, e1005767.

Mackay, I.J., Bansept-Basler, P., Barber, T., Bentley, A.R., Cockram, J., Gosman, N., Greenland, A.J., Horsnell, R., Howells, R., O'Sullivan, D.M., et al., 2014. An eight-parent multiparent advanced generation inter-cross population for winter-sown wheat: creation, properties, and validation. G3: Genes, Genomes, Genetics 4, 1603–1610.

Madritch, M.D., Kingdon, C.C., Singh, A., Mock, K.E., Lindroth, R.L., Townsend, P.A., 2014. Imaging spectroscopy links aspen genotype with below-ground processes at landscape scales. Philosophical Transactions of the Royal Society B: Biological Sciences 369, 20130194.

McFadden, B.A., 1980. A perspective of ribulose bisphosphate carboxylase/oxygenase, the key catalyst in photosynthesis and photorespiration. Accounts of Chemical Research 13, 394–399. URL: https://doi.org/10.1021%2Far50155a002, doi:10.1021/ar50155a002.

Meireles, J.E., Cavender-Bares, J., Townsend, P.A., Ustin, S., Gamon, J.A., Schweiger, A.K., Schaepman, M.E., Asner, G.P., Martin, R.E., Singh, A., Schrodt, F., Chlus, A., O'Meara, B.C., 2020. Leaf reflectance spectra capture the evolutionary history of seed plants. New Phytologist 228, 485–493. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/nph.16771, doi:10.1111/nph.16771. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/nph.16771.

Meireles, J.E., Schweiger, A.K., Cavender-Bares, J.M., 2017. spectrolab: Class and methods for hyperspectral data. r package version 0.0. 2. .

Miller, J., Steven, M., Demetriades-Shah, T., 1992. Reflection of layered bean leaves over different soil backgrounds: measured and simulated spectra. International Journal of Remote Sensing 13, 3273–3286.

Mitra, S., Baldwin, I.T., 2008. Independently silencing two photosynthetic proteins in nicotiana attenuata has different effects on herbivore resistance. Plant Physiology 148, 1128–1138.

Mitra, S., Baldwin, I.T., 2014. Ru bpc ase activase (rca) mediates growth–defense trade-offs: silencing rca redirects jasmonic acid (ja) flux from ja-isoleucine to methyl jasmonate (me ja) to attenuate induced defense responses in nicotiana attenuata. New Phytologist 201, 1385–1395.

Moreno, J.I., Martín, R., Castresana, C., 2005. Arabidopsis shmt1, a serine hydroxymethyltransferase that functions in the photorespiratory pathway influences resistance to biotic and abiotic stress. The Plant Journal 41, 451–463.

Nyholt, D.R., 2004. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. The American Journal of Human Genetics 74, 765–769.

Petibon, F., Czyż, E.A., Ghielmetti, G., Hueni, A., Kneubühler, M., Schaepman, M.E., Schuman, M.C., 2021. Uncertainties in measurements of leaf optical properties are small compared to the biological variation within and between individuals of european beech. Remote Sensing of Environment 264, 112601.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D., 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics 38, 904–909.

Qian, X., Liu, L., Croft, H., Chen, J., 2021. Relationship between leaf maximum carboxylation rate and chlorophyll content preserved across 13 species. Journal of Geophysical Research: Biogeosciences 126, e2020JG006076.

R Core Team, 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rao, M., 2020. Gene expression profile of RNA n1-methyladenosine

methyltransferases. E3S Web of Conferences 218, 03052. URL: https://doi.org/10.1051%2Fe3sconf%2F202021803052, doi:10.1051/e3sconf/202021803052.
Ravanel, S., Cherest, H., Jabrin, S., Grunwald, D., Surdin-Kerjan, Y., Douce, R., Rébeillé, F., 2001. Tetrahydrofolate biosynthesis in plants: molecular and functional characterization of dihydrofolate synthetase and three isoforms of folylpolyglutamate synthetase in arabidopsis thaliana. Proceedings of the National Academy of Sciences 98, 15360–15365.
Ray, R., Halitschke, R., Gase, K., Leddy, S.M., Schuman, M.C., Rodde, N., Baldwin, I.T., 2023. A persistent major mutation in canonical jasmonate signaling is embedded in an herbivory-elicited gene network. Proceedings of the National Academy of Sciences 120, e2308500120.
Ray, R., Li, D., Halitschke, R., Baldwin, I.T., 2019. Using natural variation to achieve a whole-plant functional understanding of the responses mediated by jasmonate signaling. The Plant Journal 99, 414–425.
Ray, R., et al., 2022. Whole genome sequence of magic parents and ril population. NCBI Sequence Read Archive. URL: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA907539. deposited 2 December 2022.
Saha, S., Perrin, L., Röder, L., Brun, C., Spinelli, L., 2022. Epi-MEIF: detecting higher order epistatic interactions for complex traits using mixed effect conditional inference forests. Nucleic Acids Research 50, e114. URL: https://doi.org/10.1093/nar/gkac715, doi:10.1093/nar/gkac715.
Sannemann, W., Huang, B.E., Mathew, B., Léon, J., 2015. Multi-parent advanced generation inter-cross in barley: high-resolution quantitative trait locus mapping for flowering time as a proof of concept. Molecular Breeding 35, 1–16.
Sarabi, M.M., Naghibalhossaini, F., 2015. Association of dna methyltransferases expression with global and gene-specific dna methylation in colorectal cancer cells. Cell biochemistry and function 33, 427–433.
Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary dna microarray. Science 270, 467–470.
Serbin, S.P., Singh, A., McNeil, B.E., Kingdon, C.C., Townsend, P.A., 2014. Spectroscopic determination of leaf morphological and biochemical traits for northern temperate and boreal tree species. Ecological Applications 24, 1651–1669.
Sero, D., Zaidi, A., Li, J., White, J.D., Zarzar, T.B.G., Marazita, M.L., Weinberg, S.M., Suetens, P., Vandermeulen, D., Wagner, J.K., et al., 2019. Facial recognition from dna using face-to-dna classifiers. Nature communications 10, 2557.
Solomon, M.J., Larsen, P.L., Varshavsky, A., 1988. Mapping proteindna interactions in vivo with formaldehyde: Evidence that histone h4 is retained on a highly transcribed gene. Cell 53, 937–947.
Somerville, C., Ogren, W., 1980. Photorespiration mutants of arabidopsis thaliana deficient in serine-glyoxylate aminotransferase activity. Proceedings of the National Academy of Sciences 77, 2684–2687.
Somerville, C., Ogren, W.L., 1981. Photorespiration-deficient mutants of arabidopsis thaliana lacking mitochondrial serine transhydroxymethylase activity. Plant physiology 67, 666–671.
Taiz, L., Zeiger, E., Møller, I.M., Murphy, A., et al., 2015. Plant physiology and development. Ed. 6, Sinauer Associates Incorporated.
Thenkabail, P.S., Lyon, J.G., Huete, A., 2018. Advances in hyperspectral remote sensing of vegetation and agricultural crops, in: Fundamentals, Sensor Systems, Spectral Libraries, and Data Mining for Vegetation. CRC press, pp. 3–37.
Verrelst, J., Malenovskỳ, Z., Van der Tol, C., Camps-Valls, G., Gastellu-Etchegorry, J.P., Lewis, P., North, P., Moreno, J., 2019. Quantifying vegetation biophysical variables from imaging spectroscopy data: a review on retrieval methods. Surveys in Geophysics 40, 589–629.
Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., Yang, J., 2017. 10 years of gwas discovery: biology, function, and translation. The American Journal of Human Genetics 101, 5–22.
Wan, L., Tang, Z., Zhang, J., Chen, S., Zhou, W., Cen, H., 2021. Upscaling from leaf to canopy: Improved spectral indices for leaf biochemical traits estimation by minimizing the difference between leaf adaxial and abaxial surfaces. Field Crops Research 274, 108330.
Wang, J., Zhang, Z., 2021. Gapit version 3: boosting power and accuracy for genomic association and prediction. Genomics, proteomics & bioinformatics 19, 629–640.
Wang, R., Gamon, J.A., 2019. Remote sensing of terrestrial plant biodiversity. Remote Sensing of Environment 231, 111218.
Wang, R., Gamon, J.A., Schweiger, A.K., Cavender-Bares, J., Townsend, P.A., Zygielbaum, A.I., Kothari, S., 2018. Influence of species richness, evenness, and composition on optical diversity: A simulation study. Remote Sensing of Environment 211, 218–228.
Wang, Y., Goh, W., Wong, L., Montana, G., the Alzheimer's Disease Neuroimaging Initiative, 2013. Random forests on Hadoop for genome-wide association studies of multivariate neuroimaging phenotypes. BMC Bioinformatics 14, S6. URL: https://doi.org/10.1186/1471-2105-14-S16-S6, doi:10.1186/1471-2105-14-S16-S6.
Wang, Z., Chlus, A., Geygan, R., Ye, Z., Zheng, T., Singh, A., Couture, J.J., Cavender-Bares, J., Kruger, E.L., Townsend, P.A., 2020. Foliar functional traits from imaging spectroscopy across biomes in eastern north america. New Phytologist 228, 494–511.
Weigelt, A., Mommer, L., Andraczek, K., Iversen, C.M., Bergmann, J., Bruelheide, H., Freschet, G.T., Guerrero-Ramírez, N.R., Kattge, J., Kuyper, T.W., Laughlin, D.C., Meier, I.C., van der Plas, F., Poorter, H., Roumet, C., van Ruijven, J., Sabatini, F.M., Semchenko, M., Sweeney, C.J., Valverde-Barrantes, O.J., York, L.M., McCormack, M.L., 2023. The importance of trait selection in ecology. Nature 618, E29–E30. URL: https://www.nature.com/articles/s41586-023-06148-8, doi:10.1038/s41586-023-06148-8. number: 7967 Publisher: Nature Publishing Group.
Wu, D., Karhade, D.S., Pillai, M., Jiang, M.Z., Huang, L., Li, G., Cho, H., Roach, J., Li, Y., Divaris, K., 2021. Machine learning and deep learning in genetics and genomics. Machine Learning in Dentistry , 163–181.
Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., et al., 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics 38, 203–208.
Zhang, F., Zhou, G., 2019. Estimation of vegetation water content using hyperspectral vegetation indices: A comparison of crop water indicators in response to water stress treatments for summer maize. BMC ecology 19, 1–12.
Zhang, Z., Ersoz, E., Lai, C.Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., Bradbury, P.J., Yu, J., Arnett, D.K., Ordovas, J.M., et al., 2010. Mixed linear model approach adapted for genome-wide association studies. Nature genetics 42, 355–360.