Manuscript title: Genome-wide associations of leaf spectral variation in MAGIC lines of *Nicotiana attenuata*

Authors: Cheng Li[1*], Ewa A. Czyż[1†], Bernhard Schmid[1], Rishav Ray[2‡], Rayko Halitschke[2], Ian T. Baldwin[2], Michael E. Schaepman[1], Meredith C. Schuman[1,3]

[1] Department of Geography, University of Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland
[2] Department of Molecular Ecology, Max Planck Institute for Chemical Ecology, Hans-Knoell-Strasse 8, 07745 Jena, Germany
[3] Department of Chemistry, University of Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland
*Corresponding author: *cheng.li@geo.uzh.ch*
Keywords: Field spectroscopy | Genome-wide association studies (GWAS) | Hierarchical Spectral Clustering with Parallel Analysis | Leaf reflectance | Multiparent Advanced Generation Inter-Cross (MAGIC) | *Nicotiana attenuata*

† Present address: Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, 91109 Pasadena, California, USA
‡ Present address: Department of Plant Biology, University of California, Davis, 605 Hutchison Drive, Green Hall 1002, 95616 Davis, California, USA

ABSTRACT

The application of in-field and aerial spectroscopy to assess functional and phylogenetic variation in plants has led to novel ecological insights and supports global assessments of plant biodiversity. Understanding how plant genetic variation influences reflectance spectra will help harness this potential for biodiversity monitoring and improve understanding of why plants differ in functional responses to environmental change. Here, we use a well-resolved genetic mapping population derived from Multi-parent Advanced Generation Inter-cross (MAGIC) lines of *Nicotiana attenuata* to associate genetic differences with differences in leaf spectra between plants in a field experiment in their natural environment. We analyzed the leaf reflectance spectra using a hand-held spectroradiometer (350–2500 nm) on 616 fully genotyped plants of *N. attenuata* grown in a randomized block design. We tested three approaches to conducting Genome-Wide Association Studies on spectral variants. We introduce a new Hierarchical Spectral Clustering with Parallel Analysis (HSC-PA) method. This method 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 734-1143 nm spectral range, spanning the red-edge and near-infrared regions that are sensitive to leaf structure and photosynthetic activity. This locus contains a candidate gene annotated as carbonic anhydrase, an enzyme involved in CO_2 hydration and regulation of photosynthetic efficiency, suggesting a physiological link between variation in leaf optical properties and carbon assimilation. In contrast, an approach treating single wavelengths as phenotypes identified the same associations as HSC-PA, but without the statistical power to pinpoint significant associations. An index-based approach, which reduces complex spectra to a few dimensionless variables, detected two significant associations for $ARDSI_C_w$ (a water-content-related index) with loci on chromosome 1 near genes annotated as a

Zeta toxin domain-containing protein, and an Exocyst subunit Exo70 family protein. While these findings are biologically plausible, they represent a very narrow subset of the spectral variation captured by HSC-PA. The HSC-PA approach supports 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 linking plant genetics with biodiversity monitoring, large-scale ecological assessment and remote-sensing applications.

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, et al. 2018). Remote sensing may revolutionize the way we study genetic diversity by offering non-invasive methods with the potential of repeated measurement on large spatial scales (Madritch et al. 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. Notably, high-throughput hyperspectral phenotyping has long been used in crop quantitative genetics to link genetic variation with spectral traits, demonstrating the heritability of multiple biochemical and physiological traits estimated from spectral signatures—thus providing valuable methodological precedents for biodiversity and ecological studies (Grzybowski et al., 2021). 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 et al., 2014). 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 et al., 2014; Cavender-Bares et al., 2016; Wang et al., 2018; Asner et al., 2015; Stasinski et al., 2021; Czyż et al., 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 et al., 2023).

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 variation. 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 et al., 2017; Santure et al., 2018). 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 et al., 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 et al.,

84 2010), are employed to correct for population structure in GWAS analyses. Phenotyping remains
85 a substantial bottleneck in genetic studies due to its labor-intensive and time-consuming nature.
86 High-throughput phenotyping, especially hyperspectral imaging, offers a powerful solution—
87 enabling rapid, scalable, and non-invasive measurements of plant traits that can accelerate
88 genotype–phenotype association studies (Furbank & Tester, 2011; Zhang & Zhang, 2018).

89 *Nicotiana attenuata* Torr. ex S. Watson is a model wild plant for molecular ecology native to the
90 Great Basin Desert of the southwestern USA. It is primarily found in large ephemeral
91 populations following fires in sagebrush and pinyon-juniper ecosystems. Its unique germination
92 behavior is stimulated by cues found in wood smoke and the removal of inhibitors from
93 unburned litter, allowing it to thrive in post-fire environments (Bahulikar et al., 2004).
94 Ecologically, *N. attenuata* is a compelling subject for its intricate defense mechanisms against
95 herbivores, its tissue-specific diurnal metabolic rhythms, and its ability to acclimate to varying
96 environmental conditions, including high UVB radiation (Glawe et al., 2003; Kim et al., 2011; Li
97 et al., 2016; INH, Galis, and Baldwin 2013). Further, it offers unique opportunities for genetic
98 research due to a well-defined MAGIC (Multiparent Advanced Generation Inter-Cross)
99 population (Ray et al., 2019, 2023). This design comprises 26 phenotypically and genetically
100 differentiated parental lines (PLs) intercrossed to produce two replicate sets each of 325
101 recombinant inbred lines (RILs; 616 genotyped plants of the total of 650 used in the present
102 study), allowing for the study of phenotype-genotype associations without the confounding
103 influence of population structure. More details on plant material, field site, and generation of
104 MAGIC RIL Population can be found in Appendix S1: Supporting Text.
105 In this study, we explore the genetic diversity of the MAGIC population to identify associations
106 between genetic variants and leaf spectral traits, advancing our understanding of genetic

influences on complex spectral characteristics. Our methodology involves three approaches: analyzing spectral indices, single wavelengths (SW), and implementing Hierarchical Spectral Clustering with Parallel Analysis (HSC-PA). Through GWAS using 168,903 SNPs that showed variation within the population, we investigate genetic links with genome annotation. Our methodological approach and the tools utilized are summarized in Figure 1.

2 MATERIALS AND METHODS

This study's methodology extends from our previous work (Li et al., 2023), with detailed description available in Appendix S1: Supporting Text. Briefly, our study utilizes two replicate sets each of 325 *N. attenuata* MAGIC RILs, derived from 26 parental lines and six generations of inbreeding, yielding a total of 650 genetically different plants of which 616 were planted in the current experiment. The RILs sequencing was performed at Novogene HK (Ray et al., 2023), and leaf optical properties (350-2500 nm) were measured using a FieldSpec 4 spectroradiometer (Li et al., 2023).

Data analysis was performed using R (R Core Team 2023), employing the spectrolab package (version 0.0.10, Meireles, et al., 2017) and a three-step filtering approach for outlier removal (Li et al., 2023). We employed a linear model to remove known environmental effects from the raw spectral data. Specifically, for each wavelength, we fitted the model:

$$Y \sim LN + REP + BA + LN:(REP + BA) + DT:CT + MO$$

where LN represents the influence of the measurement technique (the number of measured leaves), REP accounts for potential differences between the two replicate sets of RILs (that occupied different sectors of the field plot and were measured in series), BA represents spatial patch effects characterized from *N. attenuata* reference measurements (in each square of 2×2

129 planting positions one was occupied by a "phytometer" plant of a single tester genotype of the
130 species), $LN: (REP + BA)$ models the interaction between measurement technique and spatial
131 effects, $DT: CT$ captures temporal effects (DT for day and CT for time of day), and MO denotes
132 the maternal cytoplasmic effects. The residuals from each fitted model were taken as the adjusted
133 spectral values, corrected for (i.e., removing) the modeled environmental influences.
134 We tested three ways of treating the spectral phenotypic data. First, we used six selected spectral
135 indices, which are single values derived from ratios of reflectance at specific wavelengths
136 representing spectral features (details see Table 1 and Appendix S1: Supporting Text). These
137 indices span pigment-related (e.g., chlorophyll a/b, carotenoids), water-related, and structural
138 spectral features across the VIS, red-edge, NIR, and SWIR ranges, thus capturing major axes of
139 leaf functional variation. The indices were first calculated from the raw spectral data and
140 subsequently adjusted using the same linear model as described above. Second, we used single
141 wavelengths (SW), i.e., treating each wavelength in the adjusted spectrum as a phenotype. Third,
142 we used Hierarchical Spectral Clustering with Parallel Analysis (HSC-PA), a new data-driven
143 dimension reduction approach, developed from a method for the analysis of genetic associations
144 with human facial features (Claes et al., 2018; Sero et al., 2019), which accounts for correlations
145 within spectral data while retaining interpretable features. We applied HSC-PA to the adjusted
146 spectral data for downstream analyses. We associated the resulting phenotypic data with 168,903
147 SNPs using multiple GWAS models based on the Genome Association and Prediction Integrated
148 Tool (GAPIT) version 3 (Wang and Zhang 2021). The GWAS models include: General Linear
149 Model (GLM, Price et al., 2006), Mixed Linear Model (MLM, Yu et al., 2006), Fixed and
150 random model Circulating Probability Unification (FarmCPU, Liu et al., 2016), and Bayesian
151 information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK, Huang et al.,

2019). Kinship matrices were automatically calculated by GAPIT using the Zhang method (Zhang et al., 2010) from the genotype data; these matrices were applied only in models that incorporate kinship (i.e., MLM). See more details of the four models in Appendix S1 Supporting Text.

Multiple testing adjustments were conducted on 168,903 SNPs. Initially, an exploratory threshold of $p < 1 \times 10^{-5}$ was used for preliminary SNP selection. This was followed by a false discovery rate (FDR) adjustment (Benjamini and Hochberg 1995) and the application of the PhenoSpD method (Nyholt 2004; Zheng et al., 2018) to determine the significance threshold ($0.05/M_{eff}$), taking into account the effective number of independent trait calculations. The PhenoSpD method first derives the pairwise phenotypic correlation matrix from GWAS summary statistics using cross-trait LD Score regression. Specifically, for two traits $t_1$and t_2, the regression

$$\mathrm{Z_{1j}\,Z_{2j}} = \mathrm{a} + \mathrm{b} \cdot LDscore_j + \epsilon_\mathrm{j}$$

is calculated across SNPs j, where Z_{1j} and Z_{2j} are the z-scores from GWAS summary statistics for each trait, and the intercept a provides an estimate of the phenotypic correlation between traits. This phenotypic correlation matrix is then subjected to spectral decomposition (SpD), in which the eigenvalues (λ_i) are used to calculate the effective number of independent tests:

$$\mathrm{M_{eff}} = \frac{\left(\sum_{i=1}^{M} \lambda_i\right)^2}{\sum_{i=1}^{M} {\lambda_i}^2}$$

where M is the total number of phenotypes. We used $0.05/M_{eff}$ as the significant threshold. SNPs with FDR-adjusted p-values below the significance threshold were considered significantly associated. For candidate gene selection, we extracted genes within ±100 kb of significant SNPs from the *N. attenuata* genome, following the approach used in closely related *Nicotiana tabacum*

(Lai et al., 2021). We then refined the candidate gene list using the *N. attenuata* Data Hub (http://nadh.ice.mpg.de/NaDH/) to filter out genes lacking transcript accumulation evidence in leaves under various conditions. Genes with relevant functions were then identified based on annotation information.

We also conducted GWAS using indices, single wavelengths, and HSC-PA segments with raw spectral data as a general comparison to assess the effect of removing known environmental effects (see Appendix S1: Supporting Text, Results from raw spectral data).

3 RESULTS

3.1 Phenotypic analysis

3.1.1 Variance partitioning and adjustment of raw spectra

To advance our understanding of the phenotypic variation in leaf spectral traits and to remove the known environmental effects, we employed linear models. These models partitioned variance across the spectral range of 350–2500 nm, assessing contributions to cumulative multiple R^2 values (i.e. percent sum of squares) from measurement technique (LN), replicate and spatial effects and their interactions with measurement technique ($REP + BA + LN{:}(REP + BA)$), measurement time ($DT{:}CT$), and maternal cytoplasm (MO) within the 616 plants. Figure 2 illustrates these components for each wavelength as stacked bars, which—given the large number of wavelengths—merge into a continuous stacked profile that collectively accounts for the total variance. Analysis of variance results indicated that residual phenotypic variance between plants, representing genetic and residual environmental contributions accounted from 71.7% to 89.5% of the total phenotypic variance.

3.1.2 Spectral characteristics of leaves

196 Our analyses focus on calculated leaf reflectance spectra spanning from 400 to 2500 nm, as the
197 initial region (350–399 nm) has a poor signal-to-noise ratio and is commonly excluded from
198 downstream analyses (Cavender-Bares et al., 2016). This range includes the visible (VIS), near
199 infrared (NIR), and shortwave infrared (SWIR) regions. This broad spectral range captured
200 variation among the 616 plants, shown in Figure 3a as an orange shaded region indicating the
201 range of adjusted reflectance spectra for all samples. The blue shaded region indicates the range
202 of raw reflectance spectra before adjustment with linear models. The NIR region showed more
203 consistency in reflectance values among the plants, contrasting with greater variability in the VIS
204 and SWIR regions.
205 The distribution of six selected spectral indices (Figure 3b), encapsulating attributes like water
206 content and chlorophyll concentration, provided more specific information about leaf
207 physiology. Note that the y-axis represents the index values, which are specific to each index
208 calculation and therefore cannot be directly compared across indices.

209 Our subsequent application of the HSC-PA segmented the adjusted spectra into 38 distinct
210 patterns. The blank (NA) areas are regions where further clustering was halted due to the
211 retention of a singular Principal Component (PC) from a previous level. This approach allowed
212 for a finer granularity in the spectral data representation while accounting for correlations among
213 wavelengths. The hierarchical clustering process concluded at different levels, ranging from
214 level 3 to level 8, indicating the varying depths of spectral similarities among the segments.
215 Table 2 shows the detailed spectral region and potential associated features of the segments used
216 in HSC-PA-GWAS. Appendix S1: Table S1 shows the spectral regions of all 38 segments.

217 By visualizing the 20 segments that each retained only a single PC, we were able to span the
218 entire spectrum from 400 to 2500 nm. These segments, parsed into several interpretable regions,

were then used for HSC-PA-GWAS approach, offering phenotypic data for uncovering associations with underlying genetic variations.

The distributions of the six indices derived from raw spectra, along with the HSC-PA results based on raw spectral data, are shown in Appendix S1: Figure S1.

3.1.3 Effective number of independent traits (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 traits, denoted as M_{eff}, using the PhenoSpD method (Zheng et al., 2018). For the Index-GWAS approach, which employed six distinct spectral indices, the effective number of independent tests, $M_{eff_{ind}}$, was determined to be 5.0. In the Single Wavelength GWAS (SW-GWAS) approach, the effective number, $M_{eff_{sw}}$, was notably higher at 999.9. 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 10.5, reflecting the reduced dimensionality and the focus on major patterns of variation. These calculated M_{eff} values are used to set a threshold ($0.05/M_{eff}$) for significant associations.

3.2 SNP marker analysis

The kinship relationships among the 616 plants are visualized in Appendix S1: Figure S2. Through a heatmap, the genetic relatedness between individuals is presented. Most values in the histogram gravitate towards zero, indicating a consistent and minimal genetic difference between plants. This uniform distribution of kinship values affirms the thorough genetic mixing of the

239 parental lines in the MAGIC population and is consistent with the finding that this population
240 does not display obvious structure (Ray et al., 2023).

241 Appendix S1: Figure S3 presents the frequency of heterozygosity of individual and markers,
242 while Appendix S1: Figure S4 illustrates the linkage disequilibrium (LD) decay over distance.

243 3.3 Results of genome-side association studies (GWAS)

244 3.3.1 Index-GWAS results

245 The Manhattan plot presented in Figure 4a, shows the associations between the SNPs and the six
246 indices ($NDWI$, CCI, CI_{re}, $ARDSI_C_{ab}$, $ARDSI_C_{w}$, and $ARDSI_C_{m}$) using four models (GLM,
247 MLM, FarmCPU, and BLINK). Several genomic regions displayed distinct associations
248 according to the $-\log_{10}$(p-values). An exploratory threshold, defined at 1×10^{-5} (denoted by
249 the grey horizontal line), was adopted to shortlist SNPs for a detailed examination.

250 In Figure 4b, each of the 6 indices is plotted against the SNPs that exceeded the exploratory
251 threshold. The gradient color scale indicates the association's intensity, with deeper hues
252 indicating smaller FDR-adjusted p-values. This representation provides a concise overview of
253 the genomic regions that are most strongly correlated with each spectral index. In total, 12 SNPs
254 exhibited p-values less than 1×10^{-5}: five with $ARDSI_C_{w}$, three with CCI, two each with
255 $NDWI$ and CI_{re} and one each with $ARDSI_C_{m}$and $ARDSI_C_{ab}$. The index $ARDSI_C_{w}$ displayed
256 two associations lower than the significance threshold: $\frac{0.05}{5.0} = 0.01$, with two SNPs in
257 chromosome 1. The first SNP was located at chr1_193505583, near Niat3g_08138, annotated as
258 a Zeta toxin domain-containing protein. The second was at chr1_438155938, close to
259 Niat3g_17277 (TPR_REGION domain-containing protein) and Niat3g_17282 (Exocyst subunit

Exo70 family protein). Furthermore, $ARDSI_C_w$ was chosen as a representative example for the QQ plots in Figure 4c and 4d. 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. 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. Figure 4e and 4f show the phenotypic distributions for two significant markers, where 0 represents a homozygous reference allele, 1 a heterozygous allele, and 2 a homozygous alternative allele.

3.3.2 SW-GWAS and HSC-PA-GWAS results

In the SW-GWAS analysis, the BLINK model was selected to investigate the genetic correlations with 2101 spectral wavelengths spanning from 400 to 2500 nm. Figure 5a highlights the genetic associations for the spectral wavelength at 952 nm, which displayed the strongest association among all phenotypes, serving as an example. The diagnostic QQ plot for this wavelength is also presented in Figure 5c. Figure 5b offers an in-depth depiction of associations that exceeded the exploratory threshold. The X-axis represents the spectral wavelengths, the Y-axis the associated SNPs, providing a holistic view of the spectral regions correlated with each SNP. A total of 27 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 need for methods that can account for these correlations. Furthermore, correlations with the same SNP spanning more distant spectral regions may also reflect underlying biological processes and warrant further investigation. The associations span nearly the entire spectral range and are distributed across most chromosomes. The significance level, adjusted by M_{eff}, stands at 5.0×10^{-5}. Notably, no SNP met this threshold.

283 In the HSC-PA-GWAS, Figure 6a shows the associations that surpassed our exploratory
284 threshold of 1×10^{-5}. A total of 20 SNP associations distributed over 20 segments emerged
285 from the analysis, all of which were also found in the SW-GWAS. Focusing on segment
286 associations obscures spectral information, and so Figure 6b maps these associations back to
287 their corresponding spectral wavelengths for better interpretation. This visualization facilitates
288 the identification of specific spectral regions linked to each SNP. It is noteworthy that these
289 associations span a range of chromosomal regions, indicating a complex genetic basis of the
290 spectral traits. The significance level, after adjusting for M_{eff}, is set at 4.7×10^{-3}. Two
291 associations, the SNP chr1_131409683 with Segments 37 and 38 (734-1143 nm), surpassed this
292 threshold. The QQ plots for Segments 37 and 38, as depicted in Figure 6c and d, serve a dual
293 purpose of validation and diagnostic assessment. Figure 6e and f show the phenotypic
294 distributions of the significant marker (chr1_131409683) affiliated with Segments 37 and 38. In
295 this distribution, the numbers 0, 1, and 2 represent homozygous alleles as the reference,
296 heterozygous alleles, and homozygous alternative alleles, respectively. The majority of the RILs
297 are assigned reference or heterozygous alleles at this site, as evidenced by the greater number of
298 points in classes 0 and 1. A clear difference in the mean values of these three classes is evident in
299 the boxplot.
300 To further elucidate the genetic association indicated by HSC-PA-GWAS, we searched for
301 potential candidate genes within a ± 100 kb window surrounding the SNP chr1_131409683. This
302 led to the identification of one gene of interest: Niat3g_06527 (more annotation information can
303 be found in Appendix S1: Table S2). This gene is annotated as Carbonic anhydrase (CA).
304 We broadened our analysis to include the top 10 SNPs from HSC-PA GWAS with the smallest
305 FDR-adjusted P-values (see Appendix S1: Table S2). The top six associations are linked to

306 Segment 37 or Segment 38 in the red-edge and near-infrared regions. In addition to the
307 significantly associated locus (chr1_131409683), these associations involve two SNPs:
308 chr12_4508375 and chr10_23318853. Furthermore, the significantly associated locus
309 (chr1_131409683) is also associated with Segment 34 (1144–1310 nm), which lies immediately
310 after Segments 37 and 38 and remains within the NIR region.

311 # 4 DISCUSSION

312 ## 4.1 HSC-PA-GWAS for discovering genetic associations with spectral variation

313 Understanding the genetic foundation of leaf spectral diversity will facilitate advances in
314 biodiversity monitoring and phenotyping, but determining appropriate approaches to extract
315 genetic information from complex, context-dependent spectral phenotypes is a critical step (Li et
316 al., 2023). SW-GWAS, while straightforward, is limited by the high dimensionality of spectral
317 data and the spectral resolution of instruments, typically ranging from 3-10 nm. Subsampling can
318 simplify data and enhance signal-to-noise ratios, yet risks omitting key data points, and the
319 choice of bands for subsampling and the resulting number of traits require consideration and
320 testing. It is common to subsample the continuous spectral data to 1 nm intervals, as done here
321 for the SW-GWAS and as a basis for other analyses, because 1 nm is the subsampling interval
322 provided in the calibrated output data from the ASD FieldSpec and other field spectrometers.
323 Cavender-Bares and colleagues (Cavender-Bares et al., 2016) subsampled leaf spectra at a larger
324 interval of 5 nm prior to PLS-DA and evolutionary analysis and stated that the decision
325 regarding the interval width (1, 2, or 10 nm) did not significantly impact the results. Czyż and
326 colleagues used a coarser subsampling interval of 10 nm to estimate the predictive power of
327 spectral bands for genetic structure (Czyż et al., 2020). Dimensionality reduction methods like
328 PCA yield complex components challenging to biological interpretation (Li et al., 2023).

329 Similarly, spectral indices and methods like varimax rotation have limitations due to continuous
330 spectra and multiple influencing factors. Partial least squares regression (PLSR) is another
331 alternative, often combined with discriminant analysis or beta regression models, but it requires
332 orthogonal trait measurements for accurate associations and carries the risk of overfitting
333 (Cavender-Bares et al., 2016; Wang et al., 2020; Stasinski et al., 2021).

334 Here, we developed HSC-PA-GWAS as a new approach that retains the information in spectra in
335 a biologically interpretable form without losing meaningful information, and without requiring
336 an *a priori* knowledge to target specific traits, while maintaining sufficient statistical power to
337 identify potentially meaningful associations. This method stands out for its combination of
338 sensitivity and statistical power when handling spectral phenotypes. Associations identified
339 using HSC-PA-GWAS were consistent with the SW-GWAS method, as every SNP identified by
340 HSC-PA-GWAS was also detected by SW-GWAS. However, the HSC-PA method surpasses
341 SW-GWAS in several key aspects. One of its strengths is its ability to reduce data dimensionality
342 while accounting for phenotypic correlations within spectra, by transforming multiple correlated
343 wavelengths into a single value, typically represented by the first principal component. This
344 mitigates the challenges of multiple testing and correlated phenotypic data and thus increases
345 statistical power. By aggregating information across multiple wavelengths, the HSC-PA method
346 offers a more comprehensive and potentially more accurate representation of the underlying
347 traits with its ability to discover the correlation structure among wavelengths from data. The
348 resulting genetic associations encompass a broader portion of the spectrum than those derived
349 from indices, which focus on specific wavelength combinations, or from single wavelengths,
350 which do not behave independently. While indices remain valuable for highlighting proxies of
351 relevant biological processes or capturing relative changes between spectral regions, the HSC-

PA approach retains a wider spectral context for interpretation. Moreover, compared to spectral subsampling (e.g., sampling or averaging every 10 nm), which typically does not allow for (varying) covariation patterns of spectral bands and focuses only on neighboring bands, the HSC-PA method provides a more data-driven approach that accounts for empirical internal correlations.

Comparing the HSC-PA method with established methods, such as an approach combining partial least squares regression (PLSR) and linear regression to interpret spectral features in terms of other traits (Verrelst et al., 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 regression. 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, can be computationally intensive, and carry the danger of overfitting. The HSC-PA method offers a data-driven yet human-interpretable approach with the caveat that its criterion for simplification is linear (dimensionality in a PCA). 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 Spectra adjustment and genetic associations

374 In this study, we used replicated plants of a single tester genotypes of *N. attenuate* as
375 “phytometers” to account for small-scale (2x2 planting positions) spatial variation across an
376 experimental field. In addition, we accounted for differences in other environmental variables
377 such as measurement time as well as for maternal effects. By applying a linear model across the
378 full spectral range and using the residuals as adjusted spectra for downstream analyses, we were
379 able to reduce the influence of confounding environmental variables. This correction, which
380 could not have been incorporated directly into a mixed-model GWAS analysis, substantially
381 altered many results, including derived phenotypes and GWAS associations, leading to different
382 segment definitions and candidate genes in comparison to directly conducting GWAS on spectral
383 data without prior correction for non-genetic effects (Appendix S1: Supporting Text, Results
384 from raw spectral data). Such shifts underscore the importance of accounting for non-genetic
385 effects in spectral-genomic studies, as failing to do so can lead to markedly different—and
386 potentially misleading—biological conclusions. Here, we were able to do so both by using a
387 phytometer and by accounting for environmental factors dictated by the experimental and
388 measurement design; while a genetically invariable phytometer is rarely an option in ecological
389 studies, it should usually be feasible to account for quantitative environmental and experimental
390 effects on phenotype values.
391 Advanced genetic association models considering epistasis, akin to HSC-PA's handling of
392 spectral correlations, could enhance GWAS power, sensitivity, and interpretability, with
393 advances in machine learning opening new possibilities (Libbrecht and Noble 2015; Wu et al.,
394 2021). Random forest models have already been used for analyzing complex genetic effects on
395 multivariate phenotypes, though their significance interpretation differs (Brieuc et al., 2018;
396 Wang et al., 2013). The "Epi-MEIF" approach by Saha et al., (2022), using conditional inference

397 forests for epistasis in complex phenotypes, offers a method for significant association
398 identification, accommodating multiple loci testing as an add-on to conventional single-locus
399 models like those in GAPIT.

400 ### 4.3 Ecological significance of findings in *N. attenuata*

401 We discovered a significant association between the SNP chr1_131409683 and Segments 37 and
402 38 (734-1143 nm) in the HSC-PA-GWAS. One interesting candidate gene was identified within
403 100 kb of this SNP: Niat3g_06527, is annotated as a carbonic anhydrase (CA). Carbonic
404 anhydrases catalyze the reversible hydration of CO_2 to bicarbonate and protons ($CO_2 + H_2O \rightleftharpoons$
405 $HCO_3^- + H^+$) (Supuran, 2008). In plants, multiple CA isoforms are found in the chloroplast
406 stroma, thylakoid membranes, cytosol, and other compartments (Fabre et al., 2007; Rudenko et
407 al., 2015). Chloroplastic CAs are thought to facilitate the rapid supply of CO_2 to Rubisco by
408 accelerating the interconversion between dissolved inorganic carbon species, thereby supporting
409 the Calvin–Benson cycle (Badger & Price, 1994). In C_4 and CAM plants, CA activity in
410 mesophyll cells is also critical for providing bicarbonate to phosphoenolpyruvate carboxylase
411 during the initial fixation step (Hatch & Burnell, 1990). In C_3 species such as *N. attenuata*,
412 stromal and thylakoid-associated CAs have additionally been implicated in maintaining
413 photosynthetic efficiency under fluctuating light and CO_2 conditions, and in the regulation of pH
414 and ion balance in the thylakoid lumen (Tiwari et al., 2016).
415 The associated spectral range, 734-1143 nm, spans the red-edge and near-infrared regions, which
416 are sensitive to changes in leaf internal structure and chlorophyll content. Genetic variation in a
417 chloroplast-localized CA could influence CO_2 assimilation rates and downstream effects on
418 chlorophyll concentration, mesophyll cell structure, or water status, any of which can alter
419 reflectance in this part of the spectrum. The link between Niat3g_06527 and these spectral

420 regions therefore suggests that natural allelic differences in CA function may contribute to
421 variation in photosynthetic performance and leaf optical properties in *N. attenuata*.
422 In the context of the ecology of *N. attenuata*, understanding the genetic basis of spectral
423 variation can provide insights into the adaptive mechanisms of this wild plant. The significant
424 association of a SNP close to a gene annotated as carbonic anhydrase, with a spectral range
425 spanning the red-edge and near-infrared regions—wavelengths closely linked to chlorophyll
426 content, canopy structure, and photosynthetic capacity—may reflect evolutionary pressures that
427 have shaped the species' genetic architecture to optimize carbon fixation and water-use
428 efficiency, enhancing its survival in the variable and often harsh environments of its native
429 habitat. Our extended analysis of the top 10 SNPs has unveiled additional genetic associations
430 that may play important significant roles in the ecological adaptation and survival of *N.*
431 *attenuata*.

432 Two other statistically significant associations were discovered with the $ARDSI_C_w$ index, an
433 index of water content that uses the wavelengths 1080, 1240, 1360, and 1560 nm, which are
434 mostly (except for 1080 nm) in the range of wavelengths longer, *i.e.*, further into the infrared
435 region, than those comprising segments 37 and 38. The two associated SNPs, both also on
436 chromosome 1, were located in the vicinity of the gene Niat3g_08138 (chr1_193505583),
437 annotated as a Zeta toxin domain-containing protein, and the genes Niat3g_17277 annotated as
438 TPR_REGION domain-containing protein and Niat3g_17282 annotated as Exocyst subunit
439 Exo70 family protein (chr1_438155938). Proteins with these domains are documented to serve a
440 wide variety of functions in defense, growth and development, and cellular function.

441 It is interesting that the strongest associations were all identified between loci on chromosome 1
442 and wavelengths longer than those of visible light, in the red edge to infrared portions of the leaf

443 spectrum, that are influenced by micromorphology, structure, abundant non-pigment organic
444 constituents, and water content. The strongest associations were within regions of high
445 signal:noise in spectral reflectance, *i.e.*, where the detection of phenotypic differences should be
446 most robust.

447 ## 4.4 Limitations and outlook

448 One limitation of our study lies in the phenotypic interpretation for biological meanings specific
449 to the samples we used. The six spectral indices we selected are derived from existing literature
450 and have not been validated in *N. attenuata* generally, or for this dataset. Without chemical or
451 other analyses to specifically quantify the constituents represented in the indices, such as
452 chlorophyll, our interpretations remain speculative. Establishing causal links between candidate
453 genes and specific biological processes is inherently difficult, particularly in ecological systems
454 where multiple interacting factors may influence spectral traits (Hoban et al. 2016). Moreover,
455 while the HSC-PA-GWAS approach has identified potential candidate genes, any influence of
456 these genes on the traits of interest remains to be verified, e.g. through the generation and testing
457 of knock-out, knock-down, and overexpression lines.

458 While GWAS with a MAGIC population does not require replication of genotypes, the absence
459 of such replicates prevented us from testing variation between genotypes directly using the
460 analysis of variance approach. Nevertheless, by removing small-scale environmental variation
461 using replicated phytometers of a single genotype as well as other environmental variation and
462 maternal effects, the residual phenotypic differences between plants was likely due to their
463 genetic differences rather than further environmental causes.

464 Our study noted unexpectedly high heterozygosity levels (mean around 0.5, see Appendix S1:
465 Figure S3) in the dataset, potentially due to the low-pass sequencing (mean 0.5x coverage) of the
466 RILs (Ray et al., 2023). The significant data gaps (about 95% missing data) necessitated
467 imputation from parental lines, which weren't highly inbred, likely inflating heterozygosity
468 estimates in the RILs. While the imputed genotype matrix was accurate for known QTLs (Ray et
469 al., 2023), the sequencing and imputation approach introduces uncertainty in true heterozygosity
470 levels. Factors like residual heterozygosity and heterozygote hotspots, seen in species like maize
471 (Liu et al., 2018), may also influence these high heterozygosity readings in the *N. attenuata*
472 MAGIC RIL population. Future studies with higher sequencing depth could more accurately
473 determine homozygosity and heterozygosity in this population, aiding in dissecting genetic
474 architecture and validating GWAS findings.

475 Looking forward, there is great potential for deepening our understanding of leaf spectral data
476 using the *N. attenuata* system. Our analysis indicates relatively large variance in blue light
477 reflectance and perhaps related photosynthetic traits among the wild genotypes contributing to
478 the *N. attenuata* MAGIC population, as discussed previously (Li et al., 2023). Variation in
479 conserved and essential traits, which can be mediated by variants of regulatory genes embedded
480 in coexpression networks (Ray et al., 2023), may lend different genotypes a competitive
481 advantage under changing environmental conditions. The significant associations identified here
482 thus indicate genotype and phenotype targets for novel and interesting in-depth functional studies
483 that could help us to better understand plant survival strategies. A combination of natural
484 variants, genetic modification, and laboratory assays, such as gene expression analyses (Schena
485 et al., 1995), protein-protein interactions (Fields and Song 1989), and chromatin
486 immunoprecipitation assays (Solomon et al., 1988), comprise the gold standard for functional

487 testing and validation of discovered genetic associations. At this stage, the associations reported
488 here represent hypotheses that should be subjected to these tests.

489 Unraveling the genetic basis of leaf spectral properties has far-reaching implications and could
490 for example revolutionize plant biodiversity monitoring, ecophysiology, and adaptation research,
491 and help to identify genotypes resilient to stresses associated with global change. Our study sets
492 the stage for future research aimed at deciphering these links by presenting a widely applicable
493 new method for data-driven discovery of associations which can be integrated with existing and
494 novel tools to discover genetic associations and support moving from association to prediction or
495 causation.

496 5 CONCLUSIONS

497 In our study, we explored the genetic basis of leaf spectral properties in *N. attenuata* by
498 associating spectral phenotypes with SNP genotypes in GWAS during a field experiment. We
499 analyzed leaf reflectance spectra from 400 to 2500 nm, identifying regions correlating with leaf
500 composition and function. We developed the Hierarchical Spectral Clustering with Parallel
501 Analysis (HSC-PA) method to manage the data's high dimensionality and correlation, proving
502 more effective than other methods like spectral indices or individual wavelengths. This approach
503 offered greater statistical power and maintained interpretable features, despite reduced
504 dimensionality.

505 Our findings included a significant SNP association in the red-edge to near-infrared spectral
506 range (734–1143 nm), a region sensitive to chlorophyll content, canopy structure, and
507 photosynthetic efficiency. The most significant SNP lies near the Niat3g_06527 gene, annotated
508 as carbonic anhydrase, an enzyme central to photosynthetic carbon fixation and stomatal

regulation. These results require further validation, but they raise the intriguing possibility that variation in red-edge/NIR reflectance could be linked to differences in carbon assimilation and water-use strategies in *N. attenuata*, potentially reflecting adaptations to its native, resource-variable environments.

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 experiment 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 ACKNOWLEDGMENTS

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.

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: B.S., C.L., R.R. Formal analysis: B.S., C.L. Funding acquisition: I.T.B., M.E.S. Investigation: E.A.C., M.C.S. Methodology: B.S., 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: B.S., C.L., E.A.C., I.T.B., M.C.S., M.E.S., R.H., R.R.

8 CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

9 REFERENCES

Adam, Elhadi, Onisimo Mutanga, and Denis Rugege. 2010. "Multispectral and Hyperspectral Remote Sensing for Identification and Mapping of Wetland Vegetation: A Review." *Wetlands Ecology and Management* 18: 281–96.

Asner, Gregory P, Roberta E Martin, Christopher B Anderson, and David E Knapp. 2015. "Quantifying Forest Canopy Traits: Imaging Spectroscopy Versus Field Survey." *Remote Sensing of Environment* 158: 15–27.

Asner, Gregory P, Roberta E Martin, Loreli Carranza-Jiménez, Felipe Sinca, Raul Tupayachi, Christopher B Anderson, and Paola Martinez. 2014. "Functional and Biological Diversity of Foliar Spectra in Tree Canopies Throughout the Andes to Amazon Region." *New Phytologist* 204 (1): 127–39.

Badger, M. R., & Price, G. D. (1994). The role of carbonic anhydrase in photosynthesis. *Annual review of plant physiology and plant molecular biology*, *45*(1), 369-392.

Bahulikar, Rahul A, Dominic Stanculescu, Catherine A Preston, and Ian T Baldwin. 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): 1–13.

Bandillo, Nonoy, Chitra Raghavan, Pauline Andrea Muyco, Ma Anna Lynn Sevilla, Irish T Lobina, Christine Jade Dilla-Ermita, Chih-Wei Tung, et al. 2013. "Multi-Parent Advanced Generation Inter-Cross (MAGIC) Populations in Rice: Progress and Potential for Genetics Research and Breeding." *Rice* 6 (1): 1–15.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1): 289–300.

Brieuc, Marine S. O., Charles D. Waters, Daniel P. Drinan, and Kerry A. Naish. 2018. "A Practical Introduction to Random Forest for Genetic Association Studies in Ecology and Evolution." *Molecular Ecology Resources* 18 (4): 755–66. Cavender-Bares, Jeannine, Jose Eduardo Meireles, John J Couture, Matthew A Kaproth, Clayton C Kingdon, Aditya Singh, Shawn P Serbin, et al. 2016. "Associations of Leaf Spectra with Genetic and Phylogenetic Variation in Oaks: Prospects for Remote Detection of Biodiversity." *Remote Sensing* 8 (3): 221.

Claes, Peter, Jasmien Roosenboom, Julie D White, Tomek Swigut, Dzemila Sero, Jiarui Li, Myoung Keun Lee, et al. 2018. "Genome-Wide Mapping of Global-to-Local Genetic Effects on Human Facial Shape." *Nature Genetics* 50 (3): 414–23.

Czyż, Ewa A, Carla Guillén Escribà, Hendrik Wulf, Andrew Tedder, Meredith C Schuman, Fabian D Schneider, and Michael E Schaepman. 2020. "Intraspecific Genetic Variation of a Fagus Sylvatica Population in a Temperate Forest Derived from Airborne Imaging Spectroscopy Time Series." *Ecology and Evolution* 10 (14): 7419–30.

Czyż, Ewa A, Bernhard Schmid, Andreas Hueni, Maarten B Eppinga, Meredith C Schuman, Fabian D Schneider, Carla Guillén-Escribà, and Michael E Schaepman. 2023. "Genetic

Constraints on Temporal Variation of Airborne Reflectance Spectra and Their Uncertainties over a Temperate Forest." *Remote Sensing of Environment* 284: 113338.

Fabre, N., Reiter, I. M., BECUWE-LINKA, N. O. E. L. L. E., Genty, B., & Rumeau, D. (2007). Characterization and expression analysis of genes encoding α and β carbonic anhydrases in Arabidopsis. *Plant, cell & environment*, *30*(5), 617-629.

Dechant, Benjamin, Youngryel Ryu, Grayson Badgley, Yelu Zeng, Joseph A Berry, Yongguang Zhang, Yves Goulas, et al. 2020. "Canopy Structure Explains the Relationship Between Photosynthesis and Sun-Induced Chlorophyll Fluorescence in Crops." *Remote Sensing of Environment* 241: 111733.

DeLancey, Evan R, John F Simms, Masoud Mahdianpari, Brian Brisco, Craig Mahoney, and Jahan Kariyeva. 2019. "Comparing Deep Learning and Shallow Learning for Large-Scale Wetland Classification in Alberta, Canada." *Remote Sensing* 12 (1): 2.

Fields, Stanley, and Ok-kyu Song. 1989. "A Novel Genetic System to Detect Protein–Protein Interactions." *Nature* 340 (6230): 245–46.

Furbank, R. T., & Tester, M. (2011). Phenomics–technologies to relieve the phenotyping bottleneck. *Trends in plant science*, *16*(12), 635-644.

Gamon, John A, K Fred Huemmrich, Christopher YS Wong, Ingo Ensminger, Steven Garrity, David Y Hollinger, Asko Noormets, and Josep Peñuelas. 2016. "A Remotely Sensed Pigment Index Reveals Photosynthetic Phenology in Evergreen Conifers." *Proceedings of the National Academy of Sciences* 113 (46): 13087–92.

Gao, Bo-Cai. 1996. "NDWI—a Normalized Difference Water Index for Remote Sensing of Vegetation Liquid Water from Space." *Remote Sensing of Environment* 58 (3): 257–66.

Gitelson, Anatoly A, Yuri Gritz, and Mark N Merzlyak. 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 (3): 271–82.

Gitelson, Anatoly A, Mark N Merzlyak, and Olga B Chivkunova. 2001. “Optical Properties and Nondestructive Estimation of Anthocyanin Content in Plant Leaves¶.” *Photochemistry and Photobiology* 74 (1): 38–45.

Gitelson, Anatoly A, Andrés Viña, Verónica Ciganda, Donald C Rundquist, and Timothy J Arkebauer. 2005. “Remote Estimation of Canopy Chlorophyll Content in Crops.” *Geophysical Research Letters* 32 (8).

Glawe, Grit A, Jorge A Zavala, André Kessler, Nicole M Van Dam, and Ian T Baldwin. 2003. “Ecological Costs and Benefits Correlated with Trypsin Protease Inhibitor Production in Nicotiana Attenuata.” *Ecology* 84 (1): 79–90.

Grabska, Ewa, Patrick Hostert, Dirk Pflugmacher, and Katarzyna Ostapowicz. 2019. “Forest Stand Species Mapping Using the Sentinel-2 Time Series.” *Remote Sensing* 11 (10): 1197.

Grzybowski, M., Wijewardane, N. K., Atefi, A., Ge, Y., & Schnable, J. C. (2021). Hyperspectral reflectance-based phenotyping for quantitative genetics in crops: Progress and challenges. *Plant Communications*, *2*(4).

Gu, Yingxin, Jesslyn F Brown, James P Verdin, and Brian Wardlow. 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 (6).

Hatch, M. D., & Burnell, J. N. (1990). Carbonic anhydrase activity in leaves and its role in the first step of C4 photosynthesis. *Plant Physiology*, *93*(2), 825-828.

Helfenstein, Isabelle S., Fabian D. Schneider, Michael E. Schaepman, and Felix Morsdorf. 2022. "Assessing Biodiversity from Space: Impact of Spatial and Spectral Resolution on Trait-Based Functional Diversity." Remote Sensing of Environment 275: 113024. Elsevier.

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., ... & Whitlock, M. C. (2016). Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *The American Naturalist*, *188*(4), 379-397.

Huang, Meng, Xiaolei Liu, Yao Zhou, Ryan M Summers, and Zhiwu Zhang. 2019. "BLINK: A Package for the Next Level of Genome-Wide Association Studies with Both Individuals and Markers in the Millions." *Gigascience* 8 (2): giy154.

Huete, Alfredo R. 2012. "Vegetation Indices, Remote Sensing and Forest Monitoring." *Geography Compass* 6 (9): 513–32.

Huynh, Bao-Lam, Jeffrey D Ehlers, Bevan Emma Huang, María Muñoz-Amatriaín, Stefano Lonardi, Jansen RP Santos, Arsenio Ndeve, 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 (6): 1129–42.

INH, SƠN TRƯỜNG, Ivan Galis, and Ian T Baldwin. 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 (3): 590–606.

Jacquemoud, Stéphane, and Susan Ustin. 2019. *Leaf Optical Properties*. Cambridge University Press.

Karthikeyan, L, Ila Chawla, and Ashok K Mishra. 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.

Kim, Sang-Gyu, Felipe Yon, Emmanuel Gaquerel, Jyotasana Gulati, and Ian T Baldwin. 2011. "Tissue Specific Diurnal Rhythms of Metabolites and Their Regulation During Herbivore Attack in a Native Tobacco, Nicotiana Attenuata." *PloS One* 6 (10): e26214.

Knipling, Edward B. 1970. "Physical and Physiological Basis for the Reflectance of Visible and Near-Infrared Radiation from Vegetation." *Remote Sensing of Environment* 1 (3): 155–59.

Li, Cheng, Ewa A Czyz, Rayko Halitschke, Ian T Baldwin, Michael E Schaepman, and Meredith C Schuman. 2023. "Evaluating Potential of Leaf Reflectance Spectra to Monitor Plant Genetic Variation."

Li, Cheng, Ewa A Czyż, Rayko Halitschke, Ian T Baldwin, Michael E Schaepman, and Meredith C Schuman. 2023. "UZH SG Nicotiana Attenuata ASD Data." SPECCHIO. *http://sc22.geo.uzh.ch:8080/SPECCHIO_Web_Interface/search*.

Li, Dapeng, Sven Heiling, Ian T Baldwin, and Emmanuel Gaquerel. 2016. "Illuminating a Plant's Tissue-Specific Metabolic Diversity Using Computational Metabolomics and Information Theory." *Proceedings of the National Academy of Sciences* 113 (47): E7610–18.

Li, Xu, Qin Wang, Xuhong Yu, Hongtao Liu, Huan Yang, Chenxi Zhao, Xuanming Liu, 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 (51): 20844–49.

Libbrecht, Maxwell W, and William Stafford Noble. 2015. "Machine Learning Applications in Genetics and Genomics." *Nature Reviews Genetics* 16 (6): 321–32.

662 Liu, Nannan, Jianxiao Liu, Wenqiang Li, Qingchun Pan, Jie Liu, Xiaohong Yang, Jianbing Yan,
663 and Yingjie Xiao. 2018. “Intraspecific Variation of Residual Heterozygosity and Its Utility for
664 Quantitative Genetic Studies in Maize.” *BMC Plant Biology* 18: 1–15.
665 Liu, Xiaolei, Meng Huang, Bin Fan, Edward S Buckler, and Zhiwu Zhang. 2016. “Iterative
666 Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide
667 Association Studies.” *PLoS Genetics* 12 (2): e1005767.
668 Madritch, Michael D, Clayton C Kingdon, Aditya Singh, Karen E Mock, Richard L Lindroth,
669 and Philip A Townsend. 2014. “Imaging Spectroscopy Links Aspen Genotype with Below-
670 Ground Processes at Landscape Scales.” *Philosophical Transactions of the Royal Society B:*
671 *Biological Sciences* 369 (1643): 20130194.
672 Meireles, Jose Eduardo, Jeannine Cavender-Bares, Philip A. Townsend, Susan Ustin, John A.
673 Gamon, Anna K. Schweiger, Michael E. Schaepman et al. "Leaf reflectance spectra capture the
674 evolutionary history of seed plants." New Phytologist 228, no. 2 (2020): 485-493.
675 Meireles, Jose Eduardo, Anna K Schweiger, and Jeannine M Cavender-Bares. 2017. “Spectrolab:
676 Class and Methods for Hyperspectral Data. R Package Version 0.0. 2.”
677 Miller, JR, MD Steven, and TH Demetriades-Shah. 1992. “Reflection of Layered Bean Leaves
678 over Different Soil Backgrounds: Measured and Simulated Spectra.” *International Journal of*
679 *Remote Sensing* 13 (17): 3273–86.
680 Nyholt, Dale R. 2004. “A Simple Correction for Multiple Testing for Single-Nucleotide
681 Polymorphisms in Linkage Disequilibrium with Each Other.” *The American Journal of Human*
682 *Genetics* 74 (4): 765–69.

683 Price, Alkes L, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and
684 David Reich. 2006. "Principal Components Analysis Corrects for Stratification in Genome-Wide
685 Association Studies." *Nature Genetics* 38 (8): 904–9.
686 Qian, X, L Liu, H Croft, and J Chen. 2021. "Relationship Between Leaf Maximum
687 Carboxylation Rate and Chlorophyll Content Preserved Across 13 Species." *Journal of*
688 *Geophysical Research: Biogeosciences* 126 (2): e2020JG006076.
689 R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna,
690 Austria: R Foundation for Statistical Computing. *https://www.R-project.org/*.
691 Ray, R. et al. 2022. "Whole Genome Sequence of MAGIC Parents and RIL Population." NCBI
692 Sequence Read Archive. *https://www.ncbi.nlm.nih.gov/bioproject/PRJNA907539*.
693 Ray, Rishav, Rayko Halitschke, Klaus Gase, Sabrina M Leddy, Meredith C Schuman, Nathalie
694 Rodde, and Ian T Baldwin. 2023. "A Persistent Major Mutation in Canonical Jasmonate
695 Signaling Is Embedded in an Herbivory-Elicited Gene Network." *Proceedings of the National*
696 *Academy of Sciences* 120 (35): e2308500120.
697 Ray, Rishav, Dapeng Li, Rayko Halitschke, and Ian T Baldwin. 2019. "Using Natural Variation
698 to Achieve a Whole-Plant Functional Understanding of the Responses Mediated by Jasmonate
699 Signaling." *The Plant Journal* 99 (3): 414–25.
700 Rudenko, N. N., Ignatova, L. K., Fedorchuk, T. P., & Ivanov, B. N. (2015). Carbonic anhydrases
701 in photosynthetic cells of higher plants. *Biochemistry (Moscow)*, *80*(6), 674-687.
702 Saha, Saswati, Laurent Perrin, Laurence Röder, Christine Brun, and Lionel Spinelli. 2022. "Epi-
703 MEIF: Detecting Higher Order Epistatic Interactions for Complex Traits Using Mixed Effect
704 Conditional Inference Forests." *Nucleic Acids Research* 50 (19): e114.

Sannemann, Wiebke, Bevan Emma Huang, Boby Mathew, and Jens Léon. 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.

Santure, A. W., & Garant, D. (2018). Wild GWAS—association mapping in natural populations. *Molecular ecology resources*, *18*(4), 729-738.

Schena, Mark, Dari Shalon, Ronald W Davis, and Patrick O Brown. 1995. “Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray.” *Science* 270 (5235): 467–70.

Serbin, Shawn P, Aditya Singh, Brenden E McNeil, Clayton C Kingdon, and Philip A Townsend. 2014. “Spectroscopic Determination of Leaf Morphological and Biochemical Traits for Northern Temperate and Boreal Tree Species.” *Ecological Applications* 24 (7): 1651–69.

Sero, Dzemila, Arslan Zaidi, Jiarui Li, Julie D White, Tomás B González Zarzar, Mary L Marazita, Seth M Weinberg, et al. 2019. “Facial Recognition from DNA Using Face-to-DNA Classifiers.” *Nature Communications* 10 (1): 2557.

Solomon, Mark J, Pamela L Larsen, and Alexander Varshavsky. 1988. “Mapping proteinDNA Interactions in Vivo with Formaldehyde: Evidence That Histone H4 Is Retained on a Highly Transcribed Gene.” *Cell* 53 (6): 937–47.

Stasinski, Lance, Dawson M White, Peter R Nelson, Richard H Ree, and José Eduardo Meireles. 2021. “Reading Light: Leaf Spectra Capture Fine-Scale Diversity of Closely Related, Hybridizing Arctic Shrubs.” *New Phytologist* 232 (6): 2283–94.

Supuran, C. T. (2004). Carbonic anhydrases: catalytic and inhibition mechanisms, distribution and physiological roles. In *Carbonic anhydrase* (pp. 13-36). CRC press.

Taiz, Lincoln, Eduardo Zeiger, Ian Max Møller, Angus Murphy, et al. 2015. *Plant Physiology and Development.* Ed. 6. Sinauer Associates Incorporated.

Thenkabail, Prasad S, John G Lyon, and Alfredo Huete. 2018. "Advances in Hyperspectral Remote Sensing of Vegetation and Agricultural Crops." In *Fundamentals, Sensor Systems, Spectral Libraries, and Data Mining for Vegetation*, 3–37. CRC press.

Tiwari, A., Mamedov, F., Grieco, M., Suorsa, M., Jajoo, A., Styring, S., ... & Aro, E. M. (2016). Photodamage of iron–sulphur clusters in photosystem I induces non-photochemical energy dissipation. *Nature Plants*, *2*(4), 1-9.

Verrelst, Jochem, Zbyněk Malenovskỳ, Christiaan Van der Tol, Gustau Camps-Valls, Jean-Philippe Gastellu-Etchegorry, Philip Lewis, Peter North, and Jose Moreno. 2019. "Quantifying Vegetation Biophysical Variables from Imaging Spectroscopy Data: A Review on Retrieval Methods." *Surveys in Geophysics* 40: 589–629.

Visscher, Peter M, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 2017. "10 Years of GWAS Discovery: Biology, Function, and Translation." *The American Journal of Human Genetics* 101 (1): 5–22.

Wan, Liang, Zheng Tang, Jiafei Zhang, Shuobo Chen, Weijun Zhou, and Haiyan Cen. 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, Jiabo, and Zhiwu Zhang. 2021. "GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction." *Genomics, Proteomics & Bioinformatics* 19 (4): 629–40.

Wang, Ran, and John A Gamon. 2019. "Remote Sensing of Terrestrial Plant Biodiversity." *Remote Sensing of Environment* 231: 111218.

Wang, Ran, John A Gamon, Anna K Schweiger, Jeannine Cavender-Bares, Philip A Townsend, Arthur I Zygielbaum, and Shan Kothari. 2018. "Influence of Species Richness, Evenness, and Composition on Optical Diversity: A Simulation Study." *Remote Sensing of Environment* 211: 218–28.

Wang, Yue, Wilson Goh, Limsoon Wong, Giovanni Montana, and the Alzheimer's Disease Neuroimaging Initiative. 2013. "Random Forests on Hadoop for Genome-Wide Association Studies of Multivariate Neuroimaging Phenotypes." *BMC Bioinformatics* 14 (16): S6.

Wang, Zhihui, Adam Chlus, Ryan Geygan, Zhiwei Ye, Ting Zheng, Aditya Singh, John J Couture, Jeannine Cavender-Bares, Eric L Kruger, and Philip A Townsend. 2020. "Foliar Functional Traits from Imaging Spectroscopy Across Biomes in Eastern North America." *New Phytologist* 228 (2): 494–511.

Weigelt, Alexandra, Liesje Mommer, Karl Andraczek, Colleen M. Iversen, Joana Bergmann, Helge Bruelheide, Grégoire T. Freschet, et al. 2023. "The Importance of Trait Selection in Ecology." *Nature* 618 (7967): E29–30.

Wu, Di, Deepti S Karhade, Malvika Pillai, Min-Zhi Jiang, Le Huang, Gang Li, Hunyong Cho, Jeff Roach, Yun Li, and Kimon Divaris. 2021. "Machine Learning and Deep Learning in Genetics and Genomics." *Machine Learning in Dentistry*, 163–81.

Yu, Jianming, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, et al. 2006. "A Unified Mixed-Model Method for Association Mapping That Accounts for Multiple Levels of Relatedness." *Nature Genetics* 38 (2): 203–8.

Zhang, F, and G Zhou. 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.

773 Zhang, Zhiwu, Elhan Ersoz, Chao-Qiang Lai, Rory J Todhunter, Hemant K Tiwari, Michael A
774 Gore, Peter J Bradbury, et al. 2010. "Mixed Linear Model Approach Adapted for Genome-Wide
775 Association Studies." *Nature Genetics* 42 (4): 355–60.
776 Zheng, J., Richardson, T. G., Millard, L. A., Hemani, G., Elsworth, B. L., Raistrick, C. A., ... &
777 Gaunt, T. R. (2018). PhenoSpD: an integrated toolkit for phenotypic correlation estimation and
778 multiple testing correction using GWAS summary statistics. *Gigascience*, *7*(8), giy090.

10 TABLES

Table 1. Six Spectral Indices Used in Indices-GWAS

Index	Formula	Example of Validated Systems	Citations
$NDWI$	$\frac{R_{865} - R_{1614}}{R_{865} + R_{1614}}$	Crops, forests, wetlands, grasslands	(Gao 1996; F. Zhang and Zhou 2019; Karthikeyan, Chawla, and Mishra 2020; Huete 2012; Adam, Mutanga, and Rugege 2010; Gu et al. 2007; Helfenstein et al., 2022)
CI_{re}	$\frac{R_{783}}{R_{704}} - 1$	Crops, trees and vines, wetlands	(Gitelson, Merzlyak, and Chivkunova 2001; Qian et al. 2021; Gitelson et al. 2005; Gitelson, Gritz, and Merzlyak 2003; DeLancey et al. 2019; Helfenstein et al., 2022)
CCI	$\frac{R_{560} - R_{664}}{R_{560} + R_{664}}$	Crops, trees, forests	(Gamon et al. 2016; Dechant et al. 2020; Grabska et al. 2019; Helfenstein et al., 2022)
$ARDSI_C_{ab}$	$\frac{R_{750} - R_{730}}{R_{770} + R_{720}}$	Crops	(Wan et al. 2021)
$ARDSI_C_{w}$	$\frac{R_{1360} - R_{1080}}{R_{1560} + R_{1240}}$	Crops	(Wan et al. 2021)
$ARDSI_C_{m}$	$\frac{R_{2200} - R_{1640}}{R_{2240} + R_{1720}}$	Crops	(Wan et al. 2021)

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

Segment	Spectral range (nm)	Potentially associated physiological features*
5	518–614, 695–709	VIS–NIR: Chlorophyll a, chlorophyll b, carotenoids, anthocyanins; leaf structure
10	400–432	VIS: Chlorophyll a, chlorophyll b, carotenoids, anthocyanins
12	433–517	VIS: Chlorophyll a, chlorophyll b, carotenoids, anthocyanins
13	648–690	VIS: Chlorophyll a, chlorophyll b, carotenoids, anthocyanins
14	615–647, 691–694	VIS: Chlorophyll a, chlorophyll b, carotenoids, anthocyanins
19	1894–2000	SWIR: Water content
20	1880–1893, 2001–2096	SWIR: Non-pigment organic composition (e.g., lignin, cellulose, starch, protein, sugar, oil)
22	1400–1401, 1864–1879, 2097–2311	NIR–SWIR: Water content; leaf structure; non-pigment organic composition
23	2397–2500	SWIR: Water content
24	2312–2396	SWIR: Non-pigment organic composition
25	1412–1484	SWIR: Water content
26	1402–1411, 1485–1522	SWIR: Non-pigment organic composition
28	1359–1380	SWIR: Water content
29	1381–1385, 1590–1754	SWIR: Non-pigment organic composition
30	1386–1399, 1523–1589, 1755–1863	SWIR: Non-pigment organic composition
33	1311–1358	NIR: Water content
34	1144–1310	NIR: Leaf structure
36	710–733	NIR: Leaf structure
37	734–744, 946–1014, 1129–1143	NIR: Leaf structure
38	745–945, 1015–1128	NIR: Leaf structure

783 Note: The associated features are based on common findings (Jacquemoud and Ustin 2019) and
784 are intended to indicate physiological properties of leaves known to affect the given spectral
785 ranges. These interpretations are subject to various factors and should not be considered
786 definitive.

11 FIGURE CAPTIONS

Figure 1. Approaches and toolbox of this study. Photo credit: Meredith C. Schuman.

Figure 2. Variance partitioning across the spectral range of 350 to 2500 nm. Each column represents the cumulative percentage of variance for a given wavelength.

Figure 3. Spectra as phenotypes. (a) Leaf reflectance spectra for all 616 plants examined in this study. The orange shaded region illustrates the range of adjusted spectra, and the blue shaded region the range of raw spectra. The orange dotted line represents the coefficient of variation for the entire sample set. (b) Distribution of the six spectral indices used in the Index-GWAS approach. (c) Results from the Hierarchical Spectral Clustering with Parallel Analysis (HSC-PA). A total of 38 segments, differentiated by color, emerged from the clustering. Regions left blank (NA) indicate segments that retained one Principal Component (PC) in the prior level, halting further clustering. (d) Presentation of the 20 segments, shaped as example spectra, each maintaining one retained 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.

Figure 4. Index-GWAS results. (a) Manhattan plot showing the $-log_{10}(p_value)$ of SNP associations for the six indices across various models. The grey horizontal line indicates a p-value 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 six indices, the y-axis the corresponding associated SNPs. The gradient red color scale represents the FDR_adjusted p-values. (c) QQ plot contrasting the observed versus the expected $-log_{10}(p_value)$ for the GWAS analysis on the $ARDSI_C_w$ index. The diagonal line represents

809 the expected distribution under the null hypothesis. The different color schemes depict the
810 models used: red for GLM, yellow for MLM, green for FarmCPU, and blue for BLINK. (d)-(e)
811 Phenotypic distribution for two significant markers, chr1_438155938(d) or chr1_193505583(e),
812 in association with $ARDSI_C_w$. 0 represents a homozygous allele from the reference, 1 a
813 heterozygous allele, and 2 a homozygous alternative allele.

814 Figure 5. SW-GWAS results. (a) Manhattan plot showing the $-log_{10}(p_value)$ of SNP
815 associations with BLINK for wavelength 952 with the strongest association. (b) Visualization of
816 associations that have p-values less than 1×10^{-5}. The x-axis represents the spectral
817 wavelengths from 400 to 2500 nm, while the y-axis displays the corresponding associated SNPs.
818 The gradient red color scale represents the FDR_adjusted p-values. (c) QQ plot of the BLINK
819 model contrasting the observed versus the expected $-log_{10}(p_value)$ for the GWAS analysis on
820 the wavelength 952 nm. The diagonal line represents the expected distribution under the null
821 hypothesis.

822 Figure 6. HSC-PA-GWAS results. (a) Detailed visualization of associations that surpass the
823 exploratory threshold. The X-axis enumerates the 20 segments, while the Y-axis displays the
824 corresponding associated SNPs. (b) An alternative representation of the associations from (a)
825 with the X-axis showcasing the spectral wavelengths, providing insights into the specific spectral
826 regions of the associations. (c)-(d) QQ plots highlighting the observed versus expected
827 $-log_{10}(p_value)$ for GWAS analysis on Segment 37(c) or Segment 38 (d). (e)-(f) Phenotypic
828 distribution for the significantly associated marker (chr1_131409683) in association with
829 Segment 37(e) or Segment 38(f). 0 represents a homozygous allele from the reference, 1 a
830 heterozygous allele, and 2 a homozygous alternative allele.
831

12 FIGURES

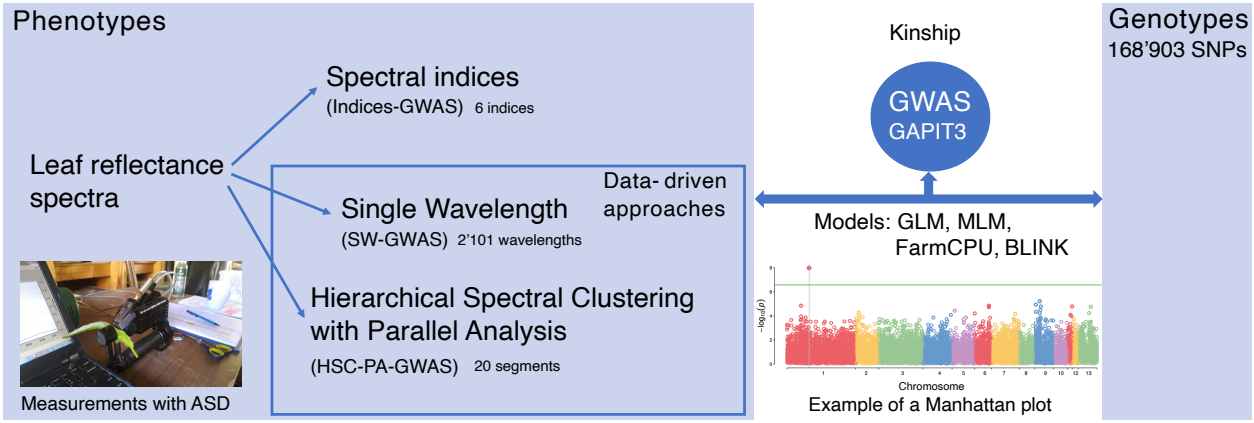

Figure 1.

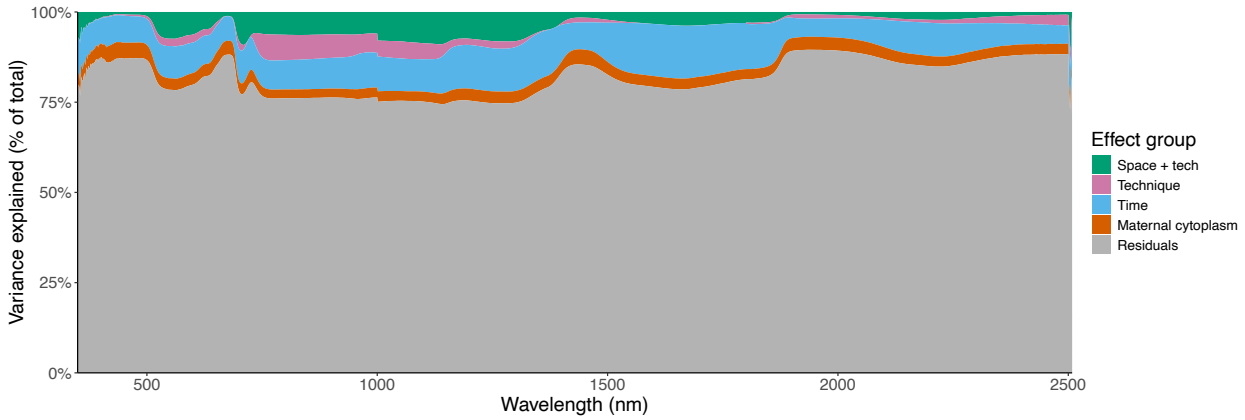
100%
75%
50%
25%
0%
Variance explained (% of total)
500
1000
1500
2000
2500
Wavelength (nm)
Effect group
Space + tech
Technique
Time
Maternal cytoplasm
Residuals

Figure 2.

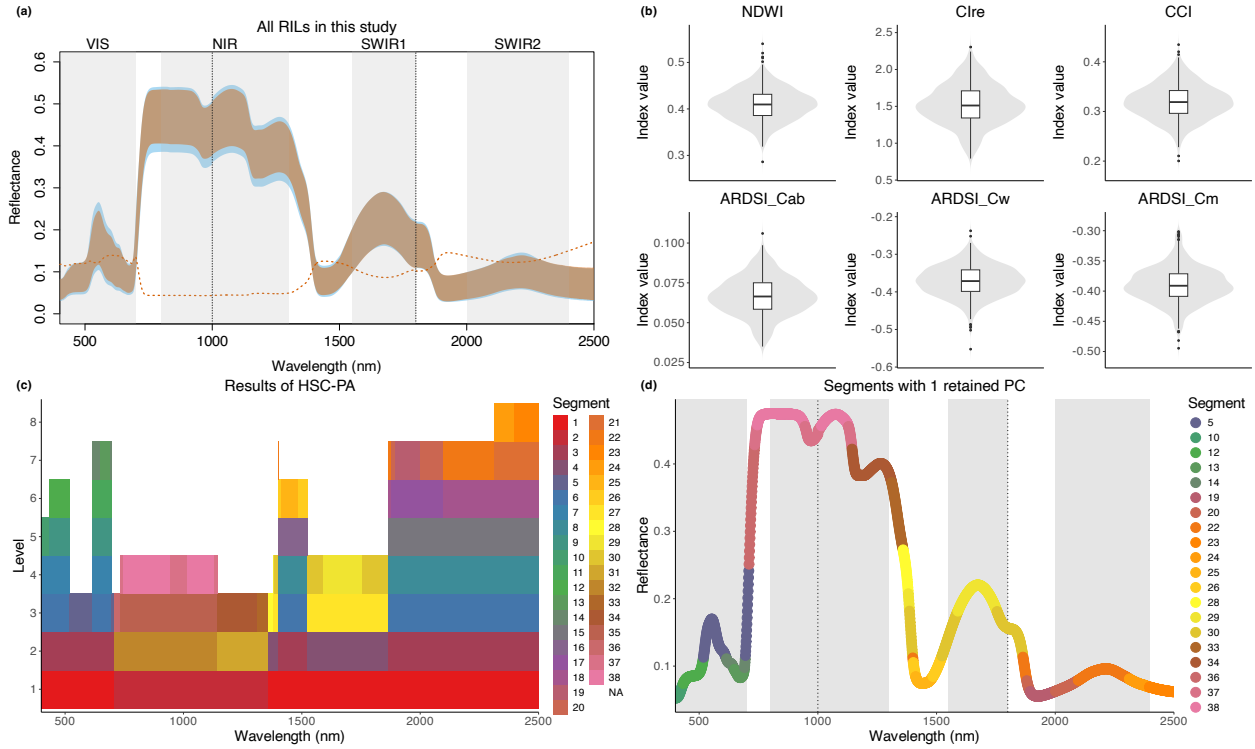
(a)
All RILs in this study
VIS
NIR
SWIR1
SWIR2
Reflectance
Wavelength (nm)
(b)
NDWI
CIre
CCI
ARDSI_Cab
ARDSI_Cw
ARDSI_Cm
Index value
(c)
Results of HSC-PA
Level
Segment
Wavelength (nm)
(d)
Segments with 1 retained PC
Reflectance
Segment
Wavelength (nm)

Figure 3.

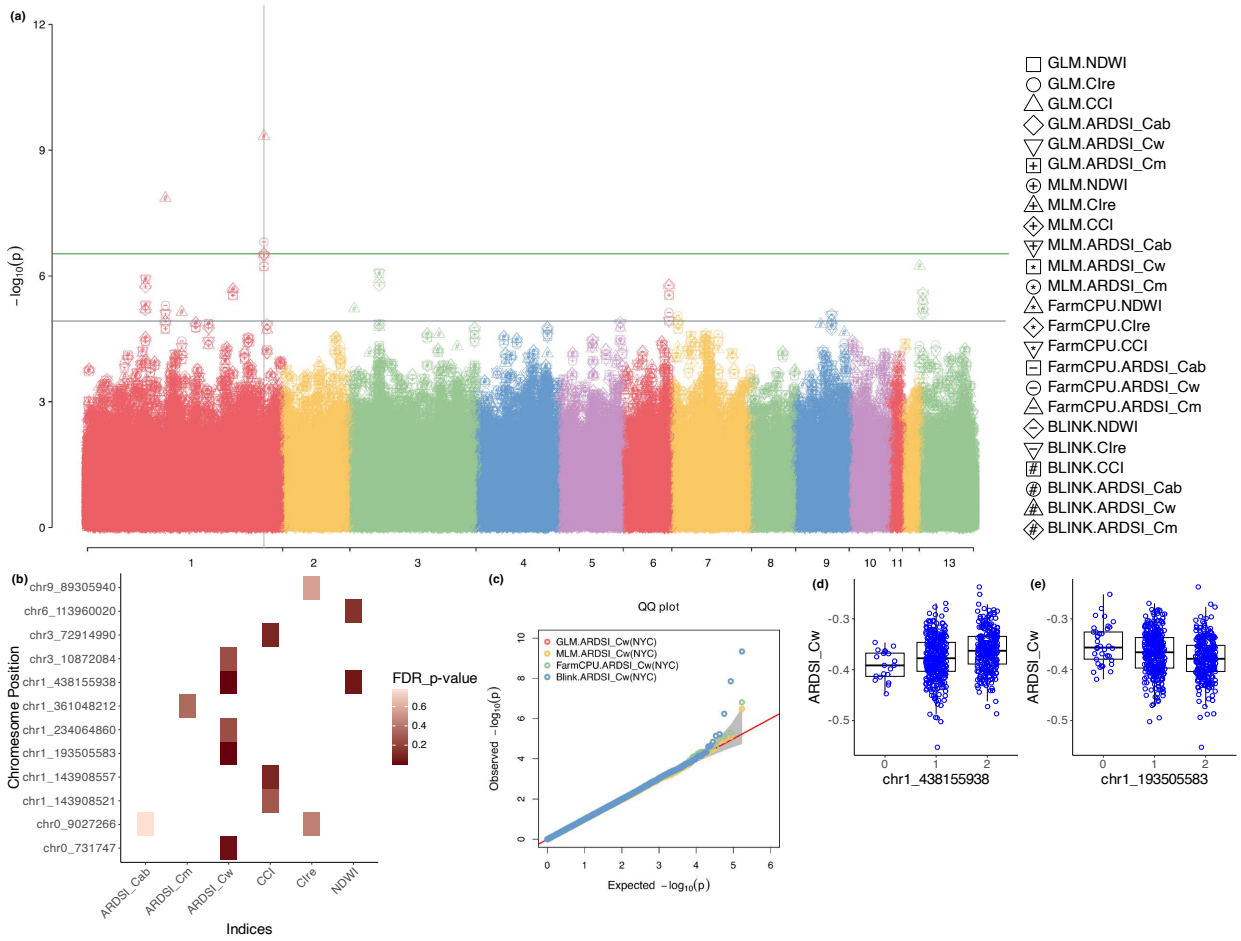
(a)
−log10(p)
GLM.NDWI
GLM.CIre
GLM.CCI
GLM.ARDSI_Cab
GLM.ARDSI_Cw
GLM.ARDSI_Cm
MLM.NDWI
MLM.CIre
MLM.CCI
MLM.ARDSI_Cab
MLM.ARDSI_Cw
MLM.ARDSI_Cm
FarmCPU.NDWI
FarmCPU.CIre
FarmCPU.CCI
FarmCPU.ARDSI_Cab
FarmCPU.ARDSI_Cw
FarmCPU.ARDSI_Cm
BLINK.NDWI
BLINK.CIre
BLINK.CCI
BLINK.ARDSI_Cab
BLINK.ARDSI_Cw
BLINK.ARDSI_Cm
(b)
Chromesome Position
chr9_89305940
chr6_113960020
chr3_72914990
chr3_10872084
chr1_438155938
chr1_361048212
chr1_234064860
chr1_193505583
chr1_143908557
chr1_143908521
chr0_9027266
chr0_731747
ARDSI_Cab
ARDSI_Cm
ARDSI_Cw
CCI
CIre
NDWI
Indices
FDR_p-value
(c)
QQ plot
GLM.ARDSI_Cw(NYC)
MLM.ARDSI_Cw(NYC)
FarmCPU.ARDSI_Cw(NYC)
Blink.ARDSI_Cw(NYC)
Observed −log10(p)
Expected −log10(p)
(d)
ARDSI_Cw
chr1_438155938
(e)
ARDSI_Cw
chr1_193505583

Figure 4.

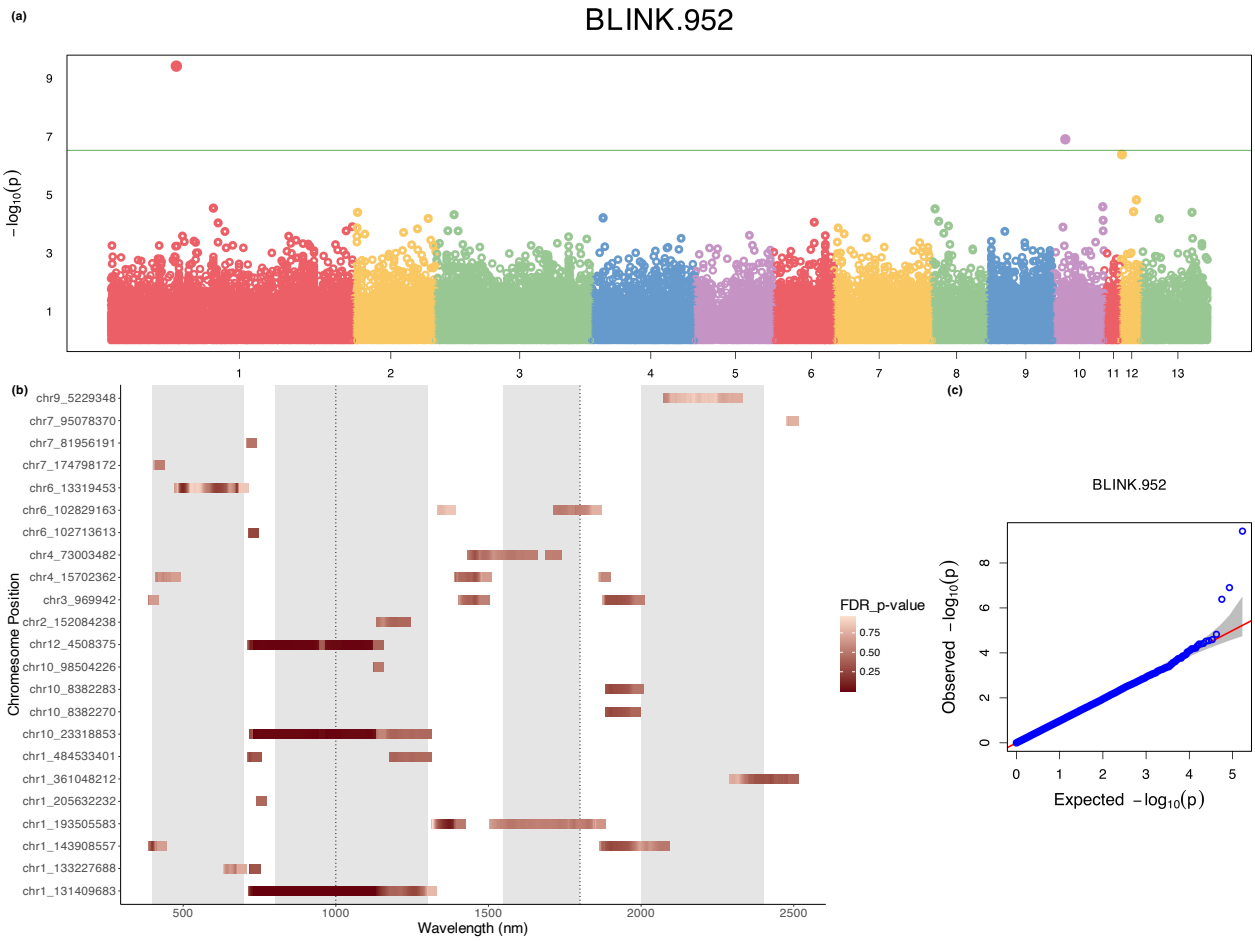
(a)
BLINK.952
$-\log_{10}(p)$
(b)
Chromosome Position
Wavelength (nm)
FDR_p-value
0.75
0.50
0.25
(c)
BLINK.952
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$

Figure 5.

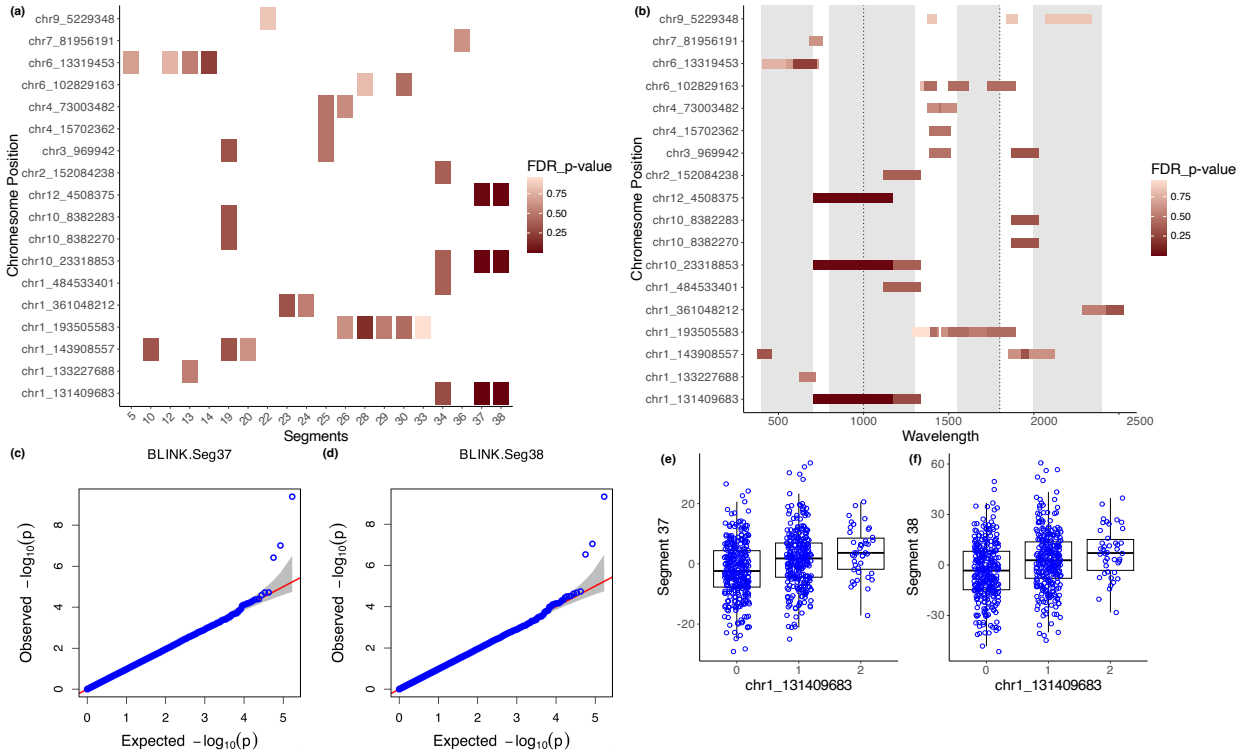
(a)
Chromesome Position
Segments
FDR_p-value
0.75
0.50
0.25
(b)
Chromesome Position
Wavelength
FDR_p-value
0.75
0.50
0.25
(c)
BLINK.Seg37
Observed −log10(p)
Expected −log10(p)
(d)
BLINK.Seg38
Observed −log10(p)
Expected −log10(p)
(e)
Segment 37
chr1_131409683
(f)
Segment 38
chr1_131409683

Figure 6.