Quantitative Plant Biology

cambridge.org/qpb

Using near-infrared spectroscopy (NIRS) to predict budbreak of the following year

Amy Watson[1], Vincent Segura[1], Yoann Bourhis[2], Guillaume Perez[1], Isabelle Farrera[1], Evelyne Costes[1] and Fernando Andrés[1,3]

[1]UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, Montpellier, France; [2]Rothamsted Research, Harpenden, UK; [3]Instituto de Biología Molecular y Celular de Plantas, Consejo Superior de Investigaciones Científicas, Universidad Politécnica de Valencia, Campus de la Universidad Politécnica de Valencia, 46022 Valencia, Spain

Original Research Article

Cite this article: A. Watson et al. Using near-infrared spectroscopy (NIRS) to predict budbreak of the following year. *Quantitative Plant Biology*, **6**:e24, 1–9
https://dx.doi.org/10.1017/qpb.2025.10019

Received: 24 December 2024
Revised: 20 May 2025
Accepted: 13 June 2025

Keywords:
apple; budbreak; near-infrared spectroscopy; partial least squares regression; prediction.

Corresponding author:
Fernando Andrés;
Email: fandres@ibmcp.upv.es

Associate Editor:
Dr. Ross Sozzani

Abstract

Near-infrared spectra (NIRS) from plant tissues can be used to predict traits owing to their relationship to internal biochemical states, shaped by both environmental and genetic components. Here, we tested the use of NIRS as predictors of budbreak the following year. We measured NIRS on leaf and bud tissue, collected at several dates during the growing season, of 240 dessert apple cultivars in 2021 and 2022. NIRS collected in 2021 and budbreak of 2022 were used to train partial least squares (PLSR) models, then tested using NIRS of 2022 to predict budbreak in 2023. A GWAS using these predictions identified a QTL, previously associated to budbreak in apple, indicating a significant genetic component was maintained in the predictions. Our results demonstrate the potential of NIRS to predict future developmental stages, such as budbreak, by detecting the metabolic states that precede them and could aid in genetic studies of difficult-to-measure traits.

1. Introduction

The timing of budbreak in temperate fruit trees has an immense bearing on final fruit production due to the interconnected relationship between seasonal climate conditions and reproductive physiology. Temperature provides one of the most important environmental signals for the beginning and end of dormancy in apple; however, its variability inevitably affects the timing of bud and floral development year to year. The ability to predict this timing facilitates management practices such as those able to mitigate damaging effects of early frosts (Cannell & Smith, 1986; Legave et al., 2008). Furthermore, as climate change advances flowering to varying degrees in many regions (Legave et al., 2013), the development of tools to improve this prediction is becoming increasingly useful.

The development of apple floral and vegetative buds begins the year prior to their eventual opening during spring, when floral induction (or not) determines the fate of meristems present on shoots produced in the spring/summer of that year (Hanke et al., 2007). Buds then form to protect the meristems over winter and, following fulfilment of the chilling and heating requirements of the cultivar to overcome endodormancy and ecodormancy (Lang et al., 1987), respectively, budbreak occurs in spring of the following season. The leaves also undergo transformation during the period from summer to autumn as the gradual reduction in temperature and photoperiod drive the re-uptake of nutrients, leading to leaf senescence and abscission. All these changes, in both buds and leaves, are the result of ongoing biochemical processes regulated by both the genetics of the cultivar and the environmental conditions. Measurements of the metabolites produced leading up to budbreak found a close relationship between their occurrence and final budbreak (Beauvieux et al., 2018; Dhuli et al., 2014).

Near-infrared spectroscopy (NIRS) refers to the region of the electromagnetic spectrum from 800 to 2500 nm and has been extensively used in plant trait prediction due to its relationship with the biochemical composition of the tissues on which it is measured. The traits predicted are usually those indicative of the current components or condition of the tissue, for example, the dry matter content of avocado (Rodríguez et al., 2023), apple flesh texture (Wang et al., 2024) or maize protein composition (Rosales et al., 2011). In addition, the high level of complex endophenotypic information captured by NIRS means a significant genetic component can

be represented in the spectra, a characteristic that underlies the developing field of phenomic prediction. Here, the kinship matrix is constructed using the NIRS spectra of the various genotypes, rather than genetic markers, to derive a breeding value associated with a particular trait (Rincent et al., 2018). This approach has shown promise in a variety of crops, for a number of traits (Brault et al., 2022; Lane et al., 2020; Rincent et al., 2018; Robert et al., 2022).

NIRS spectra can also be used to predict traits that have not yet occurred within the same plant. NIRS measurements on unpicked fruit in the field may reflect components involved in ripening or maturity, thereby enabling the prediction of the best harvest date, as in the case of peach (Minas et al., 2021). Another example is seed germination and vigour, which may be predicted due to NIRS mirroring the levels of important growth compounds, e.g. protein and starch, as has been shown in soybean (Al-Amery et al., 2018). In these two latter studies, despite the NIRS measurement and the target trait being separated temporally, the biochemical fingerprint of the tissue could be linked to the underlying physiological progression and an associated outcome predicted. We wanted to explore this link in the case of apple and attempt to extend the application of NIRS to the prediction of the timing of future flowering stages, evidence of which may be present in the biochemical components of the plant tissues. Using partial least squares regression (PLSR), a multivariate analysis able to deal effectively with high-dimensional, collinear variables, like NIRS (Wold et al., 2001), we made predictions of budbreak timing in an apple collection of diverse French cultivars using leaf and bud NIRS spectra taken the year prior. We then investigated the potential use of these predictions in genetic association studies.

2. Materials and methods

2.1. Plant material

Plant material consisted of 240 cultivars of a French dessert apple (*Malus domestica* Borkh.) core collection, as described by Lassois et al. (2016), and five commercial cultivars, including Gala, Granny Smith, Golden Delicious, Starkrimson and Condessa (list of cultivars given in Supplementary Table S1). All trees were grafted onto the M9 Pajam®2 rootstock and planted out in field conditions in 2014 at the INRAE Diascope experimental unit, near Montpellier, France. Four replicates of each collection cultivar were organised into 10 rows of 100 trees each, with two replicates planted opposite each other in adjacent rows, randomly distributed in the field. Eleven trees required replacement in 2015. Due to tree loss, twenty-four cultivars had only three replicates and two cultivars had two. Commercial cultivars were dispersed through the collection with between five and ten replicates each, making a final count of 965 trees.

2.2. Phenotyping of flowering stages

Several flowering stages were phenotyped to provide a range of potential prediction scenarios to test. The timing of phenological stages was collected in both 2022 and 2023 (Farrera et al., 2024) and was those as defined by Baggiolini (1980). The timing of budbreak (including both vegetative and floral buds) was determined as the number of days from January 1st until approximately 10% of buds on the tree had reached stage C3 (at least 10 mm of leaf tips had emerged past bud scales) and was recorded for all trees. Due to alternate bearing cycles, approximately 20% of trees did not produce any or very few floral meristems each year, thereby preventing phenotyping of flowering progression. Therefore, pink bud stage (buds were sufficiently open for the pink colouration of the petals to be seen, stage E2), full bloom (inflorescences have fully emerged from the bud, stage F2) and floral decline (first petals fall, stage G) were recorded on only 804 and 788 trees in 2022 and 2023, respectively. These four stages will now be referred to as stages C, E, F and G, respectively. Correlations between 2022 and 2023 flowering stages were calculated from the raw phenotypes of each tree as the Pearson correlation coefficient, using R statistical software (R Core Team, 2024). All further analyses were performed in this software. Only stage C was used in the subsequent model training and testing phases due to the high correlations of this trait with the other flowering stages.

2.3. Sample collection and processing

Leaf sampling was performed in the same three months in 2021 and 2022, specifically, late June (23 June, both years) to coincide with leaf expansion, and late September (23 and 28 September, respectively) and mid-November (19 and 10 November, respectively) to cover the period of senescence progression. All trees were sampled by removing eight leaves: four from each side, with two from higher in the canopy and two from lower down in order to achieve a representative sample of the whole canopy. These leaves were then stacked and, while avoiding the midribs, a handheld coring tool was used to extract eight leaf disks (10 mm diameter). These disks were pooled in a single tube and immediately snap frozen in liquid nitrogen. Buds were sampled at two timepoints: 14 December 2021 and 19 January 2022, which coincided with the approximate entry into endodormancy and the late stages of endodormancy, respectively. Ten buds were removed per tree, five per side, and buds were cut 3–4 mm below the approximate position of the meristem. All buds per tree were pooled in a single tube, already immersed in liquid nitrogen.

Bud and leaf samples were freeze-dried for 72 hours in a Cryotec lyophilizator (Cryotec, Montpellier, France) and then ground to a powder consistency with stainless steel beads using a SPEX SamplePrep 2010 Geno/Grinder tissue homogenizer (SPEX CertiPrep, Stanmore, UK). Bud samples required an initial grinding step prior to the homogenizer using a standing electric drill and drill bit. Once samples were fully ground, they were kept at room temperature.

2.4. Near-infrared spectrometry (NIRS) measurement and spectra processing

NIRS was measured on all samples from the seven collection dates using an ASD LabSpec 4 Standard-Res Lab analyser (PANalytical, Almelo, Netherlands) and the associated spectra acquisition software, IndicoPro (Version 6.5, Malvern Panalytical). Reflectance was recorded from 400 to 2500 nm with a 1 nm spectral resolution. The full dataset is available from Perez et al. (2025).

Spectra were loaded into R software using the functions asd_read_dir and asd_read from Ecarnot (2023). An initial clean-up of all sample spectra was carried out, which included a linear extrapolation adjustment at two points in the spectra in order to smooth the jumps that occur during the transition between detectors of the spectrometer (adj_asd function from Ecarnot, 2023) and discarding of reflectance readings from 400 to 500 nm, which were unstable. The following four noise-reducing, pre-processing procedures were then performed for each of the eight raw spectra datasets: standard normal variate (snv), which

centres and scales the reflectance readings to correct for light scattering; detrend (dt), using the prospectr R package (Stevens & Ramirez-Lopez, 2024), which removes low-frequency, fluctuating trends in the data, often due to instrumental shifts during multiple measurements (Barnes et al., 1989); and first and second derivative (der1 and der2, respectively), using the signal R package (Signal Developers, 2023; window length of 41 nm), which remove constant background signals from the spectra thereby improving peak resolution. Along with the raw spectra, this rendered five separate spectra matrices for each of the eight collection dates. A flowchart of the process is given in Supplementary Figure S1.

2.5. Partial least squares regression (PLSR) model selection and validation

Three PLSR models were trained to predict the timing of stage C in 2022 using spectra matrices collected on leaf tissue in June, September and November (2021) and the pls R package (Liland et al., 2023). With those models, stage C in 2023 was predicted using the spectra matrices collected in the same months (June, September and November) in 2022 as predictors. In doing so, the transferability of the spectra-based models from one year to the next was tested.

For each of the three models, the model selection procedure aimed at identifying the best noise-reducing pre-processing (raw, snv, dt, der1, der2 and all five concatenated) for the spectra matrices, as well as the optimal number of latent variables (up to 30) for the PLSR algorithm. The model selection followed a 20-iteration bootstrapping method to maximise the R^2 on the test dataset (sometimes known as q^2) while minimising the Root Mean Square Error (RMSE). The model selected was that with the lowest RMSE (Aptula et al., 2005).

The three selected models were then trained on the full dataset, i.e. not bootstrapping (NIRS 2021, stage C 2022), for subsequent application on the data of the next year (NIRS 2022, stage C 2023). Their subsequent performances are a measure of year-to-year transferability. Additionally, two other PLSR models were trained using NIRS collected on bud tissue in December 2021 and January 2022. They underwent model selection but were not advanced to application the following year due to their poor performance.

The data used for the five models is summarised in Table 1. As there was a slight variation in which trees were sampled at each collection date, only samples present in the training (2022) and testing (2023) data within each model were kept. This resulted in varying sample numbers between models, but the same number of samples in both the training and testing data within each model. For example, both the June testing and training data consisted of 929 samples, collected from the same trees in 2021 and 2022, respectively (Table 1). A flowchart of the prediction process and data used is given in Supplementary Figure S2.

Finally, to explore the practical value of the PLSR models, the predictions were also compared to those of a trivial model, where stage C timing of each tree in 2022 was considered equal to that of 2023.

2.6. Calculation of genotypic values and heritability

Predictions of stage C timing were made on a per tree basis, so the genotypic component of these values was calculated for use in genome-wide association studies (GWAS). For this, several linear mixed models were tested using the lme4 R package (Bates et al., 2015). The full model included genotype (cultivar) as a random effect and row and planting year as fixed effects. Both fixed effects were included in the final model based on the lowest Bayesian information criterion (BIC) and the random effects (Best Linear Unbiased Predictions, BLUPs) were extracted. This procedure was also carried out for the observed stage C timings of all trees in 2022 and 2023. To test the significance of the difference in stage C timing between 2022 and 2023 due to the year effect, the above mixed model, with the addition of a fixed term for year and a random interaction term for year and genotype, was fitted using data from both years, followed by an ANOVA.

The selected model was also used to calculate the broad-sense heritability (H^2) of the observations and predictions of stage C in 2023 with the variance components, adjusted for the number of replicates, as follows:

$$H^2 = \frac{VarG}{VarG + \frac{VarR}{nrep}}$$

where *VarG* is the genotypic variance, *VarR* is the residual variance and *nrep* is the number of replicates per genotype.

2.7. Single nucleotide polymorphism (SNP) markers

The SNP markers used in the GWAS analyses were an amalgamation of two datasets, including those of, firstly, an Axiom® Apple 480 K array (Bianco et al., 2016; Denancé et al., 2022) and, secondly, a dataset derived from capture sequencing of a selection of genes related to flowering time control in *Arabidopsis thaliana* (Bouché et al., 2016) as well as apple genes located within a quantitative trait locus (QTL) on chromosome 9, previously associated to budbreak (Trainin et al., 2016). A full description of the capture sequencing pipeline is provided in Watson et al. (2024) and

Table 1. Summary of the training and testing data used in partial least squares regression (PLSR) models in the selection and testing phases of stage C timing prediction

		Training data		Testing data		
Tissue	Model name	Predictor (NIRS)	Response	Predictor (NIRS)	Response	No. of samples
Leaf	June model	June 2021	Phenological stage C, 2022	June 2022	Phenological stage C, 2023	929
	Sept model	Sept 2021		Sept 2022		942
	Nov model	Nov 2021		Nov 2022		897
Bud	Dec model	Dec 2021		-	-	953
	Jan model	Jan 2022		-		946

Note: The trees sampled in the training and testing data within each model were the same each year, and the number of which is given as the number of samples.

data at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1023873. Following filtering for bi-allelic SNPs successfully mapped to the genome with a minor allele frequency (MAF) of ≥0.05 and <95% heterozygosity, the final dataset contained 290,150 SNPs (40,857 from capture sequencing and 249,293 from the array).

2.8. Genome-wide association studies (GWAS)

GWAS analyses were performed using the BLUPs of stage C 2023 predictions of 238 cultivars of the apple core collection. A single-locus, mixed model approach was chosen, performed by GEMMA software (v.0.97; Zhou & Stephens, 2012), where the following model was fitted for each SNP:

$$Y = m + X\beta + g + e,$$

where Y was a vector of genotypic values, m was the overall mean, X was a vector of SNP dosage scores (0, 1 or 2 to indicate the number of the alternate allele), β was a vector of additive effects, g was a vector of random polygenic effects and e was a vector of random residual effects. The underlying distributions of the random effects were assumed as $g \sim N\left(0, \boldsymbol{G}\sigma_g^2\right)$ and $e \sim N\left(0, \mathbf{I}\sigma_e^2\right)$, where $\boldsymbol{G}$ was the realised genomic matrix, calculated using Method 1 described in VanRaden (2008) with the rutilstimflutre R package (Flutre, 2019), $\mathbf{I}$ was an identity matrix, σ_g^2 was the genetic variance and σ_e^2 was the residual variance.

To assign statistical significance to SNP-prediction associations, the effective number of independent tests (Meff; Cheverud, 2001) was estimated using the simpleM method (Gao et al., 2008, 2010). An estimation of 85,159 independent tests resulted in a Bonferroni threshold of $-\log_{10}(p\text{-value}) = 6.23$. To test the significance (p-value ≤ 0.05) of differences in stage predictions of cultivars with different genotypes of significant SNPs, an ANOVA, followed by the Tukey method to account for multiple comparisons, was performed using the emmeans R package (Lenth, 2022).

3. Results

3.1. Flowering stages were highly correlated within and between years

All flowering stages were highly correlated in 2022 and 2023 and across both years, with correlation coefficients above 0.89 (Figure 1, left). The correlations between the two years within each flowering stage were also relatively high, with coefficients of 0.75, 0.87, 0.87 and 0.86 for stages C, E, F and G, respectively (Figure 1, right). However, for each flowering stage, there was a significant year effect (p-value = $2.2\text{x}10^{-16}$), clearly visualised in the distributions, with all stages appearing to be reached earlier in 2022 than 2023. For example, in both years, stage C commenced at the beginning of March; however, by mid-March 2022 (75 days after Jan 1st), Stage C had been reached by 27% of trees, while only 2% had reached this stage by the same time in 2023. Due to the flowering stages being so highly correlated and stage C being recorded for all trees in both years, only this stage was used in further analyses. Note that Stage C was the budbreak of both floral and vegetative buds.

3.2. Model performance was unaffected by spectra pre-processing

Model selection identified the optimal number of latent variables as well as the best pre-processing of the spectra matrices for all five models by minimising their RMSE (see Supplementary Figure S4 for detailed results). It was noted that the performance measures, R^2 and RMSE, of these selected models were not significantly affected by the type of pre-processing of the spectra used as the predictor (Supplementary Figure S5). Therefore, for parsimony, only the raw spectra were considered in the subsequent model training phase. Furthermore, the performances of the Dec and Jan models, both of which used NIRS spectra from bud tissue, were much lower than those of the three other models, all of which used NIRS spectra

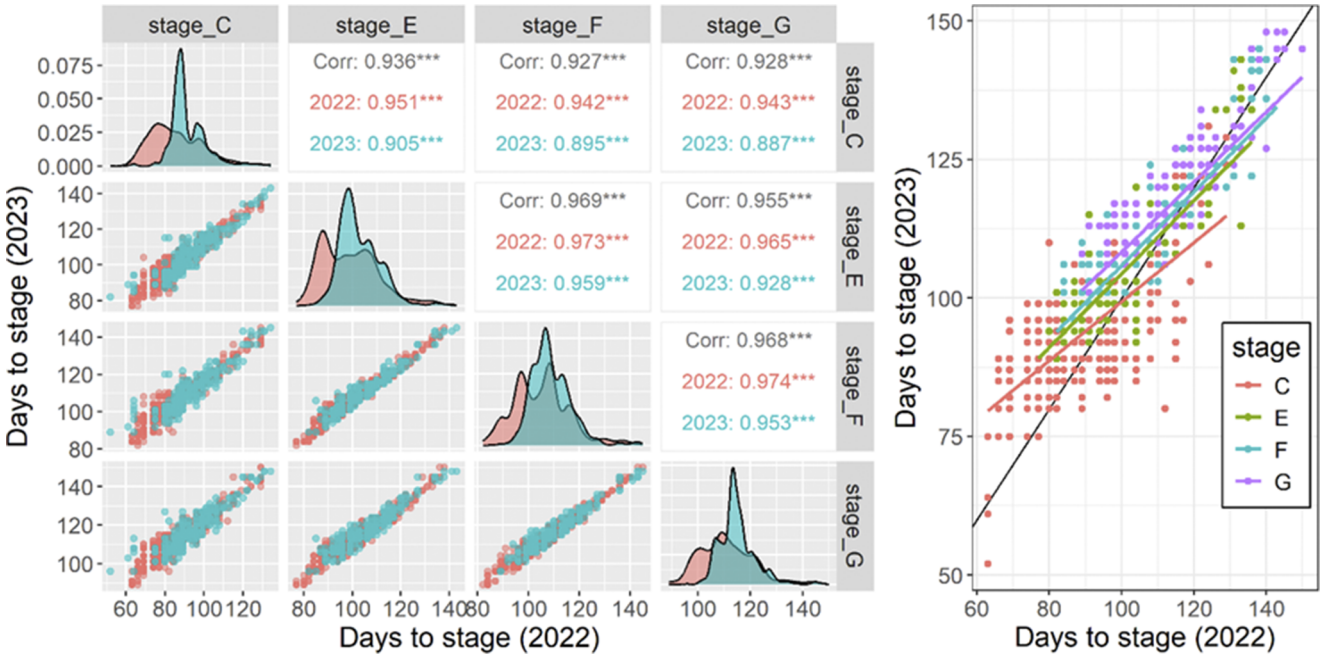

Figure 1. Left: Correlations between flowering stages in 2022 (pink), 2023 (blue) and both years together (black). *** Indicates a p-value < 0.001. Left diagonal: Density distributions of each stage in 2022 and 2023. Right: Timing of all flowering stages in 2022 versus the same stage in 2023. Values are days from January 1st of that year until the flowering stage occurred. Correlations were calculated using the phenotypes of all trees.

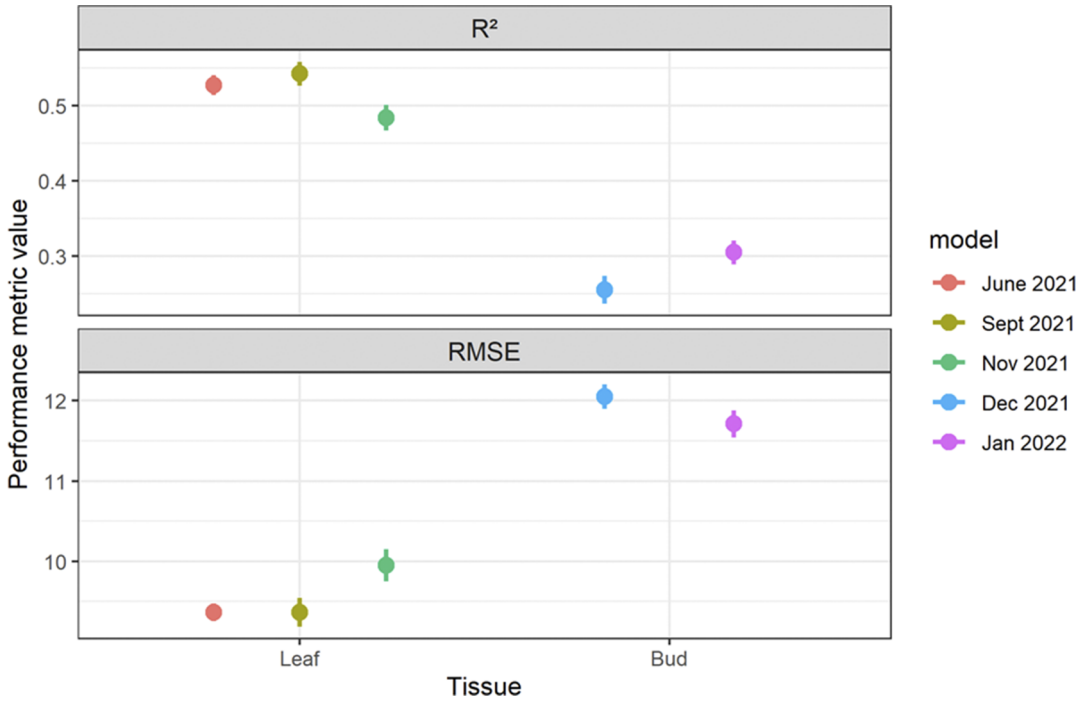

Figure 2. Performance (R^2 and RMSE) of all PLSR models from the five NIRS collection dates (three leaf and two bud collections), using the raw spectra matrices, during the model selection phase, to predict stages C in 2023. Error bars represent the 95% confidence interval.

from leaf tissue (Figure 2). Interestingly, the R^2 of the Jan model appeared slightly higher than that of the Dec model; however, due to the comparatively low performance, these two models were not advanced beyond model selection. The number of latent variables included in the optimal models selected for the June, September and November leaf collection dates was 24, 23 and 19, respectively.

3.3. NIRS collection date influenced PLSR model performance

In the model testing phase, predictions of stage C in 2023 were made using the selected trained models and the testing NIRS data of the corresponding collection date in 2022. Figure 3 shows the predicted versus observed values of stage C in 2023. In terms of R^2, the performance of the June model was much higher than that of the other two models, although the RMSE measures were all relatively high and within five days of each other.

In comparison to the spectra-based PLSR models, a trivial model was used as a prediction ability benchmark. It simply repeated the timing of stage C in 2022 as the prediction for 2023. The R^2 of this model was much higher than that of the PLSR models (Supplementary Figure S6), indicating that stage C itself was a better predictor of the same stage the next year than PLSR models using NIRS as the predictor. However, the RMSE of the trivial model was still relatively high at 11 days.

3.4. GWAS using predictions detected a known QTL associated with budbreak

To examine the extent of the genotypic variance lost through the prediction process, the H^2 and genotypic BLUPs of the prediction 2023 stage C and the observed stage C of 2023 were compared. The H^2 of stage C in 2023 was high at 0.95, much higher than those of the stage C predictions of the June, Sept and Nov PLSR models, which were 0.74, 0.66 and 0.65, respectively. Furthermore, the correlation of the genotypic BLUPs of the predictions from each model and the genotypic BLUPs of the observed stage C in 2023 also varied, with the June model having the highest correlation at 0.64 while the Sept and Nov models had similar correlations of 0.43 and 0.40, respectively (Supplementary Figure S7).

The genotypic BLUPs of the stage C predictions were also used in a GWAS analysis to further test their representation of the genetic component of this trait. A well-known QTL on chromosome 9 was detected with the predictions of the June model, although not with the predictions of the Sept or Dec models (Figure 4). Details of the significant SNPs are given in Supplementary Table S2. The most significant SNP, AX.115409139, was located slightly upstream of a peroxidase superfamily gene (MD09G1010000) and the cultivars homozygous for the A allele reached state C significantly later than those homozygous for the C allele (p-value = 0.0001; Supplementary Figure S8).

4. Discussion

4.1. NIRS prediction ability may depend on the metabolic legacy of budbreak

In an effort to explore the potential of NIRS to predict future developmental stages, we have demonstrated the ability of PLSR to predict budbreak timing in apple using NIRS spectra measured on leaf tissue. The biochemical snapshot provided by NIRS is related to the internal processes that underlie physiological progression towards a developmental stage, which are shaped by both environmental and genetic factors. Our models rely on this theoretical link. The NIRS collection dates spanned the period of full leaf expansion, just following budbreak, to leaf senescence, a process that has been found to be strongly influenced by the timing of budbreak. Fu et al. (2014) reported earlier leaf budbreak, induced by warmer temperatures, in *Quercus robur* L. and *Fagus sylvatica* L.,

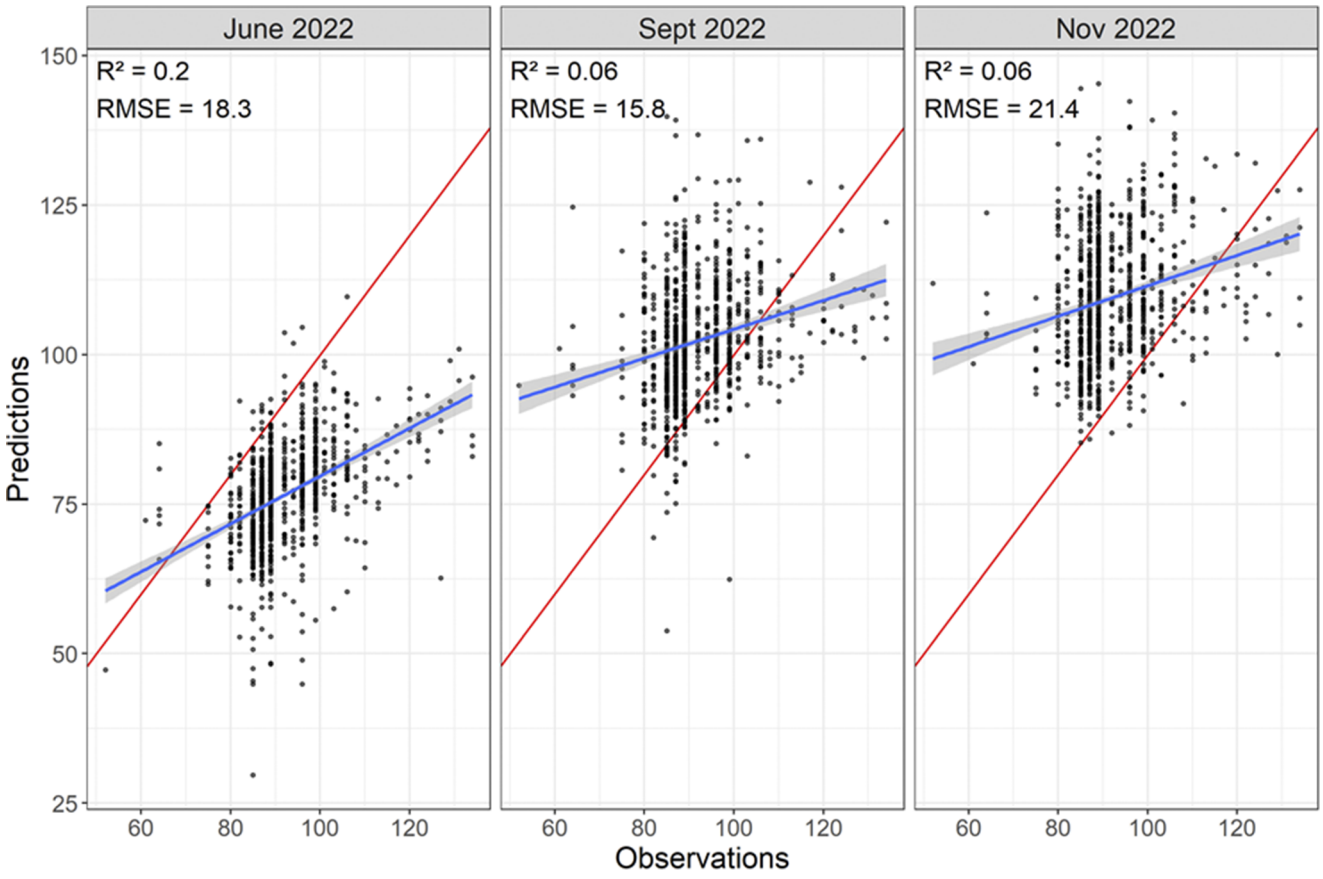

Figure 3. Timing of stage C observed in 2023 and timing of stage C predicted by the June, Sept and Nov PLSR models. Performance parameters, R^2 and RMSE (days), are indicated. Values are days from Jan 1st until the stage occurred. A linear line was fitted to each plot and grey shading indicates the standard error of the mean. The identity line (in red) marks perfect predictions.

resulting in earlier leaf senescence in the same year and budbreak in the following year. This was hypothesised to be most likely due to earlier commencement, and therefore satisfaction, of the chilling requirement to overcome endodormancy. In the current study, the relatively good prediction ability of the trivial model clearly indicated a strong relationship between the budbreak of the two years. This, in part, can be seen through the lens of the metabolic legacy of the first year budbreak influencing the duration of senescence and therefore budbreak of the next year. The performance of the June model was better at explaining the variation (i.e. R^2) in the data than the models derived from the later collection dates in September and November. This may have been due to the relationship between budbreak of the two years, with the June NIRS better capturing a metabolic state closely connected to the recent budbreak stage before senescence began. This may also explain the poor performance of PLSR models using bud tissue NIRS. While leaves undergo senescence, the buds, which are protective structures for dormancy, were sampled much later, once senescence was largely complete. Moreover, the measurement of NIRS on dried and ground tissue from complete plant organs may have affected the detectability of the relevant metabolites. Dhuli et al. (2014) demonstrated the extensive range of metabolites produced in Norway spruce (*Picea abies*) and European silver fir (*Abies alba*) in response to de-acclimation (temperature increase to simulate spring conditions), leading to budbreak. Not only were groups of metabolites identified as being specific to the process, but high correlations were found between many of these compounds and the timing of budbreak. The mixed and processed nature of our samples may have increased non-target signals in the spectra, leading to a poorer representation of the metabolic state.

4.2. Environmental factors were influential on budbreak prediction

Despite the better performance of the June model, the RMSE of all three models were relatively high and fell within five days of each other. One of the main challenges of predicting budbreak is, without doubt, the contribution of environmental factors, particularly temperature (Cannell and Smith, 1984). This impact can be seen in the relatively large RMSE of the trivial model, which largely represents the differing environmental conditions experienced between the two episodes of budbreak. Here, NIRS was measured the year before budbreak occurred, allowing an extended period during which the sequence of physiological events leading to budbreak was influenced by seasonal conditions. In our case, the presence of a significant year effect on the timing of stage C, detected between 2022 and 2023, was potentially linked, at least in part, to the differences in temperature, particularly in spring, where the warmer temperatures in 2022 (Supplementary Figure S3) may have accelerated fulfilment of the heating requirement needed to overcome ecodormancy and begin budbreak. In practice, climate models or at least temperature variables could be included in PLSR models to improve predictions. In the current study, however, the low number of years and a single location used prevented the

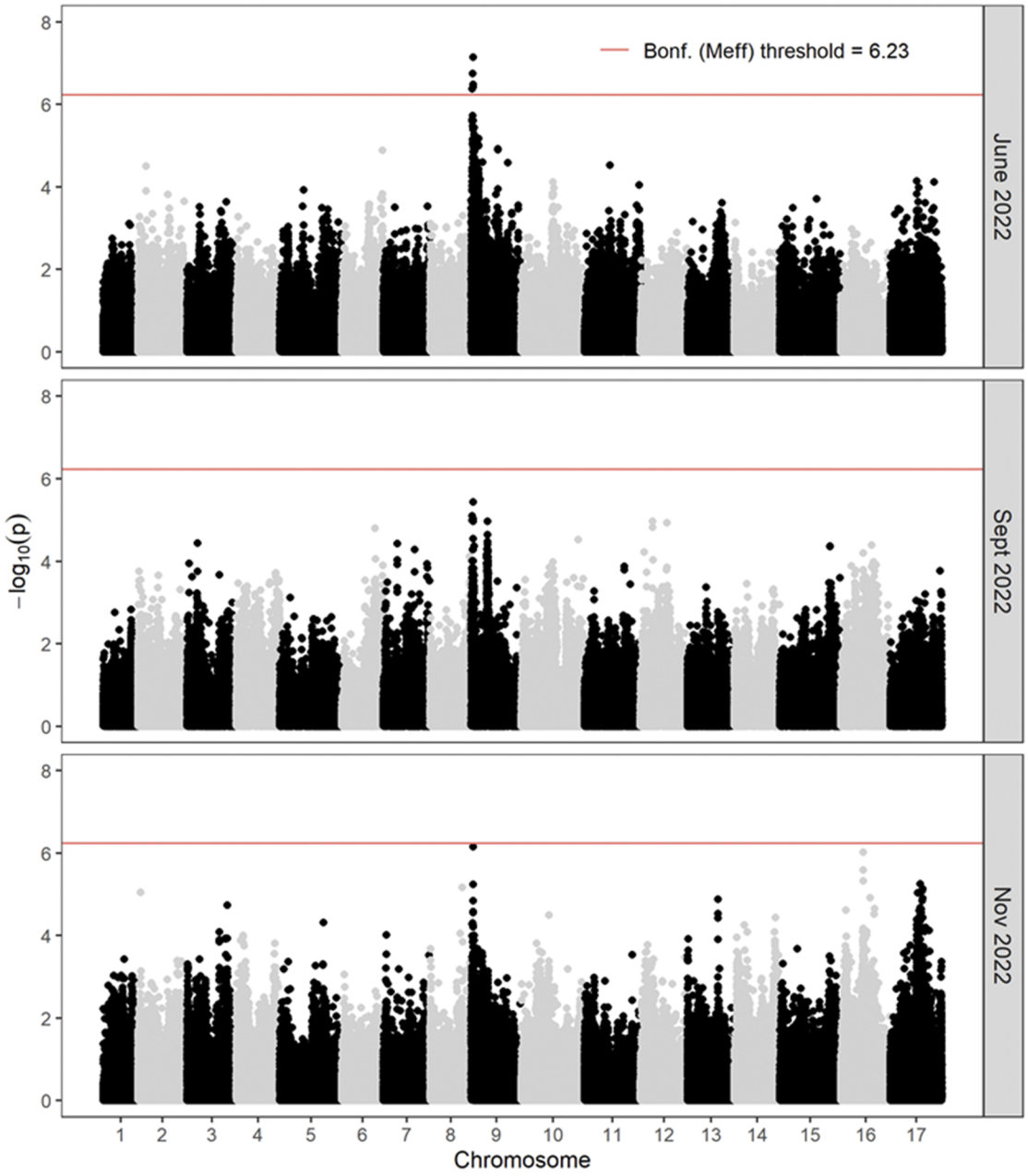

Figure 4. Manhattan plots from the GWAS analyses using the predictions of stage C 2023 from the June, Sept and Nov PLSR models. The Bonferroni (Bonf.) threshold was calculated using the effective number of independent tests (Meff).

addition of these factors to the models. Yet, their inclusion is recommended for future application, especially when the target trait is strongly affected by environmental factors, as with budbreak. In instances where the trait is less affected by this, the current approach may be sufficient to produce predictions good enough for practical applications.

4.3. NIRS-PLSR prediction offers avenues for genetic studies

Budbreak timing has a large genetic component, as evidenced by the high H^2 of the observed stage C in 2023, and although inclusion of environmental factors would likely have improved the prediction performance of the PLSR models, it appears the NIRS spectra were able to capture a significant genetic component of budbreak timing, likely through a genotype-specific chemical fingerprint (Munck et al., 2021). The H^2 of the PLSR predictions were all relatively high, suggesting a considerable portion of the genetic variation was maintained through the prediction process. This was further supported by the correlations of the genotypic BLUPs, which represent the proportion of the phenotype attributable to the genotype. The June PLSR model prediction BLUPs had the highest correlation to the observed stage C BLUPs, indicating more of the genetic variation inherent in this trait was kept by this model than by the others. In addition, the GWAS with the June model prediction BLUPs was able to detect a well-known QTL on chromosome 9, which was previously linked to budbreak in apple (Allard et al., 2016; Celton et al., 2011; Conner et al., 1998; Cornelissen et al., 2020; Miotto et al., 2019; Trainin et al., 2016; Urrestarazu et al., 2017; van Dyk et al., 2010; Watson et al., 2024). Recently, this QTL was explored more in depth with a GWAS using

the same population and genotypic dataset as the current study and stage C recorded over nine years (Watson et al., 2024). This identified *MdPEROXIDASE10* (*MdPRX10*; MD09G1010000), a peroxidase superfamily gene, as a strong candidate for involvement in budbreak, potentially via redox-mediated signalling. The most significant SNP in the GWAS analysis using the June model BLUP predictions was located very closely upstream of the same gene. This suggests that these predictions contained sufficient genetic information to be associated with the same genomic region, even gene, as the observed phenotypes.

5. Conclusion

Our results demonstrate that leaf NIRS spectra hold potential to predict the timing of future developmental stages, such as budbreak, by detecting the metabolic states that precede them. We envisage that this approach may reduce the phenotyping burden for traits that are difficult to measure. Once a PLSR model has been trained, predictions for the trait in other populations could be obtained using only NIRS, potentially with enough precision that these could then be used in genetic studies to identify new genomic regions of interest related to the trait. Using diverse populations, as we did here, may also produce models with better generalisation properties that are applicable to prediction for a wider range of genotypes. This could further be simplified with the use of handheld NIRS devices on intact plant organs, which would negate the need for sample processing and may even better communicate the complex nature of the tissue biochemical state.

Open peer review. To view the open peer review materials for this article, please visit http://doi.org/10.1017/qpb.2025.10019.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/qpb.2025.10019.

Data availability statement. The R code used for the PLSR analyses is available from gitlab.com/watsonamy1/nirs-plsr. The phenological and NIRS data can also be found at the GitLab link. The phenological data was a subset sourced from Farrera et al. (2024) and the NIRS data was a subset from Perez et al. (2025). The sequence data used for the GWAS analyses are available from the following online repositories: the capture sequencing data from the NCBI BioProject database at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1023873 and the Axiom® Apple 480 K array data from Denancé et al. 2022.

Acknowledgements

We are very grateful to the INRAE AFEF team for their contribution to in-field sampling for NIRS. This work was supported by the Biochemical Phenotyping Platform of CIRAD-AGAP Units.

Author contributions. AEW contributed to data collection, performed the analyses and wrote the manuscript. VS and YB contributed to the analyses. GP and IF performed data collection and curation. FA, VS and EC conceived the project and contributed to data collection. All authors reviewed and edited the manuscript.

Funding statement. This project received funding from ERA-NET SusCrop2 (FruitFlow, ANR-21-SUSC-0002) to support the postdoctoral post of AEW.

Competing interest. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Al-Amery, M., **Geneve, R. L.**, **Sanches, M. F.**, **Armstrong, P. R.**, **Maghirang, E. B.**, **Lee, C.**, **Vieira, R. D.**, & **Hildebrand, D. F.** (2018). Near-infrared spectroscopy used to predict soybean seed germination and vigour. *Seed Science Research*, **28**(3), 245–252. https://doi.org/10.1017/s0960258518000119.

Allard, A., **Bink, M. C. A. M.**, **Martinez, S.**, **Kelner, J. J.**, **Legave, J. M.**, **Di Guardo, M.**, et al. (2016). Detecting QTLs and putative candidate genes involved in budbreak and flowering time in an apple multi-parental population. *Journal of Experimental Botany*, **67**, 2875–2888. https://doi.org/10.1093/jxb/erw130.

Aptula, A. O., **Jeliazkova, N. G.**, **Schultz, T. W.**, & **Cronin, M. T. D.** (2005). The better predictive model: High q2 for the training set or low root mean square error of prediction for the test set? *QSAR & Combinatorial Science*, **24**(3), 385–396. https://doi.org/10.1002/qsar.200430909.

Baggiolini, M. (1980). Stades repères de la abricotier-stades repères de la pêcher. In *Stades repères du ceresier-stades repères du prunier. Guide pratique de défense des cultures*. Acta editions).

Barnes, R. J., **Dhanoa, M. S.**, & **Lister, S. J.** (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, **43**(5), 772–777. https://doi.org/10.1366/0003702894202201.

Bates, D., **Mächler, M.**, **Bolker, B.**, & **Walker, S.** (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1–48. https://doi.org/10.18637/jss.v067.i01.

Beauvieux, R., **Wenden, B.**, & **Dirlewanger, E.** (2018). Bud dormancy in perennial fruit tree species: A pivotal role for oxidative cues. *Frontiers in Plant Science*, **9**. https://doi.org/10.3389/fpls.2018.00657.

Bianco, L., **Cestaro, A.**, **Linsmith, G.**, **Muranty, H.**, **Denancé, C.**, **Théron, A.**, et al. (2016). Development and validation of the axiom® Apple480K SNP genotyping array. *The Plant Journal*, **86**, 62–74. https://doi.org/10.1111/tpj.13145.

Bouché, F., **Lobet, G.**, **Tocquin, P.**, & **Périlleux, C.** (2016). FLOR-ID: An interactive database of flowering-time gene networks in *Arabidopsis thaliana*. *Nucleic Acids Research*, **44**, 1167–1171. https://doi.org/10.1093/nar/gkv1054.

Brault, C., **Lazerges, J.**, **Doligez, A.**, **Thomas, M.**, **Ecarnot, M.**, **Roumet, P.**, **Bertrand, Y.**, **Berger, G.**, **Pons, T.**, **François, P.**, **Le Cunff, L.**, **This, P.**, & **Segura, V.** (2022). Interest of phenomic prediction as an alternative to genomic prediction in grapevine. *Plant Methods*, **18**(1). https://doi.org/10.1186/s13007-022-00940-9.

Cannell, M. G. R., & **Smith, R. I.** (1984). Spring frost damage on young Picea sitchensis 2. Predicted dates of budburst and probability of frost damage, Forestry: *An Int. J. Forest Res.*, **57**, 177–191. https://doi.org/10.1093/forestry/57.2.177.

Cannell, M. G. R., & **Smith, R. I.** (1986). Climatic warming, spring budburst and frost damage on trees. *Journal of Applied Ecology*, **23**, 177–191. https://doi.org/10.2307/2403090.

Celton, J. M., **Martinez, S.**, **Jammes, M. J.**, **Bechti, A.**, **Salvi, S.**, **Legave, J. M.**, et al. (2011). Deciphering the genetic determinism of bud phenology in apple progenies: A new insight into chilling and heat requirement effects on flowering dates and positional candidate genes. *The New Phytologist*, **192**, 378–392. https://doi.org/10.1111/nph.2011.192.issue-2.

Cheverud, J. M. (2001). A simple correction for multiple comparisons in interval mapping genome scans. *Heredity*, **87**, 52–58. https://doi.org/10.1046/j.1365-2540.2001.00901.x.

Conner, P. J., **Brown, S. K.**, & **Weeden, N. F.** (1998). Molecular-marker analysis of quantitative traits for growth and development in juvenile apple trees. *Theoretical and Applied Genetics*, **96**, 1027–1035. https://doi.org/10.1007/s001220050835.

Cornelissen, S., **Hefer, C. A.**, **Rees, D. J. G.**, & **Burger, J. T.** (2020). Defining the QTL associated with chill requirement during endodormancy in malus × domestica Borkh. *Euphytica*, **216**, 122. https://doi.org/10.1007/s10681-020-02645-3.

Denancé, C., **Muranty, H.**, & **Durel, C.-E.** (2022). FruitBreedomics apple 275K SNP genotypic data. *Recherche Data Gouv*, **V1**. https://doi.org/10.15454/F5XIVJ.

Dhuli, P., **Rohloff, J.**, & **Strimbeck, G. R.** (2014). Metabolite changes in conifer buds and needles during forced bud break in Norway spruce (Picea abies)

and European silver fir (Abies alba). *Frontiers in Plant Science*, **5**. https://doi.org/10.3389/fpls.2014.00706.

Ecarnot, M. (2023). *Miscellanous functions for NIRS and chemometrics*. https://github.com/martinEcarnot/nirsextra.

Farrera, I., **Perez, G.**, **Costes, E.**, & **Andrés, F.** (2024). Budbreak dataset from 239 cultivars of an apple core collection 2015–2023. *Recherche Data Gouv*, **V1**. https://doi.org/10.57745/WP5VGZ.

Flutre, T. (2019). *Timothee Flutre's personal R code*. https://github.com/timflutre/rutilstimflutre.

Fu, Y. S. H., **Campioli, M.**, **Vitasse, Y.**, **De Boeck, H. J.**, **Van Den Berge, J.**, **AbdElgawad, H.**, **Asard, H.**, **Piao, S.**, **Deckmyn, G.**, & **Janssens, I. A.** (2014). Variation in leaf flushing date influences autumnal senescence and next year's flushing date in two temperate tree species. *Proceedings of the National Academy of Sciences*, **111**(20), 7355–7360. https://doi.org/10.1073/pnas.1321727111.

Gao, X., **Becker, L. C.**, **Becker, D. M.**, **Starmer, J. D.**, & **Province, M. A.** (2010). Avoiding the high bonferroni penalty in genome-wide association studies. *Genetic Epidemiology*, **34**, 100–105. https://doi.org/10.1002/gepi.20430.

Gao, X., **Starmer, J.**, & **Martin, E. R.** (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology*, **32**, 361–369. https://doi.org/10.1002/gepi.20310.

Hanke, M., **Flachowsky, H.**, **Peil, A.**, and **Hättasch, C.** (2007). No flower no fruit - genetic potentials to trigger flowering in fruit trees. In *Genes, genomes and genomics*, **1**(1), 1–20. Global Science Books, Singapore. ISSN: 1749-0383

Lane, H. M., **Murray, S. C.**, **Montesinos-López, O. A.**, **Montesinos-López, A.**, **Crossa, J.**, **Rooney, D. K.**, **Barrero-Farfan, I. D.**, **De La Fuente, G. N.**, & **Morgan, C. L. S.** (2020). Phenomic selection and prediction of maize grain yield from near-infrared reflectance spectroscopy of kernels. *The Plant Phenome Journal*, **3**(1). https://doi.org/10.1002/ppj2.20002.

Lang, G. A., **Early, J. D.**, **Martin, G. C.**, & **Darnell, R. L.** (1987). Endodormancy, paradormancy, and ecodormancy: Physiological terminology and classification for dormancy research. *Horticultural Science*, **22**, 371–377. https://doi.org/10.21273/HORTSCI.22.5.701b.

Lassois, L., **Denancé, C.**, **Ravon, E.**, **Guyader, A.**, **Guisnel, R.**, **Hibrand-Saint-Oyant, L.**, et al. (2016). Genetic diversity, population structure, parentage analysis, and construction of core collections in the French apple germplasm based on SSR markers. *Plant Molecular Biology Reporter*, **34**, 827–844. https://doi.org/10.1007/s11105-015-0966-7.

Legave, J. M., **Blanke, M.**, **Christen, D.**, **Giovannini, D.**, **Mathieu, V.**, & **Oger, R.** (2013). A comprehensive overview of the spatial and temporal variability of apple bud dormancy release and blooming phenology in Western Europe. *International Journal of Biometeorology*, **57**(2), 317–331. https://doi.org/10.1007/s00484-012-0551-9.

Legave, J. M., **Farrera, I.**, **Almeras, T.**, & **Calleja, M.** (2008). Selecting models of apple flowering time and understanding how global warming has had an impact on this trait. *The Journal of Horticultural Science and Biotechnology*, **83**, 76–84. https://doi.org/10.1080/14620316.2008.11512350.

Lenth, R. (2022). *Emmeans: Estimated marginal means, aka least-squares means. R package version 1.8.3*. https://CRAN.R-project.org/package=emmeans.

Liland, K., **Mevik, B.**, **Wehrens, R.** (2023). *Pls: Partial least squares and principal component regression. R package version 2.8-3*. https://CRAN.R-project.org/package=pls.

Minas, I. S., **Blanco-Cipollone, F.**, & **Sterle, D.** (2021). Accurate non-destructive prediction of peach fruit internal quality and physiological maturity with a single scan using near infrared spectroscopy. *Food Chemistry*, **335**, 127626. https://doi.org/10.1016/j.foodchem.2020.127626.

Miotto, Y. E., **Tessele, C.**, **Czermainski, A. B. C.**, **Porto, D. D.**, **Falavigna, V.**, **da, S.**, et al. (2019). Spring is coming: Genetic analyses of the bud break date locus reveal candidate genes from the cold perception pathway to dormancy release in apple (malus × Domestica borkh.). *Frontiers in Plant Science*, **10**. https://doi.org/10.3389/fpls.2019.00033.

Munck, L., **Rinnan, Å.**, **Khakimov, B.**, **Jespersen, B. M.**, & **Engelsen, S. B.** (2021). Physiological genetics reformed: Bridging the genome-to-phenome gap by coherent chemical fingerprints – The global coordinator. *Trends in Plant Science*, **26**(4), 324–337. https://doi.org/10.1016/j.tplants.2020.12.014.

Perez, G., **Segura, V.**, **Watson, A.**, **Farrera, I.**, & **Andrés, F.** (2025). Dataset on near-infrared spectroscopy measurements of apple tree leaves_2021-2022. *Recherche Data Gouv*, **V1**. https://doi.org/10.57745/KKXXAU.

R Core Team. (2024). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.Rproject.org/.

Rincent, R., **Charpentier, J.**, **Faivre-Rampant, P.**, **Paux, E.**, **Gouis, J. L.**, **Bastien, C.**, & **Segura, V.** (2018). Phenomic selection is a low-cost and high-throughput method based on indirect predictions: Proof of concept on wheat and poplar. *G3 Genes Genomes Genetics*, **8**(12), 3961–3972. https://doi.org/10.1534/g3.118.200760.

Robert, P., **Auzanneau, J.**, **Goudemand, E.**, **Oury, F.**, **Rolland, B.**, **Heumez, E.**, **Bouchet, S.**, **Gouis, J. L.**, & **Rincent, R.** (2022). Phenomic selection in wheat breeding: Identification and optimisation of factors influencing prediction accuracy and comparison to genomic selection. *Theoretical and Applied Genetics*, **135**(3), 895–914. https://doi.org/10.1007/s00122-021-04005-8.

Rodríguez, P., **Villamizar, J.**, **Londoño, L.**, **Tran, T.**, & **Davrieux, F.** (2023). Quantification of dry matter content in hass avocado by near-infrared spectroscopy (NIRS) scanning different fruit zones. *Plants*, **12**(17), 3135. https://doi.org/10.3390/plants12173135.

Rosales, A., **Galicia, L.**, **Oviedo, E.**, **Islas, C.**, & **Palacios-Rojas, N.** (2011). Near-infrared reflectance spectroscopy (NIRS) for protein, tryptophan, and lysine evaluation in quality protein maize (QPM) breeding programs. *Journal of Agricultural and Food Chemistry*, **59**(20), 10781–10786. https://doi.org/10.1021/jf201468x.

Signal Developers. (2023). *Signal: Signal processing*. https://r-forge.r-project.org/projects/signal/.

Stevens, A., & **Ramirez-Lopez, L.** (2024). *An introduction to the prospectr package. R package vignette R package version 0.2.7*.

Trainin, T., **Zohar, M.**, **Shimoni-Shor, E.**, **Doron-Faigenboim, A.**, **Bar-Ya'akov, I.**, **Hatib, K.**, et al. (2016). A unique haplotype found in apple accessions exhibiting early bud-break could serve as a marker for breeding apples with low chilling requirements. *Molecular Breeding*, **36**, 158. https://doi.org/10.1007/s11032-016-0575-7.

Urrestarazu, J., **Muranty, H.**, **Denancé, Leforestier, D.**, **Ravon, E.**, **Guyader, A.**, et al. (2017). Genome-wide association mapping of flowering and ripening periods in apple. *Frontiers in Plant Science*, **8**. https://doi.org/10.3389/fpls.2017.01923.

van Dyk, M. M., **Soeker, M. K.**, **Labuschagne, I. F.**, & **Rees, D. J. G.** (2010). Identification of a major QTL for time of initial vegetative budbreak in apple (malus x domestica Borkh.). *Tree Genetics & Genomes*, **6**, 489–502. https://doi.org/10.1007/s11295-009-0266-1.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, **91**, 4414–4423. https://doi.org/10.3168/jds.2007-0980.

Wang, Z., **Zuo, C.**, **Wang, M.**, **Song, S.**, **Hu, Y.**, **Song, J.**, **Tu, K.**, **He, H.**, **Lan, W.**, & **Pan, L.** (2024). Optical properties related to cell wall pectin contribute to determine the firmness and microstructural changes during apple softening. *Postharvest Biology and Technology*, **218**, 113150. https://doi.org/10.1016/j.postharvbio.2024.113150.

Watson, A. E., **Guitton, B.**, **Soriano, A.**, **Rivallan, R.**, **Vignes, H.**, **Farrera, I.**, **Huettel, B.**, **Arnaiz, C.**, **Falavigna, d. S.**, **Coupel-Ledru, A.**, **Segura, V.**, **Sarah, G.**, **Dufayard, J.-F.**, **Sidibe-Bocs, S.**, **Costes, E.**, & **Andrés, F.** (2024). Target enrichment sequencing coupled with GWAS identifies MdPRX10 as a candidate gene in the control of budbreak in apple. *Frontiers in Plant Science*, **15**, 1352757. https://doi.org/10.3389/fpls.2024.1352757.

Wold, S., **Sjöström, M.**, & **Eriksson, L.** (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, **58**(2), 109–130. https://doi.org/10.1016/s0169-7439(01)00155-1.

Zhou, X., & **Stephens, M.** (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, **44**, 821–824. https://doi.org/10.1038/ng.2310.