Exploring metabolome changes in wheat (*Triticum aestivum*) under heat stress using Fourier transform infrared spectroscopy

(フーリエ変換赤外分光法を用いた高温ストレス下におけるコムギのメタボローム変動の探索)

Osman, Salma Osman Mohamedkhair

2022

Exploring metabolome changes in wheat (*Triticum aestivum*) under heat stress using Fourier transform infrared spectroscopy

(フーリエ変換赤外分光法を用いた高温ストレス下におけるコムギのメタボローム変動の探索)

A thesis submitted to the United Graduate School of Agricultural Sciences, Tottori University in partial fulfilment of requirements for the award of Doctor of Philosophy (PhD) in Global Dryland Science.

By

Osman, Salma Osman Mohamedkhair

Approved by:

Prof. **Dr Motoichiro Kodama**……

Dean, United Graduate School of Agricultural Sciences, Tottori University

Prof. **Dr Kinya Akashi**……

Chairman, supervisory committee

The United Graduate School of Agricultural Sciences, Tottori University

2022

Dedication

To everyone ambitious to leave positive impact in this life before departure.

Acknowledgements

I would like to acknowledge JICA-JST-SATREPS for offering this opportunity to study in Japan and develop skills which will flourish my future.

I really appreciate help, support and continuous encourage from my supervisor Kinya AKASHI, I really appreciate his efforts to teach me even simple things.

I am grateful for co-supervisors Dr. Hisashi Tsujimoto and Dr. Tsuyoshi Nakagawa for their guidance through scientific discussions.

I am grateful for Dr. Tanaka and Dr. Gorafi for providing wheat seeds I used in my experiments.

I am grateful for colleagues and technician in Molecular and Cellular Biology Laboratory and Venture Business Laboratory at Tottori University.

I am thankful to colleagues at Agricultural Research Corporation-Sudan, Dr. Izzat S. A. Tahir and Abu Sefyan I. Saad for their guidance and manuscripts checking.

I really appreciate Dr. Yuji Yamasaki for his efforts in checking and improving my manuscripts significantly.

I am thankful for my colleagues Abuelgasim and Almutaz for their help in R software.

I would like to convey special thanks to Ms. Alyza_Megumi for her help and support and making my life in Japan easier.

I am also thankful for all Sudanese in Tottori who make me feal I am among my sisters and brothers.

Table of contents

Contents

General Introduction

Wheat represents one of the most important crops with considerable contribution to nutrients required in human diets (Curtis et al. 2002). Increase in wheat global demand is expected as a consequence of population growth which expected to reach 9 billion by 2050 (Figueroa et al. 2018). Therefore, improving wheat production is essential but hindered by climate change and biotic stresses obstacles. Wheat yield reduction under high temperature was reported in previous literature (Mitchell et al. 1993; Stone and Nicolas, 1995; Semenov and Halford, 2009; Schittenhelm et al. 2020; Matsunaga et al. 2021a).

Understanding wheat's metabolomic response to heat stress will facilitate developing new heat tolerant varieties. Various tools for investigating metabolome are available such as liquid chromatography-mass spectrometry (LC-MS) and gas chromatography-mass spectrometry (GC-MS). Fourier transform infrared (FTIR) excels other metabolomics techniques by the capability to detect macrobiomolecules (Liu et al. 2021; Kljun et al. 2011; McCann et al. 1992).

In the light of the above-mentioned information, the objective of this thesis was to understand metabolomic response of common wheat to heat stress through:

- Establishment of FTIR spectroscopic protocol that fingerprints wheat metabolomic response under heat stress.

- Applying the established FTIR protocol to wheat genotypes with different heat tolerance level to identify potential metabolomic heat markers.

The first objective, which was detailed in Chapter 1 of this thesis, was achieved and published in International Journal of Molecular Science (Osman et al. 2022a). The second objective (Chapter 2) was also achieved and published in Agriculture (Osman et al. 2022b).

CHAPTER 1

Characterization of Heat Stress Responses in the Leaves of Common Wheat by Fourier Transform Infrared Spectroscopy

1.1. Introduction

Wheat (*Triticum aestivum* L.) is one of the most important crops globally. Together with rice, maize, and soybean, these crops supply more than half of the calories that are required for the world population (Zhao et al. 2017). Wheat yield is adversely affected by heat stress, and an approximately 6% loss in global yield is estimated with each Celsius degree increase in temperature caused by future global warming. The negative effects of heat stress on wheat yield are dependent upon the growth stages (Prasad and Djanaguiraman, 2014; Matsunaga et al. 2021a), and even a short duration of heat reduced wheat yield (Talukder et al. 2014). Therefore, the development of new climate change adaptation measures that include the optimization of wheat cultivation practices and breeding of heat-tolerant wheat varieties is essential for the thermo-stressed regions of the globe (Iizumi et al. 2021). Understanding the physiological and morphological responses to heat stress is essential for genetic and/or agronomic improvement in wheat.

Metabolome techniques have provided an important tool for understanding environmental stress tolerance mechanisms in plants (Ghatak et al. 2018; Hamany Djande et al. 2020; Thomason et al. 2018), and these techniques provide mining tools for analyzing the phenotypic/agronomic variations influenced by the environment. Metabolome-based chemical fingerprinting has been employed as a selection tool for desirable traits in crops (Hamany Djande et al. 2020). Different detection tools are available in the field of metabolomics, and multiple technologies are often required to gain comprehensive knowledge regarding the biochemical changes in each biological system (Ghatak et al. 2018). Among the various metabolomic platforms, liquid chromatography-mass spectrometry (LC-MS) and gas chromatography-mass spectrometry (GC-MS) are the commonly used technologies to date. These methodologies have been successfully used for the characterization of complex metabolic responses in wheat, including changes under post-anthesis

heat stress (Thomason et al. 2018), growth-stage-specific metabolic responses to heat (Matsunaga et al. 2021a), and the effects of post-anthesis heat stress on the metabolic profile of the grain (De Leonardis et al. 2015). Although LC-MS- and GC-MS-based metabolomics provide the advantages of higher capacities for the detection and identification of metabolites, applications of these techniques are limited to compounds possessing smaller molecular weights, and destructive samples preparation are required prior to analyses. In contrast, other metabolomic platforms, such as nuclear magnetic resonance (NMR) and Fourier transform infrared (FTIR) spectroscopy, possess the advantage of analyzing supramolecular structures such as cell walls with little pre-treatments requirements (Ghatak et al. 2018; McCann et al. 1992; Liu et al. 2021). FTIR spectroscopy excels other techniques by potential applicability to *in vivo* imaging of biological materials (Bouyanfif et al. 2017; Munz et al. 2017) and remote sensing (Li et al. 202). The FTIR spectroscopic technique has been used to study metabolic responses of plants to various environmental stresses, such as the salinity response in the beauty leaf tree (*Calophyllum inophyllum*) (Westworth et al. 2019), and differential metabolic behaviors of roots and leaves in a halophyte *Sesuvium portulacastrum* under salt stress (Nikalje et al. 2017). The FTIR spectroscopic technique has also been applied to different aspects of wheat such as metabolite distributions in the leaves under nitrate limiting conditions (Allwood et al. 2015), the oxidative-stress response of wheat roots (Zhao et al. 2013), and structural changes in gluten (Georget and Belton, 2006), as well as for phylogenetic research examining cultivated and wild wheat species (Demir et al. 2015). However, to our knowledge, no previous study has utilized this technique to detect metabolomic changes under heat stress in wheat. Therefore, the objective of the study in Chapter 1 was to establish a protocol for fingerprinting and developing chemical biomarkers that characterize the molecular responses of common wheat to heat stress.

1.2. Materials and Methods

1.2.1. Plant Materials and Growth Conditions

Non-sterilized seeds for the common wheat cultivar ‘Norin 61’ were put on a filter paper (qualitative filter paper No. 2, Advantec, Tokyo, Japan) that was trimmed to an approximately 85-mm diameter in a Petri dish (88-mm diameter), and the seeds were watered by applying 6 mL of tap water to the dish so that the paper became evenly wet. Twelve seeds were placed per dish and the seed/water mass ratio was 1:15. The dish was capped by a transparent lid to prevent water

evaporation, and the seeds were imbibed for three days at room temperature (25˚C) under a fluorescent room lamp illumination (a light intensity of approximately 10 µmol m^{-2} s^{-1}) from 9 a.m. to 5 p.m. The germinated seeds were then transferred to pots (a height of 10 cm and diameter of 5 cm) containing 120 g of commercial horticulture soil a brand "Oishii Yasaiwo Sodateru Baiyoudo", Cainz, Honjo, Saitama, Japan) containing composted bark, granular clay-like mineral, pumice, peat moss, perlite, and vermiculite. The soil was autoclaved at 121 ˚C for 30 min before planting. Pots were shifted to a growth chamber with a 14/10 h day/night regime, a relative humidity setting of 50%, a light intensity of approximately 500 µmol m^{-2} s^{-1}, and a temperature setting of 22/18˚C. Soil moisture level was maintained at 80–90% of field capacity (FC) (Assouline and Or, 2014) in duration of the experiment. The 100% FC was determined as described previously (Xu and Zhou, 2006). At the three-leaf stage and the length of the third leaf exceeded that of the second leaf samples were taken from plants and designated as C0. Six seedlings were kept in control chamber for three days (hereafter designated as C3 plants). At three-leaf stage, another six seedlings were transferred to a heat chamber with a daily maximum temperature of 42˚C and exposed to heat for three days (hereafter designated as H3 plants). In this heat chamber, the night temperature was 18˚C for 10 h, and the temperature setting was increased gradually by 5˚C per hour from the beginning of the light regime for 3 h to a maximum temperature of 42˚C and maintained for 6 h. The temperature was then dropped to 33˚C for 1 h and then decreased stepwise by 5˚C per h to 18˚C during the next 3 h.

1.2.2 Measurements of Plant Growth and Physiology

Physiological measurements were performed at three different conditions, including initial measurements on the day that treatment started (C0) and measurements at three days after the treatment for control (C3) and heat stress (H3) conditions. Canopy temperature assessment was carried out using FLIR-C2 thermal camera (FLIR system, Tallinn, Estonia). FLIR Tools software (v6.4.18039.1003) was exploited to estimate leaf surface temperature at 5 h after the beginning of the light regime. For leaf relative water content measurement, the third leaf was harvested at 5 h after the beginning of the light regime, and a 2 cm leaf segment was excised from the middle of the leaves. The fresh weight of the leaf segment was immediately measured using an electric balance, and turgid weight was measured after soaking the leaf segments in distilled water for 24 h at room temperature (25˚C). Tissue paper was used to dry leaf surfaces before the turgid weight

measurement. The leaf segments were transferred to an oven (EI-450B, ETTAS, AS-ONE, Osaka, Japan) at 70°C to achieve complete dryness, and the dry weight was measured. Following formula was applied to calculate relative water content (McCann and Huang, 2007):

$$100 \times ((Fw - Dw)/(Tw - Dw))$$

where Fw, Dw, and Tw refer to the fresh weight, dry weight, and turgid weight of the leaf segment, respectively. For the total of leaf length measurement, all leaves were harvested from the plants and scanned using a scanner (type DCP-J572N, Brother Industries, Nagoya, Japan). Leaf length was measured using ImageJ version 1.80 (ImageJ Home Page, Version 1.80. Available online: https://imagej.nih.gov/ij/index.html). In case of biomass measurement, all aboveground tissues of the individual plants were dried in an oven at 70°C until complete dryness. Their weights were taken after the samples were completely dried.

1.2.3. FTIR Measurement

The third leaf was harvested from the control and heat-treated plants and separately dried in an oven at 70°C till complete dryness. The dried leaves were powdered using an agate mortar and pestle. The ground samples (approximately 10 mg) were mixed with powdered KBr (IR grade, Nakalai, Kyoto, Japan) at a gravimetric ratio of 1:100, and approximately 10 mg of the mixture was placed into a dice of 7 mm diameter in a hydraulic press (Pixie Hydraulic Pellet Press, PIKE Technologies, Madison, WI, USA). A thin disk was formed by applying a pressure of 2.5 t cm^{-2}. Ten disks were generated from a single plant. FTIR absorbance spectra were obtained using a PerkinElmer Spectrum 65 spectrometer (Waltham, MA, USA) equipped with spectrum software version 10.4.2. Spectrum data were collected from the mid-infrared wavenumber range from 4000 to 400 cm^{-1} with a resolution of 1 cm^{-1} and 16 scans per measurement. Number of spectral measurements replications were three times per disk, with an exception of one disk from the heat stressed sample in which the measurement was performed only once. Data were collected from 60 disks where each was derived from six plants each from control and heat stressed plants, and 180 and 178 spectral data were obtained for control and heat stressed leaves, respectively.

1.2.4. Chemometrics of Spectral Data

FTIR spectra were baseline-corrected using a linear gradient of absorbance values at 4000 and 400 cm^{-1}, and the absorbance values were normalized to obtain a total value of 1 million for each spectrum. A principal component analysis (PCA) was applied using the prcomp function in the stat package (v3.6.2) in R Statistical Software (R Core Team, 2020), and ggplot function in the ggplot2 package (version 3.3.5) in R was utilized to draw the score plot and loading plot. For LDA, the 358 spectral datasets in the range from 3600 to 400 cm^{-1} wavenumber were randomly split into a training and test set at a ratio of 60% to 40% using the sample function in the base package (v3.6.2) in R and then calculated using the lda function in the MASS package (v7.3–54). For the development of spectral Fm biomarkers, a tailor-made R script was written to scan the two-candidate anchor point wavelengths in the 300 cm^{-1} range spanning the target wavenumber and for calculating the Fm values and *p* value in the student's *t*-test. The Fm values were calculated using the following formula:

$$Fm = (A_{target} - A_{anchor1}) / (A_{anchor2} - A_{anchor1})$$

where A target, A anchor1, and A anchor2 denote the normalized absorbance values for the target, and anchors 1 and 2, respectively. The R scripts were presented in appendix-1.

1.2.5. Statistical Analysis

The t.test function in the stats package (v3.6.2) was used for Student's *t*-test. Tukey's test was performed using the Astatsa.com online web statistical calculator (Astatsa. Complex Online Web Statistics Calculator. Available online: https://astatsa.com).

1.3. Results

1.3.1. Effect of heat stress in wheat growth and physiology

The common wheat cultivar 'Norin 61' was grown up to the three-leaf stage at a daily temperature of 22°C and then exposed to heat stress at a daily temperature of 42°C for three days. A significantly higher canopy temperature of 37.1°C ± 1.8 was observed in heat stressed plants (H3 plants) in comparison to 23.5 ± 1.9°C in unstressed control plants of the same age (C3 plants) and to 21.5 ± 2.0°C in the initial stage and before heat treatment (C0 plants) (Figure 1-1A). The relative water content of the leaves was comparable between (C3) and (H3) plants (81.3 ± 9.8% and 77.4 ± 7.6%, respectively), and these values were not significantly different compared to that

of (C0) plants (85.9 ± 2.5%) (Figure 1-1B). Although the total leaf length increased from 72.8 ± 6.6 cm in the (C0) plants to 88.2 ± 9.4 cm in (H3) plants, it was significantly less than that in C3 plants (102.0 ± 3.2 cm) (Figure 1-1C). Shoot biomass showed a similar trend, where the value for (H3) plants (0.120 ± 0.019 g) was decreased by 17.7% in comparison to that in (C3) plants (0.146 ± 0.009 g) (Figure 1-1 D)

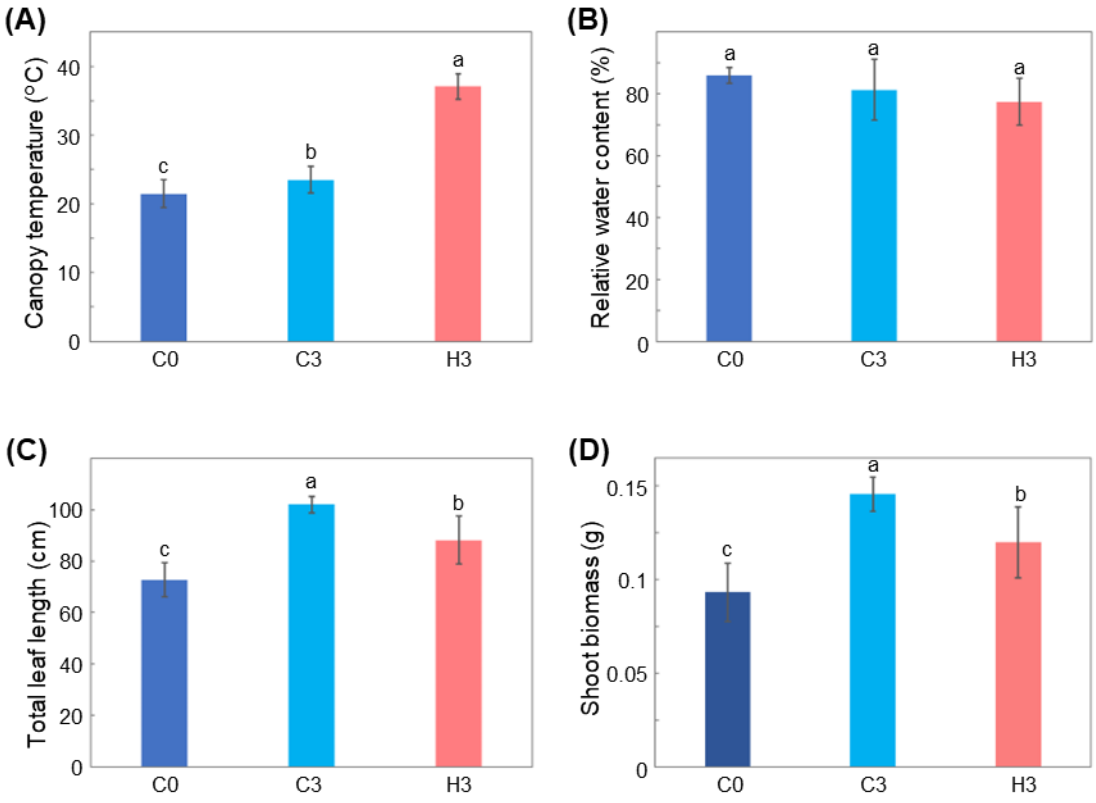

Figure 1-1. Impact of heat stress on wheat growth and physiology. (A) The canopy temperature, (B) relative water content, (C) total leaf length, and (D) shoot biomass of plants prior to heat (C0), of control plants after three days (C3), and of plants subjected to heat stress for three days (H3) are presented. Values are the average and standard deviation for 18–31 measurements from 5–6 plants in "(A)" and for 5–6 plants in "(B–D)". Statistical analysis was carried out by Tukey's range test ($p < 0.05$) and different letters (a, b, and c) were used to indicate significant differences between treatments.

1.3.2. FTIR and chemometric analysis: Principal Component Analysis:

The fully expanded third leaves of C3 and H3 plants were powdered, pressed with potassium bromide (KBr) to form pellets, and analyzed using FTIR spectroscopic technique. Figure (1-2) presents a typical example of the FTIR spectrum of each plant. These spectra presented largely similar patterns with a characteristic broad peak in the range of 2700–3700 cm^{-1}, a number of sharper peak signals at approximately 2900 cm^{-1}, complex contours in the 900–1800 cm^{-1} range, and relatively minor peak signals at approximately 400–800 cm^{-1} (Figure 1-2). The broad peak at approximately 2700–3700 cm^{-1} can be assigned as O–H, C– H, and N–H stretching, while the sharper peak signals at approximately 2900 cm^{-1} can be interpreted as C–H stretching bands from aliphatic compounds (Stuart, 2004). In the so-called "finger-printing" region ranging from 400–1800 cm^{-1} (Kamnev et al. 2018), multiple peak signals are recognizable that largely overlapped and formed complex patterns. At least 12 peaks were recognized in the spectra from both C3 and H3 plants, which can be assigned to various functional groups as shown in (Table 1-1). However, it is noteworthy that the pre-measurements sample preparation conditions employed in this study, such as drying the leaf tissues at 70°C, grinding, and the usage of a KBr matrix, might affect the wavenumber positions of maxima of some polar functional groups of biomolecules. Previous reports have demonstrated that the employment of a KBr matrix and grinding resulted in the shifts of some FTIR vibrational bands by up to 15 cm^{-1}, which might influence the band energies, affect ion exchange, and induce crystallization of metastable amorphous biopolymers (Kamnev et al. 2018; Kamnev et al. 2021). However, from visual inspection, it was not easy to identify distinguishable features between C3 and H3 plants, and this suggested that the use of chemometric techniques is required for spectral analysis.

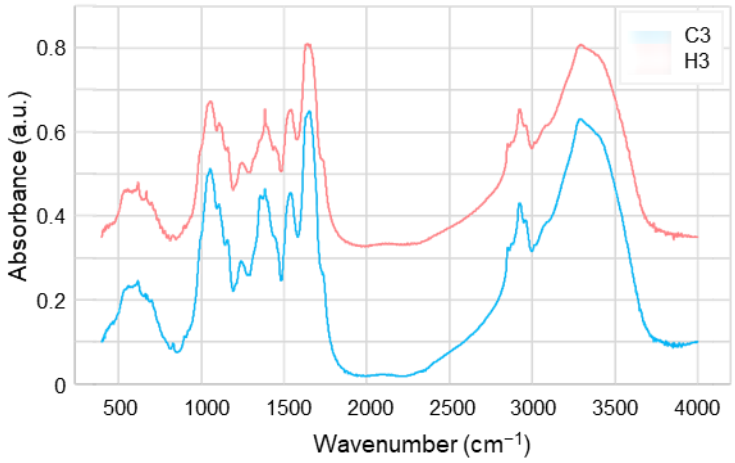

Figure 1-2. Representative FTIR spectra from the leaves of control (C3) and heat stressed (H3) wheat.

Table 1-1. Major FTIR peaks observed and their assignment to probable functional groups in wheat leaves

No.	Wavenumber (cm^{-1})	Probable Functional Groups
1	3293	O-H stretching, N-H stretching.
2	2960	C-H stretching in $-CH_3$ (antisymmetric).
3	2925	C-H stretching in $-CH_2-$(antisymmetric).
4	2852	C-H stretching in $-CH_2-$ (symmetric).
5	1651	C=C stretching, C= O stretching (amide), N-H bending (amide I).
6	1541	C=C stretching (aromatic), N-H bending (amide II), C- N stretching.
7	1385	C-H bending (antisymmetric), = C-H in-plain bending.
8	1241	C-O stretching, In-plain C-H bending (aromatic), aliphatic C-O stretching, P=O stretching (aliphatic)
9	1158	C-O stretching, C-N stretching (aliphatic), In-plain C-H bending (aromatic), aliphatic C-O stretching.
10	1106	C-O stretching, C-N stretching (aliphatic), In-plain C-H bending (aromatic), aliphatic C-O stretching.
11	1055	C-O stretching, C-N stretching (aliphatic), In-plain C-H bending (aromatic).
12	618	=C-H out-of-plane bending, =C-H bending, C-S stretching.

Assignment of wavenumbers to probable functional group are according to (Kamnev et al. 2021; Stuart, 2004; Talari et al. 2016)

Subsequently, a principal component analysis (PCA) was applied to characterize the spectral differences between C3 and H3 plants. Figure 1-3A provides the PCA score plot of 358 spectra (180 and 178 spectra from (C3) and (H3) plants, respectively) that was based upon the variables of 3601 data points (normalized absorbance values from 400 to 4000 cm^{-1} with 1 cm^{-1} resolution) for each spectrum. The PC1-PC2 space in the plot explained 81.1% of the total variance (Figure

1-3A and Figure 1-4). Consequently, spectra from (C3) and (H3) plants were mostly clustered on the PC2-positive and negative half planes, respectively, suggesting the presence of distinct spectral features between (C3) and (H3). Loading plots of the PCA showed complex patterns (Figure 1-3B–D); regions for PC2 loading over 0.5 were observed in wavenumbers of 459–484, 564–607, 610–614, 622–665, 670–752, 1177–1344, and 1351–1471 cm^{-1}, whereas PC2 loading below -0.5 were seen in the regions of 2736–2897 and 2977–3082 cm^{-1} (Figure 1-3D), indicating that absorbance of these specific positive and negative regions tended to influence separation of (C3) and (H3) plants. However, considerable numbers of (C3) and (H3) spectra were mixed in the central origin of the score plot (Figure 1-3A), suggesting that the PCA alone was not sufficient to distinguish the spectral features in heat-stressed wheat leaves.

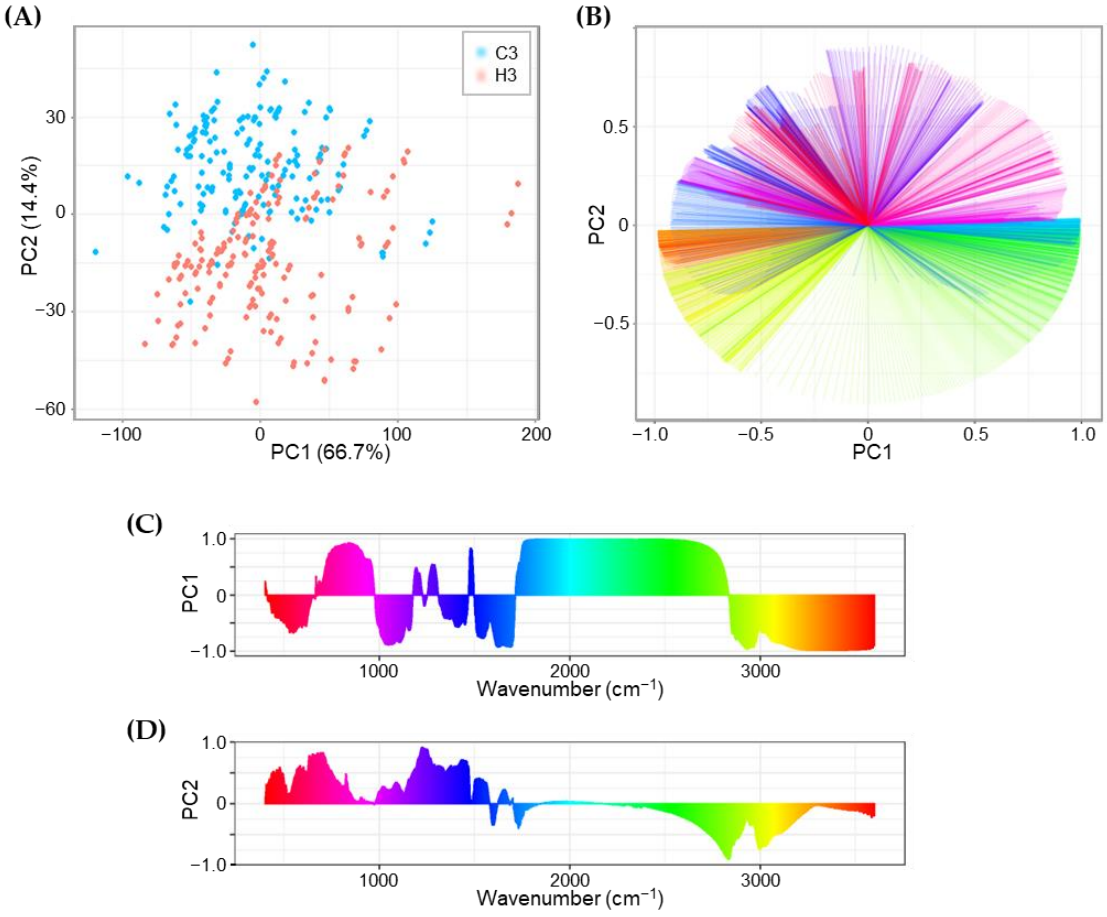

Figure 1-3. Principal component analysis of FTIR spectra. (A) A score plot showing overlapping distribution between (C3) and (H3) plants. (B) A two-dimensional loading plot. Assignment of a

color gradient to respective wavenumbers are the same as those presented in (C, D). (C–D) One-dimensional loading column plots for (C) PC1 and (D) PC2. The loading for each wavenumber is expressed using a color gradient image along their *x-axis*.

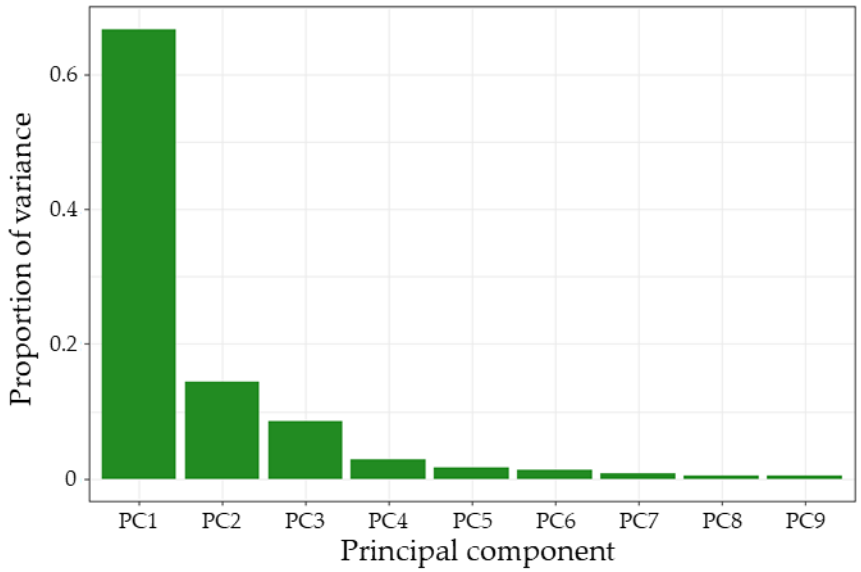

Figure 1-4. Variance explained by the first 9 components in principal component analysis.

1.3.3. Linear Discriminant Analysis

A linear discriminant analysis (LDA) was utilized to improve the discrimination of heat stressed leaves. The 358 FTIR spectra that consisted of 180 and 178 spectra from (C3) and (H3) leaves, respectively, were randomly divided into two groups at a ratio of 60:40%. The 60% group was used as a training set in the supervised machine learning process to build a linear discriminant model. The LDA algorithm successfully distinguished the training dataset into the heat stressed leaves from the controls in the histogram (Figure 1-5, where the FTIR spectra with positive and negative LD1 scores corresponded to those taken from H3 and C3 leaves, respectively. The remaining 40% of the test dataset was then applied to the model for validation, and the results presented a slightly broader frequency distribution for both C3 and H3 in the histogram compared to those in the training set, while essentially confirming a clear discrimination between heat-stressed and control leaves (Figure 1-5B). Therefore, the FTIR spectral fingerprint coupled with

the LDA approach was demonstrated to be effective in detecting discriminatory biochemical information in heat-stressed wheat leaves.

To detect which parts of the spectra were more contributing to discriminating between heat stressed and control leaves in LDA, the LDA loadings were studied. A plot of LDA loadings versus wavenumber revealed that several spectral regions, under a threshold of absolute loading intensity over 0.15, were more important in regard to the discrimination ability (Figure 1-6. The plot showed two strong positive loading peaks at 1465 cm^{-1} (loading intensity of 0.398) and 1729 cm^{-1} (0.176) that provide the highest LDA score in the H3 leaves, and four strongly negative loading minimum points of 1251 cm^{-1} (loading intensity of −0.318), 576 cm^{-1} (−0.250), 1502 cm^{-1} (−0.224), and 482 cm^{-1} (−0.183) that gives the lower LDA score in the C3 leaves. These six spectral points were in position within the multiple peaks overlapping region at 400–1800 cm^{-1} in the FTIR spectra (Figure 1-2) and corresponded to the finger-printing region (Stuart, 2004). These spectral regions may reveal changes in the chemical compositions and/or structures under heat stress that can potentially serve as spectral biomarkers for diagnosing heat-stress exposure in wheat leaves.

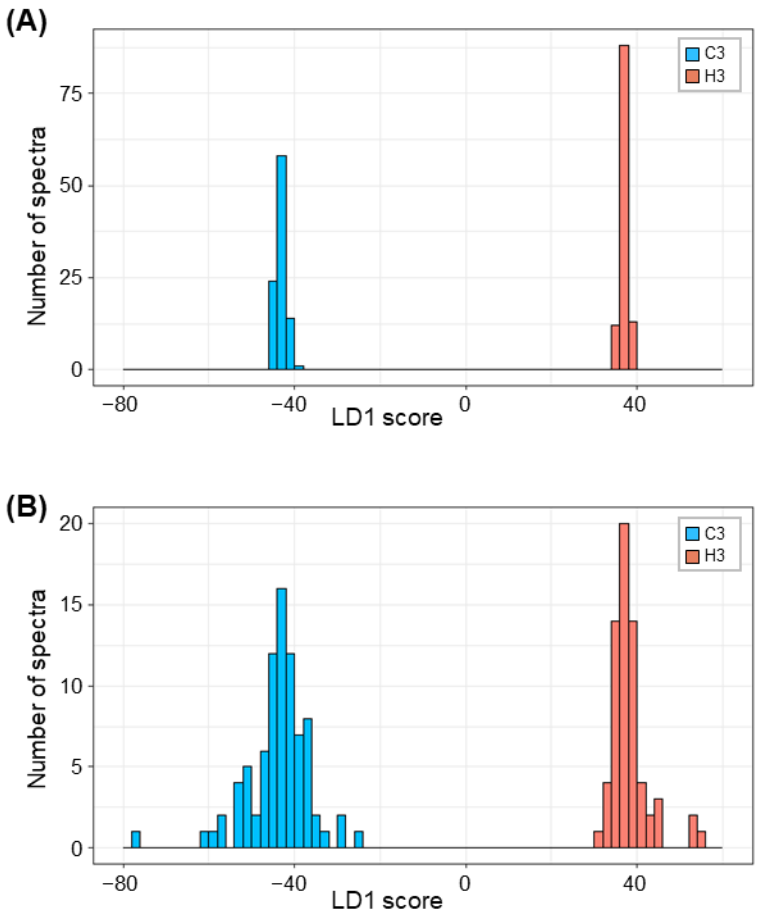

Figure 1-5. Two group histograms of the training (A) and the test (B) sets for FTIR spectra based on the LD1 score in the linear discriminant analysis, demonstrating classification performance between control and heat-stressed leaves.

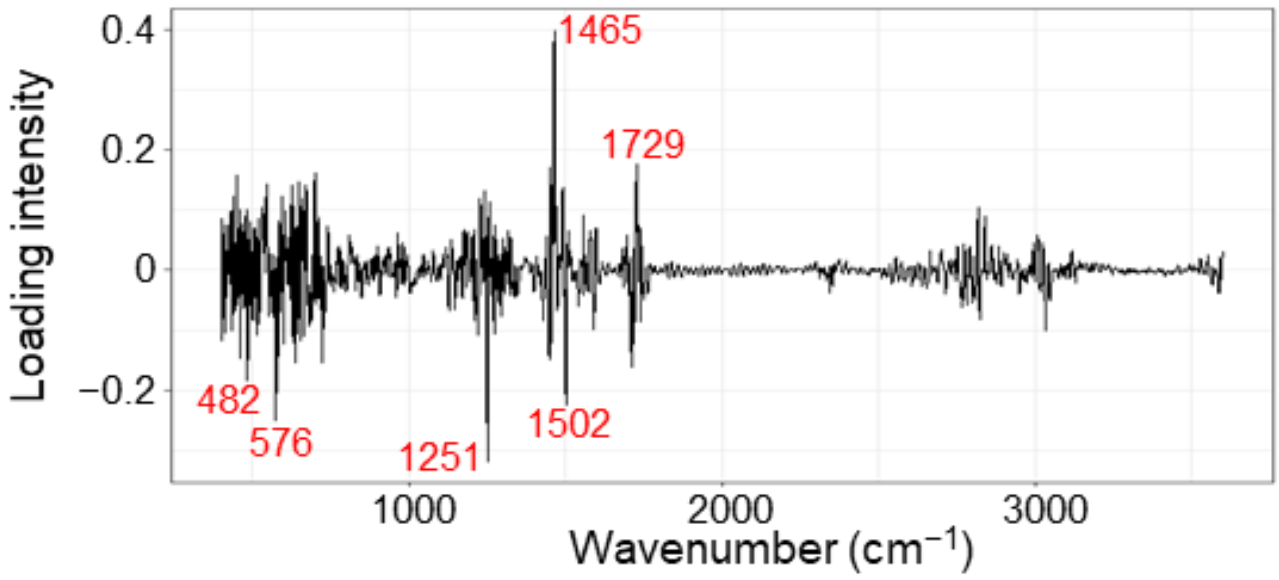

Figure 1-6. Identification of the discriminatory spectral region. Loading plot of the LDA results that were used for the classification of (H3) and (C3) leaves. Wavenumbers for the major peaks and minimum turning points are indicated by red fonts.

1.3.4. Spectral Biomarkers for Heat Stress Response

To probe the possibility of developing spectral biomarkers specific for the heat stress response, the spectral regions that were identified as the major discriminants in the LDA loading plot presented in (Figure 1-5) were further evaluated. To achieve this goal, two anchor points that encompass the target wavenumber were set, and a new term "Fm" (FTIR marker) that functions as a normalized target absorbance indicator was defined according to the compensate absorbance values of the first and second anchors as 0 and 1, respectively (described in the Materials and Methods section 1.2.4.). The two anchor points were scanned in the surrounding of the target wavenumber and selected according to the following criteria: (i) the distance between the anchor point and target was within 150 cm^{-1}; (ii) statistical significance (*p*-value) of difference by Student's *t*-test for Fm values between heat stress and control is below 0.0001; (iii) anchor points are preferably located at visually obvious landmarks such as spectral peaks and minimum or

inflection points within the spectral curves. The Fm values for the target wavenumber were calculated using 358 FTIR spectra data from (C3) (180 spectra) and (H3) (178 spectra) plants. Accordingly, anchor-1 and -2 were chosen as shown in (Table 1-2). A comparison of the averaged FTIR spectra between (C3) and (H3) plants in the magnified views showed that the target wavenumbers were mostly situated in the middle of spectral slopes (Figure 1-7). The normalized absorbance at the target wavenumbers were somewhat, but consistently, higher in Fm1465 and Fm1729 (Figure 1-7A, B), and lower in Fm1251, Fm576, Fm1502, and Fm482 (Figure 1-7C–F). Although knurl-like noises were detected in the FTIR spectra in the wavenumber range around 405–480 cm^{-1}, a difference of absorbance at the target wavenumber of 482 cm^{-1} was obviously larger than the fluctuation of the noises (Figure 1-7F). Box plots exhibited that the biomarkers Fm1465 and Fm1729 presented significantly higher Fm values in (H3) plants compared to those in (C3) plants (Figure 1-8A, B, Table 1-2), while the other biomarkers (Fm1251, Fm576, Fm1502, and Fm482) possessed statistically less Fm values in (H3) plants compared to those in (C3) plants (Figure 1-8C–F, Table 1-2). This was consistent with the positive and negative LDA loading values (Figure 1-5), respectively.

Table 1-2. Characteristics of spectral biomarkers.

Marker name	Wavenumber (cm^{-1})			Loading*[1]	Median Fm value		H3/C3 ratio *[2]	*P* *[3]
	Target	Anchor 1	Anchor 2		C3	H3		
Fm1465	1465	1480	1399	0.398	0.345	0.381	1.104	2.1×10^{-61}
Fm1729	1729	1768	1703	0.176	0.559	0.588	1.052	3.8×10^{-80}
Fm1251	1251	1241	1358	-0.318	-0.0428	-0.112	2.607	3.1×10^{-26}
Fm576	576	648	542	-0.25	1.436	0.899	0.626	1.1×10^{-4}
Fm1502	1502	1480	1615	-0.224	0.335	0.294	0.879	4.4×10^{-75}
Fm482	482	401	501	-0.183	0.741	0.666	0.899	7.9×10^{-21}

1 Loading score of LDA at the target wavenumber.

[2] H3/C3 ratio of median Fm values.

[3] Probability by *t* -test.

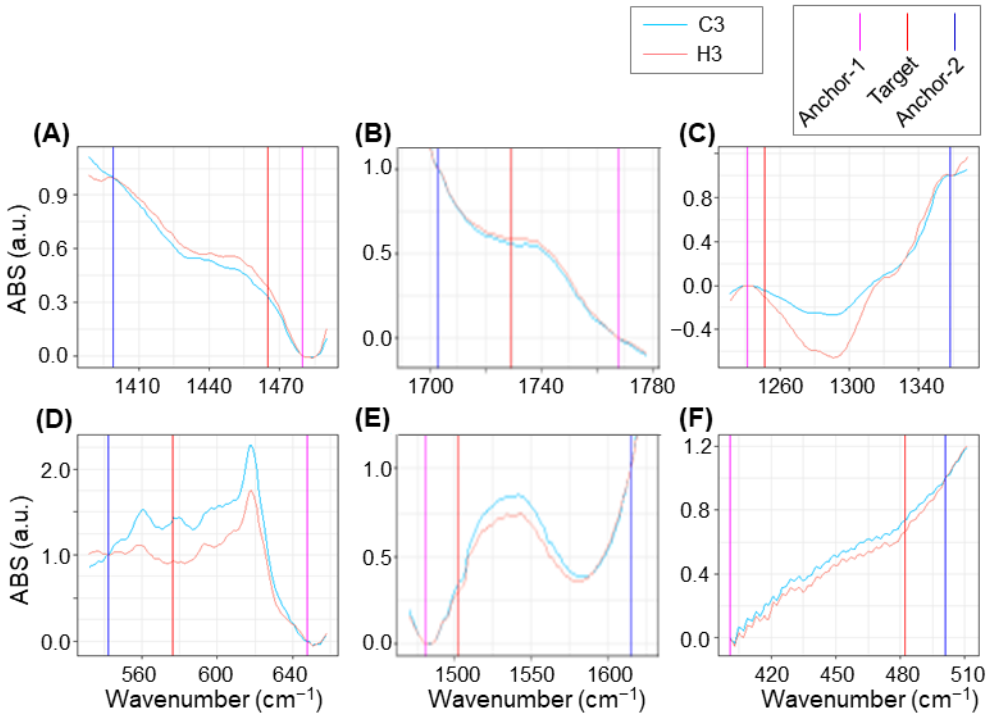

Figure 1-7. Magnified view of averaged FTIR spectra in the vicinity of Fm biomarkers. Normalized spectra for (C3) (cyan) and (H3) (orange) plants in the vicinity of (A) Fm1465, (B) Fm1729, (C) Fm1251, (D) Fm576, (E) Fm1502, and (F) Fm482 markers are shown. Red, magenta, and blue vertical lines designate the locations of wavenumbers for the target, anchor-1, and -2, respectively.

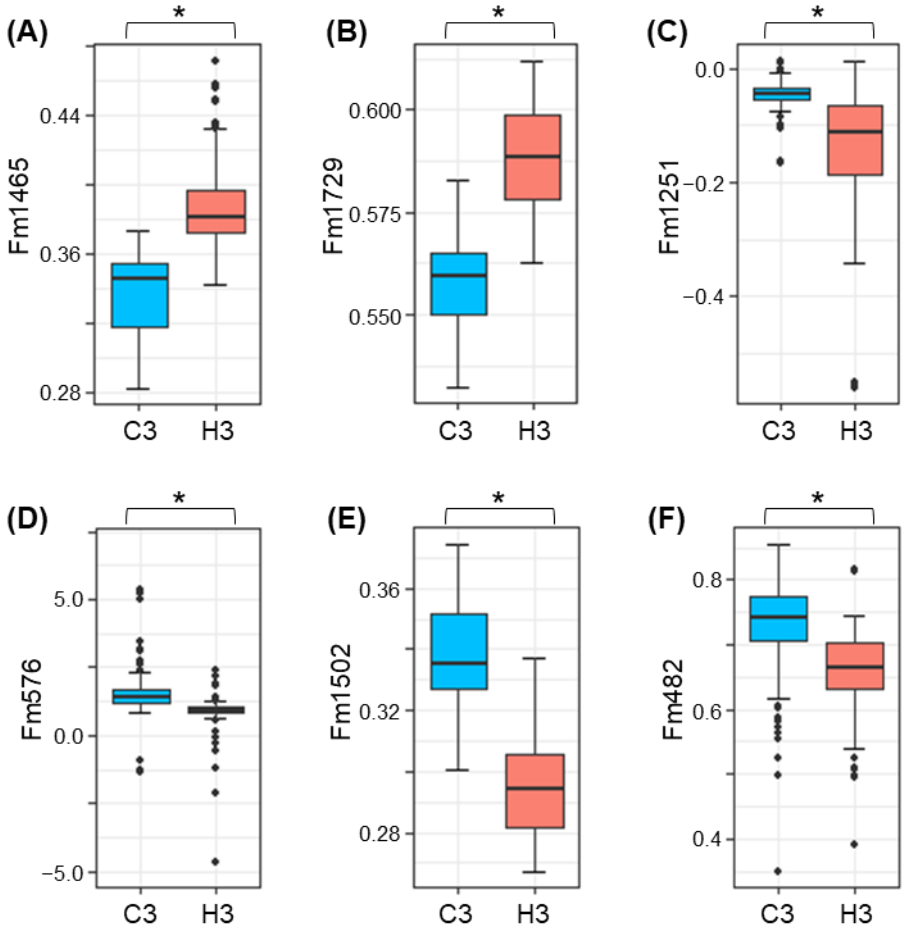

Figure 1-8. Box plots presenting the comparison of spectral biomarker values between (C3) and (H3) plants. Plots for the spectral biomarkers (A) Fm1465, (B) Fm1729, (C) Fm1251, (D) Fm576, (E) Fm1502, and (F) Fm482 are provided. The 180 and 178 spectra were used for the (C3) and (H3) plots, respectively. Asterisks represent statistically significant differences at $p < 0.001$.

1.4. Discussion

1.4.1. Sensitivity of FTIR Spectral Response in Heat-Stressed Wheat Leaves

In the current study, the FTIR spectroscopic technique was successfully applied to discriminate heat-stressed wheat leaves from those of control plants, thus revealing that this technique can serve as an analytical tool for tracing chemical changes during heat stress in wheat. The heat stress applied in this study led to delayed leaf growth and biomass production (Figure 1-1C, D), and this was similar to the observations reported in previous studies (Keles and Önce, 2002; Gupta et al. 2013) where shoot growth was persisted to some degree under stress, thus indicating that the intensity of the heat stress employed in this study was not at a lethal level. The relative water content in the leaves was statistically unchanged by heat stress in this study, unlike previously reported cases of heat-induced reduction in wheat leaves length (Ramani et al. 2017; Sattar et al. 2020). This further indicated that the stress intensity in this study was relatively modest. However, significant spectral differences were detected by FTIR spectroscopy, thus suggesting that FTIR-based fingerprinting was sensitive enough to characterize changes in the chemical constituents of wheat under nonlethal heat-stress conditions.

1.4.2. Chemometrics Using FTIR Spectra

As the FTIR spectra from heat-stressed and control plants were similar upon initial inspection (Figure 1-2), the application of chemometric methods was indispensable for obtaining a better interpretation of the FTIR spectra. Chemometric methods are commonly used to fetch more information from the obtained FTIR spectroscopic data (Allwood et al. 2015; Johnson et al. 2003; Giang et al. 2020; Rohman et al. 2020). As PCA alone was not capable to fully interpret the spectra (Figure 1-3), additional chemometric methods were included. We applied LDA, which successfully discriminated between heat stressed and control leaves, and we demonstrated the potency of the FTIR-based chemometric approach for diagnosing plant heat stress status. Many previous literatures have applied various chemometric methods. Johnson et al. (2003) utilized PCA in combination with genetic algorithms to fingerprint salt-stressed tomato varieties. Recently, Nikalje et al. (2019) applied PCA for characterizing metabolic responses of roots and leaves in a halophyte *S. portulacastrum*, demonstrating that FTIR spectroscopy differentiated different tissues and stress intensity in the PC1-PC2 plane. Cortizas and López-Costas (2020) used PCA coupled with structural equation models to study the compositional and archaeological changes in human bone collagen. Grunert et al.(2020) used PCA for factor extraction, and this was followed by the use of two types of supervised machine learning methods (PCA-LDA and PCA-Mahalanobis

discriminant analysis) for the classification of peritoneal dialysis effluent. Chemometric interpretation of FTIR spectra using the combination of PCA-LDA was capable to differentiate embryonic stem cell in murine models (Ami et al. 2010) and for the identification of spectral markers for putative stem cell regions of human intestinal crypts (Walsh et al. 2008). In harmony with these previous studies, the LDA applied in the present study successfully discriminated between heat-stressed and control leaves (Figure 1-4), thus revealing the potency of the FTIR-based chemometric approach for diagnosing plant heat-stress status.

1.4.3. FTIR-Based Biomarker for Chemical Changes under Heat Stress

The development of FTIR-based biomarkers has proven to be an effective analytical method in various scientific fields, including medical diagnosis (Grunert et al. 2020), food quality control (Giang et al. 2020; Rohman et al. 2020), and forensic analysis of cosmetic compounds (Sharma et al. 2019). In this study, we developed FTIR-based spectral biomarkers that were based on the LDA loading intensity at specific wavenumbers (Figure 1-5). The developed biomarkers successfully discriminate heat stressed leaves from controls (Table 1-1; Figure 1-7). Among the six biomarkers developed, Fm1465 and Fm1729 showed an increase under heat stress, while Fm1251, Fm576, Fm1502, and Fm482 reduced under heat stress. Among the biomarkers that increased under heat stress, the wavenumber for the marker Fm1465 was located in the major region reported as a broad and poorly resolved C–H bending and C–O stretching region (Stuart, 2004) that has been reported as a region for suberin/cutin in plant extracellular space (Stewart, 1996; Lammers et al. 2009). The peak at 1465 cm^{-1} is also located close to reported assigned signals of 1463 cm^{-1} for CH2 scissoring and 1460 cm^{-1} for CH3 asymmetric bending in lipids (Stuart, 1997) and the C–H signal in cell wall polysaccharides (Stuart, 2004; Gorgulu et al. 2007). Modifications of these candidate compounds in heat-stressed plants have been previously studied, including the complex regulation of leaf lipid composition in wheat (Narayanan et al. 2016) and heat stress-induced changes of cell-wall components in the leaves of coffee (Lima et al. 2013) and wheat (Zhang et al. 2010). Another biomarker which increased under heat stress in the current study was Fm1729. This wavenumber region can be interpreted as stretching vibrations of ester C=O groups, which (along with the aforementioned bending C–H vibrations) are typical for lipids (Kamnev et al. 2021; Stuart, 2004; Talari et al. 2016). Similar rise in peak intensity around this region were detected in pea pollen grains under heat stress (Lahlali et al. 2014), which may indicate quantitative/qualitative regulation

of the pollen exine layer under the stress. The increase in the Fm1729 value in wheat leaves in the present study may, therefore, explain the adaptive alteration of lipid composition under heat stress. Alternatively, the increase in the Fm1729 value may indicate heat-induced injury in leaf lipids. Malondialdehyde (MDA), a main product of lipid peroxidation as a consequence of oxidative stress, has a characteristic FTIR signal around 1700–1750 cm^{-1} (Oleszko et al. 2015). An increase of MDA was documented in wheat seedlings subjected to heat stress (Savicka and Škute, 2010). Among down-regulated Fm markers, the wavenumber of the Fm1251 marker was in a position in the vicinity of the previously assigned signals of 1240 cm^{-1} for hemicellulose and 1260 cm^{-1} for pectin (Stuart, 2004; Mascarenhas et al. 2000). Pectin substances in the extracellular matrix have been detected to function as a major adapting factor for cell wall porosity in soybean cells (Baron-Epel et al. 1988), thus supporting the hypothesis that adaptation to the heat environment may involve chemical rearrangement of pectins and foliar heat conductivity (Lima et al. 2013). Other Fm markers that reduced under heat stress included Fm1502. A previous study by Kurian et al. (Kurian et al. 2015) interpreted the wavenumber regions 1502–1600 cm^{-1} as aromatic skelet al. vibration of lignin. Lima et al. (2013) detected modification of lignin monomer composition after three days of heat stress in coffee leaves, suggesting that plant responses to the heat environment may include the structural modulation involving lignocellulose supramolecular structure. Nevertheless, assignments of the proposed Fm biomarkers to any specific compounds are currently premature due to the intrinsic nature of overlapping signals in FTIR spectra and cumulative steric and/or electronic effects in a given molecule that can potentially lead to a large shift in spectral signals (Stuart, 2004). To identify the molecular entities for these Fm biomarkers, further future biochemical and/or genetic studies are anticipated that may combine multifaceted approaches, including biomass fractionation, mass spectrometry, and genetic mapping.

1.4.4. Application of FTIR-Based Metabolome Profiling on Agronomy

The present study suggests that FTIR-based chemical fingerprinting can serve as a versatile tool for diagnosing plant physiological condition under various environmental conditions, including heat stress. Metabolomics has been used as a powerful analytical tool to understand the association between agronomic performance and the underlying molecular mechanisms (Ghatak et al. 2018; Hamany Djande et al. 2020; Thomason et al. 2018; Razzaq et al. 2019). The versatility of FTIR spectroscopy has been demonstrated in previous studies in regard to discriminating

genotype differences in cultivated and wild wheat species (Demir et al. 2015) and rice varieties (Giang et al. 2020). Moreover, FTIR spectroscopy provides high-throughput measurements (Shapaval et al. 2010; Ba ˘gcıo˘glu et al. 2017), promising that it can be used for chemo-typing heat stress responses in crop breeding programs. Unsophisticated setup of FTIR spectroscopic facilities in comparison to that of other metabolomic platforms may also be beneficial for applying this technology to field metabolome studies (Mandrone et al. 2021; Galleni et al. 2021). Nowadays, FTIR-based remote sensing technologies have emerged as a new tool for monitoring the surface properties of land (Li et al. 2021, Yalkun et al. 2019), and this may further broaden the possibility of developing spectrum-based plant diagnoses for crop production and breeding.

CHAPTER 2

Investigation of Differential Metabolome Responses among Wheat Genotypes to Heat Stress using FTIR Chemical Fingerprinting

2.1. Introduction

Wheat (*Triticum aestivum* L.) is one of the most important staple crops that enrich humans' diets with an important source of nutrients (Shewry and Hey, 2015). Wheat, with other cereals and soybean contribute to more than 50 percent of the calories required by the global population (Zhao et al. 2017). Among several abiotic stresses that restrict wheat production, heat stress remains one of the major challenges. The reduction in wheat yield at high temperatures is intensively studied (Mitchell et al. 1993; Stone and Nicolas, 1995; Semenov and Halford, 2009; Schittenhelm et al. 2020; Matsunaga et al. 2021). This is expected to be further deteriorated in the light of ensuing climate change. Severe global warming with a fast rate of global temperature increases of up to 5°C is predicted by the end of this century (Solomon et al. 2007). Therefore, understanding the heat response of wheat is crucial to facilitate the development of new heat-tolerant varieties.

Wheat genetic resources and their diversity have been investigated extensively (Reif et al. 2005; Gorafi et al. 2018; Balfourier et al. 2019), and wide variation in heat stress sensitivity among genotypes has been reported (Tadesse et al. 2019; Qaseem et al. 2019). Chinese Spring has been reported as a heat-sensitive genotype (Wang et al. 2018; Qin et al. 2008), whereas Norin 61 revealed heat tolerance in hot arid regions in Sudan in field studies (Elbashir et al. 2017a; Elbashir et al. 2017b). The genetic makeup of these two genotypes have been previously reported (Walkowiak et al. 2020). Imam is a heat-tolerant cultivar widely grown in Sudan, which is considered as the world's hottest wheat growing environment (Iizumi et al. 2021) and has been used as a reference genotype for detecting heat tolerance in other varieties (Elbashir et al. 2017a).

Metabolomics is one of the omics tools used to study the molecular responses of plants and has been utilized to trace metabolic responses in plants under various stresses (Ghatak et al. 2018; Hamany Djande et al. 2020; Matsunaga et al. 2021). Metabolomics has been applied to plant breeding programs because the metabolome is arguably more closely related to the phenotype than other "omics" data (Sakurai, 2022). Among the several technical platforms of metabolomics,

Fourier transform infrared (FTIR) spectroscopy is unique in that it renders an opportunity to study biological samples *in vivo* in a non-destructive manner (Bouyanfif et al. 2017; Munz et al. 2017; Petrou et al. 2018), is compatible with remote sensing in the field (Li et al. 2021; Yalkun et al. 2019), and facilitates the analysis of complex biomacromolecules such as cell wall components (McCann et al. 1997; Liu et al. 2021). FTIR spectroscopy has been applied to study the metabolome response of plants to various environmental stresses (Zhao et al. 2013; Lahlali et al. 2014; Westworth et al. 2019; Nikalje et al. 2019). In the study explained in chapter 1, the utilization of FTIR combined with chemometrics successfully identified spectral changes that distinguished heat-stressed and control leaves in the bread wheat genotype" Norin 61" (Osman et al. 2022a). Therefore, the aim of the current study was to test whether the FTIR spectroscopic technique is useful for characterizing the metabolome diversity of wheat genotypes with variable heat tolerance abilities. To reach this goal, three wheat genotypes, 'Chinese Spring', 'Imam', and 'Norin 61', with different heat tolerance capabilities, were used in this study.

2.2. Materials and Methods

2.2.1. Plant Growth Condition

Seeds of wheat genotypes Chinese Spring and Imam were kindly provided by Dr. Hiroyuki Tanaka (Faculty of Agriculture, Tottori University, Tottori, Japan). Seeds of wheat genotype Norin 61 was kindly provided by Dr. Yasir Serag Alnor Gorafi (Arid Land Research Center, Tottori University, Tottori, Japan). Total of twelve seeds each of the three wheat genotypes were distributed on top of an 85-mm diameter filter paper (Filter paper type-2, Advantec, Tokyo, Japan) in a Petri dish of 90-mm diameter and drained by adding 6 ml of tap water. The Petri dish was covered by a transparent lid and incubated for three days at room temperature (25°C). Germinated seedlings were individually transferred to pots containing 120 g of commercial horticulture soil (a brand "Oishii Yasaiwo Sodateru Baiyoudo," Cainz, Honjo, Saitama, Japan). Pots were placed in a growth chamber with light/dark regimes set at 14/10 h, light intensity of approximately 500 µmol m^{-2} s^{-1}, relative humidity 81 setting at 50%, and temperatures at 22/18°C for light/dark regimes. When the length of the third leaf become longer than that of the second leaf, half of the pots were shifted to a heat chamber with a daily temperature setting of 42/18°C under light/dark regimes. In

this heat chamber, the temperature was set to ascend stepwise from 18°C at the beginning of the light regime by 5°C/h for 3 h, then raised to the maximum temperature of 42°C and stabilized for 6 h. The temperature was then decreased to 33°C for 1 h and then decreased stepwise by 5°C/h to 18°C in the next 3 h. Heat treatment was applied for three days, and the plants were subjected to the analyses described below.

2.2. Measurement of Canopy Temperature and Plant Growth

For the measurement of canopy temperatures, the leaf surface temperature of wheat plants at 5 h after the starting of the light regime of the day was measured using a thermal camera as mentioned previously (Osman et al. 2022a). To measure leaf length, all attached leaves of an individual plant were measured using a ruler, and the values were averaged. To measure shoot biomass, the aerial parts of individual plants were collected and completely dried in an oven (EI-450B, ETTAS, AS-ONE, Osaka, Japan) at 70°C for three days, and the dry weight was measured.

2.3. FTIR spectroscopy

Fully expanded third leaves of both the unstressed and heat-treated plants were collected and completely dried in an oven at 70°C. The whole dried leaf (approximately 0.16g per leaf) was placed into a 15 ml of plastic tube containing three stainless beads with different sizes one with 10- and the other two with 5-mm diameter, respectively, and shifted to a pre-chilled aluminum block in a shaker homogenizer (Shake Master Auto, Bio Medical Science, 105 Tokyo, Japan). To obtain fine powder the machine was set at 1,100 rpm for 30 min. The powdered samples (approximately 10 mg) were mixed with 1 g of powdered KBr (IR grade, Nakalai, Kyoto, Japan), and approximately 10 mg of the mixture was transferred to a dice of 7 mm diameter in a hydraulic press (Pixie Hydraulic Pellet Press, PIKE Technologies, Madison, WI, USA). A thin disk was generated by applying a pressure of 2.5 t cm^{-2}. Three disks were made from a single plant. FTIR spectra were measured in absorbance mode using PerkinElmer Spectrum 65 (Perkin Elmer, Waltham, MA, USA) attached with Spectrum software (version 10.4.2., Perkin Elmer). The measurements were taken at mid-infrared wavenumbers from 4000 to 400 cm^{-1}, with wavenumbers interval of 1 cm^{-1}, and 16 scans were recorded and averaged for each measurement. Measurement was repeated twice for each disk; therefore, six spectra were obtained from a single

plant. Six plants were used for each genotype and environmental condition; therefore, 36 spectral data points were gained for each genotype-environment combination.

2.4. Chemometrics of FTIR Spectra and Statistical Analyses

Obtained spectrum data was subjected to baseline correction and normalization prior to analysis as described previously (Osman et al. 2022a). Chemometric analysis of the FTIR spectra were performed using R statistical software (R Core Team, 2020), applying a set of custom-made R scripts that were provided in Appendix-2. Briefly, principal component analysis (PCA) was applied to the wavenumber region between 3600 and 400 cm^{-1} using the prcomp function in the stat package (version 3.6.2) in R. To develop Fm biomarkers calculation of pair of anchor points for generating offset absorbance values was performed as described previously (Osman et al. 2022a). For linear discriminant analysis (LDA), the wavenumber region between 3600 and 400 cm^{-1} in the 216 spectral dataset made of 36 spectra each from six genotype-environment combinations (3 genotypes × 2 environment) was utilized for the development of an equation model using the lda function in the MASS package (v7.3-54) in R. Presentation of the resultant dataset, such as the score and loading plots in PCA, box plots in Fm biomarkers, LD1-LD2 biplot, and their scaling plots in LDA, were generated using the ggplot2 package (v3.3.5) in R. The Student's *t*-test was performed using the t.test function in the stat package in R. One-way ANOVA with post-hoc Tukey HSD test was performed using the Astatsa.com online statistical calculator ($p < 0.05$) (Astatsa. Complex Online Web Statistics Calculator. Available online: https://astatsa.com).

2.3. Results and Discussion

2.3.1. Impact of Heat Stress on Growth of Three Wheat Genotypes

Wheat genotypes 'Chinese Spring' (CS), 'Imam', and 'Norin 61' (N61) were grown till it reach the three-leaf stage, at a daily temperature of 22°C, and then subjected to heat stress at a daily maximum temperature of 42°C for three days. Canopy temperatures were significantly higher under heat stress in all three genotypes (Figure 2-1A, Table 2-1.); the median temperatures on day 0 (hereafter referred to as C0) were in the range of 23.0°C–26.6°C for the three genotypes and

elevated to the range of 36.6°C–37.1°C on day 3 under heat stress (H3). As a result, large differences in canopy temperatures between C3 and H3 were noticed in these genotypes; the difference in the median temperature was 12.2, 16.3, and 13.1 in CS, Imam, and N6, respectively. Total leaf length was strongly reduced under heat stress (Figure 2-1B, Table 2-2.), in which the mean values of total leaf length were suppressed by 26.9, 19.7, and 13.3% in H3 plants for CS, Imam, and N61, respectively, in comparison to their C3 counterparts. Shoot biomass also significantly decreased under stress (Figure 2-1C, Table 2-3). The mean biomass values decreased by 29.8, 25.8, and 15.7% in H3 plants for CS, Imam, and N61, respectively, in comparison to their C3 counterparts. Although the N61 genotype showed a minimum degree of biomass loss, Imam genotype still showed the highest shoot biomass on day 3 of heat stress. Similarly, high biomass production by the Imam genotype in a high temperature environment was reported in four field environments in Sudan (Elbashir et al. 2017b). These observations indicating that although the degree of heat impact differed among genotypes, all genotypes showed similar growth trends as a consequence of three days of heat stress. These growth responses were in agreement with those of previous studies. Gupta et al. (2013) observed that heat stress resulted in the reduction of shoot length in wheat seedlings. Another study (Keleş and Öncel, 2002) showed different degrees of reduction in shoot length under high day and night temperatures in different wheat seedlings.

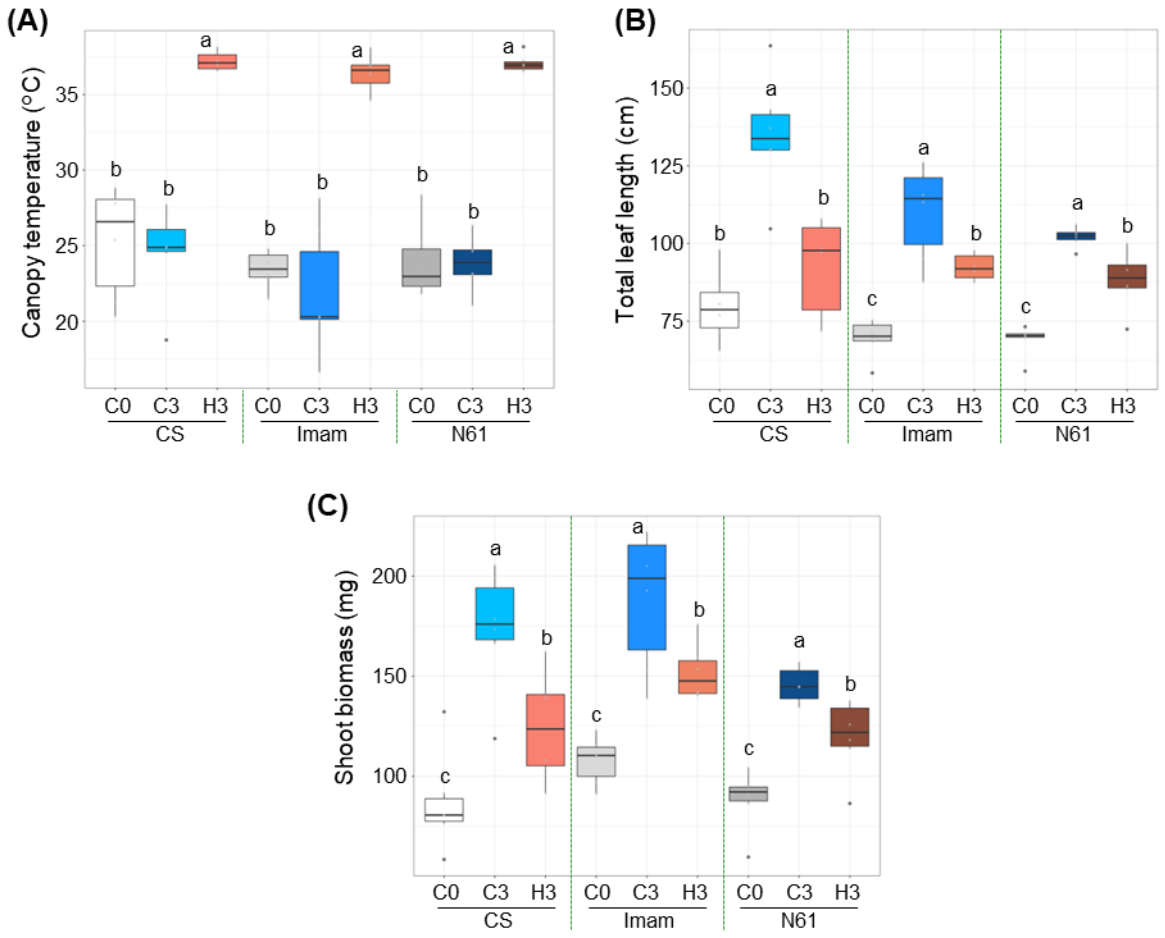

Figure 2-1. Effect of heat stress on the growth of wheat genotypes Chinese Spring (CS), Imam, and Norin 61 (N61). (A) Canopy temperature, (B) total leaf length, and (C) shoot biomass for (C0) before heat treatment, (C3) control plants after three days, and (H3) plants exposed to heat for three days are shown. The six plants each were used for the measurements. One-way ANOVA with post-hoc Tukey HSD test ($p < 0.05$) was carried out for statistical analysis within a given genotype.

Table 2-1. One-way ANOVA with post-hoc Tukey HSD test on canopy temperature.

Category	Pair	Tukey HSD Q statistic	Tukey HSD *p*-value [*1]
CS	C0 vs C3	0.5469	0.7194
	C0 vs H3	11.0873	0.001
	C3 vs H3	14.0014	0.001
Imam	C0 vs C3	1.1828	0.4225
	C0 vs H3	26.2122	0.001
	C3 vs H3	11.2751	0.001
N61	C0 vs C3	0.113	0.85
	C0 vs H3	17.9206	0.001
	C3 vs H3	24.0925	0.001
C0	CS vs Imam	1.662	0.2671
	CS vs N61	1.066	0.4683
	Imam vs N61	0.6218	0.6823
C3	CS vs Imam	1.722	0.2513
	CS vs N61	0.6856	0.6507
	Imam vs N61	1.427	0.3368
H3	CS vs Imam	2.052	0.1775
	CS vs N61	0.5831	0.7014
	Imam vs N61	1.701	0.2567

[1] The p-values <0.5 and >0.5 are labelled with light green and pink colors, respectively

Table 2- 2. One-way ANOVA with post-hoc Tukey HSD test on leaf length.

Category	Pair	Tukey HSD Q statistic	Tukey HSD *p*-value [*1]
CS	C0 vs C3	8.557	0.001
	C0 vs H3	2.236	0.145
	C3 vs H3	5.801	0.0021
Imam	C0 vs C3	8.495	0.001
	C0 vs H3	8.495	0.001
	C3 vs H3	3.845	0.0216
N61	C0 vs C3	19.3	0.001
	C0 vs H3	6.28	0.0013
	C3 vs H3	4.856	0.0064
C0	CS vs Imam	2.743	0.0811
	CS vs N61	3.002	0.06
	Imam vs N61	0.0964	0.85
C3	CS vs Imam	3.465	0.0342
	CS vs N61	5.815	0.0021
	Imam vs N61	1.778	0.2372
H3	CS vs Imam	0.0239	0.85
	CS vs N61	0.7883	0.6
	Imam vs N61	1.404	0.3443

[1] The p-values <0.5 and >0.5 are labelled with light green and pink colors, respectively.

Table 2-3. One-way ANOVA with post-hoc Tukey HSD test on biomass.

Category	Pair	Tukey HSD Q statistic	Tukey HSD *p*-value [*1]
CS	C0 vs C3	0.001	0.001
	C0 vs H3	3.567	0.0302
	C3 vs H3	4.158	0.0148
Imam	C0 vs C3	7.614	0.001
	C0 vs H3	8.352	0.001
	C3 vs H3	3.384	0.0378
N61	C0 vs C3	10.51	0.001
	C0 vs H3	4.279	0.0143
	C3 vs H3	4.233	0.0135
C0	CS vs Imam	2.679	0.0874
	CS vs N61	0.1965	0.8929
	Imam vs N61	0.1965	0.8929
C3	CS vs Imam	1.107	0.4518
	CS vs N61	3.033	0.0576
	Imam vs N61	4.147	0.015
H3	CS vs Imam	3.2572	0.0861
	CS vs N61	0.5501	0.9
	Imam vs N61	3.8073	0.0417

[1] The p-values <0.5 and >0.5 are labelled with light green and pink colors, respectively.

2.3.2. FTIR trace

FTIR spectra were taken from the fully expanded third leaves of C3 and H3 plants of each genotype. The representative spectra are shown in (Figure 2-2). The patterns of these spectra were largely similar; a broad major peak (3100–3600 cm^{-1}) was commonly visible in both the control and heat environments, which can be interpreted as O–H and/or N–H stretching bands (Osman et al. 2022a; Stuart, 2004; Talari et al. 2016; Kamnev et al. 2021). Sharper peaks were detected at wavenumbers of approximately 2960 and 2925 cm^{-1}, which can be assigned to –CH3 and –CH2– antisymmetric signals, respectively. No clear peaks were observed in the 2000–2500 cm^{-1} region. A major peak was observed at approximately 1658 cm^{-1}, which can be attributed to C=C stretching, C=O stretching (amide), and N–H bending (amide I) in proteins in all genotypes in both the control and heat stress environments. All genotypes showed another major peak at approximately 1056 cm^{-1}, which represent signals for C–O stretching, C–N stretching (aliphatic), and in-plane C–H bending (aromatic). All genotypes exhibited another broad peak at approximately 618 cm^{-1}, which can be interpreted as =C–H out-of-plane bending, =C–H bending, or C–S stretching signals. However, as presented in (Figure 2-2), obvious differences between genotypes and environments were not evident to the naked eye, inducing application of further chemometrics analysis to FTIR data, as in the previous studies (Osman et al. 2022a; Ami et al. 2010; Christou et al. 2018; Tarapoulouzi et al. 2020).

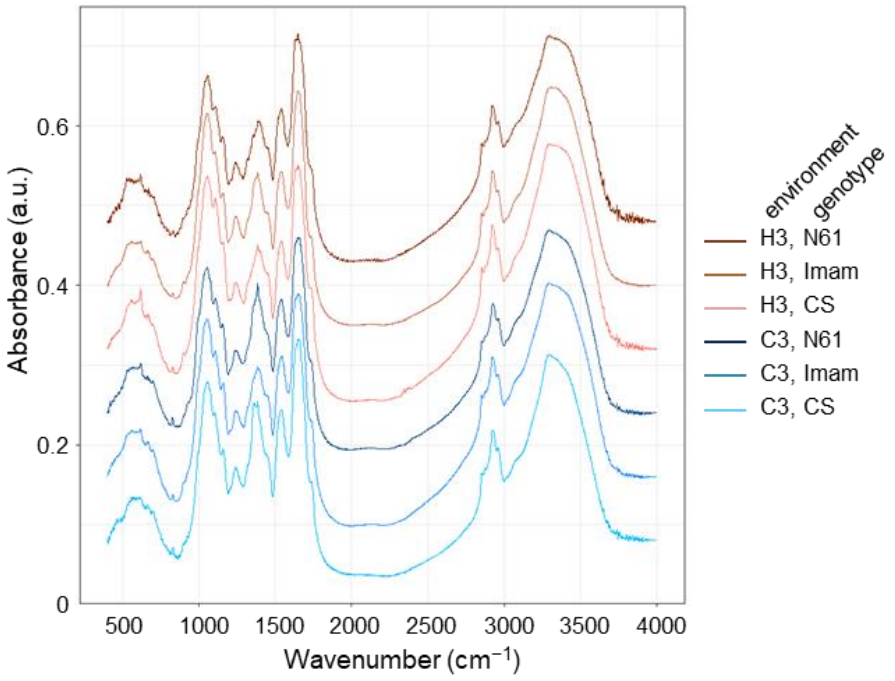

Figure 2-2. Representative FTIR spectra in the leaves of wheat genotypes Chinese Spring (CS), Imam, and Norin 61 (N61). Spectra drawn in blue and red color series represent those for control (C3) and heat stress environments (H3), respectively.

2.3.3. Principal Component Analysis

To characterize the spectral patterns of the three wheat genotypes under the C3 and H3 environments, principal component analysis (PCA) was applied. The PC1–PC2 score plot, which explained 76.2% of total variation (Figure 2-3), exhibited partial separation between genotypes and environments (Figure 2-4). For instance, the C3–CS spectra were mostly clustered in the PC2 negative range from −25 to −50, whereas the H3–CS counterparts tended to position at higher PC2 values. The C3–Imam spectra were widely scattered in the PC1 positive range of +40 to +100, whereas their H3 counterparts were mostly located at lower PC1 values between –20 and +30. The C3–N61 spectra were mostly positioned in the PC1 range between −80 and +20, and PC2 ranged between −10 and +20, while their H3 counterparts were mostly scattered around the −130 to +10 PC1 range and −10 to +45 PC2 range. Their loading plots represent a complex pattern over the entire range of 400–3600 cm^{-1} (Figure 2-3). However, overlapping patterns of different genotypes

and environmental conditions in the PC1–PC2 score plot were also evident, supporting the necessity of applying other chemometrics techniques to characterize their spectral features.

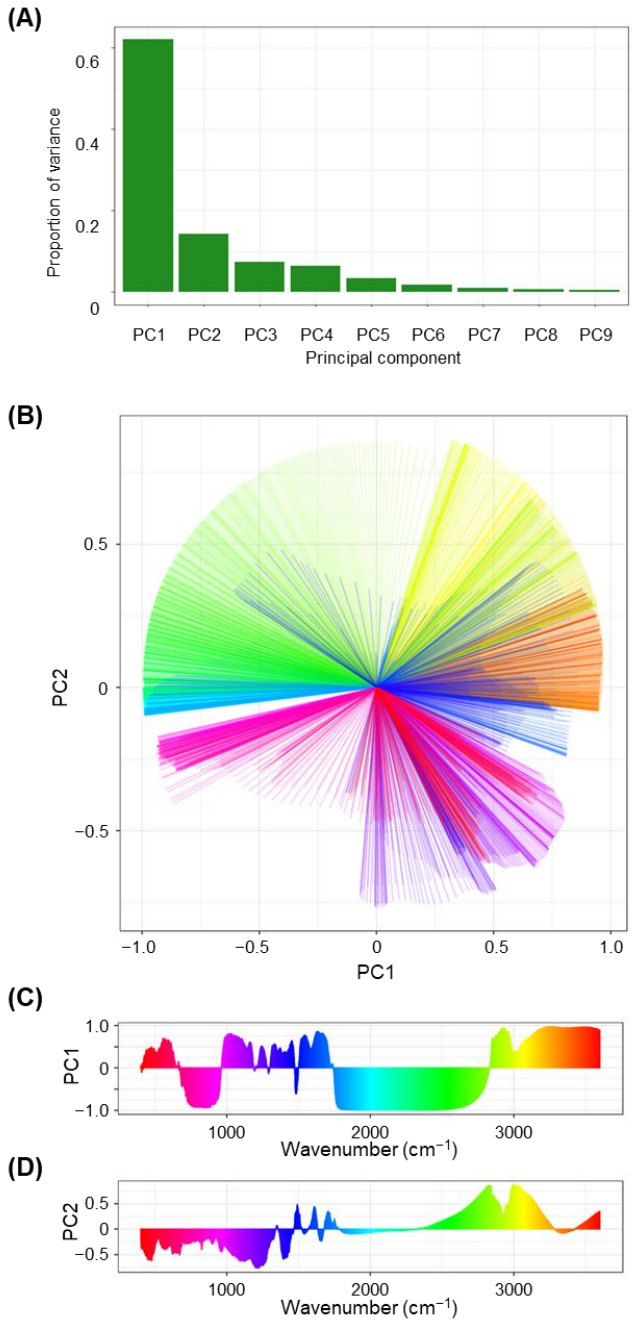

Figure 2-3. PCA of FTIR spectra. (A)Variance explained by the first nine components in PCA. (B) Two-dimensional PC1-PC2 loading plot. (C, D) One dimensional loading plot for (C) PC1 and (D) PC2. The colors of the vectors in panel (B) are the same as those in the panels (C) and (D).

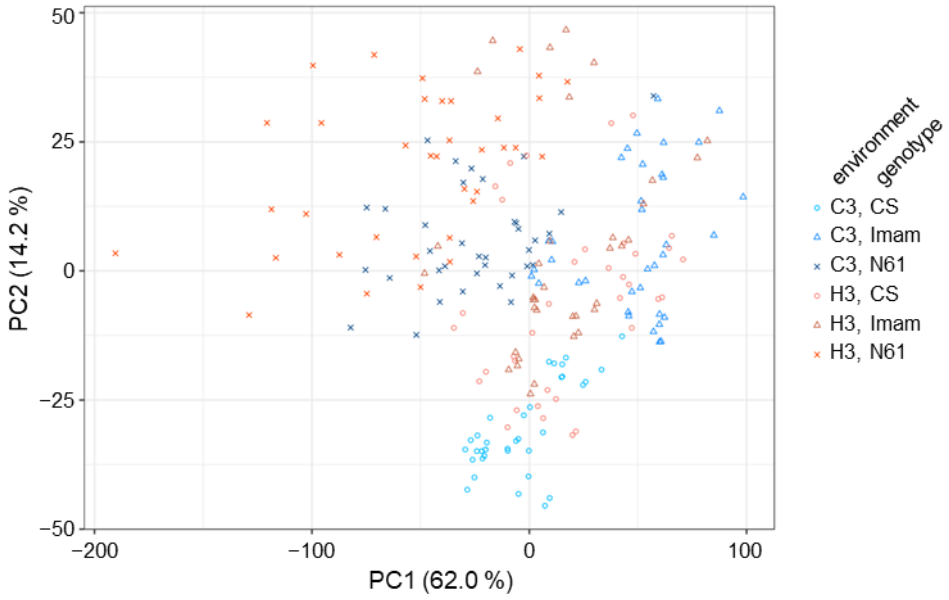

Figure 2-4. Score plot of principal component analysis showing the distribution of FTIR spectra for three wheat genotypes under two environmental conditions. Symbols for three genotypes and two environments are shown at the right of the panel.

2.3.4. Trends of FTIR biomarkers

In the study presented in chapter 1 (Osman et al. 2022a), six FTIR-based biomarkers, Fm482, Fm 576, Fm1251, Fm1465, Fm1502, and Fm1729 (Table 2-4) were generated to distinguish between control and heat-stressed leaves in N61 genotype. These markers were calculated based on the offset absorbance values at specific wavenumbers using the absorbance of the two anchor wavenumbers in the vicinity of the target wavenumber. Applying these markers to the FTIR spectra in this study showed similar responses for some markers between different genotypes (Figure 2-5). Markers Fm482 and Fm1502 were ramped in all genotypes (Figure 2-4, E), suggesting a similar chemical change between these genotypes. The wavenumber 482 cm^{-1}, a target wavenumber for the marker Fm482, was positioned outside of the" fingerprinting region" and was related to methoxy group (472/475 cm^{-1}) (Talari et al. 2016) and S–S stretching (450–550 cm^{-1}) (Stuart, 2004). The latter may be related to the heat-induced of protein disulfide isomerase in a wheat genotype Jing411 (Zhang et al. 2017), which stimulate covalent cross-linking of sulfhydryl groups of cysteine residues, leading to stabilizing structure of the cellular proteins under

heat stress. The Fm1502 was interpreted as lignin (Kurian et al. 2015; Lima et al. 2013), suggesting that physicochemical regulations in the cell wall components may occur under heat stress by same trend in these wheat genotypes, as has been reported in coffee leaves (Lima et al. 2013). Other markers, in contrast, showed contrasting behaviors between genotypes. The Fm1465 marker, which may be linked with suberin/cutin, lipids, and/or cell wall polysaccharides (Osman et al. 2022a; Stuart, 2004; Stewart, 1996; Lammers et al. 2009; Gorgulu et al. 2007), increased under heat stress in CS and N61, but reduced in Imam genotype (Figure 2-5D). The Fm576 marker elevated under heat stress in CS, but decreased in N61, and was statistically unchanged in Imam (Figure 2-5B). Information on the assignments of the wavenumber 576 cm^{-1} to chemical structures has remained relatively rare (Talari et al. 2016), except for carbon–halogen stretching (400–800 cm^{-1}), P=S stretching (500–850 cm^{-1}), and P–Cl stretching (300–600 cm^{-1}) (Stuart, 2004). Though, these observations suggested that biochemical responses to heat stress may be largely different between these genotypes. Interestingly, behaviors of the markers Fm1251 and Fm1729 were contrasting between heat-tolerant and susceptible genotypes (Figure 2-5C, F). The Fm1251 marker, which is related to hemicellulose and/or pectin (Osman et al. 2022a; Stuart, 2004; Mascarenhas et al. 2000), was decreased under heat stress in the heat tolerant Imam and N61 genotypes, while an increase was detected in the heat-sensitive CS genotype (Figure 2-5C). These notifications may suggest that chemical modulations in the extracellular matrix, which potentially work as a regulator for cell wall porosity and heat conductance (Lima et al. 2013; Baron-Epel et al. 1988), are oppositely different between heat tolerant and susceptible genotypes. The Fm1729 marker, which is located in the carbonyl ester region (1720–1760 cm^{-1}) and/or their oxidized derivatives (Lahlali et al. 2014; Osman et al. 2022a; Stuart, 2004; Talari et al. 2016; Kamnev et al. 2021; Sowa et al. 1991; Oleszko et al. 2015), was increased under heat stress in heat tolerant Imam and N61 genotypes, whereas the value was unchanged in heat susceptible CS genotype (Figure 2-5F). This spectral region gives information on the polar interfacial regions of pectin or membrane lipids (Lahlali et al. 2014; Sowa et al. 1991). The latter is similar to previous report showed that heat tolerant and susceptible genotypes showed differential lipidome responses under the stress in wheat (Narayanan et al. 2016). Therefore, the markers Fm1251 and Fm1729 may potentially serve as a tool for discriminating heat tolerant and susceptible wheat genotypes.

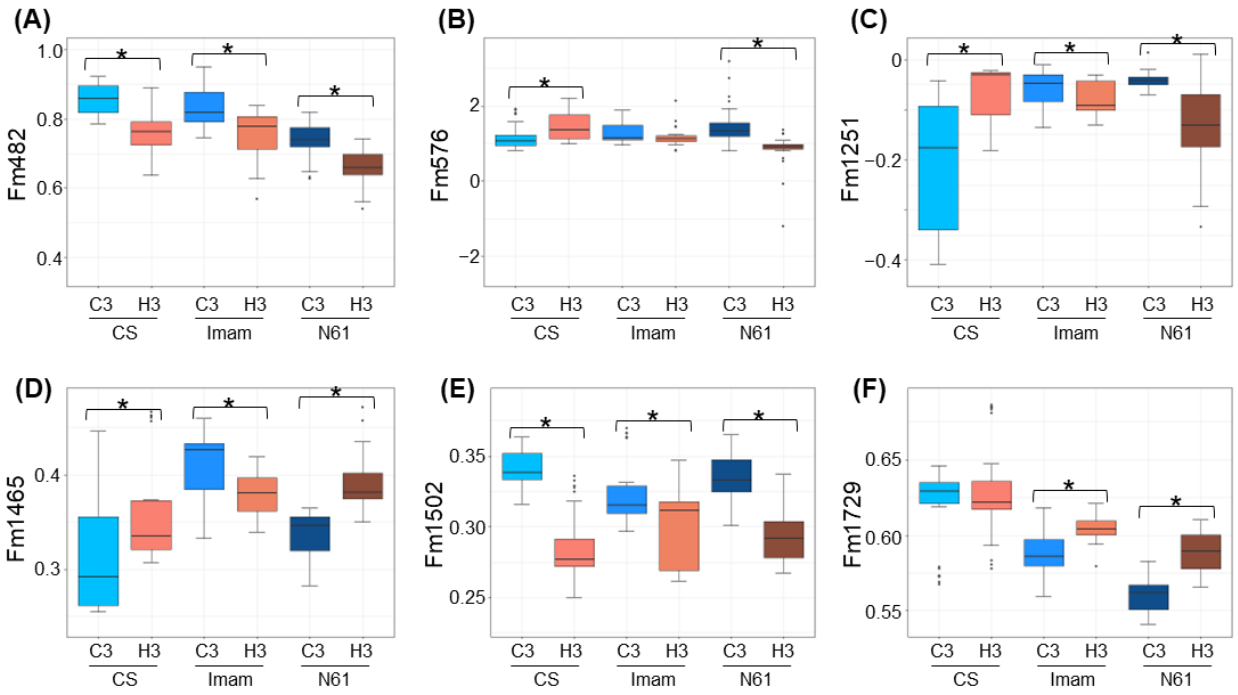

Figure 2-5. Comparison of FTIR-based biomarkers under heat stress between three wheat genotypes tested. Boxplots for (A) Fm482, (B) Fm576, (C) Fm1251, (D) Fm1465, (E) Fm1502, and 254 (F) Fm1729 are shown for CS, Imam, and N61 genotypes under C3 and H3 environments. The 36 spectra were used in each genotype–environment combination. Asterisks represent statistically significant differences at $p < 0.01$.

Table 2-4. Characteristics of spectral markers.

Marker	Wavenumber (cm^{-1})*1			CS			Imam			N61		
	Target	Anchor 1	Anchor 2	Median Fm value		p*3	Median Fm value		p*3	Median Fm value		p*3
				C3 *2	H3 *2		C3 *2	H3 *2		C3 *2	H3 *2	
Fm482	482	401	501	0.8587	0.7629	2.0×10^{-9}	0.8186	0.7788	4.3×10^{-7}	0.7401	0.6601	1.9×10^{-9}
Fm576	576	648	542	1.0621	1.3538	7.1×10^{-5}	1.1481	1.1231	8.8×10^{-2}	1.3327	0.9074	9.0×10^{-8}
Fm1251	1251	1241	1358	-0.1770	-0.0309	1.6×10^{-7}	-0.0473	-0.0915	1.6×10^{-3}	-0.0416	-0.1318	1.0×10^{-7}
Fm1465	1465	1480	1399	0.2921	0.3351	3.8×10^{-3}	0.4268	0.3806	1.9×10^{-4}	0.3463	0.3817	2.5×10^{-13}
Fm1502	1502	1480	1615	0.3381	0.2770	1.2×10^{-18}	0.3151	0.3115	1.8×10^{-4}	0.3328	0.2919	4.4×10^{-15}
Fm1729	1729	1768	1703	0.6291	0.6217	1.8×10^{-1}	0.5861	0.6040	1.1×10^{-6}	0.5615	0.5895	2.0×10^{-16}

[1] Wavenumbers for the target and the flanking two anchors, that were used for the Fm value calculation as described previously (Osman et al.., 2022a).
[2] Fm value in either C3 (control day 3) or H3 (heat day 3) condition.
[3] Significance between C3 and H3 conditions by *t*-test.

2.3.5. Linear Discriminant Analysis

Linear discriminant analysis (LDA) was performed to further characterize the FTIR spectral differences between the different genotypes. The FTIR spectra consisting of six classes (three genotypes × two environments) were used to create the LDA model. The proportion of trace values for the resultant five discriminant functions (LDs) declared that the first two LDs (LD1 and LD2) accounted for 48.2 and 24.5% of total variance, respectively (Figure 2-6). The first two LDs were used to draw a graphical distribution of each spectrum, which exhibited six distinct clusters for each 'genotype × environment' class in the scatter plot (Figure 2-7). The plot presented a typical feature of LDA, which maximizes between class variance while minimizing within class variance (Xanthopoulos et al. 2013; Harrison et al. 2018). The following features were extrapolated from the LD1–LD2 scatter plot: (i) Classes for the same genotypes were located close to each other, for example, two classes (C3 and H3) for the Imam genotype were positioned in the region spanning from –28 to 6 for LD1 and from – 42 to –21 for LD2 coordinates (Figure 2-7), which may suggest the presence of particular genotype features in the FTIR spectra. The ability of FTIR to distinguish between different genotypes is not unique to this study, and has been reported previously in many studies, including grapevine genotypes (Álvarez et al. 2020) and geographical classification of coffee (Bona et al. 2017). (ii) In all genotypes, clusters for the heat stress environment were located in higher LD1 ranges in comparison to their control counterparts, indicating that LD1 may be linked with the presence/absence of heat stress. (iii) In CS and N61 genotypes, the LD2 values for heat stress clusters were moved downward from their control counterparts, whereas an opposite upward shift of the heat cluster was detected in Imam genotype, suggesting that LD2 may partially reflect genotype-specific heat responses.

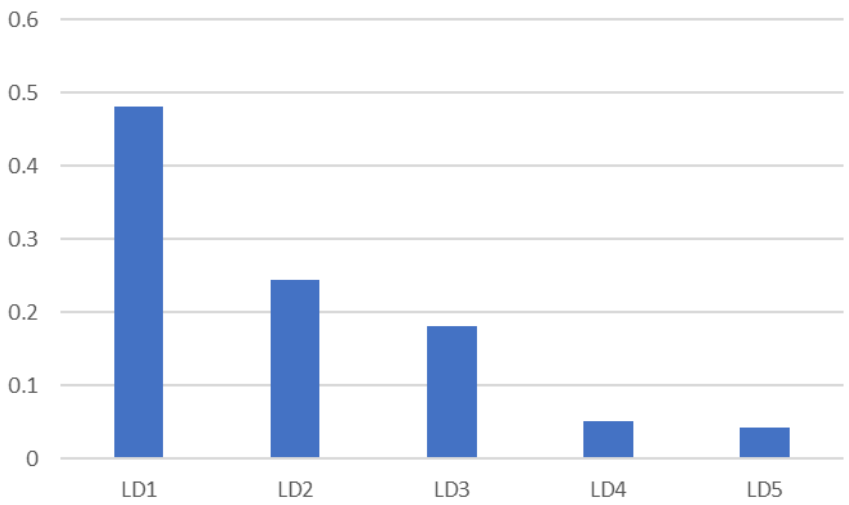

Figure 2-6. Values for proportions of trace linear discriminant analysis.

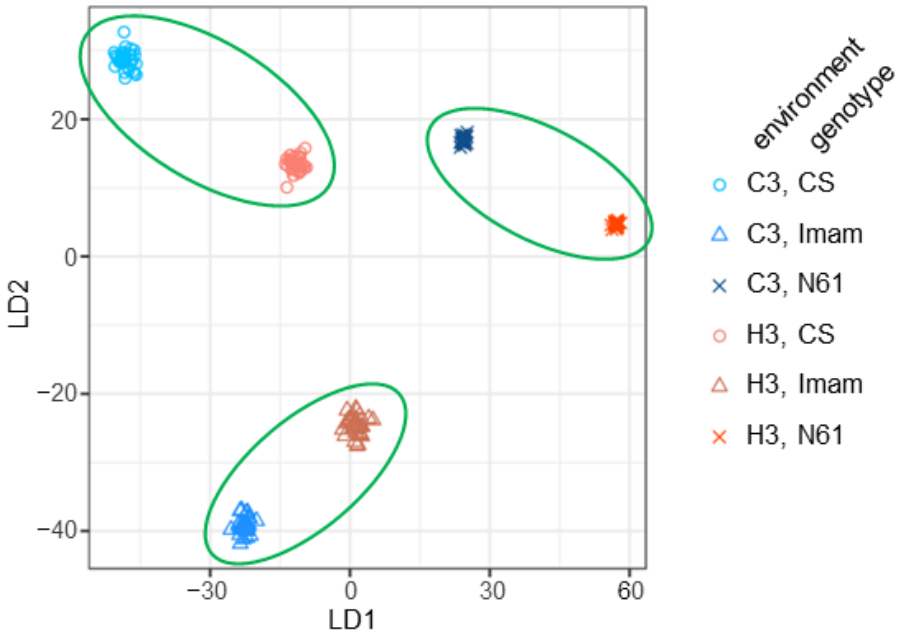

Figure 2-7. Scatter plot of LD1 and LD2 derived from the linear discriminant analysis (LDA). The LDs are the discriminate functions of the LDA model. Each point represents the LD1–LD2 coordinate for each FTIR spectrum. Symbols for three genotypes and two environmental conditions are shown at the right of the panel. Green ellipses denote the location of each wheat genotype.

To gather more information on the influential wavenumber regions in the FTIR spectra for discriminating different genotypes and environmental conditions, the coefficients of the LDs for each wavenumber were studied. The coefficient of LDs, or scaling value, indicates the weight or contribution of each wavenumber to the LD function in such a way that higher absolute values of coefficients potentially show a greater degree of contribution to the discrimination (Setser et al. 2018). A two-dimensional scatter plot of the coefficients presented that most of the wavenumbers were weakly clustered in the origin of LD1–LD2 plain, whereas considerable numbers of 'characteristic' wavenumbers were deviated from the center (Figure 2-8A). One-dimensional plots of coefficients for either LD1 or LD2 versus wavenumbers showed that several spectral regions, i.e., 400–800 cm^{-1}, 1200–1300 cm^{-1} 295 1, 1450–1550 cm^{-1}, and 1700–1800 cm^{-1} regions, had strong absolute values for either LD1 or LD2 (Figure 2-8B, C), indicating that these regions may have major participation to the spectral discrimination of different genotypes and environmental conditions. These regions contained the 600–1500 cm^{-1} region that has been called "fingerprinting" region in which infrared absorption are cumulatively influenced by small steric or electronic effects based on the nature of the molecules (Stuart, 2004). The most 'characteristic' wavenumbers that deviated from the origin of the LD1–LD2 plain (Figure 2-8A) were predominantly positioned in the ranges of 400–500 cm^{-1} and 1200–1300 cm^{-1}. Among these two regions, the 400–500 cm^{-1} range was located in the ascending curve from the spectral margin of 400 cm^{-1} which was exploited for baseline correction, thus showing less absorbance values (hence, should have a lower contribution for discrimination). Moreover, this region suffers from a jaggy shape of spectral curve, which was probably caused by noise (Figure 2-9). Therefore, in the subsequent analysis, we focused on the 1200–1300 cm^{-1} region, which is within the fingerprinting region and rich by signals from multiple functional groups, such as C–O stretching, in-plane C–H bending (aromatic), and aliphatic C–O stretching (Stuart, 2004; Talari et al. 2016).

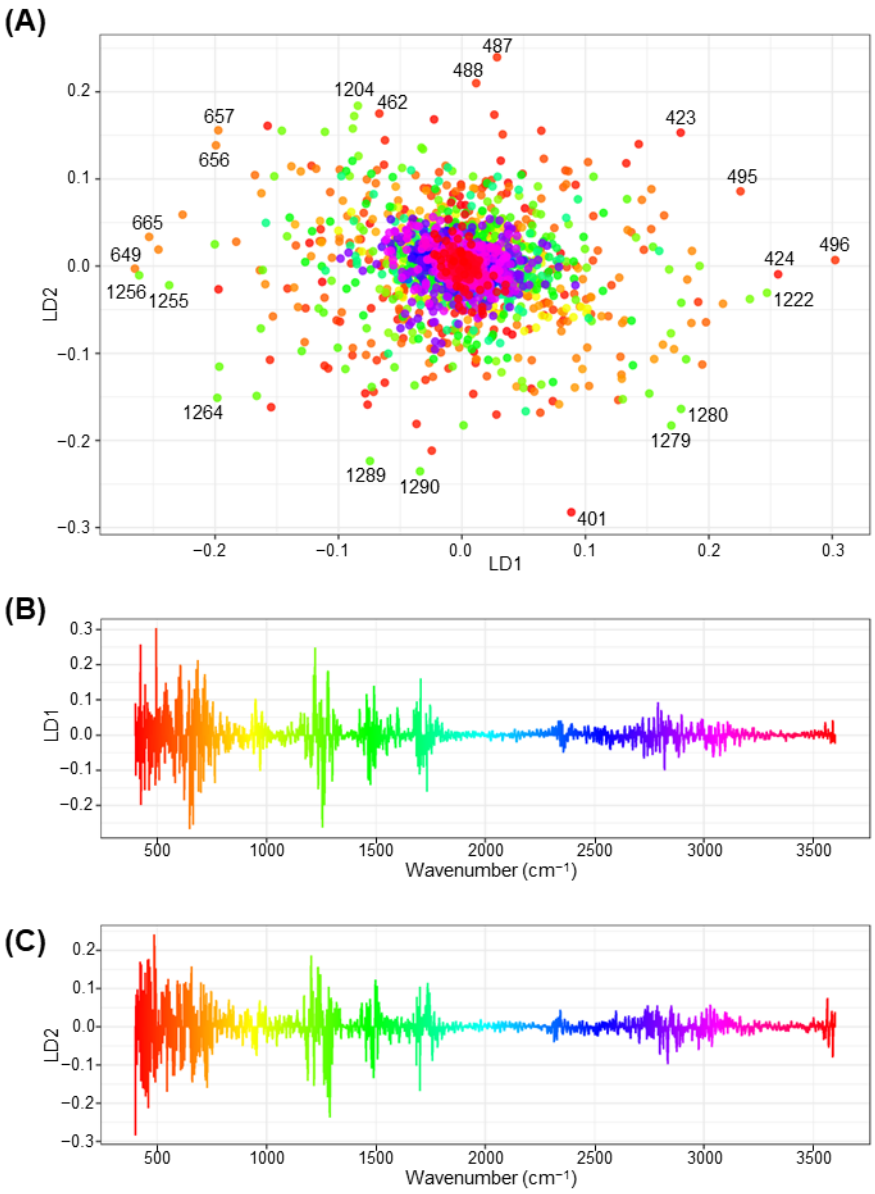

Figure 2-8. Relationships between the coefficients of the first two linear discriminants (LD1 and LD2) and wavenumbers in linear discriminant analysis (LDA). (A) Two-dimensional scatter plot for the coefficients of linear discriminant LD1 and LD2. Respective points represent the wavenumbers from 400 to 4000 cm^{-1}. Assignment of a color gradient to respective wavenumbers are the same with those presented in (B, C). Numbers in black font in the vicinity of respective color points designate wavenumbers for characteristic data points with higher absolute coefficient values. (B, C) One-dimensional column plots showing relationships between wavenumbers and coefficients of (B) LD1 and (C) LD2. The coefficient values for each wavenumber are expressed using a rainbow color gradient along their x-axes.

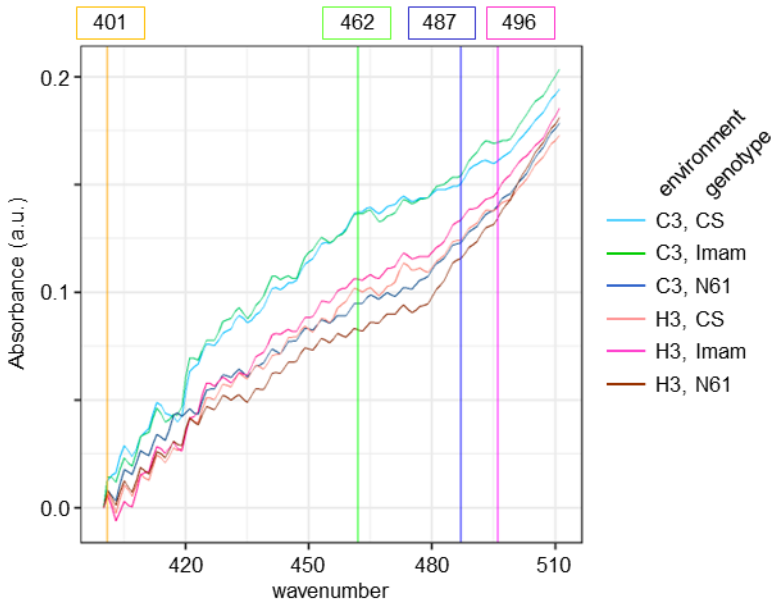

Figure 2-9. Magnified view of averaged FTIR spectra in the wavenumber ranges from 400 to 510 cm^{-1} that detected strong discriminatory variable wavenumbers in LDA. Averaged spectra for 6 genotypes × environment combinations were drawn as the color legend as depicted in the right of the panel. Colored vertical straight lines and their numbers on top of the panel denote the characteristic wavenumbers that were detected as strong discriminatory variable in LDA.

To shed light into the features of the 1200–1300 cm^{-1} region, averaged FTIR spectral curves for this region were compared among genotypes and environments (Figure 2-10). This region consists of four characteristic wavenumbers: 1222 cm^{-1} (LD1 coefficient of +0.247), 1256 cm^{-1} (LD1 coefficient of –0.262), 1204 (LD2 coefficient of +0.184), and 1290 cm^{-1} (LD2 coefficient of –0.235) (Figure 2-8A). The averaged FTIR curves revealed that the wavenumbers 1222 cm^{-1} and 1256 cm^{-1} were situated in the middle of the ascending and descending curves, respectively, to/from a peak centered at 1241 cm^{-1}, which has been tentatively interpreted as C–O stretching, in-plane C–H bending (aromatic), and aliphatic C–O stretching signals (Osman et al. 2022a). The absorbance values at these wavenumbers for the six environments × genotype combinations were in the descending order of C3–CS, C3–Imam, H3–CS, H3–Imam, C3–N61, and H3–N61 (Figure 2-7), which showed link with the ascending order of LD1 scores in LDA (Figure 2-7). This was consistent with the negative LD1 coefficient value for the wavenumber of 1256 cm^{-1} but showed contrasting trend with the positive LD1 coefficient for the wavenumber 1222 cm^{-1}. Although the reason for this discrepancy is currently unknown, one possibility for the 1222 cm^{-1} variable may

counteract the 1256 cm^{-1} variable to minimize the within-class variance in the LDA scores, which is a basic characteristic of LDA (Xanthopoulos et al. 2013; Harrison et al. 2018). Another possible explanation is that each wavenumber may exert rather small effects, and cumulative actions of multiple wavenumbers would be required for the final discrimination. Similar phenomenon were observed for the wavenumbers 1204 cm^{-1} (positive LD2 coefficient) and 1290 cm^{-1} (negative LD2 coefficient), in which the absorbance values were only partially correlated with the LD2 score (Figure 2-8 and Figure 2-10). In a conclusion, these observations suggest that further studies are required to fully elucidate the spectral behavior and underlying biochemical changes during heat stress in a variety of wheat genotypes.

In the current study, utility of FTIR-based chemical fingerprinting in association with a chemometrics was demonstrated, for characterizing metabolome responses to heat stress in the three genotypes of bread wheat with different heat tolerance. Expanding this technique to other types of climate change-related environmental stress responses in various genotypes will be expected in the future studies, such as drought stress and drought/heat combination, which led to large reduction of wheat yield worldwide (Zampieri et al. 2017; Qaseem et al. 2019).

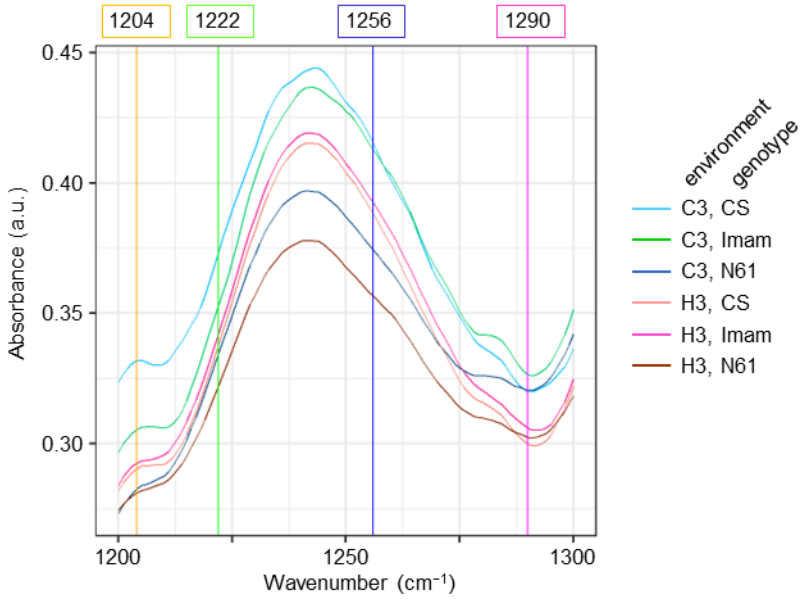

Figure 2-10. Magnified view of averaged FTIR spectra in the wavenumber ranges from 1200 to 1300 cm^{-1} for what strong discriminatory variable wavenumbers were detected in LDA. Averaged spectra for six wheat genotypes × environment combinations are shown, according to the legend on the right side. Colored vertical straight lines and their numbers on top of the panel denote the characteristic wavenumbers that were detected as strong discriminatory variables in LDA.

The PCA, spectral biomarker assays, and LDA of FTIR spectra declared the existence of common and distinct metabolic responses between the three genotypes of bread wheat with different heat tolerance. The spectral biomarker assay showed that Fm1251 and Fm1729 markers potentially distinguish heat tolerant and susceptible genotypes, suggesting that these markers may work as a selection tool for heat-tolerant genotypes. Analysis of the coefficient values in LDA indicated the presence of potential discriminatory spectral regions that were associated with genotype specific metabolic responses. The current study demonstrates the versatility and potential of the FTIR fingerprinting technique for illustrating the diversity of metabolic behaviors among diverse plants.

Summary of the study

Wheat (*Triticum aestivum* L.) is one of the most important crops globally. It contributes with rice, maize, and soybean to two-thirds of calories required for world population. Wheat is very sensitive to heat stress. An increase of 1°C temperature is estimated to reduce wheat yield by 6.0%. Therefore, understanding wheat response to heat stress is crucial for facilitating the development of new heat tolerant varieties.

In this study, metabolomic approach was chosen because it is arguably more closely related to the phenotypes than other "omics" data. Metabolomics is one of the omics tools used to analyze the molecular responses of plants, and has been utilized to study metabolic responses in plants under various stresses and genotypes differentiation. There are various tools to study plants metabolome. Among them Fourier transform infrared spectroscopy (FTIR) spectroscopy is unique in that it provides an opportunity to study biological samples *in vivo* in a non-destructive manner, is compatible with remote sensing in the field, and allows the analysis of complex biomacromolecules such as cell wall components. To our knowledge FTIR was not applied before to examine heat stress effects in wheat metabolome.

Therefore, sequences of studies were carried out starting by establishing a protocol to detect FTIR capability to characterize chemical changes of wheat metabolome under heat stress (chapter1) utilizing a genotype Norin 61 (N61). Subsequently, the established protocol was applied to wheat genotypes possessing different heat tolerance capabilities (chapter 2). Three genotypes were used in this study: Chinese Spring (CS) has been identified as a heat-sensitive genotype. Imam is a heat-tolerant cultivar widely grown in Sudan, which is regarded as the world's hottest wheat growing environment. N61 showed heat tolerance in hot regions in Sudan in field studies.

In the present studies, plants were grown in normal condition in control chamber in 18◦C for the night temperature for 10 h and the daily temperature of 22°c. Heat stress was applied when the plants reached the three-leaf stage and the length of the third leaf exceeded that of the second leaf. Under the heat stress condition, the seedlings were transferred to a heat chamber with a daily maximum temperature of 42°C. The obtained FTIR spectra from the leaves did not show visually-prominent discriminating peaks between heat stress and control conditions. Therefore, coupling the FTIR analysis with chemometric analysis was indispensable.

In the first chapter, visual inspection of FTIR spectra and their principal component analysis showed partially overlapping features between heat-stressed and control leaves in N61 genotype. In contrast, supervised machine learning through linear discriminant analysis (LDA) of the spectra demonstrated clear discrimination of heat-stressed leaves from the controls. Analysis of LDA loading suggested that several wavenumbers in the fingerprinting region (400–1800 cm^{-1}) contributed significantly to their discrimination. Six novel spectrum-based biomarkers, designated as Fm482, Fm576, Fm1251, Fm1465, Fm1502, and Fm1729, were developed using these discriminative wavenumbers, which enabled successful diagnosis of heat-stressed leaves.

In chapter 2, the metabolome responses of heat-tolerant genotypes, Imam and N61, and susceptible genotype CS were comparatively analyzed using FTIR in combination with chemometric data mining techniques. Similar to the chapter 1, principal component analysis of the FTIR data showed partially overlapping spectral feature between the three genotypes. However, the six FTIR-based markers developed in the study presented in chapter 1, together with LDA data detected contrasting metabolome behaviors between the three genotypes, demonstrating the capacity of FTIR-chemometrics approach in differentiating genotypes, environment, and their combination thereof.

The FTIR-chemometrics described above showed a wide range of metabolome changes in wheat leaves under heat stress; some of them were commonly observed in three wheat genotypes, while others were genotype-specific. The former example includes the markers Fm482 and Fm1502, which were reduced in all genotypes, indicating similar chemical response between these genotypes. Wavenumber 482 cm^{-1}, a target wavenumber for the marker Fm482, was positioned outside the" fingerprinting region" and was related to a methoxy group (472/475 cm^{-1}) and S–S stretching (450–550 cm^{-1}). The latter annotation may be related to a previously reported heat-induced protein disulfide isomerase, which promotes covalent cross-linking of sulfhydryl groups of cysteine residues, leading to stabilization of the structure of cellular proteins under heat stress. The Fm1502 marker is potentially annotated to lignin, suggesting that physicochemical modifications in cell wall compositions may occur under heat stress in these wheat genotypes.

This study identified several FTIR markers that showed differential behaviors between genotypes. The Fm1465 marker, which may be associated with suberin/cutin, lipids, and/or cell wall polysaccharides, elevated under heat stress in CS and N61, but reduced in the Imam genotype.

The Fm576 marker increased under heat stress in CS, but decreased in N61, and was statistically unchanged in Imam. No sufficient information on the assignment of the wavenumber 576 cm^{-1} to chemical structures are available, except for carbon halogen stretching (400–800 cm^{-1}), P=S stretching (500–850 cm^{-1}), and P–Cl stretching (300–600 cm^{-1}). These observations suggested that biochemical responses to heat stress may be largely different between these genotypes.

It noteworthy that, the markers Fm1251 and Fm1729 showed different responses between heat-tolerant and -susceptible genotypes. The Fm1251 marker, which is related to hemicellulose and/or pectin, decreased under heat stress in the heat-tolerant Imam and N61 genotypes, while it increased in the heat-sensitive CS genotype. Those signs may indicate that chemical modification in the extracellular matrix, which potentially functions as a controller for cell wall porosity and heat conductance, are contrastingly different between heat-tolerant and susceptible genotypes. The Fm1729 marker, which is located in the carbonyl ester region (1720–1760 cm^{-1}) and/or its oxidized derivatives, was increased under heat stress in heat-tolerant Imam and N61 genotypes, whereas the value was unchanged in the heat-susceptible CS genotype. This spectral region provides information on the polar interfacial regions of pectin or membrane lipids. Thus, the markers Fm1251 and Fm1729 may potentially serve as tools for distinguishing heat-tolerant and susceptible wheat genotypes.

Overall, in the present study, an FTIR-based fingerprint technique was applied to characterize the metabolome response of wheat leaves to heat stress. Application of chemometrics techniques to the FTIR spectral data, especially the LDA technique, revealed specific spectral regions that may reflect metabolome changes in wheat leaves under heat stress. Several spectral biomarkers were developed that correctly reflected the heat-stress status of the leaves. Application of the developed markers to wheat genotypes with different heat tolerance abilities showed common and differential metabolomic response among genotypes. Among these biomarkers; Fm1251 and Fm1729 markers potentially discriminate heat-tolerant and -susceptible genotypes, suggesting that these markers may serve as a selection tool for heat-tolerant genotypes. Overall, the present study suggests the potential of FTIR spectroscopy, coupled with chemometrics analysis, for studying the heat-stress response and tolerance mechanisms in wheat.

Japanese Summary of the study

コムギ（*Triticum aestivum* L.）は、世界的に最も重要な作物の一つである。米、トウモロコシ、大豆とともに、世界人口に必要なカロリーの 3 分の 2 を占めている。コムギは熱ストレスに非常に敏感である。気温が 1℃上昇すると、コムギの収量は 6.0%減少すると推定されている。したがって、コムギの熱ストレスに対する応答を理解することは、新しい耐熱性品種の開発を促進するために非常に重要である。

本研究では、他のオミックスデータと比較して、より表現型に近いと考えられるメタボロミクスによる研究アプローチを選択することにした。メタボロミクスは植物の分子応答を解析するためのオミクスツールの一つであり、様々なストレスや遺伝子型の違いによる植物の代謝応答を研究するために利用されてきた。植物のメタボロームを研究するための手法には様々なものがある。その中でもフーリエ変換赤外分光法（FTIR）は、生体試料を非破壊で調べられること、野外でのリモートセンシングに対応できること、細胞壁成分などの複雑な生体高分子の分析が可能であることが特徴である。私たちの知る限り、コムギのメタボロームにおける熱ストレスの影響を調べるために FTIR が適用されたことはこれまで報告されていなかった。

そこで本研究では、熱ストレス下におけるコムギのメタボロームの化学変化を FTIR で検出する実験手法を、人工気象器での実験による環境制御下で農林 61 系統 (N61)を用いて確立するための一連の実験を行った（第 1 章）。次に、耐暑性の異なるコムギ品種に本実験手法を適用した（第 2 章）。本研究では、3 種類のコムギ系統を用いた。Chinese Spring (CS)は高温感受性の遺伝子型として知られているものである。Imam は、世界で最も暑いコムギ栽培環境とされるスーダンで広く栽培されている耐暑性品種である。N61 は、スーダンの高温乾燥地帯における圃場試験で耐暑性を示した系統である。

本研究において，植物はまず人工気象器内で夜温 18℃で 10 時間、そして日中の温度 22℃の通常状態で栽培された。植物が 3 葉の段階に達し、3 葉の長さが 2 葉の長さを超えたときに、苗を日最高気温が 42℃の高温の人工気象器に移すことで熱ストレスを付与した。得られた葉の FTIR スペクトルには、熱ストレス条件と対照条件との間で視覚的には顕著なスペクトル差は確認されなかったため、FTIR 分析とケモメトリックス分析を組み合わせることが不可欠であった。

第 1 章では、FTIR スペクトルの目視での検査とその主成分分析では、N61 系統の熱ストレス葉と対照葉の間に部分的に重複した特徴があることが示された。一方、線形判別分析(LDA)による教師ありの機械学習では、熱ストレス葉と対照葉の間で明確な判別が示された。LDA の負

荷の分析により、指紋領域（400-1800 cm^{-1}）内のいくつかの波長が、その識別に大きく寄与していることが示唆された。これらの波長を用いて、Fm482, Fm576, Fm1251, Fm1465, Fm1502, Fm1729 と名付けた 6 つの新規スペクトルベースのバイオマーカーを開発し、熱ストレス葉を診断することに成功した。

第 2 章では、耐暑性系統である Imam と N61、および感受性系統である CS のメタボローム応答を、FTIR とケモメトリックデータマイニング技術を併用して比較解析した。第 1 章と同様に、FTIR データの主成分分析では、3 つの遺伝子型の間でスペクトルの特徴が一部重複していることが示された。しかし、第 1 章で開発した 6 つの FTIR ベースマーカーと LDA データを組み合わせることで、3 つの遺伝子型間で対照的なメタボロームの挙動が検出され、遺伝子型、環境、およびそれらの組み合わせを識別する FTIR-ケモメトリックスアプローチの能力が実証された。

以上のように、FTIR とケモメトリックスは、熱ストレス下のコムギの葉において、3 種の系統に共通するメタボローム挙動に加え、各系統に特異的な挙動も存在することを明らかにした。前者の例としては、Fm482 と Fm1502 というマーカーがあり、これらはすべての系統において減少しており、これらの系統で共通して同種の化学反応が起こっていることが示唆された。このうち Fm482 のターゲット波数である 482cm^{-1} は、指紋領域の外に位置し、メトキシ基（472/475 cm^{-1}）または S-S 結合の伸縮（450-550 cm^{-1}）に関連することが示唆された。後者の注釈は、以前に報告されたタンパク質ジスルフィドイソメラーゼの高温下での誘導に関連していると考えられ、システイン残基のスルフヒドリル基の共有結合による架橋を促進し、熱ストレス下で細胞内タンパク質の構造の安定化に寄与する可能性が考えられた。Fm1502 マーカーはリグニンの化学挙動と関連する可能性があり、これらのコムギ系統において熱ストレス下で細胞壁組成の物理化学的変化が起こっている可能性が示唆された。

本研究では、コムギ系統間で異なる挙動を示すいくつかの FTIR マーカーが同定された。Fm1465 マーカーは、suberin/cutin、脂質、細胞壁多糖類などに関連すると考えられ、CS と N61 では熱ストレス下で上昇したが、Imam 遺伝子型では低下した。Fm576 マーカーは熱ストレス下で CS では増加したが、N61 では減少し、Imam では統計的に変化しなかった。波数 576 cm^{-1} の化学構造へのアノテーションについては、炭素ハロゲン伸縮 (400-800 cm^{-1})、P=S 伸縮 (500-850 cm^{-1})、P-Cl 伸縮 (300-600 cm^{-1}) を除いて、十分な情報が得られていない。しかし、これらの観察から、熱ストレスに対するコムギの生化学的応答は、これらの系統間で大きく異なる可能性が示唆された。

また，Fm1251 マーカーと Fm1729 マーカーでは，耐暑性遺伝子と耐暑性遺伝子の間で異なる応答を示した。ヘミセルロースやペクチンに関連するマーカーである Fm1251 は、耐熱性の

Imam や N61 では熱ストレス下で減少したが、耐熱性の CS では増加した。 これらの結果は、細胞壁の空隙率や熱伝導率の制御因子として機能する可能性のある細胞外マトリクスの化学修飾が、耐熱性遺伝子型と感受性遺伝子型で対照的に異なることを示している可能性がある。カルボニルエステル領域（1720-1760cm^{-1}）およびその酸化誘導体のシグナル領域に存在する Fm1729 マーカーは、耐熱性 Imam および N61 遺伝子型では熱ストレス下で増加したが、耐熱性 CS 遺伝子型では値は変化しなかった。このスペクトル領域は、ペクチンや膜脂質の極性界面領域の情報を提供する。したがって、マーカーFm1251 と Fm1729 は、耐熱性コムギと感受性コムギの遺伝子型を区別するためのツールとなる可能性がある。

総括すると、本研究では、熱ストレスに対するコムギ葉のメタボローム応答を特徴付けるために、FTIR を基盤とした識別技術を適用した。FTIR スペクトルデータに対して、特に LDA 技術に代表されるケモメトリクス手法を適用することで、高温ストレス下のコムギ葉内で起こっている代謝変動を反映するスペクトル領域を明らかにすることができた。また、葉の熱ストレス状態を反映する複数のスペクトル・バイオバイオマーカーが開発された。開発されたバイオマーカーを耐熱性の異なるコムギの系統群に適用したところ、系統間で共通するメタボローム応答と、異なる応答が見られた。開発した 6 種類のバイオマーカーのうち、Fm1251 と Fm1729 は耐熱性・感受性を識別している可能性があり、これらのマーカーは耐熱性遺伝子の選抜ツールになる可能性が示唆された。本研究は、FTIR 分光法とケモメトリックス解析を組み合わせた手法が、コムギの熱ストレス応答および耐性メカニズムを研究する上で潜在的価値が高いことを示唆するものである。

Appendix-1

R-scripts for the processing of FT-IR data

Script code 1: FTIR-spectra processing

```
#salma_a2_spec_processing_ftir_211213.r
#import necessary libraries
library(conflicted)
library(dplyr)
library(ggplot2)
library(readr)
#clean up the R's brain
rm(list=ls())
#obtain date information
today <- Sys.Date()
yr <- substr(today, 3,4)
mo <- substr(today, 6,7)
day <- substr(today, 9,10)
today2 <- paste(yr, mo, day, sep="")
#obtain desktop folder information for the windows user
#change the string within "xxx" below according to your computer
desktopfolder <- "akash"
#create column names for output dataframe
wnlist <- seq(4000, 400, length=3601)
columnname <- c("filename", "condition", "genotype", "identifier", wnlist)
#name the column labels for spec data
specpile <- as.data.frame(t(columnname))
names(specpile) <- columnname
specpile <- slice(specpile, -1)
#set working directory
#setwd("C:/Users/akash/desktop/inputfolder")
#input data from .asc file that are generated by Perkin-Elmer
#obtain the filename
```

```
#obtain the list of filenames for all csv files,
 #which are transiently stored in "ftir_spec_input" folder in your desktop
pathname_inputfolder <- paste("C:/Users/",desktopfolder, "/desktop/", "ftir_spec_input",
sep="")
filelist <- list.files(path = pathname_inputfolder,
 pattern = "*.asc",
 full.names = T)
#count the number of files
fileno <- length(filelist)
#starting a loop for processing data
for (i in 1:fileno){
#obtain the new filename
 filename <- basename(filelist(i))
#obtain dataframe
#skip first 25 lines
#the 26th line does not have variable names
 rawspec <- read.table(filelist(i), skip = 25)
#quick summary
# summary(rawspec)

#plot the spectrum
# ggplot(rawspec, aes(x = V1,y = V2)) +
# geom_point()

#save the wn column for later plotting
 wn_column <- dplyr::select(rawspec, V1)

#exchange rows and columns
#(optional)keep the type as data.frame
 rawspec2 <- as.data.frame(t(rawspec))

#split the rows into wn and spec
 wn_axis <- as.data.frame(rawspec2(1,))
```

```
rawspec3 <- as.data.frame(rawspec2(2,))

#name the column labels for spec data
names(rawspec3) <- wn_axis
#smoothing of the spectrum trace
#below to fill in, but currently skip it

#obtain the baseline anchors
#this is a version to take only 4000 and 400
#the relationships between wn and column_no. is
#column_no = -wn +4001
raw400 <- rawspec3(1,3601)
raw4000 <- rawspec3(1,1)

#create a baseline data
#following is the 1st version
#line is drawn between 4000 to 400
baseline <- seq(raw4000, raw400, length=3601)
#subtract the baseline
spec4_baselined <- (rawspec3 - baseline)

#draw the baseline-corrected spectrum
#1st, exchange the rows and columns
spec4_tall <- t(spec4_baselined)
#combine the wn and spec columns
spec4_tall <- cbind(wn_column, spec4_tall)
#plot the baseline-corrected spectrum
#ggplot(spec4_tall, aes(x = V1,y = V2)) +
# geom_point(size=0.3)
#normalization of spec
#1st, sum of current spec is calculated
sum_signal_original <- sum(select(spec4_tall, V2))
#2nd, new column is generated in the spec
```

```
#spec values in ppm is calculated
spec5_tall <- dplyr::mutate(spec4_tall, ABS = V2*1000000/sum_signal_original)
#draw the normalized spectrum
#ggplot(spec5_tall, aes(x = V1,y = ABS)) +
# geom_point(size=0.3)
#row-column conversion
spec5 <- as.data.frame(t(spec5_tall))

#remove original data from spec5
spec6 <- dplyr::slice(spec5, 3)
#rownames(spec6) <- filename

#create one column at the top
#add dataname to the 1st column
spec7 <- mutate(spec6, dataname=filename, .before="4000")

#judge the treatment condition
#and add to the 2nd column
condition_id <- substring(filename, 1, 2)
spec8 <- mutate(spec7, condition=condition_id, .after="dataname")
#judge the genotype
#and add to the 3rd column
genotype_id <- substring(filename, 3, 5)
spec9 <- mutate(spec8, genotype=genotype_id, .after="condition")

#setup the identifier for later analyses
#and add to the 4rd column
identifier_column <- substring(filename, 1, 5)
spec10 <- mutate(spec9, identifier=identifier_column, .after="genotype")

#compiling the data
specpile <- rbind(specpile, spec10)
}
```

```
#export the data as csv
#data is baseline-corrected, normalized spec
filename_specpile_processed <- paste(today2, "_", "specpile_processed.csv", sep="")
filename2_specpile_processed <- paste("C:/users/",desktopfolder,"/desktop/",
filename_specpile_processed, sep="")
write.csv(specpile,
 filename2_specpile_processed, row.names=FALSE)
#prepare long-format as well, and export
#row-column conversion
long_specpile <- as.data.frame(t(specpile))
#create new column at the top
long_specpile <- mutate(long_specpile,
 variable=c("dataname","condition", "genotype", "identifier",
 seq(from=4000, to=400, by=-1)),
 .before=ABS)
#export the data as csv
#data is baseline-corrected, normalized spec
filename_specpile_processed_longformat <- paste(today2, "_",
"specpile_processed_longformat.csv", sep="")
filename2_specpile_processed_longformat <- paste("C:/users/",desktopfolder,"/desktop/",
filename_specpile_processed_longformat, sep="")
write.csv(long_specpile,
 filename2_specpile_processed_longformat, row.names=FALSE)
#End of script
```

Script code 2: Principal component analysis

```
#salma_a3_pca_ftir_211213.r
#ftir_PCA for salma's paper 1
#this is a version to analyze differences in c3-h3 samples
#imput file should be in csv format,
#typically, "specpile_processed.csv" would be selected
#1st column should be the names of original spec files
#2nd col should be treatment ID such as c0, h3, c3
#3rd col should be genotype such as n61, ima,
```

```
#4th col should be "identifier" such as c0n61, which is used for grouping
#then followed by abs values from 4000 to 400
#import necessary libraries
library(conflicted)
library(dplyr)
library(ggplot2)
library(readr)
library(psych)
#clean up the R's brain
rm(list=ls())
#obtain desktop folder information for a windows user
#you must change the string within "xxx" below according to your computer
desktopfolder <- "akash"
#obtain date information
today <- Sys.Date()
yr <- substr(today, 3,4)
mo <- substr(today, 6,7)
day <- substr(today, 9,10)
today2 <- paste(yr, mo, day, sep="")
#invoke a file-opening window, specify the file,
#input data from .csv file
#data has to be baseline-corrected and normalized
#obtain the filename
inputfile <- file.choose()
filename <- basename(inputfile)
rawspecs <- read.csv(inputfile,
 header = T)
filename
#extract values used for calculation to a new df specmatrix
#values for wn4000 and 400 (zero values) should be removed
specmatrix <- dplyr::select(rawspecs, -(3605:3605))
specmatrix <- dplyr::select(specmatrix, -(1:5))
#separate "identifier" column (category info)
```

```
id_1 <- dplyr::select(rawspecs, (4:4))
#further trimming of values in the range of wn3600-4000
specmatrix <- dplyr::select(specmatrix, -(1:399))
#perform pca analysis using prcomp
#prcomp is standard but one of the oldest
pc = prcomp(specmatrix, scale =T)
#using the new 'principal()' in psych package
#pc <- psych::principal(specmatrix, nfactors=3601,
# rotate='none')
#display the summary
summary(pc)
#preparation of score data output
pc1_score <- pc$x(,1)
pc2_score <- pc$x(,2)
scoreonly <- as.data.frame(pc$x)
score <- cbind(id_1, scoreonly)
#export the score data as csv
#data is PC1-PC2 score
filename_PC12_score <- paste(today2, "_", "PC12_score.csv", sep="")
filename2_PC12_score <- paste("C:/users/",desktopfolder,"/desktop/",
 filename_PC12_score, sep="")
write.csv(score,
 filename2_PC12_score, row.names=FALSE)
#export rotation data as csv file
rotationonly <- as.data.frame(pc$rotation)
filename_pca_rotation <- paste(today2, "_", "pca_rotation.csv", sep="")
filename2_pca_rotation <- paste("C:/users/",desktopfolder,"/desktop/",
 filename_pca_rotation, sep="")
write.csv(rotationonly,
 filename2_pca_rotation, row.names=FALSE)
#export sdev data as csv file
sdevonly <- as.data.frame(pc$sdev)
filename_pca_sdev <- paste(today2, "_", "pca_sdev.csv", sep="")
```

```
filename2_pca_sdev <- paste("C:/users/",desktopfolder,"/desktop/",
 filename_pca_sdev, sep="")
write.csv(rotationonly,
 filename2_pca_sdev, row.names=FALSE)
#calculate the loadings, and export as csv file
loadingdata <- sweep(pc$rotation, MARGIN=2, pc$sdev, FUN="*")
filename_pca_loading <- paste(today2, "_", "pca_loading.csv", sep="")
filename2_pca_loading <- paste("C:/users/",desktopfolder,"/desktop/",
 filename_pca_loading, sep="")
write.csv(loadingdata,
 filename2_pca_loading, row.names=FALSE)
#draw the pc1_pc2 scoreplot
#size of the dots in the plot can be changed
 #by modifying the location of "geom_point(size=)
dev.new()
pca_scoreplot <- ggplot(score, aes(x = PC1,y = PC2,
 color = identifier)) +
 geom_point(size=2) +
 scale_color_manual(values=c("deepskyblue", "salmon")) +
 theme_bw()
print(pca_scoreplot)
#save the plot as png format
#you can change to .jpeg, .tiff, etc
#unit is in inch
filename_pca_scoreplot <- paste(today2, "_", "pca_scoreplot.csv", sep="")
filename2_pca_scoreplot <- paste("C:/users/",desktopfolder,"/desktop/",
 filename_pca_scoreplot, ".png", sep="")
ggsave(file = filename2_pca_scoreplot,
 plot = pca_scoreplot, dpi=100,
 width=7.2, height=4.8)
#extract contribution data
contribution <- as.data.frame(t(summary(pc)$importance))
names(contribution) <-
```

```
c("standard_deviation","proportion_of_variance","cumulative_proportion")
contri_rownames <- as.data.frame(rownames(contribution))
names(contri_rownames) <- "PC"
contribution <- cbind(contribution, contri_rownames)
#save the contribution data as csv file
#write.csv(contribution, "C:/Users/akash/desktop/contribution.csv")
filename_pca_contribution <- paste(today2, "_", "pca_contribution.csv", sep="")
filename2_pca_contribution <- paste("C:/users/",desktopfolder,"/desktop/",
 filename_pca_contribution, sep="")
write.csv(contribution,
 filename2_pca_contribution, row.names=FALSE)
#extract top 9 from the contribution data, and save as a png file
top9_contribution <- dplyr::slice(contribution, 1:9)
contribution_plot <- ggplot(top9_contribution, aes(x = PC, y = proportion_of_variance)) +
 geom_bar(stat="identity", fill="forestgreen") +
 theme_bw()
print(contribution_plot)
#ggsave(file = "C:/Users/akash/desktop/contribution_plot.png",
# plot = contribution_plot, dpi = 100,
# width = 3.6, height = 2.4)
filename_pca_contribution_plot <- paste(today2, "_", "pca_contribution_plot.csv", sep="")
filename2_pca_contribution_plot <- paste("C:/users/",desktopfolder,"/desktop/",
 filename_pca_contribution_plot, ".png", sep="")
ggsave(file = filename2_pca_contribution_plot,
 plot = contribution_plot, dpi=100,
 width=7.2, height=4.8)
#End of script
```

Script code 3: Linear Discriminant Analysis

```
#salma_a4_lda_ftir_211220.r
#a4_lda_ftir
 #linear discriminant analysis of ftir spectra
#clear the brain
rm(list=ls())
```

```
#library to register
#ggplot2 and dplyr are in tidyverse
library(conflicted)
library(tidyverse)
library(MASS)
library(klaR)
library(caret)
#obtain date information
today <- Sys.Date()
yr <- substr(today, 3,4)
mo <- substr(today, 6,7)
day <- substr(today, 9,10)
today2 <- paste(yr, mo, day, sep="")
#obtain desktop folder information for a windows user
#you must change the string within "xxx" below according to your computer
desktopfolder <- "akash"
#assemble path info
pathinfo <- paste("C:/users/",desktopfolder,"/desktop/", sep="")
#import the compiled ftir csv data
#the file to choose is normally "ftir_specpile_processed.csv"
spec1 <- file.choose()
spec2 <- read.csv(spec1,
 header = T)
#remove genotype and dataname info
specmatrix <- dplyr::select(spec2, -(1:3))
#remove values at wavenumbers 4000 and 400,
#which were used for baseline anchors
#then remove those at 3999-3601, the noisy region
specmatrix <- dplyr::select(specmatrix, -3602)
specmatrix <- dplyr::select(specmatrix, -2)
specmatrix <- dplyr::select(specmatrix, -(2:400))
#treatment <- dplyr::select(rawspecs, (2:2))
#set the seednumber for randomness
```

```
set.seed(101)
#split the samples into train(60%) and test(40%)
training_sample <- sample(c(TRUE, FALSE), nrow(specmatrix), replace = T, prob = c(0.6,
0.4))
trainspec <- specmatrix(training_sample, )
testspec <- specmatrix(!training_sample, )
#perform linear discriminant analysis
lda_spec_train <- lda(identifier ~ ., trainspec)
lda_spec_train
#training results check
#1st, transform them to the values
#then one dimensional histograms
#"mar" is the margin of bottom, left, top, right
lda_spec_train_results <- predict(lda_spec_train)
dev.new()
par("mar"=c(1,1,1,1))
trainhistogram1 <- ldahist(lda_spec_train_results$x(,1), g=trainspec$identifier)
print(trainhistogram1)
#save the histogram values for the training set as csv file
class_lda_spec_train <- as.data.frame(lda_spec_train_results$class)
x1_lda_spec_train <- as.data.frame(lda_spec_train_results$x(,1))
value_lda_spec_train <- cbind(class_lda_spec_train, x1_lda_spec_train)
names(value_lda_spec_train) <- c("identifier", "LD1")
filename_value_lda_spec_train <- paste(today2, "_", "value_lda_spec_train.csv", sep="")
filename2_value_lda_spec_train <- paste(pathinfo,"/", filename_value_lda_spec_train,
sep="")
write.csv(value_lda_spec_train,
 filename2_value_lda_spec_train, row.names=FALSE)
#draw histogram of train results for publication
dev.new()
lda_train_histogram2 <- ggplot(value_lda_spec_train,
 aes(x = LD1, fill =identifier)) +
 geom_histogram(position="identity",
```

```
 colour = "black", size=0.3,
 breaks=seq(from=-80, to=60, by=2)) +
 scale_fill_manual(values=c("deepskyblue", "salmon")) +
 theme_bw()
print(lda_train_histogram2)
#save the plot as png format
#you should change the path according to your system
#you can change to .jpeg, .tiff, etc
#unit is in inch
filename_lda_train_histogram2 <- paste(today2, "_", "lda_train_histogram2.png", sep="")
filename2_lda_train_histogram2 <- paste(pathinfo,"/", filename_lda_train_histogram2,
sep="")
ggsave(file = filename2_lda_train_histogram2,
 plot = lda_train_histogram2, dpi=100,
 width=7.2, height=3.6)
#test set check
#1st, transform them to the values
#then one dimensional histograms
lda_spec_test_results <- predict(lda_spec_train, testspec)
dev.new()
par("mar"=c(1,1,1,1))
testhistogram1 <- ldahist(lda_spec_test_results$x(,1), g=testspec$identifier)
print(testhistogram1)
#save the histogram values for the test set as csv file
class_lda_spec_test <- as.data.frame(lda_spec_test_results$class)
x1_lda_spec_test <- as.data.frame(lda_spec_test_results$x(,1))
value_lda_spec_test <- cbind(class_lda_spec_test, x1_lda_spec_test)
names(value_lda_spec_test) <- c("identifier", "LD1")
filename_value_lda_spec_test <- paste(today2, "_", "value_lda_spec_test.csv", sep="")
filename2_value_lda_spec_test <- paste(pathinfo,"/", filename_value_lda_spec_test, sep="")
write.csv(value_lda_spec_test,
 filename2_value_lda_spec_test, row.names=FALSE)
#draw histogram of test results for publication
```

```
dev.new()
lda_test_histogram2 <- ggplot(value_lda_spec_test,
 aes(x = LD1, fill =identifier)) +
 geom_histogram(position="identity",
 colour = "black", size=0.3,
 breaks=seq(from=-80, to=60, by=2)) +
 scale_fill_manual(values=c("deepskyblue", "salmon")) +
 theme_bw()
print(lda_test_histogram2)
#save the plot as png format
#you should change the path according to your system
#you can change to .jpeg, .tiff, etc
#unit is in inch
filename_lda_test_histogram2 <- paste(today2, "_", "lda_test_histogram2.png", sep="")
filename2_lda_test_histogram2 <- paste(pathinfo,"/", filename_lda_test_histogram2,
sep="")
ggsave(file = filename2_lda_test_histogram2,
 plot = lda_test_histogram2, dpi=100,
 width=7.2, height=3.6)
#extract LD1 loading
scalingdata <- lda_spec_train$scaling
#transform LD1 loading to dataframe
#add wavenumber info
scalingdf <- as.data.frame(t(scalingdata))
wnlist <- seq(3600, 401, length=3200)
wnlist2 <- as.data.frame(t(wnlist))
colnames(scalingdf) <- c(seq(3600, 401, length=3200))
colnames(wnlist2) <- c(seq(3600, 401, length=3200))
scalingdf2 <- rbind(wnlist2, scalingdf)
scalingdf3 <- as.data.frame(t(scalingdf2))
names(scalingdf3)(1) <- "wavenumber"
#change the wavenumber in ascending order, and save it as csv
scalingdf4 <- arrange(scalingdf3, wavenumber)
```

```
filename_scalingdf4 <- paste(today2, "_", "LD1_loading.csv", sep="")
filename2_scalingdf4 <- paste(pathinfo,"/", filename_scalingdf4, sep="")
write.csv(scalingdf4,
 filename2_scalingdf4, row.names=FALSE)
#plot LD1 contribution, scatter plot version
dev.new()
lda_loading_scatterplot <- ggplot(scalingdf4, aes(x = wavenumber, y = LD1)) +
 geom_point(size=0.5) +
 theme_bw()
print(lda_loading_scatterplot)
filename_lda_loadingscatterplot <- paste(today2, "_", "LDA_Loading_ScatterPlot.png",
sep="")
filename2_lda_loadingscatterplot <- paste(pathinfo,"/", filename_lda_loadingscatterplot,
".png", sep="")
ggsave(file = filename2_lda_loadingscatterplot,
 plot = lda_loading_scatterplot, dpi = 100,
 width = 7.2, height = 4.8)
#plot LD1 contribution, line plot version
dev.new()
lda_loading_lineplot <- ggplot(scalingdf4, aes(x=wavenumber, y=LD1))+
 geom_line(size=0.2)+
 theme_bw()
print(lda_loading_lineplot)
filename_lda_loadinglineplot <- paste(today2, "_", "LDA_Loading_LinePlot.png", sep="")
filename2_lda_loadinglineplot <- paste(pathinfo,"/", filename_lda_loadinglineplot, ".png",
sep="")
ggsave(file = filename2_lda_loadinglineplot,
 plot = lda_loading_lineplot, dpi = 100,
 width = 7.2, height = 4.8)
#pick up peak candidate in LD1 plot
#that are higher than the threshold of 0.15
peakcandidate1 <- dplyr::filter(scalingdf4, LD1>0.15)
#check that the candidate is the higher than the neighboring wavenumbers
```

```
ncandidate <- nrow(peakcandidate1)
peakcandidate2 <- data.frame(matrix(rep(NA,8),nrow=1))(numeric(0),)
colnames(peakcandidate2) <-
 c("wavenumber","LD1","LD1m1","LD1p1","peak", "GoFurther","LargerWnBoundary",
"SmallerWnBoundary")
for(i in 1:ncandidate){
 peakcandidate_tempo <- data.frame(matrix(rep(NA,8),nrow=1))(numeric(0),)
 colnames(peakcandidate_tempo) <-
 c("wavenumber","LD1","LD1m1","LD1p1","peak","GoFurther","LargerWnBoundary",
"SmallerWnBoundary")
 wn_quest <- peakcandidate1(i,1)
 wn_quest_m1 <- wn_quest - 1
 wn_quest_p1 <- wn_quest + 1
 peakcandidate_tempo(1,1) <- wn_quest
 peakcandidate_tempo(1,2) <- scalingdf4(wn_quest-400,2)
 peakcandidate_tempo(1,3) <- scalingdf4(wn_quest_m1-400,2)
 peakcandidate_tempo(1,4) <- scalingdf4(wn_quest_p1-400,2)
 if(peakcandidate_tempo(1,2)>peakcandidate_tempo(1,3)
 & peakcandidate_tempo(1,2)>peakcandidate_tempo(1,4)){
 peakcandidate_tempo(1,5) <- 1
 }
 peakcandidate2 <- rbind(peakcandidate2, peakcandidate_tempo)
}
#save the peak candidate as csv file
#change the wavenumber in descending order of LD1, and save it as csv
peakcandidate3 <- dplyr::filter(peakcandidate2, peak==1)
peakcandidate3 <- arrange(peakcandidate3, desc(LD1))
filename_peakcandidate <- paste(today2, "_", "PeakCandidateList.csv", sep="")
filename2_peakcandidate <- paste(pathinfo,"/", filename_peakcandidate, sep="")
write.csv(peakcandidate3,
 filename2_peakcandidate, row.names=FALSE)
#pick up valley candidate in LD1 plot
#that are lower than the threshold of -0.15
```

```
valleycandidate1 <- dplyr::filter(scalingdf4, LD1 < -0.15)
#check that the candidate is the higher than the neighboring wavenumbers
n_valleycandidate <- nrow(valleycandidate1)
valleycandidate2 <- data.frame(matrix(rep(NA,8),nrow=1))(numeric(0),)
colnames(valleycandidate2) <-
 c("wavenumber","LD1","LD1m1","LD1p1","valley","GoFurther","LargerWnBoundary",
"SmallerWnBoundary")
for(i in 1:n_valleycandidate){
 valleycandidate_tempo <- data.frame(matrix(rep(NA,8),nrow=1))(numeric(0),)
 colnames(valleycandidate_tempo) <-
 c("wavenumber","LD1","LD1m1","LD1p1","valley","GoFurther","LargerWnBoundary",
"SmallerWnBoundary")
 wn_quest <- valleycandidate1(i,1)
 wn_quest_m1 <- wn_quest - 1
 wn_quest_p1 <- wn_quest + 1
 valleycandidate_tempo(1,1) <- wn_quest
 valleycandidate_tempo(1,2) <- scalingdf4(wn_quest-400,2)
 valleycandidate_tempo(1,3) <- scalingdf4(wn_quest_m1-400,2)
 valleycandidate_tempo(1,4) <- scalingdf4(wn_quest_p1-400,2)
 if(valleycandidate_tempo(1,2) < valleycandidate_tempo(1,3)
 & valleycandidate_tempo(1,2) < valleycandidate_tempo(1,4)){
 valleycandidate_tempo(1,5) <- 1
 }
 valleycandidate2 <- rbind(valleycandidate2, valleycandidate_tempo)
}
#save the valley candidate as csv file
#change the wavenumber in ascending order of LD1, and save it as csv
valleycandidate3 <- dplyr::filter(valleycandidate2, valley==1)
valleycandidate3 <- arrange(valleycandidate3, LD1)
filename_valleycandidate <- paste(today2, "_", "ValleyCandidateList.csv", sep="")
filename2_valleycandidate <- paste(pathinfo,"/", filename_valleycandidate, sep="")
write.csv(valleycandidate3,
 filename2_valleycandidate, row.names=FALSE)
```

#End of script

Script code 4: Quest anchors for Fm-markers

```
#salma_a5_quest_anchors_lda_211220.r
#quest anchors for LDA peaks
 #for identifying the most effective pair of anchor points
 #that show peaks in LDA contribution plot
#this is for Salma's data on N61 c3-h3 chamber comparison.
#clear the brain
rm(list=ls())
#library to register
#ggplot2 and dplyr are in tidyverse
library(conflicted)
library(tidyverse)
library(MASS)
library(klaR)
library(caret)
#obtain date information
today <- Sys.Date()
yr <- substr(today, 3,4)
mo <- substr(today, 6,7)
day <- substr(today, 9,10)
today2 <- paste(yr, mo, day, sep="")
#obtain desktop folder information for a windows user
#you must change the string within "xxx" below according to your computer
desktopfolder <- "akash"
#assemble path info
pathinfo <- paste("C:/users/",desktopfolder,"/desktop/", sep="")
#import a "xxxxxx_PeakCandidateXXXX.csv"
#that is modified from PeakCandidateList2.csv
#the import file is a dataframe, and it contains following 8 columns
#"wavenumber, LD1, LD1m1, LD1p1, peak, GoFurther, LargerWnBoundary,
SmallerWnBoundary"
#1st row is the column name, and data is in 2nd row
```

```
#the last two typically set at 150 larger and smaller than the target wavenumber
print("Please specify xxxxxx_a4_PeakCandidateXXXX.csv")
PeakCandidateInput <- file.choose()
PeakCandidateInput2 <- read.csv(PeakCandidateInput,
 header = T)
#import LDA-peak and boundary information
lda_peak_wn <- PeakCandidateInput2(1,1)
lda_LargerWnBoundary_wn <- PeakCandidateInput2(1,7)
lda_SmallerWnBoundary_wn <- PeakCandidateInput2(1,8)
#modify peak_wn variable to the style of column name
lda_peak_wn_string <- as.character(lda_peak_wn)
lda_peak_wn_colname <- paste("X",lda_peak_wn_string, sep="")
#import the compiled ftir csv data
#the file to choose is normally "a2_specpile_processed.csv"
print("Please specify xxxxxx_a2_specpile_processed.csv")
spec1 <- file.choose()
specmatrix <- read.csv(spec1,
 header = T)
#separate the data into c3 and h3
c3matrix <- dplyr::filter(specmatrix, identifier == "c3n61")
h3matrix <- dplyr::filter(specmatrix, identifier == "h3n61")
#now, the matrix data should have 223 columns
#consist of first 2 factorial data, followed by 221 wn data columns
#sum of wn and column number equals to 663
#thus, wn576 correspond to column 87.
#extract the wn576 data from each group
c3peakabs <- dplyr::select(c3matrix, lda_peak_wn_colname)
h3peakabs <- dplyr::select(h3matrix, lda_peak_wn_colname)
totalpeakabs <- dplyr::select(specmatrix, lda_peak_wn_colname)
#2: An empty output dataframe generated.
#rep(NA, 27) is a function to generate NA for 47 times.
quest_anchor_summary <- data.frame(matrix(rep(NA, 47), nrow=1))(numeric(0), )
colnames(quest_anchor_summary) <- c("wn1and2", "wn1", "wn2",
```

```
 "h3a_HigherMedian", "h3b_HigherMedian", "a_h_c_ratio", "b_h_c_ratio", "p_a", "p_b",
 "hih3_a_score_boxplot", "hih3_b_score_boxplot",
 "loh3_a_score_boxplot", "loh3_b_score_boxplot",
 "sand_score_a", "sand_score_b",
 "hih3_a_BoxSeparated", "hih3_c3a_med_under_h3a1stQ",
"hih3_c3a3rdQ_under_h3a_med",
 "hih3_b_BoxSeparated", "hih3_c3b_med_under_h3b1stQ",
"hih3_c3b3rdQ_under_h3b_med",
 "loh3_a_BoxSeparated", "loh3_c3a_med_over_h3a3rdQ", "loh3_c3a1stQ_over_h3a_med",
 "loh3_b_BoxSeparated", "loh3_c3b_med_over_h3b3rdQ", "loh3_c3b1stQ_over_h3b_med",
 "sand_c3a_median", "sand_c3a_1stQ", "sand_h3a_median", "sand_h3a_1stQ",
 "sand_c3b_median", "sand_c3b_1stQ", "sand_h3b_median", "sand_h3b_1stQ",
 "c3a_1stQ", "c3a_median", "c3a_3rdQ", "h3a_1stQ", "h3a_median", "h3a_3rdQ",
 "c3b_1stQ", "c3b_median", "c3b_3rdQ", "h3b_1stQ", "h3b_median", "h3b_3rdQ")
#3: Outward looping start
#loop should be from 100-higher wn from the peak, i.e., lda_LargerWnBoundary_wn,
#to 10-higher wn from the peak
#loop value is specified by column number, i.e., wn4000 is in col5, wn3999 is in col6
 #the sum of wn"xxxx" and col"x" is 4005.
 #thus, col number for lda_LargerWnBoundary_wn should be "4005-
lda_LargerWnBoundary_wn
 #col number for lda_peak_wn is "4005-Lda_peak_wn"
 #col number for 10-higher wn from the peak is "3995-lda_peak_wn"
#modify the line below to
 #"i_startpoint:i_startpoint" for pilot test,
 #and "i_startpoint:i_endpoint" for full calculation
#for (i in i_startpoint:i_endpoint){
i_startpoint <- 4005 - lda_LargerWnBoundary_wn
i_endpoint <- 3995 - lda_peak_wn
for (i in i_startpoint:i_endpoint){
 #4: Setting the anchor1 value.
 c3anchor1 <- dplyr::select(c3matrix, i)
 h3anchor1 <- dplyr::select(h3matrix, i)
```

```
anchor1total <- dplyr::select(specmatrix, i)

#5: Inward looping start.
#loop should be from 10-step downstream from the peak,
#col number for lda_peak_wn is "4005-Lda_peak_wn"
#thus col number for 10-step downstream is "4015-lda_peak_wn"
#the loop is stopped at "lda_SmallerWnBoundary_wn"
#col number for "lda_SmallerWnBoundary_wn" is "4005-lda_SmallerWnBoundary_wn"
#modify the line below to
#"j_startpoint:j_startpoint" for pilot test,
#and "j_startpoint:j_endpoint" for full calculation
#for (j in (j_startpoint:j_endpoint)){
j_startpoint <- 4015 - lda_peak_wn
j_endpoint <- 4005 - lda_SmallerWnBoundary_wn
for (j in j_startpoint:j_endpoint){
c3anchor2 <- dplyr::select(c3matrix, j)
h3anchor2 <- dplyr::select(h3matrix, j)
anchor2total <- dplyr::select(specmatrix, j)

#set up a temporary df for the results
tempo_anchor_results <- data.frame(matrix(rep(NA, 47), nrow=1))(numeric(0), )
tempo_candidate_hih3_wn1base <- data.frame(matrix(rep(NA, 44),
nrow=1))(numeric(0), )
tempo_candidate_hih3_wn2base <- data.frame(matrix(rep(NA, 44),
nrow=1))(numeric(0), )
tempo_candidate_loh3_wn1base <- data.frame(matrix(rep(NA, 44),
nrow=1))(numeric(0), )
tempo_candidate_loh3_wn2base <- data.frame(matrix(rep(NA, 44),
nrow=1))(numeric(0), )

#record the anchors info
wn1 <- colnames(specmatrix)(i)
wn2 <- colnames(specmatrix)(j)
```

```
wn1and2 <- paste(as.character(wn1), as.character(wn2), sep="and")
#define formula for ftir-marker(fm)
#set the wn1 as basepoint, calculate marker "a" value
#set the wn2 as basepoint, calculate marker "b" value
c3fma <- (c3peakabs-c3anchor1)/(c3anchor2-c3anchor1)
h3fma <- (h3peakabs-h3anchor1)/(h3anchor2-h3anchor1)
c3fmb <- (c3peakabs-c3anchor2)/(c3anchor1-c3anchor2)
h3fmb <- (h3peakabs-h3anchor2)/(h3anchor1-h3anchor2)

#6: Calculate the key statistics
c3a_summary <- summary(c3fma(,1))
c3a_1stQ <- as.numeric(c3a_summary)(2)
c3a_median <- as.numeric(c3a_summary)(3)
c3a_3rdQ <- as.numeric(c3a_summary)(5)
h3a_summary <- summary(h3fma(,1))
h3a_1stQ <- as.numeric(h3a_summary)(2)
h3a_median <- as.numeric(h3a_summary)(3)
h3a_3rdQ <- as.numeric(h3a_summary)(5)
c3b_summary <- summary(c3fmb(,1))
c3b_1stQ <- as.numeric(c3b_summary)(2)
c3b_median <- as.numeric(c3b_summary)(3)
c3b_3rdQ <- as.numeric(c3b_summary)(5)
h3b_summary <- summary(h3fmb(,1))
h3b_1stQ <- as.numeric(h3b_summary)(2)
h3b_median <- as.numeric(h3b_summary)(3)
h3b_3rdQ <- as.numeric(h3b_summary)(5)
#test 1
#judge whether the h3a_median is higher
#(in theory, it is NOT for the valley marker)
if(h3a_median > c3a_median){
h3a_HigherMedian <- 1
} else {
h3a_HigherMedian <- 0
```

```
}
#judge whether the h3b_median is lower
#(in theory, it is NOT for the valley marker)
if(h3b_median > c3b_median){
h3b_HigherMedian <- 1
} else {
h3b_HigherMedian <- 0
}

#test 2
#judge whether boxplot is separated and not overlapped
#when h3_median is higher than c3_median
if(c3a_3rdQ < h3a_1stQ){
hih3_a_BoxSeparated <- 1
} else {
hih3_a_BoxSeparated <- 0
}

if(c3a_3rdQ < h3a_1stQ){
hih3_b_BoxSeparated <- 1
} else {
hih3_b_BoxSeparated <- 0
}
#when h3_median is lower than c3_median
if(c3a_1stQ > h3a_3rdQ){
loh3_a_BoxSeparated <- 1
} else {
loh3_a_BoxSeparated <- 0
}
if(c3b_1stQ > h3b_3rdQ){
loh3_b_BoxSeparated <- 1
} else {
loh3_b_BoxSeparated <- 0
```

```
}

#test 3
#partial overlap of boxplot
#judge whether boxplot is more than 50%-separated
#test 3-1 and 3-2
#when h3_median is higher than c3_median
#test 3-1.
#median<1rdQ check
if(c3a_median < h3a_1stQ){
hih3_c3a_med_under_h3a1stQ <- 1
} else {
hih3_c3a_med_under_h3a1stQ <- 0
}

if(c3b_median < h3b_1stQ){
hih3_c3b_med_under_h3b1stQ <- 1
} else {
hih3_c3b_med_under_h3b1stQ <- 0
}

#test 3-2.
#3rdQ<median check
if(c3a_3rdQ < h3a_median){
hih3_c3a3rdQ_under_h3a_med <- 1
} else {
hih3_c3a3rdQ_under_h3a_med <- 0
}

if(c3b_3rdQ < h3b_median){
hih3_c3b3rdQ_under_h3b_med <- 1
} else {
hih3_c3b3rdQ_under_h3b_med <- 0
```

```
}

#test 3-3 and 3-4
#when h3_median is lower than c3_median
#test 3-3.
#median>3rdQ check
if(c3a_median > h3a_3rdQ){
loh3_c3a_med_over_h3a3rdQ <- 1
} else {
loh3_c3a_med_over_h3a3rdQ <- 0
}

if(c3b_median > h3b_3rdQ){
loh3_c3b_med_over_h3b3rdQ <- 1
} else {
loh3_c3b_med_over_h3b3rdQ <- 0
}
#test 3-4.
#1stQ>median check
if(c3a_1stQ > h3a_median){
loh3_c3a1stQ_over_h3a_med <- 1
} else {
loh3_c3a1stQ_over_h3a_med <- 0
}

if(c3b_1stQ > h3b_median){
loh3_c3b1stQ_over_h3b_med <- 1
} else {
loh3_c3b1stQ_over_h3b_med <- 0
}
#test 3 summary
#the xxh3_x_score_boxplot is a score that the two boxes in the plot is fully or partially
separated.
```

```
#the full score is 3.
hih3_a_score_boxplot <- hih3_a_BoxSeparated + hih3_c3a_med_under_h3a1stQ +
hih3_c3a3rdQ_under_h3a_med
hih3_b_score_boxplot <- hih3_b_BoxSeparated + hih3_c3b_med_under_h3b1stQ +
hih3_c3b3rdQ_under_h3b_med
loh3_a_score_boxplot <- loh3_a_BoxSeparated + loh3_c3a_med_over_h3a3rdQ +
loh3_c3a1stQ_over_h3a_med
loh3_b_score_boxplot <- loh3_b_BoxSeparated + loh3_c3b_med_over_h3b3rdQ +
loh3_c3b1stQ_over_h3b_med

#test 4
#sandwich status of the target between wn1 and wn2
#when it is, the fm value should be between 0 and 1
if(0 < c3a_median & c3a_median < 1){
sand_c3a_median <- 1
} else {
sand_c3a_median <- 0
}
if(0 < c3a_1stQ & c3a_1stQ < 1){
sand_c3a_1stQ <- 1
} else {
sand_c3a_1stQ <- 0
}
if(0 < h3a_median & h3a_median < 1){
sand_h3a_median <- 1
} else {
sand_h3a_median <- 0
}

if(0 < h3a_1stQ & h3a_1stQ < 1){
sand_h3a_1stQ <- 1
} else {
sand_h3a_1stQ <- 0
```

```
}

#summary of sandwich status
#the full mark is 4, but it is not the absolute requirement
sand_score_a <- sand_c3a_median + sand_c3a_1stQ + sand_h3a_median +
sand_h3a_1stQ

if(0 < c3b_median & c3b_median < 1){
sand_c3b_median <- 1
} else {
sand_c3b_median <- 0
}

if(0 < c3b_1stQ & c3b_1stQ < 1){
sand_c3b_1stQ <- 1
} else {
sand_c3b_1stQ <- 0
}

if(0 < h3b_median & h3b_median < 1){
sand_h3b_median <- 1
} else {
sand_h3b_median <- 0
}

if(0 < h3b_1stQ & h3b_1stQ < 1){
sand_h3b_1stQ <- 1
} else {
sand_h3b_1stQ <- 0
}

sand_score_b <- sand_c3b_median + sand_c3b_1stQ + sand_h3b_median +
sand_h3b_1stQ
```

```
#8: Calculate the p value by t-test
a_ttest <- t.test(c3fma(,1), h3fma(,1), var.equal=T)
p_a <- a_ttest$p.value
b_ttest <- t.test(c3fmb(,1), h3fmb(,1), var.equal=T)
p_b <- b_ttest$p.value

#calculate the c3/h3 ratio
a_h_c_ratio <- h3a_median/c3a_median
b_h_c_ratio <- h3b_median/c3b_median

#9: Record the results to the output dataframe
tempo_anchor_results <- as.data.frame(t(c(wn1and2, wn1, wn2,
h3a_HigherMedian, h3b_HigherMedian, a_h_c_ratio, b_h_c_ratio, p_a, p_b,
hih3_a_score_boxplot, hih3_b_score_boxplot,
loh3_a_score_boxplot, loh3_b_score_boxplot,
sand_score_a, sand_score_b,
hih3_a_BoxSeparated, hih3_c3a_med_under_h3a1stQ,
hih3_c3a3rdQ_under_h3a_med,
hih3_b_BoxSeparated, hih3_c3b_med_under_h3b1stQ,
hih3_c3b3rdQ_under_h3b_med,
loh3_a_BoxSeparated, loh3_c3a_med_over_h3a3rdQ, loh3_c3a1stQ_over_h3a_med,
loh3_b_BoxSeparated, loh3_c3b_med_over_h3b3rdQ, loh3_c3b1stQ_over_h3b_med,
sand_c3a_median, sand_c3a_1stQ, sand_h3a_median, sand_h3a_1stQ,
sand_c3b_median, sand_c3b_1stQ, sand_h3b_median, sand_h3b_1stQ,
c3a_1stQ, c3a_median, c3a_3rdQ, h3a_1stQ, h3a_median, h3a_3rdQ,
c3b_1stQ, c3b_median, c3b_3rdQ, h3b_1stQ, h3b_median, h3b_3rdQ)))
colnames(tempo_anchor_results) <- c("wn1and2", "wn1", "wn2",
"h3a_HigherMedian", "h3b_HigherMedian", "a_h_c_ratio", "b_h_c_ratio", "p_a",
"p_b",
"hih3_a_score_boxplot", "hih3_b_score_boxplot",
"loh3_a_score_boxplot", "loh3_b_score_boxplot",
"sand_score_a", "sand_score_b",
```

```
 "hih3_a_BoxSeparated", "hih3_c3a_med_under_h3a1stQ",
"hih3_c3a3rdQ_under_h3a_med",
 "hih3_b_BoxSeparated", "hih3_c3b_med_under_h3b1stQ",
"hih3_c3b3rdQ_under_h3b_med",
 "loh3_a_BoxSeparated", "loh3_c3a_med_over_h3a3rdQ",
"loh3_c3a1stQ_over_h3a_med",
 "loh3_b_BoxSeparated", "loh3_c3b_med_over_h3b3rdQ",
"loh3_c3b1stQ_over_h3b_med",
 "sand_c3a_median", "sand_c3a_1stQ", "sand_h3a_median", "sand_h3a_1stQ",
 "sand_c3b_median", "sand_c3b_1stQ", "sand_h3b_median", "sand_h3b_1stQ",
 "c3a_1stQ", "c3a_median", "c3a_3rdQ", "h3a_1stQ", "h3a_median", "h3a_3rdQ",
 "c3b_1stQ", "c3b_median", "c3b_3rdQ", "h3b_1stQ", "h3b_median", "h3b_3rdQ")

 #merge the generated data into output dataframe
 quest_anchor_summary = rbind(quest_anchor_summary, tempo_anchor_results)
 #10: Inward looping going out and iterate
 }
 #11: Outward looping going out and iterate
}

#12: Save the output dataframe
#write.csv(quest_anchor_summary,
# "C:/users/akash/desktop/anchor_all_data.csv", row.names=FALSE)
filename_quest_anchor_summary <- paste(today2, "_a5_quest_anchor_alldata_",
lda_peak_wn_string, ".csv", sep="")
filename2_quest_anchor_summary <- paste(pathinfo,"/", filename_quest_anchor_summary,
sep="")
write.csv(quest_anchor_summary,
 filename2_quest_anchor_summary, row.names=FALSE)
#split the data into 4 category of high or low h3, and a- or b-basepoint
hih3_a_all_data <- dplyr::filter(quest_anchor_summary, h3a_HigherMedian==1)
hih3_b_all_data <- dplyr::filter(quest_anchor_summary, h3b_HigherMedian==1)
loh3_a_all_data <- dplyr::filter(quest_anchor_summary, h3a_HigherMedian==0)
```

```
loh3_b_all_data <- dplyr::filter(quest_anchor_summary, h3b_HigherMedian==0)
#select significant candidates
#select p<0.05
hih3_a_candidate <- dplyr::filter(hih3_a_all_data, as.numeric(p_a)<0.05)
hih3_b_candidate <- dplyr::filter(hih3_b_all_data, as.numeric(p_b)<0.05)
loh3_a_candidate <- dplyr::filter(loh3_a_all_data, as.numeric(p_a)<0.05)
loh3_b_candidate <- dplyr::filter(loh3_b_all_data, as.numeric(p_b)<0.05)
#select more than 1.5 fold absolute difference in ch_ratio
#anchor_candidate2 <- dplyr::filter(anchor_candidate1,
# a_ch_ratio>1.5|a_ch_ratio<0.75|b_ch_ratio>1.5|b_ch_ratio<0.75)
#convert the type of score_boxplot to numeric
hih3_a_candidate$hih3_a_score_boxplot <-
as.numeric(hih3_a_candidate$hih3_a_score_boxplot)
hih3_b_candidate$hih3_b_score_boxplot <-
as.numeric(hih3_b_candidate$hih3_b_score_boxplot)
loh3_a_candidate$loh3_a_score_boxplot <-
as.numeric(loh3_a_candidate$loh3_a_score_boxplot)
loh3_b_candidate$loh3_b_score_boxplot <-
as.numeric(loh3_b_candidate$loh3_b_score_boxplot)
#14: Sort them according to the ranking
#sorting
#if you wish to sort in the descending order, replace "p" to "desc(p)"
hih3_a_candidate2 <- arrange(hih3_a_candidate, desc(hih3_a_score_boxplot))
hih3_b_candidate2 <- arrange(hih3_b_candidate, desc(hih3_b_score_boxplot))
loh3_a_candidate2 <- arrange(loh3_a_candidate, desc(loh3_a_score_boxplot))
loh3_b_candidate2 <- arrange(loh3_b_candidate, desc(loh3_b_score_boxplot))
#eliminate unnecessary columns for the output
hih3_a_candidate3 <- dplyr::select(hih3_a_candidate2, wn1and2, wn1, wn2,
 h3a_HigherMedian, a_h_c_ratio, p_a, hih3_a_score_boxplot, sand_score_a,
 hih3_a_BoxSeparated, hih3_c3a_med_under_h3a1stQ, hih3_c3a3rdQ_under_h3a_med,
 sand_c3a_median, sand_c3a_1stQ, sand_h3a_median, sand_h3a_1stQ,
 c3a_1stQ, c3a_median, c3a_3rdQ, h3a_1stQ, h3a_median, h3a_3rdQ)
hih3_b_candidate3 <- dplyr::select(hih3_b_candidate2, wn1and2, wn1, wn2,
```

```
 h3b_HigherMedian, b_h_c_ratio, p_b, hih3_b_score_boxplot, sand_score_b,
 hih3_b_BoxSeparated, hih3_c3b_med_under_h3b1stQ, hih3_c3b3rdQ_under_h3b_med,
 sand_c3b_median, sand_c3b_1stQ, sand_h3b_median, sand_h3b_1stQ,
 c3b_1stQ, c3b_median, c3b_3rdQ, h3b_1stQ, h3b_median, h3b_3rdQ)
loh3_a_candidate3 <- dplyr::select(loh3_a_candidate2, wn1and2, wn1, wn2,
 h3a_HigherMedian, a_h_c_ratio, p_a, loh3_a_score_boxplot, sand_score_a,
 loh3_a_BoxSeparated, loh3_c3a_med_over_h3a3rdQ, loh3_c3a1stQ_over_h3a_med,
 sand_c3a_median, sand_c3a_1stQ, sand_h3a_median, sand_h3a_1stQ,
 c3a_1stQ, c3a_median, c3a_3rdQ, h3a_1stQ, h3a_median, h3a_3rdQ)
loh3_b_candidate3 <- dplyr::select(loh3_b_candidate2, wn1and2, wn1, wn2,
 h3b_HigherMedian, b_h_c_ratio, p_b, loh3_b_score_boxplot, sand_score_b,
 loh3_b_BoxSeparated, loh3_c3b_med_over_h3b3rdQ, loh3_c3b1stQ_over_h3b_med,
 sand_c3b_median, sand_c3b_1stQ, sand_h3b_median, sand_h3b_1stQ,
 c3b_1stQ, c3b_median, c3b_3rdQ, h3b_1stQ, h3b_median, h3b_3rdQ)
#14: Save the output dataframe
#write.csv(hih3_a_candidate3,
# "C:/users/akash/desktop/hih3_a_candidate3.csv", row.names=FALSE)
filename_hih3_a_candidate <- paste(today2, "_a5_hih3_a_candidatefm_",
lda_peak_wn_string, ".csv", sep="")
filename2_hih3_a_candidate <- paste(pathinfo,"/", filename_hih3_a_candidate, sep="")
write.csv(hih3_a_candidate3,
 filename2_hih3_a_candidate, row.names=FALSE)
#write.csv(hih3_b_candidate3,
# "C:/users/akash/desktop/hih3_b_candidate3.csv", row.names=FALSE)
filename_hih3_b_candidate <- paste(today2, "_a5_hih3_b_candidatefm_",
lda_peak_wn_string, ".csv", sep="")
filename2_hih3_b_candidate <- paste(pathinfo,"/", filename_hih3_b_candidate, sep="")
write.csv(hih3_b_candidate3,
 filename2_hih3_b_candidate, row.names=FALSE)
#write.csv(loh3_a_candidate3,
# "C:/users/akash/desktop/loh3_a_candidate3.csv", row.names=FALSE)
filename_loh3_a_candidate <- paste(today2, "_a5_loh3_a_candidatefm_",
lda_peak_wn_string, ".csv", sep="")
```

```
filename2_loh3_a_candidate <- paste(pathinfo,"/", filename_loh3_a_candidate, sep="")
write.csv(loh3_a_candidate3,
 filename2_loh3_a_candidate, row.names=FALSE)
#write.csv(loh3_b_candidate3,
# "C:/users/akash/desktop/loh3_b_candidate3.csv", row.names=FALSE)
filename_loh3_b_candidate <- paste(today2, "_a5_loh3_b_candidatefm_",
lda_peak_wn_string, ".csv", sep="")
filename2_loh3_b_candidate <- paste(pathinfo,"/", filename_loh3_b_candidate, sep="")
write.csv(loh3_b_candidate3,
 filename2_loh3_b_candidate, row.names=FALSE)
#End of script
```

Script code 5: Evaluation of Fm-anchor candidates

```
#salma_a6_anchor_candi_evalu_211225.r
#anchor candidate evaluation
 #for selecting the most effective pair of anchor points
 #for the potential ftir markers
#this is for Salma's data on N61 c3-h3 chamber comparison.
#clear the brain
rm(list=ls())
#library to register
#ggplot2 and dplyr are in tidyverse
library(conflicted)
library(tidyverse)
library(MASS)
library(klaR)
library(caret)
#obtain date information
today <- Sys.Date()
yr <- substr(today, 3,4)
mo <- substr(today, 6,7)
day <- substr(today, 9,10)
today2 <- paste(yr, mo, day, sep="")
#obtain desktop folder information for a windows user
```

```
#you must change the string within "xxx" below according to your computer
desktopfolder <- "akash"
#assemble path info
pathinfo <- paste("C:/users/",desktopfolder,"/desktop/", sep="")
#import the Peak/Valley CandidateXXXX.csv"
print("Please select xxxxxx_a4_PeakCandidateXXXX.csv file")
print("Or, please select xxxxxx_a4_ValleyCandidateXXXX.csv file")
print("Please make sure that the cols 7 and 8 for search boundary is filled")
print("Typically, set the value 150 larger and smaller than the target")
print("Please select xxxxxx_a5_PeakCandidateXXXX.csv file")
print("Or, please select xxxxxx_a5_ValleyCandidateXXXX.csv file")
PeakCandidateInput <- file.choose()
PeakCandidateInput2 <- read.csv(PeakCandidateInput,
 header = T)
#extract LDA-peak and boundary information
lda_peak_wn <- PeakCandidateInput2(1,1)
lda_LargerWnBoundary_wn <- PeakCandidateInput2(1,7)
lda_SmallerWnBoundary_wn <- PeakCandidateInput2(1,8)
#modify peak_wn variable to the style of column name
lda_peak_wn_string <- as.character(lda_peak_wn)
lda_peak_wn_colname <- paste("X",lda_peak_wn_string, sep="")
#import the 2nd "xxxxxx_a5_xxh3_x_candidatefmxxxx.csv" data
print("Please select xxxxxx_a5_xxh3_x_candidatefmXXXX.csv file")
candi1 <- file.choose()
candi2 <- read.csv(candi1,
 header = T)
#import the 3rd "xxxxxx_a1_specmean.csv" data
print("Please select xxxxxx_a1_specmean.csv file")
spec1 <- file.choose()
twospec <- read.csv(spec1,
 header = T)
#arrange the spectra
twospec2 <- dplyr::select(twospec, -(c(1:1)))
```

```
colnames(twospec2) <- seq(from=4000, to=400, by=-1)
wnlist1 <- as.data.frame(t(seq(from=4000, to=400, by=-1)))
colnames(wnlist1) <- seq(from=4000, to=400, by=-1)
twospec3 <- rbind(wnlist1, twospec2)
longspec3 <- as.data.frame(t(twospec3))
colnames(longspec3) <- c("wn", "c3", "h3")
dev.new()
ggplot(longspec3, aes(x = wn ,y = c3)) +
geom_point(size=0.3)
theme.bw()
#narrow down the candidate
#according to the number of candidate,
#mask/unmask the filtering with scores
names(candi2)(7) <- "score_boxplot"
names(candi2)(8) <- "sand_score"
candi4 <- dplyr::filter(candi2, score_boxplot==3)
#candi4 <- dplyr::filter(candi4, sand_score ==4)
candi4$wn1 <- as.numeric(substr(candi4$wn1,2,5))
candi4$wn2 <- as.numeric(substr(candi4$wn2,2,5))
#trim down the spectral data
cutsite1 <- 4002 - lda_SmallerWnBoundary_wn
cutsite2 <- 4000 - lda_LargerWnBoundary_wn
spec4 <- dplyr::select(twospec3, -c(cutsite1:3601))
spec4 <- dplyr::select(spec4, -c(1:cutsite2))
longspec4 <- as.data.frame(t(spec4))
colnames(longspec4) <- c("wn", "c3", "h3")
longspec4 <- transform(longspec4, target=0)
longspec4 <- transform(longspec4, scoreL=0)
longspec4 <- transform(longspec4, scoreS=0)
longspec4 <- transform(longspec4, lower_abs_anchor=0)
longspec4 <- transform(longspec4, higher_abs_anchor=0)
longspec4 <- transform(longspec4, suffix=NA)
#normalize the spectra
```

```
longspec4$c3 <- (longspec4$c3-min(longspec4$c3))/(max(longspec4$c3)-
min(longspec4$c3))
longspec4$h3 <- (longspec4$h3-min(longspec4$h3))/(max(longspec4$h3)-
min(longspec4$h3))
#mark the target wavenumber
#mark the hih3_a_score_boxplot
nrow_longspec4 <- nrow(longspec4)
nrow_candi4 <- nrow(candi4)
for (i in 1:nrow_longspec4){
 if(longspec4(i,1)==lda_peak_wn){
 longspec4(i,4) <- 0.1
 }

 for (j in 1:nrow_candi4){
 temp_wn1 <- candi4(j,2)
 temp_wn2 <- candi4(j,3)
 if(longspec4(i,1)==temp_wn1){
 longspec4(i,5) <- longspec4(i,5)+0.01
 }
 if(longspec4(i,1)==temp_wn2){
 longspec4(i,6) <- longspec4(i,6)+0.01
 }
 }
}
#save the longspec4 as csv
filename_specanchor <- paste(today2, "_a6_", "specanchor_", lda_peak_wn_string, ".csv",
sep="")
filename2_specanchor <- paste(pathinfo,"/", filename_specanchor, sep="")
write.csv(longspec4,
 filename2_specanchor, row.names=FALSE)
#draw the anchor points
dev.new()
plotanchor4 <- ggplot(longspec4) +
```

```
 theme_light()+
 geom_line(aes(x=wn, y=c3),
 colour="deepskyblue", size=0.3)+
 geom_line(aes(x=wn, y=h3),
 colour="salmon", size=0.3)+
 geom_line(aes(x=wn, y=target),
 colour="salmon", size=0.3)+
 geom_line(aes(x=wn, y=scoreL),
 colour="black", size=0.3)+
 geom_line(aes(x=wn, y=scoreS),
 colour="black", size=0.3)
print(plotanchor4)
#save the plot as png format(you can change to .jpeg, .tiff, etc)
#unit is in inch
filename_plotanchor4 <- paste(today2, "_a6_", "plotanchor4.png", sep="")
filename2_plotanchor4 <- paste(pathinfo,"/", filename_plotanchor4, sep="")
ggsave(file = filename2_plotanchor4,
 plot = plotanchor4, dpi = 100,
 width = 7.2, height = 4.8)
#End of script
```

Script code 6: Boxplot analysis of Fm markers

```
#salma_a7_ftir_marker_boxplot_211225a.r
#a7_ftir marker boxplot
#for salma's 1st paper
#clear the brain
rm(list=ls())
#library to register
#ggplot2 and dplyr are in tidyverse
library(conflicted)
library(tidyverse)
library(MASS)
library(klaR)
library(caret)
```

```
#obtain date information
today <- Sys.Date()
yr <- substr(today, 3,4)
mo <- substr(today, 6,7)
day <- substr(today, 9,10)
today2 <- paste(yr, mo, day, sep="")
#obtain desktop folder information for a windows user
#you must change the string within "xxx" below according to your computer
desktopfolder <- "akash"
#assemble path info
pathinfo <- paste("C:/users/",desktopfolder,"/desktop/", sep="")
#import the 1st, compiled ftir csv data
print("Please specify xxxxxx_a2_specpile_processed.csv")
specpile1 <- file.choose()
specpile2 <- read.csv(specpile1,
 header = T)
#change the variable types of "condition" and "genotype"
#to factor format
specpile2$condition <- factor(specpile2$condition)
specpile2$genotype <- factor(specpile2$genotype)
#import the "a6_specanchor_xxxx_v2.csv" file
 #integer of "1" should be input at col 7 and 8
print("Please specify a6_xxxxxx_specanchor_xxxx_v2.csv")
specanchor1 <- file.choose()
specanchor2 <- read.csv(specanchor1,
 header = T)
#extract wavenumber info for target and anchors
n_specanchor2 <- nrow(specanchor2)
for(i in 1:n_specanchor2){
 if(specanchor2(i,4)==0.1){
 wn_target <- specanchor2(i,1)
 }
 if(specanchor2(i,7)==1){
```

```
 wn_lower_abs_anchor <- specanchor2(i,1)
 }
 if(specanchor2(i,8)==1){
 wn_higher_abs_anchor <- specanchor2(i,1)
 suffix_ancpair <- specanchor2(i,9)
 }
}
wn_target_col <- 4005-wn_target
wn_lower_anchor_col <- 4005-wn_lower_abs_anchor
wn_higher_anchor_col <- 4005-wn_higher_abs_anchor
wn_target_chr <- as.character(wn_target)
#calculate fm value
specpile3 <- mutate(specpile2, fm_numerator=(specpile2(,wn_target_col) -
specpile2(,wn_lower_anchor_col)))
specpile3 <- mutate(specpile3, fm_denominator=(specpile2(,wn_higher_anchor_col) -
specpile2(,wn_lower_anchor_col)))
specpile3 <- mutate(specpile3, fm=(fm_numerator/fm_denominator))
specpile3 <- mutate(specpile3, target_abs=specpile2(,wn_target_col))
specpile3 <- mutate(specpile3, lower_abs_anchor_abs=specpile2(,wn_lower_anchor_col))
specpile3 <- mutate(specpile3, higher_abs_anchor_abs=specpile2(,wn_higher_anchor_col))
specfm1 <- dplyr::select(specpile3, c(1, 2, 3606, 3607, 3608, 3609, 3610, 3611))
#save the fm info as csv
filename_specfm1 <- paste(today2, "_a7_specfm1_", wn_target_chr, suffix_ancpair, ".csv",
sep="")
filename2_specfm1 <- paste(pathinfo,"/", filename_specfm1, sep="")
write.csv(specfm1,
 filename2_specfm1, row.names=FALSE)
#save the second cv
specpile4 <- dplyr::select(specpile2, c(wn_target_col, wn_higher_anchor_col,
wn_lower_anchor_col))
filename_specpile4 <- paste(today2, "_a7_specpile4_", wn_target_chr, suffix_ancpair, ".csv",
sep="")
filename2_specpile4 <- paste(pathinfo,"/", filename_specpile4, sep="")
```

```
write.csv(specpile4,
 filename2_specpile4, row.names=FALSE)
#Reverse version
#calculate fm value for reverse version
specpile3rev <- mutate(specpile2, fm_numerator=(specpile2(,wn_target_col) -
specpile2(,wn_higher_anchor_col)))
specpile3rev <- mutate(specpile3rev, fm_denominator=(specpile2(,wn_lower_anchor_col) -
specpile2(,wn_higher_anchor_col)))
specpile3rev <- mutate(specpile3rev, fm=(fm_numerator/fm_denominator))
specpile3rev <- mutate(specpile3rev, target_abs=specpile2(,wn_target_col))
specpile3rev <- mutate(specpile3rev,
lower_abs_anchor_abs=specpile2(,wn_lower_anchor_col))
specpile3rev <- mutate(specpile3rev,
higher_abs_anchor_abs=specpile2(,wn_higher_anchor_col))
specfm1rev <- dplyr::select(specpile3rev, c(1, 2, 3606, 3607, 3608, 3609, 3610, 3611))
#save the fm info as csv
filename_specfm1rev <- paste(today2, "_a7_specfm1rev_", wn_target_chr, suffix_ancpair,
".csv", sep="")
filename2_specfm1rev <- paste(pathinfo,"/", filename_specfm1rev, sep="")
write.csv(specfm1rev,
 filename2_specfm1rev, row.names=FALSE)
#save the second cv
specpile4rev <- dplyr::select(specpile2, c(wn_target_col, wn_higher_anchor_col,
wn_lower_anchor_col))
filename_specpile4rev <- paste(today2, "_a7_specpile4rev_", wn_target_chr, suffix_ancpair,
".csv", sep="")
filename2_specpile4rev <- paste(pathinfo,"/", filename_specpile4rev, sep="")
write.csv(specpile4rev,
 filename2_specpile4rev, row.names=FALSE)
#make a boxplot
#in the following, "x" should be the grouping variable,
#usually in the category variable, such as condition
#"y" should be numerical variable such as fm.
```

```
#xlab("xxx") is for the label of figure
#for color pallet, check the following
# http://sape.inf.usi.ch/quick-reference/ggplot2/colour
wn_target_label <- paste("fm", wn_target_chr, suffix_ancpair, sep="")
dev.new()
fm_boxplot <- ggplot(specfm1, aes(x = condition, y = fm, fill=condition)) +
 stat_boxplot(geom = "errorbar", width = 0.3)+
 geom_boxplot(outlier.size=1) +
 scale_fill_manual(values=c("deepskyblue", "salmon")) +
 # geom_point(size=0.3, color='lightgray', alpha=0.5) +
 xlab("Condition") +
 ylab(wn_target_label) +
 #if you change the range of y-axis, use the follow line
 # ylim(-20, 20)+
 theme_bw()
print(fm_boxplot)
#save the same fm_boxplot as png file in the desktop
filename_fm_boxplot <- paste(today2, "_a7_", wn_target_label, "_boxplot.png", sep="")
filename2_fm_boxplot <- paste(pathinfo,"/", filename_fm_boxplot, sep="")
ggsave(file = filename2_fm_boxplot,
 plot = fm_boxplot, dpi = 100,
 width = 2.4, height = 2.4)
#reverse version
wn_target_label_rev <- paste("fm", wn_target_chr, suffix_ancpair, "_rev", sep="")
dev.new()
fmrev_boxplot <- ggplot(specfm1rev, aes(x = condition, y = fm, fill=condition)) +
 stat_boxplot(geom = "errorbar", width = 0.3)+
 geom_boxplot(outlier.size=1) +
 scale_fill_manual(values=c("deepskyblue", "salmon")) +
 # geom_point(size=0.3, color='lightgray', alpha=0.5) +
 xlab("Condition") +
 ylab(wn_target_label_rev) +
 #if you change the range of y-axis, use the follow line
```

```
 # ylim(-20, 20)+
 theme_bw()
print(fmrev_boxplot)
#save the same fm_boxplot as png file in the desktop
filename_fmrev_boxplot <- paste(today2, "_a7_", wn_target_label_rev, "_boxplot.png",
sep="")
filename2_fmrev_boxplot <- paste(pathinfo,"/", filename_fmrev_boxplot, sep="")
ggsave(file = filename2_fmrev_boxplot,
 plot = fmrev_boxplot, dpi = 100,
 width = 2.4, height = 2.4)
#End of script
```

Appendix-2

R-scripts for the processing of FT-IR data

Script code 1: FTIR-spectra processing

```
#salma c2.1a spec processing with wn400-4000 offset baseline
#import necessary libraries
library(conflicted)
library(tidyverse)
#clean up the R's brain
rm(list=ls())
#obtain date information
today <- Sys.Date()
yr <- substr(today, 3,4)
mo <- substr(today, 6,7)
day <- substr(today, 9,10)
today2 <- paste(yr, mo, day, sep="")
# !!! system check required, 1 out of 2
#obtain desktop folder information for a windows user
#you must change the string below within "xxx" according to your computer
username <- "akash"
#prepare output folder and its path
DesktopPath <- paste("C:/users/",username,"/desktop/", sep="")
setwd(DesktopPath)
if(!dir.exists(paste(today2, "_specpile/", sep=""))){
 dir.create(paste(today2, "_specpile/", sep=""), recursive=T)
}
OutputPath <- paste(DesktopPath, today2, "_specpile/", sep="")
# !!! system check required, 2 out of 2
#prepare input data folder
#subfolder below the "rawdata" will be ignored
setwd("d:/1_DataFolder/Intel/i04_Informatics_Statistics/i04b_R/trainingdata/ftir_testdata/salma
_testdata/paper3/rawdata/220330_FTIR_rawdata_c3_h3/")
#setwd("d:/1_DataFolder/Intel/i04_Informatics_Statistics/i04b_R/trainingdata/ftir_testdata/salm
```

```
a_testdata/paper3/rawdata/220327_ftir_rawdata_5dates/")
#setwd("d:/1_DataFolder/Intel/i04_Informatics_Statistics/i04b_R/trainingdata/ftir_testdata/salm
a_testdata/paper3/rawdata/01_selected")
#setwd("d:/1_DataFolder/Intel/i04_Informatics_Statistics/i04b_R/trainingdata/ftir_testdata/salm
a_testdata/paper3/rawdata/02_newdata")
pathname_inputfolder <- getwd()
pathname_inputfolder
#create column names for output dataframe
wnlist <- seq(4000, 400, length=3601)
columnname <- c("filename", "condition", "genotype", "identifier", wnlist)
#name the column labels for spec data
specpile <- as.data.frame(t(columnname))
names(specpile) <- columnname
specpile <- slice(specpile, -1)
#prepare dataframe for NG spec
NGspec <- specpile
#input data from .asc file that are generated by Perkin-Elmer
#obtain the filename
#obtain the list of filenames for all csv files,
 #which are stored in the aforementioned "rawdata" folder
filelist <- list.files(path = pathname_inputfolder,
 pattern = "*.asc",
full.names = T)
#count the number of files
fileno <- length(filelist)
#starting a loop for processing data
for (i in 1:fileno){
#for (i in 1:1){
#obtain the new filename
 filename <- basename(filelist(i))
#obtain dataframe
#skip first 25 lines
#the 26th line does not have variable names
```

```
 rawspec <- read.table(filelist(i), skip = 25)
#quick summary
# summary(rawspec)

#plot the spectrum
# ggplot(rawspec, aes(x = V1,y = V2)) +
# geom_point()

#save the wn column for later plotting
 wn_column <- dplyr::select(rawspec, V1)

#exchange rows and columns
#(optional)keep the type as data.frame
 rawspec2 <- as.data.frame(t(rawspec))

#split the rows into wn and spec
 wn_axis <- as.data.frame(rawspec2(1,))
 rawspec3 <- as.data.frame(rawspec2(2,))

#name the column labels for spec data
 names(rawspec3) <- wn_axis
 #smoothing of the spectrum trace
 #below to fill in, but currently skip it

 #obtain the baseline anchors
 #this is a version to take only 4000 and 400
 #the relationships between wn and column_no. is
 #column_no = -wn +4001
 raw400 <- rawspec3(1,3601)
 raw4000 <- rawspec3(1,1)

 #create a baseline data
 #following is the 1st version
```

```
#line is drawn between 4000 to 400
baseline <- seq(raw4000, raw400, length=3601)
#subtract the baseline
spec4_baselined <- (rawspec3 - baseline)

#draw the baseline-corrected spectrum
#1st, exchange the rows and columns
spec4_tall <- t(spec4_baselined)
#combine the wn and spec columns
spec4_tall <- cbind(wn_column, spec4_tall)
#plot the baseline-corrected spectrum
#ggplot(spec4_tall, aes(x = V1,y = V2)) +
# geom_point(size=0.3)
#normalization of spec
#1st, sum of current spec is calculated
sum_signal_original <- sum(dplyr::select(spec4_tall, V2))
#2nd, new column is generated in the spec
#spec values in ppm is calculated
spec5_tall <- dplyr::mutate(spec4_tall, ABS = V2*1000000/sum_signal_original)
#draw the normalized spectrum
#ggplot(spec5_tall, aes(x = V1,y = ABS)) +
# geom_point(size=0.3)
#row-column conversion
spec5 <- as.data.frame(t(spec5_tall))

#remove original data from spec5
spec6 <- dplyr::slice(spec5, 3)
#rownames(spec6) <- filename

#create one column at the top
#add dataname to the 1st column
spec7 <- mutate(spec6, dataname=filename, .before="4000")
```

```
#extract treatment condition info, and add to the 2nd column
condition_info <- substring(filename, 1, 2)
spec8 <- mutate(spec7, condition=condition_info, .after="dataname")
#extract genotype info, and add to the 3rd column
genotype_info <- substring(filename, 3, 5)
spec9 <- mutate(spec8, genotype=genotype_info, .after="condition")

#setup the identifier for later analyses
#and add to the 4th column
identifier_info <- paste(condition_info, genotype_info, sep="")
spec10 <- mutate(spec9, identifier=identifier_info, .after="genotype")

#eliminate, if any, NG data taken by transmission-mode,
#and maintain only OK data with absorption-mode
#the OK data has higher absorbanse at wn3400 than at wn1800
#these wn correspond to col607 and col2207, respectively, in spec10
#if(spec10(1,607)>spec10(1,2207)){
#compiling the data
specpile <- rbind(specpile, spec10)
#} else {
# NGspec <- rbind(NGspec, spec10)
#}

}
#export the data as csv
#data is baseline-corrected, normalized spec
setwd(OutputPath)
a <- getwd()
a
filename_specpile_processed <- paste(today2, "_c2.1a_specpile_offsetbaselined.csv", sep="")
write.csv(specpile,
 filename_specpile_processed, row.names=FALSE)
#export NG data as well
```

```
filename_NGspec_processed <- paste(today2, "_c2.1a_NGspecpile_offsetbaselined.csv", sep="")
write.csv(NGspec,
 filename_NGspec_processed, row.names=FALSE)
#prepare long-format as well, and export
#row-column conversion
long_specpile <- as.data.frame(t(specpile))
long_NGspec <- as.data.frame(t(NGspec))
#create new column at the top
long_specpile <- mutate(long_specpile,
 variable=c("dataname","condition", "genotype", "identifier",
 seq(from=4000, to=400, by=-1)),
 .before=ABS)
long_NGspec <- mutate(long_NGspec,
 variable=c("dataname","condition", "genotype", "identifier",
 seq(from=4000, to=400, by=-1)),
 .before=ABS)
#export the data as csv
#data is baseline-corrected, normalized spec
filename_specpile_processed_longformat <- paste(today2, "_c2.1a_speclong_offsetbaselined.csv",
sep="")
write.csv(long_specpile,
 filename_specpile_processed_longformat, row.names=FALSE)
filename_NGspec_processed_longformat <- paste(today2,
"_c2.1a_NGspeclong_offsetbaselined.csv", sep="")
write.csv(long_NGspec,
 filename_NGspec_processed_longformat, row.names=FALSE)
```

Script code 2: Averaging spectra

```
#c3.1a spec averaging ftir
 #for averaging the spectra
 #This version of script is specific to salma-3 paper
 #for genotype comparison in c3 and h3 condition.
#input file is offset-baselined "specpile_processed.csv"
#import necessary libraries
```

```
library(conflicted)
library(tidyverse)
library(psych)
#clean up the R's brain
rm(list=ls())
#obtain date information
today <- Sys.Date()
yr <- substr(today, 3,4)
mo <- substr(today, 6,7)
day <- substr(today, 9,10)
today2 <- paste(yr, mo, day, sep="")
# !!! system check required, 1 out of 2
#obtain desktop folder information for a windows user
#you must change the string below within "xxx" according to your computer
username <- "akash"
#prepare output folder and its path
DesktopPath <- paste("C:/users/",username,"/desktop/", sep="")
setwd(DesktopPath)
if(!dir.exists(paste(today2, "_drawspec/", sep=""))){
 dir.create(paste(today2, "_drawspec/", sep=""), recursive=T)
}
OutputPath <- paste(DesktopPath, today2, "_drawspec/", sep="")
# !!! system check required, 2 out of 2
#prepare input data folder
setwd("d:/1_DataFolder/Intel/i04_Informatics_Statistics/i04b_R/trainingdata/ftir_testdata/salma
_testdata/paper3/")
a <- getwd()
a
#import the 1st, compiled ftir csv data
print("Please specify c2.1a_specpile_offsetbaselined.csv")
specpile1 <- file.choose()
specpile2 <- read.csv(specpile1, header = T)
#split the data according to their identifiers
```

```
c3chs_specpile2 <- dplyr::filter(specpile2, identifier == 'c3chs')
c3ima_specpile2 <- dplyr::filter(specpile2, identifier == 'c3ima')
c3n61_specpile2 <- dplyr::filter(specpile2, identifier == 'c3n61')
h3chs_specpile2 <- dplyr::filter(specpile2, identifier == 'h3chs')
h3ima_specpile2 <- dplyr::filter(specpile2, identifier == 'h3ima')
h3n61_specpile2 <- dplyr::filter(specpile2, identifier == 'h3n61')
#separate the identifier column (category info)
c3chs_id_1 <- dplyr::select(c3chs_specpile2, (1:4))
c3ima_id_1 <- dplyr::select(c3ima_specpile2, (1:4))
c3n61_id_1 <- dplyr::select(c3n61_specpile2, (1:4))
h3chs_id_1 <- dplyr::select(h3chs_specpile2, (1:4))
h3ima_id_1 <- dplyr::select(h3ima_specpile2, (1:4))
h3n61_id_1 <- dplyr::select(h3n61_specpile2, (1:4))
#extract values used for averaging
c3chs_specmatrix <- dplyr::select(c3chs_specpile2, -(1:4))
c3ima_specmatrix <- dplyr::select(c3ima_specpile2, -(1:4))
c3n61_specmatrix <- dplyr::select(c3n61_specpile2, -(1:4))
h3chs_specmatrix <- dplyr::select(h3chs_specpile2, -(1:4))
h3ima_specmatrix <- dplyr::select(h3ima_specpile2, -(1:4))
h3n61_specmatrix <- dplyr::select(h3n61_specpile2, -(1:4))
#averaging
c3chsmean <- as.data.frame(t(apply(c3chs_specmatrix, 2, mean)))
c3imamean <- as.data.frame(t(apply(c3ima_specmatrix, 2, mean)))
c3n61mean <- as.data.frame(t(apply(c3n61_specmatrix, 2, mean)))
h3chsmean <- as.data.frame(t(apply(h3chs_specmatrix, 2, mean)))
h3imamean <- as.data.frame(t(apply(h3ima_specmatrix, 2, mean)))
h3n61mean <- as.data.frame(t(apply(h3n61_specmatrix, 2, mean)))
specmean <- rbind(c3chsmean, c3imamean, c3n61mean,
 h3chsmean, h3imamean, h3n61mean)
rownames(specmean) <- c("c3chs", "c3ima", "c3n61",
 "h3chs", "h3ima", "h3n61")
#save the specmean as csv file
setwd(OutputPath)
```

```
tempa <- getwd()
tempa
filename_specmean <- paste(today2, "_c3.1a_specmean_offset.csv", sep="")
write.csv(specmean,
 filename_specmean, row.names=TRUE)
#save a long-format of specmean as csv file
specmeanlong <- as.data.frame(t(specmean))
wn <- seq(from=4000, to=400, by=-1)
wn_col <- as.data.frame(wn)
specmeanlong2 <- cbind(wn_col, specmeanlong)
filename_specmeanlong <- paste(today2, "_c3.1a_specmeanlong_offset.csv", sep="")
write.csv(specmeanlong2,
 filename_specmeanlong, row.names=TRUE)
```

Script code 3: Drawing entire spectra

```
#salma_c3.2_draw_spec
#this is to draw averaged or representative spectra for genotype paper
#import necessary libraries
library(conflicted)
library(tidyverse)
library(caret)
library(ggpubr)
#clean up the R's brain
rm(list=ls())
#obtain date information
today <- Sys.Date()
yr <- substr(today, 3,4)
mo <- substr(today, 6,7)
day <- substr(today, 9,10)
today2 <- paste(yr, mo, day, sep="")
# !!! system check required, 1 out of 2
#obtain desktop folder information for a windows user
#you must change the string below within "xxx" according to your computer
username <- "akash"
```

```
#prepare output folder and its path
DesktopPath <- paste("C:/users/",username,"/desktop/", sep="")
setwd(DesktopPath)
if(!dir.exists(paste(today2, "_drawspec/", sep=""))){
 dir.create(paste(today2, "_drawspec/", sep=""), recursive=T)
}
OutputPath <- paste(DesktopPath, today2, "_drawspec/", sep="")
# !!! system check required, 2 out of 2
#prepare input data folder
setwd("d:/1_DataFolder/Intel/i04_Informatics_Statistics/i04b_R/trainingdata/ftir_testdata/salma
_testdata/paper3/")
pathname_inputfolder <- getwd()
pathname_inputfolder
#invoke a file-opening window, specify the input file,
#the input data should be in .csv format,
 #and has to be long-format, with baseline-corrected and normalized
#obtain the filename
print("Please specify xxxxxx_c3.1a_specmeanlong_offset.csv")
print("or, xxxxxx_c3.1b_specmeanlong_piecewise.csv")
print("or, xxxxxx_c3.1a_specrepresentativelong_offsetbaselined.csv")
longinputfile <- file.choose()
filename <- basename(longinputfile)
speclong6 <- read.csv(longinputfile,
 header = T, row.names="X")
filename
#Session 1
#drawing 18 spectra in a single panel
#arrange the spectra data
#twospec2 <- dplyr::select(twospec, -(c(1:1)))
#colnames(twospec2) <- seq(from=4000, to=400, by=-1)
#wnlist1 <- as.data.frame(t(seq(from=4000, to=400, by=-1)))
#colnames(wnlist1) <- seq(from=4000, to=400, by=-1)
#twospec3 <- rbind(wnlist1, twospec2)
```

```
#c3id <- twospec(1,6)
#h3id <- twospec(2,6)
#longspec3 <- as.data.frame(t(twospec3))
#colnames(longspec3) <- c("wn", "c3", "h3")
#draw the overlapped spec
dev.new()
plotspec6 <- ggplot(speclong6) +
 theme_bw()+
 geom_line(aes(x=speclong6(,1), y=speclong6(,2)),
 colour="deepskyblue", size=0.3)+
 geom_line(aes(x=speclong6(,1), y=speclong6(,3)),
 colour="dodgerblue", size=0.3)+
 geom_line(aes(x=speclong6(,1), y=speclong6(,4)),
 colour="dodgerblue4", size=0.3)+
 geom_line(aes(x=speclong6(,1), y=speclong6(,5)),
 colour="salmon", size=0.3)+
 geom_line(aes(x=speclong6(,1), y=speclong6(,6)),
 colour="salmon3", size=0.3)+
 geom_line(aes(x=speclong6(,1), y=speclong6(,7)),
 colour="orangered4", size=0.3)+
 xlab("wavenumber")+
 ylab("ABS")+
 ggtitle(paste(today2,"_6spec",sep=""))
print(plotspec6)
#save the plot as png format(you can change to .jpeg, .tiff, etc)
setwd(OutputPath)
b <- getwd()
b
filename_plotspec6 <- paste(today2, "_c3.2_6spec.png", sep="")
ggsave(file = filename_plotspec6,
 plot = plotspec6, dpi=100,
 width=7.2, height=2.4)
#Session 2
```

```
#draw the upward-stacked spec
speclong6a <- speclong6
speclong6a <- dplyr::mutate(speclong6a, c3chs=c3chs+400)
speclong6a <- dplyr::mutate(speclong6a, c3ima=c3ima+800)
speclong6a <- dplyr::mutate(speclong6a, c3n61=c3n61+1200)
speclong6a <- dplyr::mutate(speclong6a, h3chs=h3chs+1600)
speclong6a <- dplyr::mutate(speclong6a, h3ima=h3ima+2000)
speclong6a <- dplyr::mutate(speclong6a, h3n61=h3n61+2400)
#draw the stacked spec
dev.new()
plotspec6a <- ggplot(speclong6a) +
 theme_bw()+
 geom_line(aes(x=speclong6a(,1), y=speclong6a(,2)),
 colour="deepskyblue", size=0.3)+
 geom_line(aes(x=speclong6a(,1), y=speclong6a(,3)),
 colour="dodgerblue", size=0.3)+
 geom_line(aes(x=speclong6a(,1), y=speclong6a(,4)),
 colour="dodgerblue4", size=0.3)+
 geom_line(aes(x=speclong6a(,1), y=speclong6a(,5)),
 colour="salmon", size=0.3)+
 geom_line(aes(x=speclong6a(,1), y=speclong6a(,6)),
 colour="salmon3", size=0.3)+
 geom_line(aes(x=speclong6a(,1), y=speclong6a(,7)),
 colour="orangered4", size=0.3)+
 xlab("wavenumber")+
 ylab("ABS")+
 ggtitle(paste(today2,"_6stacked_spec",sep=""))
print(plotspec6a)
#save the plot as png format(you can change to .jpeg, .tiff, etc)
setwd(OutputPath)
b <- getwd()
b
filename_plotspec6a <- paste(today2, "_c3.2_6spec_stacked.png", sep="")
```

```
ggsave(file = filename_plotspec6a,
 plot = plotspec6a, dpi=100,
 width=7.2, height=7.2)
#instruction on the change of output filenames
print("For c3.1a_offsetbaselined data, add letter-a after c3.2 to be c3.2a")
print("For c3.1b_pairwisebaselined data, add letter-b after c3.2 to be c3.2b")
```

Script code 4: PCA

```
#c4.1a PCA_total_ftir offset for salma's genotype paper
#imput file should be in csv format,
#typically, "xxxxxx_c2.1a_specpile_offset.csv" would be selected
#1st column should be the names of original spec files
#2nd col should be treatment either c3 or h3
#3rd col should be genotype, either chs, ima, or n61
#4th col should be identifier, which combines the 2nd and 3rd
#then followed by normalized abs values from 4000 to 400
#input data should have 4+3601=3605 columns.
#import necessary libraries
library(conflicted)
library(tidyverse)
library(psych)
#clean up the R's brain
rm(list=ls())
#obtain date information
today <- Sys.Date()
yr <- substr(today, 3,4)
mo <- substr(today, 6,7)
day <- substr(today, 9,10)
today2 <- paste(yr, mo, day, sep="")
#obtain the directory info for R script location
#ScriptPath <- getwd()
#obtain desktop folder information for a windows user
#you must change the string below within "xxx"according to your computer
username <- "akash"
```

```
#prepare output folder and its path
DesktopPath <- paste("C:/users/",username,"/desktop/", sep="")
setwd(DesktopPath)
if(!dir.exists(paste(today2, "_pca/", sep=""))){
 dir.create(paste(today2, "_pca/", sep=""), recursive=T)
}
OutputPath <- paste(DesktopPath, today2, "_pca/", sep="")
#invoke a file-opening window, specify the input file,
 #the input data should be in .csv format,
 #and has to be baseline-corrected and normalized
#obtain the filename
print("Please specify xxxxxx_c2.1a_specpile_offsetbaselined.csv")
inputfile <- file.choose()
filename <- basename(inputfile)
rawspecs <- read.csv(inputfile,
 header = T)
filename
#extract values used for calculation to a new df specmatrix
#columns with zero values (wn4000, wn2600, wn2000, and wn400) should be removed
 #in "rawspecs", wn400 corresponds to col3605, thus the sum becomes 4005
 #thus, wn2000, wn2600, and wn4000 correspond to col
specmatrix <- dplyr::select(rawspecs, -(3605:3605)) #wn400
specmatrix <- dplyr::select(specmatrix, -(1:5)) #id and wn4000
#separate "identifier" column (category info)
id_2 <- dplyr::select(rawspecs, c(1,4))
#further trimming of values in the range of wn3600-4000
specmatrix <- dplyr::select(specmatrix, -(1:399))
#perform pca analysis using prcomp
#prcomp is standard but one of the oldest
pc = prcomp(specmatrix, scale =T)
#using the new 'principal()' in psych package
#pc <- psych::principal(specmatrix, nfactors=3601,
# rotate='none')
```

```
#display the summary
summary(pc)
#preparation of score data output
pc1_score <- pc$x(,1)
pc2_score <- pc$x(,2)
scoreonly <- as.data.frame(pc$x)
score <- cbind(id_2, scoreonly)
#export the score data as csv into the OutputPath
#data is PC1-PC2 score
setwd(OutputPath)
filename_PC12_score <- paste(today2,"_c4.1a_PCA_total_score_offset.csv", sep="")
write.csv(score,
 filename_PC12_score, row.names=FALSE)
#export rotation data as csv file
rotationonly <- as.data.frame(pc$rotation)
filename_pca_rotation <- paste(today2, "_c4.1a_pca_total_rotation_offset.csv", sep="")
write.csv(rotationonly, filename_pca_rotation, row.names=FALSE)
#export sdev data as csv file
sdevonly <- as.data.frame(pc$sdev)
filename_pca_sdev <- paste(today2, "_c4.1a_pca_total_sdev_offset.csv", sep="")
write.csv(sdevonly, filename_pca_sdev, row.names=FALSE)
#calculate the loadings, and export as csv file
loadingdata <- sweep(pc$rotation, MARGIN=2, pc$sdev, FUN="*")
filename_pca_loading <- paste(today2, "_c4.1a_pca_total_loading_offset.csv", sep="")
write.csv(loadingdata,
 filename_pca_loading, row.names=FALSE)
#draw the pc1_pc2 scoreplot
#size of the dots in the plot can be changed
 #by modifying the location of "geom_point(size=)
#color info can be seen in the following website
#http://sape.inf.usi.ch/quick-reference/ggplot2/colour
#point shape can be seen in the following website
#http://www.sthda.com/english/wiki/ggplot2-point-shapes
```

```
dev.new()
pca_total_scoreplot <- ggplot(score, aes(x = PC1,y = PC2,
 shape = identifier, color = identifier)) +
 geom_point(size=1) +
 scale_color_manual(values=c("deepskyblue","dodgerblue","dodgerblue4",
 "salmon","salmon3","orangered")) +
 scale_shape_manual(values=c(1,2,4,1,2,4))+
# scale_size_manual(values=c(10,1,1,10,1,1,10,1,1,10,1,1,10,1,1))+
 theme_bw()
print(pca_total_scoreplot)
#save the plot as png format
#you can change to .jpeg, .tiff, etc
#unit is in inch
filename_pca_total_scoreplot <- paste(today2, "_c4.1a_PCA_total_scoreplot_offset.png", sep="")
ggsave(file = filename_pca_total_scoreplot,
 plot = pca_total_scoreplot, dpi=600,
 width=7.2, height=4.8)
#Section 2: contribution check
#extract contribution data
contribution_total <- as.data.frame(t(summary(pc)$importance))
names(contribution_total) <-
c("standard_deviation","proportion_of_variance","cumulative_proportion")
contri_total_rownames <- as.data.frame(rownames(contribution_total))
names(contri_total_rownames) <- "PC"
contribution_total <- cbind(contribution_total, contri_total_rownames)
#save the contribution data as csv file
filename_pca_total_contribution <- paste(today2, "_c4.1a_pca_total_contribution_offset.csv",
sep="")
write.csv(contribution_total,
 filename_pca_total_contribution, row.names=FALSE)
#extract top 9 from the contribution data, and save as a png file
top9_contribution_total <- dplyr::slice(contribution_total, 1:9)
contribution_total_plot <- ggplot(top9_contribution_total, aes(x = PC, y =
```

```
proportion_of_variance)) +
 geom_bar(stat="identity", fill="forestgreen") +
theme_bw()
print(contribution_total_plot)
#save the contribution graph
filename_pca_contribution_total_plot <- paste(today2,
"_c4.1a_PCA_contribution_total_offset.png", sep="")
ggsave(file = filename_pca_contribution_total_plot,
 plot = contribution_total_plot, dpi=100,
 width=7.2, height=4.8)
#Section 3: Loading plot for all the 6 groups
#prepare data for loading plot
pc_loading <- data.frame(t(cor(pc$x,specmatrix)))
pc_score <- data.frame(pc$x)
#draw the 2D-loading plot using ggplot
#color info can be seen in the following website
#http://sape.inf.usi.ch/quick-reference/ggplot2/colour
g0 <- ggplot()
g0 <- g0 + geom_segment(data=pc_loading,
 aes(x=0,y=0,xend=(PC1*1),yend=(PC2*1)),
colour=rainbow(3200),alpha=0.2,size=0.5)
g0 <- g0 + xlab("PC1")
g0 <- g0 + ylab("PC2")
g0 <- g0 + theme_bw()
print(g0)
filename_pca_loading2d_plot <- paste(today2, "_c4.1a_PCA_loading2d_plot_offset.png", sep="")
ggsave(file = filename_pca_loading2d_plot,
 plot = g0, dpi=300, width=6.0, height=6.0)
#draw the 1D-loading barplot
#prepare x-axis data
wn_x_axis <- as.data.frame(seq(3600,401, by=-1))
names(wn_x_axis) <- c("wn")
pc_loading2 <- cbind(wn_x_axis, pc_loading)
```

```
#draw the PC1 loading barplot
g1 <- ggplot(data=pc_loading2,
 aes(x=wn, y=PC1))
g1 <- g1 + geom_bar(stat="identity", col=rainbow(3200))
g1 <- g1 + theme_bw()
print(g1)
filename_pca_PC1_loadingplot <- paste(today2, "_c4.1a_PCA_PC1_loadingplot_offset.png",
sep="")
ggsave(file = filename_pca_PC1_loadingplot,
 plot = g1, dpi=300, width=6.0, height=1.5)
#draw the PC2 loading barplot
g2 <- ggplot(data=pc_loading2,
 aes(x=wn, y=PC2))
g2 <- g2 + geom_bar(stat="identity", col=rainbow(3200))
g2 <- g2 + theme_bw()
print(g2)
filename_pca_PC2_loadingplot <- paste(today2, "_c4.1a_PCA_PC2_loadingplot_offset.png",
sep="")
ggsave(file = filename_pca_PC2_loadingplot,
 plot = g2, dpi=300, width=6.0, height=1.5)
```

Script code 5: FTIR marker boxplot

```
#c5.1_ftir marker boxplot
 #for salma's genotype paper
 #the 6 markers from the 1st paper are applied to the genotype data
#clear the brain
rm(list=ls())
#library to register
#ggplot2 and dplyr are in tidyverse
library(conflicted)
library(tidyverse)
library(MASS)
library(klaR)
library(caret)
```

```
library(ggpubr)
#obtain date information
today <- Sys.Date()
yr <- substr(today, 3,4)
mo <- substr(today, 6,7)
day <- substr(today, 9,10)
today2 <- paste(yr, mo, day, sep="")
#obtain desktop folder information for a windows user
#you must change the string below within "xxx"according to your computer
username <- "akash"
#prepare output folder and its path
DesktopPath <- paste("C:/users/",username,"/desktop/", sep="")
setwd(DesktopPath)
if(!dir.exists(paste(today2, "_MarkerBoxplot/", sep=""))){
 dir.create(paste(today2, "_MarkerBoxplot/", sep=""), recursive=T)
}
OutputPath <- paste(DesktopPath, today2, "_MarkerBoxplot/", sep="")
#redirect working directory and import the compiled ftir csv data
setwd("d:/1_DataFolder/Intel/i04_Informatics_Statistics/i04b_R/trainingdata/ftir_testdata/salma
_testdata/paper3/")
print("Please specify xxxxxx_specpile_offsetbaselined.csv")
specpile1 <- file.choose()
specpile2 <- read.csv(specpile1,
 header = T)
#change the variable types of "condition" and "genotype"
#to factor format
specpile2$condition <- factor(specpile2$condition)
specpile2$genotype <- factor(specpile2$genotype)
specpile2$identifier <- factor(specpile2$identifier)
#import the "a6_anchorinfo.csv" file
 #integer of "1" should be input at col 7 and 8
print("Please specify anchor_info.csv")
specanchor1 <- file.choose()
```

```
specanchor2 <- read.csv(specanchor1,
 header = T)
#obtain info on number of target markers
n_specanchor2 <- nrow(specanchor2)
#(section 0): separation of each spec pair
#extract values used for calculation to a new df specmatrix
#in "specpile2", wn400 corresponds to col3605, thus the sum becomes 4005
#thus, wn4000 correspond to col
#split the data according to their identifiers
c3chs_specpile2 <- dplyr::filter(specpile2, identifier == 'c3chs')
c3ima_specpile2 <- dplyr::filter(specpile2, identifier == 'c3ima')
c3n61_specpile2 <- dplyr::filter(specpile2, identifier == 'c3n61')
h3chs_specpile2 <- dplyr::filter(specpile2, identifier == 'h3chs')
h3ima_specpile2 <- dplyr::filter(specpile2, identifier == 'h3ima')
h3n61_specpile2 <- dplyr::filter(specpile2, identifier == 'h3n61')
#combine the c3-h3 pairs
chs_spec <- rbind(c3chs_specpile2, h3chs_specpile2)
ima_spec <- rbind(c3ima_specpile2, h3ima_specpile2)
n61_spec <- rbind(c3n61_specpile2, h3n61_specpile2)
#make list of id-dataframe and id-name
list_spec_pair <- list(chs_spec, ima_spec, n61_spec)
list_chr_pair <- list("chs", "ima", "n61")
#make empty output dataframe
fmpileall <- data.frame(matrix(rep(NA, 8), nrow=1))(numeric(0),)
colnames(fmpileall) <- c("dataname", "identifier", "fm_numerator", "fm_denominator",
 "fm", "abs_target", "abs_anchor1", "abs_anchor2")
#i-loop for different Fm markers
#for(i in 1:1){
for(i in 1:n_specanchor2){

 target_wn <- specanchor2(i,1)
 anchor1_wn <- specanchor2(i,2)
 anchor2_wn <- specanchor2(i,3)
```

```
ylim_min <- specanchor2(i,7)
ylim_max <- specanchor2(i,8)
#specpile2 has 6 chr col before wn4000
#thus, wn4000 corresponds to col7,
#and wn400 corresponds to col3607
#sum of these two becomes 4007
col_target_wn <- 4005 - target_wn
col_anchor1_wn <- 4005 - anchor1_wn
col_anchor2_wn <- 4005 - anchor2_wn

#prepare list for plots
bplotlist <- list()

#make empty output dataframe
fmpilesub <- data.frame(matrix(rep(NA, 8), nrow=1))(numeric(0),)
colnames(fmpilesub) <- c("dataname", "identifier", "fm_numerator", "fm_denominator",
"fm", "abs_target", "abs_anchor1", "abs_anchor2")

#j-loop for different genotype
for(j in 1:3){
#for(j in 1:1){

#calling working tissue data
wspecpile <- list_spec_pair(j)
wspecpile <- as.data.frame(wspecpile)
wpairchr <- list_chr_pair(j)
wpairchr <- unlist(wpairchr)

#calculate fm value
wspecpile2 <- dplyr::mutate(wspecpile, fm_numerator=(wspecpile(,col_target_wn) -
wspecpile(,col_anchor1_wn)))
wspecpile2 <- dplyr::mutate(wspecpile2, fm_denominator=(wspecpile2(,col_anchor2_wn) -
wspecpile2(,col_anchor1_wn)))
```

```
wspecpile2 <- dplyr::mutate(wspecpile2, fm=(fm_numerator/fm_denominator))
wspecpile2 <- dplyr::mutate(wspecpile2, abs_target=wspecpile2(,col_target_wn))
wspecpile2 <- dplyr::mutate(wspecpile2, abs_anchor1=wspecpile2(,col_anchor1_wn))
wspecpile2 <- dplyr::mutate(wspecpile2, abs_anchor2=wspecpile2(,col_anchor2_wn))
#colnames(specpile3)((3608:3613)) <- c("fm_numerator", "fm_denominator", "fm"
# "abs_target",xxxxxx)
wspecpile3 <- dplyr::select(wspecpile2, c(1, 4, 3606, 3607, 3608, 3609, 3610, 3611))

fmpileall <- rbind(fmpileall, wspecpile3)
fmpilesub <- rbind(fmpilesub, wspecpile3)

#save the wspecpile3 as csv
filename_wspecpile3 <- paste(today2, "_c5.1_fm", target_wn, "_", wpairchr, ".csv", sep="")
setwd(OutputPath)
atemp <- getwd()
atemp
write.csv(wspecpile3,
filename_wspecpile3, row.names=FALSE)

#make a boxplot
#in the following, "x" should be the grouping variable,
#usually in the category variable, such as condition
#"y" should be numerical variable such as fm.
#xlab("xxx") is for the label of figure
#for color pallet, check the following
# http://sape.inf.usi.ch/quick-reference/ggplot2/colour
boxplot_title <- paste("fm", target_wn, "_", wpairchr, sep="")
#dev.new()
fm_boxplot <- ggplot(wspecpile3, aes(x = identifier, y = fm, fill=identifier)) +
stat_boxplot(geom = "errorbar", width = 0.3)+
geom_boxplot(outlier.size=1) +
scale_fill_manual(values=c("deepskyblue", "salmon")) +
# geom_point(size=0.3, color='lightgray', alpha=0.5) +
```

```
xlab("Condition") +
ylab(boxplot_title) +
#if you change the range of y-axis, use the follow line
# ylim(-20, 20)+
theme_bw()
#print(fm_boxplot)

#save the same fm_boxplot as png file
filename_fm_boxplot <- paste(today2, "_c6.1a_", boxplot_title, "_boxplot.png", sep="")
setwd(OutputPath)
atemp <- getwd()
atemp
ggsave(file = filename_fm_boxplot,
plot = fm_boxplot, dpi = 100,
width = 2.4, height = 2.4)

bplotlist((j))<- fm_boxplot

#end of j-loop
}

#k-loop for assembling 9 boxplots in 3x3 format in 1 figure
for (k in 1:9){
allplot <- ggarrange(plotlist=c(bplotlist(1),
bplotlist(2), bplotlist(3),
bplotlist(4), bplotlist(5),
bplotlist(6), bplotlist(7),
bplotlist(8), bplotlist(9)),
nrow=3, ncol=3, align="hv")

# print allplot
setwd(OutputPath)
tempa <- getwd()
```

```
tempa
filename_allplot <- paste(today2, "_c5.1_fm", target_wn,
"_9boxplot.png", sep="")
ggsave(file = filename_allplot,
plot = allplot, dpi=100,
width=14.4, height=7.2)

#draw 3 subsets in one horizontal plot
boxplot_title <- paste("fm", target_wn, sep="")
fmpilesub2 <- transform(fmpilesub, identifier=factor(identifier,
levels=c("c3chs", "h3chs", "c3ima","h3ima", "c3n61", "h3n61")))
#dev.new()
fm_horizonplot <- ggplot(fmpilesub2, aes(x = identifier, y = fm, fill=identifier)) +
stat_boxplot(geom = "errorbar", width = 0.3)+
geom_boxplot(outlier.size=1) +
scale_fill_manual(values=c("deepskyblue", "salmon",
"dodgerblue", "salmon2", "dodgerblue4", "salmon4")) +
# geom_point(size=0.3, color='lightgray', alpha=0.5) +
xlab("Condition") +
ylab(boxplot_title) +
#if you change the range of y-axis, use the follow line
ylim(ylim_min, ylim_max)+
theme_bw()
#print(fm_horizonplot)

#save the same fm_boxplot as png file
filename_fm_horizonplot <- paste(today2, "_c5.1_", boxplot_title, "_HorizonPlot.png", sep="")
setwd(OutputPath)
atemp <- getwd()
atemp
ggsave(file = filename_fm_horizonplot,
plot = fm_horizonplot, dpi = 100,
width = 7.2, height = 4.8)
```

```
 #end of k-loop
 }

}
```

Script code 6: t-test for FTIR marker

```
#c5.2_t-test for ftir marker
 #for salma's genotype paper
 #results from c5.1_ftir_marker_boxplot are used
#clear the brain
rm(list=ls())
#library to register
#ggplot2 and dplyr are in tidyverse
library(conflicted)
library(tidyverse)
library(MASS)
library(klaR)
library(caret)
library(ggpubr)
#obtain date information
today <- Sys.Date()
yr <- substr(today, 3,4)
mo <- substr(today, 6,7)
day <- substr(today, 9,10)
today2 <- paste(yr, mo, day, sep="")
#obtain desktop folder information for a windows user
#you must change the string below within "xxx"according to your computer
username <- "akash"
#prepare output folder and its path
DesktopPath <- paste("C:/users/",username,"/desktop/", sep="")
setwd(DesktopPath)
if(!dir.exists(paste(today2, "_t-test/", sep=""))){
 dir.create(paste(today2, "_t-test/", sep=""), recursive=T)
```

```
}
OutputPath <- paste(DesktopPath, today2, "_t-test/", sep="")
#redirect working directory and import the compiled ftir csv data
setwd("d:/1_DataFolder/Intel/i04_Informatics_Statistics/i04b_R/trainingdata/ftir_testdata/salma
_testdata/paper3/ttest/")
print("Please specify xxxxxx_c5.1_fmxxx_genotype.csv")
df1 <- file.choose()
df2 <- read.csv(df1, header = T)
#perform t-test
t.test(fm ~ identifier, data=df2)
```

Script code 7: LDA

```
#c6.1 lda 2D genotype with offset baseline(400-4000) spec
 #linear discriminant analysis of ftir spectra
 #train-test sets were not prepared, and all data is used for modeling.
 #calculation using c3-h3 data in three genotypes
#this is a version for single baseline data
#clear the brain
rm(list=ls())
#library to register
#ggplot2 and dplyr are in tidyverse
library(conflicted)
library(tidyverse)
library(MASS)
library(klaR)
library(caret)
library(psych)
library(maptools)
library(ggrepel)
#obtain date information
today <- Sys.Date()
yr <- substr(today, 3,4)
mo <- substr(today, 6,7)
day <- substr(today, 9,10)
```

```
today2 <- paste(yr, mo, day, sep="")
# !!! system check required, 1 out of 2
#obtain desktop folder information for a windows user
#you must change the string below within "xxx" according to your computer
username <- "akash"
#prepare output folder and its path
DesktopPath <- paste("C:/users/",username,"/desktop/", sep="")
setwd(DesktopPath)
if(!dir.exists(paste(today2, "_lda2d/", sep=""))){
 dir.create(paste(today2, "_lda2d/", sep=""), recursive=T)
}
OutputPath <- paste(DesktopPath, today2, "_lda2d/", sep="")
# !!! system check required, 2 out of 2
#prepare input data folder
setwd("d:/1_DataFolder/Intel/i04_Informatics_Statistics/i04b_R/trainingdata/ftir_testdata/salma
_testdata/paper3/")
tempa <- getwd()
tempa
#invoke a file-opening window, specify the input file,
#the input data should be in .csv format,
 #and has to be baseline-corrected and normalized
#obtain the filename
print("Please specify xxxxxx_c2.1a_specpile_offsetbaselined.csv")
spec1 <- file.choose()
specpile <- read.csv(spec1,
 header = T)
#extract values used for calculation to a new df specmatrix
#columns with zero values (wn400, 4000) should be removed
#moreover, data around noisy region of wn4000-3600 is also removed
#in "specpile", wn400 corresponds to col3605, thus the sum becomes 4005
#likewise, wn4000 correspond to col5
specpile2 <- dplyr::select(specpile, -(3605:3605)) #wn400
#specpile2 <- dplyr::select(specpile2, -(2007:2007)) #wn2000
```

```
#specpile2 <- dplyr::select(specpile2, -(1407:1407)) #wn2600
specpile2 <- dplyr::select(specpile2, -(5:405)) #wn4000-3600
#separate the identifier column (category info)
id_1 <- dplyr::select(specpile2, (1:4))
#extract values used for lda
specmatrix <- dplyr::select(specpile2, -(1:3))
#set the seednumber for randomness
set.seed(101)
#perform linear discriminant analysis
lda_specmatrix <- lda(identifier ~ ., specmatrix)

#calculate results
#1st, transform them to the values
#then one dimensional histograms
#"mar" is the margin of bottom, left, top, right
lda_specmatrix_results <- predict(lda_specmatrix)
#convert the $x score data to dataframe
ld_score <- data.frame(identifier=id_1(,4), lda=lda_specmatrix_results$x)
#draw 2D scatter plot using LD1-LD2 plain
dev.new()
g1 <- ggplot(ld_score, aes(x=lda.LD1, y=lda.LD2, colour=identifier, shape=identifier))+
 geom_point()+
 scale_color_manual(values=c("deepskyblue","dodgerblue","dodgerblue4",
 "salmon","salmon3","orangered")) +
 scale_shape_manual(values=c(1,2,4,1,2,4))+
 theme_bw()
print(g1)
#save the plot as png format
setwd(OutputPath)
atemp <- getwd()
atemp
filename_lda2d <- paste(today2, "_c6.1_LDA2D.png", sep="")
ggsave(file = filename_lda2d,
```

```
 plot = g1, dpi=600,
 width=4.2, height=3.2)
#save the score data as csv file
ld_scoreonly <- dplyr::select(ld_score, -(1:1))
ld_score2 <- cbind(id_1, ld_scoreonly)
filename_ld_score2 <- paste(today2, "_c6.1_lda_score.csv", sep="")
setwd(OutputPath)
atemp <- getwd()
atemp
write.csv(ld_score2,
 filename_ld_score2, row.names=FALSE)

#extract LDA scalingdata
scaling1 <- lda_specmatrix$scaling

#transform LDA scaling to dataframe
#add wavenumber info
scaling2 <- as.data.frame(t(scaling1))
wnlist <- seq(from=3599, to=401, by=-1)
# wnlist_frag1 <- seq(from=3599, to=2601, by=-1)
# wnlist_frag2 <- seq(from=2599, to=2001, by=-1)
# wnlist_frag3 <- seq(from=1999, to=401, by=-1)
# wnlist <- c(wnlist_frag1, wnlist_frag2, wnlist_frag3)
wnlist2 <- as.data.frame(t(wnlist))

colnames(scaling2) <- wnlist
colnames(wnlist2) <- wnlist

scaling3 <- rbind(wnlist2, scaling2)
scaling4 <- as.data.frame(t(scaling3))
names(scaling4)(1) <- "wavenumber"

#change the wavenumber in ascending order, and save it as csv
```

```
scaling5 <- arrange(scaling4, wavenumber)
filename_scaling5 <- paste(today2, "_c6.1_lda_scaling.csv", sep="")
write.csv(scaling5,
 filename_scaling5, row.names=FALSE)

#plot LD1 scaling, scatter plot version
dev.new()
lda_scaling_LD1_scatterplot <- ggplot(scaling5, aes(x = wavenumber, y = LD1)) +
 geom_point(size=0.5) +
 theme_bw()
 print(lda_scaling_LD1_scatterplot)

 filename_lda_scaling_LD1_scatterplot <- paste(today2, "_c6.1_Scaling_LD1_ScatterPlot.png",
sep="")
 ggsave(file = filename_lda_scaling_LD1_scatterplot,
 plot = lda_scaling_LD1_scatterplot, dpi = 300,
 width = 7.2, height = 2.4)

#plot LD1 scaling, rainbow line plot version
 dev.new()
 gscale_ld1_rb <- ggplot(data=scaling5,
 aes(x = wavenumber, y = LD1))
 gscale_ld1_rb <- gscale_ld1_rb + geom_bar(stat="identity", col=rainbow(3199))
 gscale_ld1_rb <- gscale_ld1_rb + theme_bw()
 print(gscale_ld1_rb)

 filename_gscale_ld1_rb <- paste(today2, "_c6.1_Scaling_LD1_RainbowLinePlot.png", sep="")
 ggsave(file = filename_gscale_ld1_rb,
 plot = gscale_ld1_rb, dpi = 300,
 width = 7.2, height = 2.4)

#plot LD2 scaling, scatter plot version
dev.new()
```

```
lda_scaling_LD2_scatterplot <- ggplot(scaling5, aes(x = wavenumber, y = LD2)) +
 geom_point(size=0.5) +
 theme_bw()
 print(lda_scaling_LD2_scatterplot)

 filename_lda_scaling_LD2_scatterplot <- paste(today2, "_c6.1_Scaling_LD2_ScatterPlot.png",
sep="")
 ggsave(file = filename_lda_scaling_LD2_scatterplot,
 plot = lda_scaling_LD2_scatterplot, dpi = 300,
 width = 7.2, height = 2.4)

#plot LD2 scaling, rainbow line plot version
 dev.new()
 gscale_ld2_rb <- ggplot(data=scaling5,
 aes(x = wavenumber, y = LD2))
 gscale_ld2_rb <- gscale_ld2_rb + geom_bar(stat="identity", col=rainbow(3199))
 gscale_ld2_rb <- gscale_ld2_rb + theme_bw()
 print(gscale_ld2_rb)

 filename_gscale_ld2_rb <- paste(today2, "_c6.1_Scaling_LD2_RainbowLinePlot.png", sep="")
 ggsave(file = filename_gscale_ld2_rb,
 plot = gscale_ld2_rb, dpi = 300,
 width = 7.2, height = 2.4)

 #plot LD1-LD2 scaling 2D plot
 dev.new()
 g2d <- ggplot()
 g2d <- g2d + geom_point(data=scaling5,
 aes(x = LD1, y = LD2),
 colour=rainbow(3199),
 alpha=0.8, size=2)
 g2d <- g2d + labs()
 g2d <- g2d + theme_bw()
```

```
 print(g2d)


#save the plot as png format
 filename_g2d <- paste(today2, "_c6.1_Scaling_2D_LD12_ScatterPlot.png", sep="")
 ggsave(file = filename_g2d,
 plot = g2d, dpi=600,
 width=7.2, height=4.8)
```

Script code 8: Drawing magnified spectra in the vicinity of Fm marker

```
#drawspec_markervicinity for salma's genotype paper
 #for drawing spectrum in the vicinity of ftir-marker
 #this is for Salma's data on c3-h3 chamber comparison.
#input file is "specmean.csv"
#clear the brain
rm(list=ls())
#library to register
#ggplot2 and dplyr are in tidyverse
library(conflicted)
library(tidyverse)
library(MASS)
library(klaR)
library(caret)
library(ggpubr)
#obtain date information
today <- Sys.Date()
yr <- substr(today, 3,4)
mo <- substr(today, 6,7)
day <- substr(today, 9,10)
today2 <- paste(yr, mo, day, sep="")
# !!! system check required, 1 out of 2
#obtain desktop folder information for a windows user
#you must change the string below within "xxx" according to your computer
username <- "akash"
#prepare output folder and its path
```

```
DesktopPath <- paste("C:/users/",username,"/desktop/", sep="")
setwd(DesktopPath)
if(!dir.exists(paste(today2, "_drawspec_markervicinity/", sep=""))){
 dir.create(paste(today2, "_drawspec_markervicinity/", sep=""), recursive=T)
}
OutputPath <- paste(DesktopPath, today2, "_drawspec_markervicinity/", sep="")
# !!! system check required, 2 out of 2
#prepare input data folder
setwd("d:/1_DataFolder/Intel/i04_Informatics_Statistics/i04b_R/trainingdata/ftir_testdata/salma
_testdata/paper3")
a <- getwd()
a
#import a pair of spec ftir data
 #it should be 3602 cols, 1st col is an identifier, then wn4000-400
print("Please specify xxxxxx_c3.1a_specmean.csv")
specpile1 <- file.choose()
sixspec <- read.csv(specpile1,
 header = T)
#import the 2nd, "xxxxxx_newmarker_info.csv" file
print("Please specify xxxxxx_newmarker_info.csv")
anchor_info <- file.choose()
newmarker_info2 <- read.csv(anchor_info,
 header = T)
#arrange the spectra
sixspec2 <- dplyr::select(sixspec, -(c(1:1)))
colnames(sixspec2) <- seq(from=4000, to=400, by=-1)
wnlist1 <- as.data.frame(t(seq(from=4000, to=400, by=-1)))
colnames(wnlist1) <- seq(from=4000, to=400, by=-1)
sixspec3 <- rbind(wnlist1, sixspec2)
longspec3 <- as.data.frame(t(sixspec3))
colnames(longspec3) <- c("wn", "c3chs", "c3ima", "c3n61",
 "h3chs", "h3ima", "h3n61")
#draw the entire spec
```

```
#color info can be seen in the following website
#http://sape.inf.usi.ch/quick-reference/ggplot2/colour
dev.new()
glongspec3 <- ggplot(longspec3) +
 theme_bw()+
 geom_line(aes(x=wn, y=c3chs),
 colour="deepskyblue", size=0.3)+
 geom_line(aes(x=wn, y=c3ima),
 colour="springgreen3", size=0.3)+
 geom_line(aes(x=wn, y=c3n61),
 colour="dodgerblue4", size=0.3)+
 geom_line(aes(x=wn, y=h3chs),
 colour="salmon", size=0.3)+
 geom_line(aes(x=wn, y=h3ima),
 colour="deeppink", size=0.3)+
 geom_line(aes(x=wn, y=h3n61),
 colour="orangered4", size=0.3)
 print(glongspec3)
#save the plot as png format(you can change to .jpeg, .tiff, etc)
setwd(OutputPath)
b <- getwd()
b
filename_glongspec3 <- paste(today2, "_c7.1_longspec3.png", sep="")
ggsave(file = filename_glongspec3,
 plot = glongspec3, dpi=100,
 width=3.6, height=1.2)
#set the target and anchor wavenumbers
nrow_newmarker_info2 <- nrow(newmarker_info2)
#prepare list for plots
plotrawtrace <- list()
plotlistnorm <- list()
plotlistnorm2 <- list()
#i-loop for drawing respective trace around target wavenumbers
```

```
 #in the parameters below, wn_anchor1 and 2 are
 #the ones with lower and higher absorbance (or valley and peak)
for (i in 1:nrow_newmarker_info2){
#for (i in 1:1){
 wn_target1 <- newmarker_info2(i,1)
 wn_target2 <- newmarker_info2(i,2)
wn_target3 <- newmarker_info2(i,3)
 wn_target4 <- newmarker_info2(i,4)
 wn_anchor1 <- newmarker_info2(i,5)
 wn_anchor2 <- newmarker_info2(i,6)
 wn_Ledge <- newmarker_info2(i,7)
 wn_Hedge <- newmarker_info2(i,8)

 #obtaining row numbers for the target and its frame edges in longspec3
 row_target1 <- 4001 - wn_target1
 row_target2 <- 4001 - wn_target2
 row_target3 <- 4001 - wn_target3
 row_anchor1 <- 4001 - wn_anchor1
 row_anchor2 <- 4001 - wn_anchor2
 row_Ledge <- 4001 - wn_Ledge
 row_Hedge <- 4001 - wn_Hedge
 # (Option 1) raw trace without anchoring
 #slice the rows for the magnified frame
 specmag <- dplyr::slice(longspec3, row_Hedge:row_Ledge)

 #save the specmag data as csv
 setwd(OutputPath)
 b <- getwd()
 b
 filename_specmag <- paste(today2, "_c7.1_specraw_", wn_target1, "_", wn_target2, ".csv",
sep="")
 write.csv(specmag, filename_specmag, row.names=FALSE)
 #draw the sliced region of the spec
```

```
dev.new()
plotspecmag <- ggplot(specmag) +
theme_bw()+
geom_line(aes(x=wn, y=c3chs),
colour="deepskyblue", size=0.3)+
geom_line(aes(x=wn, y=c3ima),
colour="springgreen3", size=0.3)+
geom_line(aes(x=wn, y=c3n61),
colour="dodgerblue4", size=0.3)+
geom_line(aes(x=wn, y=h3chs),
colour="salmon", size=0.3)+
geom_line(aes(x=wn, y=h3ima),
colour="deeppink", size=0.3)+
geom_line(aes(x=wn, y=h3n61),
colour="orangered4", size=0.3)+
geom_vline(xintercept=wn_target1,
colour="orange",size=0.3)+
geom_vline(xintercept=wn_target2,
colour="green",size=0.3)+
geom_vline(xintercept=wn_target3,
colour="blue",size=0.3)+
geom_vline(xintercept=wn_target4,
colour="magenta",size=0.3)+
# geom_vline(xintercept=wn_anchor1,
# colour="magenta", size=0.3)+
# geom_vline(xintercept=wn_anchor2,
# colour="blue", size=0.3)+
theme(axis.text.x=element_text(angle=45, hjust=1))
print(plotspecmag)
plotrawtrace((i)) <- plotspecmag
#save the plot as png format
setwd(OutputPath)
b <- getwd()
```

```
b
filename_plotspecmag <- paste(today2, "_c7.1_specraw_",wn_target1, "_", wn_target2, ".png",
sep="")
ggsave(file = filename_plotspecmag,
plot = plotspecmag, dpi=300,
width=3.6, height=3.6)
# (Option 2) normalized trace using 2 anchors

#extract the values for anchors 1 and 2
abs_c3chs_anchor1 <- longspec3(row_anchor1,2)
abs_c3chs_anchor2 <- longspec3(row_anchor2,2)
abs_c3ima_anchor1 <- longspec3(row_anchor1,3)
abs_c3ima_anchor2 <- longspec3(row_anchor2,3)
abs_c3n61_anchor1 <- longspec3(row_anchor1,4)
abs_c3n61_anchor2 <- longspec3(row_anchor2,4)
abs_h3chs_anchor1 <- longspec3(row_anchor1,5)
abs_h3chs_anchor2 <- longspec3(row_anchor2,5)
abs_h3ima_anchor1 <- longspec3(row_anchor1,6)
abs_h3ima_anchor2 <- longspec3(row_anchor2,6)
abs_h3n61_anchor1 <- longspec3(row_anchor1,7)
abs_h3n61_anchor2 <- longspec3(row_anchor2,7)

#calculate the normalized absorbance (nabs) using Fm formula
#the formula is (A_target - A_anchor1)/(A_anchor2 - A_anchor1)

specmag2 <- mutate(specmag, nabs_c3chs=(c3chs-abs_c3chs_anchor1)/(abs_c3chs_anchor2-
abs_c3chs_anchor1))
specmag2 <- mutate(specmag2, nabs_c3ima=(c3ima-abs_c3ima_anchor1)/(abs_c3ima_anchor2-
abs_c3ima_anchor1))
specmag2 <- mutate(specmag2, nabs_c3n61=(c3n61-abs_c3n61_anchor1)/(abs_c3n61_anchor2-
abs_c3n61_anchor1))
specmag2 <- mutate(specmag2, nabs_h3chs=(h3chs-abs_h3chs_anchor1)/(abs_h3chs_anchor2-
abs_h3chs_anchor1))
```

```
 specmag2 <- mutate(specmag2, nabs_h3ima=(h3imaabs_h3ima_anchor1)/(abs_h3ima_anchor2-
abs_h3ima_anchor1))
 specmag2 <- mutate(specmag2, nabs_h3n61=(h3n61-
abs_h3n61_anchor1)/(abs_h3n61_anchor2-abs_h3n61_anchor1))
 #save the specmag2 data as csv
 setwd(OutputPath)
 b <- getwd()
 b
 filename_specmag2 <- paste(today2, "_c7.1_specmag2_", wn_target1, "_", wn_target2, ".csv",
sep="")
 write.csv(specmag2, filename_specmag2, row.names=FALSE)

 #draw the normalized spec
 dev.new()
 plotspecmag2n <- ggplot(specmag2) +
 theme_bw()+
 geom_line(aes(x=wn, y=nabs_c3chs),
 colour="deepskyblue", size=0.3)+
 geom_line(aes(x=wn, y=nabs_c3ima),
 colour="deepskyblue2", size=0.3)+
 geom_line(aes(x=wn, y=nabs_c3n61),
 colour="deepskyblue4", size=0.3)+
 geom_line(aes(x=wn, y=nabs_h3chs),
 colour="salmon", size=0.3)+
 geom_line(aes(x=wn, y=nabs_h3ima),
 colour="salmon2", size=0.3)+
 geom_line(aes(x=wn, y=nabs_h3n61),
 colour="salmon4", size=0.3)+
 geom_vline(xintercept=wn_target1,
 colour="orange",size=0.3)+
 geom_vline(xintercept=wn_target2,
 colour="green",size=0.3)+
 geom_vline(xintercept=wn_target3,
 colour="blue",size=0.3)+
```

```
 geom_vline(xintercept=wn_target4,
 colour="magenta",size=0.3)+
 theme(axis.text.x=element_text(angle=45, hjust=1))
 print(plotspecmag2n)
 plotlistnorm((i)) <- plotspecmag2n
 #save the plot as png format
 setwd(OutputPath)
 b <- getwd()
 b
 filename_plotspecmag2n <- paste(today2, "_c7.1_specmag2n_",wn_target1, "_", wn_target2,
".png", sep="")
 ggsave(file = filename_plotspecmag2n,
 plot = plotspecmag2n, dpi=300,
 width=3.6, height=3.6)
 #this is the end of magnification option 2

} # this is an end of the i-loop
```

References

Allwood, J.W.; Chandra, S.; Xu, Y.; Dunn, W.B.; Correa, E.; Hopkins, L.; Goodacre, R.; Tobin, A.K.; Bowsher, C.G. Profiling of spatial metabolite distributions in wheat leaves under normal and nitrate limiting conditions. Phytochemistry 2015, 115, 99–111. doi.org/10.1016/j.phytochem.2015.01.007

Álvarez, Á.; Yáñez, J.; Neira, Y.; Castillo-Felices, R.; Hinrichsen, P. Simple distinction of grapevine (*Vitis vinifera* L.) genotypes by direct ATR-FTIR. Food Chem. 2020, 328, 127164. doi.org/10.1016/j.foodchem.2020.127164

Ami, D.; Natalello, A.; Mereghetti, P.; Neri, T.; Zanoni, M.; Monti, M.; Doglia, S.M.; Redi, C.A. FTIR spectroscopy supported by PCA-LDA analysis for the study of embryonic stem cell differentiation. Spectroscopy 2010, 24, 89–97. doi 10.3233/SPE-2010-0411

Assouline, S.; Or, D. The concept of field capacity revisited: Defining intrinsic static and dynamic criteria for soil internal drainage dynamics. Water Resour. Res. 2014, 50, 4787–4802. doi:10.1002/2014WR015475.

Astatsa. Complex Online Web Statistics Calculator. Available online: https://astatsa.com/ (accessed on 3 January 2022 and on 3 April 2022).

Ba˘gcıo˘glu, M.; Kohler, A.; Seifert, S.; Kneipp, J.; Zimmermann, B. Monitoring of plant—environment interactions by high throughput FTIR spectroscopy of pollen. Methods Ecol. Evol. 2017, 8, 870–880. doi.org/10.1111/2041-210X.12697

Balfourier, F.; Bouchet, S.; Robert, S.; de Oliveira, R.; Rimbert, H.; Kitt, J.; Choulet, F.; Paux E. International wheat genome sequencing consortium, breed wheat consortium, worldwide phylogeography and history of wheat genetic diversity. Sci. Adv. 2019, 5, eaav0536. doi.org/10.1126/sciadv.aav0536

Baron-Epel, O.; Gharyal, P.K.; Schindler, M. Pectins as mediators of wall porosity in soybean cells. Planta 1988, 175, 389–395.

Bona, E.; Marquetti, I.; Link, J.V.; Makimori, G.Y.F.; da Costa Arca, V.; Lemes, A.L.G.; Ferreira, J.M.G.; dos Santos Scholz, M.B.; Valderrama, P.; Poppi, R.J. Support vector machines in tandem

with infrared spectroscopy for geographical classification of green arabica coffee. LWT - Food Sci. Technol. 2017, 76, 330–336. doi.org/10.1016/j.lwt.2016.04.048

Bouyanfif, A.; Liyanage, S.; Hewitt, J.E.; Vanapalli, S.A.; Moustaid-Moussa, N.; Hequet, E.; Abidi, N. FTIR imaging detects diet and genotype-dependent chemical composition changes in wild type and mutant *C. elegans* strains. Analyst 2017, 142, 4727–4736. doi 10.1039/c7an01432e

Christou, C.; Agapiou, A.; Kokkinofta, R. Use of FTIR spectroscopy and chemometrics for the classification of carobs origin. J. Adv. Res. 2018, 10, 1–8. doi.org/10.1016/j.jare.2017.12.001

Cortizas, M.A.; López-Costas, O. Linking structural and compositional changes in archaeological human bone collagen: An FTIR-ATR approach. Sci. Rep. 2020, 10, 17888. doi.org/10.1038/s41598-020-74993-y

De Leonardis, A.M.; Fragasso, M.; Beleggia, R.; Ficco, D.B.M.; de Vita, P.; Mastrangelo, A.M. Effects of heat stress on metabolite accumulation and composition, and nutritional properties of durum wheat grain. Int. J. Mol. Sci. 2015, 16, 30382–30404. doi:10.3390/ijms161226241

Demir, P.; Onde, S.; Severcan, F. Phylogeny of cultivated and wild wheat species using ATR-FTIR spectroscopy. Spectrochim. Acta A 2015, 135, 757–763. doi.org/10.1016/j.saa.2014.07.025

Elbashir, A.A.; Gorafi, Y.S.; Tahir, I.S.; Elhashimi, A.M.; Abdalla, M.G.; Tsujimoto, H. Genetic variation in heat tolerance-related traits in a population of wheat multiple synthetic derivatives. Breed. Sci. 2017, 67, 483–492. doi.org/10.1270/jsbbs.17048

Elbashir, A.A.E.; Gorafi, Y.S.A.; Tahir, I.S.A.; Kim, J.S.; Tsujimoto, H. Wheat multiple synthetic derivatives: a new source for heat stress tolerance adaptive traits. Breed. Sci. 2017, 248–256.doi.org/10.1270/jsbbs.16204

Galleni, A.; D'Ascenzo, N.; Stagnari, F.; Pagnani, G.; Xie, Q.; Pisante, M. Past and future of plant stress detection: An overview from remote sensing to position emission tomography. Front. Plant Sci. 2021, 11, 609155. doi.org/10.3389/fpls.2020.609155

Georget, D.M.R.; Belton, P.S. Effects of temperature and water content on the secondary structure of wheat gluten studied by FTIR spectroscopy. Biomacromol. 2006, 7, 469–475. doi.org/10.1021/bm050667j

Ghatak, A.; Chaturvedi, P.; Weckwerth, W. Metabolomics in plant stress physiology. Adv. Biochem. Eng. Biotechnol. 2018, 164, 187–236. doi 10.1007/10_2017_55

Giang, L.T.; Thien, T.L.T.; Yen, D.H. Rapid classification of rice in northern Vietnam by using FTIR spectroscopy combined with chemometrics methods. Viet. J. Chem. 2020, 58, 372–379. doi 10.1002/vjch.202000001

Gorafi, Y.S.A.; Kim, J.-S.; Elbashir, A.A.E.; Tsujimoto, H. A population of wheat multiple synthetic derivatives: an effective platform to explore, harness and utilize genetic diversity of *Aegilops tauschii* for wheat improvement. Theor. Appl. Genet. 2018, 131, 1615–1626. doi.org/10.1007/s00122-018-3102-x

Gorgulu, S.T.; Dogan, M.; Severcan, F. The characterization and differentiation of higher plants by Fourier transform infrared spectroscopy. Appl. Specrosc. 2007, 61, 300–308. doi 10.1366/000370207780220903

Grunert, T.; Herzog, R.; Wiesenhofer, F.M.; Vychytil, A.; Ehling-Schulz, M.; Kratochwill, K. Vibrational spectroscopy of peritoneal dialysis effluent for rapid assessment of patient characteristics. Biomol. 2020, 10, 965. doi.org/10.3390/biom10060965

Gupta, N.K.; Agarwal, S.; Agarwal, V.P.; Nathawat, N.S.; Gupta, S.; Singh, G. Effect of short-term heat stress on growth, physiology and antioxidative defense system in wheat seedlings. Acta Physiol. Plant. 2013, 35, 1837–1842. doi 10.1007/s11738-013-1221-1

Hamany Djande, C.Y.; Pretorius, C.; Tugizimana, F.; Piater, L.A.; Dubery, I.A. Metabolomics: A tool for cultivar phenotyping and investigation of grain crops. Agronomy 2020, 10, 831. doi.org/10.3390/agronomy10060831

Harrison, D.; Rivard, B.; Sánchez-Azofeifa, A. Classification of tree species based on longwave hyperspectral data from leaves, a case study for a tropical dry forest. Int. J. Appl. Earth Obs. Geoinformation 2018, 66, 93–105. doi.org/10.1016/j.jag/2017.11.009

Iizumi, T.; Ali-Babiker, I.E.A.; Tsubo, M.; Tahir, I.S.A.; Kurosaki, Y.; Kim, W.; Gorafi, Y.S.A.; Idris, A.A.M.; Tsujimoto, H. Rising temperatures and increasing demand challenge wheat supply in Sudan. Nat. Food 2021, 2, 19–27. doi.org/10.1038/s43016-020-00214-4

ImageJ Home Page, Version 1.80. Available online: https://imagej.nih.gov/ij/index.html (accessed on 10 October 2021).

Johnson, H.E.; Broadhurst, D.; Goodacre, R.; Smith, A.R. Metabolic fingerprinting of salt-stressed tomatoes. Phytochemistry 2003, 62, 919–928. doi:10.1016/S0031-9422(02)00722-7

Kamnev, A.A.; Dyatlova, Y.A.; Kenzhegulov, O.A.; Vladimirova, A.A.; Mamchenkova, P.V.; Tugarova, A.V. Fourier transform infrared (FTIR) spectroscopic analyses of microbiological samples and biogenic selenium nanoparticles of microbial origin: Sample preparation effects. Molecules 2021, 26, 1146. doi: 10.3390/molecules26041146

Kamnev, A.A.; Tugarova, A.V.; Dyatlova, Y.A.; Tarantilis, P.A.; Grigoryeva, O.P.; Fainleib, A.M.; De Luca, S. Methodological effects in Fourier transform infrared (FTIR) spectroscopy: Implications for structural analyses of biomacromolecular samples. Spectrochim. Acta. A. 2018, 193, 558–564. doi.org/10.1016/j.saa.2017.12.051

Keleş, Y.; Öncel, I. Response of antioxidative defense system to temperature and water stress combinations in wheat seedlings. Plant Sci. 2002, 163, 783–790, doi.org/10.1016/S0168-9452(02)00213-3

Kurian, J.K.; Gariepy, Y.; Orsat, V.; Raghavan, V. Microwave-assisted lime treatment and recovery of lignin from hydrothermally treated sweet sorghum bagasse. Biofuels 2015, 6, 341–355. doi10.1080/17597269.2015.1110775

Lahlali, R.; Jiang, Y.; Kumar, S.; Karunakaran, C.; Liu, X.; Borondics, F.; Hallin, E.; Bueckert, R. ATR-FTIR spectroscopy reveals involvement of lipids and proteins of intact pea pollen grains to heat stress tolerance. Front. Plant Sci. 2014, 5, 747. doi.org/10.3389/fpls.2014.00747

Lammers, K.; Arbuckle-Keil, G.; Dighton, J. FTIR study of the changes in carbohydrate chemistry of three New Jersey pine barrens leaf litters during simulated control burning. Soil Biol. Biochem. 2009, 41, 340–347. doi:10.1016/j.soilbio.2008.11.005

Leech, R. M. "synthesis of cellular components in leaves." Seminar series - Society for Experimental Biology. 1985.

Li, H.; Liu, Z.; Mamtimin, A.; Liu, J.; Liu, Y.; Ju, C.; Zhang, H.; Gao, Z. A new linear relation for estimating surface broadband emissivity in arid regions based on FTIR and MODIS products. Remote Sens. 2021, 13, 1686. doi.org/10.3390/rs13091686

Lima, R.B.; dos Santos, T.B.; Vieira, L.G.E.; Ferrarese, M.L.L.; Ferrarese-Filho, O.; Donatti, L.; Boeger, M.R.T.; Petkowicz, C.L.O. Heat stress causes alterations in the cell-wall polymers and anatomy of coffee leaves (*Coffea arabica* L.). Carbohydr. Polym. 2013, 93, 135–143. doi.org/10.1016/j.carbpol.2012.05.015

Liu, X.; Renard, G.M.G.C.; Bureau, S.; Bourvellec, C.L. Revisiting the contribution of ATR-FTIR spectroscopy to characterize plant cell wall polysaccharides. Carbohydr. Polym. 2021, 262, 117935. doi.org/10.1016/j.carbpol.2021.117935

Mandrone, M.; Chiocchio, I.; Barbanti, L.; Tomasi, P.; Tacchini, M.; Poli, F. Metabolomic study of sorghum (*Sorghum bicolor*) to interpret plant behavior under variable field conditions in view of smart agriculture applications. J. Agric. Food Chem. 2021, 69, 1132–1145. doi.org/10.1021/acs.jafc.0c06533

Mathur, S., Agrawal, D. and Jajoo, A. Photosynthesis: response to high temperature stress. J. Photochem. Photobiol. B, Biol, 2014, 137,116-126. doi: 10.1016/j.jphotobiol.2014.01.010.

Mascarenhas, M.; Dighton, J.; Arbuckle, G.A. Characterization of plant carbohydrates and changes in leaf carbohydrate chemistry due to chemical and enzymatic degradation measured by microscopic ATR-FTIR spectroscopy. Appl. Spectrosc. 2000, 54, 681–686. doi.org/10.1366/0003702001950166

Matsunaga, S.; Yamasaki, Y.; Toda, Y.; Mega, R.; Akashi, K.; Tsujimoto, H. Stage-specific characterization of physiological response to heat stress in the wheat cultivar Norin 61. Int. J. Mol. Sci. 2021a, 22, 6942. doi.org/10.3390/ijms22136942

Matsunaga, S.; Yamasaki, Y.; Mega, R.; Toda, Y.; Akashi, K.; Tsujimoto, H. Metabolome profiling of heat priming effects, senescence, and acclimation of bread wheat induced by high temperatures at different growth stages. Int. J. Mol. Sci. 2021b, 22, 13139. doi.org/10.3390/ijms222313139

McCann, M.C.; Hammouri, M.; Wilson, R.; Belton, P.; Roberts, K. Fourier transform infrared microspectroscopy is a new way to look at plant cell walls. Plant Physiol. 1992, 100, 1940–1947. doi.org/10.1104/pp.100.4.1940

McCann, S.E.; Huang, B. Effects of trinexapac-ethyl foliar application on creeping bent grass responses to combined drought and heat stress. Crop Sci. 2007, 47, 2121–2128. doi.org/10.2135/cropsci2006.09.0614

Mitchell, R.A.C.; Mitchell, V. J.; Driscoll, S.P.; Franklin, J. Lawlor, D.W. Effects of increased CO_2 concentration and temperature on growth and yield of winter wheat at two levels of nitrogen application. Plant Cell Environ. 1993, 16, 521-529. 392 doi.org/10.1111/j.1365-3040.1993.tb00899.x

Munz, E.; Rolletschek, H.; Oeltze-Jafra, S.; Fuchs, J.; Guendel, A.; Neuberger, T.; Ortleb, S.; Jakob, P.M.; Borisjuk, L. A functional imaging study of germinating oilseed rape seed. New Phytol. 2017, 216, 1181–1190. doi.org/10.1111/nph.14736

Narayanan, S.; Tamura, P.J.; Roth, M.R.; Prasad, P.V.V.; Welti, R. Wheat leaf lipids during heat stress: I. High day and night temperatures result in major lipid alteration. Plant Cell Environ. 2016, 39, 787–803. doi 10.1111/pce.12649.

Nikalje, G.C.; Kumar, J.; Nikam, T.D.; Suprasanna, P. FT-IR profiling reveals differential response of roots and leaves to salt stress in a halophyte *Sesuvium portulacastrum* (L.). Biotechnol. Rep. 2019, 23, e00352. doi.org/10.1016/j.btre.2019.e00352

Oleszko, A.; OlsztyNska-Janus, S.; Walski, T.; Grzeszczuk-Ku´c, K.; Bujok, J.; GaBecka, K.; Czerski, A.; Witkiewicz, W.; Komorowska, M. Application of FTIR-ATR spectroscopy to determine the extent of lipid peroxidation in plasma during haemodialysis. Biomed. Res. Int. 2015, 2015, 245607. doi 10.1155/2015/245607

Osman, S.O.M.; Saad, A.S.I.; Tadano, S.; Takeda, Y.; Konaka, T.; Yamasaki, Y.; Tahir, I.S.A.; Tsujimoto, H.; Akashi, K. Chemical fingerprinting of heat stress responses in the leaves of common wheat by Fourier transform infrared spectroscopy. Int. J. Mol. Sci. 2022a, 23, 2842. doi.org/10.3390/ijms23052842

Osman, S.O., Saad, A.S.I., Tadano, S., Takeda, Y., Yamasaki, Y., Tahir, I.S., Tsujimoto, H. and Akashi, K. Probing differential metabolome responses among wheat genotypes to heat stress using Fourier transform infrared-based chemical fingerprinting. Agriculture, 2022b, 12, 753. doi.org/10.3390/agriculture12060753

Paymard, P.; Yaghoubi, F.; Nouri, M.; Bannayan, M. Projecting climate change impacts on rainfed wheat yield, water demand, and water use efficiency in northeast Iran. Theor. Appl. Climatol., 2019, 138, 1361-1373. doi.org/10.1007/s00704-019-02896-8

Petrou, K.; Nielsen, D.A.; Heraud, P. Single-cell biomolecular analysis of coral algal symbionts reveals opposing metabolic responses to heat stress and expulsion. Front. Mar. Sci. 2018, 5, 110. doi.org/10.3389/fmars.2018.00110

Prasad, P.V.V.; Djanaguiraman, M. Response of floret fertility and individual grain weight of wheat to high temperature stress: Sensitive stages and thresholds for temperature and duration. Funct. Plant Biol. 2014, 41, 1261–1269. doi 10.1071/FP14061.

Qaseem, M.F.; Qureshi, R.; Shaheen, H. Effects of pre-anthesis drought, heat and their combination on the growth, yield and physiology of diverse wheat (*Triticum aestivum* L.) genotypes varying in sensitivity to heat and drought stress. Sci. Rep. 2019, 9, 6955. doi.org/10.1038/s41598-019-43477-z

Qin, D.; Wu, H.; Peng, H.; Yao, Y.; Ni, Z.; Li, Z.; Zhou, C.; Sun, Q. Heat stress-responsive transcriptome analysis in heat susceptible and tolerant wheat (*Triticum aestivum* L.) by using Wheat Genome Array. BMC genomics 2008, 9, 432. doi.org/10.1186/1471-2164-9-432

R Core Team. A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2020; Available online: http://www.r-project.org/index.html (accessed on 1 October 2020).

Ramani, H.R.; Mandavia, M.K.; Dave, R.A.; Bambharolia, R.P.; Silungwe, H.; Garaniya, N.H. Biochemical and physiological constituents and their correlation in wheat (*Triticum aestivum* L.) genotypes under high temperature at different development stages. Int. J. Plant Physiol. Biochem. 2017, 9, 1–8. doi.org/10.5897/IJPPB2015.0240

Razzaq, A.; Sadia, B.; Raza, A.; Hameed, M.K.; Saleem, F. Metabolomics: A way forward for crop improvement. Metabolites 2019, 9, 303. doi 10.3390/metabo9120303

Reif, J.C.; Zhang, P.; Dreisigacker, S.; Warburton, M.L.; van Ginkel, M.; Hoisington, D.; Bohn, M.; Melchinger, A.E. Wheat genetic diversity trends during domestication and breeding. Theor. Appl. Genet. 2005, 110, 859–864. doi.org/10.1007/s00122- 444 004-1881-8

Rohman, A.; Ghazali, M.A.B.; Windarsih, A.; Irnawati Riyanto, S.; Yusof, F.M.; Mustafa, S. Comprehensive review on application of FTIR spectroscopy coupled with chemometrics for authentication analysis of fats and oils in the food products. Molecules 2020, 25, 5485. doi 10.3390/molecules25225485.

Sakurai, N. Recent applications of metabolomics in plant breeding. Breed. Sci. 2022, 72, 56–65. doi.org/10.1270/jsbbs.21065

Sattar, A.; Sher, A.; Ijaz, M.; Ul-Allah, S.; Rizwan, M.S.; Hussain, M.; Jabran, K.; Cheema, M.A. Terminal drought and heat stress alter physiological and biochemical attributes in flag leaf of bread wheat. PLoS One 2020, 15, e0232974. doi.org/10.1371/journal.pone.0232974

Savicka, M.; Škute, N. Effects of high temperature on malondialdehyde content superoxide production and growth changes in wheat seedlings (*Triticum aestivum* L.). Ekologija 2010, 56, 26–33. doi 10.2478/v10055-010-0004-x

Schittenhelm, S.; Langkamp-Wedde, T.; Kraft, M.; Kottmann, L.; Matschiner, K. Effect of two-week heat stress during grain filling on stem reserves, senescence, and grain yield of European winter wheat cultivars. J. Agron. Crop Sci. 2020, 206, 722–733. doi.org/10.1111/jac.12410

Semenov, M.A.; Halford, N.G. Identifying target traits and molecular mechanisms for wheat breeding under a changing climate. J. Exp. Bot. 2009, 60, 2791–2804. doi.org/10.1093/jxb/erp164

Setser, A.L.; Smith, R.W. Comparison of variable selection methods prior to linear discriminant analysis classification of synthetic phenethylamines and tryptamines. Forensic Chem. 2018, 77–86. doi.org/10.1016/j.forc.2018.10.002

Shapaval, V.; Møretrø, T.; Suso, H.-P.; Åsli, A.W.; Schmitt, J.; Lillehaug, D.; Martens, H.; Böcker, U.; Kohler, A. A high-throughput microcultivation protocol for FTIR spectroscopic

characterization and identification of fungi. J. Biophoton. 2010, 3, 512–521. doi: 10.1002/jbio.201000014.

Sharma, V.; Bhardwaj, S.; Kumar, R. On the spectroscopic investigation of kohl stains via ATR-FTIR and multivariate analysis: Application in forensic trace evidence. Vib. Spectrosc. 2019, 101, 81–91. doi.org/10.1016/j.vibspec.2019.02.006

Shewry, P.R.; Hey, S.J. The contribution of wheat to human diet and health. Food Energy Secur. 2015, 4, 178–202. doi.org/ 385 10.1002/fes3.64

Solomon, S.; Qin, D.; Manning, M; Marquis, M.; Averyt, K.; Tignor, M.M.B.; LeRoy Miller H; Chen Z. 2007. Climate change 2007: the physical science basis. Contribution of working group I to the Fourth Assessment Report of the Inter govermental Panel on climate Change. New York, 2007, Cambridge University Press

Sowa, S.; Connor, K.F.; Towill, L.E. Temperature changes in lipid and protein structure measured by Fourier transform infrared spectrophotometry in intact pollen grains. Plant Sci. 1991, 78, 1–9. doi.org/10.1016/0168-9452(91)90155-2

Stewart, D. Fourier transform infrared microspectroscopy of plant tissues. Appl. Spectrosc. 1996, 50, 357–365.

Stone, P.J.; Nicolas, M.E. Effect of timing of heat stress during grain filling on two wheat varieties differing in heat tolerance. I. grain growth. Aust. J. Plant Physiol. 1995, 22, 927–934. doi.org/10.1071/PP9950927

Stuart, B. Biological application. In Infrared Spectroscopy: Fundamentals and Applications; Stuart, B., Ed.; John Wiley and Sons Ltd.: Chichester, UK, 2004; pp. (45–70; 137–165).

Tadesse, W.; Sanchez-Garcia, M.; Assefa, S.G.; Amri, A.; Bishaw, Z.; Ogbonnaya, F.C.; Baum, M. Genetic gains in wheat breeding and its role in feeding the world. Crop Breed. Genet. Genom. 2019, 1, e190005. doi.org/10.20900/cbgg20190005

Talari, A.C.S.; Martinez, M.A.G.; Movasaghi, Z.; Rehman, S.; Rehman, I.U. Advances in Fourier transform infrared (FTIR) spectroscopy of biological tissues. Appl. Spectrosc. Rev. 2016, 52, 456–506. doi.org/10.1080/05704928.2016.1230863

Talukder, A.S.M.H.M.; McDonald, G.K.; Gill, G.S. Effect of short-term heat stress prior to flowering and early grain set on the grain yield of wheat. Field Crops Res. 2014, 160, 54–63. doi.org/10.1016/j.fcr.2014.01.013

Tarapoulouzi, M.; Kokkinofta, R.; Theocharis, C.R. Chemometric analysis combined with FTIR spectroscopy of milk and Halloumi cheese samples according to species' origin. Food Sci Nutr. 2020, 8, 3262–3273. doi.org/10.1002/fsn3.1603

Thomason, K.; Babar, M.A.; Erickson, J.E.; Mulvaney, M.; Beecher, C.; MacDonald, G. Comparative physiological and metabolomics analysis of wheat (*Triticum aestivum* L.) following post-anthesis heat stress. PLoS One 2018, 13, e0197919. doi.org/10.1371/journal.pone.0197919

Walkowiak, S.; Gao, L.; Monat, C.; Haberer, G.; Kassa, M.T.; Brinton, J.; Ramirez-Gonzalez, R.H. et al.. Multiple wheat genomes reveal global variation in modern breeding. Nature 2020, 588, 277–283. doi.org/10.1038/s41586-020-2961-x

Walsh, M.J.; Fellous, T.G.; Hammiche, A.; Lin, W. R.; Fullwood, N.J.; Grude, O.; Bahrami, F.; Nicholson, J.M.; Cotte, M.; Susini, J.; et al.. Fourier transform infrared microspectroscopy identifies symmetric PO_2^- modifications as a marker of the putative stem cell region of human intestinal crypts. Stem Cells 2008, 26, 108–118. doi:10.1634/stemcells.2007-0196

Wang, X.; Hou, L.; Lu, Y.; Wu, B.; Gong, X.; Liu, M.; Wang, J.; Sun, Q.; Vierling, E.; Xu, S. Metabolic adaptation of wheat grain contributes to a stable filling rate under heat stress. J. Exp. Bot. 2018, 69, 5531–5545. doi:10.1093/jxb/ery30

Westworth, S.; Ashwath, N.; Cozzolino, D. Application of FTIR-ATR spectroscopy to detect salinity response in beauty leaf tree (*Calophyllum inophyllum* L). Energy Proc. 2019, 160, 761–768. doi 10.1634/stemcells.2007-0196

Xanthopoulos, P.; Pardalos, P.M.; Trafalis, T.B. Linear discriminant analysis. In: Robust Data Mining. pp. 27–33. Springer Briefs in Optimization, Springer, New York, NY, USA. doi.org/10.1007/978-1-4419-9878-1_4

Xu, Z.Z.; Zhou, G.S. Combined effects of water stress and high temperature on photosynthesis, nitrogen metabolism and lipid peroxidation of a perennial grass *Leymus chinensis*. Planta 2006, 224, 1080–1090. doi: 10.1007/s00425-006-0281-5.

Yalkun, A.; Mamtimin, A.; Liu, S.; Yang, F.; He, Q.; Qi, F.; Liu, Y. Coefficients optimization of the GLASS broadband emissivity based on FTIR and MODIS data over the Taklimakan Desert. Sci. Rep. 2019, 9, 18460. doi.org/10.1038/s41598-019-54982-6

Zhang, B.; Liu, W.; Chang, S.X.; Anyia, A.O. Water-deficit and high temperature affected water use efficiency and arabinoxylan concentration in spring wheat. J. Cereal Sci. 2010, 52, 263–269. doi.org/10.1016/j.jcs.2010.05.014

Zampieri, M.; Ceglar, A.; Dentener, F.; Toreti, A. Wheat yield loss attributable to heat waves, drought and water excess at the global, national and subnational scales. Environ. Res. Lett., 2017, 12, 064008. doi.org/10.1088/1748-9326/aa723b

Zhang, Y.; Pan, J.; Huang, X.; Guo, D.; Lou, H.; Hou, Z.; Su, M.; Liang, R.; Xie, C.; You, M.; Li, B. Differential effects of a post anthesis heat stress on wheat (*Triticum aestivum* L.) grain proteome determined by iTRAQ. Sci. Rep. 2017, 3468. doi.org/10.1038/s41598-017-03860-0

Zhao, C.; Liu, B.; Piao, S.; Wang, X.; Lobell, D.B.; Huang, Y.; Huang, M.; Yao, Y.; Bassu, S.; Ciais, P.; et al.. Temperature increase reduces global yields of major crops in four independent estimates. Proc. Natl. Acad. Sci. USA 2017, 114, 9326–9331. doi.org/10.1073/pnas.1701762114

Zhao, X.; Yang, X.; Shi, Y.; Chen, G.; Li, X. Protein and lipid characterization of wheat roots plasma membrane damaged by Fe and H_2O_2 using ATR-FTIR method. J. Biophys. Chem. 2013, 4, 28–35. doi 10.4236/jbpc.2013.41004

List of Publications

Chapter 1

Title: Chemical Fingerprinting of Heat Stress Responses in the Leaves of Common Wheat by Fourier Transform Infrared Spectroscopy

Authors: Osman, S. O. M., Saad, A. I., Tadano, S., Takeda, Y., Konaka, T., Yamasaki, Y., Tahir, I. S. A., Tsujimoto, H. and Akashi, K.

Journal: International Journal of Molecular Science 23: 2842. doi.org/10.3390/ijms23052842

Published online: March, 2022

Chapter 2

Title: Probing Differential Metabolome Responses among Wheat Genotypes to Heat Stress Using Fourier Transform Infrared-Based Chemical Fingerprinting

Authors: Osman, S. O. M., Saad, A. I., Tadano, S., Takeda, Y., Yamasaki, Y., Tahir, I. S. A., Tsujimoto, H. and Akashi, K.

Journal: Agriculture 12: 753. doi.org/10.3390/agriculture12060753

Published online: May, 2022