ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΒΙΟΪΑΤΡΙΚΗΣ

Επεξεργασία Μεταβολομικών δεδομένων με φασματοσκοπία Πυρηνικού Μαγνητικού Συντονισμού(NMR)

ΕΥΜΟΡΦΙΑ ΑΝΝΑ ΚΟΛΑΚΗ

Αριθμός Μητρώου: 48015049

Επιβλέπων Καθηγητής

Παναγιώτης Ζουμπουλάκης, Αναπληρωτής Καθηγητής

Αθήνα 07/07/2022

Η Τριμελής Εξεταστική Επιτροπή

Ο Επιβλέπων Καθηγητής		
Παναγιώτης Ζουμπουλάκης	Διονύσιος Κάβουρας	Μίνως Ματσούκας
Αναπληρωτής καθηγητής	Ομότιμος Καθηγητής	Αναπληρωτής Καθηγητής
Panagiotis Zoumpoulakis Digitally signed by Panagiotis Zoumpoulakis Date: 2022.08.09 09:33:00 +03'00'	Dionysios Kavouras Digitally signed by Dionysios Kavouras Date: 2022.08.09 22:16:43 +03'00'	Minos Timotheos Matsoukas Digitally signed by Minos Timotheos Matsoukas Date: 2022.08.09 12:32:53 +03'00'

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΠΤΥΧΙΑΚΗΣ/ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Η υπογραφούσα Κολάκη Ευμορφία Άννα του Παναγιώτη, με αριθμό μητρώου 48015049 φοιτήτρια του Τμήματος Μηχανικών Βιοϊατρικής της Σχολής Μηχανικών του Πανεπιστημίου Δυτικής Αττικής, δηλώνω υπεύθυνα ότι:

«Είμαι συγγραφέας αυτής της διπλωματικής εργασίας και κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος.

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του διπλώματός μου».

Ημερομηνία

19/07/2022

Η Δηλούσα

ΠΕΡΙΛΗΨΗ

Η παρούσα διπλωματική έχει ως σκοπό την εφαρμογή της γλώσσας προγραμματισμού R καθώς και μεθόδων πολυμεταβλητής στατιστικής ανάλυσης σε δεδομένα μεταβολομικής από φασματοσκοπία Πυρηνικού Μαγνητικού Συντονισμού (NMR). Στόχος ήταν η αναζήτηση του βέλτιστου τρόπου επεξεργασίας των φασματικών δεδομένων για την περαιτέρω στατιστική ανάλυση. Συγκεκριμένα χρησιμοποιήθηκε μια βάση φασματοσκοπικών δεδομένων που αναφέρονται στο μεταβολικό αποτύπωμα συγκεκριμένης ποικιλίας ντομάτας καθώς και το ελεύθερα διαθέσιμο λογισμικό NMRProcFlow το οποίο αναλύει και επεξεργάζεται φάσματα NMR χωρίς να είναι αναγκαία η γραφή κώδικα. Για την πολυμεταβλητή στατιστική ανάλυση χρησιμοποιήθηκαν μέθοδοι - όπως η PCA, η PLS-DA και η OPLS-DA. Επίσης, ταυτοποιήθηκαν και ποσοτικοποιήθηκαν οι μεταβολίτες των φασμάτων μέσω εφαρμογής αυτοματοιποιημένου αλγορίθμου. Ακολούθησε συγκριτική στατιστική ανάλυση με χρήση των συγκεντρώσεων των μεταβολιτών.

Λέξεις Κλειδιά: Μεταβολομική, Μεταβολίτες, Στατιστική ανάλυση, R, φασματοσκοπία NMR, πολυμεταβλητή στατιστική ανάλυση, PCA, PLS-DA, OPLS-DA

ABSTRACT

The purpose of this thesis is the application of R-programming as well as multivariate statistical analysis on metabolic data derived from Nuclear Magnetic Resonance (NMR) Spectroscopy. The main aim was to find the best way to process the spectral data for further statistical analysis. A spectroscopic data base which is the metabolic fingerprint of a specific species of tomato was used. Also, the freely available software called NMRPRocFLow was utilized which analyzes and processes NMR spectra, without the need to use coding skills. We conducted multivariate statistical analysis by using PCA, PLS-DA and OPL-DA methods. Quantification and identification of metabolites from the NMR spectra of tomatoes has occurred by using a specialized, automated algorithm. Lastly, in order to assess the results, a statistical analysis has been done over the concentration of metabolites.

Keywords: Metabolomics, Metabolites, Statistical Analysis, R-programming, NMR spectroscopy, multivariate analysis, PCA, PLS-DA, OPLS-DA

Ευχαριστίες**:**

Θα ήθελα να ευχαριστήσω τον Αν. Καθηγητή Παναγιώτη Ζουμπουλάκη για την βοήθεια του στη διεκπεραίωση της εργασίας και τον καθηγητή Δρ. Διονύσιο Κάβουρα που με το σεμινάριο Μηχανικής Μάθησης που οργάνωσε, μου πρόσφερε τα απαραίτητα εργαλεία για την παρούσα πτυχιακή. Ευχαριστώ επίσης τον συνάδελφο μου υποψήφιο διδάκτωρα Σωτήρη Ουζούνη για την βοήθεια και τη στήριξη του.

Περιεχόμενα

ΕΙΣΑΓΩΓΗ

Σε αυτό το κεφάλαιο θα αναφερθούν και θα αναλυθούν βασικές έννοιες στις οποίες βασίστηκε η παρούσα διπλωματική. Πιο συγκεκριμένα θα αναλυθεί η μεταβολομική ως έννοια αλλά και η επιστήμη που έχει βασιστεί στα μεταβολομικά δεδομένα. Επίσης, αναφέρεται η λειτουργία της φασματοσκοπίας Πυρηνικού Μαγνητικού Συντονισμού (Nuclear Magnetic Resonance- NMR) μέσω της οποίας και παράγονται τα δεδομένα που χρησιμοποιήθηκαν.

1.1 Βασικές έννοιες της Μεταβολομικής

Οι μεταβολίτες είναι ενώσεις χαμηλού μοριακού βάρους (<1 kDa) που είναι το αποτέλεσμα χημικών αντιδράσεων που πραγματοποιούνται μέσα στα κύτταρα και είναι απαραίτητες για τη διατήρηση, την ανάπτυξη και την κανονική τους λειτουργία. (Φιλντίση, 2018) Οι μεταβολίτες είναι κυρίως οργανικές ενώσεις δηλαδή αμινοξέα, λιπαρά οξέα, υδατάνθρακες, βιταμίνες και λιπίδια.

Το μεταβόλωμα (metabolome) ορίζεται ως το σύνολο των μεταβολιτών που υπάρχουν ή εκκρίνονται σε ένα τύπο κυττάρου ή ιστού. Το μεταβόλωμα περιγράφει, επίσης, ένα σύνολο μεταβολιτών που υπάρχουν σε ένα βιολογικό σύστημα σε μια φυσιολογική κατάσταση υπό συγκεκριμένες περιβαλλοντικές συνθήκες, καθώς και σε ένα συνολικό μεταβολικό περιεχόμενο ενός βιολογικού δείγματος. Το μεταβόνωμα (metabonome) ορίζεται ως το σύνολο των κυτταρικών μεταβολωμάτων και προϊόντων χημικών αντιδράσεων σε πολυκύτταρους οργανισμούς. (Φιλντίση, 2018) Οι έννοιες του μεταβολώματος και του μεταβονώματος αρχικά ήταν διαφορετικές μεταξύ τους, πλέον όμως είναι συνώνυμες και περιγράφουν το σύνολο των μεταβολιτών ενός βιολογικού συστήματος. Η πιο ολοκληρωμένη βάση δεδομένων σε σχέση το μεταβόλωμα του ανθρώπου (Human Metabolome Database, HMDB) περιλαμβάνει μέχρι και τον Ιούλιο 2022 περισσότερους από 250.000 μεταβολίτες.

Οι πρώτες έρευνες σε σχέση με τη μεταβολική σύνθεση των κυττάρων, των ιστών και άλλων βιολογικών υγρών είναι αρκετά πρόσφατη, καθώς ξεκίνησαν τη δεκαετία του 1980. Κατά τη διάρκεια των χρόνων έχει αλλάξει ο ορισμός που έχει δοθεί στη μεταβολομική και στη μεταβονομική ως έννοιες. Ωστόσο έχουμε καταλήξει στους εξής ορισμούς. Μεταβολομική (metabolomics) είναι η ολική, ποιοτική, ποσοτική και αμερόληπτη ανάλυση των μεταβολιτών σε ένα βιολογικό σύστημα. (Φιλντίση, 2018) Άλλος ένα ορισμός για τη μεταβολομική είναι η αμερόληπτη ποσοτικοποίηση και αναγνώριση των μεταβολιτών που υπάρχουν σε ένα βιολογικό σύστημα. Μεταβονομική (metabonomics) είναι η ποσοτική μέτρηση της δυναμικής πολυπαραμετρικής μεταβολικής απόκρισης των έμβιων συστημάτων σε γενετικές μεταβολές ή παθοφυσιολογικά ερεθίσματα. Η μεταβονομική περιγράφει το αντικείμενο της ανίχνευσης, ταυτοποίησης, ποσοτικοποίησης και καταγραφής των μεταβολικών αλλαγών ενός βιολογικού συστήματος ως απόκριση σε διάφορους ενδογενείς ή εξωγενείς παράγοντες (ο τρόπος ζωής, γενετικοί/περιβαλλοντικοί παράγοντες). Η μεταβονομική και η μεταβολομική είναι πλέον δύο συνώνυμες έννοιες καθώς αποτελούν τη συνοπτική ανάλυση των μεταβολιτών ενός βιολογικού συστήματος/δείγματος.

Το μεταβολικό προφίλ (metabolite ή metabolic profiling) είναι μία διαδικασία που υπολογίζει συγκεκριμένους μεταβολίτες ενός δείγματος, οι οποίοι επηρεάζονται υπό ορισμένες συνθήκες. Αυτές οι συνθήκες μπορεί να είναι γενετική αλλοίωση ή/και αναπτυξιακή επίδραση. Στόχος αυτής της διαδικασίας είναι η αναγνώριση και η ποσοτικοποίηση των μεταβολιτών, οι οποίοι μπορεί να είναι μία κατηγορία μεταβολιτών (π.χ. υδατάνθρακες, λιπίδια) ή μέλη ενός βιοχημικού μονοπατιού. Με δεδομένο ότι οι μεταβολίτες υπό μελέτη είναι προκαθορισμένοι, η προετοιμασία των δειγμάτων και η λήψη δεδομένων μπορεί να διαμορφωθεί με βάση τις χημικές ιδιότητες αυτών των μεταβολιτών. Η διαδικασία του μεταβολικού προφίλ είναι χρήσιμη στην περίπτωση που είναι επιθυμητή η διερεύνηση επιλεγμένων βιοχημικών μονοπατιών. Το μεταβολικό αποτύπωμα (metabolite/metabolic fingerprinting) είναι μία γρήγορη διαδικασία σάρωσης υψηλής απόδοσης με στόχο την κατηγοριοποίηση των δειγμάτων προς ανάλυση, δηλαδή το διαχωρισμό τους ως προς την προέλευσή τους ή τη βιολογική κατάσταση στην οποία αντιστοιχούν. (Φιλντίση, 2018)

Η μεταβολομική χωρίζεται σε δύο διαφορετικές προσεγγίσεις ως προς τη μελέτη της, τη στοχευμένη (targeted) και τη μη στοχευμένη (untargeted) μεταβολομική ανάλυση. Η μη στοχευμένη προσέγγιση έχει ως στόχο την ανίχνευση όσων περισσότερων μεταβολιτών χωρίς να υπάρχει εκ των προτέρων γνώση για τους μεταβολίτες που πρόκειται να ανιχνεύσει. Μέσω της μη στοχευμένης υπάρχει δυνατότητα εξερεύνησης νέων περιοχών του μεταβολισμού του οργανισμού. Αντίθετα, η στοχευμένη προσέγγιση έχει ως στόχο την καλύτερη δυνατή ανάλυση μεταβολιτών οι οποίοι είναι ήδη γνωστό ότι υπάρχουν στον επικείμενο οργανισμό. Με αυτόν τον τρόπο, χρησιμοποιούνται προκαθορισμένα σήματα για τον ακριβή υπολογισμό των συγκεντρώσεων ενός συγκεκριμένου αριθμού μεταβολιτών. Στην παρούσα εργασία ασχολούμαστε με τη μη στοχευμένη προσέγγιση μεταβολιτών σε συγκέκριμενο σετ δεδομένων, στο όποιο έχει γίνει ήδη η στοχευμένη μεταβολομική ανάλυση. Χρησιμοποιούμε αυτή τη στοχευμένη μεταβολμική ανάλυση που έχει ερευνηθεί ως καθοδηγητή για τη μη στοχευμένη μεταβολομική ανάλυση των δεδομένων.

1.2 Φασματοσκοπία Πυρηνικού Μαγνητικού Συντονισμού, NMR spectroscopy

Η Φασματοσκοπία Πυρηνικού Μαγνητικού Συντονισμού (Nuclear Magnetic Resonance, NMR) είναι ένα εργαλείο που βοηθά στον προσδιορισμό της μοριακής δομής και χρησιμοποιείται σε ένα μεγάλο εύρος χημικών συστημάτων. Η λειτουργία του βασίζεται στην ιδιότητα που έχει ύλη και ονομάζεται πυρηνικό spin. Τα πυρηνικά spin μπορούν να εξαναγκαστούν σε αλλαγή του ενεργειακού τους επιπέδου μέσω έκθεσης του δείγματος σε συγκεκριμένης συχνότητας ραδιοκύματα. Η συχνότητα μεταβάλλεται και όταν συμπέσει ακριβώς με την χαρακτηριστική συχνότητα των πυρήνων (την συχνότητα συντονισμού) παράγεται ηλεκτρικό σήμα στον ανιχνευτή. Το φάσμα NMR είναι το διάγραμμα του σήματος ως προς την συχνότητα. Οι συχνότητες εξαρτώνται από το χημικό περιβάλλον των πυρήνων. ΟΙ πυρήνες στοιχείων που μελετώνται συχνά με αυτή την τεχνική είναι οι 1H, 13C, 19F, 31P.
Ο τρόπος που λειτουργεί η φασματοσκοπία πυρηνικού μαγνητικού συντονισμού είναι η εξής. Το δείγμα εισάγεται σε ομογενές στατικό μαγνητικό πεδίο, με αποτέλεσμα οι μαγνητικές ροπές των πυρήνων του δείγματος να παίρνουν συγκεκριμένους προσανατολισμούς ως προς την ένταση του μαγνητικού πεδίου και έρχονται σε κατάσταση ισορροπίας. Στη συνέχεια, εφαρμόζονται παλμοί ραδιοσυχνοτήτων, οι πυρήνες του δείγματος απορροφούν ηλεκτρομαγνητική ενέργεια και μεταβάλλουν ενεργειακά επίπεδα.

Σε ένα σημείο διακόπτονται οι παλμοί ραδιοσυχνοτήτων και το δείγμα επιστρέφει στην κατάσταση ισορροπίας. Όσο επιστρέφει στην αρχική του θέση ισορροπίας εκπέμπει την ηλεκτρομαγνητική ακτινοβολία που απορρόφησε, από την οποία προκύπτει το σήμα NMR.
Η φασματοσκοπία NMR έχει διάφορες εφαρμογές, η κυριότερη από τις οποίες είναι η ταυτοποίηση της μοριακής δομής των χημικών ενώσεων. Υπάρχουν διάφορες κατηγορίες πειραμάτων NMR, με κυριότερο κριτήριο διαφοροποίησης τον αριθμό διαστάσεων του σήματος NMR. Τα μονοδιάστατα (1D) πειράματα έχουν ως έξοδο ένα σήμα στο πεδίο χρόνου, από το μετασχηματισμό Fourier του οποίου προκύπτει ένα φάσμα συχνοτήτων με έναν άξονα συχνότητας (x) και έναν άξονα έντασης (y) του φάσματος. Τα δισδιάστατα (2D) πειράματα έχουν ως έξοδο ένα δισδιάστατο σήμα στο πεδίο χρόνου, από το δισδιάστατο μετασχηματισμό Fourier του οποίου προκύπτει ένα φάσμα συχνοτήτων με δύο άξονες συχνοτήτων (x,y) και έναν τρίτο άξονα έντασης (z) του σήματος.
Μέσω της φασματοσκοπίας NMR είναι δυνατή η βιοχημική και μεταβολική σύνθεση ενός οργανισμού, γεφυρώνοντας το κενό μεταξύ γενοτύπου και φαινοτύπου. Πιο συγκεκριμένα μέσω του ^{1}H-NMR δηλαδή της πρωτονιακής μαγνητικής φασματοσκοπίας επιτυγχάνεται η ταυτοποίηση όλων των μεταβολιτών που διαθέτουν πρωτόνια. Το ^{1}H-NMR είναι η πιο κατάλληλη τεχνική από άποψη ταχύτητας απόδοσης αποτελεσμάτων καθώς παράγει αποτελέσματα με πολύ χρήσιμη πληροφορία με ελάχιστη προετοιμασία του δείγματος.

ΜΕΘΟΔΟΛΟΓΙΑ

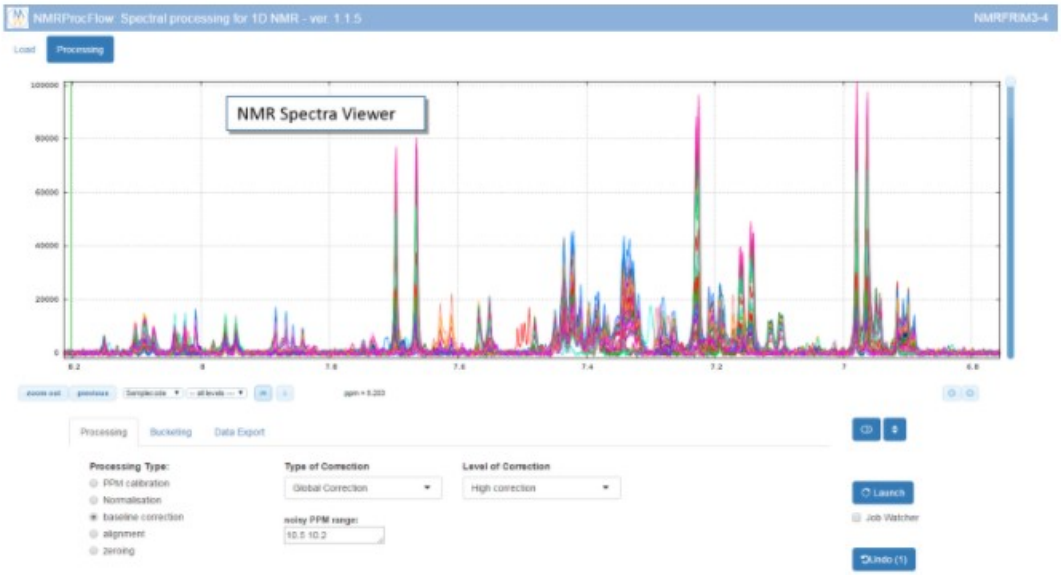
Σε αυτό το κεφάλαιο θα αναλύσουμε το κομμάτι της μεθοδολογίας της διπλωματικής.

1.3 Λογισμικό NMRProcFlow

Το NMRProcFlow είναι ένα ελεύθερο λογισμικό για την επεξεργασία και απεικόνιση των φασμάτων των μεταβολομικών δεδομένων. (Jacob, et al., 2017) Το λογισμικό είναι φιλικό προς τους νεοεισερχόμενους και μη έμπειρους στην μεταβολομική, καθώς χάρις το γραφικό περιβάλλον χρήστη (Graphical User Interface, GUI) αφαιρείται η δυσκολία εκμάθησης του προγραμματισμού. Τα δεδομένα που μπορούν να εισαχθούν είναι της μορφής σημάτων ελεύθερης επαγωγικής απόσβεσης (Free Induction Decay, FID) είτε της μορφής πινάκων. Οι πίνακες αυτοί περιέχουν δεδομένα της μορφής 1r-spectrum δηλαδή έναν πίνακα ο οποίος περιγράφεται με τη χημική μετατόπιση (parts-per-million/ ppm) και την ένταση (intensity). Μετά την είσοδο των δεδομένων μπορεί να απεικονιστεί το φάσμα δηλαδή οι κορυφές του. Ύστερα, μέσω του λογισμικού μπορεί να πραγματοποιηθεί επεξεργασία των δεδομένων. Τα είδη επεξεργασίας του φάσματος που είναι διαθέσιμα είναι τα εξής, ppm calibration, normalization (CSN/PQN), baseline correction, alignment (LS/TW/CluPA), ppm shift και zeroing. Μετά την επεξεργασία, ακολουθεί η διαδικασία της ομαδοποίησης των κορυφών (bucketing/binning), όπου υπάρχουν οι επιλογές της ομοιόμορφης τμηματοποίησης (uniform bucketing), της ευφυούς προσαρμοστικής τμηματοποίησης (Adaptive Intelligent bucketing) καθώς και τη στοχευμένη τμηματοποίηση (targeted bucketing) η οποία ορίζεται από τον ίδιο τον χρήστη. Στο τέλος της επεξεργασίας μπορούν να εξαχθούν ο πίνακας δεδομένων (data matrix), ο πίνακας ομαδοποίησης (buckets table), ο πίνακας σήματος προς θόρυβο (SNR matrix) καθώς και ένας πίνακας σε μορφή excel που περιέχει όλα τα παραπάνω.

Το NMRProcFlow είναι ένα ελεύθερα διαθέσιμο λογισμικό το οποίο αφορά την επεξεργασία H-NMR δεδομένων καθώς και την αναγνώριση μεταβολιτών. Είναι ένα εργαλείο μόνο για μεταβολομικά δεδομένα. Δημιουργήθηκε από το Γαλλικό Δημόσιο Ερευνητικό Ινστιτούτο - Institut National de la Recherche Agronomique (INRA) το οποίο απασχολείται με την γεωργική επιστήμη.

Τα δεδομένα της διπλωματικής αυτής αντλήθηκαν από το λογισμικό NMRProcFlow. Τα δεδομένα είναι μεταβολίτες που εντοπίζονται σε συγκεκριμένη ποικιλία ντομάτας σε διαφορετικά στάδια και συνθήκες ανάπτυξης. Στην έρευνα χρησιμοποιήθηκε ένα ελεύθερα διαθέσιμο λογισμικό το NMRProcFlow, το οποίο χρησιμοποιείται για την ανάλυση φασμάτων NMR με σκοπό ο βιολόγος/χημικός να μπορεί να επεξεργαστεί φάσματα χωρίς να είναι απαραίτητες γνώσεις προγραμματισμού. Μέσω της παρούσας διπλωματικής ελέγχεται η λειτουργία και οι δυνατότητες αυτού του εργαλείου.



Εικόνα 2.1.1 Το γραφικό περιβάλλον χρήστη (GUI) του NMRProcFlow

Free access: https://nmrprocflow.org/c2

1.4 Η έρευνα και η μέθοδος παραγωγής των δεδομένων από το INRA σε συγκεκριμένη ποικιλία ντομάτας

1.4.1 Περιγραφή της ντομάτας ως φυτό και φρούτο

Η ντομάτα (επιστημονικά Solanum lycopersicum) είναι φυτό της οικογένειας των Στρυχνοειδών (Solanaceae), ιθαγενές της Κεντρικής και Νοτίου Αμερικής. Υπάρχουν ντομάτες θερμοκηπίου (αναρριχώμενες) και υπαίθριες ντομάτες (ημιαναρριχώμενες και αυτοκλαδευόμενες). Οι αναρριχώμενες και οι ημιαναρριχώμενες χρειάζονται στήριξη, η οποία γίνεται είτε με σπάγκο (θερμοκήπιο από οριζόντιο σύρμα) είτε σε καλάμια όταν πρόκειται για υπαίθρια καλλιέργεια. Οι αυτοκλαδευόμενες ντομάτες δεν χρειάζονται στήριξη, διότι τυφλώνουν μόνες τους την κορυφή τους και δεν αυξάνονται προς τα πάνω.

Ο καρπός είναι σφαιρικός ή μακρόστενος, και όταν είναι ώριμος έχει έντονο κόκκινο χρώμα. Το κόκκινο χρώμα του οφείλεται στο ότι περιέχει τη χρωστική λυκοπένιο. Στα άγρια φυτά ο καρπός έχει διάμετρο 1-2 εκατοστά, αλλά στα περισσότερα είναι αρκετά μεγαλύτερος, από 5-10 εκατοστά. Το βάρος της ντομάτας φτάνει τα 250-350 γραμμάρια (μεγαλόκαρπη), ενώ υπάρχουν και μικρόκαρπα υβρίδια (cherry) τα οποία μπορούν να συγκομιστούν με το τσαμπί (και όχι μεμονωμένα) και έχουν βάρος 50-100 γραμμάρια.

Ενώ είναι ευρέως γνωστό ότι η ντομάτα είναι λαχανικό (σύμφωνα με απόφαση του Ανώτατου Δικαστηρίου των Ηνωμένων Πολιτειών το 1893), από βοτανικής άποψης η ντομάτα είναι φρούτο. Ωστόσο περιέχει πολύ χαμηλότερο επίπεδο σακχάρων σε σύγκριση με τα υπόλοιπα φρούτα. Είναι ένα διπλοειδές φυτό με 2n=24 χρωμοσώματα. (Gerszberg, et al., 2015) Η ντομάτα έχει πολλά θρεπτικά συστατικά όπως το λυκοπένιο, β-καροτίνη, φλαβονοειδή, βιταμίνη C και παράγωγα υδροξυκινναμωμικών οξέων. Επίσης, η ντομάτα φαίνεται να χρησιμοποιείται συχνά ως μοντέλο από ερευνητές. Αυτό οφείλεται γιατί η ντομάτα διαθέτει πολλά χρήσιμα για έρευνα χαρακτηριστικά (features), όπως την

ικανότητα να αναπτυχθεί σε διαφορετικές συνθήκες καλλιέργειας, έχει σχετικά μικρό κύκλο ζωής, μπορεί να παράγει σπόρους (seed) και έχει σχετικά μικρό γονιδίωμα (950 Mb). (Gerszberg, et al., 2015) Η τελική ποιότητα που έχει η ντομάτα οφείλεται στο γονιδίωμα της και στο περιβάλλον το οποίο αναπτύσσεται.

1.4.2 Το μοντέλο της οικολογίας –σχέσεις των κέντρων παραγωγής και κατανάλωσης φωτοσυνθετικών προϊόντων (source-to-sink)

Στον ανθρώπινο οργανισμό πολλές λειτουργίες του οργανισμού γίνονται αυτόματα μέσω διαφορετικών συστημάτων όπως το αναπνευστικό, το κυκλοφορικό, το αναπαραγωγικό και το πεπτικό. Έτσι και στα φυτά υπάρχουν αντίστοιχα συστήματα τα οποία φροντίζουν την σωστή λειτουργία του φυτικού οργανισμού. Μερικές βασικές λειτουργίες του φυτού είναι η αναπνοή και η φωτοσύνθεση. Η φωτοσύνθεση παρέχει την τροφή στο φυτό και άρα την απαραίτητη ενέργεια ώστε να γίνουν οι υπόλοιπες φυσικές λειτουργίες του φυτού. Το φυτό απαρτίζεται από τριών ειδών ιστούς, τους επιδερμικούς ιστούς οι οποίοι είναι χρήσιμοι για την προστασία του φυτού, τους ιστούς εδάφους οι οποίοι συμμετέχουν στον μεταβολισμό, την αποθήκευση και την στήριξη και τέλος τα αγγεία τα οποία συμμετέχουν στη μεταφορά ουσιών (κυρίως νερού και σακχαρόζης). Οι σχέσεις των κέντρων παραγωγής και κατανάλωσης φωτοσυνθετικών προϊόντων (source-to-sink) αφορά άμεσα τη μεταφορά σακχαρόζης στο φυτό όπου είναι απαραίτητο να μεταφερθεί (δεν έχει μονή κατεύθυνση). Κατά βασικό κανόνα η πηγή είναι τα φύλλα του φυτού και η κατανάλωση/αποθήκευση της παραγόμενης σακχαρόζης βρίσκεται στις ρίζες και στο βλαστό. Ωστόσο στην ερεύνα του INRA, στο μοντέλο source-to-sink ορίζεται ως πηγή (source) τα φύλλα του φυτού ενώ ως τελικός προορισμός της σακχαρόζης (sink) έχει οριστεί ο καρπός/φρούτο του φυτού δηλαδή η ντομάτα. (Bénard, et al., 2015) Η ανάπτυξη και η παραγωγή που θα επιφέρουν τα φυτά, ιδιαίτερα τα φρούτα, οφείλεται κατά βάση στο περιβάλλον και τις συνθήκες που αναπτύχθηκαν. Επίσης η ανάπτυξη των φυτών εξαρτάται σημαντικά από τους μεταβολίτες, τα μεταλλικά στοιχεία και το νερό που διατίθεται από τα όργανα του φυτού. Οι περισσότεροι μεταβολίτες παράγονται από τα φύλλα του φυτού όπου και ξεκινά η φωτοσύνθεση.

1.4.3 Μεταβολομικά δεδομένα για συγκεκριμένη ποικιλία ντομάτας

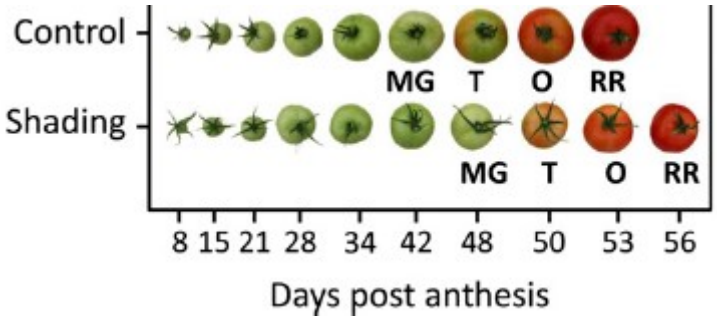
Στην ιστοσελίδα του NMRProcFlow (https://nmrprocflow.org) είναι δημοσιευμένη η μεθοδολογία για το πως παράχθηκαν τα δεδομένα τα οποία και χρησιμοποιήθηκαν σε αυτή τη εργασία. Πιο συγκεκριμένα πραγματοποιήθηκε από το INRA λεπτομερής έρευνα πάνω στις καθημερινές μεταβολές στη σύσταση των μεταβολιτών στα φύλλα και στο καρπό της ντομάτας.

Ντομάτες συγκεκριμένης ποικιλίας (*Solanum lycopersicum* cv. Moneymaker) καλλιεργήθηκαν σε θερμοκήπιο στη νοτιοδυτική Γαλλία από τον Ιούνιο μέχρι τον Σεπτέμβρη. Σε κάθε τσαμπί της ντομάτας υπάρχουν κατά μέσο όρο 6 ντομάτες. Η ολοκληρωμένη ανάπτυξη του φρούτου από την άνθηση μέχρι την ωρίμανση (red ripe stage) διήρκησε 55 μέρες. Οι συνθήκες περιβάλλοντος ορίστηκαν σε δύο, την ελεγχόμενη (control) με 276 φυτά και την υπό σκιά συνθήκη (shaded) με 138 φυτά. (Bénard, et al., 2015)

Οι καρποί και τα πλησιέστερα ώριμα φύλλα μαζεύτηκαν κατά τη διάρκεια δύο διαφορετικών ημερήσιων κύκλων (πρωί/βράδυ και υπό ηλιοφάνεια ή υπό σκιά). Παρατηρήθηκε πως αντιδρούν τα φυτά σε διαφορετικές συνθήκες για μερικές εβδομάδες. Σε αυτό το χρονικό διάστημα μετρήθηκε το ύψος των φυτών. Όσον αφορά τη

δειγματοληψία, επιλέχθηκαν 4-5 πανομοιότυπα φρούτα από διαφορετικά φυτά, τα οποία στη συνέχεια ζυγίστηκαν. Έπειτα επιλέχθηκαν φύλλα τα οποία ήταν κοντά στον καρπό. Τα φύλλα και τα φρούτα που επιλέχθηκαν ψύχθηκαν σε υγρό άζωτο και φυλάχθηκαν στους –80^{o}C μέχρι να γίνει η ανάλυση. Μετά από επεξεργασία δημιουργήθηκε παγωμένη σκόνη για ανάλυση με χρήση φασματοσκοπίας GC-MS ανάλυση και λυοφιλιωμένη (lyophilized) σκόνη για ανάλυση με ^{1}H-NMR και φασματοσκοπία μάζας (LC-MS). Για την στατιστική ανάλυση χρησιμοποιήθηκαν δείγματα μεταβολιτών DW –dry weight. Χρησιμοποιήθηκαν οι εξής αναλύσεις: PCA, ANOVA, K-means clustering, Συσχέτιση Pearson (Pearson correlation) και Γκαουσιανό γραφικό μοντέλο (Gaussian graphical model).

Μέσω της φασματοσκοπίας NMR και της φασματοσκοπίας μάζας προσδιορίστηκαν 70 μεταβολίτες στα φύλλα και 60 μεταβολίτες στα φρούτα σε συνάρτηση με οικο-φυσιολογικές μετρήσεις. Τα δεδομένα των μεταβολιτών επεξεργάστηκαν με πολυμεταβλητή ανάλυση (multivariate analyses), μονομεταβλητή ανάλυση (univariate analyses) ή ανάλυση με ομαδοποιήσεις (clustering analyses) και δίκτυα συσχετίσεων (correlation networks). Στην υπό σκιά συνθήκη τα φυτά προσάρμοσαν την περιοχή των φύλλων μειώνοντας τη ανάγκη για άνθρακα, φανερώνοντας μικρές αλλαγές στη χημική σύσταση του φυτού. Όσον αφορά τα φύλλα, τα οποία και λειτουργούσαν ως πηγές σακχαρόζης, παρατηρήθηκε ότι κατά τη διάρκεια του ημερήσιου κύκλου υπήρχε ποικιλία σε διάφορους μεταβολίτες κυρίως σε όσους συμμετείχαν και συνδέονται άμεσα με την φωτοσύνθεση και την αναπνοή των φυτών. Αυτοί οι μεταβολίτες είχαν μέγιστες τιμές (στη συγκέντρωση τους) στη μέση της ημέρας και για τις δύο συνθήκες περιβάλλοντος. Ωστόσο οι αποθήκες άνθρακα ήταν περιορισμένες στα φύλλα της ντομάτας. Στα φρούτα παρουσιάστηκαν λιγότερες διακυμάνσεις στους μεταβολίτες. Μεταξύ των μεταβολιτών που παρουσίασαν αυξομειώσεις ήταν οργανικά οξέα. Τα καθημερινά μοτίβα διέφεραν στα φύλλα της ντομάτας αλλά ακόμα περισσότερο στα φρούτα στις διαφορετικές συνθήκες περιβάλλοντος. Πολλές από τις αλλαγές στη σύσταση των φύλλων και των φρούτων φαίνονται να εξαρτώνται από την άμεση σύνδεση διάφορων μεταβολικών διαδικασιών της ντομάτας με τη στιγμιαία παροχή σακχαρόζης στο φυτό. (Bénard, et al., 2015)



Εικόνα 2.2.3.1 Στάδια ωρίμανσης της ντομάτας σε διαφορετικές συνθήκες ανάπτυξης (ελεγχόμενη και υπό σκιά) (Biais, et al., 2014)

1.5 Προεπεξεργασία binning/bucketing

Η τμηματοποίηση του φάσματος (binning/bucketing) είναι μία μέθοδος που πραγματοποιείται σε δεδομένα προκειμένου να ελαχιστοποιήσει λάθη μικρών

παρατηρήσεων. Τα αρχικά δεδομένα διαχωρίζονται σε μικρότερα τμήματα γνωστά και ως bins/buckets και αντικαθίστανται με μία αριθμητική τιμή υπολογισμένη από το ίδιο τμήμα (bin). Αυτό, εξομαλύνει τα δεδομένα τα οποία εισήχθησαν και επίσης μπορεί να μειώσει τις πιθανότητες να υπερπροσαρμοστούν σε περίπτωση μικρού όγκου δεδομένων. Η υπερπροσρμογή (overfitting) υφίσταται όταν το στατιστικό μοντέλο περιγράφει και λειτουργεί σωστά μόνο στα δεδομένα που έχουμε ορίσει κατά τη δημιουργία του μοντέλου. Αποτελεί ένα σημαντικό πρόβλημα κατά τη δημιουργία στατιστικών μοντέλων καθώς στη χρήση του μοντέλου σε διαφορετική βάση δεδομένων, τα αποτελέσματα θα είναι λανθασμένα.

Ένα φάσμα (συχνοτήτων) ^{1}H-NMR αποτελείται από ένα σύνολο χιλιάδων ζευγών συχνότητας και έντασης. Επειδή ο όγκος του φάσματος είναι πολύ μεγάλος χρειάζεται μια διαδικασία επεξεργασίας με σκοπό την διευκόλυνση εξαγωγής χρήσιμης πληροφορίας. Σύμφωνα με τη μέθοδο της τμηματοποίησης (bucketing), το φάσμα χωρίζεται σε διαστήματα, τα λεγόμενα buckets, τα οποία περιγράφουν/περιλαμβάνουν μία συχνότητα και μία τιμή έντασης. Ιδανικά, κάθε φασματική κορυφή αντιστοιχεί σε ένα τμήμα και αντίστοιχα κάθε τμήμα αντιστοιχεί σε μία φασματική κορυφή. Αυτό ωστόσο δεν είναι δυνατό να συμβεί, καθώς αποσκοπούμε στη μείωση του όγκου της πληροφορίας. Αποτέλεσμα της διαδικασίας τμηματοποίησης είναι το αρχικό φάσμα να αποτελείται πλέον από τμήματα του φάσματος τα οποία αντιπροσωπεύουν ένα μικρότερο σύνολο τιμών συχνότητας-έντασης που είναι διαχειρίσιμο.

Ο αλγόριθμος ομοιόμορφης τμηματοποίησης (Uniform bucketing) χρησιμοποιεί τμήματα με μη μεταβλητά μεγέθη και χωρίζει αυτόματα σε ίσα μέρη το φάσμα. Το πλεονέκτημα αυτής της μεθόδου είναι ότι είναι γνωστό το μέγεθος των τμημάτων, οπότε είναι γνωστό και προβλέψιμο το μέγεθος των αποτελεσμάτων και των πληροφοριών που θα προκύψουν. Ωστόσο μέσω της ομοιόμορφης τμηματοποίξσης (Uniform bucketing), ο τρόπος ομαδοποίησης του φάσματος ενδέχεται να μη δώσει ορθά αποτελέσματα καθώς στις περισσότερες των περιπτώσεων το φάσμα ενός μεταβολίτη χωρίζεται σε δύο ή και παραπάνω τμήματα.

Ο αλγόριθμος ευφυούς προσαρμοστικής τμηματοποίησης (Adaptive Intelligent bucketing) χρησιμοποιεί τμήματα με μεταβλητά μεγέθη και αποφασίζει αυτόματα πότε θα σταματήσει τη διαδικασία τμηματοποίησης των φασμάτων εισόδου. Μία ακμή του τμήματος (bin edge) είναι ένα σημείο στον άξονα συχνότητας που διαχωρίζει το τμήμα σε δύο νέα. Σε κάθε βήμα του αλγορίθμου οι ακμές των τμημάτων βρίσκονται στις ίδιες συχνότητες για όλα τα φάσματα εισόδου. Στην αρχή της διαδικασίας, ο αλγόριθμος τοποθετεί όλο το φάσμα σε ένα τμήμα. Κάθε σημείο στο εύρος συχνοτήτων ενός συγκεκριμένου τμήματος αξιολογείται ως μία πιθανή νέα ακμή. Η πιθανή ακμή τμήματος διαχωρίζει το τρέχον τμήμα σε δύο νέα, η αποδοτικότητα των οποίων συγκρίνεται με την αποδοτικότητα του τρέχοντος τμήματος. Η αποδοτικότητα ενός τμήματος αξιολογείται μέσω της τιμής του τμήματος (bin value). Η τιμή του τμήματος είναι ένας αριθμός ο οποίος λειτουργεί σαν μέσος όρος. Αυτή η τιμή προσδιορίζει το κάθε τμήμα, δηλαδή το αρχικό τμήμα, και τα δύο νέα που πρόκειται να προκύψουν από το ίδιο τμήμα. Αν η τιμή του αρχικού τμήματος είναι μεγαλύτερη τότε το τμήμα θα μείνει ίδιο και δε θα διαχωριστεί σε νέα τμήματα. Αν η τιμή των δύο νέων τμημάτων είναι μεγαλύτερη από το αρχικό τμήμα τότε διαιρείται σε δύο νέα τμήματα. Η

διαδικασία επαναλαμβάνεται για κάθε τμήμα όσο το άθροισμα των τιμών των νέων τμημάτων ξεπερνά την τιμή του παλιού. (Meyer, et al., 2008)

1.6 Στατιστική ανάλυση με την γλώσσα προγραμματισμού R

Μετά την εξαγωγή των πινάκων από το λογισμικό πρόγραμμα NMRProcFlow ακολούθησε η διαδικασία σύγκρισης αποτελεσμάτων μέσω στατιστικής ανάλυσης. Οι στατιστικές μέθοδοι που χρησιμοποιήθηκαν είναι η Ανάλυση Κύριων Συνιστωσών (PCA) και η Ανάλυση Μερικών Ελαχίστων Τετραγώνων (PLS). Για να χρησιμοποιηθούν αυτές οι στατιστικές αναλύσεις χρησιμοποιήθηκε η γλώσσα προγραμματισμού R. Στον προγραμματισμό πολλές φορές χρησιμοποιούνται πακέτα-βιβλιοθήκες οι οποίες διαθέτουν συναρτήσεις ή και δεδομένα τα οποία με την δημιουργία κώδικα μπορούν να χρησιμοποιηθούν από τον οποιοδήποτε χρήστη. Το ίδιο έγινε στην στατιστική ανάλυση της παρούσας εργασίας όπου χρησιμοποιήθηκαν τα πακέτα ASICS και ROPLS.

1.6.1 ASICS

Μία από τις δυσκολίες στην επεξεργασία μεταβολομικών δεδομένων είναι η ανάλυση του μεταβολικού προφίλ/προτύπων στην φασματοσκοπία NMR. Γεγονός που οδηγεί στην ανάγκη εύρεσης ενός γενικευμένου μοντέλου ελέγχου για τον χαρακτηρισμό του μεταβολώματος. Το πακέτο αυτό χρησιμοποιείται για την ταυτοποίηση και ποσοτικοποίηση μεταβολιτών σε σύνθετα βιολογικά μίγματα. Το πακέτο ASICS βασίζεται στη στατιστική θεωρία και μπορεί αυτόματα να πραγματοποιήσει ταυτοποίηση και ποσοτικοποίηση μεταβολιτών σε φάσματα NMR. Πιο συγκεκριμένα χρησιμοποιείται ένα γραμμικό στατιστικό μοντέλο με σκοπό να προσδιοριστούν οι μεταβολίτες μέσω μιας βιβλιοθήκης φασμάτων μεταβολιτών που υπάρχει εγκατεστημένη στο πακέτο.

1.6.2 ROPLS

Για την σύγκριση των αποτελεσμάτων είναι απαραίτητη μια στατιστική ανάλυση. Μέσω του πακέτου είναι διαθέσιμοι διάφορες μέθοδοι στατιστικής ανάλυσης όπως PCA, PLS-DA και OPLS-DA. Η ποιότητα των αποτελεσμάτων των μοντέλων μπορεί να εκτιμηθεί από τις τιμές R^2 και Q^2 που προκύπτουν. Οι τιμές αυτές περιγράφουν κατά πόσο το μοντέλο συνάδει με τα δεδομένα καθώς και τη συνοχή και τη προβλεψιμότητα που παρουσιάζει κατά τη ταξινόμηση (π χ. Στα διαφορετικά στάδια ωρίμανσης της ντομάτας). Το R^2 προσδιορίζει κατά πόσο το μοντέλο μπορεί να αναγνωρίσει τα δεδομένα στα οποία και έχει εκπαιδευτεί ενώ το Q^2 προσδιορίζει κατά πόσο το μοντέλο μπορεί να ομαδοποιήσει ένα μέρος των δεδομένων στο οποίο δεν έχει εκπαιδευτεί το μοντέλο. Επίσης με το πακέτο αυτό παράγονται τα γραφήματα που θα αναλυθούν στη παρούσα διπλωματική εργασία.

1.6.3 Ανάλυση Κύριων Συνιστωσών, PCA

Η Ανάλυση Κύριων Συνιστωσών (Principal component analysis-PCA), αποτελεί μία γραμμική μέθοδο συμπίεσης δεδομένων η οποία βασίζεται στη δημιουργία νέου συστήματος μεταβλητών με βάση τις μεταβλητές του αρχικού συνόλου δεδομένων με σκοπό την ευκολότερη ανάλυση των δεδομένων. Αυτές οι νέες μεταβλητές είναι το αποτέλεσμα ενός γραμμικού συνδυασμού προερχόμενου από τις αρχικές μεταβλητές. Η πρώτη κύρια συνιστώσα (principal component) διατηρεί περισσότερες πληροφορίες δεδομένων σε σύγκριση με τη δεύτερη η οποία δεν διατηρεί πληροφορίες οι οποίες έχουν εισέλθει στη

πρώτη συνιστώσα (αυτό ισχύει και για τις επόμενες νέες μεταβλητές). Οι κύριες συνιστώσες δεν συσχετίζονται. Ο αριθμός των κύριων συνιστωσών είναι ίσος με τον αριθμό των αρχικών μεταβλητών και παρουσιάζει τις ίδιες πληροφορίες στατιστικής. Οι πρώτες συνιστώσες διατηρούν παραπάνω από το 90% της πληροφορίας από τα αρχικά δεδομένα με αποτέλεσμα τη μείωση του αριθμού των μεταβλητών.

Ο βασικός σκοπός αυτής της μεθόδου ανάλυσης είναι η μείωση των διαστάσεων του χώρου παρατήρησης που ερευνάται και αναλύεται. Αυτό επιτυγχάνεται με τη δημιουργία νέων γραμμικών συνδυασμών των μεταβλητών που χαρακτηρίζουν τα αντικείμενα που βρίσκονται υπό ανάλυση. Οι συνδυασμοί που προκύπτουν πρέπει να ικανοποιούν συγκεκριμένες μαθηματικές και στατιστικές προϋποθέσεις.

Μέσω των κύριων συνιστωσών που προέκυψαν, μπορούν να προσδιοριστούν στη συνέχεια οι αρχικές μεταβλητές που επηρέασαν ποιοτικά και ποσοτικά αυτές τις κύριες συνιστώσες . Δηλαδή, οι μεγαλύτερες συσχετίσεις μεταξύ των κυρίων συνιστωσών και των αρχικών μεταβλητών οδηγούν στην εύρεση των μεταβλητών που περιέχουν χρήσιμη, για την έρευνα, πληροφορία. Σκοπός της ανάλυσης κυρίων συνιστωσών (PCA) είναι η εξαγωγή των σημαντικών πληροφοριών από τον πίνακα δεδομένων, συμπιέζοντας το μέγεθος του συνόλου δεδομένων και διατηρώντας μόνο τις σημαντικές πληροφορίες, προκειμένου να απλουστευτεί η περιγραφή του συνόλου των δεδομένων και να αναλυθεί η δομή των παρατηρήσεων και των μεταβλητών. Στη μεταβολομική ανάλυση, η PCA είναι ιδιαίτερα χρήσιμη για την αρχική προσέγγιση και ανάλυση των δεδομένων αλλά και των εξωκείμενων παρατηρήσεων/ έκτροπων τιμών (outliers).

1.6.4 Ανάλυση Μερικών Ελαχίστων Τετραγώνων, PLS

Η ανάλυση μερικών ελαχίστων τετραγώνων (Partial Least Squares analysis, PLS) είναι μία τεχνική η οποία είναι επηρεασμένη από την Ανάλυση Κύριων Συνιστωσών, PCA. Αρχικά, η ανάλυση μερικών ελάχιστων τετραγώνων χρησιμοποιούνταν για να οριστεί πόσο μια παρατήρηση συμμετέχει στη δημιουργία ενός προβλεπόμενου μοτίβου. Πλέον χρησιμοποιείται για την επίλυση προβλημάτων διάκρισης και ταξινόμησης των παρατηρήσεων σε ομάδες. Είναι μία μέθοδος που και αυτή βοηθά στη μείωση των αρχικών μεταβλητών και κατηγοριοποιεί τις μεταβλητές σε δύο ή παραπάνω κατηγορίες.
Δηλαδή η ανάλυση μερικών ελαχίστων τετραγώνων συνδυάζεται, με την διαχωριστική ανάλυση (Partial Least Squares-Discriminant Analysis, PLS-DA). Η PLS-DA χρησιμοποιεί την πληροφορία για την κατηγοριοποίηση των δειγμάτων και βελτιώνει τον διαχωρισμό μεταξύ των δύο συγκρινόμενων ομάδων δειγμάτων. Η μέθοδος αυτή χρησιμοποιείται τόσο για οπτικοποίηση των αποτελεσμάτων, για την ανάδειξη μεταβολικών διαφορών και για τη δημιουργία ενός μοντέλου πρόβλεψης. Το μοντέλο πρόβλεψης είναι ουσιαστικά η δημιουργία ενός αλγόριθμου που θα μπορεί αν ορίσει νέες παρατηρήσεις σε μία από τις ομάδες. Ακόμα μία επιβλεπόμενη μέθοδος κατηγοριοποίησης, που προκύπτει από την τροποποίηση του αλγορίθμου της PLS, είναι η O-PLS (Orthogonal Partial Least Squares). Η O-PLS χρησιμοποιεί ορθογώνια διόρθωση σήματος, για να απομακρύνει έναν αριθμό μεταβλητών που διαφοροποιούν τα δείγματα χωρίς να σχετίζονται με την κατηγοριοποίηση των δειγμάτων, με αποτέλεσμα να επιτυγχάνεται μεγιστοποίηση της συνδιακύμανσης. Η μέθοδος OPLS βοηθά στην ερμηνεία των αποτελεσμάτων όταν υπάρχουν παράγοντες υπεύθυνοι για τη διαφοροποίηση των δειγμάτων της ίδιας ομάδας, διαφορετικά δεν διαθέτει κάποιο πλεονέκτημα έναντι της PLS. Απαραίτητο βήμα μετά τη δημιουργία των στατιστικών μοντέλων, με τη χρήση επιβλεπόμενων μεθόδων κατηγοριοποίησης, είναι η

αξιολόγηση τους και υπάρχουν αρκετές στατιστικές προσεγγίσεις για την αξιολόγηση των αποτελεσμάτων τέτοιων μοντέλων.

1.7 Περιγραφή μεθοδολογίας

Αρχικά έγινε η σύγκριση των δύο ειδών τμηματοποίησης, της ομοιόμορφης τμηματοποίησης (uniform bucketing) και της ευφυούς προσαρμοστικής τμηματοποίησης (Adaptive Intelligence bucketing) με σκοπό να οριστεί αυτό που παρέχει το καλύτερο αποτέλεσμα για το συγκεκριμένο σετ δεδομένων στο NMRProcFlow. Σύμφωνα με τη θεωρία αναμένονται καλύτερα αποτελέσματα από την ευφυή προσαρμοστική τμηματοποίηση γιατί τα τμήματα στα οποία θα χωριστεί το φάσμα δεν είναι προκαθορισμένα, καθώς η τμηματοποίηση του φάσματος καθορίζεται από τις κορυφές του φάσματος (τα τμήματα προσαρμόζονται με τις κορυφές του φάσματος). Αντίθετα με την ομοιόμορφη τμηματοποίηση, τα τμήματα του φάσματος είναι προκαθορισμένα, γεγονός που μπορεί να οδηγήσει στην ύπαρξη δύο ή και παραπάνω κορυφών σε ένα τμήμα. Αυτό μάλιστα είναι και το βασικό πρόβλημα με την ομοιόμορφη τμηματοποίηση του φάσματος.

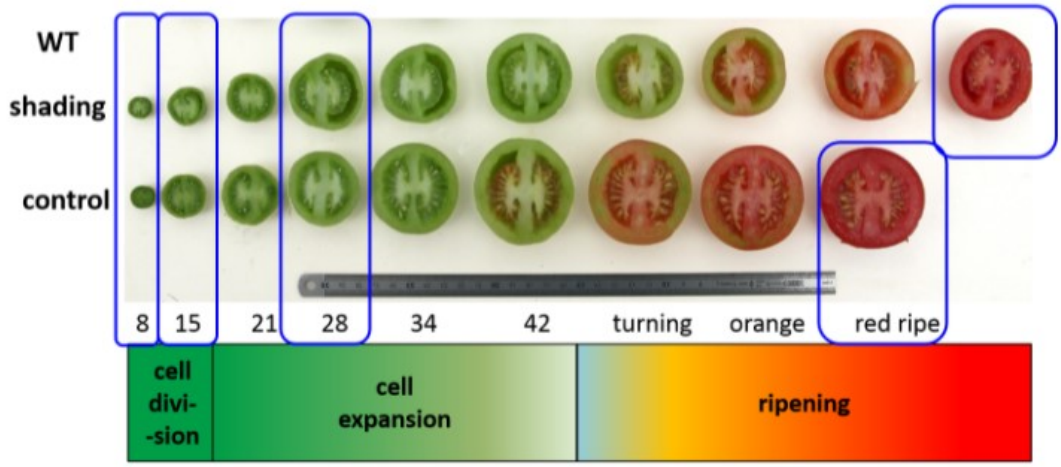
Το NMRProcFLow διαθέτει για κάθε είδος τμηματοποίησης διαφορετική διακριτική ικανότητα (resolution). Οπότε, θα πραγματοποιηθεί έλεγχος και για τα δύο είδη τμηματοποίησης, για το πόσο αυτή επηρεάζει το τελικό αποτέλεσμα που λαμβάνουμε από το NMRProcFlow. Για την ομοιόμορφη τμηματοποίηση οι τιμές της διακριτικής ικανότητας είναι από 0,01 έως 0,1 και για την ευφυή προσαρμοστική τμηματοποίηση οι τιμές της διακριτικής ικανότητας κυμαίνονται από 0,1 έως 0,5. Στις οδηγίες που προσφέρονται από την ιστοσελίδα, αναφέρεται ότι όσο μικρότερη η τιμή της διακριτικής ικανότητας (πχ. 0,01) τόσο μεγαλύτερη είναι η ευκρίνεια του προγράμματος, δηλαδή καλύτερη η διακριτική ικανότητα καθώς ‘βλέπει’ περισσότερες κορυφές. Για αυτό το λόγο, επιλέχθηκαν τιμές διακριτικής ανάλυσης που θα είχαν ως αποτέλεσμα μεγαλύτερη ευκρίνεια. Για την περαιτέρω επεξεργασία χρησιμοποιήθηκαν ο πίνακας δεδομένων (data matrix), ο πίνακας της σχέσης σήματος προς θόρυβο (Signal to Noise Ratio- SNR matrix) και ο πίνακας τμημάτων (bucket matrix), τα οποία μπορούν να εξαχθούν αυτόματα από το NMRProcFlow.

Λήφθηκαν υπόψη οι πίνακες που αναφέρθηκαν (data matrix, SNR matrix, bucket matrix), και το συμπέρασμα ήταν ότι η ευφυής προσαρμοστική τμηματοποίηση (AI bucketing) είναι η καλύτερη επιλογή (βλ. Πίνακα 3.1.1 & 3.1.2), γιατί διαχωρίζει καλύτερα το φάσμα σε τμήματα. Εφόσον επιλέχθηκε η καλύτερη μέθοδος τμηματοποίησης του φάσματος, ακολουθεί έλεγχος σε σχέση με την καλύτερη διακριτική ικανότητα μέσω της οποίας μπορεί να διαχωριστεί καλύτερα το σετ δεδομένων. Καλύτερη διακριτική ικανότητα θεωρούμε εκείνη η οποία θα έχει υψηλό SNR (Signal to Noise Ratio) και έχει καλύτερη απόδοση στις στατιστικές αναλύσεις. Πραγματοποιήθηκε σύγκριση στις διαφορετικές διακριτικές ικανότητες δηλαδή για τιμές 0,1 , 0,3 και 0,5 για την ευφυή προσαρμοστική τμηματοποίηση. Χρησιμοποιούμε αυτές τις τιμές ως ένα αντιπροσωπευτικό δείγμα ώστε να ελέγξουμε το εύρος των τιμών της διακριτικής ικανότητας και τι διαφορές προσφέρει στα τελικά αποτελέσματα. Αρχικά έγινε οπτική σύγκριση του φάσματος με σκοπό να ελεγχθεί αν η τμηματοποίηση των κορυφών δε γίνεται σωστά, δηλαδή κόβεται μια κορυφή στη μέση

και ορίζεται διαφορετικό τμήμα (Εικόνες 3.1.1-3.1.15). Επίσης έγινε σύγκριση της ταξινόμησης μεταξύ των συνθηκών όπου αναπτύχθηκε η ντομάτα (condition) και των σταδίων ωρίμανσης της ντομάτας (stage) (Πίνακας 3.2.1). Με αυτή τη σύγκριση θα βγει πόρισμα αν το σετ δεδομένων μπορεί να ομαδοποιηθεί καλύτερα με βάση το condition ή το stage, κοινώς θα γίνει κατανοητό αν το μεταβολικό προφίλ της ντομάτας καθορίζεται περισσότερο με βάση τις συνθήκες ανάπτυξης ή με βάση τα στάδια ανάπτυξης της ντομάτας (κεφάλαιο 4 - Αποτελέσματα). Για αυτή τη σύγκριση χρησιμοποιήθηκε το ROPLS πακέτο και οι στατιστικές αναλύσεις PCA, PLS-DA και OPLS-DA.

Ακολουθεί η ταυτοποίηση των μεταβολιτών μέσω του πακέτου ASICS (Πίνακας 3.3.1) και η αξιολόγηση της ταυτοποίησης με το πακέτο ROPLS (Πίνακας 3.3.2). Πραγματοποιήθηκε ταυτοποίηση των μεταβολιτών για τις διαφορετικές διακριτικές ικανότητες. Μέσω των διαφορετικών διακριτικών ικανοτήτων παρατηρήθηκε ότι αλλάζει ο αριθμός των τμημάτων, καθώς και το είδος των κορυφών που υπάρχουν μέσα στο κάθε τμήμα.

Η ανάπτυξη των φρούτων χωρίζεται σε τρεις βασικές φάσεις, στην κυτταρική διαίρεση, στην εξάπλωση των κυττάρων και στην ωρίμανση. Τα δεδομένα είναι διαχωρισμένα με βάση τις ημέρες πριν την άνθηση τους –Days Post Anthesis (DPA). Όσον αφορά τα δεδομένα στα οποία βασίζεται η διπλωματική, η πρώτη φάση αντιστοιχεί στο στάδιο 8DPA και 15DPA, η δεύτερη φάση το στάδιο 28DPA και τέλος η τρίτη η φάση αντιστοιχεί στο στάδιο 55DPA. Σύμφωνα με τις φάσεις που έχουν τα φρούτα και τα στάδια που έχουν τα δεδομένα αναμένεται όσα στάδια ανήκουν σε διαφορετικές φάσεις ανάπτυξης να είναι διακριτά διαφορετικές και να προκύπτουν ομάδες (clusters) που να μη συμπίπτουν μεταξύ τους. Σύμφωνα με όλα τα παραπάνω έχουν προκύψει τρεις διαφορετικοί τρόποι επεξεργασίας των δεδομένων από το NMRProcFlow με τους οποίους είναι διακριτός ο διαχωρισμός των διαφορετικών σταδίων στην ανάπτυξη της ντομάτας.



Εικόνα 2.5.1 Τα στάδια ανάπτυξης της ντομάτας με βάση τις ημέρες DPA και τις συνθήκες ανάπτυξης της.

Free access: https://nmrprocflow.org/ex1

Σε αυτό το σημείο επιχειρήθηκε η ταυτοποίηση και ποσοτικοποίηση των μεταβολιτών του φάσματος. Υπάρχουν διαφορετικοί τρόποι και συνδυασμοί για την ταυτοποίηση των

μεταβολιτών σε ένα σετ δεδομένων. Ένας τρόπος [Σύγκριση (1)] είναι να χρησιμοποιηθούν όλα τα δείγματα από όλα τα στάδια και να γίνει ταυτοποίηση και ποσοτικοποίηση όλων μαζί και ύστερα να γίνει σύγκριση των σταδίων ανά δύο (Εικόνα 3.4.1-6 & Πίνακας 3.4.1). Με αυτό τον τρόπο, όλα τα δεδομένα ανεξάρτητα αν χρησιμοποιηθούν στη σύγκριση έχουν συμβάλλει στα τελικά αποτελέσματα της ποσοτικοποίησης και της ταυτοποίησης των μεταβολιτών. Ένας άλλος τρόπος [Σύγκριση(2)] είναι να ταυτοποιηθεί και να ποσοτικοποιηθεί κάθε στάδιο ξεχωριστά και έπειτα να γίνει η σύγκριση των σταδίων μεταξύ τους (Εικόνα 3.4.7-12 & πίνακας 3.4.2). Με αυτό τον τρόπο, το μεταβολικό προφίλ κάθε σταδίου παραμένει ανεπηρέαστο από τα άλλα στάδια. Τέλος, μπορεί να χρησιμοποιηθούν τα δείγματα των δύο σταδίων που θέλουμε να συγκρίνουμε να πραγματοποιηθεί σε αυτά τα δύο ταυτοποίηση και ποσοτικοποίηση και ύστερα να γίνει η σύγκριση (Πίνακας 3.4.3 & 3.4.4).

ΑΠΟΤΕΛΕΣΜΑΤΑ

1.8 Πρώτη φάση δοκιμών και αποτελεσμάτων

Βάζοντας τις προκαθορισμένες παραμέτρους (default settings) για τα δύο είδη τμηματοποίησης (bucketing) προκύπτουν για την ομοιόμορφη τμηματοποίηση (uniform bucketing) 243 τμήματα ενώ για την ευφυή προσαρμοστική τμηματοποίηση (AI bucketing) 327 τμήματα.

Πίνακας 3.1.1. Αριθμός των τμημάτων για τα διαφορετικά είδη τμηματοποίησης.

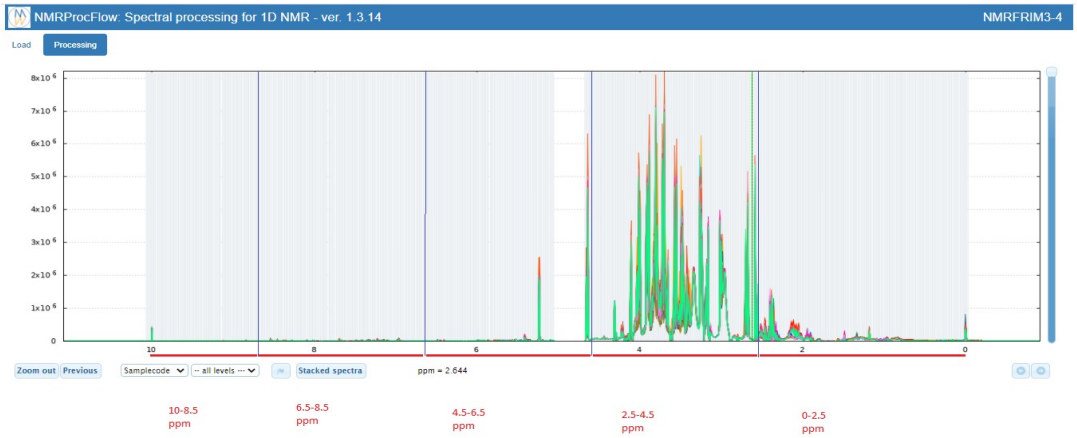
Είδη Bucketing	**Αριθμός Buckets**	**SNR**	**Προκαθορισμένη Διακριτική Ικανότητα**
Uniform	243	747,0	0,04
Adaptive Intelligence	327	528,5	0,5

Πίνακας 3.1.2. Αποτελέσματα για τα διαφορετικά είδη τμηματοποίησης στις ακραίες τιμές της διακριτικής ικανότητας.

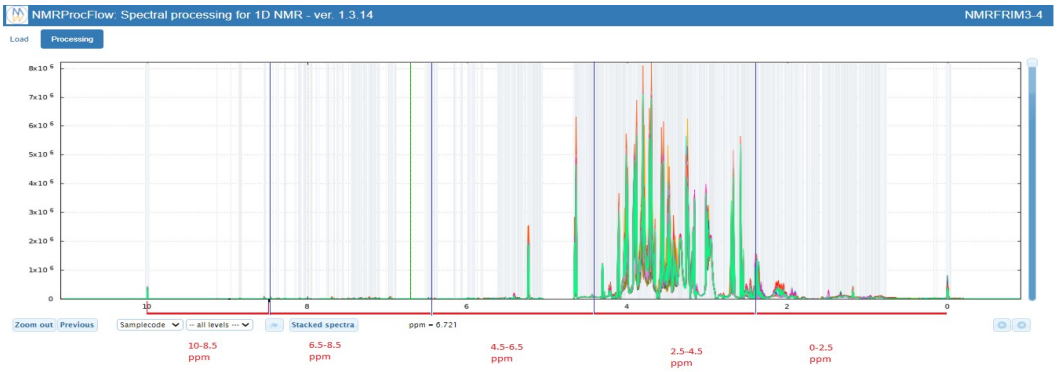
			0-2,5 ppm	2,5-4,5 ppm	4,5-6,5 ppm	6,5-8,5 ppm	8,5-10 ppm	
Ομοιόμορφη Τμηματοποίηση	Διακριτική Ικανότητα = (0.01)	Τμήματα	249	198	162	194	148	= 954
		Μέγιστη τιμή κορυφής (counts)	429,2	3138,8	1089,0	43,0	19,1	
		Μέγιστο SNR	2104,7	7559,8	4480,2	53,7	32,1	SNR πίνακας 393,6 (μέση τιμή)
	Διακριτική Ικανότητα = (0.05)	Τμήματα	50	40	34	40	30	= 194
		Μέγιστη τιμή κορυφής (counts)	1436,4	11034,1	2495,4	102,5	66,6	
		Μέγιστο SNR	1454,7	7563,2	4629,3	46,6	29,4	SNR πίνακας 693,7 (μέση τιμή)
Ευφυής Προσαρμοστική Τμηματοποίηση	Διακριτική Ικανότητα = (0.1)	Τμήματα	195	203	117	102	9	=626
		Μέγιστη τιμή κορυφής (counts)	658,0	7218,9	1522,3	64,7	25,9	
		Μέγιστο SNR	2104,7	2539,4	4629,3	52,0	29,4	SNR πίνακας 557,9 (μέση τιμή)
	Διακριτική Ικανότητα = (0.5)	Τμήματα	179	198	108	86	9	=580
		Μέγιστη τιμή κορυφής (counts)	675,0	7405,7	1376,6	450,9	26,4	
		Μέγιστο SNR	2104,7	2539,4	4629,3	22,1	29,4	SNR πίνακας 582,4 (μέση τιμή)

Με τον παραπάνω πίνακα 3.1.2 μπορούμε να κατανοήσουμε τη δομή του φάσματος της ντομάτας η οποία και απεικονίζεται στην παρακάτω εικόνα (Εικόνα 3.1.1). Στην ομοιόμορφη τμηματοποίηση παρατηρούμε μεγάλες διαφορές με την αλλαγή της διακριτικής ικανότητας. Σε υψηλή διακριτική ικανότητα (0,01) υπάρχουν περισσότερα τμήματα σε σχέση με τη μικρότερη διακριτική ικανότητα (0,05). Όσον αφορά τις μέγιστες τιμές των κορυφών παρατηρούμε ότι η μέγιστη τιμή με βάση την ομοιόμορφη τμηματοποίηση είναι στο εύρος 2,5-4,5 ppm. Μάλιστα με τη διακριτική ικανότητα 0,05 παρατηρούμε ότι η μεγαλύτερη τιμή σε όλο τον πίνακα είναι 11034,1 counts. Παρατηρούμε μεγάλη διαφορά στη μέση τιμή SNR για τις διαφορετικές διακριτικές ικανότητες στην ομοιόμορφη τμηματοποίηση.
Στην ευφυή προσαρμοστική τμηματοποίηση παρατηρούμε ότι στις διαφορετικές διακριτικές ικανότητες υπάρχει μια αναλογία στον αριθμό των τμημάτων που έχουν δημιουργηθεί. Δηλαδή και στις δύο διακριτικές ικανότητες το εύρος 2,5-4,5 ppm έχει τα περισσότερα τμήματα και το εύρος 8,5-10 ppm έχει τα λιγότερα τμήματα. Παρατηρούμε επίσης ότι το SNR είναι παρόμοιο για διαφορετικές διακριτικές ικανότητες. Επίσης, παρατηρούνται παρόμοιες τιμές στα counts. Εξαίρεση αποτελεί το εύρος 6,5-8,5 ppm όπου τα counts για υψηλή διακριτική ικανότητα (0,1) είναι 64,7 counts ενώ για χαμηλή διακριτική ικανότητα (0,5) είναι 450,9 counts

a.



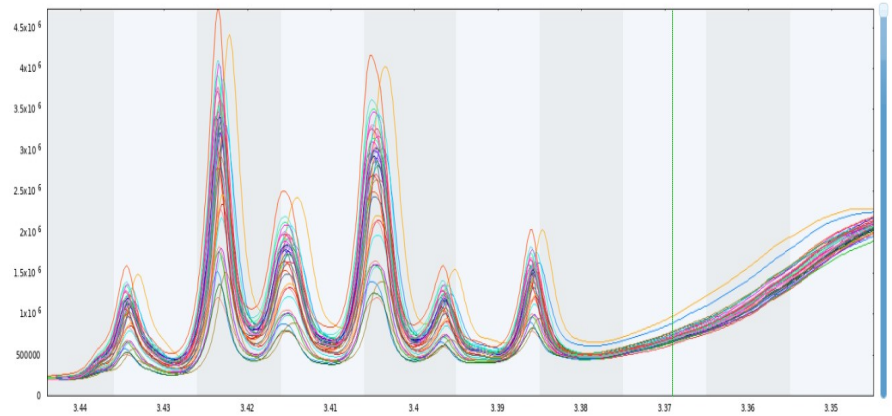
b.



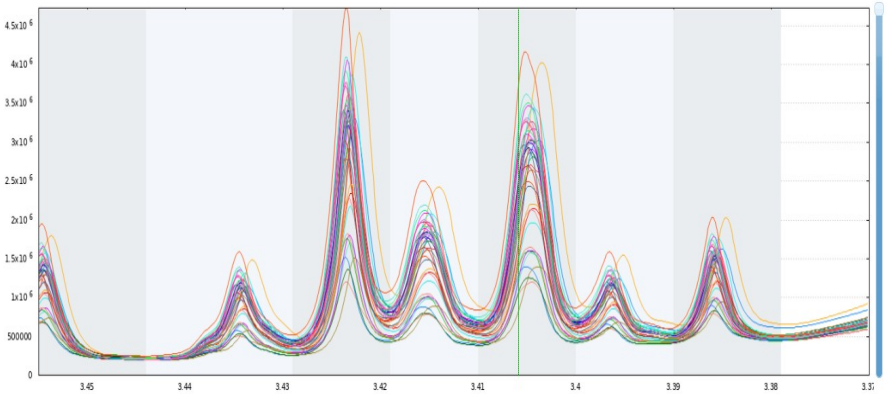
Εικόνα 3.1.1. Απεικονίζονται τα τμήματα για a. Ομοιόμορφη τμηματοποίηση, b. Ευφυής προσαρμοστική τμηματοποίηση

Παρατηρούμε στην Εικόνα 3.1.1 ότι η ομοιόμορφη τμηματοποίηση έχει πολλά τμήματα σε όλο το φάσμα ακόμα και σε ppm όπου φαίνεται ότι οι κορυφές είναι πολύ μικρές (δηλαδή πληροφορίες που μπορεί να είναι και θόρυβος-άχρηστη πληροφορία). Αντίθετα, η ευφυής προσαρμοστική τμηματοποίηση φαίνεται ότι έχει πιο στοχευμένη τμηματοποίηση και αποφεύγει να συμπεριλάβει περιοχές με χαμηλό σήμα ή και θόρυβο.

a.



b.

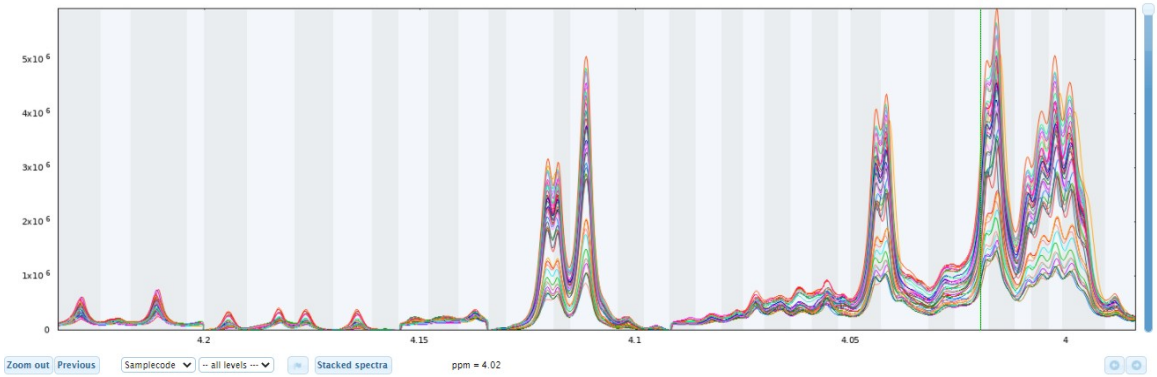


Εικόνα 3.1.2. Απεικονίζονται a.Ομοιόμορφη τμηματοποίηση με διακριτική ικανότητα 0,01 b.Ευφυής προσαρμοστική τμηματοποίηση με διακριτική ικανότητα 0,1

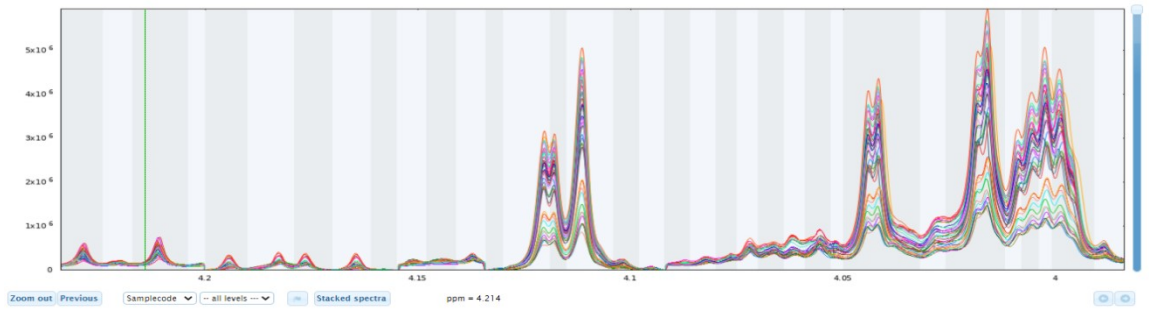
Στην Εικόνα 3.1.2 παρατηρούμε το ίδιο εύρος ppm τον τρόπο με τον οποίο τμηματοποιείται από τα δύο διαφορετικά είδη τμηματοποίησης στην μέγιστη διακριτική ικανότητα (δηλαδή έχουμε τον καλύτερο δυνατό διαχωρισμό μεταξύ των κορυφών του φάσματος).

Για τη σύγκριση μεταξύ των διαφορετικών διακριτικών ικανοτήτων για την ευφυή προσαρμοστική τμηματοποίηση πραγματοποιήθηκε οπτική εξέταση με σκοπό να παρατηρηθούν τυχόν αστοχίες στη τμηματοποίηση του φάσματος. Τα αποτελέσματα είναι τα εξής:

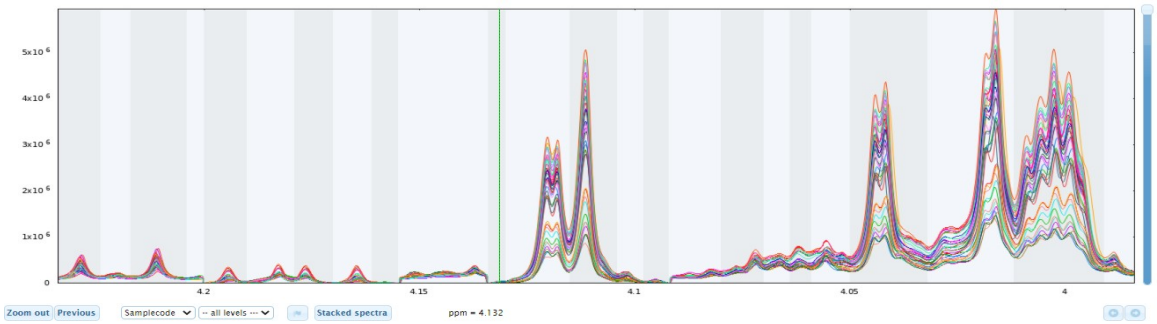
a.



b.

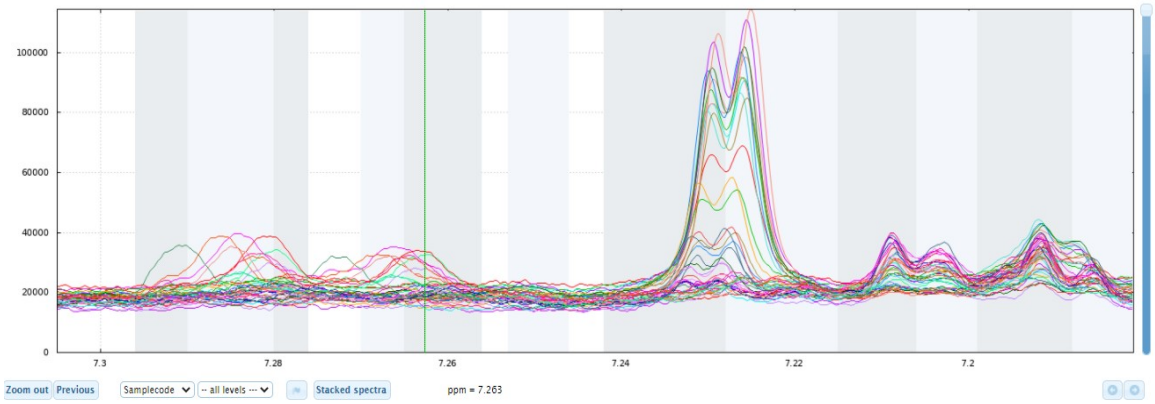


c.

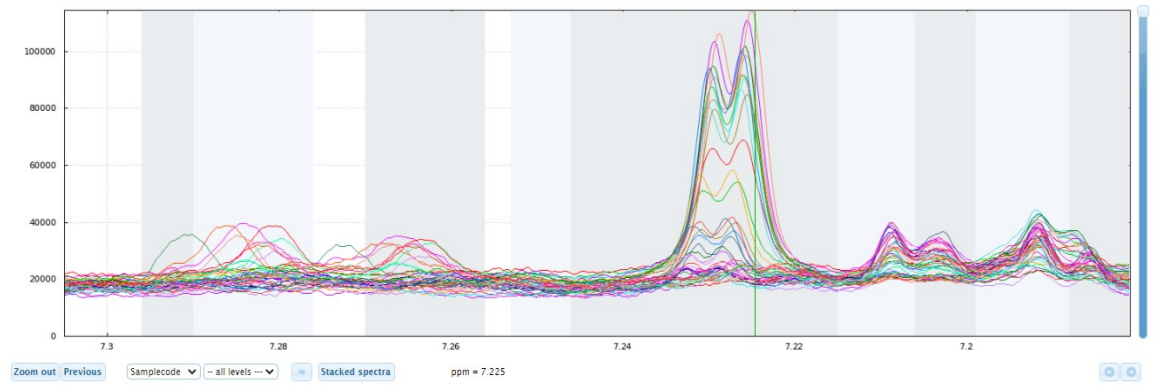


Εικόνα 3.1.3. Τα φάσματα των δειγμάτων από 0-10ppm με διαφορετικές διακριτικές ικανότητες στην ευφυή προσαρμοστική τμηματοποίηση. a.resolution 0,1 b. resolution 0,3 c.resolution 0,5

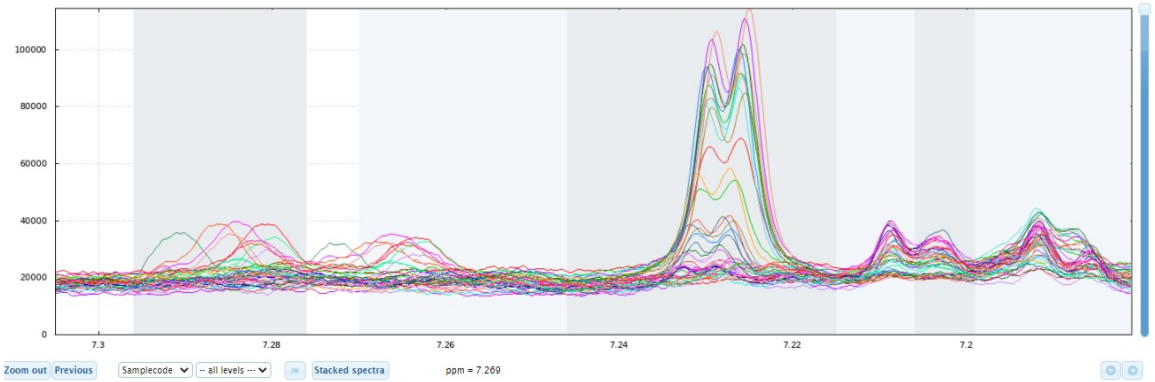
a.



b.

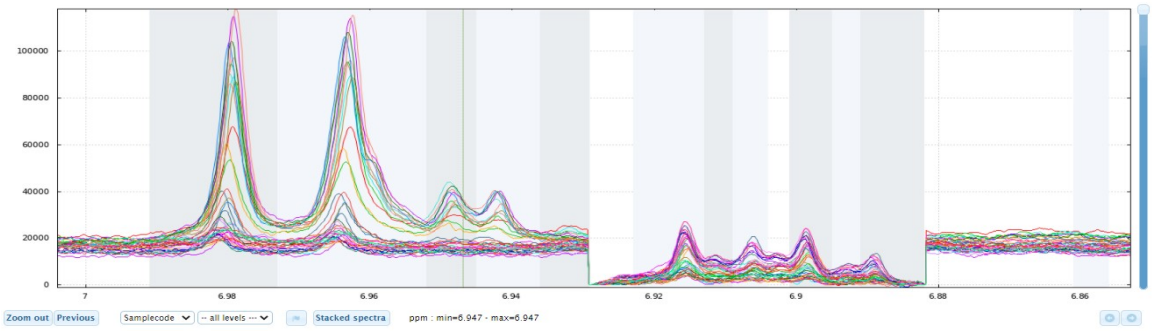


c.

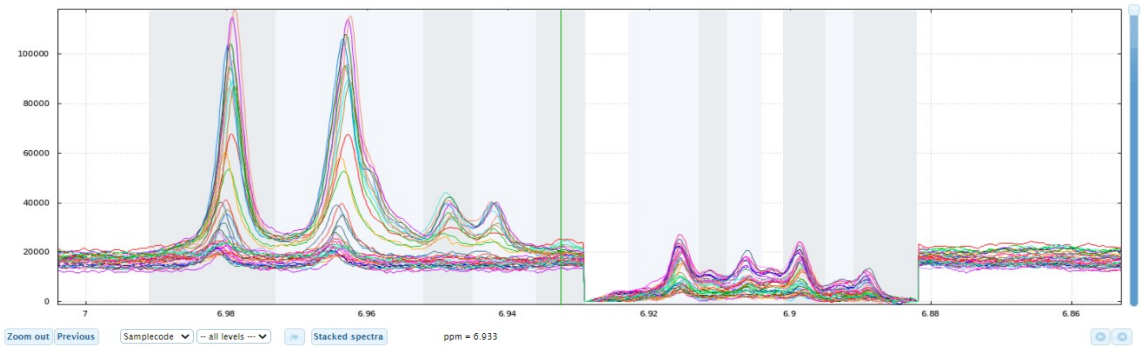


Εικόνα 3.1.4. Τα φάσματα των δειγμάτων από 7,2-7,3ppm με διαφορετικές διακριτικές ικανότητες στην ευφυή προσαρμοστική τμηματοποίηση. a.resolution 0,1 b. resolution 0,3 c.resolution 0,5

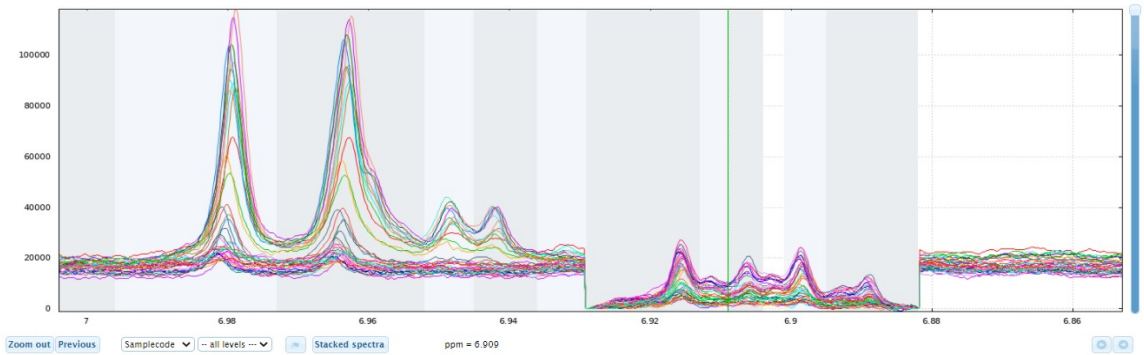
a.



b.



c.

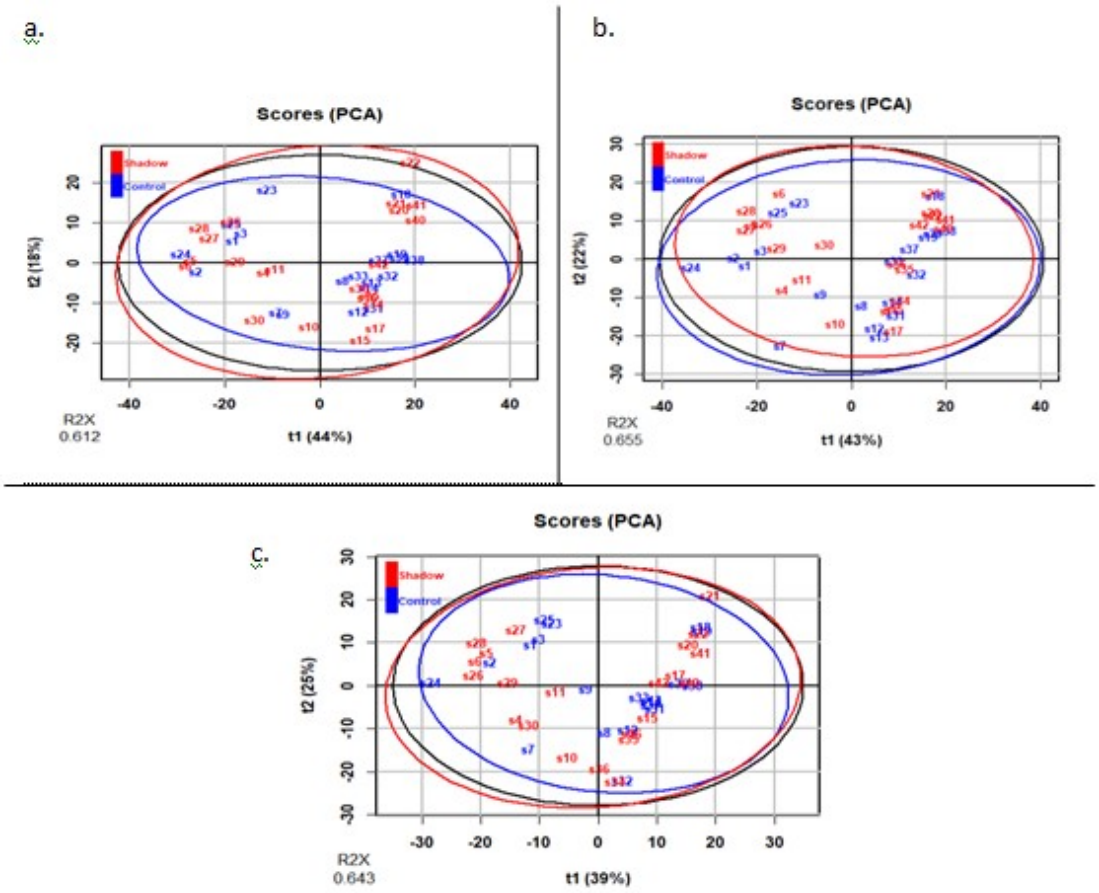


Εικόνα 3.1.5. Τα φάσματα των μεταβολιτών από 6-7ppm με διαφορετικές διακριτικές ικανότητες στην ευφυή προσαρμοστική τμηματοποίηση. a.resolution 0,1 b.resolution 0,3 c.resolution 0,5

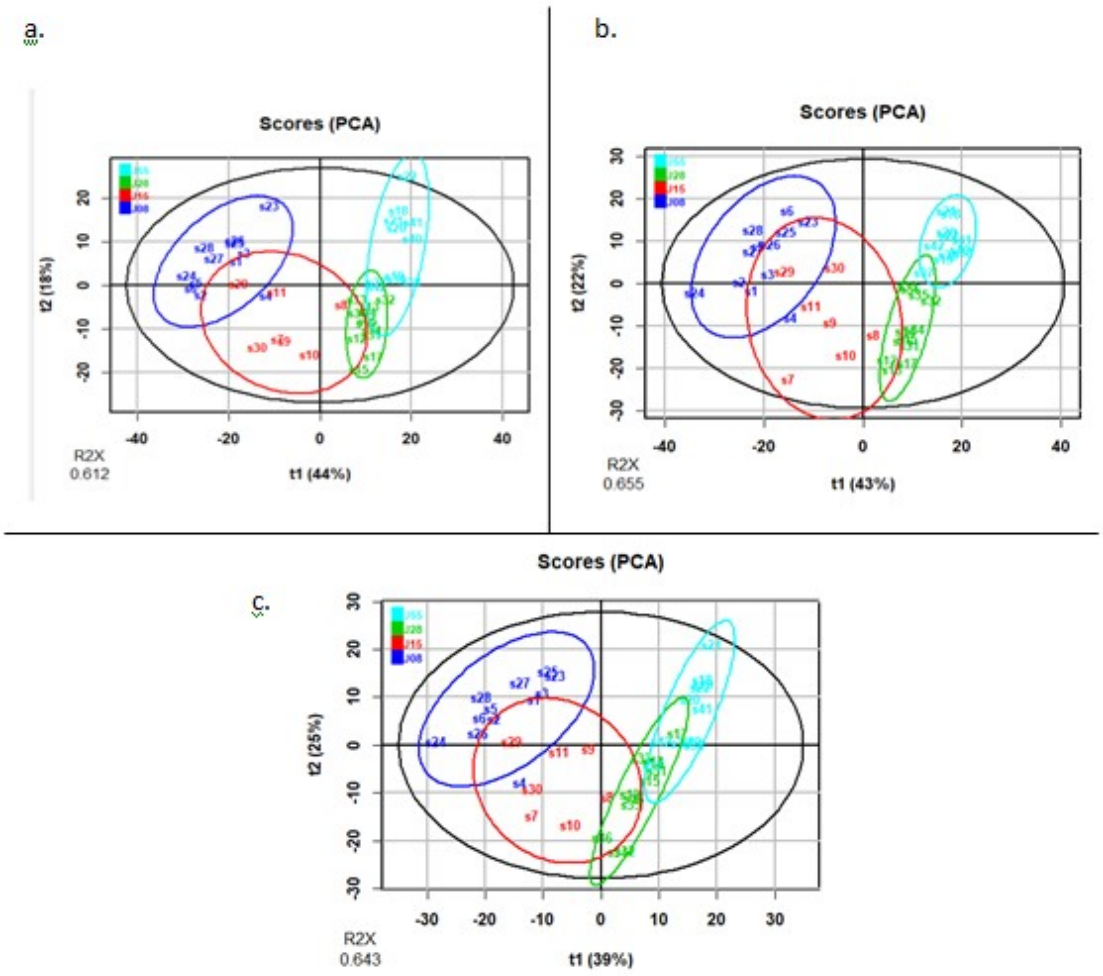
1.9 Στατιστική Ανάλυση των συνθηκών ανάπτυξης με τα στάδια ωρίμανσης της ντομάτας με τμηματοποίηση βασισμένη σε διαφορετικές διακριτικές ικανότητες

Σε αυτή την υποενότητα παρουσιάζονται τα αποτελέσματα της στατιστικής ανάλυσης των συνθηκών ανάπτυξης της ντομάτας με δεδομένα τμηματοποίησης βασισμένα σε διαφορετικές διακριτικές ικανότητες. Επίσης, παρουσιάζονται και τα αποτελέσματα της στατιστικής ανάλυσης των σταδίων ωρίμανσης της ντομάτας για διαφορετικές διακριτικές

ικανότητες. Η στατιστική ανάλυση γίνεται αφού τα δεδομένα έχουν προεπεξεργαστεί με την τεχνική της ευφυούς προσαρμοστικής τμηματοποίησης.

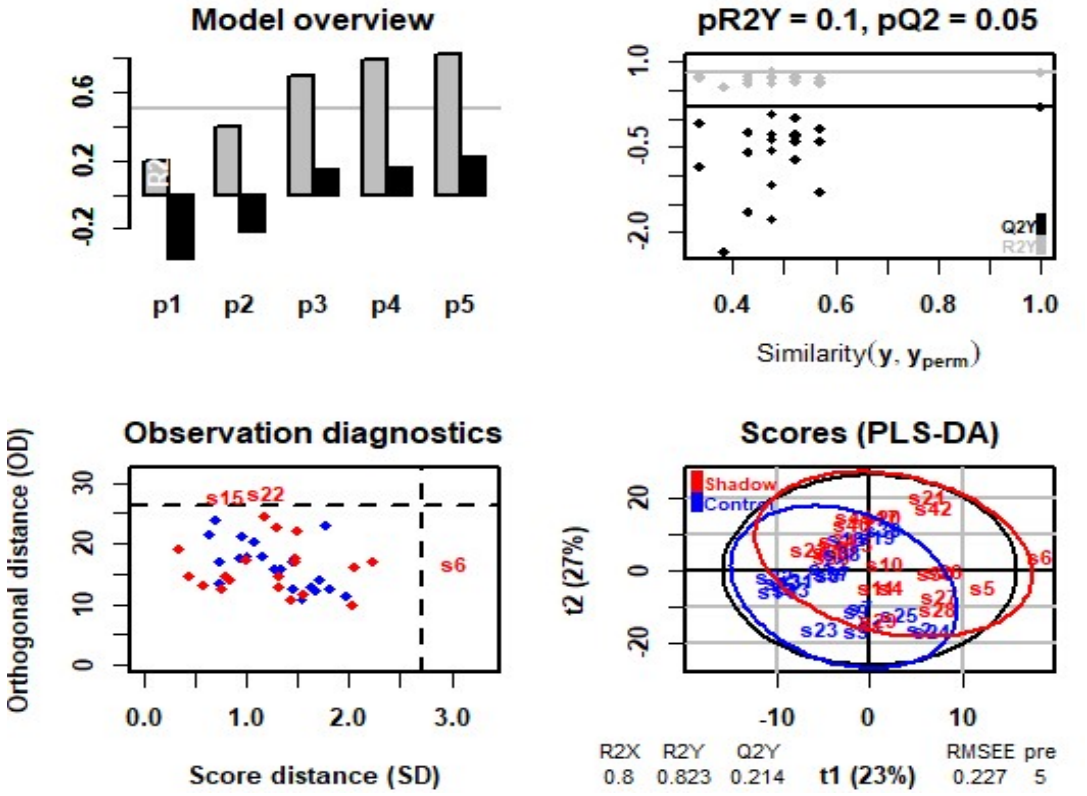


Εικόνα 3.2.1. Περιληπτικό γράφημα PCA ανάλυσης για διαφορετικές διακριτικές ικανότητες με βάση της συνθήκες ανάπτυξης της ντομάτας (condition). a.resolution 0,1
b. resolution 0,3 c. resolution 0,5

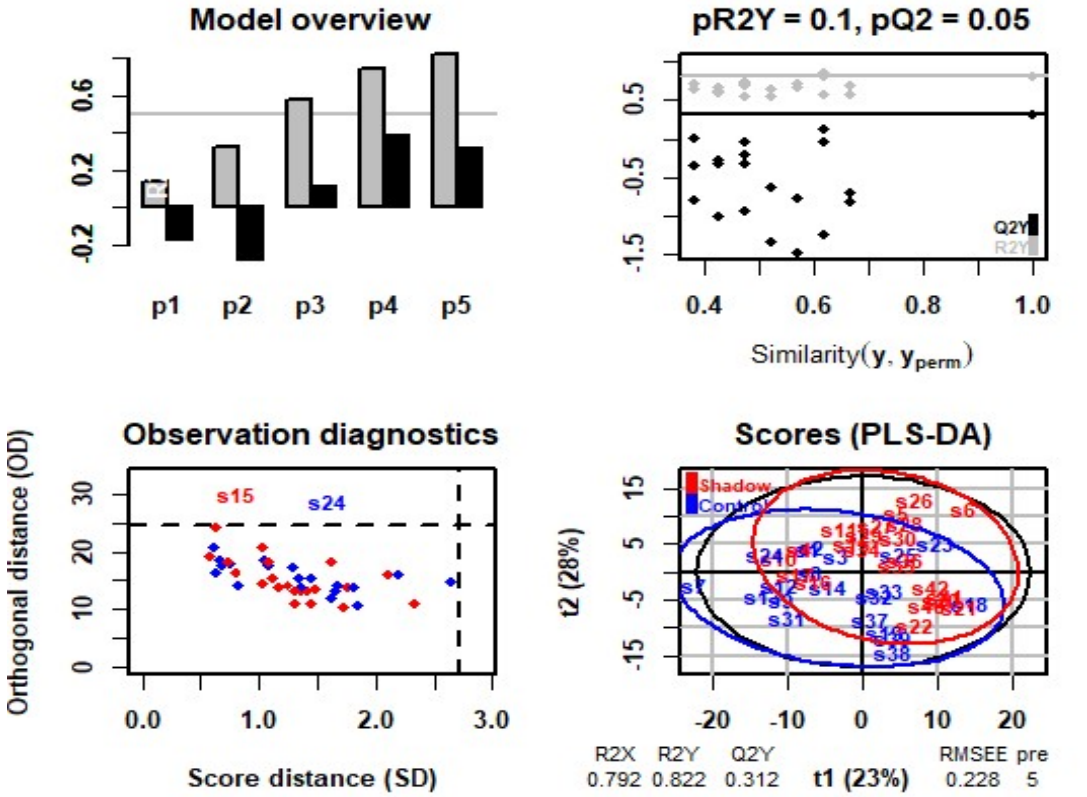


Εικόνα 3.2.2 Περιληπτικό γράφημα PCA ανάλυσης για διαφορετικές διακριτικές ικανότητες με βάση τα στάδια ωρίμανσης της ντομάτας (stage). a. resolution 0,1 b. resolution 0,3 c. resolution 0,5

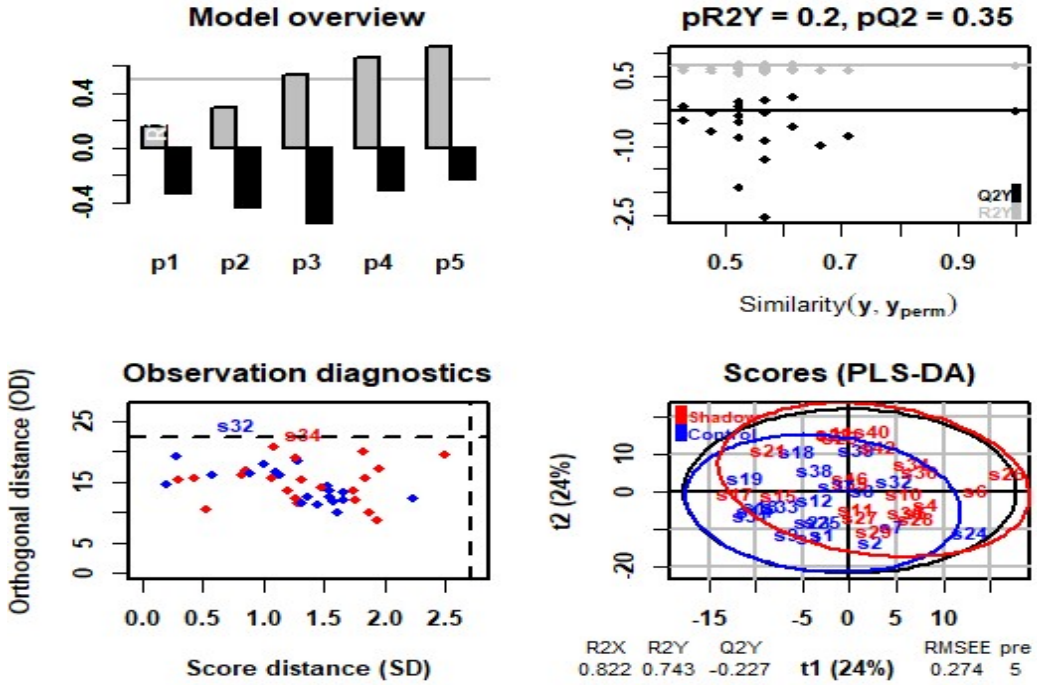
a.


Model overview
p1
p2
p3
p4
p5
pR2Y = 0.1, pQ2 = 0.05
Q2Y
R2Y
Similarity(y, y_perm)
Observation diagnostics
Orthogonal distance (OD)
Score distance (SD)
s15
s22
s6
Scores (PLS-DA)
Shadow
Control
t2 (27%)
t1 (23%)
R2X R2Y Q2Y RMSEE pre
0.8 0.823 0.214 0.227 5

b.

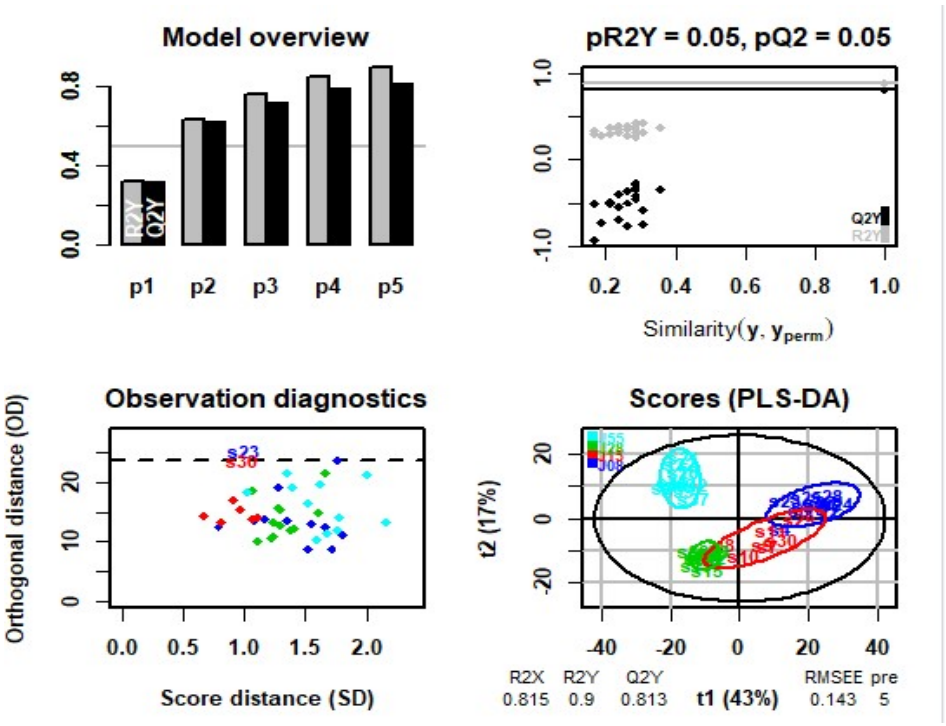

Model overview
p1
p2
p3
p4
p5
pR2Y = 0.1, pQ2 = 0.05
Q2Y
R2Y
Similarity(y, y_perm)
Observation diagnostics
Orthogonal distance (OD)
Score distance (SD)
s15
s24
Scores (PLS-DA)
Shadow
Control
t2 (28%)
t1 (23%)
R2X R2Y Q2Y RMSEE pre
0.792 0.822 0.312 0.228 5

c.

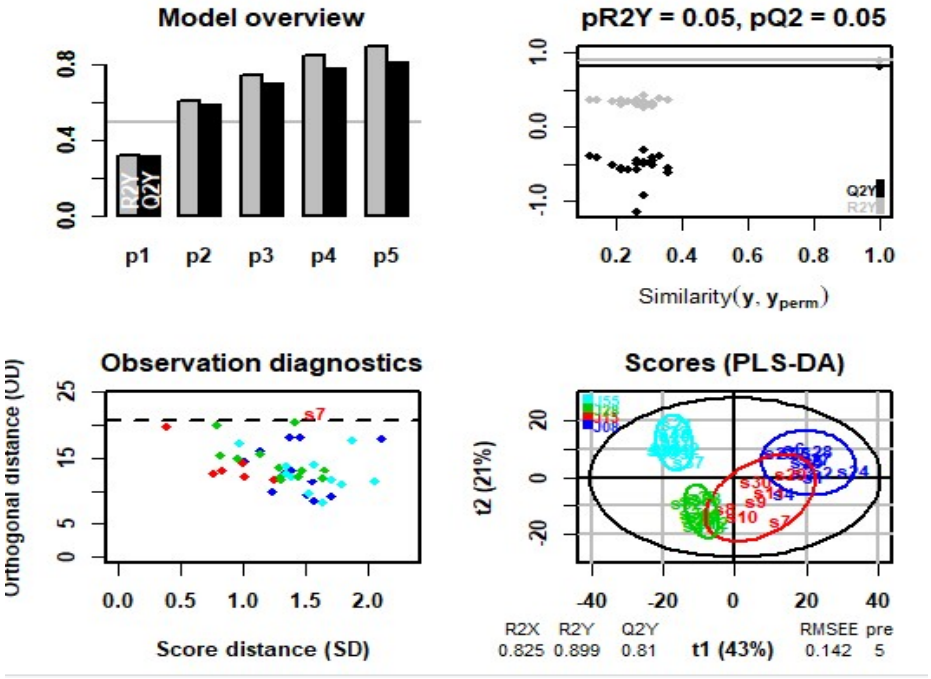


Εικόνα 3.2.3 PLS-DA μοντέλο με διαφορετικές διακριτικές ικανότητες με βάση τις συνθήκες ανάπτυξης της ντομάτας. a.resolution 0,1 b. resolution 0,3 c. resolution 0,5

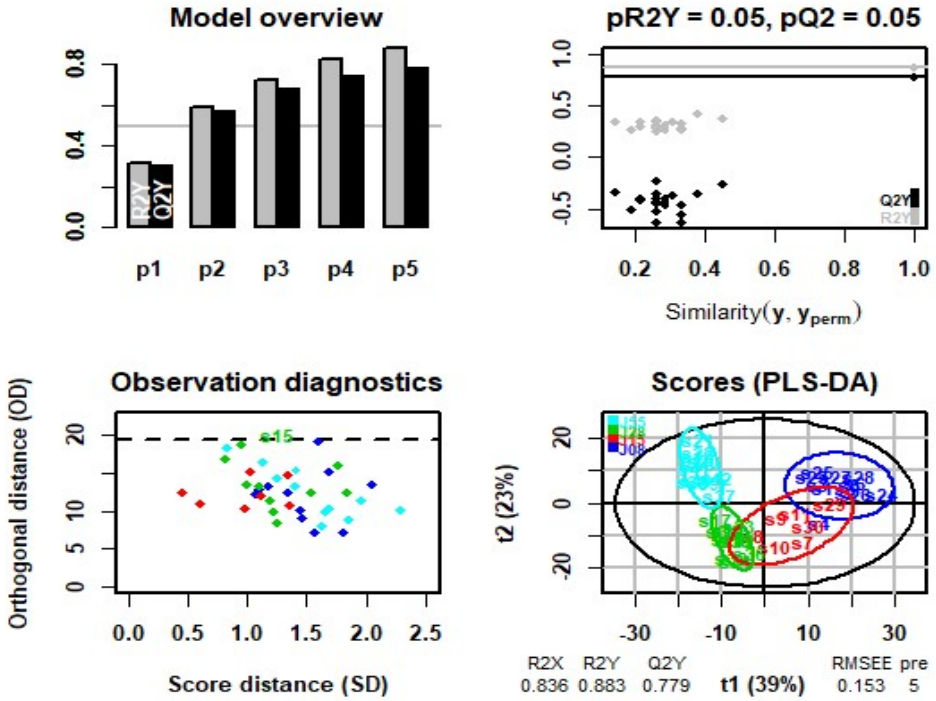
a.



b.



c.



Εικόνα 3.2.4 PLS-DA μοντέλο για διαφορετικές διακριτικές ικανότητες με βάση τα στάδια ανάπτυξης της ντομάτας. a.resolution 0,1 b. resolution 0,3 c. resolution 0,5

Πίνακας 3.2.1 Συγκεντρωτικά όλα τα αποτελέσματα από τη σύγκριση των διαφορετικών διακριτικών ικανοτήτων για τις συνθήκες ανάπτυξης (condition) και για τα στάδια ωρίμανσης (stage) της ντομάτας αντίστοιχα.

	Condition			Stage		
	0,1 resolution	0,3 resolution	0,5 resolution	0,1 resolution	0,3 resolution	0,5 resolution
PCA	R^2=0,612	R^2=0,655	R^2=0,643	R^2=0,612	R^2=0,655	R^2=0,643
PLS-DA	R^2=0,823 Q^2=0,214	R^2=0,822 Q^2=0.312	R^2=0,743 Q^2=-0,227	R^2=0,9 Q^2=0,813	R^2=0,899 Q^2=0,81	R^2=0,883 Q^2=0,779
OPLS-DA	R^2=0,395 Q^2=-0,183	R^2=0,319 Q^2=-0,026	R^2=0,299 Q^2=-0,176	-	-	-
Targeted	PLS-DA R^2=0,654 Q^2=-0,106		OPLS-DA R^2=0,249 Q^2=-0,216	PLS-DA R^2=0,771 Q^2=0,602		

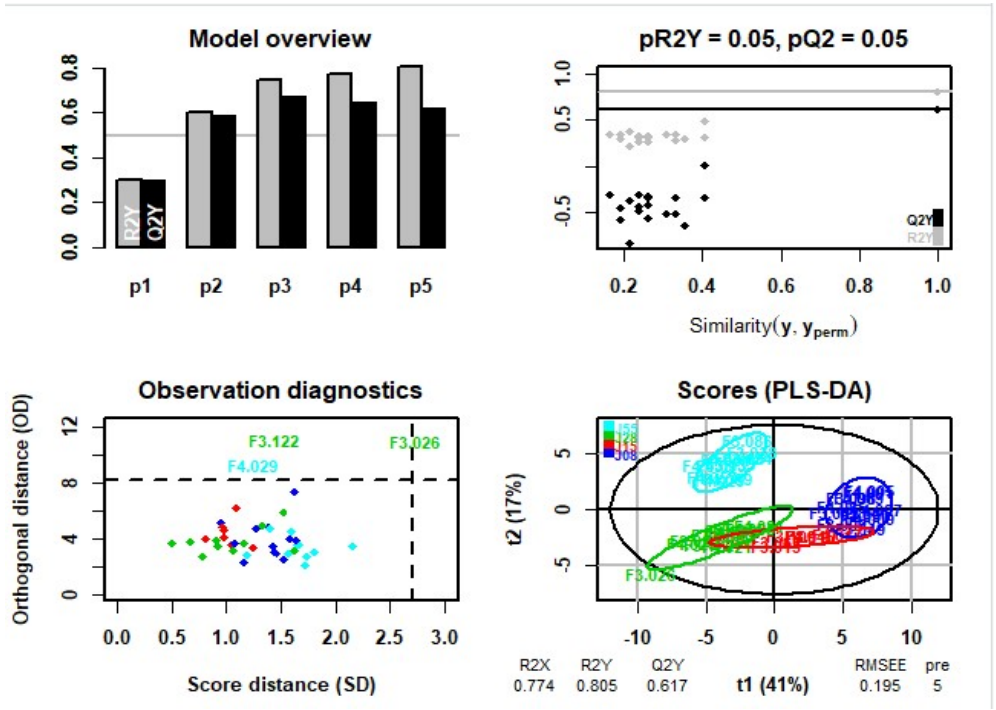
Οι παύλες στον πίνακα δηλώνουν ότι όταν τα δεδομένα εισήχθησαν στον αλγόριθμο δεν προέκυψε γράφημα-τιμή. Οπότε οι παύλες χαρακτηρίζουν τις κενές τιμές στο OPLS-DA.

1.10 Σύγκριση των αποτελεσμάτων μετά την ταυτοποίηση και ποσοτικοποίηση των μεταβολιτών στο φάσμα.

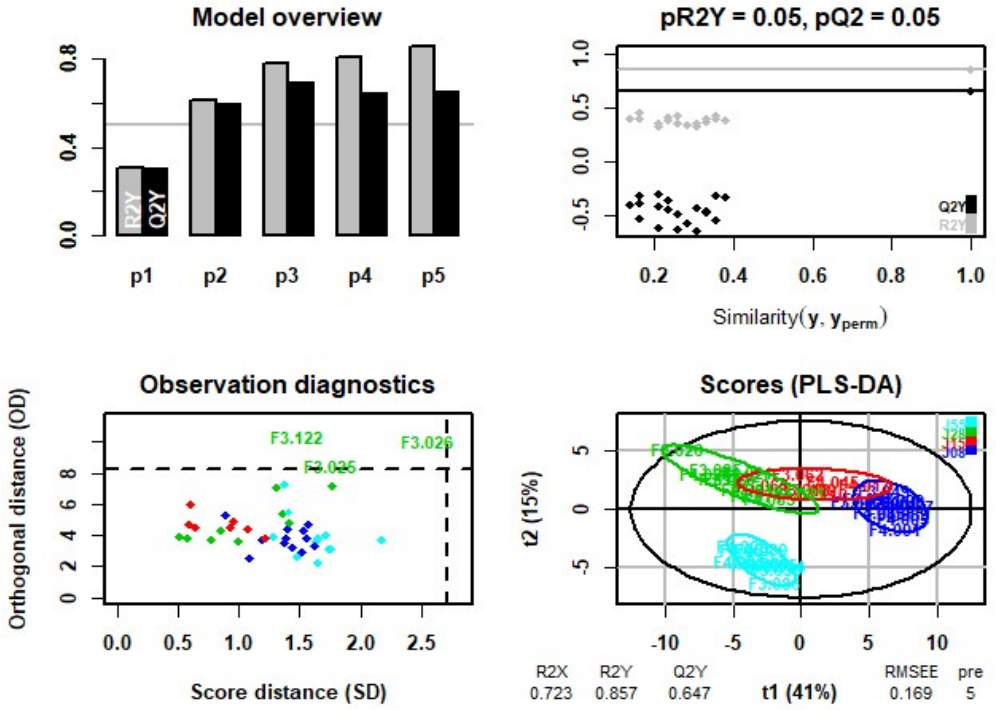
Πίνακας 3.3.1. Ο αριθμός των μεταβολιτών που ταυτοποιήθηκαν και οι έξι μεταβολίτες για κάθε διακριτική ικανότητα με τη μεγαλύτερη συνεισφορά στη διάκριση του μοντέλου.

Διακριτική Ικανότητα	Αριθμός ταυτοποιημένων μεταβολιτών	Μεταβολίτες που βρέθηκαν σε μεγαλύτερη συγκέντρωση
0,5	62	L-Asparagine, D-Glucose, D-Glucose-6-Phosphate, D-Fructose, L-Aspartate, AscorbicAcid
0,3	58	L-Asparagine, D-Glucose, Hypotaurine, L-Aspartate, Methanol, L-Proline
0,1	54	L-Asparagine, D-Glucose, Hypotaurine, L-Aspartate, Methanol, D-Glucose-6-Phosphate

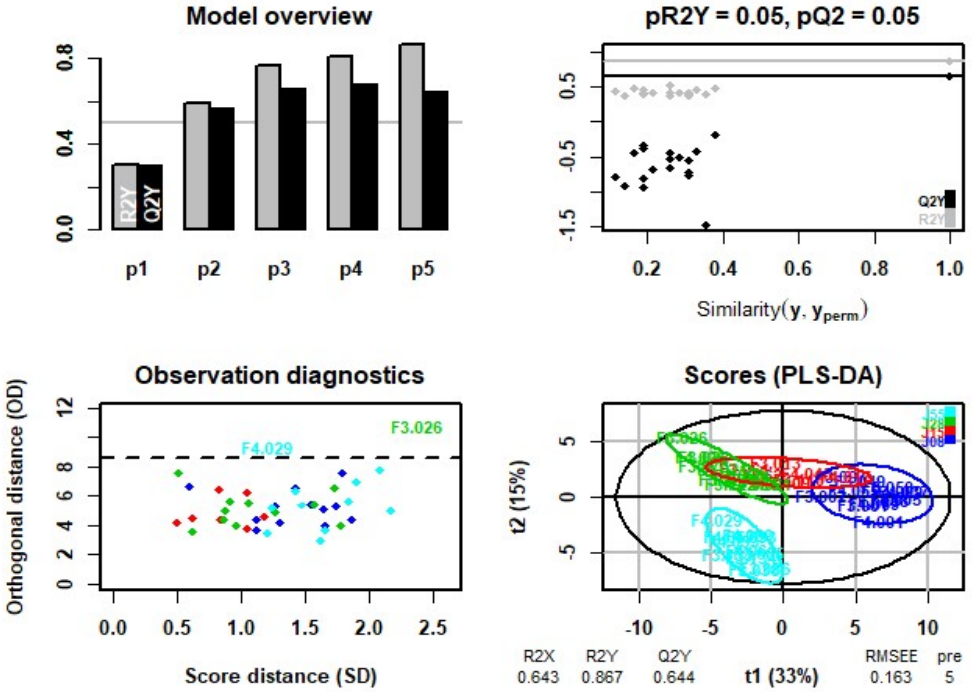
a.


Model overview
R2Y
Q2Y
p1
p2
p3
p4
p5
pR2Y = 0.05, pQ2 = 0.05
Q2Y
R2Y
Similarity(y, y_perm)
Observation diagnostics
F3.122
F3.026
F4.029
Orthogonal distance (OD)
Score distance (SD)
Scores (PLS-DA)
t2 (17%)
F3.026
R2X R2Y Q2Y RMSEE pre
0.774 0.805 0.617 t1 (41%) 0.195 5

b.


Model overview
R2Y
Q2Y
p1
p2
p3
p4
p5
pR2Y = 0.05, pQ2 = 0.05
Q2Y
R2Y
Similarity(y, y_perm)
Observation diagnostics
F3.122
F3.026
F3.025
Orthogonal distance (OD)
Score distance (SD)
Scores (PLS-DA)
t2 (15%)
R2X R2Y Q2Y RMSEE pre
0.723 0.857 0.647 t1 (41%) 0.169 5

c.



Εικόνα 3.3.1. PLS-DA μοντέλο για τον έλεγχο της ταυτοποίησης.
a. resolution 0,1 b. resolution 0,3 c. resolution 0,5

Πίνακας 3.3.2. Τα αποτελέσματα από τα μοντέλα για τον έλεγχο της ταυτοποίησης των μεταβολιτών .

	Στάδιο ανάπτυξης		
	Διακριτική Ικανότητα 0,1	Διακριτική Ικανότητα 0,3	Διακριτική Ικανότητα 0,5
PCA	R^2=0,578	R^2=0,566	R^2=0,564
PLS-DA	R^2=0,805 Q^2=0,617	R^2=0,857 Q^2=0,647	R^2=0,867 Q^2=0,644

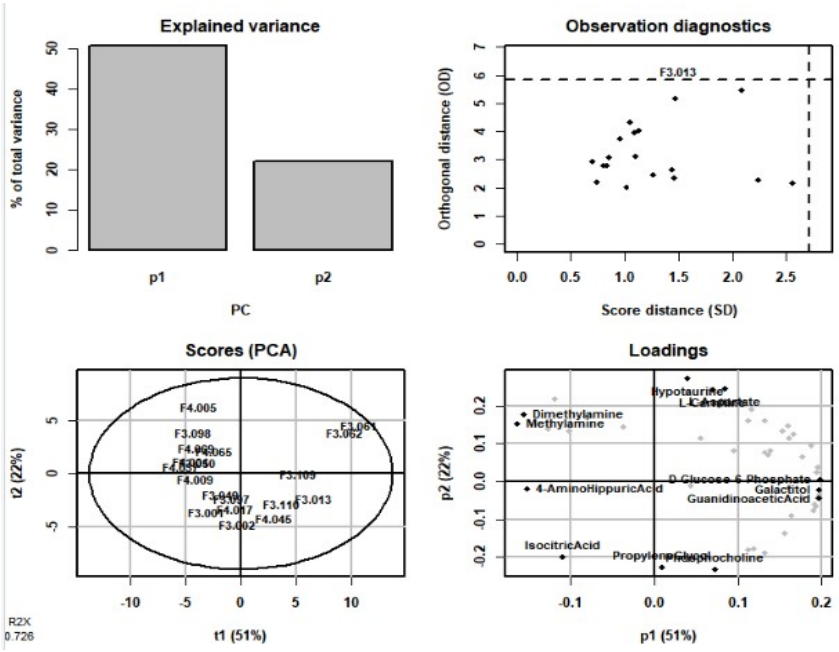
1.11 Έλεγχος των σταδίων ανάπτυξης της ντομάτας (μέσω 3 διαφορετικών μεθόδων σύγκρισης)

Όπως αναφέρθηκε υπάρχουν τρεις πιθανοί τρόποι να συγκρίνουμε τα στάδια ανάπτυξης της ντομάτας από την ημέρα άνθησης (DPA). Οι ομάδες οι οποίες ορίστηκαν από τους ερευνητές των δεδομένων είναι J08-J15-J28-J55 οι οποίες περιγράφουν τις 8 DPA, 15 DPA, 28 DPA και 55 DPA αντίστοιχα. Παρακάτω δίνονται τα αποτελέσματα δύο συγκρίσεων με πίνακες και γραφήματα. Για το βέλτιστο αποτέλεσμα σε κάθε σύγκριση μεταξύ των σταδίων, προσαρμόστηκε ο αριθμός των κύριων συνιστωσών στην PCA ανάλυση.

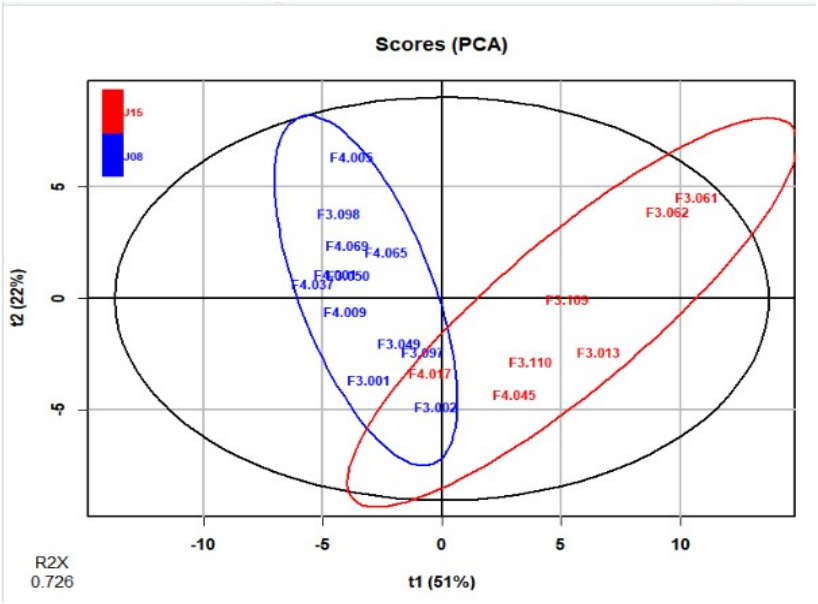
1.11.1 Πρώτη σύγκριση των σταδίων [Σύγκριση (1)]

Ο πρώτος τρόπος [Σύγκριση (1)] είναι σε όλα τα δείγματα των σταδίων να γίνει ταυτοποίηση και ποσοτικοποίηση των μεταβολιτών και ύστερα σύγκριση των σταδίων ανά δύο. Αυτό πραγματοποιήθηκε προγραμματιστικά μέσω της γλώσσας προγραμματισμού R. Δημιουργήθηκε αλγόριθμος στον οποίο παράχθηκε ένας πίνακας που περιλάμβανε όλα τα φάσματα (πίνακας με αριθμητικές τιμές). Με αυτό τον τρόπο όλες οι τιμές των φασμάτων των διαφορετικών σταδίων ωρίμανσης συμβάλλουν και επηρεάζουν την τελική δημιουργία του μοντέλου. Με γνώμονα τη μη στοχευμένη μεταβολομική επιστήμη (untargeted metabolomics) το μοντέλο αυτό που θα παραχθεί θα πρέπει να αναγνωρίζει ένα νέο δείγμα/φάσμα που θα τοποθετείται στο μοντέλο και θα πρέπει αυτόματα να το ομαδοποιεί στο σωστό στάδιο ωρίμανσης.

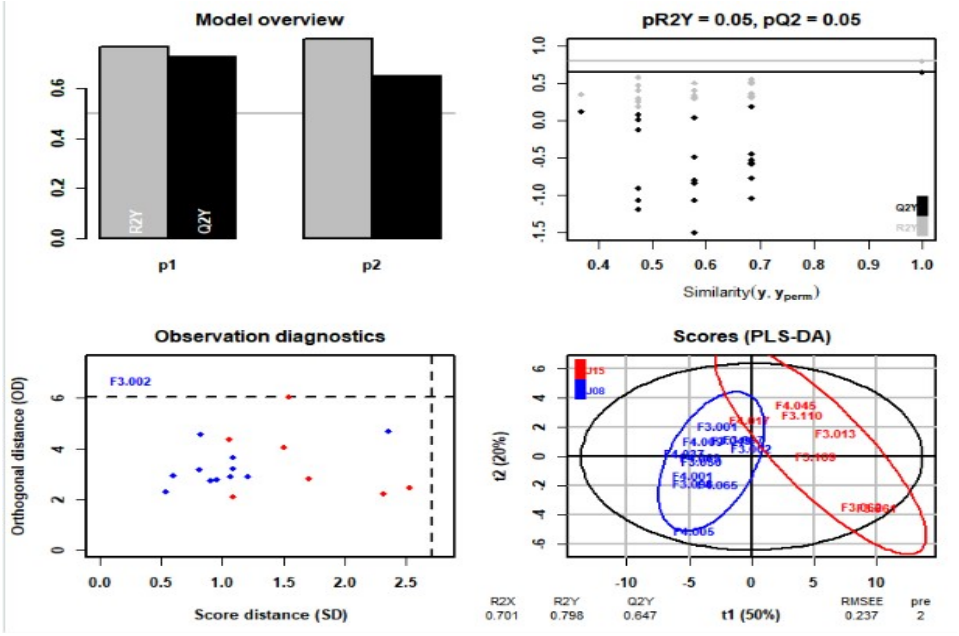
a.



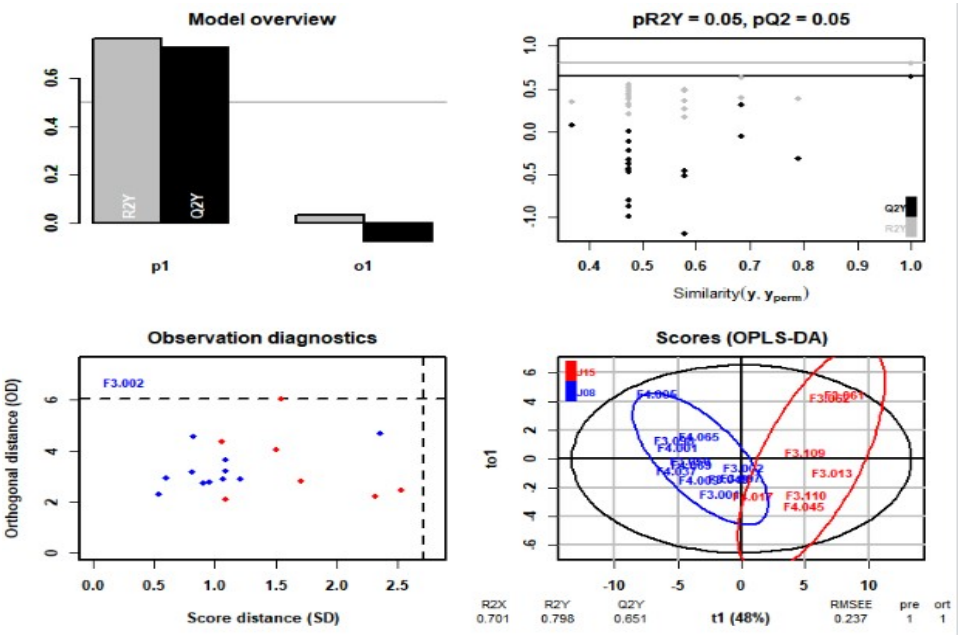
b.



c.



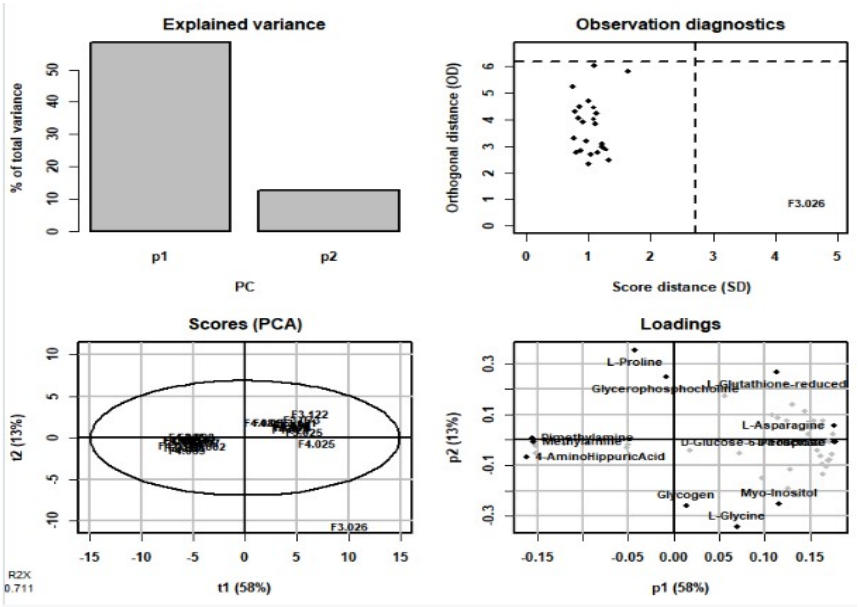
d.



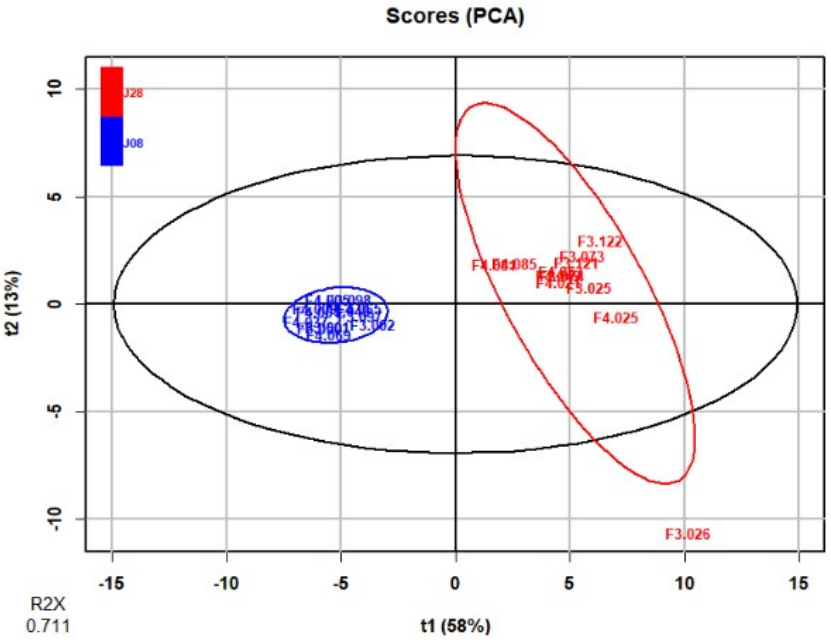
Εικόνα 3.4.1.1. Σύγκριση(1) σταδίων J08-J15
a. περιληπτικό γράφημα PCA , b. γράφημα PCA , c. PLS-DA μοντέλο , d. OPLS-DA μοντέλο

Τα στάδια J08 και J15 όπως αναφέρθηκε και σε προηγούμενο κεφάλαιο κατανέμονται στη φάση της κυτταρικής διαίρεσης. Για αυτό το λόγο αναμένεται ο διαχωρισμός των δύο αυτών σταδίων να είναι πιο δύσκολος σε σύγκριση με τους υπόλοιπους που ανήκουν σε διαφορετικές φάσεις. Από τις εικόνες παραπάνω επιβεβαιώνεται η υπόθεση διότι παρατηρείται ένα κοινό υποσύνολο δειγμάτων μεταξύ των δύο σταδίων.

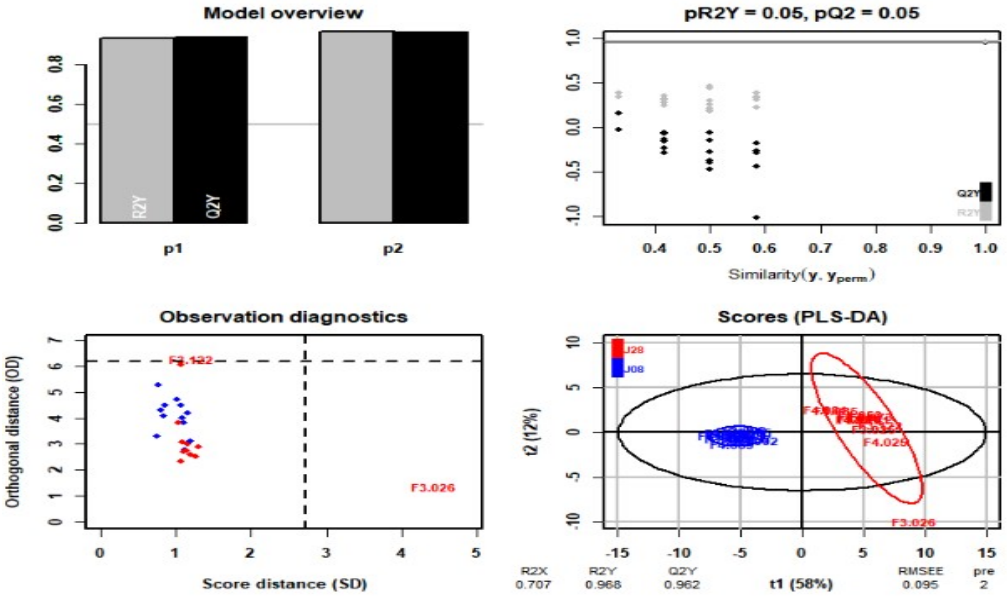
a.


Explained variance
% of total variance
p1
p2
PC
Observation diagnostics
Orthogonal distance (OD)
Score distance (SD)
F3.026
Scores (PCA)
t2 (13%)
t1 (58%)
F3.122
F4.025
F3.026
R2X
0.711
Loadings
p2 (13%)
p1 (58%)
L-Proline
Glycerophosphocholine
L-Glutathione-reduced
L-Asparagine
Phenethylamine
Methylamine
4-AminoHippuricAcid
Glycogen
Myo-Inositol
L-Glycine

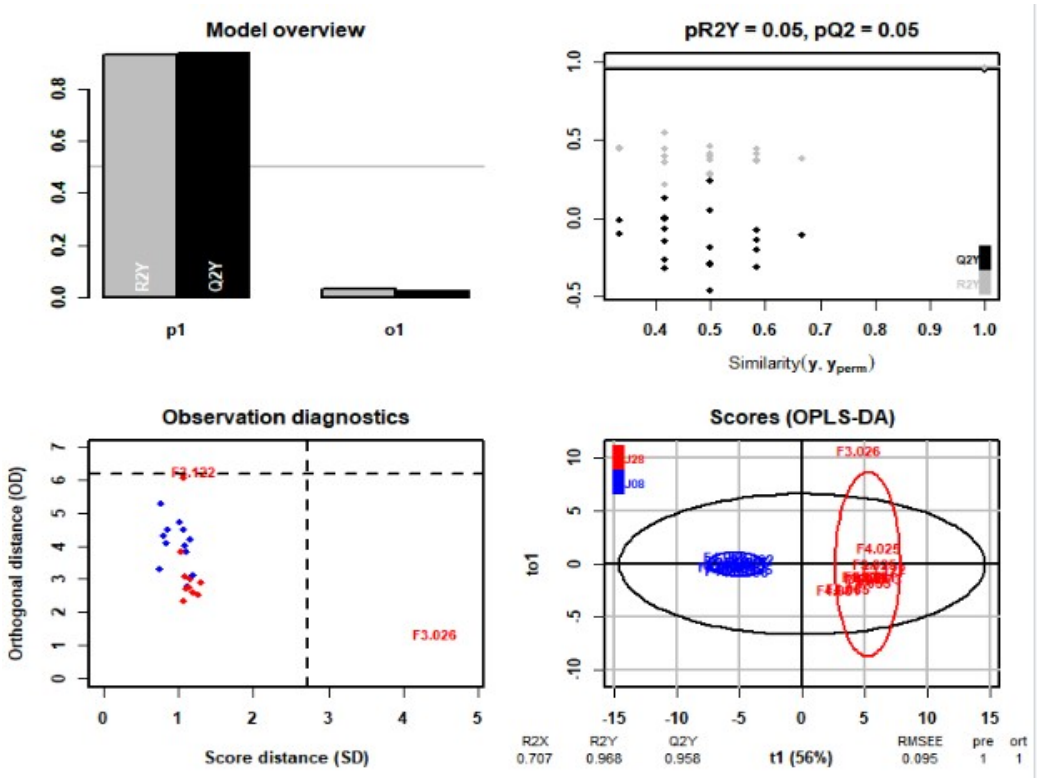
b.


Scores (PCA)
t2 (13%)
t1 (58%)
F3.122
F3.073
F3.025
F4.025
F3.026
R2X
0.711

c.



d.

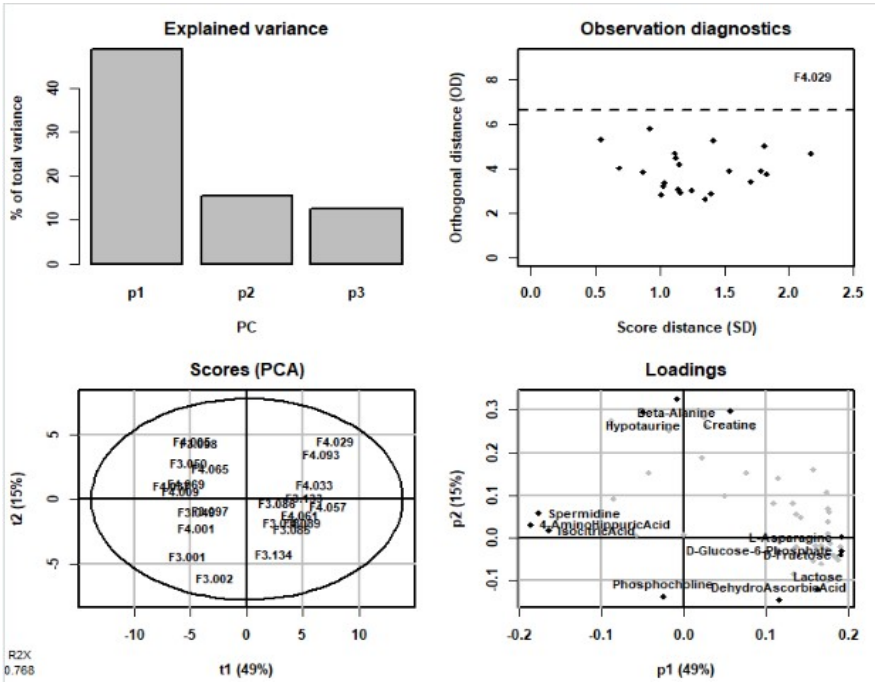


Εικόνα 3.4.1.2. Σύγκριση(1) σταδίων J08-J28
a. περιληπτικό γράφημα PCA , b. γράφημα PCA , c. PLS-DA μοντέλο , d. OPLS-DA μοντέλο

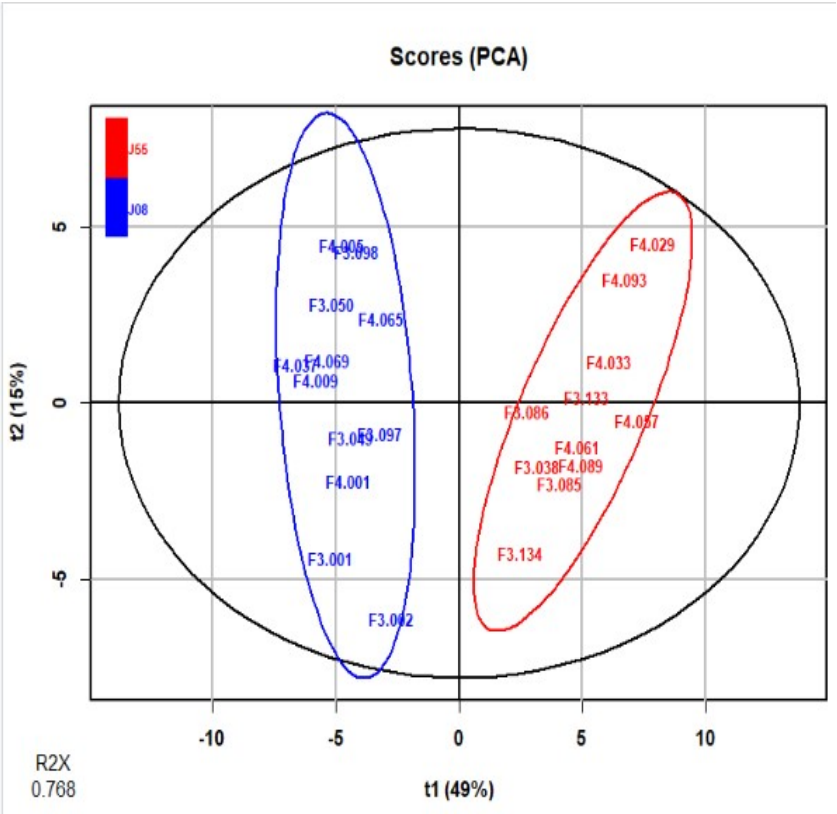
Στη σύγκριση αυτή μεταξύ των σταδίων J08 και J28 ο διαχωρισμός των δύο ομάδων είναι ξεκάθαρος. Επίσης η προβλεψιμότητα του μοντέλου είναι υψηλή. Γεγονός που φανερώνει

ότι το μοντέλο μπορεί να αναγνωρίσει με υψηλό ποσοστό επιτυχίας ακόμα και δείγματα που δεν έχουν συμβάλλει στη δημιουργία του μοντέλου.

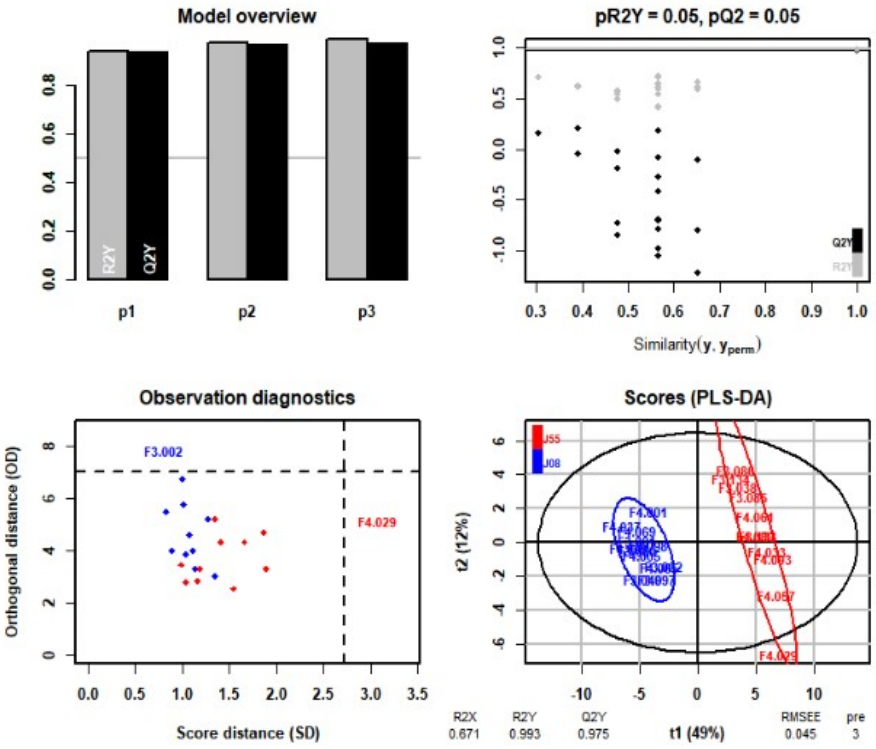
a.



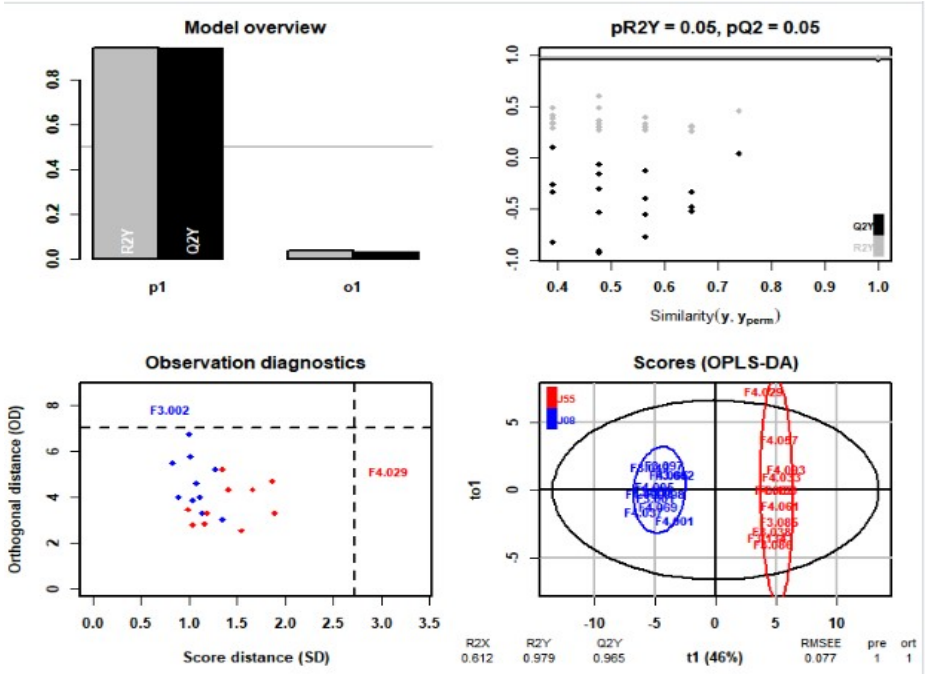
b.



c.



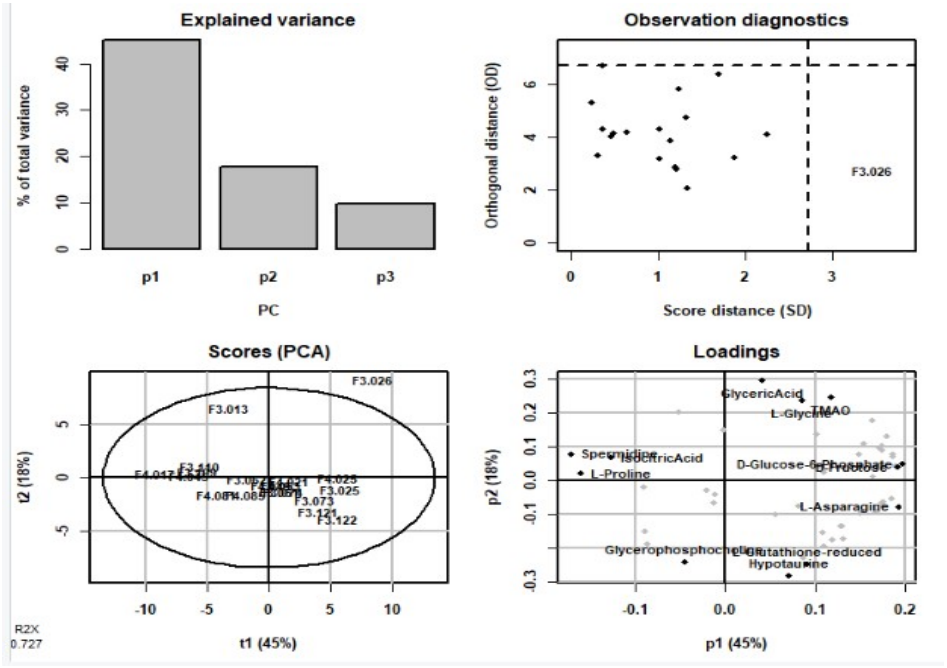
d.



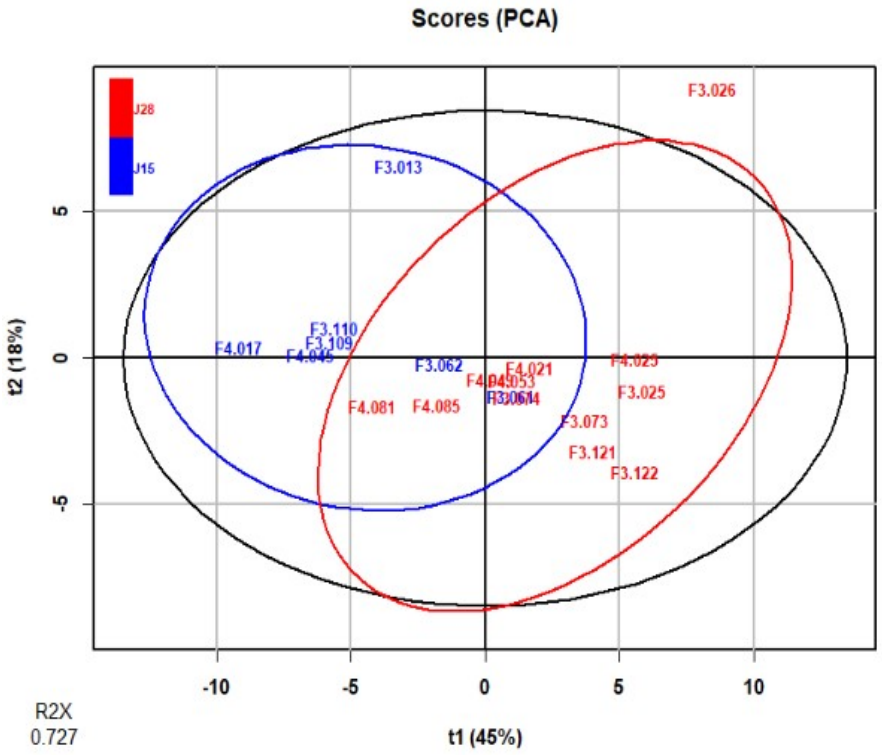
Εικόνα 3.4.1.3. Σύγκριση(1) σταδίων J08-J55
a. περιληπτικό γράφημα PCA , b. γράφημα PCA , c. PLS-DA μοντέλο , d. OPLS-DA μοντέλο

Επίσης, άλλος ένας καλός διαχωρισμός μεταξύ των σταδίων J08 και J55, το οποίο και αναμένεται καθώς είναι η σύγκριση του πρώτου σταδίου ωρίμανσης της ντομάτας με το τελικό στάδιο.

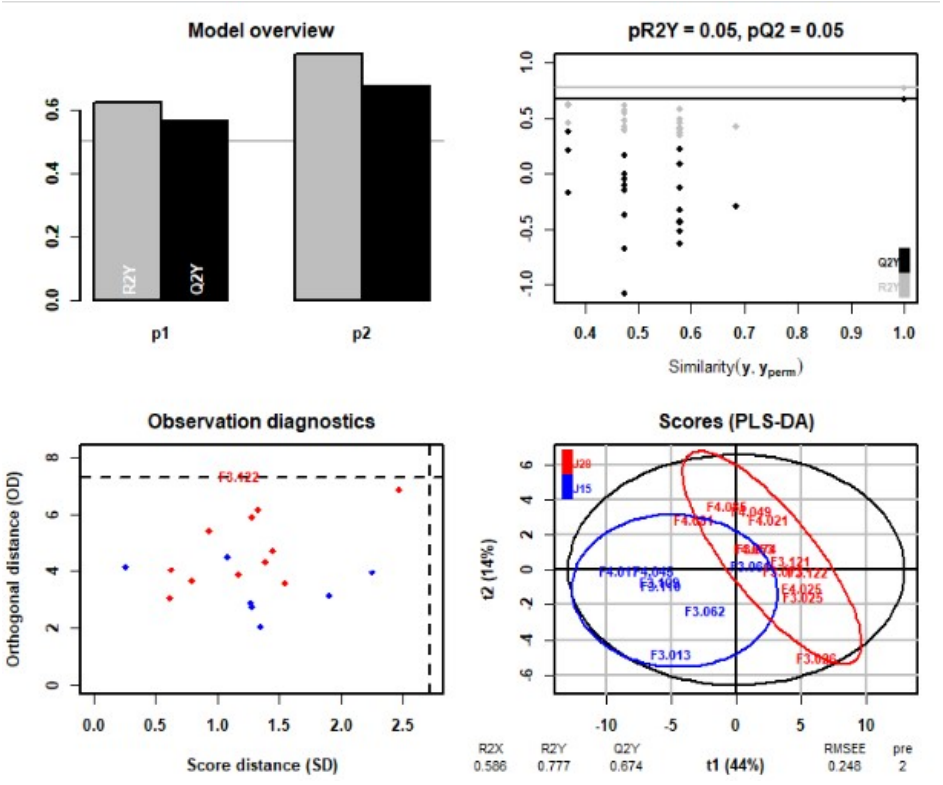
a.


Explained variance
% of total variance
p1
p2
p3
PC
Observation diagnostics
Orthogonal distance (OD)
Score distance (SD)
F3.026
Scores (PCA)
t2 (18%)
t1 (45%)
R2X
0.727
Loadings
p2 (18%)
p1 (45%)
GlycericAcid
L-Glycine
TMAO
L-Proline
L-Asparagine
Hypotaurine

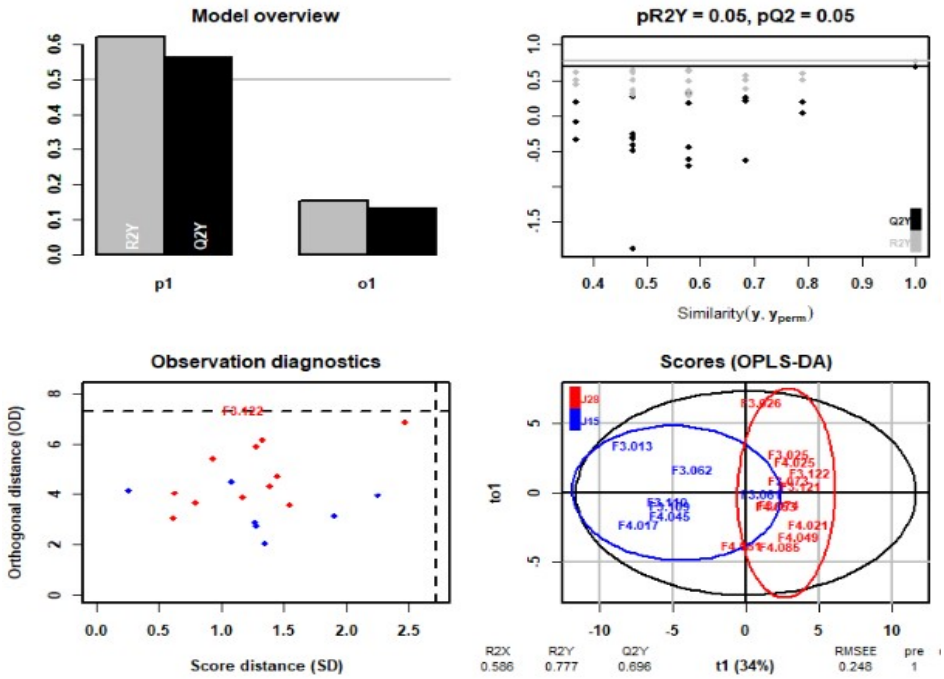
b.


Scores (PCA)
J28
J15
F3.026
F3.013
F3.110
F3.109
F4.017
F4.045
F3.062
F4.021
F4.023
F3.025
F4.081
F4.085
F3.073
F3.121
F3.122
t2 (18%)
t1 (45%)
R2X
0.727

c.



d.

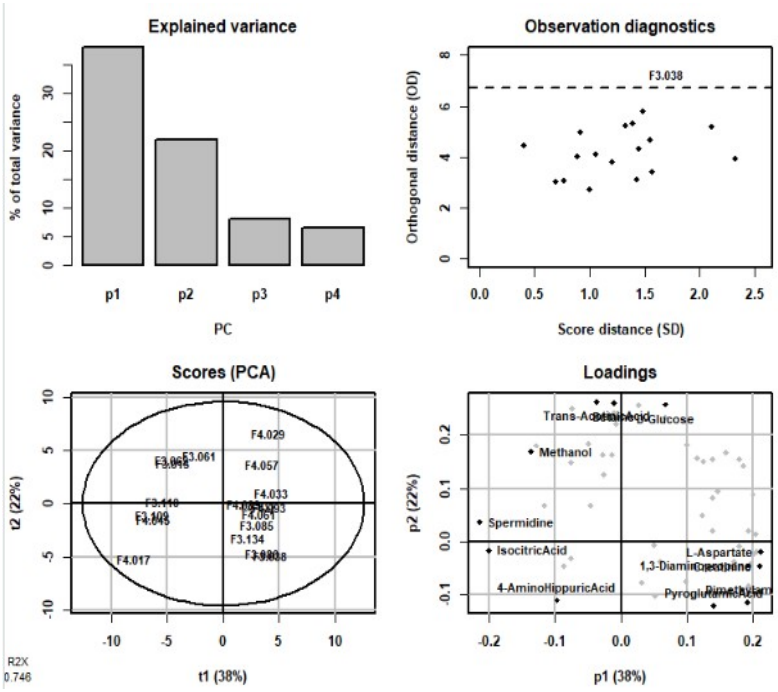


Εικόνα 3.4.1.4. Σύγκριση(1) σταδίων J15-J28
a. περιληπτικό γράφημα PCA , b. γράφημα PCA , c. PLS-DA μοντέλο , d. OPLS-DA μοντέλο

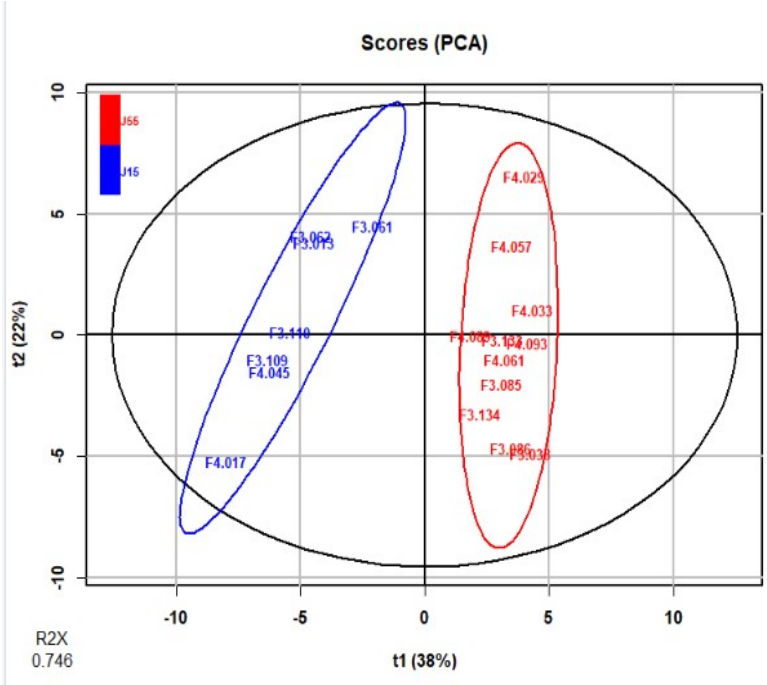
Παρατηρούμε ότι για τα στάδια J15 και J28 υπάρχει επικάλυψη των ομάδων που σχηματίζονται. Εφόσον το κάθε στάδιο ανήκει θεωρητικά σε διαφορετικές φάσεις

ανάπτυξης (J15- κυτταρική διαίρεση, J28- πολλαπλασιασμός κυττάρων) αναμενόταν δύο ξεκάθαρες ομάδες χωρίς επικαλύψεις.

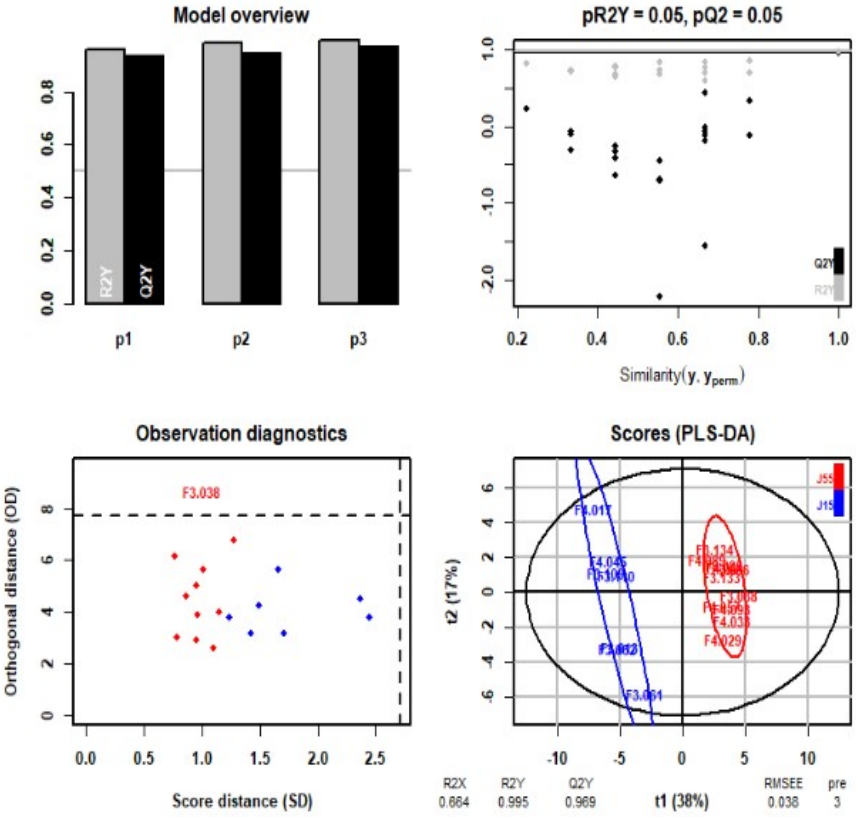
a.



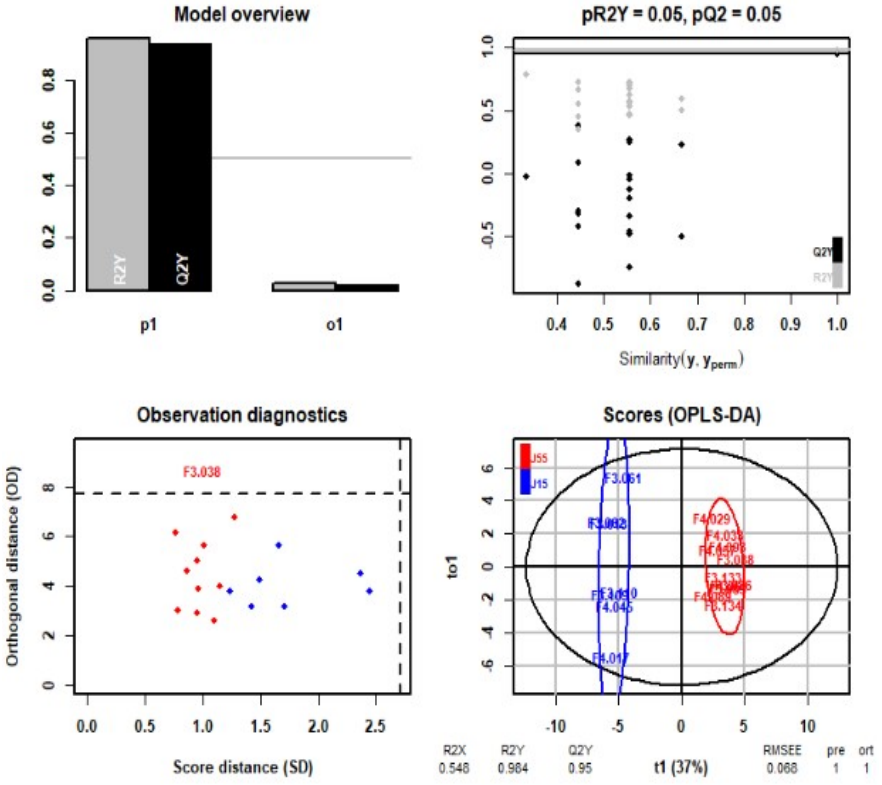
b.



c.



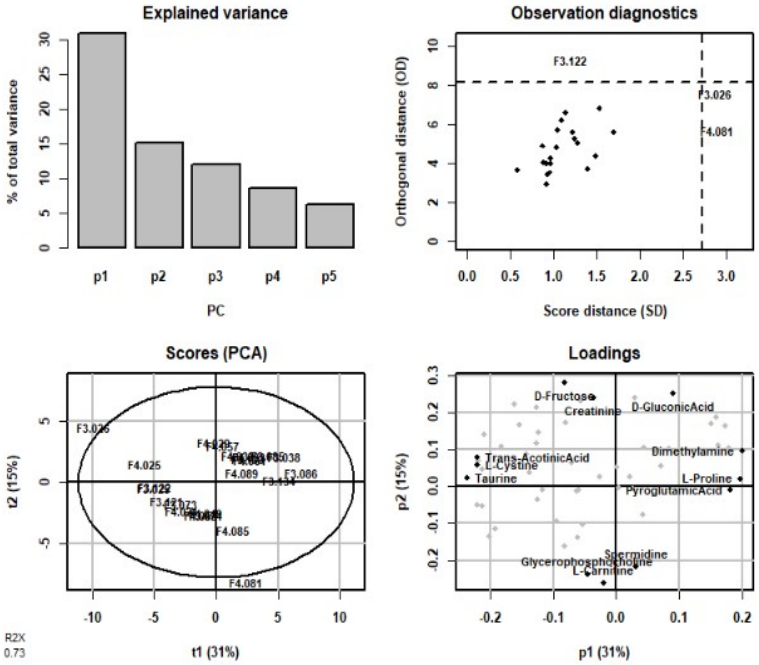
d.



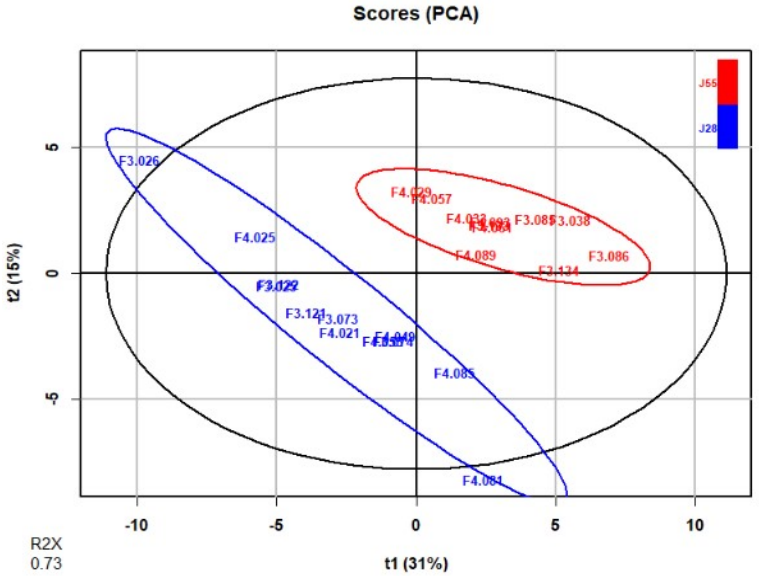
Εικόνα 3.4.1.5. Σύγκριση(1) σταδίων J15-J55 a. περιληπτικό γράφημα PCA , b. γράφημα PCA, c. PLS-DA μοντέλο , d. OPLS-DA μοντέλο

Στη σύγκριση μεταξύ των σταδίων J15 και J55 παρατηρείται επίσης πολύ καλός διαχωρισμός των ομάδων, το οποίο αναμενόταν καθώς ανήκουν στην αρχική και τελική φάση ανάπτυξης της ντομάτας αντίστοιχα.

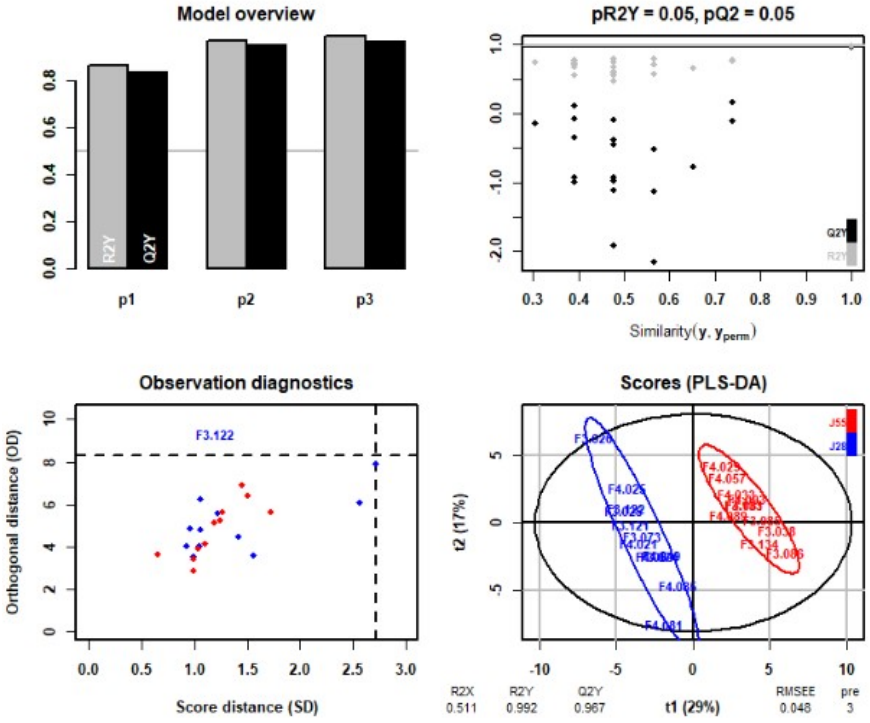
a.



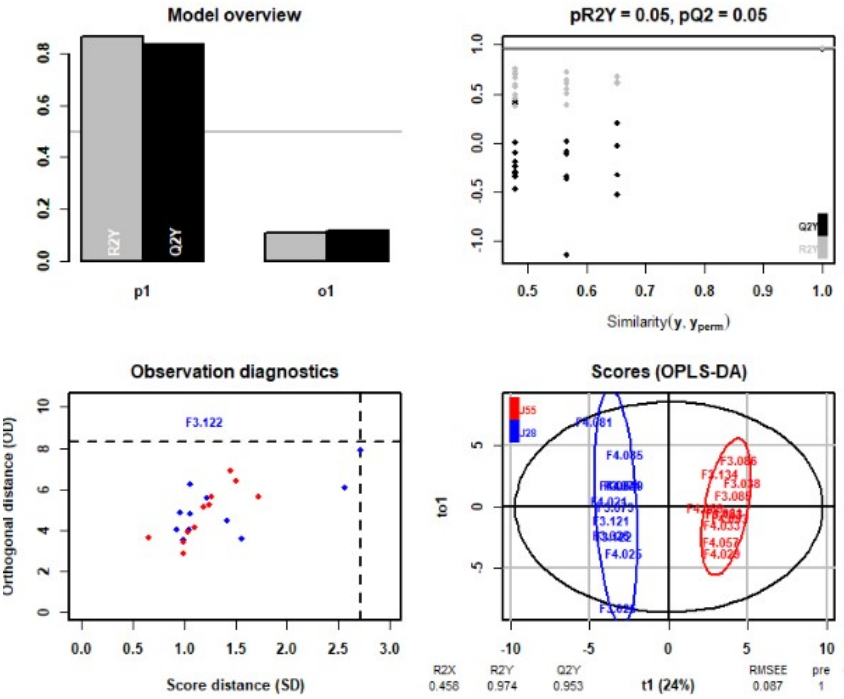
b.



c.



d.



Εικόνα 3.4.1.6. Σύγκριση(1) σταδίων J28-J55 a. περιληπτικό γράφημα PCA , b. γράφημα PCA, c. PLS-DA μοντέλο , d. OPLS-DA μοντέλο

Στην σύγκριση μεταξύ των ομάδων J28 και J55 παρατηρούμε επίσης ξεκάθαρο διαχωρισμό των δύο ομάδων. Επίσης παρατηρούμε ότι χρειάστηκαν 5 κύριες συνιστώσες ώστε να

επιτευχθεί αυτός ο διαχωρισμός. Δηλαδή χρειάστηκε παραπάνω δεδομένα-πληροφορία προκειμένου να δώσει την καλύτερη ομαδοποίηση. Ωστόσο, το πρόβλημα το οποίο δημιουργείται είναι ότι μπορεί να οδηγήσει σε φαινόμενα υπερπροσαρμογής (overfitting).

Πίνακας 3.4.1.1 Τα αποτελέσματα από τα μοντέλα για τη Σύγκριση(1) .

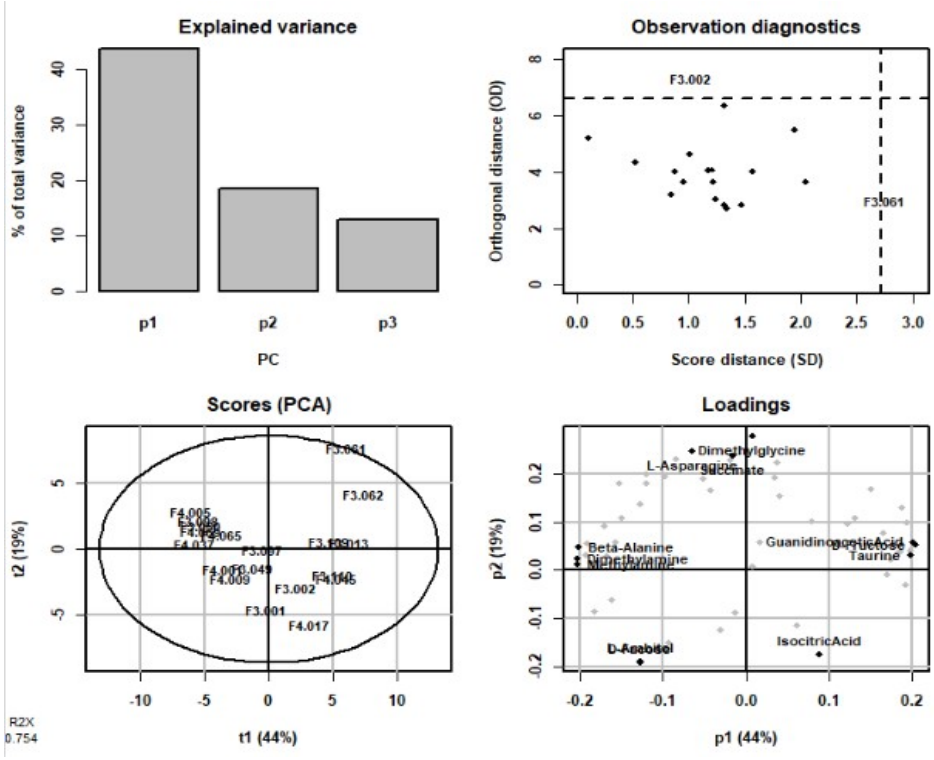
	Διακριτική Ικανότητα 0,3		
	PCA	**PLS-DA**	**OPLS-DA**
J08– J15	R^2=0,726	R^2=0,79 Q^2=0,647	R^2=0,798 Q^2=0,65
J08- J28	R^2=0,711	R^2=0,968 Q^2=0,962	R^2=0,968 Q^2=0,958
J08– J55	R^2=0,768	R^2=0,993 Q^2=0,975	R^2=0,979 Q^2=0,965
J15– J28	R^2=0,727	R^2=0,777 Q^2=0,674	R^2=0,777 Q^2=0,696
J15– J55	R^2=0,746	R^2=0,995 Q^2=0,969	R^2=0,984 Q^2=0,95
J28 –J55	R^2=0,73	R^2=0,992 Q^2=0,967	R^2=0,974 Q^2=0,953

1.11.2 Δεύτερη σύγκριση των σταδίων [Σύγκριση (2)]

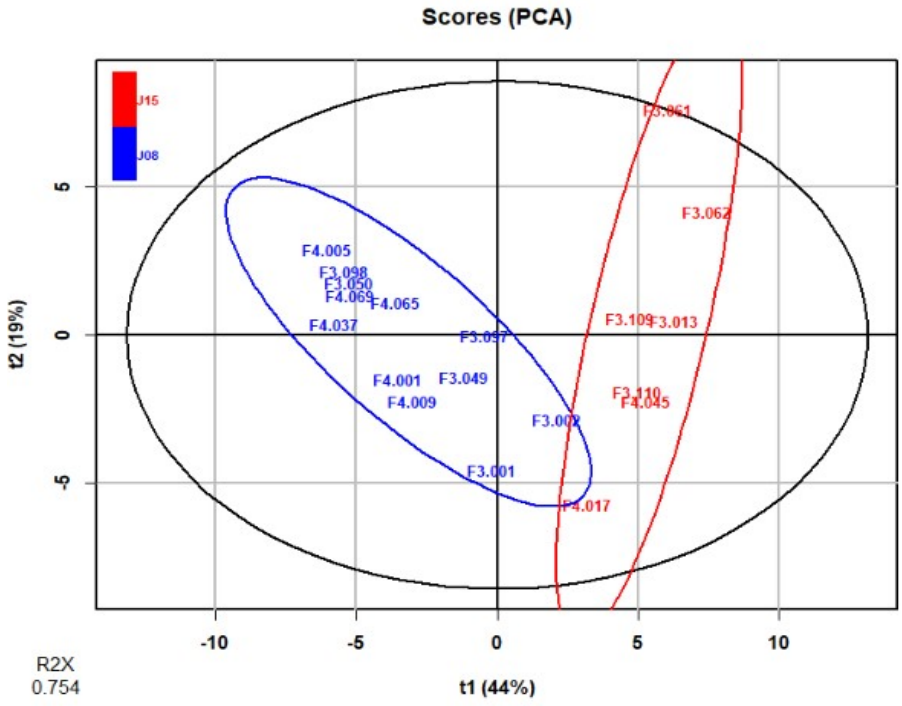
Ο δεύτερος τρόπος [Σύγκριση(2)] είναι να γίνει ταυτοποίηση και ποσοτικοποίηση των μεταβολιτών σε κάθε στάδιο ξεχωριστά και ύστερα σύγκριση ανά δύο. Και σε αυτή τη σύγκριση ορίστηκε ο αριθμός των κύριων συνιστωσών ώστε να επιτευχθεί η καλύτερη ομαδοποίηση μεταξύ των σταδίων.

Στο δεύτερο μέρος της σύγκρισης, δημιουργήθηκε νέος αλγόριθμος ώστε πλέον η κανονικοποίηση των δεδομένων να γίνει σε κάθε στάδιο ξεχωριστά. Δηλαδή, το στάδιο J08 θα κανονικοποιηθεί χωρίς οι τιμές του να επηρεαστούν από τις τιμές των άλλων σταδίων. Αυτή είναι και η βασική διαφορά με τον πρώτο τρόπο σύγκρισης που αναλύθηκε παραπάνω.

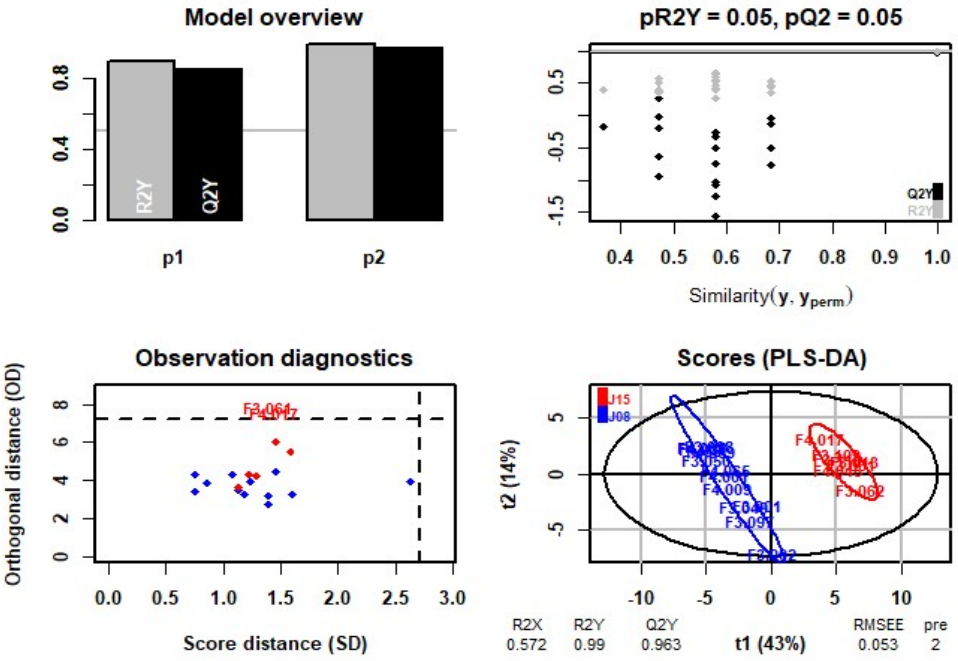
a.


Explained variance
% of total variance
p1
p2
p3
PC
Observation diagnostics
Orthogonal distance (OD)
Score distance (SD)
F3.002
F3.061
Scores (PCA)
t2 (19%)
t1 (44%)
R2X
0.754
Loadings
p2 (19%)
p1 (44%)
Dimethylglycine
L-Asparagine
Succinate
Beta-Alanine
Taurine
D-Asaibisol
IsocitricAcid

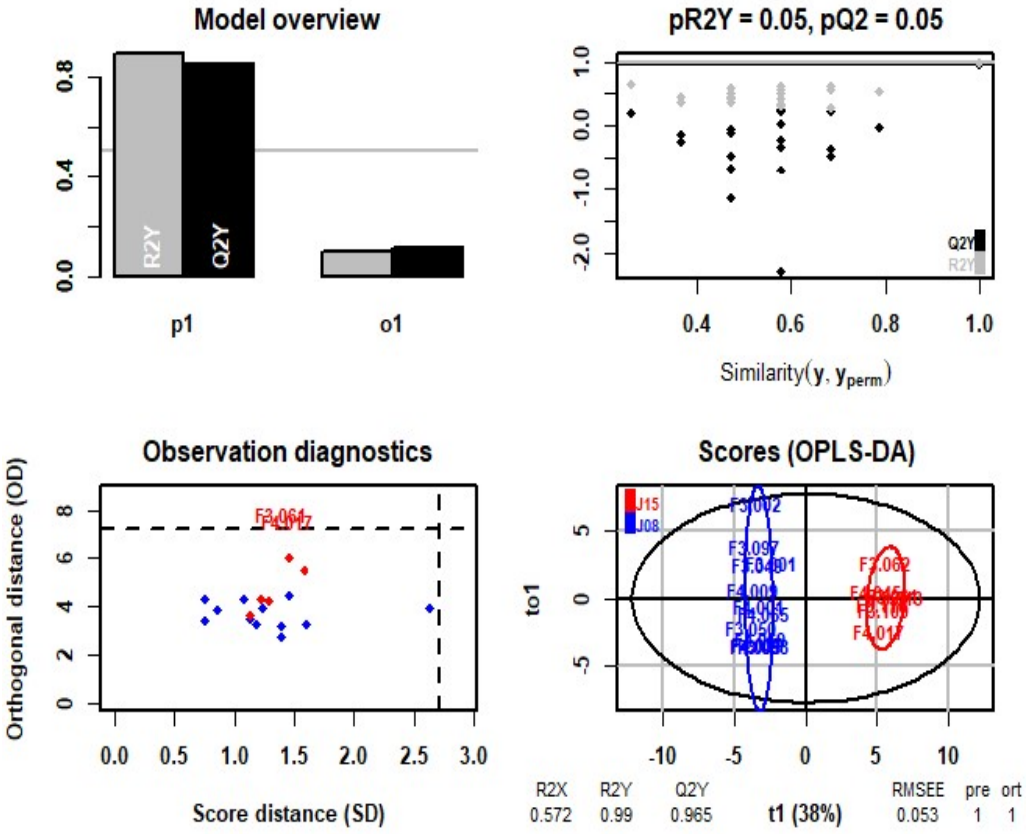
b.


Scores (PCA)
J15
J08
F3.061
F3.062
F4.005
F3.098
F3.050
F4.069
F4.065
F4.037
F3.109
F3.013
F3.057
F4.001
F3.049
F4.009
F3.110
F4.045
F3.007
F3.001
F4.017
t2 (19%)
t1 (44%)
R2X
0.754

c.



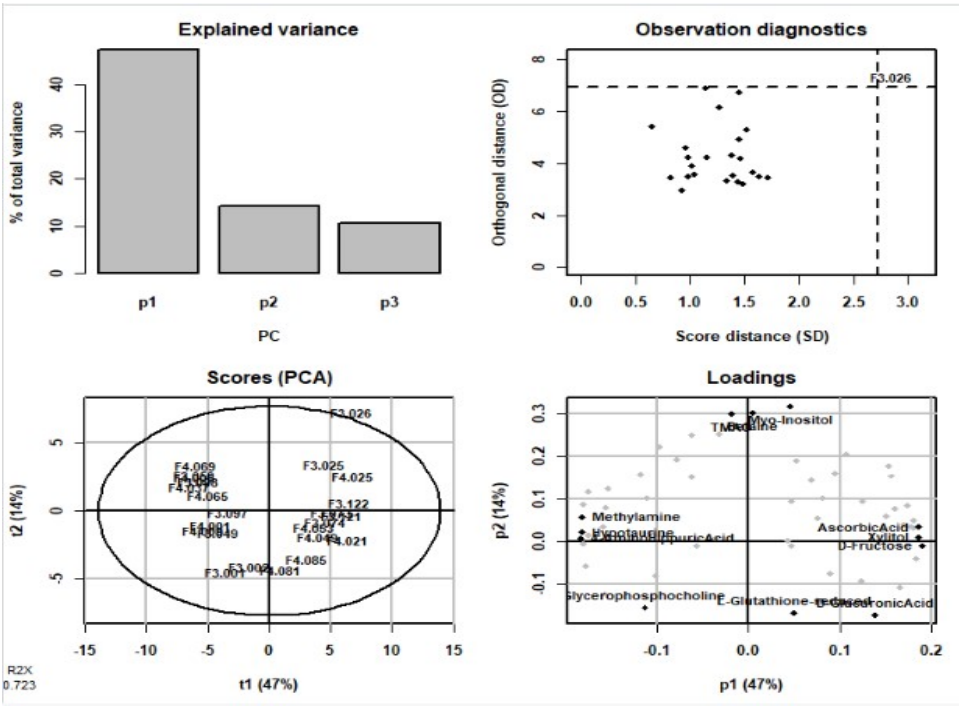
d.



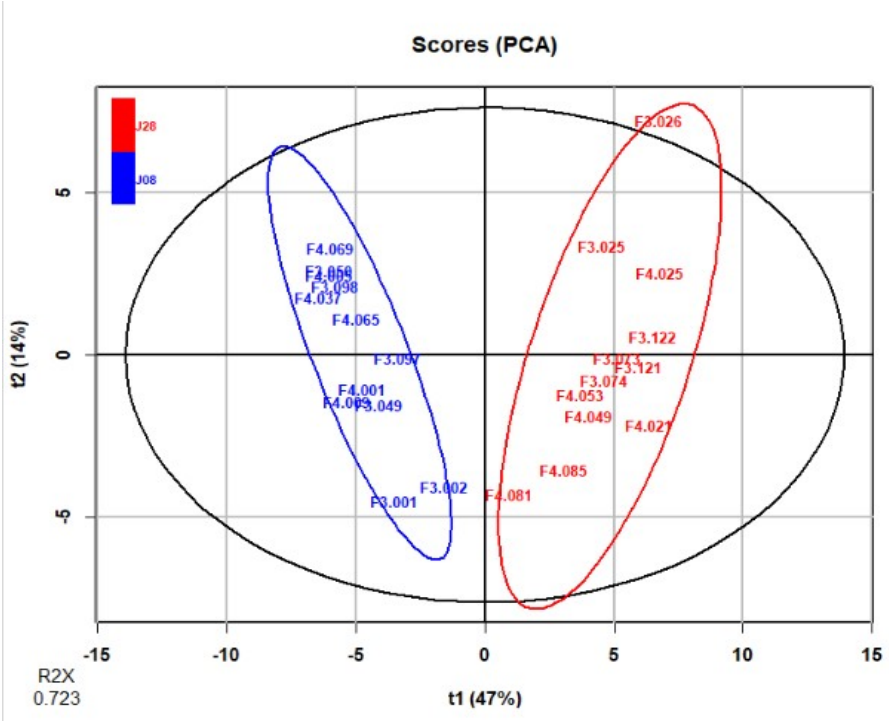
Εικόνα 3.4.2.1. Σύγκριση(2) σταδίων J08-J15 a. περιληπτικό γράφημα PCA , b. γράφημα PCA, c. PLS-DA μοντέλο , d. OPLS-DA μοντέλο

Τα στάδια J08-J15 είναι τα δυσκολότερα στον διαχωρισμό λόγω του ότι ανήκουν στην ίδια φάση ανάπτυξης όπως αναφέρθηκε και προηγουμένως. Παρατηρείται, καλύτερος διαχωρισμός στο b.γράφημα PCA καθώς και ξεκάθαρος διαχωρισμός των ομάδων στο PLS-DA και OPLS-DA μοντέλο.

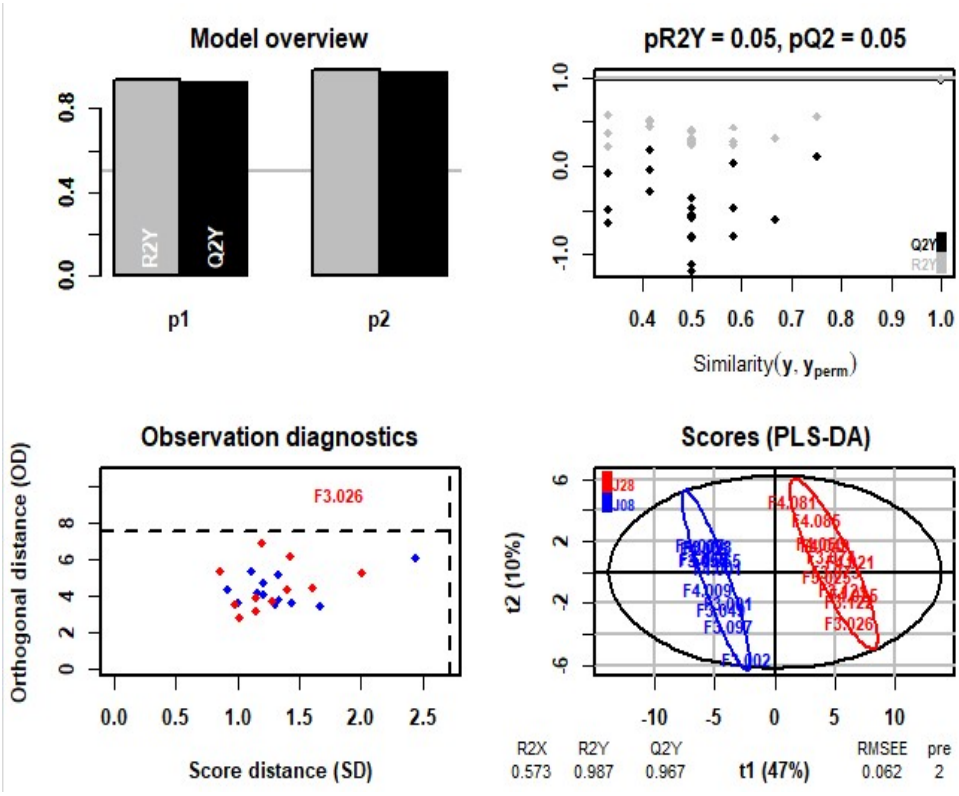
a.



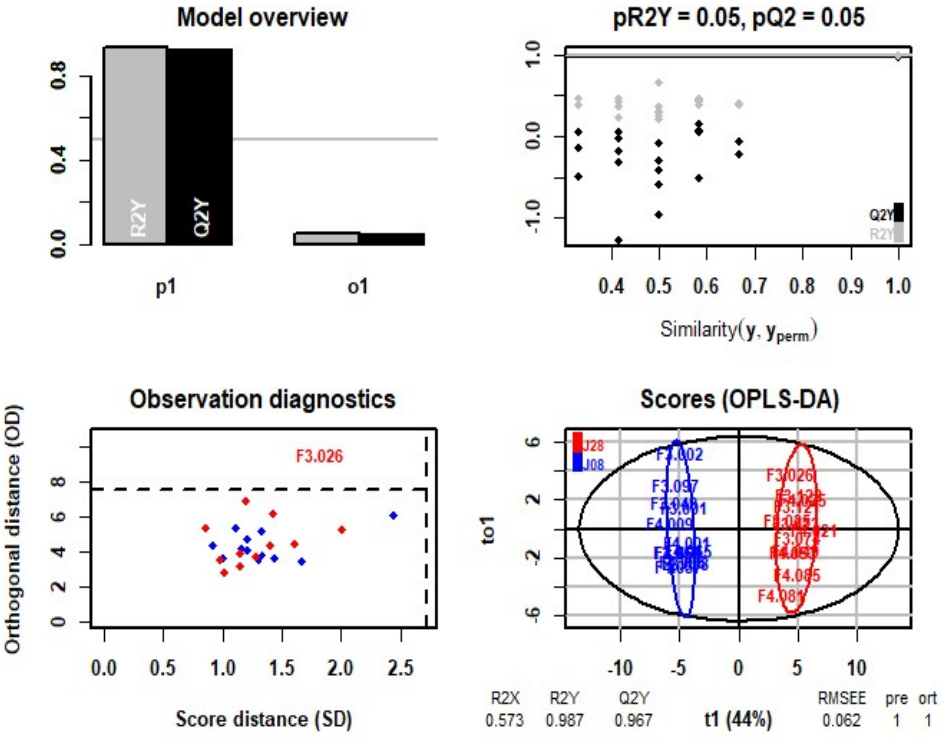
b.



c.

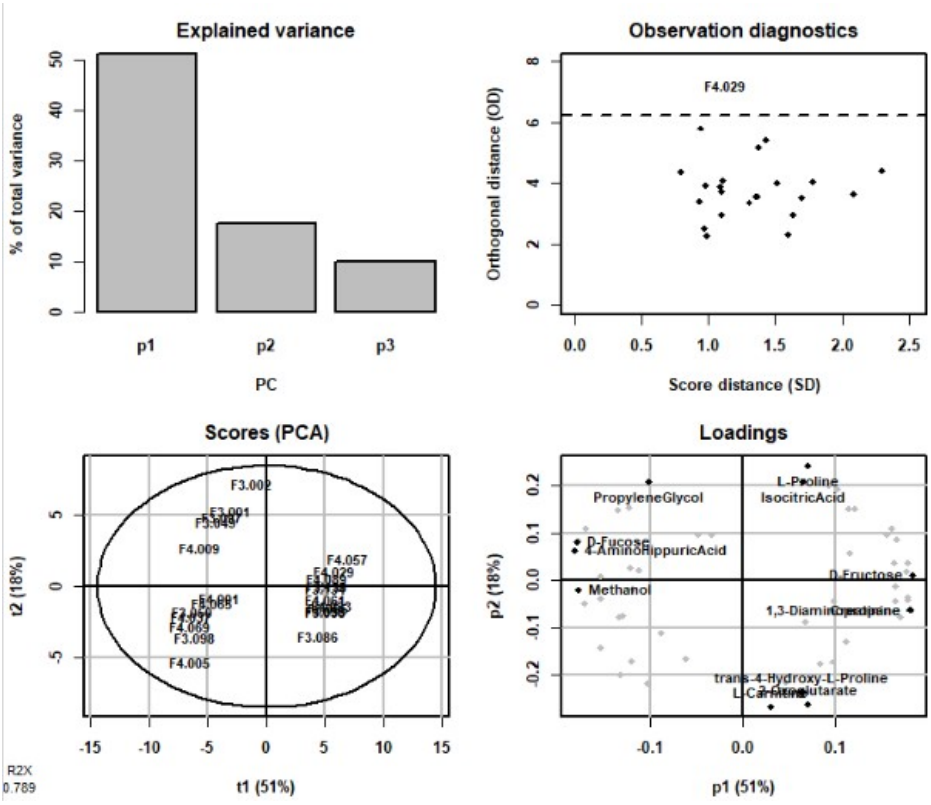


d.

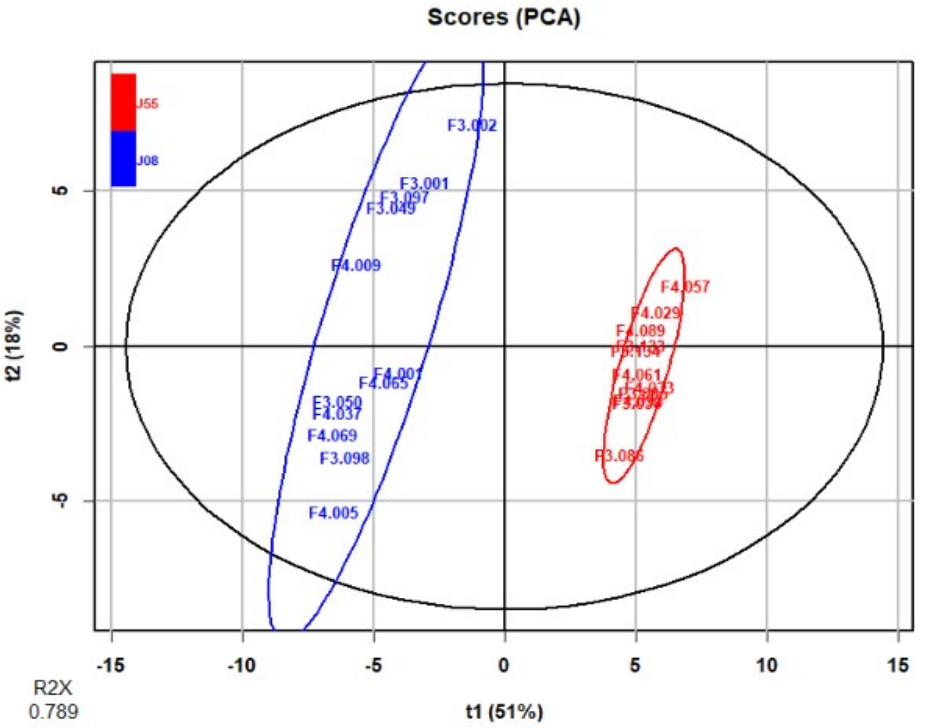


Εικόνα 3.4.2.2. Σύγκριση(2) σταδίων J08-J28 a. περιληπτικό γράφημα PCA , b. γράφημα PCA, c. PLS-DA μοντέλο , d. OPLS-DA μοντέλο

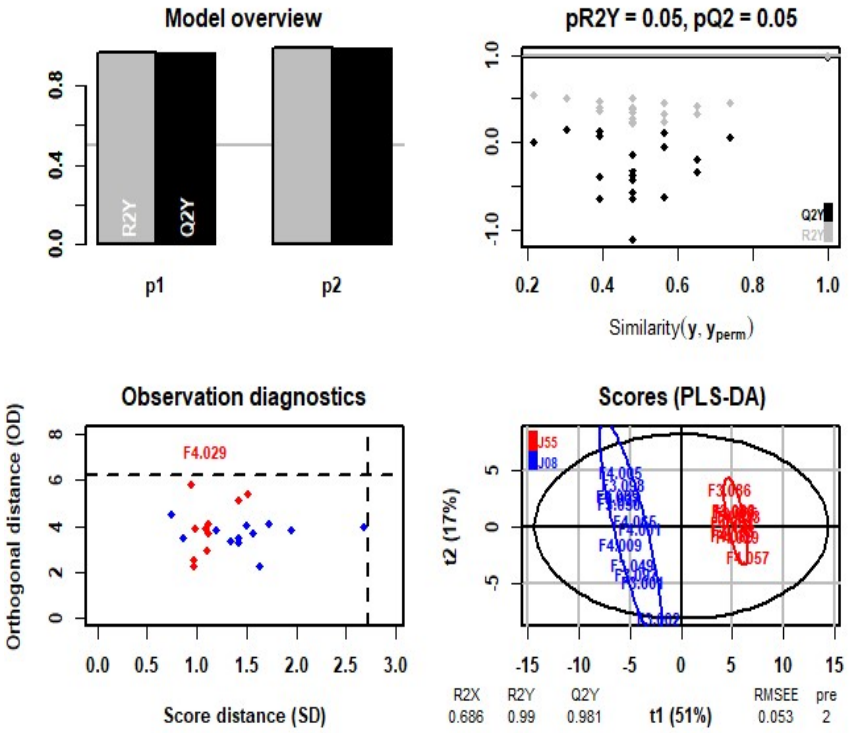
a.


Explained variance
% of total variance
p1
p2
p3
PC
Observation diagnostics
Orthogonal distance (OD)
F4.029
Score distance (SD)
Scores (PCA)
t2 (18%)
t1 (51%)
R2X
0.789
Loadings
p2 (18%)
p1 (51%)
PropyleneGlycol
L-Proline
IsocitricAcid
D-Fucose
4-AminoHippuricAcid
D-Fructose
Methanol
trans-4-Hydroxy-L-Proline

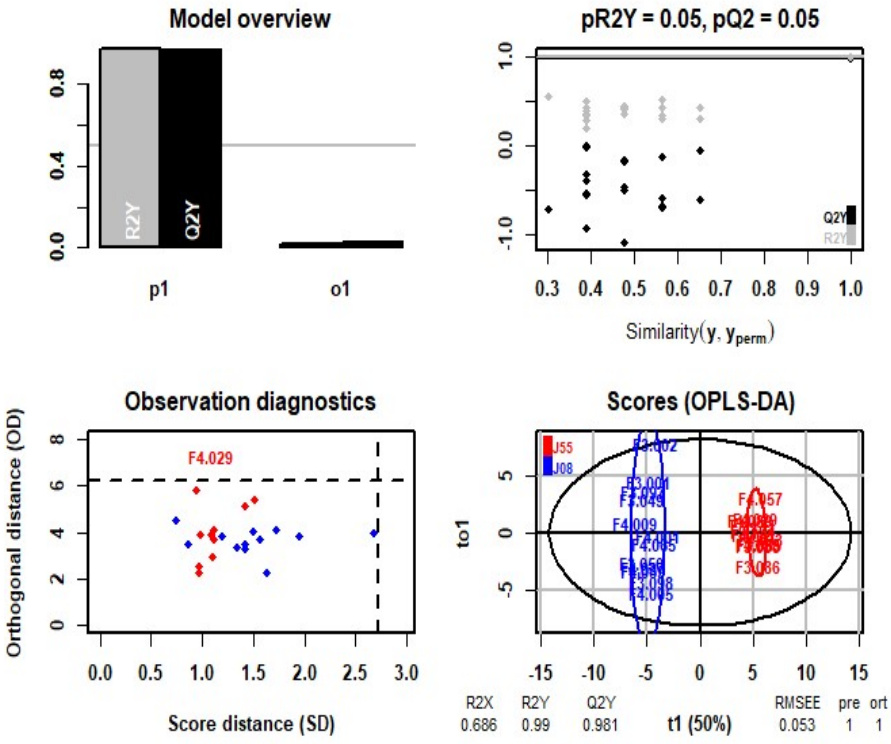
b.


Scores (PCA)
t2 (18%)
t1 (51%)
R2X
0.789

c.

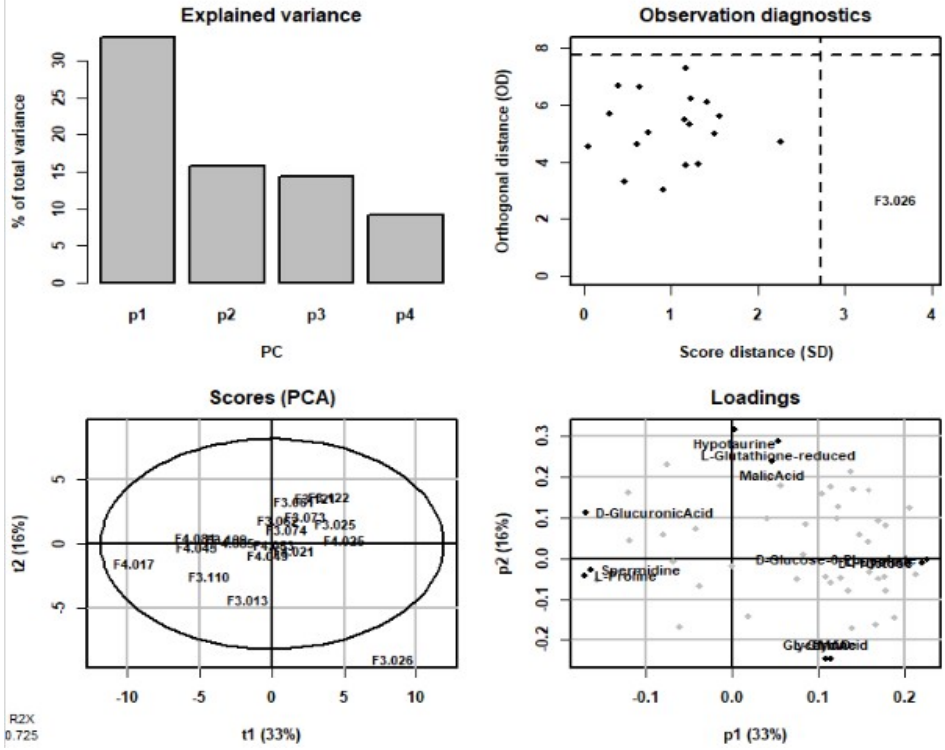


d.

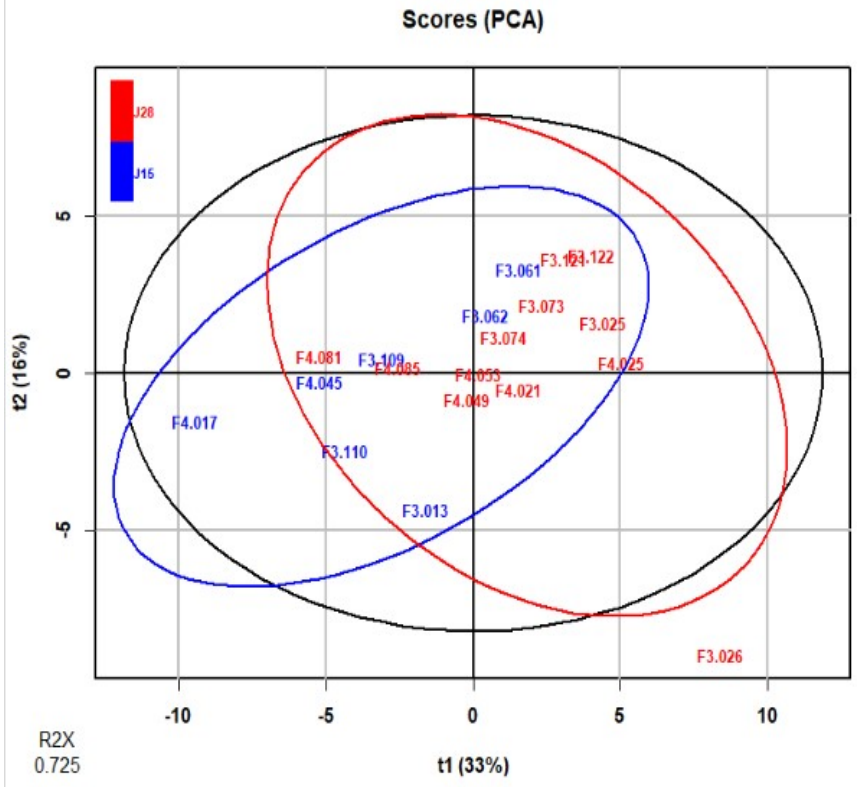


Εικόνα 3.4.2.3. Σύγκριση(2) σταδίων J08-J55 a. περιληπτικό γράφημα PCA , b. γράφημα PCA, c. PLS-DA μοντέλο , d. OPLS-DA μοντέλο

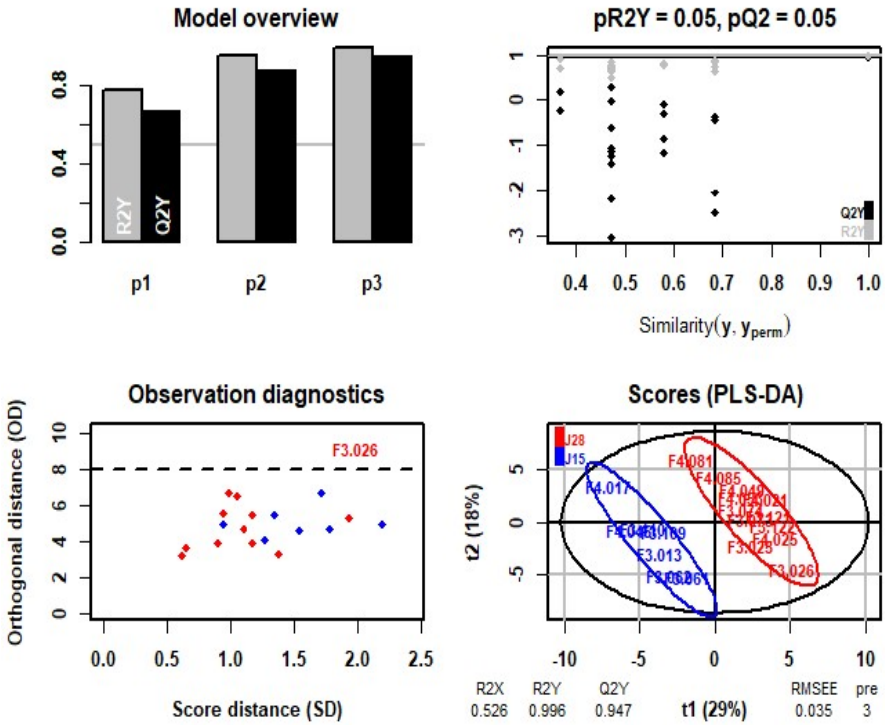
a.


Explained variance
% of total variance
p1
p2
p3
p4
PC
Observation diagnostics
Orthogonal distance (OD)
Score distance (SD)
F3.026
Scores (PCA)
t2 (16%)
t1 (33%)
R2X
0.725
Loadings
p2 (16%)
p1 (33%)
Hypotaurine
L-Glutathione-reduced
MalicAcid
D-GlucuronicAcid
Spermidine
L-Proline

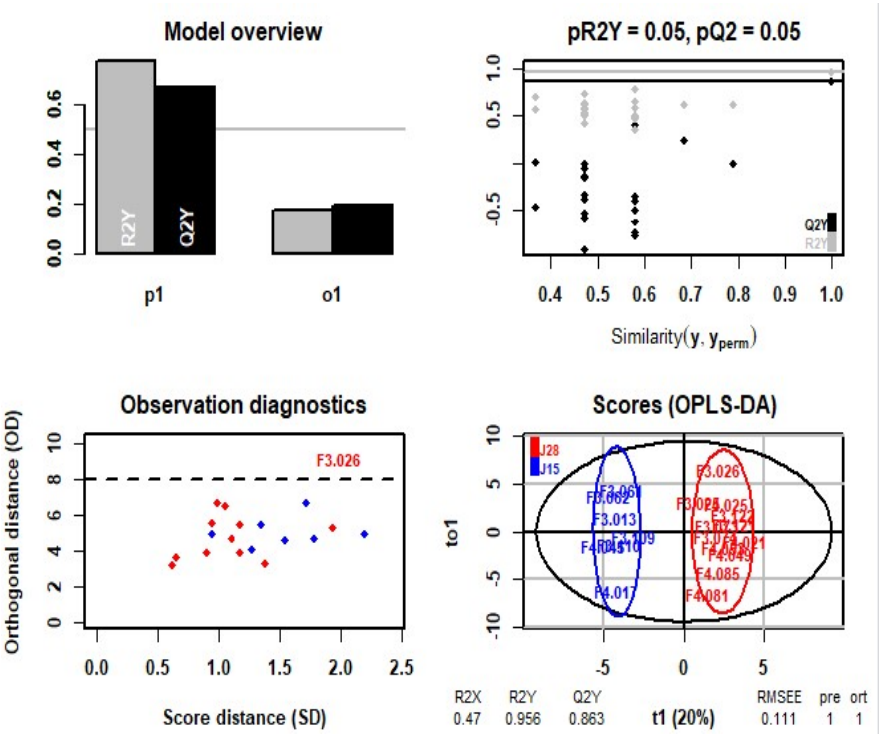
b.


Scores (PCA)
J28
J15
t2 (16%)
t1 (33%)
R2X
0.725
F3.061
F3.062
F3.073
F3.074
F3.025
F4.081
F4.045
F4.017
F3.110
F3.013
F4.021
F4.049
F4.025
F3.026

c.



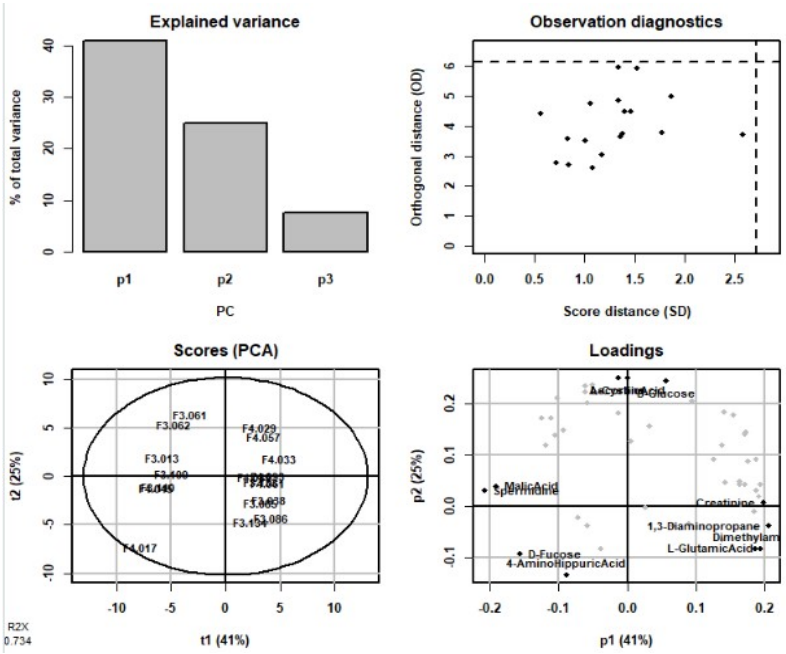
d.



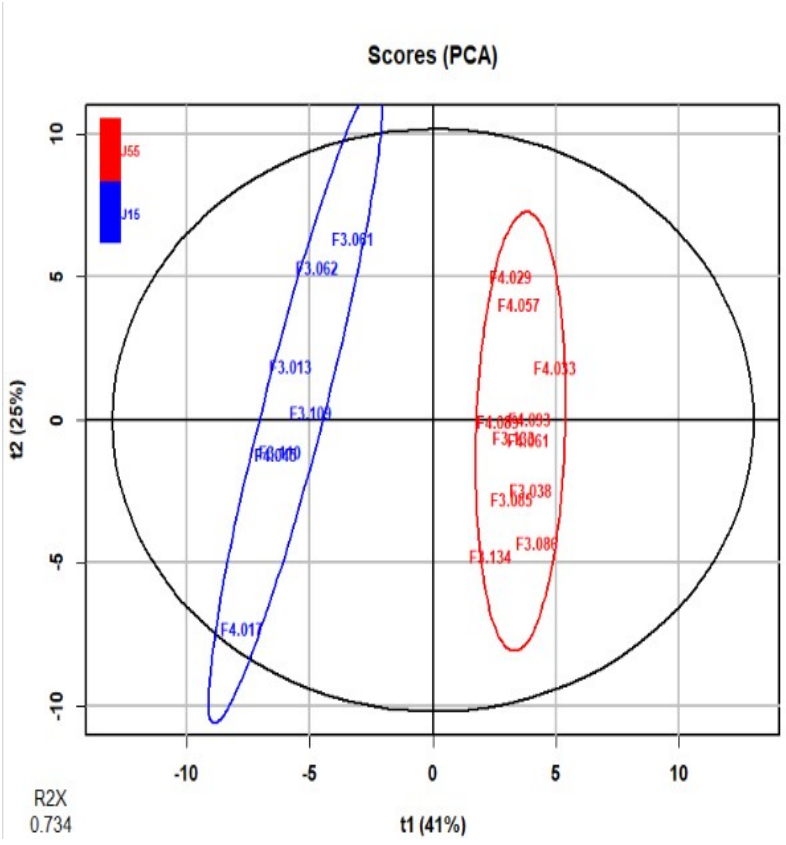
Εικόνα 3.4.2.4. Σύγκριση(2) σταδίων J15-J28 a. περιληπτικό γράφημα PCA , b. γράφημα PCA, c. PLS-DA μοντέλο , d. OPLS-DA μοντέλο

Στη σύγκριση των σταδίων J15 και J28 παρατηρούμε στο γράφημα PCA υπάρχει μεγάλη επικάλυψη μεταξύ των δύο ομάδων που προκύπτουν. Ωστόσο, στα μοντέλα PLS-DA και OPLS-DA παρατηρούμε ξεκάθαρο διαχωρισμό των ομάδων.

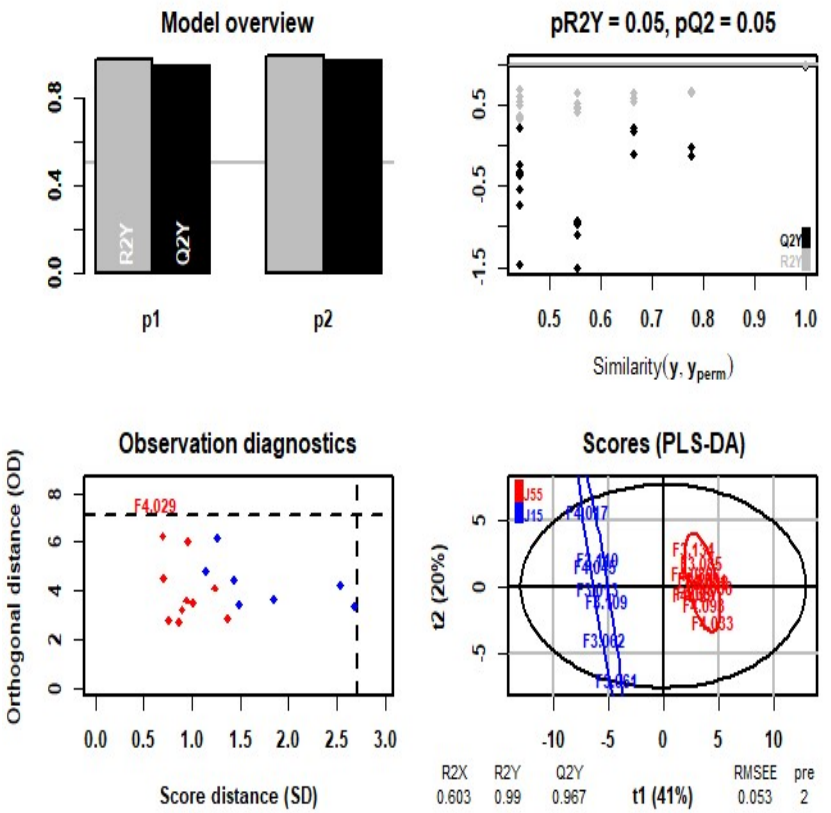
a.



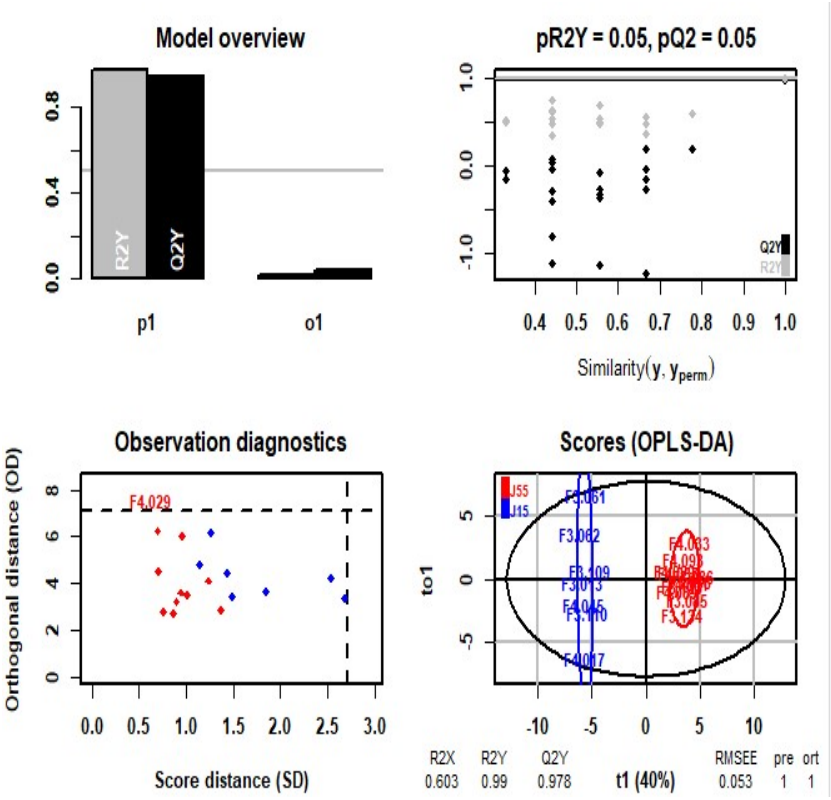
b.



c.

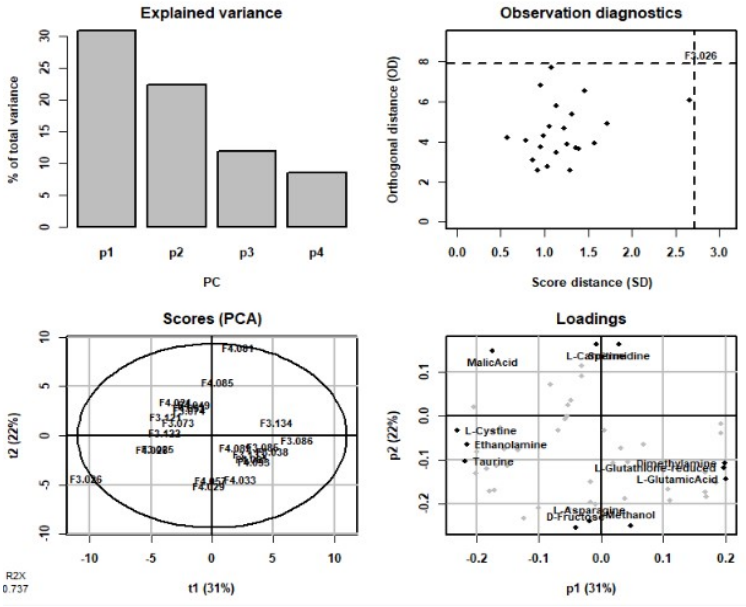


d.

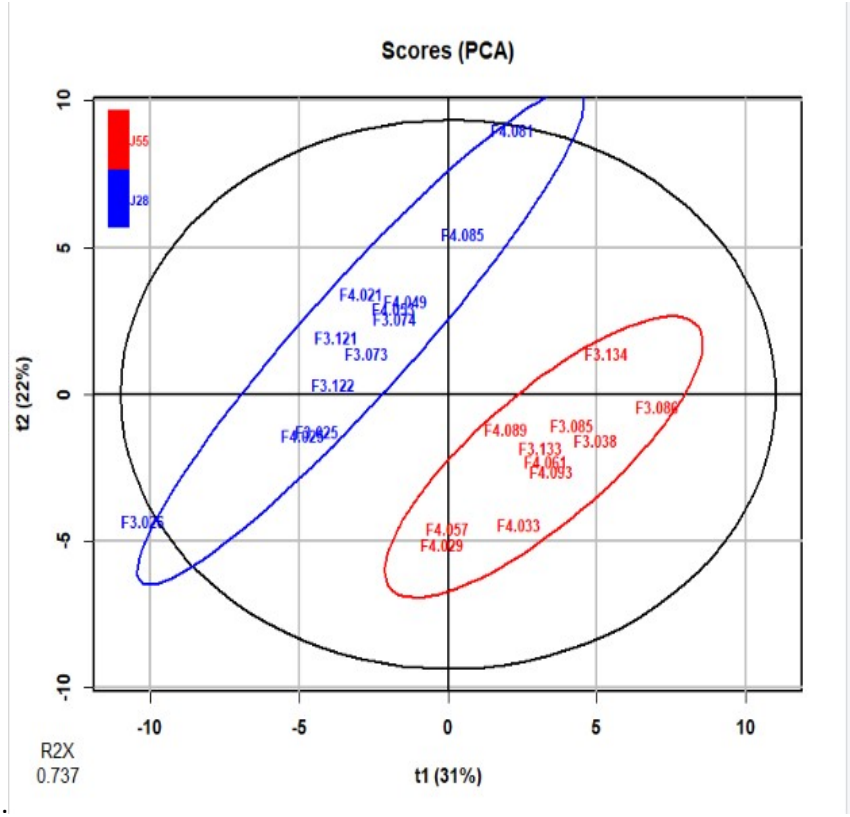


Εικόνα 3.4.2.5. Σύγκριση(2) σταδίων J15-J55 a. περιληπτικό γράφημα PCA , b. γράφημα PCA, c. PLS-DA μοντέλο , d. OPLS-DA μοντέλο

a.

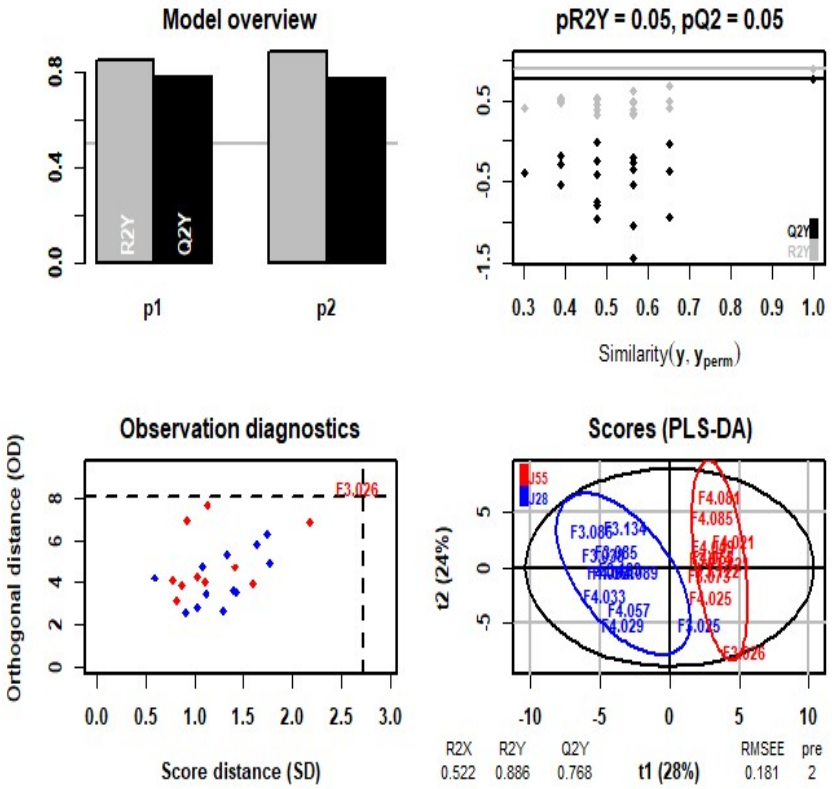

Explained variance
% of total variance
p1
p2
p3
p4
PC
Observation diagnostics
Orthogonal distance (OD)
Score distance (SD)
F3.026
Scores (PCA)
t2 (22%)
t1 (31%)
R2X
0.737
Loadings
p2 (22%)
p1 (31%)
MalicAcid
L-Cystine
Ethanolamine
Taurine
L-GlutamicAcid
L-Asparagine
D-Fructose
Methanol

b

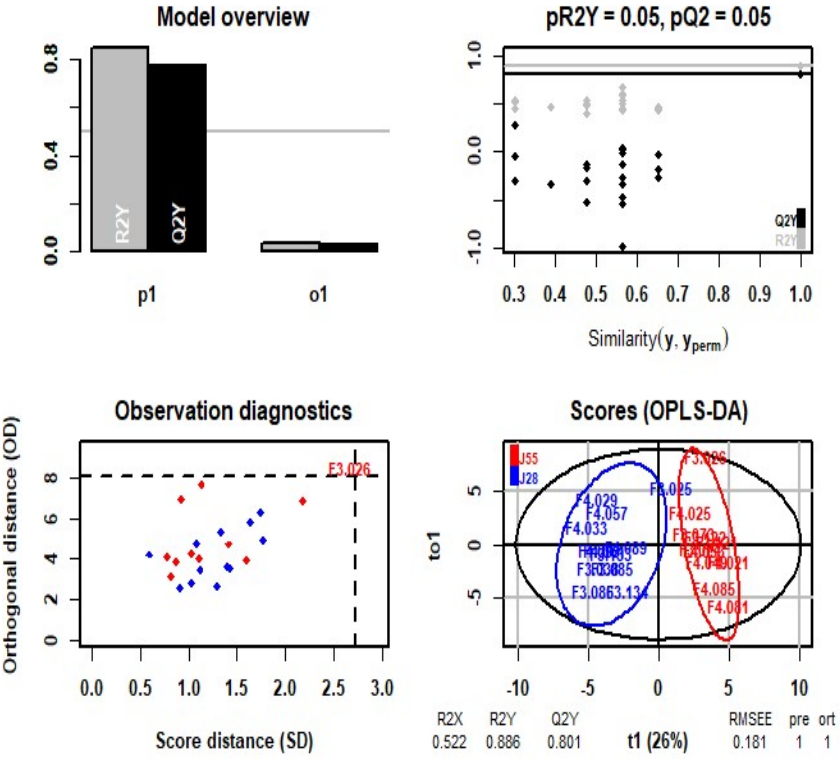

Scores (PCA)
J55
J28
t2 (22%)
t1 (31%)
R2X
0.737
F4.081
F4.085
F4.021
F4.049
F4.074
F3.121
F3.073
F3.122
F3.134
F3.086
F4.089
F3.085
F3.038
F3.133
F4.061
F4.093
F3.025
F4.057
F4.033
F4.029

.

c.



d.



Εικόνα 3.4.2.6. Σύγκριση(2) σταδίων J15-J28 a. περιληπτικό γράφημα PCA , b. γράφημα PCA, c. PLS-DA μοντέλο , d. OPLS-DA μοντέλο

Πίνακας 3.4.2.1. Αποτελέσματα από τα μοντέλα για τη Σύγκριση(2) .

	Διακριτική Ικανότητα 0,3		
	PCA	**PLS-DA**	**OPLS-DA**
J08 – J15	R^2=0,754	R^2=0,99 Q^2=0,963	R^2=0,99 Q^2=0,965
J08- J28	R^2=0,723	R^2=0,987 Q^2=0,967	R^2=0,987 Q^2=0,967
J08 – J55	R^2=0,789	R^2=0,99 Q^2=0,981	R^2=0,99 Q^2=0,981
J15 – J28	R^2=0,725	R^2=0,996 Q^2=0,947	R^2=0,956 Q^2=0,863
J15 – J55	R^2=0,734	R^2=0,99 Q^2=0,967	R^2=0,99 Q^2=0,978
J28 – J55	R^2=0,737	R^2=0,886 Q^2=0,768	R^2=0,886 Q^2=0,801

Για τη Σύγκριση(2) παρατηρούμε ότι οι τιμές στη PCA, PLS-DA και OPLS-DA είναι παρόμοιες και στις 6 διαφορετικές συγκρίσεις μεταξύ των σταδίων. Εξαιρείται η τελευταία σύγκριση J28 – J55 η οποία είναι η μόνη που έχει στα μοντέλα PLS-DA και OPLS-DA τιμές μικρότερες από 0,9.

1.11.3 Τρίτος τρόπος σύγκρισης σταδίων

Πίνακας 3.4.3.1. Τα δείγματα και οι μεταβολίτες που υπάρχουν σε κάθε στάδιο ξεχωριστά .

Στάδιο	J08	J15	J28	J55
Αριθμός Δειγμάτων	12	7	12	11
Αριθμός μεταβολιτών	51	48	50	53

Πίνακας 3.4.3.2 Δείγματα και οι μεταβολίτες που υπάρχουν ανά δύο στάδια.

Στάδιο	J08– J15	J08– J28	J08–J55	J15–J55	J15– J28	J28- J55
Αριθμός Δειγμάτων	19	24	23	18	19	23
Αριθμός μεταβολιτών	54	49	53	54	53	55

Ένας ακόμα τρόπος όπου θα μπορούσε να γίνει η σύγκριση μεταξύ των διαφορετικών σταδίων είναι να γίνει ταυτοποίηση και ποσοτικοποίηση των μεταβολιτών ανά δύο στάδια και ύστερα σύγκριση μεταξύ των δύο αυτών όπου αρχικά έγινε η ταυτοποίηση και η ποσοτικοποίηση. Ωστόσο, μέσω αυτού του τρόπου δεν παρατηρούνται διαφορές καθώς δεν περιέχεται όλο τα φάσμα όλων των σταδίων με αποτέλεσμα να γίνεται κανονικοποίηση των δεδομένων και οι διαφορές όλων των σταδίων μεταξύ τους να είναι μηδαμινές και για αυτό δεν συμπεριλήφθηκαν στην παρούσα εργασία.

1.12 Πίνακας ταυτοποιημένων μεταβολιτών για τα διαφορετικά στάδια ανάπτυξης της ντομάτας

Σε αυτήν την υποενότητα παρουσιάζονται τα αποτελέσματα που εξήχθησαν με τη βοήθεια του πακέτου ASICS. Δημιουργήθηκε κώδικας και χρησιμοποιήθηκε το πακέτο μέσω του οποίου ταυτοποιήθηκαν οι μεταβολίτες που υπάρχουν στα διαφορετικά στάδια ωρίμανσης της ντομάτας.

Πίνακας 3.5.1.Μεταβολίτες που ταυτοποιήθηκαν σε κάθε στάδιο

J08	J15	J28	J55
1-Methylhydantoin	1-Methylhydantoin	1-Methylhydantoin	1-Methylhydantoin
2-Oxoglutarate	2-Oxoglutarate	2-Oxoglutarate	2-Oxoglutarate
AscorbicAcid	AscorbicAcid	AscorbicAcid	AscorbicAcid
Beta-Alanine	Beta-Alanine	Beta-Alanine	Beta-Alanine
Betaine	Betaine	Betaine	Betaine
CholineChloride	CholineChloride	CholineChloride	CholineChloride
Creatine	Creatine	Creatine	Creatine
DehydroAscorbicAcid	DehydroAscorbicAcid	DehydroAscorbicAcid	DehydroAscorbicAcid
D-Fructose	D-Fructose	D-Fructose	D-Fructose
D-GluconicAcid	D-GluconicAcid	D-GluconicAcid	D-GluconicAcid
D-Glucose	D-Glucose	D-Glucose	D-Glucose
D-Glucose-6-Phosphate	D-Glucose-6-Phosphate	D-Glucose-6-Phosphate	D-Glucose-6-Phosphate
D-GlucuronicAcid	D-GlucuronicAcid	D-GlucuronicAcid	D-GlucuronicAcid
Dimethylamine	Dimethylamine	Dimethylamine	Dimethylamine
Dimethylglycine	Dimethylglycine	Dimethylglycine	Dimethylglycine
Dimethylsulfone	Dimethylsulfone	Dimethylsulfone	Dimethylsulfone
D-Maltose	D-Maltose	D-Maltose	D-Maltose
Ethanolamine	Ethanolamine	Ethanolamine	Ethanolamine
Galactitol	Galactitol	Galactitol	Galactitol
Glycerophosphocholine	Glycerophosphocholine	Glycerophosphocholine	Glycerophosphocholine
Glycogen	Glycogen	Glycogen	Glycogen
GuanidinoaceticAcid	GuanidinoaceticAcid	GuanidinoaceticAcid	GuanidinoaceticAcid
Hypotaurine	Hypotaurine	Hypotaurine	Hypotaurine
IsocitricAcid	IsocitricAcid	IsocitricAcid	IsocitricAcid
Lactose	Lactose	Lactose	Lactose
L-Arabitol	L-Arabitol	L-Arabitol	L-Arabitol
L-Asparagine	L-Asparagine	L-Asparagine	L-Asparagine
L-Aspartate	L-Aspartate	L-Aspartate	L-Aspartate
L-Carnitine	L-Carnitine	L-Carnitine	L-Carnitine
L-Cystine	L-Cystine	L-Cystine	L-Cystine
L-Glutathione-oxidized	L-Glutathione-oxidized	L-Glutathione-oxidized	L-Glutathione-oxidized

L-Glutathione-reduced	L-Glutathione-reduced	L-Glutathione-reduced	L-Glutathione-reduced
L-Proline	L-Proline	L-Proline	L-Proline
MalicAcid	MalicAcid	MalicAcid	MalicAcid
Methanol	Methanol	Methanol	Methanol
Methylamine	Methylamine	Methylamine	Methylamine
Myo-Inositol	Myo-Inositol	Myo-Inositol	Myo-Inositol
Phosphocholine	Phosphocholine	Phosphocholine	Phosphocholine
PyroglutamicAcid	PyroglutamicAcid	PyroglutamicAcid	PyroglutamicAcid
Succinate	Succinate	Succinate	Succinate
Taurine	Taurine	Taurine	Taurine
TMAO	TMAO	TMAO	TMAO
Trans-AcotinicAcid	Trans-AcotinicAcid	Trans-AcotinicAcid	Trans-AcotinicAcid
4-AminoHippuricAcid	4-AminoHippuricAcid	1,3-Diaminopropane	1,3-Diaminopropane
D-Fucose	D-Fucose	Xylitol	Xylitol
PropyleneGlycol	PropyleneGlycol	Pyruvic-Acid	Pyruvic-Acid
L-Cysteine	Creatinine	Creatinine	Creatinine
Spermidine	Spermidine	Spermidine	L-GlutamicAcid
GABA	-	GABA	GABA
trans-4-Hydroxy-L-Proline	-	trans-4-Hydroxy-L-Proline	trans-4-Hydroxy-L-Proline
GlycericAcid	-	GlycericAcid	-
-	-	Glycerol	-
-	-	L-Glycine	-

Παρατηρούμε ότι υπάρχει πολύ μεγάλη ομοιότητα μεταξύ των μεταβολιτών που ταυτοποιήθηκαν για κάθε στάδιο.

Πίνακας 3.5.2. Συγκεντρωτικά τα στοιχεία του πίνακα 3.5.1 με όλους τους μεταβολίτες του φάσματος που αναγνωρίστηκαν σε κάθε στάδιο συνδυαστικά.

	Κοινοί μεταβολίτες	**Όλοι οι μεταβολίτες**	**Κοινοί μεταβολίτες για όλα τα στάδια**
J08-J15	47	52	44
J08-J28	47	57	
J08-J55	45	56	
J15-J28	45	56	
J15-J55	44	54	
J28-J55	49	54	

ΣΥΖΗΤΗΣΗ

Στον πίνακα 3.1.1 παρατηρείται ότι καλύτερα αποτελέσματα έχουμε με την ομοιόμορφη τμηματοποίηση καθώς με μικρότερο αριθμό τμημάτων έχουμε μεγαλύτερη τιμή σήματος προς θόρυβο (SNR). Όταν το SNR είναι μεγάλο σημαίνει ότι ο θόρυβος, δηλαδή το μη επιθυμητό σήμα, είναι χαμηλός και δεν επηρεάζει σημαντικά την μέτρηση του πραγματικού σήματος. Ωστόσο, το αποτέλεσμα αντικρούει την αρχική υπόθεση ότι η ευφυής προσαρμοστική τμηματοποίηση θα είχε πολύ υψηλότερο SNR σε σχέση με την ομοιόμορφη τμηματοποίηση.

Στον πίνακα 3.1.2 παρατηρείται ότι το εύρος 2,5-4,5ppm διαθέτει τις μεγαλύτερες τιμές (counts) στον πίνακα. Επίσης παρατηρήθηκαν χαμηλές τιμές SNR για το εύρος 6,5-10ppm. Στην ευφυή προσαρμοστική τμηματοποίηση εμφανίζονται παρόμοιες τιμές παρά την αλλαγή στην διακριτική ικανότητα, σε αντίθεση με την ομοιόμορφη τμηματοποίηση. Αυτό οδηγεί στο συμπέρασμα ότι στην ομοιόμορφη τμηματοποίηση το SNR επηρεάζεται από τις διπλές κορυφές που μπορεί να υπάρχουν σε ένα μόνο τμήμα, γεγονός και το οποίο προκαλεί τις μεγαλύτερες τιμές. Ωστόσο, έτσι δημιουργείται μια ψευδή εικόνα για την αποτελεσματικότητα της τμηματοποίησης του φάσματος, γεγονός που φανερώνει ότι η πρώτη σύγκριση (πίνακας 3.1.1) είναι εσφαλμένη. Το πρόβλημα αυτό λύνεται με την ευφυή προσαρμοστική τμηματοποίηση καθώς σε κάθε τμήμα υπάρχει μόνο μία κορυφή, που οδηγεί σε ένα αποτέλεσμα που είναι πιο κοντά στην πραγματικότητα, δηλαδή μπορεί να επιβεβαιωθεί παρατηρώντας οπτικά το φάσμα. Συνεπώς επιλέγεται η ευφυής προσαρμοστική τμηματοποίηση ως η καλύτερη επιλογή για τη συνέχεια της έρευνας.

Από την οπτική εξέταση (Εικόνες 3.1.1-5) για τις διαφορετικές διακριτικές ικανότητες παρατηρούμε ότι στις μικρές κορυφές (Εικόνα 3.1.4) είναι δύσκολο να διακριθεί αν η τμηματοποίηση του φάσματος είναι σωστή καθώς μοιάζει πολύ με σήμα θορύβου (πολύ περίπλοκο φάσμα). Παρατηρείται, επίσης, ότι σε αυτού του τύπου φάσματα ο διαχωρισμός των κορυφών είναι σχεδόν τυχαίος, αφού σε κάποιες περιπτώσεις τοποθετεί διπλές κορυφές στο ίδιο τμήμα ενώ σε άλλες το ξεχωρίζει ανεξάρτητα από την διακριτική ικανότητα (Εικόνα 3.1.5). Ακόμα, όταν η ευθυγράμμιση των φασμάτων μεταξύ τους (alignment) δεν είναι ιδανική (δηλαδή οι κορυφές του ίδιου σήματος όλων των δειγμάτων-φασμάτων δεν συμπίπτουν ακριβώς η μία πάνω στην άλλη) τότε η τμηματοποίηση έχει αστοχίες, καθώς η μία κορυφή δεν μπορεί να οριστεί από συγκεκριμένη τιμή ppm άλλα από ένα πολύ μεγαλύτερο εύρος ppm (Εικόνα 3.1.4). Τέλος, παρατηρήθηκε ότι για μικρότερη διακριτική ικανότητα είναι πιο συχνό το φαινόμενο μια κορυφή να κόβεται στη μέση και να διασπάται σε δύο αντί για ένα τμήμα. Παραδείγματα αυτής της αστοχίας είναι στα 8,84ppm, 8,01-8,08ppm, 7,48ppm (Εικόνα 3.1.4). Το γενικό συμπέρασμα το οποίο προκύπτει από την οπτική εξέταση, σύμφωνα και με τα παραπάνω, είναι ότι σε πολύ μικρές ή σε πολύ μεγάλες τιμές διακριτικής ικανότητας δημιουργούνται αστοχίες στη τμηματοποίηση των κορυφών του φάσματος. Έπειτα λαμβάνοντας υπόψη τα αποτελέσματα των στατιστικών αναλύσεων παρατηρούμε στη PCA ότι καλύτερος διαχωρισμός μεταξύ των ομάδων (clusters) παρατηρείται για resolution 0,3 με το R^2 να είναι μεγαλύτερο (για resolution 0,3).

Έπειτα ακολούθησε έλεγχος της επίδρασης της διακριτικής ικανότητας στις ομάδες με βάση τις συνθήκες ανάπτυξης (control/ shadow) και τα στάδια ωρίμανσης της ντομάτας (J08/J15/J28/J55) (Πίνακας 3.2.1). Στη σύγκριση με βάση τις συνθήκες παρατηρείται ότι στο PLS-DA οι τιμές R^2 και Q^2 είναι μεγαλύτερες για διακριτική ικανότητα 0,3. Ωστόσο, παρατηρούμε υπερπροσαρμογή (overfitting) στο PLS-DA και στο OPLS-DA καθώς το Q^2 έχει αρνητικό πρόσημο και το R^2 λαμβάνει πολύ μεγάλη τιμή. Όταν το R^2 και Q^2 έχουν αυτές τις τιμές δηλώνει ότι υπάρχει τυχαιότητα στην επιλογή των χαρακτηριστικών άρα των τμημάτων που χρησιμοποιήθηκαν προκειμένου να βγει το τελικό αποτέλεσμα (Y). Στη σύγκριση με βάση τα στάδια ωρίμανσης της ντομάτας παρατηρούμε ότι στο PLS-DA οι τιμές R^2 και Q^2 είναι παρόμοιες για τις διακριτικές ικανότητες 0,3 και 0,1. Επίσης παρατηρούμε ότι στο PLS-DA οι τιμές R^2 και Q^2 είναι μεγάλες, γεγονός που δηλώνει την καλή προβλεψιμότητα του μοντέλου. Επιβεβαιώνεται ότι το OPLS-DA μοντέλο δεν δίνει αποτελέσματα αν υπάρχουν περισσότερες από δύο κλάσεις/ομάδες, (Πίνακας 3.2.1) καθώς με την σύγκριση των συνθηκών προκύπτουν αποτελέσματα, ενώ στα στάδια που είναι 4 δεν προκύπτουν αποτελέσματα. Τέλος, παρατηρείται ότι τα δεδομένα του φάσματος δεν μπορούν να διαχωριστούν με βάση τις συνθήκες (condition) αλλά με βάση το στάδιο ωρίμανσης της ντομάτας (stage).

Με το παραπάνω πόρισμα καταλήγουμε ότι οι μεγάλες διαφορές στα δεδομένα είναι στα διαφορετικά στάδια ωρίμανσης της ντομάτας καθώς αναπτύσσονται διαφορετικού είδους μεταβολίτες σε κάθε στάδιο. Είναι φανερό ότι οι συνθήκες επηρεάζουν την ανάπτυξη της ντομάτας αλλά όχι σε τέτοιο βαθμό ώστε να μπορούν να οριστούν δύο τελείως διακριτές ομάδες.

Από την ταυτοποίηση των μεταβολιτών (Πίνακας 3.3.1) παρατηρήθηκε ότι για το resolution 0,5 ανιχνεύτηκαν περισσότεροι μεταβολίτες σε σχέση με τις υπόλοιπες τιμές διακριτικής ικανότητας. Στον πίνακα 3.3.2 το R^2 στο PCA είναι μεγαλύτερο για διακριτική ικανότητα 0,1 και στο PLS-DA το Q^2 είναι μεγαλύτερο για το διακριτική ικανότητα 0,3. Οι ομάδες οι οποίες προκύπτουν στο PCA μοντέλο παρατηρούμε ότι διαχωρίζονται καλύτερα όταν η διακριτική ικανότητα είναι 0,3 καθώς φαίνεται τα δείγματα να είναι πιο σωστά και πιο συμπυκνωμένα σε κάθε ομάδα σε αντίθεση με τις άλλες δύο διακριτικές ικανότητες. Θεωρούμε την τιμή Q^2, δηλαδή την προβλεψιμότητα του μοντέλου πολύ σημαντική, διότι αυτή η τιμή υποδεικνύει πόσο επιτυχημένο μπορεί να είναι το μοντέλο και σε τυχαία δεδομένα. Με αυτόν τον τρόπο καταλήγουμε στο συμπέρασμα ότι η διακριτική ικανότητα 0,3 ακόμα και στην ταυτοποίηση των δεδομένων είναι η καλύτερη επιλογή ώστε να εξάγουμε τα δεδομένα μας από το NMRprocFlow.

Στη σύγκριση των σταδίων ανά δύο παρατηρούμε τα εξής συμπεράσματα. Η μεγαλύτερη διαφοροποίηση παρουσιάζεται για τα στάδια J08-J55. Επίσης στο δεύτερο τρόπο διαφοροποίησης παρατηρούμε καλύτερα αποτελέσματα στα μοντέλα PLS-DA και OPLS-DA, γεγονός που οδηγεί στο συμπέρασμα ότι ο δεύτερος τρόπος σύγκρισης [Σύγκριση(2)] μεταξύ των σταδίων, δηλαδή όταν η ταυτοποίηση και η ποσοτικοποίηση των μεταβολιτών για το κάθε στάδιο πραγματοποιείται ξεχωριστά, λειτουργεί πιο αποτελεσματικά.

ΣΥΜΠΕΡΑΣΜΑΤΑ

Μέσω της διπλωματικής αυτής έγινε χρήση στατιστικών αναλύσεων προκειμένου να βρεθεί ο καλύτερος δυνατός τρόπος επεξεργασίας και εξαγωγής μεταβολομικών δεδομένων. Μέσω του προγράμματος NMRProcFlow επεξεργαστήκαμε το φάσμα και ύστερα μέσω της στατιστικής ανάλυσης ορίσαμε την καλύτερη λύση για το συγκεκριμένο σετ μεταβολομικών δεδομένων. Η ευφυής προσαρμοστική τμηματοποίηση (AI bucketing) φέρει καλύτερα αποτελέσματα καθώς προσαρμόζει το μέγεθος των τμημάτων της. Η διακριτική ικανότητα 0,3 λειτουργεί καλύτερα στο συγκεκριμένο σετ μεταβολομικών δεδομένων καθώς ορίζει με μεγαλύτερη ακρίβεια την τμηματοποίηση. Με γνώμονα τη μηχανική μάθηση και χρησιμοποιώντας ως εργαλείο τη γλώσσα προγραμματισμού R, δημιουργήθηκε κώδικας ο οποίος ταυτοποιεί αυτόματα μεταβολίτες (untargeted metabolomics). Αναλύθηκαν οι τρόποι όπου μπορούσε να επιτευχθεί η ταυτοποίηση των μεταβολιτών ώστε να διαφοροποιούνται τα στάδια μεταξύ τους. Μετά πάλι μέσω στατιστικής ανάλυσης ορίστηκε η καλύτερη δυνατή μέθοδος. Αποδείχθηκε ότι ο καλύτερος τρόπος μέσω του οποίου θα προκύψουν 4 διαφορετικές ομάδες είναι όταν τα μεταβολομικά δεδομένα ποσοτικοποιηθούν και ταυτοποιηθούν για κάθε στάδιο ξεχωριστά. Η έρευνα αυτή, αποτελεί μια αρχή ώστε να δημιουργηθεί ένα εργαλείο το οποίο θα επεξεργάζεται και θα ταυτοποιεί αυτόματα φάσματα μεταβολιτών. Η διπλωματική βασίστηκε σε φάσματα φυτικού ιστού οπότε ενδέχεται τα αποτελέσματα να διαφέρουν για διαφορετικά υποστρώματα, δηλαδή βιολογικά υγρά όπως ούρα, αίμα κλπ..

Αναφορές-Πηγές

Beauvoit, B. P. et al., 2014. Model-Assisted Analysis of Sugar Metabolism throughout Tomato Fruit Development Reveals Enzyme and Carriew Properties in Relation to Vacuole Expansion. *The Plant Cell,* 26(1), pp. 3224-3242.

Bénard, C. et al., 2015. Metabolomic profiling in tomato reveals diel compositional changes in fruit affected by source-sink relationships. *Journal of Experimental Botany,* 66(11), pp. 3391-3404.

Biais, B. et al., 2014. Remarkable Reproducibility of Enzyme Activity Profiles in Tomato Fruits Grown under Contrasting Environments Provides a Roadmap for Studies of Fruit Metabolism. *Plant Physiology,* 164(3), pp. 1204-1221.

Colombié, S. et al., 2014. Modelling central metabolic fluxes by constraint-based optimization reveals metabolic reprogramming of developing Solanum lycopersicum (tomato) fruit. *Plant Journal,* 81(1), pp. 24-39.

Gerszberg, A., Hnatuszko, K., Kowalcyzk, T. & Kononowicz, A., 2015. Tomato (Solanum lycopersicum L.) in the service of biotechnology. *Plant Cell Tiss Organ Cult,* 1(120), pp. 881-902.

Jacob, D. et al., 2017. NMRProcFlow: a graphical and interactive tool dedicated to 1D spectra processing for NMR-based metabolomics. *Metabolomics,* 1(13), p. 36.

Lefort, G., Servien, R. & Vialaneix, N., 2020. *Bioconductor:Open Source Software for Bioinformatics.* [Online]
Available at:
https://www.bioconductor.org/packages/devel/bioc/vignettes/ASICS/inst/doc/ASICSUsersGuide.html
[Accessed 02 May 2021].

Maćkiewicz, A. & Ratajczak, W., 1993. Principal components analysis (PCA). *Computers & Geosciences,* 19(3), pp. 303-342.

Meyer, T. D. et al., 2008. NMR-Based Characterization of Metabolic Alterations in Hypertension Using an Adaptive, Intelligent Binning Algorithm. *Analytical Chemistry,* 10(80), pp. 3783-3790.

Tardivel, P. et al., 2017. ASICS: an automatic method for identification and quantification of metabolites in complex 1D 1H NMR spectra. *Metabolomics,* 13(10), p. 109.

Thevenot, E. A., 2020. *Bioconductor: Open Source Software for Bioinformatics.* [Online]
Available at:
https://master.bioconductor.org/packages/release/bioc/vignettes/ropls/inst/doc/ropls-vignette.html
[Accessed 23 April 2021].

Γεωργακοπούλου, Ι., 2018. *Μεταβολομική ανάλυση βιολογικών υγρών για την ανίχνευση βιοδεικτών, μέσω Φασματοσκοπίας NMR,* Πάτρα: Πανεπιστήμιο Πατρών.

Μπρέστα, Π., 2009. *Επίδραση της υδατικής καταπόνησης σε φυσιολογικές και ανατομικές παραμέτρους των φύλλων σε ανθεκτικές και μη ποικιλίες σίτου,* Αθήνα: Γεωπονικό Πανεπιστήμιο.

Φιλντίση, Α. Ι., 2018. *Ανάπτυξη Τεχνικών Επεξεργασίας Δεδομένων στην Αντικαρκινική Θεραπεία και τη Μεταβολομική,* Αθήνα: Εθνικό Μετσόβιο Πολυτεχνείο.