ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΛΙΑΣ

ΣΧΟΛΗ ΓΕΩΠΟΝΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΓΕΩΠΟΝΙΑΣ, ΦΥΤΙΚΗΣ ΠΑΡΑΓΩΓΗΣ & ΑΓΡΟΤΙΚΟΥ ΠΕΡΙΒΑΛΛΟΝΤΟΣ

Εργαστήριο Βιομετρίας

Πτυχιακή Διατριβή

«Εργαλεία βιοπληροφορικής σε δεδομένα ομικής (omics) με εφαρμογή στην εύρεση βιοδεικτών ανθεκτικότητας της φακής έναντι καταπονήσεων»

Επιβλέπων Καθηγητής

Νάκας Χρήστος

Μέλη Τριμελούς Επιτροπής

Παυλή Ουρανία

Λεβίζου Ευθυμία

Τρίγκα Ρόζα Μαρία

Βόλος 2021

«Εργαλεία βιοπληροφορικής σε δεδομένα ομικής (omics) με εφαρμογή στην εύρεση βιοδεικτών ανθεκτικότητας της φακής έναντι καταπονήσεων»

«Bioinformatics tools in omics data with an application to biomarkers of lentil stress resistance»

Νάκας Χρήστος, Καθηγητής,
Βιομετρία
Παυλή Ουρανία, Επίκουρη Καθηγήτρια,
Γενετική Βελτίωση Φυτών
Λεβίζου Ευθυμία, Επίκουρη Καθηγήτρια,
Φυσιολογία του Φυτού στις Γεωπονικές Επιστήμες

Βεβαιώνω ότι είμαι συγγραφέας αυτής της πτυχιακής εργασίας, η οποία εκπονήθηκε σύμφωνα με τον Κανονισμό Πτυχιακής Εργασίας του Τμήματος Γεωπονίας, Φυτικής Παραγωγής και Αγροτικού Περιβάλλοντος.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέπων καθηγητή κ. Νάκα Χρήστο καθώς και τα υπόλοιπα μέλη της τριμελούς επιτροπής Παυλή Ουρανία και Λεβίζου Ευθυμία.

Περιεχόμενα

Σκοπός

Σκοπός της παρούσας πτυχιακής διατριβής είναι η εύρεση πιθανών βιοδεικτών στη φακή έτσι ώστε να υπάρχει η δυνατότητα πρόβλεψης ανθεκτικότητας ή ασθένειας των φυτών όταν αυτά προσβληθούν από τον μύκητα *Fusarium oxysporum* f.sp. lentis.

1. Εισαγωγή

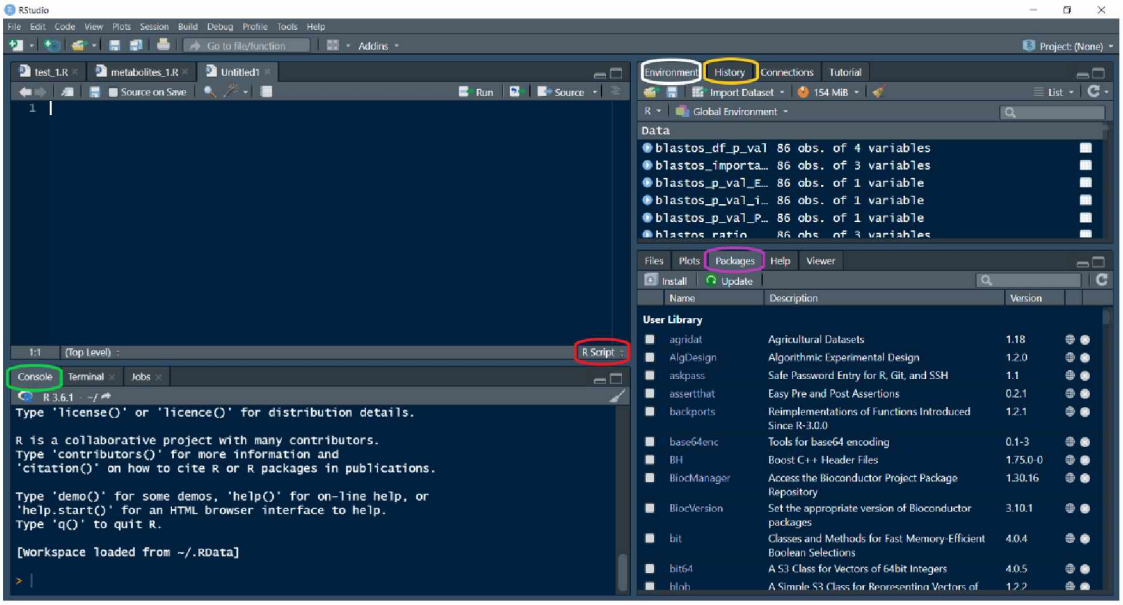
1.1 Η γλώσσα Προγραμματισμού R

Στην στατιστική ανάλυση χρησιμοποιείται ευρέως η R. Είναι μία γλώσσα προγραμματισμού υψηλού επιπέδου η οποία πρωτοεμφανίστηκε το 1993 από τους Ross Ihaka και Robert Gentleman καθηγητές του Πανεπιστημίου του Auckland της Νέας Ζηλανδίας. Πήρε το όνομα της από το κοινό αρχικό γράμμα του μικρού ονόματος των δημιουργών της. Αποτελεί απόγονο της γλώσσας προγραμματισμού S που δημιουργήθηκε από τον John Chambers και τους φοιτητές του στα Bell Laboratories. Η R δίνει την δυνατότητα ποικίλων στατιστικών αναλύσεων (γραμμικών και μη μοντέλων, στατιστικών t-test κ.α.) αλλά και την παραγωγή γραφημάτων όπως ραβδογραμμάτων, boxplots, heatmaps καθώς και την παραμετροποίηση του σχεδιασμού αυτών ανάλογα με τις προτιμήσεις του εκάστοτε χρήστη. Το λογισμικό της διατίθεται ελεύθερα τηρώντας τις προϋποθέσεις του GNU project (*What is r?* R. (n.d.). Retrieved September 21, 2021, from https://www.r-project.org/about.html). Ελεύθερο χαρακτηρίζεται ένα λογισμικό όταν ο κάθε χρήστης έχει πρόσβαση στον πηγαίο κώδικα του προγράμματος και μπορεί να αντιγράψει, να διανείμει, να τροποποιήσει αλλά και να βελτιώσει το λογισμικό χωρίς να χρειαστεί να ζητήσει άδεια από τον δημιουργό του προγράμματος (*The GNU operating system and the free software movement*. [A GNU head] . (n.d.). Retrieved September 21, 2021, from http://www.gnu.org/). Στην περίπτωση που ο προγραμματιστής ανακαλέσει αύτη την δυνατότητα από τους χρήστες τότε το λογισμικό παύει να είναι ελεύθερο. Μπορεί να τρέξει σε όλα τα λειτουργικά συστήματα όπως MacOS , Windows, Linux.

1.1.1. Περιβάλλον RStudio και Τρόπος Λειτουργίας των Πακέτων και των Συναρτήσεων.

Το R studio είναι το εργαλείο μέσω του οποίου ο χρήστης μπορεί να γράψει κώδικα, να δημιουργήσει και να εμφανίσει γραφήματα. Παρέχει την δυνατότητα εγκατάστασης πακέτων. Τα πακέτα είναι συλλογές συναρτήσεων μέσω των

οποίων μπορεί να παραχθεί ένα γράφημα (π. χ *boxplot()*), να δημιουργηθούν πίνακες (π. χ *as.matrix()*). Μέσω των πακέτων γίνεται η επέκταση χρήσης της γλώσσας. Ακόμη μέσα από το RStudio δίνεται η δυνατότητα αποθήκευσης του



Εικόνα 1. RStudio

κώδικα, προβολής του ιστορικού αλλά και του περιβάλλοντος (Environment) στο οποίο εμφανίζονται τα ονόματα των τύπων δεδομένων που έχουν δημιουργηθεί κατά τη διάρκεια δημιουργίας του κώδικα. Τέλος στην κονσόλα (Console) εμφανίζονται τα αποτελέσματα εκτέλεσης του κώδικα και προειδοποιήσεις όταν υπάρχει λάθος (Error). Υπάρχει η δυνατότητα να γραφεί κώδικας στην κονσόλα αλλά δεν αποθηκεύεται.

1.2 Η Φακή

Η φακή (*Lens culinaris* Medik.) είναι ένα από τα αρχαιότερα καλλιεργούμενα είδη στον κόσμο και η καλλιέργειά του ξεκίνησε πριν από περίπου 6.500 π.Χ. ανήκει στην οικογένεια των Fabaceae όπως η μηδική και το μπιζέλι. Η *Lens culinaris* ssp. *culinaris* είναι το μοναδικό είδος έναντι των επτά που περιέχει το γένος *Lens* που είναι καλλιεργούμενο. Είναι ετήσιο, αυτογονιμοποιούμενο φυτό, διπλοειδές με 2n = 14 και καλλιεργείται κατά τη διάρκεια της χειμερινής περιόδου του έτους. Πρόγονος της είναι η *Lens culinaris* ssp. *orientalis* που καλλιεργούνταν στη Μεσόγειο. Ανήκει στα ψυχανθή επομένως με ενσωμάτωση της φυτικής βιομάζας στο έδαφος απελευθερώνονται σημαντικές ποσότητες αζώτου στο έδαφος, βοηθώντας έτσι στη βελτίωση της οργανικής ουσίας του εδάφους. Επομένως χρειάζεται λιγότερη αζωτούχος λίπανση με αποτέλεσμα την εξοικονόμηση χρημάτων αλλά και την προστασία του περιβάλλοντος. Για τον λόγο αυτό η καλλιέργεια προτιμάται από τους παραγωγούς για αμειψισπορά σε εναλλαγές με σιτηρά καθώς διακόπτει τον κύκλο αναπαραγωγής ασθενειών και εχθρών (Cahota et al. 2019). Το *Fusarium* , *Rizoctonia* και *Pithium* είναι από τις κυριότερες ασθένειες εδάφους που προσβάλλουν την φακή. Προκαλούν σήψεις ριζών αλλά και βλαστών. Εκτός από μύκητες μπορεί να προσβληθούν και από ιούς. Σε όλες τις χώρες καλλιέργειας τη φακής τα φυτά προσβάλλονται από το *Fusarium oxysporum*. Η ασθένεια εμφανίζεται με τα καχεκτικά φύλλα από την κορυφή του φυτού και στη συνέχεια εξαπλώνεται προς τη βάση του. Το φυτό μπορεί να προσβληθεί τόσο σε νεαρή ηλικία αλλά και όταν είναι ανεπτυγμένο κατά τη διάρκεια της αναπαραγωγικής περιόδου. Στη συνέχεια τα φύλλα συρρικνώνονται επομένως το φυτό χάνει την φωτοσυνθετική του λειτουργία και κατά συνέπεια οδηγείται στον μαρασμό. Στην περίπτωση προσβολής των φυτών κατά τη διάρκεια της άνθισης τότε μειώνεται η ποσότητα του παραγόμενου σπόρου, ενώ αν προσβληθεί κατά τη διάρκεια γεμίσματος των λοβών τότε η στρεμματική απόδοση είναι σημαντικά μικρότερη (Nleya et al. 2004). Τέλος εφαρμόζονται πολλά βελτιωτικά προγράμματα με σκοπό της αύξησης της στρεμματικής απόδοσης αλλά και την αντοχή των φυτών σε εχθρούς και ασθένειες (Cahota et al. 2019).

1.3 Μεταβολίτες

Διακρίνονται σε πρωτογενείς και δευτερογενείς. Οι πρωτογενείς μεταβολίτες είναι παρόμοιοί ή ίδιοι σε όλα σχεδόν τα φυτικά είδη. Συντίθενται κατά τη διάρκεια ανάπτυξης του φυτού. Ευθύνονται για την αύξηση, ανάπτυξη αλλά και την αναπαραγωγή του φυτού. Συναντώνται ως μακρομόρια και μέσω σημάτων δίνουν εντολή ενεργοποίησης των μηχανισμών άμυνας και αναγνώρισης του παθογόνου. Μπορεί να είναι υδατάνθρακες , πρωτεΐνες ή λιπίδια αλλά εκείνοι που μπορούν να προσφέρουν ανθεκτικότητα έναντι μυκήτων, ιών και βακτηρίων είναι οι πρωτεΐνες και οι πολυσακχαρίτες (Zaynab et al. 2019). Οι δευτερογενείς μεταβολίτες είναι ενώσεις υπεύθυνες για το χρώμα αλλά και το άρωμα του φυτού. Είναι διαφορετικοί μεταξύ των οργάνων του ίδιου φυτού, στα διάφορα στάδια ανάπτυξης του φυτού, εντός του ίδιου αλλά και διαφορετικών φυτικών ειδών. Αποτελούν διαμεσολαβητές για την ανταπόκριση του φυτού στις βιοτικές και αβιοτικές καταπονήσεις. Η συγκέντρωσή τους εντός του φυτικού κυττάρου εξαρτάται από τους εξής παράγοντες : την βιοσυνθετική οδό που έχουν ακολουθήσει για να παραχθούν, τις περιβαλλοντικές συνθήκες και τις βιοτικές καταπονήσεις κάτω από τις οποίες αναπτύσσεται ο φυτικός οργανισμός και τέλος από γενετικούς παράγοντες. Στους δευτερογενείς μεταβολίτες ανήκουν τα τερπένια, οι φαινολικές και οι αζωτούχες ενώσεις (Ashraf et al. 2018).

1.4 Πειραματικά σχέδια

Η διαδικασία κατά την οποία ένα πείραμα πρέπει να σχεδιαστεί κατάλληλα ώστε ο εκάστοτε ερευνητής να αντλήσει τον απαραίτητο τύπο και όγκο δεδομένων προκειμένου να είναι σε θέση να δώσει απαντήσεις στα ερωτήματα που τέθηκαν εξαρχής ονομάζεται πειραματικό σχέδιο. Η πρώτη αναφορά στο σύγχρονο πειραματικό σχεδιασμό έγινε απο τον Ronald A. Fisher το 1935 στο βιβλίο που δημοσίευσε με τίτλο The Design of Experiments (Ramachandran et al. 2015). Στην αρχή κάθε πειράματος πρέπει να καθορίζεται ο πληθυσμός που θα μελετηθεί όπως επίσης και το μέγεθος αυτού, η μηδενική υπόθεση H_0, η

οποία υποδηλώνει ότι δεν υπάρχει στατιστικά σημαντική διαφορά ως προς ένα χαρακτηριστικό του υπό μελέτη πληθυσμού. Ένα παράδειγμα μηδενικής υπόθεσης θα ήταν ότι δεν υπάρχει στατιστικά σημαντική διαφορά στο ύψος μεταξύ της F1 και της F2 γενεάς σε φυτά τομάτας. Κριτήριο απόρριψης ή όχι της H_0 είναι η τιμή p-value η οποία σε επίπεδο σημαντικότητας α. Για επίπεδο σημαντικότητας α = 5% και p-value < 0.05 απορρίπτεται η H_0 καθώς υπάρχει στατιστικά σημαντική διαφορά μεταξύ των δειγμάτων. Στην περίπτωση που το p-value ≥ 0.05 τότε δεν έχουμε αρκετά στοιχεία έτσι ώστε να απορρίψουμε την H_0. Το επίπεδο σημαντικότητας α καθορίζεται από τον κάθε ερευνητή και οι τιμές που παίρνει συνήθως είναι 0.01 ή 0.05, στο πείραμα που πραγματοποιήθηκε το α=0.05. Τέλος πρέπει να καθορίζεται ο στόχος του πειράματος και να διατυπώνονται με σαφήνεια τα ερωτήματα που πρέπει να απαντηθούν. Στόχος ενός πειράματος μπορεί να αποτελεί η εύρεση της σχέσης με την οποία συνδέονται οι παράγοντες που υπεισέρχονται στην πειραματική διαδικασία, σύγκριση της αλληλεπίδρασης των παραγόντων αλλά και αν η αρχική υπόθεση που διατυπώθηκε ήταν σωστή ή λανθασμένη. Το πειραματικό σχέδιο που είναι καταλληλότερο και τελικά αυτό που θα εφαρμοστεί καθορίζεται από την ομοιογένεια των πειραματικών τεμαχίων, τους υπό μελέτη παράγοντες και τις αλληλεπιδράσεις που μπορεί να υπάρχουν μεταξύ τους. Σημαντικό κομμάτι στην εκτέλεση ενός πειράματος μετά την επιλογή του πειραματικού σχεδίου είναι το κομμάτι της τυχαιοποίησης. Μέσω αυτής της διαδικασίας υπάρχει ίση πιθανότητα ένα πειραματικό τεμάχιο να λάβει μια μεταχείριση. Επίσης ελέγχεται η επίδραση των συχγυτικών παραγόντων πάνω στους υπό μελέτη παράγοντες. Η τυχαιοποίηση μπορεί να πραγματοποιηθεί με αρκετούς τρόπους μερικοί εκ των οποίων είναι : χαρτάκια με τις μεταχειρίσεις που θα εφαρμοστούν στα αγροτεμάχια, τυχαιοποίηση κατά block και τέλος με τη βοήθεια κάποιου προγράμματος όπως για παράδειγμα η R. Ακολουθεί αναλυτική περιγραφή κάποιων συνηθισμένων στη γεωπονία πειραματικών σχεδίων και η διαδικασία τυχαιοποίησης και πιο σχέδιο τελικά είναι καταλληλότερο για το πείραμα που πραγματοποιήθηκε.

Πλήρως Τυχαιοποιημένο Σχέδιο (CRD)

Χρησιμοποιείται όταν υπάρχει μόνο ένας παράγοντας υπό μελέτη και ο πειραματικός αγρός είναι ομοιογενής. Ο πειραματικός αγρός διαιρείται σε πειραματικά τεμάχια, πραγματοποιείται τυχαιοποίηση των μεταχειρίσεων εντός των αγροτεμαχίων έτσι ώστε κάθε πειραματικό τεμάχιο να έχει μια μεταχείριση. Η ανάλυση που πραγματοποιείται είναι one-way ANOVA. Οι υπό μελέτη παράγοντες είναι δυο άρα το CRD απορρίπτεται (Dodge, 2008).

Σχέδιο Τυχαιοποιημένων Πλήρων Ομάδων (RCBD)

Εφαρμόζεται όταν μελετάται μόνο ένας παράγοντας και ο αγρός είναι ετερογενής. Κοντινά σημεία του πειραματικού αγρού έχουν κοινές ιδιότητες οπότε κατατάσσονται στο ίδιο block. Η διαδικασία της τυχαιοποίησης γίνεται εντός των block. Ο αριθμός των block είναι ίσος με τον αριθμό των επαναλήψεων (Dodge, 2008).

Λατινικό Τετράγωνο (Latin Square)

Εφαρμόζεται όταν υπάρχει ένα παράγοντας ενδιαφέροντος και δυο συγχυτικοί. ο ένας εκ των δύο συγχυτικών παραγόντων βρίσκεται κατά γραμμή και ο άλλος κατά στήλη. Με t αριθμό μεταχειρίσεων, η κάθε μια επαναλαμβάνεται t φορές έτσι ώστε κάθε μεταχείριση να εφαρμόζεται μια φορά κατά στήλη και γραμμή (Dodge, 2008).

Σχέδιο Υποδιαιρεμένων Τεμαχίων (Split-Plot)

Εφαρμόζεται όταν υπάρχουν δυο παράγοντες μελέτης ένας με n_1 επίπεδα και ένας με n_2, ο ένας εκ των οποίων είναι δύσκολο να τυχαιοποιηθεί πλήρως εντός των πειραματικών τεμαχίων (whole plots). Έτσι ο παράγοντας αυτός τυχαιοποιείται μόνο μεταξύ των whole plots. Κάθε whole plot χωρίζεται στη μέση (split plots) και μεταξύ των split plots γίνεται η τυχαιοποίηση του δεύτερου παράγοντα που μπορεί να τυχαιοποιηθεί εύκολα. Είναι ένα ιεραρχικό πειραματικό σχέδιο όπου θυσιάζουμε την ακρίβεια του δύσκολου παράγοντα για να έχουμε περισσότερη ακρίβεια στον εύκολο (Dodge, 2008).

Πλήρως Παραγοντικό Σχέδιο (Full Factorial Design)

Κύρια εφαρμογή του στις βιολογικές επιστήμες. Τα πειράματα είναι σχεδιασμένα έτσι ώστε να προσδιορίζουν κάθε φορά την επίδραση ενός

παράγοντα σε κάθε εκτέλεση τους. Το παραγοντικό πειραματικό σχέδιο δίνει την δυνατότητα στον εκάστοτε ερευνητή να μελετήσει την επίδραση πολλαπλών παραγόντων σε μια πειραματική εφαρμογή. Κατά συνέπεια μειώνεται ο αριθμός των πειραμάτων που θα εκτελούνταν σε άλλη περίπτωση, κερδίζεται χρόνος και πόροι. Μελετώνται όλοι οι πιθανοί συνδυασμοί των επιπέδων για όλους τους παράγοντες. Στόχος του σχεδίου είναι να βρεθεί η συνεισφορά του κάθε παράγοντα ξεχωριστά αλλά και η αλληλεπίδραση μεταξύ των παραγόντων. Το παραγοντικό σχέδιο είναι πιο αποτελεσματικό όταν υπάρχει μεγάλη και στατιστικά σημαντική αλληλεπίδραση μεταξύ των παραγόντων αλλά και η συνεισφορά του κάθε παράγοντα είναι εξίσου στατιστικά σημαντική (Woolf, 2021). Το παραγοντικό σχέδιο προσεγγίζει περισσότερο την περιγραφή του πειράματος και έτσι επιλέχθηκε για να υλοποιηθεί. Για k παράγοντες με δυο επίπεδα μελέτης ο καθένας πραγματοποιούνται 2^k πειράματα. Τα 2^k πειραματικά σχέδια χρησιμοποιούνται κυρίως στα αρχικά στάδια της πειραματικής διαδικασίας, ειδικά όταν ο αριθμός των παραγόντων που υπεισέρχονται είναι μικρότερος ή ίσος του τέσσερα. Οι παράγοντες συμβολίζονται με ένα κεφαλαίο ελληνικό γράμμα (π.χ. Α) και τα επίπεδα με ένα μικρό λατινικό γράμμα (π.χ i, k).

- A = ποικιλία και επίπεδα :
  - ✓ k1 = ILL – 6031(ποικιλία 7 , ευαίσθητη στο φουζάριο)
  - ✓ k2 = ILL – 590 (ποικιλία 8 , ανθεκτική στο φουζάριο)
- B = μεταχειρίσεις που εφαρμόστηκαν και επίπεδα :
  - ✓ i1 = Μάρτυρας
  - ✓ i2 = Μολυσμένη

Η στατιστική ανάλυση για το παραγοντικό πειραματικό σχέδιο είναι η two-way ANOVA και αναλύεται εκτενέστερα στο κεφάλαιο 2 Υλικά & Μέθοδοι.

1.5 Διερευνητική (EFA) & Επιβεβαιωτική Παραγοντική Ανάλυση(CFA)

Υπάρχουν δύο τεχνικές στατιστικής ανάλυσης η διερευνητική παραγοντική ανάλυση (EFA) και η επιβεβαιωτική παραγοντική ανάλυση (CFA). Η διερευνητική ανάλυση πραγματοποιεί διερεύνηση των δεδομένων και των μοτίβων που μπορεί να τα συνδέουν μεταξύ τους. Αντίθετα στην επιβεβαιωτική ανάλυση ο ερευνητής μπορεί να καθορίσει των αριθμό των παραγόντων που είναι αντιπροσωπευτικός για τα δεδομένα όπως επίσης να ορίσει και ποια εξαρτημένη μεταβλητή αντιστοιχεί σε ποιόν παράγοντα (Statistics Solutions, 2013). Για την ανάλυση χρησιμοποιήθηκε η EFA καθώς στόχος της πειραματικής διαδικασίας είναι να βρεθεί με ποιον τρόπο αλληλοεπιδρούν οι μεταβολίτες και μπορούν να προσδώσουν στη φακή την ιδιότητα της ανθεκτικότητας έναντι του *Fusaruim oxysporum* f.sp lentis. Η διερευνητική στατιστική διακρίνεται στην μόνο-μεταβλητή (univariate) και την πόλυ-μεταβλητή (multivariate) μέθοδο. Η μόνο-μεταβλητή μέθοδος εξετάζει κάθε εξαρτώμενη μεταβλητή μεμονωμένα και ποια είναι η επίδραση της καθεμίας στο σετ δεδομένων. Η πόλυ-μεταβλητή (multivariate) μέθοδος συγκρίνει ταυτόχρονα όλες τις μεταβλητές και υπολογίζει την επίδραση τους στο σετ δεδομένων (Shah, 2021). Τα αποτελέσματα της μεθόδου για να γίνουν κατανοητά παρουσιάζονται με τη μορφή διαγραμμάτων καθώς είναι πιο εύκολο να παρουσιαστεί ένας τόσο μεγάλος όγκος δεδομένων. Στο πακέτο της R {mixOmics}, που χρησιμοποιήθηκε για την multivariate ανάλυση των δεδομένων οι συναρτήσεις *plotIndiv()*, *plotVar()*, *cim()* και *plotloading()* οδήγησαν στην παραγωγή των αντίστοιχων γραφημάτων. Στο κεφάλαιο 2 ακολουθεί αναλυτική περιγραφή των παραπάνω συναρτήσεων.

1.6 Στατιστικά Πακέτα

Πακέτο {mixOmics}

Το πακέτο {mixOmics} επιλέχθηκε για να υλοποιηθεί η πόλυ-μεταβλητή ανάλυση. Μέσω της πόλυ-μεταβλητής ανάλυσης μπορεί να γίνει διαχείριση μεγάλου όγκου δεδομένων ιδιαίτερα όταν ο αριθμός των μεταβλητών του σετ δεδομένων (μεταβολίτες) είναι κατά πολύ μεγαλύτερος από τον αριθμό των

δειγμάτων. Μειώνει το μέγεθος των δεδομένων δημιουργώντας όλους τους πιθανούς συνδυασμούς μεταξύ τους. Οι συνδυασμοί που προκύπτουν χρησιμοποιούνται για την παραγωγή γραφημάτων μέσω των οποίων θα γίνει κατανοητός ο τρόπος με τον οποίο συνδέονται και πως συσχετίζονται. Είναι το κατάλληλο πακέτο για την ανάλυση μεταβολιτών καθώς παρέχει αρκετές πληροφορίες που βοηθούν στην κατανόηση της λειτουργίας των βιολογικών δομών. Υπάρχουν και άλλα πακέτα μέσω των οποίων μπορεί να γίνει η πολύ-μεταβλητή ανάλυση όπως για παράδειγμα το {lattice} το οποίο δίνει την δυνατότητα παραγωγής γραφημάτων αλλά δεν είναι τόσο εύκολη η παραμετροποίηση τους όσο στο {mixOmics} (Le Cao et al. 2016).

2. Υλικά & Μέθοδοι

2.1 Υλικά

Προκειμένου να πραγματοποιηθεί τεχνητή μόλυνση των φυτών της φακής ακολουθήθηκε η παρακάτω διαδικασία. Αρχικά έγινε απολύμανση των σπόρων φακής με διάλυμα χλωρίνης 10% και ακολούθησε ανάδευση για διάρκεια 5 λεπτών. Ακολούθησε ξέπλυμα των σπόρων με απεσταγμένο απιονισμένο νερό. Η διαδικασία επαναλήφθηκε άλλες τρείς φορές. Έπειτα έγινε τοποθέτηση των σπόρων σε τριβλία τύπου Petri εντός των οποίων είχε τοποθετηθεί διηθητικό χαρτί εμποτισμένο με απεσταγμένο απιονισμένο νερό. Οι σπόροι από τα τριβλία που βλάστησαν τοποθετήθηκαν σε πλαστικούς περιέκτες εντός των οποίων τοποθετήθηκε αποστειρωμένος περλίτης. Οι περιέκτες τοποθετήθηκαν σε θάλαμο ελεγχόμενων συνθηκών με θερμοκρασία 26±2˚C και διάρκεια φωτοπεριόδου 12 ώρες. Παρέμειναν στο θάλαμο μέχρι και την εμφάνιση των 4 πρώτων φύλλων. Για κάθε πιθανό συνδυασμό ποικιλίας και μεταχείρισης υπήρχαν 16 φυτά, συνολικά 64 (16 για την ILL-6031 μάρτυρα, 16 για την ILL-6031 μολυσμένη με φουζάριο, 16 για την ILL-590 μάρτυρα και 16 για την ILL-590 μολυσμένη με φουζάριο). Η τεχνητή μόλυνση έγινε με την αποικία 2.1 που απομονώθηκε από τον αγρό της Φλώρινας. Για να διαπιστωθεί η ικανότητα προσβολής του μύκητα της αποικίας πραγματοποιήθηκε μόλυνση των φυταρίων όταν αυτά είχαν 4 φύλλα. Έπειτα τοποθετήθηκαν στον θάλαμο

ελεγχόμενων συνθηκών. Με την επιτυχημένη προσβολή των φυταρίων από το μύκητα, έγινε καλλιέργεια της απομόνωσης 2.1 σε τριβλία τύπου Petri με υπόστρωμα PDA και σε δοκιμαστικούς σωλήνες. Η καλλιέργεια στα τριβλία έγινε με τη μέθοδο του streaking. Τα τριβλία και οι δοκιμαστικοί σωλήνες παρέμειναν στο θάλαμο ανάπτυξης για 5 ημέρες. Με την προσθήκη 5 mL απεσταγμένου απιονισμένου νερού πραγματοποιήθηκε διήθηση του εναιωρήματος με αποστειρωμένο τουλπάνι. Με το αιματοκυτταρόμετρο ρυθμίστηκε η επιθυμητή συγκέντρωση κονιδίων στα 10^5 κονίδια/mL. Η τεχνητή μόλυνση πραγματοποιήθηκε σε φυτάρια φακής όταν είχαν 4 φύλλα , περίπου 12-14 ημέρες μετά το φύτρωμα, με ριζοπότισμα σε εναιώρημα κονιδίων. Οι ρίζες των φυταρίων παρέμειναν στο εναιώρημα κονιδίων για 10 λεπτά, μεταφέρθηκαν σε πλαστικούς περιέκτες με αποστειρωμένο περλίτη και ριζοποτίστηκαν με 15 mL διαλύματος Hoagland. Τα φυτάρια μεταφέρθηκαν στο θάλαμο ελεγχόμενων συνθηκών για 14 ημέρες ενώ κατά τη διάρκεια παραμονής τους έγιναν και δυο ποτίσματα με απεσταγμένο απιονισμένο νερό. Ως μάρτυρες χρησιμοποιήθηκαν φυτάρια που η ρίζα εμβαπτίστηκε σε απεσταγμένο απιονισμένο νερό. Τα δείγματα φύλλων και ρίζα που συλλέχθηκαν τοποθετήθηκαν σε βαθεία κατάψυξη στους -80 ◦C για να μην υπάρξουν αλλοιώσεις.

2.2 Μέθοδοι Στατιστικής ανάλυσης

Για την ανάλυση του συνόλου των δεδομένων αρχικά υλοποιείται μόνο-μεταβλητή ανάλυση όπου εξετάζονται ξεχωριστά οι τιμές κάθε μεταβολίτη στα διάφορα επίπεδα των παραγόντων. Μετά υλοποιείται πόλυ-μεταβλητή ανάλυση και ελέγχεται ποιοι μεταβολίτες συνεισφέρουν περισσότερο στο διαχωρισμό των δειγμάτων. Οι μεταβολίτες που επαληθεύονται και από τις δύο αναλύσεις, θεωρούνται υποψήφιοι βιοδείκτες. Η διαδικασία για την επαλήθευση των βιοδεικτών δεν αναφέρεται καθώς ξεφεύγει από τα πλαίσια της συγκεκριμένης εργασίας. Ένα σημαντικό κομμάτι της συνολικής διαδικασίας της στατιστικής ανάλυσης, που συνήθως αποτελεί μία χρονοβόρα διαδικασία, είναι η προετοιμασία των δεδομένων και η μετατροπή τους σε κατάλληλη

μορφή προς επεξεργασία, αν χρειαστεί. Παρακάτω αναλύεται το φιλτράρισμα τον δεδομένων πριν την υλοποίηση της στατιστικής ανάλυσης καθώς και η μεταφορά του excel αρχείου δεδομένων σε πλαίσιο δεδομένων (data frame) της R.

Φιλτράρισμα Δεδομένων (Data Cleaning)

Ξεκινώντας η univariate ανάλυση περιλαμβάνει την εισαγωγή των δεδομένων στην R. Αρχικά με την *setwd()* συνάρτηση ορίζετε η θέση του αρχείου στον υπολογιστή. Οι μεταβολίτες και οι συγκεντρώσεις τους στις μεταχειρίσεις που εφαρμόστηκαν αποθηκεύτηκαν σε ένα αρχείο excel. Με την *read.table()* συνάρτηση γίνεται εισαγωγή των δεδομένων του excel στο περιβάλλον της R. Επίσης μπορεί να χρησιμοποιηθεί και η συνάρτηση *read.xlsx().* Ανάλογα με τον τύπο του αρχείου που έχουν αποθηκευτεί τα αρχικά δεδομένα χρησιμοποιείται και η αντίστοιχη συνάρτηση. Στην περίπτωση του .csv αρχείου χρησιμοποιείται η συνάρτηση *read.csv()*. Έτσι έχει δημιουργηθεί ένα πλαίσιο δεδομένων (dataframe) το οποίο αποθηκεύεται στην μεταβλητή data_1. Η αρχική βάση δεδομένων σε excel απεικονίζει τους μεταβολίτες ως γραμμές και τις παρατηρήσεις ως στήλες (Εικόνα 3). Για να μετατραπεί η συγκεκριμένη μορφή (wide format) στη βασική μορφή της στατιστικής επεξεργασίας, έγινε αναστροφή του πίνακα δεδομένων (Εικόνα 4) για να γίνει η βάση σε κατάλληλη μορφή (long format) επεξεργασίας από την R. Η συνάρτηση που χρησιμοποιήθηκε ήταν η *t()* όπου εκτελεί την αναστροφή του πλαισίου δεδομένων. Ως μοναδικό όρισμα στη συνάρτηση εισάγεται το πλαίσιο δεδομένων.

```
22  #kanw transport ta dataframes vlastou kai rizas
23  t_stem_data<- as.data.frame(t(stem_data))
24  t_root_data<- as.data.frame(t(root_data))
```

Εικόνα 2. Συνάρτηση που πραγματοποιεί αναστροφή δεδομένων

Με την διαδικασία της αναστροφής οι γραμμές του πίνακα έγιναν στήλες και οι στήλες γραμμές.

	A	B	C	D	E	F	G
1	**Ετικέτες γραμμής**			**ESC**	**ESC**	**ESC**	**ESC**
2	[101713] glucoheptonic acid 1 [18.985]			0	0	0	0
3	[10712] cellobiose 2 [24.7]			8.905218137	7.51849735	2.263507184	4.195851483
4	[111064] ribonic acid-gamma-lactone 2 [15.154]			0	0	0	0
5	[138] 5-aminovaleric acid 1 [14.458] +5-aminovaleric acid 2 [14.955]			0	0	0	0
6	[614] L-proline 1 [8.58]+ L-proline 1 [8.567]			0	0	0	0
7	[145742] L-proline 2 [10.321]+L-proline 2 [10.341]			3.938060629	28.95598309	155.9620027	251.7913153
8	[16219560] lactobionic acid 3 [25.406]			2.153226823	16.39058945	718.3138774	776.8441404
9	[18950] D-mannose 1 [17.287]			0	3.744231117	14.48132324	11.95622373
10	[192826] maltotriose 2 [30.668]			3.010102078	117.113912	4.80577563	65.49729188
11	[204] allantoin 2 [17.356]			0	1.532241966	0	0.719115369
12	[204] allantoin 3 [19.167]			0	0	0	0
13	[206] D (+) galactose 2 [17.662]			0.295481634	0	0.495038259	0
14	[21236] L-norleucine 1 [8.945]			1.32280749	1.180965346	0.406960522	5.543671144
15	[236] L-asparagine 1 [14.496]			17.70682603	57.70506905	11.78302594	55.43836836
16	[24749] D-glucose 1 [17.426]			0	0	0	0
17	[268779] N-methylglutamic acid 4 [15.174]			0	17.83186529	0	0
18	[2724552] tagatose 1 [17.011]			0.826017286	0	0	0
19	[3034828] palatinitol 2 [26.01]			0.24047719	1.586860119	5.908433972	11.08210602

Εικόνα 3. Δεδομένα στο format του excel

	[101713] glucoheptonic acid 1 [18.985]	[10712] cellobiose 2 [24.7]	[111064] ribonic acid-gamma-lactone 2 [15.154]	[138] 5-aminovaleric acid 1 [14.458] +5-aminovaleric acid 2 [14.955]	[614] L-proline 1 [8.58]+ L-proline 1 [8.567]	[145742] L-proline 2 [10.321]+L-proline 2 [10.341]	[16219560] lactobionic acid 3 [25.406]
ESC	0.000000	8.905218	0.00000000	0.0000000	0.0000000	3.9380606	2.1532268
ESC.1	0.000000	7.518497	0.00000000	0.0000000	0.0000000	28.9559831	16.3905895
ESC.2	0.000000	2.263507	0.00000000	0.0000000	0.0000000	155.9620027	718.3138774
ESC.3	0.000000	4.195851	0.00000000	0.0000000	0.0000000	251.7913153	776.8441404
ESM	0.000000	10.238351	0.00000000	0.0000000	0.0000000	419.6915919	31.2485567
ESM.1	0.000000	3.129988	0.00000000	0.0000000	1.9458773	152.9965187	115.6624603
ESM.2	0.000000	2.433592	0.00000000	0.0000000	0.0000000	112.2751701	111.7756239
ESM.3	0.000000	3.387618	0.00000000	0.0000000	0.3139047	266.0117440	211.4886852

Εικόνα 4. Δεδομένα μετά την αναστροφή του πίνακα στο περιβάλλον RStudio

Για να έχει νόημα η στατιστική ανάλυση τέθηκε το κριτήριο του 80% για τους μεταβολίτες. Από τις 16 παρατηρήσεις, ο αριθμός των μηδενικών τιμών δε θα μπορούσε να ξεπεράσει το 80%, δηλαδή τα 13 κατά στρογγυλοποίηση μηδενικά. Εφόσον κρατήθηκαν οι μεταβολίτες που έχουν συνεισφορά στην στατιστική ανάλυση, έγινε κανονικοποίηση των δεδομένων με την συνάρτηση *scale()* η οποία έχει μοναδικό όρισμα το πλαίσιο δεδομένων του οποίου τα δεδομένα πρέπει να κανονικοποιηθούν (Εικόνα 5) .

```
48  #kanw kanonikopoihsh(scale) ta dedomena mou
49  scaled_df_stem <- scale(unscaled_stem_data)
50  scaled_df_root <- scale(unscaled_root_data)
```

Εικόνα 5. Συνάρτηση με την οποία πραγματοποιείται κανονικοποίηση των δεδομένων

2.2.1 Μόνο-μεταβλητή Ανάλυση (Univariate Analysis)

Όπως αναφέρθηκε προηγουμένως, αρχικά αναλύθηκε κάθε μεταβολίτης ξεχωριστά ως προς τις τιμές του σε κάθε επίπεδο των παραγόντων «Ποικιλία» και «Μεταχείριση». Με αυτό τον τρόπο εκτελέστηκε ένας αρχικός έλεγχος για την εύρεση σημαντικών μεταβολιτών. Η στατιστική ανάλυση που χρησιμοποιήθηκε βασίζεται στην Ανάλυση Διακύμανσης (ANOVA) για δύο παράγοντες ενδιαφέροντος.

Two-way ANOVA

Η ANOVA (Analysis Of Variance) είναι ένα στατιστικό τεστ που χρησιμοποιείται για την σύγκριση των μέσων όρων δύο ή περισσότερων επιπέδων παραγόντων ενδιαφέροντος. Η two-way ANOVA εκτιμά πως μεταβάλλεται η ποσοτική εξαρτώμενη μεταβλητή σε σχέση με τα επίπεδα των ανεξάρτητων μεταβλητών. Στο πείραμα που πραγματοποιήθηκε η εξαρτημένη ποσοτική μεταβλητή είναι οι μεταβολίτες και οι ανεξάρτητες ποιοτικές μεταβλητές η ποικιλία και οι μεταχειρίσεις (επεμβάσεις) που εφαρμόστηκαν. Η μεταβλητή ποικιλία και μεταχειρίσεις έχουν δύο επίπεδα η κάθε μία. Η two-way ANOVA εφαρμόζεται στην περίπτωση που έχουμε δύο ανεξάρτητες μεταβλητές και έχει γίνει συλλογή δεδομένων της ποσοτικής εξαρτώμενης μεταβλητής για κάθε επίπεδο των ανεξάρτητων παραγοντικών μεταβλητών. Ωστόσο θα πρέπει και οι δυο ανεξάρτητες μεταβλητές να είναι ποιοτικές. Η two-way ANOVA με αλληλεπίδραση ελέγχει ταυτόχρονα τις αρχικές υποθέσεις (Bevans, 2021) :

- Οι μέσοι όροι των γκρουπ δεν διαφέρουν σε κανένα επίπεδο της πρώτης ανεξάρτητης μεταβλητής
- Οι μέσοι όροι των γκρουπ δεν διαφέρουν σε κανένα επίπεδο της δεύτερης ανεξάρτητης μεταβλητής

- Η επίδραση της πρώτης ανεξάρτητης μεταβλητής δεν σχετίζεται με την επίδραση της δεύτερης ανεξάρτητης μεταβλητής.

Για να μπορέσει ωστόσο να εφαρμοστεί θα πρέπει να ισχύουν οι εξής προϋποθέσεις :

- ✓ Ομοιογένεια μεταξύ των διακυμάνσεων
- ✓ Ανεξάρτητες παρατηρήσεις
- ✓ Η εξαρτώμενη μεταβλητή να ακολουθεί την κανονική κατανομή

Η παραγωγή του πίνακα ANOVA έγινε με χρήση της συνάρτησης *aov()*. Πρώτο όρισμα της συνάρτησης είναι η εξαρτημένη μεταβλητή (μεταβολίτες), οι παράγοντες που μας ενδιαφέρουν (ποικιλία και επεμβάσεις που εφαρμόστηκαν) αλλά και η αλληλεπίδραση μεταξύ των παραγόντων και δεύτερο όρισμα είναι το πλαίσιο δεδομένων (dataframe) που αντιστοιχεί στην εξαρτημένη μεταβλητή.

```
>
> p_value_stem_Poikilies[i] <- unlist(summary(aov(df_stem[ ,i]~ Poikilies +
  Epemvaseis + Poikilies*Epemvaseis, data = df_stem)))[17]
```

Εικόνα 6. Συνάρτηση για την παραγωγή του πίνακα ANOVA

Στην Εικόνα 6 παρουσιάζεται η συνάρτηση *aov()* με τα ορίσματα της, η *summary()* με την οποία εμφανίζεται ο πίνακας της ANOVA και η συνάρτηση *unlist()* με την οποία ο πίνακας μετατρέπεται σε διάνυσμα από το οποίο απομονώθηκε το σύνολο των τιμών p-value για την αλληλεπίδραση μεταξύ της ποικιλίας και των επεμβάσεων που εφαρμόστηκαν. Με κόκκινο επισημαίνεται ο τρόπος με τον οποίο συμβολίζεται η αλληλεπίδραση μεταξύ των παραγόντων. Στον πίνακα της ANOVA υπάρχουν τρεις διαφορετικές τιμές p-value : μια τιμή για την ποικιλία, μία για την επέμβαση (μεταχείριση) που εφαρμόστηκε και μία για την αλληλεπίδραση.

```
> summary(aov(df_root[,1]~ Poikilies + Epemvaseis + Poikilies*Epemvaseis,
+             data = df_root))
                   Df Sum Sq Mean Sq F value  Pr(>F)
Poikilies           1  7.537   7.537  14.035 0.00279 **
Epemvaseis          1  0.006   0.006   0.012 0.91448
Poikilies:Epemvaseis  1  1.012   1.012   1.884 0.19501
Residuals          12  6.444   0.537
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Εικόνα 7. Πίνακας ANOVA για τον μεταβολίτη cellobiose 2

Στην Εικόνα 7 παρουσιάζεται ο πίνακας ANOVA για τον μεταβολίτη cellobiose 2 όπου διακρίνονται οι τιμές p-value της ποικιλίας (0.00279) με στατιστικά σημαντική διαφορά (συμβολίζονται στον πίνακα με δύο αστερίσκους) για τον παράγοντα ποικιλία αλλά όχι για τις επεμβάσεις (0.91448) και την αλληλεπίδραση μεταξύ των παραγόντων (0.19501) οι οποίες δεν είναι στατιστικά σημαντικές. Λόγω των πολλαπλών συγκρίσεων, το σφάλμα τύπου I, δηλαδή η πιθανότητα να απορρίψουμε την αρχική υπόθεση H_0 ενώ δεν θα έπρεπε, αυξάνεται. Για το λόγο αυτό οι τιμές p-value διορθώθηκαν μέσω της *p.adjust()* με ορίσματα το πλαίσιο δεδομένων στο οποίο έχουν αποθηκευτεί οι τιμές p-value του πίνακα ANOVA και την μέθοδο με βάση την οποία έγιναν. Χρησιμοποιήθηκε η “fdr” μέθοδος (Benjamini & Hochberg 1995).

```
>
>
> p_val_adj_root_Poikilies <- p.adjust(p_value_root_Poikilies, method = "fdr")
```

Εικόνα 8. Συνάρτηση p.adjust() με όρισμα την "fdr" μέθοδο

Άλλες μέθοδοι διόρθωσης των τιμών p-value είναι η μέθοδος “Bonferroni” η οποία πολλαπλασιάζει τις τιμές p-value με τον αριθμό των συγκρίσεων που έγιναν και η μέθοδος “holm” από τον Holm (1979) που είναι λιγότερο επεμβατική, ωστόσο η “fdr” είναι η πιο ισχυρή στην περίπτωση του συγκεκριμένου πειράματος. Τέλος για επίπεδο σημαντικότητας α = 5% οι στατιστικά σημαντικοί μεταβολίτες είναι αυτοί με τιμές p-value<0.05 . Επιπλέον για την ερμηνεία του μεγέθους της διαφοράς ανά μεταβολίτη ανά επίπεδο χρησιμοποιείται το πηλίκο των μολυσμένων μεταβολιτών κάθε ποικιλίας προς τον μάρτυρα .

Ρυθμός Σφάλματος Οικογένειας Συγκρίσεων (Family-Wise Error Rate - FWER)

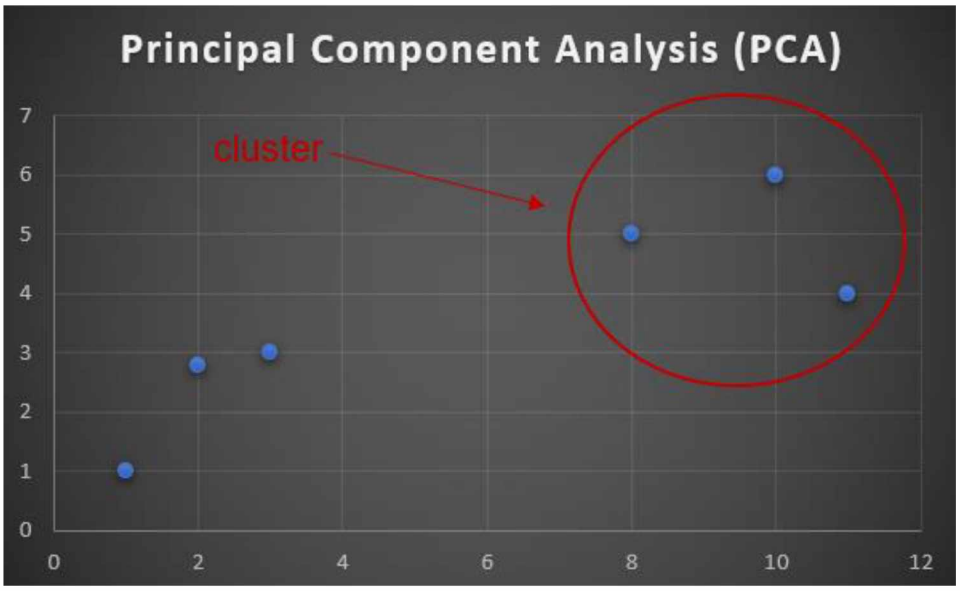
Κατά την ταυτόχρονη μελέτη πολλαπλών αρχικών υποθέσεων είναι αναγκαία η προσαρμογή των πολλαπλών τεστ ώστε να περιοριστεί όσο περισσότερο γίνεται ο αριθμός των λανθασμένων ανακαλύψεων. Τέτοιες διαδικασίες προσαρμογής πολλαπλών συγκρίσεων εφαρμόζονται ευρέως σε ομικά δεδομένα (genomics, metabolomics , proteomics). Κλασσικές μέθοδοι ελέγχου του οικογενειακού ποσοστού σφαλμάτων ή αλλιώς της πιθανότητας εύρεσης τουλάχιστον μιας λάθος ανακάλυψης έχουν αναπτυχθεί και χρησιμοποιηθεί ευρέως για την διόρθωση των πολλαπλών τεστ. Μια από αυτές είναι η διόρθωση των Benjamini & Hochberg (1995) η οποία βασίζεται στην μέθοδο ψευδούς ποσοστού ανακάλυψης (False Discovery Rate - FDR). Η FDR μέθοδος είναι μια νέα πρόταση ως νέο μέτρο στην FWER στον έλεγχο των πολλαπλών τεστ. Έχει μεγάλη ισχύ στην ανίχνευση των σφαλμάτων τύπου I καθώς αυτά μπορούν εύκολα να υπολογιστούν με μια λίστα τιμών p-value. Σε πειράματα βιολογικού ενδιαφέροντος είναι η καταλληλότερη και η πιο χρήσιμη μέθοδος (Korthauer et al. 2019). Για να εφαρμοστεί η μέθοδος, πρώτα διατάσσονται οι τιμές p-value κατά αύξουσα σειρά. Έπειτα θέτονται τάξεις για κάθε p-value. Η μικρότερη τιμή p-value έχει τάξη ένα, η αμέσως μεγαλύτερη τάξη 2 κ.ο.κ. Μετά υπολογίζεται η B-H (Benjamini & Hochberg) κρίσιμη τιμή με τον τύπο $\frac{i}{m} * q$, όπου i = η τάξη κάθε τιμής p-value, m = ο συνολικός αριθμός των τεστ που υλοποιήθηκαν, q = 0.05 (προκαθορισμένο από τη συνάρτηση της R). Τέλος γίνεται σύγκριση των σημαντικοτήτων με αυτές που βρέθηκαν με την βοήθεια του τύπου. Κρατάμε τη μεγαλύτερη τιμή p-value που είναι μικρότερη από την B-H κρίσιμη τιμή.

2.2.2 Πόλυ-μεταβλητή Ανάλυση (Multivariate Analysis)

Για τη πόλυ-μεταβλητή ανάλυση διατίθενται αρκετά εργαλεία. Δύο από αυτά είναι η ανάλυση κύριων συνιστωσών (PCA) που ανήκει στην μη ελεγχόμενη μέθοδο μηχανικής μάθησης (unsupervised machine learning) και η μέθοδος μερικών ελαχίστων τετραγώνων (PLS-DA) που ανήκει στην ελεγχόμενη μέθοδο μηχανικής μάθησης (supervised machine learning).

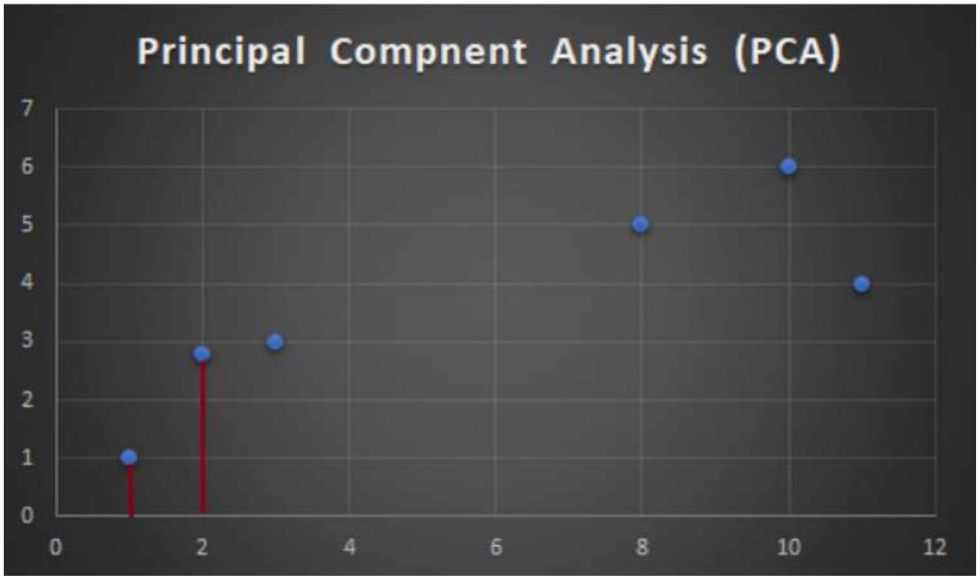
Ανάλυση Κύριων Συνιστωσών (PCA)

Η μη ελεγχόμενη ανάλυση (unsupervised) προβάλει τα δεδομένα σε όσα επίπεδα οριστούν και έπειτα κατηγοριοποιούνται. Μέρος της μη ελεγχόμενης ανάλυσης είναι η ανάλυση κύριων συνιστωσών (PCA). Η ανάλυση κύριων συνιστωσών είναι μια στατιστική διαδικασία στην οποία αναπαρίστανται ένας πίνακας συνδιακύμανσης ενός συνόλου «αρχικών» μεταβλητών μέσα από ένα διαφορετικό σύνολο «νέων» μεταβλητών οι οποίες προκύπτουν από το γραμμικό συνδυασμό των αρχικών μεταβλητών. Η PCA λειτουργεί μειώνοντας το μέγεθος των δεδομένων όπου υπάρχει μεγάλος αριθμός συγγενικών μεταβλητών ενώ επιστρέφει το μεγαλύτερο ποσοστό της παραλλακτικότητας των δεδομένων. Αναλυτικότερα έστω ότι μελετώνται k μεταβλητές σε n αριθμό δειγμάτων. Σε κάθε μεταβλητή παραχωρείται ένας άξονας. Στην περίπτωση τριών μεταβλητών η αναπαράσταση είναι δύσκολη. Η PCA τροφοδοτείται με τέσσερις ή και παραπάνω μεταβλητές και κατασκευάζει ένα δισδιάστατο διάγραμμα. Το διάγραμμα αυτό απεικονίζει τα δείγματα σε συστάδες (clusters). Μέσα από την διαδικασία της PCA αναδεικνύονται οι μεταβλητές που συνεισφέρουν περισσότερο στη δημιουργία συστάδων με τον καλύτερο διαχωρισμό στους άξονες x και y. Τέλος διαπιστώνεται η ακρίβεια του δισδιάστατου γραφήματος δίνοντας τιμές για την παραλλακτικότητα που προκύπτει. Αν για παράδειγμα μελετώνται δύο μεταβλητές, στο καρτεσιανό σύστημα συντεταγμένων αναπαρίστανται οι τιμές των δειγμάτων για τις μεταβλητές. Η μία μεταβλητή αναπαρίσταται στον άξονα x΄x και η δεύτερη μεταβλητή στον άξονα y΄y. Παρόμοια δείγματα δημιουργούν ένα cluster.

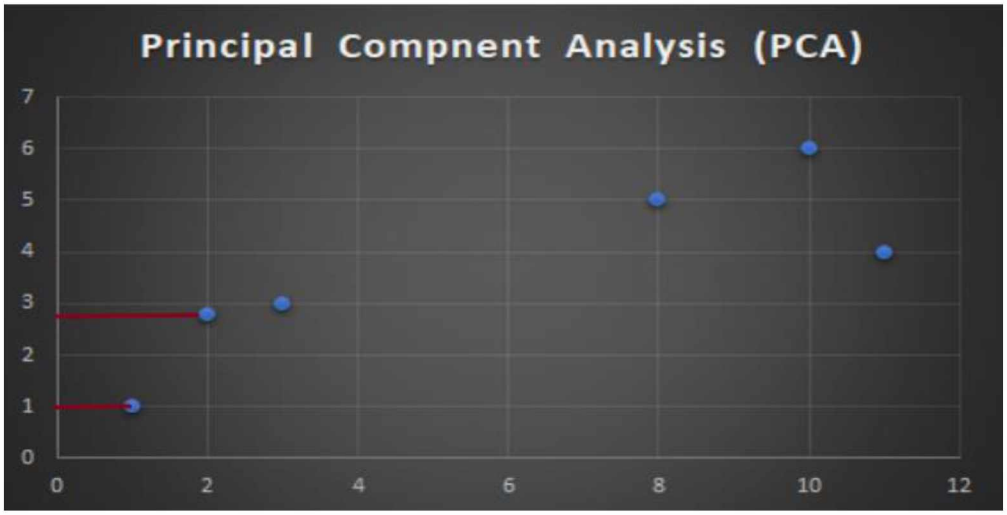


Εικόνα 9. Δείγματα σε cluster

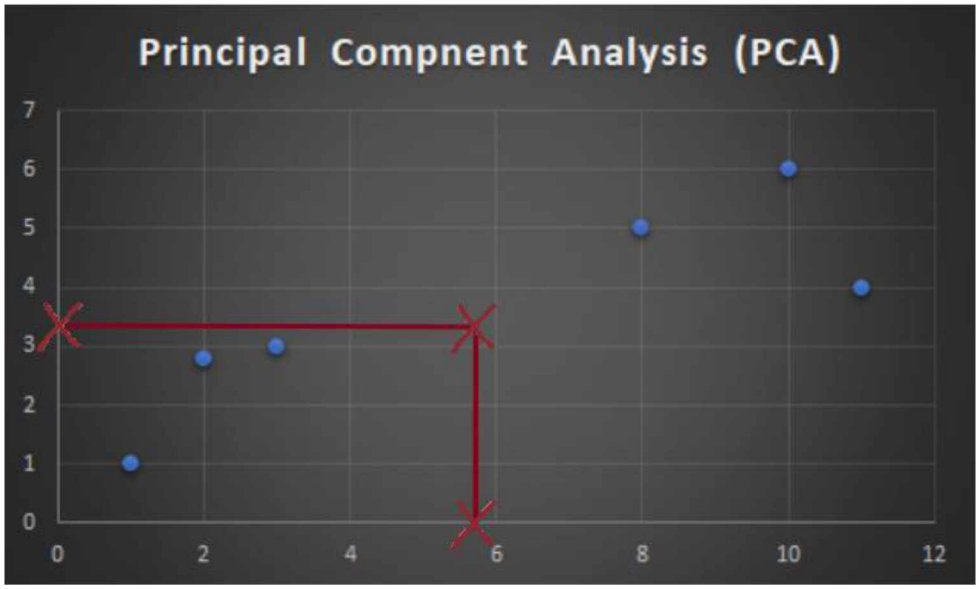
Τα σημεία (οι τιμές των δειγμάτων) προβάλλονται στους άξονες του συστήματος συντεταγμένων και έτσι υπολογίζεται ο μέσος όρος των μετρήσεων.



Εικόνα 10. Προβολή σημείων στον άξονα x΄x

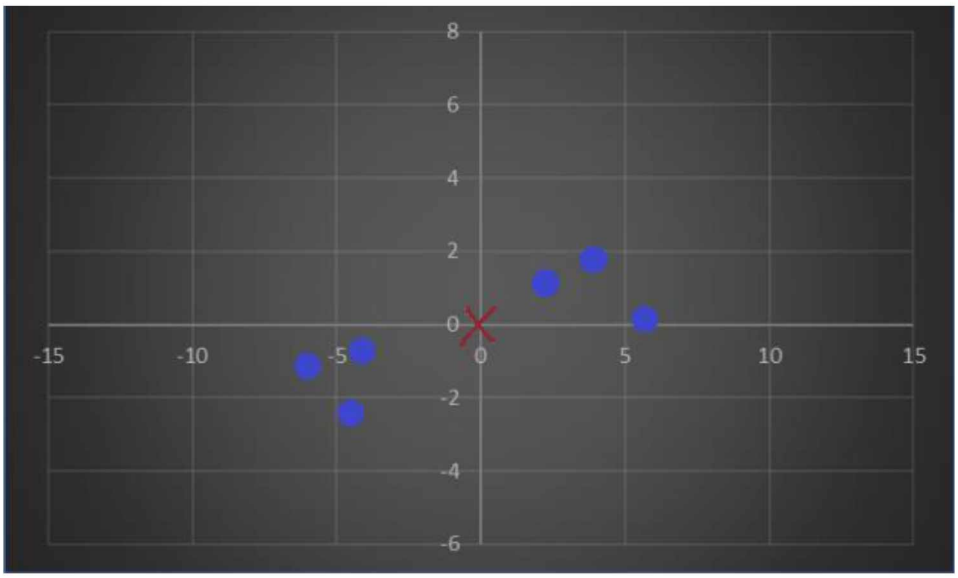


Εικόνα 11. Προβολή σημείων στον άξονα y΄y



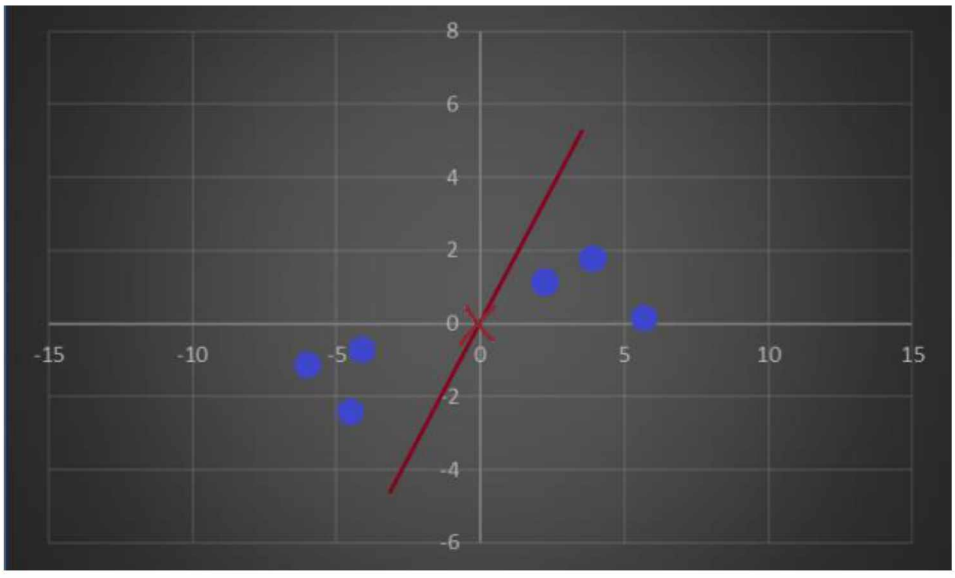
Εικόνα 12. Μέσος όρος μετρήσεων

Γίνεται μετατόπιση των σημείων μέχρις ότου το σημείο που βρίσκεται ο μέσος όρος των μετρήσεων να συμπίπτει με την αρχή των αξόνων Ο(0,0).



Εικόνα 13. Μετατοπισμένα σημεία με κέντρο την αρχή των αξόνων

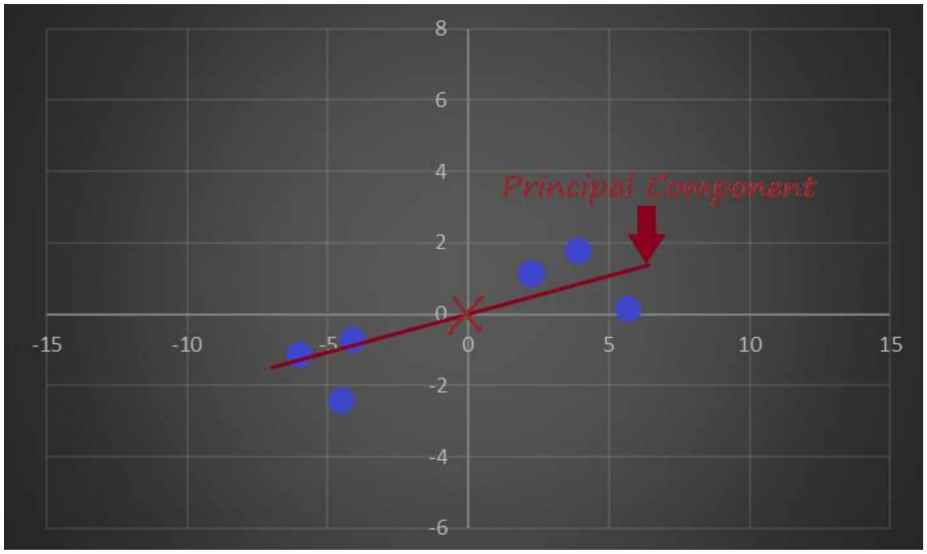
Για να μειωθεί το μέγεθος των δεδομένων θα πρέπει να γίνει απεικόνισή τους σε μία διάσταση από τις δυο διαστάσεις που απεικονίζονταν προηγουμένως. Για να επιτευχθεί αυτό σε μια τυχαία ευθεία που διέρχεται από την αρχή των αξόνων προβάλλονται οι τιμές των δειγμάτων.



Εικόνα 14. Τυχαία ευθεία που διέρχεται από την αρχή των αξόνων

Στη συνέχεια υπολογίζεται η απόσταση που απέχει κάθε τιμή από την αρχή των αξόνων και υψώνεται στο τετράγωνο. Η ίδια διαδικασία επαναλαμβάνεται για τις υπόλοιπες τιμές των δειγμάτων. Έτσι υπολογίζεται το άθροισμα των

τετραγώνων αυτών των αποστάσεων (sum of square distances SS). Η ευθεία που έχει το μεγαλύτερο SS καλείται συστατικό διαχωρισμού (Principal Component) και είναι η καλύτερη δυνατή ευθεία πάνω στην οποία αναπαρίστανται οι περισσότερες δυνατές τιμές των δειγμάτων.



Εικόνα 15. Βέλτιστη ευθεία Principal Component

Με τον τρόπο αυτό τα δεδομένα απεικονίζονται σε μια διάσταση (Principal Component) και κατά συνέπεια έχει μειωθεί και το μέγεθος τους. Τα principal components διατάσσονται κατά τέτοιο τρόπο ώστε τα περισσότερα components να διατηρήσουν όσο περισσότερη από την παραλλακτικότητα που υπάρχει στο αρχικό σύνολο δεδομένων (Jolliffe, 2002).

Διακριτή Ανάλυση Μερικών Ελαχίστων Τετραγώνων (PLS-DA)

Η ελεγχόμενη ανάλυση (supervised) εφαρμόζεται σε κατηγοριοποιημένα δεδομένα, δηλαδή σε παρατηρήσεις που γνωρίζουμε εκ των προτέρων σε ποια κατηγορία του παράγοντα ανήκουν. Εφαρμόζεται κυρίως για περιπτώσεις κατάταξης και επιλογής βιοδεικτών στις μεταβολομικές αναλύσεις καθώς είναι δυνατή η ανάλυση πολλών εξαρτημένων μεταβλητών ταυτόχρονα. Στην έρευνα για πιθανούς βιοδείκτες χρησιμοποιούνται μοντέλα μεταβολομικής ανάλυσης μεταξύ δύο κλάσεων των δειγμάτων. Η PLS-DA αποτελεί την πιο κοινή μέθοδο πόλυ-μεταβλητής διακριτής ανάλυσης μεταξύ των κλάσεων. Λειτουργεί με παρόμοιο τρόπο με αυτό της PCA αλλά η PLS-DA έχει ως στόχο να γίνει ο καλύτερος διαχωρισμός μεταξύ των κλάσεων που υπεισέρχονται στο πείραμα.

Τέλος με τον καλύτερο διαχωρισμό μεταξύ των παραγόντων είναι ευκολότερη η εξαγωγή συμπεράσματος. Στην περίπτωση στατιστικά σημαντικής διάκρισης μεταξύ των κλάσεων, δηλαδή των μολυσμένων φυτών και αυτών που αποτελούσαν τον μάρτυρα, τότε οι παράμετροι του μοντέλου έχουν αρκετά διακριτή ισχύ οπότε είναι πιθανή η εύρεση βιοδεικτών. Στα PLS-DA μοντέλα η σχέση μεταξύ του σετ δεδομένων και της εξαρτώμενης μεταβλητής y αναπτύσσεται κατά τέτοιο τρόπο ώστε η μεταβλητή y να μπορεί να προβλεφθεί και από δείγματα άγνωστης προέλευσης χρησιμοποιώντας μόνο τα δεδομένα των μεταβολιτών. Τέλος χαρακτηριστικό της μεθόδου είναι το γεγονός ότι εκτιμάται η ποιότητα των μοντέλων πρόβλεψης και των διαφορών μεταξύ των κλάσεων (Szymańska et al. 2012). Με το πακέτο "mixOmics" της R και την συνάρτηση *mixOmics ()* επιτυγχάνεται η παραγωγή λίστας με τις κατάλληλες μεταβλητές που χρησιμοποιούνται αργότερα για την παραγωγή γραφημάτων και πινάκων. Καλώντας τις συναρτήσεις του πακέτου που μας ενδιαφέρουν και έχοντας ως κύριο όρισμα τη συγκεκριμένη λίστα που παράχθηκε με τη συνάρτηση *mixOmics()* παράγουμε γραφήματα που αξιολογούν τη συνεισφορά κάθε μεταβολίτη στο διαχωρισμό των δειγμάτων σε ομάδες.

Στην παρούσα εργασία τα δεδομένα γνωρίζουμε εκ των προτέρων σε ποια κατηγορία του παράγοντα ανήκουν και για το λόγο αυτό χρησιμοποιήσαμε την supervised και την PLS-DA μέθοδο.

2.2.3 Γραφήματα Στην Πόλυ-μεταβλητή Ανάλυση

Στο παρόν κεφάλαιο γίνεται λόγος για τα γραφήματα που παράγονται στην πόλυ-μεταβλητή ανάλυση, ο τρόπος με τον οποίο κάθε διάγραμμα αναπαριστά τα δεδομένα, με ποια συνάρτηση παράγονται και τελικά τι προκύπτει από το κάθε ένα διάγραμμα για τα δεδομένα που απεικονίζονται. Στο κεφάλαιο 3 παρουσιάζονται τα διαγράμματα και ο σχολιασμός τους.

Γράφημα Διασποράς Δειγμάτων (Individual Plot)

Στα γραφήματα αυτού του είδους τα δείγματα του πειράματος, δηλαδή οι ποικιλίες μαζί με τις επεμβάσεις που εφαρμόστηκαν, αναπαρίστανται ως σημεία τα οποία είναι τοποθετημένα σύμφωνα με την προβολή τους στο μικρότερο υποσύνολο του διανύσματος των συνιστωσών των πόλυ-μεταβλητών μοντέλων. Με την οπτικοποίηση των αποτελεσμάτων και την διάταξη των δειγμάτων σε clusters γίνονται εμφανείς οι διαφορές αλλά και οι ομοιότητες που υπάρχουν μεταξύ των δειγμάτων. Η παραγωγή του γραφήματος γίνεται μέσω τις συνάρτησης *plotIndiv()*. Με την συνάρτηση *png()* δημιουργείται θέση αποθήκευσης του διαγράμματος που θα παραχθεί με την *plotIndiv()*. Στην εικόνα που ακολουθεί παρουσιάζεται ο κώδικας και για τις δύο συναρτήσεις.

```
png(filename = "plotIndiv_blastos_1.png", width = 2500,
    height = 2500, res = 300)
plotIndiv(blastos, comp.predicted = 2, ind.names = TRUE,
          col.per.group = colors,
          style = "graphics", ellipse = TRUE ,abline = TRUE ,
          legend = TRUE,
          legend.title = "Υπόμνημα",
          title = "Plot for Individuals Blastos")
dev.off()
```

Εικόνα 16. Ορίσματα συναρτήσεων png() και plotIndiv()

Γράφημα Μεταβλητών Κυκλικής Συσχέτισης (Correlation Circle Plot)

Στο γράφημα αυτό αναπαρίστανται είτε οι εξαρτώμενες Y είτε οι ανεξάρτητες X μεταβλητές μέσω της προβολής τους στο επίπεδο, που καθορίζεται από την εξαρτημένη ή την ανεξάρτητη μεταβλητή. Οι προβολές των X και Y βρίσκονται στο εσωτερικό κύκλου ακτίνας 1 με κέντρο του την αρχή O(0,0) του καρτεσιανού συστήματος συντεταγμένων. Όσο πιο μακριά βρίσκονται οι μεταβλητές από την αρχή των αξόνων τόσο ισχυρότερο είναι το γραμμικό μοντέλο που τις συνδέει. Στο γράφημα εμφανίζονται δύο ομόκεντροι κύκλοι ακτίνας 1 και 0.5 για να αποκαλυφθεί το μοντέλο συσχέτισης με το οποίο συνδέονται. Παράγεται μέσω της συνάρτησης *plotVar()*.

```
png(filename = "plotVar_blastos_1.png", width = 2500, height = 2500, res=200)
plotVar(blastos, plot = TRUE, var.names = TRUE,
        abline = TRUE, rad.in = 0.5, title = "Blastos Correlation Circle Plots",
        legend = TRUE,
        style = "ggplot2",
        overlap = FALSE)
dev.off()
```

Εικόνα 17. Ορίσματα της συνάρτησης plotVar()

Θερμικός Χάρτης (Heatmap)

Ο θερμικός χάρτης ή αλλιώς *heatmap* λειτουργεί με ταυτόχρονη ιεραρχική ομαδοποίηση των γραμμών και των στηλών των ομοιοτήτων του πίνακα που είναι στατιστικά σημαντικές. Ο πίνακας ομοιοτήτων προκύπτει από την μέθοδο PLS-DA. Είναι δύο διαστάσεων, κάθε στοιχείο του πίνακα αντιπροσωπεύεται από ένα χρώμα και οι σειρές αλλά και οι στήλες κατατάσσονται με ιεραρχική ομαδοποίηση. Παράγεται με την συνάρτηση *cim()*.

```
png(filename = "heatmap_blastos1.png", width = 2500, height = 2500, res=300)
cim(blastos, row.sideColors = colors[factor(sindiasmoi)],
    row.names = TRUE, col.names = FALSE,
    row.cex = 0.8, col.cex = 0.8, cluster = "both",
    legend = list(legend = unique(sindiasmoi),
                  title = "Ποικιλία & Μεταχείριση",
                  cex = 0.60),
    title = "Βλαστός : Heatmap", keysize = c(0.7,0.7))
dev.off()
```

Εικόνα 18. Ορίσματα της συνάρτησης cim()

Πίνακας VIP(Variable Importance Plot)

Κατασκευάζεται μέσα από το πακέτο {VIP} και την συνάρτηση *vip()* και υπολογίζει τα score/plots για κάθε μεταβλητή μέσα από συγκεκριμένα μοντέλα αλγορίθμων ελεγχόμενης μηχανικής μάθησης (Retrieved September 21, 2021, from https://www.rdocumentation.org/).

Θηκόγραμμα (Boxplot)

Δίνουν μια εικόνα για τον τρόπο με τον οποίο είναι οργανωμένα τα δεδομένα. Για να κατασκευαστεί χρειάζονται κυρίως πέντε μεταβλητές : η ελάχιστη τιμή, το πρώτο τεταρτημόριο, η διάμεσος, το τρίτο τεταρτημόριο και η μέγιστη τιμή (Retrieved September 21, 2021, from https://courses.lumenlearning.com/introstats1/chapter/box-plots/).

3. Αποτελέσματα

Παρατίθενται οι πίνακες με τις τιμές από τα πηλίκα (Μολυσμένα/Μάρτυρας) για την κάθε ποικιλία για την ρίζα και τον βλαστό αντίστοιχα.

Μεταβολίτες	Ποικιλία 7 : Μολυσμένα Ρίζα	Ποικιλία 8 : Μολυσμένα Ρίζα
[614] L-proline 1 [8.58]+ L-proline 1 [8.567]	0.656095925	2.649787131
[145742] L-proline 2 [10.321]+L-proline 2 [10.341]	0	0
[16219560] lactobionic acid 3 [25.406]	Inf	Inf
[18950] D-mannose 1 [17.287]	0.284411383	3.945763495
[192826] maltotriose 2 [30.668]	0.214406182	2.196967641
[204] allantoin 2 [17.356]	0	2.048610029
[206] D (+) galactose 2 [17.662]	2.302073752	2.444035021
[21236] L-norleucine 1 [8.945]	Inf	Inf
[236] L-asparagine 1 [14.496]	0	Inf
[268779] N-methylglutamic acid 4 [15.174]	0.01802791	0.985631647
[2724552] tagatose 1 [17.011]	33.64640605	11.24950028
[3034828] palatinitol 2 [26.01]	0	1.879390301
[33032] L-glutamic acid 1 [13.338]	0.090682361	Inf
[439193] isomaltose 2 [25.863]	2.036200242	2.477463674
[439227] pipecolic acid 1 [9.868]	0.270704862	1.744113575
[439958] D-glucose-6-phosphate 1 [21.394]+D-glucose-6-phosphate 2 [21.558]	0.644151986	0.61090939
[440658] melibiose 2 [25.784]	0.081955329	0.94516707
[441032] D (+)altrose 1 [17.397]	0.254285615	1.205875643
[441035] talose 2 [17.584]	0	0
[448388] D-allose 1 [17.278]	Inf	0.709008502
[5280335] D-sphingosine 2 [22.36]+ D-sphingosine 3 [22.527]	0	Inf
[5280451] maleamic acid 3 [13.681]	Inf	0
[5950] L-alanine 2 [11.182]	2.679663974	0.425349106
[5951] L-serine 1 [9.706]	0.327592286	0.715360793
[5960] aspartic acid 1 [12.002]	0.527628183	2.369531088
[6255] maltose 1 [24.702]	0.258357275	1.939050965
[6262] L-ornithine 1 [14.349]	0.276017855	2.801973594
[6287] L-valine 1 [7.296]	Inf	0
[6288] L-threonine 1 [10.224]	Inf	Inf
[6305] L-tryptophan 1 [20.284]	0.240599184	1.023587183
[65098] norvaline 2 [9.468]	0.22656459	1.226552278
[738] L-glutamine 1 [13.431]	1.342987146	2.795056609
[791] DL-isoleucine 1 [8.576]	Inf	Inf
[8299] acetol 1 [13.293]	1.794476576	2.913656315
[8299] acetol 3 [15.377] +acetol 4 [15.498]	0.184217467	0
[835] dehydroascorbic acid 1 [16.863]	0	0
[84571] lactose 1 [24.386]	0	0

[867] malonic acid 1 [8.919]	0.716324669	2.990769325
[1000] 2-amino-1-phenylethanol [15.668]	0	0.889494157
[1004] phosphoric acid [9.966]	0.8934779	Inf
[1045] putrescine [15.709]	7.249171163	12.34547189
[1102] spermidine 2 [20.811]	8.063736407	Inf
[14490] 6-hydroxy caproic acid [12.059]	0	0.391926693
[156807] D-lyxosylamine 2 [14.861]	0.250465813	1.366075153
[165577] leucrose [24.975]	0	Inf
[1662] 3-hydroxy-3-methylglutaric acid (dicrotalic acid) [14.232]	Inf	2.661144021
[26519] tetratriacontane [29.185]	0	Inf
[3037582] mucic acid [18.907]	Inf	0
[439746] 6-deoxy-D-glucose 1 [15.598]	7.591643292	3.516276789
[439766] citramalic acid [12.63]	0.206395269	3.683179993
[444212] trans-aconitic acid [15.842]	0	0.805336589
[487] methylmalonic acid [9.088]	Inf	1.836383617
[5984] fructose 1 [17.18]	0.335167211	1.016357965
[6057] L-tyrosine 2 [17.856]+tyrosine 2 [17.871]	0.140703652	2.405557709
[6101] p-toluenesulfonic acid [14.265]	0.135497231	1.970843066
[65080] phenyl-beta-glucopyranoside [21.186]	0	0
[68152] 3-hydroxypropanoic acid 2 [13.848]	0	Inf
[6854] carbazole 1 [17.044]	Inf	Inf
[7405] L-pyroglutamic acid [13.218]	0	0
[7427] D-(+) trehalose [24.752]	0.158830681	0.087787197
[754] glycerol 1-phosphate [16.056]	0	0
[785] hydroquinone [11.659]	1.063199513	2.999276091
[8066] 2-butyne-1,4-diol [9.446]	0.359594229	1.982860614
[8215] behenic acid [23.897]	0.501212377	3.491193868
[8742] shikimic acid [16.433]	0	0
[8897] iminodiacetic acid 2 [13.285]	1.124777735	1.440183034
[92817] melezitose [29.884]	0.929314112	Inf
[94154] arabitol [15.601]	0.00711211	0.049204754
[94214] methyl-beta-D-galactopyranoside [16.935]	0.908594504	3.122024697
[993] ribose [15.113]	0.26901898	1.17248768
[1531] 2-amino-2-methyl-1,3-propanediol 2 [10.56]	0.203952319	0.59778014
[239] Beta- alanine 2 [14.555]	Inf	0
[439240] D-lyxose 2 [14.889]	1.647123918	1.698908529
Epemvaseis	0.207338699	1.345786725
Poikilies	0	0

Μεταβολίτες	Ποικιλία 7 : Μολυσμένα Βλαστός	Ποικιλία 8 : Μολυσμένα Βλαστός
[111064] ribonic acid-gamma-lactone 2 [15.154]	0.526897751	Inf
[138] 5-aminovaleric acid 1 [14.458] +5-aminovaleric acid 2 [14.955]	3.602919689	0.556378021
[614] L-proline 1 [8.58]+ L-proline 1 [8.567]	4.607850881	Inf
[145742] L-proline 2 [10.321]+L-proline 2 [10.341]	0.301855312	Inf
[16219560] lactobionic acid 3 [25.406]	0	Inf
[18950] D-mannose 1 [17.287]	0.468368563	2.265914053
[192826] maltotriose 2 [30.668]	0	0.309564042
[204] allantoin 2 [17.356]	0.561771466	14.12331044
[206] D (+) galactose 2 [17.662]	1.598742678	1.089779174
[21236] L-norleucine 1 [8.945]	Inf	1.663226799
[236] L-asparagine 1 [14.496]	0.610775192	0.450934372
[268779] N-methylglutamic acid 4 [15.174]	0.106182754	0.300278018
[2724552] tagatose 1 [17.011]	5.415784358	23.10916451
[3034828] palatinitol 2 [26.01]	1.443108257	Inf
[33032] L-glutamic acid 1 [13.338]	1.081108157	Inf
[439193] isomaltose 2 [25.863]	1.344934996	0.77344381
[439227] pipecolic acid 1 [9.868]	0.483617356	1.565132321
[439451] galactinol 1 [26.055]	0.874207307	10.11551316
[439958] D-glucose-6-phosphate 1 [21.394]+D-glucose-6-phosphate 2 [21.558]	0.270865953	4.902705703
[440658] melibiose 2 [25.784]	0	Inf
[441032] D (+)altrose 1 [17.397]	0.527739374	0.906169585
[441035] talose 2 [17.584]	0.399765723	Inf
[448388] D-allose 1 [17.278]	0	Inf
[5280335] D-sphingosine 2 [22.36]+ D-sphingosine 3 [22.527]	0	0
[5280451] maleamic acid 3 [13.681]	0	0
[5950] L-alanine 2 [11.182]	1.803269338	1.298107945
[5951] L-serine 1 [9.706]	0.839700108	0.499104601
[5960] aspartic acid 1 [12.002]	0.210805749	2.256211401
[6106] L-leucine 1 [8.298]	0.120294863	1.198046523
[6255] maltose 1 [24.702]	0.44165807	3.448369689
[6262] L-ornithine 1 [14.349]	0	Inf
[6287] L-valine 1 [7.296]	0.224809895	Inf
[6288] L-threonine 1 [10.224]	0.641785931	Inf
[6305] L-tryptophan 1 [20.284]	0.349056782	0.368083152
[65098] norvaline 2 [9.468]	0.250991106	1.189406801
[736] gluconic acid lactone 1 [17.303]	1.962804632	3.706206557
[738] L-glutamine 1 [13.431]	0	Inf
[791] DL-isoleucine 1 [8.576]	0.373051425	Inf
[8299] acetol 1 [13.293]	0.878694056	2.185641005
[8299] acetol 3 [15.377] +acetol 4 [15.498]	0.082828427	0.073207762
[835] dehydroascorbic acid 1 [16.863]	0.775550358	Inf
[84571] lactose 1 [24.386]	0.35451225	Inf

[1000] 2-amino-1-phenylethanol [15.668]	Inf	0.853817711
[1004] phosphoric acid [9.966]	0.421861729	0.149389642
[1045] putrescine [15.709]	2.499726373	19.68865715
[1102] spermidine 2 [20.811]	3.944048454	Inf
[14490] 6-hydroxy caproic acid [12.059]	0.51688735	0.500340783
[156807] D-lyxosylamine 2 [14.861]	0.549868365	0.94780675
[165577] leucrose [24.975]	0.319790477	Inf
[1662] 3-hydroxy-3-methylglutaric acid (dicrotalic acid) [14.232]	Inf	0
[26519] tetratriacontane [29.185]	0	0
[3037582] mucic acid [18.907]	3.334203053	Inf
[439746] 6-deoxy-D-glucose 1 [15.598]	0	Inf
[439766] citramalic acid [12.63]	0.940181949	0.633738008
[444212] trans-aconitic acid [15.842]	0.47116363	4.179629695
[444972] fumaric acid [10.94]	0.681104768	Inf
[487] methylmalonic acid [9.088]	0.520982238	0.456338789
[5641] carbamic acid ethyl ester (urethane) [6.361]	0.002008093	Inf
[5984] fructose 1 [17.18]	0.270175382	0.14579391
[604] gluconic acid 2 [18.297]	Inf	0
[6057] L-tyrosine 2 [17.856]+tyrosine 2 [17.871]	0.593999368	1.236079246
[60961] adenosine [23.825]	0.854378901	0
[6101] p-toluenesulfonic acid [14.265]	0.973766361	0
[6274] L-histidine 3 [17.658]	1.929052078	Inf
[65080] phenyl-beta-glucopyranoside [21.186]	Inf	Inf
[68152] 3-hydroxypropanoic acid 2 [13.848]	0.373315521	Inf
[6854] carbazole 1 [17.044]	Inf	0.41635936
[7405] L-pyroglutamic acid [13.218]	0	Inf
[7427] D-(+) trehalose [24.752]	0.356399872	Inf
[754] glycerol 1-phosphate [16.056]	0.007783046	0
[785] hydroquinone [11.659]	0.691586284	0.814428485
[8066] 2-butyne-1,4-diol [9.446]	1.332174898	Inf
[8215] behenic acid [23.897]	0.152841604	0.409302643
[8742] shikimic acid [16.433]	0.343881323	Inf
[8897] iminodiacetic acid 2 [13.285]	1.385886123	1.070404205
[92817] D-(+)-melezitose [29.952]	1.031122839	0.559637265
[92817] melezitose [29.884]	0.523650354	0.816579851
[94154] arabitol [15.601]	0.60366762	Inf
[94214] methyl-beta-D-galactopyranoside [16.935]	0.467566583	3.557990058
[993] ribose [15.113]	1.118417022	4.131503145
[1531] 2-amino-2-methyl-1,3-propanediol 2 [10.56]	0.893440408	0.353406484
[239] Beta- alanine 2 [14.555]	Inf	0
[439240] D-lyxose 1 [14.762]+D-lyxose 2 [14.869]+D-lyxose 1 [14.741]	0.564446088	Inf
[439240] D-lyxose 2 [14.889]	1.326339606	0.764404406
Epemvaseis	0	Inf
Poikilies	0.121566953	Inf

Οι μεταβολίτες με πηλίκο >1 σημαίνει ότι βρίσκονται σε μεγαλύτερη συγκέντρωση στα μολυσμένα δείγματα σε σχέση με τον μάρτυρα, οι μεταβολίτες με πηλίκο μηδέν σημαίνει ότι η συγκέντρωσή τους στο μολυσμένο δείγμα είναι μηδέν και οι μεταβολίτες με πηλίκο άπειρο (= inf) σημαίνει ότι η συγκέντρωσή τους στο δείγμα του μάρτυρα ήταν μηδέν.

3.1 Μόνο-μεταβλητή ανάλυση

Παρακάτω δίνεται ο πίνακας με τους στατιστικά σημαντικούς μεταβολίτες ανά κατηγορία για τη ρίζα και τον βλαστό αντίστοιχα.

Ρίζα			
Μεταβολίτες	Ποικιλία : p-value	Επεμβάσεις : p-value	Αλληλεπίδραση : p-value
[10712] cellobiose 2 [24.7]	0.031980037		
[3034828] palatinitol 2 [26.01]	0.00042944	0.017228526	
[441035] talose 2 [17.584]	0.005175901	0.005175901	0.010351801
[5280451] maleamic acid 3 [13.681]	0.031980037	0.01868839	
[1045] putrescine [15.709]	0.031980037	0.001033888	
[165577] leucrose [24.975]	0.027582742		
[7427] D-(+) trehalose [24.752]	0.02085462		
[94154] arabitol [15.601]	0.04669938		
[6262] L-ornithine 1 [14.349]		0.0290796	
[1000] 2-amino-1-phenylethanol [15.668]		0.01868839	

Πίνακας 3. Στατιστικά σημαντικοί μεταβολίτες για τη ρίζα

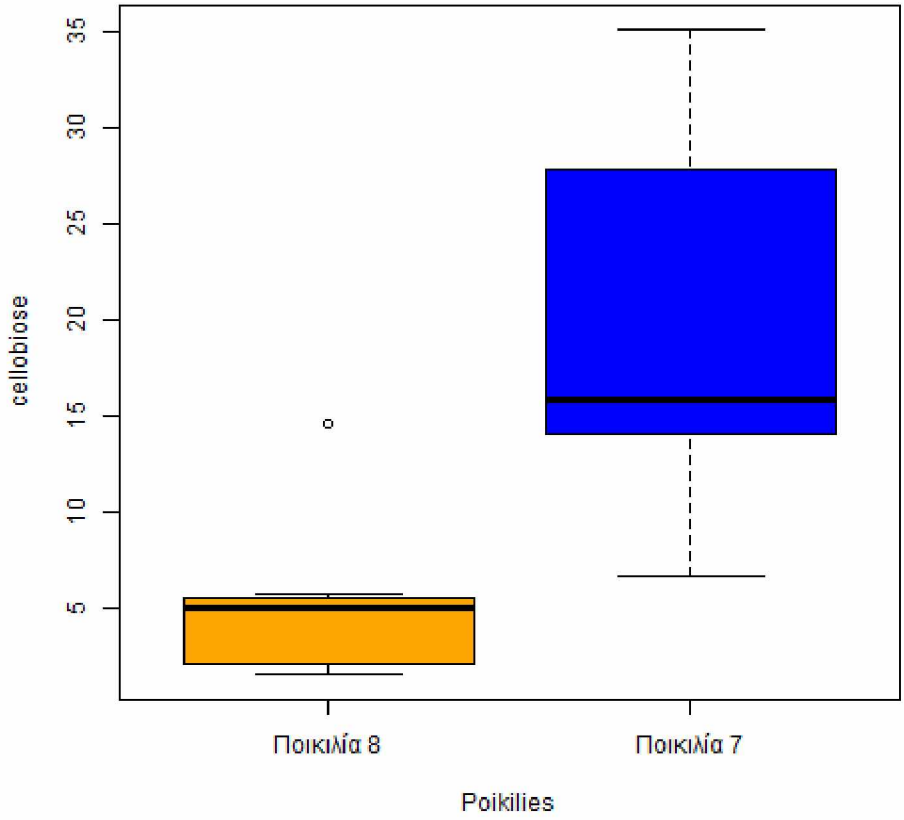
Βλαστός			
Μεταβολίτες	Ποικιλία : p-value	Επεμβάσεις : p-value	Αλληλεπίδραση : p-value
[145742] L-proline 2 [10.321]+L-proline 2 [10.341]	0.044147594		
[236] L-asparagine 1 [14.496]		0.010250096	
[8299] acetol 3 [15.377] +acetol 4 [15.498]	0.019756897		
[444972] fumaric acid [10.94]	0.001632986	0.001461508	0.001775413
[68152] 3-hydroxypropanoic acid 2 [13.848]	0.003937483	0.002624989	0.003937483
[754] glycerol 1-phosphate [16.056]		0.001461508	
[8066] 2-butyne-1,4-diol [9.446]	0.011046964		
[8897] iminodiacetic acid 2 [13.285]	0.036186683		

Πίνακας 4. Στατιστικά σημαντικοί μεαταβολίτες για τον βλαστό

Παρακάτω παρουσιάζονται τα θηκογράμματα για τους στατιστικά σημαντικούς μεταβολίτες για τον παράγοντα ποικιλία για την ρίζα και τον βλαστό και για τον παράγοντα επεμβάσεις για τη ρίζα και τον βλαστό αντίστοιχα.

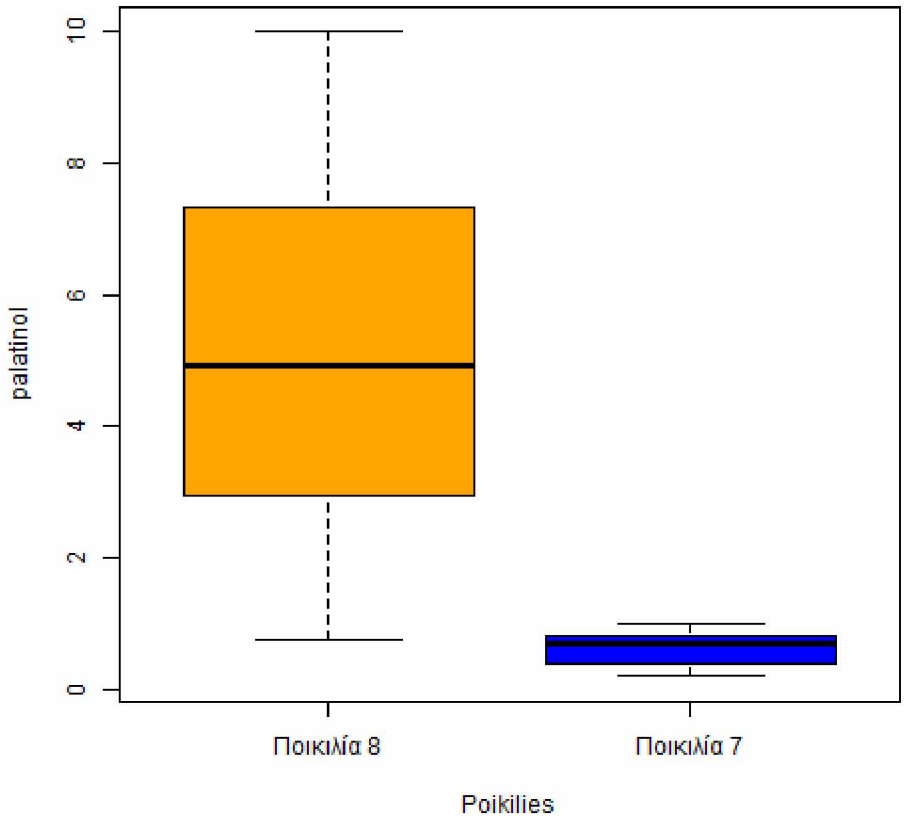
Παράγοντας Ποικιλία

- **Ρίζα**



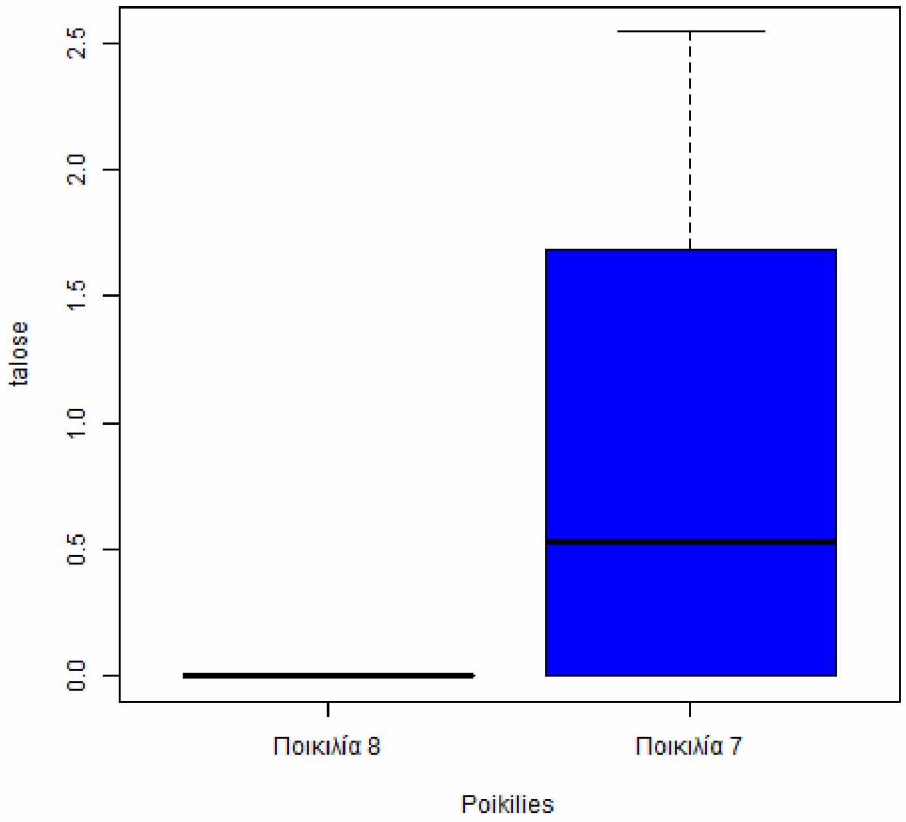
Εικόνα 19. Θηκόγραμμα για την cellobiose στην ρίζα

Παρατηρείται ότι η cellobiose 2 στην ποικιλία 7 βρίσκεται σε μεγαλύτερη συγκέντρωση σε σχέση με την ποικιλία 8. Επομένως η cellobiose έχει θετική συνεισφορά όσο αναφορά τον διαχωρισμό των ποικιλιών.



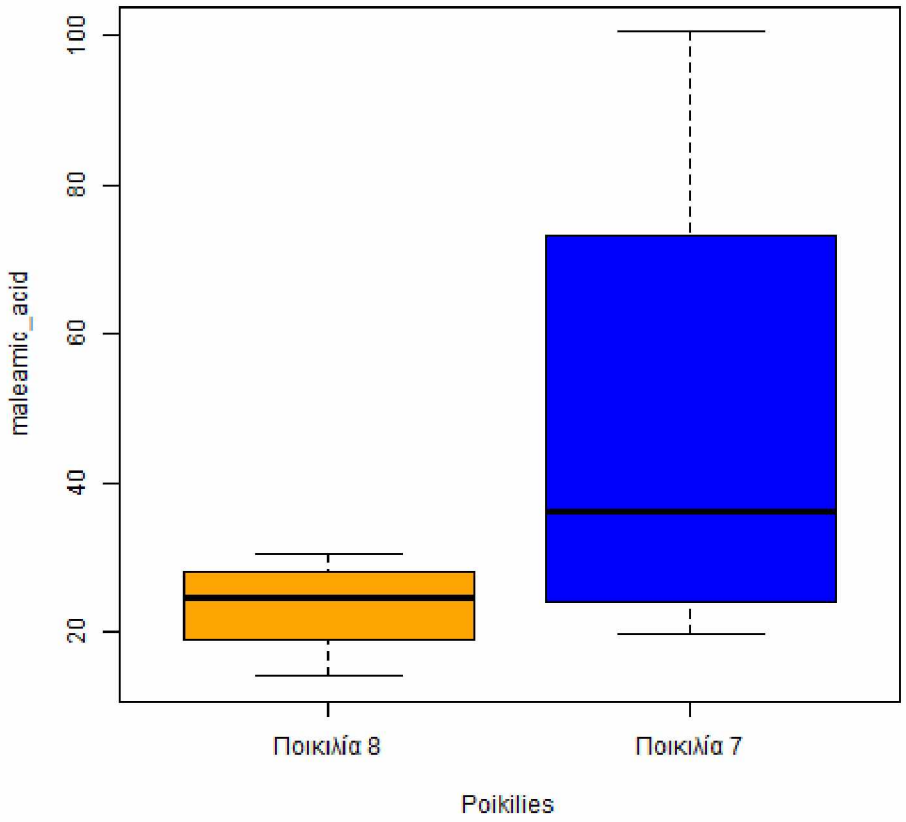
Εικόνα 20. Θηκόγραμμα για την palatinol στην ρίζα

Η palatinol βρίσκεται σε μεγαλύτερη συγκέντρωση στην ποικιλία 8 σε σχέση με την ποικιλία 7. Επομένως έχει θετική συνεισφορά στον διαχωρισμό μεταξύ των ποικιλιών.



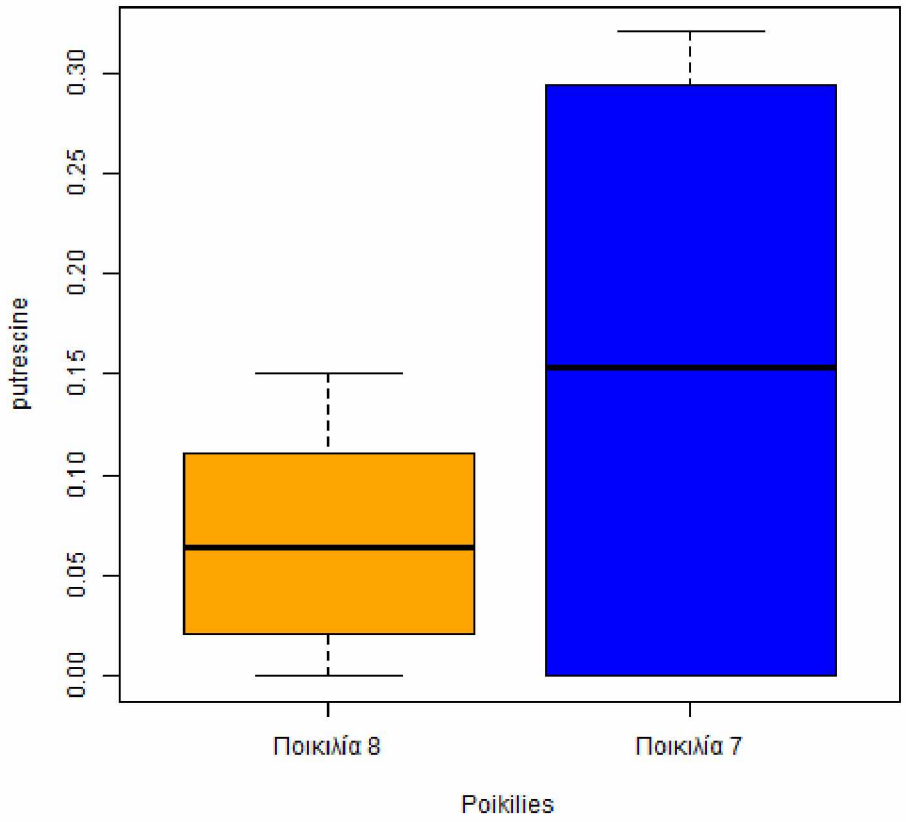
Εικόνα 21. Θηκόγραμμα για την talose στην ρίζα

Παρατηρείται μηδενική συγκέντρωση της talose στην ποικιλία 8 σε σχέση με την ποικιλία 7.



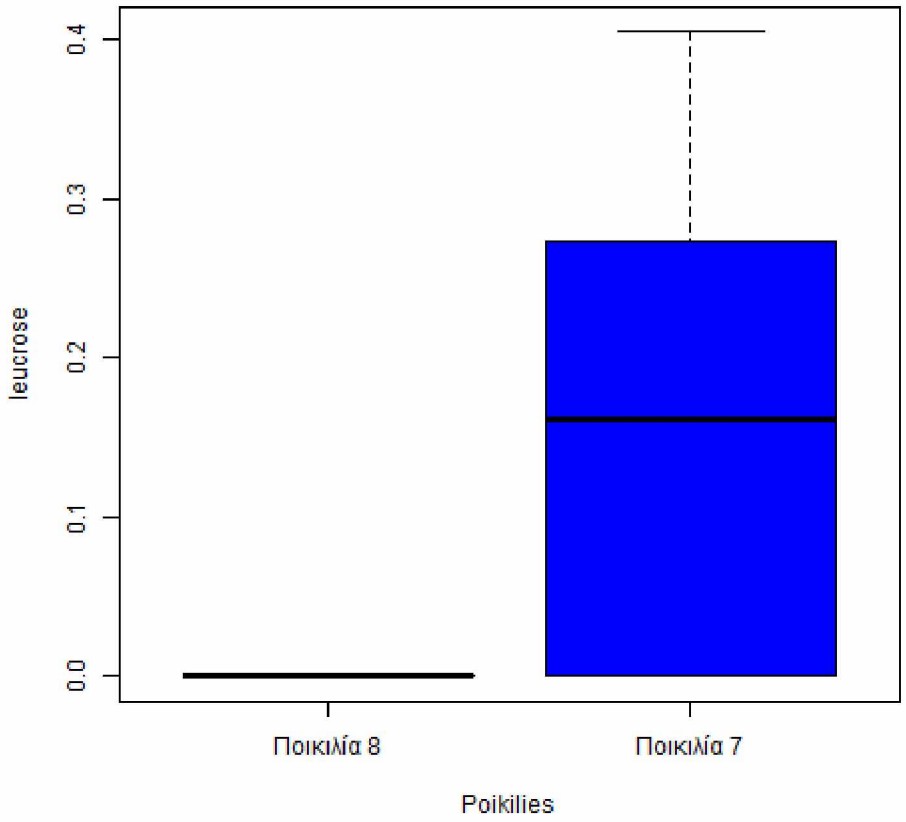
Εικόνα 22. Θηκόγραμμα για το maleamic acid στην ρίζα

Μεγαλύτερη συγκέντρωση του maleamic acid στην ποικιλία 7 σε σχέση με την ποικιλία 8. Επομένως έχει θετική συνεισφορά στον διαχωρισμό των ποικιλιών.



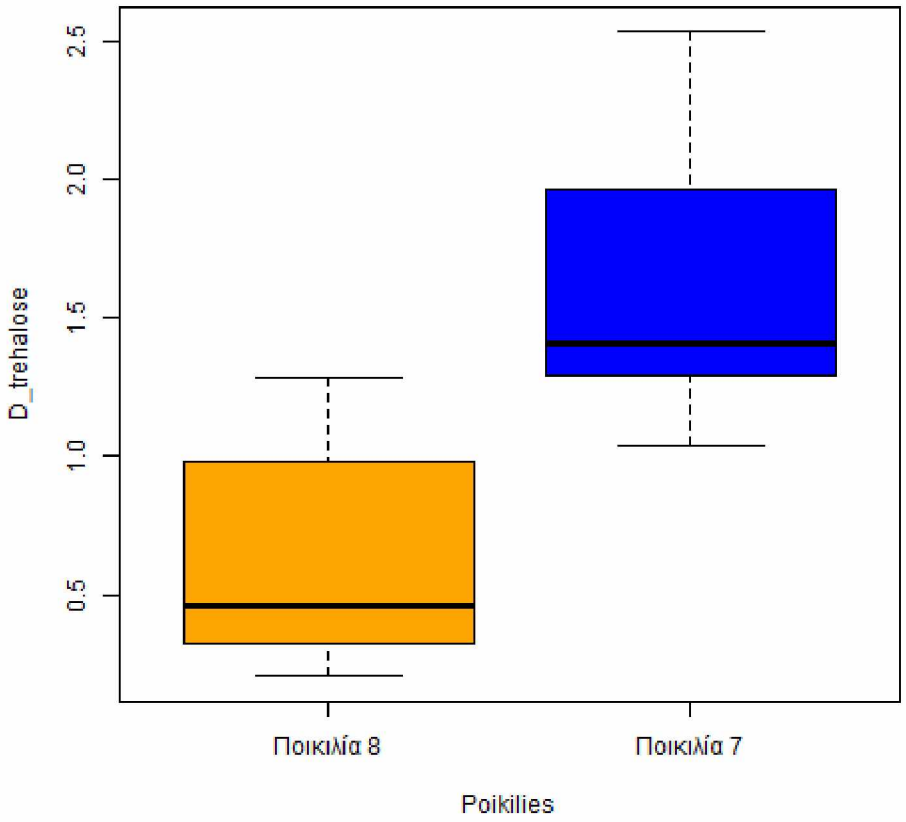
Εικόνα 23. Θηκόγραμμα για την putrescine στην ρίζα

Μεγαλύτερη συγκέντρωση της putrescine στην ποικιλία 7 σε σχέση με την ποικιλία 8. Θετική συνεισφορά της putrescine στον διαχωρισμό των ποικιλιών.



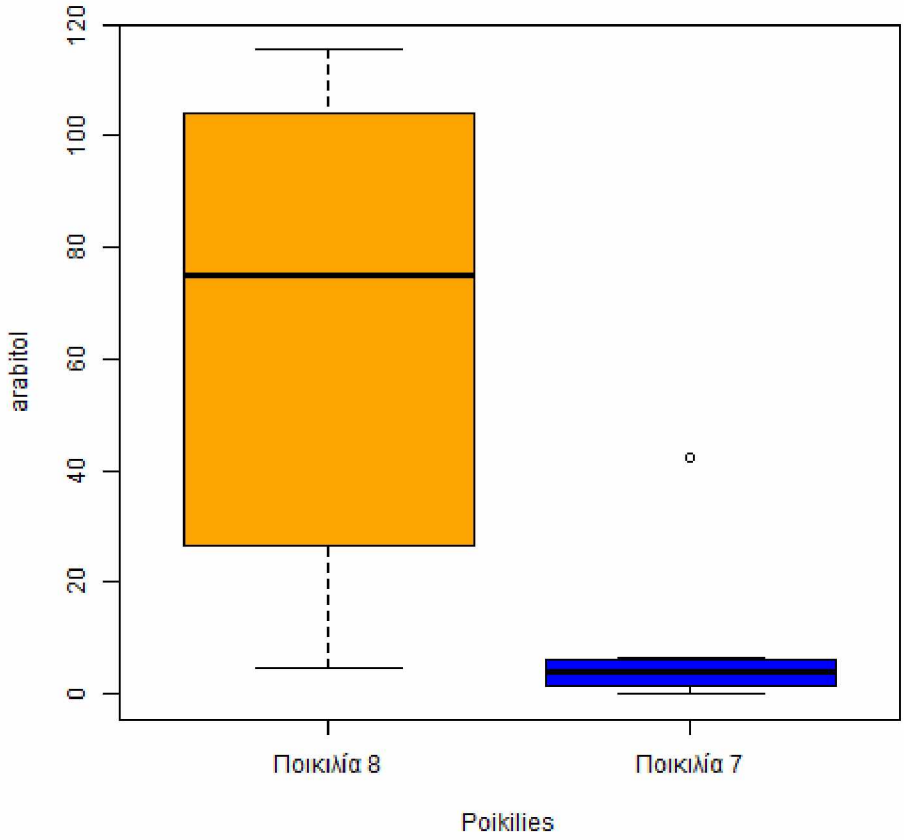
Εικόνα 24. Θηκόγραμμα για την leucrose στην ρίζα

Μηδενική συγκέντρωση της leucrose στην ποικιλία 8.



Εικόνα 25. Θηκόγραμμα για την D-trehalose στην ρίζα

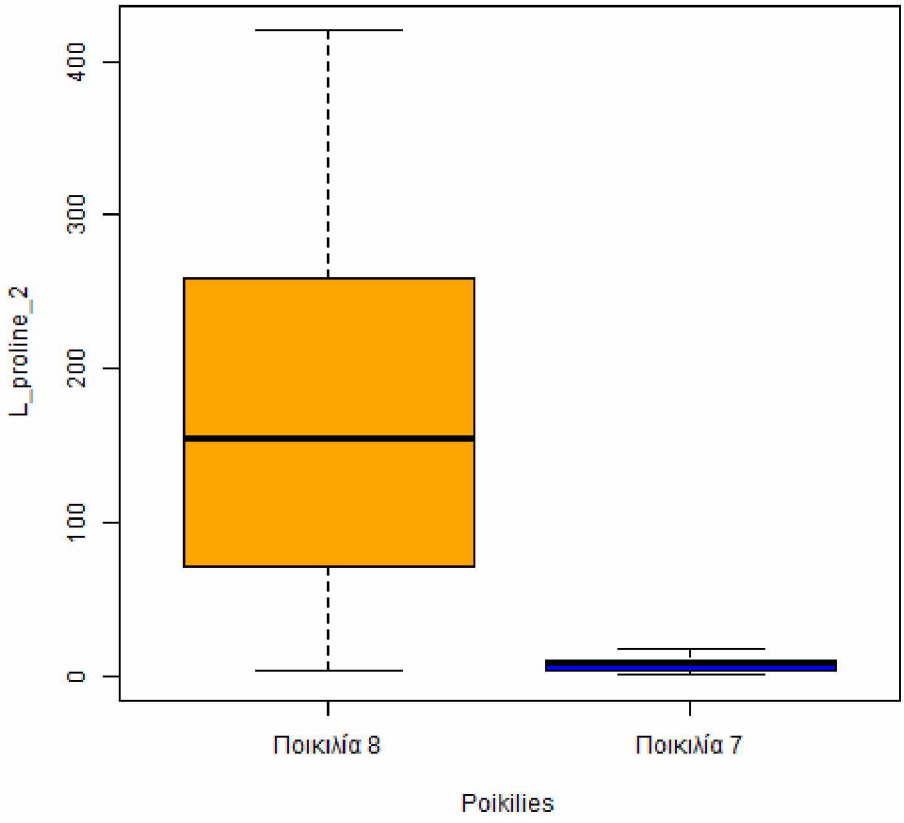
Μεγαλύτερη συγκέντρωση της D-trehalose στην ποικιλία 7 σε σχέση με την 8. Θετική συνεισφορά της D-trehalose στον διαχωρισμό των ποικιλιών.



Εικόνα 26. Θηκόγραμμα για την arabitol στην ρίζα

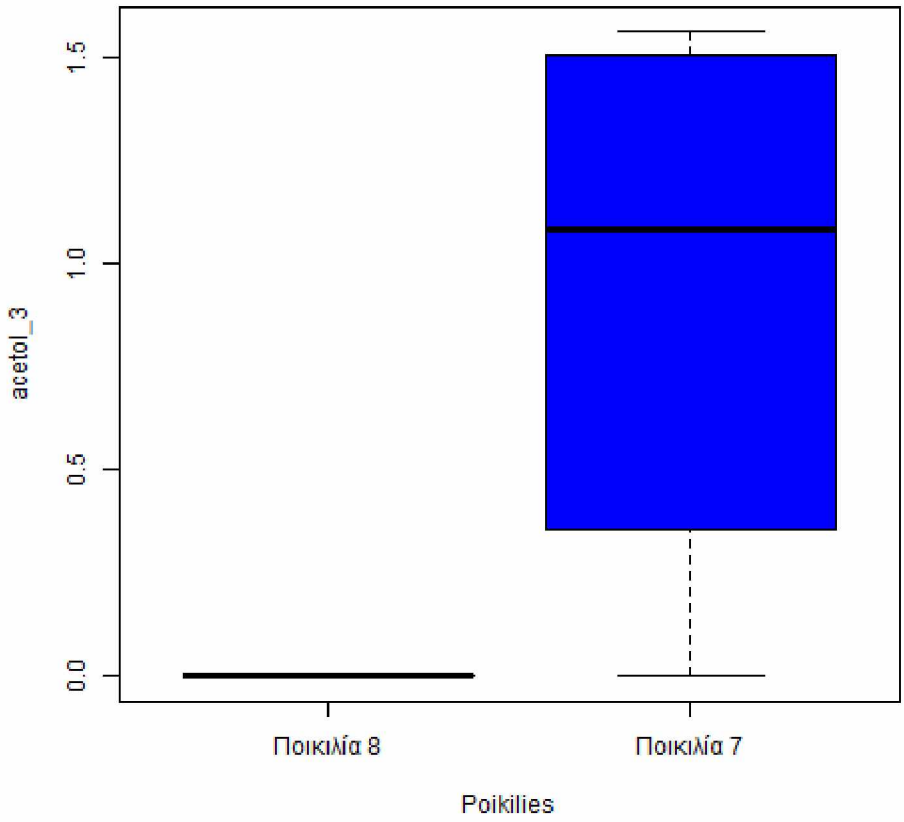
Παρατηρείται μεγαλύτερη συγκέντρωση της arabitol στην ποικιλία 8 σε σχέση με την 7. Άρα η arabitol συνεισφέρει στον διαχωρισμό των ποικιλιών.

- **Βλαστός**



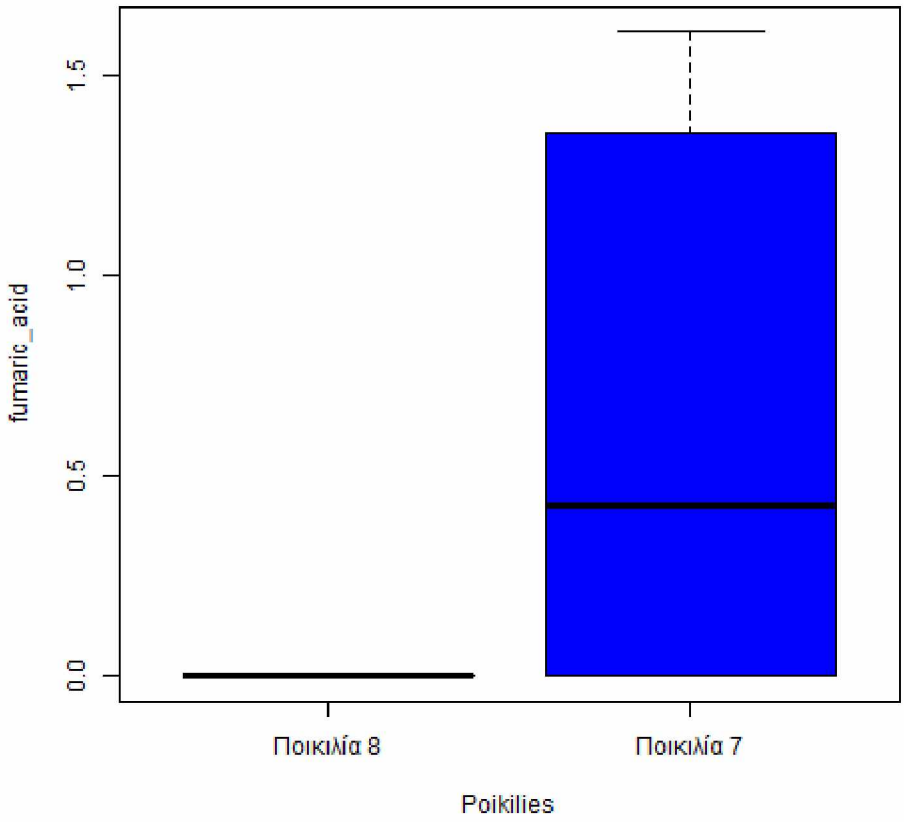
Εικόνα 27. Θηκόγραμμα για την L-proline 2 στον βλαστό

Παρατηρείται μεγαλύτερη συγκέντρωση της L-proline_2 στην ποικιλία 8. Επομένως η L-proline_2 συνεισφέρει στον διαχωρισμό των ποικιλιών.



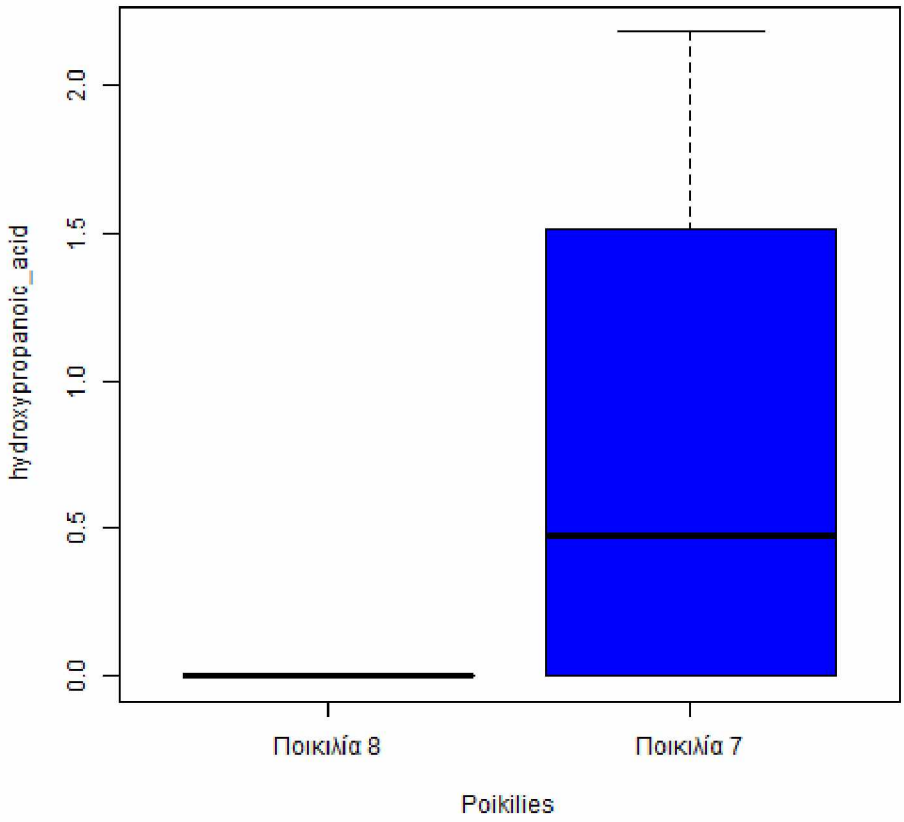
Εικόνα 28. Θηκόγραμμα για την acetol 3 στον βλαστό

Μηδενική συγκέντρωση της acetol_3 στην ποικιλία 8. Επομένως η acetol_3 συνεισφέρει στον διαχωρισμό των ποικιλιών.



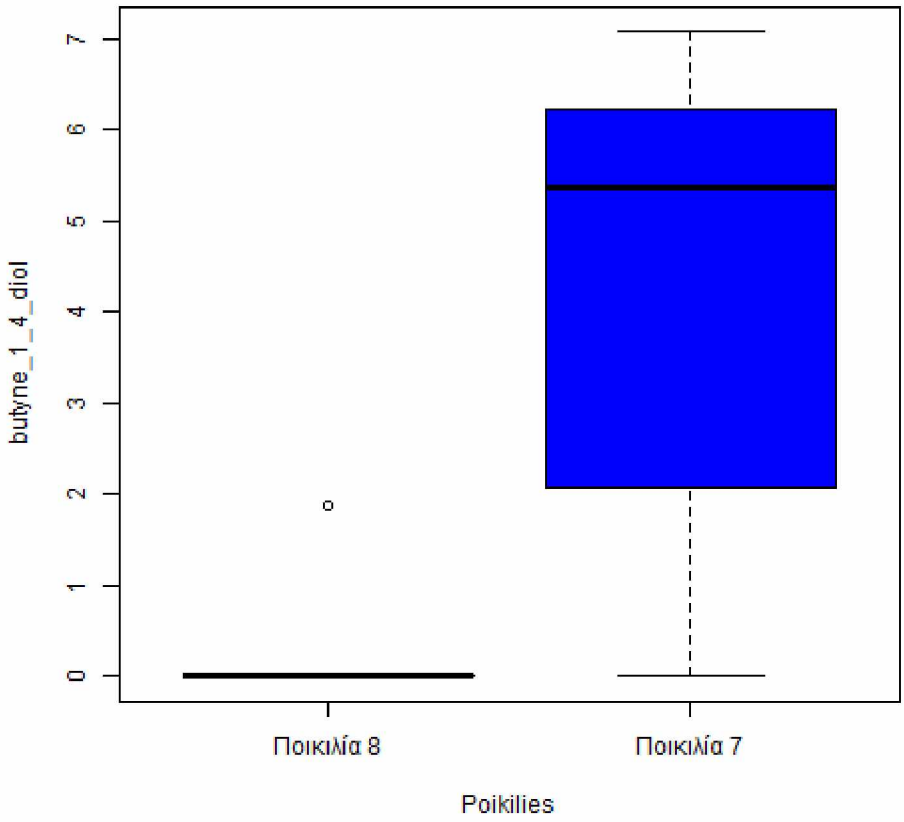
Εικόνα 29. Θηκόγραμμα για το fumaric acid στον βλαστό

Επίσης μηδενική συγκέντρωση του fumaric acid στην ποικιλία 8. Άρα το fumaric_ acid συνεισφέρει στον διαχωρισμό των ποικιλιών.



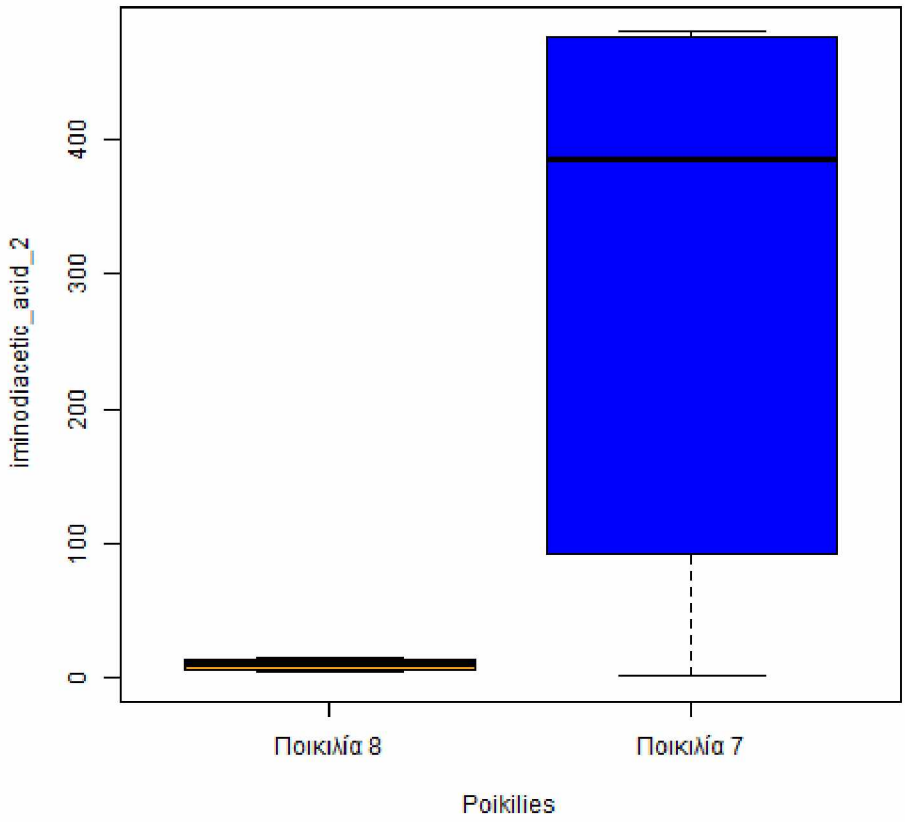
Εικόνα 30. Θηκόγραμμα για το hydroxypropanoic acid στον βλαστό

Μηδενική συγκέντρωση του hydroxypropanoic_acid στην ποικιλία 8. Άρα το hydroxypropanoic_acid συνεισφέρει στον διαχωρισμό των ποικιλιών.



Εικόνα 31. Θηκόγραμμα για το butyne-1,4-diol στον βλαστό

Μηδενική συγκέντρωση του butyne-1,4-diol στην ποικιλία 8. Επομένως συνεισφέρει στον διαχωρισμό των ποικιλιών.

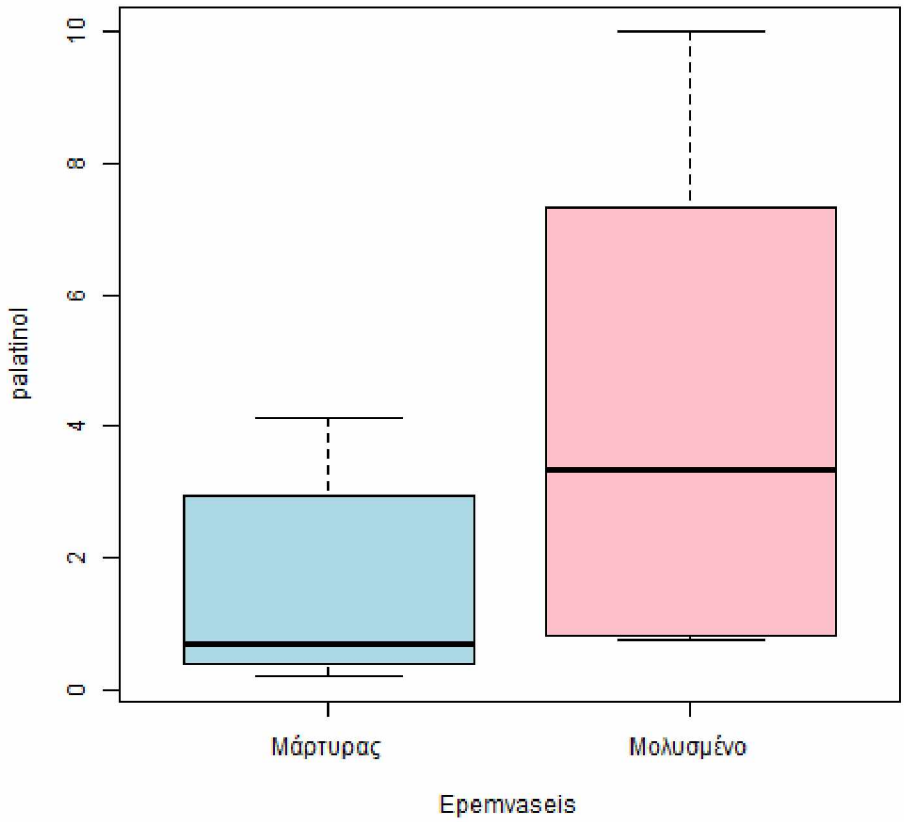


Εικόνα 32. Θηκόγραμμα για το iminodiacetic acid 2 στον βλαστό

Σχεδόν μηδενική συγκέντρωση του iminodiacetic_acid_2 στην ποικιλία 8.

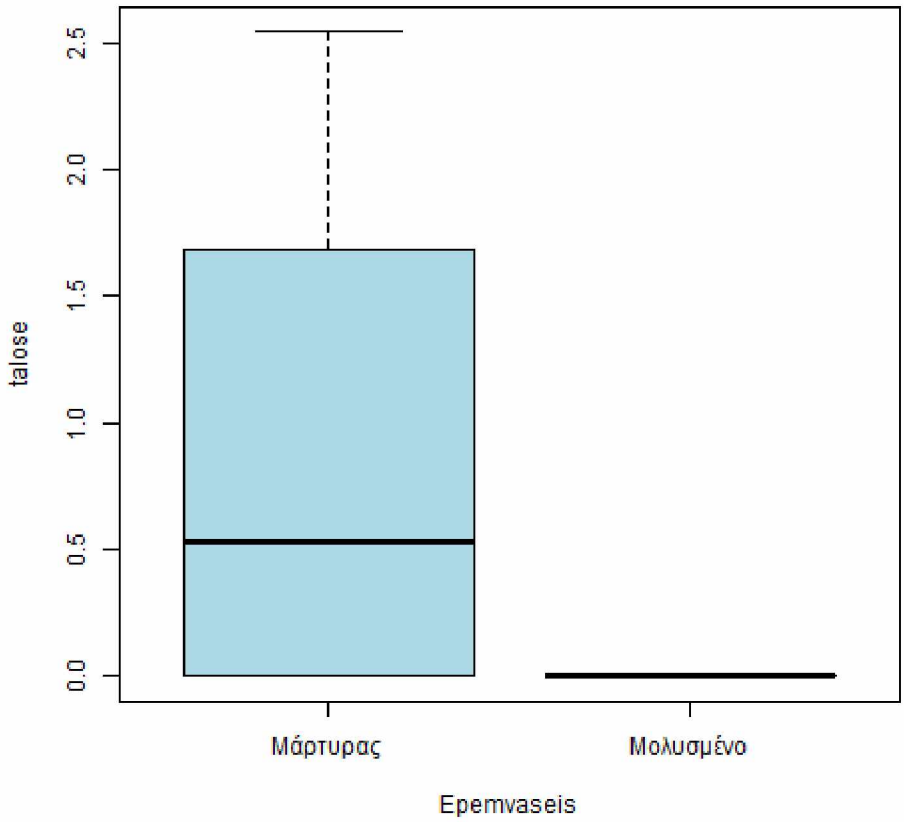
Παράγοντας Επεμβάσεις

- **Ρίζα**



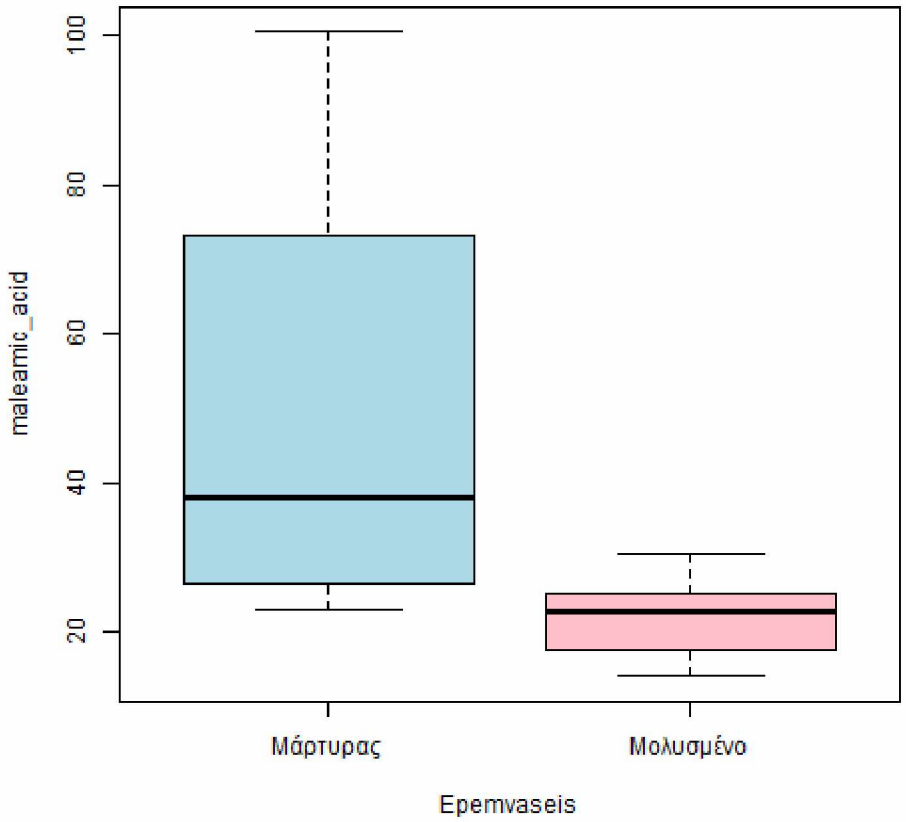
Εικόνα 33. Θηκόγραμμα για την palatinol στην ρίζα

Παρατηρείται ότι η palatinol βρίσκεται σε μεγαλύτερη συγκέντρωση στο μολυσμένο δείγμα σε σχέση με το δείγμα του μάρτυρα.



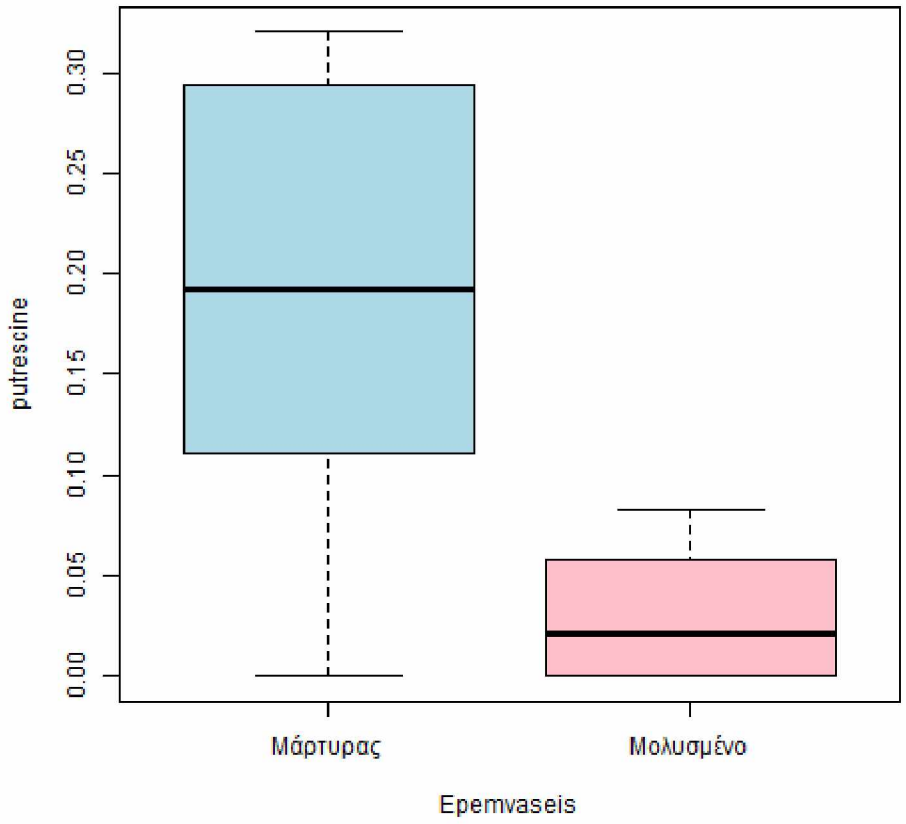
Εικόνα 34. Θηκόγραμμα για την talose στην ρίζα

Παρατηρείται μηδενική συγκέντρωση της talose στο μολυσμένο δείγμα. Επομένως η talose συνεισφέρει στον διαχωρισμό μεταξύ των δειγμάτων.



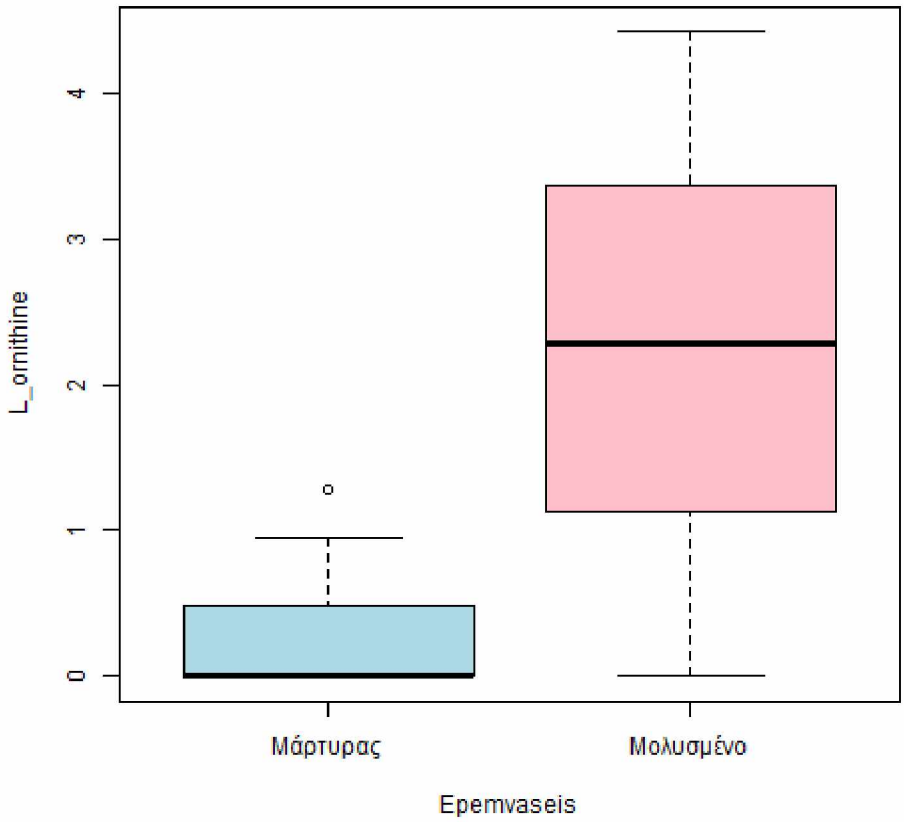
Εικόνα 35. Θηκόγραμμα του maleamic_acid στην ρίζα

Παρατηρείται μεγαλύτερη συγκέντρωση του maleamic_acid στο δείγμα του μάρτυρα σε σχέση με το μολυσμένο δείγμα.



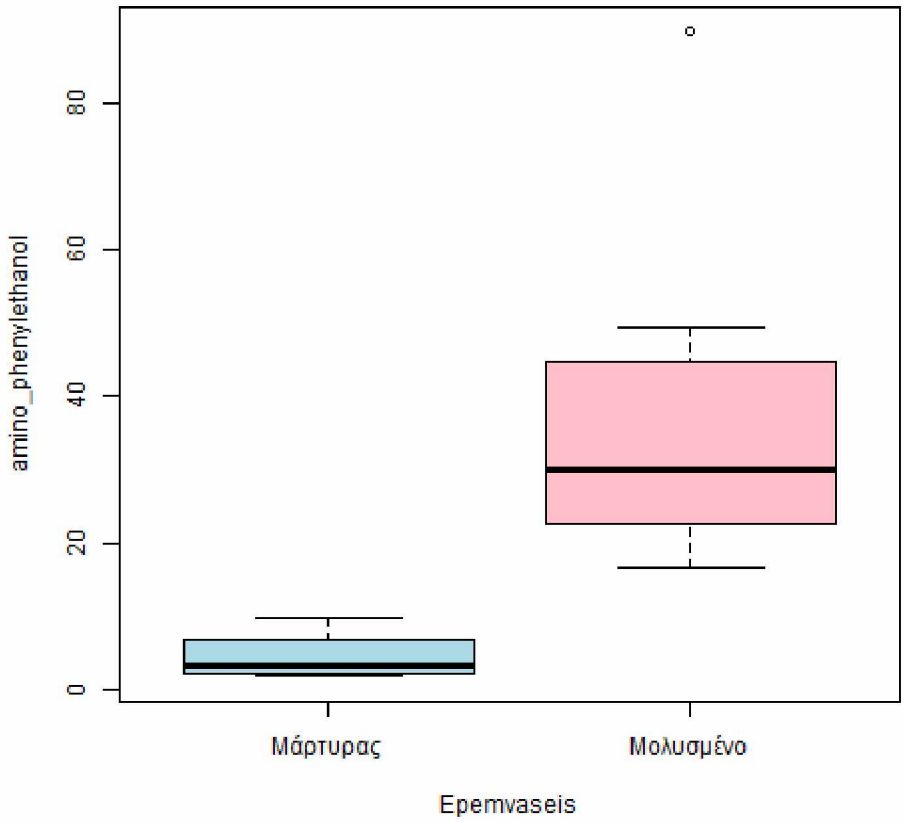
Εικόνα 36. Θηκόγραμμα για την putrescine στην ρίζα

Η συγκέντρωση της putrescine στο δείγμα του μάρτυρα είναι μεγαλύτερη σε σχέση με την συγκέντρωσή της στο μολυσμένο δείγμα.



Εικόνα 37. Θηκόγραμμα για την L-ornithine στην ρίζα

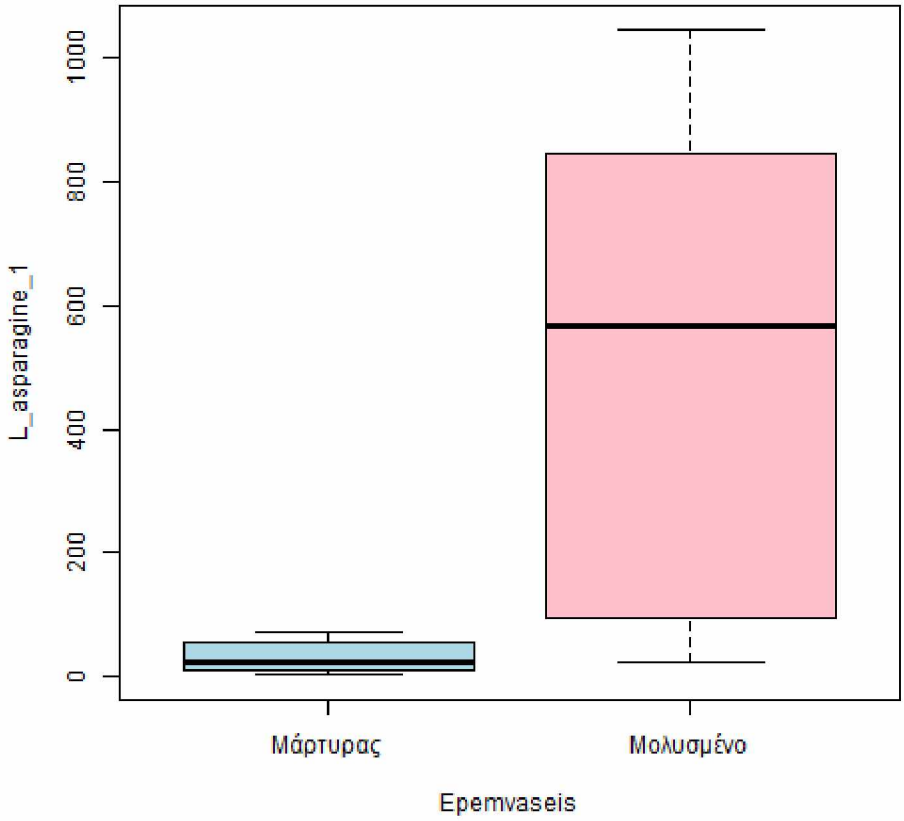
Παρατηρείται υψηλή συγκέντρωση της L-ornithine στο μολυσμένο δείγμα σε σχέση με το δείγμα του μάρτυρα.



Εικόνα 38. Θηκόγραμμα για το amino_phenylethanol για την ρίζα

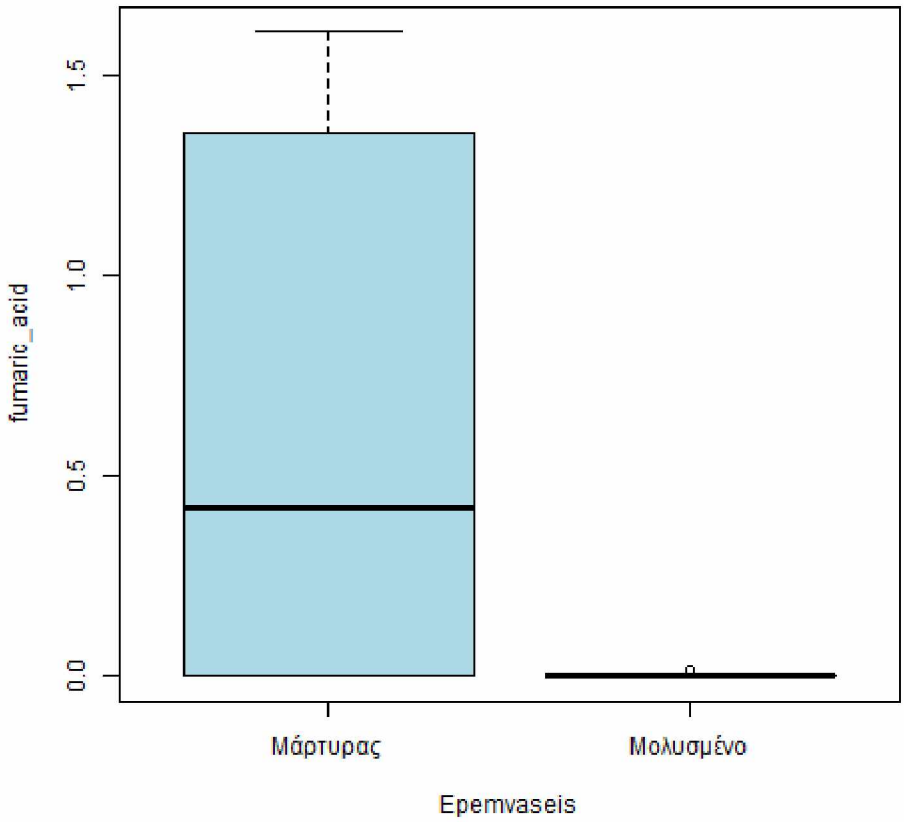
Παρατηρείται υψηλή συγκέντρωση του amino_phenylethanol στο μολυσμένο δείγμα σε σχέση με το δείγμα του μάρτυρα.

- **Βλαστός**



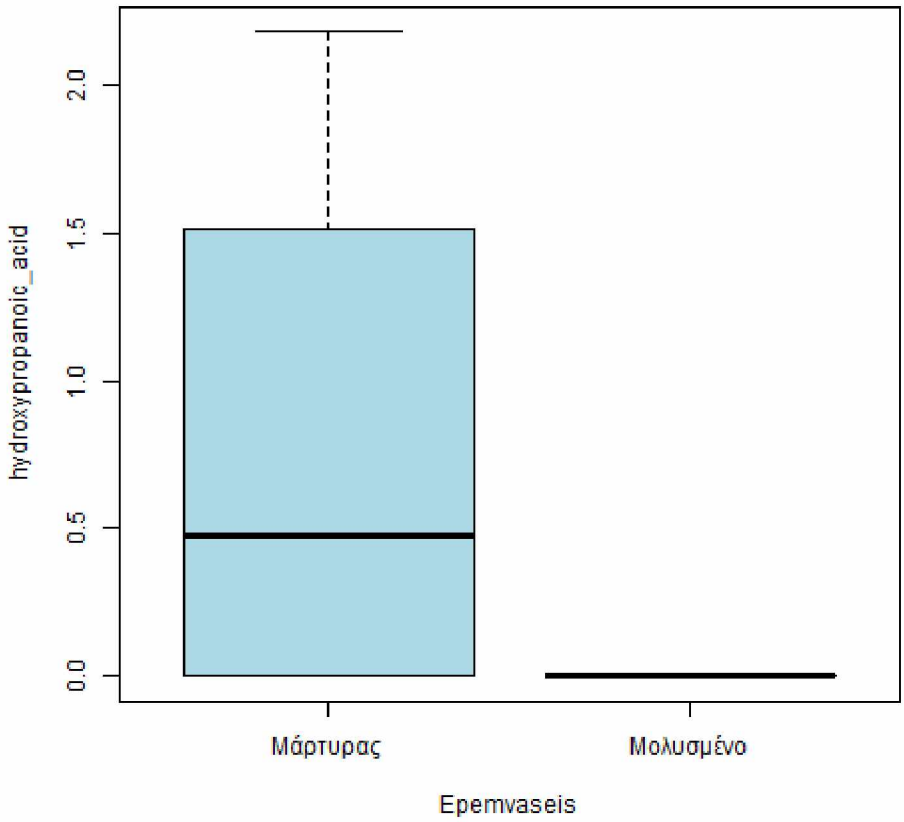
Εικόνα 39. Θηκόγραμμα για την L-asparagine-1 για τον βλαστό

Παρατηρείται υψηλή συγκέντρωση της L-asparagine-1 στο μολυσμένο δείγμα σε σχέση με το δείγμα του μάρτυρα.



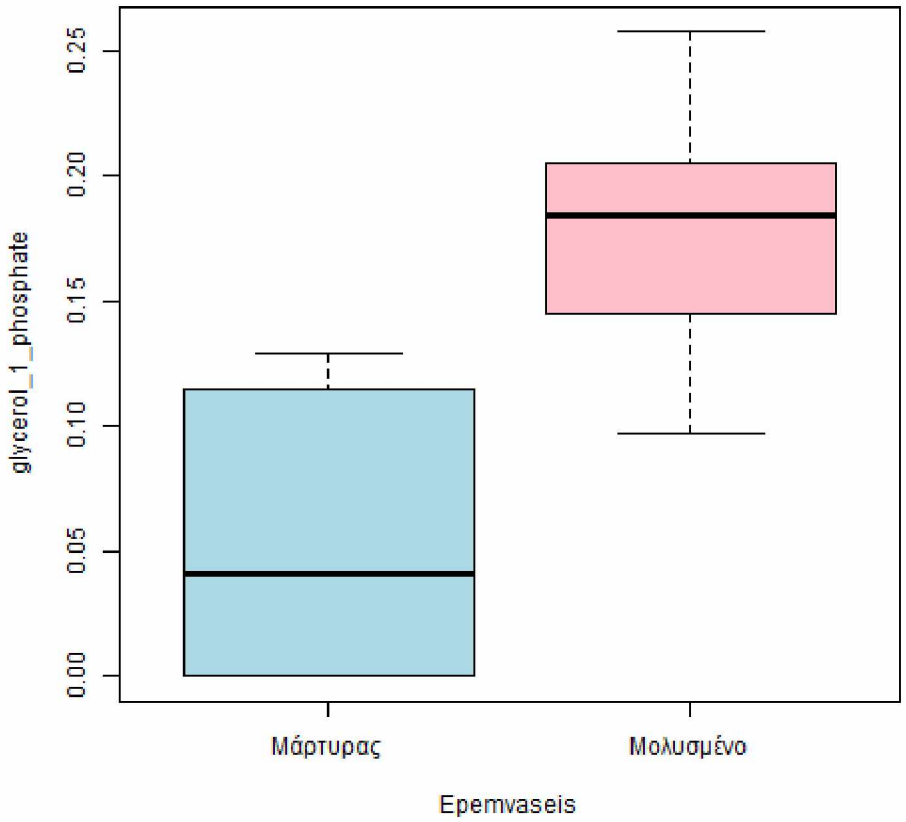
Εικόνα 40. Θηκόγραμμα για το fumaric_acid στον βλαστό

Παρατηρείται υψηλή συγκέντρωση του fumaric_acid στο δείγμα του μάρτυρα , ενώ στο μολυσμένο δείγμα η συγκέντρωση του είναι μηδενική.



Εικόνα 41.Θηκόγραμμα για το hydroxypropanoic_acid για τον βλαστό

Μηδενική συγκέντρωση του hydroxypropanoic_acid στο μολυσμένο δείγμα και υψηλή στο δείγμα του μάρτυρα.



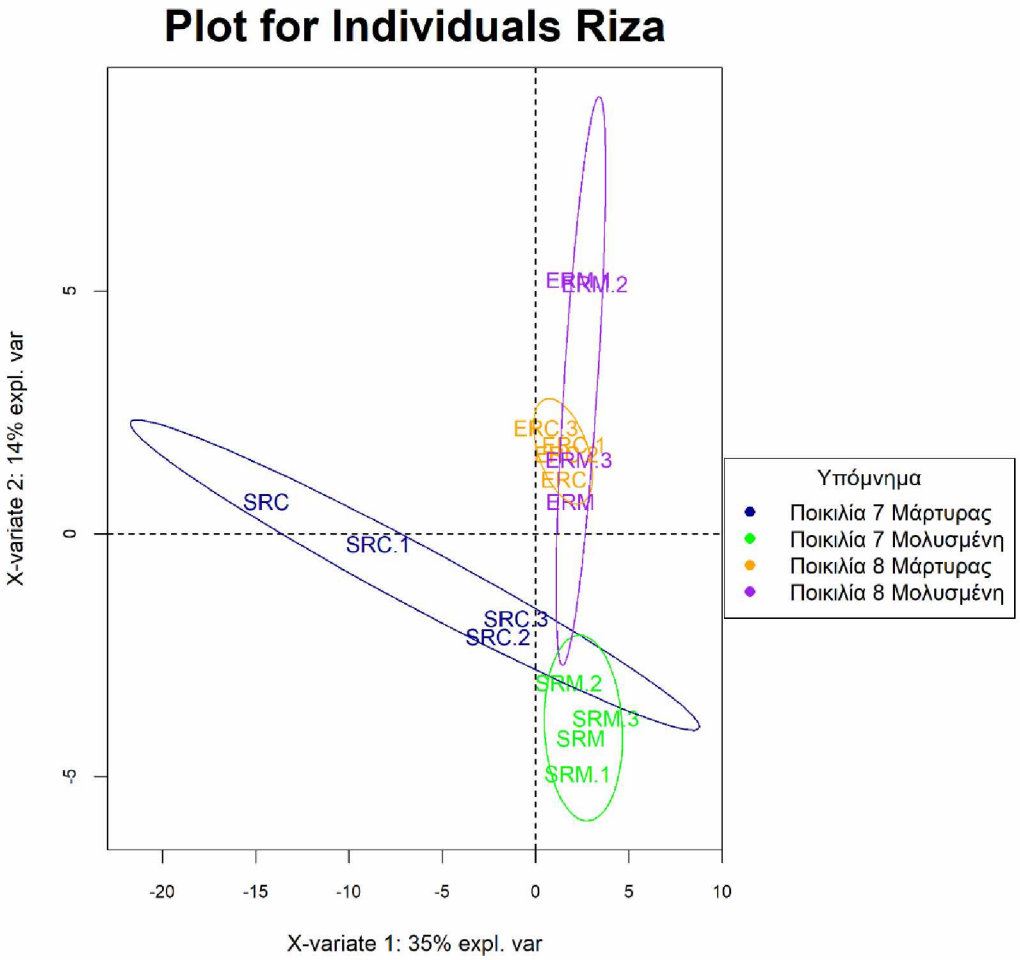
Εικόνα 42. Θηκόγραμμα για την glycerol-1-phosphate για τον βλαστό

Υψηλή συγκέντρωση της glycerol-1-phosphate στο μολυσμένο δείγμα σε σχέση με το δείγμα του μάρτυρα.

3.2 Πόλυ-μεταβλητή Ανάλυση

Η πόλυ-μεταβλητή ανάλυση εξετάζει ταυτόχρονα όλους τους μεταβολίτες. Η μέθοδος αυτή συνεισφέρει στην εξέταση αντίστοιχων σχέσεων μεταξύ των μεταβολιτών. Παρακάτω παρουσιάζονται τρία διαγράμματα και δύο πίνακες για την αξιολόγηση της συνεισφοράς κάθε μεταβολίτη για τη διάκριση κάθε ομάδας.

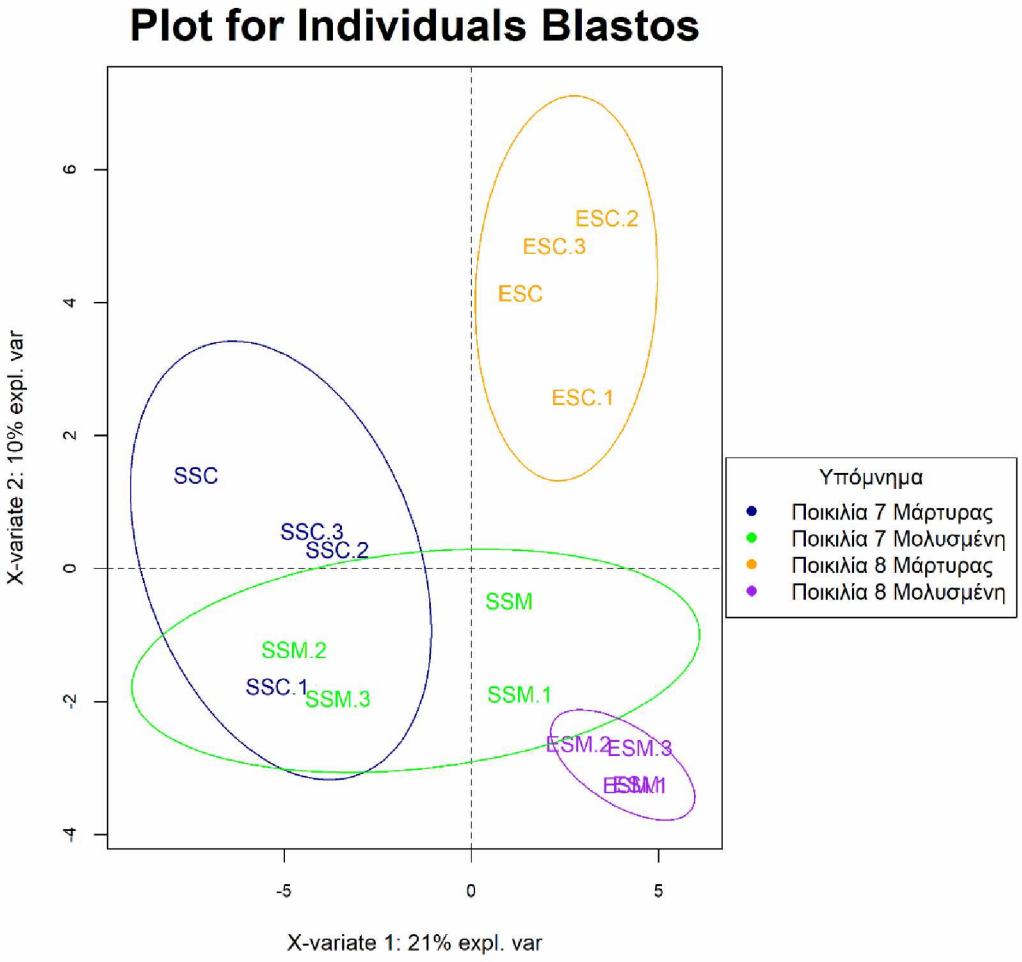
Γράφημα Διασποράς Δειγμάτων (Individual Plot)



Εικόνα 43. Γράφημα Διασποράς Ρίζας

Στην Εικόνα 43 αναπαρίστανται οι σχέσεις μεταξύ των δειγμάτων ανάλογα με τα χαρακτηριστικά τους και ο διαχωρισμός που υπάρχει μεταξύ τους.

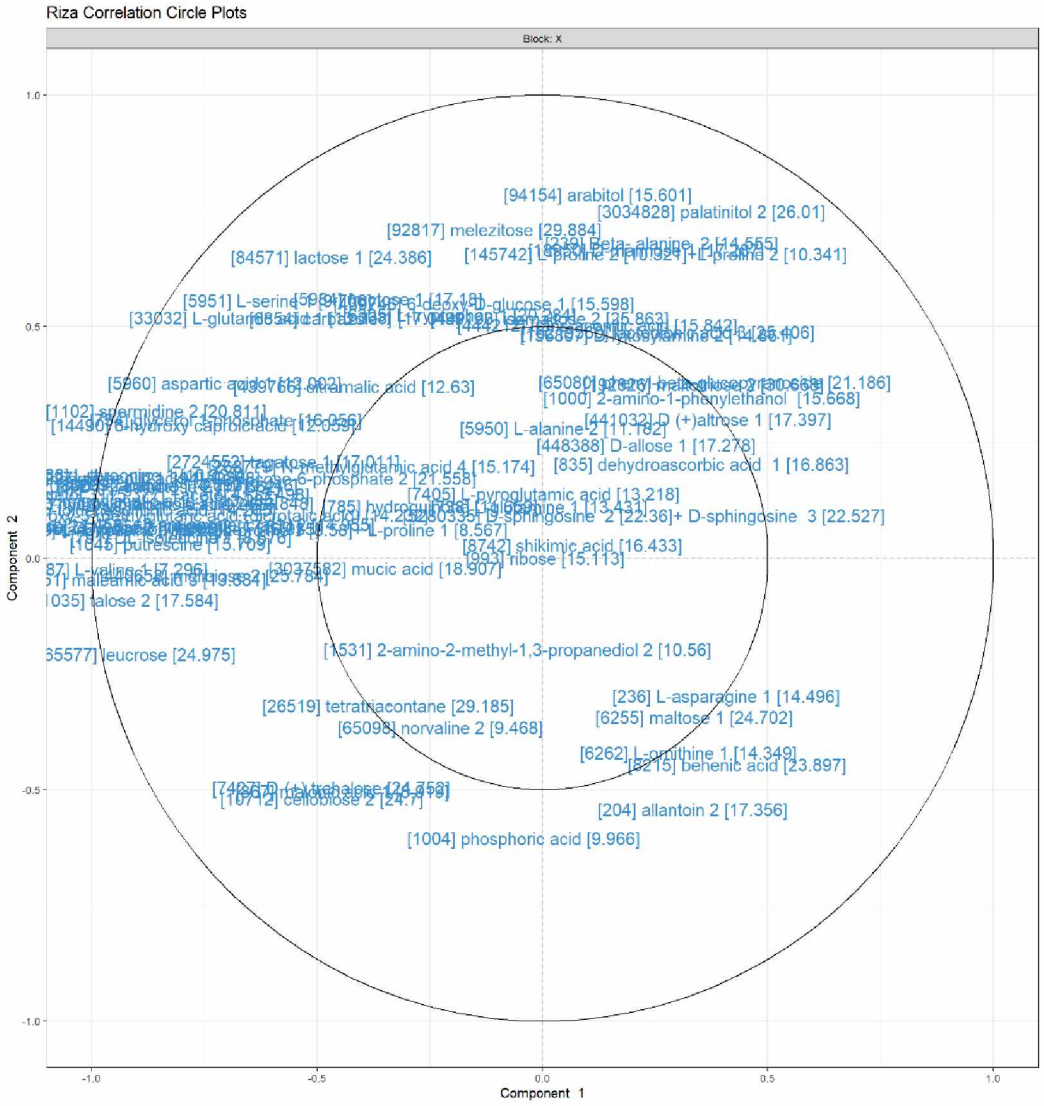
Παρατηρείται σχεδόν τέλειος διαχωρισμός της ποικιλίας 7 του δείγματος ρίζας που αποτέλεσε τον μάρτυρα (μπλε δείγμα), όπως επίσης και της ποικιλίας 7 που είναι μολυσμένη (πράσινο δείγμα) μεταξύ των άλλων δειγμάτων στον άξονα των x΄x (πρώτο component). Το ποσοστό 35% και 14% στον άξονα x΄x και y΄y αντίστοιχα εξηγεί την παραλλακτικότητα μεταξύ των ποικιλιών για το 1ο component και για το 2ο component αντίστοιχα. Στο σύνολο εξηγείται το 49% της παραλλακτικότητας.



Εικόνα 44. Γράφημα Διασποράς Βλαστού

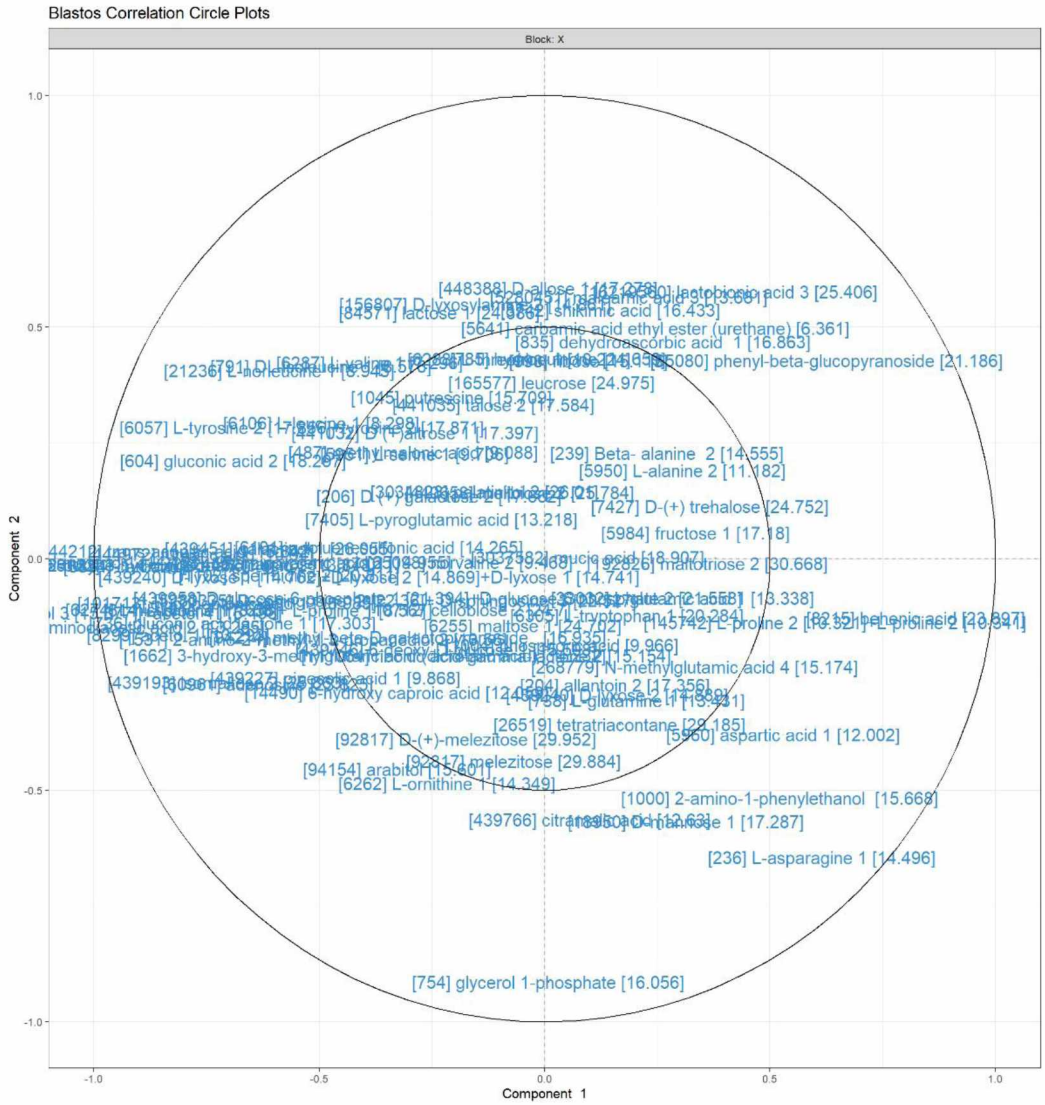
Στην Εικόνα 44 παρατηρείται σχεδόν τέλειος διαχωρισμός μεταξύ των δειγμάτων της ποικιλίας 8 τόσο του μολυσμένου όσο και του μάρτυρα (μωβ και πορτοκαλί δείγμα αντίστοιχα). Το 1ο component εξηγεί κατά 21% την παραλλακτικότητα μεταξύ των δειγμάτων και το 2ο component κατά 10%. Συνολικά εξηγείται το 31% της παραλλακτικότητας.

Γράφημα Μεταβλητών Κυκλικής Συσχέτισης (Correlation Circle Plot)



Εικόνα 45. Γράφημα Κυκλικής Συσχέτισης Ρίζας

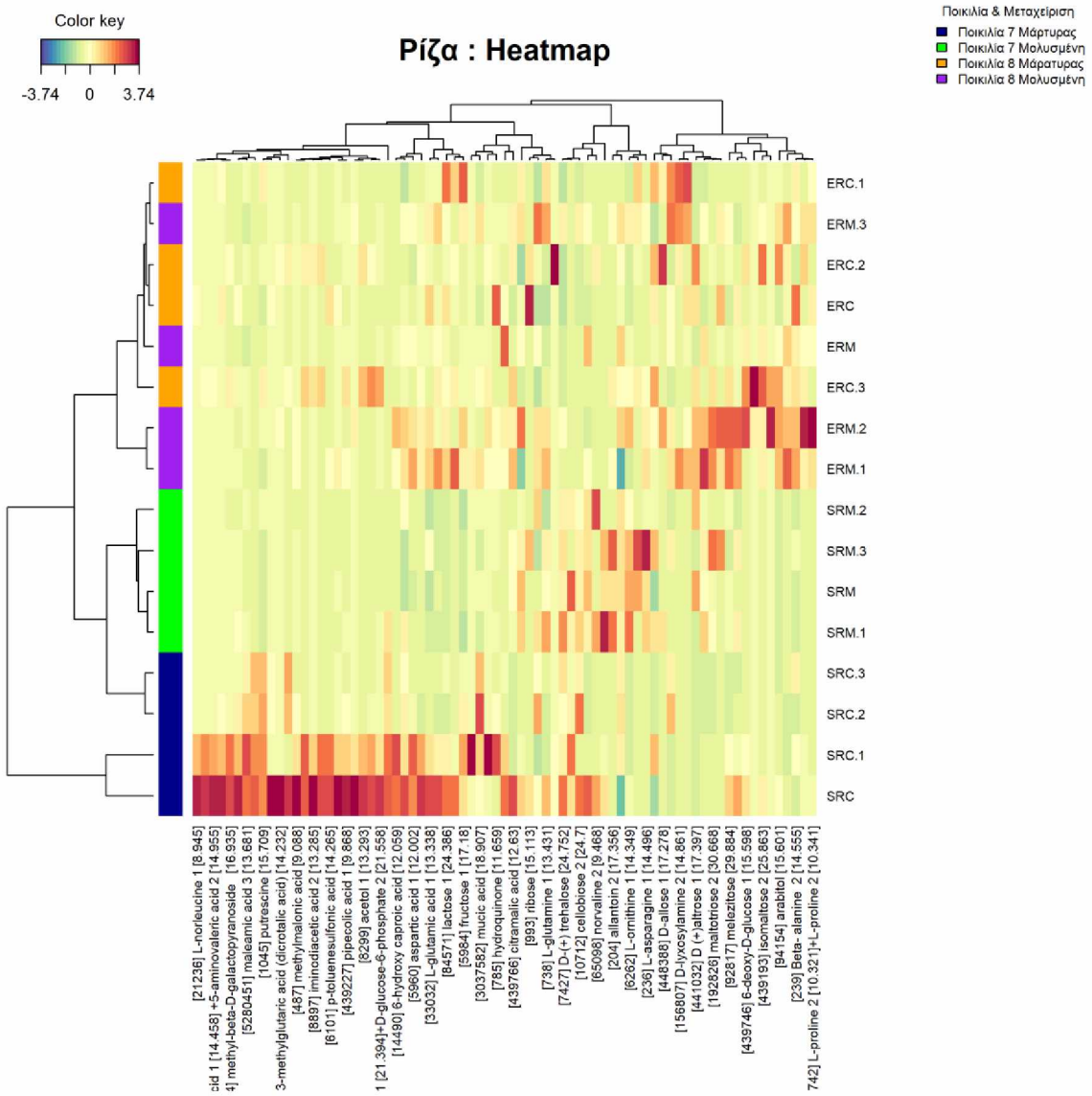
Στην Εικόνα 45 οι μεταβολίτες που βρίσκονται στον εξωτερικό κύκλο , δηλαδή η arabitol , palatinitol 2 , melezitose, lactose 1, το phosphoric acid και η allantoin 2 , συνεισφέρουν περισσότερο στον διαχωρισμό των δειγμάτων. Οι μεταβολίτες που δημιουργούν cluster έχουν παρόμοιες ιδιότητες. Τέλος η arabitol και το phosphoric acid, δηλαδή σημεία που βρίσκονται αντιδιαμετρικά έχουν αντιστρόφως ανάλογη δράση (όσο αυξάνει η arabitol τόσο μειώνεται το phosphoric acid).



Εικόνα 46. Γράφημα Κυκλικής Συσχέτισης Βλαστού

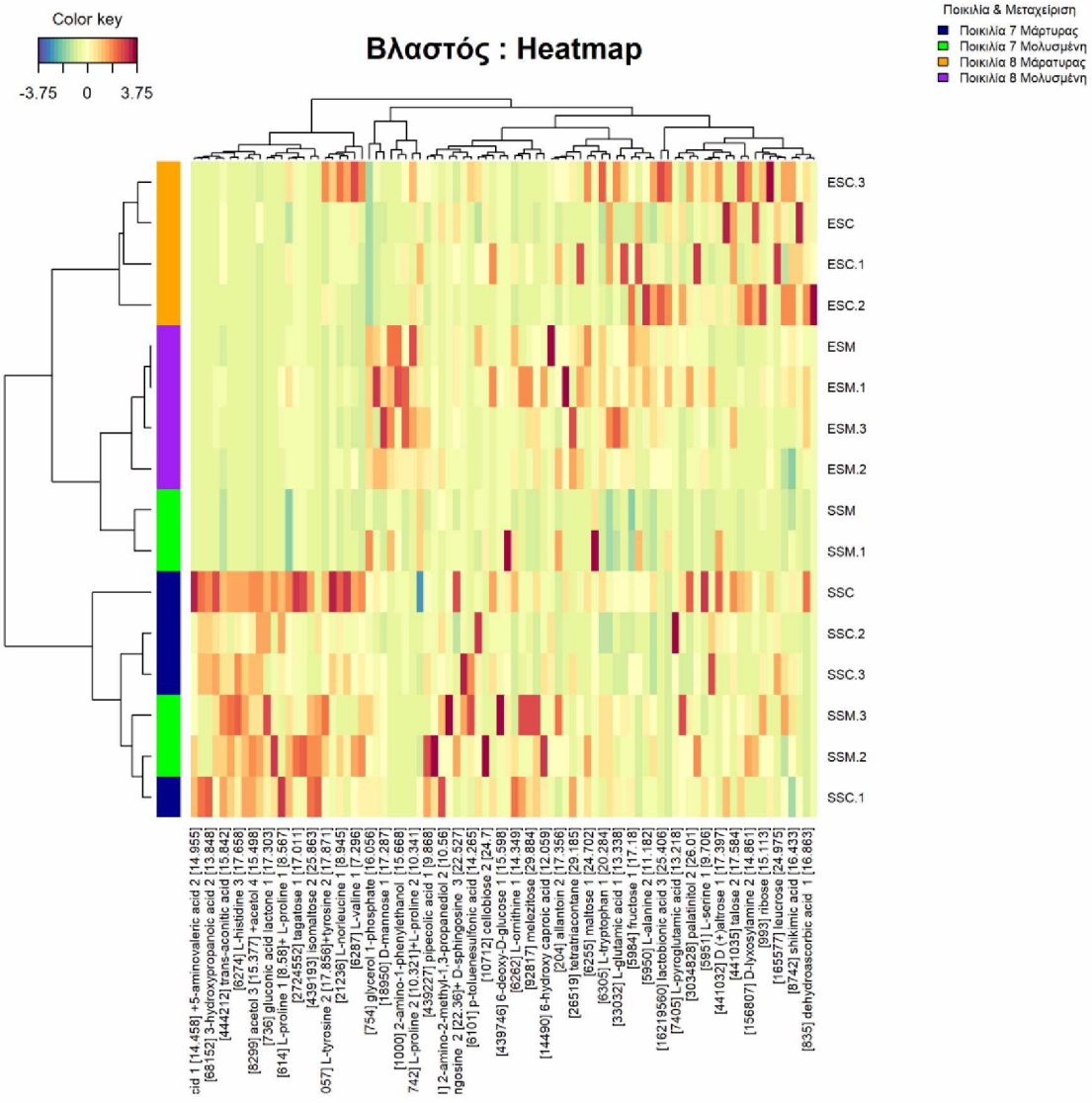
Στην Εικόνα 46 παρατηρείται ότι η glycerol 1-phosphate, το gluconic acid 2, η L-asparagine 1 και η 2-amino-1-phenylethanol συνεισφέρουν στον διαχωρισμό των δειγμάτων. Επίσης η L-asparagine 1 και η 2-amino-phenylethanol έχουν παρόμοιες ιδιότητες.

Θερμικός Χάρτης (Heatmap)



Εικόνα 47. Θερμικός Χάρτης Ρίζας

Στην Εικόνα 47 απεικονίζεται ο θερμικός χάρτης της ρίζας. Ο θερμικός χάρτης ερμηνεύεται συμπληρωματικά με το γράφημα διασποράς δειγμάτων. Παρατηρείται ότι υπάρχει σαφής διαχωρισμός του πράσινου και του μπλε δείγματος δηλαδή της ποικιλίας 7 τόσο του δείγματος μάρτυρα όσου και του μολυσμένου. Μεταβολίτες με όμοια χαρακτηριστικά δημιουργούν clusters και συμβολίζονται με το ίδιο χρώμα. Οι μεταβολίτες που συμβολίζονται με κόκκινο χρώμα έχουν θετική επίδραση στον διαχωρισμό του δείγματος σύμφωνα με το color key του χάρτη. Μεγαλύτερη συνεισφορά φαίνεται να έχουν οι μεταβολίτες της ποικιλίας 7 του μάρτυρα.



Εικόνα 48. Θερμικός Χάρτης βλαστού

Στην Εικόνα 48 απεικονίζεται ο θερμικός χάρτης του βλαστού. Παρατηρείται σαφής διαχωρισμός του πορτοκαλί και μωβ δείγματος, δηλαδή της ποικιλίας 8 τόσο του δείγματος μάρτυρα όσο και του μολυσμένου. Οι μεταβολίτες με κόκκινο χρώμα έχουν μεγαλύτερη συνεισφορά στον διαχωρισμό των δειγμάτων.

Πίνακες VIP

Το vip-score είναι το μέτρο σημαντικότητας κάθε μεταβολίτη στο PLS-DA μοντέλο. Οι μεταβολίτες ταξινομούνται κατά αύξουσα τιμή vip-score. Οι μεταβολίτες με vip-score >1.5 είναι οι σημαντικοί για την πόλυ-μεταβλητή ανάλυση.

Μεταβολίτες	value(comp = 1)
[94214] methyl-beta-D-galactopyranoside [16.935]	1.444408992
[6287] L-valine 1 [7.296]	1.511357072
[3037582] mucic acid [18.907]	1.551904909
[440658] melibiose 2 [25.784]	1.656171465
[165577] leucrose [24.975]	1.864334709
[5280451] maleamic acid 3 [13.681]	1.966836303
[1045] putrescine [15.709]	2.065986688
[441035] talose 2 [17.584]	2.083476844

Πίνακας 5. vi -score για comp = 1 στη ρίζα

Μεταβολίτες	value(comp=2)
[239] Beta- alanine 2 [14.555]	1.478541381
[94154] arabitol [15.601]	1.59614273
[3034828] palatinitol 2 [26.01]	1.794112748

Πίνακας 6. vi-score για comp = 2 στη ρίζα

Μεταβολίτες	value(comp = 1)
[65080] phenyl-beta-glucopyranoside [21.186]	1.48650773
[444212] trans-aconitic acid [15.842]	1.527641127
[138] 5-aminovaleric acid 1 [14.458] +5-aminovaleric acid 2 [14.955]	1.549949926
[8215] behenic acid [23.897]	1.63912423
[614] L-proline 1 [8.58]+ L-proline 1 [8.567]	1.693068919
[145742] L-proline 2 [10.321]+L-proline 2 [10.341]	1.696359097
[6854] carbazole 1 [17.044]	1.74651518
[6274] L-histidine 3 [17.658]	1.75880112
[8897] iminodiacetic acid 2 [13.285]	1.908175297
[68152] 3-hydroxypropanoic acid 2 [13.848]	1.957769005
[444972] fumaric acid [10.94]	1.993056863
[8299] acetol 3 [15.377] +acetol 4 [15.498]	2.014430413
[8066] 2-butyne-1,4-diol [9.446]	2.111611109

Πίνακας 7. vi-score για comp = 1 στον βλαστό

Μεταβολίτες	value(comp = 2)
[1000] 2-amino-1-phenylethanol [15.668]	1.496427215
[5280451] maleamic acid 3 [13.681]	1.645010057
[18950] D-mannose 1 [17.287]	1.671328793
[236] L-asparagine 1 [14.496]	1.869945003
[754] glycerol 1-phosphate [16.056]	2.210403799

Πίνακας 8. vi-score για comp = 2 στον βλαστό

4. Συζήτηση

Πιθανούς βιοδείκτες αποτελούν οι μεταβολίτες που είναι κοινοί στην univariate και τη multivariate ανάλυση.

Για την ρίζα κοινοί μεταβολίτες που μπορεί να είναι πιθανοί βιοδείκτες είναι : palatinol 2, talose 2, maleamic acid 3, putrescine, leucrose , arabitol, L-ornithine 1. Για τον βλαστό κοινοί μεταβολίτες που μπορεί να αποτελούν πιθανούς βιοδείκτες είναι : L-proline 2 + L-proline 2, L-asparagine 1, acetol 3 + acetol 4, fumaric acid, 3-hydropropanoic acid 2, glycerol 1- phosphate, 2-butyne-1,4-diol, iminodiacetic acid 2.

Σε μεταγενέστερο χρόνο μπορούν να πραγματοποιηθούν πειράματα με μεγαλύτερο δείγμα για την προαγωγή των πιθανών βιοδεικτών.

5. Βιβλιογραφία

Ashraf, M. A., Iqbal, M., Rasheed, R., Husain, I., Riaz, M., & Arif, M. S. (2018). Environmental Stress and Secondary Metabolites in Plants: An Overview. In *Plant metabolites and regulation under environmental stress* (pp. 153–167). essay, Academic Press.

Bevans, R. (2021, January 7). *An introduction to the TWO-WAY ANOVA*. Scribbr. Retrieved September 10, 2021, from https://www.scribbr.com/statistics/two-way-anova/

Cahota, R. K., Sharma, T. R., & Sharma, S. K. (2019). Conventional Genetic Manipulations. In *Lentils: Potential resources for enhancing genetic gains* (pp. 43–55). essay, Academic Press.

Dodge, Y. (2008). *The concise encyclopedia of statistics: With 247 tables* (2008 ed.). New York: Springer. doi:https://doi.org/10.1007/978-0-387-32833-1

DeepAI. (2019, May 17). *Univariate analysis*. DeepAI. Retrieved September 14, 2021, from https://deepai.org/machine-learning-glossary-and-terms/univariate-analysis

Jolliffe I.T. (2002) Principal Component Analysis. Springer Series in Statistics, Springer, New York.

Korthauer, K., Kimes, P.K., Duvallet, C. *et al.* A practical guide to methods controlling false discoveries in computational biology. *Genome Biol* **20,** 118 (2019). https://doi.org/10.1186/s13059-019-1716-1

Lawson, J. (2015). Split-Plot Designs. In *Design and analysis of experiments with r* (p. 307). essay, CRC Press.

Le Cao, K.A, Rohart, F., Gonzalez, I.,Dejean, S. with key contributors Benoit Gautier , Bartolo, F., contributions from Pierre Monget, Coquery, J. , Yao, F. and Liquet B.(2016). mixOmics: Omics Data Integration Project. R package version 6.1.1. https://CRAN.R-project.org/package=mixOmics

Nleya, T., Vandenberg, A., & Walley, F. L. (2004). LENTIL | Agronomy. In *Encyclopedia of grain science* (pp. 150–157). essay, Elsevier Academic Press.

Ramachandran, K. M., & Tsokos, C. P. (2015). Chapter 9 - Design of Experiments. In *Mathematical statistics with applications in r* (Second, pp. 459–494). essay, Academic Press, imprint of Elsevier.

Shah, K. (2021, April 19). *Exploratory analysis: Univariate, Bivariate, and multivariate analysis*. Analytics Vidhya. Retrieved September 21, 2021, from https://www.analyticsvidhya.com/blog/2021/04/exploratory-analysis-using-univariate-bivariate-and-multivariate-analysis-techniques/

Statistics Solutions. (2013). Confirmatory Factor Analysis . Retrieved September 21, 2021, from https://www.statisticssolutions.com/academic-solutions/resources/directory-of-statistical-analyses/confirmatory-factor-analysis/

Szymańska, E., Saccenti, E., Smilde, A.K. *et al.* Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics* **8,** 3–16 (2012). https://doi.org/10.1007/s11306-011-0330-3

The GNU operating system and the free software movement. [A GNU head] . (n.d.). Retrieved September 21, 2021, from http://www.gnu.org/

What is r? R. (n.d.). Retrieved September 21, 2021, from https://www.r-project.org/about.html

Why multivariate methods? mixomics. (n.d.). Retrieved September 14, 2021, from http://mixomics.org/

Zaynab, M., Fatima, M., Sharif, Y., Zafar, M. H., Ali, H., & Khan, K. A. (2019). Role of primary metabolites in plant defense against pathogens. *Microbial Pathogenesis*, *137*, 1–4. https://doi.org/10.1016/j.micpath.2019.103728