Πανεπιστήμιο Κρήτης
Σχολή Θετικών και Τεχνολογικών Επιστημών
Τμήμα Χημείας

Πτυχιακή Εργασία

Έλεγχος ποιότητας σε μελέτες μεταβολομικής με υγρή χρωματογραφία συζευγμένη με χρωματογραφία μάζας. Μελέτη απόδοσης αλγορίθμων επεξεργασίας δεδομένων.

Ονοματεπώνυμο φοιτήτριας: Καραπιδάκη Ανθή (2641)
Υπεύθυνη Καθηγήτρια: Χατζηιωάννου Αναστασία-Χρυσοβαλάντου
Δεύτερος Αξιολογητής: Σπύρος Απόστολος

Ηράκλειο, 2024

Ευχαριστίες:

Με την ολοκλήρωση της πτυχιακής μου εργασίας, έπειτα από ένα εξάμηνο εντατικής εργασίας και έρευνας στο εργαστήριο Ενόργανης Χημικής Βιοανάλυσης του τομέα Αναλυτικής Χημείας του τμήματος Χημείας, θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες σε όλους όσους συνέβαλλαν στην εκπόνησή της, προσφέροντας τόσο πολύτιμες γνώσεις αλλά και αμέριστη στήριξη καθ' όλη τη διάρκεια της.

Πρώτα απ' όλα, θα ήθελα να εκφράσω την ειλικρινή εκτίμησή μου στην επιβλέπουσα καθηγήτρια μου κ. Χρυσοβαλάντου- Αναστασία Χατζηιωάννου για τη χορήγηση αυτής της εξαιρετικής ευκαιρίας, την εμπιστοσύνη που μου έδειξε εξ' αρχής, την επιστημονική της καθοδήγηση, τις εποικοδομητικές υποδείξεις της, το χρόνο της και το αμείωτο ενδιαφέρον που έδειξε από την αρχή μέχρι το τέλος. Οι μέθοδοι που χρησιμοποιήθηκαν, ο τρόπος σκέψης και οργάνωσης των πειραμάτων, καθώς επίσης και η αντιμετώπιση των προβλημάτων, καλλιεργήθηκαν μέσα από την καθοδήγηση της. Είμαι λοιπόν ευγνώμων για τη βοήθεια, τις γνώσεις και τις συμβουλές της καθ' όλη τη διάρκεια της εργασίας μου, που συνέβαλαν στην απόκτηση σημαντικών εφοδίων για την μελλοντική μου ερευνητική σταδιοδρομία.

Ακόμα, ιδιαίτερες ευχαριστίες θα ήθελα να απευθύνω στον καθηγητή κ. Απόστολο Σπύρο ο οποίος προθυμοποιήθηκε να αναλάβει τη διόρθωση της παρούσας πτυχιακής εργασίας.

Φυσικά δεν θα μπορούσα να παραλείψω να ευχαριστήσω τις φίλες μου Χριστίνα Τ., Χριστίνα Φ., Χαρά, Έφη, Κέτι, Μαρίνα, Μυρτώ και Κατερίνα για την ψυχολογική υποστήριξη και συμπαράσταση σε όλη αυτή την προσπάθεια.

Τέλος, αλλά όχι λιγότερο σημαντικό, θα ήθελα εκφράσω την ευγνωμοσύνη μου στην οικογένειά μου για όλη τη στήριξη, την ενθάρρυνσή και την κατανόησή τους, σε όλη τη διάρκεια των σπουδών μου. Είμαι ειλικρινά ευγνώμων για την αγάπη και τις θυσίες τους οι οποίες μου επέτρεψαν να αφοσιωθώ σε αυτή την ερευνητική προσπάθεια.

Πίνακας περιεχομένων

Περίληψη: **4**

Εισαγωγή: **6**

Το πεδίο των -OMICS: 7

Μεταβολομική: 8

Στοχευμένη Μεταβολομική: 9

Μη- Στοχευμένη Μεταβολομική: 9

LC- HRMS: 10

MetaboAnalyst: 12

Αποδεκτές μορφές ακατέργαστων δεδομένων από MetaboAnalyst 5.0: 13

Profinder: 14

Προκατεργασία δεδομένων μη- στοχευμένης μεταβολομικής ανάλυσης: 14

Δείγματα ελέγχου ποιότητας (Quality control samples (QC)): 18

Λευκά δείγματα (Blank samples): 20

Πειραματική Διαδικασία: **21**

Πειραματικός σχεδιασμός και προκατεργασία δειγμάτων 21

Συνθήκες ανάλυσης UHPLC-QTOF-MS 22

Προκατεργασία αποτελεσμάτων 22

Στατιστική ανάλυση 24

Προσπάθεια αφαίρεσης θορύβου: 25

Διόρθωση του φαινομένου παρτίδας: 25

Σύγκριση αποτελεσμάτων MetaboAnalyst με τα αποτελέσματα του λογισμικού της AGILENT: 26

Λογισμικά: 26

Αποτελέσματα: **27**

Συμπεράσματα: **65**

Βιβλιογραφία: **69**

Περίληψη:

Στην σύγχρονη εποχή, το ενδιαφέρον έχει στραφεί προς την ολιστική ανάλυση και την συλλογή όσο το δυνατόν περισσότερων πληροφοριών από κάθε δείγμα (συνολικό «αποτύπωμα» του δείγματος). Με τον όρο μεταβολομική εννοούμε την ολοκληρωμένη και ταυτόχρονη μελέτη πληθώρας μεταβολιτών ενός δείγματος. Αυτή η μη στοχευμένη μεταβολομική ανάλυση οδηγεί συνήθως σε μεγάλα σύνολα δεδομένων, η ανάλυση των οποίων αποτελεί σημαντικό κομμάτι της ροής εργασίας τέτοιου είδους μελετών.

Η παρούσα εργασία έχει ως στόχο, αρχικά, την χρήση και τον έλεγχο των δυνατοτήτων ενός διαδικτυακού εργαλείου για την προκατεργασία και την ανάλυση ακατέργαστων δεδομένων μεταβολομικής, του MetaboAnalyst. Επιπλέον, αποσκοπεί στη σύγκριση των αποτελεσμάτων της ανάλυσης του MetaboAnalyst με αυτά που προκύπτουν από την ανάλυση με ένα εμπορικά διαθέσιμο λογισμικό της εταιρείας AGILENT το οποίο έχει χρησιμοποιηθεί κατά κόρον σε τέτοιου είδους αναλύσεις.

Τα δείγματα τα οποία επεξεργάστηκαν στην παρούσα εργασία είναι δείγματα ελέγχου ποιότητας (QC) καθώς και λευκά δείγματα (Blank) τα οποία είχαν αναλυθεί με την μέθοδο UPLC-HRMS σε μελέτη μεγάλης κλίμακας που διήρκησε περίπου 2 μήνες.

Συμπερασματικά, τα αποτελέσματα που προκύπτουν τόσο μεταξύ των διάφορων αναλύσεων του MetaboAnalyst, όσο και μεταξύ των δύο μελετούμενων λογισμικών διαφέρουν αρκετά μεταξύ τους. Επομένως, μπορεί να εξαχθεί το συμπέρασμα πως η επεξεργασία των ακατέργαστων δεδομένων σε μελέτες μεταβολομικής είναι ένα καίριο και αναπόσπαστο κομμάτι της ανάλυσης. Διαφορετικές επιλογές στο συγκεκριμένο κομμάτι της ανάλυσης μπορούν να οδηγήσουν σε πολύ διαφορετικά αποτελέσματα, τα οποία θα επηρεάσουν τη μετέπειτα στατιστική ανάλυση.

Abstract:

In the modern period, emphasis has turned toward holistic analysis and maximizing the collected information from each sample (sample's "fingerprint"). The term metabolomics refers to the comprehensive study of all metabolites in a given biological sample. Metabolomic analysis typically results in large datasets therefore, and the extraction of meaningful information out of these data sets is a significant part of the workflow.

The primary goal of this thesis was to utilize and assess the capabilities of an online tool for the preprocessing and analysis of unprocessed metabolomic data, namely MetaboAnalyst. Additionally, we compared the results of the MetaboAnalyst analysis with those obtained from the analysis using a commercially available software from AGILENT, which has been extensively used in such analyses. The samples processed in this work are quality control (QC) samples and blank samples, which were analyzed using the LC-HRMS method in a large-scale study lasting multiple weeks.

In conclusion, the results obtained from the various algorithms within MetaboAnalyst, as well as the analysis conducted in MetaboAnalyst and Agilent software exhibit significant differences. Consequently, the inference can be drawn that the preprocessing of raw data in metabolomic studies is a crucial and integral component of the analysis. Varied choices in this particular aspect of the analysis can lead to entirely different outcomes, thereby influencing dramatically the subsequent statistical analyses.

Εισαγωγή:

Ο ερευνητικός κλάδος της Αναλυτικής Χημείας σε μεγάλο βαθμό ασχολείται με την ποσοτική και ποιοτική ανάλυση γνωστού ή αγνώστου δείγματος με σκοπό την ανίχνευση των ενώσεων από τις οποίες αποτελείται και κατ΄ επέκταση την άντληση σημαντικών πληροφοριών για το ίδιο το δείγμα. Στα πρώτα βήματα ανάπτυξης της αναλυτικής χημείας μπορούσε να ταυτοποιηθεί ένας περιορισμένος αριθμός ενώσεων σε κάθε ανάλυση. Σχεδόν ολόκληρη η βιβλιογραφία αναφέρεται σε ειδικές μεθόδους με ζητούμενα χαρακτηριστικά την ευαισθησία, την αποτελεσματικότητα, την ειδικότητα και την εκλεκτικότητα. Στόχος μας ήταν η μέτρηση μιας ένωσης ή μίας ομάδας ενώσεων χωρίς παρεμποδίσεις από όλα τα υπόλοιπα συστατικά του δείγματος. Ένας από τους βασικούς στόχους για την εξέλιξη της αναλυτικής χημείας ήταν η ανάπτυξη κατάλληλων τεχνικών οι οποίες θα επέτρεπαν την ταυτόχρονη μελέτη πολλών χημικών ενώσεων σε ένα δείγμα. Ο Roger Williams ήταν ο πρώτος που, στα τέλη της δεκαετίας του 1940, έκανε λόγο για την έννοια του «Μεταβολομικού αποτυπώματος» και πως αυτό μάλλον θα μπορεί να μελετηθεί στα βιολογικά υγρά[1]. Η πρώτη αναλυτική μέθοδος που χρησιμοποιήθηκε για την απόδειξη της συγκεκριμένης θεωρίας ήταν η χρωματογραφία χάρτου. Αργότερα, η ραγδαία ανάπτυξη της τεχνολογίας επέτρεψε πέρα από ποιοτική ανάλυση και ποσοτικές μετρήσεις, ενώ η ανάπτυξη των ηλεκτρονικών υπολογιστών και της πληροφορικής έπαιξε καθοριστικό ρόλο στην ανάπτυξη των φασματοσκοπικών τεχνικών και κατ΄ επέκταση της ολιστικής προσέγγισης της Αναλυτικής Χημείας, δηλαδή την ανάλυση αγνώστων δειγμάτων με στόχο μία «ολική χαρτογράφηση». Πλέον, κατέστη δυνατόν όχι μονάχα να μετρήσουμε ορισμένες από τις ενώσεις ενός δείγματος, αλλά να λάβουμε ένα συνολικό «αποτύπωμα» του δείγματος. Στην πορεία της ολιστικής ανάλυσης συλλέγονται όσες περισσότερες πληροφορίες είναι δυνατόν και στη συνέχεια τα δεδομένα αναλύονται με προηγμένα υπολογιστικά εργαλεία όπως η στατιστική ανάλυση πολλαπλών μεταβλητών.

Οι βασικές αναλυτικές τεχνικές που έχουν βρει εφαρμογή στην ολιστική ανάλυση είναι κυρίως η φασματοσκοπία πυρηνικού συντονισμού (NMR), ο συνδυασμός υγρής χρωματογραφίας με φασματομετρία μαζών (LC-MS), η αέρια χρωματογραφία συζευγμένη με φασματοσκοπία μάζας GC-MS και η συζευγμένη αέρια χρωματογραφία (GC x GC). Το κύριο πεδίο εφαρμογής των ολιστικών τεχνολογιών είναι η χαρτογράφηση μεταβολιτών σε φυτά, τρόφιμα, περιβαλλοντικά δείγματα και κυρίως βιολογικά δείγματα. Το πεδίο αυτό, επίσης γνωστό ως μεταβολομική (metabolomics/metabonomics), είναι αναπόσπαστο τμήμα του ευρύτερου πεδίου της

βιολογίας συστημάτων (systems biology). Η σπουδαιότητα του έγκειται κυρίως στο γεγονός ότι καθορίζει τη βιολογική λειτουργία ενός οργανισμού, την ποιότητα ενός προϊόντος, την καταλληλόλητα ενός τροφίμου. Παράλληλα με τη στοχευμένη ανάλυση, απαιτούνται πια και ολιστικές αναλύσεις στην τοξικολογική έρευνα, στον έλεγχο μεταβολισμού φαρμάκων, στην εύρεση βιοδεικτών, στη διασφάλιση ποιότητας και ασφάλειας. Ελεγκτικοί οργανισμοί όπως το Food and Drug Administration των ΗΠΑ εργάζονται προς την κατεύθυνση της οριοθέτησης του νέου αυτού πεδίου και του καθορισμού προτύπων[2].

<u>Το πεδίο των -OMICS:</u>

Το πεδίο των -omics, είναι ένας τομέας που αποσκοπεί στην ποιοτική και ποσοτική ανάλυση πληθώρας μορίων σε βιολογικά δείγματα. Μέχρι στιγμής, η μεταβολική, η γονιδιωματική, η μεταγραφομική και η πρωτεομική αναγνωρίζονται ως κρίσιμα μέρη της συστημικής βιολογίας. Τα τέσσερα omics* χτίζουν επίσης μια γέφυρα για να κατανοήσουμε πλήρως έναν ζωντανό οργανισμό από το μικροεπίπεδο έως το μακροεπίπεδο και να εξηγήσουμε ολόκληρη τη διαδικασία της ζωής από το στατικό μέρος των μικροσκοπικών μορίων του DNA έως το ενεργό μέρος της έκκρισης μικρών μορίων μεταβολιτών. Επομένως, κοινό χαρακτηριστικό των συγκεκριμένων πεδίων αποτελεί η πεποίθηση πως για την πλήρη κατανόηση ενός σύνθετου συστήματος απαιτείται η σφαιρική μελέτη του ως σύνολο και όχι ως μεμονωμένα χαρακτηριστικά. Οι ολιστικές μέθοδοι, σε αντίθεση με τις κλασσικές στοχευμένες αναλυτικές τεχνικές αποσκοπούν στην συγκέντρωση όσων περισσότερων δεδομένων είναι δυνατόν που αφορούν το 'όλο', τα δεδομένα αυτά πιθανώς να μας οδηγήσουν σε μία νέα υπόθεση, η οποία θα απαιτεί επιπλέον στοχευμένη πειραματική μελέτη και επιβεβαίωση.

Οι συγκεκριμένες αναλυτικές μέθοδοι βρίσκουν εφαρμογή στην ανίχνευση γονιδίων (γονιδιωματική), μεταγραφημάτων mRNA (μεταγραφομική), πρωτεϊνών (πρωτεομική) και μεταβολιτών (μεταβολομική). Το συγκεκριμένο πεδίο βρίσκει εφαρμογή σε πολλούς ερευνητικούς τομείς. Αρχικά, χρησιμοποιείται στην ιατρική για τον χαρακτηρισμό ασθενειών με σκοπό την καλύτερη πρόβλεψη της κλινικής πορείας των οργανισμών και την αξιολόγηση της αποτελεσματικότητας των υφιστάμενων ή υπό ανάπτυξη θεραπειών, καθώς και για την γενικότερη κατανόηση του μεταβολισμού και των αλληλεπιδράσεων που λαμβάνουν χώρα. Στην επιστήμη των τροφίμων, τα omics διαδραματίζουν σημαντικό ρόλο υπό το πρίσμα της βελτίωσης της ανθρώπινης διατροφής. Στον τομέα του περιβάλλοντος, οι μελέτες omics αποσκοπούν στην αξιολόγηση των μεταβολών που

ενδέχεται να υποστούν οι οργανισμοί μετά την έκθεση σε περιβαλλοντικούς παράγοντες. Το εύρος των εφαρμογών είναι εξαιρετικά μεγάλο, με πολυάριθμες εφαρμογές σε πληθώρα πεδίων[3].

Μεταβολομική:

Η μεταβολομική, η μελέτη δηλαδή του μεταβολικού αποτυπώματος ενός συστήματος κάτω από συγκεκριμένες συνθήκες, είναι ένας κλάδος της αναλυτικής χημείας ο οποίος βρίσκει εφαρμογή σε διάφορα ερευνητικά πεδία, όπως η ιατρική, η μικροβιολογία, η τοξικολογία κ.α. Η προσέγγιση αυτή εστιάζει κυρίως στην παρατήρηση μεταβολών στην συγκέντρωση των μεταβολιτών, δηλαδή μόρια χαμηλού μοριακού βάρους (50-1500 Da), που ως επί το πλείστων συμμετέχουν σε κάποιες βιολογικές διεργασίες ως υποστρώματα ή ενδιάμεσα προϊόντα ή προϊόντα βιολογικών μονοπατιών του υπό μελέτη οργανισμού. Η μεταβολομική προσφέρει ποιοτικές και ποσοτικές πληροφορίες για τους μεταβολίτες συμβάλλοντας έτσι στον καθορισμό μεταβολικών αλλαγών που σχετίζονται με καταστάσεις ενδιαφέροντος, όπως έκθεση σε φάρμακα και άλλες μεταλλαξιογόνες ουσίες κ.α. Οι μελέτες μεταβολομικής έχουν συνήθως ως στόχο την ολοκληρωμένη μελέτη όλων των μεταβολιτών σε βιολογικά δείγματα, ώστε να δημιουργήσουν ένα στιγμιότυπο της φυσιολογικής κατάστασης ενός συστήματος και να καθορίσουν πώς ένα εξωτερικό ή εσωτερικό ερέθισμα μπορεί να μεταβάλει τη βιολογία του συστήματος που μελετάται. Ένα από τα μεγαλύτερα πλεονεκτήματα των μελετών της μεταβολομικής έναντι των υπόλοιπων πεδίων -omics είναι ότι το μεταβόλωμα συνοψίζει την τελική απόκριση του κάθε οργανισμού. Επομένως, οι μεταβολές που έχουν συμβεί σε προηγούμενα στάδια της γονιδιωματικής έκφρασης αποτυπώνονται, ίσως και ενισχυμένες, στο μεταβόλωμα[4].

Σημαντικό περιορισμό στην ολική χαρτογράφηση των μεταβολιτών ενός οργανισμού θέτουν οι δυσκολίες που προκύπτουν εξαιτίας των αναλυτικών τεχνικών αλλά και των συνθηκών ανάλυσης που επιλέγονται, τόσο για την συλλογή, όσο και για την προκατεργασία των δειγμάτων. Οι πρώτες μεταβολομικές μελέτες πραγματοποιήθηκαν με τη χρήση της αέριας χρωματογραφίας συνδυασμένης με χρωματογραφία μάζας (GC-MS) και της φασματοσκοπίας πυρηνικού μαγνητικού συντονισμού (NMR). Στην ραγδαία ανάπτυξη των μεταβολομικών αναλύσεων σημαντικό ρόλο έπαιξε η ανάπτυξη και η εφαρμογή ειδικών λογισμικών ανάλυσης. Έτσι έγινε εφικτή η ερμηνεία του τεράστιου όγκου, σύνθετων δεδομένων που προκύπτουν από τις συγκεκριμένες μελέτες. Με τον τρόπο αυτόν προκύπτουν φάσματα τα οποία περιλαμβάνουν κορυφές που

αντιστοιχούν σε όλους τους μεταβολίτες των δειγμάτων, οι οποίοι βρίσκονται σε συγκεντρώσεις που ξεπερνούν τα όρια ανίχνευσης[5].

Στοχευμένη Μεταβολομική:

Με τον όρο στοχευμένη μεταβολομική περιγράφουμε την αναλυτική τεχνική κατά την οποία γίνεται η μέτρηση της ποσότητας ενός ή περισσοτέρων συγκεκριμένων μεταβολιτών σε βιολογικά δείγματα. Επομένως, η στοχευμένη μεταβολομική εστιάζει στον ακριβή και ποσοτικό προσδιορισμό προκαθορισμένου συνόλου μεταβολιτών. Στη στοχευμένη προσέγγιση, οι χημικές ιδιότητες των υπό εξέταση ενώσεων είναι γνωστές. Είναι πιο πιθανό η ανάλυση να επικεντρωθεί στον εντοπισμό και τον ποσοτικό προσδιορισμό επιλεγμένων μεταβολιτών ή κατηγοριών μεταβολιτών, όπως υποστρώματα ενός ενζύμου, άμεσα προϊόντα μιας πρωτεΐνης, ένας συγκεκριμένος τύπος ένωσης ή μέλη ενός συγκεκριμένου μονοπατιού. Όταν χρησιμοποιείται η στοχευμένη μεταβολωμική, η προετοιμασία του δείγματος βελτιστοποιείται με τη χρήση ισοτοπικά επισημασμένων προτύπων και σαφώς καθορισμένων παραμέτρων, με στόχο την ανάπτυξη μεθόδων που χαρακτηρίζονται από μεγάλη ακρίβεια, πιστότητα κι εκλεκτικότητα.Επιπλέον, η βελτιστοποιημένη προετοιμασία του δείγματος μειώνει την κυριαρχία μορίων υψηλής αφθονίας. Ως αποτέλεσμα, οι προκαθορισμένοι κατάλογοι μεταβολιτών υπό τον φακό των μεταβαλλόμενων φυσιολογικών καταστάσεων παρέχουν την δυνατότητα σύγκρισης μεταξύ ομάδων ελέγχου και πειραματικών ομάδων. Το σημαντικότερο μειονέκτημα των στοχευμένων προσεγγίσεων είναι η περιορισμένη κάλυψη του μεταβολώματος. Με τον όρο μεταβόλωμα, εννοούμε το σύνολο των μεταβολιτών χαμηλού μοριακού βάρους που παράγονται από τα κύτταρα κατά τη διάρκεια του μεταβολισμού. Ο συγκεκριμένος περιορισμός αυξάνει τον κίνδυνο να παραβλεφθεί η μεταβολική απόκριση που ενδιαφέρει[6].

Μη- Στοχευμένη Μεταβολομική:

Με τον όρο μη- στοχευμενη μεταβολομική περιγράφουμε τον κλάδο της μεταβολομικής κατά τον οποίο γίνεται προσπάθεια δημιουργίας ενός σφαιρικού μεταβολομικού προφίλ του οργανισμού από τον οποίο προέρχονται τα βιολογικά δείγματα που αναλύθηκαν. Λόγω του ολοκληρωμένου χαρακτήρα της, η μη στοχευμένη μεταβολωμική συνδυάζεται με προηγμένες χημειομετρικές τεχνικές. Χρησιμοποιούνται φασματομετρικές μέθοδοι υψηλής ανάλυσης ώστε να καταγράφουν φάσματα που αποτελούνται από πολυάριθμα

σήματα αναλυτών. Η μη στοχευμένη προσέγγιση προσφέρει την ευκαιρία της ανακάλυψης νέων στόχων.

Η κάλυψη του μεταβολώματος περιορίζεται μόνο από τις μεθοδολογίες προετοιμασίας του δείγματος και την εγγενή ευαισθησία και ειδικότητα της χρησιμοποιούμενης αναλυτικής τεχνικής. Ωστόσο, οι κύριες προκλήσεις αυτής της προσέγγισης έγκεινται στο μεγάλο όγκο δεδομένων και πληροφορίας που αυτές οι προσεγγίσεις δημιουργούν, καθώς και στις δυσκολίες που αφορούν τον εντοπισμό και το χαρακτηρισμό άγνωστων μικρών μορίων που διαφοροποιούνται στατιστικά σημαντικά μεταξύ των μελετούμενων συνθηκών. Τα δεδομένα της συγκεκριμένης τεχνικής περιλαμβάνουν φάσματα GC/LC-MS ή NMR τα οποία περιλαμβάνουν πληροφορίες για το μεταβολικό αποτύπωμα του οργανισμού από τον οποίο προέρχεται το δείγμα για την χρονική στιγμή κατά την οποία έγινε η λήψη του βιολογικού υγρού. Στη συνέχεια, σημαντικό βήμα αποτελεί η ταυτοποίηση των ληφθέντων σημάτων, που παρόλα τα τεράστια βήματα που έχουν γίνει για την πλήρη χαρτογράφηση και ταυτοποίηση του γονιδιώματος, υπάρχει ακόμα σημαντικό ποσοστό σημάτων τα οποία δεν έχουν ταυτοποιηθεί ακόμα[7,8].

LC- HRMS:

Τα δείγματα που χρησιμοποιήθηκαν στην συγκεκριμένη μελέτη είχαν πρώτα αναλυθεί με την μέθοδο υγρής χρωματογραφίας υπερυψηλής πίεσης συζευγμένη με φασματομετρία μάζας υψηλής διακριτικής ικανότητας και ακρίβειας (UPLC-HRMS). Η υγρή χρωματογραφία συνδυασμένη με φασματομετρία μάζας, είναι μία τεχνική που τα τελευταία χρόνια χρησιμοποιείται ευρέως για την ποιοτική και ποσοτική ανίχνευση διαφόρων ενώσεων σε ένα μεγάλο φάσμα υποστρωμάτων και τον ολοκληρωμένο έλεγχο πολύπλοκων δειγμάτων. Η συγκεκριμένη μέθοδος ενδείκνυται για την ανάλυση και την ανίχνευση μεγάλου πλήθους χημικών ουσιών με διαφορετικές φυσικοχημικές ιδιότητες και, ως εκ τούτου, χρησιμοποιείται ευρέως σε μελέτες μη- στοχευμένης μεταβολομικής. Από τις αναλύσεις αυτές προκύπτουν φάσματα τα οποία περιέχουν πλήθος κορυφών.

Οι προσδιοριζόμενοι αναλύτες περιγράφονται ως ζεύγη μάζας/φορτίου (m/z) - χρόνου κατακράτησης (retention time, rt). Αυτό συμβαίνει εξαιτίας της ιδιότητας του φασματογράφου μαζών να διαχωρίζει τα μόρια με βάση τον λόγο m/z μεταξύ των ιόντων. Από την συγκεκριμένη μέθοδο προκύπτουν τρισδιάστατα δεδομένα αφού παρέχονται πληροφορίες για τον χρόνο συγκράτησης κάθε μορίου μέσα στην στήλη, την μάζα αλλά και την ένταση των κορυφών. Συγκεκριμένα, σε κάθε σημείο του χρωματογραφήματος αντιστοιχεί και ένα πλήρες φάσμα μαζών.

Ο ιονισμός των ατόμων ή των μορίων που θέλουμε να ανιχνεύσουμε, δηλαδή η προσθήκη ή η αφαίρεση ηλεκτρονίων από τα ουδέτερα άτομα ή μόρια, πραγματοποιείται εντός του φασματομέτρου μάζας. Τα ιόντα αυτά μεταφέρονται στον αναλυτή μαζών, στον οποίον διαχωρίζονται με βάση τον λόγο μάζα/φορτίο, στη συνέχεια κατευθύνονται προς τον ανιχνευτή. Οι παραπάνω διεργασίες δημιουργούν θραύσματα των ενώσεων που περιέχουν τα δείγματα. Τα θραύσματα αυτά είναι δυνατόν να ταυτοποιηθούν και κατ΄ επέκταση να ταυτοποιηθούν και οι αρχικές ενώσεις που περιέχονταν στο δείγμα. Επομένως, όπως ήδη αναφέρθηκε η φασματομετρία μάζας επιτρέπει τη μέτρηση της αφθονίας φορτισμένων μορίων (ιόντων) σε ένα δείγμα. Οι αφθονίες προσδιορίζονται ως αριθμοί ιόντων για συγκεκριμένο λόγο μάζας προς φορτίο (m/z) και το μετρούμενο σήμα αναπαρίσταται ως φάσμα: εντάσεις κατά m/z. Όμως, πολλά ιόντα έχουν το ίδιο ή ένα πολύ παρόμοιο λόγο m/z καθιστώντας δύσκολη ή αδύνατη τη διάκρισή τους. Συνεπώς, η MS συχνά συνδυάζεται με μια δεύτερη τεχνολογία για το διαχωρισμό των αναλυτών με βάση άλλες ιδιότητες εκτός από τη μάζα τους (ή μάλλον το m/z). Συνήθεις επιλογές είναι η αέρια χρωματογραφία (GC) ή η υγρή χρωματογραφία (LC). Μια τέτοια διάταξη όπως η LC-MS εκτελεί σαρώσεις σε διακριτά χρονικά σημεία με αποτέλεσμα να προκύψει ένα σύνολο φασμάτων για ένα δεδομένο δείγμα, με δυνατότητα διαχωρισμού ενώσεων (ιόντων) με βάση το m/z σε συνδυασμό με τον χρόνο κατακράτησης.

Μεταξύ των διαφόρων διαθέσιμων τεχνολογιών, η φασματομετρία μάζας υψηλής ανάλυσης (HRMS) έχει δώσει μεγάλες υποσχέσεις. Τα όργανα HRMS, όπως τα συστήματα Orbitrap ή τα συστήματα χρόνου πτήσης (TOF) σε συνδυασμό με αέρια χρωματογραφία (GC) ή υγρή χρωματογραφία (LC), μπορούν να μετρήσουν ταυτόχρονα ένα ευρύ φάσμα ενδογενών και εξωγενών ενώσεων για να χαρακτηρίσουν τον μεταβολικό φαινότυπο ενός ατόμου, τις περιβαλλοντικές εκθέσεις και τις σχετικές βιολογικές αποκρίσεις. Η LC-HRMS είναι αναμφισβήτητα η πιο ευρέως χρησιμοποιούμενη πλατφόρμα, καθώς μπορεί να μετρήσει ένα ευρύ φάσμα ενώσεων με ελάχιστη προετοιμασία του δείγματος. Ωστόσο, η σχετική επεξεργασία και ανάλυση δεδομένων παραμένει ιδιαίτερα δύσκολη για τους περισσότερους ερευνητές[9].

Τα δεδομένα που προέκυψαν με την τεχνική της υγρής χρωματογραφίας συνδυασμένη με χρωματογραφία μάζας υψηλής ανάλυσης, πρέπει να αναλυθούν περαιτέρω προκειμένου να είναι ικανοί οι ερευνητές να εξάγουν συμπεράσματα από αυτά.

MetaboAnalyst:

Ήδη αναφέρθηκε στις προηγούμενες ενότητες ότι στις σύγχρονες μεταβολομικές αναλύσεις συνδυάζονται συχνά όργανα φασματομετρίας μάζας υψηλής ανάλυσης (HRMS), όπως η orbitrap , μαζί με μεθόδους χρωματογραφίας, συνήθως υγρής και αέριας (LC-HRMS και GC-HRMS) προκειμένου να προκύψουν δεδομένα υψηλής ευαισθησίας. Με τον τρόπο αυτόν, μπορεί να μετρηθεί ταυτόχρονα ένα ευρύ φάσμα ενδογενών και εξωγενών ενώσεων προκειμένου να χαρακτηριστεί ο βιολογικός φαινότυπος ενός ατόμου, οι εκθέσεις στις οποίες έχει εκτεθεί μέσω της διατροφής, του περιβάλλοντος ή των συνηθειών του αλλά και οι σχετικές βιολογικές αποκρίσεις του. Δυσκολίες προκύπτουν στο κομμάτι που αφορά την βιοπληροφορική, δηλαδή την επεξεργασία, την στατιστική ανάλυση και την λειτουργική ερμηνεία των δεδομένων. Κατά την τελευταία δεκαετία, έχουν αναπτυχθεί αρκετές ισχυρές υπολογιστικές σειρές για τη φασματική επεξεργασία, τον σχολιασμό και την ανάλυση. Τα περισσότερα από αυτά τα εργαλεία πρέπει να εγκατασταθούν τοπικά από τους χρήστες. Παραμένουν όμως και κάποιες online επιλογές εργαλείων όπως το MetaboAnalyst, το XCMS online, W4M, NOREVA κ.α. Ωστόσο, παραμένουν σημαντικά εμπόδια όσον αφορά τις ρυθμίσεις παραμέτρων, την υπολογιστική αποτελεσματικότητα, τις επιδράσεις παρτίδας και τις λειτουργικές ερμηνείες.

Το MetaboAnalyst είναι ένα ολοκληρωμένο και ελεύθερα προσβάσιμο λογισμικό, φιλικό προς τον χρήστη, που δημιουργήθηκε, μεταξύ άλλων, για να εκτελέσει την προ-επεξεργασία των δεδομένων που λαμβάνονται από όλες τις σύγχρονες αναλυτικές πλατφόρμες. Αναπτύχθηκε πρώτη φορά το 2009 και στη συνέχεια ενημερώθηκε το 2012 (MetaboAnalyst 2.0), το 2015 (MetaboAnalyst 3.0), το 2019 (MetaboAnalyst 4.0) και το 2021 (MetaboAnalyst 5.0), το οποίο χρησιμοποιήθηκε και στις αναλύσεις της συγκεκριμένης εργασίας. Οι εκδόσεις 1.0-3.0 του MetaboAnalyst σχεδιάστηκαν κυρίως για τη γενική στατιστική και λειτουργική ανάλυση δεδομένων στοχευμένης μεταβολωμικής.[10] Ξεκινώντας από την έκδοση v4.0, η ανάπτυξη του MetaboAnalyst μετατοπίστηκε σταδιακά προς την αντιμετώπιση πιο σύνθετων βιοπληροφορικών και στατιστικών αναγκών που προκύπτουν από τα παγκόσμια δεδομένα μεταβολομικής, συμπεριλαμβανομένης της επεξεργασίας ακατέργαστων φασμάτων, της λειτουργικής ανάλυσης και της ενοποίησης με άλλα δεδομένα -omics. Το MetaboAnalyst παρέχει ολοκληρωμένα διαδικτυακά εργαλεία για ανάλυση δεδομένων μεταβολομικής, τη στατιστική ανάλυση, το λειτουργικό σχολιασμό και την οπτικοποίηση των δεδομένων. Οι παροχές που προσφέρει το συγκεκριμένο λογισμικό έχουν συγκριθεί με τις

αντίστοιχες των υπόλοιπων online εργαλείων. Το MetaboAnalyst προσφέρει τα περισσότερα χαρακτηριστικά για στατιστική και λειτουργική ανάλυση- το XCMS online και το W4M παρέχουν πιο ολοκληρωμένη υποστήριξη για την επεξεργασία ακατέργαστων δεδομένων- ενώ το NOREVA είναι αφιερωμένο στην κανονικοποίηση δεδομένων μεταβολωμικής και στην αξιολόγηση της ποιότητας. Χαρακτηριστικό γνώρισμα της διαδικτυακής πλατφόρμας του MetaboAnalyst είναι η ευκολία χρήσης του, η οποία επιτρέπει σε ερευνητές με διαφορετικό υπόβαθρο να εκτελούν διάφορες σύνθετες εργασίες ανάλυσης δεδομένων σε ένα φιλικό προς τον χρήστη περιβάλλον. Για παράδειγμα, τα περισσότερα εργαλεία απαιτούν από τους χρήστες να ρυθμίζουν χειροκίνητα πολλαπλές παραμέτρους για να επιτύχουν ικανοποιητική απόδοση για την επεξεργασία ακατέργαστων φασμάτων LC-MS. Το MetaboAnalyst 5.0 παρέχει μια σε μεγάλο βαθμό αυτοματοποιημένη ροή εργασίας μέσω της αποτελεσματικής βελτιστοποίησης των παραμέτρων από τις προεπιλεγμένες ρυθμίσεις για κοινά όργανα LC-MS. Επιπλέον, τα αποτελέσματα της επεξεργασίας φασμάτων μπορούν να μεταφερθούν απευθείας σε άλλες συμβατές ενότητες εντός του MetaboAnalyst για βελτιωμένη στατιστική και λειτουργική ανάλυση[11,12] .

Παρόλα τα πολλά πλεονεκτήματα που εμφανίζει το MetaboAnalyst, παρουσιάζει και αρκετούς περιορισμούς. Αρχικά, λόγω των ιδιόκτητων μορφοτύπων και των μεγάλων μεγεθών αρχείων που παράγονται από τα όργανα LC-HRMS, το MetaboAnalyst δεν δέχεται επί του παρόντος ακατέργαστα φάσματα που προκύπτουν από τα συγκεκριμένα όργανα. Έτσι, τα ακατέργαστα δεδομένα πρέπει πρώτα να μετατραπούν σε κατάλληλη μορφή δεδομένων χρησιμοποιώντας είτε ένα εργαλείο μετατροπής που παρέχεται από τον προμηθευτή είτε άλλα εργαλεία, όπως το ProteoWizard. Επιπλέον, εξαιτίας του περιορισμένου χώρου αλλά και του μεγάλου αριθμού χρηστών αλλά και για να διασφαλιστεί ο γρήγορος κύκλος εργασιών, η δημόσια πλατφόρμα του MetaboAnalyst περιορίζει επί του παρόντος την επεξεργασία ακατέργαστων φασμάτων έως 200 δείγματα ανά εργασία, το οποίο είναι συνήθως επαρκές για κοινές μεταβολωμικές μελέτες αλλά καθιστά το συγκεκριμένο εργαλείο ακατάλληλο για έργα μεγαλύτερης κλίμακας[13].

<u>Αποδεκτές μορφές ακατέργαστων δεδομένων από MetaboAnalyst 5.0:</u>

Για την επεξεργασία των ακατέργαστων δεδομένων μεταβολομικής που προέρχονται από αναλύσεις με την μέθοδο υγρής χρωματογραφίας υπερυψηλής πίεσης συζευγμένη με φασματομετρία μάζας υψηλής διακριτικής ικανότητας και ακρίβειας (UPLC-HRMS),

το MetaboAnalyst δέχεται μόνο μορφές ανοιχτού κώδικα (.mzML/.mzXML/.mzData) για κεντροειδή δεδομένα, και πέρα από την ηλεκτρονική φιλική προς τον χρήστη πλατφόρμα, δίνει τη δυνατότητα εγκατάστασης του πακέτου MetaboAnalystR, που βασίζεται στη γλώσσα προγραμματισμού R. Στην περίπτωση αυτή ο χρήστης δεν χρειάζεται να συμπιέσει τα ακατέργαστα δεδομένα πριν την ανάλυση, ενώ δεν υπάρχει και ο περιορισμός του μέγιστου αριθμού δειγμάτων, εφόσον το επιτρέπει η μνήμη RAM του υπολογιστή που χρησιμοποιείται. Στη διαδικτυακή πλατφόρμα υπάρχει η δυνατότητα επεξεργασίας μέχρι 200 συμπιεσμένων αρχείων σε κάθε ανάλυση που κατατίθεται[14].

Profinder:

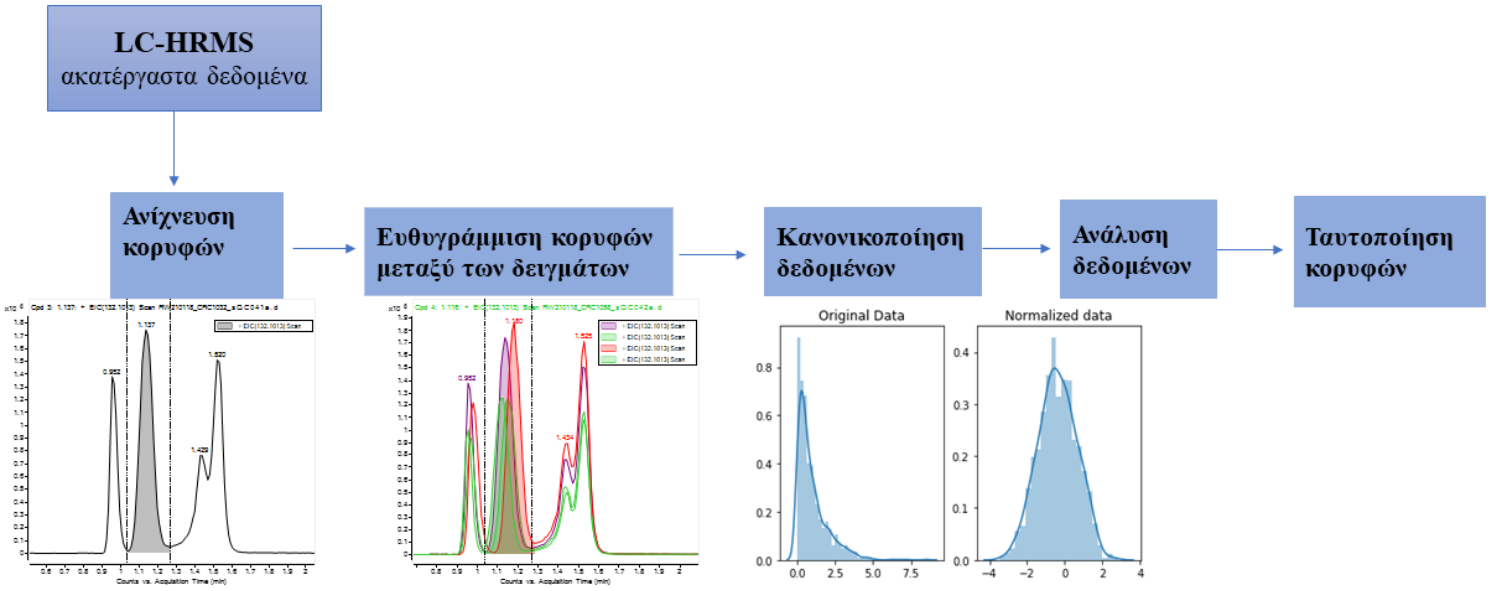
Το λογισμικό της εταιρίας AGILENT, MassHunter Profinder είναι ένα αυτόνομο πρόγραμμα, βελτιστοποιημένο για την ανίχνευση feature και την προεπεξεργασία δεδομένων που προέρχονται από αναλύσεις που έγιναν αποκλειστικά με κάποιο όργανο της ίδιας εταιρίας, GC/MSD, GC/Q-TOF, LC/TOF και LC/Q-TOF. Το συγκεκριμένο λογισμικό είναι εύκολο στη χρήση του και έχει χρησιμοποιηθεί κατά κόρον σε μεγάλες ερευνητικές μελέτες. Εμπεριέχει αλγόριθμους που επιτρέπουν την ανίχνευση κορυφών, την ομαδοποίηση και ευθυγράμμισή τους σε σχέση με τον χρόνο κατακράτησης στην χρωματογραφική στήλη, ανάμεσα σε δεδομένα πολλών αρχείων. Με την "ομαδοποίηση" των feature στο πεδίο του χρωματογραφικού χρόνου, το λογισμικό ελαχιστοποιεί την εμφάνιση τόσο ψευδών όσο και λανθασμένων θετικών και ψευδώς αρνητικών feature[15].

Προκατεργασία δεδομένων μη- στοχευμένης μεταβολομικής ανάλυσης:

Παρά τη σημαντική πρόοδο που έχει σημειωθεί τα τελευταία χρόνια στο κομμάτι της βιοπληροφορικής, παραμένουν κρίσιμα ζητήματα όσον αφορά διάφορα βασικά βήματα που εμπλέκονται στην τρέχουσα ροή εργασίας της μεταβολομικής. Η υπολογιστική ροή εργασίας για την μεταβολομική με βάση το LC-HRMS μπορεί να συνοψιστεί σε τρία γενικά βήματα: (1) επεξεργασία φασμάτων για τη μετατροπή των ακατέργαστων φασμάτων σε πίνακα έντασης κορυφών, (2) στατιστική ανάλυση δεδομένων για τον εντοπισμό σημαντικών χρωματογραφικών κορυφών και (3) σχολιασμός κορυφών για τον χαρακτηρισμό των σχέσεων κορυφών και την απόδοση ταυτοτήτων (identification)-. Η πρώτη πρόκληση αφορά την ανίχνευση των κορυφών κατά την επεξεργασία των ακατέργαστων φασμάτων. Η βελτίωση της ικανότητας επιλογής των κορυφών που

αντιστοιχούν σε μεταβολίτες έναντι του θορύβου είναι ζωτικής σημασίας για την βελτίωση των αποτελεσμάτων της μεταβολομικής. Οι προεπιλεγμένες παράμετροι που παρέχονται από τα κοινά εργαλεία επεξεργασίας φασμάτων δεν είναι βελτιστοποιημένα για την εκάστοτε οργανολογία/χρωματογραφία/συνθήκες λειτουργίας του φασματομέτρο μάζας που επιλέχθηκαν σε ένα συγκεκριμένο πείραμα και η λανθασμένη χρήση των παραμέτρων από τους χρήστες μπορεί να οδηγήσει σε σημαντικά ζητήματα στην ποιότητα των δεδομένων. Το δεύτερο καίριο ζήτημα είναι η επίδραση της παρτίδας (batch effect), το οποίο συνδέεται συνήθως με κλινικές ή πληθυσμιακές μελέτες μεγάλης κλίμακας, όταν τα δείγματα χρειάζεται να αναλυθούν σε διαφορετικές παρτίδες ή εκτείνονται σε μεγάλη χρονική περίοδο. Κατά τη διάρκεια της συλλογής δεδομένων από τα διάφορα δείγματα, οι χρωματογραφικές συνθήκες μπορεί να υποστούν μικρές μεταβολές που μπορεί να παρασύρουν τις γραμμές βάσης, δημιουργώντας διαφοροποιήσεις στις εντάσεις ή/και στους χρόνους συγκράτησης. Για την αντιμετώπιση αυτού του ζητήματος, έχουν αναπτυχθεί διάφοροι τύποι μεθόδων διόρθωσης του συγκεκριμένου φαινομένου που βασίζονται σε δείγματα ποιοτικού ελέγχου (quality control, QC), εσωτερικά πρότυπα ή κανονικοποίηση σε κλίμακα θέσης. Αυτές οι μέθοδοι βασίζονται σε διαφορετικές παραδοχές με τα δικά τους πλεονεκτήματα και περιορισμούς. Η επιλογή μιας κατάλληλης μεθόδου διόρθωσης παρτίδας είναι κρίσιμη, καθώς έχει σημαντικό αντίκτυπο στη μετέπειτα στατιστική και λειτουργική ανάλυση. Τέλος, η βιολογική ερμηνεία των δεδομένων μεταβολομικής απαιτεί συνήθως να έχει προηγηθεί η ταυτοποίηση των μεταβολιτών. Αυτή η διαδικασία είναι εξαιρετικά χρονοβόρα και παραμένει τροχοπέδη στις μελέτες μεταβολομικής.

Η προεπεξεργασία των δεδομένων μη- στοχευμένης μεταβολομικής είναι το πρώτο βήμα στην ανάλυση τέτοιων πειραμάτων. Το συγκεκριμένο βήμα αποσκοπεί αρχικά στην ποσοτικοποίηση όλων των σημάτων που λαμβάνονται από τα ιόντα των ενώσεων που υπάρχουν σε κάθε δείγμα και έπειτα στην αντιστοίχιση των συγκεκριμένων ιόντων ανάμεσα σε όλα τα δείγματα που συμμετέχουν σε κάθε μεταβολομική μελέτη. Από την παραπάνω διαδικασία προκύπτει ένας δισδιάστατος πίνακας, ο οποίος περιλαμβάνει πληροφορίες για τις χρωματογραφικές κορυφές που έχουν ομαδοποιηθεί μεταξύ των δεδομένων και χαρακτηρίζονται από τη μάζα που ανιχνεύθηκε σε έναν δεδομάνο χρωματογραφικό χρόνο (m/z@RT, feature) και τις εντάσεις των κορυφών αυτών. Τα δεδομένα του πίνακα, στη συνέχεια μπορούν να υποστούν περαιτέρω επεξεργασία όπως η κανονικοποίηση αυτών για την απομάκρυνση των διαφορών δειγματοληψίας, των επιδράσεων διαχωρισμού της παρτίδας ή των τυχόν μετατοπίσεων του σήματος που εξαρτώνται από τη σειρά έγχυσης.



Εικόνα 1.Διάγραμμα ροής στο οποίο απεικονίζονται τα βήματα επεξεργασίας των ακατέργαστων δεδομένων στο MetaboAnalyst[16].

Ένα από τα βασικότερα και πρωταρχικά βήματα της προεπεξεργασίας των δεδομένων είναι η ανίχνευση κορυφών (γνωστή και ως peak picking, Εικόνα 1). Υπάρχουν πολλοί αλγόριθμοι με βάση τους οποίους μπορεί να γίνει το συγκεκριμένο βήμα. Πιο συγκεκριμένα στην πλατφόρμα metaboanalyst παρέχονται οι αλγόριθμοι centwave, massifquant και matchedfilter. Ο αλγόριθμος Centwave που δανείζεται στην πλατφόρμα metaboanalyst από το XCMS, έχει αποδειχθεί ότι έχει καλή απόδοση στην επεξεργασία φασμάτων LC-HRMS, εξάλλου συστήνεται η χρήση του συγκεκριμένου αλγορίθμου σε κεντροειδή δεδομένα υψηλής ανάλυσης. Ωστόσο, μια βασική δυσκολία είναι το να επιλεχθούν οι κατάλληλες τιμές για διάφορες βασικές παραμέτρους, γεγονός που απαιτεί βαθιά κατανόηση τόσο των οργάνων MS και των δεδομένων που παρέχουν, όσο και του αλγορίθμου ανίχνευσης κορυφών. Για την αντιμετώπιση αυτής της πρόκλησης, έχει αναπτυχθεί και ενσωματωθεί τον Απρίλιο του 2023 στο Metaboanalyst το OptiLCMS, το οποίο επιτρέπει τον αυτοματοποιημένο καθορισμό παραμέτρων μέσω των δεδομένων που πρόκειται να προ-επεξεργαστούν.. Μετά την ανίχνευση κορυφών, πραγματοποιείται ευθυγράμμιση κορυφών για την αντιμετώπιση των διακυμάνσεων του χρόνου κατακράτησης σε όλα τα φάσματα. Αυτές οι ευθυγραμμισμένες κορυφές σχηματίζουν έναν πίνακα έντασης κορυφής με ποικίλα ποσοστά ελλειπόντων δεδομένων. Αυτές οι ελλειπούσες τιμές υποδηλώνουν ότι είτε η ανίχνευση κορυφής απέτυχε λόγω

περιορισμένων ποσοτήτων της συγκεκριμένης μάζας σε ένα δεδομένο δείγμα που δεν ξεπερνούν το όριο ανίχνευσης της μεθόδου, είτε το αντίστοιχο χαρακτηριστικό απουσιάζει από το συγκεκριμένο δείγμα[13].

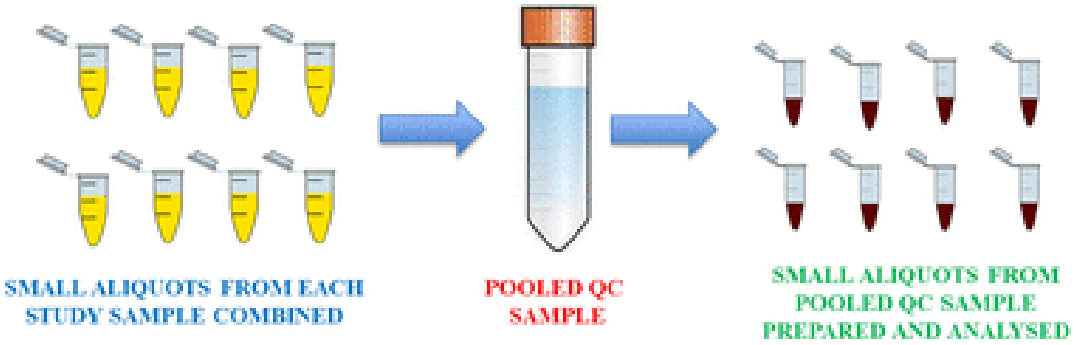
Επόμενο βήμα στην προκατεργασία των δεδομένων είναι η ταυτοποίηση των κορυφών (peak annotation). Ένα τυπικό φάσμα LC-HRMS κοινών βιολογικών υγρών (π.χ. δείγματα αίματος) μπορεί συχνά να περιλαμβάνει >10.000 κορυφές. Ωστόσο, ο αριθμός αυτός δεν ισοδυναμεί με τον αριθμό των μεταβολιτών που πραγματικά ανιχνεύονται. Η αντιστοιχία μεταξύ του αριθμού των κορυφών και του αριθμού των πραγματικών μεταβολιτών είναι δύσκολο να προβλεφθεί. Πολλαπλές κορυφές μπορεί να προέρχονται από τις ίδιες ενώσεις, λόγω του τρόπου λειτουργίας του ηλεκτροψεκασμού. Πρόκειται για πραγματικές, βιολογικά-σχετιζόμενες κορυφές και μπορεί να προκύψουν από το σχηματισμό προσθετικών προϊόντων (adducts), την ενσωμάτωση ισοτόπων ή τον κατακερματισμό κατά την προετοιμασία ή την ανάλυση LC-MS του δείγματος. Οι υπόλοιπες κορυφές θεωρούνται τώρα ότι προέρχονται σε μεγάλο βαθμό από θόρυβο υποβάθρου. Ως εκ τούτου, το πρώτο βήμα αποσκοπεί στον εντοπισμό των πραγματικών κορυφών και στην αποσαφήνιση των μεταξύ τους σχέσεων. Έχουν αναπτυχθεί πολλοί εμπειρικοί και στατιστικοί κανόνες για την αντιμετώπιση αυτού του προβλήματος, συμπεριλαμβανομένων των CAMERA και Clique τα οποία είναι δύο δημοφιλή πακέτα R. Η ταυτοποίηση κορυφών στο MetaboAnalyst 5.0 βασίζεται επί του παρόντος στο CAMERA, το οποίο διαθέτει αλγόριθμους που επιτρέπουν την αναγνώριση ισοτόπων και χημικών προσθέτων (adducts) και προτείνει ταυτοποιήσεις βάσει της ακριβούς μάζας[17]. Το επόμενο βήμα είναι η απόδοση 'ταυτοτήτων' σε αυτές τις κορυφές. Αυτό είναι ένα δύσκολο έργο, ακόμη και με όργανα υψηλής ανάλυσης, καθώς μία κορυφή μπορεί ενδεχομένως να αντιστοιχεί σε πολλές ενώσεις. Είναι σημαντικό να έχουμε κατά νου ότι, σε αντίθεση με τη στοχευμένη μεταβολομική, ο πρωταρχικός στόχος της ολιστικής μεταβολομικής ανάλυσης είναι η κατανόηση των μεταβολικών μοτίβων και ο εντοπισμός υποσχόμενων μεταβολικών μονοπατιών για τον σχεδιασμό επακόλουθων μελετών για πιο στοχευμένη ανάλυση.

Ακολουθεί η επιπλέον ανάλυση των αρχικών δεδομένων, με τη χρήση μεταδεδομένων. Τα μεταδεδομένα περιγράφουν τα δεδομένα και περιέχουν λεπτομέρειες σχετικά με τις πειραματικές συνθήκες, τις πηγές των δειγμάτων (π.χ. είδη, ιστοί), τη συλλογή των δειγμάτων (π.χ. τοποθεσία, χρόνος) και άλλους παράγοντες. Τέτοια μεταδεδομένα είναι κρίσιμα για την ερμηνεία των δεδομένων, διότι επιτρέπουν στους αναλυτές να λαμβάνουν υπόψη το βιολογικό και περιβαλλοντικό πλαίσιο καθώς και τις επιδράσεις των πιθανών τεχνικών σφαλμάτων όταν αναλύουν τα δεδομένα. Με τον τρόπο αυτόν

γίνεται επιτρεπτό σε όλους τους ερευνητές να μπορούν να έχουν πρόσβαση σε όλες τις πληροφορίες που χρειάζονται προκειμένου να ερμηνεύσουν σωστά τα δεδομένα μίας άλλης μελέτης και πιθανόν να τα συγκρίνουν με τα δικά τους, αν αυτό κρίνεται αναγκαίο. Οι στατιστικές αναλύσεις που λαμβάνουν υπόψη τις επιπλέον πληροφορίες που παρέχουν τα μεταδεδομένα μπορούν να οδηγήσουν στην δημιουργία πιο έγκυρων συμπερασμάτων σχετικά με τις σχέσεις μεταξύ της πρωτογενούς μεταβλητής και των omics δεδομένων[18].

Δείγματα ελέγχου ποιότητας (Quality control samples (QC)):

Τα δείγματα ελέγχου ποιότητας (quality control, QC) είναι ειδικά δείγματα που εισάγονται στη διαδικασία δοκιμής και αντιμετωπίζονται ακριβώς κατά τον ίδιο τρόπο με τα δείγματα ασθενών της μελέτης, καθώς εκτίθενται στις ίδιες συνθήκες προκατεργασίας και ανάλυσης. Η ανάλυση δειγμάτων QC πραγματοποιείται σε τακτά χρονικά διαστήματα για την παρακολούθηση της επαναληψιμότητας της ανάλυσης. Επιπλέον, επειδή τα δείγματα QC ακολουθούν όλα τα βήματα της προκατεργασίας, μαζί με τα πραγματικά δείγματα, καταδεικνύον προβλήματα όχι μόνο κατά την αναλυτική διαδικασία, αλλά και κατά το στάδιο της προκατεργασίας. Για την κατασκευή των QC δειγμάτων λαμβάνεται μικρός όγκος από κάθε προς ανάλυση δείγμα και συγκεντρώνεται σε ένα κοινό δοχείο (pooled QC). Στη συνέχεια, το ενιαίο αυτό δείγμα χωρίζεται σε επιμέρους τμήματα, ίσα με τον αριθμό των δειγμάτων ελέγχου ποιότητας που πρόκειται να συμπεριληφθούν στην ανάλυση, όπως φαίνεται στην Εικόνα 2.



Εικόνα 2: Οπτικοποίηση του τρόπου δημιουργίας δειγμάτων QC[19].

Η μη στοχευμένη μεταβολομική, όπως αναφέρθηκε παραπάνω, είναι μια αναλυτική προσέγγιση με πολυάριθμες εφαρμογές που στοχεύει στον χαρακτηρισμό του συνόλο των μεταβολιτών ενός βιολογικού συστήματος. Λόγω της ταυτόχρονης

παρακολούθησης χιλιάδων σημάτων σε συνήθως μεγάλο αριθμό δειγμάτων, η ποιότητα των δεδομένων μπορεί να είναι δύσκολο να αξιολογηθεί και να αποδειχθεί στα πειράματα μεταβολομικής. Η πιο ευρέως χρησιμοποιούμενη μέθοδος παρακολούθησης της ποιότητας αυτών των αναλύσεων είναι μέσω της χρήσης συγκεντρωτικών δειγμάτων ελέγχου ποιότητας (QC). Μέσω των δειγμάτων QC επιτυγχάνεται όχι μόνο η παρακολούθηση αλλά είναι δυνατή και η διόρθωση της αναλυτικής απόκλισης που εισάγεται κατά τα στάδια της προετοιμασίας των δειγμάτων και της ανάλυσης. Οι διαδικασίες ποιοτικού ελέγχου αποτελούν ιδιαίτερη πρόκληση για μη στοχευμένες αναλυτικές προσεγγίσεις. Τα τελευταία χρόνια έχουν εμφανιστεί συντονισμένες προσπάθειες για την έμφαση στα μέτρα διαχείρισης της ποιότητας. Ένα συγκεντρωτικό δείγμα QC, το οποίο παράγεται με την ανάμειξη μικρών όγκων από όλα τα δείγματα της μελέτης, μπορεί να θεωρηθεί ο "μέσος όρος" όλων των δειγμάτων. Τα δείγματα QC αναλύονται στη συνέχεια μαζί με τα δείγματα μελέτης περιοδικά καθ' όλη τη διάρκεια της σειράς εγχύσεων[20,21].

Υπάρχουν διάφορες πιθανές χρήσεις ενός συγκεντρωτικού δείγματος QC (pooled QC). :

- Ένας από τους βασικότερους ρόλους των pooled QC είναι η αξιολόγηση της μεταβλητότητας στην προετοιμασία του δείγματος και στην απόδοση του οργάνου. Τα συγκεντρωτικά δείγματα QC παρέχουν μια εκτίμηση της αναλυτικής επαναληψιμότητας των μετρήσεων.
- Μπορούν επίσης να χρησιμοποιηθούν ως αρχική αξιολόγηση της καταλληλόλητας του συστήματος πριν από μια μελέτη ή για την προετοιμασία του αναλυτικού συστήματος.
- Τέλος, είναι απαραίτητα για την διόρθωση των τεχνικών παραλλαγών εντός και μεταξύ των παρτίδων (batch effect) και την παρακολούθηση της σταθερότητας του αναλυτικού συστήματος. Έτσι, γίνεται εφικτή η ενσωμάτωση των αποτελεσμάτων από πολλαπλές αναλυτικές παρτίδες, γεγονός που επέτρεψε μελέτες μεγάλης κλίμακας (n>1000) μέσω της διόρθωσης των αναπόφευκτων τεχνικών παραλλαγών που αναπόφευκτα διαδραματίζονται σε αναλύσεις που εκτείνονται σε διάστημα αρκετών εβδομάδων έως και μηνών[22,23].

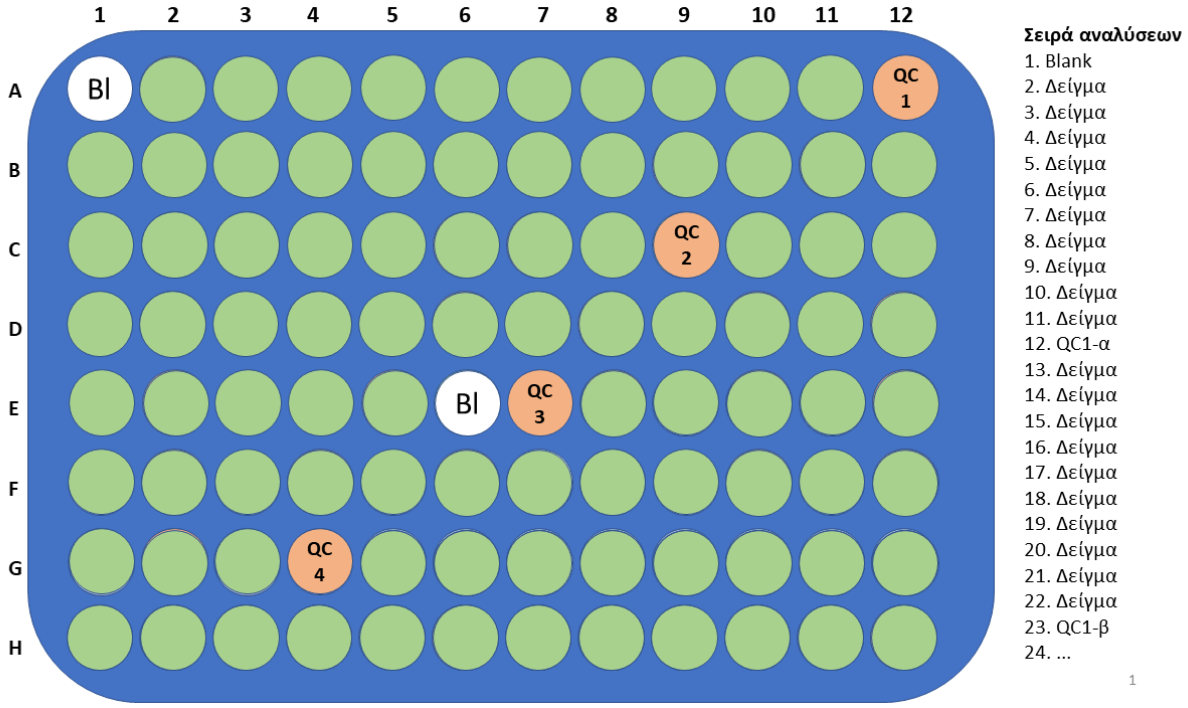
Λευκά δείγματα (Blank samples):

Λευκό ονομάζεται ένα δείγμα το οποίο περιέχει όλα τα αντιδραστήρια που χρησιμοποιήθηκαν, χωρίς όμως να περιέχει κανέναν από τους αναλύτες. Τέτοια δείγματα έχουν σημαντικό ρόλο στην αναλυτική διαδικασία, αφού μας βοηθούν να εντοπίσουμε κάποιο σφάλμα, για παράδειγμα πιθανή επιμόλυνση των δειγμάτων ή κάποια παρεμβολή στο σύστημα[24].

Πειραματική Διαδικασία:

Πειραματικός σχεδιασμός και προκατεργασία δειγμάτων

Στόχος της συγκεκριμένης εργασίας αποτέλεσε η ανάλυση 200 QC και 50 Blank δειγμάτων τα οποία προέρχονται από μία μελέτη μεγάλης κλίμακας, η οποία διήρκησε 8 εβδομάδες. Κατά τη μελέτη αυτή, πλάσμα αίματος από 2242 δείγματα αναλύθηκαν μετά την καταβύθιση των πρωτεϊνών με ανάμιξη 30 μL πλάσματος με περίσσεια ακετονιτριλίου και το φιλτράρισμα τους. Τα δείγματα QC προέκυψαν από την δημιουργία κοινού δείγματος με την ανάμιξη 10 μL από 400 τυχαία επιλεγμένα δείγματα της μελέτης. Τα λευκά δείγματα (blank) παρασκευάστηκαν κατά όμοιο τρόπο με τα υπόλοιπα δείγματα (πραγματικά και δείγματα ελέγχου), αντικαθιστώντας το πλάσμα αίματος με υπερκάθαρο νερό. Όλα τα δείγματα ακολούθησαν την ίδια προκατεργασία και συνθήκες ανάλυσης με τα πραγματικά δείγματα πλάσματος της μελέτης. Σε κάθε μικροπλάκα 96 κυψελών περιλήφθηκαν τέσσερα δείγματα QC και δύο λευκά δείγματα. Κάθε δείγμα QC αναλύθηκε δύο φορές, με διαμεσολάβηση 10-12 πραγματικών δειγμάτων, όπως φαίνεται στην Εικόνα 3. Το σύνολο των δειγμάτων αναλύθηκε σε 5 παρτίδες, κάθε μία αποτελούμενη από 5 μικροπλάκες 96 κυψελών.



Εικόνα 3: Σχηματική απεικόνιση του τρόπου τοποθέτησης των δειγμάτων στις μικροπλάκες, καθώς και της σειράς ανάλυσής τους.

Συνθήκες ανάλυσης UHPLC-QTOF-MS

Οι αναλύσεις πραγματοποιήθηκαν σε ένα UHPLC-QTOF-MS σύστημα, αποτελούμενο από έναν 1290 Binary υγρό χρωματογράφο και ένα 6550 QTOF αναλυτή μαζών με Jet Stream Electrospray Ionization source της εταιρίας Agilent Technologies στο εργαστήριο Ογκο-μεταβολομικής του Διεθνούς Οργανισμού Ερευνών για τον Καρκίνο (International Agency for Research on Cancer (IARC)). Η στήλη διαχωρισμού που χρησιμοποιήθηκε ήταν η Acquity UHPLC HSS T3 (2.1 x 100 mm, 1.8 μm) και η κινητή φάση αποτελούταν από βαθμωτή έκλουση υπερκάθαρου νερού και μεθανόλης περιεκτικότητας 0.05% (v/v) σε φορμικό οξύ.

Προκατεργασία αποτελεσμάτων

Η προκατεργασία των δειγμάτων ελέγχου και των λευκών δειγμάτων, για τους σκοπούς της παρούσας εργασίας, πραγματοποιήθηκε με τα λογισμικά Profinder 10.0.2.162 και Mass Profiler Professional B.14.9.1 της Agilent Technologies. Επιπλέον, προκειμένου να συγκρίνουμε τα αποτελέσματα που προέρχονται από τα λογισμικά της Agilent με αυτά ενός λογισμικού ανοιχτού κώδικα, τα ακατέργαστα δεδομένα που προέρχονται από τις συγκεκριμένες αναλύσεις και αφορούν αυτά τα δείγματα επεξεργάστηκαν περαιτέρω με την χρήση ενός κατάλληλου διαδικτυακού εργαλείου, του MetaboAnalyst. Με δεδομένο ότι η χωρητικότητα του συγκεκριμένου εργαλείου περιορίζεται στα 200 δείγματα, τα QC αναλύθηκαν ξεχωριστά από τα blank δείγματα, ακολουθώντας όμως τα ίδια βήματα και ταυτόσημες παραμέτρους.

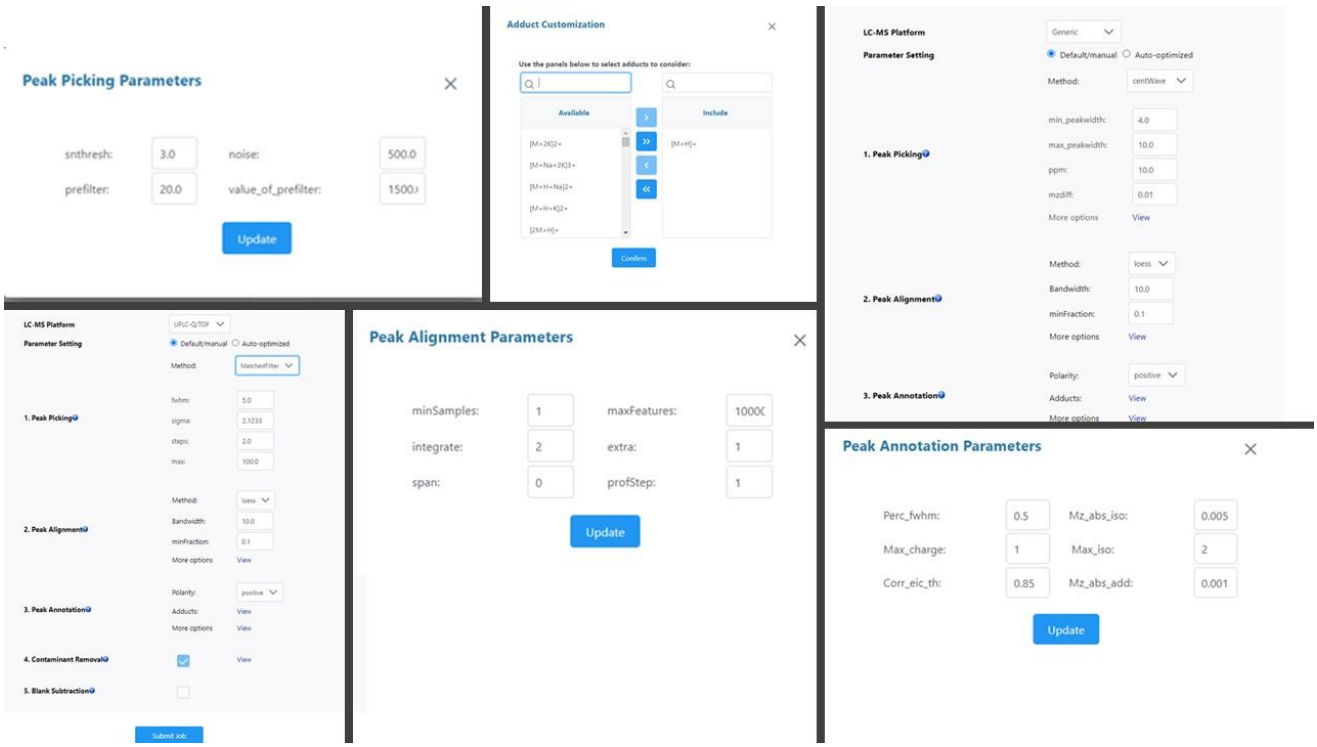
Αρχικά, πραγματοποιήθηκε η μετατροπή των δεδομένων σε κατάλληλη μορφή προκειμένου να μπορούν να γίνουν αποδεκτά από το MetaboAnalyst. Μία από τις πλέον συνηθισμένες μορφές δεδομένων για χρήση με εργαλεία ανοιχτού κώδικα είναι η‘.mzXML’ και η μετατροπή των δεδομένων ‘.d’ στη μορφή αυτή έγινε με τη χρήση του online εργαλείου proteowizard.

Ακολούθησε η έναρξη των αναλύσεων με την βοήθεια των αλγορίθμων του MetaboAnalyst. Η πρώτη ανάλυση έγινε με την επιλογή των auto- optimized παραμέτρων του προγράμματος. Με την επιλογή αυτήν, το λογισμικό επιλέγει τις καταλληλότερες παραμέτρους με τις οποίες θα γίνει η προκατεργασία των δεδομένων βάσει των δεδομένων με τα οποία τροφοδοτούμαι το λογισμικό. Στην περίπτωση αυτή ο χρήστης δε χρειάζεται να πάρει αποφάσεις που αργότερα πιθανώς επηρεάσουν την ποιότητα των εξαγόμενων αποτελεσμάτων, αλλά δεν έχει και έλεγχο στις επιλογές του συστήματος. Επιπλέον, πραγματοποιήθηκε η προκατεργασία και με χρήση άλλων αλγορίθμων για τους οποίους έγινε προσπάθεια επιλογής των καταλληλότερων παραμέτρων έτσι ώστε να ταιριάζουν στα συγκεκριμένα δείγματα. Οι παράμετροι που επιλέχθηκαν ρυθμίστηκαν όσο το δυνατόν πλησιέστερες στις αναλύσεις που έγιναν και

με τους τρείς διαθέσιμους αλγορίθμους του λογισμικού (centWave, matchedfilter και Massifquant). Στην βιβλιογραφία αναφέρεται ο αλγόριθμος centWave ως ο καταλληλότερος για την προκατεργασία δεδομένων που προέρχονται από ανάλυση LC-HRMS. Έγιναν δοκιμές όμως και με τους τρείς αλγορίθμους με τις κοντινότερες δυνατές παραμέτρους, προκειμένου να είναι δυνατή η σύγκριση μεταξύ τους, αλλά και τελικά η σύγκριση τους με τα αποτελέσματα που λήφθηκαν από τα λογισμικά που προτείνονται από την Agilent Technologies για το σκοπό αυτό.

Στην κατηγορία επιλογής κορυφών (Peak Picking) η πρώτη επιλογή που έγινε αφορούσε την επιλογή του αλγορίθμου, ανάμεσα στους τρείς αλγορίθμους που προσφέρει το συγκεκριμένο λογισμικό (CentWave, Massifquant, Matchedfilter). Πραγματοποιήθηκαν αναλύσεις και με τους τρείς αλγορίθμους προκειμένου να αποσαφηνιστεί η επίδραση του καθενός από αυτούς στα αναλυτικά αποτελέσματα. Έπειτα, για να προσδιοριστούν τα min και max_peakwidth παρατηρήθηκαν τυχαία χρωματογραφήματα από τα δείγματα ελέγχου, εκεί μελετήθηκαν κορυφές και προσδιορίστηκαν οι συγκεκριμένες τιμές. Οι παράμετροι ppm και mzdiff υποδηλώνουν την μέγιστη διαφορά μάζας που μπορεί να έχουν δύο κορυφές σε διαφορετικά χρωματογραφήματα για να θεωρηθεί πως προέρχονται από τον ίδιο μεταβολίτη. Για την ευθυγράμμιση των κορυφών επιλέχθηκε η μέθοδος Loess, η οποία ευθυγραμμίζει τις κορυφές οι οποίες εμφανίζουν μικρή απόκλιση στον χρόνο κατακράτησης ανάμεσα στις αναλύσεις των διαφόρων δειγμάτων. Επιπλέον, στην επιλογή του bandwidth, που είναι η μέγιστη διαφορά σε χρόνο κατακράτησης που μπορεί να εμφανίζουν μεταξύ τους feature τα οποία πρόκειται να 'ευθυγραμμιστούν', ορίστηκε η τιμή 10 sec[25,26].

Από κάθε αλγόριθμο και λογισμικό προέκυψε ένας πίνακας excel, στον οποίο περιέχονται πληροφορίες οι οποίες αφορούν τον λόγο μάζα/φορτίο, τον χρόνο κατακράτησης της συγκεκριμένης ουσίας στην στήλη αλλά και την ένταση του κάθε ανιχνευόμενου μεταβολίτη σε όλα τα προς ανάλυση δείγματα.



Εικόνα 4: Οι ακριβείς παράμετροι με τις οποίες ρυθμίστηκε το MetaboAnalyst για τις συγκεκριμένες αναλύσεις.

Στατιστική ανάλυση

Έγινε λήψη των συγκεκριμένων αναλυτικών αποτελεσμάτων από τα δύο λογισμικά (Profinder και MetaboAnalyst) και ακολούθησε στατιστική επεξεργασία. Για τον σκοπό αυτό, υπολογίστηκε ο συντελεστής μεταβλητότητας επί τοις εκατό (coefficient of variation (CV%)) για κάθε feature, ο οποίος είναι ένα μέσο αξιολόγησης της σταθερότητας του συγκεκριμένου feature σε όλα τα δείγματα ελέγχου.

$$CV\% = \left(\frac{Standard\ deviation\ (a{:}\,b)}{Average\ (a{:}\,b)}\right) * 100$$

Αρχικά υπολογίστηκε το συνολικό CV% μεταξύ όλων των προς ανάλυση δειγμάτων και στη συνέχει ξεχωριστά για κάθε παρτίδα (batch), επιπλέον στα δείγματα QC υπολογίστηκε και το συνολικό CV για όλα τα a και τα b δείγματα (CVa και CVb αντίστοιχα). Επιπλέον, υπολογίστηκε για κάθε feature ο αριθμός των δειγμάτων στα οποία ανιχνεύθηκε κάθε feature, αλλά και ο αριθμός των feature κάθε ανάλυσης που βρίσκονται στο 100% των δειγμάτων. Η τελευταία παράμετρος που χρησιμοποιήθηκε για την αξιολόγηση της αξιοπιστίας των αναλυτικών αλγορίθμων του MetaboAnalyst ήταν ο αριθμός των feature στα οποία το ολικό CV% έχει τιμή μικρότερη από 20%. Ακολούθησε η δημιουργία συνολικών πινάκων αλλά και η διαγραμματική απεικόνιση

των συγκεκριμένων στατιστικών υπολογισμών, με σκοπό την καλύτερη κατανόηση των αποτελεσμάτων αλλά και την ευκολότερη σύγκριση των αλγορίθμων μεταξύ τους.

Προσπάθεια αφαίρεσης θορύβου:

Το επόμενο βήμα αποσκοπούσε στην αφαίρεση των feature τα οποία πιθανόν να αντιστοιχούν σε θόρυβο, που δημιουργήθηκε είτε μέσω του αναλυτικού οργάνου είτε από τους διαλύτες που χρησιμοποιήθηκαν στην προκατεργασία των δειγμάτων (background exclusion). Στο συγκεκριμένο στάδιο κεντρικό ρόλο έπαιξαν τα feature που ανιχνεύθηκαν στα δείγματα blank. Έγινε προσπάθεια να βρεθούν τα feature που εμφανίζονται σε παρόμοιους χρόνους και m/z στα QC+ blank για να αφαιρεθούν από την διαδικασία της ανάλυσης, αφού σίγουρα δεν προέρχονται από κάποιον αναλύτη. Για τον σκοπό αυτόν έγινε χρήση ενός προγράμματος που κάνει match τα features ανάμεσα στις Peak lists για QC+ BLANK samples. Τα όρια αντιστοίχισης που τέθηκαν αρχικά στο λογισμικό (για το match feature) ήταν ελαστικά. Στη συνέχεια εξετάστηκαν οι συσχετίσεις περαιτέρω και αφαιρέθηκαν τα κοινά feature τα οποία είχαν εντοπιστεί αρχικά και είχαν διαφορά χρόνου 0,5<dt<-0.5. Τα κοινά feature ανάμεσα σε QC+ Blank, από κάθε τρόπο ανάλυσης απομονώθηκαν. Υπολογίστηκε ο μέσος όρος της έντασης κάθε feature σε όλα τα δείγματα. Έπειτα προσδιορίστηκε το fold change κάθε feature σε QC+ Blank δείγματα (Με τον ίδιο τρόπο που υπολογίζεται στο λογισμικό AGILENT , για να είναι συγκρίσιμα τα αποτελέσματα δηλ:

$$Fold\ Change = \frac{\text{Average intensity QC}}{\text{Average intensity Blank}}$$

Όσα κοινά feature είχαν fold change>5 δεν αφαιρέθηκαν από την τελική επεξεργασία, ενώ όσα feature ανιχνεύθηκαν και στα QC και στα BLANK και έχουν fold change<=5 θεωρήθηκαν θόρυβος και αφαιρέθηκαν. Επομένως, τα περαιτέρω βήματα διαδραματίζονται μόνο στα feature τα οποία ανιχνεύονται αποκλειστικά στα δείγματα QC αλλά και στα κοινά feature με fold change>5 .

Διόρθωση του φαινομένου παρτίδας:

Το τελευταίο βήμα που αφορά την επεξεργασία των αποτελεσμάτων του MetaboAnalyst αποσκοπεί στην βελτίωση του φαινομένου της παρτίδας (batch effect correction). Πραγματοποιήθηκε με την αντίστοιχη επιλογή που προσφέρει το διαδικτυακό εργαλείο. Για τον σκοπό αυτό δημιουργήθηκε ένα αρχείο στο excel τύπου ".csv" το οποίο περιείχε τα δεδομένα τοποθετημένα με κατάλληλο τρόπο, ακολουθώντας τις οδηγίες τις οποίες παρέχει το λογισμικό μέσα από τα παραδείγματά του. Στο batch

correction αρχικά επιλέχθηκε για όλους τους τρόπους ανάλυσης η επιλογή η οποία τοποθετεί στη θέση των missing values το 1/5 της μικρότερης τιμής. Στη συνέχεια για την ανάλυση που έγινε με τις Auto-optimized παραμέτρους έγινε batch correction και με τις επιλογές K-nearest neighbours και Probabilistic PCA για να γίνει σύγκριση των αποτελεσμάτων και να εξάγουμε συμπεράσματα σχετικά με την επίδραση της συγκεκριμένης παραμέτρου στα αποτελέσματα του batch correction. Στα δεδομένα από τους υπόλοιπους αλγορίθμους έγινε batch correction μόνο με την λειτουργία η οποία τοποθετεί στη θέση των missing values το 1/5 της μικρότερης τιμής. Τέλος, έγινε υπολογισμός του CV αλλά και του CV per batch στα δεδομένα που προέκυψαν μετά το batch correction.

Σύγκριση αποτελεσμάτων MetaboAnalyst με τα αποτελέσματα του λογισμικού της AGILENT:

Μετά το τέλος της επεξεργασίας όλων των δεδομένων που προέκυψαν από το δωρεάν διαδικτυακό εργαλείο MetaboAnalyst, τα ίδια δεδομένα αναλύθηκαν στο Profinder, λογισμικό της εταιρίας AGILENT. Το λογισμικό Agilent Profinder είναι ένα αυτόνομο πρόγραμμα, βελτιστοποιημένο για την ομαδική εξαγωγή feature, το οποίο χρησιμοποιείται κατά κόρον στον χώρο των μεταβολομικών αναλύσεων. Στη συγκεκριμένη ανάλυση οι παράμετροι με της οποίες ρυθμίστηκε το Profinder ήταν οι πλησιέστερες δυνατές σε σχέση με αυτές που χρησιμοποιήθηκαν στο MetaboAnalyst. Τα αποτελέσματα που προέκυψαν από τα δύο διαφορετικά λογισμικά συγκρίθηκαν μεταξύ τους αρχικά ως προς τον αριθμό των feature. Στη συνέχεια, έγινε προσπάθεια να συγκριθούν ποιοτικά τα feature μεταξύ τους ώστε να δούμε αν τα δύο λογισμικά ανιχνεύουν κορυφές σε πολύ κοντινά m/z και χρόνους κατακράτησης (RT).

Λογισμικά:

Παρακάτω θα αναφερθούν αναλυτικά όλα τα λογισμικά που χρησιμοποιήθηκαν για την εκπόνηση της συγκεκριμένης εργασίας:

- MetaboAnalyst 5.0
- MSConvert ProteoWizard 3.0.23260 64-Bit
- Profinder Agilent MassHunter Workstation Version 10.0.2
- RStudio Version 1.3.1093
- Statistics Kingdom. (2017). Multiple Linear Regression Calculator. (October 30, 2023) [web application]. https://www.statskingdom.com/410multi_linear_regression.html

- Venn Diagram Plotter Version 1.6.7458.27187

Αποτελέσματα:

Τα συνολικά αποτελέσματα από τις αναλύσεις των θετικά ιονισμένων QC και των θετικά ιονισμένων blank δειγμάτων μετά την στατιστική ανάλυση παρουσιάζονται συγκεντρωτικά στους παρακάτω πίνακες.

Πινακας 1 : Σύνοψη στατιστικών αποτελεσμάτων από τις αναλύσεις των θετικά ιονισμένων QC δειγμάτων με τους αλγορίθμους του MetaboAnalyst.

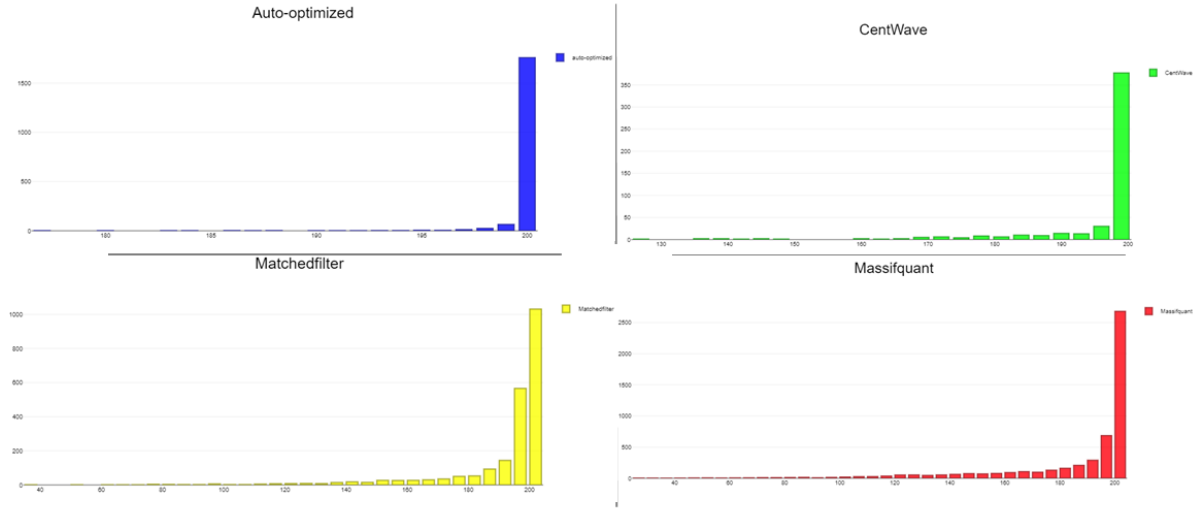
Algorithm	Number of reported features	Cv%	CV%<20	CV% per Batch <20%	Features detected in 100% of samples	Parameters
auto-optimized	1882	Min:6.18 Max:405.05	61	CV1:719 CV2:656 CV3:861 CV4:919 CV5:580	1758	Auto-optimized parameters from metaboAnalyst POSITIVE
centwave	496	Min:14.06 Max: 276.20	9	CV1:121 CV2:131 CV3: 124 CV4: 130 CV5: 91	336	Peakwidth(4,10), ppm(10) mzdiff(0.01) POSITIVE
centwave	496	Min: 14.06 Max: 276.20	9	CV1: 121 CV2: 131 CV3: 124 CV4: 130 CV5: 91	336	Peakwidth(4,10), ppm(10) mzdiff(0.005) POSITIVE
Massifquant	5128	Min:11.64 Max:1292.54	25	CV1: 601 CV2: 411 CV3: 492 CV4: 601 CV5: 534	2680	Peakwidth(4,10), ppm(10) mzdiff(0.01) POSITIVE
Matchedfilter	2192	Min: 4.256 Max: 586.37	66	CV1: 756 CV2: 875 CV3: 1035 CV4: 1019 CV5: 600	1031	Fwhm(5.0), sigma (2.1233) POSITIVE

Πινακας 2: Σύνοψη στατιστικών αποτελεσμάτων από τις αναλύσεις των θετικά ιονισμένων Blank δειγμάτων.

Algorithm	Number of reported features	Cv%	CV%<20	CV% per batch <20%	Features detected in 100% of samples	Parameters
auto-optimized	806	Min:13.196 Max:393.07	12	CV1: 321 CV2: 271 CV3: 314 CV4: 327 CV5: 176	774	Auto-optimized parameters from metaboAnalyst POSITIVE
centwave	456	Min:16.383 Max: 549.788	1	CV1: 110 CV2: 91 CV3: 132 CV4: 130 CV5: 92	327	Peakwidth(4, 10), ppm(10) mzdiff(0.01) POSITIVE
centwave	456	Min:16.383 Max: 549.788	1	CV1: 110 CV2: 91 CV3: 132 CV4: 130 CV5: 92	327	Peakwidth(4, 10), ppm(10) mzdiff(0.005) POSITIVE
Massifquant	4246	Min:18.106 Max:669.661	2	CV1: 663 CV2: 400 CV3: 462 CV4: 631 CV5: 743	2892	Peakwidth(4, 10), ppm(10) mzdiff(0.01) POSITIVE
Matchedfilter	2034	Min: 13.207 Max:662.121	11	CV1:498 CV2: 373 CV3: 526 CV4: 632 CV5: 351	769	Fwhm(5.0), sigma (2.1233) POSITIVE

Αρχικά, προσδιορίστηκε ο αριθμός των συνολικών feature που ανιχνεύονται σε κάθε ανάλυση. Επιπλέον υπολογίστηκε ο αριθμός των feature που ανιχνεύεται στο 100% των δειγμάτων. Τα προς ανάλυση δείγματα στην συγκεκριμένη εργασία είναι δείγματα ελέγχου ποιότητας και λευκά δείγματα, επομένως το αναμενόμενο θα ήταν το ποσοστό των δειγμάτων που ανιχνεύονται σε όλα τα δείγματα να ήταν αρκετά υψηλό, παραπλήσιο του 100%. Η συγκεκριμένη όμως υπόθεση δεν επαληθεύεται από τα δεδομένα των παραπάνω πινάκων για τις περισσότερες αναλύσεις. Μόνο τα αποτελέσματα από την ανάλυση με τις auto-optimized

παραμέτρους του MetaboAnalyst εμφανίζουν υψηλό ποσοστό των συνολικών feature σε όλα τα δείγματα και συγκεκριμένα 93.5% στα QC (Εικόνα 5) και 96% στα Blank.



Εικόνα 5: Διαγραμματική απεικόνιση με μορφή ιστογράμματος της συχνότητας εμφάνισης των feature προς το πλήθος των QC δειγμάτων για όλους τους αλγόριθμους του MetaboAnalyst.

Από τους παραπάνω συγκεντρωτικούς πίνακες μπορεί εύκολα να παρατηρηθεί πως τα δεδομένα που προέκυψαν από τις αναλύσεις με τον αλγόριθμο CentWave του MetaboAnalyst με mzdiff 0.01 και 0.005 μοιάζουν ίδια. Για να εξακριβωθεί αυτό, παρατηρήθηκαν ακριβώς τα feature που ανιχνεύθηκαν και στις δύο αναλύσεις και βρέθηκε πως είναι πανομοιότυπα. Πέρα από το ότι τα feature που ανιχνεύθηκαν στις δύο αναλύσεις εμφανίζονται σε ίδια m/z και RT, παρατηρήθηκε πως εμφανίζουν παραπλήσιες εντάσεις. Η διαδικασία αυτή έγινε για να προσδιοριστεί η επίδραση που έχει η επιλογή της συγκεκριμένης παραμέτρου και ελέγχθηκε στα αποτελέσματα του αλγορίθμου CentWave που αναγράφεται στη βιβλιογραφία ότι ταιριάζει καλύτερα στα δεδομένα της συγκεκριμένης εργασίας. Επομένως, η επιλογή της συγκεκριμένης παραμέτρου δεν φαίνεται κρίσιμη τουλάχιστον στις συγκεκριμένες συνθήκες λόγω της μεγάλης ακρίβειας στη μάζα του συγκεκριμένου Q-TOF που χρησιμοποιήθηκε.

Έπειτα, υπολογίστηκε το ολικό CV και στη συνέχεια για όλα τα QC δείγματα υπολογίστηκε συμπληρωματικά το συνολικό CV για όλα τα a δείγματα και στη συνέχεια για όλα τα b. Αυτό μας επιτρέπει να δούμε τη σταθερότητα και την επαναληψιμότητα της μεθόδου. Αρχικά γίνεται το injection και η a ανάλυση των δειγμάτων και έπειτα από περίπου 12 injection γίνεται ξανά injection και ανάλυση του ίδιου δείγματος για να προκύψει η b ανάλυση. Με τον συγκεκριμένο τρόπο μπορεί να ελεγχθεί τόσο η επαναληψιμότητα της μεθόδου, όσο και η σταθερότητα των δειγμάτων. Μετά την πρώτη ανάλυση ο δειγματολήπτης έχει δημιουργήσει εγκοπή στη μεμβράνη που προστατεύει τα δείγματα, οπότε πιθανόν να έχει αλλάξει λίγο η σύσταση του

μέχρι την b ανάλυση (εξαιτίας αλλοιώσεων, πτητικών διαλυτών κλπ) Παρόλα αυτά οι τιμές του CVa% και του CVb% αναμένονται παραπλήσιες για κάθε feature.

Πίνακας 3: Συγκεντρωτικός πίνακας των τιμών CV%, CVa% και CVb% για τις αναλύσεις με όλους τους αλγορίθμους.

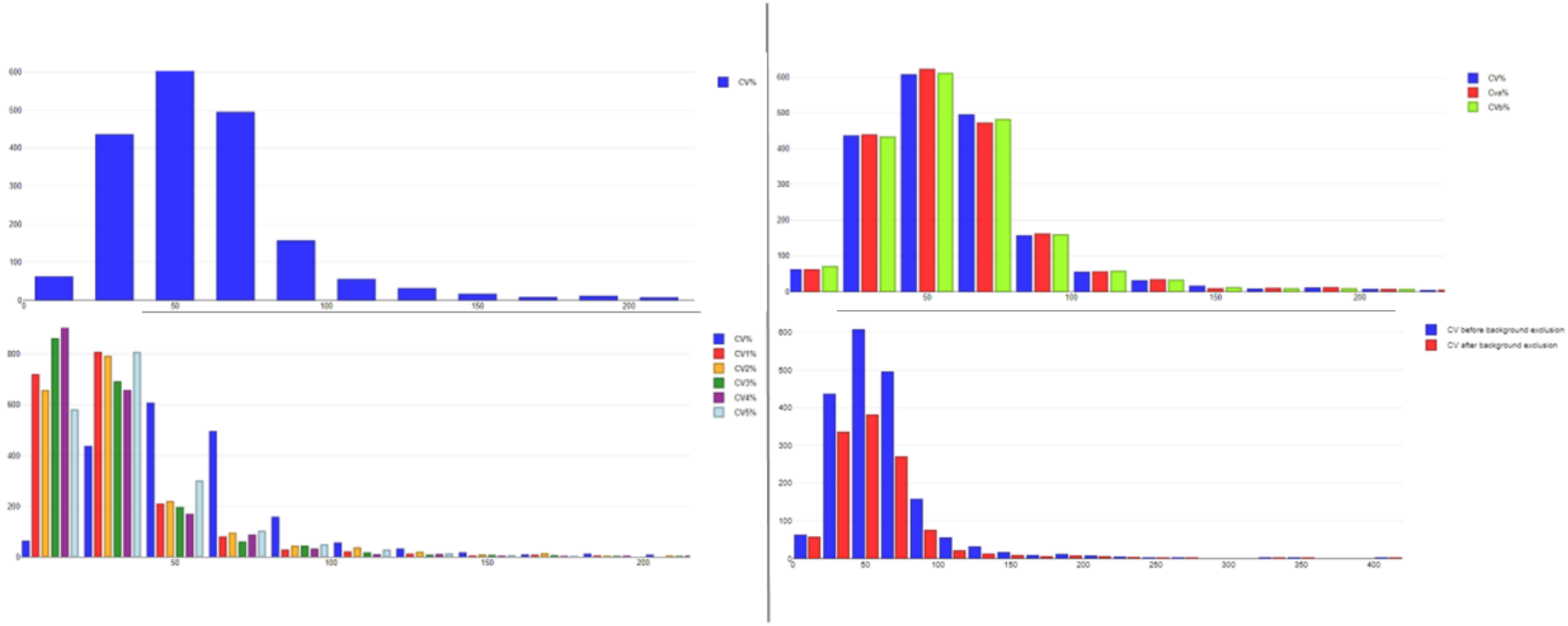
	CV%	CV%<20	Minimum CVa%/ CVb%	CVa%<20% CVb%<20%
Auto-optimized	Min:6.18	61	CVa%: 5.91 CVb%: 6.32	CVa%: 61 CVb%: 70
Centwave (mzdiff=0.01)	Min:14.06	9	CVa%: 14.26 CVb%: 8.69	CVa%: 11 CVb%: 9
Centwave (mzdiff=0.005)	Min: 14.06	9	CVa%: 8.67 CVb%: 14.31	CVa%: 9 CVb%: 11
Massifquant	Min:11.64	25	CVa%: 11.56 CVb%: 12.29	CVa%: 8 CVb%: 11
Matchedfilter	Min: 4.256	66	CVa%: 4.20 CVb%: 4.23	CVa%: 71 CVb%: 75

Όπως φαίνεται από τον παραπάνω πίνακα οι τιμές του CVa% και CVb% προσδιορίστηκαν στις περισσότερες περιπτώσεις παραπλήσιες μεταξύ τους και αρκετά κοντά με την τιμή του συνολικού CV%. Μόνο στις αναλύσεις που έγιναν με τον αλγόριθμο CentWave δεν φαίνεται να ισχύει αυτό, οι συγκεκριμένες τιμές εμφανίζουν αξιοσημείωτες διαφορές.

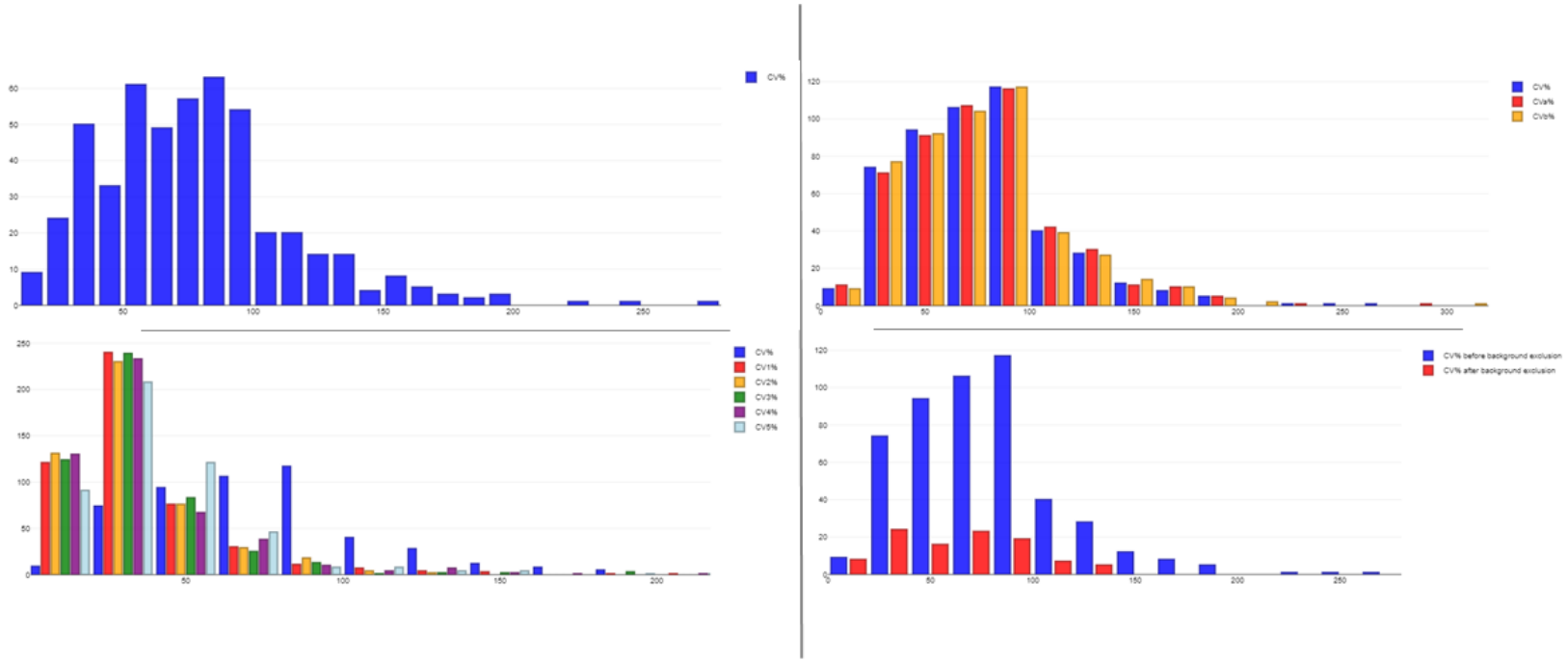
Μετά τον υπολογισμό του CV% για το συνολικό πλήθος των δειγμάτων, παρατηρήθηκε πως οι τιμές αυτές ήταν αρκετά αυξημένες σε σχέση με τις αναμενόμενες και πολύ μικρός αριθμός αυτών βρισκόταν σε τιμές <20%, γεγονός σημαντικό για την μετέπειτα στατιστική ανάλυση. Για ένα feature είναι σημαντικό η τιμή του CV% να είναι μικρότερη του 20% γιατί δείχνει πως η προσδιοριζόμενη τιμή της έντασής του είναι σχετικά σταθερή σε όλο το πλήθος των δειγμάτων. Εξαιτίας της συγκεκριμένης παρατήρησης υπολογίστηκε η τιμή του CV% per batch. Οι συγκεκριμένες τιμές ήταν πλησιέστερες στις αναμενόμενες. Επομένως, κρίθηκε χρήσιμη η προσπάθεια διόρθωσης του φαινομένου παρτίδας (batch effect correction). Πριν γίνει αυτή η διόρθωση όμως έγινε προσπάθεια να αφαιρεθούν από τα αποτελέσματα feature

που είναι πιθανόν να αντιστοιχούν σε θόρυβο (background exclusion), όπως περιγράφεται αναλυτικά στην αντίστοιχη υποενότητα της πειραματικής διαδικασίας.

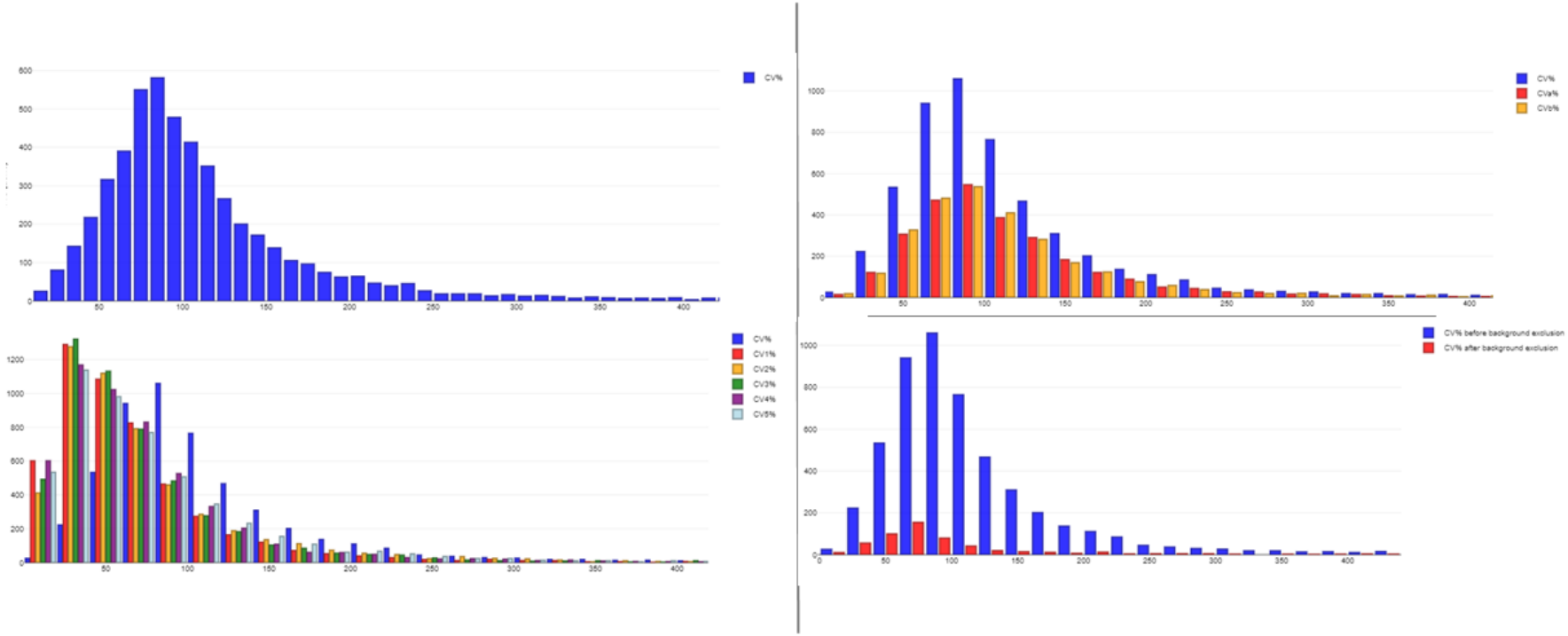
Από τον παραπάνω πίνακα παρατηρούμε πως μετά την προσπάθεια αφαίρεσης των feature που αντιστοιχούν σε θόρυβο ο συνολικός αριθμός τους αριθμός μειώθηκε αρκετά σε σύγκριση με τον αρχικό αριθμό feature που είχαν ανιχνευθεί. Παρόλα αυτά, δεν άλλαξε σχεδόν καθόλου ο αριθμός των feature με CV%<20%. Γεγονός που επιβεβαιώνει πως τα συγκεκριμένα feature είναι μεγαλύτερης σταθερότητας και είναι δύσκολο να αντιστοιχούν σε σήμα υποβάθρου, ο οποίος εμπίπτει στα τυχαία σήματα μίας αναλυτικής μεθόδου και είναι δύσκολο να εμφανίζει σταθερότητα σε τόσο μεγάλο βαθμό. Είναι πιθανό όμως feature τα οποία εμφανίζουν μεγάλη σταθερότητα να μην αντιστοιχούν σε κάποιον μεταβολίτη, αλλά σε κάποιον από τους αναλύτες που χρησιμοποιήθηκαν.



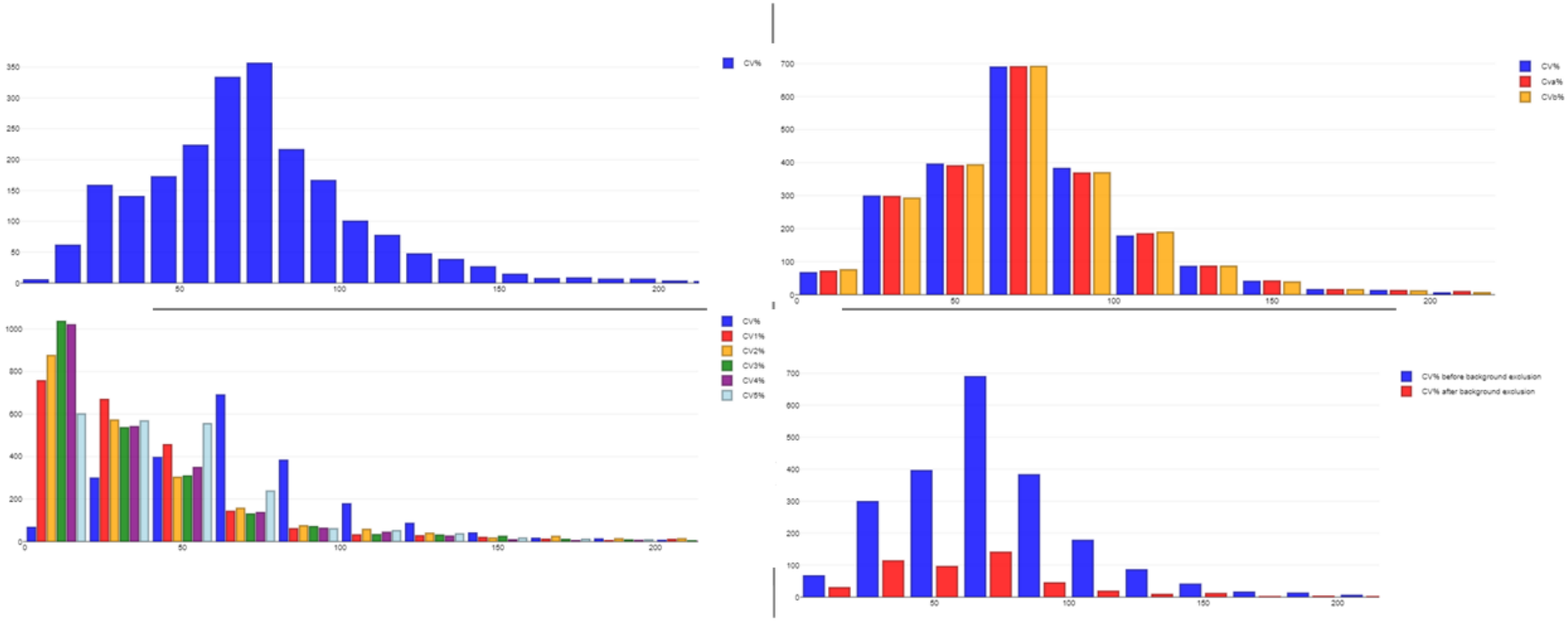
Εικόνα 6: Απεικόνιση με μορφή ιστογράμματος του συντελεστή μεταβλητότητας (CV) για την ανάλυση με τις παραμέτρους auto-optimized. a) Η κατανομή του συνολικού CV. b) Σύγκριση του συνολικού CV, με τα CVa και CVb.. c) Σύγκριση του συνολικού CV με τα CV per batch. d) Σύγκριση του CV πριν και μετά το background exclusion.



Εικόνα 7: Απεικόνιση με μορφή ιστογράμματος του συντελεστή μεταβλητότητας (CV) για την ανάλυση με τον αλγόριθμο CentWave. a) Η κατανομή του συνολικού CV. b) Σύγκριση του συνολικού CV, με τα CVa και CVb.. c) Σύγκριση του συνολικού CV με τα CV per batch. d) Σύγκριση του CV πριν και μετά το background exclusion.



Εικόνα 8: Απεικόνιση με μορφή ιστογράμματος του συντελεστή μεταβλητότητας (CV) για την ανάλυση με τον αλγόριθμο Massifquant a) Η κατανομή του συνολικού CV. b) Σύγκριση του συνολικού CV, με τα CVa και CVb.. c) Σύγκριση του συνολικού CV με τα CV per batch. d) Σύγκριση του CV πριν και μετά το background exclusion.

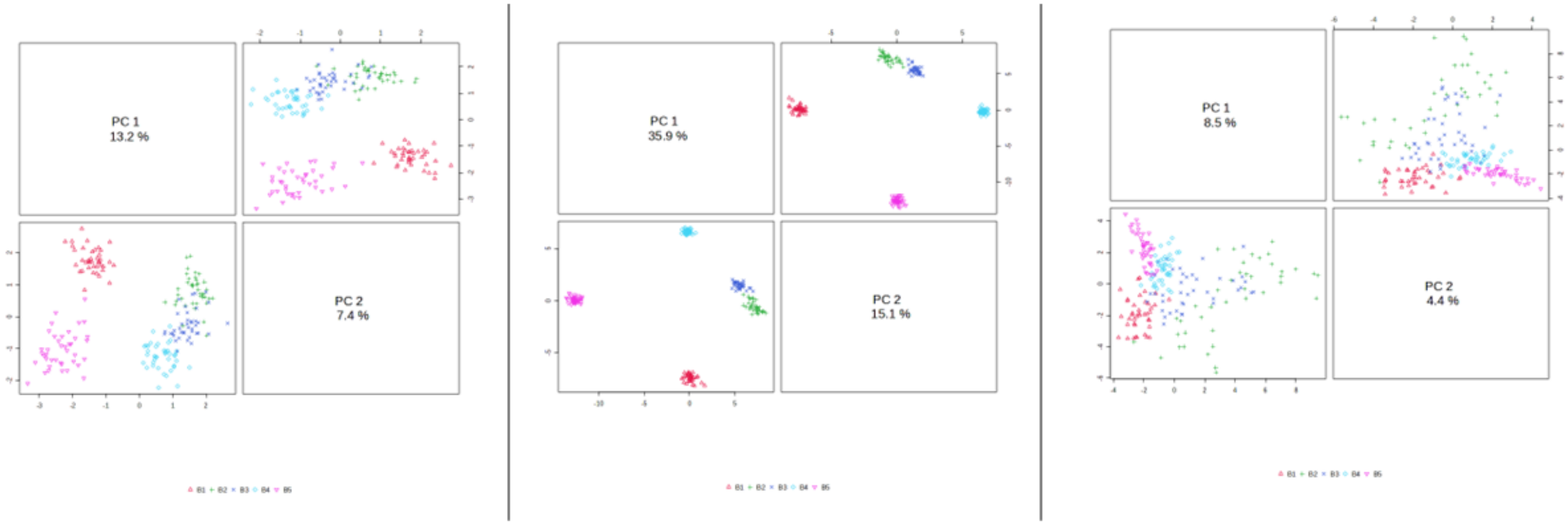


Εικόνα 9: Απεικόνιση με μορφή ιστογράμματος του συντελεστή μεταβλητότητας (CV) για την ανάλυση με τον αλγόριθμο Massifquant a) Η κατανομή του συνολικού CV. b) Σύγκριση του συνολικού CV, με τα CVa και CVb.. c) Σύγκριση του συνολικού CV με τα CV per batch. d) Σύγκριση του CV πριν και μετά το background exclusion.

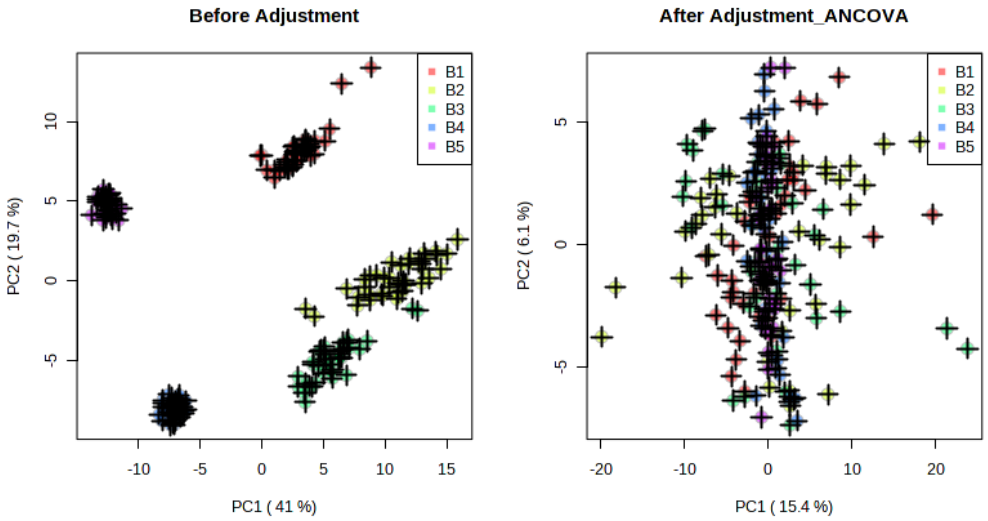
Η διόρθωση παρτίδας (batch effect correction) έγινε μέσω της επιλογής που προσφέρει το MetaboAnalyst για την διόρθωση του φαινομένου αυτού. Το λογισμικό δίνει την δυνατότητα επιλογής αρκετών παραμέτρων για τον συγκεκριμένο σκοπό, επιτρέποντας στο χρήστη επαρκή έλεγχο των συνθηκών κατά τη διόρθωση. Συγκεκριμένα, έγινε επιλογή του αλγορίθμου auto (default). Σε πρώτη φάση έγινε διόρθωση του φαινομένου παρτίδας με την επιλογή που αντικαθιστά τα ελλείποντα δεδομένα με το 1/5 της μικρότερης τιμής. Αυτό εφαρμόστηκε στα αναλυτικά αποτελέσματα όλων των αλγορίθμων.

Πίνακας 4: Σύνοψη στατιστικών αποτελεσμάτων μετά την αφαίρεση θορύβου και μετά από την διόρθωση παρτίδας για τα αποτελέσματα των αναλύσεων όλων των αλγορίθμων με την επιλογή LoDs (1/5 of the minimum) για τα ελλείποντα δεδομένα

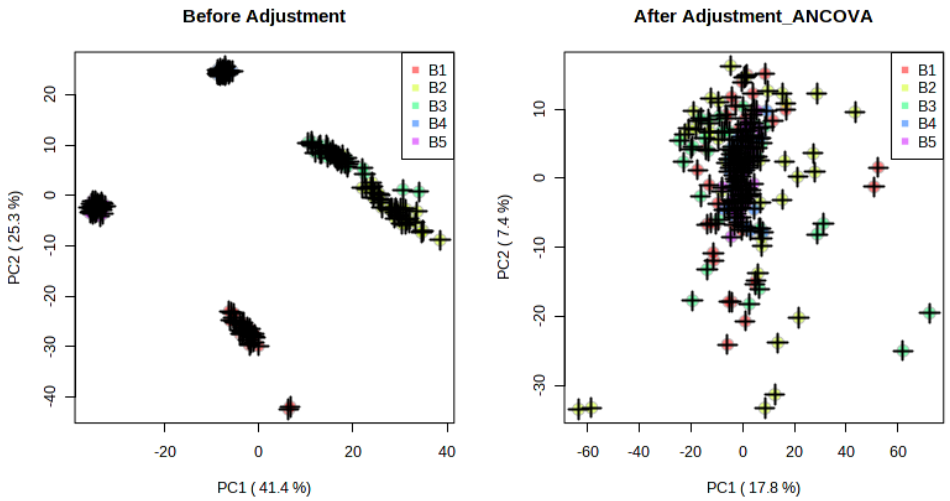
	Αρχικά					Μετά την αφαίρεση θορύβου		Μετά τη διόρθωση παρτίδων		
Αλγόριθμος	**Αριθμός συνολικών features**	**Εύρος CV%**	**Πλήθος CV%<20**	**Πλήθος CV% <20% ανά παρτίδα**	**Features που ανιχνεύθηκαν στο 100% των QC (% του συνόλου)**	**Πλήθος features**	**Πλήθος feature με CV%<20% (% feature μετά την αφαίρεση θορύβου)**	**Εύρος CV%**	**Πλήθος CV%<20**	**Πλήθος CV% <20% ανά παρτίδα**
Auto-optimized	1882	6.18 - 405.05	61	CV1:719 CV2:656 CV3:861 CV4:919 CV5:580	1758 (93%)	1172	57 (4.8%)	2.74 - 343.45	502	CV1%: 596 CV2%: 499 CV3%: 587 CV4%: 760 CV5%: 791
Centwave (mzdiff= 0.01)	496	14.06 - 276.20	9	CV1:121 CV2:131 CV3: 124 CV4: 130 CV5: 91	336 (68%)	188	8 (4.2%)	6.76 - 198.22	44	CV1%: 101 CV2%: 62 CV3%: 55 CV4%: 88 CV5%: 128
Centwave (mzdiff= 0.005)	496	14.06 - 276.20	9	CV1: 121 CV2: 131 CV3: 124 CV4: 130 CV5: 91	336 (68%)	188	8 (4.2%)	8.43 - 197.64	39	CV1%: 77 CV2%: 52 CV3%: 41 CV4%: 85 CV5%: 122
Massifquant	5128	11.64 - 1292.54	25	CV1: 601 CV2: 411 CV3: 492 CV4: 601 CV5: 534	2680 (52%)	2595	17 (0.7%)	5.05 - 683.32	163	CV1%: 886 CV2%: 391 CV3%: 434 CV4%: 734 CV5%: 1231
Matchedfilter	2192	4.26 - 586.37	66	CV1: 756 CV2: 875 CV3: 1035 CV4: 1019 CV5: 600	1031 (47%)	1154	58 (5%)	2.66 - 262.01	543	CV1%: 663 CV2%: 504 CV3%: 592 CV4%: 816 CV5%: 880



Εικόνα 10: Διαγραμματική απεικόνιση (PCA) για τα αναλυτικά αποτελέσματα με τις auto-optimized παραμέτρους α) για τα δεδομένα πριν την αφαίρεση των feature που πιθανόν να αντιστοιχούν σε θόρυβο (background exclusion) β) μετά το background exclusion γ) μετά την διόρθωση του φαινομένου παρτίδας με την επιλογή η οποία αντικαθιστά τα ελλείποντα δεδομένα με το 1/5 της μικρότερης τιμής.



Εικόνα 11: Διαγραμματική απεικόνιση PCA για τα δεδομένα της ανάλυσης με τον αλγόριθμο CentWave πριν και μετά την διόρθωση του φαινομένου παρτίδας.



Εικόνα 12: Διαγραμματική απεικόνιση PCA για τα δεδομένα της ανάλυσης με τον αλγόριθμο Matchedfilter πριν και μετά την διόρθωση του φαινομένου παρτίδας.

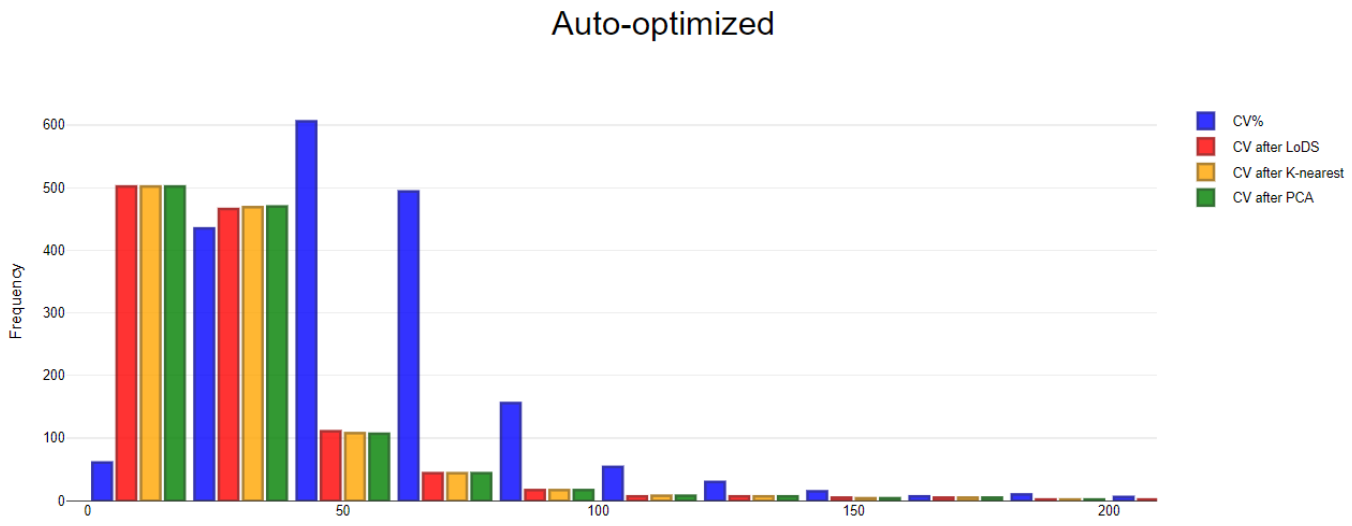
Τα αποτελέσματα που προέκυψαν μετά τη διόρθωση του φαινομένου παρτίδας ήταν εμφανώς βελτιωμένα σε όλους τους αλγορίθμους, όπως φαίνεται στον πίνακα 4. Ο αριθμός των feature που έχουν CV<20% είναι αυξημένος τόσο συνολικά, όσο και ξεχωριστά ανά παρτίδα.Το γεγονός αυτό αποδεικνύει πως η αρχική απόκλιση των τιμών CV% από τις ιδανικές είχε προκληθεί σε μεγάλο βαθμό από το συγκεκριμένο φαινόμενο. Επιπλέον, η βελτίωση της συγκεκριμένης εικόνας μετά από την χρήση του κατάλληλου

εργαλείου επιβεβαιώνει και την αποτελεσματική λειτουργία της επιλογής αυτής στον διαδικτυακό εργαλείο MetaboAnalyst.

Έπειτα, χρησιμοποιήθηκαν μόνο τα αναλυτικά αποτελέσματα που προέκυψαν από την ανάλυση με τις auto-optimized παραμέτρους, αφού αυτές φαίνονται να ανταποκρίνονται καλύτερα στα δεδομένα της συγκεκριμένης εργασίας. Στα αποτελέσματα αυτά έγινε προσπάθεια για batch effect correction και με τις τρείς επιλογές αντικατάστασης των ελλειπόντων δεδομένων με σκοπό την σύγκριση για να εξεταστεί η επίδραση που έχει στα αποτελέσματα η επιλογή που ρυθμίζει την τιμή με την οποία θα αντικατασταθούν τα missing values.

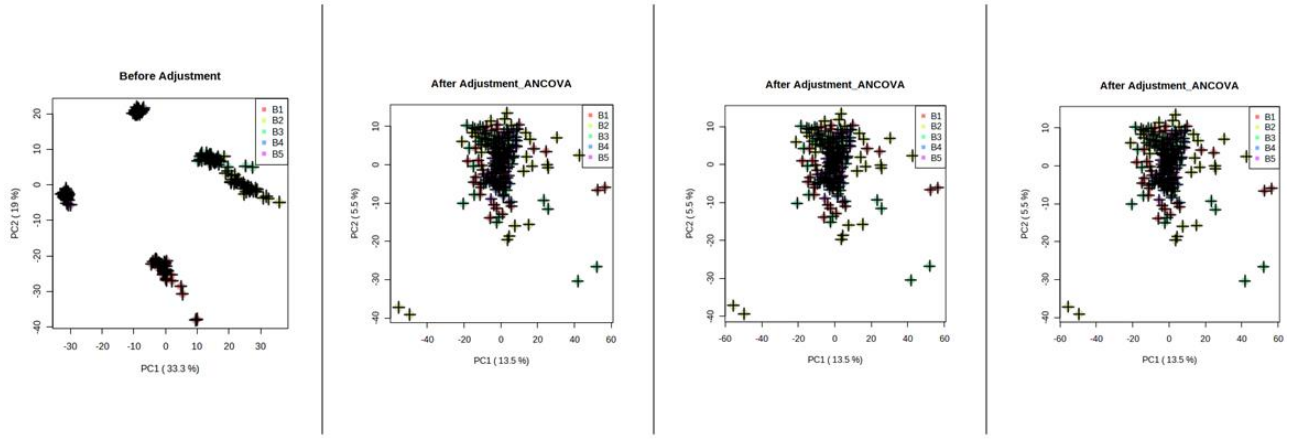
Πίνακας 5: Σύνοψη στατιστικών αποτελεσμάτων από την προσπάθεια διόρθωσης του φαινομένου παρτίδας για τα αποτελέσματα της ανάλυσης με τις auto-optimized παραμέτρους και με τις τρείς επιλογές συμπλήρωσης των ελλειπόντων δεδομένων.

	Εύρος CV%	**Πλήθος CV%< 20**	**Πλήθος CV% <20 ανά παρτίδα**
LoDs (1/5 of the minimum)	2.74 - 343.45	502	CV1%: 596 CV2%: 499 CV3%: 587 CV4%: 760 CV5%: 791
K- nearest neighbours	2.74 - 343.45	502	CV1%: 596 CV2%: 500 CV3%: 587 CV4%: 760 CV5%: 7
Probabilistic PCA	2.74 - 343.45	502	CV1%:596 CV2%:500 CV3%:587 CV4%: 760 CV5%: 786



Εικόνα 13: Σύγκριση του συνολικού CV πριν τη διόρθωση του φαινομένου παρτίδας, με τα συνολικά CV μετά την διόρθωση και με τις τρείς επιλογές συμπλήρωσης των ελλειπόντων δεδομένων.

Από την εικόνα 14 μπορούμε να διακρίνουμε πως αρχικά είναι ευδιάκριτες οι 5 παρτίδες, χωρίς όμως σαφή όρια. Αυτό συμβαίνει γιατί στα δεδομένα υπάρχουν feature τα οποία πιθανόν αντιστοιχούν σε θόρυβο και οι συγκεκριμένες κορυφές δεν παρουσιάζουν επαναληψιμότητα στην εμφάνισή τους. Αντίθετα, στο δεύτερο διάγραμμα PCA, μετά την αφαίρεση του θορύβου, οι παρτίδες έχουν σαφή όρια μεταξύ τους. Τέλος, μετά τη διόρθωση του συγκεκριμένου φαινομένου (batch effect), οι παρτίδες φαίνεται να έχουν πλησιάσει μεταξύ τους, γεγονός που επιβεβαιώνει τον σκοπό εφαρμογής της διόρθωσης.



Εικόνα 14: Διαγραμματική απεικόνιση (PCA) για τα αναλυτικά αποτελέσματα με τις auto-optimized παραμέτρους a) πριν τη διόρθωση του φαινομένου παρτίδας, b) μετά την διόρθωση του φαινομένου παρτίδας με την επιλογή LoDS (1/5 of the minimum) c) μετά την διόρθωση του φαινομένου παρτίδας με την επιλογή k-nearest neighbours, d) μετά την διόρθωση του φαινομένου παρτίδας με την επιλογή Probabilistic PCA.

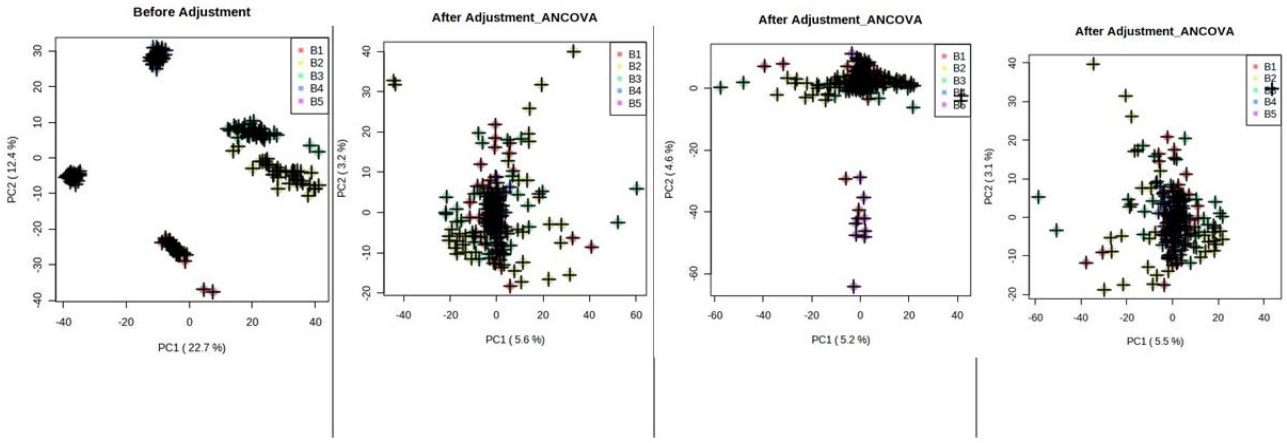
Παρατηρήθηκε πως η εφαρμογή και των τριών επιλογών για τον τρόπο αντικατάστασης των missing values στα αποτελέσματα της ανάλυσης με τις auto-optimized παραμέτρους οδήγησε στην δημιουργία αποτελεσμάτων τα οποία από τη μία ήταν σαφώς βελτιωμένα (σε σύγκριση με τα δεδομένα πριν την διόρθωση του φαινομένου παρτίδας), όμως παρόμοια μεταξύ τους και με τις τρεις επιλογές διαχείρισης των ελλειπουσών τιμών. Το γεγονός αυτό αποδόθηκε σε μεγάλο βαθμό στο ότι ο συγκεκριμένος αλγόριθμος είχε δημιουργήσει δεδομένα με ελάχιστα ελλείποντα δεδομένα (missing values). Επομένως, αποφασίστηκε να εφαρμοστούν και οι τρείς επιλογές για την αντικατάσταση των ελλείποντων δεδομένων στον αλγόριθμο του οποίου τα αποτελέσματα εμφάνιζαν τα περισσότερα ελλείποντα δεδομένα, με σκοπό να συγκριθούν τα αποτελέσματα που θα προκύψουν. Για τον σκοπό αυτό χρησιμοποιήθηκαν τα δεδομένα του αλγορίθμου massifquant.

Πίνακας 6: Σύνοψη στατιστικών αποτελεσμάτων από την προσπάθεια διόρθωσης του φαινομένου παρτίδας για τα αποτελέσματα της ανάλυσης με τον αλγόριθμο massifquant και με τις τρείς επιλογές συμπλήρωσης των ελλειπόντων δεδομένων.

	Εύρος CV%	Πλήθος CV%< 20	Πλήθος CV% <20 ανά παρτίδα
LoDs (1/5 of the minimum)	5.05 - 683.32	163	CV1%: 886 CV2%: 391 CV3%: 434 CV4%: 734 CV5%: 1231
K- nearest neighbours	5.05 - 619.82	164	CV1%: 824 CV2%: 414 CV3%: 477 CV4%: 744 CV5%: 1075
Probabilistic PCA	5.05 - 632.33	165	CV1%:790 CV2%:350 CV3%:410 CV4%: 532 CV5%: 1015

Τα παραπάνω αποτελέσματα φαίνονται βελτιωμένα σε μεγάλο βαθμό σε σύγκριση με τα αρχικά στατιστικά αποτελέσματα που προέκυψαν από την ανάλυση με τη χρήση του αλγορίθμου massifquant. Σε αντίθεση όμως με τα αποτελέσματα μετά την διόρθωση του φαινομένου παρτίδας στην ανάλυση με τις auto-optimized παραμέτρους, τα συγκεκριμένα αποτελέσματα δεν εμφανίζουν τόσο μεγάλη ομοιότητα σχετικά με την επίδραση των τριών επιλογών συμπλήρωσης των ελλειπόντων δεδομένων. Το γεγονός αυτό επαληθεύει το γεγονός ότι οι συγκεκριμένες επιλογές επηρεάζουν περισσότερο αποτελέσματα με περισσότερα

ελλείποντα δεδομένα, όμως η επιλογή τους δεν κρίνεται υψίστης σημασίας όταν πρόκειται να εφαρμοστούν σε δεδομένα με ελάχιστες ελλείπουσες τιμές.



Εικόνα 15: Διαγραμματική απεικόνιση (PCA) για τα αναλυτικά αποτελέσματα του αλγορίθμου Massifquant a) πριν τη διόρθωση του φαινομένου παρτίδας, b) μετά την διόρθωση του φαινομένου παρτίδας με την επιλογή LoDS (1/5 of the minimum) c) μετά την διόρθωση του φαινομένου παρτίδας με την επιλογή k-nearest neighbours, d) μετά την διόρθωση του φαινομένου παρτίδας με την επιλογή Probabilistic PCA

Μετά την ολοκλήρωση της ανάλυσης των QC και Blank δειγμάτων στο MetaboAnalyst, ακολούθησε η ανάλυση των συγκεκριμένων δειγμάτων στο λογισμικό της εταιρίας AGILENT, Profinder. Στο συγκεκριμένο λογισμικό δεν υπάρχει ο περιορισμός του ανώτερου αριθμού δειγμάτων, όπως υπάρχει στο MetaboAnalyst (200 δείγματα), έτσι στο Profinder η ανάλυση των QC και των Blank έγινε ταυτόχρονα.

Τα αναλυτικά αποτελέσματα που προέκυψαν από το λογισμικό της AGILENT, Profinder, εμφανίζουν τιμές CV% πιο κοντά στις αναμενόμενες. Ο αριθμός των feature που εμφανίζει τόσο συνολικό CV<20%, όσο και CV% per batch<20% είναι αρκετά αυξημένος συγκριτικά με τις αντίστοιχες τιμές του συντελεστή μεταβλητότητας που είχαν προκύψει από τα αποτελέσματα των αλγορίθμων του MetaboAnalyst. Από το γεγονός αυτό μπορούμε να συμπεράνουμε πως το Profinder ανιχνεύει feature τα οποία παρουσιάζουν μεγαλύτερη σταθερότητα.

Επιπλέον, η συνολική εικόνα των αποτελεσμάτων του λογισμικού Profinder δείχνει πως πέρα από σταθερά feature στη συγκεκριμένη μελέτη ανιχνεύει feature τα οποία εμφανίζονται σε πολύ μικρό αριθμό δειγμάτων, δηλαδή υπάρχουν feature που έχουν ανιχνευθεί μόνο σε ένα ή δύο δείγματα. Αυτό δεν παρατηρήθηκε στις αναλύσεις που έγιναν με το εργαλείο MetaboAnalyst. Τέλος, τα αποτελέσματα του Profinder εμφανίζουν μεγάλο ποσοστό ελλειπόντων δεδομένων,

συγκεκριμένα 41%, σε αντίθεση με τα αποτελέσματα που προέκυψαν από τις αναλύσεις με το MetaboAnalyst των οποίων τα ποσοστά αυτά κυμαίνονταν σε πολύ χαμηλότερες τιμές. Πιο συγκεκριμένα, το μεγαλύτερο ποσοστό το οποίο εμφανίζεται στα αποτελέσματα του MetaboAnalyst ήταν 8.2% με τον αλγόριθμο Massifquant.

Στα αποτελέσματα του Profinder έγινε προσπάθεια να αφαιρεθούν τα feature που πιθανόν να αντιστοιχούν σε θόρυβο με δύο τρόπους. Αρχικά, με τον ίδιο τρόπο που αναφέρεται παραπάνω και έπειτα με την αντίστοιχη επιλογή που προσφέρει το πακέτο της εταιρίας AGILENT. Στον παραπάνω πίνακα φαίνεται πως αν συγκρίνουμε ποσοτικά τα συγκεκριμένα αποτελέσματα δεν απέχουν σημαντικά. Όμως, το αξιοσημείωτο είναι πως μετά από την συγκεκριμένη διαδικασία και με τους δύο τρόπους αφαιρούνται πολλά από τα 'σταθερά' feature (CV<20%).

Πίνακας 7: Σύνοψη στατιστικών αποτελεσμάτων μετά την αφαίρεση θορύβου και μετά από την διόρθωση παρτίδας για τα αποτελέσματα των αναλύσεων όλων των αλγορίθμων, συμπεριλαμβανομένου του Profinder, με την επιλογή LoDs (1/5 of the minimum) για τα ελλείποντα δεδομένα

	Αρχικά					Μετά την αφαίρεση θορύβου		Μετά τη διόρθωση παρτίδων		
Αλγόριθμος	**Αριθμός συνολικών features**	**Εύρος CV%**	**Πλήθος CV%<20**	**Πλήθος CV% <20% ανά παρτίδα**	**Features που ανιχνεύθηκαν στο 100% των QC (% του συνόλου)**	**Πλήθος features**	**Πλήθος feature με CV%<20% (% feature μετά την αφαίρεση θορύβου)**	**Εύρος CV%**	**Πλήθος CV%<20**	**Πλήθος CV% <20% ανά παρτίδα**
Auto-optimized	1882	6.18 - 405.05	61	CV1:719 CV2:656 CV3:861 CV4:919 CV5:580	1758 (93%)	1172	57 (4.8%)	2.74 - 343.45	502	CV1%: 596 CV2%: 499 CV3%: 587 CV4%: 760 CV5%: 791
Centwave (mzdiff= 0.01)	496	14.06 - 276.20	9	CV1:121 CV2:131 CV3: 124 CV4: 130 CV5: 91	336 (68%)	188	8 (4.2%)	6.76 - 198.22	44	CV1%: 101 CV2%: 62 CV3%: 55 CV4%: 88 CV5%: 128
Centwave (mzdiff= 0.005)	496	14.06 - 276.20	9	CV1: 121 CV2: 131 CV3: 124 CV4: 130 CV5: 91	336 (68%)	188	8 (4.2%)	8.43 - 197.64	39	CV1%: 77 CV2%: 52 CV3%: 41 CV4%: 85 CV5%: 122
Massifquant	5128	11.64 - 1292.54	25	CV1: 601 CV2: 411 CV3: 492 CV4: 601 CV5: 534	2680 (52%)	2595	17 (0.7%)	5.05 - 683.32	163	CV1%: 886 CV2%: 391 CV3%: 434 CV4%: 734 CV5%: 1231
Matchedfilter	2192	4.26 - 586.37	66	CV1: 756 CV2: 875 CV3: 1035 CV4: 1019 CV5: 600	1031 (47%)	1154	58 (5%)	2.66 - 262.01	543	CV1%: 663 CV2%: 504 CV3%: 592 CV4%: 816 CV5%: 880
Profinder	3780	0.01 - 473.01	107	CV1:961 CV2:1020 CV3: 1201 CV4:1288 CV5:893	1101 (29%)	3035	56 (1.8%)	4.3 - 633.18	540	CV1%: 1357 CV2%: 1042 CV3%: 1058 CV4%: 1310 CV5%: 1431

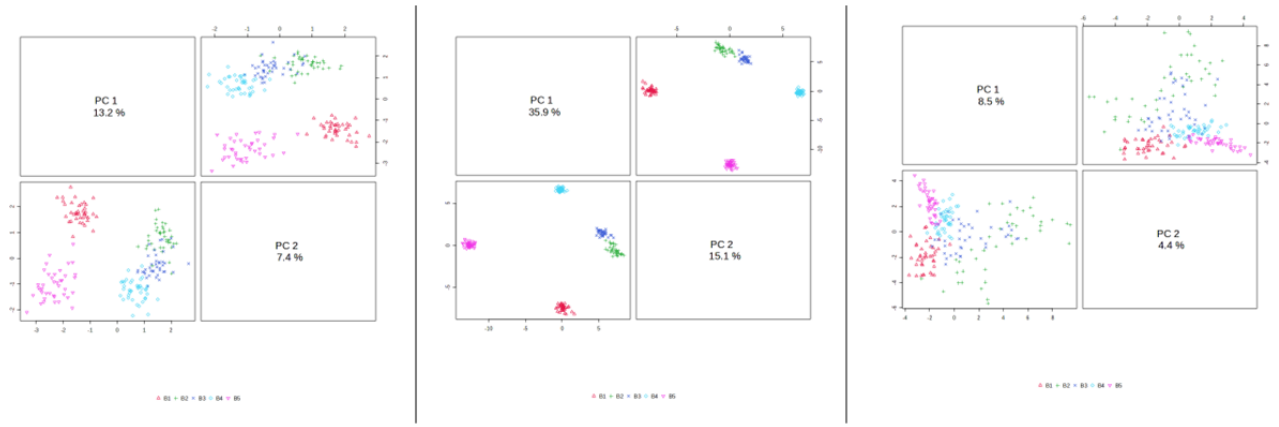
Πίνακας 8: Σύνοψη στατιστικών αποτελεσμάτων από την προσπάθεια διόρθωσης του φαινομένου παρτίδας για τα αποτελέσματα της ανάλυσης με το λογισμικό Profinder και με τις τρείς επιλογές συμπλήρωσης των ελλειπόντων δεδομένων.

	Εύρος CV%	Πλήθος CV%< 20	Πλήθος CV% <20 ανά παρτίδα
LoDs (1/5 of the minimum)	4.3 - 633.18	540	CV1%: 1357 CV2%: 1042 CV3%: 1058 CV4%: 1310 CV5%: 1431
K- nearest neighbours	4.27 - 301.34	1038	CV1%: 1759 CV2%: 1381 CV3%: 1428 CV4%: 1658 CV5%: 1730
Probabilistic PCA	0.01 - 333.76	1321	CV1%: 1660 CV2%: 1509 CV3%: 1601 CV4%: 1931 CV5%: 1940

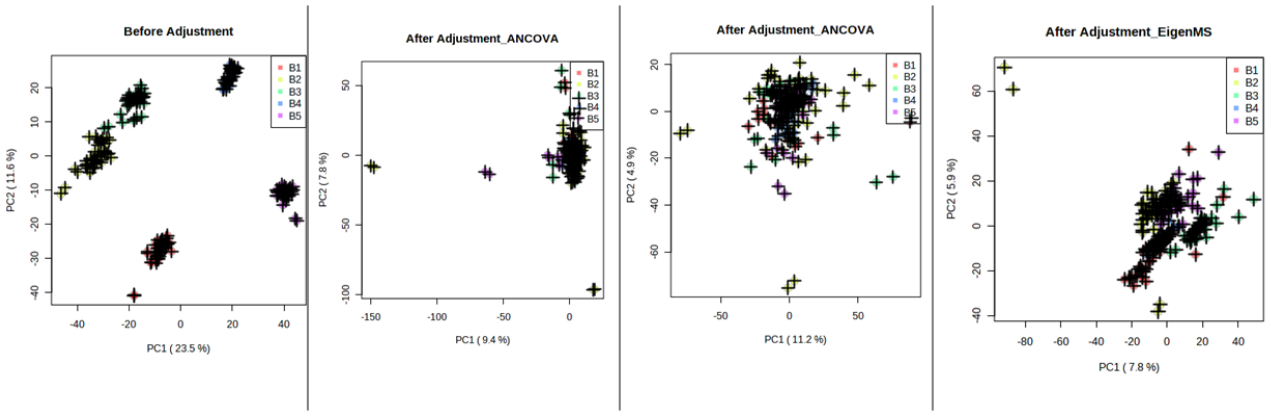
Όπως αναφέρθηκε και παραπάνω, τα αναλυτικά αποτελέσματα που προέκυψαν από το λογισμικό Profinder εμφανίζουν μεγάλα ποσοστά ελλειπόντων δεδομένων. Το γεγονός αυτό αναμένεται να επηρεάζει αρκετά τα αποτελέσματα που προκύπτουν μετά την προσπάθεια διόρθωσης του φαινομένου παρτίδας. Για να εξετασθεί το κατά πόσο οι τιμές του παραπάνω πίνακα ανταποκρίνονται στα πραγματικά αποτελέσματα και δεν έχει μεγάλη επίδραση σε αυτά η αντικατάσταση των missing values, δημιουργήθηκε ένας πίνακας με τα αντίστοιχα δεδομένα μόνο των feature που ανιχνεύονται σε ποσοστό μεγαλύτερο ή ίσο του 60% των συνολικών δειγμάτων.

Πίνακας 9: Σύνοψη στατιστικών αποτελεσμάτων από την προσπάθεια διόρθωσης του φαινομένου παρτίδας για τα αποτελέσματα της ανάλυσης με το λογισμικό Profinder και με τις τρείς επιλογές συμπλήρωσης των ελλειπόντων δεδομένων, μόνο στα feature που εντοπίζονται σε ποσοστό μεγαλύτερο ή ίσο του 60% των δειγμάτων.

	CV%	**CV%< 20**	**CV per batch< 20%**
LoDs (1/5 of the minimum)	Min: 4.3 Max: 470.1	375	CV1%: 718 CV2%: 489 CV3%: 549 CV4%: 720 CV5%: 770
K- nearest neighbours	Min: 4.26 Max: 284.8	514	CV1%: 892 CV2%: 604 CV3%: 702 CV4%: 860 CV5%: 872
Probabilistic PCA	Min: 0.01 Max: 269.3	585	CV1%: 751 CV2%: 638 CV3%: 751 CV4%: 953 CV5%: 947



Εικόνα 16: Διαγραμματική απεικόνιση (PCA) για τα αναλυτικά αποτελέσματα του λογισμικού Profinder α) για τα δεδομένα πριν την αφαίρεση των feature που πιθανόν να αντιστοιχούν σε θόρυβο (background exclusion) β) μετά το background exclusion γ) μετά την διόρθωση του φαινομένου παρτίδας με την επιλογή η οποία αντικαθιστά τα ελλείποντα δεδομένα με το 1/5 της μικρότερης τιμής.



Εικόνα 17: Διαγραμματική απεικόνιση (PCA) για τα αναλυτικά αποτελέσματα που προέκυψαν από το λογισμικό Profinder a) πριν τη διόρθωση του φαινομένου παρτίδας, b) μετά την διόρθωση του φαινομένου παρτίδας με την επιλογή LoDS (1/5 of the minimum) c) μετά την διόρθωση του φαινομένου παρτίδας με την επιλογή k-nearest neighbours, d) μετά την διόρθωση του φαινομένου παρτίδας με την επιλογή Probabilistic PCA

Τα παραπάνω αποτελέσματα δημιουργήθηκαν με την χρήση της επιλογής για διόρθωση του φαινομένου παρτίδας που προσφέρει το MetaboAnalyst. Είναι βελτιωμένα σε μεγάλο βαθμό συγκριτικά με τα αποτελέσματα πριν από τη διόρθωση παρτίδας, όπως αυτά παρουσιάζονται στον πίνακα 8. Επιπροσθέτως, παρατηρούμε πως τα αποτελέσματα μεταξύ των τριών επιλογών διόρθωσης των ελλειπόντων δεδομένων διαφέρουν αρκετά μεταξύ τους στην συγκεκριμένη περίπτωση. Η διαφορά αυτή πιθανόν να οφείλεται στο ότι η ανάλυση που έγινε μέσω του λογισμικού Profinder επέστρεψε αποτελέσματα με αρκετά αυξημένο ποσοστό ελλειπόντων δεδομένων, συγκριτικά με τις αναλύσεις που έγιναν μέσω του MetaboAnalyst. Σε αυτό το σημείο, όπου ο αριθμός των ελλειπόντων δεδομένων είναι υψηλός και η επίδραση που έχει η επιλογή αντικατάστασής τους πιο εμφανής, μπορούν με ευκολία να συγκριθούν τα

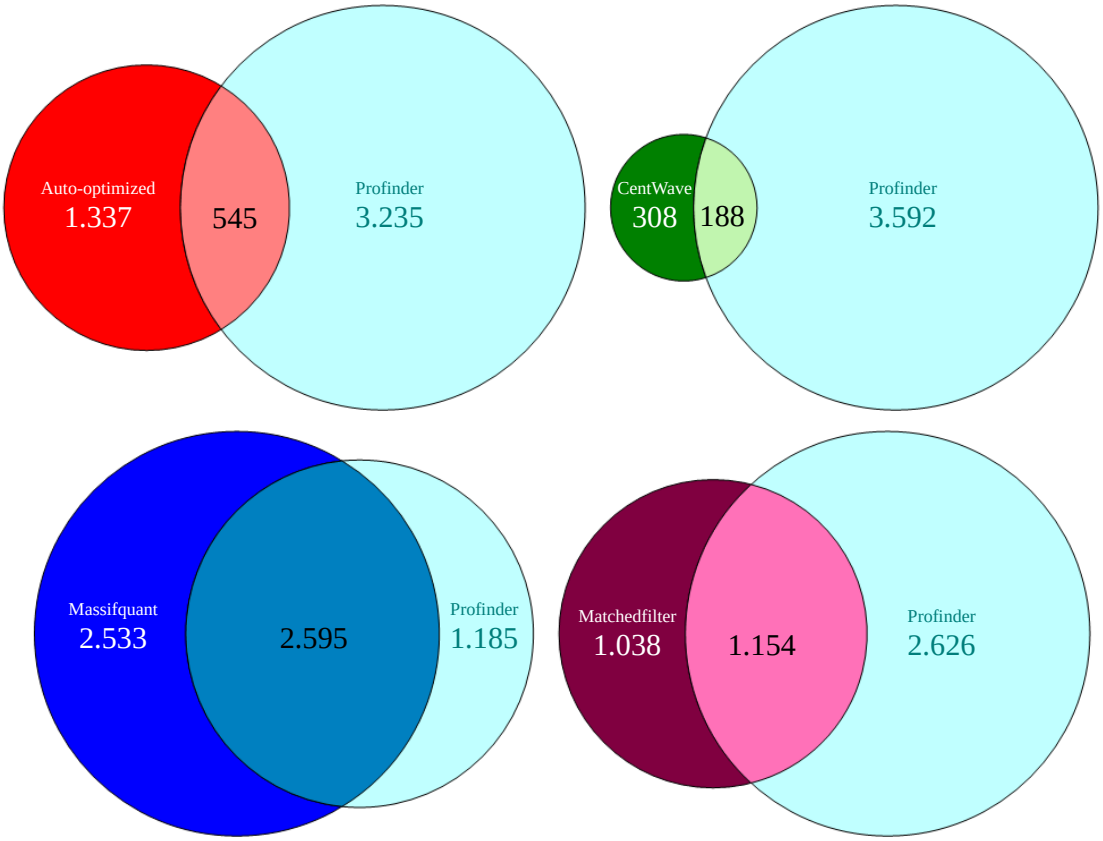
αποτελέσματά των τριών αυτών επιλογών. Από τον παραπάνω πίνακα μπορεί κανείς να παρατηρήσει πως η επιλογή Probabilistic PCA, δίνει υψηλότερες τιμές του συντελεστή μεταβλητότητας, τόσο συνολικά, όσο και ξεχωριστά ανά παρτίδα. Η εικόνα αυτή συνεχίζει να είναι εμφανής, αλλά σε μικρότερο βαθμό, στις τιμές του συντελεστή μεταβλητότητας των feature που ανιχνεύονται σε ποσοστό μεγαλύτερο ή ίσο του 60% των συνολικών δειγμάτων, όπως φαίνεται στον Πίνακα 9.

Στη συνέχεια έγινε σύγκριση των αποτελεσμάτων όλων τον αλγορίθμων του MetaboAnalyst, με αυτά του Profinder. Προσδιορίστηκαν με ακρίβεια τα κοινά feature και σε αυτά εντοπίστηκαν όσα είχαν CV<20%. Η διαδικασία σύγκρισης έγινε ανάμεσα στα αποτελέσματα του Profinder και στα αποτελέσματα που προέκυψαν με όλους τους αλγορίθμους του MetaboAnalyst, σε όλες τις περιπτώσεις συγκρίθηκαν μεταξύ τους οι λίστες των feature μετά την αφαίρεση αυτών που πιθανόν αντιστοιχούσαν σε θόρυβο.

Πίνακας 10: Σύγκριση αποτελεσμάτων μεταξύ των δύο εργαλείων (MetaboAnalyst/ Profinder).

Αλγόριθμος	Αριθμός συνολικών features	Πλήθος features μετά την αφαίρεση του θορύβου	Πλήθος κοινών feature με Profinder (% των συνολικών feature)	Πλήθος κοινών features με CV%<20%
Auto-optimized	1882	1172	545 (29%)	33 (6% των κοινών feature)
Centwave (mzdiff= 0.01)	496	188	102 (21%)	8 (7.8% των κοινών feature)
Centwave (mzdiff= 0.005)	496	188	103 (21%)	8 (7.7% των κοινών feature)
Massifquant	5128	2595	532 (10%)	10 (1.8% των κοινών feature)
Matchedfilter	2192	1154	465 (21%)	29 (6.2% των κοινών feature)

Από τον παραπάνω πίνακα μπορούμε εύκολα να παρατηρήσουμε πως την μεγαλύτερη επικάλυψη με τα δεδομένα του Profinder, εμφανίζουν αυτά που προέκυψαν με τον αλγόριθμο Matchedfilter του MetaboAnalyst.



Εικόνα 18: Διαγράμματα Venn των κοινών feature που ανιχνεύονται μεταξύ των αλγορίθμων του MetaboAnalyst και του Profinder.

Από τα διαγράμματα Venn της εικόνας 18 μπορούμε να δούμε την ποσοτική συσχέτιση ανάμεσα στα feature τα οποία ανιχνεύονται με τους αλγορίθμους του MetaboAnalyst και του Profinder. Συγκεκριμένα, στην ανάλυση με τις auto-optimized παραμέτρους από τα 1.882 συνολικά feature, τα 545 (29%) είναι κοινά με το Profinder. Στην ανάλυση με τον αλγόριθμο CentWave από τα 496, τα 188 (38%) είναι κοινά. Στην ανάλυση με τον αγόριθμο Massifquant από τα 5.128, τα 2.595 (51%) είναι κοινά. Στην ανάλυση με τον αλγόριθμο Matchedfilter από τα 2.192, τα 1.154 (53%) είναι κοινά με το Profinder. Συμπερασματικά, το μεγαλύτερο ποσοστό επικάλυψης με τα Feature του Profinder το εμφανίζει ο αλγόριθμος Matchedfilter.

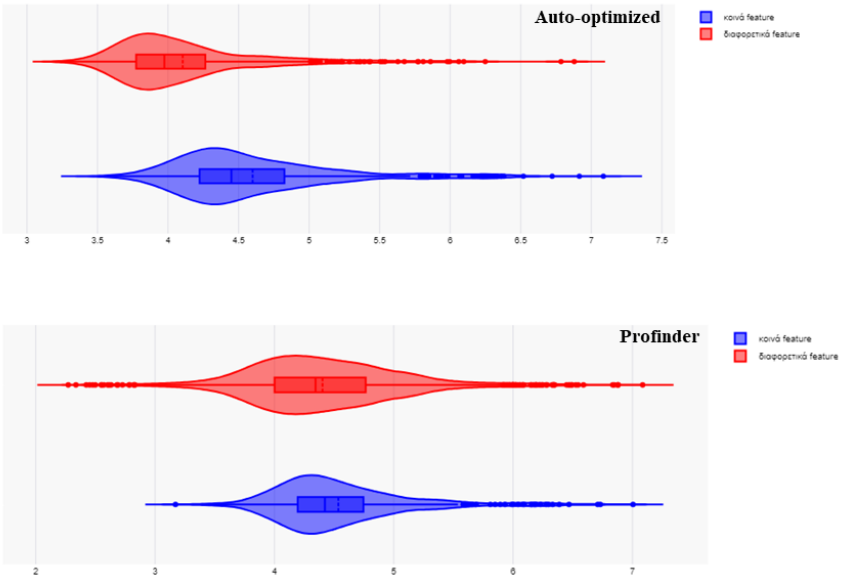
Πέρα από την ποσοτική σύγκριση των feature που ανιχνεύονται στις αναλύσεις με τους αλγορίθμους του εργαλείου MetaboAnalyst και με το Profinder και την εύρεση των κοινών feature μεταξύ τους, κρίνεται απαραίτητο να ελεγχθεί και η ποιότητα του σήματος το οποίο ανιχνεύει το κάθε εργαλείο. Για τον λόγο αυτόν, επιλέχθηκαν κάποιοι μεταβολίτες οι οποίοι είναι γνωστό ότι αποδίδονται με καλοσχηματισμένες κορυφές και σχετικά σταθερή ένταση. Η ταυτοποίηση των συγκεκριμένων feature πραγματοποιήθηκε με ανάλυση εμπορικά διαθέσιμων πρότυπων ενώσεων προκειμένου να επιβεβαιωθεί ο χρόνος κατακράτησης και η θραυσματοποίηση των εν λόγω μαζών. Έγινε αναζήτηση για αυτά τα feature στον συγκεντρωτικό πίνακα που έχει προκύψει από την ανάλυση με τις auto-optimized παραμέτρους του MetaboAnalyst και σε αυτόν από την ανάλυση με το Profinder. Αφού εντοπίστηκαν τα feature στους αντίστοιχους πίνακες, έγινε συσχέτιση των ανιχνευόμενων εντάσεων όλων των δειγμάτων στα συγκεκριμένα feature.

Πίνακας 11: Μάζα και αναμενόμενος χρόνος κατακράτησης γνωστών μεταβολιτών, μαζί με την συσχέτιση των εντάσεων τους στην ανάλυση με τις auto-optimized παραμέτρους του MetaboAnalyst και στην ανάλυση με το Profinder.

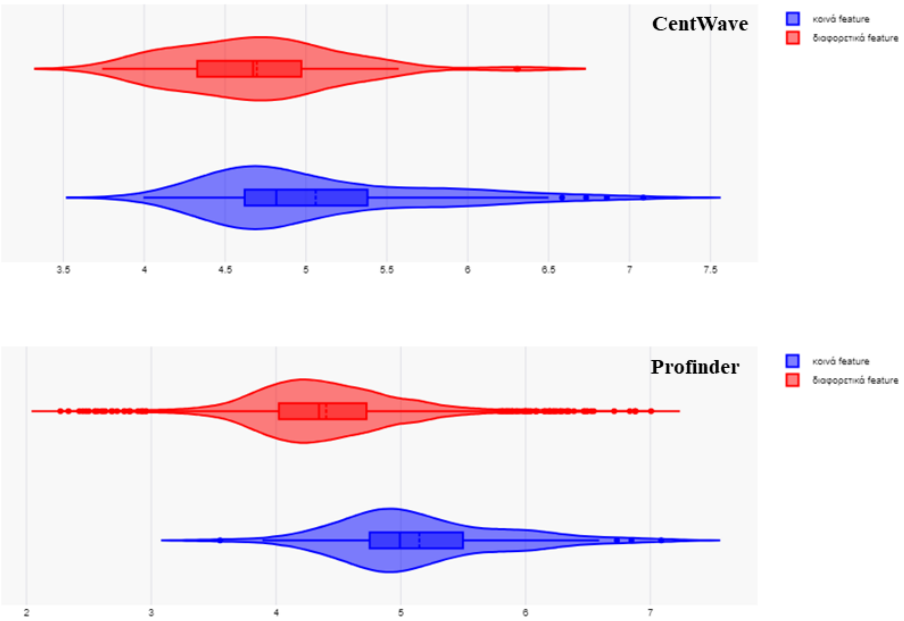
Μεταβολίτης	Μάζα	Χρόνος κατακράτησης (min)	Συντελεστής συσχέτισης R^2 Profinder/Auto-optimized
Υποξανθίνη	136.03851	1.16	0.993
Ινοσίνη	268.08077	1.66	0.998
Κυνουρενίνη	208.08479	1.9	0.999
Φαινυλαλανίνη	165.07898	2.1	0.966
Τρυπτοφάνη	204.08988	2.55	0.891
Ιππουρικό οξύ	179.05824	3.1	0.433
Ίνδολ-3-οξικό οξύ	175.06333	4.15	0.943
Κορτιζόλη	362.20932	5.19	0.996
Δεκανοϋλκαρνιτίνη	315.24096	5.28	0.999
Ρετινόλη	268.2191	7.219	0.998

Από τον παραπάνω πίνακα είναι εμφανές πως οι εντάσεις στις οποίες ανιχνεύουν και τα δύο λογισμικά τα συγκεκριμένα feature είναι παραπλήσιες, αφού η τιμές του R^2 είναι πολύ κοντά στο 1. Επομένως, συμπεραίνουμε ότι μεταβολίτες οι οποίοι δημιουργούν υψηλές εντάσεις και 'καλές' κορυφές αντιμετωπίζονται με παρόμοιο τρόπο στα δύο λογισμικά.

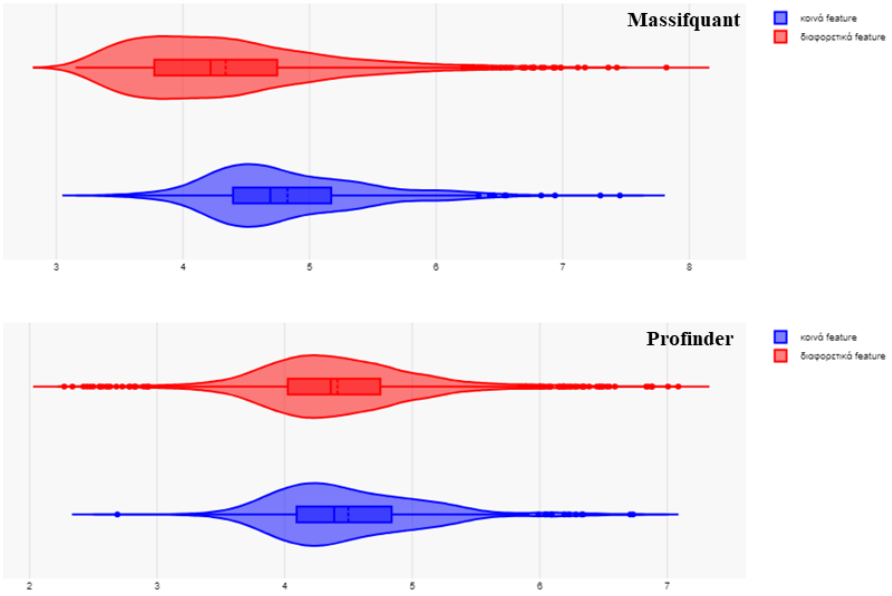
Στη συνέχεια, υπολογίστηκε ο μέσος όρος των εντάσεων των feature σε όλα τα δείγματα και έπειτα ο δεκαδικός λογάριθμος του μέσου όρου. Ακολούθησε η κατασκευή διαγραμμάτων διάχυσης για τους μέσους όρους των εντάσεων των κοινών και των διαφορετικών feature ανάμεσα στις αναλύσεις που έγιναν με τους αλγορίθμους του MetaboAnalyst και αυτή με το Profinder.



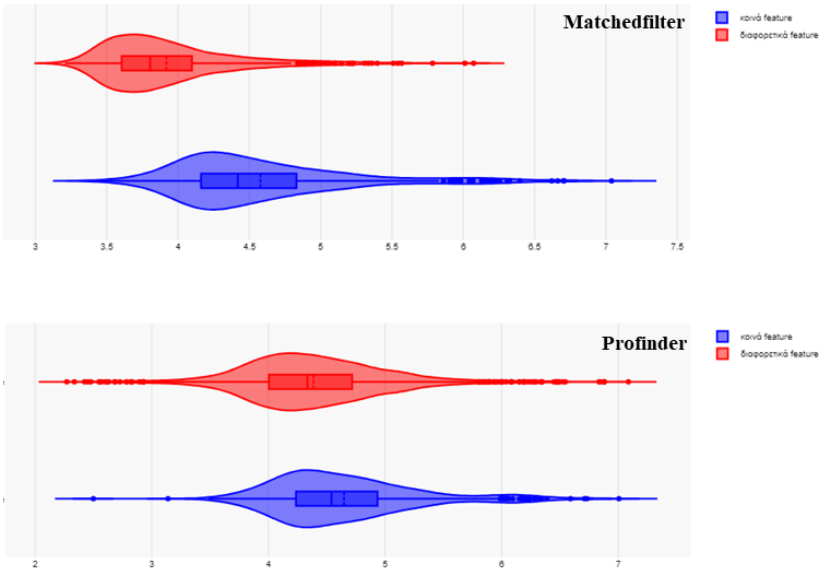
Εικόνα 19: Διάγραμμα διάχυσης του δεκαδικού λογαρίθμου του μέσου όρου των εντάσεων για τα κοινά και τα διαφορετικά feature των αναλύσεων auto-optimized- Profinder α) όπως αυτά ανιχνεύονται στην ανάλυση με το MetaboAnalyst β) όπως αυτά ανιχνεύονται στην ανάλυση με το Profinder.



Εικόνα 20: Διάγραμμα διάχυσης του δεκαδικού λογαρίθμου του μέσου όρου των εντάσεων για τα κοινά και τα διαφορετικά feature των αναλύσεων CentWave- Profinder α) όπως αυτά ανιχνεύονται στην ανάλυση με το MetaboAnalyst β) όπως αυτά ανιχνεύονται στην ανάλυση με το Profinder.

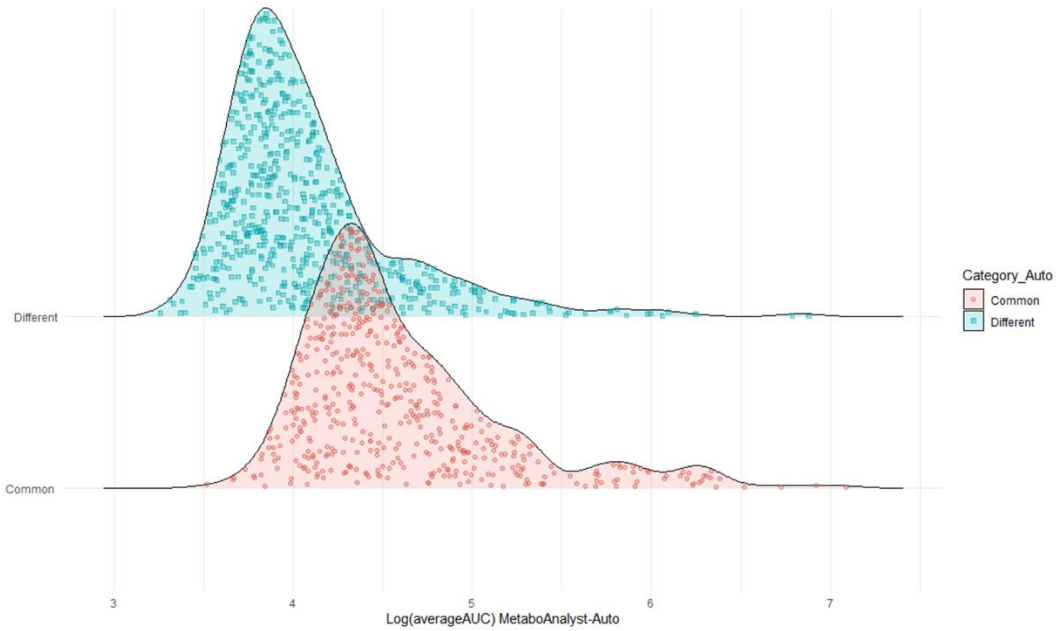


Εικόνα 21: Διάγραμμα διάχυσης του δεκαδικού λογαρίθμου του μέσου όρου των εντάσεων για τα κοινά και τα διαφορετικά feature των αναλύσεων Massifquant- Profinder α) όπως αυτά ανιχνεύονται στην ανάλυση με το MetaboAnalyst β) όπως αυτά ανιχνεύονται στην ανάλυση με το Profinder.

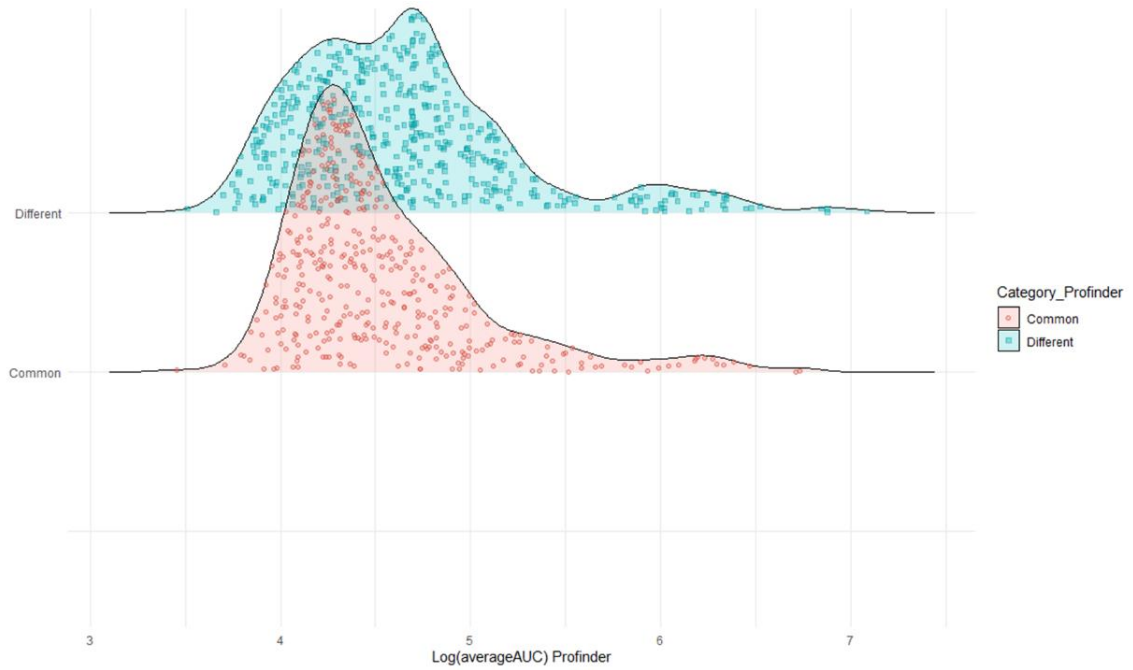


Εικόνα 22: Διάγραμμα διάχυσης του δεκαδικού λογαρίθμου του μέσου όρου των εντάσεων για τα κοινά και τα διαφορετικά feature των αναλύσεων Matchedfilter- Profinder α) όπως αυτά ανιχνεύονται στην ανάλυση με το MetaboAnalyst β) όπως αυτά ανιχνεύονται στην ανάλυση με το Profinder.

Από τα παραπάνω διαγράμματα μπορούμε εύκολα να παρατηρήσουμε πως οι τιμές του μέσου όρου των εντάσεων των feature που ανιχνεύονται και στα δύο λογισμικά είναι εμφανώς μεγαλύτερες συγκριτικά με αυτές των feature που ανιχνεύονται μόνο σε ένα από τα δύο. Στις περισσότερες περιπτώσεις το μοτίβο αυτό είναι πιο εμφανές στα διαγράμματα τα οποία οπτικοποιούν τις εντάσεις που προέρχονται από το MetaboAnalyst. Από την παρατήρηση αυτή θα μπορούσαμε ίσως να εξάγουμε το συμπέρασμα πως οι πιο υψηλές εντάσεις είναι πιο εύκολο να ανιχνευθούν από τέτοιου είδους εργαλεία, γι' αυτό τα feature που χαρακτηρίζονται από υψηλές εντάσεις συνήθως έχουν ανιχνευθεί και από τα δύο λογισμικά.

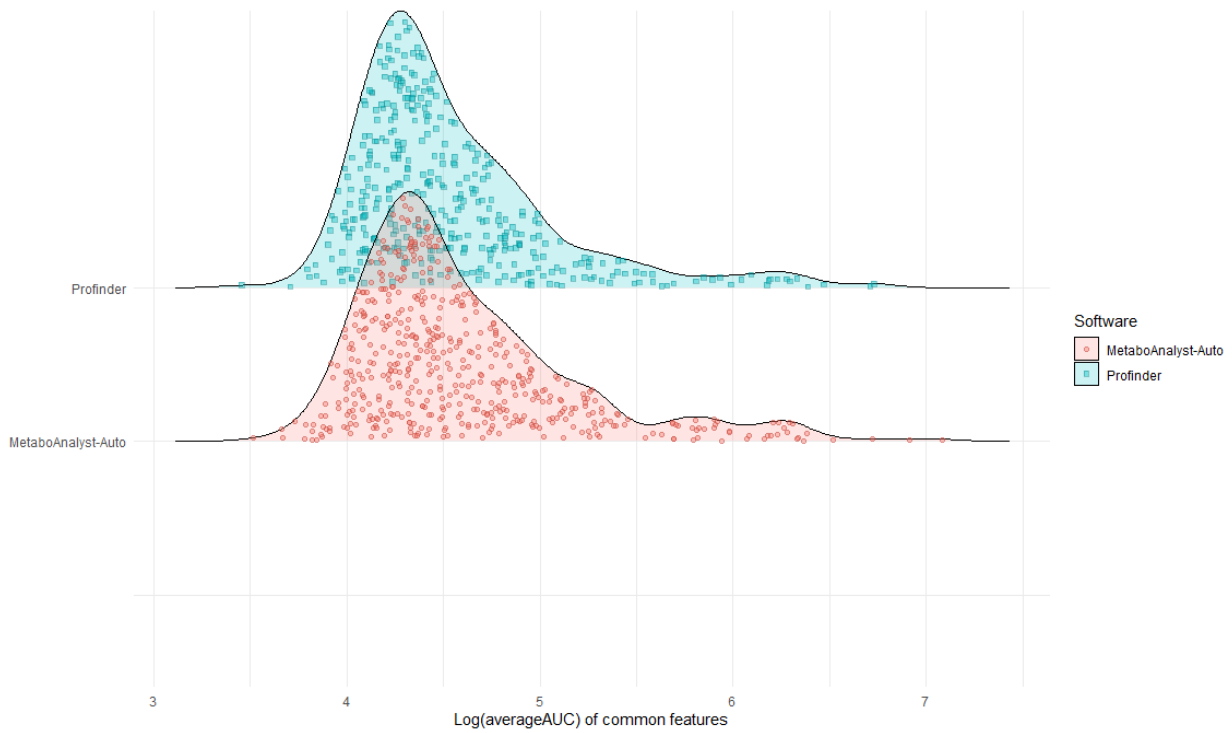


Εικόνα 23: Διάγραμμα πυκνότητας του δεκαδικού λογαρίθμου του μέσου όρου της ανιχνευόμενης έντασης από το MetaboAnalyst για τα μοναδικά feature που ανιχνεύονται από το MetaboAnalyst, αλλά και για τα feature που ανιχνεύουν και τα δύο λογισμικά.



Εικόνα 24: Διάγραμμα πυκνότητας του δεκαδικού λογαρίθμου του μέσου όρου της ανιχνευόμενης έντασης από το Profinder για τα μοναδικά feature που ανιχνεύονται από το Profinder, αλλά και για τα feature που ανιχνεύουν και τα δύο λογισμικά.

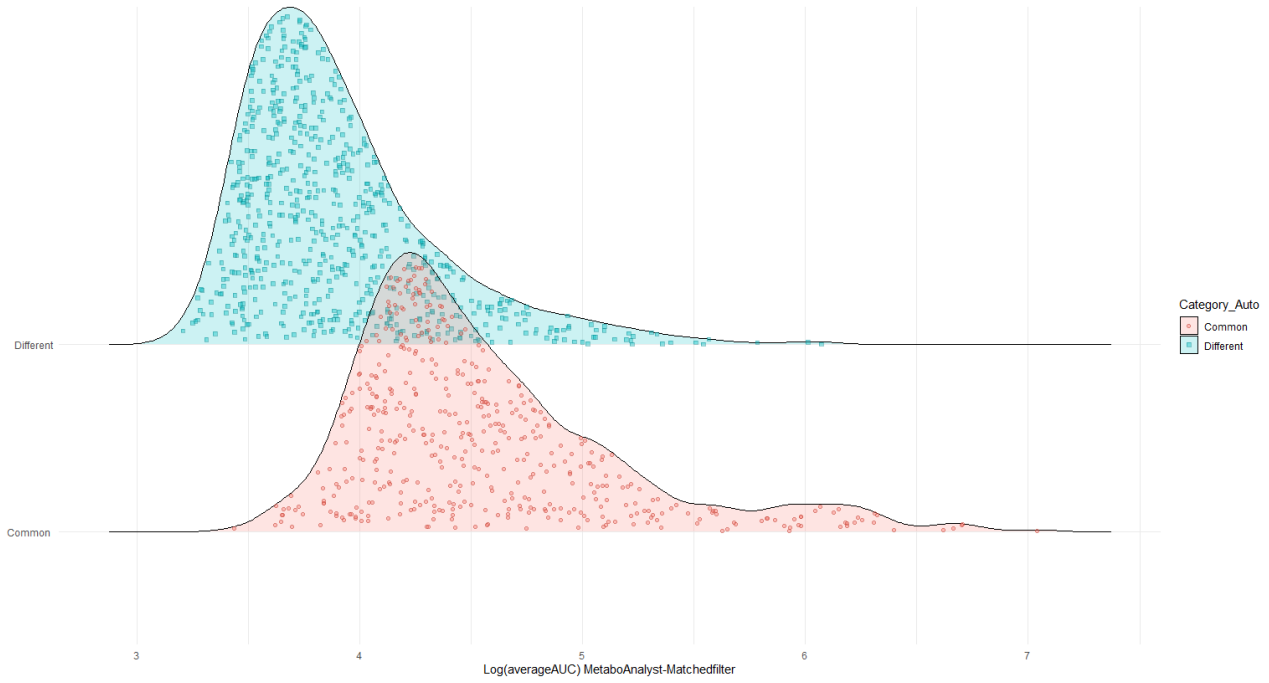
Από το διάγραμμα πυκνότητας του δεκαδικού λογαρίθμου του μέσου όρου της ανιχνευόμενης έντασης από το MetaboAnalyst για τα μοναδικά feature που ανιχνεύονται από το MetaboAnalyst, αλλά και για τα feature που ανιχνεύουν και τα δύο λογισμικά μπορούμε να παρατηρήσουμε ότι τα κοινά feature μεταξύ των δύο λογισμικών ανιχνεύονται σε μεγαλύτερες εντάσεις συγκριτικά με αυτά που εντοπίζονται μόνο από το MetaboAnalyst. Αυτό μπορεί να μας οδηγήσει στο συμπέρασμα ότι τα feature με τις μεγαλύτερες εντάσεις είναι πιο σταθερά και πιο εύκολο να ανιχνευθούν. Όμως, στη αντίστοιχη διαγραμματική απεικόνιση του λογισμικού Profinder τα feature φαίνεται να έχουν ακριβώς την αντίθετη τάση, δηλαδή τα διαφορετικά feature φαίνεται να τείνουν να ανιχνεύονται με μεγαλύτερες εντάσεις. Η παρατήρηση αυτή αντιτίθεται σε σχέση με την αρχική.



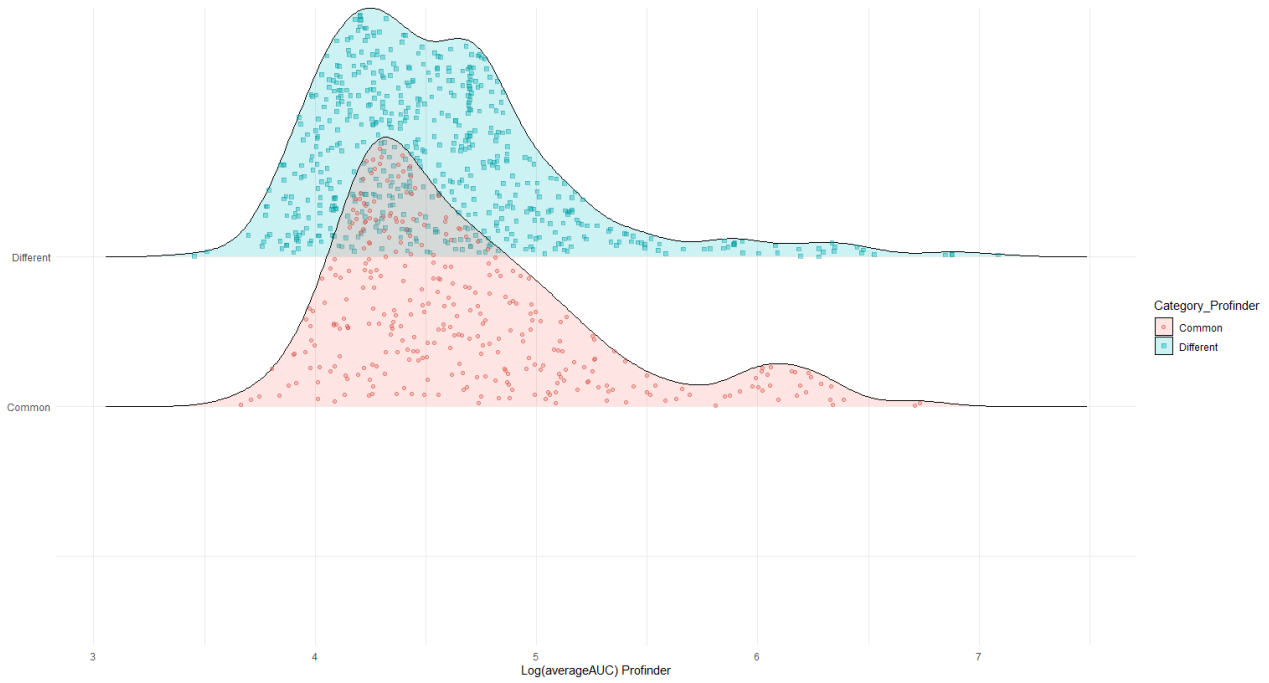
Εικόνα 25: Διάγραμμα πυκνότητας του δεκαδικού λογαρίθμου του μέσου όρου της ανιχνευόμενης έντασης για τα feature που εντοπίζονται ταυτόχρονα και στα λογισμικά MetaboAnalyst και Profinder.

Τα παραπάνω διαγράμματα πυκνότητας του δεκαδικού λογαρίθμου του μέσου όρου των εντάσεων στα κοινά feature μεταξύ των δύο λογισμικών έχουν παρόμοια κατανομή. Από το γεγονός αυτό μπορούμε να συμπεράνουμε πως τα feature που ανιχνεύονται και στα δύο λογισμικά ανιχνεύονται μάλλον σε παρόμοιες εντάσεις τόσο στο MetaboAnalyst, όσο και στο Profinder.

Έπειτα, κατασκευάστηκαν τα ίδια διαγράμματα για την σύγκριση του Profinder με τα αποτελέσματα της ανάλυσης με τον αλγόριθμο Matchedfilter στο λογισμικό MetaboAnalyst. Η επιλογή του συγκεκριμένου αλγορίθμου έγινε με κριτήριο το γεγονός πως φαίνεται στον αρχικό πίνακα να έχει την καλύτερη εφαρμογή στα δεδομένα της συγκεκριμένης εργασίας, μετά την ανάλυση με τις auto-optimized παραμέτρους. Στον αρχικό πίνακα, φαίνεται να έχει τις περισσότερες τιμές CV% στο επιθυμητό εύρος 0-20% αλλά και τα λιγότερα ελλείποντα δεδομένα.



Εικόνα 26: Διάγραμμα πυκνότητας του δεκαδικού λογαρίθμου του μέσου όρου της ανιχνευόμενης έντασης από το MetaboAnalyst για τα μοναδικά feature που ανιχνεύονται από το MetaboAnalyst, αλλά και για τα feature που ανιχνεύουν και τα δύο λογισμικά.

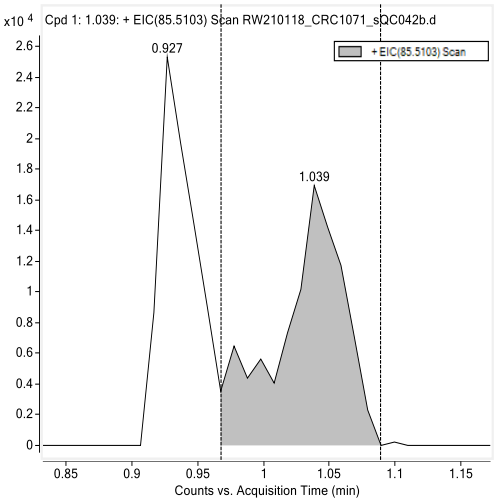


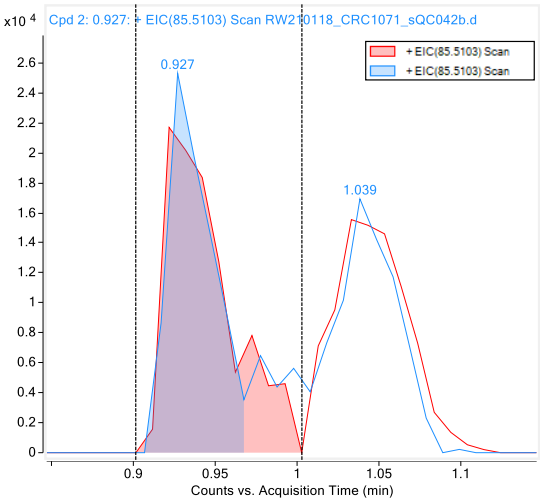
Εικόνα 27: Διάγραμμα πυκνότητας του δεκαδικού λογαρίθμου του μέσου όρου της ανιχνευόμενης έντασης από το Profinder για τα μοναδικά feature που ανιχνεύονται από το Profinder, αλλά και για τα feature που ανιχνεύουν και τα δύο λογισμικά.

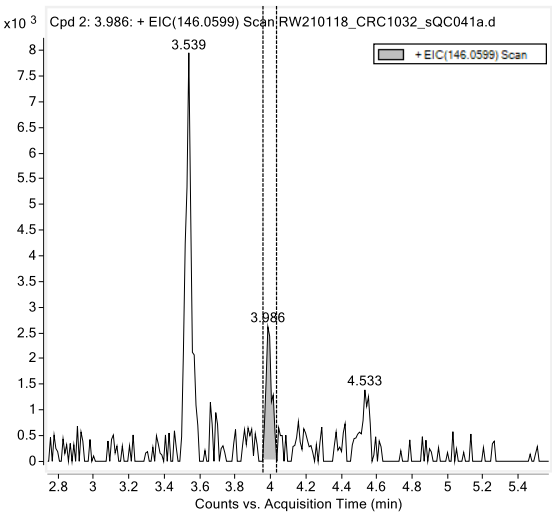
Εικόνα 28: Διάγραμμα πυκνότητας του δεκαδικού λογαρίθμου του μέσου όρου της ανιχνευόμενης έντασης για τα feature που εντοπίζονται ταυτόχρονα και στα λογισμικά MetaboAnalyst και Profinder.

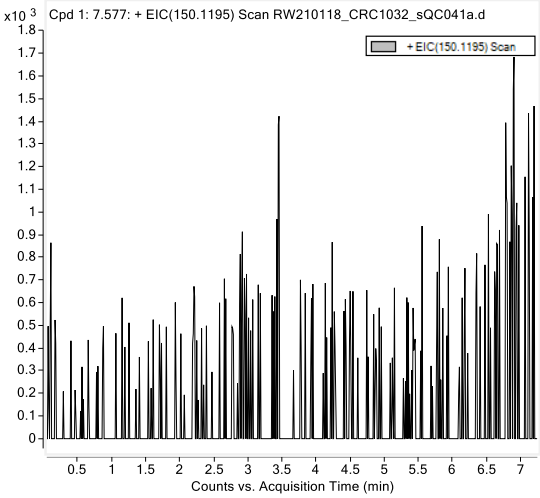
Οι παρατηρήσεις που προκύπτουν απ' αυτά τα διαγράμματα πυκνότητας της σύγκρισης Matchedfilter/ Profinder είναι πολύ κοντινές με αυτές που αναφέρθηκαν παραπάνω για την σύγκριση των δεδομένων από τις αναλύσεις Auto-optimized/ Profinder.

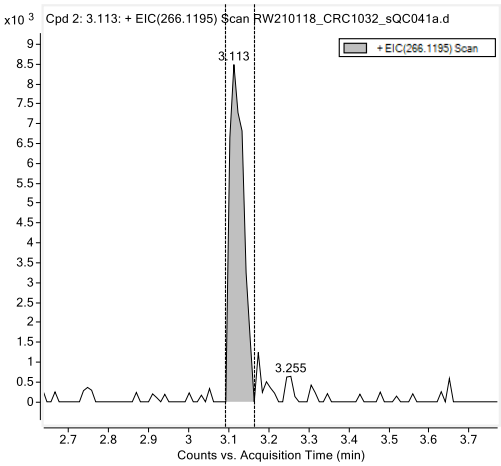
Σε αυτή τη φάση, κρίνεται σημαντική η κατανόηση του γιατί κάποια feature ανιχνεύονται μόνο από ένα λογισμικό. Η αρχική υπόθεση ήταν πως τα feature που δεν ανιχνεύονται και από τα δύο λογισμικά, πιθανόν να εντοπίζονται σε σχετικά μικρές εντάσεις. Από τα παραπάνω διαγράμματα πυκνότητας του δεκαδικού λογαρίθμου του μέσου όρου της ανιχνευόμενης έντασης φαίνεται πως οι υπόθεση αυτή επαληθεύεται για τα μοναδικά feature που ανιχνεύει το MetaboAnalyst και στις δύο αναλύσεις που εξετάσθηκαν, όχι όμως για τα μοναδικά feature που εντοπίζει το Profinder. Για τον λόγο αυτόν, αποφασίστηκε να εξετασθούν δειγματοληπτικά ορισμένα από τα μοναδικά feature κάθε τρόπου ανάλυσης όπως αυτά φαίνονται στην ακατέργαστη μορφή των δεδομένων. Η κίνηση αυτή αποσκοπεί στην ανίχνευση κάποιου κοινού μοτίβου στον τρόπου που κάθε λογισμικό επιλέγει ή απορρίπτει μία κορυφή.

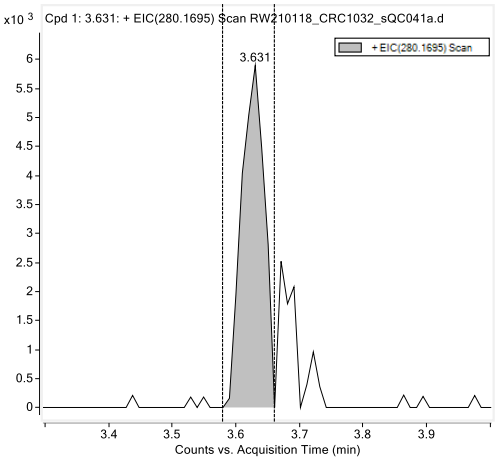

Cpd 1: 1.039: + EIC(85.5103) Scan RW210118_CRC1071_sQC042b.d
+ EIC(85.5103) Scan
0.927
1.039
Counts vs. Acquisition Time (min)

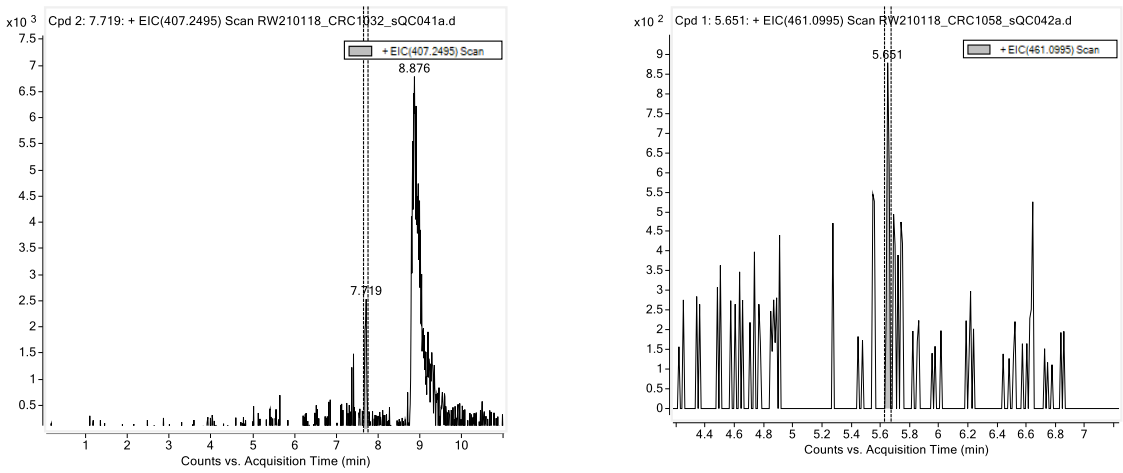

Cpd 2: 0.927: + EIC(85.5103) Scan RW210118_CRC1071_sQC042b.d
+ EIC(85.5103) Scan
+ EIC(85.5103) Scan
0.927
1.039
Counts vs. Acquisition Time (min)


Cpd 2: 3.986: + EIC(146.0599) Scan RW210118_CRC1032_sQC041a.d
+ EIC(146.0599) Scan
3.539
3.986
4.533
Counts vs. Acquisition Time (min)


Cpd 1: 7.577: + EIC(150.1195) Scan RW210118_CRC1032_sQC041a.d
+ EIC(150.1195) Scan
Counts vs. Acquisition Time (min)

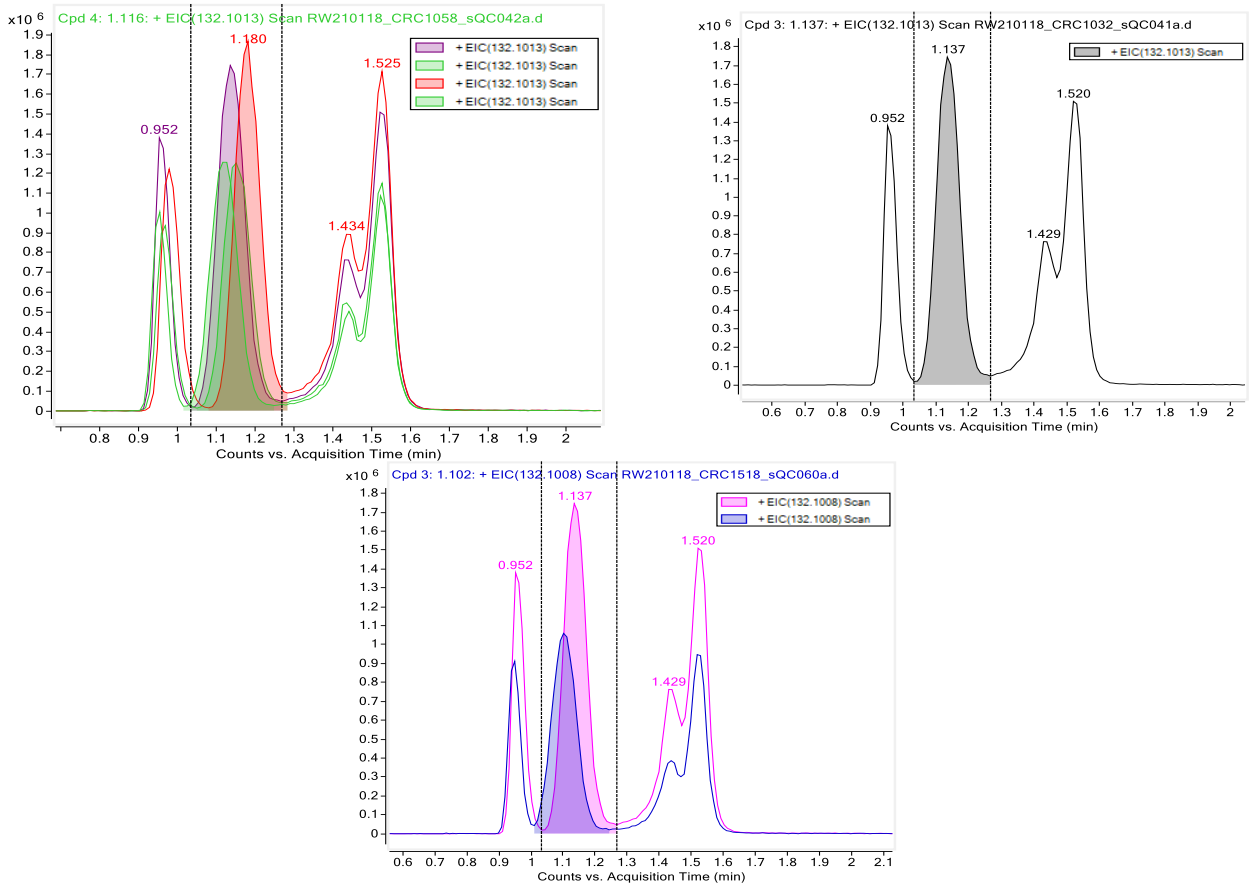

Cpd 2: 3.113: + EIC(266.1195) Scan RW210118_CRC1032_sQC041a.d
+ EIC(266.1195) Scan
3.113
3.255
Counts vs. Acquisition Time (min)

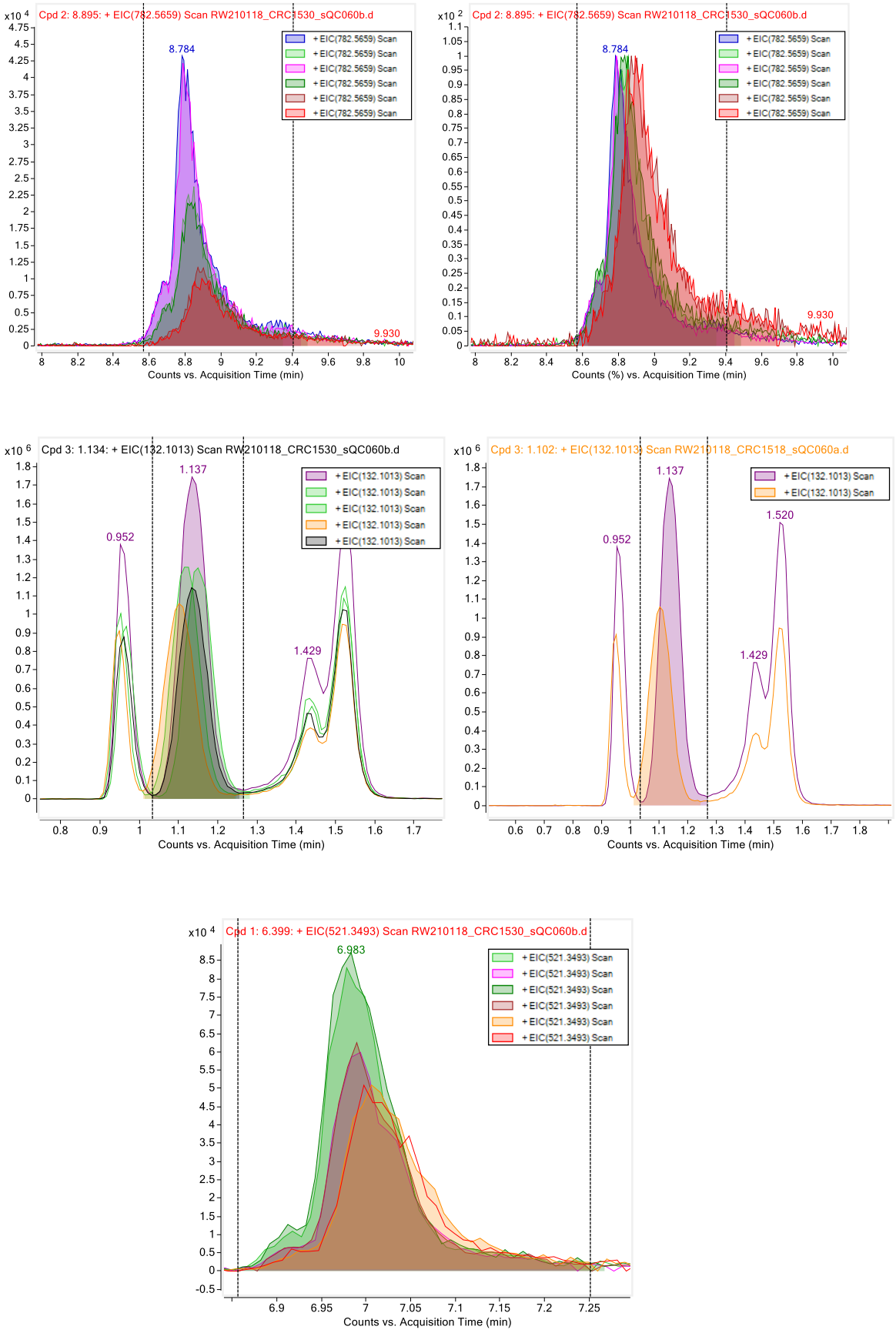

Cpd 1: 3.631: + EIC(280.1695) Scan RW210118_CRC1032_sQC041a.d
+ EIC(280.1695) Scan
3.631
Counts vs. Acquisition Time (min)



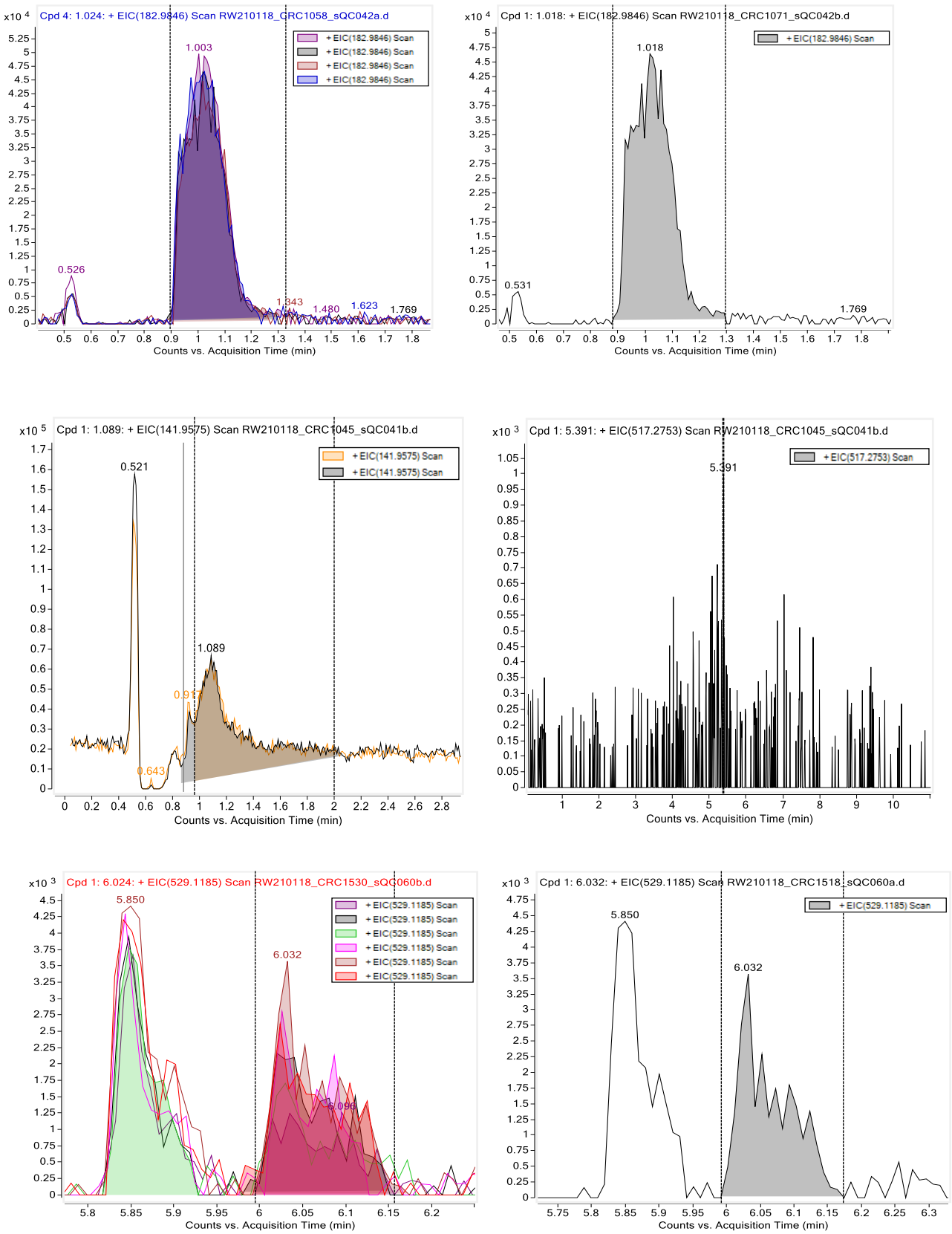
Εικόνα 29: Εικόνες από τα ακατέργαστα δεδομένα για τις κορυφές που ανιχνεύθηκαν μόνο από την ανάλυση με τις auto-optimized παραμέτρους του MetaboAnalyst και όχι από το Profinder.

Στις παραπάνω εικόνες φαίνονται κορυφές τις οποίες έχει εντοπίσει το MetaboAnalyst στην ανάλυση με τις παραμέτρους auto-optimized αλλά όχι το Profinder. Η επιλογή των κορυφών οι οποίες παρατηρήθηκαν στα ακατέργαστα δεδομένα έγινε τυχαία. Από τις παραπάνω εικόνες μπορούμε να δούμε ότι οι συγκεκριμένες κορυφές είναι μικρές και κακού σχήματος, αρκετές φορές μάλιστα μοιάζουν και με θόρυβο. Επομένως, συνδυαστικά με τα διαγράμματα πυκνότητας του δεκαδικού λογαρίθμου του μέσου όρου της έντασης μπορούμε να ισχυριστούμε πως οι κορυφές που δεν εντοπίζονται από το Profinder αντιστοιχούν σε μεταβολίτες μικρής συγκέντρωσης ή και σε θόρυβο.





Εικόνα 30: Εικόνες από τα ακατέργαστα δεδομένα για τις κορυφές που ανιχνεύθηκαν μόνο από την ανάλυση με το Profinder και όχι από αυτές με το MetaboAnalyst και παρουσιάζουν μετακίνηση ως προς τον χρόνο κατακράτησης.



Εικόνα 31: Εικόνες από τα ακατέργαστα δεδομένα για τις κορυφές που ανιχνεύθηκαν μόνο από την ανάλυση με το Profinder και όχι από αυτές με το MetaboAnalyst και παρουσιάζουν σπασίματα και δεν έχουν καλό σχήμα .

Στις παραπάνω εικόνες φαίνονται εικόνες από κάποιες κορυφές οι οποίες εντοπίζονται στο Profinder αλλά όχι στο MetaboAnalyst. Η επιλογή των κορυφών οι οποίες παρατηρήθηκαν στα ακατέργαστα δεδομένα έγινε τυχαία. Οι περισσότερες από τις παραπάνω κορυφές είναι σχετικά μεγάλης έντασης, άρα πιθανόν να αντιστοιχούν σε μεταβολίτες μεγάλης συγκέντρωσης στα βιολογικά δείγματα. Οι κορυφές αυτές παρόλο που είναι μεγάλες, παρατηρείται πως σε κάποιες περιπτώσεις εμφανίζουν σπασίματα και δεν έχουν καλό σχήμα. Επίσης, σε κάποιες περιπτώσεις φαίνεται να μην είναι πολύ σταθερές στον χρόνο, δηλαδή να εμφανίζεται αρκετά μεγάλη διακύμανση στον χρόνο κατακράτησης του συγκεκριμένου μεταβολίτη ανάμεσα στα δείγματα. Επομένως, μπορούμε να συμπεράνουμε ότι το MetaboAnalyst πιθανόν να δυσκολεύεται να αναγνωρίσει κορυφές με κακό σχήμα και κορυφές οι οποίες δεν είναι σταθερές στον χρόνο. Επιπλέον, παρατηρούμε πως το Profinder εντοπίζει feature τα οποία έχουν πολλά missing values, ακόμα και feature τα οποία εντοπίζονται σε ένα μόνο δείγμα. Αντίθετα, το MetaboAnalyst δεν φαίνεται να ανιχνεύει feature τα οποία εντοπίζονται σε τόσο λίγα δείγματα.

Συμπεράσματα:

Στη συγκεκριμένη εργασία πραγματοποιήθηκε προκατεργασία των ακατέργαστων δεδομένων που προέκυψαν από την ανάλυση 200 δειγμάτων ελέγχου ποιότητας (QC) και 50 (Blank). Τα δείγματα αυτά προέρχονται από μία μελέτη μεγάλης κλίμακας που διήρκησε 8 εβδομάδες.

Αρχικά, πραγματοποιήθηκε η επεξεργασία των συγκεκριμένων δεδομένων στο διαδικτυακό λογισμικό MetaboAnalyst με την χρήση όλων των διαθέσιμων αλγορίθμων (centWave, Massifquant, Matchedfilter), αλλά και της επιλογής των auto-optimized παραμέτρων. Από τις συγκεκριμένες αναλύσεις μπορεί να εξαχθεί το συμπέρασμα πως αλλάζοντας μόνο την επιλογή του αλγορίθμου και διατηρώντας όλες τις άλλες παραμέτρους ίδιες, τα αποτελέσματα που προκύπτουν διαφέρουν σημαντικά. Επιπλέον, αξίζει να σημειωθεί πως οι παράμετροι auto-optimized του MetaboAnalyst, ανταποκρίνονται αρκετά ικανοποιητικά στην ανάλυση των συγκεκριμένων δεδομένων.

Έπειτα, πραγματοποιήθηκε προσπάθεια αφαίρεσης των feature που πιθανόν αντιστοιχούν σε σήμα υποβάθρου είτε σε κορυφές που προέρχονται από κάποιον διαλύτη που χρησιμοποιήθηκε κατά την αναλυτική διαδικασία (background exclusion). Η συγκεκριμένη προσπάθεια φαίνεται να έχει σημαντική επίδραση. Αυτό μπορούμε να το διακρίνουμε αρχικά από το γεγονός πως ενώ ο αριθμός των συνολικών feature παρουσιάζει μεγάλη μείωση, χωρίς όμως να μειώνονται κατά πολύ τα feature τα οποία εμφανίζουν κάποια σταθερότητα (CV<20%), αλλά και από τις PCA (Εικόνα 10), στις οποίες φαίνεται πως μετά το συγκεκριμένο βήμα απομακρύνονται feature τα οποία ήταν πιο μακριά σε σχέση με αυτά της υπόλοιπης παρτίδας.

Ακολούθησε η διόρθωση του φαινομένου παρτίδας (Batch effect correction). Για τον σκοπό αυτόν, έγινε χρήση και των τριών επιλογών που προσφέρει το MetaboAnalyst για την αντικατάσταση των ελλειπόντων δεδομένων. Η συγκεκριμένη επιλογή φάνηκε ότι έχει πολύ μικρή επίδραση σε σύνολα δεδομένων με πολύ μικρό ποσοστό ελλειπόντων τιμών, όσο όμως τα ποσοστά αυτά αυξάνονται τόσο η συγκεκριμένη επιλογή θα πρέπει να είναι αποτέλεσμα περισσότερης σκέψης καθώς μπορούν να προκύψουν μεγάλες διαφορές στα αποτελέσματα. Μετά τηδιόρθωση του φαινομένου παρτίδας, παρατηρήθηκε πως τα feature πλησίασαν μεταξύ τους. Αυτό προκύπτει τόσο από τις τιμές του συντελεστή μεταβλητότητας που παρατηρήθηκε πως αυξήθηκαν κατά πολύ οι τιμές που βρίσκονται κάτω από 20%, όσο και από τα διαγράμματα PCA. Επομένως, μπορούμε να συμπεράνουμε πως η επιλογή για Batch effect correction, που προσφέρει το συγκεκριμένο λογισμικό συμβάλλει στην βελτίωση της ποιότητας των παρεχόμενων αποτελεσμάτων.

Τέλος, ακολούθησε η ανάλυση των ίδιων δεδομένων με το λογισμικό της εταιρείας AGILENT, αλλά και η σύγκριση των αποτελεσμάτων των δύο λογισμικών. Συγκρίθηκαν τα αναλυτικά αποτελέσματα και των τριών αλγορίθμων του MetaboAnalyst με αυτά του λογισμικού της AGILENT. Από την συγκεκριμένη σύγκριση προέκυψε πως πολύ μικρό ποσοστό των ανιχνευόμενων feature επικαλύπτονται μεταξύ των δύο λογισμικών. Τυχαία επιλέχθηκαν feature τα οποία ανιχνεύονται μόνο από το λογισμικό MetaboAnalyst και παρατηρήθηκε η εικόνα των συγκεκριμένων κορυφών στα ακατέργαστα δεδομένα. Οι κορυφές αυτές ήταν συνήθως μικρής έντασης και κακού σχήματος που τις περισσότερες φορές έμοιαζαν με θόρυβο. Στη συνέχεια ακολούθησε η ίδια διαδικασία για feature τα οποία ανιχνεύθηκαν μόνο από το λογισμικό της AGILENT. Οι περισσότερες από τις συγκεκριμένες κορυφές ήταν μεγάλης έντασης και καλού σχήματος, συχνά όμως παρουσίαζαν διακύμανση στον χρόνο κατακράτησης ανάμεσα στα δείγματα. Συνεπώς, μπορούμε να συμπεράνουμε πως το MetaboAnalyst μπορεί να δυσκολεύεται να αναγνωρίσει κορυφές που δεν είναι σταθερές στον χρόνο κατακράτησης.

Οι παρατηρήσεις της παρούσας εργασίας αναφορικά με τη σύγκριση των συγκεκριμένων λογισμικών συνοψίζονται στον Πίνακα 12, στον οποίο αναγράφονται τα πλεονεκτήματα και τα μειονεκτήματα του καθενός.

Πίνακας 12: Σύγκριση πλεονεκτημάτων και μειονεκτημάτων ανάμεσα στο MetaboAnalyst και στο λογισμικό της AGILENT.

AGILENT suite		**MetaboAnalyst**	
Πλεονεκτήματα	**Μειονεκτήματα**	**Πλεονεκτήματα**	**Μειονεκτήματα**
Εύκολο στη χρήση.	Υψηλό κόστος.	Εύκολο στη χρήση.	Περιορισμός στον αριθμό των προς ανάλυση δειγμάτων (μέγιστο 200 δείγματα).
Μεγαλύτερος αριθμός προς ανάλυση δειγμάτων (~500).	Γίνονται αποδεκτά μόνο αρχεία της μορφής ".d", προερχόμενα από όργανα του ίδιου κατασκευαστή (Agilent).	Προσβάσιμο σε όλους.	Συχνά προκύπτουν σφάλματα κατά την διαδικασία της ανάλυσης με αποτέλεσμα να πρέπει να ξεκινήσει από την αρχή η αναλυτική διαδικασία.
Ικανότητα ανίχνευσης κορυφών που εντοπίζονται σε πολύ μικρό αριθμό δειγμάτων με χρήση κατάλληλων παραμέτρων.	Μεγάλος χρόνος ανάλυσης (~ 4 μέρες για 250 δείγματα).	Αποδεκτές πολλές μορφές αρχείων (αρχεία ανοιχτού κώδικα).	Φαίνεται να αποτυγχάνει να ανιχνεύσει κορυφές μεγάλης έντασης.
Τα feature που δεν ανιχνεύει φαίνεται να είναι επί το πλείστον μικρής έντασης και κακού σχήματος κορυφές.	Απαραίτητη χρήση Η/Υ υψηλών δυνατοτήτων προκειμένου να ανταποκριθεί στις απαιτήσεις του λογισμικού.	Πολλές επιλογές και δυνατότητες για περαιτέρω επεξεργασία των δεδομένων στην ίδια πλατφόρμα (π.χ. στατιστική επεξεργασία)	Ο χρήστης δεν έχει σαφή εικόνα για τον τρόπο ανάλυσης στην επιλογή auto-optimized παραμέτρων.
Δυνατότητα οπτικής απεικόνισης όλων τον ανιχνευόμενων κορυφών.		Η ανάλυση γίνεται σε online server και δεν εξαρτάται από τις δυνατότητες του υπολογιστή.	Το λογισμικό δεν προσφέρει δυνατότητα οπτικής απεικόνισης των ανιχνευόμενων κορυφών.

Συμπερασματικά, το στάδιο της προκατεργασίας των ακατέργαστων δεδομένων που προκύπτουν από μελέτες μεταβολομικής, είναι ένα από τα πιο σημαντικά βήματα της αναλυτικής διαδικασίας. Στην συγκεκριμένη εργασία φαίνεται πως πολύ μικρές αλλαγές στις παραμέτρους που θα επιλεγούν, μπορεί να οδηγήσουν σε αξιοσημείωτες διαφορές στα προκύπτοντα αναλυτικά αποτελέσματα, οι οποίες θα παίξουν κεντρικό ρόλο και στην μετέπειτα στατιστική ανάλυση. Επιπροσθέτως, έπειτα από την ανάλυση των δεδομένων και με τα δύο λογισμικά και την μεταξύ τους σύγκριση μπορούμε να συμπεράνουμε αρχικά πως τα αναλυτικά αποτελέσματα τα οποία προέκυψαν απ΄αυτά δεν εμφανίζουν μεγάλο ποσοστό επικάλυψης. Τα

δύο αυτά λογισμικά εμφανίζουν διαφορετικά πλεονεκτήματα και μειονεκτήματα, όπως αποτυπώθηκαν στην παρούσα εργασία. Η ανάπτυξη και η βελτιστοποίηση των συγκεκριμένων λογισμικών επεξεργασίας ακατέργαστων δεδομένων καθώς και η κατάλληλη εκπαίδευση των χρηστών στην ορθή χρήση των εργαλείων αυτών θα μπορούσε να φανεί πολύ σημαντική στην περαιτέρω ανάπτυξη του πεδίου της μεταβολομικής. Ιδιαίτερα, η βελτιστοποίηση των δυνατοτήτων των λογισμικών ελεύθερης πρόσβασης θα μπορούσε να φανεί ιδιαίτερα βοηθητική παρέχοντας δυνατότητες επεξεργασίας των μεταβολομικών δεδομένων που υπάρχουν ηλεκτρονικά καταθετήρια ή που δημιουργούνται σε εργαστήρια που αξιοποιούν συστήματα φασματομετρίας μάζας διαφορετικών κατασκευαστών. Τέλος, στα λογισμικά αυτά, λόγω της φύσης τους ως εφαρμογές ανοιχτού κώδικα, υπάρχει η δυνατότητα άμεσης ενσωμάτωσης νέων πακέτων και δυνατοτήτων που αναπτύσσονται από την ερευνητική κοινότητα, συμβάλλοντας στην ταχύτερη ανάπτυξη του πεδίου.

Βιβλιογραφία:

1.| Quantitative metabolic profiling based on gas chromatography. Clinical Chemistry, 24(10), 1663–1673 | 10.1093/clinchem/24.10.1663. https://sci-hub.ru/10.1093/clinchem/24.10.1663.

2. The metamorphosis of analytical chemistry | Analytical and Bioanalytical Chemistry. https://link.springer.com/article/10.1007/s00216-019-02313-z.

3. Gorrochategui, E., Jaumot, J., Lacorte, S. & Tauler, R. Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow. *TrAC Trends Anal. Chem.* **82**, 425–442 (2016).

4. Metabolomics – the link between genotypes and phenotypes | Plant Molecular Biology. https://link.springer.com/article/10.1023/A:1013713905833.

5. Χατζηιωάννου, Α.-Χ. & Chatziioannou, A.-C. Ανάπτυξη ολιστικών αναλυτικών τεχνολογιών για την εύρεση βιοδεικτών σε βιολογικά δείγματα νεογνών. (Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης (ΑΠΘ), Σχολή Θετικών Επιστημών, Τμήμα Χημείας, Τομέας Φυσικής, Αναλυτικής και Περιβαλλοντικής Χημείας, Εργαστήριο Αναλυτικής Χημείας, 2019). doi:10.12681/eadd/45293.

6. Σαμαλίδου, Α. Ενσωμάτωση γονιδιακών Δεδομένων στον ηλεκτρονικό φάκελο υγείας και ιατρικής ακρίβειας. (Πανεπιστήμιο Μακεδονίας, Δεκέμβριος 2020)

7. Targeted vs Untargeted Metabolomics | Metabolon. https://www.metabolon.com/blog/targeted-vs-untargeted-metabolomics/.

8. Metabolomics Data Analysis | Thermo Fisher Scientific - GR. https://www.thermofisher.com/gr/en/home/industrial/mass-spectrometry/mass-spectrometry-learning-center/mass-spectrometry-applications-area/metabolomics-mass-spectrometry/metabolomics-data-analysis.html.

9. Scalable Analysis of Untargeted LC-HRMS Data by Means of SQL Database Archiving | Analytical Chemistry. https://pubs.acs.org/doi/10.1021/acs.analchem.2c03769.

10. MetaboAnalyst 2.0—a comprehensive server for metabolomic data analysis | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/40/W1/W127/1072165.

11. Guide to Metabolomics Analysis: A Bioinformatics Workflow - PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9032224/.

12. MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights - PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8265181/.

13. Using MetaboAnalyst 5.0 for LC–HRMS spectra processing, multi-omics integration and covariate adjustment of global metabolomics data | Nature Protocols. https://www.nature.com/articles/s41596-022-00710-w.

14. Introductions. https://www.metaboanalyst.ca/resources/vignettes/Introductions.html.

15. MassHunter Profinder Software. https://www.agilent.com/Library/usermanuals/Public/G3835-90014_Profinder_QuickStart_B.pdf

16. Xia, J., Psychogios, N., Young, N. & Wishart, D. S. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res.* **37**, W652–W660 (2009).

17. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets - PubMed. https://pubmed.ncbi.nlm.nih.gov/22111785/.

18. HPB-advancedXCMS.pdf.

19. Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies | Metabolomics. https://link.springer.com/article/10.1007/s11306-018-1367-3.

20. Broeckling, C. D. *et al.* Current Practices in LC-MS Untargeted Metabolomics: A Scoping Review on the Use of Pooled Quality Control Samples. *Anal. Chem.* **95**, 18645–18654 (2023).

21. Quality assurance and quality control reporting in untargeted metabolic phenotyping: mQACC recommendations for analytical quality management - PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9420093/.

22. Δέδα, Ό. Μέθοδοι Μεταβολομικής ανάλυσης για την εύρεση βιοδεικτών άσκησης και γήρανσης σε θηλαστικά. (Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης (ΑΠΘ), Σχολή Θετικών Επιστημών, Τμήμα Χημείας, Τομέας Φυσικής, Αναλυτικής και Περιβαλλοντικής Χημείας, Εργαστήριο Αναλυτικής Χημείας, 2017).

23. Representing the Metabolome with High Fidelity: Range and Response as Quality Control Factors in LC-MS-Based Global Profiling | Analytical Chemistry. https://pubs.acs.org/doi/10.1021/acs.analchem.0c03848.

24. The Vital Role of Blanks in Sample Preparation. https://www.chromatographyonline.com/view/vital-role-blanks-sample-preparation.

25. MetaboAnalyst Tutorial https://www.metaboanalyst.ca/resources/vignettes/Introductions.html.

26. LC-MS/MS Raw Spectral Data Processing. https://www.metaboanalyst.ca/resources/vignettes/LCMSMS_Raw_Spectral_Processing.html.