ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΙΑΤΡΙΚΗ ΣΧΟΛΗ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΣΤΗ

“ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ”

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Επικύρωση στρατηγικών ενσωμάτωσης για βιολογικά δεδομένα υψηλής απόδοσης

Ελευθέριος Παύλος

Ηράκλειο, Σεπτέμβριος 2021

Πρόλογος

Τα τελευταία χρόνια έχει παρατηρηθεί μία ραγδαία αύξηση των δεδομένων που παράγονται σε διάφορους επιστημονικούς τομείς. Ανάμεσα σε αυτούς συγκαταλέγονται και αυτοί της ιατρικής και της βιολογίας. Στον τομέα της Ιατρικής επιστήμης υπάρχει μία σύγχρονη τάση για εξατομικευμένη θεραπεία ασθενειών η οποία είναι πιο αποτελεσματική και εύστοχη σε σχέση με την υπάρχουσα προσέγγιση. Για να επιτευχθεί εξατομικευμένη θεραπεία απαιτείται η συλλογή και ανάλυση μεγάλου όγκου δεδομένων από ασθενείς. Ο τομέας της Βιοπληροφορικής είναι το επιστημονικό πεδίο που ασχολείται με την αναζήτηση νέων μεθόδων συλλογής, διαχείρισης και ανάλυσης δεδομένων. Η ανάπτυξη νέων τεχνολογιών όπως π.χ. οι τεχνολογίες αλληλούχισης επόμενης γενιάς έχει μειώσει αισθητά το κόστος και έχει διευκολύνει την παραγωγή δεδομένων όπως η αλληλούχιση του DNA, η μέτρηση της έκφρασης των γονιδίων κ.α.

Στον τομέα της Βιοπληροφορικής έχουν αναπτυχθεί πολλά εργαλεία, ευρέως χρησιμοποιούμενα, για την ανάλυση των δεδομένων ιατροβιολογικού ενδιαφέροντος. Ωστόσο τα περισσότερα εργαλεία επεξεργάζονται και αναλύουν δεδομένα ενός τύπου, όπως π.χ. τα εργαλεία που είναι προσαρμοσμένα ώστε να εντοπίζουν διαφορές ανάμεσα σε ομάδες ασθενών όσον αφορά την έκφραση γονιδίων. Σχετικά πρόσφατα έχει αρχίσει να αναπτύσσεται το πεδίο της συνδυαστικής ανάλυσης δεδομένων διαφορετικού τύπου (multi-omics analysis) η οποία αφορά δεδομένα που προέρχονται από διαφορετικές μελέτες (μεθυλίωση DNA, έκφραση γονιδίων και πρωτεϊνών, σωματικές μεταλλάξεις κ.α.). Η προσέγγιση αυτή έχει ως στόχο να καταφέρει να αποτυπώσει μία πιο ολοκληρωμένη εικόνα του υπό εξέταση βιολογικού συστήματος ή μίας ασθένειας αντλώντας επιπλέον πληροφορία από τη συνδυαστική ανάλυση των επιμέρους δεδομένων.

Στην παρούσα εργασία γίνεται παρουσίαση των πιο διαδεδομένων εργαλείων που χρησιμοποιούνται σήμερα για την ανάλυση multi-omics δεδομένων στη γλώσσα προγραμματισμού R, επαλήθευση της λειτουργίας αυτών και παράθεση των δυνατοτήτων οπτικοποίησης τους με σκοπό τη μεταξύ τους σύγκριση ως προς την ερμηνεία των αποτελεσμάτων.

Η επιλογή του συγκεκριμένου θέματος έγινε με στόχο να παρέχει μια ολοκληρωμένη εικόνα της κατάστασης που υπάρχει σε σχέση με τα διαθέσιμα εργαλεία για την ανάλυση multi-omics δεδομένων.

Επιβλέπων καθηγητής της εργασίας ήταν ο Αναπληρωτής Καθηγητής της Ιατρικής σχολής του Πανεπιστημίου Κρήτης κ. Ιωάννης Ηλιόπουλος και μέλη της εξεταστικής επιτροπής ο Ερευνητής Α' - Διευθυντής Ερευνών του Ιδρύματος Ιατροβιολογικών Ερευνών της Ακαδημίας Αθηνών κ. Ευάγγελος Ανδρεάκος και ο Ερευνητής Β' του Ερευνητικού Κέντρου Βιοϊατρικών Επιστημών «Αλέξανδρος Φλέμινγκ» κ. Γεώργιος Παυλόπουλος. Οφείλω να ευχαριστήσω θερμά τα μέλη της επιτροπής για τις χρήσιμες υποδείξεις και τη συμβολή τους στην ολοκλήρωση της διπλωματικής μου εργασίας.

Τέλος, θέλω να ευχαριστήσω την οικογένεια και τους φίλους μου για την συμπαράσταση, την κατανόηση και την έμπρακτη βοήθεια τους.

Abstract

In recent years there has been a rapid increase in data generated in various scientific fields. Among them are those of medicine and biology. In the field of medical science there is a modern trend for personalized treatment of diseases which is more effective and targeted than the existing approaches. Achieving personalized treatment requires the collection and analysis of a large volume of patient's data. The field of Bioinformatics is the scientific field that deals with the development of new methods for data collection, management and analysis. The development of new technologies such as Next Generation Sequencing (NGS) technologies have significantly reduced costs and facilitated the production of data such as DNA sequencing, measuring gene expression, etc.

In the field of Bioinformatics, many tools have been developed, widely used, for the analysis of data of biomedical interest. However, most of these tools are designed to approach data of a single type, such as tools that are tailored to detect differences between patient groups in terms of gene expression. More recently, the field of multi-omics analysis has begun to develop, which concerns data from different studies (DNA methylation, expression of genes and proteins, body mutations, etc.). This approach aims to capture a more complete picture of the biological system under consideration or of a disease by drawing additional information from the combined analysis of the individual datasets.

In the present work, the most common tools used today for the analysis of multi-omics data in the R programming language are presented, their function is verified and their visualization possibilities are listed in order to compare them in terms of interpreting the results.

Περίληψη

Η παρούσα διπλωματική εργασία αποτελεί μία παρουσίαση των πιο διαδεδομένων εργαλείων που χρησιμοποιούνται για την ανάλυση διαφορετικών τύπων βιολογικών δεδομένων (multi-omics data) και γίνεται επαλήθευση της λειτουργίας τους. Σκοπός της εργασίας είναι η διερεύνηση των δυνατοτήτων που παρέχουν στον χρήστη τα εργαλεία αυτά. Η εργασία αποτελείται από τέσσερα κεφάλαια.

Στο 1ο κεφάλαιο της εργασίας γίνεται αναφορά στα δεδομένα που υπάρχουν διαθέσιμα στο πεδίο της Βιολογίας και ειδικότερα σε αυτά που παράγονται από νέες τεχνολογίες και είναι γνωστά ως high-throughput data. Επιπλέον, αναφέρονται κάποια εισαγωγικά στοιχεία όσον αφορά τις γενικές κατηγορίες στις οποίες εμπίπτουν τα εργαλεία που χρησιμοποιούνται για την ενσωμάτωση και ανάλυση διαφορετικού τύπου δεδομένων (multi-omics).

Στο 2ο κεφάλαιο παρατίθενται οι βασικότερες μαθηματικές και στατιστικές τεχνικές στις οποίες βασίζονται τα περισσότερα μοντέλα ενσωμάτωσης διαφορετικού τύπου omics δεδομένων. Αυτές αφορούν μοντέλα παλινδρόμησης, ανάλυση συσχέτισης και κύριων συνιστωσών καθώς και την παραγοντοποίηση πινάκων. Ακόμη περιγράφονται δύο ευρέως χρησιμοποιούμενες μέθοδοι ομαδοποίησης (clustering), της ιεραρχικής ομαδοποίησης και k-means.

Στο 3ο κεφάλαιο παρουσιάζονται αναλυτικά οι μέθοδοι ενσωμάτωσης καθώς και τα αποτελέσματα που παρήχθησαν στο περιβάλλον της γλώσσας προγραμματισμού R. Τα εργαλεία που ελέγχθηκαν είναι αυτά που ανήκουν στα πακέτα mixOmics, MOFA, iCluster και MOVICS.

Τέλος, στο 4ο κεφάλαιο παρατίθενται τα συμπεράσματα που εξήχθησαν από την εφαρμογή των παραπάνω εργαλείων, η σύγκριση τους όσον αφορά τον υπολογιστικό χρόνο και τις συναρτήσεις απεικόνισης που παρέχουν καθώς επίσης και προτάσεις για την περαιτέρω ανάπτυξη παρόμοιων εργαλείων.

Πίνακας Περιεχομένων

Ευρετήριο σχημάτων

Ευρετήριο πινάκων

1. Εισαγωγή

Στο κεφάλαιο αυτό γίνεται μία αναφορά στις κατηγορίες των δεδομένων που θα αναλυθούν στο κύριο μέρος της εργασίας. Επιπλέον, υπάρχει μία εισαγωγική περιγραφή των εργαλείων που χρησιμοποιούνται σε multi-omics αναλύσεις και του τρόπου με τον οποίο αυτά λειτουργούν.

1.1.Βιολογικά δεδομένα υψηλής απόδοσης (high-throughput data)

Η ταχεία τεχνολογική πρόοδος κατέστησε διαθέσιμες τεχνολογίες υψηλής απόδοσης για τη μελέτη των βιολογικών συστημάτων, θέτοντας με αυτόν τον τρόπο τα θεμέλια για την ανάπτυξη της επονομαζόμενης εποχής των -omics και multi-omics αναλύσεων (Sandhu et al 2018). Πράγματι, η ολοκλήρωση της αλληλουχίας του ανθρώπινου γονιδιώματος (International Human Genome Sequencing Consortium 2004) και η διαθεσιμότητα τεχνολογικών εργαλείων μεγάλης κλίμακας επέτρεψαν τη μελέτη της γονιδιωματικής, της μεταγραφικής, της επιγενωμικής και άλλων -omics πεδίων σε προηγουμένως αδιανόητο επίπεδο (Sandhu et al 2018). Η ενσωμάτωση αυτών των πεδίων αυξάνει την κατανόησή μας για τις μοριακές βάσεις των ανθρώπινων ασθενειών (τόσο επίκτητων όσο και κληρονομικών), με τελικό στόχο τη βελτίωση της διάγνωσης, της παρακολούθησης και της θεραπείας τους, ενόψει ενός ακόμη πιο εξατομικευμένου τρόπου παροχής ιατρικής φροντίδας (Sandhu et al 2018). Οι τρέχουσες διαθέσιμες τεχνολογίες μπορούν να παράγουν gigabytes δεδομένων ανά ημέρα με μεγάλο επίπεδο ακρίβειας και αξιοπιστίας (Precone et al 2015). Αυτό το χαρακτηριστικό έχει ωθήσει τη μοριακή έρευνα πέρα από τους περιορισμούς που επιβάλλουν οι πιο παραδοσιακές αναλυτικές προσεγγίσεις. Ωστόσο, σύντομα έγινε σαφές ότι το ίδιο αυτό χαρακτηριστικό υποκρύπτει μια σημαντική παρενέργεια: οι τεχνολογίες υψηλής απόδοσης μπορούν να παράγουν μεγάλο όγκο δεδομένων, των οποίων η διαχείριση, ανάλυση και αποθήκευση απαιτούν συγκεκριμένες υποδομές και ανάπτυξη βιοπληροφορικής γνώσης (Kulkarni, Frommolt 2017). Ειδικότερα, η σωστή ερμηνεία των δεδομένων μέσα από αυτόν τον τεράστιο όγκο πληροφοριών και η αναζήτηση των δεδομένων που είναι σημαντικά από κλινική άποψη, αποτελούν σήμερα τη μεγαλύτερη πρόκληση. Επίσης, ηθικά ζητήματα που σχετίζονται με τυχαία ευρήματα, κατοχή και διακίνηση δεδομένων καθώς και η διασφάλιση του προσωπικού απορρήτου αποτελούν

μείζον θέμα επιστημονικής συζήτησης και πρέπει να ρυθμιστούν προσεκτικά για να αποφευχθούν οι κίνδυνοι που σχετίζονται με τη διαχείριση και ανάλυση των παραγόμενων δεδομένων.

Τα τελευταία 15 χρόνια οι τεχνολογίες αλληλούχισης επόμενης γενιάς (NGS) παρουσίασαν μεγάλη ανάπτυξη και γρήγορη διάδοση στην επιστημονική κοινότητα. Αυτές οι τεχνικές έχουν επηρεάσει κάθε τομέα μοριακής έρευνας, αναβαθμίζοντας τις τεχνολογίες που χρησιμοποιούνταν προηγουμένως και ανοίγοντας το δρόμο για τη θεμελίωση των αναλύσεων -omics δεδομένων. Πράγματι, οι μέθοδοι NGS επιτρέπουν τον προσδιορισμό της αλληλουχίας ολόκληρων γονιδιωμάτων (D'Argenio 2018), εξωμάτων (Weisz Hubshman et al 2018), ομάδων γονιδίων που σχετίζονται με κάποια συγκεκριμένη ασθένεια (Kalsner et al 2018), ή ενός γονιδίου (D'Argenio et al 2015), αλλά μπορούν επίσης να χρησιμοποιηθούν για να ερευνηθεί ολόκληρο το μεταγράφωμα, μικρά κομμάτια RNA, το επιγένωμα, και το μικροβίωμα.

Ανεξάρτητα από ορισμένα ιδιόμορφα χαρακτηριστικά που σχετίζονται με τους διαφορετικούς κατασκευαστές, οι διαθέσιμες τεχνολογίες NGS βασίζονται στην ενίσχυση μιας συγκεκριμένης βιβλιοθήκης (ή πολλαπλών βιβλιοθηκών με γραμμωτό κώδικα), δηλαδή, μιας δεξαμενής θραυσμάτων DNA που αντιπροσωπεύουν τον στόχο του οποίου πρόκειται να προσδιοριστεί η αλληλουχία, σε μία επιφάνεια ροής, ή σε μικροσκοπικά σφαιρίδια, για να ληφθούν συστάδες θραυσμάτων που στη συνέχεια θα αναλυθούν μαζικά. Οι τεχνικές αλληλούχισης επόμενης γενιάς συνδυάζουν την ικανότητα υψηλής απόδοσης και την ακρίβεια ανάγνωσης της αλληλουχίας με χαμηλό κόστος ανά βάση. Το κόστος για την αλληλουχία ολόκληρου του ανθρώπινου γονιδιώματος έχει μειωθεί από περίπου 10 εκατομμύρια δολάρια σε περίπου 1000 δολάρια μόνο τα τελευταία 10 χρόνια (Hayden 2014).

Από αυτήν την άποψη δεν προκαλεί έκπληξη το γεγονός ότι η τεχνολογία NGS τείνει να γίνει η μέθοδος αναφοράς για μοριακές αναλύσεις. Πιο συγκεκριμένα, η τεχνολογία NGS επιτρέπει την ανάλυση δειγμάτων από περισσότερους ασθενείς ταυτόχρονα, γονιδίων που σχετίζονται με συγκεκριμένες ασθένειες σε λιγότερο χρόνο και με χαμηλότερο κόστος από τις παραδοσιακές προσεγγίσεις, αλλά και τον προσδιορισμό της αλληλουχίας ομάδων γονιδίων έως και ολόκληρου του γονιδιώματος. Με αυτόν τον τρόπο, είναι δυνατόν να αυξηθεί η ευαισθησία

(sensitivity) στη διαδικασία της διάγνωσης, να ανακαλυφθούν νέα γονίδια που σχετίζονται με κάποια ασθένεια και επίσης να εξαχθούν δεδομένα σχετικά με γονίδια που ενδέχεται να επηρεάζουν τον φαινότυπο μιας νόσου. Λόγω της υψηλής ευαισθησίας και της ευελιξίας τους, οι τεχνολογίες NGS είναι επίσης χρήσιμες σε προγεννητικούς και προεμφυτευτικούς διαγνωστικούς ελέγχους (Huang et al 2017) καθώς και σε άλλες εφαρμογές, όπως η αλληλούχιση μορίων κυκλοφορούντος ελεύθερου DNA ή μεμονωμένων κυττάρων (Müller et al 2107).

Εκτός της μελέτης των παραλλαγών της αλληλουχίας σε επίπεδο DNA, οι τεχνολογίες NGS μπορούν να χρησιμοποιηθούν και για τη μελέτη της γενετικής μεταβλητότητας και των μηχανισμών που αποτελούν τη βάση της εμφάνισης συγκεκριμένων ασθενειών σε επιγενετικό, μεταγραφικό και μεταγονιδιωματικό επίπεδο. Πράγματι, αρκετοί παράγοντες, εκτός της γενετικής προδιάθεσης, όπως η διατροφή, οι περιβαλλοντικοί παράγοντες και ο τρόπος ζωής, μπορούν να επηρεάσουν το επιγένωμα, το μεταγράφωμα και το μικροβίωμα (Sandhu et al 2018, D'Argenio & Salvatore 2015). Έτσι, όλα αυτά τα συστήματα είναι δυναμικά και μπορεί να έχουν υποστεί συγκεκριμένες τροποποιήσεις που σχετίζονται με μια συγκεκριμένη παθολογική κατάσταση. Η κατανόηση τέτοιων τροποποιήσεων όχι μόνο διαφωτίζει τους μηχανισμούς που βρίσκονται πίσω από την ανάπτυξη της νόσου, αλλά μπορεί επίσης να παρέχει νέους, πιθανούς βιοδείκτες για μια πρώιμη ή/και ακριβέστερη διάγνωση, για τη διαστρωμάτωση των ασθενών σε κατηγορίες, για την παρακολούθηση ασθενειών ή/και για την ανάπτυξη στοχευμένων και κατά συνέπεια πιο αποτελεσματικών θεραπειών. Οι προσεγγίσεις που βασίζονται σε τεχνολογίες αλληλούχισης επόμενης γενιάς, παρέχουν μεγάλη κάλυψη της υπό εξέτασης αλληλουχίας όσο και αμερόληπτη “ανάγνωση” περίπλοκων συστημάτων χωρίς την ανάγκη ύπαρξης πρότερης γνώσης των στόχων που μας ενδιαφέρουν αλλά και επιβάλλουν νέα αναλυτικά πρότυπα σε αυτούς τους τομείς (Caspar et al2018, Precone et al 2015).

Για παράδειγμα στην περίπτωση της μελέτης του RNA, οι προσεγγίσεις που βασίζονται σε NGS τεχνολογίες έχουν ξεπεράσει τη χρήση μικροσυστοιχιών και επιτρέπουν την ανάλυση σχεδόν όλων των μορίων RNA, γνωστών και άγνωστων, που υπάρχουν σε ένα δείγμα, με χαμηλότερο κόστος (Precone et al 2015). Επιπλέον, μπορούν να επισημανθούν εναλλακτικές ισομορφές και μη κωδικά μόρια RNA (Su et al 2018) καθώς επίσης μπορούν να εμπλουτιστούν και να αλληλουχηθούν συγκεκριμένες κατηγορίες μικρών μορίων RNA (Nardelli et al 2017). Ακόμα,

πρόσφατες εφαρμογές δείχνουν επίσης τη δυναμική των NGS τεχνολογιών στην αλληλούχιση του RNA μεμονωμένων κυττάρων (Zong et al 2017) και όμοια στη μελέτη του επιγονιδιώματος και του μικροβιώματος. Χρησιμοποιώντας τα πρωτόκολλα παρασκευής συγκεκριμένων βιβλιοθηκών, είναι δυνατή η ανάλυση της κατάστασης μεθυλίωσης του DNA σε επίπεδο γονιδιώματος ή ακόμα και η ανάλυση σε ένα προσαρμοσμένο σύνολο γονιδιωματικών περιοχών ενδιαφέροντος (Pu et al 2017). Επιπλέον, οι προσεγγίσεις αλληλούχισης της χρωματίνης με ανοσοκαταβύθιση (ChIP-Seq) έχουν δείξει την αποτελεσματικότητά τους στη μελέτη των ρυθμιστικών δικτύων έκφρασης γονιδίων σε γενετικό επίπεδο με τον προσδιορισμό των στόχων συγκεκριμένων μεταγραφικών παραγόντων (Pavesi 2017).

Τέλος, αντικαθιστώντας την ανάγκη μικροβιακής καλλιέργειας, οι τεχνικές που βασίζονται σε NGS τεχνολογίες έδωσαν σημαντική ώθηση στη μεταγονιδιωματική για τη μελέτη των μικροβιακών σχέσεων με τη φυσιολογία και την παθολογία του ανθρώπου όπως και για τον προσδιορισμό συγκεκριμένων μικροβιακών υπογραφών που σχετίζονται με μια συγκεκριμένη ασθένεια (D'Argenio et al 2017, D'Argenio & Salvatore 2015). Έχει πλέον διαπιστωθεί ότι το ανθρώπινο μικροβίωμα παίζει ρόλο στη διατήρηση υγιούς κατάστασης (Perez-Muñoz et al 2017, D'Argenio & Salvatore 2015). Κατά συνέπεια, αλλαγές στην κατάσταση του μικροβιώματος μπορούν να συμβάλουν στην ανάπτυξη ασθενειών και μπορούν να προσφέρουν νέους σκοπούς, όχι μόνο όσον αφορά την παρακολούθηση μιας ασθένειας, αλλά κυρίως στην ανάπτυξη νέων θεραπειών.

Καθώς το κόστος των νέων αυτών τεχνολογιών συνεχίζει να μειώνεται με την πάροδο του χρόνου, γίνεται αντιληπτό πως οι εφαρμογές τους θα χρησιμοποιούνται με όλο και μεγαλύτερη συχνότητα και θα αποτελούν μέρος της κλινικής πρακτικής. Πρέπει να σημειωθεί ότι κλινικές και ερευνητικές μελέτες απαιτούν διαφορετική προσέγγιση. Πράγματι, οι κλινικές μελέτες χρειάζονται επικύρωση και αφορούν τις περισσότερες φορές μόνο τα ευρήματα που μπορούν να εφαρμοστούν στην κλινική πρακτική, ενώ οι ερευνητικές μελέτες είναι πιο ρευστές και αφορούν την ανακάλυψη νέας γνώσης. Επιπλέον, με την εμφάνιση των τεχνολογιών NGS, τα επιστημονικά πεδία που κάνουν χρήση αυτών των τεχνολογιών παρουσιάζουν ραγδαία εξέλιξη. Για παράδειγμα, ένας περιορισμός προηγούμενων τεχνολογιών ήταν το περιορισμένο μήκος ανάγνωσής αλληλουχιών, το οποίο σήμερα παύει να είναι εμπόδιο εξαιτίας της αυξημένης

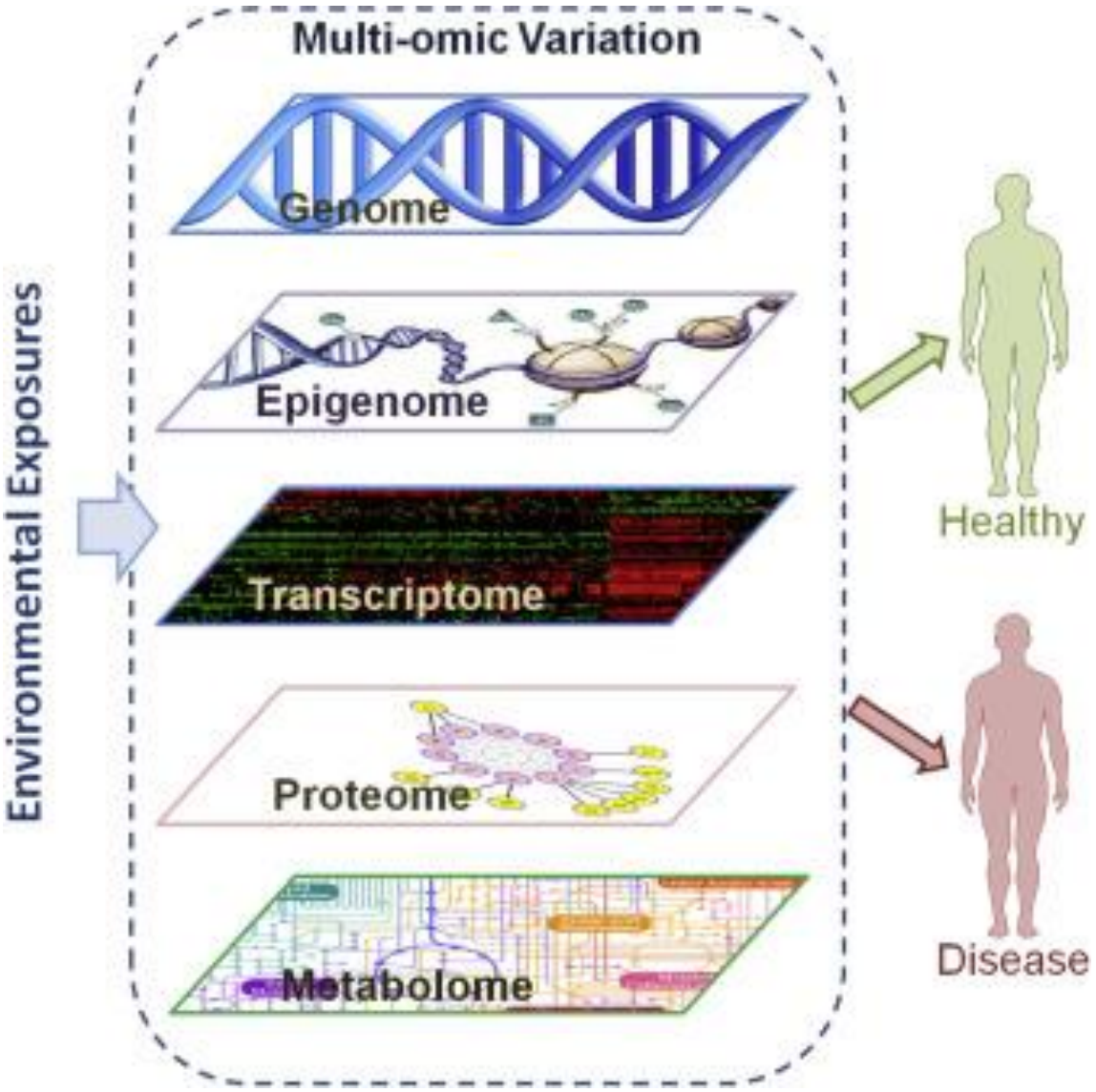
ικανότητας ανάγνωσης των τεχνολογιών επόμενης γενιάς. Οι νέες τεχνικές προσδιορισμού της αλληλουχίας του DNA υπόσχονται περαιτέρω βελτίωση αυτής της εικόνας. Για παράδειγμα, ο προσδιορισμός αλληλουχίας που βασίζεται σε νανοπόρους έχει το πλεονέκτημα πως μπορεί να αποφύγει την ενίσχυση των βιβλιοθηκών (και τα σχετικά σφάλματα) και επιτρέπει την αλληλούχιση πολύ μεγάλων αλληλουχιών (έως 950 kb) (Tyson et al 2018). Όσο αυξάνεται η ακρίβεια αυτών των μεθόδων και ελαχιστοποιείται το περιθώριο σφάλματος, θα παρατηρείται και μια νέα ώθηση στο πεδίο προσδιορισμού αλληλούχισης του DNA καθώς και περαιτέρω μείωση του κόστους αλληλούχισης για κάθε γονιδίωμα.

Εκτός από τις προαναφερθείσες βελτιώσεις, με παρόμοιο τρόπο έχουν αναπτυχθεί πλατφόρμες φασματομετρίας μάζας υψηλής απόδοσης (MS) για την εκμετάλλευση ολόκληρου του πρωτεώματος με μεγαλύτερο βάθος ή/και του μεταβολώματος των κυττάρων (Jacob et al 2019). Αντιπροσωπεύοντας τα τελικά προϊόντα που προκύπτουν από τις κυτταρικές διεργασίες, η μελέτη των πρωτεϊνών και των μεταβολιτών σε συνδυασμό με δεδομένα από άλλα -omics δεδομένα παρέχουν τη δυνατότητα καλύτερης κατανόησης των παθογενετικών μηχανισμών και ανάδειξης επιπλέον βιοδεικτών. Η εξέλιξη της φασματομετρίας μάζας επιτρέπει την ταυτόχρονη ανάλυση πολλαπλών πεπτιδίων/μεταβολιτών και επίσης δίνει τη δυνατότητα για μη στοχευμένες προσεγγίσεις ανίχνευσης νέων μορίων. Ωστόσο, η ταυτοποίηση νέων πεπτιδίων/μεταβολιτών βασίζεται στη σύγκριση των αποτελεσμάτων του υπό ανάλυση δείγματος με δεδομένα από συγκεκριμένες βάσεις βιολογικών δεδομένων που εξακολουθούν να παρουσιάζουν περιορισμούς. Καθώς θα επιτευχθεί περαιτέρω τεχνολογική πρόοδος, τόσο το πρωτεϊνικό όσο και το μεταβολωμικό προφίλ μπορούν να ενσωματώσουν γονιδιωματικά δεδομένα για καλύτερη διαγνωστική και προγνωστική ταξινόμηση.

1.2.Μέθοδοι ενσωμάτωσης πολλαπλών δεδομένων (multiomics)

Με την πάροδο του χρόνου, οι νέες τεχνολογίες εξελίσσονται με αποτέλεσμα να υπάρχουν σε σχετική αφθονία δεδομένα που παράγονται από διαφορετικές μελέτες όπως transcriptomics, proteomics κ.α. Η συνήθης διαδικασία που λαμβάνει χώρα είναι η ανάλυση αυτών των

δεδομένων ξεχωριστά και η οποία εμπλουτίζει συνεχώς τη γνώση που περιβάλλει βιολογικούς μηχανισμούς, εξέλιξη ασθενειών, ομοιότητες και διαφορές δειγμάτων. Τα τελευταία χρόνια γίνεται προσπάθεια ώστε να αναλύονται από κοινού δεδομένα διαφορετικού τύπου ώστε να παράγεται μία περισσότερο ολιστική εικόνα για ένα βιολογικό σύστημα ή την εξέλιξη μιας ασθένειας. Για να επιτευχθεί αυτό, έχουν αρχίσει να αναπτύσσονται μέθοδοι που έχουν ως στόχο την από κοινού μελέτη των δεδομένων που παράγονται (multiomics integration). Τα εργαλεία αυτά παράγονται από συγκεκριμένα εργαστήρια και ομάδες σε διαφορετικά ερευνητικά κέντρα ή/και πανεπιστήμια με σκοπό να εξυπηρετήσουν τις εκάστοτε ανάγκες και έχουν τη δυνατότητα να ενσωματώνουν περισσότερα των δύο διαφορετικά σύνολα δεδομένων που προέρχονται από -omics μελέτες.



Εικόνα 1. Σχηματική απεικόνιση multi-omics δεδομένων (Πηγή: Advances in Genetics, Elsevier)

Οι μέθοδοι που χρησιμοποιούνται για την ενσωμάτωση πολλαπλών τύπων -omics δεδομένων εμπίπτουν σε δύο γενικές κατηγορίες, αυτές της εποπτευόμενης (supervised) και της μη-εποπτευόμενης (unsupervised) ανάλυσης.

Στην κατηγορία της εποπτευόμενης ανάλυσης κύριοι στόχοι είναι η εύρεση των σημαντικών μεταβλητών οι οποίες έχουν μεγάλη διακριτική αξία και παρέχουν τη δυνατότητα διαχωρισμού των δειγμάτων σε ομάδες με συγκεκριμένα κοινά χαρακτηριστικά καθώς και η δυνατότητα χρήσης των σημαντικών αυτών μεταβλητών για τη μελέτη πρόβλεψης. Για την περίπτωση αυτή

καθίσταται δυνατή, με χρήση των σημαντικών μεταβλητών, η ταξινόμηση νέων δειγμάτων στις ομάδες που έχουν προκύψει. Ένα απλό παράδειγμα είναι το γονίδιο BRCA1 και η πρόβλεψη που μπορεί να παρέχει η αλληλούχιση του, για την πρόληψη του καρκίνου του μαστού. Συγκεκριμένες μεταλλάξεις στο γονίδιο αυτό, οι οποίες σήμερα είναι πολύ εύκολο να εντοπιστούν με την εξέλιξη της τεχνολογίας, καθιστούν το γονίδιο (μεταβλητή) ικανό ώστε να παρθεί απόφαση για μια γυναίκα και να προχωρήσει σε μαστεκτομή. Η προβλεπτική αξία του συγκεκριμένου γονιδίου για τη συγκεκριμένη μορφή καρκίνου είναι πολύ υψηλή και συνεπώς γίνεται λόγος για μία μεταβλητή η οποία μπορεί να χρησιμοποιηθεί για ταξινόμηση νέων δειγμάτων με μεγάλη ακρίβεια.

Οι μέθοδοι που εμπίπτουν στο πλαίσιο της μη-εποπτευόμενης (unsupervised) ανάλυσης παρέχουν τη δυνατότητα για ανακάλυψη νέων σχέσεων μεταξύ των δειγμάτων αλλά και μεταξύ των μεταβλητών (μετρήσεων) που έχουν παραχθεί από τεχνολογίες NGS στα διαφορετικού τύπου -omics δεδομένα. Βασικό χαρακτηριστικό αυτών των μεθόδων είναι πως είναι περισσότερο αμερόληπτες (unbiased) καθώς δεν υπεισέρχεται ο ανθρώπινος παράγοντας στον χαρακτηρισμό των δειγμάτων και η εξόρυξη γνώσης βασίζεται στην πληροφορία που φέρουν τα δεδομένα. Παραμένει βέβαια και σε αυτήν την περίπτωση η πιθανότητα σφάλματος στις μετρήσεις που πραγματοποιούνται από τα μηχανήματα.

Οι περισσότερες από τις μεθόδους αυτές αναπτύσσονται βασισμένες στην παλινδρόμηση μερικών ελαχίστων τετραγώνων, στην ανάλυση κύριων συνιστωσών, στην ανάλυση κανονικής συσχέτισης και την παραγοντοποίηση πινάκων.

Η παλινδρόμηση μερικών ελαχίστων τετραγώνων (Partial Least Square Regression-PLS) είναι μία σχετικά πρόσφατη τεχνική (Wold 1996) η οποία χρησιμοποιήθηκε για πρώτη φορά στον τομέα της οικονομίας. Σήμερα χρησιμοποιείται πολύ ευρύτερα και σε τομείς όπως η υπολογιστική βιολογία, η νευροαπεικόνιση και αλλού. Ανήκει στην ευρύτερη οικογένεια της πολλαπλής παλινδρόμησης και η χρήση της παρουσιάζει πλεονεκτήματα όταν ο αριθμός των μεταβλητών είναι μεγάλος.

Η ανάλυση κύριων συνιστωσών (Principal Component Analysis-PCA) είναι μία πολυμεταβλητή στατιστική τεχνική η οποία στηρίζεται σε αναδιάταξη των αρχικών μεταβλητών σε ένα νέο

σύστημα ορθογώνιων μεταβλητών που ονομάζονται κύριες συνιστώσες. Χρησιμοποιείται κυρίως όταν το πλήθος των αρχικών μεταβλητών είναι πολύ μεγάλο (έκφραση γονιδίων) και η προέλευση της ανάγεται στον Pearson (1901) και αργότερα στον Hotelling (1930). Η χρήση της σήμερα είναι ευρέως διαδεδομένη σε πολλά επιστημονικά πεδία.

Η Ανάλυση Κανονικής Συσχέτισης (Canonical Correlation Analysis-CCA) μπορεί να θεωρηθεί ως η πιο γενική μέθοδος των μεθόδων ελαχίστων τετραγώνων για την ανάλυση συνόλων δεδομένων. Ο στόχος της είναι να δώσει μια απλή περιγραφή της δομής της συσχέτισης μεταξύ υποσυνόλων μεταβλητών, δηλαδή επιδιώκει να προσδιορίσει και να ποσοτικοποιήσει τις σχέσεις μεταξύ δύο συνόλων μεταβλητών.

Τέλος η παραγοντοποίηση πινάκων (Matrix Factorization) αποτελεί χρήσιμο εργαλείο για την επίλυση γραμμικών συστημάτων. Ξεκινώντας από ένα γραμμικό σύστημα ο βασικός πίνακας, ο οποίος αποτελείται από τις μεταβλητές του προβλήματος, παραγοντοποιείται και με αυτόν τον τρόπο η επίλυση του αρχικού συστήματος ανάγεται στην επίλυση απλουστευμένων συστημάτων.

2. Μαθηματικές και στατιστικές τεχνικές

Σε αυτό το κεφάλαιο έχουν συγκεντρωθεί και παρουσιάζονται βασικές πληροφορίες που αφορούν το μαθηματικό υπόβαθρο των εργαλείων που χρησιμοποιούνται σε multi-omics αναλύσεις. Επίσης περιγράφονται κάποιες στατιστικές τεχνικές των οποίων η χρήση είναι ευρέως διαδεδομένη αλλά συγχρόνως αποτελεί μέρος των εργαλείων που θα εξεταστούν.

2.1. Παλινδρόμηση μερικών ελαχίστων τετραγώνων

Η παλινδρόμηση μερικών ελάχιστων τετραγώνων (PLS) είναι μια τεχνική που μειώνει το πλήθος των προγνωστικών παραγόντων (εξαρτημένες μεταβλητές) σε ένα μικρότερο σύνολο μη συσχετισμένων συνιστωσών και πραγματοποιεί παλινδρόμηση ελάχιστων τετραγώνων στις νέες συνιστώσες, αντί για τις αρχικές. Η παλινδρόμηση PLS είναι ιδιαίτερα χρήσιμη όταν οι ανεξάρτητες μεταβλητές παρουσιάζουν υψηλή συγγραμικότητα ή σε περιπτώσεις που το πλήθος των προγνωστικών παραγόντων είναι μεγαλύτερο από αυτό των παρατηρήσεων και η συνήθης παλινδρόμηση ελάχιστων τετραγώνων είτε παράγει συντελεστές με μεγάλο σφάλμα είτε αποτυγχάνει εντελώς. Η μέθοδος PLS δεν υποθέτει ότι οι προγνωστικοί παράγοντες είναι σταθεροί, σε αντίθεση με την πολλαπλή παλινδρόμηση γεγονός που σημαίνει ότι οι προγνωστικοί παράγοντες μπορούν να μετρηθούν με κάποιο σφάλμα, καθιστώντας έτσι τη συγκεκριμένη μέθοδο πιο ισχυρή απέναντι στα σφάλματα των μετρήσεων.

Η μέθοδος PLS είναι μια στατιστική μέθοδος που έχει κάποια σχέση με την παλινδρόμηση των κύριων συνιστωσών όμως αντί να αναζητά υποχώρους που μεγιστοποιούν τη διακύμανσης μεταξύ της εξαρτημένης μεταβλητής και των ανεξάρτητων μεταβλητών, δημιουργεί ένα μοντέλο γραμμικής παλινδρόμησης προβάλλοντας τις ανεξάρτητες μεταβλητές (predictors) και τις παρατηρούμενες σε ένα νέο χώρο. Επειδή και οι δύο πίνακες δεδομένων **X** (predictors) και **Y** (response) προβάλλονται σε νέους χώρους, οι μέθοδοι της κατηγορίας PLS είναι γνωστές ως διγραμμικά (bilinear) παραγοντικά μοντέλα. Η παλινδρόμηση μερικών ελάχιστων τετραγώνων – διακριτική ανάλυση (PLS-DA) είναι μια παραλλαγή που χρησιμοποιείται όταν η εξαρτημένη μεταβλητή **Y** είναι κατηγορική και στόχος είναι να κατηγοριοποιηθούν τα δεδομένα σε ένα πεπερασμένο αριθμό ομάδων. Αν και οι αρχικές εφαρμογές της μεθόδου παρατηρούνταν στις

κοινωνικές επιστήμες, η παλινδρόμηση PLS χρησιμοποιείται σήμερα και σε άλλους τομείς όπως στη βιοπληροφορική, στις νευροεπιστήμες και την ανθρωπολογία.

Το γενικό μοντέλο των πολυμεταβλητών PLS είναι το παρακάτω:

$$\begin{cases} X = TP^T + E \\ Y = UQ^T + F \end{cases} \qquad (1)$$

όπου **X** είναι ένας $n \times m$ πίνακας που περιέχει τις ανεξάρτητες μεταβλητές, **Y** ένας $n \times p$ πίνακας των εξαρτημένων μεταβλητών. Οι πίνακες **T** και **U** είναι διάστασης $n \times l$ οι οποίοι είναι προβολές του **X** και του **Y** αντίστοιχα. Αντιστοίχως, οι **P** και **Q** είναι οι ορθογώνιοι πίνακες φόρτωσης (loadings) και οι πίνακες **E**, **F** περιέχουν τα σφάλματα, που θεωρούνται ανεξάρτητες κανονικές τυχαίες μεταβλητές με όμοια κατανομή. Η παραγοντοποίηση των **X** και **Y** γίνεται με τέτοιο τρόπο ώστε να μεγιστοποιείται η συνδιακύμανση μεταξύ των **T** και **U** πινάκων.

Η μέθοδος PLS-DA είναι μία ειδική περίπτωση της παλινδρόμησης PLS, όπου ο πίνακας **Y** λαμβάνει διακριτές τιμές. Στη συνήθη περίπτωση μοντέλου πολλαπλής γραμμικής παλινδρόμησης (MLR) ισχύει

$$Y = XB + F \quad (2)$$

όπου **X** είναι ο πίνακας δεδομένων $n \times j$, **B** είναι ο πίνακας συντελεστών παλινδρόμησης $j \times 1$, **F** ο πίνακας σφαλμάτων $n \times 1$ και **Y** ο πίνακας εξαρτημένης μεταβλητής $n \times 1$. Σε αυτήν την προσέγγιση, η λύση των ελάχιστων τετραγώνων δίνεται από τη σχέση $B = (X^T X)^{-1} X^T Y$.

Σε πολλές περιπτώσεις, το πρόβλημα είναι η ανωμαλία (singularity) του πίνακα $X^T X$ (π.χ., όταν υπάρχουν προβλήματα πολλαπλής συγγραμμικότητας στα δεδομένα ή ο αριθμός των προγνωστικών μεταβλητών είναι μεγαλύτερος από τον αριθμό των παρατηρήσεων). Τόσο η μέθοδος PLS όσο και η PLS-DA αντιπαρέρχονται του συγκεκριμένου προβλήματος με την παραγοντοποίηση του πίνακα **X** σε P ορθογώνιες βαθμολογίες **T** ($n \times P$) και τον πίνακα φορτίων (loadings) **P** ($J \times P$). Ο πίνακας **Y** των εξαρτημένων μεταβλητών αναλύεται αντίστοιχα σε P ορθογώνιες βαθμολογίες **T** ($n \times P$) και στον πίνακα φορτίων **Q** ($1 \times P$). Έστω, **E** ($n \times J$) και **F** ($n \times 1$) οι πίνακες των σφαλμάτων που σχετίζονται με τους πίνακες **X** και **Y** αντίστοιχα. Τότε προκύπτουν δύο θεμελιώδεις εξισώσεις στο μοντέλο PLS-DA:

$$X = TP^T + E$$
$$Y = TQ^T + F \qquad (3)$$

Αν οριστεί ένας πίνακας βαρών (weights) **W** ($J \times P$), ο πίνακας των scores μπορεί να γραφεί ως

$$T = XW(P^TW)^{-1} \qquad (4)$$

και με αντικατάσταση στο μοντέλο PLS-DA, η δεύτερη σχέση της (3) θα γίνει

$$Y = XW(P^TW)^{-1}Q^T + F \qquad (5)$$

όπου ο πίνακας των συντελεστών παλινδρόμησης **B** δίνεται από τη σχέση

$$\hat{B} = W(P^TW)^{-1}Q^T \quad (6).$$

Με αυτόν τον τρόπο, μια άγνωστη τιμή ενός δείγματος που ανήκει στον πίνακα **Y** μπορεί να προβλεφθεί από τη σχέση $\hat{Y} = X\hat{B}$, η οποία ανάγεται τελικά στην

$$\hat{Y} = XW(P^TW)^{-1}Q^T \ (7).$$

2.2. Ανάλυση κύριων συνιστωσών

Η ανάλυση κύριων συνιστωσών αποτελεί την απλούστερη και πλέον διαδεδομένη πολυμεταβλητή ανάλυση και στοχεύει στην ανεύρεση από ένα πλήθος p μεταβλητών ορισμένων νέων ολιγάριθμων μεταβλητών οι οποίες έχουν την ιδιότητα να είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών και παράλληλα να μη συσχετίζονται μεταξύ τους. Το μεγάλο πλεονέκτημά τους έγκειται στην ιδιαιτερότητα που διαθέτουν, λόγω της ανάλυσης, να εξηγούν πολύ μεγάλο ποσοστό της ολικής μεταβλητότητας που αναπτύσσεται μεταξύ των p μεταβλητών, το οποίο τελικά κατανέμεται σε μερικές μόνο νέες μεταβλητές. Έτσι, το μέγιστο μέρος της πληροφόρησης που θα αντλούνταν αν λαμβάνονταν υπόψη οι p μεταβλητές συγκρατείται με τη δημιουργία αυτών των νέων μεταβλητών. Η διαδικασία της ανάλυσης βασίζεται στις ακόλουθες αρχές:

1. Από τις p μεταβλητές $X_1, X_2, \dots, X_p$, δημιουργούνται p συνδυασμοί αυτών $Z_1, Z_2, \dots, Z_p$, με τέτοιο τρόπο ώστε να μη συσχετίζονται μεταξύ τους. Η απουσία συσχετισμού μεταξύ των μεταβλητών Z_i προδιαθέτει ότι αυτές μετρούν διαφορετικές "διαστάσεις" των στοιχείων.
2. Οι διακυμάνσεις (μεταβλητότητα) που αναπτύσσονται μεταξύ των μεταβλητών Z_i, διαβαθμίζονται με τέτοιο τρόπο ώστε η πρώτη μεταβλητή Z_1 επιλέγεται να εξηγεί ένα όσο το δυνατόν μέγιστο ποσοστό της ολικής μεταβλητότητας, η Z_2 το δεύτερο μεγαλύτερο ποσοστό αυτής κοκ., ικανοποιώντας τη σχέση: $\lambda_1 > \lambda_2 > \cdots > \lambda_\pi$, όπου λ_i η i ποσότητα της διακύμανσης. Οι νέες μεταβλητές Z_i καλούνται κύριες συνιστώσες και με τον τρόπο αυτόν δημιουργούνται λίγες στο πλήθος Z συνιστώσες, οι οποίες ωστόσο εξηγούν μεγάλο ποσοστό της συνολικής διακύμανσης $\sum \lambda_i$. Ταυτόχρονα, πολυάριθμες δευτερεύουσες συνιστώσες εξηγούν μικρό έως ελάχιστο ποσοστό και συνεπώς το στατιστικό τους αποτέλεσμα μπορεί να αγνοηθεί χωρίς την απώλεια ουσιαστικής πληροφόρησης.

Η τεχνική των κύριων συνιστωσών έχει ως βάση, κατά τη διαδικασία υπολογισμού της, τον πίνακα των συσχετίσεων (correlation matrix) κατά ζεύγη των μεταβλητών. Κατά συνέπεια, για να θεωρείται η τεχνική επιτυχημένη, απαραίτητη προϋπόθεση είναι κάποιοι συντελεστές συσχέτισης των αρχικών μεταβλητών του πίνακα συσχετίσεων να φέρουν υψηλές τιμές θετικές ή αρνητικές (π.χ. $r \geq \pm 0.7$). Έτσι, καθίσταται δυνατό ένα σύνολο πολλών μεταβλητών να είναι σε θέση να αντιπροσωπευτεί από δύο έως τρεις κύριες συνιστώσες, αρκεί να καλύπτεται η προϋπόθεση της παρουσίας υψηλών συντελεστών στον πίνακα των συσχετίσεων. Από την άλλη πλευρά, αρχικές μεταβλητές με πολύ ισχυρές τιμές συσχετίσεων $\geq \pm 0.99$ θεωρούνται πλεονάζουσες και κάποιες από αυτές θα πρέπει να απορρίπτονται πριν από την εφαρμογή της μεθόδου.

Στη συνέχεια οι αρχικές μεταβλητές μετασχηματίζονται σε τυποποιημένες σύμφωνα με την παρακάτω σχέση:

$$(X_i - \bar{X})/s \qquad (2)$$

όπου $\bar{X}$ η μέση τιμή της μεταβλητής και s η τυπική απόκλιση.

Η πρώτη κύρια συνιστώσα προκύπτει από το γραμμικό συνδυασμό των p μεταβλητών,

$$Z_1 = \alpha_{11}X_1 + \alpha_{12}X_2 + \cdots + \alpha_{1p}X_p \quad (3)$$

όπου α_{ij} ο ειδικός συντελεστής στάθμισης (weight) της j μεταβλητής στην i κύρια συνιστώσα και με τον περιορισμό ότι

$${\alpha_{11}}^2 + {\alpha_{12}}^2 + \cdots + {\alpha_{1p}}^2 = 1 \qquad (4)$$

εξαιτίας του οποίου εξασφαλίζεται η εκτίμηση της μέγιστης διακύμανσης λ_1 της Z_1. Σε αντίθετη περίπτωση, η διακύμανση θα αυξανόταν απεριόριστα με την απλή και μόνο αύξηση ενός από τους συντελεστές στάθμισης.

Η δεύτερη συνιστώσα προκύπτει ομοίως ως συνδυασμός των $X_1, X_2, \dots, X_p$

$$Z_2 = \alpha_{21}X_1 + \alpha_{22}X_2 + \cdots + \alpha_{2p}X_p \quad (5)$$

όπου τα $\alpha_{21}, \alpha_{22}, \dots, \alpha_{2p}$ ικανοποιούν και εδώ τον περιορισμό

$${\alpha_{21}}^2 + {\alpha_{22}}^2 + \cdots + {\alpha_{2p}}^2 = 1 \qquad (6)$$

Ένας επιπρόσθετος σημαντικός περιορισμός που εισάγεται στο σημείο αυτό είναι ο συντελεστής συσχέτισης μεταξύ των συνιστωσών Z_1 και Z_2 να είναι ίσος με μηδέν.

Η τρίτη συνιστώσα υπολογίζεται ως:

$$Z_3 = \alpha_{31}X_1 + \alpha_{32}X_2 + \cdots + \alpha_{3p}X_p \quad (7)$$

Με αντίστοιχο περιορισμό για τους συντελεστές $\alpha_{31}, \alpha_{32}, \dots, \alpha_{3p}$, όπως προηγουμένως:

$${\alpha_{31}}^2 + {\alpha_{32}}^2 + \cdots + {\alpha_{3p}}^2 = 1 \qquad (8)$$

και με τη νέα προϋπόθεση πως η συνιστώσα Z_3 θα πρέπει να είναι ασυσχέτιστη με τις δύο προηγούμενες συνιστώσες Z_1 και Z_2.

Με την παραπάνω διαδικασία δημιουργούνται p συνιστώσες Z_i ίσου πλήθους με το πλήθος των αρχικών μεταβλητών X_i.

Οι συντελεστές στάθμισης α_{ij} υπολογίζονται με τη βοήθεια του πίνακα C των συνδιακυμάνσεων των αρχικών μεταβλητών,

$$C = \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1p} \\ c_{21} & c_{22} & \dots & c_{2p} \\ c_{p1} & c_{p2} & \dots & c_{pp} \end{pmatrix} \quad (9)$$

όπου τα διαγώνια στοιχεία c_{ii} είναι οι διακυμάνσεις της X_i μεταβλητής και τα στοιχεία c_{ij} του πίνακα C αποτελούν τις συνδιακυμάνσεις των μεταβλητών X_i και X_j.

Με την τυποποίηση των αρχικών μεταβλητών ο πίνακας των συνδιακυμάνσεων μετατρέπεται στον πίνακα των συσχετίσεων ως

$$C = \begin{pmatrix} 1 & c_{12} & \dots & c_{1p} \\ c_{21} & 1 & \dots & c_{2p} \\ c_{p1} & c_{p2} & \dots & 1 \end{pmatrix} \quad (10)$$

και έτσι προκύπτει $c_{ii} = 1$, ενώ $c_{ij} = c_{ji}$ είναι ο συντελεστής συσχέτισης μεταξύ των X_i και X_j. Ουσιαστικά, η ανάλυση των κύριων συνιστωσών εκτελείται με βάση τον πίνακα των συσχετίσεων.

Οι διακυμάνσεις των κύριων συνιστωσών καλούνται χαρακτηριστικές ρίζες ή ιδιοτιμές λ_i (eigenvalues) και το πλήθος τους είναι ίσο με το πλήθος των συνιστωσών p. Ακόμα ισχύει για τις ιδιοτιμές η σχέση $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$. Μία επιπλέον σημαντική ιδιότητα των χαρακτηριστικών ριζών είναι ότι το άθροισμά τους ισοδυναμεί με το άθροισμα των διακυμάνσεων των αρχικών μεταβλητών:

$$\lambda_1 + \lambda_2 + \cdots + \lambda_p = c_{11} + c_{22} + c_{pp} \quad (11)$$

Αφού c_{ii} είναι η διακύμανση της X_i και λ_i η διακύμανση της Z_i, εύκολα συνάγεται με βάση τους παραπάνω υπολογισμούς ότι οι κύριες συνιστώσες εξηγούν πλήρως όλη τη μεταβλητότητα των αρχικών δεδομένων.

Συνοψίζοντας, τα στάδια της ανάλυσης των κύριων συνιστωσών έχουν ως εξής:

1. Τυποποίηση των αρχικών μεταβλητών $X_1, X_2, \dots, X_p$, έτσι ώστε να έχουν μέσο όρο μηδέν και διακύμανση ίση με 1.

2. Υπολογισμός του πίνακα των συνδιακυμάνσεων, ο οποίος πλέον έχει την έννοια του πίνακα των συσχετίσεων.
3. Εκτίμηση των χαρακτηριστικών ριζών $\lambda_1, \lambda_2, \dots, \lambda_p$, και των συντελεστών στάθμισης α_{ij} και κατά συνέπεια των διανυσμάτων $\alpha_1, \alpha_2, \dots, \alpha_p$. Οι συντελεστές της i κύριας συνιστώσας εμφανίζονται στο διάνυσμα α_i και η διακύμανση αυτής απεικονίζεται στη χαρακτηριστική ρίζα λ_i.
4. Απορρίπτονται όλες οι συνιστώσες που εξηγούν μικρό ποσοστό της ολικής μεταβλητότητας και επιλέγονται μόνον οι πλέον σημαντικές. Για παράδειγμα, εκκινώντας με 100 αρχικές μεταβλητές είναι δυνατόν μέσω της ανάλυσης να εντοπιστούν οι 3 πρώτες κύριες συνιστώσες οι οποίες εξηγούν το 90% της ολικής μεταβλητότητας και κατά συνεπώς να αγνοηθούν οι υπόλοιπες 97, αφού το αποτέλεσμα της δράσης αυτών αθροίζει στο 10% της συνολικής μεταβλητότητας των δεδομένων και συνεπώς είναι σχεδόν ασήμαντο.

Οι συσχετίσεις r_{ij} μεταξύ των αρχικών μεταβλητών και των κύριων συνιστωσών ονομάζονται φορτία (loadings) και δείχνουν τη συμμετοχή (contribution) που των αρχικών μεταβλητών για τη δημιουργία των συνιστωσών, σε τι βαθμό είναι, δηλαδή, υπεύθυνες γι' αυτές. Η ένταση της σχέσης είναι εξαιρετικά ισχυρή σε τιμές κοντά στο ± 1.0 (ισχυρή θετική ή αρνητική συσχέτιση) και ασήμαντη σε τιμές κοντά στο μηδέν. Έτσι, όσο υψηλότερα είναι τα φορτία τόσο σημαντικότερες είναι οι υποψήφιες μεταβλητές για το σχηματισμό των κύριων συνιστωσών. Τα φορτία υπολογίζονται επίσης και από τη σχέση:

$$l_{ij} = \frac{\alpha_{ij}}{s_j} \cdot \sqrt{\lambda_i} \qquad (12)$$

όπου l_{ij} είναι το φορτίο της μεταβλητής j για την i συνιστώσα, α_{ij} είναι ο συντελεστής στάθμισης της μεταβλητής j για την i συνιστώσα επίσης, λ_i είναι η χαρακτηριστική ρίζα της i συνιστώσας και s_j είναι η τυπική απόκλιση της μεταβλητής j.

Υπενθυμίζεται ότι δύο οποιεσδήποτε κύριες συνιστώσες έχουν μηδενική συσχέτιση.

2.3. Ανάλυση κανονικής συσχέτισης

Ας υποθέσουμε ότι δύο σύνολα μεταβλητών $\boldsymbol{X_1}$ και $\boldsymbol{X_2}$ έχουν κοινή κανονική κατανομή, όπως φαίνεται στην (1)

$$\begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} \sim N_p \left[\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{pmatrix} \right] \qquad (1)$$

Η Ανάλυση Κανονικής Συσχέτισης εστιάζει στην συσχέτιση μεταξύ ενός γραμμικού συνδυασμού $\boldsymbol{a'X_1}$ των μεταβλητών σε ένα σύνολο $\boldsymbol{X_1}$, και ενός γραμμικού συνδυασμού $\boldsymbol{b'X_1}$ των μεταβλητών σε ένα άλλο σύνολο $\boldsymbol{X_2}$. Η ιδέα είναι να καθορίσουμε πρώτα το ζεύγος των γραμμικών συνδυασμών που έχει τη μεγαλύτερη συσχέτιση. Στη συνέχεια προσδιορίζουμε το ζεύγος των γραμμικών συνδυασμών που έχει τη μεγαλύτερη συσχέτιση μεταξύ όλων των ζευγών, τα οποία είναι ασυσχέτιστα με το αρχικά επιλεγμένο ζεύγος και επαναλαμβάνουμε μέχρις ότου όλες οι πιθανές συσχετίσεις εξαντληθούν. Τα ζεύγη των γραμμικών συνδυασμών ονομάζονται κανονικές μεταβλητές και οι συσχετίσεις τους ονομάζονται κανονικές συσχετίσεις. Οι κανονικές συσχετίσεις μετρούν την ισχύ την συσχέτισης μεταξύ δύο συνόλων μεταβλητών. Η πτυχή μεγιστοποίησης της τεχνικής αντιπροσωπεύει μία προσπάθεια να επικεντρώσει μία υψηλών διαστάσεων σχέση μεταξύ δύο συνόλων μεταβλητών σε μερικά ζεύγη κανονικών μεταβλητών.

<u>Ορισμός και βασικά στοιχεία</u>

Υποθέτουμε ότι έχουμε δύο σύνολα μεταβλητών με την ίδια μονάδα μέτρησης. Το πρώτο σύνολο αποτελείται από p μεταβλητές και συμβολίζεται από το διάνυσμα $\boldsymbol{X_1} = (X_{11}, X_{12}, \dots, X_{1p})$, ενώ το δεύτερο σύνολο αποτελείται από q μεταβλητές και συμβολίζεται από το διάνυσμα $\boldsymbol{X_2} = (X_{21}, X_{22}, \dots, X_{2q})$. Ας θεωρήσουμε ότι $p > q$, δηλαδή το σύνολο $\boldsymbol{X_1}$ είναι το μικρότερο σύνολο.

Θεωρούμε ότι για τα τυχαία διανύσματα $\boldsymbol{X_1}$ και $\boldsymbol{X_2}$ ισχύουν:

$$E(\boldsymbol{X_1}) = \boldsymbol{\mu_1}, \qquad Cov(\boldsymbol{X_1}) = \boldsymbol{\Sigma_{11}}$$

$$E(\boldsymbol{X_2}) = \boldsymbol{\mu_2}, \qquad Cov(\boldsymbol{X_2}) = \boldsymbol{\Sigma_{22}}$$

$$Cov(\boldsymbol{X_1}, \boldsymbol{X_2}) = \boldsymbol{\Sigma_{12}} = \boldsymbol{\Sigma'_{21}}$$

Επίσης μπορούμε να θεωρήσουμε τα τυχαία διανύσματα $\boldsymbol{X_1}$ και $\boldsymbol{X_2}$ από κοινού συνδεδεμένα σε ένα διάνυσμα $\boldsymbol{X}$ διάστασης $(p+q)\times 1$ το οποίο δίνεται παρακάτω:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X_1} \\ \boldsymbol{X_2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X_{11}} \\ \boldsymbol{X_{12}} \\ \dots \\ \boldsymbol{X_{1p}} \\ \dots \\ \boldsymbol{X_{21}} \\ \boldsymbol{X_{22}} \\ \dots \\ \boldsymbol{X_{2q}} \end{bmatrix}$$

με διάνυσμα μέσης τιμής $\boldsymbol{\mu}$ διάστασης $(p+q)\times 1$ και πίνακα συνδιακύμανσης $\boldsymbol{\Sigma}$ διάστασης $(p+q)\times(p+q)$, τα οποία δίνονται παρακάτω:

$$\boldsymbol{\mu} = E(\boldsymbol{X}) = \begin{bmatrix} E(\boldsymbol{X_1} \\ E(\boldsymbol{X_2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{bmatrix}$$

$$\boldsymbol{\Sigma} = E(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})' = \begin{bmatrix} E(\boldsymbol{X_1}-\boldsymbol{\mu_1})(\boldsymbol{X_1}-\boldsymbol{\mu_1})' & E(\boldsymbol{X_1}-\boldsymbol{\mu_1})(\boldsymbol{X_2}-\boldsymbol{\mu_2})' \\ E(\boldsymbol{X_1}-\boldsymbol{\mu_1})(\boldsymbol{X_1}-\boldsymbol{\mu_1})' & E(\boldsymbol{X_2}-\boldsymbol{\mu_2})(\boldsymbol{X_2}-\boldsymbol{\mu_2})' \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{bmatrix},$$

όπου $\boldsymbol{\Sigma_{11}}$ είναι ο πίνακας συνδιακύμανσης $p\times p$ του $\boldsymbol{X_1}$, $\boldsymbol{\Sigma_{11}}$ είναι ο πίνακας συνδιακύμανσης $p\times q$ των $\boldsymbol{X_1}$ και $\boldsymbol{X_2}$, του οποίου τα στοιχεία μετρούν την σχέση μεταξύ των δύο συνόλων και $\boldsymbol{\Sigma_{22}}$ ο πίνακας συνδιακύμανσης $q\times q$ του $\boldsymbol{X_2}$.

Ο κύριος σκοπός της Ανάλυσης Κανονικής Συσχέτισης είναι να συνοψίσει τις σχέσεις μεταξύ των συνόλων $\boldsymbol{X_1}$ και $\boldsymbol{X_2}$ από την άποψη λίγων προσεκτικά επιλεγμένων συνδιασπορών αντί όλων των συνδιασπορών του πίνακα $\boldsymbol{\Sigma_{12}}$. Όπως ήδη έχει αναφερθεί μας ενδιαφέρουν οι γραμμικοί συνδυασμοί των μεταβλητών. Οι γραμμικοί συνδυασμοί παρέχουν απλά συνοπτικά μέτρα ενός συνόλου μεταβλητών. Ας θέσουμε:

$$\begin{pmatrix} \boldsymbol{U} = \boldsymbol{a}'\boldsymbol{X_1} = a_1X_{11} + a_2X_{12} + \cdots + a_pX_{1p} \\ \boldsymbol{V} = \boldsymbol{b}'\boldsymbol{X_1} = b_1X_{21} + b_2X_{22} + \cdots + b_qX_{1q} \end{pmatrix} \qquad (2.1)$$

για κάποιο ζεύγος διανυσμάτων συντελεστών $\boldsymbol{a}, \boldsymbol{b}$. Οι διασπορές των γραμμικών συνδυασμών U και V δίνονται από τις ακόλουθες σχέσεις:

$$\begin{pmatrix} Var(U) = Var(\boldsymbol{a'X_1}) = \boldsymbol{a'}Cov(\boldsymbol{X_1})\boldsymbol{a} = \boldsymbol{a'\Sigma_{11}a} \\ Var(V) = Var(\boldsymbol{b'X_2}) = \boldsymbol{b'}Cov(\boldsymbol{X_2})\boldsymbol{b} = \boldsymbol{b'\Sigma_{22}b} \end{pmatrix} \quad (2.2)$$

ενώ η συνδιασπορά τους θα είναι:

$$Cov(U,V) = Cov(a'X_1, b') = a'Cov(X_1, X_2)b = a'\Sigma_{12}b \quad (2.3)$$

Σκοπός μας είναι η εύρεση διανυσμάτων συντελεστών $\boldsymbol{a}, \boldsymbol{b}$ έτσι ώστε ο συντελεστής συσχέτισης των U, V, ο οποίος ορίζεται από την ακόλουθη σχέση (2.4), να είναι όσο το δυνατόν μεγαλύτερος

$$\rho_{U,V} = Corr(U,V) = \frac{Cov(U,V)}{\sqrt{Var(U)}\sqrt{Var(V)}} = \frac{\boldsymbol{a'\Sigma_{12}b}}{\sqrt{\boldsymbol{a'\Sigma_{11}a}}\sqrt{\boldsymbol{b'\Sigma_{22}b}}} \quad (2.4)$$

Ορισμός 2.1 Οι μεταβλητές $U_1, U_2, \dots, U_p$ και $V_1, V_2, \dots, V_q$ ορίζονται ως κανονικές μεταβλητές (canonical variables), ενώ οι αριθμοί $\rho_i, 1 \geq \rho_1 \geq \rho_2 \geq \cdots \geq \rho_p \geq 0$, ορίζονται ως κανονικές συσχετίσεις (canonical correlations). Ας ορίσουμε ότι το πρώτο ζεύγος κανονικών μεταβλητών, το οποίο μεγιστοποιεί τον συντελεστή συσχέτισης της σχέσης (2.4) είναι το ζεύγος των γραμμικών συνδυασμών μοναδιαίας διασποράς $U_1 = \boldsymbol{a_1'X_1}$ και $V_1 = \boldsymbol{b_1'X_2}$.

Το δεύτερο ζεύγος κανονικών μεταβλητών, το οποίο μεγιστοποιεί τον συντελεστή συσχέτισης της σχέσης (2.4) και είναι ασυσχέτιστο με το πρώτο ζεύγος είναι το ζεύγος των γραμμικών συνδυασμών μοναδιαίας διασποράς $U_2 = \boldsymbol{a_2'X_1}$ και $V_2 = \boldsymbol{b_2'X_2}$.

Το k-οστό ζεύγος κανονικών μεταβλητών, το οποίο μεγιστοποιεί τον συντελεστή συσχέτισης της σχέσης (2.4) και είναι ασυσχέτιστο με τα προηγούμενα $k-1$ ζεύγη κανονικών μεταβλητών είναι το ζεύγος των γραμμικών συνδυασμών μοναδιαίας διασποράς $U_k = \boldsymbol{a_k'X_1}$ και $V_k = \boldsymbol{b_k'X_2}$.

Ας υποθέσουμε ότι $p \leq q$ και έστω τα τυχαία διανύσματα $\boldsymbol{X_1}$, διάστασης $p \times 1$ και $\boldsymbol{X_2}$, διάστασης $q \times 1$ με $Cov(\boldsymbol{X_1}) = \boldsymbol{\Sigma_{11}}, Cov(\boldsymbol{X_2}) = \boldsymbol{\Sigma_{22}}$ και $Cov(\boldsymbol{X_1}, \boldsymbol{X_2}) = \boldsymbol{\Sigma_{12}}$. Με την χρήση των διανυσμάτων $\boldsymbol{a}$ και $\boldsymbol{b}$ σχηματίζουμε τους γραμμικούς συνδυασμούς $U = \boldsymbol{a'X_1}$ και $V = \boldsymbol{b'X_2}$. Τότε η μεγιστοποίηση του συντελεστή συσχέτισης:

$$max_{\boldsymbol{a},\boldsymbol{b}} Corr(U,V) = \rho_1^*$$

επιτυγχάνεται από τον γραμμικό συνδυασμό:

$$U_1 = \boldsymbol{a_1'X_1} \text{ και } V_1 = \boldsymbol{b_1'X_2},$$

όπου $\boldsymbol{\alpha_1' = e_1'\Sigma_{11}^{-1/2}}$ και $\boldsymbol{b_1' = f_1'\Sigma_{22}^{-1/2}}$.

Αντίστοιχα, το k-οστό ζεύγος κανονικών μεταβλητών, $k = 2,3, \dots, p$,

$$U_k = \boldsymbol{a_k'X_2} \text{ και } V_k = \boldsymbol{b_k'X_2},$$

όπου

$$\boldsymbol{\alpha_k' = e_k'\Sigma_{11}^{-1/2}} \text{ και } \boldsymbol{b_k' = f_k'\Sigma_{22}^{-1/2}}$$

μεγιστοποιεί τον συντελεστή συσχέτισης:

$$Corr(U_k, V_k) = \rho_k^*$$

μεταξύ εκείνων των γραμμικών συνδυασμών που είναι ασυσχέτιστες με τις προηγούμενες 1,2,…,k κανονικές μεταβλητές. Οι αριθμοί $\rho_1^{*^2}, \rho_2^{*^2}, \dots, \rho_k^{*^2}$ αποτελούν τις ιδιοτιμές του πίνακα $\boldsymbol{\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}}$, ενώ τα $\boldsymbol{e_1, e_2, \dots, e_p}$, είναι τα αντίστοιχα ιδιοδιανύσματα. Επίσης, οι αριθμοί $\rho_1^{*^2}, \rho_2^{*^2}, \dots, \rho_k^{*^2}$ αποτελούν τις ιδιοτιμές του πίνακα $\boldsymbol{\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}}$, με αντίστοιχα ιδιοδιανύσματα τα $\boldsymbol{f_1, f_2, \dots, f_p}$.

Οι κανονικές μεταβλητές έχουν τις εξής ιδιότητες:

$$Var(U_k) = Var(V_k) = 1$$

$$Cov(U_i, U_j) = Corr(U_i, U_j) = 0, \qquad i \neq j$$

$$Cov(V_i, V_j) = Corr(V_i, V_j) = 0, \qquad i \neq j$$

$$Cov(U_i, V_j) = Corr(U_i, V_j) = 0, \qquad i \neq j$$

για $i, j = 1,2, \dots p$.

Επίσης, γνωρίζουμε ότι οι μη μηδενικές ιδιοτιμές ενός πίνακα $\boldsymbol{AB}$ είναι οι ίδιες με του πίνακα $\boldsymbol{BA}$, αν και μόνο αν οι πίνακες $\boldsymbol{AB}$ και $\boldsymbol{BA}$ είναι τετραγωνικοί, δεν ισχύει όμως το ίδιο και για τα ιδιοδιανύσματα των $\boldsymbol{AB}$ και $\boldsymbol{BA}$. Εάν θεωρήσουμε ότι $\boldsymbol{A = \Sigma_{11}^{-1}\Sigma_{12}}$ και $\boldsymbol{B = \Sigma_{22}^{-1}\Sigma_{21}}$, τότε τα

$\rho_1^{*^2}, \rho_2^{*^2}, \ldots, \rho_k^{*^2}$ μπορούν να υπολογιστούν είτε από τον πίνακα $\boldsymbol{AB} = \boldsymbol{\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}$ είτε από τον πίνακα $\boldsymbol{BA} = \boldsymbol{\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}}$. Έτσι οι ιδιοτιμές μπορούν να προκύψουν και από καθεμία από τις παρακάτω χαρακτηριστικές εξισώσεις:

$$\left|\boldsymbol{\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}} - \lambda^2\boldsymbol{I}\right| = 0 \qquad (2.5)$$

$$\left|\boldsymbol{\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}} - \lambda^2\boldsymbol{I}\right| = 0 \qquad (2.6)$$

Τα διανύσματα των συντελεστών $\boldsymbol{a_i}$ και $\boldsymbol{b_i}$ των κανονικών μεταβλητών $U_i = \boldsymbol{a_i'X_1}$ και $V_i = \boldsymbol{b_i'X_2}$ αποτελούν τα ιδιοδιανύσματα των δύο ίδιων παρακάτω πινάκων:

$$\left(\boldsymbol{\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}} - \lambda^2\boldsymbol{I}\right)\boldsymbol{a} = \boldsymbol{0} \qquad (2.7)$$

$$\left(\boldsymbol{\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}} - \lambda^2\boldsymbol{I}\right)\boldsymbol{b} = \boldsymbol{0} \qquad (2.8)$$

Έτσι, οι δύο πίνακες $\boldsymbol{\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}$ και $\boldsymbol{\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}}$ έχουν τις ίδιες μη μηδενικές ιδιοτιμές, όπως υποδεικνύεται στις σχέσεις (2.5) και (2.6), αλλά διαφορετικά ιδιοδιανύσματα, όπως φαίνεται στις σχέσεις (2.7) και (2.8). Εφόσον, το $\boldsymbol{X_1}$ είναι ένα διάνυσμα διάστασης $p \times 1$ και το $\boldsymbol{X_2}$ ένα διάνυσμα διάστασης $q \times 1$, τότε και τα $\boldsymbol{a_i}$ και $\boldsymbol{b_i}$ είναι διάστασης $p \times 1$ και $q \times 1$ αντίστοιχα. Αυτό μπορεί επίσης να φανεί στα μεγέθη των πινάκων των σχέσεων (2.7) και (2.8), όπου ο πίνακας $\boldsymbol{\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}$ είναι διάστασης $p \times p$ και ο πίνακας $\boldsymbol{\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}}$ είναι διάστασης $q \times q$. Εφόσον το p δεν είναι τυπικά ίσο με το q, ο μεγαλύτερος σε μέγεθος πίνακας θα είναι μη αντιστρέψιμος, και ο μικρότερος θα είναι αντιστρέψιμος. Όταν $p < q$, η τάξη του πίνακα $\boldsymbol{\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}$ είναι p, επειδή ο πίνακας $\boldsymbol{\Sigma_{11}^{-1}}$ έχει τάξη q και ο πίνακας $\boldsymbol{\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}}$ έχει τάξη p. Στην περίπτωση αυτή, έχουμε p μη μηδενικές ιδιοτιμές και $q - p$ μηδενικές ιδιοτιμές. Γενικά, υπάρχουν $k = \min(p, q)$ τιμές του τετραγώνου της κανονικής συσχέτισης $\rho_i^{*^2}$ με k αντίστοιχα ζεύγη των κανονικών μεταβλητών $U_i = \boldsymbol{a_i'X_1}$και $V_i = \boldsymbol{b_i'X_2}$. Για παράδειγμα, εάν $p = 3$ και $q = 7$, θα υπάρχουν τρεις κανονικές συσχετίσεις, ρ_1^*, ρ_2^* και ρ_3^*.

Για κάθε i, το ρ_i^* είναι η δειγματική συσχέτιση μεταξύ των U_i και V_i, δηλαδή $\rho_i^* = \rho_{U_i,V_i}$. Όπως, ήδη αναφέραμε τα $U_1, U_2, \ldots, U_k$ είναι ασυσχέτιστα και επίσης δεν είναι ορθογώνια γιατί τα $\boldsymbol{a_1}, \boldsymbol{a_2}, \ldots, \boldsymbol{a_k}$ είναι ιδιοδιανύσματα του πίνακα $\boldsymbol{\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}$, ο οποίος είναι μη συμμετρικός.

Ομοίως τα $V_i, i = 1,2, \dots, k$ είναι ασυσχέτιστα και κάθε U_i είναι ασυσχέτιστο με όλα τα $V_j, j \neq i$ εκτός φυσικά από το U_i .

Οι κανονικές συσχετίσεις μπορούν επίσης να υπολογιστούν από τον διαμερισμένο πίνακα συσχέτισης των $\boldsymbol{X_1}$ και $\boldsymbol{X_2}$, ο οποίος δίνεται παρακάτω:

$$\boldsymbol{\rho} = \begin{pmatrix} \boldsymbol{\rho_{11}} & \boldsymbol{\rho_{12}} \\ \boldsymbol{\rho_{21}} & \boldsymbol{\rho_{22}} \end{pmatrix},$$

όπου $\boldsymbol{\rho_{11}}$ είναι ο $p \times p$ δειγματικός πίνακας συσχέτισης των $\boldsymbol{X_1}$, $\boldsymbol{\rho_{12}}$ είναι ο $p \times q$ πίνακας των δειγματικών συσχετίσεων μεταξύ των $\boldsymbol{X_1}$ και $\boldsymbol{X_2}$, και $\boldsymbol{\rho_{22}}$ είναι ο $q \times q$ δειγματικός πίνακας συσχέτισης των $\boldsymbol{X_2}$. Ο πίνακας $\boldsymbol{\rho_{11}^{-1}\rho_{12}\rho_{22}^{-1}\rho_{21}}$ είναι ανάλογος του $\boldsymbol{\rho^2} = \rho_{12}'\boldsymbol{\rho_{22}^{-1}}\rho_{12}$ στην μονοδιάστατη περίπτωση. Οι αντίστοιχες χαρακτηριστικές εξισώσεις των σχέσεων (2.5) και (2.6) με ιδιοτιμές τις $\rho_1^{*^2}, \rho_2^{*^2}, \dots, \rho_k^{*^2}$, δηλαδή $\rho_i^* = \sqrt{\lambda_i}$, είναι

$$\left|\boldsymbol{\rho_{11}^{-1}\rho_{12}\rho_{22}^{-1}\rho_{21}} - \lambda^2\boldsymbol{I}\right| = 0, \qquad (2.9)$$

$$\left|\boldsymbol{\rho_{22}^{-1}\rho_{21}\rho_{11}^{-1}\rho_{12}} - \lambda^2\boldsymbol{I}\right| = 0 \qquad (2.10)$$

Εάν χρησιμοποιήσουμε τον διαμερισμένο πίνακα συσχέτισης αντί για τον πίνακα διακύμανσης των σχέσεων (2.7) και (2.8) θα εξασφαλίσουμε τις ίδιες ιδιοτιμές, αλλά διαφορετικά ιδιοδιανύσματα:

$$\left(\boldsymbol{\rho_{11}^{-1}\rho_{12}\rho_{22}^{-1}\rho_{21}} - \lambda^2\boldsymbol{I}\right)\boldsymbol{c} = \boldsymbol{0} \qquad (2.11)$$

$$\left(\boldsymbol{\rho_{22}^{-1}\rho_{21}\rho_{11}^{-1}\rho_{12}} - \lambda^2\boldsymbol{I}\right)\boldsymbol{d} = \boldsymbol{0} \qquad (2.12)$$

Η σχέση μεταξύ των ιδιοδιανυσμάτων $\boldsymbol{c}$ και $\boldsymbol{d}$ στις σχέσεις (2.11) και (2.12) και των ιδιοδιανυσμάτων $\boldsymbol{a}$ και $\boldsymbol{b}$ στις σχέσεις (2.7) και (2.8) είναι:

$$\boldsymbol{c} = \boldsymbol{D_x a} \text{ και } \boldsymbol{d} = \boldsymbol{D_y b},$$

όπου $\boldsymbol{D_x} = diag(s_{11}, s_{12}, \dots, s_{1p})$ και $\boldsymbol{D_y} = diag(s_{21}, s_{22}, \dots, s_{2q})$. Τα ιδιοδιανύσματα $\boldsymbol{c}$ και $\boldsymbol{d}$ είναι τα κανονικοποιημένα διανύσματα συντελεστών.

Πιο αναλυτικά, εάν οι αρχικές μεταβλητές είναι κανονικοποιημένες με $\boldsymbol{Z_1} = (Z_{11,}Z_{12}, \dots, Z_{1p})'$ και $\boldsymbol{Z_2} = (Z_{21,}Z_{22}, \dots, Z_{2q})'$ τότε οι κανονικές μεταβλητές θα έχουν τη μορφή:

$$U_k = \boldsymbol{a_k'Z_1} = \boldsymbol{e_k'\rho_{11}^{-1/2}Z_1}$$

$$V_k = \boldsymbol{b_k'Z_1} = \boldsymbol{f_k'\rho_{22}^{-1/2}Z_2}$$

Εδώ, έχουμε ότι $Cov(\boldsymbol{Z_1}) = \boldsymbol{\rho_{11}}$, $Cov(\boldsymbol{Z_2}) = \boldsymbol{\rho_{22}}$, $Cov(\boldsymbol{Z_1}, \boldsymbol{Z_2}) = \boldsymbol{\rho_{12}} = \boldsymbol{\rho_{21}'}$ και τα $\boldsymbol{e_k}, \boldsymbol{f_k}$ είναι τα ιδιοδιανύσματα των πινάκων $\boldsymbol{\rho_{11}^{-1/2}\rho_{12}\rho_{22}^{-1}\rho_{21}\rho_{11}^{-1/2}}$ και $\boldsymbol{\rho_{22}^{-1/2}\rho_{21}\rho_{11}^{-1}\rho_{12}\rho_{22}^{-1/2}}$ αντίστοιχα. Οι κανονικές συσχετίσεις ικανοποιούν, ρ_k^*, ικανοποιούν τη σχέση:

$$Corr(U_k, V_k) = \rho_k^*, \quad k = 1,2, \dots, p$$

όπου $\rho_1^{*^2} \geq \rho_2^{*^2} \geq \cdots \geq \rho_p^{*^2}$ είναι οι μη μηδενικές ιδιοτιμές του πίνακα $\boldsymbol{\rho_{11}^{-1/2}\rho_{12}\rho_{22}^{-1}\rho_{21}\rho_{11}^{-1/2}}$, ή ισοδύναμα οι μεγαλύτερες ιδιοτιμές του πίνακα $\boldsymbol{\rho_{22}^{-1/2}\rho_{21}\rho_{11}^{-1}\rho_{12}\rho_{22}^{-1/2}}$.

Παρατηρούμε ότι:

$$a_k'(X_1 - \mu_1) = a_{\kappa 1}(X_{11} - \mu_{11}) + a_{\kappa 2}(X_{12} - \mu_{12}) + \cdots + a_{\kappa p}(X_{1p} - \mu_{1p}) =$$
$$a_{\kappa 1}\sqrt{\sigma_{11}}\frac{(X_{11}-\mu_{11})}{\sqrt{\sigma_{11}}} + a_{\kappa 2}\sqrt{\sigma_{22}}\frac{(X_{12}-\mu_{12})}{\sqrt{\sigma_{22}}} + \cdots + a_{\kappa p}\sqrt{\sigma_{pp}}\frac{(X_{1p}-\mu_{1p})}{\sqrt{\sigma_{pp}}},$$

όπου $Var(X_{1i}) = \sigma_{ii}, i = 1,2, \dots, p$. Επιπλέον, οι κανονικοί συντελεστές για τις κανονικοποιημένες μεταβλητές $Z_{1i} = (X_{1i} - \mu_{1i})/\sqrt{\sigma_{ii}}$ σχετίζονται με τους κανονικούς συντελεστές των αρχικών μεταβλητών X_{1i}. Συγκεκριμένα, εάν $\boldsymbol{a_k'}$ είναι το διάνυσμα του συντελεστή της k-οστής κανονικής μεταβλητής U_k, τότε $\boldsymbol{a_k'V_{11}^{1/2}}$ είναι το διάνυσμα του συντελεστή της k-οστής κανονικής μεταβλητής που προκύπτει από τις κανονικοποιημένες μεταβλητές $\boldsymbol{Z_1}$. Εδώ, $\boldsymbol{V_{11}^{1/2}}$ είναι ο διαγώνιος πίνακας με i-οστό διαγώνιο στοιχείο $\sqrt{\sigma_{1i}}$. Ομοίως, εάν $\boldsymbol{b_k'}$ είναι το διάνυσμα του συντελεστή της k-οστής κανονικής μεταβλητής V_k, τότε $\boldsymbol{a_k'V_{22}^{1/2}}$ είναι το διάνυσμα του συντελεστή της k-οστής κανονικής μεταβλητής που προκύπτει από τις κανονικοποιημένες μεταβλητές $\boldsymbol{Z_2}$. Εδώ, $\boldsymbol{V_{11}^{1/2}}$ θα είναι είναι ο διαγώνιος πίνακας με i-οστό διαγώνιο στοιχείο $\sqrt{\sigma_{2i}} = \sqrt{Var(X_{2i})}$.

2.4. Παραγοντοποίηση πινάκων

Η παραγοντοποίηση ενός πίνακα είναι ένας τρόπος αναγωγής σε δευτερεύοντες πίνακες οι οποίοι αποτελούν συστατικά μέρη του αρχικού. Είναι μια προσέγγιση που μπορεί να απλοποιήσει πολύπλοκες πράξεις που χρειάζεται να εκτελεστούν σε έναν πίνακα καθώς η εκτέλεση τους στους νέους πίνακες είναι απλούστερη σε σχέση με τον αρχικό πίνακα. Ένα απλό παράδειγμα είναι η παραγοντοποίηση του αριθμού 10 σε 2×5. Όπως και στον υπολογισμό παραγόντων για έναν αριθμό, έτσι και για την παραγοντοποίηση πινάκων υπάρχουν πολλοί τρόποι για να επιτευχθεί η παραπάνω διαδικασία. Παρακάτω θα αναφερθούν συνοπτικά τρεις διαφορετικές προσεγγίσεις παραγοντοποίησης πινάκων.

<u>LU παραγοντοποίηση πίνακα</u>

Η παραγοντοποίηση LU χρησιμοποιείται όταν πρόκειται για αρχικό πίνακα ο οποίος είναι τετραγωνικός και τον οποίο ανασυνθέτει στους πίνακες L και U. Τετραγωνικός πίνακας θεωρείται αυτός που έχει τον ίδιο αριθμό γραμμών και στηλών. Η παρακάτω σχέση περιγράφει την παραγοντοποίηση του πίνακα A

$$A = LU$$

όπου L είναι ένας κάτω τριγωνικός πίνακας και U άνω τριγωνικός.

Κάτω τριγωνικός είναι ένας πίνακας που αποτελείται από μηδενικά στοιχεία από τη διαγώνιο και επάνω όπως ο πίνακας στην (1A), ενώ αντίστροφα άνω τριγωνικός είναι αυτός που απεικονίζεται στην (3B).

$$\begin{bmatrix} \alpha_{11} & 0 & 0 & 0 \\ \alpha_{21} & \alpha_{22} & 0 & 0 \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & 0 \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{bmatrix} \quad \text{(3A)} \qquad \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ 0 & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ 0 & 0 & \alpha_{33} & \alpha_{34} \\ 0 & 0 & 0 & \alpha_{44} \end{bmatrix} \quad \text{(3B)}$$

Η παραγοντοποίηση LU γίνεται μέσω μιας επαναληπτικής αριθμητικής διαδικασίας η οποία αποτυγχάνει όταν ο αρχικός πίνακας δεν είναι δυνατόν να παραγοντοποιηθεί.

Παραγοντοποιήσιμος είναι ένας πίνακας για τον οποίο μπορεί να υπολογιστεί ο αντίστροφος του.

Η LU παραγοντοποίηση χρησιμοποιείται συχνά για να απλοποιήσει την επίλυση συστημάτων γραμμικών εξισώσεων, όπως η εύρεση των συντελεστών σε μια γραμμική παλινδρόμηση, καθώς και για τον υπολογισμό της ορίζουσας και του αντίστροφου ενός πίνακα.

QR παραγοντοποίηση πίνακα

Η παραγοντοποίηση QR χρησιμοποιείται και για πίνακες οι οποίοι δεν είναι τετραγωνικοί και διασπά έναν πίνακα A $(m \times n)$ στους πίνακες Q και R.

$$A = QR$$

όπου Q είναι πίνακας διάστασης $(m \times m)$ και R άνω τριγωνικός πίνακας με διαστάσεις $(m \times n)$.

Η παραγοντοποίηση QR χρησιμοποιεί και αυτή μια επαναληπτική αριθμητική μέθοδο που μπορεί να αποτύχει για εκείνους τους πίνακες που δεν μπορούν να διασπαστούν. Όπως και η αποσύνθεση LU, η αποσύνθεση QR χρησιμοποιείται για την επίλυση συστημάτων γραμμικών εξισώσεων, χωρίς όμως να περιορίζεται σε τετραγωνικούς πίνακες.

2.5. k-means clustering

Δεδομένου ενός συνόλου παρατηρήσεων $x_1, x_2, \dots, x_n$, όπου κάθε παρατήρηση είναι ένα πραγματικό διάνυσμα d διάστασης, η ομαδοποίηση k-means στοχεύει στη διαίρεση των n παρατηρήσεων σε k $(\leq n)$ σύνολα $\{S_1, S_2, \dots, S_n\}$ έτσι ώστε να ελαχιστοποιηθεί το άθροισμα των τετραγώνων εντός της ομάδας (cluster) ή αλλιώς η διακύμανση. Από μαθηματικής άποψης, στόχος είναι να βρεθούν:

$$\underbrace{argmin}_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2 = \underbrace{argmin}_{S} \sum_{i=1}^{k} |S_i| Var S_i, \quad (1)$$

όπου μ_i είναι ο μέσος όρος των σημείων του S_i. Αυτό ισοδυναμεί με ελαχιστοποίηση του τετραγώνου της απόκλισης των σημείων του S_i κατά ζεύγη:

$$\underbrace{argmin}_{S}\sum_{i=1}^{k}\frac{1}{2|S_i|}\sum_{x,y\in S_i}\|x-y\|^2$$

Η παραπάνω ισοδυναμία εξάγεται με χρήση της ταυτότητας (3)

$$\sum_{x\in S_i}\|x-\mu_i\|^2=\sum_{x\neq y\in S_i}(x-\mu_i)^T(\mu_i-y) \qquad (3)$$

στη σχέση (1). Επειδή η συνολική διακύμανση είναι σταθερή, υπάρχει ισοδυναμία με τη μεγιστοποίηση του αθροίσματος των τετραγωνικών αποκλίσεων μεταξύ σημείων σε διαφορετικές ομάδες (clusters).

Ο πιο συνηθισμένος αλγόριθμος χρησιμοποιεί μια επαναληπτική τεχνική βελτίωσης. Με δεδομένο ένα αρχικό σύνολο k στο πλήθος μέσων όρων, ο αλγόριθμος προχωρά εναλλάσσοντας δύο βήματα:

Βήμα ανάθεσης: Αντιστοιχεί κάθε παρατήρηση στο cluster με τον πλησιέστερο μέσο όρο: αυτόν με τον οποίο έχει τη μικρότερη ευκλείδεια απόσταση υψωμένη στο τετράγωνο.

Βήμα ενημέρωσης: Επανυπολογισμός μέσων όρων βάσει των παρατηρήσεων που έχουν ταξινομηθεί σε κάθε cluster.

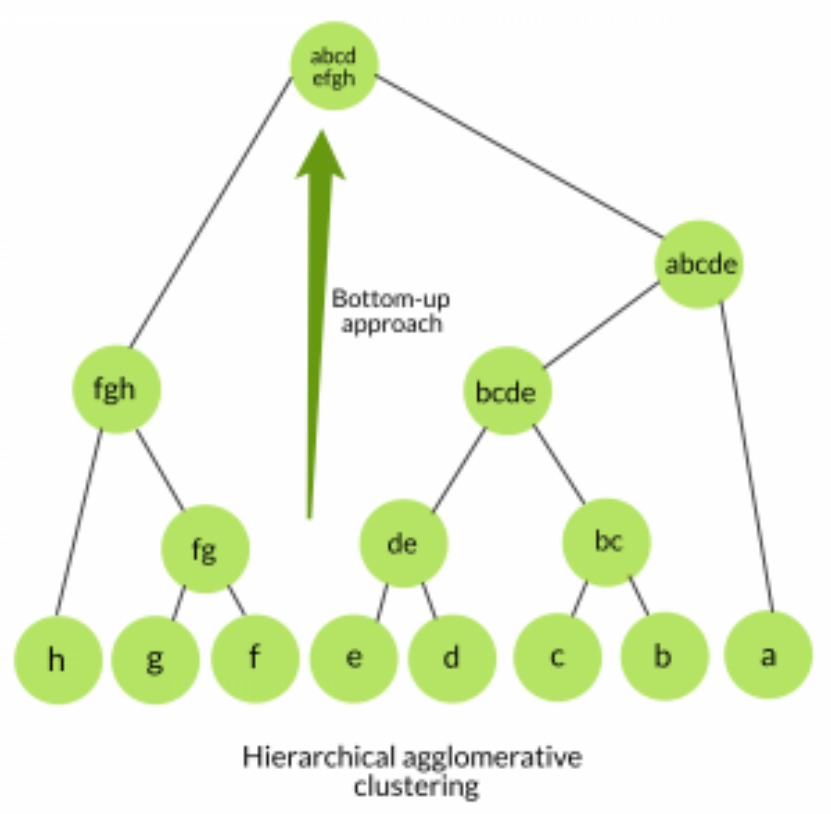
Ο αλγόριθμος συγκλίνει όταν για το βήμα της ανάθεσης δεν παρατηρούνται αλλαγές.

Οι αρχικές τιμές στους μέσους όρους μπορούν να εκχωρηθούν με διάφορους τρόπους. Ο απλούστερος είναι να δοθούν με τυχαίο τρόπο. Ακόμα μπορούν να χρησιμοποιηθούν διαφορετικές μετρικές για τον υπολογισμό των αποστάσεων, εκτός της ευκλείδειας, όπως οι Minkowski και Manhattan. Είναι σημαντικό να σημειωθεί πως η μετρική επηρεάζει σε μεγάλο βαθμό τα παραγόμενα αποτελέσματα και για αυτό το λόγο η επιλογή θα πρέπει να γίνεται βάσει των δεδομένων.

2.6. Ιεραρχική ομαδοποίηση (hierarchical clustering)

Η ιεραρχική ομαδοποίηση μπορεί να χωριστεί σε δύο κύριους τύπους: τη συσσωρευτική (agglomerative) και τη διαιρετική (divisive).

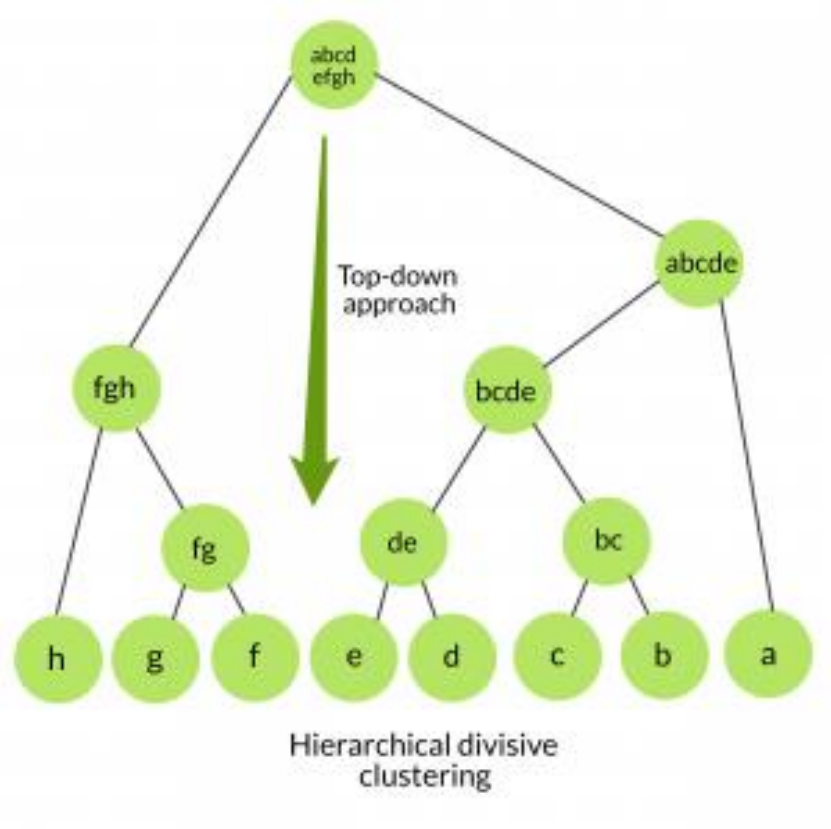
Agglomerative clustering: λειτουργεί από κάτω προς τα πάνω. Δηλαδή, κάθε αντικείμενο θεωρείται αρχικά ως cluster ενός στοιχείου. Σε κάθε βήμα του αλγορίθμου, οι δύο ομάδες που έχουν τη μεγαλύτερη ομοιότητα συνενώνονται σε ένα νέο μεγαλύτερο cluster, δημιουργώντας έναν κόμβο. Αυτή η διαδικασία επαναλαμβάνεται έως ότου όλα τα σημεία καταλήξουν να ανήκουν σε ένα μοναδικό cluster. Το αποτέλεσμα είναι ένα δέντρο που μπορεί να σχεδιαστεί ως δενδρόγραμμα (Εικόνα 2).



Εικόνα 2. Αναπαράσταση Ιεραρχικής ομαδοποίησης με συσσωρευτικό τρόπο.

Διαχωριστική ιεραρχική ομαδοποίηση: λειτουργεί από πάνω προς τα κάτω. Ο αλγόριθμος είναι μια αντίστροφη διαδικασία αυτής που περιεγράφηκε στην προηγούμενη περίπτωση. Η

διαδικασία ξεκινά με μία μόνο ομάδα (cluster), στην οποία περιλαμβάνονται όλα τα αντικείμενα. Σε κάθε βήμα επανάληψης, η ομάδα με τη μικρότερη ομοιογένεια χωρίζεται σε δύο νέες. Η διαδικασία επαναλαμβάνεται έως ότου όλα τα αντικείμενα να τοποθετηθούν το καθένα στο δικό του cluster (Εικόνα 3).



Εικόνα 3. Αναπαράσταση Ιεραρχικής ομαδοποίησης με διαιρετικό τρόπο.

Και στους δύο αυτούς τύπους ιεραρχικής ομαδοποίησης χρησιμοποιούνται διαφορετικές μετρικές ανάλογα με τον τύπο των δεδομένων. Αποδεκτές είναι μεταξύ άλλων η ευκλείδεια, η Manhattan, η μέγιστη απόσταση, ενώ σε μη αριθμητικά δεδομένα γίνεται χρήση της απόστασης Hamming. Επιπλέον, υπάρχουν ποικίλοι τρόποι για να μετρηθεί η ανομοιομορφία μεταξύ των ομάδων στοιχείων που σχηματίζονται σε κάθε βήμα. Οι πιο συνηθισμένες μέθοδοι είναι:

Ομαδοποίηση μέγιστης σύνδεσης (maximum linkage): Υπολογίζει την ανομοιότητα ανάμεσα σε όλα τα ζεύγη δύο ομάδων και θεωρεί τη μεγαλύτερη τιμή αυτών ως την απόσταση μεταξύ των δύο ομάδων.

Ομαδοποίηση ελάχιστης σύνδεσης (minimum linkage): Υπολογίζει την ανομοιότητα ανάμεσα σε όλα τα ζεύγη δύο ομάδων και θεωρεί την ελάχιστη τιμή αυτών ως κριτήριο για τη σύνδεση των ομάδων.

Ομαδοποίηση μέσης συσχέτισης (average linkage): Υπολογίζει την ανομοιότητα ανάμεσα σε όλα τα ζεύγη δύο ομάδων και θεωρεί το μέσο όρο αυτών ως την απόσταση μεταξύ των δύο ομάδων.

Μέθοδος ελάχιστης διακύμανσης Ward (Ward's minimum variance): Ελαχιστοποιεί τη συνολική διακύμανση εντός μιας ομάδας στοιχείων. Σε κάθε βήμα, το ζεύγος ομάδων με την ελάχιστη απόσταση μεταξύ τους συγχωνεύονται.

3. Μέθοδοι ενσωμάτωσης

Στο παρόν κεφάλαιο παρουσιάζονται τα αποτελέσματα της εφαρμογής κάποιων εργαλείων που εφαρμόζονται για multi-omics αναλύσεις. Η δομή που ακολουθείται για κάθε εργαλείο αποτελείται από ένα περιγραφικό κομμάτι και στη συνέχεια εκτίθενται οι δυνατότητες του καθενός από αυτά καθώς και τα αποτελέσματα που μπορούν αυτά να παράγουν.

3.1.Τεχνολογίες NGS και μέθοδοι ενσωμάτωσης

Η έλευση των νέων τεχνολογιών και η ώθηση που δόθηκε στην παραγωγή δεδομένων από διαφορετικά -omics πεδία (π.χ. transcriptomics για τη μελέτη της μεταγραφής, proteomics για τη μελέτη των πρωτεϊνών, metabolomics για μεταβολίτες) έχει δώσει νέες ευκαιρίες για βιολογικές και ιατρικές ερευνητικές ανακαλύψεις. Συνήθως, κάθε χαρακτηριστικό από κάθε πεδίο (μετάγραφα, πρωτεΐνες, μεταβολίτες, κλπ) αναλύεται ανεξάρτητα μέσω μονομεταβλητών στατιστικών μεθόδων που περιλαμβάνουν τεχνικές όπως ANOVA, γραμμικά μοντέλα ή t-test. Ωστόσο, μια τέτοια ανάλυση αγνοεί τις σχέσεις μεταξύ των διαφόρων χαρακτηριστικών και μπορεί να χάσει σημαντικές βιολογικές πληροφορίες. Πράγματι, τα βιολογικά χαρακτηριστικά δρουν συντονισμένα για να διαμορφώσουν και να επηρεάσουν τα βιολογικά συστήματα και τα βιολογικά μονοπάτια. Οι πολυμεταβλητές προσεγγίσεις, όπου τα διάφορα χαρακτηριστικά μοντελοποιούνται ως σύνολο, μπορούν να παρέχουν μια πιο ολιστική και ακριβή εικόνα ενός βιολογικού συστήματος και να συμπληρώνουν τα αποτελέσματα που προκύπτουν από μονομεταβλητές μεθόδους. Για τον παραπάνω λόγο, η μελέτη που περιλαμβάνει συνδυαστική ανάλυση δεδομένων από διαφορετικά -omics πεδία (multiomics) εφαρμόζεται όλο και περισσότερο σε διαφορετικούς βιολογικούς τομείς, συμπεριλαμβανόμενης της βιολογίας του καρκίνου (Gerstung et al, 2015; Iorio et al, 2016, Cancer Genome Atlas Research Network, 2017), της ρυθμιστικής γονιδιωματικής (Chen et al, 2016) και της μικροβιολογίας (Kim et al, 2016). Οι πιο πρόσφατες τεχνολογικές εξελίξεις επέτρεψαν ακόμα τη διενέργεια multiomics αναλύσεων στο επίπεδο μεμονωμένων κυττάρων (single cell) (Angermueller et al, 2016; Colomé-Tatché & Theis, 2018). Ένας κοινός στόχος τέτοιων εφαρμογών είναι ο χαρακτηρισμός της ετερογένειας μεταξύ των δειγμάτων, όπως εκδηλώνεται σε μία ή περισσότερες από τις μορφές δεδομένων

(Ritchie et al, 2015). Η χρήση multiomics αναλύσεων είναι ιδιαίτερα χρήσιμη στις περιπτώσεις που δεν είναι εκ των προτέρων γνωστή η πηγή της μεταβλητότητας των δεδομένων καθώς υπάρχει η δυνατότητα να ανακαλυφθούν νέες σχέσεις μεταξύ των διαφορετικών -omics δεδομένων γεγονός το οποίο μπορεί να παραλειφθεί από μελέτες που βασίζονται στη μεταβλητότητα ενός μόνο τύπου δεδομένων.

3.2. mixOmics

3.2.1. Περιγραφή

Το πακέτο mixOmics προτείνει πολυμεταβλητές μεθοδολογίες που βασίζονται στην προβολή των δεδομένων σε ένα νέο πολυδιάστατο χώρο για την ανάλυση των -omics δεδομένων, καθώς αυτές οι μεθοδολογίες παρέχουν αρκετές ιδιότητες ελκυστικές για τον αναλυτή των δεδομένων. Αφενός είναι υπολογιστικά αποδοτικοί όταν πρόκειται για ανάλυση μεγάλων συνόλων δεδομένων, όπου ο αριθμός των βιολογικών χαρακτηριστικών (συνήθως χιλιάδες) είναι πολύ μεγαλύτερος από τον αριθμό των δειγμάτων (συνήθως λιγότερα από 50) και αφετέρου εφαρμόζουν μείωση των διαστάσεων προβάλλοντας τα δεδομένα σε έναν υποχώρο λιγότερων διαστάσεων της αρχικής ενώ ταυτόχρονα “αντιλαμβάνονται” και αναδεικνύουν τις μεγαλύτερες πηγές διακύμανσης από τα δεδομένα, με αποτέλεσμα την καλύτερη απεικόνιση του υπό μελέτη βιολογικού συστήματος. Τέλος, οι παραδοχές αυτών των μεθοδολογιών σχετικά με την κατανομή των δεδομένων τις καθιστούν εξαιρετικά ευέλικτες στην ανεύρεση απαντήσεων στα ερωτήματα που τίθενται σε πολυάριθμα πεδία που σχετίζονται με τη βιολογία. Οι πολυμεταβλητές μέθοδοι που περιέχονται στο πακέτο mixOmics έχουν εφαρμοστεί με επιτυχία στην ενσωμάτωση συνόλων δεδομένων που προέρχονται από διαφορετικές βιολογικές πηγές και στον εντοπισμό βιοδεικτών σε μελέτες -omics δεδομένων όπως αυτά της μεταβολομικής, της απεικόνισης του εγκεφάλου και του μικροβιώματος.

Το πακέτο mixOmics χρησιμοποιείται στο πλαίσιο της supervised ανάλυσης, όπου στόχοι είναι η ταξινόμηση ή η διάκριση ομάδων δειγμάτων, η εύρεση του υποσυνόλου βιολογικών χαρακτηριστικών για την καλύτερη ταξινόμηση των δειγμάτων και η πρόβλεψη της κατηγορίας

στην οποία εμπίπτουν νέα δείγματα. Παρόλα αυτά, περιέχει και μεθόδους unsupervised ανάλυσης όπως η ανάλυση κύριων συνιστωσών (PCA). Τα δύο κυριότερα πλαίσια ανάλυσης που προσφέρει το πακέτο mixOmics είναι το DIABLO και το MINT. Με τη χρήση του DIABLO είναι δυνατή η ενσωμάτωση διαφορετικών -omics δεδομένων που αφορούν στα ίδια δείγματα, ενώ χρησιμοποιώντας το MINT μπορούν να ενσωματωθούν και να αναλυθούν δεδομένα που προέρχονται από διαφορετικά δείγματα για τα οποία έχουν μετρηθεί τα ίδια χαρακτηριστικά. Στο πλαίσιο του MINT υπάρχει η δυνατότητα supervised και unsupervised ανάλυσης. Τα δύο αυτά πλαίσια εργασίας βασίζονται στην ανάλυση παλινδρόμησης και συγκεκριμένα στη μέθοδο μερικών ελαχίστων τετραγώνων (PLS) αλλά και σε παραλλαγές αυτής καθώς και στην ανάλυση κανονικής συσχέτισης. Το DIABLO προσπαθεί να βρει ένα σύνολο μεταβλητών από τα διαφορετικά σύνολα omics δεδομένων οι οποίες έχουν τη δυνατότητα να διαχωρίζουν ήδη γνωστές ομάδες δειγμάτων.

3.2.2. Τύποι δεδομένων

Το πακέτο mixOmics μπορεί να χειριστεί διαφορετικούς τύπους βιολογικών δεδομένων. Τα δεδομένα μπορούν να είναι δεδομένα αλληλούχισης (π.χ. RNA-seq, 16S), τα οποία είναι “έγκυρα” μετά από κατάλληλη προετοιμασία και κανονικοποίηση προκειμένου να αναχθούν σε συνεχή κλίμακα καθώς και χαρακτηριστικά που μετρούνται σε συνεχή κλίμακα (π.χ. microarray, mass spectrometry proteomics). Γενικά, τα δεδομένα που χρησιμοποιούνται ως είσοδος σε αυτό το πακέτο θα πρέπει να είναι σε συνεχή κλίμακα.

3.2.3. Εφαρμογή και αποτελέσματα

Η εφαρμογή του DIABLO έγινε σε δεδομένα 220 ασθενών οι οποίοι χωρίστηκαν σε δύο ομάδες. Η πρώτη, στην οποία βασίστηκε η εκπαίδευση του μοντέλου (training set), αποτελούμενη από 150 ασθενείς και η δεύτερη ομάδα ελέγχου (test set) με 70 ασθενείς. Οι τύποι δεδομένων που ενσωματώθηκαν είναι τα επίπεδα έκφρασης miRNA, mRNA, proteins και ο χαρακτηρισμός των ασθενών σε υποτύπους της ασθένειας. Αφού γίνει η κατάλληλη προετοιμασία στα δεδομένα (κανονικοποίηση, μετατροπή σε συνεχή κλίμακα κ.α.), δημιουργείται το μοντέλο το οποίο είναι ένα μοντέλο παλινδρόμησης μερικών ελαχίστων τετραγώνων. Τα αποτελέσματα που

παράγονται βασίζονται στην κατασκευή των συνιστωσών που σχετίζονται με κάθε διαφορετικό σύνολο δεομένων (mRNA, miRNA κ.α.) καθώς και σε ένα σύνολο διανυσμάτων φορτίων (loading vectors) τα οποία είναι οι συντελεστές κάθε μεταβλητής για κάθε συνιστώσα. Τα διανύσματα φορτίων δημιουργούνται με τέτοιο τρόπο ώστε να μεγιστοποιείται η συνδιακύμανση του γραμμικού συνδυασμού των αρχικών μεταβλητών **X** και των μεταβλητών απόκρισης **Y.**

Τα δεδομένα που ενσωματώθηκαν στο DIABLO παρατίθενται σε μορφή υποδείγματος στους παρακάτω πίνακες.

	RTN2	NDRG2	CCDC113	FAM63A	ACADS
A0FJ	4,36218330538705	7,53346145207899	3,95612417342232	4,45717045459448	2,25681701402838
A13E	1,98449231403661	7,4551937560165	5,42762306424528	5,44095735941906	4,02881328348217
A0G0	1,72732286530513	8,07996823763621	2,22730024457496	5,54348046285294	2,62985528529214
A0SX	4,36399578822303	5,79375019285628	3,54486568640499	4,73711416287221	4,2691008476309
A143	2,44756194574285	7,15899336229585	4,69125638632212	4,80872751046921	2,44213451715903
A0DA	4,77079757149914	8,74806057576936	4,30540146408101	5,30748033386192	3,23990914570011
A0B3	3,35206181777286	5,09840404813609	0,593205644792737	5,21758513821267	3,88515338602591
A0I2	1,81038178283432	3,79196548917246	2,71916935042778	4,35591911785313	4,20024883426051
A0RT	2,09446033673917	6,32797259214932	2,35793285823601	4,04166106251926	4,12681672512024
A131	4,34091154531675	4,69995003853425	3,63905565559377	4,03018749114371	3,13569699367135

Πίνακας 1. DIABLO - Υπόδειγμα συνόλου δεδομένων mRNA.

	hsa-let-7a-1	hsa-let-7a-2	hsa-let-7a-3	hsa-let-7b	hsa-let-7c
A0FJ	11,835815644832	12,8510467162388	11,9188113263646	14,801380237107	10,9356752207516
A13E	12,8979316315316	13,9008691660534	12,9127259302698	14,7155111747103	12,0318358888277
A0G0	12,3077264364518	13,2903246797872	12,3006247974952	15,0661909515156	10,9339298922897
A0SX	12,0392931397056	13,0108071298356	12,0814070379743	14,6200326276362	11,4715272075954
A143	13,3909710853051	14,3898160002792	13,4222777526341	15,3056060804917	10,1469009274522
A0DA	12,3403716402942	13,3598599485471	12,3932000195956	14,8570542547069	11,3705322854953
A0B3	12,4376972179566	13,4058889250225	12,4644795491056	15,7665717545421	11,0587753885469
A0I2	12,528040711661	13,5186287234223	12,541838588839	15,1750692072442	12,247543225551
A0RT	12,2303801926699	13,239471139603	12,2921490197409	15,0118314558262	11,6311927537217
A131	12,6698204160879	13,6546079568221	12,6828137293762	14,7433879863392	11,1681430747175

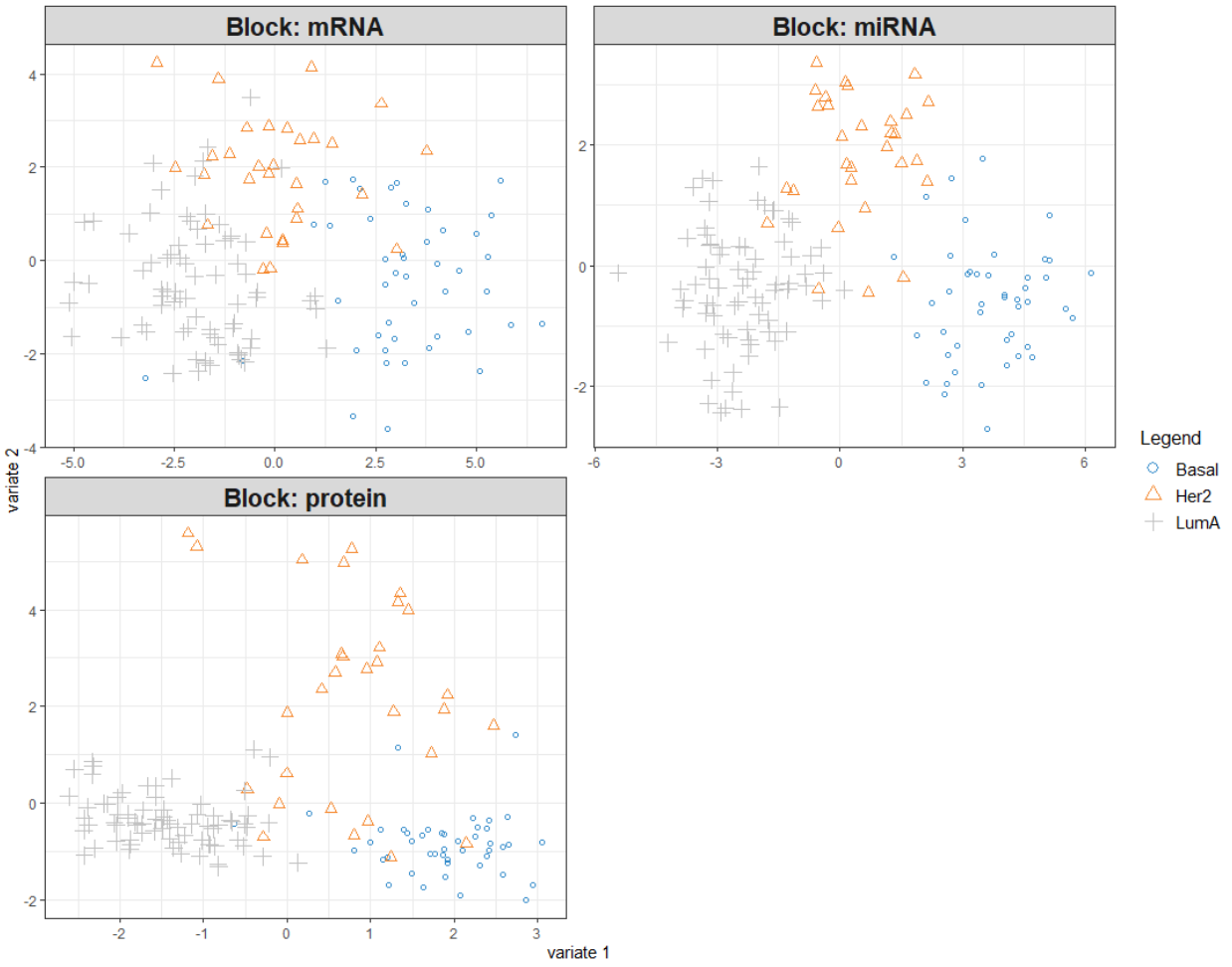
Πίνακας 2. DIABLO - Υπόδειγμα συνόλου δεδομένων miRNA.

	14-3-3_epsilon	4E-BP1	4E-BP1_pS65	4E-BP1_pT37	4E-BP1_pT70
A0FJ	0,049130778085686	0,44748623091927	-0,074321750760139	-0,381162897847251	0,026069431978311
A13E	-0,079982106355818	0,605218417826312	0,288911839658517	1,16024541957318	-0,010967170983153
A0G0	-0,03284988615576	0,894609732271096	0,891277838347467	1,40018887427333	0,265880117211022
A0SX	-0,205329492210682	-0,141322923550322	-0,018410962635847	-0,444224272119272	-0,051267718224327
A143	0,060190211021632	0,131768991570967	0,665325385153256	0,351763305169398	0,069802860336663

A0DA	0,030761714453474	0,032996799470719	0,052360640103079	0,789235974569012	0,017045552375359
A0B3	-0,107861537125951	-0,03712469104008	0,199690200772294	0,586575642836024	0,323403396755904
A0I2	0,64984396402993	-0,52148656560265	0,249978143122115	0,4634114576105	0,320128653878775
A0RT	-0,013650441460306	-0,634850632562606	-0,082575425844651	0,302828701946043	0,067672264050429
A131	0,430934243034933	1,05257162195633	0,350978693805628	0,205140253099992	0,610566788880038

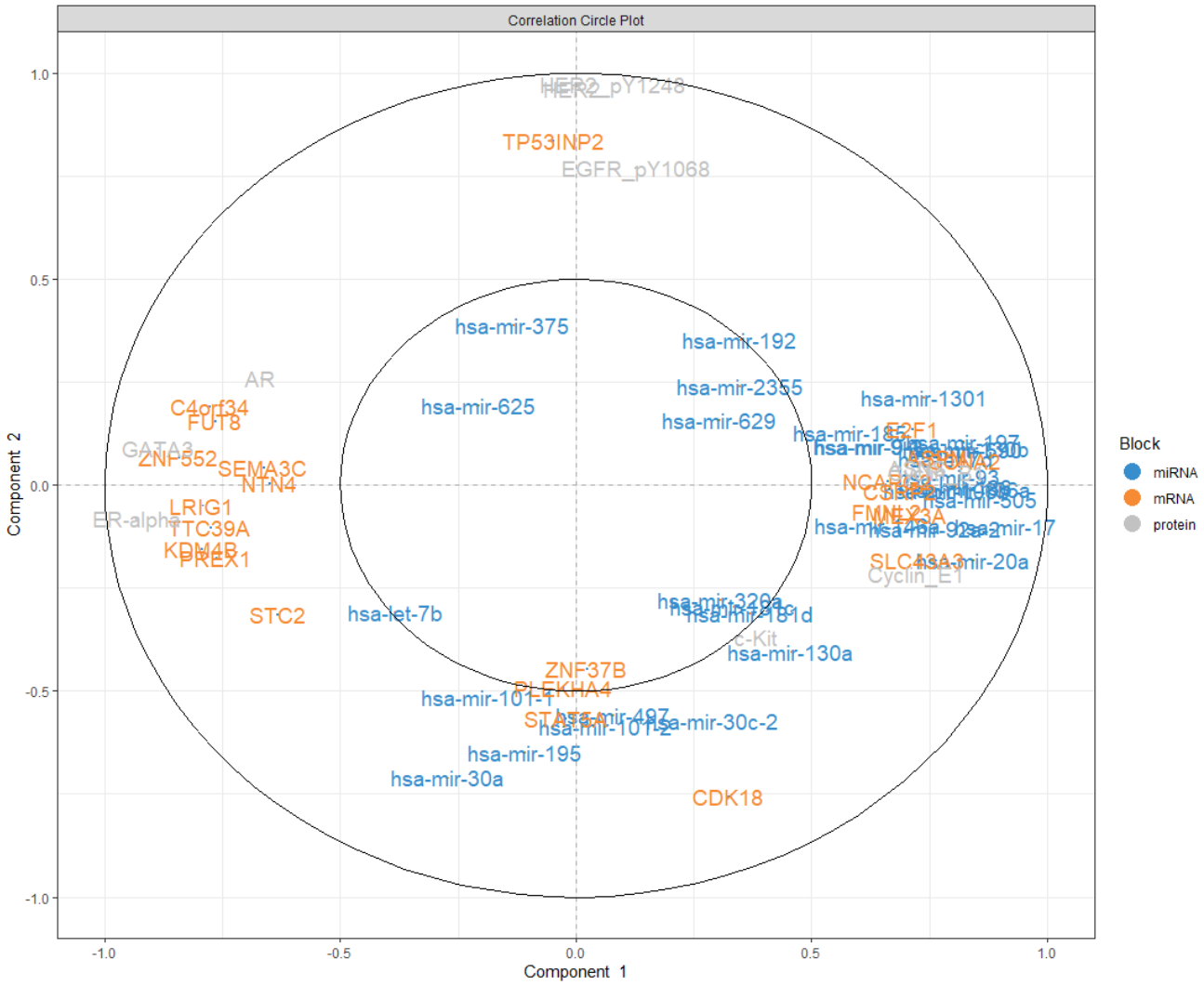
Πίνακας 3. DIABLO - Υπόδειγμα συνόλου δεδομένων protein.

Η αποτύπωση των ασθενών στις δύο συνιστώσες που δημιουργήθηκαν για κάθε διαφορετικό σύνολο δεδομένων παρουσιάζεται στην Εικόνα 4, όπου οι ασθενείς χωρίζονται σε τρεις κατηγορίες σύμφωνα με τον υποτύπο της ασθένειας (Basal, Her2, LumA). Όπως φαίνεται, η κατασκευή των συνιστωσών για κάθε σύνολο δεδομένων έχει αποδώσει σε μεγάλο βαθμό τη μεταβλητότητα που υπάρχει ανάμεσα στις διαφορετικές ομάδες ασθενών.



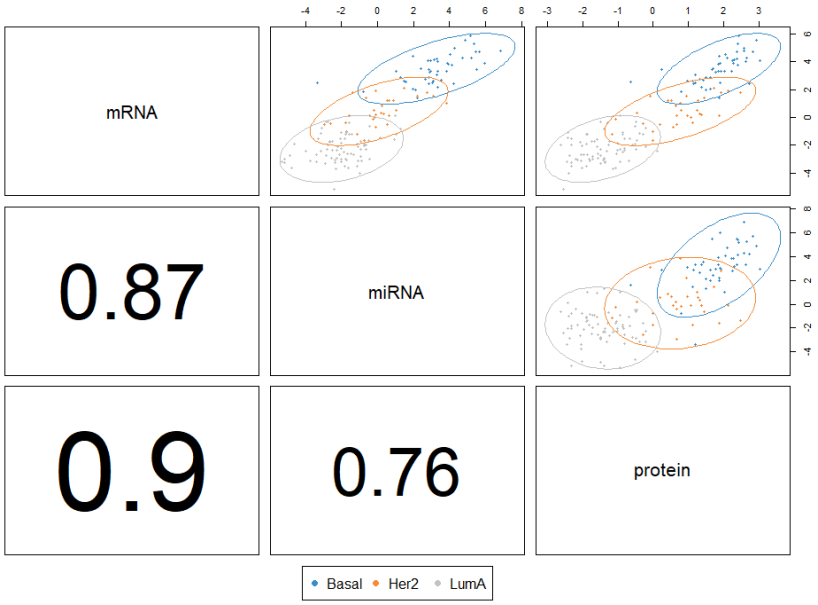
Εικόνα 4. Κατανομή των ασθενών στο χώρο των συνιστωσών για κάθε σύνολο δεδομένων ξεχωριστά. Οι ασθενείς έχουν χαρακτηριστεί σύμφωνα με τον τύπο της ασθένειας.

Στη συνέχεια, γίνεται ανάλυση της συσχέτισης μεταξύ των μεταβλητών από τα διαφορετικά σύνολα δεδομένων. Παρατηρείται συσχέτιση των μεταβλητών που προέρχονται από τα σύνολα δεδομένων miRNA και mRNA καθώς οι μεταβλητές αυτές βρίσκονται στην ίδια πλευρά ή σε αντιδιαμετρική θέση ως προς τον άξονα της πρώτης συνιστώσας.

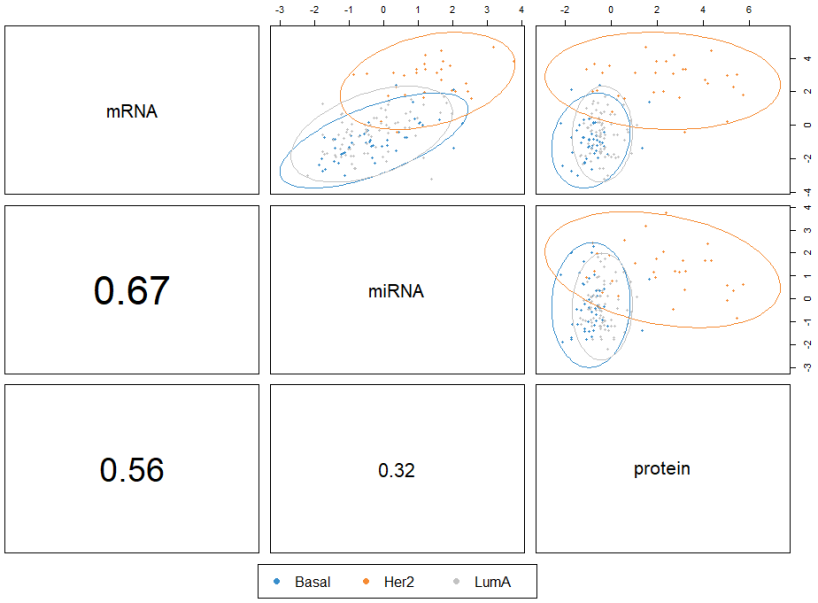


Εικόνα 5. Συσχέτιση μεταβλητών ως προς τις δύο συνιστώσες. Με διαφορετικό χρώμα απεικονίζονται μεταβλητές από διαοφρετικά σύνολα δεδομένων.

Συνολικά, η συσχέτιση μεταξύ των συνόλων δεδομένων που έχουν ενσωματωθεί στο μοντέλο απεικονίζεται στα δύο παρακάτω σχήματα όπου για την πρώτη συνιστώσα (Εικόνα 6) υπάρχει μεγάλη συσχέτιση μεταξύ των συνόλων mRNA και protein (r = 0.9) καθώς και με το σύνολο miRNA (r = 0.87). Για τη δεύτερη συνιστώσα (Εικόνα 7) δεν παρατηρούνται υψηλές συσχετίσεις μεταξύ διαφορετικών συνόλων.

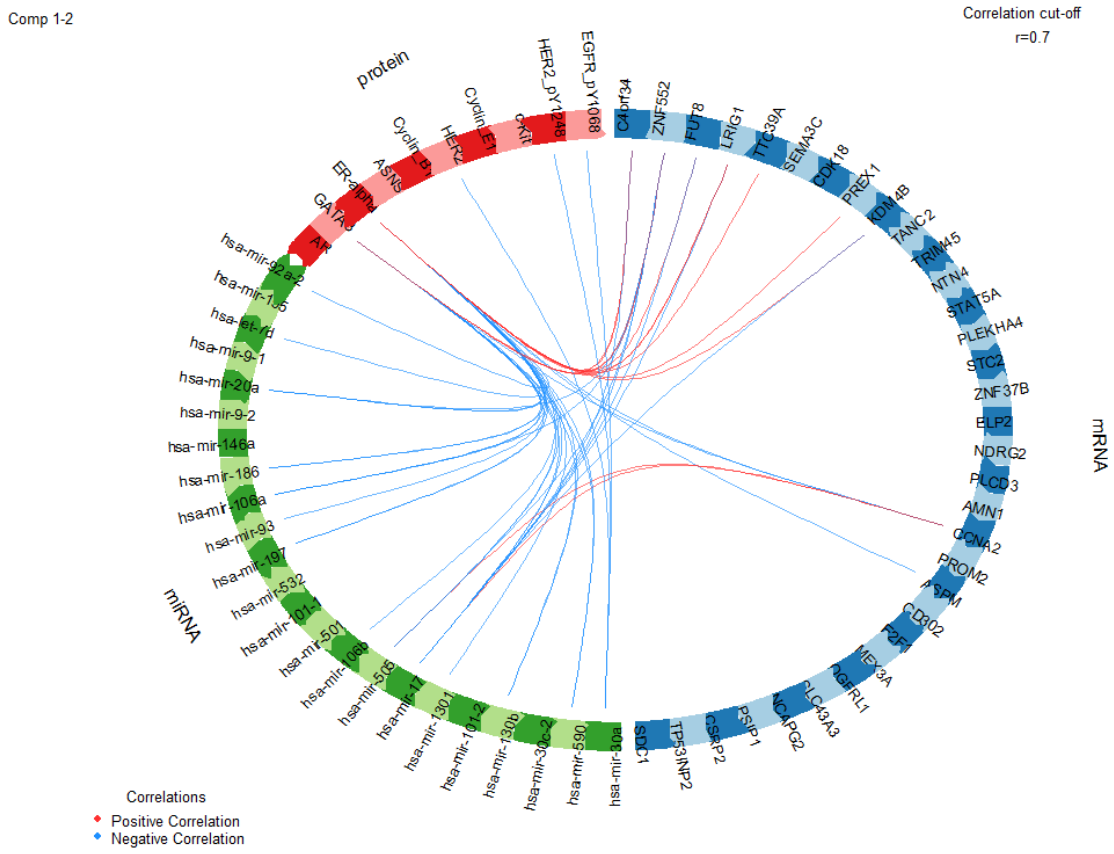


Εικόνα 6. Συνολική επισκόπηση της συσχέτισης σε επίπεδο συνιστωσών (περίπτωση πρώτης συνιστώσας). Η κλάση κάθε δείγματος παρουσιάζεται με διαφορετικό χρώμα.



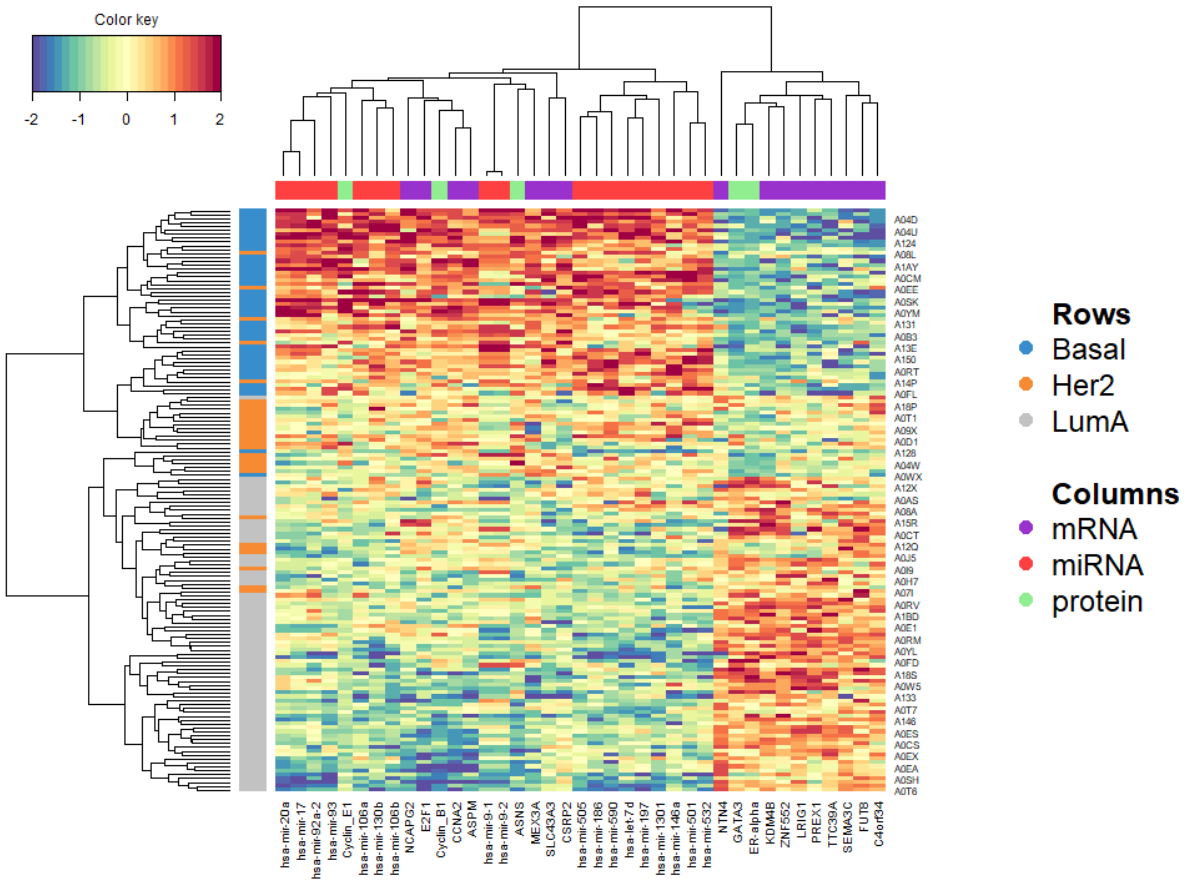
Εικόνα 7. Συνολική επισκόπηση της συσχέτισης σε επίπεδο συνιστωσών (περίπτωση δεύτερης συνιστώσας). Η κλάση κάθε δείγματος παρουσιάζεται με διαφορετικό χρώμα.

Στο επόμενο διάγραμμα αντανακλώνται οι συσχετίσεις μεταξύ μεταβλητών διαφορετικών τύπων, που αναπαρίστανται στα πλευρικά τεταρτημόρια και οι οποίες είναι μεγαλύτερες από 0.7.



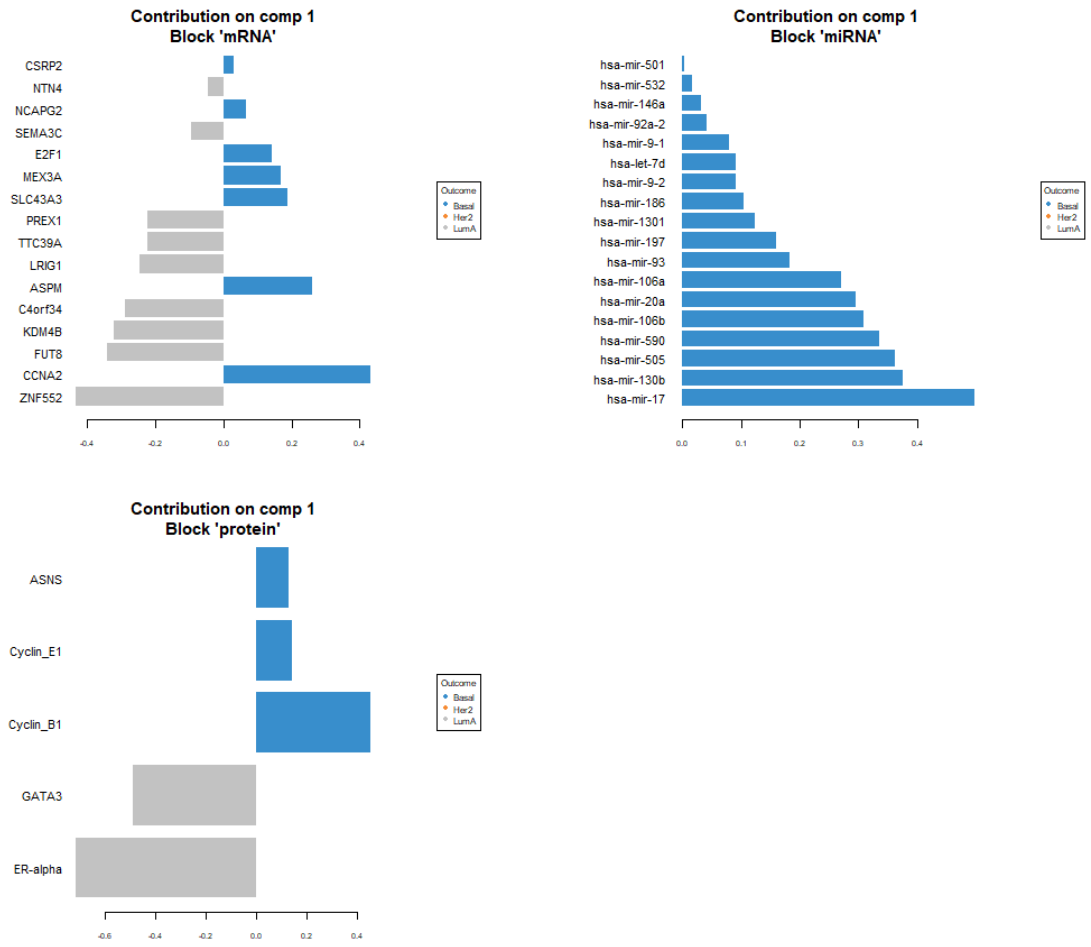
Εικόνα 8. Απεικόνιση συσχετίσεων μεταξύ διαοφρετικών ομάδων μεταβλητών. Οι θετικές συσχετίσεις απεικονίζονται με κόκκινο χρώμα, ενώ οι αρνητικές με γαλάζιο.

Η εύρεση των μεταβλητών που χαρακτηρίζουν κάθε δείγμα και κατά συνέπεια τις ομάδες δειγμάτων για κάθε τύπο της ασθένειας οδηγεί στην Εικόνα 9 όπου είναι διακριτό πως για κάθε ομάδα υπάρχει διαφοροποίηση στις μεταβλητές που προέρχονται από τα διάφορα omics σύνολα δεομένων (multi-omics molecular signature).

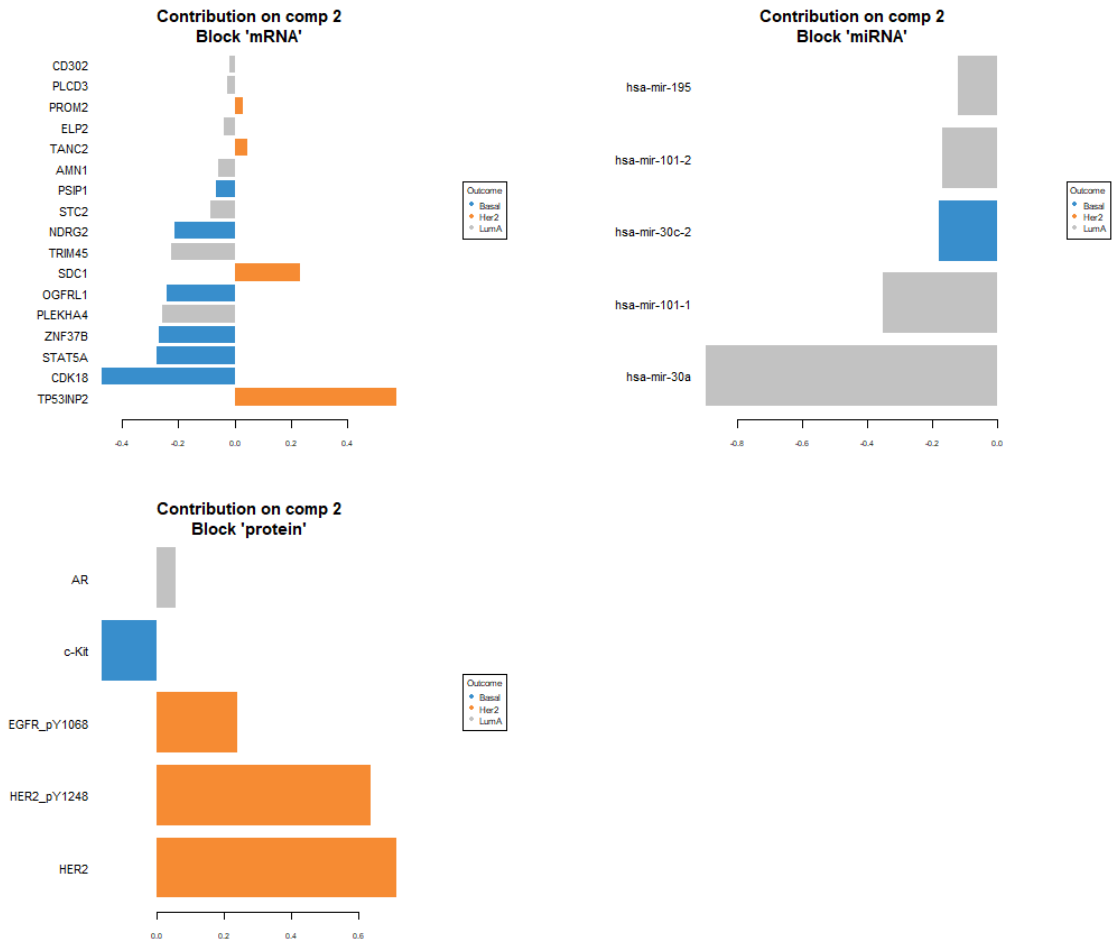


Εικόνα 9. Heatmap με ιεραρχική ομαδοποίηση (hierarchical clustering) των δειγμάτων και των μεταβλητών.

Το επόμενο βήμα της ανάλυσης είναι να βρεθούν τα φορτία των αρχικών μεταβλητών για κάθε συνιστώσα που δημιουργήθηκε. Είναι σημαντικό να αναγνωριστούν οι μεταβλητές με το μεγαλύτερο φορτίο σε κάθε συνιστώσα καθώς αυτές οι μεταβλητές μπορούν να θεωρηθούν ως αυτές οι οποίες τελικά διαχωρίζουν τις ομάδες των ασθενών με τη μεγαλύτερη ακρίβεια και επομένως μπορούν να χρησιμοποιηθούν για την πρόβλεψη του τύπου της ασθένειας σε νέα δείγματα για τα οποία δεν είναι γνωστός ο υποτύπος. Στην Εικόνα 10 παρουσιάζονται οι μεταβλητές με τα μεγαλύτερα ανά διαφορετικό τύπο omics δεδομένων για την πρώτη συνιστώσα και στην Εικόνα 11 για τη δεύτερη συνιστώσα.

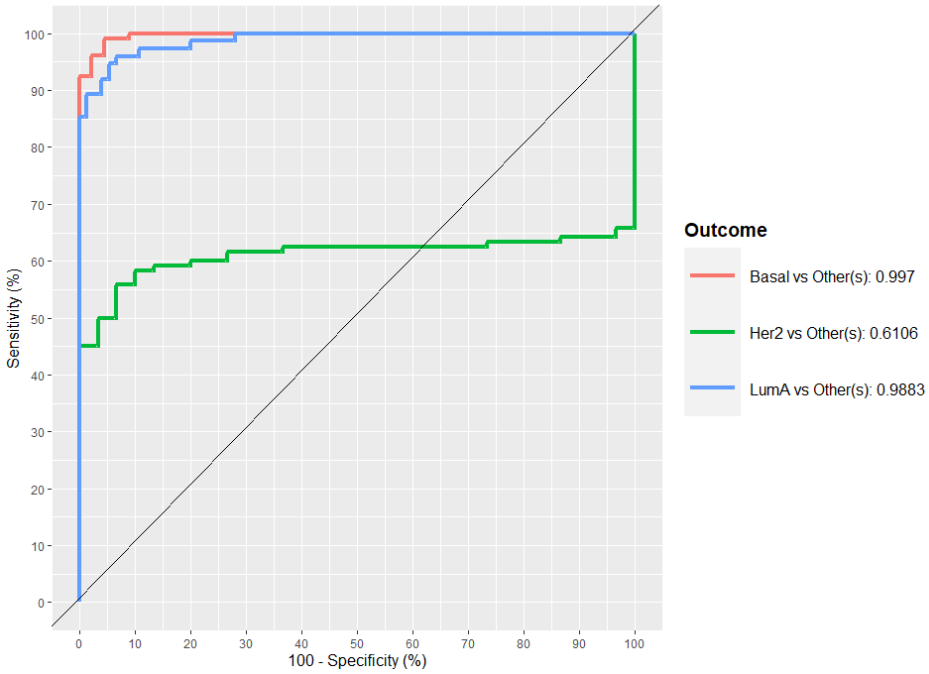


Εικόνα 10. Φορτία μεταβλητών ανά σύνολο δεδομένων για την πρώτη συνιστώσα.

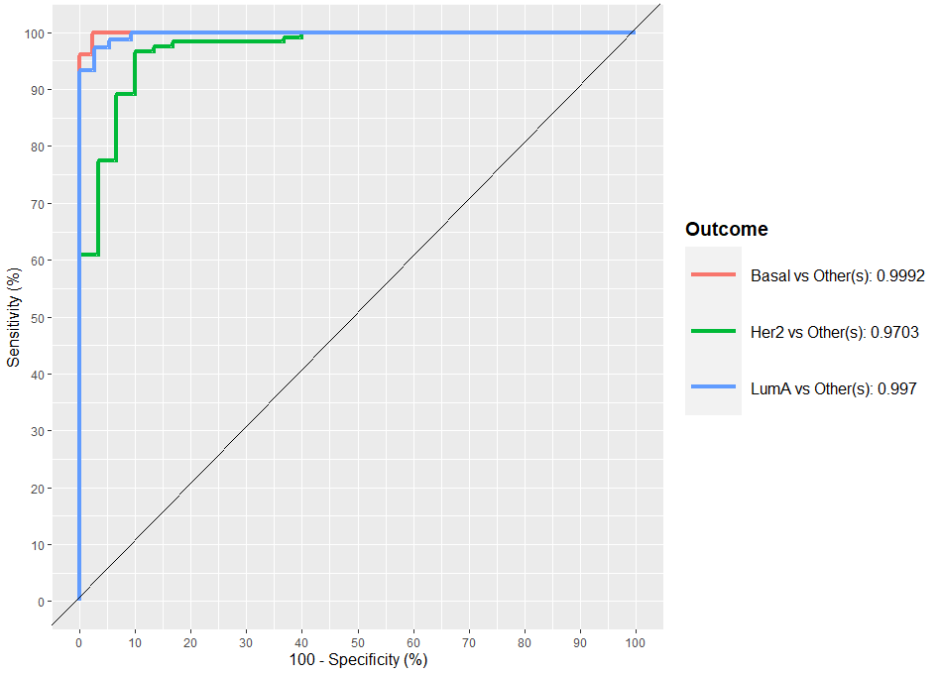


Εικόνα 11. Φορτία μεταβλητών ανά σύνολο δεδομένων για τη δεύτερη συνιστώσα.

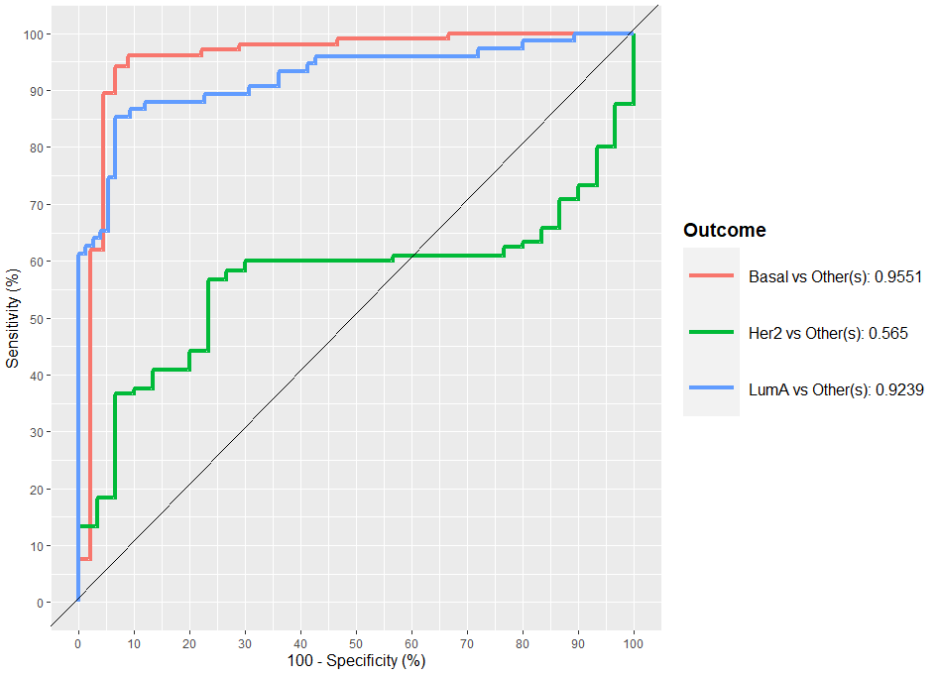
Τέλος, ελέγχεται η απόδοση του μοντέλου ως προς την ταξινόμηση (classification performance) των ασθενών. Στην περίπτωση που βρεθεί το μοντέλο να έχει ικανοποιητική απόδοση μπορεί αυτό να χρησιμοποιηθεί για μελέτη πρόβλεψης. Στις Εικόνες 9, 11 και 13 φαίνεται η απόδοση της ταξινόμησης των ασθενών ανά ομάδα (τύπος ασθένειας) ως προς την πρώτη συνιστώσα για τα σύνολα δεδομένων mRNA, miRNA και protein αντίστοιχα. Η απόδοση ως προς τη δεύτερη συνιστώσα βελτιώνεται σημαντικά (Εικόνα 12, 14 και 16) και το αποτέλεσμα αυτό επιβεβαιώνει τα ευρήματα ως προς τα φορτία των μεταβλητών για κάθε συνιστώσα. Στη δεύτερη συνιστώσα τα φορτία είναι κατανεμημένα με μεγαλύτερη ισορροπία όσον αφορά τους τύπους δεδομένων αλλά και τις ομάδες ασθενών.



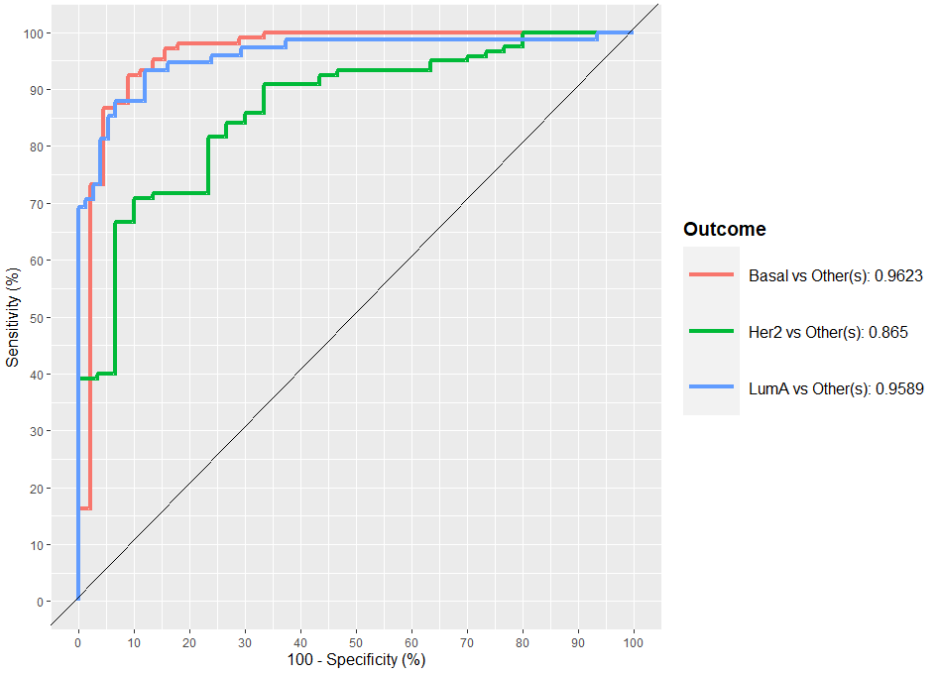
Εικόνα 12. AUC για το συνόλο δεδομένων mRNA ως προς την πρώτη συνιστώσα.



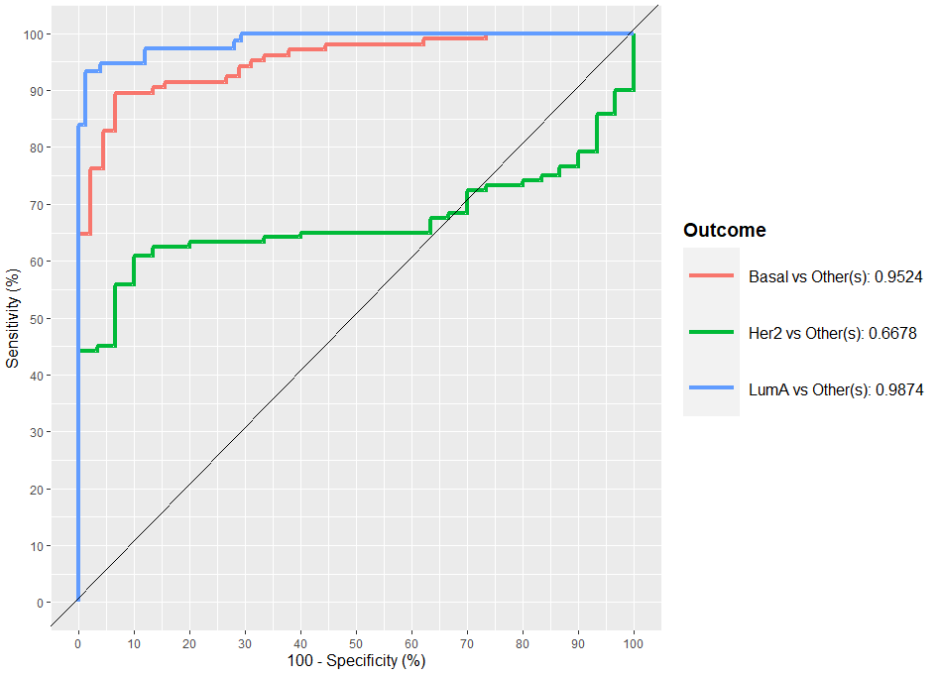
Εικόνα 13. AUC για το συνόλο δεδομένων mRNA ως προς τη δεύτερη συνιστώσα.



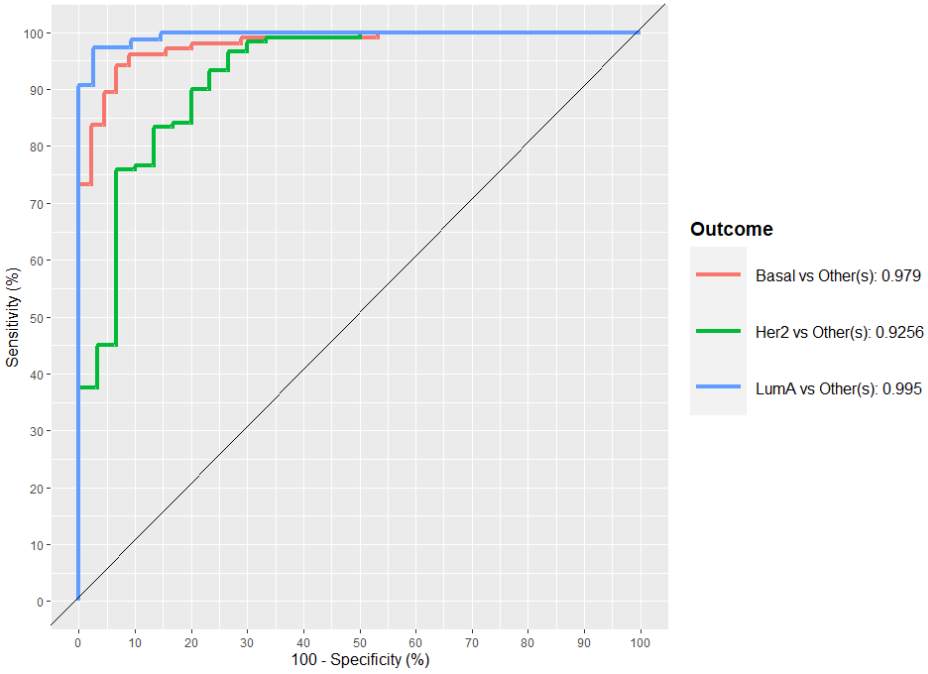
Εικόνα 14. AUC για το συνόλο δεδομένων miRNA ως προς την πρώτη συνιστώσα.



Εικόνα 15. AUC για το συνόλο δεδομένων miRNA ως προς τη δεύτερη συνιστώσα.



Εικόνα 16. AUC για το συνόλο δεδομένων protein ως προς την πρώτη συνιστώσα.



Εικόνα 17. AUC για το συνόλο δεδομένων protein ως προς τη δεύτερη συνιστώσα.

Χρησιμοποιώντας τα παραπάνω αποτελέσματα που έχουν προκύψει για το μοντέλο, διενεργήθηκε μελέτη πρόβλεψης για ασθενείς των οποίων ο τύπος της ασθένειας είναι άγνωστος. Από τον παρακάτω πίνακα συνάγεται πως για όλους τους τύπους της ασθένειας η πρόβλεψη έχει γίνει με αρκετά μεγάλη ακρίβεια έχοντας μόνο κάποια δείγματα στα οποία δεν ήταν δυνατό να προβλεφθεί ο τύπος τους.

	Predicted as Basal	Predicted as Her2	Predicted as LumA	Predicted as NA
Basal	15	1	0	5
Her2	0	11	0	3
LumA	0	0	27	8

Πίνακας 4. Πρόβλεψη για τα δείγματα που δε φέρουν χαρακτηρισμό ως προς τον τύπο της ασθένειας.

3.3.MOFA

3.3.1. Περιγραφή

Η μέθοδος MOFA, είναι μια στατιστική μέθοδος για την ενσωμάτωση πολλαπλών omics δεδομένων και εμπίπτει στην κατηγορία της μη εποπτευόμενης (unsupervised) ανάλυσης. Η μέθοδος MOFA μπορεί να θεωρηθεί ως μία γενίκευση της ανάλυσης κύριων συνιστωσών (PCA) σε omics δεδομένα διαφορετικών τύπων. Λαμβάνοντας υπόψη πίνακες δεδομένων με μετρήσεις πολλαπλών τύπων δεδομένων omics στα ίδια ή σε μερικώς επικαλυπτόμενα σύνολα δειγμάτων, η μέθοδος MOFA αναπαράγει μια νέα αναπαράσταση των δεδομένων σε μικρότερη διάσταση όπου τα αποτελέσματα είναι ερμηνεύσιμα χάρη στους παράγοντες (factors) που δημιουργούνται. Αυτοί οι παράγοντες έχουν τη βασική ιδιότητα να συλλαμβάνουν τις σημαντικές πηγές διακύμανσης στα διαφορετικά σύνολα δεδομένων omics, διευκολύνοντας έτσι τον εντοπισμό μοριακών μεταβολών ή διακριτών υποομάδων δειγμάτων. Τα φορτία των παραγόντων αυτών είναι μεταβλητές των αρχικών πινάκων και έτσι μπορούν να εξαχθούν συμπεράσματα ως προς τη σύνδεση μεταξύ των παραγόντων και των μοριακών χαρακτηριστικών. Είναι σημαντικό να αναφερθεί πως η μέθοδος αυτή μπορεί να διακρίνει αν ο κάθε παράγοντας που δημιουργείται είναι μοναδικός για κάθε τύπο omics δεδομένων ή αν σχετίζεται με διαφορετικούς τύπους δεδομένων οπότε και υπάρχει η δυνατότητα να

αναδειχθούν κοινοί άξονες μεταβλητότητας μεταξύ των διαφορετικών τύπων δεδομένων. Το μοντέλο που προκύπτει, μετά τη φάση εκπαίδευσης, μπορεί να χρησιμοποιηθεί για διαφορετικές αναλύσεις, όπως η οπτικοποίηση, η ομαδοποίηση και ταξινόμηση των δειγμάτων στο νέο χώρο χαμηλής διάστασης που συνίσταται από τους παράγοντες, καθώς και τον χαρακτηρισμό των παραγόντων με χρήση της ανάλυσης εμπλουτισμού (enrichment analysis), όπως και στον εντοπισμό ακραίων δειγμάτων (outliers) και αναπλήρωση τιμών που απουσιάζουν από τα δεδομένα (imputation).

3.3.2. Τύποι δεδομένων

Διάφοροι τύποι δεδομένων μπορούν να αναλυθούν με τη μέθοδο MOFA όπως δεδομένα έκφρασης RNA (RNA-seq), μεθυλίωσης DNA, απόκρισης σε φαρμακευτική θεραπεία και μεταδεδομένων των δειγμάτων. Γενικά δεν υπάρχει όριο στον αριθμό και τον τύπο δεδομένων με την προϋπόθεση αυτά να έχουν υποστεί μία προετοιμασία ώστε να είναι συμβατά με τη μέθοδο (π.χ. δεδομένα RNA-seq θα πρέπει να έχουν κανονικοποιηθεί και μοντελοποιηθεί με Gaussian κατανομή). Η MOFA μπορεί επίσης να χειριστεί τιμές που λείπουν από τα σύνολα δεδομένων που χρησιμοποιούνται. Τα δείγματα θα πρέπει να είναι τα ίδια σε διαφορετικά σύνολα δεδομένων omics ή τουλάχιστον να επικαλύπτονται σε κάποιον βαθμό. Ο ελάχιστος αριθμός δειγμάτων πρέπει να είναι δεκαπέντε.

3.3.3. Εφαρμογή και αποτελέσματα

Το πακέτο MOFA+ εφαρμόστηκε σε δεδομένα 200 ασθενών τα οποία προέρχονται από τέσσερα διαφορετικά -omics πεδία. Πιο συγκεκριμένα, έχουμε δεδομένα μεθυλίωσης του DNA, RNA-seq, σωματικών μεταλλάξεων και δεδομένα απόκρισης σε φάρμακα. Τα δεδομένα που χρησιμοποιήθηκαν παρουσιάζονται στους παρακάτω πίνακες.

	H045	H109	H024	H056	H079	H164	H059	H167	H113	H049
D_001_1	0,023639380942096	0,073598996822619	NA	0,058139303504587	0,020420767940 5	0,029627250975091	0,027697899610797	0,083659258571109	0,124549918646179	0,037047835507875
D_001_2	0,046232742977211	0,106230018907542	NA	0,090220277563438	0,047505428634585	0,080546282204532	0,080653563515795	0,144983594650632	0,191173183591 5	0,087363249024272
D_001_3	0,318747061101616	0,273289095997414	NA	0,232214515811653	0,363896220976851	0,472599146995332	0,476526425556326	0,501205592598379	0,280352978955372	0,500607966179189

D_001_4	0,823702723796607	0,717137940155506	NA	0,722573641422706	0,807390744665635	0,817914322627347	0,793896752456273	0,878446706895314	0,639713353446484	0,926103119015592
D_001_5	0,896277691626947	0,885000342882894	NA	0,795749657854532	0,879488646460444	0,892796131340132	0,948585330641063	1,02095915342485	0,854098973487536	0,993191441426517

Πίνακας 5. MOFA - Υπόδειγμα συνόλου δεδομένων drugs.

	H045	H109	H024	H056	H079	H164	H059	H167	H113	H049
cg10146935	1,8110858498206	-3,99750845839768	-2,84431298171718	-3,33865610898556	-0,019362027947554	-2,48599708044986	-1,46021132072567	-4,95229089986933	-2,98020882197335	-0,091712682621574
cg26837773	-5,17257225425006	1,59487016037977	0,16117048658491	-2,09343255641179	3,74897955776427	0,06053030703837838	-3,47223212392174	0,54757730382644	2,44009810817957	2,94076721198039
cg17801765	5,411526313571	5,41269252666631	0,36570594130089	0,37363416836074	5,41200956666996	5,26890755243885	-4,98999854210598	5,33708106170601	0,74954566283504	0,426492620726609
cg13244315	-0,118825084420945	1,04387064402101	-4,219236222289489	-1,592196418452645	1,41641834406155	4,6598310208793	-0,461119831045204	-1,918860866865977	-1,23701504999055	-0,421916267595282
cg06181703	5,1203837592182	1,27948027978667	0,72110040388410 8	4,04705938789 12	5,23742245977866	1,76124742209 87	4,54399748885432	4,93946326764876	4,78168325735127	5,05107299295084

Πίνακας 6. MOFA - Υπόδειγμα συνόλου δεδομένων methylation.

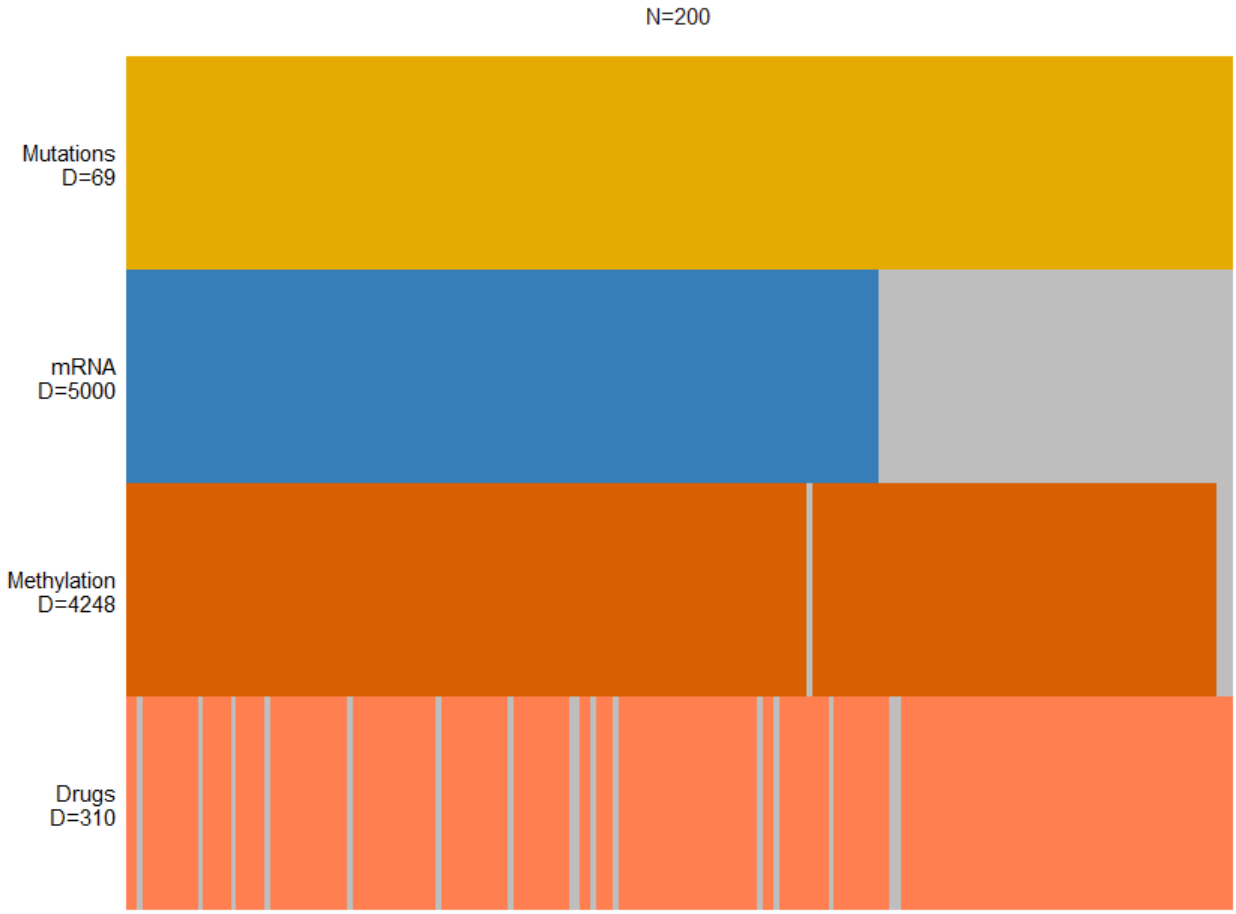
	H045	H109	H024	H056	H079	H164	H059	H167	H113	H049
ENSG00000244734	4,55864355066888	2,72151226637397	9,93845570346189	13,2780044922396	6,086873856581	2,5718393421321	4,93896055791459	1,52884844915265	2,28612220819862	2,5046992238725
ENSG00000158528	11,7418538709691	13,2874317329878	2,34100585326454	3,23287407422519	11,9408196807182	11,5068184040119	5,48367519462055	2,61886913795501	2,81280053774393	2,5046992238725
ENSG00000198478	8,92145551607793	2,72151226637397	12,3814519768193	8,10626624967317	4,88950292523214	12,756213186746	3,59389000161745	4,11949043640136	5,22004136871471	2,8848966835096
ENSG00000175445	12,6864582980058	10,9259850667711	1,52884844915265	1,52884844915265	13,3405882376542	10,885547285613	11,1940286064369	11,5999800766264	2,28612220819862	2,8848966835096
ENSG00000174469	2,6449457023189	12,6483553590615	1,52884844915265	13,5652102941789	5,47691374442363	10,9751874774257	7,94424565412073	2,61886913795501	2,28612220819862	12,9409569507993

Πίνακας 7. MOFA - Υπόδειγμα συνόλου δεδομένων mRNA.

	H045	H109	H024	H056	H079	H164	H059	H167	H113	H049
gain2p25.3	0	0	0	0	1	0	0	0	0	0
gain3q26	0	0	0	0	0	0	0	0	0	0
del6p21.2	0	0	0	0	0	0	0	0	0	0
del6q21	0	0	0	0	0	0	0	0	0	0
del8p12	0	0	0	0	0	0	0	1	0	0

Πίνακας 8. MOFA - Υπόδειγμα συνόλου δεδομένων mutations.

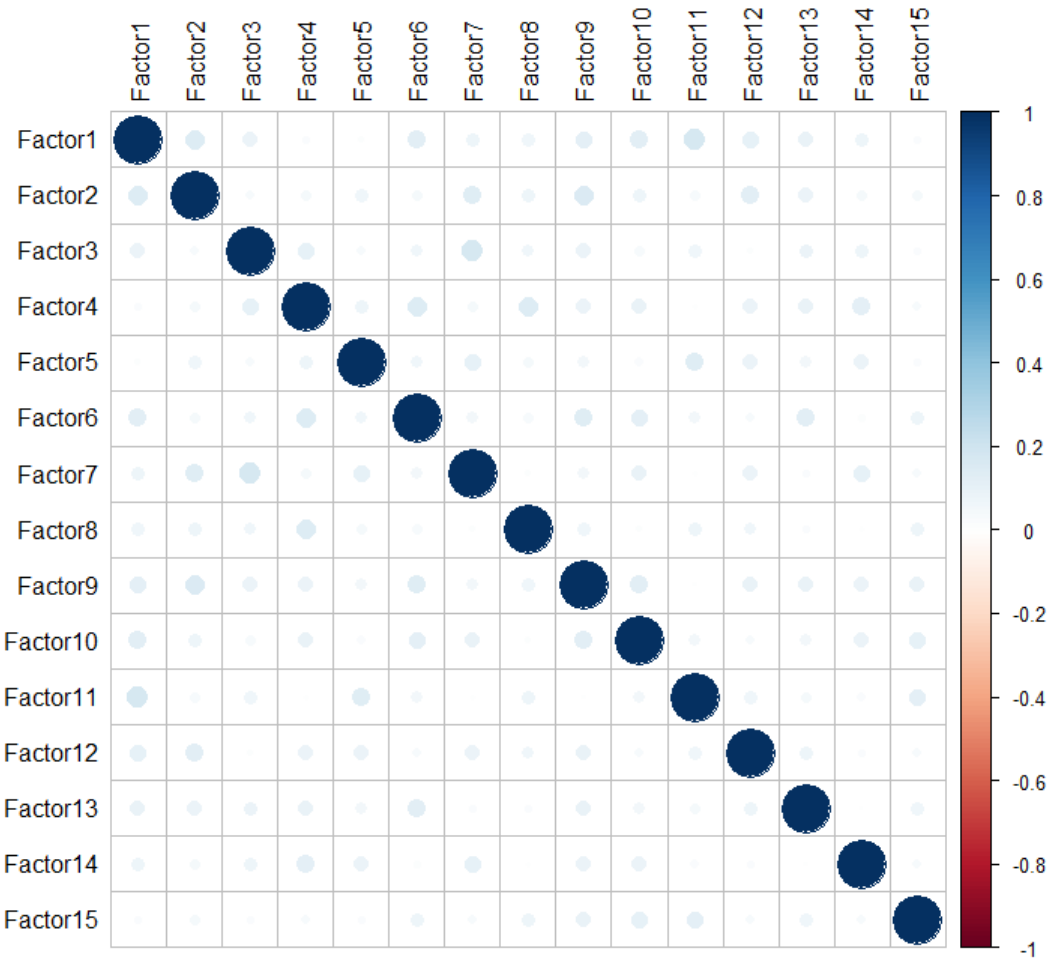
Αρχικά παρέχεται η δυνατότητα οπτικοποίησης των δεδομένων που θα αναλυθούν όπως φαίνεται στην Εικόνα 18.



Εικόνα 18. Διαθέσιμα δεδομένα για τους ασθενείς που συμμετέχουν στην ανάλυση. Ο αριθμός D αντιστοιχεί στον αριθμό μεταβλητών κάθε πεδίου. Με γκρι χρώμα απεικονίζονται τα δεδομένα που δεν είναι διαθέσιμα σε κάθε περίπτωση.

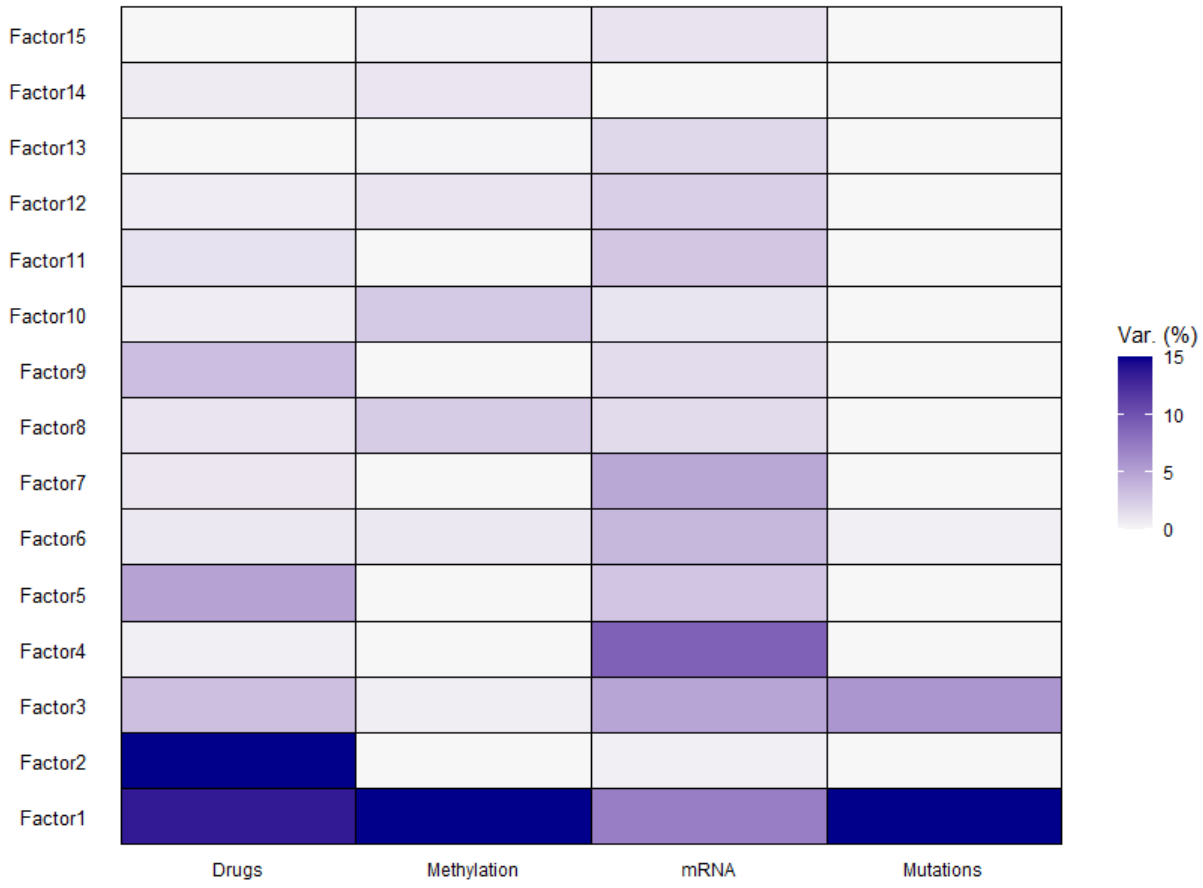
Στη συνέχεια κατασκευάζεται και εκπαιδεύεται το μοντέλο που θα παράξει τα αποτελέσματα. Εδώ μπορούν να εισαχθούν σημαντικά ορίσματα που αφορούν τα δεδομένα όπως για παράδειγμα το scaling αυτών ή ο αριθμός των παραγόντων που θα δημιουργηθούν. Συνήθως ο αριθμός των παραγόντων είναι δέκα και περισσότεροι αλλά αυτό εξαρτάται από τα δεδομένα και τον αριθμό μεταβλητών που υπάρχουν διαθέσιμες. Αφού εκπαιδευτεί το μοντέλο και για να διαπιστωθεί αν χρειάζεται να αλλάξει κάποια παράμετρος στο προηγούμενο βήμα, ελέγχουμε

τη συσχέτιση μεταξύ των παραγόντων. Όταν οι παράγοντες δεν παρουσιάζουν συσχέτιση, όπως φαίνεται στην Εικόνα 19 το μοντέλο έχει κατασκευαστεί σωστά.



Εικόνα 19. Πίνακας συσχετίσεων των παραγόντων.

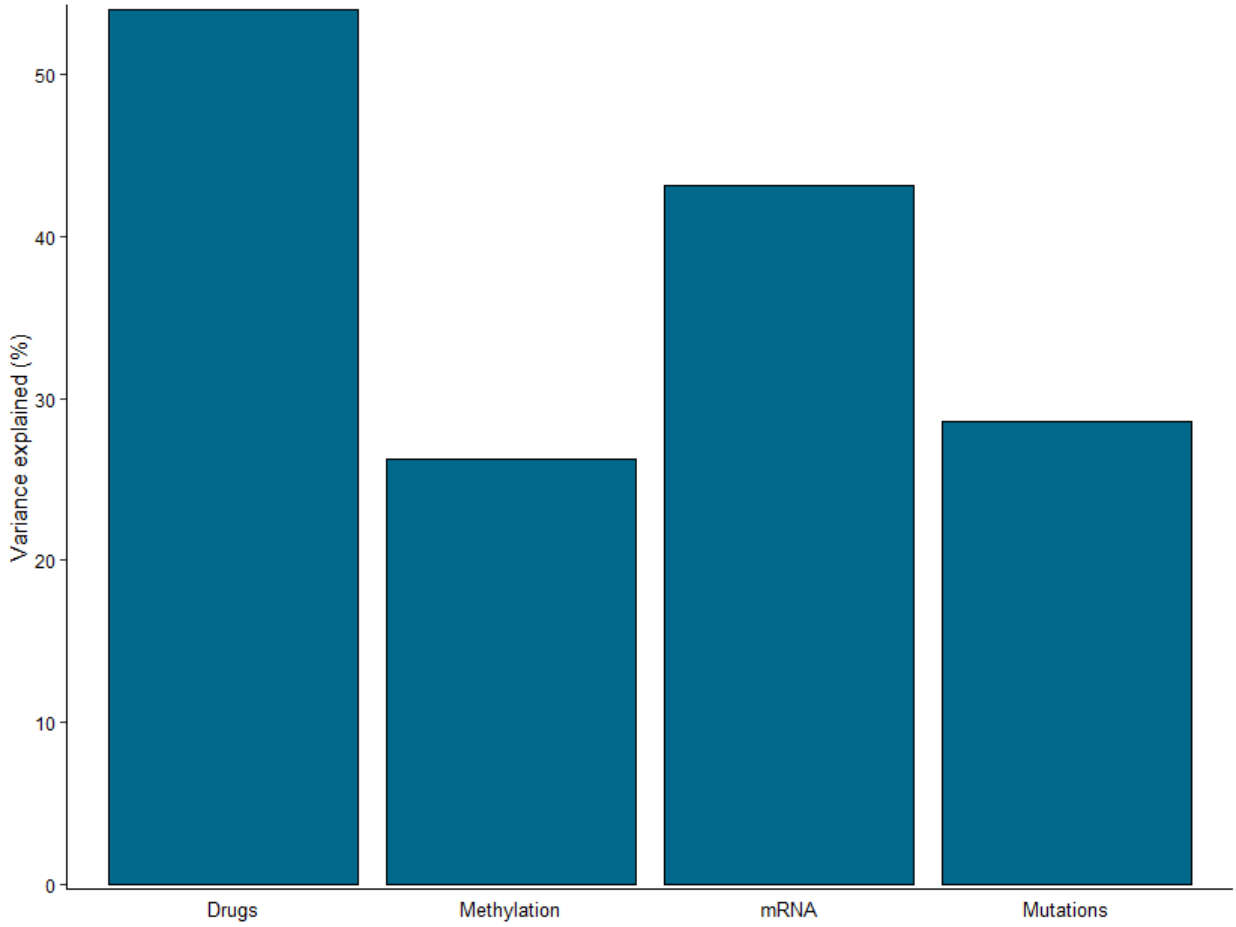
Η πιο σημαντική εικόνα που παράγει η MOFA είναι η ανάλυση των παραγόντων της διακύμανσης. Αυτό το διάγραμμα δείχνει το ποσοστό διακύμανσης που εξηγείται από κάθε παράγοντα σε κάθε τύπο δεδομένων. Συνοψίζει τις πηγές διακύμανσης από ένα πολύπλοκο ετερογενές σύνολο δεδομένων σε ένα μόνο σχήμα (Εικόνα 20).



Εικόνα 20. Ανάλυση μεταβλητότητας ανά παράγοντα. Ο πρώτος παράγοντας καταγράφει μια πηγή μεταβλητότητας που προέρχεται και από τους τέσσερις τύπους δεδομένων. Ο δεύτερος παράγοντας συλλαμβάνει μια πολύ ισχυρή πηγή διακύμανσης που είναι αποκλειστική για τα δεδομένα απόκρισης σε φάρμακα κοκ.

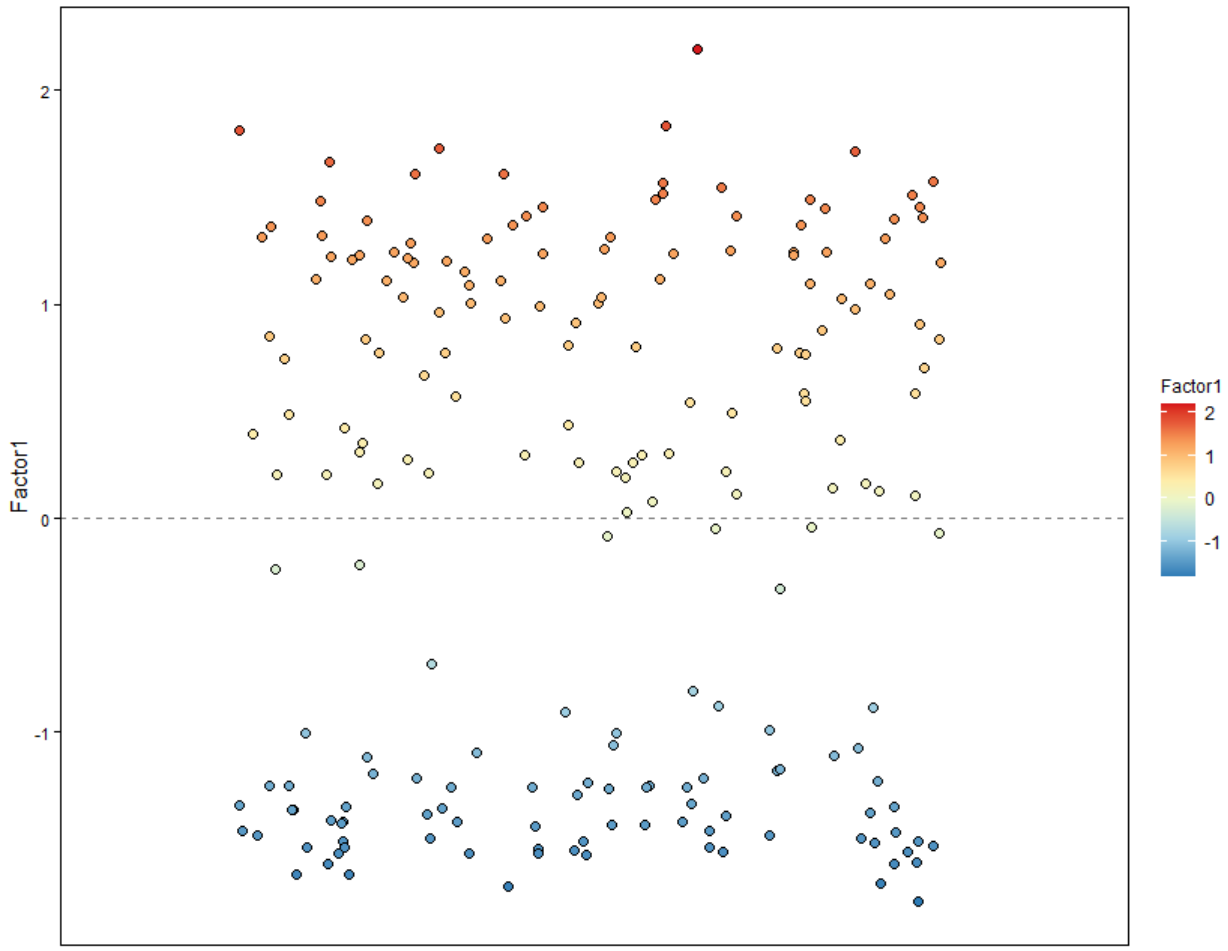
Ένα εύλογο ερώτημα είναι αν το μοντέλο παρέχει καλή προσαρμογή στα δεδομένα. Για αυτό μπορούμε να σχεδιάσουμε τη συνολική διακύμανση που εξηγείται (χρησιμοποιώντας όλους τους παράγοντες) όπως φαίνεται στην Εικόνα 21. Τα πολύπλοκα σύνολα δεδομένων με ισχυρές μη γραμμικότητες οδηγούν σε μικρή εξήγηση της διακύμανσης (<10%). Καθώς και ο μεγάλος αριθμός των δειγμάτων ενώ αντίθετα όσο μεγαλύτερος είναι ο αριθμός των παραγόντων, τόσο μεγαλύτερη είναι και η συνολική διακύμανση που εξηγείται.

Η μέθοδος MOFA στηρίζεται σε ένα γραμμικό μοντέλο. Αυτό έχει ως αποτέλεσμα την αποφυγή υπερ-προσαρμογής (overfitting) στα δεδομένα αλλά χωρίς να μπορεί να εξηγήσει το σύνολο της διακύμανσης ακόμη και αν χρησιμοποιηθούν πολλοί παράγοντες.



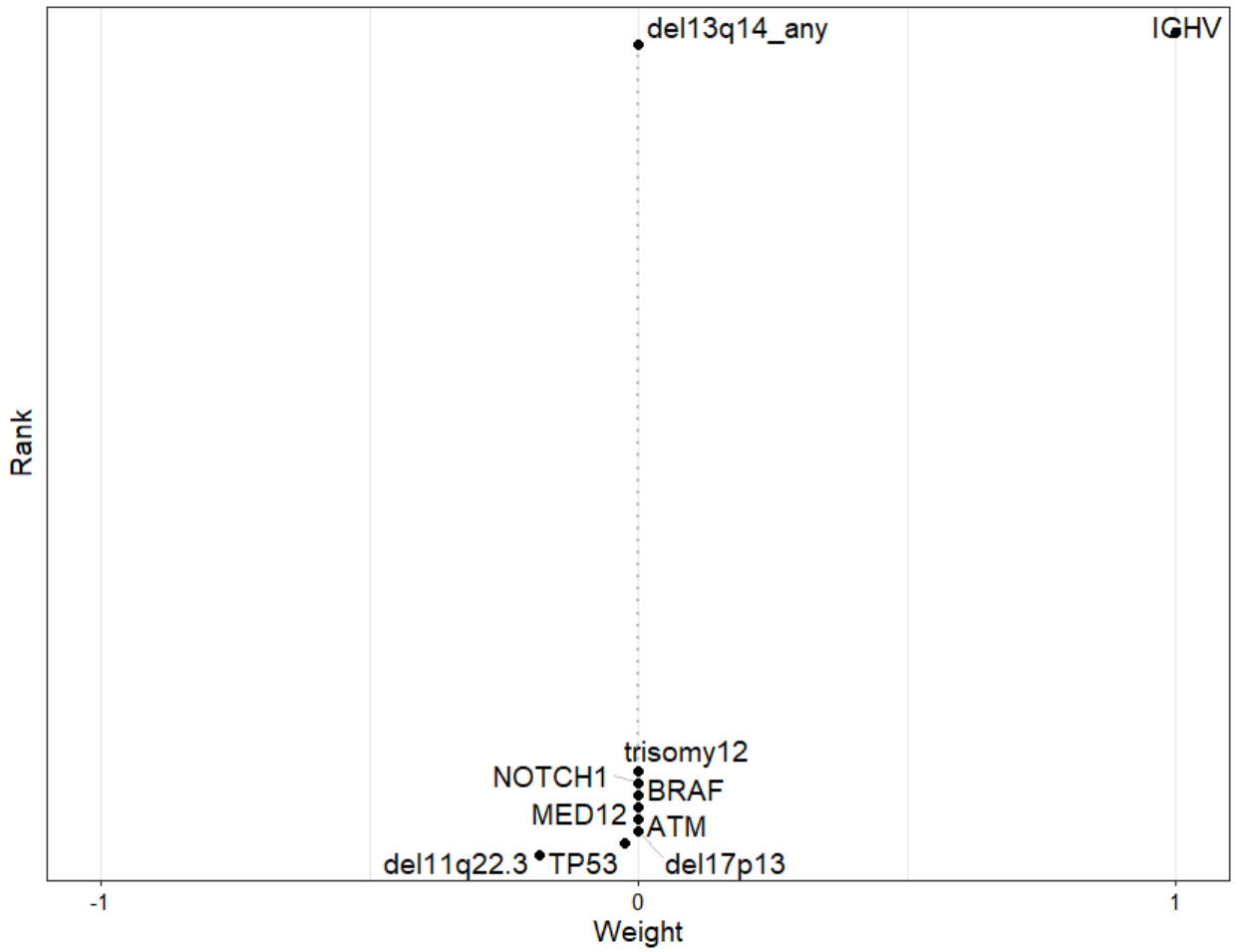
Εικόνα 21. Συνολική διακύμανση για κάθε τύπο δεδομένων που έχει αποτυπωθεί στο μοντέλο και για τους 15 παράγοντες. Περίπου 54% της διακύμανσης της απόκρισης σε φάρμακα και 42% για την περίπτωση του mRNA.

Κάθε παράγοντας συλλαμβάνει μια διαφορετική πηγή μεταβλητότητας στα δεδομένα. Από μαθηματική οπτική, κάθε παράγοντας ορίζεται ως ένας γραμμικός συνδυασμός των μεταβλητών των αρχικών δεδομένων. Δείγματα με διαφορετικό πρόσημο αποτυπώνουν αντίθετους φαινότυπους κατά μήκος του άξονα της διακύμανσης (Εικόνα 22).

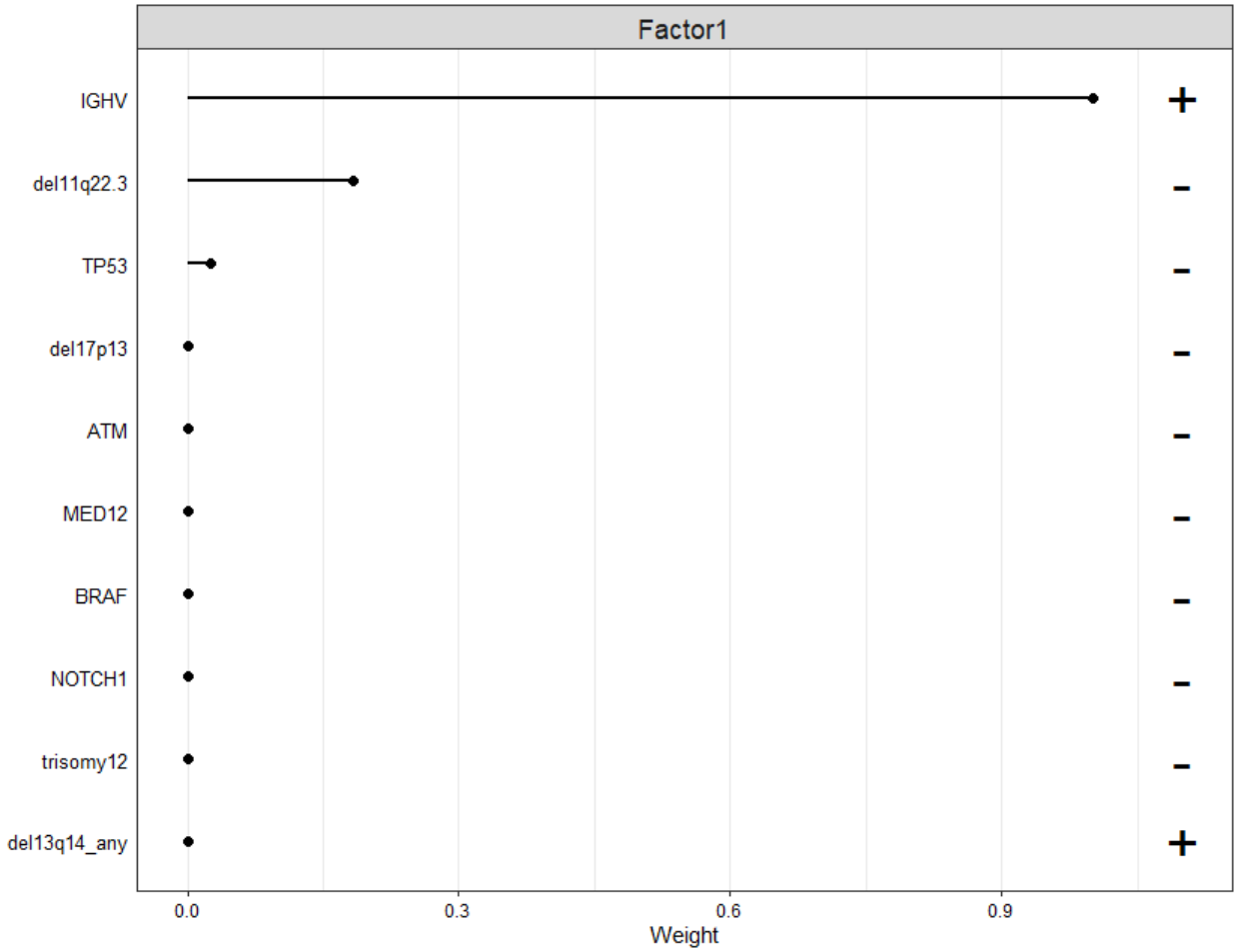


Εικόνα 22. Κατανομή των δειγμάτων ως προς τον πρώτο παράγοντα. Δείγματα με αντίθετο πρόσημο παρουσιάζουν διαφορετικούς φαινοτύπους. Όσο μεγαλύτερη είναι η απόσταση από τον άξονα τόσο ισχυρότερη είναι η σχέση ως προς το συγκεκριμένο παράγοντα.

Στη συνέχεια, ελέγχονται τα βάρη των αρχικών μεταβλητών στον πρώτο παράγοντα (Εικόνα 23) και διακρίνεται εύκολα πως ενώ τα περισσότερα βάρη βρίσκονται κοντά στο μηδέν, το βάρος του γονιδίου IGHV (immunoglobulin heavy chain variable) είναι μεγάλο. Όπως φαίνεται στην Εικόνα 24, όπου απεικονίζονται τα δέκα γονίδια με το μεγαλύτερο βάρος, το πρόσημο για το γονίδιο IGHV είναι θετικό για τον πρώτο παράγοντα που σημαίνει ότι τα δείγματα που είναι θετικά για αυτόν τον παράγοντα έχουν τη μετάλλαξη στο συγκεκριμένο γονίδιο.

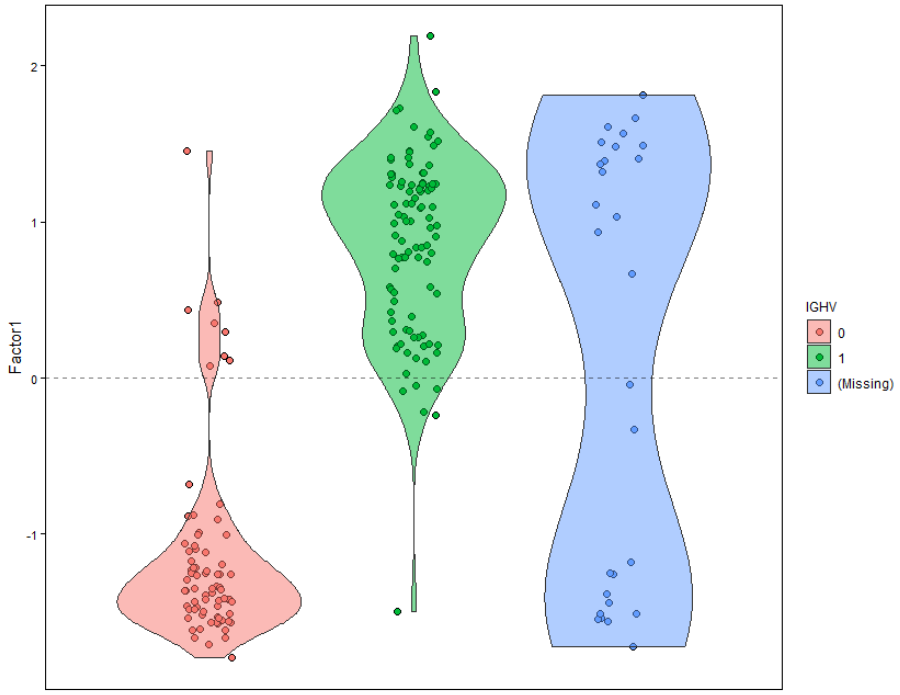


Εικόνα 23. Βάρη σωματικών μεταλλάξεων ως προς τον πρώτο παράγοντα. Το γονίδιο IGHV έχει το μεγαλύτερο βάρος, ενώ τα υπόλοιπα έχουν σχεδόν μηδενικό.

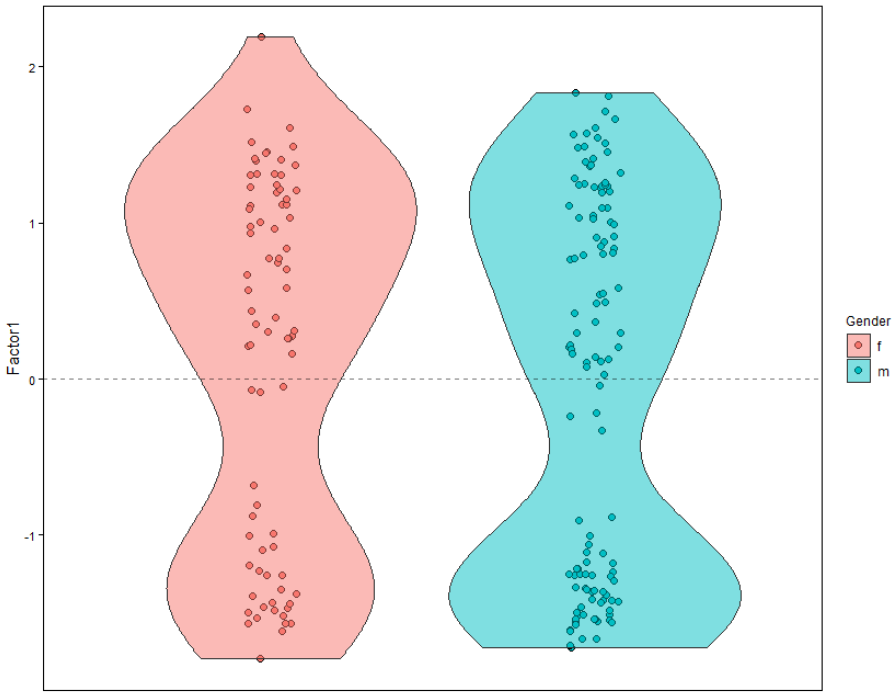


Εικόνα 24. Οι δέκα μεταβλητές από το σύνολο δεδομένων των σωματικών μεταλλάξεων με το μεγαλύτερο βάρος για τον πρώτο παράγοντα. Το θετικό/αρνητικό πρόσημο υποδεικνύει θετική/αρνητική σχέση με το συγκεκριμένο παράγοντα.

Μπορεί κανείς να αποτυπώσει το διαχωρισμό των δειγμάτων ως προς έναν παράγοντα με διαφορετικούς τρόπους. Στην Εικόνα 25, τα διαφορετικά χρώματα αντιστοιχούν σε διαφορετικές καταστάσεις όσον αφορά τη μετάλλαξη του γονιδίου που βρέθηκε να έχει το μεγαλύτερο βάρος και στην Εικόνα 26 χρησιμοποιήθηκε το φύλλο για το χρωματισμό των δειγμάτων το οποίο ανήκει στο σύνολο των μεταδεδομένων που έχουν εισαχθεί στο μοντέλο. Αντίστοιχα σχήματα μπορούν να παραχθούν για άλλα χαρακτηριστικά όπως η ηλικία ή για μεταβλητές που υπάρχουν σε κάποιο από τα αρχικά σύνολα δεδομένων.

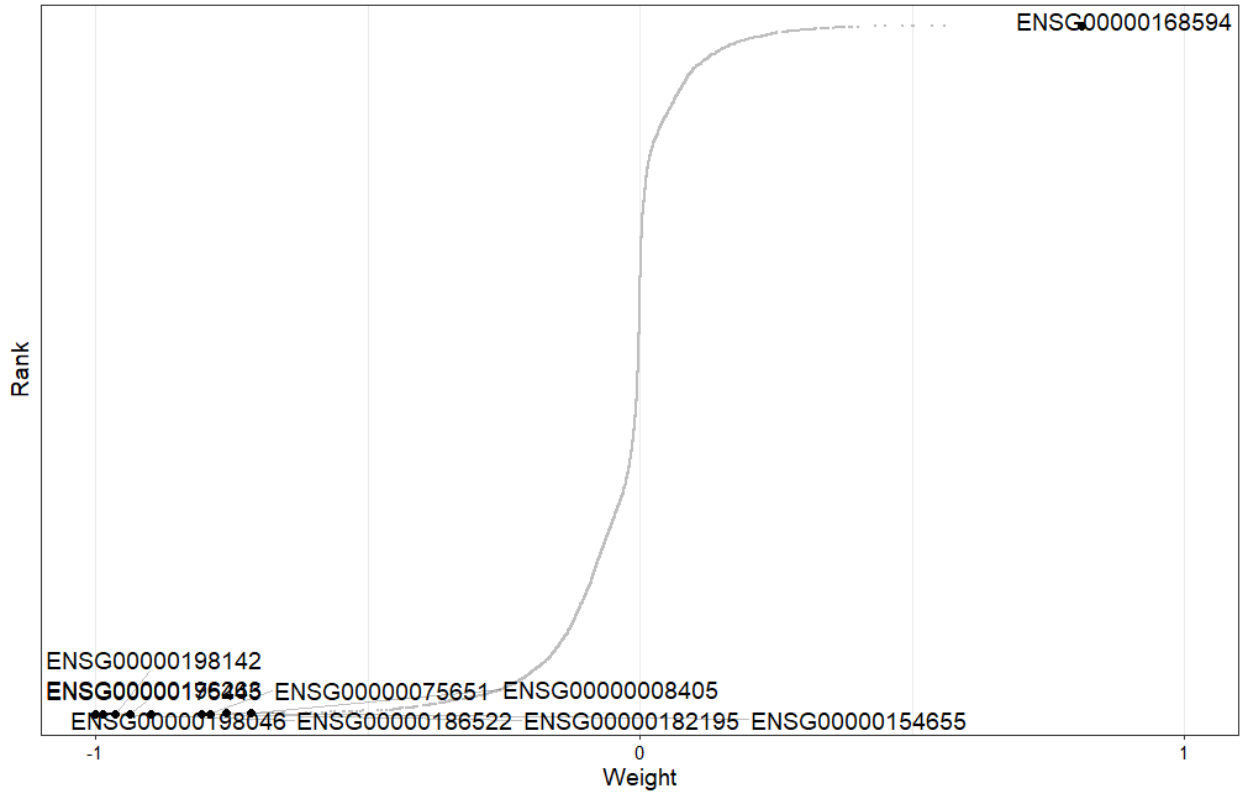


Εικόνα 25. Απεικόνιση των δειγμάτων ως προς τον πρώτο παράγοντα. Τα δείγματα με πράσινο χρώμα είναι αυτά που παρουσιάζουν την μετάλλαξη στο γονίδιο IGHV, ενώ με κόκκινο χρώμα υποδηλώνεται η απουσία μετάλλαξης.



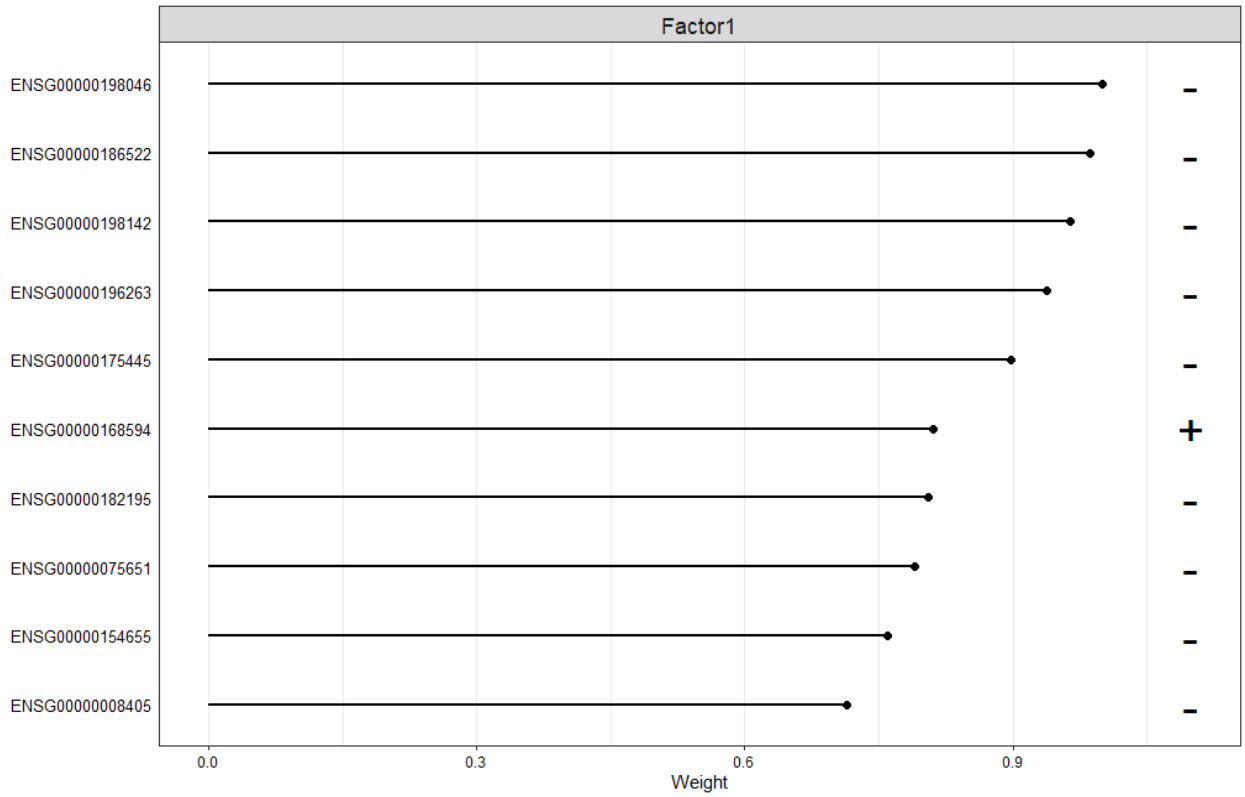
Εικόνα 26. Απεικόνιση των δειγμάτων ως προς τον πρώτο παράγοντα. Τα δείγματα έχουν ομαδοποιηθεί ως προς το φύλο. Είναι φανερό πως το φύλο δε συνδέεται με τον πρώτο παράγοντα.

Αντίστοιχη απεικόνιση μπορεί να γίνει για τα διαφορετικά σύνολα δεδομένων που έχουν ενσωματωθεί στο μοντέλο. Στην Εικόνα 27 φαίνονται τα γονίδια (mRNA) και τα αντίστοιχα βάρη τους στον πρώτο παράγοντα. Εδώ ένα μόνο γονίδιο παρουσιάζει θετικό βάρος, ενώ τα υπόλοιπα αρνητικό. Ο αριθμός των γονιδίων που εμφανίζονται κάθε φορά μπορεί να τροποποιηθεί.



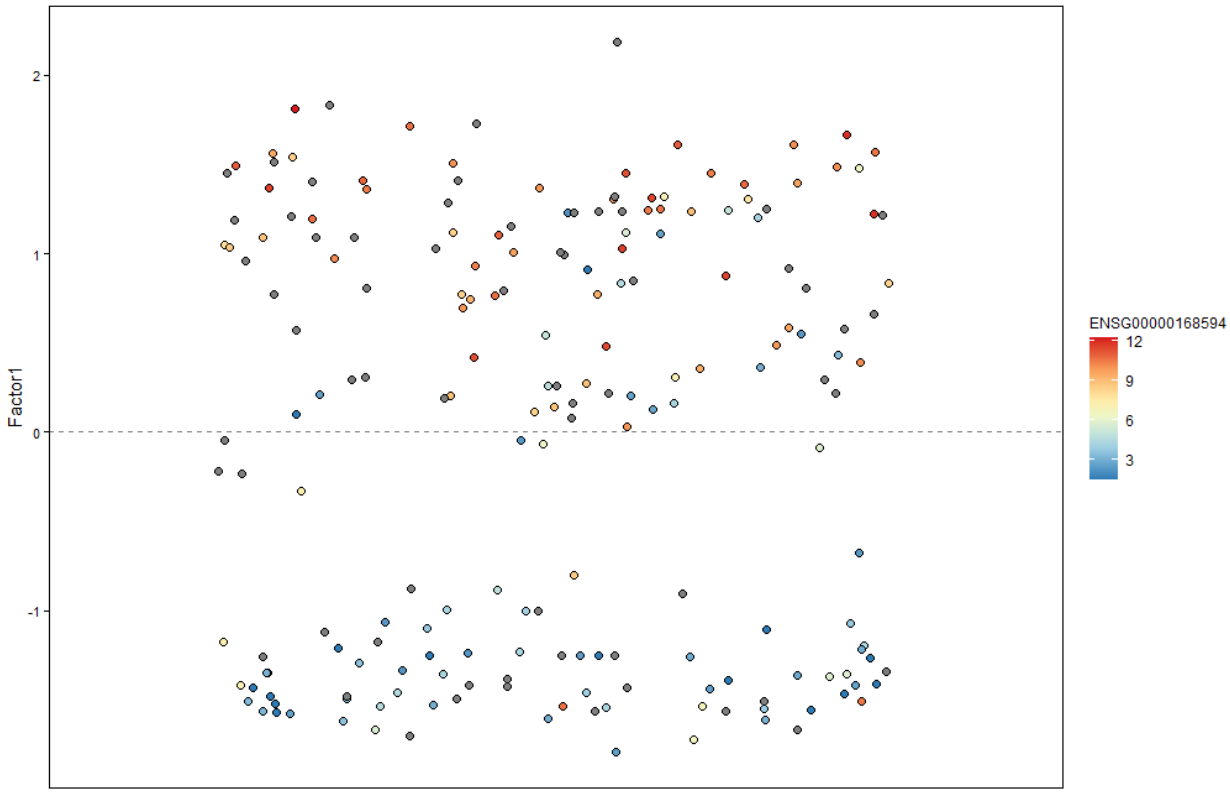
Εικόνα 27. Βάρη γονιδίων (mRNA) για τον πρώτο παράγοντα. Το γονίδιο ENSG00000168594 έχει θετικό πρόσημο και το γονίδιο με το μεγαλύτερο βάρος είναι το ENSG00000198046.

Στην Εικόνα 28 φαίνονται τα δέκα γονίδια με τα μεγαλύτερα βάρη καθώς και η σχέση τους με τον πρώτο παράγοντα. Μόνο ένα γονίδιο από αυτά έχει θετική σχέση με τον παράγοντα, ενώ τα υπόλοιπα αρνητική.



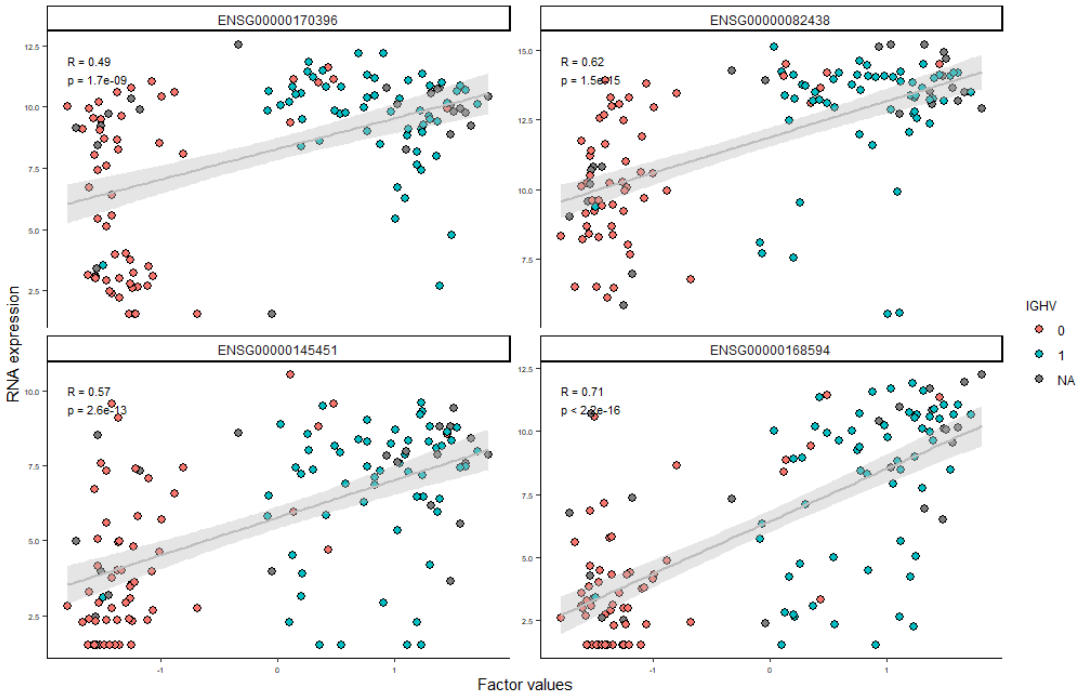
Εικόνα 28. Οι δέκα μεταβλητές από το σύνολο δεδομένων του mRNA με το μεγαλύτερο βάρος για τον πρώτο παράγοντα. Το θετικό/αρνητικό πρόσημο υποδεικνύει θετική/αρνητική σχέση με το συγκεκριμένο παράγοντα.

Καθώς οι τιμές έκφρασης, σε αντίθεση με τη μετάλλαξη IGHV ή το φύλο, λαμβάνει συνεχείς τιμές δεν είναι δυνατόν να ομαδοποιηθούν τα δείγματα παρά μόνο να απεικονιστούν με διαφορετικό χρώμα που αντιστοιχεί στις τιμές έκφρασης του συγκεκριμένου γονιδίου.

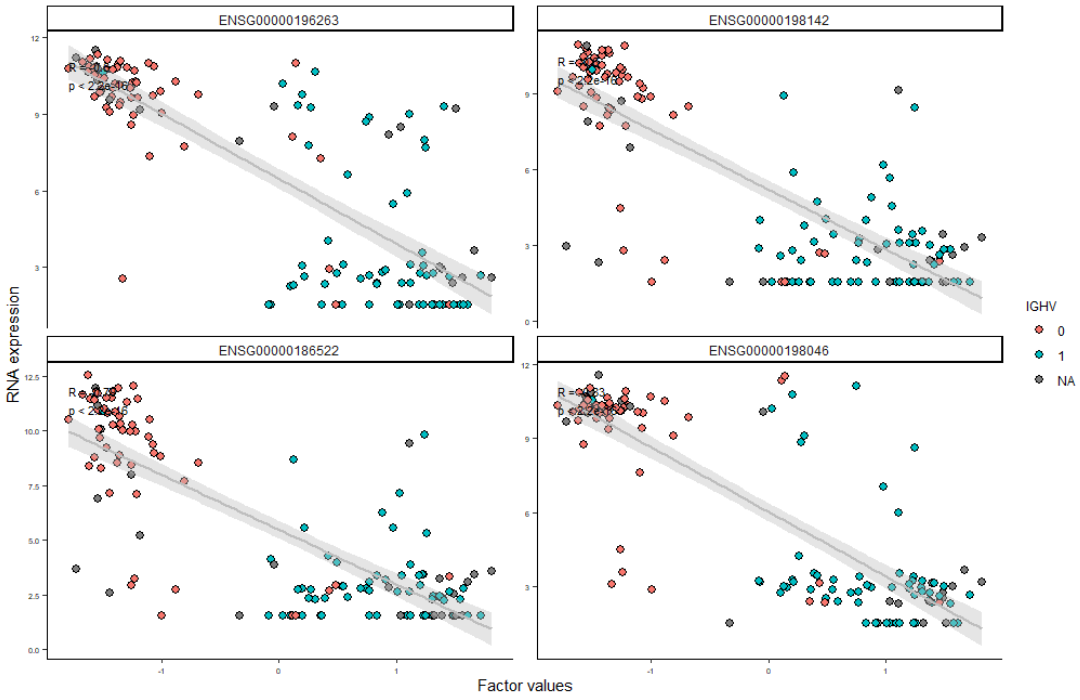


Εικόνα 29. Απεικόνιση των δειγμάτων ως προς τον πρώτο παράγοντα. Τα δείγματα χρωματίζονται βάσει των τιμών έκφρασης του γονιδίου ENSG00000168594.

Με την απεικόνιση της συσχέτισης της γονιδιακής έκφρασης ως προς τις τιμές του πρώτου παράγοντα μπορεί να διακρίνει κανείς πως τα δείγματα που έχουν την IGHV μετάλλαξη έχουν και μεγαλύτερη έκφραση των γονιδίων που έχουν θετικό βάρος για τον παράγοντα (βλ. Εικόνα 30). Αντίθετα τα γονίδια με αρνητικό βάρος για τον ίδιο παράγοντα (Εικόνα 31) παρουσιάζουν μικρότερη έκφραση για τα δείγματα που είναι αρνητικά στην μετάλλαξη IGHV.

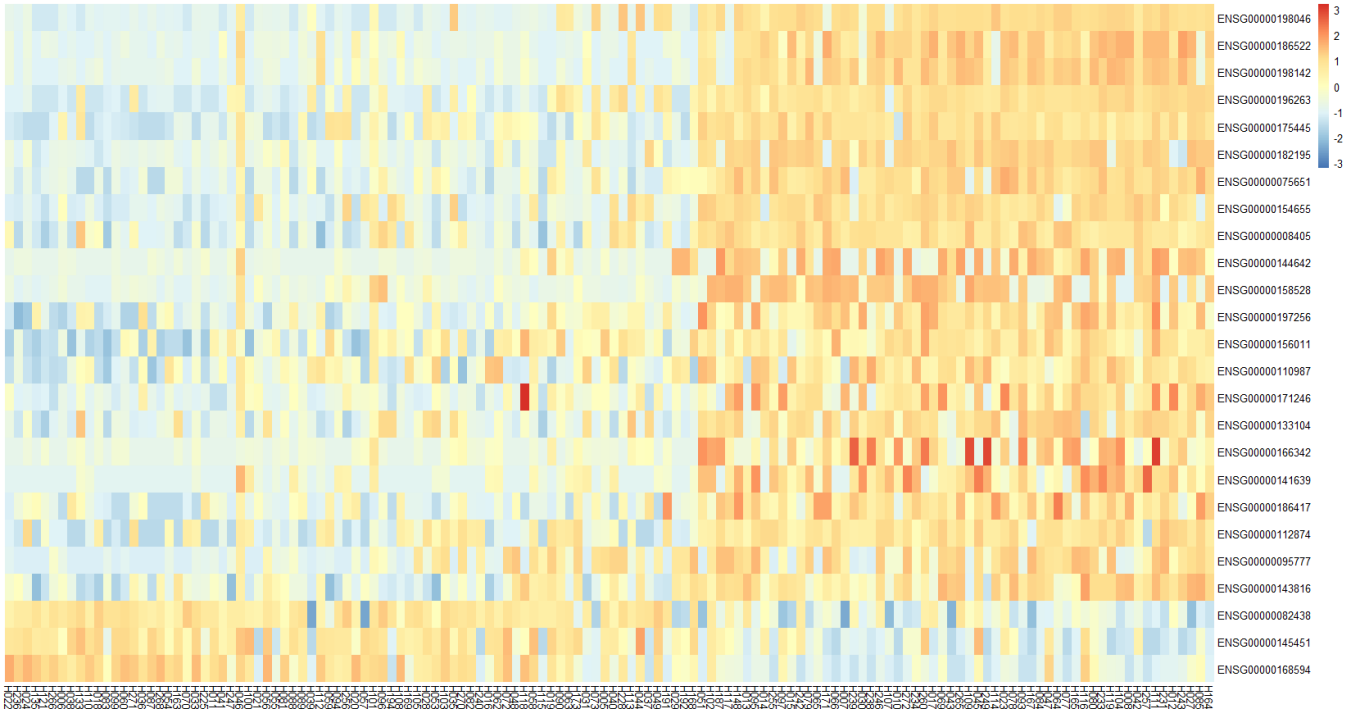


Εικόνα 30. Συσχέτιση γονιδιακής έκφρασης για τα γονίδια με το μεγαλύτερο θετικό βάρος ως προς τον πρώτο παράγοντα.



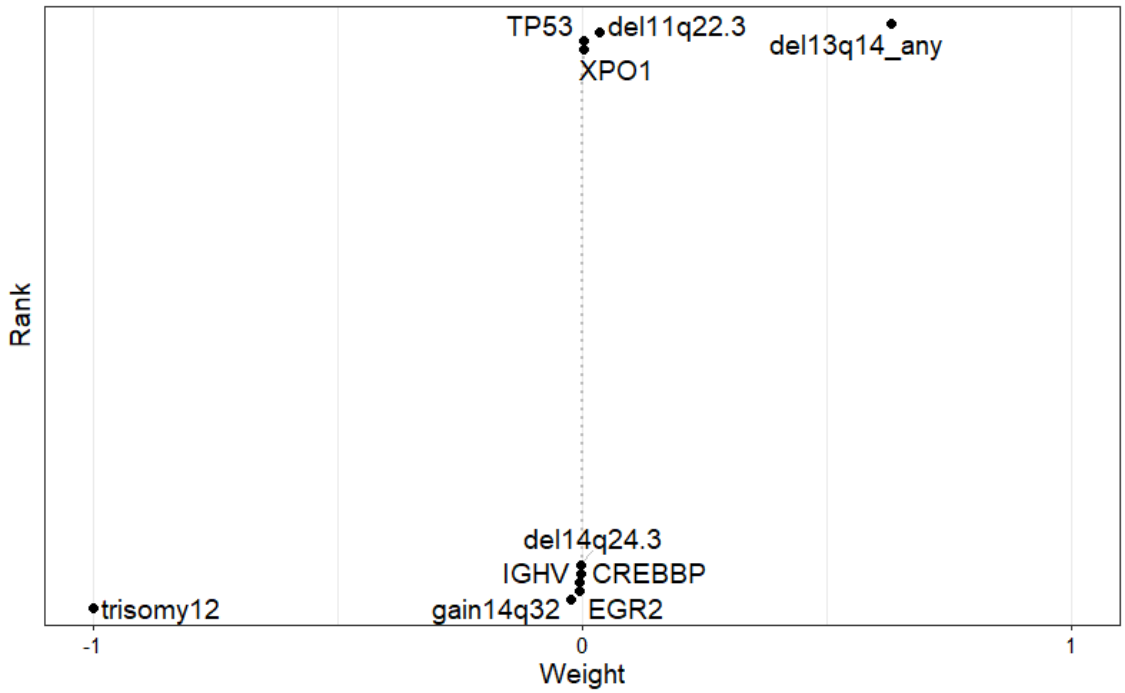
Εικόνα 31. Συσχέτιση γονιδιακής έκφρασης για τα γονίδια με το μεγαλύτερο αρνητικό βάρος ως προς τον πρώτο παράγοντα.

Ένας εναλλακτικός τρόπος απεικόνισης είναι με χρήση heatmap όπου φαίνονται εύκολα οι δύο ομάδες που προκύπτουν όσον αφορά την έκφραση των γονιδίων με τα μεγαλύτερα βάρη.

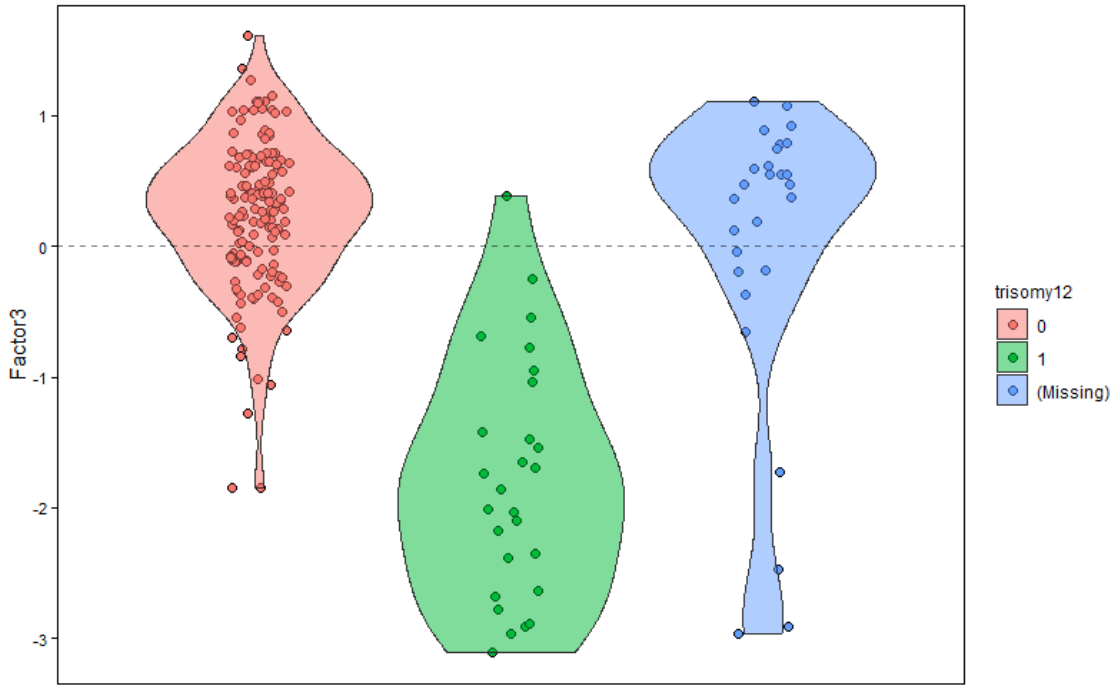


Εικόνα 32. Γονιδιακή έκφραση (mRNA) των δειγμάτων όσον αφορά τα γονίδια (25) με τα μεγαλύτερα βάρη στον πρώτο παράγοντα.

Η παραπάνω ανάλυση γίνεται για τον χαρακτηρισμό καθενός από τους σημαντικούς παράγοντες που έχει το μοντέλο (βλ. Εικόνα 20). Για να είναι δυνατή η περιγραφή κάποιων συνδυαστικών αποτελεσμάτων που παράγει το μοντέλο θα γίνει μία μικρή αναφορά στον τρίτο παράγοντα (factor 3) καθώς ο δεύτερος έχει συμπεριλάβει μέρος της διακύμανσης για μία μόνο κατηγορία δεδομένων, αυτήν της απόκρισης σε φαρμακευτική αγωγή. Ο χαρακτηρισμός του τρίτου παράγοντα, ο οποίος έχει συλλάβει μέρος της διακύμανσης της γονιδιακής έκφρασης και των σωματικών μεταλλάξεων, οδηγεί στο συμπέρασμα πως το χαρακτηριστικό με το μεγαλύτερο βάρος στη διαμόρφωση του είναι η τρισωμία 12 (Εικόνα 33). Διακρίνεται στην Εικόνα 34 η ομαδοποίηση των δειγμάτων ως προς την τρισωμία 12.

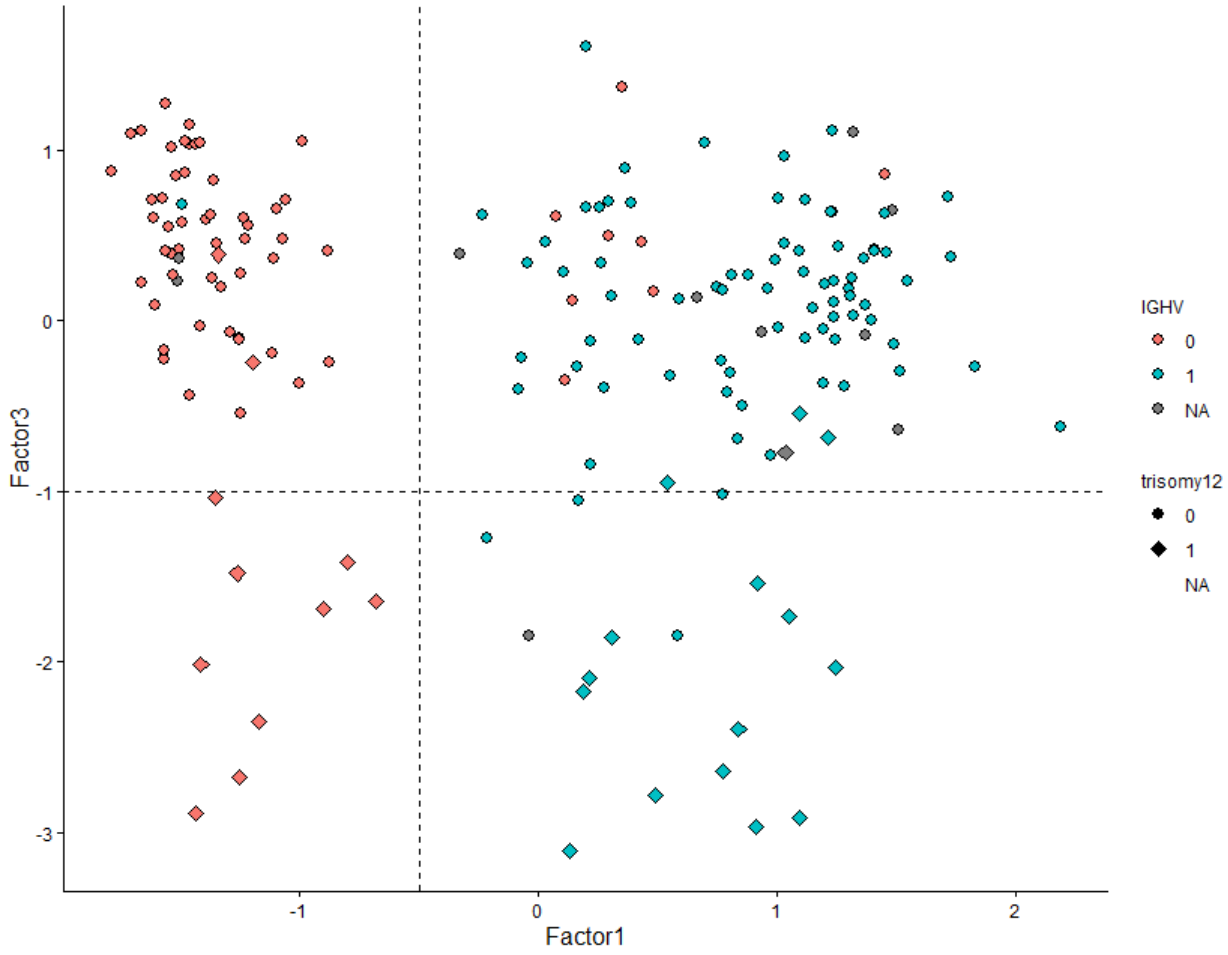


Εικόνα 33. Βάρη σωματικών μεταλλάξεων ως προς τον τρίτο παράγοντα. Η τρισωμία 12 (trisomy12) έχει το μεγαλύτερο βάρος, το οποίο είναι και αρνητικό.



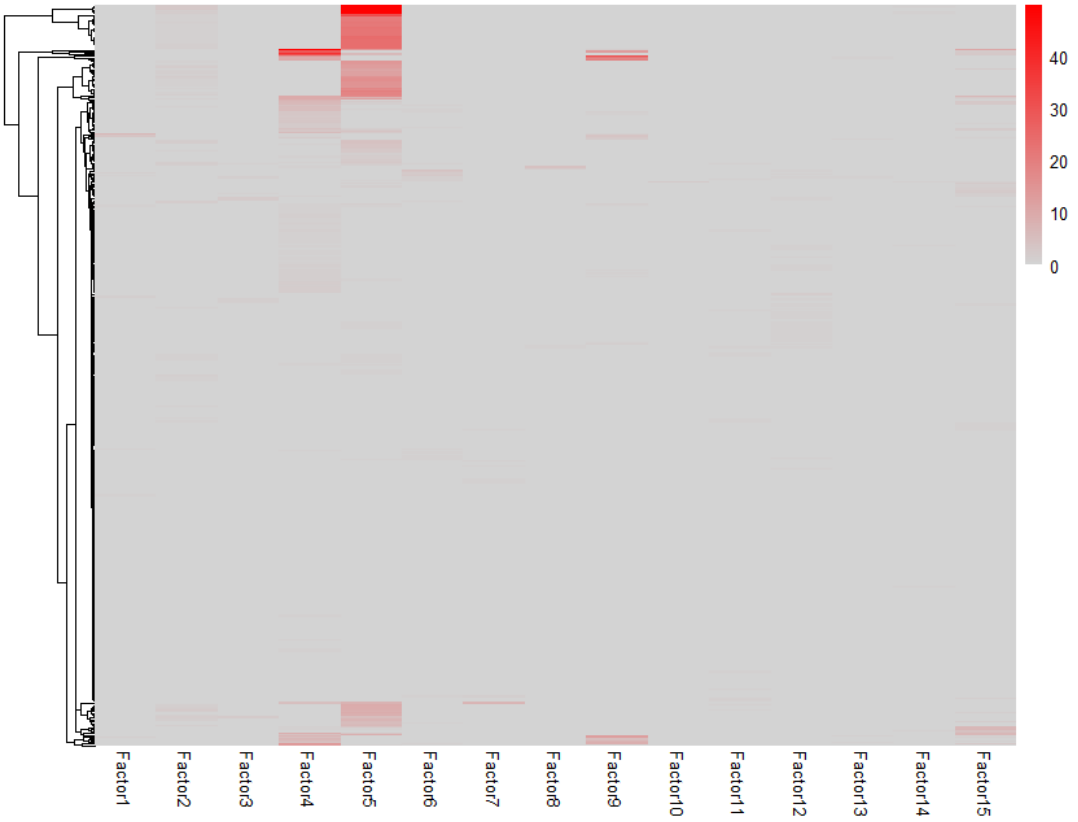
Εικόνα 34. Απεικόνιση των δειγμάτων ως προς τον τρίτο παράγοντα. Τα δείγματα με πράσινο χρώμα είναι αυτά που παρουσιάζουν τρισωμία 12, ενώ με κόκκινο χρώμα υποδηλώνεται η απουσία μετάλλαξης.

Στην Εικόνα 35 παρουσιάζεται η ταξινόμηση των δειγμάτων σε τέσσερις διαφορετικές κατηγορίες όσον αφορά τα δύο κύρια χαρακτηριστικά που έχουν τα μεγαλύτερα βάρη για τους παράγοντες 1 και 3. Η ταξινόμηση των δειγμάτων είναι βασισμένη στα διαφορετικού τύπου omics δεδομένα αφού οι παράγοντες 1,3 καταγράφουν τη μεταβλητότητα των δεδομένων έκφρασης και των σωματικών μεταλλάξεων (βλ. Εικόνα 20).



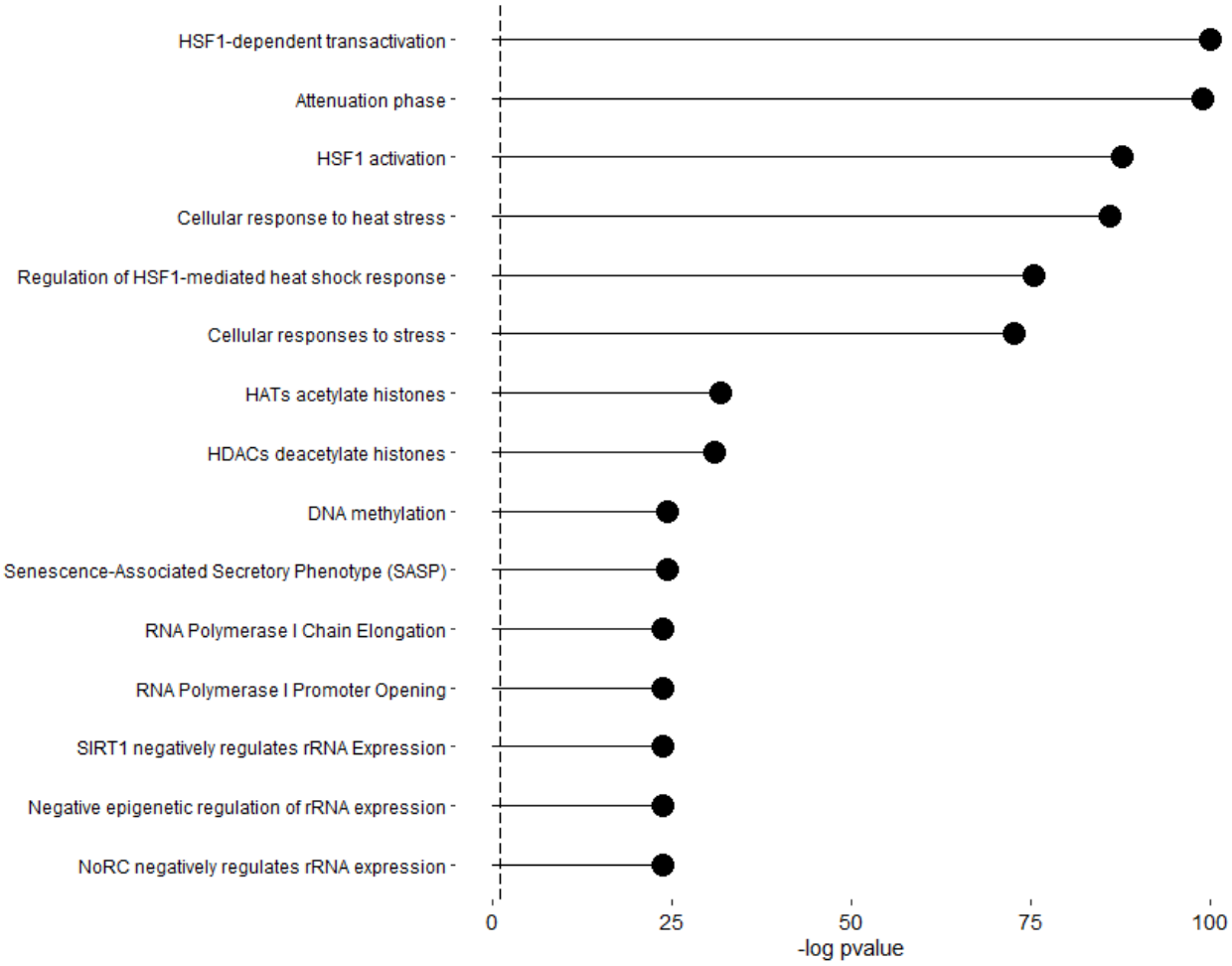
Εικόνα 35. Ταξινόμηση των δειγμάτων, βασισμένη στους παράγοντες 1 και 3, σε 4 διαφορετικές ομάδες.

Τέλος το μοντέλο παρέχει ανάλυση των βιολογικών μονοπατιών που είναι σημαντικά για κάθε παράγοντα. Για τα δεδομένα που αναλύθηκαν από το μοντέλο MOFA προέκυψε πως ο πέμπτος παράγοντας (factor 5) και τα γονίδια που έχουν θετικά βάρη σε αυτόν συμμετέχουν σε μεγάλο αριθμό βιολογικών μονοπατιών (Εικόνα 36).

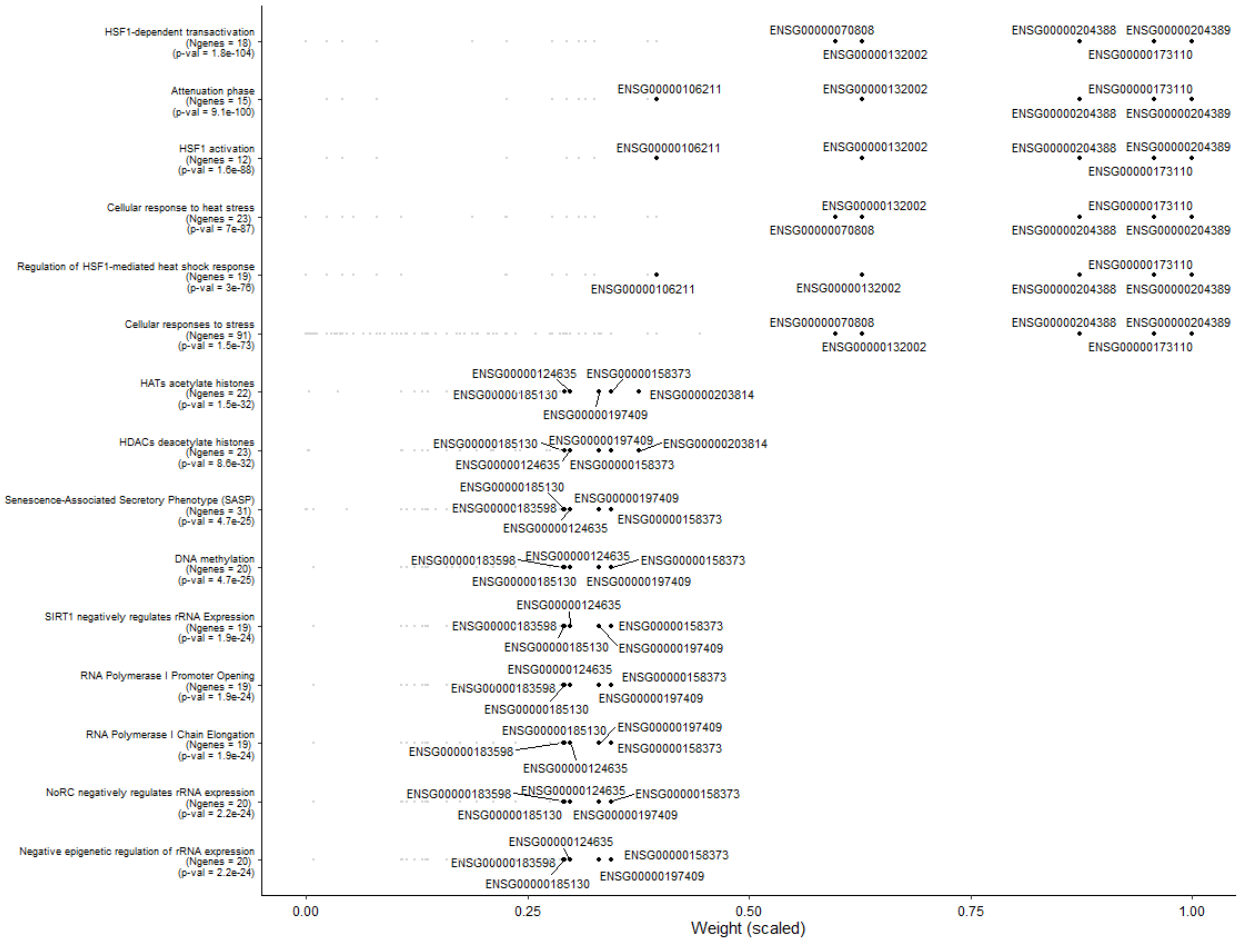


Εικόνα 36. Αριθμός βιολογικών μονοπατιών για κάθε παράγοντα του μοντέλου. Χρησιμοποιήθηκαν τα γονίδια που έχουν θετικό βάρος για κάθε παράγοντα.

Τα δεκαπέντε στατιστικά (p-value) πιο σημαντικά από τα μονοπάτια αυτά απεικονίζονται στην Εικόνα 37. Στη συνέχεια, υπάρχει η δυνατότητα να βρούμε τα γονίδια που συμμετέχουν σε κάθε μονοπάτι όπως φαίνεται στην Εικόνα 38 και με αυτόν τον τρόπο μπορούν να αποφευχθούν περιπτώσεις όπου σε πολύ μικρά μονοπάτια η συμμετοχή ενός μόνο γονιδίου, καθιστά το μονοπάτι στατιστικά σημαντικό.



Εικόνα 37. Τα δεκαπέντε (15) στατιστικά πιο σημαντικά βιολογικά μονοπάτια που σχετίζονται με τα γονίδια που είχαν θετικά βάρη στον πέμπτο παράγοντα.



Εικόνα 38. Τα μονοπάτια που είναι στατιστικά σημαντικά για τον πέμπτο παράγοντα καθώς και τα συμμετέχοντα γονίδια αυτού σε κάθε μονοπάτι.

3.4. iClusterPlus

3.4.1. Περιγραφή

Η κύρια ιδέα πίσω από τη δημιουργία του πακέτου iCluster είναι ότι οι διαφορετικοί τύποι όγκων μπορούν να μοντελοποιηθούν ως μη παρατηρούμενες (λανθάνουσες) μεταβλητές που μπορούν να εκτιμηθούν ταυτόχρονα από πολλαπλά omics δεδομένα όπως αυτά της γονιδιακής έκφρασης (mRNA), του αριθμού αντιγράφων DNA (CNV) και άλλους διαθέσιμους τύπους δεδομένων. Το μοντέλο που δημιουργείται στο πλαίσιο του iCLuster είναι ένα εννοιολογικά απλό και υπολογιστικά εφικτό μοντέλο που επιτρέπει ταυτόχρονη εξαγωγή συμπερασμάτων σε

οποιονδήποτε αριθμό και τύπο συνόλων δεδομένων. Το πακέτο iCluster επιτυγχάνει μία ολιστική προσέγγιση ομαδοποίησης (clustering) που επιτρέπει την εξαγωγή συνδυαστικών συμπερασμάτων από δεδομένα διαφορετικού τύπου και δημιουργεί τέτοια ομαδοποίηση ώστε αυτή να είναι συνεπής σε πολλούς τύπους δεδομένων.

3.4.2. Τύποι δεδομένων

Στο μοντέλο iClusterPlus, μπορούν να ενσωματωθούν δεδομένα omics χωρίς κάποιο περιορισμό όσον αφορά τον αριθμό διαφορετικού τύπου δεδομένων και μεταβλητών. Οι προϋποθέσεις αφορούν στην μορφή που έχουν τα δεδομένα. Οι δυαδικές παρατηρήσεις όπως οι σωματικές μεταλλάξεις πρέπει να έχουν μοντελοποιηθεί με βάση τη διωνυμική κατανομή, οι κατηγορικές μεταβλητές ως πολυωνυμικές τυχαίες μεταβλητές. Διακριτές μεταβλητές διαμορφώνονται ως τυχαίες διαδικασίες Poisson και συνεχείς μεταβλητές από Gaussian κατανομές.

3.4.3. Εφαρμογή και αποτελέσματα

Τα δεδομένα που ενσωματώθηκαν στο μοντέλο iCluster προέρχονται από σωματικές μεταλλάξεις, έκφραση γονιδίων και CNV (copy number variation) τα οποία αφορούν 84 δείγματα. Στους παρακάτω πίνακες φαίνεται η μορφή που έχουν τα δεδομένα για κάθε κατηγορία μετά την προεργασία τους.

	A2M	ABCC4	ADAMTSL3	ASXL1	BAI3	BCAR1	BCL11A	BCL11B
TCGA.02.0001	0	0	0	0	0	0	0	0
TCGA.02.0003	0	0	0	0	0	0	0	0
TCGA.02.0006	0	0	0	0	0	0	0	0
TCGA.02.0007	0	0	0	0	0	0	0	0
TCGA.02.0009	0	0	0	0	0	0	0	0
TCGA.02.0010	0	0	1	1	1	0	0	1
TCGA.02.0011	0	0	0	0	0	0	0	0
TCGA.02.0014	0	0	0	0	0	0	1	0
TCGA.02.0021	0	0	0	0	0	0	0	0
TCGA.02.0024	1	0	0	0	0	0	0	0

Πίνακας 9. Σωματικές μεταλλάξεις – Στον πίνακα αποτυπώνεται ένα υποσύνολο των ασθενών (84) και των μεταλλάξεων (306).

	FSTL1	MMP2	BBOX1	GCSH	EDN1	CXCR4	SALL1	MMP7
TCGA.02.0001	-0,66392	-0,27716	0,79896	0,09005	0,46557	0,30278	0,76869	0,55745
TCGA.02.0003	-0,28438	1,00445	0,19157	0,92115	1,08181	-0,0379	0,00452	-0,04971

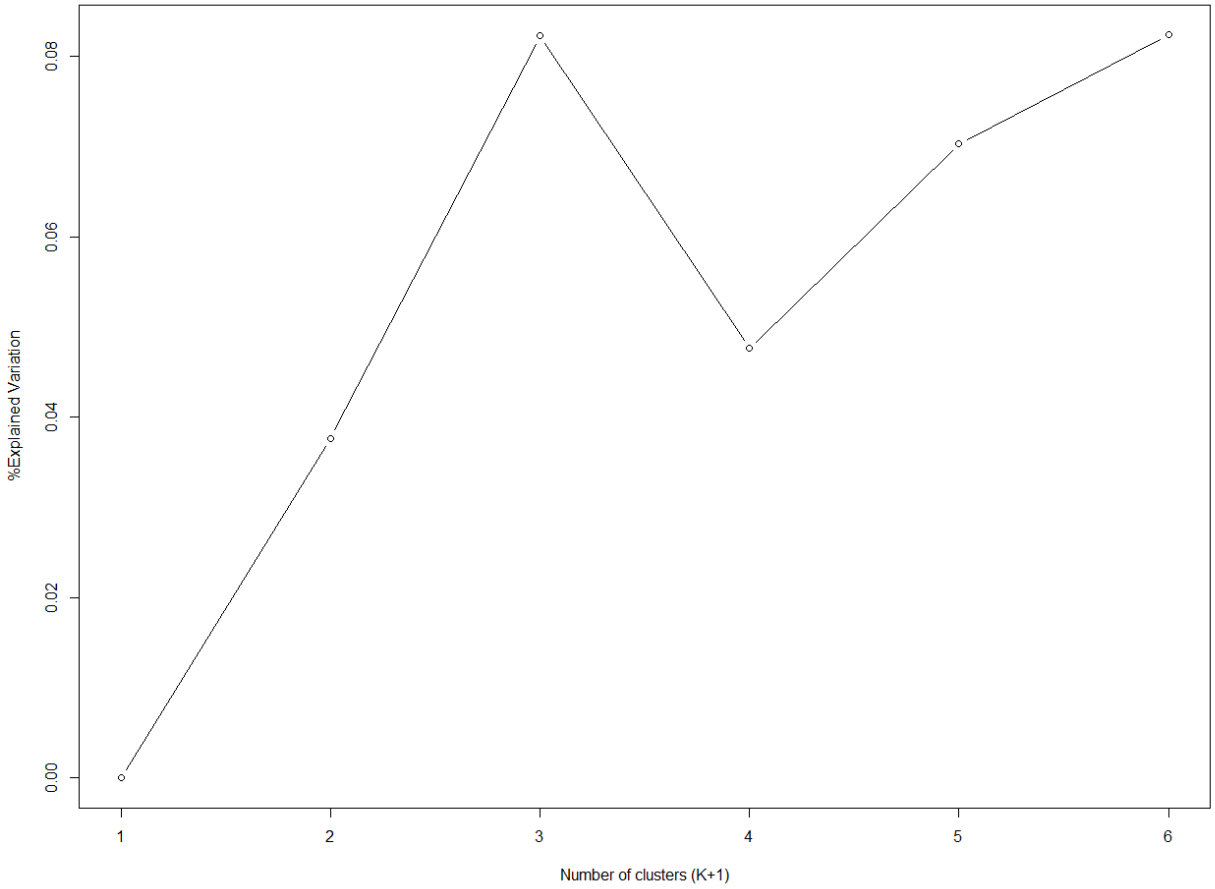
TCGA.02.0006	0,9889	0,19374	0,9383	0,49317	-0,22644	1,43145	-0,38401	1,58288
TCGA.02.0007	1,9823	-1,80818	0,58787	0,11215	-0,26708	0,3817	-0,47289	-1,6433
TCGA.02.0009	-0,22892	0,48973	0,99977	0,14226	0,80353	-0,54267	0,19177	1,28225
TCGA.02.0010	-1,52133	0,60094	-2,44385	0,30331	-0,33647	-1,62847	1,26128	1,77345
TCGA.02.0011	-1,0212	-2,06385	0,09309	0,50001	-0,01541	-1,72344	0,46225	-1,76488
TCGA.02.0014	-2,06326	0,66888	-3,11554	-0,03018	-0,9972	-3,0421	-0,14222	-1,86157
TCGA.02.0021	0,04085	1,76569	0,9542	0,59383	0,8455	0,1691	0,45267	-1,47079
TCGA.02.0024	-1,1537	0,94142	-2,52206	0,10764	-0,74209	-1,81867	0,52414	-0,13406

Πίνακας 10. Γονιδιακή έκφραση – Στον πίνακα αποτυπώνεται ένα υποσύνολο των ασθενών (84) και των γονιδίων (1740).

	chr1.554268-554287	chr1.554287-736483	chr1.736483-746956	chr1.746956-757922	chr1.757922-769590
TCGA.02.0001	0,2077	0,2077	0,2077	0,2077	0,2077
TCGA.02.0003	-0,0096	-0,0096	-0,0096	-0,0096	-0,0096
TCGA.02.0006	0,0027	0,0027	0,0027	0,0027	0,0027
TCGA.02.0007	-0,0107	-0,0107	-0,0107	-0,0107	-0,0107
TCGA.02.0009	-0,0052	-0,0052	-0,0052	-0,0052	-0,0052
TCGA.02.0010	-0,0588	-0,0588	-0,0588	-0,0588	-0,0588
TCGA.02.0011	0,0036	0,0036	0,0036	0,0036	0,0036
TCGA.02.0014	0,071	0,071	0,071	0,071	0,071
TCGA.02.0021	-0,0333	-0,0333	-0,0333	-0,0333	-0,0333
TCGA.02.0024	-0,3034	-0,3034	-0,3034	-0,3034	-0,3034

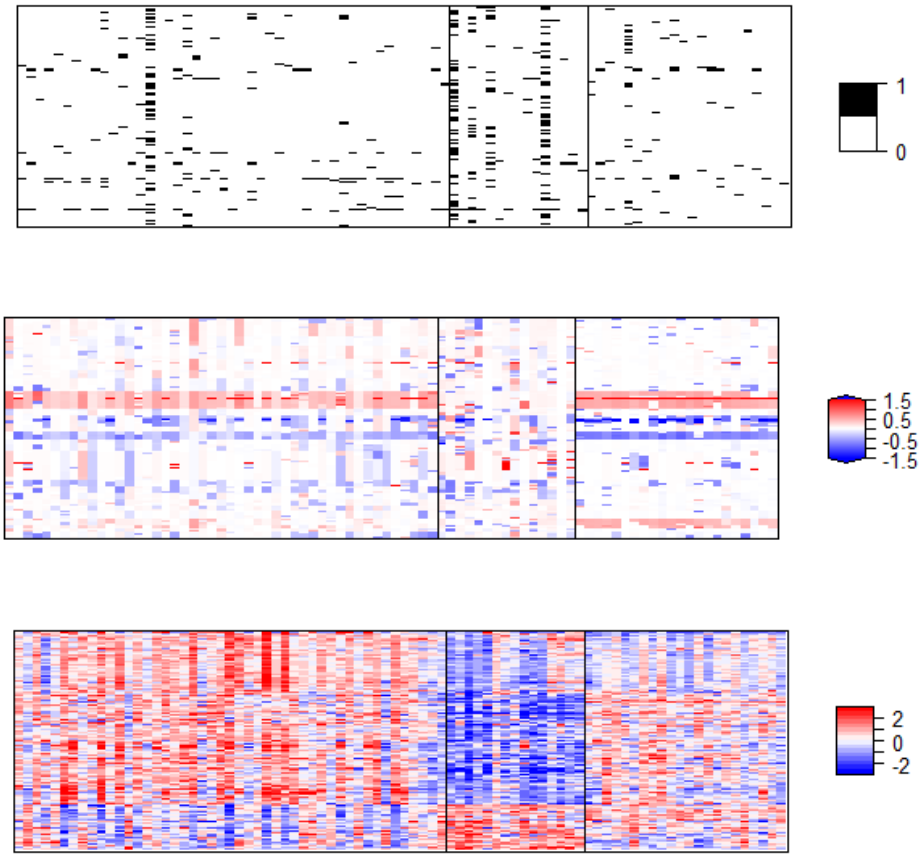
Πίνακας 11. Μεταβλητός αριθμός αντιγράφων DNA – Στον πίνακα αποτυπώνεται ένα υποσύνολο των ασθενών (84) και περιοχών του DNA (5512).

Στη συνέχεια ακολουθεί μία επαναληπτική διαδικασία εύρεσης του εύρεσης και επιλογής του καλύτερου μοντέλου, βασισμένη και στα τρία διαφορετικά σύνολα δεδομένων. Για τα δεδομένα που ενσωματώθηκαν ο βέλτιστος αριθμός ομάδων (clusters) ασθενών που λήφθηκε φαίνεται στην Εικόνα 39.



Εικόνα 39. Επεξηγούμενη μεταβλητότητα των δεδομένων ως προς τον αριθμό ομάδων (clusters).

Η επιλογή των σημαντικότερων χαρακτηριστικών τα οποία έχουν τη δυνατότητα να διαχωρίζουν τις ομάδες ασθενών με κοινά χαρακτηριστικά γίνεται στο επόμενο βήμα και τέλος δίνεται η δυνατότητα απεικόνισης των αποτελεσμάτων όπως φαίνεται στην Εικόνα 40. Τα σημαντικότερα χαρακτηριστικά που επιλέχθηκαν από το μοντέλο είναι 1378 περιοχές χρωμοσωμάτων και 435 γονίδια.



Εικόνα 40. Απεικόνιση της ομαδοποίησης των ασθενών και των σημαντικών χαρακτηριστικών για κάθε περίπτωση. Αρχικά η περίπτωση των σωματικών μεταλλάξεων, στο μέσο η περίπτωση CNV και τέλος η γονιδιακή έκφραση.

3.5.MOVICS

3.5.1. Περιγραφή

Το πακέτο MOVICS είναι το πιο ολοκληρωμένο καθώς περιέχει πλήθος εργαλείων και τεχνικών για την ανάλυση διαφορετικού τύπου omics δεδομένων. Ουσιαστικά, πρόκειται για μία συλλογή εργαλείων από άλλα πακέτα που χρησιμοποιούνται ευρέως και μεγάλη απήχηση. Σημαντικό στοιχείο που πρέπει να αναφερθεί είναι το γεγονός πως τα αποτελέσματα παράγονται συνδυαστικά με χρήση πολλών εργαλείων και κατά συνέπεια είναι πολλαπλώς επιβεβαιωμένα. Το πακέτο MOVICS υποστηρίζει την ανάλυση μέχρι έξι διαφορετικών συνόλων δεδομένων. Οι

δυνατότητες που προσφέρει αφορούν σε μεγάλο μέρος την ανάλυση δεδομένων που αφορούν τον καρκίνο αλλά πολλές τεχνικές είναι γενικευμένες και μπορούν να εφαρμοστούν και σε δεδομένα από άλλες ασθένειες. Περιλαμβάνει τεχνικές ανάλυσης διαφορικής έκφρασης γονιδίων από πακέτα όπως π.χ. limma, DESeq2, συγκριτικής μελέτης με πληροφορίες που αντλούνται από βάσεις δεδομένων όπως π.χ. για να βρεθούν τα βιολογικά μονοπάτια που σχετίζονται με τα αποτελέσματα, μεθόδους εύρεσης του βέλτιστου αριθμού clusters και τέλος πλήθος συναρτήσεων για την απεικόνιση των αποτελεσμάτων. Η παρούσα εργασία θα επικεντρωθεί μόνο στις τεχνικές που αφορούν την ενσωμάτωση και ανάλυση πολλαπλών omics συνόλων δεδομένων.

3.5.2. Τύποι δεδομένων

Τα δεδομένα που μπορούν να ενσωματωθούν στο πακέτο MOVICS για συνδυαστική ανάλυση, μπορεί να είναι δεδομένα γονιδιακής έκφρασης (mRNA), σωματικών μεταλλάξεων, πρωτεϊνών και γενικά δεν υπάρχει περιορισμός σε δεδομένα omics. Βασική προϋπόθεση είναι όλα τα δεδομένα που θα συμπεριληφθούν να αφορούν στα ίδια ακριβώς δείγματα καθώς και να μην υπερβαίνουν στο πλήθος τα έξι σύνολα.

3.5.3. Εφαρμογή και αποτελέσματα

Στην παρακάτω ανάλυση χρησιμοποιήθηκαν 643 δείγματα για τα οποία υπάρχουν διαθέσιμα δεδομένα γονιδιακής έκφρασης (mRNA), σωματικών μεταλλάξεων (somatic mutations), μεθυλίωσης (DNA methylation) και μεγάλου μήκους μη κωδικών RNA (lncRNA). Η μορφή των δεδομένων για κάθε περίπτωση φαίνεται στους επόμενους τέσσερις πίνακες.

	BRCA-A03L-01A	BRCA-A04R-01A	BRCA-A075-01A	BRCA-A08O-01A	BRCA-A0A6-01A	BRCA-A0AD-01A	BRCA-A0AU-01A	BRCA-A0AW-01A
SCGB2A2	9,2	1,42	7,24	2,41	13,97	12,65	2,58	0,67
SCGB1D2	10,11	1,95	5,88	2,2	12,27	12,75	2,48	0,41
PIP	4,54	2,59	4,35	1,97	11,03	1,35	2,67	7,66
TFF1	8,25	9,2	8,39	8,41	11,23	8,6	6,26	9,15
TFAP2B	0,64	0,36	6,39	0,84	6,32	3,86	0,04	5,96
MUCL1	4,77	1,99	4,73	1,57	5,89	0,14	3,12	6,08

NAT1	7,38	5,42	4,35	4,23	5,51	3,85	3,4	1,7
LTF	5,05	1,68	4,97	3,02	7,98	1,26	1,01	3,31
FABP4	2,51	3,56	3,76	5,66	4,21	3,99	4,87	4,39
SLC7A2	5,11	7,62	4,04	6,15	4,3	7,11	6,66	3,82

Πίνακας 12. MOVICS - Υπόδειγμα συνόλου δεδομένων mRNA.

	BRCA-A03L-01A	BRCA-A04R-01A	BRCA-A075-01A	BRCA-A08O-01A	BRCA-A0A6-01A	BRCA-A0AD-01A	BRCA-A0AU-01A	BRCA-A0AW-01A
RP11-20F24.2	4,08	0,54	4,47	6,39	6,63	5,81	1,58	4,91
RP11-20F24.4	3,53	0,45	4,49	6,45	6,3	5,66	1,86	5,65
RP11-206M11.7	8,18	1,89	1,45	5,84	1,61	0,51	0,34	8,35
AC096579.7	10,17	7,75	12,93	8,9	12,54	3,91	10,02	12,23
AC008268.1	1,22	4,17	1,01	2,97	1,87	3,71	0,92	4,33
RP11-13L2.4	4,23	6,53	3,14	6,57	3,95	3,57	3,75	0,99
RP11-431J24.2	3,98	4,26	3,98	2,12	1,53	0,04	6,41	0,24
RP11-731F5.2	3,68	0,71	5,43	1,74	6,62	1,38	5,91	6,9
RP11-321G12.1	2,29	2,74	3,45	2,67	1,67	1,12	2,58	0,54
AP004372.1	2,7	0,56	3,57	0,19	0,89	2,28	0	0

Πίνακας 13. MOVICS - Υπόδειγμα συνόλου δεδομένων lncRNA

	BRCA-A03L-01A	BRCA-A04R-01A	BRCA-A075-01A	BRCA-A08O-01A	BRCA-A0A6-01A	BRCA-A0AD-01A	BRCA-A0AU-01A	BRCA-A0AW-01A
TTBK1	0,499	0,809	0,19	0,044	0,776	0,382	0,262	0,666
VSTM2B	0,354	0,554	0,063	0,142	0,615	0,042	0,05	0,509
VWC2	0,597	0,774	0,342	0,12	0,65	0,648	0,023	0,361
EPSTI1	0,498	0,058	0,408	0,324	0,432	0,042	0,497	0,037
CPLX2	0,704	0,138	0,035	0,431	0,6	0,764	0,166	0,038
PROM1	0,603	0,803	0,024	0,683	0,569	0,359	0,641	0,431
ADAMTS12	0,124	0,807	0,043	0,792	0,649	0,323	0,665	0,656
CCDC8	0,193	0,088	0,35	0,58	0,208	0,029	0,599	0,65
FOXE3	0,12	0,751	0,022	0,177	0,46	0,652	0,02	0,586
SCGB3A1	0,58	0,266	0,04	0,629	0,401	0,576	0,61	0,182

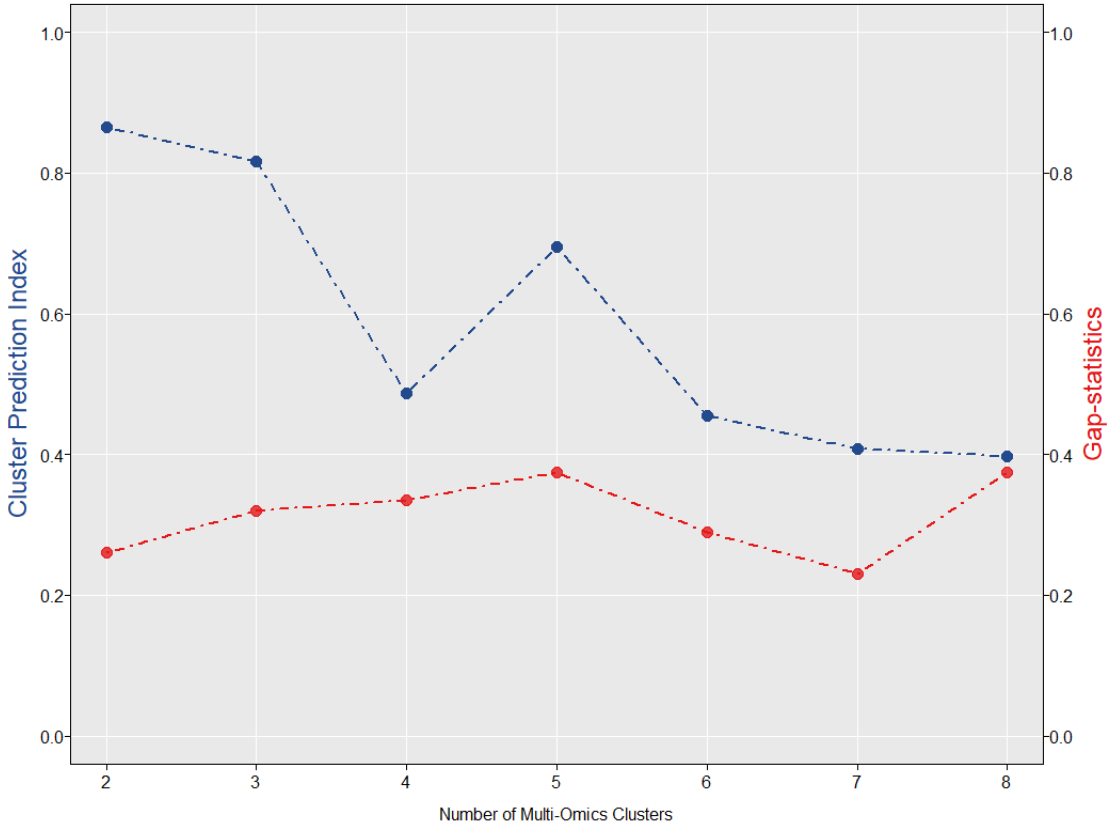
Πίνακας 14. MOVICS - Υπόδειγμα συνόλου δεδομένων methylation

	BRCA-A03L-01A	BRCA-A04R-01A	BRCA-A075-01A	BRCA-A08O-01A	BRCA-A0A6-01A	BRCA-A0AD-01A	BRCA-A0AU-01A	BRCA-A0AW-01A

PIK3CA	1	0	1	1	1	0	0	0
TP53	1	0	1	0	0	0	0	1
TTN	0	1	0	0	1	0	0	0
CDH1	0	0	0	0	1	0	0	0
GATA3	0	0	0	0	0	0	0	0
MLL3	0	0	0	0	1	0	0	0
MUC16	0	0	1	0	1	0	0	1
MAP3K1	0	0	0	0	0	0	0	0
SYNE1	0	0	0	0	1	0	0	1
MUC12	0	0	0	0	0	0	0	0

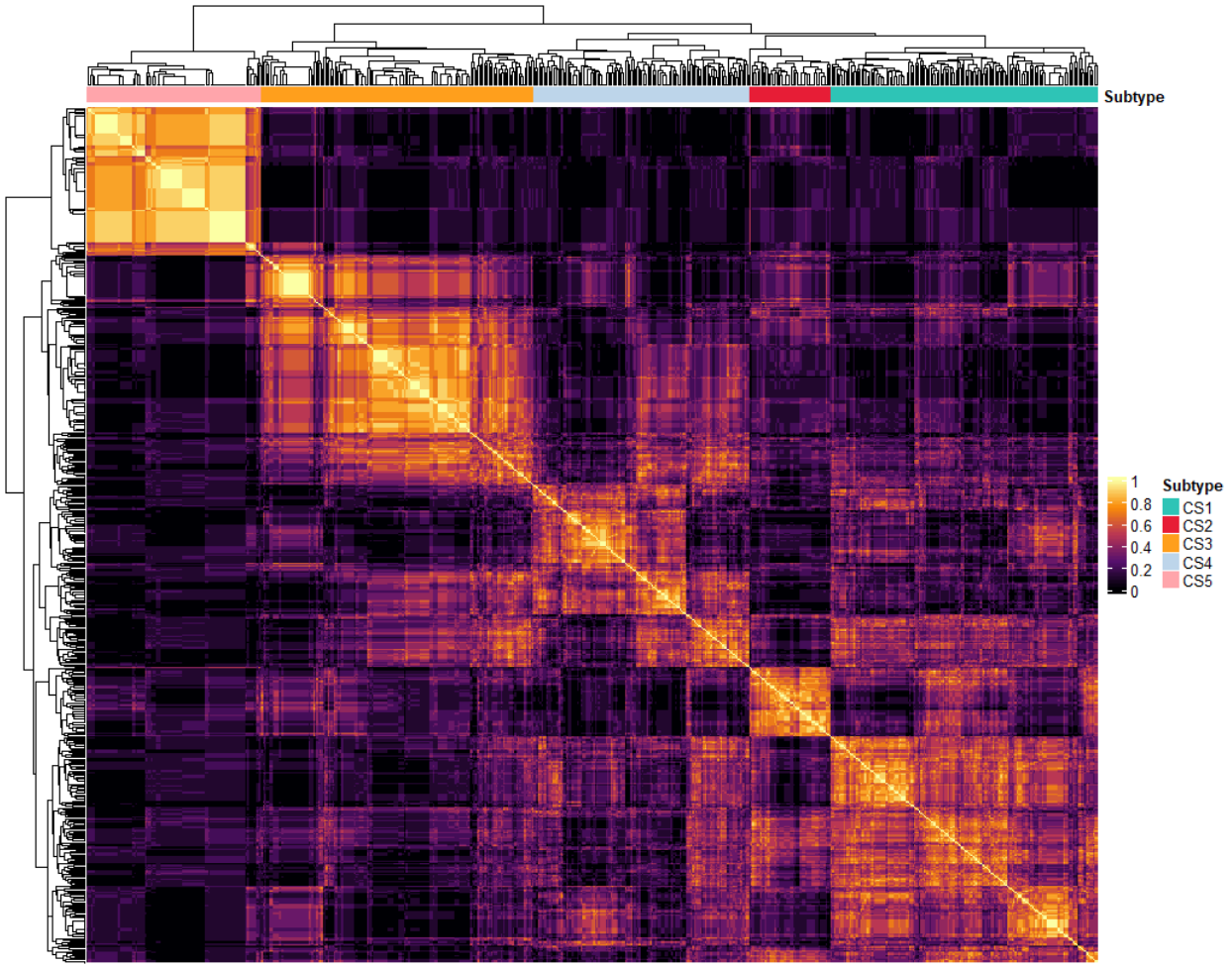
Πίνακας 15. MOVICS - Υπόδειγμα συνόλου δεδομένων somatic mutations

Η πιο σημαντική παράμετρος που πρέπει να υπολογιστεί αρχικά είναι ο βέλτιστος αριθμός ομάδων (clusters) στις οποίες θα διαχωριστούν τα δείγματα. Ο αριθμός αυτός πρέπει να είναι αφενός μικρός για να μειωθεί το φαινόμενου του θορύβου από τα δεδομένα αλλά συγχρόνως αρκετά μεγάλος ώστε να διατηρηθεί η σημαντική πληροφορία που περιέχεται σε αυτά. Το πακέτο MOVICS περιλαμβάνει δύο μεθόδους για την εύρεση του βέλτιστου αριθμού ο οποίος για τα συγκεκριμένα δεδομένα είναι πέντε καθώς στο σημείο αυτό συγκλίνουν οι τεχνικές όπως φαίνεται στην Εικόνα 41.



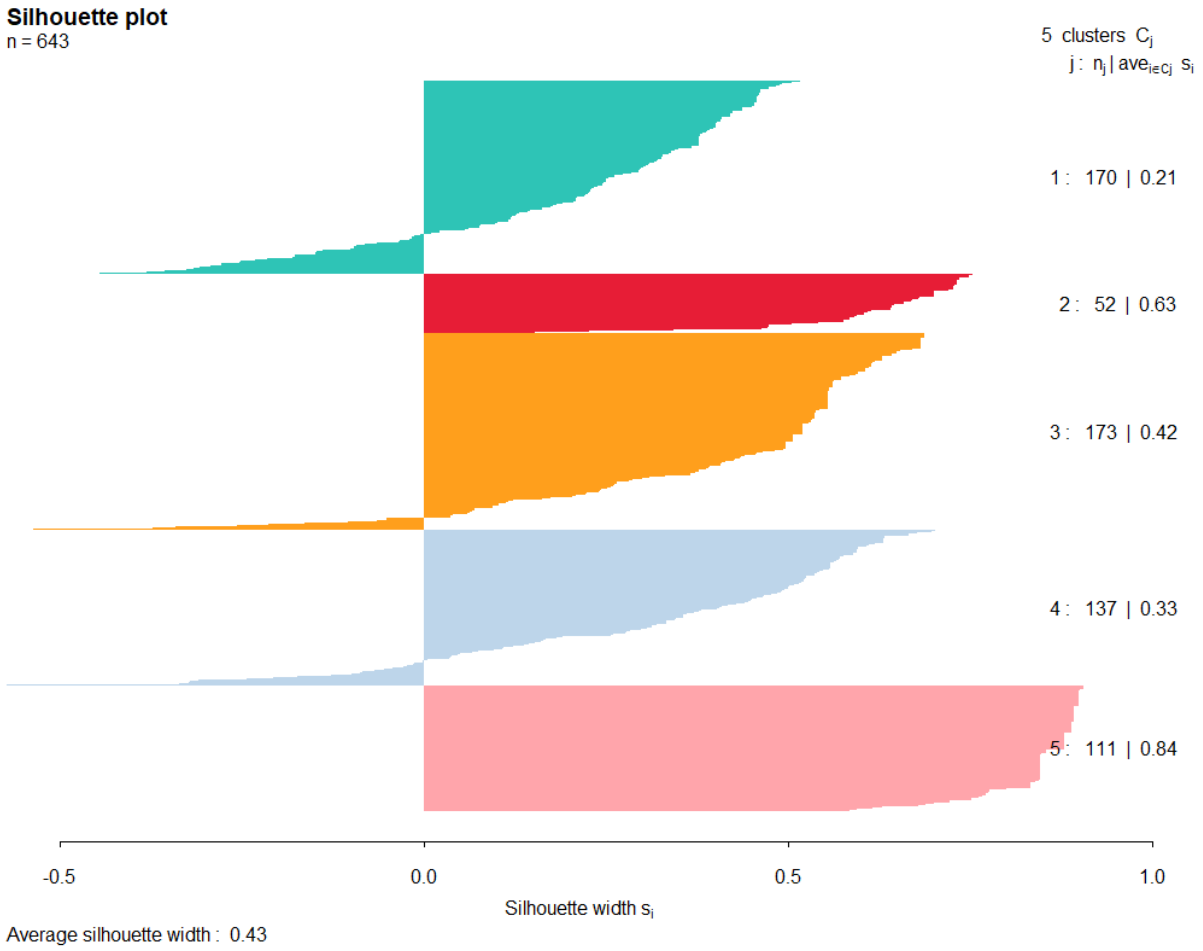
Εικόνα 41. Βέλτιστος αριθμός clusters υπολογισμένος με τη μέθοδο Gap-statistics και την Cluster Prediction Index.

Στο επόμενο βήμα εφαρμόζεται η ομαδοποίηση (clustering) των δειγμάτων από διαφορετικούς αλγορίθμους αλλά σε όλες τις περιπτώσεις χρησιμοποιείται ο αριθμός clusters που έχει επιλεχθεί προηγουμένως. Η Εικόνα 42 απεικονίζει τον συναινετικό (consensus) πίνακα που έχει προκύψει από τα αποτελέσματα διαφορετικών αλγορίθμων ομαδοποίησης που εφαρμόστηκαν. Για κάθε αλγόριθμο που εφαρμόστηκε υπολογίζεται ένα πίνακας $M_{n\times n}^{(t)}$, όπου n είναι ο αριθμός των δειγμάτων και t ο μοναδικός δείκτης κάθε αλγορίθμου με $2 < t_{max} < 10$. Κάθε στοιχείο του πίνακα $M_{ij}^{(t)} = 1$ εάν τα δείγματα έχουν κατηγοριοποιηθεί στην ίδια ομάδα ειδάλλως είναι ίσο με 0. Αφού υπολογιστούν όλοι οι πίνακες $M^{(t)}$, υπολογίζεται ο τελικός πίνακας $CM = \sum_1^{t_{max}} M^t$ (consensus matrix), κάθε στοιχείο cm_{ij} του οποίου όπως προκύπτει από τα παραπάνω θα ανήκει στο διάστημα $[0,10]$. Ο πίνακας που προκύπτει και απεικονίζεται στην παρακάτω εικόνα, είναι ένας πίνακας πιθανοτήτων που αναπαριστά την πιθανότητα δειγμάτων που ανήκουν στην ίδια ομάδα να ομαδοποιηθούν μαζί.



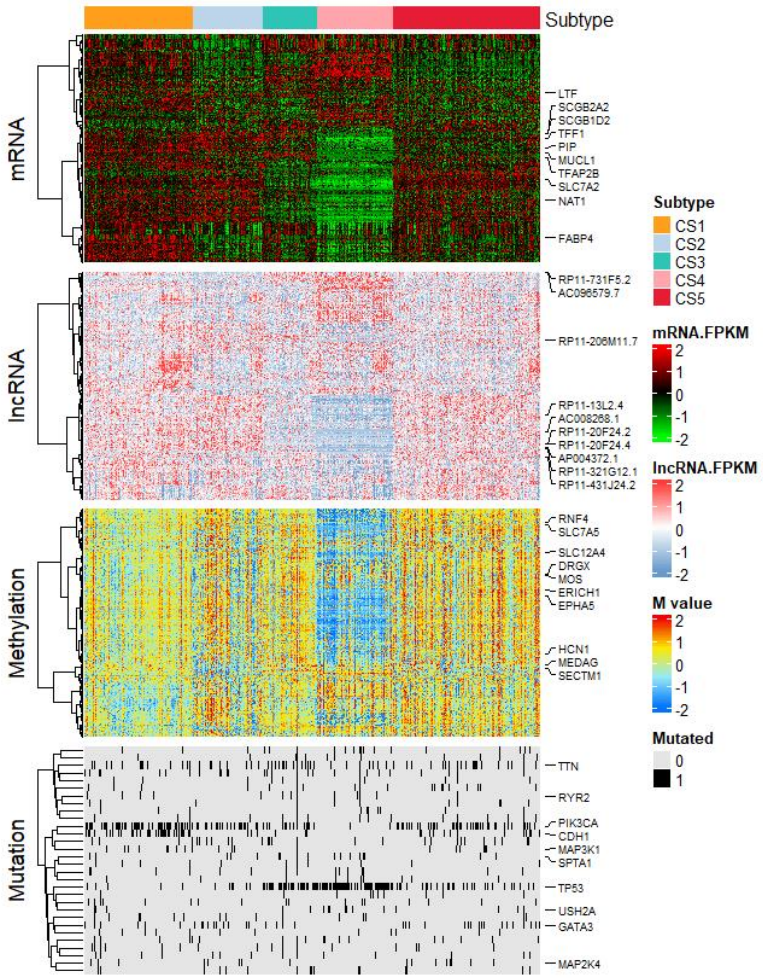
Εικόνα 42. Συναινετικός (consensus) πίνακας πιθανοτήτων ομαδοποίησης των δειγμάτων. Μεγάλες τιμές του πίνακα στην κύρια διαγώνιο υποδηλώνουν πως τα αποτελέσματα ομαδοποίησης των διαφορετικών αλγορίθμων είναι παρόμοια.

Ένας ακόμα τρόπος για να επαληθευτεί η ομαδοποίηση που προέκυψε παραπάνω, είναι η εφαρμογή της μεθόδου Silhouette (παρέχεται από το πακέτο MOVICS) η οποία δίνει τη δυνατότητα ποσοτικοποίησης και οπτικοποίησης της ομοιότητας των δειγμάτων για τη συγκεκριμένη ομαδοποίηση. Στην Εικόνα 43 παρουσιάζονται τα αποτελέσματα της μεθόδου Silhouette.



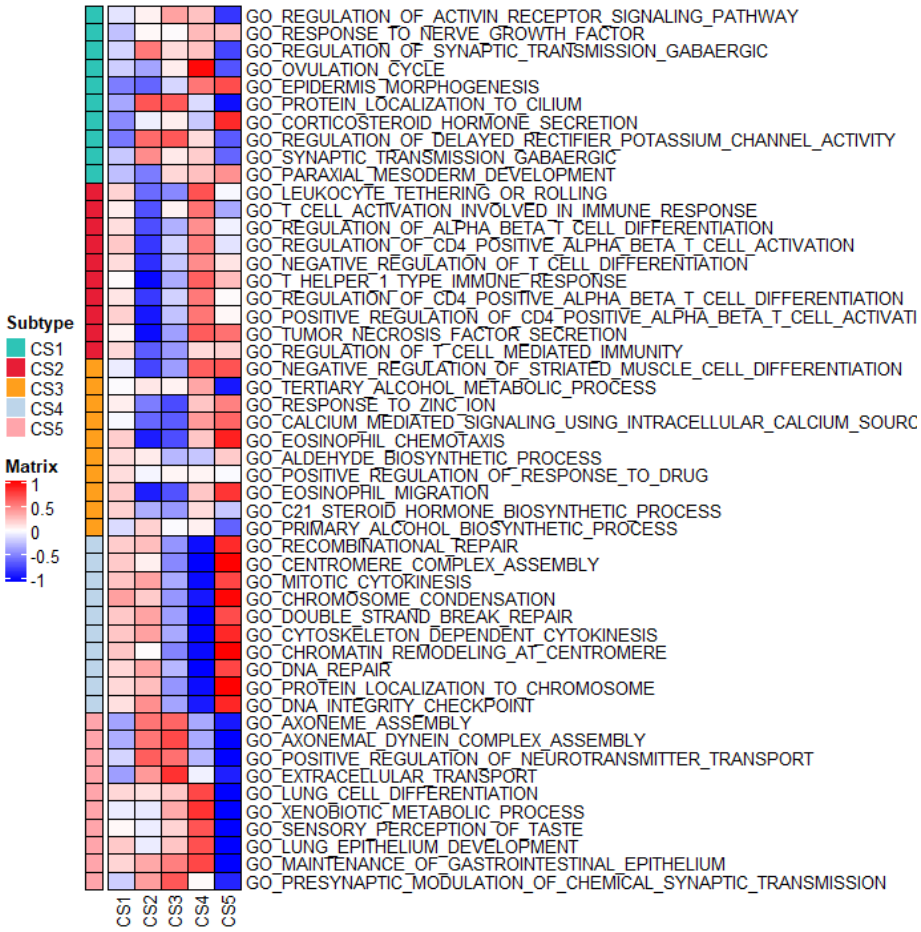
Εικόνα 43. Μέθοδος Silhouette – Ο αριθμός n_jδηλώνει το μέγεθος του κάθε cluster και ο αριθμός $ave_{i\epsilon Ci}s_i$ τη βαθμολογία του.

Για να δημιουργηθεί το heatmap στην Εικόνα 44 πρέπει να γίνει μία προεργασία στα δεδομένα. Συγκεκριμένα, τα δεδομένα που αφορούν έκφραση (π.χ. γονιδίων ή πρωτεϊνών) πρέπει να είναι κανονικοποιημένα, τα δεδομένα μεθυλίωσης DNA είναι καλό να μετατρέπονται από β σε M τιμές σύμφωνα με τη σχέση $M = log_2 \frac{\beta}{1-\beta}$ καθώς αυτή η μορφή είναι ιδανικότερη για οπτικοποίηση και τα δεδομένα σωματικών μεταλλάξεων (τιμές 0 και 1) να παραμένουν σε δυαδική μορφή. Μετά την προετοιμασία των δεδομένων επιλέγεται ο αλγόριθμος για τον οποίο θα κατασκευαστεί το heatmap της παρακάτω εικόνας.



Εικόνα 44. Heatmap δειγμάτων για κάθε κατηγορία omics δεδομένων. Η αναγραφή στοιχείων των γραμμών του σχήματος αφορά εκείνα τα οποία έχουν επιλεγεί από κάθε κατηγορία ως τα πιο σημαντικά (feature selection).

Τέλος ένα εργαλείο που έχει ενσωματωθεί στο πακέτο MOVICS είναι αυτό της εύρεσης των βιολογικών διαδικασιών που σχετίζονται με συγκεκριμένα γονίδια (gene set enrichment analysis). Αφού προηγηθεί η μελέτη των γονιδίων που διαφοροποιούνται μεταξύ των clusters που έχουν προκύψει μπορούν αυτά τα γονίδια να συσχετιστούν με συγκεκριμένες βιολογικές διαδικασίες. Η διαφορική ανάλυση έκφρασης (differential expression analysis) μπορεί να διενεργηθεί με χρήση τριών διαφορετικών πακέτων (DESeq2, limma, edgeR), συναρτήσεις των οποίων περιέχονται στο πακέτο MOVICS. Έτσι προκύπτουν τα αποτελέσματα στην Εικόνα 45 όπου φαίνονται οι διαδικασίες που σχετίζονται με κάθε cluster.



Εικόνα 45. Heatmap των βιολογικών μονοπατιών που είναι υπερ-ρυθμισμένα (upregulated) για κάθε ένα από τα clusters της ανάλυσης.

4. Συμπεράσματα

Όπως αναδείχθηκε στο κύριο μέρος της παρούσας εργασίας, στις επιστήμες της Ιατρικής και της Βιολογίας έχουν επέλθει αλλαγές οι οποίες σε μεγάλο βαθμό οφείλονται στην ανάπτυξη νέων τεχνολογιών. Με τη χρήση των νέων τεχνολογιών είναι εφικτή σήμερα η συλλογή μεγάλου όγκου δεδομένων που αφορούν βιολογικά συστήματα και ασθένειες με αρκετά μικρότερο κόστος σε σχέση με τις προηγούμενες δεκαετίες. Ο όγκος της πληροφορίας που μπορεί να εξαχθεί από τα διαθέσιμα δεδομένα απαιτεί και αντίστοιχα εξελιγμένα και προσαρμοσμένα εργαλεία που συνήθως προέρχονται από το πεδίο της Βιοπληροφορικής για τη διαχείριση και την ανάλυση των δεδομένων αυτών. Τα τελευταία χρόνια αναπτύσσονται εργαλεία, εκτός αυτών που εφαρμόζονται στη μεμονωμένη ανάλυση κάποιου συγκεκριμένου τύπου δεδομένων (π.χ. mRNA), τα οποία παρέχουν τη δυνατότητα στον ερευνητή να αναλύσει δεδομένα διαφορετικού τύπου με συνδυαστικό τρόπο με στόχο την καλύτερη κατανόηση του υπό μελέτη συστήματος με μία πιο ολιστική προσέγγιση.

Τα εργαλεία που άπτονται της συνδυαστικής ανάλυσης διαφορετικών τύπων omics δεδομένων (multi-omics integrative analysis) και τα οποία εξετάστηκαν στην παρούσα εργασία, έχουν συνήθως διττό ρόλο. Για τις περιπτώσεις όπου δεν είναι γνωστή η κατηγοριοποίηση των δειγμάτων, με χρήση μη-εποπτευόμενων τεχνικών (unsupervised), δίνεται η δυνατότητα να ομαδοποιηθούν αυτά σε συγκεκριμένες κατηγορίες οι οποίες έχουν προέλθει συγκρίνοντας ως προς την ομοιότητα τις μεταβλητές που είναι διαθέσιμες στα διαφορετικά omics δεδομένα. Όταν είναι εκ των προτέρων γνωστή η ομαδοποίηση των δειγμάτων (π.χ. ασθενείς και υγιείς), εποπτευόμενων τεχνικών (supervised), μπορούν να εξαχθούν συμπεράσματα για την σημαντικότητα των μεταβλητών οι οποίες διαφοροποιούνται ανάμεσα στις ομάδες δειγμάτων.

Τα εργαλεία που ελέγχθηκαν είναι αυτά των πακέτων mixOmics, MOFA, iCluster και MOVICS στην R.

Ως προς τον τύπο των omics δεδομένων που μπορούν να ενσωματώσουν τα παραπάνω πακέτα δεν παρουσιάζουν σημαντικές διαφορές καθώς όλα έχουν τη δυνατότητα να δεχθούν τους πιο ευρέως χρησιμοποιούμενους τύπους. Ως προς τον αριθμό των διαφορετικών συνόλων που

μπορούν να εισαχθούν σε κάθε πακέτο, δεν υπάρχει κάποιος περιορισμός με μόνη εξαίρεση το πακέτο MOVICS που μπορεί να χειριστεί έως έξι διαφορετικά σύνολα. Η σωστή προετοιμασία έχει καθοριστικό ρόλο στην εισαγωγή των δεδομένων και την εφαρμογή του κάθε πακέτου. Το πακέτο MOFA είναι το μοναδικό που μπορεί να χειριστεί σύνολα δεδομένων για τα οποία τα δείγματα δεν είναι πλήρως αλληλεπικαλυπτόμενα. Στα υπόλοιπα εργαλεία, τα δείγματα πρέπει να είναι ακριβώς τα ίδια για τα διαφορετικά σύνολα δεδομένων που εισάγονται.

Οι χρόνοι που χρειάστηκαν για την υλοποίηση της ανάλυσης για κάθε πακέτο είναι 31.5 sec για το πακέτο mixOmics, 45 sec για το πακέτο MOFA, 425 sec για το πακέτο iCLuster και 1.62 hr για το πακέτο MOVICS. Οι αναλύσεις έγιναν σε υπολογιστή με που έφερε επεξεργαστή Ryzen 5 3600XT 6-Core Processor 3.80 GHz, μνήμη 64GB, SSD αποθηκευτικό μέσο και λειτουργικό σύστημα Windows 10 pro. Οι παραπάνω χρόνοι δεν περιλαμβάνουν την προετοιμασία των δεδομένων καθώς και την παραμετροποίηση των μοντέλων.

Η περιγραφή των αποτελεσμάτων καθώς και η κατανόηση τους διαφέρει σημαντικά ανάμεσα στα πακέτα. Τα πιο ολοκληρωμένα όσον αφορά την απεικόνιση των αποτελεσμάτων είναι τα πακέτα mixOmics και MOFA καθώς περιέχουν πλήθος συναρτήσεων για απεικόνιση. Όλα τα πακέτα, εκτός του iCLuster, περιέχουν συναρτήσεις που στόχο έχουν την προβλεπτική διαδικασία για νέα δείγματα των οποίων η κατάσταση δεν είναι γνωστή και θα προβλεφθεί σύμφωνα με τα αποτελέσματα του κάθε μοντέλου. Ακόμα ένα σημαντικό χαρακτηριστικό των πακέτων MOFA και MOVICS είναι η δυνατότητα εξερεύνησης των βιολογικών μονοπατιών που σχετίζονται με τα αποτελέσματα, αφού ενσωματώνουν σχετικές συναρτήσεις. Σημαντικό στοιχείο που διαχωρίζει το πακέτο MOVICS είναι πως αποτελεί συλλογή μεθοδολογιών οι οποίες χρησιμοποιούνται ευρέως και παρέχεται η δυνατότητα χρήσης τους είτε μεμονωμένα είτε συνδυαστικά με στόχο την παραγωγή περισσότερο εύρωστων (robust) αποτελεσμάτων.

Ένα βασικό μειονέκτημα όλων των πακέτων είναι η ελλιπής περιγραφή των συναρτήσεων και ειδικότερα των παραμέτρων που απαιτείται να οριστούν για την κατασκευή του βέλτιστου μοντέλου. Επίσης τα παραπάνω πακέτα δίνουν τη δυνατότητα για ανάλυση μόνο σε ερευνητές οι οποίοι ασχολούνται με τον προγραμματισμό και δεν υπάρχει κάποιο γραφικό περιβάλλον που να απλοποιεί τη διαδικασία για μη εξειδικευμένους χρήστες.

5. Βιβλιογραφία

1. Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, Krueger F, Smallwood S, Ponting CP, Voet T, Kelsey G, Stegle O, Reik W. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. Nat Methods. 2016 Mar;13(3):229-232. doi: 10.1038/nmeth.3728. Epub 2016 Jan 11. PMID: 26752769; PMCID: PMC4770512
2. Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., & Stegle, O. (2020). MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biology 2020 21:1, 21(1), 1–17. https://doi.org/10.1186/S13059-020-02015-1
3. Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., & Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Molecular Systems Biology, 14(6). https://doi.org/10.15252/msb.20178124
4. Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., & Milanesi, L. (2016). Methods for the integration of multi-omics data: Mathematical aspects. BMC Bioinformatics, 17(2), S15. https://doi.org/10.1186/s12859-015-0857-9
5. Canzler, S., Schor, J., Busch, W., Schubert, K., Rolle-Kampczyk, U. E., Seitz, H., Kamp, H., von Bergen, M., Buesen, R., & Hackermüller, J. (2020). Prospects and challenges of multi-omics data integration in toxicology. In Archives of Toxicology (Vol. 94, Issue 2, pp. 371–388). Springer. https://doi.org/10.1007/s00204-020-02656-y
6. Caspar SM, Dubacher N, Kopps AM, Meienberg J, Henggeler C, Matyas G. Clinical sequencing: From raw data to diagnosis with lifetime value. Clin Genet. 2018 Mar;93(3):508-519. doi: 10.1111/cge.13190. PMID: 29206278
7. Chen, D., Zhang, F., Zhao, Q., & Xu, J. (2019). OmicsARules: a R package for integration of multi-omics datasets via association rules mining. BMC Bioinformatics, 20(1), 554. https://doi.org/10.1186/s12859-019-3171-0
8. Chen, L., Ge, B., Casale, F. P., Downes, K., & Pastinen, T. (2016). Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell, 167, 1398–1414. https://doi.org/10.1016/j.cell.2016.10.026
9. Chierici, M., Bussola, N., Marcolini, A., Francescatto, M., Zandonà, A., Trastulla, L., Agostinelli, C., Jurman, G., & Furlanello, C. (2020). Integrative Network Fusion: A Multi-Omics Approach in Molecular Profiling. Frontiers in Oncology, 10, 1065. https://doi.org/10.3389/fonc.2020.01065
10. Colomé-Tatché, M., & Theis, F. J. (2018). Statistical single cell multi-omics integration. Current Opinion in Systems Biology, 7, 54–59. https://doi.org/10.1016/J.COISB.2018.01.003
11. D'Argenio, V. (2018). The High-Throughput Analyses Era: Are We Ready for the Data Struggle? High-Throughput, 7(1). https://doi.org/10.3390/HT7010008

12. D'Argenio V, Esposito MV, Telese A, Precone V, Starnone F, Nunziato M, Cantiello P, Iorio M, Evangelista E, D'Aiuto M, Calabrese A, Frisso G, D'Aiuto G, Salvatore F. The molecular analysis of BRCA1 and BRCA2: Next-generation sequencing supersedes conventional approaches. Clin Chim Acta. 2015 Jun 15;446:221-5. doi: 10.1016/j.cca.2015.03.045. Epub 2015 Apr 17. PMID: 25896959
13. D'Argenio V, Salvatore F. The role of the gut microbiome in the healthy adult status. Clin Chim Acta. 2015 Dec 7;451(Pt A):97-102. doi: 10.1016/j.cca.2015.01.003. Epub 2015 Jan 10. PMID: 25584460
14. D'Argenio V, Frisso G, Precone V, Boccia A, Fienga A, Pacileo G, Limongelli G, Paolella G, Calabrò R, Salvatore F. DNA sequence capture and next-generation sequencing for the molecular diagnosis of genetic cardiomyopathies. J Mol Diagn. 2014 Jan;16(1):32-44. doi: 10.1016/j.jmoldx.2013.07.008. Epub 2013 Oct 31. PMID: 24183960
15. D'Argenio V, Notomista E, Petrillo M, Cantiello P, Cafaro V, Izzo V, Naso B, Cozzuto L, Durante L, Troncone L, Paolella G, Salvatore F, Di Donato A. Complete sequencing of Novosphingobium sp. PP1Y reveals a biotechnologically meaningful metabolic pattern. BMC Genomics. 2014 May 19;15(1):384. doi: 10.1186/1471-2164-15-384. PMID: 24884518; PMCID: PMC4059872
16. D'Argenio V, Torino M, Precone V, Casaburi G, Esposito MV, Iaffaldano L, Malapelle U, Troncone G, Coto I, Cavalcanti P, De Rosa G, Salvatore F, Sacchetti L. The Cause of Death of a Child in the 18th Century Solved by Bone Microbiome Typing Using Laser Microdissection and Next Generation Sequencing. Int J Mol Sci. 2017 Jan 6;18(1):109. doi: 10.3390/ijms18010109. PMID: 28067829; PMCID: PMC5297743
17. Fondi, M., & Liò, P. (2015). Multi -omics and metabolic modelling pipelines: Challenges and tools for systems microbiology. In Microbiological Research (Vol. 171, pp. 52–64). Elsevier GmbH. https://doi.org/10.1016/j.micres.2015.01.003
18. Gerstung, M., Pellagatti, A., Malcovati, L., Giagounidis, A., Porta, M. G. Della, Jädersten, M., Dolatshad, H., Verma, A., Cross, N. C. P., Vyas, P., Killick, S., Hellström-Lindberg, E., Cazzola, M., Papaemmanuil, E., Campbell, P. J., & Boultwood, J. (2015). Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. Nature Communications 2015 6:1, 6(1), 1–11. https://doi.org/10.1038/ncomms6901
19. Ghaemi, M. S., DiGiulio, D. B., Contrepois, K., Callahan, B., Ngo, T. T. M., Lee-Mcmullen, B., Lehallier, B., Robaczewska, A., McIlwain, D., Rosenberg-Hasson, Y., Wong, R. J., Quaintance, C., Culos, A., Stanley, N., Tanada, A., Tsai, A., Gaudilliere, D., Ganio, E., Han, X., … Aghaeepour, N. (2019). Multiomics modeling of the immunome, transcriptome, microbiome, proteome and metabolome adaptations during human pregnancy. Bioinformatics, 35(1), 95–103. https://doi.org/10.1093/bioinformatics/bty537
20. Gligorijević, V., & Pržulj, N. (2015). Methods for biological data integration: Perspectives and challenges. In Journal of the Royal Society Interface (Vol. 12, Issue 112). Royal Society of London. https://doi.org/10.1098/rsif.2015.0571

21. Haas, R., Zelezniak, A., Iacovacci, J., Kamrad, S., Townsend, S. J., & Ralser, M. (2017). Designing and interpreting "multi-omic" experiments that may change our understanding of biology. In Current Opinion in Systems Biology (Vol. 6, pp. 37–45). Elsevier Ltd. https://doi.org/10.1016/j.coisb.2017.08.009
22. Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. In Genome Biology (Vol. 18, Issue 1, pp. 1–15). BioMed Central Ltd. https://doi.org/10.1186/s13059-017-1215-1
23. Hayden EC. Technology: The $1,000 genome. Nature. 2014 Mar 20;507(7492):294-5. doi: 10.1038/507294a. PMID: 24646979
24. Huang CE, Ma GC, Jou HJ, Lin WH, Lee DJ, Lin YS, Ginsberg NA, Chen HF, Chang FM, Chen M. Noninvasive prenatal diagnosis of fetal aneuploidy by circulating fetal nucleated red blood cells and extravillous trophoblasts using silicon-based nanostructured microfluidics. Mol Cytogenet. 2017 Dec 2; 10:44. doi: 10.1186/s13039-017-0343-3. PMID: 29213331; PMCID: PMC5712079
25. Huang, S., Chaudhary, K., & Garmire, L. X. (2017). More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Frontiers in Genetics, 8(JUN), 84. https://doi.org/10.3389/fgene.2017.00084
26. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 431, 931–945 (2004). https://doi.org/10.1038/nature03001
27. Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., Cokelaer, T., Greninger, P., Dyk, E. van, Chang, H., Silva, H. de, Heyn, H., Deng, X., Egan, R. K., Liu, Q., … Garnett, M. J. (2016). A Landscape of Pharmacogenomic Interactions in Cancer. Cell, 166(3), 740–754. https://doi.org/10.1016/J.CELL.2016.06.017
28. Jacob M, Lopata AL, Dasouki M, Abdel Rahman AM. Metabolomics toward personalized medicine. Mass Spectrom Rev. 2019 May;38(3):221-238. doi: 10.1002/mas.21548. Epub 2017 Oct 26. PMID: 29073341
29. Jiang, Y., Liang, Y., Wang, D., Xu, D., & Joshi, T. (2017). IMPRes: Integrative MultiOmics pathway resolution algorithm and tool. 2260–2260. https://doi.org/10.1109/bibm.2017.8218016
30. Kalsner L, Twachtman-Bassett J, Tokarski K, Stanley C, Dumont-Mathieu T, Cotney J, Chamberlain S. Genetic testing including targeted gene panel in a diverse clinical population of children with autism spectrum disorder: Findings and implications. Mol Genet Genomic Med. 2018 Mar;6(2):171-185. doi: 10.1002/mgg3.354. Epub 2017 Dec 21. PMID: 29271092; PMCID: PMC5902398
31. Kim, K., Gitlin, L. N., & Han, H.-R. (2016). Kim et al. Respond. American Journal of Public Health, 106(8), e10. https://doi.org/10.2105/AJPH.2016.303276
32. Kim, S., Herazo-Maya, J. D., Kang, D. D., Juan-Guardela, B. M., Tedrow, J., Martinez, F. J., Sciurba, F. C., Tseng, G. C., & Kaminski, N. (2015). Integrative phenotyping framework

(iPF): integrative clustering of multiple omics data identifies novel lung disease subphenotypes. BMC Genomics, 16(1), 924. https://doi.org/10.1186/s12864-015-2170-4
33. Koh, H. W. L., Fermin, D., Vogel, C., Choi, K. P., Ewing, R. M., & Choi, H. (2019). iOmicsPASS: network-based integration of multiomics data for predictive subnetwork discovery. Npj Systems Biology and Applications, 5(1), 1–10. https://doi.org/10.1038/s41540-019-0099-y
34. Kohl, M., Megger, D. A., Trippler, M., Meckel, H., Ahrens, M., Bracht, T., Weber, F., Hoffmann, A. C., Baba, H. A., Sitek, B., Schlaak, J. F., Meyer, H. E., Stephan, C., & Eisenacher, M. (2014). A practical data processing workflow for multi-OMICS projects. Biochimica et Biophysica Acta - Proteins and Proteomics, 1844(1 PART A), 52–62. https://doi.org/10.1016/j.bbapap.2013.02.029
35. Krumsiek, J., Bartel, J., & Theis, F. J. (2016). Computational approaches for systems metabolomics. In Current Opinion in Biotechnology (Vol. 39, pp. 198–206). Elsevier Ltd. https://doi.org/10.1016/j.copbio.2016.04.009
36. Lapatas, V., Stefanidakis, M., Jimenez, R. C., Via, A., & Schneider, M. V. (2015). Data integration in biological research: an overview. Journal of Biological Research-Thessaloniki, 22(1), 9. https://doi.org/10.1186/s40709-015-0032-5
37. Lê Cao, K. A., Boitard, S., & Besse, P. (2011). Sparse PLS discriminant analysis: Biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics, 12(1), 253. https://doi.org/10.1186/1471-2105-12-253
38. Lê Cao, K. A., González, I., & Déjean, S. (2009). IntegrOmics: An R package to unravel relationships between two omics datasets. Bioinformatics, 25(21), 2855–2856. https://doi.org/10.1093/bioinformatics/btp515
39. Lê Cao, K. A., Martin, P. G. P., Robert-Granié, C., & Besse, P. (2009). Sparse canonical methods for biological data integration: Application to a cross-platform study. BMC Bioinformatics, 10(1), 34. https://doi.org/10.1186/1471-2105-10-34
40. Lock, E. F., & Dunson, D. B. (2013). Bayesian consensus clustering. Bioinformatics, 29(20), 2610–2616. https://doi.org/10.1093/bioinformatics/btt425
41. Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., Powers, R. S., Ladanyi, M., & Shen, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. Proceedings of the National Academy of Sciences of the United States of America, 110(11), 4245. https://doi.org/10.1073/PNAS.1208949110
42. Müller S, Kohanbash G, Liu SJ, Alvarado B, Carrera D, Bhaduri A, Watchmaker PB, Yagnik G, Di Lullo E, Malatesta M, Amankulor NM, Kriegstein AR, Lim DA, Aghi M, Okada H, Diaz A. Single-cell profiling of human gliomas reveals macrophage ontogeny as a basis for regional differences in macrophage activation in the tumor microenvironment. Genome Biol. 2017 Dec 20;18(1):234. doi: 10.1186/s13059-017-1362-4. PMID: 29262845; PMCID: PMC5738907
43. Nardelli C, Granata I, Iaffaldano L, D'Argenio V, Del Monaco V, Maruotti GM, Omodei D, Del Vecchio L, Martinelli P, Salvatore F, Guarracino MR, Sacchetti L, Pastore L. miR-138/miR-222 Overexpression Characterizes the miRNome of Amniotic Mesenchymal

Stem Cells in Obesity. Stem Cells Dev. 2017 Jan 1;26(1):4-14. doi: 10.1089/scd.2016.0127. Epub 2016 Nov 8. PMID: 27762728

44. O'Donnell, S. T., Ross, R. P., & Stanton, C. (2020). The Progress of Multi-Omics Technologies: Determining Function in Lactic Acid Bacteria Using a Systems Level Approach. In Frontiers in Microbiology (Vol. 10, p. 3084). Frontiers Media S.A. https://doi.org/10.3389/fmicb.2019.03084
45. Kulkarni, P, Frommolt, P (2017). Challenges in the Setup of Large-scale Next-Generation Sequencing Analysis Workflows. Computational and Structural Biotechnology Journal, 15, 471–477. https://doi.org/10.1016/J.CSBJ.2017.10.001
46. Pavesi G. ChIP-Seq Data Analysis to Define Transcriptional Regulatory Networks. Adv Biochem Eng Biotechnol. 2017;160:1-14. doi: 10.1007/10_2016_43. PMID: 28070596
47. Perez-Muñoz ME, Arrieta MC, Ramer-Tait AE, Walter J. A critical assessment of the "sterile womb" and "in utero colonization" hypotheses: implications for research on the pioneer infant microbiome. Microbiome. 2017 Apr 28;5(1):48. doi: 10.1186/s40168-017-0268-4. PMID: 28454555; PMCID: PMC5410102
48. Pinu, F. R., Beale, D. J., Paten, A. M., Kouremenos, K., Swarup, S., Schirra, H. J., & Wishart, D. (2019). Systems biology and multi-omics integration: Viewpoints from the metabolomics research community. Metabolites, 9(4). https://doi.org/10.3390/metabo9040076
49. Precone et al. (2015). Cracking the Code of Human Diseases Using Next-Generation Sequencing: Applications, Challenges, and Perspectives. BioMed Research International, 2015. https://doi.org/10.1155/2015/161648
50. Pu W, Wang C, Chen S, Zhao D, Zhou Y, Ma Y, Wang Y, Li C, Huang Z, Jin L, Guo S, Wang J, Wang M. Targeted bisulfite sequencing identified a panel of DNA methylation-based biomarkers for esophageal squamous cell carcinoma (ESCC). Clin Epigenetics. 2017 Dec 15;9:129. doi: 10.1186/s13148-017-0430-7. PMID: 29270239; PMCID: PMC5732523
51. Rajasundaram, D., & Selbig, J. (2016). More effort - more results: Recent advances in integrative "omics" data analysis. In Current Opinion in Plant Biology (Vol. 30, pp. 57–61). Elsevier Ltd. https://doi.org/10.1016/j.pbi.2015.12.010
52. Ramos, M., Schiffer, L., Re, A., Azhar, R., Basunia, A., Rodriguez, C., Chan, T., Chapman, P., Davis, S. R., Gomez-Cabrero, D., Culhane, A. C., Haibe-Kains, B., Hansen, K. D., Kodali, H., Louis, M. S., Mer, A. S., Riester, M., Morgan, M., Carey, V., & Waldron, L. (2017). Software for the integration of multiomics experiments in bioconductor. Cancer Research, 77(21), e39–e42. https://doi.org/10.1158/0008-5472.CAN-17-0344
53. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 43(7), e47–e47. https://doi.org/10.1093/NAR/GKV007
54. Rohart, F., Gautier, B., Singh, A., & Lê Cao, K.-A. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. PLOS Computational Biology, 13(11), e1005752. https://doi.org/10.1371/journal.pcbi.1005752

55. Sandhu, C., Qureshi, A., & Emili, A. (2018). Panomics for Precision Medicine. Trends in Molecular Medicine, 24(1), 85. https://doi.org/10.1016/J.MOLMED.2017.11.001
56. Shen, R., Olshen, A. B., & Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics, 25(22), 2906–2912. https://doi.org/10.1093/bioinformatics/btp543
57. Singh, A., Shannon, C. P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S. J., & Cao, K. A. L. (2019). DIABLO: An integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics, 35(17), 3055–3062. https://doi.org/10.1093/bioinformatics/bty1054
58. SM, C., N, D., AM, K., J, M., C, H., & G, M. (2018). Clinical sequencing: From raw data to diagnosis with lifetime value. Clinical Genetics, 93(3), 508–519. https://doi.org/10.1111/CGE.13190
59. Su YT, Chen R, Wang H, Song H, Zhang Q, Chen LY, Lappin H, Vasconcelos G, Lita A, Maric D, Li A, Celiku O, Zhang W, Meetze K, Estok T, Larion M, Abu-Asab M, Zhuang Z, Yang C, Gilbert MR, Wu J. Novel Targeting of Transcription and Metabolism in Glioblastoma. Clin Cancer Res. 2018 Mar 1;24(5):1124-1137. doi: 10.1158/1078-0432.CCR-17-2032. Epub 2017 Dec 18. PMID: 29254993; PMCID: PMC8108069
60. Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics Data Integration, Interpretation, and Its Application. In Bioinformatics and Biology Insights (Vol. 14). SAGE Publications Inc. https://doi.org/10.1177/1177932219899051
61. Sun, Y. V., & Hu, Y. J. (2016). Integrative Analysis of Multi-omics Data for Discovery and Functional Studies of Complex Human Diseases. Advances in Genetics, 93, 147–190. https://doi.org/10.1016/bs.adgen.2015.11.004
62. Tyson, J. R., O'Neil, N. J., Jain, M., Olsen, H. E., Hieter, P., & Snutch, T. P. (2018). MinION-based long-read sequencing and assembly extends the Caenorhabditis elegans reference genome. Genome research, 28(2), 266-274
63. Ulfenborg, B. (2019). Vertical and horizontal integration of multi-omics data with miodin. BMC Bioinformatics 2019 20:1, 20(1), 1–10. https://doi.org/10.1186/S12859-019-3224-4
64. Ulfenborg, B. (2019). Vertical and horizontal integration of multi-omics data with miodin. BMC Bioinformatics, 20(1), 1–10. https://doi.org/10.1186/s12859-019-3224-4
65. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina

N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. The sequence of the human genome. Science. 2001 Feb 16;291(5507):1304-51. doi: 10.1126/science.1058040. Erratum in: Science 2001 Jun 5;292(5523):1838. PMID: 11181995

66. Wang, Z., & Wei, P. (2020). IMIX: A multivariate mixture model approach to integrative analysis of multiple types of omics data. BioRxiv, 2020.06.23.167312. https://doi.org/10.1101/2020.06.23.167312
67. Weisz Hubshman M, Broekman S, van Wijk E, Cremers F, Abu-Diab A, Khateb S, Tzur S, Lagovsky I, Smirin-Yosef P, Sharon D, Haer-Wigman L, Banin E, Basel-Vanagaite L, de Vrieze E. Whole-exome sequencing reveals POC5 as a novel gene associated with autosomal recessive retinitis pigmentosa. Hum Mol Genet. 2018 Feb 15;27(4):614-624. doi: 10.1093/hmg/ddx428. PMID: 29272404
68. Wu, C., Zhou, F., Ren, J., Li, X., Jiang, Y., & Ma, S. (2019). A Selective Review of Multi-Level Omics Data Integration Using Variable Selection. High-Throughput, 8(1), 4. https://doi.org/10.3390/ht8010004
69. Yadav, S. P. (2007). The Wholeness in Suffix -omics, -omes, and the Word Om. Journal of Biomolecular Techniques: JBT, 18(5), 277. /pmc/articles/PMC2392988/
70. Yan, K. K., Zhao, H., & Pang, H. (2017). A comparison of graph- and kernel-based -omics data integration algorithms for classifying complex traits. BMC Bioinformatics, 18(1), 539. https://doi.org/10.1186/s12859-017-1982-4

71. Zanfardino, M., Castaldo, R., Pane, K., Affinito, O., Aiello, M., Salvatore, M., & Franzese, M. (2021). MuSA: a graphical user interface for multi-OMICs data integration in radiogenomic studies. Scientific Reports, 11(1), 1550. https://doi.org/10.1038/S41598-021-81200-Z
72. Zong CC. Single-cell RNA-seq study determines the ontogeny of macrophages in glioblastomas. Genome Biol. 2017 Dec 21;18(1):235. doi: 10.1186/s13059-017-1375-z. PMID: 29262851; PMCID: PMC5738720

6. Παράρτημα

Παρατίθενται οι βασικές εντολές του κώδικα σε R, που χρησιμοποιήθηκε για τους σκοπούς της εργασίας. Παραλείπονται τα κομμάτια που της προετοιμασίας των δεδομένων και των δοκιμών ώστε να βρεθούν οι κατάλληλες παράμετροι σε κάθε περίπτωση.

6.1. Κώδικας που χρησιμοποιήθηκε για το πακέτο mixOmics

```
if (!requireNamespace("BiocManager", quietly = TRUE))
install.packages("BiocManager")
BiocManager::install("mixOmics")
library(mixOmics)
library(rgl)
library(knitr)
```

```
library(mixOmics)
data(breast.TCGA)
X <- list(mRNA = breast.TCGA$data.train$mrna,
          miRNA = breast.TCGA$data.train$mirna,
          protein = breast.TCGA$data.train$protein)
Y <- breast.TCGA$data.train$subtype
summary(Y)
list.keepX <- list(mRNA = c(16, 17), miRNA = c(18,5), protein = c(5, 5))

MyResult.diablo <- block.splsda(X, Y, keepX=list.keepX)

plotIndiv(MyResult.diablo,
          ind.names = FALSE,
          legend=TRUE, cex=c(1,2,3),
          title = '')

plotVar(MyResult.diablo, var.names = c(TRUE, TRUE, TRUE),
        legend=TRUE, pch=c(16,16,1))

plotDiablo(MyResult.diablo, ncomp = 1)

circosPlot(MyResult.diablo, cutoff=0.7, size.variables = 0.8)

cimDiablo(MyResult.diablo, color.blocks = c('darkorchid', 'brown1', 'lightgreen'), comp = 1,
    margin=c(8,20), legend.position = "right")

plotLoadings(MyResult.diablo, comp = 2, contrib = "max", size.name = 1)

MyPerf.diablo <- perf(MyResult.diablo, validation = 'Mfold', folds = 5,
                      nrepeat = 10,
                      dist = 'centroids.dist')

# Area under the Curve (AUC)
Myauc.diablo <- auroc(MyResult.diablo, roc.block = "protein", roc.comp = 1, title = "")
as_mRNA_comp1 = Myauc.diablo$graph.mRNA$comp1
as_mRNA_comp1 = as_mRNA_comp1 + theme(legend.text = element_text(size=12),
                                      axis.title = element_text(size = 12)
                                      )
as_mRNA_comp2 = Myauc.diablo$graph.mRNA$comp2
as_mRNA_comp2 = as_mRNA_comp2 + theme(legend.text = element_text(size=12),
                                      axis.title = element_text(size = 12)
                                      )
```

```
48.
49. as_miRNA_comp1 = Myauc.diablo$graph.miRNA$comp1
50. as_miRNA_comp1 = as_miRNA_comp1 + theme(legend.text = element_text(size=12),
51.                                         axis.title = element_text(size = 12)
52. )
53. as_miRNA_comp2 = Myauc.diablo$graph.miRNA$comp2
54. as_miRNA_comp2 = as_miRNA_comp2 + theme(legend.text = element_text(size=12),
55.                                         axis.title = element_text(size = 12)
56. )
57.
58. as_protein_comp1 = Myauc.diablo$graph.protein$comp1
59. as_protein_comp1 = as_protein_comp1 + theme(legend.text = element_text(size=12),
60.                                             axis.title = element_text(size = 12)
61. )
62. as_protein_comp2 = Myauc.diablo$graph.protein$comp2
63. as_protein_comp2 = as_protein_comp2 + theme(legend.text = element_text(size=12),
64.                                             axis.title = element_text(size = 12)
65. )
66.
67. # prediction
68. X.test <- list(mRNA = breast.TCGA$data.test$mrna,
69.                miRNA = breast.TCGA$data.test$mirna)
70.
71. Mypredict.diablo <- predict(MyResult.diablo, newdata = X.test)
72. confusion.mat <- get.confusion_matrix(
73.   truth = breast.TCGA$data.test$subtype,
74.   predicted = Mypredict.diablo$MajorityVote$centroids.dist[,2])
```

6.2. Κώδικας που χρησιμοποιήθηκε για το πακέτο MOFA

```
1.  devtools::install_github("bioFAM/MOFAdata", build_opts = c("--no-resave-data"))
2.  library(MOFA2)
3.  library(MOFAdata)
4.  library(data.table)
5.  library(ggplot2)
6.  library(tidyverse
7.  utils::data("CLL_data")
8.  lapply(CLL_data,dim)
9.
10. MOFAobject <- create_mofa(CLL_data)
11. MOFAobject
12.
13. data_opts <- get_default_data_options(MOFAobject)
14. data_opts
15.
16. model_opts <- get_default_model_options(MOFAobject)
17. model_opts$num_factors <- 15
18. model_opts
19.
20. train_opts <- get_default_training_options(MOFAobject)
21. train_opts$convergence_mode <- "slow"
22.
23. MOFAobject <- prepare_mofa(MOFAobject,
24.                            data_options = data_opts,
25.                            model_options = model_opts,
26.                            training_options = train_opts
27. )
28.
29. slotNames(MOFAobject)
30. names(MOFAobject@data)
31. dim(MOFAobject@data$Drugs$group1)
32. names(MOFAobject@expectations)
```

```
dim(MOFAobject@expectations$Z$group1)
dim(MOFAobject@expectations$W$mRNA)

samples_metadata(MOFAobject) <- metadata

plot_factor_cor(MOFAobject)
plot_variance_explained(MOFAobject, max_r2=15)
plot_variance_explained(MOFAobject, plot_total = T)[[2]]


correlate_factors_with_covariates(MOFAobject,
                                  covariates = c("Gender","died","age"),
                                  plot="log_pval"
                                  )

plot_factor(MOFAobject,
            factors = 1,
            color_by = "Factor1"
            )

plot_weights(MOFAobject,
             view = "Mutations",
             factor = 1,
             nfeatures = 10,     # Top number of features to highlight
             scale = T           # Scale weights from -1 to 1
             )

plot_top_weights(MOFAobject,
                 view = "Mutations",
                 factor = 1,
                 nfeatures = 10,     # Top number of features to highlight
                 scale = T           # Scale weights from -1 to 1
                 )

plot_factor(MOFAobject,
            factors = 1,
            color_by = "IGHV",
            add_violin = TRUE,
            dodge = TRUE
            )

plot_factor(MOFAobject,
            factors = 1,
            color_by = "Gender",
            dodge = TRUE,
            add_violin = TRUE
            )

plot_weights(MOFAobject,
             view = "mRNA",
             factor = 1,
             nfeatures = 20
             )

plot_data_scatter(MOFAobject,
                  view = "mRNA",
                  factor = 1,
                  features = 4,
                  sign = "positive",
                  color_by = "IGHV"
                  ) + labs(y="RNA expression")

plot_data_scatter(MOFAobject,
                  view = "mRNA",
                  factor = 1,
```

```
                  features = 4,
                  sign = "negative",
                    color_by = "IGHV"
                    ) + labs(y="RNA expression")

plot_data_heatmap(MOFAobject,
                    view = "mRNA",
                    factor = 1,
                    features = 5,
                    cluster_rows = FALSE, cluster_cols = FALSE,
                    show_rownames = TRUE, show_colnames = TRUE,
                    scale = "row"
                    )

plot_data_heatmap(MOFAobject,
                    view = "mRNA",
                    factor = 1,
                    features = 25,
                    denoise = TRUE,
                    cluster_rows = FALSE, cluster_cols = FALSE,
                    show_rownames = TRUE, show_colnames = FALSE,
                    scale = "row"
                    )


plot_weights(MOFAobject,
              view = "Mutations",
              factor = 3,
              nfeatures = 10,
              abs = F
              )
plot_factor(MOFAobject,
             factors = 3,
             color_by = "trisomy12",
             dodge = TRUE,
             add_violin = TRUE
             )

plot_data_scatter(MOFAobject,
                    view = "Drugs",
                    factor = 3,
                    features = 4,
                    sign = "positive",
                    color_by = "trisomy12"
                    ) + labs(y="Drug response (cell viability)")

plot_data_heatmap(MOFAobject,
                    view = "mRNA",
                    factor = 3,
                    features = 25,
                    denoise = TRUE,
                    cluster_rows = TRUE, cluster_cols = FALSE,
                    show_rownames = TRUE, show_colnames = FALSE,
                    scale = "row"
                    )

p <- plot_factors(MOFAobject,
                    factors = c(1,3),
                    color_by = "IGHV",
                    shape_by = "trisomy12",
                    dot_size = 2.5,
                    show_missing = T
                    )

p <- p +
```

```
  geom_hline(yintercept=-1, linetype="dashed") +
  geom_vline(xintercept=(-0.5), linetype="dashed")

suppressPackageStartupMessages(library(randomForest))
df <- as.data.frame(get_factors(MOFAobject, factors=c(1,2))[[1]])

df$IGHV <- as.factor(MOFAobject@samples_metadata$IGHV)
model.ighv <- randomForest(IGHV ~ ., data=df[!is.na(df$IGHV),], ntree=10)
df$IGHV <- NULL

MOFAobject@samples_metadata$IGHV.pred <- stats::predict(model.ighv, df)
df$trisomy12 <- as.factor(MOFAobject@samples_metadata$trisomy12)
model.trisomy12 <- randomForest(trisomy12 ~ ., data=df[!is.na(df$trisomy12),], ntree=10)
df$trisomy12 <- NULL
MOFAobject@samples_metadata$trisomy12.pred <- stats::predict(model.trisomy12, df)

MOFAobject@samples_metadata$IGHV.pred_logical <-
  c("True","Predicted")[as.numeric(is.na(MOFAobject@samples_metadata$IGHV))+1]
p <- plot_factors(MOFAobject,
                  factors = c(1,3),
                  color_by = "IGHV.pred",
                  shape_by = "IGHV.pred_logical",
                  dot_size = 2.5,
                  show_missing = T
                  )

p <- p +
  geom_hline(yintercept=-1, linetype="dashed") +
  geom_vline(xintercept=(-0.5), linetype="dashed")

print(p)
utils::data(reactomeGS)

head(colnames(reactomeGS))
head(rownames(reactomeGS))

res.positive <- run_enrichment(MOFAobject,
                               feature.sets = reactomeGS,
                               view = "mRNA",
                               sign = "positive"
                               )

res.negative <- run_enrichment(MOFAobject,
                               feature.sets = reactomeGS,
                               view = "mRNA",
                               sign = "negative"
                               )
plot_enrichment_heatmap(res.positive)
plot_enrichment_heatmap(res.negative)
plot_enrichment(res.positive, factor = 5, max.pathways = 15)

plot_enrichment_detailed(
  enrichment.results = res.positive,
  factor = 5,
  # feature.sets = reactomeGS,
  max.pathways = 15
  )
```

6.3. Κώδικας που χρησιμοποιήθηκε για το πακέτο iClusterPlus

```
BiocManager::install("iClusterPlus")
BiocManager::install("GenomicRanges")
```

```
3.
4.  library(iClusterPlus)
5.  library(GenomicRanges)
6.  library(gplots)
7.  library(lattice)
8.
9.  data(gbm)
10. dim(gbm.mut)
11. dim(gbm.exp)
12. dim(gbm.seg)
13.
14. mut.rate=apply(gbm.mut,2,mean)
15. gbm.mut2 = gbm.mut[,which(mut.rate>0.02)]
16. gbm.mut2[1:10,1:8]
17. write.csv(x = as.data.frame(gbm.mut2[1:10,1:8]), file = "./mutations.csv")
18.
19. dim(gbm.exp)
20. gbm.exp[1:3,1:8]
21. write.csv(x = as.data.frame(gbm.exp[1:10,1:8]), file = "./expression.csv")
22.
23. dim(gbm.seg)
24. gbm.seg[1:3,]
25. write.csv(x = as.data.frame(gbm.seg[1:10,1:8]), file = "./cnv.csv")
26. data(variation.hg18.v10.nov.2010)
27. gbm.cn=CNregions(seg=gbm.seg,epsilon=0,adaptive=FALSE,rmCNV=TRUE,
28.                  cnv=variation.hg18.v10.nov.2010[,3:5],
29.                  frac.overlap=0.5, rmSmallseg=TRUE,nProbes=5)
30.
31. dim(gbm.cn)
32. gbm.cn[1:3,1:5]
33.
34. gbm.cn=gbm.cn[order(rownames(gbm.cn)),]
35. # check if all the samples are in the same order for the three data sets
36. all(rownames(gbm.cn)==rownames(gbm.exp))
37. all(rownames(gbm.cn)==rownames(gbm.mut2))
38.
39. fit.single=iClusterPlus(dt1=gbm.mut2,dt2=gbm.cn,dt3=gbm.exp,
40.                         type=c("binomial","gaussian","gaussian"),
41.                         lambda=c(0.04,0.61,0.90),K=2,maxiter=10)
42.
43. set.seed(123)
44. date()
45. for(k in 1:5){
46.   cv.fit = tune.iClusterPlus(cpus=12,dt1=gbm.mut2,dt2=gbm.cn,dt3=gbm.exp,
47.                              type=c("binomial","gaussian","gaussian"),K=k,n.lambda=185,
48.                              scale.lambda=c(1,1,1),maxiter=20)
49.   save(cv.fit, file=paste("cv.fit.k",k,".Rdata",sep=""))
50.   }
51. date()
52.
53. output=alist()
54. files=grep("cv.fit",dir())
55. for(i in 1:length(files)){
56.   load(dir()[files[i]])
57.   output[[i]]=cv.fit
58.   }
59. nLambda = nrow(output[[1]]$lambda)
60. nK = length(output)
61. BIC = getBIC(output)
62. devR = getDevR(output)
63.
64. minBICid = apply(BIC,2,which.min)
65. devRatMinBIC = rep(NA,nK)
66. for(i in 1:nK){
67.   devRatMinBIC[i] = devR[minBICid[i],i]
```

```
68.   }
69.
70. clusters=getClusters(output)
71. rownames(clusters)=rownames(gbm.exp)
72. colnames(clusters)=paste("K=",2:(length(output)+1),sep="")
73. k=2
74. best.cluster=clusters[,k]
75. best.fit=output[[k]]$fit[[which.min(BIC[,k])]]
76.
77. plot(1:(nK+1),c(0,devRatMinBIC),type="b",xlab="Number of clusters (K+1)",
78.      ylab="%Explained Variation")
79.
80. chr=unlist(strsplit(colnames(gbm.cn),"\\."))
81. chr=chr[seq(1,length(chr),by=2)]
82. chr=gsub("chr","",chr)
83. chr=as.numeric(chr)
84. cn.image=gbm.cn
85. cn.image[cn.image>1.5]=1.5
86. cn.image[cn.image< -1.5]= -1.5
87. exp.image=gbm.exp
88. exp.image[exp.image>2.5]=2.5
89. exp.image[exp.image< -2.5]= -2.5
90.
91. features = alist()
92. features[[1]] = colnames(gbm.mut2)
93. features[[2]] = colnames(gbm.cn)
94. features[[3]] = colnames(gbm.exp)
95. sigfeatures=alist()
96. for(i in 1:3){
97.   rowsum=apply(abs(best.fit$beta[[i]]),1, sum)
98.   upper=quantile(rowsum,prob=0.75)
99.   sigfeatures[[i]]=(features[[i]])[which(rowsum>upper)]
100.    }
101.  names(sigfeatures)=c("mutation","copy number","expression")
102.  #print a few examples of selected features
103.  head(sigfeatures[[1]])
104.  head(sigfeatures[[2]])
105.  head(sigfeatures[[3]])
106.
107.  bw.col = colorpanel(2,low="white",high="black")
108.  col.scheme = alist()
109.  col.scheme[[1]] = bw.col
110.  col.scheme[[2]] = bluered(256)
111.  col.scheme[[3]] = bluered(256)
112.  plotHeatmap(fit=best.fit,datasets=list(gbm.mut2,cn.image,exp.image),
113.              type=c("binomial","gaussian","gaussian"), col.scheme = col.scheme,
114.              row.order=c(F,F,T),chr=chr,plot.chr=c(T,T,T),sparse=c(T,F,T),cap=c(F,T,F))
115.
```

6.4. Κώδικας που χρησιμοποιήθηκε για το πακέτο MOVICS

```
1.  devtools::install_github("xlucpu/MOVICS")
2.  library("MOVICS")
3.
4.  load(system.file("extdata", "brca.tcga.RData", package = "MOVICS", mustWork = TRUE))
5.  load(system.file("extdata", "brca.yau.RData",  package = "MOVICS", mustWork = TRUE))
6.
7.  names(brca.tcga)
8.  names(brca.yau)
9.  Mydata   <- brca.tcga[1:4]
10.
11. write.csv(x = as.data.frame(Mydata$mRNA.expr[1:10,1:8]), file = "./mRNA_expression.csv")
```

```
write.csv(x = as.data.frame(Mydata$lncRNA.expr[1:10,1:8]), file = "./lncRNA_expression.csv")
write.csv(x = as.data.frame(Mydata$meth.beta[1:10,1:8]), file = "./methylation_beta.csv")
write.csv(x = as.data.frame(Mydata$mut.status[1:10,1:8]), file = "./mutations.csv")

count     <- brca.tcga$count
fpkm      <- brca.tcga$fpkm
maf       <- brca.tcga$maf
segment   <- brca.tcga$segment

optk.brca <- getClustNum(data        = Mydata,
                         is.binary   = c(F,F,F,T),
                         try.N.clust = 2:8,
                         fig.name    = "CLUSTER NUMBER OF TCGA-BRCA")

iClusterBayes.res <- getiClusterBayes(data        = Mydata,
                                      N.clust     = 5,
                                      type        =
    c("gaussian","gaussian","gaussian","binomial"),
                                      n.burnin    = 1800,
                                      n.draw      = 1200,
                                      prior.gamma = c(0.5, 0.5, 0.5, 0.5),
                                      sdev        = 0.05,
                                      thin        = 3)

res.list <- getMOIC(data        = Mydata,
                         methodslist = list("SNF", "PINSPlus", "NEMO", "COCA", "LRAcluster",
    "ConsensusClustering", "IntNMF", "CIMLR", "MoCluster"),
                         N.clust     = 5,
                         type        = c("gaussian", "gaussian", "gaussian", "binomial"))

res.list <- append(res.list,
                        list("iClusterBayes" = iClusterBayes.res))

save(res.list, file = "./res.list.rda")


consensus.brca <- getConsensusMOIC(res.list = res.list,
                                   fig.name      = "CONSENSUS HEATMAP",
                                   distance      = "euclidean",
                                   linkage       = "average")

getSilhouette(sil      = consensus.brca$sil, # a sil object returned by getConsensusMOIC()
              fig.path = getwd(),
              fig.name = "SILHOUETTE",
              height   = 5.5,
              width    = 5)

indata <- Mydata
indata$meth.beta <- log2(indata$meth.beta / (1 - indata$meth.beta))

plotdata <- getStdiz(data       = indata,
                     halfwidth  = c(2,2,2,NA), # no truncation for mutation
                     centerFlag = c(T,T,T,F), # no center for mutation
                     scaleFlag  = c(T,T,T,F)) # no scale for mutation

feat   <- iClusterBayes.res$feat.res
feat1  <- feat[which(feat$dataset == "mRNA.expr"),][1:10,"feature"]
feat2  <- feat[which(feat$dataset == "lncRNA.expr"),][1:10,"feature"]
feat3  <- feat[which(feat$dataset == "meth.beta"),][1:10,"feature"]
feat4  <- feat[which(feat$dataset == "mut.status"),][1:10,"feature"]
annRow <- list(feat1, feat2, feat3, feat4)

mRNA.col   <- c("#00FF00", "#008000", "#000000", "#800000", "#FF0000")
lncRNA.col <- c("#6699CC", "white"  , "#FF3C38")
meth.col   <- c("#0074FE", "#96EBF9", "#FEE900", "#F00003")
```

```
mut.col    <- c("grey90" , "black")
col.list   <- list(mRNA.col, lncRNA.col, meth.col, mut.col)

x = getMoHeatmap(data      = plotdata,
             row.title     = c("mRNA","lncRNA","Methylation","Mutation"),
             is.binary     = c(F,F,F,T), # the 4th data is mutation which is binary
             legend.name   = c("mRNA.FPKM","lncRNA.FPKM","M value","Mutated"),
             clust.res     = iClusterBayes.res$clust.res, # cluster results
             clust.dend    = NULL, # no dendrogram
             show.rownames = c(F,F,F,F), # specify for each omics data
             show.colnames = FALSE, # show no sample names
             annRow        = annRow, # mark selected features
             color         = col.list,
             annCol        = NULL, # no annotation for samples
             annColors     = NULL, # no annotation color
             width         = 10, # width of each subheatmap
             height        = 5, # height of each subheatmap
             fig.name      = "COMPREHENSIVE HEATMAP OF ICLUSTERBAYES")

MSIGDB.FILE <- system.file("extdata", "c5.bp.v7.1.symbols.xls", package = "MOVICS", mustWork
    = TRUE)

gsea.up <- runGSEA(moic.res     = consensus.brca,
                   dea.method   = "limma", # name of DEA method
                   prefix       = "TCGA-BRCA", # MUST be the same of argument in runDEA()
                   dat.path     = "C:/Users/eleft/OneDrive/Documents/Github/thesis/dea/", #
    path of Differential expression analysis files
                     res.path     = getwd(),
                     msigdb.path  = MSIGDB.FILE,
                     norm.expr    = fpkm,
                     dirct        = "up",
                     p.cutoff     = 0.05,
                     p.adj.cutoff = 0.05,
                     gsva.method  = "gsva",
                     norm.method  = "mean",
                     fig.name     = "UPREGULATED PATHWAY HEATMAP")

```