ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΓΟΝΙΔΙΑΚΗΣ ΕΚΦΡΑΣΗΣ ΜΕ ΧΡΗΣΗ ΕΜΠΕΙΡΙΚΩΝ ΜΠΕΫΖΙΑΝΏΝ ΜΕΘΟΔΩΝ ΚΑΙ ΥΠΟΛΟΓΙΣΤΙΚΗ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΚΑΡΚΙΝΙΚΩΝ ΜΕΛΑΝΩΜΑΤΙΚΩΝ ΔΕΙΓΜΑΤΩΝ ΓΙΑ ΕΠΑΝΑΤΟΠΟΘΕΤΗΣΗ ΦΑΡΜΑΚΩΝ

Διπλωματική Εργασία

Κωνσταντίνος Νταγιάντας

Επιβλέπων Καθηγητής: Αλεξόπουλος Λεωνίδας

Εργαστήριο Εμβιομηχανικής και Συστημικής Βιολογίας

Εθνικό Μετσόβιο Πολυτεχνείο

ΑΘΗΝΑ

Σεπτέμβριος 2018

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον κ. Λεωνίδα Αλεξόπουλο για την καθοδήγηση και συμπαράστασή του καθόλη τη διάρκεια της εκπόνησης της διπλωματικής μου εργασίας, καθώς επίσης και τα άτομα του εργαστηρίου για τη βοήθειά τους. Ιδιαιτέρως ευχαριστώ τον κ. Asier Antoranz, ο οποίος επέβλεψε την διπλωματική μου εργασία συμβουλεύοντάς με, και δίνοντάς μου ιδέες καθ'όλη τη διάρκειά της.

Περίληψη

Η μοντελοποίηση βιολογικών συστημάτων είναι ένα από τα πολλά παρακλάδια του ταχύτατα αναπτυσσόμενου κλάδου της υπολογιστικής βιολογίας. Με βάση τεχνολογίες που μετράνε την έκφραση των γονιδίων των κυττάρων, καθώς και προηγούμενη βιολογική γνώση, κατασκευάζονται μοντέλα τα οποία περιγράφουν τον μηχανισμό με τον οποίο διενεργούνται οι διάφορες διεργασίες των κυττάρων. Με τη νέα γνώση που προσφέρουν τα βιολογικά μοντέλα, επιταχύνεται η ανακάλυψη ουσιών και φαρμάκων για καινούργιες θεραπείες, καθώς και γίνεται δυνατή η εύρεση βιοδεικτών για πρόληψη και διάγνωση ασθενειών ή δυσλειτουργιών των κυττάρων. Στη συγκεκριμένη εργασία, γίνεται εφαρμογή αυτών των πρακτικών στην περίπτωση του καρκίνου και συγκεκριμένα στο μελάνωμα. Το μελάνωμα αποτελεί την πιο θανατηφόρα μορφή καρκίνου του δέρματος, και χαρακτηρίζεται από ιδιαίτερη ποικιλομορφία. Ο καρκίνος, λόγω του τρόπου δημιουργίας του, δεν μπορεί να μελετηθεί όπως άλλες ασθένειες, αλλά πρέπει να υπάρχει εξειδικευμένη αντιμετώπιση για κάθε είδος καρκίνου, ακόμα και για κάθε διαφορετική καρκινική κυτταροσειρά ή ιστό που μελετάται από τον ίδιο τύπο καρκίνου. Για να αυξηθεί η αξιοπιστία των αποτελεσμάτων, χρησιμοποιήθηκαν εμπειρικά Μπεϋζιανά μοντέλα που επιτρέπουν τον συνδυασμό δεδομένων από διαφορετικά πειράματα, στη συγκεκριμένη εργασία τέσσερα στον αριθμό, ώστε να καλυφθεί μεγαλύτερο εύρος των διαφορετικών μορφών καρκινικών κυττάρων μελανώματος. Στη συνέχεια, γίνεται ανάλυση ώστε να εντοπιστούν τα στατιστικώς σημαντικά γονίδια και βιολογικά μονοπάτια που διαφοροποιούν καρκινικά από υγιή κύτταρα, αλλά και τις διάφορες μεταλλάξεις του καρκίνου μεταξύ τους. Τέλος, με βάση αυτά τα αποτελέσματα, κατασκευάζεται μοντέλο που χρησιμοποιεί τα λιγότερα γονίδια ώστε να κάνει τη διάκριση ανάμεσα στα διαφορετικά κύτταρα, καθώς επίσης γίνεται χρήση του εργαλείου cMap του Broad Institute ,ώστε να γίνει εντοπισμός πιθανών ουσιών και φαρμάκων για την υποχώρηση και τη θεραπεία του μελανώματος.

Λέξεις κλειδιά: υπολογιστική βιολογία, καρκίνος, μελάνωμα, βιολογικά μονοπάτια, σύνθετα συστήματα, διαφορικώς εκφρασμένα γονίδια,βιοπληροφορική, support vector machines, επανατοποθέτηση φαρμάκων,machine learning, cMap.

Abstract

Systems biology is one of the many branches of the rapidly evolving science of computational biology. Biological models that suggest new ways and mechanisms through which cells are able to conduct complex procedures are based on new technologies that measure gene expression of cells and existing biological knowledge. These models help accelerate the discovery of drugs and substances to fight a disease, and make biomarker discovery easier and more precise. In this particular dissertation, the aforementioned practices are applied in the case of cancerous melanoma. Melanoma is the deadliest form of skin cancer, with a high patient mortality and high recurrence rate, and has a wide variance among its different types and mutations. In general, cancer constitutes a complex disease, with each form of cancer being highly distinct from one another. Hence a specialized approach is necessary for every subtype and case, even cell line. To increase the reliability of the results, empirical Bayes models were used that allow the integration of data from different experiments, in this project four in number, in an effort to cover a wide range of melanoma cell lines and mutations. Then, differential analysis is conducted that leads to differentially expressed genes and biological pathways between melanoma cell lines and healthy melanocytes, as well as among the different melanoma cell lines. These findings are then used as an input to the cMap tool of Broad Institute for drug discovery and repurposing, and also as features for machine learning models that predict the type of cell (healthy vs. cancerous) and the type of melanoma mutations among cell lines.

Key words: computational biology, cancer, melanoma, biological pathways, complex systems, differentially expressed genes, bioinformatics, drug, support vector machines, repurposing, machine learning, cMap.

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

1.1 Υπολογιστική Βιολογία

Η υπολογιστική βιολογία είναι ένας ευρύς επιστημονικός κλάδος που συνδυάζει γνώσεις από πολλές επιστήμες. Μέσα από την ανάπτυξη και την εφαρμογή θεωρητικών και εμπειρικών μεθόδων, μαθηματικών μοντέλων και υπολογιστικές προσομοιώσεις, μελετούνται βιολογικά συστήματα σε διάφορους βαθμούς, από πολύ μεγάλες κλίμακες (πχ. οικοσυστήματα, πληθυσμοί οργανισμών), μέχρι πολύ μικρές (πχ. κύτταρα, βιολογικά μονοπάτια εντός των κυττάρων). Ο κλάδος βασίζεται στη βιολογία, τη μοριακή βιολογία και τη γενετική, αλλά και στα εφαρμοσμένα μαθηματικά, τη στατιστική και την επιστήμη των υπολογιστών.Τα τελευταία χρόνια, χάρη στη ραγδαία ανάπτυξη των υπολογιστικών συστημάτων και κατ' επέκταση της βιοπληροφορικής (bioinformatics), αλλά και στις καινούργιες τεχνολογίες διεργασιών υψηλής απόδοσης (high-throughput screening), η υπολογιστική βιολογία έχει επιταχύνει τους ρυθμούς με τους οποίους γίνονται γνωστές οι περίπλοκες διεργασίες των οργανισμών και των κυττάρων τους.
Συγκεκριμένα, ο κλάδος της βιοπληροφορικής συνδυάζει υπάρχουσες βιολογικές γνώσεις, πειραματικά δεδομένα έκφρασης κυττάρων και στατιστική για να προκύψουν συμπεράσματα σχετικά με τα σημαντικά στοιχεία του γονιδιώματος και του κυττάρου τα οποία ρυθμίζουν βιολογικές λειτουργίες.

1.2 Μοριακή Βιολογία και Βιολογικά Μονοπάτια

Η παρούσα διπλωματική εργασία εφαρμόζει μαθηματικές μεθόδους σε δεδομένα που προέκυψαν από βιολογικά πειράματα, ώστε να μοντελοποιηθούν διάφορες βιολογικές λειτουργίες. Για τον λόγο αυτό, είναι αναγκαίο να επεξηγηθούν, εν συντομία, κάποιες κύριες έννοιες της μοριακής βιολογίας.

1.2.1 Το κύτταρο ως σύστημα μεταβλητών

Το πρώτο πράγμα που πρέπει να σημειωθεί είναι ο τρόπος με τον οποίο η υπολογιστική βιολογία περιγράφει το κύτταρο. Ένα μαθηματικό μοντέλο αποτελείται από μεταβλητές. Για τη βιοπληροφορική, το κύτταρο είναι ένα σύστημα στο οποίο οι μεταβλητές είναι τα γονίδια. Το γονιδίωμα ενός σύνθετου πολυκύτταρου οργανισμού, όπως είναι ο άνθρωπος,

περιέχει περίπου 22000 γονίδια [1]. Σύμφωνα με το κεντρικό δόγμα της μοριακής βιολογίας που φαίνεται στο σχήμα 1.1 [2], τα γονίδια, που είναι αλληλουχίες με συγκεκριμένη λειτουργία στο DNA, μεταγράφονται σε RNA για να καταλήξουν τελικά μετά τη μετάφραση σε πρωτεΐνες. Οι πρωτεΐνες είναι τα μόρια τα οποία εκτελούν όλες τις λειτουργίες των κυττάρων, ενώ παράλληλα έχουν τη δυνατότητα να εξέλθουν από αυτά, αποτελώντας έτσι και τον τρόπο επικοινωνίας των κυττάρων με το περιβάλλον τους και τα γειτονικά κύτταρα. Ένα γονίδιο λέμε ότι εκφράζεται σε ένα κύτταρο όταν μεταγράφεται, δηλαδή όταν το κύτταρο παράγει από τη συγκεκριμένη αλληλουχια DNA, το αντίστοιχο mRNA. Σε ένα κύτταρο, ανάλογα με τον τύπο του κυττάρου (πχ. εγκεφαλικό, ηπατικό, καρκινικό) και τη λειτουργία του, υπάρχουν συγκεκριμένα γονίδια που εκφράζονται. Επομένως, αντιμετωπίζοντας το κύτταρο ως σύστημα, οι μεταβλητές είναι στην πραγματικότητα η έκφραση των γονιδίων, η οποία αλλάζει από κύτταρο σε κύτταρο και από κατάσταση σε κατάσταση. Παρακολουθώντας έναν τύπο κυττάρου σε διαφορετικές καταστάσεις, μπορούμε να βασιστούμε στις αλλαγές που παρατηρούμε στα γονίδια ώστε να κατασκευάσουμε ένα μοντέλο που να τις ικανοποιεί, και έτσι να προταθεί ένας τρόπος με τον οποίο πραγματοποιείται η αλλαγή από τη μία κατάσταση στην άλλη. Για παράδειγμα στη συγκεκριμένη εργασία θα συγκριθούν οι εκφράσεις των γονιδίων υγιών δερματικών κυττάρων με αυτές των κυττάρων μελανώματος. Τα γονίδια που έχουν διαφορετική έκφραση, επηρεάζονται από τον καρκίνο ή/και οφείλονται για αυτόν.

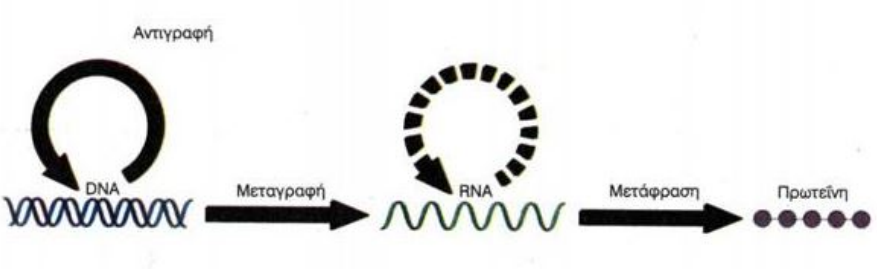
Επίσης, επειδή μιλάμε για πολύπλοκα βιολογικά συστήματα, εισάγεται σε αυτά αναγκαστικά μια τυχαιότητα, δηλαδή ακόμα και στην υποθετική περίπτωση που δημιουργηθεί ένα τέλειο μοντέλο για μια συγκεκριμένη αλλαγή, το μοντέλο δε θα αποτελεί συνάρτηση του χρόνου, δηλαδή δε θα μπορούμε γνωρίζοντας το σύστημα τη χρονική στιγμή t να υπολογίσουμε τη την κατάσταση του στη χρονική στιγμή $t+dt$. Τέτοια συστήματα ονομάζονται πολύπλοκα (complex), και μια συστηματική προσέγγιση είναι αναγκαία, σε αντίθεση με ντετερμινιστικά συστήματα που προσεγγίζονται με αναλυτικό τρόπο. Δηλαδή, το σύστημα παρατηρείται σε σταθερή κατάσταση (μετριέται η έκφραση των υγιών κυττάρων) και στη συνέχεια επιβάλλεται μία διαταραχή, ή παρατηρείται μία διαφορετική κατάσταση (μετριέται η έκφραση των κυττάρων όταν έχει εφαρμοστεί κάποιο φάρμακο σε αυτά, μετριέται η έκφραση των καρκινικών κυττάρων). Με βάση τις αλλαγές που σημειώνονται σε συγκεκριμένα γονίδια ανάμεσα στις δύο μετρήσεις, βγαίνει συμπέρασμα για το ποιες μεταβλητές του συστήματος είναι σημαντικές και ποιες όχι.

Προκειμένου να κατασκευαστεί ένα ακριβές μοντέλο σύμφωνα με την παραπάνω προσέγγιση, θα πρέπει να εισαχθούν στο σύστημα άπειρες διαταραχές όλων των μεγεθών, και στη συνέχεια με έναν υπερυπολογιστή να υπολογισθούν εμπειρικά οι σχέσεις που συνδέουν τις χιλιάδες μεταβλητές και ικανοποιούν παράλληλα τις αλλαγές των διαταραχών. Κάτι τέτοιο είναι προς το παρόν αδύνατο, αν και συνεχώς εμπλουτίζονται με καινούργια πειράματα και μελέτες τα ήδη υπάρχοντα δεδομένα, ρυθμίζοντας κατάλληλα τα προηγούμενα μοντέλα ώστε να ανταποκρίνονται στα νέα δεδομένα.

Στην παρούσα εργασία, η οποία βασίζεται σε ήδη υπάρχοντα δεδομένα από προηγούμενες μελέτες, τα κύτταρα παρατηρούνται σε δύο μόνο καταστάσεις, υγιή και καρκινικά, οπότε το μοντέλο που κατασκευάζεται ικανοποιεί μόνο τις σχέσεις που συνδέουν αυτές τις δύο καταστάσεις.

1.2.2 Βιολογικά Μονοπάτια

Όπως αναφέρθηκε στην ενότητα 1.2.1, τα κύτταρα χρησιμοποιούν πρωτεΐνες που προέρχονται από γονίδια για να εκτελέσουν τις διάφορες κυτταρικές λειτουργίες και να επικοινωνήσουν με το περιβάλλον τους. Λαμβάνοντας υπόψιν τις ιδιαίτερα περίπλοκες λειτουργίες που εκτελούνται στα κύτταρα των οργανισμών, όπως ο κυτταρικός κύκλος, ή



Σχήμα 1.1: Το κεντρικό δόγμα της μοριακής βιολογίας

η παραγωγή ινσουλίνης, γίνεται κατανοητό ότι δε δρα ένα συγκεκριμένο γονίδιο για να εκτελεστεί η κάθε μία, αλλά ένα σύνολο γονιδίων ρυθμίζεται κατάλληλα από το κύτταρο, με τα γονίδια ενεργοποιούμενα αλυσιδωτά το ένα από το άλλο. Έτσι δημιουργούνται σύνολα γονιδίων τα οποία εκτελούν συγκεκριμένες λειτουργίες, όπως η παρασκευή ενός λιπιδίου ή μιας πρωτεΐνης, η ενεργοποίηση ενός άλλου γονιδίου ή μονοπατιού, ακόμα και η εντολή προς το κύτταρο να αλλάξει θέση μέσα στο περιβάλλον του [3]. Αυτά τα σύνολα, των οποίων τα στοιχεία συνδέονται τοπολογικά μεταξύ τους, καθώς έχουμε σχέσεις ενεργοποίησης/απενεργοποίησης και συγκεκριμένη σειρά, ονομάζονται βιολογικά μονοπάτια (biological pathways) και αναπαριστούν τις λειτουργικές σχέσεις μεταξύ των πρωτεϊνικών προϊόντων των γονιδίων. Η μελέτη και η ανάλυση των βιολογικών μονοπατιών έχει κύριο ρόλο στην έρευνα της γενετικής σήμερα.
Τα βιολογικά μονοπάτια διακρίνονται σε κατηγορίες, ανάλογα με τον τύπο λειτουργίας τον οποίο ελέγχουν, οι βασικότερες από τις οποίες είναι:

Μεταβολικά μονοπάτια Υπεύθυνα για την κατασκευή ή τη διάσπαση πολύπλοκων ενζύμων ή πρωτεϊνών και ρυθμίζουν την κατανάλωση ή απορρόφηση ενέργειας από το κύτταρο. Για παράδειγμα, το πιο γνωστό μεταβολικό μονοπάτι είναι η γλυκόλυση, που αποτελεί τη διαδικασία με την οποία το κύτταρο παράγει ενέργεια από τη διάσπαση της γλυκόζης που εισέρχεται σε αυτό

Μονοπάτια ρύθμισης της γονιδιακής έκφρασης Υπεύθυνα για την ενεργοποίηση ή την απενεργοποίηση γονιδίων μέσω ρυθμιστών που μπορεί να είναι συνδυασμοί από DNA, RNA, ή πρωτεΐνες.

Σηματοδοτικά μονοπάτια Υπεύθυνα για διεργασίες κυτταρικής σηματοδότησης, μεταδίδουν πληροφορία συνήθως από το περιβάλλον (κυτταρική μεμβράνη) προς τον πυρήνα του κυττάρου και τους μεταγραφικούς παράγοντες εντός αυτού, μέσα από μία αλληλουχία μορίων και σηματοδοτικών αντιδράσεων.

1.3 Μικροσυστοιχίες DNA

Οι τεχνολογίες μικροσυστοιχιών DNA βασίζονται στη δυνατότητα να αποτεθούν δεκάδες χιλιάδες από διαφορετικές αλληλουχίες DNA σε ένα μικρό διάφανο κομμάτι γυαλί που ονομάζεται chip. Το chip είναι με τέτοιο τρόπο σχεδιασμένο, ώστε τα διαφορετικά κομμάτια DNA να τοποθετούνται σε προκαθορισμένες σειρές και στήλες και να μπορούν να ταυτοποιηθούν με βάση τη θέση τους στο πλέγμα που έχει δημιουργηθεί.
Όπως αναφέρθηκε στο 1.2.1, όταν γονίδια εκφράζονται και είναι ενεργοποιημένα, πολλά αντίγραφα mRNA που αντιστοιχούν σε αυτά τα γονίδια παράγονται μέσω τις μεταγραφής, τα οποία στη συνέχεια συνθέτουν αντίστοιχες πρωτεΐνες μέσω της μετάφρασης. Επομένως,

η παραγωγή mRNA αποτελεί την ποσοτικοποίηση της γονιδιακής έκφρασης. Για αυτό το λόγο, χρησιμοποιείται το mRNA και όχι το DNA, ως δείκτης της έκφρασης. Ο μικρός χρόνος ζωής και η γρήγορη αποσύνθεση του mRNA το καθιστούν ακατάλληλο για να μελετηθεί κατευθείαν, και έτσι πριν την απόθεση στη μικροσυστοιχία, μετατρέπεται σε cDNA. Τα κομμάτια cDNA που παράγονται κόβονται σε μικρότερα τμήματα και σηματοδοτούνται με φθορίζουσες χρωστικές. Κάθε chip έχει στην επιφάνειά του χιλιάδες ανιχνευτές (probes) [4], δηλαδή συμπληρωματικές αλληλουχίες γνωστών γονιδίων ή τμημάτων DNA, οι οποίοι εκτίθονται στο cDNA. Οι συμπληρωματικές αλληλουχίες cDNA προσδένονται στους αντίστοιχους ανιχνευτές τους, και οι υπόλοιπες απομακρύνονται. Στη συνέχεια, μέσω των χρωστικών αναγνωρίζονται οι αλληλουχίες και μετριέται η ένταση με την οποία ακτινοβολούν οι διάφοροι ανιχνευτές, ποσοτικοποιώντας με αυτόν τον τρόπο τη γονιδιακή έκφραση.

1.4 Εφαρμογή στο Μελάνωμα

Η συγκεκριμένη διπλωματική εργασία εφαρμόζει βασικές μεθόδους της υπολογιστικής βιολογίας και της βιοπληροφορικής σε συγκεκριμένα πειράματα γονιδιακής έκφρασης κυτταροσειρών μελανώματος. Όπως αναφέρθηκε στην περίληψη, το μελάνωμα αποτελεί τύπο καρκίνου, με τη μεγαλύτερη θνησιμότητα ανάμεσα στους υπόλοιπους καρκίνους του δέρματος.
Σε αντίθεση με τις περισσότερες ασθένειες οι οποίες αφαιρούν από το κύτταρο τη δυνατότητα να εκτελέσει μία συγκεκριμένη λειτουργία, όπως για παράδειγμα η παραγωγή ινσουλίνης στους διαβητικούς, ο καρκίνος έχει την ιδιαιτερότητα να αλλάζει σε μεγαλύτερο βαθμό το κύτταρο. Ο καρκίνος ουσιαστικά αποτελεί ένα άτακτο σώμα από κύτταρα, τα οποία στις περισσότερες περιπτώσεις προέρχονται από έναν πρόγονο, τα οποία έχουν χάσει βασικούς μηχανισμούς ελέγχου [5]. Ως αποτέλεσμα, αυτά τα κύτταρα επεκτείνονται συνεχώς, σχηματίζοντας μάζες, εισχωρώντας σε γειτονικούς ιστούς ή μεταναστεύοντας σε μακρινά σημεία του οργανισμού. Για να καταλήξει ένα κύτταρο να γίνει καρκινικό, πρέπει να έχουν προηγηθεί πολλές μεταλλάξεις και βλάβες στο DNA του, καθώς οι μηχανισμοί ελέγχου των κυττάρων αποτελούν από μόνοι τους ένα πολύπλοκο σύστημα που δεν καταστρέφεται τόσο εύκολα.
Όλα τα παραπάνω δείχνουν πως ο καρκίνος μπορεί να συναντηθεί σε οποιοδήποτε κύτταρο του οργανισμού, και να εκδηλωθεί διαφορετικά, ανάλογα με τα εργαλεία ελέγχου που έχουν αλλάξει ή καταστραφεί, μετατρέποντας το κύτταρο σε καρκινικό. Για αυτό, αρχικά σε επίπεδο οργάνου, η ασθένεια μελετάται ξεχωριστά (καρκίνος του μαστού, του δέρματος, των ωαρίων κλπ.) αλλά και στη συνέχεια σε επίπεδο κυτταροσειρών και ιστών. Στην εργασία, μελετώνται 144 διαφορετικές σειρές κυττάρων μελανώματος με αναφορά 6 σειρές υγιών κυττάρων δέρματος, από τέσσερα διαφορετικά πειράματα. Παρακάτω ακολουθούν πληροφορίες για το μελάνωμα από βιολογικής πλευράς.
Στον άνθρωπο, το μελάνωμα προκαλείται από βλάβες στο DNA των δερματικών κυττάρων, όπως για παράδειγμα αυτές που δημιουργούνται από την υπεριώδη ακτινοβολία των ηλιακών ακτίνων [6]. Τα καρκινικά κύτταρα προέρχονται από μελανοκύτταρα τα οποία βρισκονται στη βασική στρώση (*stratum basale*), δηλαδή στην πιο βαθιά από τις στρώσεις της επιδερμίδας. Τα μελανώματα συχνά μοιάζουν με τις απλές ελιές, και σε συγκεκριμένες περιπτώσεις αναπτύσσονται από αυτές. Στην πλειοψηφία τους είναι μαύρα ή σκούρα καφέ σημάδια στην επιδερμίδα, όπως φαίνεται στο σχήμα 1.2α΄, αλλά συνατνώνται σπανιότερα και στο χρώμα του δέρματος, ή σε κόκκινες, μωβ και μπλε αποχρώσεις. Παράλληλα, τα σχήματα των καρκινικών μελανωμάτων είναι ασύμμετρα, γεγονός που αποτελεί την κυριότερη διαφορά με τις κανονικές ελιές, οι οποίες έχουν κυκλικό ή οβάλ σχήμα, όπως

φαίνεται στο 1.2β΄.

(α΄) Ιστός μελανώματος

(β΄) Ιστός κανονικής ελιάς

Σχήμα 1.2: Τα περισσότερα μελανώματα μοιάζουν και συχνά προκύπτουν από κανονικές ελιές. Έχουν συνήθως μάυρο ή σκούρο καφέ χρώμα, και σε αντίθεση με τις ελιές, μπορεί να μην είναι συμμετρικά στο σχήμα, όπως αυτό που φαίνεται στην εικόνα. Το σχήμα και το χρώμα αποτελούν τις κύριες φαινοτυπικές διαφοροποιήσεις μεταξύ των δύο ιστών.

Από την πλευρά της μοριακής βιολογίας, έχει αποδειχθεί ότι κάποιοι τύποι του μελανώματος, *ορίζονται* από συγκεκριμένες *μεταλλάξεις-οδηγούς*, *driver* mutations, [7], οι οποίες συμβαίνουν σε πολλαπλά ογκογονίδια (δηλαδή γονίδια που συμβάλλουν στον πολλαπλασιασμό του κυττάρου, και σε περίπτωση αποτυχίας ελέγχου τους από το κύτταρο μπορεί να οδηγήσουν σε καρκίνο). Οι πιο σημαντικές από αυτές τις μεταλλάξεις είναι:

BRAF Η μετάλλαξη συμβαίνει στον κώδωνα 600 του γονιδίου BRAF. Ο όρος κώδωνας χρησιμοποιείται στη βιολογία για να περιγράψει μία αλληλουχία τριών νουκλεοτιδίων DNA ή RNA που αντιστοιχεί σε ένα συγκεκριμένο αμινοξύ, ή σε αλληλουχία τερματισμού της πρωτεϊνοσύνθεσης. Ουσιαστικά, οι κώδωνες είναι ο *μετασχηματισμός* από τη γλώσσα των νουκλεϊκών οξέων (DNA, RNA), που αποτελείται από 4 νουκλεοτίδια, στη γλώσσα των πρωτεϊνών, που αποτελείται από 20 αμινοξέα. Η μετάλλαξη που λαμβάνει χώρα έχει ως συνέπεια τη μετατροπή του αντίστοιχου αμινοξέως του κώδωνα από **βαλίνη σε γλουταμινικό οξύ**, που συμβολίζεται **V600E**. Περισσότερα από 84.6% των BRAF-μεταλλαγμένων μελανωμάτων περιέχουν τη συγκεκριμένη μετάλλαξη, ενώ λιγότερο συνήθεις BRAF-μεταλλάξεις είναι από **βαλίνη σε λυσίνη**, με συμβολισμό **V600K** (περίπου 7.7% των περιπτώσεων), και από **βαλίνη σε αργινίνη**, **V600R** (1%) [8]. Περίπου $37-50\%$ των μελανωμάτων οδηγούνται από μετάλλαξη στο BRAF γονίδιο.

NRAS Η αμέσως επόμενη πιο συχνή οδηγός-μετάλλαξη του μελανώματος συμβαίνει στο γονίδιο NRAS [9]. Το γονίδιο αυτό αποτελεί το πρώτο ογκογονίδιο μελανώματος, και σήμερα συναντάται με συχνότητα $13-25\%$ [10]. Η μετάλλαξη που λαμβάνει χώρα σε περισσότερες από 80% περιπτώσεις NRAS μετάλλαξης, οδηγεί στην αντικατάσταση **από γλουταμίνη σε λευκίνη** στον κώδωνα 61 του γονιδίου, ενώ με μικρότερη συχνότητα παρατηρούνται μεταλλάξεις στις θέσεις 12 και 13 [11].

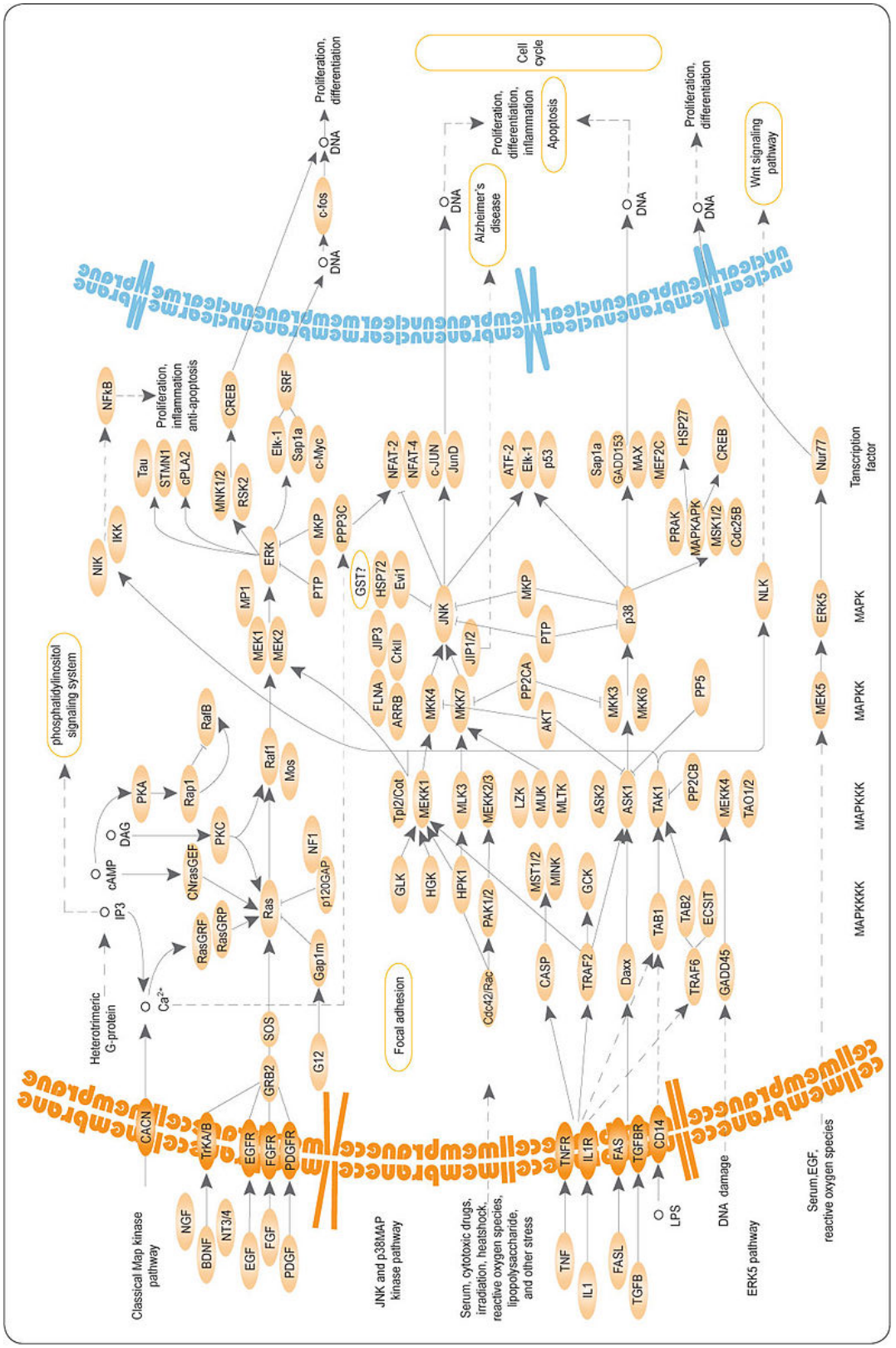
PTEN Η λιγότερο μελετημένη μετάλλαξη του γονιδίου PTEN στο μελάνωμα, δεν έχει εξακριβωμένη συχνότητα εμφάνισης, αλλά θεωρείται ότι κυμαίνεται στο $15-28\%$. Οι μεταλλάξεις που συμβαίνουν στο γονίδιο έχουν βρεθεί σε διάφορα σημεία του, εκτός από το εξόνιο 9, με σημαντικό σημείο το εξόνιο 5 [12]. Ως εξόνιο ορίζεται το τμήμα του γονιδιακού DNA το οποίο θα μεταγραφεί σε RNA αφού τα εσόνια έχουν απομα-

κρυνθεί, δηλαδή τα τμήματα που περιέχονται στο γονίδιο αλλά δε μεταγράφονται σε RNA. Οι κώδωνες του RNA αποτελούνται εξόλοκλήρου επομένως από εξόνια, δηλαδή το εξόνιο 5 του γονιδίου PTEN αντιστοιχεί και στον κώδωνα 5.

MEK1 ή αλλιώς **MAP2K1** είναι γονίδιο που κωδικοποιεί πρωτεΐνη η οποία περιλαμβάνεται στην οικογένεια των ενζύμων που ονομάζονται MAP κινάσες με μιτογόνο ενεργοποίηση. Περισσότερα για αυτές ακολουθούν παρακάτω. Οι μεταλλάξεις του τύπου MEK1 παρατηρούνται σε $6-7\%$ των μελανωμάτων.

Λοιπές μεταλλάξεις Όπως αναφέρθηκε, ο καρκίνος είναι αποτέλεσμα αρκετών μεταλλάξεων, πολλές από τις οποίες δεν έχουν ακόμη μελετηθεί ή ανακαλυφθεί. Με την εξέλιξη της μοριακής βιολογίας, καινούργια ευρήματα βοηθούν στην καλύτερη και αναλυτικότερη εικόνα που έχουμε για την ασθένεια, και οδηγούν σε βελτιωμένες θεραπείες [13]. Τυπικά αναφέρονται κάποιες μεταλλάξεις του μελανώματος με συχνότητα μικρότερα από 6%, στα γονίδια **KIT, CTNNB1, GNA11, GNAQ**.

Είναι σημαντικό να σημειωθεί ότι όλα τα παραπάνω γονίδια είναι συστατικά στοιχεία του βιολογικού μονοπατιού **MAPK\ERK**, ή αλλιώς **Ras-Raf-MEK-ERK** που αποτελεί ένα από τα πιο σημαντικά σηματοδοτικά μονοπάτια του κυττάρου, πρακτικά μία αλυσίδα πρωτεϊνών οι οποίες μεταφέρουν ένα σήμα από τους αισθητήρες της κυτταρικής μεμβράνης, στο DNA του πυρήνα. Το σήμα ξεκινά να μεταδίδεται όταν ένα μόριο αντιδρά και συνδέεται με τον αισθητήρα στην επιφάνεια του κυττάρου [3]. Ο αισθητήρας ενεργοποιεί κάποια πρωτεΐνη που στη συνέχεια ενεργοποιεί μία επόμενη, μέχρι το σήμα να φτάσει μέσω τέτοιων αλυσιδωτών σχέσεων στον πυρήνα, όπου ενεργοποιείται κάποιο αντίστοιχο τμήμα του DNA και ξεκινάει να μεταφράζεται στην πρωτεΐνη του. Οι κινάσες MAP φωσφορυλιώνουν κατάλοιπα σερίνης\θρεονίνης, δηλαδη τροποποιούν τη λειτουργία τους είτε αυξάνοντας είτε μειώνοντας τη δραστικότητά τους, και ανταποκρίνονται σε εξωκυττάρια αυξητικά ερεθίσματα, όπως η αυξητική ορμόνη, ο Epidermal Growth Factor (EGF), ο Platelet-derived Growth Factor (PDGF), και η ινσουλίνη. Το μονοπάτι MAPK\ERK ενεργοποιείται όταν στον επίπεδο του υποδοχεά οι ras GTPases ενεργοποιούν τη Raf κινάση, που ενεργοποιεί τη MEK και τέλος ενεργοποείται η ERK που είναι σε θέση να ρυθμίσει τη γονιδιακή έκφραση. Ουσιαστικά η κάθε κινάση λειτουργεί σαν ένας διακόπτης on\off για την επόμενη. Άρα, όταν ένας από αυτούς τους *διακόπτες* μείνουν σε μία θέση λόγω μετάλλαξης, το μονοπάτι κατάντι της μεταλλαγμένης κινάσης είναι μόνιμα ενεργοποιημένο, ενεργοποιώντας κατέπέκταση σημαντικά γονίδια του κυτταρικού κύκλου, και συμβάλλοντας στη μετατροπή του κυττάρου σε καρκινικό. Επομένως, τα επιμέρους γονίδιά του είναι σημαντικά στη μελέτη του καρκίνου, και αποτελούν πιθανούς στόχους για την καταστολή του. Το μονοπάτι MAPK παρουσιάζεται στο σχήμα 1.3.



Σχήμα 1.3: Το σηματοδοτικό μονοπάτι MAPK. Εικόνα από την Abgent, γραφικά από τον wikipedia user *kosigrim*

1.5 Σκοπός

Σκοπός της παρούσης διπλωματικής είναι η δημιουργία ενός εννιαίου *πάνελ* σειρών μελανώματος, αντιπροσωπευτικού της ασθένειας, και η χρήση αυτού προς ανακάλυψη ουσιών/φαρμάκων με δυνητικά θεραπευτική δράση καθώς και προς δημιουργία μοντέλου διάκρισης καρκινικών από υγιών κυττάρων και διαφορετικών μεταλλάξεων μελανώματος μεταξύ τους. Η ανάγκη για τη δημιουργία αυτού του ευρύτερου πάνελ προέρχεται από το γεγονός ότι, το εργαστήριο εμβιομηχανικής και βιοτεχνολογίας του ΕΜΠ είναι εξοπλισμένο με συγκεκριμένες κυτταροσειρές μελανώματος. Η υπολογιστική μελέτη είναι αναγκαία πριν ξεκινήσουν τα βιολογικά πειράματα, προκειμένου να εντοπιστούν οι υποψήφιες ουσίες που θα εισαχθούν στις κυτταρικές καλλιέργειες. Η υπολογιστική μελέτη που γίνεται όμως, βασίζεται σε διαθέσιμα στο κοινό δεδομένα από πειράματα της βιολογικής κοινότητας τα οποία προφανώς δεν περιέχουν απαραίτητα τις παραπάνω κυτταροσειρές. Με την ενσωμάτωση των δεδομένων από πολλά πειράματα, τα οποία περιέχουν κάποιες από τις κυτταροσειρές το καθένα, γίνεται δυνατή η υπολογιστική μελέτη. Πιο αναλυτικά, η ενσωμάτωση δεδομένων από πολλά πειράματα, αν κριθεί επιτυχής, θα αυξήσει την αξιοπιστία της ανάλυσης καθώς θα συμπεριληφθούν σε αυτή περισσότερες κυτταροσειρές και άρα περισσότερη ποικιλία καρκινογενών μηχανισμών. Στη συνέχεια, ακολουθεί επαναδειγματοληψία ώστε να εισαχθεί ισορροπία στις συγκρίσεις μεταξύ υγιών και καρκινικών δειγμάτων (υπάρχουν 144 καρκινικές κυτταροσειρές έναντι 6 υγιων), και εξάγονται συμπεράσματα για τα στατιστικά διαφορικώς εκφρασμένα γονίδια. Τα σημαντικότερα εξάυτών επιλέγονται ως features σε μοντέλο μηχανών διανυσμάτων υποστήριξης (SVM) ώστε να υπολογιστεί η ακρίβεια με την οποία είναι δυνατό να διακριθουν υγιή και καρκινικά κύτταρα, και μοντέλο γραμμικής διακριτικής ανάλυσης (LDA) ανάλογα για την κατηγοριοποίηση των κυτταροσειρών με βάση τη μετάλλαξη μελανώματος που τα χαρακτηρίζει. Σε τελικό στάδιο, τα δείγματα φιλτράρονται ώστε να γίνει ανάλυση συγκεκριμένα στις κυτταροσειρές που περιλαμβάνονται στο εργαστήριο και να βρεθούν υποψήφιες ουσίες για δοκιμή.

Κεφάλαιο 2

Μέθοδοι

2.1 Η γλώσσα προγραμματισμού R και το Bioconductor

Ολόκληρο το προγραμματιστικό και υπολογιστικό κομμάτι, καθώς και τα διαγράμματα υλοποιήθηκαν στη γλώσσα προγραμματισμού R [14]. Η R χρησιμοποιείται για στατιστικούς υπολογισμούς και δημιουργία γραφημάτων. Είναι ένα GNU project που αναπτύχθηκε στα Bell Laboratories από τους John Chambers και συνεργάτες. Ως γλώσσα προγραμματισμού, προσφέρει μία ευρεία ποικιλία στατιστικών τεχνικών (γραμμική και μη γραμμική μοντελοποιήση, κλασικά στατιστικά τεστ, ομαδοποίηση, κατηγοριοποίηση και άλλα), καθώς και γραφικών τεχνικών, δίνοντας τη δυνατότητα να απεικονισθούν τα αποτελέσματα των στατιστικών αποτελεσμάτων. Καθώς αποτελεί ένα ανοιχτό και δημόσιο project, ο κάθε χρήστης μπορεί να δημιουργήσει τους δικούς του κώδικες και εν συνεχεία να διανείμει στους υπόλοιπους χρήστες σε μορφή πακέτου (package) το οποίο ανεβαίνει στο αποθετήριο CRAN της R.
Παράλληλα με το αποθετήριο CRAN, υπάρχει το πιο εξειδικευμένο Bioconductor [15] το οποίο αποτελεί ένα σύνολο πακέτων/εργαλείων για την ανάλυση βιολογικών γενετικών δεδομένων που προέρχονται από διεργασίες υψηλής απόδοσης. Εκτός από τα πακέτα εργαλείων, υπάρχει και μεγάλος αριθμός πακέτων μετα-δεδομένων που παρέχουν σχόλια (annotations), δηλαδή επίσημες επιβεβαιωμένες βιολογικές πληροφορίες για διάφορους οργανισμούς, βιολογικά μονοπάτια, γονίδια, μικροσυστοιχίες και άλλα. Η χρήση της R στον τομέα της βιοπληροφορικής είναι πλέον άμεσα συνυφασμένη με τα πακέτα του Bioconductor, η γνώση και εξοικείωση των οποίων αποτελούν βασικές προϋποθέσεις για όλους τους επιστήμονες και ερευνητές που θέλουν να ασχοληθούν με την υπολογιστική βιολογία. Τα πακέτα της R που χρησιμοποιήθηκαν για την ανάλυση είναι περιληπτικά τα παρακάτω:

tidyverse [16]: πακέτο γενικής χρήσης στην R, για την εύκολη διαχείριση δεδομένων μεγάλου όγκου. Περιλαμβάνει τα υποπακέτα ggplot, ggplot2 [17] μέσω των οποίων έγιναν ευθέως μέσα στην R όλα τα σχήματα και διαγράμματα της παρούσας διπλωματικής εργασίας.

GEOquery [18]: πακέτο εξοπλισμένο με συναρτήσεις οι οποίες δρουν ως *γέφυρα* μεταξύ του GEO, και του Bioconductor. Με βασικές συναρτήσεις του πακέτου, ήταν εφικτή η λήψη των απαραίτητων δεδομένων των πειραμάτων, προκειμένου να ξεκινήσει η υπολογιστική μελέτη.

affy [19]: πακέτο για την εισαγωγή και την προεπεξεργασία δεδομένων που έχουν ληφθεί

από το GEO. παρέχει πλήθος μεθόδων για την αφαίρεση θορύβου και την κανονικοποίηση των δεδομένων.

sva [20]: πακέτο για τη διόρθωση σφάλματος ομάδας (batch effect), που χρησιμοποιείται για την ενσωμάτωση των πειραμάτων σε ένα εννιαίο μητρώο.

limma [21]: πακέτο για τη διαφορική γονιδιακή ανάλυση δεδομένων.

e1071 [22]: πακέτο που περιέχει πλήθος αλγορίθμων και συναρτήσεων χρήσιμων για τη στατιστική ανάλυση και τις μηχανές εκμάθησης. Στην εργασία χρησιμοποιήθηκε για την εκτέλεση του svm αλγορίθμου της ενότητας 2.8.

MASS [23]: πακέτο που περιέχει τον αλγόριθμο για την εκτέλεση γραμμικής διακριτικής ανάλυσης (LDA) για την κατηγοριοποίηση των δειγμάτων με βάση την οδηγό-μετάλλαξή τους.

2.2 Gene Expression Omnibus

Τα δεδομένα τα οποία χρησιμοποιήθηκαν στη διπλωματική εργασία δεν προέκυψαν από κάποιο πείραμα εντός του εργαστηρίου εμβιομηχανικής του κατασκευαστικού τομέα ΕΜΠ, αλλά λήφθηκαν από διαδικτυακή βάση δεδομένων. Η επιστημονική κοινότητα της υπολογιστικής βιολογίας χρησιμοποιεί συγκεκριμένες βάσεις δεδομένων για να αποθηκεύει τον τεράστιο όγκο αποτελεσμάτων που προκύπτουν από τις διεργασίες υψηλής απόδοσης, ώστε να είναι άμεσα διαθέσιμα σε όλους.
Δύο βασικές τέτοιες βάσεις δεδομένων είναι το Gene Expression Omnibus (GEO) και το ArrayExpress, στα οποία βρίσκονται αποθηκευμένα περισσότερα από 100 χιλιάδες πειράματα, η πλειοψηφία των οποίων προέρχεται από πειράματα μικροσυστοιχιών. Το κάθε πείραμα έχει συγκεκριμένη δομή, η οποία αποτελεί ξεχωριστή κλάση αντικειμένου στην R, που περιέχει πληροφορίες (metadata) για το ίδιο το πείραμα και την πλατφόρμα του πειράματος, τα γονίδια που παρακολουθούνται (feature data) και τα δείγματα που συμπεριλαμβάνονται στο πείραμα (pheno data). Με βάση αυτές τις πληροφορίες, από κάθε πείραμα, διαλέγονται τα δείγματα και τα γονίδια τα οποία χρησιμοποιούνται για την ανάλυση και τελικά προκύπτει ένας πίνακας με τα γονίδια στις γραμμές του και τα δείγματα στις στήλες του.
Το GEO παρέχει τα δεδομένα του σε συγκεκριμένες δομές, κάθε μία από τις οποίες συνδέεται με τις υπόλοιπες με συγκεκριμένη ιεραρχία. Ανάλογα με τις ανάγκες κάθε χρήστη, γίνονται διαφορετικές αναζητήσεις στο αποθετήριο, με φίλτρο τη δομή των δεδομένων και λέξεις κλειδιά. Οι δομές του GEO είναι:

- Platform
- Sample
- Series
- DataSet

Τα οποία θα αναλυθούν παρακάτω.

2.2.1 GEO Platform

Τα αρχεία δομής πλατφόρμας παρέχονται από άλλους χρήστες και περιέχουν πληροφορίες για τις τεχνολογίες που χρησιμοποιούνται, όπως οι πολλές διαφορετικές μικροσυστοιχίες

ή πιο καινούργια μηχανήματα που χρησιμοποιούν διαφορετικές μεθόδους για τη μέτρηση της γονιδιακής έκφρασης. Οι πληροφορίες που παρέχονται από τα GEO platforms είναι πολύ σημαντικές καθώς μπορεί ο χρήστης να δει κατευθείαν ποιους ανιχνευτές περιλαμβάνει η πλατφόρμα και τι είδους είναι αυτοί. Παράλληλα, κάθε πλατφόρμα περιέχει δεδομένα για κάθε ανιχνευτή της, η πιο σημαντική από τις οποίες είναι το γονίδια (ή τα γονίδια) στα οποία αντιστοιχεί, τα οποία μπορεί να δηλώνονται μέσω του επίσημου συμβόλου HUGO, του Entrez Gene ID ή άλλες ονοματολογίες. Άλλες πληροφορίες είναι ο οργανισμός στον οποίο αντιστοιχεί η συμπληρωματική αλληλουχία DNA του ανιχνευτή και η ημερομηνία στην οποία καταγράφηκε επίσημα το αντίστοιχο γονίδιο. Τα αρχεία πλατφόρμας έχουν το πρόθεμα GPL και στη συνέχεια ακολουθεί ο κωδικός της πλατφόρμας. Το GEO δίνει επίσης τη δυνατότητα προβολής όλων των δεδομένων τα οποία έχουν γίνει σε κάθε πλατφόρμα, διευκολύνοντας έτσι τον χρήστη.

2.2.2 GEO Sample

Τα αρχεία samples (δειγμάτων) του GEO παρέχονται από χρήστες όπως και τα αρχεία πλατφόρμας και περιέχουν πληροφορίες που αφορούν ένα δείγμα από κάποιο πείραμα. Συνήθως, σε πειράματα συμμετέχουν πολλά δείγματα, ώστε να αυξηθεί η στατιστική δύναμη κατά την υπολογιστική μελέτη. Για παράδειγμα, ένα δείγμα μπορεί να είναι η μέτρηση μίας συγκεκριμένης κυτταροσειράς αφού έχει εκτεθεί πρώτα σε κάποιο φάρμακο ώστε να παρατηρηθεί η δράση του φαρμάκου. Τα δεδομένα από ένα δείγμα είναι η ημερομηνία του πειράματος, οι συνθήκες υπό τις οποίες έγινε (ατμοσφαιρική πίεση, θερμοκρασία, χρόνος επώασης, μέσο καλλιέργειας και οτιδήποτε άλλο κρίνει ο χρήστης ότι μπορεί ενδεχομένως να επηρεάσει τα αποτελέσματα του πειράματος). Ανάλογα με το αρχείο, αυτό μπορεί να περιέχει και τις μετρήσεις από κάποιο πείραμα. Κάθε δείγμα ανήκει μοναδικά σε κάποιο πείραμα, δηλαδή στην περίπτωση κυτταροσειρών, μπορεί να έχουμε χιλιάδες δείγματα από κάποια ίδια συγκεκριμένη κυτταροσειρά που τα εργαστήρια αγοράζουν από προμηθευτές, αλλά ακόμα κι αν πρόκειται για βιολογικά αντίγραφα που συμμετείχαν στο ίδιο πείραμα, βρίσκονται στο GEO ως διαφορετικές καταχωρίσεις. Οι καταχωρήσεις δειγμάτων έχουν το πρόθεμα GSM το οποίο ακολουθείται από τον κωδικό της αντίστοιχης καταχώρησεις, μοναδικό για κάθε δείγμα. Έτσι, τα GEO Samples αποτελούν τη δομική μονάδα της βάσης δεδομένων του GEO.

2.2.3 GEO Series

Η πιο πολύπλοκη και πιο χρήσιμη δομή της βάσης δεδομένων του GEO είναι τα GEO Series δηλαδή οι καταχωρήσεις πειραμάτων (σειρών). Μία σειρά αποτελεί το σύνολο των αποτελεσμάτων από τις μετρήσεις διαφόρων δειγμάτων που χρησιμοποιήθηκαν σε κάθε δείγμα, και συνήθως ένα πείραμα μπορεί να περιγραφεί πλήρως μέσω μίας καταχώρησης σειράς. Όπως γίνεται κατανοητό, σε αυτές τις καταχωρήσεις περιέχονται όχι μόνο τα αποτελέσματα των μετρήσεων του πειράματος, αλλά και τα δεδομένα για τα δείγματα, καθώς και για την πλατφόρμα που χρησιμοποιήθηκε. Επομένως, τα GEO Series περιλαμβάνουν ολες τις απαραίτητες πληροφορίες για τη διεξαγωγής υπολογιστικής μελέτης ενός πειράματος. Όπως και οι προαναφερθήσες δομές, έτσι και οι σειρές παρέχονται εξόλοκλήρου από τους χρήστες, το οποίο εισάγει ένα βαθμό δυσκολίας γιατί μπορεί να δίνονται περισσότερες ή λιγότερες πληροφορίες από όσες είναι απαραίτητες, από σειρά σε σειρά.
Σε περίπτωση που γίνεται ανάλυση πολλών διαφορετικών σειρών, όπως στην παρούσα διπλωματική εργασία, επιπλέον στατιστικά εργαλεία και μέθοδοι κρίνονται απαραίτητα, προκειμένου να επεξεργαστούν τα δεδομένα σε πρώτο επίπεδο ώστε να είναι άμεσα συγκρίσιμα από πείραμα σε πείραμα. Πρέπει να σημειωθεί, ότι ακόμα κι αν ένα πείραμα

αναπαραχθεί σε ακριβώς ίδιες συνθήκες με τα ακριβώς ίδια δείγματα και τεχνολογίες, οι μετρήσεις θα είναι διαφορετικές από το αρχικό. Γίνεται έτσι κατανοητό πως οι διαφορές αυξάνονται δραματικά όταν μελετούνται πειράματα που διεξήχθησαν σε διαφορετικές πλατφόρμες υπό διαφορετικές συνθήκες. Οι μέθοδοι που χρησιμοποιούνται στη μελέτη για την ενσωμάτωση πολλών διαφορετικών πειραμάτων περιγράφονται εκτενώς στην ενότητα 2.5. Η δομή μιας σειράς του GEO αποτελείται από τρία βασικά μέρη:

Assay data: περιλαμβάνει τα δεδομένα, δηλαδή τις μετρήσεις. Αυτές δεν είναι άμεσα προσβάσιμες από τον χρήστη, αλλά με μια μικρή επεξεργασία και βασικές εντολές της R μετατρέπονται σε έναν ευανάγνωστο πίνακα με στήλες τα διαφορετικά δείγματα και γραμμές τους ανιχνευτές της πλατφόρμας στην οποία έγινε το πείραμα. Τα assay data μιας σειράς αποτελούν τα μόνα *πραγματικά* δεδομένα-μετρήσεις του πειράματος. Τα υπόλοιπα δομικά μέρη, όπως αναφέρεται ακριβώς παρακάτω, αφορούν *μετα-δεδομένα* (metadata) τα οποία φέρουν πληροφορίες σχετικές με τα συστατικά του πειράματος, όπως η πλατφόρμα, τα δείγματα.

Pheno data: ο όρος pheno παραπέμπει στο βιολογικό όρο **φαινότυπος**. Σε αυτό το κομμάτι των GEO Series αναφέρονται πληροφορίες για κάθε δείγμα, όπως ημερομηνία μέτρησης, οργανισμός, τύπος ιστού, πρώτη επίσημη δημοσίευση του δείγματος, τύπος δείγματος (υγιές, καρκινικό, με φάρμακο κλπ.) και οτιδήποτε άλλο θεωρεί χρήσιμο ή απαραίτητο ο χρήστης που ανεβάζει το πείραμα.

Feature data: ο όρος feature στην υπολογιστική βιολογία συνήθως παραπέμπει στους ανιχνευτές που παρέχονται από την πλατφόρμα του πειράματος. Στα feature data παρέχονται όλες οι πληροφορίες που μία πλατφόρμα δίνει για τους ανιχνευτές της. Η πιο σημαντική από αυτές είναι προφανώς τα γονίδια που σχετίζονται με κάθε ανιχνευτή. Ιδανικά, η αντιστοιχία γονιδίου και ανιχνευτή πρέπει να είναι 1:1, κάτι τέτοιο όμως απαιτεί μεγάλη ευαισθησία, όταν πρόκειται για 22000 ανιχνευτές. Έτσι, ένα ποσοστό των ανιχνευτών αντιστοιχούν σε περισσότερα από ένα γονίδια, και αντίστροφα, ένα ποσοστό γονιδίων εντοπίζονται από περισσότερους από έναν ανιχνευτές. Η δεύτερη περίπτωση δεν αποτελεί πρόβλημα για τη μελέτη, καθώς διαλέγεται ο μέσος όρος ή η διάμεσος των μετρήσεων των ανιχνευτών με αξιόπιστα αποτελέσματα (συνήθως τα δύο στατιστικά μεγέθη είναι πολύ κοντά μεταξύ τους και έτσι δεν έχει ιδιαίτερη σημασία για την υπόλοιπη μελέτη ποιο απο τα δύο θα επιλεχθεί). Ωστόσο, όταν ένας ανιχνευτής αντιστοιχεί σε πολλά γονίδια, δεν υπάρχει κάποιος παρόμοιος απλός τρόπος για να βγει συμπέρασμα για κάθε ένα από τα γονίδια, και οι μέθοδοι που χρησιμοποιούνται ξεπερνούν τις απαιτήσεις της συγκεκριμένης μελέτης. Μετά το φιλτράρισμα των γονιδιών που αντιστοιχούν σε έναν ανιχνευτή, ανάλογα με την πλατφόρμα, ο πίνακας των μετρήσεων (assay data) μπορεί να καταλήξει μέχρι και με τις μισές γραμμές, αλλά πλέον θα πρόκειται για πίνακα που στις γραμμές τους θα έχει γονίδια και όχι ανιχνευτές. Μόνο τότε είναι σε θέση κάποιος να συνεχίσει τη μελέτη, καθώς οι ανιχνευτές δεν έχουν κάποια βιολογική σημασία, και τα αποτελέσματα που βγάζει κάποιος με κατευθείαν μελέτη πάνω στους ανιχνευτές, δεν μπορούν να θεωρηθούν αξιόπιστα. Αξίζει να σημειωθεί ότι υπάρχουν πλατφόρμες ειδικά σχεδιασμένες για πολύ συγκεκριμένα πειράματα, οι οποίες δίνουν 1:1 αντιστοιχία μεταξύ ανιχνευτών και γονιδίων, όχι αυξάνοντας την ευαισθησία αλλά αφαιρώντας εντελώς ανιχνευτές οι οποίοι συμπληρώνουν ή συμπληρώνονται από πολλά γονίδια.

2.3 Προεπεξεργασία δεδομένων

Το πρώτο βήμα μετά τη λήψη των δεδομένων των πειραμάτων από τη βάση δεδομένων και της επιλογής των δειγμάτων και των γονιδίων προς ανάλυση είναι η προεπεξεργασία τους (preprocessing). Χωρίς την προεργασία, δεν είναι δυνατό ο χρήστης να δει απευθείας τις μετρήσεις γονιδιακής έκφρασης του πειράματος, πόσο μάλλον να διεξάγει κάποια υπολογιστική μελέτη πάνω σε αυτές. Με την εξέλιξη της στατιστικής λόγω του μεγάλου όγκου δεδομένων που παράγονται πλέον από τα πειράματα, αναπτύσσονται συνεχώς τρόποι προεπεξεργασίας γονιδιακών δεδομένων. Στα πλαίσια της εργασίας, προτιμήθηκαν αξιόπιστες και εύρωστες μέθοδοι, οι οποίες χρησιμοποιούνται ευρέως ειδικά για δεδομένα γονιδιακής έκφρασης με ικανοποιητικά αποτελέσματα.
Το πακέτο της R που χρησιμοποιήθηκε για την προεπεξεργασία είναι το affy [19] με τη συνάρτηση rma, (Robust Multi-Chip Average) [24] η οποία εκτελεί διόρθωση θορύβου και στη συνέχεια κανονικοποίηση των μετρήσεων με τις μετασχηματισμένες log2 τιμές τους.

2.3.1 Αφαίρεση θορύβου

Οι μικροσυστοιχίες αποτελούνται όπως αναφέρθηκε στο 1.3 από ανιχνευτές. Για την ακρίβεια, κάθε ανιχνευτής αποτελείται από μία αλληλουχία 16-20 βάσεων PM (perfect match) και άλλη μία σχεδόν πανομοιότυπη, με μόνο μία βάση στη μέση αλλαγμένη MM (mismatch). Για το background correction (αφαίρεση θορύβου) από τα ακατέργαστα δεδομένα μίας μικρόσυστοιχίας, ο αλγόριθμος RMA λαμβάνει υπ'όψιν μόνο την ένταση που μετριέται από τους ανιχνευτές PM [25]. Συγκεκριμένα, θεωρείται ότι η ένταση αποτελείται από το πραγματικό σήμα και από τον θόρυβο που εισάγει το σύστημα:

$$\underbrace{PM_{ijn}}_{\text{ένταση ανιχνευτή}} = \overbrace{bg_{ijn}}^{\text{θόρυβος}} + \underbrace{s_{ijn}}_{\text{σήμα ανιχνευτή}} \tag{2.1}$$

Όπου:

- i το δείγμα RNA που μετριέται
- j ο ανιχνευτής που μελετάται
- n το σετ ανιχνευτών στο οποίο ανήκει ο ανιχνευτής j

Ο θόρυβος θεωρείται ότι προέρχεται από τον οπτικό θόρυβο που εισάγεται από την κάμερα της μικροσυστοιχίας που χρησιμοποιείται για να μετρήσει την ένταση που εκπέμουν οι ανιχνευτές, καθώς και από τις μετρήσεις που προέρχονται απο μη συμπληρωματικές αλληλουχίες που συνδέθηκαν με τους ανιχνευτές. Θεωρείται πως σε κάθε μικροσυστοιχία για ένα συγκεκριμένο δείγμα, ο θόρυβος έχει μία μέση τιμή $E(bg_{ijn}) = \beta_i$. πρέπει επομένως να γίνει κάποια ρύθμιση των PM τιμών ώστε να αφαιρεθεί το background effect [26].
Ο αλγόριθμος RMA χρησιμοποιεί τη συνελιγμένη διόρθωση υποβάθρου (convolution background correction) με τον παρακάτω κλειστό μετασχηματισμό $B(\cdot)$ [27]:

$$\begin{aligned} &B(PM_{ijn}) = E[s_{ijn}|PM_{ijn}] > 0 \\ &s_{ijn} \backsim Exp(\kappa_{ijn}) \qquad bg_{ijn} \backsim N(\beta_i, \sigma_i^2) \end{aligned} \tag{2.2}$$

Δηλαδή το η μέτρηση αφού έχει αφαιρεθεί ο θόρυβος ισοδυναμεί με την αναμενόμενη τιμή του σήματος, όταν υπάρχει τιμή PM_{ijn} από τον ανιχνευτή, με την υπόθεση ότι το σήμα ακολουθεί εκθετική κατανομή, και ο θόρυβος κανονική τιμή με μέση τιμή β_i και

τυπική απόκλιση σ^2. Το μοντέλο μπορεί να βελτιστοποιηθεί, αλλά οι παραπάνω συνθήκες έχουν εμπειρικά αποδειχθεί αξιόπιστες με ικανοποιητικά αποτελέσματα. Να σημειωθεί ότι καθώς μετριέται ένταση φωσφορισμού από τους ανιχνευτές, το σήμα και επομένως και ο μετασχηματισμός είναι πάντα θετικά $s_{ijn} > 0 \Rightarrow B(PM_{ijn}) > 0$.

2.3.2 Κανονικοποίηση

Μετά την αφαίρεση του θορύβου με τη διόρθωση υποβάθρου, είναι αναγκαίο να κανονικοποιηθούν τα δεδομένα κάθε πειράματος, μέσα στο ίδιο το πείραμα. Ο λόγος για τον οποίο γίνεται αυτό είναι ότι ανάμεσα σε δείγματα, ακόμα κι αν αυτά έχουν μετρηθεί με την ίδια μικροσυστοιχία, υπάρχει διακύμανση των μετρήσεων, η οποία οφείλεται είτε σε βιολογικές διαφορές, οι οποίες πρέπει να ληφθούν υπόψιν και να διατηρηθούν στη μελέτη που θα ακολουθήσει, είτε σε μη-σημαντικές διαφορές, μη-βιολογικές, οι οποίες είναι στατιστικής φύσης και πρέπει να αφαιρεθούν μέσω της κανονικοποίησης. Όπως και με τη διόρθωση υποβάθρου και όλες τις μεθόδους προεπεξεργασίας δεδομένων, και εδώ υπάρχει πλήθος μεθόδων που χρησιμοποιούνται σε κάθε περίπτωση.
Για δεδομένα που προέρχονται από μικροσυστοιχίες τύπου Affymetrix όπως αυτά που χρησιμοποιούνται στην παρούσα διπλωματική εργασία, η κανονικοποίηση ποσοστημορίων (quantile normalization), είναι ταυτόχρονα πολύ γρήγορη, υπολογιστικά απλή και αξιόπιστη με καλά αποτελέσματα [28]. Η μέθοδος είναι πλήρως ενσωματωμένη στον αλγόριθμο rma του πακέτου affy, ακολουθώντας τη διόρθωση υποβάθρου. Ο όρος ποσοστημόριο στη στατιστική χρησιμοποιείται για να περιγράψει σημεία τα οποία χωρίζουν μία κατανομή πιθανοτήτων σε διαδοχικά μέρη με την ίδια πιθανότητα. Προφανώς, έχουμε ένα λιγότερο ποσοστημόριο από ότι μέρη που δημιουργούνται, και το εμβαδό κάθε μέρους ανάμεσα στην καμπύλη της κατανομής και τον άξονα x, το οποίο αποτελεί την πιθανότητα του συγκεκριμένου μέρους, είναι ίδιο και ίσο με $\frac{1}{q}$, όπου q ο αριθμός των ποσοστημορίων.
Με την κανονικοποίηση ποσοστημορίων, ο στόχος είναι οι κατανομές εντάσεων των ανιχνευτών σε δύο ή περισσότερα δείγματα με την ίδια μικροσυστοιχία να γίνουν ίδιες. Ένα γράφημα ποσοστημορίων (quantile-quantile plot) δείχνει ότι οι κατανομές δύο διανυσμάτων δεδομένων είναι ίδιες όταν το γράφημα είναι μία ευθεία διαγώνιος γραμμή κλίσης 1, και άνισες όταν όλα τα σημεία δε βρίσκονται πάνω σε αυτή τη διαγώνιο. Για την περίπτωση γονιδιακών δεδομένων κάθε δείγμα αποτελεί ένα διάνυσμα με δεδομένα όλους τους ανιχνευτές που έχουν μετρηθεί. Ανάλογα επεκτείνεται η παρατήρηση σε n διανύσματα, όπου κατασκευάζοντας νοητά το γράφημα ποσοστιμορίων σε n διαστάσεις, όλα τα δεδομένα βρίσκονται πάνω στη διαγώνιο που έχει διεύθυνση ίδια με το μοναδιαίο διάνυσμα $\left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)$, όταν οι κατανομές και των n δειγμάτων είναι ίδιες. Η μέθοδος της κανονικοποίησης ποσοστημορίων βασίζεται στην ιδέα ότι ένα σετ n δειγμάτων μπορεί να μετατραπεί ώστε όλα να έχουν την ίδια κατανομή, αν προβληθούν τα σημεία των n διανυσμάτων στη διαγώνιο.
Έστω $\mathbf{q}_k = (q_{k1}, \ldots, q_{kn})$ όπου το $k = 1, \ldots, p$ δηλώνει τον k από τους p συνολικούς ανιχνευτές, συμβολίζει το διάνυσμα των k-οστών ποσοστημορίων για όλα τα n δείγματα και $\mathbf{d} = \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right)$ η μοναδιαία διαγώνιος. Για να μετασχηματιστούν τα ποσοστημόρια έτσι ώστε να βρίσκοντα όλα επί της διαγωνίου, πρέπει να προβληθεί το διάνυσμα $\mathbf{q}$ πάνω στο $\mathbf{d}$:

$$\mathrm{proj}_{\mathbf{d}}\, \mathbf{q}_k = \left(\frac{1}{n}\sum_{j=1}^{n} q_{kj}, \ldots, \frac{1}{n}\sum_{j=1}^{n} q_{kj}\right) \tag{2.3}$$

Με αυτόν το μετασχηματισμό οι συγκεντρώσεις των δειγμάτων γίνονται ίδιες, αντικαθιστώντας την αρχική μέτρηση με τη μέση τιμή του αντίστοιχου ποσοστημορίου. Ο αλγόριθμος είναι απλός και κατασκευάζεται ως εξής:

1. Έστω X ο πίνακας $p \times n$ που προέκυψε από τη διόρθωση υποβάθρου, με n δείγματα μεγέθους το καθένα p, δηλαδή πρόκειται για πείραμα με n δείγματα (κυτταροσειρές ή ιστούς) και σε κάθε μικροσυστοιχία μετριούνται p ανιχνευτές.
2. Αποθηκεύεται η σειρά (rank) των δεδομένων κάθε στήλης του πίνακα X.
3. **Κάθε** στήλη του πίνακα στοιχίζεται από τη μικρότερη προς τη μεγαλύτερη μέτρηση και έτσι προκύπτει ο πίνακας X_{sorted}. Να σημειωθεί ότι ενώ κάθε γραμμή του αρχικού πίνακα X αντιστοιχούσε σε διαφορετικό ανιχνευτή, οι γραμμές του X_{sorted} δεν έχουν κάποια βιολογική σημασία.
4. Σε κάθε γραμμή του X_{sorted} υπολογίζεται ο αριθμητικός μέσος και στη συνέχεια κάθε στοιχείο της γραμμής αντικαθίσταται από αυτόν. Πρακτικά οι κατανομές των δειγμάτων χωρίστηκαν σε p ποσοστημόρια το καθένα από τα οποία περιλαμβάνει μία μέτρηση, και κάθε ποσοστημόριο αντικαθίσταται από το μέσο όσο των ποσοστημορίων. Δημιουργείται έτσι ο X^{new}_{sorted}.
5. Οι στήλες του X^{new}_{sorted} επαναδιατάσσονται επιβάλλοντας σε κάθε μία την αρχική σειρά των δεδομένων της όπως αυτή αποθηκεύτηκε στο βήμα 2. Ο τελικός κανονικοποιημένος πίνακας είναι ο $X_{normalized}$.

Κάποια μειονεκτήματα της μεθόδου γίνονται ευθέως προφανή, καθώς η μέθοδος επιβάλλει την ίδια τιμή σε όλα τα ποσοστημόρια δια μέσου των δειγμάτων, χάνοντας έτσι πιθανές βιολογικές διαφορές σε κάποιες περιπτώσεις. Επίσης, η μέθοδος είναι ευαίσθητη σε λάθος μετρήσεις. Ωστόσο, τα προβλήματα μειώνονται πρακτικά καθώς μία μικροσυστοιχία είναι πολύ απίθανο να βρίσκεται εξόλοκλήρου σε διαφορετική κλίμακα απο τις υπόλοιπες, και επίσης χρησιμοποιούνται πολλαπλοί ανιχνευτές για κάθε γονίδιο, εξομαλύνοντας πιθανά σφάλματα που μπορεί να επηρρεάζουν κάποιους από αυτούς. Η μέθοδος χρησιμοποιείται από τις αρχές του $21^{ου}$ αιώνα με πολύ ικανοποιητικά αποτελέσματα που αποδεικνύουν, έστω εμπειρικά, πως οι παραπάνω παρατηρήσεις σπανίως παίζουν ρόλο.

2.4 Ανάλυση σε κύριες συνιστώσες

Ένα μεγάλο πρόβλημα των γονιδιακών δεδομένων όταν πρέπει να γίνει ανάλυση αυτών σε πρώτο επίπεδο (Exploratory Data Analysis [29]), δηλαδή μια αναγνωριστική *ματιά* είναι ο αριθμός των διαστάσεων. Κάθε δείγμα περιγράφεται πλήρως από περίπου 10000 γονίδια (διαστάσεις), τα οποία χωρίς κάποια βιολογική γνώση που να περιγράφει τις μεταξύ τους σχέσεις, είναι ανεξάρτητα το ένα από το άλλο. Είναι επομένως αδύνατο να γίνει κατανοητή η μορφή των δεδομένων σε 10000 διαστάσεις, και επιτακτικό να μειωθούν με κάποιον τρόπο, ώστε να μπορούν να παρατηρηθούν τα δείγματα σε κάποιον *χώρο-δειγμάτων* και να βγει κάποιο συμπέρασμα για τα δεδομένα.
Ο κλάδος της στατιστικής που ασχολείται με τη μείωση διαστάσεων (dimensionality reduction), παρέχει τις απαραίτητες μεθόδους, η πιο γνωστή από τις οποίες είναι η ανάλυση σε κύριες συνιστώσες (Principal Component Analysis - **PCA**). Μέσω της PCA, οι αρχικές διαστάσεις αντικαθιστούνται από καινούργιες **συνιστώσες** οι οποίες δε σχετίζονται η μία με την άλλη και οι οποίες εξηγούν η κάθε μια ένα ποσοστό της μεταβλητότητας που αναπτύσσεται μεταξύ των μεταβλητών. Μία πολύ σημαντική ερώτηση που τίθεται αναγκαστικά είναι: *Θα χρησιμοποιηθούν ως μεταβλητές για την ανάλυση τα n δείγματα, ή οι p μετρήσεις κάθε δείγματος;* Η απάντηση εξαρτάται από το τι θέλει ο αναλυτής να μελετήσει. Η ανάλυση κυρίων συνιστωσών στα γονίδια, σημαίνει ότι θα βγει συμπέρασμα για το ποια γονίδια συντελλούν περισσότερο στη μεταβλητότητα μεταξύ των δειγμάτων. Αντίστοιχα, ανάλυση στα δείγματα θα έχει ως αποτέλεσμα την οπτικοποίηση της *απόστασης* μεταξύ

των δειγμάτων. Προφανώς, σε πρώτο επίπεδο, ο ερευνητής ενδιαφέρεται για τη δεύτερη περίπτωση, δηλαδή για το πώς τα δείγματα κατανέμονται στον χώρο των κυρίων συνιστωσών. Έτσι μπορεί να βγάλει συμπέρασμα για το ποιες *ομάδες* δειγμάτων δημιουργούνται.
Σημείωση:Στις προηγούμενες ενότητες, τα χαρακτηριστικά των δειγμάτων αποτελούσαν οι p ανιχνευτές. Πλέον, μετά την προεπεξεργασία, έχει γίνει αντιστοίχιση των ανιχνευτών στα γονίδιά τους, με την απαραίτητη αφαίρεση ανιχνευτών που ανήκουν σε περισσότερα από ένα γονίδια, και με τον υπολογισμό του μέσου όρου για ένα γονίδιο που ανήκει σε περισσότερους από έναν ανιχνευτές. Για λόγους ευκρίνειας, πλέον δε θα χαρακτηρίζουμε p τα χαρακτηριστικά, αλλά g (gene), εννοώντας τα γονίδια.
Η βασική ιδέα πίσω από την ανάλυση σε κύριες συνιστώσες είναι απλή: Οι μεταβλητές μπορούν να χωριστούν και να συνδυαστούν με κατάλληλο τρόπο, δημιουργώντας έτσι καινούργιες *μεταβλητές* οι οποίες δε σχετίζονται μεταξύ τους. Στην περίπτωση των γονιδιακών δεδομένων συνήθως ο αριθμός των δειγμάτων είναι πολύ μεγαλύτερος από τον αριθμό των γονιδίων, οπότε υπολογιστικά συμφέρει να γίνει ανάλυση ως προς τα δείγματα. Εξάλλου, όπως αναφέρθηκε στην προηγούμενη παράγραφο, μας ενδιαφέρει ο τρόπος με τον οποίο τα δείγματα κατανέμονται στον χώρο των κύριων συνιστωσών. Η ανάλυση σε πρώτους παράγοντες γίνεται γενικά ως εξής [30]:
Έστω $\mathbf{X}_{n\times p}$ πίνακας με n παρατηρήσεις κάθε μία από τις οποίες περιγράφεται από p μεταβλητές, με στοιχεία

$$x_{ij} \qquad i = 1,\ldots,n, \qquad ,j = 1,\ldots,p$$

. Οι στήλες του πίνακα (κατά μεταβλητή) κεντράρονται, δηλαδή υπολογίζεται ο μέσος όρος κάθε στήλης και αφαιρείται από την αντίστοιχη στήλη, δηλαδή:

$$x_{ij} = x_{ij} - \frac{1}{n}\sum_{i=1}^{n} x_{ij} \tag{2.4}$$

Με αυτόν το μετασχηματισμό, μπορεί να υπολογιστεί εύκολα ο πίνακας συνδιακύμανσης C. Από τη στατιστική γνωρίζουμε ότι η συνδιακύμανση δύο διανυσμάτων $\mathbf{X},\mathbf{Y}$ διάστασης n είναι ίση με:

$$\mathrm{cov}(\mathbf{X},\mathbf{Y}) = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{n-1} \tag{2.5}$$

Η οποία για δεδομενα που έχουν κεντραριστεί με τον μετασχηματισμό (2.4) γίνεται:

$$\mathrm{cov}(\mathbf{X},\mathbf{Y}) = \frac{\sum_{i=1}^{n} X_i Y_i}{n-1} \tag{2.6}$$

Επομένως, για ένα μητρώο $\mathbf{X}_{n\times p}$ που ουσιαστικά αποτελείται από p διανύσματα μήκους n το καθένα, οι συνδιακυμάνσεις όλων των στηλών με όλες τις στήλες, δίνονται από τον παρακάτω τύπο:

$$\mathbf{C}_{p\times p} = \frac{\mathbf{X}^T\mathbf{X}}{n-1} \tag{2.7}$$

Επομένως η συνδιακύμανση είναι πίνακας συμμετρικός, και επομένως αφού υπολογιστούν οι ιδιοτιμές και τα (μοναδιαία) ιδιοδιανύσματά του γράφεται:

$$\mathbf{C} = \mathbf{V}\mathbf{L}\mathbf{V}^T \tag{2.8}$$

Όπου $\mathbf{V}$ ο πίνακας των ιδιοδιανυσμάτων, με κάθε στήλη να είναι ένα ιδιοδιάνυσμα, και $\mathbf{L}$ ο διαγώνιος πίνακας που περιέχει στη διαγώνιο τις αντίστοιχες ιδιοτιμές, από τη μεγαλύτερη προς τη μικρότερη. Στην ουσία, κάθε ιδιοδιάνυσμα περιγράφει τη διεύθυνση μιας κύριας συνιστώσας, και η προβολή των δεδομένων πάνω σε αυτή αποτελεί την αντίστοιχη κύρια

συνιστώσα. Επομένως, στον $n \times p$ πίνακα $\mathbf{XV}$, η j κύρια συνιστώσα δίνεται από την j στήλη, και οι συντεταγμένες του i δείγματος στον χώρο των κύριων συνιστωσών δίνεται από την i γραμμή. Επομένως για να γίνει ανάλυση σε κύριες συνιστώσες, πρέπει να βρεθούν οι ιδιοτιμές και τα ιδιοδιανύσματα του μητρώου συνδιακύμανσης. Κάτι τέτοιο υπολογιστικά δε συμφέρει, για αυτό οι κύριες συνιστώσες υπολογίζονται μέσω της πιο σταθερής μεθόδου αποσύνθεσης μοναδικής τιμής, ή Singular Value Decomposition (SVD) [31] [32]. Μέσω της SVD το μητρώο $\mathbf{X}$ μετασχηματίζεται στο γινόμενο τριών επιμέρους μητρώων:

$$\mathbf{X}_{n\times p} = \mathbf{U}_{n\times n}\mathbf{S}_{n\times n}\mathbf{V}^T_{n\times p} \tag{2.9}$$

- $\mathbf{U}$ ένας ορθογώνιος πίνακας $n \times n$
- $\mathbf{S}$ ένας διαγώνιος πίνακας με τιμές $s_i \qquad i = 1, \ldots, n$

Είναι τώρα προφανές, αντικαθιστώντας τον τύπο (2.9), στον τύπο (2.6) ότι:

$$\mathbf{C} = \mathbf{V}\frac{\mathbf{S}^2}{n-1}\mathbf{V}^T \tag{2.10}$$

Συγκρίνοντας τον τύπο (2.10) με τον (2.8) γίνεται κατανοητό πως το μητρώο $\mathbf{V}$ περιλαμβάνει τις διευθύνσεις των κύριων συνιστωσών, ενώ ισχύει παράλληλα ότι:

$$\lambda_i = \frac{s_i^2}{n-1} \tag{2.11}$$

και οι κύριες συνιστώσες δίνονται όπως και προηγουμένως με την προβολή:

$$\mathbf{XV} = \mathbf{USV}^T\mathbf{V} = \mathbf{US} \tag{2.12}$$

Επομένως, αρκεί να βρεθούν τα μητρώα $\mathbf{U}, \mathbf{S}$ για να βρεθούν οι κύριες συνιστώσες. Στην περίπτωση γονιδιακών δεδομένων, το μητρώο που μελετάται είναι το $\mathbf{X}_{g\times n}$, με n τα δείγματα στις στήλες και g τα γονίδια στις γραμμές. Αυτό που θέλουμε να πετύχουμε με την ανάλυση σε κύριες συνιστώσες είναι να οπτικοποιηθεί η απόσταση μεταξύ των δειγμάτων και όχι μεταξύ των γονιδίων, καθώς η γονιδιακή ανάλυση θα γίνει πολύ πιο λεπτομερειακά σε επόμενο στάδιο. Άρα η PCA θα γίνει πάνω στο ανάστροφο μητρώο $\mathbf{X}^T_{n\times g}$. Για την οπτικοποίηση τα n δείγματα απεικονίζονται στις δύο πρώτες κύριες συνιστώσες, δεδομένου ότι αυτές εξηγούν ένα σχετικά μεγάλο ποσοστό της μεταβλητότητας. Για να υπολογιστεί η μεταβλητότητα που εξηγεί κάθε κύρια συνιστώσα, από τον πίνακα συνδιακύμανσης του μεταχηματισμένου πίνακα $\mathbf{C}(\mathbf{US})$ υπολογίζεται η συνολική μεταβλητότητα που είναι το άθροισμα των στοιχείων της διαγωνίου $\sum_{i=1}^{n}\mathbf{C}_{ii}$. Το ποσοστό μεταβλητότητας [33] που εξηγεί η συνιστώσα i, είναι το κλάσμα

$$\frac{\mathbf{C}(\mathbf{US})_{ii}}{\sum_{i=1}^{n}\mathbf{C}_{ii}} \tag{2.13}$$

2.5 ComBat και εμπειρικές Μπεϋζιανές μέθοδοι

Στη συγκεκριμένη εργασία έγινε ενσωμάτωση δεδομένων από τέσσερα διαφορετικά πειράματα. Ο λόγος για τον οποίο δεν αντιμετωπίστηκε κάθε πείραμα ξεχωριστά είναι η φύση του μελανώματος, που χαρακτηρίζεται από μεγάλη ποικιλομορφία. Επομένως, με την ενσωμάτωση των δεδομένων από διαφορετικά εργαστήρια και πλατφόρμες, τα αποτελέσματα γίνονται πιο αξιόπιστα αφού βασίζονται σε δεδομένα που καλύπτουν πολλές βιολογικές περιπτώσεις. Επίσης, ξεπερνιέται η δυσκολία που θέτει ο μικρός αριθμός υγιών κυττάρων,

τα οποία υπάρχουν σε δύο από τα τέσσερα πειράματα, κι έτσι καρκινικά κύτταρα από άλλες μελέτες είναι δυνατό να συγκριθούν με τα υγιή. Μητρώα μετρήσεων από διαφορετικά πειράματα δεν είναι σωστό να συνδυαστούν απευθείας σε ένα εννιαίο μητρώο χωρίς καμια επεξεργασία. Κάτι τέτοιο συμβαίνει επειδή στις μετρήσεις από μία συγκεκριμένη πλατφόρμα εισάγεται αυτόματα το συστηματικό σφάλμα (bias) της πλατφόρμας [34]. Για το συνδυασμό τέτοιων μητρώων έχουν αναπτυχθεί διάφορες μέθοδοι, οι περισσότερες από τις οποίες δίνουν ικανοποιητικά αποτελέσματα όταν κάθε πείραμα έχει μεγάλο αριθμό δειγμάτων. Στην ανάλυση που έγινε, οι σειρές του GEO που χρησιμοποιήθηκαν περιέχουν δείγματα που κυμαίνονται σε αριθμό από 6 έως 144. Επομένως πρέπει να εφαρμοστεί μία μέθοδος που να ενσωματώνει δεδομένα γονιδιακής έκφρασης ακόμα και με πειράματα μικρής έκτασης. Η πιο αξιόπιστη μέθοδος ενσωμάτωσης σε αυτή την περίπτωση είναι μέσω χρήσης **Εμπειρικών Μπεϋζιανών μοντέλων** (Empirical Bayes models **EB**) [35]. Το πακέτο της R που χρησιμοποιείται για την εφαρμογή των μεθόδων αυτών ονομάζεται **sva** μέσω της συνάρτησης **ComBat**.
Τα βασικά βήματα της μεθόδου διόρθωσης των δεδομένων με εμπειρικά Μπεϋζιανά μοντέλα ακολουθούν παρακάτω.

1. **Μοντέλο**: Αρχικά, γίνεται η υπόθεση πως όλα τα δεδομένα έχουν την παρακάτω μορφή:
$$Y_{ijg} = a_g + X\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg}, \tag{2.14}$$
όπου:
   - $i = 1, \ldots, m$ ο αριθμός των διαφορετικών ομάδων που πρέπει να συνδυαστούν
   - $j = 1, \ldots, n_i$ το δείγμα που ανήκει στην ομάδα i
   - $g = 1, \ldots, G$ το γονίδιο
   - a_g η γενική έκφραση του γονιδίου g σε όλα τα δείγματα και τις ομάδες
   - X ο πίνακας σχεδίασης (design matrix) που δίνει τις διάφορες καταστάσεις των δειγμάτων. Συνήθως αυτές περιγράφονται δυαδικά, αφού στην πλειοψηφία των πειραμάτων συγκρίνονται κύτταρα περίπτωσης (case), όπως άρρωστα κύτταρα ή κύτταρα με κάποια ενεργή ουσία ή φάρμακο, έναντι υγιών κυττάρων (control). Ωστόσο, μπορεί να έχουμε και παραπάνω από δύο καταστάσεις σε μεγάλα πειράματα.
   - β_g το διάνυσμα που σε συνδυασμό με τον X δίνει τη διορθωμένη έκφραση ενός γονιδίου όταν αυτό μετριέται σε διαφορετική κατάσταση από την control, και προκύπτει από ένα γραμμικό μοντέλο, όπως θα περιγραφεί και στην ενότητα 2.6
   - γ_{ig} το αθροιστικό αποτέλεσμα που έχει η ομάδα i στο γονίδιο g για όλα τα δείγματα της ομάδας αυτής
   - δ_{ig} το πολλαπλασιαστικό αποτέλεσμα που έχει η ομάδα i στο γονίδιο g για όλα τα δείματα της ομάδας αυτής
   - ε_{ijg} ο όρος σφάλματος, που θεωρείται ότι ακολουθεί κανονική κατανομή με μέσο όρο 0, δηλαδή $\varepsilon_{ijg} \sim N\left(0, \sigma_g^2\right)$
2. **Κανονικοποίηση των μετρήσεων**: Ένα κύριο θέμα που πρέπει να λυθεί πριν την ενσωμάτωση των ομάδων είναι το γεγονός ότι οι μετρήσεις διαφέρουν συνολικά σε μέγεθος από πλατφόρμα σε πλατφόρμα λόγων των διαφορετικών ευαισθησιών που έχουν οι ανιχνευτές και των διαφορετικών εντάσεων από κάθε γονίδιο. Επομένως, στη σχέση (2.14) τα μεγέθη $a_g, \beta_g, \gamma_g, \sigma_g^2$ διαφέρουν από γονίδιο σε γονίδιο. Αν αυτές

οι διαφορές δε ληφθούν υπ'όψιν, το αποτέλεσμα θα είναι οι κατανομές που θα υπολογιστούν από το μπεϋζιανό μοντέλο να μειώσουν τις χρήσιμες βιολογικές διαφορές των γονιδίων, θεωρώντας αυτές ως διαφορετικές ομάδες που πρέπει να ενσωματωθούν σε μία. Για να αποφευχθεί αυτό, γίνεται κανονικοποίηση των δεδομένων από άποψη γονιδίων έτσι ώστε αυτά να έχουμ περίπου ίδιο μέσο όρο και διακύμανση. Εξάλλου αυτό που ενδιαφέρει τη μετέπειτα μελέτη δεν είναι πως συμπεριφέρεται ένα γονίδιο στο ίδιο δείγμα σε σχέση με κάποιο άλλο, αλλά πως ένα γονίδιο συμπεριφέρεται ανάμεσα σε πολλά διαφορετικά δείγματα. Για να γίνει αυτό, εκτιμούνται μέ ελάχιστα τετράγωνα δια μήκους όλων των δειγμάτων οι παράμετροι του μοντέλου:

$$\begin{aligned} &\widehat{a}_g, \widehat{\beta}_g, \widehat{\gamma}_{ig} \\ &i = 1, \ldots, m, \qquad g = 1, \ldots, G \\ &\sum_i n_i \widehat{\gamma}_{ig} = 0 \qquad g = 1, \ldots, G \end{aligned}$$

Ο περιορισμός για τις παραμέτρους $\widehat{\gamma}$ ισχύει για να διαχωριστεί το σφάλμα που εισάγουν οι ομάδες από τη γενική έκφραση και τη διόρθωση μέσω του πίνακα σχεδίασης. Με βάση αυτές τις εκτιμήσεις, υπολογίζεται η διακύμανση του σφάλματος για κάθε γονίδιο

$$\widehat{\sigma}_g^2 = \frac{1}{N} \sum_{ij} \left(Y_{ijg} - \widehat{a}_g - X\widehat{\beta}_g - \widehat{\gamma}_{ig} \right)^2 \qquad N \text{ όλα τα δείγματα}$$

Και τώρα τα κανονικοποιημένα δεδομένα παίρνουν τη μορφή:

$$Z_{ijg} = \frac{Y_{ijg} - \widehat{a}_g - X\widehat{\beta}_g}{\widehat{\sigma}_g} \tag{2.15}$$

3. **Εκτιμήσεις δεδομένων με παραμετρικά εμπειρικά priors**: Έχοντας κανονικοποιήσει πλέον τα δεδομένα, θεωρείται ότι αυτά ικανοποιούν την κατανομή $Z_{ijg} \sim N\left(\gamma_{ig}, \delta_{ig}^2\right)$, και αυτές οι παράμετροι (που είναι διαφορετικές από τα προηγούμενα γ, δ έχουν a priori τις παρακάτω κατανομές:

$$\gamma_{ig} \sim N(\gamma_i, \tau_i^2) \qquad \delta_{ig}^2 \sim InverseGamma(\lambda_i, \theta_i) \tag{2.16}$$

Οι παράμετροι των κατανομών (2.16) εκτιμούνται εμπειρικά από κανονικοποιημένα δεδομένα με τη μέθοδο των ροπών. Σε περίπτωση που οι κατανομές δεν ταιριάζουν αρκετά με τα δεδομένα, ακολουθείται μη-παραμετρικό μοντέλο εκτίμησεις. Με βάση τα μοντέλα αυτά, οι εκτιμήσεις για τις παραμέτρους διόρθωσης των ομάδων είναι:

$$\gamma_{ig}^* = \frac{n_i \bar{\tau}_i^2 \widehat{\gamma}_{ig} + \delta_{ig}^{2*} \overline{\gamma}_i}{n_i \bar{\tau}_i^2 + \delta_{ig}^{2*}} \qquad \delta_{ig}^{2*} = \frac{\overline{\theta}_i + \frac{1}{2} \sum_j \left(Z_{ijg} - \gamma_{ig}^* \right)^2}{\frac{n_i}{2} + \overline{\lambda}_i - 1} \tag{2.17}$$

4. **Διόρθωση δεδομένων για συνδυασμό**: Αφού έχουν πλέον υπολογιστεί μέσω των ΕΒ οι παράμετροι διόρθωσης, έχει τελειώσει η διαδικασία και διορθώνονται οι μετρήσεις σύμφωνα με τη σχέση (2.17). Ουσιαστικά, τα $\gamma_{ig}^*, \delta_{ig}^{2*}$ αποτελούν τις εκτιμήσεις που έγιναν μέσω του Μπεϋζιανού μοντέλου για τις παραμέτρους $\gamma_{ig}, \delta ig$ του μοντέλου (2.14). Παράλληλα με τη διόρθωση, πρέπει να *επιστραφεί* σε κάθε γονίδιο η μη-κανονικοποιημένη έκφρασή του, καθώς αυτές οι διαφορές είναι επιθυμητές, βιολογικής φύσης, και το μοντέλο θα ήταν αποτυχημένο αν εκτός από σφάλματα μεταξύ των πειραμάτων διόρθωνε και βιολογικές διαφορές μεταξύ γονιδίων. Προφανώς, από αυτούς τους κανονικοποιημένους μέσους όρους πρέπει να ληφθεί υπ'όψιν

η διακύμανση που οφείλεται στην ομάδα που ανήκει το δείγμα. Έτσι η τελική έκφραση διορθωμένη με τα EB, ώστε να μπορούν τα δεδομένα να συνδυαστούν σε ένα εννιαίο πείραμα είναι:

$$\gamma^*_{ijg} = \frac{\widehat{\sigma}_g}{\widehat{\delta}^*_{ig}} \left(Z_{ijg} - \widehat{\gamma}^*_{ig}\right) + \widehat{a}_g + X\widehat{\beta}_g \tag{2.18}$$
$$= \frac{\widehat{\sigma}_g}{\widehat{\delta}^*_{ig}} \left(\frac{Y_{ijg} - \widehat{a}_g - X\widehat{\beta}_g}{\widehat{\sigma}_g} - \widehat{\gamma}^*_{ig} \right) + \widehat{a}_g + X\widehat{\beta}_g$$

Με την ομαδοποίηση των δεδομένων τελειώνει το πρώτο μεγάλο μέρος της διπλωματικής εργασίας που σχετίζεται με την *τακτοποίηση των δεδομένων* (tidying of data), και την προετοιμασία τους για επεξεργασία και ανάλυση. Κάθε μελέτη υπολογιστικής βιολογίας περιλαμβάνει το κομμάτι της τακτοποίησης δεδομένων, που επηρρεάζει άμεσα την κατάντι ανάλυση. Για τον σκοπό της παρούσης μελέτης, επιλέχθηκαν εργαλεία ήδη αποδεδειγμένα αξιόπιστα, ευρέως χρησιμοποιούμενα, τα οποία παρ'όλα αυτά πρέπει να ελεγχθούν εκ των υστέρων για το βαθμό επιτυχίας τους. Οι έλεγχοι που έγιναν ακολουθούν στο επόμενο κεφάλαιο των αποτελεσμάτων αφού αφορούν καθαρά τα δεδομένα και δε σχετίζονται με κάποια συγκεκριμένη μέθοδο. Για περισσότερες πληροφορίες, ο αναγνώστης καλείται να ανατρέξει στο [35].

2.6 Γονιδιακή Ανάλυση

Ο σκοπός μιας υπολογιστικής μελέτης γονιδιακής έκφρασης, είναι η ανάλυση των δεδομένων. Αυτό σημαίνει ότι με βάση στοιχεία που υπάρχουν για τα διάφορα δείγματα και πειράματα, ο αναλυτής πρέπει να εξάγει συμπεράσματα για τα διάφορα γονίδια. Συγκεκριμένα, αποκαλύπτονται τα γονίδια που διαφέρουν ανάμεσα σε διαφορετικά δείγματα (case vs. control), και επομένως είτε ευθύνονται για τις διαφορές, είτε έχουν επηρρεαστεί από αυτές, ενώ στις περισσότερες περιπτώσεις ισχύουν και τα δύο. Τα γονίδια αυτά ονομάζονται **διαφορικά εκφρασμένα γονίδια**, και αποτελούν τη βάση της γονιδιακής ανάλυσης. Προφανώς, επειδή έχουμε συνεχή δεδομένα που προέρχονται από μια μικροσυστοιχία, δεν είναι τόσο απλό να χαρακτηριστεί ένα γονίδιο ως διαφορικά εκφρασμένο, για αυτό όλα τα αποτελέσματα είναι στατιστικά και συνοδεύονται από τη στατιστική ισχύ του πειράματος, η οποία αυξάνεται με τον αριθμό των δειγμάτων. Δηλαδή, στην ουσία μελετούνται δύο πληθυσμοί δειγμάτων, οι οποίοι έχουν ως μεταβλητές τα γονίδια. Κάθε γονίδιο δηλαδή χαρακτηρίζεται από δύο κατανομές, την κατανομή του στον πληθυσμό *control* και την κατανομή του στον πληθυσμό *case*. Ως γνωστόν από τη Στατιστική, προκειμένου να εξαχθεί συμπέρασμα για το αν δύο κατανομές είναι στατιστικά σημαντικές ή όχι, χρησιμοποιείται το t-test και τα p-values που προκύπτουν από αυτό. Συνήθως για τη γονιδιακή ανάλυση, ένα γονίδιο διαφέρει στα case δείγματα από τα control, όταν η τιμή p-value που του αντιστοιχεί είναι μικρότερη από 0.05. Ωστόσο, αυτό το κριτήριο δεν αρκεί για να χαρακτηριστεί κάποιο γονίδιο ως διαφορικά εκφρασμένο, αφού πρέπει και η μέση έκφρασή του να είναι εμφανώς διαφορετική ανάμεσα στους δύο πληθυσμούς. Έτσι εισάγεται ένα μέγεθος πολύ σημαντικό στην επιστήμη της μοριακής βιολογίας, ο λογάριθμος του λόγου των τιμών έκφρασης case/control του γονιδίου, **Fold Change**, (FC):

$$\log_2 FC_g = \log_2 \frac{E(g)_{case}}{E(g)_{control}}$$

όπου: $E(g)_{case}$ η μέση έκφραση του γονιδίου στα *case* δείγματα, και $E(g)_{control}$ η μέση έκφραση του γονιδίου στα *control* δείγματα.

Με το μέγεθος $\log_2 FC$ είναι εύκολη η άμεση σύγκριση της έκφρασης του γονιδίου μεταξύ των δύο πληθυσμών: Αρνητικές τιμές αντιστοιχούν σε *υποεκφρασμένο* γονίδιο, ενώ θετικές τιμές σε *υπερεκφρασμένο* (down-, up-regulated gene). Επίσης, η τιμή 1 αντιστοιχεί σε διπλάσια έκφραση, ενώ -1 στη μισή. Τα γονίδια που έχουν $p - value < 0.05$ πρέπει να έχουν και $|\log_2 FC| > threshold$, όπου *threshold* μία τιμή που επιλέγει ο αναλυτής. Συνήθεις τιμές είναι $threshold = 1, 1.2, 1.5$, και εξαρτώνται από την προεπεξεργασία και την κανονικοποίηση που έχει προηγηθεί στα δεδομένα.
Η οπτικοποίηση των δύο αυτών μεγεθών που αρκούν για να περιγράψουν ένα γονίδιο ως διαφορικά εκφρασμένο γίνεται μέσω του διαγράμματος *κρατήρα ηφαιστείου* [36] (Volcano Plot). Στο διάγραμμα, ο οριζόντιος άξονας x αποτελεί τον άξονα του $\log_2 FC$ ενώ ο κατακόρυφος άξονας y το $-\log_1 0p - value$, που σημαίνει ότι όσο πιο *ψηλά* βρίσκεται το γονίδιο στον άξονα y τόσο μικρότερο είναι το p-value του και άρα τόσο πιο πιθανό η διαφορά ανάμεσα στους δύο πληθυσμούς να είναι στατιστικά σημαντική.

Το πακέτο limma της R περιέχει τις απαραίτητες συναρτήσεις για υπολογισμό των παραπάνω μεγεθών, προκειμένου να διευκολυνθεί ο αναλυτής [37]. Η γονιδιακή ανάλυση γίνεται μέσω γραμμικού μοντέλου που έχει τη μορφή:

$$y_g = \beta_{0g}x_1 + \ldots + \beta_{pg}x_p + \varepsilon_g \qquad g = 1, \ldots, g \tag{2.19}$$

Όπου:

- y_g η έκφραση του γονιδίου g
- β_{jg} η έκφραση του γονιδίου η οποία αφορά την κατάσταση j. Στην περίπτωση που μελετάται, δηλαδή σε δεδομένα τα οποία έχουν δύο καταστάσεις, $p = 1$, και άρα $j = 0$ δηλώνει τα control δείγματα και $p = 1$ δηλώνει τα case.
- $x_j \qquad i = 1, \ldots, n \qquad j = 1, \ldots, p$ ο δυαδικός συντελεστής (1 για case, 0 για control) που δηλώνει την κατάσταση του δείγματος i.
- ε_g το σφάλμα υπολογισμού της γονιδιακής έκφρασης

Με την ελαχιστοποίηση του σφάλματος ε_g μέσω ελαχίστων τετραγώνων, εκτιμούνται για κάθε γονίδιο οι παράμετροι $\widehat{\beta}_{jg}$ οι οποίοι από τη μέθοδο των ελαχίστων τετραγώνων υπολογίζονται σε μορφή διανύσματος από τη σχέση:

$$\widehat{\beta}_g = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{2.20}$$

Όπου:

- $\widehat{\beta}$ το διάνυσμα-στήλη με τις μεταβλητές $\widehat{\beta}_j$:

$$\widehat{\beta} = \begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \vdots \\ \widehat{\beta}_p \end{bmatrix}$$

- $\mathbf{X}$ ο **πίνακας σχεδίασης** που περιλαμβάνει τις τιμές (0 ή 1) που περιγράφουν την κατάσταση του δείγματος:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}$$

Δηλαδή στην περίπτωση που έχουμε $p = 2$ με τις δύο καταστάσεις control, case, η πρώτη στήλη θα περιγράφει τα control και θα έχει 1 στα control δείγματα και 0 στα case δείγματα, ενώ το αντίθετο θα συμβαίνει στη δεύτερη στήλη.

- $\mathbf{y}$ το διάνυσμα των γονιδιακών εκφράσεων για τα n δείγματα:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Και τώρα οι εκτιμώμενες τιμές γονιδιακής έκφρασης από το μοντέλο είναι:

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\beta} \tag{2.21}$$

Το τυπικό σφάλμα των εκτιμήσεων των παραμέτρων β υπολογίζεται από τη σχέση:

$$se(\widehat{\beta}_g) = \sqrt{s_g^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}} \tag{2.22}$$

Όπου s_g^2 είναι η διακύμανστη του γονιδίου g και το t-test για να υπολογιστεί το p-value του γονιδίου είναι: $t_g = \frac{\widehat{\beta}_{1g} - \widehat{\beta}_{0g}}{se(\widehat{\beta}_1 g)}$ Στο σημείο αυτό τελειώνει η ανάλυση με το γραμμικό μοντέλο, αφού έχει τελειώσει η εκτίμηση των συντελεστών με τα ελάχιστα τετράγωνα. Το πακέτο έχει ενσωματωμένες μεθόδους που χρησιμοποιούν Μπεϋζιανά μοντέλα όπως στην ενότητα 2.5 για να διορθωθούν κατά τα γονίδια οι συντελεστές $\widehat{\beta}$ και s^2 από την κατανομή της γονιδιακής έκφρασης κατά γονίδια. Το τελικό αποτέλεσμα είναι ένας πίνακας που περιέχει όλα τα αποτελέσματα της διαφορικής γονιδιακής ανάλυσης, με πιο σημαντικά το $\log(FC)$ και τη διορθωμένη τιμή $p - value$ κατά τη μέθοδο Benjamini-Hochberg.
Με το τέλος της διαφορικής ανάλυσης, έχουν εξαχθεί από τα δεδομένα των πειραμάτων τα αποτελέσματα με βάσει τα οποία μπορούν να χαρακτηριστούν τα γονίδια ως διαφορικά εκφρασμένα (υπερ- ή υπό- εκφρασμένα). Στο σημείο αυτό η κατάντι μελέτη χωρίζεται σε δύο κατευθύνσεις: α) Τα υπερεκφρασμένα και τα υποεκφρασμένα γονίδια αποθηκεύονται, κι έτσι δημιουργούνται λίστες γονιδίων που περιγράφουν την κατάσταση του δείγματος υπο μελέτη. Αυτές οι λίστες μπορούν να χρησιμοποιηθούν ως είσοδος στο εργαλείο cMap του Broad Institute ώστε να βρεθούν ουσίες οι οποίες έχουν ως αποτέλεσμα την ίδια γονιδιακή έκφραση. β) Τα διαφορικά εκφρασμένα γονίδια χρησιμοποιούνται ως features σε αλγορίθμους μηχανών εκμάθησης, ώστε να βρεθούν όσο το δυνατόν λιγότερα γονίδια γίνεται με βάση τα οποία ένας αλγόριθμος μηχανής μπορεί να ξεχωρίσει ανάμεσα σε κυτταρικές καταστάσεις, όπως ένα υγιές κύτταρο από ένα καρκινικό.

2.7 cMap

Η πρώτη κατεύθυνση αφού έχουν βρεθεί τα γονίδια που περιγράφουν την κατάσταση του κυττάρου, είναι αυτά να χρησιμοποιηθούν ως είσοδος στο λογισμικό cMap [38]. Το λογισμικό είναι διαθέσιμο μέσω του ιστότοπου clue.io το οποίο εκτός από το Query που χρησιμοποιείται στην εργασία, υπάρχει πλήθος εφαρμογών χρήσιμες στην κοινότητα της υπολογιστικής βιολογίας. Η βασική ιδέα πίσω από τη δημιουργία του cMap είναι ότι κάθε διαφορετικό κύτταρο σε κάθε διαφορετική κατάσταση έχει μια συγκεκριμένη έκφραση. Όταν δύο εκφράσεις είναι παρόμοιες, για παράδειγμα για δύο όμοια κύτταρα που έχουν εκτεθεί σε δύο διαφορετικές ουσίες, οι δύο ουσίες προκαλούν παρόμοιες αλλαγές στο κύτταρο και έτσι μπορεί η μία να αντικαταστήσει την άλλη, ή να συνδυαστεί μαζί της για να

έχουν βελτιωμένη δράση.
Για να υλοποιηθεί κάτι τέτοιο, από τη στιγμή που δεν υπάρχουν ντετερμινιστικές σχέσεις μεταξύ των γονιδίων, των καταστάσεων, και των ουσιών, είναι αναγκαία μία συστημική προσέγγιση. Το Broad Institute έχει αποθηκευμένα, και διαθέσιμα μέσω της πλατφόρμας clue.io, πάνω από 1.5 εκατομμύρια προφίλ γονιδιακής έκφρασης, με βάση κάποιες συγκεκριμένες κυτταροσειρές, όπως η μελανωματική σειρά A-375 που χρησιμοποιείται και στην παρούσα εργασία. Αυτά τα προφίλ προέρχονται από πειράματα στα οποία οι κυτταροσειρές δοκιμάζονται σε συνδυασμό με περίπου 5000 ουσίες μικρών μορίων (small-molecule compounds) και 3000 γενετικά αντιδρώντα (genetic reagents).
Όταν ο τελικός χρήστης, μέσω της διαδικτυακής εφαρμογής εισάγει ένα συγκεκριμένο γενετικό προφίλ, ουσιαστικά δηλώνοντας ποια γονίδια είναι υπερκεφρασμένα και ποια υποεκφρασμένα, το cMap συγκρίνει το προφίλ με τα προφίλ που έχει αποθηκευμένα στη βάση δεδομένων του, και επιστρέφει προφίλ τα οποία έχουν έναν συγκεκριμένο αριθμό συσχέτισης με το προφίλ του χρήστη. Συντελεστής συσχέτισης 100 σημαίνει ότι το προφίλ του χρήστη έχει ίδια έκφραση με το αντίστοιχο προφίλ του cMap, συντελεστής -100 δηλώνει αντίστροφη έκφραση, ενώ συντελεστής 0 δηλώνει οτι δεν υπάρχει συσχέτιση μεταξύ των καταχωρήσεων. Αυτά που ενδιαφέρουν τον χρήστη είναι τα προφίλ με συντελεστές συσχέτισης ± 100. Όταν μελετάται το προφίλ ενός καρκινικού κυττάρου, τα προφίλ που έχουν συντελεστή $-100 \div -95$ δηλώνουν αντιστροφή της έκφρασης. Έτσι αν το προφίλ προέρχεται από πείραμα μιας κυτταροσειράς με ένα συγκεκριμένο φάρμακο, το φάρμακο αυτό είναι δυνητικά ουσία που θα αντιστρέψει τη δράση του καρκινικού κυττάρου και ενδεχομένως θα καταστείλει τον καρκίνο. Αντίστροφα, σε προφίλ με συντελεστή $95 \div 100$, αν γνωρίζουμε τους μηχανισμούς με τους οποίους δρα η ουσία (Mechanism of Action), τότε ενδεχομένως και ο καρκίνος που αναπτύσσεται στο κύτταρο να δημιουργήθηκε από παρόμοιους μηχανισμούς, γεγονός που δίνει μεγαλύτερη επίγνωση στην ασθενεια του καρκίνου.
Όλα τα παραπάνω μπορούν προφανώς να εφαρμοστούν για οποιαδήποτε ασθένεια μελετάται, επιταχύνοντας σημαντικά σήμερα την επανατοποθέτηση φαρμάκων, και τη γνώση που υπάρχει για τις διάφορες ασθένειες και μεταλλάξεις.

2.8 Μηχανές Διανυσμάτων Υποστήριξης (SVM)

Όπως αναφέρθηκε στο τέλος της ενότητας 2.6, τα διαφορικά εκφρασμένα γονίδια χρησιμοποιούνται ως είσοδος σε έναν αλγόριθμο μηχανής εκμάθησης (machine learning), ο οποίος λειτουργεί με μηχανές διανυσμάτων υποστήριξης, Support Vector Machines, ή εν συντομία SVMs. Αποτελούν ένα εργαλείο το οποίο αν και αναπτύχθηκε το 1995 από τον *Vapnik* [39], συνεχώς αποδεικνύεται μία όλο και πιο αξιόπιστη μέθοδος κατηγοριοποίησης, με υπομεθόδους και βελτιστοποιήσεις να χρησιμοποιούνται με αυξανόμενο ρυθμό στην ανάλυση της γονιδιακής έκφρασης και στην επεξεργασία γονιδιακών δεδομένων [40]. Ο λόγος είναι ότι τα SVMs μπορούν να ρυθμιστούν κατάλληλα από τον χρήστη ώστε να καλύπτουν σχεδόν οποιαδήποτε μορφή δεδομένων, είτε αυτά παρουσιάζουν γραμμικό διαχωρισμό είτε όχι, όπως συμβαίνει συνήθως με γονιδιακά δεδομένα. Τα SVMs προγραμματίζονται κατάλληλα ώστε να *μαθαίνουν* να ξεχωρίζουν μία κλάση δεδομένων από μία άλλη, ενώ πρέπει να σημειωθεί ότι δουλεύουν για κατηγοριοποίηση ανάμεσα σε δύο κλάσεις μόνο, χαρακτηρίζοντάς τες δυαδικά ως 0 ή 1. Για την κατηγοριοποίηση ανάμεσα σε περισσότερες κλάσεις πρέπει να γίνει βηματικά η διαδικασία, διαχωρίζοντας μία κλάση από τις υπόλοιπες, στη συνέχεια από αυτές άλλη μία, κ.ο.κ., χρησιμοποιώντας κάθε φορά διαφορετικά SVMs. Στα πλαίσια της εργασίας, η κατηγοριοποίηση που πρέπει να γίνει είναι αρχικά ανάμεσα σε υγιή και καρκινικά κύτταρα. Είναι δηλαδή δυαδικής φύσης,

για αυτό χρησιμοποιήθηκε μοντέλο SVMs το οποίο εκπαιδεύεται σε ένα ποσοστό των δειγμάτων, και δοκιμάζεται η επιτυχία του στα υπόλοιπα δείγματα που απομένουν. Για την κατηγοριοποίηση ανάμεσα στις διάφορες μεταλλάξεις του μελανώματος, όπου έχουμε περισσότερες από δύο κλάσεις, επιλέχθηκε η μέθοδος Γραμμικής Διακριτικής Ανάλυσης όπως περιγράφεται στην επόμενη ενότητα 2.9.

Η βασική ιδέα των μηχανών εκμάθησης είναι να δίνει την εκτίμηση $\widehat{y}_i$ μιας τιμής y_i που μας ενδιαφέρει, μέσω ενός δοσμένου σημείου $\mathbf{x_i}$. Για να το κάνει αυτό, πρέπει πρώτα να εκπαιδευθεί σε ένα σύνολο $\mathit{\Lambda}$ σημείων $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{\mathit{\Lambda}} \in \Re^N$ τα οποία έχουν αντίστοιχα γνωστές τιμές y_i, οι οποίες στην περίπτωση των SVMs που διαχειρίζονται δυαδικές κλάσεις, παίρνουν την τιμή 1 ή -1, όπου 1 στην περίπτωση της εργασίας συμβολίζει το καρκινικό κύτταρο και -1 το υγιές. Αφού το μοντέλο εκπαιδευτεί σε αυτό το σύνολο $\mathit{\Lambda}$ σημείων που ονομάζεται το σύνολο εκμάθησης (training set), η απόδοση του μοντέλου ελέγχεται με ένα δεύτερο σύνολο, το σύνολο ελέγχου (testing set) από το οποίο χρησιμοποιούνται τα σημεία $\mathbf{x}$ για να εκτιμηθεί η τιμή $\widehat{y}$, η οποία συγκρίνεται με την πραγματική y. Ένα μοντέλο που έχει καλή απόδοση μπορεί στη συνέχεια να χρησιμοποιηθεί για την εκτίμηση της τιμής y, δεδομένων για τα οποία δεν έχουμε έτοιμη την πραγματική τιμή, παρα μόνο τα σημεία x.

Τα SVMs εντοπίζουν τις ευθείες που διαχωρίζουν με τον καλύτερο δυνατό τρόπο τα δεδομένα στον χώρο $\Re^M$. Οι ευθείες, επειδή μιλάμε για σημεία που έχουν N διαστάσεις, αποτελούν ένα υπερεπίπεδο διάστασης $N-1$, το οποίο ορίζεται από σημεία που ανήκουν στα ίδια τα δεδομένα που μελετούνται, και τα σημεία αυτά χαρακτηρίζονται ως διανύσματα υποστήριξης, με την έννοια ότι *κρατάνε/ορίζουν* το υπερεπίπεδο. Ο χώρος διαχωρισμού διαφέρει από τον χώρο στον οποίο βρίσκονται τα δεδομένα, για αυτό και χρησιμοποιήθηκε το M έναντι του N που χαρακτηρίζει τον χώρο των δεδομένων. Επομένως από τα παραπάνω, ο στόχος των SVMs είναι να εντοπίσουν το υπερεπίπεδο που επιτυγχάνει την καλύτερη κατηγοριοποίηση των δεδομένων εκμάθησης, δηλαδή ελαχιστοποιεί το σφάλμα κατάταξης και αντιπροσωπεύει το μεγαλύτερο διαχωρισμό των κλάσεων. Είναι προφανές ότι πρόκειται για ένα πρόβλημα βελτιστοποίησης, όπου αναζητείται το υπερεπίπεδο που μεγιστοποιεί την απόσταση από τα κοντινότερα σε αυτό σημεία των δύο κλάσεων.

Όπως αναφέρθηκε παραπάνω, ο διαχωρισμός των κλάσεων γίνεται όχι απαραίτητα στο χώρο των δεδομένων, αλλά αυτά μπορούν να μετασχηματιστούν από τον χώρο $\Re^N$ στον $\Re^M$ όπου είναι πιο έυκολα διαχωρίσιμα, βρίσκεται το υπερεπίπεδο, και στη συνέχεια μετασχηματίζονται όλα πίσω στον αρχικό χώρο. Ο μετασχηματισμός αυτός καλείται συνάρτηση κελύφους (kernel function) και θα αναλυθεί σε μετέπειτα παραγράφους. Παρακάτω περιγράφονται οι βασικές αρχές των SVMs [41].

Έστω ότι το σύνολο εκμάθησης αποτελείται από n σημεία της μορφής:

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$$

όπου τα y_i είναι 1 ή -1 και χαρακτηρίζει την κλάση του σημείου $\mathbf{x}_i \in \Re^N$. Στόχος είναι να βρεθεί το *υπερεπίπεδου μεγίστου περιθωρίου* που διαχωρίζει τα σημεία με κλάση $y_i = 1$ από αυτά με κλάση $y_i = -1$, και ορίζεται έτσι ώστε η απόσταση μεταξύ του υπερεπίπεδου και τα πιο κοντινά σημεία $\mathbf{x}_i$ από τις δύο κλάσεις μεγιστοποιείται. Ένα οποιοδήποτε υπερεπίπεδο στο χώρο $\Re^N$ γράφεται:

$$\mathbf{w}^T \cdot \mathbf{x} - b = 0 \tag{2.23}$$

όπου το $\mathbf{w}$ είναι το κάθετο διάνυσμα στο υπερεπίπεδο. Να σημειωθεί ότι στη μορφή αυτή το διάνυσμα $\mathbf{w}$ δεν είναι απαραίτητα κανονικοποιημένο, αλλά προς ευκολία, θα δουλέψουμε με τα κανονικοποιημένα διανύσματα, οπότε ένα σημείο x_i των δεδομένων που έχουμε ορίζει υπερεπίπεδο παράλληλο στο (2.23):

$$|\mathbf{w}^T \cdot \mathbf{x_i} - b| = 1 \tag{2.24}$$

Επομένως η απόσταση του σημείου x_i από ένα σημείο x πάνω στο υπερεπίπεδο (2.23) είναι:

$$\begin{aligned} d &= \frac{|\mathbf{w}^T(\mathbf{x}_i - \mathbf{x})|}{\|\mathbf{w}\|} \\ &= \frac{|\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}|}{\|\mathbf{w}\|} \\ &= \frac{1}{\|\mathbf{w}\|} \end{aligned} \tag{2.25}$$

Όπως αναφέρθηκε, ο σκοπός των SVMs είναι να μεγιστοποιήσουν αυτή την απόσταση, η οποία είναι η ίδια από τα σημεία της μίας και της άλλης κλάσης. Άρα το πρόβλημα βελτιστοποίησης είναι:

$$\begin{aligned} &\max \frac{1}{\|\mathbf{w}\|} \\ &s.t. \min_{n=1,\dots,n} |\mathbf{w}^T\mathbf{x}_i + b| = 1 \end{aligned} \tag{2.26}$$

Ο περιορισμός που τίθεται, δεδομένου ότι η κλάση y_i παίρνει τιμές ± 1 γράφεται

$$y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \tag{2.27}$$

Και το πρόβλημα μπορεί να ξαναγραφεί ώστε να είναι τετραγωνικής (quadratic) μορφής, οπότε τελικά γράφεται:

$$\begin{aligned} &\min \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ &s.t. y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 \Leftrightarrow \\ &\Leftrightarrow a_i\left[y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) - 1\right] \geq 0 \end{aligned} \tag{2.28}$$

Η (2.28) μετατρέπεται σε Λαγκρανζιανή μορφή σε ένα εννιαίο πρόβλημα βελτιστοποίησης Karun-Kuhn-Tucker:

$$\min L_P(\mathbf{w}, b, a) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{n} a_i\left[y_i\left(\mathbf{w}^T\mathbf{x_i} + b\right) - 1\right] \tag{2.29}$$

Η (2.29) πρέπει να ελαχιστοποιηθεί ως προς τις μεταβλητές $\mathbf{w}, b$ και να μεγιστοποιηθεί ως προς τα $a_i \geq 0$ Από την ιδιότητα ότι οι παράγωγοι ως προς $\mathbf{w}, b$ στο 0 είναι μηδενικές, έχουμε:

$$\frac{\partial L_P}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n} a_i y_i x_i = 0 \tag{2.30}$$

$$\frac{\partial L_P}{\partial b} = \sum_{i=1}^{n} a_i y_i = 0 \tag{2.31}$$

Από τις (2.30) προκύπτει άμεσα ότι

$$\mathbf{w} = \sum_{i=1}^{n} a_i y_i \mathbf{x}_i, \qquad \sum_{i=1}^{n} a_i y_i = 0 \tag{2.32}$$

και άρα η (2.29) αναπτύσσεται σε:

$$\min L_P(\mathbf{w}, b) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{n} a_i \left[y_i \left(\mathbf{w}^T \mathbf{x_i} + b \right) - 1 \right] \Leftrightarrow \tag{2.33}$$

$$\max L_D(a_i) = \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i=1}^{n} a_i a_j y_i y_j \left(\mathbf{x}_i \cdot \mathbf{x}_j \right) \tag{2.34}$$

$$s.t. \sum_{i=1}^{n} a_i y_i = 0, \qquad a_i \geq 0 \tag{2.35}$$

Πρακτικά πρόκειται για το ίδιο πρόβλημα βελτιστοποίησης αλλά από διαφορετική πλευρά παραμέτρων, για αυτό χρησιμοποιούνται οι δείκτες P και D συμβολίζοντας αντίστοιχα το primal και dual πρόβλημα. Η μορφή (2.34) της αντικειμενικής έχει μεγάλη σημασία, καθώς για τον υπολογισμό της αρκεί κάποιος να υπολογίσει το εσωτερικό γινόμενο $(\mathbf{x}_i \cdot \mathbf{x}_j)$. Επίσης, η (2.34) λύνεται εύκολα με κάποιον αλγόριθμο quadratic programming και έτσι βρίσκεται η λύση

$$\mathbf{a} = [a_1, \ldots, a_n] \tag{2.36}$$

Οπότε

$$\mathbf{w} = \sum_{i=1}^{n} a_i y_i \mathbf{x}_i \tag{2.37}$$

Να σημειωθεί ότι τα περισσότερα a_i και άρα και $\mathbf{w}_i$ θα είναι ίσα με το 0. Μόνο τα σημεία x_i τα οποία αποτελούν διανύσματα στήριξης θα έχουν μη μηδενικά a_i. Αυτό μειώνει σε σημαντικό βαθμό τη διάσταση του τελικού διανύσματος-λύση. Πλέον το μοντέλο έχει εκπαιδευθεί με τα n σημεία, και μπορεί να δώσει απάντηση για την κλάση ενός άγνωστου σημείου, έστω y^* για το οποίο γνωρίζουμε $\mathbf{x_u} = x_1, \ldots, x_N$ και άρα αρκεί να βρεθεί το πρόσημο της παρακάτω σχέσης:

$$f(x) = \mathbf{w} \cdot \mathbf{x_u} + b = \sum_{i=1}^{n} a_i y_i \mathbf{x}_i \cdot \mathbf{x_u} + b \tag{2.38}$$

Και το πρόσημο της $f(x)$ υποδεικνύει την κλάση y^*. Αν είναι αρνητικό, τότε $y^* = -1$, αν είναι θετικό τότε $y^* = 1$.
Η παραπάνω διαδικασία αφορά όμως σημεία τα οποία στις διαστάσεις που μελετούνται είναι γραμμικά διαχωριζόμενα. Κάτι τέτοιο σίγουρα δεν ισχύει για όλα τα σύνολα εκμάθησης, τα οποία ίσως δεν είναι γραμμικά διαχωριζόμενα, αλλά μπορεί να διαχωρίζονται από κάποια άλλη συνάρτηση μη γραμμική. Για παράδειγμα, δεδομένα που διαχωρίζονται από τριωνυμική συνάρτηση μπορούν να μετασχηματιστούν $x \mapsto \{x^2, x\}$ και στο μετασχηματισμένο χώρο υπάρχει γραμμικός διαχωρισμός. Στην περίπτωση αυτή όμως, έστω ϕ ο μετασχηματισμός που πρέπει να γίνει, θα πρέπει να υπολογιστεί το $\phi(x_i) \cdot \phi(x_j)$, το οποίο μπορεί να αποδειχθεί πολύ ακριβό από υπολογιστική άποψη. Το πρόβλημα αυτό λύνεται με τη συνάρτηση κελύφους K τέτοια ώστε $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$, οπότε χρειάζεται μόνο να υπολογιστεί η τιμή της K ώστε να λυθεί το πρόβημα, το οποίο μετασχηματίζεται στο:

$$L_D = \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i=1}^{n} a_i a_j y_i y_j K(x_i \cdot x_j) \tag{2.39}$$

Το ποια συνάρτηση κελύφους θα επιλεχθεί εξαρτάται από τη μορφή των δεδομένων. Οι συνηθισμένες είναι οι παρακάτω:

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y})^p \tag{2.40}$$

$$K(\mathbf{x}, \mathbf{y}) = exp\{-\frac{\|\mathbf{x} - \mathbf{y}\|}{2\sigma^2} \tag{2.41}$$

$$K(\mathbf{x}, \mathbf{y}) = \tanh(\kappa\mathbf{x} \cdot \mathbf{y} - \delta) \tag{2.42}$$

Όπου: (2.40) πολυωνυμική συνάρτηση κελύφους, με ξεχωριστή περίπτωση το απλό εσωτερικό γινόμενο, (2.41) η ακτινική, που χρησιμοποιείται για δεδομένα των οποίων οι κλάσεις διαχωρίζονται σε σφαίρες, και (2.42) η σιγμοϊδής συνάρτηση κελύφους που χρησιμοποιείται και στα νευρωνικά δίκτυα. Με τις συναρτήσεις κελύφους, ο αναλυτής μπορεί να βελτιστοποιήσει την κατηγοριοποίηση των μεταβλητών του. Σημαντικό είναι να υπάρχει προσοχή για το overfitting των δεδομένων, κάτι που εκτείνεται πέρα από τις απαιτήσεις της παρούσης εργασίας.

2.9 Γραμμική Διακριτική Ανάλυση LDA

Η γραμμική διακριτική ανάλυση μοιάζει στην κεντρική ιδέα της με την ανάλυση σε κύριους παράγοντες. Και στις δύο μεθόδους, ουσιαστικά συνδυάζονται διαφορετικά χαρακτηριστικά δύο αντικειμένων (στην περίπτωση που μελετάμε τα χαρακτηριστικά είναι τα γονίδια και τα αντικείμενα τα δείγματα), με αποτέλεσμα να προκύψουν διαφορετικά χαρακτηριστικά, όπως οι κύριοι παράγοντες στην PCA. Η διαφορά της γραμμικής διακριτικής ανάλυσης είναι ότι στην περίπτωση αυτή, τα χαρακτηριστικά συνδυάζονται με τέτοιο τρόπο ώστε να υπάρχει όσο το δυνατόν καλύτερος διαχωρισμός γίνεται μεταξύ των αντικειμένων, ενώ η ανάλυση σε κύριους παράγοντες δε λαμβάνει υπ'όψιν τις διαφορές μεταξύ διαφορετικών κατηγοριών, αλλά προσπαθεί να βρει τις διαφορετικές συνιστώσεις που μεγιστοποιούν τη διακύμανση μεταξύ των δεδομένων.
Η μέθοδος της γραμμικής διακριτικής ανάλυσης μπορεί να χρησιμοποιηθεί τόσο για την οπτικοποίηση των δεδομένων κατά την exploratory data analysis για να φανεί ο διαχωρισμός των δεδομένων όπως γίνεται και με την ανάλυση σε κύριες συνιστώσες, αλλά και για την κατασκευή ενός γραμμικού μοντέλου πρόβλεψης, στα πλαίσια των μηχανών εκμάθησης. Συγκεκριμένα, η μέθοδος υποθέτει *a priori* ότι τα δεδομένα κάθε κλάσης έχουν τη μορφή πολυδιάστατης κανονικής κατανομής (multivariate normal distribution) και ότι ο πίνακας συνδιακύμανσης είναι κοινός για όλες τις κλάσεις. Στη συνέχεια χρησιμοποιεί Bayes κατηγοριοποιητή:

$$\hat{y}_0 = \arg\max_y \hat{P}(Y = y|X = x_0) \tag{2.43}$$

όπου y οι διάφορες κλάσεις, και $\hat{y}_0$ η κλάση που μεγιστοποιεί την πιθανότητα $\hat{P}(Y = y|X = x_0)$.
Έστω ότι έχουμε n δείγματα p διαστάσεων το καθένα, κατανεμημένα σε k διαφορετικές κλάσεις.
Σύμφωνα με το θεώρημα Bayes ισχύει:

$$P(Y = k|X = x) = \frac{P(X = x|Y = k)\hat{P}(Y = k)}{\hat{P}(X = x)} \tag{2.44}$$

το οποίο με λόγια περιγράφεται: η πιθανότητα να έχουμε κλάση k όταν υπάρχει η παρατήρηση x ισούται με την πιθανότητα να έχουμε παρατήρηση x όταν υπάρχει κλάση k επί την πιθανότητα παρατήρησης της κλάσης k δια την πιθανότητα να έχουμε παρατήρηση x.

Τώρα, όπως αναφέρθηκε παραπάνω, η μέθοδος υποθέτει κανονική κατανομή σε κάθε κλάση, και a priori θεώρηση της συχνότητας των κλάσεων, δηλαδή με μαθηματικούς όρους:

$$P(X = x|Y = k) = \hat{f}_k(x) = \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\right) \tag{2.45}$$

$$\hat{P}(Y = k) = \hat{\pi}_k \tag{2.46}$$

όπου Σ η κοινή για όλες τις κλάσεις συνδιακύμανση $\Sigma = \frac{1}{n-k}\sum_{j=1}^{k}\sum_{y_i=k}(x_i-\mu_j)(x_i-\mu_j)^T$, και $\mu_k = \frac{1}{n_k}\sum_{y_i=k} x_i$ ο μέσος όρος κάθε κλάσης. Επομένως, η (2.44) σε συνδυασμό με τα παραπάνω, μετατρέπει τον κατηγοριοποιητή στον:

$$P(Y = k|X = x) = \frac{f_k(x)\pi_k}{P(X = x)} = Cf_k(x)\pi_k \tag{2.47}$$

όπου οι όροι που δεν επηρεάζονται από την επιλογή της κλάσης k συγκεντρώνονται στη σταθερά C. Αντικαθιστώντας την κατανομή (2.45) παίρνουμε:

$$P(Y = k|X = x) = \frac{C\pi_k}{(2\pi)^{\frac{p}{2}}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\right) \tag{2.48}$$

$$= C_1\pi_k \exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\right) \tag{2.49}$$

Και λογαριθμώντας από κάθε πλευρά παίρνουμε:

$$\log\left[P(Y = k|X = x)\right] = \log C_1 + \log\pi_k - \frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k) \tag{2.50}$$

$$= \log C_1 + \log\pi_k - \frac{1}{2}\left[x^T\Sigma^{-1}x + \mu_k^T\Sigma^{-1}\mu_k\right] + x^T\Sigma^{-1}\mu_k \tag{2.51}$$

$$= C_2 + \log\pi_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + x^T\Sigma^{-1}\mu_k \tag{2.52}$$

Οπότε η κλάση που μεγιστοποιεί την (2.52) είναι η κλάση στην οποία προβλέπει το μοντέλο ότι βρίσκεται η παρατήρηση x.

Κεφάλαιο 3

Αποτελέσματα

3.1 GEO series

Για την υπολογιστική μελέτη του μελανώματος, και προκειμένου να καλυφθεί μεγάλο έυρος περιπτώσεων και μορφών της ασθένειας, ήταν αναγκαίο να βρεθούν μέσα από τη βάση δεδομένων του GEO τα σχετικά πειράματα. Με τις λέξεις κλειδιά **melanoma**, **melanocyte** που αντιστοιχούν σε καρκινικά δείγματα μελανώματος και σε υγιή μελανοκύτταρα (καθώς αυτά είναι αναγκαίο να υπάρχουν για τη σύγκριση case vs. control. Η αναζήτηση έγινε με φίλτρο την εμφάνιση μόνο GEO series, δηλαδή αυτόνομα πειράματα με τα δικά τους δείγματα, ενώ παράλληλα κρίθηκε αναγκαίο ο προμηθευτής του πειράματος να παρέχει τα ακατέργαστα δεδομένα raw μορφής, ώστε να είναι δυνατή η κατάλληλη επεξεργασία (κανονικοποίηση και αφαίρεση θορύβου μέσω του αλγόριθμου rma), απαραίτητη για τη μετέπειτα ενσωμάτωση των πειραμάτων σε ένα εννιαίο χώρο.
Από τα περίπου 880 πειράματα που εμφανίστηκαν στο αποθετήριο μετά την αναζήτηση των παραπάνω όρων, μόνο 4 από αυτά κρίθηκαν κατάλληλα για τις ανάγκες της μελέτης. Αυτό οφείλεται σε τρεις λόγους:

- Τα περισσότερα από τα πειράματα που εμφανίστηκαν, περιελάμβαναν σύγκριση μεταξύ καρκινικών κυττάρων και των ίδιων κυττάρων αφού εισάχθηκε και επέδρασε πάνω τους κάποια συγκεκριμένη ουσία. Η μελέτη που επιθυμούμε πρέπει να περιλαμβάνει κύτταρα υγιή και καρκινικά μόνο.

- Πολλές υποψήφιες καταχωρήσεις περιείχαν μόνο τα δεδομένα αφού ο παροχέας έκανε την προεπεξεργασία με τις μεθόδους της επιλογής του. Τα δεδομένα που ψάχνουμε πρέπει να είναι σε ακατέργαστη raw μορφή.

- Η πλατφόρμα GEO υποστηρίζει εκτός από πειράματα σε μικροσυστοιχία, και διαφορετικά πειράματα μέτρησης της γονιδιακής έκφρασης τα οποία βασίζονται σε διαφορετικές τεχνολογίες υψηλής απόδοσης, όπως η RNA-Seq, L1000 και άλλες. Κάποιες από τις τεχνολογίες αυτές, όπως οι προαναφερθήσες, αν και πιο σύγχρονες σε σχέση με τη μικροσυστοιχία DNA και ολοένα και πιο συνηθισμένες, βασίζονται σε διαφορετικές αρχές για τον υπολογισμό της γονιδιακής έκφρασης. Για παράδειγμα, η RNA-Seq δε μετράει με κάποιο τρόπο ένα συνεχές μέγεθος όπως γίνεται στη μικροσυστοιχία, αλλά υπολογίζει την έκφραση ανάλογα με τον αριθμό αλληλουχιών RNA που προσδένονται στο γονιδίωμα αναφοράς, δηλαδή πρόκειται για διακριτές μετρήσεις. Προφανώς, τα εργαλεία για την επεξεργασία τέτοιων δεδομένων είναι

στις περισσότερες περιπτώσεις διαφορετικά από αυτά που χρησιμοποιούνται στις μικροσυστοιχίες.

- Τέλος, από τα τέσσερα πειράματα που επιλέχθηκαν, ο αριθμός των υγιών κυττάρων δέρματος είναι πολύ μικρός σε σχέση με τα καρκινικά κύτταρα. Συγκεκριμένα, περιέχονται συνολικά μόνο 6 υγιή κύτταρα έναντι των 144 καρκινικών. Η ανισορροπία που εισάγεται στο σύστημα αντιμετωπίστηκε με μεθόδους που περιγράφονται στην ενότητα 3.3, αλλά τα πειράματα περιορίστηκαν σε τέσσερα ώστε να μη μεγαλώσει κι άλλο η διαφορά στον αριθμό. Με τα 144 καρκινικά κύτταρα, στα οποία συμπεριλαμβάνονται πολλές σημαντικές κυτταροσειρές, θεωρείται ότι δημιουργείται ένα αρκετά ευρύ πεδίο περιπτώσεων, και είναι δυνατή η στοχευμένη ανάλυση, δηλαδή ξεχωριστή ανάλυση για ξεχωριστή μορφή του μελανώματος. Οι τέσσερεις σειρές του GEO που χρησιμοποιήθηκαν αναγράφονται παρακάτω.

GSE7127 [42]: Πείραμα δημοσιευμένο το 2007 στο οποίο μετρήθηκαν 63 κυτταροσειρές μελανώματος. Τα 63 δείγματα υβριδοποιήθηκαν στην πλατφόρμα **GPL570** που αντιστοιχεί στην Affymetrix Human Genome U1333 Plus 2.0 Array - [HG-U133_Plus_2].

GSE36133 [43]: Πείραμα δημοσιευμένο το 2012 με γονιδιακά δεδομένα από την Cancer Cell Line Encyclopedia (CCLE) (Εγκυκλοπαίδεια Καρκινικών Κυτταρικών Σειρών). Η CCLE είναι αποτέλεσμα συνεργασίας του Broad institute, που κατασκεύασε την πλατφόρμα clue.io και το cMap, των Novartis Institutes for Biomedical Research και του Genomics Novartis Foundation. Σκοπός του έργου είναι η κατασκευή λεπτομερών γενετικών και φαρμακολογικών προφίλ ενός μεγάλου πλαισίου καρκινικών μοντέλων. Η CCLE περιέχει περισσότερα από 900 δείγματα καρκινικών κυτταροσειρών από 36 είδη καρκίνου, συμπεριλαμβανομένου και του δερματικού μελανώματος. Τα δείγματα υβριδοποιήθηκαν στην πλατφόρμα **GPL15308**, που αντιστοιχεί στην Affymetrix Human Genome U133 Plus 2.0 Array Brainarray Version 15.0.0 - [HGU133Plus2_Hs_ENTREZG]. Αυτή αποτελεί ειδική έκδοση της GPL570 που χρησιμοποιήθηκε στο πείραμα GSE7127 αλλά επεξεργασμένη ώστε κάθε ανιχνευτής να αντιστοιχεί σε συγκεκριμένο γονίδιο, με αντιστοιχία $1:1$.

GSE22301 [44]: Το συγκεκριμένο πείραμα, δημοσιευμένο το 2010, περιέχει δείγματα καρκινικών και υγιών κυτταροσειρών. Δύο από τις τέσσερεις υγιής κυτταροσειρές αποτελούν δείγματα αθανατοποιημένων κυτταροσειρών (immortalized). Αυτές έχουν αλλαχθεί στο εργαστήριο γενετικά, ώστε να *απενεργοποιηθούν* συγκεκριμένα γονίδια που προκαλούν κυτταρικό θάνατο σε απομονωμένα κύτταρα. Τα immortalized κύτταρα δεν μπορούν πάντα να θεωρούνται ίδια με τα υγιή, αλλά στη συγκεκριμένη μελέτη συμπεριλαμβάνονται στα υγιή όπως προέκυψε από ανάλυση που περιγράφεται στην ενότητα 3.2. Τα δείγματα της GSE22301 υβριδοποιήθηκαν στην πλατφόρμα **GPL571** Affymetrix Human Genome U133A 2.0 Array [HG-U133A_2].

GSE35388 [45]: Πείραμα του 2012 στο οποίο συμμετέχουν δύο καρκινικά και δύο υγιή δείγματα, τα οποία αποτελούνται από βιολογικά αντίγραφα. Ένα δείγμα μπορεί αποτελεί βιολογικό αντίγραφο ενός άλλου όταν προέρχονται από τον ίδιο πληθυσμό κυττάρων, αλλά έχουν αναπτυχθεί σε διαφορετικό περιβάλλον (φλάσκα). Έτσι μπορούν να συμπεριληφθούν στην ανάλυση χωρίς να εισάγεται στατιστικό σφάλμα. Αντίθετα, όταν πρόκειται για τεχνητά αντίγραφα (technical replicates), μετριέται το ίδιο δείγμα περισσότερες φορές, και δεν μπορεί να θεωρηθεί ότι δεν εισάγεται στατιστικό σφάλμα στο σύστημα. Η υβριδοποίηση των κυτταροσειρών έγινε στην πλατφόρμα GPL570 όπως και το πρώτο πείραμα.

Η πλατφόρμα είναι σε ένα πείραμα το πιο σημαντικό στοιχείο που θα το διαφοροποιήσει

από ένα άλλο. Αυτό δικαιολογείται από το διαφορετικό αριθμό και είδος ανιχνευτών που αντιστοιχούν στα γονίδια που μετρούνται, και στο ότι κάθε πλατφόρμα αποδίδει ελαφρώς διαφορετικές τιμές στην ίδια φωφορίζουσα ένταση από τα κομμάτια RNA. Είναι σημαντικό να παρατηρηθεί ότι οι πλατφόρμες των πειραμάτων δε διαφέρουν σε μεγάλο βαθμό μεταξύ τους: είναι όλες διαφορετικές εκδόσεις της Affymetrix Human Genome U133 2. Περιληπτικά τα δεδομένα περιγράφονται στον πίνακα 3.1. Το πρώτο σημαντικό βήμα είναι η

Πίνακας 3.1: Πειράματα προς μελέτη

Πείραμα	Πλατφόρμα	Καρκινικά δείγματα	Υγιή δείγματα
GSE7127	GPL570	63	-
GSE36133	GPL15308	61	-
GSE22301	GPL571	18	4
GSE35388	GPL570	2	2

αντιστοίχιση των ανιχνευτών στα γονίδια, για κάθε πείραμα. Όταν πολλοί ανιχνευτές αντιστοιχούν σε ένα γονίδιο, μπορεί να επιλεχθεί είτε η διάμεσος των μετρήσεων, είτε ο μέσος όρος τους. Στην παρούσα μελέτη επιλέχθηκε ο μέσος όρος, αφού πρώτα ελέγχθηκε ότι ακραίες μετρήσεις έχουν ομαλοποιηθεί με την κανονικοποίηση. Έτσι έχουμε τον πίνακα 3.2: Καθώς θα γίνει συνδυασμός όλων των πειραμάτων, είναι προφανές πως πρέπει να

Πίνακας 3.2: Ανιχνευτές και Γονίδια

Πείραμα	Πλατφόρμα	Αριθμός ανιχνευτών	Αριθμός γονιδίων
GSE7127	GPL570	54675	20218
GSE36133	GPL15308	18988	18988
GSE22301	GPL571	22277	12421
GSE35388	GPL570	54675	20218

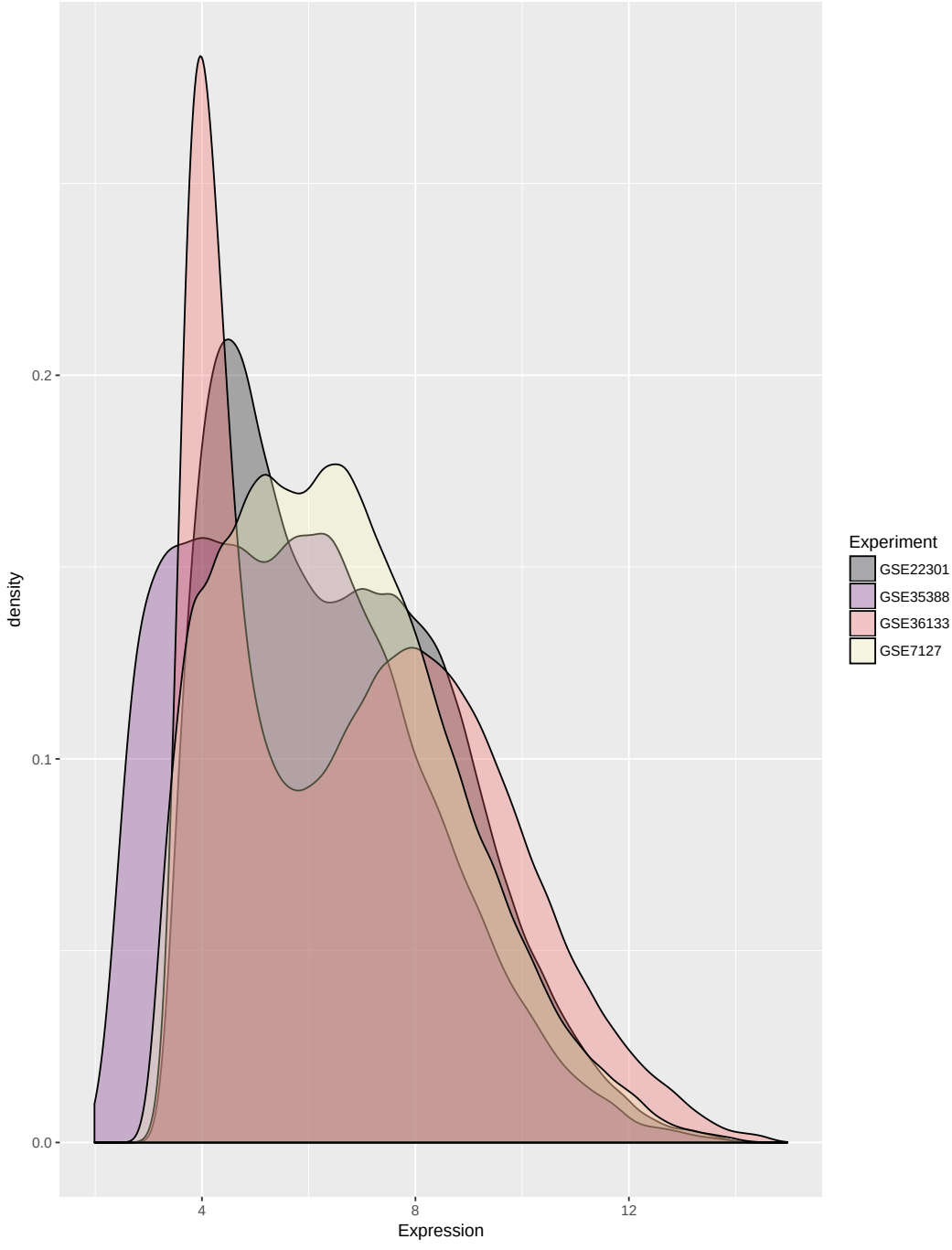
επιλεχθούν τα κοινά γονίδια όλων των πλατφόρμων, και επομένως θα αναλυθούν λιγότερα από 12421 γονίδια, τα οποία μετρούνται στο τρίτο πείραμα. Κάτι τέτοιο περιορίζει τα αποτελέσματα της γονιδιακής ανάλυσης, καθώς διάφορα γονίδια δε θα ληφθούν καθόλου υπ'όψιν.
Πριν κάποιος προχωρήσει στην ενσωμάτωση όλων των δεδομένων σε ένα εννιαίο πείραμα, πρέπει να παρατηρηθεί η αρχική μορφή των δεδομένων μετά την αφαίρεση του θορύβου και την κανονικοποίσή τους, σε κάθε πείραμα μεμονωμένα. Με αυτόν τον τρόπο βεβαιώνεται ο αναλυτής ότι η μορφή των δεδομένων είναι η αναμενόμενη, και ότι δεν υπάρχουν πιθανά σφάλματα, όπως ακραίες μετρήσεις. Στα σχήματα 3.1 έως 3.3 φαίνεται η κατανομή των μετρήσεων ανά πείραμα, τύπο κυττάρου και πλατφόρμα αντίστοιχα. Οι κατανομές που παρατηρούνται σε όλες τις περιπτώσεις είναι *δικόρυφες* ή αλλιώς *διμοδικές* με θετική κλίση positive bimodal distribution. Κάτι τέτοιο είναι αναμενόμενο σε κατανομές μετρησεων γονιδιακής έκφρασης, όπου το μεγαλύτερο ποσοστό των γονιδίων βρίσκεται στην αριστερή επικρατούσα τιμή, και τα υπερεκφραζόμενα γονίδια αποτελούν τη δεξιά επικρατούσα τιμή που είναι και μικρότερη. Με μια πρώτη ματιά φαίνονται μεγάλες διαφορές στις κορυφές και τις κατανομές ανάμεσα στα πειράματα και τις πλατφόρμες, ενώ μικρότερες είναι οι διαφορές ανάμεσα στις κατανομές κατά τύπο κυττάρου.

Στη συνέχεια εκτελείται με τις μεθόδους της ενότητας 2.4 ανάλυση σε κύριες συνιστώσες, προκειμένου να αποκαλυφθούν οι σχέσεις ανάμεσα στα δείγματα κάθε πειράματος, αλλά και ανάμεσα στα διαφορετικά πειράματα. Πρέπει να σημειωθεί πως οι κύριες συνιστώσες

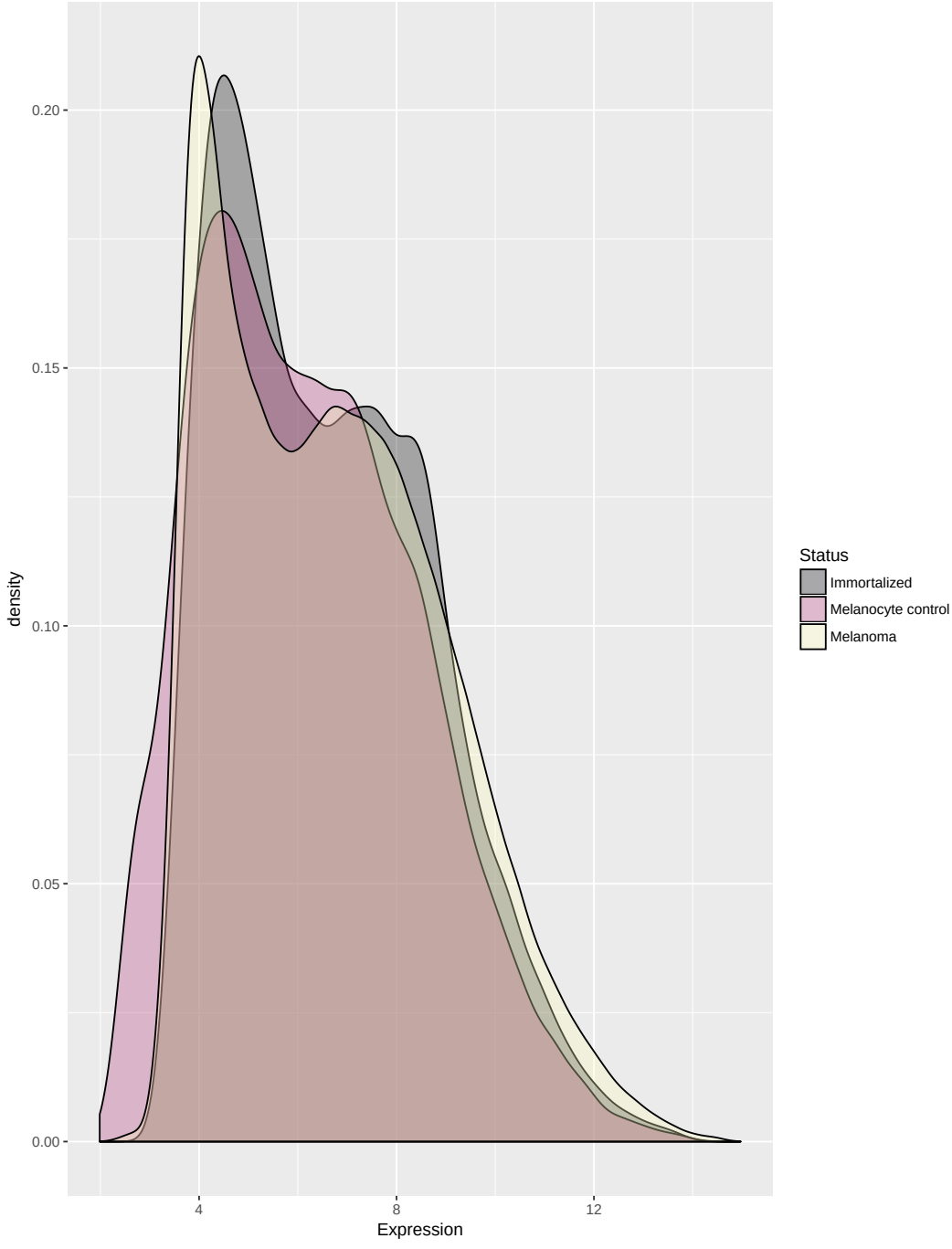
ονομάζονται με βάση τη διακύμανση που εξηγεί η κάθε μία. Δηλαδή η πρώτη κύρια συνιστώσα εξηγεί τη μεγαλύτερη διακύμανση, η δεύτερη λιγότερη κ.ο.κ. Για να θεωρηθεί η ανάλυση σε κύριες συνιστώσες επιτυχημένη και να μπορεί να χρησιμοποιηθεί στην κατάντι μελέτη, πρέπει οι συνιστώσες που θα διατηρηθούν να εξηγούν ικανοποιητικό ποσοστό της συνολικής διακύμανσης. Από τη στιγμή που γίνεται PCA ως προς τα δείγματα και όχι ως προς τα γονίδια, καθώς πρέπει να βρεθούν οι σχέσεις ανάμεσα στις κυτταροσειρές, θα υπάρχουν 150 συνιστώσες, λόγω των 150 δειγμάτων ($144 + 6$ καρκινικά/υγιή κύτταρα αντίστοιχα). Οι συνιστώσες που θα διατηρηθούν πρέπει ως ελάχιστο να εξηγούν τουλάχιστον 10% της διακύμανσης, μετά την ενσωμάτωση, δεδομένης της πολυπλοκότητας του μελανώματος.
Όπως αναφέρθηκε παραπάνω τα γονίδια στα οποια έγινε η ανάλυση πρέπει να είναι παρόντα σε όλα τα πειράματα. Μετά το φιλτράρισμα αυτό, δημιουργείται ένα εννιαίο μητρώο $\mathbf{X}_{11328\times151}$ δηλαδή διατηρούνται συνολικά 11328 γονίδια στις γραμμές και τα 150 δείγματα στις στήλες συν την πρώτη στήλη στην οποία γράφονται τα ονόματα των διαφόρων γονιδίων. Παρακάτω φαίνονται οι πρώτες πέντε γραμμές και στήλες του πίνακα $\mathbf{X}$.

SYMBOL	GSM162902	GSM162904	GSM162905	GSM162906
A1CF	3.9420	3.9905	3.9914	4.0449
A2M	6.9437	4.8343	8.0147	5.3661
A4GALT	6.8732	7.3280	6.9938	7.0655
A4GNT	4.4641	4.9375	4.6470	4.7928
AAAS	8.7370	8.0524	8.1302	7.7822

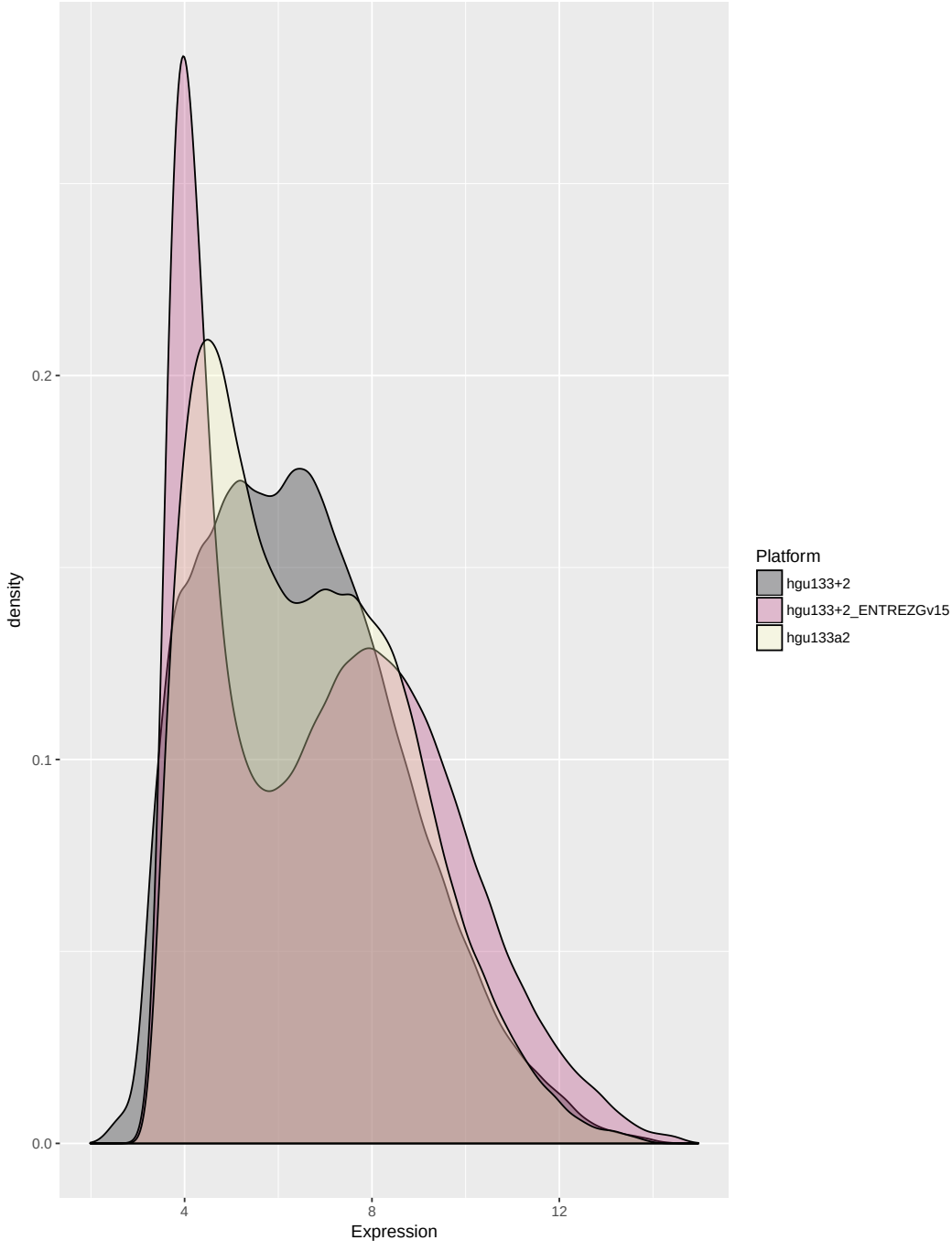
Πρέπει να σημειωθεί ότι το μητρώο $\mathbf{X}$ δεν αποτελεί το συνδυασμένο μητρώο, δηλαδή δεν έχουν διορθωθεί με τις Μπεϋζιανές μεθόδους της ενότητας 2.5 οι μετρήσεις, απλά έχουν παραταχθεί τα δείγματα σε στήλες, για να φιλτραριστούν όλα τα γονίδια που δεν είναι παρόντα και στα τέσσερα πειράματα. Στη συνέχεια στο μητρώο εκτελείται ανάλυση σε κύριες συνιστώσες, που παρουσιάζεται στα σχήματα 3.5 έως 3.7. Στον άξονα των x σχεδιάζεται η πρώτη κύρια συνιστώσα που περιγράφει περισσότερο από 30% της συνολικής διακύμανσης, και στον άξονα των y σχεδιάζεται η δεύτερη κύρια συνιστώστα που περιγράφει σχεδόν το 10%. Μαζί οι δύο συνιστώσες περιγράφουν περισσότερο από το 40% της συνολικής διακύμανσης, ακόμα και όταν γίνεται η ανάλυση σε όλα τα δείγματα ταυτόχρονα και όχι από κάθε δείγμα ξεχωριστά. Στο σχήμα 3.4 γίνεται προφανές ότι οι πρώτες κύριες συνιστώσες εξηγούν το μεγαλύτερο ποσοστό μεταβλητότητας των δεδομένων. Παρατηρείται πως και η τρίτη συνιστώσα συνισφέρει σημαντικά, σχεδόν όσο η δεύτερη, από τη στιγμή όμως που παίρνουμε αρκετές πληροφορίες από τις δύο πρώτες, που προσφέρουν καλύτερη οπτικοποίηση σε δύο διαστάσεις, δεν την υπολογίζουμε. Παρατηρείται κατευθείαν πως τα δείγματα χωρίζονται σε τρεις μεγάλες κατηγορίες. Προκειμένου να βρεθεί αν αυτή η διαφοροποίηση οφείλεται σε κάποιο στοιχείο του συστήματος και δεν είναι σφάλμα, οπτικοποιούνται στα σχήματα διαφορετικά χαρακτηριστικά των κυττάρων κάθε φορά. Στο σχήμα 3.5 παρατηρείται πως τα δείγματα δε δείχνουν να διαχωρίζονται κατά τύπο, δηλαδή τα υγιή και τα απαθανατισμένα κύτταρα βρίσκονται κοντά στα κύτταρα μελανώματος, οπότε ο τύπος κυττάρου δε μας δίνει κάποια πληροφορια για τη διαφοροποίηση. Τα αποτελέσματα είναι πιο ικανοποιητικά στο σχήμα 3.6 όπου τα πειράματα GSE22301 με κίτρινο χρώμα, και GSE36133 σχηματίζουν το καθένα από μία ομάδα. Το ίδιο δε συμβαίνει στα πειράματα GSE35388, GSE7127 όπου τα τέσσερα δείγματα του GSE35388 βρίσκονται ανάμεσα στα δειγματα του GSE7127. Ανατρέχοντας στον πίνακα 3.1 γίνεται φανερό ότι το κοινό μεταξύ των δύο πειραμάτων είναι η πλατφόρμα. Αυτό επιβεβαιώνεται στο σχήμα 3.7 όπου με τη διάκριση από πλατφόρμα σε πλατφόρμα (διαφορετικά σχήματα για κάθε μία), υπάρχει πλήρης και τέλειος διαχωρισμός των δειγμάτων. Επομένως, ο παράγοντας που επηρεάζει τις μετρήσεις και επιβάλλει τη διόρθωση των μετρήσεων είναι η πλατφόρμα.



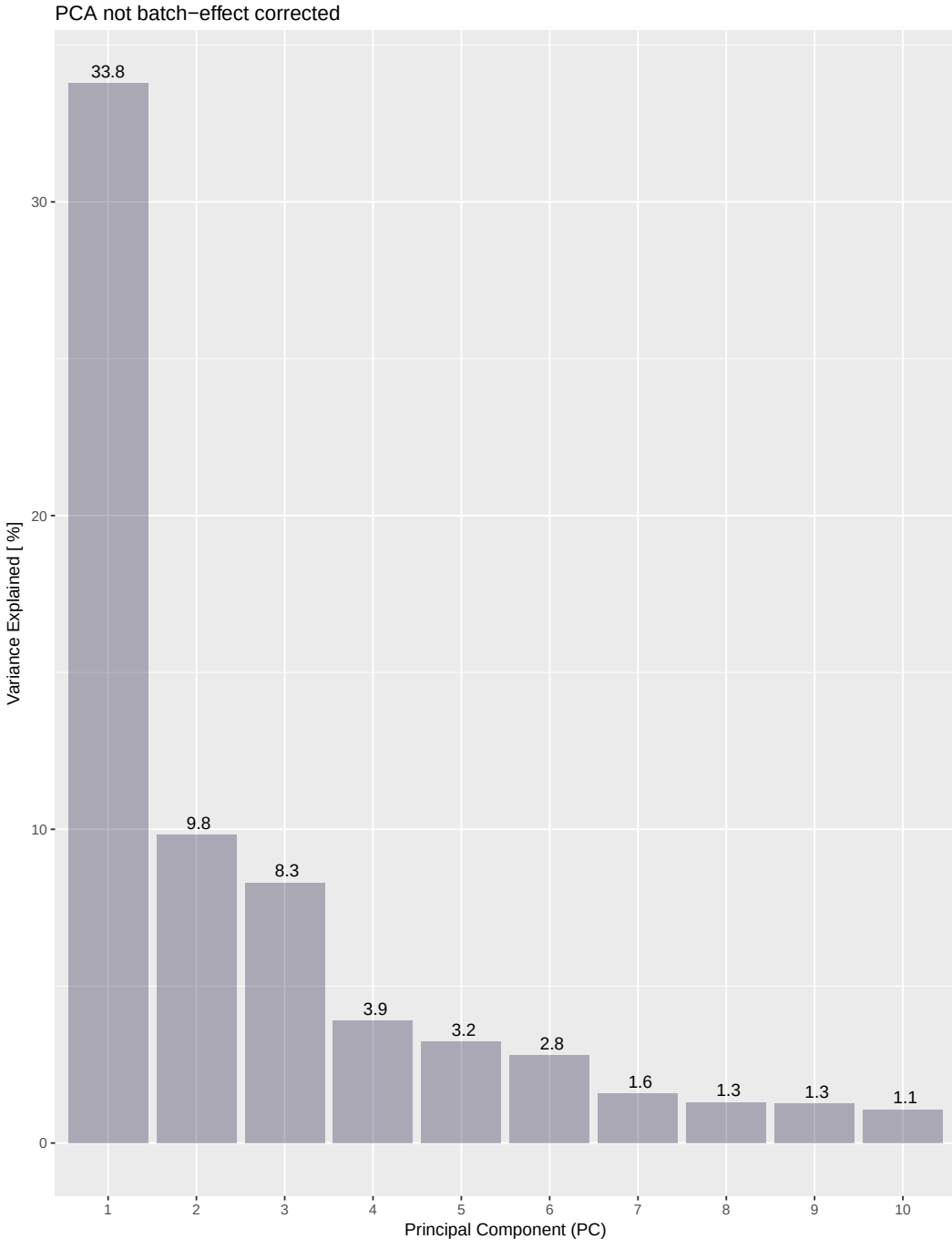
Σχήμα 3.1: Κατανομή μετρήσεων κατά πείραμα



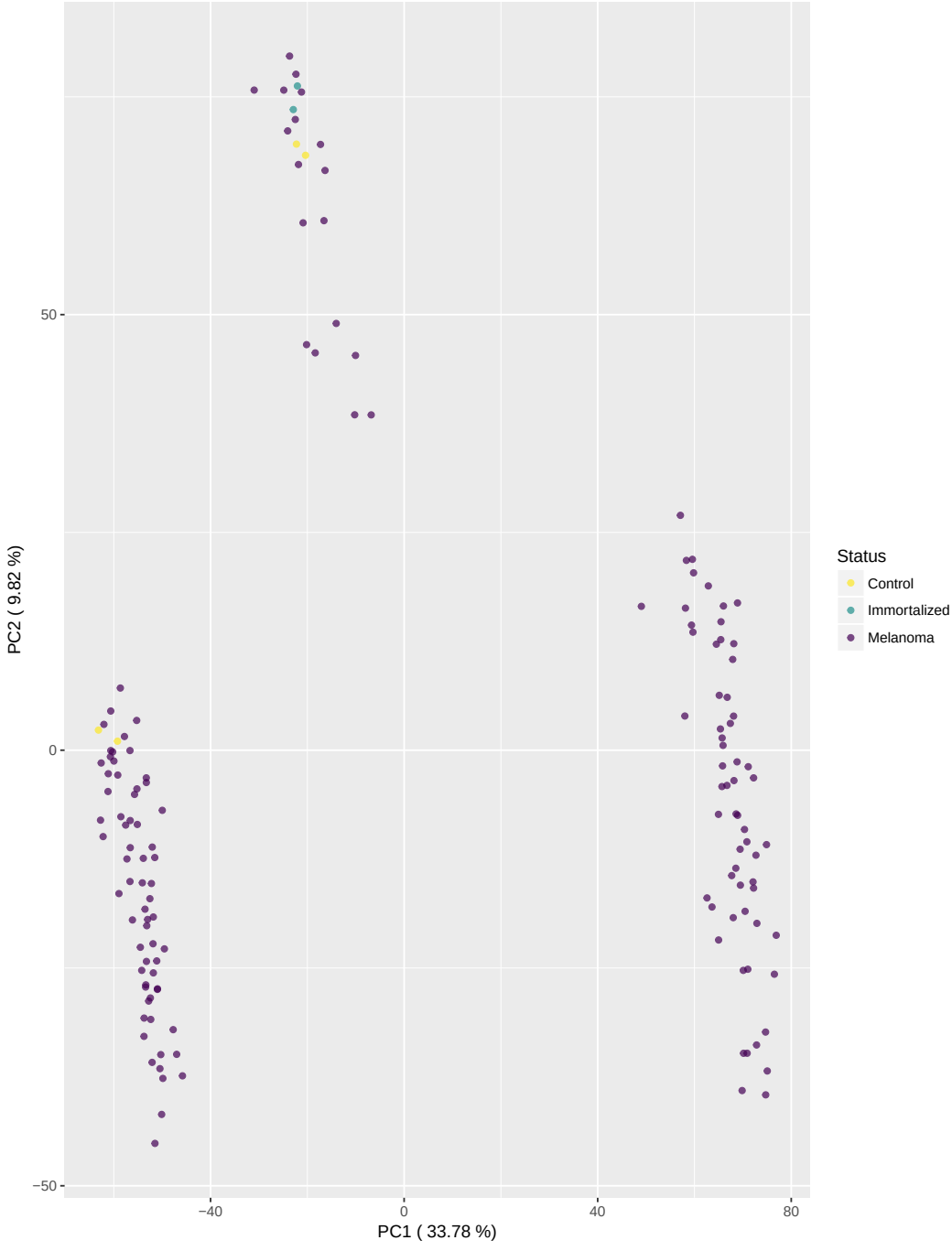
Σχήμα 3.2: Κατανομή μετρήσεων κατά τύπο κυττάρου



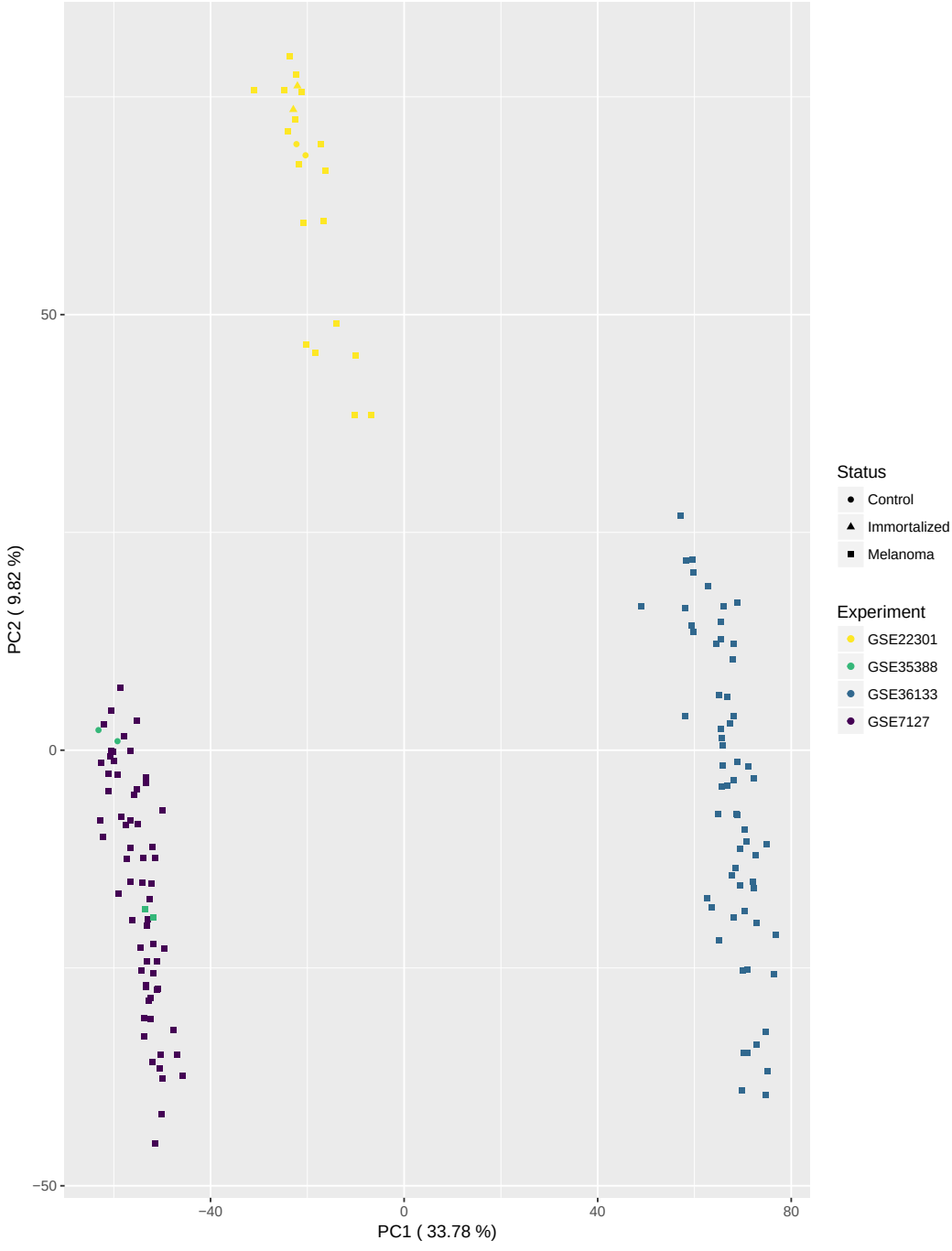
Σχήμα 3.3: Κατανομή μετρήσεων κατά πλατφόρμα



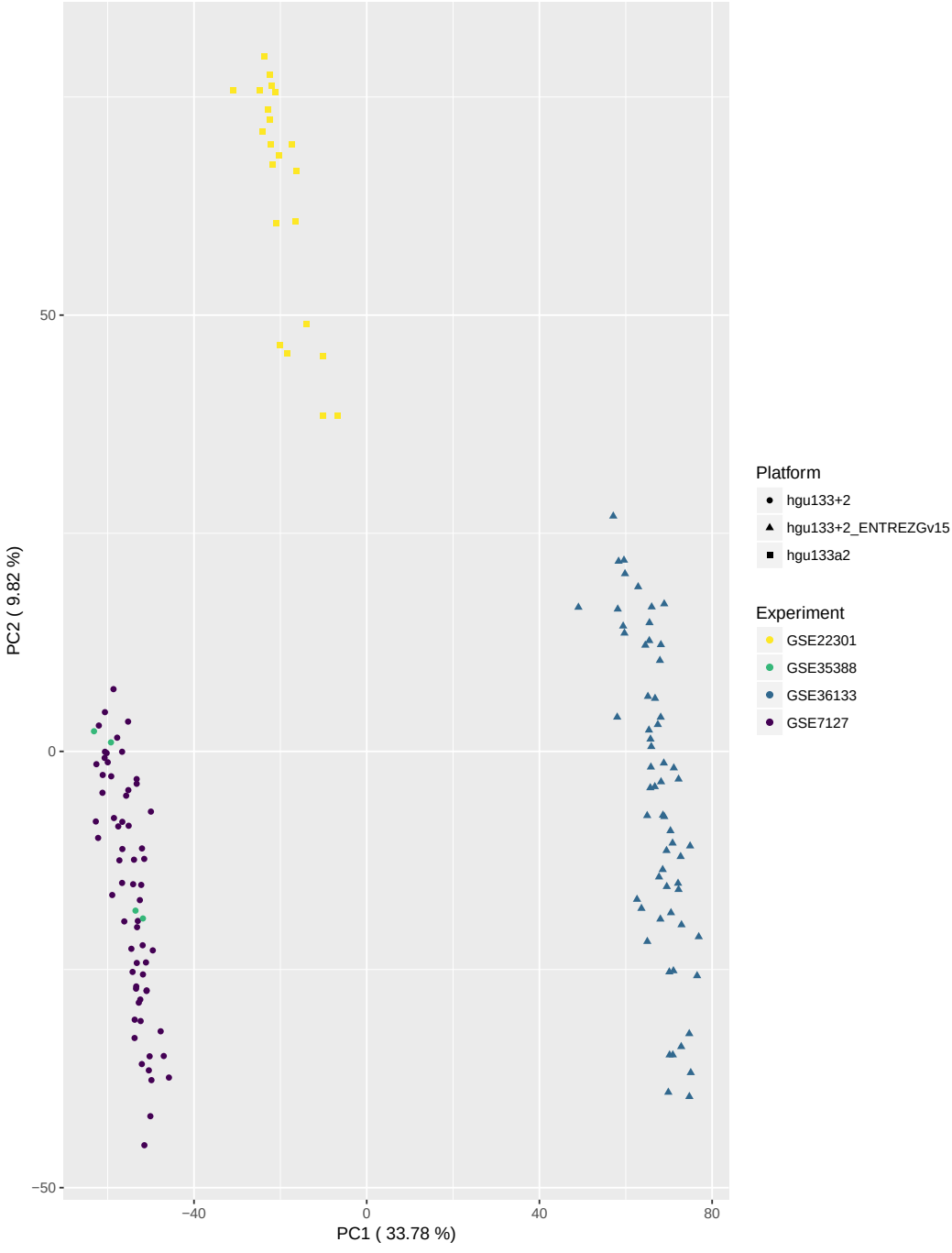
Σχήμα 3.4: Ποσοστό διακύμανσης που εξηγείται από κάθε συνιστώσα



Σχήμα 3.5: Ανάλυση σε κύριες συνιστώσες - τύπος κυττάρου



Σχήμα 3.6: Ανάλυση σε κύριες συνιστώσες - τύπος κυττάρου και πείραμα



Σχήμα 3.7: Ανάλυση σε κύριες συνιστώσες - τύπος κυττάρου και πλατφόρμα

3.2 Ενσωμάτωση δεδομένων

Όπως αναφέρθηκε στο τέλος της προηγούμενης ενότητας, τα δεδομένα προκειμένου να ενσωματωθούν σε ένα εννιαίο πλαίσιο, πρέπει να διορθωθούν με βάση την πλατφόρμα. Επομένως, στη σχέση (2.14) της ενότητας 2.5, ισχύει $i = 3$ και αντιστοιχεί στις τρεις διαφορετικές πλατφόρμες. Διορθώνοντας με βάση τη θεωρία της ενότητας, εξαλείφεται εκείνο το μέρος της έκφρασης των γονιδίων, που οφείλεται καθαρά στην πλατφόρμα. Με τη χρήση της συνάρτησης ComBat του πακέτου sva, και την εισαγωγή σε αυτή ως στοιχείο batch, δηλαδή ομάδα, την πλατφόρμα, επιστρέφεται το διορθωμένο μητρώο $\mathbf{X}_{cor}$, του οποίου οι πρώτες πεντε γραμμές και στήλες φαίνονται παρακάτω.

SYMBOL	GSM162902	GSM162904	GSM162905	GSM162906
A1CF	4.0009	4.0499	4.0508	4.1048
A2M	7.8722	4.5036	9.5826	5.3528
A4GALT	5.4583	5.8542	5.5633	5.6257
A4GNT	4.2495	4.6836	4.4172	4.5510
AAAS	8.7226	8.0385	8.1162	7.7685

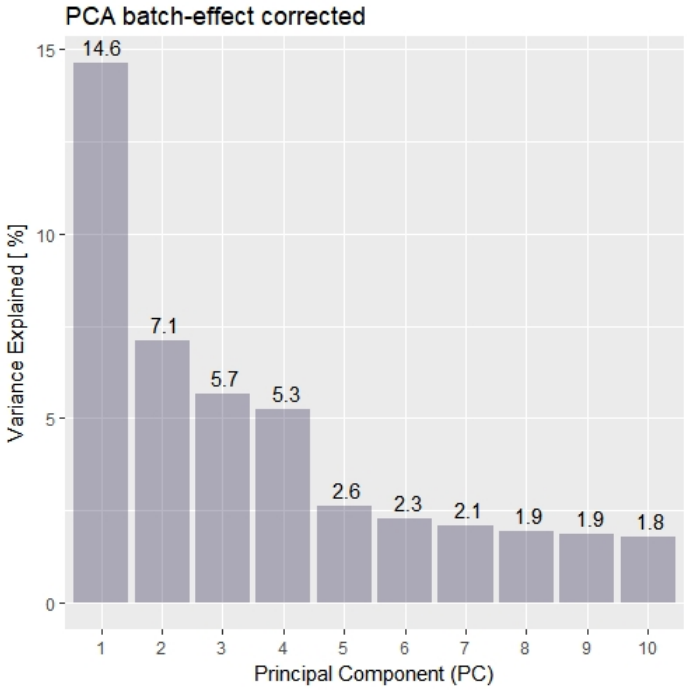
Συγκρίνοντας το $\mathbf{X}$ με το $\mathbf{X_{cor}}$ παρατηρούνται διαφορές στην έκφραση των γονιδίων για τα 4 δείγματα και 5 γονιδια που αναγράφονται. Για να γίνει φανερή η αλλαγή, ξαναεφαρμόζεται ανάλυση κυρίων παραγόντων στα ομαδοποιημένα πλέον δεδομένα. Ανάλογα με το διάγραμμα 3.4 που παρουσιάζει τη μεταβλητότητα που εξηγεί κάθε κύρια συνιστώσα πριν τη διόρθωση των δεδομένων, στο σχήμα 3.8 φαίνεται το ποσοστό διακύμανσης κάθε συνιστώσας μετά τη διόρθωση. Παρατηρείται πως πλέον οι πρώτες συνιστώσες, εξηγούν μικρότερο ποσοστό από ότι στα μη διορθωμένα δεδομένα. Καλύτερα φαίνεται αυτό στο σχήμα 3.9, όπου η ευθεία γραμμή είναι η διαγώνιος, στον άξονα x παρουσιάζεται το ποσοστό επί του 100 της διακύμανσης που εξηγεί κάθε συνιστώσα πριν τη διόρθωση, ενώ στον άξονα y αντίστοιχα μετά τη διόρθωση.

Πέραν της πρώτης κύριας συνιστώσας που πλέον εξηγεί πολύ μικρότερο ποσοστό της διακύμανσης (λιγότερο από το μισό), οι υπόλοιπες δεν έχουν αλλάξει σε μεγάλο βαθμό. Κάτι τέτοιο οφείλεται στο ότι μετά τη διόρθωση έχουμε μικρότερη διακύμανση συνολικά, όπως φαίνεται στους πίνακες 3.3 και 3.4, δηλαδή η απόσταση των δειγμάτων έχει μειωθεί στον χώρο των καινούργιων κύριων συνιστωσών, ενώ παράλληλα η πρώτη συνιστώσα δεν περιγράφει το ίδιο καλά τα δείγματα. Τα αποτελέσματα αυτά είναι αναμενόμενα, καθώς με την ενσωμάτωση των πειραμάτων σε ένα, έχει απαλειφθεί η επίδραση της πλατφόρμας, που όπως φάνηκε στην προηγούμενη ενότητα διαχώριζε σημαντικά τα δείγματα.

Πίνακας 3.3: Διακύμανση που εξηγεί η κάθε συνιστώσα πριν τη διόρθωση

Συνιστώσα	Τυπική απόκλιση προ-διόρθωσης	Διακύμανση προ-διόρθωσης [$\times 100$]
PC1	57.37	33.77
PC2	30.93	9.82
PC3	28.43	8.29
PC4	19.49	3.90
PC5	17.73	3.23
PC6	16.49	2.79

Εκτελώντας εκ νέου ανάλυση κύριων συνιστωσών στα διορθωμένα δεδομένα, προκύπτουν τα σχήματα 3.10 έως 3.12. Απο αυτά είναι φανερό ότι τα δείγματα δε διαχωρίζονται πλέον σε ομάδες με βάση την πλατφόρμα στην οποία υβριδοποιήθηκαν ή το πείραμα στο οποίο ανήκουν. Πολύ σημαντικό είναι το γεγονός ότι τα ομαδοποιημένα δεδομένα δείχνει



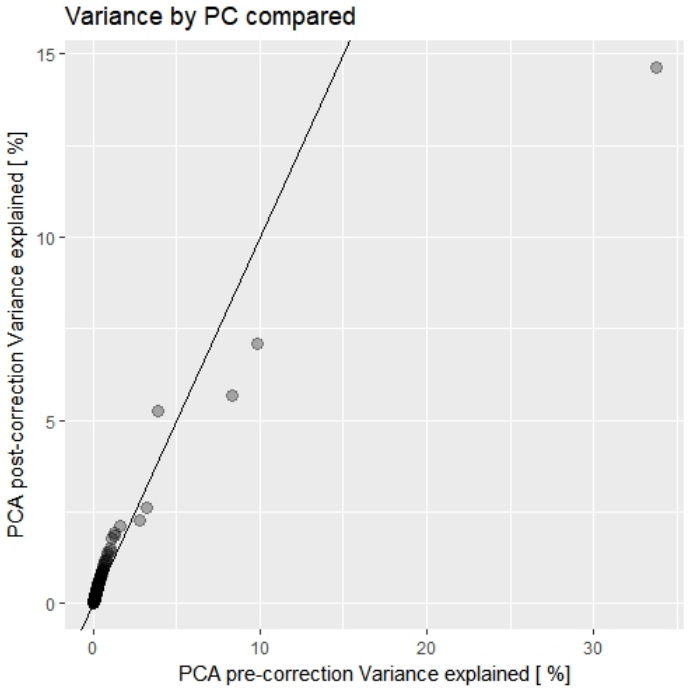
Σχήμα 3.8: Ποσοστό διακύμανσης που εξηγείται από κάθε συνιστώσα μετά τη διόρθωση

Πίνακας 3.4: Διακύμανση που εξηγεί η κάθε συνιστώσα μετά τη διόρθωση

Συνιστώσα	Τυπική απόκλιση μετά-διόρθωσης	Διακύμανση μετά-διόρθωσης [$\times 100$]
PC1	28.03	14.64
PC2	19.51	7.09
PC3	17.46	5.68
PC4	16.80	5.26
PC5	11.86	2.62
PC6	11.03	2.27

στις πρώτες δύο κύριες συνιστώσες να εμφανίζουν και χρήσιμες βιολογικές πληροφορίες, αφού τουλάχιστον τα υγιή κύτταρα και τα απαθανατισμένα, με κίτρινο και μπλε χρώμα αντίστοιχα στο σχήμα 3.10 είναι κοντά το ένα με το άλλο και δείχνει να σχηματίζουν μια γειτονιά. Επίσης, τα καρκινικά κύτταρα καλύπτουν ένα ευρύ φάσμα του χώρου των δύο συνιστωσών, κάτι που θα αποδειχθεί χρήσιμο για την εξειδικευμένη μελέτη.

Η ανάλυση σε κύριες συνιστώσες δεν αρκεί για να επιβεβαιωθεί ότι η διόρθωση με τα Μπεϋζιανά μοντέλα ήταν επιτυχής. Πρέπει να ελεγχθεί ότι δεν χάθηκαν χρήσιμες πληροφορίες κατά τη διόρθωση, δηλαδή ότι απομακρύνθηκε μόνο η επίδραση της πλατφόρμας και όχι η γονιδιακή έκφραση που βρίσκεται από πίσω. Για τον έλεγχο της υπόθεσης γίνεται χρήση γονιδιακής ανάλυσης όπως περιγράφηκε στην ενότητα 2.6, για τον υπολογισμό του μεγέθους $\log(FC)_g$ σε κάθε γονίδιο, για $g = 1, \ldots, 11328$, πριν και μετά τη διόρθωση. Ο λογάριθμος $\log(FC)_g$ δίνει την πιο χρήσιμη πληροφορία της γονιδιακής έκφρασης, αφού ακόμα και αν απόλυτα άλλαξε η έκφραση σε κάποιο δείγμα, αυτό που ενδιαφέρει τον αναλυτή στην κατάντι ανάλυση είναι το αν άλλαξε η διαφορική έκφραση,



Σχήμα 3.9: Συγκριση ποσοστών διακύμανσης κάθε συνιστώσας πριν και μετά τη διόρθωση

δηλαδή η διαφορά της έκφρασης ενός γονιδίου από τα καρκινικά δείγματα στα υγιή. Η σύγκριση γίνεται για δείγματα που προέρχονται από τα ίδια πειράματα, αφού πριν τη διόρθωση δεν είναι εφικτό να συγκριθούν δείγματα διαφορετικών πειραμάτων. Εκτός από τα σημεία που αντιστοιχούν το καθένα σε ένα γονίδιο, έχει γίνει προσέγγιση με μοντέλο GAM (γενικευμένα αθροιστικά μοντέλα) που στην περίπτωσή μας, πλησιάζει ένα απλό γραμμικό μοντέλο καθώς υπάρχει έντονη γραμμική συσχέτιση. Για την ποσοτικοποίηση της συσχέτισης αυτής, και καλύτερο έλεγχο της διόρθωσης, υπολογίζεται κάθε φορά και ο συντελεστής γραμμικής συσχέτισης κατά Pearson:

$$r = \frac{\sum_{i=1}^{G} \left(\log(FC)_{pre_i} - \overline{\log(FC)_{pre}}\right)\left(\log(FC)_{post_i} - \overline{\log(FC)_{post}}\right)}{\sqrt{\sum_{i=1}^{G} \left(\log(FC)_{pre_i} - \overline{\log(FC)_{pre}}\right)^2}\sqrt{\sum_{i=1}^{G} \left(\log(FC)_{post_i} - \overline{\log(FC)_{post}}\right)^2}}$$

Όσο πιο κοντά είναι το r στη μονάδα, τόσο πιο επιτυχημένη κρίνεται η διόρθωση των δεδομένων από το σφάλμα των διαφορετικών πλατφόρμων. Τα αποτελέσματα της παραπάνω ανάλυσης φαίνονται στα σχήματα 3.13 έως 3.17 ενώ σχεδιάζεται και η διαγώνιος, για άμεση σύγκριση με τη συνάρτηση προσέγγισης των δεδομένων.

Παρατηρώντας τους συντελεστές συσχέτισης r, ο μικρότερος έχει τιμή 0.97. Επομένως η ομαδοποίηση κρίνεται επιτυχής, αφού έχουμε σχεδόν γραμμική συσχέτιση. Πλέον, στην κατάντι ανάλυση τα δείγματα μπορούν να θεωρηθούν ότι προέρχονται από την ίδια πλατφόρμα και πείραμα, και μπορούν να συγκριθούν μεταξύ τους. Αυτό σημαίνει ότι τα υγιή κύτταρα (απαθανατισμένα ή όχι) θα χρησιμοποιηθούν ως control δείγματα για να συγκριθούν με τα καρκινικά και να βρεθούν τα σημαντικά γονίδια που είναι διαφορικά εκφρασμένα. Ένας τελευταίος έλεγχος που να το επιβεβαιώνει αυτό είναι ο υπολογισμός της συσχέτισης όταν συγκρίνονται τα γονίδια απο δείγματα ενός πειράματος πριν την ο-

μαδοποίηση με τα γονίδια από δείγματα όλων των πειραμάτων μετά την ομαδοποίηση. Ανάλογα με το σχήμα, όπως αναφέρεται στην αντίστοιχη λεζάντα, τα απαθανατισμένα κύτταρα δε διαφοροποιούνται από τα υγιή, προκειμένου να ελεγθεί αν μπορούν να θεωρηθούν control δείγματα παρά το μεταλλαγμένο γονιδίωμά τους. Έτσι προκύπτουν τα σχήματα 3.18 έως 3.20.

Το τελευταίο βήμα αφού έχει επιτυχώς εκτελεσθεί η ενσωμάτωση των δεδομένων, είναι να ερευνηθεί κατά πόσο υπάρχουν στο χώρο των δύο κύριων συνιστωσών βιολογικές πληροφορίες. Για το σκοπό αυτό, αρχικά παρουσιάζεται το σχήμα 3.21 στο οποίο παρατηρούνται τα εξής:

- Τα υγιή κύτταρα *HEMa-LP 1, HEMa-LP2 (biological replicate), HEM-N, HEM-LP* καθώς και τα απαθανατισμένα υγιή *Hermes1, Hermes 2B* έχουν πολύ μικρή απόσταση δεδομένου του εύρους που απαντάται σε όλο τον χώρο των συνιστωσών. Αυτό είναι αναμενόμενο για τα δύο βιολογικά αντίγραφα HEMa-LP 1, HEMa-LP2 αλλά είναι ικανοποιητικό το γεγονός ότι περιλαμβάνονται στη *γειτονιά* και τα υπόλοιπα υγιή δείγματα.
- Από την κυτταροσειρά *A-375* που αποτελεί πολύ αντιπροσωπευτικό δείγμα μελανώματος και έχει χρησιμοποιηθεί σε πλήθος πειραμάτων και ερευνών/δημοσιεύσεων, υπάρχουν συνολικά τέσσερα δείγματα σε τρία από τα τέσσερα πειράματα (GSE35388, GSE36133, GSE22301). Και τα τέσσερα δείγματα της κυτταροσειράς βρίσκονται κοντά το ένα με το άλλο.
- Οι κυτταροσειρές του τύπου *Hs . . . .T* σχηματίζουν τη δικιά τους γειτονιά στην κάτω αριστερά γωνία του χώρου. Επομένως, αν και απομακρυσμένα από τα υπόλοιπα καρκινικά δείγματα, οι κυτταροσειρές αυτές, μαζί με την D38 ενδεχομένως φέρουν μεταλλάξεις που διαφοροποιούν την ασθένεια και επομένως πρέπει να αναπτυχθεί εξειδικευμένη αντιμετώπιση.

Ο μεγάλος αριθμός δειγμάτων σε συνδυασμό με την πολυπλοκότητα της ασθένειας επιβάλλουν την ομαδοποίηση των δεδομένων, για ευκολότερη και ακριβέστερη γονιδιακή ανάλυση. Η ομαδοποίηση έγινε με ιεραρχικές μεθόδους *hierarchical clustering*: Στην αρχή της διαδικασίας κάθε δείγμα αποτελεί μία ξεχωριστή ομάδα. Σταδιακά, οι ομάδες συνδυάζονται σε μεγαλύτερες, μέχρι να υπάρξει μία μόνο ομάδα. Ο αναλυτής μπορεί επομένως να επιλέξει κατάλληλα, ανάλογα με την πολυπλοκότητα και τη μορφή των ομάδων που επιθυμεί, πόσες ομάδες δειγμάτων θα έχει. Η μέθοδος που χρησιμοποιήθηκε ονομάζεται complete-linkage clustering και μαθηματικά σημαίνει ότι η απόσταση D μεταξύ δύο ομάδων X, Y μετριέται με τη σχέση:

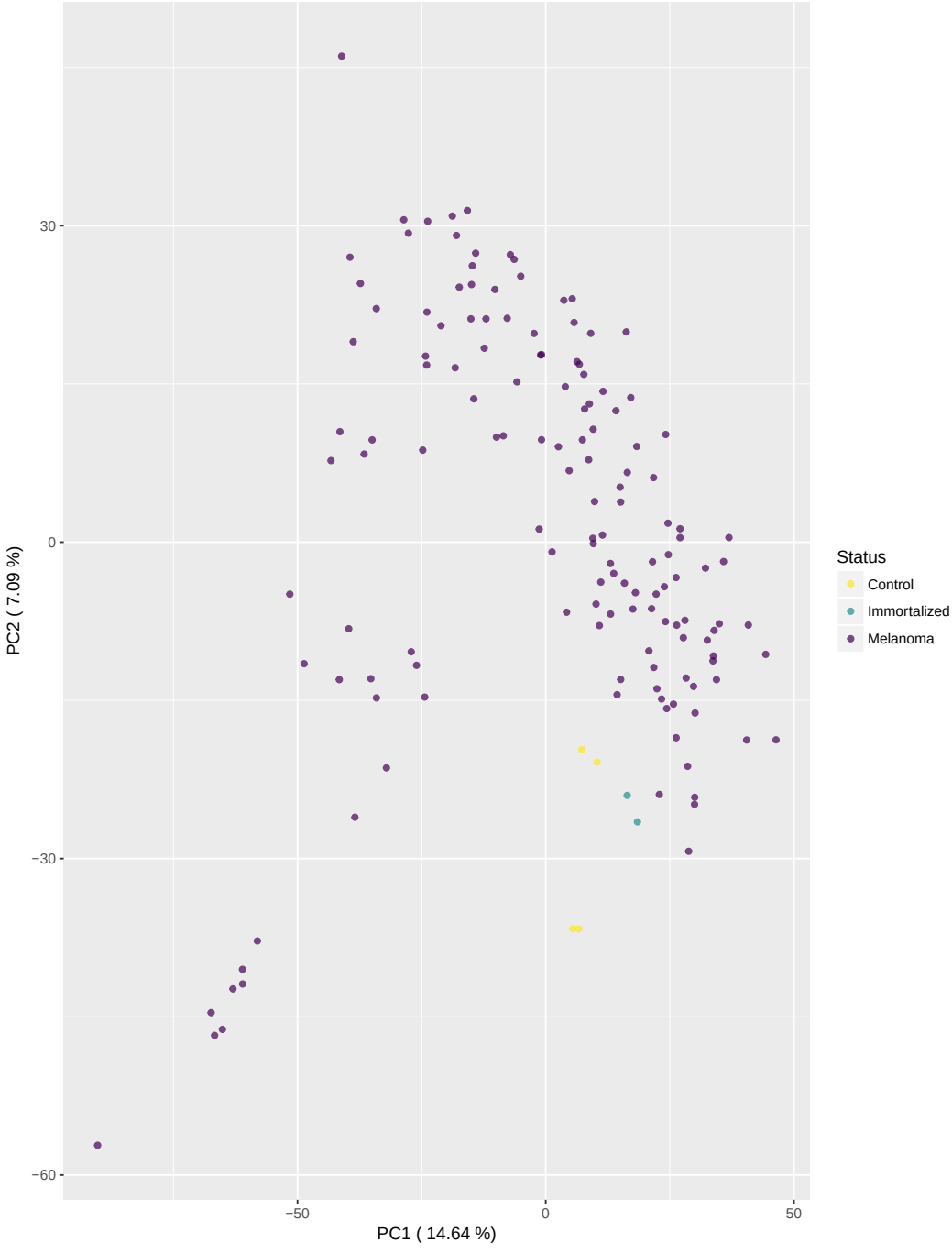
$$D(X, Y) = \max_{x \in X, y \in Y} d(x, y) \tag{3.1}$$

όπου:

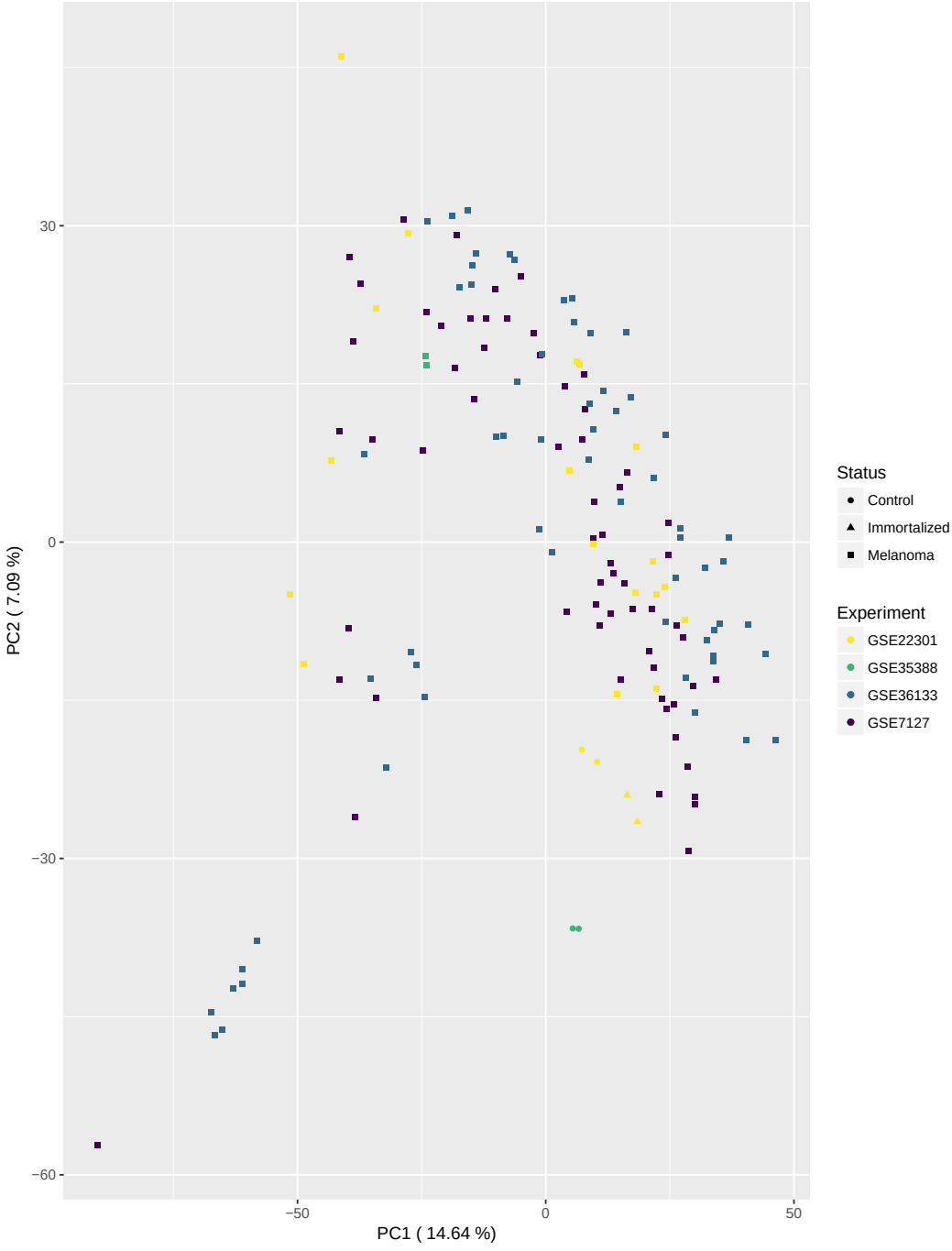
- $d(x, y) = \sqrt{\sum(x_i^2 - y_i^2}$ η απόσταση μεταξύ των στοιχείων $x \in X$ και $y \in Y$. Στην περίπτωση που μελετάμε x_i, y_i είναι η κύρια συνιστώσα i του δείγματος x, y αντίστοιχα. Επειδή η ομαδοποίηση γίνεται στο χώρο των δύο κύριων συνιστωσών, μελετάμε τα δείγματα μόνο σε αυτές τις δύο διαστάσεις, άρα $i = 2$
- X, Y σύνολα στοιχείων (εδώ δειγμάτων)

Τα αποτελέσματα της ομαδοποίησης φαίνονται στο σχήμα 3.22. Παρατηρούνται 7 ομάδες, το πλήθος των οποίων, όπως αναφέρθηκε, επιλέχθηκε για την παρούσα ανάλυση συγκεκριμένα και δεν προέκυψε από κάποιον αλγόριθμο. Αρχικά, εξετάστηκαν οι περιπτώσεις με 5, 7, και 9 ομάδες, η κατάντι ανάλυση όμως έδειξε ότι οι 7 ομάδες εξηγούν ικανοποιητικά την κατανομή των δειγμάτων, με τις 5 ομάδες να περιέχουν η κάθε μία μεγαλύτερη

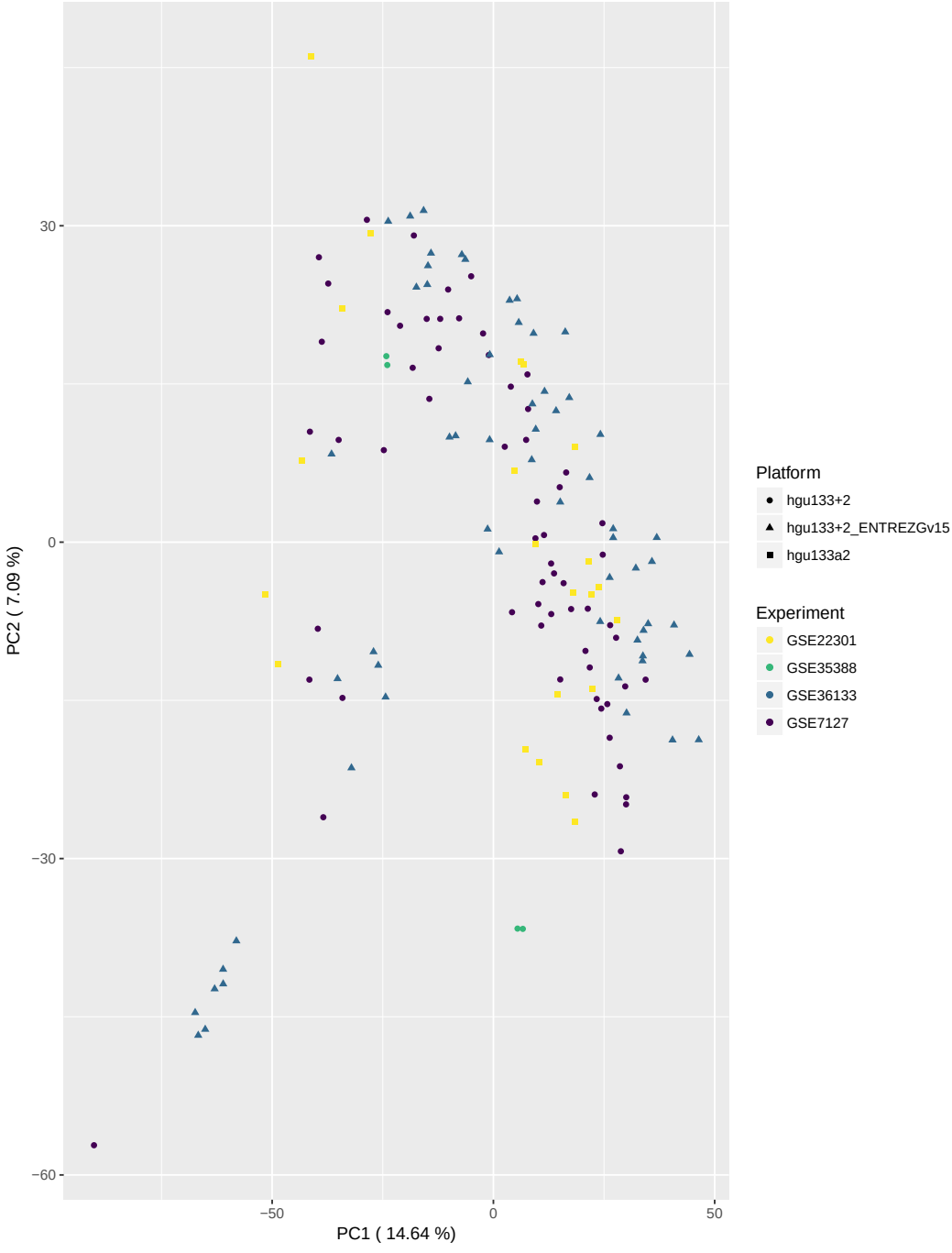
διακύμανση από την επιθυμητή, και τις 9 να δημιουργούν περιττή πολυπλοκότητα. Κύριο στοιχείο που ενθαρρύνει την επιλογή των 7 ομάδων είναι το γεγονός πως τα τέσσερα υγιή κύτταρα αποτελούν μία ομάδα από μόνα τους. Τα απαθανατισμένα κύτταρα, παρόλο που ανήκουν στην ομάδα 5 του σχήματος 3.22 μαζί με κακρινικά κύτταρα, στη γονιδιακή ανάλυση λαμβάνονται υπόψιν ως υγιή.



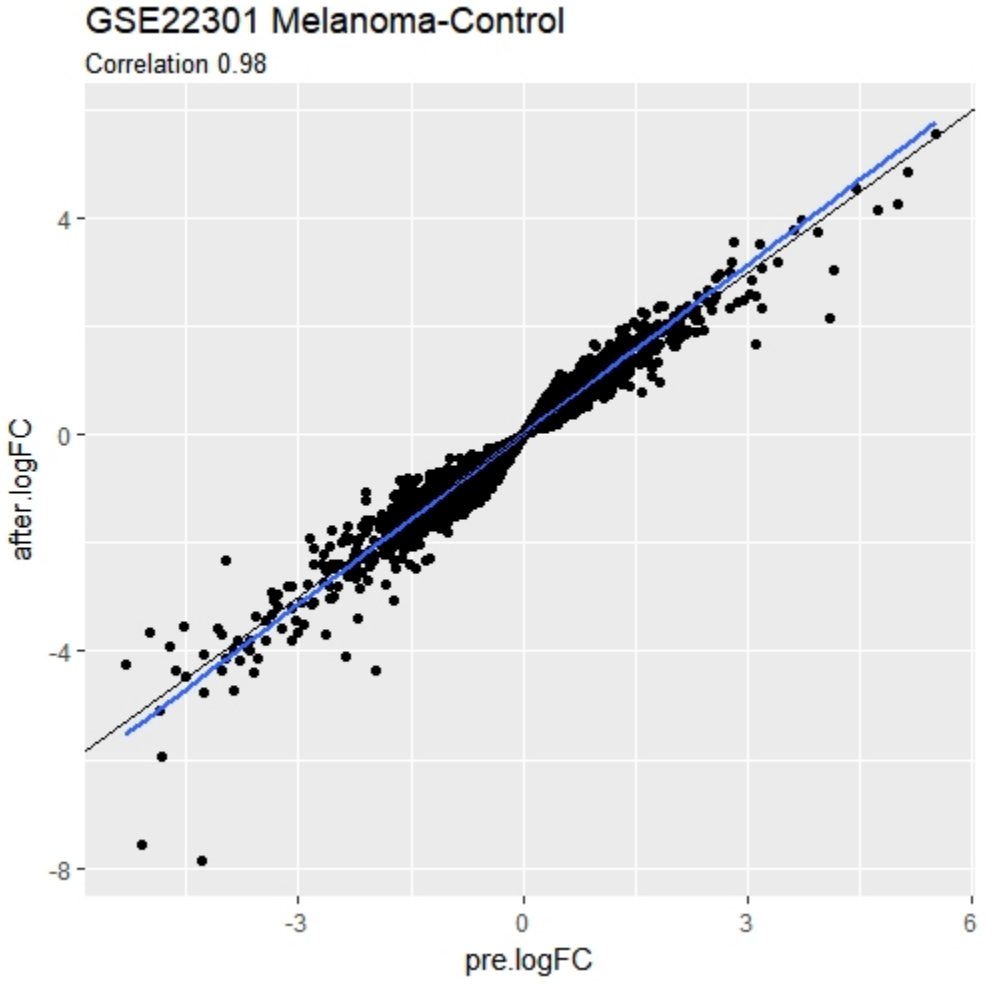
Σχήμα 3.10: Ανάλυση σε κύριες συνιστώσες - τύπος κυττάρου



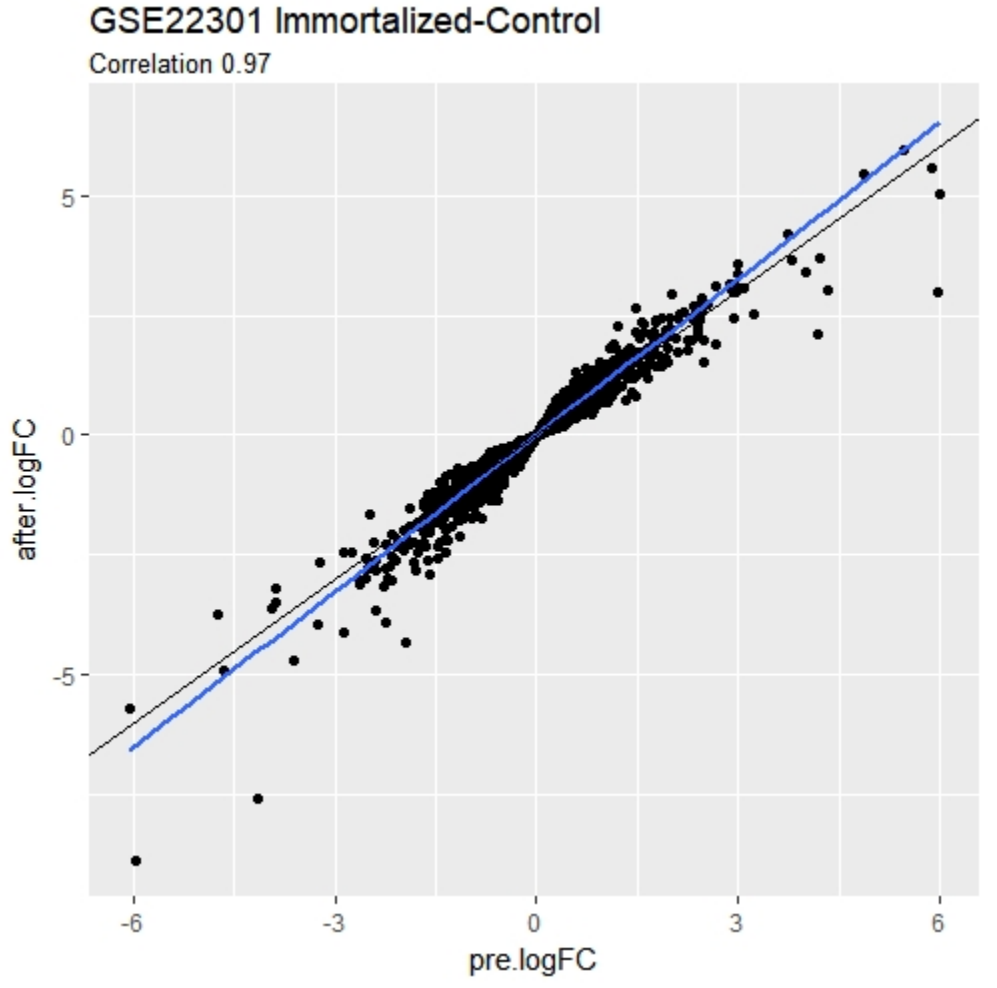
Σχήμα 3.11: Ανάλυση σε κύριες συνιστώσες - τύπος κυττάρου και πείραμα



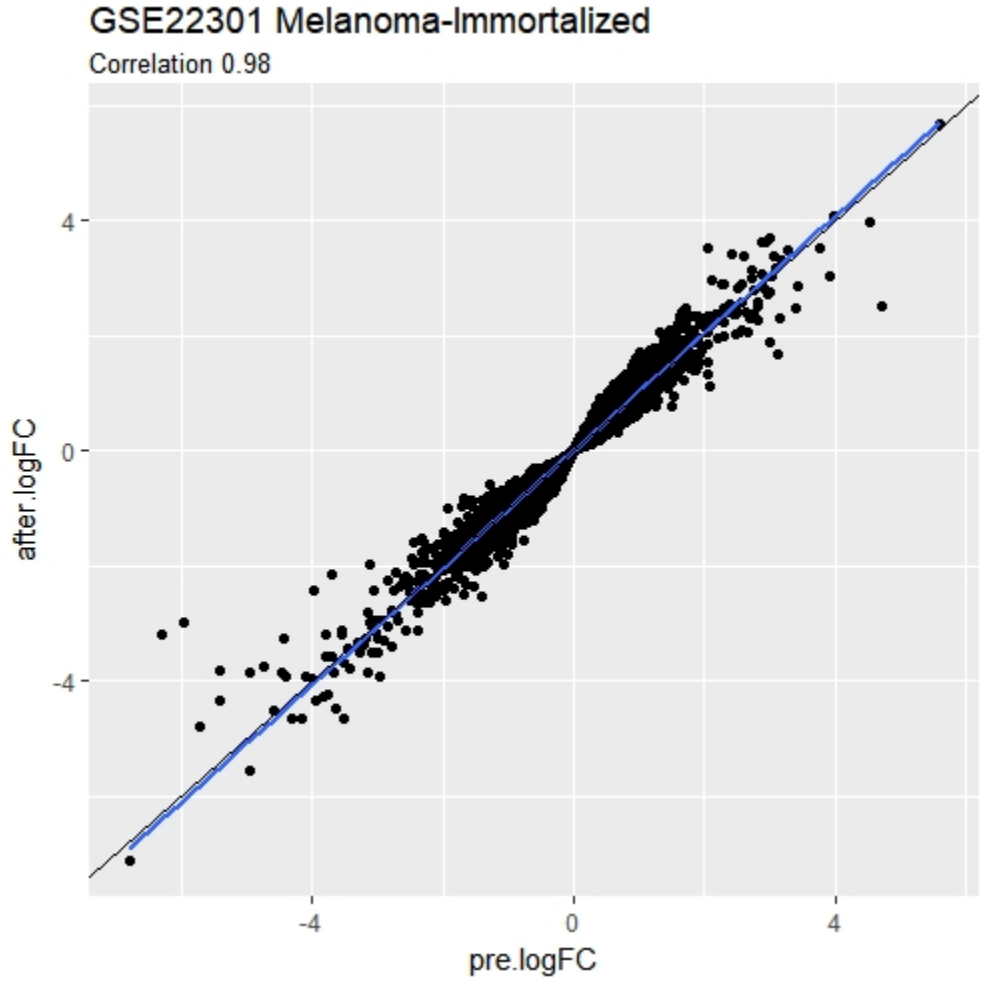
Σχήμα 3.12: Ανάλυση σε κύριες συνιστώσες - τύπος κυττάρου και πλατφόρμα



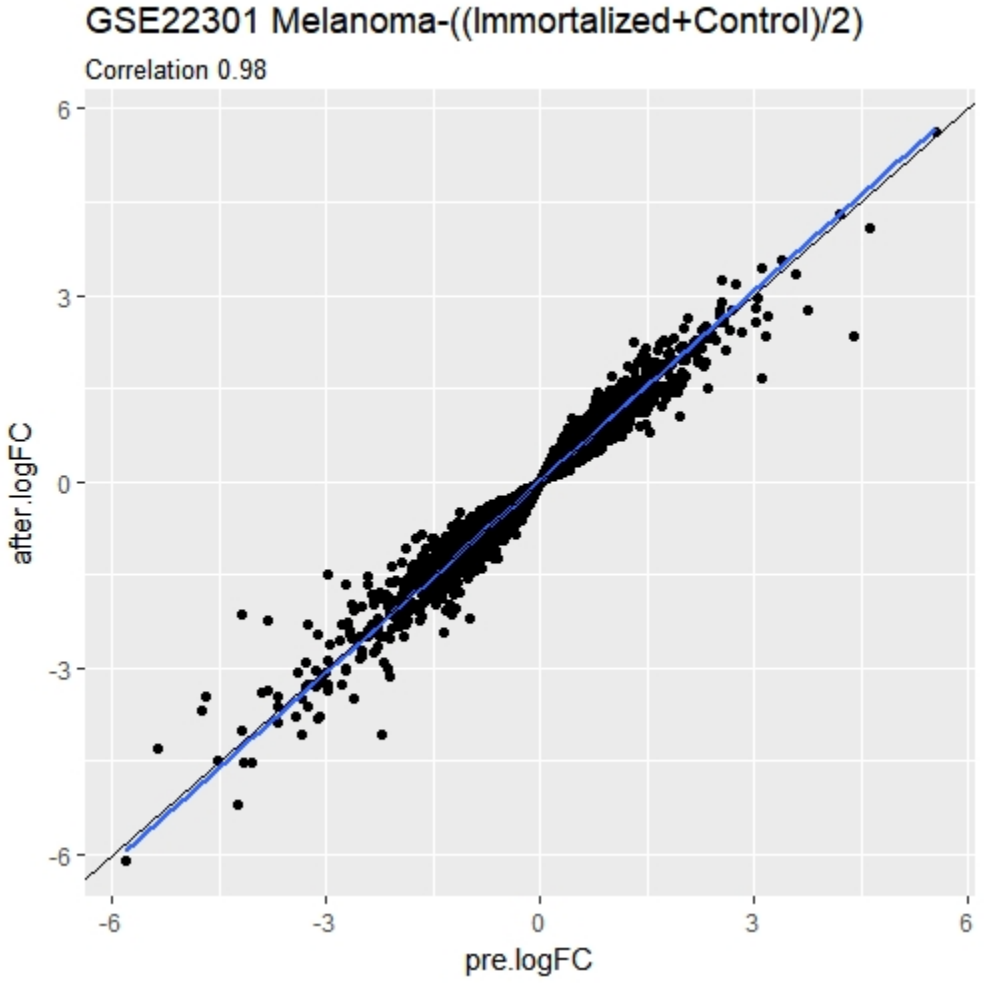
Σχήμα 3.13: Σύγκριση του λογαρίθμου $\log(FC)$ για όλα τα γονίδια για τα καρκινικά κύτταρα έναντι των υγιών πριν τη Μπεϋζιανή διόρθωση (άξονας *x*) και μετά (άξονας *y*). Συντελεστής συσχέτισης κατά Pearson 0.98.



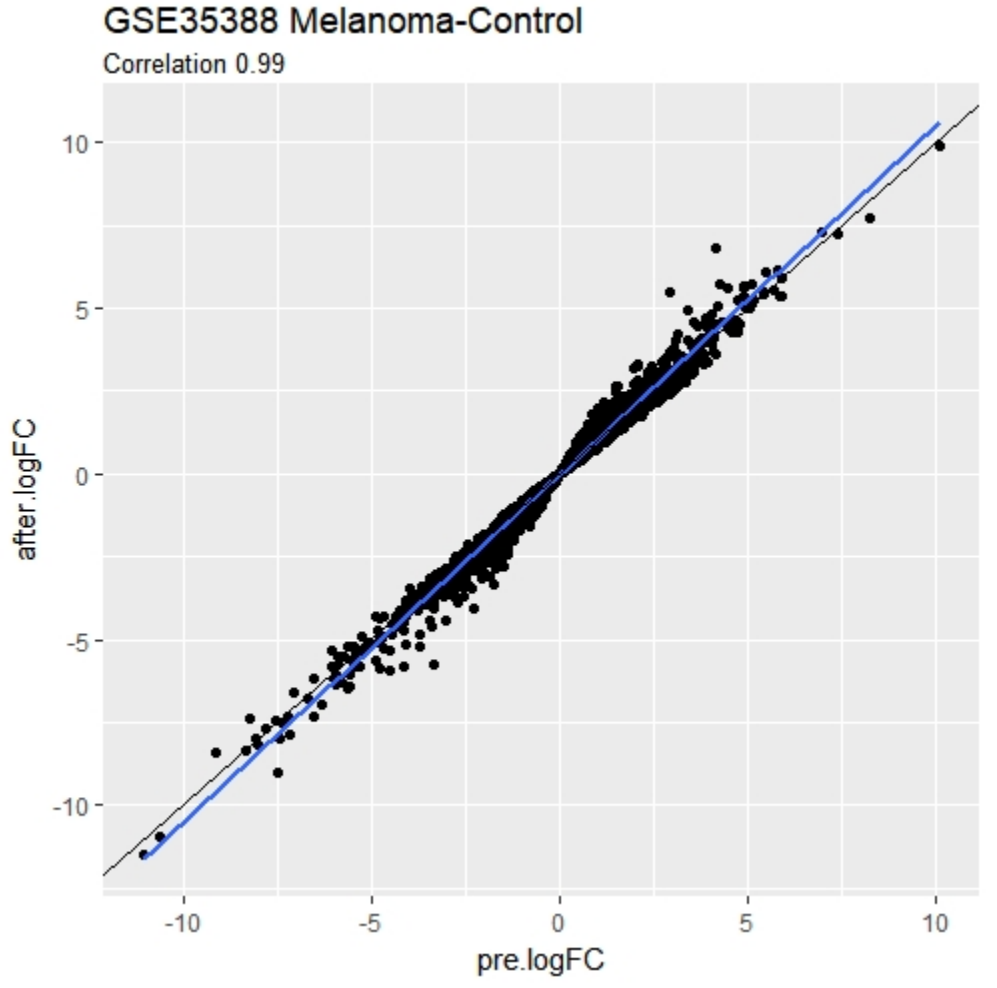
Σχήμα 3.14: Σύγκριση του λογαρίθμου $\log(FC)$ για όλα τα γονίδια για τα απαθανατισμένα κύτταρα έναντι των υγιών πριν τη Μπεϋζιανή διόρθωση (άξονας x) και μετά (άξονας y). Συντελεστής συσχέτισης κατά Pearson 0.97.



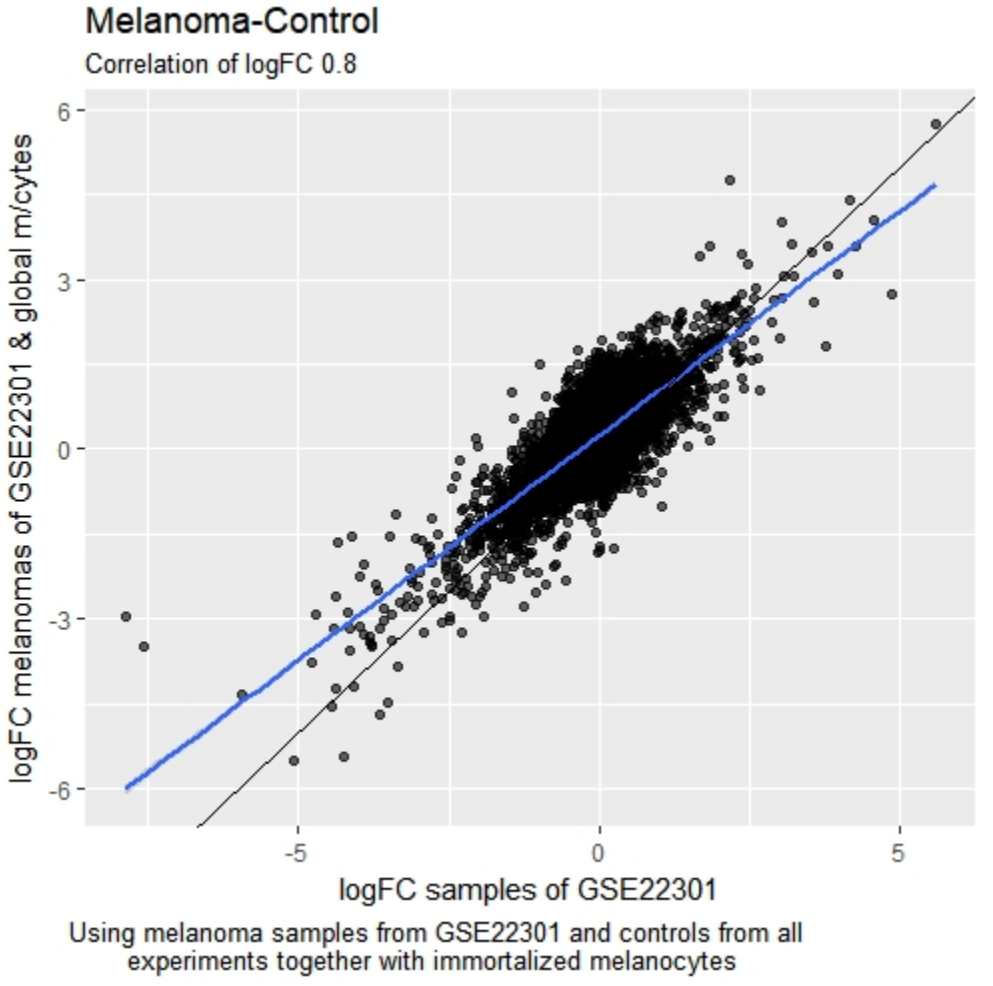
Σχήμα 3.15: Σύγκριση του λογαρίθμου $\log(FC)$ για όλα τα γονίδια για τα απαθανατισμένα κύτταρα έναντι των καρκινικών πριν τη Μπεϋζιανή διόρθωση (άξονας x) και μετά (άξονας y). Συντελεστής συσχέτισης κατά Pearson 0.98.



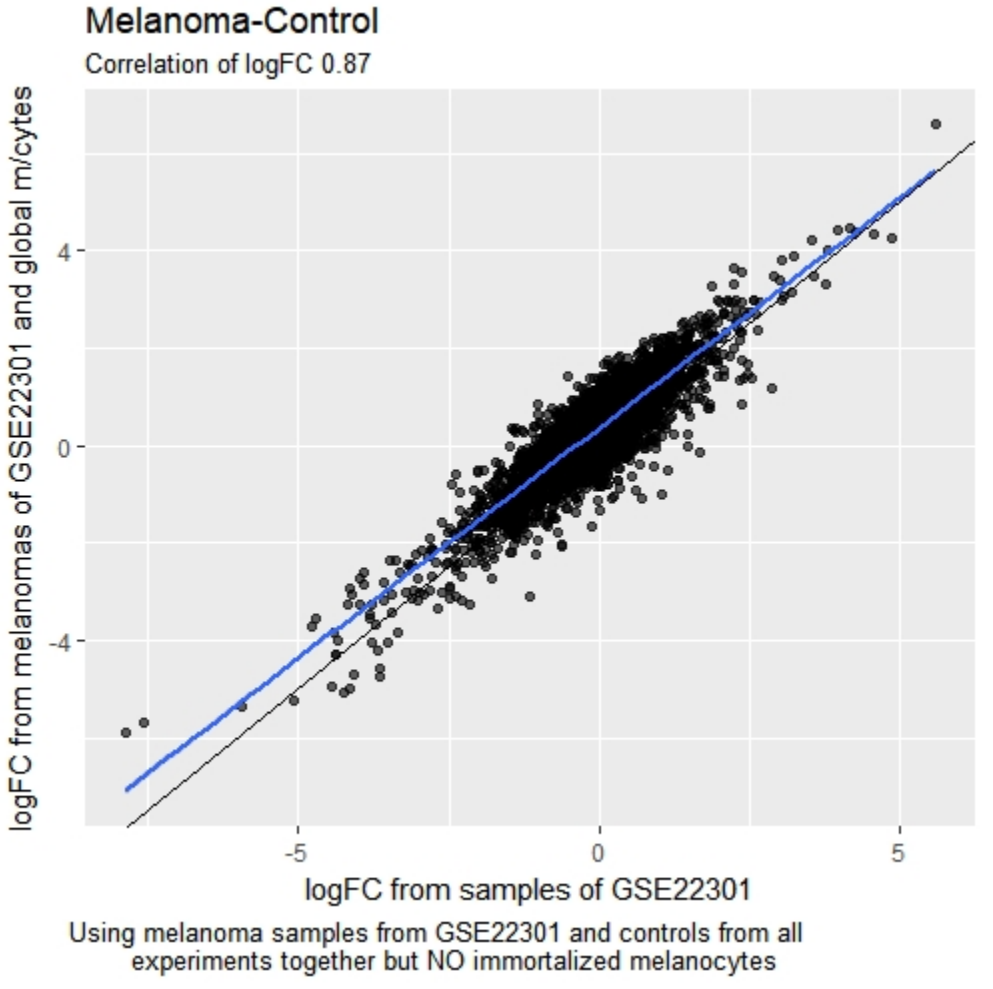
Σχήμα 3.16: Σύγκριση του λογαρίθμου $\log(FC)$ για όλα τα γονίδια για τα απαθανατισμένα και τα υγιή κύτταρα έναντι των καρκινικών πριν τη Μπεϋζιανή διόρθωση (άξονας x) και μετά (άξονας y). Συντελεστής συσχέτισης κατά Pearson 0.98.



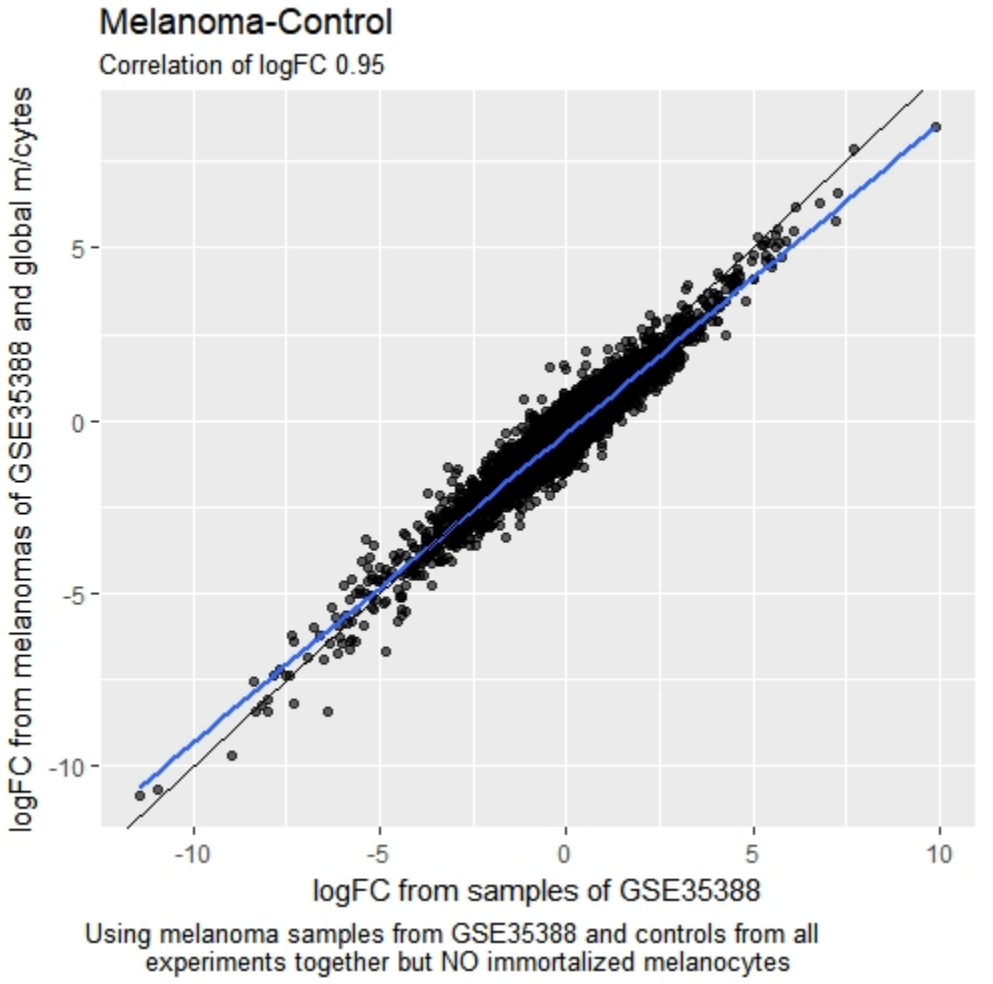
Σχήμα 3.17: Σύγκριση του λογαρίθμου $\log(FC)$ για όλα τα γονίδια για τα υγιή κύτταρα έναντι των καρκινικών πριν τη Μπεϋζιανή διόρθωση (άξονας x) και μετά (άξονας y). Συντελεστής συσχέτισης κατά Pearson 0.99.



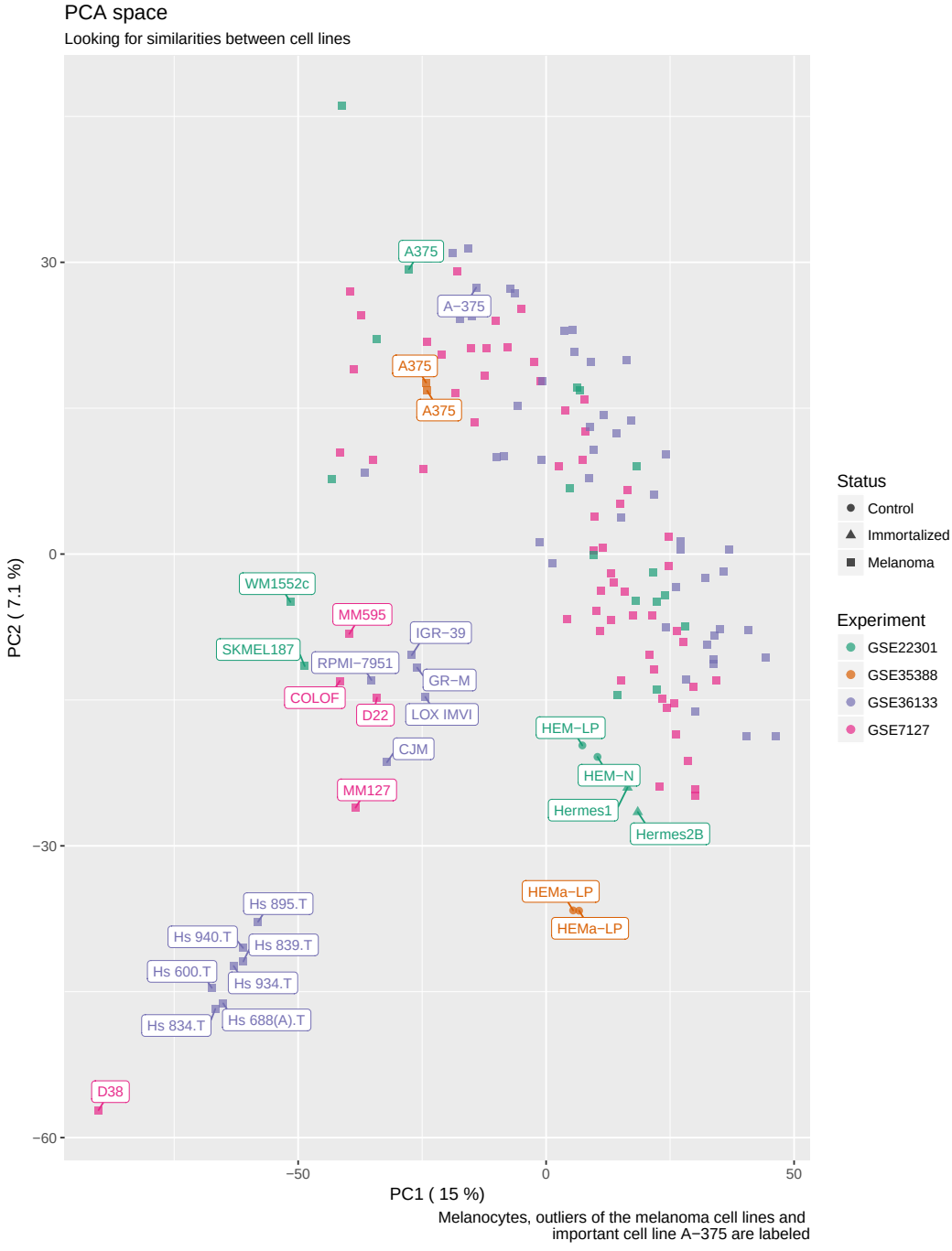
Σχήμα 3.18: Σύγκριση του λογαρίθμου $\log(FC)$ για όλα τα γονίδια για τα υγιή κύτταρα έναντι των καρκινικών στο πείραμα GSE22301 πριν τη Μπεϋζιανή διόρθωση (άξονας x) και για όλα τα γονίδια για τα υγιή (**μαζί με τα απαθανατισμένα**) κύτταρα έναντι των καρκινικών από όλα τα πειράματα μετά τη διόρθωση (άξονας y). Συντελεστής συσχέτισης κατά Pearson 0.8. Προφανώς η γραμμική συσχέτιση έχει μειωθεί καθώς πλέον η έκφραση των γονιδίων για τα καρκινικά κύτταρα έχει πολύ μεγαλύτερη διακύμανση, αφού συμπεριλαμβάνονται στη σύγκριση και τα 144 καρκινικά δείγματα έναντι των 18 που περιλαμβάνει το GSE22301. Επίσης, δεν έχουν διαφοροποιηθεί τα υγιή από τα απαθανατισμένα κύτταρα, καθώς είναι επιθυμητό αυτά τα δείγματα να αντιμετωπιστούν ως control απέναντι στις σειρές του μελανώματος. Η συσχέτιση 0.80 είναι ικανοποιητική με βάση τα πρότυπα των αναλύσεων γονιδιακής έκφρασης.



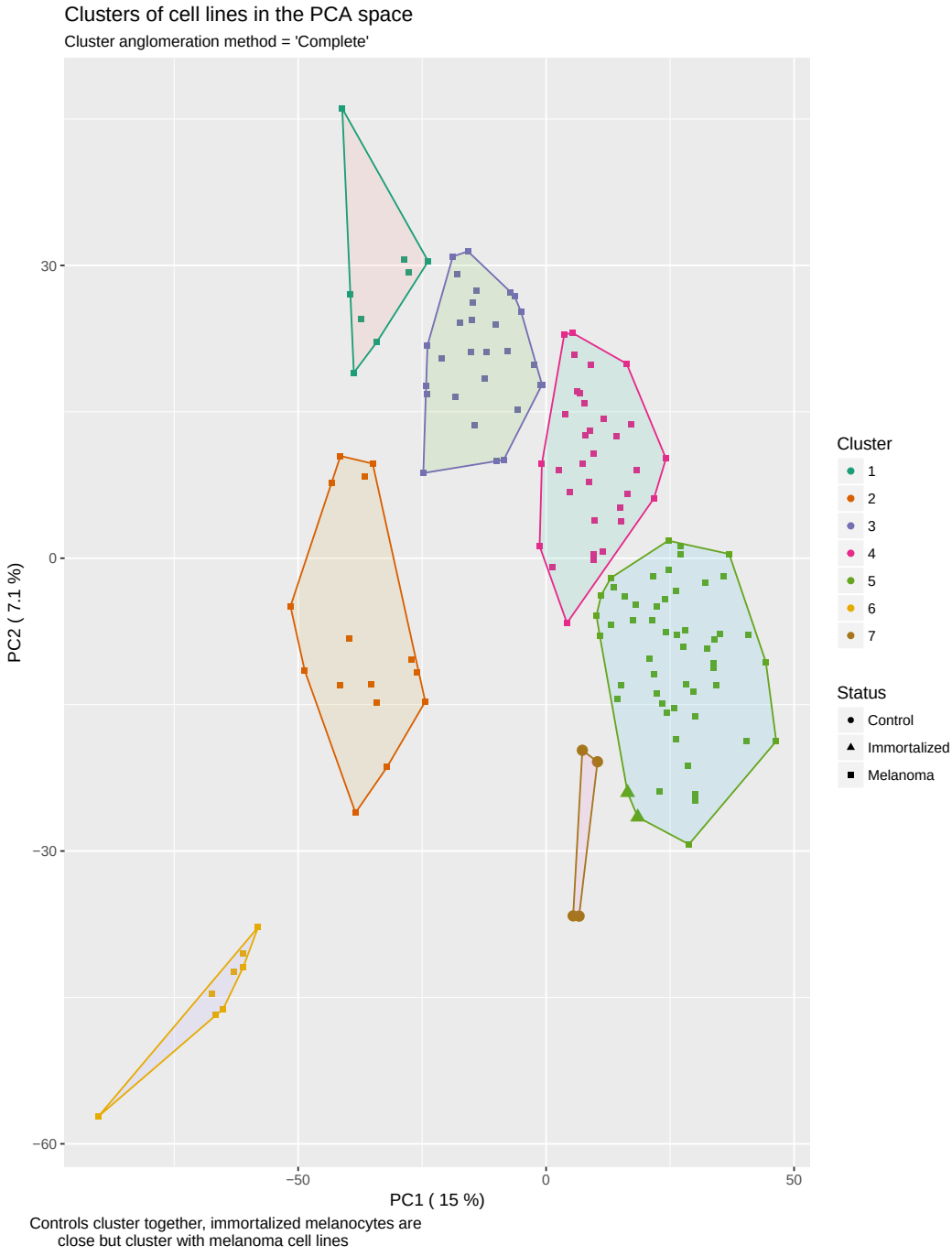
Σχήμα 3.19: Σύγκριση του λογαρίθμου $\log(FC)$ για όλα τα γονίδια για τα υγιή κύτταρα έναντι των καρκινικών στο πείραμα GSE22301 πριν τη Μπεϋζιανή διόρθωση (άξονας x) και για όλα τα γονίδια για τα υγιή (**χωρίς τα απαθανατισμένα**) κύτταρα έναντι των καρκινικών από όλα τα πειράματα μετά τη διόρθωση (άξονας y). Τα αποτελέσματα δε διαφέρουν σημαντικά από το προηγούμενο σχήμα 3.18, πέρα από την αυξημένη γραμμική συσχέτιση καθώς εδώ δεν υπολογίζονται τα απαθανατισμένα κύτταρα. Η μικρή διαφορά (7%) δεν κρίνεται αρκετή για να απορριφθούν τα δύο επιπλέον δείγματα ως control δεδομένης της εγγύτητας που παρουσιάζουν με τα υγιή στον χώρο των κύριων συνιστωσών, και της επιπλέον στατιστικής δύναμης που προσφέρουν στην κατάντι ανάλυση.



Σχήμα 3.20: Σύγκριση του λογαρίθμου $\log(FC)$ για όλα τα γονίδια για τα υγιή κύτταρα έναντι των καρκινικών στο πείραμα GSE35388 πριν τη Μπεϋζιανή διόρθωση (άξονας x) και για όλα τα γονίδια για τα υγιή (**χωρίς τα απαθανατισμένα**) κύτταρα έναντι των καρκινικών από όλα τα πειράματα μετά τη διόρθωση (άξονας y). Εδώ η γραμμική συσχέτιση είναι 95% ακόμα και περιλαμβανομένων των δειγμάτων από όλα τα πειράματα στον κατακόρυφο άξονα.



Σχήμα 3.21: Ο χώρος που δημιουργείται από τις πρώτες δύο κύριες συνιστώσες. Τα υγιή κύτταρα *HEMa-LP 1, HEMa-LP2 (biological replicate), HEM-N, HEM-LP* καθώς και τα απαθανατισμένα υγιή *Hermes1, Hermes 2B* έχουν πολύ μικρή απόσταση δεδομένου του έυρους που απαντάται σε όλο τον χώρο. Επίσης, από την κυτταροσειρά *A-375* που αποτελεί πολύ αντιπροσωπευτικό δείγμα μελανώματος και έχει χρησιμοποιηθεί σε πλήθος πειραμάτων και ερευνών/δημοσιεύσεων, υπάρχουν συνολικά τέσσερα δείγματα σε τρία από τα τέσσερα πειράματα (GSE35388, GSE36133, GSE22301). Και τα τέσσερα δείγματα της κυτταροσειράς βρίσκονται κοντά το ένα με το άλλο. Οι κυτταροσειρές του τύπου *Hs ....T* σχηματίζουν τη δικιά τους γειτονιά στην κάτω αριστερά γωνία του χώρου.



Σχήμα 3.22: Τα δείγματα κατανέμονται ανάμεσα σε 7 ομάδες, όπως φάινεται στο σχήμα. Από αυτές, η μία (ομάδα 7) αποτελείται από τις τέσσερεις υγιής κυτταροσειρές, ενώ όλες οι υπόλοιπες είναι ομάδες αμιγώς καρκινικών κυττάρων με εξαίρεση τις δύο απαθανατισμένες υγιής κύτταρικές σειρές, που ανήκουν στην καρκινική ομάδα 5.

3.3 Ανάλυση διαφορικής έκφρασης

3.3.1 Καρκινικά δείγματα έναντι υγιών δειγμάτων

Με την ομαδοποίηση των δειγμάτων στον χώρο των δύο πρώτων κύριων συνιστωσών, η ανάλυση μπορεί να προχωρήσει σε επόμενο επίπεδο, το οποίο είναι η ανάλυση της διαφορικής έκφρασης των γονιδίων. Το πρώτο *πρόβλημα* που παρουσιάζουν τα διαθέσιμα δεδομένα είναι η μεγάλη ανισορροπία που υπάρχει ανάμεσα στους δύο τύπους κυττάροσειρών: υπάρχον μόνο 6 υγιή δείγματα έναντι 144 καρκινικών δειγμάτων, δηλαδή η διαφορά είναι δύο τάξης μεγέθους παραπάνω. Κάτι τέτοιο καλεί για εξισορρόπηση των δειγμάτων, προκειμένου να έχουμε ακριβείς τιμές t που προκύπτουν από τα t-tests, καθώς αυξημένη διαφορά ανάμεσα στον αριθμό δειγμάτων και στην πολλαπλάσια τυπική απόκλιση που αναγκαστικά παρουσιάζει ο πληθυσμός καρκινικών κυττάρων έχει ως αποτέλεσμα την παραμόρφωση του στατιστικού μεγέθους, και επομένως οι p τιμές που θα υπολογίζονταν με τα γραμμικά μοντέλα δε θα ήταν αξιόπιστες και θα οδηγούσαν σε μειωμένο αριθμό διαφορικά εκφρασμένων γονιδίων. Ο αναλυτής σε αυτό το στάδιο επιθυμεί να καταλήξει με όσο περισσότερα διαφορικά εκφρασμένα γονίδια είναι στατιστικά εφικτό, προκειμένου να υπάρχει υλικό στην κατάντι ανάλυση.
Για τους παραπάνω λόγους, κατασκευάστηκε αλγόριθμος ο οποίος διαλέγει από τις δύο ομάδες που επιλέγονται 6 δείγματα. Τα δείγματα επιλέχθηκε να είναι 6 καθώς υπάρχουν 6 υγιή δείγματα, οπότε στην περίπτωση των υγιών δειγμάτων λαμβάνονται κάθε φορά τα ίδια 6. Επίσης, για να υπάρξει στοχευμένη ανάλυση, επιπλέον της περίπτωσης που τα 6 καρκινικά δείγματα λαμβάνονται από όλες τις 144 κυτταροσειρές, έγιναν άλλες 6 αναλύσεις όπου κάθε φορά τα καρκινικά δείγματα λαμβάνονται απο κάθε μία από τις 6 ομάδες του σχήματος 3.22 της προηγούμενης ενότητας. Η διαδικασία του resampling επαναλαμβάνεται από 1000 φορές σε κάθε περίπτωση, προκειμένου να καλυφθεί ικανοποιητικός αριθμός διαφορετικών πληθυσμών και διακύμανσης.
Σε κάθε τρέξιμο, υπολογίζονται και αποθηκεύονται τα γονίδια που είναι διαφορικά εκφρασμένα στα καρκινικά δείγματα από ότι στα υγιή. Ως συνθήκη για να χαρακτηριστεί κάποιο γονίδιο διαφορικά εκφρασμένο είναι η:

$$|\log(FC)| > 1 \qquad p_{value_{BH-corrected}} < 0.05 \tag{3.2}$$

Τα γονίδια που ικανοποιούν την παραπάνω συνθήκη επιστρέφονται σε μορφή πίνακα που μοιάζει με τον παρακάτω:

Gene	logFC	t	adj.P.Val	MelCells	Set
RAB33A	-4.6	-10.1	0.0014	39-54-82-132-29-128	From all
SYK	-1.6	-8.6	0.0026	39-54-82-132-29-128	From all
CD36	-3.7	-8.6	0.0026	39-54-82-132-29-128	From all
MICAL1	-2.0	-8.5	0.0026	39-54-82-132-29-128	From all
HMGA2	2.3	7.4	0.0097	39-54-82-132-29-128	From all
ZNF365	1.7	7.1	0.0123	39-54-82-132-29-128	From all

Στην πρώτη στήλη αναφέρεται το όνομα του γονιδίου και ακολουθεί το $\log(FC)$. Αντί της κανονικής τιμής p αναγράφεται η προσαρμοσμένη κατά Benjamini-Hochberg τιμή p που δίνει πιο αξιόπιστα αποτελέσματα, και είναι αναγκαίο να χρησιμοποιηθεί ιδιαίτερα όταν στην κατάντι μελέτη γίνει ανάλυση μονοπατιών, όπου συνδυάζονται τιμές p γονιδίων που ανήκουν στο μονοπάτι για να υπολογιστεί τιμή p του μονοπατιού. Τέλος, σε περίπτωση που κάποιος θέλει να ξανατρέξει τις προσομοιώσεις, δίνονται τα επιλεγμένα καρκινικά δείγματα στη στήλη MelCells, όπου ο δείκτης αντιστοιχεί στη γραμμή του δείγματος στο εννιαίο ομαδοποιημένο μητρώο $\mathbf{X}_{corr}$, καθώς και από ποια ομάδα (set) προέρχονται αυτά

τα δείγματα. Παρακάτω φαίνεται ένα διάγραμμα ηφαιστείου για ένα από τα 1000 τρεξίματα. Τα ανω-εκφρασμένα γονίδια έχουν μπλε χρώμα, ενώ με κόκκινο συμβολίζονται τα κάτω-εκφρασμένα, και με πράσινο τα γονίδια που δεν είναι διαφορικά εκφρασμένα.

Στο παράδειγμα παραπάνω, το From all αντιστοιχεί σε sampling από όλα τα καρκινικά δείγματα. Αντίστοιχα υπάρχουν τα From 1, From 2, . . . , From 6. Το τελικό αποτέλεσμα είναι ένας πίνακας με 4688609 γραμμές, στον οποίο περιλαμβάνονται όλα τα διαφορικά εκφρασμένα γονίδια από τις 1000 προσομοιώσεις. Είναι αναγκαίο να επεξεργαστεί κατάλληλα ώστε να αντληθούν περαιτέρω πληροφορίες.
Το πρώτο βήμα σε αυτή την επεξεργασία είναι η κατασκευή ενός σύντομου αλγορίθμου ο οποίος μετράει πόσα γονίδια έχουν εμφανιστεί περισσότερο από n φορές. Προφανώς ισχύει $n = 1, \ldots, 1000$. Οπτικοποιώντας το αποτέλεσμα προκύπτει το σχήμα 3.24. Όπως είναι αναμμενόμενο, στο πρώτο τέταρτο του οριζοντίου άξονα περίπου παρατηρείται αυξημένη κλίση της καμπύλης πυκνότητας σε όλες τις ομάδες. Προφανώς, η κλίση που παρατηρείται σε κάθε ομάδα σχετίζεται άμεσα με τη διακύμανση του πληθυσμού των καρκινικών δειγμάτων κάθε ομάδας. Δηλαδή, όσο πιο μεγάλη διακύμανση έχει ο πληθυσμός που εξετάζεται, τόσο πιο λίγα θα είναι τα γονίδια που βρίσκονται διαφορικά εκφρασμένα. Στην πρώτη ομαδοιποίηση το resampling γίνεται από όλα τα 144 καρκινικά κύτταρα, και είναι εμφανές ότι λόγω της μεγάλης διακύμανσης, δεν υπάρχουν καν γονίδια τα οποία να είναι διαφορικά εκφρασμένα σε περισσότερες από ~ 800 επαναλήψεις. Αντίθετα, σε μικρότερες ομάδες όπως η ομάδα 6, όπου υπάρχει μικρότερη διακύμανση όπως φαίνεται και στο σχήμα 3.22, υπάρχουν σχεδόν 1000 γονίδια που βρίσκονται διαφορικά εκφρασμένα σε όλες τις επαναλήψεις.
Σε αυτό το στάδιο, για να προχωρήσει η ανάλυση, πρέπει να επιλεχθεί ο αριθμός n των επαναλήψεων που απαιτείται για να χαρακτηριστεί ένα γονίδιο διαφορικά εκφρασμένο στην αντίστοιχη ομάδα. Αν το γονίδιο είναι διαφορικά εκφρασμένο σε παραπάνω από n επαναλήψεις, τότε το αποτέλεσμα είναι στατιστικά σημαντικό. Εκτελείται ανάλυση ευαισθησίας, όπου επιλέγονται τρεις τιμές και προκύπτουν αντίστοιχα τρεις συνθήκες:

- $n_l = 260$ χαλαρό όριο (lax threshold)
- $n_m = 540$ μεσαίο όριο (medium threshold)
- $n_s = 800$ αυστηρό όριο (strict threshold)

Η κατάντι ανάλυση θα συνεχιστεί ξεχωριστά για κάθε μία από τις τρεις συνθήκες. Εφαρμόζωντάς τες, προκύπτει το σχήμα 3.25. Η μεγάλη διαφορά ανάμεσα στην πρώτη περίπτωση From all και στις υπόλοιπες, δικαιολογείται αν αναλογιστεί κάποιος τη μορφή του κριτηρίου Student t για δείγματα άνισων πληθυσμών:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{\Delta}}}$$

$$s_{\overline{\Delta}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Εύκολα παρατηρείται πως όταν υπάρχει μεγάλει διακύμασνη s_i^2 έστω και σε ένα από τα δείγματα, το μέγεθος t μειώνεται και επομένως η αντίστοιχη τιμή p αυξάνεται, επομένως υπάρχουν λιγότερα γονίδια που ικανοποιούν τη συνθήκη (3.2).

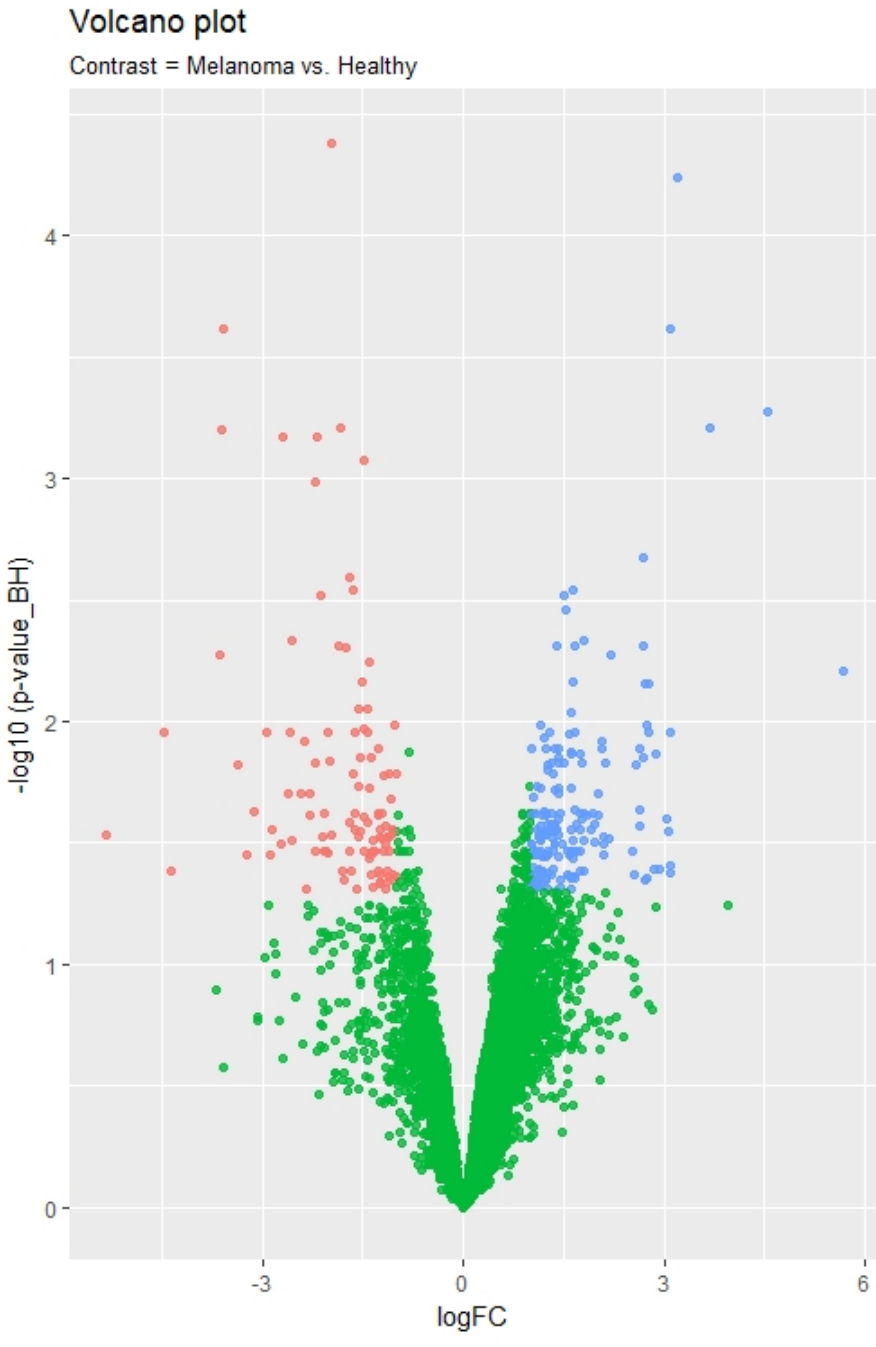
Το επόμενο βήμα αφού έχουν επιλεχθεί τα διαφορικά εκφρασμένα γονίδια από κάθε ομάδα, είναι να φιλτραριστούν κατάλληλα, ώστε να προκύψουν γονίδια που μπορούν να χρησιμοποιηθούν στην κατηγοριοποίηση των κυτταροσειρών της ενότητας 3.4. Πρέπει να καταλήξουμε με όσο το δυνατό λιγότερα γονίδια, και ταυτόχρονα καλή ακρίβεια του αλγορίθμου κατηγοριοποίησης. Διακρίνουμε τρεις διαφορετικές περιπτώσεις για την επιλογή των τελικών γονιδίων:

Περίπτωση 1 Τα γονίδια που ικανοποιούν τη συνθήκη ορίου **και είναι διαφορικά εκφρασμένα στις 6 ομάδες resampling και σε ολόκληρο τον πληθυσμό**. Η πιο αυστηρή από τις τρεις περιπτώσεις, γίνεται πρακτικά διπλός έλεγχος, αφού είναι επιθυμητό τα γονίδια να έχουν επιλεγεί στις κυτταροσειρές κάθε μικρότερης ομάδας, αλλά και από το συνολικό πληθυσμό. Περιμένουμε να απομένουν τα λιγότερα γονίδια, σε σχέση με τις άλλες περιπτώσεις.

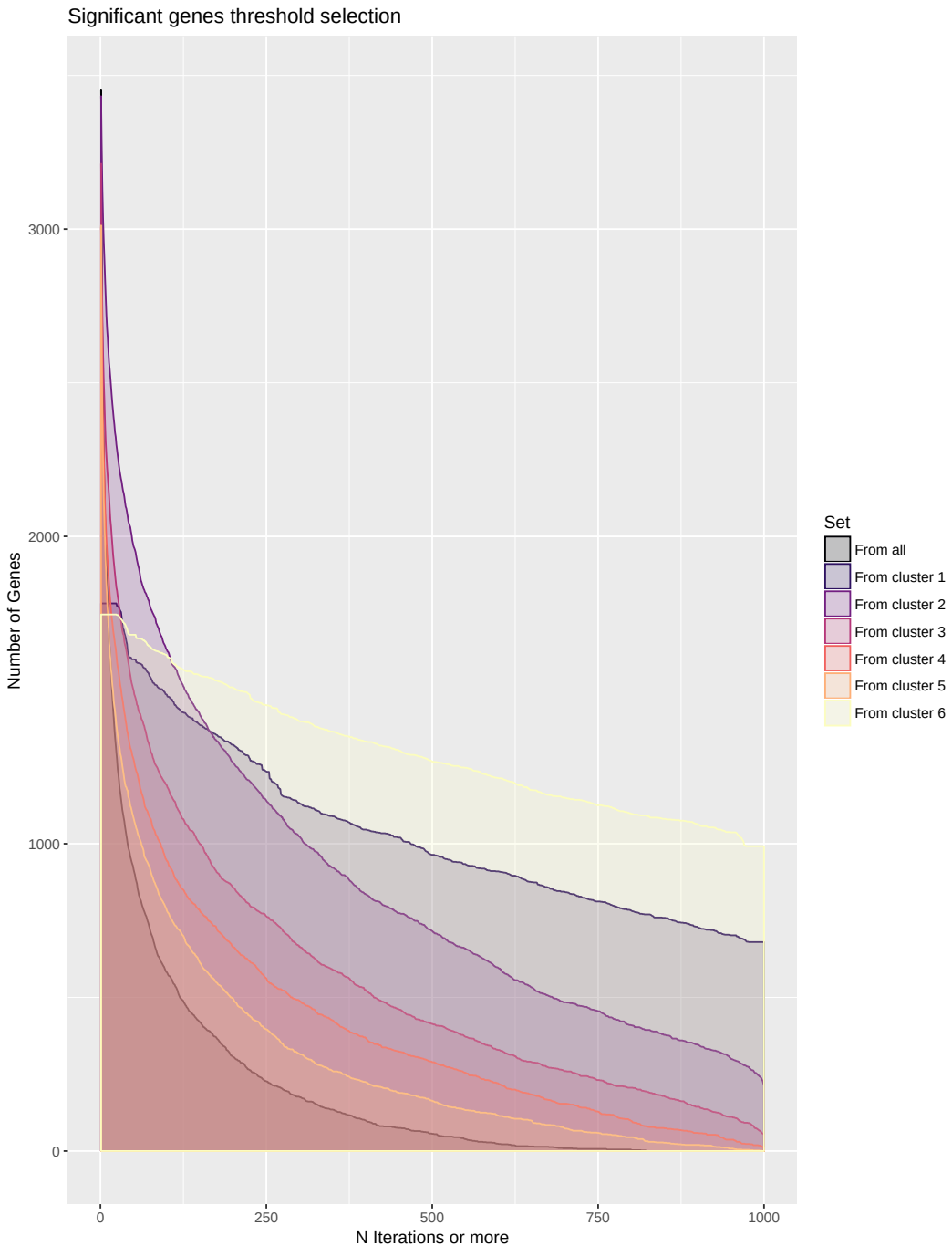
Περίπτωση 2 Τα γονίδια που ικανοποιούν τη συνθήκη ορίου **και είναι διαφορικά εκφρασμένα στις 6 ομάδες resampling** που αντιστοιχούν στις 6 ομάδες καρκινικών δειγμάτων του σχήματος 3.22. Σε αυτή την περίπτωση, περιμένουμε λίγα περισσότερα αποτελέσματα, καθώς δεν υπάρχει ο περιορισμός της περίπτωσης 1, που αυξάνει τις απαιτήσεις σημαντικά, δεδομένης της μεγάλης διακύμανσης που έχει ο πληθυσμός των καρκινικών κυτταροσειρών.

Περίπτωση 3 Τα γονίδια που ικανοποιούν τη συνθήκη ορίου **και είναι διαφορικά εκφρασμένα όταν γίνεται resampling** από όλο τον πληθυσμό. Περιμένουμε να έχουμε τα περισσότερα γονίδια στη χαλαρή και τη μεσσαία συνθήκη ορίου καθώς εξετάζεται μόνο αυτή η περίπτωση, αλλά δεδομένης της μεγάλης διακύμανσης να μειωνονται σημαντικά τα γονίδια στο αυστηρό όριο.

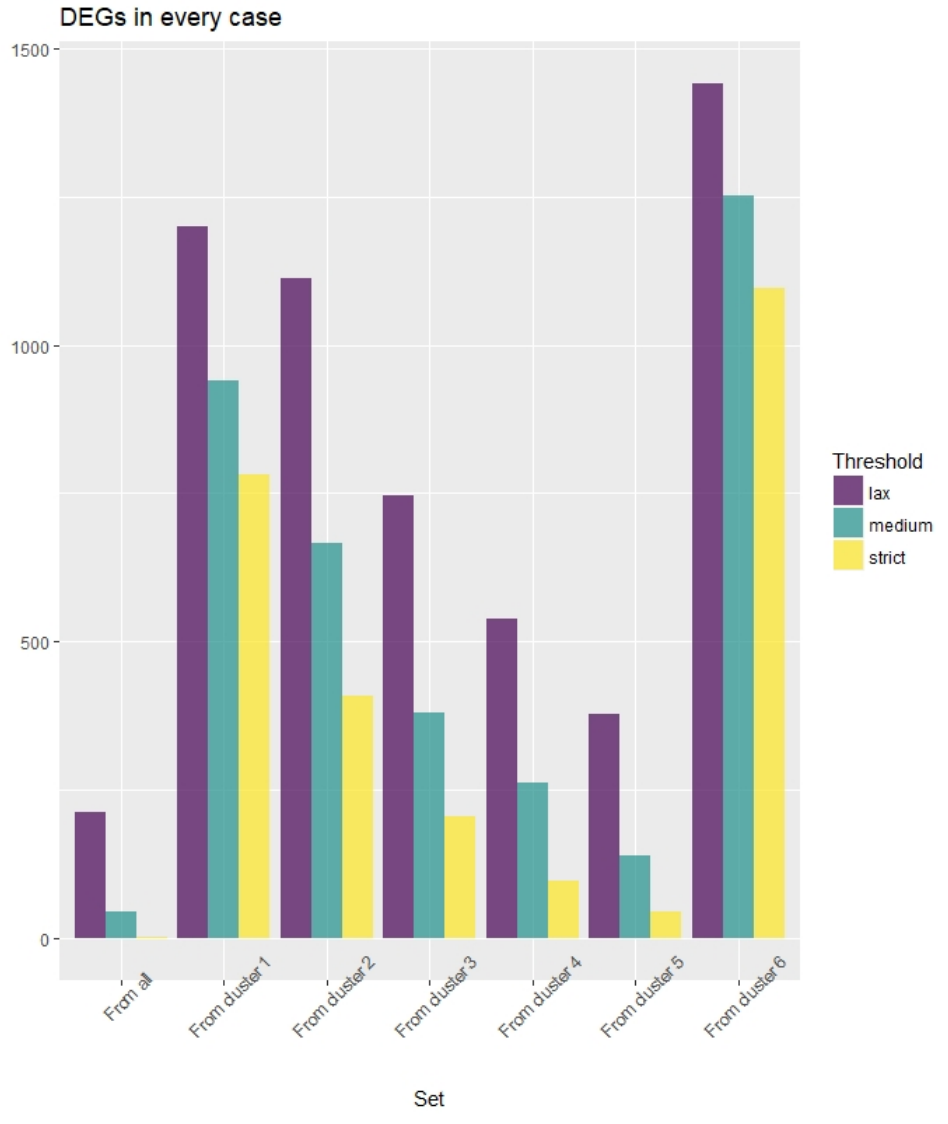
Τα αποτελέσματα παρουσιάζονται στο σχήμα 3.26 και συμφωνούν με τα προαναφερθέντα. Για να επαληθεφθούν και να γίνει έλεγχος της αξιοπιστίας σε κάθε περίπτωση, τα γονίδια που επιλέχθηκαν θα πρέπει να είναι κοινά. Για το σκοπό αυτό δημιουργούνται τα διαγράμματα Venn 3.27 έως 3.29.



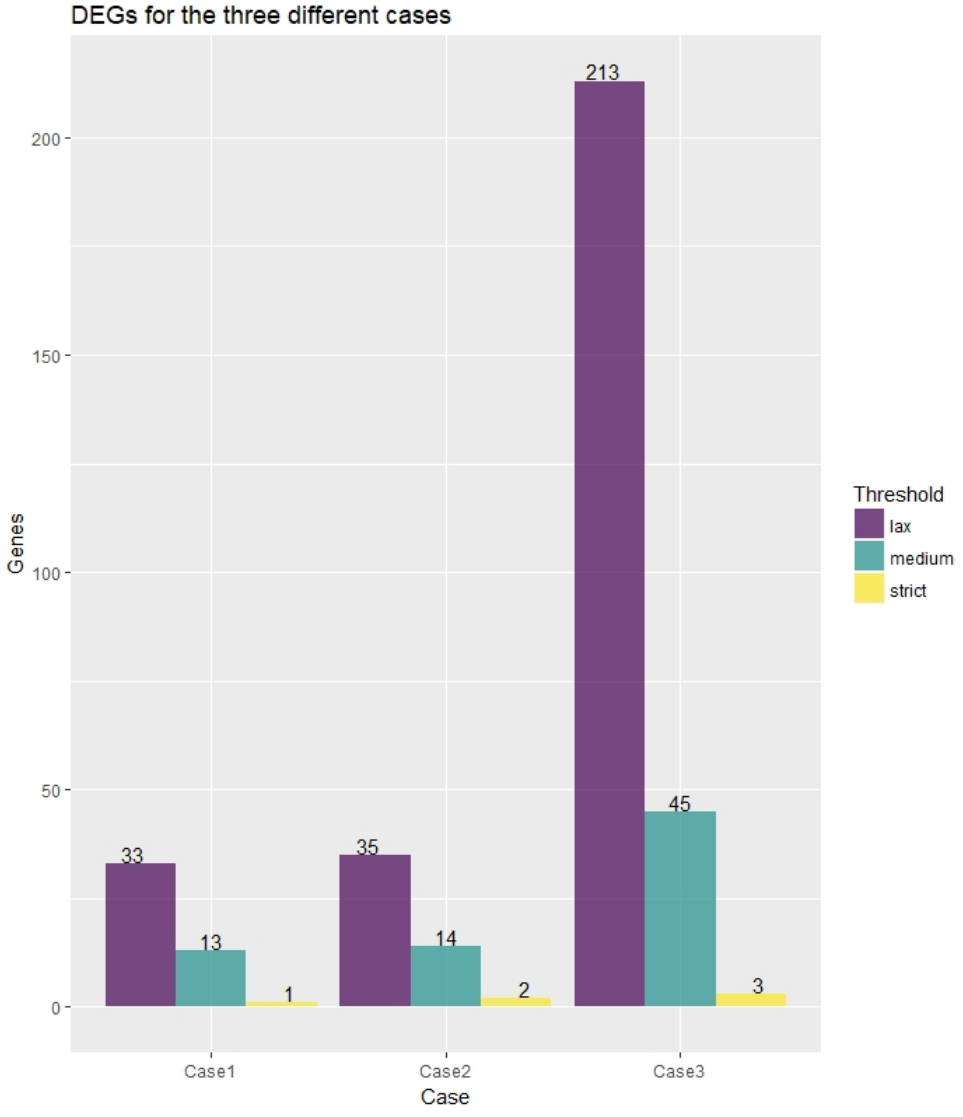
Σχήμα 3.23: Διάγραμμα ηφαιστείου για τη σύγκριση ανάμεσα σε υγιή και καρκινικά δείγματα. Στον οριζόντιο άξονα εκφράζεται η αλλαγή στην έκφραση του γονιδίου, και στον κατακόρυφο η πιθανότητα ότι αυτή η αλλαγή είναι στατιστικά σημαντική ή όχι. Ώς συνθήκη για να χαρακτηριστεί ένα γονίδιο διαφορικά εκφρασμένο είναι να έχει τουλάχιστον διπλάσια ή υποδιπλάσια έκφραση ($|\log(FC)| > 1$) και η τιμή p ανάμεσα στα καρκινικά και υγιή δείγματα για το γονίδιο να είναι μικρότερη από 0.05, δηλαδή πάνω από το σημείο 1.30 του κατακόρυφου άξονα.



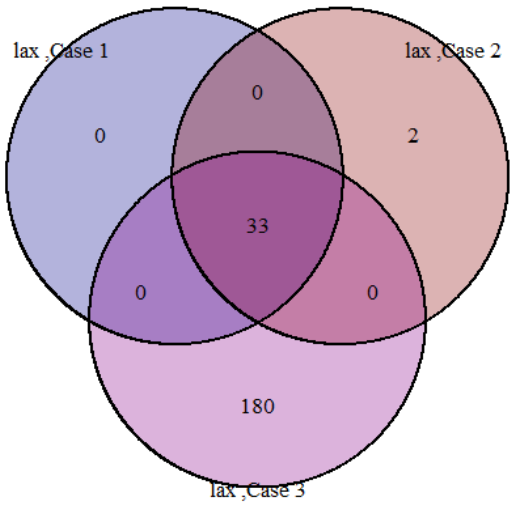
Σχήμα 3.24: Κάθε γονίδιο μπορεί να βρεθεί διαφορικά εκφρασμένο σε $1, \ldots, 1000$ επαναλήψεις του στατιστικού τεστ. Στο σχήμα φαίνεται στον άξονα x ο αριθμός των επαναλήψεων, και στον άξονα y ο αριθμός των διαφορετικών γονιδίων, από τα συνολικά 11328, που βρέθηκαν διαφορικά εκφρασμένα σε τουλάχιστον x επαναλήψεις. Έτσι, στο σημείο 1 του οριζοντίου άξονα, αντιστοιχεί ο αριθμός των γονιδίων που βρέθηκαν τουλάχιστον μία φορά διαφορικά εκφρασμένα, ενώ στο σημείο 1000 του οριζόντιου άξονα, αντιστοιχεί ο αριθμός των γονιδίων που βρέθηκαν διαφορικά εκφρασμένα σε όλες τις επαναλήψεις



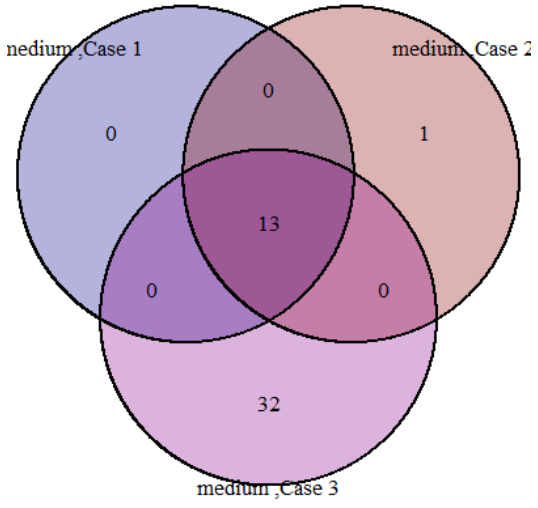
Σχήμα 3.25: Στο σχήμα γίνεται ουσιαστικά διακριτοποίηση των δεδομένων του σχήματος 3.24. Οι τρεις συνθήκες lax - medium - strict αντιστοιχούν σε $260, 540, 800$ ελάχιστες επαναλήψεις ώστε ένα γονίδιο να χαρακτηριστεί ως σημαντικά διαφορικά εκφρασμένο. Παρατηρείται πως στην πρώτη περίπτωση του resampling από όλα τα καρκινικά δείγματα, ακόμα και στο χαλαρό όριο έχουμε λιγότερα από 250 διαφορικά εκφρασμένα γονίδια. Αντίθετα, στις υπόλοιπες περιπτώσεις ο αριθμός αυξάνεται σημαντικά.



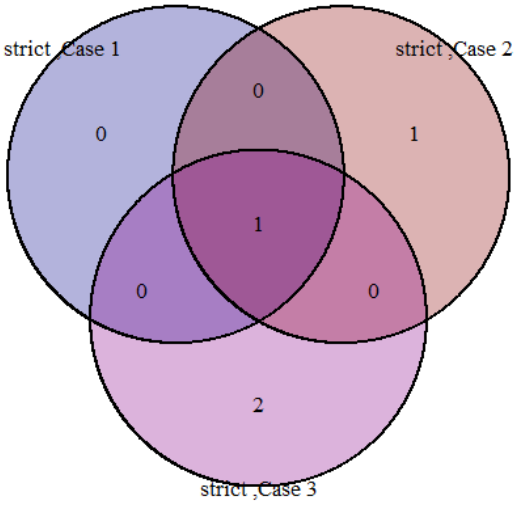
Σχήμα 3.26: Στην πρώτη περίπτωση τα γονίδια είναι διαφορικά εκφρασμένα σε όλες τις ομάδες αλλά και στον πληθυσμό. Στη δεύτερη τα γονίδια είναι διαφορικά εκφρασμένα σε όλες τις ομάδες, και δεν έχει ελεγχθεί το resampling από τον πληθυσμό



Σχήμα 3.27: Στη συνθήκη χαλαρού ορίου, όπως φαίνεται στο σχήμα 3.26, υπάρχουν αντίστοιχα για τις τρείς περιπτώσεις 33, 35 και 213 διαφορικά εκφρασμένα γονίδια. Τα 33 από αυτά υπάρχουν και στις 3 περιπτώσεις, ενώ η περίπτωση 2 έχει μόνο δύο επιπλέον γονίδια από την πρώτη. Προφανώς, τα υπόλοιπα 180 γονίδια περιέχονται μόνο στην τρίτη περίπτωση όπου η επιλογή των δειγμάτων γίνεται από ολόκληρο τον πληθυσμό.



Σχήμα 3.28: Στη συνθήκη μεσσαίου ορίου, υπάρχουν αντίστοιχα για τις τρείς περιπτώσεις 13, 14 και 45 διαφορικά εκφρασμένα γονίδια. Παρατηρείται σε σχέση με τη συνθήκη χαλαρού ορίου μεγάλη μείωση των επιλεγμένων γονιδίων, ιδιαίτερα στην τρίτη περίπτωση. Παράλληλα, στο διάγραμμα Venn παρατηρείται πάλι μεγάλη επικάλυψη στους πληθυσμούς των γονιδίων για τις δύο πρώτες περιπτώσεις, ενώ τα επιπλέον γονίδια ανήκουν πάλι μόνο στην τρίτη περίπτωση.



Σχήμα 3.29: Με την αυστηρή συνθήκη ορίου οι διαφορές είναι πολύ λιγότερες καθώς μένει μόνο ένα κοινό γονίδιο σε όλες τις περιπτώσεις, με τη δεύτερη περίπτωση να διατηρεί άλλο ένα, και την τρίτη άλλα δύο.

3.3.2 Διαφορική ανάλυση μεταλλάξεων μελανώματος

Όπως αναφέρθηκε στην ενότητα 1.4 του κεφαλαίου 1, το μελάνωμα *οδηγείται* από συγκεκριμένες μεταλλάξεις. Ανάλογα με τη μετάλλαξη, ο τύπος της ασθένειας ενδεχομένως αλλάζει, και προφανώς θα πρέπει να αντιμετωπιστεί διαφορετικά, με ουσίες που στοχεύουν την πρωτεΐνη ή το ένζυμο του αντίστοιχου μεταλλαγμένου γονιδίου. Επομένως, υπάρχει ανάγκη εύρεσης χαρακτηριστικών κάθε μετάλλαξης, το οποίο από συστημικής πλευράς σημαίνει διαφοροποίηση της γονιδιακής έκφρασης από μετάλλαξη σε μετάλλαξη. Η βασική ιδέα πίσω από αυτή την ανάλυση είναι ακριβώς ίδια με την προηγούμενη ανάλυση της ενότητας 3.3.1, με τη διαφορά ότι πριν υπήρχαν μόνο δύο διαφορετικοί τύποι κυττάρων, ενώ τώρα υπάρχουν τρεις μεταλλάξεις, όπως περιγράφεται στον πίνακα 3.5. Τα 22 δείγματα που απομένουν αποτελούν κυτταροσειρές wild-type για τις οποίες δεν υπάρχει αρκετή γνώση του μηχανισμού του μελανώματος.
Σημείωση: Όλες οι κυτταροσειρές που φέρουν τη μετάλλαξη PTEN καθώς και μία από τις 22 με τη μετάλλαξη NRAS κουβαλάνε τη μετάλλαξη BRAF [46]. Τόσο αυτό το γεγονός, όσο και το ότι πλέον πρακτικά αναλύονται δείγματα μελανώματος, σημαίνει πως οι διαφορές που θα προκύψουν θα είναι πολύ λιγότερες από ότι στη σύγκριση υγιών και καρκινικών κυτταροσειρών.

Πίνακας 3.5: Οι μεταλλάξεις στα δείγματα που αναλύονται

	BRAF	PTEN	NRAS	Wild-type
Αριθμός Δειγμάτων	84	16	22	22

Πρώτο βήμα στην ανάλυση, όπως και πριν, είναι η οπτικοποίηση των παραπάνω μεταλλάξεων στο χώρο των δύο κύριων συνιστωσών, που φαίνεται στο σχήμα 3.30.

Κατά τη σύγκριση των υγιών και των καρκινικών δειγμάτων, έγινε resampling από τον πληθυσμό των μελανωμάτνω, καθώς ο αριθμός των δύο πληθυσμών διέφερε κατά δύο τάξεις μεγέθους, και η διακύμανση των δύο πληθυσμών διέφερε σημαντικά, γεγονός που θα σήμαινε ότι το Students t-test θα έχανε την αξιοπιστία του. Το ίδιο δε συμβαίνει στους τρεις πληθυσμούς των κυρίων μεταλλάξεων, όπου ο αριθμός των δειγμάτων δε διαφέρει σημαντικά, αλλά και η διακύμανση δε μοιάζει να είναι πολύ διαφορετική, όπως φαίνεται από το σχήμα 3.30 . Για το λόγο αυτό, δεν γίνεται resampling, αλλά κατευθείαν συγκρίνεται η κάθε μετάλλαξη με τις άλλες δύο. Τα διαφορικά εκφρασμένα γονίδια είναι πολύ λίγα, και φαίνονται στους παρακάτω πίνακες.

Πίνακας 3.6: Διαφορικά γονίδια BRAF vs. NRAS

Γονίδιο	$\log(FC)$	$p-value_{BHcorrected}$
BASP1	-1.98	0.013
IL24	-2.38	0.022
IFI27	-2.74	0.022

Πίνακας 3.7: Διαφορικά γονίδια BRAF vs. PTEN

Γονίδιο	$\log(FC)$	$p-value_{BHcorrected}$
PTEN	1.40	0.007

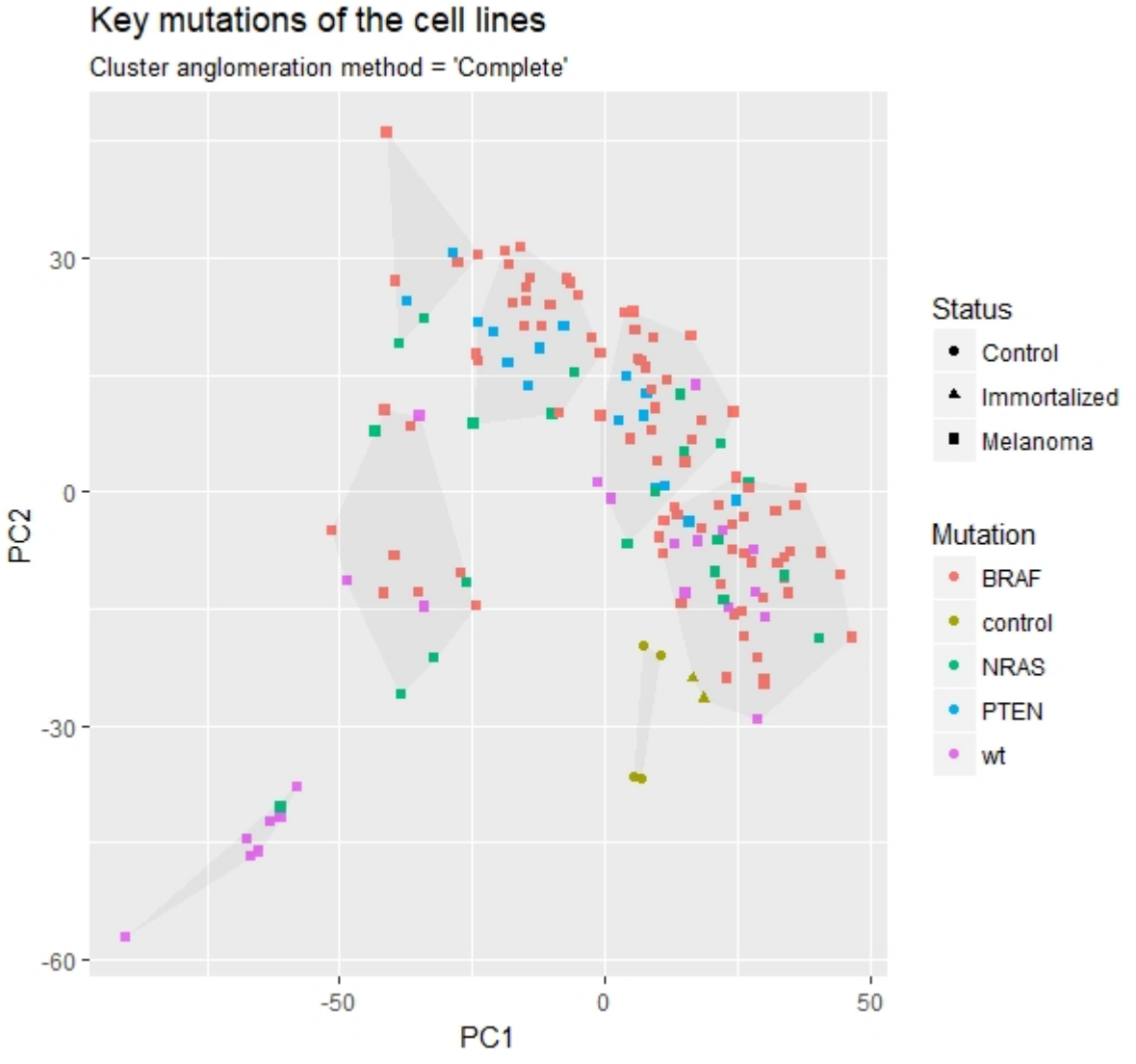
Πίνακας 3.8: Διαφορικά γονίδια NRAS vs. PTEN

Γονίδιο	$\log(FC)$	$p-value_{BHcorrected}$
PTEN	1.96	0.00013

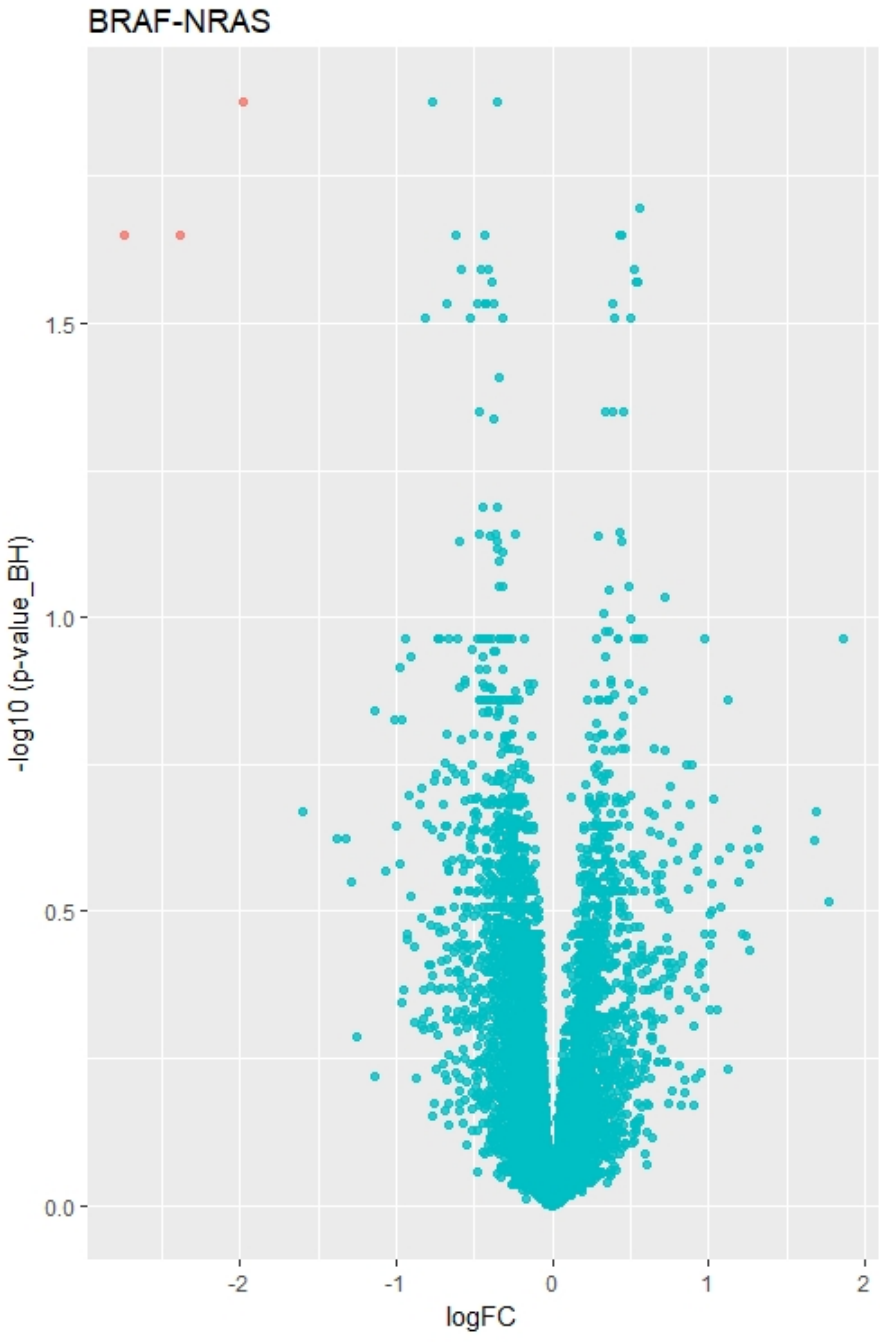
Τα γονίδια είναι πολύ λίγα σε αριθμό. Για να επιβαιωθεί αυτό παρουσιάζονται τα τρία διαγράμματα ηφαιστείου για τις τρεις περιπτώσεις αντίστοιχα, στα σχήματα 3.31 έως 3.33. Οι διαφορές είναι πολύ μικρότερες αν συγκριθούν με το σχήμα 3.23, αφού συγκρίνονται καρκινικά δείγματα μεταξύ τους.
Το διαφορικό γονίδιο στη σύγκριση των δειγμάτων PTEN με τα υπόλοιπα είναι μόνο το ίδιο το PTEN. Ακόμα και αυτή η διαφορά ενδεχομένως να αποτελεί χρήσιμη πληροφορία για τα χαρακτηριστικά του μελανώματος των αντίστοιχων δειγμάτων. Το γονίδιο κωδικοποιεί την αντίστοιχη πρωτεΐνη που αναλυτικά ονομάζεται *Phosphatase and Tensin homolog*. Είναι γνωστό ότι μεταλλάξεις του PTEN αποτελούν βήμα προς τον σχηματισμό καρκίνου. Αυτό εξηγείται από τη δράση του γονιδίου, καθώς αυτό ενεργεί ως *ογκοκατασταλτικό* γονίδιο μέσα από τη δράση της παράγωγης φωσφατάσης από την πρωτεΐνη του. Τα ογκοκατασταλτικά γονίδια έχουν αντίστροφο ρόλο από ότι τα ογκογονίδια, και είναι υπεύθυνα για τον έλεγχο και τη ρύθμιση του ρυθμού ανάπτυξης και διπλασιασμού του κυττάρου. Η απενεργοποίησή του σημαίνει μία λιγότερη (και πολύ σημαντική) απώλεια στους μηχανισμούς άμυνας του κυττάρου ενάντια στον καρκίνο. Στον πίνακα 3.9 παρουσιάζονται οι μέσες εκφράσεις του γονιδίου στου διάφορους πληθυσμούς δειγμάτων. Όπως είναι αναμενόμενο, στα υγιή δείγματα, όπου το γονίδιο δεν είναι κατεσταλμένο, η τιμή είναι η μεγαλύτερη. Στα δείγματα με οδηγό-μετάλλαξη το BRAF και το NRAS, καθώς και τα απροσδιόριστα wild-type η έκφραση είναι μικρότερη από ότι στα υγιή. Εμφανής, και στατιστικά σημαντική, διαφορά, υπάρχει στα δείγματα με τη μετάλλαξη του γονιδίου.

Πίνακας 3.9: Έκφραση του PTEN στα δείγματα

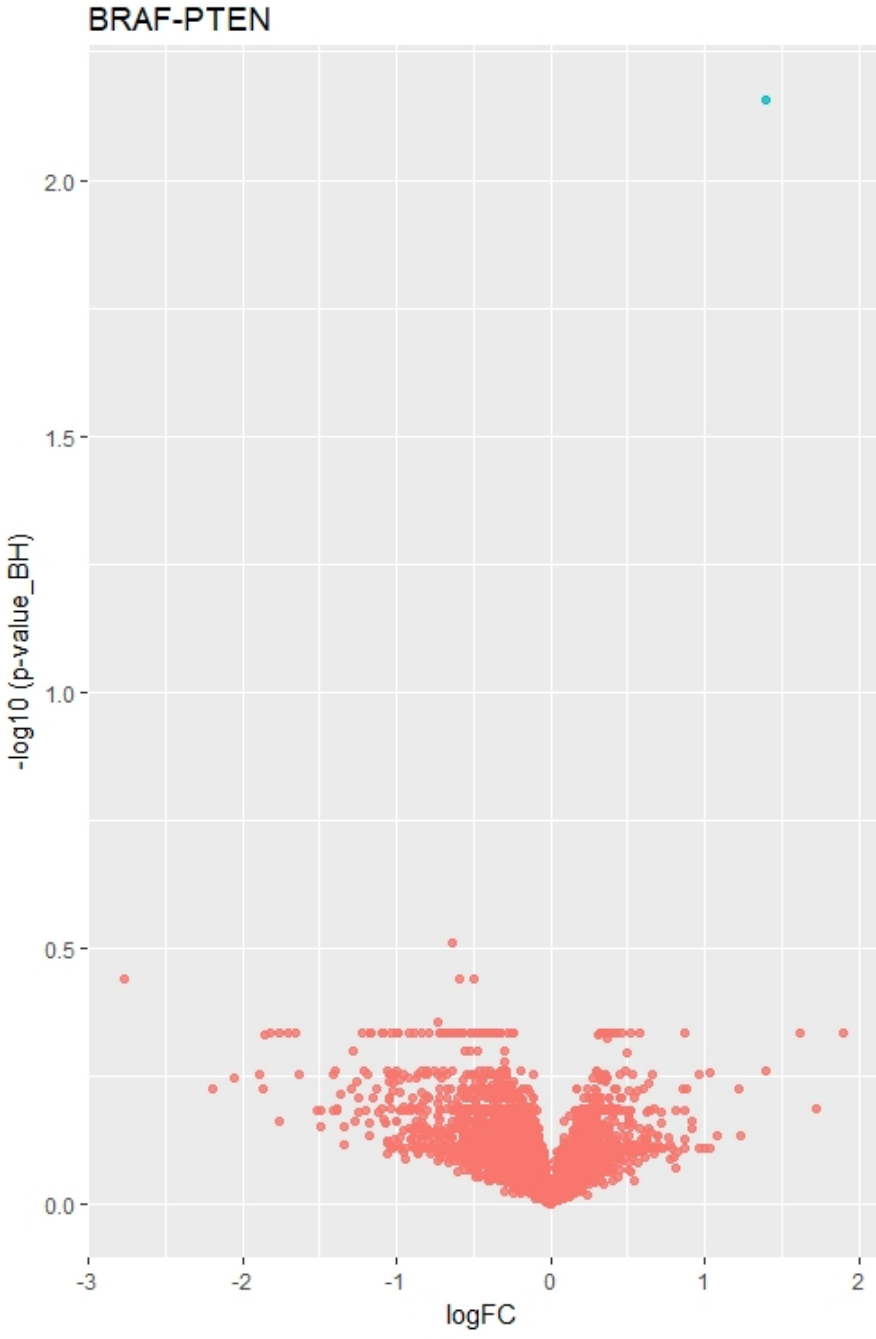
Πληθυσμός	Μέση έκφραση
Υγιή δείγματα	8.11
NRAS	7.59
Wild-type	7.51
BRAF	7.07
PTEN	5.64



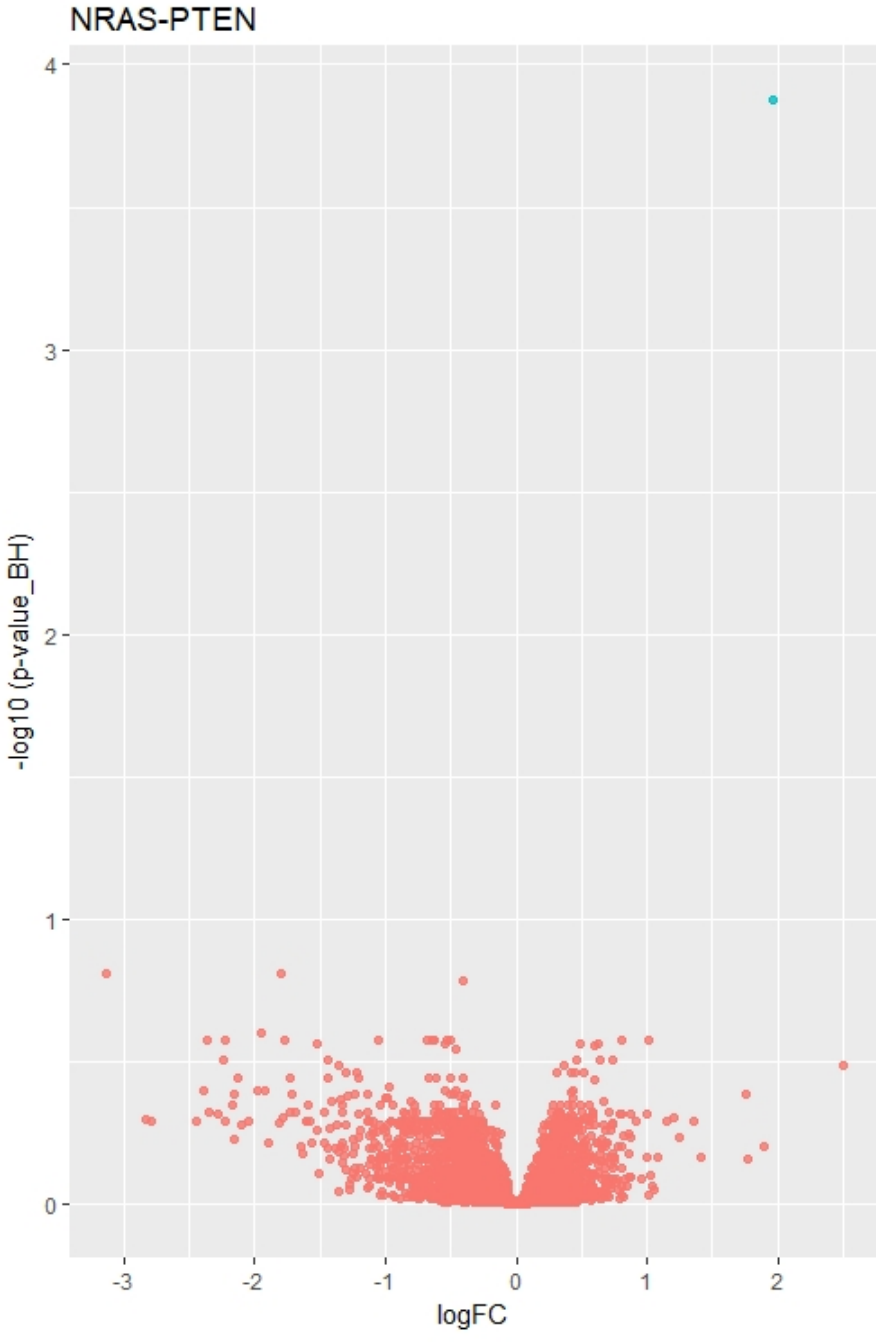
Σχήμα 3.30: Όπως αναφέρεται στην αντίστοιχη παράγραφο, οι διαφορές ανάμεσα στις διαφορετικές μεταλλάξεις BRAF-PTEN-NRAS. Δεν παρατηρούνται σχηματισμένες ομάδες μεταλλάξεων, αλλά είναι διασκορπισμένα στον χώρο των κύριων συνιστωσών. Για λόγους πληρότητας φαίνονται και τα wild-type και υγιή δείγματα.



Σχήμα 3.31: Διάγραμμα ηφαιστείου για τη σύγκριση των δειγμάτων με οδηγό-μετάλλαξη BRAF με δείγματα με οδηγό-μετάλλαξη NRAS. Είναι εμφανές ότι τα διαφορικά γονίδια είναι πολύ λιγότερα, και ιδιαίτερα η αλλαγή της έκφρασης, δηλαδή οι μετρήσεις στον άξονα x έχουν πολύ μικρότερο εύρος από σύγκριση καρκινικών-υγιών δειγμάτων, όπως στο σχήμα 3.23



Σχήμα 3.32: Διάγραμμα ηφαιστείου για τη σύγκριση των δειγμάτων με οδηγό-μετάλλαξη BRAF με δείγματα με οδηγό-μετάλλαξη PTEN. Εδώ οι διαφορές είναι ακόμα λιγότερες, αφού εξάλλου όλα τα δείγματα PTEN φέρουν και τη μετάλλαξη BRAF.



Σχήμα 3.33: Διάγραμμα ηφαιστείου για τη σύγκριση των δειγμάτων με οδηγό-μετάλλαξη NRAS με δείγματα με οδηγό-μετάλλαξη PTEN.

3.4 Κατηγοριοποίηση κυττάροσειρών

Με βάση τα γονίδια που πρέκυψαν στην παραπάνω ενότητα, πρώτα από τη σύγκριση μεταξύ υγιών και καρκινικών δειγμάτων, και στη συνέχεια από τη σύγκριση μεταξύ των καρκινικών δειγμάτων με διαφορετικές μεταλλάξεις, θα γίνει σε αυτή την ενότητα κατηγοριοποίηση. Για την πρώτη περίπτωση θα χρησιμοποιηθούν μηχανές διανυσμάτων υποστήριξης, και στη δεύτερη γραμμική διακριτική ανάλυση.

3.4.1 Καρκινικά και υγιή κύτταρα

Για τα γονίδια που χρησιμοποιούνται ως features για τη διάκριση ανάμεσα στα δείγματα μελανώματος και τις υγιείς κυτταροσειρές διακρίνονται τρεις περιπτώσεις, συγκεκριμένα οι συνθήκες ορίου της ενότητας 3.3, όπου ένα γονίδιο για να χαρακτηριστεί σημαντικά διαφορικά εκφρασμένο πρέπει να είναι διαφορικά εκφρασμένο σε 260, 540 και 800 επαναλήψεις αντίστοιχα από τα 1000 στατιστικά τεστ που γίνονται κάθε φορά.
Επιπλέον ο αναγνώστης ανατρέχοντας στην ενότητα 3.3 θα παρατηρήσει ότι έχουν διακριθεί τρία διαφορετικά σενάρια για τα σημαντικά διαφορικά εκφρασμένα γονίδια όλου του πάνελ των καρκινικών κυτταροσειρών. Αυτά ήταν:

1. Τα σημαντικά διαφορικά γονίδια πρέπει να περιέχονται σε όλες τις επιμέρους ομάδες δειγμάτων καθώς και σε ολόκληρο τον πληθυσμό (resampling από κάθε ομάδα και από όλα τα δείγματα αντίστοιχα).
2. Τα σημαντικά διαφορικά γονίδια πρέπει να περιέχονται και στις 6 ομάδες δειγμάτων, αλλά όχι στις επαναλήψεις που γίνονται για ολόκληρο τον πληθυσμό.
3. Τα σημαντικά διαφορικά γονίδια πρέπει να περιέχονται μόνο στην περίπτωση που το resampling γίνεται από ολόκληρο τον πληθυσμό.

Από τα τρία σενάρια, το δεύτερο επιλεχθηκε ως το πιο αξιόπιστο:

- Όταν τα σημαντικά διαφορικά γονίδια πρέπει να περιέχονται μόνο στην περίπτωση που εξετάζεται χωρίς ομαδοποίηση όλος ο πληθυσμός των 144 κυτταροσειρών, υπάρχει μεγάλη διαφορά στον αριθμό των γονιδίων καθώς αλλάζει η συνθήκη ορίου, όπως φαίνεται στο σχήμα 3.26. Αυτό ίσως οφείλεται στο ότι λόγω της αυξημένης διακύμανσης της κατανομής, δεν είναι κατάλληλα τα t-test που πραγματοποιούνται.
- Η περίπτωση που τα σημαντικά διαφορικά γονίδια πρέπει να περιέχονται σε κάθε μία από τις 6 ομαδοποιήσεις αλλά και σε όλο τον πληθυσμό κρίνεται ότι μπορεί να αποκλείει γονίδια τα οποία ενδεχομένως περιέχουν πληροφορία για την κατηγοριοποίηση.
- Θυμίζεται ότι τα σενάρια 1 και 2 έχουν πολύ μικρές διαφορές, με το δεύτερο σενάριο να συμπεριλαμβάνει τελικά το πολύ 2 παραπάνω γονίδια. Τα δύο αυτά γονίδια, δεν κοστίζουν υπολογιστικά για να συμπεριληφθούν ως features στην ανάλυση. Αντίθετα, αν χρησιμοποιούνταν τα επιπλέον 180 γονίδια του σεναρίου 3, το υπολογιστικό κόστος θα ανέβαινε εκθετικά, και θα αυξανόταν η πιθανότητα overfitting των δεδομένων.

Τελικά, τα γονίδια που θα χρησιμοποιηθούν φαίνονται στο σχήμα 3.34 Για την αξιολόγηση των αποτελεσμάτων είναι αναγκαίο να αναφερθούν οι παρακάτω ορισμοί:

True Positive (TP): Δηλώνει ένα αντικείμενο τάξης X που σωστά έχει κατηγοριοποιηθεί στην τάξη X.

Genes used as features for classification

Lax.threshold	Medium.threshold	Strict.threshold
ASNS, BAG2, CELF2, CITED1, CRYL1, CXCL2, DOK5, FKBP1B, GADD45A, GATA6, HMGA2, HSPA2, IER3, KCNJ13, LIPA, MAP4K3, MEIS2, MTHFD2, P2RX7, PER2, PMAIP1, PYCR1, RAB33A, SLC1A5, SLC39A14, SOBP, ST6GALNAC2, STK17A, SYK, TBX3, TRIB3, TSPAN13, VEGFA, WFDC1, ZIC1	CITED1, CRYL1, DOK5, FKBP1B, GATA6, HMGA2, KCNJ13, LIPA, MTHFD2, P2RX7, PMAIP1, PYCR1, RAB33A, ZIC1	HMGA2, PMAIP1

Σχήμα 3.34: Τα γονίδια που χρησιμοποιούνται από τα SVMs για την κατηγοριοποίηση καρκινικών και υγιών δειγμάτων. Οι τρεις περιπτώσεις αντιστοιχούν στις τρεις συνθήκες ορίου, δηλαδή κάθε γονίδιο για να είναι σημαντικά διαφορικά εκφρασμένο πρέπει να έχει βρεθεί διαφορικά εκφρασμένο σε 260 - 540 - 800 επαναλήψεις από τις 1000, και στις 6 ομάδες δειγμάτων μελανώματος.

False Positive (FP): Δηλώνει ένα αντικείμενο τάξης *όχι-*$X = X'$ που έχει κατηγοριοποιηθεί λάθος στην τάξη X.

False Negative (FN): Δηλώνει ένα αντικείμενο τάξης X που έχει κατηγοριοποιηθεί στην τάξη X'.

True Negative (TN): Δηλώνει ένα αντικείμενο τάξης X' που σωστά έχει κατηγοριοποιηθεί στην τάξη X'.

Πίνακας σύγχυσης (Confusion Matrix): Στην επιστήμη των μηχανών εκμάθησης, το πιο σημαντικό ίσως στοιχείο που επιτρέπει την οπτικοποίηση της απόδοσης ενός αλγορίθμου είναι ο πίνακας σύγχυσης, ή αλλιώς πίνακας σφαλμάτων. Κάθε γραμμή του πίνακα συμβολίζει την κατηγοριοποίηση που έγινε από τον αλγόριθμο στο σετ εκπαίδευσης, ενώ κάθε στήλη την πραγματική κατηγοριοποίηση του συνόλου. Στη διαγώνιο επομένως βρίσκοντα οι επιτυχημένες προβλέψεις, και σε όλες τις υπόλοπες θέσεις του πίνακα, οι αποτυχημένες. Για παράδειγμα παρουσιάζεται ο πίνακας 3.10 που αναφέρεται στην κατηγοριοποίηση των διαφόρων ειδών του φυτού *Iris: Setosa, Versicolor, Virginica*. Έτσι για παράδειγμα,από τα 50 είδη *Iris Setosa*, τα 40 κατηγοριοποιήθηκαν σωστά από τον αλγόριθμο, ενώ 4 χαρακτηρίστηκαν *Iris Versicolor* και 6 ως *Iris Virginica*. Στη συγκεκριμένη περίπτωση, δηλαδή για το είδος *Iris Setosa* δημιουργείται ο πίνακας 3.11:

Ακρίβεια (Acc): Ορίζεται ως το μέγεθος:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

και χρησιμοποιείται για να ποσοτικοποιηθεί η γενική απόδοση του αλγορίθμου. Ο

Πίνακας 3.10: Πίνακας σύγχυσης ενός μοντέλου μηχανής εκμάθησης

		Πραγματική τάξη		
		Set.	*Ver.*	*Vir.*
Προβλεπόμενη τάξη	*Set.*	40	4	6
	Ver.	10	38	2
	Vir.	3	4	43

Πίνακας 3.11: Πίνακας σύγχυσης για μία τάξη

		Πραγματική τάξη	
		Set.	όχι *Set.*
Προβλεπόμενη τάξη	*Set.*	40	10
	όχι *Set.*	13	87

παραπάνω λόγος με μικρές τροποποιήσεις μας δείχνει διαφορετικά στοιχεία απόδοσης του αλγορίθμου, όπως η ευαισθησία του να βρίσκει TP τιμές και άλλα. Προφανώς, όταν ο αλγόριθμος έχει επιτυχία στην πρόβλεψη μίας τάξης, αλλά αποτυγχάνει στις υπόλοιπες, η ακρίβεια δεν είναι κατάλληλο μέγεθος για την αξιολόγηση του αλγορίθμου. Για τέτοιες περιπτώσεις, και ειδικά όταν οι δύο πληθυσμοί διαφέρουν ως προς το πλήθος τους, ένα εναλλακτικό μέγεθος που δίνει την αξιολόγηση της μεθόδου είναι ο συντελεστής συσχέτισης Matthews.

Συντελεστής συσχέτισης Matthews (MCC): Χρησιμοποιείται για να μετρηθεί η ποιότητα δυαδικής κατηγοριοποίησης όπως στην περίπτωση υγιών έναντι καρκινικών δειγμάτων. Γενικά θεωρείται ότι μπορεί να χρησιμοποιηθεί και για δύο τάξεις πολύ διαφορετικών μεγεθών. Στην ουσία πρόκειται για τη συσχέτιση μεταξύ των αληθινών και των προβλεπόμενων τιμών, και επιστρέφει μία τιμή ανάμεσα στο -1 και το 1, με +1 να σημαίνει τέλεια πρόβλεψη, 0 να δηλώνει τυχαιότητα ανάμεσα στις δύο κατανομές, και -1 να δηλώνει πλήρη αντίθεση. Υπολογίζεται από τον παρακάτω τύπο:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Μία κλασική μέθοδος για την αξιολόγηση των αλγορίθμων εκμάθησης είναι το Cross Validation. Με το Cross Validation το σύνολο εκμάθησης, το οποίο μπορεί να είναι πλέον και ολόκληρο το σύνολο που μελετάμε, χωρίζεται σε k τμήματα. Στη συνέχεια ο αλγόριθμος χρησιμοποιεί τα $k-1$ τμήματα για να εκπαιδευτεί, και υπολογίζεται η απόδοσή του από την πρόβλεψη που θα κάνει στο 1 τμήμα που απομένει. Αυτό επαναλαμβάνεται κ φορές όπου κάθε φορά το μοντέλο ελέγχεται σε διαφορετικό τμήμα. Τέλος, υπολογίζονται οι μέσες τιμές των μεγεθών που χρησιμοποιούνται για την αξιολόγηση του μοντέλου.
Στον πληθυσμό που πρέπει να γίνει κατηγοριοποίηση, υπάρχουν 6 υγιή και 144 καρκινικά δείγματα. Το σύνολο που θα χρησιμοποιηθεί για το cross validation θα περιέχει προφανώς τα περιορισμένα 6 υγιή δείγματα, και κάποιον αριθμό καρκινικών κυτταροσειρών. Ο αριθμός τους ενδεχομενως θα επηρεάζει την απόδοση του αλγορίθμου, για αυτό μελετούνται όλες οι περιπτώσεις, με 6 έως 144 δείγματα μελανώματος. Επειδή υπάρχουν 6 υγιή δείγματα, και για να έχει νόημα η επαλήθευση του μοντέλου με το cross validation πρέπει σε κάθε τμήμα να περιέχεται τουλάχιστον μία υγιής κυτταροσειρά, επιλέγεται

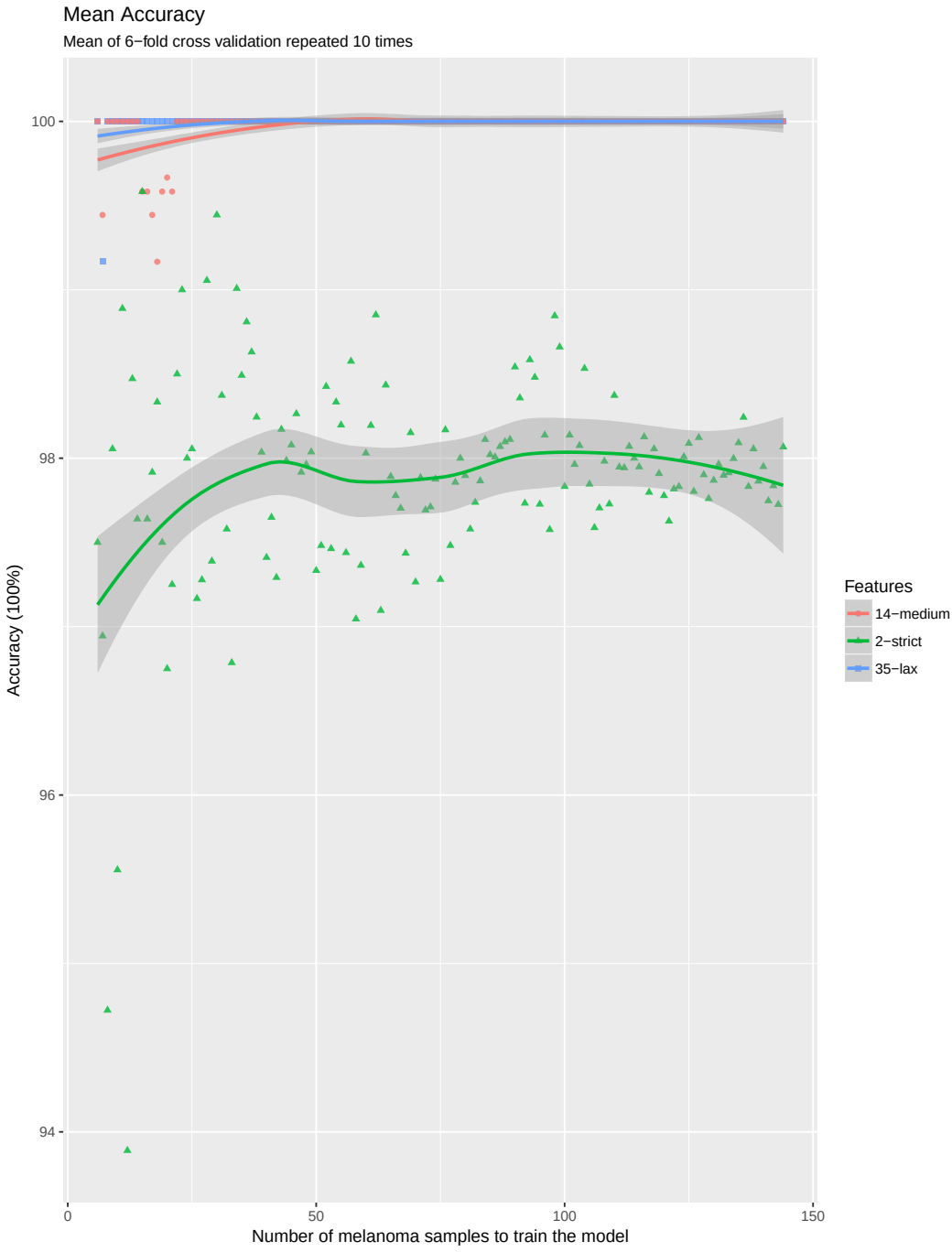
$k = 6$, για το χωρισμό του συνόλου σε τμήματα. Τελικά, το σύνολο που θα χωριστεί περιέχει 6 υγιή δείγματα και $m = 6, \ldots, 144$ δείγματα μελανώματος. Κάθε τμήμα περιέχει 1 υγιές δείγμα και τουλάχιστον $\lfloor \frac{m}{6} \rfloor$.
Ξανά, για να ελεγχθεί η καταλληλότητα των feature γονιδίων να προβλέψουν αν ένα κύτταρο είναι καρκινικό ή όχι, σε κάθε σύνολο με τα υγιή και m δείγματα μελανώματος γίνεται resampling 10 φορές, δηλαδή προκύπτουν 10 διαφορετικοί πληθυσμοί των m δειγμάτων, από τους οποίους υπολογίζονται οι μέσοι όροι των μεταβλητών που μας ενδιαφέρουν. Προφανώς στην περίπτωση που έχουμε $m = 144$ δε γίνεται resampling. Για τις παραμέτρους των SVMs, επιλέχθηκε αρχικά συνάρτηση κελύφους radial. Τα αποτελέσματα της παραπάνω διερεύνησης φαίνονται στα σχήματα 3.35 έως 3.37.

Από τα σχήματα παρατηρείται κατάρχάς ότι για τις περιπτώσεις που το μοντέλο εκπαιδεύεται με 35 ή 14 γονίδια, η διαφοροποίηση είναι σχεδόν τέλεια, με ελάχιστη ακρίβεια μεγαλύτερη του 99% και συντελεστή συχέτισης Matthews σχεδόν ίσο με 1. Ο στόχος της κατηγοριοποίησης είναι να μπορέσει το μοντέλο να εκπαιδευφθεί με όσο το δυνατόν λιγότερα γονίδια γίνεται, για αυτό ενδιαφέρει ιδιαίτερα η απόδοση με μόνο τα 2 γονίδια **HMGA2, PMAIP1**. Όταν το μοντέλο εκπαιδεύται μόνο με αυτά, η απόδοση είναι ικανοποιητική ($> 90\%$) μέχρι την περίπτωση που έχουμε 50 καρκινικά δείγματα. Απο εκεί και έπειτα, η απόδοση σταδιακά μειώνεται μέχρι το 51%. Η απόδοση προσδιορίζεται εδώ με τον MCC και όχι με το ACC, καθώς λόγω της αύξησης των δειγμάτων το τελευταίο δε δίνει αξιόπιστα αποτελέσματα αξιολόγησης. Ο υψηλότερος MCC που υπολογίζεται είναι για 15 καρκινικά δείγματα και ισούται με 0.993, ενώ η τιμή μειώνεται σταδιακά μέχρι το 0.514. Για τα 144 δείγματα μελανώματος, όταν περιλαμβάνονται όλα στο σύνολο εκμάθησης, ισχύει $MCC_{144} = 0.574$.
Από βιολογικής πλευράς, τα δύο γονίδια που χρησιμοποιούνται έχουν συνδεθεί με τον καρκίνο του ανθρώπου:

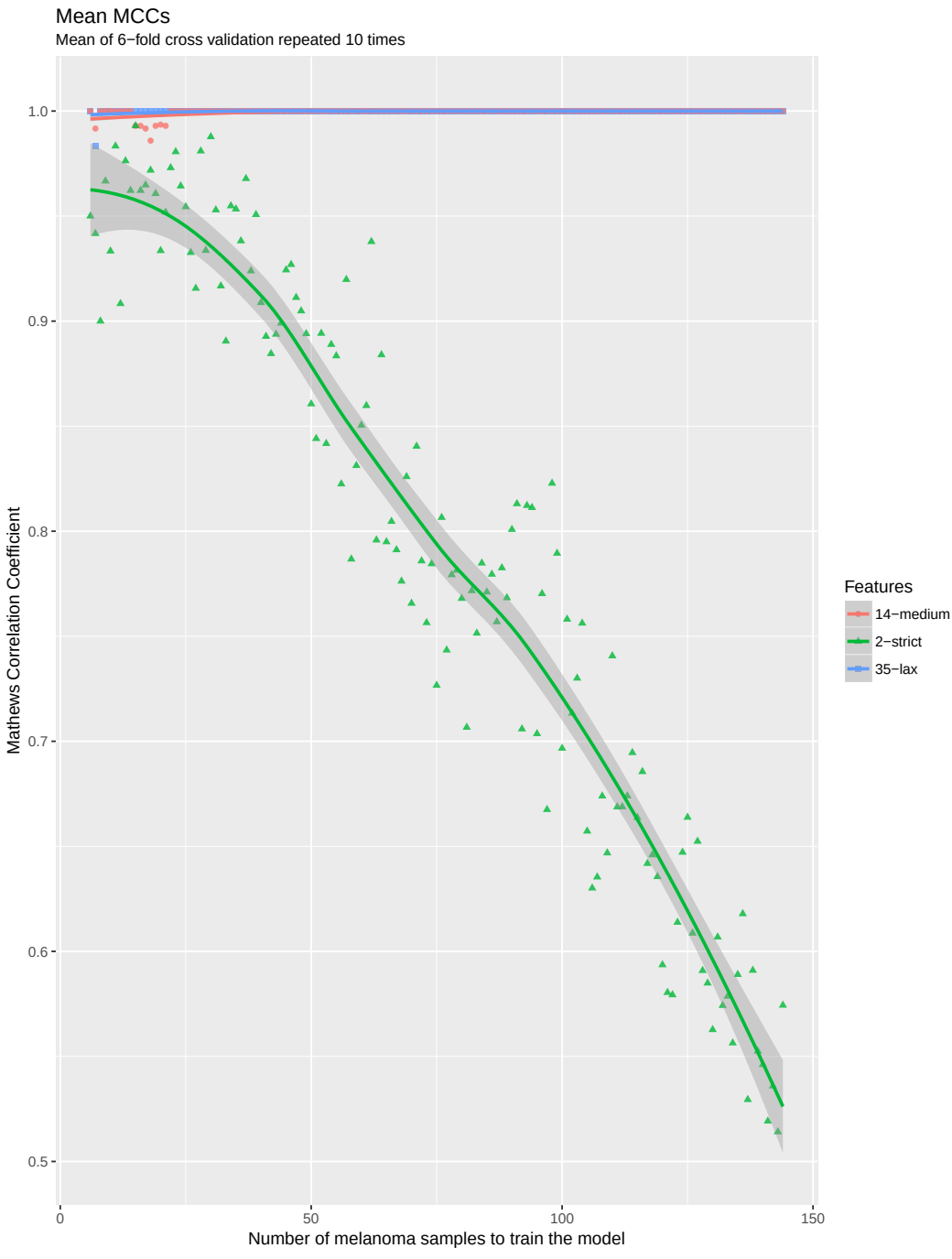
HMGA2: Κωδικοποιεί την αντίστοιχη πρωτεΐνη που ανήκει στην πρωτεϊνική οικογένεια HMG. Η συγκεκριμένη πρωτεΐνη μπορεί να δράσει ως ρυθμιστικός παράγοντας της μεταγραφής του DNA, καθώς περιέχει τομείς (domains) που συνδέονται σε αυτό εμποδίζοντας τη μεταγραφή του. Η έκφραση του γονιδίου έχει συνδεθεί με τον σχηματισμό από καλοήθεις και κακοήθεις όγκους, καθώς και με μεταλλάξεις που φαίνεται να ευνοούν τον καρκίνο [47]. Ειδικά για το μελάνωμα, το γονίδιο έχει προταθεί ως βιοδείκτης για την ανάπτυξη και την πρόγνωση του καρκίνου [48]. Πράγματι, το γονίδιο έχει αυξημένο λόγο έκφρασης logFC σε όλα τα δείγματα. Πρέπει να σημειωθέι ότι ακόμα ο μηχανισμός δράσης με τον οποίο το γονίδιο συνεισφέρει στο σχηματισμό του καρκίνου δεν είναι γνωστός.

PMAIP1: Κωδικοποιεί την αποπτοτική πρωτεΐνη με το ίδιο όνομα, η οποία είναι γνωστή και ως Noxa, μέλος της πρωτεϊνικής οικογένειας Bcl-2. Οι πρωτεΐνες αυτές συμμετέχουν, είτε ως υπερ- είτε ως αντί-αποπτιτικοί ρυθμιστές σε πολλές κυτταρικές δραστηριότητες. Η έκφραση του γονιδίου ρυθμίζεται από το ογκοκατασταλτικό γονίδιο p53 [49]. Πρέπει να σημειωθεί ότι στα δείγματα που μελετώνται, το γονίδιο είναι υπερεκφρασμένο στο μελάνωμα με μέση έκφραση που φαίνεται στο σχήμα 3.38β΄.

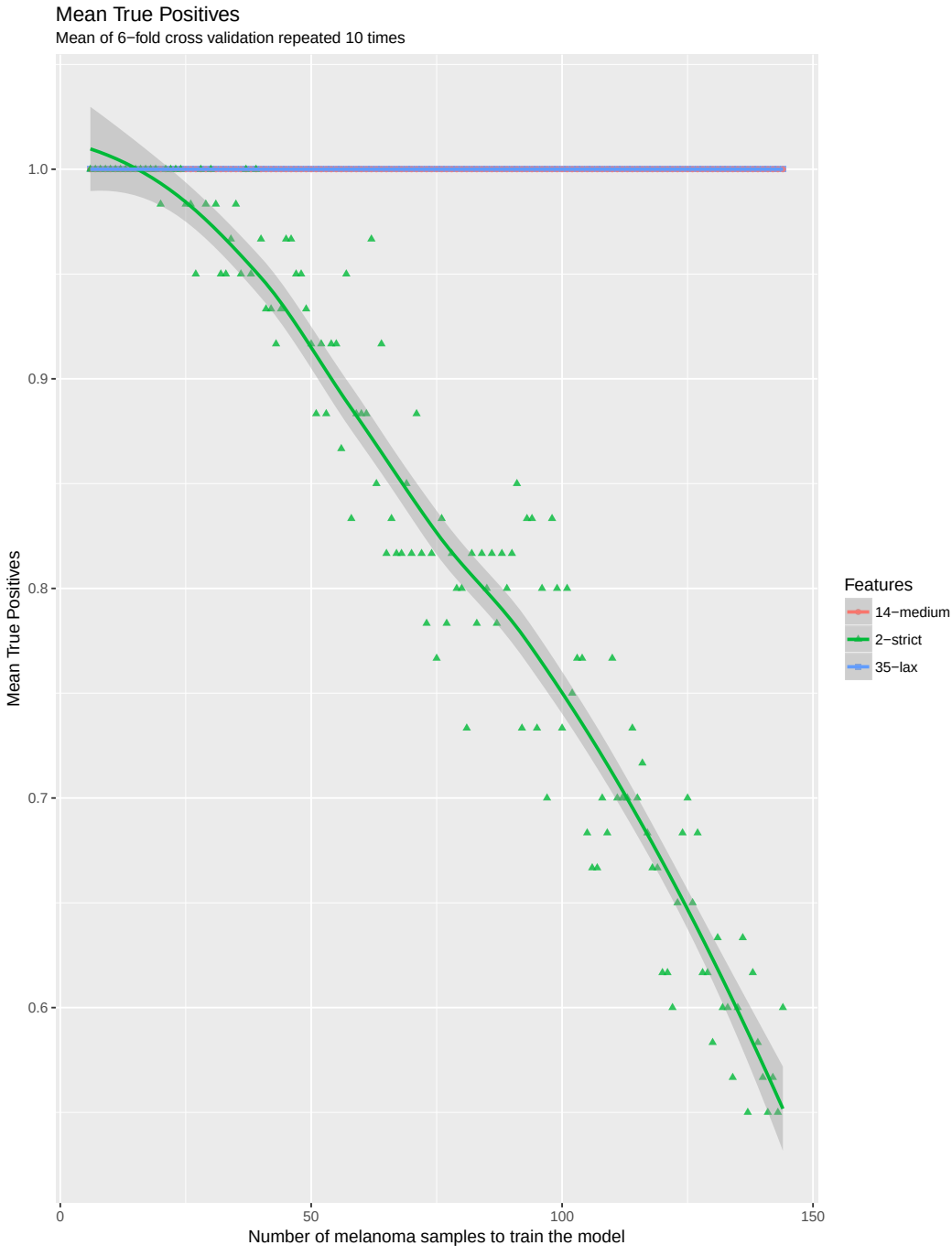
Όπως παρατηρήθηκε, ενώ με τα 33 και 13 γονίδια των πιο αυστηρών συνθηκών ορίου γίνεται σχεδόν τέλεια κατηγοριοποίηση, όταν χρησιμοποιούνται μόνο 2 γονίδια, υπάρχει σταδιακή μείωση της απόδοσης του αλγορίθμου, που αυξάνεται με την αύξηση του αριθμού δειγμάτων καρινικών κυττάρων στο οποία εκπαιδεύεται και που πρέπει να κατηγοριοποιήσει ο αλγόριθμος. Επομένως, πρέπει να βρεθεί ο ελάχιστος αριθμός γονιδίων για τον οποίο θα είναι η απόδοση 100 και ο MCC και τα True Positives ίσα με τη μονάδα. Κατασκευάστηκε σύντομος αλγόριθμος βασισμένος στην ιδέα πίσω από τους εξελικτικούς



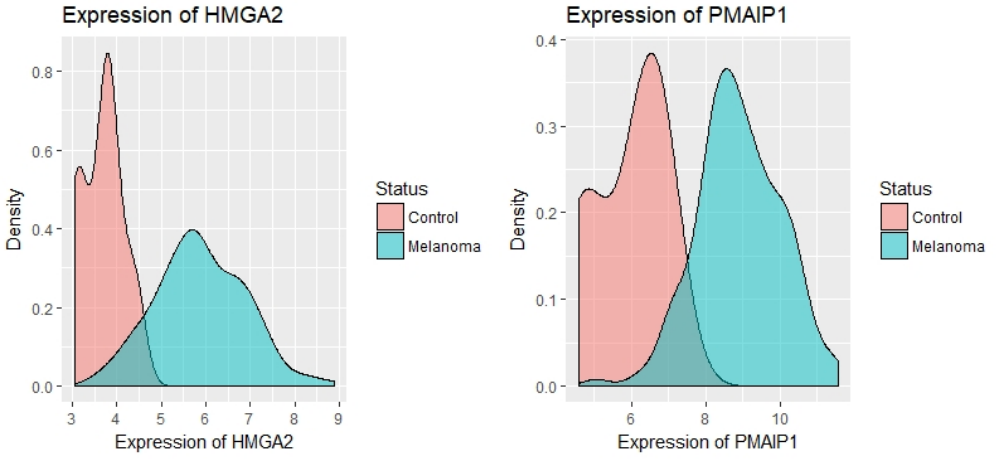
Σχήμα 3.35: Η ακρίβεια Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$. Στην περίπτωση που μελετάται, υπάρχουν δύο τάξεις για τα δείγματα, δηλάδη control και melanoma. Επομένως, ο πίνακας σύγχυσης που μελετάται αντιστοιχεί στα TP τα σωστά προβλεπόμενα υγιή και στα TN τα σωστά προβλεπόμενα καρκινικά δείγματα.



Σχήμα 3.36: Όπως αναφέρεται στην αντίστοιχη παράγραφο της ενότητας 3.4.1, όταν υπάρχει μεγάλη διαφορά στους πληθυσμούς των δύο τάξεων, ο συντελεστής συσχέτισης Matthews είναι το κατάλληλο μέγεθος για την αξιολόγηση του αλγορίθμου. Πράγματι, σε αντίθεση με το προηγούμενο σχήμα 3.35, όπου η ακρίβεια φαίνεται να είναι σχεδόν 100% ανεξάρτητα από τον αριθμό των δειγμάτων μελανώματος, με τον *MCC* παρατηρείται ότι καθώς αυξάνεται ο αριθμός των καρκινικών δειγμάτων, το μοντέλο όλο και πιο συχνά αποτυγχάνει.



Σχήμα 3.37: Όπως αναφέρθηκε στο κείμενο, λόγω του περιορισμένου αριθμού υγιών κυτταροσειρών, το μοντέλο εκπαιδεύεται με 5 υγιή δείγματα, και ελέγχεται η δυνατότητα πρόβλεψης του δείγματος που απομένει. Ουσιαστικά, με το παρόν σχήμα, γίνεται φανερό ότι ο αλγόριθμος χαρακτηρίζεται από τη δυνατότητά του να προβλέψει σωστά αυτό το δείγμα. Παρατηρείται παρόμοια συμπεριφορά της σωστής πρόβλεψης του υγιούς δείγματος, με τον συντελεστή συσχέτισης Matthews.



(α') Έκφραση του γονιδίου HMGA2 (β') Έκφραση του γονιδίου PMAIP1

Σχήμα 3.38: Τα δύο γονίδια που χρησιμοποιήθηκαν για την κατηγοριοποίηση κυτταροσειρών. Το HMGA2 έχει αποδειχθεί ότι συμβάλλει στην ανάπτυξη και τον σχηματισμό καρκίνου, αν και ακόμα είναι άγνωστος ο μηχανισμός δράσης του. Επιπλέον, έχει δυνατότητα να χρησιμοποιηθεί ως βιοδείκτης για την πρόγνωση του καρκίνου. Το PMAIP1 αντιθέτως ανήκει στα αποπτωτικά μονοπάτια του κυττάρου, και η έκφρασή του θα έπρεπε να είναι μειωμένη στα καρκινικά δείγματα. Η αντίθετη συμπεριφορά που παρατηρείται, ίσως να είναι σημάδι για αντίσταση του μελανώματος στα ογκογονίδια p53. Περαιτέρω βιολογική μελέτη πρέπεινα διεξαχθεί για να βγει συμπέρασμα.

αλγορίθμους ο οποίος εξετάζει την περίπτωση που στα 2 γονίδια προστίθεται ένα ακόμη, προερχόμενο από τα 13 γονίδια της αμέσως προηγούμενης συνθήκης ορίου. Για πιο αξιόπιστα αποτελέσματα, ο αλγόριθμος ελέγχεται και στα 150 δείγματα, ενώ επαναλαμβάνεται 10 φορές με 6 σετ cross-validation κάθε φορά. Τα αποτελέσματα φαίνονται στον πίνακα 3.12: Παρατηρείται πως με την προσθήκη ενός γονιδίου από τα **MTHFD2, RAB33A**, το μοντέλο βελτιστοποιείται και δίνει τα ακριβή αποτελέσματα. Τα δύο γονίδια έχουν βρεθεί ότι συσχετίζονται με τον καρκίνο:

MTHFD2: Το γονίδιο κωδικοποιεί ένα μιτοχονδριακό ένζυμο (με DNA κωδικοποίησης από τον πυρήνα). Τα μεταβολικά ένζυμα έχουν συνδεθεί τα τελευταία χρόνια με τον καρκίνο, και το μιτοχονδριακό μονοπάτι μεταβολισμού έχει βρεθεί ότι στοχεύεται, με στόχο το MTHFD2 από ραδιοφάρμακα για χημειοθεραπείες [50].

RAB33A: Το γονίδιο κωδικοποιεί την πρωτεΐνη Rab-33A, η οποία σχετίζεται με το Ras και ανήκει στην οικογένεια Rab των GTP-εάσων. Συγκεκριμένα για το μελάνωμα έχει βρεθεί ότι το γονίδιο είναι κανονικά εκφρασμένο στα υγιή μελανοκύτταρα, αλλά υποεκφράζεται στο μελάνωμα [51].

Πίνακας 3.12: Βελτιστοποίηση του αλγορίθμου SVM.

	Gene	Accuracy	TrueControl	Mcc
1	CITED1	99.67	0.92	0.92
2	CRYL1	97.60	0.60	0.55
3	DOK5	98.07	0.70	0.67
4	FKBP1B	99.47	0.87	0.87
5	GATA6	99.87	0.97	0.97
6	HMGA2	97.93	0.58	0.55
7	KCNJ13	98.87	0.82	0.79
8	LIPA	99.53	0.88	0.88
9	MTHFD2	100.00	1.00	1.00
10	P2RX7	99.93	0.98	0.98
11	PMAIP1	97.93	0.58	0.55
12	PYCR1	99.93	1.00	0.99
13	RAB33A	100.00	1.00	1.00
14	ZIC1	99.00	0.92	0.87

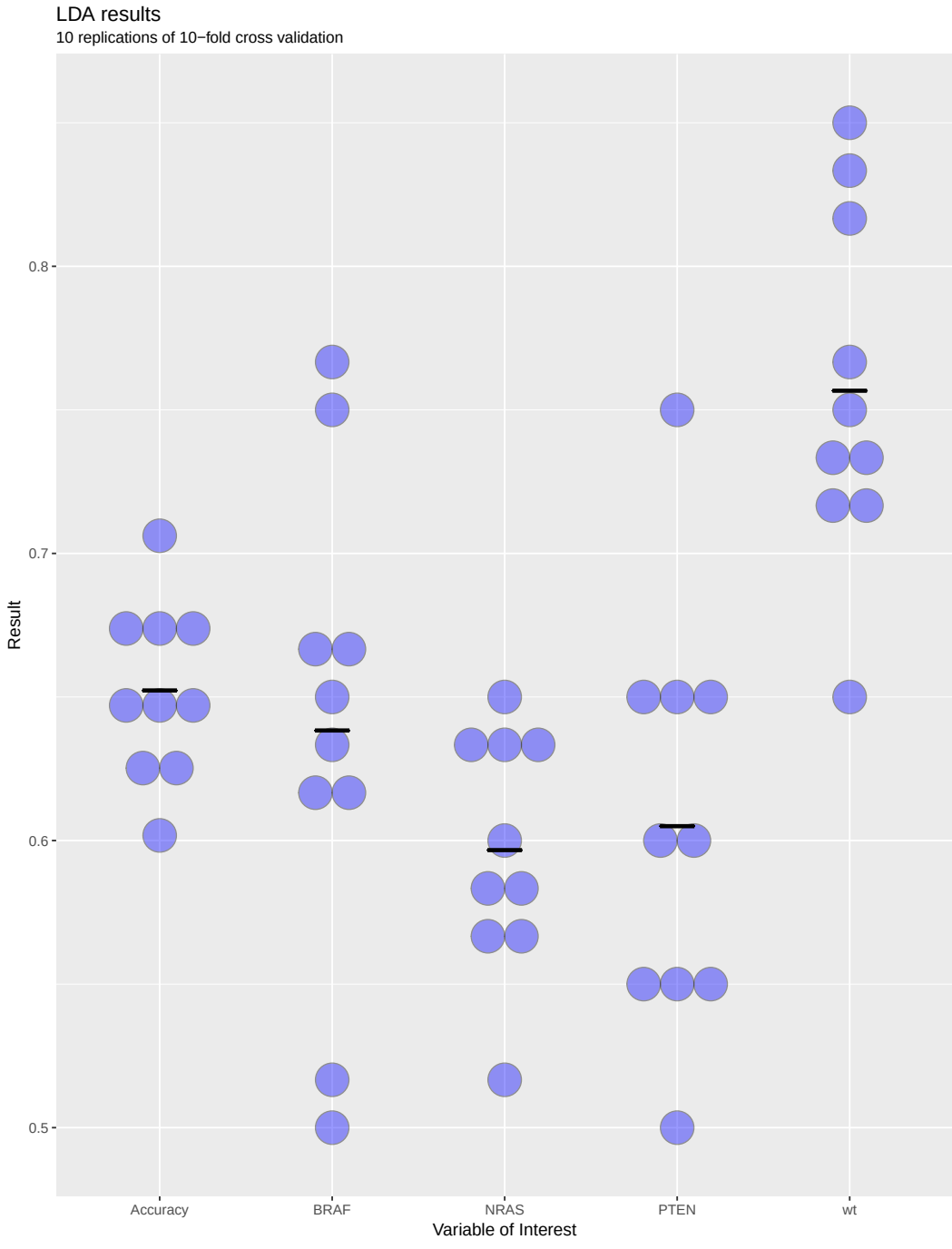
3.4.2 Κατηγοριοποίηση μεταλλάξεων

Το δεύτερο κομμάτι της κατηγοριοποίησης περιλαμβάνει μία προσπάθεια να βρεθούν γονίδια με βάση τα οποία ένας αλγόριθμος εκμάθησης μπορεί να ξεχωρίσει στα καρκινικά δείγματα, αυτά που προέρχονται από διαφορετικές οδηγούς-μεταλλάξεις. Διαφορετική οδηγός-μετάλλαξη σημαίνει ενδεχομένως διαφορετική αντιμετώπιση και θεραπεία για το μελάνωμα, είναι επομένως σημαντικό να βρεθούν βιοδείκτες για την σωστή πρόγνωση. Η εκπαίδευση του αλγορίθμου γίνεται όπως και στην προηγούμενη ενότητα 3.4.1, μέσω της μεθόδου Cross-Validation, ενώ η μέθοδος εκμάθησης είναι η γραμμική διακριτική ανάλυση (LDA). Στην περίπτωση αυτή, τα γονίδια που χρησιμοποιούνται ως features για να γίνει η διακριση, είναι τα τέσσερα διαφορικά εκφρασμένα γονίδια της ενότητας 3.3.2: **BASP1, IFI27, IL24, PTEN**. Τα πρώτα τρία διαφοροποιούν τα δείγματα με τις μεταλλάξεις BRAF, NRAS, ενώ το PTEN διαφοροποιεί τα δείγματα με την ομώνυμη μετάλλαξη με τις άλλες δύο περιπτώσεις. Εκτός από τις κυτταροσειρές με τις τρεις οδηγούς-μεταλλάξεις, στο σύνολο εκμάθησης και ελέγχου συμπεριλαμβάνονται και τα wild-type δείγματα. Στη συνάρτηση που εκτελεί το cross-validation και την LDA, γίνεται ενδιάμεσα αφαίρεση των γονιδίων τα οποία έχουν συντελεστή συσχέτισης κατά Pearson μεγαλύτερο από 0.9, καθώς θεωρείται ότι περιέχουν την ίδια πληροφορία για τη διακριτοποίηση. Καθώς η σειρά με την οποία αναγράφονται τα γονίδια στον κώδικα είναι και η σειρά σημαντικότητάς τους με αυξανόμενη τιμή p, τα γονίδια αφαιρούνται από το τέλος προς την αρχή. Ανάλογα με την προηγούμενη διαδικασία κατηγοριοποίησης, τώρα το σύνολο για να εκπαιδευτεί και να ελεγχθεί ο αλγόριθμος αποτελείται από συνολικά 82 δείγματα από τα 144, 22 από κάθε διαφορετικό τύπο δείγματος BRAF, NRAS, wild-type και τα 16 από τον τύπο PTEN. Επίσης γίνεται resampling 10 φορές, και υπολογίζεται ο μέσος όρος των μεγεθών αξιολόγησης. Από τη στιγμή που δεν υπάρχει ανισορροπία δειγμάτων, η ακρίβεια Acc αποτελεί κατάλληλο μέγεθος για την αξιολόγηση. Ταυτόχρονα υπολογίζεται και το ποσοστό TP για κάθε περίπτωση, δηλαδή πόσα από τα δείγματα κάθε είδους κατηγοριοποιήθηκαν σωστά. Τα αποτελέσματα της εκτέλεσης του προγράμματος φαίνονται στο σχήμα 3.39, το οποίο πληροφορεί όχι μόνο για τη μέση τιμή των μεγεθών, αλλά και την κατανομή των αποτελεσμάτων των 10 επαναλήψεων. Από το σχήμα φαίνεται ότι η ακρίβεια των αποτελεσμάτων, έχει μέση τιμή 0.65 και μικρή τυπική απόκλιση. Σχετικά με την επιτυχία της κατηγοριοποίησης των οδηγών-μεταλλάξεων, οι τιμές είναι μικρότερες. Πιο αναλυτική

σύνοψη γίνεται στον πίνακα 3.13.

Πίνακας 3.13: Σύνοψη αποτελεσμάτων LDA για τα 10 τρεξίματα

	Ακρίβεια (Acc) [%]	BRAF TP [%]	NRAS TP [%]	PTEN TP [%]	Wild-type TP [%]
Μέση τιμή	65.2	63.8	59.7	60.5	75.7
Τυπική απόκλιση	3.06	8.54	4.14	7.25	6.15
Μέγιστη τιμή	70.6	76.7	65.0	75.0	85.0
Ελάχιστη τιμή	60.2	50.0	51.7	50.0	65.0



Σχήμα 3.39: Τα αποτελέσματα του αλγορίθμου εκμάθησης για την κατηγοριοποίηση των δειγμάτων με κριτήριο την οδηγό-μετάλλαξή τους. Στον κατακόρυφο άξονα y φαίνεται η τιμή του μεγέθους αξιολόγησης που αντιστοιχεί στον άξονα x. Οι στήλες BRAF, NRAS, PTEN, και wild-type αντιστοιχούν στο ποσοστό των σωστά αναγνωρισμένων δειγμάτων με την αντίστοιχη οδηγό-μετάλλαξη.

3.5 Επανατοποθέτηση φαρμάκου με το cMap

Τελευταίο κομμάτι της παρούσας διπλωματικής εργασίας είναι η εύρεση ουσιών που δυνητικά αναστρέφουν την επίδραση του μελανώματος στη γονιδιακή έκφραση, και έχουν ως αποτέλεσμα την υποχώρηση της ασθένειας. Το πάνελ των 144 δειγμάτων που έχει δημιουργηθεί, αποτελείται εξόλοκλήρου από κυτταροσειρές. Από αυτές, στο εργαστήριο Εμβιομηχανικής και Βιοτεχνολογίας υπάρχουν διαθέσιμες οι παρακάτω 15: Οι παραπάνω

Πίνακας 3.14: Κυτταροσειρές που βρίσκονται στο εργαστήριο

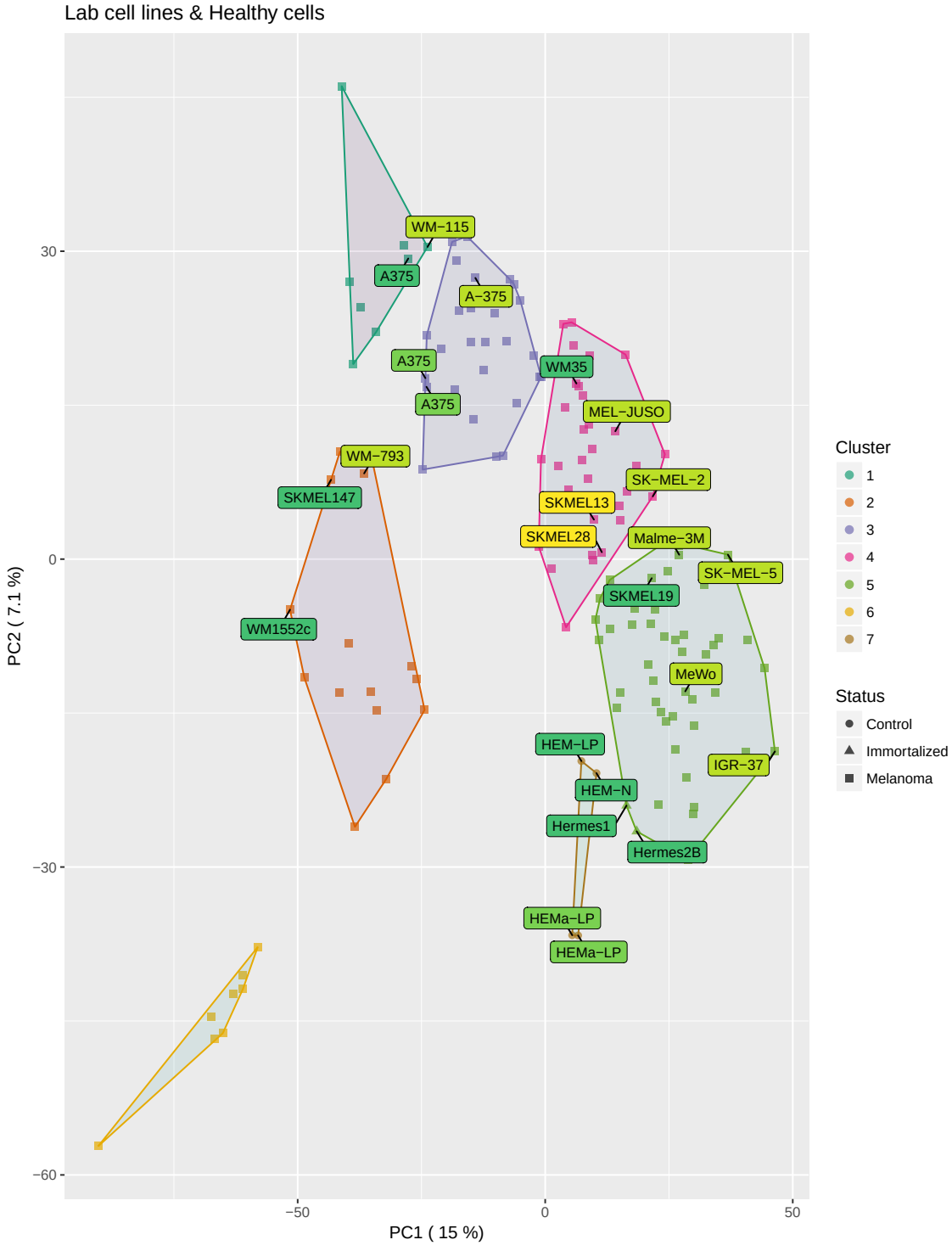
A-375	SK-MEL-13	SK-MEL-5
IGR-37	SK-MEL-147	WM-115
Malme-M	SK-MEL-19	WM-35
MEL-JUSO	SK-MEL-2	WM-1552c
MeWo	SK-MEL-28	WM-793

κυτταροσειρές κατανέμονται στις ομάδες των δειγμάτων του σχήματος 3.22, όπως φαίνεται στο σχήμα 3.40, και στον παρακάτω πίνακα.

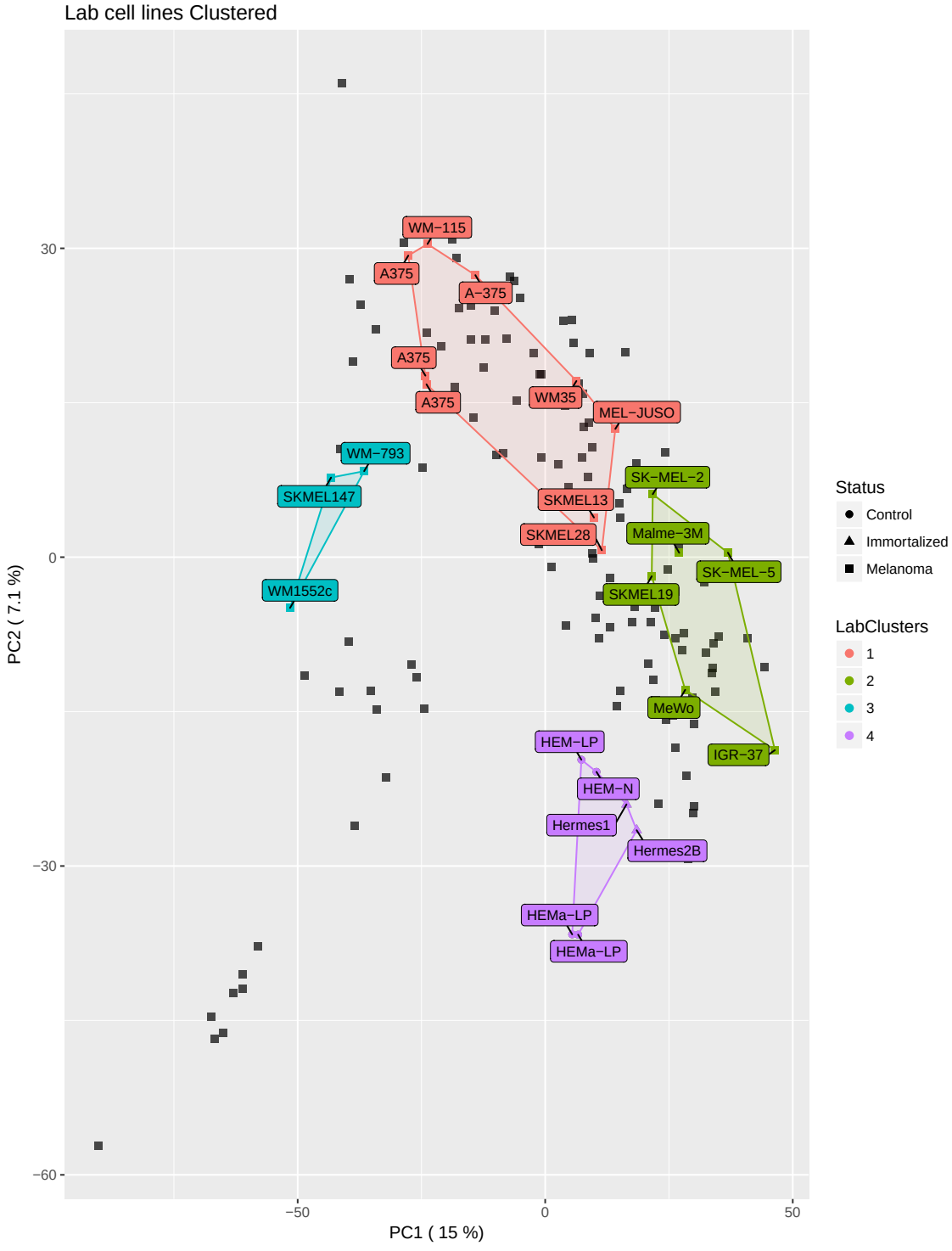
Πίνακας 3.15: Οι κυτταροσειρές του εργαστηρίου κατανεμημένες στις ομάδες του σχήματος 3.22

	Cluster	Cell line
1	1	WM-115
2	1	A375
3	2	WM-793
4	2	SKMEL147
5	2	WM1552c
6	3	A-375
7	3	A375
8	3	A375
9	4	SKMEL13
10	4	SKMEL28
11	4	MEL-JUSO
12	4	SK-MEL-2
13	4	WM35
14	5	IGR-37
15	5	Malme-3M
16	5	MeWo
17	5	SK-MEL-5
18	5	SKMEL19

Στο επόμενο βήμα της ανάλυσης, οι κυτταροσειρές του εργαστηρίου και τα υγιή δείγματα (δηλαδή συνολικά 24 δείγματα) ομαδοποιούνται με την ίδια μέθοδο που ομαδοποιήθηκαν τα 150 δείγματα ολόκληρης της μελέτης, προκειμένου να υπάρξει ξεχωριστή αντιμετώπιση για κάθε καινούργια ομάδα. Το αποτέλεσμα της ομαδοποίησης παρουσιάζεται στο σχήμα 3.41. Η μεγαλύτερη καρκινική ομάδα που δημιουργείται είναι η πρώτη με κόκκινο χρώμα, που περιέχει και τα τέσσερα δείγματα της σειράς A-375 και περιλαμβάνει 9 δείγματα. Η ανισορροπία των δειγμάτων που υπήρχε ανάμεσα σε καρκινικά και υγιή της ενότητας 3.3.1 δεν παρατηρείται εδώ, οπότε τα διαφορικά εκφρασμένα γονίδια για κάθε



Σχήμα 3.40: Οι κυτταροσειρές που υπάρχουν διαθέσιμες στο εργαστήριο, όπως κατανέμονται στις 6 καρκινικές ομάδες. Είναι σημαντικό το γεγονός ότι δεν περιέχεται καμία από αυτές στην ομάδα 6, η οποία φαίνεται να είναι και η πιο έντονα διαχωρισμένη από τις υπόλοιπες. Το χρώμα που πλαισιώνει το όνομα κάθε κυτταροσειράς συμβολίζει το πείραμα από το οποίο αυτή προέρχεται. Είναι σημαντικό να παρατηρηθεί ότι τα δείγματα προέρχονται και από τα τέσσερα διαφορετικά πειράματα.

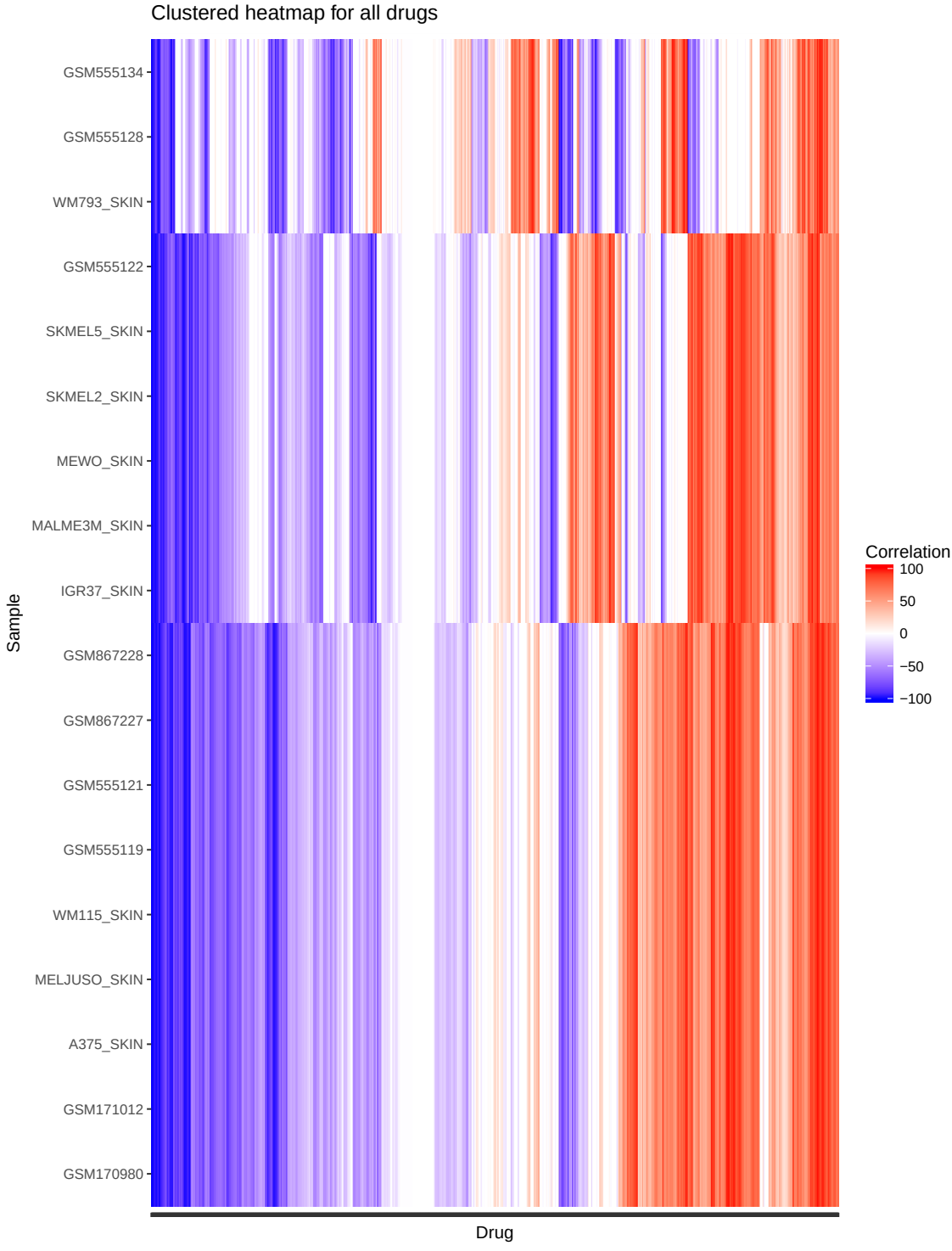


Σχήμα 3.41: Δημιουργούνται τέσσερεις ομάδες (clusters, μία από τις οποίες αποτελείται από τα υγιή δείγματα, μαζί με τα απαθανατισμένα. Η μεγαλύτερη καρκινική ομάδα που δημιουργείται είναι η πρώτη (κόκκινο χρώμα), με 9 δείγματα μελανώματος. Σε αυτή την περίπτωση, η σύγκριση με τις υγιείς κυτταροσειρές είναι ισορροπημένη, για αυτό στην εύρεση των διαφορικά εκφρασμένων γονιδίων δεν απαιτείται η τεχνική του resampling που παρουσιάστηκε στην ενότητα 3.3, αλλά χρησιμοποιείται απλό γραμμικό μοντέλο της ενότητας 2.6

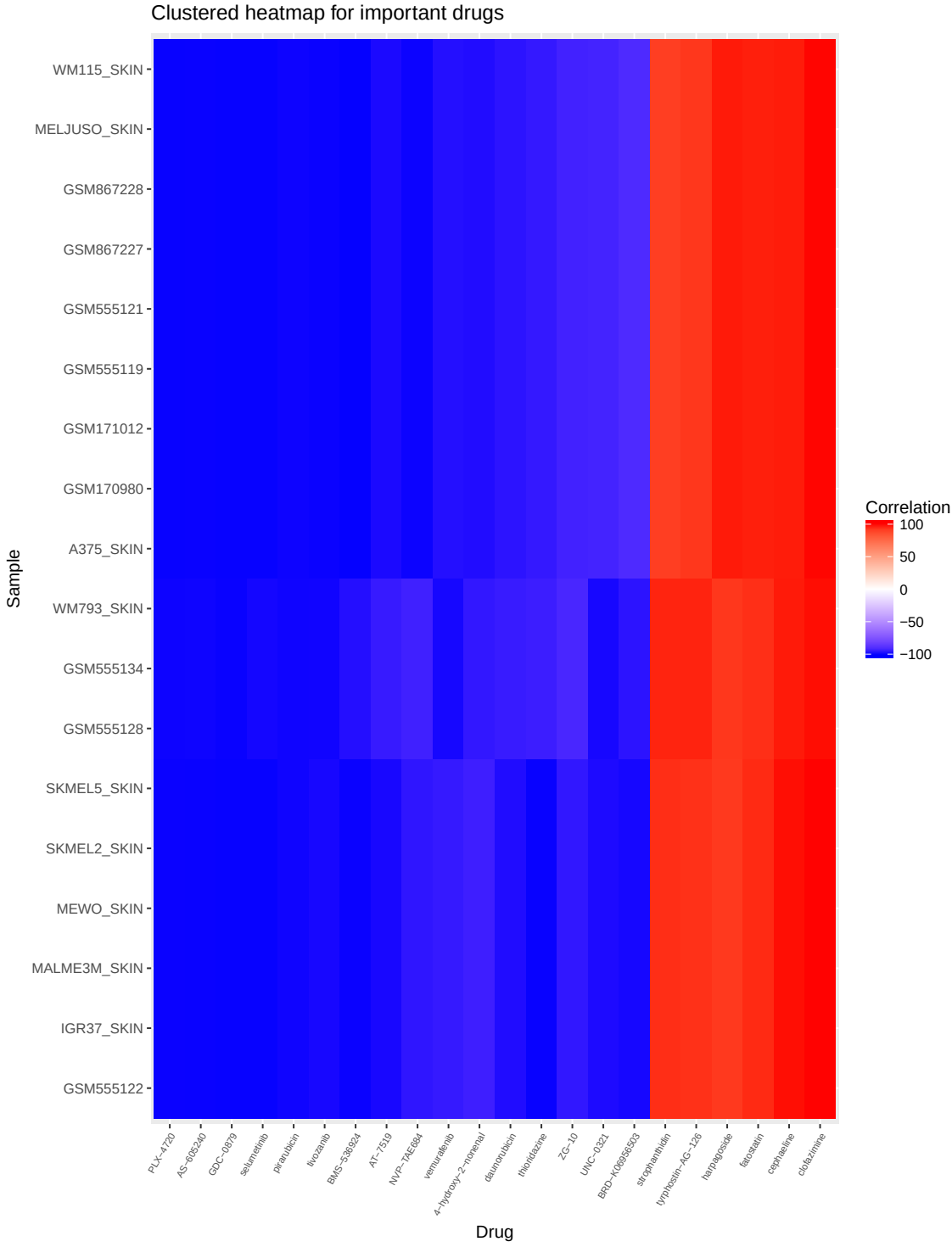
μία από τις 3 καρκινικές ομάδες με τα δείγματα του εργαστηρίου, βρίσκονται με απλό γραμμικό μοντέλο. Τα διαφορικά εκφρασμένα γονίδια διαχωρίζονται σε άνω- και κάτω-εκφρασμένα, με βάση το πρόσημο του $\log(FC)$ τους. Οι δύο κατηγορίες στη συνέχεια στοιχίζονται με αύξουσα σειρά $p - value$ και αποθηκεύονται τα 150 πρώτα γονίδια. Το φιλτράρισμα αυτό είναι αναγκαίο πριν την είσοδο των δεδομένων στο cMap, καθώς το Broad Institute προτείνει ο αριθμός των γονιδίων να είναι το ανώτερο 150 για να προκύψουν αξιόπιστα αποτελέσματα. Καθώς το cMap περιλαμβάνει στην ανάλυση που έχει γίνει, τη σειρά μελανώματος A-375, τα αποτελέσματα βασίζονται σε αυτή, ως η πιο κοντινή στα δείγματα που έχουμε. Για λόγους πληρότητας, έχει γίνει ξεχωριστή ανάλυση μόνο με τα 4 δείγματα της A-375 προκειμένου να δημιουργηθεί και ένα αρνητικό control που να μπορεί να συγκριθεί με τα λοιπά αποτελέσματα.
Από τα αποτελέσματα του cMap διατηρούνται οι ουσίες που έχουν συντελεστή συσχέτισης με τη γονιδιακή έκφραση των δειγμάτων, το πολύ -95%.
Τελικά επομένως, γίνονται 4 queries στο cMap, ένα για την κυτταροσειρά A-375 ξεχωριστά, και 3 που αντιστοιχούν στις 3 ομάδες των κυτταροσειρών του εργαστηρίου. Τα αποτελέσματα, φαίνονται στους πίνακες 4.1 έως 4.4 του παραρτήματος και περιλαμβάνουν εκτός από ουσίες, και τα διαφορετικά είδη μορίων που γενικά σχετίζονται με τη γονιδιακή έκφραση που έχει εισαχθεί. Η έρευνα του Broad Institute περιλαμβάνει 9 διαφορετικές κυτταροσειρές: **A-375, A-549, HA1E, HCC515, HEPG2, HT29, MCF7, PC3, VCAP**.
Απο τη στιγμή που μελετώνται καρκινικά κύτταρα μελανώματος, είναι προφανές πως τα αποτελέσματα της σειράς A-375 είναι τα πιο σχετικά με τα 144 δείγματα της ανάλυσης. Για αυτό το λόγο, ο συντελεστής συσχέτισης που αναγράφεται ως Score στους πίνακες 4.1 έως 4.4 του παραρτήματος, προκύπτει από τη στήλη A-375. Παρ'όλα αυτά, για λόγους πληρότητας, στους πίνακες 4.2 έως 4.4, περιλαμβάνεται και η στήλη με τη διάμεσο της κατανομής των συντελεστών συσχέτισης για τις 9 κυτταροσειρές.
Μετά από επεξεργασία των αποτελεσμάτων, βρέθηκαν οι ουσίες (μαζί με κάποιες κυτταρικές λειτουργίες), οι οποίες είναι παρούσες στις αναλύσεις και των τριών καρκινικών ομάδων. Αυτές φαίνονται στον πίνακα 3.16, και η στήλη MeanScore αντιστοιχεί στο μέσο συντελεστή συσχέτισης ανάμεσα στις 3 ομάδες.
Τα αποτελέσματα είναι ενθαρρυντικά, καθώς παρουσιάζεται με συντελεστή συσχέτισης σχεδόν -100 ο τομέας RAF inhibitor, δηλαδή το σύνολο των ουσιών και μορίων που μπορούν να καταστείλουν τη δράση του γονιδίου BRAF. Όπως έχει αναφερθεί, τα περισσότερα μελανώματα χαρακτηρίζονται από μεταλλαγμένο BRAF γονίδιο, το οποίο είναι υπερεκφρασμένο και ενεργοποιεί το σηματοδοτικό μονοπάτι MAPK το οποίο δίνει σήμα στο κύτταρο για ταχύτερο διπλασιασμό και ανάπτυξη [52]. Παράλληλα, παρατηρούνται ουσίες που καταστέλλουν τη δράση των MAP κινάσων, όπως το φάρμακο selumetinib, που στοχεύει και απενεργοποιεί κατάντι του BRAF τα γονίδια MAP2K1, MAP2K2 [53]. Επίσης, υπάρχουν καταστολείς του PI3K το οποίο αποτελεί το κέντρο του σηματοδοτικού μονοπατιού PI3K/AKT/mTOR [54] που παίζει σημαντικό ρόλο στη ρύθμιση του κυτταρικού κύκλου, και κατέπέκταση αποτελεί σημείο ενδιαφέροντος για ασθένειες όπως ο καρκίνος [55]. Αξίζει να σημειωθεί ότι ένας φυσικός καταστολέας του PI3K είναι η πρωτεΐνη του γονιδίου PTEN. Στο μελάνωμα, και ιδιαίτερα στην περίπτωση που υπάρχει μετάλλαξη του γονιδίου PTEN, αυτό υποεκφράζεται [44]. Είναι λογικό να χρειάζεται ένας εξωτερικός PI3K καταστολέας, όπως ο AS-605240 που προτείνεται από το cMap. Το TOP2A που κωδικοποιεί το ένζυμο *τοποϊσομεράση 2-άλφα* βοηθάει στην χαλάρωση της τάσης που δημιουργείται από την αντιγραφή του DNA, και έχει βρεθεί ότι η υπερέκφρασή του αποτελεί βιοδείκτη για την ανάπτυξη μελανώματος [56].

Για λόγους πληρότητας, κατασκευάστηκε σύντομος κώδικας με βάση τον οποίο γίνεται δυνατή η απεικόνιση των αποτελεσμάτων του cMap μετά από ιεραρχική ομαδοποίηση

τόσο των κυτταροσειρών του εργαστηρίου, όσο και των ουσιών που βρέθηκαν, με βάση τον πίνακα που περιέχει τους συντελεστές συσχέτισης δείγματος-ουσίας, με τα δείγματα στις γραμμές και τις ουσίες στις στήλες.



Σχήμα 3.42: Τα αποτελέσματα του cMap για όλες τις ουσίες που περιλαμβάνει. Τα δείγματα πρακτικά είναι χωρισμένα στις 3 ομάδες του σχήματος 3.41, οπότε τα αποτελέσματα είναι κοινά για τα δείγματα της ίδιας ομάδας. Τόσο οι ουσίες όσο και οι κυτταροσειρές έχουν διαταχθεί κατάλληλα μετά από ιεραρχική ομαδοποίηση.



Σχήμα 3.43: Τα αποτελέσματα του cMap για τις ουσίες που έχουν απόλυτο συντελεστή συσχέτισης μεγαλύτερο του 90. Τα δείγματα πρακτικά είναι χωρισμένα στις 3 ομάδες του σχήματος 3.41, οπότε τα αποτελέσματα είναι κοινά για τα δείγματα της ίδιας ομάδας. Τόσο οι ουσίες όσο και οι κυτταροσειρές έχουν διαταχθεί κατάλληλα μετά από ιεραρχική ομαδοποίηση.

Πίνακας 3.16: Ουσίες που αντιστρέφουν τη γονιδιακή έκφραση και στις 3 ομάδες δειγμάτων.

	Name	Description	Target	MeanScore
1	GDC-0879	RAF inhibitor	BRAF	-99.71
2	PLX-4720	RAF inhibitor	BRAF, KDR	-99.58
3	AS-605240	Phosphatidylinositol 3-kinase (PI3K) inhibitor, PI3K inhibitor	MAOB, PIK3CA, PIK3CB, PIK3CD, PIK3CG	-99.56
4	selumetinib	MEK inhibitor, MAP kinase inhibitor	MAP2K1, MAP2K2	-99.51
5	pirarubicin	topoisomerase inhibitor	TOP2A	-99.40
6	tivozanib	VEGFR inhibitor, KIT inhibitor, tyrosine kinase inhibitor	FLT1, FLT4, KDR, KIT, PDGFRA, PDGFRB	-99.28
7	RAF inhibitor			-99.09
8	BMS-536924	insulin growth factor receptor inhibitor, insulin receptor ligand	IGF1R, AKT1, CCNE1, CDK2, CYP3A4, ERBB2, INSR, KDR, LCK, MAPK1, MET, PDGFRA, PDGFRB	-99.05
9	tozasertib	Aurora kinase inhibitor, Bcr-Abl kinase inhibitor, FLT3 inhibitor, JAK inhibitor, Abl kinase inhibitor, mitotic inhibitor	AURKA, AURKB, ABL1, AURKC, BCR, FLT3, JAK2, DDR2, LCK	-98.97
10	GW-5074	RAF inhibitor, leucine rich repeat kinase inhibitor	LRRK1, LRRK2, NTRK1, RAF1	-98.01
11	daunorubicin	RNA synthesis inhibitor, topoisomerase inhibitor, DNA synthesis inhibitor, radical formation stimulant	TOP2A, TOP2B	-97.76
12	AT-7519	CDK inhibitor, cell cycle inhibitor	CDK2, CDK5, CDK1, CDK4, CDK6, CDK9	-97.47
13	vemurafenib	RAF inhibitor, protein kinase inhibitor	BRAF, CYP2C19, CYP3A4, CYP3A5, RAF1	-97.35
14	simvastatin	HMGCR inhibitor	HMGCR, CYP2C8, CYP3A4, CYP3A5, ITGB2	-97.29

Κεφάλαιο 4

Συμπεράσματα και μελλοντική εργασία

Η παρούσα διπλωματική εργασία αποτελεί μία ολοκληρωμένη μελέτη δεδομένων γονιδιακής έκφρασης που προέρχονται από μικροσυστοιχίες DNA. Αρχικά, αναπτύσσεται κώδικας ο οποίος επιτρέπει την ενσωμάτωση δεδομένων από πολλά πειράματα σε ένα εννιαίο πάνελ το οποίο μπορεί να χρησιμοποιηθεί για την κάλυψη πολλών διαφορετικών περιπτώσεων. Στην ανάλυση ενσωματώθηκαν τέσσερα διαφορετικά πειράματα, αλλά με βάση τον κατασκευασμένο κώδικα είναι δυνατό, αφού πρώτα ελεγχθεί με τη μέθοδο της ανάλυσης σε κύριες συνιστώσες ότι υπάρχει διαχωρισμός με βάση καποιο συστατικό του πειράματος (εδώ *η πλατφόρμα*), να ενσωματωθούν περισσότερα δεδομένα. Κατά την ενσωμάτωση υπάρχει η δυνατότητα να διατηρηθούν διαφορές οι οποίες δεν οφείλονται σε συστηματικό σφάλμα και πρέπει να διατηρηθούν, όπως για παράδειγμα οι βιολογικές διαφορές ανάμεσα σε καρκινικά και υγιή κύτταρα. Ώστόσο, στην παρούσα ανάλυση δεν έγινε διόρθωση με παράμετρο τη διατήρηση αυτών των στοιχείων, καθώς θεωρείται ότι μπορεί να γίνει υπερεκτίμηση των διαφορών και να εισαχθεί σφάλμα (bias) στα δεδομένα, και να επηρεαστεί η κατάντι ανάλυση [57].
Όσον αφορά στη διαφορική γονιδιακή ανάλυση, αυτή έγινε σε πρώτο επίπεδο, δηλαδή σε επίπεδο γονιδίων. Το επόμενο στάδιο μιας κλασικής διαφορικής ανάλυσης είναι η ανάλυση μονοπατιών, που περιλαμβάνει ουσιαστικά τον υπολογισμό των τιμών p κάθε βιολογικού μονοπατιού από τα επιμέρους γονίδια που αυτό περιέχει [58]. Στην εργασία δεν έγινε ανάλυση σε επίπεδο μονοπατιών, καθώς για την κατηγοριοποίηση και για την αναζήτηση γονιδιακών εκφράσεων στο cMap απαιτούνται γονίδια και όχι μονοπάτια. Με τα αποτελέσματα και τους κώδικες που κατασκευάστηκαν, μελλοντική εργασία μπορεί εύκολα να περιλάβει και την διαφορική ανάλυση μονοπατιών. Τα σημαντικότερα μονοπάτια που είναι γνωστό ότι συμμετέχουν στον καρκίνο, φαίνεται να είναι διαφοροποιημένα και στα καρκινικά δείγματα που χρησιμοποιήθηκαν, αφού στα αποτελέσματα του cMap υπάρχει πλήθος καταστολέων αυτών των μονοπατιών, όπως το MAPK\ERK και το PI3K/AKT/mTOR. Για την ανάλυση χρησιμοποιήθηκε η βασική μέθοδος resampling για να ελαχιστοποιηθεί το στατιστικό σφάλμα από άνισους πλυθησμούς. Οι συνθήκες ορίου για τις 1000 επαναλήψεις βασίστηκαν στα αντίστοιχα διαγράμματα και την κρίση του συγγραφέα, χωρίς να υπάρχει κάποια ντετερμινιστική μέθοδος για την επιλογή των ορίων 260, 540, και 800 επαναλήψεων.

Βιβλιογραφία

[1] Consortium, International Human Genome Sequencing: *Initial sequencing and analysis of the human genome*. Nature, 409:860—-921, 2001.

[2] Καψάλης, Α., Μπουρμπουχάκης Ι.Ε., Περάκη Β., και Σαλαμαστράκης Σ.: *Βιολογία γενικής παιδείας, Β΄ τάξης γενικού λυκείου*. Οργανισμός Εκδόσεων Διδακτικών Βιβλίων, 2016.

[3] Αιμιλία, Ζ.: *Διακυτταρική Επικοινωνία - Μεταγωγή σήματος*. 2018. http://eclass.uth.gr/eclass/courses/SEYC101/.

[4] Govindarajan, R., Duraiyan J., Kaliyappan K., and Palanisamy M.: *Microarray and its applications*. Journal of Pharmacy and Bioallied Sciences, Suppl 2:S310–S312, 2012. doi:10.4103/0975-7406.100283.

[5] Fink, D.J.: *Cancer overview*. Cancer Research, 39(7 Part 2):2819–2821, 1979, ISSN 0008-5472. http://cancerres.aacrjournals.org/content/39/7_Part_2/2819.

[6] Veer, L.J. Van 't, Burgering B.M., and Versteeg et al. R.: *N-ras mutations in human cutaneous melanoma from sun-exposed body sites*. Molecular and Cellular Biology, 9(7):3114–3116, 1989.

[7] Hodis, E., Watson I.R., and Kryukov et al. G.V.: *A landscape of driver mutations in melanoma*. Cell, 150(2):251–263, July 2012.

[8] Davies, H, Bignell G.R., Stephens P., Edkins S., Teague J., Woffendin H., Garnett M.J., Bottomley W., Davis N., Dicks E., Ewing R., Floyd Y., Gray K., Hall S., Hawes R., Hughes J., Kosmidou V., Menzies A., Mould C., Parker A., and Stevens et al. C.: *Mutations of the braf gene in human cancer*. Nature, 419:949—-954, 2002.

[9] Kwong, L.N., Costello J.C., Liu H., Jiang S., Helms T.L., Langsdorf A.E., Jakubosky D., Genovese G., Muller F.L., Jeong J.H., Bender R.P., Chu G.C., Flaherty K.T., Wargo J.A., Collins J.J., and Chin L.: *Oncogenic nras signaling differentially regulates survival and proliferation in melanoma*. Nature Medicine, 18:1503–1510, 2012.

[10] Muñoz-Couselo, E., Adelantado E.Z., Ortiz C., García J.S., and Perez Garcia J.: *Nras-mutant melanoma: current challenges and future prospect*. OncoTargets and therapy, 10:3941–3947, August 2017.

[11] Ball, N.J., Yohn J.J., Morelli J.G., Norris D.A., Golitz L.E., and Hoeffler J.P.: *Ras mutations in human melanoma: A marker of malignant progression*. Journal of Investigative Dermatology, 102(3):285 - 290, 1994, ISSN 0022-202X. http://www.sciencedirect.com/science/article/pii/S0022202X94976252.

[12] Aguissa-Touré, Almass-Houd, and Li G.: *Genetic alterations of pten in human melanoma*. Cellular and Molecular Life Sciences, 69(9):1475–1491, May 2012, ISSN 1420-9071. https://doi.org/10.1007/s00018-011-0878-0.

[13] Lovly, C., Pao W., and Sosman J.: *Mek1 (map2k1) in melanoma*, 2015. https://www.mycancergenome.org/content/disease/melanoma/map2k1/.

[14] R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. http://www.R-project.org, ISBN 3-900051-07-0.

[15] Huber, W., Carey V.J., Gentleman R., Anders S., Carlson M., Carvalho B.S., Bravo H.C., Davis S., Gatto L., Girke T., Gottardo R., Hahne F., Hansen K.D., Irizarry R.A., Lawrence M., Love M.I., MacDonald J., Obenchain V., Ole's A.K., Pag'es H., Reyes A., Shannon P., Smyth G.K., Tenenbaum D., Waldron L., and Morgan M.: *Orchestrating high-throughput genomic analysis with Bioconductor*. Nature Methods, 12(2):115–121, 2015. http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3252.html.

[16] Wickham, H.: *tidyverse: Easily Install and Load the 'Tidyverse'*, 2017. https://CRAN.R-project.org/package=tidyverse, R package version 1.2.1.

[17] Wickham, H.: *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009, ISBN 978-0-387-98140-6. http://ggplot2.org.

[18] Davis, S. and Meltzer P.: *Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor*. Bioinformatics, 14:1846–1847, 2007.

[19] Gautier, L., Cope L., Bolstad B.M., and Irizarry R.A.: *affy—analysis of affymetrix genechip data at the probe level*. Bioinformatics, 20(3):307–315, 2004, ISSN 1367-4803.

[20] Leek, J.T., Johnson W.E., Parker H.S., Jaffe A.E., and Storey J.D.: *The sva package for removing batch effects and other unwanted variation in high-throughput experiments*. Bioinformatics, 28(6):882–883, 2012, ISSN 0008-5472. doi:10.1093/bioinformatics/bts034.

[21] Ritchie, M.E., Phipson B., Wu D., Hu Y., Law C.W., Shi W., and Smyth G.K.: *limma powers differential expression analyses for RNA-sequencing and microarray studies*. Nucleic Acids Research, 43(7):e47, 2015.

[22] Meyer, D., Dimitriadou E., Hornik K., Weingessel A., and Leisch F.: *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2017. https://CRAN.R-project.org/package=e1071, R package version 1.6-8.

[23] Venables, W.N. and Ripley B.D.: *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. http://www.stats.ox.ac.uk/pub/MASS4, ISBN 0-387-95457-0.

[24] Irizarry, R.A., Hobbs B., Collin F., Beazer-Barclay Y.D., Antonellis K.J., Scherf U., and Speed T.P.: *Exploration, normalization, and summaries of high density oligonucleotide array probe level data*. Biostatistics, 4(2):249–264, 2003. http://dx.doi.org/10.1093/biostatistics/4.2.249.

[25] Irizarry, R.A., Bolstad B.M., Collin F., Cope L.M., Hobbs B., and Speed T.P.: *Summaries of affymetrix genechip probe level data*. Nucleic Acids Research, 31(4):e15, 2003.

[26] Wu, Z., Irizarry R.A., Gentleman R., Martinez Murillo F., and Spencer F.: *A model-based background adjustment for oligonucleotide expression arrays*. Journal of the American Statistical Association, 99(468):909–917, December 2004.

[27] Bolstadt, B.M.: *Low-level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization*. PhD thesis, University of California, Berkeley, Spring 2004.

[28] Bolstadt, B.M., Irizarry R.A., Astrand M., and Speed T.P.: *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias*. Bioinformatics, 19(2):185–193, 2003, ISSN 0008-5472. https://www.ncbi.nlm.nih.gov/pubmed/12538238#.

[29] Tykey, J.W.: *Exploratory data analysis*. Reading, Mass: Addison-Wesley Pub. Co, 1977.

[30] Πετρίδης, Δ.: *ΑΝΑΛΥΣΗ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ - ΠΑΡΑΓΟΝΤΙΚΗ ΑΝΑΛΥΣΗ. [Κεφάλαιο Συγγράμματος]. Στο Πετρίδης, Δ. 2015. Ανάλυση πολυμεταβλητών τεχνικών*, κεφάλαιο 4. Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, 2015. http://hdl.handle.net/11419/2129.

[31] Madsen, R.E., Hansen L.K., and Winther O.: *Singular value decomposition and principal component analysis*. Technical report, 2004. http://www2.imm.dtu.dk/pubdb/p.php?4000.

[32] Smith, L.I.: *A tutorial on principal components analysis (computer science technical report no. oucs-2002-12)*. Technical report, University of Otago, May 2002. http://hdl.handle.net/10523/7534.

[33] Cheplyaka, R.: *Explained variance in pca*.

[34] Gibbons, S.M., Duvallet C., and Alm E.J.: *Correcting for batch effects in case-control microbiome studies*. PLOS Computational Biology, 14(4):1–17, April 2018. https://doi.org/10.1371/journal.pcbi.1006102.

[35] Johnson, W.E., Li C., and Rabinovic A.: *Adjusting batch effects in microarray expression data using empirical bayes methods*. Biostatistics, 8(1):118–127, 2007. http://dx.doi.org/10.1093/biostatistics/kxj037.

[36] Νικολάου, Χ. και Π., Χουβαρδάς: *Υπολογιστική βιολογία*, chapter κεφ. 9 Λειτουργική Ανάλυση της Γονιδιακής Έκφρασης. Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, 2015. http://hdl.handle.net/11419/1586.

[37] Smyth, G.K.: *Linear models and empirical bayes methods for assessing differential expression in microarray experiments*. Statistical Applications in Genetics and Molecular Biology, 3(1), 2004.

[38] Lamb, J., Crawford E., Peck D., Modell J.W., Blat I.C., Wrobel M.J., Lerner J., Brunet J.P., Subramanian A., Ross K.N., Reich M., Hieronymus H., Wei G., Armstrong S.A., and Haggarty et al. S.J.: *The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease*. Science, 313(5795):1929–1935, 2006, ISSN 0036-8075. http://science.sciencemag.org/content/313/5795/1929.

[39] Cortes, C. and Vapnik V.: *Support-vector networks*. Machine Learning, 20(3):273–297, Sep 1995, ISSN 1573-0565. https://doi.org/10.1007/BF00994018.

[40] Xu, Guangru, Minghui Zhang, Hongxing Zhu, and Jinhua Xu: *A 15-gene signature for prediction of colon cancer recurrence and prognosis based on svm*. Gene, 604:33

- 40, 2017, ISSN 0378-1119. http://www.sciencedirect.com/science/article/pii/S0378111916309854.

[41] Berwick, R.: *An idiot's guid to support vector machines (svms).* Technical report.

[42] Johansson, P., Pavey S., and Hayward N.: *Confirmation of a braf mutation-associated gene expression signature in melanoma.* Pigment Cell Research, 20(3):216–221. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0749.2007.00375.x.

[43] Barretina, J., Caponigro G., Stransky N., Venkatesan K., Margolin A.A., and S. Kim et al.: *The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature, 483:603–607, 2012.

[44] Rose, A.E., Poliseno L., Wang J., Clark M., Pearlman A., and Wang et al. G.: *Integrative genomics identifies molecular alterations that challenge the linear model of melanoma progression.* Cancer Research, 71(7):2561–2571, 2011, ISSN 0008-5472. http://cancerres.aacrjournals.org/content/71/7/2561.

[45] Xiao, D., Ohlendorf J., Chen Y., Taylor D.D., Rai S., Waigel S., Zacharias W., Hao H., and McMasters K.M.: *Identifying mrna, microrna and protein profiles of melanoma exosomes.* PLOS ONE, 7(10):1–15, October 2012. https://doi.org/10.1371/journal.pone.0046874.

[46] Cheng, L., Lopez Beltran A., Massari F., MacLennan G.T., and Montironi R.: *Molecular testing for braf mutations to inform melanoma treatment decisions: a move toward precision medicine.* Modern Pathology, 31(1):24–38, November 2018.

[47] Fedele, M., Pierantoni G.M., Visone R., and Fusco A.: *Critical role of the hmga2 gene in pituitary adenomas.* Cell cycle, 5:2045–2048, October 2006.

[48] Raskin, L., Fullen D.R., Giordano T.J., Thomas D.G., Frohm M.L., Cha K.B., Ahn J., Mukherjee B., Johnson T.M., and Gruber S.B.: *Transcriptome profiling identifies hmga2 as a biomarker of melanoma progression and prognosis.* Journal of Investigative Dermatology, 133(11):2585 - 2592, 2013, ISSN 0022-202X. http://www.sciencedirect.com/science/article/pii/S0022202X15360231.

[49] Nickoloff, B.J., Hendrix M., Pollock P.M., Trent J.M., Miele L., and J. Qin": *Notch and noxa-related pathways in melanoma cells.* Journal of Investigative Dermatology Symposium Proceedings, 10(2):95 - 104, 2005, ISSN 1087-0024. http://www.sciencedirect.com/science/article/pii/S0022202X15525700.

[50] Nilsson, R., Mohit J., Madhusudhan N., Sheppard N.G., Strittmatter L., Kampf C., Huang J., Asplund A., and Mootha V.K.: *Metabolic enzyme expression highlights a key role for mthfd2 and the mitochondrial folate pathway in cancer.* Nature Communications, 5, 2014.

[51] Cheng, E., Trombeta S.E., Kovacs D., Beech R.D., Ariyan S., Reyes Mugica M., McNiff J.M., Narayan D., Kluger H.M., Picardo M., and Halaban R.: *Rab33a: Characterization, expression, and suppression by epigenetic modification.* Journal of Investigative Dermatology, 126(10):2257 - 2271, 2006, ISSN 0022-202X. http://www.sciencedirect.com/science/article/pii/S0022202X15326634.

[52] Curtin, J.A., Fridlyand J., Kageshita T., Patel H.N., Busam K.J., Kutzner H., Cho K.H., Aiba S., Bröcker E.B., LeBoit P.E., Pinkel D., and Bastian B.C.: *Distinct sets of genetic alterations in melanoma.* New England Jour-

nal of Medicine, 353(20):2135–2147, 2005. https://doi.org/10.1056/NEJMoa050092, PMID: 16291983.

[53] Shtivelman, E., Davies M.A., and Hwu et al. P.: *Pathways and therapeutic targets in melanoma*. Oncotarget, 5(7):1701–1752, 2014.

[54] Paluncic, J., Kovacevic Z., and Jansson et al. P.J.: *Roads to melanoma: Key pathways and emerging players in melanoma progression and oncogenic signaling*. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research, 1863(4):770 – 784, 2016, ISSN 0167-4889. http://www.sciencedirect.com/science/article/pii/S0167488916300155.

[55] Palmieri, G., Ombra M., Colombino M., Casula M., Sini M.C., Manca A., Paliogiannis P., Ascierto P.A., and Cossu A.: *Multiple molecular pathways in melanomagenesis: Characterization of therapeutic targets*. Frontiers in Oncology, 5:183, 2015, ISSN 2234-943X. https://www.frontiersin.org/article/10.3389/fonc.2015.00183.

[56] Song, L.iang, Robson T., Doig T., Brenn T., Mathers M., Brown E.R., Doherty V., Bartlett J.M.S., Anderson N., and Melton D.W.: *Dna repair and replication proteins as prognostic markers in melanoma*. Histopathology, 62(2):343–350, 2013. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2559.2012.04362.x.

[57] Nygaard, Vegard, Einar Andreas Rødland, and Eivind Hovig: *Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses*. Biostatistics, 17(1):29–39, 2016. http://dx.doi.org/10.1093/biostatistics/kxv027.

[58] Hess, A. and Iyer H.: *Fisher's combined p-value for detecting differentially expressed genes using affymetrix expression arrays*. BMC Genomics, 8(1):96, Apr 2007. https://doi.org/10.1186/1471-2164-8-96.

Παράρτημα

Πίνακας 4.1: Αποτελέσματα για τα δείγματα A-375

	Name	Description	Target	Score
1	Bromodomain Inhibitor			-99.46
2	atenolol	adrenergic receptor antagonist	ADRB1, ADRB2	-99.45
3	efavirenz	HIV reverse transcriptase inhibitor, non-nucleoside reverse transcriptase inhibitor, reverse transcriptase inhibitor	CYP2B6, CYP2C19, CYP2C8, CYP3A4, CYP3A5	-99.32
4	AG-490	epidermal growth factor receptor (EGFR) inhibitor, ErbB2 and JAK2 inhibitor, JAK inhibitor	JAK2, JAK3, EGFR, STAT3	-99.27
5	3-matida	glutamate receptor antagonist	GRM1	-99.11
6	SB-590885	RAF inhibitor	BRAF	-99.10
7	JAK inhibitor			-99.06
8	bisindolylmaleimide	CDK inhibitor, PKC inhibitor, leucine rich repeat kinase inhibitor	CCND1, CDK4, LRRK2, PDPK1, PIM1, PRKCA, PRKCB, PRKCI, PRKCZ	-98.98
9	SB-202190	p38 MAPK inhibitor, interleukin inhibitor, stress activated protein kinase inhibitor	MAPK14, AKT1, ALOX5, CHEK1, GSK3B, LCK, MAPK1, MAPK11, MAPK12, MAPK8, PRKCA, ROCK1, RPS6KB1, SGK1	-98.88
10	NVP-AUY922	HSP inhibitor	HSP90AA1, HSP90AA2, HSP90AB1	-98.85
11	BI-2536	PLK inhibitor, apoptosis stimulant, cell cycle inhibitor, protein kinase inhibitor	PLK1, BRD4, PLK2, PLK3	-98.83
12	PF-543	sphingosine kinase inhibitor	SPHK1	-98.79

13	nimodipine	calcium channel blocker, L-type calcium channel blocker	CACNA1C, NR3C2, AHR, CACNA1D, CACNA1F, CACNA1S, CACNB1, CACNB2, CACNB3, CACNB4, CFTR	-98.71
14	deltaline	acetylcholine receptor antagonist	CHRNA7	-98.63
15	azacyclonol	ataractive drug used to diminish hallucinations	HRH1	-98.56
16	TG-101348	JAK inhibitor, FLT3 inhibitor, RET tyrosine kinase inhibitor	JAK2, FLT3, BRD4, JAK1, JAK3, RET, TYK2	-98.55
17	STO-609	calcium/calmodulin dependent protein kinase inhibitor, calmodulin inhibitor	CAMKK1, CAMKK2	-98.55
18	tetrahydropalmatine	serotonin release inhibitor		-98.37
19	BIBU-1361	EGFR inhibitor	EGFR	-98.34
20	PKC inhibitor			-98.28
21	MG-132	proteasome inhibitor	PSMB1	-98.28
22	midostaurin	FLT3 inhibitor, KIT inhibitor, PKC inhibitor, angiogenesis inhibitor, cell cycle inhibitor, cyclin inhibitor, histamine release inhibitor, multi targeted kinase inhibitor, PDGFR tyrosine kinase receptor inhibitor, VEGFR antagonist, VEGFR inhibitor	FLT3, KIT, CCNB1, FLT1, KDR, PDGFRB, PRKCA, PRKCG, VEGFA	-98.22
23	eriochrome-black-t	azo dye used in titrations to detect metal ions		-98.16
24	FLT3 inhibitor			-98.15
25	givinostat	HDAC inhibitor, interleukin receptor antagonist, interleukin synthesis inhibitor, tumor necrosis factor receptor antagonist, tumor necrosis factor release inhibitor	HDAC2, HDAC1, HDAC3, HDAC4, HDAC5, HDAC6, HDAC7, HDAC8, HDAC9, IL1B, IL1R2, IL6R, TNF	-98.12
26	benzatropine	anticholinergic	CHRM1, HRH1, SLC6A3	-98.09
27	U-0126	MEK inhibitor, JAK inhibitor, MAP kinase inhibitor	AKT1, CHEK1, GSK3B, JAK2, LCK, MAP2K1, MAP2K2, MAP2K7, MAPK1, MAPK11, MAPK12, MAPK14, MAPK8, PRKCA, RAF1, ROCK1, RPS6KB1, SGK1	-97.99
28	FOS transcription factor family GOF			-97.95
29	PU-H71	HSP inhibitor	HSP90AA1	-97.68

30	isoreserpine	vesicular monoamine transporter inhibitor	SLC18A2, SLC18A1, SIAH1	-97.44
31	PD-184352	MEK inhibitor, MAP kinase inhibitor	MAP2K1, MAP2K2, MAP3K1, MAP3K2	-97.39
32	scriptaid	HDAC inhibitor	HDAC1, HDAC2, HDAC3, HDAC4, HDAC5, HDAC6, HDAC7, HDAC8, HDAC9	-97.30
33	IB-MECA	adenosine receptor agonist, granulocyte colony stimulating factor agonist	ADORA3, ADORA1, ADORA2A, ADORA2B	-97.24
34	trichostatin-a	HDAC inhibitor, CDK expression enhancer, ID1 expression inhibitor	HDAC7, HDAC8, HDAC1, HDAC10, HDAC2, HDAC3, HDAC4, HDAC5, HDAC6, HDAC9	-97.16
35	SA-792987	PKC inhibitor	WEE1	-97.10
36	GDC-0879	RAF inhibitor	BRAF	-96.93
37	Proteasome inhibitor			-96.83
38	SB-218078	CHK inhibitor, PKC inhibitor	CHEK1	-96.80
39	panobinostat	HDAC inhibitor, apoptosis stimulant, cell cycle inhibitor	HDAC1, HDAC2, HDAC3, HDAC4, HDAC6, HDAC7, HDAC8, HDAC9	-96.72
40	meropenem	cell wall synthesis inhibitor		-96.69
41	XMD-885	leucine rich repeat kinase inhibitor, MAP kinase inhibitor	LRRK2, MAPK7	-96.65
42	geldanamycin	HSP inhibitor	HSP90AA1, HSP90AB1	-96.63
43	JAK3-Inhibitor-II	JAK inhibitor, ALK tyrosine kinase receptor inhibitor, EGFR inhibitor	EGFR, ALK, JAK1, JAK2, JAK3	-96.50
44	GATA Zinc finger domain containing LOF			-96.42
45	CS-110266	dopamine receptor agonist	SLC6A3	-96.26
46	pirarubicin	topoisomerase inhibitor	TOP2A	-96.07
47	PLX-4720	RAF inhibitor	BRAF, KDR	-95.85
48	tyrphostin-AG-825	receptor tyrosine protein kinase inhibitor	ERBB2	-95.72
49	BRD-K06956503	glucosylceramidase inhibitor	GBA	-95.60
50	RS-17053	adrenergic receptor antagonist	ADRA1A, ADRA1D	-95.59

51	mirtazapine	adrenergic receptor antagonist, serotonin receptor antagonist	ADRA2A, HTR2A, HTR2C, ADRA2C, HTR3A, ADRA1A, ADRA1B, ADRA1D, ADRA2B, ADRB1, ADRB2, DRD1, DRD2, DRD3, DRD5, HRH1, HRH3, HTR2B, HTR7, OPRK1, SLC6A2, SLC6A3, SLC6A4	-95.56
52	vorinostat	HDAC inhibitor, cell cycle inhibitor	HDAC1, HDAC2, HDAC3, HDAC6, HDAC8, HDAC10, HDAC11, HDAC5, HDAC9	-95.51
53	HSP inhibitor			-95.39
54	NBI-27914	CRF receptor antagonist	CRHR1	-95.27
55	erythromycin	NFkB pathway inhibitor, 50S ribosomal subunit inhibitor, motilin receptor agonist, RPLV inhibitor	CYP3A4, MLNR	-95.26
56	tipifarnib-P2	farnesyltransferase inhibitor, angiogenesis inhibitor, apoptosis stimulant	FNTA, FNTB	-95.22
57	FIT	opioid receptor agonist	OPRD1	-95.21
58	THM-I-94	HDAC inhibitor, apoptosis stimulant, cell cycle inhibitor	HDAC1, HDAC10, HDAC2, HDAC3, HDAC6, HDAC8	-95.13
59	MEK inhibitor			-95.02

Πίνακας 4.2: Αποτελέσματα για την ομάδα δειγμάτων 1

	Name	Description	Target	Median	Score
1	AG-490	epidermal growth factor receptor (EGFR) inhibitor, ErbB2 and JAK2 inhibitor, JAK inhibitor	JAK2, JAK3, EGFR, STAT3	-28.21	-99.98
2	HG-5-113-01	protein kinase inhibitor	ABL1, LTK, STK10	-92.62	-99.86
3	BIBU-1361	EGFR inhibitor	EGFR	-98.49	-99.85
4	Bromodomain Inhibitor			-96.37	-99.83
5	TG-101348	JAK inhibitor, FLT3 inhibitor, RET tyrosine kinase inhibitor	JAK2, FLT3, BRD4, JAK1, JAK3, RET, TYK2	-98.10	-99.81
6	BMS-536924	insulin growth factor receptor inhibitor, insulin receptor ligand	IGF1R, AKT1, CCNE1, CDK2, CYP3A4, ERBB2, INSR, KDR, LCK, MAPK1, MET, PDGFRA, PDGFRB	-94.26	-99.80
7	BI-2536	PLK inhibitor, apoptosis stimulant, cell cycle inhibitor, protein kinase inhibitor	PLK1, BRD4, PLK2, PLK3	-96.97	-99.79
8	JAK inhibitor			-88.03	-99.75
9	selumetinib	MEK inhibitor, MAP kinase inhibitor	MAP2K1, MAP2K2	-98.17	-99.74
10	GDC-0879	RAF inhibitor	BRAF	-24.63	-99.72
11	KU-0060648	DNA dependent protein kinase, DNA dependent protein kinase inhibitor, PI3K inhibitor	PIK3CA, PIK3CB, PIK3CD, PIK3CG, PRKDC	-97.72	-99.72
12	RO-3306	CDK inhibitor	CDK1	-71.30	-99.69
13	PLX-4720	RAF inhibitor	BRAF, KDR	65.86	-99.67
14	AS-605240	Phosphatidylinositol 3-kinase (PI3K) inhibitor, PI3K inhibitor	MAOB, PIK3CA, PIK3CB, PIK3CD, PIK3CG	-91.92	-99.63
15	tivozanib	VEGFR inhibitor, KIT inhibitor, tyrosine kinase inhibitor	FLT1, FLT4, KDR, KIT, PDGFRA, PDGFRB	-93.77	-99.60
16	tozasertib	Aurora kinase inhibitor, Bcr-Abl kinase inhibitor, FLT3 inhibitor, JAK inhibitor, Abl kinase inhibitor, mitotic inhibitor	AURKA, AURKB, ABL1, AURKC, BCR, FLT3, JAK2, DDR2, LCK	-96.63	-99.58

17	SB-202190	p38 MAPK inhibitor, interleukin inhibitor, stress activated protein kinase inhibitor	MAPK14, AKT1, ALOX5, CHEK1, GSK3B, LCK, MAPK1, MAPK11, MAPK12, MAPK8, PRKCA, ROCK1, RPS6KB1, SGK1	-82.81	-99.55
18	L-690488	nositol monophosphatase inhibitor	IMPA1	-58.50	-99.53
19	NVP-TAE684	ALK tyrosine kinase receptor inhibitor, ALK tyrosine kinase receptor mutant inhibitor, leucine rich repeat kinase inhibitor	ALK, INSR	-93.72	-99.52
20	aminopurvalanol-a	CDK inhibitor, tyrosine kinase inhibitor	CDK1, CDK2, CDK5, CDK6	-87.58	-99.49
21	pirarubicin	topoisomerase inhibitor	TOP2A	-88.19	-99.48
22	doxorubicin	topoisomerase inhibitor, DNA intercalating drug	TOP2A	-93.94	-99.46
23	trichostatin-a	HDAC inhibitor, CDK expression enhancer, ID1 expression inhibitor	HDAC7, HDAC8, HDAC1, HDAC10, HDAC2, HDAC3, HDAC4, HDAC5, HDAC6, HDAC9	-75.69	-99.42
24	SB-590885	RAF inhibitor	BRAF	-88.08	-99.40
25	RAF inhibitor			-71.83	-99.39
26	givinostat	HDAC inhibitor, interleukin receptor antagonist, interleukin synthesis inhibitor, tumor necrosis factor receptor antagonist, tumor necrosis factor release inhibitor	HDAC2, HDAC1, HDAC3, HDAC4, HDAC5, HDAC6, HDAC7, HDAC8, HDAC9, IL1B, IL1R2, IL6R, TNF	-19.59	-99.37
27	NCH-51	HDAC inhibitor	HDAC1, HDAC10, HDAC11, HDAC2, HDAC3, HDAC4, HDAC5, HDAC6, HDAC7, HDAC8, HDAC9	-58.40	-99.36

28	scriptaid	HDAC inhibitor	HDAC1, HDAC2, HDAC3, HDAC4, HDAC5, HDAC6, HDAC7, HDAC8, HDAC9	-48.70	-99.32
29	IKK-16	IKK inhibitor	IKBKB	-81.35	-99.29
30	TWS-119	glycogen synthase kinase inhibitor	GSK3B, JUN, MYC	0.00	-99.21
31	trichostatin-a	HDAC inhibitor, CDK expression enhancer, ID1 expression inhibitor	HDAC7, HDAC8, HDAC1, HDAC10, HDAC2, HDAC3, HDAC4, HDAC5, HDAC6, HDAC9	-81.30	-99.21
32	NVP-AUY922	HSP inhibitor	HSP90AA1, HSP90AA2, HSP90AB1	-92.52	-99.11
33	U-0126	MEK inhibitor, JAK inhibitor, MAP kinase inhibitor	AKT1, CHEK1, GSK3B, JAK2, LCK, MAP2K1, MAP2K2, MAP2K7, MAPK1, MAPK11, MAPK12, MAPK14, MAPK8, PRKCA, RAF1, ROCK1, RPS6KB1, SGK1	-95.11	-99.09
34	vorinostat	HDAC inhibitor, cell cycle inhibitor	HDAC1, HDAC2, HDAC3, HDAC6, HDAC8, HDAC10, HDAC11, HDAC5, HDAC9	-67.47	-99.00
35	neratinib	EGFR inhibitor, receptor tyrosine protein kinase inhibitor, tyrosine kinase inhibitor	EGFR, ERBB2, ERBB4, KDR	-96.51	-99.00
36	geldanamycin	HSP inhibitor	HSP90AA1, HSP90AB1	-91.41	-98.94
37	RS-17053	adrenergic receptor antagonist	ADRA1A, ADRA1D	-80.46	-98.92
38	MK-2206	AKT inhibitor	AKT1, AKT2, AKT3	-97.29	-98.81
39	berbamine	calmodulin antagonist	CALM1	-74.45	-98.74
40	HSP inhibitor			-80.15	-98.62

41	daunorubicin	RNA synthesis inhibitor, topoisomerase inhibitor, DNA synthesis inhibitor, radical formation stimulant	TOP2A, TOP2B	-89.99	-98.57
42	PD-184352	MEK inhibitor, MAP kinase inhibitor	MAP2K1, MAP2K2, MAP3K1, MAP3K2	-89.40	-98.54
43	AT-7519	CDK inhibitor, cell cycle inhibitor	CDK2, CDK5, CDK1, CDK4, CDK6, CDK9	-88.45	-98.53
44	THM-I-94	HDAC inhibitor, apoptosis stimulant, cell cycle inhibitor	HDAC1, HDAC10, HDAC2, HDAC3, HDAC6, HDAC8	-56.55	-98.46
45	everolimus	mTOR inhibitor, angiogenesis inhibitor, cell cycle inhibitor, immunosuppressant, protein kinase inhibitor, rotamase inhibitor	MTOR, CYP3A5, FKBP1A	-74.85	-98.40
46	FLT3 inhibitor			-94.81	-98.34
47	penicillin	cell wall synthesis inhibitor		-10.20	-98.32
48	lobelanidine	acetylcholine receptor antagonist, dopamine receptor modulator, opioid receptor antagonist, vesicular monoamine transporter ligand	SLC18A2, CHRNA10, CHRNA4, CHRNA9, CHRNB2, OPRM1	0.00	-98.27
49	MEK inhibitor			-90.79	-98.21
50	HDAC inhibitor			-65.04	-98.19
51	daunorubicin	RNA synthesis inhibitor, topoisomerase inhibitor, DNA synthesis inhibitor, radical formation stimulant	TOP2A, TOP2B	-94.46	-98.11
52	ISOX	HDAC inhibitor	HDAC6	-73.05	-98.07
53	DL-PDMP	ceramide glucosyltransferase inhibitor, glucosyltransferase inhibitor	ASAH1, UGCG	-80.51	-98.06
54	4-hydroxy-2-nonenal	cytotoxic lipid peroxidation product	IKBKB	0.00	-98.05
55	pidorubicine	topoisomerase inhibitor, DNA intercalating drug	TOP2A	-92.33	-98.02
56	efavirenz	HIV reverse transcriptase inhibitor, non-nucleoside reverse transcriptase inhibitor, reverse transcriptase inhibitor	CYP2B6, CYP2C19, CYP2C8, CYP3A4, CYP3A5	9.80	-97.97
57	JAK3-inhibitor-VI	JAK inhibitor	JAK3	-56.60	-97.86

58	sunitinib	FLT3 inhibitor, KIT inhibitor, PDGFR tyrosine kinase receptor inhibitor, RET tyrosine kinase inhibitor, VEGFR inhibitor, angiogenesis inhibitor, colony stimulating factor receptor antagonist, colony stimulating factor receptor inhibitor, platelet-derived growth factor receptor (PDGFR) inhibitor, vascular endothelial growth factor receptor (VEGFR) inhibitor, vascular endothelial growth factor receptor 1 (VEGFR1) inhibitor, vascular endothelial growth factor receptor 2 (VEGFR2) inhibitor, VEGFR antagonist	FLT3, KDR, KIT, FLT4, FLT1, PDGFRA, PDGFRB, RET, CSF1R, FGFR1	-83.03	-97.83
59	XMD-885	leucine rich repeat kinase inhibitor, MAP kinase inhibitor	LRRK2, MAPK7	-59.80	-97.81
60	BRD-K30351863	APEX inhibitor	APEX1	-85.58	-97.74
61	IGF-1 inhibitor			-92.73	-97.71
62	vemurafenib	RAF inhibitor, protein kinase inhibitor	BRAF, CYP2C19, CYP3A4, CYP3A5, RAF1	19.36	-97.68
63	latrunculin-b	actin destabilizer, unidentified pharmacological activity	ACTA1, MKL1, SPIRE2	-82.90	-97.68
64	CDK inhibitor			-73.27	-97.66
65	avrainvillamide-analog-5	nucleophosmin inhibitor	NPM1	-23.91	-97.62
66	Calmodulin antagonist			-39.23	-97.62
67	VER-155008	HSP inhibitor	HSPA1A	-85.37	-97.51
68	SA-792987	PKC inhibitor	WEE1	-82.70	-97.50
69	dacinostat	HDAC inhibitor	HDAC1, HDAC2, HDAC3, HDAC4, HDAC5, HDAC6, HDAC7, HDAC8, HDAC9	-47.84	-97.47

70	mibefradil	T-type calcium channel blocker, angiogenesis inhibitor, calcium channel blocker, calcium channel inhibitor, L-type calcium channel blocker, sodium channel blocker	CACNA1G, CACNA1H, CACNA1C, CACNA1I, ANO1, CACNA1D, CACNA1F, CACNA1S, CACNB1, CACNB2, CACNB3, CACNB4, CATSPER1, CATSPER2, CATSPER3, CATSPER4, CYP3A5, CYP3A7, SCN2A, SCN4A, SCN5A, SCN9A	-52.72	-97.41
71	doconexent	PPAR receptor agonist, unidentified pharmacological activity	FFAR1, PPARA, PTGS1, PTGS2	0.00	-97.36
72	PU-H71	HSP inhibitor	HSP90AA1	-63.40	-97.30
73	AKT-inhibitor-1-2	AKT inhibitor, PI3K inhibitor	AKT1, AKT2, AKT3	-83.02	-97.27
74	nimodipine	calcium channel blocker, L-type calcium channel blocker	CACNA1C, NR3C2, AHR, CACNA1D, CACNA1F, CACNA1S, CACNB1, CACNB2, CACNB3, CACNB4, CFTR	6.70	-97.26
75	JNJ-7706621	CDK inhibitor, Aurora kinase inhibitor	CDK1, CDK2, AURKA, AURKB	-66.95	-97.19

76	staurosporine	PKC inhibitor, AKT inhibitor, BMX inhibitor, CDK inhibitor, CHK inhibitor, G protein coupled receptor agonist, glycogen synthase kinase inhibitor, leucine rich repeat kinase inhibitor, ribosomal protein inhibitor, sodium/hydrogen exchanger inhibitor	CDK2, GSK3B, CAMK2B, CDK1, CDK5, CHEK1, CHRM1, CHRM2, CHRM4, CSK, DAPK1, GPR35, IKBKB, ITK, LCK, LRRK2, MAP2K4, MAP2K6, MAPKAPK2, PAK2, PDPK1, PHKG2, PIK3CG, PIM1, PKN1, PRKACB, PRKCI, PRKCQ, RPS6KA1, STK3, SYK, TNIK, ZAP70	-72.10	-97.18
77	alvocidib	CDK inhibitor, apoptosis stimulant, BCL inhibitor, cell cycle inhibitor, MCL1 inhibitor, survivin inhibitor, XIAP inhibitor	CDK2, CDK4, CDK1, CDK6, CDK7, CDK9, CDK5, CDK8, EGFR, PYGM, BCL2, BIRC5, CCNT1, MCL1, XIAP	-87.50	-97.14
78	propranolol	adrenergic receptor antagonist	ADRB2, ADRB3, ADRB1, CYP2C19, HTR1A, HTR1B	0.00	-97.12
79	PD-0325901	MEK inhibitor, MAP kinase inhibitor, protein kinase inhibitor	MAP2K1, MAP2K2	-26.74	-97.05
80	WT-171	HDAC inhibitor	HDAC6	-55.55	-97.03
81	serdemetan	MDM inhibitor, angiogenesis stimulant, apoptosis stimulant, oncogene inhibitor	MDM2	-83.76	-96.65
82	triptolide	RNA polymerase inhibitor	CYP2C19, RELA	-92.97	-96.63
83	benzatropine	anticholinergic	CHRM1, HRH1, SLC6A3	-45.41	-96.53
84	tipifarnib-P2	farnesyltransferase inhibitor, angiogenesis inhibitor, apoptosis stimulant	FNTA, FNTB	-70.88	-96.41
85	DNA dependent protein kinase inhibitor			-94.79	-96.41
86	MEK1-2-inhibitor	MEK inhibitor	MAP2K1, MAP2K2	-80.50	-96.37
87	HC-toxin	HDAC inhibitor	HDAC1	-47.94	-96.10

88	wiskostatin	neural Wiskott-Aldrich syndrome protein inhibitor	WAS, WASL	-89.88	-96.04
89	Homeobox Gene GOF			-44.93	-95.87
90	crizotinib	ALK tyrosine kinase receptor inhibitor, c-Met inhibitor, hepatocyte growth factor receptor inhibitor, tyrosine kinase inhibitor	ALK, MET, CYP2B6, CYP3A5, MST1R, ROS1	-69.29	-95.74
91	avrainvillamide-analog-2	nucleophosmin inhibitor	NPM1	-36.64	-95.69
92	thioridazine	acetylcholine receptor ligand, dopamine receptor, dopamine receptor antagonist, mucosa associated lymphoid tissue lymphoma translocation protein 1 (MALT1) inhibitor	DRD1, DRD2, HTR2A, ADRA1A, ADRA1B, CHRNA7, DRD5, HRH1, HTR1A, HTR2C, HTR6, HTR7, KCNH2, MALT1	-80.35	-95.58
93	simvastatin	HMGCR inhibitor	HMGCR, CYP2C8, CYP3A4, CYP3A5, ITGB2	-87.01	-95.53
94	ZM-447439	Aurora kinase inhibitor	AURKA, AURKB	-45.85	-95.51
95	GW-5074	RAF inhibitor, leucine rich repeat kinase inhibitor	LRRK1, LRRK2, NTRK1, RAF1	0.00	-95.47
96	hydrocortisone	corticosteroid agonist, glucocorticoid receptor agonist, immunosuppressant, interleukin receptor antagonist	ANXA1, NOS2, NR3C1, NR3C2	-31.25	-95.24
97	BX-795	IKK inhibitor, PDK1 inhibitor, phosphoinositide dependent kinase inhibitor, serine/threonine kinase inhibitor, TBK1 inhibitor	PDPK1, CDK2, CHEK1, GSK3B, IKBKE, KDR, PDK1, TBK1	-76.03	-95.18

Πίνακας 4.3: Αποτελέσματα για την ομάδα δειγμάτων 2

	Name	Description	Target	Median	Score
1	MK-2206	AKT inhibitor	AKT1, AKT2, AKT3	-97.29	-99.89
2	BIBU-1361	EGFR inhibitor	EGFR	-98.49	-99.88
3	serdemetan	MDM inhibitor, angiogenesis stimulant, apoptosis stimulant, oncogene inhibitor	MDM2	-83.76	-99.80
4	GDC-0879	RAF inhibitor	BRAF	-24.63	-99.74
5	selumetinib	MEK inhibitor, MAP kinase inhibitor	MAP2K1, MAP2K2	-98.17	-99.72
6	thioridazine	acetylcholine receptor ligand, dopamine receptor, dopamine receptor antagonist, mucosa associated lymphoid tissue lymphoma translocation protein 1 (MALT1) inhibitor	DRD1, DRD2, HTR2A, ADRA1A, ADRA1B, CHRNA7, DRD5, HRH1, HTR1A, HTR2C, HTR6, HTR7, KCNH2, MALT1	-80.35	-99.69
7	AS-605240	Phosphatidylinositol 3-kinase (PI3K) inhibitor, PI3K inhibitor	MAOB, PIK3CA, PIK3CB, PIK3CD, PIK3CG	-91.92	-99.65
8	PLX-4720	RAF inhibitor	BRAF, KDR	65.86	-99.61
9	BMS-536924	insulin growth factor receptor inhibitor, insulin receptor ligand	IGF1R, AKT1, CCNE1, CDK2, CYP3A4, ERBB2, INSR, KDR, LCK, MAPK1, MET, PDGFRA, PDGFRB	-94.26	-99.61
10	BI-2536	PLK inhibitor, apoptosis stimulant, cell cycle inhibitor, protein kinase inhibitor	PLK1, BRD4, PLK2, PLK3	-96.97	-99.55
11	berbamine	calmodulin antagonist	CALM1	-74.45	-99.51
12	tozasertib	Aurora kinase inhibitor, Bcr-Abl kinase inhibitor, FLT3 inhibitor, JAK inhibitor, Abl kinase inhibitor, mitotic inhibitor	AURKA, AURKB, ABL1, AURKC, BCR, FLT3, JAK2, DDR2, LCK	-96.63	-99.49
13	PU-H71	HSP inhibitor	HSP90AA1	-63.40	-99.45
14	atorvastatin	HMGCR inhibitor, dipeptidyl peptidase inhibitor, tumor necrosis factor expression inhibitor	HMGCR, DPP4, AHR, CYP3A5, FASLG	-85.00	-99.45
15	triptolide	RNA polymerase inhibitor	CYP2C19, RELA	-92.97	-99.45

16	BRD-K77681376	casein kinase inhibitor, FLT3 inhibitor		-78.80	-99.40
17	dihydroergocristine	adrenergic receptor antagonist, prolactin inhibitor, adrenergic receptor partial agonist, dopamine receptor agonist, dopamine receptor partial agonist, dopamine receptor partial antagonist, serotonin receptor antagonist	HTR2A, ADRA1A, ADRB1, DRD1, DRD2, DRD3, DRD4, DRD5, HTR1A, HTR3A, HTR4, HTR5A, HTR6, HTR7	-66.50	-99.39
18	pirarubicin	topoisomerase inhibitor	TOP2A	-88.19	-99.37
19	daunorubicin	RNA synthesis inhibitor, topoisomerase inhibitor, DNA synthesis inhibitor, radical formation stimulant	TOP2A, TOP2B	-94.46	-99.35
20	trifluoperazine	breast cancer resistance protein inhibitor, calcium/calmodulin dependent protein kinase inhibitor, dopamine receptor, dopamine receptor antagonist, sodium channel blocker	DRD2, ABCG2, ADRA1A, CALM1, CALY, CAMK2A, DRD4, HRH1, HTR2A, HTR2C, S100A4, SCN4A, SCN9A, TNNC1	-70.32	-99.30
21	GW-5074	RAF inhibitor, leucine rich repeat kinase inhibitor	LRRK1, LRRK2, NTRK1, RAF1	0.00	-99.29
22	pidorubicine	topoisomerase inhibitor, DNA intercalating drug	TOP2A	-92.33	-99.25
23	doxorubicin	topoisomerase inhibitor, DNA intercalating drug	TOP2A	-93.94	-99.20
24	sirolimus	mTOR inhibitor, CCR expression inhibitor, cell cycle inhibitor, proteasome inhibitor, protein kinase inhibitor, T cell inhibitor	MTOR, FKBP1A, CCR5, FGF2	-79.00	-99.04
25	HG-5-113-01	protein kinase inhibitor	ABL1, LTK, STK10	-92.62	-99.00
26	BRD-K06956503	glucosylceramidase inhibitor	GBA	-41.59	-98.95
27	tivozanib	VEGFR inhibitor, KIT inhibitor, tyrosine kinase inhibitor	FLT1, FLT4, KDR, KIT, PDGFRA, PDGFRB	-93.77	-98.90
28	dihydroergocristine	adrenergic receptor antagonist, prolactin inhibitor, adrenergic receptor partial agonist, dopamine receptor agonist, dopamine receptor partial agonist, dopamine receptor partial antagonist, serotonin receptor antagonist	HTR2A, ADRA1A, ADRB1, DRD1, DRD2, DRD3, DRD4, DRD5, HTR1A, HTR3A, HTR4, HTR5A, HTR6, HTR7	-34.33	-98.89

29	cyclopamine	smoothened receptor antagonist, hedgehog pathway inhibitor	SMO, DHH, IHH, PTCH1	-45.99	-98.89
30	AKT-inhibitor-1-2	AKT inhibitor, PI3K inhibitor	AKT1, AKT2, AKT3	-83.02	-98.87
31	AT-7519	CDK inhibitor, cell cycle inhibitor	CDK2, CDK5, CDK1, CDK4, CDK6, CDK9	-88.45	-98.84
32	andarine	androgen receptor modulator	AR	0.00	-98.75
33	RAF inhibitor			-71.83	-98.47
34	VER-155008	HSP inhibitor	HSPA1A	-85.37	-98.45
35	UNC-0321	histone lyinse methyltransferase inhibitor, histone lysine methyltransferase inhibitor	EHMT2	-10.13	-98.41
36	doxorubicin	topoisomerase inhibitor, DNA intercalating drug	TOP2A	-94.05	-98.41
37	CGP-60474	CDK inhibitor, PKC inhibitor	CDK1, CDK2	-79.60	-98.23
38	niguldipine	adrenergic receptor antagonist, calcium channel blocker, calcium channel antagonist, calcium channel inhibitor	ADRA1A, CACNA1C, ADRA1B, ADRA1D	-78.20	-98.17
39	daunorubicin	RNA synthesis inhibitor, topoisomerase inhibitor, DNA synthesis inhibitor, radical formation stimulant	TOP2A, TOP2B	-91.44	-98.11
40	QL-XI-92	DDR1 inhibitor	DDR1	-82.40	-98.09
41	alimemazine	histamine receptor ligand	HRH1	-72.76	-98.01
42	propranolol	adrenergic receptor antagonist	ADRB2, ADRB3, ADRB1, CYP2C19, HTR1A, HTR1B	0.00	-98.00
43	alvocidib	CDK inhibitor, apoptosis stimulant, BCL inhibitor, cell cycle inhibitor, MCL1 inhibitor, survivin inhibitor, XIAP inhibitor	CDK2, CDK4, CDK1, CDK6, CDK7, CDK9, CDK5, CDK8, EGFR, PYGM, BCL2, BIRC5, CCNT1, MCL1, XIAP	-87.50	-97.81
44	boldine	acetylcholine receptor inhibitor, dopamine receptor antagonist	CHRNA4, CHRNB2, DRD1, DRD2	0.00	-97.80
45	penicillin	cell wall synthesis inhibitor		-10.20	-97.78
46	simvastatin	HMGCR inhibitor	HMGCR, CYP2C8, CYP3A4, CYP3A5, ITGB2	-87.01	-97.57
47	pseudopelletierine	stimulant reflex trigger used as an anthelmintic and antiamoeboid		12.29	-97.56

48	fluocinolone	corticosteroid agonist, glucocorticoid receptor agonist	NR3C1, SERPINA6	-49.43	-97.52
49	dactinomycin	DNA directed RNA polymerase inhibitor, nucleic acid syntheis inhibitor, protein synthesis inhibitor	POLR2A	-84.54	-97.50
50	picotamide	prostanoid receptor antagonist, thromboxane receptor antagonist, thromboxane synthase inhibitor	TBXA2R, TBXAS1	-3.36	-97.48
51	NNC-05-2090	GABA uptake inhibitor, GAT inhibitor	SLC6A11, SLC6A12, SLC6A13	-77.16	-97.47
52	fluspirilene	dopamine receptor, dopamine receptor antagonist	DRD2, HTR2A, CACNG1, HRH1, HTR1A, HTR1D, HTR1E	-67.28	-97.43
53	KU-0060648	DNA dependent protein kinase, DNA dependent protein kinase inhibitor, PI3K inhibitor	PIK3CA, PIK3CB, PIK3CD, PIK3CG, PRKDC	-97.72	-97.37
54	bisindolylmaleimide-ix	PKC inhibitor, glycogen synthase kinase inhibitor, leucine rich repeat kinase inhibitor, SIRT inhibitor	SIRT1, AKT1, GSK3B, LCK, LRRK2, MAPK1, MAPK11, MAPK12, MAPK14, MAPK8, PRKCA, ROCK1, RPS6KB1, SIRT2	-79.36	-97.35
55	Calmodulin antagonist			-39.23	-97.27
56	proxyfan	histamine receptor modulator	HRH3	0.00	-97.04
57	daunorubicin	RNA synthesis inhibitor, topoisomerase inhibitor, DNA synthesis inhibitor, radical formation stimulant	TOP2A, TOP2B	-89.99	-96.99
58	5-iodotubercidin	adenosine kinase inhibitor, nucleoside transporter inhibitor	MAPK3, ADK, CSNK1G2, GSG2, LRRK2	-75.39	-96.96
59	BRD-K53780220	casein kinase inhibitor, FLT3 inhibitor		-63.07	-96.92
60	idarubicin	topoisomerase inhibitor, radical formation stimulant, RNA synthesis inhibitor	TOP2A	-79.29	-96.83

61	epinephrine	carbonic anhydrase activator, hormone, neurotransmitter	ADRA1A, ADRA1B, ADRA1D, ADRA2A, ADRA2B, ADRA2C, ADRB1, ADRB2, ADRB3, PAH, TNF	-36.79	-96.70
62	PF-562271	focal adhesion kinase inhibitor, angiogenesis inhibitor, apoptosis stimulant	PTK2, PTK2B	-81.59	-96.69
63	AS-601245	JNK inhibitor	GSK3B, MAPK10, MAPK8, MAPK9, PIM1	-57.40	-96.62
64	RS-67333	serotonin receptor partial agonist	HTR4	-0.30	-96.58
65	BRD-K11757396	neuropeptide receptor ligand	NPSR1	-16.71	-96.53
66	NVP-TAE684	ALK tyrosine kinase receptor inhibitor, ALK tyrosine kinase receptor mutant inhibitor, leucine rich repeat kinase inhibitor	ALK, INSR	-93.72	-96.38
67	ZM-447439	Aurora kinase inhibitor	AURKA, AURKB	-45.85	-96.18
68	ZG-10	JNK inhibitor	MAPK8	-93.72	-95.97
69	arctigenin	adiponectin receptor agonist, AP inhibitor, aryl hydrocarbon receptor antagonist, HIV integrase inhibitor, MEK inhibitor, NFkB pathway inhibitor, topoisomerase inhibitor	ADIPOR1, AHR, CHUK, MAP2K1	15.69	-95.88
70	Bromodomain Inhibitor			-96.37	-95.85
71	mitoxantrone	topoisomerase inhibitor, DNA intercalating drug, HCV inhibitor, immunosuppressant, Pim kinase inhibitor	TOP2A, PIM1	-72.15	-95.80
72	IGF-1 inhibitor			-92.73	-95.76
73	benperidol	dopamine receptor antagonist	DRD2	0.00	-95.64
74	NSC-23766	RAC1 GTPase inhibitor		-11.32	-95.63
75	L-732138	tachykinin antagonist	TACR1, TACR2	24.72	-95.50
76	vemurafenib	RAF inhibitor, protein kinase inhibitor	BRAF, CYP2C19, CYP3A4, CYP3A5, RAF1	19.36	-95.46
77	ellipticine	topoisomerase inhibitor, DNA intercalating drug	TOP2A, TOP2B	-38.89	-95.46

78	tamoxifen	estrogen receptor antagonist, selective estrogen receptor modulator (SERM), estrogen receptor agonist, estrogen receptor modulator, PKC inhibitor	ESR1, ESR2, CYP3A5, EBP, GPER1, PRKCA, PRKCB, PRKCD, PRKCE, PRKCG, PRKCI, PRKCQ, PRKCZ	-85.24	-95.45
79	doxorubicin	topoisomerase inhibitor, DNA intercalating drug	TOP2A	-90.78	-95.42
80	Homeobox Gene GOF			-44.93	-95.01

Πίνακας 4.4: Αποτελέσματα για την ομάδα δειγμάτων 3

	Name	Description	Target	Median	Score
1	IB-MECA	adenosine receptor agonist, granulocyte colony stimulating factor agonist	ADORA3, ADORA1, ADORA2A, ADORA2B	0.00	-99.94
2	SB-202190	p38 MAPK inhibitor, interleukin inhibitor, stress activated protein kinase inhibitor	MAPK14, AKT1, ALOX5, CHEK1, GSK3B, LCK, MAPK1, MAPK11, MAPK12, MAPK8, PRKCA, ROCK1, RPS6KB1, SGK1	-82.81	-99.93
3	dobutamine	adrenergic receptor agonist	ADRB1, ADRB2, ADRA1A	0.00	-99.82
4	nimodipine	calcium channel blocker, L-type calcium channel blocker	CACNA1C, NR3C2, AHR, CACNA1D, CACNA1F, CACNA1S, CACNB1, CACNB2, CACNB3, CACNB4, CFTR	6.70	-99.74
5	tipifarnib-P2	farnesyltransferase inhibitor, angiogenesis inhibitor, apoptosis stimulant	FNTA, FNTB	-70.88	-99.68
6	GDC-0879	RAF inhibitor	BRAF	-24.63	-99.66
7	AG-490	epidermal growth factor receptor (EGFR) inhibitor, ErbB2 and JAK2 inhibitor, JAK inhibitor	JAK2, JAK3, EGFR, STAT3	-28.21	-99.66
8	U-0126	MEK inhibitor, JAK inhibitor, MAP kinase inhibitor	AKT1, CHEK1, GSK3B, JAK2, LCK, MAP2K1, MAP2K2, MAP2K7, MAPK1, MAPK11, MAPK12, MAPK14, MAPK8, PRKCA, RAF1, ROCK1, RPS6KB1, SGK1	-95.11	-99.65

9	PD-184352	MEK inhibitor, MAP kinase inhibitor	MAP2K1, MAP2K2, MAP3K1, MAP3K2	-89.40	-99.63
10	cimaterol	adrenergic receptor agonist		23.05	-99.62
11	MEK1-2-inhibitor	MEK inhibitor	MAP2K1, MAP2K2	-80.50	-99.61
12	MEK inhibitor			-90.79	-99.54
13	PLX-4720	RAF inhibitor	BRAF, KDR	65.86	-99.46
14	RAF inhibitor			-71.83	-99.42
15	AS-605240	Phosphatidylinositol 3-kinase (PI3K) inhibitor, PI3K inhibitor	MAOB, PIK3CA, PIK3CB, PIK3CD, PIK3CG	-91.92	-99.40
16	RO-15-4513	GABA benzodiazepine site receptor inverse agonist	GABRA1, GABRA2, GABRA3, GABRA5, GABRA4, GABRA6, GABRB2, GABRG2	0.00	-99.39
17	benzatropine	anticholinergic	CHRM1, HRH1, SLC6A3	-45.41	-99.39
18	bisindolylmaleimide	CDK inhibitor, PKC inhibitor, leucine rich repeat kinase inhibitor	CCND1, CDK4, LRRK2, PDPK1, PIM1, PRKCA, PRKCB, PRKCI, PRKCZ	-77.59	-99.37
19	pirarubicin	topoisomerase inhibitor	TOP2A	-88.19	-99.36
20	tivozanib	VEGFR inhibitor, KIT inhibitor, tyrosine kinase inhibitor	FLT1, FLT4, KDR, KIT, PDGFRA, PDGFRB	-93.77	-99.34
21	GW-5074	RAF inhibitor, leucine rich repeat kinase inhibitor	LRRK1, LRRK2, NTRK1, RAF1	0.00	-99.28
22	TWS-119	glycogen synthase kinase inhibitor	GSK3B, JUN, MYC	0.00	-99.28
23	EMF-bca1-60	caspase inhibitor		-12.09	-99.20
24	PD-0325901	MEK inhibitor, MAP kinase inhibitor, protein kinase inhibitor	MAP2K1, MAP2K2	-26.74	-99.13
25	selumetinib	MEK inhibitor, MAP kinase inhibitor	MAP2K1, MAP2K2	-98.17	-99.07
26	AM-92016	glutamate receptor antagonist, ionotropic glutamate receptor antagonist, time-dependent delayed rectifier potassium current blocker	GRIN1, GRIN2B	-18.06	-99.06
27	atenolol	adrenergic receptor antagonist	ADRB1, ADRB2	0.00	-98.97
28	vemurafenib	RAF inhibitor, protein kinase inhibitor	BRAF, CYP2C19, CYP3A4, CYP3A5, RAF1	19.36	-98.92

29	UNC-0321	histone lyinse methyltransferase inhibitor, histone lysine methyltransferase inhibitor	EHMT2	-10.13	-98.88
30	MNITMT	lymphocyte inhibitor		0.00	-98.87
31	NVP-AUY922	HSP inhibitor	HSP90AA1, HSP90AA2, HSP90AB1	-92.52	-98.85
32	niguldipine	adrenergic receptor antagonist, calcium channel blocker, calcium channel antagonist, calcium channel inhibitor	ADRA1A, CACNA1C, ADRA1B, ADRA1D	-78.20	-98.82
33	SB-590885	RAF inhibitor	BRAF	-88.08	-98.82
34	U0126	MEK inhibitor, JAK inhibitor, MAP kinase inhibitor	JAK2, MAP2K1, MAP2K2, MAP3K1, MAP3K2	-86.11	-98.81
35	simvastatin	HMGCR inhibitor	HMGCR, CYP2C8, CYP3A4, CYP3A5, ITGB2	-87.01	-98.77
36	pimozide	dopamine receptor antagonist, dopamine receptor, opioid receptor antagonist, serotonin receptor antagonist	DRD2, DRD3, CACNA1I, CALM1, HRH1, HTR1A, HTR2A, KCNA10, KCNH2	-47.96	-98.48
37	SR-27897	CCK receptor antagonist	CCKAR	0.00	-98.32
38	procaterol	adrenergic receptor agonist	ADRB2	0.00	-98.20
39	LY-364947	TGF beta receptor inhibitor, p38 MAPK inhibitor	TGFBR1	-15.76	-97.95
40	ZK-164015	estrogen receptor antagonist	ESR1, ESR2	0.00	-97.86
41	tozasertib	Aurora kinase inhibitor, Bcr-Abl kinase inhibitor, FLT3 inhibitor, JAK inhibitor, Abl kinase inhibitor, mitotic inhibitor	AURKA, AURKB, ABL1, AURKC, BCR, FLT3, JAK2, DDR2, LCK	-96.63	-97.84
42	aminolevulinic-acid	oxidizing agent	ALAD	-41.85	-97.77
43	BMS-536924	insulin growth factor receptor inhibitor, insulin receptor ligand	IGF1R, AKT1, CCNE1, CDK2, CYP3A4, ERBB2, INSR, KDR, LCK, MAPK1, MET, PDGFRA, PDGFRB	-94.26	-97.74
44	etazolate	phosphodiesterase inhibitor, alpha secretase activator, GABA receptor modulator	GABRB3, PDE4A	0.00	-97.65

45	PP-1	src inhibitor, Abl kinase inhibitor	HCK, RET, SRC	-86.88	-97.63
46	levocabastine	histamine receptor antagonist	HRH1, NTSR2	1.09	-97.61
47	panobinostat	HDAC inhibitor, apoptosis stimulant, cell cycle inhibitor	HDAC1, HDAC2, HDAC3, HDAC4, HDAC6, HDAC7, HDAC8, HDAC9	-45.45	-97.55
48	NBI-27914	CRF receptor antagonist	CRHR1	3.26	-97.36
49	nitrocaramiphen	acetylcholine receptor antagonist		23.00	-97.35
50	RO-3306	CDK inhibitor	CDK1	-71.30	-97.22
51	dasatinib	KIT inhibitor, src inhibitor, Bcr-Abl kinase inhibitor, ephrin receptor inhibitor, PDGFR tyrosine kinase receptor inhibitor, yes kinase inhibitor, Abl kinase inhibitor, Bruton's tyrosine kinase (BTK) inhibitor, discoidin domain containing receptor Inhibitor, lymphocyte specific tyrosine kinase inhibitor, tyrosine kinase inhibitor	ABL1, FYN, LCK, SRC, KIT, YES1, BCR, EPHA2, LYN, PDGFRB, ABL2, BTK, DDR1, DDR2, PDGFRA, STAT5B	-65.37	-97.16
52	daunorubicin	RNA synthesis inhibitor, topoisomerase inhibitor, DNA synthesis inhibitor, radical formation stimulant	TOP2A, TOP2B	-89.99	-97.12
53	STO-609	calcium/calmodulin dependent protein kinase inhibitor, calmodulin inhibitor	CAMKK1, CAMKK2	0.00	-97.03
54	carpindolol	adrenergic receptor antagonist, serotonin receptor antagonist		-44.01	-97.02
55	arctigenin	adiponectin receptor agonist, AP inhibitor, aryl hydrocarbon receptor antagonist, HIV integrase inhibitor, MEK inhibitor, NFkB pathway inhibitor, topoisomerase inhibitor	ADIPOR1, AHR, CHUK, MAP2K1	15.69	-96.87
56	veratridine	sodium channel activator	SCN1A, SCN3A, SCN8A, SCN9A	0.00	-96.79
57	BRD-K06956503	glucosylceramidase inhibitor	GBA	-41.59	-96.74
58	sinensetin	cyclooxygenase inhibitor		-8.91	-96.66
59	salbutamol	adrenergic receptor agonist	ADRB2, ADRB1	-26.77	-96.65
60	L-690488	nositol monophosphatase inhibitor	IMPA1	-58.50	-96.65
61	FLT3 inhibitor			-94.81	-96.64
62	mestranol	estrogen receptor agonist	ESR1	9.57	-96.51
63	VEGFR inhibitor			-50.76	-96.50

64	tamoxifen	estrogen receptor antagonist, selective estrogen receptor modulator (SERM), estrogen receptor agonist, estrogen receptor modulator, PKC inhibitor	ESR1, ESR2, CYP3A5, EBP, GPER1, PRKCA, PRKCB, PRKCD, PRKCE, PRKCG, PRKCI, PRKCQ, PRKCZ	-85.24	-96.47
65	daunorubicin	RNA synthesis inhibitor, topoisomerase inhibitor, DNA synthesis inhibitor, radical formation stimulant	TOP2A, TOP2B	-94.46	-96.44
66	PKC inhibitor			-33.29	-96.21
67	alimemazine	histamine receptor ligand	HRH1	-72.76	-96.15
68	forskolin	adenylyl cyclase activator, Adenylate cyclase stimulant, growth hormone receptor agonist, phosphokinase stimulant	ADCY2, ADCY5, GNAS	0.00	-96.08
69	4-hydroxy-2-nonenal	cytotoxic lipid peroxidation product	IKBKB	0.00	-95.94
70	JAK3-Inhibitor-II	JAK inhibitor, ALK tyrosine kinase receptor inhibitor, EGFR inhibitor	EGFR, ALK, JAK1, JAK2, JAK3	-46.24	-95.76
71	sunitinib	FLT3 inhibitor, KIT inhibitor, PDGFR tyrosine kinase receptor inhibitor, RET tyrosine kinase inhibitor, VEGFR inhibitor, angiogenesis inhibitor, colony stimulating factor receptor antagonist, colony stimulating factor receptor inhibitor, platelet-derived growth factor receptor (PDGFR) inhibitor, vascular endothelial growth factor receptor (VEGFR) inhibitor, vascular endothelial growth factor receptor 1 (VEGFR1) inhibitor, vascular endothelial growth factor receptor 2 (VEGFR2) inhibitor, VEGFR antagonist	FLT3, KDR, KIT, FLT4, FLT1, PDGFRA, PDGFRB, RET, CSF1R, FGFR1	-83.03	-95.64
72	GR-144053	integrin antagonist	ITGB3, ITGA2B	6.82	-95.60
73	BRD-A61599461	thyroid-stimulating hormone receptor inverse agonist	TSHR	16.02	-95.57
74	PD-198306	MAP kinase inhibitor, MEK inhibitor	MAP2K1, MAP2K2, MAPK1, MAPK3	-64.24	-95.47
75	AM-630	cannabinoid receptor antagonist, cannabinoid receptor inverse agonist	CNR2, CNR1	0.00	-95.18
76	RS-504393	CC chemokine receptor antagonist	CCR2, CCL2	-45.95	-95.08

77	cediranib	VEGFR inhibitor, KIT inhibitor, angiogenesis inhibitor, VEGFR antagonist	KDR, FLT1, FLT4, KIT, PDGFRB, CSF1R, FLT3, PDGFRA	-90.07	-95.05
78	AT-7519	CDK inhibitor, cell cycle inhibitor	CDK2, CDK5, CDK1, CDK4, CDK6, CDK9	-88.45	-95.03