ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικό και Καποδιστριακό
Πανεπιστήμιο Αθηνών

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ»

Δ Ι Π Λ Ω Μ Α Τ Ι Κ Η Ε Ρ Γ Α Σ Ι Α

«Ανάλυση δεδομένων προερχόμενων από RNA αλληλούχιση, με χρήση στατιστικών μεθόδων»

Μαριέτα Ρήγα

Πτυχιούχος Τμήματος Βιολογίας, Πανεπιστήμιο Κρήτης

ΑΘΗΝΑ 2019

HELLENIC REPUBLIC
National and Kapodistrian University of Athens

SCHOOL OF SCIENCE
FACULTY OF BIOLOGY

MASTER IN "BIOINFORMATICS"

Master Diploma Thesis

«Statistical analysis of RNA sequencing data»

Marieta Riga

Biology Degree, University of Crete

ATHENS 2019

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ «ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ»

Δ Ι Π Λ Ω Μ Α Τ Ι Κ Η Ε Ρ Γ Α Σ Ι Α

«Ανάλυση δεδομένων προερχόμενων από RNA αλληλούχηση, με χρήση στατιστικών μεθόδων»

Τριμελής εξεταστική επιτροπή

Καθηγητής Κωνσταντίνος Βοργιάς (Επιβλέπων)

Τομέας Βιοχημείας και Μοριακής Βιολογίας

Τμήμα Βιολογίας, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Αναπληρωτής καθηγητής Παντελής Μπάγκος

Τομέας Βιοπληροφορικής και Βιοστατιστικής

Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική,

Πανεπιστήμιο Θεσσαλίας

Ερευνητής Β' Αριστοτέλης Χατζηιωάννου

Εργαστήριο Μεταβολικής Μηχανικής και Βιοπληροφορικής,

Ινστιτούτο Βιολογίας, Φαρμακευτικής Χημείας και Βιοτεχνολογίας

Εθνικό Ίδρυμα Ερευνών

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τους καθηγητές της τριμελούς εξεταστικής επιτροπής για την προθυμία να συνεργαστούν μαζί μου κατά την πρόοδο και την ολοκλήρωση τόσο της διπλωματικής μου εργασίας, όσο και του μεταπτυχιακού μου.

Ιδιαίτερα, θα ήθελα να ευχαριστήσω των Αριστοτέλη Χατζηιωάννου, για την υπομονή, την καθοδήγηση, τον πάντα χειμαρρώδη και ενθουσιώδη λόγο, ακόμα και σε περιόδους που υπήρξα δραματικά ασυνεπής απέναντί του και απέναντι στις δυνατότητές μου. Αριστοτέλη, θερμά, ευχαριστώ.

Επίσης, την Όλγα Παπαδοδήμα και την Γεωργία Κοντογιάννη. Ένιωθα πάντα το ενδιαφέρον και την ενθάρρυνση, ακόμα και αν γινόταν με τον πιο διακριτικό τρόπο. Σας ευχαριστώ πολύ.

Ένα μεγάλο ευχαριστώ στην ομάδα της eNIOS και του Εθνικού Ιδρύματος Ερευνών. Κωνσταντίνε, Ειρήνη, Στάθη, Θοδωρή, Λευτέρη, Γεωργία, Ιλόνα, σας ευχαριστώ πολύ όλους.

Ακόμη, θα ήθελα να ευχαριστήσω την οικογένειά μου και ιδιαίτερα τον αδερφό μου Αλέξη, ο οποίος, σε περιόδους άγχους και πίεσης, ήταν εδώ να μου θυμίζει πόσο σημαντικό είναι το φαγητό, ο καφές, ο ύπνος και οι βόλτες.

Τέλος, οι φίλοι μου Μιχάλης, Γιάννης, Βαρβάρα, Κωνσταντίνα και Ευγένιος ήταν εκεί σε όλη τη διάρκεια και τώρα πλέον τελειώνει για όλους αυτή η διαδρομή. Παιδιά ευχαριστώ για όλα! Και στα επόμενα!

Κλείνω με τους υπόλοιπους συμφοιτητές μου. Το περάσαμε, το ζήσαμε, με ελάχιστες απώλειες και αναίμακτα και για αυτό, παιδιά, σας ευχαριστώ πολύ!

Περίληψη

Το δερματικό μελάνωμα είναι μία ιδιαίτερα επιθετική μορφή καρκίνου και σίγουρα ο πιο επικίνδυνος τύπος καρκίνου του δέρματος. Το 2012 καταγράφηκαν 232.000 νέες περιπτώσεις ανθρώπων με μελάνωμα, ενώ το 2015 υπήρχαν 3,1 εκατομμύρια ασθενών, οι οποίες περιπτώσεις κατέληξαν σε 59.800 θανάτους. Ως εκ τούτου, το μελάνωμα έχει 2 βασικά χαρακτηριστικά που το κάνουν κατάλληλο μοντέλο μελέτης, υπάρχει πληθώρα δεδομένων σχετικά με τη νόσο, τα οποία συχνά είναι ευρέως διαθέσιμα και είναι μια ασθένεια που σχετίζεται με τη σύγχρονη εποχή. Το μελάνωμα επιλέχθηκε ως μοντέλο για τη μελέτη διαφορικής έκφρασης γονιδίων μεταξύ δύο καταστάσεων. Χρησιμοποιήθηκε ένας αρχικός πληθυσμός 472 ασθενών / δειγμάτων, οι οποίοι διαιρέθηκαν σε ομάδες, με βάση δημογραφικά κριτήρια, φύλο και ηλικία και γενετικά κριτήρια, παρουσία ή όχι μεταλλαγής των γονιδίων της οικογένειας RAS, BRAF, NF1, ενώ δημιουργήθηκε και μια τρίτη ομάδα Triple-WT, για να υποδηλώσει ασθενείς που δε φέρουν καμιά από τις προηγούμενες μεταλλαγές. Η ανάλυση που οδήγησε στην εύρεση διαφορικά εκφραζόμενων γονιδίων μετά από διμερείς συγκρίσεις, πραγματοποιήθηκε με χρήση της γλώσσας προγραμματισμού R και συγκεκριμένα του πακέτου edgeR. Τα δεδομένα που προέκυψαν εισήχθησαν στο εργαλείο BioInfoMiner, για εμπλουτισμό και λειτουργική ανάλυση, από το οποίο προέκυψαν συμπεράσματα σχετικά με τις 3 από τις 9 συνολικά συγκρίσεις. Μια από αυτές τελικά έδωσε τα πιο ενδιαφέροντα συμπεράσματα. Η ανάλυση διαφορικής έκφρασης μεταξύ των ασθενών που φέρουν και των ασθενών που δεν φέρουν μεταλλαγή στο γονίδιο BRAF, οδήγησε σε συμπεράσματα σχετικά με την υπερ- ή -υπό έκφραση γονιδίων μεταξύ των δύο καταστάσεων και έδωσε ενδείξεις για την εμπλοκή των μηχανισμών ανοσολογικής απόκρισης και φλεγμονής, αγγειογένεσης, λειτουργίας του κυτταροσκελετού, βιολογικής και κυτταρικής προσκόλλησης. Η διεύρυνση των συγκρίσεων, η διεξοδική ανάλυση όλων των εμπλουτισμένων κυτταρικών διεργασιών και η δοκιμή επιπλέον υπολογιστικών εργαλείων, αποτελούν τους στόχους για μελλοντική έρευνα και γενίκευση των αποτελεσμάτων.

Abstract

Skin cutaneous melanoma is considered to be a very aggressive type of cancer and it definitely one of the most dangerous skin cancers Therefore, melanoma has two main characteristics that make it a perfect analysis model, there is plenty of information and data available for processing and it is considered to be a contemporary disease. Melanoma was selected as o model disease for differential expression analysis between two different situations. A group of 472 patients / samples was the initial input for the procedures, who were divided into sub-groups based on demographic criteria like age at diagnosis and sex and also, based on genetic criteria, aka the presence or not of a mutated BRAF, RAS family or NF1 genes. A fourth sub-group was also created to include all patients that did not have any of these mutations. The analysis lead to the extraction of lists of differentially expressed genes for each case and comparison. All procedures were implemented using R programming language and more precisely edgeR package. Extracted data were used as an input for BioInfoMiner tool for furthermore analysis, at the level of molecular function. The most interesting analysis was between patients that carry a BRAF mutation and patients without the condition. In this case BioInfoMiner returned results that indicate a relationship between the presence of mutant BRAF gene and the over or under expression of genes that participate in processes like blood vessels generation, inflammatory response and cell adhesion. Further investigation needs to be done.

Περιεχόμενα

1 Εισαγωγή

1.1 Δερματικό μελάνωμα

1.1.1 Σύντομη περιγραφή του καρκίνου

Ως καρκίνος περιγράφεται η παθολογική κατάσταση κατά την οποία καταρρέουν ή δυσλειτουργούν οι διαδικασίες της κυτταρικής διαίρεσης, της κυτταρικής διαφοροποίησης και του κυτταρικού θανάτου σε συγκεκριμένα κύτταρα ή ομάδες κυττάρων. Συνήθως τα καοήθη κύτταρα που σχηματίζουν όγκους, είναι επιθηλιακής προέλευσης και ονομάζονται καρκινώματα. Σε πολλά όργανα (στήθος, πνεύμονες, ουροδόχος κύστη) η πλειονότητα των καρκίνων που εμφανίζονται είναι καρκινώματα. Αυτές οι περιπτώσεις, παρόλο που εμφανίζουν ορισμένα κοινά χαρακτηριστικά, δεν παύουν να είναι ετερογενείς, να δημιουργούνται από διαφορετικά αίτια, να παρουσιάζουν διαφορετική συμπτωματολογία και να απαιτούν διαφορετική θεραπευτική προσέγγιση. Η βιολογική βάση που οδήγησε στην εμφάνιση κακοήθειας, η επίδραση περιβαλλοντικών παραγόντων και οι επιλογές πρόληψης και έγκαιρης γνωμάτευσης είναι μόνο μερικοί από τους παράγοντες που εμπλέκονται στην εμφάνιση της νόσου και την εξέλιξή της (International Agancy for Research on Cancer, n.d.).

Στο επίπεδο του πληθυσμού των ασθενών, οι συχνότερα εμφανιζόμενες μορφές καρκίνου παγκοσμίως, εξαιρουμένων των καρκίνων του δέρματος που δεν θεωρούνται μελανώματα, είναι ο καρκίνος του πνεύμονα, ο καρκίνος του μαστού και ο καρκίνος του παχέως εντέρου. Δεδομένου αυτού, έχει νόημα να εξεταστεί και η συσχέτιση ανάμεσα στη συχνότητα εμφάνισης της νόσου και τη θνησιμότητα. Έτσι παρέχεται μια πρώτη ένδειξη για την αποτελεσματική ή μη πρόληψη, καθώς παρόμοια επίπεδα εμφάνισης περιστατικών και θνησιμότητας είναι ενδεικτικά μιας θανατηφόρου νόσου που δεν έχει προληφθεί και δεν έχει διαγνωστεί στο προγενέστερο δυνατό στάδιο. Τέτοιου είδους μελέτες οδηγούν στο συμπέρασμα πως ο καρκίνος του πνεύμονα είναι η σημαντικότερη αιτία πρόκλησης θανάτου από καρκίνο στον κόσμο, από τη στιγμή που συσχετίζεται με κακή πρόγνωση. Από την άλλη πλευρά, η καίρια και σωστή παρέμβαση είναι συχνά σωτήρια στο να αποφευχθούν τα θανατηφόρα αποτελέσματα, με χαρακτηριστικό παράδειγμα την περίπτωση του καρκίνου

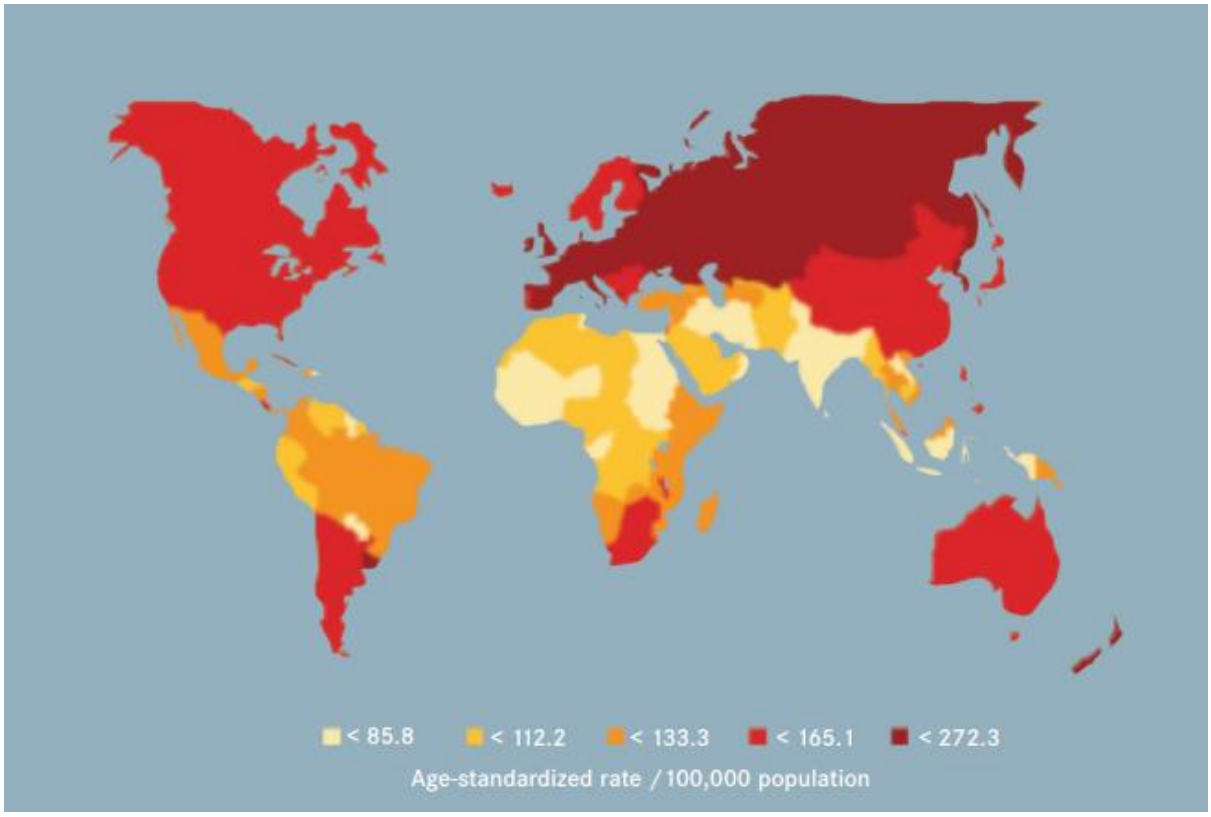
του μαστού. Η κλίμακα, λοιπόν, των περιπτώσεων καρκίνων που εμφανίζουν μεγάλο αριθμό ασθενών που καταλήγουν λόγω της προόδου της νόσου, είναι κατά σειρά ο καρκίνος των πνευμόνων, ο καρκίνος του στομάχου και ο καρκίνος του ήπατος (International Agancy for Research on Cancer, n.d.).

Στο επίπεδο του φύλου, σε χαρακτηριστικά υψηλά ποσοστά εμφανίζεται ο καρκίνος του πνεύμονα στους άνδρες, με αυτούς του οισοφάγου, του στομάχου και της ουροδόχου κύστης να δείχνουν επίσης μια «προτίμηση» στο φύλο αυτό. Αυτή η διαφοροποίηση στον πληθυσμό οφείλεται κατά κύριο λόγο στην διαφορετική έκθεση των δύο φύλων σε καρκινογόνους και μεταλλαξογόνους παράγοντες, ενώ σε πολύ μικρότερο ποσοστό και σπανιότερα σε εγγενή ευαισθησία των ατόμων της μιας ή της άλλη κατηγορίας. Σε αντίθεση με τα παραπάνω, άλλες μορφές καρκίνου, περιλαμβανομένων εκείνων του παχέως εντέρου και του παγκρέατος, παρουσιάζουν μικρές διαφορές στην εμφάνιση της νόσου στα δύο φύλα. Γενικά μιλώντας, η σχέση ανάμεσα στην ανάπτυξη της νόσου και τη θνησιμότητα δεν επηρεάζεται από το φύλο, αν και υπάρχει μια μειοψηφία περιπτώσεων που θα μπορούσε να δώσει ενδείξεις για το αντίθετο, αν εξεταστεί μονομερώς (International Agancy for Research on Cancer, n.d.).

Τα αποτελέσματα της θεραπευτικής προσέγγισης, που θα μπορούσαν ίσως να βελτιώσουν τα παραπάνω ποσοστά θνησιμότητας, ποικίλουν. Κάποιες μορφές όγκων ανταποκρίνονται πολύ καλά στη θεραπεία και αυτό κάνει παραδείγματος χάρη τους καρκίνους του μαστού, του προστάτη και του τραχήλου της μήτρας να αποτελούν αιτία θανάτου σε μία μειοψηφία των ασθενών που εμφανίζουν τη νόσο, γίνεται σωστή διάγνωση και σωστή καταγραφή των περιστατικών (International Agancy for Research on Cancer, n.d.). Η τελευταία πρόταση δίνει το έναυσμα για την εξαγωγή συμπερασμάτων σχετικά με την πρόσβαση σε προσωπικό υγείας και την καταγραφή των ασθενών με καρκίνο, οι οποίες ανά περιοχή του κόσμου ή και ομάδα πληθυσμού, δεν είναι επαρκείς.

Το καρκινικό φορτίο κατανέμεται άνισα μεταξύ του αναπτυγμένου και του αναπτυσσόμενου κόσμου, με συγκεκριμένες μορφές καρκίνου να εμφανίζουν διαφορετικά πρότυπα κατανομής. Όλη η Ευρώπη, η Ιαπωνία, η Αυστραλία, η Νέα Ζηλανδία και η Βόρεια Αμερική κατηγοριοποιούνται στις αναπτυγμένες περιοχές, ενώ η Αφρική, η Λατινική Αμερική, η

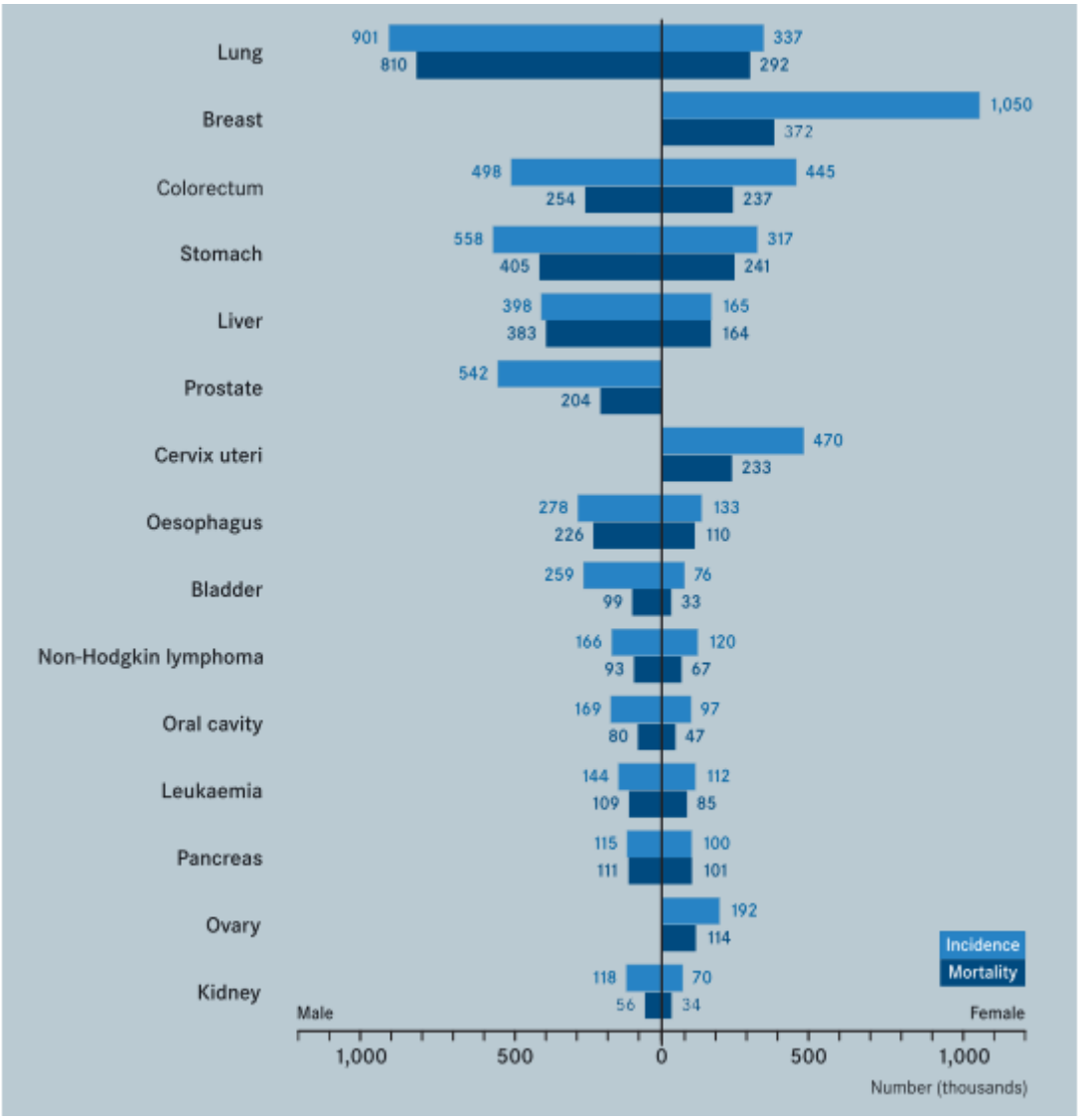
Καραϊβική, η Ασία, εκτός της Ιαπωνίας, η Μικρονησία και η Πολυνησία θεωρούνται αναπτυσσόμενες ή λιγότερο αναπτυγμένες περιοχές. Η διαφοροποίηση αυτή στην κατανομή των μορφών καρκίνου ανά περιοχή του κόσμου απεικονίζεται στον παρακάτω χάρτη.



Εικόνα 1 Απεικόνιση του ποσοστού θνησιμότητας, σε άντρες, συνδυαστικά για όλες τις μορφές καρκίνου, εκτός των δερματικών νεοπλασιών που δεν εμπίπτουν στην κατηγορία του μελανώματος. Τα υψηλότερα ποσοστά εμφανίζονται σε περιοχές του αναπτυγμένου κόσμου (International Agancy for Research on Cancer, n.d.).

Για την κατασκευή τέτοιων χαρτών είναι απαραίτητο να υπάρχει η πληροφορία του πλήθους των περιστατικών για κάθε χώρα, όπου αυτό είναι δυνατό, το οποίο συνήθως υπολογίζεται από υποδομές υπεύθυνες για τη μελέτη του καρκίνου και τη συγκέντρωση των δεδομένων, οι οποίες εξετάζουν ολόκληρο τον πληθυσμό σε εθνικό επίπεδο ή ένα αντιπροσωπευτικό του δείγμα. Ο αριθμός των φορέων καταγραφής αυτών έχει αυξηθεί τα τελευταία χρόνια και η πληροφορία που παρέχεται είναι πιο πλούσια, πιο ακριβής και σε κάποιες περιπτώσεις ειδική για υποομάδες με συγκεκριμένα χαρακτηριστικά ενδιαφέροντος. Παράλληλα, παρέχουν στατιστικά για τη θνησιμότητα από τη νόσο. Με δεδομένα για τον αριθμό των περιστατικών και την επιβίωση των ασθενών, ο αριθμός των ατόμων που πάσχουν από καρκίνο και διαγνώστηκαν μέσα στην προηγούμενη πενταετία είναι δυνατόν να υπολογιστεί προσεγγιστικά. Η θνησιμότητα των ατόμων που πάσχουν από τη νόσο, σε πολλές χώρες είναι

δυνατόν να υπολογιστεί με γνώμονα την καταχώρησης των πράξεων γέννησης και θανάτου. Παρόλα αυτά η ποιότητα των δεδομένων σε ότι αφορά την ακρίβεια, όσο και σε ότι αφορά την πληρότητα κάθε καταγραφής ποικίλλουν σημαντικά.



Εικόνα 2 Αριθμός περιστατικών και θνησιμότητα στις πιο συχνά εμφανιζόμενες μορφές καρκίνου (International Agancy for Research on Cancer, n.d.)

Η καταγραφή της θνησιμότητας των ασθενών που πάσχουν από καρκίνο γίνεται από τον Παγκόσμιο Οργανισμό Υγείας (WHO) σε εθνικό επίπεδο, παρ' όλες τις τεχνικές δυσκολίες που προκύπτουν κατά την συλλογή και επικαιροποίηση των δεδομένων, οι οποίες πολλές φορές δημιουργούν την ανάγκη για πιο σύνθετες στατιστικές αναλύσεις και μεθόδους.

Από τον Παγκόσμιο Οργανισμό Υγείας και τα δεδομένα που παρέχει, ως αίτια της νόσου παρουσιάζονται εμπεριστατωμένα το κάπνισμα, η κατανάλωση αλκοόλ, η περιοδική έκθεση σε καρκινογόνους παράγοντες, όπως είναι η έκθεση σε συγκεκριμένες χημικές ουσίες, η περιβαλλοντική μόλυνση, οι τοξίνες των τροφίμων, σπάνιες περιπτώσεις φαρμάκων, η ιονίζουσα ακτινοβολία, η ακτινοβολία UV, οι χρόνιες μολύνσεις, η δίαιτα και οι διατροφικές συνήθειες, η ανοσοκαταστολή, η γενετική προδιάθεση, συγκεκριμένες ορμόνες, που η λαμβάνονται εξωγενώς είτε εκκρίνονται από τον οργανισμό σε συγκεκριμένα στάδια και υπό συγκεκριμένες συνθήκες.

Σε ότι αφορά τον καρκίνο σε κυτταρολογικό και οργανικό επίπεδο, οι όγκοι που αναπτύσσονται αποτελούνται από κύτταρα των οποίων τα αναπτυξιακά και μορφολογικά χαρακτηριστικά διαφοροποιούνται σε σχέση με τα φυσιολογικά. Τα κριτήρια ώστε να χαρακτηρίσουμε την εμφάνιση ενός όγκου ως κακοήθεια περιλαμβάνουν τον αυξημένο ρυθμό κυτταρικής διαίρεσης, την απουσία κυτταρικής διαφοροποίησης, την ανεξέλεγκτη ανάπτυξη και την μετάσταση σε άλλα όργανα.

Οι αναμορφώσεις αυτές, προκειμένου να προκύψει η κακοήθεια, γίνονται σε πολλαπλά στάδια και συνήθως ξεκινούν από καλοήθεις σχηματισμούς που προοδευτικά καταλήγουν σε κακοήθεις όγκους. Αυτή η εξέλιξη των κυττάρων προκαλείται από την συγκέντρωση μεταλλαγών στο επίπεδο της αλληλουχίας των γονιδίων, τα οποία είναι υπεύθυνα για έλεγχο του κυτταρικού πολλαπλασιασμού, του κυτταρικού θανάτου και τη διατήρηση της κυτταρικής ακεραιότητας. Η ανάπτυξη του καρκίνου μπορεί να πυροδοτηθεί από περιβαλλοντικούς παράγοντες ή από γενετικούς παράγοντες (International Agancy for Research on Cancer, n.d.) και είναι μια πολυεπίπεδη διαδικασία που εμπλέκει πληθώρα κυτταρικών μονοπατιών που σχετίζονται με διαδικασίες της κυτταρικής διαίρεσης και κυτταρικού θανάτου.

1.1.2 Γενική ανασκόπηση του μελανώματος

Το μελάνωμα είναι μία ιδιαίτερα επιθετική μορφή καρκίνου και σίγουρα ο πιο επικίνδυνος τύπος καρκίνου του δέρματος. Το 2012 καταγράφηκαν 232.000 νέες περιπτώσεις ανθρώπων με μελάνωμα, ενώ το 2015 υπήρχαν 3,1 εκατομμύρια ασθενών, οι οποίες περιπτώσεις κατέληξαν σε 59.800 θανάτους. Η Αυστραλία και η Νέα Ζηλανδία έχουν τα υψηλότερα

ποσοστά μελανώματος στον κόσμο. Επίσης υψηλά ποσοστά έχουν η βόρεια Ευρώπη και η βόρεια Αμερική ενώ είναι πιο σπάνιο στην Ασία, την Αφρική και τη Λατινική Αμερική. Το μελάνωμα είναι ελαφρώς πιο συνηθισμένο σε άντρες σε σχέση με τις γυναίκες και έχει γίνει κοινότοπο από τη δεκαετία του 1960 και μετά, σε περιοχές που κατοικούνται κυρίως από πληθυσμούς με λευκό χρώμα δέρματος. Βγάζοντας ένα γρήγορο συμπέρασμα, από την παρατήρηση των τοπολογικών δεδομένων συνάγεται πως το μελάνωμα αποτελεί μία νόσο του αναπτυγμένου κόσμου και σχετίζεται συχνά με τις επιλογές του σύγχρονου τρόπου ζωής (“Melanoma of the Skin - Cancer Stat Facts,” n.d.).

Η εν λόγω νόσος προκαλείται από τον ανεξέλεγκτο πολλαπλασιασμό των μελανοκυττάρων ο οποίος οδηγεί σε κακοήθεια. Τα μελανοκύτταρα είναι κύτταρα του δέρματος τα οποία παράγουν μελανίνη και αποτελούν την κατηγορία κυττάρων στην οποία αναπτύσσεται η ασθένεια σε περισσότερο από 95% των περιπτώσεων. Τα μελανοκύτταρα υπάρχουν κατά κύριο λόγο στο δέρμα, όπου εμφανίζεται το 95% των περιπτώσεων μελανώματος, όμως βρίσκονται επίσης στους βλεννογόνους του σώματος, στη μύτη, στον κόλπο και στους όρχεις. Πολλές φορές το μελάνωμα μπορεί να αναπτυχθεί από μία ελιά ή φακίδα, με παρατηρήσιμες αλλαγές, όπως η αύξηση μεγέθους, ακανόνιστες άκρες, διαφοροποιημένο χρώμα και η εμφάνιση συμπτωμάτων φαγούρας και δερματοπάθειες στην περιοχή. Τα παραπάνω διαφοροποιούνται συχνά ανάλογα με την ηλικία του ασθενούς, δηλαδή αλλάζει η μορφή του σπίλου μεταξύ νεότερων και γηραιότερων ασθενών που εμφανίζουν τη νόσο (Azoury & Lange, 2014) .

Σύμφωνα με την ανασκόπηση του WHO (Παγκόσμιος Οργανισμός Υγείας) του 2014, η οποία περιλάμβανε δεδομένα για μέχρι και το 2000, υπάρχουν περίπου 133.000 νέα περιστατικά μελανώματος σε ολόκληρο τον κόσμο κάθε χρόνο, με τον αριθμό αυτό να παρουσιάζει αύξηση. Οι προβλέψεις περί αύξησης του ρυθμού επίπτωσης της ασθένειας και για της χρονολογίες μετά το 2014, επιβεβαιώνονται απόλυτα από το πλήθος των νέο-διαγνωσθέντων ασθενών το 2015, οι οποίοι όμως δεν έχουν συμπεριληφθεί ακόμα σε κάποια ενιαία αναφορά εύρους δεκαετίας, όπως είναι η παρούσα (International Agancy for Research on Cancer, n.d.).

Το κακόηθες δερματικό μελάνωμα εμφανίζεται κυρίως σε πληθυσμούς με λευκό δέρμα, Καυκάσιους, που ζουν σε χώρες όπου υπάρχει αυξημένη ένταση υπεριώδους ακτινοβολίας, όμως ως ένα βαθμό η συγκεκριμένη κακοήθεια επηρεάζει όλους τους πληθυσμούς. Σε ότι αφορά το χρώμα του δέρματος, τα περιστατικά μελανώματος είναι σημαντικά μειωμένα καθώς αυξάνονται τα επίπεδα μελανίνης και η ασθένεια είναι πολύ σπάνια σε πληθυσμούς με σκούρο χρώμα επιδερμίδας (GBD 2015 Disease and Injury Incidence and Prevalence Collaborators, 2016).

Το μεγαλύτερο πλήθος περιστατικών μελανώματος εμφανίζεται στην Αυστραλία, όπου ο πληθυσμός είναι κατά κύριο λόγω λευκής επιδερμίδας, επικρατεί ηλιοφάνεια 6 ώρες την ημέρα κατά μέσο όρο και υιοθετείται ένας τρόπος ζωής που ευνοεί την έκθεση στον ήλιο. Ως εκ τούτου, η πιθανότητα να αναπτύξει κάποιος μελάνωμα στην Αυστραλία είναι 4%-5% για τους άντρες και 3%-4% για τις γυναίκες.

Η σκουρόχρωμη επιδερμίδα προδιαθέτει αρνητικά για την εμφάνιση της νόσου. Στους πληθυσμούς της Αφρικής και της νότιας Αμερικής, το πέλμα, που είναι η περιοχή που εμφανίζει μικρότερη έκκριση μελανίνης, είναι το σημείο του σώματος στο οποίο εμφανίζεται πιο συχνά μελάνωμα. Οι ασιατικοί πληθυσμοί παρουσιάζουν επίσης μικρό ρίσκο ανάπτυξης μελανώματος, παρά την ωχρότητα του δέρματός τους.

Το ποσοστό θνησιμότητας από τη νόσο είναι ελαφρώς αυξημένο για τους άντρες σε σχέση με τις γυναίκες, η διαφορά, όμως είναι πολύ μικρή. Παρόλο, που το μελάνωμα μπορεί να εμφανιστεί οπουδήποτε στο δέρμα η πλειονότητα των μελανωμάτων στους άντρες εμφανίζεται στην πλάτη, ενώ στις γυναίκες στα πόδια. Η διαφοροποίηση αυτή στην παθούσα περιοχή δεν μπορεί να επεξηγηθεί τελείως από την διαφορική έκθεση σε υπεριώδη ακτινοβολία ανά φύλο και περιοχή δέρματος. Μια τέτοια προσπάθεια θα αποτελούσε απλούστευση της διαδικασίας, ενώ στην πραγματικότητα εξαρτάται από πολλούς παράγοντες, γενετικούς και περιβαλλοντικούς.

Αν και είναι σαφές το γενετικό πλαίσιο των πληθυσμών που είναι επιρρεπείς στο μελάνωμα και αποδεδειγμένα παίζει πολύ σημαντικό ρόλο, λίγες είναι οι περιπτώσεις μελανώματος που μπορούν να αποδοθούν εξ ολοκλήρου σε συγκεκριμένη γενετική προδιάθεση σε αυτούς τους πληθυσμούς. 10% των ασθενών με μελάνωμα έχουν έναν πρώτου βαθμού συγγενή ο

οποίος έχει επηρεαστεί από τη νόσο όμως λιγότερο από 3% των μελανωμάτων στην Αυστραλία μπορεί να αποδοθεί σε μία κληρονομήσιμη γενετική αλλαγή. Το οικογενειακό μελάνωμα είναι ακόμη σπανιότερο σε χώρες που εμφανίζουν μικρότερο αριθμό περιστατικών. Γενικά, αποτελεί ετερόζυγο χαρακτηριστικό και οι μεταλλαγές για το οικογενειακό μελάνωμα εμφανίζονται στους χρωμοσωμικούς βραχίονες 1p, 9p, 12q. Η ανάπτυξη μελανώματος πολλές φορές σχετίζεται με την απουσία λειτουργίας ή ενίσχυση λειτουργίας ογκοκατασταλτικών γονιδίων ή ογκογονιδίων της οικογένειας των κυκλινών (Bodenham, 1968; GBD 2015 Disease and Injury Incidence and Prevalence Collaborators, 2016).

Οι μεταλλαγές απώλειας λειτουργίας στον υποδοχέα της ανθρώπινης μελανωκορτίνης-1 έχουν συσχετιστεί με το κόκκινο χρώμα μαλλιών, το χλωμό δέρμα και τη μειωμένη ικανότητα μαυρίσματος, όλα φυσικά χαρακτηριστικά τα οποία επηρεάζουν την ευπάθεια στον καρκίνο του δέρματος. Περίπου 20% των οικογενειών που είναι επιρρεπείς στο μελάνωμα διαθέτουν μία κληρονομήσιμη μετάλλαξη στο γονίδιο CDKN2A, το οποίο κωδικοποιεί για την p16INK4A πρωτεΐνη. Μεταλλαγές επίσης στο γονίδιο που κωδικοποιεί για την CDK4 έχει δειχθεί πως είναι εξαιρετικά σπάνιες, όμως εμφανίζονται σε κάποιες περιπτώσεις.

Τα γονίδια τα οποία έχει βρεθεί να έχουν κάποιο ρόλο στην ανάπτυξη μελανώματος περιλαμβάνουν το CDKN2A και PTEN, ενώ οι χρωμοσωμικές περιοχές 1p, 6a, 7x, 11q, μπορεί επίσης να σχετίζονται. Περίπου 20% των μελανωμάτων εμφανίζουν μεταλλαγή στο γονίδιο p53. Τα οζώδη μελανώματα εμφανίζουν διπλασιασμό του ογκογονιδίου MYC. Υπάρχουν και άλλες γενετικές τροποποιήσεις που σχετίζονται με την εμφάνιση μελανώματος με πιο πρόσφατη τη μεταλλαγή του BRAF, το οποίο εμφανίζεται μεταλλαγμένο σε μεγάλο ποσοστό των ασθενών, έως και σε 50% των περιπτώσεων και η μετάλλαξή του έχει σαν αποτέλεσμα τη διαταραχή της σηματοδότησης στα επόμενα βήματα του μονοπατιού MAPK (Johns Hopkins University. School of Hygiene and Public Health. & Society for Epidemiologic Research (U.S.), n.d.).

1.1.3 Ιστορικά στοιχεία για το μελάνωμα

Η λέξη μελάνωμα προέρχεται από τα νέα λατινικά του δέκατου ένατου αιώνα και συνδυάζει μορφές που προέρχονται από τα αρχαία ελληνικά και τη ρίζα της λέξης «melano» για να

υποδείξει τη μελανίνη και την κατάληξη –ώμα, η οποία υποδεικνύει μία κυτταρική μάζα και πιο συγκεκριμένα μια νεοπλασία. Ακόμη πιο αναλυτικά το «μελάνω-» προέρχεται από την Ελληνική λέξη «μέλας» που σημαίνει «σκούρος» και την κατάληξη «-ωμα» που χρησιμοποιείται στο κλείσιμο λέξεων που δηλώνουν διεργασίες. Αρχικά, η λέξη «μελάνωμα» χρησιμοποιούνταν με μία πιο ευρεία έννοια και αναφερόταν σε όλους τους μελανοκυτταρικούς όγκους καλοήθεις ή κακοήθεις, αλλά συνήθως κακοήθεις. Σήμερα, όμως η λέξη «μελάνωμα» περιλαμβάνει ένα πολύ πιο στενό πεδίο και αναφέρεται στις κακοήθειες. Παράλληλα, έχει γίνει τόσο κοινή ονομασία αυτών των περιπτώσεων που οι καλοήθεις τύποι δερματικών ανωμαλιών δεν αναφέρονται καν ως μελανώματα πλέον (Dorland, 2003).

Το μελάνωμα δεν είναι μία νέα ασθένεια, όμως οι αποδείξεις σε σχέση με την συχνότητα εμφάνισής του ήταν μέχρι πρόσφατα αρκετά σπάνιες. Παρόλα αυτά τη δεκαετία του 1960 πραγματοποιήθηκε εξέταση ραδιοχρονολόγησης σε 9 από τις μούμιες που βρέθηκαν στο Περού, οι οποίες είχαν ηλικία 2.400 ετών. Στις μούμιες αυτές βρέθηκαν εμφανή σημάδια μελανώματος δηλαδή μελανωματικές μάζες στο δέρμα και διεισδυτικές μεταστάσεις στα οστά. Ο John Hunter φαίνεται να είναι ο πρώτος γιατρός ο οποίος χειρούργησε μεταστατικό μελάνωμα το 1787. Παρόλο που, προφανώς, δε γνώριζε τι ακριβώς ήταν το μελάνωμα, το περιέγραψε ως μυκητιακή έκκριση καρκινικής φύσης. Ο όγκος που αφαιρέθηκε συντηρήθηκε στο Hunterian Museum στο Royal College of Surgeons of England. Μέχρι το 1968, όμως, δεν είχε πραγματοποιηθεί μικροσκοπική εξέταση του δείγματος, η οποία όταν έγινε αποκάλυψε ότι ήταν ένας ιστός μεταστατικού μελανώματος (Johns Hopkins University. School of Hygiene and Public Health. & Society for Epidemiologic Research (U.S.), n.d.). Αργότερα, ο Γάλλος γιατρός Rene Laennec ήταν ο πρώτος που περιέγραψε το μελάνωμα ως ασθένεια (Laennec RTH, 1806). Η αναφορά του παρουσιάστηκε για πρώτη φορά το 1804 και δημοσιεύτηκε το 1806, ενώ η πρώτη αναφορά και περιγραφή στην αγγλική γλώσσα πραγματοποιήθηκε το 1820 από τον William Norris. Στην σχετική αναφορά του, το 1857 ο Norris ισχυρίστηκε πως υπάρχει οικογενειακή προδιάθεση για την ανάπτυξη μελανώματος. Ήταν επίσης πρωτοπόρος σε ότι αφορά τη σχέση ανάμεσα στο είδος και στο μελάνωμα και στην πιθανότητα να υπάρχει συσχέτιση ανάμεσα στην εμφάνιση μελανώματος και τις περιβαλλοντικές εκθέσεις, παρατηρώντας ότι οι περισσότεροι ασθενείς είχαν χλωμή

επιδερμίδα (Norris, 1820). Το 1840 ο Samuel Cooper γνώρισε πώς το μελάνωμα σε προχωρημένο στάδιο δεν είναι ιάσιμο και επισήμανε πως η μόνη πιθανότητα θεραπείας υπάρχει όταν το μελάνωμα εντοπιστεί νωρίς και αφαιρεθεί (Cooper, 1844).

1.1.4 Μελάνωμα και υπεριώδης ακτινοβολία

Υπολογίζεται πως το 80% των περιπτώσεων μελανώματος, προκαλείται από την έκθεση σε υπεριώδη ακτινοβολία (UV) και τις βλάβες που μπορεί να προκαλέσει αυτή στα δερματικά κύτταρα. Η UV ακτινοβολία μπορεί να προέρχεται είτε από τον ήλιο είτε από πηγές όπως τις συσκευές τεχνητού μαυρίσματος. Το ποσοστό αυτό αφορά περιπτώσεις με ευαίσθητο δέρμα, δηλαδή δέρμα που καίγεται εύκολα, λευκό ή κοκκινωπό με πολλές φακίδες που δεν μαυρίζει και αναπτύσσει δυσμορφίες, αφυδάτωση, έκζεμα, έγκαυμα ή άλλες αλλοιώσεις, ως αντίδραση στην έκθεση στην ηλιακή ακτινοβολία. Η πρόληψη του μελανώματος βασίζεται στον περιορισμό της έκθεσης στην ηλιακή ακτινοβολία, ειδικά τα 20 πρώτα χρόνια της ζωής (GBD 2015 Disease and Injury Incidence and Prevalence Collaborators, 2016; Kanavy & Gerstenblith, 2011).

Η υπεριώδης ακτινοβολία μπορεί να αποδειχθεί ιδιαίτερα βλαβερή όταν περιλαμβάνει σποραδική έντονη έκθεση στον ήλιο και ηλιακά εγκαύματα. Το μεγαλύτερο ποσοστό βλάβης προκαλείται από την ηλιακή ακτινοβολία που εκτίθεται το άτομο κατά τη διάρκεια της παιδικής ηλικίας και της εφηβείας, μετατρέποντας αυτά τα ηλικιακά γκρουπ στον πιο σημαντικό στόχο των προγραμμάτων πρόληψης. Άλλες πιο σπάνιες αιτίες περιλαμβάνουν την ανοσοκαταστολή και την εφαρμογή σολάριουμ, όπως προαναφέρθηκε. Δεδομένων των παραπάνω, διαφαίνεται πως οι μόνοι τρόποι πρόληψης είναι η χρήση αντηλιακού και η αποφυγή έκθεσης σε ακτινοβολία UV (International Agancy for Research on Cancer, n.d.).

Η UV ακτινοβολία μπορεί να προκαλέσει βλάβες στο DNA των κυττάρων συνήθως διμερισμό της θυμίνης, οι οποίες αν δεν επιδιορθωθούν από τους επιδιορθωτικούς μηχανισμούς του DNA, μπορεί να οδηγήσουν σε μεταλλάξεις στα γονίδια του κυττάρου. Καθώς το κύτταρο που φέρει αυτές τις μεταλλαγές διαιρείται, είναι λογικό να διαιωνίζονται στις επόμενες γενιές κυττάρων. Αν οι μεταλλαγές αυτές παρουσιαστούν σε ογκογονίδια ή ογκοκατασταλτικά γονίδια, η διαδικασία της μίτωσης στα κύτταρα που φέρουν τη μεταλλαγή μπορεί να γίνει ανεξέλεγκτος και να οδηγήσει στην δημιουργία ενός όγκου.

Γενικώς οι καρκίνοι αναπτύσσονται από καταστροφές στο DNA εξαιτίας της ακτινοβολίας UVA που δημιουργεί διμερή θυμίνης. Η UVA, επίσης ενεργοποιεί την παραγωγή ενεργών ριζών οξυγόνου και αυτές επάγουν βλάβες στο DNA διαφορετικού τύπου, κυρίως σπασίματα στον έναν από τους δύο κλώνους και οξειδωμένες πυριμιδίνες. Αν δεν επιδιορθωθούν αυτές οι βλάβες από τους επιδιορθωτικούς μηχανισμούς του κυττάρου μπορούν να οδηγήσουν σε έκφραση διαφοροποιημένων πρωτεϊνικών μορίων.

Εξαιτίας του ότι η UV ακτινοβολία έχει σημαντικό ρόλο σε ότι αφορά τις βλάβες του DNA που σχετίζονται με το μελάνωμα υπάρχει μία συγκεκριμένη κατηγορία μεταπτώσεων που θεωρείται από τους επιστήμονες ως μεταλλάξεις, που αν βρεθούν, είναι βέβαιο πως το κύτταρο έχει δεχθεί τις αρνητικές επιδράσεις της UV ακτινοβολίας. Σε αυτή την περίπτωση, το κύτταρο, άρα και το αντίστοιχο γονιδίωμα έχει UV υπογραφή. Αυτές οι μεταβολές στην αλληλουχία του DNA καταλήγουν σε φωτο-προϊόντα, σε διαγραφές βάσεων, εισαγωγή νουκλεοτιδίων και χρωμοσωμικές ανωμαλίες (Kanavy & Gerstenblith, 2011) .

1.1.5 Φυσιολογία του μελανώματος

Τα πρώτα στάδια του μελανώματος ξεκινούν όταν τα μελανοκύτταρα αρχίζουν να πολλαπλασιάζονται ανεξέλεγκτα. Τα μελανοκύτταρα βρίσκονται ανάμεσα στην εξωτερική στιβάδα του δέρματος, δηλαδή την επιδερμίδα και την επομένη στιβάδα, δηλαδή την δερμίδα ή χόριο ή κυρίως δέρμα. Τα πρώτα στάδια της ασθένειας αποτελούν τη ριζική αναπτυξιακή φάση, όταν ο όγκος έχει πάχος λιγότερο από 1mm. Εξαιτίας του ότι τα καρκινικά κύτταρα δεν έχουν ακόμη έρθει σε επαφή με τα αιμοφόρα αγγεία στις βαθύτερες στιβάδες του δέρματος, είναι απίθανο αυτό το πρώτο στάδιο μελανώματος να επεκταθεί και σε άλλες περιοχές του σώματος. Αν το μελάνωμα εντοπιστεί και διαγνωσθεί σε αυτό το στάδιο συνήθως μπορεί να αφαιρεθεί πλήρως χειρουργικά (Goldstein & Goldstein, 2001).

Από την άλλη πλευρά όταν τα καρκινικά κύτταρα αρχίσουν να κινούνται προς διαφορετική κατεύθυνση, κάθετα και προς τα πάνω, στο εσωτερικό της επιδερμίδας και μέσα στο δερματικό επιθήλιο, η κυτταρική συμπεριφορά αλλάζει δραματικά. Στο επόμενο βήμα, το μελάνωμα αποκτά τη δυνατότητα διείσδυσης σε άλλους ιστούς και τη δυνατότητα επέκτασης.

Η φάση κάθετης ανάπτυξης που ακολουθεί αποτελεί το επιθετικό μελάνωμα. Ο όγκος αποκτά την ικανότητα να αυξηθεί σε μέγεθος και να επεκταθεί στους γειτονικούς ιστούς με τη μεταφορά των καρκινικών κυττάρων μέσω των αιμοφόρων ή των λεμφικών αγγείων. Συνήθως το πάχος του όγκου είναι 1mm και όγκος αναπτύσσεται σε βαθύτερα στρώματα της επιδερμίδας. Σε αυτό το επίπεδο μπορεί να υπάρξει ανοσολογική αντίδραση από τα κύτταρα του ανοσοποιητικού συστήματος, τα οποία σε κάποιες περιπτώσεις καταστρέφουν εντελώς τον όγκο. Οπότε, πλέον μόνο ο μεταστατικός όγκος μπορεί να εντοπιστεί (Azoury & Lange, 2014; Goldstein & Goldstein, 2001).

Η θεραπεία περιλαμβάνει κατά κύριο λόγο χειρουργική αφαίρεση του σπίλου ή της ελιάς μου μεταμορφώθηκε σε πρωτογενές μελάνωμα. Στις περιπτώσεις των ασθενών με μεγαλύτερους όγκους, η τοπική λεμφική κόμβοι μπορεί να εξεταστούν για την πιθανότητα το μελάνωμα να έχει επεκταθεί και σε αυτούς. Οι περισσότεροι ασθενείς θεραπεύονται πλήρως αν αυτή η επέκταση δεν έχει συμβεί. Σε αντίθετη περίπτωση, προτείνεται ανοσοθεραπεία, βιολογική θεραπεία, ακτινοθεραπεία ή χημειοθεραπεία, οι οποίες δεν εξαλείφουν τη νόσο, αλλά βελτιώνουν την επιβίωση. Η πιθανότητα το μελάνωμα να επιστρέψει μετά τη θεραπεία ή να επεκταθεί εξαρτάται από το πόσο πυκνό είναι, το ρυθμό διαίρεσης των κυττάρων και από το αν οι κατώτερες δερματικές στιβάδες έχουν διαρρηχθεί (Allen & Spitz, 1953; Goldstein & Goldstein, 2001).

1.2 Η επανάσταση των -omics

Ο όρος “-omics” αναφέρεται στο σύνολο των τομέων της βιολογίας που λήγουν σε –omics όπως genomics, proteomics, metabolomics, transcriptomics, έννοιες οι οποίες στα ελληνικά αφορούν κατά σειρά την επιστήμη που ασχολείται με το γονιδίωμα, το πρωτέωμα, το μεταβόλωμα, το μεταγράφωμα, παρόλο που χρησιμοποιούμε και τους όρους γονιδιωματική, πρωτεωμική, μεταβολωμική. Αυτήν ακριβώς την ομαδοποίηση είναι που δηλώνει και η κατάληξη –ome ή –ωμα στην πρώτη περίπτωση. Η επανάσταση των –omics ήρθε ουσιαστικά με την ανάπτυξη τεχνολογιών αλληλούχισης και για άλλα μακρομόρια όπως το RNA και οι πρωτεΐνες, που ακολούθησαν χρονικά την αλληλούχιση γονιδιωμάτων. Η ανάγκη για ανάπτυξη της διαφορετικής ορολογίας για τους τομείς αυτούς αποτελεί καθαρή ένδειξη της στροφής προς την ποσοτική ανάλυση όλων των μακρομορίων και τη νέα

δυνατότητα ανάλυσης, όχι μόνο ενός γονιδιακού προϊόντος, αλλά όλου του συστήματος των μορίων που σχετίζονται με τον εκάστοτε στόχο.

1.2.1 Γονιδιωματική ή Γενωμική (Genomics)

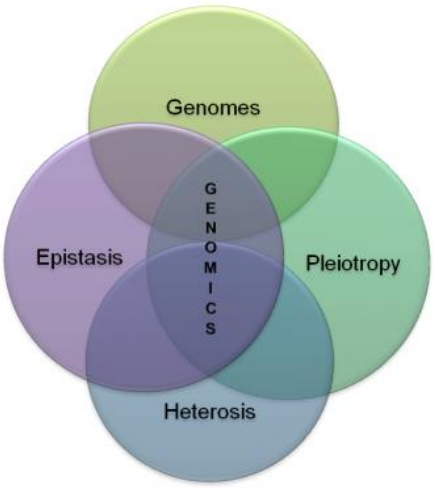
Η Γονιδιωματική είναι ένας διεπιστημονικός χώρος που εστιάζει στη μελέτη της δομής και της λειτουργίας, τη χαρτογράφηση και την επεξεργασία της γονιδιακής πληροφορίας και των γονιδιωμάτων. Το γονιδίωμα αποτελεί το πλήρες σετ DNA ενός οργανισμού συμπεριλαμβανομένων όλων του των γονιδίων.

Εν αντιθέσει με τη Γενετική, η οποία αφορά ατομικά τη μελέτη των γονιδίων και το ρόλο τους στην κληρονομικότητα, η Γονιδιωματική στοχεύει στον συλλογικό χαρακτηρισμό και την ποσοτικοποίηση της γονιδιακής και γενετικής πληροφορίας, η οποία οδηγεί στην παραγωγή πρωτεϊνών με αρωγό ένζυμα και σηματοδοτικά μόρια. Με τη σειρά τους οι πρωτεΐνες έχουν δομικό και λειτουργικό ρόλο στον οργανισμό.

Η Γονιδιωματική επίσης περιλαμβάνει την αλληλούχιση και την ανάλυση των γονιδιωμάτων μέσω της χρήσης της υψηλής απόδοσης DNA αλληλούχισης και της Βιοπληροφορικής, ώστε να συγκεντρώσει και να αναλύσει τη λειτουργία και τη δομή ολόκληρων γονιδιωμάτων οργανισμών.

Η πρόοδος της Γενωμικής έχει πυροδοτήσει την εξέλιξη της έρευνας που βασίζεται στην ανακάλυψη και τη βιολογία συστημάτων, ώστε να διευκολυνθεί η κατανόηση ακόμη και των πιο περίπλοκων βιολογικών συστημάτων, όπως ο εγκέφαλος.

Ο τομέας αυτός περιλαμβάνει ακόμη την ανάλυση φαινομένων που συμβαίνουν εντός των γονιδιωμάτων, όπως η επίσταση, οι πλειοτροπίες, η ετέρωση και άλλες αλληλεπιδράσεις μεταξύ γενετικών τόπων και αλληλομόρφων του γονιδιώματος (“EMBL-European Molecular Biology Laboratory,” n.d.; Lockhart & Winzeler, 2000).



Εικόνα 3 Η τομή της μελέτης των γονιδιωμάτων, των πλειοτροπιών, της επίστασης και της ετέρωσης είναι η Γονιδιωματική ("EMBL-European Molecular Biology Laboratory," n.d.)

Κατά συνέπεια των παραπάνω, μελετά μεγάλης κλίμακας αλλαγές στα γονιδιώματα των οργανισμών. Η Γονιδιωματική και η Μεταγραφωμική (transcriptomics), η οποία μελετά ευρείες γονιδιακές αλλαγές στο RNA που μεταγράφεται από το DNA, μελετούν πολλά γονίδια τη φορά. Η Γονιδιωματική μπορεί επίσης να περιλαμβάνει την ανάγνωση και την ευθυγράμμιση πολύ μακριών ακολουθιών του DNA ή του RNA. Η ανάλυση και ερμηνεία αυτών των μεγάλης κλίμακας σύνθετων δεδομένων απαιτεί τη βοήθεια των υπολογιστών. Το ανθρώπινο μυαλό δεν είναι σε θέση να χειριστεί τόσες πολλές πληροφορίες.

Η Γονιδιωματική μπορεί να χωριστεί σε υποκατηγορίες, μερικές εκ των οποίων είναι η Λειτουργική Γονιδιωματική, η Δομική Γονιδιωματική, η Επιγενετική ή Επιγενωμική και η Μεταγονιδιωματική.

Η Λειτουργική Γονιδιωματική είναι ένας τομέας της Μοριακής Βιολογίας, ο οποίος χρησιμοποιεί το μεγάλο εύρος των διαθέσιμων δεδομένων που έχουν προέλθει από προγράμματα αλληλούχισης, ώστε να περιγράψει τις λειτουργίες και τις αλληλεπιδράσεις των γονιδίων και των πρωτεϊνών. Ως εκ τούτου ο τομέας αυτός εστιάζει σε δυναμικές πλευρές των διαδικασιών του γονιδιώματος όπως η γονιδιακή μεταγραφή και μετάφραση και οι αλληλεπιδράσεις μεταξύ πρωτεϊνών. Η στατική πληροφορία σχετικά με το γονιδίωμα, δηλαδή η αλληλουχία του DNA και η δομή στο χώρο δεν περιλαμβάνεται στον χώρο αυτό.

Το κύριο χαρακτηριστικό της είναι ότι μελετά το γονιδίωμα με μία πιο ευρεία προσέγγιση που περιλαμβάνει μεθόδους υψηλής ανάλυσης (Hieter & Boguski, 1997).

Η Δομική Γονιδιωματική επιχειρεί να περιγράψει την τρισδιάστατη δομή στο χώρο κάθε πρωτεΐνης που κωδικοποιείται από ένα δεδομένο γονιδίωμα. Οι παραπάνω διαδικασίες μπορούν να προσεγγιστούν μέσω ανάλυσης υψηλής απόδοσης για την πρόβλεψη της δομής με συνδυασμό πειραματικών δεδομένων και αλγορίθμων μοντελοποίησης. Η βασική διαφορά της δομικής γονιδιωματικής με την παραδοσιακή πρόβλεψη της δομής είναι ότι η δομική γονιδιωματική επιχειρεί να προσδιορίσει τη δομή κάθε πρωτεΐνης που προέρχεται από το γονιδίωμα και δεν εστιάζει σε ένα συγκεκριμένο μόριο και τη δομή του. Από τη στιγμή που ολόκληρο το γονιδίωμα είναι διαθέσιμο, η πρόβλεψη της δομής μπορεί να γίνει πιο γρήγορα μέσα από τον συνδυασμό πειραματικών μεθόδων και προσεγγίσεων μοντελοποίησης, ειδικά αφού υπάρχει η πρότερη γνώση της αλληλουχίας πολλών γονιδιωμάτων και της δομής πρωτεϊνών. Συνδυαστικά αυτά τα δύο επιτρέπουν στους επιστήμονες να μοντελοποιήσουν μια πρωτεϊνική δομή με βάση τα ομόλογα μόρια που έχουν ήδη λυθεί. Συνοψίζοντας, η δομική γονιδιωματική εμπεριέχει έναν μεγάλο αριθμό προσεγγίσεων με στόχο τον προσδιορισμό της δομής στο χώρο. Πειραματικές μέθοδοι που χρησιμοποιούν τις γονιδιωματικές αλληλουχίες, μέθοδοι μοντελοποίησης βασισμένες στην αλληλουχία ή την δομή ομολόγων πρωτεϊνών, εφαρμογή χημικών και φυσικών παραμέτρων για οποία πρωτεΐνη δεν έχει ομόλογη γνωστή δομή. Συνήθως, η γνώση για την δομή μιας πρωτεΐνης μέσω της δομικής γονιδιωματικής έρχεται πριν από την γνώση της λειτουργίας του εν λόγω μορίου. Αυτό αποτελεί νέο τομέα πρόκλησης για τη δομική Βιοπληροφορική, λόγω της ανάγκης προσδιορισμού της πρωτεϊνικής λειτουργίας ως δεδομένη την τριτοταγή δομή και όχι το αντίστροφο (Vitkup, Melamud, Moult, & Sander, 2001).

Η Επιγενωμική ή Επιγενετική είναι η μελέτη όλων των επιγενετικών τροποποιήσεων του γενετικού υλικού ενός κυττάρου γνωστό και ως επιγονιδίωμα. Οι επιγενετικές τροποποιήσεις είναι αντιστρεπτές τροποποιήσεις στο DNA ενός κυττάρου ή στις ιστόνες, οι οποίες επηρεάζουν την έκφραση χωρίς να αλλάζουν τη γενετική αλληλουχία στη βάση της. Δύο από τις πιο χαρακτηριστικές επιγενετικές τροποποιήσεις είναι η μεθυλίωση του DNA και η τροποποίηση των ιστονών. Οι διαδικασίες αυτές παίζουν σημαντικό ρόλο στην γονιδιακή έκφραση και ρύθμιση και εμπλέκονται σε πολλές κυτταρικές διαδικασίες όπως η

διαφοροποίηση, η ανάπτυξη και η ογκογένεση. Η μελέτη της Επιγενετικής έγινε δυνατή χάρη στην πρόσφατη εφαρμογή των αναλύσεων υψηλής απόδοσης στη DNA αλληλούχιση (Jirtle & Skinner, 2007).

Η Μεταγονιδιωματική αναφέρεται στη μελέτη του μετά-γονιδιώματος, δηλαδή του γενετικού υλικού που παραλαμβάνεται απευθείας από περιβαλλοντικά δείγματα. Μπορεί να αναφέρεται επίσης ως περιβαλλοντική γονιδιωματική, ή πληθυσμιακή γονιδιωματική. Η παραδοσιακή μικροβιολογία και η αλληλούχιση μικροβιακών γονιδιωμάτων βασίζονται σε μονοκλωνικές καλλιέργειες, εν αντιθέσει με την αλληλούχιση μικροβιακών γονιδίων στα πλαίσια της Μεταγονιδιωματικής, η οποία εστιάζει στην κλωνοποίηση συγκεκριμένων γονιδίων, προκειμένου να εξάγει ένα προφίλ ποικιλομορφίας για ένα φυσικό δείγμα. Η διαδικασία αυτή αποκάλυψε πως η μεγάλη πλειονότητα της μικροβιακής βιοποικιλότητας δεν είχε γίνει αντιληπτή και παρατηρήσιμη από τις μεθόδους που βασίζονταν σε καλλιέργειες και χρησιμοποιούνταν αποκλειστικά μέχρι πρόσφατα (Daniel, 2005).

1.2.1.1 Το γονιδίωμα, μια απλή προσέγγιση

Το DNA είναι ένα χημικό μόριο που περιέχει τις πληροφορίες που είναι απαραίτητες για την ανάπτυξη και την καθοδήγηση των δραστηριοτήτων όλων των ζωντανών οργανισμών. Τα μόρια του DNA είναι φτιαγμένα από δύο έλικες που περιελίσσονται μεταξύ τους, στις οποίες συχνά αναφερόμαστε με τον όρο διπλή έλικα. Κάθε κλώνος του DNA είναι φτιαγμένος από 4 βασικά χημικά μόρια τα οποία ονομάζονται νουκλεοτίδια ή/και βάσεις και αποτελούν το γενετικό αλφάβητο.

Κάθε οργανισμός είχε ένα πλήρες σετ DNA το οποίο ονομάζεται γονιδίωμα. Κάθε κύτταρο ενός οργανισμού περιέχει ένα πλήρες αντίγραφο του DNA, το οποίο αποτελείται από περίπου 3 εκατομμύρια βάσεις και συνθέτει το ανθρώπινο γονιδίωμα. Αυτή η αλφάβητος των τεσσάρων γραμμάτων περιέχει όλη την πληροφορία που είναι απαραίτητη για να δομηθεί, να αναπτυχθεί και να λειτουργήσει το ανθρώπινο σώμα.

Παρόλο που δεν υπάρχει ακριβής ορισμός το τι είναι γονίδιο, παραδοσιακά το περιγράφουμε ως μια περιοχή στο DNA, η οποία έχει τις πληροφορίες για την παρασκευή μιας συγκεκριμένης πρωτεΐνης ή μιας ομάδας πρωτεϊνών. Υπολογίζεται επίσης ότι υπάρχουν περί τα 20 με 25.000 γονίδια στο ανθρώπινο γονιδίωμα, τα οποία κωδικοποιούν έναν

πολλαπλάσιο αριθμό πρωτεϊνικών μορίων. Οπότε, ξεκινώντας από το ανθρώπινο σώμα και κινούμενοι όλο και περισσότερο προς το εσωτερικό και προς μικρότερες δομές, ακολουθεί το ανθρώπινο κύτταρο, ο πυρήνας και το DNA το οποίο είναι «πακεταρισμένο» σε 23 ζεύγη χρωμοσωμάτων. Τα ζεύγη αυτά φέρουν δίπλα αντίγραφα των γονιδίων τα οποία καθοδηγούν την παραγωγή πρωτεϊνών, με τη βοήθεια ενζύμων και σηματοδοτικών μορίων. Η διαδικασία αυτή αποτελεί το βασικό δόγμα της βιολογίας, δηλαδή το DNA αντιγράφεται, με στόχο να δημιουργήσει περισσότερα αντίτυπα, απαραίτητα καθώς πολλαπλασιάζονται τα κύτταρα ενός αναπτυσσόμενου οργανισμού, επίσης μεταγράφεται σε ένα άλλο βασικό μακρομόριο, το RNA, το οποίο αποτελεί το ενδιάμεσο στάδιο μεταξύ DNA και πρωτεΐνης για να φτάσει τελικά να μεταφραστεί σε πρωτεϊνικά μόρια τα οποία έχουν δομικό, λειτουργικό και σηματοδοτικό ρόλο στα κύτταρα (Uzman, 2003).

Κατά τη Μοριακή Βιολογία και τη Γενετική, το γονιδίωμα είναι το γενετικό υλικό ενός ο οργανισμού. Η περεταίρω πληροφορία σε σχέση με τον προηγούμενο, απλούστερο ορισμό, αφορά στο ότι το γονιδίωμα αυτό εμφανίζεται στις εξής παραλλαγές, μπορεί να είναι δίκλωνο DNA, μονόκλωνο DNA ή ακόμη και μονόκλωνο RNA μόριο. Σε κάθε μια από αυτές τις περιπτώσεις εξυπηρετεί πλήρως τις λειτουργίες των οργανισμών που το φέρουν και παρέχει όλη τη απαραίτητη γενετική πληροφορία.

Στους DNA οργανισμούς, εμφανίζονται δύο μεγάλες κατηγορίες γονιδιώματος, το κωδικοποιόν DNA, δηλαδή περιοχές όπως τα γονίδια που κωδικοποιούν για κάποιο μόριο downstream μέσω τον διαδικασιών της μεταγραφής και της μετάφρασης, αλλά υπάρχουν και μεγάλες περιοχές μη κωδικοποιόντος DNA (non-coding DNA), που εξυπηρετούν άλλους ρόλους. Παράλληλα, αυτού του είδους τα μακρομόρια συναντώνται στον πυρήνα, τα μιτοχόνδρια και τους χλωροπλάστες των οργανισμών (Uzman, 2003).

Τα ευκαρυωτικά γονιδιώματα, που αποτελούν και το αντικείμενο της εργασίας αυτής, αποτελούνται από ένα ή περισσότερα γραμμικά μόρια DNA. Ο αριθμός των χρωμοσωμάτων ποικίλει ανάλογα με το είδος του οργανισμού. Ένα τυπικό ανθρώπινο κύτταρο, όπως προαναφέρθηκε, περιέχει 23 ζεύγη χρωμοσωμάτων, 22 αυτοσωμικά ζεύγη και ένα φυλετικό. Στους οργανισμούς που αναπαράγονται αμφιγονικά, όπως ο άνθρωπος, τα γαμετικά κύτταρα είναι απλοϊδικά. Τέλος, εν αντιθέσει με τους προκαρυώτες, τα ευκαρυωτικά DNA

έχουν δομή εξονίων – εσονίων, με τα πρώτα να αποτελούν τις περιοχές που μεταγράφονται ή και μεταφράζονται, ενώ τα δεύτερα είναι περιοχές επαναλαμβανόμενου DNA και αποτελούν το μεγαλύτερο τμήμα του ανθρώπινου γονιδιώματος.

1.2.1.2 Αλληλούχιση DNA – Ιστορική Αναδρομή

Η λέξη αλληλούχιση, πολύ απλά, σημαίνει τον καθορισμό της ακριβούς σειράς των βάσεων ενός κλώνου DNA. Από τη στιγμή που οι βάσεις υπάρχουν σε ζεύγη, η γνώση της ταυτότητάς της μιας από τις δύο βάσεις μπορεί να καθορίσει και ποιο θα είναι το ζευγάρι της και ως εκ τούτου δεν χρειάζεται να αλληλουχιθούν και οι δύο αλυσίδες του DNA. Η πιο συνηθισμένη αλληλούχιση που χρησιμοποιείται σήμερα ονομάζεται «αλληλούχιση κατά τη σύνθεση» όπου η DNA πολυμεράση χρησιμοποιείται για να δημιουργήσει ένα νέο κλώνο DNA από έναν κλώνο ενδιαφέροντος που λειτουργεί ως καλούπι. Κατά τη διαδικασία της αλληλούχισης το ένζυμο ενσωματώνει στο νέο κλώνο του DNA νουκλεοτίδια τα οποία έχουν σημανθεί με φθορισμό. Αφού συμβεί αυτό ο φθορισμός ενεργοποιείται από κάποια πηγή φωτός και το φθορίζον σήμα εκπέμπεται και γίνεται ανιχνεύσιμο. Το σήμα είναι διαφορετικό και εξαρτάται από το ποιο από τα 4 νουκλεοτίδια έχει σημανθεί. Αυτή η μέθοδος μπορεί να παράγει τμήματα των 125 νουκλεοτιδίων στη σειρά και δισεκατομμύρια αναγνώσεις μορίων (Olsvik et al., 1993) .

Για να συντεθεί η αλληλουχία σε ένα μεγάλο κομμάτι DNA, όπως το γονίδιο, οι ερευνητές χρειάζεται να εντοπίσουν επικαλυπτόμενες περιοχές πάνω στην αλληλουχία αυτή. Η διαδικασία επιτρέπει να συναρμολογηθεί ένα μεγαλύτερο τμήμα από τα μικρότερα αρχικά κομμάτια.

Είναι προφανές πως είναι απαραίτητο κάθε βάση να μην έχει διαβαστεί μόνο μία φορά, αλλά πολλαπλές αναγνώσεις στα επικαλυπτόμενα τμήματα είναι απαραίτητες για να εξασφαλιστεί η αξιοπιστία της αλληλούχισης (Olsvik et al., 1993).

Οι επιστήμονες χρησιμοποιούν την αλληλούχιση DNA κατά τη διερεύνηση γενετικών παραλλαγών και μεταλλάξεων που δυνητικά έχουν σημαίνων ρόλο στην εμφάνιση ή την πρόοδο μιας ασθένειας. Η αλλαγή από την οποία θα προκληθεί μία ασθένεια μπορεί να είναι κάτι αρκετά μικρό όπως μία υποκατάσταση, μία διαγραφή ή μία προσθήκη ενός μόνο

ζεύγους βάσεων ή κάτι πάρα πολύ μεγάλο όπως η διαγραφή χιλιάδων βάσεων (Pettersson, Lundeberg, & Ahmadian, 2009).

Όσον αφορά την ετυμολογία της λέξης, προέρχεται από το ελληνικό ρήμα «γεννώ» και τα παράγωγά του. Η γονιδιωματική, βασίζεται στην διαθεσιμότητα ολόκληρων DNA μορίων από οργανισμούς, ενώ η αλληλούχιση και η μελέτη αυτών των μακρομορίων έγινε δυνατή μετά την πρωτοποριακή δουλειά του Fred Sanger και την πιο πρόσφατη τεχνολογία της αλληλούχισης νέας γενιάς (Dorland, 2003).

Ιστορικά, μετά τη δημοσίευση της διπλής έλικας του DNA από τους James D. Watson και Francis Crick, βασιζόμενοι στην προγενέστερη εργασία της Rosalind Franklin, το 1953, ο Fred Sanger, υπήρξε πρωτοπόρος σε ότι αφορά μια πρωτογενή μορφή αλληλούχισης, που κατέληξε να είναι ο κύριος στόχος των μοριακών βιολόγων της εποχής. Ο Sanger δημοσίευσε τη νουκλεοτιδική αλληλουχία της ινσουλίνης το 1955 και έκτοτε, συνάδελφοί του συνέχισαν να δημοσιεύουν αλληλουχίες μακρομορίων, που κανένα δεν έφτανε σε έκταση το ανθρώπινο γονιδίωμα (Sanger, Nicklen, & Coulson, 1977).

Η μέθοδος Sanger, διδάσκεται ακόμη και σήμερα στο πανεπιστήμιο γιατί έθεσε τις βάσεις της αλληλούχισης DNA και των υπόλοιπων μακρομοριών, της χαρτογράφησης του γονιδιώματος, της διαχείρισης του όγκου τις πληροφορίας που προκύπτει από τέτοιες διαδικασίες και της Βιοπληροφορικής. Το 1975 δημοσιεύθηκε μια εργασία του σε συνεργασία με τον Alan Coulson, σχετικά με την αλληλούχιση μορίων χρησιμοποιώντας DNA πολυμεράση και ραδιενεργά σεσημασμένα νουκλεοτίδια, η οποία σταδιακά βελτιώθηκε και έφτασε να μπορεί να αλληλουχίσει τα περίπου 5300 νουκλεοτίδια, μονόκλωνου DNA του βακτηριοφάγου φΧ174 (Sanger et al., 1977).

Από τα στάδια αυτά και μετά, ήταν πλέον ανοιχτός ο δρόμος για την ανάπτυξη όλο και πιο αποτελεσματικών μεθόδων, που κάθε φορά υπερείχαν από τις προηγούμενες τεχνολογίες σε αξιοπιστία, ταχύτητα και ικανότητα. Οι πρώτες γονιδιωματικές αλληλουχίες προέρχονταν από το ανθρώπινο μιτοχόνδριο, τον χλωροπλάστη, τον ζαχαρομύκητα *Saccharomyces cerevisiae* και κάποιους ιούς, αρχαία και βακτήρια. Οι περισσότεροι από τους οργανισμούς που αλληλουχίθηκαν αρχικά, αφορούν προφανώς, οργανισμούς βιολογικού ενδιαφέροντος, είτε γιατί αποτελούν σημαντικά παθογόνα, είτε διότι είναι οργανισμοί μοντέλα, όπως η

Drasophila melanogaster και το *zebrafish* (Feldmann et al., 1994; Prober et al., 1987; Smith et al., 1986).

Ο Fred Sanger και η ομάδα του εφηύραν και καθιέρωσαν τεχνικές αλληλούχισης και χαρτογράφησης του γονιδιώματος, αποθήκευσης δεδομένων και Βιοπληροφορικής ανάλυσης, τις δεκαετίες του '70 και του '80. Η δουλειά αυτή άνοιξε το δρόμο για την έναρξη και την ολοκλήρωση του «Προγράμματος του Ανθρώπινου Γονιδιώματος», τη δεκαετία του '90. Κατά τη διάρκεια αυτού του εγχειρήματος συνεργάστηκαν εργαστήρια και επιστήμονες από όλο τον κόσμο, ενώ το αποτέλεσμα ήταν να έχουμε πλήρως αλληλουχημένο το ανθρώπινο γονιδίωμα το 2003 (National Human Genome Research Institute, 2016).

Σήμερα, η τεχνολογία αλληλούχισης επόμενης γενιάς έχει οδηγήσει σε βελτίωση στην ταχύτητα, στην ικανότητα διαχείρισης μνήμης και στην οικονομική προσιτότητα των αλληλουχίσεων γενικά. Επιπλέον, η εξέλιξη στη Βιοπληροφορική έδωσε τη δυνατότητα σε εκατοντάδες βάσεις δεδομένων που σχετίζονται με τις επιστήμες της ζωής, να παρέχουν σημαντική στήριξη στην επιστημονική έρευνα. Η πληροφορία που αποθηκεύεται και είναι οργανωμένη σε αυτές τις βάσεις δεδομένων μπορεί εύκολα να αναζητηθεί, να συγκριθεί με άλλες πληροφορίες και να αναλυθεί.

1.2.1.3 Το Πρόγραμμα του Ανθρώπινου Γονιδιώματος (The Human Genome Program)

Το Πρόγραμμα του Ανθρώπινου Γονιδιώματος υπήρξε ένα διεθνές, επιστημονικό, ερευνητικό πρόγραμμα με στόχο τον προσδιορισμό της αλληλουχίας των νουκλεοτιδικών βάσεων που απαρτίζουν το ανθρώπινο γονιδίωμα καθώς και τον προσδιορισμό του τύπου και της θέσης των γονιδίων πάνω σε αυτό, τόσο από φυσικής όσο και λειτουργικής πλευράς. Ήταν και παραμένει το μεγαλύτερο συνεργατικό πρόγραμμα. Η σύλληψη της ιδέας ανήκει στην κυβέρνηση των Ηνωμένων Πολιτειών Αμερικής το 1984, ενώ τη χρονιά αυτή ξεκινά ο προγραμματισμός και συντονισμός για να επέλθει η αρχή της υλοποίησης το 1990. Η λήξη και ολοκλήρωση του προγράμματος ανακοινώθηκε το 2003. Η χρηματοδότηση προήλθε από ποικίλες πηγές μεταξύ των οποίων το Εθνικό Ινστιτούτο Υγείας (National Health Institute, NIH) και μια πληθώρα άλλων οργανισμών ανά τον κόσμο. Ένα παράλληλο εν ενεργεία εγχείρημα υπήρξε το Celera Corporation ή Clera Genomics, το οποίο ξεκίνησε επίσημα το

1998. Το μεγαλύτερο μέρος της αλληλούχισης που πραγματοποιήθηκε με κρατικά έξοδα των ΗΠΑ, έλαβε χώρα σε Ινστιτούτα των ΗΠΑ, του Ηνωμένου Βασιλείου, της Ιαπωνίας, της Γαλλίας, της Γερμανίας και της Κίνας ("Human Genome Project Completion: Frequently Asked Questions - National Human Genome Research Institute (NHGRI)," n.d.).

Το Πρόγραμμα του Ανθρώπινου Γονιδιώματος αρχικά στόχευε στην αλληλούχιση των γονιδίων που περιλαμβάνονται στο απλοϊδικό DNA (περισσότερα από 3 δισεκατομμύρια βάσεις). Το γονιδίωμα του κάθε οργανισμού είναι μοναδικό και η χαρτογράφηση του ανθρώπινου γονιδιώματος περιέλαβε την αλληλούχιση του DNA μικρού αριθμού ατόμων η οποία στη συνέχεια συρράφτηκε για να ληφθεί και να αναπαρασταθεί τελικά η ολική αλληλουχία του κάθε ανθρώπινου χρωμοσώματος. Ως εκ τούτου, το ανθρώπινο γονιδίωμα που δόθηκε σαν αποτέλεσμα της διαδικασία είναι ένα μωσαϊκό της αυτόνομης δουλειάς πολλών ατόμων και πολλών εργαστηρίων που εργάστηκαν ταυτόχρονα, για να φέρουν εις πέρας τη διαδικασία.

Το πρόγραμμα του ανθρώπινου γονιδιώματος παρήγαγε μία υψηλής ποιότητας εκδοχή της αλληλουχίας του ανθρώπινου γονιδιώματος και είναι ελεύθερα διαθέσιμο σε βάσεις δεδομένων. Σχεδιάστηκε προκειμένου να δημιουργηθεί μία πηγή γνώσης που θα μπορούσε να χρησιμοποιηθεί από το φάσμα τον βιοχημικών επιστημών.

Μέχρι πρόσφατα οι γιατροί μπορούσαν να λάβουν υπόψη τους τη μελέτη των γονιδίων και της γενετικής πληροφορίας σε περιπτώσεις εκ γενετής προβλημάτων και για ένα περιορισμένο αριθμό ασθενειών. Συνήθως, η διαδικασία αφορούσε ασθένειες ποιες έχουν εύκολα προβλέψιμο κληρονομικό μοτίβο γιατί κάθε μία από αυτές οφείλεται στην αλλαγή ενός μόνο γονιδίου ("EMBL-European Molecular Biology Laboratory," n.d.).

Πλέον όμως χάρη στην ποσότητα δεδομένων που υπάρχουν σχετικά με το ανθρώπινο DNA, η οποία προήλθε το πρόγραμμα του ανθρώπινου γονιδιώματος και άλλες γονιδιακές μελέτες οι επιστήμονες και οι γιατροί διαθέτουν περισσότερα εργαλεία προκειμένου να μελετήσουν το ρόλο που έχουν πολλαπλοί γενετικοί παράγοντες, όταν λειτουργήσουν παράλληλα με ειδικές συνθήκες στο περιβάλλον του οργανισμού, διαδικασίες που κάνουν το μοτίβο κάποιων ασθενειών ακόμα πιο πολύπλοκο. Αυτές οι πολυπαραγοντικές ασθένειες όπως ο καρκίνος, ο διαβήτης και τα καρδιαγγειακά προβλήματα αποτελούν την πλειονότητα των

προβλημάτων υγείας στις Ηνωμένες Πολιτείες και στην Ευρώπη. Η έρευνα που βασίζεται στη γονιδιωματική δίνει ήδη τη δυνατότητα στους ερευνητές να αναπτύξουν βελτιωμένες διαγνωστικές μεθόδους, πιο αποτελεσματικές θεραπευτικές στρατηγικές, προσεγγίσεις που να μπορούν να στηριχτούν σε αποδεδειγμένα στοιχεία και να οδηγήσουν σε βελτίωση της κλινικής εικόνας ή/και της ποιότητας ζωής των ασθενών, ενώ παράλληλα να βοηθήσουν και τους ίδιους τους γιατρούς και τους ασθενείς στη λήψη αποφάσεων. Δεδομένων των παραπάνω είναι λογικό η Ιατρική να κινείται προς τον δρόμο της εξατομικευμένης θεραπείας, δηλαδή της θεραπείας που θα βασίζεται στο ατομικό γενετικό προφίλ κάθε ασθενούς. Ο ρόλος της γενετικής στον τομέα της υγείας είναι όλο και πιο προφανής και η γενετική ιατρική είναι κάτι που ήδη έχει ξεκινήσει να συμβαίνει(“EMBL-European Molecular Biology Laboratory,” n.d.).

Ένα πρώτο αποτέλεσμα του ανθρώπινου DNA, ήρθε το 2001, από το «Πρόγραμμα του Ανθρώπινου Γονιδιώματος», παρόλο που το πρόγραμμα τελικά ολοκληρώθηκε το 2003, με την ολοκληρωμένη αλληλούχιση του ανθρώπινου γονιδιώματος να έχει λάβει πλέον χώρα, μετά την επί σειρά ετών συνεργασία επιστημόνων και ινστιτούτων σε ολόκληρο τον κόσμο. Έκτοτε, οι διαδικασίες αλληλούχισης έχουν διαφοροποιηθεί δραστικά, προς το καλύτερο, κάνοντας όλο και πιο εύκολη την γνώση της ακριβούς αλληλουχίας του γονιδιώματος όλο και περισσότερων οργανισμών. Οι προεκτάσεις των επιτευγμάτων αυτών αγγίζουν όλους τους τομείς των επιστημών υγείας και ζωής (“Human Genome Project Completion: Frequently Asked Questions - National Human Genome Research Institute (NHGRI),” n.d.; National Human Genome Research Institute, 2016).

1.2.1.4 Γονιδιωματική Ανάλυση

Η αλληλούχιση DNA δεν θα είχε τόσο μεγάλη αξία αν δεν ακολουθούσε επιπλέον ανάλυση. Ο σχολιασμός του γονιδιώματος που ακολουθεί την αλληλούχιση είναι η διαδικασία κατά την οποία προστίθεται βιολογική πληροφορία στις αλληλουχίες και αποτελείται από τρία βασικά βήματα. Τον προσδιορισμό περιοχών του γονιδιώματος που δεν κωδικοποιούν για πρωτεΐνες, τον προσδιορισμό πάνω στο γονιδίωμα, μια διαδικασία που ονομάζεται πρόβλεψη γονιδίων και η προσθήκη βιολογικής πληροφορίας σε αυτά τα στοιχεία.

Αυτή τη στιγμή υπάρχουν υπολογιστικά εργαλεία που κάνουν αυτόματη πρόβλεψη και ο σχολιασμός να γίνει πολύ γρήγορα σε σχέση με τις περιπτώσεις που η διαδικασία γίνεται χειροκίνητα και απαιτεί πολύ εξειδικευμένες γνώσεις από αυτόν που την εκτελεί. Ιδανικά όμως αυτές οι δύο προσεγγίσεις θα πρέπει να συνυπάρχουν γιατί τα υπολογιστικά εργαλεία έχουν ποσοστό σφαλμάτων.

Παραδοσιακά το βασικό επίπεδο του σχολιασμού γίνεται χρησιμοποιώντας τον αλγόριθμο BLAST προκειμένου να βρεθούν ομοιότητες και μετά να γίνει ο σχολιασμός των γονιδιωμάτων με βάση την ομολογία. Διαφορετικά κάποιες βάσεις χρησιμοποιούν πληροφορίες για το γονιδίωμα σκορ ομοιότητας πειραματικά δεδομένα και άλλες πηγές προκειμένου να Παρέχουν τον απαραίτητο σχολιασμό. Επιπλέον βάσεις όπως η ensemble βασίζονται σε επιλεγμένες yes δεδομένων καθώς και σε ένα μεγάλο εύρος λογισμικού στα πλαίσια μιας ροής αυτόματου σχολιασμού του γονιδιώματος.

Από την στιγμή που ένας οργανισμός έχει συλλεχθεί, τα προγράμματα του γονιδιώματος περιλαμβάνουν τα εξής βήματα: την αλληλούχιση του DNA του οργανισμού, την προσπάθεια αναδόμησης ενός ολοκληρωμένου μορίου DNA από τα τμήματα που προκύπτουν από την αλληλούχιση και τον σχολιασμό και ανάλυση αυτής της ολοκληρωμένης πληροφορίας.

Ιστορικά, η αλληλούχιση πραγματοποιούνταν σε ειδικές εγκαταστάσεις, ειδικευμένα κέντρα από άποψη ανθρώπινου δυναμικού και τεχνολογικών μέσων. Παρόλα αυτά σήμερα, υπάρχει στην αγορά μια νέα γενιά διατάξεων αλληλούχισης, που είναι αποτελεσματική και προσβάσιμη από μεγαλύτερη μερίδα των επιστημονικών εργαστηρίων και συνεχίζει να βελτιώνεται. Σε ότι αφορά την αλληλούχιση πλήρων γονιδιωμάτων (whole genome sequencing), δύο είναι οι βασικές προσεγγίσεις, η μέθοδος shotgun και η υψηλής απόδοσης (επόμενης γενιάς) αλληλούχιση (Next Generation Sequencing, NGS). Κατά την πρώτη, το γονιδίωμα κατακερματίζεται σε τμήματα μικρού μοριακού βάρους και γίνεται κλωνοποίηση των τμημάτων αυτών σε φορείς. Στη συνέχεια, προσδιορίζεται η αλληλουχία του ένθετου τμήματος του κάθε κλώνου της βιβλιοθήκης προκειμένου να βρεθούν κοινές περιοχές ανάμεσα στους κλώνους. Τελικά, η αλληλούχιση ολόκληρού του γονιδιώματος επιτυγχάνεται βάσει των αλληλουχιών του τεράστιου αριθμού των επικαλυπτόμενων κλώνων. Η παραπάνω μέθοδος απαιτεί μεγάλη υπολογιστική ισχύ και έτσι είναι πιο

αποτελεσματική για μικρά γονιδιώματα. Από την άλλη πλευρά η αλληλούχιση επόμενης γενιάς ήρθε σαν απάντηση στην ζήτηση για αλληλουχίσεις χαμηλότερου κόστους, που όμως δε θα στερούνταν σε αξιοπιστία αποτελεσμάτων. Αυτό επιτυγχάνεται μέσω της ικανότητας διαχείρισης μεγαλύτερου όγκου δεδομένων (Pettersson et al., 2009).

Λεπτομερέστερα, η shotgun sequencing αποτελεί μία μέθοδο αλληλούχισης που σχεδιάστηκε για αναλύσεις αλληλουχιών DNA μεγαλύτερες από 1.000 ζεύγη βάσεων μέχρι και ολόκληρα χρωμοσώματα. Ονομάστηκε shotgun ώστε να γίνει παραλληλισμός με τη γρήγορη πυροδότηση ενός πυροβόλου όπλου. Η διαδικασία μοιάζει με την κλασσική ηλεκτροφόρηση σε τζελ, όπου μεγάλα μόρια DNA πρέπει να σπάσουν σε μικρότερα τυχαία τμήματα τα οποία στη συνέχεια αλληλουχούνται προκειμένου να εξασφαλιστούν αναγνώσεις. Το DNA Στόχος παραλαμβάνετε μετά από πολλούς γύρους αυτής της τμηματοποίησης αλληλουχίσεων. Είναι απαραίτητο οι αναγνώσεις να είναι πολλαπλές και επικαλυπτόμενες ώστε να εξασφαλιστεί ότι δεν θα χαθεί γονιδιακή πληροφορία. Έπειτα ειδικό λογισμικό χρησιμοποιεί επικαλυπτόμενα άκρα διαφορετικών αναγνώσεων προκειμένου να τα συνθέσει σε μία συνεχή αλληλουχία. Η αλληλούχιση shotgun είναι μία διαδικασία τυχαίων δειγμάτων που απαιτεί να έχουμε πολλαπλά δείγματα για να εξασφαλιστεί ότι ένα δεδομένο νουκλεοτίδιο προσθέτετε οπωσδήποτε στην αλληλουχία που θα παρασκευαστεί (Weber & Myers, 1997).

Η δεύτερη και πιο πρόσφατη μέθοδος αλληλούχισης είναι η αλληλούχιση υψηλής απόδοσης (High-throughput Sequencing). Η μεγάλη ανάγκη χαμηλότερου κόστους αλληλούχισης οδήγησε στην ανάπτυξη τεχνολογιών υψηλής απόδοσης παραλληλίζουν την διαδικασία αλληλούχισης και παράγουν χιλιάδες ή εκατομμύρια αλληλουχίες στον ίδιο χρόνο. Ο στόχος είναι η μείωση του κόστους της αλληλούχισης DNA κάτι που δεν ήταν εφικτό με την χρήση των προηγούμενων μεθόδων. Στη μέθοδο αλληλούχισης illumina dye τα μόρια του DNA πρώτα προσκολλώνται σε ένα πλακίδιο και μετά πολυμερίζονται από την πολυμεράση, έτσι ώστε να σχηματιστούν τοπικές μονοκλωνικές αποικίες DNA. Για να προσδιοριστεί η αλληλουχία προστίθενται τέσσερις τύποι βάσεων λήξης του πολυμερισμού, οι οποίοι μπορούν να συνδεθούν με αντιστρεπτό τρόπο, ενώ τα νουκλεοτίδια που δεν χρησιμοποιήθηκαν απομακρύνονται. Οι αλυσίδες του DNA προεκτείνονται κατά ένα νουκλεοτίδιο κάθε φορά και η λήψη της εικόνας της αλληλούχισης μπορεί να γίνεται σε κάθε

χρονική στιγμή, επιτρέποντας έτσι μεγάλες αλληλουχίες από DNA αποικίες να φωτογραφίζονται από μία μόνο κάμερα. Η παύση της ενζυμικής αντίδρασης και η δυνατότητα λήψης εικόνας στο τέλος κάθε διακριτού κύκλου, δίνει την καλύτερη δυνατή απόδοση και θεωρητικά απεριόριστη ικανότητα αλληλούχισης. Η κάμερα τραβά μία φωτογραφία από τα νουκλεοτίδια που έχουν σημανθεί με φθορισμό και έπειτα το στοιχείο που παρεμποδίζει την περαιτέρω ανάπτυξη αφαιρείται με χημικό τρόπο επιτρέποντας το ξεκίνημα του επόμενου κύκλου (Reuter, Spacek, & Snyder, 2015).

1.2.2 Μεταγραφωμική (Transcriptomics)

Το μεταγράφωμα είναι το άθροισμα ή το πλήθος όλων των RNA του κυττάρου ή ενός πληθυσμού κυττάρων ή ενός ολόκληρου οργανισμού και συχνά συγχέεται με την έννοια του εξονιώματος. Βασική τους διαφορά αποτελεί πως το μεταγράφωμα περιλαμβάνει μόνο τα μόρια RNA που απαντώνται σε συγκεκριμένο κυτταρικό πληθυσμό και συνήθως περιλαμβάνει την ποσότητα ή τη συγκέντρωση του κάθε μορίου παράλληλα με την πληροφορία της μοριακής ταυτότητας. Ο όρος μπορεί να αναφέρεται ακόμα και σε όλο το πλήθος των μεταγράφων ενός οργανισμού ή σε συγκεκριμένες υποομάδες μεταγράφων παρούσες σε συγκεκριμένους κυτταρικούς τύπους. Εν αντιθέσει με το γονιδίωμα που είναι λίγο έως πολύ καθορισμένο ανάλογα με την κυτταρική σειρά (υφίστανται τροποποιήσεις στα πλαίσια χρωμοσωμικών ανωμαλιών), το μεταγράφωμα αποτελεί δυναμική δομή, μεταβαλλόμενη ανάλογα με τις εξωγενείς περιβαλλοντικές συνθήκες στις οποίες το κύτταρο ή και ο οργανισμός καλείται να προσαρμοστεί και να επιβιώσει.

Παράλληλα από τη στιγμή που αναπαριστά όλο το πλήθος των RNA του κυττάρου, αντικατοπτρίζει τα γονίδια που εκφράζονται ενεργά σε κάθε χρονική στιγμή με την εξαίρεση να αποτελούν μόνο οι περιπτώσεις φαινομένων αποδόμησης mRNA όπως η μεταγραφική εξασθένηση.

Η μελέτη του μεταγραφώματος, η Μεταγραφωμική, εξετάζει τα επίπεδα έκφρασης του RNA σε δεδομένο πληθυσμό κυττάρων, συνήθως εστιάζεται στο mRNA, όμως κάποιες φορές περιλαμβάνονται μόρια όπως tRNA και sRNA.

Οι εργαστηριακές τεχνικές που σχετίζονται με τη μελέτη του μεταγραφώματος είναι οι τεχνολογίες Μικροσυστοιχιών (Microarrays) και η Αλληλούχιση Επόμενης Γενιάς (Next

Generation Sequencing, NGS), γενικά αναφερόμενες με τον όρο RNA-Sequencing ή συντομογραφικά RNA-Seq. Πιο ειδική κατηγορία αποτελεί η μελέτη των επιπέδων έκφρασης σε ένα μόνο κύτταρο, η οποία ονομάζεται single-cell transcriptomics, Μεταγραφωμική ενός κυττάρου.

Το μεταγράφωμα μπορεί να θεωρηθεί ως υποσύνολο του πρωτεώματος, δεδομένου ότι το δεύτερο αποτελεί το σύνολο όλων των πρωτεϊνικών μορίων που προέρχονται ή έχουν εκφραστεί από το ίδιο γονιδίωμα. Μολαταύτα, η ανάλυση της mRNA έκφρασης μπορεί να αποδειχθεί επίπονη λόγω του γεγονότος πως οι μικρές αλλαγές στο επίπεδο του RNA μεγιστοποιούνται ανά περιπτώσεις στο πρωτεϊνικό επίπεδο. Ως εκ τούτου γίνεται προφανές πως σχετικές συγκεντρώσεις των mRNA μορίων στο κύτταρο που εξάγονται από την ανάλυση των μικροσυστοιχιών δεν συνάδουν απόλυτα, ούτε αντικατοπτρίζουν την ποσότητα και τον τύπο των πρωτεϊνικών μορίων σε ένα κύτταρο. Ο αριθμός των πρωτεϊνικών μορίων που συντίθενται δεδομένου mRNA μορίου εξαρτάται σε μεγάλο βαθμό από τους μηχανισμούς μεταφραστικής ρύθμισης και ιδιαίτερα τους μηχανισμούς έναρξης της μεταγραφής, παραδείγματος χάρη, την αποτελεσματική πρόσδεση των ριβοσωμικών υπομονάδων στο προς μετάφραση mRNA και τους μηχανισμούς που την επάγουν. Το σύνολο των εκφρασμένων πρωτεϊνών του κυττάρου ονομάζεται πρωτέωμα.

Οι τεχνολογίες που μελετούν το μεταγράφωμα αναφέρονται ως μεταγραφωμικές τεχνολογίες και αφορούν τη μελέτη και ανάλυση του συνόλου των mRNA μορίων του κυττάρου ή του οργανισμού, δηλαδή η πληροφορία που είναι αποθηκευμένη στο DNA και εκφράζεται μέσω της διαδικασίας της μεταγραφής. Αναφερόμαστε αποκλειστικά στα RNA μόρια που λειτουργούν ως ενδιάμεσο στάδιο μεταξύ DNA και πρωτεΐνης και όχι στα non-coding RNA του κυττάρου που είναι επιφορτισμένα με άλλους ρόλους. Δεδομένου ότι, ανάλογα με τις κυτταρικές ανάγκες, τις ανάγκες του οργανισμού και τις περιβαλλοντικές συνθήκες, το μεταγράφωμα μπορεί να ποικίλει. Η Μεταγραφωμική μελέτη αφορά τα μόρια mRNA του κυττάρου σε δεδομένη χρονική στιγμή. Οι τεχνολογίες μελέτης του μεταγραφώματος, συνεπώς παρέχουν πληροφορίες σχετικά με τις ενεργές και ανενεργές κυτταρικές διεργασίες. Οι πρώτες προσπάθειες μελέτης ολόκληρου μεταγραφώματος οργανισμού έγιναν στις αρχές της δεκαετίας του ‘90. Από τότε και στο εξής, οι τεχνικές και ο τρόπος μελέτης αναβαθμίζονται συνεχώς και έχουν κάνει την Μεταγραφωμική πτυχή της

Βιολογίας. Οι δύο βασικές σύγχρονες τεχνικές που χρησιμοποιούνται και αφορούν την εξόρυξη πληροφορίας δεδομένης της RNA αλληλουχίας είναι η τεχνολογία μικροσυστοιχιών, η αλληλούχιση υψηλής απόδοσης και η αλληλούχιση επόμενης γενιάς με στόχο την καταγραφή του συνόλου των μεταγράφων. Με τη βελτίωση των παραπάνω τεχνικών ο όγκος των δεδομένων αυξήθηκε ραγδαία και οι μέθοδοι υπολογιστικής ανάλυσης σταθερά προσαρμόστηκαν ώστε να αναλύουν πιο αποτελεσματικά και με ακρίβεια μεγάλους όγκους δεδομένων. Παράλληλα, οι βάσεις μεταγραφικών δεδομένων έχουν αυξηθεί σε αριθμό και παρουσιάζουν διαφορετικές λειτουργικότητες καθώς η Μεταγραφωμική πληροφορία συλλέγεται από τους ερευνητές και γίνεται δημόσια προσβάσιμη για περαιτέρω ανάλυση.

Η μέτρηση της γονιδιακής έκφρασης ενός οργανισμού σε διαφορετικούς ιστούς ή συνθήκες ή σε διαφορετικό χρόνο, παρέχει πληροφορία για τη γονιδιακή ρύθμιση στην εκάστοτε κατάσταση και αποκαλύπτει λεπτομέρειες για τη φυσιολογία και την κυτταρική λειτουργία του οργανισμού. Μπορεί επίσης να χρησιμοποιηθεί για αποκαλύψει τη λειτουργία γονιδίων μέχρι πρόσφατα, χωρίς λειτουργικό σχολιασμό. Παρόμοια, η ανάλυση του μεταγραφώματος έχει κάνει δυνατή τη μελέτης της διαφορικής γονιδιακής ρύθμισης μεταξύ οργανισμών που πάσχουν και αντίστοιχων υγιών ομάδων, παρέχοντας έτσι καλύτερη κατανόηση της ασθένειας και δυνατότητα προσανατολισμού προς συγκεκριμένες πρακτικές πρόληψης και θεραπείας.

Οι μεταγραφωμικές μέθοδοι απαιτούν την απομόνωση του RNA από τον οργανισμό. Οι τεχνικές απομόνωσης και εξαγωγής RNA είναι παρόμοιες ανεξαρτήτως του είδους του οργανισμού και περιλαμβάνουν λύση των κυτταρικών ιστών, καταστολή της δράσης της RNάσης με χρήση χαοτροπικών αλάτων, λύση των μακρομορίων και των νουκλεοτιδικών συμπλοκών, διαχωρισμό των RNA μορίων από τα υπόλοιπα βιομόρια του κυττάρου και τελικά συγκέντρωση του RNA μέσω έκλουσης από μια αρχική μήτρα στην οποία έχει δεσμευθεί. Σε ορισμένες περιπτώσεις πραγματοποιείται και επώαση με DNάση.

1.2.3 Λοιπές υποκατηγορίες των –omics

Εκτός από τις βασικές υποκατηγορίες των –omics που αναφέρθηκαν στην ενότητα 1.2 «Η επανάσταση των -omics» και τις δύο κατηγορίες Genomics και Transcriptomics που σχετίζονται με το πεδίο μελέτης της μεταπτυχιακής διπλωματικής εργασίας, υπάρχουν και

επιπλέον κατηγορίες. Μάλιστα, υπάρχει η τάση να αναφέρονται με την κατάληξη –omics τομείς που είναι είτε εξαιρετικά ειδικοί και θα μπορούσαν να εμπίπτουν σε κάποια από τις ευρύτερες έννοιες, είτε τομείς που δεν σχετίζονται άμεσα και αποκλειστικά με τον χώρο της Βιολογίας, όμως απαιτούν βιολογική γνώση και βιολογικές εφαρμογές. Οι ευρύτερες και γνωστότερες κατηγορίες που δεν έχουν αναφερθεί ήδη περιλαμβάνουν τα εξής:

Πρωτεωμική ή Proteomics. Το πρωτέωμα αποτελεί το σύνολο των πρωτεϊνών ενός κυττάρου, καθώς και όλες τις τροποποιήσεις που μπορεί να συμβούν στα μόρια που παράγονται από έναν οργανισμό ή από κάποιο σύστημα. Η Πρωτεωμική, οπότε, είναι η μελέτη των πρωτεϊνών σε μεγάλη κλίμακα, ειδικά στο επίπεδο της δομής και της λειτουργίας. Πιο ειδικές κατηγορίες της Πρωτεωμικής, αποτελούν οι Ανοσο-πρωτεωμική (Immunoproteomics), μελέτη των πρωτεϊνών που εμπλέκονται σε λειτουργίες του ανοσοποιητικού συστήματος, η Διατροφο-πρωτεωμική (Nutriproteomics), ταυτοποίηση των μοριακών στόχων θρεπτικών ή μη χημικών συστατικών, η Πρωτεο-Γενωμική (Proteogenomics), ένας αναδυόμενος τομέας της Βιολογίας που συνδυάζει τη γνώση του πρωτεώματος με τον γενετικό σχολιασμό και η Δομική Πρωτεωμική/Γενωμική (Structural Genomics/Proteomics), η μελέτη της τριτοταγούς δομής ενός πρωτεϊνικού μορίου χρησιμοποιώντας παράλληλα υπολογιστική μοντελοποίηση και πειραματική προσέγγιση (Pandey & Mann, 2000).

Επιγενωμική ή Επιγενετική ή Epigenomics. Το επιγονιδίωμα είναι η υποστηρικτική δομή του γονιδιώματος και περιλαμβάνει πρωτεϊνικά μόρια και RNA προσδέτες, εναλλακτικές δομές DNA και χημικές τροποποιήσεις της DNA αλληλουχίας. Η Επιγενετική περιλαμβάνει σύγχρονες τεχνολογίες οι οποίες εντοπίζουν και αξιοποιούν την πληροφορία για γενετικές τροποποιήσεις που συμβαίνουν κατά τη διάρκεια της ζωής του ατόμου, παρόλα αυτά είναι κληρονομήσιμες (Bird, 2007).

Μεταβολική ή Metabolomics. Αποτελεί τη μελέτη χημικών διεργασιών που εμπλέκουν μεταβολίτες. Όλες οι μελέτες αυτού του χώρου γίνονται σε επίπεδο συστήματος καθώς οι μεταβολικές διαδικασίες είναι πολυπαραγοντικές και περιλαμβάνουν μικρά χημικά μόρια προερχόμενα από διεργασίες σε διάφορα μέρη του οργανισμού (Fiehn, 2002).

Lipidomics, Glycomics, Foodomics, Nutrigenomics, Pharmacogenomics, Toxigenomics είναι μερικά ακόμη από τα ονόματα των –omics που δεν αναλύονται λεπτομερέστερα σε αυτό το

πλαίσιο. Κλείνοντας, ιδιαίτερο ενδιαφέρον παρουσιάζει ο όρος “Polyomics”, ο οποίος αναφέρεται στη συστηματική μελέτη πολλών από τις παραπάνω κατηγορίες ταυτόχρονα, δεδομένου ενός κοινού υπόβαθρου, γενετικού, μοριακού ή ασθένειας, το οποίο εξετάζεται από όλες τις δυνατές πλευρές.

1.2.4 Η Βιοπληροφορική στην υπηρεσία των -omics

Η Βιοπληροφορική είναι ένα υβριδικό πεδίο που συνδυάζει τη γνώση της βιολογίας και της επιστήμης της πληροφορικής και αποτελεί ένα υπο-πεδίο της επιστήμης των υπολογιστών. Τα γονιδιώματα των οργανισμών είναι πολύ μεγάλα. Το ανθρώπινο γονιδίωμα εκτιμάται ότι έχει τρία δισεκατομμύρια ζεύγη βάσεων που περιέχουν περίπου 25.000 γονίδια. Επιπλέον, η Μεταγραφωμική (Transcriptomics) μελετά το ποια γονίδια, μεταξύ των δεκάδων χιλιάδων σε έναν οργανισμό, έχουν ενεργοποιηθεί ή απενεργοποιηθεί σε μια δεδομένη στιγμή, σε πολλαπλές χρονικές στιγμές αλλά και σε πολλαπλές πειραματικές συνθήκες σε κάθε χρονικό σημείο. Με άλλα λόγια, τα δεδομένα «-ωμικής» περιέχουν τεράστιες ποσότητες πληροφοριών που ο ανθρώπινος νους δεν μπορεί να συλλάβει χωρίς τη βοήθεια των υπολογιστικών μεθόδων της Βιοπληροφορικής. Η Βιοπληροφορική είναι σημαντική για τη γενετική έρευνα, επειδή τα γενετικά δεδομένα έχει ένα πλαίσιο, αυτό της βιολογίας. Όλες οι μορφές ζωής έχουν ορισμένους κανόνες συμπεριφοράς. Το ίδιο ισχύει και για τους ιστούς και τα κύτταρα, τα γονίδια και τις πρωτεΐνες. Αλληλεπιδρούν με συγκεκριμένους τρόπους και ρυθμίζουν το ένα το άλλο με συγκεκριμένους τρόπους. Η μεγάλη κλίμακα, σύνθετων δεδομένων που δημιουργούνται στον τομέα της γονιδιωματικής, δεν θα είχε νόημα χωρίς τη γνώση του πώς λειτουργούν οι μορφές ζωής. Τα δεδομένα που προκύπτουν από τη γονιδιωματική μπορούν να αναλυθούν με τις ίδιες μεθόδους που χρησιμοποιούνται από τους μηχανικούς και τους φυσικούς που μελετούν τις αγορές των χρηματοοικονομικών και των οπτικών ινών, αλλά αναλύοντας δεδομένα με τρόπο που να απαιτεί γνώση της βιολογίας. Έτσι, η Βιοπληροφορική έγινε ένα πολύτιμο υβριδικό πεδίο της γνώσης. Η Βιοπληροφορική είναι σε θέση να υπολογίσει δεκάδες χιλιάδες αριθμών σε λίγα λεπτά, ανάλογα με το πόσο γρήγορα ο υπολογιστής μπορεί να επεξεργαστεί τις πληροφορίες. Ο τομέας των «–omics» χρησιμοποιεί υπολογιστές για να τρέξουν οι αλγόριθμοι σε μεγάλη κλίμακα ώστε να βρει μοτίβα σε μεγάλα σύνολα δεδομένων. Κοινοί αλγόριθμοι περιλαμβάνουν λειτουργίες όπως η ιεραρχική ομαδοποίηση και ανάλυση σε κύριες

συνιστώσες. Και οι δύο είναι οι τεχνικές για να βρεθούν οι σχέσεις μεταξύ των δειγμάτων που έχουν πολλούς παράγοντες σε αυτές .

Η Βιοπληροφορική κατάφερε να μελετήσει τον τρόπο με τον οποίο ένα σύστημα που έχει χιλιάδες κινούμενα μέρη καταφέρνει να προσαρμοστεί στο επίπεδο που όλα τα μέρη κινούνται ταυτόχρονα. Παλιότερα, οι γενετιστές μελετούσαν μόνο ένα γονίδιο τη φορά. Αν και η προσέγγιση αυτή εξακολουθεί να έχει ιδιαίτερη αξία και θα συνεχίσει να έχει, η Βιοπληροφορική άνοιξε το δρόμο για νέες ανακαλύψεις και μέσω της θεωρίας συστημάτων, η βιολογία συστημάτων ερμηνεύει πλέον τα διάφορα βιολογικά συμβάντα ως σύνολο αλληλεπιδράσεων σε δυναμική ισορροπία.

1.3 Αλληλούχιση RNA και επεξεργασία πρωτογενών δεδομένων

Η αλληλούχιση επόμενης γενιάς (Next Generation Sequencing, NGS) πρόκυψε σε μεγάλο βαθμό εξαιτίας της αυξανόμενης ανάγκης για τεχνολογίες αλληλούχισης χαμηλού κόστους χωρίς να στερούνται σε ποιότητα δεδομένων. Πρόκειται για τη βιολογική τεχνολογία της “παραλληλοποίησης” της διαδικασίας αλληλούχισης, με τρόπο που να καθιστά δυνατό τον προσδιορισμό της αλληλουχίας των βάσεων των μακρομορίων RNA και DNA, εκατομμυρίων ή δεκάδων εκατομμυρίων αλληλουχιών ταυτόχρονα. Η επεκτασιμότητα, η ταχύτητα, αλλά κυρίως η σχέση κόστους-απόδοσης των NGS εφαρμογών επιτρέπουν στους ερευνητές να μελετήσουν τα βιολογικά συστήματα σε επίπεδο που δεν ήταν δυνατό μέχρι πρότινος. Η τεχνολογία αυτή περιγράφηκε εκτενέστερα στην ενότητα 1.2.1.4.

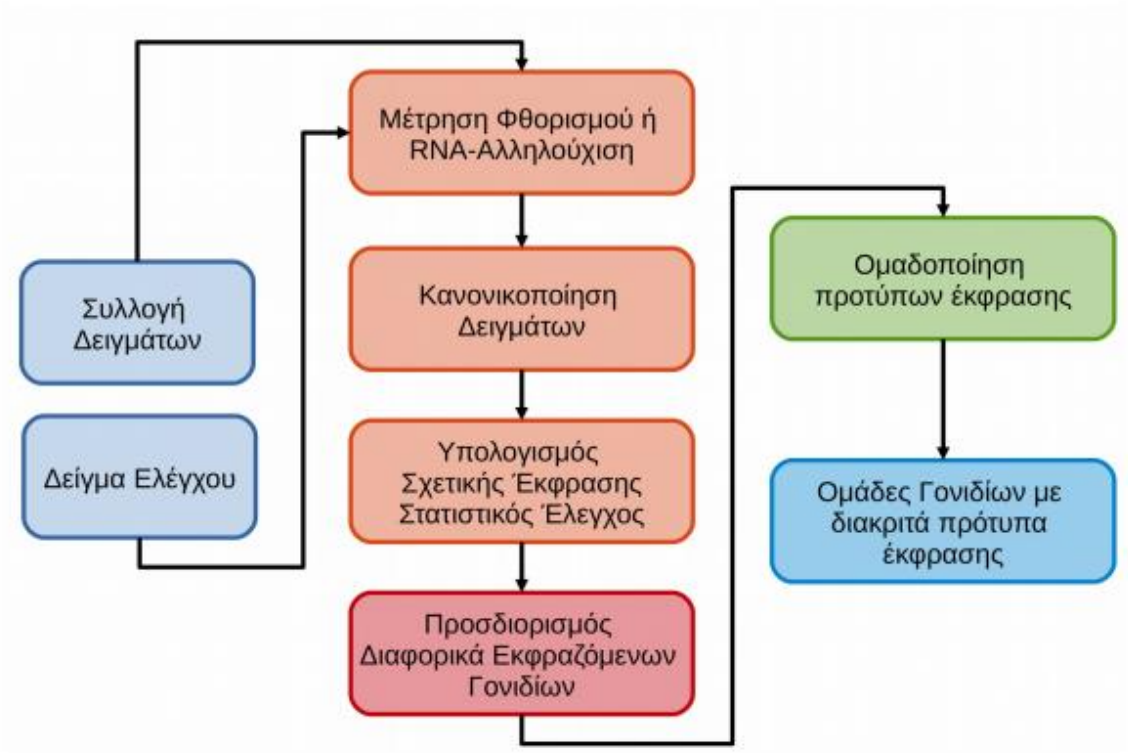
Η NGS ποσοτικοποίηση της γονιδιακής έκφρασης γίνεται με τη μαζική αλληλούχιση mRNA που αρχικά απομονώνεται από το δείγμα και στη συνέχεια μετατρέπεται σε cDNA όπως στην περίπτωση των μικροσυστοιχιών. Στη συνέχεια η διαδικασία διαφοροποιείται. Αρχικά το cDNA δείγμα υπόκειται σε ένα στάδιο κλασμάτωσης πριν την αλληλούχιση καθώς οι υπάρχουσες τεχνολογίες, στη συντριπτική τους πλειοψηφία, αποδίδουν αξιόπιστα αποτελέσματα για αλληλουχίες όχι μεγαλύτερες από 300-500 βάσεις. Στη συνέχεια και ανάλογα με την τεχνολογία που εφαρμόζεται, το cDNA ενισχύεται μέσω αλυσιδωτής αντίδρασης πολυμεράσης (PCR) (η οποία μπορεί να διενεργηθεί με διαφορετικούς τρόπους) και αλληλουχείται μαζικά συνήθως “μέσω σύνθεσης”. Αυτό σημαίνει ότι νέοι κλώνοι cDNA συντίθενται πάνω στο εκμαγείο των κλώνων του δείγματος και η διαδικασία της σύνθεσης

καταγράφεται νουκλεοτίδιο-νουκλεοτίδιο ((Wang, Gerstein, & Snyder, 2009). Το αποτέλεσμα είναι ένα αρχείο που περιέχει έναν πολύ μεγάλο αριθμό (της τάξης των δεκάδων εκατομμυρίων) αλληλουχιών μικρού μήκους (μεταξύ 100 και 500 βάσεων). Μια ακόμα βασική διαφορά με τις μεθοδολογίες που βασίζονται στην υβριδοποίηση έχει να κάνει με την ποσοτικοποίηση των αποτελεσμάτων. Αυτή γίνεται για ένα πείραμα αλληλούχισης επόμενης γενιάς μέσω των εξής βημάτων (Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008).

- Έλεγχος ποιότητας των αλληλουχιών και αποκλεισμός αυτών που δεν ικανοποιούν συγκεκριμένα κριτήρια αξιοπιστίας. Στο πρώτο αυτό στάδιο απορρίπτονται οι αλληλουχίες που δεν πληρούν τις προϋποθέσεις ποιότητας που είναι γενικώς αποδεκτές. Ποσοστά απόρριψης μεταξύ 5 και 30% είναι φυσιολογικά, ανάλογα με το πείραμα.
- Χαρτογράφηση των αλληλουχιών στο γονιδίωμα αναφοράς. Πρόκειται για τη βασική διαδικασία μέσω της οποίας ποσοτικοποιούνται τα αποτελέσματα. Για καθεμία από τα εκατομμύρια των μικρών αλληλουχιών (που ονομάζονται και “αναγνώσεις αλληλουχιών”, sequence reads ή απλά reads) εντοπίζεται η θέση του γονιδιώματος από την οποία προέρχεται (το οποίο ονομάζουμε γονιδίωμα αναφοράς, reference genome). Μετά τη χαρτογράφηση του συνόλου των reads, μπορούμε να γνωρίζουμε με ακρίβεια πόσες φορές “διαβάστηκε” κάθε νουκλεοτίδιο του υπό μελέτη γονιδιώματος.
- Ποσοτικοποίηση αριθμού αναγνώσεων ανά μετάγραφο. Παίρνοντας ως δεδομένες τις θέσεις των γονιδίων/μεταγράφων στο γονιδίωμα αναφοράς μπορούμε να υπολογίσουμε τον αριθμό των αναγνώσεων που επικαλύπτονται με κάθε γενετικό τόπο ξεχωριστά. Λαμβάνοντας υπόψη το μήκος της αντίστοιχης γονιδιωματικής περιοχής, (ή για την ακρίβεια των μεταγράφων mRNA που προκύπτουν από αυτήν), αλλά και το συνολικό αριθμό των αναγνώσεων που προέκυψαν από το πείραμα μπορούμε να καταλήξουμε σε μια αριθμητική τιμή που είναι έτσι δηλωτική της ποσότητας mRNA που υπήρχε στο αρχικό δείγμα από το συγκεκριμένο γενετικό τόπο.

Ανεξάρτητα από τη μέθοδο που χρησιμοποιούμε για την μελέτη της γονιδιακής έκφρασης, τα στάδια της ανάλυσης είναι λίγο πολύ τα ίδια και περιλαμβάνουν: α) την

απομόνωση του mRNA β) την ποσοτικοποίηση του γ) τον προσδιορισμό των διαφορικά εκφραζόμενων γονιδίων και δ) την ομαδοποίηση γονιδίων ανάλογα με τα πρότυπα έκφρασής τους. Παρακάτω απεικονίζεται γραφικά αυτή η πληροφορία.



Εικόνα 4 Σχηματική αναπαράσταση των σταδίων ενός πειράματος γονιδιακής έκφρασης από την αποκομιδή των πρωτογενών δεδομένων ως τη δημιουργία ομάδων γονιδίων με χαρακτηριστικά πρότυπα έκφρασης (τροποποιημένη από ("Kallipos Repository: Home," n.d.)).

1.3.1 Αποκομιδή και χειρισμός των πρωτογενών δεδομένων

Τα πρωτογενή δεδομένα εξαρτώνται από τη μεθοδολογία που χρησιμοποιήθηκε για την ανάλυση. Στην περίπτωση των μικροσυστοιχιών το μετρούμενο μέγεθος είναι η ένταση φθορισμού που προκύπτει από την υβριδοποίηση συμπληρωματικών στους ανιχνευτές αλληλουχιών. Στην περίπτωση του RNA-Seq οι μετρήσεις αφορούν καθαρά το πλήθος των συντιθέμενων αλληλουχιών που προέρχονται από ένα συγκεκριμένο mRNA μόριο. Και στις δύο περιπτώσεις είναι απαραίτητη μια σειρά χειρισμών των δεδομένων έτσι όπως λαμβάνονται από την πειραματική διάταξη και είναι βασικό να προηγηθεί μια διαδικασία κανονικοποίησης των πρωτογενών δεδομένων πριν περάσουμε στην περαιτέρω ανάλυσή τους. Η κανονικοποίηση είναι η διαδικασία με την οποία μετατρέπουμε δεδομένα που έχουν

προκύψει με διαφορετικούς τρόπους σε μια κλίμακα που να τα καθιστά άμεσα συγκρίσιμα. Μέσω της κανονικοποίησης, αφαιρούμε συστηματικά σφάλματα που μπορεί να προέρχονται από τους χειρισμούς των πειραματιστών, την πειραματική διάταξη ή άλλους αστάθμητους παράγοντες. Μ' αυτόν τον τρόπο δεδομένα από διαφορετικά πειράματα μπορούν να συγκριθούν στη βάση των διαφορών τους που αφορούν το βιολογικό υπόβαθρο, ελαχιστοποιώντας την επίδραση τεχνικών σφαλμάτων και εξωγενών παραγόντων.

1.3.2 RNA Sequencing και αριθμός αναγνώσεων ανά 1000 βάσεις (FRKM)

Στα πειράματα RNA-Seq η ποσοτικοποίηση της έκφρασης γίνεται στη βάση του αριθμού των reads/αναγνώσεων που βρίσκονται να επικαλύπτονται με μια συγκεκριμένη γονιδιωματική περιοχή, που αντιστοιχεί σ' ένα δεδομένο μετάγραφο (transcript) ή γονίδιο. Ωστόσο, η απόλυτη τιμή αυτής της ποσότητας δεν μπορεί να χρησιμοποιηθεί αυτούσια για δύο βασικούς λόγους. Ο πρώτος είναι ότι εξαρτάται από το μήκος του μεταγράφου. Μεγάλα γονίδια που εκτείνονται για πολλές χιλιάδες βάσεις (κάποιες φορές και εκατοντάδες χιλιάδες) θα συγκεντρώνουν μεγαλύτερο αριθμό αναγνώσεων απλώς και μόνο λόγω μεγέθους. Ο δεύτερος λόγος είναι ότι πειράματα που συνολικά παράγουν μεγαλύτερο αριθμό αναγνώσεων, επειδή το δείγμα εμπλουτίστηκε περισσότερο, επειδή η ποιότητα της αλληλούχισης ήταν καλύτερη και οδήγησε στην απόρριψη μικρότερου αριθμού αναγνώσεων ή απλώς επειδή το δείγμα αλληλουχίθηκε σε μεγαλύτερο “βάθος”, θα δίνουν συστηματικά μεγαλύτερες τιμές αναγνώσεων ανά μετάγραφο. Για τους δύο αυτούς λόγους, για την εκτίμηση του βαθμού έκφρασης μεταγράφων υπολογίζουμε μια διορθωμένη τιμή που ονομάζεται RPKM ή FPKM, από τα αρχικά “Reads/Fragments per Kilo base of gene per Million”. Η τιμή αυτή αντιστοιχεί σε μια διπλή κανονικοποίηση του αριθμού των αναγνώσεων ανά μετάγραφο, αρχικά ως προς το μήκος του (ανά χιλιάδα βάσεων) κι έπειτα ως προς το σύνολο των παραχθεισών αλληλουχιών (ανά εκατομμύριο αλληλουχιών) (Bolstad, Irizarry, Astrand, & Speed, 2003) . Αν λοιπόν το μετάγραφο t με μήκος l επικαλύπτεται με r αναγνώσεις σε σύνολο N χαρτογραφημένων αλληλουχιών, τότε η τιμή FPKM θα είναι ίση με:

$$FPKM_{(t)} = \frac{r}{\frac{l}{10^3}\frac{N}{10^6}} = \frac{r10^9}{lN}$$

Η παραπάνω μετατροπή εξασφαλίζει ότι η τιμή στην οποία αναφέρεται ο βαθμός έκφρασης ενός μεταγράφου θα είναι ανεξάρτητη τόσο του μήκους του όσο και του βάθους της αλληλούχισης.

1.3.3 Κανονικοποίηση ποσοστιμορίων (quantile normalization)

Μια μη-παραμετρική μέθοδος κανονικοποίησης που βρίσκει ευρεία εφαρμογή στην ανάλυση δεδομένων γονιδιακής έκφρασης, είναι η κανονικοποίηση ποσοστημορίων (quantile normalization) (Hansen, Irizarry, & WU, 2012). Η συγκεκριμένη μέθοδος οδηγεί σε συγκρίσιμες κατανομές τιμών ακόμα και στην περίπτωση που οι αρχικές κατανομές δεν είναι κανονικές. Η διαδικασία βασίζεται ουσιαστικά στη σύγκριση της κατάταξης των τιμών και για το λόγο αυτό είναι ανεξάρτητη των ροπών (μέση τιμή, διασπορά κλπ). Δεδομένου ενός πίνακα N τιμών από M διαφορετικά πειράματα/δείγματα, υπολογίζεται αρχικά η κατάταξη (κατά αύξουσα σειρά) των τιμών εντός του κάθε δείγματος σε έναν νέο πίνακα R[N,M]. Στη συνέχεια δημιουργείται ένα μοναδικό διάνυσμα μέσων τιμών που λαμβάνουν υπόψη τους την κατάταξη. Έτσι η τιμή Q[1] είναι η μέση τιμή των πρώτων στην κατάταξη (μικρότερων) τιμών των M δειγμάτων, η τιμή Q[2] είναι η μέση τιμή των δεύτερων στην κατάταξη τιμών κοκ. Οι τιμές Q χρησιμοποιούνται στη συνέχεια, στη θέση των τιμών με την αντίστοιχη κατάταξη στον πίνακα R. Έτσι η χαμηλότερη τιμή για το πρώτο δείγμα εξισώνεται με την Q[1], το ίδιο και η χαμηλότερη τιμή για το δεύτερο, το τρίτο κοκ. Το τελικό αποτέλεσμα είναι ένας πίνακας που περιέχει τις ίδιες ακριβώς τιμές σε κάθε στήλη με διαφορετική ωστόσο κατάταξη. Αυτός ο μετασχηματισμός, εξασφαλίζει ότι οι κατανομές είναι απολύτως συγκρίσιμες χωρίς όμως να χάνεται η εσωτερική τους δομή. Είναι προφανές ότι για την εφαρμογή της μεθόδου απαιτείται ένας πίνακας με τον ίδιο αριθμό μετρήσεων N για όλα τα πειράματα (Hansen et al., 2012).

1.4 Διαφορική έκφραση γονιδίων

Κάθε σωματικό κύτταρο περιέχει το ίδιο αντίγραφο του γονιδιώματος του οργανισμού. Αναλογιζόμενοι ότι υπάρχουν κύτταρα που συντελούν διαφορετικές λειτουργίες και παράγουν διαφορετικών ειδών προϊόντα, είναι προφανές πως δεν αξιοποιείται ολόκληρο το γονιδίωμα, δηλαδή όλη η διαθέσιμη γενετική πληροφορία, γονίδια, ρυθμιστικές περιοχές και επαναλαμβανόμενο DNA, σε κάθε είδος κυττάρου. Από όλο το γενετικό υλικό μόνο ένα

μικρό ποσοστό είναι υπεύθυνο για να σύνθεση RNA, ικανό να πραγματοποιήσει συγκεκριμένες «ειδικές» διεργασίες που άπτονται στον τύπο του κυττάρου. Ακόμα, τα γονίδια, τα οποία δεν μεταφράζονται προς δημιουργία RNA ή και πρωτεϊνών, δεν καταστρέφονται αλλά διατηρούν την ικανότητα έκφρασης που διαθέτουν. Εξάγεται έτσι το συμπέρασμα πως η γονιδιακή έκφραση υπόκειται σε κάποιας μορφής ρύθμιση ανάλογη των συνθηκών αλλά και των αναγκών που παρουσιάζονται στο κύτταρο, ή γενικεύοντας και πηγαίνοντας σε ανώτερες βαθμίδες οργάνωσης, σε ομάδες κυττάρων με όμοια δομή και λειτουργία, ιστούς, όργανα συστήματα οργάνων και ολόκληρους οργανισμούς. Το πολύπλοκο σύστημα σηματοδότησης, μεταφοράς μηνυμάτων από και προς το εσωτερικό των κυττάρων, τα οποία μηνύματα αρχικά έχουν δημιουργηθεί σαν ερέθισμα σε μια απομακρυσμένη ομάδα κυτταρικών υποδοχέων, είναι που επιτρέπει την έκφραση συγκεκριμένης πληροφορίας με την μέγιστή δυνατή ειδικότητα. Ο μηχανισμός της διαφορικής έκφρασης των γονιδίων προνοεί ώστε, για παράδειγμα, το γονίδιο της αιμοσφαιρίνης που υπάρχει στον πυρήνα των νευρικών κυττάρων να μην εκφράζεται στους νευρώνες.

Το μεταγράφωμα ενός οργανισμού αποτελεί το σύνολο του RNA που παράγεται από τα γονίδια. Παρά το γεγονός ότι και περεταίρω ρυθμίσεις συμβαίνουν στην διαδικασία της μεταγραφής – μετάφρασης, καθώς αυτή η πρώιμη μορφή του μεταγραφικού προϊόντος δεν αποτελεί το τελικό προϊόν, η μελέτη της διαφορικής γονιδιακής έκφρασης συντελεί στην αναγνώριση γονιδίων που παίζουν σημαντικό ρόλο στην έκφραση του φαινοτύπου. Χαρακτηριστικό παράδειγμα αποτελεί η μελέτη μεταξύ υγειών και ασθενών ιστών για τον καθορισμό των γενετικών μεταβολών που υπάρχουν σε κάθε περίπτωση (Finotello & Di Camillo, 2015). Άλλες περιπτώσεις είναι η κατηγοριοποίηση ασθενών με βάση δημογραφικά ή γενετικά χαρακτηριστικά και η μελέτη της διαφορικής έκφρασης γονιδίων σε αυτό το επίπεδο ή και η ανάλυση δειγμάτων από πάσχων και παρακείμενο ιστό.

Στη μελέτη του καρκίνου η διαφορική γονιδιακή έκφραση παρέχει πληροφορίες για τον εντοπισμό γονιδίων που ευθύνονται για την ογκογένεση. Δύο διαφορετικά δείγματα αναμένεται να έχουν διαφορετικό τύπο μεταγραφώματος καθώς επιτελούν ξεχωριστές λειτουργίες, άρα μεταγραφούν διαφορετικά γονίδια. Για τα καρκινικά κύτταρα η πληροφορία για τα αίτια της καρκινογένεσης βρίσκεται στη διαφορά μεγέθους του

μεταγραφώματος σε σχέση με τα υγιή κύτταρα . Χαρακτηριστικό παράδειγμα η ανεύρεση του γονιδίου p53 τη δεκαετία του 1970 μέσω της υπέρ-έκφρασης του στα καρκινικά κύτταρα έναντι των φυσιολογικών.

Η τεχνολογία στο πέρασμα των χρόνων έχει περάσει και αυτή από διάφορα μεταβατικά στάδια με εργαστηριακές τεχνικές να διαδέχονται η μία την άλλη αυξάνοντας την απόδοση των μεθόδων και την ποιότητα του αποτελέσματος. Μέθοδοι όπως ο διαφορικός συνδυασμός ανασυνδιασμένου DNA και η ηλεκτροφόρηση πρωτεϊνών που ξεκίνησαν ήδη από την δεκαετία του 1970 να διαδίδονται αντικαταστήθηκαν από τις τεχνολογίες των μικροσυστοιχιών και του GeneChip oligo microarrays.

2 Βιβλιογραφική βάση της ανάλυσης

2.1 Γονιδιωματική ταξινόμηση του δερματικού μελανώματος

Παρακάτω στην εργασία γίνεται επιλογή των κριτηρίων και των ομάδων ασθενών που θα συγκριθούν, ώστε να μελετηθεί το μεταγραφικό τους προφίλ και να δειχθεί αν υπάρχουν διαφορές ανάμεσα στους πάσχοντες με βάση δευτερεύοντα χαρακτηριστικά.

Η επιλογή κατηγοριοποίησης των ασθενών με βάση την έκφραση ή μη, συγκεκριμένων γονιδίων, έχει βάση και υποστηρίζεται βιβλιογραφικά (Akbani et al., 2015).

Για να πραγματοποιηθεί η ταξινόμηση των ασθενών με μελάνωμα με βάση το μεταγραφικό τους προφίλ, μελετήθηκε ένα μεγάλο σετ δεδομένων που περιελάμβανε 331 ασθενείς. Είναι προφανές πως πρόκειται για μια εκτενής μελέτη που μπορεί να παρέχει στιβαρά αποτελέσματα.

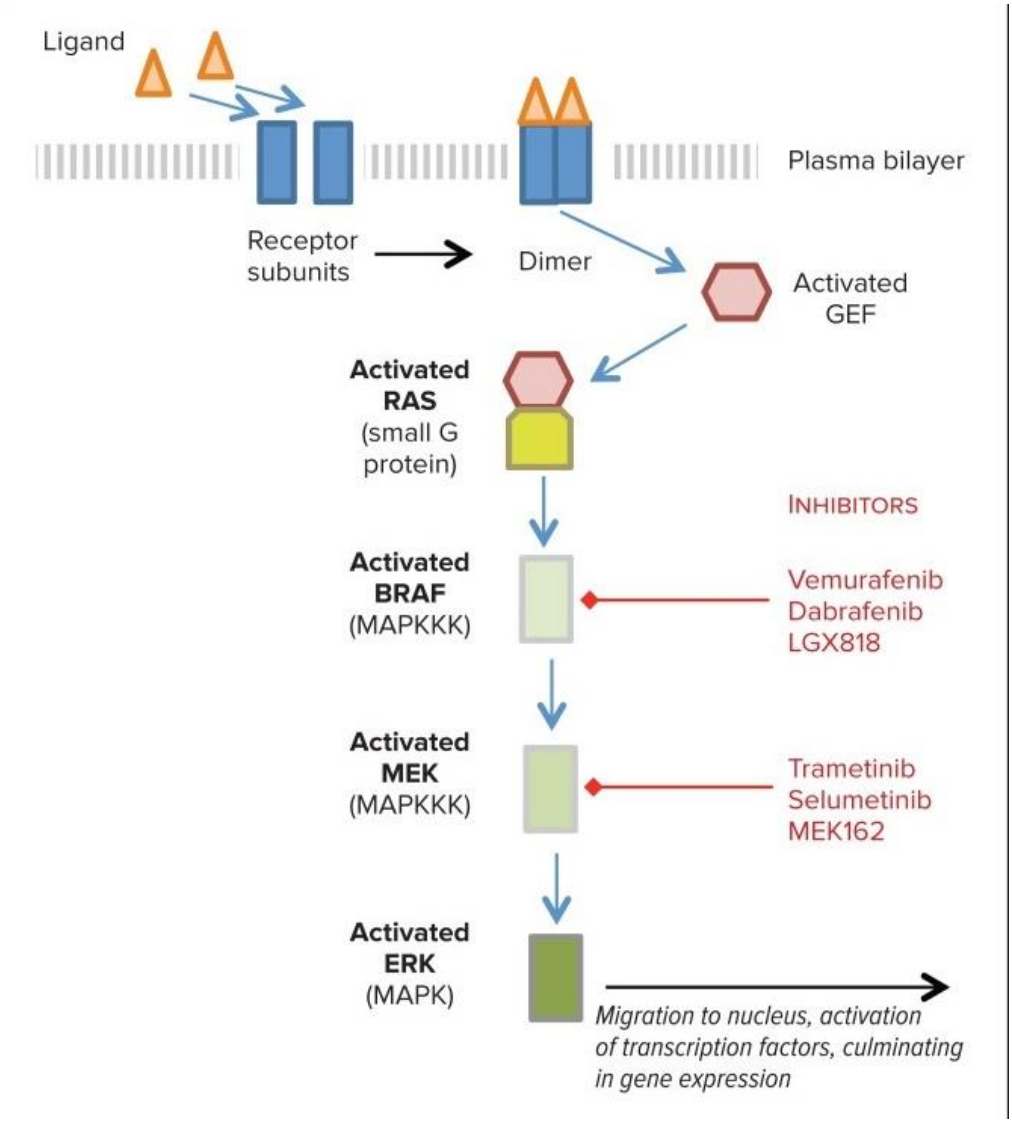
Από τους 331 ασθενείς λήφθηκαν 333 δείγματα δερματικού μελανώματος, με τα τελευταία να προέρχονται είτε από πρωτογενείς όγκους, είτε από μεταλλάξεις. Η Γενωμική ταξινόμηση έγινε σε τέσσερις υπότυπους, βασιζόμενη στα 4 κύρια πιο συχνά μεταλλασσόμενα γονίδια, μεταλλαγή στο BRAF ή σε γονίδια της οικογένειας RAS ή μεταλλαγμένο NF1 ή Triple-WT (Akbani et al., 2015).

Το γονίδιο NF1 κωδικοποιεί για την πρωτεΐνη νευροβρωμίνη (P21359, NF1_HUMAN) (UniProt Consortium, 2018), στα νευρικά κύτταρα, τα ολιγοδενδροκύτταρα, και τα κύτταρα Schwann (Rasmussen & Friedman, 2000). Η νευροβρωμίνη δρα ως καταστολέας δημιουργίας όγκων. Οι ογκοκατασταλτικές πρωτεΐνες, φυσιολογικά, εμποδίζουν τα κύτταρα να αναπτυχθούν και να διαιρεθούν ανεξέλεγκτα. Η συγκεκριμένη πρωτεΐνη δρα στην παρεμπόδιση της ανάπτυξης καρκινικών όγκων καταστέλλοντας την πρωτεΐνη Ras, που επάγει την καρκινογένεση. Τα ολιγοδενδροκύτταρα είναι τύπος νευρογλοία που παρέχουν στήριξη και μόνωση στους άξονες του Κεντρικού Νευρικού Συστήματος (ΚΝΣ) κάποιων σπονδυλωτών, δημιουργώντας μια θήκη μυελίνης με 80% λιπιδική και 20% πρωτεϊνική σύνθεση. Τέλος, τα κύτταρα Schwann αποτελούν τα κύρια γλοία του Περιφερικού Νευρικού Συστήματος.

Το BRAF ανήκει στην κατηγορία των πρωτο-ογκογονιδίων. Τα πρωτο-ογκογονίδια, αποτελούν περιοχές μεγάλης σημασίας στο γονιδίωμα, καθώς επιτελούν λειτουργίες που σχετίζονται με τον κυτταρικό πολλαπλασιασμό και την κυτταρική επιβίωση, δηλαδή ζωτικές διεργασίες για το κύτταρο και τον οργανισμό. Παρόλα αυτά, αν μεταλλαγούν, μετατρέπονται σε ογκογονίδια, επάγουν ανεξέλεγκτο πολλαπλασιασμό και αθανατοποίηση των κυττάρων και τα μετατρέπουν σε καρκινικά. Φυσιολογικά, το BRAF γονίδιο παράγει την πρωτεΐνη B-Raf, μια κινάση σερίνης-θρεονίνης, η οποία εμπλέκεται στη μεταφορά εξωκυττάριων σημάτων στον πυρήνα, μέσω του ρόλου της στο μονοπάτι ERK/MAPK.

Το μονοπάτι RAS-RAF-MEK-ERK, είναι εξαιρετικά πολύπλοκο, όμως παρακάτω περιγράφονται τα βασικά του στάδια. Η γενική δομή του μονοπατιού περιλαμβάνει μια μικρή G πρωτεΐνη (RAS) και τρεις πρωτεϊνικές κινάσες, RAF, MEK, ERK. Οι κινάσες είναι ένζυμα που καταλύουν τη μεταφορά μιας φωσφορικής ομάδας από ένα μόριο- δότη σε ένα μόριο-δέκτη, με το δεύτερο να περιλαμβάνει μια ειδική περιοχή για πρόσδεση της φωσφορικής ομάδας, η είσοδος της οποίας, συνήθως επιφέρει κάποια δομική αλλαγή στο μόριο που «ξεκλειδώνει» λειτουργικότητες που δεν είχε στην αποφωσφορυλιωμένη μορφή του, δηλαδή ενεργοποιείται. Η διαδικασία είναι αντιστρεπτή και ένας συνεχώς ενεργοποιημένος δέκτης συνδέεται με παθογένειες στο κυτταρικό επίπεδο. Συνοπτικά, η λειτουργία του μονοπατιού περιλαμβάνει τα εξής στάδια. Το εξωκυττάριο σήμα, με τη μορφή μορίου-προσδέτη π.χ. αυξητικός παράγοντας ή ορμόνη, προσδένεται στο

εξωκυττάριο τμήμα της κινάσης σερίνης-θρεονίνης. Η πρόσδεση επάγει διμερισμό του κυτταρικού υποδοχέα επιφάνειας (RTK, Receptor Tyrosine Kinase, υποδοχέας κινάσης τυροσίνης) στο κυτταροπλασματικό τμήμα του και ενεργοποίησή του. Ακολουθεί η πρόσδεση πρωτεΐνης – προσαρμογέα στην RTK, ενώ παράλληλα προσελκύονται παράγοντες ανταλλαγής (GEFs, Guanine - nucleotide Exchange Factors) στην κυτταροπλασματική μεμβράνη. Οι GEFs ενεργοποιούν μια μικρή G πρωτεΐνη, εδώ την RAS πρωτεΐνη, η οποία αρχικά έχει προσδεμένο ένα μόριο GDP (διφωσφορική γουανοσίνη) το οποίο με την ανταλλακτική δράση των GEFs, απομακρύνεται και αντικαθίσταται από το μόριο GTP (τριφωσφορική γουανοσίνη), δηλαδή προσδένει μόριο με έναν παραπάνω φώσφορο και άρα φωσφορυλιώνεται. Η ενεργοποιημένη πλέον RAS, ενεργοποιεί με τη σειρά της κάποια MAPKKK (Mitogen-Activated Protein Kinase Kinase Kinase), εδώ τη BRAF, η οποία είναι μια ειδική πρωτεϊνική κινάση σερίνης – θρεονίνης (φωσφορυλιώνει ειδικά την υρδοξυλική ομάδα μιας σερίνης ή μιας θρεονίνης και δε δρα σε άλλα αμινοξέα). Αφού γίνει αυτό, η RAS επιστρέφει στην απενεργοποιημένη της μορφή. Η μη μεταλλαγμένη RAS παραμένει ενεργή μόνο μέχρι να καλυφθούν οι ανάγκες του κυττάρου. Η ενεργοποίηση της BRAF δημιουργεί τη συνθήκη για τη συνέχιση του καταρράκτη αντιδράσεων, με την φωσφορυλίωση μιας MAPKK πρωτεΐνης (MEK). Οι MAPKK διαθέτουν διπλή ειδικότητα κινάσης. Δρουν ως κινάσες σερίνης – θρεονίνης και τυροσίνης παράλληλα. Το επόμενο μόριο στον καταρράκτη, η MAPK (εδώ ERK) χρειάζεται διπλή φωσφορυλίωση, σε δύο διαφορετικά αμινοξέα της αλληλουχίας της για να μεταβεί στην ενεργοποιημένη τη μορφή. Μετά τη φωσφορυλίωση σερίνης και γειτονικής τυροσίνης από την MAPKK, η ενεργή πλέον MAPK δρα ως ένζυμο, μεταναστεύει στον πυρήνα και ενεργοποιεί μεταγραφικούς παράγοντες ανάλογα με το αρχικό εξωκυττάριο σήμα που ενεργοποίησε τον καταρράκτη και άρα με τις ανάγκες του ιστού. Παρέμβαση σε κάποια από τα πολλαπλά στάδια του μονοπατιού που περιεγράφηκε αποτελεί σταθερά μελέτη για πιθανά αντικαρκινικά φάρμακα, που σταματούν την πρόοδο του καταρράκτη αντιδράσεων και άρα τις διαδικασίες με τις οποίες αυτός σχετίζεται (McCain, 2013).



Εικόνα 5 Διαγραμματική απεικόνιση του μονοπατιού MAPK (ERK) περιλαμβανομένων των δραστικών ουσιών που χρησιμοποιούνται συχνά ως αναστολείς σε διάφορα στάδια της διαδικασίας. Το Vemurafenib και το Dabrafenib είναι γνωστοί αναστολείς του μονοπατιού ERK και χρησιμοποιούνται στο στάδιο του ενεργοποιημένου BRAF, στην περίπτωση μεταλλαγμένου πρωτεϊνικού μορίου που φέρει τη μετάλλαξη V600E (McCain, 2013).

Το μονοπάτι MAPK/ERK ελέγχει σημαντικές κυτταρικές λειτουργίες όπως η κυτταρική διαίρεση, η κυτταρική διαφοροποίηση, η κυτταρική μετανάστευση, η απόπτωση και η φυσιολογική λειτουργία στο εμβρυικό στάδιο.

Το μελάνωμα ξεκινά στα μελανοκύτταρα, δερματικά κύτταρα που παράγουν χρωστική. Στα κύτταρα αυτά, κατά τη μεταλλαγή του BRAF V600E, η βαλίνη στη θέση 600 μετατρέπεται σε γλουταμικό και εμφανίζεται σε περίπου 50% των μελανωμάτων. Το μεταλλαγμένο BRAF αποτελεί αυτοσωμική μεταλλαγή και το αποτέλεσμα είναι η συνεχής ενεργοποίηση της BRAF

πρωτεΐνης. Μια μονή μετάλλαξη δεν μπορεί να προκαλέσει παρά μη καρκινικές δερματικές αλλαγές, άρα απαιτείται τουλάχιστον ακόμη μια μετάλλαξη, παράλληλα με το BRAF, για να εμφανιστεί μελάνωμα.

Η οικογένεια γονιδίων RAS είναι μια οικογένεια επίσης πρωτο-ογκογονιδίων που κρίσιμες μεταλλαγές τα μετατρέπουν σε ογκογονίδια. Είναι μικρές G πρωτεΐνες που συμμετέχουν στο μονοπάτι MAPK/ERK και τρία μέλη αυτής της οικογένειας συναντώνται πιο συχνά στο μελάνωμα και γενικά σε καρκίνους: N-RAS, K-RAS, H-RAS. Στην ενεργοποιημένη τους μορφή έχουν σημαντικό ρόλο στις διαδικασίες της κυτταρικής διαίρεσης, της κυτταρικής διαφοροποίησης και της επιβίωσης, όμως μια μεταλλαγμένη RAS σημαίνει πως είναι συνεχώς συνδεδεμένη με GTP και άρα συνεχώς ενεργή. Η απουσία δυνατότητας εναλλαγής ενεργής και ανενεργής μορφής με βάση τις απαιτήσεις του κυττάρου είναι που καταλήγει σε υπέρ-ενεργοποίηση του μονοπατιού που συμμετέχει από εκείνο το στάδιο και μετά.

Η ομάδα Triple-WT αποτελεί ετερογενή υπό-ομάδα καθώς περιλαμβάνει τις μεταλλαγές που δεν είναι hot-spots από τις τρεις προηγούμενες κατηγορίες και συνοπτικά, μεταλλαγές της KIT, που εμφανίζονται σε 6%-8% των μελανωμάτων, δομικές ανακατατάξεις και αναδιπλασιασμούς.

3 Μέθοδοι – Εργαλεία Λογισμικού

3.1 Πηγές Εξόρυξης Δεδομένων Αλληλούχισης RNA – GDC

Genomic Data Commons – GDC

Η βιολογική βάση δεδομένων Genomic Data Commons, GDC, του National Cancer Institute (NCI) (Grossman et al., 2016), αποτελεί ουσιαστικά ένα δίκτυο πληροφορίας επόμενης γενιάς σχετικά με τον καρκίνο, η οποία υποστηρίζει τη διάθεση και την τυποποίηση γενωμικών και κλινικών δεδομένων που προέρχονται από ερευνητικά προγράμματα σχετιζόμενα με τον καρκίνο, όπως τα TCGA, TARGET και CGCI. Η εναρμόνιση των δεδομένων αλληλούχισης με το γονιδίωμα ή το μεταγράφωμα και η εφαρμογή πρότυπων μεθόδων στα δεδομένα αυτά, προκειμένου να αντληθεί η βιολογική πληροφορία είναι οι βασικοί στόχοι τέτοιων δομών. Το τμήμα NCI ανήκει στο NIH των ΗΠΑ που σε συνδυασμό με το Department of Health and

Human Services (DHHS) δημιούργησαν την GDC, ώστε να παρέχεται στην επιστημονική κοινότητα η δυνατότητα λήψης, ποιοτικού ελέγχου, αποθήκευσης και αναδιανομής τυποποιημένων καρκινικών δεδομένων αλληλούχισης από πειράματα σχετικά με διάφορες μορφές καρκίνου (Grossman et al., 2016).

Ο βασικός και κύριος στόχος της GDC είναι να παρέχει στην ερευνητική κοινότητα του καρκίνου ένα αποθετήριο ομαδοποιημένων δεδομένων προερχόμενων από μελέτες του γονιδιώματος ασθενών με καρκίνο. Το αποθετήριο αυτό, παρέχει ένα δίκτυο ηνώσης σχετικά με τον καρκίνο και κάνει δυνατό τον προσδιορισμό σπάνιων παραγόντων που επάγουν την εμφάνιση καρκίνου, υποστηρίζει την εύρεση γενετικών παραγόντων που έχουν ρόλο τόσο στην ανάπτυξη της νόσου όσο και στην αποδοτική θεραπεία και πληροφορεί τους ιθύνοντες του τομέα κλινικών δοκιμών σχετικά με τις στοχευμένες γενετικές αλλοιώσεις (Grossman et al., 2016). Δουλεύοντας προς την επίτευξη αυτού του στόχο, η GDC παρέχει πηγές εξόρυξης δεδομένων σε διάφορες μορφές και σε κάποιες περιπτώσεις με κάποια προ-επεξεργασία η οποία αναλύεται εντός της GDC και η αντίστοιχη βιβλιογραφία είναι διαθέσιμη.

Τα δεδομένα της GDC είναι εναρμονισμένα με το γονιδίωμα αναφοράς GRCh38.

Η GDC παράγει δεδομένα υψηλής ποιότητας για κληρονομήσιμους και σωματικούς γονότυπους, ποσοτικοποίηση και δομική ανάλυση RNA – Seq δεδομένων και διάφορους τύπους γενετικού σχολιασμού (Grossman et al., 2016).

3.2 Ροή Ανάλυσης mRNA – Προ-επεξεργασία Δεδομένων

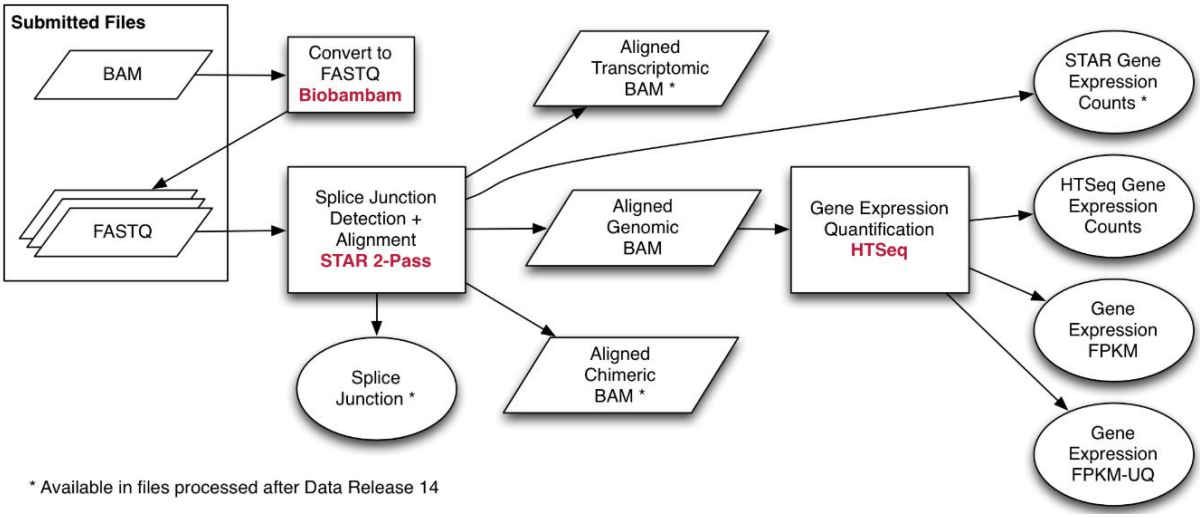
Η ροή ανάλυσης της GDC για ποσοτικοποίηση της πληροφορίας που προέρχεται από πειράματα αλληλούχισης RNA μετρά τα επίπεδα γονιδιακής έκφρασης σε αριθμό αναγνώσεων ποσοτικών δεδομένων HT-Seq, σε αριθμό αναγνώσεων ανά 1000 βάσεις (FPKM) και κανονικοποίηση ποσοστιμορίων (FPKM-UQ). Αυτές οι τιμές προκύπτουν μέσω μια ροής ανάλυσης που περιλαμβάνει τα βήματα της στοίχισης των αναγνώσεων στο γονιδίωμα αναφοράς GRCh38 και έπειτα ποσοτικοποιούνται οι στοιχισμένες αναγνώσεις. Προκειμένου να διευκολυνθεί η εναρμόνιση των δεδομένων κατά μήκος των δειγμάτων σε κάθε περίπτωση, όλες οι αναγνώσεις της RNA αλληλούχισης μεταχειρίζονται από τη GDC μη ελικοποιημένα και μονόκλωνα (“Bioinformatics Pipeline: mRNA Analysis - GDC Docs,” n.d.; Grossman et al., 2016).

3.2.1 Βήματα της επεξεργασίας των RNA – Seq δεδομένων

Ροή στοίχισης των αλληλουχιών RNA

Η ροή ανάλυσης του mRNA ξεκινά με τη στοίχιση των αναγνώσεων στο γονιδίωμα αναφοράς, το οποίο πραγματοποιείται με τη μέθοδο STAR (Dobin, 2014). Η STAR στοιχίζει κάθε ομάδα αναγνώσεων χωριστά και έπειτα ενοποιεί τα αποτελέσματα σε μια ενιαία στοίχιση. Στη συνέχεια, οι μέθοδοι που χρησιμοποιούνται από το International Genome Consortium (ICGC) περιλαμβάνουν μια διασταύρωση ματίσματος και ένα βήμα ανίχνευσης, τα οποία χρησιμοποιούνται για να εξάγουν την τελική στοίχιση. Αυτή η αλληλουχία ενεργειών δίνει σαν εξαγόμενο ένα αρχείο BAM, το οποίο περιλαμβάνει τόσο τις στοιχισμένες αναγνώσεις όσο και αυτές που δεν μπόρεσαν να στοιχηθούν (Grossman et al., 2016). Η αξιολόγηση της ποιότητας των δεδομένων πραγματοποιείται πρότερα της στοίχισης με τον FASTQC (“Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data,” n.d.) αλγόριθμο και μετά τη στοίχιση με τα εργαλεία Picard (“Picard Tools - By Broad Institute,” n.d.).

Τα αρχεία των οποίων η επεξεργασία έγινε με την Data Release 14, έχουν ήδη συσχετισμένα τα μεταγραφικά δεδομένα και τις χιμαιρικές στοιχίσεις, επιπλέον της γενωμικής στοίχισης που περιεγράφηκε στην προηγούμενη παράγραφο. Τα χιμαιρικά BAM αρχεία περιέχουν αναγνώσεις που έχουν χαρτογραφηθεί σε διαφορετικά χρωμοσώματα ή κλώνους. Τα αρχεία γενωμικής στοίχισης περιέχουν χιμαιρικά και μη στοιχισμένα δεδομένα αναγνώσεων με τις μεταγραφικές συντεταγμένες και όχι τις γενωμικές. Η στοίχιση του μεταγράφου επίσης διαχειρίζεται διαφορετικά, για να διευκολύνει την μεταγενέστερη ανάλυση.



Εικόνα 6 Ροή εργασιών και εφαρμογής αλγορίθμων προκειμένου να προκύψουν τα αρχεία δεδομένων που είναι διαθέσιμα στη GDC database. Η παραπάνω ροή υλοποιείται από τη GDC ως προ-επεξεργασία των δεδομένων για τα περισσότερα διαθέσιμα πακέτα δεδομένων της βάσης και είναι έγκυρη και τυποποιημένη (Grossman et al., 2016).

Ροή ανάλυσης της έκφρασης mRNA

Μετά τη στοίχιση, τα αρχεία BAM επεξεργάζονται μέσω της αλγοριθμικής ροής RNA Expression Workflow για να προσδιοριστούν τα επίπεδα έκφρασης του RNA. Οι αναγνώσεις που χαρτογραφούνται πάνω στο κάθε γονίδιο απαριθμούνται με τη χρήση του HT-Seq-Count. Οι τιμές έκφρασης παρέχονται σε tab-delimited μορφοποίηση αρχείου και χρησιμοποιείται το εργαλείο GENCODEv22 για να γίνει ο γονιδιακός σχολιασμός (Grossman et al., 2016). Οπότε η είσοδος του αλγορίθμου HT-Seq (HT-Seq-0.6.1p1) είναι οι στοιχισμένες αλληλουχίες και το αποτέλεσμα είναι ένα αρχείο με την ποσοτικοποιημένη γονιδιακή έκφραση ανά γονίδιο.

Κανονικοποίηση HT-Seq της έκφρασης mRNA

Οι μετρήσεις των επιπέδων έκφρασης των αναγνώσεων της RNA αλληλούχισης παράγονται από τον αλγόριθμο HT-Seq-Counts και κανονικοποιούνται χρησιμοποιώντας δύο μεθόδους που μοιάζουν μεταξύ τους, την FPKM και την FPKM-UQ. Οι κανονικοποιημένες τιμές για τα δεδομένα πρέπει να χρησιμοποιούνται μόνο με υπόβαθρο ολόκληρου του σετ γονιδίων του οργανισμού. Ο χρήστης θα πρέπει να χρησιμοποιήσει κάποια επιπλέον κανονικοποίηση, αν ασχολείται μόνο με κάποια υποομάδα των γονιδιών.

Οι δύο παραπάνω μέθοδοι έχουν αναλυθεί στην εισαγωγή και εδώ γίνεται απλή αναφορά τους (ενότητες 1.3.2, 1.3.3).

Type	Description	Format
RNA-Seq Alignment	RNA-Seq reads that have been aligned to the GRCh38 build. Reads that were not aligned are included to facilitate the availability of raw read sets	BAM
HT-Seq Read Counts	The number of reads aligned to each gene, calculated by HT-Seq	TXT
STAR Read Counts	The number of reads aligned to each gene, calculated by STAR	TSV
FPKM	A normalized expression value that takes into account each gene length and the number of reads mapped to all protein-coding genes	TXT
FPKM-UQ	A modified version of the FPKM formula in which the 75th percentile read count is used as the denominator in place of the total number of protein-coding reads	TXT

Εικόνα 7 Συνοπτικός πίνακας των μορφών δεδομένων που μπορούν να αντληθούν από την GDC (Grossman et al., 2016).

Ανάλυση της ροής εργασιών που επιτελεί ο αλγόριθμος HT-Seq-Counts

Δεδομένου του αρχείου που περιέχει τις στοιχισμένες αναγνώσεις και μιας λίστας με τα γονιδιωματικά χαρακτηριστικά, ένα κοινό και λογικό βήμα είναι η μέτρηση του πόσες αναγνώσεις χαρτογραφούνται σε κάθε χαρακτηριστικό (Grossman et al., 2016), ενώ αυτό το χαρακτηριστικό μπορεί να είναι ένα γονίδιο, μια μεταλλαγή, ένα SNP κοκ δηλαδή μια καθορισμένη περιοχή πάνω στο χρωμόσωμα ή ένα σύνολο τέτοιων περιοχών.

Στην περίπτωση της αλληλούχισης RNA, τα χαρακτηριστικά είναι τυπικά γονίδια, όπου κάθε γονίδιο θεωρείται ένα σύνολο από όλα τα αντίστοιχα εξώνια, παρόλο που και κάθε εξώνιο θα μπορούσε, υπό άλλες συνθήκες μελέτης να θεωρηθεί και αυτό ως χαρακτηριστικό υπό μελέτη, παραδείγματος χάρη στην περίπτωση μελέτης του εναλλακτικού ματίσματος. Σαν ένα διαφορετικό παράδειγμα, στην περίπτωση του ChIP-Seq, το χαρακτηριστικό θα μπορούσε να αναχθεί σε μια περιοχή πρόσδεσης από μια προκαθορισμένη λίστα.

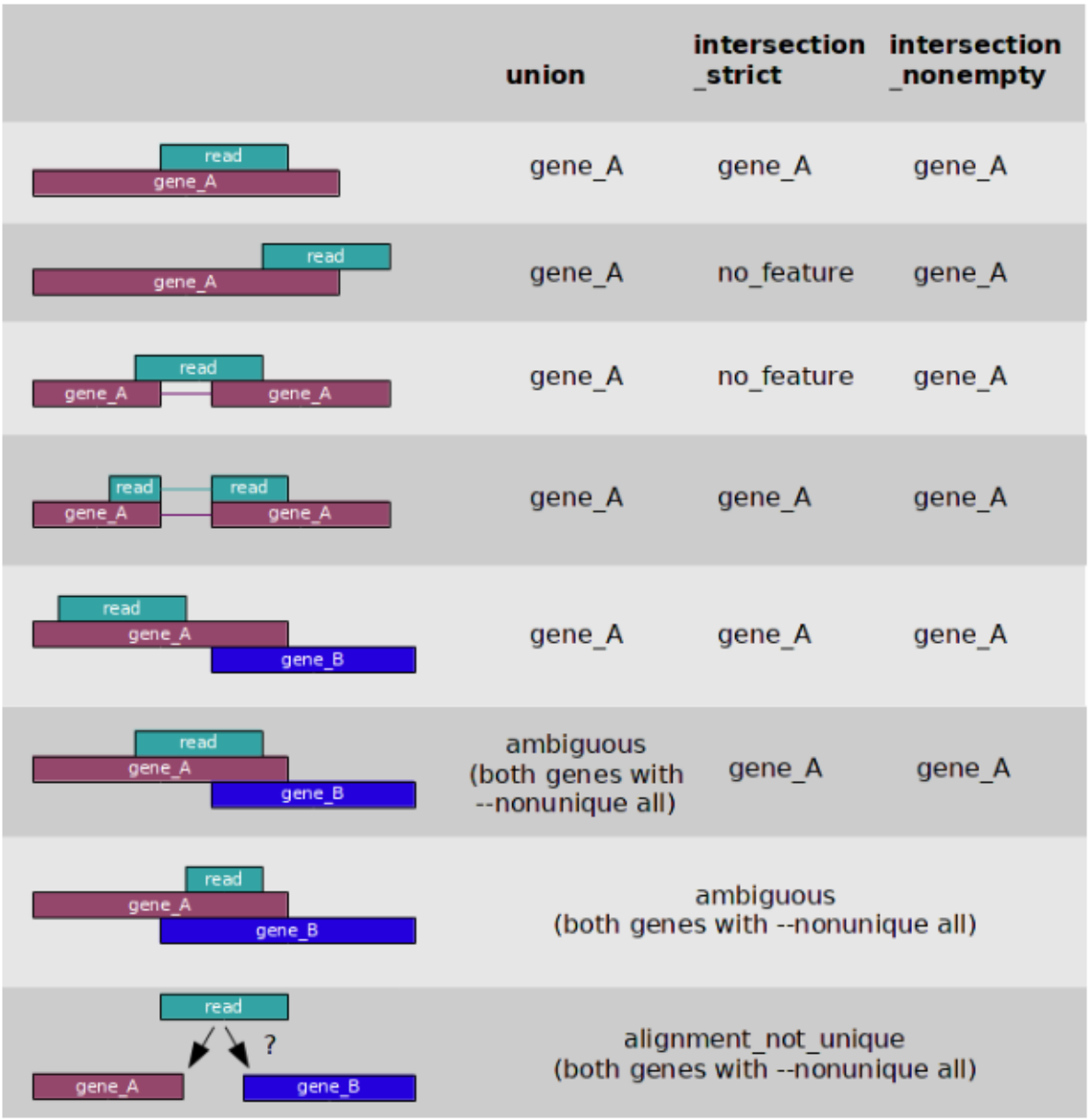
Ιδιαίτερη προσοχή πρέπει να δοθεί στην απόφαση του πως θα γίνει η διαχείριση των αναγνώσεων που αντιστοιχίζονται σε περισσότερα από ένα χαρακτηριστικά ενδιαφέροντος. Η αλληλουχία εντολών του htseq-count επιτρέπει την εφαρμογή ανάμεσα σε τρεις επιλογές. Φυσικά, αν καμιά από αυτές τις επιλογές δεν ταιριάζει με τις ανάγκες του πειράματος, υπάρχει δυνατότητα παρέμβασης στην κώδικα των εντολών.

Η λειτουργία του htseq-count έγκειται στην εξής βάση: για κάθε θέση i της ανάγνωσης, ένα σετ S(i) προσδιορίζεται ως ένα σύνολο όλων των χαρακτηριστικών που συμπίπτουν στη θέση i. Έπειτα το S σύνολο, όπου το i σταδιακά περνά από όλες τις θέσεις της ανάγνωσης, μπορεί να είναι:

- Η μονάδα/ένωση όλων των σετ S(i) για την επιλογή union. Η επιλογή αυτή συστήνεται για τις περισσότερες περιπτώσεις.
- Η τομή όλων των σετ s(i) για την επιλογή interception-strict
- Η τομή όλων των μη κενών σετ S(i) για την επιλογή intersection-nonempty.

Αν το S εμπεριέχει ακριβώς ένα χαρακτηριστικό, η ανάγνωση ή το σετ αναγνώσεων προσμετράται για αυτό το χαρακτηριστικό και μόνο. Αν το σύνολο S είναι κενό, η ανάγνωση προσμετράται σαν να μην υπάρχει χαρακτηριστικό, no_feature. Αν το S περιέχει περισσότερα από ένα χαρακτηριστικά, η htseq-count συμπεριφέρεται διαφορετικά, δεδομένης της επιλογής –nonunique.

Περεταίρω λεπτομέρειες θα ήταν πολύ τεχνικές και ξεφεύγουν από το πλάισιο ενδιαφέροντος της διπλωματικής εργασίας. Παρόλα αυτά, στην παρακάτω εικόνα (), αναπαριστάται γραφικά η λειτουργία του htseq-count ανάλογα τις διάφορες επιλογές που μπορεί να ακολουθήσει ο εκάστοτε χρήστης.



Εικόνα 8 Γραφική αναπαράσταση των επιλογών που δίνει η συνάρτηση htseq-count ανάλογα με τις ιδιαιτερότητες του σετ δεδομένων και των αναγνώσεων, για την ποσοτικοποίηση της γονιδιακής έκφρασης (Anders, Reyes, & Huber, 2012).

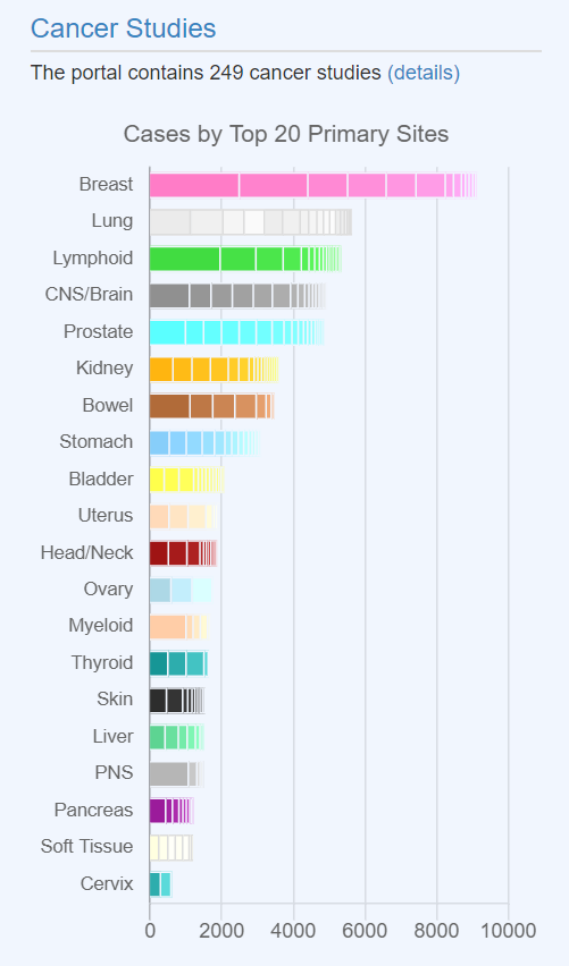
3.3 Αποθετήριο δεδομένων γονιδιώματος cBioportal

Η εφαρμογή cBioPortal αναπτύχθηκε και υποστηρίζεται από το Computational Biology Center του Memorial Sloan-Kettering Cancer Center και του i-Vis (Information Visualization) Research Group του Computer Engineering Department του Bilkent University.

Αποτελεί έναν χώρο αποθήκευσης ολοκληρωμένων δεδομένων του γονιδιώματος μεγάλης κλίμακας. Προσφέρει τη δυνατότητα για απλή μεταφόρτωση δεδομένων ανάλογα με αυτά που ορίζει ο χρήστης.

Μέσα από την εφαρμογή αυτή μπορεί να γίνει ανάλυση για την αναζήτηση μεγάλης κλίμακας καρκινικών γενωμικών δεδομένων. Μέσα από την εφαρμογή αυτή μπορούμε να βρούμε γονιδιωματικές αλλαγές σε ένα σύνολο των ασθενών, τύπων καρκίνου και την εκτέλεση επιβίωσης και ανάλυσης του δικτύου. Επιπλέον δίνεται η δυνατότητα εξερεύνησης του μονοπατιού, κατόπιν επιλογής μας, σε έναν ή περισσότερους τύπους καρκίνου. Ακόμα είναι προσβάσιμα τα κλινικά δεδομένα του ασθενούς, εφόσον τα στοιχεία δίνονται.

Για να πραγματοποιηθεί αποθήκευση δεδομένων πρέπει η εκάστοτε μελέτη να έχει δημοσιοποιηθεί. Επίσης δεν γίνεται αποθήκευση προσωρινών (provisional) δεδομένων τα οποία παράγονται από το πρόγραμμα TCGA (Cerami et al., 2012; Gao et al., 2013).



Εικόνα 9 Διαγραμματική απεικόνιση των μελετών που περιέχονται στο ηλεκτρονικό αποθετήριο cBioportal ανά καρκινικό τύπο (Cerami et al., 2012; Gao et al., 2013).

3.3.1 OncoPrint

Το Oncoprint είναι ένα συμπαγές μέσο οπτικοποίησης του γονιδιώματος με διακριτές μεταβολές. Μπορεί να εμφανίσει σωματικές μεταλλάξεις, αριθμό μεταβολών, και αλλαγές έκφρασης του mRNA σε ένα σύνολο περιπτώσεων. Επιπλέον, είναι χρήσιμο για την οπτικοποίηση ενός συνόλου γονιδίων και ενός μεταβολικού μονοπατιού σε μια σειρά περιπτώσεων, καθώς και των οπτικών κατευθύνσεων, όπως αυτών σε αμοιβαία αποκλειστικότητα ή συν-εμφάνιση μεταξύ ζευγών γονιδίων μέσα σε ένα σύνολο γονιδίων. Τα μεμονωμένα γονίδια αναγράφονται ως σειρές ενώ οι μεμονωμένες περιπτώσεις ή ασθενείς ως στήλες. Η σειρά εμφάνισης των γονιδίων καθορίζεται από την σειρά τοποθέτησης τους στο αρχικό πεδίο ερωτήματος (Cerami et al., 2012; Gao et al., 2013).

3.3.2 Πηγές δεδομένων του cBioportal

Τα δεδομένα προέρχονται από το PathwayCommons που περιλαμβάνει δύο οδούς:

- Binary or Pairwise Interaction Networks: αποτελείται από κόμβους του γονιδίου και αλληλεπιδράσεις στην άκρη, και δεν αντιπροσωπεύουν την πλήρη σημασιολογική πολυπλοκότητα των βιολογικών οδών.
- BioPax Netwokrs: είναι μια τυπική γλώσσα με στόχο την ολοκλήρωση, ανταλλαγή, οπτικοποίηση και ανάλυση των στοιχείων μιας βιολογικής οδού. Μπορεί να εκπροσωπήσει την πολυπλοκότητα των πολλαπλών τύπων βιολογικών μονοπατιών, μεταβολικών οδών, μοριακών αλληλεπιδράσεων, οδών σηματοδότησης, ρύθμιση των γονιδίων και των γενετικών αλληλεπιδράσεων.

3.4 Το πακέτο edgeR

Έχοντας δει τα βασικά βήματα για την αρχική ανάλυση των δεδομένων έκφρασης η διαδικασία που έπεται είναι αυτή με την οποία θα εκτιμήσουμε τη διαφορική έκφραση γονιδίων από ένα πείραμα που διενεργείται σε γονιδιωματική κλίμακα. Το ερώτημα είναι: Δεδομένων τιμών έκφρασης για N διαφορετικά γονίδια του ίδιου οργανισμού για δύο διαφορετικές συνθήκες, πώς θα προσδιορίσουμε ποια γονίδια είναι ενεργοποιημένα, ποια κατεσταλμένα και ποια εκείνα των οποίων η έκφραση δε μεταβάλλεται; Στην περίπτωση πειραμάτων μεγάλης κλίμακας ο αριθμός N μπορεί να βρίσκεται μεταξύ 3000 (για ένα απλό προκαρυωτικό γονιδίωμα) και 60000 περίπου (για ένα σύνθετο ευκαρυωτικό γονιδίωμα στο οποίο εξετάζουμε και εναλλακτικές μορφές μεταγράφων). Οι διαφορετικές συνθήκες μπορεί να είναι στάδια στην ανάπτυξη ενός οργανισμού, διαφορετικοί κυτταρικοί τύποι, ο ίδιος κυτταρικός τύπος πριν και μετά την επίδραση μιας ουσίας ή ενός τροποποιητικού παράγοντα ή ένας πληθυσμός υγιών έναντι παθολογικών δειγμάτων. Σε κάθε περίπτωση θα πρέπει να σημειωθεί ότι η σύγκριση γίνεται μεταξύ δύο καταστάσεων, καθώς η διαφορική έκφραση βασίζεται στην έννοια της μεταβολής.

Ένας πίνακας που περιέχει αυτού του είδους την πληροφορία, δηλαδή την κωδικοποίηση των δειγμάτων / ατόμων, την μεταγραφική τους πληροφορία και δευτερεύοντα χαρακτηριστικά με βάση τα οποία επιλέγεται να γίνει ο διαχωρισμός του αρχικού πλήθους δειγμάτων σε δύο διακριτές υπο – ομάδες, αποτελεί το αρχείο εισόδου για τον αλγόριθμο

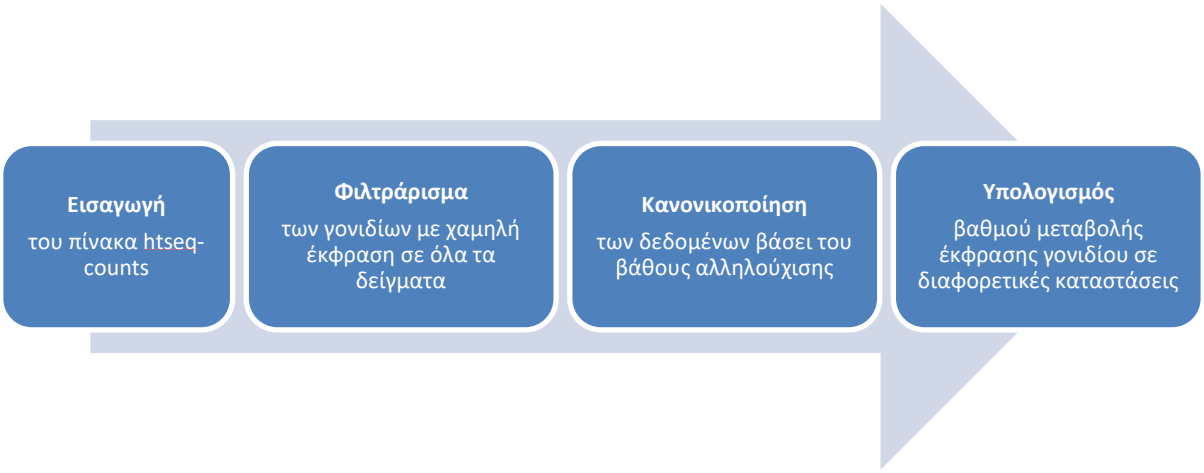
edgeR (Robinson, McCarthy, & Smyth, 2010). Στο σημείο αυτό θα πρέπει να αναφερθεί ότι δεν υπάρχει κάποιο εμπορικό ή ακαδημαϊκό προγραμματιστικό πακέτο το οποίο να διενεργεί αυτοματοποιημένα την ανάλυση διαφορικής έκφρασης γονιδίων με το EdgeR. Ο χρήστης είναι απαραίτητο να έχει γνώσεις προγραμματισμού της γλώσσας R ώστε να μπορεί να καλεί τις μεθόδους του πακέτου edgeR και να υλοποιήσει αλγοριθμικά την εν λόγω ανάλυση. Για της ανάγκες της συγκεκριμένης διπλωματικής εργασίας, σχεδιάστηκε και υλοποιήθηκε μια ροή εντολών της γλώσσας R, η οποία απαιτούσε να είναι εγκατεστημένο το προγραμματιστικό περιβάλλον της γλώσσας προγραμματισμού R, εδώ το R.studio και όλες τις απαραίτητες για τη λειτουργία του αλγόριθμου βιβλιοθήκες (όπως το Bioconductor, η limma και το edgeR). Αυτό σημαίνει πως ένας χρήστης που δεν έχει εμπειρία στη γλώσσα προγραμματισμού R, δεν μπορεί να εκτελέσει την ανάλυση διαφορικής έκφρασης γονιδίων χωρίς να χρειάζεται να σχεδιάσει και να υλοποιήσει προγραμματιστικά έναν αναλυτικό αλγόριθμο.

Το edgeR είναι ένα στατιστικό πακέτο της γλώσσας προγραμματισμού R το οποίο μπορεί να χρησιμοποιηθεί για τον στατιστικό έλεγχο της διαφορικής έκφρασης γονιδίων. Το πακέτο edgeR, όπως θα αναλυθεί με λεπτομέρειες στη συνέχεια, αξιοποιεί μεθόδους φιλτραρίσματος και κανονικοποίησης των δεδομένων και για τον λόγο αυτό ο χρήστης θα πρέπει να εισάγει μη κανονικοποιημένα δεδομένα στο αρχείο εισόδου (raw counts). Στη συνέχεια ακολουθεί η περιγραφή των βημάτων που εκτελεί ο αλγόριθμος για να εκτελέσει τον στατιστικό έλεγχο διαφορικής έκφρασης γονιδίων.

3.4.1 Φιλτράρισμα των δεδομένων

Σε πρώτο βήμα ο αλγόριθμος φιλτράρει όλα τα γονίδια τα οποία παρουσιάζουν πολύ χαμηλή έκφραση σε όλα τα δείγματα, καθώς δημιουργούν τεχνικό θόρυβο. Σε βιολογική βάση, το συγκεκριμένο βήμα φιλτραρίσματος δικαιολογείται καθώς για την μετάφραση ενός μεταγράφου σε πρωτεΐνη, αυτό θα πρέπει να εκφράζεται στο κύτταρο σε ικανή ποσότητα. Έτσι, ο αλγόριθμος απορρίπτει πριν ξεκινήσει η ανάλυση τα μετάγραφα τα οποία παρουσιάζουν πολύ χαμηλή έκφραση. Σαν κατώφλι ορίζεται διαφορετικός αριθμός μεταγράφων ανά βιβλιοθήκη, ανάλογα με το πλήθος των δειγμάτων που ανήκει σε κάθε μια από τις κατηγορίες που συγκρίνονται, το οποίο όμως δεν ξεπερνά ποτέ τα δείγματα που

υπάρχουν στη μικρότερη κατηγορία ανά βιβλιοθήκη. Οποιαδήποτε τιμή μικρότερη θεωρείται ως μη εκφρασμένο γονίδιο και απορρίπτεται.



Εικόνα 10 Διαγραμματική απεικόνιση της ροής διαδικασιών από την εισαγωγή του πίνακα htseq – counts στο περιβάλον της R για ανάλυση με το πακέτο edgeR έως και τον υπολογισμό του βαθμού μεταβολής έκφρασης του κάθε γονιδίου στις δύο διαφορετικές καταστάσεις. Αρχικά εισάγεται στην R ο προαναφερθείς πίνακας ως αντικείμενο RDGEList. Στη συνέχεια το βήμα του φιλτραρίσματος πραγματοποιείται με βάση τις τιμές cpm και lcpm, ενώ το βήμα τις κανονικοποίησης με τη μέθοδο TMM, Trimmed mean of M – values. Τέλος, ο βαθμός μεταβολής δίνεται από τον στατιστικό έλεγχο της τιμής log2FC, ο οποίος αναλύεται στην επόμενη ενότητα.

3.4.2 Κανονικοποίηση δεδομένων

Αρχικά το πακέτο edgeR διενεργεί κανονικοποίηση των δεδομένων βάσει του βάθους αλληλούχισης (sequence depth). Η διόρθωση είναι βασικό μέρος του αλγορίθμου και εκφράζεται στα μεγέθη fold – change και p – value, χωρίς ο χρήστης να χρειάζεται να ορίσει τις παραμέτρους κανονικοποίησης.

Ένα από τα κυριότερα τεχνικά προβλήματα που ενδέχεται να επηρεάσει τον υπολογισμό της έκφρασης των γονιδίων σε ότι αφορά στις RNA – seq βιβλιοθήκες, είναι ότι γίνεται ο υπολογισμός της σχετικής περιεκτικότητας σε RNA για κάθε δείγμα, ωστόσο δεν υπολογίζεται το ολικό περιεχόμενο σε RNA μετάγραφα ανά κύτταρο. Το γεγονός αυτό είναι σημαντικό πρόβλημα, ιδιαίτερα στις περιπτώσεις όπου ένας μικρός αριθμός γονιδίων βρίσκεται υπερεκφρασμένος σε ένα δείγμα, αλλά όχι και στα υπόλοιπα. Τα υψηλά εκφρασμένα γονίδια είναι πιθανό να δεσμεύσουν μεγάλο τμήμα της RNA – seq βιβλιοθήκης, οδηγώντας έτσι σε υποεκτίμηση των υπολοίπων γονιδίων που βρίσκονται στο συγκεκριμένο δείγμα. Το edgeR διαθέτει σαν προεπιλογή την διόρθωση αυτού του προβλήματος μέσω της

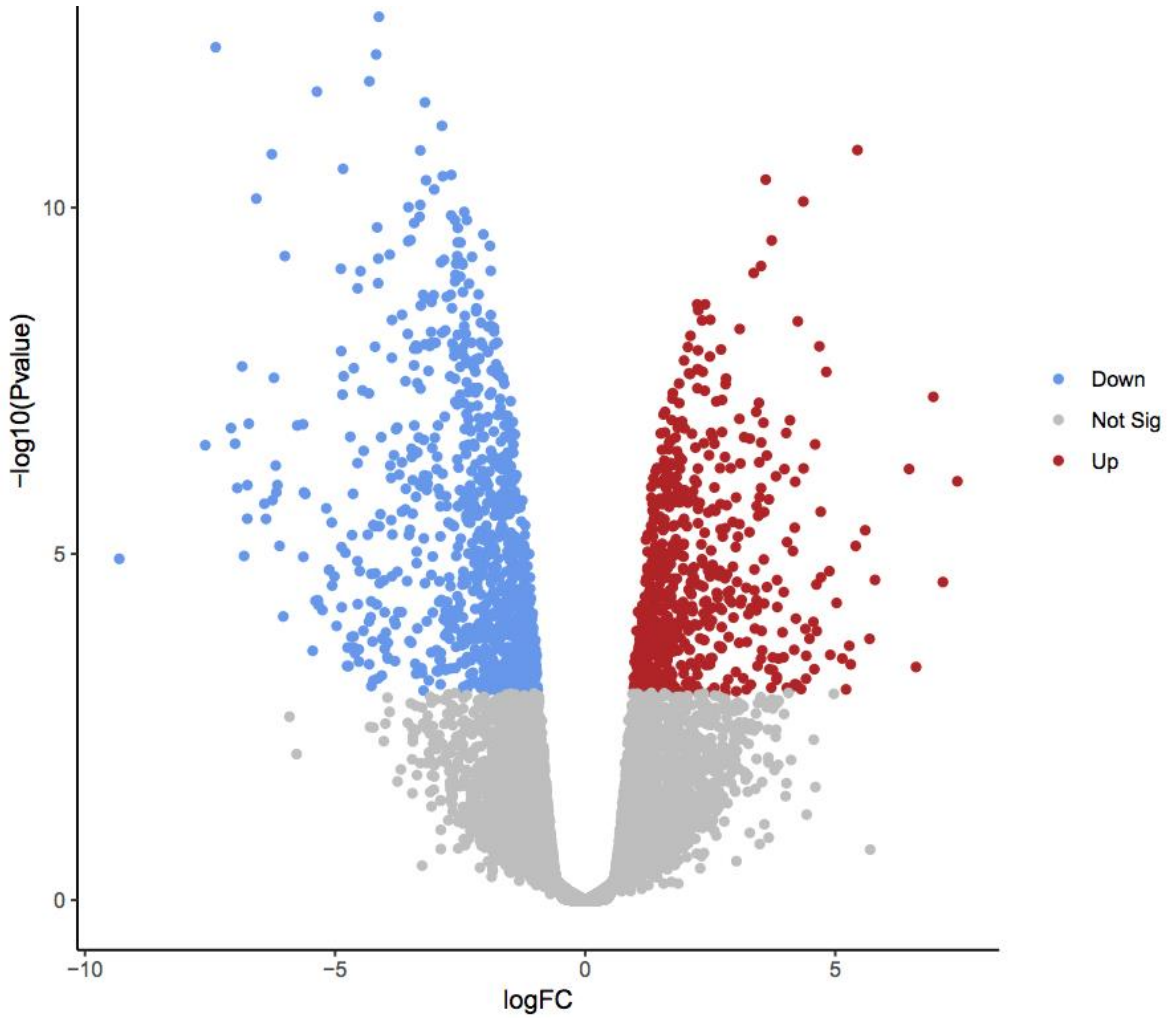
μεθόδου κανονικοποίησης «trimmed mean of M – values», TMM (Robinson and Oshlack, 2010).

3.4.3 Στατιστική Ανάλυση Διαφορικής Έκφρασης

Η σχετική έκφραση ποσοτικοποιείται σε μια λογαριθμική κλίμακα λόγων. Πώς όμως μπορούμε να προσδιορίσουμε κατά πόσο μια τιμή log2FC=1 είναι αρκετά μεγάλη ώστε να κατατάξουμε το γονίδιο στο οποίο αντιστοιχεί στα ενεργοποιημένα; Θα χρειαστούμε ένα στατιστικό έλεγχο της τιμής αυτής ώστε να αξιολογήσουμε σε ποιο βαθμό η διπλάσια έκφραση αντανακλά ένα υπαρκτό βιολογικό φαινόμενο ή μια τυχαία διακύμανση στα επίπεδα του mRNA. Οπωσδήποτε, ακραίες τιμές log2FC είναι καταρχάς ενδεικτικές πραγματικών διαφορών, ωστόσο ένα απλό διαισθητικά ερώτημα σχετίζεται με το κατά πόσο μια τιμή είναι επαναλήψιμη. Πιο απλά, αν θα παίρναμε την ίδια (ή σχεδόν την ίδια) τιμή log2FC αν επαναλαμβάναμε το πείραμα. Η έννοια της επαναληψιμότητας είναι βασική για το στατιστικό (και ουσιαστικό) έλεγχο κάθε πειράματος, αλλά στην περίπτωση των πειραμάτων έκφρασης έχει ιδιαίτερα κεντρικό ρόλο. Είναι πρακτικά αδύνατο να αξιολογήσουμε τη διαφορική έκφραση μ' ένα μοναδικό πείραμα. Κι αυτό γιατί ο στατιστικός έλεγχος απαιτεί να συγκρίνουμε όχι ένα ζεύγος τιμών για κάθε γονίδιο αλλά ένα ζεύγος κατανομών των τιμών αυτών. Απ' αυτήν τη σκοπιά είναι σημαντικό να γνωρίζουμε ότι κάθε πείραμα έκφρασης θα πρέπει να διενεργείται τουλάχιστο σε τρεις επαναλήψεις, που σημαίνει ότι για κάθε συνθήκη και για κάθε γονίδιο θα πρέπει να έχουμε τουλάχιστον τρεις μετρήσεις έκφρασης (Lee, Poh, & Loke, 2002). Ο λόγος γι' αυτόν τον περιορισμό είναι ότι για να αξιολογήσουμε τη διαφορική έκφραση στατιστικά, θα πρέπει να υπολογίσουμε μια πιθανότητα που να εκτιμά τη διαφορά μεταξύ των μέσων τιμών των μετρήσεων για κάθε συνθήκη (και για μια μέση τιμή καλό είναι να έχουμε τουλάχιστον τρεις μετρήσεις). Σε πειράματα έκφρασης, όπου εξετάζουμε μια συνθήκη μελέτης με μια συνθήκη ελέγχου, ο στατιστικός έλεγχος γίνεται με τη χρήση του ελέγχου t (Student's t-test ή πιο απλά t-test) που αποτελεί ένα στατιστικό έλεγχο υποθέσεων. Η υπόθεση αυτή που στη στατιστική ονομάζεται και “μηδενική υπόθεση” (null hypothesis) και συμβολίζεται με H0 είναι, στην περίπτωση του t-test, ότι οι μέσες τιμές δύο συνόλων τιμών είναι ταυτόσημες . Πιο αναλυτικά, δεδομένων δύο συνόλων τιμών X1 και X2, το t-test υπολογίζει ένα μέγεθος t το οποίο είναι μικρό αν οι μέσες τιμές μ1 και μ2 είναι παραπλήσιες. Όσο μεγαλύτερο είναι το t τόσο μικρότερη είναι η

πιθανότητα οι δύο μέσες τιμές να ταυτίζονται. Η διενέργεια του t-test γίνεται πολύ γρήγορα σε όλα τα διαθέσιμα στατιστικά προγράμματα (R, SPSS, Matlab κλπ) και σε κάθε περίπτωση η αποδιδόμενη τιμή είναι η πιθανότητα ταύτισης των δύο μέσων τιμών, που αντιστοιχίζεται σε μια τιμή p (p-value). Όσο μικρότερη είναι η τιμή p - value τόσο μικρότερη είναι η πιθανότητα οι δύο μέσες τιμές να είναι ίδιες και συνεπώς τα δύο δείγματα X1 και X2 να προέρχονται από την ίδια κατανομή. Κατ' αυτόν τον τρόπο, μικρές τιμές p-value σε πειράματα έκφρασης είναι ισχυρή ένδειξη ότι οι διαφορές που παρατηρούνται στα επίπεδα έκφρασης ενός γονιδίου είναι βιολογικά σημαντικές. Στην Εικόνα 11 μπορούμε να δούμε πώς οι δύο αυτές τιμές μπορούν να αναπαρασταθούν γραφικά για να περιγράψουν τη γενικότερη εικόνα ενός πειράματος διαφορικής έκφρασης που συγκρίνει δύο συνθήκες. Η Εικόνα 11 αναπαριστά ένα “διάγραμμα ηφαιστείου” ή volcano plot όπως είναι γνωστό. Πρόκειται ουσιαστικά για ένα διάγραμμα σκέδασης όπου κάθε σημείο αντιστοιχεί σε ένα

γονίδιο. Οι καρτεσιανές



Εικόνα 11 Αναπαράσταση Volcano Plot για τη συσχέτιση των τιμών log2FC και p – value.

συντεταγμένες του κάθε γονιδίου είναι η τιμή log2FC στον οριζόντιο άξονα και ο αρνητικός δεκαδικός λογάριθμος του p-value στον κάθετο. Η εγγενής, αναμενόμενη τάση που έχουν οι ακραίες τιμές log2FC να αντιστοιχούν σε μικρές τιμές p-value οδηγεί στο χαρακτηριστικό σχήμα που προσομοιάζει έναν κρατήρα ηφαιστείου. Μ' αυτόν τον τρόπο, όσο ψηλότερα στον κάθετο άξονα βρίσκεται ένα σημείο (μεγάλη τιμή αρνητικού λογαρίθμου του p - value) τόσο πιο σημαντική στατιστικά είναι η διαφορική του έκφραση, ενώ όσο πιο μακριά από το σημείο 0 στον οριζόντιο άξονα, τόσο πιο μεγάλη είναι η έντασή της. Πώς μπορούμε να χρησιμοποιήσουμε αυτήν την αναπαράσταση για να εκτιμήσουμε τη διαφορική έκφραση; Στην πράξη, ο προσδιορισμός των γονιδίων με διαφορική έκφραση γίνεται ορίζοντας κάποια όρια τιμών log2FC και p-value. Στη βιβλιογραφία θα συναντήσουμε συχνά την τιμή p-

value<=0.05 ως όριο σημαντικότητας και την αντίστοιχη απόλυτη τιμή |log2FC| > 1.5 ως όριο διαφορικής έκφρασης. Αυτό σημαίνει ότι γονίδια με log2FC > 1.5 ή log2FC και < -1.5 που ταυτόχρονα έχουν p-value<=0.05 προσδιορίζονται ως ενεργοποιημένα και κατεσταλμένα αντίστοιχα. Ονομάζουμε αυτά τα γονίδια διαφορικά εκφραζόμενα (diffentially expressed genes, DEG). Τα συγκεκριμένα όρια είναι αυτά που έχουν χρησιμοποιηθεί στην Εικόνα 11 για να χρωματίσουν με διαφορετικό τρόπο τα ενεργοποιημένα (κόκκινα) από τα κατεσταλμένα (πράσινα) γονίδια. Φυσικά τα όρια αυτά είναι αυθαίρετα και όχι σπάνια, μπορεί κανείς να συναντήσει διαφορετικά (περισσότερο ή λιγότερο αυστηρές τιμές κατωφλίων), η γενική λογική όμως είναι ότι για να προσδιορίσουμε ένα γονίδιο ως διαφορικά εκφραζόμενο χρειάζεται μια επαρκώς μεγάλη (σε απόλυτη τιμή) log2FC και μια αντίστοιχη τιμή p-value όχι μεγαλύτερη από 0.05.

Ακόμα όμως και χαμηλές τιμές p-value θα πρέπει να αντιμετωπίζονται με προσοχή. Στην περίπτωση που υπολογίζουμε μεγάλο αριθμό από p-values (όπως στην περίπτωση ενός πειράματος γονιδιακής έκφρασης) αυτό που κάνουμε είναι να διενεργήσουμε τον ίδιο στατιστικό έλεγχο πολλές φορές, να κάνουμε δηλαδή αυτό που στη στατιστική ονομάζεται “έλεγχος πολλαπλών υποθέσεων”. Αυτό που συμβαίνει σε αυτές τις περιπτώσεις είναι ότι για καθαρά στατιστικούς λόγους κάποιες τιμές p-values, που είναι αρκετά μικρές ώστε να χαρακτηρίσουν στατιστικά σημαντικές αλλαγές στα επίπεδα έκφρασης, μπορούν να έχουν προκύψει τυχαία. Αυτές θα είναι περισσότερες όσο μεγαλύτερος είναι ο αριθμός των πολλαπλών υποθέσεων (Dudoit, Shaffer, & Block, 2003).

3.4.4 Προσδιορισμός διαφορικά εκφρασμένων γονίδιων

Ο υπολογισμός του βαθμού της μεταβολής της έκφρασης του ίδιου γονιδίου μεταξύ δύο διαφορετικών συνθηκών γίνεται με τη χρήση του λογαρίθμου του λόγου των τιμών έκφρασης στη συνθήκη μελέτης (test) προς τη συνθήκη ελέγχου (control). Η απόδοση των συνθηκών γίνεται από τον πειραματιστή και βασίζεται στο βιολογικό ερώτημα. Υπολογίζεται ο λογάριθμος του λόγου τους ως εξής:

$$log_2FC(g) = log2\frac{E(g)_{test}}{E(g)_{control}}$$

Η χρήση του λόγου των τιμών έκφρασης είναι προφανής. Τιμές του λόγου >1 θα είναι ενδεικτικές μεγαλύτερης έκφρασης στη συνθήκη μελέτης και συνεπώς ενεργοποίησης του γονιδίου ενώ τιμές <1 θα είναι ενδεικτικές καταστολής του. Η εφαρμογή του λογαρίθμου γίνεται για δύο λόγους. Αρχικά, για να μειώσει τη διασπορά των τιμών λόγων έκφρασης, με τον ίδιο τρόπο που αναφέρθηκε παραπάνω για τις καθαρές τιμές έκφρασης. Κατά δεύτερο λόγο, για να μετατρέψει το "ουδέτερο" σημείο της ποσότητας από το 1 στο 0. Οι τιμές log2FC είναι θετικές στην περίπτωση της ενεργοποίησης του γονιδίου g, της αύξησης δηλαδή των επιπέδων έκφρασής του στη συνθήκη μελέτης σε σχέση με τη συνθήκη ελέγχου και αρνητικές στην περίπτωση καταστολής. Μηδενικές μεταβολές αντιστοιχίζονται στην τιμή 0. Επιπλέον, η χρήση του δυαδικού λογάριθμου επιτρέπει μια απευθείας ανάγνωση του βαθμού της διαφορικής έκφρασης. Μια τιμή log2FC=1 σημαίνει διπλάσια έκφραση σε σχέση με τη συνθήκη ελέγχου, ενώ μια τιμή log2FC=-1 υποδηλώνει μείωση στο μισό. Η χρήση της λογαριθμικής κλίμακας επιτρέπει γενικότερα ευκολότερη ερμηνεία των αποτελεσμάτων. Στην περίπτωση που η έκφραση του κάθε γονιδίου έχει μετρηθεί περισσότερες από μία φορές σε επαναλήψεις του ίδιου πειράματος, η εξίσωση που περιγράφηκε δεν αλλάζει αλλά στη θέση των E(g)test και E(g)control χρησιμοποιούνται οι αντίστοιχες μέσες τιμές των επαναλήψεων. Η σημασία της πραγματοποίησης επαναλήψεων του ίδιου πειράματος είναι πολύ μεγάλη για λόγους στατιστικής αξιολόγησης. Σημειώνεται εδώ ότι η ζευγαρωτή σχέση συνθήκης μελέτης/ελέγχου δε σημαίνει ότι σε κάθε πείραμα υπάρχει μόνο μια συνθήκη μελέτης. Κρατώντας σταθερή τη συνθήκη ελέγχου κανείς μπορεί να υπολογίσει τη σχετική έκφραση σε μια σειρά από καταστάσεις. Έτσι π.χ. μπορεί κανείς να συγκρίνει παθολογικά δείγματα με δείγματα στα οποία οι ασθενείς υποβάλλονται σε διαφορετικές θεραπείες ή να μελετήσει τη διαφορική έκφραση σε διαφορετικά στάδια μιας διαδικασίας εφ' όσον όλα συγκρίνονται με ένα αρχικό χρονικό σημείο t=0 κλπ.

3.4.5 Μοντελοποίηση και έκφραση διαφορικής έκφρασης γονιδίων

Μετά την κανονικοποίηση των δεδομένων, ακολουθεί η μοντελοποίηση με τη χρήση στατιστικών μοντέλων αρνητικής διωνυμικής κατανομής (Negative Binomial models – NB). Αυτό σημαίνει ότι το μοντέλο δημιουργείται και εφαρμόζεται στα δεδομένα με την παραδοχή ότι αυτά ακολουθούν την αρνητική διωνυμική κατανομή για το γονίδιο g και το δείγμα *i*.

$$Ygi \sim NB(Mipgi, \Phi_g)$$

Το *Mi* αναφέρεται στο μέγεθος της βιβλιοθήκης (συνολικός αριθμός ενδείξεων), Φ_g στη διασπορά και *pgi* στη σχετική περιεκτικότητα (relative abundance) του γονιδίου g στην πειραματική ομάδα *j* όπου ανήκει το δείγμα *i*. Έτσι η NB κατανομή έχει μέσο $\mu_{gi} = Mipgi$ και διασπορά $\mu_{gi}(1+\mu_{gi}\Phi_g)$.

Η NB κατανομή καταπίπτει σε Poisson όταν το $\Phi_g = 0$ και αφού το Φ_g αντιπροσωπεύει τον συντελεστή διακύμανσης της βιολογικής διακύμανσης μεταξύ δειγμάτων, η τεχνική διακύμανση μπορεί να αντιμετωπιστεί ως Poisson. Έτσι, το μοντέλο που εφαρμόζεται στα δεδομένα είναι εύκολο να διαχωρίσει την τεχνική από την βιολογική διακύμανση.

Πριν γίνει η εφαρμογή του μοντέλου στα δεδομένα, θα πρέπει πρώτα να υπολογιστεί ο δείκτης της βιολογικής διακύμανσης (BCV). Έστω *Ygi* ο αριθμός των ενδείξεων του δείγματος i τα οποία στοιχίζονται στο γονίδιο g και πgi το τμήμα όλων των θραυσμάτων cDNA στο δείγμα i που προέρχονται από το γονίδιο g. Έστω G ο συνολικός αριθμός γονιδίων, έτσι ώστε να ισχύει $\Sigma\pi_{giGg=1} = 1$ για κάθε δείγμα και $\sqrt{\Phi_g}$ ο δείκτης διακύμανσης (CV – η τυπική απόκλιση διαιρεμένη με τον μέσο) του π_{gi} για κάθε δείγμα. Ο ολικός αριθμός των στοιχισμένων ενδείξεων της βιβλιοθήκης *i* ορίζεται ως *Ni* και ισχύει:

$$E(Ygi) = \mu_{gi} = Ni\pi_{gi}$$

Με βάση την παραδοχή ότι το Ygi ακολουθεί κατανομή Poisson για επαναλήψεις πειραμάτων αλληλούχισης του ίδιου δείγματος RNA, ισχύει:

$$var(Y_{gi}) = E_{\pi}[var(Y|\pi)] + var_{\pi}[E(Y|\pi)] = \mu_{gi} + \Phi_g\mu_{gi2}$$

Διαιρώντας κατά μέλη με $\mu gi_{2,}$ προκύπτει:

$$CV_2(Ygi) = 1 / \mu_{gi} + \Phi_g$$

Με βάση την προηγούμενη παραδοχή, μπορεί να αποδειχθεί ότι ο συνολικός δείκτης διακύμανσης υψωμένος στο τετράγωνο ισούται με το άθροισμα της τεχνικής και βιολογικής διακύμανσης των ενδείξεων για κάθε δείγμα:

$$TotalCV_2 = TechnicalCV_2 + BiologicalCV_2$$

Ο δείκτης βιολογικής διακύμανσης (Biological CV – BCV) είναι ένα μέτρο της (μη παρατηρούμενης) αληθινής περιεκτικότητας σε γονίδια μεταξύ όλων των δειγμάτων RNA.

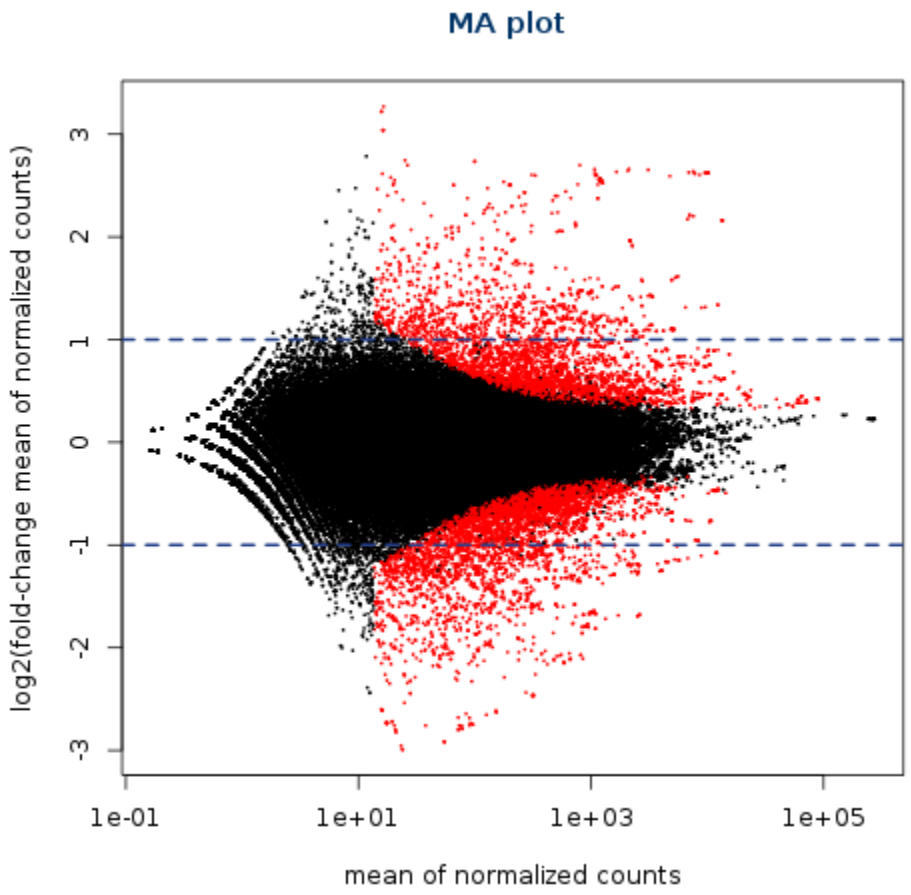
Αντιπροσωπεύει τον δείκτη διακύμανσης μεταξύ όλων των δειγμάτων, αν το βάθος αλληλούχισης έτεινε στο άπειρο.

Μετά τον υπολογισμό του BCV, δημιουργείται κ εφαρμόζεται στα δεδομένα ένα μοντέλο που ανήκει στην οικογένεια των γενικευμένων γραμμικών μοντέλων (GLMs), με την μορφή:

$$\log \mu_{gi} = X_{iT}\beta_g + \log N_i$$

Όπου X_{iT} είναι το διάνυσμα που περιέχει τους πειραματικούς παράγοντες (μετα-δεδομένα πειραματικού σχεδιασμού) για το δείγμα *i* και β_g ένα διάνυσμα που περιέχει τις ανεξάρτητες μεταβλητές της παλινδρόμησης για το γονίδιο *g*.

Τέλος, η οπτικοποίηση των αποτελεσμάτων γίνεται μέσω διαγραμμάτων MA.



Εικόνα 12 Παράδειγμα MA plot. Στον άξονα Y αναπαρίστανται οι λογαριθμημένες τιμές του "Fold change" και στον άξονα X οι μέσοι των κανονικοποιημένων μετρήσεων. Με κόκκινο χρωματίζονται τα γονίδια που παρουσιάζουν στατιστικά σημαντική διαφορά μεταξύ των καταστάσεων που εξετάζονται (πχ μεταξύ δειγμάτων μαρτύρων και ασθενών). Θετικές τιμές παίρνουν τα γονίδια που βρίσκονται να είναι υπερεκφρασμένα στην μία κατηγορία δειγμάτων που χρησιμοποιήθηκε ως

κατάσταση αναφοράς στο γενικευμένο γραμμικό μοντέλο (πχ δείγματα από μάρτυρες) και αρνητικές τιμές παίρνουν τα γονίδια που βρίσκονται να είναι υπερεκφρασμένα στην δεύτερη κατηγορία σύγκρισης (πχ δείγματα ασθενών).

3.5 Ανάλυση Εμπλουτισμού – Λειτουργική Ανάλυση – BioInfoMiner

Για την ανάλυση εμπλουτισμού και την σημασιολογική ανάλυση των αποτελεσμάτων, χρησιμοποιήθηκε η ψηφιακή πλατφόρμα BioInfoMiner (Koutsandreas et al., 2016). Το BioInfoMiner χρησιμοποιεί στατιστικές μεθόδους ανάλυσης και μεθόδους ανάλυσης σχεσιακών δικτύων για να εντοπίσει και να ταξινομήσει, βάσει της σημαντικότητάς τους, τις διαδικασίες που παρουσιάζουν αλλαγές μεταξύ των καταστάσεων που ελέγχονται καθώς και τον γονιδίων που έχουν κομβικό ρόλο στην σύνδεση αυτών των διαδικασιών.

Το BioInfoMiner δέχεται ως είσοδο μία λίστα σημαντικών γονιδίων και επιστρέφει σαν αποτελέσματα λίστες ταξινομημένων γονιδίων και γραφήματα που αφορούν στον εμπλουτισμό των γονιδίων σε συγκεκριμένα μονοπάτια.

Η ψηφιακή πλατφόρμα BioInfoMiner είναι διαθέσιμη στον ιστότοπο https://bioinfominer.com .

- Ιεραρχική Ταξινόμηση Μονοπατιών

  Το Bioinfominer χρησιμοποιεί μεθόδους επαναδειγματοληψίας, αντί για μεθόδους διόρθωσης πολλαπλών ελέγχων, για την απομάκρυνση μη σημαντικών όρων. Έτσι, με την μέθοδο επαναδειγματοληψίας bootstrap από μία κατανομή ομοδοποιημένων όρων οντολογίας βάσει του εμπλουτισμού τους, εντοπίζει και ταξινομεί στατιστικά σημαντικούς εμπλουτισμούς και όχι οντολογικούς όρους. Δεν χρησιμοποιεί παραμετρικούς ελέγχους και δεν κάνει παραδοχές για την κατανομή των μεταβλητών, αντίθετα ελέγχει την παρατηρούμενη κατανομή των εμπλουτισμένων όρων. Για τον λόγο αυτό είναι πρακτικά δυνατό να εφαρμοστεί με επιτυχία σε οποιοδήποτε λεξικό οντολογικών όρων.

- Ιεραρχική Ταξινόμηση Στόχων

  Η αρχική λίστα εισόδου σημαντικών γονιδίων χρησιμοποιείται για την κατάταξή τους σε μονοπάτια βάσει των λεξικών οντολογίας. Ωστόσο, η ύπαρξη μεροληπτικών σφαλμάτων στα λεξικά οντολογίας, δημιουργεί προβλήματα στην αναγνώριση και ταξινόμηση των γονιδίων – στόχων. Το BioInfoMiner χρησιμοποιεί μεθόδους μη

επιβλεπόμενης μάθησης (όπως τα δέντρα αποφάσεων), για να ομαδοποιήσει με στατιστική εγκυρότητα τα γονίδια – στόχους σε κλίκες που βασίζονται στην σημασιολογική τους ερμηνεία. Η μέθοδος αυτή έχει σκοπό την μείωση του θορύβου που προκαλείται λόγω των μεροληπτικών σφαλμάτων στα λεξικά οντολογίας, μέσω της ομαδοποίησης και ταξινόμησης των στόχων με πιθανοθεωρητικές προσεγγίσεις.

- Ιεραρχική Ταξινόμηση Λίστας Γονιδίων

  Η τελική λίστα γονιδίων που επιστρέφει το BioInfoMiner είναι ιεραρχικά ταξινομημένη και παρουσιάζει πολύ χαμηλό λόγο σήματος προς θόρυβο (Signal to Noise – SR ratio). Τα αποτελέσματα οδηγούνται εξολοκλήρου με βάση τα δεδομένα εισόδου (data – driven), ενώ υπάρχει η δυνατότητα επέκτασης της λίστας (scalable) για μεγαλύτερη εμβάθυνση στην βιολογική πληροφορία που φέρει. Επίσης, η λίστα παρουσιάζει σχετικά μικρή διαστασιμότητα (περίπου 2 κλίμακες μικρότερη από την αρχική λίστα γονιδίων) ενώ ταυτόχρονα είναι αρκετά περιεκτική με του σημαντικότερους στόχους που ελέγχουν τα βιολογικά μονοπάτια. Το γεγονός αυτό καθιστά εύκολη την συνθετική ανάλυση γονοτυπικών και φαινοτυπικών χαρακτηριστικών, ενώ ταυτόχρονα μπορεί να διαχειριστεί μεγάλης διαστασιμότητας δεδομένα.

4 Ανάλυση

4.1 Επιλογή του τύπου δεδομένων και της μορφής της ανάλυσης

Η βιοπληροφορική ανάλυση γενωμικών δεδομένων περιλαμβάνει τα επίπεδα της Ποσοτικής Ανάλυσης, της Δομικής ανάλυσης και της Λειτουργικής Ανάλυσης. Η έρευνά μου εστιάζει στην Ποσοτική ανάλυση, η οποία αφορά στον εντοπισμό της διαφορικής έκφρασης γονιδίων μεταξύ δύο καταστάσεων καθώς και στην Λειτουργική Ανάλυση που αφορά στη μελέτη μονοπατιών που επηρεάζονται από υπέρ- και υπό- εκφρασμένα γονίδια. Για τα επίπεδα αυτά της ανάλυσης ακολουθούνται συγκεκριμένα βήματα, με χρήση υπολογιστικών εργαλείων, καθώς και αλγορίθμων και πακέτων της γλώσσας R.

Η Ποσοτική Ανάλυση μεταγραφικών δεδομένων περιλαμβάνει τα εξής βήματα: 1) φιλτράρισμα των μεταγράφων με βάση την ποιότητα των αλληλουχιών, 2) ευθυγράμμιση

των αλληλουχιών σε μεταγράφωμα - αναφοράς, με χρήση και του γονιδιώματος, 3) συγκέντρωση των μεταγράφων και ποσοτικοποίηση της έκφρασης τους, 4) συναρμολόγηση των μεταγράφων για όλα τα δείγματα σε ένα κύριο μετάγραφο, 5) σύγκριση του επιπέδου έκφρασης των γονιδίων μεταξύ δύο ή περισσότερων καταστάσεων αλλά και μεταξύ των διαφορετικών ισομορφών που μπορεί να υπάρχουν μεταξύ των καταστάσεων και 6) λειτουργική ανάλυση που στοχεύει στην εύρεση κυτταρικών μονοπατιών που επηρεάζονται από τις συγκεκριμένες μεταλλαγές και στην ιεράρχηση των κυρίαρχων γονιδίων. Τα πρώτα τέσσερα βήματα λαμβάνουν σαν είσοδο ακατέργαστα δεδομένα (raw data) προερχόμενα από αλληλούχιση RNA και πραγματοποιούνται με την χρήση εργαλείων όπως Trimmomatic, Bowtie2 και το πακέτο εργαλείων Samtools ή και αντίστοιχα εργαλεία, ενώ αποτελούν μέρος της προ-επεξεργασίας (pre-processing) των γενωμικών δεδομένων, πριν την κυρίως στατιστική ανάλυση. Για την πραγματοποίηση της σύγκρισης των επιπέδων έκφρασης των γονιδίων (DEG, Analysis of Differentially Expressed Genes) υπάρχουν διάφορα έτοιμα υπολογιστικά εργαλεία, όμως βασική επιδίωξη αποτελεί η διεκπεραίωση του βήματος της στατιστικής ανάλυσης με χρήση αποκλειστικά πακέτων της R καθώς και μέσω άλλων δυνατοτήτων της γλώσσας, χωρίς να χρησιμοποιηθεί κάποιο έτοιμο εργαλείο ή σουίτα.

Η επιλογή του κλινικού υπόβαθρου των προς μελέτη περιπτώσεων, δηλαδή η επιλογή της προς μελέτη ασθένειας και του αντίστοιχου γενετικού προφίλ των ασθενών, έγινε με βάση προηγούμενα ενδιαφέροντα του εργαστηρίου Μεταβολικής Μηχανικής και Βιοπληροφορικής, του Ινστιτούτου Βιολογίας, Φαρμακευτικής Χημείας και Βιοτεχνολογίας, του Εθνικού Ιδρύματος Ερευνών. Το σετ δεδομένων που επιλέχθηκε να αναλυθεί, περιλαμβάνει δείγματα από ασθενείς με μελάνωμα. Η ιδιαιτερότητα των συγκεκριμένων δεδομένων έγκειται στο ότι η έρευνα δεν αφορά την αναζήτηση διαφορικά εκφρασμένων γονιδίων μεταξύ δειγμάτων ασθενών και δειγμάτων ελέγχου από υγιή άτομα, καθώς δείγματα ελέγχου δεν είναι διαθέσιμα. Πρόκειται για ένα σύνολο 470 περιπτώσεων ασθενών, οι οποίοι πάσχουν από μελάνωμα και ο διαχωρισμός τους σε ομάδες σύγκρισης γίνεται με βάση δημογραφικά και γενετικά χαρακτηριστικά.

Τα πρώτα τέσσερα βήματα της Ποσοτικής Ανάλυσης έχουν πραγματοποιηθεί εκ των προτέρων, ως μια διαδικασία ρουτίνας. Ως εκ τούτου, το αρχικό σετ δεδομένων που χρησιμοποιείται για την έρευνα, δεν αποτελείται από τα άμεσα αποτελέσματα της

αλληλούχισης RNA (raw data), αλλά τα δεδομένα αυτά έχουν υποστεί ήδη την προ-επεξεργασία και η έρευνά μου εστιάζει στην στατιστική και λειτουργική ανάλυσή τους. Το ότι τα δεδομένα για το μελάνωμα δεν ήταν διαθέσιμα στην ακατέργαστη (raw) μορφή τους, δεν αποτέλεσε τροχοπέδη για την συνέχιση της έρευνας, καθώς η προ-επεξεργασία τείνει να καθιερωθεί ως μέρος και φυσική συνέχεια της αλληλούχισης, δίνοντας αξιόπιστα αριθμητικά αποτελέσματα για περαιτέρω ανάλυση.

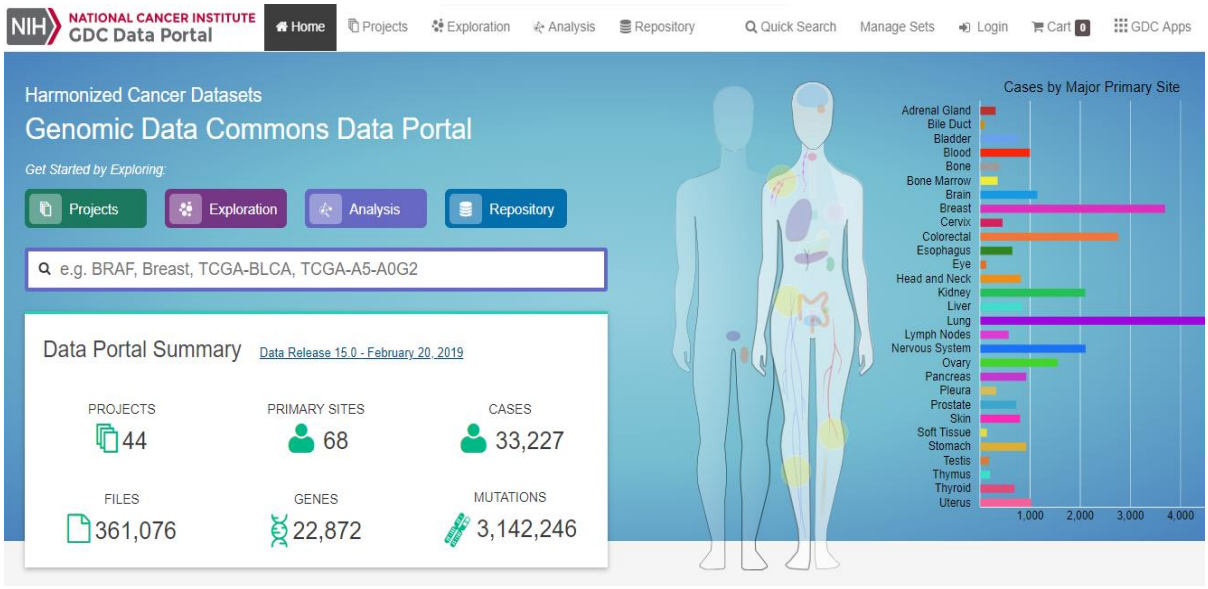
Τα αναμενόμενα αποτελέσματα αφορούν τον εντοπισμό διαφορικής έκφρασης γονιδίων σε υποομάδες των ασθενών με μελάνωμα και την ανάλυση του λειτουργικού ρόλου των γονιδίων αυτών, προκείμενου να εντοπιστούν οι αλλαγές εκείνες στα επίπεδα έκφρασής τους, που δυνητικά οδηγούν σε παθογένεια.

4.2 Εξόρυξη HT-Seq-Counts δεδομένων της GDC

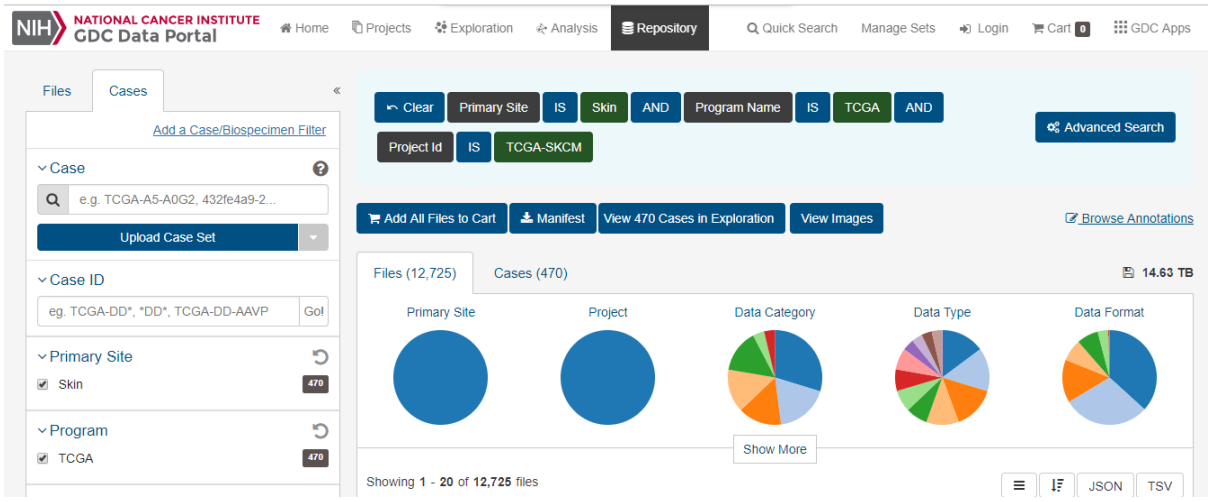
4.2.1 Διερεύνηση των χαρακτηριστικών του σετ δεδομένων μέσω της GDC

Για τις ανάγκες της διπλωματικής εργασίας έγινε εξόρυξη δεδομένων από τη βάση δεδομένων GDC, τα οποία έχουν υποστεί τα βήματα προ-επεξεργασίας που περιγράφηκαν στην ενότητα 3.2 και βρίσκονται στη μορφή των HT-Seq-Counts.

Το γραφικό περιβάλλον και η το αποθετήριο δεδομένων της βάσης GDC, παρέχουν στο χρήστη εύκολη περιήγηση και τη δυνατότητα να λάβει τα δεδομένα στη μορφή αρχείου που επιθυμεί, ανάλογα με τα παρεχόμενα, από τον ισότοπο: https://portal.gdc.cancer.gov/ και κάνοντας τις απαραίτητες επιλογές.



Εικόνα 13 Αρχική σελίδα της βάσης δεδομένων GDC για εξόρυξη δεδομένων σχετικά με τον καρκίνο.



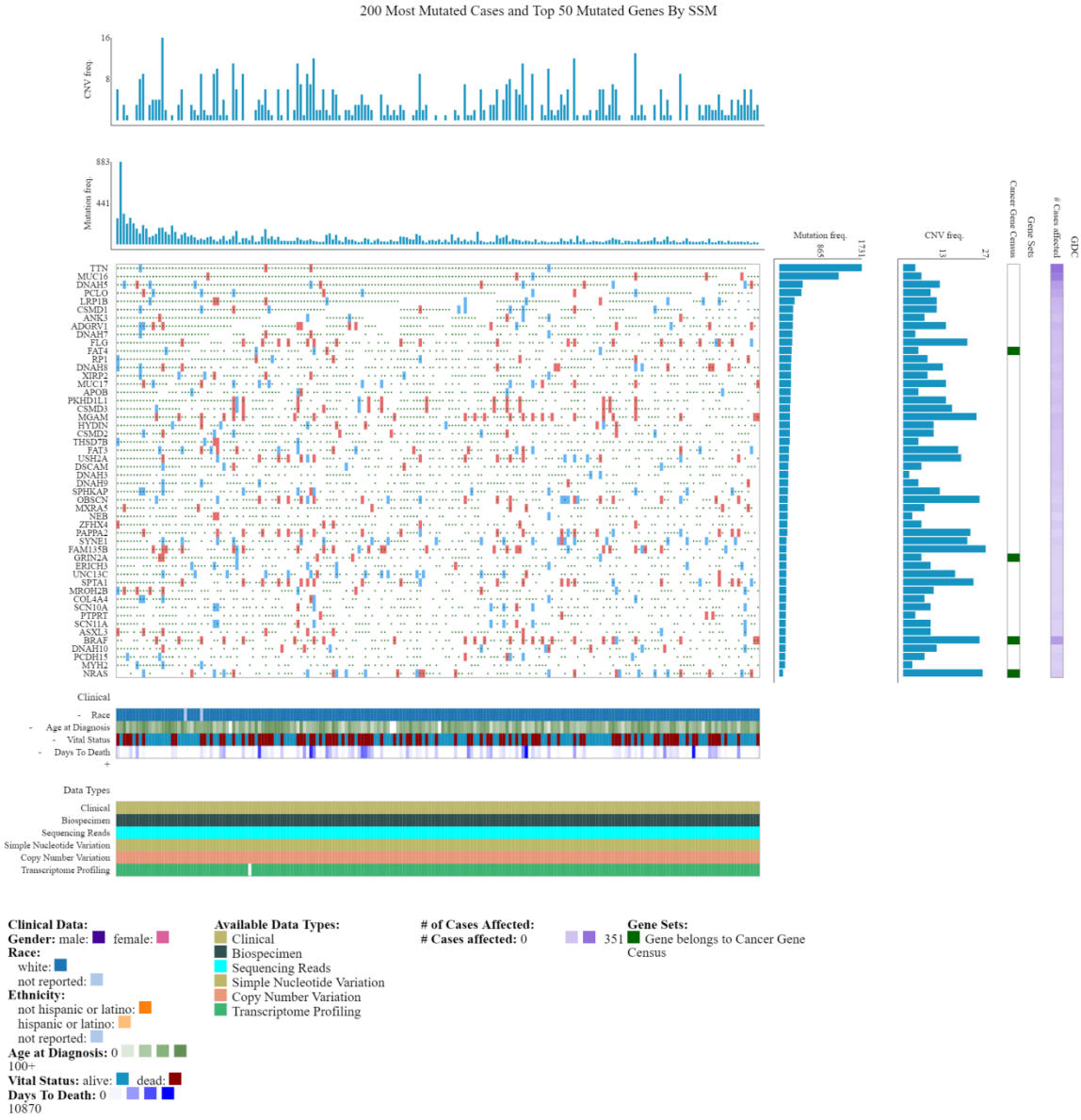
Εικόνα 14 Αποθετήριο της GDC. Οι επιλογές που έχουν γίνει σε ότι αφορά τον τύπο των δεδομένων που επιλέγονται, φαίνονται στο μέσο και πάνω μέρος της σελίδας.

Η πληροφορία για τα χαρακτηριστικά των ασθενών του σετ δεδομένων, αρχικά προέρχεται από την πλοήγηση στη βάση δεδομένων από την οποία εξορύχτηκαν και επιβεβαιώνονται από την διερεύνηση των ιδιοτήτων του σετ μέσω της γλώσσας R και της απλής απεικόνισης του Notepad++ λογισμικού. Η GDC database κάνει διαχωρισμό των δεδομένων της βάση σε “projects”, κατηγορία δεδομένων, τύπο δεδομένων, πειραματικού πρωτοκόλλου και in silico πρωτοκόλλου για προ-επεξεργασία των δεδομένων (αν έχει γίνει). Στην προκειμένη περίπτωση, τα δεδομένα είναι μέρος του εγχειρήματος TCGA-SKCM (Skin Cutaneous Melanoma). Στην πραγματικότητα αντλήθηκαν τα δεδομένα όλων των ασθενών του

συγκεκριμένου πακέτου που υπακούν και στις παρακάτω απαιτήσεις. Σε ότι αφορά την κατηγορία δεδομένων, θα έπρεπε να είναι διαθέσιμο το μεταγραφικό προφίλ των συμμετεχόντων και πιο συγκεκριμένα να παρέχεται η πληροφορία της γονιδιακής έκφρασης. Η πειραματική τεχνική είναι η RNA αλληλούχιση και η προ-επεξεργασία που έχει γίνει έγκειται στο γεγονός ότι δεν αντλούνται τα πρωτογενή δεδομένα της αλληλούχισης, αλλά τα δεδομένα της μέτρησης αναγνώσεων, που έρχεται ως δεύτερο βήμα της διαδικασίας (HTseq-counts).

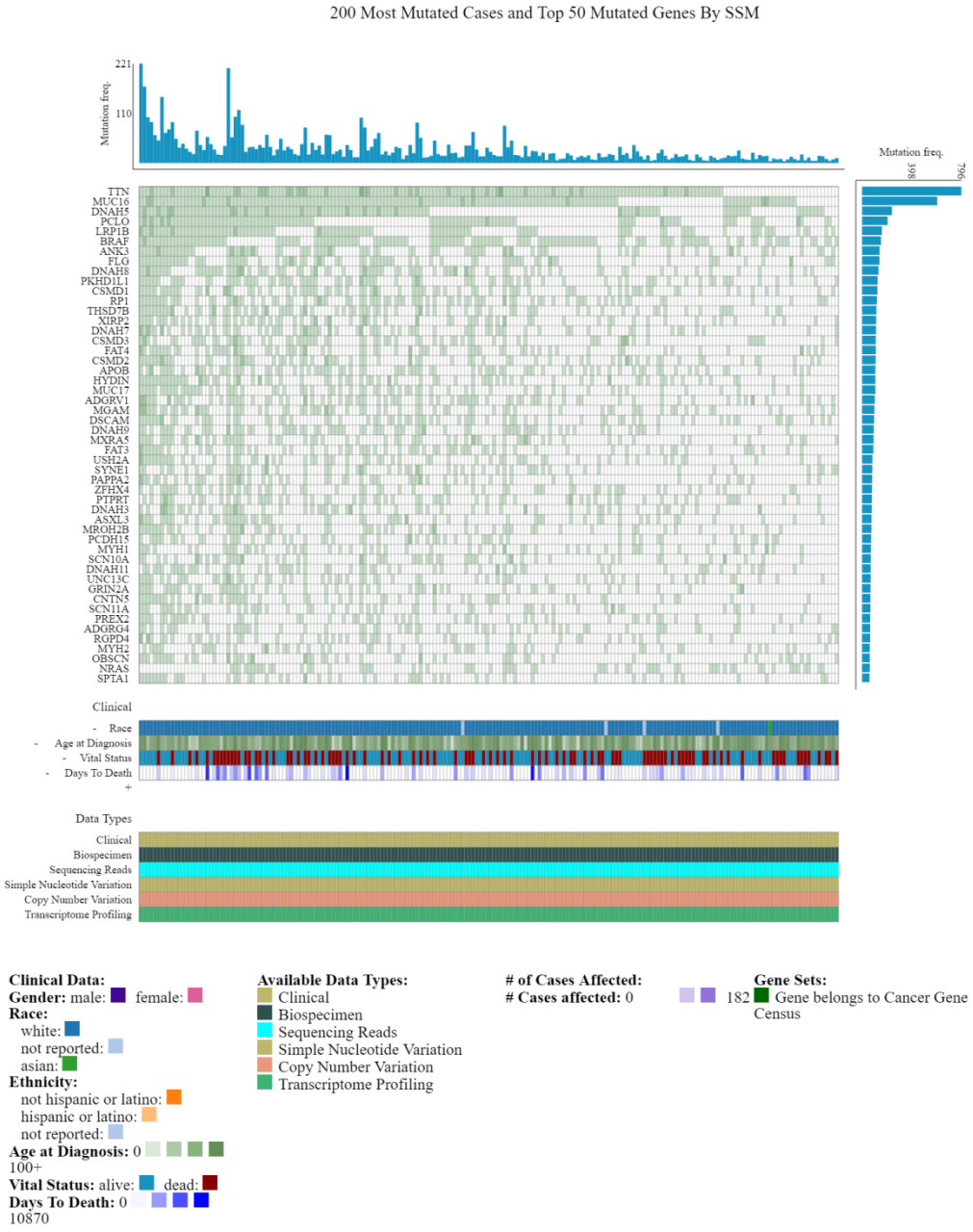
Τα αποτελέσματα αφορούν 470 περιπτώσεις ασθενών μεταξύ των οποίων υπάρχει εκπροσώπηση και των δύο φύλων, περίπου οι μισοί εκ των ασθενών βρίσκονται ακόμη εν ζωή ή βρίσκονταν εν ζωή την τελευταία φορά που ελέγχθηκε η κατάστασή τους. Η μεγάλη πλειονότητάς τους είναι λευκοί, ενώ υπάρχουν λίγες περιπτώσεις (12) ασιατών και 1 στην κατηγορία «Αφροαμερικάνος ή έγχρωμος ». Με βάση την εθνικότητα, οι ασθενείς χωρίζονται σε δύο υποκατηγορίες, «όχι ισπανοί ή λατίνοι», «ισπανοί ή λατίνοι». Το πλήθος που εμπίπτει στη δεύτερη κατηγορία είναι οριακά αμελητέο. Με βάση τα παραπάνω δημογραφικά χαρακτηριστικά πλην του φύλου, δε θα μπορούσε να γίνει κάποιος διαχωρισμός των ασθενών, δεδομένου ότι το μεγαλύτερο πλήθος τους σε κάθε περίπτωση εμπίπτει σε μια από τις κατηγορίες που παρουσιάζονται σε κάθε επίπεδο και τα αποτελέσματα δε θα παρουσίαζαν την αναμενόμενη στατιστική σημαντικότητα.

Παράλληλα, από την GDC παρέχονται διαγράμματα σχετικά με το μεταλλακτικό προφίλ των ασθενών, διαγράμματα σχετικά με την οπτικοποίηση όλων των δημογραφικών χαρακτηριστικών καθώς και των κλινικών χαρακτηριστικών. Για να δοθεί μια γενική εικόνα της ομοιογένειας ή και ετερογένειας που υπάρχει μεταξύ των δειγμάτων ανά κατηγορία, μερικά από τα διαγράμματα αυτά παρατίθενται παρακάτω. Σε αυτό το σημείο θα πρέπει να διευκρινιστεί πως δεν είναι γνωστό με ποια στατιστικά εργαλεία, με ποιες παραδοχές και παραμέτρους και με ποια εργαλεία οπτικοποίησης έχουν προκύψει οι παρακάτω εικόνες και η παρουσίασή τους ως έγκυρης πληροφορίας βασίζεται στην εγκυρότητα της GDC ως βάσης δεδομένων που χαίρει παγκόσμιας αναγνώρισης και χρησιμοποιείται από χιλιάδες χρήστες ανά τον κόσμο.



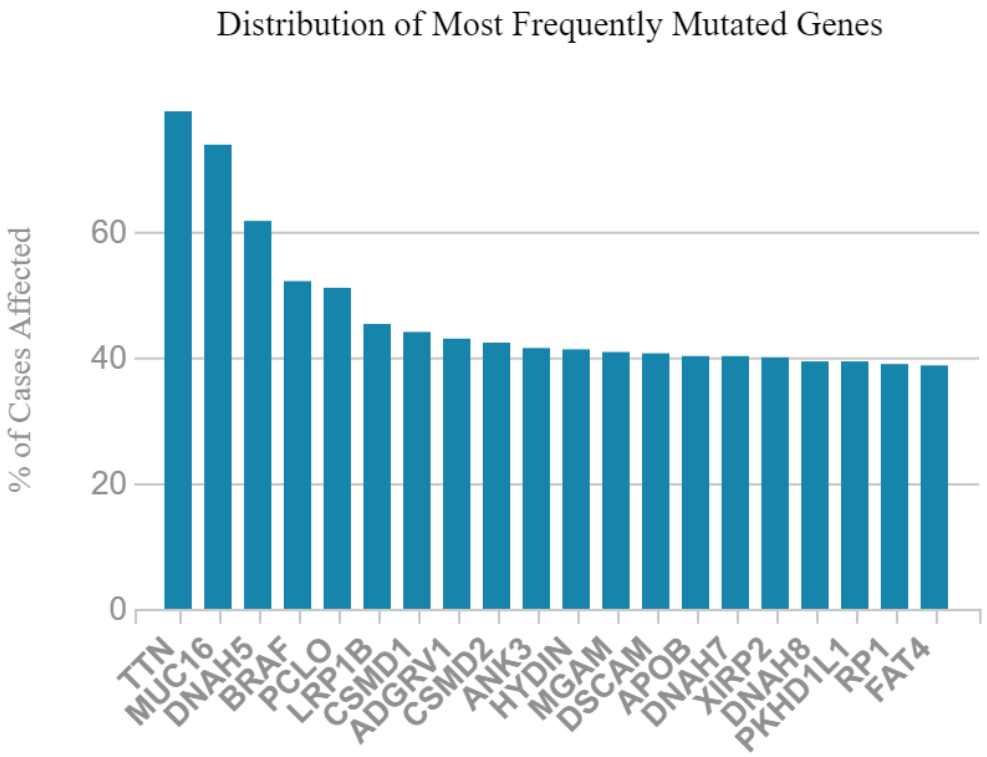
Εικόνα 15 Αναπαράσταση των δεδομένων μεταλλάξεων για το σετ των 470 ασθενών της GDC. Η εξαγόμενη πληροφορία αφορά τα 50 από τα 200 συχνότερα μεταλλαγμένα γονίδια των ασθενών του σετ δεδομένων, τις μεταλλαγές που αυτά φέρουν και τα δημογραφικά χαρακτηριστικά του κάθε ασθενούς (φυλή, ηλικία κατά τη διάγνωση, κατάσταση (εν ζωή ή αποθανών), ημέρες από την ημερομηνία διάγνωσης μέχρι την ημερομηνία θανάτου) που αφορά το αντίστοιχο κελί του OncoGrid. Παράλληλα παρέχεται η πληροφορία των διαθέσιμων τύπων δεδομένων.

Στα παραπάνω δεδομένα επιτρέπεται και απεικόνιση μορφής θερμικού πίνακα (heatmap), για την καθαρότερη απεικόνιση και ποσοτικοποίηση του μεταλλακτικού φορτίου των ασθενών.



Εικόνα 16 Θερμικός χάρτης που αναπαριστά το ποσοστικοποιημένο μεταλλακτικό φορτίο των ασθενών με μελάνωμα.

Στο παρακάτω διάγραμμα γίνεται πιο προφανής η πληροφορία σχετικά με τα πιο συχνά μεταλλασσόμενα γονίδια των ασθενών.

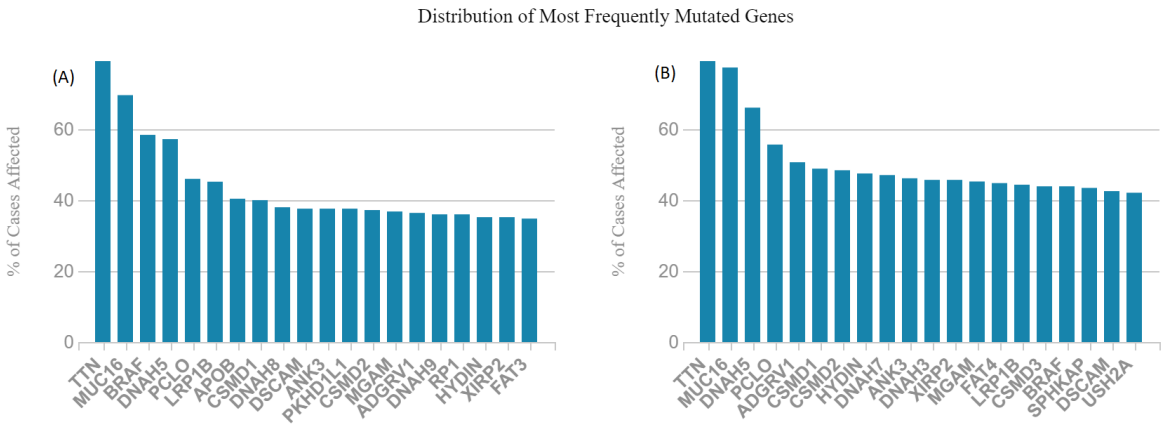


Εικόνα 17 Αριθμός περιπτώσεων (%) που επηρεάζονται από τη μεταλλαγή του γονιδίου που αναφέρεται στον οριζόντιο άξονα. Τα 4 πρώτα γονίδια, TTN, MUC16, DNAH5, BRAF διαφέρουν σημαντικά από τα υπόλοιπα, τα οποία βρίσκονται σε παρόμοια επίπεδα με καθοδική τάση στη συχνότητα εμφάνισης της μεταλλαγμένης του μορφής. Το διάγραμμα αυτό θα μπορούσε να συσχετιστεί με την αυξημένη πιθανότητα τα γονίδια αυτά να έχουν ρόλο στην έναρξη της νόσου. Το BRAF και συγκεκριμένα η μεταλλαγή στη θέση 600 είναι άμεσα συσχετισμένη με το μελάνωμα.

Βιβλιογραφικά, είναι ισχυρές οι ενδείξεις πως από τους ασθενείς με μελάνωμα, εκείνοι που διαθέτουν τη μεταλλαγή BRAF, τείνουν να είναι νεότεροι σε σχέση με αυτούς που δεν φέρουν ή που φέρουν σπανιότερα τη μεταλλαγή. Παράλληλα, μια ηλικία που θα μπορούσε να γίνει ο διαχωρισμός μεταξύ νεότερων και γηραιότερων ασθενών είναι η ηλικία των 60 ετών. Προτείνεται πως η φυσιολογία του μελανώματος και το μεταλλακτικό προφίλ των ασθενών αλλάζει σε εύρος ±10 ετών από αυτή την ηλικία, ενώ πληθυσμιακά, η εμφάνιση μελανώματος κορυφώνεται γύρω από την ηλικία των 50 ετών, με σταδιακά μικρότερα ποσοστά εμφάνισης πριν και μετά από αυτή. Ως εκ τούτου, το μελάνωμα εμφανίζεται

συχνότερα σε άτομα ηλικίας 50-60 ετών και αναμένεται αυτή η αλλαγή να εξηγείται και γενετικά και να αντικατοπτρίζεται στο προφίλ μεταλλάξεων των ασθενών της κάθε κατηγορίας (Akbani et al., 2015).

Πράγματι τα δεδομένα της GDC σχετικά με τη συχνότητα μεταλλαγής συγκεκριμένων γονιδίων σε καθένα από τα εν λόγω ηλικιακά γκρουπ, επιβεβαιώνει την περίπτωση του BRAF, όμως μένει να διερευνηθεί περεταίρω ποια είναι τα διαφορικά εκφρασμένα γονίδια στις δύο ηλικιακές κατηγορίες.



Εικόνα 18 (A) Ασθενείς ηλικίας 0-60 ετών. Τα δύο πρώτα σε συχνότητα εμφάνισης μεταλλαγμένα γονίδια, συμπίπτουν με τα δεδομένα για τον γενικό πληθυσμό των ασθενών, όμως εδώ έρχεται τρίτο σε σειρά το BRAF. (B) Ασθενείς ηλικίας 60-100 ετών. Σε ότι αφορά τα δύο πρώτα πιο συχνά μεταλλασσόμενα γονίδια, το προφίλ των ασθενών συμπίπτει με το γενικό πληθυσμό, όμως εδώ δεν εμφανίζεται το BRAF, παρά στη 17η θέση ανάμεσα στις 20 πιο συχνά εμφανιζόμενες μεταλλαγές, υποδεικνύοντας πως ο αριθμός των ασθενών που φέρουν την εν λόγω μεταλλαγή είναι σημαντικά μικρότερος σε σχέση με το ηλικιακό γκρουπ 0-60. Στην πραγματικότητα, αυτός ο αριθμός είναι ακόμα μικρότερος, δεδομένων των hot-spot mutations του BRAF και όχι τις μεταλλαγές σε όλες τις δυνατές θέσεις.

4.2.2 Εξόρυξη πρωτογενών μετρήσεων αναγνώσεων (raw read counts)

Η εξόρυξη των δεδομένων έγινε, τελικά, με χρήση του πακέτου TCGAbiolinks της γλώσσας προγραμματισμού R. Το εν λόγω πακέτο έχει τη δυνατότητα να εξάγει δεδομένα από τη βάση Genomic Data Commons (GDC) του National Cancer Institute (NCI) των ΗΠΑ, εκμεταλλευόμενο την πρόσβαση στο Application Programming Interface (API) της GDC, για την αναζήτηση, εξόρυξη και προετοιμασία των GDC δεδομένων για περεταίρω ανάλυση στην R (Colaprico et al., 2016).

TCGAbiolinks, Query, Download, Prepare

GDCQuery() function

Ερώτημα (Query) στην GDC μέσω των συναρτήσεων του πακέτου TCGAbiolinks.

>query<-GDCquery (project = "TCGA-SKCM", data.category = "Transcriptome Profiling", data.type= "Gene Expression Quantification", workflow.type= "HTSeq - Counts")

GDCdownload() function

>GDCdowload(query, method="api")

Δομή φακέλων δεδομένων της GDC:

Documents> GDCdata> TCGA_SKCM> harmonized> Transcripotome_Profiling> Gene_Expression_Quantification> one file per folder> GZ file (compressed)> εφαρμογή 7.zip> file_name.htseq.counts (472 files)

Μαζική αποσυμπίεση των αρχείων, μέσω των λειτουργιών των Windows, ώστε να προκύψει φάκελος με τα 472 .htseq.counts αρχεία.

*Gene_Expression_Quantification> search bar> *.gz (select every folder that ends in ".gz")> unzip folders> search bar> *.counts (select every file that ends in ".counts")> paste in GDCdataArxiko(folder) (delete all previous versions)*

GDCprepare function ()

>GDCprepare(query)

Total Output: Documents>Human_genes_GRCh38_p12_.rda (RDA file, 914 MB)

4.3 Εξόρυξη του προφίλ μεταλλαγών - cBioportal

Από τη βάση δεδομένων cBioportal και χρησιμοποιώντας τις επιλογές που δίνει ο αντίστοιχος ιστότοπος, αντλήθηκαν δεδομένα για τις μεταλλαγές που φέρουν οι ασθενείς ενδιαφέροντος. Με την είσοδο στη βάση επιλέγεται η καρτέλα "Dowload Data" και έπειτα το εγχείρημα (project) που μπορεί να συνοψίζει καλύτερα το σύνολο ασθενών της μελέτης ενδιαφέροντος. Σχετικά με το μελάνωμα υπάρχουν δύο μεγάλα σετ δεδομένων το "Skin Cutaneous Melanoma (TCGAPanCancer Atlas)" και το "Skin Cutaneous Melanoma (Provisional)". Και οι δύο περιπτώσεις ελέγχθηκαν προκειμένου να βρεθεί πιο από τα δύο

παρουσιάζει τη μεγαλύτερη επικάλυψη σε επίπεδο ταυτοποίησης των ασθενών (TCGA submitter_ids) και τελικά επιλέχθηκε το πρώτο. Το επόμενο βήμα της διαδικασίας είναι η εισαγωγή των γονιδίων ενδιαφέροντος στο πεδίο “Enter Genes”. Τα γονίδια που επιλέχθηκαν είναι τα: PTK2B, CTNNB1, NOTCH1, LRRK2, DMD, BRAF, RELN, ATM, PDPK1, EPHA2, ZC3H12A, ANGPT1, TP53, HSF1, NR1H4, KDR, CLU, CDKN1B, TLR4, HNF1A, CASP8, GSN, ROCK1, ANK3, TERT, DCN, PPP1R9A, AKAP6, ROBO2, PKP2, NRAS, KRAS, HRAS, NF1. Η επιλογή των παραπάνω βασίζεται σε προηγούμενη δουλειά μελών του εργαστηρίου Μεταβολικής Μηχανικής και Βιοπληροφορικής, του Ινστιτούτου Βιολογίας, Φαρμακευτικής Χημείας και Βιοτεχνολογίας, του Εθνικού Ιδρύματος Ερευνών και πιο συγκεκριμένα στη έρευνα των Kontogianni et.al (Kontogianni, Piroti, Maglogiannis, Chatziioannou, & Papadodima, 2018), η οποία αναλύθηκε εκτενώς στην ενότητα 2, δεδομένου ότι αποτελεί βασικό μέρος της βιβλιογραφικής βάσης της ανάλυσης.

Το εξαγώμενο προϊόν που προκύπτει είναι ένα πίνακας με τα 34 γονίδια επιλογής σε στήλες και τους κωδικούς ταυτοποίησης 415 ασθενών σε γραμμές (TCGA submitter_id code) οι οποίοι διαιρούνται σε κατηγορίες με βάση το μεταλλακτικό τους προφίλ και πιο συγκεκριμένα με βάση το αν εμφανίζουν μεταλλαγή σε κάποιο από τα γονίδια BRAF, την οικογένεια γονιδίων RAS, NRAS, HRAS και KRAS και το NF1 ή την ετερογενή ομάδα Triple-WT, η οποία δεν παρουσιάζει καμιά από τις επόμενες μεταλλάξεις. Οι ασθενείς είναι λιγότεροι από το αρχικό σετ δεδομένων των 470 περιπτώσεων διότι δεν υπάρχουν δεδομένα μεταλλάξεων για όλους τους ασθενείς.

Αναλυτικά για την κάθε ομάδα:

- **BRAF ασθενείς** (μεταλλαγή V600+). Το γονιδίο BRAF παρουσιάζει ένα εύρος μεταλλαγών σε διάφορες θέσεις της έκτασής του. Σημαντικές θέσεις μεταλλαγών για το μελάνωμα, αναφερόμενες ως hot-spot mutations για την ασθενεία, δεν είναι όλο το εύρος των μεταλλαγών που μπορούν να εμαφανιστούν, αλλά το υποσύνολο των μεταλλαγών στη θέση V600+, με το “+” να συμβολίζει το τοιδήποτε μπορεί να ακουλουθήσει, δηλαδή τη μετάπτωση της βαλίνης στη θέση 600 σε οποιοδήποτε άλλο αμινοξύ, συχνά γλουταμικό οξύ, E, ή λυσίνη, K σε ότι αφορά το επίπεδο του εκφρασνμένου mRNA μορίου στην αντιστοιχη πρωτεΐνη.

- **RAS ασθενείς**: σε αυτή την κατηγορία ενάτσσονται ασθενείς που φέρουν μεταλλαγή σε τουλάχιστον 1 από τα γονίδια μεταλλαγή NRAS, HRAS, KRAS. Σε πολύ μεγαλυτερη συχνότητα σε σχέση με τα υπόλοιπα, εμφανίζονται μεταλλαγές στο γονίδο NRAS, από αυτή την καητγορία.
- **NF1 ασθενείς**: ασθενείς που παρουσιάζουν μεταλλαγή στο γονίδιο NF1
- **Triple_wt ασθενείς**: ομάδα ασθενών που παρουσιάζει μεταλλαγές σε οποιοδήποτε από τα γονίδια της αρχικής λίστας, χωρίς όμως να εμπίπτει σε καμιά από τις παραπάνω διακριτές κατηγορίες.

Προκειμένου να μετρηθεί ο αριθμός των δειγμάτων που ανταποκρίνονται σε κάθε μια από τις παραπάνω κατηγορίες, χρησιμοποιήθηκε η γλώσσα προγραμματισμού R και ενδεικτικά τα πακέτα edgeR (Robinson et al., 2010) για την εισαγωγή των αρχικό δεδομένων στο περιβάλλον της γλώσσας και τα πακέτα dplyr (Hadley Wickham, Romain François, 2018), tidyr (Henry, 2018) για την μορφοποίηση των πινάκων σε κατάλληλη μορφή ώστε να είναι εύκολα διαχειρίσιμοι και να μπορούν να συζευχθούν με τα προϋπάρχοντα δεδομένα της GDC.

Αρχικά, με την εντολή read.delim(), το αρχείο .txt, το οποίο είναι το εξαγόμενο από το cBioportal, εισάγεται στο περιβάλλον της R, ως αντικείμενο. Με τις κατάλληλες τροποποιήσεις, αφαιρούνται οι γραμμές με ελλείψεις (missing values) και διαμορφώνεται ένας πίνακας με ονόματα στηλών τα ονόματα των γονιδίων ενδιαφέροντος (κοινή ονομασία) και ονόματα γραμμών τους TCGA κωδικούς (submitter ids) των ασθενών. Δείγμα του πίνακα παρουσιάζεται παρακάτω. Η πληροφορία μέσα στα κελιά, αναπαριστά είτε την απουσία μεταλλαγής στο συγκεκριμένο γονίδιο («NaN» value) για τον αντίστοιχο ασθενή, είτε την παρουσία μεταλλαγής και συγκεκριμένα τη θέση της μεταλλαγής με τον κωδικό του ενός γράμματος για τα αμινοξέα στο μεταφραστικό προϊόν.

Πίνακας 1 Παράδειγμα πίνακα με τα δεδομένα μεταλλαγών. Πίνακας παρόμοιας μορφής προκύπτει από τον ιστότοπο cBioportal και έρχεται σε αυτή τη μορφή μετά την εφαρμογή πακέτων της γλώσσας R για οργάνωση και μορφοποίηση δεδομένων. Περιλαμβάνονται τα κοινά ονόματα των γονιδίων σε στήλες και το στοιχείο ταυτοποίησης του κάθε ασθενούς σε γραμμή. Η πληροφορία στο κάθε κελί δηλώνει αν ο ασθενής τη κάθε γραμμής παρουσιάζει μεταλλαγή στο αντίστοιχο γονίδιο της στήλης. Αν ναι, εμφανίζεται ο τύπος της μεταλλαγής. Αν όχι, η τιμή "NaN".

Submitter_ids/Genes	AKAP6	ANGPT1	ANK3	ATM	BRAF
TCGA-3N-A9WB	NaN	NaN	NaN	NaN	H725Y

TCGA-3N-A9WC	P1839L	NaN	E2513K,P2141S,P1892L	NaN	H574Y
TCGA-3N-A9WD	NaN	R391K	NaN	NaN	NaN
TCGA-BF-A1PU	NaN	NaN	E154K	NaN	V600E
TCGA-BF-A1PV	S81L	NaN	NaN	NaN	NaN
TCGA-BF-A1PX	P1244H	NaN	NaN	NaN	V600E
TCGA-BF-A1PZ	NaN	NaN	NaN	NaN	NaN
TCGA-BF-A1Q0	Q1427K	NaN	NaN	W3052L	NaN

Μελετώντας τον εξαγόμενο πίνακα προκύπτουν τα εξής δεδομένα για μη αποκλειστικές μεταλλαγές, δηλαδή για μεταλλαγές τις οποίες μπορεί να φέρει το γονίδιο ενδιαφέροντος της κάθε στήλης ανεξάρτητα από το αν φέρει παράλληλα μεταλλαγή και σε κάποιο από τα υπόλοιπα γονίδια.

Ως μεταλλαγμένο BRAF γονίδιο λογίζεται εκείνο που περιλαμβάνει απαραίτητα μεταλλαγή στη θέση 600. Η μεταλλαγή αυτή εμφανίζεται σε 8 επίπεδα της μεταβλητής BRAF, δηλαδή μπορεί να συνυπάρχει με άλλες περιπτώσεις μεταλλαγών είτε να αφορά μετάπτωση προς διαφορετικό κάθε φορά αμινοξύ. Οπότε, όλες αυτές οι περιπτώσεις συνυπολογίζονται και είναι προσμετρήσιμες για να δοθεί το τελικό σύνολο των ασθενών που φέρουν μεταλλαγή του BRAF σχετική με το μελάνωμα.

Σε ότι αφορά τη μεταλλαγή RAS, θεωρείται ότι εμφανίζεται σε κάθε περίπτωση όπου οποιοδήποτε από τα 3 RAS γονίδια που έχουν συμπεριληφθεί στο σετ δεδομένων παρουσιάζει κάποια μεταλλαγή.

Αντιστοίχως για τη μεταλλαγή στο NF1 γονίδιο.

Σε ότι αφορά την κατηγορία ασθενών Triple_WT, είναι μια ετερογενής ομάδα που περιλαμβάνει όλες τις περιπτώσεις ασθενών που δεν εμφανίζουν καμιά από τις παραπάνω μεταλλαγές, δηλαδή ασθενείς χωρίς μεταλλαγές στα γονίδια N/K/HRAS, NF1 και χωρίς hot-spot μεταλλαγές του γονιδίου BRAF, οπότε μεταλλαγές εκτός της θέσης 600 ή κελιά με τιμή “NaN”.

Συνοπτικά προκύπτουν τα παρακάτω δεδομένα:

Πίνακας 2 Αριθμητικά δεδομένα που προκύπτουν μετά τη διαίρεση των ασθενών σε υπο-ομάδες ανάλογα με την παρουσία ή μη μεταλλαγών σε συγκεκριμένα γονίδια.

mutated genes	BRAF(V600+)	RAS(N/F/H-RAS)	NF1	Triple_WT (none of the others)
Number of cases	198	140	79	41
percentage	47.71084337	33.73493976	19.03614	9.879518072

Βιβλιογραφικά το BRAF και τα γονίδια της οικογένειας RAS, με μεγαλύτερες πιθανότητες για το NRAS είναι αλληλοαποκλειόμενα (Akbani et al., 2015). Σε μια μειονότητα των περιπτώσεων εμφανίζονται ασθενείς που να φέρουν ταυτόχρονα μεταλλαγή στη θέση 600 του BRAF και στο NRAS, αλλά και γενικότερα σε γονίδια της οικογένειας RAS. Αυτός ο αμοιβαίος αποκλεισμός δεν αποτελεί ανεξήγητο φαινόμενα, καθώς τόσο το BRAF όσο και τα RAS αποτελούν και τα δύο συστατικά του μονοπατιού MAPK (ενότητα 2.1).

Ο αριθμός ασθενών που προκύπτει με μεταλλαγή στη θέση 600 του BRAF και ταυτόχρονη μεταλλαγή σε οποιοδήποτε από τα γονίδια RAS, είναι 7 ασθενείς. Αν αναπροσαρμοστεί η εντολή για να συμπεριλάβει μόνο το NRAS γονίδιο, ο αριθμός αυτός γίνεται 4. Και οι δύο πληθυσμοί αποτελούν μειοψηφίες σε σχέση με το μέγεθος των 415 ασθενών με δεδομένα μεταλλαγών, άρα τα εν λόγω δεδομένα συμπίπτουν στο μεγαλύτερο δυνατό βαθμό με τις βιβλιογραφικές αναφορές.

Για λόγους εγκυρότητας αναφέρεται πως υπάρχουν συνολικά 15 ασθενείς που φέρουν μεταλλαγμένο οποιοδήποτε RAS γονίδιο και παράλληλα μεταλλαγμένο BRAF, σε οποιοδήποτε σημείο και όχι μόνο στη θέση 600 που εξετάζεται παραδοσιακά για το μελάνωμα. Δεδομένου ότι η μεταλλαγή του BRAF εκτός της θέσης 600 θεωρείται πως δεν σχετίζεται με την ασθένεια, δε λαμβάνεται υπόψη ως μεταλλαγή.

Στον παρακάτω πίνακα γίνεται παρουσίαση των πιο εξειδικευμένων ομάδων στις οποίες θα μπορούσαν να χωριστούν οι ασθενείς με βάση το μεταλλακτικό τους προφίλ, ώστε έπειτα οι ομάδες αυτές να χρησιμοποιηθούν σε δυαδικές συγκρίσεις και να ελεγχθούν για την παρουσία διαφορικά εκφρασμένων γονιδίων.

Πίνακας 3 Συνοπτικός πίνακας των αποτελεσμάτων της μέτρησης των ασθενών που ανήκουν σε κάθε υπο-ομάδα ανάλογα με την παρουσία ή μη μεταλλαγής σε κάποιο από τα γονίδια BRAF, N/K/H-RAS, NF1, συμπεριλαμβανομένων των δεδομένων για μορφές αποκλεισμού.

	BRAF (V600+)	RAS (N/F/H-RAS)	NF1	Triple-WT (none of the others)	BRAF-nonRAS	RAS-nonBRAF	nonBRAF
Number of patients	198	140	79	41	191	133	217
percentage	47.71084337	33.73493976	19.03614	9.879518072	46.02409639	32.04819277	52.289157

4.4 Ανάλυση Διαφορικής Έκφρασης ανά παράμετρο αναφοράς

Τα βήματα που θα περιγραφούν παρακάτω αποτελούν επαναλαμβανόμενη διεργασία για όλες τις ομάδες συγκρίσεων που πραγματοποιήθηκαν στη διπλωματική εργασία και οδήγησαν στην εξαγωγή αποτελεσμάτων και στη συνέχεια συμπερασμάτων σχετικά με τη μελέτη της διαφορικής έκφρασης γονιδίων σε ασθενείς που πάσχουν από δερματικό μελάνωμα.

4.4.1 Δημιουργία αντικειμένου GDEList της R – πακέτο edgeR

Αρχικά χρησιμοποιείται το πακέτο edgeR για τη δημιουργία DGEList αντικειμένου της R, το οποίο περιλαμβάνει τρεις διακριτούς πίνακες δεδομένων. Ο πρώτος περιέχει τα στοιχεία ταυτοποίησης του κάθε ασθενούς, το μέγεθος των βιβλιοθηκών αναγνώσεων για το κάθε δείγμα/ασθενή, μια λίστα με συντελεστές κανονικοποίησης, οι οποίοι είναι όλοι αρχικά 1 και μια στήλη “group” η οποία τελικά, ακριβώς πριν την αρχή της διαδικασίας ανάλυσης της γονιδιακής έκφρασης, θα περιέχει τα επίπεδα με βάση τα οποία διαφοροποιούνται οι ασθενείς που εξετάζονται για διαφορική έκφραση γονιδίων.

Πίνακας 4 Παράδειγμα της δομής του πίνακα των δειγματών της δομής DGEList Obgect

submitter_ids	group	lib.size	norm.factors
TCGA-FS-A1YW	1	101768225	1
TCGA-EE-A2M7	1	93973154	1
TCGA-D9-A1JW	1	88110007	1
TCGA-ER-A3ET	1	53452503	1
TCGA-GN-A267	1	71666436	1

Ο δεύτερος πίνακας περιέχει στοιχεία σχετικά με τις μετρήσεις των αναγνώσεων για καθένα από τα δείγματα, δηλαδή το πόσες φορές έχει τραυτοποιηθεί το κάθε γονίδιο σε κάθε δείγμα. Στις γραμμές του πίνακα αναφέρονται τα γονίδια με τον κωδικό της ENSEMBL και στης στήλες τα στοιχεία ταυτοποίσησης των ασθενών, εδώ το όνομα του αρχείου από το οποίο προήλθαν τα δεδομένα για την των htseq-counts για τον κάθε ασθενή. Άρα οι πληροφορία στο κάθε κελί αναφέρει των αριθμό των κανονικοποιημένων counts για το κάθε γονίδιο, στον κάθε ασθενή. Εννοείται πως υπάρχουν γονίδια τα οποία δεν εκφράζονται σε κανένα από τα δείγματα και τα οποία αφαιρούνται σε επόμενο βήμα, αλλά μεταξύ των αρχικών βημ'άτων της αναλύσης. Είναι ένας ιδιαίτερα εκτενής πίνακας που δεν είναι εύκολο να δοθεί παραδειγματικά, παρόλα αυτά, ένα εξαιρετικά απλό πρότυπο της δομής του δίνεται παρακάτω.

Πίνακας 5Παράδειγμα δομής πίνακα μετρήσεων του αντικειμένου DGEList Object της R.

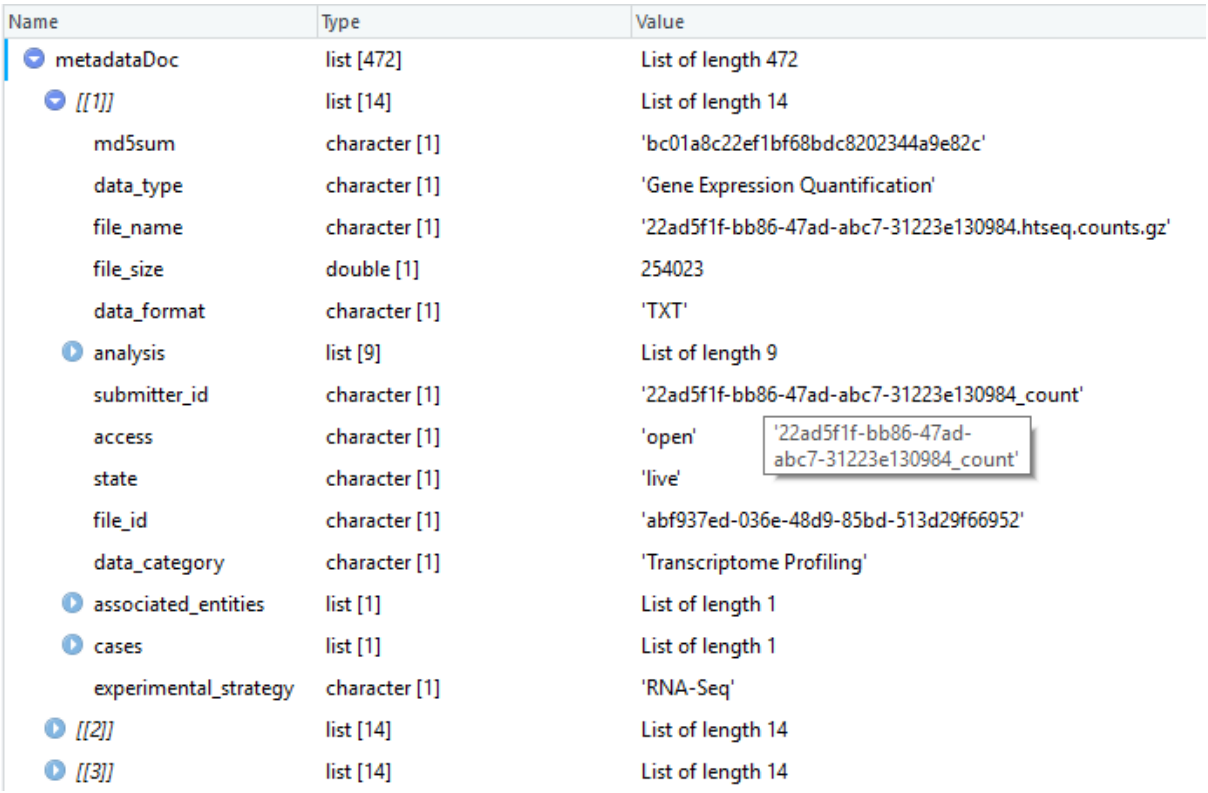
	TCGA-FS-A1YW	TCGA-EE-A2M7	TCGA-D9-A1JW	TCGA-ER-A3ET	TCGA-GN-A267
ENSG…1	123	124	125	126	127
ENSG…2	0	0	0	0	0
ENSG…3	78	2000	4	100	5

Ο τρίτος και τελευταίος πίνακας δημιουργείται μετά τη χρήση του πακέτου Homo.sapiens της R, το οποίο είναι υπεύθυνο για το σχολιασμό των γονιδίων και προσθέτει ένα πίνακα με το ENSEMBL ID των γονιδίων, την κοινή ονομασία και το χρωμόσωμα στο οποίο βρίσκεται το εκάστοτε γονίδιο (Team, 2015).

4.4.2 Εξόρυξη δεδομένων σχετικά με την παράμετρο βάσει της οποίας θα γίνει η ανάλυση διαφορικής έκφρασης γονιδίων.

Τα δεδομένα σχετικά με το προφίλ μεταλλάξεων αντλούνται από τη βάση δεδομένων cBioportal όπως περιεγράφηκε στην ενότητα 3.3, ενώ τα δημογραφικά δεδομένα βάση των οποίων έγιναν αναλύσεις, ηλικία και φύλο, αντλήθηκαν από το αρχείο metadata***.json το οποίο παρέχεται από τη βάση GDC και διαμορφώνεται ανάλογα με τα αρχεία που έχουν επιλεγεί για λήψη και περεταίρω επεξεργασία.

Το αρχείο των metadata αποτελεί ουσιαστικά μια λίστα από λίστες, όπου το πρώτο επίπεδο αποτελεί τον αύξων αριθμό του δείγματος και παράλληλα είναι μια λίστα που χωρίζεται ξανά σε πολλαπλά επίπεδα λιστών, ανάλογα με την εσωτερική κατηγοριοποίηση των δεδομένων για τον εκάστοτε ασθενή. Παρακάτω ακολουθεί μια σύντομη απεικόνιση του αρχείου .json για να δοθεί περιληπτικά ο βαθμός πολυπλοκότητας του αρχείου, του οποίου τα διάφορα επίπεδα είναι προσβάσιμα μέσω δόμησης κατάλληλων εντολών της γλώσσας R. Καθώς ο χρήστης μεταβαίνει όλο και σε μεγαλύτερο βάθος στα δεδομένα του αρχείου, καταλήγει σε ένα τελικό σύνολο από πάνω από 20 διαφορετικές λίστες για τον κάθε ασθενή.



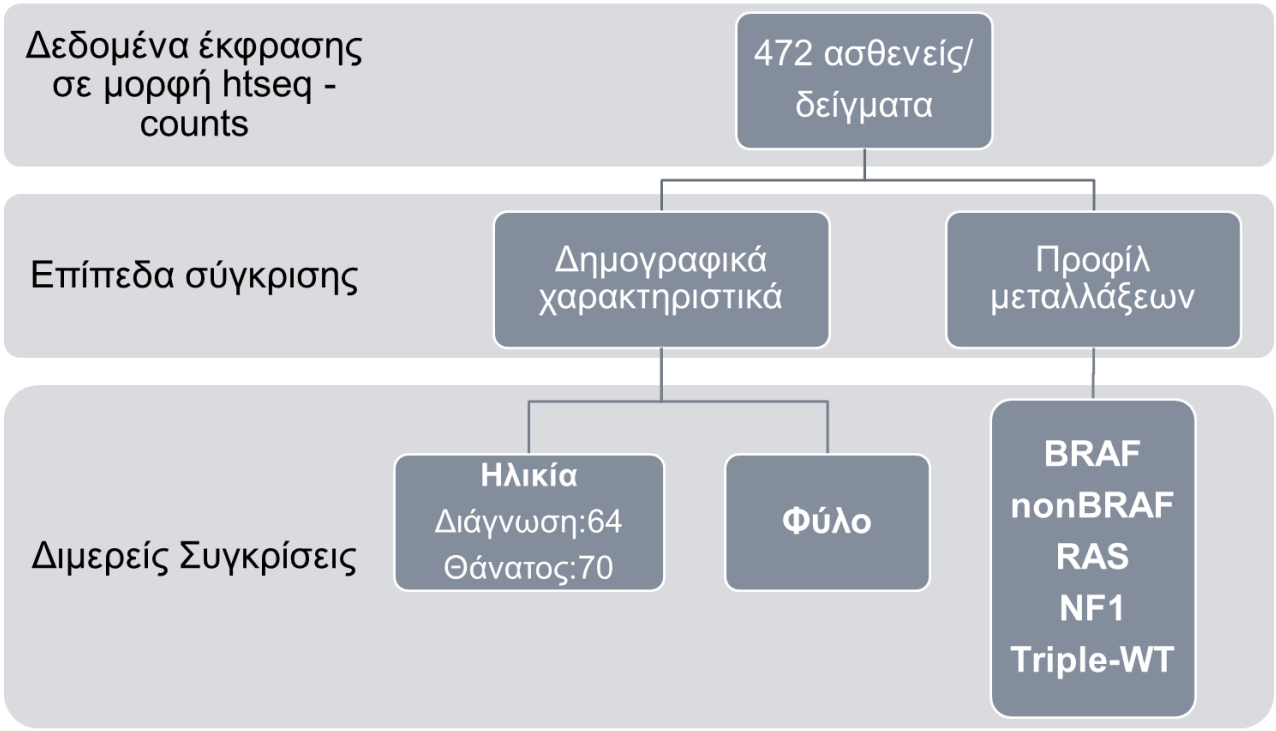
Εικόνα 19 Απεικόνιση της δομής του αρχείου metadata***.json, το οποίο περιέχει τις συνολικές πληροφορίες για τα δημογραφικά, γενετικά και κλινικά χαρακτηριστικά των ασθενών που περιλαμβάνονται στο σετ δεδομένων, δεδομένα για συγκεκριμένες περιβαλλοντικές εκθέσεις των ασθενών, πειραματικά δεδομένα της μελέτης από την οποία έχουν προέλθει τα δεδομένα αλληλούχισης, τα ευρύτερα χαρακτηριστικά της μελέτης της GDC που υπάγεται ο εν λόγω ασθενής κ.α.

Η πρόσβαση στο αρχείο εκτελέστηκε με εντολή στη γλώσσα R που με μικρές τροποποιήσεις είναι η παρακάτω για την εξόρυξη της πληροφορίας της ηλικίας των ασθενών και έχει πανομοιότυπη δομή για την εξόρυξη οποιουδήποτε από τα δεδομένα που υπάρχουν διαθέσιμα στο project της GDC.

age_at_diagnosis<-sapply(metadataDoc, function(x) x[[13]][[1]][[1]][[1]]$age_at_diagnosis)

Από την παραπάνω δομή σημαντικό είναι στο στοιχείο *x[[13]][[1]][[1]][[1]]$age_at_diagnosis* που ερμηνεύεται ως εξής: από τον ασθενή με τον εκάστοτε αύξων αριθμό από το 1 έως και το τέλος της δομής metadataDoc που περιέχει το αρχείο metadata***.json, γίνεται μετάβαση κατά σειρά στις λίστες 13, 1, 1, 1, η κάθε μια από τις οποίες αποτελεί εσωτερική λίστα του ακριβώς ανώτερου επιπέδου, ενώ από την τελευταία στη σειρά λίστα 1, αντλείται το στοιχείο *age_at_diagnosis*, ηλικία κατά την οποία έγινε διάγνωση της νόσου.

Εκτός από την ηλικία των ασθενών, άλλες παράμετροι που ενσωματώθηκαν στο αντικείμενο της R με το οποίο εργάστηκα είναι το φύλο των ασθενών και το προφίλ μεταλλάξεων με βάση το οποίο οι ασθενείς χωρίζονται στις κατηγορίες BRAF, non – BRAF, RAS, NF1 και Triple – WT. Από τα δεδομένα αυτά και με βάση αυτές τις παραμέτρους, πραγματοποιούνται διμερείς συγκρίσεις.

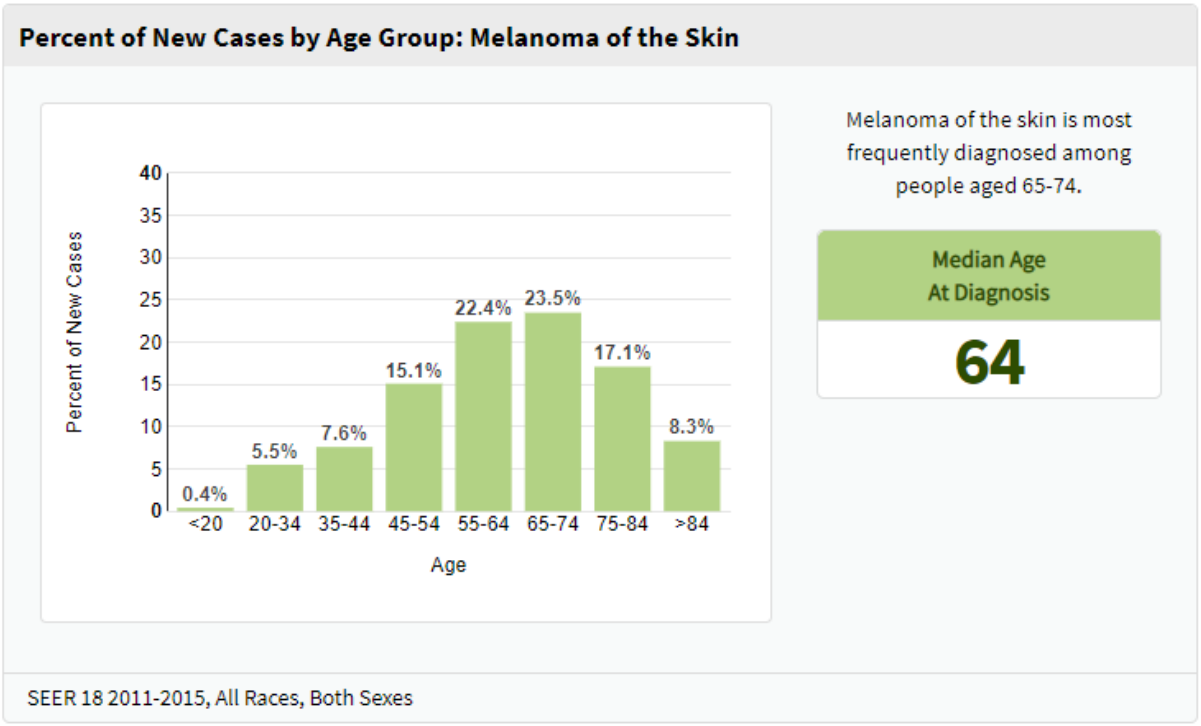


Εικόνα 20 Διαγραμματική απεικόνιση των επιπέδων σύγκρισης και των διμερών συγκρίσεων που πραγματοποιήθηκαν με βάση τον διαχωρισμό των ασθενών σε: άνω των 60 – κάτω των 60, άντρες – γυναίκες, ασθενείς με μεταλλαγή BRAF – ασθενείς χωρίς μεταλλαγή BRAF, ασθενείς με μεταλλαγή RAS – ασθενείς με μεταλλαγή BRAF, καθώς και όλοι οι δυνατοί συνδυασμοί αμάδων ασθενών στο επίπεδο του προφίλ μεταλλάξεων.

4.4.2.1 Επιλογή της ηλικίας ως παράμετρο σύγκρισης

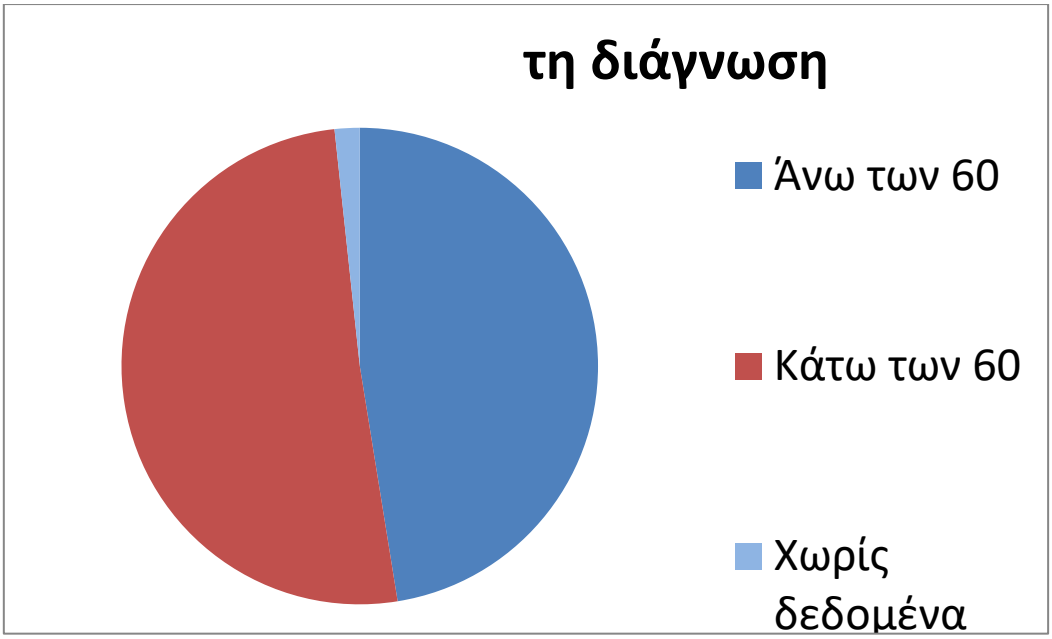
Σύμφωνα με τον Εθνικό Ινστιτούτο Υγείας των ΗΠΑ, η ηλικία των 64^{ων} ετών είναι κομβική για το μελάνωμα. Τα 64 έτη αποτελούν τη μέση ηλικία διάγνωσης του μελανώματος, ενώ αν θελήσουμε να διευρύνουμε, η δεκαετία 64 έως 75 αποτελεί το εύρος στο οποίο γίνεται

συχνότερα η διάγνωση της ασθένειας, με ποσοστό 23,5%, ενώ ακολουθεί η δεκαετία 55 – 64 με ποσοστό 22,4%. Ουσιαστικά, το παραπάνω σημαίνει πως το μεγαλύτερο ποσοστό από ασθενείς που πάσχει από μελάνωμα, διαγιγνώσκεται κατά την ηλικιακή φάση των 64ων έως 75 ετών, ανεξάρτητα από το πότε συνέβη η έναρξη της ασθένειας. Παρακάτω, αναπαρίσταται γραφικά η πληροφορία αυτή.



Εικόνα 21 Γραφική απεικόνιση του αριθμού των νέων περιστατικών μελανώματος ανά ηλικιακό γκρουπ, σε ποσοστιαία κλίμακα. Τα δεδομένα προέρχονται από το Εθνικό Ινστιτούτο Υγείας των ΗΠΑ και αντλήθηκαν των Μάρτιο του 2019 ("National Institutes of Health (NIH) | Turning Discovery Into Health," n.d.).

Τα παραπάνω δεδομένα συμφωνούν με την πληροφορία που συλλέγεται σχετικά με την ηλικία από το σετ δεδομένων του πειράματος.



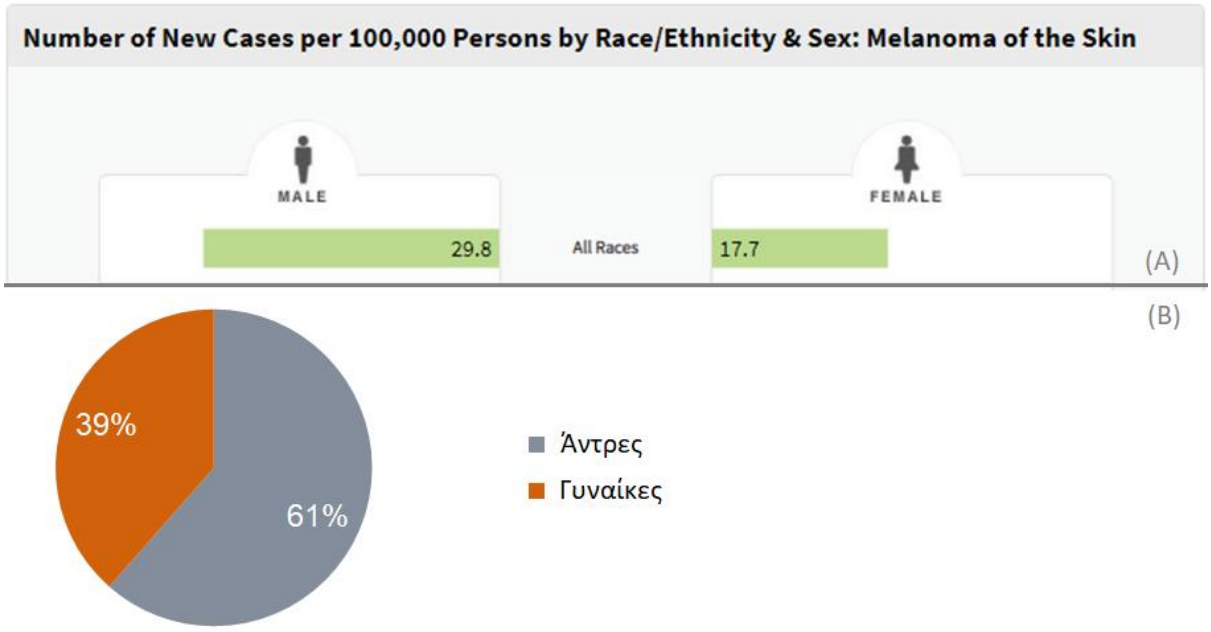
Εικόνα 22 Διάγραμμα μορφής πίττας που αναπαριστά το μέρος των ασθενών που ανήκει σε καθένα από τα ηλιακά γκρουπ άνω των 60 και κάτω των 60. Ο πληθυσμός είναι σχεδόν ισόποσα μοιρασμένος, με εξαίρεση μια μικρή περιοχή ασθενών για τους οποίους δεν υπάρχει η πληροφορία της ηλικίας διάγνωσης και εξαιρούνται από οποιαδήποτε ανάλυση βασίζεται σε αυτό το χαρακτηριστικό.

Όπως είναι προφανές και από το παραπάνω διάγραμμα, η ηλικία των 60 ετών είναι κομβική για τη διάγνωση του μελανώματος, όχι μόνο γιατί πριν και μετά από αυτή ο πληθυσμός είναι σχεδόν ισόποσα διαμοιρασμένος, αλλά και διότι υπάρχουν βιβλιογραφικά δεδομένα που υποστηρίζουν ότι και η ίδια η φυσιολογία του σπίλου του μελανώματος αλλάζει κατά τη δεκαετία των 60 σε σχέση με τους νεότερους ασθενείς, όπως αναφέρθηκε αναλυτικά και στην Εισαγωγή. Δεδομένων αυτών, η ηλικία των 60 ετών επιλέγεται ως ηλικία διαχωρισμού των ασθενών σε δύο επιμέρους κατηγορίες με βάση την ηλικία διάγνωσης της ασθένειας, τους «ασθενείς άνω των 60 ετών» και τους «ασθενείς κάτω των 60 ετών». Η ηλικία των 60 περιλαμβάνεται στο δεύτερο γκρουπ.

Για ένα μικρό ποσοστό των ασθενών της τάξης του 2%, δεν είναι διαθέσιμα τα δεδομένω για την ηλικία διάγνωσης. Αυτοί οι ασθενείς εξαιρούνται από την κύρια ανάλυση και όλες τις δοκιμαστικές αναλύσεις που γίνονται με κριτήριο την ηλικία.

4.4.2.2 *Επιλογή του φύλου ως παράμετρο σύγκρισης*

Το μελάνωμα είναι μια ασθένεια που φαίνεται να δείχνει μια ελαφρά «προτίμηση» στους άντρες σε σχέση με τις γυναίκες, αν και τα ποσοστό εμφάνισης στα δύο φύλα τείνουν σταδιακά να εξισωθούν. Παρακάτω (Α) παρουσιάζονται διαγραμματικά τα αποτελέσματα του Εθνικού Ινστιτούτο Υγείας των Ηνωμένων Πολιτειών για τη συσχέτιση του φύλου με το δερματικό μελάνωμα. Τα δεδομένα αυτά, δεν αφορούν μόνο τον πληθυσμό των Ηνωμένων Πολιτειών, αλλά είναι παγκόσμιας κλίμακας και αφορούν την επίπτωση της ασθένειας σε άντρες και γυναίκες όλων των φυλών ανά πληθυσμό 100.000 ατόμων. Είναι προφανές πως οι άντρες νοσούν συχνότερα από ότι οι γυναίκες, διαφορά, όμως που σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας (WHO), τείνει να εξισωθεί με την πάροδο του χρόνου, συμπέρασμα που προκύπτει από την μελέτη του ρυθμού επιπολασμού της ασθένειας στα δύο φύλα κατά τις τελευταίες δεκαετίες (GBD 2015 Disease and Injury Incidence and Prevalence Collaborators, 2016; International Agancy for Research on Cancer, n.d.).



Εικόνα 23 (Α) Αριθμός νεο-διαγνωσθέντων ασθενών με μελάνωμα ανά 100.000 άτομα, όλων των φυλών / εθνικοτήτων, στο επίπεδο του φύλου (εικόνα τροποποιημένη από International Agancy for Research on Cancer, n.d.). (Β) Διαχωρισμός του υπό μελέτη σετ δεδομένων με βάση το φύλο και παρουσίαση των αποτελεσμάτων σε ποσοστιαία κλίμακα.

Στο δεύτερο μέρος της παραπάνω εικόνας (Β) παρουσιάζεται σε ποσοστιαία κλίμακα ο αριθμός των ανδρών και γυναικών του υπό μελέτη σετ δεδομένων. Από τους 472 ασθενείς

με δερματικό μελάνωμα, οι 184 είναι γυναίκες και οι 288 άντρες. Υπάρχει, λοιπόν, συνέπεια των δεδομένων της βιβλιογραφίας και των δεδομένων της GDC.

Το φύλο αποτελεί μια παράμετρο με βάση την οποία μπορεί να γίνει διμερισμός του αρχικού πλήθους ασθενών, δεδομένου ότι οι πληθυσμιακή διαφορά μεταξύ των υπο – ομάδων που δημιουργούνται δεν είναι απαγορευτική για τη σωστή συνέχιση του πειράματος και την υπονόμευση των τελικών αποτελεσμάτων.

Άλλα δημογραφικά στοιχεία όπως η φυλή, το ιστορικό καπνίσματος και καταχρήσεων και το στάτους των ασθενών κατά την τελευταία επικαιροποίηση των δεδομένων (εν ζωή ή αποθανών), είναι μερικές από τις δημογραφικές παραμέτρους που δεν μπόρεσαν να χρησιμοποιηθούν, είτε διότι οι υπο – ομάδες που προέκυπταν παρουσίαζαν μεγάλη διαφορά πλήθους ασθενών είτε λόγω έλλειψης δεδομένων για ένα μεγάλο μέρος του αρχικού σετ δεδομένων.

4.4.2.3 Επιλογή του προφίλ μεταλλάξεων ως παράμετρο σύγκρισης

Η διαίρεση των ασθενών σε ομάδες με βάση το προφίλ μεταλλάξεων και πιο συγκεκριμένα με βάση την παρουσία ή όχι μεταλλαγής σε συγκεκριμένα γονίδια που θεωρούνται σημαντικά για το μελάνωμα, υποστηρίζεται βιβλιογραφικά (Akbani et al., 2015; Kontogianni et al., 2018). Αναλυτικά, η πληροφορία αυτή έχει συζητηθεί στην ενότητα 2.1 και οι υπο-πληθυσμοί των ασθενών με βάση το προφίλ μεταλλάξεων προκύπτουν δεδομένης της πληροφορίας για μεταλλαγές στα γονίδια BRAF, RAS, NF1.

Παράλληλα από το αποθετήριο cBioportal, γίνεται εξόρυξη της πληροφορίας μεταλλαγών σε άλλα 20 γονίδια με γνωστή βιβλιογραφικά συσχέτιση με το μελάνωμα, τα οποία όμως δεν χρησιμοποιούνται ως κριτήριο διαίρεσης του πληθυσμού, αλλά χρησιμοποιούνται μεταγενέστερα στην ανάλυση και συγκρίνονται με την ιεραρχημένη βάσει εμπλουτισμού λίστα με τα σημαντικότερα γονίδια που εμφανίζουν διαφορικά εκφρασμένα μετάγραφα, που προκύπτει μετά τη χρήση του εργαλείου BioInfoMiner, για γονιδιακό σχολιασμό και εύρεση σχετιζόμενων κυτταρικών μονοπατιών.

4.4.3 Χρήση του πακέτου edgeR της γλώσσας προγραμματισμού R

Προκειμένου να πραγματοποιηθεί η ανάλυση διαφορικής έκφρασης γονιδίων μεταξύ των ομάδων που δημιουργούνται ανά παράμετρο κατηγοριοποίησης, χρησιμοποιήθηκε η γλώσσα προγραμματισμού R, το περιβάλλον R studio και πιο συγκεκριμένα, το πακέτο edgeR για στατιστική ανάλυση αλληλουχιών, καθώς και άλλα βοηθητικά πακέτα συναρτήσεων. Με αυτόν τον τρόπο προέκυψε μια αλγοριθμική ροή, η οποία μπορεί να εφαρμοστεί με επιτυχία και σε άλλα σετ δεδομένων της βάσης δεδομένων GDC, με ίδια δομή.

Η εισαγωγή της παραμέτρου διαίρεσης του αρχικού πληθυσμού γίνεται χειροκίνητα στο αρχικό σύστημα πινάκων, παρόλα αυτά, μέσω της ροής συναρτήσεων εξασφαλίζεται η σωστή αντιστοίχιση της πληροφορίας αυτής με την αρχική λίστα ασθενών / δειγμάτων / κωδικών.

Το φιλτράρισμα στο πρώτο επίπεδο της ανάλυσης πραγματοποιείται με βάση τις τιμές cpm και logcpm, που προκύπτουν από τις αντίστοιχες συναρτήσεις του πακέτου edgeR και οι οποίες αντιστοιχούν στις τιμές μετρήσεων αναγνώσεων ανά εκατομμύριο (counts per million) και την αντίστοιχη λογαριθμημένη τιμή της. Ένα γονίδιο θεωρείται εκφρασμένο όταν παρουσιάζει τιμή cpm > 1 και lcpm > 0. Παράλληλα, σε αυτό το επίπεδο εισάγεται το κριτήριο της έκφρασης των γονιδίων σε πλήθος δειγμάτων τουλάχιστον όσο το πλήθος της μικρότερης από τις δύο υπό – ομάδες που συγκρίνονται. Με αυτόν τον τρόπο εξασφαλίζεται ότι δεν θα υπάρξει κανένα γονίδιο που δεν εκφράζεται σε καμιά από τις δύο ομάδες δειγμάτων / ατόμων / ασθενών που έχουν οριστεί και άρα, το κάθε γονίδιο από αυτά που απομένουν εκφράζεται τουλάχιστον σε ένα επαρκές πλήθος του πληθυσμού. Το τελευταίο κριτήριο είναι αρκετά αυστηρό για αναλύσεις όπου του πλήθος των δειγμάτων προέρχεται εξ ολοκλήρου από πάσχοντες ιστούς, παρόλο που προτείνεται βιβλιογραφικά για αναλύσεις όπου υπάρχουν δείγματα ελέγχου από υγιή ιστό και αναλύεται η διαφορική έκφραση σε σύγκριση με καρκινικούς ιστούς. Παρόλα αυτά, επιλέχθηκε σε αυτή την περίπτωση ως το αυστηρότερο όριο που θα μπορούσε να οριστεί, ώστε να δειχθεί σε μεταγενέστερο στάδιο, αν εξάγονται ασφαλή συμπεράσματα ή έχει απορριφθεί και μέρος χρήσιμης πληροφορίας.

Για τον προσδιορισμό των διαφορικά εκφρασμένων γονιδίων επιλέχθηκαν τα εξής όρια για τις τιμές log2FC και την αντίστοιχη τιμή στατιστικής σημαντικότητας p – value της τιμής

αυτής: log2FC > 0,5 κατά απόλυτη τιμή και p – value < 0,05. Συνοπτικά το log2FC αποτελεί το μέτρο μεταβολής μεταξύ των διαφορετικών καταστάσεων και το αντίστοιχο p – value είναι η στατιστική σημαντικότητα της τιμής log2FC. Σημειώνεται πως, συχνά προτείνονται μεγαλύτερα όρια για την τιμή log2FC, τα οποία δοκιμάστηκαν και οδήγησαν στο συμπέρασμα πως στερούν σημαντική πληροφορία μεταβολής από τα αποτελέσματα, δεδομένου ότι έχουν χρησιμοποιηθεί αρκετά αυστηρά όρια και στο στάδιο του φιλτραρίσματος.

5 Αποτελέσματα

Παρακάτω παρουσιάζονται τα αποτελέσματα της ανάλυσης διαφορικής έκφρασης γονιδίων που πραγματοποιήθηκε σε σετ 472 ασθενών που πάσχουν από δερματικό μελάνωμα. Τα αρχικά δεδομένα αντλήθηκαν από τη βάση δεδομένων GDC, η οποία αποτελεί αυτή τη στιγμή το μεγαλύτερο αποθετήριο δεδομένων αλληλούχισης και γονιδιακής πληροφορίας παγκοσμίως.

Η μελέτη διαφορικής έκφρασης πραγματοποιήθηκε στο επίπεδο του φύλου, της ηλικίας και του προφίλ μεταλλάξεων. Από τις 9 στο σύνολο αναλύσεις που πραγματοποιήθηκαν, οι 3 οδήγησαν στη δημιουργία επαρκών και αξιόπιστων αποτελεσμάτων, μεταξύ ασθενών άνω των 60 και κάτω των 60, μεταξύ ασθενών που φέρουν μεταλλαγή BRAF και που δεν φέρουν μεταλλαγή BRAF και μεταξύ ασθενών που φέρουν μεταλλαγή BRAF και ασθενών που φέρουν μεταλλαγή RAS. Από τις παραπάνω, όλες κατέληξαν σε λίστα με αριθμό διαφορικά εκφρασμένων γονίδιων ανάμεσα στις δύο καταστάσεις που είχαν οριστεί. Τελικά, μία από αυτές κατέληξε σε ιδιαίτερα ενδιαφέροντα συμπεράσματα, μετά και την ανάλυση εμπλουτισμού και παρουσιάζεται εκτενέστερα παρακάτω.

Μελλοντικά, οι υπόλοιπες συγκρίσεις θα πραγματοποιηθούν με διαφοροποιημένες παραμέτρους, που ίσως να ανταποκρίνονται καλύτερα στην εκάστοτε ομαδοποίηση.

5.1 Αποτελέσματα σύγκρισης ασθενών ηλικίας άνω των 60 και ασθενών ηλικίας κάτω των 60 ετών

Στον πίνακα που ακολουθεί παρουσιάζονται να γονίδια που βρέθηκαν διαφορικά εκφρασμένα μεταξύ των ασθενών με ηλικία διάγνωσης της νόσου μετά τα 60 έτη και των ασθενών με ηλικία διάγνωσης της νόσου πριν την ηλικία των 60 ετών.

Πίνακας 6 Διαφορικά εκφρασμένα γονίδια μεταξύ των ασθενών με ηλικία διάγνωσης του μελανώματος μετά τα 60 έτη και των ασθενών με ηλικία διάγνωσης του μελανώματος πριν τα 60 έτη.

ENSEMBL	SYMBOL	logFC	PValue	logCPM	TXCHROM
ENSG00000171094	ALK	1,479749	9,73996E-16	3,1664318	chr2
ENSG00000154277	UCHL1	1,180092	3,36066E-07	4,2800352	chr4
ENSG00000162493	PDPN	1,158783	1,25014E-10	3,4986593	chr1
ENSG00000125851	PCSK2	1,110052	1,208E-05	4,2073961	chr20
ENSG00000100399	CHADL	1,071062	2,01281E-08	3,1129061	chr22
ENSG00000019991	HGF	-1,00396	6,14848E-07	2,6742376	chr7
ENSG00000165092	ALDH1A1	-1,03789	1,51707E-10	5,8911819	chr9
ENSG00000111907	TPD52L1	-1,06422	2,58958E-09	3,6713656	chr6
ENSG00000158869	FCER1G	-1,11226	5,04235E-13	5,6945023	chr1
ENSG00000137558	PI15	-1,11411	4,3066E-07	5,2985687	chr8
ENSG00000198734	F5	-1,14299	4,58942E-06	4,0789484	chr1
ENSG00000101938	CHRDL1	-1,14982	3,9263E-06	3,9040794	chrX

ENSG00000082196	C1QTNF3	-1,16628	2,3084E-06	4,9981904	chr5
ENSG00000166825	ANPEP	-1,18456	9,69502E-11	4,9663781	chr15
ENSG00000013297	CLDN11	-1,20577	3,75101E-08	3,1224351	chr3
ENSG00000164283	ESM1	-1,33187	1,65808E-13	3,0533364	chr5
ENSG00000157368	IL34	-1,33739	8,06949E-16	2,270664	chr16
ENSG00000123689	G0S2	-1,34367	3,93333E-13	2,7006488	chr1
ENSG00000143248	RGS5	-1,35139	6,97273E-12	6,6435746	chr1
ENSG00000196611	MMP1	-1,3535	1,45556E-07	5,1507084	chr11
ENSG00000159167	STC1	-1,39121	4,92502E-11	5,6216936	chr8
ENSG00000091513	TF	-1,42385	3,32027E-08	5,7205437	chr3
ENSG00000133063	CHIT1	-1,57596	9,61033E-09	4,8111988	chr1

Στον πίνακα περιλαμβάνονται οι στήλες: *ENSEMBL* που αναφέρεται στο ensemble id του κάθε γονιδίου, η στήλη *SYMBOL* που αφορά το σύμβολο του κάθε γονιδίου με κωδικό τριών ή τεσσάρων γραμμάτων, η στήλη *logFC* η οποία αναπαριστά την τιμή log fold change, δηλαδή τον μέτρο διαφοράς μεταξύ των δύο καταστάσεων και όπως φαίνεται είναι πάντοτε μεγαλύτερη του 1 κατά απόλυτη τιμή και για τη συγκεκριμένη ανάλυση και μόνο, η στήλη *PValue*, που αφορά της στατιστική σημαντικότητα της τιμής logFC και περιέχει μόνο τιμές μικρότερες του 0,05, η στήλη *logCPM*, δηλαδή η τιμή με βάση την οποία γίνεται το πρώτο βήμα του φιλτραρίσματος των δεδομένων και είναι πάντοτε θετική και τέλος, η στήλη *TXCHROM*, η οποία δίνει την χρωμοσωμική θέση του κάθε γονιδίου.

Είναι εμφανές πως όλα τα φίλτρα του αρχικού σετ δεδομένων έχουν λειτουργήσει και τα γονίδια που προκύπτουν παρουσιάζουν διαφορετική έκφραση μεταξύ των δύο κατηγοριών ασθενών.

Να σημειωθεί πως το μεγαλύτερο ποσοστό των γονιδίων υπο – εκφράζεται στους ασθενείς άνω των 60 ετών σε σχέση με τους νεότερους, ενώ ένα χαρακτηριστικό παράδειγμα είναι η περίπτωση της *IL34* ιντερλευκίνης, η οποία παρουσιάζει διακριτή υπο – έκφραση στη δεύτερη κατηγορία σε σχέση με την πρώτη. Ένα τέτοιο εύρημα θα μπορούσε να υποδηλώνει έναν διαφορετικό μηχανισμό ανοσολογικής ενεργοποίησης και φλεγμονής στα δύο ηλικιακά γκρουπ, στους οποίους μηχανισμούς εμπλέκονται οι ιντερλευκίνες. Συμπερασματικά τα δεδομένα αυτά χρήζουν εμπλουτισμού, λειτουργικής ανάλυσης και περαιτέρω διερεύνησης.

5.2 Αποτελέσματα σύγκρισης ασθενών με BRAF μεταλλαγή και ασθενών χωρίς BRAF μεταλλαγή

Παρόμοια με παραπάνω, στον πίνακα που ακολουθεί παρουσιάζονται στα δεδομένα που αφορούν τα διαφορικά εκφρασμένα γονίδια μεταξύ των ασθενών που φέρουν μεταλλαγή BRAF (και όχι ταυτόχρονη RAS) και ασθενών που φέρουν μεταλλαγή RAS (και όχι ταυτόχρονη BRAF).

Πίνακας 7 Διαφορικά εκφραζόμενα γονίδια μεταξύ των καταστάσεων παρουσία μεταλλαγής BRAF – παρουσία μεταλλαγής RAS. Ένας μικρός αριθμός ασθενών που παρουσίασε και τις δύο μεταλλαγές ταυτόχρονα, έχει εξαιρεθεί από τις συγκρίσεις.

ENSEMBL	SYMBOL	logFC	PValue	logCPM	TXCHROM
ENSG00000007062	PROM1	-1,42142	5,87846E-08	2,35645065	chr4
ENSG00000019991	HGF	0,523018	0,035520572	3,009154979	chr7
ENSG00000036448	MYOM2	0,560201	0,003079979	1,054181968	chr8
ENSG00000044524	EPHA3	0,788603	0,00053222	3,274092135	chr3
ENSG00000047648	ARHGAP6	-0,81641	0,000273108	2,057136729	chrX

ENSG00000062282	DGAT2	-0,6783	0,000200827	3,141226013	chr11
ENSG00000066230	SLC9A3	-1,15823	9,15859E-05	3,191710809	chr5
ENSG00000069535	MAOB	-0,74517	0,002033099	3,935660695	chrX
ENSG00000070669	ASNS	-0,54449	0,00093812	4,329896744	chr7
ENSG00000071242	RPS6KA2	-1,05989	1,77806E-07	5,332944882	chr6
ENSG00000074410	CA12	0,733564	0,001616987	4,349537379	chr15
ENSG00000077782	FGFR1	0,526447	0,000786866	6,578302107	chr8
ENSG00000077943	ITGA8	1,15064	1,12012E-07	2,254883964	chr10
ENSG00000077984	CST7	-0,50904	0,018634626	3,746852593	chr20
ENSG00000078295	ADCY2	-0,64092	0,033592007	3,538149431	chr5
ENSG00000078596	ITM2A	0,748268	7,33334E-05	3,741093257	chrX
ENSG00000078900	TP73	-0,51089	0,005169514	1,200341437	chr1
ENSG00000084207	GSTP1	-0,54094	0,000487877	8,872639793	chr11
ENSG00000085563	ABCB1	0,579812	0,000339326	1,737203989	chr7
ENSG00000088881	EBF4	0,547597	0,004006203	2,410661761	chr20
ENSG00000089472	HEPH	0,955783	3,84532E-09	2,463904437	chrX
ENSG00000091513	TF	-0,7705	0,013814868	6,099726172	chr3

ENSG00000095303	PTGS1	-0,56598	0,001418256	3,569811951	chr9
ENSG00000096696	DSP	-0,92515	0,00409137	6,219097407	chr6
ENSG00000099284	H2AFY2	0,592121	0,000841604	2,369991216	chr10
ENSG00000101115	SALL4	0,54078	0,007256524	1,218176102	chr20
ENSG00000101213	PTK6	-0,79166	0,000413904	1,867975588	chr20
ENSG00000101938	CHRDL1	0,582477	0,045999162	4,198432212	chrX
ENSG00000102890	ELMO3	-0,50575	0,001528466	1,400457097	chr16
ENSG00000103044	HAS3	-0,52167	0,003907254	1,462873225	chr16
ENSG00000103355	PRSS33	-0,70597	0,02441636	2,653863751	chr16
ENSG00000103740	ACSBG1	-0,65492	0,006736608	2,170792732	chr15
ENSG00000104881	PPP1R13L	-0,67773	9,44631E-06	2,827434492	chr19
ENSG00000105376	ICAM5	0,645694	0,001737802	1,440728131	chr19
ENSG00000105519	CAPS	0,832763	7,27245E-05	4,76163108	chr19
ENSG00000105711	SCN1B	-0,5787	0,000366445	3,630174525	chr19
ENSG00000105989	WNT2	0,842432	0,000265106	0,89075259	chr7
ENSG00000106236	NPTX2	-0,81591	0,020009314	5,774285948	chr7
ENSG00000106366	SERPINE1	0,549237	0,012927752	6,499245453	chr7

ENSG00000106484	MEST	0,65171	0,010020672	4,439446158	chr7
ENSG00000106537	TSPAN13	0,537063	0,004386829	4,228693976	chr7
ENSG00000107984	DKK1	1,001935	0,000298607	2,719381795	chr10
ENSG00000108352	RAPGEFL1	-0,53923	0,000568945	2,751567792	chr17
ENSG00000109472	CPE	0,666958	0,000373594	4,365185586	chr4
ENSG00000109956	B3GAT1	-0,51972	0,03182789	2,005086319	chr11
ENSG00000110693	SOX6	-0,60783	0,002164941	4,649129606	chr11
ENSG00000111215	PRH1-PRR4	-0,60923	0,008147258	1,767216116	chr12
ENSG00000112378	PERP	-0,59116	0,003520811	6,358378473	chr6
ENSG00000112936	C7	0,730158	0,005295265	4,433819429	chr5
ENSG00000113296	THBS4	0,828524	0,000218405	3,793859054	chr5
ENSG00000113594	LIFR	1,200209	1,42257E-06	4,334936964	chr5
ENSG00000114251	WNT5A	1,589798	4,08121E-13	2,86141027	chr3
ENSG00000114270	COL7A1	0,726261	0,001437443	4,495219042	chr3
ENSG00000114646	CSPG5	0,556948	0,010662771	2,954665286	chr3
ENSG00000115474	KCNJ13	-0,87682	0,00131919	3,165624275	chr2
ENSG00000116299	KIAA1324	-1,09538	2,03519E-07	1,276266337	chr1

ENSG00000117152	RGS4	0,8205	0,000179116	2,523492074	chr1
ENSG00000117595	IRF6	-0,7956	0,002160527	2,288962931	chr1
ENSG00000117600	PLPPR4	0,697976	0,004679924	3,053725841	chr1
ENSG00000118473	SGIP1	0,515169	0,000859344	0,573821128	chr1
ENSG00000118898	PPL	-0,71943	0,005266804	4,510614189	chr16
ENSG00000118946	PCDH17	0,975962	4,62245E-10	2,490805089	chr13
ENSG00000120693	SMAD9	0,524825	0,001113154	1,443648448	chr13
ENSG00000120708	TGFBI	0,602016	0,001981116	7,941923322	chr5
ENSG00000120820	GLT8D2	0,762574	2,15522E-05	2,072652293	chr12
ENSG00000121207	LRAT	-0,6027	0,012660974	2,785039234	chr4
ENSG00000121552	CSTA	-0,97756	0,000541002	3,869627206	chr3
ENSG00000121769	FABP3	-0,60721	6,12098E-05	3,257781908	chr1
ENSG00000121898	CPXM2	0,849286	0,000672179	4,225121769	chr10
ENSG00000122254	HS3ST2	-1,05091	5,04731E-05	2,316255573	chr16
ENSG00000122367	LDB3	-0,64128	0,002501998	3,829482193	chr10
ENSG00000123500	COL10A1	0,561513	0,014658474	1,964733353	chr6
ENSG00000123689	G0S2	0,856662	0,000204772	3,051114472	chr1

ENSG00000124466	LYPD3	-1,89629	9,5858E-09	4,864261293	chr19
ENSG00000124731	TREM1	1,348963	7,05361E-08	2,631330945	chr6
ENSG00000125538	IL1B	1,320164	8,32995E-08	2,878881676	chr2
ENSG00000125740	FOSB	-0,56721	0,006462029	4,52014998	chr19
ENSG00000125775	FKBP1A-SDCBP2	-0,61362	0,000106929	1,416857672	chr20
ENSG00000126500	FLRT1	-0,58519	0,00622552	1,286060754	chr11
ENSG00000127824	TUBA4A	-0,91584	2,17917E-05	4,278695142	chr2
ENSG00000129244	ATP1B2	-0,51534	0,044097901	4,677705867	chr17
ENSG00000129993	CBFA2T3	-0,68275	0,00040639	1,790022532	chr16
ENSG00000130396	AFDN	-0,59684	4,7257E-06	5,069996589	chr6
ENSG00000130558	OLFM1	-0,50443	0,035985172	4,008263225	chr9
ENSG00000130600	H19	0,869227	4,14981E-05	4,781369709	chr11
ENSG00000130829	DUSP9	0,681898	0,001286311	1,659992232	chrX
ENSG00000130940	CASZ1	-0,74019	2,26909E-05	1,592402802	chr1
ENSG00000131037	EPS8L1	-0,50293	0,000806104	3,071429865	chr19
ENSG00000131386	GALNT15	0,505265	0,008528193	1,104824336	chr3

ENSG00000132031	MATN3	0,792922	0,000242759	1,487128382	chr2
ENSG00000133019	CHRM3	0,827281	0,000888148	2,516430723	chr1
ENSG00000133048	CHI3L1	-1,39039	3,81276E-06	6,707994411	chr1
ENSG00000133063	CHIT1	-0,8523	0,012539886	5,046696143	chr1
ENSG00000133800	LYVE1	0,507561	0,011235674	1,787595522	chr11
ENSG00000134317	GRHL1	-0,81381	9,31517E-05	1,423295004	chr2
ENSG00000134755	DSC2	-0,69322	0,006277085	3,15102243	chr18
ENSG00000134827	TCN1	-1,19144	0,000126122	4,278364105	chr11
ENSG00000134853	PDGFRA	0,503332	0,001210349	3,581720737	chr4
ENSG00000135144	DTX1	-0,5265	0,006734935	2,362376204	chr12
ENSG00000136002	ARHGEF4	-0,90764	0,000207887	2,021407775	chr2
ENSG00000136689	IL1RN	-1,02312	1,34343E-06	3,075901108	chr2
ENSG00000136943	CTSV	-0,51859	0,003781487	2,059320331	chr9
ENSG00000136999	NOV	0,55666	0,012783589	5,499764541	chr8
ENSG00000137077	CCL21	0,697198	0,026296447	5,710387193	chr9
ENSG00000137857	DUOX1	-0,78311	0,002031161	2,787790716	chr15
ENSG00000137868	STRA6	0,60538	0,01297989	4,738374981	chr15

ENSG00000137976	DNASE2B	-0,93035	0,001648776	2,546644741	chr1
ENSG00000138316	ADAMTS14	0,816939	9,05584E-06	2,061482084	chr10
ENSG00000138356	AOX1	1,122098	4,1464E-09	2,102661067	chr2
ENSG00000138435	CHRNA1	-0,90603	0,000505071	2,229455247	chr2
ENSG00000138722	MMRN1	0,908712	1,67303E-05	1,101469772	chr4
ENSG00000138829	FBN2	1,42231	2,28886E-11	3,064392598	chr5
ENSG00000138944	SHISAL1	-0,5266	0,033504592	2,147432838	chr22
ENSG00000140450	ARRDC4	-0,65054	5,69081E-05	4,870228147	chr15
ENSG00000140519	RHCG	-1,88202	2,26695E-09	3,083635999	chr15
ENSG00000142156	COL6A1	0,556581	0,000156094	8,868407715	chr21
ENSG00000142173	COL6A2	0,510303	0,001082884	8,840623435	chr21
ENSG00000143127	ITGA10	1,316254	3,3247E-07	6,119750918	chr1
ENSG00000143248	RGS5	-0,56567	0,024832661	6,955971143	chr1
ENSG00000143546	S100A8	-1,35923	0,000119988	6,262337083	chr1
ENSG00000143631	FLG	-1,03072	0,001459968	4,90871753	chr1
ENSG00000144355	DLX1	0,550359	0,022226076	3,356044123	chr2
ENSG00000144730	IL17RD	0,692327	0,000293105	4,256455068	chr3

ENSG00000144810	COL8A1	0,613626	0,006153088	4,363333422	chr3
ENSG00000145358	DDIT4L	1,011605	1,08931E-05	1,927027439	chr4
ENSG00000146197	SCUBE3	0,678421	0,007139112	3,908994675	chr6
ENSG00000147257	GPC3	1,689568	1,58238E-13	3,522444439	chrX
ENSG00000147889	CDKN2A	-0,52973	0,038629262	4,532877207	chr9
ENSG00000148053	NTRK2	0,552096	0,015361741	3,016520924	chr9
ENSG00000148468	FAM171A1	0,574467	0,001298156	2,790895322	chr10
ENSG00000149090	PAMR1	-0,8396	3,49258E-05	1,418060877	chr11
ENSG00000149418	ST14	-0,78141	4,48876E-05	3,507702913	chr11
ENSG00000149970	CNKSR2	0,519755	0,00544365	1,219512867	chrX
ENSG00000150048	CLEC1A	0,613945	0,004413448	2,036393703	chr12
ENSG00000150893	FREM2	1,232478	2,72073E-06	4,406364331	chr13
ENSG00000151090	THRB	0,56685	0,006003781	1,655298178	chr3
ENSG00000151617	EDNRA	0,704939	6,92143E-06	2,467270602	chr4
ENSG00000151790	TDO2	1,88315	1,88415E-19	2,281979202	chr4
ENSG00000153162	BMP6	0,954937	6,63784E-06	2,10222522	chr6
ENSG00000154122	ANKH	0,50779	0,005654281	6,315268944	chr5

ENSG00000154175	ABI3BP	0,578948	0,001636606	3,370835995	chr3
ENSG00000154262	ABCA6	0,529065	0,03233022	3,538131564	chr17
ENSG00000154330	PGM5	0,772965	1,09399E-05	1,643674939	chr9
ENSG00000154928	EPHB1	0,619162	0,001023239	1,218391031	chr3
ENSG00000155093	PTPRN2	-0,65324	9,67649E-05	1,315901095	chr7
ENSG00000156219	ART3	1,039713	7,90927E-05	2,389606052	chr4
ENSG00000156510	HKDC1	-0,60007	0,020235326	3,462061891	chr10
ENSG00000157734	SNX22	0,6921	0,001875879	4,077130607	chr15
ENSG00000158246	TENT5B	-0,51816	0,002207514	3,942748771	chr1
ENSG00000158458	NRG2	-0,54012	0,000411137	2,256970673	chr5
ENSG00000158683	PKD1L1	-0,59327	0,001075421	0,943728523	chr7
ENSG00000159166	LAD1	-1,0447	0,00013791	4,246032983	chr1
ENSG00000159167	STC1	-0,92988	0,000505605	6,024918807	chr8
ENSG00000159871	LYPD5	-0,59452	0,006687581	2,481748527	chr19
ENSG00000160117	ANKLE1	0,505061	0,003322737	0,935207349	chr19
ENSG00000160161	CILP2	0,755457	0,00020401	2,017048952	chr19
ENSG00000160716	CHRNB2	0,708014	0,001990253	1,45234913	chr1

ENSG00000162493	PDPN	-1,32411	3,5975E-08	3,71553924	chr1
ENSG00000162614	NEXN	0,503793	0,000385585	1,412239225	chr1
ENSG00000162706	CADM3	0,864234	0,001208027	3,110998369	chr1
ENSG00000162981	FAM84A	-1,09628	4,8725E-06	2,48699196	chr2
ENSG00000163220	S100A9	-1,58253	7,52225E-07	7,002567736	chr1
ENSG00000163235	TGFA	0,575824	0,030045131	3,559153552	chr2
ENSG00000163347	CLDN1	-0,7581	0,004583258	4,549227975	chr3
ENSG00000163359	COL6A3	0,529627	0,000501922	8,328355876	chr2
ENSG00000163472	TMEM79	-0,52631	5,24902E-05	3,471632401	chr1
ENSG00000163485	ADORA1	0,597347	0,003076586	2,197914035	chr1
ENSG00000163536	SERPINI1	-0,52844	0,004731826	3,425028743	chr3
ENSG00000163599	CTLA4	0,63784	0,001520976	2,405204032	chr2
ENSG00000163661	PTX3	0,577002	0,006191374	1,807633904	chr3
ENSG00000163814	CDCP1	-0,56097	0,024957556	4,783482657	chr3
ENSG00000164176	EDIL3	0,876207	0,000136675	3,349295517	chr5
ENSG00000164251	F2RL1	1,028799	6,04203E-07	1,539559126	chr5
ENSG00000164484	TMEM200A	0,552584	0,004600588	1,557632058	chr6

ENSG00000164687	FABP5	-0,54582	0,001035429	4,899706794	chr8
ENSG00000164733	CTSB	-0,50879	0,000178193	11,07645395	chr8
ENSG00000164761	TNFRSF11B	0,691985	0,002767772	1,858866077	chr8
ENSG00000165272	AQP3	-1,14432	8,38601E-05	4,902644137	chr9
ENSG00000165474	GJB2	-1,01382	0,000931869	6,087804023	chr13
ENSG00000166106	ADAMTS15	0,832986	1,58366E-05	1,827583666	chr11
ENSG00000166250	CLMP	0,540929	0,007623964	3,693890227	chr11
ENSG00000166825	ANPEP	1,523789	8,1663E-12	5,26581661	chr15
ENSG00000167157	PRRX2	-0,53157	0,008950738	1,950455607	chr9
ENSG00000167244	IGF2	1,0198	8,63122E-07	4,693401887	chr11
ENSG00000167306	MYO5B	-0,58332	0,011916487	3,841253331	chr18
ENSG00000167642	SPINT2	-0,76918	3,99847E-05	3,544944513	chr19
ENSG00000167653	PSCA	-0,79095	0,004232253	2,872454462	chr8
ENSG00000168079	SCARA5	0,785428	0,001260371	1,886765395	chr8
ENSG00000168309	FAM107A	0,738022	5,46201E-05	1,694457175	chr3
ENSG00000168481	LGI3	-0,73629	0,005224505	4,784378588	chr8
ENSG00000168528	SERINC2	-0,59166	0,000214235	5,149911256	chr1

ENSG00000168646	AXIN2	0,522747	0,002356061	2,22296584	chr17
ENSG00000168843	FSTL5	0,637939	0,025056087	3,124220716	chr4
ENSG00000169184	MN1	1,261496	8,1177E-13	1,636989309	chr22
ENSG00000169213	RAB3B	0,779793	0,002529162	2,716388483	chr1
ENSG00000169946	ZFPM2	0,580395	0,000256149	1,267074772	chr8
ENSG00000170412	GPRC5C	-0,50242	0,018532219	2,450155555	chr17
ENSG00000170689	HOXB9	0,767101	0,007578099	3,632159162	chr17
ENSG00000170961	HAS2	0,888881	8,10156E-06	4,671222552	chr8
ENSG00000171033	PKIA	0,704728	0,001044313	2,018938513	chr8
ENSG00000171476	HOPX	-0,57142	0,010888319	2,325299923	chr4
ENSG00000172382	PRSS27	-0,99618	9,35556E-08	1,287579977	chr16
ENSG00000173156	RHOD	-0,57864	0,013036261	2,25436031	chr11
ENSG00000173530	TNFRSF10D	0,676731	0,005784795	2,429984967	chr8
ENSG00000173801	JUP	-0,94461	2,09495E-06	6,670826469	chr17
ENSG00000175198	PCCA	-0,59345	1,25305E-05	3,920523273	chr13
ENSG00000175445	LPL	0,667616	0,004810112	5,488342124	chr8
ENSG00000175946	KLHL38	-0,58216	0,001389418	1,3660514	chr8

ENSG00000176402	GJC3	1,04839	1,64453E-08	1,465529801	chr7
ENSG00000176907	TCIM	0,627816	0,000628706	2,312332453	chr8
ENSG00000177283	FZD8	0,821085	0,00041683	2,627284267	chr10
ENSG00000180155	LYNX1	-0,74409	0,000620071	3,924239011	chr8
ENSG00000181449	SOX2	0,568799	0,040515074	2,7043437	chr3
ENSG00000181458	TMEM45A	-0,72567	5,12642E-05	4,538633289	chr3
ENSG00000181617	FDCSP	0,922414	0,003981261	4,894679069	chr4
ENSG00000182168	UNC5C	0,810233	0,000438409	2,075195922	chr4
ENSG00000182379	NXPH4	-0,61389	0,0030521	2,41638523	chr12
ENSG00000182463	TSHZ2	0,71455	7,13468E-05	2,041366822	chr20
ENSG00000184226	PCDH9	0,656915	0,002662316	3,882252418	chr13
ENSG00000185070	FLRT2	0,52521	0,006611294	3,471614784	chr14
ENSG00000185518	SV2B	-1,23759	2,69491E-06	2,528695762	chr15
ENSG00000185633	NDUFA4L2	-0,82693	1,64758E-06	4,112350107	chr12
ENSG00000185896	LAMP1	-0,53976	1,65397E-06	8,251502879	chr13
ENSG00000186081	KRT5	-1,70822	0,000348981	8,311945626	chr12
ENSG00000186395	KRT10	-0,67204	0,024308475	7,732244005	chr17

ENSG00000186407	CD300E	0,533032	0,001501326	0,696654303	chr17
ENSG00000187775	DNAH17	-0,50955	0,000145034	1,128566551	chr17
ENSG00000188015	S100A3	0,752361	2,98017E-05	1,466566913	chr1
ENSG00000188153	COL4A5	0,539385	0,031124647	2,969971644	chrX
ENSG00000188242	PP7080	-0,71418	3,47733E-05	3,614114789	chr5
ENSG00000188322	SBK1	0,533214	0,005661624	2,363081178	chr16
ENSG00000188573	FBLL1	-0,60672	0,006004994	1,062303817	chr5
ENSG00000189184	PCDH18	0,894655	1,19465E-07	2,89883503	chr4
ENSG00000196139	AKR1C3	0,639695	0,001334672	1,275816687	chr10
ENSG00000196754	S100A2	-1,82747	6,73774E-09	4,645895553	chr1
ENSG00000197249	SERPINA1	0,675436	0,00046604	5,475273862	chr14
ENSG00000197467	COL13A1	0,871121	2,16539E-07	1,557167324	chr10
ENSG00000197614	MFAP5	0,622594	0,007771878	2,697941131	chr12
ENSG00000197766	CFD	-0,61231	0,000647568	2,775497057	chr19
ENSG00000198734	F5	1,009744	0,000494272	4,27682011	chr1
ENSG00000203867	RBM20	-0,50373	0,031507189	2,064057214	chr10
ENSG00000211445	GPX3	-0,91793	5,07704E-05	6,436040399	chr5

ENSG00000214548	MEG3	0,732598	0,002873374	2,973991846	chr14
ENSG00000223573	TINCR	-1,06592	4,59363E-05	2,204470631	chr19
ENSG00000225614	ZNF469	0,590751	6,97233E-05	2,605164669	chr16
ENSG00000231528	FAM225A	0,567087	0,001232798	0,891175305	chr9
ENSG00000235750	KIAA0040	0,64743	0,000134096	4,583644713	chr1
ENSG00000243449	C4orf48	-0,50095	0,000674137	2,603588891	chr4
ENSG00000244734	HBB	0,649561	0,008477976	3,073762043	chr11
ENSG00000250337	PURPL	-0,65882	0,016476372	1,192053561	chr5
ENSG00000253159	PCDHGA12	-0,64671	0,00948703	4,061816722	chr5
ENSG00000253873	PCDHGA11	-0,65956	0,00523411	2,354241655	chr5
ENSG00000255346	NOX5	-1,41537	4,88951E-07	2,275317893	chr15
ENSG00000255690	TRIL	0,709151	0,002439024	4,41302849	chr7
ENSG00000256087	ZNF432	0,521163	3,68471E-05	3,152678924	chr19
ENSG00000270885	RASL10B	0,764767	0,000207339	1,587606616	chr17
ENSG00000272398	CD24	-0,8127	0,003635334	2,944862759	chrY
ENSG00000275385	CCL18	-0,5507	0,01590115	4,228126087	chr17
ENSG00000275395	FCGBP	-0,5959	0,00635779	3,290102828	chr19

259 γονίδια εμφανίζονται διαφορικά εκφραζόμενα, με τιμές logFC > 0,5 κατά απόλυτη τιμή και αντίστοιχο p – value < 0,05. Η τιμή logFC τροποποιείται σε σχέση με τις συγκρίσεις στο επίπεδο των δημογραφικών χαρακτηριστικών, καθώς εδώ, πιο αυστηρά κριτήρια δοκιμάστηκαν και προκύπτει πλήθος αισθητά μικρότερο, το οποίο δεν μπορεί να δώσει στιβαρά αποτελέσματα μετά τον εμπλουτισμό. Από την άλλη πλευρά, αυτή η εκτενής λίστα μπορεί να οδηγήσει σε υπόδειξη κάποιων επηρεασμένων μονοπατιών, όμως η εξαγωγή συμπερασμάτων δεν υπήρξε εύκολη, λόγω της πολυπλοκότητας των αντίστοιχων οπτικοποιήσεων και πινάκων και χρήζει περεταίρω προσοχής και μελέτης.

5.3 Αποτελέσματα σύγκρισης ασθενών με BRAF μεταλλαγή και ασθενών χωρίς BRAF μεταλλαγή

Οι ασθενείς είναι χωρισμένοι σε δύο υποκατηγορίες ανάλογα με το προφίλ μεταλλάξεων που αντλήθηκε από τη βάση GDC. Η αντίστοιχη ανάλυση οδήγησε σε μια λίστα 140 διαφορικά εκφρασμένων γονιδίων η οποία παρατίθεται παρακάτω.

Πίνακας 8 Διαφορικά εκφραζόμενα γονίδια μεταξύ των ασθενών που φέρουν BRAF μεταλλαγή και αυτών που δεν φέρουν μεταλλαγή στο εν λόγω γονίδιο.

ENSEMBL	SYMBOL	logFC	PValue	logCPM	TXCHROM
ENSG00000003096	KLHL13	0.599288933	0.003236083	2.22200087	chrX
ENSG00000005108	THSD7A	1.027886787	7.77898E-09	1.7383542	chr7
ENSG00000006638	TBXA2R	-0.62750221	2.38621E-05	2.71061787	chr19
ENSG00000019169	MARCO	-0.63326228	0.002630039	3.59721983	chr2
ENSG00000019991	HGF	-1.0627842	3.99438E-07	2.78185333	chr7
ENSG00000033327	GAB2	0.531335532	3.21454E-05	6.54928184	chr11

ENSG00000062282	DGAT2	0.503650344	0.000955608	3.10950172	chr11
ENSG00000071242	RPS6KA2	-0.66157244	8.67088E-05	5.31353734	chr6
ENSG00000073756	PTGS2	0.831884118	2.10177E-05	2.07994267	chr1
ENSG00000077943	ITGA8	0.655061441	0.000705046	2.16237431	chr10
ENSG00000091129	NRCAM	0.531494679	0.006023539	4.15581004	chr7
ENSG00000091513	TF	-1.15500191	1.48072E-05	5.85432357	chr3
ENSG00000099953	MMP11	0.924293107	1.0487E-08	4.39431246	chr22
ENSG00000101230	ISM1	0.576669713	0.004881381	2.24616781	chr20
ENSG00000103888	CEMIP	0.692946475	0.000244598	5.1323451	chr15
ENSG00000105929	ATP6V0A4	-0.54962735	0.013784491	2.69229948	chr7
ENSG00000107317	PTGDS	-0.91162907	3.85213E-05	6.76838572	chr9
ENSG00000107968	MAP3K8	0.586145101	1.45025E-05	1.9373006	chr10
ENSG00000112936	C7	-0.50050948	0.032732076	4.47599378	chr5
ENSG00000113296	THBS4	1.512050814	7.81137E-11	4.57120815	chr5
ENSG00000115232	ITGA4	-0.51224929	0.002073614	4.18266032	chr2
ENSG00000117245	KIF17	-0.57171591	0.006037837	3.15584608	chr1
ENSG00000117600	PLPPR4	0.53681629	0.014956951	3.11521379	chr1

ENSG00000119703	ZC2HC1C	-0.5900193	8.16738E-05	1.79395877	chr14
ENSG00000121552	CSTA	1.052766979	2.34937E-05	4.02975865	chr3
ENSG00000122176	FMOD	0.728099074	0.000741001	6.15341581	chr1
ENSG00000122861	PLAU	0.60846971	0.000298543	5.22560341	chr10
ENSG00000123243	ITIH5	0.632488797	0.021257605	6.74177426	chr10
ENSG00000123689	G0S2	0.767668502	0.000118211	2.81977982	chr1
ENSG00000124466	LYPD3	1.371908315	1.62647E-06	5.12515002	chr19
ENSG00000124731	TREM1	-0.7468616	0.000653874	2.39445021	chr6
ENSG00000125398	SOX9	0.703634384	0.000229829	4.49655539	chr17
ENSG00000125538	IL1B	1.097822001	4.96368E-07	2.65555916	chr2
ENSG00000125740	FOSB	0.687109972	0.000145582	4.57440644	chr19
ENSG00000125775	FKBP1A-SDCBP2	0.553691827	0.000128456	1.56216078	chr20
ENSG00000125851	PCSK2	0.992756554	0.00028241	4.17238824	chr20
ENSG00000127328	RAB3IP	0.929288498	5.23327E-09	2.80050843	chr12
ENSG00000127824	TUBA4A	0.83662561	2.05328E-05	4.49057743	chr2
ENSG00000127951	FGL2	-0.56742171	0.001568773	5.06238264	chr7
ENSG00000128564	VGF	-0.740015	0.006784174	8.52388799	chr7

ENSG00000129009	ISLR	-0.6813796	4.16093E-05	5.5478197	chr15
ENSG00000130600	H19	0.767899336	3.63052E-05	4.69927265	chr11
ENSG00000130707	ASS1	0.581848022	0.000568995	2.9229796	chr9
ENSG00000132692	BCAN	-0.66658436	0.006952738	7.63216909	chr1
ENSG00000133048	CHI3L1	-1.41224839	1.45521E-08	6.44473369	chr1
ENSG00000133063	CHIT1	0.854021701	0.003132249	4.95058763	chr1
ENSG00000133110	POSTN	0.691759504	0.000158642	8.18409081	chr13
ENSG00000133401	PDZD2	0.563903366	0.002387596	3.21907155	chr5
ENSG00000134755	DSC2	0.898155294	5.76606E-05	3.28073209	chr18
ENSG00000134827	TCN1	0.758325611	0.005314697	4.42076832	chr11
ENSG00000135744	AGT	-0.86148589	0.000397157	3.75605858	chr1
ENSG00000136689	IL1RN	1.042467779	4.2913E-08	3.24295828	chr2
ENSG00000136960	ENPP2	-0.53343418	0.00084511	6.3654225	chr8
ENSG00000137857	DUOX1	0.548595742	0.015061795	3.06104088	chr15
ENSG00000137877	SPTBN5	0.61712992	1.36286E-05	1.20639861	chr15
ENSG00000138615	CILP	0.760260297	0.000934389	3.85830731	chr15
ENSG00000138829	FBN2	0.601473576	0.001844572	3.06024418	chr5

ENSG00000140450	ARRDC4	-0.67987247	3.20799E-07	4.82396543	chr15
ENSG00000143127	ITGA10	-0.66844663	0.005517216	5.96788238	chr1
ENSG00000143248	RGS5	-1.03395879	5.73051E-07	6.73675945	chr1
ENSG00000143320	CRABP2	0.574556874	0.008292555	4.39396627	chr1
ENSG00000143546	S100A8	1.397999284	3.40243E-06	6.22739246	chr1
ENSG00000143631	FLG	-0.80835213	0.003197379	4.90313264	chr1
ENSG00000144810	COL8A1	-0.65056879	0.002043209	4.760291	chr3
ENSG00000145506	NKD2	0.784036843	2.35541E-06	1.47530292	chr5
ENSG00000146197	SCUBE3	-0.75242597	0.000688018	3.85874411	chr6
ENSG00000146648	EGFR	0.821112413	1.98973E-05	3.42828116	chr7
ENSG00000147394	ZNF185	0.777601964	2.07942E-07	3.55508615	chrX
ENSG00000147889	CDKN2A	0.565864606	0.008184193	4.52848771	chr9
ENSG00000148053	NTRK2	0.537107105	0.008664328	3.14943652	chr9
ENSG00000149090	PAMR1	-0.70131107	4.96867E-05	1.41652957	chr11
ENSG00000149269	PAK1	0.553827718	2.79593E-08	5.31663682	chr11
ENSG00000151790	TDO2	1.606793804	1.4305E-16	2.23495047	chr4
ENSG00000151892	GFRA1	-0.66434444	0.003825293	3.51394259	chr10

ENSG00000152207	CYSLTR2	0.876418425	0.000297408	2.93630952	chr13
ENSG00000157368	IL34	-1.41324091	9.2227E-16	2.31443637	chr16
ENSG00000157680	DGKI	-0.60222291	0.001719819	3.64617366	chr7
ENSG00000158869	FCER1G	-1.09166733	3.50631E-11	5.85213805	chr1
ENSG00000158887	MPZ	-0.5509898	0.023750418	6.67215892	chr1
ENSG00000159166	LAD1	0.830428662	0.000592981	4.34452121	chr1
ENSG00000159167	STC1	1.223929703	4.13492E-08	5.77184963	chr8
ENSG00000159871	LYPD5	1.165530744	3.07925E-09	2.59315821	chr19
ENSG00000160161	CILP2	0.607287947	0.002928305	2.53260869	chr19
ENSG00000162493	PDPN	-1.2948493	2.89453E-11	3.55961758	chr1
ENSG00000162706	CADM3	-1.44315326	2.6171E-09	3.40812449	chr1
ENSG00000162981	FAM84A	0.567597901	0.005346864	2.40740884	chr2
ENSG00000163017	ACTG2	-0.53948601	0.001033768	3.18501959	chr2
ENSG00000163220	S100A9	1.494214744	4.12407E-08	7.00217151	chr1
ENSG00000164220	F2RL2	1.010347905	3.20476E-05	3.51751311	chr5
ENSG00000164283	ESM1	0.608279152	0.00150639	3.16769105	chr5
ENSG00000164484	TMEM200A	0.568580162	0.000687913	1.45895379	chr6

ENSG00000164949	GEM	0.570975927	0.000257164	3.9491881	chr8
ENSG00000165434	PGM2L1	0.59517108	3.91904E-05	3.52346277	chr11
ENSG00000165474	GJB2	0.678168018	0.010447104	6.04155308	chr13
ENSG00000166145	SPINT1	0.506303918	0.003636056	4.33310338	chr15
ENSG00000166250	CLMP	0.72608358	3.68844E-05	3.62638718	chr11
ENSG00000166825	ANPEP	-0.72296202	0.000239667	5.06782754	chr15
ENSG00000167244	IGF2	-0.66251621	0.001647956	5.30867817	chr11
ENSG00000167619	TMEM145	-0.50530832	0.00110057	1.65764559	chr19
ENSG00000168309	FAM107A	0.697688142	9.6356E-06	1.62687716	chr3
ENSG00000168621	GDNF	-0.57063425	0.017333429	3.32526459	chr5
ENSG00000168754	FAM178B	-0.52143912	0.02362578	4.48715244	chr2
ENSG00000169184	MN1	0.706872512	1.05925E-05	1.6233622	chr22
ENSG00000169252	ADRB2	0.762218913	1.71139E-07	0.89735221	chr5
ENSG00000169429	CXCL8	0.916677755	9.23522E-07	3.70830017	chr4
ENSG00000170545	SMAGP	0.676147773	5.35206E-06	2.06830274	chr12
ENSG00000170745	KCNS3	-0.50501246	0.001107014	3.38544121	chr2
ENSG00000172382	PRSS27	0.878206839	8.95636E-08	1.37734612	chr16

ENSG00000173156	RHOD	0.534760144	0.009589335	2.32339554	chr11
ENSG00000173557	C2orf70	-0.51190495	0.009403362	1.11694271	chr2
ENSG00000173801	JUP	0.626956628	0.000480752	6.76631201	chr17
ENSG00000176402	GJC3	-0.80906862	1.40927E-06	1.32160414	chr7
ENSG00000176971	FIBIN	0.727894117	7.84389E-05	1.46878903	chr11
ENSG00000177685	CRACR2B	0.514875482	0.000588829	1.18699799	chr11
ENSG00000179862	CITED4	0.566838259	0.001891891	2.39609101	chr1
ENSG00000181617	FDCSP	1.205536021	2.66717E-05	4.8397211	chr4
ENSG00000184254	ALDH1A3	-0.84288622	6.0848E-05	6.86332038	chr15
ENSG00000185008	ROBO2	-0.7954025	0.001379939	2.65370411	chr3
ENSG00000186395	KRT10	-0.66358578	0.00906033	7.72467742	chr17
ENSG00000187240	DYNC2H1	0.534778618	6.02592E-07	3.94836778	chr11
ENSG00000188015	S100A3	0.728176375	3.75798E-06	1.32248138	chr1
ENSG00000188488	SERPINA5	-0.54838384	0.025127397	4.57800197	chr14
ENSG00000188643	S100A16	0.582809265	0.000146088	6.87673682	chr1
ENSG00000189223	PAX8-AS1	0.547004342	0.002394591	2.75869267	chr2
ENSG00000196154	S100A4	0.57162728	0.001551987	7.23204339	chr1

ENSG00000196376	SLC35F1	-0.63029461	0.00282155	3.46519973	chr6
ENSG00000196611	MMP1	0.999247246	0.000250377	5.18681846	chr11
ENSG00000196754	S100A2	1.578086073	4.47855E-09	4.71565084	chr1
ENSG00000197467	COL13A1	0.966111717	1.78858E-09	1.69753975	chr10
ENSG00000204622	HLA-J	-0.50632267	0.001507464	2.14065691	chr6
ENSG00000211445	GPX3	0.807167757	6.80816E-05	6.53910232	chr5
ENSG00000214548	MEG3	1.327177834	6.84231E-09	3.1616005	chr14
ENSG00000224389	C4B	-0.60624492	0.000425305	1.79240473	chr6_ssto_hap7
ENSG00000240583	AQP1	-0.61732284	0.000271975	6.65647434	chr7
ENSG00000241644	INMT	-0.51113124	0.001612754	1.17322174	chr7
ENSG00000244734	HBB	0.574133606	0.0085478	3.08136072	chr11
ENSG00000246174	KCTD21-AS1	0.676131884	1.0853E-07	1.17716763	chr11
ENSG00000254122	PCDHGB7	0.656324981	0.000406365	2.07653998	chr5
ENSG00000272398	CD24	0.819576936	0.000790708	3.10845681	chrY
ENSG00000275395	FCGBP	-0.50383485	0.009414201	3.51819475	chr19

Οι στήλες «SYMBOL», «LogFC», «p-value», χρησιμοποιήθηκαν σαν είσοδος για το εργαλείο BioInfoMiner, το οποίο οδήγησε σε μια λίστα με γονίδια σε προτεραιότητα, στην ανάλυση

τον οντολογιών στις οποίες εμπλέκονται τα γονίδια ενδιαφέροντος και στις αντίστοιχες επιλογές οπτικοποίησης που παρέχει το εργαλείο.

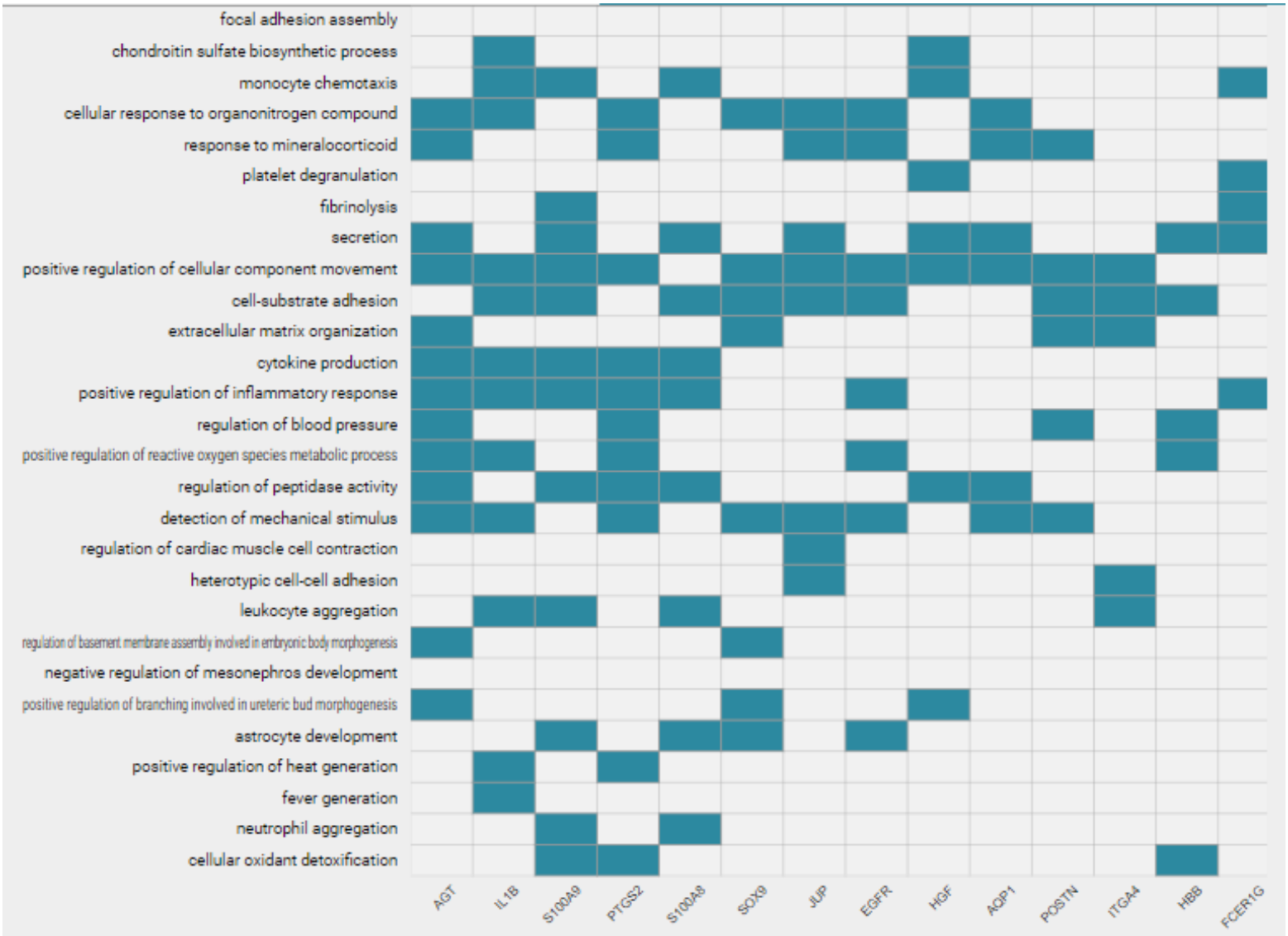
Gene Prioritization

Rank ↑	Gene Symbol	Definition	Clusters	Interactions	Fold Change	Pvalue
1	AGT	angiotensinogen	13	0	-0.86	0.000397
2	IL1B	interleukin 1 beta	12	0	1.1	4.96e-7
3	S100A9	S100 calcium binding protein A9	12	2	1.49	4.12e-8
4	PTGS2	prostaglandin-endoperoxide synthase 2	11	0	0.83	0.000021
5	S100A8	S100 calcium binding protein A8	9	1	1.4	0.0000034
6	SOX9	SRY-box 9	8	0	0.7	0.00023
7	JUP	junction plakoglobin	8	2	0.63	0.000481
8	EGFR	epidermal growth factor receptor	8	10	0.82	0.0000199
9	HGF	hepatocyte growth factor	7	1	-1.06	3.99e-7
10	AQP1	aquaporin 1 (Colton blood group)	6	0	-0.62	0.000272
11	POSTN	periostin	6	0	0.69	0.000159
12	ITGA4	integrin subunit alpha 4	5	1	-0.51	0.002
13	HBB	hemoglobin subunit beta	5	0	0.57	0.009
14	FCER1G	Fc fragment of IgE receptor Ig	5	0	-1.09	3.51e-11
15	RHOD	ras homolog family member D	3	1	0.53	0.01

Εικόνα 24 Πίνακας με διαφορικά εκφραζόμενα γονίδια σε προτεραιότητα μεταξύ των ασθενών που φέρον BRAF μεταλλαγή και των ασθενών που δεν φέρουν την εν λόγω μεταλλαγή. Αποτέλεσμα της χρήσης του εργαλείου BioInfoMiner για γονιδιακό και οντολογικό εμπλουτισμό.

Είναι προφανές ότι εμπλουτίζονται γονίδια τα οποία σχετίζονται με τις διαδικασίες φλεγμονής και ανοσολογικής απόκρισης, την αγγειογένεση, την κυτταρική αύξηση, καθώς και λειτουργίες του κυτταροσκελετού, της κυτταρικής και βιολογικής προσκόλλησης.

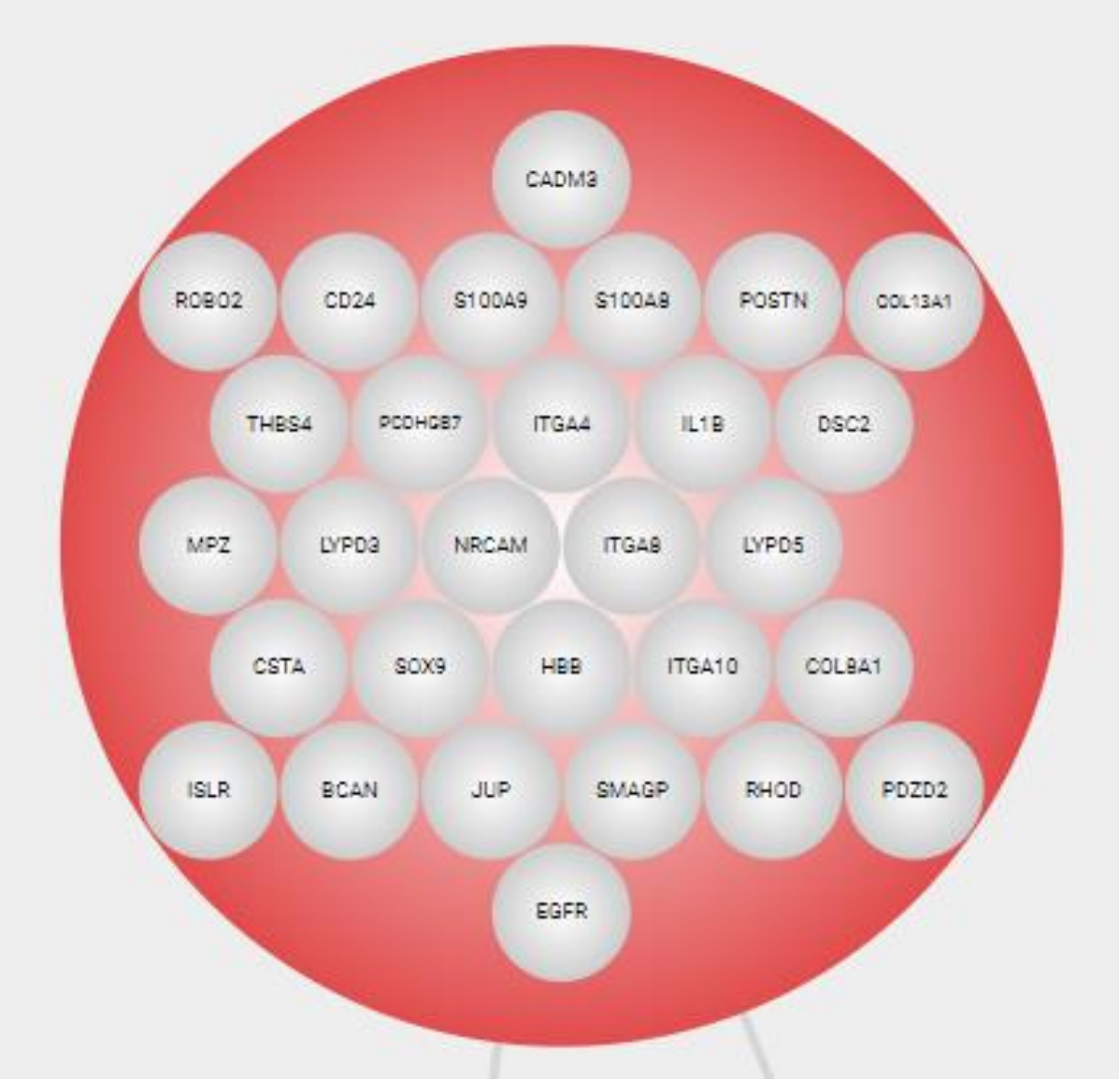
Το παρακάτω διάγραμμα δείχνει τον εμπλουτισμό των διαδικασιών ανάλογα με τα γονίδια σε προτεραιότητα. Κάθε μπλε κελί αναπαριστά την διαφορική έκφραση του εκάστοτε γονιδίου που σχετίζεται με την εν λόγω διαδικασία της αριστερής στήλης του διαγράμματος.



Εικόνα 25Διαγραμματική απεικόνιση του BioInfoMiner των διαφορικά εκφραζόμενων γονιδίων που έχουν τεθεί σε προτεραιότητα από το εργαλείο καθώς και των των αντίστοιχων εμπλουτισμένων οντολογιών, στην περίπτωση της σύγκρισης ασθενών που φέρουν BRAF μεταλλαγή και αυτών που δεν φέρουν μεταλλαγή στο εν λόγω γονίδιο.

Σε συνέχεια του παραπάνω διαγράμματος, παρέχεται και απεικόνιση τύπου δενδρογράμματος, με υπόδειξη των οντολογιών ενδιαφέροντος με κόκκινο χρώμα και με δυνατότητα μελέτης του μονοπατιού που η κάθεμια από αυτές συμμετέχει. Το εν λόγω διάγραμμα είναι αρκτεά σύνθετο και δεν έχει νόημα να παρουσιαστεί σε στατική μορφή. Παρόλα αυτά, ένα ενδιαφέρον χαρακτηριστικό του είναι η δυνατότητα να επικεντρωθεί κανείς σε κάποια από τις οντολογίες ενδιαφέροντος και επιλέγοντας την, να δει τα γονίδια

που υπήρχαν στο αρχικό σετ δεδομένων και σχετίζονται με αυτή την κυτταρική λειτουργία. Ενδεικτικά παρατίθεται η παρακάτω εικόνα για τη διαδικασία της κυτταρικής προσκόλλησης.



Εικόνα 26 Ομαδοποίηση γονιδίων που σχετίζονται με τη διαδικασία της κυτταρικής προσκόλλησης, από το αρχικό σετ διαφορικά εκφρασμένων γονιδίων που αφορά ασθενείς που φέρουν BRAF μεταλλαγή και ασθενείς που δε φέρουν την εν λόγω μεταλλαγή.

6 Συμπεράσματα

Τα αποτελέσματα που προέκυψαν από την τελευταία σύγκριση, μεταξύ της ομάδας ασθενών που πάσχουν από μελάνωμα και φέρουν μεταλλαγή στο γονίδιο BRAF και ασθενών που πάσχουν από μελάνωμα και δεν φέρουν μεταλλαγή στο γονίδιο BRAF, είναι και αυτά που

οδηγούν στα πιο ενδιαφέροντα συμπεράσματα και αναλύονται παρακάτω. Να σημειωθεί πως, το BRAF είναι ένα γονίδιο που συσχετίστηκε πρόσφατα με την εμφάνιση δερματικού μελανώματος και εμφανίζεται μεταλλαγμένο σε περίπου 50% των ασθενών, ενώ το ποσοστό αυτό συμφωνεί και με τα δεδομένα που χρησιμοποιήθηκαν και αναλύθηκαν σε αυτή την εργασία.

Από τους συνολικούς 415 ασθενείς για τους οποίους υπήρχαν δεδομένα μεταλλάξεων και αντλήθηκαν από τη βάση cBioportal, το 48% φέρει BRAF μεταλλαγή, εν αντιθέσει με το 52%. Από το σύνολο αυτό και την ανάλυσή του προέκυψε μια λίστα 140 γονιδίων των οποίων έγινε εισαγωγή στο εργαλείο BioInfoMiner ώστε να προκύψει μια ιεραρχημένη βάσει εμπλουτισμού λίστα με τα σημαντικότερα γονίδια που εμφανίζουν διαφορικά εκφρασμένα μετάγραφα.

Οι κυριότερες κυτταρικές διεργασίες που επηρεάζονται είναι η κυτταρική προσκόλληση και κίνηση, οι λειτουργίες του κυτταροσκελετού, ο μηχανισμός της νοσολογικής αντίδρασης και της φλεγμονής και η αγγειογένεση με χαρακτηριστικές περιπτώσεις γονιδίων S100A9, S100A8 – calcium binding proteins integrins (κυτταρική και βιολογική προσκόλληση), IL1B – interleukin 1 beta (μηχανισμός φλεγμονής), ATG – angiotensinogen (αγγειογένεση).

Η IL1B ιντερλευκίνη υπερ-εκφράζεται στους ασθενείς που φέρουν BRAF μεταλλαγή και οδηγεί στο συμπεράσμα πως στην ομάδα αυτή ασθενών οι μηχανισμοί φλεγμονής και ανοσολογικής απόκρισης είναι εντονότεροι. Παράλληλα, δεδομένου ότι οι ασθενείς BRAF τείνουν να είναι νεότεροι (Akbani et al., 2015), θα μπορούσε η ηλικία να συσχετιστεί και να αναλυθεί ταυτόχρονα με τις άλλες δύο παραμέτρους ως μια επιπλέον παράμετρος, για να δειχθεί αν η ανάλυση οδηγεί σε παραπλήσια συμπεράσματα.

Από την άλλη πλευρά, η υπερ-έκφραση της ATG αγγειοτενσίνης στους nonBRAF ασθενείς, καταδεικνύει την εντονότερη αγγειογένεση σε αυτή την ομάδα. Η ATG σχετίζεται, εκτός από τη φυσιολογική αγγειογένεση και με την αγγείωση και αιμάτωση του όγκου, κάνοντας έτσι την εγκαθίδρυσή του πιο έντονη και την αντιμετώπιση από το ανοσοποιητικό πιο δύσκολη, καθώς αυξάνεται σε μέγεθος, εμβαθύνει στα στρώματα του δέρματος και επεκτείνεται σε παρακείμενους ιστούς. Στην περίπτωση λοιπόν της υπερέκφρασης του συγκεκριμένου

γονιδίου, αναφερόμαστε σε πιο διεισδυτικές μορφές μελανώματος, με γρηγορότερη ανάπτυξη.

Τέλος, ενδιαφέρον παρουσιάζει η υπερ-έκφραση του HGF αυξητικό παράγοντα στους nonBRAF ασθενείς και η υπερ-έκφραση του EGFR υποδοχέα στους BRAF ασθενείς. Αν αναμορφωθεί αυτή την πρόταση, φαίνεται πως οι ασθενείς με μεταλλαγή BRAF εκφράζουν εντονότερα τον υποδοχέα του δερματικού αυξητικού παράγοντα (EGFR), ενώ ταυτόχρονα, μειώνεται η έκφραση του ηπατικού αυξητικού παράγοντα (HGF). Μια πιθανή εξήγηση είναι πως στους BRAF ασθενείς, η δερματική επέκταση του μελανώματος συμβαίνει με γρήγορους ρυθμούς, όμως ο όγκος είναι επιφανειακός και μη διεισδυτικός. Παράλληλα, δεδομένου ότι το μελάνωμα συχνά παρουσιάζει μεταστάσεις στο ήπαρ, είναι αξιοσημείωτο ότι, σε αυτή την περίπτωση, ο HGF είναι μειωμένος, οπότε ίσως οι ασθενείς BRAF να παρουσιάζουν μικρότερη πιθανότητα εμφάνισης μετάστασης στο ήπαρ.

Η ακριβής βιολογική σημασία όλων των παραπάνω θα μπορούσε να δειχθεί αποτελεσματικά με την επανάληψη πειραμάτων, δοκιμές διαφορετικών τιμών για τις παραμέτρους και καλύτερη μελέτη του βιολογικού υπόβαθρου της ασθένειας, ενέργειες που δεν εμπίπτουν στο εύρος της παρούσας διπλωματικής εργασίας, αποτελούν όμως προτάσεις για μελλοντική διερεύνηση.

7 Βιβλιογραφικές αναφορές

Akbani, R., Akdemir, K. C., Aksoy, B. A., Albert, M., Ally, A., Amin, S. B., … Zou, L. (2015). Genomic Classification of Cutaneous Melanoma. *Cell*, *161*(7), 1681–1696. https://doi.org/10.1016/j.cell.2015.05.044

Allen, A. C., & Spitz, S. (1953). Malignant melanoma. A clinicopathological analysis of the criteria for diagnosis and prognosis. *Cancer*, *6*(1), 1–45. https://doi.org/10.1002/1097-0142(195301)6:1<1::AID-CNCR2820060102>3.0.CO;2-C

Anders, S., Reyes, A., & Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Research*, *22*(10), 2008–2017. https://doi.org/10.1101/gr.133744.111

Azoury, S. C., & Lange, J. R. (2014). Epidemiology, Risk Factors, Prevention, and Early Detection of Melanoma. *Surgical Clinics of North America*, *94*(5), 945–962. https://doi.org/10.1016/j.suc.2014.07.013

Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. (n.d.). Retrieved March 3, 2019, from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bioinformatics Pipeline: mRNA Analysis - GDC Docs. (n.d.). Retrieved March 3, 2019, from https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/

Bird, A. (2007). Perceptions of epigenetics. *Nature*, *447*(7143), 396–398. https://doi.org/10.1038/nature05913

Bodenham, D. C. (1968). A study of 650 observed malignant melanomas in the South-West region. *Annals of the Royal College of Surgeons of England*, *43*(4), 218–239. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/5698493

Bolstad, B. M., Irizarry, R. A., Astrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)*, *19*(2), 185–193. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12538238

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., … Schultz, N. (2012). The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data: Figure 1. *Cancer Discovery*, *2*(5), 401–404. https://doi.org/10.1158/2159-8290.CD-12-0095

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., … Noushmehr, H. (2016). *TCGAbiolinks* : an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*, *44*(8), e71–e71. https://doi.org/10.1093/nar/gkv1507

Cooper, S. (1844). *The First Lines of the Theory and Practice of Surgery: Including the Principle Operations*. Retrieved from https://www.google.com/books?hl=en&lr=&id=EW1GAQAAMAAJ&oi=fnd&pg=PA13&dq=+The+First+Lines+of+the+Theory+and+Practice+of+Surgery:+Including+the+Principle

+Operations&ots=XOyQqtX9h5&sig=QqUZfTJA6e33ys_-C9zR4CdAS5o

Daniel, R. (2005). The metagenomics of soil. *Nature Reviews Microbiology*, *3*(6), 470–478. https://doi.org/10.1038/nrmicro1160

Dobin, A. (2014). *STAR manual 2.4.0.1*. Retrieved from https://github.com/alexdobin/

Dorland, W. A. N. (William A. N. (2003). *Dorland's illustrated medical dictionary.* Saunders.

Dudoit, S., Shaffer, J. P., & Block, J. C. (2003). Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science*, *18*(1), 71–103. https://doi.org/10.1214/ss/1056397487

EMBL-European Molecular Biology Laboratory. (n.d.). Retrieved from https://www.embl.de/

Feldmann, H., Aigle, M., Aljinovic, G., André, B., Baclet, M. C., Barthe, C., … Kleine, K. (1994). Complete DNA sequence of yeast chromosome II. *The EMBO Journal*, *13*(24), 5795–5809. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7813418

Fiehn, O. (2002). Metabolomics — the link between genotypes and phenotypes. In *Functional Genomics* (pp. 155–171). https://doi.org/10.1007/978-94-010-0448-0_11

Finotello, F., & Di Camillo, B. (2015). Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Briefings in Functional Genomics*, *14*(2), 130–142. https://doi.org/10.1093/bfgp/elu035

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., … Schultz, N. (2013). Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Science Signaling*, *6*(269), pl1–pl1. https://doi.org/10.1126/scisignal.2004088

GBD 2015 Disease and Injury Incidence and Prevalence Collaborators, G. 2015 D. and I. I. and P. (2016). Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet (London, England)*, *388*(10053), 1545–1602. https://doi.org/10.1016/S0140-6736(16)31678-6

Goldstein, B. G., & Goldstein, A. O. (2001). Diagnosis and management of malignant melanoma. *American Family Physician*, *63*(7), 1359–1368, 1374. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11310650

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., & Staudt, L. M. (2016). Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine*, *375*(12), 1109–1112. https://doi.org/10.1056/NEJMp1607591

Hadley Wickham, Romain François, L. H. and K. M. (2018). *dplyr: A Grammar of Data Manipulation. R package version 0.7.6.*

Hansen, K. D., Irizarry, R. A., & WU, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*, *13*(2), 204–216. https://doi.org/10.1093/biostatistics/kxr054

Henry, H. W. and L. (2018). *tidyr: Easily Tidy Data with "spread()" and "gather()" Functions. R*

package version 0.8.2.

Hieter, P., & Boguski, M. (1997). Functional genomics: it's all how you read it. *Science (New York, N.Y.)*, *278*(5338), 601–602. https://doi.org/10.1126/SCIENCE.278.5338.601

Human Genome Project Completion: Frequently Asked Questions - National Human Genome Research Institute (NHGRI). (n.d.). Retrieved March 3, 2019, from https://www.genome.gov/11006943/

International Agancy for Research on Cancer, W. H. O. (n.d.). World Cancer Report – IARC, "The global burden of cancer." Retrieved March 1, 2019, from https://www.iarc.fr/cards_page/world-cancer-report/

Jirtle, R. L., & Skinner, M. K. (2007). Environmental epigenomics and disease susceptibility. *Nature Reviews Genetics*, *8*(4), 253–262. https://doi.org/10.1038/nrg2045

Johns Hopkins University. School of Hygiene and Public Health., & Society for Epidemiologic Research (U.S.). (n.d.). *American journal of epidemiology.* School of Hygiene and Public Health of the Johns Hopkins University.

Kallipos Repository: Home. (n.d.). Retrieved March 2, 2019, from https://repository.kallipos.gr/?locale=en

Kanavy, H. E., & Gerstenblith, M. R. (2011). Ultraviolet Radiation and Melanoma. *Seminars in Cutaneous Medicine and Surgery*, *30*(4), 222–228. https://doi.org/10.1016/j.sder.2011.08.003

Kontogianni, G., Piroti, G., Maglogiannis, I., Chatziioannou, A., & Papadodima, O. (2018). Dissecting the Mutational Landscape of Cutaneous Melanoma: An Omic Analysis Based on Patients from Greece. *Cancers*, *10*(4), 96. https://doi.org/10.3390/cancers10040096

Koutsandreas, T., Binenbaum, I., Pilalis, E., Valavanis, I., Papadodima, O., & Chatziioannou, A. (2016). Analyzing and visualizing genomic complexity for the derivation of the emergent molecular networks. *International Journal of Monitoring and Surveillance Technologies Research (IJMSTR)*.

Laennec RTH. (1806). Sur les melanoses. *Bulletin de La Faculte de Medecine de Paris*, *1*, 24–16.

Lee, Y.-S., Poh, L. K.-S., & Loke, K.-Y. (2002). A Novel Melanocortin 3 Receptor Gene (*MC3R*) Mutation Associated with Severe Obesity. *The Journal of Clinical Endocrinology & Metabolism*, *87*(3), 1423–1426. https://doi.org/10.1210/jcem.87.3.8461

Lockhart, D. J., & Winzeler, E. A. (2000). Genomics, gene expression and DNA arrays. *Nature*, *405*(6788), 827–836. https://doi.org/10.1038/35015701

McCain, J. (2013). The MAPK (ERK) Pathway: Investigational Combinations for the Treatment Of BRAF-Mutated Metastatic Melanoma. *P & T : A Peer-Reviewed Journal for Formulary Management*, *38*(2), 96–108. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23599677

Melanoma of the Skin - Cancer Stat Facts. (n.d.). Retrieved March 1, 2019, from

https://seer.cancer.gov/statfacts/html/melan.html

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621–628. https://doi.org/10.1038/nmeth.1226

National Human Genome Research Institute, N. (2016). An Overview of the Human Genome Project - National Human Genome Research Institute (NHGRI). *NIH*. Retrieved from https://www.genome.gov/12011238/an-overview-of-the-human-genome-project/

National Institutes of Health (NIH) | Turning Discovery Into Health. (n.d.). Retrieved May 8, 2019, from https://www.nih.gov/

Norris, W. (1820). Case of Fungoid Disease. *Edinburgh Medical and Surgical Journal*, *16*(65), 562–565. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/30332089

Olsvik, O., Wahlberg, J., Petterson, B., Uhlén, M., Popovic, T., Wachsmuth, I. K., & Fields, P. I. (1993). Use of automated sequencing of polymerase chain reaction-generated amplicons to identify three types of cholera toxin subunit B in Vibrio cholerae O1 strains. *Journal of Clinical Microbiology*, *31*(1), 22–25. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7678018

Pandey, A., & Mann, M. (2000). Proteomics to study genes and genomes. *Nature*, *405*(6788), 837–846. https://doi.org/10.1038/35015709

Pettersson, E., Lundeberg, J., & Ahmadian, A. (2009). Generations of sequencing technologies. *Genomics*, *93*(2), 105–111. https://doi.org/10.1016/j.ygeno.2008.10.003

Picard Tools - By Broad Institute. (n.d.). Retrieved March 3, 2019, from http://broadinstitute.github.io/picard/

Prober, J. M., Trainor, G. L., Dam, R. J., Hobbs, F. W., Robertson, C. W., Zagursky, R. J., … Baumeister, K. (1987). A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science (New York, N.Y.)*, *238*(4825), 336–341. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2443975

Rasmussen, S. A., & Friedman, J. M. (2000). NF1 Gene and Neurofibromatosis 1. *American Journal of Epidemiology*, *151*(1), 33–40. https://doi.org/10.1093/oxfordjournals.aje.a010118

Reuter, J. A., Spacek, D. V., & Snyder, M. P. (2015). High-Throughput Sequencing Technologies. *Molecular Cell*, *58*(4), 586–597. https://doi.org/10.1016/J.MOLCEL.2015.05.004

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/271968

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., … Hood, L. E. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature*, *321*(6071), 674–679. https://doi.org/10.1038/321674a0

Team, B. C. (2015). *Homo.sapiens: Annotation package for the Homo.sapiens object. R package version 1.3.1.*

UniProt Consortium, T. (2018). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, *46*(5), 2699–2699. https://doi.org/10.1093/nar/gky092

Uzman, A. (2003). Molecular biology of the cell (4th ed.): Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. *Biochemistry and Molecular Biology Education*, *31*(4), 212–214. https://doi.org/10.1002/bmb.2003.494031049999

Vitkup, D., Melamud, E., Moult, J., & Sander, C. (2001). Completeness in structural genomics. *Nature Structural Biology*, *8*(6), 559–566. https://doi.org/10.1038/88640

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Weber, J. L., & Myers, E. W. (1997). Human whole-genome shotgun sequencing. *Genome Research*, *7*(5), 401–409. https://doi.org/10.1101/GR.7.5.401