



**ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ  
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ  
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ  
ΒΙΟΙΑΤΡΙΚΗ**

Ανάλυση πειραμάτων Knockout CRISPR που έχουν διεξαχθεί σε καρκινικές κυτταρικές σειρές A375 μελανώματος για την ανίχνευση απαραίτητων γονιδίων και γονιδίων υπευθύνων για ανθεκτικότητα σε φάρμακα.

**Καμζόλα Αναστασία**

**ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ  
Υπεύθυνος  
Μπάγκος Παντελεήμων  
Καθηγητής**

Λαμία, 2019





**ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ  
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ  
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ  
ΒΙΟΙΑΤΡΙΚΗ**

Ανάλυση πειραμάτων Knockout CRISPR που έχουν διεξαχθεί σε καρκινικές κυτταρικές σειρές A375 μελανώματος για την ανίχνευση απαραίτητων γονιδίων και γονιδίων υπευθύνων για ανθεκτικότητα σε φάρμακα.

**Καμζόλα Αναστασία**

**ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ**

**Επιβλέποντες  
Μπάγκος Παντελεήμων, Καθηγητής  
Πετσαλάκη Ευαγγελία, ΕΒΙ**

**Λαμία, 2019**

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις <sup>(1)</sup>, που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί **χωρίς να τα περικλείω σε εισαγωγικά** και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάστηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.
2. Δέχομαι ότι η αυτολεξεί **παράθεση χωρίς εισαγωγικά**, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.
3. Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια
4. Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.

Ημερομηνία: ...../...../20.....

Ο – Η Δηλ.

(Υπογραφή)

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

Ανάλυση πειραμάτων Knockout CRISPR που έχουν διεξαχθεί σε καρκινικές κυτταρικές σειρές A375 μελανώματος για την ανίχνευση απαραίτητων γονιδίων και γονιδίων υπευθύνων για ανθεκτικότητα σε φάρμακα.

Καμζόλα Αναστασία

**Τριμελής Επιτροπή:**

Μπάγκος Παντελεήμων, Καθηγητής

Μπράλιου Γεωργία, Επίκουρος Καθηγήτρια

Χατζηγεωργίου Άρτεμις, Καθηγήτρια



Η παρούσα πτυχιακή εργασία έγινε σε συνεργασία με το European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI, Cambridge, United Kingdom )



## Περιεχόμενα

Κατάλογος εικόνων.....	10
Κατάλογος πινάκων .....	11
Περίληψη .....	13
Abstract.....	15
Κεφάλαιο 1: Εισαγωγή .....	17
1.1. Το CRISPR-Cas9 και η λειτουργία του .....	17
1.1.1.Μηχανισμοί επιδιόρθωσης του κυττάρου .....	17
1.1.2. Εφαρμογές του CRISPR: .....	18
1.2. Τί είναι τα CRISPR screens .....	19
1.2.1. Essential (απαραίτητα) γονίδια .....	20
1.2.2. Negative selection screens .....	21
1.2.3. Positive selection screens.....	21
1.3. Η χρήση των CRISPR για την μελέτη των μονοπατιών κυτταρικής σηματοδότησης.....	21
1.4. Μελάνωμα.....	23
Κεφάλαιο 2: Στόχος και κίνητρα .....	25
Στόχοι της πτυχιακής μου εργασίας:.....	25
Κεφάλαιο 3: Υλικά και Μέθοδοι .....	26
3.1. Τα προγραμματισμένα βήματα της εργασίας μου είναι:.....	26
3.2. Συλλογή δεδομένων .....	28
3.3. ΜΕΘΟΔΟΙ/Επεξεργασία δεδομένων.....	29
3.3.2 ΛΙΓΑ ΛΟΓΙΑ ΓΙΑ ΤΟ MAGeCK .....	29
Μέθοδοι/Επεξεργασία δεδομένων (συνέχεια) .....	31
3.4.1. Ανάλυση των negative selection screens .....	32
3.4.2. Ανάλυση των positive selection screens .....	34
Κεφάλαιο 4: Αποτελέσματα .....	35
4.1. Αποτελέσματα της ανάλυσης των negative selection screens στις A375 κυτταρικές σειρές του μελανώματος:.....	35
α)Αξιολόγηση της απόδοσης κάθε βιβλιοθήκης στο να αναγνωρίζει γνωστά essential και non-essential (απαραίτητα και μη-απαραίτητα) γονίδια .....	35
β)Αξιολόγηση της απόδοσης κάθε βιβλιοθήκης στο να αναγνωρίζει τα γνωστά essential μονοπάτια (pathways).....	37
γ)Εύρεση των essential γονιδίων στις A375 κυτταρικές σειρές. ....	38
4.2. Αποτελέσματα των positive selection screens στις A375 κυτταρικές σειρές του μελανώματος. ....	41



Εύρεση των γονιδίων που είναι ανθεκτικά στους MEK και BRAF αναστολείς. ....	41
Κεφάλαιο 5: Συζήτηση-Συμπεράσματα-Βιβλιογραφία .....	43
5.1. Συζήτηση-Συμπεράσματα.....	43
5.2. Βιβλιογραφία .....	44
ΠΑΡΑΡΤΗΜΑ.....	49
ΛΙΣΤΕΣ ΜΕ ΤΑ ΓΟΝΙΔΙΑ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΗΣΑ .....	49
ΚΩΔΙΚΕΣ ΓΙΑ ΤΗΝ ΑΝΑΛΥΣΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ .....	55



## Κατάλογος εικόνων

<b>Εικόνα 1:</b> Επεξεργασία του γονιδιώματος μέσω του μηχανισμού CRISPR-Cas9 και επιδιόρθωση της βλάβης στην διπλή έλικα του DNA .	<b>1818</b>
<b>Εικόνα 2:</b> Positive και negative selection screens	<b>20</b>
<b>Εικόνα 3:</b> Ενεργοποίηση του MAPK σηματοδοτικού μονοπατιού μέσω της BRAFV600E μετάλλαξης.....	<b>22</b>
<b>Εικόνα 4:</b> Διάγραμμα ροής των βημάτων για την εκπόνησης της πτυχιακής εργασίας μου. .	<b>27</b>
<b>Εικόνα 5:</b> Αναγνώριση των essential γονιδίων με βάση το p-value και τον a-RRA αλγόριθμο	<b>299</b>
<b>Εικόνα 6:</b> Παράδειγμα χρήσης του προγράμματος MAGeCK για την επεξεργασία της μελέτης με pubmed id 24336571. ....	<b>30</b>
<b>Εικόνα 7:</b> Έξοδος του MAGeCK.....	<b>300</b>
<b>Εικόνα 8:</b> Παράδειγμα .gene_summary.txt αρχείου. ....	<b>31</b>
<b>Εικόνα 9:</b> Διαγράμματα πυκνότητας για την απεικόνιση των αποτελεσμάτων ανά μελέτη και βιβλιοθήκη. ....	<b>36</b>
<b>Εικόνα 10:</b> Καμπύλες ROC για την οπτικοποίηση των αποτελεσμάτων ανά μελέτη και βιβλιοθήκη. ....	<b>37</b>
<b>Εικόνα 11:</b> Heatmaps για την οπτικοποίηση της απόδοσης των βιβλιοθηκών.....	<b>38</b>
<b>Εικόνα 12:</b> Γονίδια του MAPK σηματοδοτικού μονοπατιού τα οποία είναι απαραίτητα στην A375 κυτταρική σειρά του μελανώματος. ....	<b>399</b>
<b>Εικόνα 13:</b> Essential γονίδια στο MAPK σηματοδοτικό μονοπάτι	<b>39</b>
<b>Εικόνα 14:</b> Heatmap για το TGF beta σηματοδοτικό μονοπάτι	<b>40</b>
<b>Εικόνα 15:</b> Heatmap για το MTOR σηματοδοτικό μονοπάτι	<b>40</b>
<b>Εικόνα 16:</b> Heatmap για το Melanoma σηματοδοτικό μονοπάτι	<b>41</b>
<b>Εικόνα 17:</b> Heatmap με όλα τα γονίδια (γνωστά και μη) που είναι ανθεκτικά στους MEK και BRAF αναστολείς	<b>42</b>
<b>Εικόνα 18:</b> Heatmap μετά την αφαίρεση των γνωστών από πριν γονιδίων.....	<b>42</b>

## **Κατάλογος πινάκων**

<b>Πίνακας 1:</b> CRISPR-Cas9 sgRNA βιβλιοθήκες με κάποια χαρακτηριστικά τους. ....	<b>28</b>
<b>Πίνακας 2:</b> Μελέτες ανά βιβλιοθήκη στην negative selection. ....	<b>32</b>
<b>Πίνακας 3:</b> Μελέτες ανά βιβλιοθήκη στην positive selection. ....	<b>34</b>



## Περίληψη

Σκοπός της παρούσας πτυχιακής εργασίας είναι η μετα-ανάλυση δεδομένων από CRISPR screens που έχουν διεξαχθεί στις A375 καρκινικές κυτταρικές σειρές του μελανώματος για την ανακάλυψη essential (απαραίτητων) γονιδίων και γονιδίων ανθεκτικών σε συγκεκριμένα φάρμακα.

Αφού συνέλλεξα τα δεδομένα από τις CRISPR screens, τα οποία αφορούν κυρίως βιβλιοθήκες sgRNAs, τα επεξεργάστηκα έτσι ώστε να εντοπίσω την καλύτερη βιβλιοθήκη που μπορεί να χρησιμοποιηθεί για screening, βρήκα ένα σύνολο γονιδίων essential (απαραίτητων) στις A375 κυτταρικές σειρές και σηματοδοτικά μονοπάτια, και τέλος εντόπισα τα γονίδια που προσφέρουν αντίσταση στους MEK και BRAF αναστολείς. Αυτοί οι αναστολείς χρησιμοποιούνται συνήθως στο μελάνωμα.

Συνοπτικά, τα βασικά βήματα της ανάλυσης είναι τα εξής:

1. Συλλογή δεδομένων από τις A375 κυτταρικές σειρές του μελανώματος.
2. Επεξεργασία των δεδομένων μέσω του στατιστικού πακέτου MAGeCK.
3. Περαιτέρω ανάλυση των δεδομένων χρησιμοποιώντας την γλώσσα προγραμματισμού R.
4. Οπτικοποίηση των αποτελεσμάτων με διαγράμματα πυκνότητας, καμπύλες ROC και heatmaps.

Όλα τα καινούργια ευρήματα θα χρησιμοποιηθούν στο μέλλον από το group του EBI.

Λέξεις-Φράσεις κλειδιά: CRISPR-Cas9, CRISPR screens, sgRNA, A375 κυτταρικές σειρές, μελάνωμα, essential genes (απαραίτητα γονίδια), MEK-BRAF αναστολείς.



## Abstract

The purpose of this thesis is the meta-analysis of data from CRISPR screens conducted in A375 melanoma cancer cell lines for the identification of essential and drug resistance genes.

After collecting the data from the CRISPR screens, which mainly conclude sgRNA libraries, I will process them to identify the best library that can be used for screening, find a set of essential genes in the A375 cell lines and signaling pathways, and finally I will identify genes that cause resistance to MEK and BRAF inhibitors. These inhibitors are usually used in melanoma.

More specifically, the main steps of this study are the following:

1. Data collection from the A375 melanoma cell lines.
2. Data processing through the MAGeCK statistical package.
3. Further data analysis using the R programming language.
4. Visualization of the results using density plots, ROC curves and heatmaps.

All the new findings are going to be used in the future by the EBI group.





## Κεφάλαιο 1: Εισαγωγή

### 1.1. Το CRISPR-Cas9 και η λειτουργία του

Το σύστημα CRISPR-Cas9 είναι ένα εργαλείο επεξεργασίας του γονιδιώματος το οποίο μας διευκολύνει στο να επεξεργαστούμε τμήματα του γονιδιώματος μέσω της αφαίρεσης, της προσθήκης ή της αλλαγής περιοχών της DNA ακολουθίας.

Προέρχεται από το ανοσοποιητικό σύστημα των βακτηρίων και αποτελεί συντομογραφία για το Clustered Regularly Interspaced Short Palindromic Repeats (Συσσωρευμένες, τακτικά παρεμβλλόμενες, βραχείες παλινδρομικές επαναλήψεις). Τα CRISPR είναι κομμάτια του γενετικού κώδικα τα οποία περιέχουν μικρές επαναλήψεις βάσεων και ακολουθούνται από spacer DNA τμήματα. Το spacer DNA συντίθεται από κομμάτια διαφορετικών ικών γονιδιωμάτων τα οποία είχαν προσβάλει το κύτταρο στο παρελθόν<sup>[1]</sup>. Επομένως κάθε φορά που ο βακτηριοφάγος μολύνει κάποιο κύτταρο, το κύτταρο παίρνει ένα τμήμα του και το εισάγει στο spacer DNA. Σε κάθε επόμενη πιθανή απόπειρα εισβολής του το βακτήριο θα είναι έτοιμο να τον εξουδετερώσει και κατά συνέπεια να ζήσει.

Το σύστημα CRISPR-Cas9 αποτελείται από 2 μόρια-κλειδιά τα οποία εισάγουν μία αλλαγή (μετάλλαξη) στο DNA. Αυτά είναι η πρωτεΐνη Cas9 και ένα κομμάτι RNA που ονομάζεται guide RNA (gRNA, που συντίθεται από μία crRNA και tracrRNA ακολουθία) ή single guide RNA (sgRNA) (Εικόνα 1)<sup>[2]</sup>.

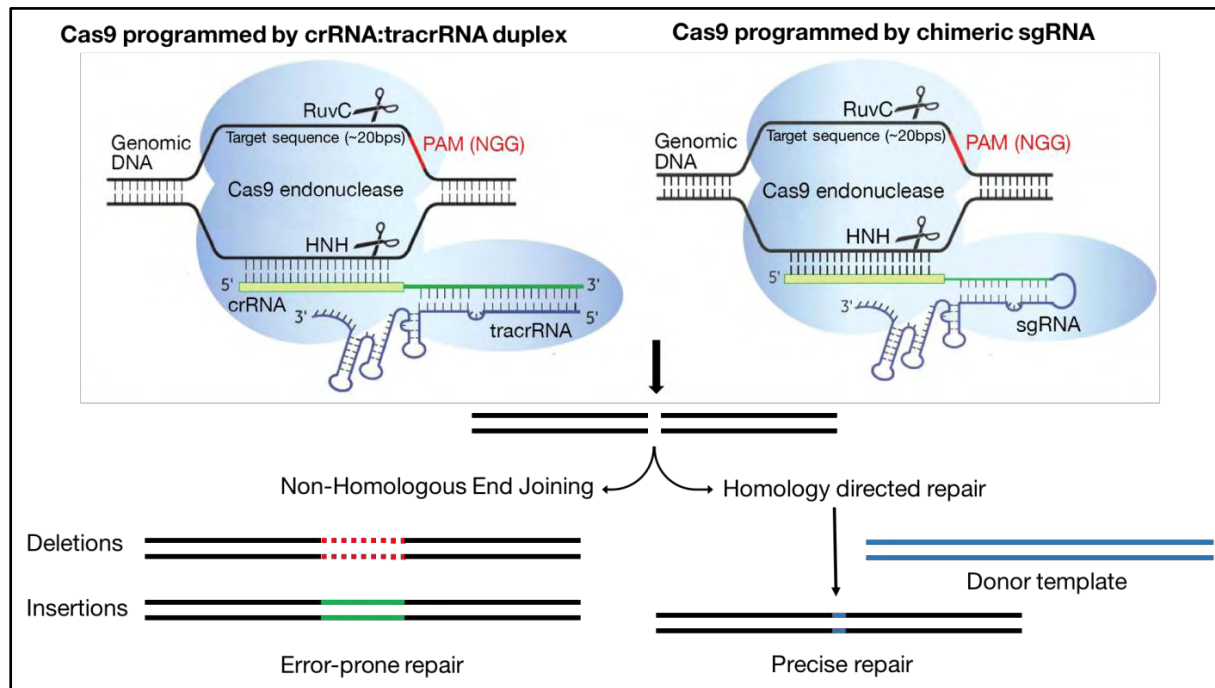
Η πρωτεΐνη Cas9 ενεργεί σαν «μοριακό ψαλίδι» το οποίο κόβει το δίκλωνο DNA σε μία συγκεκριμένη περιοχή στο γονιδίωμα έτσι ώστε μέσω της διαδικασίας επιδιόρθωσης του DNA να εισαχθούν ή αφαιρεθούν μικρές ακολουθίες DNA. Τα Cas γονίδια χρησιμοποιούνται στη δημιουργία Cas πρωτεϊνών. Οι Cas πρωτεΐνες αρχικά δρουν σαν ελικάσες, οι οποίες είναι πρωτεΐνες που ξετυλίγουν το DNA, και έπειτα σαν νουκλεάσες κόβοντας το DNA στην περιοχή που θέλουμε να εισαχθεί η μετάλλαξη. Το gRNA αποτελείται από ένα μικρό κομμάτι προσχεδιασμένης RNA ακολουθίας (με περίπου 20 βάσεις μήκος) που βρίσκεται μεταξύ μιας μακρύτερης RNA σκαλωσιάς (scaffold) και περιέχει RNA βάσεις συμπληρωματικές στο «DNA στόχο» του γονιδιώματος. Αυτό είναι σχεδιασμένο ώστε να εντοπίσει και να προσδεθεί σε μία συγκεκριμένη ακολουθία στο DNA και να οδηγήσει την Cas9 στο να ξετυλίξει τις 2 έλικες του DNA και να κόψει μία συγκεκριμένη περιοχή. Με αυτό τον τρόπο είναι σίγουρο ότι η Cas9 θα κόψει το γονιδίωμα επακριβώς<sup>[3]</sup>.

#### 1.1.1. Μηχανισμοί επιδιόρθωσης του κυττάρου

Αφού το Cas9 κόψει το DNA, το κύτταρο αναγνωρίζει ότι το DNA έχει υποστεί βλάβη και προσπαθεί να το επιδιορθώσει μέσω δύο πιθανών μονοπατιών, είτε με τη χρήση της σύνδεσης μη-ομόλογων άκρων (Non-Homologous End Joining: NHEJ) είτε με σύνδεση ομόλογων άκρων (Homology directed repair: HDR).

Το NHEJ είναι μία διαδικασία επιρρεπής σε σφάλματα καθώς εισάγει λάθη μέσω εισδοχών ή διαγραφών βάσεων (indels) τα οποία αποδιοργανώνουν την λειτουργία του γονιδίου. Αντίθετα, η HDR συμβαίνει μόνο κατά τη διάρκεια του κυτταρικού κύκλου όπου υπάρχει η

αδερφή χρωματίδη και γι' αυτό είναι πολύ ακριβής μηχανισμός <sup>[4]</sup>. Γι' αυτό συνήθως στην περίπτωση των CRISPR screens, όπου σκοπός μας είναι να εισάγουμε αλλαγές σε ένα ή περισσότερα γονίδια στο γονιδίωμα του κυττάρου που μας ενδιαφέρει, είναι προτιμότερη η χρήση των NHEJ επιδιορθώσεων.



**Εικόνα 1:** Επεξεργασία του γονιδιώματος μέσω του μηχανισμού CRISPR-Cas9 και επιδιόρθωση της βλάβης στην διπλή έλικα του DNA <sup>[5]</sup>. Ο προγραμματισμός του Cas9 μπορεί να γίνει είτε μέσω ενός διμερούς RNA το οποίο έχει παραχθεί από crRNA και tracrRNA, είτε ένα χμιαρικό sgRNA (ενώνοντας τεχνητά τα δύο μέρη RNA σε ένα sgRNA) το οποίο μιμείται το διμερές RNA. Το Cas9 κόβει το DNA στο κομμάτι που είναι συμπληρωματικό με τις 20 περίπου βάσεις του gRNA και του προκαλεί μία δίκλωνη θραύση. Αυτή η βλάβη θα διορθωθεί από το κύτταρο μέσω του μηχανισμού σύνδεσης μη-ομόλογων άκρων (Non-Homologous End Joining: NHEJ) ή με σύνδεση ομόλογων άκρων (Homology directed repair: HDR). Κατά τη χρήση του CRISPR-Cas9 ως εργαλείο, προτιμότερη είναι η σύνδεση μη-ομόλογων άκρων καθώς εισάγει λάθη στο γονιδίωμα και βασικός στόχος όλης αυτής της διαδικασίας είναι οι αλλαγές στο γονιδίωμα του κυττάρου-στόχου.

### 1.1.2. Εφαρμογές του CRISPR:

Το CRISPR επιτρέπει στους ειδικούς να τροποποιούν εργαστηριακά τα γονίδια όλων των ζωικών και φυτικών οργανισμών και βελτιώνει τομείς όπως η υγεία και το περιβάλλον.

Επιστήμονες σε όλο τον κόσμο έχουν αντιληφθεί τη χρησιμότητά του σε σύντομο χρονικό διάστημα και έχουν ξεκινήσει τα πειράματα για την θεραπεία γενετικών ασθενειών καθώς επίσης και την ανάπτυξη διάφορων εφαρμογών στην γεωργία. Για παράδειγμα ερευνητές στο Ηνωμένο Βασίλειο έχουν ξεκινήσει την εφαρμογή μεθόδων για την καταπολέμηση του καρκίνου. Συγκεκριμένα χρησιμοποιούν μία τεχνική για να μελετήσουν την βιολογία των εγκεφαλικών καρκινικών όγκων με σκοπό την εύρεση θεραπείας του <sup>[6]</sup>. Επίσης τεράστιο επίτευγμα αποτέλεσε η επιτυχής αφαίρεση του HIV από τα ανθρώπινα ανοσοκύτταρα, η εξάλειψη του DNA του από ανθρώπινες εργαστηριακές καλλιέργειες και το γεγονός ότι

επόμενη έκθεση των κυττάρων στον ιό δεν προκαλεί καμία μόλυνση <sup>[7]</sup>. Πολύ σημαντική είναι και η συνεισφορά του στην βελτίωση των εξωσωματικών γονιμοποιήσεων και την μείωση των αποβολών <sup>[8]</sup>. Όσον αφορά το περιβάλλον Αμερικανοί επιστήμονες κατάφεραν να δημιουργήσουν μεταλλαγμένα κουνούπια, τα οποία όχι μόνο δεν μολύνονται από το παράσιτο της ελονοσίας αλλά και όταν αναπαράγονται μεταξύ τους μεταδίδουν κληρονομικά αυτή την ανοχή στους απογόνους τους <sup>[9]</sup>. Επιπλέον με την επιτυχή παραγωγή της διπλάσιας ποσότητας πετρελαίου μέσω της επεξεργασίας του γονιδιώματος του θαλάσσιου οργανισμού φύκι θα πραγματοποιηθεί η παραγωγή βιοκαυσίμων και η ανάπτυξη εναλλακτικών πηγών ενέργειας <sup>[10]</sup>. Τέλος στόχο αποτελεί η παραγωγή άσπορων φρούτων και λαχανικών που θα καλλιεργούνται εξολοκλήρου στα εργαστήρια και θα αποφεύγουν κακές περιβαλλοντικές συνθήκες όπως η ανεπαρκής επικοινωνία <sup>[11]</sup>.

## 1.2. Τί είναι τα CRISPR screens

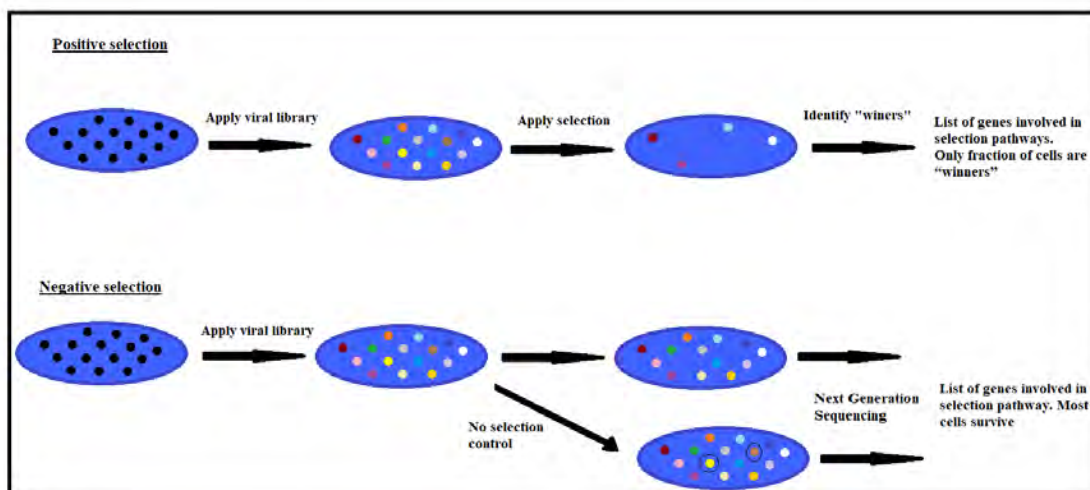
Οι CRISPR-Cas9 screens είναι μία υψηλής απόδοσης συστηματική προσέγγιση για την ταυτοποίηση γονιδίων που εμπλέκονται σε μία βιολογική διαδικασία και μας παρέχουν την ευκαιρία να εξετάσουμε χιλιάδες γενετικές διαταραχές σε ένα μόνο πείραμα <sup>[12]</sup>. Το CRISPR επιτρέπει την δημιουργία μεταλλάξεων σε μεγάλη κλίμακα και χρησιμοποιείται στις knockout screens, στις οποίες κάθε γονίδιο στοχεύεται από διάφορα sgRNAs <sup>[13]</sup>.

Για να κάνουμε screens χρειάζεται να εισάγουμε στα κύτταρα-στόχους βιβλιοθήκες που περιέχουν sgRNA (μέσω lentiviruses ή άλλους ρετροϊούς). Οι ιοί αυτοί ενσωματώνονται στο γονιδιώμα τους και το μαρκάρουν. Στην συνέχεια διεξάγουμε ένα τεστ διαχωρισμού των κυττάρων με βάση τον φαινότυπο που θα εξετασθεί. Το ποιες διαταραχές οδήγησαν σε έναν συγκεκριμένο φαινότυπο καθορίζονται παίρνοντας την ακολουθία τους από το γονιδιωματικό DNA και υπολογίζοντας την αφθονία της κάθε διαταραχής σε gRNAs <sup>[14]</sup>.

Η δημιουργία μιας βιβλιοθήκης επιτυγχάνεται με την παρακάτω διαδικασία: Τα gRNA ολιγονουκλεοτίδια συντίθενται ως μία γονιδιακή δεξαμενή (pool) και στη συνέχεια κλωνοποιούνται για να δημιουργήσουν μία βιβλιοθήκη πλασμιδίων η οποία χρησιμεύει στην παραγωγή των ιών <sup>[14]</sup>.

Υπάρχουν διαφορετικά είδη βιβλιοθηκών οι οποίες χρησιμοποιούνται σήμερα. Στην παρούσα πτυχιακή εργασία θα χρησιμοποιήσω σύνολα δεδομένων που προέρχονται από «Knockout» βιβλιοθήκες, όπως οι Whitehead [15], Brunello [16], Yusa [17], Toronto KnockOut <sup>[13]</sup>, Avana <sup>[18]</sup>, GeCko <sup>[19]</sup> κ.ά. Οι συγκεκριμένες βιβλιοθήκες περιέχουν μέχρι 10 gRNAs ανά γονίδιο και το μήκος του κάθε gRNA κυμαίνεται από 19 μέχρι 20 βάσεις <sup>[20]</sup>.

Υπάρχουν 2 είδη screening. Το Positive selection screening και το Negative selection screening. Σε μία positive screen στόχος είναι η εύρεση των κυττάρων που επιβιώνουν με την πάροδο του χρόνου <sup>[14]</sup> μετά την εφαρμογή εξελικτικής πίεσης. Σε αυτά τα screens μόνο ένα μικρό ποσοστό επιβιώνει συνήθως, τα οποία στην συνέχεια συλλέγονται έτσι ώστε να βρεθεί το γονίδιο που στόχευαν. Σε μία negative screen στόχος είναι ο εντοπισμός των κυττάρων που δεν επιβίωσαν μετά τον μηχανισμό επιλογής <sup>[21]</sup>.



**Εικόνα 2: Positive και negative selection screens.** Στην positive selection, την μόλυνση των κυττάρων με ρετροϊούς διαδέχεται η εφαρμογή της εξελικτικής πίεσης και η εύρεση των κυττάρων που επιβίωσαν καθ' όλη την διαδικασία. Στην negative selection, μετά την μόλυνση των κυττάρων με ρετροϊούς ακολουθεί η αλληλούχιση νέας γενιάς (Next Generation Sequencing) στα κύτταρα που επιβίωσαν και η παραγωγή της λίστας με τα gRNA που δεν επιβίωσαν.

Μετά την παραγωγή της sgRNA βιβλιοθήκης, τα κύτταρα προσβάλλονται με ιούς. Στόχος είναι ένα χαμηλό Multiplicity of Infection (MOI) έτσι ώστε να μειωθεί η πιθανότητα ένα κύτταρο να μολυνθεί από περισσότερα gRNAs τη φορά. Για αυτόν τον σκοπό χρησιμοποιείται μεγάλος αριθμός κυττάρων σε σχέση με τους ιούς. Οι ιοί μαζί με το sgRNA μεταφέρουν και ένα selection marker το οποίο χρησιμοποιείται στη συνέχεια για να γίνει αφαίρεση των κυττάρων που δεν έχουν gRNA. Ο marker μπορεί να είναι γονίδιο που δημιουργεί αντίσταση σε συγκεκριμένο χημικό μόριο ή φθορίζουσα πρωτεΐνη (fluorescent protein) που επιτρέπει τη συλλογή των κυττάρων με FACS sorting (Fluorescent-Activated Cell Sorting) <sup>[22]</sup>. Τα κύτταρα που έχουν λάβει gRNA στη συνέχεια επωάζονται σε κατάλληλες συνθήκες για να πολλαπλασιαστούν <sup>[23]</sup>. Τέλος, γίνεται σύγκριση της αφθονίας των gRNAs χρησιμοποιώντας στατιστικά τεστ για να εξακριβωθεί εάν τα κύτταρα που έχουν knockouts του κάθε γονιδίου έχουν αυξηθεί (enriched) ή έχουν μειωθεί (depleted) σε αριθμό.

### 1.2.1. Essential (απαραίτητα) γονίδια

Τα γονίδια αποτελούνται από αλληλουχίες βάσεων DNA και έχουν αποθηκευμένη μία συγκεκριμένη γενετική πληροφορία. Μερικά είναι σημαντικά για την επιβίωση ενός οργανισμού ενώ άλλα όχι. Αυτό οδηγεί στην δημιουργία δύο ομάδων γονιδίων, των essential (απαραίτητων) και non-essential (μη απαραίτητων). Τα essential (απαραίτητα) αναφέρονται στα θεμελιώδη γονίδια ενός κυττάρου τα οποία είναι κρίσιμα για την επιβίωσή του <sup>[24]</sup>. Ωστόσο, αυτό εξαρτάται από τις συνθήκες που ζει ο οργανισμός. Για παράδειγμα, ένα γονίδιο που απαιτείται για την πέψη του αμύλου είναι essential (απαραίτητο) εάν το άμυλο είναι η μόνη διαθέσιμη πηγή ενέργειας <sup>[25]</sup>. Αντίθετα, τα non-essential (μη-απαραίτητα) δεν οδηγούν στο θάνατο του οργανισμού εάν αφαιρεθούν. Ακόμη μερικά γονίδια είναι απαραίτητα στην πλειοψηφία των κυτταρικών σειρών και είναι γνωστά ως "core essential" (π.χ. ριβοσώματα) ενώ άλλα μόνο σε κυτταρικές σειρές με συγκεκριμένο υπόβαθρο (π.χ. γενετικό ή ιστό) και ονομάζονται "context-dependent essential" (π.χ. το γονίδιο BRAF είναι

essential στις κυτταρικές σειρές του μελανώματος, όπου υπάρχει η μεταλλαγμένη μορφή του (BRAFFV600E) <sup>[26]</sup>.

### 1.2.2. Negative selection screens

Οι Shalem O. και et. al. <sup>[27]</sup> χρησιμοποίησαν negative selection screens με σκοπό να δουν την αποτελεσματικότητα της GeCKO βιβλιοθήκης στο να αναγνωρίζει essential (απαραίτητα) γονίδια στα καρκινικά κύτταρα και στα πολυδύναμα βλαστοκύτταρα. Η συγκεκριμένη βιβλιοθήκη χρησιμοποιήθηκε για να προσβάλει την ανθρώπινη κυτταρική σειρά μελανώματος A375 και την ανθρώπινη σειρά βλαστικών κυττάρων HUES62, με MOI=0.3. Μετά από 14 μέρες από την προσβολή παρατήρησαν μία σημαντική μείωση στην ποικιλία των sgRNAs στα A375 και HUES62 κύτταρα που είχαν επιβιώσει. Αυτό υποδεικνύει ότι τα περισσότερα από τα sgRNAs που είχαν μειωθεί στόχευαν essential (απαραίτητα) γονίδια, όπως ριβοσωμικά δομικά συστατικά.

### 1.2.3. Positive selection screens

Στην positive selection <sup>[27]</sup> στόχο αποτέλεσε η εύρεση εκείνων των γονιδίων των οποίων η απώλεια κάνει το κύτταρο ανθεκτικό στον «αναστολέα του BRAF» (protein kinase inhibitor vemurafenib -PLX) στο μελάνωμα, χρησιμοποιώντας όπως και στην negative selection την GeCKO βιβλιοθήκη. Η βεμουραφενίμπη (vemurafenib) είναι αναστολέας της BRAF μιας πρωτεΐνης που συμμετέχει στην ενεργοποίηση της κυτταρικής διαίρεσης <sup>[28]</sup>. Η συνολικής 14 ημερών έκθεση των κυττάρων στο PLX οδήγησε στην αναστολή της ανάπτυξης των κυττάρων που έφεραν την V600E μετάλλαξη, ενώ αντίθετα μόνο για ένα μικρό ποσοστό γονιδίων πραγματοποιήθηκε ο πολλαπλασιασμός των sgRNAs που στόχευαν τα γονίδια ανθεκτικών στο συγκεκριμένο φάρμακο κυττάρων. Αυτό το μικρό υποσύνολο είναι και αυτό που έκανε το κύτταρο περισσότερο ανθεκτικό στο PLX. Συνολικά, από τα ανθεκτικά γονίδια, εντοπίστηκαν κάποια που ήταν ήδη γνωστό ότι αντέχουν στην έκθεση στο PLX, ενώ βρέθηκαν και αρκετά καινούργια.

## 1.3. Η χρήση των CRISPR για την μελέτη των μονοπατιών κυτταρικής σηματοδότησης

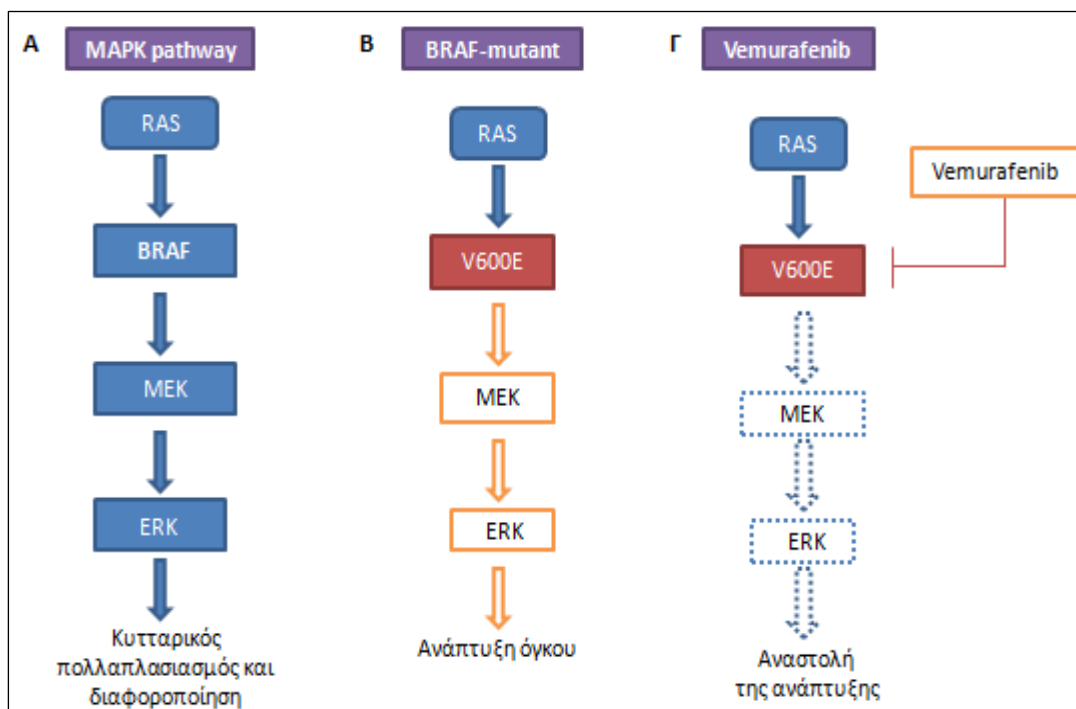
Κυτταρική σηματοδότηση ονομάζεται η διαδικασία με την οποία ένα σήμα αλληλεπιδρά με ειδικούς υποδοχείς των κυττάρων και οδηγεί σε κυτταρική απόκριση που εκφράζεται με μεταβολή χημικών διεργασιών. Είναι ζωτικής σημασίας για πολλές βιολογικές διαδικασίες, όπως η κυτταρική αύξηση και διαφοροποίηση, η απόπτωση κ.α. και επιτυγχάνεται μέσω των μονοπατιών κυτταρικής σηματοδότησης. Η μελέτη των μονοπατιών κυτταρικής σηματοδότησης είναι πολύ σημαντική γιατί μπορεί να οδηγήσει στην εύρεση και εξάλειψη ενός πλήθους ανθρώπινων ασθενειών. Συγκεκριμένα, τα ελαττώματα σε ένα σηματοδοτικό μονοπάτι συνδέονται με ένα πλήθος ανθρώπινων ασθενειών και η πλήρης κατανόηση τους μπορεί να οδηγήσει στην δημιουργία θεραπευτικών στόχων <sup>[6]</sup>.

Ένα πολύ καλό παράδειγμα αποτελεί η περίπτωση του μελανώματος αφού η BRAFFV600E μετάλλαξη που το προκαλεί μπορεί να αποτελέσει στόχο καταστολέων φαρμάκων. Το BRAF είναι μία πρωτεΐνη η οποία βοηθά στην μετάδοση των χημικών σημάτων από το εξωτερικό του κυττάρου στον πυρήνα του. Αυτή η πρωτεΐνη είναι μέρος ενός σηματοδοτικού



μονοπατιού γνωστό και ως RAS/MAPK, το οποίο ρυθμίζει κυτταρικές λειτουργίες όπως αύξηση, διαφοροποίηση και απόπτωση [Εικόνα 3]. Το γονίδιο BRAF ανήκει σε μια κατηγορία γονιδίων γνωστά ως ογκογονίδια, τα οποία όταν υποστούν μετάλλαξη έχουν την τάση να μετατρέπουν τα υγιή κύτταρα σε καρκινικά. Αυτά τα φάρμακα λοιπόν παρεμποδίζουν το BRAF που έχει υποστεί μετάλλαξη από το να ενεργοποιήσει το υπόλοιπο μονοπάτι και με αυτόν τον τρόπο θεραπεύεται ο συγκεκριμένος τύπος καρκίνου [36].

Γενικότερα η διαδικασία της κυτταρικής σηματοδότησης είναι περίπλοκη καθώς σε αυτήν λαμβάνουν μέρος πολλές πρωτεΐνες και δημιουργούνται περίπλοκα δίκτυα και έτσι καθίσταται ακόμη πιο δύσκολη η συστηματική της ανάλυση από τους επιστήμονες. Οι πρώτες προσπάθειες που έγιναν γι' αυτόν το σκοπό οδηγούσαν σε αποτελέσματα τα οποία δεν ήταν απόλυτα σωστά. Λύση σε αυτές τις δυσκολίες έδωσαν οι μελέτες των κυτταρικών δικτύων εφαρμόζοντας διαταραχές των γονιδίων με την τεχνολογία του CRISPR και των CRISPR screens. Με αυτόν τον τρόπο δίνονται με βεβαιότητα απαντήσεις σχετικά με το ποια γονίδια συνδέονται με αλλαγές στην κυτταρική σηματοδότηση [6].



**Εικόνα 3: Ενεργοποίηση του MAPK σηματοδοτικού μονοπατιού μέσω της BRAFV600E μετάλλαξης.** (Α) Κυτταρική σηματοδότηση του μονοπατιού σε ένα υγιές κύτταρο. Το σήμα περνά μέσω των πρωτεϊνών στον πυρήνα και οδηγεί στην φυσιολογική ανάπτυξη και διαφοροποίηση του κυττάρου. (Β) Κύτταρο με μετάλλαξη στη BRAF πρωτεΐνη. Το μονοπάτι μετά την μετάλλαξη είναι πάντα ενεργοποιημένο και έχει σαν αποτέλεσμα την ανάπτυξη όγκου. (Γ) Αποτελέσματα της χρήσης καταστολέων φαρμάκων (vemurafenib). Οι καταστολείς παρεμποδίζουν την BRAF πρωτεΐνη από το να ενεργοποιήσει και το υπόλοιπο μονοπάτι και οδηγεί στην αναστολή του.

Αξιοσημείωτη είναι και η χρήση των CRISPR για την μελέτη της κυτταρικής σηματοδότησης χρησιμοποιώντας signaling reporters . Μέχρι στιγμής έχουν γίνει αρκετές τέτοιες προσπάθειες σε διάφορες κυτταρικές σειρές στις οποίες προστίθενται signaling reporters στα γονίδια, δηλαδή φθορίζουσες πρωτεΐνες ή δείκτες ανθεκτικοί στα αντιβιοτικά, που λειτουργούν σαν μοριακές «ετικέτες» για να γίνει αντιληπτή η λειτουργία που μας ενδιαφέρει. Χρησιμοποιώντας φθορίζουσες πρωτεΐνες μπορούν για παράδειγμα να ανιχνευθούν αλλαγές στην κυτταρική σηματοδότηση μετά την χρήση κάποιου φαρμάκου που κάνει τα κύτταρα ανθεκτικά σε αυτό το φάρμακο. Από την άλλη μεριά, οι δείκτες που είναι ανθεκτικοί στα αντιβιοτικά βοηθούν στην εύρεση των γονιδίων που είναι ανθεκτικά σε κάποιο αντιβιοτικό και συνετέλεσαν στην αντοχή του κυττάρου σε αυτό το αντιβιοτικό και κατά συνέπεια στην ανάπτυξη του <sup>[29]</sup> . Τα κύτταρα που στοχεύουν γονίδια τα οποία συμβάλουν στην έκφραση των γονιδίων που έχουν «ετικέτες» απομονώνονται μέσω FACS sorting (Fluorescent-Activated Cell Sorting) τεχνικών <sup>[22]</sup> .

Πολλές μελέτες έχουν χρησιμοποιήσει τις παραπάνω τεχνικές για να εντοπίσουν τα γονίδια που ελέγχουν την έκφραση ενός ή περισσότερων γονιδίων, ή οδών του κυτταρικού συστήματος σηματοδότησης. Παραδείγματος χάρη, η FACS τεχνική χρησιμοποιήθηκε για τον εντοπισμό των γονιδίων που είναι απαραίτητα για την ενεργοποίηση της έκφρασης ενός παράγοντα νέκρωσης όγκου, ο οποίος χρησιμοποιήθηκε σαν signaling reporter, υπό την διέγερση ενός λιποπολυσακχαρίτη <sup>[30]</sup> . Ακόμη, χρήση αυτών των τεχνικών έκαναν και οι DeJesus R. et al. για να βρουν τους ρυθμιστές της SQSTM1 πρωτεΐνης υποβοηθούμενοι από την H4 κυτταρική σειρά του νευρογλιώματος η οποία εκφράζει την Cas9 πρωτεΐνη καθώς και ταγκαρισμένες με GFP SQSTM1 πρωτεΐνες [31]. Επιπλέον, οι Pushupati GV et al. βρήκαν τους ρυθμιστές του Sonic Hedgehog μονοπατιού, το οποίο ρυθμίζει ένα μεγάλο πλήθος λειτουργιών κατά την διάρκεια της ανάπτυξης των ζώων , όχι όμως προσθέτοντας μία φθορίζουσα ουσία σε μία πρωτεΐνη, αλλά δημιουργώντας μία ολόκληρη κυτταρική σειρά που εκφράζει την φθορίζουσα ουσία <sup>[32]</sup> .

Τέλος, μεταξύ των υπολοίπων, εφαρμογή όλων των παραπάνω γίνεται και στην ομάδα του EBI για την εύρεση γονιδίων που επηρεάζουν τα signaling pathways στις A375 κυτταρικές σειρές του μελανώματος.

#### 1.4. Μελάνωμα

Το μελάνωμα αποτελεί έναν από τους πιο σοβαρούς και επικίνδυνους τύπους καρκίνου του δέρματος καθώς προκαλεί το 75% περίπου όλων των θανάτων από καρκίνο του δέρματος. Αναπτύσσεται όταν τα μελανοκύτταρα, τα οποία είναι ειδικά κύτταρα που βρίσκονται στην επιδερμίδα και παράγουν μελανίνη, παύουν να είναι υγιή και πολλαπλασιάζονται ανεξέλεγκτα προκαλώντας την εμφάνιση καρκινικού όγκου <sup>[33]</sup> .

Κυρίαρχο ρόλο στην εμφάνισή του διαδραματίζει η υπεριώδης ακτινοβολία (UV), δηλαδή η έκθεση στο ηλιακό φως ή στο φως τεχνητού μαυρίσματος, η οποία καταστρέφει το DNA και δημιουργεί μεταλλάξεις που οδηγούν στον ανεξέλεγκτο πολλαπλασιασμό των κυττάρων του δέρματος. Συχνά, αυτά τα κύτταρα παράγουν καλοήθεις σπίλους, οι οποίοι μπορεί να εξελιχθούν σε κακοήθες μελάνωμα και να εξαπλωθούν και σε άλλα σημεία του σώματος. Επίσης σημαντική είναι και η γενετική προδιάθεση του κάθε ατόμου για την ανάπτυξη μελανώματος. Γενικά δεν αποτελεί μία κληρονομική νόσο καθώς στην πλειοψηφία των περιπτώσεων δεν υπάρχει οικογενειακό ιστορικό <sup>[34]</sup> .



Μεταλλάξεις του γονιδίου BRAF (BRAFFV600E) παρατηρούνται με μεγάλη συχνότητα στα μελανώματα (περίπου 50%) και ενοχοποιούνται για μεταστάσεις και σε άλλα όργανα π.χ. το ήπαρ. Γι' αυτό και μία από τις κύριες εξετάσεις που πραγματοποιούνται για την διάγνωση του μελανώματος είναι η ανίχνευση της συγκεκριμένης μετάλλαξης (BRAFFV600E) προκειμένου να ελεγχθεί η δυνατότητα θεραπείας με αναστολείς BRAF. Οι αναστολείς ελέγχουν τον καρκίνο, μπλοκάρουν την μετάστασή του σε άλλα όργανα και αυξάνουν δυνητικά την επιβίωση των ασθενών <sup>[35]</sup>.

Η έρευνα γύρω από το μελάνωμα διεξάγεται με εντατικούς ρυθμούς, προκειμένου να γίνουν περισσότερο κατανοητοί οι μηχανισμοί εξέλιξης της νόσου και να ανακαλυφθούν νέες θεραπευτικές μέθοδοι για την αντιμετώπισή του. Μεγάλη συμβολή σε αυτή την προσπάθεια έχει η χρήση της μεθόδου CRISPR καθώς μπορεί να χρησιμοποιηθεί για να δημιουργηθούν ακριβή μοντέλα μελανώματος στο εργαστήριο. Συγκεκριμένα, οι ερευνητές χρησιμοποιούν αυτή την τεχνολογία για να εισάγουν μεταλλάξεις οι οποίες καθιστούν την A375 κυτταρική σειρά του μελανώματος ανθεκτική σε συγκεκριμένα φάρμακα και για να δημιουργήσουν μία καινούργια κυτταρική σειρά ανθεκτική στους αναστολείς MEK και BRAF (MEK και BRAF inhibitors) ευαίσθητη σε θεραπείες που στοχεύουν γονίδια του Ras/Raf-MAPK signaling pathway, καθιστώντας την ιδανική για screening και αξιολόγηση νέων θεραπειών για το μελάνωμα <sup>[36]</sup>. Επιπλέον, μία ακόμη ομάδα επιστημόνων με επικεφαλής τον Robert Judson <sup>[37]</sup>, για να καταλάβουν το πώς σχηματίζεται το μελάνωμα, εισήγαγε σε υγιή ανθρώπινα δερματικά κύτταρα, χρησιμοποιώντας το CRISPR, τις ακολουθίες DNA από ασθενείς με καρκίνο του δέρματος. Μια ενδιαφέρουσα παρατήρηση ήταν το γεγονός ότι οι όγκοι δεν δημιουργήθηκαν μόνο από μεταλλάξεις στο γονίδιο καταστολής όγκων CDKN2A όπως είχαν προβλέψει, αλλά και από την ενεργοποίηση του μεταγραφικού παράγοντα BRN2 ο οποίος οδήγησε τα επηρεασμένα κύτταρα να γίνουν επιρρεπή στο να εξαπλώνονται και σε άλλα μέρη του σώματος <sup>[38]</sup>.

## Κεφάλαιο 2: Στόχος και κίνητρα

Η παρούσα πτυχιακή εργασία αποτελεί τμήμα ενός project που διεξάγεται στο European Bioinformatics Institute (EBI) στην ομάδα της Δρ Ευαγγελίας Πετσαλάκη από την μεταδιδάκτορα Sumana Sharma. Στο ευρύτερο project γίνεται χρήση των A375 κυτταρικών σειρών για τον σχεδιασμό genome-wide loss-of-function screens για την εύρεση γονιδίων που επηρεάζουν τα σηματοδοτικά μονοπάτια και για το πώς αυτά αλλάζουν με την παρουσία φαρμάκων που κάνουν τα γονίδια ανθεκτικά. Μέχρι σήμερα υπάρχουν δεκάδες δημοσιευμένες μελέτες που έχουν χρησιμοποιήσει τις A375 κυτταρικές σειρές για να εκτελέσουν CRISPR screens και να εντοπίσουν γονίδια, όμως δεν τα συνδέουν με τα μονοπάτια κυτταρικής σηματοδότησης κάτι το οποίο επιδιώκει το project συνολικά. Με λίγα λόγια, η ομάδα έχει φτιάξει ένα reporter cell line για να μελετήσει τις αλλαγές που συμβαίνουν στην σηματοδότηση μετά την χρήση κάποιου φαρμάκου που κάνει το κύτταρο ανθεκτικό, με εφαρμογή στις κυτταρικές σειρές του μελανώματος

Το project μου εστιάζει στην ανάλυση των CRISPR screens που έχουν διεξαχθεί στις A375 κυτταρικές σειρές του μελανώματος για την ανίχνευση απαραίτητων γονιδίων και γονιδίων υπευθύνων για ανθεκτικότητα σε φάρμακα. Η επιλογή των A375 κυτταρικών σειρών δεν είναι τυχαία. Αντίθετα, είναι ένα ιδανικό σύστημα μοντέλου για την μελέτη του μελανώματος καθώς φέρουν μία μετάλλαξη στο BRAF γονίδιο (BRAFFV600E) ίδια με αυτή του 50%-60% των ασθενών με μελάνωμα. Επιπλέον είναι δημοφιλείς εξαιτίας του ότι είναι πολύ εύκολο να μολυνθούν με λεντιϊούς και να καλλιεργηθούν σε μεγάλες ποσότητες.

### Στόχοι της πτυχιακής μου εργασίας:

1. Ο εντοπισμός της καλύτερης βιβλιοθήκης για την εφαρμογή CRISPR screens στις A375 κυτταρικές σειρές.
  - 2.1. Η εύρεση των essential (απαραίτητων) γονιδίων στις A375 κυτταρικές σειρές και
  - 2.2. Η εύρεση των essential (απαραίτητων) γονιδίων σε μονοπάτια κυτταρικής σηματοδότησης που μας ενδιαφέρουν και σχετίζονται με το μελάνωμα.
3. Η εύρεση των γονιδίων που προκαλούν αντίσταση στους αναστολείς MEK και BRAF. Αυτοί οι αναστολείς χρησιμοποιούνται συνήθως στο μελάνωμα.

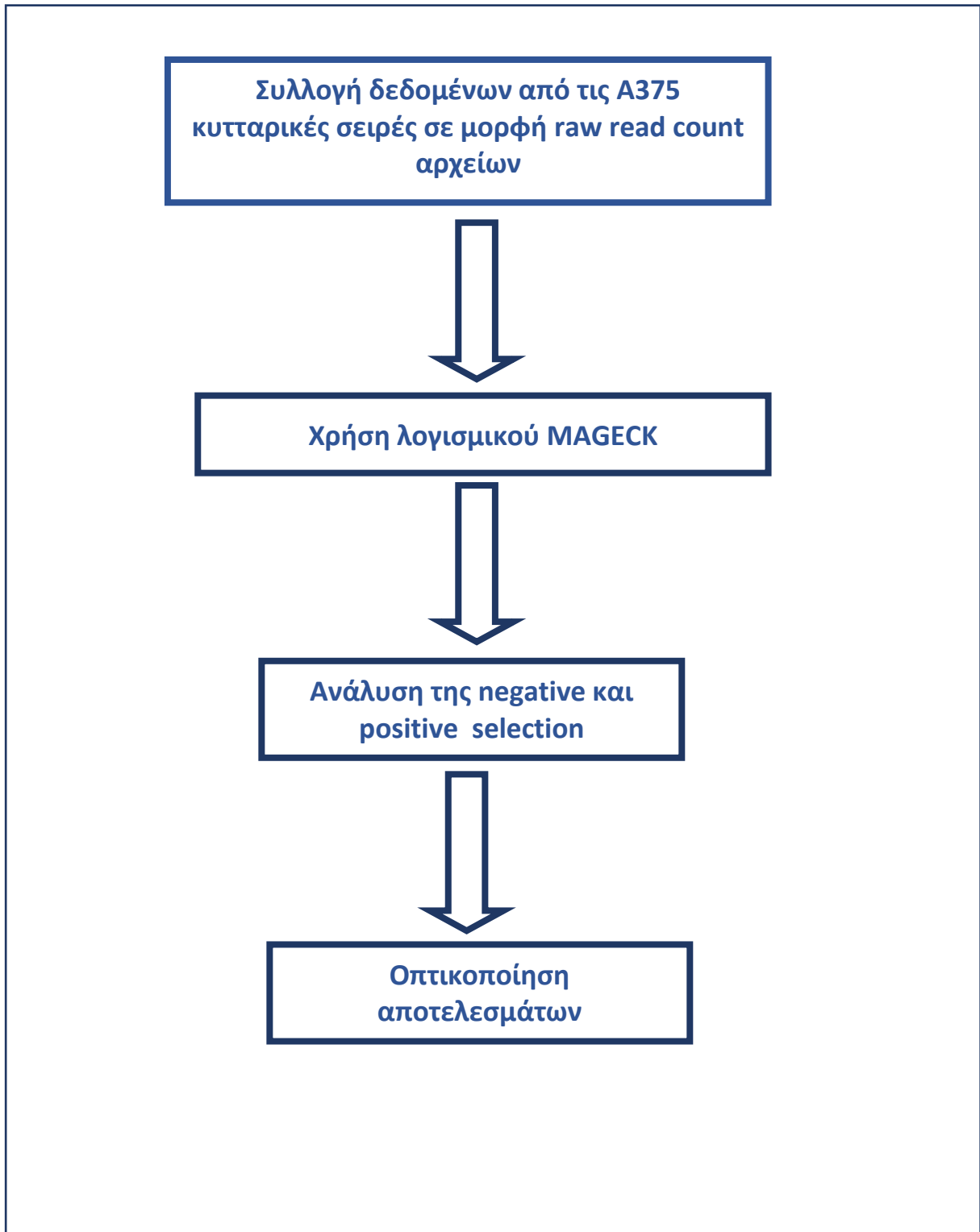
Απώτερος σκοπός είναι η χρήση όλων των παραπάνω ευρημάτων σαν αναφορά σε CRISPR screens που θα γίνουν από την ομάδα του EBI στο μέλλον.

## Κεφάλαιο 3: Υλικά και Μέθοδοι

### 3.1. Τα προγραμματισμένα βήματα της εργασίας μου είναι:

- 1.1. Η εύρεση όλων των CRISPR screens που έχουν διεξαχθεί στις A375 κυτταρικές σειρές του μελανώματος.
- 1.2. Η εύρεση των χαρακτηριστικών των sgRNA βιβλιοθηκών που έχουν χρησιμοποιηθεί σε κάθε μελέτη.
- 2.1 Ο εντοπισμός των μελετών που έχουν σαν στόχο να βρουν τα essential (απαραίτητα) γονίδια στις A375 κυτταρικές σειρές.
- 2.2 Η εύρεση των γονιδίων που εμφανίζονται από κοινού σε διαφορετικές βιβλιοθήκες.
- 2.3 Η αξιολόγηση των βιβλιοθηκών χρησιμοποιώντας ένα προκαθορισμένο set γνωστών απαραίτητων και μη απαραίτητων γονιδίων με στόχο να δούμε πόσο καλή είναι η μελέτη της ομάδας μας.
- 3.1 Ο εντοπισμός των μελετών που έχουν εφαρμόσει drug screens στις A375 κυτταρικές σειρές με μεγαλύτερο ενδιαφέρον σε αυτές που έχουν εστιάσει σε καταστολείς MAPK.
- 3.2 Η εύρεση των καταστολέων κινασών που χρησιμοποιήθηκαν σε αυτή την έρευνα και των γονιδίων που εντόπισαν, η αναζήτηση του εάν τα είδαμε και σε άλλες μελέτες και τέλος πόσο καλά σχετίζονται με τα hits που είχαμε εντοπίσει.

Τα βήματα που ακολούθησα για την διεκπεραίωση της πτυχιακής μου εργασίας συνοψίζονται στο ακόλουθο διάγραμμα ροής (Εικόνα 4).



**Εικόνα 4:** Διάγραμμα ροής των βημάτων για την εκπόνηση της πτυχιακής εργασίας μου.

### 3.2. Συλλογή δεδομένων

Αρχικά, συγκέντρωσα δεδομένα από CRISPR screens, negative και positive, οι οποίες διεξήχθησαν στις A375 κυτταρικές σειρές του μελανώματος. Η μορφή τους είναι read count αρχεία και υποδεικνύουν την σχετική αφθονία των sgRNA στην ακολουθία.

Η συλλογή τους έγινε μέσω της διαδικτυακής βάσης δεδομένων GenomeCRISPR<sup>[39]</sup>, η οποία περιέχει δεδομένα για περίπου 700,000 sgRNAs τα οποία χρησιμοποιήθηκαν σε περίπου 500 διαφορετικά πειράματα 421 διαφορετικών ανθρώπινων κυτταρικών σειρών. Τα συγκεκριμένα χρησιμοποίησαν τις βιβλιοθήκες GeCKO\_V1, GeCKO\_V2 και Avena. Επιπλέον υλικό βρήκα και στο paper των (Li W et al., 2014)<sup>[40]</sup> οι οποίοι χρησιμοποίησαν την GeCKO\_V1 βιβλιοθήκη. Τέλος πολύτιμα ήταν και τα δεδομένα που μου δόθηκαν από τον διδάκτορα John Doench<sup>[41]</sup> για τις βιβλιοθήκες Brunello\_V1, Brunello\_V2 και τα αδημοσίευτα μέχρι στιγμής δεδομένα της μεταδιδάκτορος Sumana Sharma από την Yusa\_V1 βιβλιοθήκη.

Συνολικά το υλικό αφορά 43 μελέτες. 21 μελέτες negative selection οι οποίες χρησιμοποίησαν 5 βιβλιοθήκες gRNA (Brunello\_V1, Brunello\_V2, GeCKO\_V1, GeCKO\_V2, Yusa\_V1) και 22 positive selection που έκαναν χρήση 4 βιβλιοθηκών (Yusa\_V1, Brunello, GeCKO\_V1, GeCKO\_V2) (Πίνακας 1).

Βιβλιοθήκη	Αριθμός γονιδίων	Αριθμός gRNAs	Μήκος των gRNA (σε βάσεις)	Μελέτες ανά βιβλιοθήκη (negative selection)	Μελέτες ανά βιβλιοθήκη (positive selection)	Παραπομπή (PubMedID)
Brunello	19114	77441	20	-	6	26780180
Brunello_v1	19114	77442	20	2	-	26780180
Brunello_v2	19114	77442	20	3	-	26780180
GeCKO_v1	18080	64751	20	8	10	25075903
GeCKO_v2	19050	123411	20	4	2	25075903
Yusa_v1	18010	90709	19	4	4	27760321
Σύνολο				21	22	

**Πίνακας 1: CRISPR-Cas9 sgRNA βιβλιοθήκες με κάποια χαρακτηριστικά τους.** Σχεδόν κάθε βιβλιοθήκη περιέχει διαφορετικό αριθμό γονιδίων και gRNA. Το μήκος των gRNA είναι 20 βάσεις εκτός από την βιβλιοθήκη Yusa\_v1 στην οποία είναι 19 εξαιτίας του ότι η ακολουθία ξεκινάει πάντα με το γράμμα G και συνεχίζει με τις υπόλοιπες 19 θέσεις να διαφέρουν πάντα. Έχει αποδειχθεί ότι αυτή η διαφοροποίηση έγκειται στο ότι το G στην αρχή κάνει το σύστημα περισσότερο αποδοτικό.

Επιπλέον, για την ανάλυση των δεδομένων μου χρειάστηκα και υλικό που αφορούσε ένα προκαθορισμένο σύνολο ήδη γνωστών essential (απαραίτητων, 684 συνολικά) και non-essential (μη-απαραίτητων, 927 συνολικά) γονιδίων, το οποίο βρήκα στη δημοσίευση των Hart et al.<sup>[42]</sup>. Τέλος, χρησιμοποίησα τη δυαδικτική βάση δεδομένων KEGG<sup>[43]</sup> για να βρω γονίδια του MAPK pathway (217 συνολικά), του TGF beta (86 συνολικά), του MTOR (52 συνολικά) και του MELANOMA (71 συνολικά) (Όλα τα γονίδια που αναφέρονται παραπάνω βρίσκονται στο Παράρτημα).

### 3.3. ΜΕΘΟΔΟΙ/Επεξεργασία δεδομένων

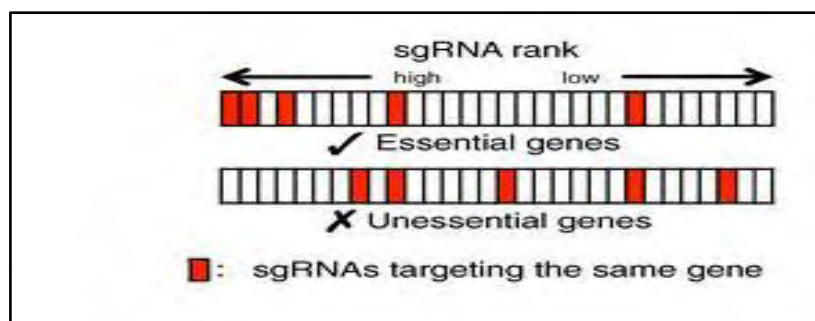
Αφού πρώτα συνέλλεξα όλα τα δεδομένα των negative και positive selection μελετών και τα έβαλα σε ξεχωριστά comma-separated values (.csv) αρχεία για την ευκολότερη προσπέλασή τους (Κώδικας 1), προχώρησα στην ανάλυσή τους με το στατιστικό πακέτο MAGeCK<sup>[40]</sup> (έκδοση v0.5.7).

#### 3.3.2 ΛΙΓΑ ΛΟΓΙΑ ΓΙΑ ΤΟ MAGeCK

Το MAGeCK (Model-based Analysis of Genome-wide CRISPR/Cas9 knockout)<sup>[40]</sup> αποτελεί ένα χρήσιμο εργαλείο για την επεξεργασία των δεδομένων που έχουν προκύψει από CRISPR/Cas9 knockout screens. Αναγνωρίζει ταυτόχρονα τα θετικώς και αρνητικώς επιλεγμένα γονίδια (positively and negatively selected genes) και αφότου κάνει την σύγκριση του δείγματος ελέγχου με το δείγμα μετά το πείραμα (comparing treatments with controls) εντοπίζει τα σημαντικά για το κύτταρο γονίδια και μονοπάτια.

Ο αλγόριθμός του αποτελείται από τέσσερα βήματα:

1. Την κανονικοποίηση των read count δεδομένων.
2. Την εφαρμογή Negative Binomial (NB) μοντέλου για την σύγκριση του πλήθους των sgRNA μεταξύ του αρχικού δείγματος με αυτού μετά το πείραμα.
3. Τον υπολογισμό των p-value των sgRNA και την ταξινόμηση τους με βάση αυτό.
4. Την χρήση  $\alpha$ -Robust Ranking Aggregation ( $\alpha$ -RRA) αλγορίθμου για την εύρεση γονιδίων που έχουν εμπλουτιστεί είτε θετικά είτε αρνητικά. Αυτός ο αλγόριθμος θεωρεί ότι εάν ένα γονίδιο δεν έχει καμία επίδραση στην εφαρμογή εξελικτικής πίεσης (selection), τότε τα sgRNA που στοχεύουν αυτό το γονίδιο είναι ομοιόμορφα κατανεμημένα στην λίστα που περιέχει όλα τα γονίδια (Εικόνα 4).



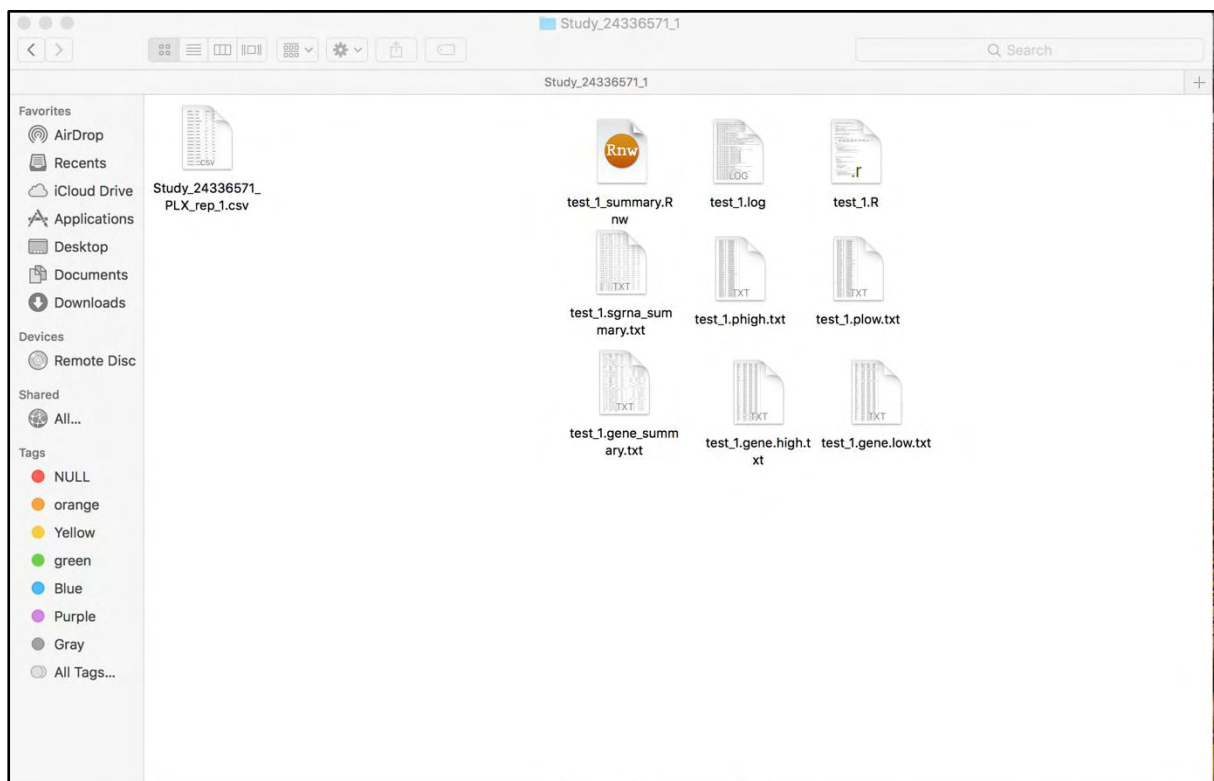
**Εικόνα 5:** Αναγνώριση των essential γονιδίων με βάση το p-value και τον  $\alpha$ -RRA αλγόριθμο

Το MAGeCK δέχεται σαν είσοδο raw read counts αρχεία της μορφής .csv ή .txt και τα επεξεργάζεται μέσω του command line (Εικόνα 6). Η εντολή έχει συγκεκριμένη μορφή και κάθε φορά παραλλάσσεται ανάλογα με τα χαρακτηριστικά του εκάστοτε αρχείου (π.χ. όνομα).

```
Study_24336571_1 -- -bash -- 189x21
~/Desktop/papers_A375_cell_lines/PMID_24336571/resistance_to_PLX_after_7_days/Study_24336571_1 -- -bash
Last login: Thu Nov 29 12:07:55 on ttys000
DGKTVHMSJ1GP:~ natasa$ cd /Users/natasa/Desktop/papers_A375_cell_lines/PMID_24336571/resistance_to_PLX_after_7_days/Study_24336571_1
DGKTVHMSJ1GP:Study_24336571_1 natasa$ ls
Hits(0.1)_from_24336571_Vemurafenib_reistance_Replicate_1.csv  test_1.gene.low.txt
Hits_from_24336571_Vemurafenib_reistance_Replicate_1.csv     test_1.gene_summary.txt
Study_24336571_PLX_rep_1.csv                                  test_1.log
my_data_negFDR.csv                                           test_1.phigh.txt
my_data_posFDR.csv                                           test_1.plow.txt
test_1.R                                                       test_1.sgrna_summary.txt
test_1.gene.high.txt                                          test_1_summary.Rnw
DGKTVHMSJ1GP:Study_24336571_1 natasa$ mageck test -k Study_24336571_PLX_rep_1.csv -t Final -c Initial -n test_1
```

**Εικόνα 6:** Παράδειγμα χρήσης του προγράμματος MAGeCK για την επεξεργασία της μελέτης με pubmed id 24336571. Η εκτέλεση κάθε εντολής γίνεται μέσω του command line.

Το αποτέλεσμα αυτής της επεξεργασίας είναι 9 αρχεία από τα οποία για την μετέπειτα ανάλυση μέσω της R χρησιμοποιώ κυρίως το .gene\_summary.txt αρχείο (Εικόνα 7). Ενδεικτικά αξίζει να αναφέρω ότι σε περίπτωση που θέλουμε να δούμε κάποιον από τους κωδικές μας μπορούμε να ανατρέξουμε στο αρχείο με κατάληξη .log (Εικόνα 7).



**Εικόνα 7:** Έξοδος του MAGeCK. Σε κάθε περίπτωση έξοδος του προγράμματος είναι τα 9 αρχεία που φαίνονται στο δεξί μέρος της εικόνας.



Το `.gene_summary.txt` αρχείο περιέχει πληροφορίες σχετικές με την negative selection και την positive selection (Εικόνα 8).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	id	num	neg score	neg p-value	neg fdr	neg rank	neg goodsgr	neg lfc	pos score	pos p-value	pos fdr	pos rank	pos goodsgr	pos lfc	
1	EIF6	7	1.13E-07	9.26E-07	0.008663	1	6	-1.4732	0.99265	0.99265	1	15679	0	-1.4732	
2	EIF2	4	3.64E-07	2.16E-06	0.008663	2	4	-1.7002	1	1	1	16030	0	-1.7002	
3	NIP7	4	4.71E-07	1.54E-06	0.008663	3	3	-2.032	1	1	1	16029	0	-2.032	
4	POLR3H	4	6.18E-07	2.16E-06	0.008663	4	4	-1.5645	1	1	1	16028	0	-1.5645	
5	SF1	4	9.03E-07	2.78E-06	0.008911	5	3	-1.8536	0.94814	0.94823	1	14748	0	-1.8536	
6	HNRNPJ	4	1.11E-06	3.40E-06	0.009076	6	4	-1.9805	1	1	1	16027	0	-1.9805	
7	ACTL6A	4	1.35E-06	4.01E-06	0.009194	7	4	-1.6766	1	1	1	16026	0	-1.6766	
8	NAA30	4	1.82E-06	5.25E-06	0.01052	8	4	-1.514	1	1	1	16025	0	-1.514	
9	PPAN	3	2.35E-06	6.49E-06	0.011551	9	3	-1.7582	1	1	1	16024	0	-1.7582	
10	DNAJA3	3	2.76E-06	7.72E-06	0.012376	10	3	-1.9135	1	1	1	16023	0	-1.9135	
11	SRP54	5	3.23E-06	1.20E-05	0.014542	11	4	-1.5852	0.99101	0.9911	1	15623	0	-1.5852	
12	PSMG3	4	3.30E-06	1.08E-05	0.014542	12	4	-1.2978	1	1	1	16022	0	-1.2978	
13	KPNB1	4	3.63E-06	1.20E-05	0.014542	13	4	-1.8032	0.99999	0.99998	1	16013	0	-1.8032	
14	HARS	4	3.93E-06	1.39E-05	0.014542	14	4	-1.7714	1	0.99999	1	16021	0	-1.7714	
15	YY1	4	4.17E-06	1.45E-05	0.014542	15	4	-1.6709	0.98682	0.98685	1	15493	0	-1.6709	
16	HEATR1	6	4.54E-06	2.38E-05	0.019059	16	6	-0.93313	1	1	1	16019	0	-0.93313	
17	CMPK1	6	4.54E-06	2.38E-05	0.019059	17	6	-1.0359	1	1	1	16020	0	-1.0359	
18	PRPF38B	3	4.63E-06	1.45E-05	0.014542	18	3	-1.6645	1	1	1	16018	0	-1.6645	
19	RBBP5	4	5.75E-06	2.38E-05	0.019059	19	3	-1.9169	0.42921	0.42859	1	6609	0	-1.9169	
20	PSMA6	3	5.88E-06	2.13E-05	0.019059	20	3	-1.5852	0.99999	0.99999	1	16017	0	-1.5852	
21	NSL1	4	6.03E-06	2.56E-05	0.019566	21	4	-1.728	0.98773	0.9878	1	15526	0	-1.728	
22	COPB1	4	6.59E-06	2.87E-05	0.020927	22	3	-1.7454	0.95927	0.95941	1	14953	0	-1.7454	
23	DCTN1	10	6.89E-06	6.83E-05	0.030391	23	10	-0.67006	0.99998	0.99998	1	16009	0	-0.67006	
24	EIF5	4	7.55E-06	3.43E-05	0.023892	24	3	-1.7847	0.71284	0.71239	1	10986	0	-1.7847	
25	WDR74	4	9.31E-06	4.48E-05	0.027558	25	3	-1.6411	0.96528	0.96539	1	15058	0	-1.6411	
26	DHX15	5	9.83E-06	4.54E-05	0.027558	26	5	-1.3535	0.99999	0.99999	1	16016	0	-1.3535	
27	FAM96B	4	1.09E-05	5.03E-05	0.027558	27	3	-1.6618	0.94346	0.9436	1	14668	0	-1.6618	
28	CSTF3	7	1.10E-05	7.94E-05	0.030528	28	7	-1.0263	0.99966	0.99967	1	15951	0	-1.0263	
29	PMF1	5	1.11E-05	5.10E-05	0.027558	29	5	-1.5675	0.99999	0.99999	1	16015	0	-1.5675	
30	POLR3A	4	1.13E-05	5.16E-05	0.027558	30	4	-1.2854	0.99999	0.99998	1	16014	0	-1.2854	
31	PFON2	6	1.19E-05	5.84E-05	0.029253	31	6	-0.98683	0.99999	0.99999	1	16012	0	-0.98683	
32	ATPGV0B	8	1.33E-05	9.73E-05	0.0339	32	6	-1.1951	0.60353	0.60323	1	9315	0	-1.1951	
33	PCNA	3	1.34E-05	4.97E-05	0.027558	33	3	-1.4846	0.99999	0.99998	1	16011	0	-1.4846	
34	XRN1	4	1.45E-05	6.70E-05	0.030391	34	4	-1.3486	0.99999	0.99998	1	16010	0	-1.3486	
35	LAS1L	6	1.58E-05	8.00E-05	0.030528	35	5	-1.3598	0.99246	0.9924	1	15678	0	-1.3598	
36	DBR1	5	1.61E-05	7.50E-05	0.030528	36	3	-1.7201	0.55927	0.5599	1	8612	0	-1.7201	
37	VPS25	3	1.72E-05	6.02E-05	0.029253	37	3	-2.0229	0.99954	0.99953	1	15939	0	-2.0229	
38	EIF2B5	3	1.72E-05	6.02E-05	0.029253	38	3	-2.0229	0.98806	0.98819	1	15541	0	-2.0229	
39	HJURP	4	1.83E-05	8.49E-05	0.030941	39	4	-1.728	0.99979	0.9998	1	15973	0	-1.728	
40	RPAIN1	3	2.09E-05	6.84E-05	0.030391	40	3	-1.5587	0.99998	0.99998	1	16008	0	-1.5587	

**Εικόνα 8:** Παράδειγμα `.gene_summary.txt` αρχείου. Κάθε τέτοιο αρχείο αποτελείται από δεδομένα που αφορούν την ανάλυση και της negative (στήλες C,D,E,F,G,H) και της positive (I,J,K,L,M,N) selection.

## Μέθοδοι/Επεξεργασία δεδομένων (συνέχεια)

Αρχικά υπολόγισα τα hits. Στην πράξη, τα hits είναι τα γονίδια που η τιμή του positive και negative fdr τους βρίσκεται εντός ενός δεδομένου κατωφλίου (π.χ.  $<0,25$ ). Για κάθε ανάλυση που γίνεται με το MAGeCK παίρνω σαν αποτέλεσμα ποια γονίδια είναι θετικά ή αρνητικά εμπλουτισμένα. Για να δω εάν ένα γονίδιο είναι θετικά εμπλουτισμένο κοιτάζω την τιμή `pos.fdr` ενώ για το αρνητικά εμπλουτισμένο την `neg.fdr`.

Στην συγκεκριμένη μελέτη χρησιμοποίησα σαν κατώφλι  $FDR < 0.25$ .

Στην συνέχεια διαχώρισα τα δεδομένα των negative από αυτά των positive μελετών (`.gene_summary` αρχεία) και χρησιμοποιώντας την γλώσσα προγραμματισμού R (έκδοση 3.5.2.) προχώρησα στην μετέπειτα ανάλυσή τους για την κατασκευή σχημάτων και διαγραμμάτων για την οπτικοποίηση των αποτελεσμάτων. Αυτός ο διαχωρισμός ήταν απαραίτητος εξαιτίας του ότι το σύνολο των negative μελετών θα χρησιμοποιηθεί για την εξαγωγή διαφορετικών αποτελεσμάτων/συμπερασμάτων από αυτό των positive.



### 3.4.1. Ανάλυση των negative selection screens

Για την negative selection χρησιμοποιήθηκαν οι βιβλιοθήκες που φαίνονται στον Πίνακα 2.

	A	B
1	Μελέτη	Βιβλιοθήκη
2	Study_4_Day21	GeCKO_V2
3	Study_3_Day21	GeCKO_V2
4	Study_2_Day21	GeCKO_V2
5	Study_1_Day21	GeCKO_V2
6	Study_plasmidDay7R2	GeCKO_V1
7	Study_plasmidDay7R1	GeCKO_V1
8	Study_plasmidDay14R2	GeCKO_V1
9	Study_plasmidDay14R1	GeCKO_V1
10	Study_2_Day7	GeCKO_V1
11	Study_2_Day14	GeCKO_V1
12	Study_1_Day7	GeCKO_V1
13	Study_1_Day14	GeCKO_V1
14	original_tracr_rep2	Brunello_V1
15	original_tracr_rep1	Brunello_V1
16	modified_tracr_rep3	Brunello_V2
17	modified_tracr_rep2	Brunello_V2
18	modified_tracr_rep1	Brunello_V2
19	Day9_A375_WT_Dropout	Yusa_V1
20	Day9_A375_MAPK_dropout	Yusa_V1
21	Day7_A375_WT_Dropout	Yusa_V1
22	Day18_A375_MAPK_dropout	Yusa_V1
23		
24		

**Πίνακας 2:** Μελέτες ανά βιβλιοθήκη στην negative selection. Στην στήλη A παρουσιάζεται η κάθε μελέτη και στην στήλη B η αντίστοιχη βιβλιοθήκη της. Τα δεδομένα αφορούν την negative selection.

Αρχικά, ξεκίνησα με την αξιολόγηση της απόδοσης κάθε βιβλιοθήκης στο να αναγνωρίζει τα γνωστά essential και non-essential genes (απαραίτητα και μη-απαραίτητα γονίδια). Χρησιμοποίησα μία προκαθορισμένη λίστα γνωστών απαραίτητων και μη-απαραίτητων γονιδίων (684 απαραίτητων και 927 μη-απαραίτητων) <sup>[42]</sup>, την συνέκρινα με τα γονίδια των μελετών που συγκέντρωσα και έφτιαξα διαγράμματα πυκνότητας (density plots) για την οπτικοποίησή τους. (Κώδικας 2).

Επίσης έφτιαξα ROC καμπύλες (ROC curves) και υπολόγισα την περιοχή κάτω από κάθε μία καμπύλη (Area Under the Curve) για να επαληθεύσω τα αποτελέσματά μου (Κώδικας 4, Εικόνα 9).

Στη συνέχεια, μέσω των FDR που είχα ήδη υπολογίσει, βρήκα τα εμπλουτισμένα μονοπάτια που βρίσκονται τα γονίδια που ήδη υπολόγισα σαν hits. Στην ουσία, αυτά τα μονοπάτια περιέχουν τα απαραίτητα γονίδια που βρήκα.

Έπειτα, έκανα ανάλυση εμπλουτισμού (pathway enrichment analysis) με την χρήση του στατιστικού πακέτου «Enrichr» της R <sup>[44]</sup> και της διαδικτυακής βάσης δεδομένων KEGG\_2016<sup>[43]</sup>, προσπελάσιμη μέσω του ακόλουθου συνδέσμου: <https://www.genome.jp/kegg/>. Στην KEGG βρίσκονται όλα τα σηματοδοτικά μονοπάτια που θέλω να δω εάν είναι εμπλουτισμένα. Χρησιμοποιώντας αυτή τη βάση δεδομένων εντόπισα τα κύρια μονοπάτια που είχαν αφαιρεθεί από το screen λόγω εξελικτικής επιλογής. (Κώδικας 2).

Τέλος, για να αξιολογήσω τα essential γονίδια των A375 κυτταρικών σειρών, ξέροντας την εξάρτηση τους από τα MAPK pathways, κοίταξα στα συγκεκριμένα pathways για να βρω τις εξαρτήσεις. Χρησιμοποιώντας την KEGG\_2016 βάση δεδομένων βρήκα τα γονίδια που εμπλέκονται στο MAPK pathway (217 γονίδια συνολικά) (Παράρτημα) και έπειτα διέκρινα αυτά που εμφανίζονται σε τουλάχιστον μία μελέτη και έφτιαξα heatmaps. (Κώδικας 3, Εικόνα 12).

Την ίδια διαδικασία ακολούθησα και για άλλα ενδιαφέροντα σηματοδοτικά μονοπάτια, τα MTOR, TGFbeta και MELANOMA (Κώδικας 6, 7, 8, Εικόνα 14,15,16).

Συγκεκριμένα, αφού χώρισα το αρχείο Interesting pathways (Παράρτημα) σε 3 διαφορετικά αρχεία (KEGG\_TGF\_BETA\_SIGNALING\_PATHWAY.csv, KEGG\_MTOR\_SIGNALING\_PATHWAY.csv, KEGG\_MELANOMA.csv) με τα γονίδια που συμμετέχουν σε κάθε μονοπάτι (MTOR, TGF beta και Melanoma) προχώρησα σε περαιτέρω επεξεργασία για το καθένα.

Αρχικά για το αρχείο KEGG\_TGF\_BETA\_SIGNALING\_PATHWAY.csv έφτιαξα ένα βοηθητικό αρχείο με όνομα Identified\_KEGG\_TGF\_BETA\_all με όλα τα γνωστά γονίδια από τις 21 μελέτες που το ID τους αντιστοιχίζεται με αυτό των γονιδίων από το KEGG\_TGF\_BETA\_SIGNALING\_PATHWAY.csv αρχείο. Στην συνέχεια χώρισα σε κατηγορίες τα γονίδια ανάλογα με το *fdr* τους. Με αυτό τον τρόπο ξέρω ποια από τα γονίδια είναι σημαντικά και πόσο. Αφαίρεσα τα γονίδια που ανήκαν στην κατηγορία D και εμφανίζονταν σε παραπάνω από 20 μελέτες και τέλος έφτιαξα τον heatmap χάρτη για την απεικόνιση τους (Κώδικας 6, Εικόνα 17,18).

Την παραπάνω διαδικασία ακολούθησα και για τα άλλα 2 μονοπάτια με την διαφορά ότι κάθε φορά άλλαζα τα αντίστοιχα ονόματα των φακέλων (Κώδικας 7,8).

### 3.4.2. Ανάλυση των positive selection screens

Για την positive selection χρησιμοποιήθηκαν οι βιβλιοθήκες που φαίνονται στον Πίνακα 3.

	A	B
1	Μελέτη	Βιβλιοθήκη
2	Study_4_GeCKOv1_lentiguide	GeCKO_V1
3	Study_3_GeCKOv1_lentiguide	GeCKO_V1
4	Study_2_lentiCRISPRv2	Brunello
5	Study_2_lentiCRISPRv1	Brunello
6	Study_2_GeCKOv2lentiguide	GeCKO_V2
7	Study_2_GeCKOv1_lentiguide	GeCKO_V1
8	Study_2_Avana_lentiguide	Avana
9	Study_1_lentiCRISPRv2	Brunello
10	Study_1_lentiCRISPRv1	Brunello
11	Study_1_GeCKOv2lentiguide	GeCKO_V2
12	Study_1_GeCKOv1_lentiguide	GeCKO_V1
13	Study_1_Avana_lentiguide	Avana
14	Study_Day7	GeCKO_V1
15	Study_Day14	GeCKO_V1
16	Study_2_Day7	GeCKO_V1
17	Study_2_Day14	GeCKO_V1
18	Study_1_Day7	GeCKO_V1
19	Study_1_Day14	GeCKO_V1
20	Day27_A375_MAPK_Resistant_count	Yusa_V1
21	Day25_A375_WT_Trametinib_treated	Yusa_V1
22	Day24_A375_WT_Trametinib_treated	Yusa_V1
23	Day18_A375_MAPK_trametinib_treated	Yusa_V1
24		

**Πίνακας 3:** Μελέτες ανά βιβλιοθήκη στην positive selection. Στην στήλη A παρουσιάζεται η κάθε μελέτη και στην στήλη B η αντίστοιχη βιβλιοθήκη της. Τα δεδομένα αφορούν την positive selection

Έχοντας ως δεδομένα τα γονίδια από τις μελέτες και την λίστα με τα γονίδια του MAPK pathway βρήκα τα γονίδια που είναι κοινά και στις δύο λίστες. Έπειτα τα κατέταξα σε κατηγορίες ανάλογα με το FDR τους και το σε πόσες μελέτες εμφανίζεται το καθένα. Σαν έξοδο πήρα ένα αρχείο που περιέχει τα γονίδια, την κατηγορία fdr, την μελέτη και την συχνότητα που εμφανίζεται το κάθε γονίδιο. Αυτό το χρησιμοποίησα για να φτιάξω heatmaps (Κώδικας 10). Επιπλέον, ξέροντας κάποια γονίδια που είναι ήδη γνωστά στο να προσφέρουν ανθεκτικότητα, έφτιαξα έναν ακόμη heatmap χωρίς αυτά (Κώδικας 10).

Τα αποτελέσματα όλων των παραπάνω διαδικασιών επεξηγούνται στο επόμενο κεφάλαιο.

## Κεφάλαιο 4: Αποτελέσματα

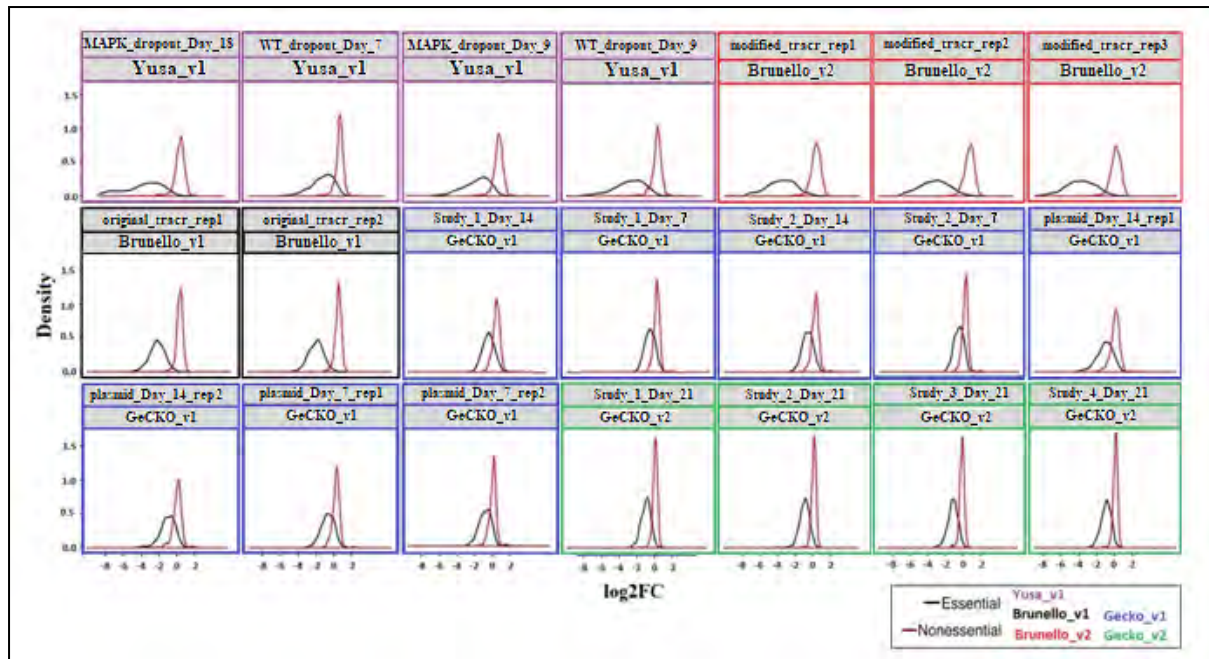
### 4.1. Αποτελέσματα της ανάλυσης των negative selection screens στις A375 κυτταρικές σειρές του μελανώματος:

#### α) Αξιολόγηση της απόδοσης κάθε βιβλιοθήκης στο να αναγνωρίζει γνωστά essential και non-essential (απαραίτητα και μη-απαραίτητα) γονίδια

Δεδομένης της λίστας με τα γονίδια από τις μελέτες που συνέλεξα και των γνωστών essential και non-essential γονιδίων (απαραίτητων και μη-απαραίτητων), έφτιαξα διαγράμματα πυκνότητας και καμπύλες ROC, προκειμένου να οπτικοποιήσω την απόδοση της κάθε βιβλιοθήκης στο να διαχωρίζει τις 2 κατηγορίες γονιδίων.

Σε κάθε ένα από τα 21 διαφορετικά διαγράμματα πυκνότητας (Εικόνα 9) παρατηρούμε την απόδοση της κάθε βιβλιοθήκης μέσω των μετατοπίσεων των καμπυλών. Στον άξονα x απεικονίζεται το  $\log_2FC$  (Fold Change) το οποίο περιγράφει το πόσο έχει διαφοροποιηθεί το control (δηλαδή τα δεδομένα πριν το πείραμα) από το treatment (δηλαδή τα δεδομένα μετά το πέρας του πειράματος). Όσο μεγαλύτερη είναι η τιμή του  $\log_2FC$  τόσο μεγαλύτερη είναι και η αλλαγή του. Στον άξονα y φαίνεται η πυκνότητα.

Γενικά στα διαγράμματα πυκνότητας προσδοκούμε μεγάλη μετατόπιση της καμπύλης προς τα αριστερά. Στα διαγράμματά μου καλύτερες είναι οι περιπτώσεις στις οποίες οι γραφικές παραστάσεις των essential γονιδίων είναι κατά πολύ μετατοπισμένες προς τα αριστερά σε σχέση με αυτές των μη essential. Αυτό φαίνεται να είναι έντονο στις περιπτώσεις των Yusa v1 και Brunello βιβλιοθηκών. Έτσι, οι μελέτες που έκαναν χρήση της Yusa v1 και Brunello βιβλιοθήκης είναι καλύτερες καθώς η μετατόπιση των 2 παραστάσεων είναι μεγαλύτερη και η επικάλυψή τους ελάχιστη. Χειρότερες φαίνονται να είναι οι παραστάσεις των Gecko v1 και Gecko v2 βιβλιοθηκών για τον αντίθετο λόγο.

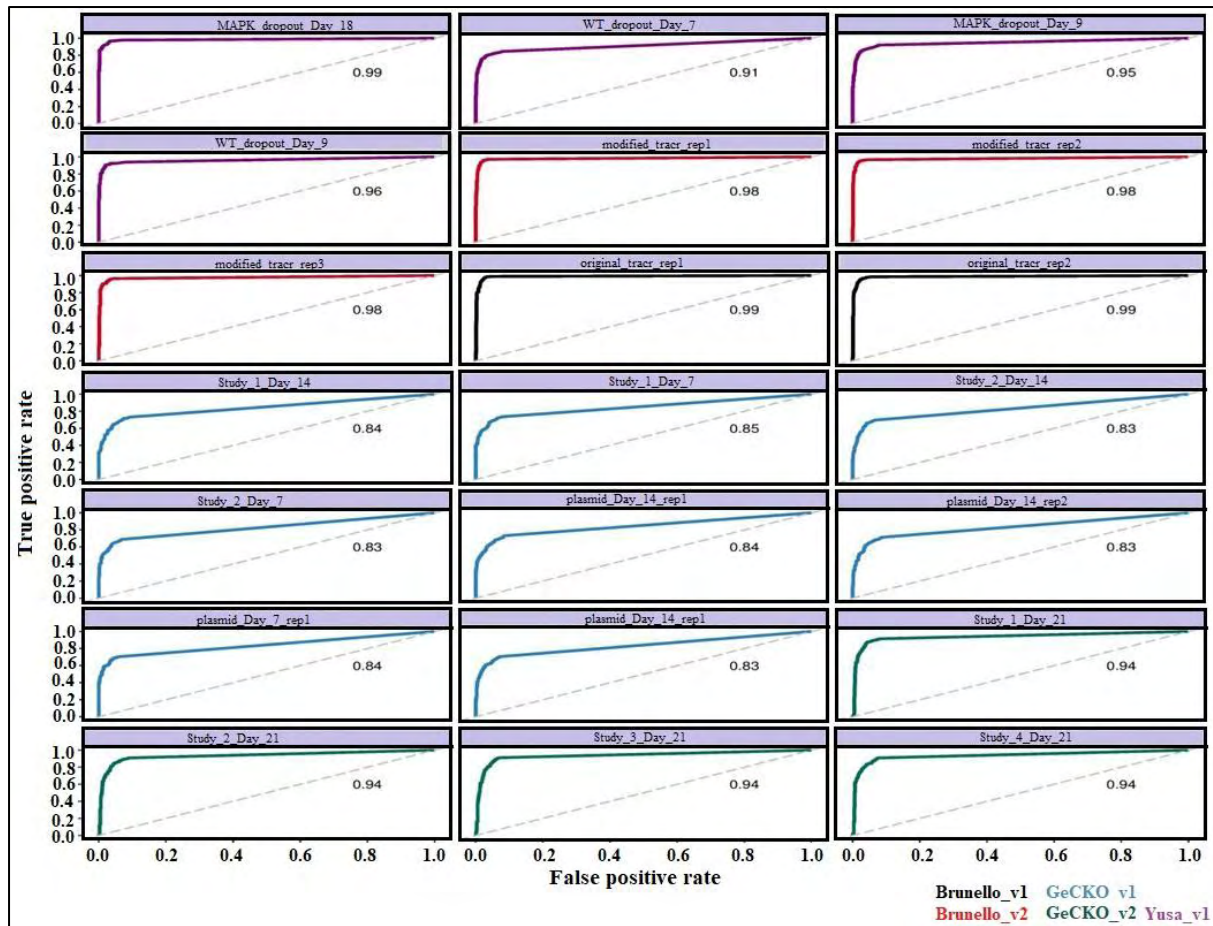


**Εικόνα 9: Διαγράμματα πυκνότητας για την απεικόνιση των αποτελεσμάτων ανά μελέτη και βιβλιοθήκη.** Καλύτερες είναι οι περιπτώσεις στις οποίες η μετατόπιση των καμπυλών για τα non-essential γονίδια είναι όσο το δυνατόν πιο αριστερά από αυτές των essential και οι επικαλύψεις οι μικρότερες δυνατές.

Οι καμπύλες ROC είναι καμπύλες πιθανότητας οι οποίες υποδηλώνουν το πόσο καλό είναι ένα μοντέλο στο να διαχωρίζει δύο πράγματα, στην δική μας περίπτωση πόσο καλή είναι μία βιβλιοθήκη στο να κατατάσσει τα essential (απαραίτητα) γονίδια ως essential και τα non-essential (μη απαραίτητα) ως non-essential. Πολύ χρήσιμος είναι και ο υπολογισμός του εμβαδού κάτω από την καμπύλη (Area Under the Curve-AUC) καθώς εκφράζει το πόσο επιτυχημένη είναι η δοκιμασία. Οι τιμές του κυμαίνονται από 0 έως 1, με καλύτερη περίπτωση την AUC=1 που ισοδυναμεί με μία απόλυτα επιτυχημένη πρόβλεψη, και χειρότερη την AUC=0.

Στην περίπτωση μου, οι καμπύλες ROC χρησιμοποιήθηκαν για την ποσοτικοποίηση των μετρήσεων και μέσω των αντίστοιχων περιοχών κάτω από τις καμπύλες (Area Under the Curve-AUC) που υπολόγισα για κάθε μία βιβλιοθήκη ξεχωριστά, έγινε ξεκάθαρη το ποια βιβλιοθήκη είναι καλύτερη και ποια χειρότερη (Εικόνα 10). Συγκεκριμένα, καλύτερες είναι οι Yusa\_v1 και Brunello βιβλιοθήκες, ενώ για ακόμη μία φορά χειρότερη είναι η GeCKO\_v1 αφού είναι αυτή με το μικρότερο AUC. Επειδή η GeCKO V1 κρίθηκε ως χειρότερη βιβλιοθήκη κρίθηκε απαραίτητο να αφαιρεθεί από την συνέχεια της ανάλυσης.



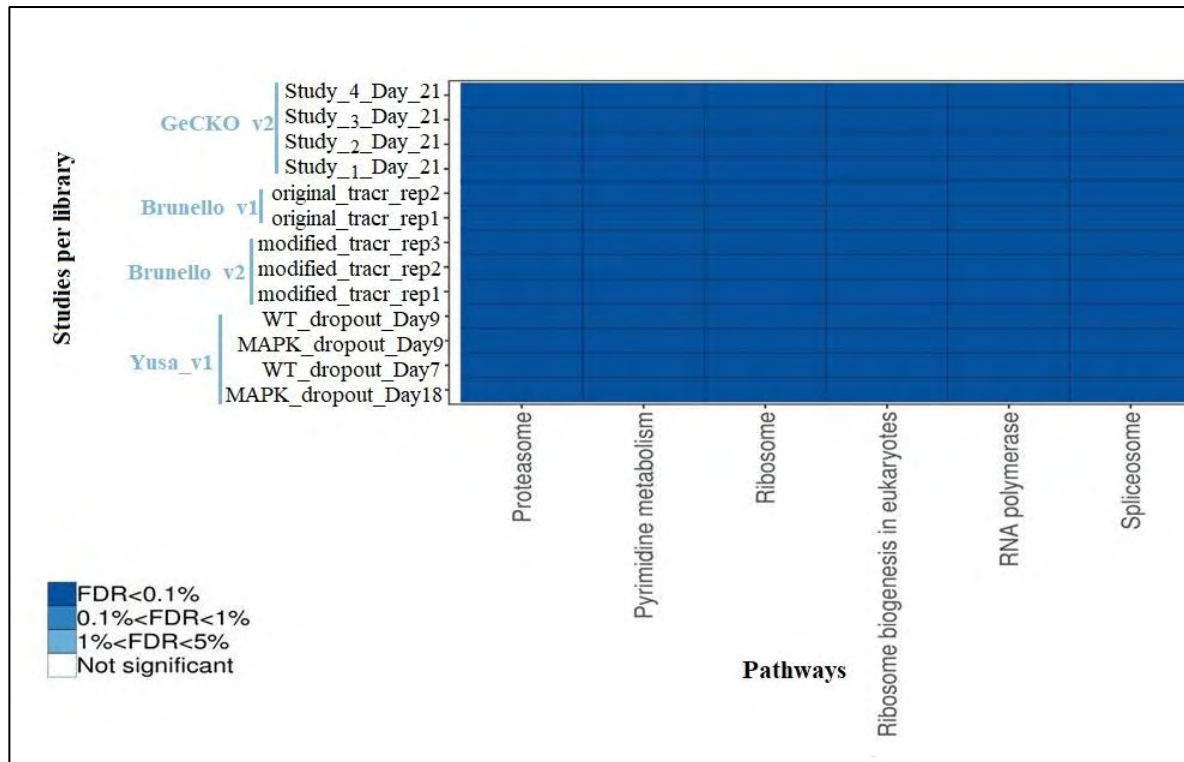


**Εικόνα 10: Καμπύλες ROC για την οπτικοποίηση των αποτελεσμάτων ανά μελέτη και βιβλιοθήκη. Καλύτερες είναι οι περιπτώσεις στις οποίες η περιοχή κάτω από την καμπύλη είναι μεγαλύτερη.**

### β) Αξιολόγηση της απόδοσης κάθε βιβλιοθήκης στο να αναγνωρίζει τα γνωστά essential μονοπάτια (pathways)

Για την εύρεση των μονοπατιών βασίστηκα στο fdr που υπολόγισα και για την οπτικοποίηση των αποτελεσμάτων έφτιαξα heatmaps (Εικόνα 10). Στον οριζόντιο άξονα αναπαριστώνται τα σηματοδοτικά μονοπάτια που περιέχουν τα περισσότερα απαραίτητα γονίδια από τις A375 κυτταρικές σειρές του μελανώματος. Στον άξονα χ απεικονίζονται τα pathways που έχουν εντοπιστεί σε περισσότερες από μία μελέτες, ενώ στον άξονα y οι μελέτες που είχα στην διάθεση μου. Κάθε τιμή μπορεί να πάρει μία απόχρωση του μπλε ή λευκού, ανάλογα με την τιμή του FDR που υπολόγισα. Εδώ θέλουμε το κάθε κουτί να είναι όσο το δυνατόν περισσότερο σκούρο μπλε, έτσι ώστε να έχουμε τα βέλτιστα αποτελέσματα.

Όλες οι μελέτες έχουν εντοπίσει τα γνωστά essential μονοπάτια στον βέλτιστο βαθμό. Τα συγκεκριμένα αποτελέσματα είναι αναμενόμενα λόγω των διαγραμμάτων πυκνότητας (Εικόνα 9) και της ROC καμπύλης (Εικόνα 10), τα οποία ανέδειξαν και αυτά την Yusa ως την καλύτερη βιβλιοθήκη. Ουσιαστικά οι heatmaps χρησιμοποιήθηκαν ως μία επαλήθευση του προηγούμενου σταδίου.

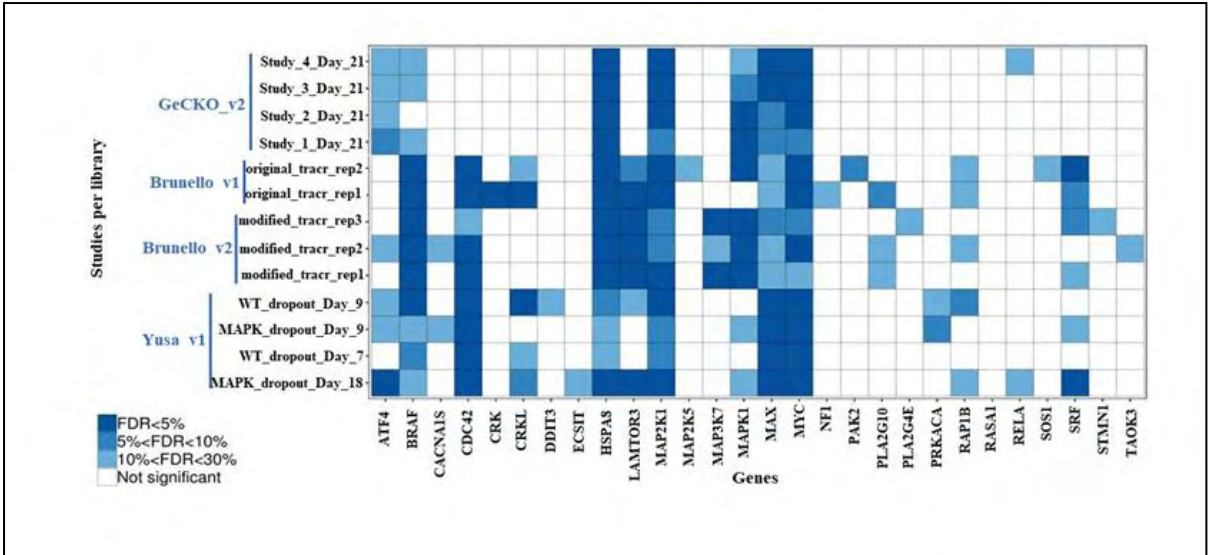


**Εικόνα 11:** Heatmaps για την οπτικοποίηση της απόδοσης των βιβλιοθηκών.

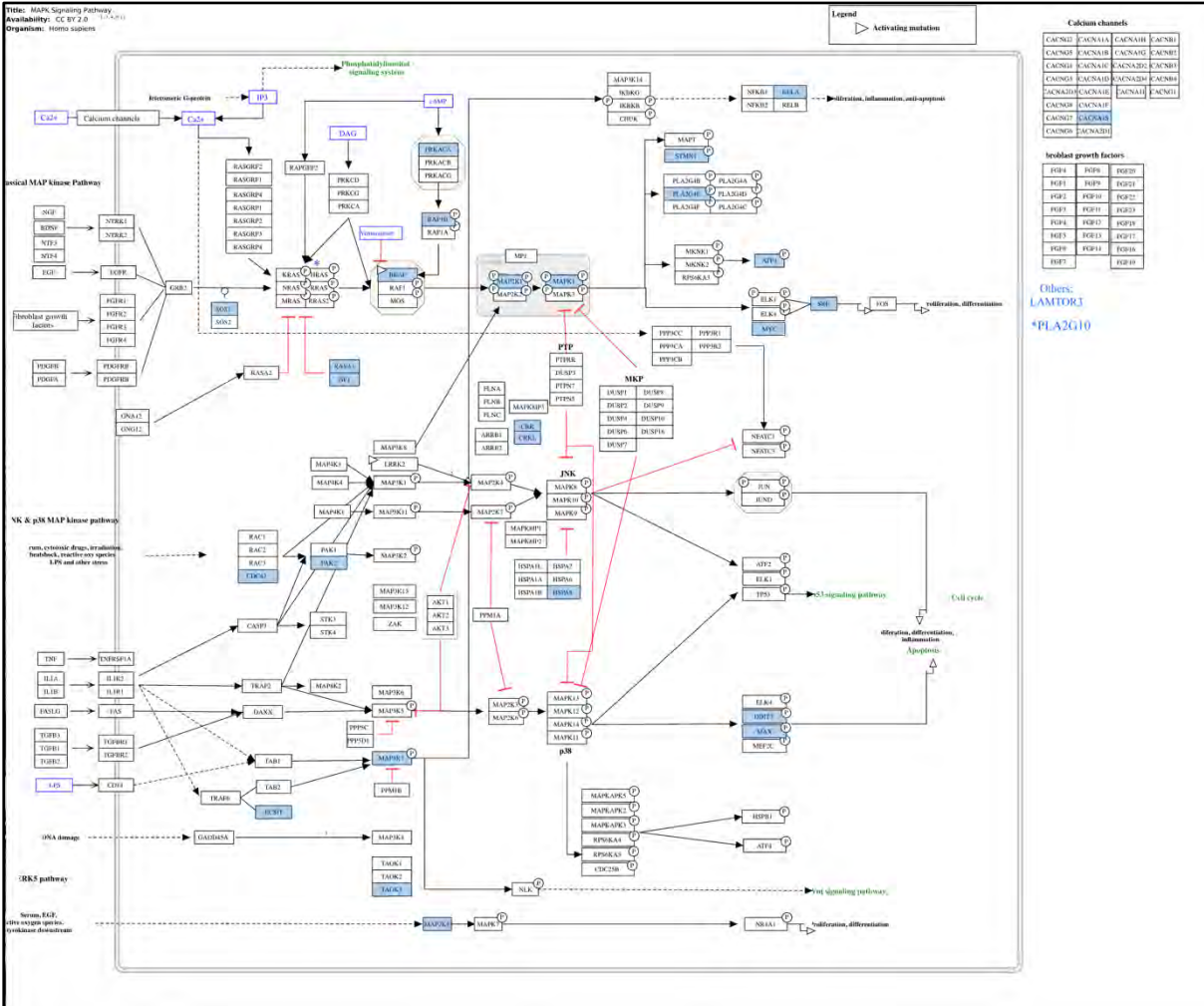
### γ)Εύρεση των essential γονιδίων στις A375 κυτταρικές σειρές.

Γνωρίζοντας ότι τα γονίδια των A375 είναι εξαρτημένα στο MAPK σηματοδοτικό μονοπάτι, έψαξα σε αυτό για να βρω τις εξαρτήσεις (Εικόνα 11). Επίσης, την ίδια αναζήτηση έκανα και στα μονοπάτια κυτταρικής σηματοδότησης MTOR, TGF beta και ‘Melanoma’ για να εντοπίσω και εκεί αν υπάρχουν essential γονίδια. Τα αποτελέσματα της αναζήτησης φαίνονται στους παρακάτω heatmaps (Εικόνες 12,13,14). Στον οριζόντιο άξονα αναπαριστώντα τα γονίδια που έχουν εμφανιστεί σε τουλάχιστον μία μελέτη και ανάλογα με το FDR τους πόσο σημαντικά είναι για κάθε μελέτη.

Στην (Εικόνα 11) βλέπουμε μέσω του heatmap τα γονίδια που είναι απαραίτητα στο MAPK μονοπάτι. Το BRAF γονίδιο είναι hit στις περισσότερες μελέτες. Αυτό είναι αναμενόμενο επειδή τα κύτταρα έχουν μετάλλαξη σε αυτό και εξαρτώνται από αυτό. Αντίθετα, γονίδια όπως το MYC μπορεί να έχουν εντοπιστεί σε πολλές μελέτες αλλά δεν λαμβάνονται υπόψη γιατί ανήκουν στην κατηγορία των “Core-essential”, δηλαδή των γονιδίων που είναι σημαντικά για όλα τα κύτταρα και όχι μόνο στα κύτταρα με το μεταλλαγμένο BRAF. Επιπλέον, εντοπίστηκε και ένα σύνολο άλλων γονιδίων όπως τα MAP2K1, MAPK1, SRF τα οποία είναι essential (Εικόνα 12).



**Εικόνα 12:** Γονίδια του MAPK σηματοδοτικού μονοπατιού τα οποία είναι απαραίτητα στην A375 κυτταρική σειρά του μελανώματος.



**Εικόνα 13:** Essential γονίδια στο MAPK σηματοδοτικό μονοπάτι (τονισμένα με γαλάζιο χρώμα)

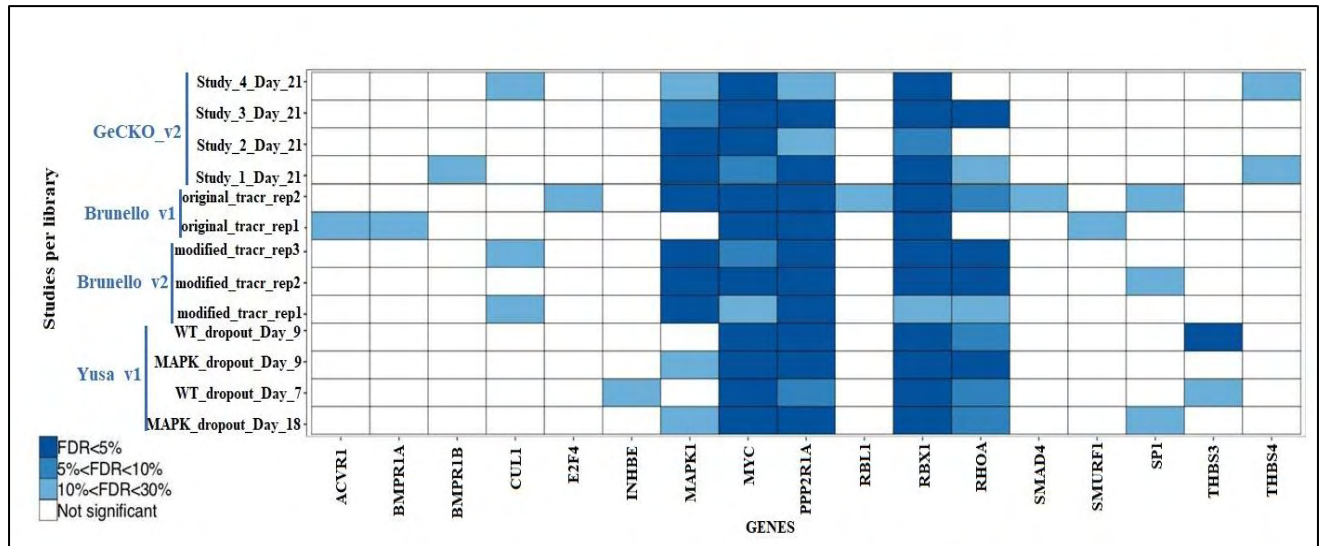


Στο MAPK (Εικόνα 12) και TGF beta pathway (Εικόνα 14) έχουν εντοπιστεί πολλά κοινά γονίδια ως essential. Τα MYC και RBX1 γονίδια είναι core-essential και τα RHOA έχουν εντοπιστεί να είναι essential και σε άλλες κυτταρικές σειρές καρκίνου του δέρματος.

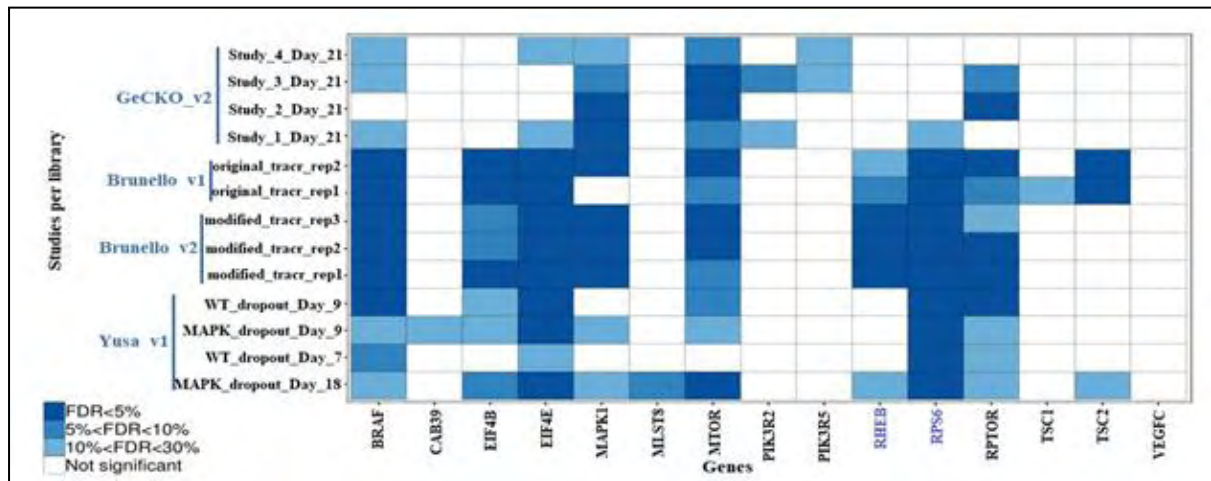
Στο MAPK (Εικόνα 12) και MTOR pathway (Εικόνα 15) και πάλι υπάρχουν επικαλύψεις, δηλαδή κοινά γονίδια, π.χ. τα BRAF, MAPK1.

Στο MAPK (Εικόνα 11) και Melanoma pathway (Εικόνα 16) εμφανίζονται επικαλύψεις στα γονίδια BRAF, MAP2K1 κ.α.

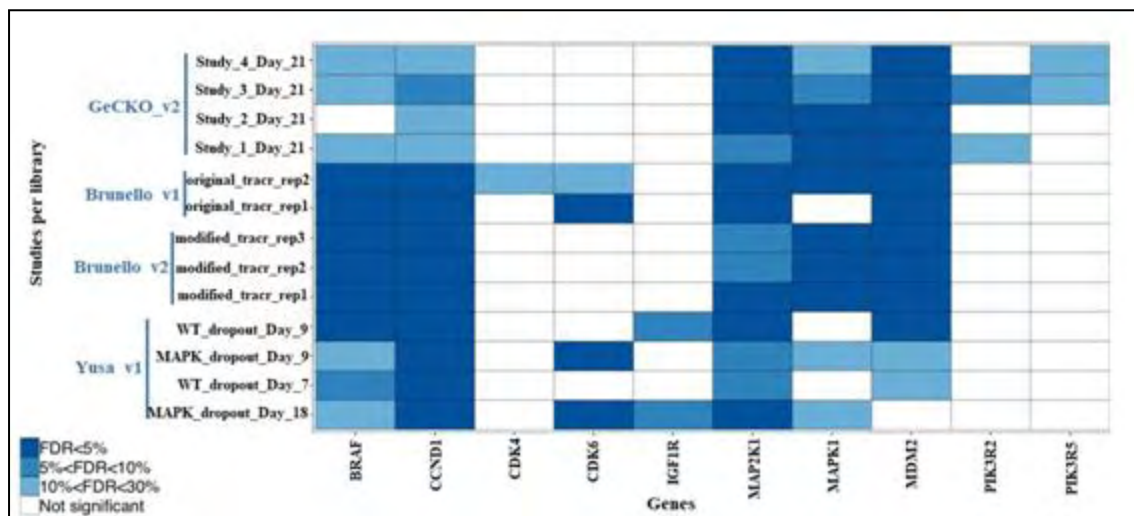
Οι 3 heatmaps εμφανίζονται στις παρακάτω εικόνες (Εικόνα 14, 15, 16).



**Εικόνα 14: Heatmap για το TGF beta σηματοδοτικό μονοπάτι.**



**Εικόνα 15: Heatmap για το MTOR σηματοδοτικό μονοπάτι. Τα γονίδια που είναι γραμμένα με μπλε χρώμα ανήκουν στην κατηγορία των “essential” στην A375 κυτταρική σειρά του μελανώματος.**



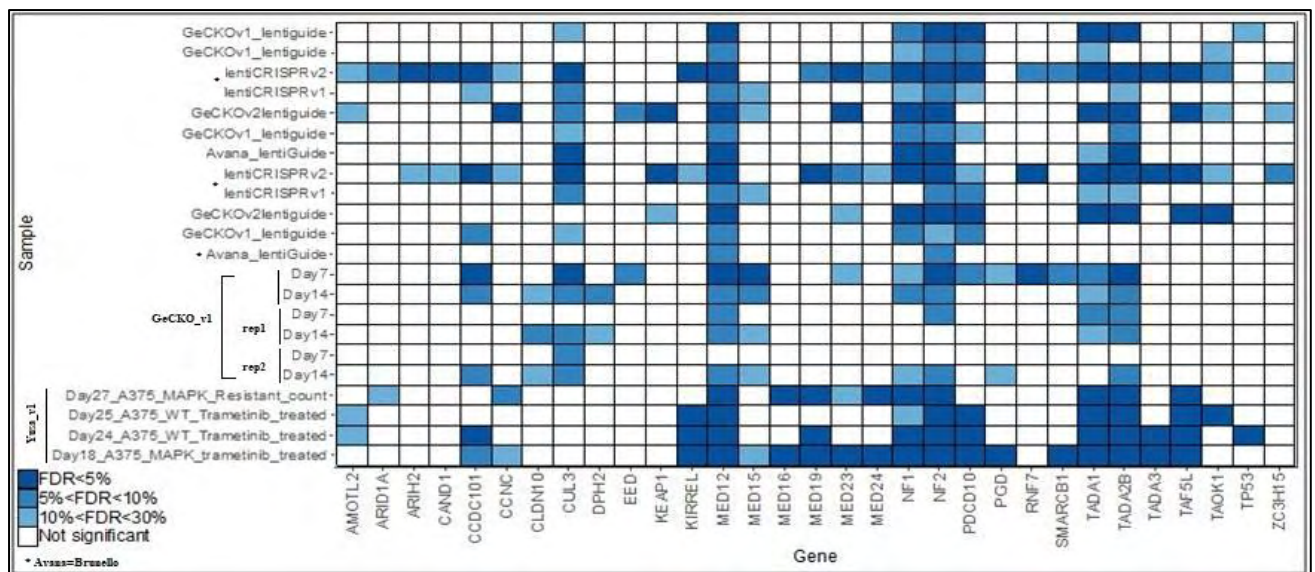
**Εικόνα 16: Heatmap για το Melanoma σηματοδοτικό μονοπάτι.**

#### 4.2. Αποτελέσματα των positive selection screens στις A375 κυτταρικές σειρές του μελανώματος.

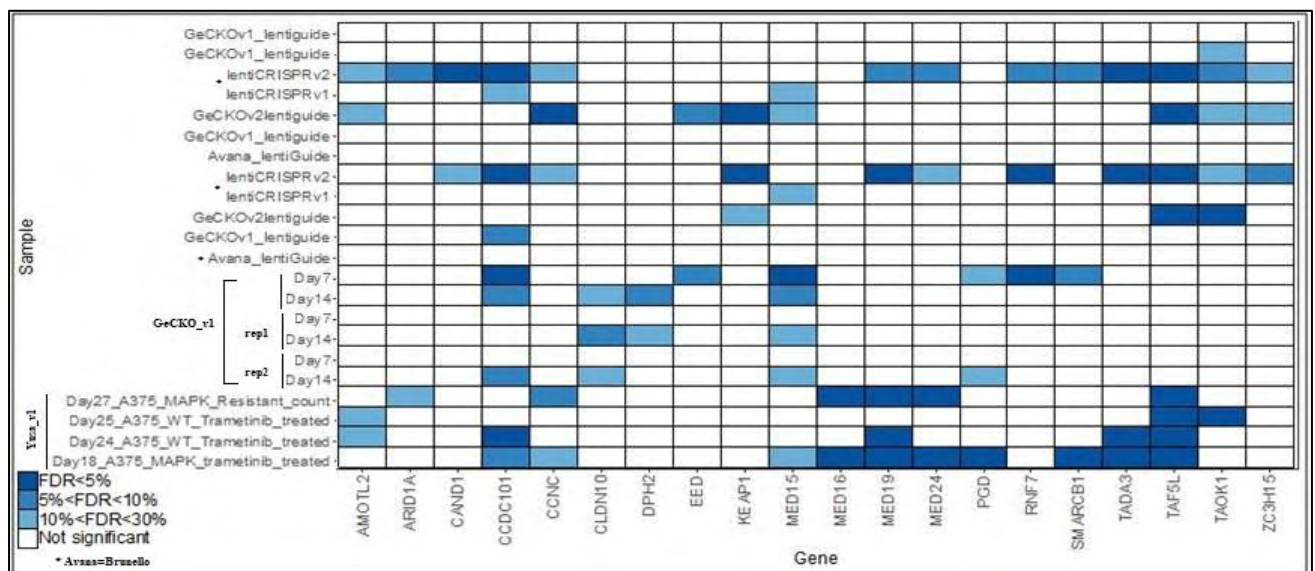
##### Εύρεση των γονιδίων που είναι ανθεκτικά στους MEK και BRAF αναστολείς.

Οι heatmaps που πήρα σαν αποτέλεσμα (Εικόνα 17, Εικόνα 18) έχουν στον άξονα x τα γονίδια που προσφέρουν ανθεκτικότητα και στον άξονα y τις μελέτες. Το χρώμα που είναι ζωγραφισμένο το κάθε κουτάκι συμβολίζει την κατηγορία που ανήκει το κάθε γονίδιο ανάλογα με το *fd*r. Όσο περισσότερο σκούρο είναι το κάθε κουτάκι τόσο πιο σημαντικό γονίδιο περιέχει.

Η Εικόνα 17 έχει όλα τα γονίδια, συμπεριλαμβανομένων εκείνων που ανακαλύψαμε στο εργαστήριο από τα δικά μας πειράματα, ενώ από την Εικόνα 18 έχουν αφαιρεθεί αυτά που ξέραμε εκ των προτέρων από προηγούμενες μελέτες<sup>[27]</sup> ότι προσφέρουν ανθεκτικότητα στο κύτταρο (Παράρτημα).



**Εικόνα 17:** Heatmap με όλα τα γονίδια (γνωστά και μη) που είναι ανθεκτικά στους MEK και BRAF αναστολείς.



**Εικόνα 18:** Heatmap μετά την αφαίρεση των γνωστών από πριν γονιδίων.

## Κεφάλαιο 5: Συζήτηση-Συμπεράσματα-Βιβλιογραφία

### 5.1. Συζήτηση-Συμπεράσματα

Συλλέγοντας δεδομένα από 43 μελέτες που έχουν προκύψει από CRISPR screens στις A375 κυτταρικές σειρές του μελανώματος και κάνοντας μετα-ανάλυση σε αυτά κατέληξα στα εξής συμπεράσματα. Οι Yusa\_v1 και Brunello βιβλιοθήκες είναι οι καλύτερες βιβλιοθήκες στο να αναγνωρίζουν τα γνωστά essential (απαραίτητα) γονίδια και σηματοδοτικά μονοπάτια στις A375 κυτταρικές σειρές και αποτελούν ένα αξιόπιστο εργαλείο έτσι ώστε να χρησιμοποιούνται στις CRISPR screens. Αυτά τα αποτελέσματα ήταν αναμενόμενα καθώς η Yusa\_v1 βιβλιοθήκη είναι μία από τις πιο καινούργιες βιβλιοθήκες sgRNA που υπάρχουν και επιπλέον τα αποτελέσματα ταυτοποιήθηκαν μεταξύ τους μέσω των διαγραμμάτων πυκνότητας και των καμπυλών ROC. Επίσης, επόμενο εύρημα είναι και ένα υψηλής ποιότητας σετ απαραίτητων γονιδίων στις ίδιες κυτταρικές σειρές. Σε αυτά τα γονίδια ανήκουν τα MYC, RBX1, RPS6, CCND1, CDC42. Τέλος, βρήκα και ένα σύνολο γονιδίων τα οποία είναι ανθεκτικά στους BRAF και MEK αναστολείς.

Όλα τα παραπάνω ευρήματα μπορούν να χρησιμοποιηθούν στο μέλλον από το group. Συγκεκριμένα, θα συνεχίσουν να χρησιμοποιούν την Yusa\_v1 βιβλιοθήκη που αποδείχτηκε ότι είναι η καλύτερη για screening στις A375 κυτταρικές σειρές του μελανώματος, και όχι και τις υπόλοιπες βιβλιοθήκες. Ακόμη, επειδή πολλά από τα γονίδια που επηρεάζουν τα μονοπάτια που μας ενδιαφέρουν είναι essential, για να μπορέσουν να δουν την επίδρασή τους στα σηματοδοτικά μονοπάτια που μελετούν θα κάνουν screen σε arrayed μορφή (δηλαδή θα εξετάζουν ένα sgRNA σε κάθε σύνολο κυττάρων) και όχι σε pooled (χρησιμοποιώντας ολόκληρη την βιβλιοθήκη των sgRNA) όπως πριν. Τέλος, για να μελετήσουν τον τρόπο που τα μονοπάτια αναπροσαρμόζονται σε περίπτωση ανθεκτικότητας σε φάρμακα θα κάνουν screen σε arrayed μορφή σε όλο το σύνολο των γονιδίων (τα γνωστά από πριν και τα καινούργια που βρήκαμε ) που είναι ανθεκτικά στους BRAF και MEK αναστολείς και θα δουν το πώς επηρεάζουν τα μονοπάτια μετρώντας τις αλλαγές στην φωσφορυλίωση συγκεκριμένων pathway markers (π.χ. χημικά μόρια ή φθορίζουσα πρωτεΐνη).



## 5.2. Βιβλιογραφία

- [1] Barrangou R(2015). “ The roles of CRISPR-Cas systems in adaptive immunity and beyond”. *Current Opinion in Immunology*. 32: 36-41  
doi:10.1016/j.coi.2014.12.008
- [2] Sun L, Lutz B & Tao Y, “The CRISPR/Cas9 system for gene editing and its potential application in pain research”, *Translational perioperative and pain medicine*, 1(3), 22-33, (2016)
- [3] Kirkpatrick B, “Cut Out the Hype: Gene Editing With CRISPR and the Truth about Superhuman ‘Designer Babies’”. *What is Epigenetics*, (2017)
- [4] Lin S, Staahl B, Alla R & Doudna J, “Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery”, *Elife*, 3, e04766, (2014)  
doi:10.7554/eLife.04766
- [5] Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA & Charpentier E, “A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity”, *Science*, 337(6096), 816-21, (2012)  
doi:10.1126/science.1225829
- [6] Bressan RB, Dewari PS, Kalantzaki M, Gangoso E, Matjusaitis M, Garcia-Diaz C, Blin C, Grant V, Bulstrode H, Gogolok S, Skarnes WC & Pollard SM, “Efficient CRISPR/Cas9-assisted gene targeting enables rapid and precise genetic manipulation of mammalian neural stem cells”, *Development*, 144(4), 635-648, (2017)
- [7] Kaminski R, Chen Y, Fischer T, Tedaldi E, Napoli A, Yonggang Z, Karn J, Hu W & Khalili K, “Elimination of HIV-1 Genomes from Human T-lymphoid Cells by CRISPR/Cas9 Gene Editing”, *Scientific reports*, 6(28213), (2016)  
doi:10.1038/srep28213
- [8] Cyranoski D & Ledford H, “Genome-edited baby claim provokes international outcry”, *Nature*, 563(7733), 607-608, (2018)  
doi: 10.1038/d41586-018-07545-0
- [9] Hammond A, Galizi R, Kyrou K, Simoni A, Siniscalchi C, Katsanos D, Gribble M, Baker D, Marois E, Russel S, Burt A, Windbichler N, Crisanti A & Nolan T, “A CRISPR-Cas9 gene drive system targeting female reproduction in the malaria mosquito vector *Anopheles gambiae*”, *Nature Biotechnology*, 34, 78-83, (2016)
- [10] Jeon S, Lim JM, Lee HG, Shin SE, Kang N, Park Y, Oh H, Jeong WJ, Jeong Br & ChangY, “Current status and perspectives of genome editing technology for microalgae”, *Biotechnology for Biofuels*, 10:267, (2017)
- [11] Jaganathan D, Ramasamy K, Sellamuthu G, Jayabalan S & Venkataraman, “CRISPR for Crop Improvement: An Update Review”, *Frontiers in plant science*, 9:985, (2018)  
doi:10.3389/fpls.2018.00985

[12] Sharma S & Petsalaki E, “Application of CRISPR-Cas9 Based Genome-wide Screening Approaches to Study Cellular Signalling Mechanisms”, *International Journal of Molecular Sciences*, 19 (4), 933, (2018)

[13] Hart T, Tong AHY, Chan K, Van Leeuwen K, Seetharaman A, Aregger M, Chandrashekar M, Hustedt N, Seth S, Noonan A, Habsid A, Sizova O, Nedyalkova L, Climie R, Tworzyanski L, Lawson K, Sartori MA, Alibeh S, Tieu D, Masud S, Mero P, Weiss A, Brown KR, Usaj M, Billmann M, Rahman M, Constanzo M, Myers CL, Andrews BJ, Boone C, Durocher D & Moffat J, “Evaluation and Design of Genome-Wide CRISPR-Cas9 Knockout Screens”, *G3 (Bethesda)*, 7(8), 2719-2727, (2017)  
doi:10.1534/g3.117.041277

[14] Doench J, “Am I ready for CRISPR? A user’s guide to genetic screens”, *Nature Reviews Genetics*, 19, 67-80, (2018)

[15] Ong SH, Li Y, Koike-Yusa H & Yusa K, “Optimised metrics for CRISPR-KO screens with second-generation gRNA libraries”, *Scientific Reports*, 7(1), (2017)  
doi: 10.1038/s41598-017-07827-z

[16] Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I, Tothova Z, Wilen C, Orchard R, Virgin HW, Listgarten J & Root DE, “Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9”, *Nature Biotechnology*, 34(2), 184-191, (2016)  
doi: 10.1038/nbt.3437

[17] Tzelepis K, Koike-Yusa H, De Braekeleer E, Li Y, Metzakopian E, Dovey OM, Mupo A, Grinkevich V, Li M, Mazan M, Gozdecka M, Ohnishi S, Cooper J, Patel M, McKerrell T, Chen B, Domingues AF, Gallipoli P, Teichmann S, Ponstingl H, McDermott U, Saez-Rodriguez J, Huntly BJP, Iorio F, Pina C, Vassiliou GS & Yusa K, “A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia”, *Cell Reports*, 17(4), 1193-1205, (2016)  
doi: 10.1016/j.celrep.2016.09.079

[18] Meyers RM, Bryan JG, McFarland JM, Weir BA, Sizemore AE, Xu H, Dharia NV, Montgomery PG, Cowley GS, Pantel S, Goodale A, Lee Y, Ali LD, Jiang G, Lubonja R, Harrington WF, Strickland M, Wu T, Hawes DC, Zhivich VA, Wyatt MR, Kalani Z, Chang JJ, Okamoto M, Stegmaier K, Golub TR, Boehm JS, Vazquez F, Root DE, Hahn WC & Tsherniak A, “Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells”, *Nature Genetics*, 49(12), 1779-1784, (2017)  
doi: 10.1038/ng.3984

[19] Sanjana NE, Shalem O & Zhang F, “Improved vectors and genome-wide libraries for CRISPR screening”, *Nature Methods*. 11(8), 783-784, (2014)  
doi: 10.1038/nmeth.3047

- [20] Sanson K R, Hanna R E, Hegde M, Donovan K F, Strand C, Sullender M E, Vaimberg E W, Goodale A, Root D E, Piccioni F & Doench J G, “Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities” *Nature communications*, 9(1), 5416, (2018) doi:10.1038/s41467-018-07901-8
- [21] Wang T, Wei J, Sabatini D & Lander E, “Genetic Screens in Human Cells Using the CRISPR-Cas9 System”, *Science*, 343(6166), 80-84, (2014) doi:10.1126/science.1246981
- [22] Duncan LM, Timms RT, Zavodszky E, Cano F, Dougan G, Randow F & Lehner PJ, “Fluorescence-based phenotypic selection allows forward genetic screens in haploid human cells”, *PloS One*, 7(6):e39651, (2012)
- [23] Joung J, Konermann S, Gootenberg J, Abudayyeh O, Platt R, Brigham M, Sanjana N & Zhang F, “Genome-scale CRISPR-Cas9 Knockout and Transcriptional Activation Screening”, *Nature Protocols*, 12, 828-863, (2017)
- [24] Zhang Z & Ren Q, “Why are essential genes essential? – The essentiality of *Saccharomyces* genes. ”, *Microbial Cell*, 2(8), 280-287, (2015) doi:10.15698/mic2015.08.218
- [25] Ning K, Ng HK, Srihari S, Leong HW & Nesvizhskii AI, “Examination of the relationship between essential genes in PPI network and hub proteins in reverse nearest neighbour topology”, *BMC Bioinformatics*, vol. 11 505, (2010)
- [26] Zhan T, Rindtorff N, Betge J, Ebert MP & Boutros M, *CRISPR/Cas9 for cancer research and therapy*, 2018 doi: 10.1016/j.semcancer.2018.04.001
- [27] Shalem O, Sanjana N, Hartenian E, Shi X, Scott D, Mikkelsen T, Heckl D, Ebert B, Root D, Doench J & Zhang F, “Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells”, *Science*, 343 (6166), 84-87, (2014)
- [28] Michaelis M, Rothweiler F, Nerretter T, Van Rikxoort M, Sharifi M, Wiese M, Ghafourian T & Cinatl J, “Differential effects of the oncogenic BRAF inhibitor PLX4032 (vemurafenib) and its progenitor PLX4720 on ABCB1 function”, *Journal of pharmacy & pharmaceutical sciences*, 17(1), 154-168, (2014) doi:10.1186/1471-2105-11-505
- [29] Huddleston JR, “Horizontal gene transfer in the human gastrointestinal tract: potential spread of antibiotic resistance genes”, *Infection and Drug Resistance*, 7, 167–176, (2014) doi:10.2147/IDR.S48820
- [30] Parnas O, Jovanovic M, Eisenhaure TM, Herbst RH, Dixit A, Ye CJ, Przybylski D, Platt RJ, Tirosh I, Sanjana NE, Shalem O, Satija R, Raychowdhury R, Mertins P, Carr SA, Zhang F, Hacohen N & Regev A., “A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks” *Cell*, 162(3):675-86, (2015) doi: 10.1016/j.cell.2015.06.059

- [31] [DeJesus R, Moretti F, McAllister G, Wang Z, Bergman P, Liu S, Frias E, Alford J, Reece-Hoyes JS, Lindeman A, Kelliher J, Russ C, Knehr J, Carbone W, Beibel M, Roma G, Ng A, Tallarico JA, Porter JA, Xavier RJ, Mickanin C, Murphy LO, Hoffman GR & Nyfeler B, “Functional CRISPR screening identifies the ufmylation pathway as a regulator of SQSTM1/p62”, *Elife*, 5, e17290, (2016) doi: 10.7554/eLife.17290
- [32] Pusapati GV, Kong JH, Patel BB, Krishnan A, Sagner A, Kinnebrew M, Briscoe J, Aravind L & Rohatgi R, “CRISPR Screens Uncover Genes that Regulate Target Cell Sensitivity to the Morphogen Sonic Hedgehog”, *Developmental Cell*, 44(1), 113-129, (2018), doi: 10.1016/j.devcel.2017.12.003.
- [33] Apalla Z, Nashan D, Weller RB & Castellsague X, “Skin Cancer: Epidemiology, Disease Burden, Pathophysiology, Diagnosis, and Therapeutic Approaches”, *Dermatology and Therapy*, 7(Suppl 1), 5-19, (2017)  
doi: 10.1007/s13555-016-0165-y
- [34] Leiter U, Eigentler T & Garbe C, “Epidemiology of skin cancer”, *Adv Exp Med Biol.*, 810:120-40, (2018)
- [35] Lorentzen HF, “Targeted therapy for malignant melanoma”, *Elsevier*, 46, 116-121, (2019)  
doi: 10.1016/j.coph.2019.05.010
- [36] Liu Y & Sheikh M, “Melanoma: Molecular Pathogenesis and Therapeutic Management”, *Mol Cell Pharmacol.*, 6(3), 228, (2014)
- [37] Shain A, Joseph N, Yu R, Benhamida J, Liu S, Prow T, Ruben B, North J, Pincus L, Yeh I, Judson R & Bastian B, “Genomic and transcriptomic analysis reveals incremental disruption of key signaling pathways during melanoma evolution”, 34(1), 45-55, (2018)
- [38] Zeng H, Jorapur A, Shain AH, Lang UE, Torres R, Zhang Y, McNeal AS, Botton T, Lin J, Donne M, Bastian IN, Yu R, North JP, Pincus L, Ruben BS, Josheph NM, Yeh I, Bastian BC & Judson RL, “Bi-allelic Loss of CDKN2A Initiates Melanoma Invasion via BRN2 Activation”, *Cancer cell*, 34 (1), 56-68, (2018)  
doi: 10.1016/j.ccell.2018.05.014
- [39] Rauscher B, Heigwer F, Breining M, Winter J & Boutros M, “GenomeCRISPR -a database for high-throughput CRISPR/Cas9 screens”, *Nucleic acids research*, 45(D1), D679-D686  
doi: 10.1093/nar/gkw997
- [40] Li W, Xu H, Xiao T, Cong L, Love M, Zhang F, Irizarry R, Liu J, Brown M & Liu X, “MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens”, *Genome Biology*, 15, 554, (2014)



[41] Doench J G, Fusi N, Sullender M, Hedge M, Vaimberg E W, Donovan K F, Smith I, Tothova Z, Wilen C, Orchard R, Virgin H W, Listgarten J & Root D E, “Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9”, *Nature Biotechnology*, 34(2), 184-191, (2016)

[42] Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, Mis M, Zimmermann M, Fradet-Turcotte A, Sun S, Mero P, Dirks P, Sidhu S, Roth FR, Rissland OS, Durocher D, Angers S & Moffat J, “High-Resolution CRISPR Screens reveal fitness genes and genotype-specific cancer liabilities”, *Cell*, 163(6), 1515-1526, (2015)  
doi: 10.1016/j.cell.2015.11.015

[43] Kanehisa M, Furumichi M, tanabe M, Sato Y & Morishima K, “KEGG: new perspectives on genomes, pathways, diseases and drugs”, *Nucleic Acids Research*, 45(D1), D353-D361, (2016)  
doi: 10.1093/nar/gkw1092

[44]Chen E Y, Tan C M, Kou Y, Duan Q, Wang Z, Meirelles G V & Ma’ayan A, “Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool.”, *BMC Bioinformatics*, 14(128) (2013)  
doi:10.1186/1471-2105-14-128

## ΠΑΡΑΡΤΗΜΑ

### ΛΙΣΤΕΣ ΜΕ ΤΑ ΓΟΝΙΔΙΑ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΗΣΑ

JUN	RASGRP3	DUSP5	FGF2	RASGRF2
MEF2C	NFKB2	DUSP6	PTPN5	FGFR4
ELK4	NFKB1	DUSP3	FGF21	MAP2K6
ELK1	MYC	TGFB3	IL1R2	MAPK8IP3
JUND	NFATC4	PLA2G4E	MAPK9	MAP3K6
GADD45B	MAPK14	DUSP4	CACNA2D3	CASP3
ZAK	FLNC	CACNG5	MAPK10	MAP3K12
STMN1	FLNA	CACNG4	MAPK11	FGFR3
RRAS2	KRAS	NF1	RASGRP2	FGFR1
MAP3K5	FLNB	PLA2G12A	PLA2G2A	RASA1
MAP3K1	PRKX	NFATC2	MAP2K2	FGF14
MAP3K3	TRAF6	RASGRP4	PLA2G4A	RPS6KA2
MAP3K4	TGFB2	MAP3K2	MAP2K3	RPS6KA3
MAP3K7	DUSP1	MAX	MECOM	FGF17
MAP3K8	DUSP2	DUSP10	PLA2G5	FGF16
AKT1	TGFB1	FGF9	MAPK13	FGF10
AKT2	TRAF2	FGF8	MAP2K1	GRB2
ARRB2	BDNF	FGF7	FGFR2	FGF11
CD14	TAB2	FGF6	MAP2K7	FGF12
ARRB1	ECSIT	FGF5	RASA2	FGF13
NRAS	TGFBR2	FGF3	MAPK8IP2	PLA2G1B
DUSP16	DUSP7	FGF4	RASGRF1	RPS6KA1
CHP2	TGFBR1	FGF1	MAP2K5	MAPKAPK3

**Λίστα 1:** Λίστα με τα γονίδια του MAPK signaling pathway

IKBKG	RPS6KA6	IL1A	NLK
HRAS	PPP3R2	PLA2G2F	PDGFRA
CACNG2	CACNG8	DAXX	PRKACA
FGF23	PPP5C	AKT3	PRKACB
CACNG3	PPP3CC	GADD45G	PAK1
MKNK2	CACNG6	FGF20	PRKACG
FGF18	PPP3R1	RELB	CRK
STK4	CACNG7	MAPKAPK5	CDC25B
STK3	MAP2K4	MAPK12	CRKL
MAPK8IP1	ATF2	RELA	MAP3K13
MOS	PDGFRB	GNA12	PLA2G2D
RAP1A	JMJD7-PLA2G4B	HSPA8	CDC42
MAPT	MAP4K3	HSPB1	RASGRP1
RAP1B	PLA2G6	PTPRR	CACNA2D1
PPP3CB	PLA2G2E	LAMTOR3	CACNB1
PPP3CA	PLA2G10	GADD45A	SRF
CACNA1H	MAP4K4	NGF	CACNB2
CACNA1G	RPS6KA5	DDIT3	SOS2
CACNA1I	BRAF	MAP3K14	CACNB3
ATF4	IKBKB	TAOK1	CACNB4
TAB1	PLA2G4B	PDGFA	CHUK
FOS	MAP3K11	FGF22	CACNG1
TAOK2	CACNA2D2	PDGFB	PRKCB

**Λίστα 1:** Λίστα με τα γονίδια του MAPK signaling pathway (συνέχεια)

RAF1	CACNA1F	RRAS
PRKCA	CACNA1S	TAOK3
TNF	MAP4K1	GNG12
PAK2	TNFRSF1A	NTF4
MKNK1	DUSP9	IL1B
PLA2G3	PPM1A	MRAS
PRKCG	PPM1B	NTF3
PTPN7	MAPK3	IL1R1
RAPGEF2	RPS6KA4	TP53
HSPA1L	CACNA2D4	FAS
CACNA1A	HSPA6	NR4A1
HSPA1B	MAPK7	PLA2G2C
RAC2	FGF19	FASLG
HSPA2	RAC1	NTRK2
CACNA1D	SOS1	NTRK1
CACNA1E	MAPK1	CHP
RAC3	DUSP14	
CACNA1B	MAP4K2	
HSPA1A	MAPK8	
PLA2G12B	EGFR	
CACNA1C	MAPKAPK2	
DUSP8	EGF	

**Λίστα 1:** Λίστα με τα γονίδια του MAPK signaling pathway (συνέχεια)

TFDP1	SMAD2	SP1
NOG	MAPK1	BMP7
TNF	SMURF2	BMP8B
GDF7	SMURF1	ZFYVE9
INHBB	EP300	BMP5
INHBC	BMP8A	BMP6
COMP	GDF5	ZFYVE16
INHBA	SKP1	THBS3
THBS4	CHRD	INHBE
RHOA	BMP2	THBS2
CREBBP	ACVR1C	DCN
ROCK1	E2F4	THBS1
ID1	SP1	
ID2	BMP7	
RPS6KB1	BMP8B	
RPS6KB2	ZFYVE9	
CUL1	BMP5	
LOC728622	BMP6	
ID4	ZFYVE16	
SMAD	THBS3	
MAPK3	INHBE	
RBL2	THBS2	
SMAD4	DCN	
RBL1	THBS1	
NODAL	BMP2	
SMAD1	ACVR1C	
MYC	E2F4	

**Λίστα 2:** Λίστα με τα γονίδια του TGF beta signaling pathway

TSC2	AKT3	FIGF
IGF1	RPS6KB2	DDIT4
RPS6KA6	TSC1	PIK3CA
MTOR	ULK2	PIK3CB
EIF4B	CAB39	PIK3CD
EIF4E	MAPK3	PIK3CG
EIF4EBP1	CAB39L	PIK3R1
RICTOR	RPS6	PIK3R2
INS	MAPK1	
PGF	RPS6KA1	
ULK1	PIK3R5	
RHEB	ULK3	
BRAF	EIF4E1B	
HIF1A	STK11	
PDPK1	EIF4E2	
RPTOR	VEGFA	
AKT1	PRKAA2	
AKT2	VEGFC	
RPS6KA2	PRKAA1	
RPS6KA3	VEGFB	
RPS6KB1	STRADA	
MLST8	PIK3R3	

**Λίστα 3:** Λίστα με τα γονίδια του MTOR signaling pathway

E2F1	RB1	KRAS	MITF
RAF1	AKT1	CDKN1A	PIK3CG
FGF9	AKT2	EGF	PIK3R1
HGF	FGFR1	FGF23	PIK3R2
FGF8	FGF14	FGF18	CDH1
FGF7	FGF17	ARAF	
IGF1	FGF16	MET	
FGF6	AKT3	CCND1	
FGF5	NRAS	PTEN	
PDGFD	FGF10	PIK3R3	
IGF1R	FGF11	PDGFA	
FGF3	FGF12	FGF22	
FGF4	FGF13	PDGFB	
FGF1	MAPK3	PDGFRA	
FGF2	FGF19	CDK4	
FGF21	MAPK1	TP53	
BAD	PIK3R5	E2F3	
PDGFRB	FGF20	E2F2	
MAP2K2	HRAS	PIK3CA	
BRAF	EGFR	PIK3CB	
MAP2K1	PDGFC	CDK6	
MDM2	CDKN2A	PIK3CD	

**Λίστα 4:** Λίστα με τα γονίδια του MELANOMA signaling pathway

ARIH2
CUL3
KIRREL
MED12
MED23
NF1
NF2
PCDC10
PTEN
SUPT20H
TADA1
TADA2B
TAF6L
TP53
UBE2F

**Λίστα 5:** Λίστα με τα γονίδια που είναι ήδη γνωστό (από προηγούμενες μελέτες) ότι προσφέρουν ανθεκτικότητα στο κύτταρο

## ΚΩΔΙΚΕΣ ΓΙΑ ΤΗΝ ΑΝΑΛΥΣΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

```
CRISPR_files<-read.csv("/Users/natasa/Desktop/papers_A375_cell_lines/GenomeCRISPR_full05112017.csv", header=T)
A375_cells<- CRISPR_files[CRISPR_files$cellline=='A375',]
A375_cells<- CRISPR_files[CRISPR_files$cas=='hspCas9',]
write.csv(A375_cells, file="/Users/natasa/Desktop/papers_A375_cell_lines/All_studies_from_GenomeCRISPR.csv")

#Study 27260156
Study_27260156<- A375_cells[A375_cells$pubmed=='27260156',]
write.csv(Study_27260156, file="/Users/natasa/Desktop/papers_A375_cell_lines/Study_27260156.csv")

#Study 26780180
Study_26780180<- A375_cells[A375_cells$pubmed=='26780180',]
write.csv(Study_26780180, file="/Users/natasa/Desktop/papers_A375_cell_lines/study_26780180.csv")

#Study 24336571
Study_24336571<- A375_cells[A375_cells$pubmed=='24336571',]
write.csv(Study_24336571, file="/Users/natasa/Desktop/papers_A375_cell_lines/Study_24336571.csv")
```

**Κώδικας 1:** Κώδικας για την εξαγωγή των δεδομένων από την βάση δεδομένων «genomeCRISPR»

```
#Merging all the gene files to make plots:

#Looking at all essential genes:
library(ggplot2)
library(reshape2)
library(enrichR)

Essential_genes<-read.csv('/Users/natasa/Desktop/CRIPSR_Screens/Essential_genes_TKO.csv', header=T)
Non_essential_genes<- read.csv('/Users/natasa/Desktop/CRIPSR_Screens/Non_Essential_genes_TKO.csv', header=T)

mypath<- '/Users/natasa/Desktop/papers_A375_cell_lines/All_negative_selection_gene_summary_files/all_studies_4/'
setwd(mypath)
file.names <- dir(mypath, pattern = ".txt")

#Essential/non essential genes
Identified_genes_all<- NULL
for(i in 1:length(file.names)){
  Sample_ID<- gsub(".gene_summary.txt","",file.names[i])
  Identified_genes<- read.table(file.names[i], header=TRUE)
  Identified_genes_Essential<-Identified_genes[Identified_genes$id %in% Essential_genes$Gene,]
  Identified_genes_Essential$Sample<- paste0('Essential_', Sample_ID)
  Identified_genes_non_Essential<- Identified_genes[Identified_genes$id %in% Non_essential_genes$Gene,]
  Identified_genes_non_Essential$Sample<- paste0('Non_essential_', Sample_ID)
  Identified_genes<- rbind(Identified_genes_Essential, Identified_genes_non_Essential)
  Identified_genes_all <- rbind(Identified_genes_all, Identified_genes)
}

write.csv(Identified_genes_all, file= paste0(mypath, 'All_essentiality.csv'))
```

**Κώδικας 2:** Κώδικας για την pathway enrichment analysis και την δημιουργία density plot



```

#Plot
pdf('density_plot.pdf', width=10, height=10)

p <- ggplot(Identified_genes_all, aes(neg.lfc, color= sample))
p1 <- p + geom_density(show.legend=FALSE)+
stat_density(aes(x=neg.lfc, colour=sample),geom="line",position="identity")+
xlab(expression("log"[2]*~"FC"))+
ylab("density")+
theme_bw()+
scale_color_manual(values=c("#a50f15", "#de2d26", "#fb6a4a", "#fc9272", "#a50f15", "#de2d26", "#fb6a4a",
"#fc9272", "#a50f15", "#de2d26", "#fb6a4a", "#fc9272", "#a50f15", "#de2d26",
"#fb6a4a", "#fc9272", "#252525", "#636363", "#969696", "#bdbdbd", "#252525",
"#636363", "#969696", "#bdbdbd", "#252525", "#636363", "#969696", "#bdbdbd"))+
theme(axis.text.y = element_text(angle = 0, vjust = 0.5,size=8),
axis.text.x = element_text(angle = 0, size=8, hjust=0.5,vjust=0.5),
axis.title.x = element_text(size=8),
axis.title.y = element_text(size=8),
plot.background = element_blank() ,
panel.grid.major = element_blank(),
panel.grid.minor = element_blank()) +
scale_x_continuous(breaks=seq(-8,2,2))+
theme(legend.position=c(.3,.8))+
theme(legend.text=element_text(size=8))+
theme(legend.title=element_blank())
print(p1)
dev.off()

```

**Κώδικας 2:** Κώδικας για την pathway enrichment analysis και την δημιουργία density plot

```

#####Pathway enrichment analysis#####
#Essential/non essential genes
Pathway_all<- NULL
for(i in 1:length(file.names)){
  Sample_ID<- gsub(".",gene_summary.txt","",file.names[i])
  Gene_summary_file<- read.table(file.names[i], header=TRUE)
  Gene_summary_file<- Gene_summary_file[Gene_summary_file$neg.fdr<=0.10,]
  Gene_names<- as.character(Gene_summary_file$id)
  Pathway<- enrichr(Gene_names, databases = 'KEGG_2016')
  Pathway<- Pathway$KEGG_2016
  Pathway<- dplyr::select(Pathway, Term, old.Adjusted.P.value)
  Pathway$Sample<- Sample_ID
  Pathway$Sample<- paste0('Pathway_', Sample_ID)
  Pathway$pvalue_group<- NA
  Pathway[Pathway$old.Adjusted.P.value<=0.001, 'pvalue_group']<- 'A'
  Pathway[Pathway$old.Adjusted.P.value>0.001 & Pathway$old.Adjusted.P.value<=0.01, 'pvalue_group']<- 'B'
  Pathway[Pathway$old.Adjusted.P.value>0.01 & Pathway$old.Adjusted.P.value<=0.05, 'pvalue_group']<- 'C'
  Pathway[Pathway$old.Adjusted.P.value>0.05, 'pvalue_group']<- 'Non_significant'
  Pathway_all <- rbind(Pathway_all, Pathway)
}

write.csv(Pathway_all, file= paste0(mypath, 'All_pathway.csv'))

Pathway_all$Term<-gsub("_.*", "", Pathway_all$Term)
Count_terms<- table(Pathway_all$Term, Pathway_all$pvalue_group)
Count_terms<- data.frame(Count_terms)
Count_terms<- dcast(Count_terms, Var1~Var2)
Count_terms$sum<- Count_terms$A+Count_terms$B+Count_terms$C
Count_terms<- Count_terms[order(Count_terms$sum, decreasing=T),]
Count_terms<-Count_terms[1:6,]
write.csv(Count_terms, file= paste0(mypath, 'Pathway_count.csv'))

Pathway_all<- Pathway_all[Pathway_all$Term %in% Count_terms$Var1,]

```

**Κώδικας 3:** Κώδικας για την pathway enrichment analysis και την δημιουργία heatmap

```
pdf('Heatmap_pathways.pdf', width=14, height=10)

p <- ggplot(Pathway_all, aes(Term,sample, fill= pvalue_group))
p1 <- p + geom_tile(color='black')+
  xlab('Pathways')+
  ylab("Sample")+
  theme_bw()+
  scale_fill_manual(values=c( "#08519c", "#3182bd", "#6baed6", "#ffffff"))+
  theme(axis.text.y = element_text(angle = 0, vjust = 0.5,size=15),
        axis.text.x = element_text(angle = 90, size=15, hjust=1,vjust=0.5),
        axis.title.x = element_text(size=15),
        axis.title.y = element_text(size=15),
        plot.background = element_blank() ,
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
  theme(legend.text=element_text(size=15))+
  theme(legend.title=element_blank())
print(p1)

dev.off()
```

### Κώδικας 3: Κώδικας για την pathway enrichment analysis και την δημιουργία heatmap

```
install.packages("ggplot2")
library(ggplot2)
install.packages("dplyr")
library(dplyr)
install.packages("ROCR")
library(ROCR)

essential_genes<-read.csv('/Users/natasa/Desktop/CRISPR_Screens/Essential_genes_TKO.csv', header = T)
non_essential_genes<-read.csv('/Users/natasa/Desktop/CRISPR_Screens/Non_Essential_genes_TKO_copy.csv', header = T)
libraries_ID<- read.csv('/Users/natasa/Desktop/Libraries_ID.csv', header = T)

BOTH_GENES<- rbind.data.frame(essential_genes,non_essential_genes)
#write.csv(both.genes, file= "/Users/natasa/Desktop/try.csv")

mypath<- '/Users/natasa/Desktop/All_the_studies/All_Negative_Selection_Gene_Summary_Files_21STUDIES/All_Studies_21/'
setwd(mypath)
file.names <- dir(mypath, pattern = ".txt")

DATA_NEW<- NULL
AUC_new<-NULL

for(i in 1:length(file.names)){
  sample_id<- gsub(".gene_summary.txt","",file.names[i])
  Identified_GENES<- read.table(file.names[i], header=TRUE)
  Identified_GENES_New<-Identified_GENES[Identified_GENES$id %in% BOTH_GENES$Gene,]

  Identified_GENES_New$essential_05<-as.numeric(-log(Identified_GENES_New$neg.fdr))
  Identified_GENES_New<-dplyr::select(Identified_GENES_New, id, essential_05)

  #function yes/no
  Identified_GENES_New$yes= ifelse(Identified_GENES_New$id %in% essential_genes$Gene,1,0)

  PRED<- prediction(Identified_GENES_New$essential_05, Identified_GENES_New$yes)
  auc<- performance(PRED, measure='auc')
  auc<-auc@y.values[[1]]
  PERF<- performance(PRED, "tpr", "fpr")
  Data<- data.frame(PERF@x.values, PERF@y.values)

  colnames(Data)<- c('Recall', 'Precision')
  Data$sample<- sample_id
  AUC<- data.frame(cbind(sample_id, auc))
  AUC_new<- rbind(AUC, AUC_new)
  DATA_NEW<- rbind(Data, DATA_NEW)
}
```

### Κώδικας 4: Κώδικας για την δημιουργία ROC\_curve

```

AUC_values<-AUC_new
AUC_values$auc<-as.numeric(as.character(AUC_values$auc))
AUC_values$auc<-round(AUC_values$auc, digits = 2)
AUC_values$sample_id<-as.factor(AUC_values$sample_id)
colnames(AUC_values)<- c('sample', 'auc')

COMBINE_ROC<-DATA_NEW
write.csv(DATA_NEW, file= "/Users/natasa/Desktop/COMBINE_ROC.csv")

COMBINE_ROC$Library<- libraries_ID$Library[match(COMBINE_ROC$sample,libraries_ID$sample)]

#ROC curve plot

pdf('ROC_CURVE.pdf', width=12, height=12)

p <- ggplot(COMBINE_ROC, aes(x=Recall, y=Precision))
p1 <- p + geom_line(aes(color =Library), size = 1.2) +
  xlab("False positive rate")+
  ylab("True positive rate")+
  theme_bw() +
  facet_wrap(~sample, nrow=7)+
  theme(axis.text.y = element_text(angle = 0, vjust = 0.5,size=16),
        axis.text.x = element_text(angle = 0, size=16, hjust=0.5,vjust=0.5),
        axis.title.x = element_text(size=18),
        axis.title.y = element_text(size=18),
        plot.background = element_blank() ,
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
  scale_x_continuous(breaks=seq(0,1,0.2))+
  scale_y_continuous(breaks=seq(0,1,0.2))+
  theme(legend.text=element_text(size=14))+
  geom_text(data = AUC_values, aes(x = 0.8, y = 0.6, label = auc))+
  theme(legend.position="bottom")+
  geom_abline(intercept = 0, slope = 1, color='gray',lty="longdash")+
  theme(legend.title=element_blank())
print(p1)

dev.off()

```

#### Κώδικας 4: Κώδικας για τη δημιουργία ROC\_curve (συνέχεια)

```

#Now we look at MAPK Pathway genes
#####
library(ggplot2)
library(reshape2)

MAPK_List<- read.csv('/Users/natasa/Desktop/CRIPSR_Screens/MAPK_Pathway.csv', header = T)

mypath<- '/Users/natasa/Desktop/All_the_studies/All_Negative_Selection_Gene_Summary_Files_21STUDIES/MAPK_All_Studies_21/'
setwd(mypath)
file.names <- dir(mypath, pattern = ".txt")

Identified_MAPKs_all<- NULL
for(i in 1:length(file.names)){
  Sample_ID<-gsub(".gene_summary.txt","",file.names[i])
  Identified_MAPKs<- read.table(file.names[i], header=TRUE)
  Identified_MAPKs<-Identified_MAPKs[Identified_MAPKs$id %in% MAPK_List$Gene,]
  Identified_MAPKs$Sample<- paste0('MAPK_', Sample_ID)
  Identified_MAPKs$FDR_group<- NA
  Identified_MAPKs[Identified_MAPKs$neg.fdr<=0.05, 'FDR_group']<- 'A'
  Identified_MAPKs[Identified_MAPKs$neg.fdr>0.05 & Identified_MAPKs$neg.fdr<=0.1, 'FDR_group']<- 'B'
  Identified_MAPKs[Identified_MAPKs$neg.fdr>0.1 & Identified_MAPKs$neg.fdr<=0.30, 'FDR_group']<- 'C'
  Identified_MAPKs[Identified_MAPKs$neg.fdr>0.30, 'FDR_group']<- 'D'
  Identified_MAPKs<- rbind(Identified_MAPKs)
  Identified_MAPKs_all <- rbind(Identified_MAPKs_all, Identified_MAPKs)
}

#write the MAPK
write.csv(Identified_MAPKs_all, file = paste0(mypath, 'All_MAPKs_full.csv'))

#Extract relevant columns:
MAPK_hit_list<- dplyr::select(Identified_MAPKs_all, id, Sample, FDR_group)
MAPK_hit_list_new<- dcast(MAPK_hit_list, id~ Sample)
keep <- apply(MAPK_hit_list_new[2:16], 1, function(x) length(unique(x[!is.na(x)])) != 1)
MAPK_hit_list_new<- MAPK_hit_list_new[keep, ]
MAPK_hit_list<- melt(MAPK_hit_list_new, id.vars=c('id'))
MAPK_hit_list<- MAPK_hit_list[complete.cases(MAPK_hit_list), ]

```

#### Κώδικας 5: Γονίδια του MAPK μονοπατιού τα οποία είναι απαραίτητα στην A375 κυτταρική σειρά.

```
pdf('heatmap.pdf', width=14, height=9)

p <- ggplot(MAPK_hit_list, aes(id,variable, fill= value))
p1 <- p + geom_tile(color='black')+
  xlab('Gene')+
  ylab("Sample")+
  theme_bw()+
  scale_fill_manual(labels= c("FDR<5%", "5%<FDR<10%", "10%<FDR<30%", "Not significant"),
                    values=c( "#08519c", "#3182bd", "#6baed6", "#ffffff"))+
  theme(axis.text.y = element_text(angle = 0, vjust = 0.5,size=13),
        axis.text.x = element_text(angle = 90, size=14, hjust=1,vjust=0.5),
        axis.title.x = element_text(size=15),
        axis.title.y = element_text(size=15),
        plot.background = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
  theme(legend.text=element_text(size=15))+
  theme(legend.title=element_blank()+
  theme(legend.position = c(-0.25, -0.09))
print(p1)

dev.off()
```

**Κώδικας 5:** Γονίδια του MAPK μονοπατιού τα οποία είναι απαραίτητα στην A375 κυτταρική σειρά.

```
#Now we look at other signaling pathways
#####
install.packages("ggplot2")
library(ggplot2)
KEGG_MTOR<- read.csv('/Users/natasa/Desktop/CRISPR_Screens/KEGG_MTOR_SIGNALING_PATHWAY.csv', header = T)

mypath<- '/Users/natasa/Desktop/All_the_studies/All_Negative_Selection_Gene_Summary_Files_21STUDIES/MAPK_All_Studies_21'
setwd(mypath)
file.names <- dir(mypath, pattern = ".txt")

Identified_KEGG_MTOR_all<- NULL
for(i in 1:length(file.names)){
  Sample_ID<-gsub(".gene_summary.txt","",file.names[i])
  Identified_KEGGS_MTOR<- read.table(file.names[i], header=TRUE)
  Identified_KEGGS_MTOR<-Identified_KEGGS_MTOR[Identified_KEGGS_MTOR$id %in% KEGG_MTOR$KEGG_MTOR_SIGNALING_PATHWAY,]
  Identified_KEGGS_MTOR$Sample<- paste0('KEGG_', Sample_ID)
  Identified_KEGGS_MTOR$FDR_group<- NA
  Identified_KEGGS_MTOR[Identified_KEGGS_MTOR$neg.fdr<=0.05, 'FDR_group']<- 'A'
  Identified_KEGGS_MTOR[Identified_KEGGS_MTOR$neg.fdr>0.05 & Identified_KEGGS_MTOR$neg.fdr<=0.1, 'FDR_group']<- 'B'
  Identified_KEGGS_MTOR[Identified_KEGGS_MTOR$neg.fdr>0.1 & Identified_KEGGS_MTOR$neg.fdr<=0.30, 'FDR_group']<- 'C'
  Identified_KEGGS_MTOR[Identified_KEGGS_MTOR$neg.fdr>0.30, 'FDR_group']<- 'Non_significant'
  Identified_KEGGS_MTOR<- rbind(Identified_KEGGS_MTOR)
  Identified_KEGG_MTOR_all <- rbind(Identified_KEGG_MTOR_all, Identified_KEGGS_MTOR)
}

#write the MAPK
write.csv(Identified_KEGG_MTOR_all, file = paste0(mypath, 'All_KEGGS_MTOR.csv'))
```

**Κώδικας 6:** Κώδικας για τη δημιουργία heatmap του KEGG\_MTOR pathway

```

#Extract relevant columns:
KEGG_MTOR_hit_list<- dplyr::select(Identified_KEGG_MTOR_all, id, Sample, FDR_group)
KEGG_MTOR_hit_list_new<- dcast(KEGG_MTOR_hit_list, id~ Sample)

KEGG_MTOR_hit_list_new$frequency<- rowSums(KEGG_MTOR_hit_list_new=="Non_significant")
KEGG_MTOR_hit_list_new <- subset(KEGG_MTOR_hit_list_new, frequency<21)
KEGG_MTOR_hit_list_new<- melt(KEGG_MTOR_hit_list_new, id.vars= c('id'))
KEGG_MTOR_hit_list_new <- subset(KEGG_MTOR_hit_list_new, value>9)

pdf('KEGG_MTOR.pdf', width=14, height=9)

p <- ggplot(KEGG_MTOR_hit_list_new, aes(id,variable, fill= value))
p1 <- p + geom_tile(color='black')+
  xlab('Gene')+
  ylab("Sample")+
  theme_bw()+
  scale_fill_manual(labels= c("FDR<5%", "5%<FDR<10%", "10%<FDR<30%", "Not significant"),
    values=c( "#08519c", "#3182bd", "#6baed6", "#ffffff"))+
  theme(axis.text.y = element_text(angle = 0, vjust = 0.5,size=13),
    axis.text.x = element_text(angle = 90, size=14, hjust=1,vjust=0.5),
    axis.title.x = element_text(size=15),
    axis.title.y = element_text(size=15),
    plot.background = element_blank() ,
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()) +
  theme(legend.text=element_text(size=15))+
  theme(legend.title=element_blank()+
  theme(legend.position = c(-0.25, -0.06))
print(p1)

dev.off()

```

## Κώδικας 6: Κώδικας για τη δημιουργία heatmap του KEGG\_MTOR pathway

```

#Now we look at other signaling pathways
#####
install.packages("reshape2")
library(reshape2)
install.packages("ggplot2")
library(ggplot2)
KEGG_TGF_BETA<- read.csv('/Users/natasa/Desktop/CRIPSR_Screens/KEGG_TGF_BETA_SIGNALING_PATHWAY.csv', header = T)

mypath<- '/Users/natasa/Desktop/All_the_studies/All_Negative_Selection_Gene_Summary_Files_21STUDIES/MAPK_All_Studies_21'
setwd(mypath)
file.names <- dir(mypath, pattern = ".txt")

Identified_KEGG_TGF_BETA_all<- NULL
for(i in 1:length(file.names)){
  Sample_ID<-gsub(".gene_summary.txt","",file.names[i])
  Identified_KEGGS_TGF<- read.table(file.names[i], header=TRUE)
  Identified_KEGGS_TGF<-Identified_KEGGS_TGF[Identified_KEGGS_TGF$id %in% KEGG_TGF_BETA$KEGG_TGF_BETA_SIGNALING_PATHWAY,]
  Identified_KEGGS_TGF$Sample<- paste0('KEGG_', Sample_ID)
  Identified_KEGGS_TGF$FDR_group<- NA
  Identified_KEGGS_TGF[Identified_KEGGS_TGF$neg.fdr<=0.05, 'FDR_group']<- 'A'
  Identified_KEGGS_TGF[Identified_KEGGS_TGF$neg.fdr>0.05 & Identified_KEGGS_TGF$neg.fdr<=0.1, 'FDR_group']<- 'B'
  Identified_KEGGS_TGF[Identified_KEGGS_TGF$neg.fdr>0.1 & Identified_KEGGS_TGF$neg.fdr<=0.30, 'FDR_group']<- 'C'
  Identified_KEGGS_TGF[Identified_KEGGS_TGF$neg.fdr>0.30, 'FDR_group']<- 'Non_significant'
  Identified_KEGGS_TGF<- rbind(Identified_KEGGS_TGF)
  Identified_KEGG_TGF_BETA_all <- rbind(Identified_KEGG_TGF_BETA_all, Identified_KEGGS_TGF)
}

#write the MAPK
write.csv(Identified_KEGG_TGF_BETA_all, file = paste0(mypath, 'All_KEGGS_TGF.csv'))

```

## Κώδικας 7: Κώδικας για την δημιουργία heatmap του KEGG\_TGF\_BETA pathway

```

#Extract relevant columns:
KEGG_hit_list<- dplyr::select(Identified_KEGG_TGF_BETA_all, id, Sample, FDR_group)
KEGG_hit_list_new<- dcast(KEGG_hit_list, id~ Sample)

KEGG_hit_list_new$frequency<- rowSums(KEGG_hit_list_new=="Non_significant")
KEGG_hit_list_new <- subset(KEGG_hit_list_new, frequency<21)
KEGG_hit_list_new<- melt(KEGG_hit_list_new, id.vars= c('id'))
KEGG_hit_list_new <- subset(KEGG_hit_list_new, value>9)

pdf('KEGG_TGF_BETA.pdf', width=17, height=8)

p <- ggplot(KEGG_hit_list_new, aes(id,variable, fill= value))
p1 <- p + geom_tile(color='black')+
  xlab('Gene')+
  ylab("Sample")+
  theme_bw()+
  scale_fill_manual(labels= c("FDR<5%", "5%<FDR<10%", "10%<FDR<30%", "Not significant"),
                    values=c( "#08519c", "#3182bd", "#6baed6", "#ffffff"))+

  theme(axis.text.y = element_text(angle = 0, vjust = 0.5,size=13),
        axis.text.x = element_text(angle = 90, size=14, hjust=1,vjust=0.5),
        axis.title.x = element_text(size=13),
        axis.title.y = element_text(size=13),
        plot.background = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
  theme(legend.text=element_text(size=15))+
  theme(legend.title=element_blank()+
  theme(legend.position = c(-0.2, -0.07))
print(p1)

dev.off()

```

### Κώδικας 7: Κώδικας για την δημιουργία heatmap του KEGG\_TGF\_BETA pathway

```

#Now we look at other signaling pathways
#####
install.packages("ggplot2")
library(ggplot2)
install.packages("reshape2")
library(reshape2)
KEGG_MELANOMA<- read.csv('/Users/natasa/Desktop/CRIPSR_Screens/KEGG_MELANOMA.csv', header = T)

mypath<- '/Users/natasa/Desktop/All_the_studies/All_Negative_Selection_Gene_Summary_Files_21STUDIES/MAPK_All_Studies_21'
setwd(mypath)
file.names <- dir(mypath, pattern = ".txt")

Identified_KEGG_MELANOMA_all<- NULL
for(i in 1:length(file.names)){
  Sample_ID<-gsub(".gene_summary.txt","",file.names[i])
  Identified_KEGGS_MELANOMA<- read.table(file.names[i], header=TRUE)
  Identified_KEGGS_MELANOMA<-Identified_KEGGS_MELANOMA[Identified_KEGGS_MELANOMA$id %in% KEGG_MELANOMA$KEGG_MELANOMA,]
  Identified_KEGGS_MELANOMA$Sample<- paste0('KEGG_', Sample_ID)
  Identified_KEGGS_MELANOMA$FDR_group<- NA
  Identified_KEGGS_MELANOMA[Identified_KEGGS_MELANOMA$neg.fdr<=0.05, 'FDR_group']<- 'A'
  Identified_KEGGS_MELANOMA[Identified_KEGGS_MELANOMA$neg.fdr>0.05 & Identified_KEGGS_MELANOMA$neg.fdr<=0.1, 'FDR_group']<- 'B'
  Identified_KEGGS_MELANOMA[Identified_KEGGS_MELANOMA$neg.fdr>0.1 & Identified_KEGGS_MELANOMA$neg.fdr<=0.30, 'FDR_group']<- 'C'
  Identified_KEGGS_MELANOMA[Identified_KEGGS_MELANOMA$neg.fdr>0.30, 'FDR_group']<- 'Non_significant'
  Identified_KEGGS_MELANOMA<- rbind(Identified_KEGGS_MELANOMA)
  Identified_KEGG_MELANOMA_all <- rbind(Identified_KEGG_MELANOMA_all, Identified_KEGGS_MELANOMA)
}

#write the MAPK
write.csv(Identified_KEGG_MELANOMA_all, file = paste0(mypath, 'All_KEGGS_MELANOMA.csv'))

```

### Κώδικας 8: Κώδικας για τη δημιουργία heatmap του KEGG\_MELANOMA pathway



```

#Extract relevant columns:
KEGG_MELANOMA_hit_list<- dplyr::select(Identified_KEGG_MELANOMA_all, id, Sample, FDR_group)
KEGG_MELANOMA_hit_list_new<- dcast(KEGG_MELANOMA_hit_list, id~ Sample)

KEGG_MELANOMA_hit_list_new$frequency<- rowSums(KEGG_MELANOMA_hit_list_new=="Non_significant")
KEGG_MELANOMA_hit_list_new <- subset(KEGG_MELANOMA_hit_list_new, frequency<21)
KEGG_MELANOMA_hit_list_new<- melt(KEGG_MELANOMA_hit_list_new, id.vars= c('id'))
KEGG_MELANOMA_hit_list_new <- subset(KEGG_MELANOMA_hit_list_new, value>9)

pdf('KEGG_MELANOMA.pdf', width=15, height=9)

p <- ggplot(KEGG_MELANOMA_hit_list_new, aes(id,variable, fill= value))
p1 <- p + geom_tile(color='black')+
  xlab('Gene')+
  ylab("Sample")+
  theme_bw()+
  scale_fill_manual(labels= c("FDR<5%", "5%<FDR<10%", "10%<FDR<30%", "Not significant"),
                    values=c( "#08519c", "#3182bd", "#6baed6", "#ffffff"))+
  theme(axis.text.y = element_text(angle = 0, vjust = 0.5,size=13),
        axis.text.x = element_text(angle = 90, size=14, hjust=1,vjust=0.5),
        axis.title.x = element_text(size=15),
        axis.title.y = element_text(size=15),
        plot.background = element_blank() ,
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
  theme(legend.text=element_text(size=15))+
  theme(legend.title=element_blank()+
  theme(legend.position = c(-0.25, -0.07))
print(p1)

dev.off()

```

## Κώδικας 8: Κώδικας για τη δημιουργία heatmap του KEGG\_MELANOMA pathway

```

#Now we look at MAPK Pathway genes
#####
install.packages("reshape2")
library(reshape2)
install.packages("ggplot2")
library(ggplot2)

MAPK_full_list<- read.csv('/Users/natasa/Desktop/CRIPSR_Screens/MAPK_Pathway.csv', header = T)

mypath<- '/Users/natasa/Desktop/All_the_studies/All_Positive_Selection_Gene_Summary_Files_22STUDIES/MAPK_All_Studies_22/'
setwd(mypath)
file.names <- dir(mypath, pattern = ".txt")

Identified_MAPKs_all<- NULL
for(i in 1:length(file.names)){
  Sample_ID<-gsub(".gene_summary.txt","",file.names[i])
  Identified_MAPKs<- read.table(file.names[i], header=TRUE)
  Identified_MAPKs<-Identified_MAPKs[Identified_MAPKs$id %in% MAPK_full_list$Gene,]
  Identified_MAPKs$Sample<- paste0('MAPK_', Sample_ID)
  Identified_MAPKs$FDR_group<- NA
  Identified_MAPKs[Identified_MAPKs$pos.fdr<=0.05, 'FDR_group']<- 'A'
  Identified_MAPKs[Identified_MAPKs$pos.fdr>0.05 & Identified_MAPKs$pos.fdr<=0.1, 'FDR_group']<- 'B'
  Identified_MAPKs[Identified_MAPKs$pos.fdr>0.1 & Identified_MAPKs$pos.fdr<=0.30, 'FDR_group']<- 'C'
  Identified_MAPKs[Identified_MAPKs$pos.fdr>0.30, 'FDR_group']<- 'D'
  Identified_MAPKs<- rbind(Identified_MAPKs)
  Identified_MAPKs_all <- rbind(Identified_MAPKs_all, Identified_MAPKs)
}

#Write the MAPK
write.csv(Identified_MAPKs_all, file = paste0('/Users/natasa/Desktop/', 'All_MAPKs_full_correct.csv'))

```

## Κώδικας 9: Positive\_selection ( MAPK pathway genes)

```

#Extract relevant columns:
MAPK_hit_list<- dplyr::select(Identified_MAPKs_all, id, Sample, FDR_group)
MAPK_hit_list_new<- dcast(MAPK_hit_list, id~ sample)
keep <- apply(MAPK_hit_list_new[2:16], 1, function(x) length(unique(x[!is.na(x)])) != 1)
MAPK_hit_list_new<- MAPK_hit_list_new[keep, ]
MAPK_hit_list<- melt(MAPK_hit_list_new, id.vars=c('id'))
MAPK_hit_list<- MAPK_hit_list[complete.cases(MAPK_hit_list), ]

pdf('heatmap.pdf', width=14, height=9)

p <- ggplot(MAPK_hit_list, aes(id,variable, fill= value))
p1 <- p + geom_tile(color='black')+
  xlab('Gene')+
  ylab("Sample")+
  theme_bw()+
  scale_fill_manual(labels= c("FDR<5%", "5%<FDR<10%", "10%<FDR<30%", "Not significant"),
    values=c( "#08519c", "#3182bd", "#6baed6", "#ffffff"))+
  theme(axis.text.y = element_text(angle = 0, vjust = 0.5,size=13),
    axis.text.x = element_text(angle = 90, size=14, hjust=1,vjust=0.5),
    axis.title.x = element_text(size=15),
    axis.title.y = element_text(size=15),
    plot.background = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()) +
  theme(legend.text=element_text(size=15))+
  theme(legend.title=element_blank()+
  theme(legend.position = c(-0.25, -0.09))
print(p1)
dev.off()

```

### Κώδικας 9: Positive\_selection ( MAPK pathway genes)

```

#Positive selection
install.packages("dplyr")
library(dplyr)
install.packages("ggplot2")
library(ggplot2)

mypath<- '/Users/natasa/Desktop/All_the_studies/All_Positive_Selection_Gene_Summary_Files_22STUDIES/All_Studies_22/'
setwd(mypath)
file.names <- dir(mypath, pattern = ".txt")

Identified_categories<-NULL
for(i in 1:length(file.names)){
  sample<- gsub(".gene_summary.txt","",file.names[i])
  identified_genes<- read.table(file.names[i], header=TRUE)
  identified_genes$study<- sample
  identified_genes$FDR_category<- NA
  identified_genes[identified_genes$pos.fdr<=0.001, 'FDR_category']<- 'A'
  identified_genes[identified_genes$pos.fdr>0.001 & identified_genes$pos.fdr<=0.01, 'FDR_category']<- 'B'
  identified_genes[identified_genes$pos.fdr>0.01 & identified_genes$pos.fdr<=0.05, 'FDR_category']<- 'c'
  identified_genes[identified_genes$pos.fdr>0.05, 'FDR_category']<- 'D'
  Identified_categories<-rbind(Identified_categories, identified_genes)
}

write.csv(Identified_categories, file = paste0(mypath, 'All_the_categories.csv'))

Identified_categories<- dplyr::select(Identified_categories, id, Study, FDR_category)
Identified_categories$FDR_category<- as.factor(Identified_categories$FDR_category)
Identified_categories$Study<- as.factor(Identified_categories$Study)

#Rearrange data
Identified_categories_new<- dcast(Identified_categories, id ~ Study)
keep <- apply(Identified_categories_new[2:23], 1, function(x) length(unique(x[!is.na(x)])) != 1)
Identified_categories_new<- Identified_categories_new[keep, ]

```

### Κώδικας 10: Positive\_selection1-heatmap με τα γνωστά και μη γνωστά γονίδια



```

#Remove rows with more than 20 "D"
#Find the frequency of "D" for each row, if it's more than 20 then delete the row!
Identified_categories_new$frequency<- rowSums(Identified_categories_new=="D")

Identified_categories_new <- subset(Identified_categories_new, frequency<21)
write.csv(new, file = paste0(mypath, 'new.csv'))

write.csv(Identified_categories_new, file = paste0(mypath, 'All_the_categories_all_D_removed.csv'))

Identified_categories_new<- melt(Identified_categories_new, id.vars= c('id'))
new<- melt(new, id.vars= c('id'))

new <- subset(Identified_categories_new, value>9)

pdf('heatmap.pdf', width=50, height=9)

p <- ggplot(new, aes(id,variable, fill= value))
p1 <- p + geom_tile(color='black')+
  xlab('Gene')+
  ylab("Sample")+
  theme_bw()+
  scale_fill_manual(labels= c("FDR<5%", "5%<FDR<10%", "10%<FDR<30%", "Not significant"),
    values=c( "#08519c", "#3182bd", "#6baed6", + "#ffffff"))+

  theme(axis.text.y = element_text(angle = 0, vjust = 0.5,size=13),
    axis.text.x = element_text(angle = 90, size=14, hjust=1,vjust=0.5),
    axis.title.x = element_text(size=15),
    axis.title.y = element_text(size=15),
    plot.background = element_blank() ,
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()) +
  theme(legend.text=element_text(size=15))+
  theme(legend.title=element_blank()+
  theme(legend.position = c(-0.25, -0.09))
print(p1)

dev.off()

```

**Κώδικας 10:** Positive selection1-heatmap με τα γνωστά και μη γνωστά γονίδια

```

newest=read.csv('/Users/natasa/Desktop/All_the_studies/All_Positive_Selection_Gene_Summary_Files_22STUDIES/
All_Studies_22/new_without_the_known.csv', header=T)

newest<-melt(newest, id.vars = c('id'))
newest <- subset(newest, value>9255)
newest <- subset(newest, value>943)

#####NEW HEATMAP##### WITHOUT THE KNOWN GENES
pdf('newest.pdf', width=50, height=9)

p <- ggplot(newest, aes(id,variable, fill= value))
p2 <- p + geom_tile(color='black')+
  xlab('Gene')+
  ylab("Sample")+
  theme_bw()+
  scale_fill_manual(labels= c("FDR<5%", "5%<FDR<10%", "10%<FDR<30%", "Not significant"),
    values=c( "#08519c", "#3182bd", "#6baed6", "#ffffff"))+

  theme(axis.text.y = element_text(angle = 0, vjust = 0.5,size=13),
    axis.text.x = element_text(angle = 90, size=14, hjust=1,vjust=0.5),
    axis.title.x = element_text(size=15),
    axis.title.y = element_text(size=15),
    plot.background = element_blank() ,
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()) +
  theme(legend.text=element_text(size=15))+
  theme(legend.title=element_blank()+
  theme(legend.position = c(-0.25, -0.09))
print(p2)

dev.off()

```

**Κώδικας 10:** Positive selection1-heatmap χωρίς τα γνωστά γονίδια

```

52
53 pdf('heatmap.pdf', width=50, height=9)
54
55 p <- ggplot(new, aes(id,variable, fill= value))
56 p1 <- p + geom_tile(color='black')+
57   xlab('Gene')+
58   ylab("Sample")+
59   theme_bw()+
60   scale_fill_manual(labels= c("FDR<5%", "5%<FDR<10%", "10%<FDR<30%", "Not significant"), values=c( "#08519c", "#3182bd", "#6baed6",
61     + "#ffffff"))+
62   theme(axis.text.y = element_text(angle = 0, vjust = 0.5,size=13),
63     axis.text.x = element_text(angle = 90, size=14, hjust=1,vjust=0.5),
64     axis.title.x = element_text(size=15),
65     axis.title.y = element_text(size=15),
66     plot.background = element_blank() ,
67     panel.grid.major = element_blank(),
68     panel.grid.minor = element_blank()) +
69   theme(legend.text=element_text(size=15))+
70   theme(legend.title=element_blank()+
71     theme(legend.position = c(-0.25, -0.09))
72   print(p1)
73
74 dev.off()
75
76 newest=read.csv('/Users/natasa/Desktop/All_the_studies/All_Positive_Selection_Gene_Summary_Files_22STUDIES/All_Studies_22/
77   new_without_the_known.csv', header=T)
78 newest<-melt(newest, id.vars = c('id'))
79 newest <- subset(newest, value>9255)
80 newest <- subset(newest, value>943)
81
82 #####NEW HEATMAP##### WITHOUT THE KNOWN GENES
83 pdf('newest.pdf', width=50, height=9)
84
85 p <- ggplot(newest, aes(id,variable, fill= value))
86 p2 <- p + geom_tile(color='black')+
87   xlab('Gene')+
88   ylab("Sample")+
89   theme_bw()+
90   scale_fill_manual(labels= c("FDR<5%", "5%<FDR<10%", "10%<FDR<30%", "Not significant"), values=c( "#08519c", "#3182bd", "#6baed6",
91     + "#ffffff"))+
92   theme(axis.text.y = element_text(angle = 0, vjust = 0.5,size=13),
93     axis.text.x = element_text(angle = 90, size=14, hjust=1,vjust=0.5),
94     axis.title.x = element_text(size=15),
95     axis.title.y = element_text(size=15),
96     plot.background = element_blank() ,
97     panel.grid.major = element_blank(),
98     panel.grid.minor = element_blank()) +
99   theme(legend.text=element_text(size=15))+
100   theme(legend.title=element_blank()+
101     theme(legend.position = c(-0.25, -0.09))
102   print(p2)
103
104 dev.off()
105
106

```

**Κώδικας 10:** Positive selection2-heatmap χωρίς τα γνωστά γονίδια (συνέχεια)

