ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Εύρεση της σχέσης μετάλλαξης ::miRNA ::mRNA με καρκίνο του πνεύμονα μέσω ανάλυσης άρθρων και χρήσης γνωστών εργαλείων και πρόβλεψης στόχων

Ευσταθία Ιωάννα Βαβάλου

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Υπεύθυνος

Χατζηγεωργίου Άρτεμις-Γεωργία

Λαμία 2023

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

Εύρεση της σχέσης μετάλλαξης ::miRNA ::mRNA με καρκίνο του πνεύμονα μέσω ανάλυσης άρθρων και χρήσης γνωστών εργαλείων και πρόβλεψης στόχων

Ευσταθία Ιωάννα Βαβάλου

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επιβλέπουσα

Χατζηγεωργίου Άρτεμις-Γεωργία

Καθηγήτρια

Λαμία 2023

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

1. *Δενπαραθέτω κομμάτια βιβλίων ή άρθρωνή εργασιώνάλλωναυτολεξεί* ***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 επεξεργάσθηκαν, 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: 15/02/202 3

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.

Εύρεση της σχέσης μετάλλαξης ::miRNA ::mRNA με καρκίνο του πνεύμονα μέσω ανάλυσης άρθρων και χρήσης γνωστών εργαλείων και πρόβλεψης στόχων

Τριμελής Επιτροπή:

Χατζηγεωργίου Άρτεμις-Γεωργία, Καθηγήτρια (επιβλέπουσα)

Παντελεήμων Μπάγκος, Καθηγητής

Μπράλιου Γεωργία, Επίκουρος Καθηγήτρια

Ευχαριστίες

Θα ήθελα να ευχαριστήσω την επιβλέπουσα καθηγήτρια κυρία Χατζηγεωργίου Γ.Άρτεμις που μου έδωσε την ευκαιρία να ασχοληθώ με ένα τόσο ενδιαφέρον αντικείμενο ανακλύπτοντας νέες διεξόδους επιστημονικού ενδιαφέροντος. Επιπλέον θα ήθελα να ευχαριστήσω ιδιατέρως την υποψήφια διδάκτορα κυρία Ζαχαροπούλου Ελίζα για την πολύτιμη καθοδήγηση της καθόλη τη διάρκεια εκπονησης της εργασίας μου. Τέλος θα ήθελα να εχαριστήσω την οικογένεια μου και τους δικούς μου ανθρώπους που στάθηκαν δίπλα μου όλον αυτόν τον καιρό προσφέροντας μου τη στήριξη και συμπαράσταση τους θυμίζοντας μου συνεχώς πόσο σημαντικό είναι να επιμένεις και να προσπαθείς για να πετύχεις τα όνειρα σου.

Περίληψη

Τα microRNAs είναι μικρά μόρια RNA μεγέθους μεταξύ 19-25 νουκλεοτιδίων που στοχεύουν τα mRNAs με αποτέλεσμα να ρυθμίζουν την έκφραση των γονιδίων. Μελέτες των τελευταίων δεκατιών αποδεικνύουν την άρρηκτη σχέση μεταξύ των μεταλλάξεων και την υπερέκφραση διαφόρων microRNAs με ασθένειες, όπως και στον καρκίνο του πνεύμονα. Δημιουργούν λοιπόν ένα πεδίο έρευνας για την ικανότητά τους ως ρυθμιστές στην εξέλιξη και στην μετάστασή του.Προκειμένου να ερευνήσουμε από την βιβλιογραφία την ύπαρξη συσχέτισης των microRNAs με ογκογονίδια και τον καρκίνο του πνεύμονα, δημιουργήθηκε με τη βοήθεια της Entrez ένα σύστημα αυτοματοποιημένης αναζήτησης. Με την βοήθεια ενός ερωτήματος (query), την επεξεργασία κειμένου και τη χρήση μοντέλων ανάλυσης ονομάτων-οντοτήτων του scispaCy, επιστράφηκε ένα σύνολο από δημοσιεύσεις που περιέχει την επιθυμητή πληροφορία. Για την κατάληξη σε άρθρα που περιείχαν όσο το δυνατόν πιο σχετική πληροφορία με το αντικείμενο μελέτης, δημιουργήθηκαν κάποια φίλτρα τα οποία έχουν ως στόχο να απομονώσουν τα άρθρα ενδιαφέροντος. Μετά την εφαρμογή των δύο φίλτρων τα τελικά άρθρα που προέκυψαν πέρασαν από μια διαδικασία χειροκίνητου σχολιασμού για να ενσωματωθούν σε μία βάση εύκολης αναζήτησης. Τα μεταδεδομένα που προέκυψαν από αυτά τα άρθρα όπως ο τίτλος του άρθρου, το ID του και πληροφορίες σχετικά με τη συσχέτιση όπως το microRNA, το γονίδιο, η μετάλλαξη και η ασθένεια διατηρήθηκαν λοιπόν σε έναν κατάλογο αποτελεσμάτων.Η αναζήτηση στην Entrez μας επέστρεψε 221 άρθρα από τα οποία τα 146 είχαν διαθέσιμο κείμενο στην Pubmed Central και τα 106 από αυτά διέθεταν πλήρες κείμενο ώστε να είναι δυνατό να υποβληθούν σε διαδικασία ανάλυσης και επεξεργασίας κειμένου. Τα 61 από αυτά τα άρθρα πέρασαν τη διαδικασία του φιλτραρίσματος καταλήγοντας στον κατάλογο των αποτελεσμάτων τα 23 από αυτά. Στόχος της παρούσας εργασίας είναι να γίνει περισσότερο αποδοτική η αυτοματοποιημένη αναζήτηση και η διαδικασία εξαγωγής πληροφορίας καθώς η βιβλιογραφία διαθέτει τεράστιο όγκο δεδομένων τη δεδομένη χρονική στιγμή. Τα αποτελέσματα της παρούσας μελέτης που πραγματοποιήθηκε μπορούν να αποτελέσουν χρήσιμο εργαλείο για την επιστημονική κοινότητα.

Abstract

MicroRNAs are small RNA molecules containing 19 to 25 nucleotides which their target are mRNAs. As a result they regulate gene expression. Studies of the last decades demonstrate the closely related relationship between mutations and overexpression of various microRNAs with diseases, such as lung cancer. Therefore they create a field of research for their capacity as regulators of evolution and its transition.In order to search from the literature, the existence of an association of microRNAs with oncogenes and lung cancer, an automated search system was created through Entrez. A set of publications containing the desired information was enlisted by a query, text processing and the use of scispaCy name-entity analysis models. In order to find articles which contained as much relevant information as possible to the subject of study,some filters were created. These isolate the articles of interest.The final articles, that came through the two filters, went through a manual curation process in order to be incorporated into an easily searchable database. Metadata derived from these articles such as article title, article ID, and information which has to do with association such as microRNA, gene, mutation, and disease were maintained in a results list.Entrez search returned 221 articles. 146 had text available in Pubmed Central and 106 of them had full text so it can be possible for it to be submitted for the analysis procedure and text processing. 61 of these articles passed the filtering process and as a result 23 of them listed in the results. The aim of this study is to become the automated search and information extraction process more efficient because the literature has a huge amount of data currently. The results of the present study can be a useful tool for the scientific community.

Keywords: lung cancer, miRNA, SNP, μεταλλάξεις, text mining

Περιεχόμενα

1. Εισαγωγή

1.1 ΚΑΡΚΙΝΟΣ ΤΟΥ ΠΝΕΥΜΟΝΑ

1.1.1 ΑΙΤΙΑ

Ο καρκίνος του πνεύμονα είναι ως επί το πλείστον ο συχνότερος κακοήθης όγκος. Χωρίζεται σε 4 κατηγορίες βάσει των ιστολογικών διαφορών : 1)τον μικρο- κυτταρικό καρκίνο του πνεύμονα (Small cell lung cancer) ή αλλιώς (SCLC) ,2) τον μη μικροκυτταρικό καρκίνο του πνεύμονα (Non-small cell lung cancer) ή αλλιώς (NSCLC), 3) το αδενοκαρκίνωμα (AdenoCa) και το 4)πλακώδες καρκίνωμα (squamous cell lung cancer) ή αλλιώς (SQCLC). Τα αίτια της ασθένειας αυτής φαίνεται να είναι ένας συνδυασμός έκθεσης σε συγκεκριμένους αιτιολογικούς παράγοντες και η ευαισθησία του εκάστοτε ατόμου στους παράγοντες αυτούς.Το κάπνισμα αποτελεί την κυριότερη αιτία εμφάνισης καρκίνου του πνεύμονα. Επίσης η έκθεση σε μέταλλα όπως το χρώμιο, το νικέλιο , το αρσενικό και οι αρωματικοί πολυκυκλικοί υδρογονάνθρακες σχετίζονται με την εμφάνιση του καρκίνου του πνεύμονα. Ο άσβεστος και το ραδώνιο αποτελούν σημαντικούς παράγοντες κινδύνου για τον καρκίνο του πνεύμονα [1] .

1.1.2 ΣΥΜΠΤΩΜΑΤΑ

Οι περισσότεροι ασθενείς εμφανίζουν τα συμπτώματα όταν γίνεται η διάγνωση της ασθένειας. Γενικά δεν υπάρχουν συγκεκριμένα συμπτώματα για πρόωρη διάγνωση. Τα συμπτώματα του καρκίνου του πνεύμονα είναι κατά κύριο λόγο βήχας, αιμόπτυση, δύσπνοια, απώλεια βάρους ,κούραση και πόνος στο στήθος [2] .

1.1.3 ΣΤΑΔΙΑ

Υπάρχει ένα διεθνές μοντέλο που απεικονίζει τα στάδια του όγκου και την εξάπλωσή του. Το “TNM staging system”.Η κατηγορία Τ περιγράφει το μέγεθος και την εξάπλωση του κύριου όγκου καθώς και την ακριβή τοποθεσία του όγκου που δημιουργήθηκε ξεχωριστά σε σχέση με τον πρωταρχικό όγκο.

Το στάδιο Ν(ταξινόμηση των μεταστατικών κόμβων) απεικονίζει την εξάπλωση του περιφερειακού λεμφικού κόμβου βασιζόμενο στη δομική θέση των εμπλεκόμενων λεμφαδένων και χωρίζεται σε τέσσερις υποκατηγορίες (N0-N3).

Το στάδιο Μ περιγράφει την παρουσία ή απουσία κοντινής μεταστατικής εξάπλωσης . Οι πιο συχνές θέσεις απομακρυσμένων μεταστάσεων είναι ο εγκέφαλος, το ήπαρ , ο σκελετός , τα πνευμόνια και τα επινεφρίδια [3] [4] .

1.1.4 Γενετικοί παράγοντες

Η ευαισθησία στον καρκίνο του πνεύμονα καθορίζεται εν μέρη από γενετικούς παράγοντες. Άτομα που έχουν γενετική ευαισθησία στον καρκίνο του πνεύμονα είναι πιο πιθανό να εκδηλώσουν την ασθένεια αν καπνίσουν κατά τη διάρκεια της ζωής τους. Αν υπάρχει οικογενειακό ιστορικό για καρκίνο του πνεύμονα υπάρχει 1,7 φορές μεγαλύτερη πιθανότητα το άτομο να αναπτύξει καρκίνο του πνεύμονα κάποια στιγμή. Η μελέτη του γονιδιώματος (GWAS) έχει συσχετίσει τις περιοχές των χρωμοσωμάτων 5p15,15q25-26 και 6q21 με αυξημένο κίνδυνο εμφάνισης καρκίνου του πνεύμονα [5][2] .

1.1.5 Περιβαλλοντικοί παράγοντες

Οι περιβαλλοντικοί παράγοντες αποτελούν σημαντικό παράγοντα κινδύνου για την εμφάνιση της νόσου. Μπορούν να είναι:φυσικοί όπως ιονίζουσα και μη ιονίζουσα ακτινοβολία, η έκθεση στο ραδόνιο, η υπεριώδης ακτινοβολία χημικοί όπως ο άσβεστος, οι διοξίνες, μέταλλα όπως το αρσενικό, το νικέλιο, το χρώμιο και το κοβάλτιο καθώς και άλλοι ρύποι όπως βιομηχανικές εκπομπές ρύπων και καπνός σπιτιών.[6]

1.1.6 Επιδημιολογία

Ο καρκίνος του πνεύμονα είναι η κύρια αιτία θανάτου στους άνδρες ήδη από τις αρχές της δεκαετίας του 1950. Το 1987 ξεπέρασε σε ποσοστά θνησιμότητας και τον καρκίνο του μαστού στις Η.Π.Α. Έχει τους περισσότερους θανάτους σε άνδρες και γυναίκες από οποιαδήποτε άλλη μορφή καρκίνου[8]. Οι τιμές του καρκίνου του πνεύμονα στους άνδρες είναι σημαντικά υψηλότερες στις αναπτυγμένες χώρες σε σχέση με τις μη αναπτυγμένες. Στις γυναίκες ο καρκίνος του πνεύμονα έχει επίσης μεγαλύτερη συχνότητα εμφάνισης στις αναπτυγμένες χώρες σε σχέση με τις αναπτυσσόμενες και σχετίζεται με το κάπνισμα. Η συχνότητα εμφάνισης του καρκίνου του πνεύμονα στις γυναίκες ολοένα και αυξάνεται ενώ το 2017 ξεπέρασε τη συχνότητα εμφάνισης του καρκίνου του μαστού. Στην Ασία υπάρχει επίσης μεγαλύτερο ποσοστό εμφάνισης καρκίνου του πνεύμονα στις γυναίκες εξαιτίας της μόλυνσης του αέρα και της έκθεσης τους σε βλαβερούς περιβαλλοντικούς παράγοντες που τους επιβάλλει η εργασίας τους. Οι χώρες με υψηλότερο εισόδημα φαίνεται να έχουν καλύτερα ποσοστά επιβίωσης σε σχέση με τις χώρες με χαμηλότερο εισόδημα [8] .

1.2 MICRORNAS

1.2.1 Βιογένεση

Η βιογένεση διαχωρίζεται σε κανονικά και μη-κανονικά μονοπάτια. Η συμβατική οδός βιογένεσης αποτελείται από δύο γεγονότα διάσπασης, ένα πυρηνικό και ένα κυτταροπλασματικό.Ωστόσο υπάρχουν εναλλακτικές οδοί βιογένεσης που διαφέρουν ως προς τον αριθμό γεγονότων διάσπασης και στα ένζυμα που ευθύνονται για τη διάσπαση αυτή.[10] Η κανονική βιογένεση είναι το κυρίαρχο μονοπάτι στο οποίο γίνεται η επεξεργασία των miRNAs. Σε αυτό, τα pri-miRNAs μεταγράφονται από τα γονίδιά τους με τη βοήθεια της RNA-πολυμεράσης II . Περίπου τα μισά από όλα τα miRNAs είναι ενδογονιδιακά οπότε επεξεργάζονται κυρίως από ιντρόνια και σε μικρότερο βαθμό από εξώνια γονιδίων που κωδικοποιούν πρωτεΐνες. Τα υπόλοιπα είναι διαγονιδιακά και μεταγράφονται ανεξάρτητα από ένα γονίδιο ξενιστή και ρυθμίζονται από τους δικούς τους υποκινητές [10] .
Έπειτα όσο βρίσκονται μέσα στον πυρήνα επεξεργάζονται σε pre-miRNAs με τη βοήθεια ενός συμπλέγματος μικροεπεξεργαστή που αποτελείται από μια πρωτεΐνη πρόσδεσης την DiGeorge Syndrome Critical Region 8 (DGCR8) και ένα ένζυμο ριβονουκλεάσης III, την Drosha .Η Drosha διασπά το δίκλωνο pri-miRNA στη βάση της χαρακτηριστικής δομής της φουρκέτας του pri-miRNA και η DGCR8 αναγνωρίζει μια N6 μεθυλαδενυλιωμένη GGAC και άλλα μοτίβα που βρίσκονται μέσα στο pri- miRNA. Αυτό οδηγεί στο σχηματισμό μιας 3’ εξοχής 2 νουκλεοτιδίων στο pre-miRNA [10] .
Κατά την εξαγωγή τους στο κυτταρόπλασμα το πρώιμο miRNA διασπάται από τη Dicer κοντά στον τερματικό βρόχο απελευθερώνοντας ένα μικρό δίκλωνο RNA.Η έξοδός τους γίνεται με τη βοήθεια ενός συμπλόκου exportin 5(XPO5)/RanGTP και αναλαμβάνει την επεξεργασία τους η RNA III ενδονουκλεάση Dicer.Σε αυτή την επεξεργασία απομακρύνεται ο τερματικός κλώνος και δημιουργείται ένα ώριμο δίκλωνο pri-miRNA.Η κατεύθυνση του miRNA κλώνου δηλώνει και το όνομα του ώριμου miRNA [10][11] .

Η Dicer συνδέεται με το pre-miRNA με προτίμηση στο 3’ άκρο μήκους 2 νουκλεοτιδίων που δημιουργήθηκε αρχικά από την Drosha 46. Στα θηλαστικά και στα πτηνά υπάρχει ένας πρόσθετος μηχανισμός για τον προσδιορισμό της θέσης διάσπασης του pre-miRNA.Η Dicer συνδέεται με το 5’ φωσφορυλιωμένο άκρο του pre-miRNA και διασπάται σε 22 νουκλεοτίδια μακριά από το 5΄άκρο [10] .

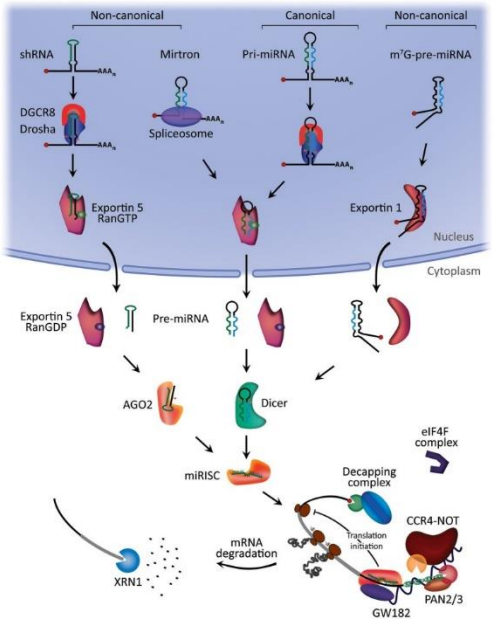
Οι συμπληρωματικοί κλώνοι που προκύπτουν από το δίκλωνο ώριμο miRNA μπορούν να εισαχθούν στην οικογένεια πρωτεϊνών Argonaute (AGO) με έναν ATP-εξαρτώμενο τρόπο. Για κάθε δοθέν miRNA η αναλογία του κλώνου που έχει εισαχθεί η AGO ποικίλλει ανάλογα με τον τύπο του κυττάρου η το κυτταρικό περιβάλλον που κυμαίνονται από σχεδόν ίσες αναλογίες από το ένα ή το άλλο. Η επιλογή 3΄ή 5΄άκρου είναι εν μέρει βασισμένη στη θερμοδυναμική σταθερότητα των 5΄άκρων των δίκλωνων

miRNAs ή η ύπαρξη του νουκλεοτιδίου της ουρακίλης στο 5’ άκρο στη θέση του νουκλεοτιδίου 1 [10] .

Συνήθως ο κλώνος με τη λιγότερη σταθερότητα στο 5’ άκρο ή με νουκλεοτίδιο ουρακίλης στο 5΄άκρο εισάγεται στην πρωτεΐνη AGO και θεωρείται ο κλώνος οδηγός. Ο κλώνος που δεν εισάγεται στην πρωτεΐνη ονομάζεται κλώνος επιβάτης. Οι κλώνοι επιβάτες των miRNAs που δεν περιέχουν διαφορές διασπώνται από την AGO2 και υποβαθμίζονται από τον κυτταρικό μηχανισμό, κάτι το οποίο μπορεί να οδηγήσει σε μεροληψία κλώνου. Τέλος, αν υπάρχουν δίκλωνα miRNAs με κεντρικές αναντιστοιχίες ή miRNAs που δεν έχουν εισαχθεί στην πρωτεΐνη AGO2 τότε ξετυλίγονται παθητικά και υποβαθμίζονται .

<u>Μη κανονικά μονοπάτια βιογένεσης :</u>

Τα μη συμβατικά μονοπάτια της βιογένεσης κάνουν χρήση διαφορετικών συνδυασμών των πρωτεϊνών που εμπλέκονται στα συμβατικά μονοπάτια βιογένεσης, κυρίως η Drosha η Dicer, exportin 5 και η AGO2. Γενικά τα μονοπάτια της μη- συμβατικής βιογένεσης μπορούν να χωριστούν στο ανεξάρτητο μονοπάτι Drosha/DGCR8 και στο ανεξάρτητο μονοπάτι Dicer. Τα pre-miRNAs παράγονται από το ανεξάρτητο μονοπάτι Drosha/DGCR8 όμοιο με τα υποστρώματα της Dicer. Παράδειγμα τέτοιων Pre-miRNAs αποτελούν τα μιρτρόνια, που παράγονται από τα ιντρόνια του mRNA κατά τη διάρκεια του ματίσματος. Άλλο ένα αντίστοιχο παράδειγμα αποτελεί η 7-μεθυλο γουανοσινη(m^7G) pre-miRNA. Αυτά τα εκκολαπτόμενα RNAs εξάγονται αμέσως στο κυτταρόπλασμα μέσω της exportin-1 χωρίς να έχουν ανάγκη από τη διάσπαση της Drosha. Πολλές φορές μπορεί να υπάρξει μεροληψία κλώνου 3p πιθανόν εξαιτίας του καπακιού m^7G που εμποδίζει τον 5p κλώνο να εισαχθεί στην πρωτεΐνη AGO. Από την άλλη πλευρά τα ανεξάρτητα από την Dicer miRNAs υφίστανται επεξεργασία από την Drosha από ενδογενή μετάγραφα της κοντής φουρκετας RNA . Αυτά τα pre- miRNAs απαιτούν την AGO2 ώστε να ολοκληρώσουν την ωρίμανση τους μέσα στο κυτταρόπλασμα εξαιτίας του μη- ικανοποιητικού μεγέθους τους ώστε να είναι υποστρώματα της Dicer. Αυτό προκαλεί την εισαγωγή ολόκληρου του pre-miRNA στην AGO2 και στο εξαρτώμενη AGO2 μάτισμα του σκέλους 3p. Το 3’-5’ κούρεμα του κλώνου 5p ολοκληρώνουν την ωρίμανσή τους [10] .



Εικόνα 1βιογένεση των microRNAs [18]

1.2.3 RISC

Το ελάχιστο σύμπλοκο αποσιώπησης αποτελείται από τον κλώνο οδηγό και την πρωτεΐνη AGO. Η εξειδίκευση του στόχου του mi-RISC οφείλεται στην αλληλεπίδραση του με συμπληρωματικές αλληλουχίες στο mRNA στόχο που ονομάζονται στοιχεία απόκρισης miRNA(MRES) [10] [11] .

1.2.4 Ρύθμιση γονιδιακής έκφρασης

Τα microRNAs είναι υπεύθυνα για τη ρύθμιση της έκφρασης των γονιδίων μετά τη μεταγραφή. Γενικά δεσμεύονται στην 3' αμετάφραστη περιοχή των mRNA στόχων τους και καταστέλλουν την παραγωγή πρωτεΐνης από σταθεροποιώντας το mRNA και τη μεταφραστική σίγαση [12] .

1.2.5 Ο Ρόλος των microRNAs στον καρκίνο του πνεύμονα

Πολλά ανθρώπινα microRNAs (miRNAs) φαίνεται να βρίσκονται σε γονιδιωματικές περιοχές που σχετίζονται με τον καρκίνο του πνεύμονα ή σε κομβικά σημεία και να ελέγχουν σημαντικές διεργασίες όπως ο πολλαπλασιασμός κυττάρων, η κυτταρική προσκόλληση, η απόπτωση, η αγγειογένεση και η δυσλειτουργία. Τα οποία παίζουν πολύ σημαντικό ρόλο στην εξέλιξη και τη μετάσταση του καρκίνου. Τα miRNAs έχουν τη δυνατότητα να δρουν ως ογκογονίδια ή ογκοκατασταλτικά γονίδια το οποίο εξαρτάται από το κυτταρικό περιεχόμενο και τα γονίδια στόχους. Oncomirs είναι ογκογονίδια miRNA που έχουν περιστασιακό ρόλο στην επίθεση ή την επικράτηση των φαινοτύπων του καρκίνου. Η οικογένεια Let-7 ήταν το πρώτο σύνολο Oncomirs που αποδείχθηκε ότι ρυθμίζει την έκφραση του γονιδίου Ras, το οποίο είναι το τρίτο πιο γνωστό γονίδιο που σχετίζεται με τον καρκίνο στον άνθρωπο [13] . Το oncomir mir-21 στοχεύει την προώθηση, την εισβολή και μετανάστευση των γονιδίων PTEN και SPRY2 καθώς και την ογκογένεση με αναστολή των αρνητικών ρυθμιστών του μονοπατιού RAS/MEK/ERK. Έχουν βρεθεί αυξημένα επίπεδα έκφρασης του mir-21 σε ένα διαφορετικό υποσύνολο καρκινικών κυτταρικών σειρών και ιστών όπως ο καρκίνος του μαστού, ο καρκίνος του παγκρέατος, ο καρκίνος του μαστού και άλλο τύποι καρκίνου . Το miR 34a ως ρυθμιστής κλειδί για την καταστολή όγκων ελέγχει την έκφραση πληθώρας πρωτεϊνών – στόχων που εμπλέκονται στον κυτταρικό κύκλο, την κυτταρική διαφοροποίηση και την απόπτωση. Η έκφραση του miR-96 μειώνει την εισβολή και τη μετανάστευση των καρκινικών κυττάρων και επιβραδύνει την ανάπτυξη του όγκου, που σχετίζεται με την καταστολή του γονιδίου KRAS. Τα ογκοκατασταλτικά miRNAs όπως το miR-96 και το miR-34a καταστέλλονται σε πρωταρχικούς όγκους , όμως ορός του miR-34a βρέθηκε να υπερεκφράζεται στους ασθενείς με καρκίνο του στομάχου [7] .

1.3 Μεταλλάξεις

Οι πιο γνωστές μεταλλάξεις που σχετίζονται με τον καρκίνο του πνεύμονα είναι οι EGFR mutation η οποία αποτελεί μια από τις πιο συχνές μεταλλάξεις που εντοπίζονται στον NSCLC και είναι πιο συχνές σε μη καπνίζοντες πληθυσμούς της Ασίας. Συναντώνται συχνότερα σε αδενοκαρκινώματα με πολύ λιγότερες πιθανότητες εμφάνισης σε πλακώδη κυτταρικά καρκινώματα. TP53 mutation είναι από τις επικρατέστερες μεταλλάξεις και συναντάται σε πολλούς τύπους καρκίνου κυρίως σε SCC με ποσοστό 43% και έπειτα σε AC με ποσοστό 35%. KRAS mutation διάφοροι τύποι της KRAS μετάλλαξης βρίσκονται σε διαφορετικούς τύπους καρκίνου όπου 19% των ασθενών με AC να έχει μετάλλαξη KRAS και μόνο ένα 5% με SCC να έχει αυτή τη μετάλλαξη. Οι ασθενείς με μετάλλαξη KRAS στον NSCLC , δεν επωφελούνται από τη χημειοθεραπεία στα αρχικά στάδια της νόσου και εμφανίζουν χειρότερα

κλινικά αποτελέσματα στη θεραπεία με αναστολείς κινάσης τυροσίνης καθώς το γονίδιο KRAS αποτελεί μέρος της οδού σηματοδότησης του EGFR.Η μετάλλαξη PIK3CA σχετίζεται με ταχύτατη εξέλιξη της νόσου και κάποιες κλινικές έρευνες δείχνουν ότι η μετάλλαξη αυτή εμφανίζει αντίσταση στη θεραπεία με αναστολείς κινάσης τυροσίνης (TKIS) .Η μετάλλαξη BRAF εμπλέκεται στην παθογένεση πολλών τύπων καρκίνων, με πιο συχνή τη μετάλλαξη V600E που ευθύνεται για το 90% των μεταλλάξεων στο μελάνωμα. Μια όχι και τόσο συχνή μετάλλαξη είναι η CDKN2A που αφορά ένα ογκοκατασταλτικό γονίδιο που παίζει σημαντικό ρόλο στη ρύθμιση του κυτταρικού κύκλου και στην πρωτεΐνη τελεστή TP53 . Ένα ποσοστό 7-8% των τύπων καρκίνου AC και SCC εμφανίζονται να έχουν αυτή τη μετάλλαξη [9].

1.4 Στόχος εργασίας

Οι βιολογικές βάσεις δεδομένων όπως είναι η PubMed και η PubMed Central απαρτίζονται τη δεδομένη χρονική στιγμή από πολύ μεγάλους αριθμούς εγγραφών που ξεπερνούν τα 28 εκατομμύρια . Οπότε και όγκος της πληροφορίας που υπάρχει είναι μεγάλος καθιστώντας την αναζήτηση και απομόνωση της μια ιδιαίτερα χρονοβόρα και πολύπλοκη διαδικασία . Έτσι η συνεισφορά της αυτοματοποιημένης αναζήτησης και εξαγωγής κειμένου φαίνεται να είναι πολύ χρήσιμη καθώς στοχεύει στην εξαγωγή καλύτερης ποιότητας πληροφορίας σε μικρό χρονικό διάστημα .

Για να μπορέσει να ερευνηθεί η σχέση μετάλλαξης miRNA::Mrna με την ασθένεια του καρκίνου του πνεύμονα έπρεπε να γίνει συλλογή της πληροφορίας από την υπάρχουσα βιβλιογραφία . Για το σκοπό αυτό αναπτύχθηκε ένα σύστημα αυτοματοποιημένης αναζήτησης που μέσα από την ανάλυση ονομάτων οντοτήτων και φιλτραρίσματος των αποτελεσμάτων καταλήξαμε στα σχετικά για την πληροφορία που ψάχναμε άρθρα στα οποία έγινε μελέτη και χειροκίνητος σχολιασμός .

2. Μέθοδοι

2.1 Αναζήτηση του ερωτήματος (query) στην Entrez μέσω της Esearch():

Για να μπορέσουμε να συλλέξουμε όσο το δυνατόν πιο στοχευμένη πληροφορία από τη βιβλιογραφία χρειάστηκε να δημιουργηθεί το ερώτημα (query) το οποίο μέσω της αυτοματοποιημένης αναζήτησης στο Entrez θα μας επέστρεφε τα επιστημονικά άρθρα που μας ενδιαφέρουν. Για να χαρακτηριστεί ένα επιστημονικό άρθρο ως άρθρο ενδιαφέροντος πρέπει να πληροί 3 βασικές προϋποθέσεις : Να αναφέρεται η ασθένεια που μελετάμε (καρκίνος του πνεύμονα), ένας τύπος microRNA και ένας τύπος μετάλλαξης. Το query περιέχει τους όρους της αναζήτησης που ενώνονται μεταξύ τους με λογικό ΚΑΙ (AND) , Ή(OR) και ΟΧΙ(NOT). Ο τελεστής OR χρησιμοποιήθηκε για να εξασφαλίσει τους διαφορετικούς τρόπους με τους οποίους μπορεί να αναφέρεται κάποιος στη μελέτη. Για την εύρεση του όρου της μετάλλαξης (mutation) χρησιμοποιήθηκαν επίσης οι όροι variant, SNP, polymorphism. Ο τελεστής AND χρησιμοποιήθηκε για να ενώσει τους διαφορετικούς όρους του ερωτήματος (query) στην αναζήτηση. Ενώ ο τελεστής NOT για να εξασφαλίσουμε ότι δεν θέλουμε να μας επιστραφεί η συγκεκριμένη μορφή ενός όρου από την αναζήτηση μας. Στην προκειμένη δε θέλαμε τα άρθρα μας να έχουν ως τύπο δημοσίευσης το REVIEW, δηλαδή να στηρίζεται σε μεθόδους και συμπεράσματα προηγούμενων μελετών. Για αυτόν τον λόγο και χρησιμοποιούμε τον τελεστή NOT Review με ετικέτα [ptyp] που αντιπροσωπεύει τον τύπο δημοσίευσης (publication type). Επίσης για να εξασφαλίσουμε περισσότερη ακρίβεια αποτελεσμάτων συμπεριλάβαμε και τις ετικέτες Title/Abstract. Τελικά το query στο οποίο καταλήξαμε ήταν της μορφής:

('"lung cancer[Title/Abstract]**AND** (mutation[Title/Abstract] **OR** polymorphism[Title/Abstra ct] **OR** variant[Title/Abstract]**OR**snp[Title/Abstract])**AND**microrna[Title/ABSTRACT]**OR** mirna[Title/Abstract])**NOT** Review[ptyp]')

Τα αποτελέσματα του query που δημιουργήσαμε παραπάνω μας επιστράφηκαν μέσω της συνάρτησης esearch() η οποία αναζητά και ανακτά κύρια αναγνωριστικά τα οποία δύναται να διατηρήσει για μελλοντική χρήση στο περιβάλλον του χρήστη. Ως ορίσματα της η esearch() δέχεται:

- τη βάση δεδομένων στην οποία θα γίνει η αναζήτηση(db), στην περίπτωση μας είναι η Pubmed.
- ένα συνολικό αριθμό των αναγνωριστικών (ID) που θα μας επιστραφούν(retmax).Επειδή θέλουμε να πάρουμε το μέγιστο αριθμό άρθρων που ικανοποιούν το αντίστοιχο query θέτουμε το retmax με μια ακραία μεγάλη τιμή της τάξης των 100.000.
- το ερώτημα της αναζήτησης (query) το οποίο εισάγεται στη μεταβλητή term.

Τα αποτελέσματα επιστρέφονται σε μορφή XML(Extensive Markup Language) και χρησιμοποιούνται στη συνάρτηση read() ως είσοδος ώστε να αναλυθούν και επιστρέφονται με μια πολύ επίπεδη δομή δεδομένων λιστών και λεξικών της Python.

Η συνάρτηση efetch() ανακτά εγγραφές στην απαιτούμενη μορφή από μια λίστα με μία ή περισσότερες διεπαφές χρήστη ή εγγραφές που προέρχονται από το περιβάλλον του χρήστη. Ως ορίσματα δεχεται

- τη βάση δεδομένων στην οποία θα γίνει η αναζήτηση (db) , στην προκειμένη περίπτωση πρόκειται για την pubmed.
- μια λίστα με τις εγγραφές / αναγνωριστικά (ids), εδώ πρόκειται για τη λίστα με τα ids που επιστράφηκαν από την αναζήτηση της esearch().
- το επόμενο όρισμα (retmode) αφορά τη μορφή με την οποία επιλέγουμε να επιστραφούν τα δεδομένα των εγγραφών και στην περίπτωσή μας πρόκειται για το XML.
- Τέλος το rettype καθορίζει την προβολή της εγγραφής που επιστρέφεται, εδώ βάλαμε full .

Τα δεδομένα που μας επιστράφηκαν από την αναζήτηση στην Entrez σε μορφή XML αποθηκεύονται σε ένα αρχείο με ονομασία filenew.xml.

2.2 Name Entity Recognition(NER)

Το query που κατασκευάσαμε και αναζητήσαμε αρχικά μας επιστρέφει ένα ευρύ φάσμα επιστημονικών άρθρων που δεν μπορούμε να είμαστε σίγουροι για το πόσο αξιόπιστη είναι η πληροφορία που μας επιστρέφεται. Για να περιορίσουμε το εύρος των αποτελεσμάτων μας και να έχουμε μεγαλύτερη ακρίβεια χρησιμοποιήσαμε τη διαδικασία αναγνώρισης ονομάτων οντοτήτων (NER). Η διαδικασία αναγνώρισης ονομάτων οντοτήτων είναι μια διαδικασία εξαγωγής πληροφοριών και σχέσεων. Το NER έχει ως στόχο να εντοπίσει και να ταξινομήσει επώνυμες οντότητες που αναφέρονται σε μη δομημένο κείμενο. Προκειμένου να εφαρμόσουμε τη διαδικασία αναγνώρισης ονομάτων-οντοτήτων χρησιμοποιούμε το πακέτο της Python scispaCy που περιέχει μοντέλα για επεξεργασία επιστημονικών ή βιοϊατρικών κειμένων. Για την παρούσα μελέτη χρησιμοποιήθηκε ένα μοντέλο του scispacy και συγκεκριμένα το en_ner_bionlp13cg_md, το οποίο χρησιμοποιείται για όρους που αφορούν μετάλλαξη, microRNA και γονίδιο στην κλάση GENE_OR_GENE_PRODUCT. Ως είσοδος για το μοντέλο bio_nlp13cg_md δημιουργήσαμε το αρχείο filenew1.xml στο οποίο αποθηκεύονται όλα τα ids που απομονώθηκαν από την PMC(PubMed Central). Η εύρεση των άρθρων αυτών έγινε μέσα από την εξαγωγή των δεδομένων που υπάρχουν xml αρχείο filenew.xml.

Για την εύρεση των άρθρων που είχαν id στην PMC χρησιμοποιήθηκε το πακέτο bs4.BeautifulSoup το οποίο εξάγει δεδομένα από ένα xml αρχείο ανάλογα με τον αναλυτή που επιλέγεται. Στη δική μας περίπτωση επιλέξαμε τον lxml parser. Η συνάρτηση findall() αναζητά στο αρχείο το όρισμα ArticleID

με Id Type την pmc και επιστρέφονται όλες οι δημοσιεύσεις που πληρούν αυτή την προϋπόθεση.Επίσης για την εύρεση όρων βιοϊατρικού ενδιαφέροντος χρησιμοποιήθηκε και το pipeline en_core_sci_sm το οποίο αναγνωρίζει βιοιατρικά δεδομένα εντός των προτάσεων. Η σωλήνωση παίρνει ως είσοδο μια γλώσσα και με επεξεργασία φυσικής γλώσσας επιστρέφει μια λίστα οντοτήτων βιοϊατρικού ενδιαφέροντος που εντοπίζει τις προτάσεις. Οι προτάσεις απομονώθηκαν με τη βοήθεια του εργαλείου tokenize από το λίστα πακέτων και βιβλιοθηκών NLTK(Natural Language ToolKit) που χρησιμοποιείται για τη συμβολική και στατιστική επεξεργασία φυσικής γλώσσας για κείμενα που είναι γραμμένα στα αγγλικά και στη γλώσσα προγραμματισμού της Python. Η ανάλυση των αποτελεσμάτων διατηρήθηκε σε ένα αρχείο με όνομα sentence_seg.csv το οποίο αποτελείται από το αναγνωριστικό (id) του κάθε άρθρου στην PMC (PMCID) , τον τίτλο του κάθε άρθρου (Title), τις προτάσεις που επιλέχθηκαν (SENTENCE) και τους όρους των επιστρεφόμενων προτάσεων (DOC_ENTS).

Φίλτρα Αναζήτησης

Τα τελικά άρθρα που μας ενδιαφέρουν θέλουμε να πληρούν 2 βασικές προϋποθέσεις:

- Η πρώτη προϋπόθεση είναι σε κάθε άρθρο να αναφέρεται τουλάχιστον μια φορά η ασθένεια για την οποία ερευνάμε, ένα γονίδιο, ένας τύπος microRNA και ένας πολυμορφισμός μονονουκλεοτιδίου(SNP single nucleotide polymorphism).
- Η δεύτερη προϋπόθεση είναι να αναφέρονται τουλάχιστον 2 από τους προαναφερθέντες όρους σε μια τουλάχιστον πρόταση του άρθρου.

Το πρώτο φίλτρο που κατασκευάσαμε έγινε με σκοπό ο έλεγχος να ικανοποιεί την πρώτη προϋπόθεση. Ως είσοδο δέχεται το αρχείο entities_seg.csv και ελέγχει για κάθε id αν ικανοποιούνται τα κριτήρια της πρώτης προϋπόθεσης. Προκειμένου να εντοπίσει το πρόγραμμα τον τύπο των microRNAs που υπάρχουν στο άρθρο ή την ασθένεια χρησιμοποιήσαμε ως αναγνωριστικό κανονικές εκφράσεις. Για τα γονίδια ο έλεγχος έγινε μέσω αρχείου από την HGNC το οποίο περιέχει όλες τις απαραίτητες πληροφορίες που σχετίζονται με τα γονίδια. Το σύστημα λοιπόν ελέγχει 3 φορές ξεχωριστά αν ο τύπος GENE_OR_GENE_PRODUCT που υπάρχει στην τελευταία στήλη του αρχείου entities αντιστοιχεί σε κάποιο από τους τύπους δεδομένων που ψάχνουμε.Υπάρχουν 3 συνθήκες ελέγχου μια για κάθε όρο όπου αν η συνθήκη είναι αληθής τότε και το αναγνωριστικό(flag) που υπάρχει έξω από τα σημεία γίνεται κι αυτό αληθές. Μετά από αυτές τις συνθήκες ελέγχου υπάρχει και μια τέταρτη όπου αν κι αυτή είναι αληθής τότε το επιβεβαιωμένο id του άρθρου καταχωρείται σε μια λίστα μαζί με όλα τα id που πέρασαν επιτυχώς το πρώτο φίλτρο με ονομασία id_list.

Το δεύτερο φίλτρο λαμβάνει ως είσοδο το αρχείο sentence_seg.csv. Εκεί γίνεται έλεγχος μόνο για τα id που περιέχονται στην id_list, τη λίστα με τα ids πέρασαν με επιτυχία το πρώτο φίλτρο. Στο αρχείο που μας επιστράφηκε με τα entities η λίστα DOC_ENTS περιέχει όρους των επιστρεφόμενων προτάσεων. Οπότε ελέγχεται κάθε όρος του DOC_ENTS αν αντιστοιχεί σε SNP, microRNA ή ασθένεια.

Για τη μετάλλαξη (SNP) χρησιμοποιήθηκαν πάλι κανονικές εκφράσεις. Οπότε εφόσον η συνθήκη ελέγχου για την ύπαρξη rs είναι αληθής και βρίσκει τουλάχιστον έναν από τους 3 όρους microRNA,gene και ασθένεια μέσω ενός counter τότε η συνθήκη ελέγχου είναι αληθής και μας επιστρέφεται ένα τελικό αρχείο τύπου txt με όνομα finalresults.txt που περιέχει όλα τα ids των άρθρων που έχουν περάσει και τα δυο φίλτρα και αναγράφεται ο τίτλος τους και μια μικρή περίληψη για το τι πραγματεύεται το κάθε άρθρο.

Χειροκίνητος σχολιασμός αποτελεσμάτων

Το finalresults.txt που επιστρέφεται στο τέλος του προηγούμενου βήματος μας χρησιμεύει ώστε να προχωρήσουμε σε χειροκίνητο φιλτράρισμα των αποτελεσμάτων καθώς μπορεί να μην πρόκειται για μια επιβεβαιωμένη επιστημονική μελέτη αλλά για μια στατιστική ανάλυση που στηρίχθηκε σε αποτελέσματα προυπάρχουσων ερευνών. Επίσης μπορεί να αναφέρονται οι όροι ενδιαφέροντος δηλαδή γονίδιο,microRNA,SNP χωρίς όμως να αποδεικνύεται πειραματικά στην παρούσα μελέτη η σχέση τους με την ασθένεια. Ο κατάλογος του χειροκίνητου σχολιασμού αποτελείται από τις εξής στήλες:

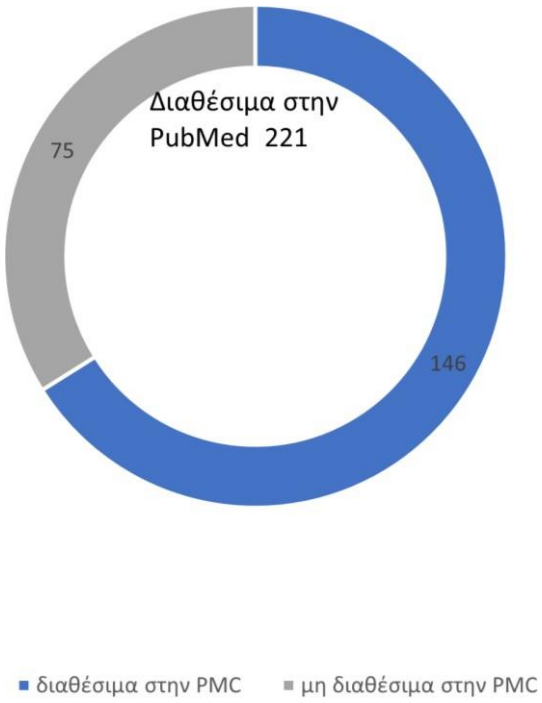
- Pmid: το αναγνωριστικό του άρθρου
- MicroRNA: τον κάθε τύπο miRNA
- Gene: το όνομα του γονιδίου
- Association: την ύπαρξη σχέσης ανάμεσα στον τύπο miRNA, το αντίστοιχο γονίδιο , το SNP και την ασθένεια. Αν υπάρχει σχέση ανάμεσα σε αυτούς τους όρους παίρνει τιμή TRUE αλλιώς FALSE.
- Risk:αν υπάρχει αύξηση η μείωση του ρίσκου της ασθένειας σε σχέση με κάποιον από τους τύπους που μελετώνται.Αν υπάρχει αύξηση του ρίσκου στον πίνακα αναγράφεται increased αλλιώς decreased.
- Variant:το id του κάθε πολυμορφισμού
- Gene Biotype: εδώ παραθέτουμε την ομάδα στην οποία ανήκει το γονίδιο όπως είναι καταγεγραμμένο στη Genecards
- Variant region:η περιοχή στην οποία εντοπίζεται ο πολυμορφισμός όπως και το αν βρίσκεται σε miRNA ή στο γονίδιο.
- Disease:η ονομασία της ασθένειας
- Drug:αν υπάρχει φαρμακευτική αγωγή
- Drug response:η απόκριση στην φαρμακευτική αγωγή, αν είχε θετική επίδραση βάζουμε την τιμή POSITIVE αλλιώς NEGATIVE
- Population:τον πληθυσμό στον οποίο έγινε η μελέτη
- Cells:τις κυτταρικές σειρές
- Study:τι είδους μελέτη πραγματοποιήθηκε

- Sentences:τις προτάσεις που μας επιστράφηκαν από το προηγούμενο βήμα κατά τη διάρκεια του φιλτραρίσματος όπως αναφέρονται στο αρχείο sentences_seg.csv
- Comments: κάποια επιπλέον σχόλια που θεωρήσαμε πως έπρεπε να παραθέσουμε

3.Αποτελέσματα

3.1 Αποτελέσματα αυτοματοποιημένης αναζήτησης στην Entrez και φιλτραρίσματος

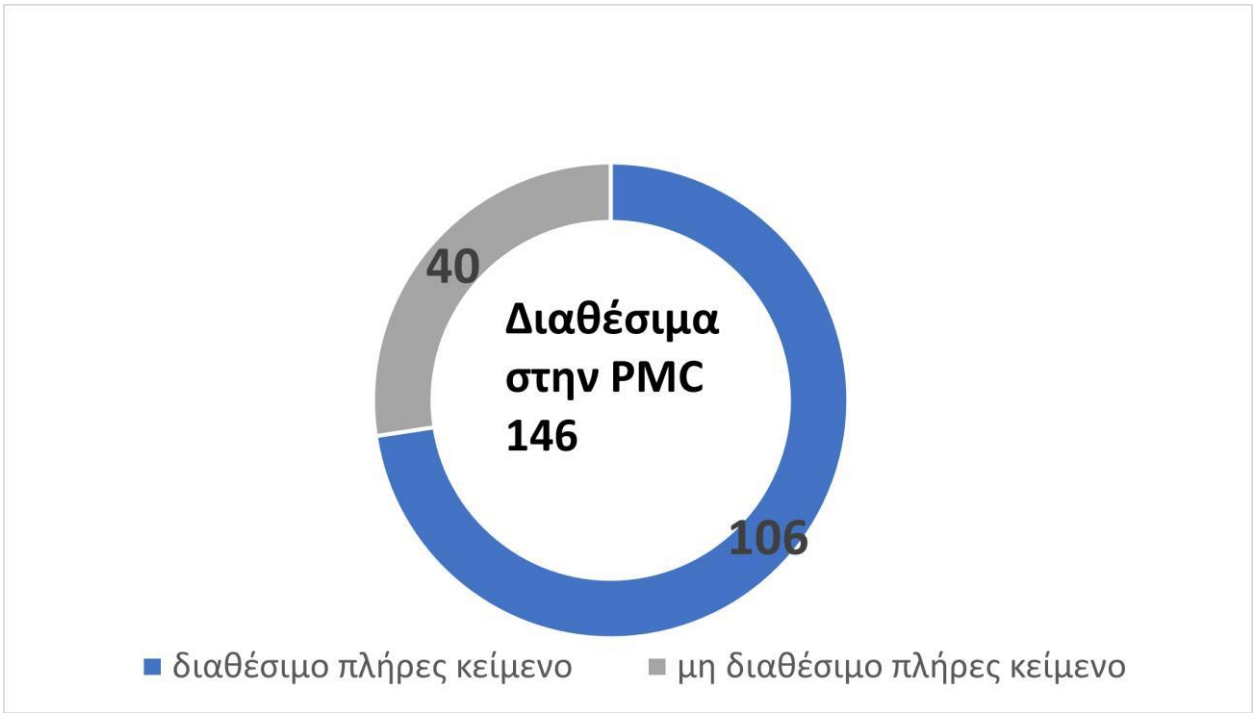
Η αναζήτηση του ερωτήματος (query) στην Entrez μας επέστρεψε 221 άρθρα από την Pubmed. Από αυτά τα 146 ανήκαν στην PMC(Εικόνα1). Από αυτά τα 106 βρέθηκαν να έχουν πλήρες κείμενο διαθέσιμο σε xml μορφή (full text)(Εικόνα2). Με την αναγνώριση ονομάτων-οντοτήτων και την εφαρμογή του μοντέλου του scispacy βρήκαμε ότι για τα 106 άρθρα επιστρέφονται 36.823 οντότητες που αποθηκεύτηκαν σε ένα αρχείο με όνομα entities_seg.csv και 27.023 προτάσεις που αποθηκεύτηκαν σε ένα αρχείο με ονομασία sentence_seg.csv. Μετά την εφαρμογή των φίλτρων μας επιστράφηκαν 61 άρθρα. Τα παραπάνω αρχεία χρησιμοποιήθηκαν για το τελικό φιλτράρισμα των αποτελεσμάτων όπου βρέθηκε ότι 61 άρθρα από τα 106 που είχαν πλήρες κείμενο πέρασαν επιτυχώς και τα 2 φίλτρα του συστήματος(Εικόνα 3). Το τελικό αρχείο finalresults.txt περιείχε τα 61 άρθρα που πέρασαν από την εφαρμογή των φίλτρων μαζί με κάποια περίληψη για το περιεχόμενο των άρθρων αυτών και τον τίτλο της κάθε δημοσίευσης με το αντίστοιχο PMCID της.



Εικόνα 2 Τα άρθρα που ήταν διαθέσιμα στην PMC

3.2Αποτελέσματα του χειροκίνητου σχολιασμού των αποτελεσμάτων

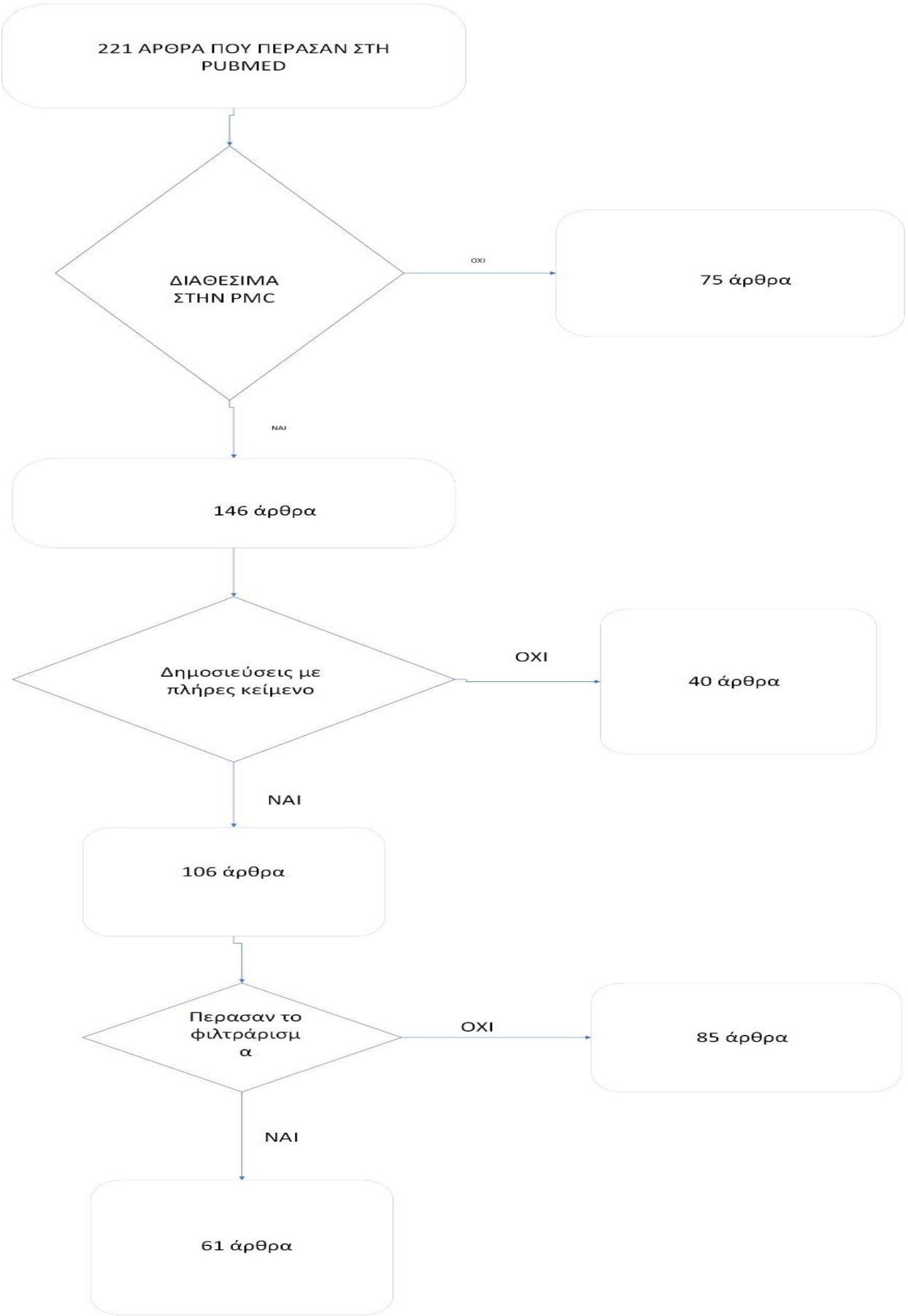
Με τον χειροκίνητο σχολιασμό των αποτελεσμάτων καταλήξαμε σε 49 εγγραφές οι οποίες απαρτίζουν τον πίνακα μας. Από τις 49 αυτές εγγραφές προέκυψε ότι 23 δημοσιεύσεις από τις 61 είχαν την κατάλληλη πληροφορία και κατέληξαν στον τελικό κατάλογο(Εικόνα 4). Οι δημοσιεύσεις οι οποίες δεν πέρασαν τον χειροκίνητο σχολιασμό ενλω πληρούσαν τα κριτήρια των φίλτρων που εφαρμόστηκαν στο προηγούμενο βήμα της έρευνας μας απορρίφθηκαν για δύο λόγους. Ο πρώτος λόγος είναι ότι παρόλο που αναφερόταν μέσα η τριάδα ενδιαφέροντος μας, δεν είχε πραγματοποιηθεί στη συγκεκριμένη δημοσίευση κάποια έρευνα και στηριζόταν σε συμπεράσματα προηγούμενων ερευνών. Ο δεύτερος λόγος είναι ότι μπορεί για το φίλτρο του προηγούμενου βήματος να αναφέρονταν και οι 4 όροι ενδιαφέροντος, ώστε να γίνει η διερεύνηση της συσχέτισης miRNA::μετάλλαξη::γονίδιο::ασθένεια. Μέσα από τον χειροκίνητο σχολιασμό διαπιστώσαμε ότι ενώ αναφερόντουσαν οι 3 όροι ενδιαφέροντος που υποδηλώνουν συσχέτιση με την ασθένεια που ερευνούμε κάποιος από αυτούς δεν ανήκε στη συγκεκριμένη τριπλέτα ή αναφερόταν ως απλή αναφορά προηγούμενης έρευνας οπότε δεν μπορούσε να διεξαχθεί εμπεριστατωμένο συμπέρασμα για την ύπαρξη σχέσης γιατί τα στοιχεία ήταν ελλιπή. Εν κατακλείδι διαπιστώσαμε ότι 33 microRNAs που αναφέρονται στον τελικό κατάλογο έχουν σημαντική συσχέτιση με την ασθενείς που μελετάμε τον καρκίνο του πνεύμονα , επίσης βρήκαμε 26 γονίδια που έχουν άμεση συσχέτιση με τον καρκίνο του πνεύμονα και 42 μεταλλάξεις προστέθηκαν στον κατάλογο του χειροκίνητου σχολιασμού ως άρρηκτα συνδεδεμένες με τον καρκίνου του πνεύμονα.



Εικόνα 3 Τα άρθρα με PMCID που είχαν διαθέσιμο κείμενο

Εικόνα 4Τα άρθρα που πέρασαν τα τελικά φίλτρα

Εικόνα 5 Δημοσιεύσεις που πέρασαν τον χειροκίνητο σχολιασμό.



Εικόνα 6 Διάγραμμα ροής του αριθμού των άρθρων που επιλέχθηκαν σε κάθε βήμα.

4.ΣΥΖΗΤΗΣΗ

Όπως αναφέραμε και παραπάνω τα microRNAs διαδραματίζουν σημαντικό ρόλο στη γονιδιακή ρύθμιση και για αυτό ερευνώνται ως βιοδείκτες και πιθανοί θεραπευτικοί στόχοι σε διάφορες μορφές καρκίνου. Συνεχώς εμφανίζονται νέες έρευνες που συσχετίζουν μεταλλάξεις που αφορούν την αλληλεπίδραση miRNA::mRNA με τις διαφορετικές μορφές του καρκίνου του πνεύμονα. Οποιαδήποτε μελέτη πάνω στην επίδραση των μεταλλάξεων στο microRNA είναι χρήσιμη για να κατανοήσουμε πλήρως τη δράση τους με σκοπό την έγκαιρη διάγνωση και θεραπεία της ασθένειας.

Όσον αφορά την ανάλυση και την επεξεργασία κειμένου που εφαρμόστηκε στην παρούσα μελέτη επιδέχεται βελτιώσεις ώστε να γίνει πιο αποδοτική όσον αφορά τη συλλογή της κατάλληλης πληροφορίας. Διερευνητικά θα ήταν θεμιτό να αναλυθούν τα αποτελέσματα και από άλλα μοντέλα state-of-the-art όπως το stanza[15] και το BERT για την αναζήτηση ονομάτων-οντοτήτων [16].

Επόμενο στάδιο θα ήταν η εύρεση σχέσης μεταξύ των οντοτήτων, δηλαδή η αναζήτηση λέξεων κλειδιών που να αποδεικνύουν τη σχέση μεταξύ των οντοτήτων. Επίσης η χειροκίνητη εργασία έδειξε ότι τα κεφάλαια ενός άρθρου που δίνουν την περισσότερη πληροφορία είναι τα αποτελέσματα και οι μεθοδολογίες. Ενώ συνηθίζεται η εισαγωγή του άρθρου να περιέχει πληροφορίες για προηγούμενες μελέτες ως αναφορές. Επομένως ένα μοντέλο αναγνώρισης των κεφαλαίων και στόχευσης για ανάλυση μόνο στα σημεία ενδιαφέροντος θα έδινε μεγαλύτερη ακρίβεια. Παράλληλα θα μπορούσαν να χρησιμοποιηθούν για την επεξεργασία των αποτελεσμάτων που κατέληξαν στον κατάλογο in silico υπολογιστικοί αλγόριθμοι πρόβλεψης στόχων όπως ο αλγόριθμος Targetscan [17] και ο MicroT-CDS [14].

Στην παρούσα εργασία για να φτάσουμε στην ύπαρξη σχέσης μεταξύ των miRNA::mRNA αναπτύξαμε ένα σύστημα αυτοματοποιημένης αναζήτησης σε βάσεις με πληθώρα βιοϊατρικών δεδομένων, ενώ στη συνέχεια προχωρήσαμε στην ανάλυση ονομάτων-οντοτήτων των αποτελεσμάτων αυτών. Για την ακριβέστερη διαλογή, τα αποτελέσματα φιλτραρίστηκαν αλγοριθμικά και σχολιάστηκαν χειροκίνητα. Οι προτάσεις για βελτίωση που παρατίθενται παραπάνω έχουν ως στόχο να συνδράμουν στη καλύτερη δυνατή συγκέντρωση της πληροφορίας και τους τρόπους απομόνωσης της. Προκειμένου να δώσουν στην επιστημονική κοινότητα τρόπους και εργαλεία ώστε να γίνει ενδελεχής μελέτη των microRNAs και της σχέσης τους με τις διάφορες μορφές του καρκίνου. Φυσικά το παραπάνω σύστημα δίνει τη δυνατότητα έρευνας και μελέτης για διάφορες ασθένειες πέραν του καρκίνου που υπάρχει υποψία ότι σχετίζονται με μεταλλάξεις των microRNAs και να υπάρξουν νέα ευρήματα προς μελέτη.

ΠΑΡΑΡΤΗΜΑ

pmcid	miRNA	Gene	Association	Risk	Variant	Gene Biotype	Variant region	Disease
9046731	mir-126	EGFL7	TRUE	increased	rs2297538	protein coding	EGFL7 : Missense Variant	lung cancer
9046731	mir-126	EGFL7	TRUE	increased	rs2297538	protein coding	EGFL7 : Missense Variant	lung cancer
9046731	miR-126	EGFL7	TRUE	increased	rs2297538	protein coding	EGFL7 : Missense Variant	lung cancer
7773458	Mir-608	TFAP4	TRUE	__	rs4919510	protein coding	MIR608 : Non Coding Transcript Variant	lung cancer
7751538	let-7	KRAS	TRUE	increased	rs61764370	protein coding	KRAS : 3 Prime UTR Variant	lung cancer
7427888	mir-539-5p	ENO1	TRUE	increased	rs140618127	protein coding	ARHGAP27P1-BPTFP1-KPNA2P3 : Non Coding Transcript Variant	lung cancer
7191703	hsa-mir-1197	RIPK1	TRUE	increased	rs17548629	protein coding	RIPK1 : 3 Prime UTR Variant	lung cancer
7191703	hsa-mir-1197	RIPK1	FALSE	__	rs77736895	protein coding	RIPK1 : 3 Prime UTR Variant	lung cancer
6325744	hsa-mir-144-3p and has-mir-182-5p	LPP	TRUE	increased	rs1064607	protein coding	LPP : Non Coding Transcript Variant	lung cancer
6325744	hsa-mir-144-3p and has-mir-182-5p	LPP	TRUE	increased	rs3796283	protein coding	LPP : Non Coding Transcript Variant	lung cancer
6325744	hsa-mir-144-3p and has-mir-182-5p	LPP	TRUE	increased	rs2378456	protein coding		lung cancer
6325744	hsa-mir-144-3p and has-mir-182-5p	NR5A2	__	__	rs2246209	protein coding	NR5A2 : 3 Prime UTR Variant	lung cancer
6325744	hsa-mir-144-3p and has-mir-182-5p	NR5A2	TRUE	__	rs1056426	protein coding	NR5A2 : 3 Prime UTR Variant	lung cancer
6325744	hsa-mir-144-3p and has-mir-182-5p	TNS3	__	__	rs3750163	protein coding	TNS3 : 3 Prime UTR Variant	lung cancer
6325744	hsa-mir-144-3p and has-mir-182-5p	TNS3	TRUE	__	rs9876	protein coding	TNS3 : 3 Prime UTR Variant	lung cancer
5963629	hsa-mir-29	SLC31A1	__		rs4979223	protein coding	SLC31A1 : 2KB Upstream Variant	lung cancer
5963629	hsa-mir-29	SLC31A1	__		rs4978536	protein coding	SLC31A1 : 2KB Upstream Variant	lung cancer
5963629	hsa-mir-29	SLC31A1	__		rs2233914	protein coding	SLC31A1 : 2KB Upstream Variant	lung cancer

5963629	hsa-mir-29	SLC31A1	__		rs10817464	protein coding	SLC31A1 : Intron Variant	lung cancer
5963629	hsa-mir-29	SLC31A1	__		rs10981699	protein coding	SLC31A1 : Intron Variant	lung cancer
5963629	hsa-mir-29	SLC31A1	__		rs10817465	protein coding	SLC31A1 : Intron Variant	lung cancer
5963629	hsa-mir-29	SLC31A1	__		rs10513102	protein coding	CLSTN2 : Intron Variant	lung cancer
5963629	hsa-mir-29	SLC31A1	TRUE	___	rs10759637	protein coding	SLC31A1 : 3 Prime UTR Variant	lung cancer
5668488	hsa-miR-652	POLR2A	TRUE	___	rs2071504	protein coding	POLR2A : Synonymous Variant	lung cancer
5668488	has-miR-196a	NR2F6	TRUE	___	rs2288539	protein coding	NR2F6 : Synonymous Varian	lung cancer
5668488	has-let-7b	ADCK2	___	___	rs1140034	protein coding	ADCK2 : Missense Variant	lung cancer
5668488	hsa-miR-92a	ACADS	___	___	rs2229534	protein coding	ACADS : 3 Prime UTR Variant	lung cancer
5668488	hsa-miR-92a	CD3EAP	___	___	rs3212986	protein coding	POLR1G : Stop Gained	lung cancer
5567228	miR-204	FZD4	TRUE	___	rs713065	protein coding	FZD4 : 3 Prime UTR Variant	lung cancer
5547930	miR-324-5p, miR-296-5p	B7-H1	TRUE	increased	rs2297136	protein coding	rs2297136	lung cancer
5547930	miR-570, miR-7-1 , miR-495,miR-298	B7-H1	FALSE	___	rs4143815	protein coding	CD274 : Non Coding Transcript Variant	lung cancer
5547930	mir-138	B7-H1	TRUE	increased	rs4742098	protein coding	CD274 : Non Coding Transcript Variant	lung cancer
5834125	miR-29a	IREB2	TRUE	increased	rs1062980	protein coding	IREB2 : 3 Prime UTR Variant	lung cancer

5356798	miR-146a	TRAF6	TRUE	decreased	rs2910164	protein coding	MIR146A : Non Coding Transcript Variant	lung cancer
5347756	miR-149	TOP1	TRUE	___	rs2292832	protein coding	MIR149 : Non Coding Transcript Variant	lung cancer
5369965	mir-887-3p	MDM4	TRUE	___	rs4245739	protein coding	MDM4 : 3 Prime UTR Variant	lung cancer
5369965	mir-3173	MDM4	TRUE	___	rs10900598	protein coding	MDM4 : 3 Prime UTR Variant	lung cancer
4544728	let-7g	KRAS	TRUE	increased	rs712	protein coding	KRAS : 3 Prime UTR Variant	lung cancer
4315715	mir-516a-3p	CXCR2	TRUE	decreased	rs1126579	protein coding	CXCR2 : 3 Prime UTR Variant	lung cancer
4315715	mir-516a-3p	CXCR2	TRUE	decreased	rs1126579	protein coding	CXCR2 : 3 Prime UTR Variant	lung cancer
3819265	pre-mir-27a	HOXA10		increased	rs895819	Protein coding	MIR27A : Non Coding Transcript Variant	lung cancer
3795636	mir-502	SET8	TRUE	decreased	rs16917496	protein coding	KMT5A : Non Coding Transcript Variant	lung cancer
3438626	mir-502	SET8	FALSE	___	rs16917496	protein coding	KMT5A : Non Coding Transcript Variant	lung cancer
3416793	mir-126	PI3K	FALSE	___	rs4636297	protein coding	MIR126 : 500B Downstream Variant	lung cancer

Disease	Drug	Drug response	Population	Cells
lung cancer	___	__	Han-Chinese	H1299
lung cancer	___	__	Han-Chinese	BEAS-2B
lung cancer	___	__	Han-Chinese	SPC-A1
lung cancer	DOX	positive	__	A549
lung cancer	___	___	Iranian	__
lung cancer	___	___	Chinese	A549,PC9,BEAS2B
lung cancer	___	___	Han-Chinese	A549,293T
lung cancer	___	___	han-Chinese	A549,293T
lung cancer	___	___	Han-Chinese	__
lung cancer	___	___	Han-Chinese	__
lung cancer	___	___	Han-Chinese	__
lung cancer	___	___	Han-Chinese	__
lung cancer	___	___	Han-Chinese	__
lung cancer	___	___	Han-Chinese	__
lung cancer	___	___	Han-Chinese	__
lung cancer	platinum based chemotherapy		Chinese	__
lung cancer	platinum based chemotherapy		Chinese	__
lung cancer	platinum based chemotherapy		Chinese	__
lung cancer	platinum based chemotherapy		Chinese	__
lung cancer	platinum based chemotherapy		Chinese	__
	__	___	Chinese	16HBE
lung cancer	platinum based chemotherapy		Chinese	__
	__	___	Chinese	16HBE
lung cancer	platinum based chemotherapy		Chinese	__
lung cancer	platinum based chemotherapy		Chinese	16HBE
lung cancer	18F-Fluorodeoxyglucose	___	__	A549
lung cancer	___	___	Korean	__
lung cancer	___	___	Korean	__
lung cancer	___	___	Korean	__
lung cancer	___	___	Korean	__

lung cancer	___	___	Korean	__
lung cancer	___	___	__	H1299,H322
lung cancer	___	___	__	A549
lung cancer	___	___	__	A549
lung cancer	___	___	__	A549
lung cancer	___	___	___	__
lung cancer	___	___	Chinese	A549
lung cancer	___	___	___	A549
lung cancer	platinum based chemotherapy	positive	___	A549
lung cancer	platinum based chemotherapy	positive	___	A549
lung cancer	___	___	___	__
lung cancer	___	___	Japanese	__
lung cancer	___	___	European American	
lung cancer	platinum based chemotherapy	negative	Chinese	__
lung cancer	___	___	Chinese	A549
lung cancer	___	___	___	__
lung cancer	___	___	___	A549
lung cancer	taxanes:paclitaxel and docetaxel	positive	___	H196

ΚΩΔΙΚΑΣ

Υλοποιήση της ατοματοποιημένης αναζήτησης του ερωτήματος (query) στην Entrez μέσω της esearch():

```
from Bio import Entrez
from bs4 import BeautifulSoup
from bs4 import BeautifulSoup as bs
import re
import scispacy
import spacy
import nltk
import en_core_sci_sm
import en_ner_bionlp13cg_md




Entrez.email = "iwanna.vav@gmail.com"
handle = Entrez.esearch(db="pubmed", retmax=1000000, term='"lung
cancer"[Title/Abstract] AND (mutation[Title/Abstract] '
                                                        'OR
polymorphism[Title/Abstract] OR variant[Title/Abstract]'
                                                        ' OR
snp[Title/Abstract]) '
                                                        'AND (
microrna[Title/ABSTRACT] OR mirna[Title/Abstract]) '
                                                        'NOT Review[ptyp]')

record = Entrez.read(handle)
r = record["IdList"]
print(r)
print(len(r))


handle = Entrez.efetch(db="pubmed", id=r, retmax=1000000,retmode= "xml",
rettype= "")

with open('filenew.xml', 'wb') as f:
    f.write(handle.read())

with open('filenew.xml', 'r') as file:
    file_contents = file.read()
    #soup = BeautifulSoup(file, 'lxml')
soup = BeautifulSoup(file_contents,'xml')
articles = soup.findAll('Article')
for article in articles:
    print(article.get_text(separator= " "))


pmcs = soup.findAll('ArticleId', {'IdType': 'pmc'})
pmcid_list = []
for pmc in pmcs:
    print(pmc.get_text())
    pmcid_list.append(pmc.get_text())
print(len(pmcid_list))

handle = Entrez.efetch(db= "pmc",resetmode = 'xml', id=pmcid_list,
rettype="full")
with open('filenew1.xml', 'wb') as f:
    f.write(handle.read())
```

```
with open('filenew1.xml', 'r') as file:
    soup = BeautifulSoup(file, 'lxml')
all = soup.findAll('article')
```

Name Entity Recognition(NER)

```
import pandas as pd
from bs4 import BeautifulSoup

import spacy
import scispacy
import nltk
nltk.download('punkt')

import en_ner_bionlp13cg_md
nlp_bi = en_ner_bionlp13cg_md.load()



nlp_core = spacy.load("en_core_sci_sm")

f = open('sentence_seg.csv','w')
fi = open('entities_seg.csv','w')


with open('filenew1.xml', 'r') as file:
    soup = BeautifulSoup(file, 'lxml')
all = soup.findAll('article')

table = {"ID": [], "Entity": [], "Class": []}
meta = { "PMCID": [], "TITLE": [], "SENTENCE": [], "DOCENTS": []}

for article in all:
    pmcid = article.find("article-id", attrs={"pub-id-type": "pmc"}).text
    article_title = article.find("article-title").text
    paragraph_list = article.findAll('p')
    caption_list = article.findAll('caption')


    for paragraph in paragraph_list:

        p = paragraph.text

        tokens = nltk.sent_tokenize(p)
        for token in tokens:
            doc = nlp_core(token)
            meta["SENTENCE"].append(token)
            meta["DOCENTS"].append(doc.ents)
            meta["PMCID"].append(pmcid)
            meta["TITLE"].append(article_title)

        doc = nlp_bi(p)
        ent = {}
        for x in doc.ents:
            ent[x.text] = x.label_
        for k in ent:
            table["ID"].append(pmcid)
            table["Entity"].append(k)
            table["Class"].append(ent[k])

    for caption in caption_list:
```

```
        cap = caption.text
        tokens = nltk.sent_tokenize(cap)
        for token in tokens:
            doc = nlp_core(token)
            meta["SENTENCE"].append(token)
            meta["DOCENTS"].append(doc.ents)
            meta["PMCID"].append(pmcid)
            meta["TITLE"].append(article_title)

        doc = nlp_bi(cap)
        ent = {}
        for x in doc.ents:
            ent[x.text] = x.label_
        for k in ent:
            table["ID"].append(pmcid)
            table["Entity"].append(k)
            table["Class"].append(ent[k])

trans_df = pd.DataFrame(meta)
trans_df.to_csv('sentence_seg.csv', index=False)


trans_df = pd.DataFrame(table)
trans_df.to_csv('entities_seg.csv', index=False)
```

Φίλτρα αναζήτησης:

```
import csv
import re
import sys


def clear_empty_lines():
    output = ""
    with open("finalresults.txt",encoding='utf-8') as f:
        for line in f:
            if line.startswith("ID"):
                output += "\n" + line
            elif not line.isspace():
                output += line

        f = open("finalresults.txt", "w",encoding='utf-8')
        f.write(output)


results = open("finalresults.txt", "w+",encoding='utf-8')
file = open('entities_seg.csv',encoding='utf-8')
type(file)

csvreader = csv.reader(file)

rows = []
for row in csvreader:
        rows.append(row)


meta = open('sentence_seg.csv',encoding='utf-8')

type(meta)

csvreader_meta = csv.reader(meta)
rows_meta = []
for row_meta in csvreader_meta:
        rows_meta.append(row_meta)
```

```
id_init= rows[1][0]
id_list = []
flag1 = False
flag2 = False
flag3 = False
flag4 = False
listtest = []

for ele in range(1, len(rows)):
    print(rows[ele][1])

    if id_init != rows[ele][0]:
        id_init = rows[ele][0]



    if (rows[ele][2] == "GENE_OR_GENE_PRODUCT") and (re.search("mir|let-",
rows[ele][1].lower()) != None):
        flag4 = True


    if (rows[ele][2] == "CANCER") and (re.search(
            "lung|canc|neop",
            rows[ele][1].lower().replace(" ", "")) != None):
        flag3 = True


    with open('hgnc-search-1666005137067.txt',encoding='utf-8') as infile:

        for line in infile:
            if (rows[ele][2] == "GENE_OR_GENE_PRODUCT") and
((line.split()[1]).find(rows[ele][1]) >= 0):
                print(line.split()[1])
                print(rows[ele][1])
                flag2 = True



    if (flag4 == True) and (flag3 == True)  and(flag2==True) :
        id_list.append(rows[ele][0])
        flag1 = False
        flag2 = False
        flag3 = False
        flag4 = False


print(id_list)
pr = False
id_init_meta = rows_meta[1][0]
printlist = []

for ele_meta in range(1, len(rows_meta)):

    counter = 0
    flag_RS = False
    flag_MIR = False
    flag_DISEASE = False
    flag_GENE = False

    if rows_meta[ele_meta][0] in id_list:
        if id_init_meta != rows_meta[ele_meta][0]:
            id_init_meta = rows_meta[ele_meta][0]
            pr = False

        if pr == False:
```

```
            printlist.append(rows_meta[ele_meta][0])
            results.write("\n\n")
            pr = True

        a = str(rows_meta[ele_meta][3]).replace("(", "").replace(")",
"").replace(" ", "")
        lista = a.split(",")
        for element in lista:

            if (re.search("rs[0-9]+", element) != None) and (flag_RS ==
False):
                counter += 1

                flag_RS = True

            if (re.search("mir|let-", element.lower()) != None) and (flag_MIR
== False):
                counter += 1

                flag_MIR = True


            if (re.search(
                    "lung|canc|neop",
                    element.lower().replace(" ", "")) != None) and
(flag_DISEASE == False):
                counter += 1

                flag_DISEASE = True


            with open('hgnc-search-1666005137067.txt',encoding='utf-8') as
infile:
                for line in infile:
                    flag_GENE = False
                    if ((line.split()[1]).find(element) >= 0) and flag_GENE
== True:
                        print(line.split()[1])
                        print(element)
                        counter += 1
                        flag_GENE = True



        if (counter >= 2 and flag_RS == True) :
            if len(printlist) > 0:
                results.write("ID:" + rows_meta[ele_meta][0] + " " + "Title:"
+ rows_meta[ele_meta][1] + "\n")
                results.write(rows_meta[ele_meta][2] + "\n")
                printlist.clear()
            else:
                results.write(rows_meta[ele_meta][2])
            flag_RS = False
            flag_MIR = False
            flag_DISEASE = False
            flag_GENE = False

results.close()
print(len(listtest))

clear_empty_lines()
```

Βιβλιογραφία

[1] Siddiqui F, Vaqar S, Siddiqui AH. Lung Cancer. 2022 Dec 5. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan–. PMID: 29493979.

[2] Hammerschmidt S, Wirtz H. Lung cancer: current diagnosis and treatment. Dtsch Arztebl Int. 2009 Dec;106(49):809-18; quiz 819-20. doi: 10.3238/arztebl.2009.0809. Epub 2009 Dec 4. PMID: 20038979; PMCID: PMC2797332.

[3[, Wirtz H. Lung cancer: current diagnosis and treatment. Dtsch Arztebl Int. 2009 Dec;106(49):809-18; quiz 819-20. doi: 10.3238/arztebl.2009.0809. Epub 2009 Dec 4. PMID: 20038979; PMCID: PMC2797332.

[4] Lababede O, Meziane MA. The Eighth Edition of TNM Staging of Lung Cancer: Reference Chart and Diagrams. Oncologist. 2018 Jul;23(7):844-848. doi: 10.1634/theoncologist.2017-0659. Epub 2018 Apr 12. PMID: 29650687; PMCID: PMC6058324.

[5] de Groot PM, Wu CC, Carter BW, Munden RF. The epidemiology of lung cancer. Transl Lung Cancer Res. 2018 Jun;7(3):220-233. doi: 10.21037/tlcr.2018.05.06. PMID: 30050761; PMCID: PMC6037963.

[6] Shankar A, Dubey A, Saini D, Singh M, Prasad CP, Roy S, Bharati SJ, Rinki M, Singh N, Seth T, Khanna M, Sethi N, Kumar S, Sirohi B, Mohan A, Guleria R, Rath GK. Environmental and occupational determinants of lung cancer. Transl Lung Cancer Res. 2019 May;8(Suppl 1):S31-S49. doi: 10.21037/tlcr.2019.03.05. PMID: 31211104; PMCID: PMC6546634.

[7] Zarredar H, Ansarin K, Baradaran B, Shekari N, Eyvazi S, Safari F, Farajnia S. Critical microRNAs in Lung Cancer: Recent Advances and Potential Applications. Anticancer Agents Med Chem. 2018;18(14):1991-2005. doi: 10.2174/1871520618666180808125459. PMID: 30088452.

[8] de Groot PM, Wu CC, Carter BW, Munden RF. The epidemiology of lung cancer. Transl Lung Cancer Res. 2018 Jun;7(3):220-233. doi: 10.21037/tlcr.2018.05.06. PMID: 30050761; PMCID: PMC6037963.

[9] Feng H, Wang X, Zhang Z, Tang C, Ye H, Jones L, Lou F, Zhang D, Jiang S, Sun H, Dong H, Zhang G, Liu Z, Dong Z, Guo B, Yan H, Yan C, Wang L, Su Z, Li Y, Nandakumar V, Huang XF, Chen SY, Liu D. Identification of Genetic Mutations in Human Lung Cancer by Targeted Sequencing. Cancer Inform. 2015 Jun 29;14:83-93. doi: 10.4137/CIN.S22941. PMID: 26244006; PMCID: PMC4489668.

[10] Hammond SM. An overview of microRNAs. Adv Drug Deliv Rev. 2015 Jun 29;87:3-14. doi: 10.1016/j.addr.2015.05.001. Epub 2015 May 12. PMID: 25979468; PMCID: PMC4504744.

[11] Ha, M., Kim, V. Regulation of microRNA biogenesis. *Nat Rev Mol Cell Biol* 15, 509–524 (2014). https://doi.org/10.1038/nrm3838.

[12] Cannell IG, Kong YW, Bushell M. How do microRNAs regulate gene expression? Biochem Soc Trans. 2008 Dec;36(Pt 6):1224-31. doi: 10.1042/BST0361224. PMID: 19021530.

[13] Du X, Zhang J, Wang J, Lin X, Ding F. Role of miRNA in Lung Cancer-Potential Biomarkers and Therapies. Curr Pharm Des. 2018 Feb 12;23(39):5997-6010. doi: 10.2174/1381612823666170714150118. PMID: 28714414.

[14] Paraskevopoulou MD, Georgakilas G, Kostoulas N, Vlachos IS, Vergoulis T, Reczko M, Filippidis C, Dalamagas T, Hatzigeorgiou AG. DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W169-73. doi: 10.1093/nar/gkt393. Epub 2013 May 16. PMID: 23680784; PMCID: PMC3692048.

[15] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 101–108, Online. Association for Computational Linguistics.

[16] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. https://doi.org/10.48550/arXiv.1810.04805

[17] Witkos TM, Koscianska E, Krzyzosiak WJ. Practical Aspects of microRNA Target Prediction. Curr Mol Med. 2011 Mar;11(2):93-109. doi: 10.2174/156652411794859250. PMID: 21342132; PMCID: PMC3182075.

[18] O'Brien J, Hayder H, Zayed Y, Peng C. Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. Front Endocrinol (Lausanne). 2018 Aug 3;9:402. doi: 10.3389/fendo.2018.00402. PMID: 30123182; PMCID: PMC6085463.