



Πέτρος Κυρές
Ανάλυση Ελεύθερων Δεδομένων Γονιδιακής Έκφρασης
στο Ακανθοκυτταρικό Καρκίνωμα του Στόματος.



ΕΛΛΗΝΙΚΟ
ΑΝΟΙΚΤΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ



ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Σχολή Θετικών Επιστημών και
Τεχνολογίας

Τμήμα Πληροφορικής

Βιοπληροφορική και Νευροπληροφορική

Μεταπτυχιακή Διπλωματική Εργασία

**Ανάλυση Ελεύθερων Δεδομένων Γονιδιακής Έκφρασης στο
Ακανθοκυτταρικό Καρκίνωμα του Στόματος**

Πέτρος Κυρές

Επιβλέπων Καθηγητής: Λευτέρης Κουμάκης

Πάτρα, Ιούνιος 2021

© 2021 All rights reserved

Η παρούσα εργασία αποτελεί πνευματική ιδιοκτησία του/της φοιτητή («συγγραφέας/δημιουργός») που την εκπόνησε. Στο πλαίσιο της πολιτικής ανοικτής πρόσβασης ο/η συγγραφέας/δημιουργός εκχωρεί στο ΕΑΠ, μη αποκλειστική άδεια χρήσης του δικαιώματος αναπαραγωγής, προσαρμογής, δημόσιου δανεισμού, παρουσίασης στο κοινό και ψηφιακής διάχυσής τους διεθνώς, σε ηλεκτρονική μορφή και σε οποιοδήποτε μέσο, για διδακτικούς και ερευνητικούς σκοπούς, άνευ ανταλλάγματος και για όλο το χρόνο διάρκειας των δικαιωμάτων πνευματικής ιδιοκτησίας. Η ανοικτή πρόσβαση στο πλήρες κείμενο για μελέτη και ανάγνωση δεν σημαίνει καθ' οιονδήποτε τρόπο παραχώρηση δικαιωμάτων διανοητικής ιδιοκτησίας του/της συγγραφέα/δημιουργού ούτε επιτρέπει την αναπαραγωγή, αναδημοσίευση, αντιγραφή, αποθήκευση, πώληση, εμπορική χρήση, μετάδοση, διανομή, έκδοση, εκτέλεση, «μεταφόρτωση» (downloading), «ανάρτηση» (uploading), μετάφραση, τροποποίηση με οποιονδήποτε τρόπο, τμηματικά ή περιληπτικά της εργασίας, χωρίς τη ρητή προηγούμενη έγγραφη συναίνεση του/της συγγραφέα/δημιουργού. Ο/Η συγγραφέας/δημιουργός διατηρεί το σύνολο των ηθικών και περιουσιακών του δικαιωμάτων.



Πέτρος Κυρές
Ανάλυση Ελεύθερων Δεδομένων Γονιδιακής Έκφρασης
στο Ακανθοκυτταρικό Καρκίνωμα του Στόματος.



ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ



ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

Ανάλυση Ελεύθερων Δεδομένων Γονιδιακής Έκφρασης στο Ακανθοκυτταρικό Καρκίνωμα του Στόματος.

Πέτρος Κυρές

Επιτροπή Κρίσης

Επιβλέπων Καθηγητής

Συν-Επιβλέπουσα Καθηγήτρια

Συν-Επιβλέπουσα Καθηγήτρια

Λευτέρης Κουμάκης

Σουλτάνα Μαρκοπούλου

Μαρία Χατζηνικολάου
Καθηγήτρια

Μέλος ΣΕΠ

Μέλος ΣΕΠ

Ελληνικό Ανοικτό
πανεπιστήμιο

Ελληνικό Ανοικτό
Πανεπιστήμιο

Ελληνικό Ανοικτό
Πανεπιστήμιο

Πάτρα , Ιούνιος 2021

ΠΕΡΙΛΗΨΗ

Το ακανθοκυτταρικό καρκίνωμα αποτελεί τον συχνότερο ιστολογικό τύπο καρκίνου του στόματος. Στην σύγχρονη εποχή η μελέτη της γονιδιακής έκφρασης γίνεται με μεθόδους υψηλής ανάλυσης όπως είναι οι μικροσυστοιχίες και η αλληλούχιση RNA, όμως δυστυχώς η μετα-ανάλυση δεδομένων από τέτοια πειράματα είναι δύσκολη και υπάρχουν πολλές διαφορετικές προσεγγίσεις στην αντιμετώπιση αυτού του προβλήματος χωρίς καμία να θεωρείται περισσότερο αξιόπιστη από την άλλη.

Σκοπός αυτής της εργασίας ήταν η σκιαγράφηση του προφίλ γονιδιακής έκφρασης στο ακανθοκυτταρικό καρκίνωμα του στόματος όπως προκύπτει από ελεύθερα διαθέσιμα δεδομένα από μικροσυστοιχίες στην βάση δεδομένων GEO. Χρησιμοποιήθηκαν δύο προσεγγίσεις με βάση την ζ -κανονικοποίηση και μία νέα προτεινόμενη μέθοδος μετα-ανάλυσης πειραμάτων μικροσυστοιχιών που την ονομάσαμε διασπαστική μέθοδο. Έγινε και στις τρεις παραπάνω προσεγγίσεις, αναζήτηση διαφορικά εκφραζόμενων γονιδίων με τρεις στατιστικές τεχνικές (Student's t-test, Moderated t-test και Significance Analysis of Microarrays - SAM) και λειτουργική ανάλυση εμπλουτισμού τους σε κατηγορίες της Kegg. Ακολούθησε στην συνέχεια αξιολόγηση των συνολικά τριών μεθόδων ως προς την παραμόρφωση των δεδομένων με μία προσέγγιση επιβλεπόμενης και μη-επιβλεπόμενης ανάλυσης αλλά και με βάση τα αποτελέσματα της λειτουργικής τους ανάλυσης. Επίσης έγινε και περαιτέρω ανάλυση κλάσεων για όσα πειράματα μικροσυστοιχιών είχαν επαρκή μεταδεδομένα με την χρήση ANOVA. Από τις τρεις παραπάνω προσεγγίσεις επιλέχτηκε με βάση τα αποτελέσματα της αξιολόγησης, η διασπαστική τεχνική, με τα αποτελέσματα της οποίας έγινε περαιτέρω διερεύνηση με ανάλυση κύριων συνιστωσών, μεθόδους μηχανικής μάθησης που βασίζονται σε δέντρα και δάση, χαρτογράφηση των διαφορικά εκφραζόμενων γονιδίων σε χάρτες βιολογικών μονοπατιών της Kegg και τέλος ανάλυση τοπολογίας δικτύου για τα υπερεκφραζόμενα γονίδια.

Η διασπαστική τεχνική απέδωσε καλύτερα σε σχέση με τις προσεγγίσεις με ζ -κανονικοποίηση. Η εύρεση διαφορικά εκφραζόμενων γονιδίων με τις μεθόδους t-test και moderated t-test έδωσαν ταυτόσημα αποτελέσματα ενώ η μέθοδος SAM φαίνεται είχε χειρότερη απόδοση με πολλά ψευδώς θετικά ευρήματα. Ο αλγόριθμος C-tree για την δημιουργία δένδρων απόφασης είχε καλύτερη απόδοση από τον αλγόριθμο CART, αλλά

χρησιμοποιεί πολύ περισσότερα γονίδια. Η έκφραση της MMP1 φαίνεται να είναι καθοριστική στην απόφαση αν ένα δείγμα είναι καρκινικό ή όχι. Από πλευράς τοπολογίας δικτύων ωστόσο η έκφραση των γονιδιών για τα STAT1, MMP9, CXCL1, CXCL10, CDK1 και την οστεονεκτίνη (SPARC) έχουν τον πιο κομβικό ρόλο ανάμεσα στα υπερεκφραζόμενα γονίδια. Τέλος φαίνεται ότι η κλινική σταδιοποίηση ενός ακανθοκυτταρικού καρκινώματος δεν μπορεί να προσδιοριστεί ή να προβλεφθεί βάσει του προφίλ γονιδιακής έκφρασης του.

Τα αποτελέσματα που εξάχθηκαν σε γενικές γραμμές συμφωνούν με τα δεδομένα από την βιβλιογραφία και δείχνουν νέες κατευθύνσεις στις οποίες θα πρέπει να δωθεί περισσότερη έμφαση σε μελλοντικές ερευνητικές προσπάθειες σχετικά με το ακανθοκυτταρικό καρκίνωμα του στόματος.

Λέξεις κλειδιά : μετα-ανάλυση, ακανθοκυτταρικό καρκίνωμα, μικροσυστοιχίες, μηχανική μάθηση, βιολογικά δίκτυα, γονιδιακή έκφραση

ABSTRACT

Squamous cell carcinoma is the most frequent histological diagnosis in oral cancer. Nowadays, research regarding gene expression is accomplished with high throughput methods and techniques like microarrays and RNA sequencing. However, meta-analysis of data extracted by those kinds of experiments is difficult and although there are many approaches for facing this problem, none of them is considered to be more reliable than others.

The purpose of this thesis was to outline the gene expression profile of oral squamous cell carcinoma using publicly available data from GEO database. We utilized two z-transformation approaches plus a new proposed microarray meta-analysis method which in the context of this thesis was named as “disintegrative” method. Using those 3 approaches, extraction of the differentially expressed genes was performed based on three popular statistical methods (Student’s t-test, Moderated t-test and Significance analysis of Microarrays –SAM) and after that a differential genes functional analysis of enrichment in Kegg categories was done. This was followed by a validation of the aforementioned approaches for data manipulation based on a combination of unsupervised and supervised learning techniques but also taking into consideration the results of differential genes functional analysis. Further ANOVA class analysis was done using only datasets with adequate metadata available. Taking these initial results into consideration, the best performing meta-analysis technique was selected (which was proven to be the “disintegrative” approach) for further more advanced analysis using PCA data exploration, tree and forest-based machine learning methods, mapping of results on Kegg biological pathways and last but not least network topology analysis for the differentially overexpressed genes.

Disintegrative technique performed better compared to z-scaling based methods. Differentially expressed genes extracted with either Student’s t-test and Moderated t-test were identical, but the SAM technique appeared to perform worse with many false positive results. Decision tree algorithm C-tree had a better approach compared to CART but it is taking into consideration many more gene expressions. MMP1 expression seems to be the

most important parameter in deciding whether a sample is cancerous or not. Network topology analysis however uncovered the critical role of STAT1, MMP9, CXCL1, CXCL10, CDK1 and osteonectin (SPARC) among the differentially overexpressed genes. Lastly, it seems that clinical staging of an oral squamous cell carcinoma tumour, cannot be determined or predicted based solely on its gene expression profile.

The results of this research thesis in general terms appear to be in accordance with findings from the scientific literature and they propose potential new directions for future studies in oral squamous cell carcinoma.

Keywords: meta-analysis, squamous cell carcinoma, microarrays, machine learning, biological networks, gene expression

Πίνακας Περιεχομένων

ΠΕΡΙΛΗΨΗ	iv
ABSTRACT	VI
ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ	VIII
ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ	XI
ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ	XV
ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ ΚΑΙ ΑΚΡΩΝΥΜΙΑ.....	XVI

I. ΓΕΝΙΚΟ ΜΕΡΟΣ

ΚΕΦΑΛΑΙΟ 1. ΙΣΤΟΛΟΓΙΑ & ΚΥΤΤΑΡΙΚΗ ΒΙΟΛΟΓΙΑ ΤΟΥ ΦΥΣΙΟΛΟΓΙΚΟΥ ΣΤΟΜΑΤΙΚΟΥ ΒΛΕΝΟΓΟΝΝΟΥ	2
1.1 Εισαγωγικά Στοιχεία	2
1.2 Το Στοματικό Καλυπτικό Επιθήλιο	4
1.3 Η Διασύνδεση Επιθηλίου-Συνδετικού Ιστού	9
ΚΕΦΑΛΑΙΟ 2. ΚΥΤΤΑΡΙΚΟΣ ΚΥΚΛΟΣ ΚΑΙ ΑΡΧΕΣ ΠΑΘΟΒΙΟΛΟΓΙΑΣ ΚΑΡΚΙΝΟΥ.....	12
2.1 Κυτταρικός Κύκλος	12
2.2 Προγραμματισμένος Κυτταρικός Θάνατος	16
2.3 Αρχές Παθοβιολογίας Καρκίνου.....	18
ΚΕΦΑΛΑΙΟ 3. ΑΚΑΝΘΟΚΥΤΤΑΡΙΚΟ ΚΑΡΚΙΝΩΜΑ ΤΟΥ ΣΤΟΜΑΤΟΣ	23
3.1 Γενικά Στοιχεία.....	23
3.2 Ιστοπαθολογία.....	26

3.3 Σταδιοποίηση **28**

3.4 Θεραπεία **30**

3.5 Μοριακή Παθοβιολογία **31**

ΚΕΦΑΛΑΙΟ 4. ΜΕΤΑ-ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΓΟΝΙΔΙΑΚΗΣ ΕΚΦΡΑΣΗΣ **37**

4.1 Τεχνικές υψηλής ανάλυσης στην μελέτη της γονιδιακής έκφρασης **37**

4.2 Δυσκολίες στην μετα-ανάλυση ελεύθερων δεδομένων μικροσυστοιχιών..... **37**

4.3 Στρατηγικές Επίλυσης Προβλημάτων Ομογενοποίησης **47**

II. ΕΙΔΙΚΟ ΜΕΡΟΣ

ΚΕΦΑΛΑΙΟ 1. ΣΚΟΠΟΣ ΤΗΣ ΕΡΓΑΣΙΑΣ **54**

ΚΕΦΑΛΑΙΟ 2. ΥΛΙΚΟ ΚΑΙ ΜΕΘΟΔΟΣ.......... **55**

2.1 ΣΥΛΛΟΓΗ ΣΧΕΤΙΚΩΝ ACCESSION NUMBERS KAI ΑΡΧΙΚΟ ΦΙΛΤΡΑΡΙΣΜΑ **55**

2.2 Αρχική επεξεργασία δεδομένων και συλλογή των datasets.......... **57**

2.3 Δημιουργία τυποποιημένων αρχείων , προβλήματα και χειρισμοί που χρησιμοποιήθηκαν στην κατασκευή τους **59**

2.4 Γραφική Παρατήρηση των δεδομένων.......... **60**

2.5 Προεπεξεργασία Δεδομένων των επιλεγχέντων datasets.......... **62**

2.6 Στατιστική και Λειτουργική Ανάλυση **65**

2.7 Εφαρμογή Προχωρημένων Τεχνικών Ανάλυσης **70**

ΚΕΦΑΛΑΙΟ 3. ΑΠΟΤΕΛΕΣΜΑΤΑ.......... **73**

3.1 Στατιστική Ανάλυση για εύρεση ΔΕ γονιδίων **73**

3.2 Αξιολόγηση μεθόδων βάσει ικανότητας διαχωρισμού δειγμάτων **74**

3.3 Γραφικές Απεικονίσεις στατιστικής ανάλυσης ΔΕ γονιδίων **76**

3.3.1 Volcano Plots

3.3.2 Διαγράμματα Venn..... **77**

3.3.3 Θερμικοί Χάρτες (Heatmaps).....	79
3.4 Αποτελέσματα ανάλυσης κλάσεων με βάση τα μεταδεδομένα σταδιοποίησης	82
3.5 Λειτουργική Ανάλυση γονιδιακής έκφρασης	86
3.6 Χαρτογράφηση ΔΕ γονιδίων σε γράφους της Kegg	90
3.7 Ανάλυση Κύριων Συνιστώσων (PCA)	100
3.8 Μηχανική Μάθηση I. Τυχαιοποιημένα Δάση.....	101
3.9 Μηχανική Μάθηση II. Δένδρα Απόφασης	103
3.10 Ανάλυση Βιολογικών Δικτύων.....	107
ΚΕΦΑΛΑΙΟ 4. ΣΥΖΗΤΗΣΗ	111
ΚΕΦΑΛΑΙΟ 5. ΣΥΜΠΕΡΑΣΜΑΤΑ	126
ΒΙΒΛΙΟΓΡΑΦΙΑ.....	127
III.ΠΑΡΑΡΤΗΜΑΤΑ	
ΠΑΡΑΡΤΗΜΑ 1. ΑΡΧΙΚΑ GEO DATASETS	149
ΠΑΡΑΡΤΗΜΑ 2. – ΑΠΟΔΕΚΤΑ ΣΤΗΝ ΠΡΩΤΗ ΦΑΣΗ DATASETS	154
ΠΑΡΑΡΤΗΜΑ 3 – ΙΔΙΟΤΗΤΕΣ.....	156
ΠΑΡΑΡΤΗΜΑ 4 – ΕΝΔΕΙΚΤΙΚΟ ΠΑΡΑΔΕΙΓΜΑ ΠΡΟΕΠΕΞΕΡΓΑΣΙΑΣ ΚΑΙ ΤΥΠΟΠΟΙΗΣΗΣ ΕΝΟΣ ΣΥΝΟΛΟΥ .CEL FILES ΜΕ AFFY	158
ΠΑΡΑΡΤΗΜΑ 5 – ΓΡΑΦΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΕΠΕΞΕΡΓΑΣΙΑΣ ΤΩΝ CEL ΑΡΧΕΙΩΝ	166

Κατάλογος Εικόνων

ΕΙΚΟΝΑ 1. ΓΡΑΦΙΚΗ ΑΝΑΠΑΡΑΣΤΑΣΗ ΙΣΤΟΛΟΓΙΚΗΣ ΕΙΚΟΝΑΣ ΣΤΟΜΑΤΙΚΟΥ ΒΛΕΝΝΟΓΟΝΟΥ ΑΔΕΙΑ ΧΡΗΣΗΣ: COMMON LICENSE ΑΠΟ HTTPS://COMMONS.WIKIMEDIA.ORG/WIKI/FILE:ORAL_MUCOSA.JPG	3
ΕΙΚΟΝΑ 2. ΙΣΤΟΛΟΓΙΚΑ ΕΙΔΗ ΕΠΙΘΗΛΙΟΥ. ΑΠΟ ΑΡΙΣΤΕΡΑ ΠΡΟΣ ΤΑ ΔΕΞΙΑ ΕΧΟΥΜΕ ΙΣΤΟΛΟΓΙΚΕΣ ΕΙΚΟΝΕΣ ΜΗ ΚΕΡΑΤΙΝΟΠΟΙΗΜΕΝΟΥ, ΠΑΡΑΚΕΡΑΤΙΝΟΠΟΙΗΜΕΝΟΥ ΚΑΙ ΟΡΘΟΚΕΡΑΤΙΝΟΠΟΙΗΜΕΝΟΥ ΕΠΙΘΗΛΙΟΥ. (ΕΙΚΟΝΑ ΑΠΟ ΤΟ ΒΙΒΑΙΟ ΤΩΝ HAND & FRANK, 2014)	9
ΕΙΚΟΝΑ 3. ΣΧΗΜΑΤΙΚΗ ΑΠΕΙΚΟΝΙΣΗ ΤΗΣ ΔΙΑΣΥΝΔΕΣΗΣ ΕΠΙΘΗΛΙΟΥ-ΣΥΝΔΕΤΙΚΟΥ ΙΣΤΟΥ (ΑΠΟ ALBERTS ET AL, 2015)	11
ΕΙΚΟΝΑ 4. ΚΥΤΤΑΡΙΚΟΣ ΚΥΚΛΟΣ ΚΑΙ ΣΗΜΕΙΑ ΕΛΕΓΧΟΥ ΤΟΥ (ΑΠΟ DANG ET AL, 2021).....	14
ΕΙΚΟΝΑ 5. ΤΟ ΜΟΝΟΠΑΤΙ JAK/STATS. (ΑΠΟ BARATA & OLIVEIRA , 2019).....	21
ΕΙΚΟΝΑ 6. ΤΟ ΜΟΝΟΠΑΤΙ RAS/MEK/ERK (ΑΠΟ BARATA & OLIVEIRA , 2019) ..	22
ΕΙΚΟΝΑ 7. ΤΟ ΜΟΝΟΠΑΤΙ RAS/MEK/ERK (ΑΠΟ BARATA & OLIVEIRA , 2019)	22
ΕΙΚΟΝΑ 8. ΚΛΙΝΙΚΗ ΕΙΚΟΝΑ ΑΠΟ ΕΝΑ ΑΚΑΝΘΟΚΥΤΤΑΡΙΚΟ ΚΑΡΚΙΝΩΜΑ ΤΗΣ ΓΛΩΣΣΑΣ (ΑΠΟ NEVILLE ET AL, 2008).....	26
ΕΙΚΟΝΑ 9. ΙΣΤΟΛΟΓΙΚΗ ΕΙΚΟΝΑ ΑΚΑΝΘΟΚΥΤΤΑΡΙΚΟΥ ΚΑΡΚΙΝΩΜΑΤΟΣ ΚΑΛΗΣ ΔΙΑΦΟΡΟΠΟΙΗΣΗΣ ΣΕ 2 ΜΕΓΕΝΘΥΝΣΕΙΣ, ΣΤΗΝ ΠΑΝΩ ΜΕ ΜΙΚΡΗ ΜΕΓΕΝΘΥΝΣΗ ΚΑΙ ΣΤΗΝ ΚΑΤΩ ΜΕ ΜΕΓΑΛΗ, ΟΠΟΥ ΒΛΕΠΟΥΜΕ ΚΑΙ ΜΑΡΓΑΡΙΤΑΡΙΑ ΚΕΡΑΤΙΝΗΣ (ΑΠΟ NEVILLE ET AL, 2008)	28
ΕΙΚΟΝΑ 10. ΠΡΟΣΠΑΘΕΙΑ ΠΡΟΕΠΕΞΕΡΓΑΣΙΑΣ RAW ΔΕΔΟΜΕΝΩΝ ΕΝΟΣ DATASET ΜΕ ΣΧΕΤΙΚΑ ΜΙΚΡΟ ΑΡΙΘΜΟ ΔΕΙΓΜΑΤΩΝ ΣΕ ΥΠΟΛΟΓΙΣΤΗ ΓΕΝΙΚΗΣ ΧΡΗΣΗΣ.....	39
ΕΙΚΟΝΑ 11. ΠΡΟΣΠΑΘΕΙΑ ΠΡΟΕΠΕΞΕΡΓΑΣΙΑΣ ΕΝΟΣ ΠΟΛΥ ΜΙΚΡΟΥ DATASET (9 ΔΕΙΓΜΑΤΑ) ΣΤΟ ΟΠΟΙΟ ΕΠΡΕΠΕ ΝΑ ΓΙΝΕΙ ΕΠΕΞΕΡΓΑΣΙΑ ΜΕ ΤΟ ΠΑΚΕΤΟ OLIGO ΤΗΣ AFFYMETRIX ΣΕ ΥΠΟΛΟΓΙΣΤΗ ΜΕ ΛΕΙΤΟΥΡΓΙΚΟ ΣΥΣΤΗΜΑ 32-BIT.....	40
ΕΙΚΟΝΑ 12. ΑΞΙΟΛΟΓΗΣΗ ΠΟΙΟΤΗΤΑΣ ΔΕΔΟΜΕΝΩΝ ΜΕ NUSE KAI RLE ΣΕ NORMALIZED DATASET ΑΠΟ ΤΟ GEO . ΤΟ ΔΕΙΓΜΑ 7 ΦΑΙΝΕΤΑΙ ΝΑ ΕΙΝΑΙ ΜΕ ΣΙΓΟΥΡΙΑ ΕΝΑ ΔΕΙΓΜΑ ΚΑΚΗΣ ΠΟΙΟΤΗΤΑΣ. ΤΟ ΔΕΙΓΜΑ 11 ΦΑΙΝΕΤΑΙ ΕΠΙΣΗΣ ΑΜΦΙΒΟΛΟΥ ΠΟΙΟΤΗΤΑΣ.....	41
ΕΙΚΟΝΑ 13. ΔΙΑΓΡΑΜΜΑ Q-Q «ΚΑΝΟΝΙΚΟΠΟΙΗΜΕΝΟΥ» DATASET. Η ΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ ΑΝΑΠΑΡΙΣΤΑΤΑΙ ΑΠΟ ΤΗΝ ΜΠΛΕ ΓΡΑΜΜΗ, ΕΝΩ Η ΚΑΤΑΝΟΜΗ ΤΟΥ ΔΕΙΓΜΑΤΟΣ ΑΠΟ ΤΗΝ ΜΑΥΡΗ ΚΑΜΠΥΛΗ.....	42
ΕΙΚΟΝΑ 14. ΔΕΔΟΜΕΝΑ ΣΤΑ ΟΠΟΙΑ ΑΝΑΦΕΡΕΤΑΙ ΟΤΙ ΕΧΕΙ ΓΙΝΕΙ QN	43
ΕΙΚΟΝΑ 15. ΤΑ ΔΕΔΟΜΕΝΑ ΤΙΣ ΕΙΚΟΝΑΣ 10 ΜΕΤΑ ΑΠΟ ΕΝΑ ΑΠΛΟ QN ΚΑΙ ΛΟΓΑΡΙΘΜΗΣΗ ΜΕ LOG2.....	43

ΕΙΚΟΝΑ 16. Q-Q ΔΙΑΓΡΑΜΜΑ ΑΠΟ ΔΕΔΟΜΕΝΑ ΜΕΤΑ ΑΠΟ GCRMA (ΑΡΙΣΤΕΡΑ) ΚΑΙ ΙΔΙΑ ΔΕΔΟΜΕΝΑ ΜΕΤΑ ΑΠΟ RMA ΚΑΝΟΝΙΚΟΠΟΙΗΣΗ ΤΩΝ CEL FILES (ΔΕΞΙΑ).....	44
ΕΙΚΟΝΑ 17. ΤΟ ΠΡΟΒΛΗΜΑ «MANY TO MANY» ΣΤΑ MICROARRAYS ΣΕ ENA R STUDIO TERMINAL ΚΑΤΑ ΤΗΝ ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΑΥΤΗ ΤΗΝ ΔΙΠΛΩΜΑΤΙΚΗ.....	46
ΕΙΚΟΝΑ 18. 200 ΔΕΙΓΜΑΤΑ ΑΠΟ ΔΙΑΦΟΡΕΤΙΚΑ DATASETS, ΕΠΕΞΕΡΓΑΣΜΕΝΑ ΜΕ ΤΗΝ ΠΡΟΣΕΓΓΙΣΗ ΤΗΣ ΑΠΛΗΣ Z-ΚΑΝΟΝΙΚΟΠΟΙΗΣΗΣ ΚΑΤΑ ΣΤΗΛΕΣ ΚΑΙ ANYΨΩΜΕΝΑ ΜΕ ΤΕΤΟΙΟ ΤΡΟΠΟ ΩΣΤΕ ΝΑ ΕΙΝΑΙ ΟΛΑ ΘΕΤΙΚΑ.....	49
ΕΙΚΟΝΑ 19. BOXPLOT ΤΩΝ ΠΡΩΤΩΝ 200 ΔΕΙΓΜΑΤΩΝ ΣΤΑ ΔΕΔΟΜΕΝΑ ΜΑΣ, ΜΕΤΑ ΑΠΟ ΔΙΠΛΗ Z-ΚΑΝΟΝΙΚΟΠΟΙΗΣΗ, ΕΙΝΑΙ ΕΜΦΑΝΕΙΣ ΟΙ ΜΕΓΑΛΕΣ ΤΥΠΙΚΕΣ ΑΠΟΚΛΙΣΕΙΣ ΠΟΥ ΔΗΜΙΟΥΡΓΟΥΝΤΑΙ.....	50
ΕΙΚΟΝΑ 20. BOXPLOT ΜΕ ΤΑ ΠΡΩΤΑ 200 ΔΕΔΟΜΕΝΑ ΟΠΩΣ ΑΥΤΑ ΠΡΟΕΚΥΨΑΝ ΑΠΟ ΤΗΝ ΔΙΑΣΠΑΣΤΙΚΗ ΜΕΘΟΔΟ.....	51
ΕΙΚΟΝΑ 21. PIPELINE ΒΗΜΑΤΩΝ ΣΥΛΛΟΓΗΣ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑΣ ΑΡΧΙΚΩΝ ΔΕΔΟΜΕΝΩΝ. ΟΙ ΑΡΙΘΜΟΙ ΑΠΕΙΚΟΝΙΖΟΥΝ ΤΟΝ ΑΡΙΘΜΟ ΤΩΝ DATASET ΣΤΑ ΕΠΙΜΕΡΟΥΣ ΣΤΑΔΙΑ.....	60
ΕΙΚΟΝΑ 22. PIPELINE ΓΡΑΦΙΚΗΣ ΑΠΕΙΚΟΝΙΣΗ ΚΑΤΑΝΟΜΩΝ ΚΑΙ ΦΙΛΤΡΑΙΣΜΑ ΤΩΝ ΔΕΔΟΜΕΝΩΝ. ΟΙ ΑΡΙΘΜΟΙ ΥΠΟΔΗΛΩΝΟΥΝ ΤΟΝ ΑΡΙΘΜΟ DATASETS ΑΝΑ ΣΤΑΔΙΟ.....	62
ΕΙΚΟΝΑ 23. PIPELINE ΠΡΟΕΠΕΞΕΡΓΑΣΙΑΣ RAW ΔΕΔΟΜΕΝΩΝ ΜΕ Z- ΚΑΝΟΝΙΚΟΠΟΙΗΣΗ.....	64
ΕΙΚΟΝΑ 24. PIPELINE ΠΡΟΕΠΕΞΕΡΓΑΣΙΑΣ ΜΕ ΤΗΝ ΔΙΑΣΠΑΣΤΙΚΗ ΤΕΧΝΙΚΗ	65
ΕΙΚΟΝΑ 25. BOXPLOT ΤΩΝ ΠΡΩΤΩΝ 200 SAMPLES ΜΕΤΑ ΑΠΟ Z- ΚΑΝΟΝΙΚΟΠΟΙΗΣΗ ΣΕ ΣΤΗΛΕΣ	66
ΕΙΚΟΝΑ 26. BOXPLOT 200 ΠΡΩΤΩΝ ΔΕΙΓΜΑΤΩΝ ΜΕ ΤΗΝ ΔΙΑΣΠΑΣΤΙΚΗ ΚΑΙ ΕΠΙΜΕΡΟΥΣ ΣΤΑΔΙΑ	66
ΕΙΚΟΝΑ 27. BOXPLOT ΠΡΩΤΩΝ 200 ΔΕΙΓΜΑΤΩΝ ΜΕ ΤΗΝ ΔΙΑΣΠΑΣΤΙΚΗ ΤΕΧΝΙΚΗ	67
ΕΙΚΟΝΑ 28. PIPELINE ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΛΕΙΤΟΥΡΓΙΚΗΣ ΑΝΑΛΥΣΗΣ.....	70
ΕΙΚΟΝΑ 29. PIPELINE ΣΤΑΤΙΣΤΙΚΗΣ ΑΝΑΛΥΣΗΣ ΚΛΑΣΕΩΝ ΜΕ ΒΑΣΕΙ ΤΑ ΥΠΑΡΧΟΝΤΑ ΜΕΤΑΔΕΔΟΜΕΝΑ.....	70
ΕΙΚΟΝΑ 30. PIPELINE ΠΡΟΧΩΡΗΜΕΝΗΣ ΕΠΕΞΕΡΓΑΣΙΑΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ	72
ΕΙΚΟΝΑ 31. Z-SCORE ΤΕΧΝΙΚΗ ΜΟΝΟ ΣΕ ΣΤΗΛΕΣ ΜΕΤΑΞΥ ΤΟΥ T-TEST ΚΑΙ ΤΟΥ MODERATED T-TEST.....	76
ΕΙΚΟΝΑ 32. VOLCANO PLOTS ΜΕ ΤΗΝ ΤΕΧΝΙΚΗ ΤΟΥ ΔΙΠΛΟΥ Z-SCORE ΓΙΑ ΤΑ MODERATED-T ΚΑΙ T-TEST	76
ΕΙΚΟΝΑ 33. VOLCANO PLOT ΑΠΟ ΤΗΝ ΔΙΑΣΠΑΣΤΙΚΗ ΤΕΧΝΙΚΗ ΜΕΤΑΞΥ ΤΟΥ T- TEST ΚΑΙ ΤΟΥ MODERATED T-TEST	77
ΕΙΚΟΝΑ 34. ΔΙΑΓΡΑΜΜΑ VENN ΔΕ ΓΟΝΙΔΙΩΝ ΣΤΗΝ ΜΟΝΟ ΚΑΤΑ ΣΤΗΛΕΣ Z- SCORE ΤΕΧΝΙΚΗ	77

ΕΙΚΟΝΑ 35. VENN ΔΙΑΓΡΑΜΜΑ ΔΕ ΓΟΝΙΔΙΩΝ ΜΕ ΤΗΝ ΤΕΧΝΙΚΗ ΤΗΣ ΔΙΠΛΗΣ Z- ΚΑΝΟΝΙΚΟΠΟΙΗΣΗΣ	78
ΕΙΚΟΝΑ 36. ΔΙΑΓΡΑΜΜΑ VENN ΔΕ ΓΟΝΙΔΙΩΝ ΣΤΗΝ ΔΙΑΣΠΑΣΤΙΚΗ ΤΕΧΝΙΚΗ	78
ΕΙΚΟΝΑ 37. ΗΕΑΤΜΑΡ ΤΩΝ ΔΕ ΓΟΝΙΔΙΩΝ ΣΤΗΝ ΤΕΧΝΙΚΗ Z-SCORE MONO ΚΑΤΑ ΣΤΗΛΕΣ ΚΑΙ T-TESTS	79
ΕΙΚΟΝΑ 38. ΗΕΑΤΜΑΡ ΤΩΝ ΔΕ ΓΟΝΙΔΙΩΝ ΜΕ ΔΙΠΛΗ Z-ΚΑΝΟΝΙΚΟΠΟΙΗΣΗ ΚΑΙ T-TESTS	80
ΕΙΚΟΝΑ 39. ΗΕΑΤΜΑΡ ΤΩΝ ΔΕ ΓΟΝΙΔΙΩΝ ΜΕ ΔΙΑΣΠΑΣΤΙΚΗ ΤΕΧΝΙΚΗ ΚΑΙ T- TESTS	80
ΕΙΚΟΝΑ 40. ΗΕΑΤΜΑΡ ΔΕ ΓΟΝΙΔΙΩΝ ΜΕ Z-SCORE ΤΕΧΝΙΚΗ MONO ΚΑΤΑ ΣΤΗΛΕΣ ΚΑΙ SAM	81
ΕΙΚΟΝΑ 41. ΗΕΑΤΜΑΡ ΔΕ ΓΟΝΙΔΙΩΝ ΜΕ ΔΙΠΛΗ Z-ΚΑΝΟΝΙΚΟΠΟΙΗΣΗ ΤΩΝ ΔΕ ΓΟΝΙΔΙΩΝ ΚΑΙ SAM	81
ΕΙΚΟΝΑ 42. ΗΕΑΤΜΑΡ ΔΕ ΓΟΝΙΔΙΩΝ ΜΕ ΔΙΑΣΠΑΣΤΙΚΗ ΤΕΧΝΙΚΗ ΚΑΙ SAM.	82
ΕΙΚΟΝΑ 43. VENN ΔΙΑΓΡΑΜΜΑ ΔΕ ΓΟΝΙΔΙΩΝ ΜΕ ΑΝΑ ΣΤΗΛΗ Z- ΚΑΝΟΝΙΚΟΠΟΙΗΣΗ ΚΑΙ ANOVA	83
ΕΙΚΟΝΑ 44. VENN ΔΙΑΓΡΑΜΜΑ ΔΕ ΜΕ ΔΙΠΛΗ Z-ΚΑΝΟΝΙΚΟΠΟΙΗΣΗ ΚΑΙ ANOVA	84
ΕΙΚΟΝΑ 45. ΔΙΑΓΡΑΜΜΑΤΑ VENN ΔΕ ΓΟΝΙΔΙΩΝ ΜΕ ΤΗΝ ΔΙΑΣΠΑΣΤΙΚΗ ΤΕΧΝΙΚΗ ΚΑΙ ANOVA	84
ΕΙΚΟΝΑ 46. ΗΕΑΤΜΑΡ ΜΕ Z-ΚΑΝΟΝΙΚΟΠΟΙΗΣΗ ΑΝΑ ΣΤΗΛΗ ΚΑΙ ANOVA.....	85
ΕΙΚΟΝΑ 47. ΗΕΑΤΜΑΡ ΜΕ ΔΙΠΛΗ Z-ΚΑΝΟΝΙΚΟΠΟΙΗΣΗ ΚΑΙ ANOVA	85
ΕΙΚΟΝΑ 48. ΗΕΑΤΜΑΡ ΜΕ ΤΗΝ ΔΙΑΣΠΑΣΤΙΚΗ ΠΡΟΣΕΓΓΙΣΗ ΚΑΙ ANOVA	86
ΕΙΚΟΝΑ 49. ΑΛΛΗΛΕΠΙΔΡΑΣΕΙΣ ΚΥΤΤΑΡΙΚΩΝ ΥΠΟΔΟΧΕΩΝ- ΕΞΩΚΥΤΤΑΡΙΑΣ ΜΗΤΡΑΣ	90
ΕΙΚΟΝΑ 50. ΜΕΤΑΒΟΛΙΣΜΟΣ RETINΟΛΗΣ	91
ΕΙΚΟΝΑ 51. ΠΡΟΣΒΟΛΗ ΑΠΟ ΙΟΥΣ HPV	92
ΕΙΚΟΝΑ 52. ΚΟΙΝΟΙ ΑΝΟΣΟΛΟΓΙΚΟΙ ΜΗΧΑΝΙΣΜΟΙ ΤΗΣ ΑΜΟΙΒΑΔΩΣΗΣ ΜΕ ΤΟ AKK	93
ΕΙΚΟΝΑ 53. ΣΗΜΑΤΟΔΟΤΙΚΟ ΜΟΝΟΠΑΤΙ AGE-RAGE	94
ΕΙΚΟΝΑ 54. ΣΗΜΑΤΟΔΟΤΙΚΟ ΜΟΝΟΠΑΤΙ IL-17	95
ΕΙΚΟΝΑ 55. ΣΗΜΑΤΟΔΟΤΙΚΟ ΜΟΝΟΠΑΤΙ TLR (TOLL-LIKE RECEPTOR)	96
ΕΙΚΟΝΑ 56. ΜΟΝΟΠΑΤΙ ΕΣΤΙΑΚΗΣ ΣΥΓΚΟΛΛΗΣΗΣ	97
ΕΙΚΟΝΑ 57. ΣΗΜΑΤΟΔΟΤΙΚΟ ΜΟΝΟΠΑΤΙ PI3/AKT/MTOR	98
ΕΙΚΟΝΑ 58. ΑΛΛΗΛΕΠΙΔΡΑΣΕΙΣ ΚΥΤΟΚΙΝΩΝ ΚΑΙ ΟΙΚΕΙΩΝ ΥΠΟΔΟΧΕΩΝ ΤΟΥΣ	99
ΕΙΚΟΝΑ 59. ΔΙΑΓΡΑΜΜΑΤΑ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΣΤΗΝ PCA	100
ΕΙΚΟΝΑ 60. ΠΡΟΣΔΙΟΡΙΣΜΟΣ ΤΗΣ ΥΠΕΡΠΑΡΑΜΕΤΡΟΥ MTRY ΓΙΑ ΤΟ ΜΕΓΕΘΟΣ ΤΩΝ ΕΠΙΜΕΡΟΥΣ ΔΕΝΤΡΩΝ	102
ΕΙΚΟΝΑ 61. ΑΠΟΔΟΣΗ ΑΛΓΟΡΙΘΜΟΥ ΤΥΧΑΙΟΠΟΙΗΜΕΝΩΝ ΔΑΣΩΝ.	102
ΕΙΚΟΝΑ 62. ΚΑΜΠΥΛΗ ROC ΓΙΑ ΤΑ ΤΥΧΑΙΟΠΟΙΗΜΕΝΑ ΔΑΣΗ	103

ΕΙΚΟΝΑ 63. ΠΡΟΣΔΙΟΡΙΣΜΟΣ ΥΠΕΡΠΑΡΑΜΕΤΡΟΥ ΠΕΡΙΠΛΟΚΟΤΗΤΑΣ ΚΑΙ ΤΟ ΔΕΝΔΡΟ ΑΠΟΦΑΣΗΣ ΜΕ ΤΟΝ ΑΛΓΟΡΙΘΜΟ CART	104
ΕΙΚΟΝΑ 64. ΕΠΙΛΟΓΗ ΥΠΕΡΠΑΡΑΜΕΤΡΟΥ MINCRITERION.....	105
ΕΙΚΟΝΑ 65. ΔΕΝΔΡΟ ΑΠΟΦΑΣΗΣ ΜΕ ΤΟΝ ΑΛΓΟΡΙΘΜΟ C-TREE	105
ΕΙΚΟΝΑ 66. ΚΑΜΠΥΛΕΣ ROC ΓΙΑ ΤΟΥΣ 2 ΑΛΓΟΡΙΘΜΟΥΣ ΔΕΝΔΡΩΝ ΑΠΟΦΑΣΗΣ	106
ΕΙΚΟΝΑ 67. ΒΙΟΛΟΓΙΚΟ ΔΙΚΤΥΟ ΤΩΝ ΥΠΕΡΕΚΦΡΑΖΟΜΕΝΩΝ ΓΟΝΙΔΙΩΝ ΜΕ ΛΕΙΤΟΥΡΓΙΚΗ ΣΥΝΔΕΣΗ	108
ΕΙΚΟΝΑ 68. ΚΑΤΑΝΟΜΗ ΕΝΔΙΑΜΕΣΗΣ ΕΚΚΕΝΤΡΟΤΗΤΑΣ	109
ΕΙΚΟΝΑ 69. ΚΥΚΛΟΤΕΡΕΙΣ ΓΡΑΦΟΙ ΤΩΝ CLUSTERS. ΤΟ ΚΑΤΩ ΜΕΡΟΣ ΕΧΕΙ 2 ΗΜΙΚΥΚΛΙΚΟΥΣ ΚΑΙ 1 ΚΥΚΛΙΚΟ ΓΡΑΦΟ.	109
ΕΙΚΟΝΑ 70. ΚΑΤΑΝΟΜΗ ΒΑΘΩΝ ΚΟΜΒΩΝ	110
ΕΙΚΟΝΑ 71. ΓΡΑΦΟΣ ΜΕ ΤΑ ΣΥΧΝΟΤΕΡΑ ΓΟΝΙΔΙΑ ΣΤΗΝ ΒΙΒΛΙΟΓΡΑΦΙΑ, ΟΙ ΑΚΜΕΣ ΑΝΑΠΑΡΙΣΤΟΥΝ ΤΙΣ ΛΕΙΤΟΥΡΓΙΚΕΣ ΤΟΥΣ ΣΧΕΣΕΙΣ ΣΕ ΕΠΙΠΕΔΟ ΠΡΩΤΕΪΝΗΣ.....	125
ΕΙΚΟΝΑ 72. ΓΟΝΙΔΙΑ ΠΟΥ ΑΝΑΦΕΡΟΝΤΑΙ ΣΥΧΝΑ ΣΤΗΝ ΒΙΒΛΙΟΓΡΑΦΙΑ ΆΛΛΑ ΔΕΝ ΕΧΟΥΝ ΚΑΠΟΙΑ ΛΕΙΤΟΥΡΓΙΚΗ ΣΧΕΣΗ ΜΕΤΑΞΥ ΤΟΥΣ Η ΜΕ ΤΑ ΓΟΝΙΔΙΑ ΤΗΣ ΠΡΟΗΓΟΥΜΕΝΗΣ ΕΙΚΟΝΑΣ.....	125

Κατάλογος Πινάκων

ΠΙΝΑΚΑΣ 1. ΚΥΚΛΙΝΕΣ ΚΑΙ ΣΥΜΠΛΕΓΜΑΤΑ ΤΟΥΣ (ΠΡΟΣΑΡΜΟΓΗ ΑΠΟ ALBERTS ET AL , 2015).....	15
ΠΙΝΑΚΑΣ 2. ΤΑΞΙΝΟΜΗΣΗ ΤΝΜ ΓΙΑ ΤΟΥΣ ΚΑΚΟΗΘΕΙΣ ΟΓΚΟΥΣ ΚΕΦΑΛΗΣ ΚΑΙ ΤΡΑΧΗΛΟΥ (ΠΡΟΣΑΡΜΟΓΗ ΑΠΟ NEVILLE ET AL, 2008).....	29
ΠΙΝΑΚΑΣ 3. ΤΕΛΙΚΑ GEO ACCESSION NUMBERS.	61
ΠΙΝΑΚΑΣ 4. ΠΙΝΑΚΑΣ ΣΥΓΧΥΣΗΣ ΜΕ Z-SCORE MONO ΑΝΑ ΣΤΗΛΗ ΚΑΙ T-TESTS	74
ΠΙΝΑΚΑΣ 5. ΠΙΝΑΚΑΣ ΣΥΓΧΥΣΗΣ ΜΕ Z-SCORE MONO ΑΝΑ ΣΤΗΛΗ ΚΑΙ SAM..	75
ΠΙΝΑΚΑΣ 6. ΠΙΝΑΚΑΣ ΣΥΓΧΥΣΗΣ ΜΕ ΤΗΝ ΜΕΘΟΔΟ ΔΙΠΛΗΣ Z-ΚΑΝΟΝΙΚΟΠΟΙΗΣΗΣ ΜΕ T-TESTS.....	75
ΠΙΝΑΚΑΣ 7. ΠΙΝΑΚΑΣ ΣΥΓΧΥΣΗΣ ΜΕ ΤΗΝ ΜΕΘΟΔΟ ΔΙΠΛΗΣ Z-ΚΑΝΟΝΙΚΟΠΟΙΗΣΗΣ ΚΑΙ SAM	75
ΠΙΝΑΚΑΣ 8. ΠΙΝΑΚΑΣ ΣΥΓΧΥΣΗΣ ΜΕ ΔΙΑΣΠΑΣΤΙΚΗ ΜΕΘΟΔΟ ΚΑΙ T-TESTS....	75
ΠΙΝΑΚΑΣ 9. ΠΙΝΑΚΑΣ ΣΥΓΧΥΣΗΣ ΜΕ ΔΙΑΣΠΑΣΤΙΚΗ ΜΕΘΟΔΟ ΚΑΙ SAM.....	75
ΠΙΝΑΚΑΣ 10. ΛΕΙΤΟΥΡΓΙΚΗ ΑΝΑΛΥΣΗ ΔΕ ΓΟΝΙΔΙΩΝ ΣΤΗΝ Z-ΚΑΝΟΝΙΚΟΠΟΙΗΣΗ ΜΟΝΟ ΣΕ ΣΤΗΛΕΣ	87
ΠΙΝΑΚΑΣ 11. ΛΕΙΤΟΥΡΓΙΚΗ ΑΝΑΛΥΣΗ ΔΕ ΓΟΝΙΔΙΩΝ ΜΕ ΤΗΝ ΜΕΘΟΔΟ ΤΗΣ ΔΙΠΛΗΣ ΚΑΝΟΝΙΚΟΠΟΙΗΣΗΣ	88
ΠΙΝΑΚΑΣ 12. ΛΕΙΤΟΥΡΓΙΚΗ ΑΝΑΛΥΣΗ ΤΩΝ ΔΕ ΓΟΝΙΔΙΩΝ ΣΤΗΝ ΔΙΑΣΠΑΣΤΙΚΗ ΤΕΧΝΙΚΗ	89
ΠΙΝΑΚΑΣ 13. ΑΡΙΘΜΟΙ ENTREZ ΓΙΑ ΤΑ 30 ΣΗΜΑΝΤΙΚΟΤΕΡΑ ΑΠΟ ΤΑΞΙΝΟΜΙΚΗΣ ΑΠΟΨΗΣ ΓΟΝΙΔΙΑ	101
ΠΙΝΑΚΑΣ 14. ΠΙΝΑΚΑΣ ΣΥΓΧΥΣΗΣ ΤΥΧΑΙΟΠΟΙΗΜΕΝΑ ΔΑΣΗ.....	103
ΠΙΝΑΚΑΣ 15. ΠΙΝΑΚΑΣ ΣΥΓΧΥΣΗΣ ΑΛΓΟΡΙΘΜΟΥ CART	104
ΠΙΝΑΚΑΣ 16. ΠΙΝΑΚΑΣ ΣΥΓΧΥΣΗΣ ΓΙΑ ΤΟΝ ΑΛΓΟΡΙΘΜΟ CTREE	106
ΠΙΝΑΚΑΣ 17. ΟΙ 10 ΠΙΟ ΣΗΜΑΝΤΙΚΟΙ ΚΟΜΒΟΙ ΑΠΟ ΑΠΟΨΗ ΕΝΔΙΑΜΕΣΗΣ ΕΚΚΕΝΤΡΟΤΗΤΑΣ	110

Συντομογραφίες και Ακρωνύμια

AGE	Advanced Glycation End products
ALDH	Aldehyde Dehydrogenase
ANOVA	Analysis of Variance
Apaf-1	Apoptotic Protease Activating Factor-1
APC/C	Anaphase Promoting Complex/Cyclosome
aSMA	actine Smooth Muscle Antibody
AURKB	Aurora Kinase B
BAK	BCL-2 Antagonist Killer
BAX	BCL associated X protein
BCL	B-cell Lymphoma
BIRC	Baculoviral IAP Repeat Containing
BMI	Polycomb complex protein
BP	Bullous Pemphigoid
CAF	Cancer Associated Fibroblast
CAF	Cancer Associated Fibroblasts
CAK	Cdk-Activating Kinase
CART	Classification and Regression Trees
CASP	Caspase
CCN	Cyclin
CD	Cluster of differentiation
Cdc	Cell Division Cycle
Cdh	Cadherine

CdK	Cyclin – dependent kinases
CDKN2A	Cyclin-Dependent Kinase Inhibitor 2A
Cip/Kip	CDK interacting protein/Kinase inhibitory protein
CK	Cytokeratin
COL17A1	Collagen Type XVII Alpha 1 Chain
CRAN	Compehensive R Archive Network
CSF	Colony Stimulating Factor
CTTN	Cortactine
Cul	Cullin
CXCL	C-X-C motif Ligand
DES	Desaturase
DISC	Death –inducing signaling complex
DNA	Deoxyribonucleic Acid
ECM	Extracellular Matrix
EGF	Epidermal Growth Factor
EGFR	Epidermal Growth Factor Receptor
EMT	Epithelial Mesenchymal Transition
FADD	Fas-associated protein with death domain
FAK	Focal Adhesion Kinase
Fas	FS-7-Associated Surface antigen
FAT	Fatty acid-binding Protein
FBXW	F-Box And WD Repeat Domain Containing
FDA	Food and Drug Administration
FDR	False Discovery Rate
FU	Fluorouracil

gcRMA	GeneChip RMA
GEO	Gene Expression Omnibus
GPX	Glutathione Peroxidase
HGF	Hepatocyte Growth Factor
HOX	Homeobox
HPV	Human Papillomavirus
HSD	Honestly Significant Difference
HST	Histone
IFN	Interferone
IL	Interleukine
INK4	Inhibitors of CdK4
INT	Integrator Complex
IQR	Interquartile Range
Kegg	Kyoto Encyclopedia of Genes and Genomes
kNN	k Nearest Neighbours
KGF	Keratinocyte Growth Factor
LIG	Ligase
logFC	logarithmic fold change
MET	Mesenchymal Epithelial Transition
MICAL2	Microtubule Associated Monooxygenase, Calponin & LIM Domain Containing 2
miRNA	micro RNA
MMP	Matrix Metalloproteinase
mTOR	mammalian Target of Rapamycin
MYT	Myelin Transcription Factor
ncRNA	non - coding RNA

NFE2L2	Nuclear Factor, Erythroid 2 Like 2
NSD	Nuclear Receptor Binding SET Domain Protein
NUSE	Normalized Unscaled Standard Error
PCA	Principal Components Analysis
PD	Programmed cell death Protein
PDGF	Platelet Derived Growth Factor
PIK3CA	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha
PSMB2	Proteosome Subunit B2
QN	Quantile Normalization
RAGE	Receptor AGE
RIP	Receptor Interacting Protein
RLE	Relative log Expression
RMA	Robust Microarray Analysis
RNA	Ribonucleic Acid
RNAseq	RNA sequence
ROC	Receiver operating characteristic
RPA	Replication Protein A
RRAGD	Ras-Related GT binding D
SAM	Significance analysis of Microarrays
SCF	Skp, Cullin, F-box containing complex
Skp	S-phase kinase-associated protein
SMAD	Small-warm Mothers Against Decapentaplegic
SNP	Single Nucleotide Polymorphism
SOX	SRY (sex determining region Y)-box
SPARC	Secreted Protein Acidic and Cysteine Rich

SPRR	Small Proline Rich Protein
TF	Transcription Factor
TGF	Tumour Growth Factor
TNF	Tumor Necrosis Factor
TNM	Tumour Nodes Metastasis
TP	Tumour Protein
TRAF	TNF Receptor Associated Factor
TWIST	Twist Related protein
VEGF	Vascular Endothelial Growth Factor
VIM	Vimentin
Wnt	Wingless-related Integration Site
YAP	Yes Associated Protein
AKK	Ακανθοκυτταρικό Καρκίνωμα
ΔΕ	Διαφορικά Εκφραζόμενα





I.

ΓΕΝΙΚΟ ΜΕΡΟΣ

Κεφάλαιο 1. Ιστολογία & Κυτταρική Βιολογία του Φυσιολογικού Στοματικού Βλεννογόννου

1.1 Εισαγωγικά Στοιχεία

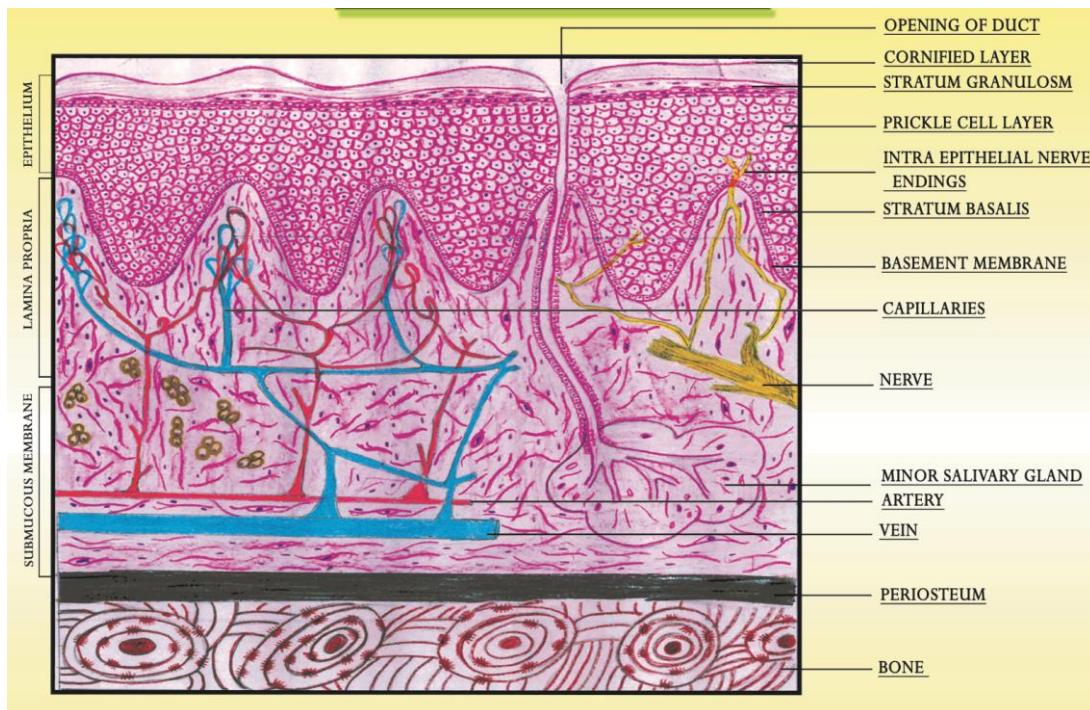
Ος βλεννογόνος ορίζεται το ενυδατωμένο επένδυμα της γαστρεντερικής οδού, ανώτερης αναπνευστικής οδού και άλλων κοιλοτήτων του ανθρώπινου σώματος που έρχονται σε επαφή με το εξωτερικό περιβάλλον (Nanci, 2018). Ο στοματικός βλεννογόνος αποτελεί ανατομικά, τμήμα του ανώτερου γαστρεντερικού συστήματος και διαχωρίζεται σε 3 τύπους:

- **Μασητικός:** Κερατινοποιημένος βλεννογόνος με εντόπιση στα ούλα και στην σκληρή υπερώα.
- **Καλυπτικός:** Λιγότερο κερατινοποιημένος βλεννογόνος με εντόπιση σε παρειά, γλώσσα, έδαφος του στόματος και στην μαλακή υπερώα.
- **Εξειδικευμένος στοματικός βλεννογόνος:** Αισθητικές θηλές της γλώσσας.

Στην εργασία αυτή επικεντρωνόμαστε στον καλυπτικό βλεννογόννο καθώς αυτός αποτελεί κατά συντριπτική πλειοψηφία το κυριότερο πεδίο εντόπισης του ακανθοκυτταρικού καρκινώματος (AKK) του στόματος (Neville, Damm, Allen & Bouquot, 2008). Ο στοματικός βλεννογόνος επιτελεί διάφορες λειτουργίες, οι πιο σημαντικές είναι η προστασία των υποκείμενων ανατομικών μορίων καθώς και διάφορες αισθητικές λειτουργίες μέσω ελεύθερων ή συνδεδεμένων με σωμάτια Merkel, ελεύθερων αμύνελων νευρικών απολήξεων (Kumar, 2015). Ανάμεσα στον στοματικό βλεννογόνο και στα υποκείμενα ανατομικά μόρια (οστά, μύες) υπάρχει μία ενδιάμεση στιβάδα που αποτελείται από χαλαρό συνδετικό ιστό, λιποκύτταρα, λεμφοειδή κύτταρα και ελάσσονες σιαλογόνους αδένες και ονομάζεται υποβλεννογόνος (Prasad & Anuthama, 2019 ; Kumar, 2015).

Ο βλεννογόνος του στόματος χωρίζεται σε 2 μικροσκοπικά διακριτές στιβάδες, το χόριο και την επιθηλιακή στιβάδα (Εικόνα 1). Το χόριο διαιρείται νοητά σε 2 ζώνες, την δικτυωτή και την θηλώδη, περιέχει κυρίως συνδετικό ιστό με άφθονες κολλαγόνες και ελαστικές ίνες (Αγγελόπουλος, Παπανικολάου & Αγγελοπούλου, 2000). Περιέχει όμως και άλλα ιστολογικά στοιχεία όπως νεύρα, αγγεία, λεμφαγγεία και κύτταρα του ανοσοποιητικού συστήματος. Η επιθηλιακή στιβάδα περιέχει κερατινοκύτταρα, μελανοκύτταρα, κύτταρα

Merkel, κύτταρα Langerhans και σποραδικά ανευρίσκονται και άλλα κύτταρα του ανοσοποιητικού συστήματος όπως λεμφοκύτταρα (Prasad & Anuthama, 2019 ; Kumar, 2015).



Εικόνα 1. Γραφική αναπαράσταση ιστολογικής εικόνας Στοματικού Βλεννογόνου Αδεια χρήσης: Common License από https://commons.wikimedia.org/wiki/File:Oral_mucosa.jpg.

Στα πλαίσια της μελέτης του ΑΚΚ θα επικεντρωθούμε στην επιθηλιακή στιβάδα του στοματικού βλεννογόνου καθώς αυτή αποτελεί την ιστοπαθολογική προέλευση του, αλλά θα αναφερθούμε και στην διασύνδεση επιθηλίου-συνδετικού ιστού διότι παίζει κρίσιμο ρόλο στην μετάβαση ενός καρκινώματος από καρκίνωμα *in situ* σε διηθητικό καρκίνωμα με επέκταση στο χόριο (Neville et al, 2008).

1.2 Το Στοματικό Καλυπτικό Επιθήλιο

Ο τύπος του επιθηλιακού ιστού στο στόμα περιγράφεται στα εγχειρίδια οδοντιατρικής ιστολογίας ως «πολύστιβο πλακώδες επιθήλιο». Η εμβρυολογική του προέλευση είναι το εξώδερμα, εξαίρεση αποτελεί η οπίσθια μοίρα της γλώσσας που προέρχεται από το ενδόδερμα (Αγγελόπουλος και συν., 2000). Ο κυτταροσκελετός του επιθηλιακού κυττάρου αποτελείται κυρίως από ενδιάμεσα ινίδια κυτταροκερατίνης, τα οποία κάποιοι συγγραφείς τα ονομάζουν και τονοϊνίδια (Hand & Frank, 2014; Chiego & Chiedo, 2013). Υπάρχουν στον άνθρωπο 20 είδη κυτταροκερατίνης, όμως μία αδρή κατηγοριοποίηση είναι ο διαχωρισμός σε βασικές κυτταροκερατίνες (CK 1-8) και όξινες κυτταροκερατίνες (CK 9-20). Ένα ενδιάμεσο ινίδιο σε σταθερή κατάσταση αποτελείται από ένα ζεύγος CKs στο οποίο η μία πρέπει να ανήκει υποχρεωτικά στις όξινες και η άλλη στις βασικές CK. Αν δεν έχουμε συνδυασμό των 2 αυτών ομάδων τότε προκύπτουν ασταθείς CK που είναι ευάλωτες σε αποδόμηση. Στις επιφανειακές στιβάδες του επιθηλίου παρατηρούνται κυρίως CK1 και CK10, ενώ στην περιοχή της βασικής και υπερβασικής στιβάδας έχουμε ζεύγη CK4 και CK13. Βέβαια εδώ πρέπει να σημειωθεί ότι το είδος CK διαφέρει από την μία περιοχή του στόματος στην άλλη, π.χ. στην κάτω επιφάνεια της γλώσσας έχουμε συνδυασμούς των CK 5,6,14 ενώ στην μαλακή υπερώα συνδυασμούς CK 7,8,18 (Kumar, 2015).

Παραδοσιακά το επιθήλιο τόσο στο στόμα όσο και σε άλλες περιοχές του σώματος, διαιρείται σε κερατινοποιημένο και μη κερατινοποιημένο. Ωστόσο αυτός ο διαχωρισμός ίσως να είναι παραπλανητικός αφού όλα τα επιθήλια περιέχουν ενδιάμεσα ινίδια κερατίνης, και επομένως όλα τα επιθήλια είναι λίγο-πολύ κερατινοποιημένα. Παρακάτω θα δούμε ότι υπάρχουν διαφορές στην δομή, οργάνωση και ποσότητα της κερατίνης στα 2 αυτά παραδοσιακά είδη επιθηλίου. Επιπλέον το κερατινοποιημένο επιθήλιο διακρίνεται σε ορθοκερατινοποιημένο και παρακερατινοποιημένο ανάλογα με το αν υπάρχουν ορατά με το οπτικό μικροσκόπιο υπολείμματα πυρήνα ή όχι. Στην Εικόνα 2 παρατίθεται ιστολογικές εικόνες από στοματικό επιθήλιο (Hand & Frank, 2014).

Στο κερατινοποιημένο επιθήλιο έχουμε από εν τω βάθει προς την επιφάνεια 4 στιβάδες, την βασική, την ακανθωτή, την κοκκιώδη και την κερατίνη στιβάδα. Από την άλλη μεριά το μη-κερατινοποιημένο επιθήλιο έχει 3 στιβάδες την βασική, την ενδιάμεση και την επιφανειακή στιβάδα (Αγγελόπουλος και συν., 2000). Από την άποψη της μελέτης του AKK η βασική στιβάδα είναι η πιο σημαντική και για τα 2 είδη επιθηλίου.

Στην βασική/υπερβασική περιοχή υπάρχουν 2 είδη κερατινοκυττάρων:

- 1.Επιθηλιακά βλαστοκύτταρα (epithelial stem cells) που πολλαπλασιάζονται σχετικά αργά και πιστεύεται ότι βρίσκονται στην περιοχή ακριβώς πάνω από την κορυφή των θηλών του χορίου.
- 2.Μεταβατικά ενισχυτικά κύτταρα (transient amplifying cells) που έχουν μεγάλη δραστηριότητα πολλαπλασιασμού και πιστεύεται ότι βρίσκονται στα κύτταρα της βασικής στιβάδας που αντιστοιχούν στις επιθηλιακές καταδύσεις (Hand & Frank, 2014).

Και τα 2 παραπάνω είδη κυττάρων της βασικής στιβάδας έχουν ως αποστολή την παραγωγή επιθηλιακών κυττάρων, ωστόσο για τα μεταβατικά ενισχυτικά κύτταρα οι γνώσεις μας μέχρι σήμερα είναι λίγες, ενώ αντίθετα για τα επιθηλιακά κύτταρα υπάρχουν ανοσοϊστοχημικοί βιοδείκτες όπως οι CD44, Bmi1, Sox2 και CK14 που μπορούν να τα εντοπίσουν (Papagerakis, Pannone, Zheng, About, Taqi, Nguyen, ... & Papagerakis, 2014). Τα κύτταρα της βασικής μεμβράνης έχουν κυβοειδές σχήμα με μεγάλη αναλογία πυρήνα/κυτταροπλάσματος. Συνδέονται με την βασική μεμβράνη κατά κύριο λόγο με ειδικές δομές που ονομάζονται ημιδεσμοσώματα και μεταξύ τους με παρόμοιες (αλλά εις διπλούν) δομές που ονομάζονται δεσμοσώματα (Kumar, 2015). Περισσότερα για τα είδη σύνδεσης των επιθηλιακών κυττάρων μεταξύ τους θα δούμε στο τέλος αυτής της ενότητας, ενώ με τα ημιδεσμοσώματα θα ασχοληθούμε στην επόμενη ενότητα που ασχολείται με την διασύνδεση επιθηλίου/συνδετικού ιστού.

Η παραγωγή επιθηλιακών κυττάρων ξεκινά με την διαίρεση ενός επιθηλιακού βλαστοκυττάρου (stem cell) σε 2 θυγατρικά κύτταρα (daughter cells), από αυτά το ένα θυγατρικό κύτταρο διατηρεί απολύτως τα χαρακτηριστικά του αρχικού βλαστοκυττάρου και το δεύτερο γίνεται μεταβατικό ενισχυτικό κύτταρο. Αυτό μετά την παραγωγή του από το μητρικό βλαστοκύτταρο, πορεύεται πλάγια και προς τα κάτω στις επιθηλιακές καταδύσεις και μετά από μία περίοδο έντονου πολλαπλασιασμού, ακολουθεί την οδό της ωρίμανσης/διαφοροποίησης κατά την οποία το κύτταρο παύει να διαιρείται πλέον. Η σηματοδότηση αυτής της φάσης γίνεται με την αποκοπή από την βασική μεμβράνη μέσω μετατροπής της ιντεγκρίνης α6β4 των ημιδεσμοσωμάτων σε ιντεγκρίνη α3β1 (Hand & Frank, 2014). Λέμε το κύτταρο που ακολουθεί αυτή την τελευταία οδό «αποφασισμένο κύτταρο» (determined cell), το κύτταρο αυτό με την βοήθεια κυτοκινών και αυξητικών παραγόντων (IL-1, EGF ,

KGF, TGF κτλ) μεγαλώνει και κατά κάποιο τρόπο ενηλικιώνεται , η μοίρα του είναι να ωριμάσει . Πλέον δεν μπορεί να διαιρεθεί, μεταναστεύει παθητικά λόγω υπερ-συσσώρευσης κυττάρων στην βασική/υπερβασική περιοχή και εν τέλει ο τελικός προορισμός του είναι να νεκρωθεί στις επιφανειακές επιθηλιακές στιβάδες με μία διαδικασία προγραμματισμένου κυτταρικού θανάτου που θα περιγραφεί στην συνέχεια (Nanci, 2018 ; Kumar, 2015). Ο επιθηλιακός ιστός του στόματος ανανεώνεται κατά μέσο όρο κάθε 14-24 μέρες και είναι ταχύτερος από αυτόν του δέρματος, αλλά λιγότερο ταχύς σε σχέση με το εντερικό επιθήλιο. Όσο πιο κερατινοποιημένο είναι το επιθήλιο τόσο πιο αργός είναι ο ρυθμός ανανέωσης του. Ως ρυθμός ανανέωσης του επιθηλίου ορίζεται ο χρόνος ο οποίος χρειάζεται ένα αποφασισμένο επιθηλιακό κύτταρο από την δημιουργία του στην βασική ή υπερβασική περιοχή μέχρι να φτάσει πλέον στην επιφανειακότερη επιθηλιακή στιβάδα όπου νεκρώνεται με προγραμματισμένο κυτταρικό θάνατο και τελικά αποπίπτει και απομακρύνεται με μια διαδικασία που ονομάζεται απολέπιση (αγγλ. desquamation) (Papagerakis et al, 2014 ; Nanci, 2018).

Στην ακανθωτή στιβάδα τα επιθηλιακά κύτταρα μεγαλώνουν σε μέγεθος και έχουν μικρότερη αναλογία πυρήνα/κυτταροπλάσματος σε σχέση με τα κύτταρα της βασικής στιβάδας και αποκτούν ένα αστεροειδές σχήμα που προκύπτει από την σύνδεση τους με τα υπόλοιπα επιθηλιακά κύτταρα με εξειδικευμένες συνδετικές δομές που ονομάζονται δεσμοσώματα. Αυτές οι περιοχές είναι ορατές με μεγάλες μεγενθύνσεις στο οπτικό μικροσκόπιο με τις συνήθεις χρώσεις αιματοξυλίνης-ηωσίνης. Τα δεσμοσώματα αποτελούνται από ένα ενδοκυτταρικό τμήμα που είναι η πλάκα πρόσφυσης των μικρο-ινιδίων ακτίνης και των ενδιάμεσων ινιδίων κερατίνης (πλακόδιο), αυτή αποτελείται από τις πρωτεΐνες δεσμοπλακίνη και πλακοσφαιρίνη. Από το πλακόδιο εκτείνονται εξωκυτταρικά πλέον οι ειδικές διαμεμβρανικές πρωτεΐνες πρόσφυσης που είναι η δεσμογλεΐνη και η δεσμοκολλίνη. Αυτές ανήκουν στην οικογένεια των καδερινών (cadherines), που είναι πρωτεΐνες κυτταρικής συγκόλλησης (Hand & Frank, 2014). Η αντίστοιχη της ακανθωτής στιβάδας στα μη-κερατινοποιημένα επιθήλια είναι η ενδιάμεση στιβάδα, με την οποία μοιράζονται πολλά κοινά μορφολογικά χαρακτηριστικά. Ένα χαρακτηριστικό παράδειγμα μίας ασθένειας που προκύπτει λόγω αυτοάνοσης καταστροφής των δεσμοσωμάτων είναι η πέμφιγα (Αγγελόπουλος και συν, 2000) .

Οι επόμενες 2 στιβάδες, η κοκκώδης και η κερατίνη δεν υπάρχουν αντίστοιχα στα μη κερατινοποιημένα επιθήλια. Η κοκκώδης στιβάδα ονομάζεται έτσι λόγω των βασεόφιλων κοκκίων κερατοϋαλίνης (περιέχουν τριχοϋαλίνη και φιλαγκρίνη), αλλά και λόγω άλλης μιας κατηγορίας κοκκίων που ονομάζονται πλακοειδή κοκκία (lamellar granules ή κερατόσωματα ή σώματα του Odland ,περιέχουν γλυκολιπίδια) που παίζουν καθοριστικό ρόλο στην στεγανότητα του διακυτταρικού φραγμού σε ανώτερες στιβάδες. Στην κοκκώδη στιβάδα τα κύτταρα χάνουν το αστεροειδές σχήμα που είχαν στην ακανθωτή στιβάδα και γίνονται αποπεπλατυσμένα (πλακώδης μορφολογία). Η εμφάνιση των παραπάνω κοκκίων στα επιθηλιακά κύτταρα προκύπτει από απότομη μείωση του PH, από 7.2 σε 5, και σηματοδοτεί την επικείμενη απόπτωση του κυττάρου (Hand & Frank, 2014).

Η κερατίνη στιβάδα σχηματίζεται με την δημιουργία του κερατινικού κυτταρικού φακέλου. Η φιλαγκρίνη των κοκκίων κερατοϋαλίνης απελευθερώνεται εντός του κυττάρου και σχηματίζει ένα συσσωμάτωμα ενδιάμεσων ινών κερατίνης. Κάποιοι συγγραφείς (Lippens, Denecker, Ovaere, Vandenabeele, ... & Declercq, 2005) λένε ότι τα κύτταρα από αυτό το σημείο και μετά παύουν να ονομάζονται κερατινοκύτταρα και λέγονται εφεξής κερατοκύτταρα (αγγλ. corneocytes). Στην διαδικασία αυτή εμπλέκονται και άλλες πρωτεΐνες όπως η ινβιολουκρίνη, λορικρίνη, τριχοϋαλίνη, περιπλακίνη και μικρές πρωτεΐνες πλούσιες σε προλίνη (αγγλ. SPRRs) που συνδέονται διασταυρούμενα με την βοήθεια τρανσγλουταμινασών. Τα πλακοειδή κοκκία εικρίνουν τα περιεχόμενα σε αυτά γλυκολιπίδια στον εξωκυττάριο χώρο βοηθώντας στην δημιουργία ενός αδιαπέραστου φραγμού ενώ ταυτόχρονα ακολουθεί η σταδιακή αποδόμηση των δεσμοσωμάτων (Nanci, 2018). Τα κύτταρα στις ανώτερες περιοχές της κερατίνης στιβάδας πλέον είναι φανερά γερασμένα, αποτελούνται από πυκνές στίβες παχέων ινών κερατίνης, με ένα αποδομημένο ή στην περίπτωση της παρακερατίνης ανενεργό πυκνωτικό πυρήνα. Σε αυτή την στιβάδα πλέον τα κυτταρικά οργανίδια είναι ανύπαρκτα ή μερικώς αποδομημένα και λόγω της κατάρρευσης του κυτταροσκελετού τα κύτταρα εμφανίζονται εντελώς επίπεδα (πλακώδη). Τελικά αυτά τα κύτταρα μετά την πλήρη αποσύνθεση των δεσμοσωμάτων τους απομακρύνονται από το επιθήλιο ως «κυτταρικά απόβλητα» με την διαδικασία της απολέπισης.

Όπως ειπώθηκε παραπάνω στα μη κερατινοποιημένα επιθήλια, αντί για κοκκώδη και κερατίνη στιβάδα έχουμε την επιφανειακή στιβάδα. Σε αυτήν έχουμε κύτταρα με μικρότερο περιεχόμενο κερατίνης, τα ενδιάμεσα ινίδια κερατίνης είναι λιγότερο πυκνά και πιο χαλαρά

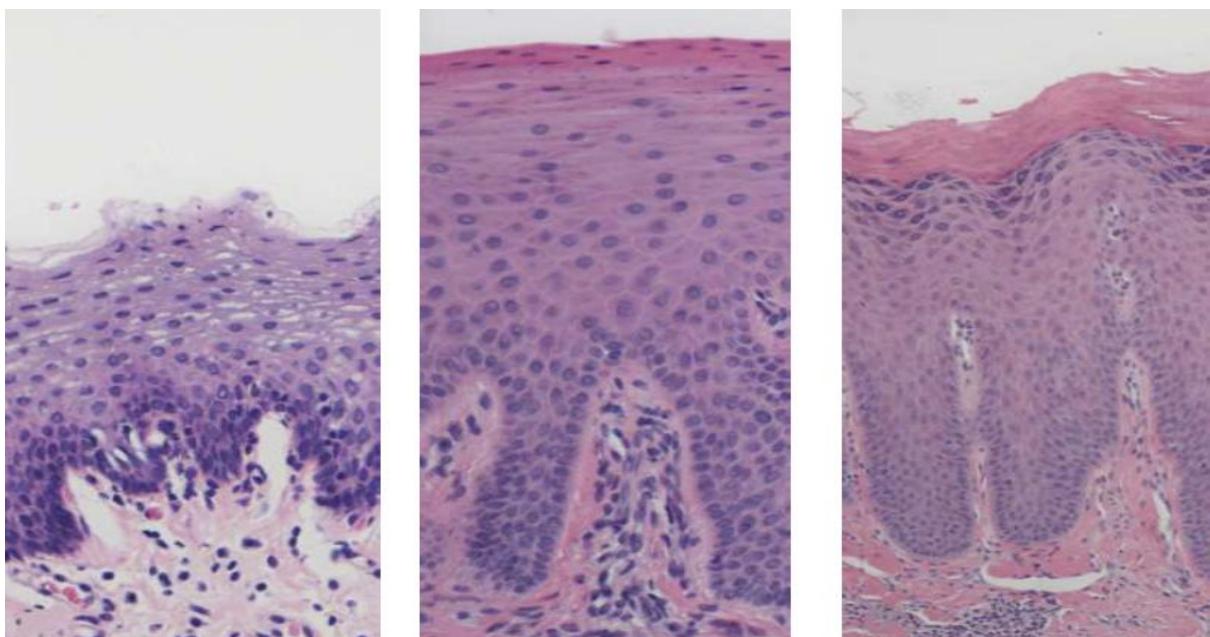
συνδεδεμένα μεταξύ τους, δεν σχηματίζουν παχιές δέσμες κερατίνης όπως σχηματίζονται στα κύτταρα των 2 επιφανειακών στιβάδων του κερατινοποιημένου επιθηλίου και γενικά οι αλλαγές προς την απολέπιση είναι λιγότερο δραματικές (Hand & Frank, 2014).

Εκτός από κερατινοκύτταρα ο επιθηλιακός ιστός του στοματικού βλεννογόνου περιέχει 3 άλλα είδη κυττάρων που ονομάζονται συνολικά ως μη-κερατινοκύτταρα (non-keratinocytes) τα οποία αποτελούν το 10% περίπου των κυττάρων του επιθηλίου (Chiego & Chiedo, 2013). Τα κύτταρα αυτά εμφανίζονται ως διαυγή κύτταρα στο οπτικό μικροσκόπιο, δηλαδή δεν χρωματίζονται με την χρώση αιματοξυλίνης/ηωσίνης και με αυτόν τον τρόπο ξεχωρίζουν από τα υπόλοιπα κερατινοκύτταρα (Prasad & Anuthama, 2019). Αυτά είναι τα:

- **Μελανοκύτταρα** : Είναι κύτταρα με δενδριτικές αποφυάδες, οι οποίες όμως δεν φαίνονται στο οπτικό μικροσκόπιο με συνήθεις χρώσεις αιματοξυλίνης/ηωσίνης (χρειάζονται ειδικές χρώσεις όπως η χρώση αργύρου), και κάθε ένα μελανοκύτταρο, έρχεται σε επαφή με 30-40 κερατινοκύτταρα μέσω αυτών των αποφυάδων. Η αποστολή τους είναι να παρέχουν κοκκία μελανίνης στα κερατινοκύτταρα που έρχονται σε επαφή. Είναι σημαντικό να πούμε ότι άνθρωποι με σκούρο χρώμα στον βλεννογόνο δεν έχουν μεγαλύτερο αριθμό μελανοκυττάρων σε σχέση με αυτούς που έχουν πιο ανοικτό χρώμα, αλλά τα υπάρχοντα μελανοκύτταρα έχουν σε αυτά τα άτομα μεγαλύτερη δραστηριότητα και παράγουν περισσότερη μελανίνη. Σε περιπτώσεις φλεγμονής στο επιθήλιο είναι δυνατόν η μελανίνη να πάει στο χόριο οπου εκεί φαγοκυτταρώνεται από μακροφάγα κύτταρα δημιουργώντας έτσι στον βλεννογόνο μία μελάγχρωση λόγω φλεγμονής (Αγγελόπουλος και συν., 2000).
- **Κύτταρα Langerhans** : Είναι επίσης κύτταρα με δενδριτικές αποφυάδες που δεν φαίνονται στο οπτικό μικροσκόπιο με συνήθεις χρώσεις (φαίνονται με ειδικές χρώσεις όπως χλωριούχου χρυσού) και βρίσκονται στα ανώτερα στρώματα του στοματικού επιθηλίου. Πρόκειται για κύτταρα του ανοσοποιητικού συστήματος με αντιγονοπαρουσιαστικό ρόλο και ενεργοποιούνται με κυτοκίνες από τα κερατινοκύτταρα όταν αυτά προσβάλλονται από κάποιο αντιγόνο (Hand & Frank, 2014).
- **Κύτταρα Merkel** : Πρόκειται για αισθητικούς μηχανοϋποδοχείς που ανιχνεύουν ερεθίσματα πίεσης και αφής και βρίσκονται συχνά στο μασητικό βλεννογόνο και σπανιότερα στον καλυπτικό. Εντοπίζονται κυρίως στην βασική και υπερβασική

στιβάδα του επιθηλίου και έρχονται σε επαφή με μη ορατές στο οπτικό μικροσκόπιο ελεύθερες νευρικές απολήξεις (Chiego & Chiedo, 2013 ; Kumar, 2015).

Πέρα από τα παραπάνω συχνότερα απαντώμενα μη κερατινοκύτταρα, σποραδικά μπορεί να διακρίνουμε στο επιθήλιο και άλλα κύτταρα του ανοσοποιητικού όπως λεμφοκύτταρα και σπανιότερα πολυμορφοπόρηνα. Αυτά δεν έχουν διαυγές κυτταρόπλασμα και έχουν ακριβώς την ίδια εμφάνιση όπως και στο χόριο και έτσι διακρίνονται εύκολα από τις κύριες κατηγορίες που μόλις αναφέραμε (Chiego & Chiedo, 2013).



Εικόνα 2. Ιστολογικά είδη επιθηλίου. Από αριστερά προς τα δεξιά έχουμε ιστολογικές εικόνες μη κερατινοποιημένου, παρακερατινοποιημένου και όρθοκερατινοποιημένου επιθηλίου. (Εικόνα από το βιβλίο των Hand & Frank, 2014)

1.3 Η Διασύνδεση Επιθηλίου-Συνδετικού Ιστού

Παρατηρώντας με το οπτικό μικροσκόπιο την περιοχή ένωσης του επιθηλίου με τον υποκείμενο συνδετικό ιστό βλέπουμε ότι αυτή έχει ένα κυματοειδές μοτίβο, αυτό δεν είναι τυχαίο γιατί έτσι αυξάνεται η επιφάνεια διεπαφής του συνδετικού ιστού με το επιθήλιο με αποτέλεσμα κατά πρώτον την καλύτερη και πιο στερεή πρόσφυση των 2 ιστών μεταξύ τους

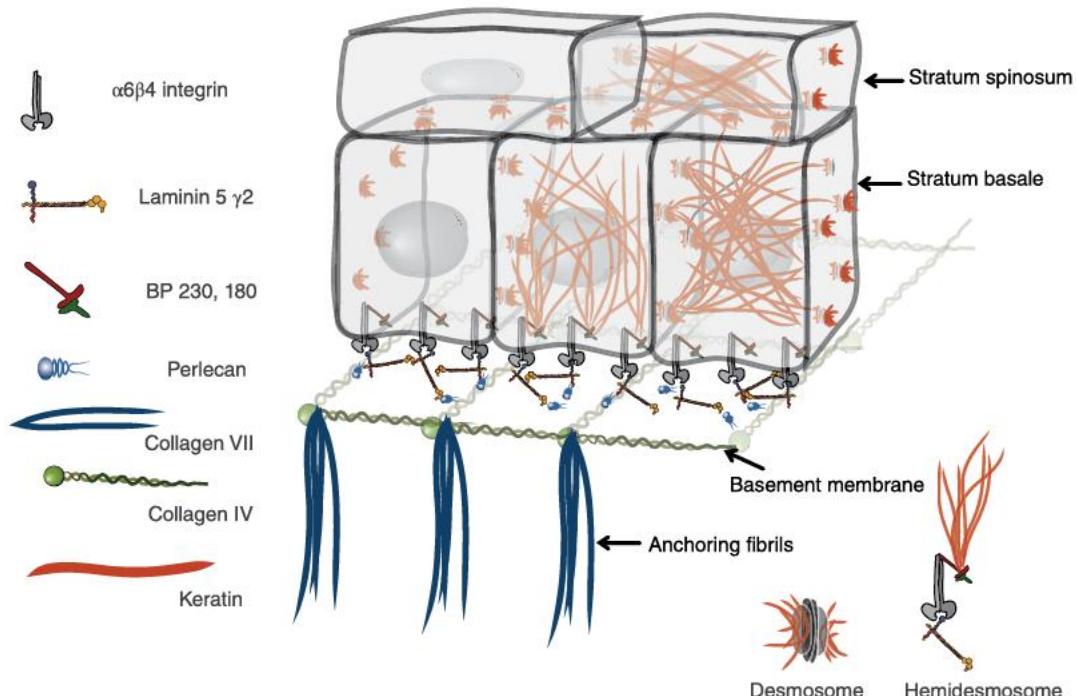
και κατά δεύτερον την καλύτερη διάχυση οξυγόνου και θρεπτικών συστατικών από τον αγγειούμενο συνδετικό ιστό στον μη αγγειούμενο επιθηλιακό ιστό. Τα τμήματα του επιθηλίου που εισέρχονται εντός του συνδετικού ιστού ονομάζονται επιθηλιακές καταδύσεις ή κατ'άλλους επιθηλιακές ακρολοφίες. Ενώ τα τμήματα του χορίου που εισέρχονται εντός του επιθηλίου ονομάζονται θηλές του χορίου (Αγγελόπουλος και συν, 2000 ; Kumar, 2015).

Υπερμικροσκοπικά ο επιθηλιακός ιστός διαχωρίζεται από τον συνδετικό με μία σκληρή μεμβράνη πλούσια σε γλυκοζαμινογλυκάνες και κολλαγόνο τύπου IV, πάχους 1-4μμ που ονομάζεται βασική μεμβράνη ή βασικός υμένας. Πρόκειται για μία ετερογενή ακύτταρη δομή που αποτελείται από 3 υπερμικροσκοπικά διακριτές ζώνες , την φωτεινή (lamina lucida) προς την πλευρά του επιθηλίου, την σκοτεινή (lamina densa) κάτωθεν της φωτεινής και τέλος την δικτυωτή ζώνη (lamina reticularis) που πρόσκειται στο χόριο. Η φωτεινή ζώνη περιέχει κολλαγόνο τύπου IV, πολυσακχαρίτες, το ετεροτριμερές λαμινίνη-V και τα αντιγόνα του πομφολυγώδους πεμφιγοειδούς και έχει πάχος 20-40nm. Η σκοτεινή ζώνη περιέχει ένα δίκτυο από γλυκοζαμινογλυκάνες και θεϊκή ηπαρίνη που συμπλέκεται με το κολλαγόνο τύπου IV, η πρωτεΐνη που κάνει την διασύνδεση αυτή ονομάζεται περλεκάνη (perlecan). Τέλος η δικτυωτή ζώνη αποτελείται κυρίως από φιμπρονεκτίνη και θεωρείται από κάποιους τμήμα του συνδετικού ιστού. Μία σημαντική δομή που παρατηρείται ανάμεσα στην βασική μεμβράνη και τα υπερκείμενα επιθηλιακά κύτταρα είναι τα ημιδεσμοσώματα που εμφανίζονται στο ηλεκτρονικό μικροσκόπιο ως πυκνωτικές σκούρες δομές και συνδέουν την βασική μεμβράνη με την υπερκείμενη βασική στιβάδα του επιθηλίου (Kumar, 2015). Σε αυτό το σημείο θα ήταν σημαντικό να προσέξουμε να μην μπερδεύουμε αυτές τις 2 ξεχωριστές οντότητες, είναι διαφορετικό πράγμα η βασική στιβάδα του επιθηλίου που πρόκειται ουσιαστικά για δομική οντότητα του επιθηλίου και διαφορετικό η βασική μεμβράνη που αποτελεί την δομή διασύνδεσης του επιθηλίου με τον υποκείμενο συνδετικό ιστό.

Τα ημιδεσμοσώματα αποτελούνται από ένα σύνθετο σύμπλεγμα πρωτεΐνών , με κεντρική μια διαμεμβρανική ετεροδιμερή πρωτεΐνη, την ιντεγκρίνη α6β4. Αυτή εντός του κυττάρου συνδέεται με ινίδια ακτίνης και ενδιάμεσα ινίδια κερατίνης σε μία οργανωμένη ενδοκυτταρική δομή την πλάκα πρόσφυσης του ημιδεσμοσώματος, η οποία αποτελείται κυρίως από τα αντιγόνα πομφολυγώδους πεμφιγοειδούς (BP230 και BP180, ονομάζονται εναλλακτικά και κολλαγόνο τύπου XVII) τα οποία πήραν το όνομα τους από την νόσο που

προκαλεί η καταστροφή τους (Αγγελόπουλος και συν., 2000). Η ιντεγρίνη α6β4 εξωκυτταρικά συνδέεται με την λαμινίνη-V της βασικής μεμβράνης.

Η βασική μεμβράνη συνδέεται με το υποκείμενο χόριο μέσω κάθετων σε αυτό ινώδων αγκυλών κολλαγόνου IV που εν συνεχείᾳ διαπερνώνται από παράλληλες ίνες κολλαγόνου I, φτιάχνοντας έτσι ένα ισχυρό πλέγμα που ενισχύει την σύνδεση των 2 αυτών δομών (Hand & Frank, 2014). Τα παραπάνω βασικά δομικά στοιχεία της διασύνδεσης επιθηλίου και συνδετικού ιστού απεικονίζονται στην Εικόνα 3.



Εικόνα 3. Σχηματική απεικόνιση της διασύνδεσης επιθηλίου-συνδετικού ιστού (Από Alberts et al, 2015)

Κεφάλαιο 2. Κυτταρικός Κύκλος και Αρχές Παθοβιολογίας Καρκίνου

2.1 Κυτταρικός Κύκλος

Ο κυτταρικός κύκλος είναι ουσιαστικά τα γεγονότα που συμβαίνουν στον κύκλο ζωής ενός κυττάρου από την στιγμή που αυτό προκύπτει ως θυγατρικό κύτταρο από ένα προγονικό, μέχρι και την στιγμή που αυτό με την σειρά του ως προγονικό κύτταρο διαιρείται σε 2 θυγατρικά κύτταρα (Alberts, Johnson, Lewis, Morgan, Raff, Roberts & Walters, 2015). Η σημασία του είναι θεμελιώδης στην κατανόηση του παθογενετικού μηχανισμού του καρκίνου γιατί αυτό που συμβαίνει κυρίως σε αυτή την περίπτωση είναι μία μεταβολή του τρόπου ρύθμισης του που έχει ως αποτέλεσμα κάποιες φορές την αθανατοποίηση των καρκινικών κυττάρων αλλά και τον ανεξέλεγκτο πολλαπλασιασμό τους.

Πρόκειται γενικά για μία διαδικασία στην οποία συμμετέχουν πολλές πρωτεΐνες και συμπλέγματα πρωτεϊνών και περιλαμβάνει ένα ακόμα περιπλοκότερο σύστημα ελέγχου το οποίο δρα ως ρυθμιστής και ενορχηστρωτής της όλης διαδικασίας. Το σημείο έναρξης του κυτταρικού κύκλου στην βιβλιογραφία αναφέρεται με τον όρο «σημείο Start» όταν αναφερόμαστε σε κατώτερους ευκαρυωτικούς οργανισμούς και ως σημείο περιορισμού (Restriction Point) στα θηλαστικά και στον άνθρωπο. Οι περισσότεροι ερευνητές τοποθετούν το σημείο περιορισμού σε μία χρονική στιγμή λίγο πριν την έναρξη της αντιγραφής του DNA (αλλά όχι ακριβώς στην έναρξη), κατά το οποίο συμβαίνει η συναρμολόγηση των απαιτούμενων μοριακών συμπλεγμάτων που τελικά οδηγούν με αλυσιδωτή αντίδραση (αγγ. cascade) στην έναρξη αναδιπλασιασμού του γενετικού υλικού (Zetterberg, Larsson & Wiman, 1995 ; Johnson & Scottheim, 2013).

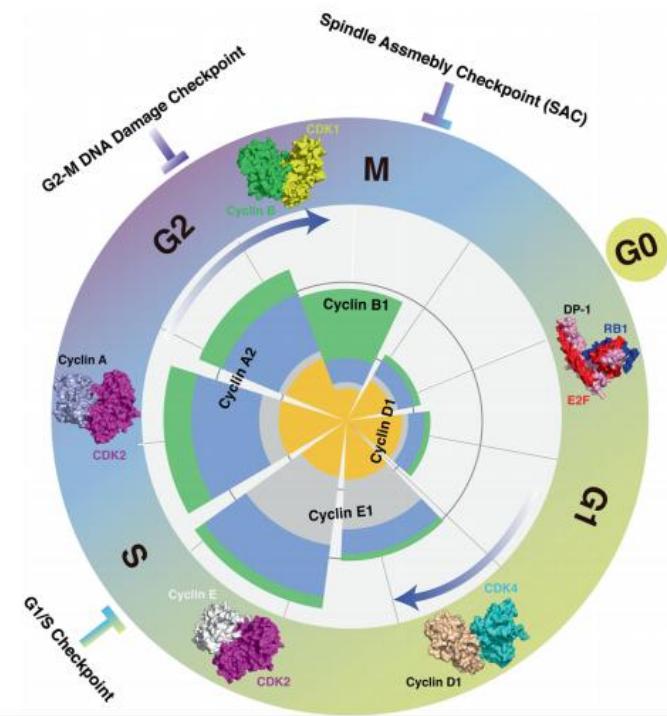
Ο κυτταρικός κύκλος απεικονίζεται σχηματικά στην Εικόνα 4 και περιλαμβάνει 4 φάσεις, οι οποίες είναι με την σειρά αρχίζοντας από το σημείο περιορισμού οι εξής :

1. Φάση S. Φάση κατά την οποία συμβαίνει ο αναδιπλασιασμός του DNA
2. Φάση G2. Πρόκειται για ενδιάμεση φάση πλην όμως αρκετά ενεργή μεταξύ του διπλασιασμού του DNA και της μίτωσης του κυττάρου.

3. Φάση M. Φάση μίτωσης του κυττάρου. Η φάση αυτή διαιρείται σε 4 επιμέρους φάσεις την πρόφαση, την μετάφαση, την ανάφαση και την τελόφαση και είναι η πιο ενεργή φάση του κυτταρικού κύκλου.
4. Φάση G1. Φάση σχετικής ηρεμίας του κυττάρου.

Κάποια κύτταρα (π.χ. νευρικά, μυικά) παραμένουν για πάντα στην φάση G1 η οποία σε αυτή την περίπτωση λέγεται και φάση G0 (Alberts et al, 2015). Από πλευράς μελέτης της βιολογίας του καρκίνου οι πιο σημαντικές φάσεις είναι όλες οι υπόλοιπες φάσεις. Η φάση G2 αν και θεωρείται μία ενδιάμεση φάση ανάμεσα στις κρίσιμες φάσεις S και M, παρόλα αυτά γεγονότα που συβαίνουν σε αυτήν την φάση και έχουν να κάνουν με τον σχηματισμό της μιτωτικής ατράκτου και ιδιαίτερα των μικροσωληνίσκων ατράκτου, από το κεντρόσωμα., φαίνεται να είναι μεγάλης σημασίας. Διαταραχές είτε στον αριθμό κεντροσωμάτων είτε στην λειτουργία τους, που σχετίζεται με την διάταξη και σύνδεση των χρωμοσωμάτων στην μιτωτική άτρακτο, έχουν ενοχοποιηθεί τόσο για την ανευπλοειδία όσο και για την γενετική αστάθεια που συναντάται στην παθοβιολογία του καρκίνου (Levine, Bakker, Boeckx, Moyett, Lu, Vitre & Holland, 2017).

Το πέρασμα του κυττάρου από την μία φάση του κύκλου στην άλλη ελέγχεται από ένα περίπλοκο σύστημα/δίκτυο που περιλαμβάνει συνδυασμό πολλών μορίων και ονομάζεται σύστημα ελέγχου του κυτταρικού κύκλου. Κάποιες λεπτομέρειες σε σχέση με το σύστημα αυτό παραμένουν αδιευκρίνιστες μέχρι σήμερα (Dang, Nie & Wei, 2021) αλλά φαίνεται ότι το παραπάνω σύστημα ελέγχου είναι ζωτικής σημασίας για το κύτταρο, και τυχόν βλάβες σε αυτό έχει μοιραία αποτελέσματα είτε για το κύτταρο (π.χ. προγραμματισμένος κυτταρικός θάνατος) είτε για ολόκληρο τον οργανισμό (π.χ. ανάπτυξη καρκινικών όγκων). Στην παρούσα εργασία θα γίνει μία αρκετά αδρή περιγραφή αυτού του μηχανισμού ελέγχου, γιατί η αναλυτική περιγραφή του θα χρειαζόταν πιθανότατα μία (ή περισσότερες) διπλωματική εργασία από μόνο του.



Εικόνα 4. Κυτταρικός Κύκλος και Σημεία Ελέγχου του (Από Dang et al, 2021)

Οι βασικές μοριακές οντότητες του συστήματος ελέγχου του κυτταρικού κύκλου είναι οι πρωτεΐνικές κινάσες εξαρτώμενες από τις κυκλίνες (Cdk ή CCNs). Οι Cdk έχουν μία σχετικά σταθερή, μη μεταβαλλόμενη συγκέντρωση καθ' όλη την διάρκεια του κυτταρικού κύκλου, όμως αυτές ενεργοποιούνται από τις κυκλίνες που είναι πρωτεΐνες που σχηματίζουν σύμπλοκα με τις Cdk και όπως δείχνει και το όνομα τους η συγκέντρωση τους έχει περιοδική διακύμανση στον χρόνο κατά την διάρκεια του κυτταρικού κύκλου. Στον Πίνακα 1 φαίνονται τα είδη κυκλινών και Cdk καθώς και των συμπλεγμάτων που δημιουργούν, να σημειωθεί ότι το πρώτο συνθετικό στο όνομα του συμπλέγματος είναι η κύρια φάση του κυτταρικού κύκλου στην οποία το συγκεκριμένο εκάστοτε σύμπλεγμα ασκεί ρυθμιστικό έλεγχο (Alberts et al, 2015).

Cdk	Κυκλίνη	Σύμπλεγμα
Cdk4,Cdk6	Κυκλίνες D1,D2,D3	G1-Cdk
Cdk2	Κυκλίνη E	G1/S-Cdk
Cdk2,Cdk1	Κυκλίνη A	S-Cdk
Cdk1	Κυκλίνη B	M-Cdk

Πίνακας 1. Κυκλίνες και συμπλέγματα τους (Προσαρμογή από Alberts et al , 2015)

Το σύστημα ελέγχου του κυτταρικού κύκλου δρα στρατηγικά σε 3 σημεία του κυτταρικού κύκλου τα οποία ονομάζονται στην βιβλιογραφία checkpoints και πιο συγκεκριμένα:

- Στο σημείο περιορισμού μέσω του συμπλέγματος G1/S-Cdk.
- Στο σημείο έναρξης του αναδιπλασιασμού του DNA μέσω του S-Cdk.
- Στο σημείο έναρξης της Μίτωσης του κυττάρου μέσω του M-Cdk.

Το σύμπλεγμα G1-Cdk δε συμμετέχει απευθείας στην ρύθμιση του κυτταρικού κύκλου αλλά έχει περισσότερο ρυθμιστικό ρόλο στην δράση του G1/S-Cdk. Επίσης πρέπει να τονιστεί ότι από μόνο του το εκάστοτε σύμπλεγμα κυκλίνης-Cdk δεν είναι ενεργό αλλά ρυθμίζεται και αλλοστερικά με έναν μηχανισμό φωσφορυλίωσης/αποφωσφορυλίωσης . Πιο συγκεκριμένα ενεργοποιείται από άλλα μόρια όπως η Ενεργοποιητική Κινάση του Cdk (CAK) και η Cdc25 (μέσω αποφωσφορυλίωσης σε ένα αμινοξύ κοντά στο ενεργό κέντρο της Cdk) και απενεργοποιείται από άλλα μόρια όπως είναι οι κινάσες Wee1 και MYT1 (Alberts et al, 2015). Πέρα από τον μηχανισμό φωσφορυλίωσης/αποφωσφορυλίωσης που αναφέρθηκε παραπάνω υπάρχουν και άλλες πρωτεΐνες που έχουν κύριο έργο την αναστολή των συμπλεγμάτων κυκλίνης –Cdk που ονομάζονται συνολικά ανασταλτικοί παράγοντες των Cdk. Κύριοι εκπρόσωποι των τελευταίων είναι μόρια που ανήκουν στις πρωτεΐνικές οικογένειες των Cip/Kip (p21,p27,p57) και INK4 πρωτεϊνών (Fonseca & Bettencourt-Dias, 2019).

Τα 3 αυτά σημεία ελέγχου λειτουργούν ως μοριακοί διακόπτες (on/off) και εφόσον το κύτταρο περάσει ένα σημείο ελέγχου τότε η εκάστοτε κυτταρική διεργασία δεν μπορεί να αναστραφεί ή να σταματήσει μέχρι το επόμενο σημείο ελέγχου (Alberts et al, 2015).

Η συγκέντρωση των κυκλινών εξαρτάται κατά κύριο λόγο από τον μηχανισμό αποδόμησης τους μέσω ουβικούτινωσης και τελικά καταστροφής τους στο πρωτεόσωμα. Η σύνδεση της

ουβικουιτίνης με τις κυκλίνες ρυθμίζεται από άλλα ρυθμιστικά μόρια τις Ε3 λιγάσες της ουβικουιτίνης. Ίσως το σπουδαιότερο από αυτά είναι το σύμπλεγμα προώθησης του κυττάρου στην ανάφαση (APC/C) το οποίο μεταξύ άλλων συμμετέχει και σε άλλες λειτουργίες του κυτταρικού κύκλου όπως η καταστροφή της survivin, μίας πρωτεΐνης που συγκρατεί τα διπλασιασμένα χρωμοσώματα μεταξύ τους πριν τον διαχωρισμό τους στην μίτωση. Ένα άλλο σημαντικό μόριο με παρόμοια λειτουργία είναι και το πρωτεΐνικό σύμπλοκο Skp1-Cul1-F-box (SCF) (Dang et al, 2021).

Τα μόρια αυτά που αναφέρθηκαν ότι ρυθμίζουν την δράση των συμπλεγμάτων κυκλίνης-Cdk ρυθμίζονται και αυτά με την σειρά τους από άλλα μόρια και σηματοδοτικά σήματα (λιγάση CSF,Cdc20,Cdh1) δημιουργώντας πολλαπλά επίπεδα ελέγχου που λειτουργούν ως πολλές δικλείδες ασφαλείας (Albert et al 2015).

Αντιλαμβανόμαστε λοιπόν ότι η ρύθμιση του κυτταρικού κύκλου είναι πολυεπίπεδη , πράγμα που αποκαλύπτει και το γιατί η λεπτομερής περιγραφή αυτού του δαιδαλώδους μηχανισμού ξεφεύγει από τα όρια της παρούσας ενότητας που έχει ως σκοπό κυρίως να δώσει μία γενική περιγραφή ώστε να κατανοηθεί σε επόμενη ενότητα ο παθοβιολογικός μηχανισμός του καρκίνου.

2.2 Προγραμματισμένος Κυτταρικός Θάνατος

Ο προγραμματισμένος κυτταρικός θάνατος είναι ένας περίπλοκος κυτταρικός μηχανισμός, ο οποίος διατηρείται ανενεργός καθ' όλη την διάρκεια ζωής του κυττάρου, με τον οποίο το κύτταρο κατά κάποιον τρόπο αυτοκτονεί μέσω διαφόρων τρόπων όπως με αποδόμηση των πρωτεΐνών του κυτταροσκελετού, αποδόμηση πυρηνικής μεμβράνης κ.α. Αυτός ο μηχανισμός μπορεί να ενεργοποιηθεί είτε ως απάντηση σε κάποιο εξωκυτταρικό σήμα (εξωγενής οδός) είτε χωρίς να υπάρξει ερέθισμα (ενδογενής οδός). Οι οδοί αυτοί δεν φαίνεται να είναι διακριτοί δεδομένου ότι ένα εξωκυτταρικό σήμα μπορεί να προκαλέσει ενεργοποίηση και της ενδογενούς οδού (Alberts et al, 2015). Υπάρχουν 3 είδη προγραμματισμένου κυτταρικού θανάτου :

- Τύπος I ή απόπτωση
- Τύπος II ή αυτοφαγία και
- Τύπος III ή προγραμματισμένη νέκρωση

Από τους 3 παραπάνω τύπους αυτός που μας ενδιαφέρει στην μελέτη της παθοβιολογίας του καρκίνου είναι ο Τύπος I. Οι τύποι II και III συναντώνται σε άλλες παθολογικές καταστάσεις όπως η ασιτία και η απόκριση σε ιογενή λοίμωξη (Sun & Peng, 2009).

Οι κύριες πρωτείνες που πραγματοποιούν τον προγραμματισμένο κυτταρικό θάνατο είναι οι κασπάσες (Casp). Αυτές βρίσκονται στο κύτταρο σε ανενεργή μορφή (προκασπάσες) και μέσω κατάλληλων εξωγενών ή ενδογενών ερεθισμάτων, ενεργοποιούνται με πρωτεόλυση. Οι κασπάσες όταν ενεργοποιηθούν ουσιαστικά δρουν ως πρωτεολυτικά ένζυμα διασπώντας πρωτεΐνες μέσω μίας κυστεΐνης στο ενεργό κέντρο τους με ένα κατάλοιπο ασπαραγινικού οξέος στην πρωτεΐνη-στόχο. Η ενεργοποίηση έστω και μίας κασπάσης πυροδοτεί μία αλυσιδωτή ενεργοποίηση πολλαπλών άλλων προκασπασών σε κασπάσες δημιουργώντας έναν μηχανισμό θετικής ανατροφοδότησης. Όταν ενεργοποιηθεί ο παραπάνω μηχανισμός τότε έχουμε μη αντιστρεπτή πορεία του κυττάρου προς τον θάνατο (Alberts et al, 2015).

Η εξωγενής οδός πραγματοποιείται μεσω διαμεμβρανικών ομοτριμερών υποδοχέων της οικογένειας των παραγόντων νέκρωσης όγκων (TNF) με κυριότερο εκπρόσωπο την πρωτεΐνη Fas. Μετά από κατάλληλο εξωγενές σήμα έχουμε σύνδεση της ενδοκυτταρικής περιοχής του TNF με μία πρωτεΐνη προσαρμογής (Adaptor protein) με κυριότερο εκπρόσωπο την πρωτεΐνη FADD, η οποία συνδέεται εν συνεχείᾳ με την Κασπάση-8 φτιάχνοντας ένα σύμπλεγμα που ονομάζεται DISC (Kauffman, Straser & Jost, 2012). Η ενδογενής οδός συμβαίνει μέσω απελευθέρωσης του κυτοχρώματος c, από τα μιτοχόνδρια στο κυτταρόπλασμα. Εκεί ενώνεται με την πρωτεΐνη προσαρμογής Apaf1 η οποία εν συνεχείᾳ πολυμερίζεται και φτιάχνει ένα θανατηφόρο πολυμερές το «αποπτώσωμα» (apoptosom) που ενεργοποιεί την Κασπάση-9. Ακολουθεί αλυσιδωτή ενεργοποίηση πολλαπλών Κασπασών με πρωτεόλυση και με τον μηχανισμό αυτόν θετικής ανατροφοδότησης επέρχεται ο κυτταρικός θάνατος (Dorstyn, Akey & Kumar, 2018).

Η απόπτωση ρυθμίζεται από μία οικογένεια πρωτεΐνων που ονομάζονται Bcl-2. Κάποιες από αυτές έχουν ενεργοποιητικό (προ-αποπτωτικές πρωτεΐνες) και κάποιες ανασταλτικό ρόλο στην έναρξη της απόπτωσης. Οι κυριότερες προ-αποπτωτικές πρωτεΐνες είναι οι Bax και Bak ενώ οι κυριότερες ανασταλτικές πρωτεΐνες είναι η Bcl-2 και η Bcl-xL. Ο σύνδεσμος αυτών των 2 κατηγοριών είναι οι BH3-only πρωτεΐνες, οι οποίες έχουν την ικανότητα να απενεργοποιούν τις ανασταλτικές Bcl-2 πρωτεΐνες δίνοντας έτσι στις ενεργοποιητικές την

ικανότητα να προκαλέσουν την απελευθέρωση του κυτοχρώματος c από τα μιτοχόνδρια προς το κυτταρόπλασμα (Alberts et al, 2015).

Σε περίπτωση βλάβης του DNA τότε ενεργοποιείται ένας περίπλοκος μηχανισμός επιδιόρθωσης του DNA στα πλαίσια του οποίου εμπλέκεται ο γνωστός ογκοκατασταλτικός παράγοντας p53 ο οποίος προκαλεί μεταγραφή των γονιδίων προγραμματισμένου κυτταρικού θανάτου Puma και Noxa, που είναι BH3-only πρωτεΐνες, με τελικό αποτέλεσμα την επικράτηση των προ-αποπτωτικών πρωτεϊνών και την θανάτωση του κυττάρου πριν αυτό εξελιχθεί σε καρκινικό. Αυτό εξηγεί εν μέρει γιατί όταν υπάρχουν μεταλλάξεις στο γονίδιο του p53, αυτό μπορεί να οδηγήσει σε ένα κύτταρο που είναι πλέον αθάνατο και κατ' επέκταση δυνητικά καρκινογόνο (Park, Jeong & Jang, 2012).

2.3 Αρχές Παθοβιολογίας Καρκίνου

Η ανάπτυξη καρκίνου σε έναν οργανισμό φαίνεται να είναι προϊόν συνδυασμού πολλών μεταλλάξεων αλλά και πιο εκτεταμένων γενετικών ανακατατάξεων στον καρυότυπο που προκύπτουν ως αποτέλεσμα μίας κατάστασης γενετικής αστάθειας στην οποία περιέπεσε το κύτταρο.

Οι καρυοτυπικές μεταβολές μπορεί να συνίστανται σε διπλασιασμούς γενετικών περιοχών ή βραχιόνων χρωμοσωμάτων ή και των χρωμοσωμάτων ολόκληρων κάποιες φορές (ανευπλοειδία) προκαλώντας με αυτόν τον τρόπο ενίσχυση και αύξηση της έκφρασης των γονιδίων που υπάρχουν σε αυτές. Επίσης μπορεί να παρουσιαστούν απαλοιφές (deletions) προκαλώντας μείωση της έκφρασης ή και απουσία συγκεκριμένων γονιδίων στο κύτταρο. Τέλος μπορεί να υπάρξουν και μεταθέσεις γενετικών περιοχών με αποτέλεσμα την δημιουργία νέων υβριδικών γονιδίων με πιθανή καρκινογόνο δράση όπως συμβαίνει π.χ. στην χρόνια μυελοειδή λευχαιμία με το χρωμόσωμα της Φιλαδέλφειας (Duesberg, Fabarius & Hehlmann, 2005).

Επίσης η ύπαρξη και συσσώρευση μη-συνόνυμων μεταλλάξεων σταδιακά κάνουν κρίσιμα γονίδια του κυτταρικού κύκλου να χάνουν την λειτουργία τους και να καθίσταται ανενεργά ή να υπολειτουργούν, ή αντιστρόφως να χάνουν την λειτουργία τους οι άμεσοι ή έμμεσοι αναστολείς αυτών των γονιδίων με αποτέλεσμα την απώλεια των σημείων checkpoint στον κυτταρικό κύκλο. Αυτό εξηγεί εν μέρει γιατί συχνότερα ο καρκίνος ως ασθένεια χτυπά

ανθρώπους μεγαλύτερης ηλικίας στους οποίους έχει περάσει ένα χρονικό διάστημα συσσώρευσης γονιδιακών δυσλειτουργιών και μεταλλάξεων (Alberts et al, 2015).

Πολλά από τα παθολογικά αυτά κύτταρα αντιμετωπίζονται από το ανοσοποιητικό σύστημα του οργανισμού. Όμως αν κάποιο αποσυντονισμένο κύτταρο αναπτύξει μηχανισμούς που το κάνουν μη ανιχνεύσιμο από το ανοσοποιητικό σύστημα τότε μπορεί να αναπτυχθεί καρκινογένεση. Τέτοιοι μηχανισμοί εκμεταλλεύονται φυσιολογικά σημεία ελέγχου του ανοσολογικού συστήματος που έχουν να κάνουν με την αποτροπή αυτοάνοσων νοσημάτων (Povoa & Fior, 2019). Τέτοια ανοσολογικά σημεία ελέγχου είναι 2 αναστολείς του υποδοχέα B7 (CD28); το CTLA-4, που εκκρίνεται φυσιολογικά από τα T-reg κύτταρα και ο υποδοχέας PD-1 που βρίσκεται κυρίως στα αντιγονοπαρουσιαστικά κύτταρα. Και τα 2 αυτά μόρια έχουν γίνει αντικείμενο μελέτης για την κατασκευή στοχευμένων αντικαρκινικών θεραπειών (Seidel, Otsuka & Kabashima, 2018).

Μία από τις πρώτες υποθέσεις που αναπτύχθηκαν ήταν αυτή του διπλού χτυπήματος. Γνωρίζουμε ότι κάθε γονίδιο υπάρχει σε 2 αντίγραφα σε όλα τα σωματικά κύτταρα, η θεωρία του διπλού χτυπήματος υποστηρίζει ότι αρχικά γίνεται κάποια μετάλλαξη που προκαλεί λειτουργική διαταραχή στο ένα αντίγραφο, η οποία όμως δεν διαφαίνεται στον φαινότυπο σε αυτό το στάδιο, διότι η λειτουργική διαταραχή αντισταθμίζεται από το 2^o λειτουργικό αντίγραφο. Σε δεύτερο χρόνο γίνεται μία μετάλλαξη και στο 2^o λειτουργικό αντίγραφο το οποίο με την σειρά του καθίσταται και αυτό μη λειτουργικό, με αποτέλεσμα πλέον για το κύτταρο ολική απώλεια του συγκεκριμένου γονιδίου. Άν το διπλό χτύπημα αφορά ένα κρίσιμο γονίδιο για τον κυτταρικό κύκλο, τότε είναι δυνατόν να προκύψει δυσλειτουργία του, με πιθανή έκβαση τον ανεξέλεγκτο πολλαπλασιασμό κυττάρων. Τα 2 χτυπήματα μπορούν να μοντελοποιηθούν με κατανομή Poisson (Knudson, 1971). Η θεωρία του διπλού χτυπήματος δεν έχει ανατραπεί μέχρι σήμερα, και σε σύγχρονη επιστημονική ορολογία ονομάζεται απώλεια ετεροζυγωτίας (Loss of Heterozygosity).

Τα γονίδια που πλήγονται στον καρκίνο μπορεί να ανήκουν σε πολλές κατηγορίες όμως οι κυριότερες είναι τα ογκογονίδια (που προκύπτουν από μεταλλάξεις στα πρωτο-ογκογονίδια) και τα ογκοκατασταλτικά γονίδια (όπως το p53) που έχουν ως στόχο την επιδιόρθωση του ελλατωματικού DNA όταν αυτό υποστεί κάποια βλάβη ή όταν αυτό δεν μπορεί να γίνει, την ενεργοποίηση του μηχανισμού της απόπτωσης που αναφέρθηκε με λεπτομέρεια στο προηγούμενο κεφάλαιο. Ένας άλλος μηχανισμός που συχνά αναφέρεται στην βιβλιογραφία

είναι ότι τα κύτταρα γίνονται «αθάνατα» δηλαδή μπορούν να πολλαπλασιάζονται για πάντα, μηχανισμός που κυρίως αποδίδεται σε ενεργοποίηση των τελομερασών, με αποτέλεσμα το μοριακό ρολόι της μείωσης των τελομερών των χρωμοσωμάτων να μην σταματά ποτέ πλέον αλλά να είναι αιώνιο (Carlos, 2019).

Το κυριότερο και πιο δυσεπίλυτο πρόβλημα που συμβαίνει στον καρκίνο είναι ότι το κύτταρο οδηγείται σε μία κατάσταση ολοένα και αυξανόμενου ρυθμού εμφάνισης μεταλλάξεων, που στην βιβλιογραφία ονομάζεται **κυτταρική αστάθεια** (Duijf, Nanayakkara, Nones, Srihari, Kalimutho & Khanna, 2019). Αυτή συνίσταται στην αδυναμία του κυττάρου πλέον να επιδιορθώσει με την ίδια αποτελεσματικότητα τις μεταλλάξεις που εμφανίζονται στο DNA κατά την διάρκεια ζωής του. Αυτό λειτουργεί με θετική συνέργεια σε συνδυασμό με το γεγονός ότι όταν το κύτταρο βρίσκεται σε κυτταρική αστάθεια, το σύμπλεγμα της DNA πολυμεράσης κάνει συχνότερα λάθη κατά την αντιγραφή του γενετικού υλικού. Τέλος το τρίτο χαρακτηριστικό της κυτταρικής αστάθειας είναι η αύξηση του ρυθμού που συμβαίνουν γενετικές ανακατατάξεις του καρυοτύπου (δηλαδή ανακατατάξεις από χρωμόσωμα σε χρωμόσωμα) με περισσότερες μεταθέσεις, αντιστροφές, διπλασιασμούς γενετικών περιοχών αλλά και διαγραφές αυτών (Alberts et al, 2015).

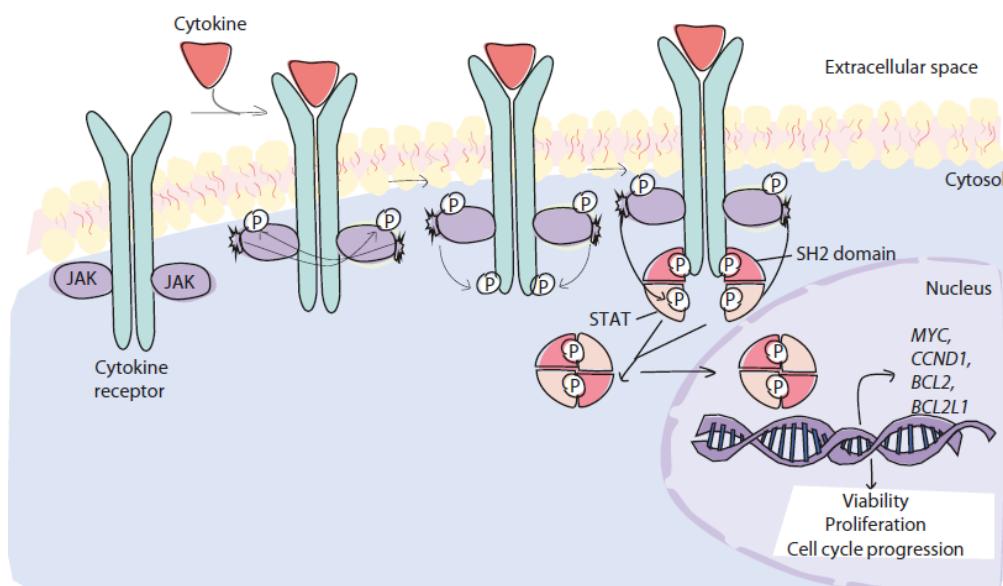
Κατά την ανάπτυξη μίας κακοήθους νεοπλασίας, επηρεάζονται σε όλες τις περιπτώσεις κρίσιμα σηματοδοτικά βιολογικά μονοπάτια τα οποία σε μεγάλο βαθμό είναι γνωστά. Αυτά που επηρεάζονται στα περισσότερα είδη καρκίνων είναι:

- το μονοπάτι JAK/STAT (Janus Kinase/signal transducers and activators of transcription)
- το μονοπάτι MAPK στο οποίο σημαντικότερο ρόλο στον καρκίνο παίζει η δυσλειτουργία στον κλάδο Ras/MEK/Erk (ή αλλιώς classical extracellular signal-regulated kinase pathway).
- Το μονοπάτι των PI3K/Akt/mTOR ή πιο απλά μονοπάτι mTOR (mammalian target of rapamycin).

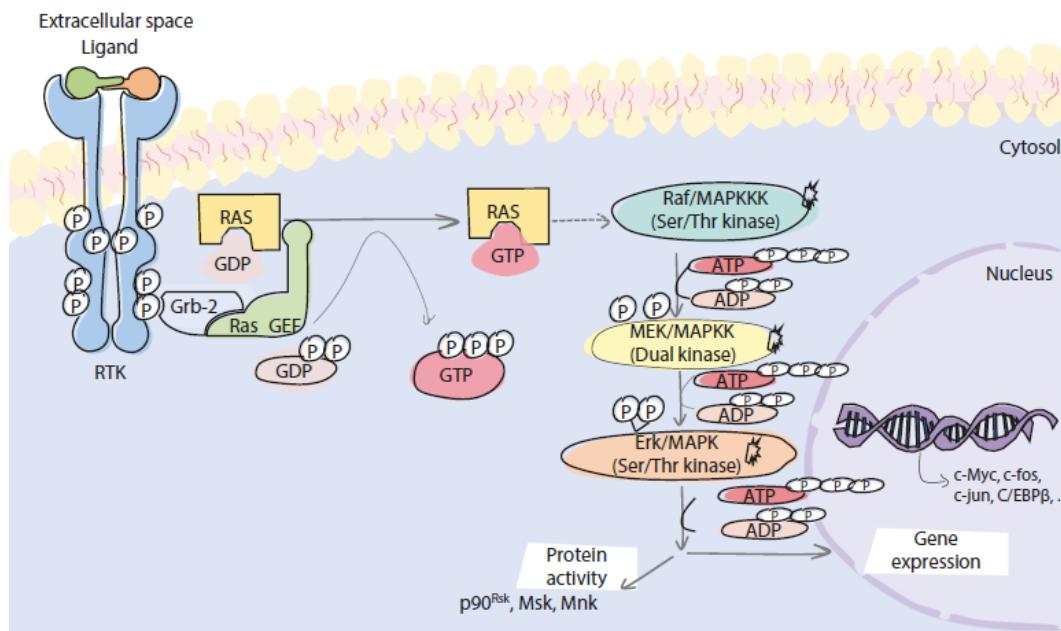
Τα παραπάνω μονοπάτια επηρεάζουν την απόπτωση, την διαφοροποίηση και τον μεταβολισμό του κυττάρου. Είναι δηλαδή μονοπάτια με ευρύ αντίκτυπο στις διάφορες εκφάνσεις της κυτταρικής λειτουργίας και αποτελούν μοριακούς στόχους πολλών νέων στοχευμένων θεραπειών κατά του καρκίνου χρησιμοποιούμενα κατά κύριο λόγο σε

αιματολογικές κακοήθειες αλλά και σε άλλα κακοήθη νοσήματα (Barata & Oliveira, 2019). Άλλα μονοπάτια που επίσης φαίνεται να παίζουν ρόλο είναι τα Wnt, NOTCH και HedgeHog σηματοδοτικά μονοπάτια (Takebe, Miele, Harris, Jeong, Bando, Kahn, ... & Ivy, 2015). Η λεπτομερής επεξήγηση όλων των παραπάνω μονοπατιών είναι εξαιρετικά περίπλοκη και ξεφεύγει μακράν των στόχων αυτού του γενικού μέρους. Μία στοιχειώδης απεικόνιση των κυριότερων μορίων και συμπλόκων που συμμετέχουν στα 3 πρώτα αναφερόμενα μονοπάτια παρατίθεται στις εικόνες 5,6 και 7.

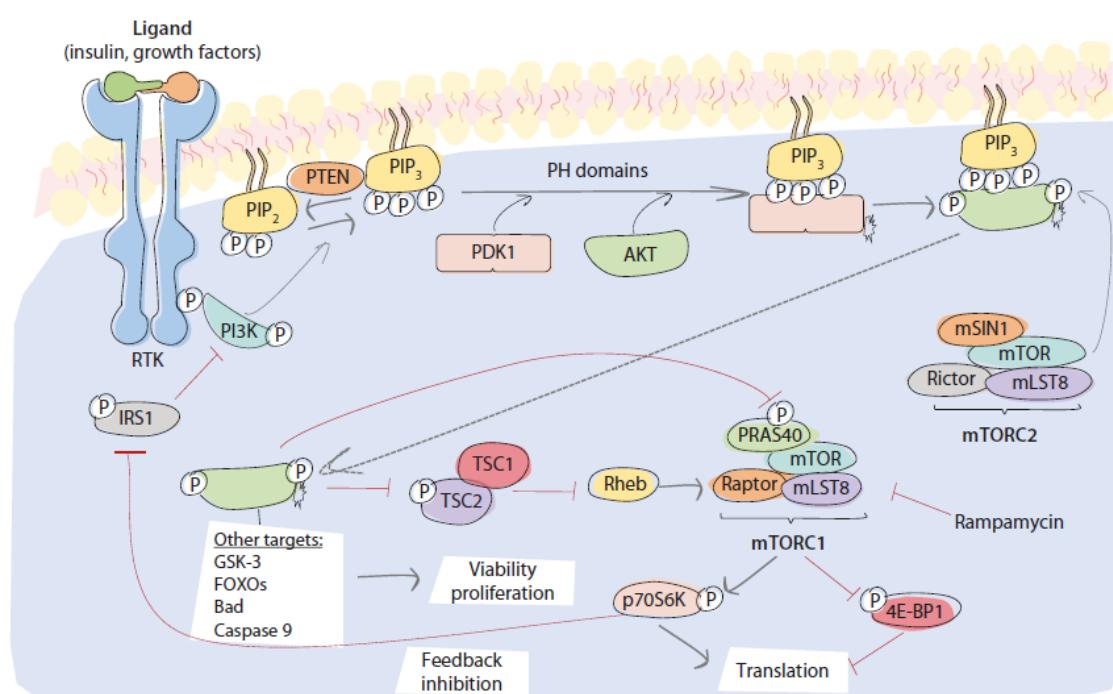
Τα τελευταία χρόνια έχει αναδειχθεί ο ρόλος ενός παράγοντα που παλαιότερα θεωρούνταν να έχει ένα παθητικό ρόλο στην όλη καρκινική εξεργασία. Αυτός είναι το καρκινικό μικροπεριβάλλον ή μικροπεριβάλλον του όγκου. Ο καρκινικός όγκος δεν αποτελείται μόνο από ένα είδος κυττάρου αλλά πρόκειται για ένα ετερογενές κυτταρικό σύνολο. Αποτελείται εκτός από τα καθ'εαυτού καρκινικά κύτταρα και από κύτταρα του ανοσοποιητικού συστήματος τα οποία σε κάποιες περιπτώσεις βοηθούν τον όγκο και τον ενισχύουν με διάφορους μηχανισμούς. Επίσης στο μικροπεριβάλλον όγκων επιθηλιακής προέλευσης όπως το AKK είναι δυνατόν να ανευρίσκονται και κύτταρα του συνδετικού οστού (όπως οι CAFs) αλλά και ενδοθηλιακά κύτταρα που παίζουν καίριο ρόλο στην καρκινική αγγειογένεση που συμβάλει στην ομαλή θρέψη του όγκου όταν αυτός ξεπεράσει ένα συγκεκριμένο μέγεθος (Zilhao & Neves, 2019).



Εικόνα 5. Το Μονοπάτι JAK/STATS. (Από Barata & Oliveira , 2019)



Εικόνα 6. Το μονοπάτι Ras/MEK/Erk (Από Barata & Oliveira , 2019)



Εικόνα 7. Το μονοπάτι PI3K/Akt/mTOR (Από Barata & Oliveira , 2019)

Κεφάλαιο 3. Ακανθοκυτταρικό Καρκίνωμα του στόματος

3.1 Γενικά Στοιχεία

Το ΑΚΚ είναι μία κακοήθης νεοπλασία που όπως αναφέρει το όνομα του ανήκει στην κατηγορία των καρκινωμάτων που σημαίνει ότι προέρχεται από τον επιθηλιακό ιστό. Ωστόσο αν και το όνομα του υποδηλώνει προέλευση από τα κύτταρα της ακανθωτής στιβάδας, το καρκίνωμα αυτό ξεκινά από την βασική στιβάδα του επιθηλίου. Στα αγγλικά ονομάζεται “squamous cell carcinoma” (Αγγελόπουλος και συν., 2000).

Πρόκειται για το πιο συχνό ιστολογικό τύπο κακοήθειας στην περιοχή κεφαλής και τραχήλου με ποσοστό κοντά στο 90% και συγκαταλέγεται στα πρώτα 10 συχνότερα είδη καρκίνου στον κόσμο (Scully, 2012). Σύμφωνα με στοιχεία του υπουργείου υγείας των ΗΠΑ, η επίπτωση του καρκίνου του στόματος και φάρυγγα ήταν για το 2017, 11.68 ανά 100.000 κατοίκους, και οι άντρες είχαν σχεδόν τριπλάσια επίπτωση (17.28/100.000) σε σχέση με τις γυναίκες (6.76/100.000). Από τα παραπάνω ποσοστό 78.5%, αφορά σε ιστολογική διάγνωση ΑΚΚ. Το ΑΚΚ ως ιστολογικός τύπος εντοπίζεται συχνά εκτός από το στόμα και σε άλλες ανατομικές περιοχές της κεφαλής και του τραχήλου όπως στον φάρυγγα, στις αμυγδαλές, στον λάρυγγα, στον οισοφάγο και στο δέρμα. Το ποσοστό των ΑΚΚ που εντοπίζονται στο στόμα κυμαίνεται περίπου στο 30% (SEER Preliminary cancer incidence rate estimates, 2017).

Το ΑΚΚ προσβάλλει συνήθως άντρες και η επίπτωση του αυξάνεται με την ηλικία. Ωστόσο τα τελευταία χρόνια φαίνεται μία αυξανόμενη τάση και σε μικρότερες ηλικίες. Γεωγραφικά υπάρχουν σημαντικές διαφορές στην κατανομή κρουσμάτων από χώρα σε χώρα. Οι χώρες της νοτιοανατολικής Ασίας και η Βραζιλία φαίνεται να έχουν την μεγαλύτερη συχνότητα του συγκεκριμένου καρκίνου (Neville et al, 2008).

Η αιτιολογία του συγκεκριμένου όγκου όπως συμβαίνει και για τα περισσότερα είδη καρκίνου φαίνεται να είναι πολύ-παραγοντική αλλά έχουν αναγνωριστεί κάποιοι παράγοντες κινδύνου. Τον κυριότερο ρόλο φαίνεται ότι παίζει η χρήση διαφόρων μορφών καπνού ακολουθούμενη σε λιγότερο βαθμό από την κατάχρηση αλκοόλ. Φαίνεται ότι σε περιπτώσεις που έχουμε ταυτόχρονα κατάχρηση αλκοόλ και καπνού υπάρχει θετική

συνέργεια στην εκδήλωση καρκίνου στο στόμα, αλλά και γενικότερη ανάπτυξη καρκινωμάτων στην ευρύτερη περιοχή κεφαλής και τραχήλου (Neville et al, 2008). Ο καπνός γενικά περιέχει γύρω στα 60 καρκινογόνα ανάμεσα τους το βενζοπυρένιο (ένας πολύ-κυκλικός υδρογονάνθρακας) και οι νικοτινικές νιτροζαμίνες. Η μεγαλύτερη συχνότητα AKK στην κάτω επιφάνεια και τα πλάγια χείλη της γλώσσας αλλά και στο έδαφος του στόματος φαίνεται να σχετίζεται με την λίμναση των καρκινογόνων προϊόντων του καπνού μέσω του σάλιου και τις βαρύτητας σε αυτές τις περιοχές (Warnakulasuriya & Tilakaratna, 2013). Ένα είδος άκαπνου καρπού το areca (ή betel quid), που χρησιμοποιείται σε χώρες της 'Απω Ανατολής, έχει επίσης αποδειχθεί ότι προκαλεί AKK αλλά και μία παραλλαγή του, που ονομάζεται ακροχορδονώδες καρκίνωμα του στόματος (Κυρές, Κορώνης, Τόσιος, Νικητάκης & Σκλαβούνου, 2008). Η χρήση του areca γίνεται με την τοποθέτηση του καρπού στην ουλοπαρειακή αύλακα, με αποτέλεσμα ο καρκίνος σε αυτή την περίπτωση να εκδηλώνεται σε μεγαλύτερη ποσοστό στην παρειά (Neville et al, 2008).

Το αλκοόλ μεταβολίζεται σε ένα γνωστό καρκινογόνο την ακεταλδεύδη. Θεωρείται ότι η παραγωγή ακεταλδεύδης είναι μεγαλύτερη σε άτομα με κακή στοματική υγιεινή. Αυτό οφείλεται σε αυξημένο μεταβολισμό της αλκοόλης από τα βακτήρια της οδοντικής μικροβιακής πλάκας του στόματος, παραπροϊόν του οποίου είναι η ακεταλδεύδη. Ένας άλλος τρόπος με τον οποίον επιδρά η αλκοόλη στην καρκινογένεση του στόματος θεωρείται ότι μπορεί να είναι και η αφυδάτωση του στοματικού επιθηλίου με αποτέλεσμα την ευκολότερη είσοδο μέσα σε αυτό των καρκινογόνων παραγόντων του καπνού (Warnakulasuriya & Tilakaratna, 2013).

Ένας άλλος αρκετά σημαντικός παράγοντας ανάπτυξης AKK που έχει απασχολήσει πολύ την βιβλιογραφία είναι η προσβολή από ιούς HPV (το 90% των συσχετισμών αφορά το στέλεχος 16, και ένα μεγάλο μέρος των υπολοίπων το στέλεχος 18) η οποία φαίνεται να παίζει έναν σημαντικό ρόλο ιδιαίτερα στην ανάπτυξη AKK σε νεότερες από το μέσο όρο ηλικίες (Scully, 2012). Η HPV θετικότητα του όγκου σχετίζεται με καλύτερη πρόγνωση αλλά και καλύτερη ανταπόκριση στην ακτινοθεραπεία (Panarese, Aquino, Ronchi, Longo, Montella, Cozzolino, ... & Franco, 2019).

Άλλοι προδιαθεσικοί παράγοντες μπορεί να είναι, χωρίς ωστόσο επαρκή στοιχεία και βιβλιογραφικά δεδομένα η ανεπάρκεια βιταμίνης A και ρετινοϊκού οξέως, η ανοσοκαταστολή, η ανεπάρκεια σιδήρου (Ravikiran & Praveen, 2014), το ιστορικό

θεραπευτικής ακτινοβόλησης, οι φαινολικοί παράγονες, η τριτογενής σύφιλη και τέλος η υπερπλαστική καντιντίαση (Neville et al, 2008).

Αρκετές φορές πριν την εμφάνιση ενός ιστολογικά επιβεβαιωμένου ΑΚΚ, εμφανίζονται στο στόμα κάποιες βλάβες που ονομάζονται προ-καρκινικές, ακριβώς γιατί η παρουσία τους αυξάνει την πιθανότητα καρκίνου στην περιοχή εντόπισης τους. Με άλλα λόγια ο καρκίνος συχνά αναπτύσσεται σε έδαφος αυτών των προ-καρκινικών βλαβών, αυτές είναι με σειρά συχνότητας οι εξής:

- Ερυθροπλακία
- Λευκοπλακία
- Ακτινική Χειλίτιδα
- Υποβλεννογόνια ίνωση
- Ομαλός Λειχήνας

Κλινικά το ΑΚΚ εμφανίζεται συνήθως σε έδαφος δυσπλαστικού επιθηλίου, συχνά ως μετεξέλιξη κάποιας από τις παραπάνω προκαρκινικές βλάβες του στοματικού βλεννογόνου (Αγγελόπουλος και συν, 2000). Το ΑΚΚ του στόματος δεν έχει ομοιόμορφη κλινική εικόνα. Μπορεί να εμφανιστεί είτε ως ενδοφυτικός όγκος (με την μορφή έλκους) είτε ως εξωφυτικός (με την μορφή μάζας). Στην Εικόνα 8 απεικονίζεται ένα ενδοφυτικό ΑΚΚ με εντόπιση στην γλώσσα. Στα αρχικά στάδια και εφόσον δεν έρχεται σε επαφή με μείζονα νευρικά στελέχη είναι ανώδυνος και είναι σύνηθες οι ασθενείς να αναζητούν ιατρική βοήθεια μετά από αρκετούς μήνες ή και χρόνια από την έναρξη της νόσου. Αυτό έχει ως αποτέλεσμα οι μισοί ασθενείς κατά την στιγμή της τελικής διάγνωσης να έχουν ήδη κάνει μετάσταση σε επιχώριους λεμφαδένες, ενώ δυστυχώς ένας στους 10 έχει ήδη δώσει απομακρυσμένη μετάσταση (Scully, 2012).

Η πιο συχνή θέση εκδήλωσης ΑΚΚ στο στόμα είναι η γλώσσα με ποσοστό 45-50% και ακολουθεί το έδαφος του στόματος με ποσοστό 35% (Scully, 2012 ; Neville et al, 2008). Όταν αυτός ο καρκίνος δίνει απομακρυσμένη μετάσταση συνήθως την κάνει στους πνεύμονες και ακολουθούν το ήπαρ και ο εγκέφαλος, να σημειωθεί όμως ότι είναι δυνατόν να εμφανιστεί μετάσταση από ΑΚΚ σε οποιαδήποτε περιοχή του σώματος (Neville et al, 2008). Ο όγκος αυτός έχει σκληρή σύσταση; όταν έχει δώσει μετάσταση σε επιχώριους λεμφαδένες αυτοί εμφανίζονται ως ακινητοποιημένοι στους υποκείμενους σε αυτούς ιστούς

(χωρίς ωστόσο αυτό να είναι απόλυτα απαραίτητο), ανώδυνοι και σκληρής σύστασης, σε αντίθεση με την λεμφαδενίτιδα λόγω κάποιας φλεγμονής όπου είναι ευκίνητοι, εμφανίζουν ευαισθησία στην ψηλάφηση και με μαλακή σύσταση. Η τελική διάγνωση τίθεται πάντοτε με βιοψία και ιστολογική εξέταση με τις συνήθεις χρώσεις αιμοτοξυλίνης/ηωσίνης. Σε περίπτωση αρκετά αναπλαστικών όγκων μπορούν να χρησιμοποιηθούν ανοσοϊστοχημικές τεχνικές για να προσδιοριστεί ο προέλευση του όγκου και να αποκλειστεί το ενδεχόμενο ο υπό εξέταση όγκος να είναι μεταστατικός όγκος από με πρωτογενή εστία σε άλλη περιοχή του σώματος (Αγγελόπουλος και συν., 2000).



Εικόνα 8. Κλινική Εικόνα από ένα Ακανθοκυτταρικό Καρκίνωμα της γλώσσας (από Neville et al, 2008)

3.2 Ιστοπαθολογία

Η συμπεριφορά του όγκου εξαρτάται πολλές φορές από την ιστολογική του εικόνα, αν τα κύτταρα ενός όγκου μοιάζουν με τα υπόλοιπα υγιή επιθηλιακά κύτταρα τότε λέμε ότι έχουμε έναν καλά διαφοροποιημένο όγκο. Αν αντίθετα τα κύτταρα εμφανίζονται φαινοτυπικά πολύ διαφορετικά από τα κλασικά επιθηλιακά κύτταρα τότε λέμε ότι έχουμε έναν όγκο χαμηλής διαφοροποίησης (ή αναπλαστικός όγκος). Συνήθως οι καλά διαφοροποιημένοι όγκοι έχουν καλύτερη βιολογική συμπεριφορά, με αργή εξέλιξη και δίνουν μεταστάσεις μετά από πολύ

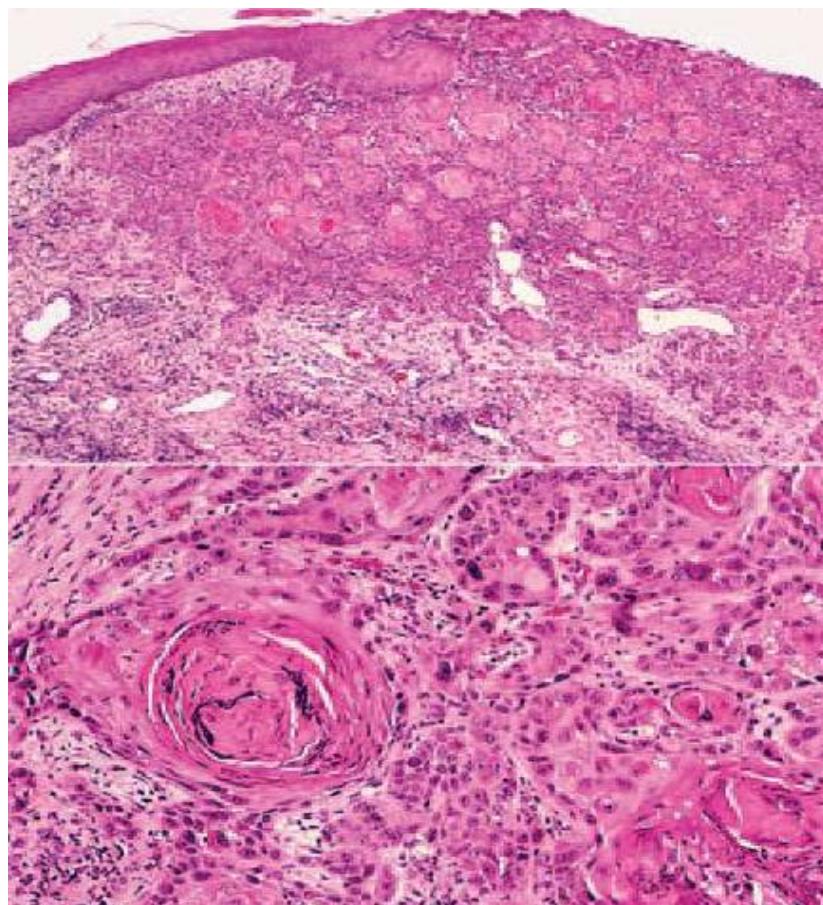
καιρό. Αντίθετα οι χαμηλής διαφοροποίησης είναι συχνά επιθετικοί όγκοι και δίνουν μεταστάσεις σε σχετικά σύντομο διάστημα (Αγγελόπουλος και συν., 2000).

Ιστοπαθολογικά σε αρχικά στάδια παρατηρείται τοπική κλωνική μεταπλασία των κυττάρων της βασικής στιβάδας του επιθηλίου η οποία εν συνεχεία σταδιακά επεκτείνεται σε ολόκληρο το πάχος του επιθηλίου. Στους όγκους υψηλής διαφοροποίησης διαφαίνονται οι διακυτταρικές δεσμοσωματικές γέφυρες και διατηρείται η γενική αρχιτεκτονική του επιθηλιακού ιστού. Γενικότερα όμως διακρίνεται σαφής μεταλλαγή των κυτταρικών χαρακτηριστικών με χαρακτηριστική κυτταρική πλειομορφία, ήπια αύξηση του αριθμού των οπτικά διαφαινόμενων κυτταρικών διαιρέσεων, αύξηση της αναλογίας πυρήνα/κυτταροπλάσματος και περισσότερο σκούρο χρώμα πυρήνα (υπερχρωμάτωση) με εμφανές πυρηνίσκο (ή/και ενίοτε με 2 πυρηνίσκους). Συνήθως οι καλά διαφοροποιημένοι όγκοι σχηματίζουν περιοχές κερατίνης μέσα στην ακανθωτή στιβάδα που ονομάζονται μαργαριτάρια κερατίνης (ή κατ' άλλους κύτταρα φαντάσματα). Τα μαργαριτάρια κερατίνης αφορίζονται από κυκλοτερώς διατεταγμένα καρκινικά κύτταρα και η παρουσία τους είναι σημάδι καλής διαφοροποίησης και μικρότερης επιθετικότητας του όγκου (Ravikiran & Praveen, 2014 ; Srivastava, 2008).

Σε χαμηλής διαφοροποίησης όγκους έχουμε πλέον μεγάλη διαφορά στο μέγεθος από κύτταρο σε κύτταρο με ανώμαλες και πολυάριθμες μιτώσεις, απουσία δεσμοσωμάτων και τα κύτταρα μορφολογικά είναι δυνατό να μην μπορούν να αναγνωριστούν με βεβαιότητα ως προς το σε ποια κατηγορία ιστού ανήκουν (Ravikiran & Praveen, 2014). Ακόμα έχουμε διάρρηξη της ακεραιότητας της διασύνδεσης επιθηλίου και συνδετικού ιστού στην βασική μεμβράνη, και είσοδο των καρκινικών κυττάρων εντός του συνδετικού ιστού. Όλα αυτά δημιουργούν εμφανή απώλεια της γενικότερης αρχιτεκτονικής του επιθηλιακού ιστού (Srivastava, 2008).

Όταν εντοπίζεται το ΑΚΚ σε αρχικά στάδια και δεν εμφανίζεται να διηθεί το υποκείμενο χόριο τότε λέμε ότι έχουμε ένα καρκίνωμα *in situ*, αυτό όμως δεν είναι και τόσο συχνό λόγω του ότι συχνότερα η διάγνωση επιβεβαιώνεται αρκετούς μήνες μετά το αρχικό αυτό στάδιο. Συνήθως έχουμε ήδη πλέον στην στιγμή της διάγνωσης, ιστοπαθολογικά σημάδια διήθησης των καρκινικών κυττάρων στην περιοχή του χορίου ή ακόμα και σε βαθύτερες δομές όπως ο υποβλεννογόνος, το οστό και οι μύες (Neville et al, 2008). Συνήθως διακρίνονται άφθονα φλεγμονώδη κύτταρα στο χόριο όπως λεμφοκύτταρα, πλασματοκύτταρα και ιστιοκύτταρα, και αυτό υποδηλώνει ανοσολογική αντίδραση του οργανισμού στην κακοήθη νεοπλασία

(Αγγελόπουλος και συν., 2000). Όμως είναι ενδιαφέρον ότι όσο πιο χαμηλής διαφοροποίησης είναι ο όγκος τόσο πιο αραιή εμφανίζεται η παρουσία φλεγμονώδων κυττάρων πέριξ των ιστολογικών ορίων του (Srivastava, 2008). Μία ιστολογική εικόνα σε μικρή και μεγάλη μεγένθυνση ενός καλά διαφοροποιημένου ΑΚΚ απεικονίζεται στην Εικόνα 9.



Εικόνα 9. Ιστολογική εικόνα Ακανθοκυτταρικού καρκινώματος καλής διαφοροποίησης σε 2 μεγενθύνσεις, Στην πάνω με μικρή μεγένθυνση και στην κάτω με μεγάλη, οπου βλέπουμε και μαργαριτάρια κερατίνης (Από Neville et al, 2008)

3.3 Σταδιοποίηση

Όπως σε όλους του υπόλοιπους καρκίνους η πιο συχνά χρησιμοποιούμενη ταξινόμηση είναι η TNM (tissue-nodes-metastasis). Αυτή είναι μία καθαρά κλινική σταδιοποίηση και δεν σχετίζεται με τα ιδιαίτερα ιστοπαθολογικά χαρακτηριστικά του όγκου. Η ταξινόμηση TNM παρουσιάζεται στον πίνακα 2.

TNM	T	N	M
0	Χωρίς κλινικά στοιχεία όγκου	Χωρίς μετάσταση επιχώριων λεμφαδένων.	Όχι απομακρυσμένη μετάσταση
1	Όγκος διαμέτρου <2cm	Μετάσταση ομόπλευρου επιχώριου λεμφαδένα με διάμετρο <3cm	Απομακρυσμένη μετάσταση.
2	Όγκος διαμέτρου >2cm αλλά <4cm	Μετάσταση σε λεμφαδένα ομόπλευρα >3cm αλλά μικρότερο από 6cm ή μετάσταση παραπάνω από έναν ομόπλευρο λεμφαδένα.	
3	Όγκος διαμέτρου >4cm χωρίς διήθηση παρακείμενων ανατομικών μορίων.	Μετάσταση σε επιχώριο λεμφαδένα >6cm ή μετάσταση σε ετερόπλευρο λεμφαδένα	
4	Όγκος διαμέτρου >4cm με διήθηση παρακείμενων ανατομικών μορίων.		

Πίνακας 2. Ταξινόμηση TNM για τους κακοήθεις όγκους κεφαλής και τραχήλου (Προσαρμογή από Neville et al, 2008).

Ένας άλλος αμιγώς κλινικός τρόπος σταδιοποίησης περισσότερο αδρός αλλά ευρέως χρησιμοποιούμενος στην καθημερινή κλινική πράξη είναι η συνένωση των παραπάνω TNM χαρακτηριστικών σε διακεκριμένα κλινικά στάδια (Neville et al, 2008) :

- Στάδιο 1: T1N0M0
- Στάδιο 2: T2N0M0
- Στάδιο 3: T3N0M0 ή T[1-3]N1M0
- Στάδιο 4: Κάθε άλλος όγκος.

Από άποψης ιστοπαθολογικής εικόνας του AKK ένα ευρέως χρησιμοποιούμενο σύστημα βαθμονόμησης είναι αυτό του Broder (Gho & Gho, 2014 ; Roland,Caslin,Nash & Stell, 1992) και πιο συγκεκριμένα:

- Grade 1: Όγκοι καλής διαφοροποίησης με παρουσία αναπλαστικών κυττάρων μέχρι 25% και αρκετά μαργαριτάρια κερατίνης.
- Grade 2: Όγκοι μέτριας διαφοροποίησης με παρουσία αναπλαστικών κυττάρων 25-50% με πλέον λίγα σποραδικά μαργαριτάρια κερατίνης.
- Grade 3: Όγκοι μέτριας διαφοροποίησης χωρίς μαργαριτάρια κερατίνης με παρουσία αναπλαστικών κυττάρων 50-75%
- Grade 4: Όγκοι χαμηλής διαφοροποίησης που πλέον μπορεί και να μην αναγνωρίζεται ο ιστός προέλευσης τους με παρουσία αναπλαστικών κυττάρων μεγαλύτερη από 75%.

3.4 Θεραπεία

Η αντιμετώπιση του AKK είναι κατά κύριο λόγο η χειρουργική εξαίρεση, μόνη της ή σε συνδυασμό με ακτινοθεραπεία. Σε περιπτώσεις όμως που ο καρκίνος είναι πλέον προχωρημένου σταδίου, τότε μπορεί να εφαρμοστεί και χημειοθεραπεία (Huang & O' Sullivan, 2013). Στα πλαίσια αυτής της διπλωματικής εργασίας θα αναφερθούμε συνοπτικά μόνο στην εφαρμογή χημειοθεραπευτικών ή στοχευμένων μοριακών θεραπειών στο AKK.

Οι πιο συνηθισμένοι χημειοθεραπευτικοί παράγοντες που χρησιμοποιούνται είναι η σισπλατίνη (cisplatin) και η 5 φθοριο-ουρακίλη (5-FU) χορηγούμενοι συχνά συνδυαστικά . Σπανιότερα μπορεί να χρησιμοποιηθούν και άλλοι χημειοθεραπευτικοί παράγοντες (συνήθως συνδυαστικά) όπως η πακλιταξελη, η μεθοτρεξάτη, η μιτομυκίνη-C, η μπλεομυκίνη, η επιρουβικίνη και η καρβοπλατίνα (Scully, 2012).

Τα τελευταία χρόνια έχουν αναπτυχθεί και στοχευμένες θεραπείες με μονοκλωνικά αντισώματα, οι οποίες έχουν το πλεονέκτημα ότι έχουν λιγότερες και ηπιότερες ανεπιθύμητες ενέργειες, τις περισσότερες φορές όμως χρησιμοποιούνται και σε συνδυασμό με τους χημειοθεραπευτικούς παράγοντες αλλά σε αυτή την περίπτωση οι ανεπιθύμητες ενέργειες είναι εντονότερες και περισσότερες.

Οι κυριότερες κατηγορίες στοχευμένων φαρμακευτικών προσεγγίσεων είναι:

- Οι Αναστολείς EGFR: Cetuximab, Panitumamab, Lapatinib, συνδυασμός Erlotinib+Gemcitabin.
- Οι Αναστολείς mTOR: Rapamycin (ή σιρόλιμους), Δεφορόλιμους, Τεμσιρόλιμους, Εβερόλιμους.
- Οι Αναστολείς VEGF/PDGF: Σουνιτινίμπη, Ιματινίμπη, Bevacizumab (Scully, 2012).
- Οι Αναστολείς PD-1: Nivolumab και Pembrolizumab (Chai, Lim & Cheong, 2020).

Άλλοι παράγοντες είναι η Τραστοζουμάμπη που είναι αναστολέας HER-2 και η Σοραφενίμπη που είναι αναστολέας RAF πολυκινασών (Scully, 2012).

Από τις παραπάνω μόνο, 3 στοχευμένες μοριακές θεραπείας έχουν έγκριθεί από την FDA των ΗΠΑ για το AKK και αφορούν τα μονοκλωνικά αντισώματα Cetuximab, Nivolumab και Pembrolizumab, και αυτά μόνο για μεταστατικά και ανθεκτικές στις κλασικές θεραπείες περιπτώσεις, αφού φαίνεται αυτά να μην έχουν τελικά πολύ θεαματικά αποτελέσματα (Chai, Lim & Cheong, 2020).

3.5 Μοριακή Παθοβιολογία

Μεταλλάξεις σε διαφορετικά κρίσιμα αλλά και γνωστά από άλλους καρκίνους ογκογονίδια και ογκοκατασταλτικά γονίδια που συσσωρεύονται με τον χρόνο και με την συνεχιζόμενη παρουσία παραγόντων κινδύνου και ως εκ τούτου έκθεση του στοματικού βλεννογόνου σε οξειδωτικούς παράγοντες και ελεύθερες ρίζες, έχουν ως αποτέλεσμα την ανάπτυξη AKK (Neville et al, 2008). Στο AKK όμως φαίνεται ότι επηρεάζονται περισσότερο τα ογκοκατασταλτικά γονίδια ενώ σχετικά λίγα πρωτο-ογκογονίδια προσβάλλονται και μεταλλάσσονται σε ογκογονίδια. Εκτός από σημειακές μεταλλάξεις υπάρχουν όπως και στα περισσότερα κακοήθη νεοπλάσματα εκτεταμένες καρυοτυπικές μεταβολές με διπλασιασμούς, απαλοιφές και μεταθέσεις γενετικών περιοχών σε πολλά χρωμοσώματα που προκύπτουν ως αποτέλεσμα γενετικής αστάθειας του γονιδιώματος (Ali, Sabiha, Jan, Haider, Khan & Ali, 2017).

Η σημαντικότερη ερευνητική προσπάθεια εξέτασης γονιδιακών χαρακτηριστικών για το AKK είναι αναμφισβήτητα η δημοσίευση στο Nature από την ομάδα του Genome Atlas

Network τον Ιανουάριο του 2015. Πρόκειται για ανάλυση γονιδιωματικών δεδομένων από RNAseq, miRNA microarrays και ανάλυση μεθυλίωσης σε 279 AKK, από μία ομάδα 100+ κορυφαίων ερευνητών στο τομέα τους. Σε αυτό το paper φάνηκε ότι υπάρχει πραγματικά διαφορά στην γονιδιακή έκφραση των HPV(+) και των HPV(-) AKK. Πιο συγκεκριμένα στους HPV(+) όγκους παρατηρήθηκε απαλοιφή ή απενεργοποιητική μετάλλαξη στο TRAF3 και ενίσχυση του γονιδίου E2F1. Από την άλλη μεριά στους HPV(-) όγκους, υπάρχει ενίσχυση των περιοχών q13 και q22 στο χρωμόσωμα 11, το οποίο προκαλεί αύξηση της έκφρασης των γονιδίων Κυκλίνη D1, FADD, CTTN και BIRC2, YAP1 αντίστοιχα. Απαλοιφές στους HPV(-) όγκους παρατηρήθηκαν πιο συχνά στα NSD1, FAT1, NOTCH1, SMAD4, CKN2A. Ιδιαίτερα στα NSD1, FAT1 και στα NOTCH1 επιπρόσθετα παρατηρούνται και απενεργοποιητικές μεταλλάξεις. Απενεργοποιητικές μεταλλάξεις παρατηρούνται επίσης στο γονίδιο απόπτωσης CASP8 ενώ αντίθετα παρατηρήθηκαν σε μεγάλη συχνότητα ενεργοποιητικές μεταλλάξεις στα HRAS και NFE2L2. Το τελευταίο γονίδιο συνδέεται στενά με την χρήση καπνού και αποτελεί γονιδιακή υπογραφή AKK που οφείλεται στο κάπνισμα.

Σε όλα τα AKK (HPV (-) και HPV(+)) σε αυτή την πολύ σημαντική μελέτη αναφέρεται αύξηση είτε μέσω γενετικού διπλασιασμού είτε μέσω ενεργοποιητικής μετάλλαξης του γονιδίου PIK3CA. Φαίνεται ότι οι περισσότερες χρωμοσωμικές ανακατατάξεις γίνονται σε κάθε περίπτωση στα χρωμοσώματα 3 και 8. Πιο συγκεκριμένα υπάρχει ενίσχυση με διπλασιασμό των περιοχών 3q26 και 3q28 με επακόλουθη αύξηση της έκφρασης των γονιδίων TP63 και SOX2 αλλά και του PIK3CA. Επίσης στις περιοχές 3p και 8p υπάρχει συχνά διαγραφή, όπως διαγραφή υπάρχει και στο βραχύ βραχίονα στο χρωμόσωμα 4 στην περιοχή που κωδικοποιεί το γονίδιο FBXW7. Συχνά επίσης παρατηρείται διπλασιασμός στον μακρύ βραχίονα του χρωμοσώματος 5. Οι παραπάνω γενετικές ανακατατάξεις γίνονται ανεξάρτητα αν έχουμε προσβολή από HPV ή όχι (Cancer Genome Atlas Network, 2015).

Ίσως ο σημαντικότερος παθογενετικός μηχανισμός στο AKK είναι η ενεργοποίηση του βιολογικού μονοπατιού του EGFR. Η υπερέκφραση μορίων του EGFR pathway συνδέεται με χειρότερη πρόγνωση και χαμηλά ποσοστά πενταετούς επιβίωσης. Υπάρχουν 4 ογκογονίδια της οικογένειας του EGFR τα οποία είναι υποδοχείς τυροσινικών κινασών και στην βιβλιογραφία βρίσκονται είτε ως ErbB1-4 είτε ως HER1-4. Το ErbB1/HER1 είναι το EGFR. Αυτοί οι υποδοχείς συνδέονται όπως δείχνει το ονομα τους με τον EGF, με μέλη της

ευρύτερης οικογένειας των EGF (Αμφιρεγούλινη, Epigen κτλ) και με τον TGF-a . Τελικά μέσω αλυσιδωτών σηματοδοτικών οδών ενεργοποιούνται όλα τα σημαντικά για την καρκινική παθογένεση βιολογικά μονοπάτια που αναφέρθηκαν στο προηγούμενο κεφάλαιο (Panarese, Aquino, Ronchi, Longo, Montella, Cozzolino, ... & Franco, 2019).

Από πλευράς γονιδιακής έκφρασης γενικά το AKK κεφαλής και τραχήλου έχει διαιρεθεί σε 4 γονιδιακούς υπότυπους (Walter, Yin, Wilkerson, Cabanski, Zhao, Du, ... & Hayes, 2013):

- **Βασικός (Basal)** : στον οποίο υπερεκφράζονται γονίδια της ομάδας του EGFR (EGFR, COL17A1, TP63, TGF-a). Είναι ο συχνότερος γονιδιακός υπότυπος και σε αυτόν ανήκουν περίπου τα μισά AKK.
- **Μεσεγχυματικός (Mesenchymal)** : στον οποίο υπερεκφράζονται γονίδια που προκαλούν Επιθηλιακή-Μεσεγχυματική Μετάβαση (βλ. παρακάτω) όπως το VIM, DES, TWIST1 και HGF. Αποτελεί τον δεύτερο συχνότερο γονιδιακό υπότυπο και σε αυτόν ανήκει περίπου το 1/3 των AKK.
- **Κλασικός (Classical)** : Σχετιζόμενα με το κάπνισμα και χημικές ουσίες υπερεκφράζόμενα γονίδια (GPX2,NFE2L2). Σε αυτόν ανήκουν 1 στα 10 AKK.
- **Ατυπικός (Atypical)** : στον οποίο υπερεκφράζονται γονίδια σχετικά με προσβολή από τον HPV (CKs, CDKN2A,LIG1,RPA2) . Αποτελεί τον σπανιότερο γονιδιακό τύπο και σε αυτόν ανήκει περίπου 1 στα 20 AKK.

Δυστυχώς δεν έχει διαπιστωθεί συσχέτιση της πρόγνωσης και της τάσης για μετάστασης του AKK με βάση την παραπάνω ταξινόμηση (Chai, Lim & Cheong, 2020). Παρόλα αυτά είναι ιδιαίτερα αποδεκτή από πολλούς ερευνητές μεταξύ των οποίων είναι και η ερευνητική ομάδα του Cancer Genome Atlas Network (Cancer Genome Atlas Network, 2015). Μία πιθανή χρησιμότητα θα μπορούσε να είναι η εξατομικευμένη εφαρμογή στοχευμένων θεραπειών με βάσει το παραπάνω γονιδιακό προφίλ και ανάλογα με τα σηματοδοτικά μονοπάτια που ενισχύονται στον κάθε υπότυπο. Μένει να αποδειχθεί στην πράξη με περισσότερες ερευνητικές προσπάθειες προς αυτή την κατεύθυνση.

Πέρα από τα αποτελέσματα του Cancer Genome Netwok σύμφωνα με την βιβλιογραφία το πιο συχνά μεταλλαγμένο γονίδιο στο AKK είναι το p53 στο 40-80% των περιπτώσεων (Panarese, Aquino, Ronchi, Longo, Montella, Cozzolino, ... & Franco, 2019). Το Μεταλλαγμένο p53 είναι λιγότερο συχνό σε Ασιατικές χώρες , ενώ από την άλλη πλευρά

η μετάλλαξη του FAT1 παρατηρείται συχνότερα σε Ασιατικές χώρες σε σχέση με σχέση με καυκάσιους πληθυσμούς (Chai, Lim & Cheong, 2020).

Κατά την διήθηση των καρκινικών κυττάρων στο χόριο παρατηρείται διάσπαση της βασικής μεμβράνης με αποδόμηση του κολλαγόνου IV που περιέχει η οποία οφείλεται στην παραγωγή διαφόρων μεταλλοπρωτεΐνασών (εξωκυττάριας) μήτρας (MMPs), κυρίως MMP 1,2,3,7,9,11 και 13 από τα καρκινικά κύτταρα (Baker, Leaper, Hayter & Dickenson, 2006 ; De Vicente, Lequerica-Fernández,, Santamaría & Fresno, 2007). Επίσης συγκεκριμένα η παρουσία των MMP 2 και 9 έχει συνδεθεί με μεγαλύτερη τάση λεμφαδενικής μετάστασης του όγκου (Patel, Shah, Rawal, Desai Shah, Rawal & Patel, 2005 ; de Vicente, Fresno, Villalain, Vega & Vallejo, 2005).

Σημαντικό ρόλο όμως για την τάση για διηθητικότητα στο ΑΚΚ φαίνεται να παίζει και το μικροπεριβάλλον του. Αρκετά συστατικά του μικροπεριβάλλοντος μεταξύ άλλων οι ινοβλάστες που σχετίζονται με τον καρκίνο (CAF) και τα ανοσολογικά κύτταρα που περιέχει. Όσον αφορά τα πρώτα είναι κύτταρα με ινοβλαστικό φαινότυπο αλλά με αυξημένη έκφραση α-ακτίνης λείων μυών (aSMA) και ενεργοποιητικής πρωτεΐνης ινοβλαστών (FAP). Η φύση των CAF είναι σε μεγάλο βαθμό άγνωστη μέχρι σήμερα και αποτελεί πεδίο ενεργούς ερευνητικής προσπάθειας. Η παρουσία τους συνδέεται με μεγάλη διηθητικότητα σε διάφορα είδη καρκίνου ενώ στην περίπτωση του ΑΚΚ συνδέεται και με αυξημένη οστική διήθηση (Elmusrati,Pilborough, Khurram & Lambert 2017 ; Sahai,Astsaturov, Cukierman,DeNardo,Egeblad,Evans., ... & Werb, 2020).

Όσον αφορά τα ανοσολογικά κύτταρα του καρκινικού μικροπεριβάλλοντος έχουν υπάρξει αρκετές προσπάθειες ταξινόμησης ΑΚΚ στην βιβλιογραφία σε σχέση με το αν εκφράζονται γονίδια σχετικά με την ενεργοποίηση του ανοσολογικού μηχανισμού όπως η IFN-γ. Τα περισσότερα ΑΚΚ έχουν ενεργοποιημένους τους ανοσολογικούς μηχανισμούς τους αλλά σε ποσοστό 26-30% οι όγκοι χαρακτηρίζονται ως χαμηλής ανοσολογικής απόκρισης ή κατ' άλλους “immune cold”. Ένα μοτίβο που έχει παρατηρηθεί είναι ότι όγκοι που σχετίζονται με HPV λοίμωξη είναι κατά συντριπτική πλειοψηφία υψηλής ανοσολογικής απόκρισης (Chai, Lim & Cheong, 2019).

Άλλα γονίδια των οποίων η υπερέκφραση συνδέεται με φτωχότερη πρόγνωση στην βιβλιογραφία είναι γονίδια της οικογένειας Homeobox , ιδιαίτερα του HOXB7 (Rivera & Rivera, 2014 ; Marcinkiewicz & Gudas, 2014), τα γονίδια INT-2 και HST-1 (Lese, Rossie,

Appel, Reddy, Johnson, Myers & Gollin, 1995), το STAT3 από το JAK/STAT βιολογικό μονοπάτι, η πρωτεΐνη survivin (Panarese et al, 2019) και οι πρωτεΐνες θερμικού σοκ HSP-27 και HSP-70 (Popli, Sircar, Chowdhry & Rani, 2015). Γονίδια τελομερασών επίσης φαίνονται μεταλλαγμένα, ενώ τελευταία ερευνάται και η συμμετοχή των miRNA στις μεταβολές γονιδιακής έκφρασης (Scully, 2012 ; Cancer Genome Atlas Network, 2015).

Ο ρόλος των επιθηλιακών βλαστοκυττάρων έχει επίσης διερευνηθεί. Οι κυριότεροι βιοδείκτες για αυτά όσον αφορά το AKK της περιοχής κεφαλής και τραχήλου είναι το CD44 και η ALDH, η υπερέκφραση των οποίων συνδέεται με κακή πρόγνωση. Η ALDH εμπλέκεται και στον μεταβολισμό του ρετινοϊκού οξέως, του οποίου η έλλειψη όπως προαναφέρθηκε θεωρείται πιθανός παράγοντας κινδύνου για το AKK (Papagerakis et al, 2014).

Ένας πολύ σημαντικός μηχανισμός που θεωρείται ότι συμβάλλει αποφασιστικά στην δημιουργία μεταστάσεων είναι η Επιθηλιακή- Μεσεγχυματική Μετάβαση (EMT), ένας μηχανισμός που παίζει σημαντικό ρόλο στην Εμβρυογένεση , αλλά και σε άλλες εκφάνσεις της ομοιόστασης του ενήλικου οργανισμού όπως η απόκριση των επιθηλιακών κυττάρων σε φλεγμονή ή σε τραυματισμό. Σε κακοήθεις νεοπλασίες επιθηλιακής προέλευσης όμως ο αρχέγονος αυτός μηχανισμός γίνεται εργαλείο στην αύξηση της κινητικότητας των καρκινικών κυττάρων και εν τέλει στην μετάσταση τους. (Zilhao & Neves, 2019)

Στην EMT το επιθηλιακό κύτταρο αποκτά χαρακτηριστικά μεσεγχυματικού κυττάρου. Αυτό συμβαίνει μέσω βιοχημικών μετασχηματισμών στον κυτταροσκελετό (μείωση τονοϊνιδίων και αυξηση ινιδίων F-ακτίνης) αλλά και μέσω ορισμένων μεταγραφικών παραγόντων (EMT-TFs) αλλαγή στην έκφραση συγκεκριμένων πρωτεϊνών, όπως η μείωση έκφρασης της E-Καδερίνης (E-Cadherin) και η αύξηση της Βιμεντίνης (VIM) . Σημαντικά σηματοδοτικά μονοπάτια που συμβάλλουν στα παραπάνω είναι για το AKK, η κανονική οδός Wnt (ο όρος κανονική υποδηλώνει την συμμετοχή β-κατενίνης που είναι ένας σημαντικός μεταγραφικός παράγοντας), το NOTCH σηματοδοτικό μονοπάτι αλλά και μονοπάτια σχετιζόμενα με το EGFR και το TGFβ (Bai, Sha & Kano, 2020 ; Jayanthi, Varun & Selvaraj, 2020).

Συνοπτικά ο μηχανισμός μετάστασης που προτείνεται βάσει του παραπάνω μοντέλου είναι ότι μετά την EMT το κύτταρο αποκτά κινητικότητα και περνά μέσω του ενδοθηλίου στην αιματική κυκλοφορία. Ωστόσο να γίνει μετάσταση πρέπει το συγκεκριμένο κύτταρο να περάσει από τα αντίξοα περιβάλλοντα της αιματικής κυκλοφορίας και να προσαρμοστεί εν

τέλει στον νέο περιβάλλον της μέλλουσας μεταστατικής περιοχής. Εν συνεχείᾳ θα πρέπει να λάβει χώρα το αντίστροφο φαινόμενο από την Επιθηλιακή –Μεσεγχυματική Μετάβαση, που εύστοχα ονομάζεται Μεσεγχυματική – Επιθηλιακή Μετάβαση (MET). Δηλαδή το νέο-μετασταθέν καρκινικό κύτταρο αποκτά ξανά χαρακτηριστικά επιθηλιακού κυττάρου μέσω αντιστροφής των μηχανισμών EMT που προαναφέρθηκαν. Τελικά συμβαίνει κλωνικός ανεξέλεγκτος πολλαπλασιασμός του κυττάρου αυτού που θα οδηγήσει τελικά στην δημιουργία της κλινικά επιβεβαιωμένης μετάστασης (Zilhao & Neves, 2019).

Κεφάλαιο 4. Μετα-ανάλυση δεδομένων γονιδιακής έκφρασης

4.1 Τεχνικές υψηλής ανάλυσης στην μελέτη της γονιδιακής έκφρασης

Η αποκρυπτογράφηση του ανθρώπινου γονιδιώματος στις αρχές του 21^{ου} αιώνα αποτέλεσε το άνοιγμα ενός νέου κεφαλαίου στην βιολογία, την λεγόμενη βιολογία μεγάλων δεδομένων (Collins, Morgan & Patrinos, 2003). Πλέον στην νέα αυτή εποχή έχουν αναπτυχθεί μέθοδοι υψηλής ανάλυσης (high throughput) που μας επιτρέπουν να μελετάμε ολόκληρο το γονιδίωμα (γονιδιωματική) αλλά και το σύνολο κυτταρικών βιομορίων όπως των mRNA (μεταγραφωμική), των πρωτεϊνών (πρωτεωμική), των μεταβολιτών (μεταβολωμική), των μοτίβων μεθυλίωσης (μεθυλωμική) κοκ. (Schneider & Orchard, 2011).

Στην μελέτη της μεταγραφωμικής σε επίπεδο υψηλής ανάλυσης υπάρχουν 2 κύριες τεχνικές που χρησιμοποιούνται σήμερα, οι μικροσυστοιχίες (microarrays) και η τεχνική αλληλουχισης RNA (RNAseq). Αυτές οι 2 τεχνολογίες και περισσότερο η RNAseq παράγουν τεράστιο όγκο δεδομένων (Νικολάου & Χουβαρδάς, 2015). Αν και υπάρχει μία τάση εγκατάλειψης των μικροσυστοιχιών στην ανάλυση της γονιδιακής έκφρασης με την ζυγαριά να γέρνει σταδιακά προς μεθόδους RNAseq, παρόλα αυτά δεν έχει έρθει ακόμα η ώρα να εγκαταλειφθούν οι μικροσυστοιχίες και συνεχίζουν ακόμα να δημοσιεύονται πειράματα βασισμένα σε μικροσυστοιχίες μέχρι σήμερα (Gonzalo & Sanchez, 2018). Για την διενέργεια μίας μετα-ανάλυσης στην περίοδο που διανύουμε φαίνεται ότι υπάρχουν περισσότερα δεδομένα στα βασικά αποθετήρια με πειράματα βασισμένα σε μικροσυστοιχίες παρά σε πειράματα RNAseq. Για παράδειγμα στο GEO (Gene expression Omnibus) για το AKK υπάρχουν περίπου 20 datasets RNAseq αλλά πάνω από 100 datasets με μικροσυστοιχίες (<https://www.ncbi.nlm.nih.gov/gds>).

4.2 Δυσκολίες στην μετα-ανάλυση ελεύθερων δεδομένων μικροσυστοιχιών

Τα dataset που υπάρχουν δημοσιευμένα σε αποθετήρια μικροσυστοιχιών όπως το GEO (Gene Expression Omnibus), έχουν κατά κανόνα μικρό αριθμό δειγμάτων. Έτσι είναι πολύ φυσικό κάποιος ερευνητής να σκεφτεί να συνδυάσει αυτά τα δεδομένα με σκοπό να έχει μεγαλύτερη στατιστική ισχύ στα αποτελέσματα του (Hu, Greenwood & Beyene, 2006). Ωστόσο αυτό το εγχείρημα αν και φαίνεται απλό εκ πρώτης όψεως, αποδεικνύεται τελικά ένα

εξαιρετικά δύσκολο πρόβλημα γιατί συνδυάζονται πολλαπλασιαστικά ετερογενή θορυβώδη δεδομένα με αποτέλεσμα η μετα-ανάλυση σε datasets από μικροσυστοιχίες να αποτελεί ένα δυσεπίλυτο επιστημονικό πρόβλημα. Μέχρι και σήμερα δεν έχει βρεθεί κοινώς αποδεκτή λύση προς αυτή την κατεύθυνση.

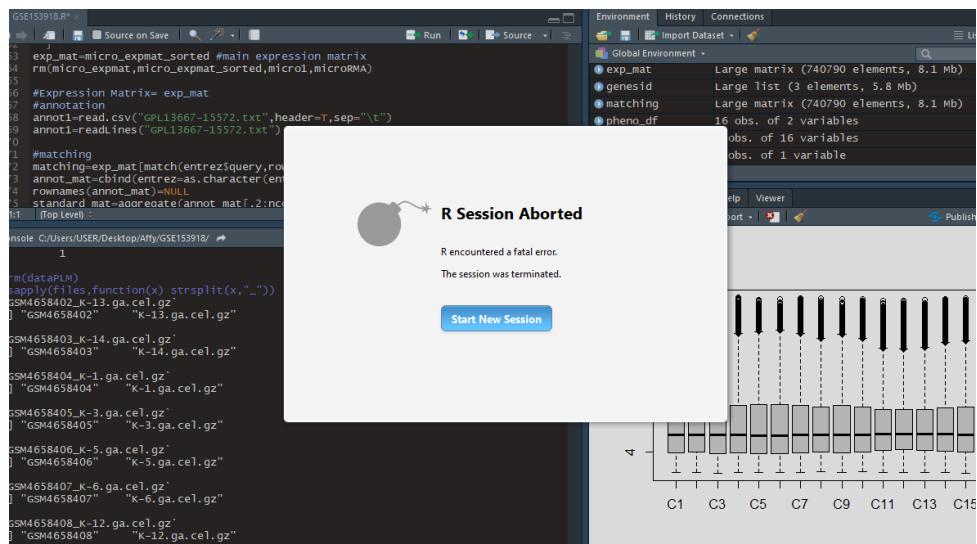
Σε αποθετήρια μικροσυστοιχιών οι επιστήμονες που κάνουν κάποια δική τους καταχώρηση από δεδομένα που παράγουν στα εργαστήρια τους, πρέπει να τηρούν κάποιες προϋποθέσεις που να συμφωνούν με ευρέως αποδεκτά κριτήρια αποδοχής όπως τα MIAME (Minimum Information About a Microarray Experiment). Δύο από αυτές τις προϋποθέσεις είναι να συμπεριλάβουν τόσο τα πρωτογενή δεδομένα (RAW data) όσο και τα κανονικοποιημένα, χωρίς όμως να προσδιορίζεται το είδος της κανονικοποίησης (Meyer, 2011). Συνεπώς κάποιος που καταχωρεί βιολογικά δεδομένα σε αποθετήρια όπως το GEO, μπορεί να χειριστεί τα δεδομένα του όπως θέλει πριν τα υποβάλλει.

Πολλοί βιοεπιστήμονες μπορεί να μπουν στον πειρασμό να παραλείψουν το «μαύρο κουτί» της προεπεξεργασίας πρωτογενών δεδομένων και μάλιστα κάποιοι φτάνουν και στο σημείο ακόμα και σήμερα να χρησιμοποιήσουν στις δημοσιεύσεις τους naïve προσεγγίσεις χρησιμοποιώντας αλόγιστα χωρίς κάποιον περαιτέρω έλεγχο τα δεδομένα των series matrices του GEO (Su, Wang, Zheng, Wei & Song, 2018). Τα πρωτογενή δεδομένα είναι δεδομένα που είναι αρκετά μεγάλα σε μνήμη (της τάξεως εκατοντάδων MB για μικρά dataset και μερικών GB στην περίπτωση των μεγαλύτερων), και η επεξεργασία τους έχει ακόμα μεγαλύτερες υπολογιστικές απαιτήσεις. Αυτό έχει ως αποτέλεσμα να μην μπορεί εύκολα (ή καθόλου) να γίνει preprocessing των RAW δεδομένων με έναν κοινό υπολογιστή για καθημερινή χρήση (Εικόνα 10). Για αυτή την δουλειά χρειάζεται ισχυρός υπολογιστής με μεγάλη RAM και έναν επεξεργαστή πολλών πυρήνων με ταχύτητα ρολογιού που να μπορεί να στηρίξει την χρήση μεγάλης μνήμης RAM. Το πρόβλημα αυτό γίνεται ακόμα εντονότερο σε κάποια πιο καινούρια πακέτα προεπεξεργασίας όπως το oligo της Affymetrix (Εικόνα 11), πακέτο που χρησιμοποιείται για cDNA oligonucleotide arrays (Carvalho & Irizzary 2010), που είναι αρκετά πιο απαιτητικά σε μνήμη από παλαιότερα, με αποτέλεσμα ακόμα και datasets πολύ μικρού μεγέθους να μην μπορούν υποστούν προεπεξεργασία σε υπολογιστές καθημερινής χρήσης. Σε μεγαλύτερα dataset η προεπεξεργασία RAW δεδομένων γίνεται αδύνατη. Ειδικά στην περίπτωση της R και του R studio, χρειάζεται οπωσδήποτε μηχάνημα 64-bit (R Core Team-2020). Γιατί ακόμα και αν ένα μηχάνημα 32-bit μπορεί να στηρίξει 4

GB RAM, που είναι το όριο για τα μηχανήματα 32-bit (2^{32} bit RAM). Οι εκδόσεις R/Rstudio για λειτουργικά συστήματα 32-bit μπορούν να αντέξουν το πολύ μέχρι 2GB. Περισσότερα για αυτό μπορεί κάποιος να βρει στο link:

<https://www.rdocumentation.org/packages/utils/versions/3.6.2/topics/memory.size>

Πιθανώς μία λύση στο παραπάνω πρόβλημα με τα λειτουργικά συστήματα 32-bit είναι η χρήση compiled γλωσσών προγραμματισμού (C/C++, Java), και αποφυγή χρήσης scripting languages όπως είναι η R και η Python, φυσικά με το ακριβό αντίτιμο της πολλαπλάσιας ποσότητας κώδικα που πρέπει να γραφτεί.



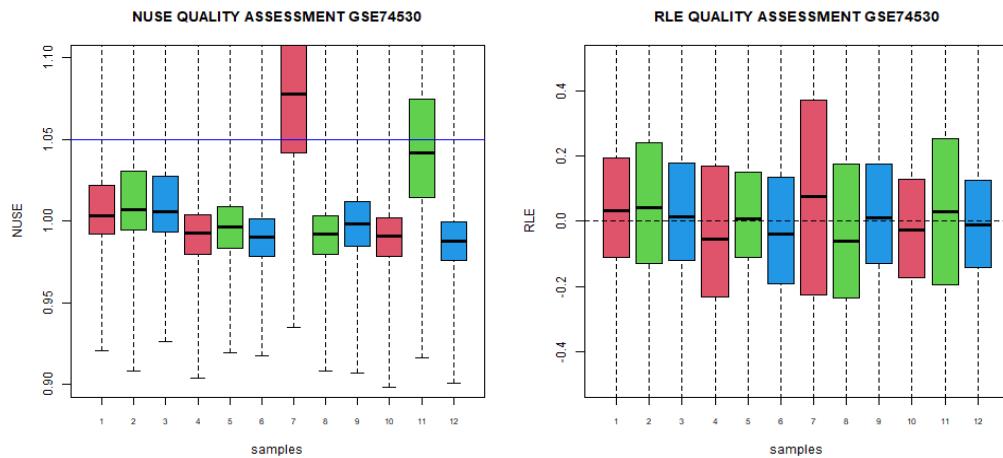
Εικόνα 10. Προσπάθεια προεπεξεργασίας RAW δεδομένων ενός dataset με σχετικά μικρό αριθμό δειγμάτων σε υπολογιστή γενικής χρήσης.

```
** building package indices
** testing if installed package can be loaded from temporary location
No methods found in package 'RSQLite' for request: 'dbListFields' when loading 'oligo'
** testing if installed package can be loaded from final location
No methods found in package 'RSQLite' for request: 'dbListFields' when loading 'oligo'
** testing if installed package keeps a record of temporary installation path
* DONE (pd.huex.1.0.st.v2)

The downloaded source packages are in
  'C:\Users\USER\AppData\Local\Temp\Rtmp4qcvye\downloaded_packages'
Loading required package: pd.huex.1.0.st.v2
Loading required package: RSQLite
Loading required package: DBI
Platform design info loaded.
Reading in : GSM2770759_CON01_Biopsy-2_BB428_HuEx-1_0-st-v2_.CEL.gz
Reading in : GSM2770761_CON03_Biopsy-2_BB432_HuEx-1_0-st-v2_.CEL.gz
Reading in : GSM2770763_CON04_Biopsy-2_BB434_HuEx-1_0-st-v2_.CEL.gz
Reading in : GSM2770765_CON05_Biopsy-2_BB436_HuEx-1_0-st-v2_.CEL.gz
Reading in : GSM2770767_CON06_Biopsy-2_BB438_HuEx-1_0-st-v2_.CEL.gz
Reading in : GSM2770769_CON07_Biopsy-2_BB440_HuEx-1_0-st-v2_.CEL.gz
Reading in : GSM2770771_CON08_Biopsy-2_BB442_HuEx-1_0-st-v2_.CEL.gz
Reading in : GSM2770773_CON10_Biopsy-2_BB446_HuEx-1_0-st-v2_.CEL.gz
Reading in : GSM2770775_CON11_Biopsy-2_BB471_HuEx-1_0-st-v2_.CEL.gz
Error: cannot allocate vector of size 450.0 Mb
> |
```

Εικόνα 11. Προσπάθεια προεπεξεργασίας ενός πολύ μικρού dataset (9 δείγματα) στο οποίο έπρεπε να γίνει επεξεργασία με το πακέτο oligo της Affymetrix σε υπολογιστή με λειτουργικό σύστημα 32-bit.

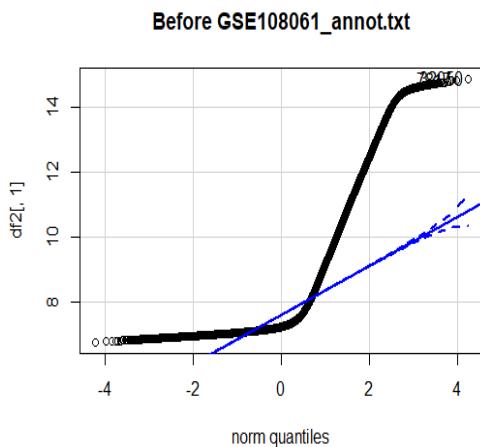
Ωστόσο η επεξεργασία των «κανονικοποιημένων» dataset φαίνεται τελικά στην πορεία να έχει αρκετά περισσότερα μειονεκτήματα παρά πλεονεκτήματα. Στα πλαίσια αυτής της διπλωματικής έγινε προσπάθεια αρχικά να χρησιμοποιηθούν τα έτοιμα αυτά “series expression matrix”. Ένα πρώτο μειονέκτημα είναι ότι σε αυτά τα δεδομένα συχνά δεν έχει γίνει έλεγχος ποιότητας των δεδομένων κάθε δείγματος (McCall, Murakami, Lukk , Huber & Irizarry, 2011). Αυτό έχει ως αποτέλεσμα αυτόματα τα αποτελέσματα μελετών που στηρίχτηκαν εν μέρει στα συγκεκριμένα δείγματα να τίθεται υπό αμφισβήτηση. Εύκολα μπορεί να αντιληφθεί κάποιος ότι το φαινόμενο της συμπερίληψης κακής ποιότητας δεδομένων έχει πολλαπλασιαστικά καταστροφικά αποτελέσματα στην περίπτωση της μετα-ανάλυσης πολλών dataset που οδηγεί αναπόφευκτα σε διαστρεβλωμένα συμπεράσματα. Στην εικόνα 12 παρατηρούμε τους δείκτες ποιότητας RLE (Relative log expression) και NUSE (normalized unscaled standard error) σε ένα μικρό dataset που λήφθηκε από το GEO στα πλαίσια αυτής της διπλωματικής, με microarrays της Affymetrix, όπου φαίνεται ότι το δείγμα 7 είναι κακής ποιότητας. Οι παραπάνω δείκτες ποιότητας έχουν εφαρμογή μόνο σε microarrays της Affymetrix , άλλες εταιρίες έχουν άλλους δείκτες ποιότητας και βασίζονται περισσότερο σε MA plots (Bolstad, 2007) .



Εικόνα 12. Αξιολόγηση ποιότητας δεδομένων με NUSE και RLE σε normalized dataset από το GEO . Το δείγμα 7 φαίνεται να είναι με σιγουριά ένα δείγμα κακής ποιότητας. Το δείγμα 11 φαίνεται επίσης αμφιβόλου ποιότητας.

Το δεύτερο μειονέκτημα της χρήσης των έτοιμων κανονικοποιημένων δεδομένων είναι ότι αν και αυτά ονομάζονται «κανονικοποιημένα» στην πραγματικότητα πολλά από αυτά δεν ακολουθούν ούτε κατά προσέγγιση την κανονική κατανομή. Στην εικόνα 13 φαίνεται ένα διάγραμμα Q-Q που συγκρίνει την κανονική κατανομή (μπλε γραμμή) με την κατανομή του δείγματος (μαύρη γραμμή) σε ένα microarray της εταιρίας Illumina και στο οποίο αναφέρεται για το συγκεκριμένο dataset ότι έχει γίνει κανονικοποίηση ποσοστημορίων (quantile normalization, QN). Τα συγκεκριμένα δεδομένα είχαν υποστεί πράγματι QN, και αυτό διαπιστώθηκε γιατί έγινε απόπειρα εφαρμογής QN στα παραπάνω «κανονικοποιημένα δεδομένα» και το αποτέλεσμα ήταν ακριβώς το ίδιο, ωστόσο σε καμία περίπτωση δεν μπορεί να θεωρηθεί κανονικής κατανομής. Αρχίζει κάποιος ίσως από τώρα να αντιλαμβάνεται ότι πλέον δεν παίζει μόνο ρόλο το τι προεπεξεργασία έχει γίνει σε κάποιο dataset, αλλά εξίσου σημαντικό ρόλο είναι και το από ποιά πλατφόρμα προέρχεται και τι δεδομένα παράγονται τελικά από αυτήν. Ακόμα και αν κάποιος επιχειρήσει να χρησιμοποιήσει την ίδια μέθοδο κανονικοποίησης σε όλα τα RAW δεδομένα του, η διαφορετική χημεία που χρησιμοποιούν διαφορετικές εταιρίες (ή και κάποιες φορές διαφορετικές πλατφόρμες της ίδιας εταιρίας), κάνει δεδομένα με την ίδια στρατηγική preprocessing να έχουν τελικά μία τελείως διαφορετική κατανομή με αποτέλεσμα η ενσωμάτωση των δεδομένων τους σε αυτή την περίπτωση να είναι αδύνατη (Dubitzky, Granzow & Berrar, 2007).

Επομένως ενδέχεται τελικά ο όρος «κανονικοποίηση» (normalization) στην ορολογία των μικροσυστοιχιών να είναι εντελώς άστοχος και αδόκιμος και ίσως θα πρέπει να αντικατασταθεί με κάποιον άλλον πιο εύστοχο όρο όπως π.χ. «εναρμόνιση δεδομένων» ή «ομογενοποίηση δεδομένων» ή «τυποποίηση δεδομένων».

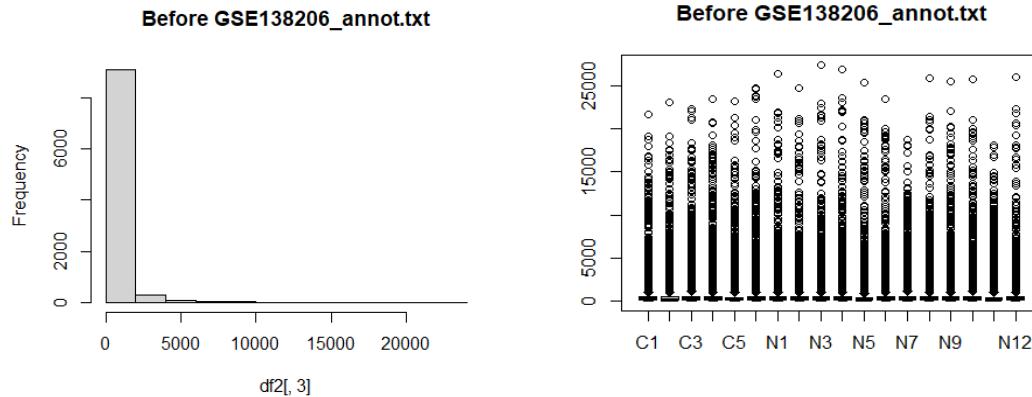


Εικόνα 13. Διάγραμμα Q-Q «κανονικοποιημένου» dataset. Η κανονική κατανομή αναπαριστάται από την μπλε γραμμή, ενώ η κατανομή του δείγματος από την μαύρη καμπύλη.

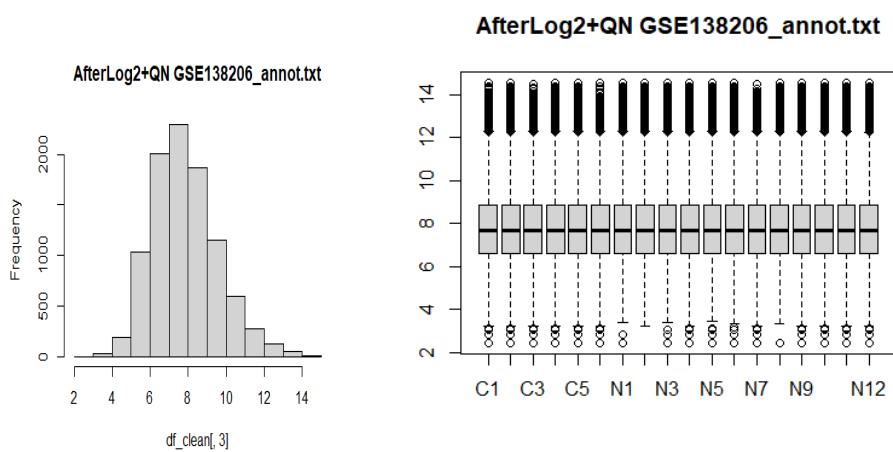
Το παραπάνω πρόβλημα γίνεται ακόμα πιο έντονο όταν, ενώ αναφέρεται στις πρώτες γραμμές ενός κανονικοποιημένου πίνακα έκφρασης που λήφθηκε από ένα αποθετήριο (series matrix annotation για το GEO) ότι αυτό έχει υποστεί κάποια επεξεργασία, τελικά αυτό έχει υποστεί μόνο ένα μέρος αυτής ή/και καμία επεξεργασία. Στις Εικόνες 14 και 15 είναι αισθητή η βελτίωση μετά από μία στοιχειώδη επεξεργασία πράγμα που σημαίνει ότι τα συγκεκριμένα δεδομένα εξ' αρχής δεν είχαν υποστεί καμία μέθοδο κανονικοποίησης. Γι' αυτό τον λόγο άλλωστε και το ίδιο το GEO αναφέρει:

“values are **assumed** to be submitted as normalized signal count data”.
[\(<https://www.ncbi.nlm.nih.gov/geo/info/faq.html>\)](https://www.ncbi.nlm.nih.gov/geo/info/faq.html)

Επομένως πάντοτε, πριν από οποιαδήποτε περαιτέρω ανάλυση συνιστάται να παρατηρούνται γραφικά τα δεδομένα, ώστε να είναι κανείς βέβαιος τι είδους δεδομένα έχει στην διάθεση του αλλά και με ποιον τρόπο αυτά πραγματικά κατανέμονται.



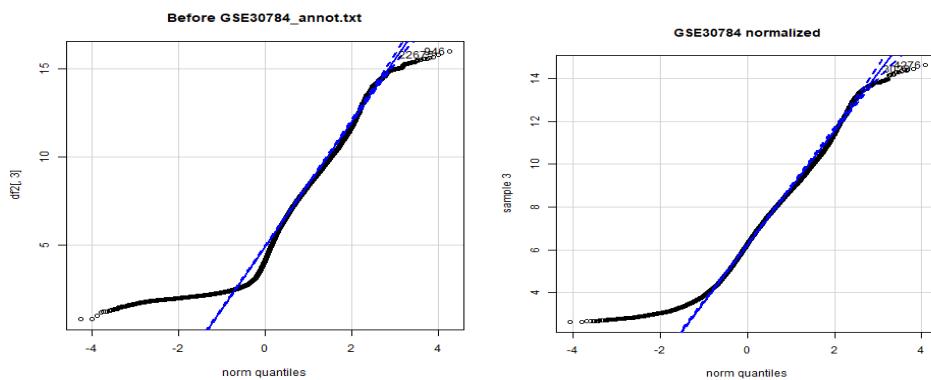
Εικόνα 14. Δεδομένα στα οποία αναφέρεται ότι έχει γίνει QN.



Εικόνα 15. Τα δεδομένα τις εικόνας 10 μετά από ένα απλό QN και λογαρίθμηση με log2.

Το τρίτο μειονέκτημα της χρήσης «κανονικοποιημένων» πινάκων έκφρασης, είναι ότι κάποιες φορές έχουν υποστεί φθορά (corruption). Αυτό σημαίνει ότι έχουν υποβληθεί δεδομένα που έχουν πολλές κενές τιμές (πιθανώς λόγω κάποιας προσπάθειας λογαρίθμησης σε μη επεξεργασμένα δεδομένα), με αποτέλεσμα κάποιος που προσπαθεί να τα αναλύσει να χρειαστεί είτε να αφαιρέσει κάποια δείγματα με βάση ένα αυθαίρετο ποσοστό κενών τιμών ανά γονίδιο ή ανά δείγμα, είτε να καταφύγει σε κάποια μέθοδο αποκατάστασης όπως π.χ. με μέσο όρο δείγματος ή γονιδίου αλλά και πιο εξεζητημένες μεθόδους όπως kNN imputation (Aittokallio, 2010). Το φαινόμενο αυτό φαίνεται να εμφανίζεται πιο σπάνια με τα RAW data.

Το τελευταίο και ίσως σημαντικότερο μειονέκτημα είναι ότι υπάρχει μεγάλη ετερογένεια στο πως έχει γίνει επεξεργασία ακόμα και σε δεδομένα που προέρχονται από ίδιες πλατφόρμες. Ένα τέτοιο παράδειγμα είναι π.χ. η προεπεξεργασία με RMA (Robust Microarray Analysis) σε κάποια dataset από microarray της Affymetrix και με gcRMA (GeneChip RMA) σε άλλα dataset από microarray της ίδιας εταιρίας. Αυτό συμβαίνει γιατί η gcRMA θεωρούνταν (ή και ακόμα ίσως θεωρείται από κάποιους) καλύτερη μέθοδος κανονικοποίησης γιατί λαμβάνει υπόψη της και τα διαφορετικά background noise effects που έχουν διαφορετικοί ανιχνευτές πάνω σε ένα microarray εξαιτίας του διαφορετικού GC περιεχομένου του cDNA τους (Wu & Irizarry, 2004). Το παραπάνω αυτό «πλεονέκτημα» δυστυχώς έχει αντίκτυπο στην κανονικότητα των δεδομένων. Το gcRMA δίνει κατανομές που είναι λιγότερο κανονικές σε σχέση με κατανομές από το απλό RMA (Εικόνα 12). Επίσης παρά τον αρχικό ενθουσιασμό για αυτή την κανονικοποίηση, η συγκεκριμένη μέθοδος έχει διαπιστωθεί ότι προσθέτει περισσότερο θόρυβο τελικά στα δεδομένα σε σύγκριση με την απλή RMA πράγμα που αποδίδεται σε platform-independent artifacts που σχετίζονται με διασταυρούμενους υβριδισμούς εντός της ίδιας μικροσυστοιχίας (Lim, Wang, Lefebvre & Califano, 2007). Για τον παραπάνω λόγο η απλή RMA θεωρείται μέχρι σήμερα η δημοφιλέστερη μέθοδος κανονικοποίησης και το “gold standard” για μικροσυστοιχίες της Affymetrix.



Εικόνα 16. Q-Q διάγραμμα από Δεδομένα μετά από gcRMA (αριστερά) και ίδια δεδομένα μετά από RMA κανονικοποίηση των CEL files (δεξιά)

Ανεξάρτητα από τα δεδομένα («κανονικοποιημένα» ή RAW) που θα πάρει κάποιος η ενσωμάτωση δεδομένων από πολλά dataset έχει και άλλες προκλήσεις. Μία από αυτές είναι το διαφορετικό annotation στα features (probes/γονίδια). Σε αυτή την εργασία

χρησιμοποιήθηκε annotation με Entrez IDs που είναι μια κεντρική και ευρέως αποδεκτή μορφή ταυτοποίησης, που χρησιμοποιείται και ως είσοδος σε πολλά εργαλεία βιοπληροφορικής ανάλυσης. Παρατηρείται λοιπόν το φαινόμενο “Many to Many” (Okoniewski, Yates, Dibben & Miller, 2007 ; Ramasamy, Mondry, Holmes & Altman, 2008), που σημαίνει ότι είναι δυνατόν ένα probe από ένα microarray να αντιστοιχεί σε πολλά Entrez IDs αλλά και αντίστροφα ένα Entrez ID να αντιστοιχεί σε πολλά probes. Ένα ενδεικτικό τέτοιο παράδειγμα απεικονίζεται στην Εικόνα 17, όπου τα probes 11715100_at, 11715101_s_at και 11715102_x_at αντιστοιχούν στο γονίδιο με Entrez ID 8355; αντιστρόφως το probe 11715107_s_at αντιστοιχεί σε γονίδια με Entrez IDs 474383, 474384 και 8263. Η αντιμετώπιση αυτού του φαινομένου είναι δύσκολη και μπορεί να έχει πολλές προσεγγίσεις. Στην παρούσα εργασία σε probes που αντιστοιχούν στο ίδιο Entrez, για το Entrez αυτό λήφθηκε η μέγιστη τιμή (ένας άλλος τρόπος θα μπορούσε να είναι ο μέσος όρος όλων). Αντίστροφα σε ένα probe που αντιστοιχούν πολλά Entrez, θεωρήθηκαν όλα τα Entrez να έχουν την τιμή αυτού του probe, με άλλα λόγια αν π.χ. υπάρχει probe με 3 entrez IDs, αυτά υποδηλώνονται στον τελικό πίνακα έκφρασης ως 3 γραμμές με πανομοιότυπες τιμές κατά μήκος των στηλών (ένας άλλος πιο «αυθαίρετος» τρόπος θα ήταν να ληφθεί ένα μόνο από αυτά π.χ. αυτό με το μικρότερο αριθμό, καθότι μεγάλες τιμές Entrez συχνά αντιστοιχούν σε τμήματα του DNA που κωδικοποιούν μη κωδικά RNA-ncRNAs). Πάντως γενικά δεν υπάρχει και εδώ ένας ευρέως αποδεκτός τρόπος για τον χειρισμό αυτού του προβλήματος.

ID	entrez
1 11715100_at	8355
2 11715101_s_at	8355
3 11715102_x_at	8355
4 11715103_x_at	126282
5 11715104_s_at	92736
6 11715105_at	284099
7 11715106_x_at	340307
8 11715107_s_at	474383 /// 474384 /// 8263
9 11715108_x_at	285501
10 11715109_at	344658
11 11715110_at	645432
12 11715111_s_at	1082 /// 114335 /// 114336 /// 93659 /// 94027 /// 94115
13 11715112_at	127254
14 11715113_x_at	55199
15 11715114_x_at	55199
16 11715115_s_at	8346
17 11715116_s_at	8367
18 11715117_x_at	8331
19 11715118_s_at	8343
20 11715119_s_at	388125

Εικόνα 17. Το πρόβλημα «Many to Many» στα microarrays σε ένα R studio Terminal κατά την επεξεργασία δεδομένων για αυτή την διπλωματική.

Ένα τελευταίο αλλά ίσως το σημαντικότερο και το πιο δυσεπίλυτο πρόβλημα είναι το καθ'εαυτού πρόβλημα ενσωμάτωσης και εναρμόνισης διαφορετικών δεδομένων μικροσυστοιχιών με τελικό σκοπό μια μετα-ανάλυση. Και εδώ έχουν περιγραφεί στην βιβλιογραφία αρκετοί τρόποι προσέγγισης, που όμως κανένας τρόπος δεν θεωρείται σήμερα αξιόπιστος. Οι περισσότεροι από αυτούς γενικά λαμβάνουν μία σειριακή προσέγγιση, κατά την οποία λαμβάνονται υπ’όψη είτε το p-value ή σταθμισμένα p-value κάθε γονιδίου σε κάθε μελέτη (Hu et al, 2006) και ranking αυτών των p-values. Μία άλλη γνωστή τεχνική εμπνευσμένη από μη παραμετρικές στατιστικές προσεγγίσεις είναι το καθ' εαυτού ranking που έχει η γονιδιακή έκφραση σε ένα γονίδιο σε σχέση με τα άλλα σε κάθε μελέτη, ενδεικτικές τέτοιες μέθοδοι είναι η Rankproduct και το RankSum (Hong et al, 2006). Κάποιοι προτείνουν την χρήση του αλγορίθμου ComBat (Muller et al, 2016), όμως αυτό χρειάζεται ως είσοδο τιμές batch effect για κάθε συγκεκριμένη πλατφόρμα πράγμα που γίνεται πειραματικά σε σύγκριση με πλατφόρμα με γνωστό batch effect distribution, και επομένως η χρήση του σε μία μετα-ανάλυση ελεύθερων δεδομένων καθίσταται δύσκολη ή και αδύνατη. Επιπλέον έρευνες δείχνουν ότι μία προσέγγιση με ζ-κανονικοποίηση έχει ίδια ή και καλύτερη απόδοση από το ComBat (Yasrebi, 2016). Αλγόριθμοι που παλαιότερα είχαν προταθεί όπως ο XPN (Taminau et al, 2012) με την ελπίδα να λύσουν αυτό το πρόβλημα, δεν

φαίνεται να είναι αποδεκτοί σήμερα. Είναι προφανές ότι η ποικιλομορφία των εργαλείων μετα-ανάλυσης που έχουν αναπτυχθεί για τα microarray είναι ενδεικτική της δυσκολίας του προβλήματος, αλλά και ενδεικτική του ότι δεν έχει βρεθεί ακόμα κάποια πραγματικά αξιόπιστη μέθοδος μετα-ανάλυσης δεδομένων από πολλά τέτοια datasets.

Στις περισσότερες από τις παραπάνω μεθόδους το τελικό αποτέλεσμα είναι ουσιαστικά μία λίστα ΔΕ γονιδίων με κάποιο μη παραμετρικό στατιστικό αποτέλεσμα. Όμως με αυτή την τακτική είναι αδύνατη η εφαρμογή μεθόδων μηχανικής μάθησης (π.χ. Random Forest, Support Vector Machine, Τεχνητά Νευρωνικά δίκτυα, Decision Trees) που είναι εντελώς απαραίτητοι για συγκεκριμένους στόχους όπως παραδειγματος χάριν για την διερεύνηση κατάλληλων βιοδεικτών (Azuaje, 2010).

Συμπερασματικά δεν υπάρχει ούτε εδώ ευρέως αποδεκτός τρόπος αντιμετώπισης του προβλήματος και ακόμα και σχετικά σύγχρονες δημοσιευμένες έρευνες χρησιμοποιούν απλά τα series matrix του GEO για μία μετα-ανάλυση με δεδομένα από μικροσυστοιχίες (Sass, Dabrowski, Charzynska & Sachadyne, 2017).

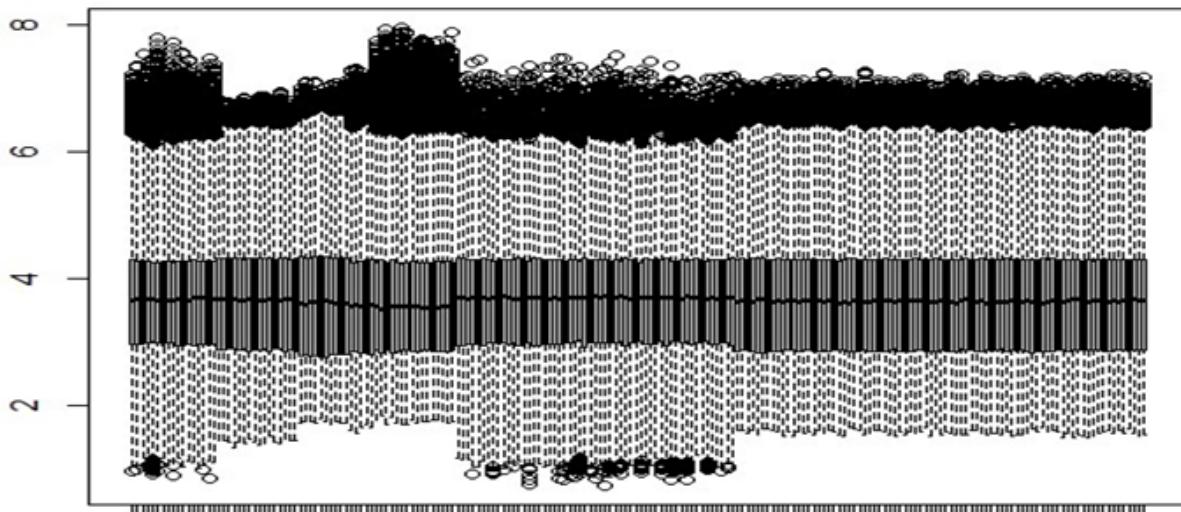
4.3 Στρατηγικές Επίλυσης Προβλημάτων Ομογενοποίησης

Για την επίλυση των παραπάνω προβλημάτων, και μετά από προσεκτικό έλεγχο των δεδομένων με τα οποία ασχολείται αυτή η διπλωματική επιλέχθηκαν 2 βασικές προσεγγίσεις στρατηγικών αντιμετώπισης. Αρχικά πρέπει να βρεθεί ένα είδος κατανομής που να μπορεί να συγκριθεί με κατανομές του ίδιου είδους, αλλά διαφορετικών παραμέτρων. Ένα τέτοιο είδος κατανομής είναι η κανονική κατανομή και οι παράμετροι της (μέσος όρος, τυπική απόκλιση) που μπορούν ικανοποιητικά να «τυποποιηθούν» έτσι ώστε όλες να έχουν τον ίδιο μέσο όρο και η εκάστοτε τυπική απόκλιση να μετριέται πλέον σε «μονάδες τυπικής απόκλισης». Αυτό μπορεί να γίνει με την z – κανονικοποίηση (Cheadle, Cho-Chung, Becker & Vawter, 2003 ; Cheadle,Vawter,Freed & Becker, 2003). Η συγκεκριμένη μεθοδος κανονικοποίησης έχει ξεχαστεί από την επιστημονική κοινότητα των μικροσυστοιχιών υπέρ της QN, μία μέθοδος σαφώς ισχυρότερη. Μπορεί εύκολα να αντιληφθεί κάποιος ότι σε μία μετα-ανάλυση όταν έχει ήδη γίνει quantile normalization κανονικοποίηση σε επίπεδο dataset (within dataset), μία δεύτερη quantile normalization σε επίπεδο συνόλου dataset (between datasets) θα οδηγούσε σίγουρα σε μία μη επιθυμητή παραμόρφωση των δεδομένων και πιθανόν και σε απώλεια

πληροφορίας. Αυτός είναι και ο λόγος που η πρώτη προσέγγιση αντιμετώπισης του προβλήματος ομογενοποίησης περιλαμβάνει μεν QN στα πλαίσια της προεπεξεργασίας με RMA σε κάθε dataset, αλλά για την ανομοιομορφία σε επίπεδο συνόλου πολλών dataset έγινε z- κανονικοποίηση, η οποία βέβαια δεν είναι τόσο ισχυρή όσο η QN με την έννοια ότι δεν έχουμε απόλυτη ευθυγράμμιση τυπικών αποκλίσεων και μέσω όρων συνολικά. Όμως παρόλα αυτά παραμορφώνει τα δεδομένα σε μικρότερο βαθμό από την επιλογή της διαδοχικής διπλής QN (successive within and between datasets) ώστε να μπορούν να αναδειχθούν οι διαφοροποιήσεις στην γονιδιακή έκφραση που θέλουμε να μελετήσουμε. Η μοναδική αλλά αναγκαία προϋπόθεση για την εφαρμογή z-κανονικοποίησης είναι τα δεδομένα να είναι ήδη σε κανονική κατανομή ή τουλάχιστον να έχουμε μία καλή προσέγγιση σε αυτήν (Νικολάου & Χουβαρδάς, 2015). Η εφαρμογή z-κανονικοποίησης θεωρείται από κάποιες μελέτες κατάλληλη ιδιαίτερα για τα dataset που χρησιμοποιούν μικροσυστοιχίες της εταιρίας Affymetrix μιας και η κατανομή τους προσομοιάζει περισσότερο την κανονική κατανομή (Cheadle et al, 2003).

Με τον παραπάνω τρόπο δημιουργείται πλέον ένας μεγάλος πίνακας έκφρασης στον οποίον έχουμε πολλά δείγματα. Σε αυτό μπορούν να γίνουν υψηλότερου επιπέδου πλέον αναλύσεις σαν να είχαμε ένα dataset με πολλά δείγματα. Ασφαλώς δεν πρόκειται για έναν ιδανικό τρόπο, η δυσκολία να ενσωματωθούν dataset από πολλές και διαφορετικών εταιριών πλατφόρμες παραμένει. Όπως και παραμένει το γεγονός ότι εξαιρετικά σπάνια έχουμε μία κατανομή σε ένα microarray που να είναι εντελώς κανονική, πράγμα που μπορεί να διαπιστωθεί με κάποιο τεστ κανονικότητας όπως το Shapiro-Wilk ή το Kolmogorov-Smirnov test (Gerald, 2018). Αυτός είναι ίσως και ο μεγαλύτερος περιορισμός αυτής της πρώτης προσέγγισης. Στην εικόνα 18, παρατηρούμε μία γραφική αναπαράσταση με διάγραμμα τύπου Boxplot, των δεδομένων που μελετήθηκαν σε αυτή την διπλωματική και που έχουν υποστεί προεπεξεργασία με την παραπάνω προσέγγιση.

Boxplot z-transformation



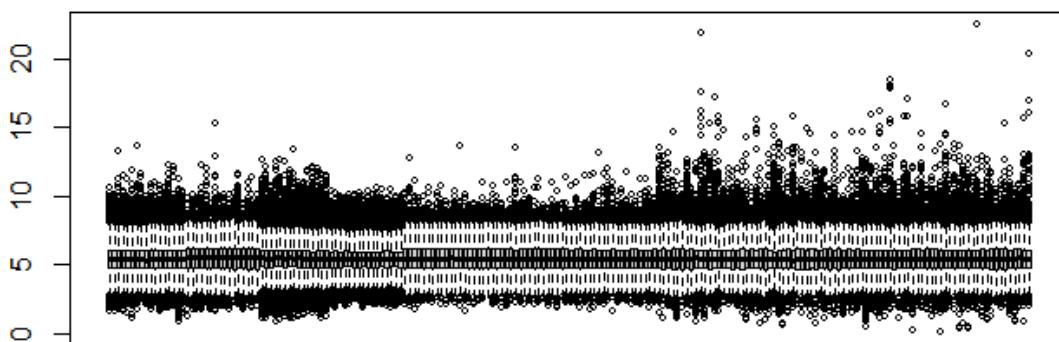
Εικόνα 18. 200 δείγματα από διαφορετικά datasets, επεξεργασμένα με την προσέγγιση της απλής ζ-κανονικοποίησης κατά στήλες και ανυψωμένα με τέτοιο τρόπο ώστε να είναι όλα θετικά.

Μία παραλλαγή του της απλής ζ-κανονικοποίησης ανά στήλη δείγματος, είναι η εφαρμογή ζ-κανονικοποίησης πρώτα σε κάθε γονίδιο (σειρά πίνακα έκφρασης) και εν συνεχείᾳ ανά στήλη. Η σειρά αυτή έχει σημασία γιατί αν γίνει κανονικοποίηση πρώτα στα δείγματα και μετά στα γονίδια προκαλείται εμφανή σε boxplot διαταραχή τόσο στους μέσους όρους όσο και στις κατανομές σε κάθε δείγμα. Αυτή η παραλλαγή ονομάζεται σε αυτή την διπλωματική εργασία «διπλή ζ-κανονικοποίηση».

Γενικά η λογική της διπλής ζ-κανονικοποίησης έγκειται στην ισοβαρή αντιμετώπιση των μεταβλητών του dataset ανεξαρτήτως της κλίμακας στην οποία αυτές μετρώνται. Έστω ότι έχουμε ένα dataset στο οποίο θέλουμε να συγκρίνουμε έναν πληθυσμό με βάσει ας πούμε 2 μεταβλητές, το ύψος σε εκατοστά και την ηλικία. Αυτά τα 2 μεγέθη είναι διαφορετικής κλίμακας αλλά και διαφορετικής διασποράς τιμών. Το ύψος έχει σίγουρα πολύ μεγαλύτερες τιμές σε εκατοστά από ότι η ηλικία, αυτό έχει ως αποτέλεσμα το ύψος να έχει μεγαλύτερη βαρύτητα από ότι η ηλικία και επομένως σε μία απόπειρα σύγκρισης διαφορετικών ατόμων, θα υπερνικούσε σε βαρύτητα την μεταβλητή της ηλικίας. Επομένως για να αποφευχθεί αυτό τα δεδομένα πριν συγκριθούν κανονικοποιύνται. Αν θέσουμε το παραπάνω πρόβλημα σε περισσότερες διαστάσεις φτάνουμε στο συμπέρασμα ότι ιδανικά θα πρέπει να κανονικοποιήσουμε δεδομένα και ανά γραμμές και ανά στήλες. Ωστόσο όπως ήδη έχει

αναφερθεί οι πολλές κανονικοποιήσεις δημιουργούν αναπόφευκτα παραμόρφωση των δεδομένων και πιθανή απώλεια πληροφορίας. Και ενώ για το παραπάνω ακραίο παράδειγμα έχουμε μία σημαντική διαφορά κλίμακας, δεν φαίνεται να είναι και τόσο συνηθισμένη μία τόσο μεγάλη διαφορά κλίμακας να είναι εμφανής μεταξύ διαφορετικών γονιδίων σε μία σειρά από μικροσυστοιχίες. Παρόλα αυτά και αυτή η παραλλαγμένη ζ -κανονικοποίηση εφαρμόστηκε και αξιολογήθηκε στα δεδομένα μας, όπως φαίνεται στην εικόνα 19. Οι μεγάλες τυπικές αποκλίσεις που προκύπτουν όμως μας προδιαθέτουν αρνητικά και μειώνουν την αισιοδοξία μας ως προς την έκβαση αυτής της προσέγγισης.

Boxplot double-z



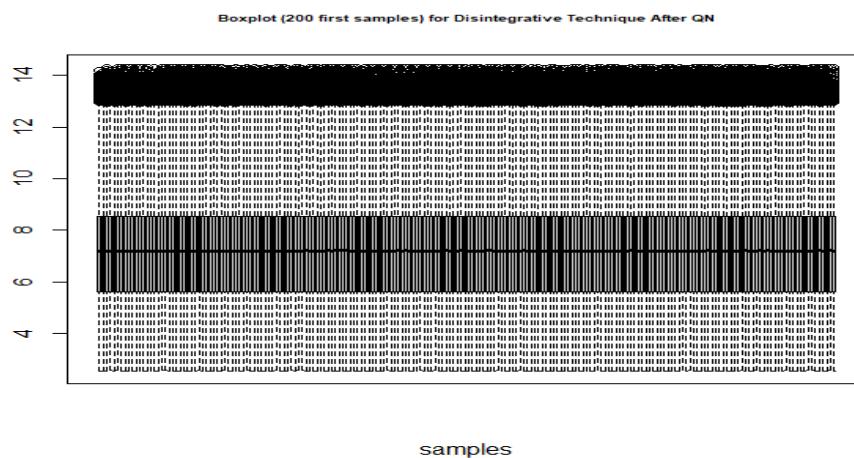
Εικόνα 19. Boxplot των πρώτων 200 δειγμάτων στα δεδομένα μας, μετά από διπλή ζ -κανονικοποίηση, είναι εμφανείς οι μεγάλες τυπικές αποκλίσεις που δημιουργούνται.

Μία δεύτερη διαφορετική αυτή την φορά προσέγγιση θα μπορούσε να είναι μία μεταβολή στην προσέγγιση για ομογενοποίηση σε επίπεδο μετατροπής ενός αντικειμένου από RAW τιμές έντασης φθορισμού στο microarray laser σκάνερ (για την Affymetrix ένα τέτοιο αντικείμενο λέγεται AffyBatch) σε ένα αντικείμενο τύπου πίνακα γονιδιακής έκφρασης (για την Affymetrix ένα τέτοιο αντικείμενο λέγεται ExpressionSet). Σε μικροσυστοιχίες της Affymetrix η κυρίως διαδικασία προεπεξεργασίας που συνιστάται είναι όπως

προαναφέρθηκε η μέθοδος RMA (Irrizary et al, 2003). Η μέθοδος αυτή περιλαμβάνει 4 στάδια:

1. Log2 transformation
2. Background Correction
3. Summarization
4. QN

Η δεύτερη προσέγγιση λοιπόν αφορά την εφαρμογή σε επίπεδο dataset όλων των παραπάνω σταδίων εκτός από το τελευταίο με το QN. Έν συνεχεία όταν τοποθετηθούν κατά την σύμπτυξη τα dataset το ένα δίπλα στο άλλο και δημιουργηθεί πλέον ένας μεγάλος ενιαίος πίνακας έκφρασης, τότε σε αυτό το στάδιο μπορεί να εφαρμοστεί QN ώστε να ευθυγραμμιστούν τα δεδομένα. Θα ονομάσουμε αυτή την δεύτερη προσέγγιση σε αυτή την διπλωματική εργασία ως «Διασπαστική» (Disintegrative) μέθοδο μετα-ανάλυσης μικροσυστοιχιών. Το boxplot του τελικού πίνακα έκφρασης παρατίθεται στην Εικόνα 20, όπου αντίθετα με την πρώτη προσέγγιση φαίνεται πλήρης εναρμόνιση των δεδομένων. Από όσο γνωρίζω, η συγκεκριμένη προσέγγιση, αν και διαισθητικά λογική δεν αναφέρεται πουθενά στην βιβλιογραφία ως μέθοδος μετα-ανάλυσης μικροσυστοιχιών.



Εικόνα 20. Boxplot με τα πρώτα 200 δεδομένα όπως αυτά προέκυψαν από την διασπαστική μέθοδο.



Πρέπει να τονιστεί εδώ ότι και στις 2 προσεγγίσεις, χρειάζεται προσεκτικός έλεγχος των δεδομένων για να διαπιστωθεί ότι έχουν παρόμοιες κατανομές. Δεδομένα με διαφορετικές κατανομές οφείλονται σε διαφορετικά «batch effects» στις διάφορες πλατφόρμες, με άλλα λόγια σε διαφοροποιήσεις από πλατφόρμα σε πλατφόρμα ακόμα και αν αυτά είναι της ίδιας εταιρίας. Γι' αυτό άλλωστε σε αυτή την διπλωματική έγινε γραφική απεικόνιση όλων των δεδομένων ώστε να διαπιστωθεί ότι δεν υπάρχουν σημαντικές αποκλίσεις μεταξύ τους και να αντιμετωπιστεί το υπαρκτό πρόβλημα των διαφορετικών batch effect.

II.

Ειδικό Μέρος

Κεφάλαιο 1. Σκοπός της εργασίας

Η πραγματοποίηση αυτής της εργασίας έγινε με σκοπό την σκιαγράφηση του προφίλ γονιδιακής έκφρασης του ακανθοκυτταρικού καρκινώματος του στόματος (ΑΚΚ) όπως προκύπτει από ελεύθερα διαθέσιμα δεδομένα στην γνωστή βάση δεδομένων από μικροσυστοιχίες, Gene Expression Omnibus. Για να γίνει αυτό ήταν απαραίτητο αρχικά να επιλεγεί μια κατάλληλη μέθοδος μετα-ανάλυσης τέτοιων δεδομένων. Η αξιολόγηση των μεθόδων μετα-ανάλυσης μικροσυστοιχιών και ιδιαίτερα η νέα τεχνική μετα-ανάλυσης που εισάγαμε πιθανώς θα βοηθήσουν άλλους ερευνητές στο μέλλον και ευχόμαστε ότι θα αποτελέσουν εφαλτήρες για νέες και συναρπαστικές ανακαλύψεις πάνω στους τομείς της μοριακής βιολογίας και μεταγραφωμικής. Τα αποτελέσματα μας όσον αφορά συγκεκριμένα το ΑΚΚ ευελπιστούμε να βοηθήσει όσους ερευνητές ασχολούνται με την παθολογία και την χειρουργική στόματος στο να εστιάσουν σε συγκεκριμένα γονίδια και βιολογικά μονοπάτια για περαιτέρω μελέτη και διερεύνηση.

Η μετα-ανάλυση των δεδομένων από μικροσυστοιχίες στο ακανθοκυτταρικό καρκίνωμα του στόματος πραγματοποιήθηκε με την βοήθεια:

- Στατιστικών Μεθόδων
- Μεθόδων λειτουργικής ανάλυσης
- Εργαλείων χαρτογράφησης σε βιολογικούς χάρτες
- Μεθόδων μηχανικής μάθησης που στηρίζονται σε δέντρα και δάση
- Μεθόδων ανάλυσης τοπολογίας δικτύων
- Αντιπαραβολής των παραπάνω αποτελεσμάτων με την βιβλιογραφία.

Κεφάλαιο 2. Υλικό και Μέθοδος

2.1 Συλλογή σχετικών Accession numbers και αρχικό φιλτράρισμα

Τα δεδομένα προέρχονται από την βάση δεδομένων Gene Expression Omnibus (GEO) (<https://www.ncbi.nlm.nih.gov/gds>) (Barret, Wilhite, Ledoux, Evangelista, Kim, Tomashevsky , ... & Soboleva, 2013). Οι όροι αναζήτησης ήταν:

1.GEO Datasets search: “oral squamous cell carcinoma” και “squamous cell carcinoma” AND “oral” , επίσης μία ξεχωριστή αναζήτηση έγινε και για τα ακανθοκυτταρικά καρκινώματα της γλώσσας με τους όρους αναζήτησης “squamous cell carcinoma” AND “tongue”. Αυτό έγινε γιατί σε περιλήψεις με καρκινώματα αυτής της εντόπισης, ο όρος “oral” δεν χρησιμοποιούνταν ιδιαίτερα, επίσης να σημειωθεί ότι η γλώσσα είναι η συχνότερη εντόπιση του ακανθοκυτταρικού καρκινώματος στο στόμα.

2.Study Type : Expression Profiling by Array

3.Organism: Homo Sapiens

4.Number of Samples=6-100.000

Η παραπάνω αναζήτηση έβγαλε 74 αποτελέσματα στην πρώτη περίπτωση και 118 στην δεύτερη, ενώ στη περίπτωση της γλώσσας είχαμε 62 αποτελέσματα. Τα παραπάνω αποτελέσματα έχουν αλληλοεπικάλυψη , τελικά εξετάστηκαν συνολικά 119 αποτελέσματα, τα οποία βρίσκονται στο παράρτημα 1, το παράρτημα βρίσκεται σε μορφή tab separated file ώστε να μπορεί εύκολα να γίνει επεξεργασία του με χρήση κάποιας γλώσσας προγραμματισμού ύστερα από αντιγραφή επικόλληση σε έναν οποιοδήποτε απλό text editor κατάλληλο για αυτή την δουλειά (π.χ. notepad ή notepad ++ κτλ.)

Στο παράρτημα 1 βλέπουμε 7 στήλες

- *Geo*: που είναι το GEO accession number κάθε dataset (π.χ. GSE97251)
- *Osc*: ο αριθμός των ακανθοκυτταρικών καρκινωμάτων που εντοπίζεται στην περιοχή που εξετάζουμε.
- *Normal*: ο αριθμός δειγμάτων φυσιολογικού βλενογόννου.
- *No_platforms*: ο αριθμός των πλατφόρμων που περιέχει το κάθε dataset

- *Primary_platform*: η πλατφόρμα που μας ενδιαφέρει σε αυτή την μελέτη, που είναι πλατφόρμες έκφρασης γονιδίων. Δεν χρησιμοποιήθηκαν σε αυτή την φάση, πλατφόρμες όπως μεθυλίωσης ή SNP (single nucleotide polymorphism).
- *Acception*: η επιμηγορία βάσει των κριτηρίων που αναλύονται αμέσως παρακάτω.
- *Info*: ο λόγος απόρριψης αν πρόκειται για απορριφθέν dataset ή άλλες ενδιαφέρουσες πληροφορίες, που πιθανόν να είναι χρήσιμες σε κάποιο άλλο στάδιο (ή άλλη σχετική μελέτη).

Από αυτά έγινε χειροκίνητη επισκόπηση, μελετήθηκαν δηλαδή ένα – ένα και αποκλείστηκαν τα εξής αποτελέσματα:

1. Μελέτες στα οποία ο ιστός που μελετάται δεν ήταν συμβατός με ΑΚΚ του στόματος.
2. Μελέτες που αφορούν δείγματα ΑΚΚ από βιοψίες λεμφαδένων.
3. Κυτταρικές σειρές ΑΚΚ στις οποίες έχει γίνει στοχευμένη επιμόλυνση από τροποποιημένο ιϊκό γονιδίωμα (transfection).
4. Μελέτες με διάφορες κυτταρικές σειρές (cell-lines).
5. Μελέτες στις οποίες έγινε knock-out τροποποίηση κάποιων γονιδίων του γενετικού υλικού.
6. Μελέτες που αφορούσαν περιοχή εκτός στοματικού βλεννογόνου. Ως στοματικός βλεννογόνος ορίζεται η ανατομική περιοχή που αφορά τον βλεννογόνο της στοματικής κοιλότητας που αφορίζεται πρόσθια από το ερυθρό κράσπεδο των χειλέων χωρίς να το περιλαμβάνει και οπίσθια από το γλωσσο-υπερώιο τόξο (δεν συμπεριλαμβάνεται ο φάρυγγας και οι αμυγδαλές). Εδώ να πούμε ότι καρκινώματα βάσης της γλώσσας παρόλο που δεν ανήκουν αυστηρά στις παραπάνω περιοχές λήφθηκαν υπόψη γιατί πρόκειται για συμπαγές όργανο του στόματος, το οποίο εκτείνεται σε ένα μικρό μέρος του πέραν τις ανατομικής περιοχής που ορίστηκε. Επίσης στις μελέτες πολλές φορές αναφέρεται ως «καρκίνος της γλώσσας» χωρίς να διευκρινίζεται σε ποιο μέρος της γλώσσας εντοπίζεται ο όγκος. Το ίδιο ισχύει και στην περίπτωση της μαλακής υπερώας όπου δείγματα από την μαλακή υπερώα λήφθηκαν υπ’όψη στην μελέτη.
7. Μελέτες με ανεπαρκή αριθμό δειγμάτων, ως επαρκής αριθμός δειγμάτων θεωρήθηκε πάνω από 4 ώστε σε αρχική φάση να μπορεί να γίνει τουλάχιστον κάποια στατιστική δοκιμασία ελέγχου υποθέσεων.

Να διευκρινιστεί ότι δείγματα από δυσπλασίες, λευκοπλακίες ή δείγματα εκτός της περιοχής που εξετάζουμε απορρίφθηκαν μεν αλλά όπου υπήρχαν έγινε σημείωση στην στήλη Info του παραρτήματος 1. Επίσης καρκινώματα in-situ, συμπεριελήφθηκαν και έγιναν δεκτά στην μελέτη αυτή ως ακανθοκυτταρικά καρκινώματα.

2.2 Αρχική επεξεργασία δεδομένων και συλλογή των datasets

Για όλα τα υπόλοιπα στάδια που θα περιγραφούν έγινε χρήση της γλώσσας προγραμματισμού R γιατί είναι μία γλώσσα με εκτεταμένη σειρά βιβλιοθηκών βιοπληροφορικής (Bioconductor). Επιλέχθηκαν όλα τα δεκτά αρχεία (“accepted”) βάσει της στήλης «acception» με δείγματα από 10 και πάνω, όπου GEO_initial.txt αρχείο που κάναμε αντιγραφή/επικόλληση σε editor από το παράρτημα 1. Το αποτέλεσμα του παρακάτω απλού κώδικα ήταν ουσιαστικά 39 datasets οι λεπτομέρειες των οποίων φαίνονται στο παράρτημα 2. Επίσης προς το τέλος του παραρτήματος 2 φαίνεται και ο πολύ απλός κώδικας που χρησιμοποιήθηκε για το εν λόγω φίλτραρισμα των δεδομένων. Σε αυτό το σκέλος της έρευνας δεν δόθηκε σημασία στο αν το dataset έχει τόσο επαρκή φυσιολογικά όσο και επαρκή καρκινικά δείγματα. Με άλλα λόγια έγιναν δεκτά datasets τα οποία είχαν μόνο AKK χωρίς καθόλου φυσιολογικά δείγματα, και το αντίστροφο. Αυτό συνέβη γιατί θέλουμε αρχικά να κάνουμε ένα matrix με πολλαπλά ενσωματωμένα δείγματα. Για αυτό τον λόγο επιλέχθηκε τελικά σε αυτή την φάση να έχουμε έναν επαρκή αριθμό δειγμάτων (από 10 και πάνω) ώστε σε αυτό το τελικό matrix να μην υπάρχει μεγάλος βαθμός ετερογένειας των δεδομένων. Με βάση τα παραπάνω προέκυψαν 1531 παθολογικά δείγματα και 301 φυσιολογικά. Όμως πρέπει να τονιστεί εδώ ότι στα επόμενα στάδια φιλτραρίσματος αυτά θα μειωθούν περαιτέρω (π.χ. λόγω dataset με κάποια μεμονωμένα δείγματα μη καλά εναρμονισμένα, ή κάποια δείγματα έχουν πολλές κενές τιμές και επομένως δεν μπορούν να χρησιμοποιηθούν κτλ).

Δεδομένα των παραπάνω dataset, που συμπεριλαμβάνονται στο παράρτημα 2, έγιναν download από το GEO website. Πιο συγκεκριμένα αποθηκεύτηκαν τα series matrix, που εμπεριέχουν τις «κανονικοποιημένες τιμές». Θα πρέπει να τονιστεί ότι ο όρος «κανονικοποίηση» χρησιμοποιείται στα microarrays με την πολύ ευρεία έννοια του όρου,

πολύ σπάνια τα δεδομένα είναι πραγματικά δεδομένα με κανονική κατανομή ακόμα και με τις πιο επιθετικές κανονικοποιήσεις, έτσι λοιπόν σε αυτή την εργασία όταν αναφέρουμε ότι τα δεδομένα είναι κανονικοποιημένα το εννοούμε με την πολύ ευρεία έννοια του όρου (βλ. γενικό μέρος για αναλυτική περιγραφή).

Τα δεδομένα αυτά εξετάστηκαν με γραφικές μεθόδους αλλά και μελετώντας τις πληροφορίες που αναφέρονται σχετικά με την στατιστική επεξεργασία των δεδομένων όπως αναφέρονται αυτές στις πρώτες σειρές των series matrix files.

Προέκυψαν όπως προαναφέρθηκε τελικά 39 datasets από (από τα οποία τα 2 ήταν διπλής πλατφόρμας, εξετάστηκαν δηλαδή και οι 2 πλατφόρμες που εμπειριείχαν) από αυτά απορρίφθηκαν 2 λόγω χαμηλού αριθμού παραμέτρων-γονιδίων και πιο συγκεκριμένα απορρίφθηκαν τα GSE148944 γιατί είχε μόνο 784 probe IDs και GSE30788-GPL13953 με 4096 probe IDs. Αυτό έγινε για να μην χάσουμε πληροφορία ή διαφορετικά γιατί θεωρούμε ότι αυτά τα λίγα probes μπορεί και να μην είναι απολύτως αντιπροσωπευτικά για την ταξινόμηση των δεδομένων σε ένα συνολικό επίπεδο.

Ακολούθησε η αναλυτική καταγραφή των ιδιοτήτων του κάθε dataset, αυτό έγινε επίσης με εξέταση των μεταδεδομένων που συμπεριλαμβάνονται στις ιστοσελίδες κάθε dataset στο GEO. Ως ιδιότητα ορίζεται το ποια δείγματα ακριβώς σε ένα dataset μπορούν να χρησιμοποιηθούν, ποια είναι τα παθολογικά και ποια τα φυσιολογικά δείγματα. Αυτά παίρνουν όλα την μορφή ενός διανύσματος. Στο παράρτημα 3 φαίνεται το αρχείο ιδιοτήτων που χρησιμοποιείται σε αυτή την εργασία. Το αρχείο έχει 3 στήλες και είναι επίσης tab separated για να μπορεί να χρησιμοποιηθεί ως αρχείο εισόδου στην γλώσσα προγραμματισμού R:

- 1.GEO_ID: το GEO accession number κάθε dataset (π.χ. GSE97251).
- 2.Properties: που είναι ένα διάνυσμα γραμμένο σε γλώσσα R, το οποίο είναι μία ακολουθία από 3 σύμβολα. 1 για τους καρκίνους, 0 για τα φυσιολογικά δείγματα και X για τα δείγματα που δεν θα χρησιμοποιηθούν τελικά (για τα οποία τα μεταδεδομένα δείχνουν ότι είναι τελικά δείγματα από άλλη περιοχή από αυτή που μας ενδιαφέρει «αμυγδαλή», «υποφάρυγγας» κτλ όπου τα αρχικά και πριν την προσεκτική ανάλυση των δεδομένων είχαν αρχικά οριστεί απλά ως π.χ. oral cancer).

3.Info : χρήσιμες πληροφορίες σχετικά με κάθε dataset που τελικά κάποιες από αυτές χρησιμοποιήθηκαν σε περαιτέρω ανάλυση με ANOVA (Analysis of Variance) σε δεύτερο χρόνο. Αυτές αφορούν είτε σε σταδιοποίηση , είτε σε ιστολογικά χαρακτηριστικά (grading) του όγκου, είτε στην παρουσία ιών HPV ή όχι. Δεν έχουν όλα τα dataset τις ίδιες πληροφορίες, αυτές οι χρήσιμες πληροφορίες υπάρχουν δυστυχώς σε λίγα datasets και έχουν μεγάλο βαθμό ανομοιομορφίας, το ένα dataset έχει πληροφορία για HPV θετικότητα, άλλο για grading άλλο για συνδυασμό πληροφοριών κτλ. Όπως θα δούμε παρακάτω το χαρακτηριστικό που απαντάται συχνότερα στα dataset που συλλέξαμε ήταν η κλινική σταδιοποίηση του όγκου και πάνω σε αυτό το χαρακτηριστικό εφαρμόστηκε επιμέρους στατιστική ανάλυση με ANOVA.

2.3 Δημιουργία τυποποιημένων αρχείων , προβλήματα και χειρισμοί που χρησιμοποιήθηκαν στην κατασκευή τους

Πέρα από τις ιδιότητες κάθε αρχείου εξετάστηκε και το είδος των χαρακτηριστικών (feature data) του κάθε GEO dataset. Δηλαδή ποιά γονίδια αυτό εξετάζει και με ποιο είδος ID τα περιλαμβάνει. Εμείς σε αυτή την εργασία εργαστήκαμε αποκλειστικά με Entrez IDs. Για να τυποποιήσουμε ένα series matrix ωστόσο πρέπει να τακτοποιηθούν και οι στήλες με τα δείγματα. Ο κώδικας που χρησιμοποιήθηκε σχεδιάστηκε έτσι ώστε στις πρώτες στήλες να μπαίνουν οι καρκίνοι με ονοματολογία (C1,C2,...) και εν συνεχείᾳ ακολουθούν οι φυσιολογικοί ιστοί με ονοματολογία (N1,N2...). Αυτό θα κάνει σε επόμενα στάδια ευκολότερη την ενσωμάτωση τους αλλά και τυποποιούν κατά κάποιον τρόπο τους πίνακες έκφρασης.

Ένα τυποποιημένο series matrix στην εργασία αυτή έχει μία στήλη που είναι τα Entrez IDs (δηλαδή πρόκειται στην ουσία για data frame) και οι επόμενες στήλες, είναι με την ονοματολογία όπως ορίστηκε παραπάνω. Σε ένα τυποποιημένο series matrix είναι δυνατόν να υπάρχουν σειρές με πανομοιότυπες τιμές στις διάφορες στήλες γιατί το πρόβλημα “Many to Many” προσεγγίστηκε με κώδικα που διπλασιάζει τα feature IDs που αντιστοιχούν σε πολλαπλά Entrez IDs. Τέλος σε όσα probes δεν υπάρχει αντιστοίχηση σε κανένα Entrez ID αυτόματα οι σειρές αυτές απορρίπτονται από την περαιτέρω ανάλυση. Να επισημανθεί εδώ ότι από εδώ και πέρα οι όροι Entrez ID και γονίδια θα είναι ταυτόσημες έννοιες για αυτή την

διπλωματική. Τα αρχικά βήματα για την συλλογή των δεδομένων συνοψίζονται στην Εικόνα 21.



Εικόνα 21. Pipeline Βημάτων συλλογής και επεξεργασίας αρχικών δεδομένων. Οι αριθμοί απεικονίζουν τον αριθμό των dataset στα επιμέρους στάδια.

2.4 Γραφική Παρατήρηση των δεδομένων

Τα 37 datasets που προέκυψαν εν συνεχείᾳ παρατηρήθηκαν γραφικά. Ωστόσο μετά από την ανάλυση που περιγράφηκε στην παραπάνω ενότητα είχαμε περαιτέρω, μείωση σε κάποια dataset των αριθμού των γονιδίων, που εξετάζονται. Αυτό συνέβη λόγω συγχώνευσης πολλών probes που αντιστοιχούσαν σε ένα μόνο αριθμό Entrez ID. Τα dataset των οποίων ο αριθμός εξεταζόμενων γονιδίων έπεσε κάτω από το αρχικό όριο που είχαμε ορίσει (10.000), αφαιρέθηκαν και αυτά από την συνέχεια αυτής της εργασίας. Τα dataset αυτά ήταν τα GSE13601 (με 9629 Entrez IDs) , GSE138206 (με 6754 Entrez IDs) και GSE51010-GPL201 (με 8969 Entrez IDs). Από τα 34 dataset που παρέμειναν, έγιναν Ιστογράμματα (στο 3^ο δείγμα κάθε dataset, επιλέχθηκε τυχαία αυτός ο αριθμός γιατί τα δεδομένα ήταν

ομογενοποιημένα πράγμα που επαληθεύτηκε και γραφικά με τα boxplots) και Q-Q διαγράμματα προκειμένου να εξετασθεί η κατανομή τους και κατά πόσο αυτά προσεγγίζουν την κανονική κατανομή. Επίσης έγιναν και boxplots όπως προαναφέρθηκε για πιο συνολική θεώρηση του dataset. Διαπιστώθηκε ότι τα microarray της Affymetrix ιδιαίτερα όταν αυτά είχαν προ-επεξεργαστεί με την τεχνική προεπεξεργασίας RMA (Robust Microarray Analysis) προσεγγίζουν σε μεγάλο βαθμό την κανονική κατανομή ενώ από τα microarray της Agilent μόνο ένα dataset (GSE40774) προσέγγιζε την κανονική κατανομή. Επίσης ένα dataset (GSE4676) με microarray της εταιρίας GE Healthcare επίσης προσέγγιζε την κανονική κατανομή. Ταυτόχρονα παρατηρήθηκε ότι η συντριπτική πλειοψηφία των dataset αφορούσε δεδομένα από Affymetrix ενώ ακολουθούν κατά πλειοψηφία τα dataset με δικάναλα microarray της εταιρίας Agilent. Στο φως των παραπάνω παρατηρήσεων αποφασίστηκε η μετα-ανάλυση να συνεχιστεί με τα εναπομείναντα (και πολυαριθμότερα) dataset με Affymetrix microarrays τα οποία ήταν 16 και παρουσιάζονται στον πίνακα 3. Τα παραπάνω βήματα συνοψίζονται στην Εικόνα 22.

GSE2280 (O' Donnel et al, 2005)	GSE52915 (Eslami et al, 2015)
GSE3524 (Toruner et al, 2004)	GSE56532
GSE25104 (πλ. GPL5175) (Peng et al, 2004)	GSE78060
GSE30784 (Chen et al, 2008)	GSE103412
GSE31056 (Reis et al, 2011)	GSE107591 (Saconi et al, 2020)
GSE41116 (Pickering et al, 2013)	GSE109756
GSE41613 (Lohavanichbutr et al, 2013)	GSE153918
GSE51010 (πλ. GPL570) (Saeed et al, 2015)	GSE74530 (10 δείγματα, διασπαστική μόνο) (Oghumu et al, 2016)

Πίνακας 3. Τελικά GEO Accession Numbers.

Υπολογισμός του αριθμού feature μετά την
συγχώνευση και φίλτραρισμα ξανά

34

Γραφική Μελέτη και διατήρηση μόνο όσων
προσέγγιζαν την κανονική κατανομή

18

Επιλογή εταιρίας με τα περισσότερα arrays

16

Εικόνα 22. Pipeline γραφικής απεικόνιση κατανομών και φίλτραρισμα των δεδομένων. Οι αριθμοί υποδηλώνουν τον αριθμό datasets ανά στάδιο.

2.5 Προεπεξεργασία Δεδομένων των επιλεχθέντων datasets

Από τα 16 datasets , έγινε download όλων των .CEL αρχείων που αφορούσαν τα συγκεκριμένα datasets. Σε 13 από αυτά χρησιμοποιήθηκε το πακέτο προεπεξεργασίας Affy (Gautier, Cope, Bolstad & Irizarry, 2004), ενώ σε 3 χρησιμοποιήθηκε το πακέτο προεπεξεργασίας για νεότερους τύπους microarray cDNA ολιγονουκλεοτιδίων της Affymetrix , oligo (Carvalho & Irizarry, 2010). Το πακέτο oligo γενικά έχει μεγαλύτερες απαιτήσεις σε μνήμη και σε μεγάλα dataset τείνει να δημιουργεί προβλήματα που σχετίζονται με ανεπάρκεια μνήμης ακόμα και σε υπολογιστές με σχετικά μεγάλη RAM (16GB). Από αυτά τα CEL αρχεία και βάσει του αρχείου properties.txt, διαγράφηκαν samples τα οποία δεν ανταποκρίνονταν στα κριτήρια που είχαμε αρχικά θέσει (π.χ. μπορεί να είναι samples από λαρυγγα), αυτά τα είχαμε σταμπάρει με X στο παραπάνω αρχείο. Είναι σημαντικό να τονιστεί ότι διαγράφουμε από το αρχείο properties τα αντίστοιχα X για να τρέξει ένα πρόγραμμα όπως αυτό στο παράρτημα 4, πριν το επικολλήσουμε στην αντίστοιχη μεταβλητή. Στο παράρτημα 4 υπάρχει ένα ενδεικτικό πρόγραμμα επεξεργασίας με το πακέτο Affy για το dataset GSE78060 , επιλεχθήκε αυτό το dataset γιατί περίεχει τόσο καρκίνους

όσο και φυσιολογικά δείγματα αλλά κυρίως λόγω του μικρού μεγέθους του μπορεί να τρέξει σε οποιονδήποτε υπολογιστή.

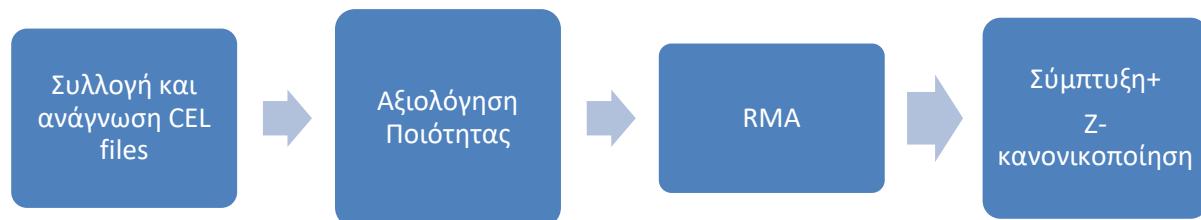
Στην συνέχεια ακολούθησε ανάλυση ποιότητας των επιμέρους δειγμάτων σε επίπεδο probes , των dataset με την βοήθεια εργαλείων που εμπεριέχονται στην βιβλιοθήκη AffyPLM (Bolstad, 2004 ; Bolstad, Collin, Brettschneider, Simpson, Cope, Irizarry & Speed, 2005 ; Brettschneider, Collin, Bolstad, Speed, 2007) και oligo. Εδώ πρέπει να διευκρινιστεί ότι κάθε εταιρία έχει διαφορετικά εργαλεία για την ποιοτική αξιολόγηση των μικροσυστοιχιών της. Στην περίπτωση της Affymetrix τα πιο διαδεδομένα είναι τα RLE και NUSE, με τα οποία έγινε τελικά και η αξιολόγηση ποιότητας. Αυτή έγινε με την γραφική επισκόπηση των παραπάνω παραμέτρων με την βοήθεια boxplots, αν το boxplot φαίνοταν αρκετά ανυψωμένο στο διάγραμμα NUSE ή/και αν φαίνοταν με μεγάλο Interquartile Range (IQR) ή με απόκλιση του μέσου όρου στο διάγραμμα RLE, τότε αυτό απορρίφθηκε για να αποφύγουμε να προκληθεί φθορά των δεδομένων (corruption) όταν ομογενοποιηθούν κατά την διαδικασία της κανονικοποίησης, αλλά και για να μειώσουμε δραστικά τα προβλήματα με outliers τα οποία αυτά δημιουργούν.

Εν συνεχεία τα δεδομένα κανονικοποιήθηκαν με RMA. Στην πρώτη προσέγγιση (z-score) έγιναν όλα τα στάδια της RMA, ενώ στην δεύτερη (διασπαστική) έγιναν όλα τα στάδια πλην της κανονικοποίησης. Στην συνέχεια έγινε annotation των features με την βοήθεια της βιβλιοθήκης του Bioconductor, mygene, ωστόσο σε μία περίπτωση (GSE 153918) η παραπάνω βιβλιοθήκη έβγαλε σφάλμα και το annotation έγινε με ειδικό κώδικα λαμβάνοντας υπ’οψη το annotation της πλατφόρμας όπως αναφέρεται από το GEO. Το πρόβλημα “Many to Many” αντιμετωπίστηκε λαμβάνοντας υπ’ όψη τη μέγιστη τιμή σε περιπτώσεις όπου ένα Entrez ID αντιστοιχεί σε πολλά probes ενώ στην περίπτωση που ένα probe αντιστοιχεί σε πολλά Entrez IDs τότε έγινε πολλαπλασιασμός της συγκεκριμένης σειράς με τον αριθμό των αντιστοιχούντων Entrez IDs.

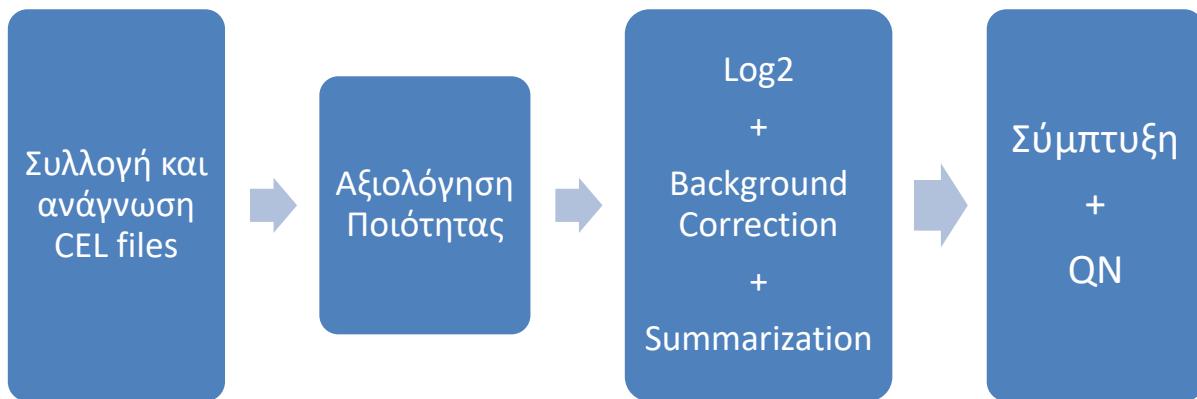
Ακολούθησε γραφική αναπαράσταση των αποτελεσμάτων στην πρώτη προσέγγιση ώστε να διασφαλιστεί η απαραίτητη προϋπόθεση της κανονικότητας για την εφαρμογή του z-score. Πιο συγκεκριμένα έγιναν boxplot, Ιστόγραμμα (σε ένα τυχαίο δείγμα π.χ. εδώ στο τρίτο) και Q-Q διάγραμμα (στο ίδιο δείγμα με αυτό στο Ιστόγραμμα) τα οποία υπάρχουν αναλυτικά όλα στο παράρτημα 5. Από την παραπάνω γραφική αναπαράσταση προέκυψε ότι το dataset GSE 74530 που αποτελείται από 10 δείγματα πληρούσε πολύ οριακά το κριτήριο της κατά

προσέγγιση κανονικότητας και για να διασφαλιστεί η εγκυρότητα των αποτελεσμάτων αφαιρέθηκε από την περαιτέρω ανάλυση στην πρώτη προσέγγιση με την ζ -κανονικοποίηση. Ωστόσο χρησιμοποιήθηκε στην διασπαστική προσέγγιση όπου η κανονικότητα δεν παίζει τόσο κρίσιμο ρόλο. Επίσης επιπλέον log2 μετασχηματισμοί χρησιμοποιήθηκαν όπου ήταν απαραίτητο για την κατά το δυνατόν προσέγγιση στην κανονικότητα (π.χ. GSE3524).

Μετά την παραπάνω επεξεργασία ανά dataset ακολούθησε σύμπτυξη των επιμέρους dataset σε ένα μεγάλο πίνακα έκφρασης (expression matrix). Σε σχέση με τα features, επιλέχθηκαν μόνο εκείνα που ήταν κοινά σε όλα τα dataset τα οποία ήταν τελικά 12.407 Entrez IDs. Τα ονόματα των στηλών περιλαμβάνουν το όνομα της στήλης όπως είναι στο τυποποιημένο expression matrix το οποίο ακολουθείται με το όνομα του dataset για λόγους αναγνωρισμότητας. Τελικά προέκυψαν 506 δείγματα ακανθοκυτταρικού καρκινώματος με 172 για την πρώτη προσέγγιση και για την δεύτερη (στην οποία συμπεριλήφθηκε ένα επιπλέον dataset) 510 και 177 αντίστοιχα. Οι Εικόνες 23 και 24 συνοψίζουν τα παραπάνω βήματα.



Εικόνα 23. Pipeline προεπεξεργασίας RAW δεδομένων με ζ -κανονικοποίηση.



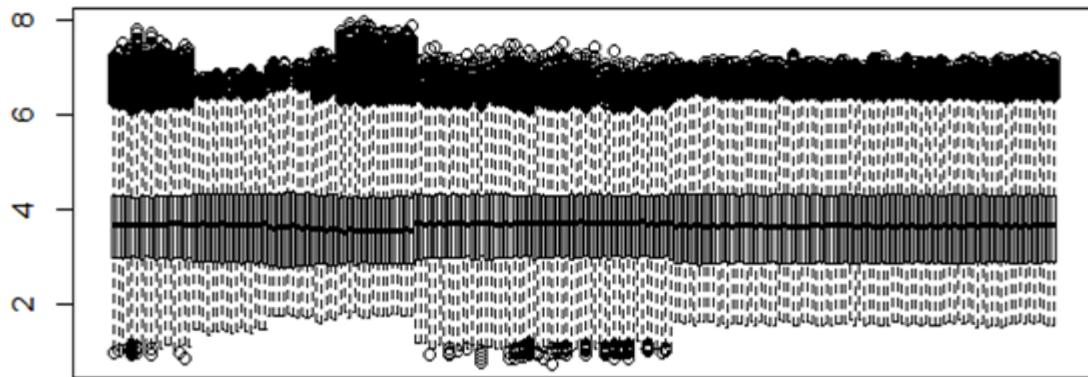
Εικόνα 24. Pipeline προεπεξεργασίας με την διασπαστική τεχνική

2.6 Στατιστική και Λειτουργική Ανάλυση

Για την στατιστική ανάλυση στην προσέγγιση όπου έγινε πλήρες RMA τα δεδομένα είναι ήδη κανονικοποιημένα και σε αυτά έγινε μία z-κανονικοποίηση για την εναρμόνιση τους, στις στήλες μόνο στην πρώτη προσέγγιση και σε γραμμές και σε στήλες στην δεύτερη προσέγγιση. Αυτό έγινε γιατί η επακόλουθη κανονικοποίηση και στις γραμμές συνετέλεσε σε αρκετή μείωση της ειδικότητας στην φάση της αξιολόγησης των αποτελεσμάτων (βλ. παρακάτω). Μετά από αυτή την z-κανονικοποίηση έγινε «ανύψωση» των δεδομένων για την αποφυγή αρνητικών ή/και μηδενικών τιμών που θα επηρέαζαν σημαντικά την διενέργεια χειρισμών για το feature selection όπως π.χ. στον υπολογισμό του logFC.

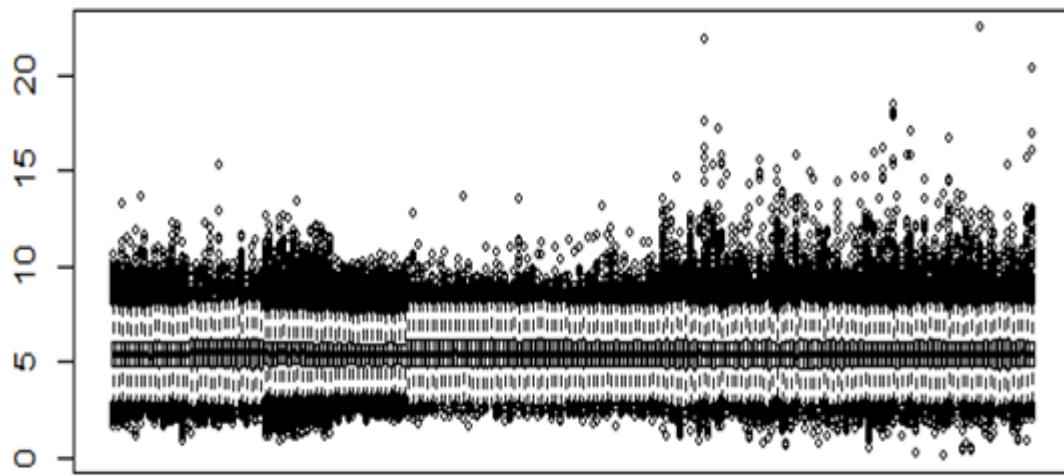
Στην δε δεύτερη περίπτωση της διασπαστικής τεχνική (όπου δεν έγινε το τελικό στάδιο QN του RMA) χρησιμοποιούνται τα δεδομένα όπως προέκυψαν από την επεξεργασία που περιγράφηκε στο προηγούμενο βήμα που είναι πλέον έτοιμα προς χρήση. Τα boxplot που προκύπτουν είναι για την απλή ζ-κανονικοποίηση, την διασπαστική τεχνική και την διπλή ζ-κανονικοποίηση οι εικόνες 25,26 και 27 αντίστοιχα.

Boxplot z-score samples 1-200

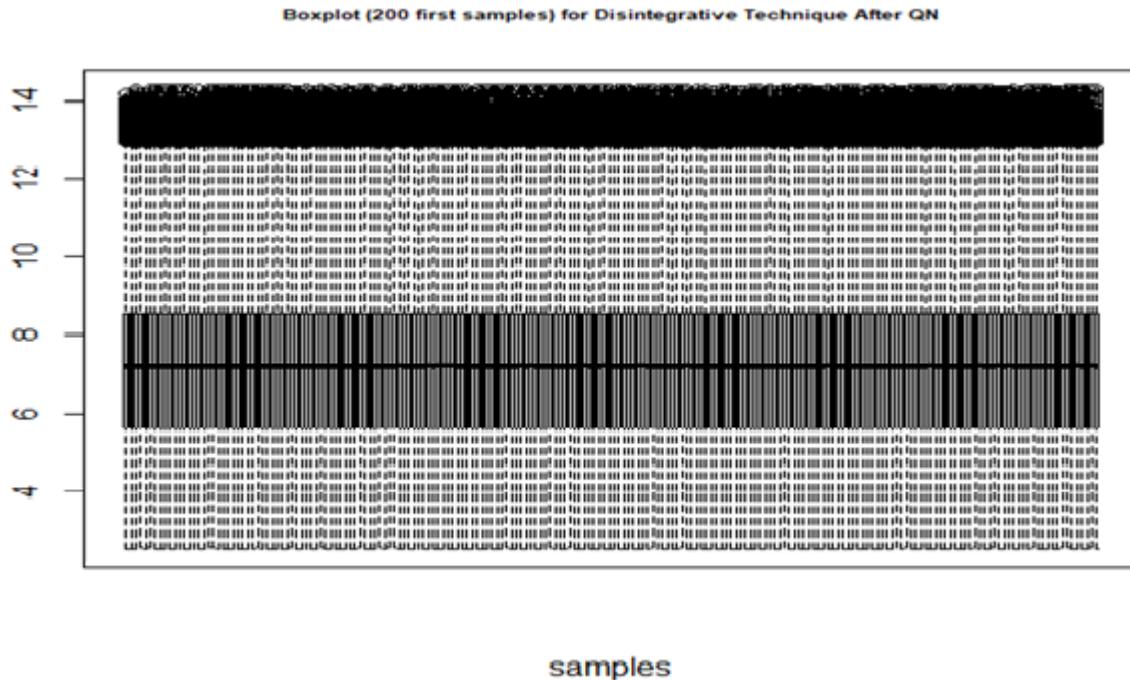


Εικόνα 25. Boxplot των πρώτων 200 samples μετά από ζ -κανονικοποίηση σε στίλες.

Boxplot double-z



Εικόνα 26. Boxplot 200 πρώτων δειγμάτων με την διπλή ζ -κανονικοποίηση



Εικόνα 27. Boxplot πρώτων 200 δειγμάτων με την διασπαστική τεχνική.

Αρχικά έγινε ένα μη ειδικό φιλτράρισμα ώστε να μην χρησιμοποιηθούν Entrez IDs με πολύ μικρή διακύμανση (Interquartile Range – IQR) (Bolstad, 2021). Τέθηκε ως όριο η τιμή $IQR = 0.5$ όπως αναφέρεται σε παρόμοιες προσεγγίσεις στην βιβλιογραφία (Beyer, Mallmann, Xue, Staratschek-Jox, Vorholt D et al, 2012). Αυτό το μη ειδικό φιλτράρισμα έχει ως στόχο να μειώσει την επίδραση του προβλήματος πολλαπλών συγκρίσεων, και θεωρείται καλή και ευρέως διαδεδομένη πρακτική, όπως περιγράφεται στο γενικό μέρος.

Στην συνέχεια έγινε αναζήτηση των ΔΕ γονιδίων με ένα απλό Student's t-test με έλεγχο πολλαπλών υποθέσεων και προσαρμογή του p-value με την τεχνική FDR και όριο FDR το 0.05. Η αναζήτηση ΔΕ γονιδίων επαναλήφθηκε με ίδιο έλεγχο πολλαπλών υποθέσεων και με moderated t-test, που είναι άλλωστε και η προσέγγιση που προτιμάται σε microarrays. Χρησιμοποιήθηκε για αυτόν τον σκοπό το πακέτο limma από το Bioconductor (Ritchie, Phipson, Wu, Hu, Law, Shi & Smyth, 2015). Το φιλτράρισμα έγινε με βάσει το p-value με όριο στατιστικής σημαντικότητας 5% και με την μέτρηση του logFC με όριο απόλυτης τιμής του το 1. Εδώ πρέπει να τονίσουμε ότι τα δεδομένα και στις 2 περιπτώσεις είναι ήδη λογαριθμημένα με την RMA και επομένως το logFC υπολογίστηκε με απλή αφαίρεση των μέσων όρων των φυσιολογικών και των καρκινικών δειγμάτων.

Ακολούθησε αναζήτηση ΔΕ γονιδίων με την μέθοδο SAM (Significance analysis of Microarrays) χρησιμοποιώντας το πακέτο samr από το Bioconductor (Tibshirani, Chu, Hastie, Narasimhan & Tibshirani, 2011), στην οποία όμως λόγω του ότι η συγκεκριμένη μέθοδος δίνει πολλά ψευδώς θετικά ευρήματα επιλέχθηκε p-value για το FDR το 1%. Όπως θα δούμε στα αποτελέσματα ακόμα και με τόσο αυστηρό όριο ελέγχου πολλαπλών υποθέσεων η μέθοδος SAM βγάζει μεγάλο αριθμό από ΔΕ γονίδια.

Στην συνέχεια έγιναν Venn Diagrams (Chen & Boutros, 2011) και για τις 3 αυτές προσεγγίσεις, volcano plots μόνο για τα t-test και moderated t-test (για την SAM δεν έγινε γιατί ο αριθμός ΔΕ γονιδίων που βγάζει είναι πάνω από το μισά γονίδια του συνόλου γονιδίων που εξετάστηκαν). Επίσης έγιναν θερμικοί χάρτες (Heatmaps) (Warnes, Bolker, Bonebakker, Gentleman & Huber, 2016) , με δενδρογράμματα και σε γραμμές και σε στήλες, σχεδιασμένα με τέτοιο τρόπο ώστε να φαίνονται στα δείγματα ποιά είναι καρκίνος και ποιά φυσιολογικά, γιατί η ιεραρχική ομαδοποίηση με τα δενδρογράμματα αλλάζει την σειρά των δειγμάτων με βάσει τα cluster που αυτά δημιουργούν. Ως μέθοδος ιεραρχικής ομαδοποίησης επιλέχθηκε η Ward.D2 γιατί θεωρείται αν και λίγο πιο περίπλοκη περισσότερο αξιόπιστη , αφού λαμβάνει υπ'οψη της και την στάθμιση κάθε δείγματος σε σχέση με τα υπόλοιπα (Murtagh & Legendre, 2014)

Στην συνέχεια έγινε γραφική παρατήρηση των θερμικών χαρτών, και ιδιαίτερα στο ύψος των ιεραρχικών ομαδοποιήσεων όπου υπάρχουν 2 clusters. Γενικότερα υπήρχε διαχωρισμός στο t-test και moderated t-test αλλά όχι στο SAM όμως περισσότερα για αυτό θα αναφερθούν στα αποτελέσματα της μελέτης. Σε όλες τις περιπτώσεις διαπιστώθηκε ότι υπήρχε στην μία ομάδα πλειοψηφία των καρκινικών δειγμάτων και στην άλλη πλειοψηφία των φυσιολογικών δειγμάτων. Λόγω του ότι η ιεραρχική ομαδοποίηση είναι κατά βάση μη επιβλεπόμενη μέθοδος, θεωρήθηκε ότι από τα 2 cluster που προκύπτουν, το ένα που είχε περισσότερα καρκινικά ως το cluster των καρκινικών δειγμάτων και το cluster με περισσότερα φυσιολογικά ως το cluster των φυσιολογικών δειγμάτων. Με αυτόν τον τρόπο έγινε δυνατόν να αξιολογηθεί η αποτελεσματικότητα τόσο των στατιστικών τεστ για την ανεύρεση ΔΕ γονιδίων, όσο και ο αρχικός χειρισμός των δεδομένων με τις 2 τεχνικές προσεγγίσεις που προαναφέραμε. Αν για παράδειγμα με κάποια τεχνική προσέγγιση προέκυπτε ανεπαρκής διαχωρισμός των 2 cluster αυτό θα σήμαινε ότι η προσέγγιση αυτή παραμόρφωσε τα δεδομένα σε τέτοιο βαθμό που να είναι εν τέλει άχρηστα για περαιτέρω ανάλυση. Στην

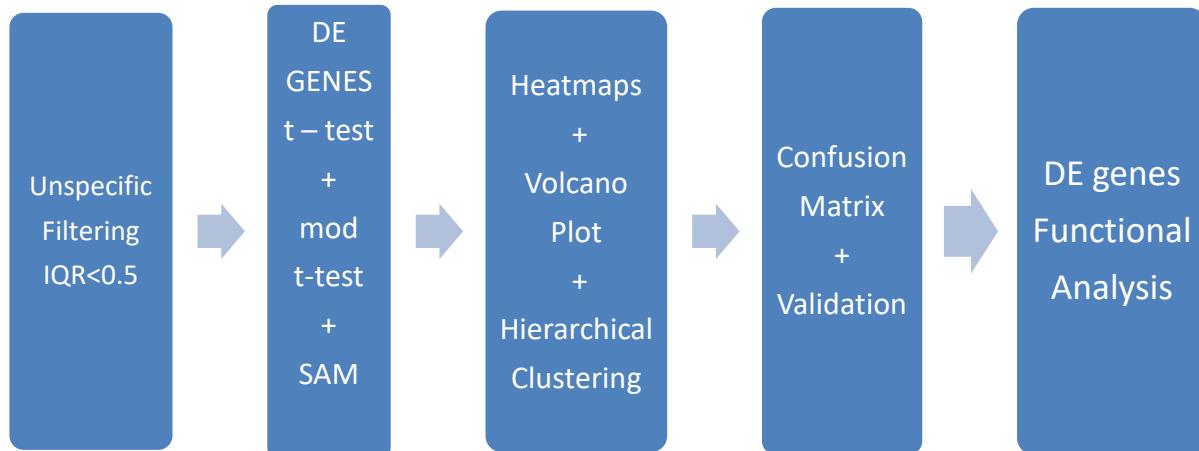
συνέχεια διενεργήθηκε με την πιο πάνω λογική, μέτρηση της ευαισθησίας και της ειδικότητας κάθε προσέγγισης αλλά και κάθε μεθόδου προσδιορισμού των ΔΕ γονιδίων.

Στην συνέχεια ακολούθησε λειτουργική ανάλυση ξεχωριστά των υπερεκφραζόμενων και των υποεκφραζόμενων ΔΕ γονιδίων με την βοήθεια του πακέτου gprofiler2 από το Bioconductor (Kolberg, Raudvere, Kuzmin, Vilo & Peterson, 2020) και αξιολογήθηκαν τα αποτελέσματα της. Ως όριο στατιστικής σημαντικότητας για την λειτουργική ανάλυση τέθηκε το 1%. Τα στάδια που ακολουθήθηκαν παραπάνω απεικονίζονται παραστατικά στην εικόνα 6.

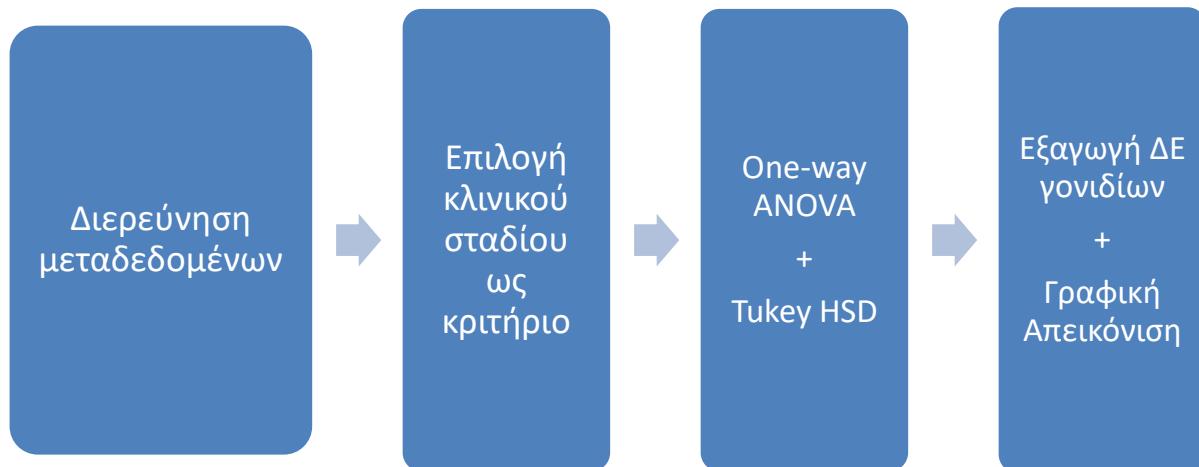
Τέλος αναζητήθηκαν στα μεταδεδομένα των 16 dataset τα πιο συχνά στοιχεία που αναφέρονται εκεί ώστε να μπορεί να γίνει περαιτέρω ανάλυση σε διάφορες κλάσεις που αφορούν το ΑΚΚ του στόματος. Από την μελέτη των μεταδεδομένων διαπιστώθηκε ότι ο πιο συχνά αναφερόμενος διαχωρισμός των καρκινωμάτων ήταν με βάσει την κλινική σταδιοποίηση τους που αναφερόταν σε 6 datasets. Παρόλα αυτά στο μεγαλύτερο από αυτά δεν αναφέρονταν ξεκάθαρα οι κατηγορίες σταδιοποίησης αλλά αναφερόταν ως στάδιο I/II και στάδιο III/IV. Τελικά αποφασίστηκε η ανάλυση κλάσεων να γίνει με 2 κατηγορίες, μία πρώιμου σταδίου (Στάδια I/II) και μία προχωρημένου σταδίου (Στάδια III/IV).

Λήφθηκαν οι ήδη επεξεργασμένοι με τους 3 διαφορετικούς τρόπους που αναφέραμε από το βήμα 5, πίνακες έκφρασης για τα 6 dataset. Αφού έγινε σύμπτυξη τους έγινε γραφικός έλεγχος τους, όπου διαπιστώθηκε ότι ένα δείγμα ήταν outlier και αφαιρέθηκε. Επειδή τα δείγματα είχαν σχετικά λίγα φυσιολογικά δείγματα προστέθηκαν φυσιολογικά δείγματα από ένα από τα dataset που είχε 45 φυσιολογικά δείγματα (GSE30784). Αυτό έγινε για λόγους ισορροπίας (balance) του αριθμού των δειγμάτων ανά κλάση και αύξηση της στατιστικής εγκυρότητας με την μέθοδο ANOVA (Analysis of Variance). Έτσι προέκυψαν 105 δείγματα προχωρημένου σταδίου, 65 πρώιμου και 66 φυσιολογικά δείγματα.

Ακολούθησε δοκιμασία ANOVA και Tukey's HSD (Honestly Significant Difference). Τα αποτελέσματα απεικονίστηκαν με θερμικό χάρτη και αναζητήθηκαν τα ΔΕ γονίδια με τα ίδια κριτήρια feature selection που τέθηκαν, και στις δοκιμασίες t-test και moderated t-test ($|logFC| > 1$, adj p-value<0.05). Οι διαφορές σε κάθε συνδυασμό των 3 κλάσεων, απεικονίστηκαν σε διαγράμματα Venn. Τα στάδια που ακολουθήθηκαν στην ανάλυση κλάσεων απεικονίζονται παραστατικά στις Εικόνες 28 και 29.



Εικόνα 28. Pipeline στατιστικής και λειτουργικής ανάλυσης.



Εικόνα 29. Pipeline Στατιστικής Ανάλυσης Κλάσεων με βάσει τα υπάρχοντα μεταδεδομένα.

2.7 Εφαρμογή Προχωρημένων Τεχνικών Ανάλυσης

Μετά το βήμα που περιγράφηκε στην προηγούμενη ενότητα έγινε αξιολόγηση των 3 προσεγγίσεων τόσο σε επίπεδο ευαισθησίας και ειδικότητας όσο και στα αποτελέσματα της λειτουργικής ανάλυσης. Διαπιστώθηκε ότι η διπλή ζ-κανονικοποίηση (κατά γονίδια και

δείγματα) αποδίδει λιγότερο καλά σε σχέση με την κανονικοποίηση μόνο ανά στήλες και την διασπαστική τεχνική. Επιπρόσθετα παρατηρήθηκε κατά την λειτουργική ανάλυση η εύστοχη και σαφώς καλύτερη περιγραφή σε επίπεδο λειτουργικών κατηγοριών από την διασπαστική τεχνική. Υπό το πρίσμα των παραπάνω αποφασίστηκε η συνέχιση της ανάλυσης των δεδομένων με μεθόδους μηχανικής μάθησης καθώς και σε επίπεδο ανάλυσης δικτύων με τα αποτελέσματα της διασπαστικής τεχνικής.

Αρχικά έγινε PCA (Principal Components Analysis) (Wold, Espensen & Geladi, 1987) ως ένα πρώτο βήμα εξερεύνησης των δεδομένων και στην συνέχεια ακολούθησε επιβλεπόμενη ανάλυση με τυχαιοποιημένα δάση (Random Forest) (Liaw & Wiener, 2002). Ακολούθησε η επιλογή των 30 πιο σημαντικών από πλευράς ταξινομικής απόδοσης γονιδίων και αυτό γιατί αυτό είναι το όριο που θέτουν κάποια εγχειρίδια βιοπληροφορικής ως ένα μέγιστο όριο εύρεσης βιοδεικτών (Dziuda, 2010) και η τελική επεξεργασία τους έγινε με 2 αλγορίθμους δένδρων απόφασης (Decision Trees). Τα δένδρα απόφασης από μόνα τους δεν βγάζουν τόσο αξιόπιστα αποτελέσματα καθώς αποτελούν παραδείγματα άπληστων αλγορίθμων. Όμως έχουν το πλεονέκτημα της εύκολης χρήσης και ερμηνείας τους στην κλινική πράξη που τα κάνει κατανοητά ακόμα και από τον μη-ειδικό στην βιοπληροφορική επιστήμονα.

Τα ολικά δείγματα διαχωρίστηκαν με αναλογία 0.8:0.2, σε δείγματα εκπαίδευσης (training data) και δείγματα δοκιμής (testing data). Το tuning στον αλγόριθμο τυχαιοποιημένου δάσους έγινε με την εντολή “tuneRF” που είναι για αυτό το σκοπό στο πακέτο του CRAN, RandomForest. Εν συνεχείᾳ χρησιμοποιήθηκαν οι αλγόριθμοι CART (Therneau, Atkinson, Ripley, ... & Ripley, 2015) με την βοήθεια του πακέτου “caret” (Kuhn, 2009) και Conditional inference tree (C-tree) με την βοήθεια του πακέτου “partykit” (Hothorn & Zeileis, 2015) και εύρεση της υπερπαραμέτρου με το “caret”. Ο αλγόριθμος CART παίρνει ως υπερπαράμετρο το “complexity parameter” ή cp, ενώ ο C-tree έχει επίσης μόνο μία υπερπαράμετρο, το mincriterion που αναπαριστά το βαθμό σημαντικότητας του δένδρου. Αυτοί οι παράμετροι τέθηκαν με κατάλληλες εντολές στο “caret” αλλά και εξερευνήθηκαν οι πιθανές τιμές τους με αλλεπάλληλες δοκιμές τύπου trial and error. Και στους 2 αλγόριθμους έγινε cross-validation μέσω control handling με bootstrap προσέγγιση. Αφού βρέθηκαν οι τελικοί υπερπαράμετροι, εφαρμόστηκαν τα training μοντέλα, και η ακολούθησε η απεικόνιση τους ως decision tree plots. Τέλος έγινε έλεγχος του μοντέλου τόσο με πίνακες

σύγχυσης (confusion matrices) όσο και με καμπύλες ROC (Robin, Turck, Hainard, Tiberti, Lisacek, Sanchez & Müller, 2011).

Ως τελευταίο βήμα έγινε ανάλυση δικτύων όσον αφορά τα ΔΕ γονίδια. Σε πρώτη φάση έγινε απλή χαρτογράφηση σε γράφους της βάσης δεδομένων Kegg. Για αυτό τον σκοπό χρησιμοποιήθηκε η βιβλιοθήκη pathview (Luo, Weijun, Brouwer & Cory, 2013) από το Bioconductor χρησιμοποιώντας τα πιο αντιπροσωπευτικά pathways όπως καθορίστηκαν από την λειτουργική ανάλυση. Στην συνέχεια χρησιμοποιήθηκε η γνωστή εφαρμογή ανάλυσης δικτύων Cytoscape (v. 3.8.2) (Lopes, Franz, Kazi, Donaldson, Morris & Bader, 2010) και μέσω χρήσης της εφαρμογής stringApp (Doncheva, Morris, Gorodkin & Jensen, 2018) αναλύθηκαν τα υπερεκφρασμένα γονίδια με την εφαρμογή CytoHubba (Chin, Chen, Wu, Ho, Ko & Lin, 2014) ως προς βασικά χαρακτηριστικά τοπολογίας δικτύων όπως ο βαθμός κόμβου και η ενδιάμεση εκκεντρότητα των κόμβων (betweenness centrality). Οι κόμβοι στο δίκτυο Cytoscape, χρωματίστηκαν από ανοικτό κίτρινο προς σκούρο ανάλογα με την τιμή logFC που υπολογίστηκε κατά την στατιστική ανάλυση των αποτελεσμάτων. Με άλλα λόγια όσο πιο σκούρος είναι ένας κόμβος τόσο περισσότερο ΔΕ είναι το γονίδιο που αναπαριστά. Τέλος έγινε και ένας γράφος Cytoscape με βάση τα αποτελέσματα της αναζήτησης βιβλιογραφίας στο Pubmed website, όσον αφορά τα 30 πιο συχνά αναφερόμενα γονίδια ώστε να βοηθήσει στην αντιπαραβολή των αποτελεσμάτων με την βιβλιογραφία. Οι ακμές σε όλους τους γράφους Cytoscape απεικονίζουν λειτουργικές σχέσεις πρωτεΐνης-πρωτεΐνης (pp). Τα παραπάνω βήματα συνοψίζονται στην Εικόνα 30.



Εικόνα 30. Pipeline προχωρημένης επεξεργασίας των δεδομένων

Κεφάλαιο 3. Αποτελέσματα

3.1 Στατιστική Ανάλυση για εύρεση ΔΕ γονιδίων

Από την αναζήτηση ΔΕ γονιδίων και βάσει των κριτηρίων p-value και logFC που αναφέρθηκαν στο προηγούμενο κεφάλαιο διαπιστώθηκε ότι το Student's t-test και το moderated t-test βρίσκουν ακριβώς τα ίδια ΔΕ γονίδια και στις 3 προσεγγίσεις μετα-ανάλυσης. Επίσης και στις 3 προσεγγίσεις τα ΔΕ γονίδια που βρέθηκαν με την μέθοδο SAM ήταν πολλά περισσότερα σε σχέση με τις άλλες 2 στατιστικές μεθόδους.

Στην μεν προσέγγιση με την ζ-κανονικοποίηση μόνο στις στήλες, αφού έγινε μη ειδικό φιλτράρισμα έμειναν 4.172 Entrez IDs. Με τα moderated και Student's t-test βρέθηκαν 109 ΔΕ γονίδια από τα οποία 45 ήταν υπερεκφραζόμενα και τα 64 ήταν υποεκφραζόμενα. Με την μέθοδο SAM βρέθηκαν συνολικά 2.515 ΔΕ γονίδια από τα οποία τα υπερεκφραζόμενα ήταν 1.229 και τα υποεκφραζόμενα 1.286.

Στην δε προσέγγιση με την διπλή κανονικοποίηση αφού έγινε μη ειδικό φιλτράρισμα έμειναν 11.114 Entrez IDs. Με τα moderated και Student's t-test βρέθηκαν 1.152 ΔΕ γονίδια από τα οποία 525 ήταν υπερεκφραζόμενα και 627 ήταν υποεκφραζόμενα. Με την μέθοδο SAM βρέθηκαν συνολικά 7986 ΔΕ γονίδια από τα οποία τα υπερεκφραζόμενα ήταν 3789 και τα υποεκφραζόμενα 4197.

Τέλος με την διασπαστική τεχνική αφού έγινε μη ειδικό φιλτράρισμα έμειναν 12.400 Entrez IDs. Με τα moderated και Student's t-test βρέθηκαν 855 ΔΕ γονίδια από τα οποία 385 ήταν υπερεκφραζόμενα και 470 ήταν υποεκφραζόμενα. Με την μέθοδο SAM βρέθηκαν συνολικά 6881 ΔΕ γονίδια από τα οποία τα υπερεκφραζόμενα ήταν 3.252 και τα υποεκφραζόμενα 3.629.

Θα ακολουθήσει σε επόμενη ειδικά αφιερωμένη για αυτόν τον σκοπό ενότητα η γραφική παρουσίαση αυτών των αποτελεσμάτων.

3.2 Αξιολόγηση μεθόδων βάσει ικανότητας διαχωρισμού δειγμάτων

Από την ανάλυση ευαισθησίας /ειδικότητας και με βάση την ιεραρχική ομαδοποίηση των δειγμάτων με την μέθοδο Ward.D2 που προέκυψαν για όλες τις μεθόδους εύρεσης ΔΕ γονιδίων που χρησιμοποιήθηκαν, διαπιστώθηκε μία σαφώς καλύτερη απόδοση με τις μεθόδους t-test και moderated t-test. Να σημειωθεί ότι τα αποτελέσματα των παραπάνω 2 μεθόδων ήταν πανομοιότυπα και στους 3 τρόπους μετα-ανάλυσης. Η μέθοδος SAM απέδωσε πολύ χειρότερα από τις t-test και moderated t-test και για τους 3 τρόπους μετα-ανάλυσης, και αυτή η χειρότερη απόδοση αφορά τόσο την ευαισθησία όσο και την ειδικότητα.

Τώρα όσον αφορά τους τρόπους μετα-αναλυτικής προσέγγισης που χρησιμοποιήθηκαν για την κατασκευή του τελικού πίνακα έκφρασης, διαπιστώθηκε ότι η μέθοδος ζ-κανονικοποίησης που αφορά μόνο τις στήλες και η διασπαστική μέθοδος αποδίδουν πολύ καλά με αποτελέσματα ευαισθησίας/ειδικότητας πάνω από 90% αμφότερα. Παρατηρείται βέβαια μία μικρότερη μείωση στην απόδοση της διασπαστικής μεθόδου της τάξεως του 1.5% σε σχέση με την ζ-κανονικοποίηση στηλών, όμως η διαφορά σίγουρα δεν είναι αξιοσημείωτη. Η μέθοδος της διπλής κανονικοποίησης, είχε την καλύτερη ευαισθησία από τις άλλες 2 μεθόδους με διαφορά 4-5% αλλά όσον αφορά την ειδικότητα η επιτυχία της ήταν μόνο 59.54% και είναι σαφώς κατώτερη από τα ποσοστά ειδικότητας που καταγράφηκαν για τις άλλες 2 μεθόδους.

Πιο αναλυτικά και με βάση την ιεραρχική ομαδοποίηση των δειγμάτων προέκυψαν για όλες τις μεθόδους που χρησιμοποιήθηκαν παρουσιάζονται τα παρακάτω αποτελέσματα αξιολόγησης.

Για την μέθοδο **ζ-κανονικοποίησης μόνο ανα στήλες** με τα 2 είδη t-test είχαμε ευαισθησία 94,86% και ειδικότητα 98.27% καθώς και τον πίνακα σύγχυσης όπως φαίνεται στον Πίνακα 4. Για την μέθοδο SAM οι τιμές ευαισθησίας και ειδικότητας ήταν 76.48% και 29.48% αντίστοιχα και ο πίνακας σύγχυσης φαίνεται στο Πίνακα 5.

Πρόβλεψη/Πραγματικότητα	AKK	Υγιής
AKK	480	3
Υγιής	26	170

Πίνακας 4. Πίνακας Σύγχυσης με z-score μόνο ανα στήλη και t-tests.

Πρόβλεψη/Πραγματικότητα	AKK	Υγιής
AKK	387	122
Υγιής	119	51

Πίνακας 5. Πίνακας Σύγχυσης με z-score μόνο ανά στήλη και SAM.

Για την μέθοδο της **διπλής ζ-κανονικοποίησης** με τα 2 είδη t-test είχαμε ευαισθησία 98.02% και ειδικότητα 59.54% καθώς και τον πίνακα σύγχυσης όπως φαίνεται στον Πίνακα 6. Ενώ αντίστοιχα για το SAM 67.79% και 34.1% και ο αντίστοιχος πίνακας Σύγχυσης φαίνεται στον Πίνακα 7.

Πρόβλεψη/Πραγματικότητα	AKK	Υγιής
AKK	496	70
Υγιής	10	103

Πίνακας 6. Πίνακας Σύγχυσης με την μέθοδο διπλής z-κανονικοποίησης με t-tests.

Πρόβλεψη/Πραγματικότητα	AKK	Υγιής
AKK	343	114
Υγιής	163	59

Πίνακας 7. Πίνακας Σύγχυσης με την μέθοδο διπλής z-κανονικοποίησης και SAM

Για την **διασπαστική μέθοδο** με τα 2 είδη t-test είχαμε ευαισθησία 93.53% και ειδικότητα 97.18% καθώς και τον παρακάτω πίνακα σύγχυσης όπως φαίνεται στον Πίνακα 8. Ενώ για την μέθοδο SAM 76.48% και 29.48% αντίστοιχα ενώ ο πίνακας σύγχυσης φαίνεται στον Πίνακα 9.

Πρόβλεψη/Πραγματικότητα	AKK	Υγιής
AKK	477	5
Υγιής	33	172

Πίνακας 8. Πίνακας Σύγχυσης με διασπαστική μέθοδο και t-tests.

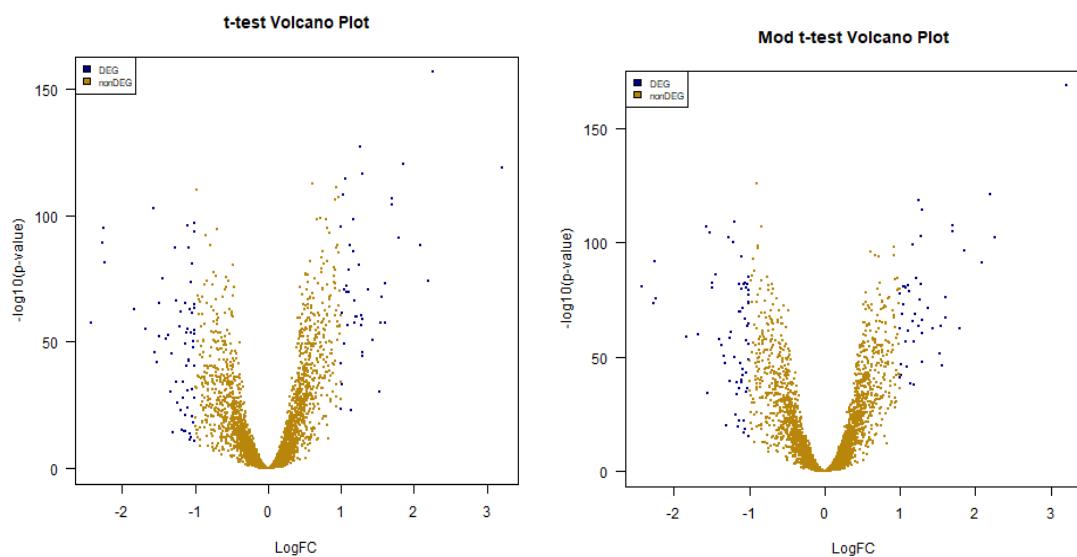
Πρόβλεψη/Πραγματικότητα	AKK	Υγιής
AKK	394	126
Υγιής	116	51

Πίνακας 9. Πίνακας Σύγχυσης με διασπαστική μέθοδο και SAM.

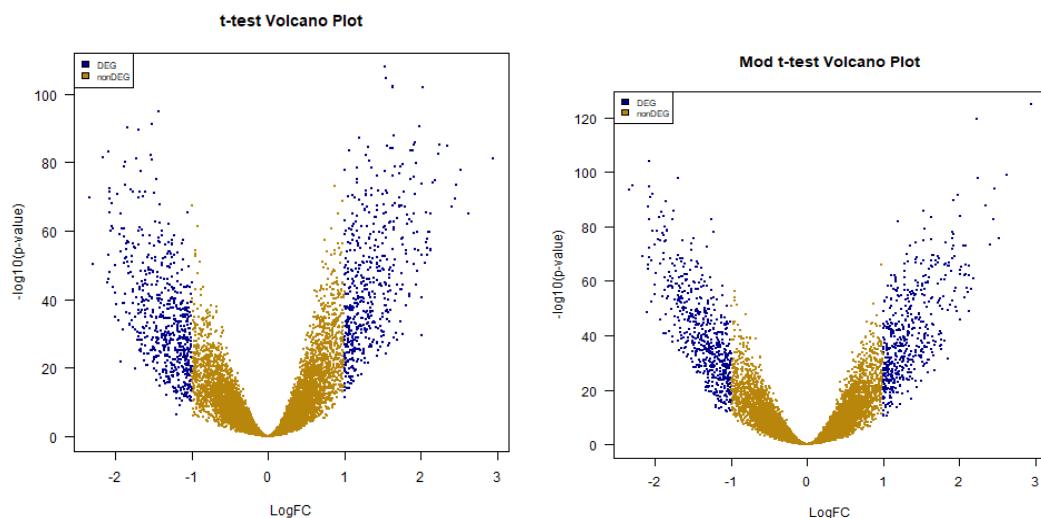
3.3 Γραφικές Απεικονίσεις στατιστικής ανάλυσης ΔΕ γονιδίων

3.3.1 Volcano Plots

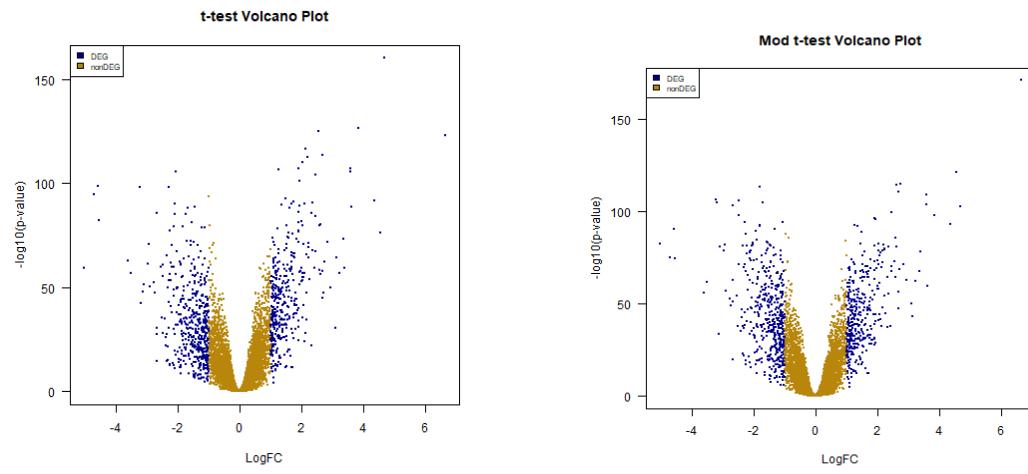
Τα ΔΕ γονίδια εδώ έχουν απεικονιστεί με μπλε χρώμα. Παρατηρούμε και στις 3 περιπτώσεις ότι οι κατανομές των ΔΕ ανάμεσα στο t-test και στο moderated t-test είναι ελάχιστες. Να σημειωθεί ότι απεικονίζονται στα volcano plots μόνο τα γονίδια που έχουν περάσει το αρχικό φιλτράρισμα με $IQR > 0.5$.



Εικόνα 31. z-score τεχνική μόνο σε στήλες μεταξύ του t-test και του moderated t-test.



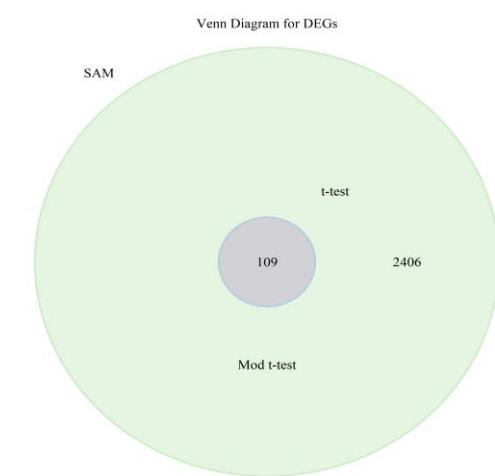
Εικόνα 32. Volcano plots με την τεχνική του διπλού z-score για τα moderated-t και t-test



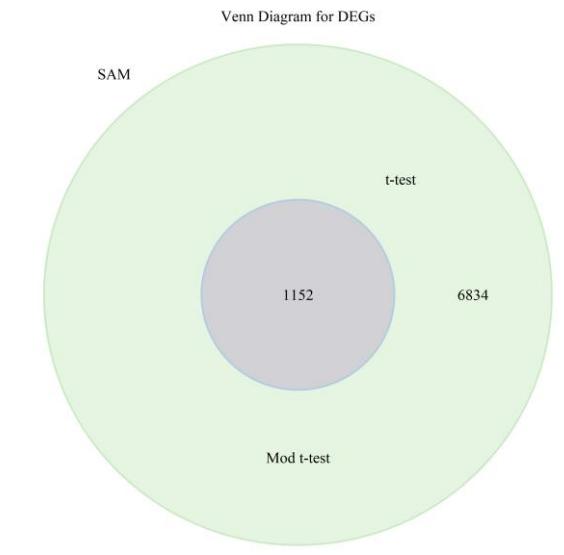
Εικόνα 33. Volcano plot από την διασπαστική τεχνική μεταξύ του t-test και του moderated t-test.

3.3.2 Διαγράμματα Venn

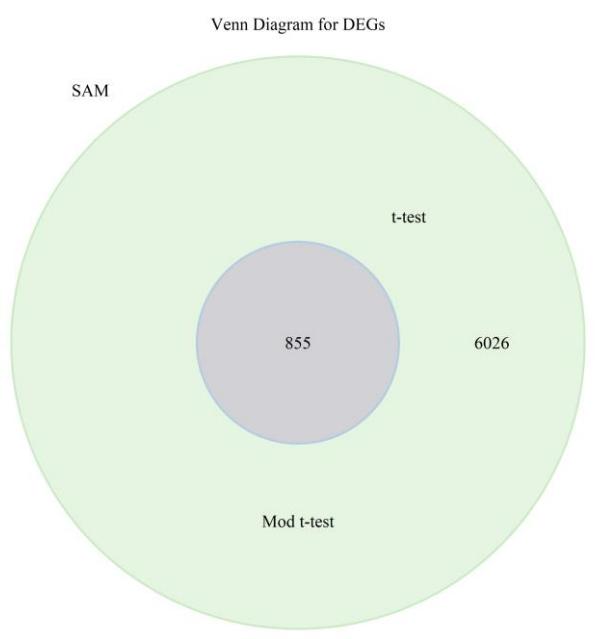
Τα διαγράμματα Venn και στις 3 προσεγγίσεις (Εικόνες 34,35 και 36) έχουν το ίδιο μοτίβο το οποίο μοιάζει με «ηλιακό σύστημα», με τα αποτελέσματα από το t-test και το moderate t-test να είναι ο Ήλιος, δηλαδή στο κέντρο και τα αποτελέσματα από το SAM να είναι το ηλιακό σύστημα. Παρατηρείται ότι ο Ήλιος έχει διπλό περίγραμμα και αυτό γιατί τα ΔΕ γονίδια από το t-test και το moderate t-test ταυτίζονται απόλυτα.



Εικόνα 34. Διάγραμμα Venn ΔΕ γονιδίων στην μόνο κατά στήλες z-score τεχνική .



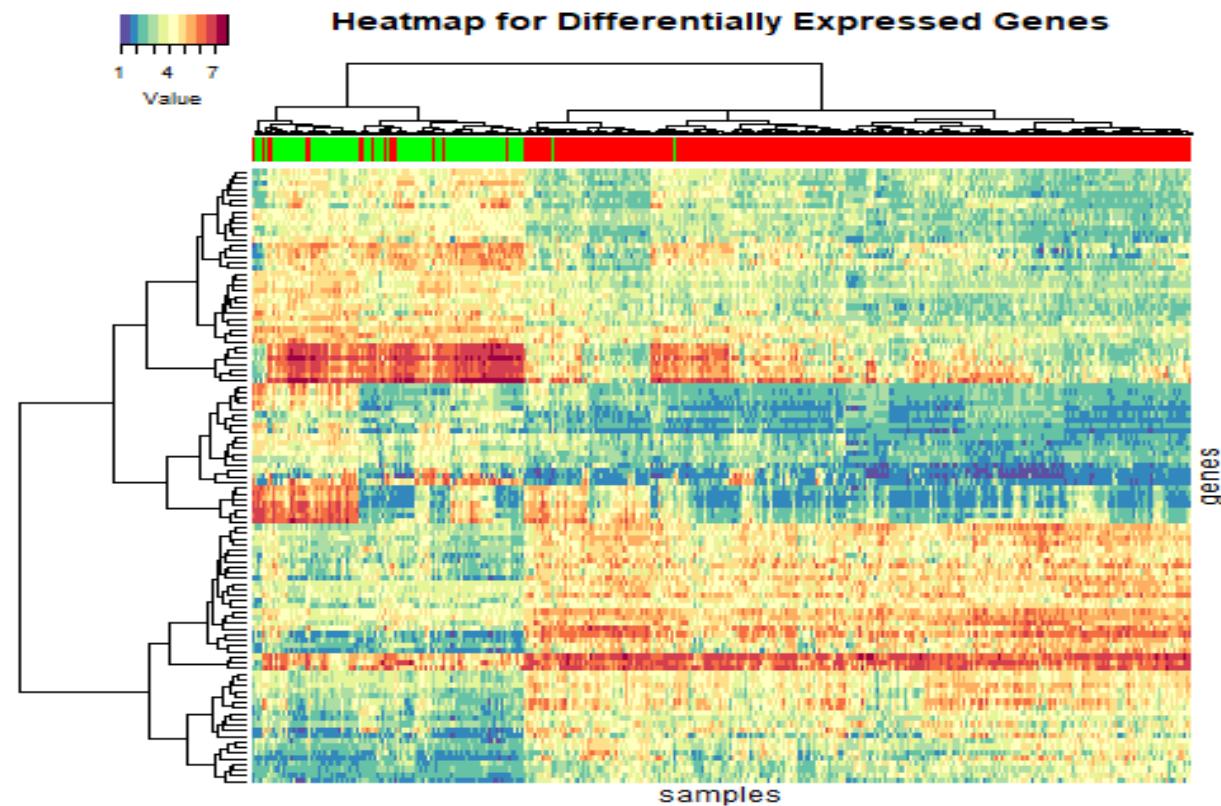
Εικόνα 35. Venn διάγραμμα ΔΕ γονιδίων με την τεχνική της διπλής ζ-κανονικοποίησης.



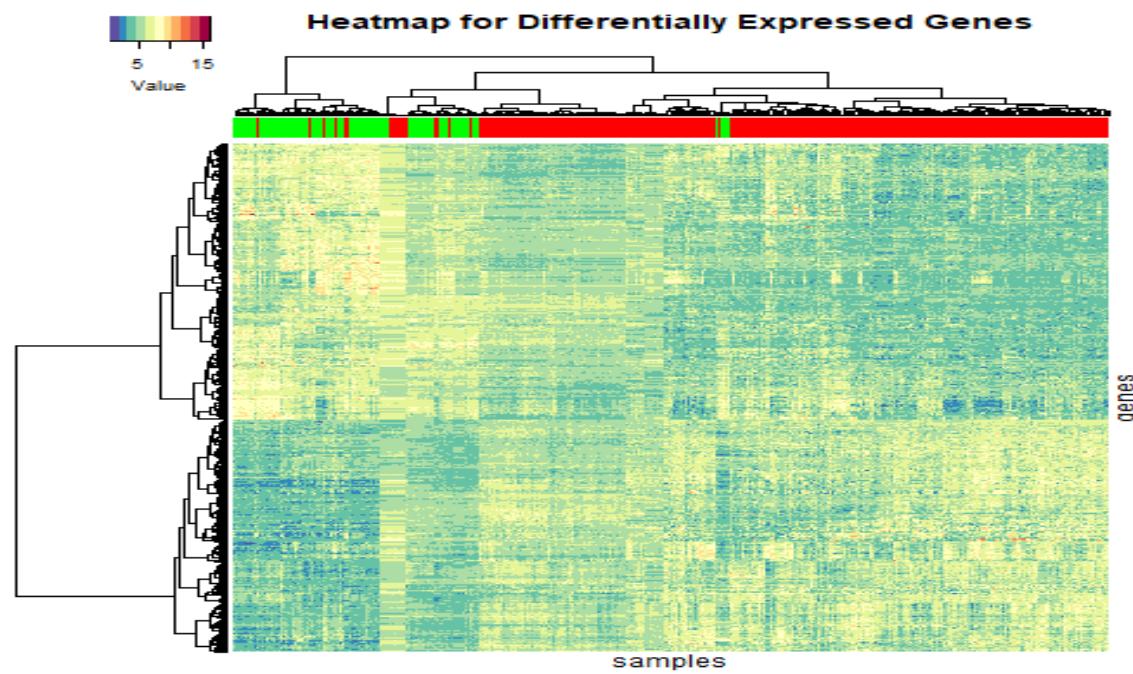
Εικόνα 36. Διάγραμμα Venn ΔΕ γονιδίων στην διασπαστική τεχνική.

3.3.3 Θερμικοί Χάρτες (Heatmaps)

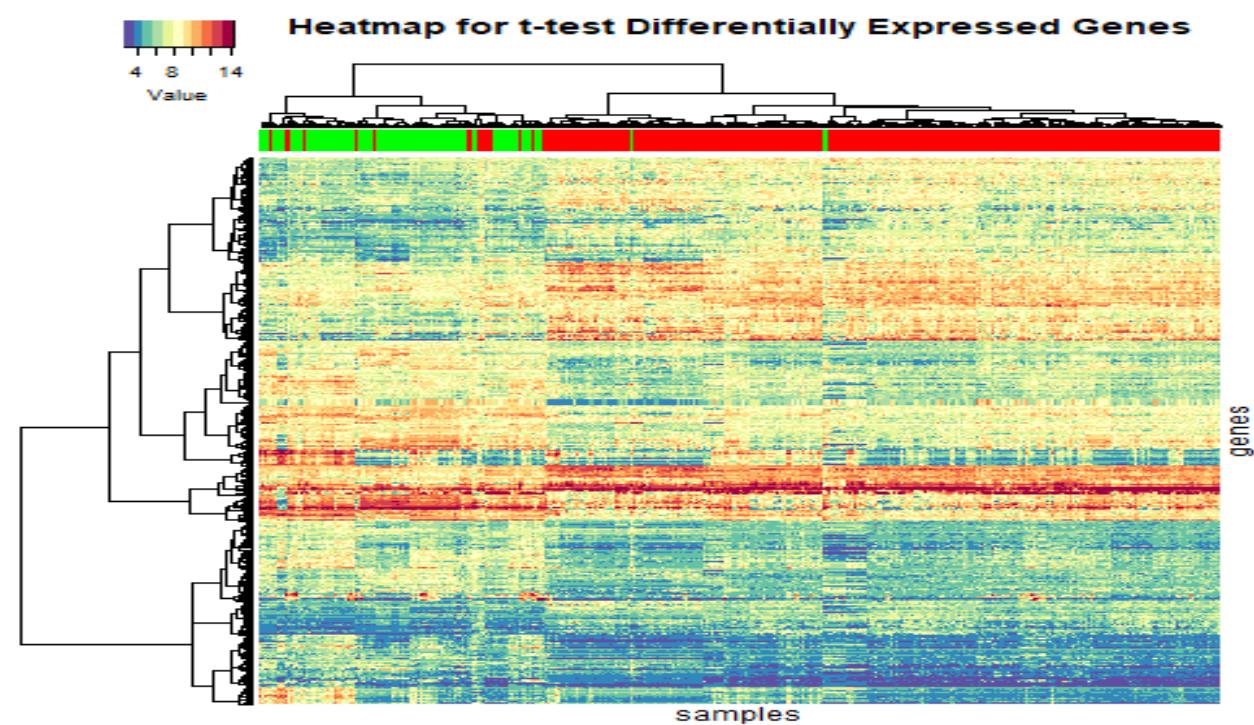
Οι θερμικοί χάρτες (Εικόνες 37-42) είναι τα σημαντικότερα διαγράμματα όσον αφορά την αξιολόγηση των 3 προσεγγίσεων αλλά και την αξιολόγηση των μεθόδων ανεύρεσης των ΔΕ γονιδίων. Το δένδρο ιεραρχικής ομαδοποίησης στον οριζόντιο άξονα κάτω από τα δείγματα έχει μία χρωματική μπάρα. Σε αυτήν τα δείγματα ΑΚΚ χρωματίζονται με κόκκινο ενώ τα δείγματα φυσιολογικού βλεννογόνου χρωματίζονται με πράσινο χρώμα. Είναι προφανές ότι η SAM διαχωρίζει τις 2 παραπάνω κλάσεις σε μη ικανοποιητικό βαθμό. Ένα άλλο σημείο που παρατηρείται είναι ότι οι χρωματικές αντιθέσεις στον χάρτη είναι πολύ πιο έντονες στην διασπαστική τεχνική και την ζ-κανονικοποίηση ανά στήλες μόνο, ενώ στην περίπτωση της διπλής ζ-κανονικοποίησης οι αντιθέσεις αμβλύνονται πιθανώς λόγω παραμόρφωσης και «ισοπέδωσης» των δεδομένων. Είναι ενδιαφέρον να παρατηρήσουμε ότι η ίδια άμβλυνση των αντιθέσεων υπάρχει στους θερμικούς χάρτες με ΔΕ γονίδια που εξάχθηκαν με την SAM τεχνική.



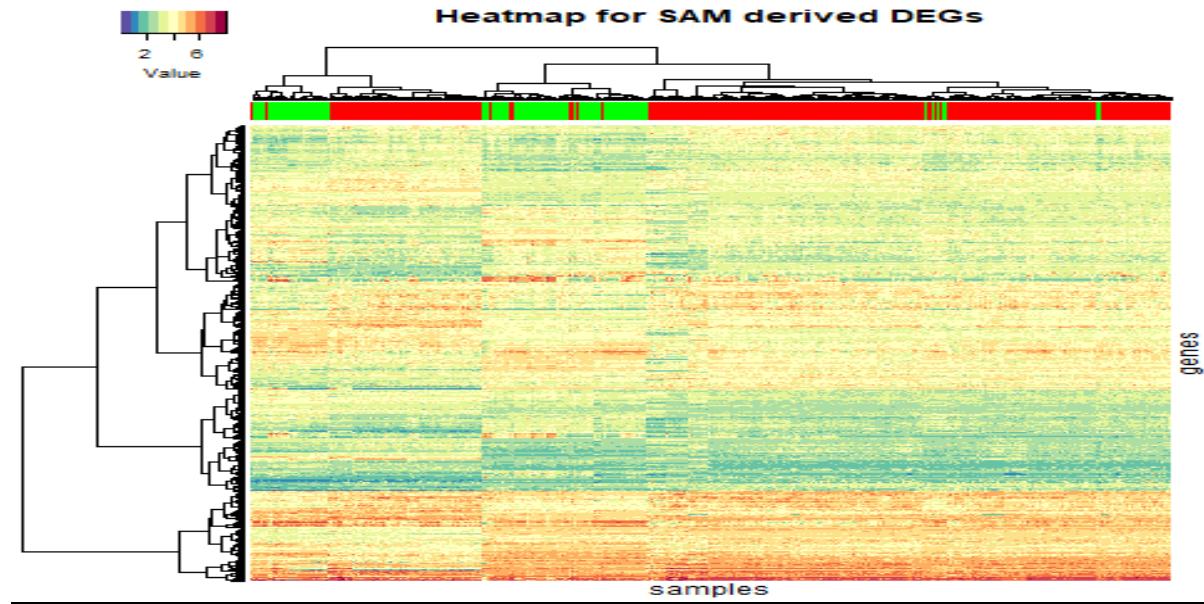
Εικόνα 37. Heatmap των ΔΕ γονιδίων στην τεχνική z-score μόνο κατά στήλες και t-tests



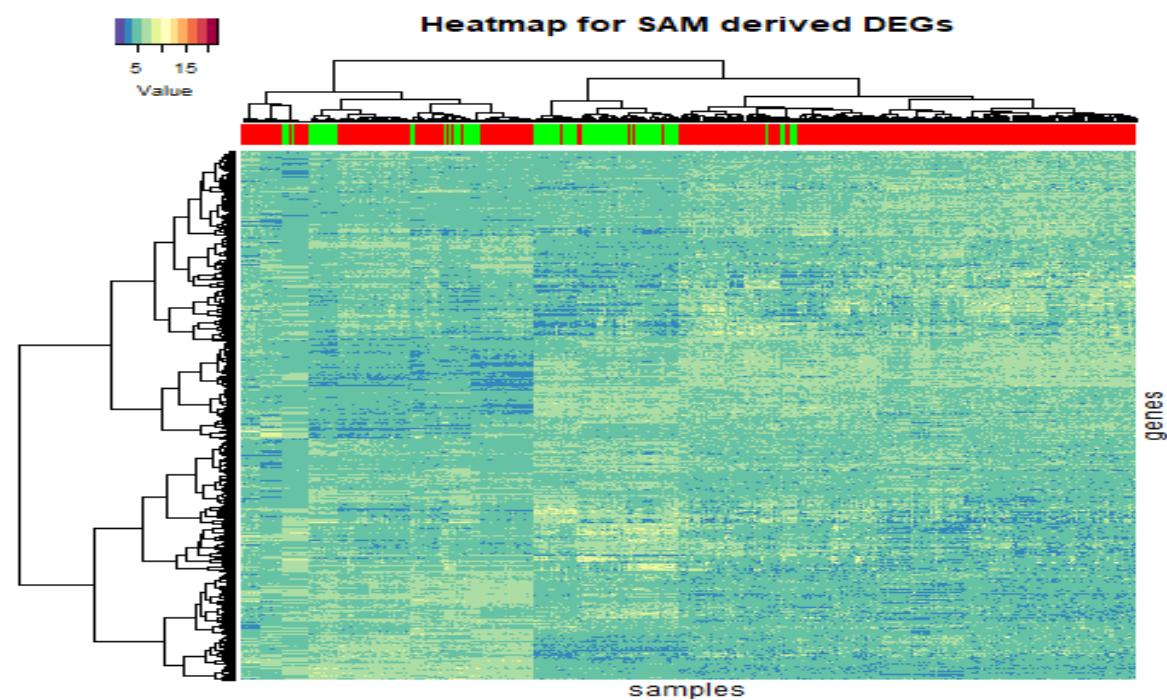
Εικόνα 38. Heatmap των ΔΕ γονιδίων με διπλή ζ-κανονικοποίηση και t-tests



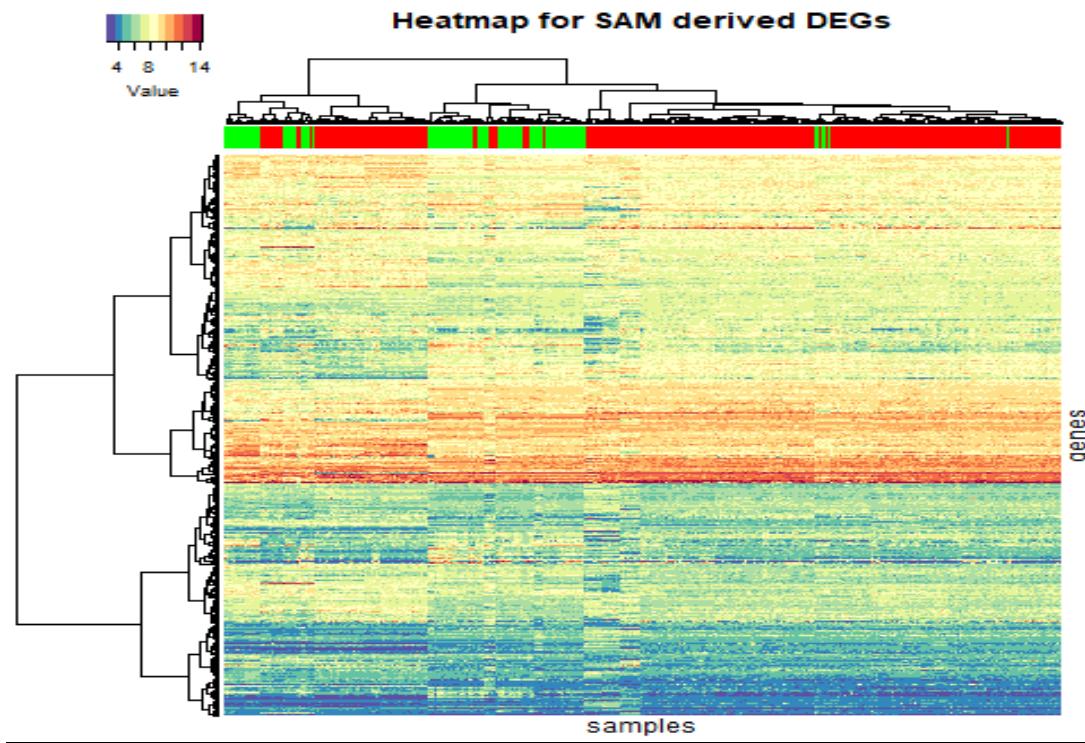
Εικόνα 39. Heatmap των ΔΕ γονιδίων με διασπαστική τεχνική και t-tests.



Εικόνα 40. Heatmap ΔΕ γονιδίων με z-score τεχνική μόνο κατά στήλες και SAM.



Εικόνα 41. Heatmap ΔΕ γονιδίων με διπλή ζ-κανονικοποίηση των ΔΕ γονιδίων και SAM



Εικόνα 42. Heatmap ΔE γονιδίων με διασπαστική τεχνική και SAM.

3.4 Αποτελέσματα ανάλυσης κλάσεων με βάση τα μεταδεδομένα σταδιοποίησης

Από την ανάλυση κλάσεων με ANOVA γενικά δεν διαπιστώθηκε κάποιος διαχωρισμός των πρώιμων και των προχωρημένων κλινικών σταδίων του ΑΚΚ σε επίπεδο γονιδιακής έκφρασης με καμία από τις 3 προσεγγίσεις μετα-ανάλυσης.

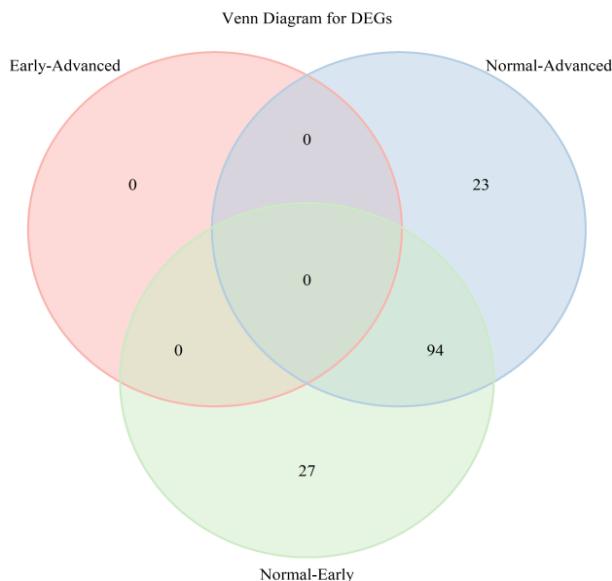
Με την ζ -κανονικοποίηση ανά στήλη μόνο, είχαμε 60 υπερεκφραζόμενα και 57 υποεκφραζόμενα γονίδια ανάμεσα στο προχωρημένο στάδιο σε σχέση με τους φυσιολογικούς ιστούς. Ανάμεσα στο πρώιμο στάδιο και στους φυσιολογικούς είχαμε 67 υπερεκφραζόμενα και 54 υποεκφραζόμενα γονίδια. Δεν υπήρξε κανένα ΔE γονίδιο ανάμεσα στους καρκίνους προχωρημένου και πρώιμου σταδίου.

Με την διπλή ζ -κανονικοποίηση, είχαμε 860 υπερεκφραζόμενα και 830 υποεκφραζόμενα γονίδια ανάμεσα στο προχωρημένο στάδιο σε σχέση με τους φυσιολογικούς ιστούς.

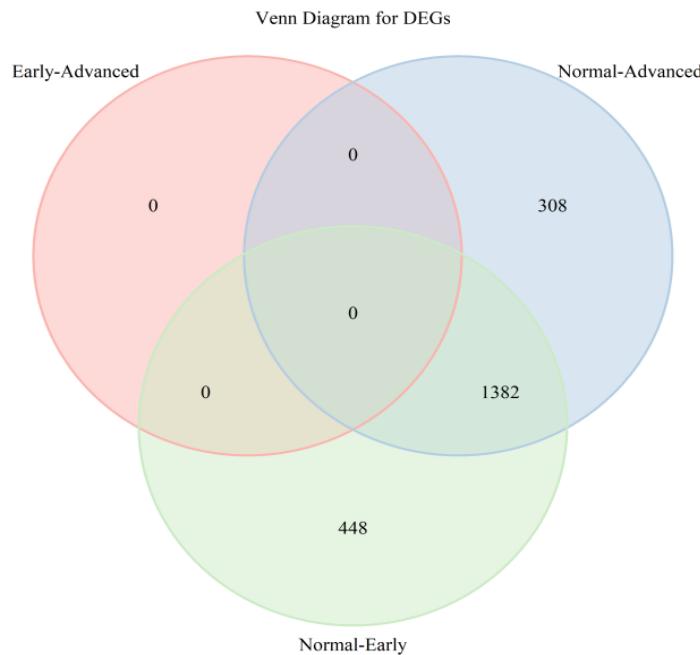
Ανάμεσα στο πρώιμο στάδιο και στους φυσιολογικούς είχαμε 1036 υπερεκφραζόμενα και 794 υποεκφραζόμενα γονίδια. Δεν υπήρξε και εδώ κανένα ΔΕ γονίδιο ανάμεσα στους καρκίνους προχωρημένου και πρώιμου σταδίου.

Τέλος με την διασπαστική προσέγγιση, είχαμε 474 υπερεκφραζόμενα και 470 υποεκφραζόμενα γονίδια ανάμεσα στο προχωρημένο στάδιο σε σχέση με τους φυσιολογικούς ιστούς. Ανάμεσα στο πρώιμο στάδιο και στους φυσιολογικούς είχαμε 546 υπερεκφραζόμενα και 530 υποεκφραζόμενα γονίδια. Ωστόσο εδώ βρέθηκαν 2 υπερεκφραζόμενα γονίδια στο προχωρημένο στάδιο καρκίνου σε σχέση με το πρώιμο, τα γονίδια αυτά είχαν αριθμούς Entrez 3008 και 1301.

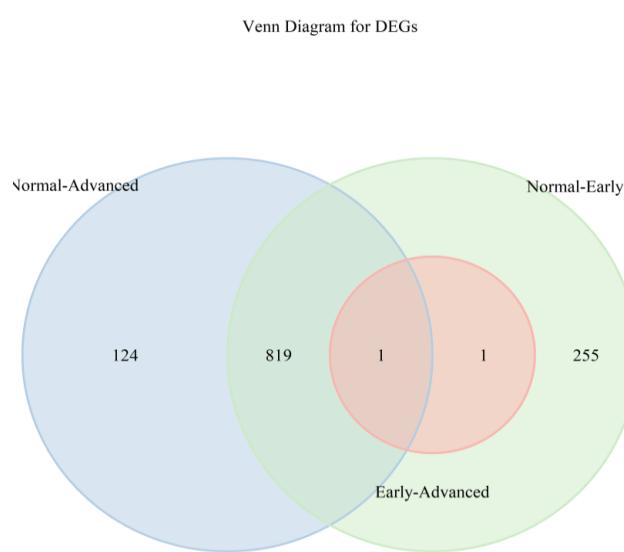
Όπως ήταν αναμενόμενο από τα παραπάνω, γραφικά στον θερμικό χάρτη με την ιεραρχική ομαδοποίηση είναι φανερή η πλήρης αδυναμία διαχωρισμού ανάμεσα στα 2 καρκινικά στάδια και με τις 3 προσεγγίσεις μετα-ανάλυσης. Τα παραπάνω αποτελέσματα φαίνονται παραστατικά με Venn Διαγράμματα και Θερμικούς Χάρτες στις Εικόνες 43-48. Στους θερμικούς χάρτες με κόκκινο απεικονίζονται τα AKK προχωρημένου σταδίου, με κίτρινο του πρώιμου σταδίου ενώ με πράσινο οι φυσιολογικοί ιστοί.



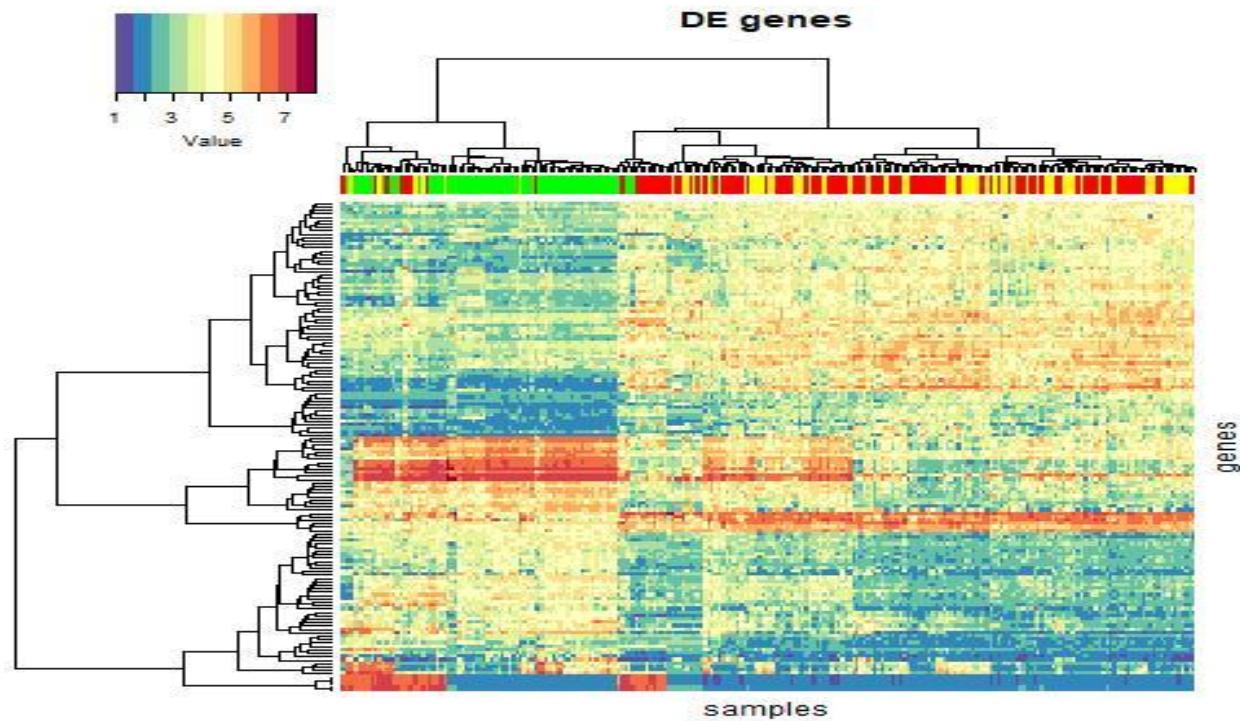
Εικόνα 43. Venn διάγραμμα ΔΕ γονιδίων με ανά στήλη ζ-κανονικοποίηση και ANOVA



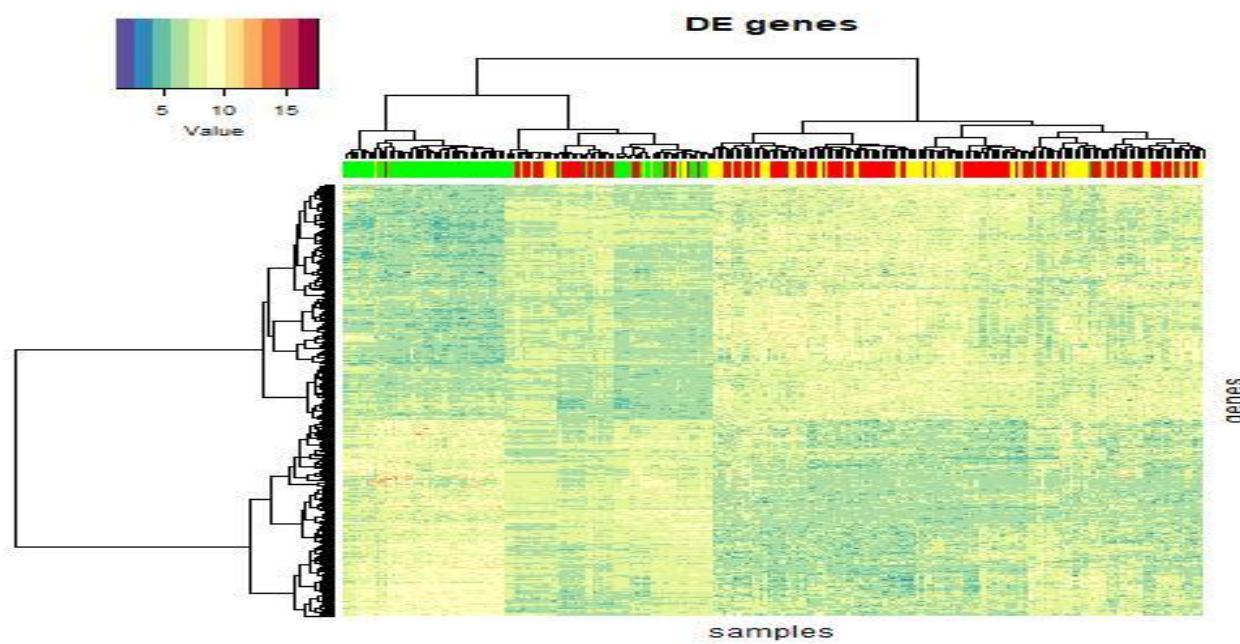
Εικόνα 44. Venn διάγραμμα ΔΕ με διπλή ζ-κανονικοποίηση και ANOVA



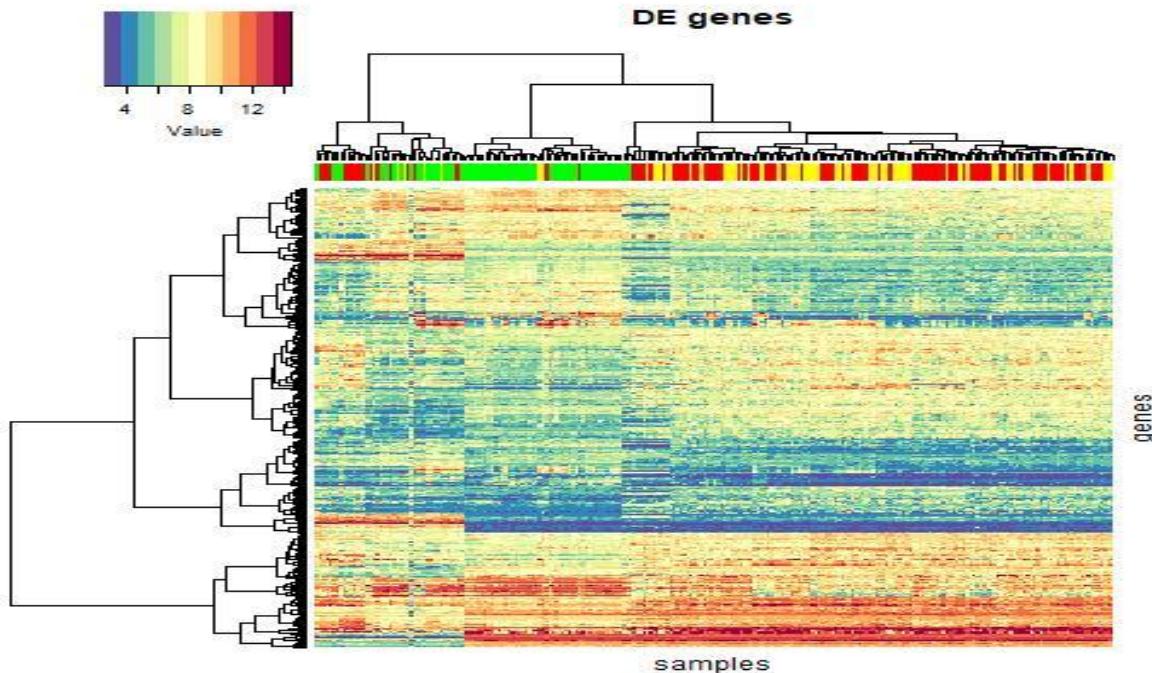
Εικόνα 45. Διαγράμματα Venn ΔΕ γονιδίων με την διασπαστική τεχνική και ANOVA



Εικόνα 46. Heatmap με ζ -κανονικοποίηση ανά στήλη και ANOVA



Εικόνα 47. Heatmap με διπλή ζ -κανονικοποίηση και ANOVA



Εικόνα 48. Heatmap με την διασπαστική προσέγγιση και ANOVA

3.5 Λειτουργική Ανάλυση γονιδιακής έκφρασης

Τα αποτελέσματα της λειτουργικής ανάλυσης γονιδιακής έκφρασης φαίνονται στους πίνακες 10-12. Σε αυτούς η προτελευταία στήλη περιέχει τα p-value που αντιστοιχούν στα υπερεκφραζόμενα γονίδια ενώ η τελευταία εκείνα των υποεκφραζόμενων. Η προσέγγιση με την ζ-κανονικοποίηση μόνο σε στήλες έδωσε τις λιγότερες κατηγορίες που είναι εμπλουτισμένες σε στατιστικά σημαντικό βαθμό, όλες οι κατηγορίες που αναφέρονται με αυτόν τον τρόπο κανονικοποίησης υπάρχουν και στους άλλους 2 πίνακες με εξαίρεση τον μεταβολισμό Τυροσίνης στα υποεκφραζόμενα γονίδια. Οι άλλες 2 προσεγγίσεις έδωσαν παρόμοιο αριθμό λειτουργικών κατηγοριών με την διασπαστική τεχνική να έχει ελαφρά περισσότερες. Σε γενικές γραμμές σε όλες τις προσεγγίσεις υπάρχουν αρκετές κατηγορίες που έχουν σχέση με την ανοσολογική απόκριση στα υπερεκφραζόμενα γονίδια. Παρόλα αυτά τα αποτελέσματα της διασπαστικής τεχνικής και της διπλής ζ-κανονικοποίησης συμφωνούν καλύτερα με την φύση του AKK όπως περιγράφεται στην βιβλιογραφία, με μικρή υπεροχή της διασπαστικής τεχνικής.

id	term_id	term_name	term_size	p_value query_1	p_value query_2
1	KEGG:04657	IL-17 signaling pathway	94	3.1e-06	1.0e+00
2	KEGG:04061	Viral protein interaction with cytokine and cytokine receptor	98	4.1e-05	1.0e+00
3	KEGG:04512	ECM-receptor interaction	88	6.0e-04	1.0e+00
4	KEGG:00350	Tyrosine metabolism	35	1.0e+00	2.6e-03
5	KEGG:04060	Cytokine-cytokine receptor interaction	293	7.5e-03	1.0e+00

gProfiler (biit.cs.ut.ee/gprofiler)

Πίνακας 10. Λειτουργική Ανάλυση ΔΕ γονιδίων στην ζ-κανονικοποίηση μονο σε στήλες

id	term_id	term_name	term_size	p_value query_1	p_value query_2
1	KEGG:04512	ECM-receptor interaction	88	1.7e-08	1.0e+00
2	KEGG:05146	Amoebiasis	101	4.0e-07	1.0e+00
3	KEGG:00982	Drug metabolism - cytochrome P450	68	1.0e+00	1.1e-06
4	KEGG:04510	Focal adhesion	200	6.2e-06	1.0e+00
5	KEGG:05204	Chemical carcinogenesis	80	1.0e+00	2.6e-05
6	KEGG:00053	Ascorbate and aldarate metabolism	30	1.0e+00	7.3e-05
7	KEGG:00980	Metabolism of xenobiotics by cytochrome P450	73	1.0e+00	1.1e-04
8	KEGG:04933	AGE-RAGE signaling pathway in diabetic complications	100	2.9e-04	1.0e+00
9	KEGG:04974	Protein digestion and absorption	103	4.0e-04	1.0e+00
10	KEGG:05222	Small cell lung cancer	92	5.5e-04	1.0e+00
11	KEGG:04976	Bile secretion	90	1.0e+00	8.3e-04
12	KEGG:05323	Rheumatoid arthritis	89	1.1e-03	1.0e+00
13	KEGG:05165	Human papillomavirus infection	331	1.2e-03	1.0e+00
14	KEGG:00830	Retinol metabolism	68	1.0e+00	3.2e-03
15	KEGG:05205	Proteoglycans in cancer	205	8.0e-03	1.0e+00
16	KEGG:03320	PPAR signaling pathway	75	9.9e-01	9.6e-03

g:Profiler (biit.cs.ut.ee/gprofiler)

Πίνακας 11. Λειτουργική ανάλυση ΔΕ γονιδίων με την μέθοδο της διπλής κανονικοποίησης

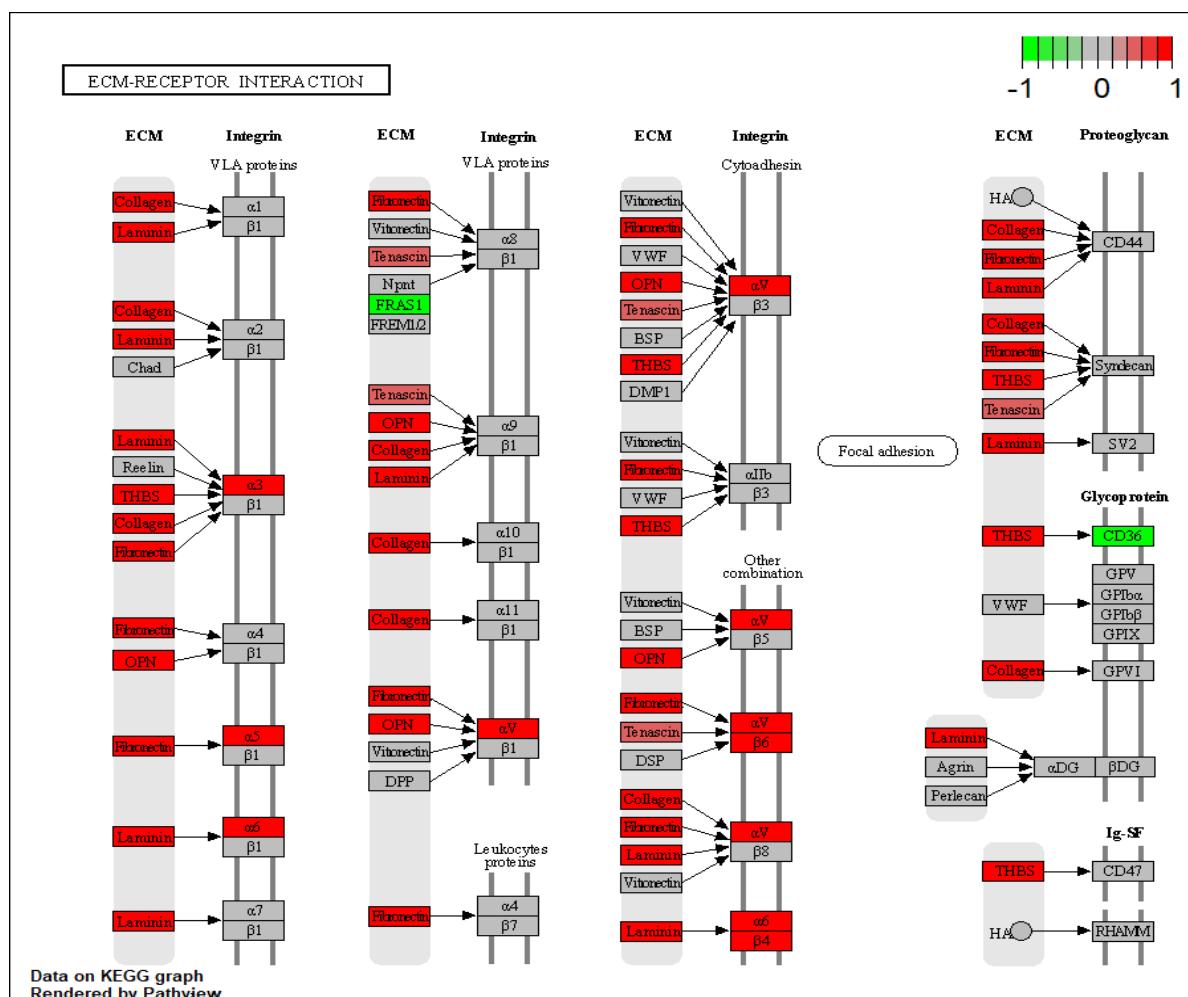
id	term_id	term_name	term_size	p_value query_1	p_value query_2
1	KEGG:04512	ECM-receptor interaction	88	2.6e-10	1.0e+00
2	KEGG:05146	Amoebiasis	101	7.2e-08	1.0e+00
3	KEGG:00053	Ascorbate and aldarate metabolism	30	1.0e+00	4.3e-07
4	KEGG:00982	Drug metabolism - cytochrome P450	68	1.0e+00	4.8e-07
5	KEGG:04657	IL-17 signaling pathway	94	5.9e-07	1.0e+00
6	KEGG:00980	Metabolism of xenobiotics by cytochrome P450	73	1.0e+00	1.7e-06
7	KEGG:04510	Focal adhesion	200	5.0e-06	1.0e+00
8	KEGG:05204	Chemical carcinogenesis	80	1.0e+00	8.6e-06
9	KEGG:04974	Protein digestion and absorption	103	1.3e-05	1.0e+00
10	KEGG:04061	Viral protein interaction with cytokine and cytokine receptor	98	1.6e-05	1.0e+00
11	KEGG:04933	AGE-RAGE signaling pathway in diabetic complications	100	6.9e-05	1.0e+00
12	KEGG:00830	Retinol metabolism	68	1.0e+00	1.0e-04
13	KEGG:05165	Human papillomavirus infection	331	1.7e-04	1.0e+00
14	KEGG:04620	Toll-like receptor signaling pathway	102	2.4e-04	1.0e+00
15	KEGG:00040	Pentose and glucuronate interconversions	34	1.0e+00	2.7e-04
16	KEGG:00983	Drug metabolism - other enzymes	78	1.0e+00	4.8e-04
17	KEGG:00860	Porphyrin and chlorophyll metabolism	42	1.0e+00	5.9e-04
18	KEGG:05222	Small cell lung cancer	92	6.1e-04	1.0e+00
19	KEGG:00140	Steroid hormone biosynthesis	60	1.0e+00	1.0e-03
20	KEGG:04060	Cytokine-cytokine receptor interaction	293	2.1e-03	1.0e+00
21	KEGG:05323	Rheumatoid arthritis	89	3.9e-03	1.0e+00
22	KEGG:04062	Chemokine signaling pathway	190	5.0e-03	1.0e+00
23	KEGG:04151	PI3K-Akt signaling pathway	353	5.1e-03	1.0e+00

gProfiler (biit.cs.ut.ee/gprofiler)

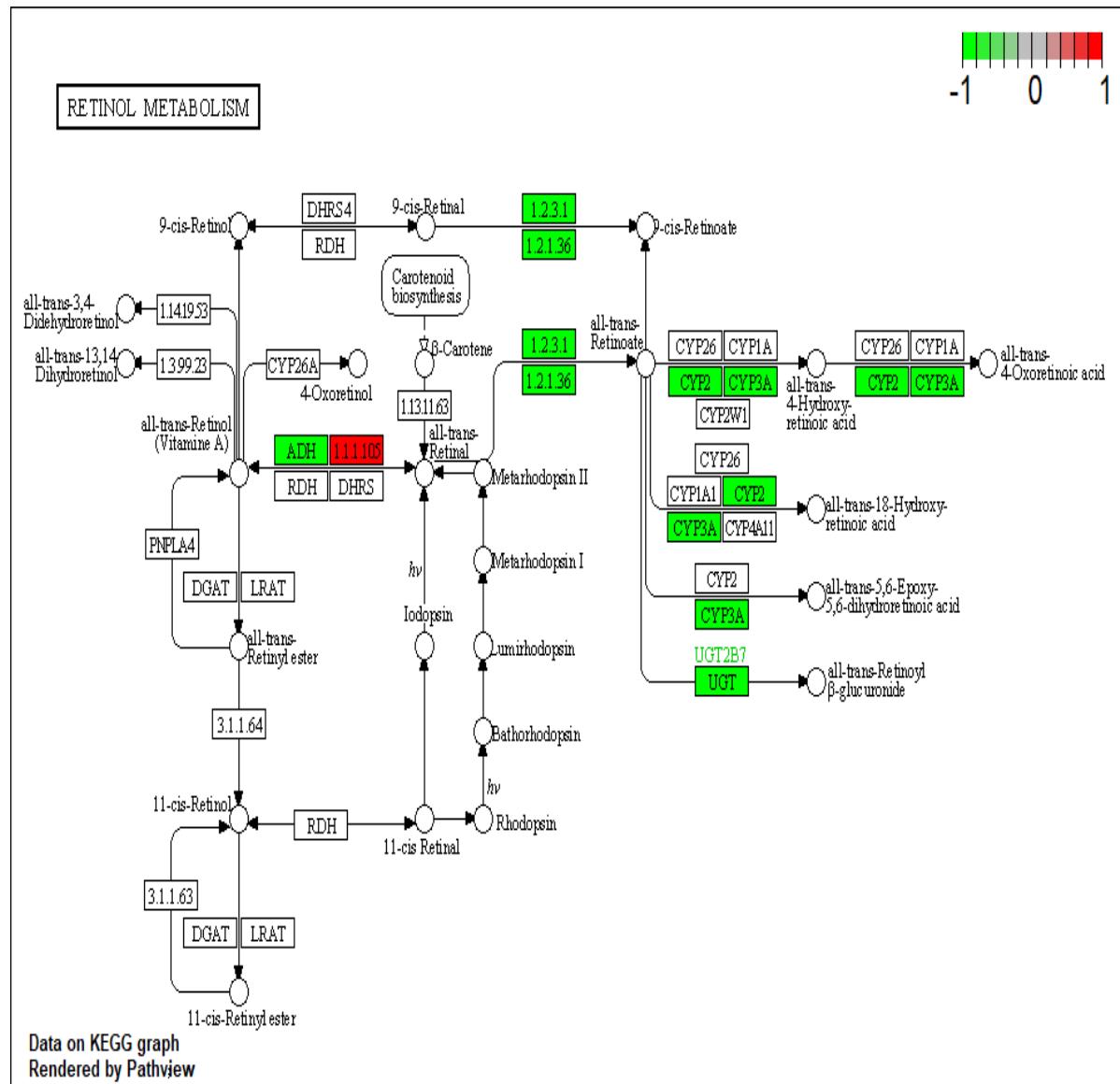
Πίνακας 12. Λειτουργική ανάλυση των ΔΕ γονιδίων στην διασπαστική τεχνική

3.6 Χαρτογράφηση ΔΕ γονιδίων σε γράφους της Kegg

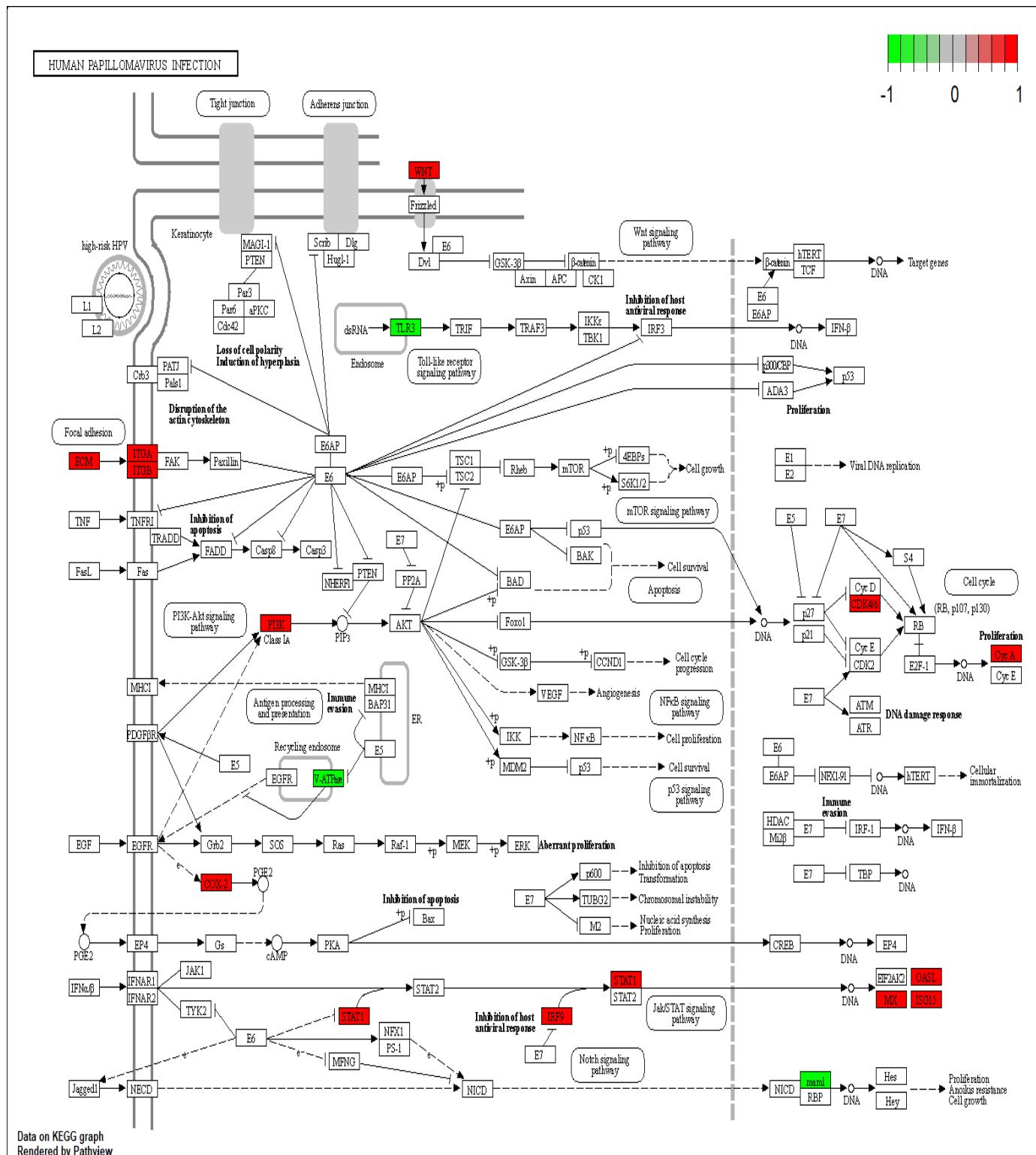
Από τα παραπάνω επιλέχθηκε ως πιο έγκυρη η προσέγγιση με την διασπαστική τεχνική η οποία από πλευράς τόσο ευαισθησίας και ειδικότητας όσο και από πλευράς αποτελεσμάτων κατά την λειτουργική ανάλυση φαίνεται να έδωσε τα πιο αξιόπιστα αποτελέσματα. Στις Εικόνες 49-58 φαίνεται η απεικόνιση των αποτελεσμάτων γονιδιακής έκφρασης στους χάρτες της Kegg σε επιλεγμένες κατηγορίες οι οποίες αναγράφονται άνω αριστερά στον κάθε χάρτη. Σε αυτούς του χάρτες με πράσινο απεικονίζονται τα υπο-εκφραζόμενα γονίδια ενώ με βάση την τιμή logFC των γονιδίων όπως αυτή προσδιορίστηκε στο στάδιο της στατιστικής ανάλυσης. Περισσότερα για την σχέση των παρακάτω κατηγοριών με το AKK του στόματος με παράθεση βιβλιογραφικών δεδομένων αναφέρονται στην συζήτηση.



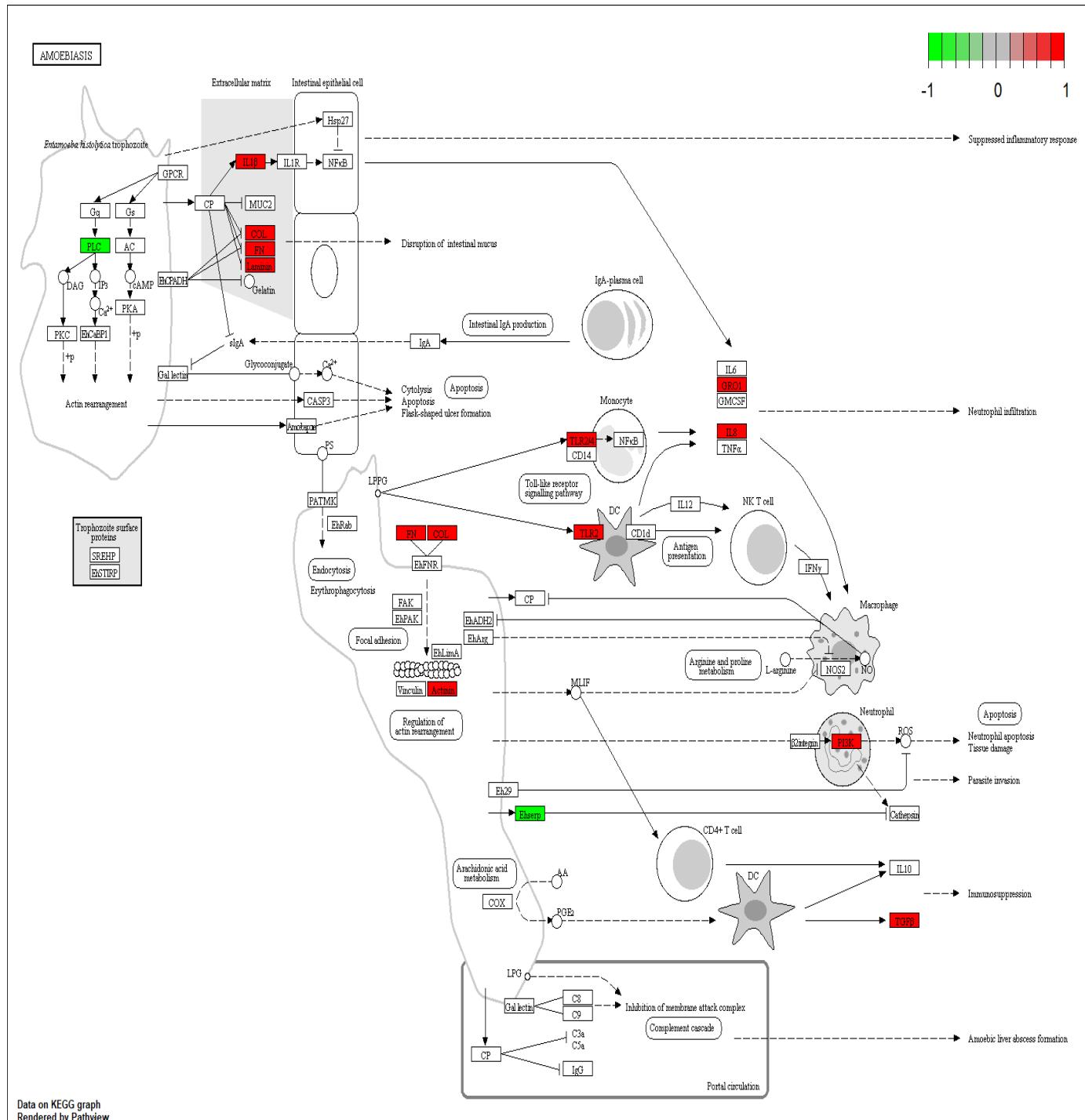
Εικόνα 49. Αλληλεπιδράσεις κυτταρικών υποδοχέων- εξωκυττάριας μήτρας



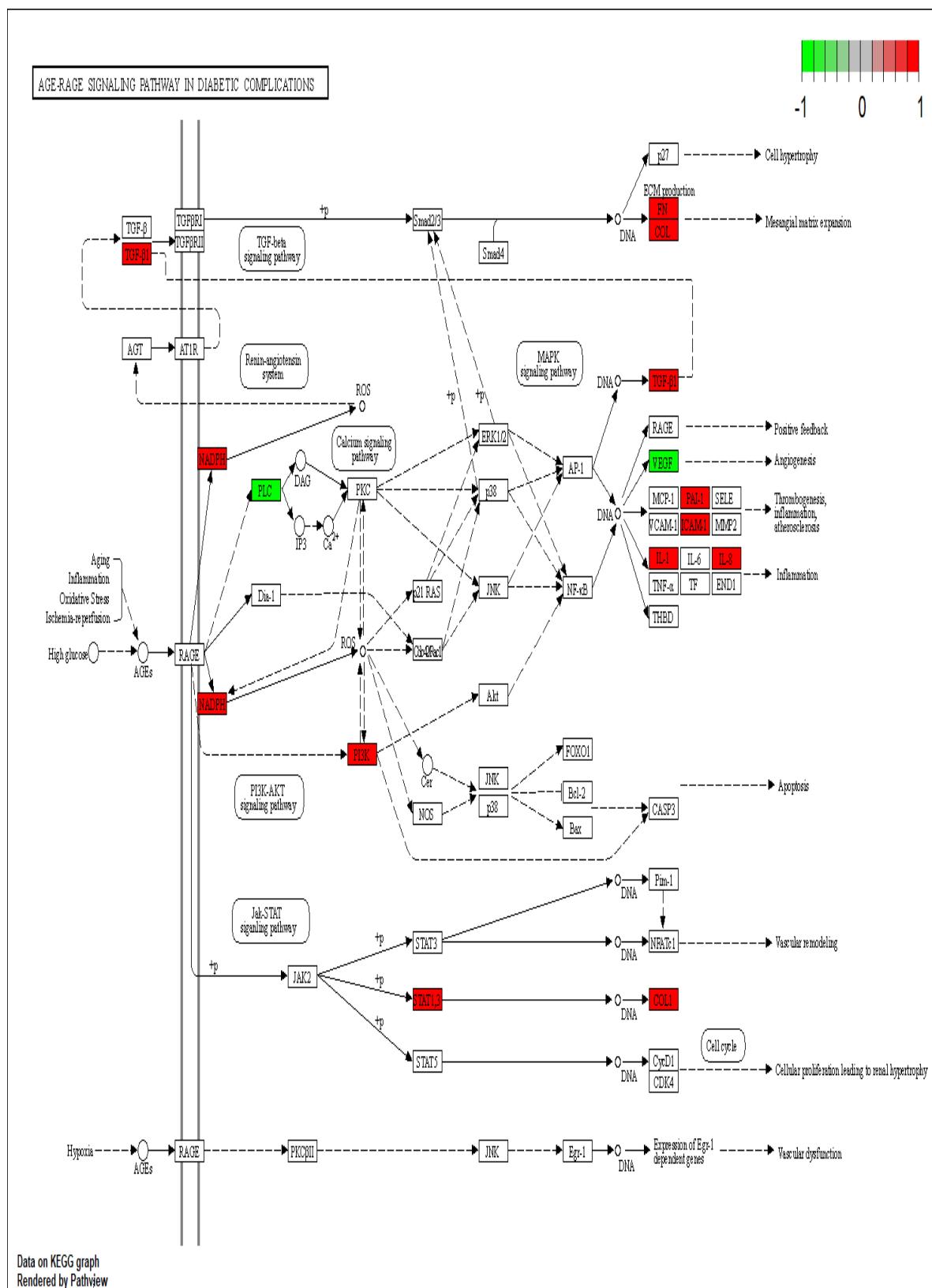
Εικόνα 50. Μεταβολισμός Ρετινόλης



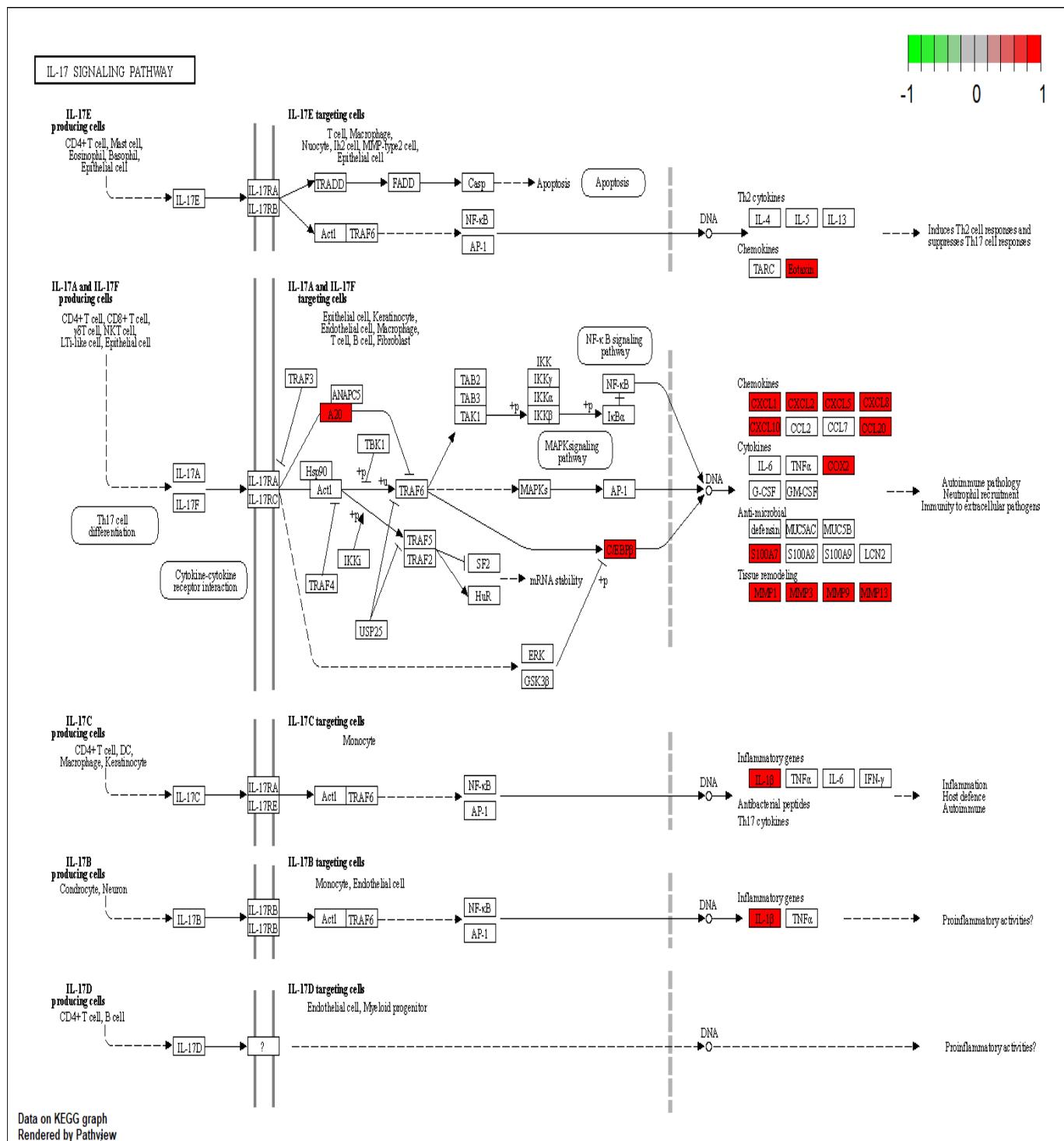
Εικόνα 51. Προσβολή από ιούς HPV



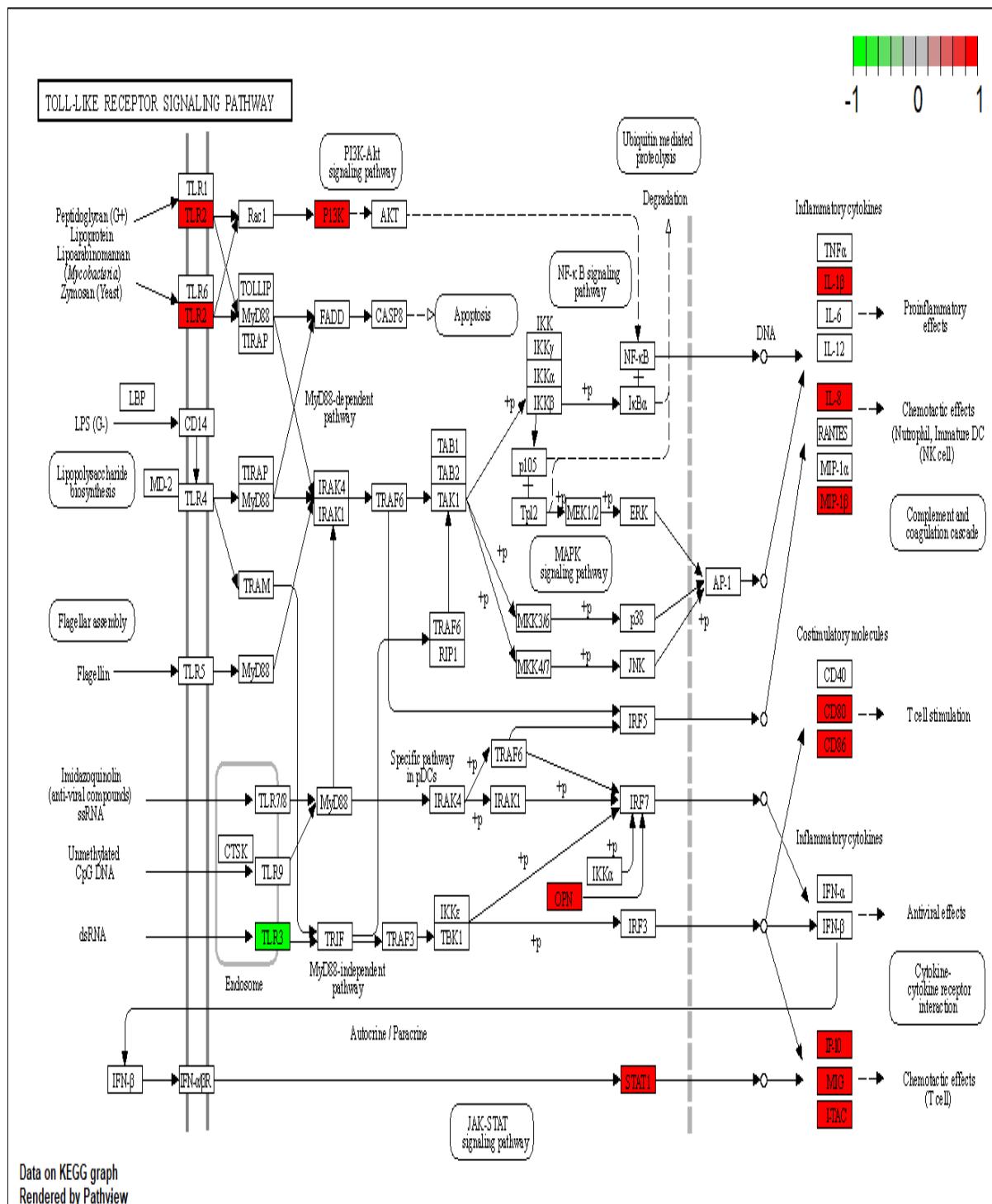
Εικόνα 52. Κοινοί ανοσολογικοί μηχανισμοί της αμοιβάδωσης με το ΑΚΚ



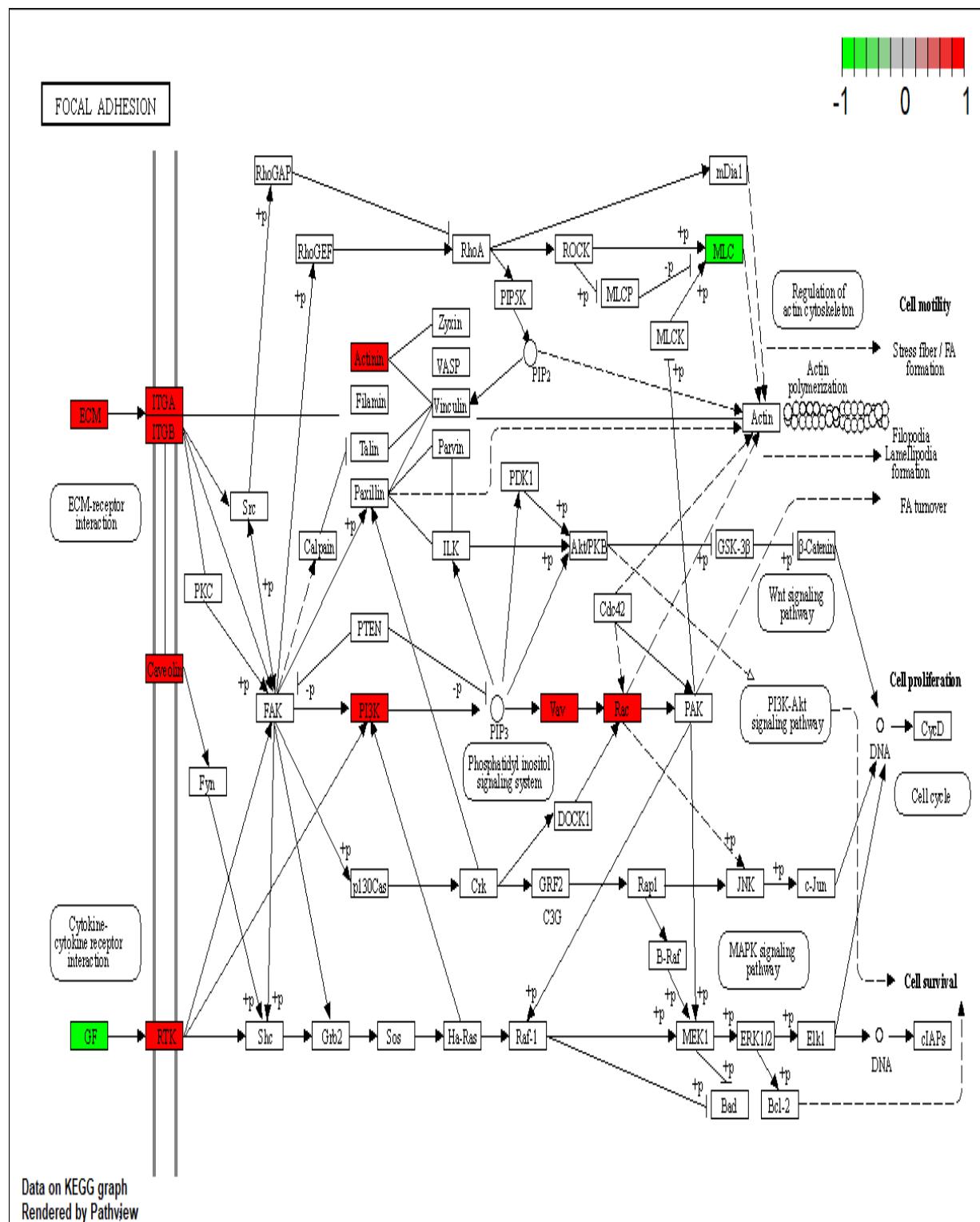
Εικόνα 53. Σημαντοδοτικό μονοπάτι AGE-RAGE



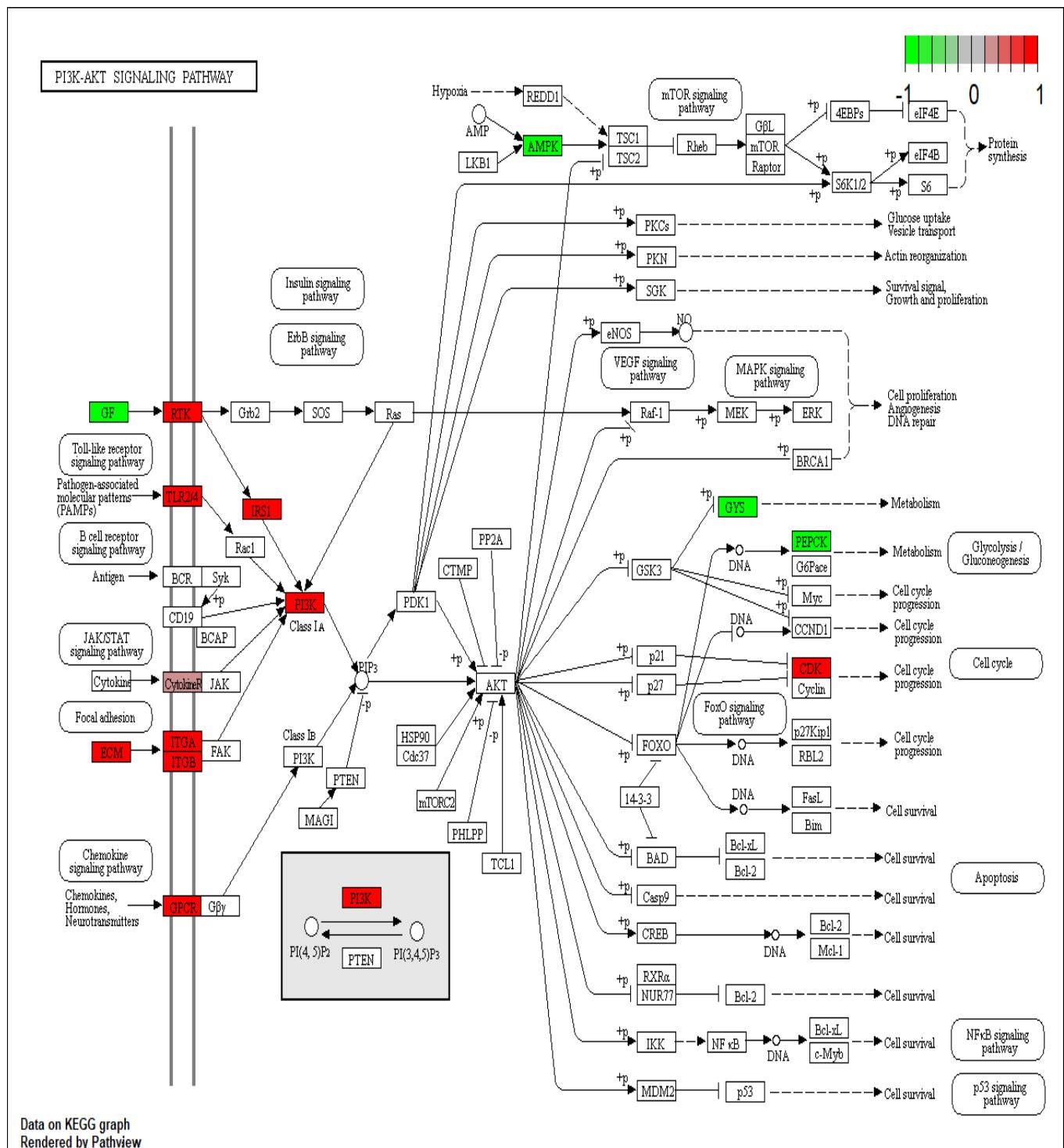
Εικόνα 54. Σηματοδοτικό μονοπάτι IL-17



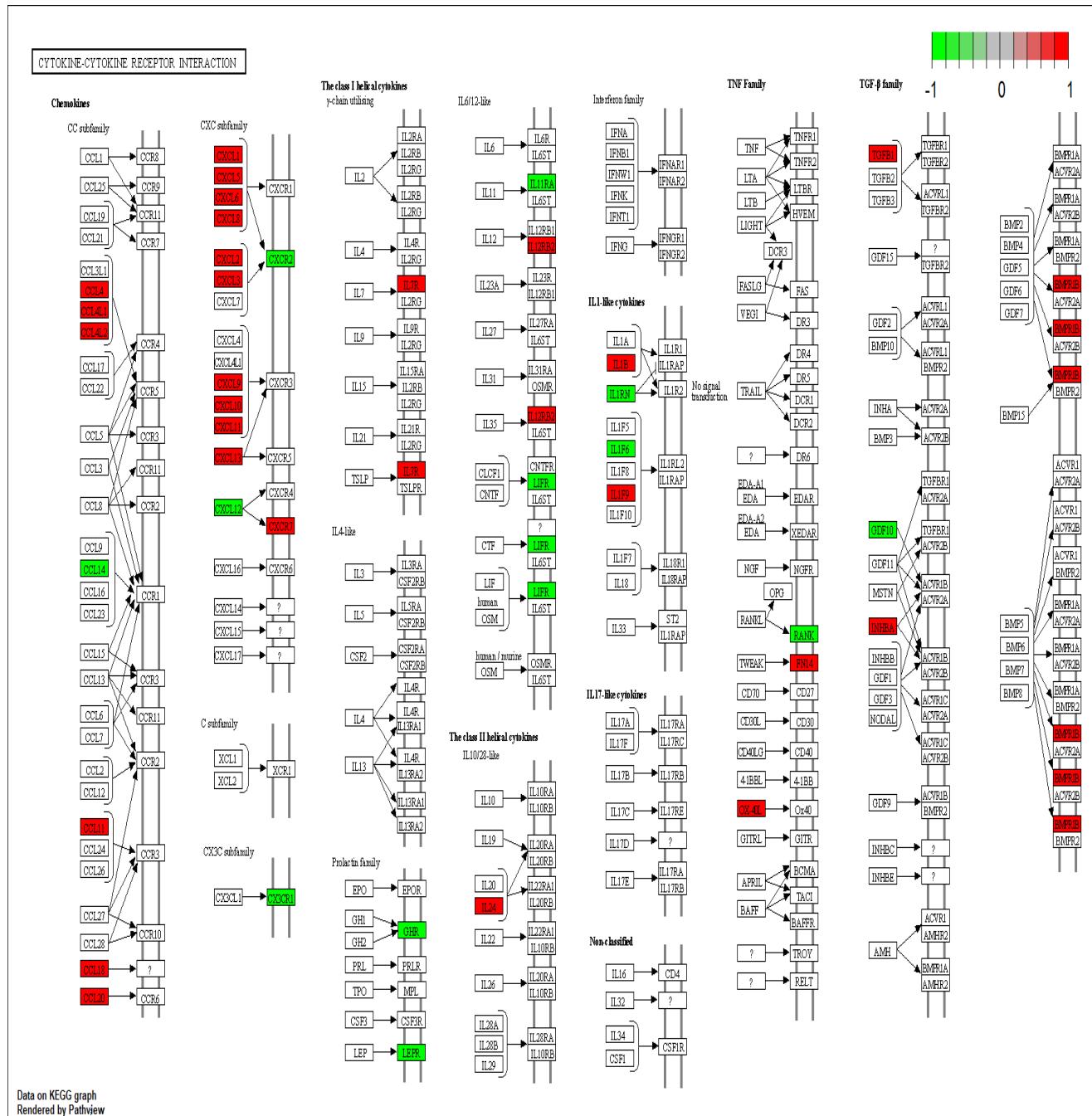
Εικόνα 55. Σηματοδοτικό μονοπάτι TLR (Toll-like Receptor)



Εικόνα 56. Μονοπάτι Εστιακής Συγκόλλησης



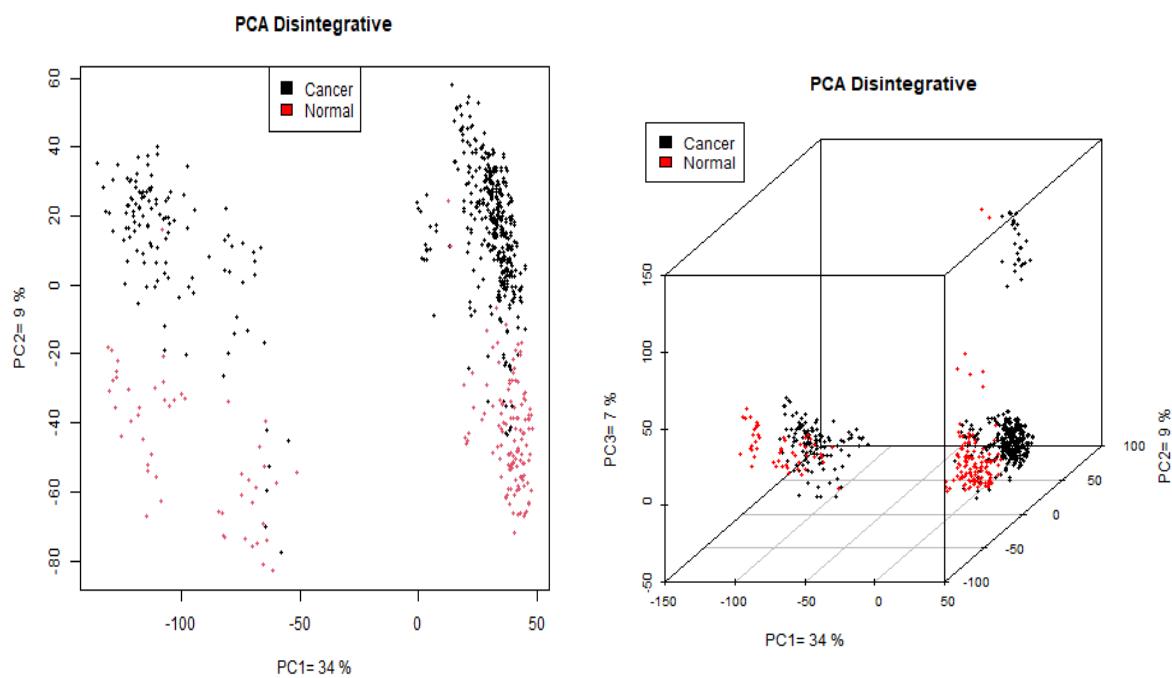
Εικόνα 57. Σηματοδοτικό μονοπάτι PI3/AKT/mTOR



Εικόνα 58. Αλληλεπιδράσεις κυτοκινών και οικείων υποδοχέων τους

3.7 Ανάλυση Κύριων Συνιστώσων (PCA)

Στην Ανάλυση κύριων συνιστώσων στις 2 διαστάσεις με τις κύριες συνιστώσες να εκφράζουν το 34% και 9% (αθροιστικά 43%) επί της συνολικής διασποράς των δεδομένων παρατηρούνται ξεκάθαρα 2 μη ομοιογενή clusters, το ένα μακριά από το άλλο τα οποία διαχωρίζονται με μια νοητή κάθετη γραμμή στο μέσο του διαγράμματος. Με μία οριζόντια νοητή γραμμή ωστόσο θα μπορούσε υπάρχει καλός διαχωρισμός (αλλά όχι τέλειος) ανάμεσα στα καρκινικά και μη καρκινικά δείγματα. Αν προστεθεί και Τρίτη διάσταση η οποία προσδίδει +7% στην συνολική διασπορά, φαίνονται 2 ανομοιογενή μεγάλα clusters και 1 μικρότερο που αποτελείται κυρίως από καρκινικά δείγματα. Τα 2 clusters που παρατηρούνται κατά μήκος της πρώτης κύριας συνιστώσας δεν μπορούν να εξηγηθούν στα πλαίσια της κλινικής διάγνωσης ή όχι του όγκου, και ενδεχομένως κάποιο άλλο απροσδιόριστο αίτιο προκαλεί αυτό τον διαχωρισμό.



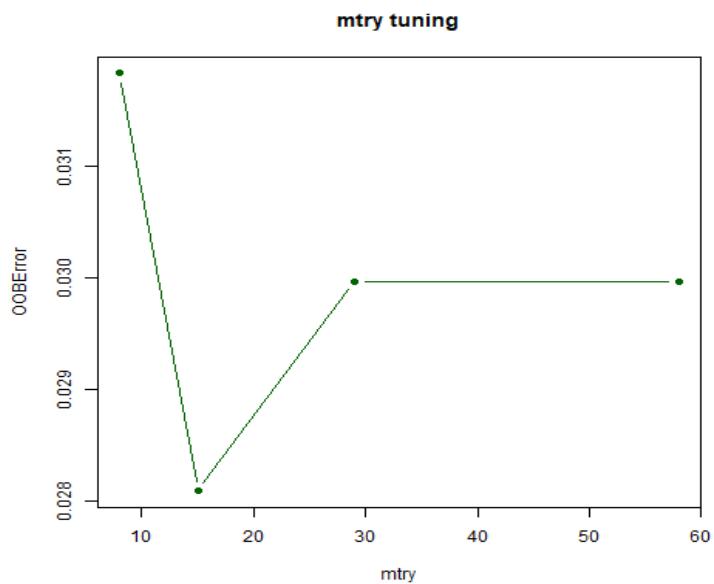
Εικόνα 59. Διαγράμματα αποτελεσμάτων στην PCA

3.8 Μηχανική Μάθηση I. Τυχαιοποιημένα Δάση

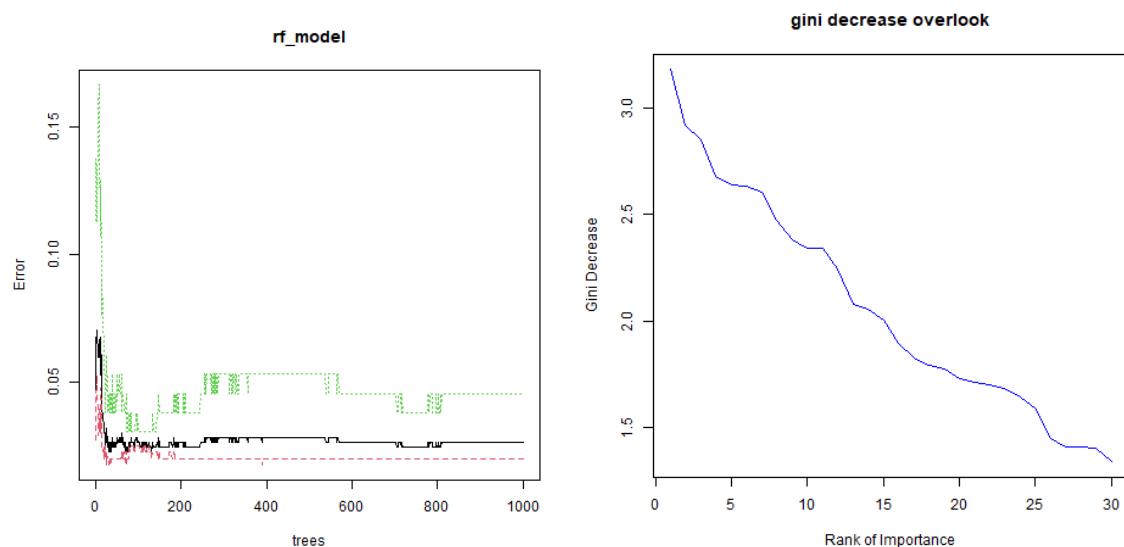
Τα αποτελέσματα εφαρμογής του αλγορίθμου τυχαιοποιημένων δασών φαίνονται στον Πίνακα 13. Στις εικόνες 60,61 και 62 παρακάτω φαίνεται η απόδοση, ο προσδιορισμός των οικείων υπερπαραμέτρων και τέλος παρατίθεται ο πίνακας σύγχυσης του αλγορίθμου στον Πίνακα 14. Η ευαισθησία ενός ταξινομητή που αποτελείται και από τα 30 γονίδια είναι 95.4 % ενώ η ειδικότητα είναι 100%.

Entrez ID	Genes	Description	Gini Decrease	Είδος ΔΕ
4312	MMP1	matrix metallopeptidase 1	3.179965	Up
8038	ADAM12	ADAM metallopeptidase domain 12	2.921605	Up
4321	MMP12	matrix metallopeptidase 12	2.852965	Up
5328	PLAU	plasminogen activator, urokinase	2.678626	Up
3851	KRT4	keratin 4	2.638946	Down
9645	MICAL2	microtubule associated monooxygenase, calponin and LIM domain containing 2	2.633087	Up
3371	TNC	tenascin C	2.603699	Up
4651	MYO10	myosin X	2.472079	Up
3624	INHBA	inhibin subunit beta A	2.380489	Up
5507	PPP1R3C	protein phosphatase 1 regulatory subunit 3C	2.342454	Down
4319	MMP10	matrix metallopeptidase 10	2.342138	Up
5361	PLXNA1	plexin A1	2.245871	Up
5690	PSMB2	proteasome 20S subunit beta 2	2.079216	Up
81617	CAB39L	calcium binding protein 39 like	2.058417	Down
4318	MMP9	matrix metallopeptidase 9	2.006228	Up
4118	MAL	mal, T cell differentiation protein	1.895065	Down
2012	EMP1	epithelial membrane protein 1	1.829282	Down
1001	CDH3	cadherin 3	1.795067	Up
1675	CFD	complement factor D	1.777362	Down
51660	MPC1	mitochondrial pyruvate carrier 1	1.728949	Down
23171	GPD1L	glycerol-3-phosphate dehydrogenase 1 like	1.714	Down
3909	LAMA3	laminin subunit alpha 3	1.702188	Up
4306	NR3C2	nuclear receptor subfamily 3 group C member 2	1.685726	Down
3872	KRT17	keratin 17	1.643351	Up
4430	MYO1B	myosin IB	1.594615	Up
5744	PTHLH	parathyroid hormone like hormone	1.450012	Up
7410	VAV2	vav guanine nucleotide exchange factor 2	1.413447	Up
3918	LAMC2	laminin subunit gamma 2	1.408899	Up
58528	RRAGD	Ras related GTP binding D	1.404603	Down
49860	CRNN	cornulin	1.3413	Down

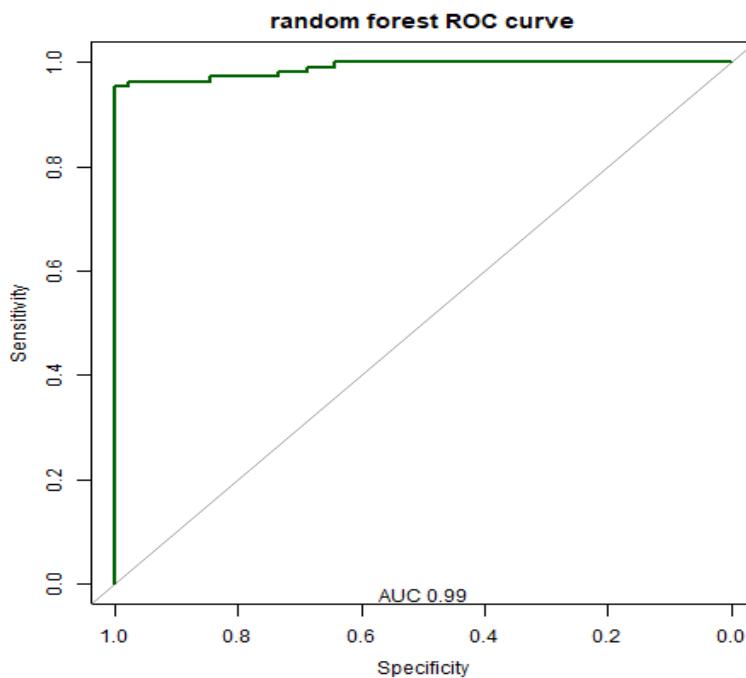
Πίνακας 13. Αριθμοί Entrez για τα 30 σημαντικότερα από ταξινομικής άποψης γονίδια



Εικόνα 60. Προσδιορισμός της υπερπαραμέτρου mtry για το μέγεθος των επιμέρους δέντρων



Εικόνα 61. Απόδοση Αλγορίθμου τυχαιοποιημένων Δασών.



Εικόνα 62. Καμπύλη ROC για τα Τυχαιοποιημένα Δάση

Pred/Reference	Cancer	Normal
Cancer	103	0
Normal	5	45

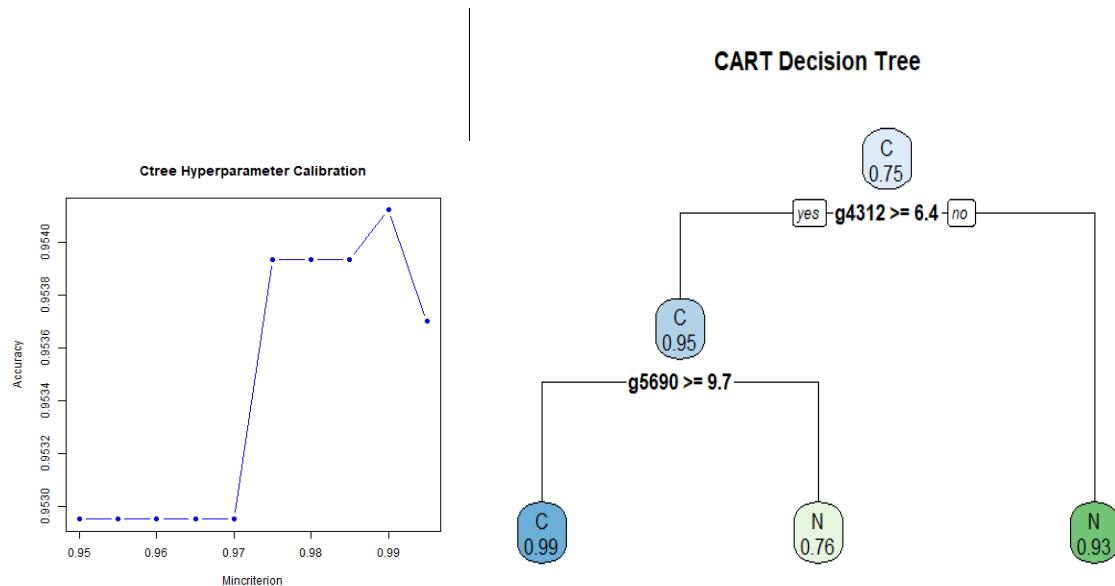
Πίνακας 14. Πίνακας Σύγχυσης Τυχαιοποιημένα Δάση

3.9 Μηχανική Μάθηση II. Δένδρα Απόφασης

Τα αποτελέσματα εφαρμογής των αλγορίθμων δένδρων απόφασης φαίνονται στις Εικόνες 63,64 και 65 που παρατίθενται παρακάτω, τα αντίστοιχα Entrez IDs είναι μετά το γράμμα g στους κόμβους του δένδρου. Στις εικόνες φαίνεται ο προσδιορισμός των οικείων υπερπαραμέτρων, το δένδρο απόφασης και τέλος παρατίθεται ο πίνακας σύγχυσης του κάθε αλγορίθμου (Πίνακες 15 και 16). Τέλος Παρατίθενται και οι καμπύλες ROC για κάθε αλγόριθμο (Εικόνα 66).

Η ευαισθησία για τον αλγόριθμο CART στον οποίον χρησιμοποιούνται μόνο 2 γονίδια είναι 91.7% ενώ η ειδικότητα 93.3%. Ο Αλγόριθμος C-tree αποδίδει λίγο καλύτερα με ευαισθησία

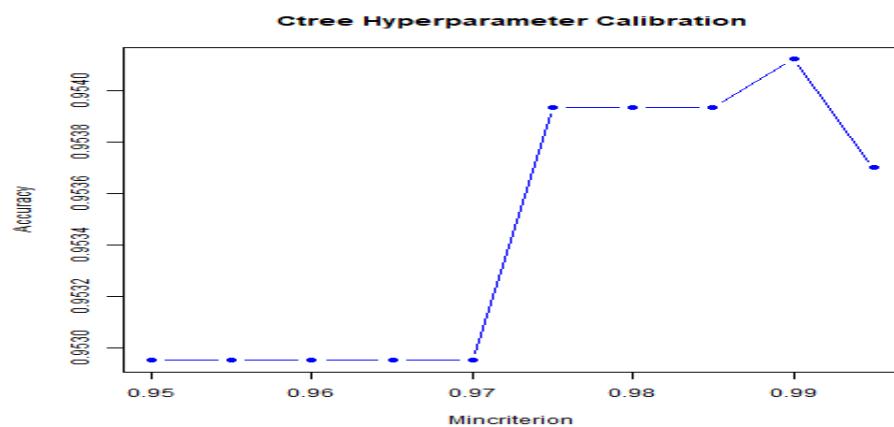
93.5% και ειδικότητα 95.6%, αλλά χρησιμοποιεί περισσότερα γονίδια. Οι επιφάνειες κάτω από την καμπύλη (Area Under the Curve) ήταν 92.8% και 94.6% για τους αλγόριθμους CART και C-tree αντίστοιχα.



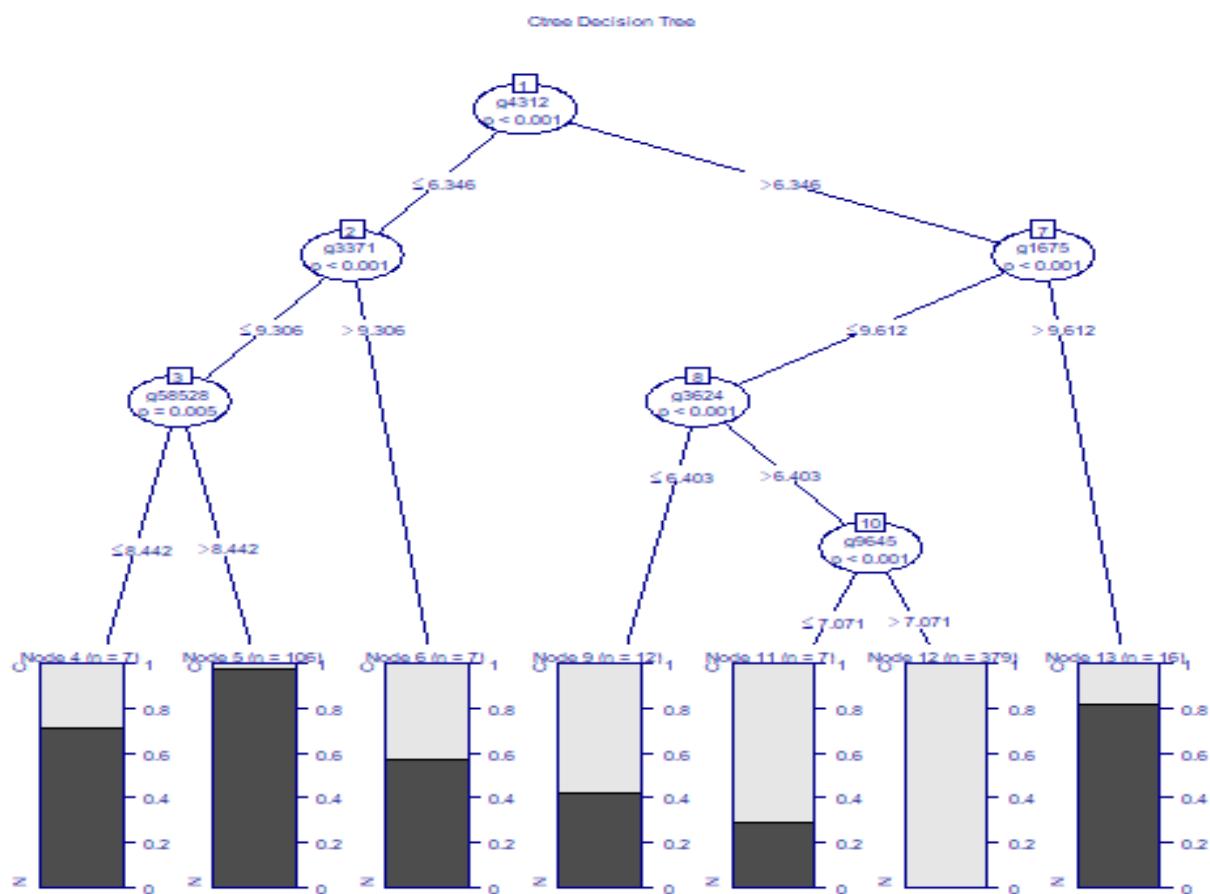
Εικόνα 63. Προσδιορισμός υπερπαραμέτρου περιπλοκότητας και το δένδρο απόφασης με τον αλγόριθμο CART

Pred/Reference	Cancer	Normal
Cancer	99	3
Normal	9	42

Πίνακας 15. Πίνακας Σύγχυσης αλγόριθμου CART



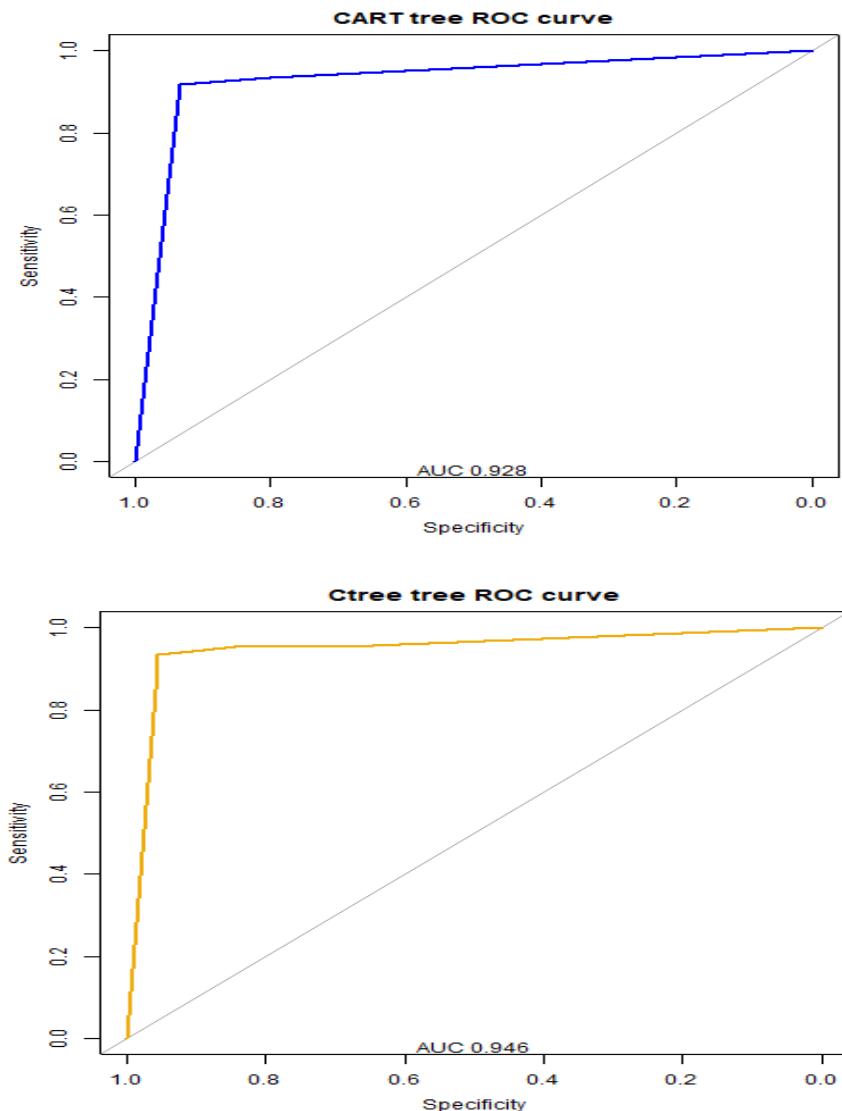
Εικόνα 64. Επιλογή υπερπαραμέτρου mincriterion



Εικόνα 65. Δένδρο Απόφασης με τον Αλγόριθμο C-tree

Pred/Reference	Cancer	Normal
Cancer	101	2
Normal	7	43

Πίνακας 16. Πίνακας Σύγχυσης για τον αλγόριθμο Ctree



Εικόνα 66. Καμπύλες ROC για τους 2 αλγορίθμους δένδρων απόφασης

3.10 Ανάλυση Βιολογικών Δικτύων

Στην ανάλυση δικτύων με το Cytoscape για τα υπερεκφραζόμενα γονίδια παρατηρήθηκαν 2 μη συνδεδεμένα μεγάλα clusters (Εικόνα 67). Το ένα από αυτά ήταν σχετικά συμπαγές (στην εικόνα το άνω δεξιά ορφανό cluster) χωρίς να μπορούν να διακριθούν ευδιάκριτα modules και περιλαμβάνει γονίδια που εμπλέκονται στον κυτταρικό κύκλο και ιδιαίτερα στην μιτωτική δραστηριότητα, σε αυτό το cluster κεντρικό ρόλο παίζει η CDK1 η οποία έχει τον μεγαλύτερο βαθμό κόμβου από όλα τα απεικονιζόμενα γονίδια στον γράφο και επίσης μεγάλη ενδιάμεση εκκεντρότητα.

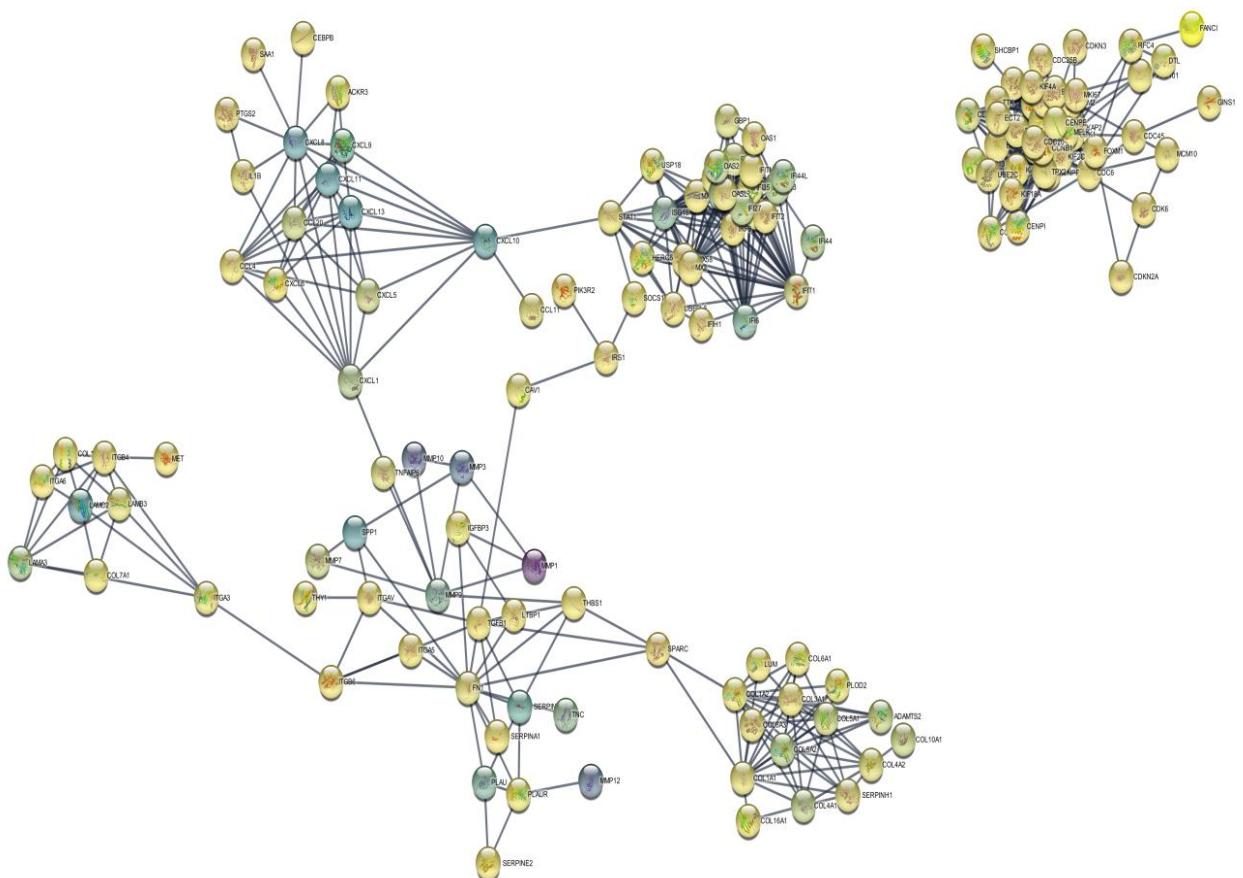
Στο 2^o μεγαλύτερο cluster διακρίνονται 5 modules. Το άνω αριστερά module όπως φαίνεται στην εικόνα 67, αποτελείται από υποδοχείς κυτοκινών/χημειοκινών. Το άνω δεξιά module αποτελείται από γονίδια που έχουν να κάνουν με ανοσολογική απόκριση σε ιογενές αντιγόνο. Το κεντρικό κάτω module αφορά σε γονίδια που παίζουν ρόλο στην Επιθηλιακή – Μεσεγχυματική Μετάβαση (EMT). Το δεξιό και το αριστερό κάτω module αποτελούνται από γονίδια με σημαντικό ρόλο στην οργάνωση της εξωκυττάριας μήτρας. Πιο συγκεκριμένα το δεξιό συμμετέχει στην οργάνωση των κολλαγόνων ινών και του υποστρώματος της εξωκυτταρικής μήτρας ενώ το αριστερό έχει να κάνει με συνδέσεις τόσο των επιθηλιακών κυττάρων μεταξύ τους αλλά περισσότερο με σύνδεση των επιθηλιακών κυττάρων με τον συνδετικό ιστό μέσω ημιδεσμοσωμάτων.

Τα 5 modules που προαναφέραμε συνδέονται μεταξύ τους με συγκεκριμένα bottleneck hub γονίδια-κόμβους. Αυτά τα bottleneck γονίδια είναι όπως απεικονίζονται στην εικόνα 67 τα εξής:

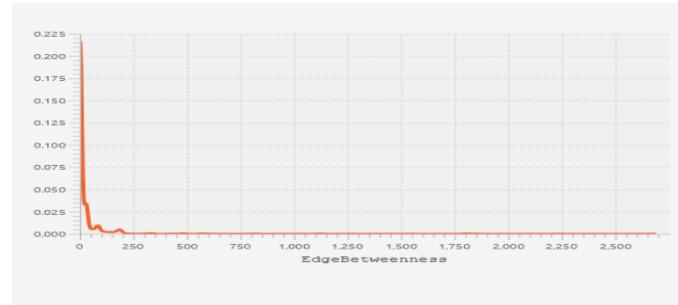
- Τα 2 άνω modules συνδέονται μεταξύ τους με το CXCL1 μέσω του CXCL10 και του STAT1.
- Το άνω αριστερά module συνδέεται με το κεντρικό κάτω επίσης με το CXCL1 και το MMP9.
- Το άνω δεξιά συνδέεται με το κεντρικό κάτω μέσω του STAT1.
- Το αριστερό κάτω module συνδέεται με το κεντρικό κάτω module μέσω του ITGB6
- Το δεξιό κάτω module συνδέεται με το κεντρικό κάτω module μέσω του SPARC (οστεονεκτίνη).

Από όλα τα bottleneck hub γονίδια αυτό με το μεγαλύτερο logFC , δηλαδή με την μεγαλύτερη διαφορά στην γονιδιακή έκφραση είναι το CXCL10 ακολουθούμενο από την MMP9 και τρίτη την CXCL1. Τα modules προσομοιάζουν στα χαρακτηριστικά ενός τυχαιοποιημένου δικτύου (Εικόνα 69), ενώ ολόκληρος ο γράφος έχει χαρακτηριστικά iεραρχικού δικτύου ελεύθερης κλίμακας (scale – free network) (Εικόνες 68 και 70).

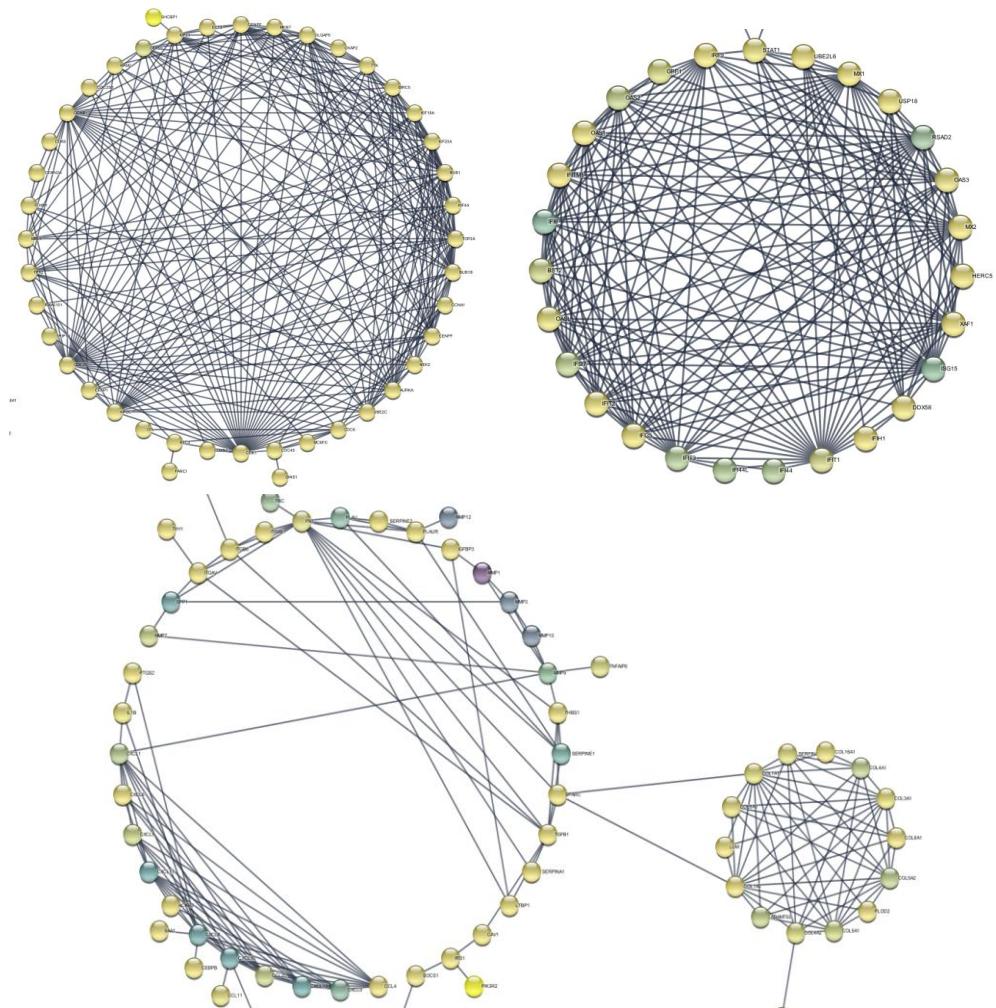
Λεπτομέρειες σχετικά με τις ιδιότητες του δικτύου των υπερεκφραζόμενων γονιδίων είναι στον πίνακα 17.



Εικόνα 67. Βιολογικό δίκτυο των υπερεκφραζόμενων γονιδίων με λειτουργική συνδεση



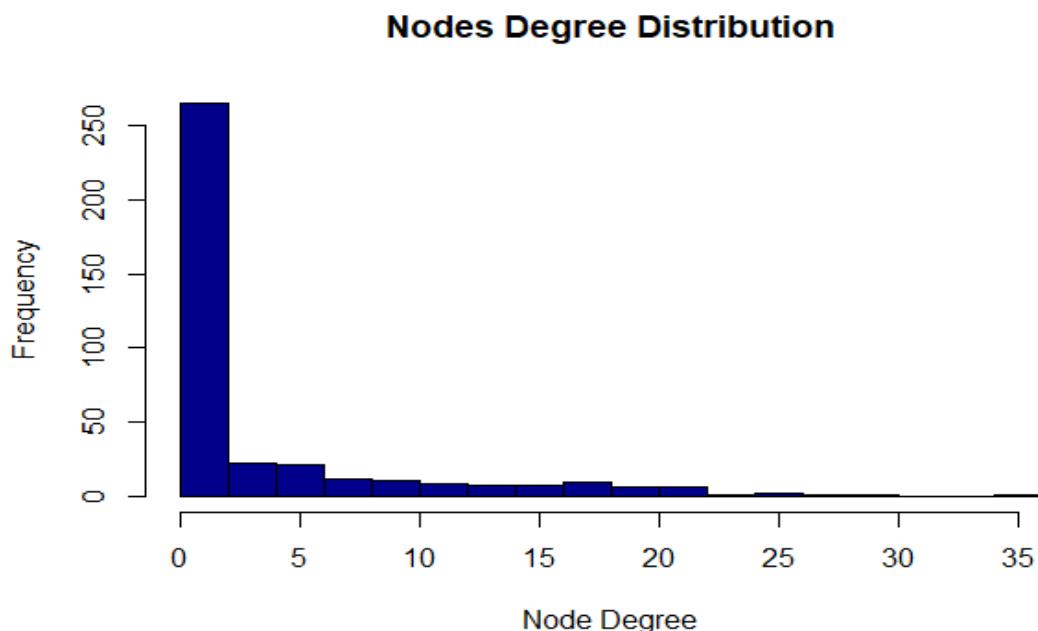
Εικόνα 68. Κατανομή Ενδιάμεσης Εκκεντρότητας



Εικόνα 69. Κυκλοτερείς γράφοι των clusters. Το κάτω μέρος έχει 2 ημικυκλικούς και 1 κυκλικό γράφο.

gene	Degree	Betweeness	Clust. Coeff
STAT1	11	0.419	0.527
MMP9	7	0.325	0.095
CXCL1	10	0.303	0.533
CDK1	35	0.281	0.361
SPARC	5	0.271	0.4
TGFB1	9	0.265	0.278
CXCL10	10	0.251	0.533
THBS1	5	0.247	0.5
FN1	14	0.18	0.176
ITGB6	5	0.178	0.4

Πίνακας 17. Οι 10 πιο σημαντικοί κόμβοι από άποψη ενδιάμεσης εκκεντρότητας



Εικόνα 70. Κατανομή βαθμών κόμβων

Κεφάλαιο 4. Συζήτηση

Στην διπλωματική αυτή εργασία έγινε προσπάθεια εναρμόνισης των δεδομένων από όλα τα dataset με τρόπους που τα καθιστούν καλή πρώτη ύλη για μία περαιτέρω ανάλυση σε επίπεδο τόσο στατιστικής όσο και μηχανικής μάθησης και ανάλυσης δικτύων. Υπάρχουν 2 προσεγγίσεις για αυτόν τον σκοπό η σειριακή (ή ύστερη εναρμόνιση) όπου γίνεται εξαγωγή των ΔΕ γονιδίων από κάθε dataset ξεχωριστά και μετά αξιολογούνται με διάφορους τρόπους τα κοινά σημεία τους, και η παράλληλη (ή πρώιμη εναρμόνιση) στην οποία κατασκευάζεται ένας τεράστιος πίνακας γονιδιακής έκφρασης για περαιτέρω ανάλυση (Walsh, Hu, Batt & Santos, 2015). Ενώ η σειριακή προσέγγιση είναι πιο εύκολο να εφαρμοστεί αποκλείει εκ προοιμίου την εφαρμογή μεθόδων μηχανικής μάθησης και επομένως δεν θεωρείται κατάλληλη για μία μετα-ανάλυση που θέλει να εφαρμόσει τέτοιες μεθόδους. Από την άλλη μεριά η παράλληλη προσέγγιση είναι πιο δύσκολο να σχεδιαστεί και να εφαρμοστεί αλλά το αποτέλεσμα της είναι κατάλληλο για μία περαιτέρω ανάλυση.

Οι 2 πρώτοι τρόποι, που εφαρμόστηκαν στα πλαίσια αυτής της διπλωματικής, χρησιμοποιούν ζ-κανονικοποίηση και χρησιμοποιούνται στην βιβλιογραφία για την μετα-ανάλυση δεδομένων μικροσυστοιχιών, ενώ ο τρίτος τρόπος που ονομάστηκε με τον περιγραφικό όρο «Διασπαστική τεχνική» είναι επινόηση του συγγραφέα αυτής της διπλωματικής εργασίας. Σε όλες τις περιπτώσεις προηγήθηκε ενδελεχής προκαταρκτικός έλεγχος των dataset και επιλέχθηκαν μόνο αυτά που πληρούσαν το κριτήριο της έστω και κατά προσέγγισης κανονικής κατανομής. Φαίνεται από την βιβλιογραφία αλλά και από την προκαταρκτική μελέτη μας επί συνόλου 37 dataset από διάφορες πλατφόρμες ότι αυτό το κριτήριο πληρείται κατά προσέγγιση σε πειράματα που χρησιμοποιούν Affymetrix μικροσυστοιχίες και επομένως η ζ-κανονικοποίηση είναι πολύ λογικό να εφαρμόζεται στην μίξη τέτοιων πειραμάτων (Cheadle, Cho-Chung, Becker, & Vawter, 2003).

Όσον αφορά τους 2 πρώτους τρόπους με τις 2 παραλλαγές εφαρμογής ζ-κανονικοποίησης, ο ένας είναι κατά στήλες μόνο ενώ ο άλλος κατά σειρές και στήλες. Να θυμίσουμε ότι στους πίνακες έκφρασης γονιδίων οι στήλες αφορούν τα δείγματα και οι σειρές τα γονίδια. Στην βιβλιογραφία δεν υπάρχει σαφής τρόπος με τον οποίο επιλέγεται πως ακριβώς θα εφαρμοστεί η ζ-κανονικοποίηση για την σύμπτυξη δεδομένων μετα-ανάλυσης. Υπάρχουν άλλοι που κάνουν μόνο σε γονίδια (Roden DL, Sewell, Loble, Levine, Smith et al, 2014) αλλά και

άλλοι που κάνουν και σε γονίδια και σε δείγματα (Yasrebi, 2016). Τέλος κάποιοι συγγραφείς κάνουν z-κανονικοποίηση μόνο σε δείγματα (Cheadle, Vawter, Freed & Becker, 2003 ; Chearle et al, 2003 ; Annavarapu, Dara & Banka, 2016).

Φαίνεται λοιπόν ότι υπάρχει μεγάλη ανομοιομορφία στον τρόπο που κάποιος μπορεί να εφαρμόσει μία z-κανονικοποίηση για να εναρμονίσει πολλά πειράματα μικροσυστοιχιών. Για την ακρίβεια αν κάποιος ανατρέξει στην βιβλιογραφία λίγο πολύ θα καταλάβει ότι κάθε συγγραφέας έχει και μια διαφορετική προσέγγιση για αυτό τον σκοπό. Άλλες τεχνικές μετα-ανάλυσης όπως η Combat έχει δειχθεί ότι δεν αποδίδουν καλύτερα αποτελέσματα σε σχέση με την z-κανονικοποίηση (Yasrebi, 2016). Από τα παραπάνω καταλαβαίνει κάποιος ότι δεν έχει βρεθεί μια πραγματικά αξιόπιστη μέθοδος με την οποία μπορεί κάποιος με ασφάλεια αν προβεί σε μία μετα-ανάλυση πολλών πειραμάτων γονιδιακής έκφρασης σε μικροσυστοιχίες.

Σε όλα τα datasets έγινε ανάλυση ποιότητας με τα εργαλεία RLE και NUSE που προτείνονται από την Affymetrix για την ανάλυση των μικροσυστοιχιών της. Το πακέτο MetaQC που χρησιμοποιούνταν σε αρκετές έρευνες πλέον έχει αποσυρθεί και δεν υποστηρίζεται λόγω προβλημάτων που έχουν αναδειχτεί σε αυτό και δεν διορθώθηκαν (<https://cran.r-project.org/web/packages/MetaQC/index.html>) .

Οι 3 τρόποι που προτείναμε υποβλήθηκαν σε έναν ιδιότυπο και πρωτότυπο πλην όμως εντελώς λογικό τεστ αξιοπιστίας με έναν συνδυασμό επιβλεπόμενης και μη επιβλεπόμενης προσέγγισης όμως αξιολογήθηκαν και τα αποτελέσματα τους και σε επίπεδο των λειτουργικών κατηγοριών που προκύπτουν ύστερα από την ανάλυση τους. Εφόσον γνωρίζουμε ποια είναι η κλινική διάγνωση για κάθε δείγμα, αν κάνουμε μία ιεραρχική ομαδοποίηση των δειγμάτων βάσει των ΔΕ γονιδίων, ιδανικά και εφόσον δεν έχουν παραμορφωθεί τα δεδομένα κατά τους χειρισμούς μετα-ανάλυσης θα είχαμε 2 cluster με το ένα να είναι καρκινικά και το άλλο φυσιολογικά δείγματα. Όμως αφού μιλάμε για ιεραρχική ομαδοποίηση που είναι μία μη επιβλεπόμενη τεχνική πως ξέρουμε ποιο cluster αντιπροσωπεύει καρκινικά και ποιο φυσιολογικά δείγματα? Εδώ λοιπόν κάναμε μία υπόθεση που ασφαλώς αποτελεί περιορισμό σε αυτήν την μελέτη αλλά που είναι αρκετά λογική. Αυτή η υπόθεση είναι ότι το cluster χαρακτηρίζεται με βάσει την πλειοψηφία των δειγμάτων που το απαρτίζουν. Αν είχαμε μία αναλογία καρκινικών/φυσιολογικών δειγμάτων ανά cluster κοντά στο 50:50 σίγουρα ο παραπάνω τρόπος θα ήταν επισφαλής, όμως σε όλα τα δεδομένα μας κάθε cluster είχε μία ξεκάθαρα επικρατή κατηγορία.

Ένας δεύτερος περιορισμός της μελέτης είναι ότι λαμβάνουμε ως δεδομένο ότι τα ΔΕ γονίδια είναι και τα πληροφοριακά γονίδια αγνοώντας εντελώς γονίδια που δεν είναι ΔΕ. Ο περιορισμός αυτός δεν είναι παράλογος και αποτελεί υπόθεση όλων των προσεγγίσεων που αφορούν την ανάλυση μεγάλων δεδομένων για τον προσδιορισμό του προφύλ γονιδιακής έκφρασης (Dubitzy et al, 2007).

Άλλοι τρόποι να γίνει τεστ αξιοπιστίας αφορούν κυρίως συγκριτικές προσεγγίσεις ως προς το κατά πόσον συμφωνούν με την βιβλιογραφία τα εξαχθέντα αποτελέσματα (Campain & Yang, 2010) ή και κατά πόσον συμφωνούν με ένα dataset το οποίο θεωρείται «αληθές» βάσει βιβλιογραφίας (Hong & Breitling , 2008). Ενώ αυτοί οι τρόποι έχουν μία λογική εντούτοις εμπεριέχουν μεγάλο βαθμό αυθαιρεσίας αφού εκ προοιμίου θεωρούν ένα dataset ως αυθεντία ενώ στην πραγματικότητα ενδεχομένως όλα τα dataset να εμπεριέχουν κάποια συστηματικά σφάλματα. Η επιλογή του “dataset αυθεντία” επίσης είναι δύσκολη και ως ένα βαθμό παρακινδυνεύμενη ιδιαίτερα όταν έχουμε να κάνουμε με πολλά dataset με μεγάλη διαφορά στον αριθμό δειγμάτων τους. Η σύγκριση με την βιβλιογραφία ασφαλώς και είναι ένα καλό μέτρο σύγκρισης, και για αυτόν τον λόγο σε αυτή την διπλωματική έγινε λειτουργική ανάλυση και στις 3 προσεγγίσεις , ώστε να επιλεχτεί τελικά εκείνη με την καλύτερη απόδοση βιβλιογραφικά βάσει τους εμπλουτισμού τους στις διάφορες κατηγορίες. Όμως η προσέγγιση συνδυασμού επιβλεπόμενων και μη επιβλεπόμενων προσεγγίσεων αποτελούν έναν καθαρότερο από μαθηματικής άποψης και πιο computation-based τρόπο αξιολόγησης.

Συστηματική σύγκριση όλων των παραπάνω προσεγγίσεων σε μεγαλύτερη έκταση αποτελούν προφανώς ένα ερευνητικό πεδίο που ξεφεύγει μακράν των στόχων αυτής της διπλωματικής, και ίσως αποτελέσουν αντικείμενο μίας άλλης (ογκώδους) διπλωματικής εργασίας ή κάποιας διδακτορικής διατριβής.

Τα αποτελέσματα από το moderated t-test και από το απλό Student's t-test ήταν ακριβώς τα ίδια. Το moderated t-test βασίζεται σε μία Bayesian προσέγγιση στο t-test με την χρήση υπερπαραμέτρων που ισορροπούν την αβεβαιότητα της ικανοποίησης της προϋπόθεσης για κανονική κατανομή με όμοια διασπορά των δεδομένων, που συχνά προκύπτει λόγω ανεπαρκούς αριθμού δειγμάτων (Dunkler, Sánchez-Cabo & Heinze, 2011 ; Demissie, Mascialino, Calza & Pawitan, 2008 ; Sartor, Tomlinson, Wesselkamper, Sivaganesan, Leikauf & Medvedovic, 2006). Η προσέγγιση αυτή εύρεσης ΔΕ γονιδίων είναι πολύ δημοφιλής σε μελέτες με μικροσυστοιχίες λόγω ακριβώς του γεγονότος ότι κατά κανόνα

υπάρχει μικρός αριθμός δειγμάτων. Σε μελέτες με μεγάλο αριθμό δειγμάτων όμως δεν φαίνεται να υπάρχει διαφορά ανάμεσα στις 2 στατιστικές μεθόδους πράγμα που επιβεβαιώνεται και από τα αποτελέσματα αυτής της διπλωματικής εργασίας.

Ένα άλλο ενδιαφέρον στοιχείο είναι η φτωχή απόδοση της μεθόδου SAM στην εύρεση ΔΕ γονιδίων. Φαίνεται ότι αυτή η μέθοδος δίνει πολλά ψευδώς θετικά ΔΕ γονίδια γιατί παρόλο που μειώσαμε το threshold για FDR στο 0.01 η μέθοδος βγάζει πολλαπλάσια γονίδια από τις μεθόδους κατηγορίας t-test ως διαφορικά εκφραζόμενα και μάλιστα αν και πολλαπλάσια αποδίδουν φτωχά στην αξιολόγηση με τον συνδυασμό επιβλεπόμενης και μη επιβλεπόμενης μεθόδου. Αυτό επιβεβαιώνει την μη καταλληλότητα αυτής της μεθόδου τουλάχιστον για τα δεδομένα που χρησιμοποιήθηκαν σε αυτή την διπλωματική. Υπάρχουν άρθρα στην βιβλιογραφία που επιβεβαιώνουν αυτή την ανεπάρκεια (Larsson, Wahlestedt & Timmons, 2005 ; Zhang, 2007) και κάποιοι προτείνουν προσαρμογές του ώστε να μπορεί να είναι πιο ασφαλής η χρήση του σε ανομοιογενή δεδομένα όπως αυτά που χρησιμοποιούνται σε μία μετα-ανάλυση (Tzeng, 2021). Βέβαια η ανομοιογένεια είναι όρος σχετικός και ασαφής και εφόσον πρόκειται για μετα-ανάλυση στην οποία υπάρχουν εγγενείς ανομοιογένειες όπως π.χ. η HPV θετικότητα ενδεχομένως αυτό να έπαιξε ρόλο στην μη καταλληλότητα αυτής της τεχνικής εύρεσης ΔΕ γονιδίων.

Στην λειτουργική ανάλυση τα αποτελέσματα σε αυτή την διπλωματική εργασία συμφωνούν σε πολύ μεγάλο βαθμό με αποτελέσματα άλλων ερευνών που χρησιμοποίησαν μικροσυστοιχίες για να διερευνήσουν το γονιδιακό προφίλ του ΑΚΚ του στόματος (Li, Wang, Xu, Wang, Guo & Guo, 2020 ; Zou, Wang, Xu, Yuan, Meng & Zhang, 2019).

Τα αποτελέσματα έβγαλαν στατιστικά σημαντική υπερεκπροσώπηση γονιδίων που ανήκουν στην λειτουργική κατηγορία «Extracellular Matrix – Receptor interaction». Πρόκειται για μια σημαντική λειτουργική κατηγορία ευρέως φάσματος (“umbrella term”) που έχει άμεση συσχέτιση με αυξημένη δραστηριότητα διάφορων σηματοδοτικών βιολογικών μονοπατιών μέσα στο κύτταρο. Οι ιντεγκρίνες, τις οποίες εξετάσαμε στην ενότητα διασύνδεσης επιθηλίου – συνδετικού ιστού, είναι τα βασικά μέλη που ανήκουν σε αυτή την λειτουργική κατηγορία (Sainio & Järveläinen, 2020). Η παραπάνω κατηγορία είναι αρκετά συχνά συσχετιζόμενη με τις κακοήθεις νεοπλασίες. Καρκινώματα σε άλλες περιοχές κυρίως ενδοδερμικής προέλευσης όπως στον πνεύμονα, διάφορες περιοχές του γαστρεντερικού βλεννογόνου και στον μαστό έχουν επίσης αυξημένη λειτουργική συσχέτιση με αυτή την

κατηγορία ενδοκυτταρικών μονοπατιών (Meucci, Keilholz, Heim, Klauschen & Cacciatore, 2018 ; Bao, Wang, Shi, Yun, Liu, Chen, Chen, Ren & Jia, 2019 ; Yeh, Tzeng, Fu, You, Chang, Ger & Tsai, 2018). Αρκετοί ερευνητές έχουν επίσης επιβεβαιώσει την αυξημένη συσχέτιση AKK κεφαλής και τραχήλου με την παραπάνω κατηγορία (Zou et al, 2019 ; Xu., Li, Hu, Chen, Yu,Dong,Sun & Han, 2018).

Μία σχετικά αναπάντεχη κατηγορία με μεγάλη μάλιστα στατιστική σημαντικότητα που εμφανίζεται στα αποτελέσματα είναι η αυξημένη έκφραση γονιδίων που ενεργοποιούνται κατά την αμοιβάδωση, μία λοίμωξη που οφείλεται στο πρωτόζωο *E. histolytica*. Μικρότερου εύρους από την δική μας μετα-αναλύσεις δεδομένων γονιδιακής έκφρασης από μικροσυστοιχίες έχουν επίσης διαπιστώσει μεγάλο εμπλούτισμό αυτής της κατηγορίας στο AKK (Wang,Fan & Zheng, 2018). Έχει αναφερθεί στην βιβλιογραφία περίπτωση με αμοιβαδικό απόστημα σε ασθενή με υποτροπιάζοντα όγκο ακανθοκυτταρικού καρκινώματος του στόματος (Cacheux, Servois, Paulmier, Girodet, Farkhondeh & Petras, 2011) προφανώς όμως αυτό δεν αποδεικνύει τίποτα και το πιο πιθανό είναι ότι πρόκειται για σύμπτωση. Υπάρχουν κάποια case reports για βλάβες του γεννητικού βλεννογόνου οφειλόμενες σε αμοιβάδωση που προσομοιάζουν κλινικά με την εικόνα του AKK (Ríos-Burgueño, Velarde-Félix & García, 2017 ; Hejase, Bahrle, Castillo & Coogan, 1996 ; Arroyo & Elgueta, 1989). Επίσης η αμοιβάδωση ως λοίμωξη έχει μεγάλη τάση διασποράς στον οργανισμό και μπορεί να κάνει αποστήματα σε απομακρυσμένες περιοχές από την πρωτογενή λοίμωξη όπως π.χ. στον εγκέφαλο (Skappak, Akierman, Belga, Novak, Chadee, Urbanski,... & Beck, 2014 ; Ohnishi, Murata, Kojima, Takemura, Tsuchida & Tachibana, 1994). Όλα τα παραπάνω μας οδηγούν ίσως σε μία υποψία ότι η αμοιβάδα να χρησιμοποιεί τους ίδιους μηχανισμούς διασποράς με το AKK, όμως είναι εξαιρετικά απίθανο να υπάρχει σχέση αιτίου και αιτιατού ανάμεσα στις 2 αυτές παθολογικές οντότητες.

Η ιντερλευκίνη-17 (IL-17), είναι μία προφλεγμονώδης κυτοκίνη και είναι το κυριότερο μέλος της οικογένειας των σχετιζόμενων με την IL-17 σηματοδοτική οδό. Η υπερέκφραση IL-17 έχει διαπιστωθεί στην καντιντίαση του στόματος (Conti, Whibley, Coleman, Garg, Jaycox & Gaffen 2015). Στο ακανθοκυτταρικό καρκίνωμα του δέρματος αυξημένη έκφραση IL-17 συνδέεται με αυξημένη παρουσία γδ Τ-λεμφοκυττάρων στο μικροπεριβάλλον του όγκου και η παρουσία τους είναι εντονότερη σε καρκίνους προχωρημένου σταδίου (Lo Presti, Toia, Oieni, Buccheri, Turdo, Mangiapane...& Dieli, 2017). Επίσης σε άλλους

καρκίνους όπως του μαστού, η αυξημένη έκφραση της IL-17 συνδέεται με μειωμένη πρόγνωση και αυξημένη πιθανότητα λεμφαδενικής ή απομακρυσμένης μετάστασης (Song, Wei & Li, 2021). Τέλος η IL-17 ενεργοποιεί μέλη που ανήκουν σε σχετιζόμενα με τον καρκίνο μονοπάτια όπως το STAT6, το NOTCH1 αλλά και τον EGFR (Brevi, Cogrossi, Grazia, Masciovecchio, Impellizzieri, Lacanfora, Grioni & Bellone, 2020). Στα αποτελέσματα αυτής της διπλωματικής εργασίας η λειτουργική αυτή κατηγορία είναι σημαντικά εμπλουτισμένη και επομένως επιβεβαιώνει έναν σημαντικό ρόλο της IL-17 στην διαμόρφωση ενός ευνοϊκού για τα καρκινικά κύτταρα μικροπεριβάλλοντος.

Το βιολογικό μονοπάτι εστιακής συγκόλλησης (focal adhesion) αποτελεί έναν φυσιολογικό μηχανισμό αλληλεπίδρασης των ινιδίων ακτίνης του κυτταροσκελετού με την εξωκυττάρια μήτρα, η αυξημένη έκφραση γονιδίων εστιακής συγκόλλησης έχει συσχετιστεί με το AKK του στόματος αλλά και με άλλους καρκίνους. Συνδέεται άμεσα με την λειτουργική κατηγορία «ECM-receptor interactions» που αναφέραμε παραπάνω και αποτελεί λειτουργική συνέχεια της (Kotb, Hyndman & Patel, 2018). Η αυξημένη έκφραση και ιδιαίτερα η φωσφορυλιωμένη (ενεργή) μορφή της κινάσης εστιακής συγκόλλησης (FAK) έχει συσχετιστεί ανοσοϊστοχημικά με αυξημένη τάση του AKK για λεμφαδενική και απομακρυσμένη μετάσταση και συνδέεται άμεσα με την πρόγνωση του όγκου (Kato, Kato, Miyazawa, Kobayashi, Noguchi & Kawashiri, 2020 , de Vicente, Rosado, Lequerica-Fernández, Allonca, Villallaín & Hernández-Vallejo, 2013). Αυτό συμβαίνει μέσω σύνδεσης της FAK με την RIP (receptor interacting protein) με αποτέλεσμα η τελευταία να μην μπορεί να συνδράμει στην δημιουργία του DISC(death inducing signaling complex). Το τελευταίο είναι ένα σημαντικό σύμπλοκο που προκαλεί μία μορφή προγραμματισμένου κυτταρικού θανάτου (Anoikis) η οποία ενεργοποιείται ως αποτέλεσμα της αποκόλλησης των κυττάρων από την εξωκυττάρια μήτρα (Bunek, Kamarajan & Kapila, 2011). Μία άλλη πρωτεΐνη αυτού του βιολογικού μονοπατιού η ζυξίνη (zyxin) έχει αναφερθεί ότι επηρεάζεται από την πρωτεΐνη E6 του HPV-6 σε καρκίνους του ουρογεννητικού συστήματος. Πιο συγκεκριμένα η E6 συνδέεται με την ζυξίνη και εν συνεχείᾳ αυτό προκαλεί μετανάστευση της στον πυρήνα όπου οδηγεί σε μεταγραφική αύξηση της έκφρασης γονιδίων που σχετίζονται με τον κυτταρικό πολλαπλασιασμό (Kotb, Hyndman & Patel, 2018).

Αρκετά εμπλουτισμένο βρέθηκε στα αποτελέσματα μας και το μονοπάτι AGE/RAGE. Το AGE/RAGE (Advaced glycation end products / Receptor AGE) έχει σχέση με την παρουσία

ελεύθερων σακχάρων και συνδέεται με άλλα γνωστά σχετιζόμενα με τον καρκίνο βιολογικά μονοπάτια όπως το RAS/MEK/ERK και το PI3/Akt/mTOR αλλά και με την απόπτωση αφού έχει δειχθεί ότι το συγκεκριμένο μονοπάτι επηρεάζει αρνητικά την έκφραση των προαποπτωτικών κασπασών 3 και 9 και ακολούθως προκαλώντας αύξηση έκφρασης της αντι-αποπτωτικής πρωτεΐνης BCL-2 οδηγεί τελικά στην αναστολή των μηχανισμών απόπτωσης. Αυξημένη δραστηριότητα του AGE/RAGE είναι συχνό εύρημα σε πολλούς καρκίνους και το συγκεκριμένο μονοπάτι είναι στόχος νεότερων στοχευμένων αντικαρκινικών θεραπειών όπως η courcumin για το ρινοφαρυγγικό καρκίνωμα και η withaferin A για το AKK κεφαλής και τραχήλου (Waghela, Vaidya, Ranjan, Chhipa, Tiwari & Pathak, 2021 ; Park, Min, Kim & Kwon, 2015). Η κατηγορία AGE/RAGE, έχει βρεθεί εμπλουτισμένη και σε αποτελέσματα ερευνών σε AKK του οισοφάγου (Chen, Chu, Li, Shi, Xu, Gu, ... & Sun, 2020 ; Xue, Jia, Ren & Xin, 2021). Φαίνεται ότι το παραπάνω βιολογικό μονοπάτι ενεργοποιείται μέσω αλληλεπίδρασης μη κωδικών RNA (lncRNA & miRNA), και η αυξημένη εκπροσώπηση του είναι εμφανής και σε άλλους καρκίνους όπως του πνεύμονα και του στομάχου στους οποίους έχει αναφερθεί υπερέκφραση των γονιδίων αυτής της σηματοδοτικής οδού (Liao, Cao, Zhang, Zhou, Xu, Zhao, ... & Wu, 2021 ; Li, Xu, Jing & Li, 2021).

Η βιταμίνη Α θεωρείται ότι έχει αντιοξειδωτική δράση και συμβάλλει σημαντικά στην πρόληψη του καρκίνου γενικότερα (Chen, He & Zhou, 2019). Υπάρχουν αρκετές μελέτες που επιβεβαιώνουν ότι η έλλειψη βιταμίνης Α κάνει τον οργανισμό πιο ευάλωτο στην δημιουργία διαφόρων κακοήθων νεοπλασιών όπως του προστάτη (Cao, Meng, Li, Xin, Ben, Cheng, ... & Cheng, 2020) και του παγκρέατος (Huang, Gao, Zhi, Ta, Jiang & Zheng, 2016). Στο AKK η ανεπάρκεια βιταμίνης Α ήταν γνωστή από πολύ παλιά και κάποια κλασικά εγχειρίδια στοματολογίας την κατατάσσουν ως έναν παράγοντα κινδύνου (Neville et al, 2008). Η αντινεοπλασματική της δράση στο AKK θεωρείται ότι επιτυγχάνεται μέσω διαφορετικών βιοχημικών μηχανισμών. Φαίνεται ότι επιτυγχάνει μείωση της PDL1 λόγω αναστολής από την βιταμίνη Α της σηματοδοτικής οδού JAK/STAT και πιο συγκεκριμένα της STAT3 και JAK2. Μείωση της ERK επίσης αναφέρεται στην ίδια μελέτη (Chen, He, Zhou, 2019). Σε άλλη μελέτη αναφέρεται ότι προκαλεί αύξηση των μεταγραφικών παραγόντων FOXO1, FOXO3A και των υποδοχέων TRAIL και επακόλουθη μείωση της VEGF-A, AURKB και του μεταγραφικού παράγοντα FOXM1. Ο FOXM1 ενεργοποιεί την

Wnt-σηματοδοτική οδό της οποία η δράση προκαλεί μετανάστευση της β-κατενίνης στον πυρήνα όπου δρα ως μεταγραφικός παράγοντας για πολλά ογκογονίδια (Osei-Sarfo & Gudas, 2019), όμως έχει αναφερθεί ότι προκαλεί και μία επιπρόσθετη επιγενετική επίδραση στην δημιουργία του AKK (Teh, Gemenetidis, Patel, Tariq, Nadir, Bahta,... & Hutchison, 2012). Στην διπλωματική αυτή εργασία πράγματι διαφαίνεται ότι υπάρχει εμπλουτισμός σε υποεκφραζόμενα γονίδια του προστατευτικού αυτού μηχανισμού και επομένως και εδώ έχουμε συμφωνία με την βιβλιογραφία.

Από τα αποτελέσματα μας 2 λειτουργικές κατηγορίες (Viral protein interaction with Cytokine and cytokine receptor και HPV infection) έχουν άμεση σχέση με την παθογενετική σχέση του AKK του στόματος με τον ιό HPV. Ο ιός HPV είναι η κυριότερη αιτία καρκίνου της μήτρας και αυτό είναι γνωστό εδώ και δεκαετίες (Munoz, Castellsagué, González & Gissmann, 2006). Για τον σκοπό αυτό έχει αναπτυχθεί και το εμβόλιο έναντι του HPV το οποίο είναι πλέον πολύτιμος και αποτελεσματικός σύμμαχος στην πρόληψη του καρκίνου της μήτρας (Chrysostomou, Stylianou, Constantinidou & Kostrikis, 2018 ; Lei, Ploner, Elfström, Wang, Roth, Fang, ... & Sparén, 2020). Η παρουσία HPV και ιδιαίτερα συγκεκριμένων στελεχών όπως το HPV-16 (D'Souza & Dempsey, 2011) έχει πλέον αποδειχθεί ότι ευθύνεται για ένα μεγάλο μέρος της αύξησης του επιπολασμού του καρκίνου του στόματος σε νεότερες σε σχέση με το παρελθόν ηλικίες τα τελευταία χρόνια. Όμως το θετικό είναι ότι τα AKK που συνδέονται παθογενετικά με τους ιούς HPV έχουν καλύτερη πρόγνωση και ανταποκρίνονται καλύτερα στις σύγχρονες θεραπείες σε σχέση με τα HPV αρνητικά AKK (Economopoulou, Kotsantis & Psyrrī, 2020 ; Santacroce, Cosola, Bottalico, Topi, Charitos, Ballini, ... & Dipalma, 2021). Ο παθογενετικός μηχανισμός που δρα αυτός ο ιός είναι σε μεγάλο βαθμό γνωστός και περιλαμβάνει σε γενικές γραμμές την απενεργοποίηση των ογκοκατασταλτικών γονιδίων p53 και pRb από τα προϊόντα των ογκογόνων ιογενών γονιδίων E6 και E7, με αποτέλεσμα να προκαλείται γενετική αστάθεια. Ο εμβολιασμός έναντι του HPV τόσο σε κορίτσια όσο και σε αγόρια προβλέπεται ότι θα συμβάλλει σε μείωση στον επιπολασμό του AKK στις αναπτυγμένες χώρες, ιδιαίτερα στις νεότερες ηλικίες (Näslund, Du & Dalianis, 2020 ; Osazuwa-Peters, Boakye, Mohammed, Tobo, Geneus & Schootman, 2017).

Πολλές από τις λειτουργικές κατηγορίες που βρέθηκαν έχουν να κάνουν με το ανοσολογικό σύστημα, ο ρόλος του οποίου στο AKK αναλύθηκε διεξοδικά στο γενικό μέρος. Πιο

συγκεκριμένα οι κατηγορίες Toll-Like Receptor Signaling Pathway, Cytokine-Cytokine receptor interaction, Rheumatoid Arthritis, Chemokine Signaling Pathway, IL-17 signaling pathway. Σε όλα τα γονίδια που ανήκουν στις παραπάνω κατηγορίες είχαμε υπερέκφραση και τα δεδομένα αυτά συμφωνούν με την βιβλιογραφία (Costa, Valadares, Souza, Mendonça, Oliveira, Silva & Batista, 2013 ; Arantes, Costa, Mendonça, Silva & Batista, 2016).

Όσον αφορά τα αποτελέσματα από τα δέντρα απόφασης, βλέπουμε ότι και στους 2 αλγορίθμους την κυριότερη σημασία την έχει η υπερέκφραση της MMP1. Η έκκριση μεταλλοπρωτεΐνασών μήτρας είναι ένας από τους βασικούς μηχανισμούς διήθησης των καρκινικών κυττάρων στον υποκείμενο συνδετικό ιστό και κατά καιρούς έχει αναφερθεί ότι υπάρχει αυξημένη έκφραση αρκετών MMP στα AKK (Miguel, Mello, Melo & Rivero, 2020 ; George, Ranganathan & Rao, 2010). Ωστόσο στην βιβλιογραφία έχει αναδειχθεί ο ρόλος των MMP-2 και MMP-9 ως πιο κρίσιμος για την παθογένεση του AKK του στόματος (Wang, Zheng, Pang, Zhang, Yu, Wu, ... & Liang, 2019). Μία μετα-ανάλυση των μεταλλοπρωτεΐνασών μήτρας σε διάφορα είδη καρκίνου αναφέρει ότι μεταλλάξεις στην περιοχή του υποκινητή των MMP1 και MMP7 έχουν συσχετιστεί με αυξημένη επιθετικότητα του όγκου ενώ από την άλλη μεριά οι μεταλλάξεις στην περιοχή υποκινητών των MMP2 και MMP9, δηλαδή η μειωμένη έκφραση τους, ενδεχομένως να δρουν ανασταλτικά στην επιθετική πορεία του. (Li, Qu, Zhong, Zhao, Chen & Daru, 2013). Στα αποτελέσματα μας φαίνεται ότι η αυξημένη έκφραση του MMP1 θέτει με μεγαλύτερη αξιοπιστία την διάγνωση καρκίνου σε σχέση με τις MMP2 και MMP9, παρόλα αυτά το MMP9 βρέθηκε στην μελέτη μας να έχει σημασία από πλευράς τοπολογίας δικτύου, και αυτό δείχνει τον κομβικό ρόλο και αυτής της μεταλλοπρωτεΐνασης στην βιοπαθολογία του AKK.

Στον ταξινομητή που προκύπτει από τον αλγόριθμο CART παρατηρούμε ότι η αυξημένη έκφραση της PSMB2 επίσης παίζει μεγάλο ρόλο στην τελική διάγνωση. Αυτή είναι μία πρωτεΐνη που είναι βασικό δομικό συστατικό του πρωτεοσώματος του οποίου ο ρόλος είναι η αποδόμηση των ενδοκυτταρικών πρωτεΐνων. Όμως αυτό που δεν ήταν αναμενόμενο είναι ότι ενώ σε άλλα είδη καρκίνων το σύνηθες είναι να υπάρχει μειωμένη έκφραση του PSMB2 και η μειωμένη παραγωγή πρωτεοσωμάτων (Banno, Garcia, van Baarsel, Metz, Fisch, Widjaja, ... & Chang, 2016); στο δένδρο απόφασης CART η αυξημένη έκφραση του συγκεκριμένου γονιδίου οδηγεί σε κανόνα που γνωμοδοτεί AKK. Παρόλα αυτά στο AKK του στόματος φαίνεται ότι η αναστολή της λειτουργίας του πρωτεοσώματος οδηγεί σε

απόπτωση των καρκινικών κυττάρων μέσω πρωτεϊνών της οικογένειας Cip/Kip που αναφέρθηκαν στο γενικό μέρος και επομένως οι κανόνες του δένδρου απόφασης συμφωνούν και εδώ με την βιβλιογραφία με την παρατήρηση ίσως ότι η σημασία του πρωτεοσώματος στο AKK δεν εκτιμάται τόσο πολύ από την συνολική βιβλιογραφία και υπάρχουν λίγες σχετικά μελέτες πάνω στον ρόλο του συγκεκριμένου στοιχείου στο AKK (Kudo, Takata, Ogawa, Kaneda, Sato, Takekoshi, ... & Nikai, 2000 ; Yoshiha, Iwase, Kurihara, Uchida, Kurihara, Watanabe & Shintani, 2011).

Στον ταξινομητή που προκύπτει από τον αλγόριθμο C-tree φαίνεται η σημασία της Τενασκίνης C. Πρόκειται για μια πρωτεΐνη που παίζει ρόλο στην εμβρυογένεση αλλά της οποίας ο ρόλος στην καρκινογένεση μόλις τα τελευταία χρόνια αρχίζει να διερευνάται όμως φαίνεται ότι παίζει ρόλο σε πολλά είδη καρκίνου. Η συγκεκριμένη πρωτεΐνη προάγει την μετάσταση μέσω ανοσοκατασταλτικής δράσης της στα κύτταρα της μυελοειδούς σειράς. (Spenlé, Loustau, Murdamoothoo, Erne, Beghelli-de la Forest , Veber., ... & Orend, 2020). Η παρουσία της τόσο στο AKK όσο και σε προκαρκινικές βλάβες εχει επιβεβαιωθεί και ανοσοϊστοχημικά (Mane, Kale & Belaldavar, 2017). Μένει να αποδειχθεί εάν η παρουσία της σε προκαρκινικές βλάβες αυξάνει την πιθανότητα ανάπτυξης καρκίνου. Αυτό μέχρι σήμερα είναι άγνωστο.

Η Ras related GTP binding D (RRAGD) έχει ρόλο στην απενεργοποίηση του γνωστού στην καρκινογένεση μονοπατιού mTOR (Oshiro, Rapley, & Avruch, 2014). Στο δένδρο απόφασης που συντάξαμε φαίνεται ότι η μειωμένη έκφραση σε αυτή την πρωτεΐνη αυξάνει την πιθανότητα η τελική διάγνωση να είναι AKK και αυτό το εύρημα συμφωνεί με την βιβλιογραφία. Είναι ενδιαφέρον το γεγονός ότι η συγκεκριμένη πρωτεΐνη απενεργοποιείται μέσω μεθυλίωσης στα AKK του οισοφάγου ; από την άλλη άλλα ιστολογικά είδη καρκίνου του οισοφάγου όπως το αδενοκαρκίνωμα δεν εμφανίζουν απενεργοποίηση αυτής της πρωτεΐνης. Φαίνεται λοιπόν ότι η υπερέκφραση της αποτελεί χαρακτηριστικό γνώρισμα ενός ακανθοκυτταρικού καρκινικού υπότυπου (Jin, Z., Feng, X., Jian, Q., Cheng, Y., Gao, Y., Zhang ,... & Meltzer, 2013). Η σίγαση της RRAGD έχει επίσης αναφερθεί σε καρκίνους του πνεύμονα και του μαστού (Suzuki, Shigematsu, Shames, Sunaga, Takahashi, Shivapurkar & Gazdar, 2007).

Ο παράγοντας του συμπληρώματος D είναι ένας παράγοντας της εναλλακτικής οδού του συμπληρώματος και ενισχύει την ανοσολογική απόδοση των B και T λεμφοκυττάρων. Στα

αποτελέσματα μας η αυξημένη παρουσία του συμπληρώματος D μειώνει την πιθανότητα ο υπό εξέταση ιστός να είναι AKK που επίσης συμφωνεί με την βιβλιογραφία (Hu, Sun, Shi, Ni & Jiang, 2020). Η Inhibin B και το ομοδιμερές της, η Activin, είναι πρωτεΐνες που αποτελούν μέλη της οικογένειας του TGF-b και θεωρούνται ότι παίζουν ρόλο στην EMT. Αναφέρεται στην βιβλιογραφία αυξημένη έκφραση της Inhibin B και της Activin στα AKK κάτι που φαίνεται και στο δένδρο απόφασης C-tree (Kita, Kasamatsu, Nakashima, Endo-Sakamoto, Ishida, Shimizu , ... & Uzawa, 2017). Τέλος η MICAL2 που εμφανίζεται χαμηλά στο ίδιο δέντρο απόφασης είναι μία πρωτεΐνη που μόλις τα τελευταία 10 χρόνια εμφανίζεται στην βιβλιογραφία να σχετίζεται με τον καρκίνο (Mariotti, Barravecchia, Vindigni, Pucci, Balsamo, Libro, ... & Angeloni, 2016) και παίζει ρόλο στην MET. Μέχρι σήμερα η συγκεκριμένη πρωτεΐνη δεν έχει μελετηθεί στο AKK του στόματος και ίσως αποτελεί καλό μελλοντικό ερευνητικό πεδίο για πληρέστερη κατανόηση της MET σε αυτό.

Κατά την ανάλυση δικτύων αναδείχθηκαν 3 bottleneck hub γονίδια τα οποία υπερεκφράζονται σημαντικά περισσότερο από τα υπόλοιπα γονίδια που επίσης έχουν κομβικό ρόλο και αυτά είναι το CXCL10, το CXCL1 και το MMP9. Για τον ρόλο των μεταλλοπρωτεΐνασών μήτρας αναφέρθηκαμε και παραπάνω. Η CXCL10 είναι η bottleneck πρωτεΐνη με την μεγαλύτερη ΔΕ στο AKK, όμως ταυτόχρονα και η CXCL1 παρουσιάζεται πολύ σημαντική από πλευράς τοπολογίας δικτύων. Πρόκειται για κυτοκίνες που φαίνεται να έχουν συντονιστικό ρόλο στο ανοσοποιητικό σύστημα. Τα τελευταία χρόνια έχουν αναδειχθεί σε σημαντικούς παράγοντες που παίζουν κεντρικό ρόλο στην παθογένεση διαφόρων ειδών καρκίνου.

Η CXCL1 θεωρείται ότι αλληλεπιδρά στο AKK του στόματος με τους ινοβλάστες σχετιζόμενους με τον καρκίνο (CAFs) συμβάλλοντας και αυξάνοντας την έκκριση MMP-1 οδηγώντας σε καρκίνους με κακή πρόγνωση (Wei, Lee, Yeh, Yang, Kok, Ko & Chia, 2019). Επιπλέον έχει αναφερθεί υπερέκφραση της σε πολλά είδη καρκίνου όπως στον πνεύμονα, στο μαστό και στο παχύ έντερο.(Łukaszewicz-Zajac, Paćzek, Mroczko, & Kulczyńska-Przybik, 2020). Η CXCL10 παίζει ρόλο και σε αυτοάνοσα νοσήματα (Karin & Razon, 2018). Πρόσφατα η συγκεκριμένη ουσία χαρακτηρίστηκε ως ένα μέλος μίας ομάδας 6 γονιδίων που αποτελούν γονιδιακή υπογραφή πρόγνωσης στο AKK (Wang, Wang, Kong, Han, Song, Chen, Bu, Wang, Yue & Ma, 2020) και κάποιοι έχουν συμπεράνει σε πρόσφατες μελέτες ότι ίσως εμπλέκεται σε αυξημένο κίνδυνο AKK σε ασθενείς που πάσχουν από

περιοδοντίτιδα (Geng, Wang, Li, Liu, Zhang, Zhang & Pan, 2019 ; Mitsuwan & Nissapatorn, 2020). Η παρουσία της CXCL10 και ο ρόλος της στο AKK και ίσως και σε αυτοάνοσα νοσήματα του στόματος που προδιαθέτουν σε AKK όπως ο ομαλός λειχήνας πρέπει να διερευνηθούν περαιτέρω με ανοσοϊστοχημικές μεθόδους και ίσως και με προσπάθειες σχεδιασμού στοχευμένων θεραπευτικών προσεγγίσεων.

Το STAT1 έχει έναν σχετικά αβέβαιο ρόλο στον καρκίνο και δεν είναι ξεκάθαρο αν πρόκειται για ογκογονίδιο ή ογκοκατασταλτικό γονίδιο (Zhang & Liu, 2017 ; Meissl, Macho-Maschler, Müller & Strobl, 2017). Άλλα φαίνεται ότι η αυξημένη έκφραση του βελτιώνει την συνολική πενταετή επιβίωση των ασθενών που πάσχουν από AKK. (Zhang, Wang, Liu & Xu, 2020). Ωστόσο στην δική μας εργασία το STAT1 βρέθηκε περισσότερο εντός του AKK παρά στα υγιή χειρουργικά όρια ενώ από την βιβλιογραφία η ανοσοϊστοχημική εξέταση τέτοιων όγκων φαίνεται να δείχνει το αντίθετο (Pappa, Nikitakis, Vlachodimitropoulos, Avgoustidis, Oktseloglou & Papadogeorgakis, 2014).

Η οστεονεκτίνη (SPARC) ένα άλλο γονίδιο που αναδείχτηκε κατά την ανάλυση δικτύων σε αυτή την διπλωματική εργασία εκφράζεται στα επιθηλιακά κύτταρα και παίζει ρόλο την επούλωση των μαλακών ιστών αλλά παίζει και κεντρικό ρόλο στην EMT που αποτελεί βασικό παθοβιολογικό μηχανισμό για το AKK (Kothari, Arffa, Chang, Blackwell, Syn, Zhang, ... & Kuo, 2016). Και η οστεονεκτίνη όπως και αρκετά από τα γονίδια που ήδη αναφέραμε εκφράζεται σε διάφορους καρκίνους αλλά και στο AKK (Aquino, Sabatino, Cantile, Aversa, Ionna, Botti, ... & Longo, 2013). Άλλα bottleneck hub γονίδια στα αποτελέσματα μας έχουν αναφερθεί επίσης ως σημαντικά γονίδια για το AKK (Nagata, Noman, Suzuki, Kurita, Ohnishi, Ohyama ,... & Takagi, 2013 ; Wei et al, 2019).

Στην ανάλυση δικτύων η πρωτεΐνη με τον μεγαλύτερο βαθμό κόμβου είναι η CDK1 η οποία όπως είδαμε στο γενικό μέρος παίζει κεντρικό ρόλο στην προαγωγή της μίτωσης στον κυτταρικό κύκλο. Αυξημένη έκφραση της έχει βρεθεί και σε διάφορα καρκινώματα όπως το μικροκυτταρικό καρκίνωμα του πνεύμονα και προάγει τον ανεξέλεγκτο πολλαπλασιασμό των καρκινικών κυττάρων (Zhao, Wang, Ma, Kuang, Liang & Yuan, 2020). Ο ρόλος της υπερέκφρασης της CDK1 είναι παρόμοιος και στο AKK (Chen, Zhang, Chen, Wang, Wang, He & Zhou, 2015).

Πάντως ένας πιθανός μηχανισμός που θα μπορούσε να προταθεί για τον μηχανισμό διήθησης στο AKK όπως αυτός διαφαίνεται από την τοπολογία δικτύων είναι ότι αρχικά τα

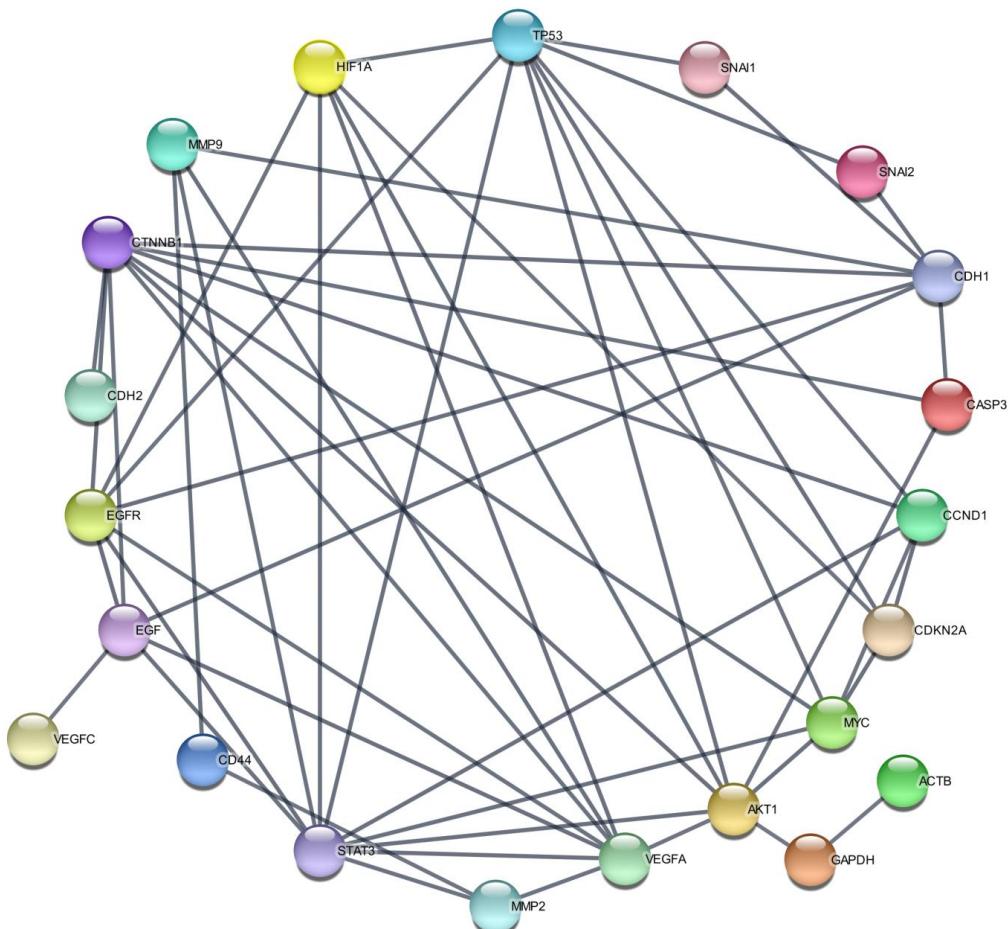
καρκινικά κύτταρα εκκρίνουν μεταλλοπρωτεΐνασες και με αυτόν τον τρόπο διαλύουν την εξωκυττάρια μήτρα ή και τα παρακείμενα φυσιολογικά κύτταρα. Εν συνεχεία ακολουθεί πολλαπλασιασμός των καρκινικών κυττάρων για την αναπλήρωση του δημιουργούμενου κενού χώρου και πρόσδεση τους με τον υποκείμενο συνδετικό ιστό μέσω ημιδεσμοσωμάτων. Ο παραπάνω κύκλος επαναλαμβάνεται συνεχώς και με αυτόν τον τρόπο μπορεί ενδεχομένως να εξηγηθεί ο τρόπος τοπικής επέκτασης του συγκεκριμένου τύπου καρκίνου.

Στην μελέτη του Cancer Genome Atlas φαίνεται ότι το PIK3CA παίζει σημαντικό ρόλο στην βιοπαθολογία του AKK, όμως στην δική μας έρευνα αυτό το γονίδιο δεν είναι ΔΕ (Cancer Genome Atlas, 2015). Επίσης παρόλο την πλούσια βιβλιογραφία σχετικά με τον παθογνωμονικό ρόλο της αυξημένης έκφρασης του p53 στο AKK του στόματος στα αποτελέσματα αυτής της διπλωματικής δεν φαίνεται η έκφραση του να παίζει ρόλο στον διαχωρισμό καρκινικών από φυσιολογικούς ιστούς με κριτήρια γονιδιακής έκφρασης και δεν βρέθηκε ΔΕ (Cutilli, Leocata, Dolo & Altobelli, 2016). Αυτό αποδίδεται στο γεγονός ότι αυξημένη έκφραση p53 έχει βρεθεί αυξημένη στα κύτταρα των χειρουργικών ορίων επί υγιείς ιστοπαθολογικά ιστούς και στην μετα-ανάλυση μας δεχτήκαμε τα υγιή χειρουργικά όρια ως δείγματα φυσιολογικού βλεννογόνου (Kamat, Rai, Puranik & Datar, 2020). Στις περισσότερες μελέτες με μικροσυστοιχίες στο AKK του στόματος συγκρίνεται το γονιδιακό προφίλ του όγκου με τα υγιή ιστολογικά χειρουργικά όρια και επομένως ένα φιλτράρισμα πειραμάτων που έχουν έναν σχεδιασμό που περιλαμβάνει δείγματα από χειρουργικά όρια, θα είχε σημαντικό αντίκτυπο στον αριθμό των δειγμάτων που θα έφταναν στο στάδιο της τελικής ανάλυσης. Το παραπάνω λοιπόν αποτελεί άλλον έναν περιορισμό αυτής της διπλωματικής εργασίας.

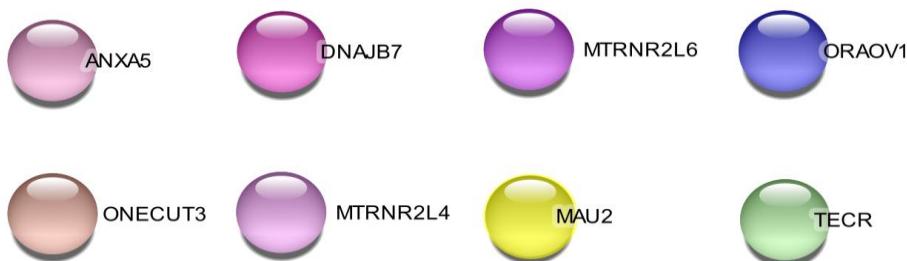
Στην παρούσα εργασία έγινε προσπάθεια διερεύνησης της γονιδιακής έκφρασης στο AKK ανά κλινικό στάδιο με βάση την σύντομη σταδιοποίηση που έχει περιγραφεί στο Γενικό Μέρος. Δυστυχώς τα μεταδεδομένα από το GEO δεν μας επέτρεψαν να κάνουμε μελέτη όλων των δειγμάτων που πέρασαν τα αρχικά κριτήρια που θέσαμε. Μόνο το 1/3 των δεδομένων είχαν στα μεταδεδομένα τους την σταδιοποίηση του κάθε όγκου και επειδή σε πολλά dataset υπήρχε ασαφής προσδιορισμός σταδιοποίησης (όπως π.χ. στάδιο I/II και στάδιο III/IV) αυτό επέβαλε στον σχεδιασμό αυτού του μέρους της διπλωματικής εργασίας τον κλινικό διαχωρισμό των δειγμάτων σε 2 μόνο κλινικά στάδια, το πρώιμο και το προχωρημένο. Επίσης έγινε προσπάθεια να γίνει διερεύνηση της γονιδιακής έκφρασης βάσει

άλλων παραμέτρων πέραν της σταδιοποίησης, όπως η HPV θετικότητα και η ιστολογική σταδιοποίηση των AKK (grading). Δυστυχώς τέτοια δεδομένα υπήρχαν μόνο σε 1-2 datasets και επομένως μία τέτοια μετα-ανάλυση δεν θα είχε νόημα λόγω μειωμένης στατιστικής ισχύος. Αυτό υπογραμμίζει την σημασία της καταγραφής εκτενέστερων μεταδεδομένων και ίσως στο μέλλον στα αποθετήρια να μπορούν να γίνουν και τέτοιες πραγματικά χρήσιμες μετα-αναλύσεις που ενδεχομένως και να εξηγούν πληρέστερα και clusters δειγμάτων όπως αυτά που εντοπίσαμε στην ανάλυση με PCA.

Φαίνεται ότι το κλινικό στάδιο δεν σχετίζεται με το προφίλ γονιδιακής έκφρασης στο AKK του στόματος. Με άλλα λόγια το μέγεθος αλλά και το αν έχει δώσει επιχώρια λεμφαδενική ή/και απομακρυσμένη μετάσταση δεν μπορούν να προβλεφτούν με βάσει το ποιά γονίδια αποκλίνουν στην έκφραση τους. Αυτό έρχεται σε αντίθεση με την βιβλιογραφία στην οποία υπάρχουν αρκετά άρθρα που υποστηρίζουν ότι το εκάστοτε μοτίβο γονιδιακής έκφρασης αποτελεί προγνωστικό βιοδείκτη του AKK (Qadir, Lalli, Dar, Hwang, Aldehlawi, Ma, Dai, Waseem & Teh, 2019 ; Zhao, Xu, Yin, Sun, Shi & Li, 2009 ; Méndez, Lohavanichbutr, Fan, Houck, Rue, Doody, ... & Chen, 2011). Αν και κάποιες έρευνες που χρησιμοποιήθηκαν για αυτή την μετα-ανάλυση επίσης αναφέρουν την δυνατότητα πρόβλεψης μετάστασης του καρκίνου βάσει του γονιδιακού προφίλ (O'Donnell, Kupferman, Wei, Singhal, Weber, O'Malley, Cheng, Putt, Feldman, Ziobor & Muschel, 2005) εν τούτοις λαμβάνοντας υπ'οψη όλα τα datasets αυτό διαψεύδεται εδώ, πράγμα που εμμέσως καταδεικνύει την μεγαλύτερη σημασία των μετα-αναλύσεων στην ιεραρχία της κλινικής πρακτικής βασισμένη σε στοιχεία (evidence-based clinical practice) σε σχέση με τις επιμέρους έρευνες που ενδεχομένως να το αποτελούν. Παρόλα αυτά είναι γεγονός να αναγνωριστεί ότι ανεξάρτητα τι γράφεται στην ακαδημαϊκή βιβλιογραφία στην καθημερινή κλινική πράξη κανένα guideline δεν λαμβάνει υπ'οψη του την γονιδιακή έκφραση κάποιου γονιδίου για να προβλέψει την κλινική σταδιοποίηση στο AKK. Σύμφωνα με τα ευρήματα σε αυτή την διπλωματική μάλλον είναι απίθανο στο κοντινό μέλλον να υπάρξει κάποια τέτοια προσέγγιση με ικανοποιητική απόδοση στην εφαρμοσμένη κλινική πράξη. Σε επίπεδο γονιδίων τα συχνότερα αναφερόμενα για το AKK του στόματος στην βιβλιογραφία γονίδια αναπαριστώνται στούς γράφουνς Cytoscape στις Εικόνες 71 και 72.



Εικόνα 71. Γράφος με τα συχνότερα γονίδια στην βιβλιογραφία, οι ακμές αναπαριστούν τις λειτουργικές τους σχέσεις σε επίπεδο πρωτεΐνης



Εικόνα 72. Γονίδια που αναφέρονται συχνά στην βιβλιογραφία αλλά δεν έχουν κάποια λειτουργική σχέση μεταξύ τους ή με τα γονίδια της προηγούμενης εικόνας

Κεφάλαιο 5. Συμπεράσματα

Εν κατακλείδει, από αυτή την διπλωματική εργασία και λαμβάνοντας υπ' όψη τις επιφυλάξεις που εκπορεύονται από τους περιορισμούς στον σχεδιασμό της, βγαίνουν τα παρακάτω συμπεράσματα:

1. Η διασπαστική προσέγγιση μετα-ανάλυσης μελετών με μικροσυστοιχίες της Affymetrix δίνει παρόμοια ή/και καλύτερα αποτελέσματα σε σύγκριση με άλλες τεχνικές που βασίζονται στην ζ-κανονικοποίηση.
2. Το απλό Student's t-test δίνει ακριβώς τα ίδια αποτελέσματα με το moderated t-test όταν υπάρχει σχετικά μεγάλος αριθμός δειγμάτων.
3. Η μέθοδος SAM φαίνεται να είναι λιγότερο αξιόπιστη σε μετα-αναλύσεις σε σχέση με μεθόδους βασισμένες στο t-test, λόγω του μεγάλου αριθμού ψευδώς θετικών ευρημάτων.
4. Υπάρχει σαφής διαχωρισμός των καρκινικών από τους φυσιολογικούς ιστούς όσον αφορά την γονιδιακή έκφραση τους.
5. Το γονιδιακό προφίλ δεν μπορεί να καθορίσει την κλινική σταδιοποίηση του AKK, δηλαδή το μέγεθος του όγκου και την ύπαρξη τυχόν λεμφαδενικών ή/και απομακρυσμένων μεταστάσεων.
6. Η πρωτεΐνη MMP-1 παίζει τον κυριότερο ρόλο στον διαχωρισμό AKK από τους φυσιολογικούς ιστούς.
7. Τα CXCL1, CXCL10, MMP9, STAT1 και CDK1 έχουν τον σημαντικότερο ρόλο από πλευράς τοπολογίας δικτύων.

Σε γενικές γραμμές τα αποτελέσματα αυτής της διπλωματικής εργασίας συμφωνούν με τα μέχρι σήμερα υπάρχοντα βιβλιογραφικά δεδομένα.

Βιβλιογραφία

- Aittokallio, T. (2010). Dealing with missing values in large-scale studies: microarray data imputation and beyond. *Briefings in bioinformatics*, 11(2), 253-264.
- Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., Walters, P. (2015) Molecular Biology of the Cell. 6th Edition, Garland Science, Taylor and Francis Group, New York.
- Ali, J., Sabiha, B., Jan, H. U., Haider, S. A., Khan, A. A., & Ali, S. S. (2017). Genetic etiology of oral cancer. *Oral oncology*, 70, 23–28.
<https://doi.org/10.1016/j.oraloncology.2017.05.004>
- Annavarapu, C. S., Dara, S., & Banka, H. (2016). Cancer microarray data feature selection using multi-objective binary particle swarm optimization algorithm. *EXCLI journal*, 15, 460–473. <https://doi.org/10.17179/excli2016-481>
- Aquino, G., Sabatino, R., Cantile, M., Aversa, C., Ionna, F., Botti, G., La Mantia, E., Collina, F., Malzone, G., Pannone, G., Losito, N. S., Franco, R., & Longo, F. (2013). Expression analysis of SPARC/osteonectin in oral squamous cell carcinoma patients: from saliva to surgical specimen. *BioMed research international*, 2013, 736438.
<https://doi.org/10.1155/2013/736438>
- Arantes, D. A. C., Costa, N. L., Mendonça, E. F., Silva, T. A., & Batista, A. C. (2016). Overexpression of immunosuppressive cytokines is associated with poorer clinical stage of oral squamous cell carcinoma. *Archives of oral biology*, 61, 28-35.
- Arroyo G, Elgueta R. Squamous cell carcinoma associated with amoebic cervicitis. Report of a case. Acta Cytol. 1989 May-Jun;33(3):301-4. PMID: 2728784.
- Azuaje, F. (2010). Bioinformatics and biomarker discovery. *Wiley Online Library*, 1, 19.
- Baker, E. A., Leaper, D. J., Hayter, J. P., & Dickenson, A. J. (2006). The matrix metalloproteinase system in oral squamous cell carcinoma. *British Journal of Oral and Maxillofacial Surgery*, 44(6), 482-486.
- Banno, A., Garcia, D. A., van Baarsel, E. D., Metz, P. J., Fisch, K., Widjaja, C. E., Kim, S. H., Lopez, J., Chang, A. N., Geurink, P. P., Florea, B. I., Overkleeft, H. S., Ovaa, H., Bui, J. D., Yang, J., & Chang, J. T. (2016). Downregulation of 26S proteasome catalytic activity promotes epithelial-mesenchymal transition. *Oncotarget*, 7(16), 21527–21541.
<https://doi.org/10.18632/oncotarget.7596>
- Bai, Y., Sha, J., & Kanno, T. (2020). The Role of Carcinogenesis-Related Biomarkers in the Wnt Pathway and Their Effects on Epithelial-Mesenchymal Transition (EMT) in Oral

Squamous Cell Carcinoma. *Cancers*, 12(3), 555.

<https://doi.org/10.3390/cancers12030555>

Bao, Y., Wang, L., Shi, L., Yun, F., Liu, X., Chen, Y., Chen, C., Ren, Y., & Jia, Y. (2019). Transcriptome profiling revealed multiple genes and ECM-receptor interaction pathways that may be associated with breast cancer. *Cellular & molecular biology letters*, 24, 38. <https://doi.org/10.1186/s11658-019-0162-0>

Barata, J & Oliveira, M. (2019). Cell Signalling in Cancer. In Fior, R., & Zilhão, R. (Eds.). *Molecular and cell biology of cancer*. Springer International Publishing. pp 31-49

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. (2013). NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.* 2013 Jan;41(Database issue):D991-5.

Beyer M, Mallmann MR, Xue J, Staratschek-Jox A, Vorholt D, et al. (2012) High-Resolution Transcriptome of Human Macrophages PLOS ONE 7(9):e45466. <https://doi.org/10.1371/journal.pone.0045466>

Bolstad B (2021). *preprocessCore: A collection of pre-processing functions*. R package version 1.52.1, <https://github.com/bmbolstad/preprocessCore>.

Bolstad BM (2004). *Low Level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization*. Ph.D. thesis, University of California, Berkeley.

Bolstad BM, Collin F, Brettschneider J, Simpson K, Cope L, Irizarry RA, Speed TP (2005). “Quality Assessment of Affymetrix GeneChip Data.” In Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S (eds.), *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, chapter 3, 33–47. Springer, New York.

Bolstad. Preprocessing DNA Microarray Data. In Dubitzky, W., Granzow, M., & Berrar, D. P. (Eds.). (2007). *Fundamentals of data mining in genomics and proteomics*. Springer Science & Business Media. pp 51-78

Brettschneider J, Collin F, Bolstad BM, Speed TP (2007). “Quality assessment for short oligonucleotide arrays.” *Technometrics*, In press.

Brevi, A., Cogrossi, L. L., Grazia, G., Masciovecchio, D., Impellizzieri, D., Lacanfora, L., Grioni, M., & Bellone, M. (2020). Much More Than IL-17A: Cytokines of the IL-17 Family Between Microbiota and Cancer. *Frontiers in immunology*, 11, 565470. <https://doi.org/10.3389/fimmu.2020.565470>

Bunek, J., Kamarajan, P., & Kapila, Y. L. (2011). Anoikis mediators in oral squamous cell carcinoma. *Oral diseases*, 17(4), 355–361. <https://doi.org/10.1111/j.1601-0825.2010.01763.x>

Cacheux, W., Servois, V., Paulmier, B., Girodet, J., Farkhondeh, F., & Petras, S. (2011). Amoebic abscess diagnosed on fluorodeoxyglucose positron emission tomography scan in patient with recurrent oropharyngeal squamous cell carcinoma. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 29(13), e365–e366. <https://doi.org/10.1200/JCO.2010.33.3237>

Campain, A., & Yang, Y. H. (2010). Comparison study of microarray meta-analysis methods. *BMC bioinformatics*, 11, 408. <https://doi.org/10.1186/1471-2105-11-408>

Cancer Genome Atlas Network. (2015). Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, 517(7536), 576.

Cao, D., Meng, Y., Li, S., Xin, J., Ben, S., Cheng, Y., Wang, M., Hua, L., & Cheng, G. (2020). Association study between genetic variants in retinol metabolism pathway genes and prostate cancer risk. *Cancer medicine*, 9(24), 9462–9470.
<https://doi.org/10.1002/cam4.3538>

Carlos A. Genomic Instability: DNA repair and cancer. In Fior, R., & Zilhão, R. (Eds). *Molecular and cell biology of cancer*. Springer International Publishing. pp 75-96.

Carvalho BS, Irizarry RA (2010). “A Framework for Oligonucleotide Microarray Preprocessing.” *Bioinformatics*, 26(19), 2363-7. ISSN 1367-4803, doi: 10.1093/bioinformatics/btq431.

Chai, A. W. Y., Lim, K. P., & Cheong, S. C. (2020, April). Translational genomics and recent advances in oral squamous cell carcinoma. In *Seminars in cancer biology* (Vol. 61, pp. 71-83). Academic Press.

Cheadle, C., Cho-Chung, Y. S., Becker, K. G., & Vawter, M. P. (2003). Application of z-score transformation to Affymetrix data. *Applied bioinformatics*, 2(4), 209–217.

Cheadle, C., Vawter, M. P., Freed, W. J., & Becker, K. G. (2003). Analysis of microarray data using Z score transformation. *The Journal of molecular diagnostics : JMD*, 5(2), 73–81. [https://doi.org/10.1016/S1525-1578\(10\)60455-2](https://doi.org/10.1016/S1525-1578(10)60455-2)

Chen, C., Méndez, E., Houck, J., Fan, W., Lohavanichbutr, P., Doody, D., Yueh, B., Futran, N. D., Upton, M., Farwell, D. G., Schwartz, S. M., & Zhao, L. P. (2008). Gene expression profiling identifies genes predictive of oral squamous cell carcinoma. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 17(8), 2152–2162. <https://doi.org/10.1158/1055-9965.EPI-07-2893>

Chen, F., Chu, L., Li, J., Shi, Y., Xu, B., Gu, J., Yao, X., Tian, M., Yang, X., & Sun, X. (2020). Hypoxia induced changes in miRNAs and their target mRNAs in extracellular vesicles of esophageal squamous cancer cells. *Thoracic cancer*, 11(3), 570–580.
<https://doi.org/10.1111/1759-7714.13295>

Chen, H., & Boutros, P. C. (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC bioinformatics*, 12(1), 1-7.

Chen, X. J., He, M. J., & Zhou, G. (2019). All-trans retinoic acid induces anti-tumor effects via STAT3 signaling inhibition in oral squamous cell carcinoma and oral dysplasia. *Journal of oral pathology & medicine : official publication of the International Association of Oral Pathologists and the American Academy of Oral Pathology*, 48(9), 832–839. <https://doi.org/10.1111/jop.12931>

Chen, X., Zhang, F. H., Chen, Q. E., Wang, Y. Y., Wang, Y. L., He, J. C., & Zhou, J. (2015). The clinical significance of CDK1 expression in oral squamous cell carcinoma. *Medicina oral, patología oral y cirugía bucal*, 20(1), e7–e12.
<https://doi.org/10.4317/medoral.19841>

Chiego, D. J., & Chiedo, D. J. (2013). Essentials of oral histology and embryology, a clinical approach. 4th edition, Elsevier. pp.166-183.

Chin, C. H., Chen, S. H., Wu, H. H., Ho, C. W., Ko, M. T., & Lin, C. Y. (2014). cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC systems biology*, 8(4), 1-7.

Chrysostomou, A. C., Stylianou, D. C., Constantinidou, A., & Kostrikis, L. G. (2018). Cervical cancer screening programs in Europe: the transition towards HPV vaccination and population-based HPV testing. *Viruses*, 10(12), 729.

Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: lessons from large-scale biology. *Science*, 300(5617), 286-290.

Conti, H. R., Whibley, N., Coleman, B. M., Garg, A. V., Jaycox, J. R., & Gaffen, S. L. (2015). Signaling through IL-17C/IL-17RE is dispensable for immunity to systemic, oral and cutaneous candidiasis. *PloS one*, 10(4), e0122807.
<https://doi.org/10.1371/journal.pone.0122807>

Costa, N. L., Valadares, M. C., Souza, P. P. C., Mendonça, E. F., Oliveira, J. C., Silva, T. A., & Batista, A. C. (2013). Tumor-associated macrophages and the profile of inflammatory cytokines in oral squamous cell carcinoma. *Oral oncology*, 49(3), 216-223.

Cutilli, T., Leocata, P., Dolo, V., & Altobelli, E. (2016). p53 as a prognostic marker associated with the risk of mortality for oral squamous cell carcinoma. *Oncology letters*, 12(2), 1046–1050. <https://doi.org/10.3892/ol.2016.4742>

- Dang, F., Nie, L., & Wei, W. (2021). Ubiquitin signaling in cell cycle control and tumorigenesis. *Cell death and differentiation*, 28(2), 427–438.
<https://doi.org/10.1038/s41418-020-00648-0>.
- de Vicente, J. C., Fresno, M. F., Villalain, L., Vega, J. A., & Vallejo, G. H. (2005). Expression and clinical significance of matrix metalloproteinase-2 and matrix metalloproteinase-9 in oral squamous cell carcinoma. *Oral oncology*, 41(3), 283-293.
- De Vicente, J. C., Lequerica-Fernández, P., Santamaría, J., & Fresno, M. F. (2007). Expression of MMP-7 and MT1-MMP in oral squamous cell carcinoma as predictive indicator for tumor invasion and prognosis. *Journal of oral pathology & medicine*, 36(7), 415-424.
- de Vicente, J. C., Rosado, P., Lequerica-Fernández, P., Allonca, E., Villallaín, L., & Hernández-Vallejo, G. (2013). Focal adhesion kinase overexpression: correlation with lymph node metastasis and shorter survival in oral squamous cell carcinoma. *Head & neck*, 35(6), 826–830. <https://doi.org/10.1002/hed.23038>
- Demissie, M., Mascialino, B., Calza, S., & Pawitan, Y. (2008). Unequal group variances in microarray data analyses. *Bioinformatics*, 24(9), 1168-1174.
- Doncheva, N. T., Morris, J. H., Gorodkin, J., & Jensen, L. J. (2018). Cytoscape StringApp: network analysis and visualization of proteomics data. *Journal of proteome research*, 18(2), 623-632.
- Dorstyn, L., Akey, C. W., & Kumar, S. (2018). New insights into apoposome structure and function. *Cell death and differentiation*, 25(7), 1194–1208.
<https://doi.org/10.1038/s41418-017-0025-z>
- D'Souza, G., & Dempsey, A. (2011). The role of HPV in head and neck cancer and review of the HPV vaccine. *Preventive medicine*, 53, S5-S11.
- Dubitzky, W., Granzow, M., & Berrar, D. P. (Eds.). (2007). *Fundamentals of data mining in genomics and proteomics*. Springer Science & Business Media.
- Duesberg, P., Li, R., Fabarius, A., & Hehlmann, R. (2005). The chromosomal basis of cancer. *Analytical Cellular Pathology*, 27(5, 6), 293-318.
- Duijf, P. H., Nanayakkara, D., Nones, K., Srihari, S., Kalimutho, M., & Khanna, K. K. (2019). Mechanisms of genomic instability in breast cancer. *Trends in molecular medicine*, 25(7), 595-611.
- Dunkler, D., Sánchez-Cabo, F., & Heinze, G. (2011). Statistical analysis principles for omics data. In *Bioinformatics for Omics Data* (pp. 113-131). Humana Press.

Dziuda, D. M. (2010). *Data mining for genomics and proteomics: analysis of gene and protein expression data* (Vol. 1). John Wiley & Sons.

Economopoulou, P., Kotsantis, I., & Psyrra, A. (2020). Special Issue about Head and Neck Cancers: HPV Positive Cancers. *International journal of molecular sciences*, 21(9), 3388. <https://doi.org/10.3390/ijms21093388>.

Elmusrati, A. A., Pilborough, A. E., Khurram, S. A., & Lambert, D. W. (2017). Cancer-associated fibroblasts promote bone invasion in oral squamous cell carcinoma. *British journal of cancer*, 117(6), 867-875.

Eslami, A., Miyaguchi, K., Mogushi, K., Watanabe, H., Okada, N., Shibuya, H., Mizushima, H., Miura, M., & Tanaka, H. (2015). PARVB overexpression increases cell migration capability and defines high risk for endophytic growth and metastasis in tongue squamous cell carcinoma. *British journal of cancer*, 112(2), 338–344.
<https://doi.org/10.1038/bjc.2014.590>

Fonseca, I. & Bettencourt-Dias, M. (2019). The Cell Cycle, Cytoskeleton and Cancer. In Fior, R., & Zilhão, R. (Eds.). *Molecular and cell biology of cancer*. Springer International Publishing. pp 51-74.

Gautier L, Cope L, Bolstad BM, Irizarry RA (2004). “affy—analysis of Affymetrix GeneChip data at the probe level.” *Bioinformatics*, 20(3), 307–315. ISSN 1367-4803, doi: 10.1093/bioinformatics/btg405.

Geng, F., Wang, Q., Li, C., Liu, J., Zhang, D., Zhang, S., & Pan, Y. (2019). Identification of Potential Candidate Genes of Oral Cancer in Response to Chronic Infection With *Porphyromonas gingivalis* Using Bioinformatical Analyses. *Frontiers in oncology*, 9, 91. <https://doi.org/10.3389/fonc.2019.00091>

George, A., Ranganathan, K., & Rao, U. K. (2010). Expression of MMP-1 in histopathological different grades of oral squamous cell carcinoma and in normal buccal mucosa - an immunohistochemical study. *Cancer biomarkers : section A of Disease markers*, 7(6), 275–283. <https://doi.org/10.3233/CBM-2010-0191>

Gerald, B. (2018). A brief review of independent, dependent and one sample t-test. *International Journal of Applied Mathematics and Theoretical Physics*, 4(2), 50.

Ghom, A. G., & Ghom, S. A. L. (Eds.). (2014). *Textbook of oral medicine*. JP Medical Ltd. pp 24-34

Gonzalo, R., & Sánchez, A. (2018). Introduction to microarrays technology and data analysis. In *Comprehensive Analytical Chemistry* (Vol. 82, pp. 37-69). Elsevier.

Hand, A. R., & Frank, M. E. (2014). *Fundamentals of oral histology and physiology*. John wiley & sons. pp 169-190.

Hejase, M. J., Bahrle, R., Castillo, G., & Coogan, C. L. (1996). Amebiasis of the penis. *Urology*, 48(1), 151–154. [https://doi.org/10.1016/s0090-4295\(96\)00108-2](https://doi.org/10.1016/s0090-4295(96)00108-2)

Hong F, Breitling R. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*. 2008 Feb 1;24(3):374-82. doi: 10.1093/bioinformatics/btm620. Epub 2008 Jan 18. PMID: 18204063.

Hong, F., Breitling, R., McEntee, W.C., Wittner, B.S., Nemhauser, J.L., Chory, J. (2006). “RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis.” *Bioinformatics*, 22, 2825-2827.

Hothorn T, Zeileis A (2015). “partykit: A Modular Toolkit for Recursive Partitioning in R.” *Journal of Machine Learning Research*, 16, 3905-3909.
<https://jmlr.org/papers/v16/hothorn15a.html>.

Hu, P., Greenwood, C. M., & Beyene, J. (2006). Statistical methods for meta-analysis of microarray data: a comparative study. *Information Systems Frontiers*, 8(1), 9-20.

Hu, X., Sun, G., Shi, Z., Ni, H., & Jiang, S. (2020). Identification and validation of key modules and hub genes associated with the pathological stage of oral squamous cell carcinoma by weighted gene co-expression network analysis. *PeerJ*, 8, e8505.
<https://doi.org/10.7717/peerj.8505>

Huang SH, O'Sullivan B. Oral cancer: Current role of radiotherapy and chemotherapy. *Med Oral Patol Oral Cir Bucal*. 2013 Mar 1;18(2):e233-40. doi: 10.4317/medoral.18772. PMID: 23385513; PMCID: PMC3613874.

Huang, X., Gao, Y., Zhi, X., Ta, N., Jiang, H., & Zheng, J. (2016). Association between vitamin A, retinol and carotenoid intake and pancreatic cancer risk: Evidence from epidemiologic studies. *Scientific reports*, 6, 38936. <https://doi.org/10.1038/srep38936>

Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., & Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic acids research*, 31(4), e15-e15.

Jayanthi, P., Varun, B. R., & Selvaraj, J. (2020). Epithelial-mesenchymal transition in oral squamous cell carcinoma: An insight into molecular mechanisms and clinical implications. *Journal of oral and maxillofacial pathology : JOMFP*, 24(1), 189.
https://doi.org/10.4103/jomfp.JOMFP_334_19

Jin, Z., Feng, X., Jian, Q., Cheng, Y., Gao, Y., Zhang, X., Wang, L., Zhang, Y., Huang, W., Fan, X., Chen, S., Yu, H., Zhao, Z., Dong, M., Liu, J., Mori, Y., & Meltzer, S. J.

(2013). Aberrant methylation of the Ras-related associated with diabetes gene in human primary esophageal cancer. *Anticancer research*, 33(11), 5199–5203.

Johnson, A., & Skotheim, J. M. (2013). Start and the restriction point. *Current opinion in cell biology*, 25(6), 717-723.

Kamat, M. S., Rai, B. D., Puranik, R. S., & Datar, U. V. (2020). Immunoexpression of p53 in histologically negative surgical margins adjacent to oral squamous cell carcinoma: A preliminary study. *Journal of oral and maxillofacial pathology : JOMFP*, 24(1), 184. https://doi.org/10.4103/jomfp.JOMFP_288_19

Karin, N., & Razon, H. (2018). Chemokines beyond chemo-attraction: CXCL10 and its significant role in cancer and autoimmunity. *Cytokine*, 109, 24–28. <https://doi.org/10.1016/j.cyto.2018.02.012>

Kato, A., Kato, K., Miyazawa, H., Kobayashi, H., Noguchi, N., & Kawashiri, S. (2020). Focal Adhesion Kinase (FAK) Overexpression and Phosphorylation in Oral Squamous Cell Carcinoma and their Clinicopathological Significance. *Pathology oncology research : POR*, 26(3), 1659–1667. <https://doi.org/10.1007/s12253-019-00732-y>

Kaufmann, T., Strasser, A., & Jost, P. J. (2012). Fas death receptor signalling: roles of Bid and XIAP. *Cell death and differentiation*, 19(1), 42–50. <https://doi.org/10.1038/cdd.2011.121>

Kita, A., Kasamatsu, A., Nakashima, D., Endo-Sakamoto, Y., Ishida, S., Shimizu, T., Kimura, Y., Miyamoto, I., Yoshimura, S., Shiiba, M., Tanzawa, H., & Uzawa, K. (2017). Activin B Regulates Adhesion, Invasiveness, and Migratory Activities in Oral Cancer: a Potential Biomarker for Metastasis. *Journal of Cancer*, 8(11), 2033–2041. <https://doi.org/10.7150/jca.18714>

Knudson, A. G. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences*, 68(4), 820-823.

Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J., & Peterson, H. (2020). gprofiler2--an R package for gene list functional enrichment analysis and namespace conversion toolset g: Profiler. *F1000Research*, 9.

Kotb, A., Hyndman, M. E., & Patel, T. R. (2018). The role of zyxin in regulation of malignancies. *Heliyon*, 4(7), e00695. <https://doi.org/10.1016/j.heliyon.2018.e00695>

Kothari, A. N., Arffa, M. L., Chang, V., Blackwell, R. H., Syn, W. K., Zhang, J., Mi, Z., & Kuo, P. C. (2016). Osteopontin-A Master Regulator of Epithelial-Mesenchymal Transition. *Journal of clinical medicine*, 5(4), 39. <https://doi.org/10.3390/jcm5040039>

- Kudo, Y., Takata, T., Ogawa, I., Kaneda, T., Sato, S., Takekoshi, T., Zhao, M., Miyauchi, M., & Nikai, H. (2000). p27Kip1 accumulation by inhibition of proteasome function induces apoptosis in oral squamous cell carcinoma cells. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 6(3), 916–923.
- Kuhn, M. (2009). The caret package. *Journal of Statistical Software*, 28(5).
- Kumar, G. S. (2015). *Orban's oral histology and embryology*. 14th edition Elsevier India. pp.194-236
- Larsson, O., Wahlestedt, C., & Timmons, J. A. (2005). Considerations when using the significance analysis of microarrays (SAM) algorithm. *BMC bioinformatics*, 6(1), 1-6.
- Lee, C. H., Chang, J. S., Syu, S. H., Wong, T. S., Chan, J. Y., Tang, Y. C., Yang, Z. P., Yang, W. C., Chen, C. T., Lu, S. C., Tang, P. H., Yang, T. C., Chu, P. Y., Hsiao, J. R., & Liu, K. J. (2015). IL-1β promotes malignant transformation and tumor aggressiveness in oral cancer. *Journal of cellular physiology*, 230(4), 875–884.
<https://doi.org/10.1002/jcp.24816>
- Lei, J., Ploner, A., Elfström, K. M., Wang, J., Roth, A., Fang, F., ... & Sparén, P. (2020). HPV vaccination and the risk of invasive cervical cancer. *New England Journal of Medicine*, 383(14), 1340-1348.
- Lese, C. M., Rossie, K. M., Appel, B. N., Reddy, J. K., Johnson, J. T., Myers, E. N., & Gollin, S. M. (1995). Visualization of INT2 and HST1 amplification in oral squamous cell carcinomas. *Genes, Chromosomes and Cancer*, 12(4), 288-29
- Levine, M. S., Bakker, B., Boeckx, B., Moyett, J., Lu, J., Vitre, B., ... & Holland, A. J. (2017). Centrosome amplification is sufficient to promote spontaneous tumorigenesis in mammals. *Developmental cell*, 40(3), 313-322.
- Li, Q., Wang, Y., Xu, L., Wang, L., Guo, Y., & Guo, C. (2020). High level of CD10 expression is associated with poor overall survival in patients with head and neck cancer. *International journal of oral and maxillofacial surgery*, S0901-5027(20)30369-6.
Advance online publication. <https://doi.org/10.1016/j.ijom.2020.07.037>
- Li, X., Qu, L., Zhong, Y., Zhao, Y., Chen, H., & Daru, L. (2013). Association between promoters polymorphisms of matrix metalloproteinases and risk of digestive cancers: a meta-analysis. *Journal of cancer research and clinical oncology*, 139(9), 1433–1447.
<https://doi.org/10.1007/s00432-013-1446-9>
- Li, Z., Xu, D., Jing, J., & Li, F. (2021). Network pharmacology-based study to explore the mechanism of the Yiqi Gubiao pill in lung cancer treatment. *Oncology letters*, 21(4), 321. <https://doi.org/10.3892/ol.2021.12583>

Liao, Y., Cao, W., Zhang, K., Zhou, Y., Xu, X., Zhao, X., Yang, X., Wang, J., Zhao, S., Zhang, S., Yang, L., Liu, D., Tian, Y., & Wu, W. (2021). Bioinformatic and integrated analysis identifies an lncRNA-miRNA-mRNA interaction mechanism in gastric adenocarcinoma. *Genes & genomics*, 10.1007/s13258-021-01086-z. Advance online publication. <https://doi.org/10.1007/s13258-021-01086-z>

Liaw A. & Wiener M. (2002). Classification and Regression by randomForest. *R News* 2(3), 18-22.

Lim, W. K., Wang, K., Lefebvre, C., & Califano, A. (2007). Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics*, 23(13), i282-i288.

Lippens, S., Denecker, G., Ovaere, P., Vandenabeele, P., & Declercq, W. (2005). Death penalty for keratinocytes: apoptosis versus cornification. *Cell Death & Differentiation*, 12(2), 1497-1508.

Lo Presti, E., Toia, F., Oieni, S., Buccheri, S., Turdo, A., Mangiapane, L. R., Campisi, G., Caputo, V., Todaro, M., Stassi, G., Cordova, A., Moschella, F., Rinaldi, G., Meraviglia, S., & Dieli, F. (2017). Squamous Cell Tumors Recruit $\gamma\delta$ T Cells Producing either IL17 or IFN γ Depending on the Tumor Stage. *Cancer immunology research*, 5(5), 397–407. <https://doi.org/10.1158/2326-6066.CIR-16-0348>

Lohavanichbutr, P., Méndez, E., Holsinger, F. C., Rue, T. C., Zhang, Y., Houck, J., Upton, M. P., Futran, N., Schwartz, S. M., Wang, P., & Chen, C. (2013). A 13-gene signature prognostic of HPV-negative OSCC: discovery and external validation. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 19(5), 1197–1203. <https://doi.org/10.1158/1078-0432.CCR-12-2647>

Lopes, C. T., Franz, M., Kazi, F., Donaldson, S. L., Morris, Q., & Bader, G. D. (2010). Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, 26(18), 2347-2348.

Łukaszewicz-Zająć, M., Pączek, S., Mroczko, P., & Kulczyńska-Przybik, A. (2020). The Significance of CXCL1 and CXCL8 as Well as Their Specific Receptors in Colorectal Cancer. *Cancer management and research*, 12, 8435–8443. <https://doi.org/10.2147/CMAR.S267176>

Luo, Weijun, Brouwer, Cory (2013). “Pathview: an R/Bioconductor package for pathway-based data integration and visualization.” *Bioinformatics*, 29(14), 1830-1831. doi: 10.1093/bioinformatics/btt285.

Lyu, J., Wang, J., Miao, Y., Xu, T., Zhao, W., Bao, T., & Zhu, H. (2021). KLF7 is associated with poor prognosis and regulates migration and adhesion in tongue

cancer. *Oral diseases*, 10.1111/odi.13767. Advance online publication.
<https://doi.org/10.1111/odi.13767>

Mane, D. R., Kale, A. D., & Belaldavar, C. (2017). Validation of immunoexpression of tenascin-C in oral precancerous and cancerous tissues using ImageJ analysis with novel immunohistochemistry profiler plugin: An immunohistochemical quantitative analysis. *Journal of oral and maxillofacial pathology : JOMFP*, 21(2), 211–217.
https://doi.org/10.4103/jomfp.JOMFP_234_16

Marcinkiewicz, K. M., & Gudas, L. J. (2014). Altered epigenetic regulation of homeobox genes in human oral squamous cell carcinoma cells. *Experimental cell research*, 320(1), 128–143. <https://doi.org/10.1016/j.yexcr.2013.09.011>

Mariotti, S., Barravecchia, I., Vindigni, C., Pucci, A., Balsamo, M., Libro, R., Senchenko, V., Dmitriev, A., Jacchetti, E., Cecchini, M., Roviello, F., Lai, M., Broccoli, V., Andreazzoli, M., Mazzanti, C. M., & Angeloni, D. (2016). MICAL2 is a novel human cancer gene controlling mesenchymal to epithelial transition involved in cancer growth and invasion. *Oncotarget*, 7(2), 1808–1825. <https://doi.org/10.18632/oncotarget.6577>

Mayer, B. (Ed.). (2011). *Bioinformatics for omics data: methods and protocols* (No. 57: 004 BIO). New York: Humana Press.

McCall, M. N., Murakami, P. N., Lukk, M., Huber, W., & Irizarry, R. A. (2011). Assessing affymetrix GeneChip microarray quality. *BMC bioinformatics*, 12(1), 1-10. Med J 85:134–140

Meissl, K., Macho-Maschler, S., Müller, M., & Strobl, B. (2017). The good and the bad faces of STAT1 in solid tumours. *Cytokine*, 89, 12–20.
<https://doi.org/10.1016/j.cyto.2015.11.011>

Méndez, E., Lohavanichbutr, P., Fan, W., Houck, J. R., Rue, T. C., Doody, D. R., Futran, N. D., Upton, M. P., Yueh, B., Zhao, L. P., Schwartz, S. M., & Chen, C. (2011). Can a metastatic gene expression profile outperform tumor size as a predictor of occult lymph node metastasis in oral cancer patients?. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 17(8), 2466–2473.
<https://doi.org/10.1158/1078-0432.CCR-10-0175>

Meucci, S., Keilholz, U., Heim, D., Klauschen, F., & Cacciatore, S. (2018). Somatic genome alterations in relation to age in lung squamous cell carcinoma. *Oncotarget*, 9(63), 32161–32172. <https://doi.org/10.18632/oncotarget.25848>

Miguel, A., Mello, F. W., Melo, G., & Rivero, E. (2020). Association between immunohistochemical expression of matrix metalloproteinases and metastasis in oral squamous cell carcinoma: Systematic review and meta-analysis. *Head & neck*, 42(3), 569–584. <https://doi.org/10.1002/hed.26009>

Mitsuwan, W., & Nissapatorn, V. (2020). Interaction of human oral cancer and the expression of virulence genes of dental pathogenic bacteria. *Microbial pathogenesis*, 149, 104464. <https://doi.org/10.1016/j.micpath.2020.104464>

Müller, C., Schillert, A., Röthemeier, C., Trégouët, D. A., Proust, C., Binder, H., ... & Ziegler, A. (2016). Removing batch effects from longitudinal gene expression-quantile normalization plus ComBat as best approach for microarray transcriptome data. *PloS one*, 11(6), e0156594.

Munoz, N., Castellsagué, X., de González, A. B., & Gissmann, L. (2006). HPV in the etiology of human cancer. *Vaccine*, 24, S1-S10

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion?. *Journal of classification*, 31(3), 274-295.

Nagata, M., Noman, A. A., Suzuki, K., Kurita, H., Ohnishi, M., Ohyama, T., Kitamura, N., Kobayashi, T., Uematsu, K., Takahashi, K., Kodama, N., Kawase, T., Hoshina, H., Ikeda, N., Shingaki, S., & Takagi, R. (2013). ITGA3 and ITGB4 expression biomarkers estimate the risks of locoregional and hematogenous dissemination of oral squamous cell carcinoma. *BMC cancer*, 13, 410. <https://doi.org/10.1186/1471-2407-13-410>

Nanci A.(2018): Ten Cate's Oral Histology - Development , Structure and Function. 9th edition. Elsevier pp.260-272.

Näsman, A., Du, J., & Dalianis, T. (2020). A global epidemic increase of an HPV-induced tonsil and tongue base cancer—potential benefit from a pan-gender use of HPV vaccine. *Journal of internal medicine*, 287(2), 134-152.

Neville, B. W., Damm, D. D., Allen, C. M., & Bouquot, J. (2008). *Oral and maxillofacial pathology*. Elsevier Health Sciences.409-421

O'Donnell, R. K., Kupferman, M., Wei, S. J., Singhal, S., Weber, R., O'Malley, B., Cheng, Y., Putt, M., Feldman, M., Ziobor, B., & Muschel, R. J. (2005). Gene expression signature predicts lymphatic metastasis in squamous cell carcinoma of the oral cavity. *Oncogene*, 24(7), 1244–1251. <https://doi.org/10.1038/sj.onc.1208285>

Oghumu S, Knobloch TJ, Terrazas C, Varikuti S, Ahn-Jarvis J, Bollinger CE, Iwenofu H, Weghorst CM, Satoskar AR. Deletion of macrophage migration inhibitory factor inhibits murine oral carcinogenesis: Potential role for chronic pro-inflammatory immune mediators. *Int J Cancer*. 2016 Sep 15;139(6):1379-90. doi: 10.1002/ijc.30177. Epub 2016 Jun 16. PMID: 27164411; PMCID: PMC4939094.

- Ohnishi, K., Murata, M., Kojima, H., Takemura, N., Tsuchida, T., & Tachibana, H. (1994). Brain abscess due to infection with *Entamoeba histolytica*. *The American journal of tropical medicine and hygiene*, 51(2), 180-182.
- Okoniewski, M. J., Yates, T., Dibben, S., & Miller, C. J. (2007). An annotation infrastructure for the analysis and interpretation of Affymetrix exon array data. *Genome biology*, 8(5), 1-9.
- Osazuwa-Peters, N., Boakye, E. A., Mohammed, K. A., Tobo, B. B., Geneus, C. J., & Shootman, M. (2017). Not just a woman's business! Understanding men and women's knowledge of HPV, the HPV vaccine, and HPV-associated cancers. *Preventive medicine*, 99, 299-304.
- Osei-Sarfo, K., & Gudas, L. J. (2019). Retinoids induce antagonism between FOXO3A and FOXM1 transcription factors in human oral squamous cell carcinoma (OSCC) cells. *PloS one*, 14(4), e0215234. <https://doi.org/10.1371/journal.pone.0215234>
- Oshiro, N., Rapley, J., & Avruch, J. (2014). Amino acids activate mammalian target of rapamycin (mTOR) complex 1 without changing Rag GTPase guanyl nucleotide charging. *The Journal of biological chemistry*, 289(5), 2658–2674. <https://doi.org/10.1074/jbc.M113.528505>
- Panarese, I., Aquino, G., Ronchi, A., Longo, F., Montella, M., Cozzolino, I., ... & Franco, R. (2019). Oral and Oropharyngeal squamous cell carcinoma: prognostic and predictive parameters in the etiopathogenetic route. *Expert review of anticancer therapy*, 19(2), 105-119.
- Papagerakis, S., Pannone, G., Zheng, L., About, I., Taqi, N., Nguyen, N. P., ... & Papagerakis, P. (2014). Oral epithelial stem cells—implications in normal development and cancer metastasis. *Experimental cell research*, 325(2), 111-129.
- Pappa, E., Nikitakis, N., Vlachodimitropoulos, D., Avgoustidis, D., Oktseloglou, V., & Papadogeorgakis, N. (2014). Phosphorylated signal transducer and activator of transcription-1 immunohistochemical expression is associated with improved survival in patients with oral squamous cell carcinoma. *Journal of oral and maxillofacial surgery : official journal of the American Association of Oral and Maxillofacial Surgeons*, 72(1), 211–221. <https://doi.org/10.1016/j.joms.2013.06.198>
- Park JW, Min K-J, Kim DE, Kwon TK (2015) Withaferin A induces apoptosis through the generation of thiol oxidation in human head and neck cancer cells. *Int J Mol Med* 35:247–252
- Park, S. Y., Jeong, M. S., & Jang, S. B. (2012). In vitro binding properties of tumor suppressor p53 with PUMA and NOXA. *Biochemical and biophysical research communications*, 420(2), 350-356.

Patel, B. P., Shah, P. M., Rawal, U. M., Desai, A. A., Shah, S. V., Rawal, R. M., & Patel, P. S. (2005). Activation of MMP-2 and MMP-9 in patients with oral squamous cell carcinoma. *Journal of surgical oncology*, 90(2), 81-88.

Peng, C. H., Liao, C. T., Peng, S. C., Chen, Y. J., Cheng, A. J., Juang, J. L., Tsai, C. Y., Chen, T. C., Chuang, Y. J., Tang, C. Y., Hsieh, W. P., & Yen, T. C. (2011). A novel molecular signature identified by systems genetics approach predicts prognosis in oral squamous cell carcinoma. *PloS one*, 6(8), e23452.
<https://doi.org/10.1371/journal.pone.0023452>

Pickering, C. R., Zhang, J., Yoo, S. Y., Bengtsson, L., Moorthy, S., Neskey, D. M., Zhao, M., Ortega Alves, M. V., Chang, K., Drummond, J., Cortez, E., Xie, T. X., Zhang, D., Chung, W., Issa, J. P., Zweidler-McKay, P. A., Wu, X., El-Naggar, A. K., Weinstein, J. N., Wang, J., ... Frederick, M. J. (2013). Integrative genomic characterization of oral squamous cell carcinoma identifies frequent somatic drivers. *Cancer discovery*, 3(7), 770–781. <https://doi.org/10.1158/2159-8290.CD-12-0537>

Popli, D. B., Sircar, K., Chowdhry, A., & Rani, V. (2015). Role of heat shock proteins in oral squamous cell carcinoma: A systematic review. *Biomedical Papers*, 159(3), 366-371.

Povoa V, Fior T (2019).Cancer Immunoediting and hijacking the immune system. In Fior, R., & Zilhão, R. (Eds). *Molecular and cell biology of cancer*. Springer International Publishing. pp 117-139

Prasad H, Anuthama K (2019). *Atlas of oral histology*, Elsevier. pp 65-78.

Qadir, F., Lalli, A., Dar, H. H., Hwang, S., Aldehlawi, H., Ma, H., Dai, H., Waseem, A., & Teh, M. T. (2019). Clinical correlation of opposing molecular signatures in head and neck squamous cell carcinoma. *BMC cancer*, 19(1), 830. <https://doi.org/10.1186/s12885-019-6059-5>

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>

Rafael Ríos-Burgueño, E., Velarde-Félix, J. S., & Flores García, A. (2017). Penile amebiasis mimicking penile squamous cell carcinoma. *International journal of dermatology*, 56(5), 573–575. <https://doi.org/10.1111/ijd.13524>

Ramasamy A, Mondry A, Holmes CC, Altman DG (2008) Key Issues in Conducting a Meta-Analysis of Gene Expression Microarray Datasets. PLOS Medicine 5(9): e184. <https://doi.org/10.1371/journal.pmed.0050184>

Ravikiran O, Praveen B. (2014). Textbook of oral medicine, oral diagnosis and oral radiology. 2nd edition, Elsevier. pp 380-402.

Reis, P. P., Waldron, L., Perez-Ordonez, B., Pintilie, M., Galloni, N. N., Xuan, Y., Cervigne, N. K., Warner, G. C., Makitie, A. A., Simpson, C., Goldstein, D., Brown, D., Gilbert, R., Gullane, P., Irish, J., Jurisica, I., & Kamel-Reid, S. (2011). A gene signature in histologically normal surgical margins is predictive of oral carcinoma recurrence. *BMC cancer*, 11, 437. <https://doi.org/10.1186/1471-2407-11-437>

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). “limma powers differential expression analyses for RNA-sequencing and microarray studies.” *Nucleic Acids Research*, 43(7), e47. doi: 10.1093/nar/gkv007.

Rivera, C., & Rivera, C. (2014). Histological and molecular aspects of oral squamous cell carcinoma (Review). *Oncology Letters*, 8, 7-11. <https://doi.org/10.3892/ol.2014.2103>

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J, Müller M (2011). “pROC: an open-source package for R and S+ to analyze and compare ROC curves.” *BMC Bioinformatics*, 12, 77.

Roden DL, Sewell GW, Lobley A, Levine AP, Smith AM, et al. (2014). ZODET: Software for the Identification, Analysis and Visualisation of Outlier Genes in Microarray Expression Data. *PLOS ONE* 9(1): e81123. <https://doi.org/10.1371/journal.pone.0081123>

Roland, N. J., Caslin, A. W., Nash, J., & Stell, P. M. (1992). Value of grading squamous cell carcinoma of the head and neck. *Head & neck*, 14(3), 224-229.

Sacconi A, Donzelli S, Pulito C, Ferrero S et al (2020). TMPRSS2, a SARS-CoV-2 internalization protease is downregulated in head and neck cancer patients. *J Exp Clin Cancer Res* 2020 Sep 23;39(1):200. PMID: 32967703

Saeed, A. A., Sims, A. H., Prime, S. S., Paterson, I., Murray, P. G., & Lopes, V. R. (2015). Gene expression profiling reveals biological pathways responsible for phenotypic heterogeneity between UK and Sri Lankan oral squamous cell carcinomas. *Oral oncology*, 51(3), 237–246. <https://doi.org/10.1016/j.oraloncology.2014.12.004>

Sahai, E., Astsaturov, I., Cukierman, E., DeNardo, D. G., Egeblad, M., Evans, R. M., ... & Werb, Z. (2020). A framework for advancing our understanding of cancer-associated fibroblasts. *Nature Reviews Cancer*, 20(3), 174-186.

Sainio, A., & Järveläinen, H. (2020). Extracellular matrix-cell interactions: focus on therapeutic applications. *Cellular signalling*, 66, 109487.

Santacroce, L., Di Cosola, M., Bottalico, L., Topi, S., Charitos, I. A., Ballini, A., Inchingolo, F., Cazzolla, A. P., & Dipalma, G. (2021). Focus on HPV Infection and the Molecular Mechanisms of Oral Carcinogenesis. *Viruses*, 13(4), 559. <https://doi.org/10.3390/v13040559>

- Sartor, M. A., Tomlinson, C. R., Wesselkamper, S. C., Sivaganesan, S., Leikauf, G. D., & Medvedovic, M. (2006). Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments. *BMC bioinformatics*, 7(1), 1-17.
- Sass, P. A., Dąbrowski, M., Charzyńska, A., & Sachadyn, P. (2017). Transcriptomic responses to wounding: meta-analysis of gene expression microarray data. *BMC genomics*, 18(1), 1-12.
- Schneider M & Orchard S. Omics Technologies, Data and Bioinformatics Principles in Mayer, B. (Ed.). (2011). *Bioinformatics for omics data: methods and protocols* (No. 57: 004 BIO). New York: Humana Press. pp 3- 31.
- Scully, C. (2012). *Oral and Maxillofacial Medicine-E-Book: The Basis of Diagnosis and Treatment*. Elsevier Health Sciences.204-216
- SEER Preliminary Cancer Incidence Rate Estimates for 2017, and diagnosis years 2000 to 2017, SEER 18, National Cancer Institute. Bethesda, MD, <https://seer.cancer.gov/statistics/preliminary-estimates/>, based on the February 2019 SEER data submission and the November 2018 SEER data submission. Posted to the SEER web site, September 2019.
- Seidel, J. A., Otsuka, A., & Kabashima, K. (2018). Anti-PD-1 and Anti-CTLA-4 Therapies in Cancer: Mechanisms of Action, Efficacy, and Limitations. *Frontiers in oncology*, 8, 86. <https://doi.org/10.3389/fonc.2018.00086>
- Skappak, C., Akierman, S., Belga, S., Novak, K., Chadee, K., Urbanski, S. J., ... & Beck, P. L. (2014). Invasive amoebiasis: a review of Entamoeba infections highlighted with case reports. *Canadian Journal of Gastroenterology and Hepatology*, 28(7), 355-359.
- Song, X., Wei, C., & Li, X. (2021). The potential role and status of IL-17 family cytokines in breast cancer. *International immunopharmacology*, 95, 107544. Advance online publication. <https://doi.org/10.1016/j.intimp.2021.107544>
- Spénlé, C., Loustau, T., Murdamoothoo, D., Erne, W., Beghelli-de la Forest Divonne, S., Veber, R., Petti, L., Bourdely, P., Mörgelin, M., Brauchle, E. M., Cremel, G., Randrianarisoa, V., Camara, A., Rekima, S., Schaub, S., Nouhen, K., Imhof, T., Hansen, U., Paul, N., Carapito, R., ... Orend, G. (2020). Tenascin-C Orchestrates an Immune-Suppressive Tumor Microenvironment in Oral Squamous Cell Carcinoma. *Cancer immunology research*, 8(9), 1122–1138. <https://doi.org/10.1158/2326-6066.CIR-20-0074>
- Srivastava, G. (2008). *Essentials of Oral Medicine*. JAYPEE BROTHERS PUBLISHERS. pp.80-98.

Su, L., Wang, C., Zheng, C., Wei, H., & Song, X. (2018). A meta-analysis of public microarray data identifies biological regulatory networks in Parkinson's disease. *BMC medical genomics*, 11(1), 40.

Sun, Y., & Peng, Z. L. (2009). Programmed cell death and cancer. *Postgraduate medical journal*, 85(1001), 134–140. <https://doi.org/10.1136/pgmj.2008.072629>

Suzuki, M., Shigematsu, H., Shames, D. S., Sunaga, N., Takahashi, T., Shivapurkar, N., Iizasa, T., Minna, J. D., Fujisawa, T., & Gazdar, A. F. (2007). Methylation and gene silencing of the Ras-related GTPase gene in lung and breast cancers. *Annals of surgical oncology*, 14(4), 1397–1404. <https://doi.org/10.1245/s10434-006-9089-6>

Takebe, N., Miele, L., Harris, P. J., Jeong, W., Bando, H., Kahn, M., ... & Ivy, S. P. (2015). Targeting Notch, Hedgehog, and Wnt pathways in cancer stem cells: clinical update. *Nature reviews Clinical oncology*, 12(8), 445.

Taminau J, Meganck S, Lazar C, Steenhoff D, Coletta A, Molter C, Duque R, Schaetzen Vd, Solis DYW, Bersini H and Nowe A (2012). “Unlocking the potential of publicly available microarray data using inSilicoDb and inSilicoMerging R/Bioconductor packages.” *BMC Bioinformatics*, 13, pp. 355.

Teh, M. T., Gemenetzidis, E., Patel, D., Tariq, R., Nadir, A., Bahta, A. W., Waseem, A., & Hutchison, I. L. (2012). FOXM1 induces a global methylation signature that mimics the cancer epigenome in head and neck squamous cell carcinoma. *PloS one*, 7(3), e34329. <https://doi.org/10.1371/journal.pone.0034329>

Therneau, T., Atkinson, B., Ripley, B., & Ripley, M. B. (2015). Package ‘rpart’. Available online: cran.ma.ic.ac.uk/web/packages/rpart/rpart.pdf (accessed on 20 April 2016).

Tibshirani, R., Chu, G., Hastie, T., Narasimhan, B., & Tibshirani, M. R. (2011). The samr Package.

Toruner, G. A., Ulger, C., Alkan, M., Galante, A. T., Rinaggio, J., Wilk, R., Tian, B., Soteropoulos, P., Hameed, M. R., Schwalb, M. N., & Dermody, J. J. (2004). Association between gene expression profile and tumor invasion in oral squamous cell carcinoma. *Cancer genetics and cytogenetics*, 154(1), 27–35. <https://doi.org/10.1016/j.cancergenryo.2004.01.026>

Tzeng I. S. (2021). Modified Significance Analysis of Microarrays in Heterogeneous Diseases. *Journal of personalized medicine*, 11(2), 62. <https://doi.org/10.3390/jpm11020062>

Verduci, L., Ferraiuolo, M., Sacconi, A., Ganci, F., Vitale, J., Colombo, T., Paci, P., Strano, S., Macino, G., Rajewsky, N., & Blandino, G. (2017). The oncogenic role of circPVT1 in head and neck squamous cell carcinoma is mediated through the mutant

p53/YAP/TEAD transcription-competent complex. *Genome biology*, 18(1), 237.
<https://doi.org/10.1186/s13059-017-1368-y>

Waghela, B. N., Vaidya, F. U., Ranjan, K., Chhipa, A. S., Tiwari, B. S., & Pathak, C. (2021). AGE-RAGE synergy influences programmed cell death signaling to promote cancer. *Molecular and cellular biochemistry*, 476(2), 585–598.
<https://doi.org/10.1007/s11010-020-03928-y>

Walsh, C. J., Hu, P., Batt, J., & Santos, C. C. (2015). Microarray Meta-Analysis and Cross-Platform Normalization: Integrative Genomics for Robust Biomarker Discovery. *Microarrays (Basel, Switzerland)*, 4(3), 389–406.
<https://doi.org/10.3390/microarrays4030389>

Walter, V., Yin, X., Wilkerson, M. D., Cabanski, C. R., Zhao, N., Du, Y., ... & Hayes, D. N. (2013). Molecular subtypes in head and neck cancer exhibit distinct patterns of chromosomal gain and loss of canonical cancer genes. *PloS one*, 8(2), e56823.

Wang Y, Fan H, Zheng L. Biological information analysis of differentially expressed genes in oral squamous cell carcinoma tissues in GEO database. J BUON. 2018 Nov-Dec;23(6):1662-1670. PMID: 30610792.

Wang, S. S., Zheng, M., Pang, X., Zhang, M., Yu, X. H., Wu, J. B., Gao, X. L., Wu, J. S., Yang, X., Tang, Y. J., Tang, Y. L., & Liang, X. H. (2019). Macrophage migration inhibitory factor promotes the invasion and metastasis of oral squamous cell carcinoma through matrix metalloprotein-2/9. *Molecular carcinogenesis*, 58(10), 1809–1821.
<https://doi.org/10.1002/mc.23067>

Warnakulasuriya, S., & Tilakaratna, W. M. (Eds.). (2013). *Oral medicine & pathology: a guide to diagnosis and management*. JP Medical Ltd. pp.301-330.

Warnes, M. G. R., Bolker, B., Bonebakker, L., Gentleman, R., & Huber, W. (2016). Package ‘gplots’. *Various R programming tools for plotting data*.

Wei, L. Y., Lee, J. J., Yeh, C. Y., Yang, C. J., Kok, S. H., Ko, J. Y., Tsai, F. C., & Chia, J. S. (2019). Reciprocal activation of cancer-associated fibroblasts and oral squamous carcinoma cells through CXCL1. *Oral oncology*, 88, 115–123.
<https://doi.org/10.1016/j.oraloncology.2018.11.002>

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3), 37-52.
Wu, Z., & Irizarry, R. A. (2004). Preprocessing of oligonucleotide array data. *Nature biotechnology*, 22(6), 656-658.

Xu, X., Li, M., Hu, J., Chen, Z., Yu, J., Dong, Y., Sun, C., & Han, J. (2018). Expression profile analysis identifies a two-gene signature for prediction of head and neck squamous

cell carcinoma patient survival. *Journal of cancer research and therapeutics*, 14(7), 1525–1534. https://doi.org/10.4103/jcrt.JCRT_557_18

Xue, J., Jia, E., Ren, N., & Xin, H. (2021). Identification of prognostic miRNA biomarkers for esophageal cancer based on The Cancer Genome Atlas and Gene Expression Omnibus. *Medicine*, 100(7), e24832. <https://doi.org/10.1097/MD.00000000000024832>

Yasrebi H. (2016). Comparative study of joint analysis of microarray gene expression data in survival prediction and risk assessment of breast cancer patients. *Briefings in bioinformatics*, 17(5), 771–785. <https://doi.org/10.1093/bib/bbv092>

Yeh, M. H., Tzeng, Y. J., Fu, T. Y., You, J. J., Chang, H. T., Ger, L. P., & Tsai, K. W. (2018). Extracellular Matrix-receptor Interaction Signaling Genes Associated with Inferior Breast Cancer Survival. *Anticancer research*, 38(8), 4593–4605. <https://doi.org/10.21873/anticanres.12764>

Yoshiba, S., Iwase, M., Kurihara, S., Uchida, M., Kurihara, Y., Watanabe, H., & Shintani, S. (2011). Proteasome inhibitor sensitizes oral squamous cell carcinoma cells to TRAIL-mediated apoptosis. *Oncology reports*, 25(3), 645–652. <https://doi.org/10.3892/or.2010.1127>

Zetterberg, A., Larsson, O., & Wiman, K. G. (1995). What is the restriction point?. *Current opinion in cell biology*, 7(6), 835-842.

Zhang S. (2007). A comprehensive evaluation of SAM, the SAM R-package and a simple modification to improve its performance. *BMC bioinformatics*, 8, 230. <https://doi.org/10.1186/1471-2105-8-230>

Zhang, J., Wang, F., Liu, F., & Xu, G. (2020). Predicting STAT1 as a prognostic marker in patients with solid cancer. *Therapeutic advances in medical oncology*, 12, 1758835920917558. <https://doi.org/10.1177/1758835920917558>

Zhang, Y., & Liu, Z. (2017). STAT1 in cancer: friend or foe?. *Discovery medicine*, 24(130), 19–29.

Zhao, E., Xu, J., Yin, X., Sun, Y., Shi, J., & Li, X. (2009). Detection of deregulated pathways to lymphatic metastasis in oral squamous cell carcinoma. *Pathology oncology research : POR*, 15(2), 217–223. <https://doi.org/10.1007/s12253-008-9102-4>

Zhao, S., Wang, B., Ma, Y., Kuang, J., Liang, J., & Yuan, Y. (2020). NUCKS1 Promotes Proliferation, Invasion and Migration of Non-Small Cell Lung Cancer by Upregulating CDK1 Expression. *Cancer management and research*, 12, 13311–13323. <https://doi.org/10.2147/CMAR.S282181>

Zilhao R & Neves H. Tumor Niche disruption and metastasis. The role of Epithelial-Mesenchymal transition. In Fior, R., & Zilhão, R. (Eds). *Molecular and cell biology of cancer*. Springer International Publishing. pp 159-189.

Zou, B., Wang, D., Xu, K., Yuan, D. Y., Meng, Z., & Zhang, B. (2019). Integrin α-5 as a potential biomarker of head and neck squamous cell carcinoma. *Oncology letters*, 18(4), 4048–4055. <https://doi.org/10.3892/ol.2019.10773>

Αγγελόπουλος Α., Παπανικολάου Σ., Αγγελοπούλου Ε.(2000): Σύγχρονη στοματική και γναθοπροσωπική παθολογία. 3^η έκδοση , Εκδόσεις Λίτσας.

Κυρές Π, Κορώνης Σ, Τόσιος Κ, Νικητάκης Ν, Σκλαβούνον A (2008): Ακροχορδωνώδες καρκίνωμα του στόματος , κλινικοπαθολογική μελέτη 45 περιπτώσεων. Στοματολογία 65(2), 55-66, 2008.

Νικολάου X, Χουβαρδάς Π. (2015) Υπολογιστική Βιολογία. Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών.

Υπεύθυνη Δήλωση Συγγραφέα: Δηλώνω ρητά ότι, σύμφωνα με το άρθρο 8 του Ν.1599/1986, η παρούσα εργασία αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης.

III.

ΠΑΡΑΡΤΗΜΑΤΑ

Παράρτημα 1. Αρχικά GEO Datasets

!part1."oral squamous cell carcinoma" part2."squamous cell carcinoma" AND "oral"
part3."squamous cell carcinoma" AND "tongue"

geo,oscc,normal,no_platforms,primary_platform,acception,info
GSE97251,3,3,1,GPL21827,rejected,blood_samples
GSE130441,15,0,1,GPL145550,rejected,cell-line
GSE160395,8,0,1,GPL21185,rejected,cell-line
GSE153918,16,0,1,GPL13667,accepted,
GSE108712,62,0,1,GPL24460,rejected,4-genes_only
GSE138206,6,12,1,GPL570,accepted,contralateral-marginal
GSE87593,8,0,1,GPL14550,accepted,+8_LN
GSE127770,3,0,1,GPL16686,rejected,transfected
GSE114928,6,0,1,GPL17077,rejected,cell-line
GSE108851,6,0,1,GPL16699,rejected,transfected
GSE95704,18,0,1,GPL17930,rejected,cell-line
GSE115801,0,0,1,GPL20844,rejected,lung_OSCC
GSE115116,6,0,1,GPL13497,rejected,transfected
GSE113860,8,0,1,GPL21185,rejected,cell-line
GSE106791,2,0,1,GPL20844,rejected,cell-line
GSE74530,12,0,1,GPL570,accepted,
GSE84846,99,0,1,GPL6480,accepted,
GSE72118,8,0,1,GPL10558,rejected,cell-line
GSE100746,8,0,1,GPL20844,rejected,transfected
GSE100589,6,0,1,GPL570,rejected,transfected
GSE89923,45,45,1,GPL570,rejected,cell-line
GSE87539,0,12,2,GPL22516-GPL22517,rejected,nonOSCC-gingiva

GSE85195,35,1,1,GPL6480,accepted,+15_leucoplakia
GSE97569,8,0,1,GPL20844,rejected,cell-line
GSE85966,6,0,1,GPL10904,rejected,cell-line
GSE83346,15,0,1,GPL17077,rejected,cell-line
GSE83276,13,0,1,GPL17077,rejected,cell-line
GSE72938,10,0,1,GPL15207,rejected,cell-line
GSE72935,15,0,1,GPL16699,rejected,duplicateGSE72938
GSE87587,6,0,1,GPL13912,rejected,mouse
GSE83981,16,0,2,GPL11487-GPL14550,rejected,mouse
GSE80347,6,0,1,GPL2895,rejected,transfected
GSE70301,6,0,1,GPL570,rejected,transfected
GSE66498,32,0,1,GPL16699,rejected,transfected
GSE79795,9,0,2,GPL8432-GPL18461,accepted,same_samples_with_2_array_types
GSE62636,12,0,1,GPL14550,rejected,cell-line
GSE64216,2,2,1,GPL10558,rejected,inadeqate_samples
GSE74580,12,0,1,GPL10558,rejected,cell-line
GSE70604,7,1,1,GPL2986,rejected,LN
GSE64271,7,5,1,GPL6885,rejected,mouse
GSE75127,8,0,1,GPL570,rejected,transfected
GSE59069,36,0,1,GPL14555,accepted,
GSE66949,12,0,1,GPL17889,rejected,transfected
GSE38823,4,0,1,GPL6883,rejected,cell-line
GSE57819,6,0,1,GPL17425,rejected,cell-line
GSE58162,6,0,1,GPL10904,rejected,cell-line
GSE56532,10,6,1,GPL10739,accepted,
GSE51125,11,11,1,GPL14746,rejected,rat
GSE59407,11,7,2,GPL6480-GPL18947,rejected,cell-line
GSE59238,0,8,1,GPL6480,rejected,cell-line

GSE31277,15,15,3,GPL9770,rejected,custom-array

GSE52795,6,0,1,GPL18001,rejected,cell-line

GSE57083,9,0,1,GPL570,rejected,cell-line

GSE51010,48,8,2,GPL570-GPL201,accepted,Normal_GPL570-OSCC_GPL201

GSE38517,15,5,1,GPL570,rejected,fibroblasts

GSE46802,10,10,2,GPL6480,accepted,+10Dysplasia

GSE37991,40,40,1,GPL6883,accepted,

GSE41613,97,0,1,GPL570,accepted

GSE41117,43,0,3,GPL5175,rejected,duplicateGSE1116

GSE41116,43,0,1,GPL5175,accepted,

GSE41495,8,0,1,GPL13497,rejected,cell-line

GSE36090,10,3,2,GPL2986,accepted,

GSE23558,27,4,1,GPL6480,accepted,

GSE31853,8,3,2,GPL96-GPL570,rejected,cell-line

GSE25104,57,22,2,GPL5175,accepted,

GSE31056,23,73,1,GPL10526,accepted,

GSE30784,167,45,1,GPL570,accepted,+17Dysplasia

GSE11357,9,0,1,GPL1261,rejected,cell-line

GSE9792,9,0,1,GPL6248,accepted,

GSE9844,26,12,1,GPL570,rejected,duplicateGSE30784

GSE10121,35,6,1,GPL6353,rejected,duplicateGSE30784

GSE4676,49,0,1,GPL2891,accepted,

GSE7073,8,0,1,GPL1291,rejected,cell-line

GSE3524,16,4,1,GPL96,accepted,

, , , ,

GSE89146,40,0,1,GPL4133,accepted,

GSE148944,38,6,1,GPL28285,rejected,***only 784 rows***

GSE122272,5,0,1,GPL23249,accepted,

GSE110996,15,0,2,GPL570,accepted,
GSE101491,19,0,1,GPL149561,rejected,unspecified_sample_location
GSE89217,12,0,1,GPL10558,rejected,cell-line
GSE108062,26,0,2,GPL9128,rejected,duplicateGSE108061
GSE108061,26,0,1,GPL9128,accepted,
GSE103412,0,10,1,GPL23978,accepted,only_10_normal-tonsil
GSE107591,16,16,1,GPL6244,accepted,
GSE85446,66,0,1,GPL6480,accepted,+Reanalysis_GSE30788
GSE94833,3,0,1,GPL6244,rejected,laser_transformed
GSE65021,19,0,1,GPL14951,accepted,
GSE75540,7,8,2,GPL10904,rejected,duplicateGSE75339
GSE75539,7,8,1,GPL10904,accepted,
GSE75338,7,8,1,GPL10904,rejected,duplicateGSE75339
GSE64217,6,0,1,GPL10558,rejected,cervical_OSCC
GSE57760,6,0,1,GPL571,rejected,cell-line
GSE40774,25,0,1,GPL13497,accepted,
GSE58194,26,0,1,GPL13376,rejected,xenograft_mouse
GSE29330,13,5,1,GPL570,rejected,cell-line
GSE55550,155,0,1,GPL17077,rejected,unspecified_samples_location
GSE55548,4,4,1,GPL17077,rejected,unspecified_samples_location
GSE55546,13,6,1,GPL17077,rejected,unspecified_samples_location
GSE55545,24,0,1,GPL17077,rejected,unspecified_samples_location
GSE55544,23,0,1,GPL17077,rejected,unspecified_samples_location
GSE55542,36,0,1,GPL17077,rejected,unspecified_samples_location
GSE39376,26,2,1,GPL201,rejected,cell-line
GSE34881,8,0,1,GPL4133,rejected,transfected
GSE45153,31,0,1,GPL570,rejected,mice_transplanted
GSE44170,5,0,1,GPL7280,accepted,

GSE39366,55,0,1,GPL9053,accepted,
GSE39368,55,0,1,GPL9053,rejected,duplicatedGSE39366
GSE33493,24,0,1,GPL6883,rejected,transfected
GSE30788,273,0,2,GPL13952,accepted,
GSE12472,63,1,GPL1708,rejected,lung
GSE13601,29,29,1,GPL8300,accepted,
GSE12428,31,31,1,GPL1708,rejected,lung
GSE10063,30,30,1,GPL570,rejected,cell-line
GSE3168,9,0,3,GPL127-GPL128-GPL129,rejected,cell-line
GSE2280,22,0,1,GPL96,accepted,
GSE2109,2,0,1,GPL570,rejected,inadequate_sample
GSE686,55,0,1,GPL503,rejected,unspecified_sample_location
,,,""
GSE111390,33,0,1,GPL6480,accepted,
GSE78060,26,4,1,GPL570,accepted,
GSE52915,27,0,1,GPL570,accepted,
GSE56142,13,0,1,GPL10558,accepted,

Παράρτημα 2 – Αποδεκτά στην πρώτη φάση datasets

"geo","oscc","normal","no_platforms","primary_platform","acception","info"
"GSE153918",16,0,"1","GPL13667","accepted","","
"GSE138206",6,12,"1","GPL570","accepted","contralateral-marginal"
"GSE74530",12,0,"1","GPL570","accepted","","
"GSE84846",99,0,"1","GPL6480","accepted","","
"GSE85195",35,1,"1","GPL6480","accepted","+15_leucoplakia"
"GSE59069",36,0,"1","GPL14555","accepted","","
"GSE56532",10,6,"1","GPL10739","accepted","","
"GSE51010",48,8,"2","GPL570-GPL201","accepted","Normal_GPL570-OSCC_GPL201"
"GSE46802",10,10,"2","GPL6480","accepted","+10Dysplasia"
"GSE37991",40,40,"1","GPL6883","accepted","","
"GSE41613",97,0,"1","GPL570","accepted","","
"GSE41116",43,0,"1","GPL5175","accepted","","
"GSE36090",10,3,"2","GPL2986","accepted","","
"GSE23558",27,4,"1","GPL6480","accepted","","
"GSE25104",57,22,"2","GPL5175","accepted","","
"GSE31056",23,73,"1","GPL10526","accepted","","
"GSE30784",167,45,"1","GPL570","accepted","+17Dysplasia"
"GSE4676",49,0,"1","GPL2891","accepted","","
"GSE3524",16,4,"1","GPL96","accepted","","
"GSE89146",40,0,"1","GPL4133","accepted","","
"GSE148944",38,6,"1","GPL28285","accepted","***784 probes***"
"GSE110996",15,0,"2","GPL570","accepted","","
"GSE108061",26,0,"1","GPL9128","accepted","","
"GSE103412",0,10,"1","GPL23978","accepted","only_10_normal-tonsil"
"GSE107591",16,16,"1","GPL6244","accepted","","

"GSE85446",66,0,"1","GPL6480","accepted","+Reanalysis_GSE30788"
"GSE65021",19,0,"1","GPL14951","accepted","","
"GSE75539",7,8,"1","GPL10904","accepted","","
"GSE40774",25,0,"1","GPL13497","accepted","","
"GSE39366",55,0,"1","GPL9053","accepted","","
"GSE30788",273,0,"2","GPL13952-GPL13853","accepted","","
"GSE13601",29,29,"1","GPL8300","accepted","","
"GSE2280",22,0,"1","GPL96","accepted","","
"GSE111390",33,0,"1","GPL6480","accepted","","
"GSE78060",26,4,"1","GPL570","accepted","","
"GSE52915",27,0,"1","GPL570","accepted","","
"GSE56142",13,0,"1","GPL10558","accepted","","

Απλός Κώδικας σε R από τον οποίο προέκυψαν τα παραπάνω

```
#create a "GEO_initial.txt" file with appendix 1
file="GEO_initial.txt"
df1=read.csv(file,header=T,skip=1,sep=",")
df2=df1[df1$acceptance=="accepted",]
df3=df2[df2$oscc+df2$normal>9,]
sum(df3$oscc)
sum(df3$normal)
GSE=df3$geo
write.table(df3,file = "accepted_samples.txt",sep=",",row.names = F)
```

Παράρτημα 3 - Ιδιότητες

Παράρτημα 4 – Ενδεικτικό παράδειγμα προεπεξεργασίας και τυποποίησης ενός συνόλου .CEL files με Affy

```
#-----CEL files preprocessing with Affy Package-----#  
  
#libraries to load, install those packages before loading  
  
library(affy)  
  
library(affyPLM)  
  
library(mygene)  
  
library(car)  
  
## INSTRUCTIONS #####  
  
#1.create a GSE specific directory called "GSE78060" inside working directory  
and put all the CEL files of a specific GEO dataset in there  
  
#-----#  
  
filename="GSE78060" # <-- set filename/GEO dataset ID  
  
#-----#  
  
#-----PREPARATION-----#  
#####  
  
files=list.files(path=paste0('./',filename),pattern="GSM.*")  
  
#Assess data from GEO in order to see which sample is cancer and which sample  
is normal
```



```
#Quality Assessment and QA boxplots, if there are more than 30 samples, more  
than one boxplot will be created for better visibility of the plots.
```

```

dataPLM=fitPLM(micro1)

nuse1=NUSE(dataPLM,type="values")

rle1=RLE(dataPLM,type="values")

if (nrow(pheno_df)>30){

  i=1

  j=1

  while ((i+29)<=length(properties)) {

    nuse2=nuse1[,i:(i+29)]

    rle2=rle1[,i:(i+29)]


    nuse_title=paste("NUSE",j,".png",sep="")

    png(file=paste(filename,nuse_title,sep="/"))

    boxplot(nuse2,col=2:4,main=paste("NUSE QUALITY
ASSESSMENT",filename,"samples",i,'-
',i+29),xaxt="n",ylim=c(0.9,1.1),xlab="samples",ylab="NUSE",outline=F)

    axis(1,1:30,labels = i:(i+29) ,las=1,cex.axis=0.7)

    abline(h=1.05,col="blue")

    dev.off()

  rle_title=paste("RLE",j,".png",sep = "")

  png(file=paste(filename,rle_title,sep="/"))

  boxplot(rle2,col=2:4,main=paste("RLE QUALITY
ASSESSMENT",filename,"samples",i,'-',i+29),xaxt='n',ylim=c(-
0.5,0.5),xlab="samples",ylab="RLE",outline=F)

  axis(1,1:30,labels=i:(i+29) ,las=1,cex.axis=0.7)

  abline(h=0,lty=2)

  dev.off()
}
}

```

```

i=i+30

j=j+1}

nuse_rest=nuse1[,i:length(properties)]
rle_rest=rle1[,i:length(properties)]

png(file=paste(filename,"NUSE-rest.png",sep="/"))

boxplot(nuse_rest,col=2:4,main=paste("NUSE QUALITY
ASSESSMENT",filename,"samples",i,'-
',length(nrow(pheno_df))),xaxt="n",xlab="samples",ylab="NUSE",outline=F,ylim=c
(0.9,1.1))

axis(1,1:(length(properties))-i+1),cex.axis=0.7,labels=i:length(properties),las=1)

abline(h=1.05,col="blue")

dev.off()

png(file=paste(filename,"RLE-rest.png",sep="/"))

boxplot(rle_rest,col=2:4,main=paste("RLE QUALITY
ASSESSMENT",filename,"samples",i,'-
',length(nrow(pheno_df))),xaxt='n',xlab="samples",ylab="RLE",outline=F)

axis(1,1:(length(properties))-i+1),cex.axis=0.7,labels=i:length(properties),las=1)

abline(h=0,lty=2)

dev.off()

}

if (nrow(pheno_df)<=30){

png(file=paste(filename,"NUSE.png",sep="/"))

boxplot(dataPLM,col=2:4,main=paste("NUSE QUALITY
ASSESSMENT",filename),xaxt="n",ylim=c(0.9,1.1),xlab="samples",ylab="NUSE")

```

```

axis(1,1:length(properties), las=1,cex.axis=0.6)

abline(h=1.05,col="blue")

dev.off()

png(file=paste(filename,"RLE.png",sep="/"))

Mbox(dataPLM,col=2:4,main=paste("RLE QUALITY
ASSESSMENT",filename),xaxt='n',ylim=c(-0.5,0.5),xlab="samples",ylab="RLE")

axis(1,1:length(properties), las=1,cex.axis=0.6)

abline(h=0,lty=2)

dev.off()

}

rm(dataPLM)

#-----
-----#####
#-----!!!! STOP to set rejected files otherwise set rejection=0 !!!!-
-----#####

#Assess RLE and NUSE diagrams and remove bad quality ones in this case samples
1,4,27 have a bad quality.

rejection=c(1,4,27) # <<---- SET rejected samples in a vector form e.g.
c(3,5)

#-----
-----#####
#-----NORMALIZATION-----#
#####
#####
```

```
#RMA and expressionSet object creation

if (all(rejection!=0)) {

  microRMA=rma(micro1[,-rejection])

  new_properties=properties[-rejection]

} else {

  microRMA=rma(micro1)

  new_properties=properties

}

rm(micro1)

micro_expmat=exprs(microRMA)

#Assigning column names according to modified properties(properties without
the rejected samples)

colnames(micro_expmat)=new_properties

cancer=which(new_properties=='cancer')

normal=which(new_properties=="normal")



if (length(cancer)>0 & length(normal)>0){

  micro_expmat_sorted=micro_expmat[,c(cancer,normal)]


  colnames(micro_expmat_sorted)=c(paste("C",1:length(cancer),sep=""),paste("N",1
  :length(normal),sep=""))} else {

  if (length(cancer)==0) {

    micro_expmat_sorted=micro_expmat

    colnames(micro_expmat_sorted)=paste("N",1:length(normal),sep="")

  if (length(normal)==0) {

    micro_expmat_sorted=micro_expmat

    colnames(micro_expmat_sorted)=paste("C",1:length(cancer),sep="")}
```

}

```
# Appointing exp_mat as the normalized Expression Matrix variable and removing redundant variables

exp_mat=micro_expmat_sorted

rm(micro_expmat,micro_expmat_sorted,microRMA)

#-----GENES ANNOTATION-----
#####
#####mygene query

library(mygene)

genesid=queryMany(rownames(exp_mat),scopes="reporter",species="human",returnal
l=T,fields="entrezgene")

entrez=genesid$response

entrez1=entrez$entrezgene

#matching and aggregation of annotations for dealing with "Many to Many"
effect

matching=exp_mat[match(entrez$query,rownames(exp_mat)),]

annot_mat=cbind(entrez=as.character(entrez1),as.data.frame(matching))

rownames(annot_mat)=NULL

standard_mat=aggregate(annot_mat[,2:ncol(annot_mat)],by=list(annot_mat$entrez)
,function(x) max(x,na.rm=T))

colnames(standard_mat)[1]="entrez"

standard_mat=standard_mat[order(as.numeric(standard_mat$entrez)),]

rownames(standard_mat)=NULL
```

```
#-----FINAL STEPS of preprocessing-----#
#####
#visualization of final results with boxplot, sample 3 histogram and Q-Q plot

library(car) # improved qqplot graphics library

png(file=paste(filename,"_box_normalized.png",sep="/"))

boxplot(standard_mat[,-1],col=2:4,main=paste(filename,"normalized"),xlab="samples")
dev.off()

png(file=paste(filename,"_qq_normalized.png",sep="/"))

qqPlot(standard_mat[,3],main=paste(filename,"normalized"),ylab="sample 3")
dev.off()

png(file=paste(filename,"_hist_normalized.png",sep="/"))

hist(standard_mat[,3],col=2,main=paste(filename,"normalized"),xlab="sample 3")
dev.off()

#write series_matrix.txt table in the GEO file subfolder

checking_file=paste(filename,"_series_matrix.txt",sep="")

write.table(standard_mat,file=paste(filename,checking_file,sep="/"),row.names = F,sep="\t")

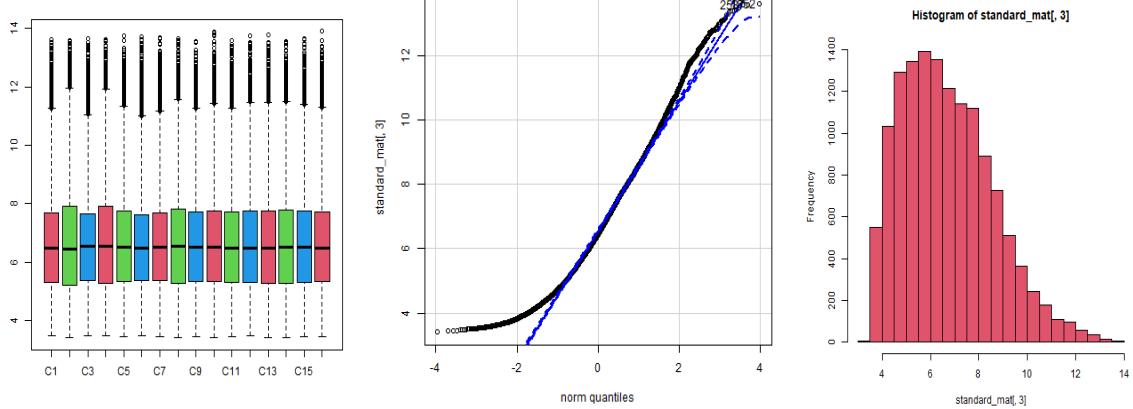
#testing - final check of series_matrix.txt file

final_check=read.csv(paste(filename,checking_file,sep="/"),header=T,sep="\t")

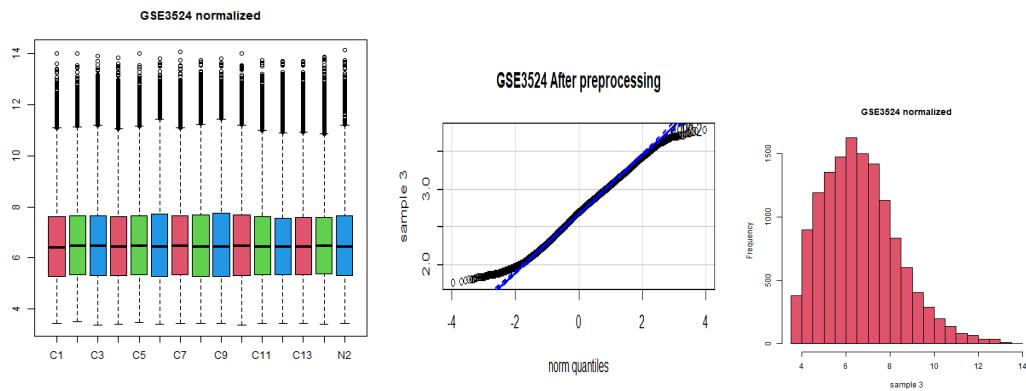
head(final_check)
```

Παράρτημα 5 – Γραφικά αποτελέσματα επεξεργασίας των CEL αρχείων

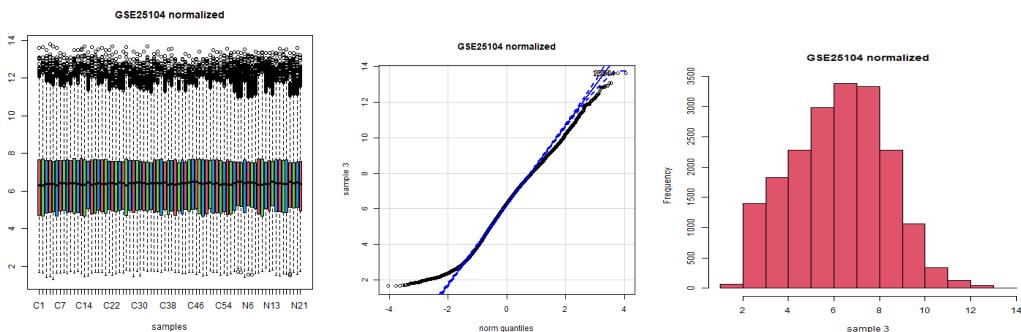
GSE2280



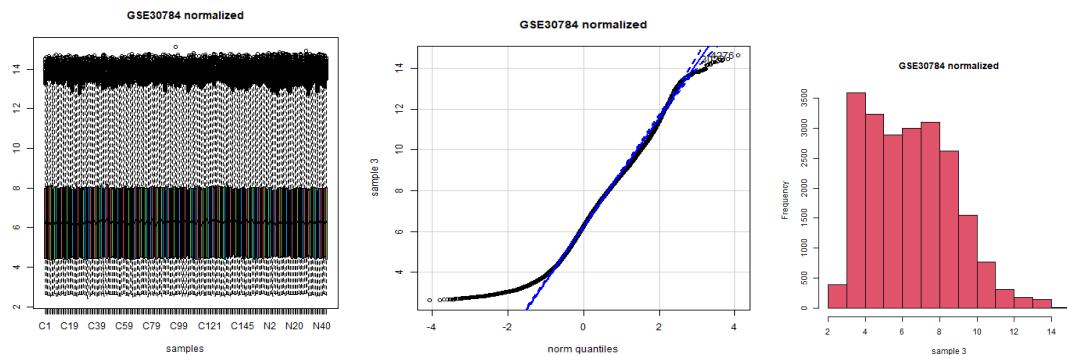
GSE3524



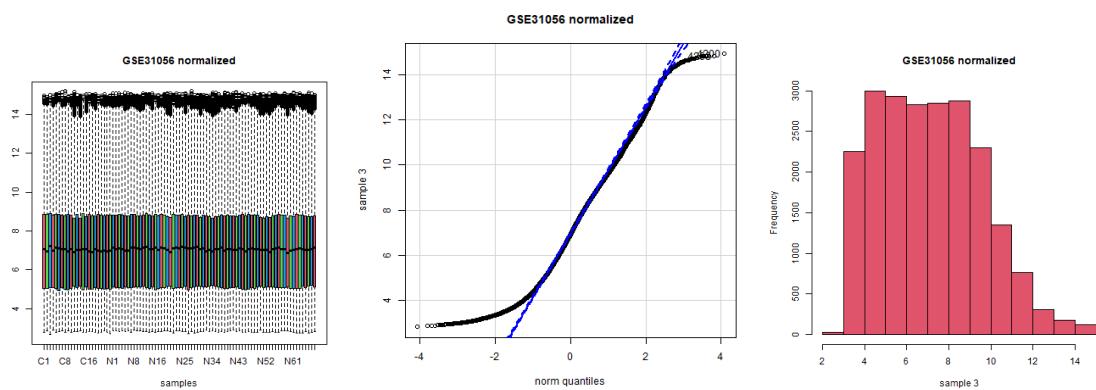
GSE25104



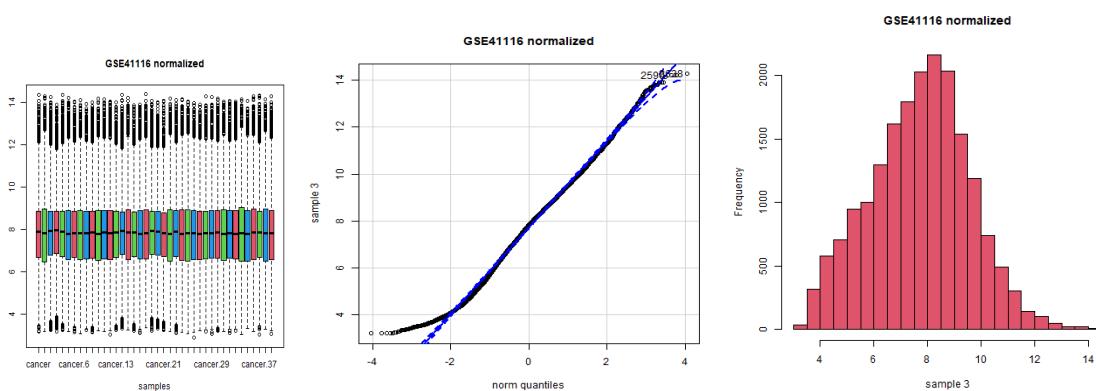
GSE30784



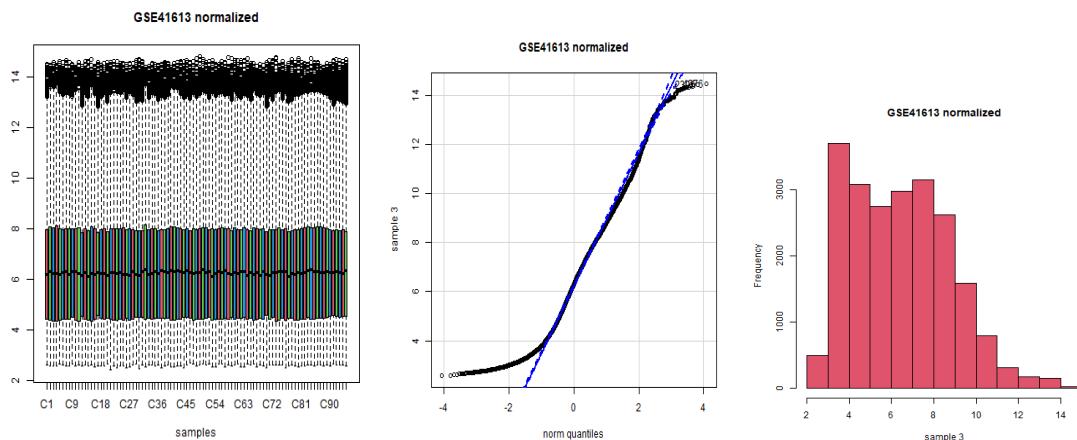
GSE31056



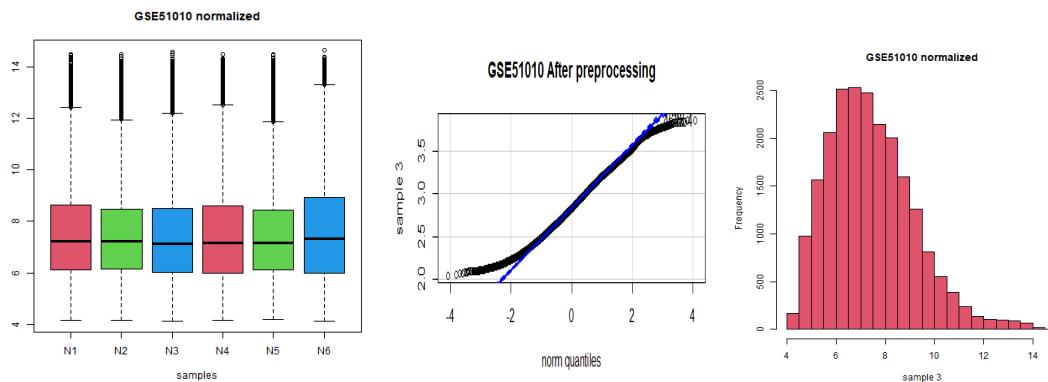
GSE41116



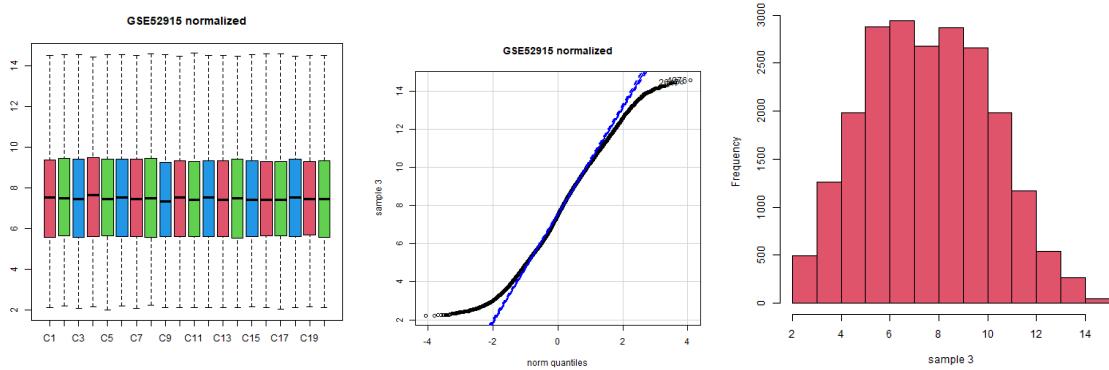
GSE41613



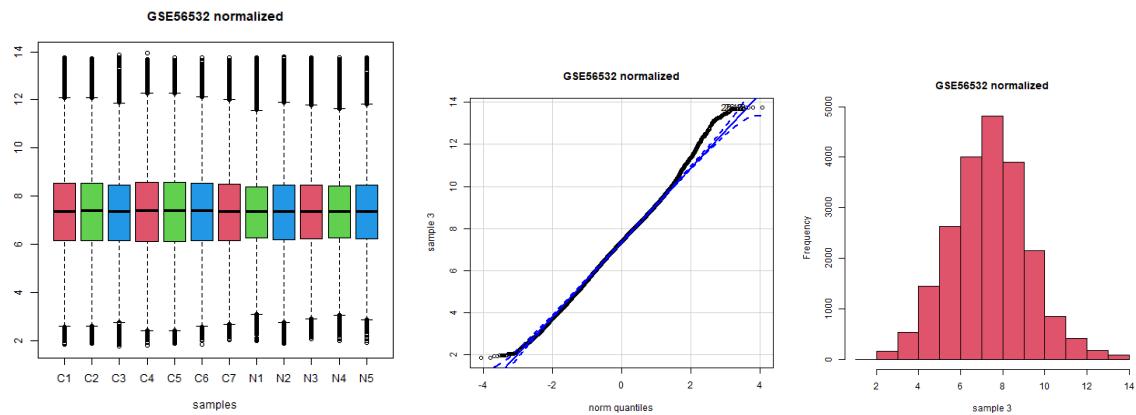
GSE51010



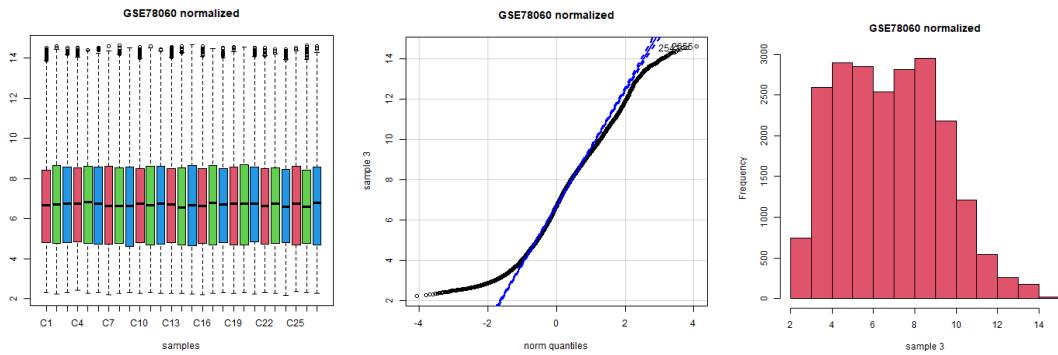
GSE52915



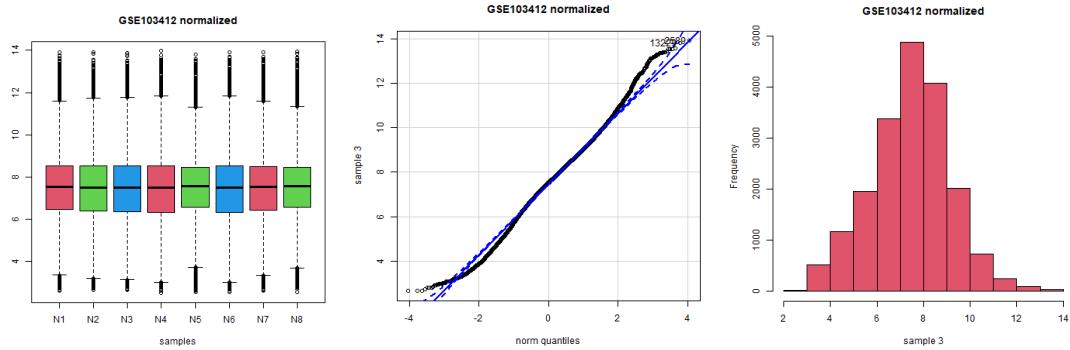
GSE56532



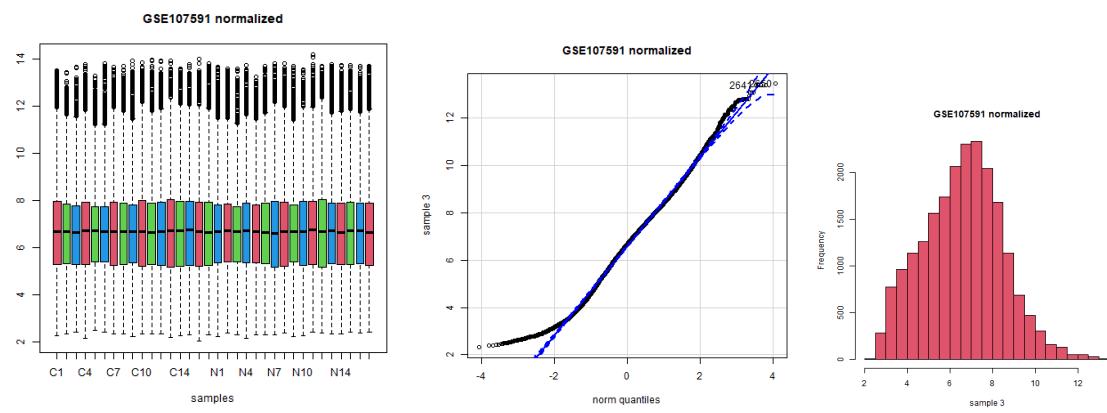
GSE78060



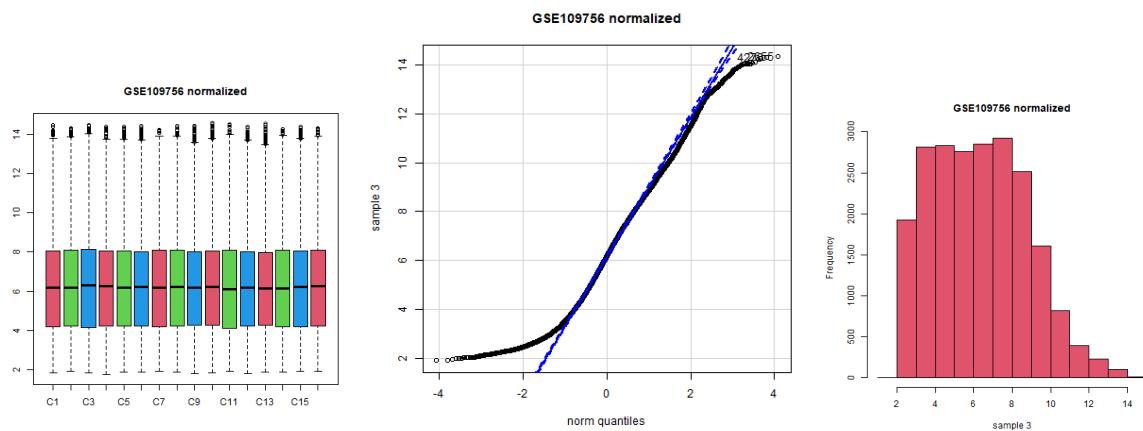
GSE103412



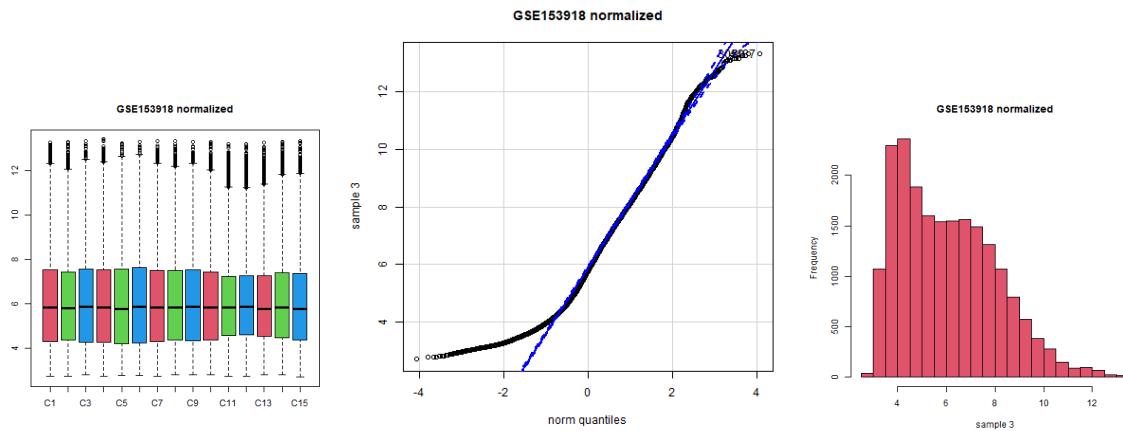
GSE107591



GSE109756



GSE153918



GSE74530 (σετ που έγινε δεκτό αρχικά(κάτω διαγράμματα των series matrix) αλλά απορρίφθηκε μετά την γραφική απεικόνιση μετά από RMA των CEL δεδομένων του)

