ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ, ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ

ΜΟΡΙΑΚΗ ΒΙΟΛΟΓΙΑ ΚΑΙ ΒΙΟΪΑΤΡΙΚΗ

ΚΑΙ

ΙΔΡΥΜΑ ΤΕΧΝΟΛΟΓΙΑΣ ΚΑΙ ΕΡΕΥΝΑΣ

ΙΝΣΤΙΤΟΥΤΟ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ ΚΑΙ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ

ΥΠΟΛΟΓΙΣΤΙΚΗ ΠΡΟΒΛΕΨΗ ΓΟΝΙΔΙΩΝ ΤΑΞΙΝΟΜΗΤΩΝ ΚΑΙ MICRORNA ΣΤΟΝ ΚΑΡΚΙΝΟ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΑΝΑΣΤΑΣΗΣ ΟΥΛΑΣ

ΗΡΑΚΛΕΙΟ 2009

Στοιχεία μελών εξεταστικής επιτροπής

ΠΟΪΡΑΖΗ ΠΑΝΑΓΙΩΤΑ (Επιβλέπουσα Ερευνήτρια ΙΜΒΒ)

ΠΑΠΑΜΑΤΘΑΙΑΚΗΣ ΙΩΣΗΦ (Επιβλέπων Καθηγητής)

ΛΟΥΗΣ ΧΡΗΣΤΟΣ (Καθηγητής)

ΗΛΙΟΠΟΥΛΟΣ ΑΡΙΣΤΕΙΔΗΣ (Αναπλ. Καθηγητής Τμ. Ιατρικής)

ΚΑΡΔΑΣΣΗΣ ΔΗΜΗΤΡΗΣ (Αναπλ. Καθηγητής Τμ. Ιατρικής)

ΤΟΛΛΗΣ ΙΩΑΝΝΗΣ (Καθηγητής Τμ. Επιστ. Υπολογιστών)

ΚΑΛΑΝΤΙΔΗΣ ΚΡΙΤΩΝΑΣ (Επικ. Καθηγητής)

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Ευχαριστίες

Θα ήθελα να ευχαριστήσω ένα μεγάλο αριθμό ανθρώπων για την υποστήριξη και βοήθεια τους κατά τη διάρκεια της διδακτορικής μου διατριβής. Ένα πολύ μεγάλο ευχαριστώ στην επιβλέπουσα ερευνήτρια της διδακτορικής μου διατριβής Γιώτα Ποϊράζη, για την εμπιστοσύνη που μου έδειξε στο ξεκίνημα της ερευνητικής μου πορείας καθώς επίσης και για τις αξιόλογες και χρήσιμες συμβουλές στο συγκεκριμένο επιστημονικό πεδίο. Θα ήθελα επίσης να την ευχαριστήσω για την ενθάρρυνση της να συμμετάσχω σε συνέδρια και σχολεία καθώς και τη ηθική υποστήριξη της σε όλα τα χρόνια της διδακτορικής μου διατριβής όπως επίσης και για την εξασφάλιση της διδακτορικής υποτροφίας μου μέσω του ΠΕΝΕΔ 03Ε_842.

Thank you Yiota!!! For everything!!!

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή και μέλος της τριμελούς επιτροπής της διδακτορικής μου διατριβής, Ιωσήφ Παπαματθαιάκη για το ενδιαφέρον που έδειξε και της καίριες παρατηρήσεις του κατά τη διάρκεια των συμβουλευτικών επιτροπών. Ένα μεγάλο ευχαριστώ στον καθηγητή Χρήστο Λούη και μέλος της τριμελούς επιτροπής για τις ουσιαστικές παρεμβάσεις και τη διάρκεια των αξιολογήσεων προόδου.

Την εκτίμηση και ευγνωμοσύνη μου θα ήθελα να εκφράσω στον επίκουρο καθηγητή Κρίτωνα Καλαντίδη και μέλος της εξεταστικής επιτροπής της διδακτορικής μου διατριβής, την μεταδιδακτορικό συνεργάτη Αλεξάνδρα Μπούτλα καθώς επίσης και τα υπόλοιπα μέλη του εργαστηρίου του για την πολύ πετυχημένη συνεργασία μας που οδήγησε σε μια πολύ αξιόλογη δημοσίευση.

Τις ευχαριστίες μου θα ήθελα να εκφράσω και στα υπόλοιπα μέλη της εξεταστικής επιτροπής της διδακτορικής μου διατριβής, Αριστείδη Ηλιόπουλο, Γιάννη Τόλλη και Δημήτρη Καρδάσση.

Επίσης ευχαριστώ τον Δρ. Δημήτρη Καφετζόπουλο και το πρόγραμμα Prognochip για την οικονομική υποστήριξη των διδακτορικών μου ερευνών.

Ένα τεράστιο ευχαριστώ στον Dr. Martin Reczko για τις πολύτιμες συμβουλές πάνω σε ερευνητικά θέματα καθώς επίσης και την αξιόλογη συνεργασία που είχαμε όλα αυτά τα χρόνια. Thanks Martin!!

Στον στενό συνεργάτη και φίλο από το Πανεπιστήμιο του Cambridge, Λώρενς Π. Πεταλίδη, θα ήθελα να εκφράσω την ευγνωμοσύνη μου για την εμπιστοσύνη του και την καθοδήγηση του στην πορεία μιας πολύ αξιόλογης δημοσίευσης.

Θερμά ευχαριστώ τους φίλους και μέλη του εργαστήριου μας, Μαρία Μανιουδάκη, Κική Σιδηροπούλου, Ελευθερία Τζαμαλή, Νάση Παπουτσή, Ξένια Κωνσταντουδάκη, Σίσση Τζάνου, Κατερίνα Γκίρτζου, Νέστορα Καραθανάση, για την ηθική τους υποστήριξη και όχι μόνο. Ένα ιδιαίτερο ευχαριστώ στην Μαρία Μανιουδάκη για τη κατανόηση και τεράστια βοήθεια κατά τη διάρκεια της διδακτορικής μου διατριβής. Thanks!!! Επίσης θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες σε παλιά μέλη του εργαστηρίου μας και του ICS και αγαπητούς φίλους, τον Παύλο Παυλίδη, την Ελευθερία Πισσαδάκη, την Αγάπη Σταματάκη και τον Αλέξη Σταματάκη για τις απίστευτες εκδρομές και ταξίδια σε όλη την Κρήτη, όπως επίσης και την Κάσση Μπαλάφα, τον Θοδωρή Πάτκο και τον Αντώνη Μπικάκη για την ανεκτίμητη βοήθεια τους. Στους πιό νέους φίλους και ελπίζω μελλοντικούς συνεργάτες, Γιάννη Ηλιόπουλο, Νίκο Παπανικολάου και Βαγγέλη Παφίλη θα ήθελα να πω ένα μεγάλο ευχαριστώ για την υποστήριξη που μου έδωσαν τους τελευταίους μήνες της διδακτορικής μου διατριβής. Επίσης δεν θα μπορούσα να παραλείψω να ευχαριστήσω τον κολλητό φίλο Γιάννη Κρητικό (Thanks brother!).

Καταλήγοντας θα ήθελα να ευχαριστήσω του γονείς μου Φοίβο και Γιολάντα και την αδερφή μου Αιμιλίανα για την αγάπη τους. Τέλος ένα απέραντο ευχαριστώ και όλη μου την αγάπη στην Σοφία Τσάδαρη. Thank you Sophia mou, could not have done it without you!!! Lots of love!!!

Περίληψη

Η παρούσα διατριβή συνδυάζει υπολογιστικές και πειραματικές προσεγγίσεις για τον εντοπισμό μορίων που εμπλέκονται σε διάφορα είδη καρκίνου. Συγκεκριμένα, η διατριβή εστιάζεται στην ανάπτυξη μεθόδων μηχανικής μάθησης και την εφαρμογή τους σε βιολογικά δεδομένα με στόχο:

(α) να παρουσιάσει στοιχεία που υποστηρίζουν μια νέα κατηγοριοποίηση για συγκεκριμένους τύπους καρκίνου του εγκεφάλου, και να ταυτοποιήσει πιθανούς καρκινικούς γονιδιακούς δείκτες

(β) να προσδιορίσει νέα miRNA γονίδια που εμπλέκονται στη διαδικασία ανάπτυξης καρκινικών όγκων και

(γ) να συμβάλλει στη βελτιστοποίηση της πρόβλεψης στόχων των miRNA.

Ο πρώτος στόχος είχε ως κίνητρο το γεγονός ότι η ιστοπαθολογική διαβάθμιση των αστροκυττωματικών όγκων που βασίζεται στα τρέχοντα κριτήρια του Παγκόσμιου Οργανισμού Υγείας (ΠΟΥ) είναι υπερ-απλουστευμένη, και συχνά ανεπαρκής στην πρόβλεψη του κλινικού φαινότυπου. Για την επίτευξη του στόχου αυτού αναπτύχθηκε ένα Τεχνητό Νευρωνικό Δίκτυο (ΤΝΔ) και εφαρμόστηκε σε ένα νέο σύνολο δεδομένων γονιδιακής έκφρασης από αστροκυττωματικούς όγκους (n=65) τα οποία συνοδεύονταν από ιστοπαθολογικά και κλινικά στοιχεία. Η μελέτη αυτή οδήγησε στην επιτυχημένη διαβάθμιση ανθρώπινων αστροκυττωματικών όγκων με τη χρήση του προφίλ έκφρασης επιλεγμένων γονιδίων (πιθανοί καρκινικοί γονιδιακοί δείκτες), την εξαγωγή συγκεκριμένων μεταγραφικών αποτυπωμάτων από ιστοπαθολογικούς υπότυπους αστροκυττωματικών όγκων και τον καθορισμό προγνωστικών υποκατηγοριών επιβίωσης με την χρήση των μοριακών αυτών αποτυπωμάτων.

Τα miRNA έχουν πρόσφατα συσχετιστεί με διάφορα είδη καρκίνου. Για τον εντοπισμό νέων miRNA που εμπλέκονται σε καρκινικές διαδικασίες στον άνθρωπο αναπτύχθηκε ένα υπολογιστικό εργαλείο που χρησιμοποιεί *Profile* Κρυφά Μαρκοβιανά Μοντέλα (HMM) και συνδυάζει πληροφορίες αλληλουχίας, δομής και συντήρησης επιτυγχάνοντας υψηλή ακρίβεια πρόβλεψης miRNA γονιδίων. Το εργαλείο εφαρμόστηκε σε καρκινικά σχετιζόμενες γενωμικές περιοχές (CAGR) και τέσσερα από τα υποψήφια miRNA γονίδια τα οποία είχαν υψηλή έκφραση σε δεδομένα από tiling arrays επιβεβαιώθηκαν πειραματικά χρησιμοποιώντας ανάλυση Northern blot. Η μελέτη αυτή οδήγησε στην δημιουργία ενός αξιόπιστου εργαλείου

για τον εντοπισμό νέων υποψήφιων miRNA γονιδίων καθώς και την ανακάλυψη τεσσάρων τέτοιων μορίων τα οποία πιθανά να παίζουν ρόλο στην καρκινογένεση.

Επόμενο βήμα για την διερεύνηση αυτού του ρόλου είναι ο εντοπισμός πιθανών στόχων. Τα περισσότερα υπάρχοντα υπολογιστικά εργαλεία είναι πολύ αποτελεσματικά στην πρόβλεψη πραγματικών στοχευμένων περιοχών (υψηλή ευαισθησία) αλλά ταυτόχρονα επιδεικνύουν έναν εξαιρετικά μεγάλο αριθμό συνολικών προβλέψεων (χαμηλή ειδικότητα). Αντίθετα, άλλα εργαλεία επιδεικνύουν συνολικά υψηλή ειδικότητα αλλά χαμηλή ευαισθησία. Για την αντιμετώπιση του προβλήματος αυτού αναπτύχθηκε ένα πιο εξελιγμένο εργαλείο (*Targetprofiler*) πρόβλεψης στόχων που επιτυγχάνει μια καλύτερη ισορροπία μεταξύ ευαισθησίας και ειδικότητας σε σύγκριση με τα υπάρχοντα εργαλεία.

Συμπερασματικά, τα αποτελέσματα της παρούσα διατριβής συνεισφέρουν σημαντικά στην προσπάθεια κατανόησης των μοριακών μηχανισμών της κακρινογένεσης ενώ ταυτόχρονα παρέχουν αξιόπιστα και ελεύθερα διαθέσιμα υπολογιστικά εργαλεία.

Κεφάλαιο 1

1 Γενική Εισαγωγή

1.1 Χρήση μικροσυστοιχιών DNA για την ταξινόμηση καρκινικών δειγμάτων και την ταυτοποίηση γονιδιακών ταξινομητών

1.1.1 Θεωρητικό Υπόβαθρο

Κατά την προηγούμενη δεκαετία, η επιστημονική έρευνα στο πεδίο της υπολογιστικής πρόβλεψης και μοντελοποίησης των μοριακών μονοπατιών του καρκίνου παρουσίασε συνεχή ανάπτυξη. Η ανάλυση της έκφρασης των γονιδίων με μικροσυστοιχίες DNA άσκησε ιδιαίτερη επίδραση στο πεδίο της ταξινόμησης και διάγνωσης του καρκίνου. Καθώς οι μικροσυστοιχίες εμπλουτίζονται με νέα γονίδια/ανιχνευτές και η τεχνολογία βελτιώνεται, αναμένεται ότι οι εν λόγω δομές θα συνεισφέρουν καταλυτικά στην κατανόηση των ανθρώπινων μοριακών προφίλ.

Οι μικροσυστοιχίες DNA έδωσαν την ευκαιρία στους ερευνητές να πραγματοποιήσουν την ταξινόμηση πολλών τύπων καρκίνου βάση της γονιδιακής έκφρασης, όπως επί παραδείγματι του καρκίνου του μαστού (1-6), του εγκεφάλου (7,8), των ωοθηκών (9-11), των πνευμόνων (12,13), του παχέως εντέρου (14), του προστάτη (15), του στομάχου (16) διαφόρων ειδών λευχαιμίας (17-19) και λεμφώματος (20,21). Μερικές λειτουργικές ομάδες γονιδίων φαίνεται ότι τροποποιούνται συστηματικά με συνέπεια τα υγιή κύτταρα να μετατρέπονται σε κακοήθη. Σε αυτά συμπεριλαμβάνονται γονίδια τα οποία εμπλέκονται στον πολλαπλασιασμό, την πρόσφυση (adhesion), την αγγειογένεση και την απόπτωση των κυττάρων (22). Κατά συνέπεια, παρά την μορφολογική και μοριακή ετερογένεια η οποία υφίσταται στους εκάστοτε καρκινικούς τύπους, υπάρχουν ομοιότητες εξαιτίας των οποίων κάποιοι γονιδιακοί δείκτες χαρακτηρίζουν μία ευρύτερη κατηγορία καρκίνων. Η κατανόηση της ογκολογικής ετερογένειας, η οποία ενυπάρχει τόσο μεταξύ όσο και εντός των διαφορετικών τύπων όγκων, ίσως συνιστά την πιο ισχυρή πρόκληση στη διάγνωση και την αντιμετώπιση του καρκίνου. Οι μικροσυστοιχίες συχνά αποδεικνύονται αποτελεσματικές στην διάκριση των επιπέδων ετερογένειας

έναντι αυτών που προβλέπονται μέσω της ιστοπαθολογικής ανάλυσης αφού το γονιδιακό προφίλ έκφρασης συχνά διαφέρει σε περιπτώσεις που τα ιστοπαθολογικά ευρήματα είναι πανοποιότυπα. Άρα, αναμένεται ότι ένα εξειδικευμένο σχήμα ταξινόμησης το οποίο θα βασίζεται σε «αποτυπώματα» έκφρασης γονιδίων πιθανά να παρέχει ακριβέστερη πρόγνωση για την εμφάνιση/επέκταση της νόσου καθώς και καλύτερη πρόβλεψη της απόκρισης κάθε ασθενούς στη προτεινόμενη θεραπευτική αγωγή.

1.1.2 Επιβλεπόμενη και Μη-επιβλεπόμενη Ταξινόμηση

Στην παρούσα μελέτη, η ταξινόμηση των όγκων θα προσεγγισθεί χρησιμοποιώντας δύο μεθοδολογίες μηχανικής μάθησης. Η πρώτη αφορά στην «επιβλεπόμενη» ανάλυση (supervised analysis), ή διαφορετικά εκμάθηση με διδασκαλία, στην οποία αναζητούνται γονίδια που τα προφίλ έκφρασής τους συσχετίζονται με κάποια εξωτερική παράμετρο (π.χ. στάδιο της νόσου). Οι εξωτερικές παράμετροι που χρησιμοποιούνται συνήθως προέρχονται από τα ιστοπαθολογικά ευρήματα όπως παραδείγματος χάριν η ύπαρξη μεταστάσεων, η χρονική διάρκεια επιβίωσης και ο βαθμός ανταπόκρισης στη θεραπεία. Αλγόριθμοι όπως οι *weighted-voting* (23), *k-nearest-neighbour* (15), *support vector machines* (24,25) και τα τεχνητά νευρωνικά δίκτυα (*artificial neural networks)* (5,26) μπορούν να εφαρμοστούν σε ένα σύνολο γονιδίων προκειμένου να υλοποιηθούν μοντέλα ταξινομητές ικανά να προβλέπουν την κατηγορία συγκεκριμένων δειγμάτων (π.χ. καρκινικά ή υγιεί). Η αξιοπιστία του μοντέλου ταξινομητή εκτιμάται συχνά με την μέθοδο ανάλυσης *leave-one-out cross-validation* (2,7,15,18,25), κατά την οποία ένα δείγμα από το σύνολο των δειγμάτων εκπαίδευσης παραλείπεται επαναληπτικά και στη συνέχεια το μοντέλο προβλέπει την κατηγορία στην οποία ανήκει το παραλειπόμενο δείγμα. Η δεύτερη μεθοδολογία αφορά την «μη-επιβλεπόμενη» μάθηση, κατά την οποία δεν χρησιμοποιείται καμία εξωτερική παράμετρος προκειμένου να καθοδηγήσει την διαδικασία της κατηγοριοποίησης. Αντιθέτως, τα δεδομένα χρησιμοποιούνται για την εύρεση μοτίβων (π.χ. ομοιοτητα στο προφιλ έκφρασης γονιδίων) χωρίς *a priori* εκτίμηση των βέλτιστων αποτελεσμάτων. Η ιεραρχική ομαδοποίηση, *hierarchical clustering* (27), αποτελεί την συχνότερη μη-επιβλεπόμενη μέθοδο οργάνωσης δειγμάτων σε ομάδες με βάση κάποιο κριτήριο ομοιότητας. Ωστόσο μέθοδοι ανάλυσης όπως: K-means (28), K-medians (29), SOMs (Self Organized Maps) (30), Principal Component Analysis

(PCA) (31) και SOTA (Self Organizing Tree Algorithms) (32) έχουν χρησιμοποιηθεί. Κάθε αναλυτική μέθοδος έχει τις δικές της αδυναμίες και πλεονεκτήματα, κατά συνέπεια η απόκτηση ενός αξιόπιστου ευρήματος απαιτεί την επαλήθευση του με τη χρήση ενός πλήθους μεθόδων.

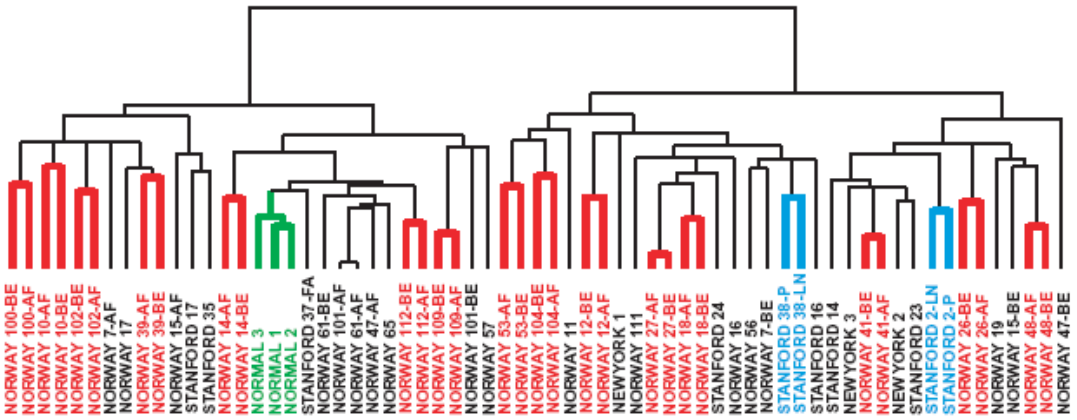
1.1.3 Επιλογή Γονιδίων-Ταξινομητών

Η **επιλογή γονιδίων ταξινομητών** αναφέρεται στον εντοπισμό γονιδίων των οποίων το προφίλ έκφρασης διαφέρει συστηματικά ανάμεσα στις κατηγορίες που μας ενδιαφέρουν, π.χ. σε υγιεί και καρκινικά δείγματα. Επειδή οι μικροσυστοιχίες περιλαμβάνουν χιλιάδες γονίδια, πολλά εκ των οποίων δεν προσφέρουν κάποια πληροφορία όσον αφορά την κατηγοριοποίηση δειγμάτων, το φιλτράρισμα αυτό είναι μία λεπτή αλλά και ιδιαίτερα σημαντική και απαραίτητη διαδικασία. Εντοπίζοντας τα 'πληροφοριακά' γονιδία μπορεί κανείς να διευκολύνει την κατανόηση του υπό μελέτη βιολογικού φαινομένου και πιθανά να προτείνει τη χρήση των γονιδίων αυτών (γονίδια ταξινομητές) σε κλινικές εξετάσεις. Πολλές στατιστικές μέθοδοι έχουν επιτυχώς χρησιμοποιηθεί σε επιβλεπόμενες μελέτες, συμπεριλαμβάνοντας μετρικές συσχέτισης, TNOM (33), SAM (34), το *t*-test και τη μέθοδο *signal-to-noise* (7,15,21) για την επιλογή ή αναγνώριση γονιδίων ταξινομητών. Δοκιμές αντιμετάθεσης *(permutation testing)* (2,17,18) συχνά χρησιμοποιούνται για να αποδώσουν στατιστικά σημαντικά *p*-values. Σε πολλές περιπτώσεις η επιλογή γονιδίων ταξινομητών συνδέεται με την εκμάθηση του μοντέλου, όπου επιλέγεται η λίστα γονιδίων η οποία φέρει τη μέγιστη ακρίβεια πρόβλεψης κατά τη διαδικασία *leave-one-out cross-validation* (2,15,18,21). Για την επίτευξη πλήρους επαλήθευσης, η λίστα των γονιδίων υπόκειται σε δεύτερο έλεγχο σε δεδομένα δοκιμής (γενίκευσης) τα οποία δεν έχουν συμπεριληφθεί στη διαδικασία εκμάθησης, (*blind test set*).

1.1.4 Πρόβλεψη ενός Μοναδικού «Αποτυπώματος» Έκφρασης για κάθε ασθενή

Οι ασθενείς που πάσχουν από καρκίνο θεωρούνται μοναδικοί ως προς τη γενετική έκφραση της ασθένειας και το περιβάλλον στο οποίο ζουν. Η υπόθεση ότι κάθε όγκος αποτελεί μία μοναδική περίπτωση προτάθηκε από τους *Perou et al 2000* (35) κατά

την ταξινόμηση περιπτώσεων καρκίνου του στήθους. Ειδικότερα, η επαναληπτική βιοψία από τον ίδιο όγκο, πριν και μετά την χημειοθεραπεία, ή ως ζεύγος όγκου χωρίς μετάσταση και όγκου με μετάσταση, έδειξε ότι τα συγκεκριμένα ζεύγη είχαν πολύ μεγαλύτερη ομοιότητα μεταξύ τους, συγκρινόμενα με οποιοδήποτε άλλο όγκο της μελέτης (35) με βάση το γονιδιακό προφίλ έκφρασης. Το σημαντικό αυτό αποτέλεσμα έδειξε ότι τα τεχνικά χαρακτηριστικά των μικροσυστοιχιών είναι αξιόπιστα και περεταίρω υποδεικνύουν ότι είναι όντως πιθανό να αποκτήσουμε ένα μοναδικό γονιδιακό προφίλ έκφρασης ή «αποτύπωμα» ενός όγκου (Σχήμα 1.1).



Σχήμα 1.1 Δενδρόγραμμα Ιεραρχικής ομαδοποίησης βασιζόμενο σε προφίλ έκφρασης. Όλα τα ζευγάρια «πριν και μετά» την χημειοθεραπεία ομαδοποιούνται μαζί (κόκκινο χρώμα), καθώς επίσης και τα 2 ζευγάρια των όγκων με μετάσταση στους λεμφαδένες (μπλε χρώμα), και τα 3 δείγματα από μη-καρκινικό ιστό του μαστού (πράσινο χρώμα).

Η μοναδικότητα μεμονωμένων όγκων, όπως αυτή αναλύθηκε από μεθόδους ιεραρχικής ομαδοποίησης και άλλες μεθόδους συσχέτισης έχει πλέον επιβεβαιωθεί σε άλλους τύπους όγκων όπως: πνευμονικών, ήπατος και *diffuse large B-cell lymphomas* (6,12,20,36).

Πέραν της μοναδικότητάς τους, οι όγκοι παρουσιάζουν πολλές ομοιότητες οι οποίες αποδίδονται στον κυτταρικό τύπο από τον οποίο προέρχονται. Σε πρόσφατη μελέτη, οι Pomeroy *et al.* (7), εξέτασαν 99 δείγματα όγκων του κεντρικού νευρικού

συστήματος και εντόπισαν έναν μοριακό ταξινομητή *(descriptor)* χρησιμοποιώντας τη μέθοδο *Principal Component Analysis (PCA)* και 50 γονίδια. Ο ταξινομητής αυτός διαφοροποίησε με επιτυχία όγκους από *medulloblastomas* από ιστολογικά παρόμοιους όγκους του εγκεφάλου.

Ίσως μία από τις μεγαλύτερες προκλήσεις στην ιατρική του καρκίνου αποτελεί η αναγνώριση εκείνων των όγκων οι οποίοι είναι κλινικά διακριτοί, ωστόσο μη διαχωρίσιμοι με βάσει την ιστοπαθολογική τους εικόνα. Για παράδειγμα, επιθετικού τύπου καρκίνος του στήθους εμφανίζεται ομογενής κατά την ιστοπαθολογική εξέταση αλλά παρουσιάζει ετερογενή κλινική συμπεριφορά. Σε μελέτη με επίκεντρο τον επιθετικό καρκίνο του στήθους *(ductal breast carcinomas),* τουλάχιστον πέντε διακριτοί υπότυποι όγκων αναγνωρίστηκαν με τη χρήση προφίλ γονιδιακής έκφρασης (1,6). Η κατηγοριοποίηση με ιεραρχική ομαδοποίηση διαχώρισε τα δείγματα σε δύο μεγάλες ομάδες: τις ομάδες θετικές ως προς τον υποδοχέα οιστρογόνου *(ER)* και σε εκείνες αρνητικές για *ER* (1).

Αρκετή έρευνα έχει επίσης πραγματοποιηθεί στην ταξινόμηση καρκίνου του πνεύμονα με τη χρήση μικροσυστοιχιών (12,13). Δύο μελέτες στις οποίες χρησιμοποιήθηκαν διαφορετικές πλατφόρμες μικροσυστοιχιών (Affymetrix and cDNA microarrays) εντόπισαν παρόμοια μοτίβα έκφρασης τα οποία αναπαριστούν με ακρίβεια τους τέσσερις ιστοπαθολογικούς υπότυπους του καρκίνου του πνεύμονα (12,37). Επιπλέον, ορισμένες περιπτώσεις αστροκυττωματικών όγκων παρουσιάζουν παρόμοιο ιστοπαθολογικό προφίλ με αυτό της χαμηλότερης βαθμίδας αστροκυττώματος, αλλά έχουν την κλινική εικόνα και το προφίλ έκφρασης μίας υψηλότερης βαθμίδας αστροκυττώματος (βλ. κεφάλαιο 2).

Το προσδόκιμο ζωής ή χρόνος επιβίωσης είναι μία άλλη παράμετρος που χρησιμοποιείται συχνά προκειμένου να ανιχνευθεί η διαφορά μεταξύ χαρακτηριστικών υπο-ομάδων καρκίνου όπως αυτές διαμορφώνονται από το προφίλ έκφρασης. Οι Shipp *et al* (21) πραγματοποίησαν μία μελέτη *large B-cell lymphomas (DLBCLs)* με *Affymetrix* μικροσυστοιχίες, από τις οποίες ανέπτυξαν έναν προγνωστικό αποτύπωμα από 13 γονίδια βασισμένο σε κλινικά χαρακτηριστικά των ασθενών. Τα άτομα με DLBCL διαιρέθηκαν σε δύο ομάδες: στην ομάδα των ατόμων οι οποίοι θεραπεύτηκαν (n = 32), και στην ομάδα των ατόμων με θανάσιμη ή επανερχόμενη νόσο (n = 26). Οι καμπύλες επιβίωσης Kaplan–Meier και οι υπολογισμοί με το *log-rank test* έδειξαν ότι τα άτομα της πρώτης ομάδας είχαν σημαντικά μεγαλύτερο χρόνο επιβίωσης από τα άτομα της δεύτερης ομάδας

(συνολικός χρόνος επιβίωσης: πέντε χρόνια, 70% έναντι 12%). Οι Beer *et al* (13), επίσης ταυτοποίησαν ομάδες πρόγνωσης εντός των ομάδων με αδενοκαρκίνωμα, χρησιμοποιώντας το προσδόκιμο ζωής ως επιβλέπουσα παράμετρο.

Η επανεμφάνιση του καρκίνου είναι ένα σύνηθες χαρακτηριστικό της καρκινικής ανάπτυξης. Δυστυχώς, δεν υπάρχουν δείκτες οι οποίοι προβλέπουν με ακρίβεια πότε και αν ένας όγκος θα επανέλθει, γεγονός που δυσκολεύει την επιλογή της βέλτιστης θεραπείας/αντιμετώπισης. Προκειμένου να ελαχιστοποιηθεί η πιθανότητα επανεμφάνισης της νόσου, τα άτομα με καρκίνο σε πρώιμο στάδιο, συχνά υποβάλλονται σε μετα-χειρουργικές τοξικές θεραπείες οι οποίες συνοδεύονται από επώδυνες παρενέργειες. Οι van ‘t Veer *et al* (2) μελέτησαν το συχνό αυτό πρόβλημα με την ταυτοποιήση ενός προφίλ γονιδιακής έκφρασης πρώιμων καρκίνων του στήθους (primary breast tumours) και την πραγματοποίηση μίας επιβλεπόμενης ανάλυσης στην οποία τα άτομα διαχωρίστηκαν σε δύο ομάδες: στην ομάδα των ατόμων που ανέπτυξαν μετάσταση σε λιγότερο από πέντε χρόνια (άστοχη πρόγνωση), και στην ομάδα των ατόμων που δεν ανέπτυξαν μετάσταση για περισσότερο από πέντε χρόνια (επιτυχής πρόγνωση). Η μελέτη ανέδειξε ένα διακριτό σύνολο 70 γονιδίων τα οποία εμφάνισαν 81% επιτυχία όταν δοκιμάστηκαν σε ανάλυση *leave-one-out cross-validation* στο σύνολο εκμάθησης, *(training set)*, και 89% ακρίβεια στο σύνολο δοκιμής *(test set)*. Το συγκεκριμένο σύνολο μοριακών δεικτών βρέθηκε να είναι μία στατιστικά σημαντική παράμετρος πρόβλεψης σε μία πολυπαραμετρική ανάλυση, αποδεικνύοντας ότι ένας έλεγχος βασιζόμενος στην έκφραση γονιδίων μπορεί να συνεισφέρει στην τρέχουσα λίστα κλινικών ελέγχων. Μελέτες σε καρκινώματα νεφρικών κυττάρων (renal cell carcinomas) (38) και όγκων του προστάτη (15) έχουν εξακριβώσει μοτίβα γονιδιακής έκφρασης με προγνωστική σημασία. Επίσης, είναι πολύ πιθανό ότι θα αποκαλυφθούν υπότυποι με κλινική σπουδαιότητα και για άλλους τύπους όγκων, χωρίς βεβαιωμένο προφίλ.

Η απόκριση σε κάποια θεραπεία σπάνια επαληθεύεται και γενικότερα παρουσιάζει διακυμάνσεις μεταξύ των ασθενών. Οι περισσότεροι κλινικοί δείκτες είναι «προγνωστικοί» (δηλαδή προβλέπουν το αποτέλεσμα), ωστόσο περισσότερο χρήσιμοι είναι εκείνοι οι δείκτες που είναι «προβλεπτικοί» για την θεραπευτική απόκριση του ασθενούς. Μόνο ένα μικρό πλήθος προβλεπτικών δεικτών χρησιμοποιούνται καθημερινά για τη θεραπεία του καρκίνου. Επί παραδείγματι, στον καρκίνο του στήθους, η παρουσία ER προβλέπει την απόκριση στο tamoxifen (39).

Θα πρέπει να σημειωθεί ότι ο θετικός στον ER καρκινικός υπότυπος του στήθους παρουσιάζει ένα διακριτό προφίλ έκφρασης σε διατάξεις μικροσυστοιχιών (1,2,4-6). Συγγενείς μεταλλαγές γαμετικής σειρας σε μερικά γονίδια προδιαθέτουν την ανάπτυξη όγκων σε συγκεκριμένους ιστούς. Επί παραδείγματι, αν και η πρωτεΐνη *BRCA1* εκφράζεται ευρέως, γυναίκες με μεταλλαγές γαμετικής σειρας σε *BRCA1* πολύ συχνά παρουσιάζουν καρκινώματα των ωοθηκών και/ή του μαστού. Εκείνοι οι όγκοι οι οποίοι προέρχονται από τις προαναφερθήσες μεταλλάξεις (με προδιάθεση) ενδέχεται να έχουν μοναδικά μοριακά προφίλ. Είναι γεγονός ότι οι germline φορείς των *BRCA1* μεταλλάξεων αναπτύσουν όγκους του στήθους ή των ωοθηκών των οποίων τα μοτίβα γονιδιακής έκφρασης διακρίνονται από τα μοτίβα των πιο σποραδικών όγκων (2,3,10).

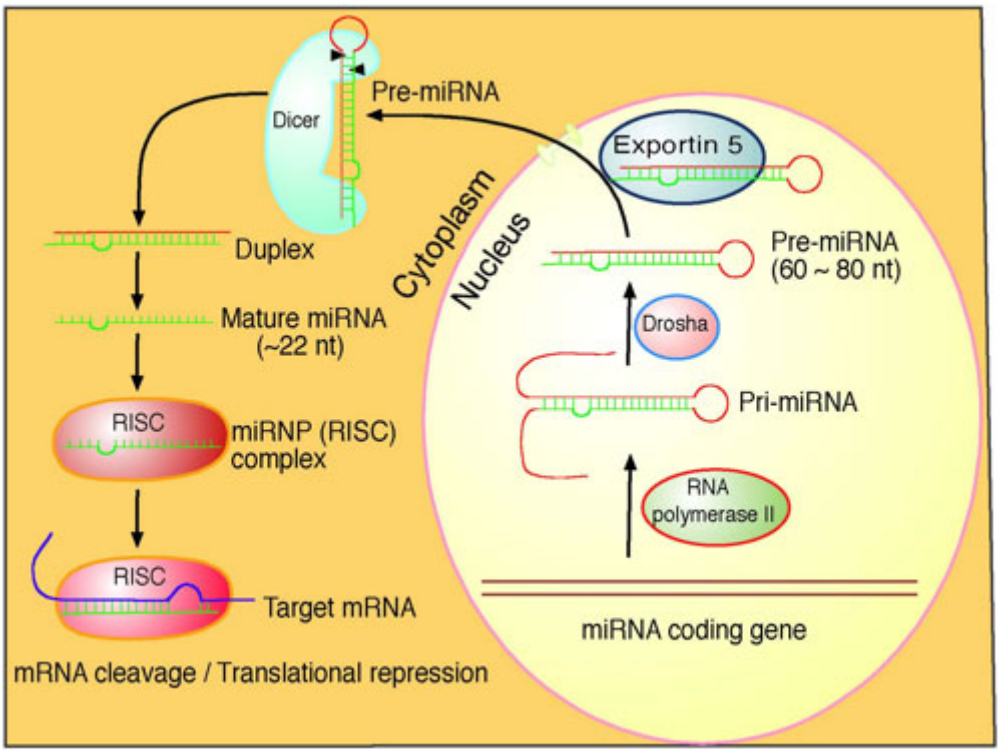
1.2 MicroRNAs και Ογκογένεση

1.2.1 Βιογένεση και Τρόπος Δράσης των MicroRNA

Τα microRNA (miRNA) ανήκουν σε μια πρόσφατα χαρακτηρισμένη ομάδα της μεγάλης οικογένειας των μη κωδικοποιών RNAs (40). Το ώριμο miRNA, το οποίο συνήθως έχει μήκος 17-27 νουκλεοτίδια, προέρχεται από ένα μεγαλύτερο πρόδρομο μόριο, ~60-70 νουκλεοτιδίων, το οποίο σχηματίζει μια ατελή δομή φουρκέτας (stem-loop). Στα ζώα, αυτά τα πρόδρομα μόρια miRNA (pre-miRNA) προέρχονται από την αποικοδόμηση του πρωτογενούς miRNA (pri-miRNA) μεταγράφου από το πολυπρωτεϊνικό σύμπλοκο που αποτελείται από την *Drosha RNase III* (41) και την *Pasha* (έταιρο του *Drosha*) η οποία συνδέεται με δίκλωνα μόρια RNA (42,43). Μετά την αποικοδόμηση, το πρόδρομο miRNA (pre-miRNA) μεταφέρεται στο κυτταρόπλασμα μέσω της *Exportin 5* (44,45). Στη συνέχεια, το πρόδρομο miRNA (pre-miRNA) κόβεται σε ένα ατελές δίκλωνο μόριο RNA μέσω μιας άλλης *RNase III ενδονουκλεάσης*, που ονομάζεται *Dicer* (41,46). Αυτό το δίκλωνο μόριο αποτελείται από το ώριμο miRNA και το συμπληρωματικό του κλώνο, που συμβολίζεται ως miRNA*.

Ο τρόπος δράσης του ώριμου miRNA σε θηλαστικά συστήματα εξαρτάται από τη συμπληρωματικότητα των βάσεων του με το 3'UTR του mRNA στόχου,

προκαλώντας την αναστολή της μετάφρασης ή/και την αποικοδόμηση του mRNA (Σχήμα 1.2).



Σχήμα 1.2 Το μονοπάτι βιογένεσης των MicroRNAs στα ζώα. Σχήμα που χρησιμοποιήθηκε στην ιστοσελίδα του Καθ. R. Padgett στο Waksman Institute of Microbiology.

1.2.2 Ο Ρόλος των miRNA στον Καρκίνο

Τα miRNA έχουν πρόσφατα συσχετιστεί με την εμφάνιση πολλών διαφορετικών τύπων καρκίνου. Σύμφωνα με την μελέτη των Calin et al. (47) όπου χρησιμοποιήθηκε πληροφορία από βάσεις δεδομένων άμεσα διαθέσιμες στο διαδίκτυο, πάνω από 50 γνωστά miRNA του ανθρώπου συσχετίστηκαν με γενωμικές περιοχές οι οποίες συνδέονται με κάποιο είδος καρίνου (Cancer Associated Genomic Regions-CAGR) καθώς επίσης και με εύθραυστες περιοχές (47). Συχνά αυτές οι περιοχές αποτελούν επίμαχες θέσεις για την παρουσία ογκοκατασταλτικών ή ογκογονιδίων (π.χ. το c-myc στην 17q22-t(8;17) CAGR, το οποίο σχετίζεται με την ανάπτυξη προ-λεμφοκκυταρικής λευχαιμίας (48)). Οι ανακαλύψεις αυτές υποδηλώνουν την ύπαρξη

κάποιας σύνδεσης μεταξύ της γενωμικών περιοχών των miRNA και των περιοχών που είναι επιρρεπείς στη διαφοροποίηση και έχουν συνδεθεί με καρκίνο. Πολλές από αυτές τις αλληλοσυνδέσεις καθώς και άλλες συσχετίσεις μεταξύ miRNA και καρκίνου έχουν επιβεβαιωθεί πειραματικά συμπεριλαμβανομένου και ενός πλήθους miRNA τα οποία είχαν πρώτα προβλεφθεί υπολογιστικά. Κάποιες από αυτές τις πειραματικά επιβεβαιωμένες μελέτες συνοψίζονται παρακάτω.

Μία από τις πρώτες, πειραματικά επιβεβαιωμένες συσχετίσεις μεταξύ καρκινικά σχετιζόμενων γενωμικών περιοχών (CAGR) και miRNA αφορά το σύμπλεγμα *mir-15a* και *mir-16-1* το οποίο εντοπίζεται εντός μιας ελάχιστης περιοχής απώλειας ετεροζυγότητας στον γενωμικό τόπο 13q14.3. Η περιοχή αυτή συχνά διαγράφεται στα B-CLLs (49). Παρόμοια αποτελέσματα έχουν προκύψει και από τη μελέτη των (50), οι οποίοι έχουν επιβεβαιώσει πειραματικά ότι 28 miRNA εκφράζονται διαφορετικά σε αδενοκαρκίνωμα του παχέως εντέρου συγκρινόμενα με φυσιολογικά δείγματα βλεννώδων ιστών. Ανάμεσα σε αυτά, τα *mir-143* και *mir-145* κατ' επανάληψη παρουσιάζονται με μειωμένα επίπεδα σε colorectal καρκίνο. Ο ρόλος των miRNA στον καρκίνο ενισχύεται περαιτέρω από τη χρήση μίας array-based comparative genomic hybridization (aCGH) μεθόδου, η οποία μπορεί να δώσει πληροφορία για την παρουσία μεταλλάξεων που σχετίζονται με miRNA σε καρκίνους διαφορετικών τύπων. Από τα 283 γνωστά miRNA που υποβλήθηκαν σε αυτήν την ανάλυση, 37.1% βρέθηκαν να εκδηλώνουν τροποποιήσεις σε επίπεδο αριθμού DNA αντιγράφων (copy number) σε καρκίνο ωοθηκών, 72.8% σε καρκίνο του μαστού και 85.9% σε μελάνωμα, παρουσιάζοντας ένα χαρακτηριστικό πρότυπο γενωμικών τροποποιήσεων για κάθε καρκινικό τύπο (51). Αυτή η μελέτη παρουσίασε περισσότερες αποδείξεις για τον πιθανό ρόλο των miRNA σαν γονίδια ογκοκαταστολείς η ογκογονίδια.
Η *let-7* οικογένεια miRNA επίσης σχετίζεται με μια εύθραυστη περιοχή (21p11.1) η οποία εμπλέκεται στον καρκίνο των πνευμόνων (47). Πειραματικά δεδομένα έχουν δείξει ότι τα *let-7* miRNA ρυθμίζονται κατασταλτικά σε σημαντικό βαθμό στον καρκίνο των πνευμόνων ενώ η υπερ-έκφραση τους σε κυτταρικές σειρές αδενώματος του πνεύμονα αποδεικνύει ότι είναι ικανά να αναστείλουν την ανάπτυξη του κυττάρου, υποδεικνύοντας ότι μπορούν και λειτουργούν ως όγκο-καταστολείς (52). Επιπρόσθετα, τα επίπεδα έκφρασης των μελών της οικογένειας let-7 miRNA στον καρκίνο του πνεύμονα είναι αρνητικά συσχετιζόμενα με το RAS, ένα πολύ γνωστό ογκογονίδιο, κάνοντας πιθανή τη καταστολή του RAS από τη συγκεκριμένη

οικογένεια miRNA (53). Η ρύθμιση του γνωστού ογκογονίδιου c-MYC από τα μέλη της οικογένειας *let-7* miRNA έχει επίσης αποδειχτεί πειραματικά (54).

Μία ευρέως γνωστή γενωμική μετάλλαξη στον καρκίνο του ανθρώπου είναι η συχνά ενισχυμένη περιοχή 13q31 στο λέμφωμα. Το ογκογονίδιο που εμπλέκεται σε αυτού του είδους τον καρκίνο δεν είχε προσδιοριστεί μέχρι τη στιγμή που οι επιστήμονες εστίασαν στο μη-κωδικοποιό γονίδιο C13orf25, το οποίο αποδείχτηκε ότι εμπεριέχει το σύμπλεγμα miRNA *mir-17-92*. Σε μελέτη που έγινε αργότερα, οι He et al βρήκαν ότι η έκφραση 6 miRNA συσχετίζεται σε σημαντικό βαθμό με την ποσότητα του C13orf25 γονιδίου (55). 5 από αυτά τα miRNA ανήκουν στο σύμπλεγμα *mir-17-92*. Πράγματι, το πρόδρομο miRNA του *mir-17-92* παρουσιάζει περισσότερο από πενταπλάσια αύξηση σε περίπου 65% των περιπτώσεων B cell λεμφώματος. Παράλληλα, με αυτή τη μελέτη, μια ανεξάρτητη ερευνητική ομάδα βρήκε ότι το *mir-17-92* εμπλέκεται σε B cell λέμφωμα. Ανάλυση των προφίλ έκφρασης miRNA της ανθρώπινης κυτταρικής σειράς B, ανέδειξε την σημαντική υπερ-έκφραση των miRNA του συμπλέγμα*τος mir-17-92* κατά την ενεργοποίηση του *c-myc*, ενός πολύ αποτελεσματικού ογκογονιδίου) (56).

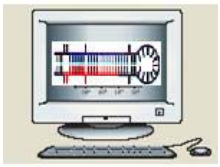
Η υπερ-έκφραση του *mir-155*, το οποίο επίσης βρίσκεται σε εύθραυστη περιοχή, πρωτοπαρουσιάστηκε σε παιδιά με Burkitt λέμφωμα (57). Περαιτέρω μελέτες που έγιναν κατόπιν έδειξαν ότι το *mir-155* υπερ-εκφράζεται επίσης στο B cell λέμφωμα, στο Hodgkin's λέμφωμα (58,59), καθώς επίσης και στο καρκίνο του μαστού (60) και του εντέρου (61). Το *mir-155* μεταγράφεται μαζί με το μη-κωδικοποιό γονίδιο *bic*, το οποίο έχει δειχθεί ότι υπερ-εκφράζεται και επίσης ότι προάγει το B cell λέμφωμα σε συνεργασία με το *myc* (62,63).

Ιδιαίτερο ενδιαφέρον αποτελεί το γεγονός ότι ένας μεγάλος αριθμός miRNA που συνδέονται με διάφορους τύπους καρκίνου έχουν εντοπιστεί από υπολογιστικές μεθόδους. Βασικά πλεονεκτήματα των μεθόδων αυτών είναι η χαμηλού κόστους γρήγορη και αποτελεσματική πρόβλεψη νέων miRNA καθώς και των στόχων τους. Αντιπροσωπευτικά παραδείγματα, που είχαν προβλεφθεί μέσω της ομολογίας των αλληλουχιών τους με ήδη κλωνοποιημένα miRNA στον ποντικό, περιλαμβάνουν το *mir-143* (64) που σχετίζεται με καρκίνο του παχέος εντέρου (50), τα *mir-125β* (lin-4) και *mir-145* που εμπλέκονται στο καρκίνο του μαστού (60) , το *mir-106α* που πιστεύεται ότι διαδραματίζει ρυθμιστικό ρόλο στο παχύ έντερο, στο πάγκρεας και στον καρκίνο του προστάτη (61) και το *mir-155* το οποίο συνδέεται με HL, BCL, παιδιατρικό BL, καρκίνο του μαστού και των πνευμόνων καθώς και χαμηλά ποσοστά

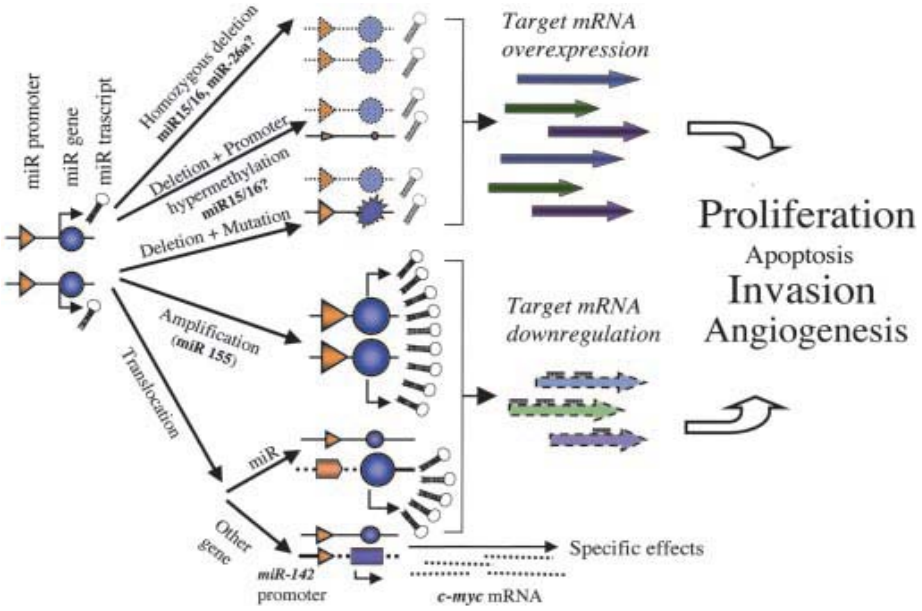
επιβίωσης των ασθενών (57-59,65,66). Επιπλέον, μια μεγάλη μελέτη προέβλεψε πολλαπλά γονίδια miRNA στα σπονδυλωτά χρησιμοποιώντας τη συντήρηση αλληλουχιών στο ποντίκι και στο «Fugu rubripes» (67) και τη βαθμολογία που δίνεται από το πρόγραμμα MiRscan.

Τα miRNA που εντοπίστηκαν σε αυτήν την μελέτη συμπεριλαμβάνουν το *mir-221/222* το όποιο εμπλέκεται σε θυρεοειδικό καρκίνωμα (Papillary thyroid carcinoma) (68) και γλοιοβλαστώματα (69), το *mir-192* το όποιο δείχνει μειωμένη έκφραση σε colorectal neoplasia (50), το *mir-196a-1* το όποιο κλωνοποιήθηκε από ανθρώπινα κύτταρα osteoblast sarcoma (70) και το *mir-210* το οποίο παίζει σημαντικό ρόλο στην μόλυνση από Kaposi's sarcoma-associated herpes virus (71).

Λόγω της εμπλοκής τους σε πολλές ασθένειες, συμπεριλαμβανομένου του καρκίνου, αυτά τα μικροσκοπικά μόρια είναι ήδη το επίκεντρο εντατικής έρευνας για νέες φαρμακολογικές παρεμβάσεις.



miRNA Gene Prediction



Σχήμα 1.3 miRNA στον καρκίνο. Η υπολογιστική πρόβλεψη αποτελεί την έναρξη της αναζήτησης για miRNA τα οποία θα παίζουν σημαντικό ρόλο στην καρκινογένεση. Ορισμένοι από αυτούς τους μηχανισμούς που προτείνονται έχουν αποδειχθεί πειραματικά, όπως η διαγραφή του συμπλέγματος (cluster) miR-15a/miR-16a στην B-CLL (49,72), η υπερέκφραση του c-myc με την επανατοποθέτηση του

κοντά σε ένα υποτιθέμενο εκκινητή ενός miRNA (47), και η μείωση της έκφρασης των miR143/miR-145 στον καρκίνο του παχέος εντέρου (50). Σχήμα που χρησιμοποιήθηκε από τους Callin et al, 2004 (47).

1.2.3 Υπολογιστική Προβλεψη miRNA Γονιδίων - Σύγκριση Εργαλείων και Κριτήρια Επιτυχίας

Δεδομένης της διαθεσιμότητας πολλών μεθόδων για τη πρόβλεψη miRNA γονιδίων, είναι αναγκαία η συγκριτική μελέτη των δυνατοτήτων τους και των περιορισμών τους. Γενικά, ο βαθμός επιτυχίας (ακρίβεια πρόβλεψης) τέτοιων εργαλείων εξαρτάται κατά πολύ από τις βιολογικές πληροφορίες που χρησιμοποιούνται ως δεδομένα εισόδου κατά την ανάπτυξή τους. Όπως φαίνεται στο Πίνακα 1.1., τα εργαλεία που αναπτύχθηκαν αρχικά και χρησιμοποιούσαν μόνο την αλληλουχία και τη συντήρηση της σε συγγενείς οργανισμούς δεν είχαν πολύ υψηλή ακρίβεια πρόβλεψης. Μεταγενέστερες μεθοδολογίες οι οποίες έλαβαν υπόψη επιπλέον βιολογικές πληροφορίες όπως η δομή και η συντήρηση σε περισσότερο αποκλίνοντα είδη, προκειμένου να φιλτράρουν τις προβλέψεις τους είχαν σαν αποτέλεσμα σημαντικές βελτιώσεις. Η ταυτόχρονη ενσωμάτωση αυτών των πληροφοριών είναι ένα άλλο σημαντικό κριτήριο επιτυχίας και έχει υιοθετηθεί από πλήθος εξελιγμένων αλγόριθμων μηχανικής μάθησης. Λαμβάνοντας υπόψη όλα αυτά τα γνωρίσματα συγχρόνως είναι περισσότερο αποτελεσματικό από την υιοθέτηση μιας σειριακής προσέγγισης (pipeline) που είχε χρησιμοποιηθεί αρχικά, με την οποία διαφορετικά γνωρίσματα χρησιμοποιούνται διαδοχικά (μέθοδος brute-force) για την πρόβλεψη νέων miRNA (βλέπε Πίνακα 1.1).

Ιδιαίτερη προσοχή κατά την εκμάθηση αυτών των αλγορίθμων, πρέπει να δοθεί στην επιλογή των θετικών και αρνητικών δειγμάτων εκπαίδευσης, καθώς οι διαδυκτιακές βάσεις δεδομένων είναι δυνατόν να περιέχουν μη-πραγματικά θετικά παραδείγματα ενώ ο ορισμός των αλληλουχιών που αποτελούν τα αρνητικά δείγματα παραμένει ασαφής. Οι περισσότερες δημοσιεύσεις χρησιμοποιούν τις 3'UTR περιοχές για να εξάγουν τα αρνητικά δείγματα καθώς το μεγαλύτερο ποσοστό των περιοχών αυτών δεν περιέχουν κανένα miRNA. Επιπλέον, η ευαισθησία και η ειδικότητα (βλεπε *Sensitivity, Specificity and Mathew's Correlation* στο παρατημα) επηρεάζονται άμεσα από τον αριθμό καθώς επίσης και από την ποιότητα των θετικών και αρνητικών δειγμάτων (73), έτσι αρχικά αποτελέσματα από μια μελέτη μπορεί να αλλάξουν εάν χρησιμοποιηθεί το σύνολο δεδομένων από άλλη μελέτη και αντίθετα. Ο Πίνακας 1.1

συνοψίζει τη βέλτιστη απόδοση που επιτυγχάνουν 12 εργαλεία πρόβλεψης miRNA γονιδίων καθώς επίσης και τον οργανισμό στον οποίο ειδικεύονται.

Παρόλο που έχει γίνει μεγάλη πρόοδος κατά τη διάρκεια της τελευταίας δεκαετίας στον υπολογιστικό προσδιορισμό των miRNA γονιδίων, υπάρχει άφθονος χώρος για βελτίωση. Με τη διάθεση διαρκώς αυξανόμενων βιολογικών πληροφοριών που αφορούν τη miRNA βιογένεση και ρύθμιση, τα υπολογιστικά εργαλεία βελτιώνονται όσον αφορά την ακρίβεια πρόβλεψης. Ένας ανασταλτικός παράγοντας στην *in silico* πρόβλεψη των miRNA είναι ο προσδιορισμός της αλληλουχίας του ώριμου miRNA στο πρόδρομο miRNA, καθώς το μικρό μέγεθος του ώριμου miRNA (~22nt) περιορίζει την εξαγωγή πληροφορίας που σχετίζεται με την αλληλουχία, τη δομή και τη συντήρηση. Βελτιστοποίηση της πρόβλεψης προκύπτει επίσης με την ενσωμάτωση νέων γνωρισμάτων όπως η περιοχή εκατέρωθεν του γονιδίου miRNA και ποιες πληροφορίες μπορεί αυτή να παρέχει ως υπόδειξη της παρουσίας ενός γονιδίου miRNA. Εργαλεία ικανά να προβλέπουν τριτοταγή δομή (όπως τα pseudoknot, (74)) των miRNA προσφέρουν μια περισσότερο ολοκληρωμένη εικόνα. Το τελευταίο μετατρέπει ένα δισδιάστατο πρόβλημα σε τρισδιάστατο, αντικατοπτρίζοντας με μεγαλύτερη ακρίβεια τις συνθήκες που συναντώνται στο κύτταρο. Σαν τελική παρατήρηση σε αυτή την ενότητα, είναι σημαντικό να αναφερθεί ότι η ανάπτυξη των εργαλείων αυτών είναι στενά συνδεδεμένη με τη βιολογική έρευνα. Η επιτυχημένη εξέλιξη των εργαλείων απαιτεί από τους υπεύθυνους ανάπτυξής τους να ενημερώνονται για τα νέα βιολογικά ευρήματα, τα οποία αλλάζουν συχνά την πληροφορία που μπορεί να χρησιμοποιηθεί. Ένα χαρακτηριστικό παράδειγμα είναι το εργαλείο Microprocessor (75) το οποίο χρησιμοποιεί περιοχές επεξεργασίας της Drosha, ενώ πρόσφατα αποδείχθηκε πως πρόδρομα miRNA που βρίσκονται σε ιντρόνια μπορούν να παρακάμψουν την επεξεργασία Drosha (76).

Πίνακας 1.1 Σύγκριση 12 εργαλείων πρόβλεψης γονιδίων miRNA. Ο πίνακας δείχνει τα γνωρίσματα που χρησιμοποιεί κάθε εργαλείο, την απόδοση που επιτυγχάνει και τον οργανισμό στον οποίο ειδικεύεται. Πίνακας που χρησιμοποιήθηκε από τους Oulas et al, 2009 (77).

Features	Cloning	MiRscan	miRseeker	phylogenetic shadowing	Blatting	miRAlign	ProMir	Bayes-classifier	Xue	Sewer	RNAmicro	Microprocessor SVM
Sequence												
Directly[3]	X					X	X	X	X	X	X	X
Indirectly		X	X	X								
Structure												
Base pairing		X				X	X	X	X	X	X	X
Hairpin		X				X	X	X	X	X	X	X
Bulges/Loops		X				X	X	X	X	X	X	X
Mature location		X					X					
Thermodymamic temp			X	X	X	X	X	X	X	X	X	X
Conservation												
Pairwise	X	X	X									
Conserved synteny					X							
Conserved clustering										X		
Multiple species				X			X	X				
Methodology												
Bruteforce		X	X									
Homology based				X	X							
SVM									X	X	X	X
Probablistics						X	X	X				
Complement other tools											X	X
Performance												
sensitivity		0.74%	75%				73%	97%	93.3%	64%	90%	90%
specificity							96%	91%	88.1%	64%		78%
species		Human + Nematode	Drosophila				human	mouse	Multi species	Human,mouse, rat	human	human
Relevant Reference		[32]	[40]	[41]	[42]	[43]	[45]	[34]	[48]	[33]	[31]	[30]

[3] Directly in the sense that the nucleotide distribution in the sequence is taken into consideration i.e.GC content. Indirectly refers to the use of sequence to derive structure.

1.2.4 Υπολογιστική Πρόβλεψη Στόχων των miRNA

Όπως αναφέρθηκε προηγούμενος ο τρόπος δράσης του ώριμου miRNA σε θηλαστικά συστήματα εξαρτάται εν μέρει από τη συμπληρωματικότητα των βάσεων του με το 3'UTR του mRNA στόχου, προκαλώντας την αναστολή της μετάφρασης ή / και την αποικοδόμηση του mRNA. Παρά το πλήθος των στοιχείων που υποστηρίζουν το ρόλο που παίζουν τα miRNA στον καρκίνο, ο ακριβής μηχανισμός με τον οποίον δρουν παραμένει άγνωστος, κυρίως εξαιτίας του γεγονότος ότι οι mRNA στόχοι δεν έχουν διευκρινιστεί. Προς αυτή την κατεύθυνση, η υπολογιστική πρόβλεψη miRNA στόχων μπορεί να προσφέρει μία πρώτη ένδειξη για το ποια γονίδια στόχοι ρυθμίζονται από miRNA και κατά συνέπεια να προσφέρει περαιτέρω γνώση σχετικά με τον τρόπο που λειτουργούν καθώς επίσης και να κατευθύνει νέα πειράματα.
Η πειραματική επιβεβαίωση των υπολογιστικά παραγόμενων προβλέψεων αποτελεί τον απώτερο στόχο για τον προσδιορισμό των ρυθμιστικών μοριακών μονοπατιών των miRNA και τον τρόπο που αυτά εμπλέκονται στον καρκίνο.

Γνωρίσματα που Χαρακτηρίζουν τις miRNA::mRNA αλληλεπιδράσεις

Οι miRNA στοχευμένες περιοχές μπορούν να ταξινομηθούν σε τρεις κύριες κατηγορίες (78) (i) 5′-dominant canonical, (ii) 5′-dominant seed και (iii) 3′-compensatory. Η κάθε μια από αυτές τις κατηγορίες παρουσιάζει ένα χαρακτηριστικό πρότυπο/γνώρισμα συμπληρωματικότητας μεταξύ του ώριμου miRNA και την στοχευόμενη περιοχή στο 3’UTR και αναλύεται λεπτομερώς στο κεφάλαιο 4.
Οι miRNA::mRNA αλληλεπιδράσεις επίσης χαρακτηρίζονται χρησιμοποιώντας άλλα γνωρίσματα όπως: εξελικτική συντήρηση υπολογισμοί ελεύθερης ενέργειας (όπως προβλέπονται από προγράμματα σαν το RNAcofold (79) και RNAhybrid (80)), και συνεργασιμότητα στην πρόσδεση (78,81,82). Πρόσφατα η δευτεροταγής δομή του 3’UTR που περιβάλλει τη στοχευμένη περιοχή έχει επίσης χρησιμοποιηθεί σε αυτήν τη διαδικασία (83).

Γενική Επισκόπηση Εργαλείων πρόβλεψης στόχων

Μία αναλυτική επισκόπηση των υπαρχόντων υπολογιστικών προσεγγίσεων για την αναγνώριση των miRNA στόχων για θηλαστικά γίνεται στην αναφορά (78). Μερικά από τα αναφερόμενα εργαλεία είναι: το TargetScan 4.0 (82), το PicTar (84), το PITA (83), το DIANA-MicroT (81) και το Miranda (85). Στα πλαίσια της μελέτης τους, οι Sethupathy et al χρησιμοποίησαν πειραματικά επικυρωμένες miRNA::mRNA αλληλεπιδράσεις από διαδυκτιακές βάσεις δεδομένων ως ένα θετικό σύνολο δεδομένων. Χρησιμοποίησαν τα παραπάνω εργαλεία πρόβλεψης για να αναλύσουν όλες τις ανθρώπινες 3'UTR αλληλουχίες για νέα miRNA. Κατέγραψαν τον αριθμό των θετικών/πραγματικών miRNA τα οποία προβλέφθηκαν σωστά (ευαισθησία) και επίσης τον συνολικό αριθμό επιτυχιών/προβλέψεων (ειδικότητα) που αποκομίστηκαν από κάθε εργαλείο πρόβλεψης. Τα αποτελέσματα της παραπάνω μελέτης αναλύονται με λεπτομέρεια στο κεφάλαιο 4 και δείχνουν ότι γενικό πρόβλημα της πλειονότητας των εργαλείων αυτών είναι η χαμηλή ακρίβεια πρόβλεψης θετικών αλληλεπιδράσεων.

Ανάγκη για την Υλοποίηση ενός Νέου Εργαλείου Πρόβλεψης Στόχων miRNA

Οι υπάρχουσες υπολογιστικές μέθοδοι πρόβλεψης στόχων miRNA διαφέρουν ως προς τον αλγόριθμο που χρησιμοποιούν, και μπορούν να διατυπωθούν διάφορες απόψεις σχετικά με τα πλεονεκτήματα και τις αδυναμίες του καθενός από τους αλγορίθμους αυτούς. Είναι όμως γεγονός ότι όλες οι μέθοδοι υστερούν σημαντικά στο να συλλάβουν όλη τη πληροφορία σχετικά με τις φυσικές, χρονικές, και χωρικές απαιτήσεις των βιολογικά σημαντικών miRNA::mRNA αλληλεπιδράσεων. Η απόδοση των υπαρχόντων εργαλείων βασίζεται σε μεγάλο ποσοστό στο συνολικό αριθμό των προβλεφθέντων στόχων (προβλέψεις). Μερικά εργαλεία μπορεί να είναι πολύ αποτελεσματικά στην πρόβλεψη πραγματικών στοχευμένων περιοχών (υψηλή ευαισθησία) αλλά ταυτόχρονα επιδεικνύουν έναν εξαιρετικά μεγάλο αριθμό συνολικών προβλέψεων (χαμηλή ειδικότητα). Αντίθετα, άλλα εργαλεία επιδεικνύουν συνολικά υψηλή ειδικότητα αλλά χαμηλή ευαισθησία. Είναι φανερή η μεγάλη αναγκαιότητα για ένα πιο εξελιγμένο εργαλείο πρόβλεψης στόχων που θα πετυχαίνει μια ισορροπία μεταξύ ευαισθησίας και ειδικότητας.

1.3 Στόχοι της Παρούσας Διατριβής

Σκοπός της παρούσας διατριβής είναι να παράσχει στοιχεία που υποστηρίζουν ένα νέο σχέδιο κατηγοριοποίησης για συγκεκριμένους τύπους καρκίνου του εγκεφάλου καθώς επίσης και να προσδιορίσει νέα miRNA γονίδια που εμπλέκονται στη διαδικασία ανάπτυξης καρκινικών όγκων. Η ανάπτυξη υπολογιστικών τεχνικών και αλγορίθμων αποτελεί βασική προϋπόθεση για την επιτυχή επίτευξη των παραπάνω στόχων. Στο κεφάλαιο 2 δίνεται ιδιαίτερη έμφαση σε αστροκυττωματικούς όγκους του εγκέφαλου. Μια πρόσφατη συνεργασία με το ***Department of Pathology, Division of Molecular Histopathology, University of Cambridge, Addenbrooke's Hospital, United Kingdom***, μας έδωσε την δυνατότητα να επεξεργαστούμε ένα μεγάλο σύνολο δεδομένων έκφρασης Affymetrix από αστροκυττωματικούς όγκους. Στα κεφάλαια 3 και 4 η έρευνα επικεντρώνεται στα γονίδια miRNA και τους πιθανούς ρυθμιστικούς στόχους τους.

Τρία κυρίως θέματα θα παρουσιαστούν:

1. Η υπολογιστική ανάλυση δεδομένων έκφρασης γονιδίων με τη χρήση τεχνητών νευρωνικών δικτύων (ΤΝΔ).
2. Η υπολογιστική πρόβλεψη νέων miRNA τα οποία βρίσκονται σε καρκινικά σχετιζόμενες γενωμικές περιοχές (CAGR) με τη χρήσης Κρυφών Μαρκοβιανών Μοντέλων (ΚΜΜ).
3. Η υλοποίηση ενός νέου αλγορίθμου για την πρόβλεψη πιθανών miRNA στόχων.

Για το πρώτο από τα τρία θέματα που αναλύονται σε αυτή η διατριβή, χρησιμοποιούμε ένα νέο σύνολο δεδομένων γονιδιακής έκφρασης από 65 αστροκυττωματικούς όγκους για τους οποίος διαθέτουμε εκτεταμένη κλινική και μοριακή πληροφορία. Στόχοι μας είναι:

i. Η χρήση των ως τώρα διαθέσιμων καθώς επίσης και η ανάπτυξη νέων αλγορίθμων επιβλεπόμενης (σε μορφή τεχνητών νευρωνικών δικτύων - ΤΝΔ) και μη-επιβλεπόμενης μάθησης.
ii. Η εξαγωγή συγκεκριμένων μεταγραφικών αποτυπωμάτων από ιστοπαθολογικούς υπότυπους αστροκυττωματικών όγκων
iii. Ο προσδιορισμός γονίδιων ταξινομητών.

iv. Η επιτυχής διαβάθμιση των διαφόρων βαθμίδων αστροκυττωματικών όγκων με την χρήση δεδομένων προφίλ έκφρασης γονιδίων.

v. Καθορισμός προγνωστικών υποκατηγοριών επιβίωσης

Ο απώτερος σκοπός είναι η ταυτοποίηση νέων βιολογικών σχέσεων μεταξύ ρυθμιστικών μονοπατιών αστροκυττωματικών όγκων, ο προσδιορισμός γονιδίων τα οποία θα μπορούσαν να αποτελέσουν μοριακούς καθώς επίσης και προγνωστικούς δείκτες για αστροκυττωματικούς όγκους και τέλος η ανάπτυξη ενός περισσότερο αποτελεσματικού συστήματος διαβάθμισης αστροκυττωματικών όγκων. Τέλος ένας γενικός στόχος της διατριβής αυτής είναι να παρέχει υπολογιστικές τεχνικές, μεθοδολογία και αλγόριθμους που θα μπορέσουν να γίνουν ευρέως εφαρμόσιμοι σε κάθε τύπο καρκίνου

Το δεύτερο θέμα αφορά:

(i) Την υλοποίηση ενός ελεύθερα διαθέσιμου υπολογιστικού εργαλείου πρόβλεψης νέων miRNA, χρησιμοποιώντας αλγόριθμους επιβλεπόμενης μάθησης (σε μορφή Κρυφών Μαρκοβιανών Μοντέλων (ΚΜΜ)) οι οποίοι εκπαιδεύονται στην αναγνώριση βιολογικών γνωρισμάτων της βιογένεσης και συντήρησης των miRNA,

(ii) Την υπολογιστική πρόβλεψη νέων υποψήφιων miRNA σε καρκινικά σχετιζόμενες γονιδιακές περιοχές (CAGR)

(iii) Η αξιοποίηση διαθέσιμων δεδομένων απο μεθόδους ευρείας κλίμακας, μαζικής ανάλυσης (large scale, high throughput) όπως tiling arrays, για να δοθεί μεγαλύτερο βάρος στις υπολογιστικές προβλέψεις.

(iv) Την επιβεβαίωση των υπολογιστικών προβλέψεων χρησιμοποιώντας πειραματικές τεχνικές όπως Northern blot analysis.

Ο προσδιορισμός νέων miRNA γονιδίων μέσα σε καρκινικά σχετιζόμενες γονιδιακές περιοχές (CAGR) αποτελεί ένα ιδιαίτερα σημαντικό κομμάτι στην αποκάλυψη νέων υποψήφιων γονίδιων με ρυθμιστική επίδραση σε διαφορετικούς τύπους καρκίνου, συμβάλλει στην καλύτερη κατανόηση των μοριακών μονοπατιών που εμπλέκονται στην ογκογένεση και παρέχει δυνητικούς στόχους για θεραπευτική παρέμβαση.

Το τρίτο θέμα αφορά την υλοποίηση ενός νέου, βελτιωμένου αλγόριθμου-εργαλείου ικανό να προβλέπει γονίδια στόχους των miRNA, καθως, οι υπάρχουσες μέθοδοι πρόβλεψης γονιδίων στόχων των miRNA βρίσκονται ακόμα σε πρωταρχικό στάδιο. Αυτό απαιτεί:

(i) Την υλοποίηση ενός υπολογιστικού μοντέλου πρόβλεψης γονιδίων στόχων των miRNA

(ii) Σωστά επιλεγμένα σύνολα δεδομένων για εκπαίδευση και επαλήθευση

(iii) Σύγκριση εργαλείων πρόβλεψης MiRNA στόχων

(iv) Επιβεβαίωση της βελτιωμένης επίδοσης του υπολογιστικού μοντέλου πρόβλεψης γονιδίων στόχων των miRNA σε σχέση με υπάρχοντα εργαλεία.

Ο απώτερος στόχος σε αυτό το θέμα είναι η ενσωμάτωση του εργαλείου πρόβλεψης miRNA γονίδιων και αυτού της πρόβλεψης στόχων των miRNA, σε ένα ενιαίο, ελευθέρα διαθέσιμο υπολογιστικό εργαλείο.

Κεφάλαιο 2

2 Βελτιωμένη Διαβάθμιση και Πρόβλεψη Επιβίωσης για Αστροκυττωματικούς Ανθρώπινους Όγκους του Εγκεφάλου με Ανάλυση Δεδομένων Έκφρασης Γονίδιων από Μικροσυστοιχίες DNA με την χρήση Τεχνητών Νευρωνικών Δικτύων

2.1 Εισαγωγή

Αστροκυττωματικοί όγκοι με βαθμίδες κακοήθειας II έως IV αποκαλούνται συνολικά διεισδυτικοί αστροκυττωματικοί όγκοι (diffusely infiltrating astrocytomas), και περιλαμβάνουν το αστροκύττωμα (diffuse astrocytoma - βαθμίδος κακοήθειας II, καλούμενο ως 'A'), αναπλαστικό αστροκύττωμα (anaplastic astrocytoma - βαθμίδος κακοήθειας III, καλούμενο ως 'AA') και γλοιοβλάστωμα (glioblastoma - βαθμίδος κακοήθειας IV, καλούμενο ως 'GB'). Ένα σύνολο τεσσάρων βαθμίδων κακοήθειας αναγνωρίζονται από τον Παγκόσμιο Οργανισμό Υγείας (ΠΟΥ) για τους αστροκυττωματικούς όγκους με αυξανόμενη επιθετικότητα όγκων για τις βαθμίδες I έως IV (86,87). Τα γλοιοβλαστώματα συνήθως αναπτύσσονται *de novo* (αποκαλούνται επίσης πρωτογενή γλοιοβλαστώματα) μπορεί όμως να εμφανιστούν από την ανάπτυξη όγκων χαμηλότερης βαθμίδας. Τα γλοιοβλαστώματα επιδεικνύουν το μεγαλύτερο εύρος γενετικών ανωμαλιών, με συνήθεις αλλαγές στους *de novo* όγκους να περιλαμβάνουν ομοζυγωτική εξάλειψη των CDKN2A, CDKN2B και $p14^{ARF}$ (9p21), απώλεια του ενός αλληλόμορφου γονιδίου PTEN (10q23) και μετάλλαξη του εναπομείναντος και ενίσχυση του EGFR γονιδίου (7p12) (87).

Η χρήση των δεδομένων έκφρασης γονίδιων από Μικροσυστοιχίες DNA για ταξινόμηση/ομαδοποίηση όγκων εγκεφάλου (23) και πρόγνωση επιβίωσης (88-90)

έχει κερδίσει σημαντικό ενδιαφέρον τα τελευταία χρόνια. Η ανάλυση περιλαμβάνει στατιστικές μεθόδους για την εύρεση σημαντικών γονιδίων και ταξινόμηση όγκων (91), principal component analysis (PCA) και t-test για την επιλογή γονιδίων με σημαντικά διαφορετική έκφραση που εμπλέκονται στην ανάπτυξη των αστροκυττωμάτων (31), k-means μαζί με multidimensional scaling για διάκριση μεταξύ γλοιοβλαστώματος, χαμηλότερης βαθμίδας αστροκυττώματος και άλλων τύπων γλοιόματος όπως oligodendrogliomas (28), ιεραρχική ομαδοποίηση (23,28,88,92), k-nearest-neighbour για την ταξινόμηση γλοιόματος υψηλότερης βαθμίδας και πρόγνωση επανεμφάνισης της νόσου (89), gene voting για πρόγνωση επιβίωσης σε ασθενείς με διεισδυτικό γλοίομα (diffusely infiltrating gliomas) (88) και πολλές άλλες. Η ανάλυση του γονιδιακού προφίλ έκφρασης έχει προσδιορίσει τόσο μοριακούς όσο και γενετικούς υπότυπους που σχετίζονται με τη διαβάθμιση και την ανάπτυξη όγκων καθώς και την επιβίωση των ασθενών (31,92). Ενώ οι αστροκυττωματικοί όγκοι εξακολουθούν να καθορίζονται από ιστολογικά κριτήρια, υπάρχει σημαντικός αριθμός αναφορών που δείχνουν ότι η ανάλυση των προφίλ έκφρασης προβλέπει την επιβίωση ασθενών καλύτερα από την ιστοπαθολογική διαβάθμιση (88,89,93). Αυτό υποστηρίζει ισχυρά την υπόθεση ότι οι μορφολογικά καθορισμένοι όγκοι αντιπροσωπεύουν ένα μείγμα μοριακών γενετικών υπότυπων.

Στις περισσότερες από αυτές τις μελέτες όμως η σύγκριση έγινε μεταξύ διεισδυτικούς αστροκυττωματικούς όγκους και όγκους ποικίλης ή μη-αστροκυττωματικής προέλευσης μη συμπεριλαμβάνοντας όγκους χαμηλότερης βαθμίδας (II) (88,89,91,93), ή έχει περιορίσει η ανάλυση σε μία μοναδική βαθμίδα όγκων (23,90). Σε πολλές από τις παραπάνω μελέτες γίνεται σύγκριση μεταξύ καρκινικού και μη-καρκινικού ιστού, σύγκριση αμφίβολης αποτελεσματικότητας αν λάβει κάνεις υπόψη του τις μεγάλες διαφορές στην κυτταρική σύνθεση ανάμεσα στους δύο ιστούς. Επιπρόσθετα, οι μελέτες έχουν επικεντρωθεί σε ερωτήματα που σχετίζονται με τη χρήση δεδομένων έκφρασης περισσότερο στη γενική ταξινόμηση όγκων εγκεφάλου παρά στη καθ'αυτή διαβάθμιση κακοήθειας διεισδυτικών αστροκυττωματών όγκων (diffusely infiltrating astrocytic tumours). Τέλος, η χρήση κλινικών ή/και μοριακών δεδομένων δεν έχει χρησιμοποιηθεί ιδιαίτερα για τη διαλεύκανση των αναντιστοιχιών μεταξύ ιστοπαθολογικής ταξινόμησης και ταξινόμησης βασισμένης σε δεδομένα έκφρασης για ένα δοσμένο σύνολο όγκων.

Χρησιμοποιώντας ένα νέο σύνολο δεδομένων γονιδιακής έκφρασης από 65 όγκους για τους οποίος διαθέτουμε εκτεταμένη κλινική και μοριακή πληροφορία και ένα

απλό αλγόριθμο τεχνητού νευρωνικού δικτύου (ΤΝΔ) σε μορφή single-layer perceptron, προσεγγίζουμε τη διαβάθμιση ανθρώπινων αστροκυττωματικών όγκων, εξάγουμε συγκεκριμένα μεταγραφικά αποτυπώματα από ιστοπαθολογικούς υπότυπους αστροκυττωματικών όγκων και εξετάζουμε κατά πόσο τα μοριακά αυτά αποτυπώματα καθορίζουν προγνωστικές υποκατηγορίες επιβίωσης. Τεκμηριώνουμε την προσέγγιση μας χρησιμοποιώντας ανεξάρτητα σύνολα δεδομένων και προφέρουμε πολύτιμη πληροφορία για τη βιολογία του όγκου και τη διαβάθμιση αστροκυττωματικών όγκων βασισμένη σε δεδομένα γονιδιακής έκφρασης.

2.2 Υλικά και Μέθοδοι

2.2.1 Πειραματικές Μέθόδοι

Δείγματα Όγκων, Απομόνωση RNA και Υβριδοποίηση στα Affymetrix U133A GeneChips

Τα δείγματα όγκων αποτελούνται από 2 pilocytic αστροκυττώματα (ΠΟΥ, βαθμίδα I, 'PA'), 5 αστροκυττώματα (ΠΟΥ βαθμίδα II, 'A'), 15 αναπλαστικά αστροκυττώματα (ΠΟΥ βαθμίδα III, 'AA') and 39 γλοιοβλαστώματα (ΠΟΥ βαθμίδα IV, 'GB'). Αυτή η κατανομή των δειγμάτων αντικατοπτρίζει την διαθεσιμότητα των ιστών και την σχετική συχνότητα της διάγνωσης ανά βαθμίδα όγκου. 4 επιπρόσθετα δείγματα διαβαθμισμένα ως AA ήταν ιδιαίτερα δύσκολο να διαβαθμιστούν ιστοπαθολογικά και έτσι αντιμετωπίστηκαν σαν «ιδιαίτερες» περιπτώσεις. Η ιστοπαθολογική διάγνωση έγινε σύμφωνα με τα κριτήρια του ΠΟΥ (86) από τον Δρ. Collins VP. RNA από τα 65 ανθρώπινα αστροκυττωματικά δείγματα όγκων εξήχθηκαν χρησιμοποιώντας υπερ-φυγοκέντριση με guanidine isothiocyanate όπως έχει περιγραφεί προγενέστερα (94). Η ποιότητα του RNA επιβεβαιώθηκε χρησιμοποιώντας Agilent Bioanalyzer 2100 (Agilent technologies). Για κάθε δείγμα όγκου, χρησιμοποιήθηκαν 7 μg RNA για να παραχθεί cDNA διπλής έλικας το οποίο στην συνέχεια μεταγράφηκε *in vitro* σε cRNA σημασμένο με biotin χρησιμοποιώντας το ENZO BioArray HighYield kit. 15 μg cRNA κατακερματίστηκαν και υβριδοποιήθηκαν στα Affymetrix HG-U133A genechips (Affymetrix, Inc, Santa Clara, CA). Τα GeneChips πλύθηκαν, βάφηκαν,

και σαρώθηκαν όπως περιγράφεται στο εγχειρίδιο του κατασκευαστή. Η ποιότητα του προ- και μετά-κατακερματισμένου cRNA πιστοποιήθηκε χρησιμοποιώντας τον Agilent Bioanalyzer 2100 (Agilent technologies).

Επαλήθευση των αποτελεσμάτων χρησιμοποιώντας Quantitative PCR (qPCR)

Η QPCR πραγματοποιήθηκε σε συσκευή LightCycler (Roche) χρησιμοποιώντας DNA master SYBR Green I (Roche Molecular Biochemicals, ή Sigma) σύμφωνα με το πρωτόκολλο του κατασκευαστή. Οι primers παραγγέλθηκαν από τους MWG. Το δίκλωνο cDNA που χρησιμοποιήθηκε σαν υπόστρωμα ήταν το ίδιο που χρησιμοποιήθηκε για την παρασκευή του cRNA στόχου. 1 μl από αυτό το cDNA υπόστρωμα αραιώθηκε με αναλογία 1:200 για την παρασκευή του τελικού υποστρώματος που χρησιμοποιήθηκε. Η επαλήθευση πραγματοποιήθηκε σε ένα υποσύνολο 23 όγκων (15 GB, 8 AA) που ήταν τμήμα του αρχικού συνόλου δειγμάτων και η διαδικασία πραγματοποιήθηκε εις διπλούν. Τα ακατέργαστα δεδομένα από την qPCR αναφέρονται στον αριθμό τον κύκλων που απαιτούνται προκειμένου οι αντιδράσεις να φτάσουν σε εκθετική φάση, όπως καθορίζεται από το λογισμικό RelQuant (Roche). Για την κανονικοποίηση των δεδομένων QPCR χρησιμοποιήθηκε η έκφραση της *MYO1C*. Η μέση τιμή της διαφοράς έκφρασης μεταξύ των ομάδων όγκων υπολογίστηκε χρησιμοποιώντας τη μέθοδο $2^{-\Delta\Delta C_T}$ (95). Αλληλουχίες Primer: *PEA15*, 5'-GAGCAGCCAGCGTTAGATGC-3', 3'-GGAGGTGTTCACAAGACCAGGG-5'; *ADM*,5'-GCAGAAGAATCCGAGTGTTTGC-3', 3'-AATCAGTTTGTGGGCGAGCACG-5'.

Παρασκευή συστοιχίας ιστού και Ανοσοϊστοχημεία

Πυρήνες (n=2, για 57 όγκους από το σύνολο δεδομένων μας) διαμέτρου 0.6 mm ελήφθησαν από ιστούς όγκων εμποτισμένους με παραφίνη και στοιχίστηκαν σε ένα φρέσκο κύβο παραφίνης χρησιμοποιώντας χειροκίνητο στοιχιστή ιστών (Beecher Instruments, Silver Spring, MD, USA). Οι περιοχές που χρησιμοποιήθηκαν ήταν εκείνες που βρέθηκαν να είναι πλούσιες σε κύτταρα όγκου σε τομές βαμμένες με hematoxylin και eosin. Συμπεριλήφθησαν επίσης 10 μη-νεοπλαστικοί πυρήνες ιστών με ελάχιστο η καθόλου περιεχόμενο σε καρκινικά κύτταρα. Ανοσοϊστοχημεία για ADM (1:50, Abcam, ab18092) και PEA15 (1:500) πραγματοποιήθηκε όπως έχει αναφερθεί προγενέστερα (96,97).

2.2.2 Υπολογιστικές Μέθοδοι

Ανάλυση Έκφρασης Μικροσυστοιχιών

Τα ακατέργαστα δεδομένα εισήχθησαν στο περιβάλλον 'R' το όποιο είναι ελεύθερα διαθέσιμο πακέτο υπολογιστικής στατιστικής (98). Η κανονικοποίηση και ο υπολογισμός των μετρήσεων γονιδιακής έκφρασης πραγματοποιηθήκαν χρησιμοποιώντας τη συνάρτηση justRMA στο πακέτο Affy του Bioconductor (99). Όλα τα δεδομένα έκφρασης έχουν υποβληθεί στο GEO (100) με έναν MIAME-συμβατό τρόπο (accession number GSE1993). Η περιγραφεί της λίστας των probes έγινε με βάση το σύστημα EASE (101).

Επιλογή γονιδίων

Στατιστικά σημαντικά γονίδια επιλέχτηκαν με την μέθοδο σήμα-προς-θόρυβο όπως περιγράφεται στο (102). Γονίδια με τιμή κοντά στο 1 ή -1 σύμφωνα με τη έξοδο αυτή περιγράφονται ως τα πλέον σημαντικά.

$$S2N = \frac{(\mathrm{Avg}_1 - \mathrm{Avg}_2)}{(\sigma_1 + \sigma_2)}$$

Όπου: - Avg_1 και σ_1 είναι ο μέσος όρος και η τυπική απόκλιση των τιμών έκφρασης για το γονίδιο χ σε όλα τα δείγματα της κλάσης 1 και όμοια Avg_2 and σ_2 για τα δείγματα της κλάσης 2.

Για λόγους σύγκρισης, γονίδια με σημαντικά διαφοροποιημένες τιμές έκφρασης μεταξύ των τριών βαθμιδώσεων αστροκυττωματικών όγκων ανιχνεύτηκαν επίσης χρησιμοποιώντας Bayesian εμπειρική ανάλυση όπως έχει αναπτυχθεί στο πακέτο LIMMA (Linear Models for Microarray Data). Στο πακέτο αυτό πιθανά λανθασμένες εκτιμήσεις ελέχθησαν με τη μέθοδο των Benjamini & Hochberg (103). Μόνο οι ομάδες γονιδίων με εντάσεις μεγαλύτερες από 50 μονάδες και στα 65 δείγματα, γεωμετρικό μέσο όρο διαφοροποίησης της γονιδιακής έκφρασης μεγαλύτερος από 2

και σχετική Bayesian p-value τιμή μικρότερη από 0.001 θεωρήθηκαν σημαντικά διαφοροποιημένα σε κάθε μια από τις συγκρίσεις που πραγματοποιήθηκαν μεταξύ των 3 γκρουπ (GB-AA, GB-A and A-AA).

2.2.3 Το Μοντέλο ΤΝΔ και Στατιστική Ανάλυση

Προκειμένου να εκπαιδευτεί ένα νευρωνικό δίκτυο, δείγματα όγκων χωρίστηκαν τυχαία σε δύο ομάδες, κατά τρόπο ώστε να διατηρηθεί περίπου η κατανομή των δειγμάτων μεταξύ κάθε βαθμίδας όγκου. Τα πρώτα 20 GB, 10 AA και 3 A χρησιμοποιήθηκαν σαν ομάδα εκπαίδευσης (training set), και 19 GB, 5 AA και 2 A χρησιμοποιήθηκαν ως ομάδα δοκιμής (γενίκευσης) (test set). Μια επιπλέον ομάδα 6 αστροκυττωματικών όγκων, αποτελούμενη από 4 AA, για τα οποία η ιστοπαθολογική τους αξιολόγηση αποδείχθηκε ιδιαίτερα δύσκολη, και από 2 δείγματα που ανήκουν στην κατηγορία (grade) I, pilocytic αστροκυττώματα (PA) χρησιμοποίηθηκε για να διαπιστωθεί η ικανότητα γενίκευσης του ΤΝΔ.

Ένας single-layer perceptron χρησιμοποιήθηκε για την διαβάθμιση των δειγμάτων από ιστούς όγκων. Ο αριθμός τον δεδομένων εισόδου ήταν ίσος με τον αριθμό τον γονιδίων-ταξινομητών και το επίπεδο εξόδου αποτελείται από ένα μοναδικό νευρώνα με σιγμοειδή συνάρτηση ενεργοποίησης. Οι τιμές των βαρών αρχικοποιήθηκαν τυχαία και η εκμάθηση έγινε χρησιμοποιώντας standard gradient descent learning rule (ή Delta rule) με learning rate η = 0.05. Η βαθμονόμηση έγινε με την μέθοδο leave-one-out cross-validation. Οι τιμές των βαρών αλλάζουν σε κάθε δείγμα και η βαθμονόμηση τερματίζεται μετά από 100 κύκλους (epochs) σε ολόκληρο το σύνολο δεδομένων εκμάθησης. Οι τελικές παράμετροι για μια ολοκληρωμένη εκμάθηση καθορίζουν ένα «μοντέλο» (βλ. παρακάτω). Ο πηγαίος κώδικας για το ΤΝΔ και τις μεθόδους οπτικοποίησης είναι διαθέσιμος από το: http://www.imbb.forth.gr/people/poirazi/software.html

Βαθμονόμηση και μέθοδος leave-one-out cross validation των ΤΝΔ

Η βαθμονόμηση αναφέρεται στη διαδικασία της βελτιστοποίησης των βαρών και των παραμέτρων ενός δικτύου που πραγματοποιείται κατά τη διάρκεια της εκμάθησης και έχει σαν στόχο την επίτευξη των επιθυμητών τιμών εξόδου από το δίκτυο σαν

συνάρτηση συγκεκριμένων τιμών εισόδου. Για να βαθμονομηθούν τα ΤΝΔ χρησιμοποιήθηκαν για την εκμάθηση προκαθορισμένα δείγματα από το σύνολο των δεδομένων που πρόκειται να αναλυθούν. Η εκμάθηση πραγματοποιείται με τη μέθοδο leave-one-out cross-validation όπου κάθε φορά ένα δείγμα αφήνεται εκτός εκμάθησης και η επιλογή γνωρισμάτων (γονίδια/probe set) καθώς και η εκμάθηση του αλγορίθμου πραγματοποιείται στο τμήμα των δεδομένων που απομένει. Είναι σημαντικό να σημειωθεί ότι για να είναι η μέθοδος leave-one-out cross-validation αξιόπιστη, η ανίχνευση των γονιδίων και η επιλογή πρέπει να πραγματοποιούνται για κάθε κύκλο της μεθόδου. Σε κάθε κύκλο η ανάλυση αποκαλύπτει διαφορετικές υποομάδες γονιδίων οι οποίες, παρόλο που είναι ισχυρά επαναλαμβανόμενες, διαφέρουν σε κάποια γονίδια αντανακλώντας έτσι την ποικιλομορφία και την ετερογένεια στα προφίλ έκφρασης. Σε προηγούμενες μελέτες (104),(25) η επιλογή των γονιδίων πραγματοποιούνταν χρησιμοποιώντας όλα τα δεδομένα εκμάθησης, μια μέθοδος που εισάγει όμως ισχυρή πόλωση στα αποτελέσματα επιβεβαίωσης προκαλώντας 'overfitting' του μοντέλου και επηρεάζοντας την αποτελεσματικότητα του δικτύου. Το overfitting αναφαίρεται στην ανικανότητα των εκπαιδευμένων μοντέλων να γενικεύσουν μεταξύ δειγμάτων εκμάθησης και δειγμάτων δοκιμής. Τα αποτελέσματα γίνονται ακόμα ποιο διφορούμενα στις περιπτώσεις όπου τα δειγμάτων δοκιμής έχουν συμπεριληφθεί στην διαδικασία επιλογής γονιδίων (105). Στην παρούσα μελέτη εξασφαλίστηκε ότι κάθε φορά ένα δείγμα έμενε εκτός και χρησιμοποιούνταν μονό για επιβεβαίωση, ενώ η εκμάθηση και η επιλογή των γονιδίων πραγματοποιούνταν στα υπόλοιπα δείγματα. Το δείγμα που έμενε εκτός κατηγοριοποιήθηκε έπειτα με βάση τα επιλεγμένα γονίδια, βάρη και παραμέτρους του δικτύου που είχαν ήδη αποθηκευτεί κατά την εκμάθηση.

Εκμάθηση και Ταξινόμηση[a]

Η μέθοδος leave-one-out cross validation πραγματοποιήθηκε χρησιμοποιώντας αυξανόμενο αριθμό γονιδίων, όπως ταξινομήθηκαν σύμφωνα με τη διαδικασία σήμα-προς-θόρυβος (βλέπε Επιλογή Γονιδίων) και με άνω όριο ομάδες των 20 γονιδίων για κάθε δείγμα που μένει εκτός μάθησης, ενώ κάθε φορά πραγματοποιούνταν ένα

[a] Ο όρος ***Ταξινόμηση*** αναφέρεται στη γενικότερη ορολογία που αφορά στους αλγορίθμους μηχανικής μάθησης, όπως τα ΤΝΔ. Για τους σκοπούς της παρούσας μελέτης η ταξινόμηση αναφέρεται στην διαβάθμιση διαφορετικών κλάσεων όγκων, βαθμίδων όγκων (βαθμίδωση) και/ή κανονικοί vs. ογκογόνοι ιστοί. Ο όρος *κλάση* χρησιμοποιείται σε ένα γενικό πλαίσιο και μπορεί να αναφέρεται σε μια συγκεκριμένη βαθμίδα όγκου, όπως στην περίπτωση των δεδομένων από αστροκυττωματικούς όγκους.

προκαταρκτικό βήμα όπου τα προφίλ έκφρασης των επιλεγμένων γονιδίων κανονικοποιούνταν ώστε να ισχύει μηδενικός μέσος όρος και μοναδιαία τυπική απόκλιση. Βρέθηκε ότι τα βέλτιστα και περισσότερο συνεπή αποτελέσματα επιτεύθηκαν με 100 επαναλήψεις της μεθόδου leave-one-out cross validation για κάθε δείγμα εκμάθησης και παίρνοντας μετά τον μέσο όρο των 100 αποτελεσμάτων. Τα νευρωνικά δίκτυα διαθέτουν την δυνατότητα να αποθηκεύουν τα εκπαιδευμένα μοντέλα, σαν εκείνα που παρήχθησαν κατά τη leave-one-out cross validation μέθοδο, ούτως ώστε να χρησιμοποιηθούν αργότερα σε άγνωστα για το μοντέλο δείγματα ή δείγματα τα οποία δεν έχουν χρησιμοποιηθεί για την εκμάθηση. Ο συνολικός αριθμός των εκπαιδευμένων μοντέλων εξαρτάται από τον αριθμό των δειγμάτων που χρησιμοποιούνται για εκπαίδευση. Έτσι, για κάθε τύπο μοντέλου ο συνολικός αριθμός των εκπαιδευμένων μοντέλων ήταν 100 x N_i όπου N_i είναι ο αριθμός δειγμάτων εκπαίδευσης.

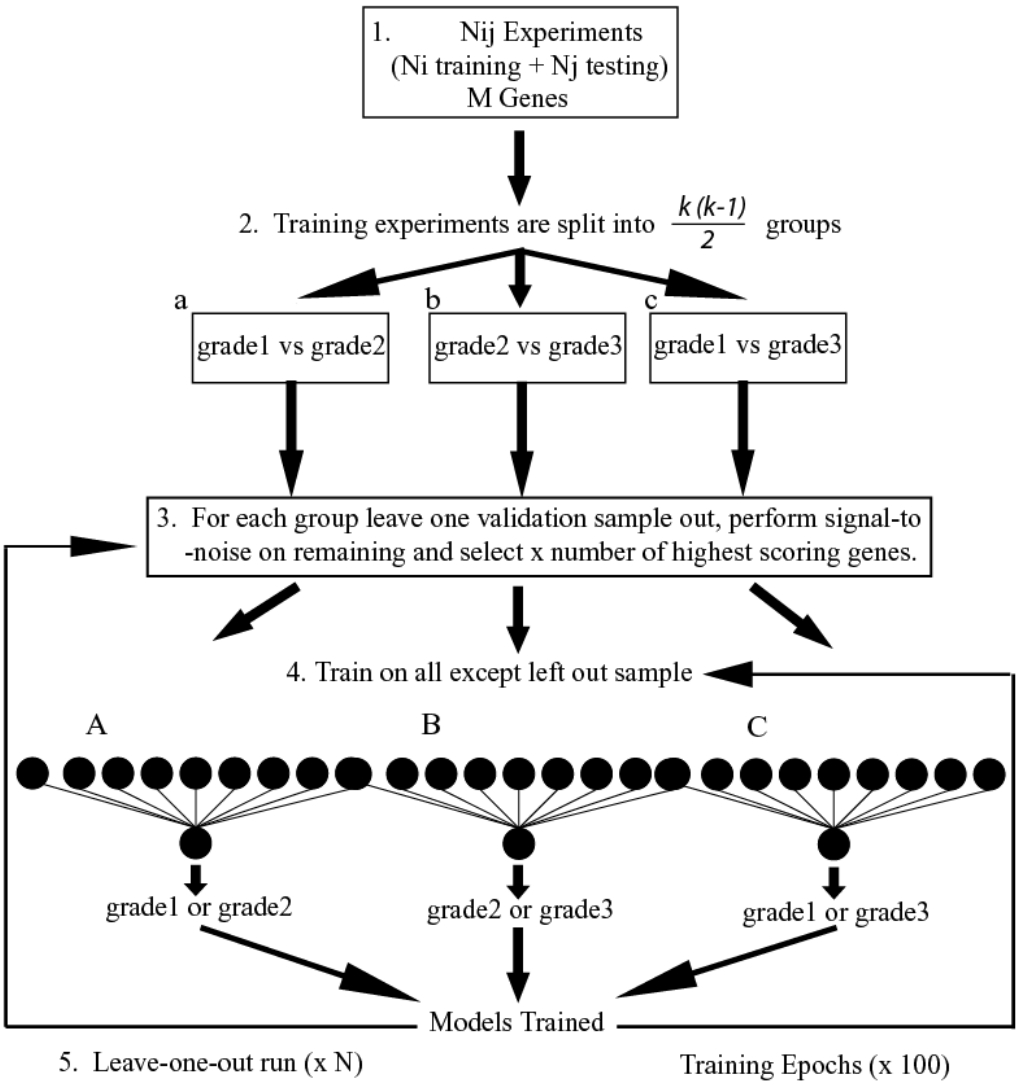
Το πρόβλημα της κατηγοριοποίησης διαχωρίστηκε σε (k(k-1))/2 διαφορετικά προβλήματα δυο κλάσεων/βαθμίδων όπου k είναι ο αριθμός των κλάσεων/βαθμίδων (προσέγγιση ***all-pairs*** – Σχήμα 2.1). Συνεπώς το δίκτυο χρησιμοποιήθηκε για να πάρουμε 3 διαφορετικά είδη μοντέλων δυαδικών ταξινομητών.

Για κάθε δείγμα προς κατηγοριοποίηση στο πρόβλημα δυο κλάσεων/βαθμίδων που μελετάμε, υπολογίστηκε ο μέσος όρος από 100 εξόδους μοντέλων. Τα δείγματα κατηγοριοποιήθηκαν ως κλάσεις 0 και 1 ανάλογα με το αν ο μέσος όρος ήταν πλησιέστερα στο 0 ή στο 1 με κατώφλι για την απόφαση αυτή το 0.5.

Η χρήση συγκεντρωτικών νευρωνικών δικτύων για να πάρουμε ένα μέσο όρο κατηγοριοποίησης είναι ένα πολύ ισχυρό εργαλείο στην επίτευξη καλύτερης κατηγοριοποίησης κάτι που ερμηνεύεται καλύτερα στατιστικά όταν χρησιμοποιούνται πολύπλοκα δεδομένα όπως είναι τα δεδομένα από μικροσυστοιχίες. Ένα από τα πλεονεκτήματα της χρήσης τεχνητών νευρωνικών δικτύων είναι ότι μπορούν να εκπαιδευτούν για την κατηγοριοποίηση δειγμάτων σε πολλαπλές κλάσεις ή βαθμίδες αυξάνοντας τον αριθμό των νευρώνων εξόδου στο τελικό επίπεδο του δικτύου. Τα μειονεκτήματα έγκειται στην προσπάθεια να κατηγοριοποιηθούν μικρότεροι αριθμοί κλάσεων/βαθμίδων (όπως στην περίπτωση των δεδομένων από αστροκυττωματικούς όγκους) καθώς οι νευρώνες συναγωνίζονται μεταξύ τους με μια προτίμηση για νευρώνες που αντιστοιχούν σε πολυπληθέστερες κλάσεις/βαθμίδες. Σε μια προσπάθεια να υιοθετηθεί ένα νευρωνικό δίκτυο το οποίο αναγνωρίζει επιτυχώς

τα όρια πολλαπλών κλάσεων/βαθμίδων, είναι πιο ασφαλές να χωριστούν τα δεδομένα σε πολλαπλά δυαδικά προβλήματα και αυτά στη συνέχεια να χρησιμοποιηθούν για να την εκμάθηση πολλαπλών δικτύων ώστε να ληφθεί ένας μέσος όρος από εκπαιδευμένους 'ειδικούς'. Έτσι, στην περίπτωση των δεδομένων από αστροκυττωματικούς όγκους, αντί να αυξήσουμε απλά τον αριθμό των νευρώνων εξόδου για να συμπίπτει με τον αριθμό των κλάσεων ή βαθμίδων, χρησιμοποιούμε την προσέγγιση ***all-pairs***. Η μέθοδος αυτή χρησιμοποιείται συνήθως με σκοπό να προκύψει ένας δυαδικός ταξινομητής εφαρμόσιμος σε προβλήματα πολλαπλών κλάσεων και έχει χρησιμοποιηθεί πρόσφατα σε support vector machines (25). Έχει δειχθεί (104) ότι χωρίζοντας ένα πρόβλημα σε διαφορετικά υπο-έργα και λαμβάνοντας έτσι ένα πλήθος από 'ειδικούς' βελτιώνει την κατηγοριοποίηση σε ικανοποιητικό βαθμό.

Τέλος, τα εκπαιδευμένα μοντέλα από το leave-one-out cross-validation με τα καλύτερα αποτελέσματα για ένα συγκεκριμένο αριθμό επιλεγμένων γονιδίων χρησιμοποιήθηκαν αργότερα για την κατηγοριοποίηση άγνωστων όγκων (γενίκευση).



Σχήμα 2.1 Σχηματική αναπαράσταση της προσέγγισης all-pairs χρησιμοποιώντας τα προτεινόμενα τεχνητά νευρωνικά δίκτυα σε ένα τριών βαθμίδων πρόβλημα όπως αυτό στην περίπτωση των δεδομένων αστροκυττώματος. Τα δεδομένα χωρίστηκαν σε ένα σύνολο δοκιμής και ένα σύνολο εκπαίδευσης (1). Έπειτα, τα πειράματα τεστ απομονώθηκαν και τα πειράματα εκπαίδευσης μοιράστηκαν σε (k(k-1))/2 ομάδες, στην συγκεκριμένη περίπτωση 3, όπου k είναι ο αριθμός των βαθμίδων (2): *α-βαθμίδα1 με βαθμίδα2, b-βαθμίδα2 με βαθμίδα3 και c βαθμιδα1 με βαθμίδα3.* Κάθε μία από αυτές τις ομάδες χρησιμοποιήθηκε για να εκπαιδεύσει τρεις διαφορετικούς τύπους μοντέλων – A, B και C αντίστοιχα. Η διαδικασία leave-one-out cross-validation χρησιμοποιήθηκε για επιβεβαίωση και βαθμονόμηση των δικτύων. Ένας αριθμός χ γονιδίων επιλέχθησαν χρησιμοποιώντας τη μέθοδο signal-to-noise, όπου το

χ μπορεί να παίρνει οποιαδήποτε άρτια τιμή μεταξύ 2 και 20 (3). Για κάθε δείγμα που μένει εκτός, τα μοντέλα των ΤΝΔ βαθμονομήθηκαν χρησιμοποιώντας ως είσοδο χ γονίδια με την υψηλότερη τιμή όπως αποδίδεται από τη μέθοδο signal-to-noise επιλεγμένα από τα εναπομείναντα δείγματα και ως έξοδο τη βαθμίδωση των όγκων. Εδώ αναπαρίσταται η περίπτωση των 9 γονιδίων (χ = 9) καθώς υπάρχουν 9 κόμβοι εισόδου (4). Για κάθε τύπο μοντέλου (*A, B και C)* η βαθμονόμηση βελτιστοποιήθηκε στους 100 επαναληπτικούς κύκλους (φάσεις). Το επόμενο δείγμα εκμάθησης στη σειρά αφέθηκε εκτός και ολόκληρη η διαδικασία εκμάθησης επαναλήφθηκε (5). Για κάθε δείγμα που αφήνεται εκτός βαθμονομούνται 100 μοντέλα φτάνοντας σε ένα σύνολο 100 x N_{ic} εκπαιδευμένα μοντέλα για κάθε ομάδα (όπου το c αντιστοιχεί στις διάφορες βαθμίδες c= 1,2,3 και N_{ic} είναι ο αριθμός των εκπαιδευμένων μοντέλων για κάθε βαθμίδα). Τα πειράματα δοκιμής (γενίκευσης) σε δείγματα που δεν συμπεριλήφθηκαν στην διαδικασία leave-one-out cross-validation έγιναν χρησιμοποιώντας όλα τα βαθμονομημένα/εκπαιδευμένα μοντέλα.

Γενίκευση σε άγνωστα δείγματα όγκων (δείγματα δοκιμής)

Η διαδικασία γενίκευσης πραγματοποιήθηκε επιλέγοντας τους αντίστοιχους δείκτες γονιδίων από το σύνολο δοκιμής (γενίκευσης), κανονικοποιόντας τα δεδομένα δοκιμής και έπειτα περνώντας κάθε δείγμα από όλα τα εκπαιδευμένα μοντέλα. Τα γονίδια που αντιστοιχούν σε διαφορετικά εκπαιδευμένα μοντέλα ίσως διαφέρουν καθώς διαφορετικό δείγμα αφέθηκε εκτός κατά τη διάρκεια του leave-one-out cross-validation και η επιλογή των γονιδίων πραγματοποιήθηκε στα εναπομείναντα δείγματα. Έτσι, κατά τη διάρκεια της γενίκευσης τα συγκεκριμένα γονίδια που αντιστοιχούν σε ένα δεδομένο μοντέλο επιλέχθησαν από το σύνολο δοκιμής (γενίκευσης) κάθε φορά που ένα δείγμα περνούσε μέσα από καθένα μοντέλο. Έπειτα υπολογίστηκε ο μέσος όρος από τις εξόδους των μοντέλων. Εάν ο μέσος όρος από τα εκπαιδευμένα μοντέλα είναι μεγαλύτερος από 0.5 το δείγμα κατηγοριοποιήθηκε στη κλάση 1, αν ήταν μικρότερος από 0.5 το δείγμα θεωρήθηκε ότι ανήκει στην κλάση 0. Εάν ο μέσος όρος των εξόδων από τα εκπαιδευμένα μοντέλα ήταν ακριβώς 0.5 το δείγμα θεωρήθηκε ότι είναι μη καθορισμένο. Η μέθοδος αυτή κατηγοριοποιεί κάθε δείγμα δοκιμής (γενίκευσης) σε μία από τις δύο κλάσεις που είναι παρούσες στα εκπαιδευμένα μοντέλα. Για τη μέθοδο all-pairs που χρησιμοποιήθηκε εδώ με σκοπό να τοποθετηθούν τα δείγματα δοκιμής (γενίκευσης) σε μια από τις 3 κατηγορίες, μία συνψηφισμένη απόφαση πάρθηκε από τους 3 τύπους μοντέλων. Για κάθε κλάση

υπάρχουν δυο σχετικά μοντέλα που τα διαχωρίζουν από τα υπόλοιπα. Παρόλα αυτά και εφόσον έχουμε άγνοια για το ποιά είναι αυτά, οι έξοδοι από όλους τους δυαδικούς ταξινομητές κατεγράφησαν και κάθε δείγμα κατηγοριοποιήθηκε στην κλάση εκείνη όπου η πλειονότητα των εξόδων συμφωνούν με αυτό. Σε περιπτώσεις όπου και οι τρεις προβλέψεις διαφωνούσαν, το δείγμα τοποθετήθηκε στην κλάση εκείνου του μοντέλου με την πλέον σημαντική τιμή εξόδου (βλέπε Πίνακα Α στο παράρτημα).

Ιεραρχική Ομαδοποίηση (Hierarchical clustering) των Νευρωνικών Εξόδων

Η εκπαίδευση μέσω του leave-one-out cross validation κατέληξε σε ένα σύνολο 6600 εκπαιδευμένων μοντέλων δικτύων (3000: GBvsAA; 2300: GBvsA; 1300: AAvsA). Με τη προσπέλαση των 59 δειγμάτων εκπαίδευσης και δοκιμής (γενίκευσης) μέσα από τα ΤΝΔ μοντέλα, προέκυψαν 6600 διαφορετικοί έξοδοι για κάθε δείγμα. Αυτοί κατέληξαν σε πίνακα 6600-by-59, ο οποίος στη συνέχεια χρησιμοποιήθηκε στην ομαδοποίησης, έτσι ώστε να βελτιωθεί η οπτικοποίηση των ΤΝΔ μας και να παραχθεί μια πιο πληροφοριακή αναπαράσταση των αποτελεσμάτων. Επιπλέον, σε αυτή τη μορφή, τα αποτελέσματα μπορούν εύκολα να συγκριθούν με αποτελέσματα ομαδοποίησης τιμών γονιδιακής έκφρασης.

Ανάλυση Επιβίωσης

Η μέθοδος Kaplan-Meier χρησιμοποιήθηκε για την εκτίμηση των κατανομών επιβίωσης (106). Το τεστ Log rank χρησιμοποιήθηκε για να αξιολογηθεί η διαφορά μεταξύ των ομάδων επιβίωσης. Για όλες τις αναλύσεις θεωρήθηκε σημαντική τιμή του $p < 5.0e^{-2}$. Η στατιστική ανάλυση πραγματοποιήθηκε με το λογισμικό πακέτο ελεύθερης πρόσβασης R.

2.3 Αποτελέσματα

2.3.1 Επιλογή Γονιδίων Ταξινομητών

Εκπαίδευση του ΤΝΔ στην διάκριση μεταξύ διαφόρων βαθμίδων αστροκυττωματικών όγκων και επιλογή των γονιδίων ταξινομητών

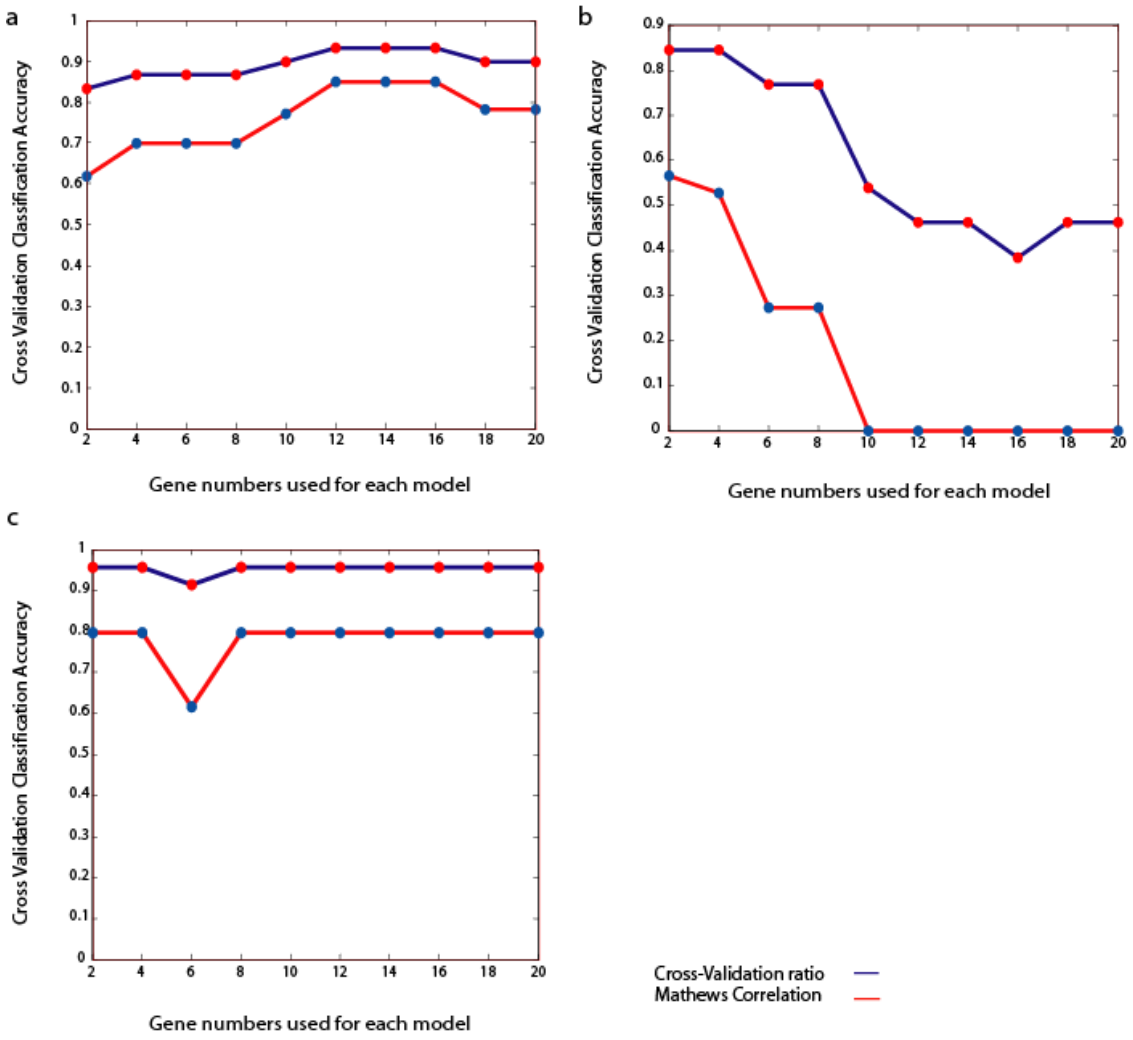
Η εκπαίδευση / βαθμονόμηση πραγματοποιήθηκε με μια all-pairs προσέγγιση, όπου το μεμονωμένο πρόβλημα διαφοροποίησης μεταξύ τριών κατηγοριών (GB, AA, and A), οριοθετήθηκε σε πολλαπλά προβλήματα 2 κατηγοριών (Σχήμα 2.1, Υλικά και Μέθοδοι). Τα 33 δείγματα που χρησιμοποιήθηκαν για την εκπαίδευση, χωρίστηκαν σε τρεις ομάδες δειγμάτων, κάθε μια εκ' των οποίων αποτελούνταν από 2 βαθμίδες όγκων, δηλαδή (α) GB-AA, (β) AA-A και (γ) GB-A. Στη συνέχεια, τρεις διαφορετικοί τύποι μοντέλων ΤΝΔ (Α, Β και Γ) εκπαιδεύτηκαν, κάθε ένας εκ' των οποίων αντιστοιχούσε στην αντίστοιχη ομάδα εκπαίδευσης. Για κάθε έναν από τους τύπους μοντέλων, τα γονίδια τα οποία εμφάνιζαν διαφορετική έκφραση μεταξύ των δύο κατηγοριών υπο εξέταση, επιλέγονταν βάσει της μεθόδου σήματος-προς-θόρυβο (107) σε ολόκληρο το U133A chip. Η επίδοση κατά την εκπαίδευση και ο βέλτιστος αριθμός γονιδίων που απαιτούνταν για τη διαβάθμιση προσδιορίστηκε βάσει της μεθόδου leave-one-out cross-validation. Για κάθε leave-one-out κύκλο, τα γονίδια ταξινομούνται σύμφωνα με την τιμή του σήματος-προς-θόρυβο (για όλα εκτός από το δείγμα που εξαιρείται), και έπειτα η ακρίβεια διαβάθμισης προσδιορίζεται με τη χρήση αυξανόμενου αριθμού των γονίδιων υπό κατάταξη. Οι βέλτιστες τιμές που επιτεύχθηκαν κατά τη διαδικασία της Leave-one-out cross-validation ήταν 93.3%, 84.6% και 95.6% χρησιμοποιώντας 44, 9 και 7 γονίδια για τις GB-AA, AA-A και GB-A ομάδες βαθμίδων αντίστοιχα (για λεπτομέρειες βλέπε επόμενη ενότητα).

Γραφήματα Leave-one-out cross-validation για την βαθμονόμηση αστροκυττωματικών όγκων

Για την δοκιμασία GB vs. AA η διαδικασία της leave-one-out cross-validation έφτασε σε μέγιστη τιμή επιτυχίας όταν επιλέχτηκαν ομάδες των 12, 14 και 16 γονιδίων σε κάθε leave-one-out κύκλο. Το ποσοστό ακρίβειας της leave-one-out cross-validation που επιτεύχθηκε ήταν της τάξεως του 93.33%, το οποίο αναλογεί σε 2/30 δείγματα με

λανθασμένη διαβάθμιση[b] (AA76 και GB133) (Σχήμα 2.2, α). Σε αυτή τη δυαδική διαβάθμιση η ακρίβεια του ποσοστού και του Mathews Correlation (βλέπε *Sensitivity, Specificity and Mathews Correlation* στο παράρτημα) ήταν η ίδια και για τις τρεις ομάδες γονιδίων, έτσι προκειμένου να επιλεγεί το βέλτιστο leave-one-out cross-validation μοντέλο, χρησιμοποιήθηκε το error margin score (βλέπε *Error Margin (EM) Score* στο παράρτημα). Τα μοντέλα επαλήθευσης, εκπαιδευμένα χρησιμοποιώντας ομάδες των 16 γονιδίων, είχαν το χαμηλότερο error margin score και χρησιμοποίθηκαν στα δεδομένα δοκιμής (γενίκευσης). Για την δοκιμασία *AA vs. A* (Σχήμα 2.2, b), η μέγιστη ακρίβεια της leave-one-out cross-validation ήταν 84.62% η οποία αντιστοιχεί σε 2/13 δείγματα με λανθασμένη διαβάθμιση. Αυτό το επίπεδο ακρίβειας επιτεύχθηκε όταν σε κάθε leave-one-out κύκλο επιλέχθηκαν ομάδες γονιδίων των δύο και των τεσσάρων με το υψηλότερο σήμα-προς-θόρυβο score. Ωστόσο, εξετάζοντας το Mathews Correlation ήταν εμφανές ότι το επίπεδο ακρίβειας ήταν σημαντικότερο για τις ομάδες γονιδίων των δύο (2 δείγματα με λανθασμένη διαβάθμιση, ένα για κάθε βαθμίδα, A26 και AA92, έναντι 2 δείγματα με λανθασμένη διαβάθμιση, και τα δύο για το grade "A", για την περίπτωση των ομάδων των τεσσάρων γονιδίων). Για την δοκιμασία *GB vs. A,* (Σχήμα 2.2, c) το μέγιστο leave-one-out cross-validation score ήταν 95.65%, το οποίο αντιστοιχεί σε ένα δείγμα με λανθασμένη διαβάθμιση (A9) από το σύνολο των 23. Αυτό το επίπεδο ακρίβειας επιτεύχθηκε όταν επιλέχθηκαν ομάδες των 2, 4, 8, 10, 12, 14, 16, 18 και 20 γονιδίων με το υψηλότερο σήμα-προς-θόρυβο score σε κάθε leave-one-out κύκλο. Όπως και στη δοκιμασία *GB vs. AA,* το Mathews Correlation δεν υπέδειξε σημαντική διαφορά, ωστόσο, το μοντέλο που περιελάμβανε τις ομάδες των δύο γονιδίων είχε το χαμηλότερο error margin score (τα δεδομένα δεν παρουσιάζονται), έτσι επιλέχθηκε για τη διαδικασία γενίκευσης.

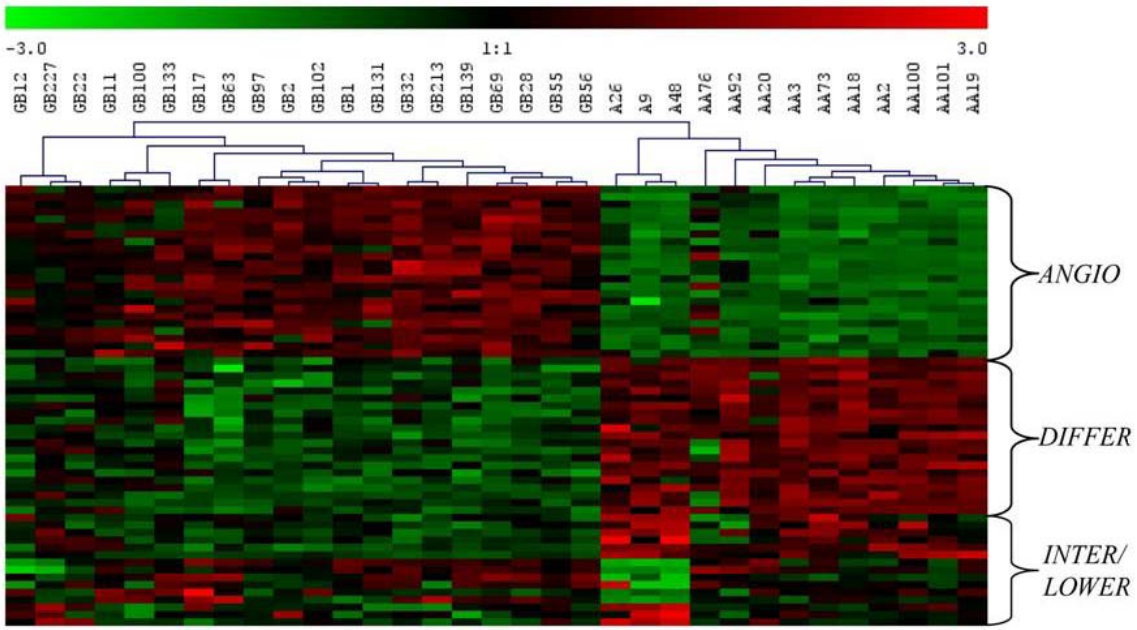
[b]Ο όρος λανθασμένη διαβάθμιση (mis-graded) χρησιμοποιείται για να χαρακτηρίσει δείγματα για τα οποία δεν υπάρχει συμφωνία μεταξύ της εξόδου του μοντέλου και της αρχική ιστοπαθολογικής διάγνωσης.



Σχήμα 2.2 Διαγράμματα της ανάλυσης leave-one-out cross-validation για τους τρεις διαφορετικούς τύπους μοντέλων που εκπαιδεύτηκαν. **α. (GB vs. AA)** – Στο διάγραμμα παρουσιάζεται η επίδοση της leave-one-out cross-validation για την ομάδα 'a' των δειγμάτων εκπαίδευσης. Όπως φαίνετε η leave-one-out cross-validation προσεγγίζει το ποσοστό μέγιστης επιτυχίας όταν σε κάθε leave-one-out γύρο επιλέγονται ομάδες των 12, 14 και 16 γονιδίων με το υψηλότερο signal-to-noise score. Το Error margin score υπέδειξε τα μοντέλα με ομάδες των 16 γονιδίων για να χρησιμοποιηθούν για δοκιμή. **β. (AA vs. A)** – Στο διάγραμμα παρουσιάζεται η επίδοση στη leave-one-out cross-validation για την ομάδα 'b' των δειγμάτων εκπαίδευσης. Τόσο το ποσοστό ακρίβειας της leave-one-out cross-validation, όσο και το Mathews Correlation αποκάλυψαν ότι η leave-one-out είχε το μέγιστο ποσοστό επιτυχίας όταν σε κάθε leave-one-out κύκλο επιλέχθηκαν ομάδες των 2 γονιδίων με το υψηλότερο σήμα-προς-θόρυβο score. **γ. (GB vs. A)** – επίδοση στη leave-one-out cross-validation για την ομάδα 'c' των δειγμάτων εκπαίδευσης. Το βέλτιστο ποσοστό

επιτυχίας επιτεύχθηκε όταν σε κάθε leave-one-out κύκλο επιλέχθηκαν ομάδες των 2, 4, 8, 10, 12, 14, 16, 18, 20 γονιδίων με το υψηλότερο σήμα-προς-θόρυβο score. Το μοντέλο των 2 γονιδίων επιλέχθηκε για τη διαδικασία γενίκευσης λόγω του γεγονότος ότι είχε το χαμηλότερο error margin score.

Συνάζοντας όλα τα γονίδια/ probe sets και ταυτόχρονα εξαλείφοντας επαναλήψεις, είχε ως αποτέλεσμα ένα σύνολο από 59 γονίδια/probe. Όπως ήταν αναμενόμενο, η ιεραρχική ομαδοποίηση όλων των δειγμάτων εκπαίδευσης με την χρήση των προαναφερόμενων probes έδειξε καθαρές διαφοροποιήσεις μεταξύ των βαθμίδων όγκων GB, AA και A, και προσδιόρισε 3 λειτουργικές κλάσεις, οι οποίες περιγράφουν 3 μοριακούς υπότυπους όγκων. (Σχήμα 2.3 – για λεπτομέρειες βλέπε επόμενη ενότητα). Στη συνέχεια τα εκπαιδευμένα / βαθμονομημένα μοντέλα ΤΝΔ (βλέπε Υλικά και Μέθοδοι) χρησιμοποιήθηκαν για τη διαβάθμιση των δειγμάτων στην ομάδα δοκιμής (γενίκευσης).



Σχήμα 2.3 Ιεραρχική ομαδοποίηση των 33 δειγμάτων εκπαίδευσης (20GB, 10 AA and 3 A) με τη χρήση 59 probe, τα οποία επιλέχτηκαν από το S2N. Το υπολογιστικό εργαλείο MeV (108) χρησιμοποιήθηκε για να πραγματοποιήσει την ιεραρχική ομαδοποίηση, με τη χρήση Ευκλείδειας απόστασης και τον αλγόριθμο complete linkage. Τα δείγματα επισημαίνονται με τις ανάλογες διαβαθμίσεις και τα γονίδια επισημαίνονται συμφώνα με τους μοριακούς υπότυπους όγκων (*ANGIO*, *DIFFER* or *INTER/LOWER*) που τους χαρακτηρίζουν. Οι τιμές της έκφρασης των γονιδίων είναι κανονικοποιημένες ώστε να έχουν μέσο όρο μηδέν και μοναδιαία τυπική απόκλιση. Με κόκκινο χρώμα αναπαριστάται η αυξανόμενη έκφραση ενώ με πράσινο χρώμα η μειωμένη έκφραση.

Τα προφίλ έκφρασης των γονιδίων ταξινομητών που επιλέχθηκαν κατά την εκπαίδευση, προσδιορίζουν τρεις μοριακούς υπότυπους όγκων

Διεξοδική εξέταση των επιλεγμένων γονιδίων ταξινομητών, τα περισσότερα εκ' των οποίων προσδιορίστηκαν και με Bayesian ανάλυση (Πίνακας 2.1), αποκάλυψε δύο ενδιαφέροντα χαρακτηριστικά. Κατά πρώτον, τα ομαδοποιημένα γονίδια εμπίπτουν σε τρεις κύριες λειτουργικές κατηγορίες, και κατά δεύτερον, αυτές οι λειτουργικές κατηγορίες προσδιορίζουν τρεις μοριακούς υπότυπους όγκων.

Ο πρώτος υπότυπος παρουσίαζε σημαντικά αυξημένη έκφραση γονιδίων που εμπλέκονται σε: i) επούλωση τραύματος (*ADM*, *PDGFa*, *EFEMP2*), ii) extracellular

matrix constituents και remodelling machinery (*LGALS1* and *3*, *PLAT*, *TIMP1*, *COL5A2*) και iii) κυτταρική προσκόλληση (*PARD3*, *DAG1, Kindlin1*, *ZYX* και *ALCAM*). Καθώς και οι τρεις αυτές λειτουργίες είναι απαραίτητες για τις αγγειογενετικές angiogenic ιδιότητες των κυττάρων, αυτός ο υπότυπος ονομάστηκε *ANGIO* και ήταν χαρακτηριστικός των GB δειγμάτων, του grade IV. Η επόμενη ομάδα ήταν ένα κράμα από ιστοπαθολογικούς και μοριακούς υπότυπους, και παρουσίασε αυξημένη έκφραση γονιδίων που εμπλέκονται σε i) κυτταρική σηματοδότηση και ανάπτυξη (*BMP2, ABI1, REPS2, ADCY2, NET1*), ii) βιοσύνθεση πρωτεϊνών (*RPL22, ZMYND11*) και στον iii) κυτταρικό κύκλο (*PARD3, ZMYND11, CLASP2*). Αυτή η ομάδα, που ονομάστηκε *DIFFER,* χαρακτηρίζει τα δείγματα των grade II και III, τα οποία παρά το γεγονός ότι είναι ενεργά σε ανάπτυξη και νευρωνική διαφοροποίηση, δεν έχουν ακόμα αποκτήσει angiogenic ιδιότητες. Αυτή η ομάδα αναλύθηκε περαιτέρω, με τη χρήση μιας ομάδας γονιδίων που κωδικοποιούν τις ankyrin repeat proteins (*ANK3*, *ANKS1B*), solute carrier proteins (*SLCO1A2, SLC34A1*), μια πρωτεΐνη η οποία εμπλέκεται στην απόπτωση (*DNAJA3*) και την *PEA15*, μια cytostatic και αντι-αποπτωτική φωσφοπρωτεΐνη που εμφανίζεται σε μεγάλες ποσότητες στα αστροκύτταρα (109). Αυτή η ανάλυση οδήγησε στο διαχωρισμό της ομάδας *DIFFER* στον υπότυπο *INTER* (*Intermediate*), ο οποίος ήταν χαρακτηριστικός των δειγμάτων του επιπέδου III, και στον υπότυπο *LOWER*, ο οποίος ήταν χαρακτηριστικός των δειγμάτων του επιπέδου II.

Πίνακας 2.1 Τρεις ομάδες επιλεγμένων γονιδίων, η κάθε μία από τις οποίες προέκυψε από μία από τις τρεις συγκρίσεις όγκων που πραγματοποιήθηκαν ανα δυάδες (pairwise): a) GB-AA (γονίδια *ANGIO/DIFFER*), b) GB-A (γονίδια *INTER/LOWER*), c) AA-A (γονίδια *INTER/LOWER*).

a)

Gene Symbol	*Gene Name*	*Bayesian p value*	*Mean expression fold change*	*Gene class*
ADM	Adrenomedullin	2.62E-05	11.79	*ANGIO*
TIMP1	tissue inhibitor of metalloproteinase 1	1.51E-08	11.56	*ANGIO*
FABP5	fatty acid binding protein 5	1.85E-04	9.41	*ANGIO*
EMP3	epithelial membrane protein 3	5.27E-07	7.58	*ANGIO*
PDPN	Podoplanin	3.22E-05	6.00	*ANGIO*
LGALS3	lectin galactoside-binding soluble 3 (galectin 3)	1.02E-05	5.86	*ANGIO*
LGALS1	lectin galactoside-binding soluble 1 (galectin 1)	1.02E-05	4.41	*ANGIO*
PDGFA	platelet-derived growth factor alpha polypeptide	2.03E-05	4.09	*ANGIO*
PLAT	plasminogen activator tissue	6.60E-05	3.97	*ANGIO*
EFEMP2	EGF-containing fibulin-like extracellular matrix protein 2	2.20E-06	3.92	*ANGIO*
COL5A2	collagen type V alpha 2	3.71E-05	3.72	*ANGIO*
COL5A2	collagen type V alpha 2	1.02E-05	3.60	*ANGIO*
DDA3	differential display and activated by p53	2.06E-05	3.53	*ANGIO*
TAGLN2	transgelin 2	3.15E-05	3.19	*ANGIO*
DUSP6	dual specificity phosphatase 6	5.49E-05	3.14	*ANGIO*
LDHA	lactate dehydrogenase A	7.84E-05	2.84	*ANGIO*
PLP2	proteolipid protein 2	6.34E-05	2.74	*ANGIO*
EFEMP2	EGF-containing fibulin-like extracellular matrix protein 2	7.34E-05	2.43	*ANGIO*
CENTD3	centaurin delta 3	2.57E-04	2.42	*ANGIO*
KIAA0495	KIAA0495	1.69E-04	2.15	*ANGIO*
DAG1	dystroglycan 1 (dystrophin-associated glycoprotein 1)	2.73E-05	1.80	*ANGIO*
ZYX	Zyxin	1.46E-04	1.78	*ANGIO*
OSBPL10	oxysterol binding protein-like 10	2.63E-04	1.75	*ANGIO*
CUTC	cutC copper transporter homolog	1.74E-05	-1.59	*DIFFER*
TNKS2	TRF1-interacting ankyrin-related ADP-ribose polymerase 2	6.60E-05	-1.67	*DIFFER*
HSA9761	dimethyladenosine transferase	8.27E-05	-1.71	*DIFFER*
KIAA1279	KIAA1279	2.03E-05	-1.76	*DIFFER*
RPL22	ribosomal protein L22	2.83E-04	-1.81	*DIFFER*
ENAH	enabled homolog	6.26E-05	-1.82	*DIFFER*
ZMYND11	zinc finger MYND domain containing 11	4.53E-05	-1.87	*DIFFER*
HNRPH3	heterogeneous nuclear ribonucleoprotein H3	3.65E-05	-1.88	*DIFFER*
RPL22	ribosomal protein L22	2.62E-05	-1.93	*DIFFER*
CLASP2	cytoplasmic linker associated protein 2	2.81E-04	-2.05	*DIFFER*
USH1C	Usher syndrome 1c (autosomal recessive severe)	1.02E-05	-2.10	*DIFFER*
RAP2A	RAP2A	7.90E-04	-2.10	*DIFFER*
ALCAM	activated leukocyte cell adhesion molecule	4.72E-03	-2.14	*DIFFER*
ABI1	abl-interactor 1	8.27E-05	-2.16	*DIFFER*
PARD3	par-3 partitioning defective 3 homolog	3.43E-06	-2.30	*DIFFER*
CRYAB	crystallin, alpha B	5.80E-05	-2.72	*DIFFER*
NAP1L3	nucleosome assembly protein 1-like 3	1.77E-04	-2.88	*DIFFER*
NET1	neuroepithelial cell transforming gene 1	2.20E-06	-2.93	*DIFFER*
C20ORF42	chromosome 20 open reading frame 42	2.87E-04	-3.09	*DIFFER*
BMP2	bone morphogenetic protein 2	5.49E-05	-3.27	*DIFFER*

ADCY2	adenylate cyclase 2 (brain)	3.05E-05	-3.79	*DIFFER*

b)

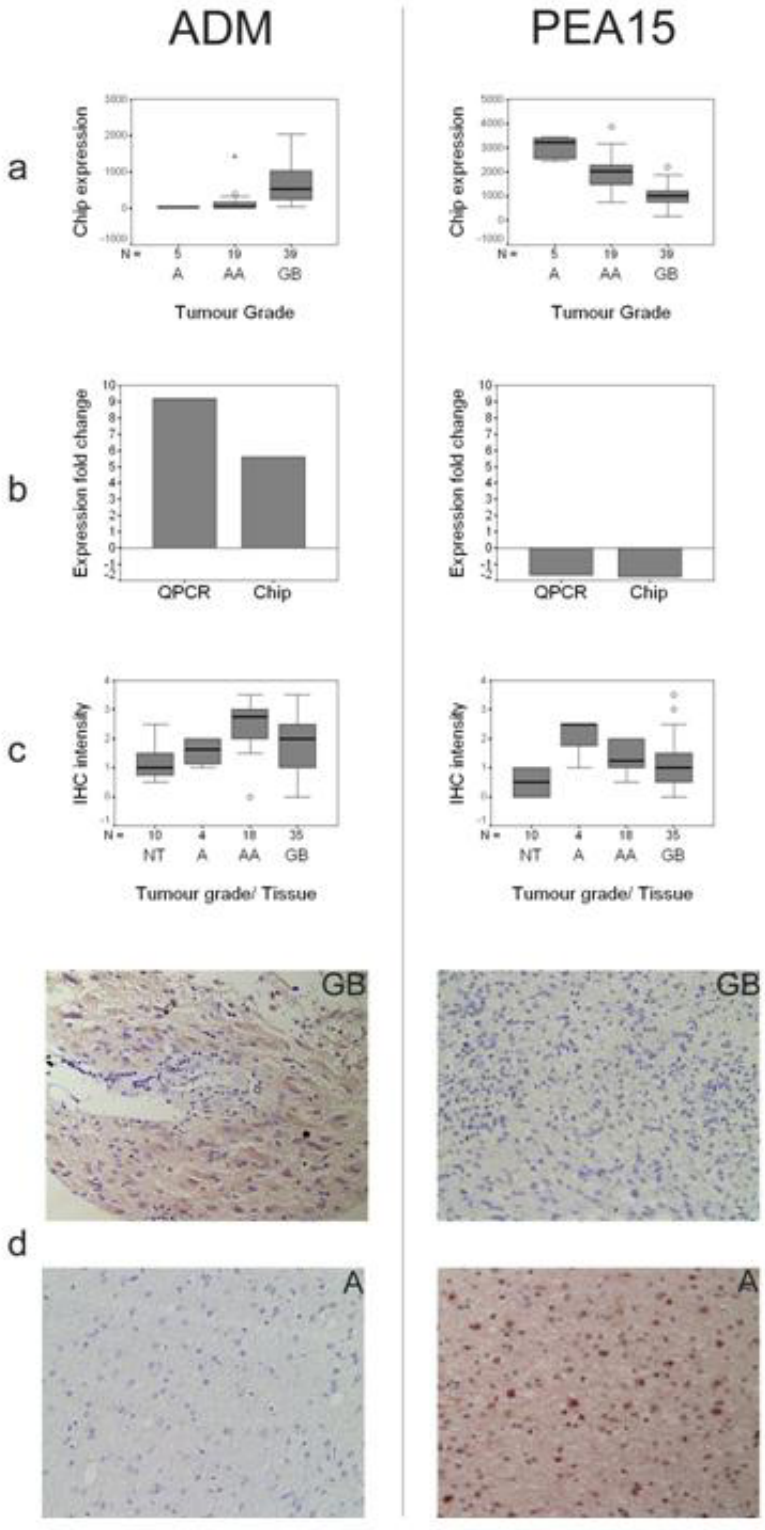
Gene Symbol	***Gene Name***	***Bayesian p value***	***Mean expression fold change***	***Gene class***
SLC34A1	solute carrier family 34 (sodium phosphate) member 1	1.46E-02	-1.35	*LOWER*
RSNL2	restin-like 2	1.88E-02	-1.39	*LOWER*
REPS2	RALBP1 associated EPS domain containing 2	1.53E-03	-1.94	*LOWER*
SLCO1A2	solute carrier organic anion transporter family member 1A2	4.46E-03	-2.06	*LOWER*
PEA15	phosphoprotein enriched in astrocytes 15	1.93E-03	-2.12	*LOWER*
USH1C	Usher syndrome 1C	1.53E-03	-3.49	*LOWER*

c)

Gene Symbol	***Gene Name***	***Bayesian p value***	***Mean expression fold change***	***Gene class***
B2M	beta-2-microglobulin	3.69E-01	2.50	*INTER*
SCP2	sterol carrier protein 2	4.79E-01	1.81	*INTER*
DDOST	dolichyl-diphosphooligosaccharide-protein glycosyltransferase	5.68E-01	1.79	*INTER*
NPTN	neuroplastin	8.49E-01	1.39	*INTER*
TAP2	transporter 2 ATP-binding cassette sub-family B	6.92E-01	1.39	*INTER*
DNAJA3	DnaJ (hsp40) homolog subfamily A, member 3	6.48E-01	1.25	*INTER*
ANKS1b	ankyrin repeat and sterile alpha motif domain containing 1b	5.17E-01	-1.21	*LOWER*
---	DKFZp434M083	5.68E-01	-1.22	*LOWER*
ANK3	ankyrin 3 node of Ranvier	9.07E-02	-1.93	*LOWER*

Ταξινομήσεις γονιδίων ιδιαίτερου βιολογικού ενδιαφέροντος

Δύο γονίδια ταξινομητές παρουσίασαν ιδιαίτερο βιολογικό ενδιαφέρον: η εμπλουτισμένη σε αστροκύτταρα 15 φωσφοπρωτείνη (*PEA15*, 1q21.1 - *LOWER*) και η adrenomedullin (*ADM*, 11p15.4 - *ANGIO*). Αυτά τα γονίδια βρέθηκε ότι εκφράζονται διαφορετικά μεταξύ των καρκινικών σταδίων GB-A και/ή GB-AA μέσω εμπειρικής Bayesian ανάλυσης. Οι μεταβολές έκφρασης επιβεβαιώθηκαν και με QPCR και με ανοσοιστοχημεία (IHC) (Σχήμα 2.4). 23 επιπλέον διαφορικά εκφραζόμενα γονίδια που ταυτοποιήθηκαν με Bayesian ανάλυση επιβεβαιώθηκαν επιτυχώς μέσω QPCR. Η συσχέτιση (R^2) μεταξύ Affymetrix και QPCR GB/AA expression fold changes γι' αυτά τα γονίδια ήταν μεγαλύτερη από 0.8 (τα δεδομένα δεν παρουσιάζονται).


ADM
PEA15
a
Chip expression
N =
5
19
39
A
AA
GB
Tumour Grade
b
Expression fold change
QPCR
Chip
c
IHC intensity
N =
10
4
18
35
NT
A
AA
GB
Tumour grade/ Tissue
d
GB
A

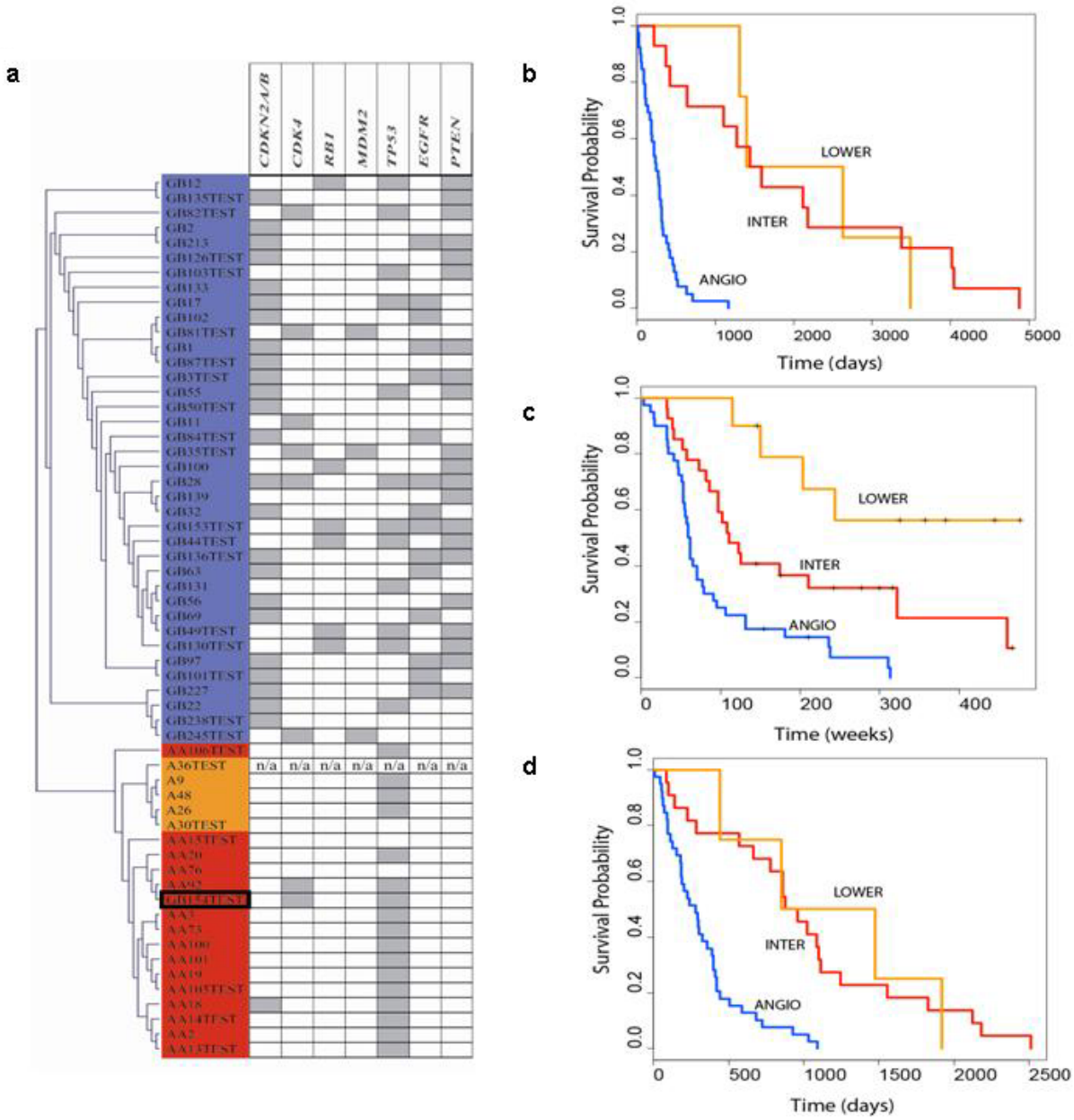
Σχήμα 2.4 Η έκφραση των ADM και PEA15 στα διαφορετικά στάδια των όγκων των αστροκυττάρων σε επίπεδο μεταγραφής και σε επίπεδο πρωτεΐνης. **(α)** GeneChip ('chip') τιμές εκφράσεις για *ADM* and *PEA15* στα διαφορετικά στάδια των όγκων. Τα δείγματα PA68 and PA67 (όγκοι βαθμίδας I - grade I tumours – see text) δεν έχουν συμπεριληφθεί σε αυτήν την ανάλυση. Οι μεταβολές έκφρασης και για τα δύο γονιδιακά προϊόντα είναι στατιστικώς σημαντικές (για τις συγκρίσεις AA-GB and A-GB αντίστοιχα: *ADM*, p = 1.1e^{-6} and 1.6e^{-4}; *PEA15*, p = 4.8e^{-5} and 8.3e^{-5}). **(β)** Επιβεβαίωση των μεταβολών έκφρασης με τη χρήση QPCR. Η μέση έκφραση των μεταβολών μεταξύ των GB και AA βαθμίδων ('Expression fold change') παρουσιάζεται όπως καθορίστηκε κα με GeneChip ('chip') και με QPCR expression technology. **(γ)** Εντάσεις ανοσοαντίδρασης (IHC immunoreactivity intensities) για ADM and PEA15 στα διαφορετικά στάδια των όγκων και φυσιολογικών εγκεφαλικών ιστών (non-neoplastic 'NT', normal tissue). Χρησιμοποιήθηκε ένα σύστημα βαθμολόγησης με 5 βαθμίδες: '0' χωρίς ανοσοαντίδραση, '4' έντονη ανοσοαντίδραση. Για κάθε καρκινικό στάδιο και για το σύνολο φυσιολογικών ιστών, καθορίστηκε ένας μέσος βαθμός ανοσοαντίδρασης από replicate tissue cores που είναι διαθέσιμοι στην tissue array. Οι διαφορές στην ανοσοαντίδραση IHC (Mann-Whitney μη-παραμετρική δοκιμή, p<5.0e^{-2}) ήταν σημαντικές για τις συγκρίσεις των ομάδων των όγκων A-GB and A-AA (PEA15), και A-AA (ADM). **(δ)** Αντιπροσωπευτικά αποτελέσματα του IHC για τις ADM και PEA15 σε τομές όγκων GB και A. Η ανοσοαντίδραση και για τα δύο γονιδιακά προϊόντα ήταν εμφανής μόνο σε κύτταρα όγκων (καρκινικά κύτταρα). Η ADM έβαψε το κυτταρόπλασμα και η PEA15 έβαψε και το κυτταρόπλασμα και τον πυρήνα. Η πυρηνική/κυτταροπλασματική κατανομή της ανοσοαντίδρασης της PEA15 δεν ήταν σταθερή για όλα τα καρκινικά κύτταρα σε ένα δεδομένο δείγμα όγκου.

2.3.2 Διαβάθμιση (Grading) με την χρήση Εκπαιδευμένων Τεχνητών Νευρωνικών Δικτύων (ΤΝΔ)

Η διαβάθμιση δειγμάτων ελέγχου με τη χρήση εκπαιδευμένων μοντέλων ΤΝΔ σε υπότυπους όγκων (tumour subtypes) συμφωνεί με προηγούμενη ιστοπαθολογική διαβάθμιση.

Η διαβάθμιση των δειγμάτων δοκιμής (γενίκευσης) (test set) (n=26) πραγματοποιήθηκε περνώντας το κάθε δείγμα από όλα τα μοντέλα που αποθηκεύτηκαν κατά τη διάρκεια της διαδικασίας εκπαίδευσης (για λεπτομέρειες

βλέπε Υλικά και Μέθοδοι). Με αυτό τον τρόπο τα 59 γονίδια που επιλέχθηκαν κατά τη διάρκεια της εκπαίδευσης μπορούν να χρησιμοποιηθούν για την διαβάθμισης ενός «τυφλού» (δεν έχουν χρησιμοποιηθεί κατά τη διαδικασίας εκπαίδευσης) συνόλου δειγμάτων (όγκων). Για κάθε δείγμα ελέγχου, έγινε ένα αρχικό ψήφισμα (voting) μέσω των *ANGIO/DIFFER* εκπαιδευμένων μοντέλων. Στα δείγματα που χαρακτηρίστηκαν ως DIFFER πραγματοποιήθηκε μία επακολουθούμενη (follow-up) διαβάθμιση μέσω των *INTER/LOWER* εκπαιδευμένων μοντέλων για να γίνει διάκριση μεταξύ των υπότυπων INTER και LOWER (βλέπε Πίνακα Α στο παράρτημα). Η ιστοπαθολογική διαβάθμιση των δειγμάτων ελέγχου φάνηκε να συμφωνεί με τους υπότυπους όγκων που παρατηρήθηκαν κατά τη διάρκεια της εκπαίδευσης. Πιο συγκεκριμένα, όλοι οι GB (εκτός του δείγματος GB154) και χαμηλότερης βαθμίδας αστροκυττωματικοί όγκοι (A and AA) έδειξαν αυξημένη έκφραση των γονιδίων ANGIO και DIFFER αντίστοιχα. Επιπλέον, όλα τα δείγματα Α διακρίθηκαν από τα ΑΑ μέσω της διαφορετικής έκφρασης των γονιδίων INTER/LOWER. Συνολικά, η διαβάθμιση των όγκων από το ΤΝΔ σε διαφορετικούς υπότυπους ήταν σε συμφωνία με προηγούμενη ιστοπαθολογική διαβάθμιση κατά: 94.74%, 100% και 100% ακρίβεια για τους ‘GB’, ‘AA’ και ‘A’ όγκους αντίστοιχα. Η Απεικόνιση των εξόδων του δικτύου χρησιμοποιώντας διαθέσιμους αλγόριθμους ομαδοποίησης (110,111) για όλα τα δείγματα εκπαίδευσης και δοκιμής (γενίκευσης) συμπεριλαμβανομένου γονιδιωματικών μετα-δεδομένων) φαίνεται στο Σχήμα 2.5a (βλέπε επίσης Υλικά και Μέθοδοι).

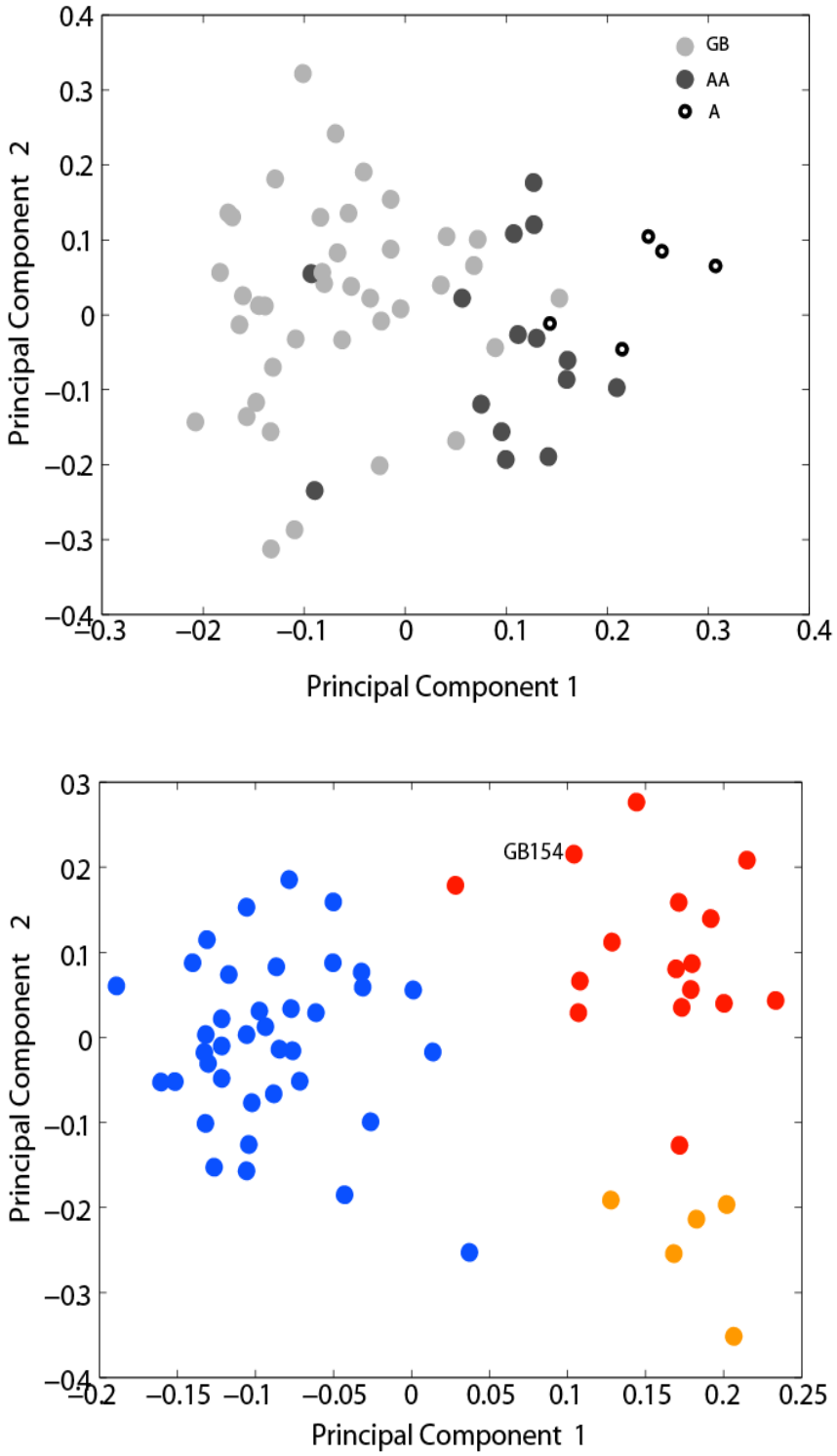


Σχήμα 2.5 (α) Απεικόνιση των αποτελεσμάτων του δικτύου για τα 33 δείγματα εκπαίδευσης και τα 26 δείγματα δοκιμής (γενίκευσης) (39GB, 15 AA and 5 A) με τη χρήση 59 γονιδίων που επιλέχτηκαν κατά τη διάρκεια της εκπαίδευσης. Ιεραρχική ομαδοποίηση των εξόδων του δικτύου (ευκλείδεια απόσταση, αλγόριθμος single linkage). Η απεικόνιση αντιπροσωπεύει αποτελέσματα από την προσπέλαση όλων των δειγμάτων (εκπαίδευσης και ελέγχου) διαμέσου των εκπαιδευμένων μοντέλων. Η χρωματική αντιστοίχηση βασίζεται σε 3 μοριακούς υπότυπους οι οποίοι

χαρακτηρίζουν καλύτερα τα δείγματα (μπλε: *ANGIO*, κόκκινο: *INTER*, πορτοκαλί: *LOWER*). Ο μόνος όγκος όπου η διαβάθμιση με ΤΝΔ δεν συμφωνεί με την ιστοπαθολογία είναι το GB154 (τονισμένο στο μαύρο κουτί). Το AA106 φαίνεται να ομαδοποιείται ξεχωριστά από τα υπόλοιπα των AA αλλά δεν ανήκει στην ομάδα των LOWER. Για τους όγκους διατίθεται γενωμική πληροφορία για ένα σύνολο από 7 γενετικούς τόπους (loci) σε προηγούμενες δημοσιεύσεις (112-116). Αυτοί οι τόποι είναι γνωστό ότι παίζουν ρολό σε αστροκυττωματική ογκογένεση και/η εξέλιξη της νόσου. Τα γκρι κουτιά δείχνουν ομοζυγωτική εξάλειψη (*CDKN2A/CDKN2B/p14*^{ARF}), ενίσχυση (CDK4, MDM2 και EGFR) ή την απώλεια ενός αλληλόμορφου γονιδίου με τη μεταλλαγή του απομείναντος αλληλόμορφου γονιδίου (RB1, TP53 και PTEN). Τα δείγματα δοκιμής (γενίκευσης) χαρακτηρίζονται με την ετικέτα "TEST". **(b)** Kaplan-Meier διάγραμμα επιβίωσης από τους 59 αστροκυττωματικούς όγκους μας όπως καθορίζονται από τη διαβάθμιση με ΤΝΔ. *ANGIO* - μπλε γραμμή, *INTER* - κόκκινη γραμμή, *LOWER* – πορτοκαλί γραμμή. **(c)** Kaplan-Meier διάγραμμα επιβίωσης από τα 76 δείγματα από τους *Phillips et al* όπως καθορίζονται από τη διαβάθμιση με ΤΝΔ χρησιμοποιώντας τα 59 γονίδια ταξινομητές. *ANGIO* - μπλε γραμμή, *INTER* - κόκκινη γραμμή, *LOWER* – πορτοκαλί γραμμή. **(d)** Kaplan-Meier διάγραμμα επιβίωσης από τα 64 δείγματα από τους *Freije et al.* Ο υπότυπος ANGIO περιείχε 38/50 GB, ο *INTER* αποτελείτε από 6/15 AA και 12/50 GB, ενώ τα υπόλοιπα 4/15 AA αποτέλεσαν τον υπότυπο *LOWER.*

Principal Component Ανάλυσης (PCA)

Με σκοπό να απεικονίσουμε την διακριτική δύναμη των γονίδιων ταξινομητών κάναμε PCA ανάλυση στα 33 δείγματα εκπαίδευσης και στα 26 δείγματα δοκιμής (γενίκευσης) χρησιμοποιώντας όλα τα γονίδια που υπήρχαν στο chip (Σχήμα 2.6α) και τα 59 γονίδια ταξινομητές (Σχήμα 2.6b). Τα αποτελέσματα συμπίπτουν με τα αποτελέσματα από τα ΤΝΔ (Σχήμα 2.6α) σύμφωνα με τα οποία μόνο ένα δείγμα GB (GB154) βρίσκεται σε ομάδα μαζί με τα υπόλοιπα δείγματα INTER.



Σχήμα 2.6 Principal Component analysis πριν και μετά την επιλογή γονιδίων.

Η ανάλυση των πρώτων δύο principal components για τα 33 δείγματα εκπαίδευσης και τα 26 δείγματα δοκιμής (γενίκευσης): **(α)** Χρησιμοποιώντας τις τιμές έκφρασης από 22,382 probes. Η χρωματική αντιστοίχηση για τα δείγματα ιστού είναι (γκρι, GB;

μαύρο, ΑΑ; άσπρο, Α), **(β)** Χρησιμοποιώντας μονό τα 59 γονίδια ταξινομητές που επιδέχθηκαν κατά την διαδικασία της εκπαίδευσες. Η χρωματική κωδικοποίηση βασίζεται στους 3 μοριακούς υπότυπους που χαρακτηρίζουν καλύτερα τα δείγματα (μπλε: ANGIO, κόκκινο: INTER, πορτοκαλί: LOWER).

Διαβάθμιση ενός ανεξάρτητου συνόλου από αστροκυττωματικούς όγκους με τη χρήση κοινών γονιδίων-ταξινομητών

Για να επιβεβαιώσουμε περεταίρω την ικανότητα διαβάθμισης των γονίδιων ταξινομητών, χρησιμοποιήσαμε ένα ανεξάρτητο σύνολο δεδομένων γονιδιακής έκφρασης από αστροκυττωματικούς όγκους που δημοσιεύτηκε από τους Shai *et al* το 2003 (28). Από τα 59 γονίδια ταξινομητές που επιλέχτηκαν κατά τη διάρκεια της εκπαίδευσης από τα HG-U133A genechips, τα 38 γονίδια είχαν >96% ταυτοποίηση με γονίδια στο U95Av2 genechip που χρησιμοποιήθηκε από τον Shai *et al*, 2003 (28). Απο αυτά, επιλέξαμε τα 20 γονίδια που εμφανίστηκαν περισσότερες από μία φορά κατά τη διαδικασία leave-one-out cross-validation, επιβεβαιώνοντας έτσι ότι μόνο τα πιο στατιστικά σημαντικά γονίδια χρησιμοποιήθηκαν στην ανάλυση (cross-chip analysis). Από αυτά, τα 17 γονίδια εκφράζονταν διαφορικά στην σύγκριση GB-AA (που αποτελούνταν από τα γονίδια *ANGIO* και *DIFFER),* 1 στην σύγκριση GB-A (*INTER/LOWER* γονίδια) και 2 στην σύγκριση AA-A (*INTER/LOWER* γονίδια). Εκπαιδεύσαμε ξανά τα ΤΝΔ με τα αρχικά μας δεδομένα εκπαίδευσης αλλά, χρησιμοποιώντας μόνο τα συγκεκριμένα 20 γονίδια (για ονόματα γονιδίων και περιγραφή, βλέπε παράτημα). Εξαιτίας του περιορισμένου αριθμού του συνόλου των γονιδίων που ήταν διαθέσιμα για την εκτίμηση του GB-A, χωρίσαμε την διαβάθμιση σε δύο συγκρίσεις ανα ζεύγη (pair-wise). Τα ΤΝΔ «τύπου 1» εκπαιδεύτηκαν για την διάκριση μεταξύ των αστροκυττωματικών όγκων βαθμίδας IV και χαμηλότερής βαθμίδας με τη χρήση των 17 γονιδίων *ANGIO/DIFFER* και τα μοντέλα «τύπου 2» εκπαιδεύτηκαν για τη διάκριση μεταξύ όγκων βαθμίδας II και III χρησιμοποιώντας τα 3 γονίδια *INTER/LOWER*. Μονό στα δείγματα που χαρακτηρίστηκαν ως όγκοι *DIFFER* χαμηλότερής βαθμίδας από τα μοντέλα τύπου 1 πραγματοποιήθηκε μία επακολουθούμενη (follow-up) διαβάθμιση μέσω των εκπαιδευμένων μοντέλων τύπου 2. Τα 23 (18 GB, 3 AA, 2 A) δείγματα που προέκυψαν από το σύνολο δεδομένων των Shai *et al*, 2003 χρησιμοποιήθηκαν σαν δείγματα δοκιμής (γενίκευσης) και ταξινομήθηκαν χρησιμοποιώντας τα εκπαιδευμένα μοντέλα. Παρατηρήθηκε αξιοσημείωτη συνέπεια μεταξύ των δύο συνόλων δεδομένων έκφρασης με τη χρήση

των 20 κοινών γονιδίων, όπου η ιστοπαθολογική διαβάθμιση και η διαβάθμιση βασιζόμενη στα ΤΝΔ συμπίπτουν σε ποσοστό 100% (2/2), 100% (3/3) και 88.89% (16/18) για τους Α (*LOWER*), ΑΑ (*INTER*) και GB (*ANGIO*) όγκους της μελέτης των Shai *et al* αντίστοιχα (βλέπε Πίνακα Γ στο παράρτημα).

Διαβάθμιση επιπρόσθετων δειγμάτων που ήταν δύσκολο να διαγνωστούν ιστοπαθολογικά και εκτίμηση των αποτελεσμάτων των ΤΝΔ χρησιμοποιώντας κλινικά, ιστοπαθολογικά και γονιδιωματικά δεδομένα

Μετά την επιβεβαίωση της ικανότητας διαβάθμισης των μοριακών υπότυπων, τους χρησιμοποιήσαμε για να ταυτοποιήσουμε τη βαθμίδα των δειγμάτων μας που ήταν ιδιαιτέρως δύσκολο να διαγνωστούν ιστοπαθολογικά. Η ιστοπαθολογική ταυτοποίηση των Pilocytic αστροκυττωμάτων (grade I) και η προσπάθεια διαβάθμισης αστροκυττωματικών όγκων οι οποίοι είχαν υποβληθεί σε θεραπεία με ακτινοβολία και/ή χημειοθεραπεία, μπορεί να έχει αμφίβολα αποτλέσματα. Επομένως εξετάσαμε τα δεδομένα έκφρασης 6 τέτοιων προβληματικών περιπτώσεων χρησιμοποιώντας τα εκπαιδευμένα ΤΝΔ.

Οι δύο pilocytic αστροκυττωματικοί (PA) όγκοι (PA68 and PA67) διαβαθμίστηκαν σαν *ANGIO* (GB-rich) and *INTER* (AA-rich) αντίστοιχα από τα εκπαιδευμένα ΤΝΔ (βλέπε Συζήτηση). Αυτοί οι όγκοι ήταν ιστολογικά τυπικοί (117) και προέκυψαν από ασθενείς με εξαιρετική επιβίωση (ζωντανοί στο τέλος της παρακολούθησης - βλέπε Πίνακα Ι στο παράρτημα). Τα δείγματα AA49 και AA86, είχαν υποβληθεί σε ακτινοβολία και χημειοθεραπεία και ήταν δύσκολο να διαβαθμιστούν. Δύο άλλοι ΑΑ όγκοι, ο AA29 και ο AA93, ήταν επίσης δύσκολο να διαβαθμιστούν ιστολογικά. Η διαβάθμιση αυτών των δειγμάτων με τη χρήση των εκπαιδευμένων ΤΝΔ δεν συμφωνούσε με την ιστοπαθολογική διαβάθμιση και κατέταξε και τα 4 ΑΑ δείγματα ως *ANGIO* (GB-rich subtype) (βλέπε Πίνακα Β στο παράρτημα).

Προκειμένου να διερευνηθούν πιθανές αιτίες αυτής της ασυμφωνίας, αποτιμήσαμε διαθέσιμες πληροφορίες και για τους 4 διφορούμενους όγκους του συνόλου των δεδομένων μας, όπως επίσης για το GB154 και τα δύο τύπου Ι pilocytic αστροκυττώματα. Μαζί με την ιστοπαθολογική διάγνωση οι διαθέσιμες πληροφορίες συμπεριελάμβαναν i) κλινικά δεδομένα (ηλικία χειρουργικής επέμβασης, φύλο, πρωτογενής ή δευτερογενής όγκος, θέση του όγκου), ii) δεδομένα επιβίωσης και iii) προδημοσιευμένες γονιδιακές πληροφορίες για ένα σύνολο 9 γονιδίων (*CDKN2A,*

CDKN2B, *p14^{ARF}*, *CDK4*, *RB1*, *MDM2*, *EGFR*, *PTEN* and *TP53)* που ήταν γνωστό ότι επηρεάζονται σε αστροκυττώματα (βλέπε Πίνακα Ι στο παράρτημα).

Η κατάταξη βαθμίδας κακοήθειας των όγκων AA49 και AA86 βασισμένη σε ιστολογικές εξετάσεις (histology), ήταν αμφίβολής αποτελεσματικότητας καθώς οι προηγούμενες θεραπείες περιπλέκουν σε μεγάλο βαθμο τα ευρήματα. Το AA49 παρουσιάζει ένα καθαρό GB γενετικό προφίλ (ομοζυγωτική εξάλειψη των CDKN2A, CDKN2B και p14^{ARF}, EGFR ενίσχυση και καθόλου wild-type PTEN), ενώ το AA86 παρουσιάζει επιπλέον γενετικές ανωμαλίες που συνήθως εμφανίζονται σε γλοιοβλαστώματα: έλλειψη wild-type CDKN2A, p14^{ARF}, ή TP53. Στην περίπτωση του όγκου AA29 κλινικές, ιστοπαθολογικές και γενετικές ενδείξεις παρουσιάζουν σημαντική ομοιότητα με GB (υποψία αλλά όχι καθαρές ενδείξεις νέκρωσης, καθώς και wild-type *PTEN*). Ο όγκος AA93 έχει την ιστολογική και κλινική εμφάνιση AA, αλλά μοιράζεται το ίδιο κλασσικό GB γενετικό προφίλ που παρουσιάζει και ο AA49. Η μόνη γενετική διαφορά ανάμεσα στους δύο όγκους σχετίζεται με διατήρηση ενός wild-type αντίγραφου του *PTEN*.

Οι 4 διφορούμενοι AA όγκοι, που ταξινομούνται από το τεχνητό νευρωνικό μας δίκτυο ως *ANGIO,* περιλαμβάνουν 100% (2/2) από EGFR ενίσχυση, 100% (2/2) PTEN μεταλλάξεις και 66% (2/3) από CDKNA/B nullizygosity, που συναντά κανείς σε όλα τα AA δείγματα που αξιολογήθηκαν στα δεδομένα μας. Αν εξαιρέσουμε το ένα *INTER* δείγμα με διαβάθμιση AA το όποιο φέρει CDKN2A/B nullizygosity, μεταλλάξεις στη γενετική περιοχή του αναστολέα της κυκλίνης (cyclin inhibitor) απουσίαζαν εντελώς σε όλα τα εναπομείναντα AA δείγματα των δεδομένων μας. Και στις 3 αυτές περιπτώσεις των AA όγκων όπου υπήρχαν διαθέσιμα δεδομένα επιβίωσης, οι ασθενείς απεβίωσαν μέσα σε 2 χρόνια.

Καμία προφανής εξήγηση δεν μπορεί να βρεθεί για την διαφωνία ανάμεσα στην ιστοπαθολογική και στην διαβάθμιση με ΤΝΔ του GB όγκου (GB154). Παρόλο που το GB154 παρουσίαζε μερικά μη κλασσικά GB χαρακτηριστικά, η παρουσία ενίσχυσης του *CDK4,* η νέκρωση και ο μικρο-αγγειακός πολλαπλασιασμός (microvascular proliferation), με το τελευταίο να είναι το σημαντικότερο ιστολογικό κριτήριο για την αξιολόγηση του ως γλοιοβλάστωμα, υποστηρίζουν την αρχική ιστοπαθολογική διάγνωση. Η επιβίωση σε αυτήν την περίπτωση ήταν επίσης κάτω από 2 χρόνια.

2.3.3 Ανάλυση Επιβίωσης

Η ανάλυση επιβίωσης χρησιμοποιώντας τα επιλεγμένα γονίδια ταξινομητές αποφέρει μια προγνωστική αξία για τους υπότυπους των όγκων.

Για να εξερευνήσουμε τις προγνωστικές ικανότητες για επιβίωση των γονιδίων ταξινομητών μας, εκτελούμε μια ανάλυση επιβίωσης στα 59 δείγματα όπως διαβαθμίστηκαν από την ιστοπαθολογική εξέταση και κατόπιν, όπως ορίστηκαν από τα εκπαιδευμένα ΤΝΔ, στους τρεις υπότυπους όγκων. Παρόλο που υπήρχε μια πολύ μικρή διαφορά ανάμεσα στην διαβάθμιση με ΤΝΔ και στην ιστοπαθολογική διαβάθμιση (η διαφορά στο δείγμα – GB154), η ανάλυση επιβίωσης με βάση την διαβάθμιση των ΤΝΔ αποδεικνύεται στατιστικά σημαντικότερη (p= 8.76e^{-7}) από την διαβάθμιση βασιζόμενη αποκλειστικά σε ιστοπαθολογικά δεδομένα (p = 2.088e^{-6}), όπως αυτό ορίζεται από το log rank test (Σχήμα 2.5b). Παρόμοια αποτελέσματα αποκτήθηκαν και από την ανάλυση επιβίωσης στα δεδομένα του Shai *et al* in 2003 (28) . Η προγνωστική αξία των υπότυπων όπως αυτοί ορίζονται από τα μοντέλα των ΤΝΔ μας είναι εξίσου στατιστικά σημαντική (p =6.0e^{-3}) όσο και αυτή που ορίζεται με βάση τα ιστοπαθολογικά χαρακτηριστικά (p =6.0e^{-3}).

Η ανάλυση επιβίωσης τεκμηριώνει την διαβάθμιση των δεδομένων με βάση τα ΤΝΔ όταν αυτή δεν συμπίπτει με την προηγούμενη ιστοπαθολογική διαβάθμιση.

Προς έκπληξή μας, για δύο ανεξάρτητα σύνολα δεδομένων, τα ΤΝΔ κατηγοριοποίησαν τα δείγματα με διαφορετικό τρόπο απ' ότι τα ιστοπαθολογικά κριτήρια. Παρ' ολα αυτά, και στις δύο αυτές περιπτώσεις, η ανάλυση επιβίωσης τεκμηρίωσε την διαβάθμιση με βάση τα ΤΝΔ. Η ανάλυση επιβίωσης για τα συγκεκριμένα σύνολα δειγμάτων περιγράφεται παρακάτω.

Τα δεδομένα των Phillips et al. 2006 (93) περιλαμβάνουν 100 MDA δείγματα (76 εκ των οποίων υπήρχε διαθέσιμη πληροφορία όσο αφορά την επιβίωση). Σε αυτή την μελέτη τα δείγματα είχαν χωριστεί σε 3 "υποκατηγορίες" αναπαριστώντας τα εξελικτικά στάδια των αστροκυττωματικών όγκων. Οι υποκατηγορίες ορίστηκαν από τους συγγραφείς ως Proneural (*PN*), proliferative (*Prolif*) and Mesenchymal (*Mes*), με αυξανόμενη κακοήθεια από την *PN* στην *Mes.* Εφόσον αυτά τα δείγματα περιλάμβαναν όγκους των σταδίων III και IV, χρησιμοποιήσαμε τα ΤΝΔ που εκπαιδεύσαμε με *ANGIO/DIFFER* γονίδια για να ταξινομήσουμε τα 100 MDA δείγματα στους αντίστοιχους υπότυπους. Ο υπότυπος *ANGIO* περιλαμβάνει 50/76 GB

και 4/24 AA, ενώ ο υπότυπος *DIFFER* 22/76 GB και 12/24 AA. Η ομάδα *ANGIO* αποτελείται από 30/35 Mes δείγματα από τα δεδομένα των Phillips *et al 2006* (93) και σύμφωνα με προηγούμενη μελέτη, έχει δειχθεί ότι οι Mes όγκοι παρουσιάζουν υπερ-έκφραση σε δείκτες αγγειο-γένεσης (93). Η ομάδα *DIFFER* αποτελείται από 33/37 από *PN* δείγματα, όπου πάλι σύμφωνα με προηγούμενες αναφορές υπάρχουν ενδείξεις ότι τα δείγματα PN παρουσιάζουν υπερ-έκφραση σε δείκτες νευρωνικής διαφοροποίησης και ανάπτυξης. Επιπλέον ανάλυση των δειγμάτων *DIFFER* έγινε χρησιμοποιώντας τα δικά μας *INTER/LOWER* γονίδια για των διαχωρισμό του υπότυπου *INTER*, που αποτελείται από 8/76 GB και 4/76 AA, και του υπότυπου *LOWER,* που αποτελείται από 4/76 GB και 8/24 AA. Η προσέγγιση αυτή ομαδοποιεί τα δείγματα του Phillips *et al 2006* (93) σε τρεις πολύ σημαντικές προγνωστικές υποκατηγορίες (Σχήμα 2.5c, $p = 1.922e^{-7}$), που για μια ακόμη φορά ξεπερνάει σε απόδοση την προηγούμενη υποκατηγοριοποίηση όπως αυτή έχει οριστεί από την μελέτη των Phillips *et al* 2006 (93) (p = 1.0e-4). Τα Prolif δείγματα των Phillips *et al 2006* (93) , που σύμφωνα με την μελέτη τους αναπαριστούν ενδιάμεσα στάδια της εξέλιξης της νόσου και είναι πολύ πλούσια σε δείκτες πολλαπλασιασμού, δεν είναι τόσο καλά ορισμένα στην δική μας υποκατηγοριοποίηση όγκων. Παρόλα αυτά, *20/28* βρέθηκαν μέσα στον υπότυπο ANGIO όπως έχει οριστεί από τα ΤΝΔ, (το οποίο είναι πλούσιο σε Mes δείγματα) και 8/28 στον υπότυπο INTER (το οποίο είναι πλούσιο σε PN δείγματα)(βλέπε Πίνακα Ε στο παράρτημα).

Τα αποτελέσματα αυτά συμφωνούν με τα προηγούμενα δημοσιευμένα αποτελέσματα, τα οποία δείχνουν κατά μέσο όρο παρόμοιο χρονικό διάστημα επιβίωσης για τις ομάδες του Mes και Prolif των Phillips et al 2006 (93) και μια υψηλή ένδειξη αγγειογένεσης για την ομάδα *Prolif* σε σχέση με τους όγκους *PN* (93). Επιπλέον, υπάρχει συμφωνία με την παρατήρηση ότι το μοριακό αποτύπωμα της ομάδας *Prolif* είναι λιγότερο αποκλειστικό/καθορισμένο και το ποσοστό των αστροκυττωματικών όγκων με αυτό το αποτύπωμα ποικίλει ανάμεσα στα δείγματα που λήφθηκαν από διαφορετικά ερευνητικά ιδρύματα (93).

Τέλος, μια σύγκριση μεταξύ των 59 γονιδίων ταξινομητών που χρησιμοποιήθηκαν σε αυτήν την ανάλυση και τα τελικά 35 γονίδια που προσδιορίζονται από τους Phillips et al, 2006, έδειξε ότι δεν υπήρχαν κοινά στοιχεία μεταξύ των δύο συνόλων γονιδίων, τονίζοντας για άλλη μια φορά την καινοτομία στην ταυτοποίηση των γονίδιων ταξινομητών μας.

Παρόμοια αποτελέσματα έδωσε ένα άλλο ανεξάρτητο σύνολο δεδομένων που περιέχει 65 αστροκυττωματικούς όγκους (15 βαθμίδας III και 50 βαθμίδας IV) που έχουν δημοσιεύσει οι Frieje et al (88). Πιο συγκεκριμένα, η υποκατηγοριοποίηση μας προσδιορίζει ομάδες με στατιστικά σημαντικότερη επιβίωση (Σχήμα 2.5d, p = 8.13e-8) σε σχέση με αυτή που αποδίδουν οι Freije et al (88) στην αντίστοιχη δημοσίευση (p = 2.2e-4).

Γονίδια που προβλέπουν το χρόνο επιβίωσης και γονίδια που προβλέπουν την ιστοπαθολογική βαθμονόμηση.

Για να διερευνηθεί αυτή την απροσδόκητη απόδοση στα δεδομένα των Phillips et al. και Freije et al., όπου γονίδια προσδιορισμένα με βάση ιστοπαθολογικά κριτήρια λειτουργούσαν ως προγνωστικά αποτυπώματα επιβίωσης, αποφασίσαμε να συγκρίνουμε γονίδια πρόβλεψης του χρόνου επιβίωσης (γονίδια συσχετιζόμενα με επιβίωση) και γονίδια πρόβλεψης της ιστοπαθολογικής βαθμονόμησης (γονίδια βασιζόμενα στην ιστοπαθολογική ταυτοποίηση του όγκου) τόσο μεταξύ του δικού μας συνόλου δεδομένων, όσο και μεταξύ άλλων δύο μεγάλων συνόλων δεδομένων. Αρχικά πραγματοποιήσαμε ομαδοποίηση χρησιμοποιώντας τα πρώτα 80 γονίδια τα οποία βρέθηκαν να είναι θετικά συσχετιζόμενα και αρνητικά συσχετιζόμενα με τον χρόνο επιβίωσης (Pearsons συσχέτισης των τιμών έκφρασης έναντι επιβίωση > 0,55 ή <-0,55), όπου και παρατηρήθηκαν τρεις μεγάλες ομάδες. Στη συνέχεια, βαθμονομήσαμε εκ νέου τα ΤΝΔ μας έτσι ώστε να αποδίδουν βέλτιστα αποτελέσματα χρησιμοποιώντας την τεχνική του leave-one-out cross-validation για τις ομάδες με γονίδια συσχετιζόμενα με την επιβίωση και καταλήξαμε σε ένα βέλτιστο σύνολο 37 γονιδίων (βλέπε Πίνακα G στο παράρτημα). Επίσης, πραγματοποιήσαμε την ίδια ανάλυση στα δύο ανεξάρτητα σύνολα δεδομένων, όπως περιγράφεται παραπάνω για το δικό μας σύνολο δεδομένων και επιλέχθηκαν τα αντίστοιχα γονίδια βασιζόμενα στην ιστοπαθολογική βαθμονόμηση.

Οι Log rank αξιολογήσεις χρησιμοποιώντας ιστοπαθολογικά βασιζόμενα ή συσχετιζόμενα με την επιβίωση γονίδια φαίνονται στον πίνακα 2.2 Παραδόξως, βρήκαμε ότι τα γονίδια με διαφοροποιημένη έκφραση ανάμεσα σε ιστοπαθολογικές βαθμίδες, όπως εντοπίζονται βάσει της μεθόδου signal-to-noise (σήματος προς θόρυβο), προέβλεπαν καλύτερα το χρόνο επιβίωσης από ότι τα αντίστοιχα γονίδια που συσχετίζονται με το χρόνο επιβίωσης σε όλα τα σύνολα δεδομένων που αναλύσαμε.

Πίνακας 2.2 Σύγκριση των ιστοπαθολογικών γονιδίων και τα γονίδια που έχουν προβλεφθεί από την συσχέτιση με την επιβίωση. *P*-τιμές έχουν υπολογιστεί/προβλεφθεί χρησιμοποιώντας το log-rank τεστ μετά την ομαδοποίηση των δειγμάτων σε ομάδες επιβίωσης.

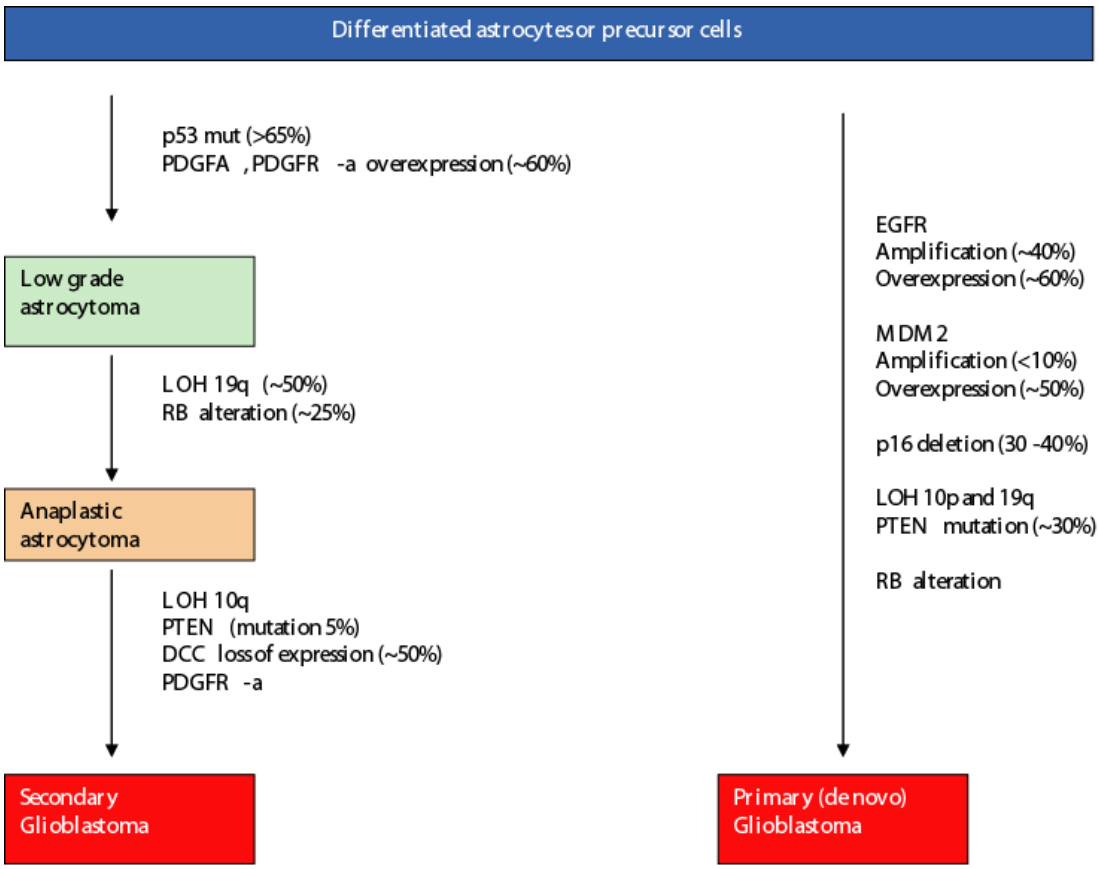
Dataset	Genes Used		
	Histopathology genes		
	Our 59 genes	Phillips genes[c]	Freijer genes[d]
Our	$p = 8.76e^{-7}$	$p = 4.258e^{-5}$	$p = 6.871e^{-6}$
Phillips	$p = 1.922e^{-7}$	$p = 8.808e^{-8}$	$p = 1.133e^{-4}$
Freijer	$p = 8.13e^{-8}$	$p = 4.55e^{-8}$	$p = 2.318e^{-8}$
	Survival correlated genes		
	Our 37 genes	Phillips 35 genes[e]	Freijer 44 genes[f]
Our	$p = 4.377e^{-6}$	$p = 1.871e^{-4}$	$p = 4.431e^{-5}$
Phillips	$p = 2.081e^{-5}$	$p = 1.0e^{-4}$	$p = 3.169e^{-5}$
Freijer	$p = 1.103e^{-5}$	$p = 1.191e^{-1}$	$p = 2.2e^{-4}$

Οι TP53 μεταλλάξεις διαχωρίζουν την βαθμίδα IV GB σε δύο ομάδες επιβίωσης

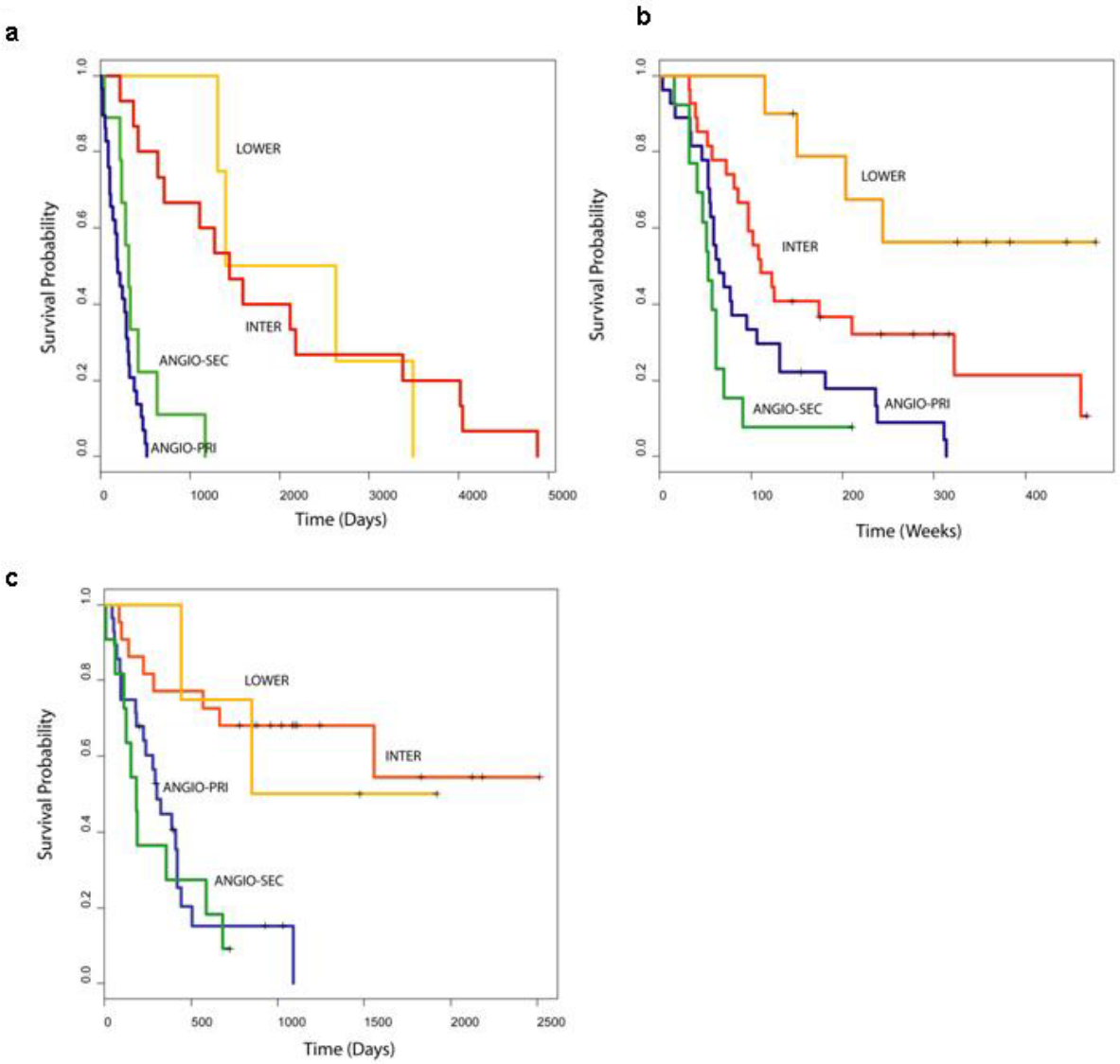
Οι TP53 μεταλλάξεις παρατηρούνται σε πάνω από 65% της δευτερογενούς GB και θεωρούνται ένα σημαντικό χαρακτηριστικό που ορίζει τα ξεχωριστά μοριακά μονοπάτια, που είναι υπεύθυνα για την ανάπτυξη του δευτερογενούς και του πρωτογενούς (de novo) GB (Σχήμα 2.7). Για να εντοπιστούν γονίδια με διακεκριμένες υπογραφές για τις δύο αυτές ξεχωριστές πορείες, εκτελούμε ένα leave-one-out cross-validation (πολλαπλής επικύρωσης) χρησιμοποιώντας μόνο τα GB, διαχωρισμένα σε αυτά που φέρουν TP53 μετάλλαξη και wild-type και προσδιορίζεται ένα βέλτιστο σύνολο από 11 probes (βλέπε Πίνακα F στο παράρτημα). Χρησιμοποιώντας αυτά τα γονίδια, τα ΤΝΔ διαχωρίζουν τον ANGIO υπότυπο μας σε δύο ομάδες, την *ANGIO-PRI* και την *ANGIO-SEC*. Αυτή η διάκριση ήταν πιο σημαντική για την πρόβλεψης επιβίωσης (p = 3.325e-2), από τον αντίστοιχο

[c]These *p*-values are the result of the grouping of the samples into 2 groups
[d]These *p*-values are the result of the grouping of the samples into 2 groups
[e]These *p*-values are the result of the grouping of the samples into 3 groups (23/35 genes were present on our gene chip as we only used the HGU133A gene chip)
[f]These *p*-values are the result of the grouping of the samples into 3 groups (41/44 genes were present on our gene chip as we only used the HGU133A gene chip)

διαχωρισμό με βάση το TP53 (p = 7.082e-1). Στο σύνολο δεδομένων του Phillips, διαπιστώσαμε ότι η ομάδα *ANGIO-SEC* αποτελείτω από 16/28 *Prolif* δείγματα και μόλις 3 / 35 *Mes* δείγματα, ενώ η ομάδα *ANGIO-PRI* αποτελείτω από 12/28 *Prolif* και 32/35 *Mes* δείγματα. Το αποτέλεσμα αυτό συμφωνεί με προηγούμενες αναφορές (118) που δείχνουν ότι τα δευτερογενή GB υφίστανται επιθετικό πολλαπλασιασμό (όπως είναι η περίπτωση των *Prolif* δειγμάτων) σε αντίθεση με τα πρωτογενή GB, που δείχνουν υπερέκφραση γονιδίων αγγειο-γένεσης (όπως συμβαίνει και με τα *Mes* δείγματα). Η ανάλυση επιβίωσης χρησιμοποιώντας τα 59 γονίδια ταξινομητές μας, όπως επίσης και το αποτύπωμα των 11 γονίδιων που περιγράφονται εδώ, για τρία σύνολα δεδομένων, φαίνεται στο Σχήμα 2.8.



Σχήμα 2.7. Λεπτομερής ορισμός των ξεχωριστών μοριακών μονοπατιών, που είναι υπεύθυνα για την ανάπτυξη του δευτερογενούς και του πρωτογενούς (de novo) γλοιοβλάστωμα.



Σχήμα 2.8 Ανάλυση επιβίωσης των αστροκυττωμάτων (συμπεριλαμβανομένου και του γονιδιακού αποτυπώματος των 11 πρωτογενών/δευτερογενών γονίδιων). **(a)** Γράφημα επιβίωσης Kaplan-Meier των 59 αστροκυττωμάτων που προσδιορίστηκαν από τα αποτελέσματα της διαβάθμισης με ΤΝΔ. *Πρωτογενή--ANGIO – μπλε γραμμή, δευτερογενή-ANGIO – πράσινη INTER –κόκκινη γραμμή, LOWER- πορτοκαλί γραμμή.* **(b)** Γράφημα επιβίωσης Kaplan-Meier των 76 δειγμάτων από τους Phillips et al, όπως προσδιορίστηκαν από τα αποτελέσματα της διαβάθμισης με ΤΝΔ, χρησιμοποιώντας τα 59 γονίδια ταξινομητές + το γονιδιακό αποτύπωμα των 11 πρωτογενών/δευτερογενών γονίδιων. **(c)** Γράφημα επιβίωσης Kaplan-Meier των 65 δειγμάτων από Freije, χρησιμοποιώντας τα 59 γονίδια ταξινομητές + το γονιδιακό αποτύπωμα των 11 πρωτογενών/δευτερογενών γονίδιων.

2.4 Συζήτηση

Στην παρούσα μελέτη χρησιμοποιήσαμε μία απλή, προσέγγιση βασισμένη στα ΤΝΔ προκειμένου να εξάγουμε συγκεκριμένα μεταγραφικά αποτυπώματα από τους διάφορους ιστοπαθολογικούς υπότυπους των αστροκυττωμάτων. Στη συνέχεια, αξιολογήσαμε αυτά τα μοριακά αποτυπώματα αναφορικά με την ικανότητά τους να οριοθετήσουν προγνωστικές υποκατηγορίες επιβίωσης. Τα αποτελέσματά μας έδειξαν ότι τα επιλεγμένα γονίδια ταξινομητές εμπίπτουν σε τρεις διαφορετικές λειτουργικές κατηγορίες, οι οποίες χαρακτηρίζουν τρεις μοριακούς υπότυπους όγκους: τους ANGIO, INTER και LOWER. Οι βασισμένες στα ΤΝΔ διαβαθμίσεις σε τρεις υπότυπους όγκων, τόσο στα δικά μας δεδομένα έκφρασης, όσο και σε ένα ανεξάρτητο σύνολο δεδομένων (28), βρέθηκαν να συμπίπτουν με ακρίβεια σε σχέση με προηγούμενες ιστοπαθολογικές διαβαθμίσεις. Ωστόσο, η ίδια διαδικασία που ακολουθήθηκε για δύο ακόμα σύνολα δεδομένων (88,93) δεν έδειξε αντίστοιχα αποτελέσματα. Προκειμένου να διερευνηθεί αυτή η ανακολουθία, πραγματοποιήσαμε μία εκτεταμένη σύγκριση των γονιδίων που σχετίζονται με το χρονικό διάστημα επιβίωσης και των γονιδίων που βασίζονται στην ιστοπαθολογική βαθμονόμηση. Βρήκαμε ότι, αναφορικά με την πρόβλεψη του χρόνου επιβίωσης (α) τα βασιζόμενα σε ιστοπαθολογικά κριτήρια γονίδια υπερτερούν των αντίστοιχων γονιδίων που σχετίζονται με το χρόνο επιβίωσης σε κάθε σύνολο δεδομένων και (β) τα δικά μας 59 ιστοπαθολογικά βασιζόμενα γονίδια υπερτερούν των γονιδίων που σχετίζονται με το χρόνο επιβίωσης σε *όλα* τα ελεγμένα σύνολα δεδομένων. Τέλος, η ανάλυση με ΤΝΔ των *TP53* μεταλλαγμένων και μη-μεταλλαγμένων δειγμάτων ταυτοποίησε ένα γενετικό αποτύπωμα το οποίο διαχωρίζει περαιτέρω τον *ANGIO* υπότυπο σε δύο ομάδες, κατοπτρίζοντας τα πρωτογενή και δευτερογενή GB.

Προγνωστικοί δείκτες για την αγγειογένεση (*VEGF*, *flt1/VEGFR1*, *kdr/VEGFR2*, *PECAM1*) και δείκτες του πολλαπλασιασμού (*PCNA* and *TOP2A*) (119-122), χρησιμοποιούνται ευρέως από τους παθολόγους για την ταξινόμηση/διαβάθμιση των αστροκυττωμάτων. Σ' αυτή την έρευνα προσδιορίσαμε μία καινούργια ομάδα γονιδίων που χαρακτηρίζει τον *ANGIO* υπότυπο και φαίνεται να ελέγχει την αγγειογένεση. Η γενική τάση του βαθμού IV, GB να αποτελεί μέρος του *ANGIO* υπότυπου είναι σε συμφωνία με αυτές τις έρευνες. Η οριοθέτηση της πλειονότητας των *Mes* δειγμάτων, όπως αυτά χαρακτηρίστηκαν από τους Phillips *et al 2006* (93), μέσα στον *ANGIO* υπότυπο επιβεβαιώνει περαιτέρω τα αποτελέσματα, καθώς τα

δείγματα αυτά υπερεκφράζουν αγγειο-γονιδιακούς δείκτες, όπως ο *VEGF*. Τα χαρακτηριστικά της ανάπτυξης και της διαφοροποίησης των χαμηλότερων βαθμών ΑΑ και Α είναι σε συμφωνία με την παρατήρηση ότι αυτοί οι όγκοι ανήκουν στη *DIFFER* ομάδα. Η γενική τάση των *PN* δειγμάτων από τους Phillips *et al 2006* (93) να ομοιάζουν με τα *DIFFER* δείγματα είναι επίσης σε συμφωνία με αναφορές που δείχνουν ότι τα *PN* δείγματα υπερεκφράζουν δείκτες της νευρογένεσης και της νευρωνικής διαφοροποίησης (93). Ο *Prolif* υπότυπος από τους Phillips *et al 2006* (93) δεν οριοθετήθηκε με βάση τους δικους μας υπότυπους όγκους, αλλά προσδιορίστηκε μερικώς από τα 11 γονίδια που χρησιμοποιήθηκαν για τη διαφοροποίηση μεταξύ των πρωτογενών και δευτερογενών GB. Το χαρακτηριστικό των *Prolif* δειγμάτων από τους Phillips et al να μην είναι καλά καθορισμένα, επιβεβαιώνει προηγούμενες παρατηρήσεις που αναφέρουν ένα λιγότερο ειδικό φαινότυπο αυτών των δειγμάτων, καθώς και μία μεγαλύτερη ποικιλομορφία των δειγμάτων αυτών που προέρχονται από διαφορετικά ιδρύματα (88,93). Ιδιαίτερου ενδιαφέροντος ήταν η ταυτοποίηση μιας ομάδας γονιδίων (συμπεριλαμβανομένου και του *PEA*), που φαίνεται να διαχωρίζει την ομάδα *DIFFER (*κατώτερος βαθμός ΙΙ, Α και βαθμός ΙΙΙ, ΑΑ) στους *INTER* (βαθμός ΙΙΙ, ΑΑ) και *LOWER* (βαθμός ΙΙ, Α) υπότυπους και προσδιορίζει περαιτέρω μία προγνωστική κατηγορία με την υψηλότερη πιθανότητα επιβίωσης (*LOWER*).

Η ανάλυση επιβίωσης υποδηλώνει ότι η ιστοπαθολογική διαβάθμιση, αν και κατηγορηματική και υπερ-απλουστευμένη, παρέχει μία γενική τάση, μέσω της οποίας μπορούν να ταυτοποιηθούν προγνωστικά για την επιβίωση γονίδια, των οποίων η προγνωστική αξία φαίνεται να είναι μεγαλύτερη από την ιστοπαθολογική διαβάθμιση *per se*. Πρόγνωση της επιβίωσης μπορεί να πραγματωθεί είτε ανεξάρτητα, είτε, όπως στην περίπτωση των δικών μας δεδομένων, σε συνδυασμό με την ιστοπαθολογική πρόγνωση. Σύγκριση των σχετιζόμενων με την επιβίωση και των ιστοπαθολογικών βασιζόμενων γονιδίων έδειξε ότι τα τελευταία ήταν πιο αποτελεσματικά στην πρόγνωση του χρόνου επιβίωσης. Αυτό παρατηρήθηκε τόσο για τα δικά μας δεδομένων, όπως επίσης και για αλλά δύο ανεξάρτητα σύνολα δεδομένων που ερευνηθήκαν. Μία πιθανή εξήγηση γι' αυτό τα μη-διαισθητικό αποτέλεσμα σχετίζεται με τη μεθοδολογία που χρησιμοποιήθηκε για να αποκτηθούν οι προγνωστικές ομάδες επιβίωσης. Αυτή περιλαμβάνει την πρόβλεψη των σχετιζόμενων με την επιβίωση γονιδίων και την παράλληλη ομαδοποίηση των δειγμάτων των όγκων χρησιμοποιώντας αυτά τα γονίδια. Οι προσδιορισμένες ομάδες θεωρήθηκαν ως προγνωστικές ομάδες και ένα μοναδικό γονιδιακό αποτύπωμα αποκτήθηκε για κάθε

ομάδα. Αυτή η μεθοδολογία εξαρτάται σε μεγάλο βαθμό από τις τεχνικές ομαδοποίησης και μπορεί να είναι λιγότερο ακριβής από τη χρήση των ιστοπαθολογικών ομάδων για το καθορισμό των αποτυπωμάτων της γονιδιακής έκφρασης. Άλλοι λόγοι περιλαμβάνουν τους πολυάριθμους εξωτερικούς παράγοντες που επηρεάζουν την πιθανότητα επιβίωσης και δεν σχετίζονται άμεσα με τον καρκίνο, όπως η ηλικία του ασθενούς, η φυσική και νευρολογική επίδοση κ.τ.λ. Γονίδια που κωδικοποιούν για τέτοιους παράγοντες θα εμφάνιζαν υψηλή συσχέτιση με την επιβίωση σε μικρές ομάδες δειγμάτων, όπως αυτές που χρησιμοποιούνται συχνά σε μελέτες μικροσυστοιχιών, χωρίς, όμως, να έχουν κάποια σχέση με τον καρκίνο *per se.* Ωστόσο, τέτοια γονίδια έχουν περιορισμένη προγνωστική χρησιμότητα όταν εφαρμοστούν σε άλλα καρκινικά δείγματα. Από την άλλη μεριά, τα προφίλ έκφρασης των γονιδίων βασιζόμενων στην ιστοπαθολογία είναι άμεσα συνδεδεμένα με τον καρκίνο και αναμένεται να είναι πιο σταθερά μεταξύ διαφορετικών ασθενών και, συνεπώς, να έχουν μεγαλύτερη χρησιμότητα για την πρόγνωση. Αν και υπάρχει σημαντική διαφοροποίηση μεταξύ των μελετών όσον αφορά στην επεξεργασία των δειγμάτων και στην ανάλυση και ετερογένεια των ιστών, που πιθανόν να επηρεάζουν την ταυτοποίηση των γονιδίων ταξινομητών, τα αποτελέσματά μας δείχνουν ότι είναι δυνατόν να χρησιμοποιηθούν δεδομένα έκφρασης για να ταυτοποιηθούν γονίδια με προγνωστική χρησιμότητα που επεκτείνεται σε πολλαπλά ανεξάρτητα σύνολα δεδομένων.

Στην παρούσα μελέτη επιλέχθηκαν δύο γονίδια ιδιαίτερου ενδιαφέροντος (τα *PEA15* και *ADM)* για περαιτέρω ανάλυση. Συγκεκριμένα, το γονίδιο PEA15 έχει προταθεί να έχει ογκοκατασταλτικές λειτουργίες (123). Το γονίδιο αυτό καταστέλλει την μεσολαβούμενη από τη DISC ενεργοποίηση της κασπάσης 8, περιορίζει την είσοδο στον κυτταρικό κύκλο και δεν έχει προηγουμένως σχετιστεί με την εξέλιξη των αστροκυττωμάτων. Φυσιολογικά επίπεδα έκφρασης του PEA15 σε καλλιέργειες αστροκυττάρων περιορίζουν την ERK κινάση στο κυτταρόπλασμα και παρεμποδίζουν την εξαρτώμενη από την ERK c-Fos μεταγραφή και τον κυτταρικό πολλαπλασιασμό (109). Πιθανά ογκοκατασταλτικά γονίδια, όπως το PEA, μπορεί να συνιστούν σημεία αναχαίτισης της εξέλιξης του όγκου. Είναι, λοιπόν, δυνατόν η υποέκφραση τέτοιων γονιδίων να συμβάλει άμεσα σε μία αλυσιδωτή σειρά γεγονότων που να οδηγούν στην εξέλιξη των πρώιμων βαθμίδων όγκων σε όψιμους πιο κακοήθεις φαινότυπους. Το *PEA 15* γονίδιο επιλέχθηκε για περαιτέρω ανάλυση, προκειμένου να μελετηθεί η υποκυτταρική του εντόπιση, αλλά και σε μία

προκαταρτική προσπάθεια να διαφωτιστούν πιθανές συσχετίσεις μεταξύ της έκφρασης του γονιδίου *PEA 15* και τον προγραμματισμένο κυτταρικό θάνατο των αστροκυττωματικών όγκων. Το ADM είναι ένα πεπτίδιο που αποτελείται από 52 αμινοξέα και θεωρείται ότι επηρεάζει την ανάπτυξη του όγκου μέσω άμεσων μιτογόνων επιδράσεων στα κύτταρα του όγκου, αλλά και με αγγειο-γονιδιακούς μηχανισμούς που σχετίζονται με την αγγείωση του οργάνου (124). Η έκφραση του γονιδίου ADM έχει δειχθεί στα αστροκυττώματα και οι Tso *et al* 2006 (118) πρόσφατα ανακάλυψαν ότι το γονίδιο αυτό εμφανίζει αυξανόμενη έκφραση κατά την εξέλιξη της βαθμίδας του όγκου. Η επιλογή του γονίδιου ADM πραγματοποιήθηκε προκειμένου να επιβεβαιωθούν προηγούμενες υποθέσεις που σχετίζουν το πεπτίδιο αυτό με την αγγειογένεση και επειδή πολύ λίγες δημοσιεύσεις σχολίαζαν την ακριβή υποκυτταρική ή ιστοειδική του εντόπιση.

Η παρούσα μελέτη παρουσιάζει ένα νέο, μεγάλο σύνολο δεδομένων έκφρασης των αστροκυττωμάτων και χρησιμοποιεί μία καινούργια, βασισμένη στα ΤΝΔ, διαβάθμιση αυτών των όγκων σε μοριακούς υπότυπους. Δείξαμε ότι είναι δυνατόν να συνάγουμε μεταγραφικά αποτυπώματα από την τριμερή ιστοπαθολογική διαβάθμιση που χρησιμοποιήθηκε για την εκπαίδευση του μοντέλου ΤΝΔ. Επιπλέον, αυτά τα αποτυπώματα επιτυγχάνουν μία πιο σημαντική πρόγνωση του χρόνου επιβίωσης όταν συγκριθούν με την ιστοπαθολογική διαβάθμιση, καθώς και με αναφορές υπότυπων όγκων από άλλες μελέτες. Ελπίζουμε ότι η ταυτοποίηση αυτής της νέας ομάδας γονιδίων που χαρακτηρίζουν αυτή την διαβάθμιση θα οδηγήσει σε μία πιο ποσοτική προσέγγιση στη διάγνωση των όγκων, σύμφωνα με την οποία τα οι όγκοι θα θεωρούνται στα πλαίσια ενός φάσματος κακοήθειας που περιλαμβάνει δείγματα απ' όλα τα στάδια της εξέλιξης της νόσου. Πιστεύουμε, επίσης, ότι η διαβάθμιση και οι προσπάθειες ομαδοποίησης που βασίζονται σε δεδομένα γονιδιακής έκφρασης θα πρέπει να επιτελούνται σε συνδυασμό με την εκτενή, σε όσο γίνεται περισσότερα επίπεδα, περιγραφή του όγκου. Ο συγκερασμός με κλινικές, γονοτυπικές και ιστοπαθολογικές περιγραφές μπορεί να μεγιστοποιήσει την αξία των δεδομένων έκφρασης γονιδίων, να αυξήσει την κατανόηση μας στην παθολογία των όγκου και να επεκτείνει περαιτέρω τις σύγχρονες διαγνωστικές και θεραπευτικές προσεγγίσεις.

Κεφάλαιο 3

3 Πρόβλεψη νέων γονιδίων microRNA σε καρκινικά σχετιζόμενες γενωμικές περιοχές (CAGR) - μια υπολογιστική και πειραματική προσέγγιση.

3.1 Εισαγωγή

Τα microRNA (miRNA) ανήκουν σε μία πρόσφατα προσδιορισμένη ομάδα της ευρείας οικογένειας των μη κωδικοποιών RNA (125). Το ώριμο (mature) miRNA έχει μήκος συνήθως 19-27 νουκλετίδια και προέρχεται από έναν μεγαλύτερο πρόδρομο μόριο το οποίο αναδιπλώνεται (folds into) σε μία ατελή δομή φουρκέτας (δομή hairpin ή stem – loop).

Ο τρόπος δράσης του ώριμου miRNA στο σύστημα των θηλαστικών καθορίζεται από τη συμπληρωματικότητα των βάσεων (complementary base pairing), πρώτιστα στη 3'UTR περιοχή του στοχευόμενου mRNA, προκαλώντας στη συνέχεια την παρεμπόδιση της μετάφρασης ή/και την αποικοδόμηση του mRNA.

Με βάση πρόσφατες εκτιμήσεις, ενώ πάνω από το 30% του γονιδιώματος των σπονδυλωτών μεταγράφεται (126), μόνο το 1% αποτελείται από κωδικοποιούντα γονίδια, υποδηλώνοντας ότι τα υπόλοιπα πρέπει να είναι διάφοροι τύποι μη-κωδικοποιών RNA γονιδίων. Επιπλέον, περισσότερες από 6000 miRNA αλληλουχία φουρκέτας (hairpin) περιέχονται ήδη σε βάσεις δεδομένων miRNA, από τις οποίες το 57% έχουν επιβεβαιωθεί πειραματικά και αναμένεται ότι μπορεί να υπάρξουν χιλιάδες ακόμη. Συνεπώς, η αναζήτηση νέων γονιδίων miRNA μέσα σε αυτό το τεράστιο πλήθος δεδομένων, είναι μια πολύπλοκη διεργασία για την οποία απαιτούνται γρήγορες, ευέλικτες και αξιόπιστες μέθοδοι. Για το σκοπό αυτό, οι διαθέσιμες σήμερα πειραματικές προσεγγίσεις είναι σύνθετες και όχι απαραίτητα οι βέλτιστες. Οι ανεπάρκειες είναι πολλαπλές και πηγάζουν από τη δυσκολία να απομονωθούν συγκεκριμένα miRNA με κλωνοποίηση λόγω της χαμηλής έκφρασης, της σταθερότητας, και της ιδιομορφίας του ιστού και γενικότερα τις τεχνικές

δυσκολίες κατά τη διαδικασία της κλωνοποίησης. Επίσης, η επιλογή για εξέταση της σωστής γενωμικής περιοχής είναι συχνά από μόνη της μια απαιτητική διεργασία. Η υπολογιστική πρόβλεψη των miRNA γονιδίων από γενωμικές αλληλουχίες είναι μια εναλλακτική τεχνική η οποία προσφέρει έναν πολύ πιο γρήγορο, χαμηλού κόστους και αποτελεσματικό τρόπο προσδιορισμού υποθετικών miRNA γονιδίων. Επιπλέον, προβλέποντας τη θέση των υποθετικών miRNA γονιδίων, αυτές οι μέθοδοι επιτρέπουν στους ερευνητές να επικεντρώνουν τις προσπάθειές τους στις γενωμικές εκείνες περιοχές στις οποίες υπάρχει μεγαλύτερη πιθανότητα να βρεθούν miRNA γονίδια, με αποτέλεσμα να διευκολύνεται έτσι η διαδικασία εύρεσης νέων miRNA γονιδίων.

Η ακριβής πρόβλεψη νέων miRNA προϋποθέτει ότι θα ληφθούν υπόψιν ορισμένες χαρακτηριστικές ιδιότητες των μορίων αυτών, γνωστές είτε από πειραματικά (67,127,128) είτε από υπολογιστικά δεδομένα (73,75,129-131), οι οποίες μπορεί να χρησιμοποιηθούν για τη δημιουργία ενός συστήματος ταξινόμησης ή ενός προγνωστικού μοντέλου. Οι γενικές αυτές ιδιότητες ή γνωρίσματα περιλαμβάνουν την αλληλουχία βάσεων, τη δευτεροταγή δομή και τη συντήρηση. Η πρόβλεψη miRNA γονιδίων μπορεί να επιτευχθεί μέσω της χρήσης επιβλεπόμενων αλγορίθμων που εκπαιδεύονται στα βιολογικά χαρακτηριστικά γνωρίσματα των ήδη γνωστών miRNA και χρησιμοποιούνται έπειτα για να προσδιορίσουν νέα, υποψήφια miRNA γονίδια. Εναλλακτικά η πρόβλεψη μπορεί να γίνει μέσω μη- επιβλεπόμενων αλγορίθμων όπως η ευθυγράμμιση δυο αλληλουχιών (Blast) ή η συντήρηση. Η μεθοδολογία πρόβλεψης συχνά ποικίλει σημαντικά μεταξύ διαφορετικών μελετών και μπορεί να εκτελεσθεί με διάφορους τρόπους: μέσω σάρωσης για αναζήτηση δομών φουρκέτας (hairpins) σε αλληλουχίες συντηρημένες ανάμεσα σε συγγενείς οργανισμούς (*C.elegans* και *C.briggsae*) (130,132), μέσω αναζήτησης για περιοχές ομολογίας μεταξύ γνωστών miRNA και άλλων περιοχών σε ευθυγραμμισμένα γονιδιώματα (π.χ ανθρώπινο και από ποντίκι) (133), είτε μέσω αναζήτησης συντηρημένων περιοχών με γονίδια διευθετημένα σε σειρά (conserved regions of systeny) στα γονιδιώματα συγγενών οργανισμών (133). Μέθοδοι αναζήτησης miRNA χρησιμοποιώντας αλληλουχίες από ισχυρά αποκλίνοντες οργανισμούς (π.χ. ποντίκι και fugu) έχουν επίσης προταθεί, βασιζόμενες στα *Profile* (134) και στην ευθυγράμμιση δευτεροταγούς δομής (secondary structure alignment) (135). Νέα miRNA έχουν επίσης προβλεφθεί με τη χρήση Support vector machines, μέθοδος που λαμβάνει υπόψη πολλαπλά χαρακτηριστικά γνωρίσματα όπως η ελεύθερη ενέργεια,

οι συμπληρωματικές βάσεις (paired bases), το μήκος φουρκέτας (loop length), η συντήρηση (stem conservation) κλπ. (75,129,136). Πολλές από αυτές τις προγνωστικές μεθόδους υιοθετούν μια σειριακή προσέγγιση (pipeline approach). Κατά τη διαδικασία αυτή χρησιμοποιούνται αυστηρά κατώφλια έτσι ώστε να αποκλείονται σταδιακά υποψήφιες αλληλουχίες (130,132), το οποίο έχει ως αποτέλεσμα την απόρριψη και πολυάριθμων αληθινών miRNA. Άλλες προσεγγίσεις εντοπίζουν νέα miRNA χρησιμοποιώντας ομολογία με γνωστά miRNA (133-135). Αυτές οι μέθοδοι προφανώς αποτυγχάνουν όταν σαρώνουν ισχυρά αποκλίνοντα γονιδιώματα και όταν το νέο miRNA δεν παρουσιάζει ομολογία με κάποιο άλλο, γνωστό, γονίδιο. Δύο πρόσφατες μελέτες (73,137), επεξεργάζονται ταυτόχρονα πληροφορίες αλληλουχίας και δομής χρησιμοποιώντας Κρυφά Μαρκοβιανά Μοντέλα (ΚΜΜ) και Bayesian ταξινομητές, αντίστοιχα, για να προσδιορίσουν νέα πρόδρομα miRNA γονίδια (pre-miRNA). Στους αλγόριθμους όμως αυτούς δεν ενσωματώνεται πληροφορία συντήρησης, ένα χαρακτηριστικό γνώρισμα της πλειοψηφίας (>50%) των miRNA γονιδίων. Τέλος, η πιο πρόσφατα δημοσιευμένη μελέτη που χρησιμοποιεί ΚΜΜ και επεξεργάζεται ταυτόχρονα πληροφορίες δομής και συντήρησης (138), αποδείχτηκε πολύ αποτελεσματική στην πρόβλεψη γονιδιών miRNA στον άνθρωπο επιτυγχάνοντας ~70% και ~90% ευαισθησία και ειδικότητα αντίστοιχα.

Εκτός από υπολογιστικά εργαλεία, μέθοδοι ευρείας κλίμακας, μαζικής ανάλυσης (large scale, high throughput) όπως tiling arrays και deep sequencing έχουν πρόσφατα χρησιμοποιηθεί για τον προσδιορισμό νέων miRNA γονιδίων (139-141). Αυτές οι μέθοδοι είναι ιδιαίτερα χρήσιμες καθώς μπορούν να παρέχουν ένα πολύ λεπτομερή και αξιόπιστο χάρτη έκφρασης για τα μικρά RNAs στο γονιδίωμα. Επιπλέον, εάν αυτά τα δεδομένα συνδυαστούν με τα υπολογιστικά εργαλεία, μπορεί να διευκολύνουν τη γρήγορη και έμπιστη ανακάλυψη νέων miRNA, προσφέροντας συγχρόνως μεγαλύτερη αξιοπιστία στις υπολογιστικές προβλέψεις.

Όπως αναλύθηκε στο Κεφάλαιο 2, τα miRNA παίζουν σημαντικό ρυθμιστικό ρόλο σε πολλές μοριακές διαδικασίες, συμπεριλαμβανομένου και του καρκίνου, είτε μέσω της διαφορποιημένης τους έκφρασης, είτε επειδή στοχεύουν κακρινικά γονίδια ή επειδή εμφανίζονται σε καρκινικά σχετιζόμενες γενωμικές περιοχές (Cancer Associated Genomic Regions - CAGR) (77,142). Οι CAGR παίρνουν τη μορφή (*i*) των

ελάχιστων περιοχών απώλειας ετεροζυγώτητας (loss of heterozygosity (LOH)) ενδεικτικές της παρουσίας γονιδίων ογκοκαταστολής, (*ii*) των ελάχιστων περιοχών ενίσχυσης, ενδεικτικές της παρουσίας ογκογονιδίων και (*iii*) των σημείων που εμφανίζουν συχνά θραύσεις (common breakpoint regions) είτε μέσα είτε κοντά σε πιθανά ογκογονίδια ή γονίδια ογκοκαταστολής. Ο προσδιορισμός νέων miRNA γονιδίων μέσα σε αυτές τις περιοχές είναι πολύ σημαντικός καθώς μπορεί να αποκαλύψει νέα υποθετικά γονίδια με ρυθμιστική επίδραση σε διαφορετικούς τύπους καρκίνου, να συμβάλλει στην καλύτερη κατανόηση των μοριακών μονοπατιών που εμπλέκονται στην ογκογένεση και να παρέχει δυνητικούς στόχους για θεραπευτική παρέμβαση.

Στο κεφάλαιο αυτό παρουσιάζουμε ένα αποτελεσματικό και ελεύθερα διαθέσιμο εργαλείο πρόβλεψης νέων miRNA (*SSCprofiler)*, το οποίο χρησιμοποιεί *Profile* KMM ώστε να αναγνωρίζει βασικά βιολογικά γνωρίσματα των miRNA όπως η αλληλουχία, η δομή και η συντήρηση. Αρχικά για την εκπαίδευση του εργαλείου χρησιμοποιήθηκαν 249 γνωστά ανθρώπινα πρόδρομα miRNA και στη συνέχεια το εκπαιδευμένο μοντέλο εφαρμόστηκε σε CAGR για αναζήτηση νέων miRNA γονιδίων. Οι προβλέψεις ταξινομούνται βάσει των πληροφοριών έκφρασης από μια πρόσφατα δημοσιευμένη μελέτη tiling array που καλύπτει όλο το ανθρώπινο γονιδίωμα (140). Τα τέσσερα υποψήφια miRNA γονίδια με την υψηλότερη έκφραση επιβεβαιώθηκαν πειραματικά χρησιμοποιώντας ανάλυση Northern blot.

3.2 Υλικά και Μέθοδοι

3.2.1 Σύνολα Δεδομένων

Για την εκπαίδευση και δοκιμή των μοντέλων KMM χρησιμοποιήσαμε ανθρώπινα πρόδρομα miRNA από την online βάση δεδομένων miRBase (έκδοση 12.0 - http://microrna.sanger.ac.uk/sequences/). Αρχικά εφαρμόστηκε ο αλγόριθμος BLASTclust (143) για την ομαδοποίηση των miRNA αλληλουχιών σε ομάδες με βάση την ομοιότητα των πρόδρομων αλληλουχιών και το πιο συντηρημένο μέλος (σύμφωνα με τα multiz αρχεία) χρησιμοποιήθηκε για να αναπαραστήσει την ομάδα. Η μέθοδος αυτή χρησιμοποιήθηκε έτσι ώστε να καταργηθούν τα επαναλαμβανόμενα πρόδρομα miRNA και να αποφευχθεί η υπερβολική αναπαράσταση όμοιων

προδρόμων. Έπειτα από χρήση μιας σειράς από παραμέτρους φιλτραρίσματος καταλήγουμε σε 249 αλληλουχίες (οι οποίες αποτελούν ένα υποσύνολο της miRBase έκδοση 8.0). Αυτές οι αλληλουχίες χρησιμοποιήθηκαν για την εκμάθηση των ΚΜΜ και ένα σύνολο από 219 αλληλουχίες (οι οποίες δεν αποτελούν υποσύνολο της miRBase έκδοση 8.0) χρησιμοποιήθηκαν για δοκιμή. Οι αρνητικές αλληλουχίες παράχθηκαν χρησιμοποιώντας ένα κυλιόμενο παράθυρο (sliding window) 104nt, το οποίο μετακινούνταν 11nt τη φορά, πάνω στις 3'UTR περιοχές του ανθρώπινου γονιδιώματος (έκδοση - Μάιος 2004). Για κάθε μετακίνηση εκτελέστηκε το πρόγραμμα RNAfold και σημειώθηκε η ελεύθερη ενέργεια της δευτεροταγούς δομής. Επιλέχθηκαν μόνο οι αλληλουχίες των οποίων η ενέργεια δεν υπερέβαινε το κατώφλι των -14.44kcal/mol [g] και είχαν τουλάχιστον το 14% [h] των νουκλεοτιδίων τους συντηρημένα. Καταλήξαμε έτσι σε 35,000 αρνητικές αλληλουχίες προδρόμων miRNA.

3.2.2 Βιολογικά Γνωρίσματα

Το SSCprofiler λαμβάνει υπόψη τρία διαφορετικά βιολογικά γνωρίσματα: την αλληλουχία, τη δομή και τη συντήρηση των πρόδρομων miRNA. Στην παρούσα μελέτη, δεδομένα συντήρησης ανακτήθηκαν από τις multiz (144) πλήρεις γενωμικές ευθυγραμμίσεις του ανθρώπινου γονιδιώματος (έκδοση Μάιος 2004, hg17) και 7 ακόμα γονιδιώματα σπονδυλωτών: Ποντικός (Μάιος 2004, mm5), Αρουραίος (Ιούνιος 2003, rn3), Σκύλος (Ιούλιος 2004, canFam1), Κοτόπουλο (Φεβρουάριος 2004, galGal2), Fugu (Αύγουστος 2002, fr1), Zebrafish (Νοέμβριος 2003, danRer1). Το γονιδίωμα του χιμπαντζή δεν συμπεριλήφθηκε εξαιτίας της υψηλής ποσοστιαίας ομοιότητας (~95%) με το ανθρώπινο. Η πρόβλεψη δευτεροταγούς δομής RNA επιτελέστηκε με χρήση της συνάρτησης RNAfold του πακέτου Vienna-RNA (79). Η εκπαίδευση των ΚΜΜ απαιτεί την ευθυγράμμιση πολλαπλών αλληλουχιών (multiple sequence alignment - msa), η οποία επιτεύχθηκε χρησιμοποιώντας ένα σταθερό παράθυρο 104nts (βλ. Εκπαίδευση και Επαλήθευση και Δοκιμή). Για τις αλληλουχίες που ήταν μικρότερες από το παράθυρο λήφθησαν υπόψιν τα επιπλέον νουκλεοτίδια εκατέρωθεν των γονιδίων miRNA, ενώ αντίστοιχα για τις αλληλουχίες που υπερέβαιναν το καθορισμένο παράθυρο αφαιρέθηκαν νουκλεοτίδια. Το μήκος του

[g] Η υψηλότερη τιμή ελεύθερης ενέργειας που χαρακτηρίζει τα miRNA στην miRASE v12.
[h] Η τιμή αυτή ορίστηκε κατά σύμβαση.

παραθύρου συνεπώς χρησιμοποιήθηκε αφενός ως μήκος του μοντέλου εκπαίδευσης και αφετέρου ως μέγεθος παραθύρου για σάρωση γενωμικών αλληλουχιών.

3.2.3 Φιλτράρισμα

Με σκοπό την ελαχιστοποίηση του χώρου αναζήτησης και την ελάττωση του υπολογιστικού φόρτου, τα δεδομένα πρώτα φιλτραρίστηκαν με χρήση διάφορων γνωρισμάτων δευτεροταγούς δομής των miRNA. Τα αποτελέσματα του φιλτραρίσματος αναπαραστήθηκαν σε ιστογράμματα που παρουσιάζουν τις σχετικές κατανομές των θετικών και αρνητικών δεδομένων σε σχέση με 8 παραμέτρους:

1. Hairpin – το πλήθος των hairpins (φουρκέτα)
2. Bulges – το πλήθος των bulges
3. Βρόγχοι – το πλήθος των βρόγχων
4. Ασυμμετρία – διαφορά σε βρόγχους + bulges on either side of the hairpin.
5. Bulges-Loops – άθροισμα των βρόγχων και bulges
6. Hairpin Length – το μήκος του hairpin
7. Folding min energy – ελάχιστη ενέργεια ως ορίζεται από το RNAfold
8. Συντήρηση – σύμφωνα με τα multiz full genome alignment files

Η αναπαράσταση των κατανομών δεδομένων για τις διάφορες παραμέτρους φιλτραρίσματος πραγματοποιήθηκε για τη διευκόλυνση της διαδικασίας φιλτραρίσματος, επιτρέποντας τη ρύθμιση των κατωφλιών ανάλογα με το συγκεκριμένο σύνολο δεδομένων. Το κατώφλι για καθένα από αυτά τα γνωρίσματα είναι τροποποιήσιμο τόσο πριν όσο και μετά τη διαδικασία εκπαίδευσης (βλ. Αποτελέσματα, Σχήμα 3.4).

3.2.4 Συνδυασμός Αλληλουχίας, Δομής και Συντήρησης

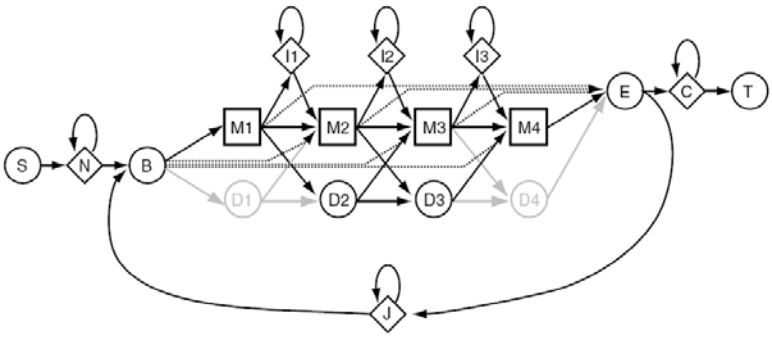
Με στόχο να ληφθούν υπόψη ταυτόχρονα πολλαπλά βιολογικά γνωρίσματα, αναπτύχθηκε ένας κώδικας 16 χαρακτήρων, που ενοποιεί πληροφορία αλληλουχίας, δομής και συντήρησης για κάθε θέση νουκλεοτιδίου σε μια δεδομένη γενωμική αλληλουχία. Συγκεκριμένα, κάθε θέση της γενωμικής αλληλουχίας αντικαθίσταται από 1 εκ των 16 γραμμάτων, με βάση τρεις παράγοντες: 1) Αλληλουχία (A,C,U,G), 2) Δομή, (M=match, L=loop); και 3) Συντήρηση (*= συντηρημένο, ‘ ‘=μη-συντηρημένο), όπως περιγράφεται λεπτομερώς στον Πίνακα 3.1.

Πίνακας 3.1 Ο κώδικας 16 γραμμάτων που χρησιμοποιήθηκε κατά την εκπαίδευση και την πρόβλεψη.

	sequence			
conservation & structure	**A**	**C**	**G**	**U**
*** M**	L	M	N	P
‘ ‘M	C	D	E	F
*** L**	Q	R	S	T
‘ ‘L	G	I	H	K

3.2.5 *Profile* Κρυφά Μαρκοβιανά Μοντέλα (KMM)

Για την κατασκευή ενός KMM ικανού να προβλέπει miRNA *Profiles* χρησιμοποιήθηκε το πακέτο λογισμικού HMMER (145). Τα KMM είναι *γενικά πιθανοκρατικά μοντέλα* που χρησιμοποιούνται συχνά για την επίλυση σημαντικών θεωρητικών προβλημάτων. Για ορθή εξαγωγή στατιστικού συμπεράσματος, είναι απαραίτητος ο υπολογισμός μιας κατανομής πιθανότητας $P(S|M)$ για την πιθανότητα αλληλουχιών S δεδομένου ενός μοντέλου M, και η αναγωγή αυτής της ποσότητας στη μονάδα στο «χώρο» όλων των αλληλουχιών. Τα generative μοντέλα λειτουργούν απαριθμώντας *αναδρομικά* τις δυνατές αλληλουχίες από ένα πεπερασμένο σύνολο κανόνων – κανόνες που σε ένα KMM αναπαριστώνται με καταστάσεις, μεταβάσεις καταστάσεων και πιθανότητες εκπομπής συμβόλων. Το HMMER χρησιμοποιεί μια αρχιτεκτονική *Profile* KMM ονόματι Plan 7, που παρουσιάζεται στο Σχήμα 3.1. Τα *Profile* KMM είναι στατιστικά μοντέλα από ευθυγραμμίσεις πολλαπλών αλληλουχιών. Αντλούν πληροφορία σχετικά με το πόσο συντηρημένη είναι η κάθε στήλη της ευθυγράμμισης και ποια κατάλοιπα είναι πιο πιθανά και στην συνέχεια μπορούν να χρησιμοποιηθούν για να αποδώσουν μία βαθμολογία (HMM score ή likelihood score) για κάποια ανεξάρτητη αλληλουχία

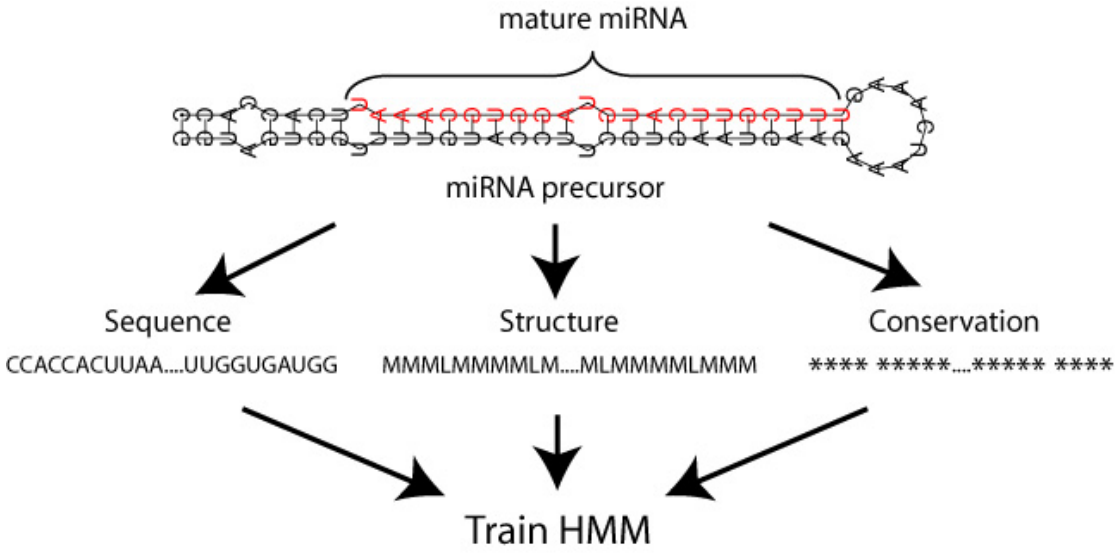


Σχήμα 3.1 Η αρχιτεκτονική HMMER Plan 7. Τα τετράγωνα υποδηλώνουν καταστάσεις ταιριάσματος (modelling consensus positions in the alignment). Οι ρόμβοι υποδηλώνουν καταστάσεις εισαγωγής (modelling insertions relative to consensus) και ειδικές καταστάσεις εκπομπής τυχαίας αλληλουχίας. Οι κύκλοι υποδηλώνουν καταστάσεις εξάλειψης (modelling deletions relative to consensus) και ειδικές καταστάσεις αφετηρίας/τερματισμού. Τα βέλη υποδηλώνουν μεταβάσεις καταστάσεων.

3.2.6 Εκπαίδευση, Επαλήθευση και Γενίκευση

Οι αλγόριθμοι μηχανικής μάθησης, όπως τα ΚΜΜ απαιτούν προσεκτικά επιλεγμένα σύνολα δεδομένων εκπαίδευσης και επαλήθευσης, ώστε να επιτύχουν μέγιστη απόδοση. Το *SSCprofiler* επιτρέπει το διαχωρισμό από τον χρήστη των εισαγόμενων δεδομένων σε σύνολα εκπαίδευσης και επαλήθευσης, με σκοπό να εκτελεστεί μια διαδικασία επαλήθευσης (boosting validation). Αυτό επιτυγχάνεται με τον τυχαίο καταμερισμό των θετικών δεδομένων σε *K* υποσύνολα, ορισμένα εκ των οποίων χρησιμοποιούνται για εκπαίδευση και άλλα για επαλήθευση. Τα αρνητικά δεδομένα χρησιμοποιούνται αποκλειστικά για επαλήθευση και δεν συμπεριλαμβάνονται στα σύνολα εκπαίδευσης. Ο καταμερισμός αυτός επαναλαμβάνεται 100 φορές και υπολογίζεται μια μέση τιμή απόδοσης για την επαλήθευση. Τα αποτελέσματα εκπαίδευσης/επαλήθευσης αναπαριστώνται σε διαγράμματα ευαισθησίας και ειδικότητας, με σκοπό να αποκομιστεί μια ένδειξη για το πόσο καλά αποδίδουν τα εκπαιδευμένα ΚΜΜ στο συγκεκριμένο σύνολο δεδομένων (βλ. Σχήμα 3.5). Ο άξονας x σε αυτά τα διαγράμματα αναπαριστά την βαθμολογία (ή κατώφλι) που αποδίδουν τα ΚΜΜ και ο άξονας y είναι η μέση ευαισθησία και ειδικότητα για κάθε κατώφλι

για τις 100 εκτελέσεις επαλήθευσης. Η εκπαίδευση επιτελείται στα βιολογικά γνωρίσματα που έχουν επιλεγεί εκ των προτέρων. Μια σύνοψη της διαδικασίας εκπαίδευσης παρουσιάζεται σχηματικά στο διάγραμμα ροής του Σχήματος 3.2. Στο τέλος της διαδικασίας εκπαίδευσης/επαλήθευσης όλα τα πραγματικά miRNA συνδυάζονται για να εκπαιδεύσουν ένα τελικό μοντέλο, που χρησιμοποιείται στη συνέχεια για τη σάρωση γενωμικών περιοχών. Ο χρήστης μπορεί να επιλέξει το κατώφλι εκείνο του KMM στο οποίο οι μέσες τιμές ευαισθησίας και ειδικότητας είναι βέλτιστες και να το χρησιμοποιήσει ως κατώφλι για την ταξινόμηση αλληλουχιών ως θετικές (πραγματικά miRNA) ή αρνητικές (βλ. παρακάτω).



Σχήμα 3.2 Η επιβλεπόμενη διαδικασία εκπαίδευσης KMM για την αναγνώριση πρόδρομων miRNA. Βιολογικά γνωρίσματα της βιογένεσης των miRNA και συντήρησης σε ισχυρά αποκλίνοντα γονιδιώματα χρησιμοποιούνται ως είσοδος για εκπαίδευση. Αρχικά, διενεργείται πρόβλεψη δευτεροταγούς δομής με χρήση του προγράμματος RNAfold. Κάθε θέση νουκλεοτιδίου εφεξής αναπαριστάται με ένα 'M' για match/ταίριασμα και ένα 'L' για loop/βρόγχο. Αυτή η πληροφορία ενοποιείται με πληροφορία συντήρησης και αλληλουχίας για κάθε θέση νουκλεοτιδίου. Ο κώδικας 16 χαρακτήρων που φαίνεται στον Πίνακα 3.1 χρησιμοποιείται στη συνέχεια για να αναπαραστήσει κάθε μία από τις θέσεις αυτές με ένα μοναδικό γράμμα. Οι προκύπτουσες σειρές χαρακτήρων για τα πραγματικά miRNA ευθυγραμμίζονται σε σχέση με τις δομές φουρκέτας/hairpins τους και χρησιμοποιούνται ως σύνολο εκπαίδευσης για το KMM. Εφόσον εκπαιδευτούν, τα KMM, μπορούν να

χρησιμοποιηθούν για να αναλυθούν αλληλουχίες επιθυμητού μήκους και να αποδοθεί μια βαθμολογία. Όσο υψηλότερη είναι αυτή η τιμή τόσο αυξάνεται η πιθανότητα μια υποψήφια αλληλουχία να αποτελεί ένα πραγματικό πρόδρομο miRNA.

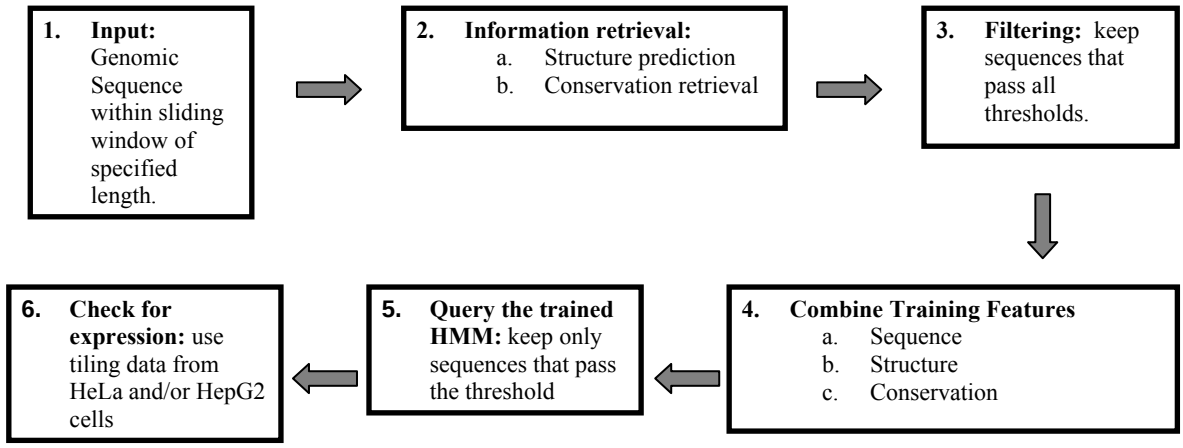
3.2.7 Εκτίμηση του Επιπέδου Έκφρασης των Προβλεφθέντων υποψηφίων MiRNA με Χρήση Δεδομένων Tiling Array

Το *SSCprofiler*, ως επιπλέον υποστήριξη στις υπολογιστικές προβλέψεις, εξετάζει αν οι περιοχές που εχουν προβλεφθεί ως υποψήφια miRNA εκφράζονται σε κύτταρα HeLa ή HepG2. Η πληροφορία έκφρασης βασίζεται σε μια πρόσφατα δημοσιευμένη μελέτη, η οποία χρησιμοποιώντας tiling arrays, παρέχει ένα χάρτη έκφρασης για μικρά RNA (20-200nt) σε ανάλυση 5nt για τις δύο αυτές κυτταρικές σειρές (140). Το *SSCprofiler* επιτρέπει στο κατώφλι έκφρασης να διαμορφωθεί σε εύρος από 1 έως 2000, με σκοπό να διατηρηθούν υποψήφια miRNA που ξεπερνούν ένα καθορισμένο κατώφλι έκφρασης. Για τα αποτελέσματα που παρατίθενται εδώ χρησιμοποιήθηκε ως κατώφλι η τιμή 200.

3.2.8 Σάρωση Γενωμικών Περιοχών για miRNA Profiles

Η διαδικασία σάρωσης γενωμικών περιοχών για Profile πρόδρομων miRNA απαρτίζεται από 6 βήματα, που παρουσιάζονται στο Σχήμα 3.3. Βήμα 1 – Ένα κυλιόμενο παράθυρο επιλεγμένου μήκους μετακινείται κατά μήκος της γενωμικής αλληλουχίας με βήμα 1nt. Βήμα 2 – Σε κάθε μετακίνηση του παραθύρου ανακτάται πληροφορία σχετικά με τη δομή και τη συντήρηση της αλληλουχίας, ανάλογα με τα καθορισμένα γνωρίσματα εκπαίδευσης, δηλαδή επιτελείται πρόγνωση δομής και αντλείται πληροφορία συντήρησης από τα αρχεία multiz. Βήμα 3 – Στη συνέχεια η αλληλουχία εντός του κυλιόμενου παραθύρου φιλτράρεται, με βάση τις προκαθορισμένες παραμέτρους φιλτραρίσματος (Hairpin Length, Ασυμμετρία κλπ). Βήμα 4 – Το βήμα αυτό αφορά την ενοποίηση γνωρισμάτων που χρησιμοποιούνται στην εκπαίδευση (αλληλουχία, δομή και συντήρηση) με τη χρήση του κλειδιού 16 γραμμάτων που περιγράφηκε νωρίτερα. Αυτό επιτρέπει την ταυτόχρονη ενσωμάτωση πληροφορίας για κάθε θέση νουκλεοτιδίου στη γενωμική αλληλουχία. Βήμα 5 – Το εκπαιδευμένο μοντέλο KMM χρησιμοποιείται για να αποδοθεί μια βαθμολογία για κάθε γενωμική αλληλουχία εντός του κυλιόμενου παραθύρου. Ως τιμή κατωφλιού

ΚΜΜ ορίζεται η τιμή όπου η ευαισθησία και η ειδικότητα από τη διαδικασία εκπαίδευσης/επαλήθευσης ήταν βέλτιστες. Βήμα 6 – Τα υποψήφια miRNA τα οποία παρουσιάζουν αλληλοεπικάλυψη ≤50nt ομαδοποιούνται και το υποψήφιο miRNA με τη υψηλότερη βαθμολογία όπως αποδίδεται από το ΚΜΜ, χρησιμοποιείται για να αναπαραστήσει την ομάδα. Τέλος, γίνεται ταξινόμηση των υποψηφίων miRNA ανάλογα με την έκφρασή τους σε κύτταρα HeLa ή HepG2 με χρήση δεδομένων από tiling arrays.



Σχήμα 3.3. Τα 6 βήματα της διαδικασίας σάρωσης γενωμικών περιοχών για νέα υποψήφια miRNA γονίδια.

3.2.9 Απομόνωση RNA και Northern Blot Ανάλυση

Από την κυτταρική σειρά HeLa πραγματοποιήθηκε ολική απομόνωση RNA χρησιμοποιώντας Trizol. 80μg συνολικού RNA αναλύθηκε σε ένα 15% αποδιατακτικό gel πολυακρυλαμίδης το οποίο περιείχε 8M ουρίας και μεταφέρθηκε σε μεμβράνη Νιτροκυτταρίνης (Schleicher &Schuell, Germany). Οι μεμβράνες ήταν σημασμένες με συγκεκριμένα DNA ολιγονουκλεοτίδια, συμπληρωματικά και στις δυο πολικότητες του RNA στόχου. Εξαιτίας της δυσκολίας πρόβλεψης της ακριβής τοποθεσίας του ώριμου πάνω στο πρόδρομο miRNA, επιλέξαμε και τις δυο stem αλληλουχίες (μέγιστο μέγεθος 50nt) από την δομή φουρκέτας (stem-loop) των υποψήφιων miRNA γονιδίων (Πίνακας 3.2). Τα ολιγονουκλεοτίδια (10 pmol από το

καθένα) ανιχνευτές σημάνθηκαν στα άκρα τους με [γ-32P]ATP χρησιμοποιώντας πολυνουκλεοτιδική κινάση T4. Η προ-υβριδοποίηση των φίλτρων πραγματοποιήθηκε σε 7% SDS, 5x SSC, 1x Denhardt's διάλυμα και 0.02M Na_2HPO_4 pH 7.2. Οι υβριδοποιήσεις πραγματοποιήθηκαν στο ίδιο διάλυμα σε θερμοκρασία 50^oC έπειτα από την προσθήκη των ραδιο-σημασμένων DNA ολιγονουκλεοτιδίων. Έπειτα από μια ολονύκτια υβριδοποίηση, οι μεμβράνες πλύθηκαν 2 φορές για 30 λεπτά στους 50^oC σε ήπιο ρυθμιστικό διάλυμα [2x SSC, 0.3% SDS] (54). Οι μεμβράνες πλύθηκαν 2 φορές για 30 λεπτά στους 80^oC σε ένα ισχυρό ρυθμιστικό διάλυμα (stripping buffer- 0.1x SSC and 0.5% SDS) και ξανα-σημάνθηκαν με τα ολιγονουκλεοτίδια αρνητικής πολικότητας.

Πίνακας 3.2 DNA ολιγονουκλεοτίδια. Πίνακας που χρησιμοποιήθηκε από τους Oulas et al, 2009 (146).

	Oligo stem 1	Oligo stem 2
1	5'-ATCTACCAGGTCCTGGGCTTCGGGCCGCGTTCCCAAGGCAAGC-3'	5'-ACCCGCGGCGAGGACACGGCCGACCGCCCGCCTGCGCCC-3'
2	5'-GCCCAGGAGGAGGTGGCACATCTGGGCTCCAGTCCTCGCAC-3'	5'-GTGCGGGCACCGCGCGAGCCTCGCCCCTTCCCACCTGCGC-3'
3	5'-ACCTCTCCCCCTGCCAGGTTCCACCAGGGGACACCGTGTGTGT-3'	5'-CGAGCAGGGCTCCCCCACCTGAGTACCTGACCATGGGCTTTGGAGAGGC-3'
4	5'-TACGCCCACAGCCCCCAGGCCCCCGAAGACAGGTGTCATGGA-3'	5'-TCCAAGAGCATCAAGCAGCAGGGGCTGGGGGAGCCAGCAGG-3'

3.3 Αποτελέσματα

3.3.1 Ακρίβεια Πρόβλεψης των Ανθρώπινων Πρόδρομων miRNA

Τα ανθρώπινα πρόδρομα miRNA από την RNA βάση δεδομένων mirBase έκδοσης 12.0 χρησιμοποιήθηκαν για να εξετάσουμε την ακρίβεια πρόβλεψης του *SSCprofiler*. Για λόγους αξιολόγησης χρησιμοποιήθηκαν πάνω από 35,000 αρνητικές αλληλουχίες προδρόμων miRNA. Οι αρνητικές αλληλουχίες ήταν δομές φουρκέτας προερχόμενες από τις 3'UTR περιοχές. Οι περιοχές αυτές επιλέχθηκαν λόγω του ότι δεν είχε προαναφερθεί να περιέχουν miRNA γονίδια. Χρησιμοποιήσαμε φίλτρα ελεύθερης ενέργειας και συντήρησης, για να διασφαλίσουμε ότι οι αρνητικές αλληλουχίες μοιάζουν σε μεγάλο βαθμό με τα πραγματικά πρόδρομα miRNA, όσον αφορά τη δομή και την εξελικτική συντήρηση έτσι ώστε να χρησιμεύσουν ως ένα αξιόπιστο σύνολο ελέγχου (control set),. Προτού προχωρήσουμε στην εκμάθηση όλα τα θετικά και αρνητικά δείγματα υποβλήθηκαν σε μια διαδικασία φιλτραρίσματος βάση διαφόρων παραμέτρων. Σε ένα αρχικό φιλτράρισμα με κατώφλι ελάχιστης ενέργειας -

25.44kcal/mol προέκυψαν 258 miRNA και ~8,000 αρνητικές αλληλουχίες. Στη συνέχεια, χρησιμοποιήθηκαν 7 επιπλέον παράμετροι φιλτραρίσματος (βλέπε Υλικά και Μέθοδοι) έτσι ώστε να εξαλειφθούν όσο το δυνατόν περισσότερα από τα μη πραγματικά θετικά (false positives). Στο σχήμα 3.4 αναπαρίστανται τα ιστογράμματα κατανομής πριν το φιλτράρισμα όπως παράχθησαν από τον *SSCprofiler* λαμβάνοντας υπόψην τρεις διαφορετικές παραμέτρους φιλτραρίσματος: Μήκος φουρκέτας, Ασσυμετρία και τα Bulges-Loops. Παρόμοιες κατανομές παράχθηκαν και για τις επτά παραμέτρους φιλτραρίσματος με σκοπό τον καθορισμό των βέλτιστων οριακών τιμών για τον διαχωρισμό των πραγματικών miRNA γονιδίων από τα αρνητικά δεδομένα. Συνεπώς, διατηρήθηκαν οι αλληλουχίες που εκπληρούσαν τα παρακάτω κριτήρια:

1. Hairpin = 1
2. Bulges ≤ 16
3. Βρόγχοι ≤ 32
4. Ασυμμετρία ≤ 13
5. Bulges-Loops ≤ 37
6. Hairpin Length ≤ 16
7. Folding min energy ≤ -25.44kcal/mol
8. Συντήρηση ≥ 25% των νουκλεοτιδίων συντηρημένα

Από την προαναφερόμενη διαδικασία φιλτραρίσματος προέκυψαν 249 πραγματικά miRNA και 2330 αρνητικές αλληλουχίες. Συνεπώς, τα ΚΜΜ μοντέλα εκπαιδεύτηκαν μονάχα στα πραγματικά miRNA χρησιμοποιώντας μια διαδικασία 5-fold (3/5 για εκμάθηση, 2/5 για επαλήθευση) boosting validation (όπως περιγράφεται στα Υλικά και Μέθοδοι). Η διαδικασία επαναλήφθηκε για διαφόρους συνδυασμούς βιολογικών γνωρισμάτων και υπολογίστηκε ο μέσος όρος των ΚΜΜ για το πόσο ακριβής είναι η επίδοση της κάθε περίπτωσης. Στο Σχήμα 3.6 απεικονίζονται οι καμπύλες ROC, οι οποίες παρουσιάζουν τον μέσο όρο της επίδοσης επαλήθευσης όπως αποδίδεται από τα ΚΜΜ, εκπαιδευμένα με όλους τους πιθανούς συνδυασμούς βιολογικών γνωρισμάτων (αλληλουχία, δομή και συντήρηση).Υπήρξε μια σημαντική βελτίωση στην ακρίβεια της πρόβλεψης για το σύνολο επαλήθευσης καθώς αυξανόταν ο αριθμός των γνωρισμάτων (σχήμα 3.6), υποδεικνύοντας τη σημασία της ταυτόχρονης ενσωμάτωσης επιπρόσθετων βιολογικών πληροφοριών κατά τη διαδικασία εκμάθησης.

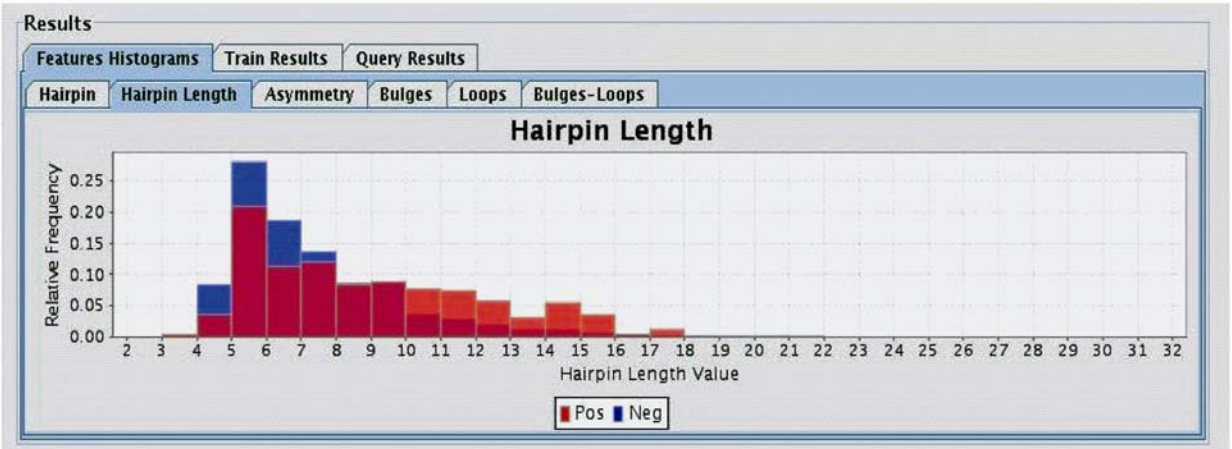
Όπως απεικονίζεται στο σχήμα 3.5C (αριστερά) και 3.6 (καμπύλη SeStCo), τα καλύτερα αποτελέσματα προέκυψαν όταν χρησιμοποιήθηκαν και τα τρία βιολογικά γνωρίσματα στην εκπαίδευση των ΚΜΜ, επιτυγχάνοντας μέσο όρο ευαισθησίας 88.95% και ειδικότητας 84.16% στο σύνολο επαλήθευσης, για ένα κατώφλι τιμής 3. Εφόσον επιτεύχθηκε μια καλή επίδοση στα σύνολα εκπαίδευσης/επαλήθευσης, συλλέχθησαν και τα 249 πρόδρομα πραγματικά miRNA και χρησιμοποιήθηκαν στην εκμάθηση ενός τελικού ΚΜΜ μοντέλου λαμβάνοντας υπόψη τον ίδιο συνδυασμό βιολογικών γνωρισμάτων και παραμέτρων φιλτραρίσματος. Το τελικό μοντέλο χρησιμοποιήθηκε για την εφαρμογή της διεπαφής του *SSCprofiler* που πραγματοποιεί τη σάρωση γενωμικών περιοχών.

Για να αποδείξουμε την ικανότητα γενίκευσης του *SSCprofiler* σε δεδομένα τα οποία δεν έχουν χρησιμοποιηθεί στην διαδικασία εκπαίδευσης/επαλήθευσης, χρησιμοποιήσαμε 373 miRNA τα οποία έχουν προστεθεί πρόσφατα στην miRBase έκδοση 12. Από τα παραπάνω, 329 πρόδρομα πέρασαν τα φίλτρα του *SSCprofiler* και 219 είχαν άνω του 25% των νουκλεοτιδίων τους συντηρημένα σύμφωνα με τα multiz full genome alignments. Ο πίνακας 3.3 αναπαριστά την επίδοση της ταξινόμησης που επιτεύχθηκε από τον *SSCprofiler* στα 249 δεδομένα εκμάθησης/επαλήθευση και στα 219 νέα πρόδρομα miRNA για διαφορετικά κατώφλια των ΚΜΜ. Η ταξινόμηση των 219 νέων προδρόμων εκτελέστηκε χρησιμοποιώντας την διεπαφή σάρωσης του *SSCprofiler* και θεωρήθηκε ότι τα πρόδρομα miRNA ταυτοποιούνται εφόσον υπάρχει μια θετική πρόβλεψη στις αντίστοιχες γενωμικές τους περιοχές. Για το λόγο αυτό αναφέρουμε μόνο την ακρίβεια πρόβλεψης για αυτό το σύνολο δεδομένων. Όπως φαίνεται στον πίνακα 3.3 η επίδοση γενίκευσης είναι μέγιστη για το κατώφλι 1 των ΚΜΜ.

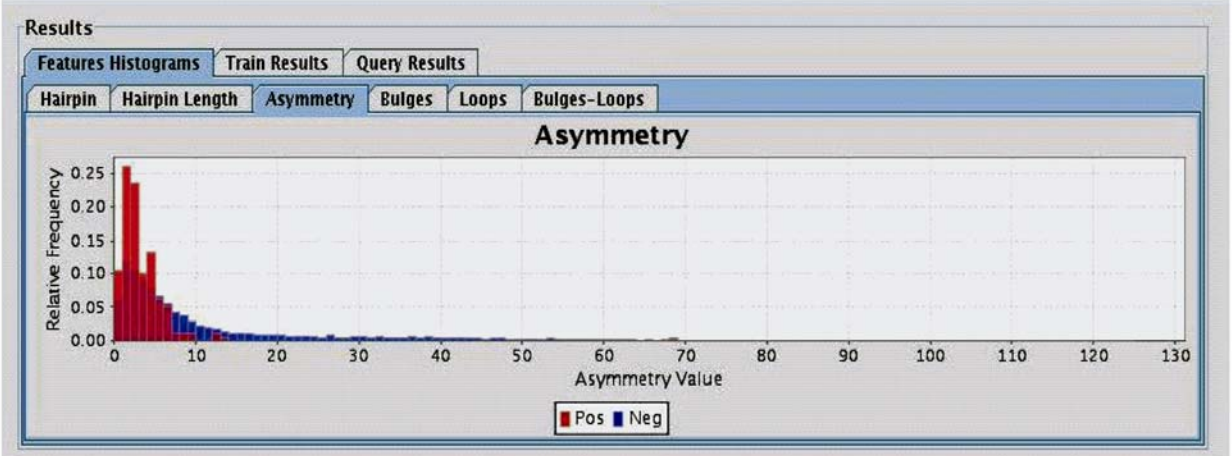
Πίνακας 3.3 Τιμές ακρίβειας πρόβλεψης για τα δεδομένα επαλήθευσης (Validation) και δοκιμής (Test) για 3 βέλτιστα κατώφλια του SSCprofiler.

Threshold	Validation		Test
	Sensitivity	Specificity	Pred. Acc
3	88.95	84.16	72.15
2	90.07	81.77	78.08
1	91.3	78.9	85.84

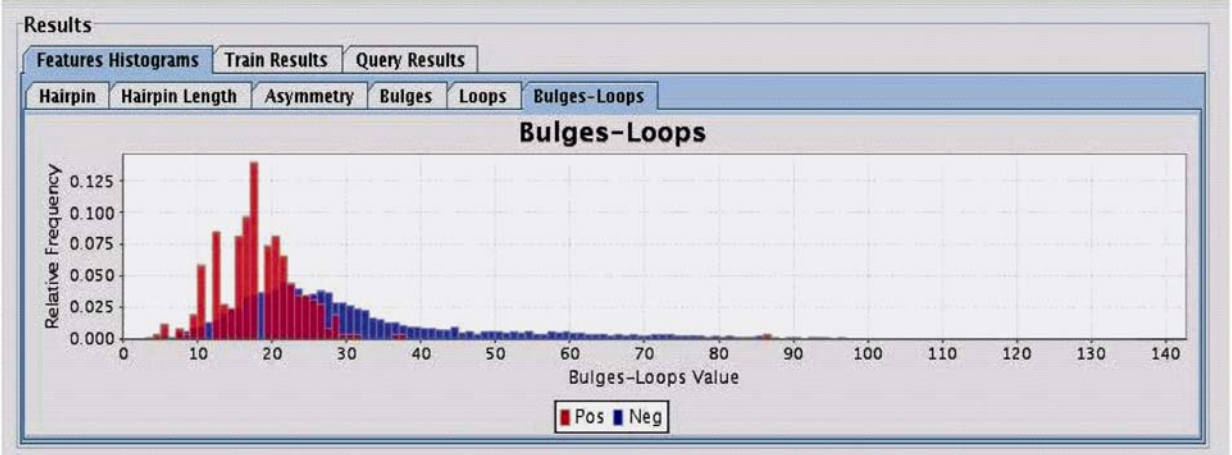
a



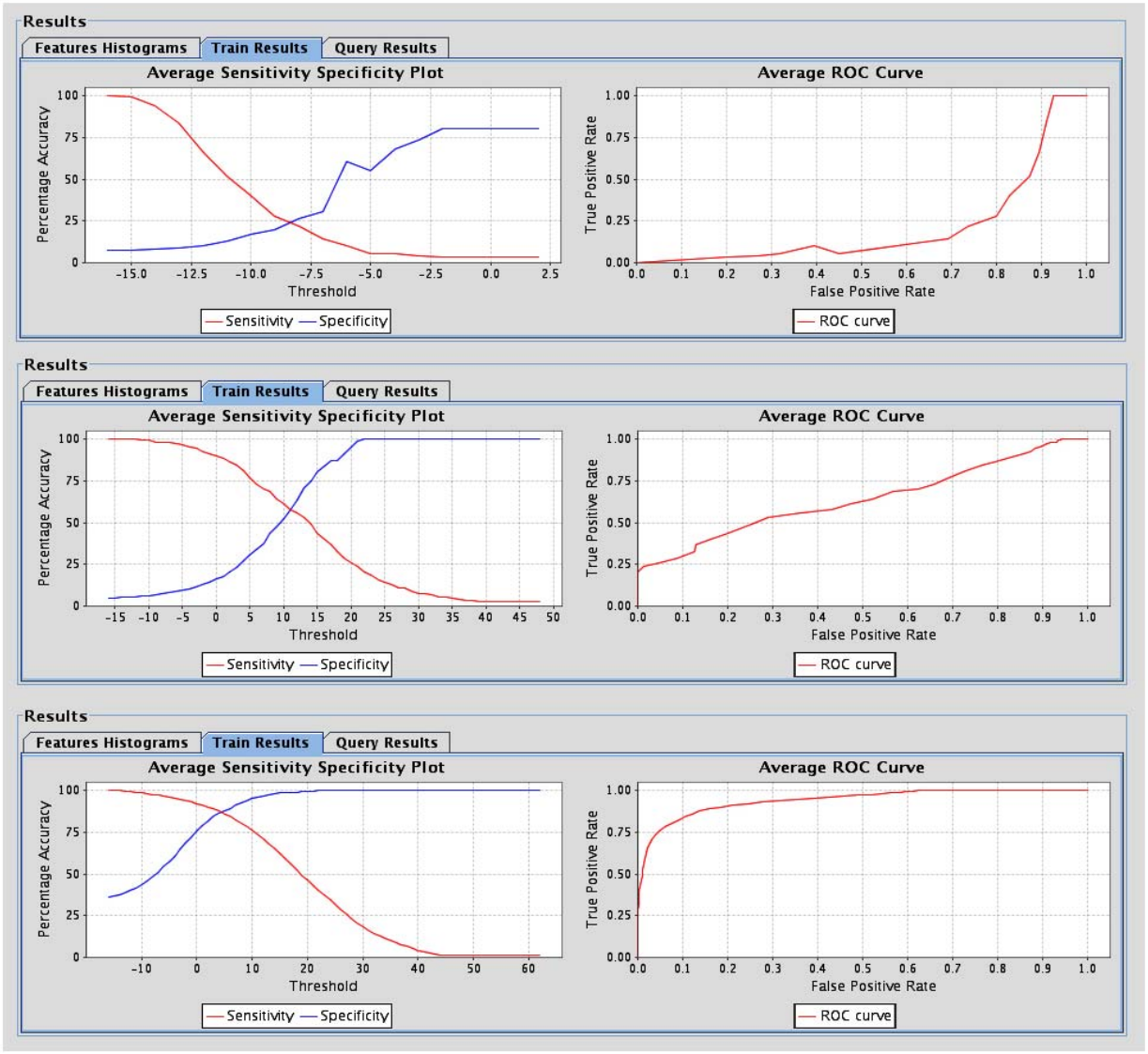
b



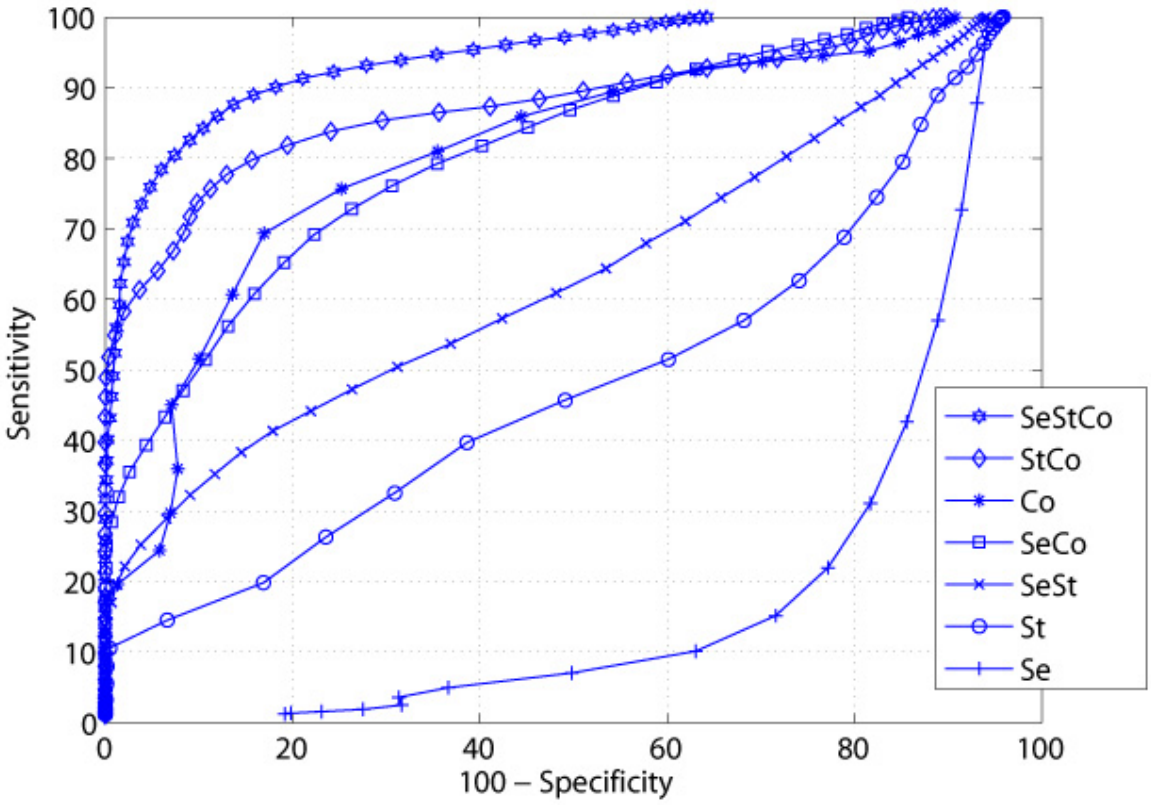
c



Σχήμα 3.4 Ιστογράμματα κατανομής των ανθρώπινων miRNA (Κόκκινο-Pos) και αρνητικές αλληλουχίες (Μπλε- Neg) όπως καθορίζονται από τον *SSCprofiler*. Απεικονίζονται τρεις από τις επτά παραμέτρους φιλτραρίσματος: (a) Hairpin Length, (b) Asymmetry και (c) Bulges-Loops count. Κοιτάζοντας τις κατανομές των θετικών και αρνητικών δεδομένων, δίνεται η δυνατότητα στο χρήστη να επιλέξει κατώφλια για τη βέλτιστη διαφοροποίηση των δυο κατανομών, έτσι ώστε να χρησιμοποιηθούν για το φιλτράρισμα των δεδομένων.



Σχήμα 3.5 Γραφήματα ευαισθησίας-ειδικότητας (αριστερά) και ROC καμπύλες (δεξιά) όπως καθορίζονται από τον SSCprofiler για τα ανθρώπινα miRNA και τις αρνητικές αλληλουχίες. Τα γραφήματα δείχνουν την θετική επίδραση της ενσωμάτωσης επιπλέον βιολογικής πληροφορίας στην διαδικασία εκμάθησης για την επίτευξη της καλύτερης ακρίβειας των ΚΜΜ. Αριστερά: Στον χ-άξονα απεικονίζεται η τιμή των κατωφλιών των ΚΜΜ και στον y-άξονα ο μέσος όρος ευαισθησίας (κόκκινο) και ειδικότητας (μπλε) όλων των μοντέλων εκμάθησης ΚΜΜ για 100 γύρους επαλήθευσης. Υπάρχει μια εξαιρετική βελτίωση στα ποσοστά ακρίβειας με την προσθήκη επιπλέον πληροφορίας ξεκινώντας απλά με την αλληλουχία (a) και καταλήγοντας σε αλληλουχία, δομή και συντήρηση (c).



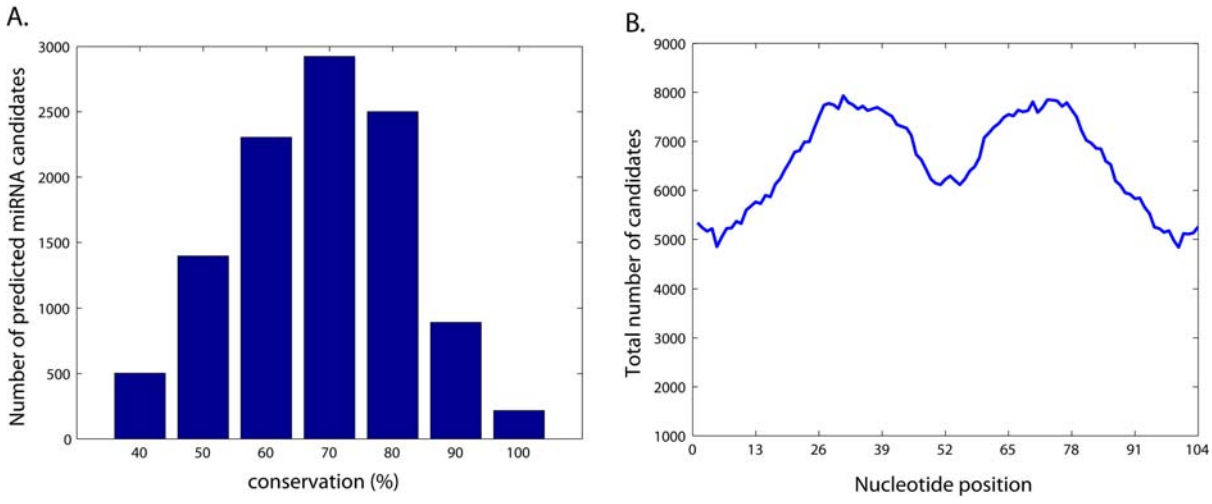
Σχήμα 3.6 Καμπύλες ROC, οι οποίες παρουσιάζουν τον μέσος όρος της επίδοσης επαλήθευσης όπως αποδίδεται από τα ΚΜΜ, εκπαιδευμένα με όλους τους πιθανούς συνδυασμούς βιολογικών γνωρισμάτων (αλληλουχία (Se), δομή (St) και συντήρηση (Co)). Όπως φαίνεται από το σχήμα, η επιφάνεια κάτω από την καμπύλη (Area Under Curve - AUC) βελτιστοποιείται όταν χρησιμοποιηθούν και τα τρία γνωρίσματα στην εκπαίδευση των ΚΜΜ (καμπύλη SeStCo).

3.3.2 Πρόβλεψη νέων miRNA Γονιδίων σε Καρκινικά Σχετιζόμενες Γενωμικές Περιοχές (CAGR).

Σύμφωνα με την πρόσφατη μελέτη των *Calin et al* (47), υπάρχει μια μεγάλη πιθανότητα οι καρκινικά σχετιζόμενες γενωμικές περιοχές να περιέχουν miRNA γονίδια. Η υπόθεση αυτή βασίζεται στην ανακάλυψη ότι 98 γνωστά miRNA γονίδια βρίσκονται μέσα σε CAGR. Σε αυτά συμπεριλαμβάνονται 80 miRNA των οποίων η θέση βρίσκεται σε περιοχές όπου χάνεται η ετεροζυγότητα (LOH) ή σε περιοχές που

παρουσιάζουν ενίσχυση και αναφέρονται σε μια ποικιλία όγκων όπως του μαστού, του πνεύμονα, των ωοθηκών, του παχέος εντέρου, του γαστρικού και του ηπατοκυτταρικού καρκινώματος, καθώς και σε περιπτώσεις λευχαιμίας και λεμφωμάτων. Προκειμένου να διερευνηθεί η παραπάνω υπόθεση, χρησιμοποιήσαμε το τελικό εκπαιδευμένο μοντέλο του *SSCprofiler* για την εύρεση νέων υποψήφιων miRNA γονιδίων σε αυτές τις περιοχές. Σαρώσαμε τη θετική και αρνητική αλυσίδα του DNA για έναν αριθμό περιοχών που αντιπροσωπεύουν πάνω από 350MB του ανθρώπινου γονιδιώματος. Οι περιοχές αυτές χαρακτηρίζονται από εξάλειψη ή αύξηση (εξαλειφόμενες ή αυξητικές περιοχές) σε περισσότερους από 20 διαφορετικούς τύπους καρκίνου. Χρησιμοποιήθηκαν οι προαναφερόμενοι παράμετροι φιλτραρίσματος καθώς και η πληροφορία σχετικά με τη συντήρηση της κάθε περιοχής.

Η διαδικασία εύρεσης νέων miRNA (Υλικά και Μέθοδοι, Σχήμα 3.3) διήρκησε περίπου 8 μέρες (πραγματικός χρόνος) χρησιμοποιώντας ένα παράλληλο PC cluster με 10 επεξεργαστές dual opteron. Ένα παράδειγμα των αποτελεσμάτων της σάρωσης όπως απεικονίζεται από τον *SSCprofiler* δίνεται στο σχήμα 3.8. Το Σχήμα 3.7 απεικονίζει την συντήρηση για όλα υποψήφια miRNA τα οποία απέδωσαν βαθμολογία του KMM υψηλότερη του 3. Όπως φαίνεται στο Σχήμα 3.7A τα περισσότερα προβλεπόμενα υποψήφια miRNA είχαν πάνω από 50% των νουκλεοτιδίων τους συντηρημένα στους 7 διαφορετικούς οργανισμούς που εξετάσαμε. Επιπλέων, η συντήρηση για κάθε ένα από τα 104nt των υποψήφιων αλληλουχιών, παρουσίασε σημαντική μείωση στην περιοχή της φουρκέτας (loop) (Σχήμα 3.7B)



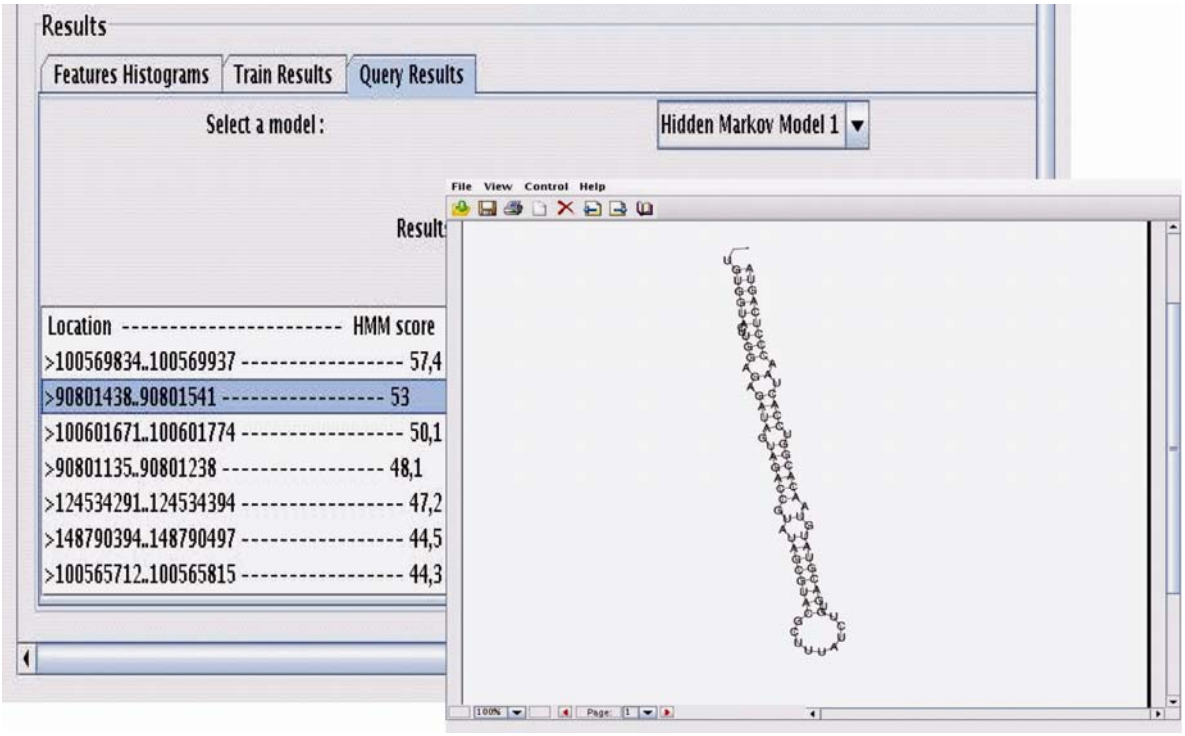
Σχήμα 3.7 Συντήρηση για τα 10511 υποψήφια miRNA (κατώφλι ΚΜΜ >=3) για 7 διαφορετικούς οργανισμούς. Α. Ιστόγραμμα συντηρημένων υποψήφιων miRNA όπως φαίνεται τα περισσότερα προβλεπόμενα υποψήφια miRNA είχαν πάνω από 50% των νουκλεοτιδίων τους συντηρημένα. Β. Η κατανομή συντηρημένων νουκλεοτιδίων για όλα τα υποψήφια miRNA για κάθε ένα από τα 104nt. Όπως είναι προφανές από το διάγραμμα, οι πλειοψηφία των υποψήφιων miRNA παρουσίασε σημαντική μείωση στην περιοχή της φουρκέτας (loop), ενώ οι περιοχές εκατέρωθεν της φουρκέτας (περιοχές όπου βρίσκεται το ώριμο miRNA) παρουσιάζουν μια συμμετρική συντήρηση υψηλού επιπέδου.

Η ταυτοποίηση των 98 γνωστών miRNA γονιδίων σε αυτές τις περιοχές αξιολογήθηκε ως συνάρτηση της βαθμολογίας που αποδίδει το ΚΜΜ καθώς φαίνεται στον πίνακα 3.4. Όπως αναμενόταν, ο αριθμός των υποψήφιων miRNA γονιδίων μειώνεται με την αύξηση της βαθμολογίας του ΚΜΜ. Συνεπώς, καθώς η βαθμολογία του ΚΜΜ αυξάνεται, η ευαισθησία μειώνεται ενώ η ειδικότητα αυξάνεται. Σύμφωνα με τις μεθόδους εκμάθησης και δοκιμής που προαναφέρθηκαν, το βέλτιστο κατώφλι της βαθμολογίας του ΚΜΜ κυμαίνεται μεταξύ 1-3 (Πίνακας 3.3). Παρόλα αυτά ακόμα και για μια μέση τιμή ειδικότητας της τάξεως του ~85% (κατώφλι 3) όταν σαρώνονται μεγάλες γενωμικές αλληλουχίες, είναι πιθανόν να συσσωρεύονται πολλαπλά μη πραγματικά θετικά. Καθώς η πειραματική επιβεβαίωση των νέων υποψήφιων γονιδίων είναι μια δαπανηρή και χρονοβόρα διαδικασία επιλέγουμε υποψήφια γονίδια που έχουν μία σημαντικά υψηλότερη βαθμολογία του ΚΜΜ, προκειμένου να αποκομίσουμε τα πιο πιθανά miRNA υποψήφια γονίδια.

Πίνακας 3.4 Προβλεπόμενα miRNA γονίδια ως συνάρτηση της βαθμολογίας που αποδίδεται από το ΚΜΜ (HMM score). Ο πίνακας δείχνει των αριθμό των προβλεπόμενων πρόδρομων miRNA (δεύτερη στήλη) για κάθε δεδομένη βαθμολογία του ΚΜΜ εύρους 3-41 (πρώτη στήλη). Ο αριθμός των αληθινών προδρόμων miRNA που περιλαμβάνεται στην προβλεπόμενη λίστα φαίνεται επίσης σαν συνάρτηση όλων των αληθινών προδρόμων miRNA μέσα στα CAGR (τρίτη στήλη). Ο αριθμός των υποψήφιων και πραγματικών miRNA γονιδίων τα όποια ξεπέρασαν το κατώφλι έκφρασης 200 σε κύτταρα HeLa και/ή HepG2, παρουσιάζεται στην τέταρτη και πέμπτη στήλη αντίστοιχα. Οι 421 αλληλουχίες των προβλεπόμενων υποψήφιων

miRNA για ένα κατώφλι 21 του ΚΜΜ επιλέχθηκαν για περεταίρω επεξεργασία (φαίνεται με γκρί χρώμα στον πίνακα).

HMM score (>=)	Candidate precursors	Identified miRNA precursors /Total miRNA precursors in CAGRs	Candidates exceeding expression threshold (>200) in HeLa and/or HepG2	True miRNA exceeding expression threshold (>200) in HeLa and/or HepG2	Sensitivity/ specificity according to 5-fold boosting validation
3	10511	98/98	1229	45	88.95/ 84.16
5	9947	97/98	1171	44	85.96/ 88.02
7	8866	97/98	1017	43	82.56/ 90.96
9	7450	95/98	872	42	78.41/ 93.99
11	5862	94/98	667	41	73.44/ 96.13
13	3906	87/98	439	38	68.18/ 97.60
15	2498	82/98	290	36	62.26/ 98.39
17	1467	75/98	154	32	56.00/ 98.78
19	819	64/98	85	29	49.15/ 99.15
21	421	64/98	38	28	43.18/ 99.48
23	230	60/98	17	26	37.18/ 99.82
25	116	52/98	8	22	31.66/ 99.96
27	62	45/98	3	22	25.85/ 100.00
29	31	40/98	0	20	20.66/ 100.00
31	16	37/98	0	15	16.39/ 100.00
33	12	29/98	0	13	12.89/ 100.00
35	4	28/98	0	13	10.13/ 100.00
37	0	21/98	0	11	7.56/ 100.00
39	0	14/98	0	8	5.09/ 100.00
41	0	9/98	0	4	3.05/ 100.00



Σχήμα 3.8 Τα αποτελέσματα σάρωσης όπως παρουσιάζονται από τον SSCprofiler. Τα αποτελέσματα δείχνουν τη γενωμική τοποθεσία της περιοχής που έχει σαρωθεί καθώς επίσης και την βαθμολογία του KMM (HMM score). Τα υποψήφια γονίδια κατατάσσονται κατά φθίνουσα σειρά με βάση την βαθμολογία του KMM και με το εργαλείο γίνεται δυνατή η απεικόνιση της δευτεροταγούς δομής του υποψήφιου miRNA γονιδίου.

3.3.3 Πειραματική Επιβεβαίωση των πιο Υψηλόβαθμων Υποψηφίων

Σύμφωνα με τις μετρήσεις ευαισθησίας και ειδικότητας, το *SSCprofiler* επιτυγχάνει πολύ υψηλή ακρίβεια πρόβλεψης κατ' αναλογία με τον προσδιορισμό νέων miRNA γονιδίων. Παρόλα αυτά, τα στατιστικά κριτήρια αξιολόγησης εξαρτώνται ισχυρά από τα δεδομένα που χρησιμοποιούνται για την εκμάθηση και την επαλήθευση του υπολογιστικού μοντέλου. Η πειραματική επιβεβαίωση των υποψήφιων miRNA γονιδίων που έχουν προσδιοριστεί, αποτελεί τη βέλτιστη μέθοδο για την αξιολόγηση της ακρίβειας του μοντέλου. Για το σκοπό αυτό, επόμενο βήμα ήταν η πειραματική επιβεβαίωση μερικών από τα πιο υψηλόβαθμα υποψήφια γονίδια, δηλαδή αυτά με το ψηλότερο KMM score. Το κατώφλι του KMM για τα γονίδια αυτά επιλέχθηκε

σύμφωνα με τα παρακάτω κριτήρια: (α) αρκετά υψηλό ώστε να μειώνει τον αριθμό των μη-αληθινών θετικών και συγχρόνως (β) αρκετά χαμηλό ώστε να συμπεριλάβει ένα μεγάλο αριθμό από τα αληθινά miRNA.

Θεωρήσαμε τελικά το 21[i] ως το κατώφλι που πληρεί τα παραπάνω κριτήρια και στο οποίο 421 υποψήφια γονίδια προβλέφθηκαν με 65.31% (64/98) ακρίβεια για τα αληθινά miRNA. Στο κατώφλι αυτό η λίστα των υποψήφιων γονιδίων περιλαμβάνει γονίδια με μερική μόνο συντήρηση συγκρινόμενα με πίο υψηλόβαθμα γονίδια. Η έκφραση και των 421 υποψηφίων miRNA σε κύτταρα HeLa αξιολογήθηκε χρησιμοποιώντας πρόσφατα δεδομένα από ένα tiling array επί του συνολικού ανθρώπινου γονιδιώματος (140), το οποίο παρέχει ένα χάρτη έκφρασης για μικρά RNA. Τα υποψήφια γονίδια με έκφραση πάνω από 200 (ν=38) στην περιοχή του stem, ήταν εκείνα που επιλέχθηκαν για περεταίρω ανάλυση. Από αυτά, τα τέσσερα γονίδια με την υψηλότερη τιμή έκφρασης (πίνακας 3.5), εξετάστηκαν πειραματικά με την μέθοδο northern blot σε κυτταρική καλλιέργεια HeLa (βλ. Υλικά και Μέθοδοι). Τα αποτελέσματα από τα northern blot συμφωνούν με δεδομένα έκφρασης από το tiling array για τα τέσσερα υποψήφια γονίδια που εξετάστηκαν. Όπως φαίνεται στο σχήμα 3.9, όλα τα υποψήφια γονίδια παρείχαν ισχυρά σήματα (bands) μέσα στο εύρος 19-27nt, ενδεικτικά για τα ώριμα miRNA, ενώ σε κάποιες περιπτώσεις ανιχνεύτηκε και το πρόδρομο miRNA. Επιπρόσθετα, βρέθηκε ότι και στα τέσσερα υποψήφια γονίδια μόνο μια αλυσίδα (strand) του πρόδρομο miRNA έδωσε ειδικό σήμα, γεγονός που ενίσχυσε την υπόθεση μας ότι πρόκειται για αληθινά πρόδρομα miRNA.

Η ανάλυση Blat (147) για τα τέσσερα υποψήφια γονίδια που μελετάμε σε σχέση με το ανθρώπινο γονιδίωμα παρείχε επιπρόσθετες ενδείξεις. Βρέθηκε ότι τα γονίδια αυτά είναι συντηρημένα σε ποσοστό μεγαλύτερο το 45% σε οκτώ άλλους οργανισμούς και βρίσκονται μέσα σε εκφραζόμενες δια-γονιδιακές (intergenic) περιοχές, όπως παρατηρείται στη πλειοψηφία των miRNA (148,149). Επιπλέον, κατά την αναζήτηση Blat υπήρξε 100% αντιστοιχία στο επίπεδο των 20-26nt σε άλλες περιοχές του γονιδιώματος και οι προβλεπόμενες δευτεροταγείς δομές για τις αλληλουχίες εκατέρωθεν των περιοχών αυτών, παρουσίαζουν την δομή φουρκέτας χαρακτηριστική των πρόδρομων miRNA σε 5/9 περιπτώσεις, προτείνοντας την ύπαρξη και άλλων ομόλογων miRNA γονιδίων.

[i] Η τιμή αυτή ορίστηκε κατά σύμβαση.

Πίνακας 3.5 Υποψήφια γονίδια miRNA επιβεβαιωμένα με ανάλυση northern blot τα οποία εντοπίζονται σε περιοχές εξάλειψης οι οποίες εμπλέκονται σε διάφορα ειδή καρκίνων.

Candidate	Candidate Information[j]	CAGR	Type of Cancer	Closest miRNA	Expression in HeLa
1	chr9:123327358-123327460 st-	chr9:121153509-128793509	bladder ca	miR-181a; miR-199b	1667.5
2	chr5:148958951-148959053 st-	chr5:144121683-156051683	prostate ca aggressiveness	miR145/ miR-143	363.5
2	chr5:148958951-148959053 st-	chr5:148181683-151101683	myelodysplastic syndrome	miR145/ miR-143	363.5
3	chr22:40863894-40863996 st+	chr22:31530000-43583971	colorectal ca,	miR-33a	345.0
3	chr22:40863894-40863996 st+	chr22:31530000-42193557	astrocytomas	miR-33a	345.0
4	chr5:149984684-149984786 st-	chr5:144121683-156051683	prostate ca aggressiveness	miR145/ miR-143	264.0
4	chr5:149984684-149984786 st-	chr5:148181683-151101683	myelodysplastic syndrome	miR145 /miR-143	264.0

3.3.4 Σύγκριση Εργαλείων

Τελικά προκειμένου να εξετάσουμε την ακρίβεια πρόβλεψης του εργαλείου μας σε σχέση με άλλους υπάρχοντες αλγόριθμους, χρησιμοποιήσαμε τα τέσσερα υποψήφια γονίδια ως είσοδο σε τρία υπάρχοντα εργαλεία πρόβλεψης miRNA γονιδίων. Από αυτά, τα MiRRim (138) και ProMir II (150) (αλγόριθμοι KMM) και το BayesMiRNAfind (73) (ταξινομητής Bayes) απέτυχαν στον εντοπισμό των γονιδίων, ενώ το TripletSVM (136) (ταξινομητής SVM) πέτυχε να προβλέψει 1 από τα 4 γονίδια (candidate 1). Είναι σημαντικό να σημειώσουμε ότι όλα αυτά τα εργαλεία θεωρούνται υψηλής ακρίβειας σύμφωνα με μετρήσεις ευαισθησίας/ειδικότητας (77), οι οποίες σε ορισμένες περιπτώσεις υπερβαίνουν αυτές του *SSCprofiler*.

[j] Chromosomal location and strand (st+ or st-)

Το πιο πρόσφατο από τα εργαλεία αυτά, το MiRRim (138) παρουσιάζει παρόμοια χαρακτηριστικά με το *SSCprofiler* καθώς χρησιμοποιεί έναν αλγόριθμο KMM που λαμβάνει υπόψη γνωρίσματα δομής και συντήρησης για την πρόβλεψη νέων υποψήφιων miRNA γονιδίων. Μια λεπτομερής σύγκριση μεταξύ των δύο εργαλείων δίνεται στον Πίνακα 3.6.

Πίνακας 3.6. Σύγκριση μεταξύ των SSCprofiler και miRRim

	SSCprofiler	**miRRim**
Biological Features of miRNAs	Sequence, structure and conservation,	Structure and conservation,
Negative data	Selected from 3'UTRs, and filtered according to conservation and a minimum energy score to ensure that they resemble true miRNAs in both structure and conservation.	Randomly selected 200nt-long genomic regions with different degrees of conservation. No requirements for resemblance with true miRNAs.
Sensitivity/ specificity (validation set)	HMM score:3 Sens: 88.95% Spec: 84.16%	Sens: ~70% Spec: ~90%.
Generalization (blind test set)	Identification of 219 previously unseen miRNAs with an accuracy of 72.15%	No evaluation of performance on a blind test set
Scanning procedure	104nt sliding window, shifted 1nt at a time for positive as well as negative data.	Size of sliding window and shift step unclear. Positive data ranged between 160-236nt, negative data were 200nt, implying a larger window and thus a smaller search space.
Total number of hits	For a coverage of 96.0% (HMM threshold 11) ~5,800 miRNA hits for 350MB (CAGRs) of the human genome	For a coverage of 91.0%, ~4,000 miRNAs hits for the whole human genome.
Expression information using high throughput methods	Tiling array data from HeLa and HepG2 cells	No expression information is provided
Experimental verification	Successful verification of 4 top scoring miRNA candidates via northern blot	No experimental verification is provided

Όπως φαίνεται στον Πίνακα 3.6, μια εκτενής σύγκριση του *SSCprofiler* με το miRRim (138) αναδεικνύει τα πλεονεκτήματα του εργαλείου μας. Η διαδικασία που ακολουθήσαμε για την επιλογή των αρνητικών δεδομένων εξασφαλίζει ότι τόσο οι

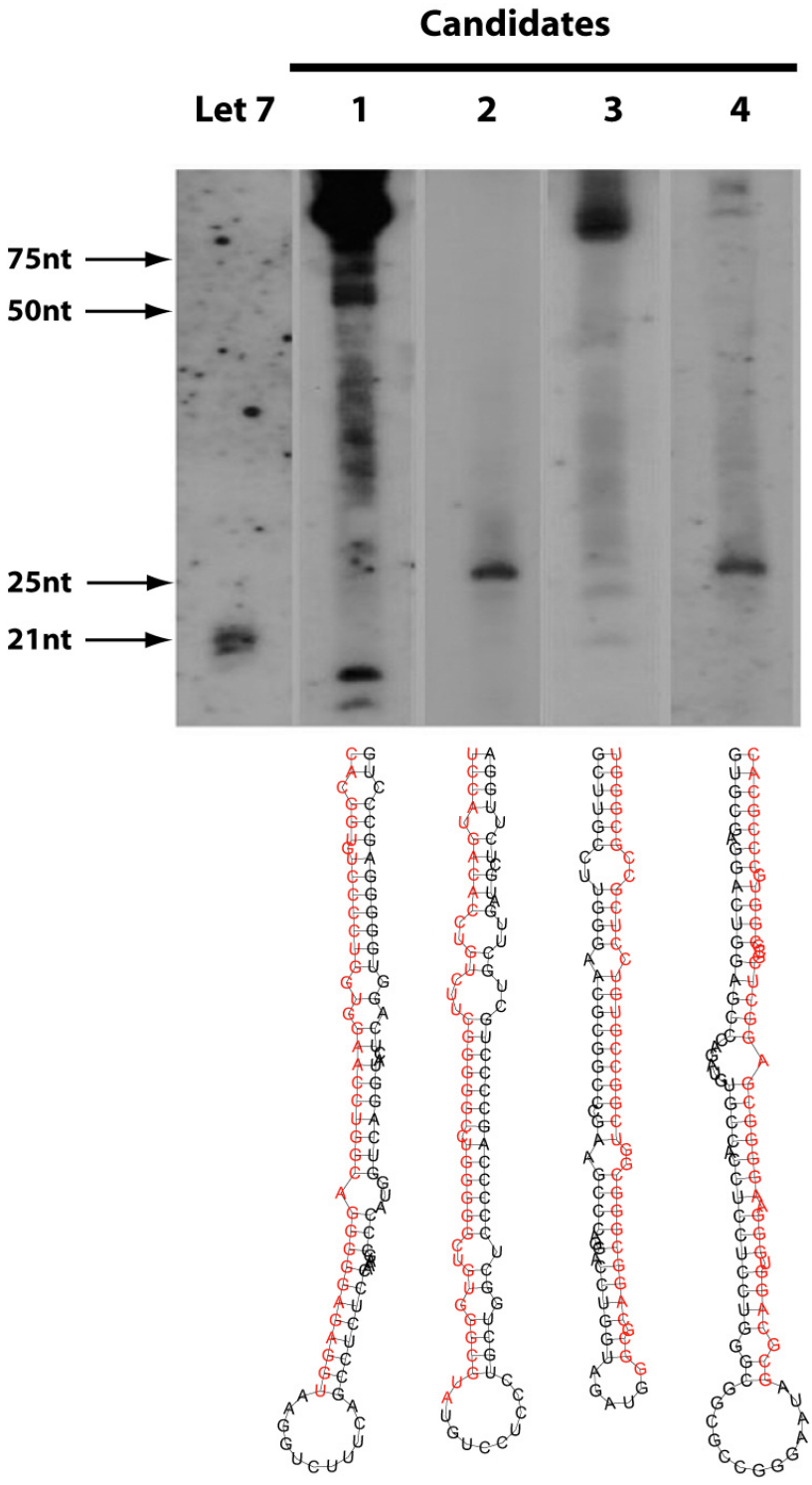
θετικές όσο και οι αρνητικές αλληλουχίες είναι όμοιες ως προς τη δομή και το βαθμό συντήρησής τους. Αποκλειστική χρήση δεδομένων συντήρησης, όπως έγινε στη μελέτη των *Terai et al* (138) για το miRRim, μπορεί να προϊδεάσει τα αποτελέσματα και να κάνει ευκολότερη τη διάκριση μεταξύ των δύο κατηγοριών, με αποτέλεσμα οι μετρήσεις ευαισθησίας/ειδικότητας να είναι υψηλότερες για τα σύνολα εκπαίδευσης και επαλήθευσης. Το miRRim δεν κάνει χρήση ενός συνόλου δοκιμής για να επιβεβαιώσει την ακρίβεια γενίκευσης (generalization performance), οπότε δεν είναι σαφές κατά πόσο η ακρίβεια πρόβλεψης κατά τη διαδικασία σάρωσης γενωμικών περιοχών είναι εξίσου υψηλή με την επίδοση επαλήθευσης.

Μια βασική διαφορά μεταξύ των δύο εργαλείων είναι ο αριθμός των προβλεπόμενων υποψήφιων miRNA γονιδίων. Το *SSCprofiler* δεν χρησιμοποιήθηκε για τη σάρωση ολόκληρου του γονιδιώματος, αλλά κατ' αναλογία με τις προβλέψεις για τις CAGR, για ποσοστό ευαισθησίας 96.0%, υποθέτουμε ότι θα προέβλεπε ~58,000[k] υποψήφια miRNA (κατώφλι KMM, 11) σε αντίθεση με τα ~4,000 που προβλέπει το miRRim. Η ασυμφωνία αυτή μπορεί να αποδοθεί σε διαφορές στη διαδικασία σάρωσης που ακολουθήθηκε από κάθε εργαλείο. Υπάρχει ασάφεια στην διαδικασία που χρησιμοποιήθηκε από το miRRim για τη σάρωση του γονιδιώματος (1nt ή 50nts), αλλά δεομένου ότι το μέγεθος των δεδομένων εκπαίδευσης κυμαίνεται μεταξύ 160-236nt, το παράθυρο σάρωσης θα πρέπει να είναι σημαντικά μεγαλύτερο από το παράθυρο σάρωσης των 104nt, που χρησιμοποιήθηκε από το *SSCprofiler*. Αυτό έχει ως αποτέλεσμα ένα σημαντικά μικρότερο χώρο αναζήτησης και επακόλουθα ένα μικρότερο αριθμό γενωμικών περιοχών που θα μπορούσαν να προσδιοριστούν ως πιθανά miRNA γονίδια. Μια πρόσφατη εκτίμηση του συνολικού αριθμού γονιδίων miRNA στο ανθρώπινο γονιδίωμα που πραγματοποιήθηκε από τους *Miranda et al* (151) τα προσδιορίζει σε ~55,000[k]. Ο αριθμός αυτός είναι κατά πολύ μεγαλύτερος από τα 4,000 γονίδια που προβλέπονται από το miRRim ενώ είναι αντίστοιχος με τα ~58,000 που προβλέπονται από το *SSCprofiler*. Επιπλέον, δεδομένα από μελέτες tilling arrays και deep sequencing (140,141) δείχνουν ότι ο αριθμός των miRNA στο ανθρώπινο γονιδίωμα είναι πιθανών πολύ μεγαλύτερος από τη συντηρητική εκτίμηση των 4,000. Τέλος, πρέπει να σημειωθεί ότι τα CAGR είναι γνωστό ότι περιέχουν ένα

[k] Σημείωση ότι πολλά από αυτά τα υποψήφια miRNA πιθανόν να μην είναι πραγματικά miRNA.

δυσανάλογα μεγάλο αριθμό miRNA (20% του συνόλου των miRNA από τη miRBase 12.0), οπότε μια αναλογία μεταξύ των περιοχών αυτών και του συνολικού ανθρώπινου γονιδιώματος δεν είναι έγκυρη.

Ακόμα ένα πλεονέκτημα του *SSCprofiler* είναι ότι επιτρέπει στο χρήστη να φιλτράρει ακόμα περισσότερο τα υποψήφια γονίδια χρησιμοποιώντας ένα σύνολο δεδομένων έκφρασης από tiling array (140). Η δυνατότητα αυτή δεν διατίθεται στο miRRim. Τέλος, κανένα από τα υποψήφια γονίδια που προβλέπονται από το miRRim δεν επιβεβαιώθηκε πειραματικά, ενώ και τα τέσσερα υποψήφια γονίδια που προβλέπονται από το *SSCprofiler* εμφάνισαν σήμα στο εύρος της μπάντας miRNA κατά τη ανάλυση Northern blot. Πιστεύουμε ότι προκειμένου μια μελέτη υπολογιστικής πρόβλεψης miRNA να αποδείξει την αξιοπιστία της, ιδιαίτερα όταν προορίζεται για τη βιολογική κοινότητα, είναι απαραίτητο να προσφέρει μια ολοκληρωμένη διαδικασία ανάλυσης που ξεκινάει από την υπολογιστική πρόβλεψη και καταλήγει στην πειραματική επιβεβαίωση για τουλάχιστον τα υποψήφια γονίδια με την υψηλότερη βαθμολογία KMM.



Σχήμα 3.9 Η ανάλυση northern blot δίνει ένα συγκεκριμένο σήμα για κάθε ένα από τα τέσσερα υποψήφια miRNA γονίδια. Το let7 probe υβριδοποιείται σε πολλαπλά μέλη της let7 οικογένειας miRNA γονιδίων με αποτέλεσμα να εμφανίζονται τρεις ζώνες στην μεμβράνη αναφοράς. Οι μεμβράνες 1-4 εκπροσωπούν τα υποψήφια γονίδια miRNA που προβλέπονται από το *SSCprofiler* με την ίδια σειρά που

παρουσιάζονται στον Πίνακα 3.5. Οι μπάντες των τεσσάρων miRNA γονιδίων ομοιάζουν αυτές γνωστών miRNA γονιδίων, με σήμα στην περιοχή των 19-27nt. Οι μπάντες που βρίσκονται ψηλότερα αντιπροσωπεύουν τα πρόδρομα miRNA μόρια (περιοχή ~70nt). Απεικονίζονται επίσης οι δευτεροταγείς δομές για τα τέσσερα υποψήφια miRNA γονίδια όπως προβλέπονται από το RNAfold. Η αλληλουχία από κάθε πρόδρομο μόριο που παρουσίαζει συμπληρωματικότητα με των ανιχνευτή εμφανίζεται με κόκκινο.

3.4 Συζήτηση

Στη μελέτη αυτή παρουσιάσαμε ένα αποτελεσματικό εργαλείο πρόβλεψης miRNA γονιδίων (SSCprofiler) βασισμένο σε KMM, το οποίο αξιολογήθηκε υπολογιστικά χρησιμοποιώντας ένα σύνολο πρόσφατα προσδιορισμένων γονιδίων miRNA και πειραματικά με την επιβεβαίωση τεσσάρων υποψήφιων γονιδίων με υψηλή βαθμολογία KMM. Το εργαλείο παρέχεται τόσο ως μία φιλική προς το χρήστη εκπαιδεύσιμη διεπαφή, όσο και ως μια web-based εφαρμογή σάρωσης η οποία μπορεί να χρησιμοποιηθεί για την αναζήτηση γενωμικών περιοχών. Σε αμφότερες τις περιπτώσεις ο χρήστης έχει ένα μεγάλο βαθμό ευελιξίας όσον αφορά τον προσδιορισμό των δεδομένων και τον καθορισμό των παραμέτρων φιλτραρίσματος και εκμάθησης.

Ο αλγόριθμος που αναπτύξαμε συνδυάζει πληροφορίες που αφορούν αλληλουχία, δομή και συντήρηση στο επίπεδο των νουκλεοτιδίων κατά μήκος του πρόδρομου miRNA. Δείχνουμε ότι η ενσωμάτωση πολλαπλών βιολογικών γνωρισμάτων προσφέρει σαφές πλεονέκτημα στην ακρίβεια πρόβλεψης και υποστηρίζουμε ότι η συνδιασμένη χρήση αυτών των γνωρισμάτων είναι περισσότερο αποτελεσματική από άλλες προσεγγίσεις. Η ενσωμάτωση πληροφορίας από δεδομένα έκφρασης είναι ένα εξίσου σημαντικό πλεονέκτημα που παρέχει το εργαλείο μας. Η χρήση δεδομένων tiling array του συνολικού γονιδιώματος (140), τα οποία παρέχουν ένα χάρτη έκφρασης για μικρά RNA σε κύτταρα HeLa και HepG2 σε ανάλυση 5nt, αυξάνει την αξιοπιστία των προβλέψεων του μοντέλου και είναι ιδιαίτερα χρήσιμη κατά την επιλογή υποψήφιων miRNA γονιδίων για πειραματική επιβεβαίωση.

Η αποτελεσματικότητα του *SSCprofiler* στην αναγνώριση ανθρώπινων miRNA γονιδίων αποδείχθηκε με τη χρήση ενός συνόλου τυφλής δοκιμής (blind test set) 219

πρόσφατα αναγνωρισμένων ανθρώπινων miRNA γονιδίων από την τελευταία έκδοση της βάσης miRBase (version 12). Για κατώφλι KMM 1, επιτεύχθηκε 85.84% ακρίβεια πρόβλεψης στο σύνολο τυφλής δοκιμής, παρόμοια με τις επιδόσεις ευαισθησίας και ειδικότητας στο σύνολο επαλήθευσης (91.3% και 78.9% αντίστοιχα). Η ικανότητα του εργαλείου να προσδιορίζει νέα miRNA γονίδια ερευνήθηκε επίσης σε καρκινικά σχετιζόμενες γενωμικές περιοχές μεγέθους 350 MB (47). Καθώς η πειραματική επιβεβαίωση είναι μια δαπανηρή και χρονοβόρα διαδικασία επιλέγουμε υποψήφια γονίδια που έχουν μεγαλύτερη πιθανότητα επιτυχίας. Για αυτόν τον λόγο χρησιμοποιούμε ένα υψηλότερο κατώφλι KMM (βαθμολογία 21) στο οποίο προσδιορίστηκε ένα σύνολο από 421 υποψήφια γονίδια τα οποία ταξινομήθηκαν σύμφωνα με την έκφρασή τους σε κύτταρα HeLa. Από αυτά μόνο 20 έδειξαν υψηλή έκφραση σε HeLa. Ανάλυση northern blot στα 4 υποψήφια γονίδια με την υψηλότερη έκφραση επιβεβαίωσε την παρουσία συγκεκριμένων μορίων RNA στο εύρος 19-27nt (χαρακτηριστικό για miRNA), που δείχνει την παρουσία μικρών μη-κωδικοποιών RNA. Μελλοντικός στόχος είναι η επέκταση της πειραματικής επιβεβαίωσης σε υποψήφια γονίδια με χαμηλότερη βαθμολογία. Αύτη η ανάλυση θα μας δώσει ένα ακριβές κατώφλι για την αξιόπιστη πρόβλεψη υποψήφιων νέων miRNA γονιδίων.

Αναφορικά με τον προσδιορισμό νέων γονιδίων miRNA, τα ευρήματά μας συμφωνούν με το ενοποιημένο σύστημα υπομνηματισμού (annotation) γονιδίων miRNA (152). Το κριτήριο βιογένεσης του miRNA ικανοποιείται με την πρόβλεψη μιας πιθανής δομής φουρκέτας στο πρόδρομο μόριο που περιλαμβάνει μια αλληλουχία ~22nt στον ένα κλώνο της. Η φουρκέτα επιδεικνύει πολύ χαμηλή τιμή ελεύθερης ενέργειας, όπως προβλέφθηκε από το πρόγραμμα RNAfold και μόνο ένας κλώνος του πρόδρομου μορίου παρουσιάζει σήμα στην ανάλυση northern blot. Τα υποψήφια γονίδια δεν περιέχουν εσωτερικά loops ή μεγάλα ασύμμετρα bulges και όλα έχουν ένα εύρος ~60–100nt, χαρακτηριστικό του μήκους των πρόδρομων miRNA που έχει αναφερθεί στα ζώα. Παρατηρείται επίσης φυλογενετική συντήρηση μεγάλου ποσοστού της αλληλουχίας του πρόδρομου miRNA για τα 4 υποψήφια γονίδια στους 7 οργανισμούς που εξετάστηκαν. Τα υποψήφια γονίδια πληρούν επίσης το κριτήριο της miRNA έκφρασης. Ένα συγκεκριμένο μεταγράφημα μήκους ~22nt ανιχνεύεται με υβριδισμό σε απομονωμένο RNA με ανάλυση northern blot για τα 4 γονίδια. Επιπρόσθετα, η έκφραση μεταγραφημάτων RNA μήκους ~22nt από την ενεργό περιοχή του κλώνου των υποψήφιων γονιδίων, παρατηρείται σε κύτταρα

HeLa χρησιμοποιώντας δεδομένα από tiling arrays. Τα κριτήρια αυτά αποτελούν ισχυρή ένδειξη ότι τα 4 υποψήφια γονίδια είναι αληθινά πρόδρομα μόρια miRNA και, συνεπώς, ότι το *SSCprofiler* αποτελεί ένα αξιόπιστο και αποτελεσματικό εργαλείο για την πρόβλεψη νέων miRNA γονιδίων. Είναι ενδιαφέρον να σημειώσουμε ότι μια συγκριτική μελέτη με 4 άλλα εργαλεία πρόβλεψης miRNA γονιδίων (73,136,138,150) δείχνει ότι αυτά αδυνατούν να προβλέψουν 3 από τα 4 επιβεβαιωμένα υποψήφια γονίδια miRNA. Μόνο 1 από τα 4 εργαλεία (136) κατάφερε να προβλέψει 1 από τα 4 επιβεβαιωμένα υποψήφια γονίδια miRNA. Η σύγκριση αυτή αναδεικνύει την υπεροχή του *SSCprofile*r και επιπλέον τεκμηριώνει τη σημασία του ως εργαλείο πρόβλεψης γονιδίων miRNA. Η διαθεσιμότητα του *SSCprofiler* τόσο ως μία εκπαιδεύσιμη διεπαφή όσο και ως μια web-based εφαρμογή σάρωσης, διευκολύνει επιπλέον τη χρήση του ως μέρος της διαδικασίας πρόβλεψης νέων γονιδίων miRNA, μειώνοντας το χρόνο, το κόστος και την προσπάθεια που απαιτούνται.

Ένα σημαντικό εύρημα της παρούσας εργασίας είναι ο προσδιορισμός τεσσάρων νέων υποψηφίων miRNA γονιδίων τα οποία εντοπίζονται σε γενωμικές περιοχές που εμπλέκονται σε πολυάριθμα είδη καρκίνων (CAGR). Παρόλο που ο πειραματικός χαρακτηρισμός της λειτουργίας των ώριμων miRNA είναι ακόμα σε εκκρεμότητα, τα μόρια αυτά φαίνεται να διαδραματίζουν σημαντικό ρόλο στη ρύθμιση της καρκινογένεσης, πιθανά δρώντας ως ‘ογκογονίδια’ ή ‘ογκοκαταστολείς’. Οι CAGR που αντιστοιχούν σε κάθε υποψήφιο miRNA είναι συχνά εξαλειμμένες σε διάφορους τύπους καρκίνου (Πίνακας 3.5). Η εξάλειψη μιας περιοχής που εμπεριέχει ένα γονίδιο miRNA αποτρέπει την έκφραση του λειτουργικού miRNA. Το αποτέλεσμα είναι ότι τα γονίδια που ρυθμίζονται από αυτό το miRNA λειτουργούν ανεξέλεγκτα, μια διαδικασία η οποία μπορεί να καταλήξει σε μια αλληλουχία γεγονότων που οδηγεί σε ογκογένεση. Καθώς ο συνήθης τρόπος δράσης των ώριμων miRNA είναι η καταστολή των γονιδίων-στόχων, πιθανόν τα υποψήφια γονίδια που έχουμε προσδιορίσει να έχουν ογκοκατασταλτικό ρόλο αποτρέποντας το κύτταρο από το να προσπεράσει το σημείο καμπής πέρα από το οποίο γίνεται ογκογόνο. Παρόλο που αρκετές μελέτες υποστηρίζουν τόσο το θετικό όσο και τον αρνητικό ρυθμιστικό ρόλο των miRNA (153), η επίδραση των υποψηφίων miRNA στα γονίδια-στόχους είναι δύσκολο να προβλεφθεί.

Προγράμματα πρόβλεψης miRNA στόχων παρέχουν ένα σημείο εκκίνησης για τον προσδιορισμό πιθανών γονιδίων-στόχων για τα υποψήφια miRNA, βοηθώντας τον χαρακτηρισμό του ρόλου τους σε συγκεκριμένα είδη καρκίνου. Για το σκοπό αυτό

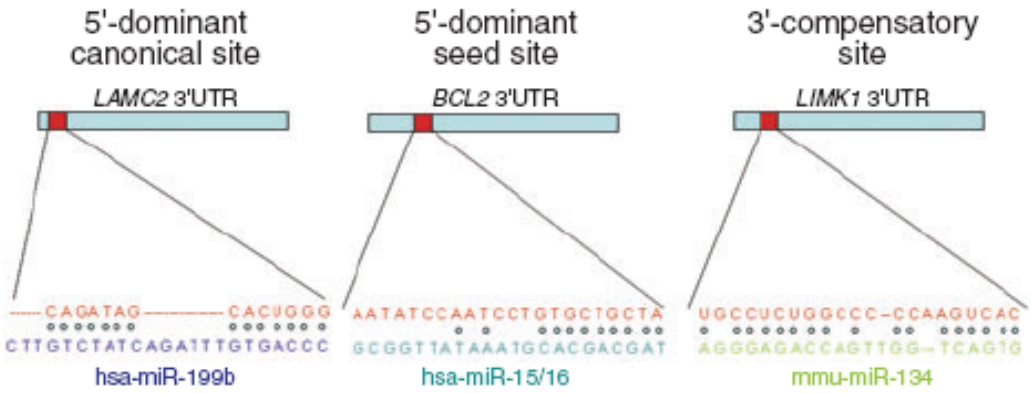
ένας αλγόριθμος πρόβλεψης διμερών RNA-RNA θα ενσωματωθεί σε μελλοντικές εκδόσεις του *SSCprofiler*. Αυτό μπορεί να επιτευχθεί με τη χρήση προγραμμάτων όπως το RNAcofold (79) το οποίο υπολογίζει δευτεροταγείς δομές από δύο RNA αλληλουχίες με τη μορφή δίκλωνου υβρίδιου. Αυτή η προσθήκη θα προσφέρει τη δυνατότητα στο χρήστη να εκπαιδεύσει profile KMM ικανά να αναγνωρίζουν τους κανόνες υβριδοποίησης μεταξύ δύο μορίων RNA, επιτρέποντας την πρόβλεψη νέων αλληλεπιδράσεων miRNA::mRNA που υπόκεινται σε παρόμοιους κανόνες. Ο τελικός μας στόχος είναι η ανάπτυξη μιας ολοκληρωμένης εφαρμογής με τη δυνατότητα πρόβλεψης νέων miRNA γονιδίων καθώς επίσης και τους πιθανούς στόχους τους. Το εργαλείο αυτό θα προσφέρει μια πιό συνοπτική βιολογική εικόνα των μονοπατιών και των γονιδίων που ρυθμίζονται από τα τέσσερα νέα υποψήφια miRNA γονίδια. Η πειραματική επιβεβαίωση είναι απαραίτητη για τον προσδιορισμό του ώριμου miRNA, την τεκμηρίωση ότι οι προβλεπόμενες αλληλεπιδράσεις λαμβάνουν χώρα στο σύστημα που μελετάμε και ότι οι λειτουργικές αλληλεπιδράσεις σχετίζονται ισχυρά με την εμφάνιση ενός καρκινικού φαινότυπου.

Κεφάλαιο 4

4 Ανάπτυξη ενός Εργαλείου Πρόβλεψης Στόχων των MicroRNA

4.1 Εισαγωγή

Όπως αναφέρθηκε στα προηγούμενα κεφάλαια, ο τρόπος δράσης του ώριμου miRNA εξαρτάται από τη συμπληρωματικότητα των βάσεων του με το 3'UTR του mRNA στόχου. Οι miRNA στοχευμένες περιοχές μπορούν να ταξινομηθούν σε τρεις κύριες κατηγορίες (78): (i) 5′-dominant canonical, (ii) 5′-dominant seed και (iii) 3′-compensatory (Σχήμα 4.1). Η περιοχή seed ορίζεται ως το συνεχόμενο διάστημα 7 νουκλεοτιδίων ξεκινώντας από το πρώτο ή το δεύτερο νουκλεοτίδιο στο 5′ άκρο ενός miRNA. Οι 5′-dominant canonical περιοχές έχουν τέλεια συμπληρωματικότητα στο 5’ άκρο (seed site) και εκτεταμένη στο 3’ακρο. Οι περιοχές 5′-dominant seed έχουν απόλυτη συμπληρωματικότητα τουλάχιστον στην περιοχή seed του 5′ άκρου του miRNA και περιορισμένη συμπληρωματικότητα στο 3′ άκρο του miRNA. Οι 3′-compensatory έχουν εκτεταμένη συμπληρωματικότητα στο 3′ άκρο του miRNA προκειμένου να αντισταθμίσουν την περιορισμένη συμπληρωματικότητα με την περιοχή seed του miRNA. Οι miRNA::mRNA αλληλεπιδράσεις χαρακτηρίζονται επίσης από μία σειρά άλλων γνωρισμάτων όπως: (α) συντήρηση, δηλαδή το ποσοστό συντήρησης στοχευμένων περιοχών στα γονιδιώματα των θηλαστικών, (β) υπολογισμούς ελεύθερης ενέργειας, δηλαδή το γεγονός ότι οι G:U ταλαντώσεις είναι λιγότερο συχνές στο 5’άκρο μιας miRNA::mRNA αλληλεπίδρασης, (γ) συνεργασιμότητα στην πρόσδεση, δηλαδή το γεγονός ότι πολλαπλά miRNA μπορούν να προσδεθούν με ένα mRNA στόχο, και ένα miRNA μπορεί να προσδεθεί με πολλές διαφορετικές στοχευμένες περιοχές mRNA (78,81,82) και (δ) τη δευτερεύουσα δομή του 3’UTR που περιβάλλει τη στοχευμένη περιοχή, η οποία έχει επίσης εμπλακεί σε αυτήν τη διαδικασία (83).



Σχήμα 4.1 Η τρεις κατηγορίες πειραματικά υποστηριζόμενων στοχευμένων περιοχών: 5′-dominant canonical (αριστερά), 5′-dominant seed (στο μέσο) και 3′-compensatory (δεξιά). Το Σχήμα προέρχεται από τη μελέτη των *Sethupathy et al 2006* (78)

Υπάρχει μεγάλος αριθμός υπολογιστικών εργαλείων για την πρόβλεψη miRNA::mRNA αλληλεπιδράσεων, τα οποία όμως υστερούν σε διάφορους τομείς (βλέπε 4.2). Η απόδοση των υπαρχόντων εργαλείων για παράδειγμα βασίζεται σε μεγάλο ποσοστό στο συνολικό αριθμό των προβλεφθέντων στόχων (προβλέψεις). Μερικά εργαλεία μπορεί να είναι πολύ αποτελεσματικά στην πρόβλεψη πραγματικών στοχευμένων περιοχών (84,85,154) (υψηλή ευαισθησία) αλλά ταυτόχρονα επιδεικνύουν έναν εξαιρετικά μεγάλο αριθμό συνολικών προβλέψεων (χαμηλή ειδικότητα). Αντίθετα, άλλα εργαλεία επιδεικνύουν συνολικά υψηλή ειδικότητα αλλά χαμηλή ευαισθησία (81,82). Συνεπώς, υπάρχει μεγάλη ανάγκη για την δημιουργία ενός πιο εξελιγμένου εργαλείου πρόβλεψης στόχων που θα πετυχαίνει μια ισορροπία μεταξύ ευαισθησίας και ειδικότητας.

Στο κεφάλαιο αυτό αναλύουμε αρχικά τα πλεονεκτήματα και τα προτερήματα υπαρχόντων εργαλείων. Στη συνέχεια περιγάφουμε την ανάπτυξη ενός νέου εργαλείου (*Targetprofiler*) που χρησιμοποιεί Κρυφά Μαρκοβιανά Μοντέλα (ΚΜΜ), προκειμένου να μεγιστοποιήσει την ευαισθησία και την ειδικότητα σε βαθμό ώστε να υπερτερεί έναντι των υπαρχόντων εργαλείων πρόβλεψης στόχων.

4.2 Επισκόπηση Υπαρχόντων Εργαλείων Πρόβλεψης MiRNA Στόχων

4.2.1 Βιολογικά γνωρίσματα που χρησιμοποιούνται από υπάρχοντα εργαλεία

Τα προαναφερθέντα βιολογικά γνωρίσματα χρησιμοποιούνται με διαφορετικούς τρόπους και σε διαφορετικούς συνδιασμούς από τα υπάρχονται υπολογιστικά εργαλεία με στόχο την αναγνώριση των miRNA στόχων στα θηλαστικά (78). Στον Πίνακα 4.1 συνοψίζονται τα γνωρίσματα 5 τέτοιων εργαλείων οργανωμένα σε τρεις κατηγορίες: (i) Αλληλουχία – το γνώρισμα αυτό επιτρέπει την ανίχνευση αλληλεπιδράσεων μεταξύ του miRNA και της στοχευμένης περιοχής και επιπλέον αναδεικνύει την ακριβή περιοχή της υβριδοποίησης καθώς και την κατηγορία (για παράδειγμα 5′-dominant seed ή 3′-compensatory). (ii) Θερμοδυναμική – η ελάχιστη ελεύθερη ενέργεια (ΔG) της miRNA::mRNA υβριδικής δομής είναι πολύ σημαντική για την πρόβλεψη στοχευμένων περιοχών. Προγράμματα όπως το RNAcofold (79) ή το RNAhybrid (80) μπορούν επιτυχώς να προβλέψουν υβριδικές δομές μεταξύ δύο μορίων RNA βασιζόμενα στην συμπληρωματικότητα και επιπλέον αποδίδουν την ελάχιστη ελεύθερη ενέργεια για τη συνολική δομή. Καθώς τα miRNA συνδέονται με τους στόχους τους με έναν πολύ συγκεκριμένο και σταθερό τρόπο, αναμένεται ότι η ΔG θα είναι μικρή. Οι υπολογισμοί ελεύθερης ενέργειας από τις πειραματικά επιβεβαιωμένες στοχευμένες περιοχές επιδεικνύουν μία πολύ χαμηλή τιμή (155). (iii) Συντήρηση – οι στοχευμένες περιοχές είναι λειτουργικές αλληλουχίες στο 3’UTR του μεταγραφόμενου RNA. Το γεγονός αυτό έχει ως αποτέλεσμα οι περιοχές αυτές να υπόκεινται σε εξελικτική συντήρηση σε διάφορους οργανισμούς. Η ανάλυση συντήρησης που χρησιμοποιεί τις multiz πλήρεις γονιδιακές ευθυγραμμίσεις δείχνει ότι περίπου το 70% των πειραματικά επιβεβαιωμένων στοχευμένων περιοχών συντηρούνται σε 8 άλλους οργανισμούς που μελετήθηκαν. Έτσι, η χρήση της συντήρησης ως γνώρισμα μπορεί να παρέχει πρόσθετη υποστήριξη για την πρόβλεψη νέων στοχευμένων περιοχών.

Πίνακας 4.1 Σύνοψη των γνωρισμάτων που χρησιμοποιούνται από εργαλεία πρόβλεψης στόχων των miRNA για θηλαστικά. Ο πίνακας προέρχεται από τη μελέτη των *Sethupathy et al 2006* (78).

Features	TargetScan	D-microT	miRanda	TargetScanS	PicTar
Sequence					
Perfect seed match rule	×			×	
Preference for perfect seed match[a]					×
Empirically determined binding rules		×			
Dynamic programming alignment score cutoff			×		
Seed 5′ and/or 3′ flank requirements				×	
Thermodynamics					
ΔG calculations based on traditional RNA folding programs	×	×	×		
ΔG calculations based on programs for short nucleic acid hybridizations					×
Conservation					
Only between human and rodent species		×	×[b]		
Among human, chimp, rodent, and dog	×		×[b]	×	×
Residing in an 'island' of conservation				×	

[a]PicTar[10,25] does predict targets with imperfect seed matches, but preferentially predicts targets with perfect seed matches. [b]miRanda provides the option of running the program under both parameters. The comparative study presented in this paper uses the "only human and rodent" version of miRanda.

4.2.2 Σύγκριτική Περιγραφή Επιλεγμένων Εργαλείων

Παρακάτω περιγράφονται αναλυτικά κάποια ευρέως χρησιμοποιούμενα εργαλεία πρόβλεψης miRNA στόχων με ιδιαίτερη έμφαση στις κατηγορίες τον στοχευμένων περιοχών στις οποίες εξειδικεύεται το κάθε εργαλείο.

PicTar

Το PicTar (http://pictar.bio.nyu.edu/cgi-bin/new_PicTar_mouse.cgi) είναι ένα πρόγραμμα κατάλληλο για πρόβλεψη "5' dominant" στόχων. Ωστόσο προβλέπει και στόχους οι οποίοι δεν έχουν τέλεια συμπληρωματικότητα στο seed site εφόσον η ενέργεια της αλληλεπίδρασης ξεπερνάει κάποιο όριο. Χρησιμοποιεί μία διαδικασία υπολογισμού της μέγιστης πιθανότητας αλληλεπίδρασης μεταξύ miRNA και πιθανού στόχου για να ενσωματώσει την συνδυαστική φύση της στόχευσης των miRNA. Ακόμα παίρνει υπόψη του τη συντήρηση της θέσης αλληλεπίδρασης με το seed site και απαιτεί συντήρηση σε τουλάχιστον 5 είδη. Μία πιο πρόσφατη έκδοση του προγράμματος παρέχει έτοιμες προβλέψεις που βασίζονται σε συγκριτική ανάλυση 17 γονιδιωμάτων σπονδυλωτών (84) .

TargetScan/TargetScanS

Το TargetScan (http://www.targetscan.org/index.html) είναι επίσης κατάλληλο για την πρόβλεψη "5' dominant" στόχων. Απαιτεί τέλεια συμπληρωματικότητα στη θέση πρόσδεσης του seed και 5 γενικά γνωρίσματα στην περιοχή γύρω από αυτή που βοηθούν τη λειτουργικότητα της αλληλεπίδρασης:

- ύπαρξη πολλών AU στην περιοχή,
- εγγύτητα στις θέσεις-στόχους για συνεκφραζόμενα miRNA,
- εγγύτητα στις βάσεις του UTR που αλληλεπιδρούν με τα νουκλεοτίδια που βρίσκονται στις θέσεις 13 έως 16 του miRNA,
- η θέση αλληλεπίδρασης να βρίσκεται τουλάχιστον 15 νουκλεοτίδια μακριά από το κωδικόνιο λήξης μέσα στο 3'UTR,
- η θέση αλληλεπίδρασης να βρίσκεται μακριά από το κέντρο του 3'UTR.

Επιπλέον, αυτό το εργαλείο αναγνωρίζει μη συντηρημένους στόχους διότι προβλέπει τη λειτουργικότητα του site χωρίς να λαμβάνει υπόψιν του την εξελικτική συντήρηση (82).

Miranda

Το Miranda (http://cbio.mskcc.org/mirnaviewer/) αναγνωρίζει την σημαντικότητα της θέσης πρόσδεσης του seed αλλά δεν απαιτεί τέλεια συμπληρωματικότητα. Επίσης απαιτεί συντήρηση μόνο ανάμεσα στον ποντικό και στον άνθρωπο, οπου συντηρημένη περιοχή θεωρείται αυτή που παρουσιάζει παραπάνω από 90% ομοιότητα σε αντίστοιχες θέσεις της 3'UTR ευθυγράμμισης των δυο οργανισμων (85).

PITA

Το PITA (http://genie.weizmann.ac.il/pubs/mir07/) χρησιμοποιεί ευρέως χρησιμοποιούμενες μεθόδους (όπως συμπληρωματικότητα στην περιοχή seed ή εκτεταμένη συμπληρωματικότητα στο 3′ άκρο του miRNA προκειμένου να αντισταθμίσουν την περιορισμένη συμπληρωματικότητα με την περιοχή seed) για να αναγνωρίσει αρχικά πιθανές θέσεις πρόσδεσης του seed (seed sites) για κάθε miRNA στο 3'UTR. Στη συνέχεια εφαρμόζει ένα μοντέλο προσβασιμότητας του στόχου (το οποιο βασίζεται σε υπολογισμούς ελεύθερης ενέργειας της δευτερεύουσας δομής του 3'UTR που περιβάλλει τη στοχευμένη περιοχή) και συνδυάζει τις πιθανές θέσεις

αλληλεπίδρασης για το ίδιο miRNA για να ορίσει ένα συνολικό βαθμό της αλληλεπίδρασης του miRNA με το 3'UTR. Επίσης αυτό το εργαλείο εισάγει και άλλη μία διάσταση στη πρόβλεψη miRNA στόχων, τα "flank sites". Τα "flank sites" είναι περιοχές γύρω από το seed site οι οποίες δεν σχηματίζουν δεσμούς με άλλες βάσεις. Βρέθηκε ότι "flanks" 3 πριν (upstream) και 15 μετά (downstream) το seed site οδηγούν στα καλύτερα αποτελέσματα του εργαλείου, που ξεπερνούν όλες τις άλλες μεθόδους. Η απόδοση του PITA είναι καλύτερη από άλλες μεθόδους ακόμα και χωρίς τη χρήση της επιλογής "3-15 flank"(83).

Diana-microT

Το Diana-microT (http://www.diana.pcbi.upenn.edu/) χρησιμοποιεί ένα αλγόριθμο δυναμικού προγραμματισμού και όπως το Miranda αναγνωρίζει την σημαντικότητα της θέσης πρόσδεσης του seed αλλά δεν απαιτεί τέλεια συμπληρωματικότητα. Επίσης απαιτεί συντήρηση ανάμεσα στον ποντικό και στον άνθρωπο αλλά προσφέρει την επιπλέον δυνατότητα ο χρήστης να επιλέξει η συντήρηση να παρουσιάζεται σε περισσότερους οργανισμους (81).

Σε μία πρόσφατη μελέτη, οι Sethupathy et al (78) έκαναν μία σύγκριση 5 εργαλείων πρόβλεψης στόχων: του TargetScan 4.0 (82), του TargetScanS (154), του PicTar (84), του DIANA-MicroT (81) και του Miranda (85). Στα πλαίσια της μελέτης χρησιμοποιήθηκαν 84 πειραματικά επιβεβαιωμένες αλληλεπιδράσεις miRNA::mRNA και 32 διαφορετικά miRNA από τη βάση δεδομένων Tarbase ως θετικό σύνολο δεδομένων. Όλα τα εργαλεία πρόβλεψης χρησιμοποιήθηκαν για να σαρώσουν όλες τις ανθρώπινες 3'UTR αλληλουχίες για νέους miRNA στόχους. Για κάθε εργαλείο πρόβλεψης καταγράφηκε ο αριθμός των θετικών/πραγματικών miRNA που προβλέφθηκαν σωστά (ευαισθησία) και ο συνολικός αριθμός επιτυχιών/προβλέψεων (ειδικότητα). Ο πίνακας 4.2 περιέχει μία αναπαράσταση των αποτελεσμάτων. Εξετάζοντας τα αποτελέσματα για το κάθε εργαλείο ξεχωριστά, προκύπτει ότι τα TS και DT πετυχαίνουν πολύ χαμηλά ποσοστά (20.8% και 9.5% αντιστοίχως) από σωστές προβλέψεις πειραματικά επιβεβαιωμένων αλληλεπιδράσεων miRNA-στοχευμένων γονιδίων (μέτρο ευαισθησίας). Πετυχαίνουν επίσης μικρό αριθμό συνολικών προβλέψεων (μέτρο ειδικότητας). Ο αριθμός των πραγματικών συνολικών στόχων για κάθε miRNA διαφέρει σημαντικά στα θηλαστικά. Γενικά έχει προταθεί ότι για κάθε miRNA υπάρχουν κατά μέσο όρο 7.1 αλληλεπιδράσεις miRNA-

στοχευμένων γονιδίων (156). Το μέτρο αυτό χρησιμοποιήθηκε για να αξιολογηθεί το κάθε εργαλείο και εμφανίζεται στην τελευταία στήλη του πίνακα 4.2. Η αξιολόγηση που προκύπτει δείχνει ότι τα εργαλεία που αναφέρθηκαν παραπάνω επιδεικνύουν ένα ρεαλιστικό συνολικό αριθμό προβλέψεων, λαμβάνοντας υπόψη τις βιολογικές ενδείξεις. Κοιτάζοντας τα αποτελέσματα για τα υπόλοιπα τρία εργαλεία (miR, TSS και PT) είναι φανερό ότι πετυχαίνουν σημαντικά μεγαλύτερα ποσοστά (48.8%, 47.6% και 47.6% αντιστοίχως) εντοπισμού πειραματικά επιβεβαιωμένων αλληλεπιδράσεων miRNA-στοχευμένων γονιδίων. ‘Ομως, ο συνολικός αριθμός προβλέψεων που αποκομίστηκε από τα εργαλεία αυτά είναι πολύ μεγάλος. Αυτό φαίνεται παρατηρώντας τον αριθμό αλληλεπιδράσεων miRNA-στοχευμένων γονιδίων για κάθε miRNA. Και τα τρία αυτά εργαλεία επιδεικνύουν έναν μη ρεαλιστικό αριθμό αλληλεπιδράσεων miRNA-στοχευμένων γονιδίων για κάθε miRNA, υπερβαίνοντας την προταθείσα τιμή (7.1) κατά 50 περίπου φορές.

Συνοψίζοντας, ο πίνακας με τα αποτελέσματα παρακάτω δείχνει ότι μολονότι κάποια εργαλεία όπως το miR, TSS και το PT μπορεί να πετυχαίνουν σχετικά υψηλές τιμές ευαισθησίας, αποτυγχάνουν να μεγιστοποιήσουν την ειδικότητα και ως αποτέλεσμα παράγουν έναν πολύ μεγάλο και μη ρεαλιστικό αριθμό πιθανών λανθασμένων προβλέψεων. Αντίθετα, άλλα εργαλεία (TS, DT) πετυχαίνουν έναν βιολογικά σημαντικό αριθμό συνολικών προβλέψεων, εις βάρος όμως της ευαισθησίας.

Πίνακας 4.2 Σύγκριση εργαλείων πρόβλεψης στόχων miRNA. Για κάθε πρόγραμμα φαίνεται η ευαισθησία και ο συνολικός αριθμός προβλέψεων. Ο πίνακας προέρχεται από τη μελέτη των *Sethupathy et al 2006* (78).

	Program	Percentage of experimentally supported miRNA–target gene interactions predicted[a]	Percentage of conserved and experimentally supported miRNA–target gene interactions predicted[b]	Number of total miRNA–target gene interactions predicted	Number of total miRNA–target gene interactions predicted per miRNA
Individual programs					
1	TS	20.8% (4.7%)	29.6% (7.3%)	278	9
2	DT	9.5% (1.3%)	13.1% (1.9%)	95	3
3	miR	48.8% (48.8%)	67.2% (67.2%)	18,289	572
4	TSS	47.6% (45.6%)	66.1% (64.3%)	10,351	323
5	PT	47.6% (45.0%)	65.6% (63.2%)	11,259	352
Program unions					
6	PT, TSS	52.4% (51.8%)	72.3% (71.7%)	14,583	456
7	PT, TSS, TS	57.1% (56.1%)	78.7% (78.0%)	14,690	459
8	PT, TSS, DT	57.1% (55.6%)	78.7% (77.6%)	14,632	457
9	PT, TSS, miR	66.7% (66.7%)	91.8% (91.8%)	26,800	838
10	PT, TSS, miR, TS	70.2% (69.9%)	96.7% (96.7%)	26,881	840
11	PT, TSS, miR, TS, DT	72.6% (71.6%)	100% (100%)	26,915	841
Program intersections					
12	PT, TSS	41.7% (41.0%)	57.3% (56.7%)	7,036	220
13	PT, TSS, TS	11.9% (11.9%)	16.4% (16.4%)	119	4
14	PT, TSS, DT	3.6% (3.6%)	4.9% (4.9%)	25	<1
15	PT, TSS, miR	28.6% (28.6%)	39.3% (39.3%)	3,895	122
16	PT, TSS, miR, TS	9.5% (9.5%)	13.1% (13.1%)	103	3
17	PT, TSS, miR, TS, DT	0.0% (0.0%)	0.0% (0.0%)	2	~0

TS, TargetScan; DT, DIANA-microT; miR, miRanda; TSS, TargetScanS; PT, PicTar.

[a]Percentages represent the sensitivity using the current data set. Percentages in parentheses represent the sensitivity test using the unbiased current data sets. [b]Percentages represent the sensitivity using the conserved current data set. Percentages in parentheses represent the sensitivity test using the conserved unbiased current data sets.

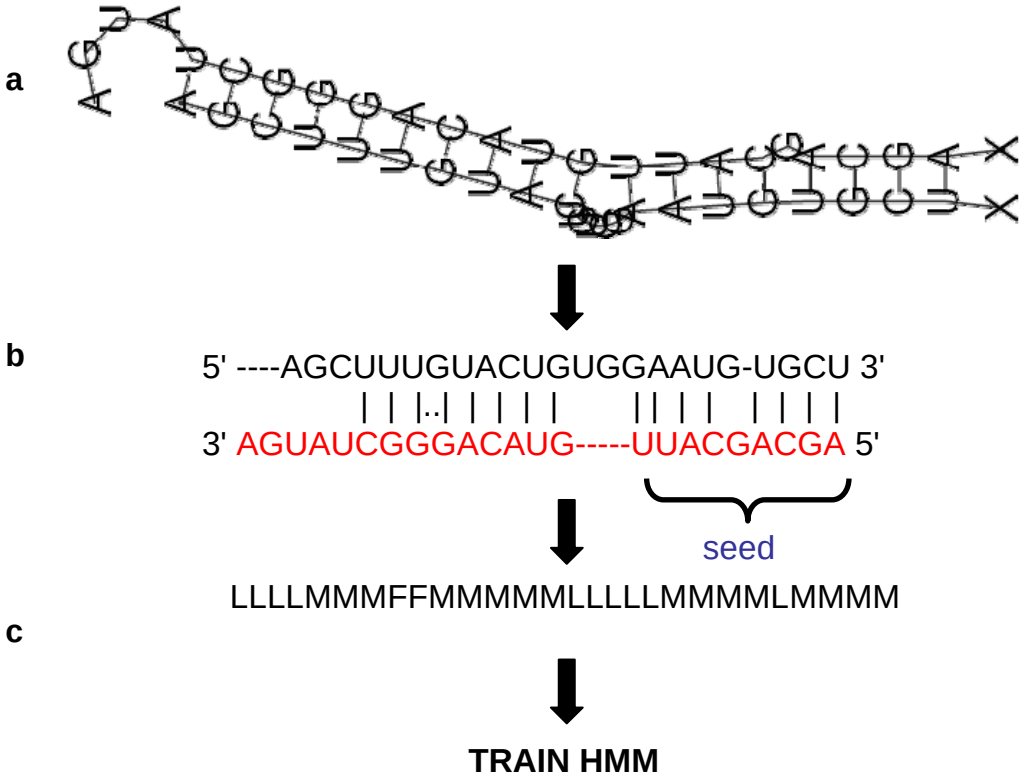
4.3 Υλικά και Μέθοδοι

4.3.1 Σύνολα Δεδομένων

Όλοι οι διαθέσιμοι στόχοι των miRNA για τους οποίους υπήρχε πειραματική επιβεβαίωση ανακτήθηκαν από την πιο πρόσφατη έκδοση της βάσης δεδομένων Tarbase (version 4). Για τους σκοπούς της μελέτης αυτής, χρησιμοποιήθηκαν μόνο τα ανθρώπινα δεδομένα, τα οποία αποτελούνται από 129 πειραματικά επιβεβαιωμένες miRNA στοχευμένες περιοχές (90 από τις οποίες έχουν γονιδιακή επισήμανση) και 57 miRNA. Το πειραματικό σύνολο που χρησιμοποιήθηκε από τους *Sethupathy et al* αποτελούνταν από 84 πειραματικά επιβεβαιωμένες αλληλεπιδράσεις miRNA:mRNA, και 32 διαφορετικά miRNA.

4.3.2 Εκπαίδευση του ΚΜΜ για την Αναγνώριση Γνωρισμάτων των Αλληλεπιδράσεων miRNA::Στοχευμένων Περιοχών

Για το σχεδιασμό ενός βελτιωμένου εργαλείου πρόβλεψης στόχων χρησιμοποιούμε τους ήδη υλοποιημένους αλγορίθμους ΚΜΜ που περιγράφονται στο *SSCprofiler*. Διαθέσιμοι αλγορίθμοι όπως ο RNAcofold (79), χρησιμοποιούνται για τον προσδιορισμό της δευτεροταγούς δομής των πειραματικά επιβεβαιωμένων αλληλεπιδράσεων miRNA::mRNA. Οι πληροφορίες σχετικά με τη δευτεροταγή δομή χρησιμοποιούνται στη συνέχεια για την εκπαίδευση του Μαρκοβιανού μας Μοντέλου. Επιπρόσθετα, ενσωματώνονται επιπλέον βιολογικές πληροφορίες όπως η εξελικτική συντήρηση στην προβλεφθείσα περιοχή και η συνεργασιμότητα (cooperativity of binding) στην πρόσδεση. Αρχικά, αυτό ερευνήθηκε χρησιμοποιώντας πειραματικά επιβεβαιωμένες περιοχές. Το Σχήμα 4.2 απεικονίζει πως η έξοδος του RNAcofold μετατρέπεται σε μία αναπαράσταση συμβολοσειράς από L (loops) και M (matches). Στη συνέχεια κατασκευάζεται μία ευθυγράμμιση πολλαπλών αλληλουχιών (msa) της LM αναπαράστασης συμβολοσειράς, όπου όλες οι αλληλεπιδράσεις miRNA-στοχευμένων γονιδίων ευθυγραμμίζονται σύμφωνα με την 5’ τους περιοχή και χρησιμοποιούνται στη συνέχεια για την εκπαίδευση του ΚΜΜ.



Σχήμα 4.2 RNAcofold σε KMM. Το **(a)** απεικονίζει την δομή της αλληλεπίδρασης του miRNA (πάνω αλληλουχία) με τη στοχευμένη περιοχή (κάτω αλληλουχία) όπως παράγονται από το RNAcofold. Το **(b)** απεικονίζει μία αναπαράσταση σε κείμενο της εξόδου που εμφανίζεται στο **(a)**. Τέλος το **(c)** απεικονίζει τη μετατροπή σε συμβολοσειρά της εξόδου από το RNAcofold. Οι δομές αυτές από όλες τις διαθέσιμες πειραματικά υποστηριζόμενες περιοχές από την Tarbase ευθυγραμμίζονται ως προς την 5' τους περιοχή για την εκπαίδευση του KMM.

4.3.3 Φιλτράρισμα

Χρησιμοποιούμε ένα σύνολο κανόνων για να φιλτράρουμε τα δεδομένα μας συνδυάζοντας την βαθμολογία από τα KMM και την ελεύθερη ενέργεια που προβλέπεται από την αναδίπλωση του RNA. Οι κανόνες αυτοί επιτρέπουν την πρόβλεψη και την κατηγοριοποίηση της 3′-compensatory κατηγορίας των miRNA στοχευμένων περιοχών. Όσο μεγαλύτερή η βαθμολογία του KMM τόσο καλύτερη η υβριδοποίηση στην περιοχή του seed. Έτσι, υποψήφια γονίδια που επιδεικνύουν υβριδικές δομές ανάλογες με τις 5′-dominant canonical ή 5′-dominant seed κατηγορίες των miRNA στοχευμένων περιοχών λαμβάνουν υψηλή βαθμολογία όπως

αποδίδεται από το KMM (>4). Εάν ένας υποψήφιος miRNA-στόχος έχει χαμηλή βαθμολογία (παράδειγμα 2, περιμένουμε ότι η συμπληρωματικότητα των βάσεων στην 3' περιοχή αντισταθμίζει για την έλλειψη συμπληρωματικότητας στην περιοχή του seed. Έτσι για να πάρουμε αλληλουχίες που ανήκουν στην 3' compensatory κατηγορία των miRNA στοχευμένων περιοχών εφαρμόζουμε τους ακόλουθους κανόνες φιλτραρίσματος:

1. IF **Score** > **4** THEN **Energy** <= **-8.0 dG**
2. IF **Score** > **3** THEN **Energy** <= **-12.0 dG**
3. IF **Score** > **2** THEN **Energy** <= **-14.0 dG**
4. IF **Score** > **1** THEN **Energy** <= **-16.0 dG**
5. IF **Score** > **0** THEN **Energy** <= **-18.0 dG**

Οι κανόνες αυτοί βασίζονται στην γραμμική συνάρτηση της βαθμολογίας όπως αποδίδεται από το KMM (score) σε σχέση με την ελάχιστη ελεύθερη ενέργεια (Energy). Όσο η βαθμολογία του KMM μειώνεται (λιγότερη συμπληρωματικότητα περιοχή seed) η ενέργεια μειώνεται επίσης, εξασφαλίζοντας ότι οι υποψήφιοι miRNA-στόχοι με περιοχές χωρίς καλή συμπληρωματικότητα στην περιοχή seed, θα κατηγοριοποιηθούν ως σωστοί στόχοι αν το ΔG είναι αρκετά μικρό που να δεικνύει υβριδοποιήση του τύπου 3′-compensatory .

4.3.4 Σάρωση Γενωμικών Περιοχών 3'UTR για miRNA Στόχους

Με την εκπαίδευση των KMM και την εφαρμογή των προαναφερόμενων παραμέτρων φιλτρασιμαρτος, η μεθοδολογία μας ενσωματώνει όλα τα σημαντικά βιολογικά γνωρίσματα που εμπλέκονται στην αλληλεπίδραση miRNA::mRNA. Το επόμενο βήμα που ακολουθεί είναι η σάρωση ή σκανάρισμα όλων των ανθρώπινων περιοχών 3'UTR για νέους miRNA στόχους. Αυτή η διαδικασία πρόβλεψης απαιτεί μεγάλη υπολογιστική ισχύ, ώστε να ελαχιστοποιηθεί ο χρόνος εκτέλεσης (CPU time). Για τον λόγο αυτό χρησιμοποιήσαμε μία συστοιχία υπολογιστών (PC cluster) για την εκτέλεση των υπολογιστικών αυτών πειραμάτων.

Η εξίσωση του cumulative score λαμβάνει υπόψη το ποσοτό ακρίβειας πρόβλεψης για τους πειραματικά επιβεβαιωμένους miRNA-στόχους και τον αριθμό

προβλεπόμενων στόχων για κάθε miRNA. Με αυτό τον τρόπο γίνεται μια περισσότερο αξιόπιστη αξιολόγηση για την αποτελεσματικότητα του κάθε εργαλείου πρόβλεψης στόχων των miRNA.

$$CS = 10 - \left[\left(0.5 \frac{100 - PE}{100} + 0.5 \frac{|Nm - 7|}{Max_{Nm} - 7} \right) \cdot 10 \right]$$

Όπου

PE: Percentage of Experimentally supported targets correctly predicted (100%)

Nm: Number of targets predicted per miRNA.

Max_{Nm}: Maximum number of targets predicted per miRNA by one tool -7.

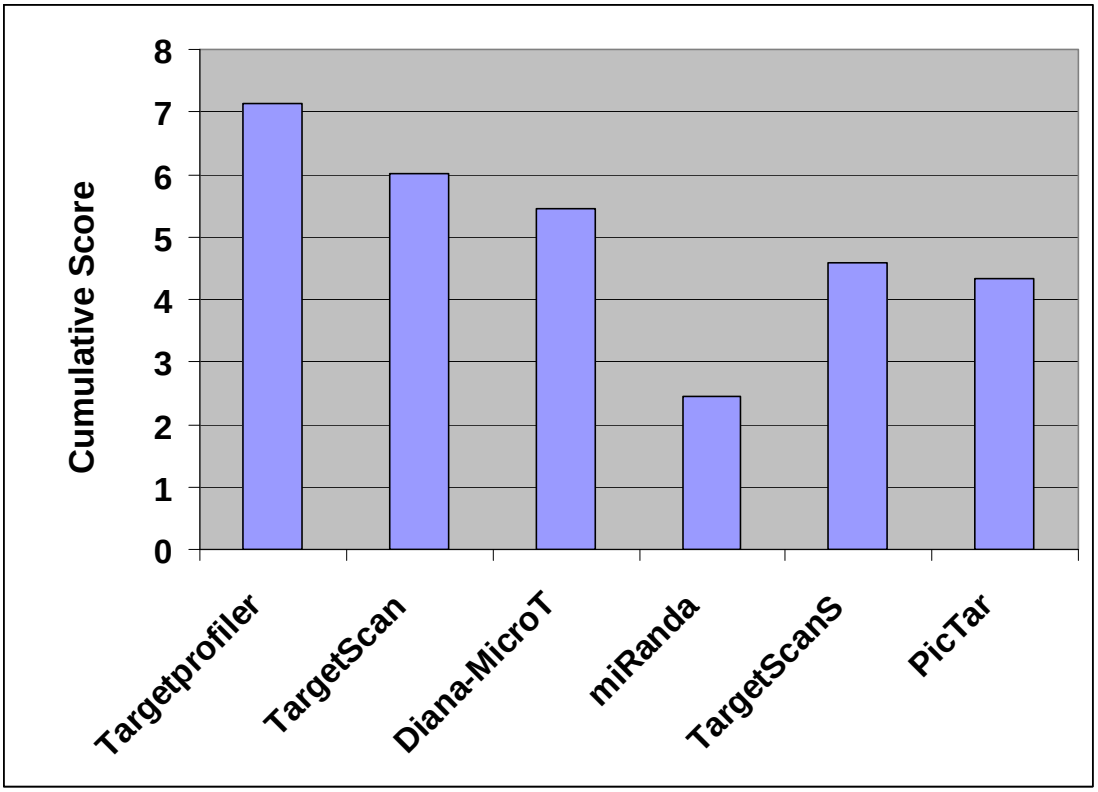
4.3 Αποτελέσματα

Σκανάραμε όλα τα ανθρώπινα 3'UTR χρησιμοποιώντας τα εκπαιδευμένα μας ΚΜΜ και εφαρμόζοντας τους κανόνες φιλτραρίσματος και τα αποτελέσματα που πήραμε φαίνονται στον πίνακα 4.3. Για να συγκρίνουμε τον ΚΜΜ αλγόριθμο (*Targetprofiler*) πρόβλεψης θέσεων στόχων με άλλα εργαλεία απεικονίζουμε τα αποτελέσματά μας σύμφωνα με τους *Sethupathy et al.* Όπως φαίνεται στον πίνακα, για μία βαθμολογία >= 6 το εργαλείο μας έχει ακρίβεια πρόβλεψης 45.74% για τους πειραματικά επιβεβαιωμένους miRNA-στόχους ενώ ο συνολικός αριθμός προβλεπόμενων αλληλεπιδράσεων είναι 1308. Αυτό αντίστοιχει σε 22.95 προβλεπόμενους στόχους για κάθε miRNA. Η τελευταία αυτή τιμή υπερβαίνει το μέσο όρο των στόχων ανά miRNA (προταθείσα τιμή 7 (156)) αλλά εξακολουθεί να είναι σχετικά κοντά στην πραγματικότητα αν αναλογιστεί κανείς ότι η τυπική απόκλιση της κατανομής των τιμών στόχων ανά miRNA είναι 4.7 (156). Επιπλέον, όπως αποδεικνύεται με την συνεχή ανακάλυψη περισσότερων miRNA στόχων, οι αριθμοί αυτοί μπορεί να αποδειχθούν υποτιμημένοι. Για να προυσίασουμε περισσότερα στοιχεία για την βελτιωμένη αρκρίβεια προβλεψης στόχων των miRNA του *Targetprofiler,* εξάγουμε μία εξίσωση η οποία λαμβάνει εξίσου υπόψη το ποσοστό ακρίβειας πρόβλεψης για τους πειραματικά επιβεβαιωμένους miRNA-

στόχους και τον αριθμό προβλεπόμενων στόχων για κάθε miRNA. Η τιμή που λαμβάνεται από αυτήν την εξίσωση (cumulative score) συγχωνεύει δυο σημαντικά πρότυπα της προβλεψής στόχων των miRNA και δίνει μια περισσότερο αντιπροσωπευτική ένδειξη της αποτελεσματικότητας του εργαλείου. Στο Σχήμα 4.3 παρουσίάζουμε ένα γράφημα το οποίο συγκρίνει τις τιμές cumulative score από διάφορα εργαλεία συμπεριλαμβανομένου και του *Targetprofiler*.

Πίνακας 4.3 Αποτελέσματα από την εξερεύνηση όλων των ανθρώπινων 3' UTR χρησιμοποιώντας τους εκπαιδευμένους αλγορίθμους KMM (*Targetprofiler*) και εφαρμόζοντας τους κανόνες φιλτραρίσματος.

Hmm Score (>=)	**Experimentally supported targets**	**Percentage of Experimentally supported targets correctly predicted**	**Predicted Targets**	**Percentage of Experimentally Supported targets in predicted**	**Number of targets predicted per miRNA**
0	90	69.77	49565	0.18	869.56
1	90	69.77	46495	0.19	815.70
2	90	69.77	40695	0.22	713.95
3	90	69.77	34801	0.26	610.54
4	89	68.99	19976	0.45	350.46
5	79	61.24	6339	1.25	111.21
6	59	45.74	1308	4.51	22.95



Σχήμα 4.3 Το σχήμα απεικονίζει τις τιμές που έλαβαν 6 εργαλεία σύμφωνα με την εξίσωση του cumulative score. Όσο μεγαλύτερη η τιμη του cumulative score τόσο πιο αποτελεσματικό είναι ένα εργαλείο, καθώς συνοψίζει ικανοποιητικά ποσοτά ακρίβειας πρόβλεψης για τους πειραματικά επιβεβαιωμένους miRNA-στόχους και αριθμό προβλεπόμενων στόχων για κάθε miRNA κοντά στο βέλτιστο (τιμή 7). Όπως φαίνεται από το γράφημα το *Targetprofiler* εχει την υψηλότερη τιμη cumulative score σε σύγκριση με άλλα 5 εργαλεία, καθως επιτυγχάνει ακρίβεια πρόβλεψης 45.74% και αριθμό προβλεπόμενων στόχων για κάθε miRNA 22.95.

4.4 Συζήτηση

Τα αποτελέσματα που λήφθησαν από τον αλγόριθμο πρόβλεψης στόχων - *Targetprofiler* μπορούν τώρα να συγκριθούν με τα αποτελέσματα από άλλα αντίστοιχα εργαλεία. Είναι σαφές ότι η ευαισθησία του εργαλείου μας (45.74% για κατώφλι ΚΜΜ στο 6) είναι στην ίδια κλίμακα με την δεύτερη ομάδα εργαλείων πρόβλεψης στόχων (miR, TSS και PT – ευαισθησίες: 48.8%, 47.6% και 47.6% αντίστοιχα) που αναλύθηκαν στο υπο-κεφάλαιο 4.2. Επιπρόσθετα, η ευαισθησία του

εργαλείου μας (45.74%) είναι ιδιαιτέρως υψηλότερη σε σχέση με την πρώτη ομάδα εργαλείων (TS και DT – ευαισθησίες: 20.8% και 9.5% αντίστοιχα). Επιπλέων των διαφορών τους στην ευαισθησία, οι δύο αυτές ομάδες εργαλείων έχουν σημαντικές διαφορές και στον αριθμό των στόχων που προβλέπουν για κάθε miRNA. Βλέπουμε ότι η πρώτη ομάδα εργαλείων (χαμηλή ευαισθησία) βρίσκουν μια ρεαλιστική τιμή (<=9) για αυτή τη μέτρηση σύμφωνα με πρόσφατα στοιχεία (156), ενώ η δεύτερη ομάδα (υψηλή ευαισθησία) αποδίδει μια εξαιρετικά υψηλή και μη ρεαλιστική τιμή (<= 572) για τη μέτρηση αυτή. Η τιμή για τον αριθμό των στόχων που προβλέπεται για κάθε miRNA σύμφωνα με τα αποτελέσματα του δικού μας αλγόριθμου είναι 22.95, που είναι πολύ κοντά στο στατιστικό εύρος για τη προταθείσα μέτρηση (μέσος όρος 7.1, τυπική απόκλιση 4.7) όπως έχει υπολογιστεί για πραγματικά miRNA (156). Έτσι τα αποτελέσματα του δικού μας αλγόριθμου πρόβλεψης είναι αρκετά κοντά στη προταθείσα τιμή για την συγκεκριμένη μέτρηση. Συμπερασματικά, ο αλγόριθμός μας επιτυγχάνει τόσο υψηλή ευαισθησία (σε σύγκριση με άλλα εργαλεία) όσο και χαμηλό (ρεαλιστικό) αριθμό στόχων για κάθε miRNA. Τα κριτήρια αυτά δεν καλύπτονται από άλλα υπάρχοντα εργαλεία πρόβλεψης miRNA στόχων που αναλύθηκαν στην προαναφερθείσα μελέτη.

Κεφάλαιο 5

5 Μελλοντικές Κατευθύνσεις

5.1 Βασικό επίκεντρο της μελλοντικής δουλειάς

Μια μελέτη που δημοσιεύτηκε πρόσφατα (146), η οποία περιγράφει την τελευταία δουλειά που πραγματοποιήθηκε στα πλαίσια αυτής της διδακτορικής διατριβής, περιγράφει την εφαρμογή του *SSCprofiler*, ενός αποτελεσματικού εργαλείου πρόβλεψης γονιδίων miRNA. Στην εργασία αυτή (περιγράφεται αναλυτικά στο Κεφάλαιο 3), εκπαιδεύτηκαν ΚΜΜ για την αναγνώριση βιολογικών γνωρισμάτων των miRNA όπως η αλληλουχία, η δομή και η συντήρηση. Η στατιστική ανάλυση έδειξε ότι με τη μέθοδο αυτή επιτυγχάνονται υψηλά ποσοστά τόσο ευαισθησίας (88.95%) όσο και ειδικότητας (84.16%). Το *SSCprofiler* χρησιμοποιήθηκε για την ταυτοποίηση νέων υποψήφιων miRNA με ιδιαίτερη βιολογική σημασία μέσα σε CAGR. Επιπλέον, τα τέσσερα υποψήφια miRNA γονίδια με την υψηλότερη έκφραση επιβεβαιώθηκαν με ανάλυση northern blot σε κύτταρα HeLa. Η επιβεβαίωση των γονιδίων αυτών, τα οποία συχνά εξαλείφονται σε διάφορους τύπους καρκίνου, όπως προστάτη, colorectal και αστροκυττώματος, αποτελεί κίνητρο για μελλοντική δουλειά. Ο βασικός στόχος της δουλειάς αυτής θα είναι η αποσαφήνιση των μοριακών μηχανισμών που διαστρεβλώνονται ως αποτέλεσμα της εξάλειψης των υποψήφιων αυτών miRNA γονιδίων και συνδέονται με ογκογενετικές διαδικασίες. Επιπλέον η δουλειά αυτή θα προσφέρει αναμφισβήτητα στοιχεία ότι τα υποψήφιων αυτά miRNA γονίδια είναι πραγματικά νέα miRNA που εμπλέκονται στην ογκογένεση.

5.2 In vitro πειραματική επιβεβαίωση των γονιδίων στόχων των miRNA.

Αφού προβλεφθούν στόχοι με τη χρήση βιοπληροφορικών μεθόδων για τα 4 νεα miRNA το επόμενο βήμα είναι η επιβεβαίωση τους με τη χρήση πειραματικών μεθόδων. Πρώτα θα ερευνήσουμε αν η έκφραση των προβλεπόμενων γονιδίων στόχων παρεμποδίζεται με οποιοδήποτε τρόπο από την παρουσία των αντίστοιχων

miRNA. Αυτό μπορεί να πραγματοποιηθεί με τη χρήση κυτταρικών σειρών. Από τη στιγμή που έχουμε αδιάσειστα στοιχεία ότι τα miRNA εκφράζονται στην κυτταρική σειρά HeLa το πρώτο βήμα θα είναι να ελέγξουμε τα επίπεδα έκφρασης του mRNA και της πρωτεΐνης των προβλεπόμενων γονιδίων στόχων μέσω qRT-PCR και ανάλυση Western Blot, αντίστοιχα. Προκειμένου να παρατηρήσουμε την επίδραση της καταστολής των ενδογενών miRNA στα επίπεδα mRNA και πρωτεΐνης των εν λόγω γονιδίων θα πρέπει να συγκριθούν τα επίπεδα πριν και μετά την καταστολή τους. Τα miRNA μπορούν να κατασταλούν με τη χρήση 2'-O-Metlyl-modified μόρια συμπληρωματικών RNA, ειδικά για τα δικά μας miRNA. Τα συμπληρωματικά RNA αφού εισαχθούν στα κύτταρα μέσω επιμόλυνσης υβριδοποιούνται με τα miRNA και καταστέλλουν τη δράση τους. Αναμένεται να παρατηρηθεί αύξηση της έκφρασης των γονιδίων με την παρουσία του συμπληρωματικού RNA και χαμηλότερη έκφραση χωρίς αυτό, σίγουρα σε επίπεδο πρωτεΐνης και πιθανά και σε επίπεδο RNA, αναλόγως αν η αλληλεπίδραση miRNA-mRNA εκτός από μεταφραστική καταστολή οδηγεί και σε αποικοδόμηση του mRNA.

Μέσω των παραπάνω πειραμάτων ωστόσο δεν θα μπορούσε να εξαχθεί το συμπέρασμα ότι η μείωση της έκφρασης που παρατηρείται οφείλεται σε άμεση αλληλεπίδραση του miRNA με το mRNA. Για να διερευνήσουμε αυτό το ενδεχόμενο θα πραγματοποιήσουμε πειράματα με τη χρήση του γονιδίου της λουσιφεράσης ως γονίδιο αναφοράς (luciferase reporter gene assays). Αυτά τα πειράματα θα δώσουν ατράνταχτα στοιχεία ότι η αλληλεπίδραση του mRNA με το miRNA είναι άμεση. Αυτά τα πειράματα συνήθως πραγματοποιούνται με την ένωση της 3' αμετάφραστης περιοχής του γονιδίου, που πιστεύεται ότι ελέγχεται από κάποιο miRNA, σε ένα γονίδιο λουσιφεράσης. Στη συνέχεια επιμολύνονται κύτταρα με τη χρήση αυτού του κατασκευάσματος (construct) και ενός πλασμιδίου (control-vector) για τον έλεγχο της αποτελεσματικότητας της επιμόλυνσης. Υψηλή ενεργότητα λουσιφεράσης δηλώνει μικρή ή καθόλου λειτουργική αλληλεπίδραση μεταξύ του miRNA και της 3' αμετάφραστης περιοχής, ενώ χαμηλή ενεργότητα σημαίνει ισχυρή αλληλεπίδραση. Στην καθημερινή πρακτική, για την απόδειξη της άμεσης αλληλεπίδρασης και την εύρεση του ακριβούς σημείου πρόσδεσης του miRNA χρησιμοποιούνται και κατασκευάσματα (construct) με μεταλλαγμένο το σημείο πρόσδεσης του miRNA με την προϋπόθεση ότι με αυτόν τον τρόπο θα ακυρωθεί η αλληλεπίδραση του miRNA με την 3' αμετάφραστη περιοχή.

5.3 Λειτουργικός χαρακτηρισμός της ρύθμισης miRNA::mRNA-Στόχου – Ανάλυση In vitro

Προκειμένου να βρεθεί ο συνδετικός κρίκος της επιβεβαιωμένης αλληλεπίδρασης miRNA::mRNA με τον καρκίνο, θα χρησιμοποιήσουμε πειραματικές βιοαναλύσεις για να εξακριβωθούν οι μοριακές διαδικασίες που εμπλέκονται στον καρκίνο και στην ογκογένεση. Διαδικασίες όπως ο πολλαπλασιασμός, η απόπτωση, η επιθετικότητα και η αγγειογένεση είναι οι βασικοί μηχανισμοί που υπαινίσσονται καρκινικό φαινότυπο. Αρχικά θα ερευνήσουμε το βαθμό στον οποίο η αλληλεπίδραση miRNA::mRNA επιρρεάζει στις διαδικασίες αυτές. Με τη χρήση τεχνολογίας antisense RNA θα απενεργοποιήσουμε (knockout) τα miRNA σε κυτταρικές καλλιέργειες και θα παρατηρήσουμε την επίδρασή τους στις ογκογενετικές διαδικασίες που προαναφέρθηκαν. Υπάρχει πλήθος βιοαναλύσεων που προσφέρουν in vitro αποτίμηση των μοριακών αυτών μηχανισμών (72,157-159). Πιο συγκεκριμένα, αναλύσεις για ανάπτυξη anchorage-ανεξάρτητων κυτταρικών σειρών επιτρέπουν μια εκτίμηση του ογκογενετικού μετασχηματισμού που περιλαμβάνει τον ανεξέλεγκτο πολλαπλασιασμό και την επιθετικότητά τους. Η διερεύνηση του μηχανισμού αυτού πραγματοποιείται σε κυτταρικές καλλιέργειες που αναπτύσσονται σε soft agar (158). Η έκταση στην οποία τα κύτταρα σχηματίζουν αποικίες στο περιβάλλον αυτό είναι ανάλογη με την προδιάθεσή τους για ογκογενετικό μετασχηματισμό. Βιοαναλύσεις απόπτωσης περιλαμβάνουν αρχικά τον εντοπισμό θραυσμάτων DNA με ανάλυση Southern blot καθώς μόνο κύτταρα σε φάση απόπτωσης εμφανίζουν τη χαρακτηριστική εικόνα DNA ladder (72). Για την κατανόηση του συγκεκριμένου αποπτωτικού μονοπατιού που ενεργοποιείται, θα χρησιμοποιήσουμε ανάλυση immunoblot για εξειδικευμένες αποπτωτικές πρωτεΐνες. Η εμφάνιση μιας μπάντας σε ανάλυση Western blot μετά από ανάλυση immunoblot είναι χαρακτηριστική για τις πρωτεΐνες APAF-1, caspase 9 ή PARP και επιβεβαιώνει την ενεργοποίηση του μονοπατιού caspase 9. Εναλλακτικά, η εμφάνιση μιας μπάντας χαρακτηριστική για την πρωτεΐνη caspase 8 επιβεβαιώνει την ενεργοποίηση του αντίστοιχου μονοπατιού. Ο κυτταρικός θάνατος μπορεί να μετρηθεί με βαφή με Tryphan blue. Επίσης μια βιοανάλυση TUNEL (72) προσφέρει ένα αποπτωτικό δείκτη στο επίπεδο ενός κυττάρου και πιο συγκεκριμένα παρέχει μια μέτρηση του

αριθμού των αποπτωτικών πυρήνων μέσα σε μια κύτταρο-καλλιέργεια συγκρινόμενων με τον συνολικό αριθμό κυττάρων.

5.4 Ανάλυση In vivo

Μετά την ολοκλήρωση της προαναφερόμενης δουλειάς, θα βρισκόμαστε σε στάδιο όπου θα μπορούμε να συνεχίσουμε την μελέτη μας με *in vivo* πειράματα. Στο στάδιο αυτό θα χρησιμοποιήσουμε κύτταρα τα οποία έχουν επιμολυνθεί (όπως περιγράφεται στα vitro πειράματα) και επιπλέον εκδηλώνουν ένα σταθερό καρκινικό φαινότυπο (ανεξέλεγκτος πολλαπλασιασμός, αντοχή σε απόπτωση, επιθετική δράση κλπ). Τα κύτταρα αυτά θα εισαχθούν αρχικά με ένεση σε ποντίκια και μετά με τη χρήση τεχνολογίας imaging, θα εξετάσουμε το σημείο εισαγωγής. Τα σημεία στα οποία έχει γίνει εισαγωγή επιμολυσμένων κυττάρων αναμένεται να παρουσιάσουν ογκογένεση, σε αντίθεση με τα σημεία στα οποία έχουν εισαχθεί μη-επιμολυσμένα κύτταρα (control) για τα οποία δεν αναμένεται ανάπτυξη όγκου (158). Εναλλακτικά, μια πρόσφατη μελέτη παρουσίασε μια νέα τάξη από χημικά τροποποιημένα ολιγονουκλεοτίδια, τα ‘antagomirs’, τα οποία εμφανίζουν αποτελεσματική και εκλεκτική δράση στη διαδικασία σίγησης ενδογενούς miRNA στα ποντίκια (160). Εισαγωγή μέσω ένεσης των antagomirs εκλεκτικά για miR-16, miR-122, miR-192 και miR-194, είχε ως αποτέλεσμα σημαντική μείωση των αντίστοιχων miRNA σε διάφορα όργανα και ιστούς όπως συκώτι, πνεύμονες, νεφρά, καρδιά, έντερο, ωοθήκες, αδένες, δέρμα, λίπος, μυελό των οστών και μύες.

Κατάλογος Δημοσιεύσεων

1. **Anastasis Oulas,** Alexandra Boutla, Katerina Gkirtzou, Martin Reczko, Kriton Kalantidis and Panayiota Poirazi. "Prediction of novel microRNA genes in cancer associated genomic regions – a combined computational and experimental approach". ***Nucleic Acid Research***. 2009, 1-12
2. **Oulas A**, Reczko M, Poirazi P. "MicroRNAs and cancer-the search begins!" ***IEEE Trans Inf Technol Biomed***. 2009 Jan;13(1):67-77.
3. Petalidis LP, **Oulas A**, Backlund M, Wayland MT, Liu L, Plant K, Happerfield L, Freeman TC, Poirazi P, Collins VP. "Improved grading and survival prediction of human astrocytic brain tumors by artificial neural network analysis of gene expression microarray data". ***Molecular Cancer Therapeutics***. 2008 May;7(5):1013-24. Epub 2008 Apr 29.
4. Martin Reczko, Panayiota Poirazi, **Anastasis Oulas**, Eleftheria Tzamali, Maria Manioudaki, Vassilis Tsiaras and Ioannis Tollis. ***ERCIM News***, April 2007**,** Special theme: The Digital Patient.

Ανακοινώσεις σε Συνέδρια και Workshops

1. 2nd International Advanced Research Workshop on In SilicoOncology, Kolympari, Chania, Greece , 25th and 26th September 2006, "Prediction of novel miRNAs and their gene targets with implications in tumourigenesis", **Anastasis Oulas** , Martin Reczko, Panayiota Poirazi.
2. 1st Cretan Bioinformatics Forum, FORTH. Building Scientific Networks (**Invited Talk**)*,* "Analysis of Microarray Data and Prediction of miRNA genes using Computational tools" , June 19, 2006
3. ISMB/ECCB 2004: Glasgow, Scotland, UK. Poster Title: Biologically Inspired Neural Network and Gentic algorithms for microarray data classification and Identification of informative genes. **Anastasis Oulas**, Martin Reczko, Panayiota Poirazi
4. Computational Biology Workshop, Mediterranean Institute for Life Sciences (MedILS), Split, Croatia, July 25 – July 29, 2007.

5. The Fifth BioSapiens European School in Bioinformatics was held in Budapest (Hungary). Talk: “Computational Prediction of miRNA in Cancer”, Sep 4-8 2006
6. Workshop in Bioinformatics, Jointly organized by The Department of Biology, University of Crete(UoC) The Institute of Molecular Biology and Biotechnology (IMBB) of the Foundation for Research and Technology (FORTH), Tutorial Presentation on Microarray Analysis Tools, Greece, Sep 2004.

Βιβλιογραφία

1. Sorlie, T., Perou, C., Tibshiranie, R., Aasf, T., Geislerg, S., Johnsen, H., Hastiee, T., Eisen, M., Van de Rijni, M., Jeffreyj, S., Thorsenk, T., Quistl, H., Matesec, J., Brown, P., Botstein, D., Lonningg, P.E., and Borresen-Dale, A. (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. US*, **98**, 10869-10874.
2. Van 't Veer L., D., H., Van de Vijver, M., He, Y., Hart, A., Mao, M., Petesre, H., Van der Kooy, K., Marton, M., Witteveen, A., Schreiber, G., Kerkhoven, R., Roberts, C., Linsley, P., Bernards, R., Friend, S. (2002) Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer. *Nature*, **415**, 530-536.
3. Hedenfalk, I., Ringne, M., Ben-Dor, A., Yakhini, Z., Chen, Y., Chebil, G., Ach, R., and Loman, N., Olsson, H., Meltzer, P., Borg, E., and Trent, J. (2003) Molecular classification of familial non- BRCA1/BRCA2 breast cancer. *PNAS*, **100**, 2532-2537.
4. Mike West, C.B., Holly Dressman, Erich Huang, Seiichi Ishida, Rainer Spang, Harry Zuzan, John A. Olson, Jr. Jeffrey R. Marks, and Joseph R. Nevins. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl Acad. Sci. USA*, 11462–11467.
5. Gruvberger Sofia, M.R.r., Yidong Chen, Sujatha Panavally, Lao H. Saal, Åke Borg, Mårten Ferno¨, and Carsten Peterson, a.P.S.M. (2001) Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Research*, **61**, 5979–5984.
6. Perou, C., Jeffrey, S., Van De Rijni, M, Rees, C, Eisen, M, Ross, D., Pergamenschikov, A., Williams, C., Zhu, S., Lee, J., Lashkarii, D., Shalon, D., Brown, P., Botstein, D. (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. US*, **96**, 9212-9217.
7. Pomeroy, S.L.e.a. (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 436–442.
8. MacDonald, T.J.e.a. (2001) Expression profiling of medulloblastoma: PDGFRA and the RAS/MAPK pathway as therapeutic targets for metastatic disease. *Nature Genet.*, 143–152.
9. Wang, K.e.a. (1999) Monitoring gene expression profile changes in ovarian carcinomas using cDNA microarray. *Gene*, 101–108.
10. Jazaeri, A.A.e.a. (2002) Gene expression profiles of BRCA1-linked, BRCA2-linked, and sporadic ovarian cancers. *J. Natl Cancer Inst.* , 990–1000.
11. Welsh, J.B.e.a. (2001) Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proc. Natl Acad. Sci. USA*, 1176–1181.
12. Garber, M.E.e.a. (2001) Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl Acad. Sci. USA*, 13784–13789.
13. Beer, D.G.e.a. (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Med*, 816–824.
14. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., and Levine, A.J. (1999) Broad patterns of gene expression revealed by clustering

analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS*, **96**, 6745-6750.

15. Singh, D., Febbo, F.G., Kenneth, R., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., D'Amico, A., Richie, J., Lander, E., Loda, M., Kantoff, P., Golub, T., and Sellers, W. (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203-209.
16. Hippo, Y.e.a. (2002) Global gene expression analysis of gastric cancer by oligonucleotide microarrays. *Cancer Res.*, 233–240.
17. Ferrando, A.A.e.a. (2002) Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. *Cancer Cell*, 75–87.
18. Yeoh, E.J.e.a. (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. . *Cancer Cell*, 133–143.
19. Hofmann, W.K.e.a. (2002) Relation between resistance of Philadelphia-chromosomepositive acute lymphoblastic leukaemia to the tyrosine kinase inhibitor STI571 and gene-expression profiles: a gene-expression study. *Lancet*, 481–486.
20. Alizadeh, A., Eisen, M., Davis, R., Ma, C., Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., Powell, J., Yang, L., Marti, G., Moore, T., Hudson, J., Lu, L., Lewis, D., Tibshirani, R., Sherlock, G., Chan, W., Greiner, G., Weisenburger, D., Armitage, J., Warnke, R., Levy, R., Wilson, W., Grever, M., Byrd, J., Botstein, D., Brown,P., Staudt, L. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503-511.
21. Shipp, M.A.e.a. (2002) Diffuse large B-cell lymphoma outcome prediction by geneexpression profiling and supervised machine learning. *Nature Med*, 68–74.
22. Hanahan, D.W., R.A. . (2000) The hallmarks of cancer. *Cell* 57–70
23. Freije, W.A., Castro-Vargas, F.E., Fang, Z., Horvath, S., Cloughesy, T., Liau, L.M., Mischel, P.S. and Nelson, S.F. (2004) Gene expression profiling of gliomas strongly predicts survival. *Cancer Res*, **64**, 6503-6510.
24. Guyon, I.e.a. (2002) Gene Selection for Cancer using Support Vector Machines. *Machine Learning*, 389-422.
25. Ramaswamy, S., Tamayo Pablo., Rifkin, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J., P., , Poggio, T. *et al.* (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *PNAS*, **98**, 15149-15154.
26. Khan, J.E., J. S.,. Ringer,M., Saal,L. H., Ladanyi,M., Westermann,F.,Berthold,F.,Scwab,M.,Antonescu C.R., Pterson,C. and Meltzer,P.S. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Med*, 579-673.
27. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. . (1998) Cluster analysis and display of genome-wide expression patterns. . *Proc. Natl Acad. Sci. USA*, 14863–14868.
28. Shai, R., Shi, T., Kremen, T.J., Horvath, S., Liau, L.M., Cloughesy, T.F., Mischel, P.S. and Nelson, S.F. (2003) Gene expression profiling identifies molecular subtypes of gliomas. *Oncogene*, **22**, 4918-4923.
29. Soukas, A., Cohen, P., Socci, N.D. and Friedman, J.M. (2000) Leptin-specific patterns of gene expression in white adipose tissue. *Genes Dev*, **14**, 963-980.

30. Tamayo, P., et al. (1999) Interpreting patterns of gene expression with self-organizing maps:Methods and application to hematopoietic differentiation. *PNAS*, **96**, 2907-2912.
31. van den Boom, J., Wolter, M., Kuick, R., Misek, D.E., Youkilis, A.S., Wechsler, D.S., Sommer, C., Reifenberger, G. and Hanash, S.M. (2003) Characterization of Gene Expression Profiles Associated with Glioma Progression Using Oligonucleotide-Based Microarray Analysis and Real-Time Reverse Transcription-Polymerase Chain Reaction. *Am J Pathol*, **163**, 1033-1043.
32. Herrero, J., Valencia, A. and Dopazo, J. (2001) A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, **17**, 126-136.
33. Ben-Dor, A., Friedman, N.,and Yakhini, Z. (2003) Overabundance Analysis and Class Discovery in Gene Expression Data. *(submitted)*.
34. Tusher, V., Tibshirani, R., Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. US*, **98**, 5116-5121.
35. Perou, C.M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., et al. . (2000) Molecular portraits of human breast tumours. *Nature* **406**.
36. Chen, X.e.a.G.e.p.i.h.l.c.M.B. (2002).) *Cell* 1929–1939.
37. Bhattacharjee, A.e.a. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. . *Proc. Natl Acad. Sci. USA* 13790–13795.
38. Takahashi, M.e.a. (2001) Gene expression profiling of clear cell renal cell carcinoma:gene identification and prognostic classification. . *Proc. Natl Acad. Sci. USA*, 9754–9759.
39. Fisher, B.e.a. (1989) A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen receptorpositive tumors. . *N. Engl. J. Med.*, 479–484.
40. Lee, R., Feinbaum, R. and Ambros, V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. . *Cell*, 843–854.
41. Lee Y, A.C., Han J, Choi H, Kim J, et al. (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature*, 415–419.
42. Denli, A.M., Tops, B.B., Plasterk, R.H., Ketting, R.F. and Hannon, G.J. (2004) Processing of primary microRNAs by the Microprocessor complex. *Nature*, **432**, 231-235.
43. Gregory, R.I., Yan, K.P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N. and Shiekhattar, R. (2004) The Microprocessor complex mediates the genesis of microRNAs. *Nature*, **432**, 235-240.
44. Lund, E., Guttinger, S., Calado, A., Dahlberg, J.E. and Kutay, U. (2004) Nuclear export of microRNA precursors. *Science*, **303**, 95-98.
45. Yi R, Q.Y., Macara IG, Cullen BR . (2003) Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev*, 3011–3016.
46. Hutva´gner G, Z.P. (2002a) A microRNA in a multiple-turnover RNAi enzyme complex. *Science* 2056–2060.
47. Calin, G.A., Sevignani., C., Dumitru., C.D., Hyslop., T., Noch., E., Yendamuri., S., Shimizu., M., Rattan., S., Bullrich., F., Negrini., M. *et al.*

(2004) Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *PNAS*, **101**, 2999–3004.

48. Lagos-Quintana, M., Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W. & Tuschl, T. . (2002) Identification of Tissue-Specific MicroRNAs from Mouse. *Curr. Biol.* , 735–739.
49. Calin, G.A., Dumitru, C.D., Shimizu, M., Bichi, R., Zupo, S., Noch, E., Aldler, H., Rattan, S., Keating, M., Rai, K. *et al.* (2002) Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc Natl Acad Sci U S A*, **99**, 15524–15529.
50. Michael, M.Z., SM, O.C., van Holst Pellekaan, N.G., Young, G.P. and James, R.J. (2003) Reduced Accumulation of Specific MicroRNAs in Colorectal Neoplasia. *Mol Cancer Res* **1**, 882–891.
51. Zhang, L., Huang, J., Yang, N., Greshock, J., Megraw, M.S., Giannakakis, A., Liang, S., Naylor, T.L., Barchetti, A., Ward, M.R. *et al.* (2006) microRNAs exhibit high frequency genomic alterations in human cancer. *Proc Natl Acad Sci U S A*, **103**, 9136-9141.
52. Takamizawa, J., Konishi, H., Yanagisawa, K., Tomida, S., Osada, H., Endoh, H., Harano, T., Yatabe, Y., Nagino, M., Nimura, Y. *et al.* (2004) Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. *Cancer Res*, **64**, 3753-3756.
53. Johnson, S.M., Grosshans, H., Shingara, J., Byrom, M., Jarvis, R., Cheng, A., Labourier, E., Reinert, K.L., Brown, D. and Slack, F.J. (2005) RAS is regulated by the let-7 microRNA family. *Cell*, **120**, 635-647.
54. Koscianska, E., Baev, V., Skreka, K., Oikonomaki, K., Rusinov, V., Tabler, M. and Kalantidis, K. (2007) Prediction and preliminary validation of oncogene regulation by miRNAs. *BMC Mol Biol*, **8**, 79.
55. He, L., Thomson, J.M., Hemann, M.T., Hernando-Monge, E., Mu, D., Goodson, S., Powers, S., Cordon-Cardo, C., Lowe, S.W., Hannon, G.J. *et al.* (2005) A microRNA polycistron as a potential human oncogene. *Nature*, **435**, 828-833.
56. O'Donnell, K.A., Wentzel, E.A., Zeller, K.I., Dang, C.V. and Mendell, J.T. (2005) c-Myc-regulated microRNAs modulate E2F1 expression. *Nature*, **435**, 839-843.
57. Metzler M, W.M., Busch K, Viehmann S, Borkhardt A. . (2003) High Expression of precursor microRNA-155/BIC RNA in children with Burkitt lymphoma. *Genes Chrom Cancer.*, 167–169.
58. Kluiver, J., Poppema, S., de Jong, D., Blokzijl, T., Harms, G., Jacobs, S., Kroesen, B.J. and van den Berg, A. (2005) BIC and miR-155 are highly expressed in Hodgkin, primary mediastinal and diffuse large B cell lymphomas. *J Pathol*, **207**, 243-249.
59. Eis, P.S., Tam, W., Sun, L., Chadburn, A., Li, Z., Gomez, M.F., Lund, E. and Dahlberg, J.E. (2005) Accumulation of miR-155 and BIC RNA in human B cell lymphomas. *Proc Natl Acad Sci U S A*, **102**, 3627-3632.
60. Iorio, M.V., Ferracin, M., Liu, C.G., Veronese, A., Spizzo, R., Sabbioni, S., Magri, E., Pedriali, M., Fabbri, M., Campiglio, M. *et al.* (2005) MicroRNA gene expression deregulation in human breast cancer. *Cancer Res*, **65**, 7065-7070.
61. Volinia, S., Calin, G.A., Liu, C.G., Ambs, S., Cimmino, A., Petrocca, F., Visone, R., Iorio, M., Roldo, C., Ferracin, M. *et al.* (2006) A microRNA

expression signature of human solid tumors defines cancer gene targets. *Proc Natl Acad Sci U S A*, **103**, 2257-2261.
62. Clurman, B.E. and Hayward, W.S. (1989) Multiple proto-oncogene activations in avian leukosis virus-induced lymphomas: evidence for stage-specific events. *Mol Cell Biol*, **9**, 2657-2664.
63. Tam, W., Ben-Yehuda, D. and Hayward, W.S. (1997) bic, a novel gene activated by proviral insertions in avian leukosis virus-induced lymphomas, is likely to function through its noncoding RNA. *Mol Cell Biol*, **17**, 1490-1502.
64. Lagos-Quintana, M., Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W. and Tuschl, T. (2002) Identification of tissue-specific microRNAs from mouse. *Curr Biol*, **12**, 735-739.
65. Kluiver, J., Haralambieva, E., de Jong, D., Blokzijl, T., Jacobs, S., Kroesen, B.J., Poppema, S. and van den Berg, A. (2006) Lack of BIC and microRNA miR-155 expression in primary cases of Burkitt lymphoma. *Genes Chromosomes Cancer*, **45**, 147-153.
66. Yanaihara, N., Caplen, N., Bowman, E., Seike, M., Kumamoto, K., Yi, M., Stephens, R.M., Okamoto, A., Yokota, J., Tanaka, T. *et al.* (2006) Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. *Cancer Cell*, **9**, 189-198.
67. Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B. and Bartel, D.P. (2003) Vertebrate microRNA genes. *Science*, **299**, 1540.
68. He, H., Jazdzewski, K., Li, W., Liyanarachchi, S., Nagy, R., Volinia, S., Calin, G.A., Liu, C.G., Franssila, K., Suster, S. *et al.* (2005) The role of microRNA genes in papillary thyroid carcinoma. *Proc Natl Acad Sci U S A*, **102**, 19075-19080.
69. Ciafre, S.A., Galardi, S., Mangiola, A., Ferracin, M., Liu, C.G., Sabatino, G., Negrini, M., Maira, G., Croce, C.M. and Farace, M.G. (2005) Extensive modulation of a set of microRNAs in primary glioblastoma. *Biochem Biophys Res Commun*, **334**, 1351-1358.
70. Lagos-Quintana, M., Rauhut, R., Meyer, J., Borkhardt, A. and Tuschl, T. (2003) New microRNAs from mouse and human. *Rna*, **9**, 175-179.
71. Cai, X., Lu, S., Zhang, Z., Gonzalez, C.M., Damania, B. and Cullen, B.R. (2005) Kaposi's sarcoma-associated herpesvirus expresses an array of viral microRNAs in latently infected cells. *Proc Natl Acad Sci U S A*, **102**, 5570-5575.
72. Cimmino, A., Calin, G.A., Fabbri, M., Iorio, M.V., Ferracin, M., Shimizu, M., Wojcik, S.E., Aqeilan, R.I., Zupo, S., Dono, M. *et al.* (2005) miR-15 and miR-16 induce apoptosis by targeting BCL2. *Proc Natl Acad Sci U S A*, **102**, 13944-13949.
73. Yousef, M., Nebozhyn, M., Shatkay, H., Kanterakis, S., Showe, L.C. and Showe, M.K. (2006) Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. *Bioinformatics*, **22**, 1325-1334.
74. Reeder, J. and Giegerich, R. (2004) Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics*, **5**, 104-115.
75. Helvik, S.A., Snove, O., Jr. and Saetrom, P. (2006) Reliable prediction of Drosha processing sites improves microRNA gene prediction. *Bioinformatics*, **23**, 142-149.
76. Ruby, J.G., Jan, C.H. and Bartel, D.P. (2007) Intronic microRNA precursors that bypass Drosha processing. *Nature*, **448**, 83-86.

77. Oulas, A., Reczko, M. and Poirazi, P. (2008) MicroRNAs and Cancer – The Search Begins! *IEEE Trans Inf Technol Biomed*, **13**, 67-77.
78. Sethupathy, P., Megraw, M. and Hatzigeorgiou, A.G. (2006) A guide through present computational approaches for the identification of mammalian microRNA targets. *Nat Methods*, **3**, 881-886.
79. Hofacker, I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res*, **31**, 3429-3431.
80. Rehmsmeier, M., Steffen, P., Hochsmann, M. and Giegerich, R. (2004) Fast and effective prediction of microRNA/target duplexes. *Rna*, **10**, 1507-1517.
81. Kiriakidou, M., Nelson, P.T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z. and Hatzigeorgiou, A. (2004) A combined computational-experimental approach predicts human microRNA targets. *Genes Dev*, **18**, 1165-1178.
82. Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P. and Burge, C.B. (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787-798.
83. Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. and Segal, E. (2007) The role of site accessibility in microRNA target recognition. *Nat Genet*, **39**, 1278-1284.
84. Krek, A., Grun, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M. *et al.* (2005) Combinatorial microRNA target predictions. *Nat Genet*, **37**, 495-500.
85. Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C. and Marks, D.S. (2003) MicroRNA targets in Drosophila. *Genome Biol*, **5**, R1.
86. Kleihues, P. and Cavenee, W.K. (2000) *Pathology and genetics of tumours of the nervous system*. IARCPress, Lyon.
87. Ichimura, K., Ohgaki, H., Kleihues, P. and Collins, V.P. (2004) Molecular pathogenesis of astrocytic tumours. *Journal of Neuro-Oncology*, **70**, 137-160.
88. Freije, W.A., Castro-Vargas, F.E., Fang, Z., Horvath, S., Cloughesy, T., Liau, L.M., Mischel, P.S. and Nelson, S.F. (2004) Gene Expression Profiling of Gliomas Strongly Predicts Survival. *Cancer Res*, **64**, 6503-6510.
89. Nutt, C.L., Mani, D.R., Betensky, R.A., Tamayo, P., Cairncross, J.G., Ladd, C., Pohl, U., Hartmann, C., McLaughlin, M.E., Batchelor, T.T. *et al.* (2003) Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res*, **63**, 1602-1607.
90. Liang, Y., Diehn, M., Watson, N., Bollen, A.W., Aldape, K.D., Nicholas, M.K., Lamborn, K.R., Berger, M.S., Botstein, D., Brown, P.O. *et al.* (2005) Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme. *Proc Natl Acad Sci U S A*, **102**, 5814-5819.
91. Kim, S., Dougherty, E.R., Shmulevich, I., Hess, K.R., Hamilton, S.R., Trent, J.M., Fuller, G.N. and Zhang, W. (2002) Identification of Combination Gene Sets for Glioma Classification. *Mol Cancer Ther*, **1**, 1229-1236.
92. Rickman, D.S., Bobek, M.P., Misek, D.E., Kuick, R., Blaivas, M., Kurnit, D.M., Taylor, J. and Hanash, S.M. (2001) Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis. *Cancer Res*, **61**, 6885-6891.
93. Phillips, H.S., Kharbanda, S., Chen, R., Forrest, W.F., Soriano, R.H., Wu, T.D., Misra, A., Nigro, J.M., Colman, H., Soroceanu, L. *et al.* (2006) Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell*, **9**, 157-173.

94. Ekstrand, A.J., James, C.D., Cavenee, W.K., Seliger, B., Pettersson, R.F. and Collins, V.P. (1991) Genes for epidermal growth factor receptor, transforming growth factor alpha, and epidermal growth factor and their expression in human gliomas in vivo. *Cancer Res*, **51**, 2164-2172.
95. Livak, K.J. and Schmittgen, T.D. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods*, **25**, 402-408.
96. Oehler, M.K., Fischer, D.C., Orlowska-Volk, M., Herrle, F., Kieback, D.G., Rees, M.C. and Bicknell, R. (2003) Tissue and plasma expression of the angiogenic peptide adrenomedullin in breast cancer. *Br J Cancer*, **89**, 1927-1933.
97. Sharif, A., Renault, F., Beuvon, F., Castellanos, R., Canton, B., Barbeito, L., Junier, M.P. and Chneiweiss, H. (2004) The expression of PEA-15 (phosphoprotein enriched in astrocytes of 15 kDa) defines subpopulations of astrocytes and neurons throughout the adult mouse brain. *Neuroscience*, **126**, 263-275.
98. Ihaka, R. and Gentleman, R. (1996) R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, **5**, 299-314.
99. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, **5**, R80.
100. Barrett, T., Suzek, T.O., Troup, D.B., Wilhite, S.E., Ngau, W.C., Ledoux, P., Rudnev, D., Lash, A.E., Fujibuchi, W. and Edgar, R. (2005) NCBI GEO: mining millions of expression profiles--database and tools. *Nucleic Acids Res*, **33 Database Issue**, D562-566.
101. Hosack, D., Dennis, G., Sherman, B., Lane, H. and Lempicki, R. (2003) Identifying biological themes within lists of genes with EASE. *Genome Biology*, **4**, R70.
102. Golub, T.R., Slonim, K.D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A. *et al.* (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, **286**, 531-537.
103. Hochberg, Y. and Benjamini, Y. (1990) More powerful procedures for multiple significance testing. *Stat Med*, **9**, 811-818.
104. Linder, R., Dew, D., Sudhoff, H., Theegarten D., Remberger, H., Poppl, S., J., and Wagner, M. (2004) The "Subsequent Artificial Neural Network" (SANN) Approach Might Bring More Classification Power To ANN-based Microarray Analysis. *BMC Bioinformatics*, **20**, 3544-3552.
105. Xu, Y., Selaru, F., M., , Yin, J., Zou, T., T., , Shustiva, V., Mori, Y., Sato, F., Liu, T., C., , Olaru, A., wang, S. *et al.* (2002) Artificial Neural networks and Gene Filtering Distinguishes Between Global Gene Expression Profiles of Barrett's Esophagus and Esophageal Cancer. *Cancer Research*, **62**, 3493-3497.
106. Kaplan E, M.P. (1958) Nonparametric estimation from incomplete observations. *J AmStat Assoc*, 457–481.
107. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.

108. Saeed, A., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M. *et al.* (2003) TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, **34**, 374-378.
109. Renault, F., Formstecher, E., Callebaut, I., Junier, M.-P. and Chneiweiss, H. (2003) The multifunctional protein PEA-15 is involved in the control of apoptosis and cell cycle in astrocytes. *Biochemical Pharmacology*, **66**, 1581-1588.
110. Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, **4**, 406-425.
111. Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, **95**, 14863-14868.
112. Ichimura, K., Bolin, M.B., Goike, H.M., Schmidt, E.E., Moshref, A. and Collins, V.P. (2000) Deregulation of the p14ARF/MDM2/p53 pathway is a prerequisite for human astrocytic gliomas with G1-S transition control gene abnormalities. *Cancer Res*, **60**, 417-424.
113. Reifenberger, G., Reifenberger, J., Ichimura, K., Meltzer, P.S. and Collins, V.P. (1994) Amplification of multiple genes from chromosomal region 12q13-14 in human malignant gliomas: preliminary mapping of the amplicons shows preferential involvement of CDK4, SAS, and MDM2. *Cancer Res*, **54**, 4299-4303.
114. Ichimura, K., Schmidt, E.E., Goike, H.M. and Collins, V.P. (1996) Human glioblastomas with no alterations of the CDKN2A (p16INK4A, MTS1) and CDK4 genes have frequent mutations of the retinoblastoma gene. *Oncogene*, **13**, 1065-1072.
115. Liu, L., Ichimura, K., Pettersson, E.H., Goike, H.M. and Collins, V.P. (2000) The complexity of the 7p12 amplicon in human astrocytic gliomas: detailed mapping of 246 tumors. *J Neuropathol Exp Neurol*, **59**, 1087-1093.
116. Schmidt, E.E., Ichimura, K., Goike, H.M., Moshref, A., Liu, L. and Collins, V.P. (1999) Mutational profile of the PTEN gene in primary human astrocytic tumors and cultivated xenografts. *J Neuropathol Exp Neurol*, **58**, 1170-1183.
117. Burger, P.C. and Scheithauer, B.W. (1994) *Atlas of Tumor Pathology*. Armed Forces Institute of Pathology, Washington, D.C.
118. Tso, C.L., Freije, W.A., Day, A., Chen, Z., Merriman, B., Perlina, A., Lee, Y., Dia, E.Q., Yoshimoto, K., Mischel, P.S. *et al.* (2006) Distinct transcription profiles of primary and secondary glioblastoma subgroups. *Cancer Res*, **66**, 159-167.
119. Ho, D.M., Hsu, C.Y., Ting, L.T. and Chiang, H. (2003) MIB-1 and DNA topoisomerase IIa could be helpful for predicting long-term survival of patients with glioblastoma. *Am J Clin Pathol*, **119**, 715–722.
120. Hsu, S.C., Volpert, O.V., Steck, P.A., Mikkelsen, T., Polverini, P.J., Rao, S., Chou, P. and Bouck, N.P. (1996) Inhibition of angiogenesis in human glioblastomas by chromosome 10 induction of thrombospondin-1. *Cancer Res* **56**, 5684–5691.
121. Osada, H., Tokunaga, T., Nishi, M., Hatanaka, H., Abe, Y., Tsugu, A., Kijima, H., Yamazaki, H., Ueyama, Y. and Nakamura, M. (2004) Overexpression of the neuropilin 1 (NRP1) gene correlated with poor prognosis in human glioma. *Anticancer Res*, **24**, 547–552.

122. Godard, S., Getz, G., Delorenzi, M., Farmer, P., Kobayashi, H., Desbaillets, I., Nozaki, M., Diserens, A.C., Hamou, M.F. and Dietrich, P.Y. (2003) Classification of human astrocytic gliomas on the basis of gene expression: a correlated group of genes with angiogenic activity emerges as a strong predictor of subtypes. *Cancer Res*, **63**, 6613–6625.
123. Gaumont-Leclerc, M.F., Mukhopadhyay, U.K., Goumard, S. and Ferbeyre, G. (2004) PEA-15 is inhibited by adenovirus E1A and plays a role in ERK nuclear export and Ras-induced senescence. *J Biol Chem*, **279**, 46802-46809.
124. Benes, L., Kappus, C., McGregor, G.P., Bertalanffy, H., Mennel, H.D. and Hagner, S. (2004) The immunohistochemical expression of calcitonin receptor-like receptor (CRLR) in human gliomas. *J Clin Pathol*, **57**, 172-176.
125. Lee, R.C., Feinbaum, R.L. and Ambros, V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, **75**, 843–854.
126. FANTOM, C. (2005) The Transcriptional Landscape of the Mammalian Genome. *Science*, **309**, 1559 -1563.
127. Khvorova, A., Reynolds, A. and Jayasena, S.D. (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115**, 209–216.
128. Lee, Y., Jeon, K., Lee, J.T., Kim, S. and Kim, V.N. (2002) MicroRNA maturation: Stepwise processing and subcellular localization. *Embo Journal*, **21**, 4663–4670.
129. Hertel, J. and Stadler, P.F. (2006) Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics*, **22**, e197-202.
130. Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A. and Yekta, S. (2003b) The microRNAs of Caenorhabditis elegans. *Genes Dev* **16**, 991–1008.
131. Sewer, A., Paul, N., Landgraf, P., Aravin, A., Pfeffer, S., Brownstein, M.J., Tuschl, T., van Nimwegen, E. and Zavolan, M. (2005) Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics*, **6**, 267-281.
132. Lai, E.C., Tomancak, P., Williams, R.W. and Rubin, G.M. (2003) Computational identification of Drosophila microRNA genes. *Genome Biol*, **4**, R42-61.
133. Weber, M.J. (2005) New human and mouse microRNA genes found by homology search. *FEBS J.*, **272**, 59–73.
134. Legendre, M., Lambert, A. and Gautheret, D. (2004) Profile-based detection of microRNA precursors in animal genomes. *Bioinformatics*, **21**, 841–845.
135. Wang, X., Zhang., J., Li., F., Gu., J., He., T., Zhang., X. and Li., Y. (2005) MicroRNA identification based on sequence and structure alignment. *Bioinformatics*, **21**, 3610–3614.
136. Xue, C., Li, F., He, T., Liu, G.P., Li, Y. and Zhang, X. (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, **6**, 310-316.
137. Nam, J.W., Shin, K.R., Han, J., Lee, Y., Kim, V.N. and Zhang, B.T. (2005) Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acid Res*, **33**, 3570-3581.
138. Terai, G., Komori, T., Asai, K. and Kin, T. (2007) miRRim: a novel system to find conserved miRNAs with high sensitivity and specificity. *Rna*, **13**, 2081-2090.

139. Friedlander, M.R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S. and Rajewsky, N. (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol*, **26**, 407-415.
140. Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermuller, J., Hofacker, I.L. *et al.* (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, **316**, 1484-1488.
141. Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A.O., Landthaler, M. *et al.* (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**, 1401-1414.
142. Sassen, S., Miska, E.A. and Caldas, C. (2008) MicroRNA: implications for cancer. *Virchows Arch*, **452**, 1-10.
143. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol*, **215**, 403-410.
144. Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*, **14**, 708-715.
145. Eddy, S., R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755-763.
146. Oulas, A., Boutla, A., Gkirtzou, K., Reczko, M., Kalantidis, K. and Poirazi, P. (2009) Prediction of novel microRNA genes in cancer-associated genomic regions--a combined computational and experimental approach. *Nucleic Acids Res*.
147. Kent, W.J. (2002) BLAT--the BLAST-like alignment tool. *Genome Res*, **12**, 656-664.
148. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. and Tuschl, T. (2001) Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853-858.
149. Lee, R.C. and Ambros, V. (2001) An extensive class of small RNAs in Caenorhabditis elegans. *Science*, **294**, 862-864.
150. Nam, J.W., Kim, J., Kim, S.K. and Zhang, B.T. (2006) ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs. *Nucleic Acids Res*, **34**, W455-458.
151. Miranda, K.C., Huynh, T., Tay, Y., Ang, Y.S., Tam, W.L., Thomson, A.M., Lim, B. and Rigoutsos, I. (2006) A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell*, **126**, 1203-1217.
152. Ambros, V., Bartel, B., Bartel, D.P., Burge, C.B., Carrington, J.C., Chen, X., Dreyfuss, G., Eddy, S.R., Griffiths-Jones, S., Marshall, M. *et al.* (2003) A uniform system for microRNA annotation. *Rna*, **9**, 277-279.
153. Vasudevan, S., Tong, Y. and Steitz, J.A. (2007) Switching from repression to activation: microRNAs can up-regulate translation. *Science*, **318**, 1931-1934.
154. Lewis, B.P., Burge, C.B. and Bartel, D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15-20.
155. Rusinov, V., Baev, V., Minkov, I.N. and Tabler, M. (2005) MicroInspector: a web tool for detection of miRNA binding sites in an RNA sequence. *Nucleic Acids Res*, **33**, W696-700.

156. John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C. and Marks, D.S. (2004) Human MicroRNA targets. *PLoS Biol*, **2**, e363.
157. Ma, L., Teruya-Feldstein, J. and Weinberg, R.A. (2007) Tumour invasion and metastasis initiated by microRNA-10b in breast cancer. *Nature*, **449**, 682-688.
158. Mayr, C., Hemann, M.T. and Bartel, D.P. (2007) Disrupting the pairing between let-7 and Hmga2 enhances oncogenic transformation. *Science*, **315**, 1576-1579.
159. Sylvestre, Y., De Guire, V., Querido, E., Mukhopadhyay, U.K., Bourdeau, V., Major, F., Ferbeyre, G. and Chartrand, P. (2007) An E2F/miR-20a autoregulatory feedback loop. *J Biol Chem*, **282**, 2135-2143.
160. Krutzfeldt, J., Rajewsky, N., Braich, R., Rajeev, K.G., Tuschl, T., Manoharan, M. and Stoffel, M. (2005) Silencing of microRNAs in vivo with 'antagomirs'. *Nature*, **438**, 685-689.

Παράρτημα

Πίνακας Α. Results from the propagation of 26 test samples through each of our 3 trained model types A, B and C. Individual model and committee votes are shown for each sample. The committee vote was taken as the vote that the majority of the models concluded upon. Each type of model placed each test sample in one of the two tumour grades it had been trained to recognize. For example, type A models gave a score <0.5 if the sample in question had an expression profile similar to *ANGIO* (GB) and a score >0.5 if the sample's expression profile was closer to *DIFFER* (AA + A). Only *DIFFER* samples received a follow-up grading to differentiate the *INTER* (AA) and the *LOWER* (A) subtypes. Samples were assigned the *LOWER* subtype only if both *LOWER* trained models (B and C) graded them as *LOWER*, otherwise they were assigned to the INTER subtype. The highlighted row corresponds to the single tumour whose histopathological diagnosis and ANN grading did not concur.

Test tumour ID	GB vs AA Model A	Model A Voting	AA vs A Model B	Model B Voting	GB vs A Model C	Model C Voting	**Committee Vote**	**His/path Grade**
GB135	0.001	ANGIO	-	-	-	-	ANGIO	**GB**
GB136	0.0168	ANGIO	-	-	-	-	ANGIO	**GB**
GB81	0.0092	ANGIO	-	-	-	-	ANGIO	**GB**
GB82	0.1688	ANGIO	-	-	-	-	ANGIO	**GB**
GB84	0.0379	ANGIO	-	-	-	-	ANGIO	**GB**
GB87	0.0001	ANGIO	-	-	-	-	ANGIO	**GB**
GB44	0.0524	ANGIO	-	-	-	-	ANGIO	**GB**
GB154	0.9771	DIFFER	0.0508	INTER	0.0671	INTER	INTER	**GB**
GB153	0.0092	ANGIO	-	-	-	-	ANGIO	**GB**
GB126	0.06	ANGIO	-	-	-	-	ANGIO	**GB**
GB35	0.043	ANGIO	-	-	-	-	ANGIO	**GB**
GB50	0.0135	ANGIO	-	-	-	-	ANGIO	**GB**
GB49	0.0025	ANGIO	-	-	-	-	ANGIO	**GB**
GB130	0.0029	ANGIO	-	-	-	-	ANGIO	**GB**
GB238	0.017	ANGIO	-	-	-	-	ANGIO	**GB**
GB103	0.0099	ANGIO	-	-	-	-	ANGIO	**GB**
GB3	0.001	ANGIO	-	-	-	-	ANGIO	**GB**
GB101	0.0014	ANGIO	-	-	-	-	ANGIO	**GB**
GB245	0.381	ANGIO	-	-	-	-	ANGIO	**GB**
A36	0.9729	DIFFER	0.8113	LOWER	0.7479	LOWER	LOWER	**A**
A30	0.9986	DIFFER	0.9153	LOWER	0.8828	LOWER	LOWER	**A**
AA105	1.0	DIFFER	0.0172	INTER	0.539	LOWER	INTER	**AA**
AA106	0.9987	DIFFER	0.7499	LOWER	0.361	INTER	INTER	**AA**
AA15	0.9999	DIFFER	0.3098	INTER	0.8289	LOWER	INTER	**AA**
AA14	0.9512	DIFFER	0.053	INTER	0.3594	INTER	INTER	**AA**
AA13	0.9754	DIFFER	0.1722	INTER	0.3444	INTER	INTER	**AA**

Malignancy grading of additional samples difficult to grade histopathologically

Πίνακας Β. ANN grading of the 6 additional samples that proved to be difficult to grade by histopathological analysis. These include the two Piloytic Astrocytoma tumour samples (PA68 and PA67) as well as the 4 difficult Anaplastic Astrocytoma samples (AA29, AA86, AA93 and AA49). Sample PA67 was assigned to a different tumour grade by each of the models, but since the two *LOWER* trained model did not agree in their grading it was assigned to the *INTER* subtype.

Test tumour ID	GB vs AA Model A	Model A Voting	AA vs A Model B	Model B Voting	GB vs A Model C	Model C Voting	**Committee Vote**	**Hist/path Class**
PA68	0.0428	ANGIO	0.0834	-	0.0056	-	ANGIO	**PA**
PA67	0.7619	DIFFER	0.723	LOWER	0.0067	INTER	INTER	**PA**
AA29	0.0025	ANGIO	0.134	-	0.1916	-	ANGIO	**AA**
AA86	0.0621	ANGIO	0.1451	-	0.0086	-	ANGIO	**AA**
AA93	0.0041	ANGIO	0.0052	-	0.0072	-	ANGIO	**AA**
AA49	0.0253	ANGIO	0.2941	-	0.0384	-	ANGIO	**AA**

Grading of independent dataset using across-array gene classifiers

Πίνακας C. Independent data set analysis of 23 test samples from the Shai et al, 2003(28) dataset using model types 1 and 2 (see text). Individual model and overall voting is shown for each sample. Type 1 models gave a score <0.5 if the sample in question had an expression profile similar to *ANGIO* (GB) and a score >0.5 if the sample's expression profile was closer to *DIFFER* (lower grades II and III). Similarly; type 2 models score tumours resembling *INTER* (AA) with a value <0.5 and tumours resembling *LOWER* (A) with a value >0.5. Highlighted rows correspond to test samples with overall votes that did not agree with original histopathological diagnosis. The gene names of the cross–chip gene classifiers are TYPE1 Genes: LDHA, PEA15, LGALS1, TIMP1, PLAT, ZMYND11, EMP3, NAP1L3, USH1C, DAG1, PDGFA, LGALS3, HNRPH3, CRYAB, EFEMP2, KIAA1279, RPL22, PDPN, TYPE2 genes: SCP2, B2M

Test tumour Index	ANGIO/ DIFFER Model 1	Model 1 Voting	INTER / LOWER Model 2	Model 2 Voting	**Overall Vote**	**His/path Grade**
1	0.4746	ANGIO	-	-	ANGIO	**GB**
2	0.998	DIFFER	0.2283	INTER	INTER	**AA**
3	0.4486	ANGIO	-	-	ANGIO	**GB**
4	0.9967	DIFFER	0.045	INTER	INTER	**GB**
5	0.7472	DIFFER	0.037	INTER	INTER	**GB**
6	0.0209	ANGIO	-	-	ANGIO	**GB**
7	0.9961	DIFFER	0.5934	LOWER	LOWER	**A**
8	0.0113	ANGIO	-	-	ANGIO	**GB**
9	0.0165	ANGIO	-	-	ANGIO	**GB**
10	0.0106	ANGIO	-	-	ANGIO	**GB**
11	0.0847	ANGIO	-	-	ANGIO	**GB**
12	0.0133	ANGIO	-	-	ANGIO	**GB**
13	0.0229	ANGIO	-	-	ANGIO	**GB**
14	0.1739	ANGIO	-	-	ANGIO	**GB**
15	0.0249	ANGIO	-	-	ANGIO	**GB**
16	0.0044	ANGIO	-	-	ANGIO	**GB**
17	0.0207	ANGIO	-	-	ANGIO	**GB**
18	0.9838	DIFFER	0.8986	LOWER	LOWER	**A**
19	0.9936	DIFFER	0.0556	INTER	INTER	**AA**
20	0.0573	ANGIO	-	-	ANGIO	**GB**
21	0.9999	DIFFER	0.4409	INTER	INTER	**AA**
22	0.0655	ANGIO	-	-	ANGIO	**GB**
23	0.1514	ANGIO	-	-	ANGIO	**GB**

Survival Analysis using independent datasets and across-array gene classifiers

Πίνακας D. Independent data set analysis of 65 samples from the Freije et al, 2004 (88) dataset using ANN models trained with our 59 gene classifiers.(see text). The initial grading of the models classified the 65 tumour samples into two groups (*ANGIO* and *DIFFER*), which, resembled the "Survival Clusters" (SC1 and SC2) obtained in the Freije study with 86.15% similarity.

ANGIO/ DIFFER	**INTER / LOWER**	**INTER / LOWER**	**Survival Days**	**Freije subtyping**	**ANN subtypeing**	**Histo/Path**
0	-	-	1089	SC2	ANGIO	GBM 1043
0	-	-	420	SC2	ANGIO	GBM 1354
0	-	-	293	SC2	ANGIO	GBM 1398
0	-	-	153	SC2	ANGIO	GBM 1469
0	-	-	1031	SC2	ANGIO	GBM 1516
0	-	-	683	SC2	ANGIO	GBM 1675
0	-	-	723	SC2	ANGIO	GBM 1798
0	-	-	325	SC2	ANGIO	GBM 1902
0	-	-	396	SC2	ANGIO	GBM 2013

0	-	-	396	SC2	ANGIO	GBM 2015
0	-	-	298	SC2	ANGIO	GBM 2079
0	-	-	203	SC2	ANGIO	GBM 2098
0	-	-	7	SC2	ANGIO	GBM 597
0	-	-	185	SC2	ANGIO	GBM 604
0	-	-	112	SC2	ANGIO	GBM 660
0	-	-	356	SC2	ANGIO	GBM 697
0	-	-	188	SC2	ANGIO	GBM 712
0	-	-	53	SC2	ANGIO	GBM 749
0	-	-	182	SC2	ANGIO	GBM 931
0	-	-	418	SC2	ANGIO	GBM 976
0	-	-	443	SC2	ANGIO	MIXED III 799
0.01	-	-	71	SC2	ANGIO	GBM 1495
0.01	-	-	588	SC2	ANGIO	GBM 1667
0.01	-	-	279	SC2	ANGIO	GBM 1900
0.01	-	-	237	SC2	ANGIO	GBM 2017
0.02	-	-	95	SC2	ANGIO	GBM 2158
0.05	-	-	389	SC2	ANGIO	GBM 1905
0.05	-	-	302	SC2	ANGIO	GBM 585
0.05	-	-	412	SC2	ANGIO	GBM 636
0.06	-	-	64	SC2	ANGIO	GBM 1414
0.06	-	-	186	SC2	ANGIO	GBM 824
0.08	-	-	224	SC2	ANGIO	GBM 1342
0.24	-	-	90	SC2	ANGIO	GBM 1032
0.25	-	-	126	SC2	ANGIO	GBM 1022
0.25	-	-	56	SC2	ANGIO	GBM 995
0.33	-	-	927	SC1	ANGIO	GBM 1681
0.4	-	-	96	SC1	ANGIO	GBM 1423
0.43	-	-	506	SC1	ANGIO	GBM 706
0.49	-	-	43	SC2	ANGIO	GBM 746
0.54	0.53	0.14	98	SC2	INTER	GBM 932
0.7	0.41	0.25	877	SC2	INTER	ASTRO III 1704
0.71	0.67	0.27	140	SC2	INTER	GBM 782
0.74	0.76	0.2	961	SC1	INTER	GBM 1656
0.82	0.95	0.06	286	SC1	INTER	GBM 839
0.96	0.81	0.11	569	SC1	INTER	GBM 938
0.98	0.36	0.32	1098	SC1	INTER	GBM 2166
0.99	0.78	0.36	1114	SC1	INTER	ASTRO III 1425
0.99	0.47	0.1	85	SC2	INTER	GBM 1511
0.99	0.6	0.3	664	SC1	INTER	MIXED III 886
1	0.36	0.34	858	SC1	INTER	ASTRO III 1723
1	0.52	0.36	2516	SC1	INTER	ASTRO III 587
1	0.42	0.4	2125	SC1	INTER	ASTRO III 671
1	0.23	0.22	1557	SC1	INTER	ASTRO III 672
1	0.84	0.35	1830	SC1	INTER	ASTRO III 747
1	0.49	0.09	1247	SC1	INTER	GBM 1038
1	0.48	0.64	1088	SC1	INTER	GBM 1478
1	0.54	0.36	1022	SC1	INTER	GBM 1521
1	0.29	0.27	780	SC2	INTER	GBM 1745
1	0.93	0.39	223	SC2	INTER	GBM 2028
1	0.8	0.38	861	SC1	INTER	MIXED III 1721
1	0.6	0.1	2185	SC1	INTER	MIXED III 664

0.99	0.86	0.53	442	SC1	LOWER	ASTRO III 659
1	0.71	0.61	850	SC1	LOWER	MIXED III 615
1	0.7	0.82	1918	SC1	LOWER	MIXED III 713
1	0.57	0.7	1474	SC1	LOWER	MIXED III 912

Πίνακας Ε. Analysis of Phillips et 2006 samples

ANGIO/ DIFFER	INTER/ LOWER	INTER/ LOWER	SURVIVAL WEEKS	PHILLIPS SUBTYPEING	ANN SUBTYPEING	Histo/Path
0.14	-	-		*Mes*	ANGIO	IV
0.15	-	-	51	*Prolif*	ANGIO	IV with necrosis
0.15	-	-	65	*Mes*	ANGIO	IV with necrosis
0.16	-	-	59	*Mes*	ANGIO	IV with necrosis
0.16	-	-	70	*Prolif*	ANGIO	IV with necrosis
0.17	-	-		*Mes*	ANGIO	IV
0.19	-	-		*Prolif*	ANGIO	IV
0.19	-	-		*Mes*	ANGIO	IV
0.19	-	-		*Prolif*	ANGIO	IV
0.2	-	-	32	*Prolif*	ANGIO	IV with necrosis
0.2	-	-	12	*Prolif*	ANGIO	IV with necrosis
0.21	-	-	236	*Mes*	ANGIO	IV with necrosis
0.21	-	-	16	*Prolif*	ANGIO	IV with necrosis
0.21	-	-	154	*Mes*	ANGIO	IV with necrosis
0.21	-	-	47	*Mes*	ANGIO	III
0.22	-	-	181	*Mes*	ANGIO	IV with necrosis
0.22	-	-	55	*Prolif*	ANGIO	IV with necrosis
0.22	-	-	33	*Mes*	ANGIO	IV with necrosis
0.23	-	-	77	*Mes*	ANGIO	IV with necrosis
0.23	-	-	3	*Mes*	ANGIO	IV with necrosis
0.24	-	-	95	*Prolif*	ANGIO	IV with necrosis
0.24	-	-		*Mes*	ANGIO	IV
0.25	-	-	313	*Mes*	ANGIO	IV with necrosis
0.25	-	-		*Mes*	ANGIO	IV
0.26	-	-	131	*Mes*	ANGIO	IV with necrosis
0.26	-	-	56	*Mes*	ANGIO	IV with necrosis
0.27	-	-	62	*Prolif*	ANGIO	IV with

						necrosis
0.27	-	-	210	*Prolif*	ANGIO	IV with necrosis
0.27	-	-	41	*Prolif*	ANGIO	IV with necrosis
0.28	-	-	311	*Mes*	ANGIO	IV with necrosis
0.28	-	-		*Mes*	ANGIO	IV
0.28	-	-		*Mes*	ANGIO	IV
0.29	-	-	79	*Prolif*	ANGIO	IV with necrosis
0.29	-	-		*Mes*	ANGIO	III
0.3	-	-	32	*Prolif*	ANGIO	IV with necrosis
0.3	-	-	62	*Mes*	ANGIO	IV with necrosis
0.31	-	-	238	*Prolif*	ANGIO	IV with necrosis
0.31	-	-	91	*Prolif*	ANGIO	IV with necrosis
0.32	-	-		*Prolif*	ANGIO	IV
0.33	-	-	59	*Prolif*	ANGIO	IV with necrosis
0.34	-	-	57	*Prolif*	ANGIO	IV with necrosis
0.34	-	-	53	*Mes*	ANGIO	IV without necrosis
0.35	-	-	131	*Mes*	ANGIO	IV with necrosis
0.35	-	-	106	*Mes*	ANGIO	IV with necrosis
0.35	-	-	53	*Mes*	ANGIO	IV without necrosis
0.36	-	-		*Mes*	ANGIO	IV
0.37	-	-	70	*Prolif*	ANGIO	IV with necrosis
0.4	-	-	53	*Mes*	ANGIO	IV with necrosis
0.42	-	-	17	*PN*	ANGIO	III
0.43	-	-		*Mes*	ANGIO	IV
0.44	-	-	34	*PN*	ANGIO	IV with necrosis
0.47	-	-		*Mes*	ANGIO	IV
0.48	-	-	62	*PN*	ANGIO	IV with necrosis
0.49	-	-	46	*PN*	ANGIO	III
0.51	0.58	0.24	125	*Mes*	INTER	IV with necrosis
0.52	0.11	0.86	111	*Prolif*	INTER	IV with necrosis
0.54	0.09	0.1	52	*Prolif*	INTER	IV with necrosis
0.56	0.3	0.56	467	*PN*	INTER	III
0.56	0.26	0.69	460	*PN*	INTER	III
0.58	0.28	0.31	39	*Prolif*	INTER	IV with necrosis
0.6	0.18	0.39	32	*Mes*	INTER	IV with

						necrosis
0.61	0.16	0.88	57	*Mes*	INTER	IV with necrosis
0.62	0.19	0.24	300	*PN*	INTER	III
0.64	0.17	0.13	97	*Mes*	INTER	IV with necrosis
0.65	0.14	0.8		*Prolif*	INTER	IV
0.65	0.13	0.35	242	*Prolif*	INTER	IV with necrosis
0.66	0.08	0.16	277	*PN*	INTER	IV without necrosis
0.66	0.28	0.15	145	*PN*	INTER	IV without necrosis
0.67	0.25	0.32	174	*PN*	INTER	III
0.7	0.14	0.8	33	*Prolif*	INTER	IV with necrosis
0.73	0.58	0.54		*PN*	LOWER	IV
0.73	0.07	0.54		*Mes*	INTER	III
0.76	0.34	0.28	86	*PN*	INTER	IV
0.76	0.77	0.99	383	*PN*	LOWER	IV without necrosis
0.79	0.36	0.54		*PN*	INTER	III
0.79	0.4	0.29	73	*PN*	INTER	III
0.8	0.16	0.2		*PN*	INTER	IV
0.8	0.18	0.31	108	*PN*	INTER	IV
0.82	0.31	0.63		*PN*	INTER	III
0.87	0.91	0.9		*PN*	LOWER	IV
0.88	0.37	0.5		*PN*	INTER	IV with necrosis
0.89	0.52	0.37	175	*PN*	INTER	IV with necrosis
0.89	0.9	0.85	146	*PN*	LOWER	III
0.9	0.65	0.85	477	*PN*	LOWER	III
0.92	0.36	0.18		*PN*	INTER	III
0.93	0.27	0.66	97	*Prolif*	INTER	IV with necrosis
0.95	0.65	0.97	150	*PN*	LOWER	III
0.96	0.16	0.25	123	*Prolif*	INTER	IV without necrosis
0.96	0.43	0.21	81	*PN*	INTER	IV
0.97	0.91	0.62	203	*PN*	LOWER	III
0.97	0.52	0.19	316	*PN*	INTER	III
0.97	0.65	0.93	445	*PN*	LOWER	IV with necrosis
0.97	0.55	0.12	41	*PN*	INTER	III
0.98	0.44	0.59	210	*PN*	INTER	IV with necrosis
0.98	0.72	0.8	325	*PN*	LOWER	IV with necrosis
0.99	0.97	0.64	357	*PN*	LOWER	III
0.99	0.55	0.17	102	*PN*	INTER	III
0.99	0.7	0.57	115	*PN*	LOWER	III
0.99	0.76	0.71	244	*PN*	LOWER	III
0.99	0.39	0.21	322	*PN*	INTER	III

The jar file for the ANN can be downloaded at:
http://www.imbb.forth.gr/people/poirazi/software.html

Πίνακας F. Gene/probe names and annotation for 11 genes used for primary and secondary tumour differentiation.

203353_s_at "gb:NM_015846.1 /DEF=Homo sapiens methyl-CpG binding domain protein 1 (MBD1), transcript variant 1, mRNA. /FEA=mRNA /GEN=MBD1 /PROD=methyl-CpG binding domain protein 1, isoform 1 /DB_XREF=gi:7710138 /UG=Hs.6211 methyl-CpG binding domain protein 1 /FL=gb:AF078830.1 gb:NM_015846.1"
203597_s_at Consensus includes gb:AI734228 /FEA=EST /DB_XREF=gi:5055341 /DB_XREF=est:zb57d10.y5 /CLONE=IMAGE:307699 /UG=Hs.28307 WW domain binding protein 4 (formin binding protein 21) /FL=gb:AF071185.1 gb:NM_007187.2
206611_at "gb:NM_013310.1 /DEF=Homo sapiens hypothetical protein (AF038169), mRNA. /FEA=mRNA /GEN=AF038169 /PROD=hypothetical protein /DB_XREF=gi:9558718 /UG=Hs.145567 hypothetical protein /FL=gb:AF038169.1 gb:NM_013310.1"
207254_at "gb:NM_005073.1 /DEF=Homo sapiens solute carrier family 15 (oligopeptide transporter), member 1 (SLC15A1), mRNA. /FEA=mRNA /GEN=SLC15A1 /PROD=solute carrier family 15 (oligopeptidetransporter), member 1 /DB_XREF=gi:4827007 /UG=Hs.2217 solute carrier family 15 (oligopeptide transporter), member 1 /FL=gb:AF043233.1 gb:NM_005073.1 gb:U21936.1 gb:U13173.1"
208973_at "gb:BC001072.1 /DEF=Homo sapiens, clone MGC:2683, mRNA, complete cds. /FEA=mRNA /PROD=Unknown (protein for MGC:2683) /DB_XREF=gi:12654484 /UG=Hs.151032 hypothetical protein MGC2683 /FL=gb:BC001072.1 gb:BC004456.1"
210896_s_at "gb:AF306765.1 /DEF=Homo sapiens junctate mRNA, complete cds. /FEA=mRNA /PROD=junctate /DB_XREF=gi:11991236 /UG=Hs.283664 aspartate beta-hydroxylase /FL=gb:AF306765.1"
212141_at Consensus includes gb:AA604621 /FEA=EST /DB_XREF=gi:2445485 /DB_XREF=est:no84b08.s1 /CLONE=IMAGE:1113495 /UG=Hs.154443 minichromosome maintenance deficient (S. cerevisiae) 4
212142_at Consensus includes gb:AI936566 /FEA=EST /DB_XREF=gi:5675436 /DB_XREF=est:wd29e06.x1 /CLONE=IMAGE:2329570 /UG=Hs.154443 minichromosome maintenance deficient (S. cerevisiae) 4
216028_at Consensus includes gb:AL049980.1 /DEF=Homo sapiens mRNA; cDNA DKFZp564C152 (from clone DKFZp564C152). /FEA=mRNA /GEN=DKFZp564C152 /PROD=hypothetical protein /DB_XREF=gi:4884230 /UG=Hs.184216 DKFZP564C152 protein
218692_at "gb:NM_017786.1 /DEF=Homo sapiens hypothetical protein FLJ20366 (FLJ20366), mRNA. /FEA=mRNA /GEN=FLJ20366 /PROD=hypothetical protein FLJ20366 /DB_XREF=gi:8923340 /UG=Hs.8358 hypothetical protein FLJ20366 /FL=gb:NM_017786.1"
219511_s_at "gb:NM_005460.1 /DEF=Homo sapiens synuclein, alpha interacting protein (synphilin) (SNCAIP), mRNA. /FEA=mRNA /GEN=SNCAIP /PROD=synuclein alpha interacting protein /DB_XREF=gi:4885602 /UG=Hs.24948 synuclein, alpha interacting protein (synphilin) /FL=gb:AF076929.1 gb:NM_005460.1"

Πίνακας G. Gene/probe names and annotations for 37 survival correlated genes.

201291_s_at Consensus includes gb:AU159942 /FEA=EST /DB_XREF=gi:11021463 /DB_XREF=est:AU159942 /CLONE=Y79AA1000724 /UG=Hs.156346 topoisomerase (DNA) II alpha (170kD) /FL=gb:J04088.1 gb:NM_001067.1
201664_at "gb:AL136877.1 /DEF=Homo sapiens mRNA; cDNA DKFZp434F205 (from clone DKFZp434F205); complete cds. /FEA=mRNA /GEN=DKFZp434F205 /PROD=hypothetical protein /DB_XREF=gi:6807670 /UG=Hs.50758 SMC4 (structural

maintenance of chromosomes 4, yeast)-like 1 /FL=gb:AB019987.1 gb:NM_005496.1 gb:AL136877.1"
201666_at "gb:NM_003254.1 /DEF=Homo sapiens tissue inhibitor of metalloproteinase 1 (erythroid potentiating activity, collagenase inhibitor) (TIMP1), mRNA. /FEA=mRNA /GEN=TIMP1 /PROD=tissue inhibitor of metalloproteinase 1precursor /DB_XREF=gi:4507508 /UG=Hs.5831 tissue inhibitor of metalloproteinase 1 (erythroid potentiating activity, collagenase inhibitor) /FL=gb:BC000866.1 gb:M12670.1 gb:M59906.1 gb:NM_003254.1"
201752_s_at Consensus includes gb:AI763123 /FEA=EST /DB_XREF=gi:5178790 /DB_XREF=est:wi06f09.x1 /CLONE=IMAGE:2389481 /UG=Hs.324470 adducin 3 (gamma) /FL=gb:D67031.1 gb:NM_019903.1
202912_at "gb:NM_001124.1 /DEF=Homo sapiens adrenomedullin (ADM), mRNA. /FEA=mRNA /GEN=ADM /PROD=adrenomedullin /DB_XREF=gi:4501944 /UG=Hs.394 adrenomedullin /FL=gb:NM_001124.1 gb:D14874.1"
203554_x_at "gb:NM_004219.2 /DEF=Homo sapiens pituitary tumor-transforming 1 (PTTG1), mRNA. /FEA=mRNA /GEN=PTTG1 /PROD=pituitary tumor-transforming protein 1 /DB_XREF=gi:11038651 /UG=Hs.252587 pituitary tumor-transforming 1 /FL=gb:NM_004219.2 gb:AF095287.1 gb:AF062649.1 gb:AF075242.1"
204713_s_at "Consensus includes gb:AA910306 /FEA=EST /DB_XREF=gi:3049596 /DB_XREF=est:ok83a04.s1 /CLONE=IMAGE:1520526 /UG=Hs.30054 coagulation factor V (proaccelerin, labile factor) /FL=gb:NM_000130.2 gb:M16967.1 gb:M14335.1"
205137_x_at "gb:NM_005709.1 /DEF=Homo sapiens PDZ-73 protein (PDZ-73NY-CO-38), mRNA. /FEA=mRNA /GEN=PDZ-73NY-CO-38 /PROD=PDZ-73 protein /DB_XREF=gi:5031978 /UG=Hs.132945 PDZ-73 protein /FL=gb:AF039700.1 gb:NM_005709.1 gb:AB018687.1"
209155_s_at "gb:BC001595.1 /DEF=Homo sapiens, 5-nucleotidase (purine), cytosolic type B, clone MGC:1109, mRNA, complete cds. /FEA=mRNA /PROD=5-nucleotidase (purine), cytosolic type B /DB_XREF=gi:12804388 /UG=Hs.138593 5-nucleotidase (purine), cytosolic type B /FL=gb:BC001595.1"
210512_s_at "gb:AF022375.1 /DEF=Homo sapiens vascular endothelial growth factor mRNA, complete cds. /FEA=mRNA /PROD=vascular endothelial growth factor /DB_XREF=gi:3719220 /UG=Hs.73793 vascular endothelial growth factor /FL=gb:M32977.1 gb:AF022375.1 gb:NM_003376.1 gb:AB021221.1 gb:AF091352.1"
211184_s_at "gb:AB006955.1 /DEF=Homo sapiens mRNA for AIE-75, complete cds. /FEA=mRNA /GEN=aie-75 /PROD=AIE-75 /DB_XREF=gi:5152287 /UG=Hs.132945 PDZ-73 protein /FL=gb:AB006955.1"
211527_x_at "gb:M27281.1 /DEF=Human vascular permeability factor mRNA, complete cds. /FEA=mRNA /DB_XREF=gi:340300 /UG=Hs.73793 vascular endothelial growth factor /FL=gb:M27281.1"
211964_at "Consensus includes gb:X05610.1 /DEF=Human mRNA for type IV collagen alpha (2) chain. /FEA=mRNA /PROD=alpha (2) chain /DB_XREF=gi:29550 /UG=Hs.75617 collagen, type IV, alpha 2"
211966_at "Consensus includes gb:AA909035 /FEA=EST /DB_XREF=gi:3048440 /DB_XREF=est:ol11h01.s1 /CLONE=IMAGE:1523185 /UG=Hs.75617 collagen, type IV, alpha 2"
211980_at "Consensus includes gb:AI922605 /FEA=EST /DB_XREF=gi:5658569 /DB_XREF=est:wm90c05.x1 /CLONE=IMAGE:2443208 /UG=Hs.119129 collagen, type IV, alpha 1 /FL=gb:NM_001845.1"
212171_x_at Consensus includes gb:H95344 /FEA=EST /DB_XREF=gi:1102977 /DB_XREF=est:yu21b08.s1 /CLONE=IMAGE:234423 /UG=Hs.73793 vascular endothelial growth factor /FL=gb:AF214570.1
213217_at Consensus includes gb:AU149572 /FEA=EST /DB_XREF=gi:11011093 /DB_XREF=est:AU149572 /CLONE=NT2RM4002598 /UG=Hs.2352 adenylate cyclase 2 (brain)
213388_at Consensus includes gb:H15535 /FEA=EST /DB_XREF=gi:880355 /DB_XREF=est:ym27c01.s1 /CLONE=IMAGE:49385 /UG=Hs.52792 Homo sapiens mRNA; cDNA DKFZp586I1823 (from clone DKFZp586I1823)
213433_at Consensus includes gb:AF038193.1 /DEF=Homo sapiens clone 23608 mRNA sequence. /FEA=mRNA /DB_XREF=gi:2795913 /UG=Hs.6220 Homo sapiens clone

23608 mRNA sequence
213689_x_at Consensus includes gb:AL137958 /FEA=EST /DB_XREF=gi:6854638 /DB_XREF=est:DKFZp761C1715_r1 /CLONE=DKFZp761C1715 /UG=Hs.180946 ribosomal protein L5
213900_at Consensus includes gb:AA524029 /FEA=EST /DB_XREF=gi:2264957 /DB_XREF=est:ng32f02.s1 /CLONE=IMAGE:936507 /UG=Hs.77889 Friedreich ataxia region gene X123
214930_at Consensus includes gb:AW449813 /FEA=EST /DB_XREF=gi:6990589 /DB_XREF=est:UI-H-BI3-akm-b-07-0-UI.s1 /CLONE=IMAGE:2734788 /UG=Hs.58009 KIAA0918 protein
216044_x_at "Consensus includes gb:AK027146.1 /DEF=Homo sapiens cDNA: FLJ23493 fis, clone LNG01831, highly similar to HSU66589 Human ribosomal protein L5 pseudogene mRNA. /FEA=mRNA /DB_XREF=gi:10440199 /UG=Hs.180946 ribosomal protein L5"
218891_at "gb:NM_024541.1 /DEF=Homo sapiens hypothetical protein FLJ13114 (FLJ13114), mRNA. /FEA=mRNA /GEN=FLJ13114 /PROD=hypothetical protein FLJ13114 /DB_XREF=gi:13375700 /UG=Hs.9444 hypothetical protein FLJ13114 /FL=gb:NM_024541.1"
219410_at "gb:NM_018004.1 /DEF=Homo sapiens hypothetical protein FLJ10134 (FLJ10134), mRNA. /FEA=mRNA /GEN=FLJ10134 /PROD=hypothetical protein FLJ10134 /DB_XREF=gi:8922242 /UG=Hs.104800 hypothetical protein FLJ10134 /FL=gb:NM_018004.1"
219978_s_at "gb:NM_018454.1 /DEF=Homo sapiens uncharacterized bone marrow protein BM037 (BM037), mRNA. /FEA=mRNA /GEN=BM037 /PROD=uncharacterized bone marrow protein BM037 /DB_XREF=gi:8922094 /UG=Hs.283649 uncharacterized bone marrow protein BM037 /FL=gb:AF217513.1 gb:NM_018454.1"
220005_at "gb:NM_023914.1 /DEF=Homo sapiens G protein-coupled receptor 86 (GPR86), mRNA. /FEA=mRNA /GEN=GPR86 /PROD=G protein-coupled receptor 86 /DB_XREF=gi:13194202 /UG=Hs.13040 G protein-coupled receptor 86 /FL=gb:AF295368.1 gb:NM_023914.1 gb:AF178982.1 gb:AF345565.1"
221011_s_at "gb:NM_030915.1 /DEF=Homo sapiens hypothetical protein DKFZp566J091 (DKFZP566J091), mRNA. /FEA=mRNA /GEN=DKFZP566J091 /PROD=hypothetical protein DKFZp566J091 /DB_XREF=gi:13569871 /FL=gb:NM_030915.1"
221362_at "gb:NM_024012.1 /DEF=Homo sapiens 5-hydroxytryptamine (serotonin) receptor 5A (HTR5A), mRNA. /FEA=CDS /GEN=HTR5A /PROD=5-hydroxytryptamine (serotonin) receptor 5A /DB_XREF=gi:13236496 /UG=Hs.248137 5-hydroxytryptamine (serotonin) receptor 5A /FL=gb:NM_024012.1"
221527_s_at "gb:AF196185.1 /DEF=Homo sapiens atypical PKC isotype-specific interacting protein long variant mRNA, complete cds. /FEA=mRNA /PROD=atypical PKC isotype-specific interactingprotein long variant /DB_XREF=gi:13491609 /UG=Hs.72249 three-PDZ containing protein similar to C. elegans PAR3 (partitioning defect) /FL=gb:AF196185.1"
221729_at "Consensus includes gb:AL575735 /FEA=EST /DB_XREF=gi:12937190 /DB_XREF=est:AL575735 /CLONE=CS0DI070YK23 (3 prime) /UG=Hs.82985 collagen, type V, alpha 2 /FL=gb:NM_000393.1"
203891_s_at "gb:NM_001348.1 /DEF=Homo sapiens death-associated protein kinase 3 (DAPK3), mRNA. /FEA=mRNA /GEN=DAPK3 /PROD=death-associated protein kinase 3 /DB_XREF=gi:4557510 /UG=Hs.25619 death-associated protein kinase 3 /FL=gb:AB007144.1 gb:NM_001348.1 gb:AB022341.1"
204563_at "gb:NM_000655.2 /DEF=Homo sapiens selectin L (lymphocyte adhesion molecule 1) (SELL), mRNA. /FEA=mRNA /GEN=SELL /PROD=selectin L /DB_XREF=gi:5713320 /UG=Hs.82848 selectin L (lymphocyte adhesion molecule 1) /FL=gb:M25280.1 gb:NM_000655.2"
206932_at "gb:NM_003956.1 /DEF=Homo sapiens cholesterol 25-hydroxylase (CH25H), mRNA. /FEA=mRNA /GEN=CH25H /PROD=cholesterol 25-hydroxylase /DB_XREF=gi:4502498 /UG=Hs.194687 cholesterol 25-hydroxylase /FL=gb:AF059214.1 gb:NM_003956.1"
212623_at Consensus includes gb:AU153138 /FEA=EST /DB_XREF=gi:11014659

/DB_XREF=est:AU153138 /CLONE=NT2RP3002507 /UG=Hs.174905 KIAA0033 protein
216150_at "Consensus includes gb:AK021609.1 /DEF=Homo sapiens cDNA FLJ11547 fis, clone HEMBA1002934. /FEA=mRNA /DB_XREF=gi:10432824 /UG=Hs.306605 Homo sapiens cDNA FLJ11547 fis, clone HEMBA1002934"
216984_x_at "Consensus includes gb:D84143.1 /DEF=Human immunoglobulin (mAb59) light chain V region mRNA, partial sequence. /FEA=mRNA /PROD=immunoglobulin light chain V-J region /DB_XREF=gi:1255613 /UG=Hs.121508 Human immunoglobulin (mAb59) light chain V region mRNA, partial sequence"

Πίνακας H. Tumour Annotations for our 65 Astrocytic Brain Tumours
(i) Metagenomic Data

Tumour identifier	Unique Identifier used by present study	Grade	CDKN2A/B	CDK4	RB1	MDM2	p53	EGFR	PTEN
A26	101	A	no loss	no AMP	wt	no AMP	MUT	no AMP	wt
A30	74	A	no loss	no AMP	wt	no AMP	wt	no AMP	wt
A36	71	A							
A48	3	A	no loss	no AMP	wt	no AMP	MUT	no AMP	wt
A53	51	A	no loss	no AMP		no AMP	wt	no AMP	
A9	16	A	no loss	no AMP	wt	no AMP	MUT	no AMP	wt
AA100	31	AA	no loss	no AMP	wt	no AMP	MUT	no AMP	wt
AA101	40	AA	no loss	no AMP	wt	no AMP	MUT	no AMP	wt
AA102	7	AA	Hemi	no AMP	wt	no AMP	wt	no AMP	wt
AA105	70	AA	no loss	no AMP	wt	no AMP	MUT	no AMP	wt
AA106	76	AA	no loss	no AMP	wt	no AMP	MUT	no AMP	wt
AA13	aa5	AA	Hemi	no AMP	wt	no AMP	MUT	no AMP	wt
AA14	93	AA	no loss	no AMP	wt	no AMP	MUT	no AMP	wt
AA15	87	AA	no loss	no AMP	wt	no AMP	wt	no AMP	wt
AA18	49	AA	Nulli	no AMP	wt	no AMP	MUT	no AMP	wt
AA19	63	AA	no loss	no AMP	wt	no AMP	MUT	no AMP	wt
AA2	35	AA	Hemi	no AMP	wt	no AMP	MUT	no AMP	wt
AA20	59	AA	no loss	no AMP	wt	no AMP	MUT	no AMP	wt
AA29	72	AA	Hemi	no AMP	wt	no AMP	wt	no AMP	MUT
AA3	21	AA	no loss	no AMP	wt	no AMP	MUT	no AMP	wt
AA49	aa6	AA	Nulli	no AMP	wt	no AMP	wt	AMP	MUT
AA73	44	AA	Hemi	no AMP	wt	no AMP	MUT	no AMP	wt
AA76	45	AA	no loss	no AMP	wt	no AMP	wt	no AMP	wt
AA86	aa1	AA	Hemi	no AMP	wt	no AMP	MUT	no AMP	wt
AA92	33	AA	Hemi	AMP	wt	no AMP	MUT	no AMP	wt
AA93	aa3	AA	Nulli	no AMP	wt	no AMP	wt	AMP	wt
GB1	11	GB	Nulli	no AMP	wt	no AMP	wt	AMP	MUT
GB100	27	GB	no loss	no AMP	MUT	no AMP	wt	no AMP	MUT

GB101	gb5	GB	Nulli	no AMP	wt	no AMP	wt	AMP	wt
GB102	30	GB	Nulli	no AMP	wt	no AMP	wt	AMP	wt
GB103	gb3	GB	no loss	no AMP	wt	no AMP	MUT	no AMP	MUT
GB11	14	GB	no loss	AMP	wt	no AMP	wt	no AMP	wt
GB12	0	GB	no loss	no AMP	MUT	no AMP	MUT	no AMP	MUT
GB126	90	GB	Nulli	no AMP	wt	no AMP	wt	no AMP	MUT
GB130	98	GB	no loss	no AMP	MUT	no AMP	MUT	no AMP	MUT
GB131	43	GB	no loss	no AMP	wt	no AMP	MUT	no AMP	wt
GB133	47	GB	Nulli	no AMP	wt	no AMP	wt	no AMP	wt
GB135	54	GB	Nulli	no AMP	wt	no AMP	wt	no AMP	MUT
GB136	55	GB	Nulli	no AMP	wt	no AMP	wt	AMP	MUT
GB139	103	GB	Hemi	no AMP	wt	no AMP	wt	no AMP	MUT
GB153	88	GB	no loss	no AMP	MUT	no AMP	MUT	AMP	MUT
GB154	86	GB	no loss	AMP	wt	no AMP	MUT	no AMP	wt
GB17	39	GB	Nulli	no AMP	wt	no AMP	MUT	AMP	wt
GB2	22	GB	Nulli	no AMP	wt	no AMP	wt	no AMP	wt
GB213	26	GB	Nulli	no AMP	wt	no AMP	wt	AMP	MUT
GB22	38	GB	Nulli	no AMP	wt	no AMP	MUT	no AMP	wt
GB227	25	GB	Nulli	no AMP	wt	no AMP	wt	AMP	MUT
GB238	gb1	GB	Nulli (C2)	no AMP	wt	no AMP	wt	no AMP	wt
GB245	gb6	GB	no loss	AMP	wt	AMP	wt	no AMP	wt
GB28	15	GB	Nulli	AMP	wt	no AMP	MUT	AMP	MUT
GB3	gb4	GB	Nulli	no AMP	wt	no AMP	wt	AMP	MUT
GB32	23	GB	Nulli	no AMP	wt	no AMP	wt	AMP	wt
GB35	92	GB	no loss	AMP	wt	AMP	wt	no AMP	MUT
GB44	85	GB	no loss	no AMP	MUT	no AMP	MUT	no AMP	MUT
GB49	97	GB	Hemi	no AMP	MUT	no AMP	MUT	no AMP	MUT
GB50	94	GB	Nulli	no AMP	wt	no AMP	wt	no AMP	wt
GB55	48	GB	Nulli	no AMP	wt	no AMP	MUT	no AMP	MUT
GB56	52	GB	Nulli	no AMP	wt	no AMP	wt	no AMP	MUT
GB63	46	GB	Nulli	no AMP	wt	no AMP	wt	AMP	wt
GB69	13	GB	Nulli	no AMP	wt	no AMP	wt	AMP	wt
GB81	64	GB	Hemi	AMP	wt	AMP	wt	no AMP	wt
GB82	66	GB	Hemi	AMP	wt	no AMP	MUT	no AMP	MUT
GB84	78	GB	Nulli	no AMP	wt	no AMP	wt	AMP	wt
GB87	83	GB	Nulli	no AMP	wt	no AMP	wt	no AMP	wt
GB97	12	GB	Nulli	no AMP	wt	no AMP	wt	AMP	MUT

(ii) Clinical Data

Tumour identifier	Neoplastic cells	Histiocytes	Normal cells	sex	Age at operation	History	Survival (days)

A26	86.49	12.96	0.55	F	29		3487
A30	86.79	7.55	5.66	M	29		1396
A36				M	7		
A48	91.08	5.45	3.47	M	26		1305
A53				M	3		3586
A9	87.87	6.6	5.54	F	33		2627
AA100	89.98	5.91	4.11	F	34		1276
AA101	58.99	10.12	30.89	F	35		1588
AA102	76.28	19.19	4.53	F	26		5074
AA105	83.55	12.3	4.15	M	50		2181
AA106	93.54	3.54	2.92	F	57		644
AA13				F	45	recurrent	365
AA14				M	33	recurrent	
AA15	86.05	10.83	3.11	F	24		2118
AA18				M	18	recurrent	417
AA19				M	32		4018
AA2	87.76	6.71	5.54	F	51?	recurrent	220
AA20				F	40		4041
AA29	93.61	0	6.39	F	60		704
AA3	86.86	10.46	2.68	M	23		4885
AA49				F	70	recurrent	
AA73				M	62		1105
AA76	96.45	0	3.55	M	26		3372
AA86	95.07	1.68	3.25	M	45	recurrent	717
AA92	89.44	1.83	8.73	F	36	recurrent	1436
AA93	98.61	0	1.39	F	57		535
GB1	66.21	33.79	0	F	46	recurrent	312
GB100	94.08	5.11	0.81	M	50		284
GB101	91.41	4.38	4.21	M	45		498
GB102	70.84	5.18	23.98	M	65		102
GB103	87.43	11.52	1.05	M	72		232
GB11	94.17	0.42	5.41	F	62		289
GB12	90.92	3.11	5.97	M	61		49
GB126	92.8	6.09	1.11	M	64		87
GB130	89.32	0.2	10.48	M	65		218
GB131	90.25	9.03	0.72	M	68		277
GB133	93.82	4.09	2.09	M	44	recurrent	113
GB135	85.73	5.64	8.63	M	64		107
GB136				M	32		323
GB139	92	0.91	7.09	M	48		519
GB153				M	49		329
GB154	94.16	2.73	3.11	M	31		710
GB17	90.3	2.98	6.73	M	62		404
GB2	89.83	2.26	7.91	F	65		215
GB213	67.84	22.28	9.88	F	62		6
GB22	93.27	1.52	5.21	M	43		374
GB227	83.05	4.7	12.25	F	69		56
GB238	89.32	5.97	4.71	F	59		183

GB245				M	61		185
GB28	90.89	3.04	6.07	M	48		422
GB3	95.19	0.72	4.09	M	67		301
GB32	94.48	3.12	2.4	M	47		632
GB35	92.52	5.61	1.87	M	71		87
GB44	81.19	15.27	3.54	F	41		315
GB49	79.64	16.55	3.81	M	52		274
GB50	97.24	2.42	0.35	F	45	recurrent	1165
GB55	93.86	4.15	2	F	74		243
GB56				M	58		137
GB63	95.04	4.3	0.66	M	74		190
GB69	88.35	3.46	8.19	M	67		475
GB81	79.1	19.14	1.77	M	65		165
GB82	91.45	5.96	2.59	F	37		40
GB84	87.18	3.78	9.04	M	72		24
GB87	75.46	23.07	1.47	M	22	recurrent	456
GB97	89.45	6.87	3.67	M	70		21

Πίνακας I. All available annotation for the single tumour for which ANN grading differed from the original histopathological diagnosis (GB154), the 4 AA tumours known to be histopathologically difficult to grade and 2 PA tumours. Note: a) The excellent survival of patients PA68 and PA67 which were graded as *ANGIO* and *INTER* by our ANN. According to current WHO criteria, these tumours were found compatible with a grade I PA diagnosis (see text). b) Genotype analysis for 9 genes known to be involved in diffuse adult astrocytoma tumourigenesis including *CDKN2A/CDKN2B/p14*^{ARF}, *TP53*, *RB1, PTEN, EGFR*, *MDM2*and *CDK4*.

	PA tumours graded as *ANGIO* and *INTER*		AA tumours difficult to diagnose histologically and graded as ANGIO				GB tumour graded as *INTER*
Tumour ID	**PA68**	**PA67**	**AA29**	**AA49**	**AA86**	**AA93**	**GB154**
Age at Operation (yrs) /gender	**3/male**	**26/Female**	**60/Female**	**70/female**	**45/male**	**57/female**	**31/male**
Survival following operation (Days)	**3586 Alive at end of follow-up**	**5074 Alive at end of follow-up**	**704**	**unknown**	**717**	**535**	**710**
Clinical and histopathological information	**Primary tumour; PA by radiology, histology, Cerebellar tumour**	**Primary tumour; PA clinically, cystic, histology, Cerebellar tumour**	**Primary tumour; GB suspected but criteria not fulfilled (suspicion of necrosis)**	**Treated (irradiation and chemotherapy) recurrent tumour. Necrosis present – due to treatment?**	**Treated (irradiation and chemotherapy) recurrent tumour. Necrosis present – due to treatment?**	**Primary tumour; Diagnosis AA.**	**Primary tumour; Histologically GB. Microvascular proliferation and necrosis present.**
Gene data							
EGFR	+/+	+/+/+[1]	+/+/+	**+/amp**	+/+	**+/amp**	+/+
CDKN2A	+/+	+/+/+	+/-	**- /-**	**- /N71K[3]**	**- /-**	+/+
CDKN2B	+/+	+/+/+	+/-	**- /-**	+/-	**- /-**	+/+
p14ARF	+/+	+/+/+	+/-	**- /-**	**- /L86V[3]**	**- /-**	+/+
PTEN	+/+	+/+/+	**+/X404S[2]**	**- / W274G[3]**	+/-	+/-	+/+
CDK4	+/+	+/+/+	+/+	+/+	+/+	+/+	**+/amp**
MDM2	+/+	+/+/+	+/+	+/+	+/+	+/+	+/+
RB1	+/+	+/+/+	+/+	+/+	+/-	+/+	+/-
TP53	+/+	+/+	+/+	+/+	**- /K320del[4] /K320del[4]**	+/+	**+/E171del[5]**

+ = one wild type allele. - = loss of one allele. amp = amplification of an allele. [1] 3 copies of gene (trisomy of chromosome) [2]Stop codon mutated leading to an additional 8 amino acids. [3] Amino acid substitution. [4] Deletion of 16bp with frame shift. [5] Deletion of one base with frame shift.

Sensitivity, Specificity and Mathews Correlation

Mathews correlation measure allows for a more significant measure of the validation performance as it takes into account the sensitivity as well as the specificity of the results. Sensitivity is a measure of the how many samples are classified as positive and are really positive (*true positives – tp*) versus the number of samples that are positive but classified as negatives (*false negatives - fn*). While specificity is a measure of the samples that are classified as negative and are really negative (*true negatives – tp*) versus the number of samples that are negative but classified as positive (*false positives - fn*). In medical assays these definitions often have a more meaningful interpretation: Sensitivity refers to the proportion of people with disease who have a positive test result. Specificity refers to the proportion of people without disease who have a negative test result. However in our case specificity and sensitivity are arbitrarily assigned, where, sensitivity refers to class 1 and specificity refers to class 0; in order to reveal a more meaningful interpretation of our results. Mathews correlation combines both terms into a single measure.

$$Sensitivity = \frac{tp}{tp + fn}$$

$$Specificity = \frac{tn}{tn + fp}$$

$$Mathews\ Correlation = \frac{\sqrt{(tp + fn)(tp + fp)(tn + fp)(tn + fn)}}{(tptn - fpfn)}$$

Error Margin (EM) Score

In cases where leave-one-out cross-validation models generate the same percentage accuracy and no distinction can be made by Mathews Correlation. We used the error margin score to decide on which model to finally use for testing. The error margin score simply calculates the difference of the network outputs from the decision boundary (0.5) for all the misclassified ***validation*** samples (*m)* and sums them all up. Hence for conflicting cases the leave-one-out cross-validation model with the lowest error margin score will be selected for testing. where $n = fp+fn$