ΑΚΑΔΗΜΙΑ ΑΘΗΝΩΝ

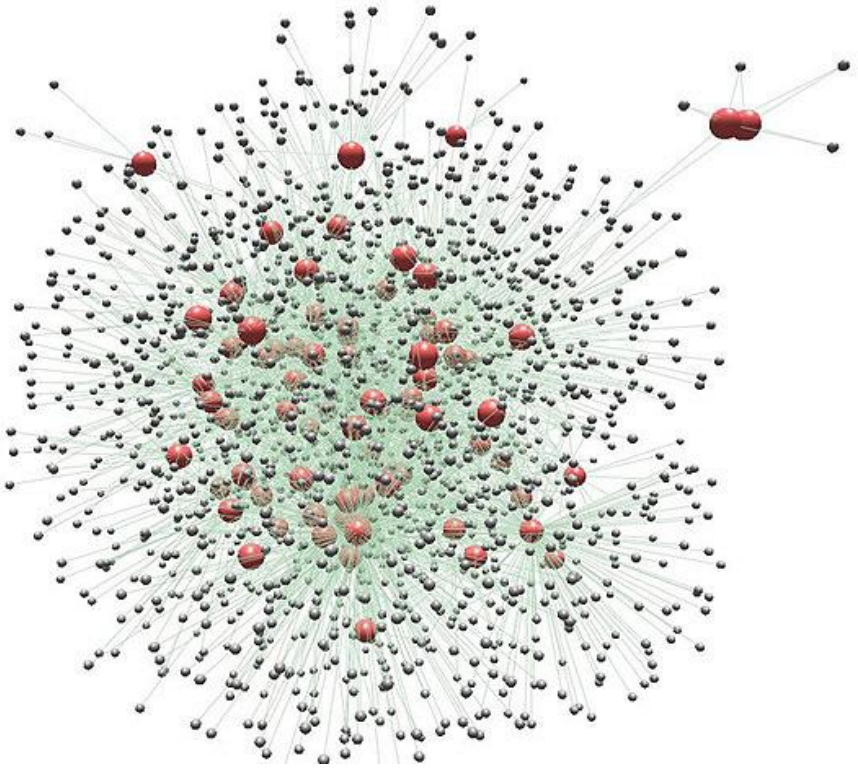
ΙΙΒΕΑΑ
ΙΔΡΥΜΑ ΙΑΤΡΟΒΙΟΛΟΓΙΚΩΝ ΕΡΕΥΝΩΝ
ΑΚΑΔΗΜΙΑΣ ΑΘΗΝΩΝ

Εθνικό και Καποδιστριακό
ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Μεταπτυχιακό Πρόγραμμα «Μοριακή Ιατρική»

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

"ΔΙΚΤΥΑ ΜΕΤΑΓΡΑΦΙΚΩΝ ΠΑΡΑΓΟΝΤΩΝ ΚΑΙ ΕΠΑΝΑΠΡΟΓΡΑΜΜΑΤΙΣΜΟΣ ΤΟΥ ΑΝΘΡΩΠΙΝΟΥ ΓΟΝΙΔΙΩΜΑΤΟΣ ΜΕΤΑ ΑΠΟ ΙΙΚΗ ΜΟΛΥΝΣΗ"



Αυγέρης Ευθύμιος Γεώργιος

ΕΡΓΑΣΤΗΡΙΟ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ

ΙΙΒΕΑΑ

ΑΘΗΝΑ, 2014

ΤΡΙΜΕΛΗΣ ΣΥΜΒΟΥΛΕΥΤΙΚΗ ΕΠΙΤΡΟΠΗ:

Επιβλέπων Μέλος ΔΕΠ: **Δημήτριος Στραβοπόδης**, Επίκουρος Καθηγητής Βιολογίας Κυττάρου & Ανάπτυξης, Τμήμα Βιολογίας Πανεπιστημίου Αθηνών.

Μέλος τριμελούς Επιτροπής (1): **Δημήτριος Θάνος,** Ερευνητής Α', Πρόεδρος Επιστημονικού Συμβουλίου. Ίδρυμα Ιατροβιολογικών Ερευνών Ακαδημίας Αθηνών.

Μέλος τριμελούς Επιτροπής (2): **Ιωάννης Ταλιανίδης,** Ερευνητής Α', Διευθυντής Ινστιτούτου Μοριακής Βιολογίας & Γενετικής , Ερευνητικό Κέντρο Βιοϊατρικών Επιστημών «Αλέξανδρος Φλέμινγκ».

Πρόλογος

Η παρούσα διπλωματική εργασία εκπονήθηκε στο εργαστήριο Μοριακής Βιολογίας του Ιδρύματος Ιατροβιολογικών Ερευνών της Ακαδημίας Αθηνών, στα πλαίσια του προγράμματος των Μεταπτυχιακών Σπουδών της «Μοριακής Ιατρικής» κατά το Ακαδημαϊκό Έτος 2013-2014. Με την ολοκλήρωσή της θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου, Dr. Δημήτριο Θάνο, Ερευνητή Α', που μου έδωσε τη δυνατότητα να εκπονήσω τη διπλωματική μου εργασία όπως και για την αμέριστη ηθική και υλική συμπαράσταση και την ουσιαστική καθοδήγησή του. Επιπλέον θα ήθελα να ευχαριστήσω τα μέλη της συντονιστικής επιτροπής, τον Καθηγητή Δημήτριο Στραβοπόδη και τον Dr. Ιωάννη Ταλιανίδη, Ερευνητή Α΄, που μου έκαναν την τιμή να αποτελούν μέλη της εξεταστικής επιτροπής μου.

Θα ήθελα να ευχαριστήσω θερμά δύο ξεχωριστούς ανθρώπους που έπαιξαν το σημαντικότερο ρόλο σε όλη αυτή τη διαδρομή. Πρόκειται για τον Μεταδιδακτορικό Ερευνητή Αλέξανδρο Πολύζο και την Μεταδιδακτορική Ερευνήτρια Μαρία Καπασά. Ήταν αυτοί οι οποίοι με δίδαξαν και μοιράστηκαν μαζί μου όλες τις γνώσεις βιοπληροφορικής και στατιστικής ανάλυσης δεδομένων τις οποίες απέκτησα τα χρόνια στα οποία συνεργαστήκαμε. Μέσα από χαρές, λύπες, σκέψεις και συμβουλές μου μετέδωσαν τον επιστημονικό τρόπο σκέψης μου έμαθαν πώς να προσεγγίζω και να πραγματοποιώ μεθοδικά τις αναλύσεις των πειραμάτων.

Θα ήθελα τέλος να ευχαριστήσω την οικογένειά μου, τον Δημήτρη, την Μαρία και την Αριάνα, που βρίσκονται πάντα κοντά μου και με στηρίζουν· η αγάπη τους είναι για μένα η μεγαλύτερη επιβράβευση και η δύναμη για τη συνέχεια.

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΛΗΨΗ

Οι αποκρίσεις των κυττάρων σε παθογόνους μικροοργανισμούς, ιδιαίτερα σε ιούς, αποτελούν τις περισσότερο μελετημένες περιπτώσεις κυτταρικών αποκρίσεων σε εξωτερικά ερεθίσματα. Οι φαινοτυπικές αλλαγές των κυττάρων-ξενιστών συνοδεύονται με χαρακτηριστικές αλλαγές στη γονιδιακή τους έκφραση. Τεχνικές που χρησιμοποιούν DNA μικροσυστοιχίες αλλά και ακόμη πιο σύγχρονες τεχνικές αλληλούχησης των μεταγράφων (RNA-seq) παρέχουν τη δυνατότητα εντοπισμού με ακρίβεια των γονιδίων που εκφράζονται αλλά και του βαθμού στον οποίο αλλάζει η έκφρασή τους κατόπιν μόλυνσης με κάποιο ιό ή βακτήριο. Η ανάλυση τέτοιων πειραμάτων προσφέρει σημαντικές πληροφορίες για τα γονίδια που εμπλέκονται στους μηχανισμούς με τους οποίους αποκρίνονται τα κύτταρα-ξενιστές στη μόλυνση και καθορίζουν την επιβίωσή τους.

Η ρύθμιση αυτών των γονιδίων πραγματοποιείται από τη συντονισμένη αλληλεπίδραση των μεταγραφικών παραγόντων αλλά και των τοπικών επιλεκτικών τροποποιήσεων των ιστονών με το DNA. Με την τεχνική της ανοσοκατακρήμνισης και της ακόλουθης αλληλούχησης των DNA τμημάτων που απομονώνονται (ChIP-seq) μπορούν να εντοπιστούν οι θέσεις πρόσδεσής των μεταγραφικών παραγόντων και οι τροποποιήσεις των ιστονών σε διάφορα γονίδια. Ο συνδυασμός πληροφοριών από τις παραπάνω τεχνικές μπορεί να αποκαλύψει τους περίπλοκους μηχανισμούς που ενεργοποιούνται στα κύτταρα μετά από ιικές μολύνσεις. Μπορεί επίσης να εντοπιστούν κάποια γονίδια που υπερεκφράζονται στην πλειοψηφία των κυτταρικών σειρών κατόπιν μόλυνσης με παθογόνους μικροοργανισμούς. Στην παρούσα εργασία συλλέχθηκαν δεδομένα πειραμάτων DNA μικροσυστοιχιών και RNA-seq από τις βάσεις δεδομένων GeoDatasets, Array Express και SRA όπου είχαν προηγηθεί μολύνσεις ανθρώπινων κυττάρων με ιούς και βακτήρια. Επιπλέον χρησιμοποιήθηκαν δεδομένα από ChIP-seq πειράματα από τη βάση δεδομένων του ENCODE. Από την ανάλυση αυτών των δεδομένων εντοπίστηκε μία ομάδα 348 γονιδίων των οποίων η έκφραση εμφανίζεται να αυξάνεται κατόπιν μόλυνσης σε διάφορες ανθρώπινες κυτταρικές σειρές. Χαρακτηρίστηκαν επίσης οι μεταγραφικοί παράγοντες αλλά και οι ιστονικές τροποποιήσεις που πιθανώς ρυθμίζουν την έκφραση αυτών των γονιδίων σε τρείς κυτταρικές σειρές.

ABSTRACT

The responses of host cells to pathogenic microorganisms are among the most-well studied examples of cellular responses to external stimuli. Pathogen-induced phenotypic changes in host cells are often accompanied by marked changes in gene expression. DNA microarrays and the most recent RNA-seq technologies provide us with the ability to monitor the changes in the abundance of transcripts after virus or bacterial infection. Analysis of such experiments offers important information for the genes that are involved in the cellular response mechanisms upon the infection and are considered crucial for the survival of the cells.

Genes are regulated by the coordinated interactions of transcription factors and histone modification with the DNA. ChIP-seq technology unravels the binding positions of the transcription factors and the histone modifications of the surrounding chromatin of the genes. The combination of the data coming from the above experiments can illuminate the complicated mechanisms that are activated as the cells response to the infection. Moreover some central genes that are upregulated in most cell lines after the virus or bacterial infection can be identified. In this thesis data from DNA microarrays and RNA-seq experiments on human cell lines infected with viruses and bacteria were used. The data were downloaded from GeoDatasets, Array Express and SRA databases, while data of ChIP-seq experiments from the ENCODE database were also gathered. From the analyses of those data, 348 genes were identified that were overexpressed after virus and bacterial infection in the majority of the studied cell lines. Finally transcription factors and histone modifications that regulate those genes were identified in 3 cell lines.

ΕΙΣΑΓΩΓΗ

2.1 ΙΙΚΕΣ ΜΟΛΥΝΣΕΙΣ

Οι ιοί και οι ξενιστές τους έχουν συν-εξελιχθεί για εκατομμύρια χρόνια. Για να αναπαράγουν με επιτυχία το γονιδίωμά τους, οι ιοί πρέπει να χρησιμοποιήσουν το βιοσυνθετικό μηχανισμό του κυττάρου-ξενιστή. Ανάλογα με την πολυπλοκότητα και τη φύση του γονιδιώματος, η αντιγραφή μπορεί να περιλαμβάνει ένα σχετικά μεγάλο υποσύνολο ιικών προϊόντων. Ο πολλαπλασιασμός του ιού πραγματοποιείται σε διάφορα υποκυτταρικά διαμερίσματα συμπεριλαμβανομένου του πυρήνα και του κυτταροπλάσματος. Ως εκ τούτου, οι ιοί πρέπει να εξασφαλίσουν τη σωστή υποκυτταρική κατανομή τους και να είναι σε θέση να αποφύγουν τους κυτταρικούς αμυντικούς μηχανισμούς.

Έχει γίνει σαφές εδώ και πολλά χρόνια ότι το κάθε άτομο διαφέρει σημαντικά στην ευαισθησία του σε μολυσματικές ασθένειες σε σχέση με άλλα άτομα. Οι πρόσφατες εξελίξεις στον τομέα της γονιδιωματικής έχουν οδηγήσει στη δραματική αύξηση της ισχύος των διαθέσιμων τεχνικών για τον εντοπισμό των γονιδίων που συμμετέχουν στους αμυντικούς μηχανισμούς. Η ανθρώπινη γενετική ποικιλότητα ασκεί σημαντική επιρροή στην πορεία νόσων που προκαλούνται από πολλούς μολυσματικούς παράγοντες. Η εν λόγω αλληλεπίδραση γονιδίων παθογόνου-ξενιστή είναι γενικού βιολογικού ενδιαφέροντος, καθώς διέπει τη διατήρηση της γενετικής ποικιλότητας. Τέτοια συνεξελικτική αλληλεπίδραση έχει καλύτερα μελετηθεί σε μολυσματικές ασθένειες του ανθρώπου, όπου τόσο το παθογόνο όσο και το γονιδίωμα του ξενιστή είναι καλά χαρακτηρισμένα.

2.2 ΙΟΙ ΚΑΙ ΚΥΤΤΑΡΑ

Οι ιοί είναι ενδοκυττάρια παράσιτα που προκαλούν πολύπλοκες αντιδράσεις στα κύτταρα ξενιστές που μολύνουν. Οι ιοί αποτελούν μια άμεση απειλή για το γονιδίωμα των ξενιστών, όπου κατά τη διάρκεια της λοίμωξης λαμβάνει χώρα μια μάχη στην οποία κάθε γονιδίωμα πρέπει να προστατεύεται για τη διατήρηση της γενετικής ακεραιότητάς του. Για τη διατήρηση της ακεραιότητας του γονιδιώματος τους, τα κύτταρα διαθέτουν ένα εξελιγμένο δίκτυο επιτήρησης για την ανίχνευση και την επισκευή των βλαβών του DNA. Η απόκριση βλάβης DNA (DNA Damage Response-DDR) εμποδίζει μια σειρά ασθενειών του ανθρώπου και όταν απορυθμιστεί μπορεί να οδηγήσει σε γενετική αστάθεια και στην εμφάνιση καρκίνου (Guan et al. 2003).

Οι ιοί για να επιβιώσουν και να πολλαπλασιαστούν επιτυχώς στα κύτταρα του ξενιστή έχουν αναπτύξει μηχανισμούς με τους οποίους κατευθύνουν την κυτταρική απόκριση σε βλάβη του DNA και τους κυτταρικούς μηχανισμούς που συμμετέχουν στην επιδιόρθωση του DNA. Οι ιοί και οι ιογενείς ογκοπρωτεΐνες στοχεύουν πολλαπλούς κυτταρικούς παράγοντες που συμμετέχουν στη διατήρηση της ακεραιότητας του γονιδιώματος. Ένας αριθμός ιών έχει ενοχοποιηθεί για την παθογένεση των ανθρώπινων κακοηθειών. Η επαγωγή γονιδιωματικής αστάθειας είναι ένα κρίσιμο γεγονός στην ιική ογκογένεση. Το DDR αποτελεί ένα σκέλος της εγγενούς κυτταρικής άμυνας κατά των ιογενών λοιμώξεων. Οι μηχανισμοί επιδιόρθωσης κυτταρικού DNA μπορεί να αναγνωρίσουν το ιικό γενετικό υλικό ως κατεστραμμένο DNA και να περιορίσουν την ιογενή λοίμωξη. Οι ιοί έχουν αναπτύξει στρατηγικές για την αξιοποίηση της κυτταρικής απόκρισης βοηθώντας τον διπλασιασμό τους. Μελετώντας την επιδιόρθωση του DNA, μαζί με τη φυσική διαδικασία της μόλυνσης από τον ιό έχει ανοίξει ένα νέο πεδίο στη βιολογία των αλληλεπιδράσεων ιού-ξενιστή και παρέχει γνώσεις σχετικά με κυτταρικές διεργασίες που εμπλέκονται στην αναγνώριση και την επεξεργασία των βλαβών του DNA (Ambeker et al. 2000).

Ως ενδοκυττάρια παράσιτα, όλοι οι ιοί εξαρτώνται από τον μηχανισμό βιοσύνθεσης του κυττάρου ξενιστή προκειμένου να αναπαράγουν το γονιδίωμά τους και να δημιουργήσουν σωματίδια απογόνους του ιού. Ένα κοινό χαρακτηριστικό μεταξύ πολλών διαφορετικών ιών είναι η δημιουργία εξειδικευμένων μεμβρανωδών διαμερισμάτων μέσα στο κυτταρόπλασμα του μολυσμένου κυττάρου.

Η ακεραιότητα του γονιδιώματος είναι υπό συνεχή απειλή ενδογενών και εξωγενών παραγόντων. Τα κύτταρα έχουν αναπτύξει συστήματα για τον εντοπισμό και την επιδιόρθωση κατεστραμμένου DNA και

σηματοδοτούν την παρουσία του, προκειμένου να αποτραπεί η μετάδοση γενετικών μεταλλάξεων σε ακόλουθες κυτταρικές διαιρέσεις. Αυτά τα μονοπάτια επισκευής εντοπίζουν μια σειρά από βλάβες του DNA όπως αντικαταστάσεις/προσθήκες βάσεων του DNA, μεταλλαγμένες βάσεις, φωτοπροϊόντα UV, εγκοπές στη μονή αλυσίδα του DNA και θραύσεις στη διπλή έλικα του DNA (Double Strand Breaks-DSBs). Η γονιδιωματική αστάθεια είναι ένα χαρακτηριστικό γνώρισμα του κακοήθους μετασχηματισμού και της δημιουργίας όγκων. Τα κληρονομικά ελαττώματα στα γονίδια που συμμετέχουν στη κυτταρική DDR εμπλέκονται σε ποικίλες ασθένειες του ανθρώπου και προδιαθέτουν τα άτομα στην εμφάνιση καρκίνου. Ξεχωριστές ανθρώπινες παθήσεις μπορεί να προκύψουν από την ύπαρξη μεταλλάξεων στα ενεργά κέντρα πρωτεϊνών όπως ATM, ATR, NBS1 και MRE11.

Υπάρχουν παραδείγματα στα οποία η DDR ενεργοποιείται κατά τη διάρκεια της μολύνσεως με ιούς φυσικού τύπου (wild type). Σε ορισμένες περιπτώσεις ο ιός φυσικού τύπου δεν οδηγεί σε επαγωγή βλάβης αλλά αυτή μπορεί να προκύψει κατόπιν μόλυνσης με μεταλλαγμένους ιούς. Με τον τρόπο αυτό οι μεταλλαγμένοι ιοί προσφέρουν σημαντικές πληροφορίες για τους τρόπους με τους οποίους το DDR λειτουργεί κατά τη διάρκεια λοιμώξεων από τον ιό και συνέβαλαν στον προσδιορισμό των ιικών παραγόντων που στοχεύουν κατά των κυτταρικών ρυθμιστικών λειτουργιών.

Η φυσιολογική κατάσταση των ζωντανών κυττάρων έχει σημαντική επίδραση επί του αποτελέσματος της λοίμωξης από τον ιό, καθώς το κύτταρο ξενιστής παρέχει το βιοσυνθετικό μηχανισμό και τα βασικά ρυθμιστικά μόρια για τη δημιουργία ιικών πρωτεϊνών και νουκλεϊνικών οξέων. Το βέλτιστο ενδοκυτταρικό περιβάλλον για τον πολλαπλασιασμό του ιού αναπτύσσεται μέσα από τα γεγονότα που αρχίζουν να λαμβάνουν χώρα με την προσκόλληση του ιού στην κυτταρική μεμβράνη. Η πρόσδεση του ιού στον υποδοχέα της κυτταρικής μεμβράνης μπορεί να ακολουθείται από μια πληθώρα γεγονότων που συνδέονται με βιοχημικές, φυσιολογικές και μορφολογικές αλλαγές στα κύτταρα . Ο υποδοχέας του ιού είναι ένα συστατικό της κυτταρικής μεμβράνης που συμμετέχει στην δέσμευση του ιού, διευκολύνει την ιική μόλυνση και είναι ένας καθοριστικός παράγοντας δυνατότητα του ιού να μολύνει συγκεκριμένους ξενιστές. Ορισμένοι ιοί αναγνωρίζουν περισσότερους από έναν κυτταρικό υποδοχέα (π.χ. ο ιός HIV-Human Immunodeficiency Virus ή οι αδενοϊοί).

2.3 ΕΠΙΔΡΑΣΕΙΣ ΣΤΗ ΒΙΟΧΗΜΕΙΑ ΤΩΝ ΚΥΤΤΑΡΩΝ

Η ιική σύνδεση στην κυτταρική μεμβράνη σε συνδυασμό με την ύπαρξη άμεσου πρώιμου ιού (π.χ. IE πρωτεΐνες των ερπητοϊών), πρώιμων μη-δομικών πρωτεϊνών (π.χ. E6, E7 από HPVs) μπορεί να μεσολαβήσει σε μια σειρά από βιοχημικές αλλαγές που βελτιστοποιούν το ενδοκυτταρικό περιβάλλον για την ανάπτυξη του ιού. Μελέτες της μεταγραφικής ρύθμισης των ιικών γονιδίων και της μετα-μεταγραφικής τροποποίησης των γονιδιακών προϊόντων αποδεικνύουν ότι η φύση των βασικών βιοχημικών διεργασιών για την αντιγραφή του ιού είναι παρόμοιες με τους μηχανισμούς που χρησιμοποιούνται για να ρυθμίζουν την έκφραση των κυτταρικών γονιδίων. Το DNA των ιών περιέχει αρκετές θέσεις πρόσδεσης μεταγραφικών παραγόντων του κυττάρου που ρυθμίζουν την έκφραση των γονιδίων του ιού.

Τέτοιοι μεταγραφικοί παράγοντες είναι οι NF-κB , Sp1, AP- 1, OCT-1, NP- 1 (Hill et al. 1987). Αυτοί οι μεταγραφικοί παράγοντες σε συντονισμό με τις ρυθμιστικές ιικές πρωτεΐνες ενεργοποιούν ή καταστέλλουν ιικά και κυτταρικά γονίδια και αναπτύσσουν λανθάνουσα και επίμονη λοίμωξη. Αυτές οι διεργασίες ενεργοποίησης μπορεί να επιτευχθούν ως αποτέλεσμα μιας σειράς γεγονότων που ξεκινούν από τον ιό και την αλληλεπίδραση του με τον υποδοχέα του κυττάρου. Ακολουθούν γεγονότα όπως ο σχηματισμός δευτερογενών αγγελιοφόρων (φωσφατιδυλο-ινοσιτόλων, διακυλογλυκερόλων, cAMP, cGMP, κ.λπ.) , ενεργοποίηση κινασών και ενεργοποίηση καναλιών ιόντων (π.χ. , Ca2 +).

Για τη διατήρηση των διαδικασιών ενεργοποίησης των κυττάρων κατά τη μεταγραφή, οι ιοί έχουν εξελίξει ειδικούς μηχανισμούς για τη ρύθμιση αυτών των κυτταρικών διαδικασιών, προσαρμόζοντας τις πρωτεΐνες τους με τέτοιο τρόπο ώστε να αλληλεπιδρούν με κυτταρικές πρωτεΐνες. Παραδείγματα περιλαμβάνουν την ένωση των γονιδιακών προϊόντων πρόωρων ιών (π.χ. E6, E7 των ιών θηλώματος, τις IE πρωτεΐνες των ερπητοϊών, τα SV40 T αντιγόνο) με τον Rb καταστολέα όγκων που οδηγεί σε απελευθέρωση του μεταγραφικού παράγοντα E2F που απαιτείται για την τροποποίηση (ενεργοποίηση / αναστολή) των

κυτταρικών βιοχημικών οδών, για τη σύνθεση του DNA του ιού ή την έναρξη κυτταρικών αποπτωτικών διεργασιών.

Σε ορισμένες περιπτώσεις, ο ιός ενσωματώνει άμεσα κυτταρικές βιοχημικές ρυθμιστικές στρατηγικές προκαλώντας τα κύτταρα να υπερπαράγουν και να εκκρίνουν ρυθμιστικά μόρια (π.χ. αυξητικοί παράγοντες μετασχηματισμού, παράγοντες νέκρωσης όγκου, ιντερλευκίνες) τα οποία μπορεί να ενεργοποιήσουν κατά ένα αυτοκρινή τρόπο τις κυτταρικές βιοχημικές αντιδράσεις που εμπλέκονται στην αντιγραφή του ιού (π.χ. HIV, ιούς του έρπητα, ιοί των ανθρωπίνων θηλωμάτων), στη συντήρηση ή στην επανενεργοποίηση από την λανθάνουσα κατάσταση του ιού. Από την άλλη πλευρά, αυτά τα διαλυτά κυτταρικά ρυθμιστικά μόρια μπορούν να ενεργοποιήσουν τις βιοχημικές αντιδράσεις των κυττάρων του ανοσοποιητικού συστήματος και με έναν παρακρινή τρόπο να επιτεθούν στα μολυσμένα κύτταρα.

Κατά τη διάρκεια της λοίμωξης ο ιός ευνοεί τη σύνθεση δικών του πρωτεϊνών και των νουκλεϊκών οξέων αναστέλλοντας τη σύνθεση κυτταρικών προϊόντων. Αυτή η αναστολή λαμβάνει χώρα με χαρακτηριστικό τρόπο. Στον ιό της πολιομυελίτιδας ή σε λοιμώξεις του απλού έρπητα, η επιλεκτική αναστολή της σύνθεσης των πρωτεϊνών του ξενιστή λαμβάνει χώρα πριν από τη σύνθεση ιικών πρωτεϊνών (Newport et al. 1994). Σε ορισμένες περιπτώσεις τα ιικά προϊόντα αναστέλλουν τόσο τη σύνθεση πρωτεϊνών όσο και των νουκλεϊκών οξέων.

Η πλήρης αναστολή της σύνθεσης προϊόντων του ξενιστή μπορεί επίσης να συμβεί όταν τα ιογενή προϊόντα συσσωρεύονται στο κύτταρο στον ιικό κύκλο πολλαπλασιασμού. Ορισμένοι πικορναϊοί (picornavirus) παράγουν μια πρωτεΐνη που προκαλεί βλάβη στα κύτταρα ανεξάρτητα από τις ιικές πρωτεΐνες που αναστέλλουν την κυτταρική μακρομοριακή σύνθεση. Σε αυτή την περίπτωση το κυτταρικό mRNA μπορεί να αποικοδομηθεί. Για παράδειγμα στον ιό της γρίπης και σε λοιμώξεις του ιού του απλού έρπητα το κυτταρικό mRNA σταματά τη σύνδεση με τα ριβοσώματα για το σχηματισμό πολυριβοσωμάτων. Η σύνθεση του DNA των κυττάρων αναστέλλεται στις περισσότερες μολύνσεις με κυτταρολυτικό ιό.

Αυτό μπορεί να επιτευχθεί με την επαγόμενη απόπτωση του ιού ή από μια μείωση στην κυτταρική σύνθεση πρωτεϊνών. Οι ρεοϊοί και μερικοί ερπητοϊοί μπορεί να αποτελούν εξαιρέσεις διότι προκαλούν μια μείωση στη σύνθεση του DNA των κυττάρων πριν από μια μείωση στην κυτταρική πρωτεϊνική σύνθεση που λαμβάνει χώρα.

2.4 ΧΡΩΜΟΣΩΜΙΚΗ ΒΛΑΒΗ

Η χρωμοσωμική βλάβη μπορεί να προκληθεί άμεσα από το σωματίδιο του ιού ή έμμεσα από γεγονότα που συμβαίνουν κατά τη σύνθεση των νέων ιικών μακρομορίων (RNA, DNA, πρωτεΐνη). Η χρωμοσωμική βλάβη μπορεί ή δεν μπορεί να επισκευαστεί πιστά και από αυτό εξαρτάται η επιβίωση του μολυσμένου κυττάρου. Όταν το κύτταρο επιβιώνει, το γονιδίωμα του ιού μπορεί να παραμείνει εντός του κυττάρου, οδηγώντας σε συνεχιζόμενη αστάθεια του γενωμικού κυτταρικού υλικού ή την αλλαγμένη έκφραση των γονιδίων του (π.χ. ογκογονίδια). Η γενωμική αστάθεια εξαιτίας του ιού σχετίζεται με τη συσσώρευση μεταλλάξεων οι οποίες οδηγούν στην κυτταρική αθανατοποίηση και στον ογκογόνο μετασχηματισμό (Sugimoto M et al. 2004).

2.5 ΒΙΟΛΟΓΙΚΕΣ ΕΠΙΔΡΑΣΕΙΣ

Οι βιολογικές συνέπειες της λοίμωξης του ιού προκύπτουν από τις προαναφερθείσες βιοχημικές, φυσιολογικές, δομικές, μορφολογικές και γενετικές αλλαγές. Οι κυτταρικές αλλαγές που προκαλούνται από τον ιό μπορούν να οδηγήσουν σε ασθένειες (π.χ. υποξεία σκληρυντική πανεγκεφαλίτιδα μετά τη μόλυνση από ιό ιλαράς), σε κυτταρικές γενετικές βλάβες (π.χ. ιός της ηπατίτιδας Β), σε αθανατοποίηση (π.χ. ιός Epstein-Barr), ή σε κακοήθη μετασχηματισμό (π.χ. HTLV-1, HTLV-2, ή τον ιό της ηπατίτιδας Β).

2.6 ΚΥΤΤΑΡΙΚΗ ΕΠΙΔΡΑΣΗ ΣΤΗΝ ΙΟΓΕΝΗ ΠΑΘΟΓΕΝΕΙΑ

Αν και τα περισσότερα από τα γεγονότα που βλάπτουν ή τροποποιούν το κύτταρο ξενιστή κατά τη διάρκεια της λυτικής μόλυνσης είναι δύσκολο να διαχωριστούν από την αντιγραφή του ιού, τα αποτελέσματα δεν συνδέονται άμεσα με την παραγωγή των ιοσωματίων απογόνων. Για παράδειγμα, μπορεί να υπάρξουν αλλαγές στο μέγεθος των κυττάρων, στο σχήμα και στις φυσιολογικές λειτουργίες πριν παραχθούν οι απόγονοι ιοσωματίων (Jubier-Maurin et al. 2001) ή ακόμη και πολλές πρωτεΐνες του ιού. Αυτές οι μεταβολές στη δομή και στη λειτουργία των κυττάρων μπορεί να είναι σημαντικές πτυχές της παθογένεσης ενός αριθμού ιογενών λοιμώξεων. Για παράδειγμα, μέσω των κυτταρικών επιπτώσεων τους πολλοί ιοί (π.χ. ροταϊοί, καλυκοϊοί, ιοί Norwalk) επάγουν ένα εύρος γαστρεντερικών συμπτωμάτων (από ήπια μεταβολή στην απορρόφηση των ιόντων έως σοβαρή διάρροια).

Οι κυτταροκτόνες ιογενείς λοιμώξεις (π.χ., ιούς έρπητα, φλαβοϊούς τογκαϊών, Bunyaviruses) του κεντρικού νευρικού συστήματος σχετίζονται με νέκρωση, φλεγμονή ή φαγοκυττάρωση στα υποστηρικτικά κύτταρα. Οι λοιμώξεις από τον ιό της ερυθράς σχετίζονται με την απομυελίνωση χωρίς το νευρικό εκφυλισμό. Οι μακροπρόθεσμες επιδράσεις των λοιμώξεων μπορούν επίσης να σχετίζονται με προοδευτικές ασθένειες όπως αθηροσκλήρωση και απομυελίνωση στη σκλήρυνση κατά πλάκας.

2.7 ΕΠΙΜΟΝΕΣ ΛΟΙΜΩΞΕΙΣ

Σε μια επίμονη λοίμωξη ο ιός δεν έχει εξαλειφθεί από το σύνολο των ιστών του ξενιστή μετά την αρχική λοίμωξη ούτε κατά τη διάρκεια της οξείας φάσης της νόσου. Οι διάφοροι τύποι της επίμονης λοίμωξης (χρόνια, αργή, λανθάνουσα) διαφέρουν στους μηχανισμούς ελέγχου της παθογένεσής τους.

• Σε χρόνιες λοιμώξεις, ένας περιορισμένος αριθμός κυττάρων (σε όργανα στόχους) μολύνεται. Αυτά τα μολυσμένα κύτταρα μπορεί να δημιουργήσουν μία κυτταροπαθογόνο επίδραση, να συνθέτουν τα μακρομόρια του ιού, και τελικά να απελευθερώσουν του απογόνους του μολυσματικού ιού. Η εξάπλωση της μόλυνσης περιορίζεται από παράγοντες του ξενιστή, όπως χυμική και κυτταρική ανοσοαπόκριση, ιντερφερόνες και άλλους μη ειδικούς αναστολείς.

• Οι αργές μολύνσεις που προκαλούνται από συμβατικούς (π.χ. ιλαρά) ή μη συμβατικούς ιούς (π.χ. πρίον) χαρακτηρίζονται από μακρές περιόδους επώασης, που προηγούνται της εμφάνισης των κλινικών συμπτωμάτων (Jubier-Maurin et al. 2001).

• Σε λανθάνουσες λοιμώξεις, ο μολυσματικός ιός σπάνια ανιχνεύεται μεταξύ των κλινικών επεισοδίων της νόσου. Στα λίγα κύτταρα που έχουν μολυνθεί, η έκφραση του ιού είναι περιορισμένη. Κοινά χαρακτηριστικά της λανθάνουσας λοίμωξης είναι η ικανότητά της να επανενεργοποιηθεί σε απόκριση προς διάφορα περιβαλλοντικά ερεθίσματα (π.χ. θερμότητα, υπεριώδη ακτινοβολία). Η ανοσοκαταστολή που προκαλείται από τη μόλυνση από ετερόλογο ιό (π.χ. HIV) ή χημειοθεραπεία, συχνά σχετίζεται με τη μεταμόσχευση οργάνων.

Η αυτοάνοση βλάβη και άλλες μορφές κυτταρικής βλάβης μπορεί να προκύψουν κατά τη διάρκεια επίμονων μολύνσεων. Τα νεοσυντιθέμενα ιοσωμάτια και τα ιικά πεπτίδια που σχετίζονται με την κυτταρική μεμβράνη αλλάζουν τα αντιγονικά χαρακτηριστικά του κυττάρου, έτσι ώστε το ανοσοποιητικό σύστημα να μπορεί να το αναγνωρίσει ως ξένο. Το κύτταρο, στη συνέχεια, μπορεί να δεχθεί επίθεση από το χυμικό και το κυτταρικό ανοσοποιητικό σύστημα του ξενιστή.

Η ανοσολογική απόκριση μπορεί επίσης να προκαλέσει το σχηματισμό συμπλεγμάτων αντιγόνου-αντισώματος που περιλαμβάνουν ιικά αντιγόνα. Αυτά τα σύμπλοκα μπορεί να προκαλέσουν φλεγμονή ενεργοποιώντας την κλασική οδό του συμπληρώματος. Η μακροπρόθεσμη σύνδεση του ιού με συγκεκριμένα κύτταρα στόχους μπορεί να οδηγήσει σε αλλοιωμένη λειτουργία ή αποκρίσεις. Αυτό το είδος του μηχανισμού πιστεύεται ότι είναι υπεύθυνο για την προοδευτική νευρολογική νόσο που σχετίζεται με λοιμώξεις αργών ιών όπως η ασθένεια kuru, η ασθένεια Creutzfeldt-Jakob, και η υποξεία σκληρυντική πανεγκεφαλίτιδα.

2.8 ΜΕΤΑΣΧΗΜΑΤΙΣΤΙΚΕΣ ΛΟΙΜΩΞΕΙΣ

Ο όρος ογκογόνος μετασχηματισμός αναφέρεται στη διαδικασία μέσω της οποίας ο έλεγχος του πολλαπλασιασμού των κυττάρων έχει απορυθμιστεί και το κύτταρο μετατρέπεται σε καρκινικό. Στο πλαίσιο των αλληλεπιδράσεων ιού-κυττάρου, τα κύτταρα μπορούν επίσης να υποβληθούν σε διάφορους τύπους κληρονομικών μεταβολών, με αποτέλεσμα να παρουσιάζονται βιοχημικές, αντιγονικές, μορφολογικές και φυσιολογικές μεταβολές, που ονομάζονται μη-ογκογόνοι μετασχηματισμοί (Kohler et al. 2005).

2.9 ΣΤΑΔΙΑ ΚΑΙ ΜΗΧΑΝΙΣΜΟΙ ΚΥΤΤΑΡΙΚΟΥ ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΥ

Ο μετασχηματισμός ενός κυττάρου περιλαμβάνει δύο βασικά στάδια:

1. Το κύτταρο αποκτά την ικανότητα για απεριόριστες κυτταρικές διαιρέσεις. Τελικά απενεργοποιούνται οι μηχανισμοί της απόπτωσης και του κυτταρικού θανάτου μετατρέποντας τα κύτταρα σε αθανατοποιημένα (αθανατοποίηση).

2. Τα αθανατοποιημένα κύτταρα αποκτούν την ικανότητα να παράγουν έναν όγκο σε ένα κατάλληλο ξενιστή (ογκογένεση)(Lestrate et al. 2003).

Ορισμένα ιικά προϊόντα μπορούν να αθανατοποιήσουν τα κύτταρα όπως τα Τ αντιγόνα παποβαϊών (π.χ. πολυώματος, JC, SV40),οι πρώιμες πρωτεΐνες των ιών θηλώματος και πρωτεΐνες του ιού Epstein-Barr (π.χ. EBNA-5). Σε αυτές τις περιπτώσεις οι ιικές πρωτεΐνες μπορούν να αλληλεπιδράσουν και να αδρανοποιήσουν μία ή περισσότερες κυτταρικές ογκοκατασταλτικές πρωτεΐνες (π.χ. Rb, p53) (Lestrate et al 2003) με αποτέλεσμα την απορύθμιση του κυτταρικού κύκλου. Κατά τη διάρκεια του διαταραγμένου κυτταρικού κύκλου, η συσσώρευση μεταλλάξεων μπορεί να συμβεί είτε αυθόρμητα είτε ως αποτέλεσμα άλλων παραγόντων (ιός, χημικοί παράγοντες, ακτινοβολία) σε κυτταρικά ογκογονίδια (π.χ., HRAS, KRAS, c-Myc), σε αντι-ογκογονίδια (π.χ. p53, Rb), ή άλλα κυτταρικά γονίδια.

2.10 ΔΟΜΗ ΧΡΩΜΑΤΙΝΗΣ ΚΑΙ ΙΣΤΟΝΙΚΕΣ ΤΡΟΠΟΠΟΙΗΣΕΙΣ

Το γονιδίωμα των ευκαρυωτικών οργανισμών μπορεί να φτάσει την τάξη μεγέθους των 10^9 ζευγών βάσεων. Ωστόσο αυτό το τεράστιο σε μήκος DNA έχει τη δυνατότητα να συμπυκνώνεται ώστε να εμπεριέχεται στον πολύ μικρό πυρήνα διαμέτρου περίπου 6mm. Το πρώτο επίπεδο της οργάνωσης της χρωματίνης είναι το νουκλεόσωμα το οποίο αποτελείται από ένα οκταμερές που περιέχει δύο μόρια της καθεμιάς από τις τέσσερεις ιστόνες, H2A, H2B, H3 και H4 γύρω από τις οποίες περιελίσσονται με 146 ζεύγη βάσεων (bp) του DNA (Oudet P et al. 1975). Το οκταμερές των ιστονών που αποτελεί και τον πυρήνα του νουκλεοσώματος εμφανίζει υδρόφοβο χαρακτήρα, με τα θετικά φορτισμένα αμινοτελικά άκρα των ιστονών να προεκβάλλουν στην εξωτερική επιφάνεια του συμπλόκου. Τα μόρια των ιστονών διαθέτουν πανομοιότυπο δομικό πρότυπο που εμφανίζει 3 α-έλικες ενωμένες με 2 βρόγχους, ενώ τα μόρια H2A-H2B και H3-H4 αντίστοιχα σχηματίζουν ετεροδιμερή. Η κρυσταλλογραφική απεικόνιση της δομής του νουκλεοσώματος (Luger et al. 1997) δείχνει πως η διπλή έλικα του DNA βρίσκεται σε επαφή με το οκταμερές ανά 10 ζεύγη βάσεων, ενώ η αλληλεπίδραση σε ορισμένα σημεία επαφής είναι ισχυρότερη.

Έχει πραγματοποιηθεί συσχέτιση των ιστονών ή των τροποποιήσεών τους με τη δομή της χρωματίνης και κατ' επέκταση με τη γονιδιακή ρύθμιση. Υπάρχει μια ποικιλία γνωστών μετα-μεταφραστικών τροποποιήσεων των ιστονών, με πιο κοινές τις ακετυλίωση, μεθυλίωση, φωσφορυλίωση και ουβικιτινίωση. Οι τροποποιήσεις συμβαίνουν κυρίως στις αμινο-τελικές ουρές των τεσσάρων ιστονών. Έχει δειχθεί ότι διαφορετικές τροποποιήσεις μπορεί να υπάρχουν στις ίδιες ιστόνες και ο συνδυασμός τους καθορίζει την έκφραση ή την αποσιώπηση των γονιδίων.

Για παράδειγμα οι μόνο,δι και τρι-μεθυλιώσεις της λυσίνης στο άκρο των H3, H4 και H2B σχετίζονται με την ενεργοποίηση γονιδίων, ενώ οι δι-και τρι-μεθυλίωση της H3K9 και H3K27 σχετίζονται με την αποσιώπηση των γονιδίων. Μία από τις πιο σημαντικές τροποποιήσεις των ιστονών είναι η ακετυλίωση των καταλοίπων λυσίνης με ακετυλοτρανσφεράσες ιστόνης που οδηγεί στην ενεργοποίηση των γονιδίων. Η ακετυλίωση της H3K9 και H3K14 σχετίζεται με την ενεργοποίηση της μεταγραφής γονιδίων (Akey and Luger 2003).

Όπως προαναφέρθηκε, η ακετυλίωση λυσίνης της ιστόνης σχετίζεται με την ενεργοποίηση των γονιδίων και είναι μια από τις πιο καλά χαρακτηρισμένες τροποποιήσεις των ιστονικών ουρών. Η ακετυλίωση είναι καλύτερα γνωστή για το ρόλο της στην εξουδετέρωση των θετικών φορτίων στις ουρές των ιστονών και ως εκ τούτου τη διατάραξη των μεσονουκλεοσωμικών αλληλεπιδράσεων οδηγώντας στη χαλάρωση των ινών της χρωματίνης (Hanna et al. 2008). Πολλαπλές μελέτες έχουν δείξει ότι η ακετυλίωση στις ουρές της ιστόνης συσχετίζεται με αποσυμπύκνωση της χρωματίνης, μαζί με την μεταβολή της δομής του νουκλεοσώματος (Akey and Luger 2003).

Οι ακέτυλο-ομάδες προστίθενται από ακετυλοτρανσφεράσες ιστόνης και μπορούν να απομακρυνθούν με ιστονικές αποακετυλάσες (HDACs). Η ακετυλίωση των ιστονών έχει συσχετιστεί με την ενεργοποίηση της μεταγραφής και η απομάκρυνση των ομάδων ακετυλίου είναι ένας μηχανισμός μεταγραφικής καταστολής (McBryant et al, 2003). Η ακετυλίωση των ιστονών αυξάνει την ικανότητα των μεταγραφικών παραγόντων να προσδένονται στο DNA (Haushalter and Kadonaga 2003). Επιπλέον, η παρουσία μιας ομάδας ακετυλίου μπορεί να ενεργοποιήσει τη μεταγραφή ευνοώντας τη δέσμευση των μεταγραφικών παραγόντων, που περιέχουν υπομονάδες που αναγνωρίζουν τις ακετυλο-ομάδες, στο DNA (Georges et al. 2003). Οι ακέτυλο-ομάδες μπορούν επίσης να μεταβάλλουν τις ιδιότητες της χρωματίνης και να προωθούν την αποσυμπύκνωσή της μέσω της εξασθένισης των αλληλεπιδράσεων των ουρών των ιστονών και των γειτονικών νουκλεοσωμάτων.

2.11.1 ΜΙΚΡΟΣΥΣΤΟΙΧΙΕΣ ΚΑΙ ΒΙΟΛΟΓΙΑ

Οι μικροσυστοιχίες αποτελούν μια σύγχρονη μέθοδο ποσοτικής μέτρησης του μεταγραφώματος, πρωτεώματος κ.α. που χρησιμοποιούνται ευρέως στη μοριακή βιολογία και στην ιατρική. Υπάρχουν διάφορα είδη μικροσυστοιχιών όπως DNA μικροσυστοιχίες, πρωτεϊνικές μικροσυστοιχίες, πεπτιδικές μικροσυστοιχίες, μικροσυστοιχίες ιστού, αντισωμικές μικροσυστοιχίες, φαινοτυπικές μικροσυστοιχίες. Αυτές χρησιμοποιούνται για τη μέτρηση των επιπέδων έκφρασης των γονιδίων, τις πρωτεϊνικές αλληλεπιδράσεις, τον εντοπισμό καρκινικών ιστών και την αλληλεπίδραση αντισώματος-αντιγόνου.

Η τεχνολογία των DNA μικροσυστοιχιών εξελίχθηκε από την τεχνική Southern blotting στην οποία θραύσματα DNA είναι προσκολλημένα σε ένα υπόστρωμα και στη συνέχεια σημαίνονται με μια γνωστή DNA ακολουθία. Η χρήση ολιγονουκλεοτιδίων DNA σε σειρά για τη δημιουργία προφίλ γονιδιακής έκφρασης περιγράφηκε το 1987 και χρησιμοποιήθηκαν για την αναγνώριση γονιδίων των οποίων η έκφραση ρυθμίζεται από την ιντερφερόνη (Kulesh DA et al. 1987). Οι πρώιμες συστοιχίες δημιουργήθηκαν με την εναπόθεση cDNA σε ημιδιαπερατό χαρτί. Η χρήση μικροσυστοιχιών για τη δημιουργία προφίλ γονιδιακής έκφρασης πραγματοποιήθηκε το 1995 (Schena et al. 1995) και το πρώτο πλακάκι μικροσυστοιχιών που περιείχε το συνολικό γονιδίωμα του Σακχαρομύκητα δημιουργήθηκε το 1997 (Lashkari DA et al. 1997).

Η μικροσυστοιχία είναι μια πλακέτα η οποία είναι φτιαγμένη από γυαλί, πλαστικό ή σιλικόνη πάνω στην οποία εναποτίθενται χιλιάδες ή εκατομμύρια ολιγονουκλεοτιδικές αλληλουχίες με μήκος 25-200 βάσεις DNA και είναι σχεδιασμένες ώστε η κάθε μια αλληλουχία να είναι συμπληρωματική προς ένα μοναδικό γονίδιο. Η μικροσυστοιχία αυτή ονομάζεται chip, σχεδιάζεται από εταιρείες όπως η Affymetrix και η Agilent (Εικόνα 1) και περιέχει ολιγονουκλεοτίδια συμπληρωματικά προς το γονιδίωμα του προς μελέτη οργανισμού. Η πρόοδος της τεχνολογίας επιτρέπει την ταυτόχρονη παρακολούθηση της έκφρασης περισσότερων από 30.000 μεταγράφων.

Εικόνα 1: Affymetrix chips

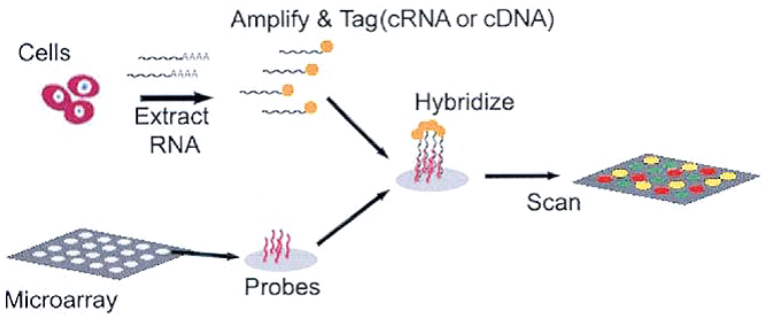
Η τεχνολογία μικροσυστοιχιών επιτρέπει την εφαρμογή τους για διαφορετικούς ερευνητικούς σκοπούς όπως:

1) Δημιουργία γενετικού προφίλ και μελέτης της αλλαγής της γονιδιακής έκφρασης μεταξύ υγιούς και ασθενούς κατάστασης, κατόπιν επίδρασης σε φυσιολογικές κυτταρικές σειρές με κάποιο ερέθισμα όπως χημικοί παράγοντες , βακτηριακή μόλυνση , ιική μόλυνση κ.α.
2) Σύγκριση του γονιδιώματος διαφορετικών κυττάρων.
3) Αναγνώριση του γονιδιώματος κάποιου οργανισμού για την ανίχνευση του παθογόνου μικροοργανισμού που προκάλεσε κάποια ασθένεια ή την ανίχνευση μικροοργανισμών σε τρόφιμα.
4) Προσδιορισμός πολυμορφισμών (SNP) για ερευνητικούς ή διαγνωστικούς σκοπούς.
5) Εντοπισμός εναλλακτικού ματίσματος (exon junction array).
6) Εντοπισμός συγχωνευμένων γονιδίων (fusion genes microarray) για παράδειγμα η συγχώνευση των γονιδίων BCR-ABL στον καρκίνο του προστάτη.
7) Εντοπισμός τμημάτων DNA που εκφράζονται (Tiling array).

Οι πιο διαδεδομένες από τις μικροσυστοιχίες είναι οι DNA μικροσυστοιχίες οι οποίες χρησιμοποιούνται για την ταυτοποίηση της αλλαγής της έκφρασης των γονιδίων μεταξύ διαφορετικών καταστάσεων, για παράδειγμα σύγκριση δειγμάτων από υγιή και ασθενή άτομα, επίδραση κάποιου χημικού παράγοντα, βακτηριακής ή ιικής μόλυνσης. Αυτές οι μικροσυστοιχίες χρησιμοποιήθηκαν και στην παρούσα διπλωματική εργασία. Από την ανάλυση των δεδομένων από τις μικροσυστοιχίες είναι δυνατή η κατανόηση των μοριακών μηχανισμών και των σηματοδοτικών μονοπατιών που χρησιμοποιούν τα κύτταρα σαν απόκριση των ερεθισμάτων που δέχονται και προκειμένου να ενεργοποιήσουν διάφορες λειτουργίες όπως η απόπτωση , η ενεργοποίηση της άμυνας , η έκκριση κυτταροκινών κ.α.

Η πειραματική διαδικασία που ακολουθείται είναι η εξής (Εικόνα 2):

Ολικό RNA εξάγεται από τον ιστό του δείγματος που μας ενδιαφέρει. Απομονώνουμε το mRNA το οποίο αποτελεί ένα κλάσμα του ολικού RNA και αντιστοιχεί στα γονίδια που εκφράζονται. Στη συνέχεια το πολλαπλασιάζουμε και προσθέτουμε φθορίζοντες ιχνηθέτες στο κάθε μόριο. Όλα αυτά τα μόρια επωάζονται με το chip των μικροσυστοιχιών και τα μόρια mRNA που φέρουν τους φθορίζοντες ιχνηθέτες υβριδοποιούνται με όσα ολιγονουκλεοτίδια του chip έχουν συμπληρωματική αλληλουχία. Η υβριδοποίηση γίνεται σε αυστηρές συνθήκες για να αποφευχθεί η τυχαία ένωση mRNA και ολιγονουκλεοτιδίων που δεν είναι πλήρως συμπληρωματικά. Στη συνέχεια η μικροσυστοιχία τοποθετείται μέσα σε έναν σαρωτή ο οποίος καταγράφει την ένταση της φθορίζουσας ουσίας. Η ένταση του φθορισμού έχει αναλογική σχέση με τον αριθμό των μορίων RNA που έχουν υβριδοποιηθεί και κατά συνέπεια αντιπροσωπεύουν το επίπεδο της έκφρασης γονιδίων. Έτσι μπορεί να δημιουργηθεί το προφίλ της γονιδιακής έκφρασης για κάποιο ιστό από διαφορετικά άτομα.



Εικόνα 2: Σχηματική αναπαράσταση της διαδικασίας των DNA μικροσυστοιχιών.

Ακολούθως πραγματοποιείται η βιοπληροφορική και στατιστική ανάλυση των δεδομένων.

2.11.2 ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ ΑΝΑΛΥΣΗ ΜΙΚΡΟΣΥΣΤΟΙΧΙΩΝ

Τα πρωταρχικά δεδομένα (Raw data) πρέπει να ελεγχθούν για τη ποιότητά τους ώστε να διασφαλισθεί η ακεραιότητά τους. Τα πρωταρχικά δεδομένα έχουν κάποιο επίπεδο θορύβου και πρέπει να επεξεργασθούν για την αφαίρεση μη επιθυμητών πηγών θορύβου για να επιτευχθεί υψηλότερο ποσοστό ακρίβειας.

ΔΙΟΡΘΩΣΗ ΥΠΟΒΑΘΡΟΥ(BACKGROUND CORRECTION)

Αρχικά διορθώνεται το υπόβαθρο των εντάσεων φθορισμού του κάθε σημείου (spot). Ο φθορισμός του υποβάθρου μπορεί να προκύψει από πολλές πηγές όπως τη μη ειδική πρόσδεση σημασμένου δείγματος στην επιφάνεια των μικροσυστοιχιών, επίδραση διαφόρων παραγόντων όπως ιζήματα που παραμένουν ύστερα από τη διαδικασία των διαδοχικών ξεπλυμάτων ή οπτικού θορύβου από τη σάρωση της μικροσυστοιχίας. Διαφορετικοί αλγόριθμοι χρησιμοποιούν διαφορετικές μεθοδολογίες διόρθωσης του υποβάθρου. Για παράδειγμα το αλγόριθμος RMA χρησιμοποιεί ένα γραμμικό μοντέλο για τη κανονικοποίηση του σήματος και τη κατανομή του θορύβου.

ΚΑΝΟΝΙΚΟΠΟΙΗΣΗ (NORMALIZATION)

Στη συνέχεια πραγματοποιείται η κανονικοποίηση. Ο σκοπός αυτού του βήματος είναι η προσαρμογή των δεδομένων ως προς τον θόρυβο εξαιτίας του οποίου θα εμφανιστούν στατιστικά σημαντικές διαφορές μεταξύ των δειγμάτων προς μελέτη οι οποίες δεν είναι αποτέλεσμα της ύπαρξης διαφορών στα επίπεδα του mRNA για τα συγκεκριμένα δείγματα. Για παράδειγμα οι διαφορές στην ποσότητα του RNA στα δείγματα μάρτυρα και στα δείγματα κατόπιν ιικής μόλυνσης που χρησιμοποιείται, η χρονική διάρκεια που παραμένει το δείγμα για να υβριδοποιηθεί με τη μικροσυστοιχία ή ο όγκος του δείγματος μπορούν να εισάγουν σημαντικό θόρυβο μεταξύ των 2 δειγμάτων προς σύγκριση. Ακόμη και μικρές φυσικές διαφορές μεταξύ των πλακετών ή των σαρωτών που χρησιμοποιούνται για να ανιχνεύσουν την ένταση του φθορισμού μπορεί να έχουν κάποια επίδραση.

Συνοπτικά η κανονικοποίηση διασφαλίζει ότι στη περίπτωση της σύγκρισης των διαφορετικών επιπέδων γονιδιακής έκφρασης μεταξύ δύο δειγμάτων οι διαφοροποιήσεις προέρχονται από βιολογικές παραμέτρους και δεν επηρεάζονται από τα επίπεδα θορύβου της μικροσυστοιχίας.

ΔΙΟΡΘΩΣΗ PM(PM CORRECTION)

Οι ολιγονουκλεοτιδικοί ιχνηθέτες ακριβούς στοίχισης (PM probe) χρησιμοποιούνται για τη μέτρηση της έκφρασης ενός γονιδίου η οποία προκύπτει όταν ένα mRNA προσδένεται σε έναν ιχνηθέτη. Οι ιχνηθέτες μέτριας στοίχισης (MM probe) είναι σχεδιασμένοι για να μετράνε τη μη ειδική πρόσδεση. Κατόπιν η τιμή από έναν ιχνηθέτη μέτριας στοίχισης πρέπει να αφαιρεθεί από τον αντίστοιχό του ιχνηθέτη ακριβούς στοίχισης.

Στη πραγματικότητα όμως το 30% των ιχνηθετών μέτριας στοίχισης έχουν μεγαλύτερη τιμή από τους ιχνηθέτες ακριβούς στοίχισης. Πολλές από τις πιο διαδεδομένες μεθόδους επεξεργασίας λύνουν αυτό το πρόβλημα αγνοώντας τους ιχνηθέτες μέτριας στοίχισης και χρησιμοποιούν άλλες μεθόδους για να διορθώνουν τη μη ειδική πρόσδεση στους ιχνηθέτες ακριβούς στοίχισης.

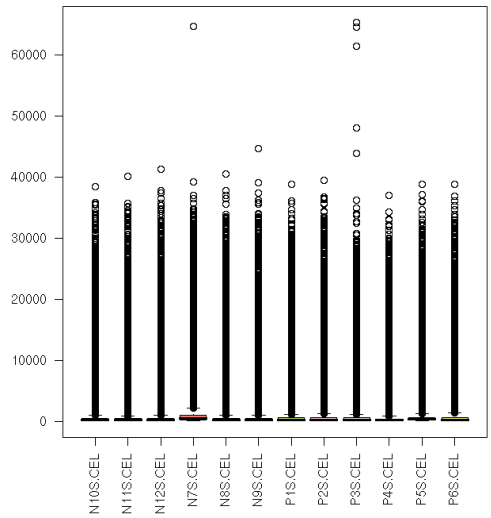
ΑΝΑΣΥΓΚΡΟΤΗΣΗ (SUMMARIZATION)

Μια πλακέτα DNA μικροσυστοιχιών αποτελείται από 11 ζεύγη ιχνηθετών ακριβούς και μέτριας στοίχισης που στοχεύουν το mRNA ενός γονιδίου. Το τελικό στάδιο της ανάλυσης είναι να συνοψισθούν τα δεδομένα από τα 11 διαφορετικά ζεύγη ιχνηθετών για να δώσουν μια τιμή για την έκφραση του γονιδίου που μετρούν.

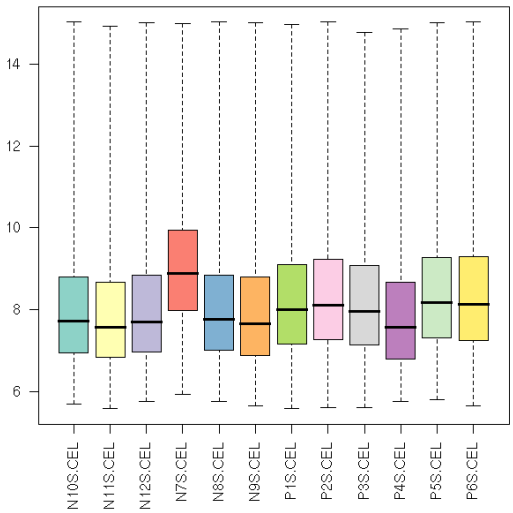
2.11.3 ΠΟΙΟΤΙΚΟΣ ΕΛΕΓΧΟΣ ΜΙΚΡΟΣΥΣΤΟΙΧΙΩΝ

ΘΗΚΟΓΡΑΜΜΑΤΑ

Το θηκόγραμμα είναι ένας εύχρηστος τρόπος για τη σύγκριση του επιπέδου της έντασης του φθορισμού μεταξύ διαφορετικών πλακετών μικροσυστοιχιών. Η γραμμή στη μέση του τετραγώνου αντιπροσωπεύει τη διάμεσο. Οι οριζόντιες γραμμές που συνδέονται με διακεκομμένες γραμμές με το τετράγωνο υποδεικνύουν την ελάχιστη και τη μέγιστη τιμή που δεν θεωρούνται εκτός ορίων. Εάν μια πλακέτα ή περισσότερες έχουν σημαντικά διαφορετικά επίπεδα έντασης φθορισμού από τις υπόλοιπες πλακέτες με τις οποίες συγκρίνονται και είναι ενδεικτικό της ύπαρξης κάποιου προβλήματος. Τέτοιου είδους προβλήματα συχνά διορθώνονται με την κανονικοποίηση. Για δεδομένα από μικροσυστοιχίες χρησιμοποιούνται οι λογαριθμισμένες τιμές των εντάσεων των ιχνηθετών ώστε τα διαγράμματα να είναι αναγνώσιμα (Εικόνες 3,4).

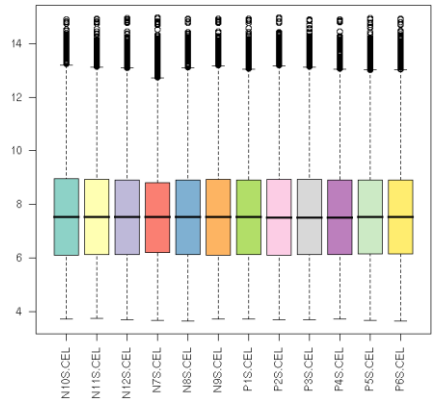


Εικόνα 3:Θηκόγραμμα με τις πρωταρχικές τιμές έντασης του φθορισμού των ιχνηθετών.

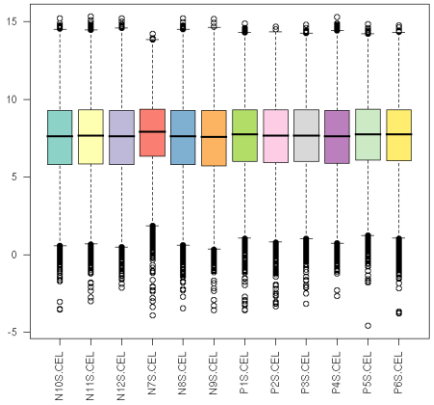


Εικόνα 4:Θηκόγραμμα με τις λογαριθμισμένες τιμές έντασης του φθορισμού των ιχνηθετών.

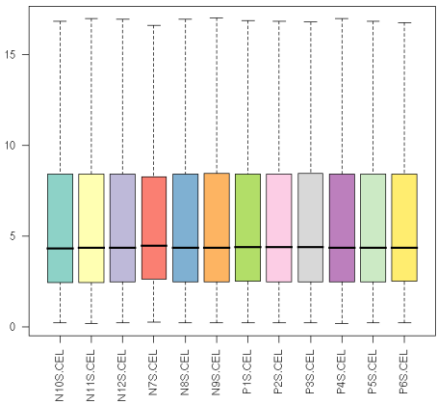
Στο παραπάνω θηκόγραμμα (Εικόνα 4) παρατηρείται ότι η τέταρτη πλακέτα από αριστερά έχει υψηλότερες τιμές εντάσεων από τις υπόλοιπες, ενδεικτικό κάποιου προβλήματος στη συγκεκριμένη. Στη συνέχεια θα πραγματοποιηθεί κανονικοποίηση που πιθανώς να διορθώσει αυτό το πρόβλημα. Στις επόμενες εικόνες φαίνονται τα θηκογράμματα των εντάσεων των ιχνηθετών αφού χρησιμοποιήθηκαν οι αλγόριθμοι MAS5, RMA και GCRMA (Εικόνες 5-7).



Εικόνα 5: Επεξεργασμένες εντάσεις φθορισμού με τον αλγόριθμο MAS5.



Εικόνα 6: Επεξεργασμένες εντάσεις φθορισμού με τον αλγόριθμο RMA.



Εικόνα 7: Επεξεργασμένες εντάσεις φθορισμού με τον αλγόριθμο GCRMA.

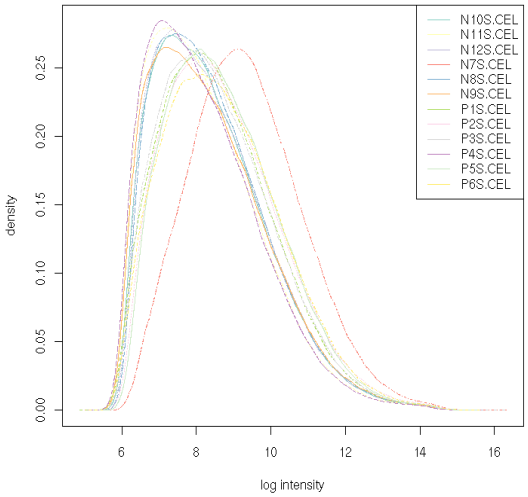
Στις παραπάνω εικόνες φαίνεται πως οι διαφορετικοί αλγόριθμοι επηρεάζουν σημαντικά τα πρωταρχικά δεδομένα με διαφορετικό τρόπο.

ΙΣΤΟΓΡΑΜΜΑΤΑ

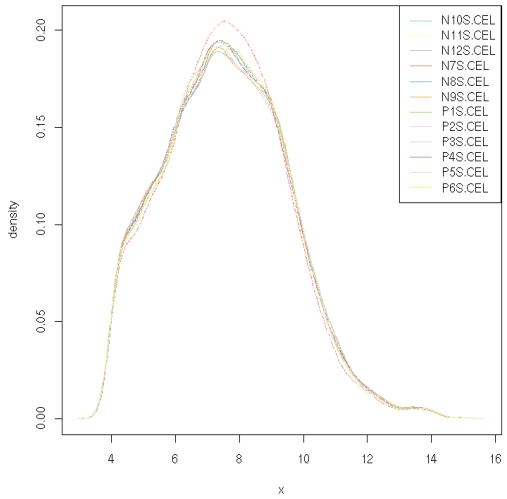
Ένα ιστόγραμμα των επιπέδων έντασης του φθορισμού μιας πλακέτας παρέχει παρόμοιες πληροφορίες με το θηκόγραμμα. Χρησιμοποιείται για να οπτικοποιήσει την κατανομή των δεδομένων και να συγκρίνει τις εντάσεις των ιχνηθετών μεταξύ των πλακετών ενός πειράματος. Ο άξονας των Y αντιπροσωπεύει την πυκνότητα των ιχνηθετών, ενώ ο άξονας των X αντιπροσωπεύει την ένταση των ιχνηθετών (Εικόνα 8). Αυτό το διάγραμμα μας παρέχει μια πιο λεπτομερή εικόνα για τα δεδομένα και μπορούν να εξαχθούν διάφορα συμπεράσματα όπως:

- μια δικόρυφη κατανομή των πρωταρχικών δεδομένων συχνά υποδεικνύει ότι μια πλακέτα περιέχει ένα χωρικό πλασματικό εύρημα (spatial artifact)
- ένα γράφημα που είναι μετατοπισμένο προς τα δεξιά συνήθως περιέχει μη φυσιολογικές υψηλές παρεμβολές θορύβου από το υπόβαθρο

Όπως φαίνεται στο παρακάτω διάγραμμα η πλακέτα με το όνομα NS7.CEL εμφανίζει πρόβλημα καθώς είναι μετατοπισμένη προς τα δεξιά. Στην Εικόνα 9 φαίνεται το ιστόγραμμα κατόπιν της χρήσης του αλγορίθμου RMA για την κανονικοποίηση και η συγκεκριμένη πλακέτα εξακολουθεί να έχει υψηλές τιμές.



Εικόνα 8: Ιστόγραμμα με τις πρωταρχικές τιμές έντασης του φθορισμού των ιχνηθετών.



Εικόνα 9: Ιστόγραμμα κατόπιν κανονικοποίησης με τον RMA αλγόριθμο.

2.11.4 ΑΛΓΟΡΙΘΜΟΣ MAS5

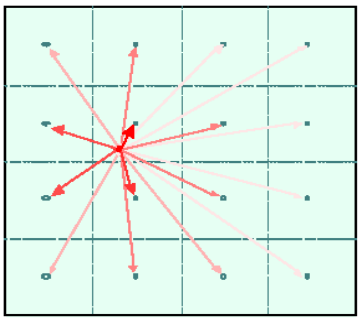
Ο MAS5 αλγόριθμος αναπτύχθηκε από την εταιρεία Affymetrix (Statistical Algorithms Description Document 2002). Συνδυάζει τα σήματα από τους PM και MM ιχνηθέτες σε μια τιμή και μετρά με ευαισθησία και ακρίβεια την έκφραση του γονιδίου. Αυτό το πετυχαίνει αυτό υπολογίζοντας ένα εύρωστο μέσο όρο των λογαριθμισμένων PM και MM (Pepper SD et al. 2007). Ο αλγόριθμος λαμβάνει υπόψη τις τιμές των PM και MM ιχνηθετών και διορθώνει το υπόβαθρο. Στη συνέχεια οι τιμές των MM μετατρέπονται σε ιδανική αντιστοιχία όπου η τιμή είναι πάντα μικρότερη από αυτή των PM. Όπως αναφέραμε και προηγουμένως στην περίπτωση όπου η τιμή των MM είναι μεγαλύτερη από την τιμή του αντίστοιχου PM τότε η τιμή MM δεν λαμβάνεται υπόψη.

Στη συνέχεια ο αλγόριθμος παράγει μια τιμή ανίχνευσης p-value, για να υπολογίζει τις τιμές που θεωρούνται μετρήσιμες. Αυτές οι τιμές καθορίζουν εάν ένα μετάγραφο είναι αξιόπιστα ανιχνεύσιμο (Present), μη ανιχνεύσιμο (Absent) ή οριακά ανιχνεύσιμο (Marginal). Επιπλέον μια τιμή σήματος υπολογίζεται και καθορίζει τη σχετική αφθονία του μεταγράφου (Statistical reference Guide).

Η τιμή που μετρά ο αλγόριθμος υπολογίζεται από τον τύπο MS=N+P+S

όπου MS=Measure value,τιμή υπολογισμού N=Noise(θόρυβος), P=Probe effect (επίδραση ιχνηθέτη -μη ειδικός υβριδισμός) και S=Signal (σήμα)

Για να υπολογίσει ο αλγόριθμος το υπόβαθρο διαιρεί τη μικροσυστοιχία σε ζώνες (Εικόνα 10).



Εικόνα 10: Διαίρεση της μικροσυστοιχίας σε ζώνες.

Επιλέγει το 2% των χαμηλότερων εντάσεων φθορισμού, υπολογίζει την τυπική απόκλιση αυτών των τιμών η οποία ορίζεται ως η μεταβλητότητα της ζώνης. Το υπόβαθρο σε οποιαδήποτε θέση είναι το άθροισμα από το υπόβαθρο όλων των ζωνών σταθμισμένο με $\frac{1}{d^2 + ff}$ όπου d η απόσταση και ff= fudge factor (παράγοντας παραποίησης).

Στη συνέχεια υπολογίζεται ο θόρυβος στη μικροσυστοιχία. Χρησιμοποιούνται οι ίδιες ζώνες όπως στον υπολογισμό του υποβάθρου. Έπειτα επιλέγεται το 2% των χαμηλότερων τιμών υποβάθρου, υπολογίζεται η τυπική απόκλιση και η τιμή αυτή ορίζεται ως ο θόρυβος της ζώνης. Ο θόρυβος σε οποιαδήποτε θέση υπολογίζεται από το άθροισμα όλων των ζωνών όπως παραπάνω.

Η προσαρμοσμένη ένταση υπολογίζεται από τον τύπο:
A(x,y)=max(I'(x,y) – b(x,y), NoiseFrac*n(x,y))
όπου I'(x,y)=max(I'(x,y), 0.5)
A = προσαρμοσμένη ένταση
I = ένταση που μετράται
b =υπόβαθρο
NoiseFrac: έχει τιμή 0.5 (παράγοντας παραποίησης)

Ο MAS5 υπολογίζει τη διακρίνουσα για κάθε ζεύγος ιχνηθετών σύμφωνα με το τύπο R=(PM-MM)/(PM+MM). Στη συνέχεια χρησιμοποιεί το τεστ Wilcoxon one sided ranked test για να συγκρίνει τις τιμές R και Tau και καθορίζει την τιμή p-value.

Οι τιμές χαρακτηρίζονται ως Present/Marginal/Absent με βάση τις τιμές p-value ως εξής:
Present $\leq$ alpha1
alpha1 < **Marginal** < alpha2
Alpha2 $\leq$ **Absent**
όπου alpha1=0.04, alpha2=0.06, tau=0.015

Συνοπτικά ο MAS5 είναι χρήσιμος για την ανάλυση μεμονωμένων πλακετών και δίνει τιμές p-value για τα δεδομένα της γονιδιακής έκφρασης. Είναι η πιο διαδεδομένη μέθοδος ανάλυσης για πλακέτες της Affymetrix και βασίζεται στις τιμές των MM ιχνηθετών.

2.11.5 ΑΛΓΟΡΙΘΜΟΣ RMA

Ο RMA αλγόριθμος -Robust Multi-array Average (Irizarry et al.2003) αναπτύχθηκε για να αντικαταστήσει τον MAS5 της Affymetrix. Η διαφορά αυτών των αλγορίθμων είναι ότι ο RMA δε λαμβάνει υπόψη τις τιμές των MM ιχνηθετών. Ο λόγος που συμβαίνει αυτό είναι ότι οι MM ιχνηθέτες προσθέτουν περισσότερο θόρυβο στο πείραμα. Ο RMA αλγόριθμος ρυθμίζει τον θόρυβο του υποβάθρου στις μη επεξεργασμένες τιμές της έντασης φθορισμού. Στη συνέχεια λαμβάνεται η διορθωμένη τιμή του κάθε PM ιχνηθέτη αφού λογαριθμισθεί ($\log_2$) και ύστερα αυτές οι τιμές κανονικοποιούνται με την κανονικοποίηση τεταρτημορίων (quantile normalization).

Αναλυτικά ο αλγόριθμος περιλαμβάνει 3 βήματα (Analysis of Microarray Data, Genestat):

1. Διόρθωση υποβάθρου βασισμένη σε ένα γραμμικό μοντέλο όπου ένα σφάλμα στην ένταση του φθορισμού υπολογίζεται και εξαλείφεται.
2. Μη-γραμμική κανονικοποίηση τεταρτημορίων όπου η κάθε πλακέτα κανονικοποιείται για να έχει την ίδια συσσωρευτική συχνότητα κατανομής
3. Ανασυγκρότηση μικροσυστοιχίας όπου η τιμή της διαμέσου για κάθε σετ ιχνηθετών ρυθμίζεται για διαφορές στο πλακάκι και υπολογίζεται.

ΔΙΟΡΘΩΣΗ ΥΠΟΒΑΘΡΟΥ

Ο αλγόριθμος χρησιμοποιεί το μέσο του 2% των χαμηλότερων κελιών σε 16 ζώνες σε όλη τη μικροσυστοιχία και στη συνέχεια δημιουργεί το σταθμισμένο μέσο όρο των κελιών με τη στάθμιση να εξαρτάται από το τετράγωνο της απόστασης από το κέντρο του κελιού. Στη συνέχεια ο αλγόριθμος χρησιμοποιεί ένα μοντέλο θορύβου στις εντάσεις των PM ιχνηθετών. Ο RMA αλγόριθμος αγνοεί τις τιμές των MM ιχνηθετών. Η διόρθωση του υποβάθρου μέσω του RMA βασίζεται σε ένα ελικοειδές μοντέλο. Το μοντέλο αυτό υποστηρίζει ότι το σήμα S που μετράται από τον μάρτυρα, αποτελείται από δύο συνιστώσες S = X + Y, όπου X είναι το σήμα και Y ο θόρυβος. Βασιζόμενο στο τυπικό γράφημα πυκνότητας των εντάσεων, το ελικοειδές μοντέλο υποθέτει ότι το σήμα X κατανέμεται με βάση την εκθετική κατανομή exp(α) και ότι το Y κατανέμεται με βάση την κανονική κατανομή N(μ, σ^2). Η ένταση που υπολογίζεται για κάθε μάρτυρα μετά τη διόρθωση του θορύβου δίνεται από την αναμενόμενη τιμή του X δεδομένης της παρατηρηθείσας τιμής, σύμφωνα με τον τύπο:

$$E(X / S) = \alpha + b\frac{\phi(\alpha / b)}{\Phi(\alpha / b)},$$

όπου $a = s - \mu - \sigma^2\alpha$ και $b = \sigma$.

ΚΑΝΟΝΙΚΟΠΟΙΗΣΗ

Το επόμενο βήμα στην ανάλυση με τον RMA αλγόριθμο είναι η κανονικοποίηση του διορθωμένου υποβάθρου έτσι ώστε η κάθε πλακέτα να έχει την ίδια συσσωρευτική πυκνότητα (κανονικοποίηση τεταρτημορίων). Ο αλγόριθμος το πραγματοποιεί αυτό κατατάσσοντας τις εντάσεις σε κάθε πλακέτα, υπολογίζει το μέσο όρο των αποτελεσμάτων κάθε τάξης (χρησιμοποιώντας τη διάμεσο, το μέσο ή το γεωμετρικό μέσο) και στη συνέχεια αντικαθιστά τις τιμές σε όλες τις πλακέτες με τους μέσους όρους.

ΑΝΑΣΥΓΚΡΟΤΗΣΗ ΜΙΚΡΟΣΥΣΤΟΙΧΙΑΣ

Η ανασυγκρότηση του RMA αλγόριθμου περιλαμβάνει τον υπολογισμό των εντάσεων έκφρασης με τη χρήση εξομάλυνσης των τιμών με βάση τη διάμεσο (median polish). Ο αλγόριθμος του median polish είναι μία μέθοδος εφαρμογής του ακόλουθου μοντέλου:

$$\log_2(y_{n_{ij}}) = \mu_n + \theta_{n_j} + a_{n_i} + \varepsilon_{n_{ij}}$$

όπου ισχύει:

$median_i(\theta_{n_{\xi}}) = median_i(a_{n_i}) = 0$ και $median_i(\varepsilon_{n_{ij}}) = median_i(\varepsilon_{n_i}) = 0$

για οποιαδήποτε τιμή του n. Οι RMA $\log_2$ τιμές έκφρασης δίνονται από τον τύπο $\hat{\beta}_{nj} = \hat{\mu}_n + \hat{\theta}_{nj}$.

Τα πλεονεκτήματα του αλγορίθμου αυτού είναι ότι αποτελεί μία γρήγορη και αξιόπιστη μέθοδο και μπορεί να χρησιμοποιηθεί σε μεγάλο αριθμό μικροσυστοιχιών ταυτόχρονα.

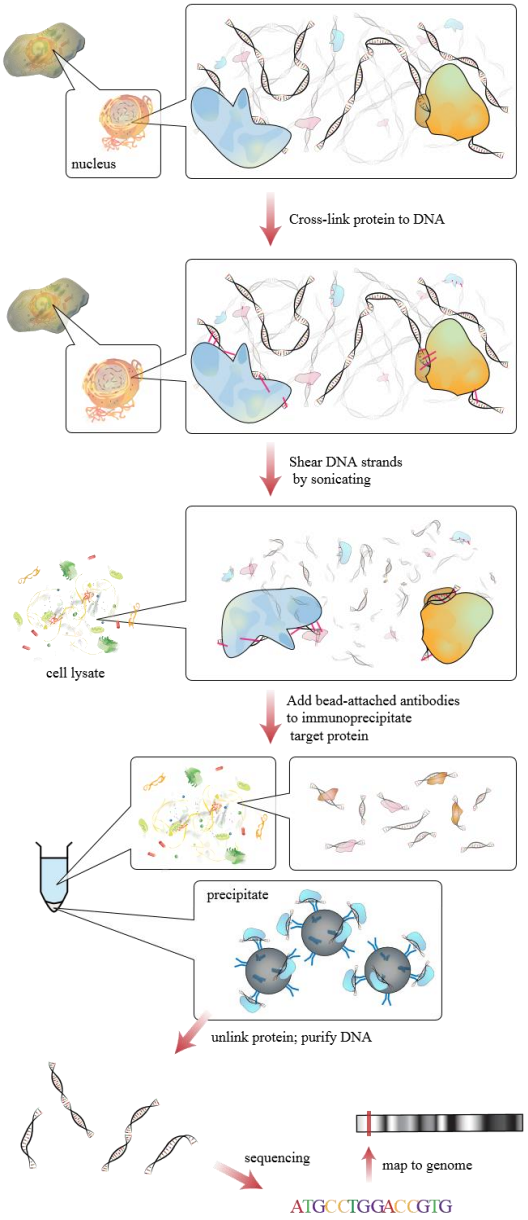
2.12.1 ChIP-seq

Οι μεταγραφικοί παράγοντες και άλλες πρωτεΐνες που σχετίζονται με τη χρωματίνη είναι ιδιαίτερα σημαντικοί καθώς έχουν συσχετιστεί με βιολογικούς μηχανισμούς και την εκδήλωση παθολογικών φαινοτύπων. Ο καθορισμός του τρόπου αλληλεπίδρασης αυτών των παραγόντων με το DNA , με τα οποία ρυθμίζεται η γονιδιακή έκφραση, είναι ουσιώδης για την πλήρη κατανόηση πολλών βιολογικών δραστηριοτήτων και τα στάδια διαφόρων ασθενειών. Αυτές οι πληροφορίες είναι συμπληρωματικές του γονοτύπου και των αναλύσεων της γονιδιακής έκφρασης.

Η τεχνική ChIP-seq (ανοσοκατακρήμνιση χρωματίνης ακολουθούμενη από αλληλούχηση του DNA) είναι μια μέθοδος ανάλυσης πρωτεϊνικών αλληλοεπιδράσεων με το DNA. Η τεχνική αυτή συνδυάζει την ανοσοκατακρήμνιση της χρωματίνης με τη μαζική παράλληλη αλληλούχηση του DNA για να αναγνωρισθούν οι θέσεις πρόσδεσης των πρωτεϊνών στο DNA. Η τεχνική αποτελεί εξέλιξη της ChIP-on-ChIP τεχνικής που χρησιμοποιείτο για τον ίδιο σκοπό. Παραδοσιακές τεχνικές όπως η ChIP-on-ChIP , μπορούν να ανιχνεύσουν τα σημεία πρόσδεσης των μεταγραφικών παραγόντων και συγκεκριμένες DNA σχετιζόμενες πρωτεϊνικές τροποποιήσεις, όμως οι τεχνικές αυτές δεν έχουν την ευαισθησία και την ακρίβεια που διαθέτουν οι σύγχρονες τεχνολογίες αλληλούχησης DNA. Η τεχνολογία των ChIP-seq παρέχει στους ερευνητές τη δυνατότητα να επεκτείνουν το εύρος των μελετών τους και να αναγνωρίσουν τις θέσεις πρόσδεσης των μεταγραφικών παραγόντων σε ολόκληρο το γονιδίωμα με υψηλή ευκρίνεια και χωρίς περιορισμούς.

ΠΕΙΡΑΜΑΤΙΚΗ ΔΙΑΔΙΚΑΣΙΑ

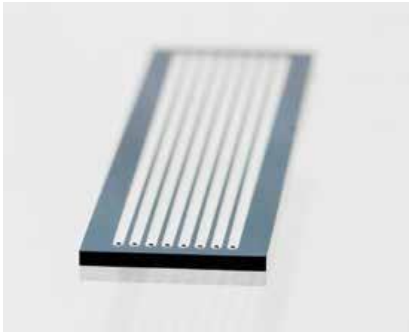
Αρχικά πραγματοποιείται η ανοσοκατακρήμνιση της χρωματίνης. Τα προς μελέτη κύτταρα δέχονται την επίδραση ενός παράγοντα διασύνδεσης (cross-linking), συνήθως φορμαλδεΰδη, ο οποίος προσηλώνει τις πρωτεΐνες στο DNA υπόστρωμά τους μέσα στο κύτταρο. Τα χρωμοσώματα στη συνέχεια απομονώνονται και τμηματοποιούνται με φυσικό τρόπο, με υπερήχους ή με ενζυματική πέψη. Συγκεκριμένες ακολουθίες DNA οι οποίες είναι συνδεδεμένες με κάποια πρωτεΐνη, την οποία θέλουμε να μελετήσουμε, απομονώνονται με καθαρισμό ανοσο-συγγένειας (immune-affinity purification) χρησιμοποιώντας ειδικό αντίσωμα εναντίων της πρωτεΐνης. Το αντίσωμα είναι προσδεμένο με σφαιρίδια και κατακρημνίζεται ύστερα από φυγοκέντριση. Ακολουθεί απομόνωση του συμπλόκου σφαιρίδια – αντίσωμα – πρωτεΐνη – DNA αλληλουχία. Στη συνέχεια αντιστρέφεται η διασταυρωτή σύνδεση (cross-linking)και απομονώνονται τα τμήματα DNA. Ακολουθούν βήματα πολλαπλασιασμού του DNA με αλυσιδωτή αντίδραση πολυμεράσης (PCR) και δημιουργούνται οι βιβλιοθήκες DNA που στη συνέχεια θα αλληλουχηθούν (Εικόνα 11).



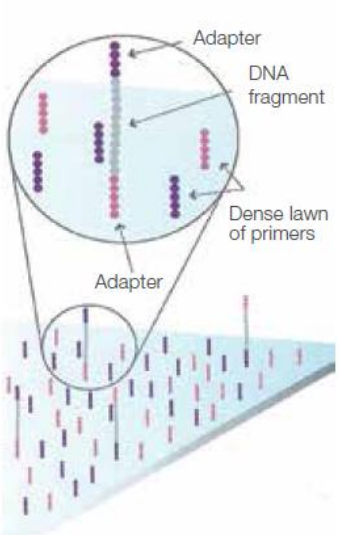
Εικόνα 11: Σχηματική αναπαράσταση της πειραματικής διαδικασίας του ChIP-seq.

Η Illumina είναι μια εταιρεία που πρωτοστατεί στην τεχνολογία των next generation sequencing. Ο τρόπος με τον οποίο αλληλουχεί τις ακολουθίες DNA από το ChIP-seq είναι ο ακόλουθος:
Τα τμήματα DNA προς αλληλούχηση ακινητοποιούνται σε μια επιφάνεια κυτταρικής ροής (Εικόνα 12) που είναι σχεδιασμένη να διατηρεί το DNA με τέτοιο τρόπο ώστε να διευκολύνεται η πρόσβασή του σε ένζυμα και παράλληλα να διασφαλίζει υψηλή σταθερότητα και χαμηλή μη ειδική πρόσδεση φθοριζόντων νουκλεοτιδίων. Τα τμήματα DNA τα οποία έχουν συνδεθεί με τα ειδικά προσαρτήματα (adapters) κατά το στάδιο της δημιουργίας των βιβλιοθηκών (Εικόνα 13) συνδέονται στην επιφάνεια αυτή. Τα μονόκλωνα τμήματα DNA μετατρέπονται σε δίκλωνα (Εικόνα 14). Στη συνέχεια προστίθενται μη σημασμένα νουκλεοτίδια και ένζυμα για να αρχίσει η διαδικασία του πολλαπλασιασμού τύπου γέφυρας (bridge amplification) (Εικόνα 15). Ακολούθως αποδιατάσσονται οι δίκλωνες αλυσίδες αφήνοντας μονόκλωνα DNA τμήματα στην επιφάνεια (Εικόνα 16). Κατόπιν αυτά τα μονόκλωνα τμήματα χρησιμοποιούνται σαν υπόστρωμα για να δημιουργηθούν πυκνά συμπλέγματα δίκλωνων αλυσίδων DNA (Εικόνα 17). Κατά τη διαδικασία αυτή παράγονται τουλάχιστον 1000 πανομοιότυπα αντίτυπα του κάθε αρχικού τμήματος DNA. Στη συνέχεια πραγματοποιείται η αλληλούχηση αυτών των τμημάτων με την τεχνολογία αλληλούχησης κατά την σύνθεση, χρησιμοποιώντας τα τέσσερα νουκλεοτίδια του DNA σημασμένα με διαφορετικού χρώματος φθορίζουσας χρωστικής. Με αυτόν τον τρόπο αλληλουχούνται δεκάδες εκατομμύρια συμπλέγματα DNA στην επιφάνεια παράλληλα.

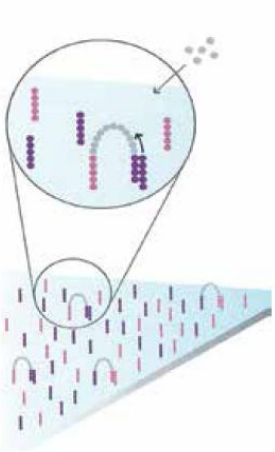
Ο πρώτος κύκλος αλληλούχησης πραγματοποιείται προσθέτοντας τα τέσσερα σημασμένα τερματικά νουκλεοτίδια , εκκινητές (primers) και DNA πολυμεράση. Μετά ακολουθεί διέγερση με laser και ο φθορισμός ανιχνεύεται από έναν σαρωτή (Εικόνα 18). Έτσι καθορίζεται η πρώτη βάση του DNA τμήματος. Επαναλαμβάνεται η ίδια διαδικασία μέχρι τον πλήρη καθορισμό της αλληλουχίας. Η αλληλουχίες που έχουν ανιχνευθεί καλούνται ‘‘διαβάσματα’’ (reads).Τα αρχεία που παράγονται από αυτή τη διαδικασία είναι στη μορφή bcl η οποία μετασχηματίζεται σε fastq, τα οποία θα χρησιμοποιηθούν για την περαιτέρω ανάλυση των δεδομένων.



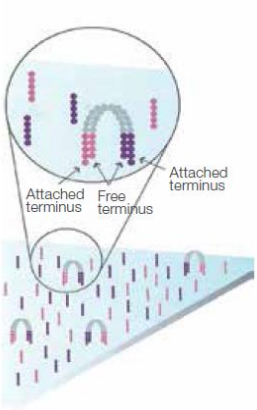
Εικόνα 12: Επιφάνεια κυτταρικής ροής της Illumina όπου μπορούν να αλληλουχηθούν έως οκτώ δείγματα ταυτόχρονα.



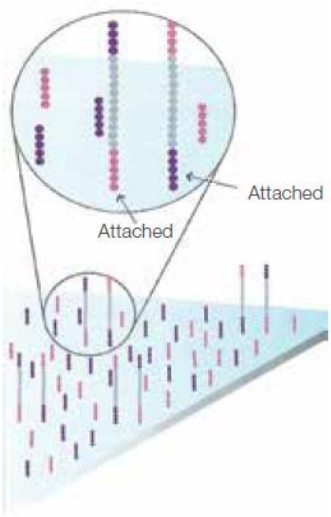
Εικόνα 13: Σύνδεση των τμημάτων DNA με τους υποδοχείς σε τυχαίες θέσης της επιφάνειας.



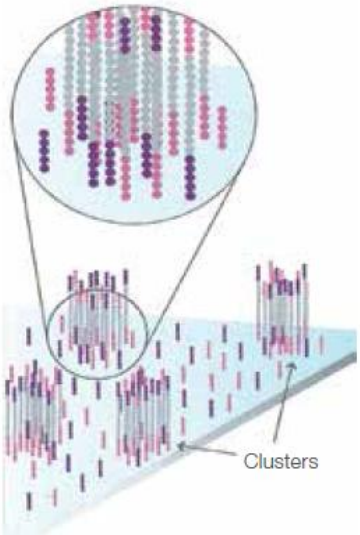
Εικόνα 14: Δημιουργία δίκλωνων αλυσίδων DNA.



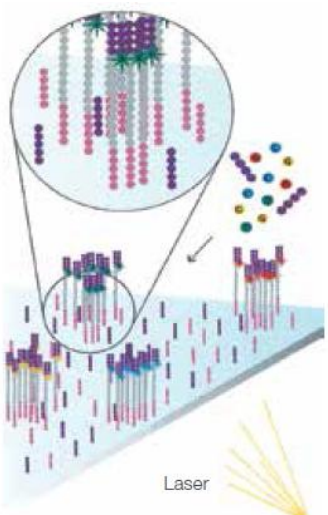
Εικόνα 15: Πολλαπλασιασμός τύπου γέφυρας.



Εικόνα 16: Αποδιάταξη δίκλωνων DNA αλυσίδων.



Εικόνα 17: Δημιουργία πυκνών συστάδων δίκλωνων DNA αλυσίδων.



Εικόνα 18: Ανίχνευση της πρώτης βάσης της DNA αλληλουχίας από τον φθορισμό των ιχνηθετημένων νουκλεοτιδίων.

2.12.2 ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ ΑΝΑΛΥΣΗ ChIP-seq

Η ανάλυση για τα δεδομένα από πειράματα ChIP-seq ξεκινά με τα αρχεία fastq. Τα αρχεία αυτά περιέχουν πληροφορίες για τη βιολογική ακολουθία (στην περίπτωση των ChIP-seq , νουκλεοτιδική αλληλουχία) και μια αντίστοιχη ποιοτική τιμή. Το αρχείο fastq χρησιμοποιεί τέσσερεις γραμμές για κάθε αλληλουχία .

- Η πρώτη γραμμή ξεκινά με ένα χαρακτήρα '' @ '' και ακολουθεί ένα αναγνωριστικό της αλληλουχίας και σε κάποιες περιπτώσεις μια προαιρετική περιγραφή.
- Η δεύτερη γραμμή περιλαμβάνει την ακολουθία .
- Η τρίτη γραμμή αρχίζει με ένα χαρακτήρα '' + '' και προαιρετικά ακολουθεί το ίδιο αναγνωριστικό της αλληλουχίας.
- Η τέταρτη γραμμή κωδικοποιεί τις τιμές της ποιότητας για την ακολουθία της δεύτερης γραμμής και πρέπει να περιέχει τον ίδιο αριθμό συμβόλων όσα και τα νουκλεοτίδια της αλληλουχίας.

Παρακάτω φαίνεται η δομή ενός τέτοιου αρχείου.

@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65

Η αποτελεσματική ανάλυση των ChIP-seq δεδομένων χρειάζεται επαρκή κάλυψη από ''διαβάσματα'' αλληλουχιών (sequence reads). Το απαιτούμενο βάθος αλληλούχησης εξαρτάται κυρίως από το μέγεθος του γονιδιώματος, τον αριθμό και το μέγεθος των σημείων πρόσδεσης της πρωτεΐνης. Εμπειρικά έχει δειχθεί ότι 20 εκατομμύρια ''διαβάσματα'' μπορούν επαρκώς να εντοπίσουν μεταγραφικούς παράγοντες από θηλαστικά. Πρωτεΐνες με περισσότερα σημεία πρόσδεσης (όπως η RNApol II), παράγοντες που έχουν ευρύτερα σημεία πρόσδεσης ή τροποποιήσεις της χρωματίνης χρειάζονται 60 εκατομμύρια ''διαβάσματα'' για ChIP-seq σε θηλαστικά (Chen Y et al. 2012).

Στα πειράματα ChIP-seq χρησιμοποιείται συχνά σαν αρνητικό δείγμα μάρτυρα μια ποσότητα χρωματίνης στην οποία δεν έχει πραγματοποιηθεί η ανοσοκατακρήμνιση με το αντίσωμα και το δείγμα αυτό καλείται "input". Πραγματοποιείται αλληλούχηση και του συγκεκριμένου δείγματος οπότε κατά την εύρεση των σημείων πρόσδεσης της πρωτεΐνης πιθανά ψευδώς θετικά σημεία πρόσδεσης που εντοπίζονται και στο δείγμα input αφαιρούνται. Συνήθως επιθυμείται το δείγμα μάρτυρα να περιέχει περισσότερες αλληλουχίες από ότι το πείραμα ChIP-seq (Stephen L. et al. 2012). Αυτή η προϋπόθεση εξασφαλίζει επαρκή κάλυψη ενός σημαντικού τμήματος του γονιδιώματος και τη δημιουργία ενός πλήρους χάρτη αναφοράς – αναζήτησης των διαφόρων θέσεων πρόσδεσης ενός μεταγραφικού παράγοντα.

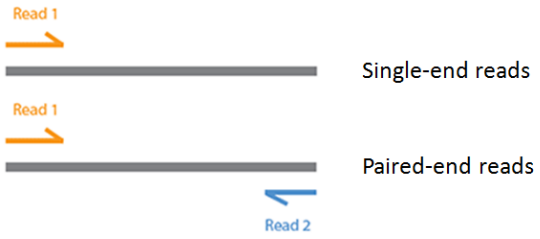
ΧΑΡΤΟΓΡΑΦΗΣΗ ΤΩΝ ''ΔΙΑΒΑΣΜΑΤΩΝ'' ΚΑΙ ΜΕΤΡΗΣΗ ΠΟΙΟΤΗΤΑΣ

Πριν από τη χαρτογράφηση των ''διαβασμάτων'' στο γονιδίωμα αναφοράς , τα ''διαβάσματα'' πρέπει να φιλτραριστούν εφαρμόζοντας κάποια ποιοτικά όρια τα οποία είναι ενδεικτικά της ύπαρξης σφαλμάτων στην αλληλουχία. Τα ''διαβάσματα'' των οποίων οι τιμές ποιότητες δεν ξεπερνούν το κατώφλι τιμών που έχει τεθεί σαν κατώτατο όριο απορρίπτονται. Στη συνέχεια τα φιλτραρισμένα ''διαβάσματα'' χαρτογραφούνται με κάποιο ειδικό αλγόριθμο όπως το Bowtie, BWA , SOAP ή MAQ. Είναι σημαντικό να ληφθεί υπόψη το ποσοστό των μοναδικά χαρτογραφημένων ''διαβασμάτων'' από τον αλγόριθμο. Το ποσοστό διαφέρει μεταξύ των οργανισμών. Από πειράματα ChIP-seq στον άνθρωπο, στο ποντίκι ή στο φυτό Arabidopsis έχει δειχθεί ότι σε ένα αξιόπιστο πείραμα το ποσοστό των μοναδικά χαρτογραφημένων περιοχών κυμαίνεται στο 70% ενώ όταν το ποσοστό πέφτει κάτω από 50% αμφισβητείται η αξιοπιστία του πειράματος. Ένα χαμηλό ποσοστό μοναδικά χαρτογραφημένων ''διαβασμάτων'' συχνά οφείλεται σε υπερβολικό πολλαπλασιασμό στο στάδιο της αλυσιδωτής αντίδρασης πολυμεράσης (PCR), σε ανεπαρκές

μήκος ‘‘διαβασμάτων’’ ή σε προβλήματα στη πλατφόρμα αλληλούχησης. Μόνο σε περιπτώσεις όπου οι πρωτεΐνες προσδένονται με μεγάλη συχνότητα σε επαναλαμβανόμενο DNA τα ποσοστά αυτά είναι αποδεκτά. Οι αλγόριθμοι χαρτογράφησης είναι σχεδιασμένοι να επιτρέπουν έναν αριθμό σφαλμάτων στα ‘‘διαβάσματα’’.

ΑΝΙΧΝΕΥΣΗ ΘΕΣΕΩΝ ΠΡΟΣΔΕΣΗΣ-ΚΟΡΥΦΩΝ (PEAK CALLING)

Μια βασική ανάλυση για τα δεδομένα των ChIP-seq είναι η ανίχνευση των περιοχών όπου η πρωτεΐνη είναι προσδεμένη. Τέτοιες περιοχές είναι εκείνες στις οποίες εντοπίζεται ένας στατιστικά σημαντικός αριθμός χαρτογραφημένων ‘‘διαβασμάτων’’ (peaks). Είναι απαραίτητο να επιλεχθεί ο κατάλληλος αλγόριθμος και η μέθοδος κανονικοποίησης για να εξασφαλισθεί τόσο η ευαισθησία όσο και η ειδικότητα ανάλογα με την πρωτεΐνη που έχει χρησιμοποιηθεί στο πείραμα. Συνίσταται να χρησιμοποιούνται χαρτογραφημένα ‘‘διαβάσματα’’ από ένα δείγμα μάρτυρα (input) για να αξιολογηθούν τα ψευδώς θετικά σημεία πρόσδεσης της πρωτεΐνης όπως αναφέρθηκε πιο πάνω. Υπάρχουν 2 διαφορετικοί τρόποι για να πραγματοποιηθεί η αλληλούχηση των DNA τμημάτων (Εικόνα 19). Στον πρώτο τρόπο αλληλουχείται το DNA τμήμα από το ένα άκρο (single end read) και στο δεύτερο τρόπο γίνεται διπλή αλληλούχηση ξεκινώντας από το κάθε ένα άκρο (paired end read).



Εικόνα 19: Διαφορετικοί τρόποι αλληλούχησης των DNA “διαβασμάτων”.

Παρόλο που κάποιοι αλγόριθμοι υποστηρίζουν και μονά ‘‘διαβάσματα’’ (single) και ‘‘διαβάσματα’’ από τα δύο άκρα (paired end) κάποιοι άλλοι αλγόριθμοι υπολογίζουν το συντελεστή συσχέτισης Pearson (Pearson Correlation Coefficient-PCC) των χαρτογραφημένων ‘‘διαβασμάτων’’ σε μια γενωμική περιοχή (Bardet AF et al 2011). Ο MACS είναι από τους πιο σύγχρονους και διαδεδομένους αλγορίθμους που χρησιμοποιείται για αυτή την ανάλυση.

ΧΑΡΑΚΤΗΡΙΣΜΟΣ ΤΩΝ ΚΟΡΥΦΩΝ (PEAK ANNOTATION)

Ο σκοπός του χαρακτηρισμού είναι να συσχετίσει τις θέσεις πρόσδεσης της πρωτεΐνης από το πείραμα ChIP-seq με λειτουργικές γενωμικές περιοχές όπως υποκινητές γονιδίων, σημεία έναρξης της μεταγραφής και μεσογονιδιακές περιοχές. Αρχικά τα αρχεία με τις θέσεις πρόσδεσης πρέπει να μεταφορτωθούν στην κατάλληλη μορφή (BED ή GFF αρχεία) σε ένα πρόγραμμα γενωμικής περιήγησης (Genome Browser), όπως αυτό του UCSC. Στον περιηγητή αυτό ο ερευνητής μπορεί να εντοπίσει γενωμικές περιοχές στις οποίες υπάρχει προσδεμένη η πρωτεΐνη που μελετάται. Εάν υπάρχουν δεδομένα από ChIP πειράματα για τις θέσεις πρόσδεσης άλλων πρωτεϊνών ή την ύπαρξη ιστονικών τροποποιήσεων μπορούν να συγκριθούν στον περιηγητή αυτό. Μπορεί επίσης να πραγματοποιηθεί μια συστηματική ανάλυση χρησιμοποιώντας υπολογιστικά πακέτα όπως τα BEDTOOLS για να υπολογιστεί η απόσταση των κορυφών από τα πλησιέστερα λειτουργικά σημεία του γονιδιώματος (όπως TSS) ή να αναγνωρισθούν τα γονίδια που βρίσκονται κοντά σε αυτές τις θέσεις πρόσδεσης. Τα αποτελέσματα από αυτές τις αναλύσεις τοποθεσίας, τα οποία λαμβάνονται από προγράμματα όπως το CEAS ή το ChIP peakAnno του Bioconductor, μπορεί να συσχετισθούν με δεδομένα γονιδιακής έκφρασης ή να ακολουθήσει ανάλυση γονιδιακής οντολογίας. Η ανάλυση γονιδιακής οντολογίας μπορεί να πραγματοποιηθεί με προγράμματα όπως το DAVID,το GREAT ή το GSEA.

ΑΝΑΛΥΣΗ ΜΟΤΙΒΩΝ (MOTIF ANALYSIS)

Η ανάλυση μοτίβων είναι χρήσιμη όχι μόνο για να αναγνωρισθούν τα συνηθισμένα DNA μοτίβα πρόσδεσης των μεταγραφικών παραγόντων αλλά και για την επιβεβαίωση της επιτυχίας του πειράματος. Ακόμη και αν το μοτίβο της πρωτεΐνης δεν είναι γνωστό, η αναγνώριση ενός κεντρικά τοποθετημένου μοτίβου σε ένα μεγάλο ποσοστό των περιοχών πρόσδεσης της υπό μελέτη πρωτεΐνης που εντοπίστηκε, είναι χαρακτηριστική της επιτυχίας του πειράματος. Η ανάλυση μοτίβων μπορεί επίσης να αναγνωρίσει μοτίβα DNA πρόσδεσης άλλων πρωτεϊνών στο DNA, οι οποίες αποτελούν σύμπλοκο με την πρωτεΐνη του πειράματός μας και να βοηθήσει στην κατανόηση των μηχανισμών ρύθμισης της μεταγραφής. Η ανάλυση μοτίβων είναι χρήσιμη και για τη μελέτη των ιστονικών τροποποιήσεων διότι μπορεί να ανακαλύψει νέες περιοχές που σχετίζονται με αυτές τις τροποποιήσεις.
Για την ανάλυση μοτίβων χρησιμοποιούνται αρχεία με τις περιοχές πρόσδεσης της προς μελέτη πρωτεΐνης, όπως εντοπίστηκαν από τους αλγόριθμους ανίχνευσης κορυφών. Είναι προτιμότερο να χρησιμοποιηθούν δύο ή περισσότεροι αλγόριθμοι για την ανακάλυψη μοτίβων ανάλογα με τα πλεονεκτήματα και τα μειονεκτήματα τους. Μερικοί αλγόριθμοι δημιουργούν βήματα ανάλυσης που περιλαμβάνουν εύρεση μοτίβων βασισμένα σε γράμματα και μπορούν να αναγνωρίσουν μοτίβα που βρίσκονται σε μικρό ποσοστό των σημείων πρόσδεσης της πρωτεΐνης. Στη συνέχεια τα μοτίβα που εντοπίστηκαν συγκρίνονται με ήδη υπάρχοντα μοτίβα για να αναγνωριστεί η παρουσία μεταγραφικών παραγόντων που εντοπίζονται στις ίδιες θέσεις με την πρωτεΐνη που μελετάται.

2.12.3 ΑΛΓΟΡΙΘΜΟΣ BOWTIE

Ο αλγόριθμος Bowtie (Langmead B et al. 2009) είναι ένας ταχύτατος και αποδοτικός σε σχέση με την υπολογιστική μνήμη αλγόριθμος χαρτογράφησης ο οποίος στοιχίζει τα ‘‘διαβάσματα’’ από τα πειράματα ChIP-seq σε γονιδιώματα αναφοράς. Ο αλγόριθμος κατηγοριοποιεί το γονιδίωμα με τη μέθοδο Burrows-Wheeler για να διατηρήσει χαμηλές τις απαιτήσεις σε υπολογιστική μνήμη. Για να επιταχυνθεί η ταχύτητα ανάλυσης μπορούν να χρησιμοποιηθούν πολλοί υπολογιστικοί πυρήνες παράλληλα. Τα αποτελέσματα του αλγορίθμου εξάγονται σε μορφή SAM κάτι που του επιτρέπει να είναι διαλειτουργικός με άλλα εργαλεία που χρησιμοποιούν τα αρχεία SAM, όπως τα SAMtools. Ο αλγόριθμος λειτουργεί σε λειτουργικό σύστημα Windows, MAC OS X και Linux.

Οι βάσεις δεδομένων χρησιμοποιούν ευρετήρια τα οποία είναι δομημένα με τέτοιο τρόπο ώστε να βελτιώνεται η ταχύτητα με την οποία πραγματοποιούνται διάφορες λειτουργίες, όπως η αναζήτηση συγκεκριμένων ομάδων δεδομένων. Με τον τρόπο αυτό αποφεύγεται η αναζήτηση σε κάθε μια ξεχωριστή γραμμή του πίνακα των δεδομένων διαδοχικά, διαδικασία η οποία απαιτεί περισσότερο χρόνο. Τα ευρετήρια που χρησιμοποιούνται περιέχουν κατηγοριοποιημένα τα δεδομένα χρησιμοποιώντας διαφορετικές μεθόδους κατηγοριοποίησης.

Ο Bowtie αλγόριθμος κατηγοριοποιεί τα δεδομένα χρησιμοποιώντας ένα σύστημα που βασίζεται στο μετασχηματισμό Burrows-Wheeler (Burrows M et al. 1994) και την κατηγοριοποίηση FM (Feragina et al. 2000). Η συνηθισμένη μέθοδος για αναζήτηση σε ένα πίνακα κατηγοριοποίησης FM είναι η χρήση του αλγορίθμου ακριβής αντιστοίχισης (EXACTMATCH) Feragina and Manzini (Feragina et al. 2000).

Ο Bowtie δεν εφαρμόζει απλώς τον αλγόριθμο ακριβούς αντιστοίχισης διότι δεν θα επιτρεπόντουσαν τα λάθη στην αλληλούχηση και δεν θα μπορούσε να εντοπιστεί η γενετική ποικιλότητα. Ο Bowtie περιλαμβάνει

1) έναν αλγόριθμο υπαναχώρησης (backtracking) που επιτρέπει τα σφάλματα και ευνοεί τα χαρτογραφημένα ‘‘διαβάσματα’’ υψηλής ποιότητας.
2) μια στρατηγική δημιουργίας διπλών ευρετηρίων για να αποφεύγεται η υπερβολική υπαναχώρηση.

ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ BURROWS-WHEELER (BURROWS-WHEELER INDEXING-BWT)

Ο μετασχηματισμός BWT είναι μια αναστρέψιμη μετάθεση των χαρακτήρων σε ένα κείμενο. Αρχικά χρησιμοποιήθηκε για τη συμπίεση δεδομένων αλλά ο τρόπος κατηγοριοποίησης σε πίνακες BWT επιτρέπει να γίνεται γρήγορα αναζήτηση σε μεγάλα αρχεία διατηρώντας χαμηλές τις απαιτήσεις σε υπολογιστική μνήμη. Για το λόγο αυτό χρησιμοποιείται σε βιοπληροφορικές εφαρμογές, όπως η χαρτογράφηση ολόκληρου του γονιδιώματος, ο σχεδιασμός ιχνηθετών για μικροσυστοιχίες κ.α.

Ο μετασχηματισμός BWT ενός κειμένου T, δομείται ως εξής. Ο χαρακτήρας $ είναι συνημμένος στο T , όπου το $ δεν βρίσκεται στο T και είναι λεξικογραφικά λιγότερος από όλους τους χαρακτήρες στο T. Ο πίνακας Burrows-Wheeler του T περιέχει σειρές που περιλαμβάνουν όλες τις κυκλικές περιστροφές του T$. Οι σειρές είναι συνήθως ταξινομημένες λεξικογραφικά. BWT(T) είναι η αλληλουχία των χαρακτήρων της πιο δεξιάς στήλης του πίνακα Burrows-Wheeler. Αυτός ο πίνακας έχει μια ιδιότητα που ονομάζεται ‘’ τελευταία πρώτη χαρτογράφηση’’ (last first mapping-LF mapping). Αυτή την ιδιότητα χρησιμοποιούν οι αλγόριθμοι που αξιοποιούν την κατηγοριοποίηση BWT για να αναζητούν στο κείμενο συγκεκριμένα δεδομένα.

Η χαρτογράφηση LF χρησιμοποιείται επίσης και στην ακριβή αντιστοίχιση. Επειδή ο πίνακας είναι ταξινομημένος λεξικογραφικά, οι σειρές ξεκινούν με μια αλληλουχία που εμφανίζεται διαδοχικά. Σε μια σειρά βημάτων ο αλγόριθμος ακριβούς αντιστοίχισης υπολογίζει το εύρος των σειρών του πίνακα που ξεκινούν με μεγαλύτερες αλληλουχίες από αυτές που έχουν ζητηθεί να εντοπιστούν. Σε κάθε βήμα το μέγεθος του εύρους μικραίνει ή παραμένει σταθερό. Όταν ο αλγόριθμος ολοκληρώνεται οι γραμμές που αρχίζουν με S_0 (ολόκληρη η αναζήτηση) αντιστοιχούν στην ακριβή αλληλουχία που αναζητείται στο κείμενο. Εάν το εύρος είναι κενό το κείμενο δεν περιέχει την αλληλουχία της αναζήτησης.

ΑΝΑΖΗΤΗΣΗ ΓΙΑ ΜΗ ΑΚΡΙΒΕΙΣ ΑΝΤΙΣΤΟΙΧΙΣΕΙΣ

Ο αλγόριθμος EXACTMATCH δεν είναι αποτελεσματικός για τη στοίχιση ‘‘διαβασμάτων’’ διότι μπορεί να περιέχουν βάσεις που δεν αντιστοιχούν στο γονιδίωμα και οφείλονται σε σφάλματα κατά την αλληλούχισή τους, σε πραγματικές διαφορές μεταξύ του γονιδιώματος και του προς μελέτη οργανισμού ή και τα δύο. Ο Bowtie αλγόριθμος πραγματοποιεί μια αναζήτηση υπαναχώρησης για να βρίσκει γρήγορα τις αντιστοιχήσεις που υπόκεινται σε κάποιους προεπιλεγμένους κανόνες. Κάθε βάση στο ‘‘διάβασμα’’ έχει μια αριθμητική τιμή ποιότητας, με τις χαμηλότερες τιμές να υποδεικνύουν τη μεγαλύτερη πιθανότητα μια αλληλουχία να περιέχει κάποιο σφάλμα. Οι κανόνες που εφαρμόζονται από τον αλγόριθμο επιτρέπουν έναν περιορισμένο αριθμό αναντιστοιχιών και προτιμούν αντιστοιχήσεις όπου το άθροισμα των ποιοτικών τιμών σε όλες τις θέσεις αναντιστοιχίας είναι χαμηλό.

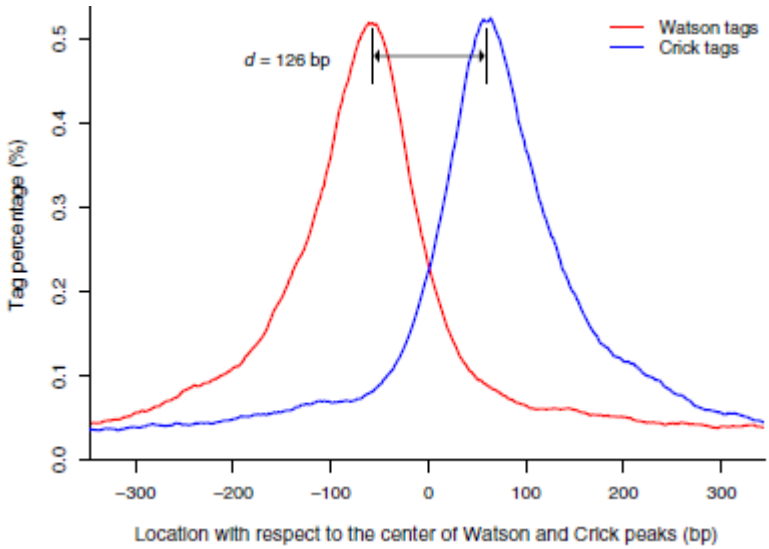
Η αναζήτηση στη συνέχεια υπολογίζει το εύρος του πίνακα για να εντοπίσει μακρύτερες καταλήξεις από το ερώτημα που έχει τεθεί. Εάν το εύρος μείνει κενό, ο αλγόριθμος επιλέγει τη θέση ενός ήδη αντιστοιχισμένου ‘’διαβάσματος’’ και αντικαθιστά μια βάση εισάγοντας μια αναντιστοιχία. Ύστερα συνεχίζει την αναζήτηση από εκεί που έμεινε. Ο αλγόριθμος επιλέγει μόνο αντικαταστάσεις που τηρούν τους κανόνες της αντιστοίχισης και βρίσκονται τουλάχιστον μια φορά στο κείμενο. Εάν υπάρχουν πολλαπλές πιθανές επιλογές για την αντικατάσταση μιας βάσης επιλέγεται εκείνη με την ελάχιστη ποιοτική τιμή. Η υπαναχώρηση συμβαίνει όταν ένα σύνολο δεδομένων αυξάνεται ύστερα από μια αντικατάσταση και συρρικνώνεται όταν ο αλγόριθμος απορρίπτει όλες τις υποψήφιες αντιστοιχήσεις για μια αντικατάσταση από αυτές που υπάρχουν στο σύνολο.

2.12.4 ΑΛΓΟΡΙΘΜΟΣ MACS (MODEL-BASED ANALYSIS FOR ChIP-Seq)

Ο αλγόριθμος MACS (Zhang Y et al. 2008) χρησιμοποιείται για να αναγνωρισθούν οι DNA περιοχές όπου προσδένονται οι μεταγραφικοί παράγοντες και να εντοπιστούν DNA περιοχές εμπλουτισμένες με ιστονικές τροποποιήσεις. Ο αλγόριθμος αυτός βασίζεται στην κατανομή Poisson για να βελτιώσει την ευκρίνεια των περιοχών πρόσδεσης. Ο τρόπος που πραγματοποιείται αυτή η ανάλυση περιγράφεται παρακάτω.

ΕΝΤΟΠΙΣΜΟΣ ΘΕΣΕΩΝ ΠΡΟΣΔΕΣΗΣ ΤΗΣ ΠΡΩΤΕΙΝΗΣ (PEAKS)

Για μελέτες με πειράματα μάρτυρες (controls), ο αλγόριθμος MACS φέρνει στο ίδιο επίπεδο το συνολικό αριθμό των αλληλουχιών από το πείραμα μάρτυρα με το συνολικό αριθμό των αλληλουχιών από το ChIP πείραμα. Μερικές φορές οι ίδιες αλληλουχίες μπορεί να αλληλουχηθούν επανειλημμένα, περισσότερες φορές από ότι θα αναμενόταν από μια τυχαία κατανομή των αλληλουχιών στο γονιδίωμα. Τέτοιες αλληλουχίες θα μπορούσαν να προκύψουν από σφάλματα κατά τον πολλαπλασιασμό του DNA του ChIP πειράματος με PCR και της προετοιμασίας των βιβλιοθηκών DNA προσθέτοντας έτσι θόρυβο που ίσως οδηγήσει στον εντοπισμό ψευδώς θετικών θέσεων πρόσδεσης. Ως εκ τούτου, ο MACS αφαιρεί τις διπλές αλληλουχίες . Με την τρέχουσα κάλυψη του γονιδιώματος που χρησιμοποιείται στα περισσότερα ChIP-Seq πειράματα, η κατανομή των αλληλουχιών κατά μήκος του γονιδιώματος μπορεί να μοντελοποιηθεί με την κατανομή Poisson. Το πλεονέκτημα αυτού του μοντέλου είναι ότι μια παράμετρος, λ_{BG}, μπορεί να προσδιορίσει τόσο τη μέση τιμή όσο και τη διακύμανση της κατανομής. Αφού ο MACS μετατοπίζει κάθε αλληλουχία κατά d/2 (όπου d είναι η απόσταση μεταξύ των κέντρων των κορυφών της Watson και Crick αλυσίδας – Εικόνα 20), και σαρώνει το γονιδιώμα για να εντοπίσει υποψήφιες κορυφές με σημαντικό εμπλουτισμό σε αλληλουχίες (κατανομή Poisson p-value με βάση την παράμετρο λ_{BG}, προεπιλογή 10^{-5}). Οι επικαλυπτόμενες εμπλουτισμένες κορυφές συγχωνεύονται και κάθε θέση αλληλουχίας εκτείνεται κατά d βάσεις από το κέντρο της. Η θέση με την υψηλότερη συσσώρευση τμημάτων αναφέρεται ως σύνοδος κορυφής και προβλέπεται ως η ακριβής θέση πρόσδεσης.



Εικόνα 20: Κατανομή των κορυφών στοιχισμένων ως προς το κέντρο τους στις αλυσίδες Watson και Crick.

Πολλές πιθανές πηγές για σφάλματα, που αναφέρθηκαν και προηγουμένως, περιλαμβάνουν σφάλματα που προέρχονται από τη δομή της χρωματίνης, σφάλματα κατά το πολλαπλασιασμό του DNA με PCR και την αλληλούχηση. Ως εκ τούτου ο αλγόριθμος MACS εφαρμόζει μια δυναμική παράμετρο, λ_{local}, η οποία ορίζεται για κάθε υποψήφια κορυφή ως:

$$\boldsymbol{\lambda_{local} = max(\lambda_{BG}, [\lambda_{1k},] \lambda_{5k}, \lambda_{10k})}$$

όπου $\boldsymbol{\lambda_{1k}}$, $\boldsymbol{\lambda_{5k}}$, **και** $\boldsymbol{\lambda_{10k}}$, είναι λ εκτιμώμενες σε απόσταση 1kb, 5kb και 10kb από το κέντρο της κορυφής στο δείγμα μάρτυρα, ή στο δείγμα ChIP-Seq όταν δεν χρησιμοποιούνται δείγματα μάρτυρες (σε αυτή την περίπτωση η μεταβλητή $\boldsymbol{\lambda_{10k}}$ δε χρησιμοποιείται). Η παράμετρος λ_{local} λαμβάνει υπόψη τα σφάλματα στις δομές της χρωματίνης και είναι "εύρωστη"(robust) όταν χαμηλός αριθμός αλληλουχιών στοιχίζονται σε μικρές τοπικές περιοχές. Ο MACS χρησιμοποιεί τη μεταβλητή λ_{local} για να υπολογίσει την p-value τιμή για κάθε υποψήφια κορυφή και αφαιρεί πιθανώς ψευδώς θετικές κορυφές εξαιτίας σφαλμάτων που προέρχονται από τη δομή της χρωματίνης (πρόκειται για κορυφές κάτω από την τιμή της $\boldsymbol{\lambda_{BG}}$, αλλά όχι κάτω από την $\boldsymbol{\lambda_{local)}}$. Εντοπίζονται υποψήφιες κορυφές με τιμές κάτω από ένα όριο για την p-value τιμή που έχει οριστεί από το χρήστη (η προεπιλογή είναι 10^{-5}), και ο λόγος μεταξύ του αριθμού των ChIP-Seq αλληλουχιών και της μεταβλητής λ_{local}, αναφέρεται ως λόγος εμπλουτισμού.

Για ένα πείραμα ChIP-Seq όπου χρησιμοποιούνται δείγματα μάρτυρες, ο MACS εμπειρικά εκτιμά τα ψευδή ποσοστά ανακάλυψης (False Discovery Rate-FDR) για κάθε εντοπισμένη κορυφή. Για κάθε p-value, ο MACS χρησιμοποιεί τις ίδιες παραμέτρους για να εντοπίσει κορυφές του Chip πειράματος σε σχέση με το δείγμα μάρτυρα και το αντίστροφο (η διαδικασία αυτή ορίζεται ως ανταλλαγή δείγματος - sample swap). Η εμπειρική FDR ορίζεται ως ο λόγος του αριθμού των κορυφών του μάρτυρα προς τον αριθμό των κορυφών του Chip πειράματος. Ο MACS μπορεί επίσης να εφαρμοστεί για τον έλεγχο της διαφορικής πρόσδεσης μεταξύ δυο συνθηκών χρησιμοποιώντας το ένα από τα δύο δείγματα σαν μάρτυρα. Δεδομένου ότι οι κορυφές από κάθε δείγμα μπορεί να είναι εξίσου σημαντικής βιολογικής σημασίας σε αυτή την περίπτωση, δεν μπορεί να χρησιμοποιηθεί η ανταλλαγή δείγματος για να υπολογιστεί η FDR, και η ποιότητα των δεδομένων κάθε δείγματος πρέπει να αξιολογηθεί σε σχέση με τον πραγματικό δείγμα μάρτυρα.

2.12.5 ΑΛΓΟΡΙΘΜΟΣ CEAS (CIS-REGULATORY ELEMENT ANNOTATION SYSTEM)

Ο αλγόριθμος CEAS (Shin H et al. 2009) χρησιμοποιείται για τον χαρακτηρισμό του σήματος από πειράματα ChIP-seq και τη συσχέτιση των περιοχών πρόσδεσης της προς μελέτη πρωτεΐνης με λειτουργικές περιοχές του γονιδιώματος, όπως υποκινητές και γονίδια.

Ο CEAS περιλαμβάνει 3 λειτουργίες:

1. Χαρακτηρισμός των περιοχών του ChIP (ChIP region annotation)
2. Γονιδιακά επικεντρωμένο χαρακτηρισμό (gene center annotation)
3. Προφίλ του μέσου όρου του σήματος μέσα και κοντά σε λειτουργικές γονιδιωματικές περιοχές (average signal profiling within and near important genomic features)

Ο CEAS απαιτεί 2 είδη αρχείων:

1. Ένα πίνακα με τις συντεταγμένες των γονιδίων στο γονιδίωμα όπως το Refseq αρχείο από το UCSC.
2. Ένα BED αρχείο με τις κορυφές του ChIP πειράματος.

Το BED αρχείο πρέπει να αποτελείται από 3 στήλες (το χρωμόσωμα, τη θέσης έναρξης και τη θέση τερματισμού του κάθε σημείου πρόσδεσης της πρωτεΐνης στο DNA).
Το αποτέλεσμα του αλγορίθμου είναι ένα αρχείο PDF με τη γραφική απεικόνιση των αποτελεσμάτων από την επισήμανση των περιοχών του ChIP και τα προφίλ του μέσου όρου του σήματος και ένα αρχείο XLS με τα αποτελέσματα του γονιδιακά επικεντρωμένου χαρακτηρισμού.

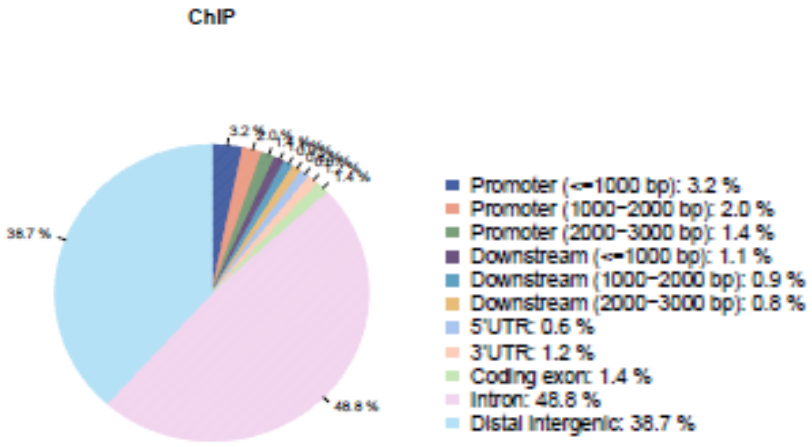
ΧΑΡΑΚΤΗΡΙΣΜΟΣ ΤΩΝ ΠΕΡΙΟΧΩΝ ΤΟΥ CHIP (CHIP REGION ANNOTATION)

Ο CEAS υπολογίζει το ποσοστό των λειτουργικών γονιδιωματικών περιοχών στις οποίες εντοπίζονται αρκετά “διαβάσματα”. Για να το πραγματοποιήσει αυτό, υπολογίζει το ποσοστό των περιοχών του ChIP που βρίσκονται σε μια από τις ακόλουθες 4 κατηγορίες:

1. Υποκινητές
2. Αμφίδρομοι υποκινητές
3. Περιοχές καθοδικά του γονιδίου
4. Σώμα του γονιδίου (3’ UTRs , 5’UTRs, κωδικά εξώνια και εσώνια)

Οι υποκινητές αντιστοιχούν στην περιοχή ανοδικά του σημείου έναρξης της μεταγραφής (TSS) του γονιδίου. Στο χρήστη δίνεται η δυνατότητα καθορισμού 3 περιοχών διαφορετικού μεγέθους ως υποκινητή (η προεπιλογή είναι 1kb, 3kb, 10kb). Για παράδειγμα εάν θέσουμε το μέγεθος του υποκινητή να είναι 1kb, 3kb και 10kb ανοδικά του σημείου έναρξης της μεταγραφής ο αλγόριθμος θα υπολογίσει τα συσσωρευτικά ποσοστά των περιοχών του ChIP που βρίσκονται ≤1 kb, ≤3 kb και ≤10 kb.

Αμφίδρομοι υποκινητές είναι οι περιοχές που βρίσκονται μεταξύ αποκλινόντων μεταγραφόμενων γονιδίων των οποίων τα σημεία έναρξης της μεταγραφής βρίσκονται αρκετά κοντά μεταξύ τους. Ομοίως και εδώ ο χρήστης έχει την δυνατότητα να ορίσει την απόσταση των αμφίδρομων υποκινητών (η προεπιλογή για 2,5kb και 5kb). Το σώμα των γονιδίων διαχωρίζεται σε 3’ και 5’ UTRs, σε κωδικά εξώνια και εσώνια. Αφού υπολογιστούν τα ποσοστά των περιοχών του ChIP που βρίσκονται στις παραπάνω κατηγορίες συγκρίνονται με τα ποσοστά του γονιδιωματικού υποβάθρου για τις ίδιες κατηγορίες και υπολογίζονται οι τιμές p-value χρησιμοποιώντας ένα μονόπλευρο διωνυμικό τεστ. Ο αλγόριθμος CEAS σχηματίζει ένα γράφημα πίτας όπου παρουσιάζεται η κατανομή των περιοχών του ChIP σε σχέση με λειτουργικές γονιδιωματικές περιοχές (Εικόνα 21). Εάν κάποια περιοχή δεν τοποθετηθεί σε κάποια από τις 4 κατηγορίες χαρακτηρίζεται ως απομακρυσμένη από τα γονίδια περιοχή (distal intergenetic).



Εικόνα 21: Διαγραμματική απεικόνιση της κατανομής των περιοχών πρόσδεσης του μεταγραφικού παράγοντα σε σχέση με λειτουργικές γονιδιωματικές περιοχές.

ΓΟΝΙΔΙΑΚΑ ΕΠΙΚΕΝΤΡΩΜΕΝΟΣ ΧΑΡΑΚΤΗΡΙΣΜΟΣ (GENE CENTERED ANNOTATION)

Για την αναγνώριση των γονιδίων που συνδέονται με τις περιοχές του ChIP με αξιοπιστία είναι σημαντικό να εντοπιστούν οι άμεσοι γονιδιακοί στόχοι όπου προσδένονται οι παράγοντες τους οποίου μελετούμε. Ο αλγόριθμος CEAS παρέχει τις αποστάσεις από το κέντρο των πλησιέστερων περιοχών του ChIP όπου βρίσκονται ανοδικά και καθοδικά του κάθε γονιδιακού σημείου έναρξης της μεταγραφής. Αυτή η πληροφορία για παράγοντες με ακριβές μοτίβο πρόσδεσης είναι αρκετή για να καθοριστούν τα γονίδια στόχοι. Για πρωτεΐνες των οποίων τα σημεία πρόσδεσης καλύπτουν όλο ή μέρος του γονιδιακού σώματος είναι χρήσιμο να γνωρίζουμε το ποσοστό του γονιδίου που καλύπτεται από την περιοχή του ChIP. Τέλος ο αλγόριθμος χωρίζει το κάθε γονίδιο σε ίσα τμήματα, εξάγει τα εξώνια και υπολογίζει τα ποσοστά κάλυψης της κάθε περιοχής.

ΠΡΟΦΙΛ ΜΕΣΟΥ ΟΡΟΥ ΣΗΜΑΤΟΣ ΕΝΤΟΣ/ΚΟΝΤΑ ΣΗΜΑΝΤΙΚΩΝ ΛΕΙΤΟΥΡΓΙΚΩΝ ΓΟΝΙΔΙΩΜΑΤΙΚΩΝ ΠΕΡΙΟΧΩΝ

Εφόσον η περιοχή ChIP και ο γονιδιακά επικεντρωμένος χαρακτηρισμός λειτουργούν σε διαφορετικές περιοχές ChIP οι οποίες αναγνωρίζονται από έναν αλγόριθμο εντοπισμού κορυφών (peak-calling algorithm), ορισμένα ακριβή μοτίβα πρόσδεσης μπορεί να μην εντοπιστούν, ανάλογα με το ανώτατο όριο που χρησιμοποιείται από τον αλγόριθμο εντοπισμού κορυφών. Συνεπώς, η εφαρμογή CEAS εμφανίζει το συνεχόμενα ενισχυμένο σήμα ChIP μέσα και κοντά σε σημαντικές λειτουργικές γονιδιωματικές περιοχές, προκειμένου οι βιολόγοι να απεικονίσουν το μέσο όρο των προτύπων πρόσδεσης σε αυτές τις περιοχές. Η εφαρμογή CEAS σχεδιάζει το μέσο όρο των σημάτων γύρω από τα TSSs και TTSs εντός ενός εύρους τιμών που έχει οριστεί από τον ίδιο το χρήστη (η προεπιλογή είναι ±3 kb από τα TSSs και TTSs). Επιπρόσθετα, υπολογίζει το μέσο όρο των σημάτων στο "μετά-γονίδιο", " μετασυνεχόμενο-εξώνιο", " μετασυνεχόμενο-εσώνιο", "μετά-εξώνια" και "μετά-εσώνια", όπου το πρόθεμα "μετά" δηλώνει ότι κάθε στοιχείο κανονικοποιείται ώστε να έχει το ίδιο μήκος. Η διαφορά μεταξύ του μετασυνεχόμενου-εξώνιου και των μετά-εξωνίων είναι ότι το πρώτο συνενώνει όλα τα εξώνια ενός γονιδίου, πριν υπολογίσει το μέσο προφίλ του γονιδίου, ενώ το τελευταίο υπολογίζει το μέσο εξώνιο προφίλ όλων των εξωνίων. Επιπλέον, η εφαρμογή CEAS σχεδιάζει ένα μέσο όρο από τα σήματα ChIP υποομάδων γονιδίων από πολλαπλούς χρήστες, επιτρέποντας έτσι τη σύγκριση των σημάτων μεταξύ διαφορετικών ομάδων γονιδίων.

2.13.1 RNA-seq

Η τεχνική RNA αλληλούχησης (RNA-seq) είναι μια τεχνολογία που χρησιμοποιεί τις δυνατότητες της αλληλούχησης επόμενης γενιάς (next generation sequencing) για να αποκαλύψει τη στιγμιαία παρουσία RNA.

Το μεταγράφωμα ενός κυττάρου έχει δυναμική φύση και μεταβάλλεται ανάλογα με τα ερεθίσματα που δέχεται. Οι πρόσφατες εξελίξεις στην τεχνολογία αλληλούχησης νέας γενιάς επιτρέπουν την εξαγωγή περισσότερων πληροφοριών από την αλληλούχηση των δειγμάτων. Έτσι παρέχεται η δυνατότητα εύρεσης του εναλλακτικού ματίσματος των γονιδίων,των μετα-μεταγραφικών αλλαγών, της σύντηξης γονιδίων, των μεταλλαγών/SNP και των αλλαγών της γονιδιακής έκφρασης. Η τεχνολογία του RNA-seq χρησιμοποιείται και σε μελέτες αλλαγής της γονιδιακής έκφρασης κατόπιν ιικής μόλυνσης και σε καρκινικά δείγματα.

Ένας πληθυσμός RNA απομονώνεται από τα προς μελέτη κύτταρα και μετατρέπεται σε βιβλιοθήκη cDNA κομματιών στα οποία προσκολλούνται προσαρτήματα (adapters) στο ένα ή και στα δύο άκρα τους. Στη συνέχεια πραγματοποιείται αλληλούχηση των μορίων με παρόμοιο τρόπο όπως περιγράφεται παραπάνω στην αλληλούχηση των ChIP-seq για να ληφθούν ‘‘διαβάσματα’’ που έχουν αλληλουχηθεί από το ένα άκρο (single-end sequencing) ή και από τα δύο άκρα (paired-end sequencing). Τα ‘‘διαβάσματα’’ έχουν συνήθως μέγεθος 36 – 1000 βάσεων ανάλογα με την τεχνολογία DNA αλληλούχησης που χρησιμοποιείται.

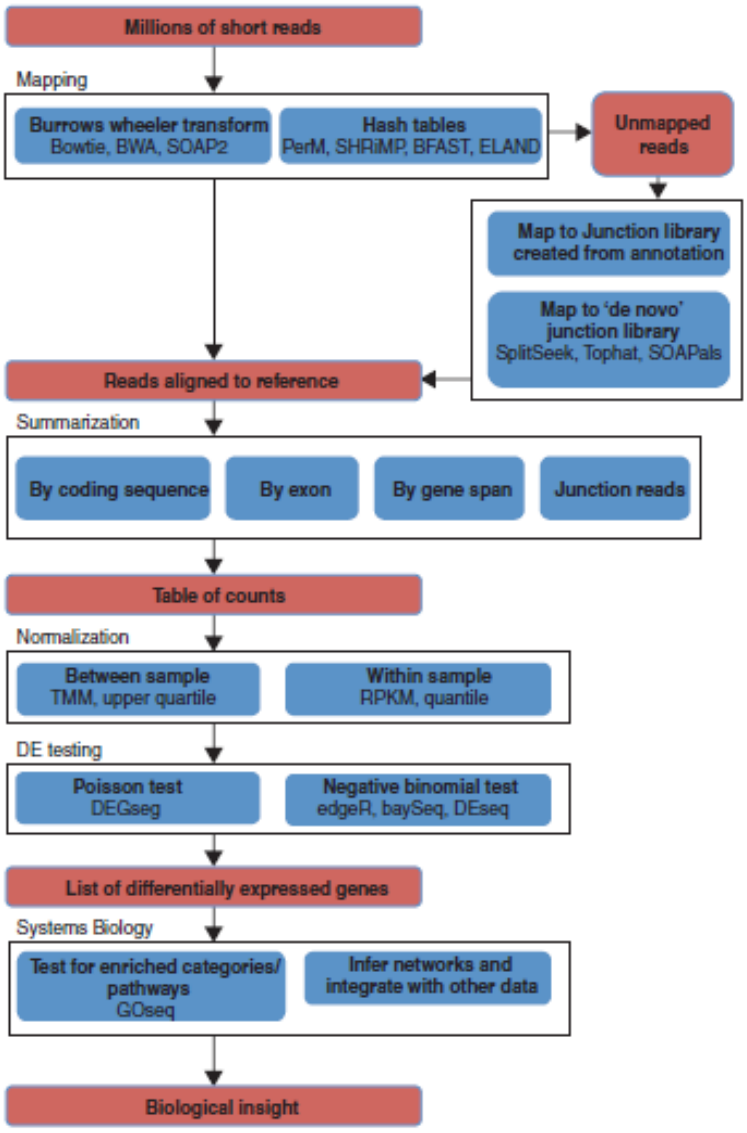
Η τεχνολογία του RNA-seq έχει πολλά προτερήματα σε σχέση με τις ήδη υπάρχουσες τεχνολογίες.

Αρχικά σε αντίθεση με τις τεχνολογίες που βασίζονται στον υβριδισμό, το RNA-seq δεν περιορίζεται στην ανίχνευση μεταγράφων που αντιστοιχούν σε υπάρχουσες γονιδιωματικές αλληλουχίες. Αυτό καθιστά τη συγκεκριμένη τεχνική χρήσιμη για τη μελέτη οργανισμών οι οποίοι δεν είναι μοντέλα και η γονιδιωματική τους αλληλουχία δεν έχει καθοριστεί. Επιπλέον με το RNA-seq μπορούν να καθοριστούν τα όρια του μεταγράφου με ακρίβεια μιας βάσης. Ακόμη ‘‘διαβάσματα’’ με μήκος 30 ζεύγη βάσεων μπορούν να δώσουν πληροφορίες για το πώς 2 εξώνια είναι συνδεδεμένα, ενώ μεγαλύτερα ‘‘διαβάσματα’’ μπορούν να χρησιμοποιηθούν για να καθοριστεί η συνδεσιμότητα πολλαπλών εξωνίων.

Ένα δεύτερο πλεονέκτημα του RNA-seq σε σχέση με τις DNA μικροσυστοιχίες είναι ότι στο RNA-seq υπάρχει πολύ χαμηλό , εάν όχι καθόλου, σήμα υποβάθρου διότι οι DNA αλληλουχίες μπορούν να χαρτογραφηθούν με ακρίβεια σε μοναδικές περιοχές του γονιδιώματος. Ακόμη η τεχνική RNA-seq έχει αποδειχθεί ότι έχει υψηλότερη ακρίβεια στην ποσοτικοποίηση των επιπέδων έκφρασης όπως καθορίστηκε από PCR ποσοτικοποίησης-qPCR (Nagalakshmi U et al. 2008). Τα αποτελέσματα από RNA-seq πειράματα έχει αποδειχθεί ότι έχουν υψηλό ποσοστό επαναληψιμότητας για βιολογικά πειράματα (replicates). Τέλος στο RNA-seq χρησιμοποιείται μικρότερη ποσότητα δείγματος RNA.

2.13.2 ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ ΑΝΑΛΥΣΗ RNA-SEQ

Για να αναλυθούν δεδομένα από πειράματα που έχει χρησιμοποιηθεί η RNA-seq τεχνολογία για τον εντοπισμό διαφορετικά εκφραζόμενων γονιδίων είναι απαραίτητο να ποσοτικοποιηθούν τα πρωτογενή δεδομένα (15-60 εκατομμύρια ‘‘διαβάσματα’’). Αυτό πραγματοποιείται με μια σειρά βημάτων η οποία περιλαμβάνει τη χρήση διαφορετικών προγραμμάτων και αλγορίθμων που συνοπτικά παρουσιάζονται στην Εικόνα 22.



Εικόνα 22: Διαδοχικά βήματα βιοπληροφορικής ανάλυσης RNA-seq δεδομένων.

ΧΑΡΤΟΓΡΑΦΗΣΗ (MAPPING)

Η χαρτογράφηση των ‘‘διαβασμάτων’’ από RNA-seq πειράματα αποτελεί το πρώτο βήμα στην ανάλυση. Η διαδικασία αυτή περιλαμβάνει την εύρεση των περιοχών του γονιδιώματος με τις οποίες αντιστοιχούν πλήρως οι αλληλουχίες από ένα RNA-seq πείραμα. Στην πραγματικότητα όμως το γονιδίωμα αναφοράς δεν αντιπροσωπεύει πάντα την ακριβή βιολογική πηγή του RNA που αλληλουχήθηκε καθώς υπάρχουν

προσθήκες ή ελλείψεις βάσεων, μεταλλάξεις μονού νουκλεοτιδίου (SNPs) αλλά και νέα μετάγραφα προέρχονται από εναλλακτικό μάτισμα.

Τέλος τα ‘‘διαβάσματα’’ αυτά μπορεί να αντιστοιχούν σε πολλαπλές περιοχές του γονιδιώματος και να περιέχουν σφάλματα αλληλούχησης τα οποία πρέπει να ληφθούν υπόψη. Για αυτό ο πραγματικός στόχος της χαρτογράφησης είναι να εντοπιστούν οι περιοχές του γονιδιώματος αναφοράς στις οποίες αντιστοιχούν καλύτερα τα ‘‘διαβάσματα’’ διαμορφώνοντας τον αλγόριθμο αναζήτησης κάθε φορά ανάλογα με το πείραμα και το ερευνητικό ερώτημα.

Σχεδόν όλοι οι αλγόριθμοι χαρτογράφησης χρησιμοποιούν την ίδια στρατηγική όπου πρώτα πραγματοποιείται μία γρήγορη αναζήτηση των αλληλουχιών σε ένα σύνολο τυχαίων τμημάτων του γονιδιώματος αναφοράς. Στη συνέχεια ακολουθεί μια αναλυτική αξιολόγηση των υποψηφίων περιοχών αντιστοίχισης που βασίζεται σε ένα περίπλοκο αλγόριθμο τοπικής στοίχισης (local alligment). Η διαδικασία αυτή μειώνει σημαντικά την υπολογιστική δύναμη και το χρόνο που απαιτείται για τη χαρτογράφηση. Οι αλγόριθμοι πραγματοποιούν τη γρήγορη αναζήτηση είτε χρησιμοποιώντας πίνακες κατακερματισμού (hash table)(Hach F et al. 2010) είτε τον μετασχηματισμό Burrows-Wheeler.

Οι αλγόριθμοι χαρτογράφησης διαφέρουν επίσης και στον τρόπο που διαχειρίζονται τα πολλαπλά σημεία στοίχισης των ‘‘διαβασμάτων’’. Οι περισσότεροι αλγόριθμοι απορρίπτουν τις πολλαπλές θέσεις ταιριάσματος, τις διανέμουν τυχαία ή τις διανέμουν στηριζόμενοι σε ένα υπολογισμό της τοπικής κάλυψης (Cloonan N et al. 2008). Τα ‘‘διαβάσματα’’ που αλληλουχήθηκαν και από τα 2 άκρα (paired-end read) μειώνουν το πρόβλημα των πολλαπλών θέσεων ταιριάσματος . Από τους πιο διαδεδομένους αλγόριθμους χαρτογράφησης είναι ο BOWTIE, το SOAP, το BWA και το GEM. Στις περισσότερες περιπτώσεις χρησιμοποιείται το γονιδίωμα σαν αρχείο αναφοράς για τη χαρτογράφηση των ‘‘διαβασμάτων’’. Σε αυτή την περίπτωση όμως ‘‘διαβάσματα’’ που καλύπτουν περιοχές όπου βρίσκονται τα όρια των εξωνίων δεν μπορούν να χαρτογραφηθούν. Μεγαλύτερα ‘‘διαβάσματα’’ διασχίζουν συχνότερα περιοχές με όρια εξωνίων με αποτέλεσμα να αυξάνονται τα ‘‘διαβάσματα’’ συμβολής (junction reads). Με σκοπό να υπολογίζονται αυτά τα ‘‘διαβάσματα’’ είναι συνηθισμένο να δημιουργούνται βιβλιοθήκες κόμβων εξωνίων στις οποίες έχουν κατασκευασθεί αλληλουχίες αναφοράς χρησιμοποιώντας τα όρια μεταξύ των εξωνίων. Ένας άλλος τρόπος για να χαρτογραφηθούν αυτά τα ‘‘διαβάσματα’’ είναι η de novo συναρμολόγηση του μεταγραφώματος για χρήση ως γονιδίωμα αναφοράς. Όλες οι de novo μέθοδοι μπορούν να αναγνωρίζουν νέα μετάγραφα και είναι η μόνη επιλογή για οργανισμούς για τους οποίους δεν υπάρχει γονιδίωμα αναφοράς.

ΣΥΝΟΨΙΣΗ ΤΩΝ ΧΑΡΤΟΓΡΑΦΗΜΕΝΩΝ ΔΙΑΒΑΣΜΑΤΩΝ (SUMMARIZING MAPPED READS)

Αφού έχουν βρεθεί οι γονιδιωματικές θέσεις των περισσότερων ‘‘διαβασμάτων’’, το επόμενο βήμα είναι να συναθροιστούν τα ‘‘διαβάσματα’’ σε κάποιες μονάδες βιολογικής σημασίας, όπως εξώνια, μετάγραφα ή γονίδια. Η πιο απλή και συνηθέστερη προσέγγιση είναι η μέτρηση του αριθμού των ‘‘διαβασμάτων’’ που αντιστοιχούν σε μετάγραφα ενός γονιδίου. Ένας από τους πιο διαδεδομένους αλγόριθμους που πραγματοποιεί αυτή τη διαδικασία είναι ο HTSeq-count.

ΑΝΑΛΥΣΗ ΔΙΑΦΟΡΙΚΗΣ ΕΚΦΡΑΣΗΣ (DIFFERENTIAL EXPRESSION ANALYSIS)

Ο σκοπός της ανάλυσης διαφορικής έκφρασης είναι η ταυτοποίηση των γονιδίων των οποίων η έκφραση έχει αλλάξει σημαντικά μεταξύ δύο δειγμάτων. Επομένως εφαρμόζονται διάφορα στατιστικά τεστ στο σύνολο των χαρτογραφημένων ‘‘διαβασμάτων’’. Στα περισσότερα μοντέλα για τον υπολογισμό της διαφορετικής έκφρασης χρησιμοποιείται η κατανομή Poisson. Για να ληφθεί υπόψη η βιολογική μεταβλητότητα έχει χρησιμοποιηθεί η αρνητική διωνυμική κατανομή (negative binomial distribution-NB) στην οποία χρειάζεται να υπολογισθεί μια επιπλέον παράμετρος διασποράς. Ο αλγόριθμος DESeq βασίζεται σε ένα μοντέλο αρνητικής διωνυμικής κατανομής για να υπολογίσει τη διαφορετική γονιδιακή έκφραση.

2.13.3 ΑΛΓΟΡΙΘΜΟΣ TOPHAT

Ο αλγόριθμος TopHat (Trapnell C et al. 2009) είναι ένας αλγόριθμος χαρτογράφησης μικρών DNA ‘‘διαβασμάτων’’ και αναγνώρισης των σημείων ματίσματος για δεδομένα από πειράματα RNA-Seq. Ο αλγόριθμος είναι δομημένος με τη γλώσσα προγραμματισμού C++ και μπορεί να χρησιμοποιηθεί σε περιβάλλον Linux ή Mac OS X . Η χαρτογράφηση των DNA ‘‘διαβασμάτων’’ σε ένα γονιδίωμα αναφοράς πετυχαίνει 2 βασικούς στόχους του RNA-seq πειράματος :

1) Αναγνώριση νέων μεταγράφων από τις θέσεις που καλύπτονται κατά τη χαρτογράφηση.
2) Υπολογισμός της αφθονίας των μεταγράφων από το βάθος της κάλυψης στη χαρτογράφηση.

Ο αλγόριθμος αυτός χαρτογραφεί ‘‘διαβάσματα’’ από RNA-seq πειράματα στο ανθρώπινο γονιδίωμα με ρυθμό 2,2εκατομμύρια ‘‘διαβάσματα’’ για κάθε ώρα CPU. Ο αλγόριθμος βρίσκει τους κόμβους (junctions) μεταξύ των εξωνίων χαρτογραφώντας τα ‘‘διαβάσματα’’ στο γονιδίωμα αναφοράς σε δύο στάδια.

Στο πρώτο πραγματοποιείται η χαρτογράφηση των ‘‘διαβασμάτων’’ στο γονιδίωμα αναφοράς χρησιμοποιώντας τον αλγόριθμο Bowtie. Όσα ‘‘διαβάσματα’’ δε χαρτογραφούνται συλλέγονται σε μια δεξαμενή με ‘‘διαβάσματα’’ που ονομάζονται “αρχικά αχαρτογράφητα διαβάσματα’’ (initially unmapped reads-‘IUM’). Ο Bowtie εντοπίζει για κάθε διάβασμα, μία ή περισσότερες περιοχές στοίχισης στο γονιδίωμα αναφοράς οι οποίες περιέχουν μέχρι κάποιο όριο σφαλμάτων (η προεπιλογή είναι 2) στο 5΄-άκρο του διαβάσματος. Το υπόλοιπο τμήμα του διαβάσματος στο 3΄-άκρο μπορεί να περιέχει επιπλέον σφάλματα. Αυτή η μέθοδος εφαρμόζεται λόγω της εμπειρικής παρατήρησης ότι το 5΄-άκρο του διαβάσματος περιέχει λιγότερα σφάλματα αλληλούχησης από το 3΄-άκρο(Hillier LW et al.2008). Ο TopHat επιτρέπει στον Bowtie αλγόριθμο να εντοπίζει περισσότερες από μια θέσεις στοίχισης στο γονιδίωμα αναφοράς (η προεπιλογή είναι μέχρι 10) και να αποκρύπτει τις περιπτώσεις όπου υπάρχουν περισσότερες τέτοιες θέσεις από τον προεπιλεγμένο αριθμό.

Στο δεύτερο στάδιο ο αλγόριθμος TopHat συναρμολογεί τα χαρτογραφημένα ‘‘διαβάσματα’’ χρησιμοποιώντας το μοντέλο συναρμολόγησης του αλγόριθμου MAQ (Li H et al 2008). Ο TopHat εξάγει τις αλληλουχίες για τις ‘’νησίδες ‘’ συνεχόμενων αλληλουχιών από την σποραδική συνένωση και αναφέρεται σε αυτές ως υποθετικά εξώνια. Για τη δημιουργία αυτών των ‘’νησίδων’’ ο Tophat χρησιμοποιεί την υποεντολή του Maq, assemple η οποία δημιουργεί ένα πυκνό αρχείο με συνενώσεις, το οποίο περιέχει τις βάσεις που έχουν αλληλουχηθεί και τις αντίστοιχες βάσεις του γονιδιώματος αναφοράς. Επειδή οι συνενώσεις μπορεί να περιέχουν λανθασμένες βάσεις λόγω σφαλμάτων στην αλληλούχηση περιοχών με χαμηλή κάλυψη, τέτοιες ‘’νησίδες’’ μπορεί να αποτελούν ψευδοσυνενώσεις. Για κάθε περιοχή χαμηλής κάλυψης ή χαμηλής ποιότητας ο αλγόριθμος χρησιμοποιεί το γονιδίωμα αναφοράς για να αναφέρει τις βάσεις. Τα περισσότερα ‘‘διαβάσματα’’ που καλύπτουν τα άκρα των εξωνίων θα καλύπτουν επίσης κόμβους ματίσματος. Τα άκρα των εξωνίων στις ‘’ψευδονησίδες’’ αρχικά θα καλύπτονται από λίγα ‘‘διαβάσματα’’ και σαν αποτέλεσμα οι ‘’ψευδονησίδες’’ των εξωνίων πιθανόν θα έχουν έλλειψη αλληλουχιών στο κάθε άκρο τους.

Τα γονίδια που μεταφράζονται σε χαμηλό επίπεδο, όταν αλληλουχούνται θα έχουν χαμηλή κάλυψη και για τα εξώνια μπορεί να έχουν κενά. Για αυτές τις περιπτώσεις ο αλγόριθμος περιέχει μια παράμετρο που ελέγχει εάν δύο ξεχωριστά αλλά γειτονικά εξώνια θα συγχωνευθούν σε ένα μοναδικό εξώνιο.
Για να χαρτογραφηθούν τα ‘‘διαβάσματα’’ σε κόμβους ματίσματος ο αλγόριθμος πρώτα καταμετρά όλα τα σημεία κανονικών δεκτών και δοτών στις αλληλουχίες των ‘’νησίδων’’. Κατόπιν λαμβάνει υπόψη όλα τα ζεύγη αυτών των σημείων που μπορούν να σχηματίσουν κανονικά (GT-AG) εσώνια μεταξύ γειτονικών ‘’νησίδων’’. Κάθε πιθανό εσώνιο ελέγχεται σε σχέση με τα IUM ‘‘διαβάσματα’’ για ‘‘διαβάσματα’’ που καλύπτουν κόμβους ματίσματος όπως περιγράφεται παρακάτω.

Χρησιμοποιώντας την παράμετρο προεπιλογής, ο αλγόριθμος ελέγχει μόνο εσώνια με μήκος μεταξύ 70bp-20.000bp. Αυτές οι τιμές περιγράφουν τη συντριπτική πλειοψηφία των γνωστών ευκαρυωτικών εσωνίων. Για να βελτιωθεί ο χρόνος που απαιτείται για την ανάλυση και να αποφευχθεί η αναφορά ψευδών θετικών, το πρόγραμμα εξαιρεί τα ζεύγη δεκτών –δοτών που βρίσκονται εξ ολοκλήρου σε μοναδικές ‘’νησίδες ’’, εκτός και αν η ‘’νησίδα’’ είχε μεγάλο βάθος αλληλούχησης.

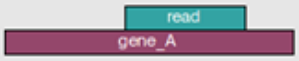
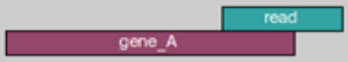
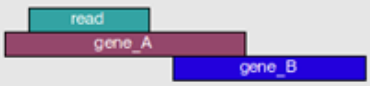
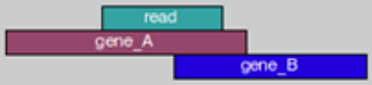
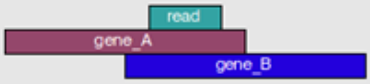
Για κάθε κόμβο ματίσματος ο TopHat ψάχνει στα IUM ‘’διαβάσματα’’ με στόχο να βρει ‘’διαβάσματα’’ που καλύπτουν κόμβους χρησιμοποιώντας μια στρατηγική ‘’αρχικής αλληλουχίας και επέκτασης ‘’(seed and extent). Η διαδικασία περιλαμβάνει την κατάταξη των IUM ‘’διαβασμάτων’’ χρησιμοποιώντας έναν απλό Πίνακα επιταχύνοντας έτσι τη διαδικασίας αναζήτησης αλληλουχιών ματίσματος σε πολλά ‘’διαβάσματα’’. Ο αλγόριθμος βρίσκει τα ‘’διαβάσματα’’ που καλύπτουν κόμβους ματίσματος με τουλάχιστον κ βάσεις σε κάθε πλευρά (όπου κ=5bp από προεπιλογή). Έτσι ο πίνακας είναι εμπλουτισμένος με 2κ-μερή όπου 2κ-μερές συνδέεται με τα ‘’διαβάσματα’’ που περιέχουν αυτά τα 2κ-μερές. Για κάθε διάβασμα ο πίνακας περιέχει (S-2κ+1) εγγραφές που αντιστοιχούν σε πιθανές θέσεις όπου ένα σημείο ματίσματος μπορεί να βρίσκεται μέσα σε ένα διάβασμα όπου S είναι το μήκος της περιοχής υψηλής ποιότητας του 5΄-άκρου. Κατόπιν ο αλγόριθμος παίρνει το κάθε πιθανό κόμβο ματίσματος και δημιουργεί μια 2κ-μερή ‘’αρχική αλληλουχία’’ (seed) συνενώνοντας τις κ βάσεις καθοδικά από τον δέκτη με τις κ βάσεις ανοδικά του δότη. Ύστερα ο πίνακας κατάταξης με τα IUM ‘’διαβάσματα’’ εξετάζεται με αυτά τα 2κ-μερή για να βρεθούν όλα τα ‘’διαβάσματα’’ που περιέχουν την ‘’αρχική αλληλουχία’’. Το 2κ-μερές με τέλειο ταίριασμα επεκτείνεται για να βρεθούν όλα τα ‘’διαβάσματα’’ που καλύπτουν κόμβους ματίσματος. Για να επεκτείνει τα τέλεια ταιριάσματα των περιοχών με ‘’αρχικές αλληλουχίες ’’ ο αλγόριθμος ευθυγραμμίζει το τμήμα του διαβάσματος στα δεξιά και αριστερά της ‘’αρχικής αλληλουχίας’’ με την αριστερή και δεξιά ‘’νησίδα’’ αντίστοιχα επιτρέποντας έναν αριθμό σφαλμάτων καθορισμένο από το χρήστη.

Ο αλγόριθμος αναφέρει όσες ευθυγραμμίσεις ματίσματος βρίσκει και στη συνέχεια δημιουργεί μια ομάδα μη-πλεοναζόντων κόμβων ματίσματος χρησιμοποιώντας αυτές τις ευθυγραμμίσεις. Για κάθε κόμβο, ο μέσος όρος βάθους κάλυψης του διαβάσματος υπολογίζεται από την αριστερή και δεξιά πλευρική περιοχή του κάθε κόμβου ξεχωριστά. Ο αριθμός των χαρτογραφημένων ‘’διαβασμάτων’’ που διασχίζουν τον κόμβο διαιρείται από την κάλυψη των πιο βαθιά επικαλυμμένων πλευρών για να βρεθεί μια εκτίμηση της ελάχιστης συχνότητας των ισομορφών. Εάν ο TopHat εκτιμήσει ότι οι κόμβοι ματίσματος υπάρχουν σε ποσοστό μικρότερο του 15% του βάθους επικάλυψης των εξωνίων που βρίσκονται πλευρικά από αυτά τότε το μάτισμα δεν αναφέρεται.

2.13.4 ΑΛΓΟΡΙΘΜΟΣ HTSeq-count

Όταν πραγματοποιηθεί η χαρτογράφηση των ‘‘διαβασμάτων’’, μια κοινή ανάλυση είναι ο υπολογισμός των ‘‘διαβασμάτων’’ που αντιστοιχούν σε μια περιοχή ενδιαφέροντος. Τέτοιες περιοχές είναι ένα διάστημα σε ένα χρωμόσωμα ή η ένωση τέτοιων περιοχών. Στην περίπτωση του RNA-seq περιοχές ενδιαφέροντος μπορεί να είναι τα γονίδια, όπου γονίδιο εδώ θεωρείται η συνένωση όλων των εξωνίων του. Σαν περιοχή ενδιαφέροντος μπορεί να χρησιμοποιηθεί και το κάθε εξώνιο ώστε να μελετηθεί το εναλλακτικό μάτισμα. Ο αλγόριθμος χρησιμοποιεί ένα αρχείο GFF το οποίο περιέχει πληροφορίες για τις θέσεις των γονιδίων στα χρωμοσώματα του οργανισμού προς μελέτη, δομημένα σε ένα πίνακα με 9 στήλες. Οι στήλες αυτές περιέχουν πληροφορίες για το όνομα της αλληλουχίας, το πρόγραμμα από το οποίο εξήχθη η αλληλουχία αυτή, τη θέση έναρξης και λήξης της αλληλουχίας, μια τιμή που καθορίζει την ένταση του χρώματος που θα απεικονισθεί αυτή η αλληλουχία στον περιηγητή και την αλυσίδα του DNA στην οποία είναι τοποθετημένες. Ο HTSeq αλγόριθμος δίνει τη δυνατότητα 3 διαφορετικών τρόπων χαρτογράφησης ενός ‘‘διαβάσματος’’ στα εξώνια.

Εάν το ‘‘διάβασμα’’ περιέχει ένα μοναδικό χαρακτηριστικό (εξώνιο) το διάβασμα αντιστοιχεί στο συγκεκριμένο. Εάν περιέχει περισσότερα από 1 χαρακτηριστικά θεωρείται αμφίβολο και εάν δεν περιέχει κάποιο χαρακτηριστικό τότε το διάβασμα θεωρείται χωρίς χαρακτηριστικό (no-feature). Στον Πίνακα 1 φαίνεται τρόπος με τον οποίο χαρακτηρίζονται τα ‘‘διαβάσματα’’ ανάλογα με τη θέση του στις διάφορες λειτουργίες.

	Συνένωση	Αυστηρή-διατομή	Κενή-διατομή
	Γονίδιο Α	Γονίδιο Α	Γονίδιο Α
	Γονίδιο Α	χωρίς χαρακτηριστικό	Γονίδιο Α
	Γονίδιο Α	χωρίς χαρακτηριστικό	Γονίδιο Α
	Γονίδιο Α	Γονίδιο Α	Γονίδιο Α
	Γονίδιο Α	Γονίδιο Α	Γονίδιο Α
	Αμφίβολο	Γονίδιο Α	Γονίδιο Α
	Αμφίβολο	Αμφίβολο	Αμφίβολο

*Πίνακας 1: Τρόπος χαρακτηρισμού των διαβασμάτων ανάλογα με την επιλεγμένη λειτουργία (*Συνένωση, *αυστηρή-διατομή ή κενή-διατομή).*

2.13.5 ΑΛΓΟΡΙΘΜΟΣ DESeq

Ένα βασικό στάδιο στην ανάλυση δεδομένων RNA-seq είναι η ανίχνευση γονιδίων των οποίων η έκφραση αλλάζει σημαντικά μεταξύ των 2 συνθηκών-πειραμάτων που μελετούνται. Το πακέτο DESeq (Anders S et al. 2010) παρέχει μεθόδους ελέγχου της διαφορετικής έκφρασης των γονιδίων βασισμένο στην αρνητική διωνυμική κατανομή (NB) και σε έναν εκτιμητή συρρίκνωσης για τη διακύμανση διανομής. Σαν αρχικό αρχείο επεξεργασίας το πρόγραμμα δέχεται ένα πίνακα με τις ακέραιες τιμές του αριθμού των ‘‘διαβασμάτων’‘ που έχουν στοιχηθεί σε κάθε γονίδιο.

ΠΕΡΙΓΡΑΦΗ ΠΡΟΓΡΑΜΜΑΤΟΣ

Θεωρείτε ότι ο αριθμός των ‘‘διαβασμάτων’’ σε ένα δείγμα j που έχουν στοιχηθεί σε ένα γονίδιο i μπορεί να μοντελοποιηθεί από μια αρνητική διωνυμική κατανομή,

$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \sigma_{ij}^2)$,

η οποία έχει δυο παραμέτρους, το μέσο μ_{ij} και τη διακύμανση σ_{ij}^2. Οι αναγνωσμένες μετρήσεις K_{ij} είναι ακέραιες τιμές μη αρνητικές. Η κατανομή NB χρησιμοποιείται συνήθως για να μοντελοποιήσει δεδομένα με μεγάλη διασπορά.

Στην πράξη, οι παράμετροι μ_{ij} και σ_{ij}^2 δεν είναι γνωστές και πρέπει να εκτιμηθούν από τα δεδομένα. Τυπικά, ο αριθμός των επαναλήψεων είναι μικρός, και περαιτέρω υποθέσεις μοντελοποίησης χρειάζεται να γίνουν ώστε να προκύψουν χρήσιμες εκτιμήσεις.

Η μέση παράμετρος μ_{ij}, η οποία είναι, η αναμενόμενη τιμή των παρατηρούμενων μετρήσεων για το γονίδιο i στο δείγμα j, είναι προϊόν μιας εξαρτώμενης συνθήκης ανά τιμή γονιδίου $q_{i,\rho(j)}$ (όπου ρ(j) είναι η πειραματική συνθήκη ενός δείγματος j) και ένας συντελεστής μεγέθους s_j,

$\mu_{ij} = q_{i,\rho(j)} s_j$

Η παράμετρος $q_{i,\rho(j)}$ είναι ανάλογη της αναμενόμενης τιμής της πραγματικής (αλλά άγνωστης) συγκέντρωσης των τμημάτων από το γονίδιο i υπό συνθήκη ρ(j). Ο συντελεστής μεγέθους s_j αναπαριστά την κάλυψη, ή βάθος δείγματος, της βιβλιοθήκης j, και θα χρησιμοποιηθεί ο όρος *συνήθης κλίμακα* για ποσότητες, όπως $q_{i,\rho(j)}$, οι οποίες προσαρμόζονται για κάλυψη διαιρώντας με s_j.

Η διακύμανση σ_{ij}^2 είναι το άθροισμα μιας παραμέτρου θορύβου (shot noise term) και μιας παραμέτρου διακύμανσης που δεν έχει υποστεί επεξεργασία (raw variance term),

$$\sigma_{ij}^2 = \underbrace{\mu_{ij}}_{\text{shot noise}} + \underbrace{s_j^2 v_{i,\rho(j)}}_{\text{raw variance}}$$

Τέλος θεωρείται ότι η ανά γονίδιο μη τροποποιημένη παράμετρος διακύμανσης $u_{i,\rho(j)}$ είναι μια ομαλή συνάρτηση ως προς q_i, ρ,

$u_{i,\rho(j)} = u_\rho\,(q_{i,\rho(j)})$.

Αυτή η υπόθεση είναι απαραίτητη καθώς ο αριθμός των επαναλήψεων είναι τυπικά πολύ μικρός ώστε να είναι εφικτή μια ακριβής εκτίμηση της διακύμανσης για το γονίδιο i από τα διαθέσιμα. Αυτή η υπόθεση επιτρέπει να συγκεντρωθούν τα δεδομένα από γονίδια με παρόμοια δύναμη έκφρασης για το σκοπό της εκτίμησης της διακύμανσης.

ΕΛΕΓΧΟΣ ΓΙΑ ΔΙΑΦΟΡΙΚΗ ΕΚΦΡΑΣΗ

Έστω ότι υπάρχουν m_A επαναλαμβανόμενα δείγματα για βιολογική συνθήκη A και m_B για συνθήκη B. Για κάθε γονίδιο i, θα πρέπει να υπολογιστεί η επίπτωση στα δεδομένα για διαφορική έκφραση αυτού του γονιδίου ανάμεσα στις δυο συνθήκες. Συγκεκριμένα, επιδιώκεται η εξέταση της μηδενικής υπόθεσης $q_{iA}=q_{iB}$ όπου q_{iA} είναι η παράμετρος ‘’δύναμης έκφρασης’’ για τα δείγματα της συνθήκης A, και q_{iB} για τη συνθήκη B. Για το σκοπό αυτό, ορίζεται, ως στατιστική εξέταση (test statistic), το σύνολο των μετρήσεων σε κάθε συνθήκη

$$K_{iA} = \sum_{j:\rho(j)=A} K_{ij}, \quad K_{iB} = \sum_{j:\rho(j)=B} K_{ij}$$

και το ολικό τους άθροισμα $K_{iS} = K_{iA} + K_{iB}$. Από το σφάλμα του μοντέλου που περιγράφηκε σε προηγούμενη ενότητα, φαίνεται παρακάτω ότι υπό μηδενική υπόθεση μπορούν να υπολογιστούν οι πιθανότητες των γεγονότων $K_{iA} = a$ και $K_{iB} = b$ για κάθε ζεύγος αριθμών a και b. Αυτή η πιθανότητα δηλώνεται ως p(a,b). Η τιμή P ενός ζεύγους παρατηρούμενων μετρούμενων αθροισμάτων (K_{iA}, K_{iB}) είναι τότε το άθροισμα όλων των πιθανοτήτων ίσο ή μικρότερο της p(K_{iA}, K_{iB}), δεδομένου ότι το συνολικό άθροισμα είναι K_{iS}:

$$p_i = \frac{\sum_{\substack{a+b=k_{iS} \\ p(a,b)\le p(k_{iA},k_{iB})}} p(a,b)}{\sum_{a+b=k_{iS}} p(a,b)}$$

Οι μεταβλητές a και b στα παραπάνω αθροίσματα παίρνουν τιμές 0,..., K_{iS}.

ΥΠΟΛΟΓΙΣΜΟΣ ΤΗΣ ΠΙΘΑΝΟΤΗΤΑΣ p(a,b)

Αρχικά θεωρείται ότι υπό τη μηδενική υπόθεση, μετρήσεις από διαφορετικά δείγματα είναι ανεξάρτητες. Τότε, p(a,b)=Pr(K_{iA}=a) Pr(K_{iB}=b). Το πρόβλημα όμως είναι ο υπολογισμός της πιθανότητας του γεγονότος K_{iA} = a, και, αναλογικά, του K_{iB} = b. Η τυχαία μεταβλητή K_{iA} είναι το άθροισμα των m_A NB-κατανεμημένων τυχαίων μεταβλητών. Η κατανομή του προσεγγίζεται από μια κατανομή NB της οποίας οι παράμετροι λαμβάνονται από εκείνες της K_{ij}. Για το σκοπό αυτό, υπολογίζεται πρώτα η μέση τιμή από τις μετρήσεις και των δυο συνθηκών,

$$\hat{q}_{io} = \sum_{j:\rho(j)\in\{A,B\}} k_{ij} / s_j$$

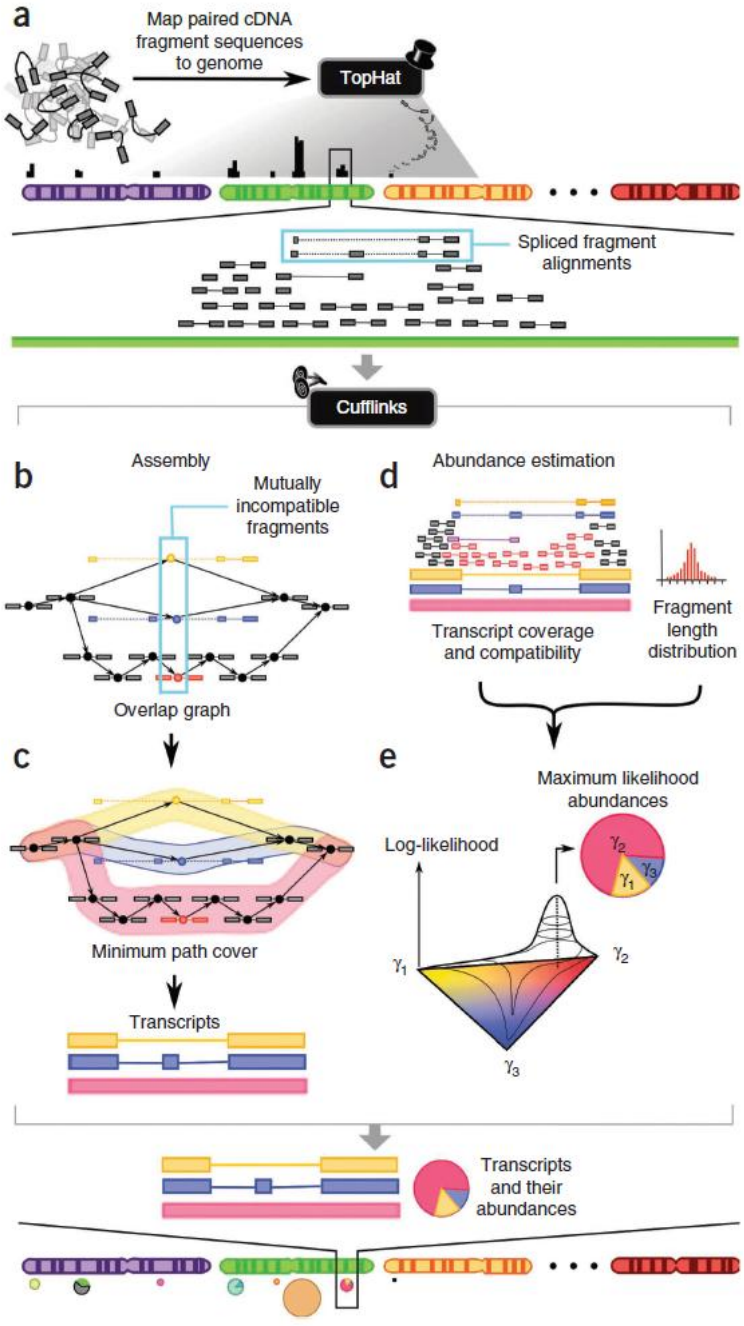
Η αθροισμένη μέση τιμή και διακύμανση για τη συνθήκη A είναι:

$$\hat{\mu}_{iA} = \sum_{j\in A} s_j \hat{q}_{io}$$

$$\hat{\sigma}^2_{iA} = \sum_{j\in A} \hat{s}_j \hat{q}_{io} + \hat{s}_j^2 \hat{v}_A(\hat{q}_{io}).$$

2.13.6 ΑΛΓΟΡΙΘΜΟΣ CUFFLINKS

Ο Cufflinks (Trapnell C et al. 2010) είναι ένας αλγόριθμος ο οποίος συναρμολογεί χαρτογραφημένα RNA-Seq ‘‘διαβάσματα’’ σε μετάγραφα, εκτιμά την αφθονία τους και εξετάζει τη διαφορετική έκφραση των γονιδίων εξάγοντας συμπεράσματα για τη ρύθμιση των επιπέδων μεταγραφής. Ο Cufflinks υποστηρίζεται τόσο από λειτουργικό σύστημα Linux όσο και από Mac OS. Στην Εικόνα 23 παρουσιάζεται συνοπτικά η λειτουργία του αλγορίθμου.



Εικόνα 23: Συνοπτική παρουσίαση της λειτουργίας του Cufflinks.

Ο αλγόριθμος χρησιμοποιεί ως αρχεία εισόδου αλληλουχίες τμημάτων cDNA οι οποίες έχουν χαρτογραφηθεί στο γονιδίωμα αναφοράς από αλγορίθμους όπως ο TopHat. Με τα διαβάσματα με συζευγμένα άκρα RNA-Seq (paired end reads), ο Cufflinks αντιμετωπίζει κάθε ζεύγος ‘‘διαβασμάτων’’ ως μια μοναδική αντιστοίχιση στο γονιδίωμα. Ο αλγόριθμος συγκεντρώνει επικαλυπτόμενες δέσμες των χαρτογραφημένων τμημάτων ξεχωριστά, μειώνοντας έτσι το χρόνο ανάλυσης και την ανάγκη σε μνήμη, καθώς κάθε δέσμη περιέχει τυπικά τα τμήματα από μερικά μόνο γονίδια. Ο Cufflinks τότε εκτιμά τις αφθονίες των συναρμολογημένων μεταγράφων. Το πρώτο βήμα στη συναρμολόγηση των τμημάτων είναι να αναγνωριστούν τα μη συμβατά τμήματα τα οποία πρέπει να έχουν προέλθει από διακριτές ισομορφές mRNA. Τα τμήματα όταν είναι συμβατά συνδέονται σε ένα "επικαλυπτόμενο γράφημα". Κάθε τμήμα έχει έναν κόμβο στο γράφημα και μια ‘‘σύνδεση’’, η οποία κατευθύνεται από τα αριστερά προς τα δεξιά κατά μήκος του γονιδιώματος και τοποθετείται μεταξύ κάθε ζεύγους συμβατών τμημάτων. Στο παράδειγμα της εικόνας 23, τα κίτρινα, μπλε και κόκκινα τμήματα πρέπει να έχουν προέλθει από ξεχωριστές ισομορφές, αλλά κάθε άλλο τμήμα θα μπορούσε να έχει προέλθει από το ίδιο μετάγραφο όπως ένα από αυτά τα τρία. Οι ισομορφές τότε συγκεντρώνονται από το επικαλυπτόμενο γράφημα (Εικόνα 23c). Μονοπάτια κατά μήκος του γραφήματος ανταποκρίνονται σε σύνολα αμοιβαία συμβατών τμημάτων τα οποία θα μπορούσαν να έχουν συγχωνευθεί σε ολοκληρωμένες ισομορφές. Το επικαλυπτόμενο γράφημα μπορεί να καλύπτεται ελάχιστα από τρία μονοπάτια (με σκίαση στο κίτρινο, μπλε και κόκκινο), αντιπροσωπεύοντας το καθένα μια διαφορετική ισομορφή. Ο Cufflinks παράγει ένα ελάχιστο σύνολο μονοπατιών που καλύπτουν όλα τα τμήματα στο επικαλυπτόμενο γράφημα, εντοπίζοντας το μέγιστο σύνολο ‘‘διαβασμάτων’’ για τα οποία ισχύει ότι δυο μονοπάτια δεν μπορούν να έχουν προέλθει από την ίδια ισομορφή. Έπειτα, εκτιμάται η αφθονία των μεταγράφων. Τα θραύσματα αντιστοιχίζονται στα μετάγραφα από τα οποία θα μπορούσαν να είχαν δημιουργήσει. Το μωβ θραύσμα θα μπορούσε να έχει προκύψει από τη μπλε ή την κόκκινη ισομορφή. Τα τμήματα με γκρι χρώμα θα μπορούσαν να έχουν προέλθει από καθένα από τα τρία που απεικονίζονται. Ο Cufflinks εκτιμά τις αφθονίες των μεταγράφων εφαρμόζοντας ένα στατιστικό μοντέλο στο οποίο η πιθανότητα παρατήρησης κάθε τμήματος είναι γραμμική συνάρτηση των αφθονιών των μεταγράφων από τα οποία θα μπορούσαν να προκύψουν. Επειδή μόνο τα άκρα κάθε θραύσματος μπαίνουν σε αλληλουχία, το μήκος καθενός ίσως να μην είναι γνωστό. Η ανάθεση ενός τμήματος σε διαφορετικές ισομορφές συχνά υποδεικνύει ένα διαφορετικό μήκος για αυτό. Ο Cufflinks ενσωματώνει την κατανομή των μηκών του τμήματος προκειμένου να συμβάλλει στην ανάθεση των τμημάτων σε ισομορφές.

ΥΠΟΛΟΓΙΣΜΟΣ ΤΗΣ ΑΦΘΟΝΙΑΣ ΤΩΝ ΜΕΤΑΓΡΑΦΩΝ

Σε πειράματα RNA-Seq, τα τμήματα cDNA αφού αλληλουχηθούν χαρτογραφούνται πίσω σε γονίδια και ιδανικά, σε μοναδικά μετάγραφα. Στη συνέχεια το άθροισμα των αλληλουχιών από το RNA-Seq μπορεί να χρησιμοποιηθεί ως μέτρο της σχετικής αφθονίας των μεταγραφών και ο Cufflinks μετράει την αφθονία των μεταγράφων σε “θραύσματα ανά 1000 βάσεις εξωνίου ανά εκατομμύριο διαβασμάτων που έχουν χαρτογραφηθεί” [**F**ragments **P**er **K**ilobase of exon per **M**illion fragments mapped (**FPKM**)], που είναι ανάλογο με το μοναδικό ‘‘διάβασμα’’ "RPKM" (Mortazavi A et al. 2008)

Στα ‘‘διαβάσματα’‘ RNA-Seq όπου πραγματοποιήθηκε αλληλούχηση και των 2 άκρων παρέχονται δυο ‘‘διαβάσματα’‘ ένα για κάθε τμήμα. Προκειμένου να εκτιμηθούν τα επίπεδα αφθονίας των ισομορφών, πρέπει να αντιστοιχηθούν τα τμήματα σε μοναδικά μετάγραφα. Αυτή η ανάλυση μπορεί να είναι δύσκολη, καθώς ένα διάβασμα μπορεί να στοιχηθεί με πολλαπλές ισομορφές του ίδιου γονιδίου.

Ο Cufflinks χρησιμοποιεί ένα στατιστικό μοντέλο σε αυτά τα ‘‘διαβάσματα’’ για να υπολογίσει την πιθανότητα για αφθονίες ενός συνόλου μεταγράφων. Αυτή η συνάρτηση πιθανότητας μπορεί να αναπαρασταθεί ώστε να έχει ένα μοναδικό μέγιστο, το οποίο ο Cufflinks εντοπίζει χρησιμοποιώντας έναν αριθμητικό αλγόριθμο βελτιστοποίησης. Στη συνέχεια το πρόγραμμα πολλαπλασιάζει αυτές τις πιθανότητες για να υπολογίσει τη συνολική πιθανότητα ότι κάποιος θα παρατηρούσε τα τμήματα στο πείραμα, δεδομένου των προτεινόμενων αφθονιών των μεταγράφων.

Εφαρμόζοντας αυτή τη στατιστική μέθοδο, ο Cufflinks μπορεί να εκτιμήσει τις αφθονίες των ισομορφών που υπάρχουν στο δείγμα, είτε χρησιμοποιώντας το γονιδίωμα αναφορά, είτε μετά από μια εξαρχής συγκέντρωση των μεταγραφών χρησιμοποιώντας μόνο το γονιδίωμα αναφοράς.

ΕΚΤΙΜΗΣΗ ΤΗΣ ΚΑΤΑΝΟΜΗΣ ΤΟΥ ΜΗΚΟΥΣ ΤΩΝ ΤΜΗΜΑΤΩΝ

Η κατανομή πιθανοτήτων του μήκους των θραυσμάτων παίζει σημαντικό ρόλο στη συναρμολόγηση, στην εκτίμηση της αφθονίας και στην διόρθωση των σφαλμάτων. Η ορθότητα της κατανομής θα έχει σημαντική επίδραση στην αξιοπιστία των αποτελεσμάτων. Εξαιτίας αυτού, γίνεται προσπάθεια να γίνει γνωστή αυτή η κατανομή από τα πρωταρχικά αντί να βασίζεται σε μια προσεγγιστική της κανονικής κατανομής, όποτε αυτό είναι δυνατόν.

- Αν παρέχονται μόνο ‘‘διαβάσματα’’ ενός άκρου, δεν υπάρχει τρόπος να εκτιμηθεί εμπειρικά η κατανομή, και επομένως ο Cufflinks θα πρέπει να χρησιμοποιήσει μια προσέγγιση της κανονικής κατανομής, με μεταβλητές είτε προκαθορισμένες είτε ορισμένες από το χρήστη.
- Εάν το αρχείο ευθυγράμμισης περιέχει ‘‘διαβάσματα’’ από αλληλούχηση των δύο άκρων και παρέχεται μια διάταξη, ο Cufflinks είναι ικανός να αντιληφθεί την κατανομή από τα ‘‘διαβάσματα’’ που χαρτογραφούνται σε γονίδια με μια ισομορφή. Μπορούν να αφαιρεθούν εσώνια που υπάρχουν ανάμεσα σε ζεύγη ‘‘διαβασμάτων’’, προσφέροντας πιο αξιόπιστη εκτίμηση.
- Εάν δοθούν ‘‘διαβάσματα’’ από αλληλούχηση των δύο άκρων και όχι η διάταξη, ο Cufflinks θα αναζητήσει μεγάλα "ανοικτά εύρη" όπου οι αντιστοιχήσεις δεν περιέχουν συνενώσεις μεταξύ των ‘‘διαβασμάτων’’. Μέσα σε αυτά τα εύρη, ο Cufflinks χρησιμοποιεί το μήκος του γονιδιώματος των ‘‘διαβασμάτων’’ αυτών για την εκτίμηση της κατανομής.

ΠΟΛΛΑΠΛΑ ΧΑΡΤΟΓΡΑΦΗΜΕΝΑ ΔΙΑΒΑΣΜΑΤΑ

Κάποια ‘‘διαβάσματα’’ μερικές φορές θα χαρτογραφηθούν σε πολλαπλές θέσεις στο γονιδίωμα εξαιτίας ύπαρξης επαναλήψεων στην αλληλουχία του ‘‘διαβάσματος’’ και ομολογίας σε διαφορετικά σημεία του γονιδιώματος. Εξ' ορισμού, ο Cufflinks θα διαχωρίσει ομοιόμορφα κάθε πολλαπλά χαρτογραφημένο ‘‘διάβασμα’’ σε όλες τις θέσεις που μπορεί να χαρτογραφηθεί. Με άλλα λόγια, μια χαρτογράφηση διαβάσματος σε 10 θέσεις θα μετρήσει σαν 10% πιθανότητα το ‘‘διάβασμα’’ να εντοπιστεί στην κάθε θέση.

Αν η διόρθωση πολλαπλά χαρτογραφημένων ‘‘διαβασμάτων’’ είναι ενεργοποιημένη, ο Cufflinks θα βελτιώσει την εκτίμησή του με έναν τρόπο ο οποίος είναι εμπνευσμένος από τη μέθοδο "διάσωσης" (Mortazavi A et al. 2008).

Ο Cufflinks θα υπολογίσει την αρχική εκτίμηση αφθονίας για όλα τα μετάγραφα εφαρμόζοντας ένα ομοιόμορφο σύστημα διαίρεσης. Στη συνέχεια, θα επανεκτιμήσει τις αφθονίες διαχωρίζοντας πολλαπλά χαρτογραφημένα ‘‘διαβάσματα’’ πιθανολογικά με βάση την αρχική εκτίμηση αφθονίας των γονιδίων που χαρτογραφεί, το μήκος θραύσματος που προκύπτει και τη διαστρέβλωση θραύσματος (αν η διόρθωση διαστρέβλωσης είναι ενεργοποιημένη).

Cuffdiff

Ο Cuffdiff (Trapnell C et al. 2013) είναι ένα πρόγραμμα που χρησιμοποιεί τη μηχανή ποσοτικοποίησης των μεταγραφών του Cufflinks για να υπολογίσει τα επίπεδα έκφρασης των γονιδίων σε διαφορετικά πειράματα και να υπολογίσει αν υπάρχουν σημαντικές διαφορές. Μπορεί να χρησιμοποιηθεί για να εντοπίσει διαφορετικά εκφρασμένα γονίδια και μετάγραφα, όπως επίσης γονίδια που ρυθμίζονται διαφορετικά σε μεταγραφικό και μετά-μεταγραφικό επίπεδο.

ΥΠΟΛΟΓΙΣΜΟΣ ΔΙΑΦΟΡΕΤΙΚΗΣ ΕΚΦΡΑΣΗΣ ΓΟΝΙΔΙΩΝ ΜΕ ΤΟΝ Cuffdiff

Η ύπαρξη σφαλμάτων μέτρησης, τεχνικής διακύμανσης και βιολογικής διακύμανσης μεταξύ των πειραμάτων επανάληψης (replicate) ίσως οδηγήσει σε μια παρατηρούμενη ψευδή αλλαγή της έκφρασης. Ο Cuffdiff αξιολογεί τη σημαντικότητα των τιμών χρησιμοποιώντας ένα μοντέλο μεταβλητότητας. Ο Cuffdiff κατασκευάζει, για κάθε συνθήκη, έναν πίνακα ο οποίος προβλέπει πόση διακύμανση υπάρχει στον αριθμό των ‘‘διαβασμάτων’‘ που προέρχονται από ένα γονίδιο ή μεταγραφή. Ο πίνακας σταθεροποιείται από το μέσο όρο των ‘‘διαβασμάτων’’ όλων των επαναλήψεων. Για να αναζητήσει τη διακύμανση για ένα μετάγραφο, ο Cuffdiff εκτιμά τον αριθμό των ‘‘διαβασμάτων’‘ που προέρχονται από αυτό το μετάγραφο και στη συνέχεια θέτει ερωτήματα στον πίνακα για να ανακτήσει τη διακύμανση για αυτό τον αριθμό των ‘‘διαβασμάτων’‘. Έπειτα, ο Cuffdiff αναφέρει για χαρτογράφηση ανάγνωσης και αβεβαιότητα εκχώρησης προσομοιώνοντας μια πιθανή ανάθεση των αναγνώσεων που έχουν χαρτογραφηθεί σε μια θέση με τις ισόμορφες ενώσεις για αυτή τη θέση. Στο τέλος της διαδικασίας εκτίμησης, ο Cuffdiff λαμβάνει μια εκτίμηση του αριθμού των ‘‘διαβασμάτων’‘ που προήλθαν από κάθε γονίδιο και μετάγραφο, μαζί με τις διακυμάνσεις για εκείνες τις εκτιμήσεις. Οι μετρήσεις των ‘‘διαβασμάτων’’ αναφέρονται μαζί με τις τιμές FKPM και τις διακυμάνσεις τους. Αλλαγή στην έκφραση αναφέρεται ως αλλαγή στο αρχείο καταγραφής σε FKPM, και οι διακυμάνσεις FKPM επιτρέπουν στο πρόγραμμα να εκτιμήσει από μόνο του τη διακύμανση στην αλλαγή του αρχείου καταγραφής. Ένα γονίδιο το οποίο έχει υψηλή διακύμανση έκφρασης θα έχει υψηλή διακύμανση στον λόγο αλλαγής της έκφρασης.

2.14 ΑΝΑΛΥΣΗ ΣΥΣΤΑΔΩΝ (CLUSTER ANALYSIS)

Η ανάλυση συστάδων ή συσταδοποίηση είναι η διαδικασία ομαδοποίησης ενός συνόλου αντικειμένων κατά τέτοιον τρόπο ώστε τα αντικείμενα της ίδιας ομάδας (ονομαζόμενα και ως συστάδες) να είναι κατά μια έννοια περισσότερο όμοια μεταξύ τους σε σχέση με τα αντικείμενα άλλων ομάδων. Αποτελεί κύριο στόχο στη διερεύνηση δεδομένων εξόρυξης και συνήθη τεχνική στη στατιστική ανάλυση δεδομένων, τα οποία χρησιμοποιούνται σε πολλούς τομείς συμπεριλαμβανομένων εκείνων της εκμάθησης μηχανών (machine learning), αναγνώρισης προτύπων (pattern recognition), ανάλυσης εικόνων και της βιοπληροφορικής.

Η ανάλυση συστάδων δεν αποτελεί έναν συγκεκριμένο αλγόριθμο, αλλά μια γενική διαδικασία η οποία πρέπει να ολοκληρωθεί. Για την επιτυχία της ανάλυσης μπορούν να χρησιμοποιηθούν ποικίλοι αλγόριθμοι, οι οποίοι διαφέρουν σημαντικά μεταξύ τους ως προς τον τρόπο με τον οποίο ορίζουν από τι συνίστανται οι συστάδες και το πως μπορούν αποτελεσματικά να τις εντοπίσουν. Για τη δημιουργία μιας συστάδας υπολογίζονται οι αποστάσεις μεταξύ των δεδομένων οι οποίες καθορίζουν την ομοιότητας μεταξύ τους. Οι πιο δημοφιλείς μέθοδοι συσταδοποίησης περιλαμβάνουν ομάδες με μικρές αποστάσεις μεταξύ των μελών της συστάδας, πυκνές περιοχές του χώρου των δεδομένων, μεσοδιαστήματα ή συγκεκριμένες στατιστικές κατανομές. Η συσταδοποίηση μπορεί επομένως να διατυπωθεί ως ένα πολύ-κριτηριακό πρόβλημα βελτιστοποίησης. Ο κατάλληλος αλγόριθμος συσταδοποίησης και οι απαραίτητες ρυθμίσεις παραμέτρων εξαρτώνται από το σύνολο των μεμονωμένων δεδομένων και την προβλεπόμενη χρήση των αποτελεσμάτων. Ως εκ τούτου, η ανάλυση συστάδων δεν είναι μια αυτοματοποιημένη εργασία αλλά μια επαναληπτική διαδικασία ανακάλυψης γνώσης ή μια διαδραστική πολυκριτηριακή (multi-objective) βελτιστοποίηση που θα περιλαμβάνει τη δοκιμή και την αποτυχία. Συχνά θα είναι απαραίτητο να γίνουν τροποποιήσεις στην προεπεξεργασία των δεδομένων και στις παραμέτρους των μοντέλων, μέχρι το αποτέλεσμα να έχει τις επιθυμητές ιδιότητες.

ΣΥΣΤΑΔΕΣ ΚΑΙ ΣΥΣΤΑΔΟΠΟΙΗΣΕΙΣ (CLUSTERS AND CLUSTERINGS)

Η έννοια της συστάδας δεν μπορεί να οριστεί επακριβώς και αυτός είναι ένας από τους λόγους που εξηγεί την ύπαρξη πολυάριθμων αλγόριθμων συσταδοποίησης. ΄Ετσι διαφορετικοί ερευνητές χρησιμοποιούν διαφορετικά μοντέλα συστάδας και για καθένα από αυτά τα μοντέλα μπορούν να δοθούν διαφορετικοί αλγόριθμοι. Η έννοια της συστάδας, όπως προέκυψε από διαφορετικούς αλγόριθμους, ποικίλλει ανάλογα με τις ιδιότητές του. Η κατανόηση αυτών των μοντέλων συστάδων αποτελεί το κλειδί για την κατανόηση των διαφορών που εντοπίζονται μεταξύ των διαφορετικών τύπων αλγόριθμων.

Τα τυπικά μοντέλα συστάδας περιλαμβάνουν:

1. Μοντέλα συνδεσιμότητας: για παράδειγμα η ιεραρχική συσταδοποίηση δημιουργεί μοντέλα τα οποία βασίζονται στην απόσταση συνδεσιμότητας.
2. Κεντροειδή μοντέλα: για παράδειγμα ο αλγόριθμος k-means, αναπαριστά κάθε συστάδα με ένα μοναδικό διανυσματικό μέσο.
3. Μοντέλα κατανομών: οι συστάδες μοντελοποιούνται χρησιμοποιώντας στατιστικές κατανομές, όπως η πολυμεταβλητή κανονική κατανομή χρησιμοποιείται από τον αλγόριθμο μεγιστοποίησης προσδοκίας (the Expectation-maximization algorithm).
4. Μοντέλα πυκνότητας: για παράδειγμα DBSCAN και OPTICS ορίζει τις συστάδες ως συνδεδεμένες πυκνές περιοχές στο χώρο των δεδομένων.
5. Μοντέλα subspace: στην συσταδοποίηση Biclustering οι συστάδες μοντελοποιούνται με τα μέλη της συστάδας και τις σχετικές ιδιότητες.
6. Μοντέλα ομάδας: ορισμένοι αλγόριθμοι δεν παρέχουν ένα προηγμένο μοντέλο για τα αποτελέσματά τους αλλά μόνο τις πληροφορίες ομαδοποίησης.
7. Μοντέλα που βασίζονται σε γραφήματα: ένα υποσύνολο κόμβων σε ένα γράφημα ώστε κάθε δυο κόμβοι του υποσυνόλου να συνδέονται με μια ακμή, μπορεί να θεωρηθεί ως μια πρωτότυπη μορφή συστάδας.

ΣΥΣΤΑΔΟΠΟΙΗΣΗ ΒΑΣΙΣΜΕΝΗ ΣΤΗ ΣΥΝΔΕΣΙΜΟΤΗΤΑ (CONNECTIVITY BASED CLUSTERING-HIERARCHICAL CLUSTERING)

Η ομαδοποίηση βασισμένη στη συνδεσιμότητα, επίσης γνωστή και ως ιεραρχική συσταδοποίηση, βασίζεται στην κεντρική ιδέα ότι ένα στοιχείο σχετίζεται περισσότερο με τα γειτονικά του στοιχεία από ότι με πιο απομακρυσμένα. Αυτός ο αλγόριθμος συνδέει στοιχεία για να σχηματίσει τις συστάδες βασισμένος στην απόστασή τους. Μια συστάδα μπορεί να περιγραφεί ευρέως από τη μέγιστη απόσταση που απαιτείται για να συνδέσει τα στοιχεία που την αποτελούν. Σε διαφορετικές αποστάσεις, διαφορετικές συστάδες θα δημιουργηθούν και μπορούν να συμβολισθούν με τη χρήση ενός δενδρογράμματος. Αυτός ο αλγόριθμος δεν παρέχει ένα μοναδικό διαχωρισμό των δεδομένων αλλά παρέχει μια εκτενή ιεράρχηση των συστάδων που συγχωνεύονται μεταξύ τους σε συγκεκριμένες αποστάσεις. Σε ένα δενδρόγραμμα ο άξονας των Y αντιπροσωπεύει την απόσταση στην οποία συγχωνεύονται οι συστάδες και τα στοιχεία είναι τοποθετημένα στον άξονα των X ώστε οι συστάδες να μην αναμειγνύονται.

Η ιεραρχική συσταδοποίηση είναι μια ομάδα μεθόδων που διαφέρουν στον τρόπο με τον οποίο υπολογίζονται οι αποστάσεις. Εκτός από την επιλογή των αποστάσεων ο χρήστης πρέπει να επιλέξει και τα κριτήρια συνδεσιμότητας των στοιχείων μιας συστάδας. Συνηθισμένες επιλογές είναι η μονή συνδεσιμότητα συστάδας (single-linkage clustering), η ολοκληρωμένη συνδεσιμότητα συστάδας (complete linkage clustering) και ο μέσος όρος συνδεσιμότητας συστάδων (average linkage clustering).

Παρακάτω (Πίνακας 2) φαίνονται οι συνηθισμένοι τρόποι υπολογισμού της απόστασης στην ιεραρχική συσταδοποίηση.

Όνομα	**Τύπος**
Ευκλείδεια απόσταση	$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$
Ευκλείδεια απόσταση στο τετράγωνο	$\|a-b\|_2^2 = \sum_i (a_i - b_i)^2$
Απόσταση Manhattan	$\|a-b\|_1 = \sum_i \|a_i - b_i\|$
Μέγιστη απόσταση	$\|a-b\|_\infty = \max_i \|a_i - b_i\|$
Απόσταση Mahalanobis	$\sqrt{(a-b)^\top S^{-1}(a-b)}$
Συνημίτονο ομοιότητας	$\frac{a \cdot b}{\|a\|\|b\|}$

Πίνακας 2: Τρόποι υπολογισμού της απόστασης μεταξύ των ομάδων στην ιεραρχική συσταδοποίηση.

Μερικά συνηθισμένα κριτήρια συνδεσιμότητας μεταξύ των παρατηρήσεων A και B είναι τα εξής (Πίνακας 3):

Όνομα	Τύπος
Ολοκληρωμένη συνδεσιμότητα συστάδας	$\max \{ d(a,b) : a \in A, b \in B \}.$
Μονή συνδεσιμότητα συστάδας	$\min \{ d(a,b) : a \in A, b \in B \}.$
Μέσος όρος συνδεσιμότητας συστάδων, ή UPGMA	$\frac{1}{\|A\|\|B\|} \sum_{a \in A} \sum_{b \in B} d(a,b).$
Συσταδοποίηση ελάχιστης ενέργειας	$\frac{2}{nm} \sum_{i,j=1}^{n,m} \|a_i - b_j\|_2 - \frac{1}{n^2} \sum_{i,j=1}^{n} \|a_i - a_j\|_2$

Πίνακας 3: Κριτήρια συνδεσιμότητας στην ιεραρχική συσταδοποίηση.

Επιπλέον η ιεραρχική συσταδοποίηση μπορεί να είναι συσσωρευτική, δηλαδή να αρχίζει από μοναδικά στοιχεία και να τα συσσωρεύει σε μια ομάδα, ή μπορεί να είναι διαιρετική, δηλαδή να αρχίζει με το σύνολο των δεδομένων και να τα διαχωρίζει σε συστάδες.

Η επιλογή της κατάλληλης μεθόδου υπολογισμού της απόστασης θα καθορίσει το σχήμα της συστάδας, καθώς κάποια στοιχεία μπορεί να είναι κοντά σε κάποια άλλα και απομακρυσμένα από άλλα. Για παράδειγμα σε επίπεδο 2 διαστάσεων η απόσταση μεταξύ του σημείου (1,0) και τη αρχής (0,0) είναι πάντα 1 σύμφωνα με τους συνήθεις κανόνες αλλά η απόσταση μεταξύ του σημείου (1,1) και της αρχής (0,0) μπορεί να είναι 2, $\sqrt{2}$ ή 1 σύμφωνα με την απόσταση Manhattan, την Ευκλείδεια απόσταση ή τη μέγιστη απόσταση αντίστοιχα.

ΚΕΝΤΡΟΕΙΔΗΣ ΒΑΣΙΣΜΕΝΗ ΣΥΣΤΑΔΟΠΟΙΗΣΗ (CENTROID-BASE CLUSTERING- k-MEANS CLUSTERING)

Στη κεντροειδή βασισμένη συσταδοποίηση, οι συστάδες αντιπροσωπεύονται από ένα κεντρικό διάνυσμα το οποίο δε χρειάζεται να ανήκει στα δεδομένα. Όταν ο αριθμός των συστάδων είναι σταθεροποιημένος στο K, η συσταδοποίηση του μέσου K πραγματοποιεί την εξής διαδικασία που θα ορίσει τη δομή των συστάδων:
εντοπίζει τα K κέντρα των συστάδων και προσδιορίζει τα στοιχεία στα πλησιέστερα κέντρα τους έτσι ώστε το τετράγωνο της απόστασης από τη συστάδα να ελαχιστοποιείται.

Η διαδικασία αυτή καλείται NP-hard. Μια μέθοδος για την πραγματοποίηση αυτής της διαδικασίας είναι η χρήση του αλγορίθμου Lloyd's, ο οποίος αναφέρεται και σαν αλγόριθμος του K-μέσου. Ο αλγόριθμος αυτός υπολογίζει μόνο το τοπικό βέλτιστο και συνήθως τρέχει πολλαπλές φορές με διαφορετικές τυχαίες αρχικές ρυθμίσεις. Η ποικιλότητα του K-μέσου περιλαμβάνει αρχικές ρυθμίσεις όπως η επιλογή από τις καλύτερες πολλαπλές αναλύσεις, αλλά επίσης περιορίζει τον αριθμό των στοιχείων του κέντρου επιλέγοντας τη διάμεσο. Οι περισσότεροι αλγόριθμοι K-μέσου απαιτούν να είναι καθορισμένος από την αρχή ο αριθμός των K συστάδων. Επιπλέον οι αλγόριθμοι αυτοί επιλέγουν συστάδες παρόμοιου μεγέθους καθώς τοποθετούν ένα στοιχείο στο πλησιέστερο κέντρο. Αυτό συχνά οδηγεί σε λανθασμένα όρια απόρριψης μεταξύ των συστάδων.

ΣΤΟΧΟΣ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Η τεχνολογία και η πληροφορική αποτελούν πλέον αναγκαίο εργαλείο για την ανάλυση και τη διαχείριση των δεδομένων που εξάγονται από τις σύγχρονες τεχνολογίες αλληλούχησης DNA, RNA αλλά και των μικροσυστοιχιών DNA. Μέσω αυτών υπάρχει δυνατότητα υπολογισμού των σχετικών επιπέδων έκφρασης όλων των γονιδίων ενός οργανισμού καθώς και η ταυτοποίηση των θέσεων πρόσδεσης των μεταγραφικών παραγόντων που ρυθμίζουν την ενεργοποίηση ή την αποσιώπηση συγκεκριμένων γονιδίων. Η συντονισμένη δράση των πρωτεϊνών αυτών και η στοχευμένη έκφραση συγκεκριμένων ομάδων γονιδίων είναι καθοριστική για την κυτταρική επιβίωση, την ομοιόσταση και την αλληλεπίδραση γειτονικών κυττάρων και ιστών. Οι ρυθμιστικοί μηχανισμοί οι οποίοι καθορίζουν όλες αυτές τις διαδικασίες διαφέρουν ανάλογα με τον κυτταρικό τύπο και τα εξωτερικά ερεθίσματα που δέχονται. Τέλος οι διαφορές στον αριθμό των γονιδίων και στα επίπεδα έκφρασης αυτών, πιθανώς να οφείλονται σε διαφορετικές χημικές τροποποιήσεις π.χ. ακετυλίωση, μεθυλίωση των ιστονών, στην αρχιτεκτονική της χρωματίνης, στην ακριβή θέση του νουκλεοσώματος σε σχέση με τον υποκινητή του ίδιου γονιδίου και στην παρουσία ισομορφών ιστονών σε αυτό.

Αναλύσεις δεδομένων δημοσιευμένων εργασιών αλλά και δεδομένων διαθέσιμων στην ερευνητική κοινότητα μέσω βάσεων δεδομένων όπως η GEO, η EBI, η SRA θα δώσουν νέες πληροφορίες για τη γονιδιακή έκφραση αλλά και τους παράγοντες που ρυθμίζουν την έκφραση συγκεκριμένων ομάδων γονιδίων σε ένα κύτταρο, ιστό ή έναν ολόκληρο οργανισμό. Η συσχέτιση των επιπέδων έκφρασης των γονιδίων από πειράματα μικροσυστοιχιών ή αλληλούχηση RNA νέας γενιάς (RNA-seq) με τις περιοχές πρόσδεσης μεταγραφικών παραγόντων από αντίστοιχα πειράματα αλληλούχησης χρωματινικής ανοσοκατακρήμνισης (ChIP-seq), καθιστά δυνατή τη δημιουργία λειτουργικών ρυθμιστικών δικτύων που εμπεριέχουν όλες τις αλληλεπιδράσεις μεταξύ των cis και των trans ρυθμιστικών στοιχείων της μεταγραφής. Στην παρούσα διπλωματική εργασία πραγματοποιήθηκε συσχέτιση των θέσεων πρόσδεσης διαφόρων μεταγραφικών παραγόντων με τα επίπεδα έκφρασης των γονιδίων σε ανθρώπινες κυτταρικές σειρές και ιστούς μετά από ιική μόλυνση.

Στόχος της παρούσας διπλωματικής εργασίας είναι η δημιουργία ενός πληροφορικού λειτουργικού δικτύου το οποίο θα συνδέει τα γονίδια, τα οποία αποκρίνονται με μεγαλύτερη συχνότητα μετά από ιική μόλυνση σε ανθρώπινα κύτταρα ή ιστούς, με μεταγραφικούς παράγοντες αλλά και με τα επίπεδα μεθυλίωσης ή ακετυλίωσης των ιστονών που βρίσκονται κοντά σε ρυθμιστικές περιοχές των γονιδίων αυτών. Μέσω του δικτύου αυτού θα μπορέσουμε να εντοπίσουμε ‘’διακόπτες’’ μεταγραφικούς παράγοντες και γονίδια τα οποία οδηγούν στην ενεργοποίηση εναλλακτικών μονοπατιών ως απόκριση των κυττάρων στην ιική μόλυνση.

ΜΕΘΟΔΟΛΟΓΙΑ

4.1 ΑΝΑΛΥΣΗ DNA ΜΙΚΡΟΣΥΣΤΟΙΧΙΩΝ

Η συγκέντρωση δεδομένων από DNA μικροσυστοιχίες προς ανάλυση πραγματοποιήθηκε με αναζήτηση στις βάσεις δεδομένων GEO Datasets, SRA και Array Express Archive του EMBL. Αυτές οι βάσεις δεδομένων περιέχουν χιλιάδες δεδομένα. Η πλατφόρμα που επιλέχθηκε να μελετηθεί είναι η HG-U133_Plus_2 η οποία έχει χρησιμοποιηθεί ευρέως τα τελευταία χρόνια σε τουλάχιστον 3500 μελέτες όπου πραγματοποιήθηκαν περίπου 100.000 πειράματα και ανιχνεύει το σύνολο του μεταγραφώματος. Η πλατφόρμα αυτή είναι μια μικροσυστοιχία της εταιρείας Affymetrix που περιέχει 54.000 γονίδια και ESTs. Για τη παρούσα εργασία χρησιμοποιήθηκαν DNA μικροσυστοιχίες από 24 δημοσιευμένες εργασίες στις οποίες πραγματοποιήθηκαν συνολικά 60 πειράματα. Τα πειράματα αυτά πραγματοποιήθηκαν σε 16 κυτταρικές σειρές και χρησιμοποιήθηκαν 12 διαφορετικά ιικά και 3 βακτηριακά στελέχη (Πίνακας 4).

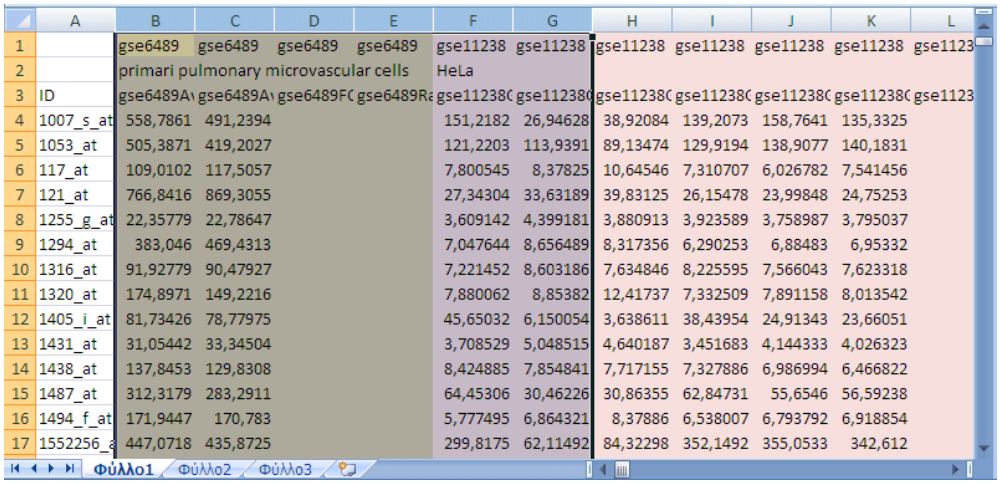
Αρχικά ελήφθησαν τα μη επεξεργασμένα δεδομένα σε μορφή CEL. Τα CEL αρχεία έχουν αποθηκευμένα τα αποτελέσματα από τους υπολογισμούς της έντασης του φθορισμού των pixel που είχαν ανιχνευθεί κατά τη σάρωση των μικροσυστοιχιών. Κατόπιν τα αρχεία αυτά αναλύθηκαν στο πρόγραμμα Expression Console της Affymetrix μέσω του οποίου ελέγχθηκε η ποιότητα των δεδομένων με θηκογράμματα και ιστογράμματα. Το πρόγραμμα αυτό παρέχει τη δυνατότητα χρήσης αλγορίθμων όπως ο MAS5, ο RMA και ο GCRMA για τη διόρθωση του υποβάθρου-θορύβου και την κανονικοποίηση των δειγμάτων.

Ο αλγόριθμος που επιλέχθηκε να χρησιμοποιηθεί ήταν ο RMA ο οποίος χρησιμοποιεί πολλαπλές πλακέτες για να πραγματοποιήσει την κανονικοποίηση, το βήμα αυτό κρίνεται απαραίτητο, αφού αναλύθηκαν δεδομένα από πολλαπλές πλακέτες. Με τον αλγόριθμο αυτό διορθώθηκε ο θόρυβος του υποβάθρου των εντάσεων φθορισμού των ιχνηθετών. Κατόπιν δημιουργήθηκαν θηκογράμματα και ιστογράμματα και ελέγχθηκε η ποιότητα των μικροσυστοιχιών.

Η εφαρμογή στατιστικών τεστ για τον εντοπισμό της διαφορικής γονιδιακής έκφρασης πραγματοποιήθηκε στο excel και το MEV. Οι λογαριθμισμένες τιμές απολογαριθμίστηκαν και πραγματοποιήθηκε κανονικοποίηση με τη διάμεσο (median normalization) όπου κάθε τιμή της έντασης ενός γονιδίου διαιρέθηκε με τη διάμεση τιμή των εντάσεων όλων των γονιδίων της συγκεκριμένης μικροσυστοιχίας. Κατόπιν πραγματοποιήθηκε students t-test στο πρόγραμμα MEV, το οποίο χρησιμοποιείται για να καθορίσει εάν 2 ομάδες δεδομένων διαφέρουν σημαντικά μεταξύ τους και πραγματοποιείται όταν τα δεδομένα ακολουθούν κανονική κατανομή. Επιλέχθηκε ως κατώφλι για την σημαντικότητα των γονιδίων η τιμή p-value $p \leq 0.05$ για το t-τεστ. Με βάση αυτό το τεστ επιλέχθηκαν τα στατιστικώς σημαντικά γονίδια με τα οποία συνεχίστηκε η ανάλυση των δεδομένων. Έπειτα υπολογίστηκε ο μέσος όρος από τις τιμές των πειραμάτων που πραγματοποιήθηκαν σε επαναλήψεις (replicates). Χρησιμοποιώντας αυτούς τους μέσους όρους βρέθηκαν οι τιμές του λόγου αλλαγής έκφρασης (fold change) του κάθε γονιδίου διαιρώντας την τιμή του μέσου όρου των δειγμάτων που είχαν μολυνθεί με ιό με την τιμή του μέσου όρου των μαρτύρων. Ως κατώφλι για το λόγο αλλαγής της έκφρασης των γονιδίων χρησιμοποιήθηκε η τιμή 1.5 η οποία χρησιμοποιείται στις περισσότερες μελέτες. Ύστερα πραγματοποιήθηκε στοίχιση και ενοποίηση των δεδομένων από όλα τα πειράματα με τη χρήση της γλώσσα προγραμματισμού SQL. Ως πρωτεύων κλειδί χρησιμοποιήθηκε ο κωδικός του κάθε γονιδίου που είναι κοινός σε όλα τα πειράματα και έτσι δημιουργήθηκε μια βάση δεδομένων με τους λόγους αλλαγής της έκφρασης όλων των γονιδίων σε όλα τα πειράματα (Εικόνα 24).

GEO accession number-paper title	Κυτταρική σειρά	Μόλυνση
gse6489	primari pulmonary microvascular cells	Human herpesvirus-8 (HHV-8)
gse11238	HeLa	Vaccinia virus (VV)
gse12108	peripheral blood monocyte	Francisella tularensis tularensis
gse12806	DC	Chlamydia pneumoniae
gse13637	HUVEC	PR8, FPV or H5N1 virus
gse16354	lymphatic endothelial cell and blood vessel endothelial cell	KSHV virus (HHV8)
gse17400	Bronchial epithelial cell line 2B4	SARS-CoV virus and Dhori virus (DHOV)
gse18906	EPCs(erythoid progenitor cells)	B19V NS1 virus
gse19810	trophoblast cell monolayers	Porphyromonas gingivalis
gse20948	Hepatoma cells	Hepatitis C Virus
gse24132	DC	Respiratory syncytial virus (RSV)
gse24897	macrophages	Porphyromonas gingivalis
gse30723	ATII and AM	Influenza virus PR8
gse31471	A549	Influenza virus A/Duck/Malaysia/01 (H9N2)
gse31518	A549	Influenza virus A/Singapore/478/2009 (pH1N1)
gse34628	HUVEC	Dengue virus
gse35283	monocytes	Low (PR8) and high pathogenic influenza viruses (FPV and H5N1)
gse37715	Human hepatocytes	Hepatitis C virus (HCV)
gse38941	liver	Hepatitis B virus (HBV)
gse40281	HUVEC	Avian influenza virus A/FPV/Bratislava/79 (H7N7)
gse42088	Dendritic Cell	Leishmania major
gse46528	huh7 (differentiated hepatocyte derived cellular carcinoma cell line)	Hepatitis C Virus
gse48466	lung epithelial cells	H1N1 influenza virus BN/59,KY/136 or KY/180 strain
Genomic analysis reveals a novel nuclear factor-κB (NF-κB)-binding site in Alu-repetitive elements	HeLa	Sendai virus

Πίνακας 4: Πειράματα DNA μικροσυστοιχιών που χρησιμοποιήθηκαν.



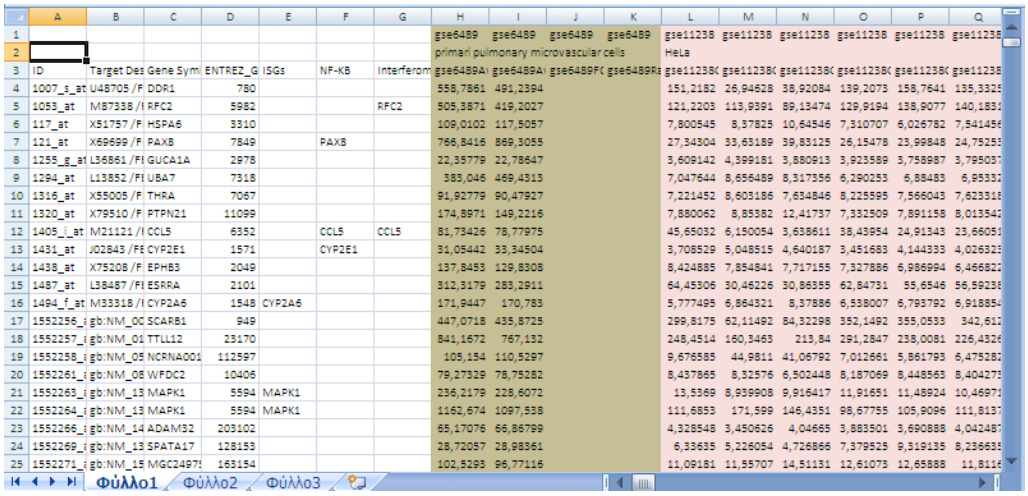
	A	B	C	D	E	F	G	H	I	J	K	L
1		gse6489	gse6489	gse6489	gse6489	gse11238	gse11238	gse11238	gse11238	gse11238	gse11238	gse1123
2		primari pulmonary microvascular cells				HeLa						
3	ID	gse6489A\	gse6489A\	gse6489FC	gse6489Ra	gse11238C	gse11238C	gse11238C	gse11238C	gse11238C	gse11238C	gse1123
4	1007_s_at	558,7861	491,2394			151,2182	26,94628	38,92084	139,2073	158,7641	135,3325	
5	1053_at	505,3871	419,2027			121,2203	113,9391	89,13474	129,9194	138,9077	140,1831	
6	117_at	109,0102	117,5057			7,800545	8,37825	10,64546	7,310707	6,026782	7,541456	
7	121_at	766,8416	869,3055			27,34304	33,63189	39,83125	26,15478	23,99848	24,75253	
8	1255_g_at	22,35779	22,78647			3,609142	4,399181	3,880913	3,923589	3,758987	3,795037	
9	1294_at	383,046	469,4313			7,047644	8,656489	8,317356	6,290253	6,88483	6,95332	
10	1316_at	91,92779	90,47927			7,221452	8,603186	7,634846	8,225595	7,566043	7,623318	
11	1320_at	174,8971	149,2216			7,880062	8,85382	12,41737	7,332509	7,891158	8,013542	
12	1405_i_at	81,73426	78,77975			45,65032	6,150054	3,638611	38,43954	24,91343	23,66051	
13	1431_at	31,05442	33,34504			3,708529	5,048515	4,640187	3,451683	4,144333	4,026323	
14	1438_at	137,8453	129,8308			8,424885	7,854841	7,717155	7,327886	6,986994	6,466822	
15	1487_at	312,3179	283,2911			64,45306	30,46226	30,86355	62,84731	55,6546	56,59238	
16	1494_f_at	171,9447	170,783			5,777495	6,864321	8,37886	6,538007	6,793792	6,918854	
17	1552256_a	447,0718	435,8725			299,8175	62,11492	84,32298	352,1492	355,0533	342,612	

Φύλλο1 Φύλλο2 Φύλλο3

Εικόνα 24 : Βάση δεδομένων με το σύνολο των τιμών του λόγου αλλαγής της έκφρασης των γονιδίων σε όλα τα πειράματα.

Στη συνέχεια με το εργαλείο Gene ID Conversion από το DAVID (http://david.abcc.ncifcrf.gov) χρησιμοποιήθηκαν οι κωδικοί Affymetrix ID των γονιδίων για να μετατραπούν στα επίσημα ονόματα των γονιδίων (official gene symbol) και σε κωδικούς πρόσβασης Entrez. Επιπλέον δόθηκε και η σύντομη περιγραφή του γονιδίου (Target Description). Οι διαφορετικοί κωδικοί για τα γονίδια βοηθούν στην ταχύτερη εύρεση κάποιου γονιδίου αλλά και στη συνοπτική περιγραφή του. Κατόπιν χρησιμοποιήθηκαν οι βάσεις δεδομένων ISGs (Interferone Stimulated Genes http://www.lerner.ccf.org/labs/williams/), NF-KB.org και INTERFEROME (http://www.interferome.org) για να ταυτοποιηθούν και να ενοποιηθούν τα γονίδια που ρυθμίζονται από τον μεταγραφικό παράγοντα NF-kB, τις ιντερφερόνες και τα γονίδια που αυτές επάγουν. (Εικόνα 25).

Η INTERFEROME είναι μια βάση δεδομένων ανοικτής πρόσβασης η οποία περιέχει γονίδια που ρυθμίζονται από τύπου I, II και III ιντερφερόνες και η πληροφορία για αυτά συλλέχθηκε από την ανάλυση δεδομένων γονιδιακής έκφρασης από πειράματα διαφόρων κυτταρικών σειρών που είχαν εκτεθεί σε ιντερφερόνες. Αυτή η βάση δεδομένων ενοποιεί πληροφορίες από πειράματα υψηλής απόδοσης (High throughput), όπως μικροσυστοιχίες και πρωτεομικές τεχνικές. Αυτή η βάση δεδομένων περιέχει πληροφορίες γονιδιακής οντολογίας, αλληλούχησης ορθολόγων από 37 είδη, μοτίβα έκφρασης σε ιστούς και πληροφορίες ρύθμισης γονιδίων που επιτρέπουν τη μελέτη των μοριακών μηχανισμών της βιολογίας των ιντερφερονών (Samarajiwa SA et al. 2009).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1								gse6489	gse6489	gse6489	gse6489	gse11238	gse11238	gse11238	gse11238	gse11238	gse11238
2								primari pulmonary microvascular cells				HeLa					
3	ID	Target Des	Gene Sym	ENTREZ_G	ISGs	NF-KB	Interferom	gse6489A\	gse6489A\	gse6489FC	gse6489Ra	gse11238C	gse11238C	gse11238C	gse11238C	gse11238C	gse11238
4	1007_s_at	U48705 /F	DDR1	780				558,7861	491,2394			151,2182	26,94628	38,92084	139,2073	158,7641	135,3325
5	1053_at	M87338 /I	RFC2	5982			RFC2	505,3871	419,2027			121,2203	113,9391	89,13474	129,9194	138,9077	140,1831
6	117_at	X51757 /F	HSPA6	3310				109,0102	117,5057			7,800545	8,37825	10,64546	7,310707	6,026782	7,541456
7	121_at	X69699 /F	PAX8	7849		PAX8		766,8416	869,3055			27,34304	33,63189	39,83125	26,15478	23,99848	24,75253
8	1255_g_at	L36861 /FI	GUCA1A	2978				22,35779	22,78647			3,609142	4,399181	3,880913	3,923589	3,758987	3,795037
9	1294_at	L13852 /FI	UBA7	7318				383,046	469,4313			7,047644	8,656489	8,317356	6,290253	6,88483	6,95332
10	1316_at	X55005 /F	THRA	7067				91,92779	90,47927			7,221452	8,603186	7,634846	8,225595	7,566043	7,623318
11	1320_at	X79510 /F	PTPN21	11099				174,8971	149,2216			7,880062	8,85382	12,41737	7,332509	7,891158	8,013542
12	1405_i_at	M21121 /I	CCL5	6352		CCL5	CCL5	81,73426	78,77975			45,65032	6,150054	3,638611	38,43954	24,91343	23,66051
13	1431_at	J02843 /FI	CYP2E1	1571		CYP2E1		31,05442	33,34504			3,708529	5,048515	4,640187	3,451683	4,144333	4,026323
14	1438_at	X75208 /F	EPHB3	2049				137,8453	129,8308			8,424885	7,854841	7,717155	7,327886	6,986994	6,466822
15	1487_at	L38487 /FI	ESRRA	2101				312,3179	283,2911			64,45306	30,46226	30,86355	62,84731	55,6546	56,59238
16	1494_f_at	M33318 /I	CYP2A6	1548	CYP2A6			171,9447	170,783			5,777495	6,864321	8,37886	6,538007	6,793792	6,918854
17	1552256_a	gb:NM_00	SCARB1	949				447,0718	435,8725			299,8175	62,11492	84,32298	352,1492	355,0533	342,612
18	1552257_a	gb:NM_01	TTLL12	23170				841,1672	767,132			248,4514	160,3463	213,84	291,2847	238,0081	226,4326
19	1552258_a	gb:NM_05	NCRNA001	112597				105,154	110,5297			9,676585	44,9811	41,06792	7,012661	5,861793	6,475282
20	1552261_a	gb:NM_08	WFDC2	10406				79,27329	78,75282			8,437865	8,32576	6,502448	8,187069	8,448563	8,404275
21	1552263_a	gb:NM_13	MAPK1	5594	MAPK1			236,2179	228,6072			13,5369	8,939908	9,916417	11,91651	11,48924	10,46971
22	1552264_a	gb:NM_13	MAPK1	5594	MAPK1			1162,674	1097,538			111,6853	171,599	146,4351	98,67755	105,9096	111,8137
23	1552266_a	gb:NM_14	ADAM32	203102				65,17076	66,86799			4,328548	3,450626	4,04665	3,883501	3,690888	4,042487
24	1552269_a	gb:NM_13	SPATA17	128153				28,72057	28,98361			6,33635	5,226054	4,726866	7,379525	9,319135	8,236635
25	1552271_a	gb:NM_15	MGC2497	163154				102,5293	96,77116			11,09181	11,55707	14,51131	12,61073	12,65888	11,8116

Φύλλο1 Φύλλο2 Φύλλο3

Εικόνα 25: Βάση δεδομένων με πληροφορίες από την INTERFEROME,NF-kB.org και ISGs.

4.2 ΑΝΑΛΥΣΗ ChIP-seq ΔΕΔΟΜΕΝΩΝ

Για τη δημιουργία ενός δικτύου μεταγραφικών παραγόντων και των θέσεων ακετυλίωσης και μεθυλίωσης των ιστονών ήταν απαραίτητο να συλλεχτούν δεδομένα ChIP-seq πειραμάτων. Τα δεδομένα αυτά αντλήθηκαν από το ENCODE (The **En**cyclopedia **O**f **D**NA **E**lements). Το πρόγραμμα αυτό ξεκίνησε από το ινστιτούτο NHGRI (US National Human Genome Research Institute) το Σεπτέμβρη του 2003 και στόχος ήταν η αναγνώριση όλων των λειτουργικών στοιχείων του ανθρώπινου γονιδιώματος με τεχνικές αλληλούχησης επόμενης γενιάς (next generation sequencing). Το πρόγραμμα περιλαμβάνει μια ομάδα επιστημονικών ομάδων από όλο τον κόσμο. Τα αποτελέσματα των μελετών τους τοποθετούνται σε μια κοινή βάση δεδομένων (https://genome.ucsc.edu/ENCODE/) στην οποία έχει ελεύθερη πρόσβαση η επιστημονική κοινότητα. Για την εξαγωγή αυτών των αποτελεσμάτων χρησιμοποιήθηκαν τεχνικές υψηλής απόδοσης όπως ChIP-seq και RNA-seq. Πειράματα πραγματοποιήθηκαν σε 147 ανθρώπινες κυτταρικές σειρές και συνολικά μελετήθηκαν 119 μεταγραφικοί παράγοντες και 13 ιστονικές τροποποιήσεις (The ENCODE Project Consortium 2012).

Αυτή η βάση δεδομένων παρέχει μια πλατφόρμα (experimental matrix) για τη γρήγορη εύρεση των δεδομένων που μας ενδιαφέρουν. Μας παρέχεται η δυνατότητα να επιλέξουμε το είδος των αρχείων που θέλουμε να χρησιμοποιήσουμε fastq, BED κ.λπ. Επιλέξαμε να χρησιμοποιήσουμε τα BED αρχεία που περιέχουν τις συντεταγμένες των θέσεων πρόσδεσης των πρωτεϊνών στο DNA (peaks). Αυτά τα αρχεία δημιουργήθηκαν από επεξεργασία των αρχικών fastq αρχείων με τον αλγόριθμο BOWTIE για τη χαρτογράφηση των ‘‘διαβασμάτων’’ και στη συνέχεια με τον αλγόριθμο MACS για την εύρεση των θέσεων πρόσδεσης των πρωτεϊνών.

Τα BED αρχεία είναι αρχεία με δομή text και περιέχουν 3 υποχρεωτικές στήλες και 9 επιπλέον προαιρετικές στήλες. Η πρώτη στήλη περιέχει το χρωμόσωμα στο οποίο βρίσκονται οι θέσεις πρόσδεσης της πρωτεΐνης (π.χ. chr3). Η δεύτερη στήλη έχει έναν αριθμό που αντιπροσωπεύει την έναρξη της θέσης πρόσδεσης (peakstart) και η τρίτη στήλη περιέχει τον αριθμό που αντιστοιχεί στο τέλος της θέσης πρόσδεσης (peakEnd). Οι υπόλοιπες στήλες περιέχουν πληροφορίες σχετικά με την αλυσίδα του DNA που βρίσκεται η θέση πρόσδεσης (- ή +), μια τιμή που καθορίζει τη χρώση που θα έχει η κορυφή στον περιηγητή, τον αριθμό των κορυφών σε μια γραμμή του αρχείου κ.λπ. Η δομή ενός αρχείου BED φαίνεται παρακάτω:

```
 track name="ItemRGBDemo" description="Item RGB demonstration" visibility=2
itemRgb="On"
chr7  127471196 127472363 Pos1 0 + 127471196 127472363 255,0,0
chr7  127472363 127473530 Pos2 0 + 127472363 127473530 255,0,0
chr7  127473530 127474697 Pos3 0 + 127473530 127474697 255,0,0
chr7  127474697 127475864 Pos4 0 + 127474697 127475864 255,0,0
```

Οι κορυφές κατατάσσονται με βάση την απόλυτη τιμή του σήματος ή τη σημαντικότητα του εμπλουτισμού (Stephen L et al. 2014).

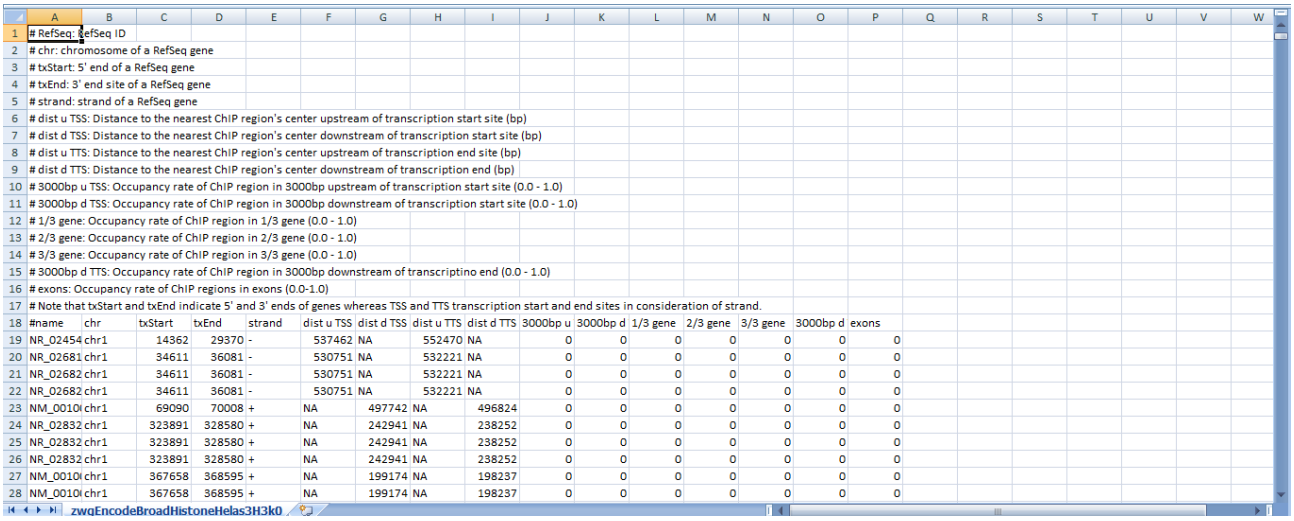
Για την περαιτέρω ανάλυση χρησιμοποιήθηκαν δεδομένα για 3 κυτταρικές σειρές για τις οποίες υπήρχαν και δεδομένα διαφορικής έκφρασης γονιδίων από πειράματα DNA μικροσυστοιχιών και RNA-seq. Για κύτταρα A546 χρησιμοποιήθηκαν τα BED αρχεία από 32 μεταγραφικούς παράγοντες, για κύτταρα HUVEC χρησιμοποιήθηκαν δεδομένα για 8 μεταγραφικούς παράγοντες και για κύτταρα HeLa χρησιμοποιήθηκαν δεδομένα για 58 μεταγραφικούς παράγοντες και 11 ιστονικές τροποποιήσεις.

Με τη χρήση του αλγορίθμου CEAS (Cis-regulatory element annotation system) εντοπίστηκαν οι θέσεις πρόσδεσης των μεταγραφικών παραγόντων και των ιστονικών τροποποιήσεων σε σχέση με τα γονίδια. Ο αλγόριθμος αυτός χρησιμοποιεί ένα αρχείο αναφοράς με τις συντεταγμένες των γονιδίων και βάση αυτού

συγκρίνει τις κορυφές του BED αρχείου για να υπολογίσει τις αποστάσεις τους από τα διάφορα γονίδια. Η εντολή που χρησιμοποιήθηκε στο τερματικό για τη λειτουργία του αλγορίθμου ήταν η εξής:

ceas –g hg19.refgene –b tf.bed
όπου –g hg19.refgene είναι το μονοπάτι για το αρχείο αναφοράς με τις συντεταγμένες των γονιδίων και
–b tf.bed το αρχείο BED από το ENCODE με τις συντεταγμένες των θέσεων πρόσδεσης της πρωτεΐνης.

Το αποτέλεσμα αυτού του αλγορίθμου είναι ένα αρχείο XLS με τις συντεταγμένες των γονιδίων και τις θέσεις πρόσδεσης των πρωτεϊνών σε σχέση με αυτά.
Τα αρχεία XLS που προέκυψαν από αυτή την ανάλυση (Εικόνα 26) είναι ταξινομημένα με βάση τους κωδικούς REFSEQ των γονιδίων.

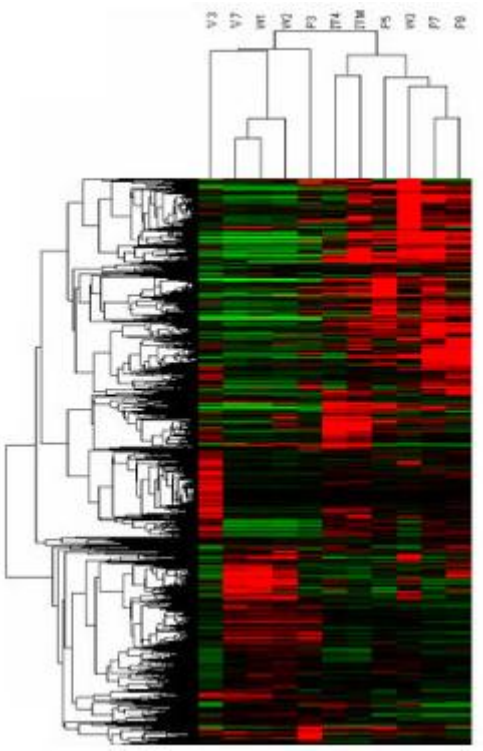


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	# RefSeq: RefSeq ID																						
2	# chr: chromosome of a RefSeq gene																						
3	# txStart: 5' end of a RefSeq gene																						
4	# txEnd: 3' end site of a RefSeq gene																						
5	# strand: strand of a RefSeq gene																						
6	# dist u TSS: Distance to the nearest ChIP region's center upstream of transcription start site (bp)																						
7	# dist d TSS: Distance to the nearest ChIP region's center downstream of transcription start site (bp)																						
8	# dist u TTS: Distance to the nearest ChIP region's center upstream of transcription end site (bp)																						
9	# dist d TTS: Distance to the nearest ChIP region's center downstream of transcription end (bp)																						
10	# 3000bp u TSS: Occupancy rate of ChIP region in 3000bp upstream of transcription start site (0.0 - 1.0)																						
11	# 3000bp d TSS: Occupancy rate of ChIP region in 3000bp downstream of transcription start site (0.0 - 1.0)																						
12	# 1/3 gene: Occupancy rate of ChIP region in 1/3 gene (0.0 - 1.0)																						
13	# 2/3 gene: Occupancy rate of ChIP region in 2/3 gene (0.0 - 1.0)																						
14	# 3/3 gene: Occupancy rate of ChIP region in 3/3 gene (0.0 - 1.0)																						
15	# 3000bp d TTS: Occupancy rate of ChIP region in 3000bp downstream of transcriptino end (0.0 - 1.0)																						
16	# exons: Occupancy rate of ChIP regions in exons (0.0-1.0)																						
17	# Note that txStart and txEnd indicate 5' and 3' ends of genes whereas TSS and TTS transcription start and end sites in consideration of strand.																						
18	#name	chr	txStart	txEnd	strand	dist u TSS	dist d TSS	dist u TTS	dist d TTS	3000bp u	3000bp d	1/3 gene	2/3 gene	3/3 gene	3000bp d	exons							
19	NR_02454	chr1	14362	29370	-	537462	NA	552470	NA	0	0	0	0	0	0	0							
20	NR_02681	chr1	34611	36081	-	530751	NA	532221	NA	0	0	0	0	0	0	0							
21	NR_02682	chr1	34611	36081	-	530751	NA	532221	NA	0	0	0	0	0	0	0							
22	NR_02682	chr1	34611	36081	-	530751	NA	532221	NA	0	0	0	0	0	0	0							
23	NM_0010(	chr1	69090	70008	+	NA	497742	NA	496824	0	0	0	0	0	0	0							
24	NR_02832	chr1	323891	328580	+	NA	242941	NA	238252	0	0	0	0	0	0	0							
25	NR_02832	chr1	323891	328580	+	NA	242941	NA	238252	0	0	0	0	0	0	0							
26	NR_02832	chr1	323891	328580	+	NA	242941	NA	238252	0	0	0	0	0	0	0							
27	NM_0010(	chr1	367658	368595	+	NA	199174	NA	198237	0	0	0	0	0	0	0							
28	NM_0010(	chr1	367658	368595	+	NA	199174	NA	198237	0	0	0	0	0	0	0							

zwgEncodeBroadHistoneHelas3H3k0

Εικόνα 26: Αρχείο XLS που παράγεται από τον αλγόριθμο CEAS το οποίο περιέχει τα γονίδια (REFSEQ κωδικοί αριστερά) καθώς και την απόσταση των θέσεων πρόσδεσης της πρωτεΐνης από τις λειτουργικές περιοχές τους.

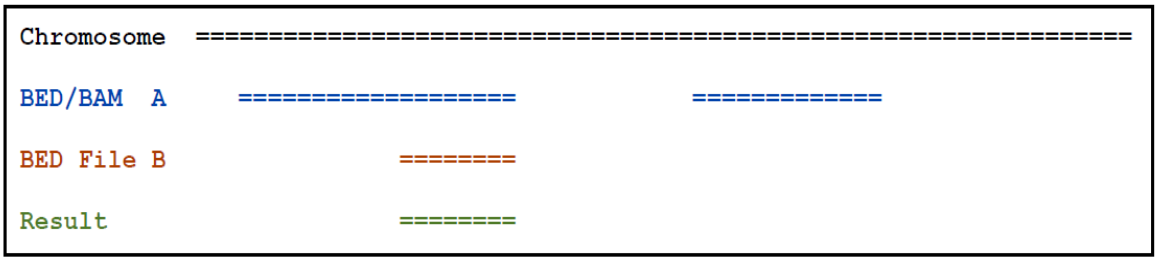
Χρησιμοποιώντας τη βάση δεδομένων DAVID οι κωδικοί πρόσβασης REFSEQ μετατράπηκαν σε επίσημα ονόματα των γονιδίων (official gene symbol). Τα XLS αρχεία όλων των μεταγραφικών παραγόντων ενώθηκαν χρησιμοποιώντας τους κωδικούς πρόσβασης REFSEQ των γονιδίων σαν στήλη αναφοράς και δημιουργήθηκε ένα κοινό αρχείο με όλους τους μεταγραφικούς παράγοντες προς μελέτη αλλά και των θέσεων πρόσδεσης τους στο DNA ανοδικά και καθοδικά του σημείου έναρξης της μεταγραφής (uTSS και dTSS) του κάθε γονιδίου για την κάθε κυτταρική σειρά που μελετήθηκε. Ακολούθως επιλέχθηκαν οι μεταγραφικοί παράγοντες που βρίσκονται σε απόσταση έως 5000bp (ζεύγη βάσεων) ανοδικά ή καθοδικά από το σημείο έναρξης της μεταγραφής. Για να επιλεχθούν οι συγκεκριμένες θέσεις χρησιμοποιήθηκε στο excel ο τύπος f(x)= if(B2≤5000;B2;"@"). Ο Πίνακας που δημιουργήθηκε χρησιμοποιήθηκε για την παρουσίαση των μεταγραφικών παραγόντων που έχουν θέσεις πρόσδεσης κοντά στο σημείο έναρξης της μεταγραφής των γονιδίων που εντοπίστηκαν να υπερεκφράζονται σε όλες τις κυτταρικές σειρές κατόπιν μόλυνσης με τη μορφή θερμικού χάρτη (heatmap). Ο θερμικός χάρτης δημιουργήθηκε με τη βοήθεια του προγράμματος T-MEV στο οποίο πραγματοποιήθηκε ιεραρχική ομαδοποίηση των δεδομένων χρησιμοποιώντας τον συντελεστή συσχέτισης PEARSON για τον υπολογισμό των αποστάσεων στο δενδρόγραμμα. Με αυτούς του χάρτες (Εικόνα 27) ομαδοποιήθηκαν κάποιοι μεταγραφικοί παράγοντες που προσδένονται στα ίδια γονίδια από αυτά που μελετήθηκαν.



Εικόνα 27: Ιεραρχική ταξινόμηση ''χωρίς επίβλεψη''(unsupervised hierarchical clustering) σε μορφή θερμικού χάρτη με την ύπαρξη πρόσδεσης των μεταγραφικών παραγόντων σε απόσταση +/-5000bp από το σημείο έναρξης της μεταγραφής για τα γονίδια που μελετήθηκαν.

Τέλος χρησιμοποιήθηκαν τα αρχεία BED με τις θέσεις πρόσδεσης των μεταγραφικών παραγόντων και τις ιστονικές τροποποιήσεις στη κυτταρική σειρά HeLa για να φτιαχτεί ένας χάρτης συσχέτισης της πρόσδεσης των μεταγραφικών παραγόντων και των ιστονικών τροποποιήσεων. Για να δημιουργηθεί αυτός ο χάρτης χρησιμοποιήθηκε η εντολή intersectBed των BEDtools.

Τα BEDtools (Quinlan AR et al. 2010) είναι ένα σύνολο εργαλείων που αναπτύχθηκαν για να παρέχουν τη δυνατότητα να συγκριθούν μεγάλα σετ γενωμικών δεδομένων. Για παράδειγμα η εντολή intersectBed ελέγχει εάν δυο γενωμικές περιοχές παρουσιάζουν επικάλυψη. Το αποτέλεσμα αυτή της εντολής είναι ένα αρχείο με τις θέσεις πρόσδεσης που παρουσιάζουν επικάλυψη όπως φαίνεται στην Εικόνα 28 και το παράδειγμα 1.



Εικόνα 28: Σχηματική απεικόνιση της λειτουργίας της εντολής intersectBed για τη σύγκριση γενωμικών περιοχών από δύο αρχεία BED.

```
$ cat A.bed
chr1    100     200
chr1    1000    2000

$ cat B.bed
chr1    150     250

$ intersectBed –a A.bed –b .bed
chr1    150     200
```

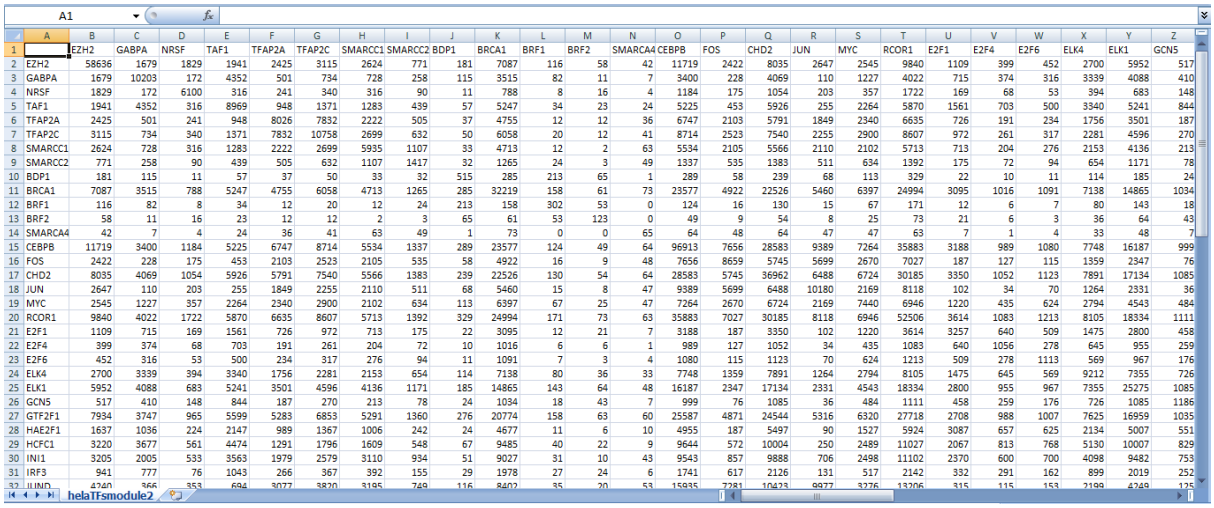
Παράδειγμα 1: Αναπαράσταση λειτουργίας της εντολής intersectBed για τη σύγκριση γενωμικών περιοχών από δύο αρχεία BED.

Η εντολή που χρησιμοποιήθηκε είναι:

intersectBed –a bed1.bed –b bed2.bed –wa –u > intersection.bed

όπου –a bed1.bed είναι το αρχείο BED του πρώτου μεταγραφικού παράγοντα
–b bed2.bed είναι το αρχείο BED του δεύτερου μεταγραφικού παράγοντα
–wa το αποτέλεσμα θα είναι η περιοχή του πρώτου αρχείου BED που παρατηρείται επικάλυψη
–u εξασφαλίζει ότι όταν εντοπιστούν πολλαπλά σημεία επικάλυψης διατηρείται μόνο μία περιοχή
intersection.bed το τελικό αρχείο που παράγεται και αποτελείται από τις κοινές περιοχές με σημεία πρόσδεσης των πρωτεϊνών.

Δημιουργήθηκε έτσι ένα αρχείο σε μορφή text που περιείχε 3 στήλες. Η πρώτη στήλη περιλαμβάνει το χρωμόσωμα όπου εντοπίστηκαν οι κοινές περιοχές πρόσδεσης, το σημείο έναρξης και το σημείο λήξης αυτών των περιοχών. Στη συνέχεια στο νέο αρχείο με το όνομα intersection.bed υπολογίστηκε ο αριθμός των γραμμών που περιείχε με την εντολή wc –l intersection.bed και αντιστοιχεί στον αριθμό των κοινών περιοχών των δύο BED αρχείων όπου υπάρχει πρόσδεση του μεταγραφικού παράγοντα. Με αυτές τις τιμές δημιουργήθηκε ένας πίνακας συσχέτισης της πρόσδεσης όλων των μεταγραφικών παραγόντων και των ιστονικών τροποποιήσεων (Εικόνα 29).

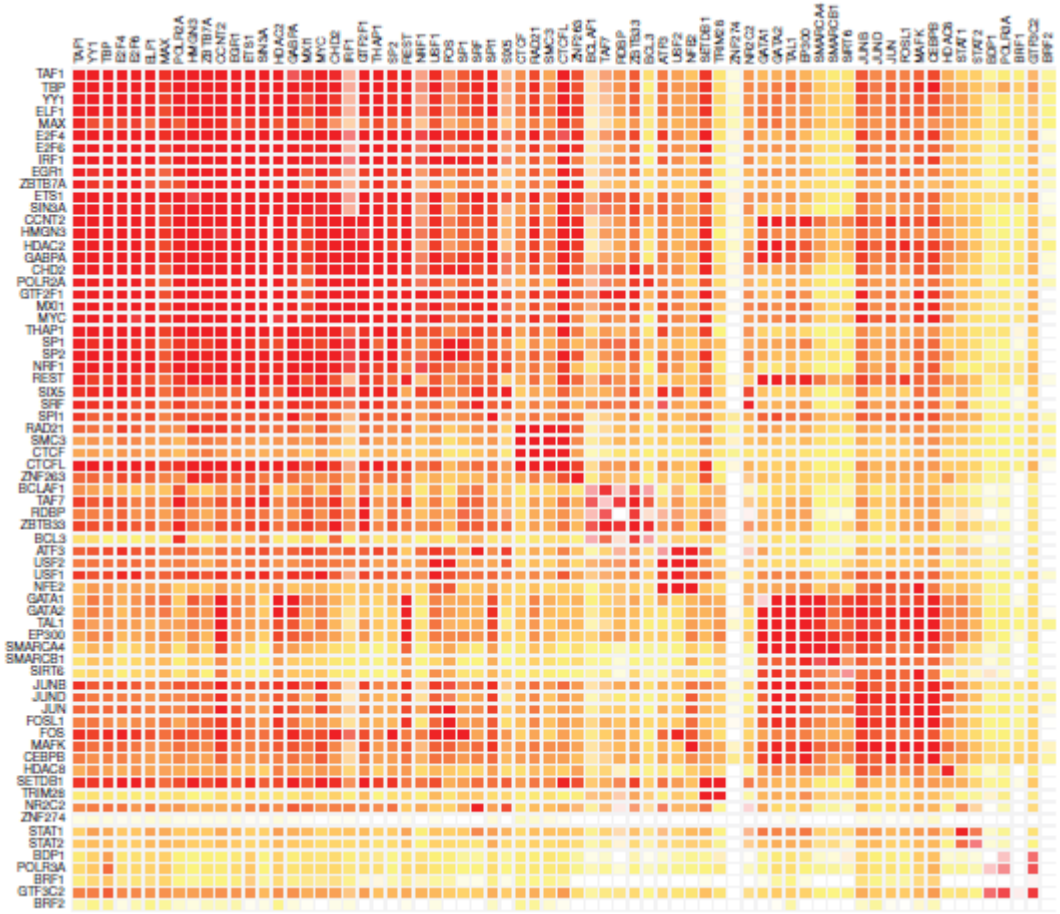


	EZH2	GABPA	NRSF	TAF1	TFAP2A	TFAP2C	SMARCC1	SMARCC2	BDP1	BRCA1	BRF1	BRF2	SMARCA4	CEBPB	FOS	CHD2	JUN	MYC	RCOR1	E2F1	E2F4	E2F6	ELK4	ELK1	GCN5
EZH2	58636	1679	1829	1941	2425	3115	2624	771	181	7087	116	58	42	11719	2422	8035	2647	2545	9840	1109	399	452	2700	5952	517
GABPA	1679	10203	172	4352	501	734	728	258	115	3515	82	11	7	3400	228	4069	110	1227	4022	715	374	316	3339	4088	410
NRSF	1829	172	6100	316	241	340	316	90	11	788	8	16	4	1184	175	1054	203	357	1722	169	68	53	394	683	148
TAF1	1941	4352	316	8969	948	1371	1283	439	57	5247	34	23	24	5225	453	5926	255	2264	5870	1561	703	500	3340	5241	844
TFAP2A	2425	501	241	948	8026	7832	2222	505	37	4755	12	12	36	6747	2103	5791	1849	2340	6635	726	191	234	1756	3501	187
TFAP2C	3115	734	340	1371	7832	10758	2699	632	50	6058	20	12	41	8714	2523	7540	2255	2900	8607	972	261	317	2281	4596	270
SMARCC1	2624	728	316	1283	2222	2699	5935	1107	33	4713	12	2	63	5534	2105	5566	2110	2102	5713	713	204	276	2153	4136	213
SMARCC2	771	258	90	439	505	632	1107	1417	32	1265	24	3	49	1337	535	1383	511	634	1392	175	72	94	654	1171	78
BDP1	181	115	11	57	37	50	33	32	515	285	213	65	1	289	58	239	68	113	329	22	10	11	114	185	24
BRCA1	7087	3515	788	5247	4755	6058	4713	1265	285	32219	158	61	73	23577	4922	22526	5460	6397	24994	3095	1016	1091	7138	14865	1034
BRF1	116	82	8	34	12	20	12	24	213	158	302	53	0	124	16	130	15	67	171	12	6	7	80	143	18
BRF2	58	11	16	23	12	12	2	3	65	61	53	123	0	49	9	54	8	25	73	21	6	3	36	64	43
SMARCA4	42	7	4	24	36	41	63	49	1	73	0	0	65	64	48	64	47	47	63	7	1	4	33	48	7
CEBPB	11719	3400	1184	5225	6747	8714	5534	1337	289	23577	124	49	64	96913	7656	28583	9389	7264	35883	3188	989	1080	7748	16187	999
FOS	2422	228	175	453	2103	2523	2105	535	58	4922	16	9	48	7656	8659	5745	5699	2670	7027	187	127	115	1359	2347	76
CHD2	8035	4069	1054	5926	5791	7540	5566	1383	239	22526	130	54	64	28583	5745	36962	6488	6724	30185	3350	1052	1123	7891	17134	1085
JUN	2647	110	203	255	1849	2255	2110	511	68	5460	15	8	47	9389	5699	6488	10180	2169	8118	102	34	70	1264	2331	36
MYC	2545	1227	357	2264	2340	2900	2102	634	113	6397	67	25	47	7264	2670	6724	2169	7440	6946	1220	435	624	2794	4543	484
RCOR1	9840	4022	1722	5870	6635	8607	5713	1392	329	24994	171	73	63	35883	7027	30185	8118	6946	52506	3614	1083	1213	8105	18334	1111
E2F1	1109	715	169	1561	726	972	713	175	22	3095	12	21	7	3188	187	3350	102	1220	3614	3257	640	509	1475	2800	458
E2F4	399	374	68	703	191	261	204	72	10	1016	6	6	1	989	127	1052	34	435	1083	640	1056	278	645	955	259
E2F6	452	316	53	500	234	317	276	94	11	1091	7	3	4	1080	115	1123	70	624	1213	509	278	1113	569	967	176
ELK4	2700	3339	394	3340	1756	2281	2153	654	114	7138	80	36	33	7748	1359	7891	1264	2794	8105	1475	645	569	9212	7355	726
ELK1	5952	4088	683	5241	3501	4596	4136	1171	185	14865	143	64	48	16187	2347	17134	2331	4543	18334	2800	955	967	7355	25275	1085
GCN5	517	410	148	844	187	270	213	78	24	1034	18	43	7	999	76	1085	36	484	1111	458	259	176	726	1085	1186
GTF2F1	7934	3747	965	5599	5283	6853	5291	1360	276	20774	158	63	60	25587	4871	24544	5316	6320	27718	2708	988	1007	7625	16959	1035
HAE2F1	1637	1036	224	2147	989	1367	1006	242	24	4677	11	6	10	4955	187	5497	90	1527	5924	3087	657	625	2134	5007	551
HCFC1	3220	3677	561	4474	1291	1796	1609	548	67	9485	40	22	9	9644	572	10004	250	2489	11027	2067	813	768	5130	10007	829
INI1	3205	2005	533	3563	1979	2579	3110	934	51	9027	31	10	43	9543	857	9888	706	2498	11102	2370	600	700	4098	9482	753
IRF3	941	777	76	1043	266	367	392	155	29	1978	27	24	6	1741	617	2126	131	517	2142	332	291	162	899	2019	252
JUND	4740	366	353	694	3077	3820	3195	749	116	8402	35	20	53	15935	7281	10423	9977	3276	13206	315	115	153	2199	4249	125

Εικόνα 29: Πίνακας συσχέτισης πρόσδεσης μεταγραφικών παραγόντων.

.

Ο πίνακας αυτός χρησιμοποιήθηκε για τη δημιουργία ενός θερμικού χάρτη στο πρόγραμμα T-MEV όπου αρχικά πραγματοποιήθηκε αφαίρεση της διαμέσου των τιμών της κάθε γραμμής από την κάθε επιμέρους τιμή της. Ύστερα πραγματοποιήθηκε ιεραρχική ομαδοποίηση των τιμών χρησιμοποιώντας τον συντελεστή συσχέτισης PEARSON για τον υπολογισμό των αποστάσεων στο δενδρόγραμμα.. Το αποτέλεσμα φαίνεται στην Εικόνα 30 με ομαδοποιημένους κάποιους μεταγραφικούς παράγοντες και ιστονικές τροποποιήσεις οι οποίοι συνεντοπίζονται στο γονιδίωμα.



Εικόνα 30: Ιεραρχική ταξινόμηση ‘’χωρίς επίβλεψη’’(unsupervised hierarchical clustering) σε μορφή θερμικού χάρτη των μεταγραφικών παραγόντων και των ιστονικών τροποποιήσεων όπου τα χρώματα αντιστοιχούν στο πλήθος των κοινών θέσεων πρόσδεσης στο DNA (με κόκκινο παρουσιάζονται οι μεταγραφικοί παράγοντες που έχουν περισσότερες κοινές θέσεις πρόσδεσης, με κίτρινο παρουσιάζονται οι μεταγραφικοί παράγοντες που έχουν οι λιγότερες κοινές θέσεις πρόσδεσης).

4.3 ΑΝΑΛΥΣΗ RNA-seq ΔΕΔΟΜΕΝΩΝ

Για τη συλλογή δεδομένων RNA-seq από ανθρώπινες κυτταρικές σειρές οι οποίες είχαν μολυνθεί με ιούς χρησιμοποιήθηκαν οι βάσεις δεδομένων GEODataSets και ArrayExpress του EMBL στις οποίες υπάρχουν τα δεδομένα από δημοσιευμένες εργασίες. Βρέθηκαν και χρησιμοποιήθηκαν τα δεδομένα από τις εργασίες με κωδικούς E-GEOD-38006 και E-MTAB-1277 όπου πραγματοποιήθηκαν ιικές μολύνσεις με HIV-1 και αδενοϊό σε CD4+ και HeLa κύτταρα αντίστοιχα. Τα δεδομένα βρίσκονται σε μορφή fastq και η ανάλυση πραγματοποιήθηκε σε υπολογιστή με λειτουργικό σύστημα MAC OS.

Αρχικά πραγματοποιήθηκε η χαρτογράφηση των ‘‘διαβασμάτων’’ σε ένα γονιδίωμα αναφοράς. Για την πραγματοποίηση αυτής της ανάλυσης χρησιμοποιήθηκε ο αλγόριθμος TOPHAT που χαρτογραφεί ταχύτατα DNA ‘‘διαβάσματα’’ και βασίζεται στον αλγόριθμο BOWTIE. Η εντολή που χρησιμοποιήθηκε στο τερματικό είναι η εξής:

tophat –r 80 –p 5 –m 1 referensegenome fastq1 fastq2 > newfile.sam

όπου –r 80 είναι η αναμενόμενη απόσταση μεταξύ των συζευγμένων “διαβασμάτων” (paired-end reads).
–p 5 είναι ο αριθμός των πυρήνων του υπολογιστή που χρησιμοποιήθηκαν για την ανάλυση.
–m 1 ο αριθμός των σφαλμάτων στις βάσεις που επιτρέπεται κατά τη χαρτογράφηση
referensegenome το γονιδίωμα αναφοράς.
fastq1 το μονοπάτι για το πρώτο αρχείο fastq με τα paired-end DNA ‘‘διαβάσματα’’.
fastq2 το μονοπάτι για το δεύτερο αρχείο fastq με τα paired-end DNA’‘διαβάσματα’’.
newfile.sam το αποτέλεσμα της χαρτογράφησης σε μορφή SAM η οποία είναι μια μορφή κειμένου που περιέχει πληροφορίες για τη θέση των αλληλουχιών σε σχέση με το γονιδίωμα αναφοράς.

Η χαρτογράφηση πραγματοποιήθηκε για κάθε αρχείο επαναλήψεων (replicates) ξεχωριστά. Το επόμενο βήμα στην ανάλυση ήταν ο υπολογισμός του αριθμού των ‘‘διαβασμάτων’’ που αντιστοιχούν σε κάθε γονίδιο και η κατασκευή πινάκων μετρήσεων με αυτές τις τιμές (counts tables). Για την πραγματοποίηση αυτής της ανάλυσης χρησιμοποιήθηκε ο αλγόριθμος HTSeq-count. Η εντολή που χρησιμοποιήθηκε ήταν:

htseq-count –s no samfile.sam gtffile.gtf > newfile.txt

όπου –s no χρησιμοποιήθηκε για να καθορίσει ότι τα δεδομένα δεν προέρχονται από μεθοδολογία ειδική προς μια αλυσίδα του DNA.
samfile.sam το αρχείο που προέκυψε από τη χαρτογράφηση των DNA ‘‘διαβασμάτων’’ με τον αλγόριθμο TOPHAT.
gtffile.gtf το αρχείο με τις γονιδιακές συντεταγμένες που χρησιμοποιήθηκε σαν αρχείο αναφοράς για να υπολογιστεί το σύνολο των ‘‘διαβασμάτων’’ που υπήρχαν σε κάθε γονίδιο.
newfile.txt το νέο αρχείο που δημιουργήθηκε από τον αλγόριθμο με τον πίνακα μετρήσεων.

Τα αρχεία με τους πίνακες μετρήσεων για τα αρχεία των πειραμάτων επανάληψης (replicates) ενοποιήθηκαν ώστε να προκύψει ένα κοινό αρχείο όπου περιέχονταν όλες οι τιμές από τα ξεχωριστά πειράματα παράλληλα σε διαφορετικές στήλες. Για την πραγματοποίηση αυτής της ανάλυσης χρησιμοποιήθηκε το πρόγραμμα multimergeRNA.R που είναι δομημένο στη γλώσσα προγραμματισμού R. Τοποθετήθηκε το σετ των αρχείων των επαναλήψεων κάθε πειράματος στον ίδιο φάκελο και χρησιμοποιήθηκε η εντολή multimergeRNA.R. Το αρχείο countstable.txt που δημιουργήθηκε περιείχε στην πρώτη του στήλη τα ονόματα των γονιδίων και στις επόμενες στήλες τον αριθμό των ‘‘διαβασμάτων’’ που αντιστοιχούν στα συγκεκριμένα γονίδια για κάθε πείραμα.

Κατόπιν χρησιμοποιήθηκε το πρόγραμμα DESeq για να βρεθούν τα γονίδια των οποίων η έκφραση αλλάζει μεταξύ του υγιούς και του μολυσμένου με ιό δείγματος.
Το πρόγραμμα είναι δομημένο στην υπολογιστική γλώσσα R. Τοποθετήθηκε το αρχείο countstable.txt σε ένα φάκελο και χρησιμοποιήθηκε η εντολή:

Rscript /NGS_Scripts/DESeqWrapper.R control,test

όπου

/NGS_Scripts/DESeqWrapper.R είναι το μονοπάτι για το πρόγραμμα
control, test καθορίζουν εάν τα πειράματα από τα οποία προέρχονται οι διαδοχικές στήλες του πίνακα περιέχουν δεδομένα από δείγματα μάρτυρα ή δείγματα μολυσμένα με ιό.

Τέλος για τη μετα-ανάλυση των δεδομένων πραγματοποιήθηκε ανάλυση της γονιδιακής οντολογίας στον δικτυακό τόπο DAVID και ελέγχθηκε ο ρόλος των γονιδίων σε διάφορες βιολογικές διεργασίες, σε μοριακές διαδικασίες και στα μεταβολικά μονοπάτια (KEGG Pathways) στα οποία συμμετέχουν .

Παράλληλα τα δεδομένα αναλύθηκαν με τη χρήση του αλγορίθμου Cufflinks για τον εντοπισμό της διαφορικής έκφρασης των γονιδίων. Για τη χαρτογράφηση των ‘‘διαβασμάτων’’ χρησιμοποιήθηκε η ίδια μεθοδολογία με τον αλγόριθμο TOPHAT που περιγράφηκε παραπάνω. Στη συνέχεια χρησιμοποιήθηκε το πρόγραμμα Cuffdiff του Cufflinks μέσω του οποίου εντοπίστηκαν τα διαφορικά εκφραζόμενα γονίδια κατόπιν μόλυνσης.

Η εντολή που χρησιμοποιήθηκε είναι:

cuffdiff –p 6 gtf.file control1,control2,control3 test1,test2,test3

όπου –p 6 είναι ο αριθμός των πυρήνων που χρησιμοποιήθηκαν για την ανάλυση(6)
gtf.file είναι το μονοπάτι για το αρχείο αναφοράς με τις συντεταγμένες των γονιδίων
control 1,2,3 τα αρχεία SAM για τα δείγματα μαρτύρων που χρησιμοποιήθηκαν (3 αρχεία replicates)
test1,2,3 τα αρχεία SAM για τα δείγματα με ιική μόλυνση που χρησιμοποιήθηκαν (3 αρχεία replicates)

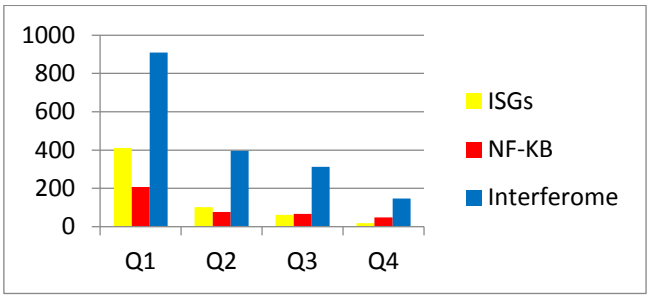
Τέλος στα δεδομένα πραγματοποιήθηκε ανάλυση της γονιδιακής οντολογίας στον δικτυακό τόπο DAVID.

ΑΠΟΤΕΛΕΣΜΑΤΑ

5.1 ΑΝΑΛΥΣΗ ΔΙΑΦΟΡΙΚΗΣ ΓΟΝΙΔΙΑΚΗΣ ΕΚΦΡΑΣΗΣ ΣΕ ΑΝΘΡΩΠΙΝΕΣ ΚΥΤΤΑΡΙΚΕΣ ΣΕΙΡΕΣ ΚΑΤΟΠΙΝ ΙΙΚΗΣ ΜΟΛΥΝΣΗΣ ΜΕ DNA ΜΙΚΡΟΣΥΣΤΟΙΧΙΕΣ

Από την ανάλυση των πειραμάτων DNA μικροσυστοιχιών ιικής μόλυνσης σε ανθρώπινες κυτταρικές σειρές δημιουργήθηκε μια βάση δεδομένων των γονιδίων των οποίων η έκφραση μεταβάλλεται στατιστικά σημαντικά. Στη βάση αυτή ταξινομήθηκαν τα γονίδια με βάση τη συχνότητα εμφάνισης τους στα διάφορα πειράματα.

Από τα 20.829 γονίδια που περιέχονται στη βάση δεδομένων τα 15496 εμφανίζουν αλλαγή στην έκφραση τους σε 1 έως 37 πειράματα που μελετήθηκαν. Τα γονίδια αυτά χωρίστηκαν σε 4 ομάδες με βάση τη συχνότητα εμφάνισής τους στα πειράματα. Για παράδειγμα στην πρώτη ομάδα (Q1) περιέχονται τα γονίδια που εμφανίζουν αλλαγή στην έκφρασή τους σε 8-37 πειράματα ενώ στην τελευταία ομάδα περιέχονται γονίδια που εμφανίζουν αλλαγή στην έκφρασή τους σε 1 ή 2 πειράματα. Αξιοποιώντας τις βάσεις δεδομένων ISGs, Nf-KB.org και INTERFEROME εντοπίστηκε ο αριθμός των γονιδίων των προηγούμενων ομάδων που ρυθμίζονται από τον Nf-KB και τις Ιντερφερόνες (Εικόνα 31). Στην Εικόνα 31 παρατηρούμε ότι τα γονίδια που επηρεάζονται συχνότερα από την μόλυνση σε σχέση με όσα επηρεάζονται λιγότερο συχνά εμφανίζουν μεγαλύτερη συσχέτιση με τον Nf-KB και τις ιντερφερόνες.



Εικόνα 31: Αριθμός γονιδίων των τεσσάρων ομάδων (τεταρτημορίων) που υπάρχουν στις βάσεις δεδομένων ISGs, NF-κB και Interferome.

Επιλέγθηκαν τα πρώτα 500 γονίδια τα οποία εμφανίζουν αλλαγή της έκφρασής τους στα περισσότερα πειράματα (13-37 πειράματα) και πραγματοποιήθηκε ανάλυση της γονιδιακής τους οντολογίας στον διαδικτυακό τόπο DAVID. Από αυτή την ανάλυση προέκυψε ότι τα γονίδια αυτά εμπλέκονται στις βιολογικές λειτουργίες της άμυνας του οργανισμού όπως η ανοσολογική απόκριση, η απόκριση σε ιό αλλά και η ρύθμιση της απόπτωσης και του κυτταρικού θανάτου. Στον Πίνακα 5 φαίνονται οι βιολογικές λειτουργίες στις οποίες εμπλέκονται τα 500 αυτά γονίδια.

Βιολογικές λειτουργίες για τα 500 πρώτα γονίδια
immune response
response to virus
defense response
response to organic substance
regulation of cell proliferation
response to wounding
regulation of apoptosis
regulation of programmed cell death
regulation of cell death
inflammatory response

Πίνακας 5: Οι 10 στατιστικά σημαντικότερες βιολογικές λειτουργίες στις οποίες συμμετέχουν τα 500 γονίδια που εμφανίζουν αλλαγή στην έκφρασή τους πιο συχνά κατόπιν μόλυνσης.

Επιπλέον πραγματοποιήθηκε ανάλυση των σηματοδοτικών μονοπατιών στα οποία συμμετέχουν αυτά τα γονίδια. Μερικά από αυτά είναι το σηματοδοτικό μονοπάτι των Toll υποδοχέων, των χημειοκινών και των JAK-STAT τα οποία είναι γνωστό ότι εμπλέκονται στην άμυνα του οργανισμού (Πίνακας 6).

Σηματοδοτικά μονοπάτια στα οποία συμμετέχουν τα 500 πρώτα γονίδια
Antigen processing and presentation
NOD-like receptor signaling pathway
Cytokine-cytokine receptor interaction
Toll-like receptor signaling pathway
Cytosolic DNA-sensing pathway
Chemokine signaling pathway
RIG-I-like receptor signaling pathway
Jak-STAT signaling pathway
Graft-versus-host disease
Allograft rejection

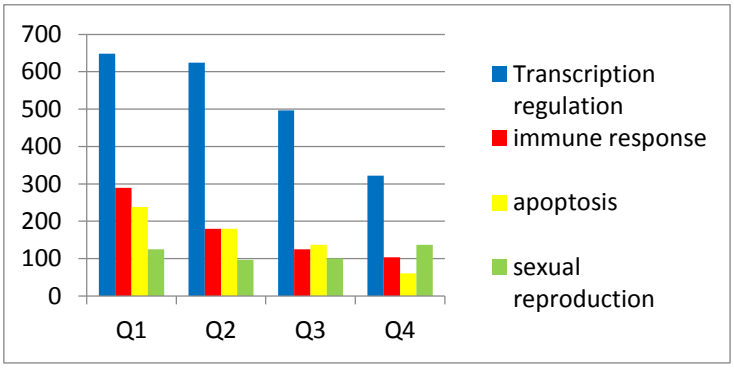
Πίνακας 6: Τα 10 στατιστικά σημαντικότερα σηματοδοτικά μονοπάτια στα οποία εντοπίζονται τα 500 γονίδια που εμφανίζουν αλλαγή στην έκφρασή τους πιο συχνά κατόπιν μόλυνσης.

Παράλληλα πραγματοποιήθηκε ανάλυση της γονιδιακής οντολογίας των 500 γονιδίων που εμφανίζουν αλλαγή στην έκφρασή τους λιγότερο συχνά για να διαπιστωθεί εάν σχετίζονται με αμυντικές λειτουργίες. Στον Πίνακα 7 παρουσιάζονται οι βιολογικές λειτουργίες στις οποίες συμμετέχουν τα 500 αυτά γονίδια και από όπου προκύπτει ότι δε σχετίζονται με την ανοσολογική απόκριση. Με τον τρόπο αυτό επιβεβαιώθηκε ότι η μεθοδολογία ανάλυσης των DNA μικροσυστοιχιών που εφαρμόστηκε είναι αξιόπιστη και η πλειοψηφία των γονιδίων που εμφανίζονται συχνότερα να αλλάζουν έκφραση σχετίζονται με ανοσολογικές αποκρίσεις.

Βιολογικές λειτουργίες στις οποίες συμμετέχουν τα 500 τελευταία γονίδια
sulfur compound biosynthetic process
sexual reproduction
proteolysis
glutathione biosynthetic process
transmembrane transport
spermatogenesis
male gamete generation
organic acid transport
peptide biosynthetic process
positive regulation of protein ubiquitination

Πίνακας 7: Οι 10 στατιστικά σημαντικότερες βιολογικές λειτουργίες στις οποίες συμμετέχουν τα 500 γονίδια που εμφανίζουν αλλαγή στην έκφρασή τους λιγότερο συχνά κατόπιν μόλυνσης.

Όπως και προηγουμένως, πραγματοποιήθηκε ανάλυση της κατανομής των γονιδίων σε 4 βιολογικές λειτουργίες, τη ρύθμιση της μεταγραφής, την ανοσολογική απόκριση, την απόπτωση και τη σεξουαλική αναπαραγωγή. Η τελευταία λειτουργία χρησιμοποιήθηκε σαν μάρτυρας και δε σχετίζεται με ανοσολογικές λειτουργίες. Όπως φαίνεται στην Εικόνα 32 το πρώτο τεταρτημόριο των γονιδίων περιέχει την πλειοψηφία των γονιδίων που συμμετέχουν στις 3 πρώτες βιολογικές λειτουργίες. Ο αριθμός των γονιδίων που συμμετέχουν σε αυτές τις λειτουργίες μειώνεται σταδιακά στα υπόλοιπα τεταρτημόρια. Δεν παρατηρείται αντίστοιχη αλλαγή για τη βιολογική λειτουργία της σεξουαλικής αναπαραγωγής όπου όλα τα τεταρτημόρια έχουν παρόμοιο αριθμό γονιδίων που εμπλέκονται σε αυτή τη λειτουργία. Στην πλειοψηφία των πειραμάτων ο αριθμός των γονιδίων που εμφανίζουν σημαντική αλλαγή στην έκφρασή τους κυμαίνεται από 500 έως 6000.



Εικόνα 32: Αριθμός γονιδίων της κάθε ομάδας τα οποία συμμετέχουν στις βιολογικές λειτουργίες της ρύθμισης της μεταγραφής, της ανοσολογικής απόκρισης, της απόπτωσης και της σεξουαλικής αναπαραγωγής.

Για τη μετα-ανάλυση των δεδομένων που συλλέχθηκαν εφαρμόστηκε η ίδια ακριβώς μεθοδολογία σε όλα τα πειράματα. Τα πρωταρχικά δεδομένα σε μορφή CEL κανονικοποιήθηκαν με τον αλγόριθμο RMA και στη συνέχεια πραγματοποιήθηκε t-test με τιμή p-value $p \leq 0.05$ ως κατώφλι σημαντικότητας για να καθοριστούν τα στατιστικώς σημαντικά γονίδια. Στη συνέχεια υπολογίστηκε ο λόγος αλλαγής της έκφρασης των γονιδίων στις κυτταρικές σειρές που είχαν δεχθεί την ιική επίδραση σε σχέση με τα κύτταρα μάρτυρες και επιλέχθηκαν τα γονίδια των οποίων η έκφραση αλλάζει περισσότερο από 1,5 φορές.

Αναλυτικά παρακάτω παρουσιάζονται τα αποτελέσματα για κάθε μελέτη ξεχωριστά:

Από την ανάλυση DNA μικροσυστοιχιών που πραγματοποιήθηκαν σε λευκοκύτταρα μονοκύτταρα τα οποία μολύνθηκαν με το βακτήριο Francisella tularensis (Butchar JP et al. 2008) συνολικά εντοπίστηκαν να αλλάζουν έκφραση 4313 γονίδια εκ των οποίων τα 1490 εμφανίζουν αύξηση της έκφρασής τους περισσότερο από 1,5 φορές κατόπιν της μόλυνσης. Σε αυτά πραγματοποιήθηκε ανάλυση της γονιδιακής οντολογίας και οι σημαντικότερες βιολογικές λειτουργίες στις οποίες συμμετέχουν τα γονίδια αυτά είναι η ανοσολογική και αμυντική απόκριση, η ρύθμιση του κυτταρικού θανάτου και η απόπτωση (Πίνακας 8). Γονίδια τα οποία εμφανίζουν τη μεγαλύτερη αύξηση στην έκφρασή τους είναι οι ιντερφερόνες IFNA1, IFNA13, IFNW1, οι ιντερλευκίνες IL1B, IL6, IL8, IL10 και οι χημειοκίνες CCL5, CXCL1, CCL4, CCL3, CCL20, CCL18 γονίδια τα οποία συμμετέχουν στην ανοσολογική απόκριση. Επίσης σημαντική αύξηση στην έκφρασή τους παρουσίασαν τα γονίδια CASP8, CASP4, CASP5, FAS και TNF τα οποία συμμετέχουν στον κυτταρικό θάνατο και την απόπτωση.

Βιολογικές λειτουργίες
immune response
regulation of apoptosis
regulation of programmed cell death
regulation of cell death
cell adhesion
defense response
cell morphogenesis
cell-cell adhesion
regulation of cell activation
positive regulation of biosynthetic process
apoptosis

Πίνακας 8: Βιολογικές λειτουργίες στις οποίες συμμετέχουν τα 1490 γονίδια των οποίων η έκφραση αυξάνεται, κατόπιν μόλυνσης ανθρώπινων λευκοκύτταρων μονοκύτταρων με το βακτήριο F. tularensis.

Από την ανάλυση της διαφορικής γονιδιακής έκφρασης με DNA μικροσυστοιχίες σε ανθρώπινα δενδριτικά κύτταρα κατόπιν βακτηριακής μόλυνσης με το βακτήριο Chlamydia pneumonia (Njau F et al. 2009) εντοπίστηκαν 604 γονίδια να εμφανίζουν αλλαγή στην έκφρασή τους με τα 380 από αυτά να υπερεκφράζονται. Τα γονίδια που εμφανίζουν τη μεγαλύτερη αύξηση στην έκφρασή τους είναι γονίδια 2-5A ολιγοαδενυλοσυνθετασών OAS1, OAS2, OAS3, OASL, χημειοκίνες όπως οι CCL5, CCL19, CXCL9, CXCL10, οι ιντερφερόνες IL6, IL18, IL12B και ο NFKB2. Από την ανάλυση της γονιδιακής οντολογίας (Πίνακας 9) φαίνεται ότι οι σημαντικότερες βιολογικές λειτουργίες στις οποίες συμμετέχουν τα γονίδια που υπερεκφράζονται είναι η ανοσολογική και αμυντική απόκριση και η ρύθμιση της απόπτωσης.

Βιολογικές λειτουργίες
immune response
regulation of apoptosis
regulation of programmed cell death
regulation of cell death
positive regulation of immune system process
defense response
regulation of lymphocyte activation
regulation of leukocyte activation
regulation of cell activation
positive regulation of leukocyte activation
apoptosis

Πίνακας 9: Βιολογικές λειτουργίες στις οποίες συμμετέχουν τα 380 γονίδια των οποίων η έκφραση αυξάνεται κατόπιν μόλυνσης ανθρώπινων δενδριτικών κυττάρων με το βακτήριο C. pneumonia.

Ο ιός της γρίπης έχει χρησιμοποιηθεί για την ανάλυση της γονιδιακής έκφρασης σε μολυσμένες κυτταρικές σειρές. Από την ανάλυση DNA μικροσυστοιχιών που πραγματοποιήθηκαν σε κύτταρα HUVEC που μολύνθηκαν με τα στελέχη H5N1, PR8 και FPV του ιού της γρίπης (Schmolke M et al. 2009), εντοπίστηκαν συνολικά 3965 γονίδια να εμφανίζουν στατιστικά σημαντική αλλαγή στην έκφρασή τους κατόπιν της μόλυνσης με το στέλεχος H5N1. Η μόλυνση με τα στελέχη FPV και PR8 προκαλεί αλλαγή στην έκφραση 3980 και 2662 γονιδίων αντίστοιχα.

Από τα γονίδια που υπερεκφράζονται 191 παρατηρείται να έχουν σημαντική αύξηση στην έκφρασή τους και στις τρείς διαφορετικές μολύνσεις. Από την ανάλυση της γονιδιακής οντολογίας εντοπίστηκαν οι βιολογικές λειτουργίες στις οποίες συμμετέχουν όπως η απόκριση σε ιό, η ανοσολογική και αμυντική απόκριση, η απόπτωση και η απόκριση σε φλεγμονή (Πίνακας 10). Γονίδια όπως η IFNB1, CCL5, IRF1, IRF7, ISG20, IFIT3, CXCL10, CXCL1, MX1, OAS1, STAT1 εμφανίζουν την μεγαλύτερη αύξηση στην έκφρασή τους. Συγκριτικά μεταξύ των διαφορετικών μολύνσεων το στέλεχος H5N1 προκαλεί σε μεγαλύτερο βαθμό αύξηση της έκφρασης των κοινών υπερεκφραζόμενων γονιδίων κατόπιν της ιικής μόλυνσης καθώς είναι το πιο παθογόνο στέλεχος από τα άλλα δύο.

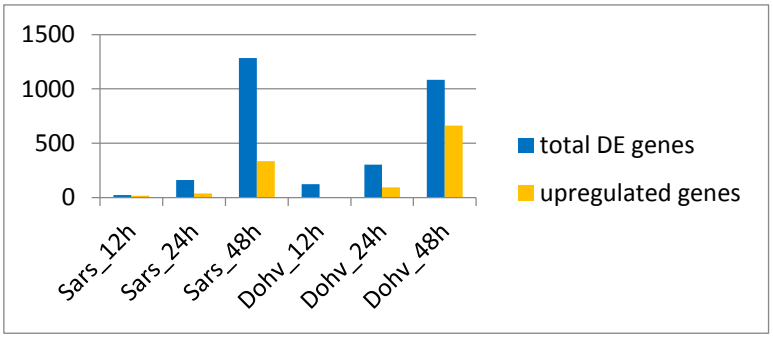
Βιολογικές λειτουργίες
response to virus
immune response
defense response
regulation of apoptosis
regulation of programmed cell death
regulation of cell death
apoptosis
cell death
programmed cell death
death
inflammatory response

Πίνακας 10: Οι 11 στατιστικά σημαντικότερες βιολογικές λειτουργίες στις οποίες συμμετέχουν τα 191 γονίδια των οποίων η έκφραση αυξάνεται κατόπιν μόλυνσης των κυττάρων και με τα διαφορετικά ιικά στελέχη.

Σε μια άλλη μελέτη μολύνθηκαν ανθρώπινα βρογχικά επιθηλιακά κύτταρα με τους ιούς Sars και Dhori και μελετήθηκε η αλλαγή της γονιδιακής έκφρασης με DNA μικροσυστοιχίες σε χρονικά διαστήματα 12, 24 και 48 ώρες κατόπιν της ιικής μόλυνσης (Yoshikawa T et al. 2010). Στον Πίνακα 11 και την Εικόνα 33 φαίνεται ο αριθμός του συνόλου των γονιδίων που αλλάζουν έκφραση και ο αριθμός των γονιδίων που υπερεκφράζονται σε κάθε χρονικό διάστημα σε σχέση με τα δείγματα που δεν μολύνθηκαν με τους ιούς.

	Sars_12h	Sars_24h	Sars_48h	Dohv_12h	Dohv_24h	Dohv_48h
total DE genes	23	160	1284	123	302	1083
upregulated genes	17	36	334	3	93	663

Πίνακας 11: Αριθμός γονιδίων των οποίων η έκφραση αλλάζει και όσων υπερεκφράζονται στην πορεία του χρόνου (0-48 ώρες) κατόπιν μόλυνσης με τους ιούς Sars και Dohv.



Εικόνα 33: Αριθμός γονιδίων των οποίων η έκφραση αλλάζει και όσων υπερεκφράζονται στην πορεία του χρόνου (0-48 ώρες) κατόπιν μόλυνσης με τους ιούς Sars και Dohv.

Τα γονίδια των οποίων η έκφραση αυξάνεται σημαντικά κατόπιν της μόλυνσης συμμετέχουν σε λειτουργίες όπως η ανοσολογική απόκριση, η απόκριση σε ιούς, η απόκριση σε φλεγμονή καθώς επίσης και στη ρύθμιση της απόπτωσης σύμφωνα με την ανάλυση της γονιδιακής τους οντολογίας. Γονίδια τα οποία εμφανίζουν τη μεγαλύτερη αύξηση στην έκφρασή τους είναι ο FAS, NFKBIA η κασπάση CASP1, ιντερφερόνες (IFNB1, IFNA1), πρωτεΐνες επαγόμενες από ιντερφερόνες (IFIT2, IFIH1, IFIT2, IFIT3, IFI44L), διάφορες χημειοκίνες (CXCL10, CCL5), πρωτείνες της οικογένειας των 2-5A συνθετασών (OASL, OAS1, OAS2), ιντερλευκίνες (IL2A, IL29, IL7) και γονίδια ρύθμισης των ιντερφερονών (IRF1, IRF2, IRF7 και IRF9).

Ανθρώπινα ερυθροειδή προγονικά κύτταρα (EPCs) μολύνθηκαν με τον ιό B19V και μελετήθηκαν οι αλλαγές στη γονιδιακή έκφραση 12, 24 και 48 ώρες κατόπιν της μόλυνσης (Wan Z et al. 2010) αναλύθηκε η διαφορική γονιδιακή έκφραση με DNA μικροσυστοιχίες. Στις 48 ώρες εμφανίζεται να αλλάζει η έκφραση στα περισσότερα γονίδια. Συνολικά 1359 γονίδια εμφανίζουν μεταβολή στην έκφραση τους μεγαλύτερη από 1,5 φορές και 795 από αυτά υπερεκφράζονται. Οι βιολογικές λειτουργίες στις οποίες συμμετέχουν αυτά τα γονίδια είναι ο κυτταρικός κύκλος, η μίτωση , η οργάνωση του κυτταροσκελετού και η απόκριση της έμφυτης ανοσίας (Πίνακας 12). Γονίδια όπως ο CPM, JAC, APOLD1, CAV, IL1R1, C5, IL13A εμφανίζουν την μεγαλύτερη αύξηση στην έκφρασή τους και τα δεδομένα αυτά υποστηρίζουν ότι ο ιός επάγει την κυτταρική διαίρεση των ερυθροειδών προγονικών κυττάρων.

Βιολογικές λειτουργίες
cell cycle
cytoskeleton organization
cell cycle phase
cell cycle process
mitotic cell cycle
nuclear division
mitosis
protein-DNA complex assembly
innate immune response
nucleosome assembly

Πίνακας 12: Βιολογικές λειτουργίες στις οποίες συμμετέχουν τα 795 υπερεκφραζόμενα γονίδια κατόπιν μόλυνσης EPC κυττάρων με τον ιό B19V.

Σε δύο μελέτες αναλύθηκαν οι μεταβολές της γονιδιακής έκφρασης σε ηπατικά κύτταρα και καρκινικά ηπατικά κύτταρα Huh7 κατόπιν μόλυνσης με τον ιό της ηπατίτιδας C (Blackham S et al. 2010, Li Q et al. 2013). Σε αυτά τα πειράματα ανιχνεύθηκαν 540 και 249 γονίδια να υπερεκφράζονται αντίστοιχα κατόπιν της μόλυνσης. Οι βιολογικές λειτουργίες στις οποίες συμμετέχουν τα κοινά γονίδια των οποίων η έκφραση αυξάνεται κατόπιν μόλυνσης με τον ιό της ηπατίτιδας C και στις δύο μελέτες είναι η απόκριση στη φλεγμονή, η ρύθμιση του μεταβολισμού των λιπιδίων και η απόκριση στις τοξίνες (Πίνακας 13). Γονίδια τα οποία υπερεκφράζονται σε μεγαλύτερο ποσοστό και στα δύο πειράματα είναι το FOS, SORBS1, LGALS2, MST1, MBL2, SULT1C2.

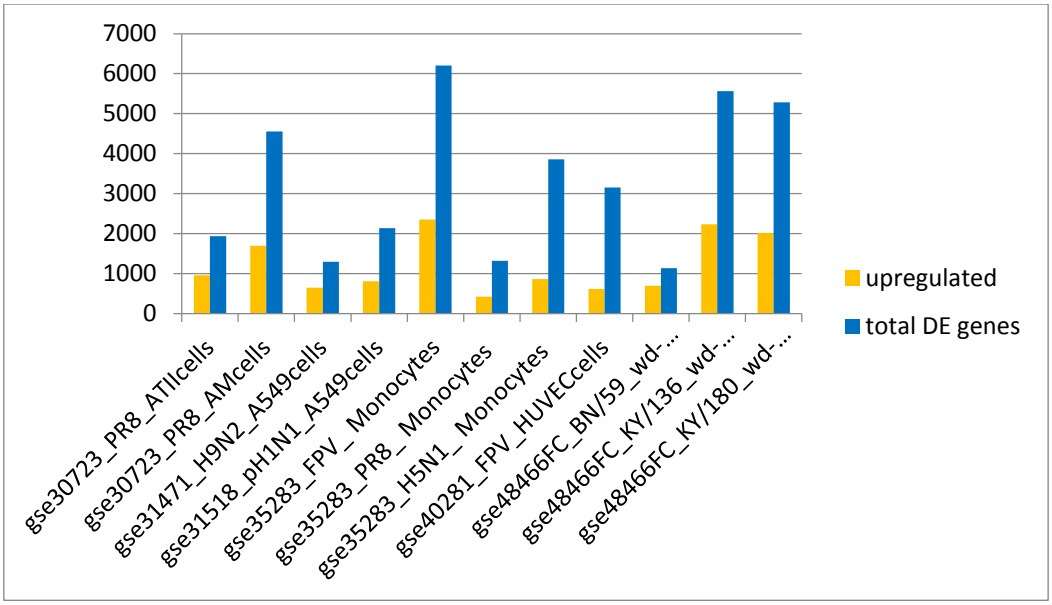
Βιολογικές λειτουργίες
regulation of lipid metabolic process
enzyme linked receptor protein signaling pathway
positive regulation of lipid metabolic process
regulation of catabolic process
regulation of glycogen biosynthetic process
response to wounding
regulation of glucan biosynthetic process
regulation of polysaccharide biosynthetic process
regulation of polysaccharide metabolic process
response to toxin

Πίνακας 13: Βιολογικές λειτουργίες στις οποίες συμμετέχουν τα κοινά υπερεκφραζόμενα γονίδια κατόπιν μόλυνσης ηπατικών κυττάρων με τον ιό της ηπατιτίδας C.

Σε αρκετές μελέτες χρησιμοποιήθηκαν στελέχη του ιού της γρίπης για να μολυνθούν ανθρώπινες κυτταρικές σειρές και να εντοπιστεί η μεταβολή της γονιδιακής έκφρασης (Viemann D et al. 2011, Wang j et al. 2012, Sutejo R et al. 2012, Gerlach RL et al. 2013 και Börgeling Y et al. 2014). Στον Πίνακα 14 και την Εικόνα 34 παρουσιάζεται ο αριθμός των γονιδίων των οποίων αλλάζει η έκφραση και όσων υπερεκφράζονται.

	upregulated	total DE genes
gse30723_PR8_ATIIcells	957	1940
gse30723_PR8_AMcells	1694	4558
gse31471_H9N2_A549cells	651	1297
gse31518_pH1N1_A549cells	811	2137
gse35283_FPV_ Monocytes	2353	6206
gse35283_PR8_ Monocytes	422	1318
gse35283_H5N1_ Monocytes	864	3862
gse40281_FPV_HUVECcells	615	3158
gse48466FC_BN/59_wd-NHBEcells	697	1138
gse48466FC_KY/136_wd-NHBEcells	2231	5562
gse48466FC_KY/180_wd-NHBEcells	2021	5286

Πίνακας 14: Αριθμός γονιδίων των οποίων η έκφραση αλλάζει και όσων υπερεκφράζονται στα διαφορετικά πειράματα των ιικών μολύνσεων στις ανθρώπινες κυτταρικές σειρές.



Εικόνα 34: Αριθμός γονιδίων των οποίων η έκφραση αλλάζει και όσων υπερεκφράζονται στα διαφορετικά πειράματα των ικών μολύνσεων στις ανθρώπινες κυτταρικές σειρές.

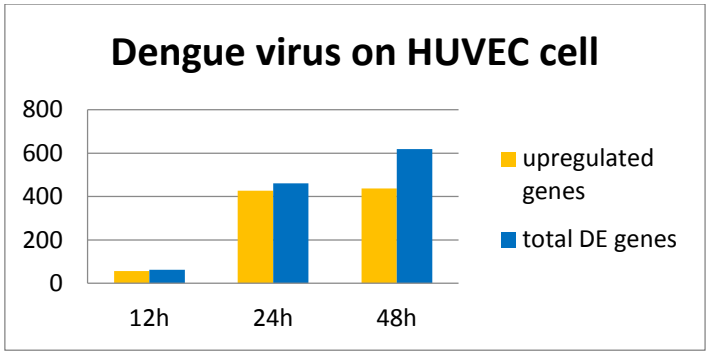
Κατόπιν πραγματοποιήθηκε ανάλυση της γονιδιακής οντολογίας αυτών των γονιδίων στο δικτυακό τόπο DAVID. Εντοπίστηκαν έτσι οι κυριότερες βιολογικές λειτουργίες στις οποίες συμμετέχουν και οι οποίες παρουσιάζονται στον Πίνακα 15.

Βιολογικές λειτουργίες
response to virus
immune response
regulation of I-kappaB kinase/NF-kappaB cascade
positive regulation of I-kappaB kinase/NF-kappaB cascade
regulation of protein kinase cascade
positive regulation of protein kinase cascade
defense response
regulation of apoptosis
regulation of programmed cell death
regulation of cell death

Πίνακας 15: Βιολογικές λειτουργίες στις οποίες συμμετέχουν τα κοινά υπερεκφραζόμενα γονίδια.

Οι σημαντικότερες λειτουργίες σχετίζονται με την άμυνα του οργανισμού, τον κυτταρικό θάνατο και την απόπτωση. Κοινά γονίδια τα οποία εμφανίζουν τη μεγαλύτερη αύξηση στην έκφρασή τους είναι τα IFNB1, DDX58, ISGL5, CCL5, IFNB1, IRF7, IRF9, ISG20, IL6, MX2, IL28A, STAT1, STAT2, ILR3, IRIM5, CASP1, CASP8, FAS, BCL12, JUN, JAK2.

Μελετήθηκαν οι αλλαγές στη γονιδιακή έκφραση σε ανθρώπινα ενδοθηλιακά κύτταρα HUVEC που μολύνθηκαν με τον δάγκειο ιό (Dalrymple NA et al. 2012) 12, 24 και 48 ώρες κατόπιν της μόλυνσης. Στην Εικόνα 35 παρουσιάζεται ο αριθμός των γονιδίων που εντοπίστηκαν να αλλάζουν έκφραση και όσων υπερεκφράζονται.



Εικόνα 35: Αριθμός γονιδίων των οποίων η έκφραση μεταβάλλεται και όσων υπερεκφράζονται 12, 24 και 48 ώρες κατόπιν μόλυνσης κυττάρων HUVEC με τον δάγκειο ιό.

Στους Πίνακες 16-18 φαίνονται οι βιολογικές λειτουργίες στις οποίες συμμετέχουν τα γονίδια που υπερεκφράζονται στα τρία χρονικά διαστήματα που πραγματοποιήθηκαν οι DNA μικροσυστοιχίες.

12h post infection	gene number
response to virus	8
immune response	11
blood vessel development	5
vasculature development	5
response to nutrient	4
response to vitamin	3
skeletal system development	5
negative regulation of gene expression	6
response to steroid hormone stimulus	4
response to nutrient levels	4

Πίνακας 16: Βιολογικές λειτουργίες στις οποίες συμμετέχουν τα υπερεκφραζόμενα γονίδια 12 ώρες κατόπιν της ιικής μόλυνσης.

24h post infection	gene number
immune response	74
response to virus	31
defense response	55
inflammatory response	32
positive regulation of response to stimulus	27
response to molecule of bacterial origin	17
response to cytokine stimulus	16
positive regulation of immune system process	26
positive regulation of defense response	15
response to bacterium	22
response to organic substance	45
positive regulation of immune response	19

Πίνακας 17: Βιολογικές λειτουργίες στις οποίες συμμετέχουν τα υπερεκφραζόμενα γονίδια 24 ώρες κατόπιν της ιικής μόλυνσης.

48h post infection	gene number
immune response	90
response to virus	29
defense response	62
antigen processing and presentation of peptide antigen via MHC class I	14
antigen processing and presentation of peptide antigen	15
antigen processing and presentation	20
innate immune response	22
inflammatory response	33
positive regulation of immune system process	28
response to cytokine stimulus	17
positive regulation of response to stimulus	27
response to wounding	41

Πίνακας 18: Βιολογικές λειτουργίες στις οποίες συμμετέχουν τα υπερεκφραζόμενα γονίδια 48 ώρες κατόπιν της ιικής μόλυνσης.

Οι κύριες βιολογικές λειτουργίες στις οποίες συμμετέχουν είναι η ανοσολογική απόκριση, η απόκριση σε ιό και η αμυντική απόκριση. Τα γονίδια IFNB1, IFIT1, OAS1, OAS2, CCL5, IFIT2, IFIT3, IF16, ISG20, KL4, MX1 εμφανίζουν τη μεγαλύτερη αύξηση στην έκφρασή τους σε όλα τα χρονικά διαστήματα ύστερα από την ιική μόλυνση και αποτελούν κεντρικά γονίδια της ανοσολογικής απόκρισης.

Σε μια κλινική μελέτη μελετήθηκαν οι αλλαγές της γονιδιακής έκφρασης σε ήπαρ από άτομα μολυσμένα με τον ιό της ηπατίτιδας B (Nissim O et al. 2012) τα οποία υποβλήθηκαν σε μεταμόσχευση ήπατος και συγκρίθηκαν με ήπαρ από υγιή άτομα. Εντοπίστηκαν 1177 γονίδια τα οποία υπερεκφράζονται και από την ανάλυση της γονιδιακής οντολογίας εντοπίστηκαν οι βιολογικές λειτουργίες στις οποίες συμμετέχουν. Στον Πίνακα 19 παρουσιάζονται οι 10 σημαντικότερες από αυτές τις βιολογικές λειτουργίες. Τη μεγαλύτερη αλλαγή στην έκφραση τους παρουσιάζουν τα γονίδια CCL5, LCT, ICK, ARSF, ELK1, CTF8, EGR2, IL1RAP, TAF1, OAF, FADD, SOX4.

Βιολογικές λειτουργίες
negative regulation of transport
ectoderm development
muscle system process
muscle contraction
negative regulation of transporter activity
cation transport
metal ion transport
di-, tri-valent inorganic cation transport
behavior
vesicle-mediated transport

Πίνακας 19: 10 σημαντικότερες βιολογικές λειτουργίες στις οποίες συμμετέχουν τα γονίδια των οποίων η έκφραση αυξάνεται στο ήπαρ ασθενών με ηπατίτιδα Β.

Τέλος σε μια μελέτη που πραγματοποιήθηκε στο εργαστήριό μας μολύνθηκαν HeLa κύτταρα με τον ιό Sendai (Antonaki A et al. 2011) και πραγματοποιήθηκαν DNA μικροσυστοιχίες για τον εντοπισμό των γονιδίων των οποίων η έκφραση μεταβάλλεται. Από την ανάλυση της γονιδιακής έκφρασης εντοπίστηκαν 1267 γονίδια να υπερεκφράζονται κατόπιν της μόλυνσης. Στα γονίδια αυτά πραγματοποιήθηκε ανάλυση της γονιδιακής τους οντολογίας και στον Πίνακα 20 παρουσιάζονται οι 10 σημαντικότερες λειτουργίες στις οποίες συμμετέχουν. Τέτοιες είναι η απόκριση σε ιό, η ρύθμιση του κυτταρικού θανάτου, η ανοσολογική απόκριση, η σηματοδότηση των υποδοχέων toll και η απόπτωση. Τα γονίδια που εμφανίζουν την

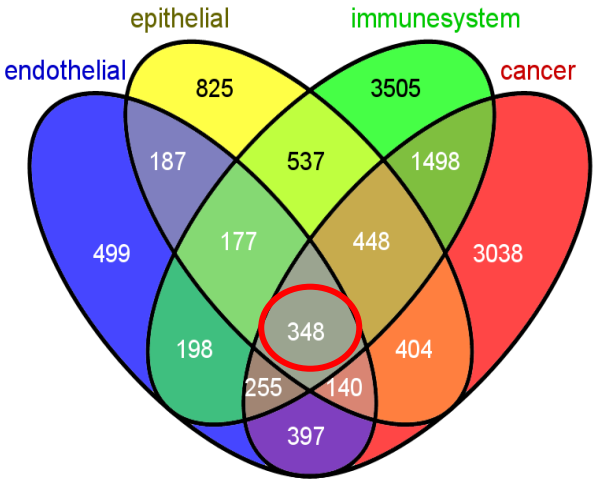
μεγαλύτερη αύξηση στην έκφρασή τους είναι ο CCL5, IFNB1, IFIT1, IFIT2, IFIT3, IFIH1, OAS1, OAS3, OASL, MX1, CXCL11, CXCL10, ISG20, IL7R, IRF1, IRF7, IRF9.

Βιολογικές λειτουργίες
response to virus
ribosome biogenesis
positive regulation of defense response
ribonucleoprotein complex biogenesis
positive regulation of innate immune response
positive regulation of response to stimulus
regulation of innate immune response
regulation of programmed cell death
toll-like receptor signaling pathway
regulation of cell death

Πίνακας 20: Οι 10 στατιστικά σημαντικότερες βιολογικές λειτουργίες στις οποίες συμμετέχουν τα υπερεκφραζόμενα γονίδια στα κύτταρα HeLa κατόπιν μόλυνσης με τον ιό Sendai.

5.2 ΤΑΥΤΟΠΟΙΗΣΗ 348 ΓΟΝΙΔΙΩΝ ΤΩΝ ΟΠΟΙΩΝ Η ΕΚΦΡΑΣΗ ΑΥΞΑΝΕΤΑΙ ΚΑΤΟΠΙΝ ΜΟΛΥΝΣΗΣ ΣΕ ΟΛΕΣ ΤΙΣ ΚΥΤΤΑΡΙΚΕΣ ΣΕΙΡΕΣ

Οι κυτταρικές σειρές που μολύνθηκαν χωρίστηκαν σε τέσσερις ομάδες ανάλογα με τον τύπο τους σε ενδοθηλιακά, επιθηλιακά, καρκινικά και κύτταρα του ανοσοποιητικού συστήματος. Σε αυτές τις ομάδες επιλέχθηκαν τα γονίδια που υπερεκφράζονται κατόπιν της μόλυνσης. Για κάθε ομάδα τα γονίδια αυτά ενοποιήθηκαν και στη συνέχεια πραγματοποιήθηκε σύγκριση των γονιδίων των ομάδων αυτών. Με τη χρήση διαγράμματος Venn εντοπίστηκαν συνολικά 348 γονίδια (Εικόνα 36) τα οποία υπερεκφράζονται σε όλες τις κυτταρικές σειρές.



Εικόνα 36: Σύγκριση των υπερεκφραζόμενων γονιδίων των τεσσάρων κυτταρικών ομάδων με τη χρήση διαγραμμάτων Venn.

Τα γονίδια αυτά (Πίνακας 23) αποτελούν τον πυρήνα των γονιδίων που εμπλέκονται στην κυτταρική απόκριση κατόπιν μόλυνσης με ιούς αλλά και βακτήρια σε όλες τις κυτταρικές σειρές. Σε αυτά τα γονίδια ανήκει η IFNB1 και πολλά γονίδια ρύθμισης των ιντερφερονών (π.χ IRF2, IRF7) όπως επίσης διάφορες χημειοκίνες (π.χ. CCL5, CXCL10), ιντερλευκίνες (π.χ. IL7, IL15) γονίδια του JAK-STAT μονοπατιού (π.χ. JAK1, STAT1) και κασπάσες (CASP1, CASP4) που εμπλέκονται στη διαδικασία της απόπτωσης. Από την ανάλυση της γονιδιακής οντολογίας σε αυτά τα γονίδια εντοπίστηκε ότι οι στατιστικά σημαντικότερες βιολογικές λειτουργίες στις οποίες συμμετέχουν είναι η ανοσολογική απόκριση, η απόκριση σε ιό και η απόπτωση (Πίνακας 21). Μελετήθηκαν επίσης και τα σηματοδοτικά μονοπάτια που εμφανίζονται να επηρεάζονται σε στατιστικά σημαντικό βαθμό κατόπιν μόλυνσης, όπως το σηματοδοτικό μονοπάτι των υποδοχέων Toll, των JAK-STAT και της απόπτωσης (Πίνακας 22).

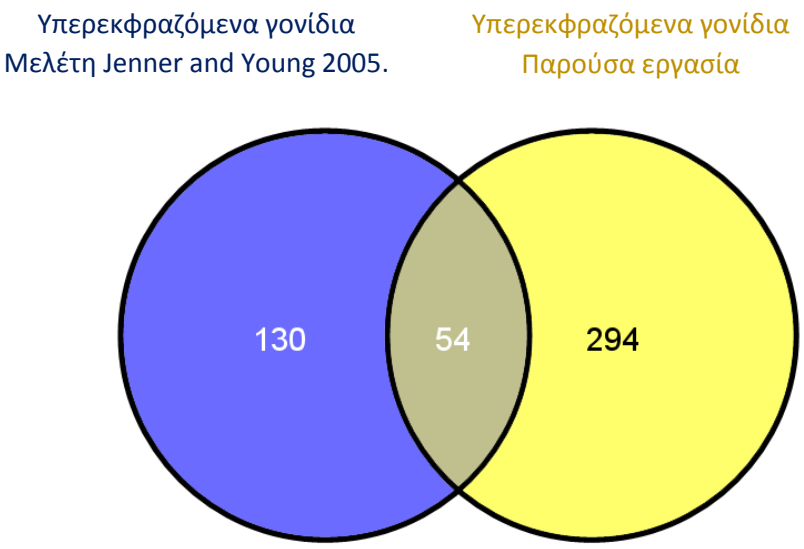
Βιολογικές λειτουργίες
immune response
response to virus
defense response
response to cytokine stimulus
positive regulation of response to stimulus
immune effector process
inflammatory response
regulation of apoptosis
cell death
antigen processing and presentation

Πίνακας 21: 10 στατιστικά σημαντικότερες βιολογικές λειτουργίες στις οποίες συμμετέχουν τα 348 υπερεκφραζόμενα γονίδια σε όλες τις κυτταρικές σειρές.

Σηματοδοτικά μονοπάτια στα οποία συμμετέχουν τα 348 γονίδια
Jak-STAT signaling pathway
Cytosolic DNA-sensing pathway
RIG-I-like receptor signaling pathway
Toll-like receptor signaling pathway
Antigen processing and presentation
Cytokine-cytokine receptor interaction
Proteasome
Apoptosis
NOD-like receptor signaling pathway
Allograft rejection

Πίνακας 22: Σηματοδοτικά μονοπάτια στα οποία εντοπίστηκαν να συμμετέχουν τα 348 κοινά γονίδια που υπερεκφράζονται σε όλες τις κυτταρικές σειρές κατόπιν μόλυνσης.

Από τα 348 αυτά γονίδια 54 (Εικόνα 37) εντοπίστηκαν να υπερεκφράζονται και σε μια μεγάλη μελέτη μετα-ανάλυσης δεδομένων DNA μικροσυστοιχιών σε ανθρώπινες κυτταρικές σειρές οι οποίες μολύνθηκαν με διάφορους μολυσματικούς παράγοντες όπως βακτήρια, μύκητες και ιοί (Jenner RG and Young RA 2005).

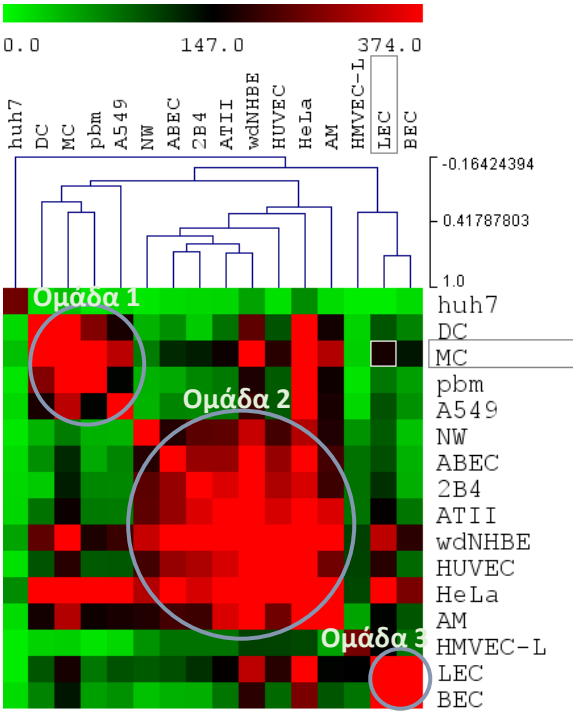


Εικόνα 37: Σύγκριση των κοινών υπερεκφραζόμενων γονιδίων σε όλα τα πειράματα της παρούσας εργασίας με τα υπερεκφραζόμενα γονίδια στη μελέτη των Jenner and Young 2005.

348 κοινά υπερεκφραζόμενα γονίδια σε όλες τις κυτταρικές σειρές							
ACTR2	CASP4	ETNK1	IFITM3	MXD1	PNPT1	SLC22A4	TP53INP2
ADAR	CASP7	ETS1	IFNB1	MYD88	PODXL	SLC25A23	TPBG
ADM	CBR3	ETS2	IL15	N4BP1	POMP	SLC25A28	TPK1
ADPRHL2	CBX4	FAM26F	IL15RA	NAMPT	PPM1K	SLC25A30	TRAF1
AFAP1L1	CCL5	FAM40B	IL28A	NAPA	PPP1R15A	SLC25A37	TRAF3IP2
ANKH	CCNE2	FAS	IL29	NBN	PRDM1	SLC38A5	TRAFD1
APOL1	CCPG1	FBXO6	IL6ST	NCOA7	PRIC285	SLC3A2	TREX1
APOL2	CCRN4L	FNDC3B	IL7	NDUFB2	PRRG4	SLC8A1	TRIM14
APOL3	CD47	FOS	IL7R	NEDD1	PSMB10	SLCO3A1	TRIM21
APOL6	CDC42SE2	FRMD8	IRAK2	NEDD9	PSMB8	SLFN5	TRIM22
ARHGAP27	CDCP1	FST	IRF2	NFKB2	PSMB9	SMCHD1	TRIM25
ASNS	CDKN2A	FZD5	IRF7	NFKBIA	PSME1	SNORD89	TRIM26
ASPH	CDKN2B	GBP1	IRF9	NFKBIZ	PSME2	SOCS1	TRIM38
ATF3	CEBPD	GBP3	IRS1	NLRC5	PXK	SOD2	TRIM5
ATPIF1	CFB	GCA	ISG15	NMI	RARRES3	SP100	TRIM69
ATRIP	CFI	GCH1	ISG20	NRP1	RBCK1	SP110	TSPAN13
AZI2	CFLAR	GGPS1	JAK1	NRP2	RGS2	SP140L	TYMP
BAG1	CH25H	GIMAP2	JAK2	NT5C3	RGS20	SPHK1	UBA6
BATF2	CMPK2	GLRX	KIAA1217	NT5E	RHEBL1	SPPL2A	UBA7
BCL2L11	CNIH4	GRK5	KRT10	NUB1	RHOC	SPTBN1	UBE2F
BHLHE40	CNP	GTPBP1	KYNU	OAS1	RIPK2	SPTLC2	UBE2J1
BIRC3	CPEB2	HAS2	LAP3	OAS2	RNF114	SQRDL	UBE2L6
BLZF1	CPM	HDX	LGALS3BP	OAS3	RNF149	SRD5A1	USP15
BMPR2	CREM	HERC5	LGALS8	OASL	RNF19B	STAT1	USP18
BTC	CSGALNACT2	HERC6	LGMN	OBFC2A	RNF213	STAT2	WARS
BTN3A3	CSNK1A1	HES4	LOC728855	OPTN	ROD1	STOM	XAF1
C15orf39	CST3	HIST1H2BK	LRG1	OSMR	RORA	STS	XRN1
C19orf66	CSTF3	HLA-E	LRP10	OSTM1	RPL35A	STX11	YAP1
C1R	CXCL10	HLA-F	LRRC8C	PANX1	RSAD2	STX17	YIPF1
C1S	CXCL11	HLA-G	LY6E	PARP10	RTP4	SYNPO2	ZC3HAV1
C2	DCP1A	HSPA1A	LYN	PARP12	SAMD9	TAP1	ZCCHC2
C21orf91	DDX52	ICAM1	MAFF	PARP14	SAMD9L	TAP2	ZFP36L2
C3orf23	DDX58	IFI16	MANEA	PARP9	SAMHD1	TAPBPL	ZNFX1
C3orf38	DDX60	IFI27	MAP2K3	PCDH17	SAP18	TBC1D1	
C4orf3	DHX58	IFI30	MAP3K8	PCGF5	SAP30L	TCEB3	
C4orf34	DPH3	IFI35	MASTL	PDCD2	SAT1	TDRD7	
C5orf13	DSP	IFI44	MGAT4A	PDZD2	SCARB2	TFPI2	
C5orf28	DTX3L	IFI6	MLEC	PECAM1	SCD	TMEM140	
C6orf150	DUSP4	IFIH1	MLKL	PELO	SEL1L	TMEM47	
C6orf192	EGFR	IFIT1	MMAA	PHC3	SEPW1	TMEM62	
C7orf42	EGR1	IFIT2	MMP14	PHF11	SERPINB1	TNFAIP3	
CACNA1A	EIF2AK2	IFIT3	MOBKL2C	PI4K2B	SERPINB8	TNFAIP6	
CALM3	ERAP1	IFIT5	MT1M	PITPNC1	SETX	TNFSF10	
CARD16	ERAP2	IFITM1	MX1	PLSCR1	SIPA1L2	TNFSF13B	
CASP1	ERO1L	IFITM2	MX2	PMAIP1	SLC17A5	TNIP1	

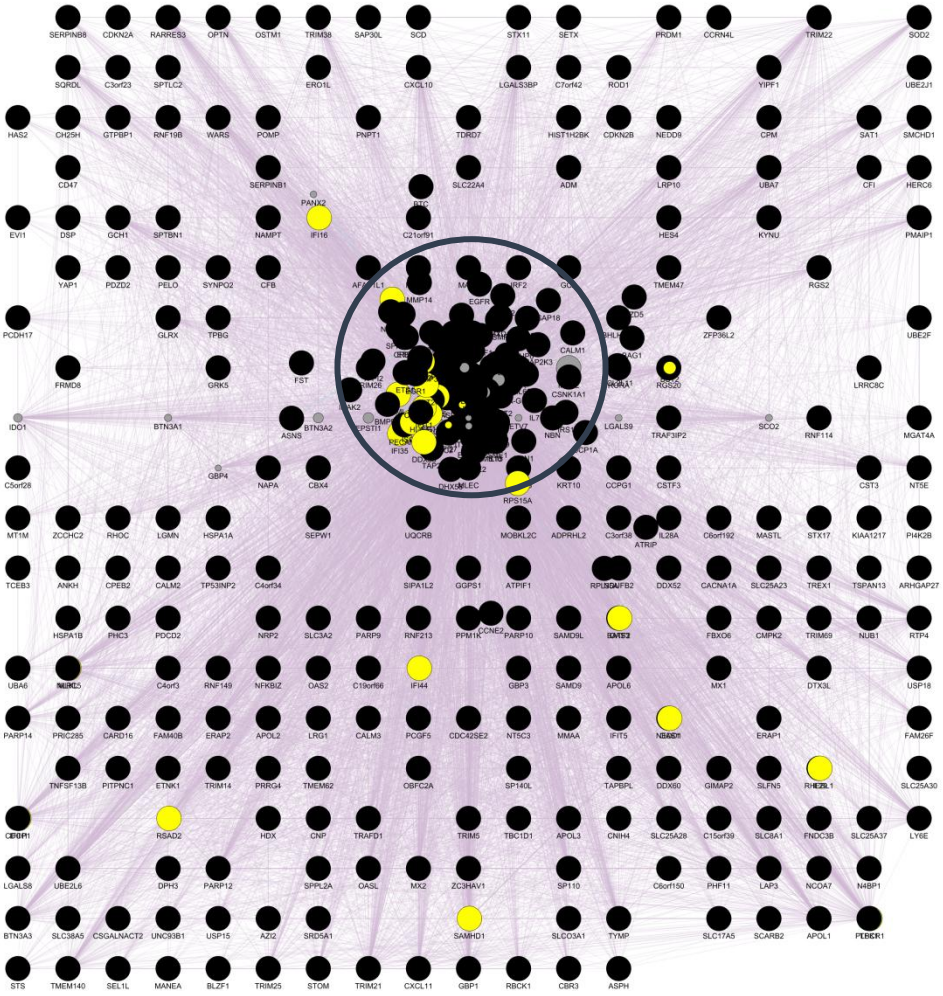
Πίνακας 23: 348 γονίδια των οποίων η έκφραση αυξάνεται σε όλες τις κυτταρικές σειρές κατόπιν μόλυνσης με διαφορετικά ικά στελέχη.

Πραγματοποιήθηκε ιεραρχική ταξινόμηση των σχέσεων των επιπέδων έκφρασης των 348 κοινά υπερεκφραζόμενων γονιδίων στις διάφορες κυτταρικές σειρές η οποία αναπαραστάθηκε με μορφή θερμικού χάρτη. Η ομαδοποίηση οδήγησε στον εντοπισμό 3 ομάδων κυτταρικών σειρών που εμφανίζουν σημαντική ομοιότητα μεταξύ τους ως προς τα γονίδια που υπερεκφράζονται (Εικόνα 38). Η πρώτη αποτελείται από δενδριτικά κύτταρα και μονοκύτταρα δηλαδή κύτταρα του ανοσοποιητικού συστήματος. Στη δεύτερη ομάδα υπάρχουν επιθηλιακά βρογχικά κύτταρα, επιθηλιακά πνευμονικά κύτταρα και επιθηλιακά κύτταρα HeLa. Στην τρίτη ομάδα ανήκουν ενδοθηλιακά κύτταρα των αιμοφόρων αγγείων και λεμφικά ενδοθηλιακά κύτταρα.

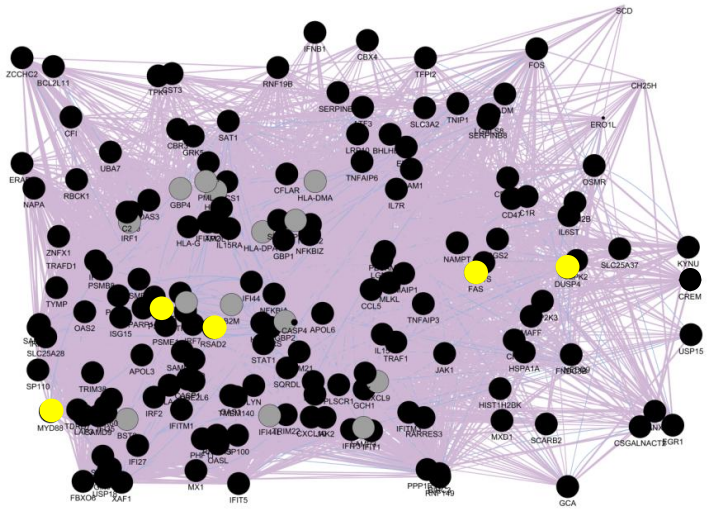


Εικόνα 38: Θερμικός χάρτης με ομαδοποιημένες τις κυτταρικές σειρές ως προς την ύπαρξη κοινών υπερεκφραζόμενων γονιδίων.

Για να καθοριστούν οι σχέσεις μεταξύ των 348 αυτών γονιδίων δημιουργήθηκε ένα δίκτυο με αυτά τα γονίδια όπου φαίνονται οι αλληλεπιδράσεις που υπάρχουν μεταξύ τους με βάση δημοσιευμένες μελέτες από πρωτεομικά δεδομένα, δεδομένων συνέκφρασης και συνεντοπισμού. Για τη δημιουργία του δικτύου αυτού χρησιμοποιήθηκε η βάση δεδομένων Genemania η οποία εγκαταστάθηκε στο πρόγραμμα Cytoscape. Όπως φαίνεται στην Εικόνα 39 τα περισσότερα από τα 348 αλληλεπιδρούν με τα υπόλοιπα. Μεταξύ αυτών ενδιαφέρον παρουσιάζει μια ομάδα 105 γονιδίων όπου παρατηρούνται οι περισσότερες αλληλεπιδράσεις μεταξύ τους (Εικόνα 40).



Εικόνα 39: Δίκτυο 348 γονιδίων όπου παρουσιάζονται οι αλληλεπιδράσεις των γονιδίων αυτών μεταξύ τους με βάση πρωτεομικά δεδομένα και δεδομένα συνεντοπισμού και συνέκφρασης από δημοσιευμένες μελέτες (με κίτρινο χρώμα επισημαίνονται τα γονίδια του σηματοδοτικού μονοπατιού των Toll υποδοχέων).



Εικόνα 40: 105 γονίδια με τις περισσότερες αλληλεπιδράσεις μεταξύ τους.

5.3 ΑΝΑΛΥΣΗ ΔΙΑΦΟΡΙΚΗΣ ΓΟΝΙΔΙΑΚΗΣ ΕΚΦΡΑΣΗΣ ΣΕ ΑΝΘΡΩΠΙΝΕΣ ΚΥΤΤΑΡΙΚΕΣ ΣΕΙΡΕΣ ΚΑΤΟΠΙΝ ΙΙΚΗΣ ΜΟΛΥΝΣΗΣ ΜΕ RNA-seq

Χρησιμοποιώντας πειράματα RNA-seq σε CD4+ (Chang ST et al. 2011) και HeLa (Evans VC et al. 2011) κύτταρα κατόπιν μόλυνσης με τον ιό HIV και τον αδενοϊό αντίστοιχα μελετήθηκε η διαφορική έκφραση των γονιδίων.

Από την ανάλυση των RNA-seq πειραμάτων που πραγματοποιήθηκαν στα CD4+ κύτταρα εντοπίστηκαν 19 γονίδια τα οποία υπερεκφράζονται 12 ώρες κατόπιν της ιικής μόλυνσης και 1583 γονίδια υπερεκφράζονται 24 ώρες κατόπιν της μόλυνσης. Από την ανάλυση γονιδιακής οντολογίας που πραγματοποιήθηκε εντοπίστηκαν οι βιολογικές λειτουργίες στις οποίες συμμετέχουν αυτά τα γονίδια (Πίνακας 24). Οι σημαντικότερες από αυτές είναι οι κυτταρική προσκόλληση και η κυτταρική μορφογένεση. Αξιοσημείωτο είναι ότι αυτές οι βιολογικές λειτουργίες δεν συμμετέχουν στην ανοσολογική απόκριση επομένως περαιτέρω διευκρίνιση κρίνεται αναγκαία.

Βιολογικές λειτουργίες
cell adhesion
biological adhesion
cell projection organization
cell morphogenesis involved in neuron differentiation
cell morphogenesis involved in differentiation
cell morphogenesis
cell projection morphogenesis
cellular component morphogenesis
cell part morphogenesis
response to cAMP
response to wounding

Πίνακας 24: Βιολογικές λειτουργίες των 1583 υπερεκφραζομένων γονιδίων κατόπιν μόλυνσης των CD4+ κυττάρων με τον ιό HIV.

Στα πειράματα RNA-seq που πραγματοποιήθηκαν σε ανθρώπινα HeLa κύτταρα κατόπιν μόλυνσης με τον αδενοϊό εντοπίστηκαν 63 γονίδια των οποίων η έκφραση μεταβάλλεται σημαντικά. Οι βιολογικές λειτουργίες στις οποίες συμμετέχουν σχετίζονται με την απόπτωση και τον κυτταρικό θάνατο (Πίνακας 25) και τα γονίδια που εμφανίζουν τις μεγαλύτερες αλλαγές είναι τα BMF, DDIT4, FOSL2, ACTC1, CARD6, GAS1, SGK1.

Βιολογικές λειτουργίες
cell death
death
programmed cell death
regulation of transcription from RNA polymerase II promoter
apoptosis
regulation of cell cycle
response to protein stimulus
regulation of apoptosis
heart development
regulation of programmed cell death

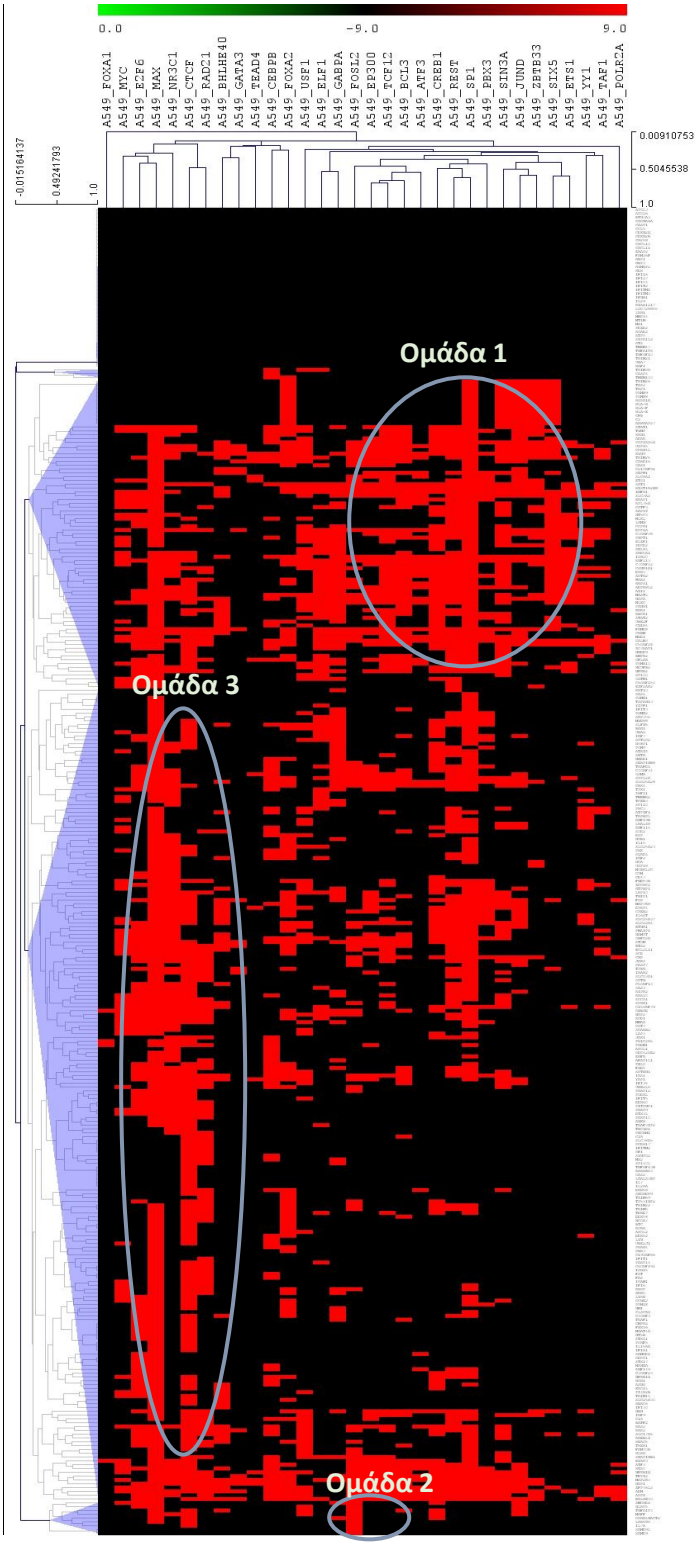
Πίνακας 25: Βιολογικές λειτουργίες των 63 υπερεκφραζομένων γονιδίων κατόπιν μόλυνσης των HeLa κυττάρων με τον αδενοϊό.

5.4 ΣΥΣΧΕΤΙΣΗ ΠΡΟΣΔΕΣΗΣ ΜΕΤΑΓΡΑΦΙΚΩΝ ΠΑΡΑΓΟΝΤΩΝ ΜΕ ΔΙΑΦΟΡΙΚΗ ΕΚΦΡΑΣΗ ΓΟΝΙΔΙΩΝ

Αξιοποιήθηκε η βάση δεδομένων του ENCODE από όπου συλλέχθηκαν δεδομένα από ChIP-seq πειράματα για 3 κυτταρικές σειρές για τις οποίες υπήρχαν και δεδομένα διαφορικής γονιδιακής έκφρασης. Για κύτταρα A549 χρησιμοποιήθηκαν δεδομένα από 32 μεταγραφικούς παράγοντες, για κύτταρα HUVEC χρησιμοποιήθηκαν δεδομένα για 8 μεταγραφικούς παράγοντες και για κύτταρα HeLa χρησιμοποιήθηκαν δεδομένα για 58 μεταγραφικούς παράγοντες και 11 ιστονικές τροποποιήσεις.

A549 ΚΥΤΤΑΡΑ

Για τα επιθηλιακά A549 κύτταρα, τα BED αρχεία με τις θέσεις πρόσδεσης των μεταγραφικών παραγόντων στο DNA επεξεργάσθηκαν με τον αλγόριθμο CEAS ώστε να εντοπιστούν οι αποστάσεις αυτών των θέσεων από τα γονίδια του οργανισμού. Επιλέχθηκαν μόνο τα γονίδια στα οποία υπάρχει πρόσδεση σε απόσταση 5000 βάσεις ανοδικά ή καθοδικά από το σημείο έναρξης της μεταγραφής. Στη συνέχεια εξετάστηκαν οι θέσεις πρόσδεσης των 32 μεταγραφικών παραγόντων για τα 348 γονίδια, που αποτελούν τον πυρήνα των γονιδίων τα οποία υπερεκφράζονται κατόπιν της ιικής μόλυνσης. Η ύπαρξη πρόσδεσης σε απόσταση έως 5000 βάσεις ανοδικά ή καθοδικά του σημείου έναρξης της μεταγραφής παρουσιάζεται με μορφή θερμικού χάρτη στην Εικόνα 41. Το κόκκινο χρώμα αντιπροσωπεύει την ύπαρξη θέσης πρόσδεσης στη συγκεκριμένη απόσταση σε σχέση με το σημείο έναρξης της μεταγραφής ενός γονιδίου. Στα δεδομένα αυτά πραγματοποιήθηκε ιεραρχική ομαδοποίηση των μεταγραφικών παραγόντων που προσδένονται στα ίδια γονίδια και πιθανώς ρυθμίζουν από κοινού την έκφραση αυτών. Συνολικά εντοπίστηκαν 3 ομάδες μεταγραφικών παραγόντων οι οποίοι παρουσιάζουν πιθανή συνέργεια στη ρύθμιση των 348 γονιδίων. Στους Πίνακες 26-28 παρουσιάζονται οι μεταγραφικοί παράγοντες και τα αντίστοιχα γονίδια που ρυθμίζουν.



Εικόνα 41: Ιεραρχική ταξινόμηση ‘’χωρίς επίβλεψη’’(unsupervised hierarchical clustering) σε μορφή θερμικού χάρτη με ομαδοποιημένους τους 32 μεταγραφικούς παράγοντες της κυτταρικής σειράς A549 ως προς την ύπαρξη θέσης πρόσδεσης σε απόσταση έως 5000bp ανοδικά η καθοδικά του TSS των 348 γονιδίων.

ΟΜΑΔΑ 1		
TFs	***GENES***	
ELF1	TRIM26	PNPT1
GABPA	TAP2	BLZF1
FOSL2	TAP1	PDCD2
EP300	PSMB9	SEL1L
TCF12	PSMB8	SRD5A1
BCL3	HSPA1A	ISG20
ATF3	HLA-G	RNF213
CREB1	HLA-F	C7ORF42
REST	HLA-E	CSNK1A1
SP1	CFB	DPH3
PBX3	C2	ACTR2
SIN3A	ARHGAP27	WARS
JUND	STAT1	GGPS1
ZBTB33	TYMP	ADPRHL2
SIX5	XRN1	AZI2
ETS1	ADAR	MASTL
YY1	PPP1R15A	NAPA
TAF1	USP15	MLEC
	CCRN4L	PANX1
	KYNU	EGR1
	TRIM25	RBCK1
	CARD16	STAT2
	OAS1	UBE2F
	C21ORF91	CNIH4
	SEPW1	FRMD8
	SLC8A1	CREM
	ETS1	MXD1
	SAT1	CALM3
	HIST1H2BK	C5ORF28
	ZNFX1	ZC3HAV1
	SLC3A2	NEDD9
	ERAP1	BMPR2
	RPL35A	CFLAR
	CSTF3	PSMB10
	SAP18	NDUFB2
	NT5C3	NFKB2
	MLKL	SP100
	LGMN	OSTM1
	CCPG1	C6ORF150
	DCP1A	EIF2AK2
	C3ORF38	

Πίνακας 26: Μεταγραφικοί παράγοντες και γονίδια που ρυθμίζουν στην ομάδα 1.

ΟΜΑΔΑ 2	
TFs	***GENES***
FOSL2	KRT10
EP300	OASL
TCF12	PSME1
BCL3	TSPAN13
ATF3	YIPF1
	IFIT3
	PSME2
	SAP30L
	MYD88
	SLFN5
	BAG1
	UBA6
	IRF7
	SPTLC2

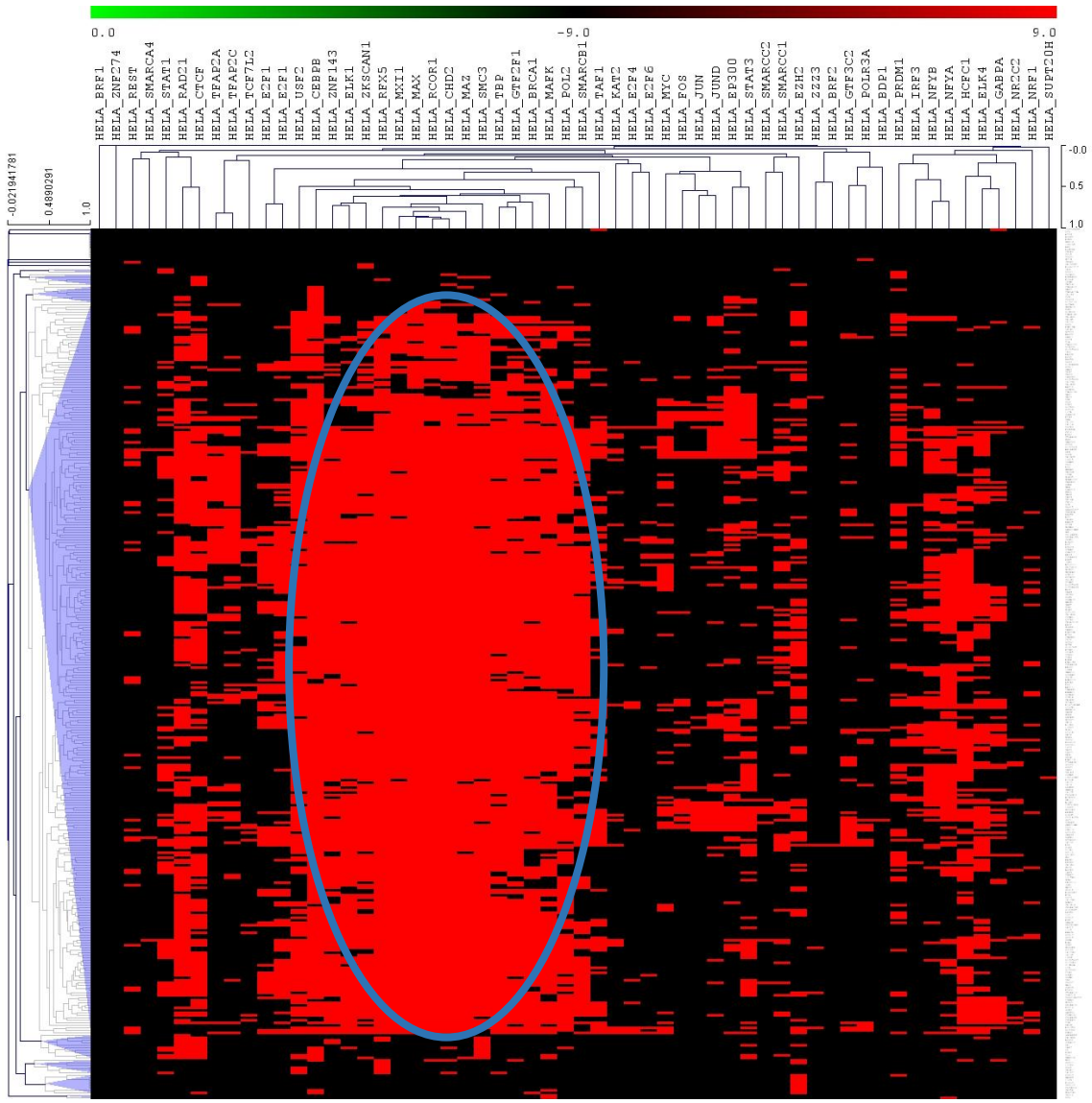
Πίνακας 27: Μεταγραφικοί παράγοντες και γονίδια που ρυθμίζουν στην ομάδα 2.

ΟΜΑΔΑ 3					
TFs	***GENES***				
MYC	DSP	C5ORF13	PECAM1	ICAM1	RGS2
E2F6	NUB1	OAS3	C1R	IFI6	SLC17A5
MAX	IL15	RIPK2	SLC38A5	RHOC	RHEBL1
NR3C1	SLC25A23	RGS20	PCDH17	GRK5	HERC5
CTCF	PXK	SOCS1	IFITM2	LY6E	TREX1
RAD21	SQRDL	SPHK1	CFI	CCNE2	FAM40B
	IRF2	C15ORF39	SYNPO2	PPM1K	GLRX
	GCA	CEBPD	MX2	NMI	SERPINB1
	USP18	NRP2	SP140L	PLSCR1	BIRC3
	MOBKL2C	ROD1	TNFSF13B	C4ORF3	ATF3
	CPM	MMAA	RARRES3	TRAF1	HES4
	CD47	CST3	OAS2	CMPK2	NFKBIA
	FNDC3B	SCARB2	LGALS3BP	FBXO6	TFPI2
	ZCCHC2	LAP3	IL7	MGAT4A	MAP2K3
	GTPBP1	JAK1	IL28A	NT5E	NRP1
	LRP10	PRIC285	DHX58	STX11	ZFP36L2
	TNIP1	PRDM1	SNORD89	PCGF5	ADM
	FOS	APOL1	TRIM69	IL15RA	ASPH
	MAP3K8	CDC42SE2	TP53INP2	IFIH1	BHLHE40
	DUSP4	EGFR	TRIM22	SAMHD1	
	CPEB2	AFAP1L1	TRIM5	CDCP1	
	IL6ST	PELO	TDRD7	STX17	
	SLC25A37	FZD5	DDX58	MANEA	
	SLC22A4	SPTBN1	NCOA7	RNF149	
	ETNK1	IRS1	BTC	C3ORF23	
	PMAIP1	YAP1	RORA	NFKBIZ	
	NAMPT	IFI35	APOL2	GCH1	
	OBFC2A	UBE2L6	DDX52	ASNS	
	STOM	PARP12	LYN	ERO1L	
	ETS2	PODXL	UBE2J1	PI4K2B	
	BCL2L11	IFIT5	PRRG4	TRIM14	
	SCD	DDX60	CBR3	SLC25A30	
	CNP	PITPNC1	C19ORF66	HERC6	
	JAK2	PARP9	IFIT1	IFI30	
	CASP7	DTX3L	PARP14	NBN	
	TPBG	PARP10	C6ORF192	IRF9	
	IRAK2	ANKH	ISG15	C1S	
	SLCO3A1	TRAF3IP2	FST	BATF2	
	OPTN	TBC1D1	FAS	HAS2	

Πίνακας 28: Μεταγραφικοί παράγοντες και γονίδια που ρυθμίζουν στην ομάδα 3.

HeLa ΚΥΤΤΑΡΑ

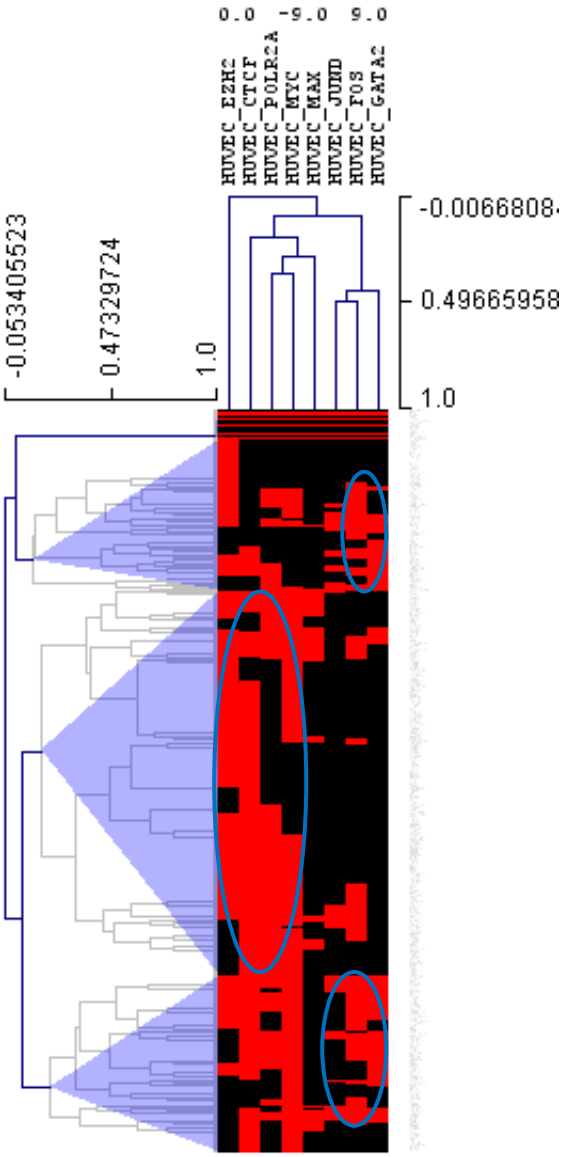
Αντίστοιχα με τα A549 κύτταρα τα BED αρχεία των HeLa κυττάρων με τις θέσεις πρόσδεσης των μεταγραφικών παραγόντων στο DNA επεξεργάσθηκαν με τον αλγόριθμο CEAS ώστε να εντοπιστούν οι αποστάσεις αυτών των θέσεων από τα γονίδια του οργανισμού. Στη συνέχεια δημιουργήθηκε ο θερμικός χάρτης (Εικόνα 42) με την ύπαρξη θέσης πρόσδεσης των 58 μεταγραφικών παραγόντων σε απόσταση έως 5000 βάσεων ανοδικά ή καθοδικά από το σημείο έναρξης της μεταγραφής των 348 γονιδίων και πραγματοποιήθηκε ιεραρχική συσταδοποίηση των δεδομένων. Από την ανάλυση αυτή προέκυψε μια μεγάλη ομάδα στην οποία καταδεικνύονται 18 μεταγραφικοί παράγοντες να προσδένονται κοντά στο σημείο έναρξης της μεταγραφής 266 γονιδίων.



Εικόνα 42: Ιεραρχική ταξινόμηση ''χωρίς επίβλεψη''(unsupervised hierarchical clustering) σε μορφή θερμικού χάρτη με ομαδοποιημένους τους 58 μεταγραφικούς παράγοντες της κυτταρικής σειράς HeLa ως προς την ύπαρξη θέσης πρόσδεσης σε απόσταση έως 5000bp ανοδικά ή καθοδικά του TSS των 348 γονιδίων.

HUVEC ΚΥΤΤΑΡΑ

Από τη βάση δεδομένων του ENCODE χρησιμοποιήσαμε δεδομένα από ChIP-seq πειράματα για τα σημεία πρόσδεσης 8 μεταγραφικών παραγόντων στα ενδοθηλιακά κύτταρα HUVEC. Τα BED αρχεία με τις συγκεκριμένες πληροφορίες αναλύθηκαν με τον αλγόριθμο CEAS και έτσι εντοπίστηκαν οι αποστάσεις στις οποίες προσδένονται αυτοί οι παράγοντες σε σχέση με τα γονίδια του οργανισμού. Όπως και στις προηγούμενες κυτταρικές σειρές δημιουργήθηκε ο θερμικός χάρτης με την ύπαρξη πρόσδεσης των παραγόντων αυτών σε απόσταση έως +/- 5000 βάσεις από το σημείο έναρξης της μεταγραφής των 348 γονιδίων (Εικόνα 43). Από την ιεραρχική ταξινόμηση των δεδομένων αυτών εντοπίστηκαν 3 ομάδες μεταγραφικών παραγόντων οι οποίοι φαίνεται να συνδέονται στη ρυθμιστική περιοχή κάποιων από τα 348 γονίδια.



Εικόνα 43: Ιεραρχική ταξινόμηση ''χωρίς επίβλεψη''(unsupervised hierarchical clustering) σε μορφή θερμικού χάρτη με ομαδοποιημένους τους 8 μεταγραφικούς παράγοντες της κυτταρικής σειράς HUVEC ως προς την ύπαρξη θέσης πρόσδεσης σε απόσταση έως 5000bp ανοδικά ή καθοδικά του TSS των 348 γονιδίων.

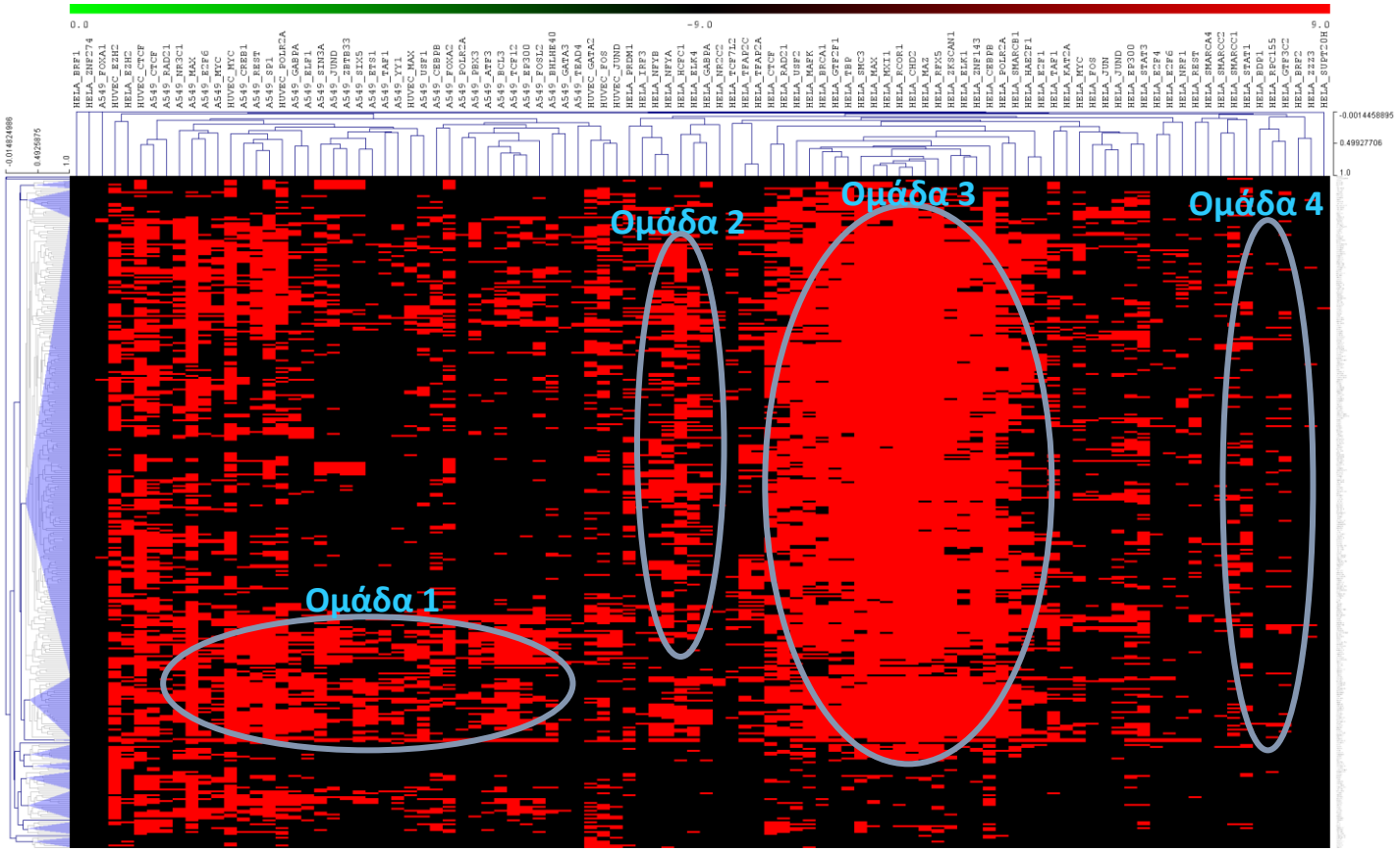
Στη συνέχεια δημιουργήθηκε ένας θερμικός χάρτης με την πρόσδεση όλων των μεταγραφικών παραγόντων των 3 αυτών κυτταρικών σειρών σε απόσταση +/- 5000 βάσεων από το σημείο έναρξης της μεταγραφής των 348 γονιδίων. Πραγματοποιήθηκε ιεραρχική ταξινόμηση και ομαδοποιήθηκαν οι μεταγραφικοί παράγοντες της κάθε κυτταρικής σειράς ως προς την ομοιότητα πρόσδεσής τους σε αυτά τα γονίδια (Εικόνα 44). Από την ιεραρχική ταξινόμηση εντοπίστηκαν 4 ομάδες μεταγραφικών παραγόντων.

Στη πρώτη ομάδα ανήκουν 7 μεταγραφικοί παράγοντες από τη κυτταρική σειρά HUVEC και 31 μεταγραφικοί παράγοντες από την A549 κυτταρική σειρά και ρυθμίζουν 57 από τα 248 γονίδια. Σε αυτή την ομάδα εντοπίζονται σε κοντινές αποστάσεις οι παράγοντες MYC και CTCF καθώς επίσης και οι παράγοντες YY1 και SP1 οι οποίοι ανήκουν στην ίδια οικογένεια παραγόντων των zing-finger-domain πρωτεϊνών.

Στην δεύτερη ομάδα ανήκουν 8 μεταγραφικοί παράγοντες της HeLa κυτταρικής σειράς τα οποία ρυθμίζουν 194 γονίδια. Σε αυτή την ομάδα ανήκουν οι μεταγραφικοί παράγοντες PRDM1 και IRF3 οι οποίοι εμπλέκονται στη ρύθμιση της έκφρασης της IFNB1. Επίσης εντοπίστηκαν οι παράγοντες NFYA και NFYB οι οποίοι αποτελούν μονάδες ενός τριμερούς συμπλόκου που ρυθμίζει την έκφραση αρκετών γονιδίων. Τέλος σε αυτή την ομάδα ανήκουν οι παράγοντες ELK4 και GABPA οι οποίοι ανήκουν στη οικογένεια των παραγόντων τύπου ETS.

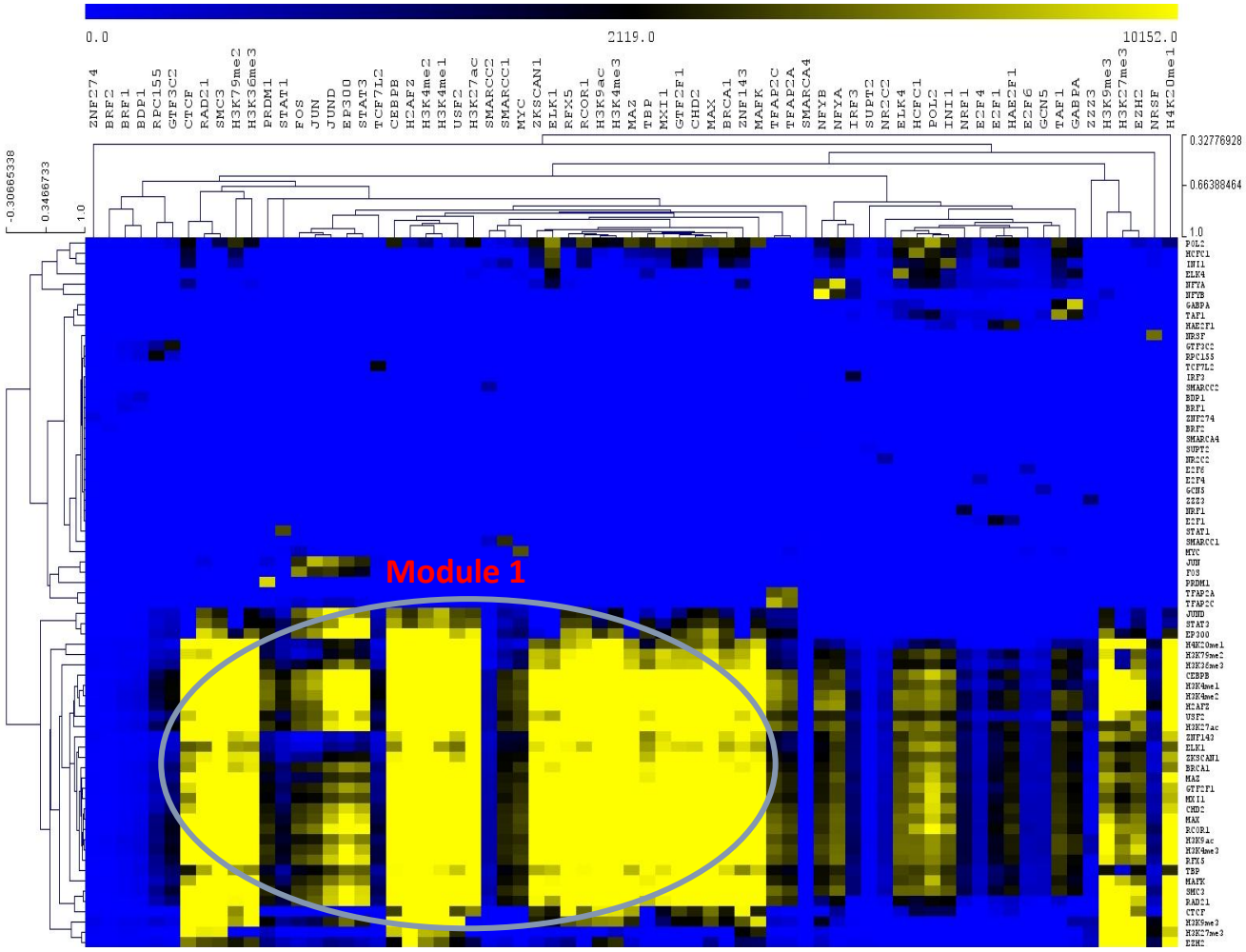
Στη τρίτη ομάδα ανήκουν 36 μεταγραφικοί παράγοντες που ξεχώρισαν από τα δεδομένα της HeLa κυτταρικής σειράς οι οποίοι ρυθμίζουν 271 γονίδια. Σε αυτή την ομάδα ανήκει ο MAX ο οποίος δημιουργεί ομοδιμερή ή ετεροδιμερή με τους MYC και MXI1. Επιπλέον εντοπίστηκαν οι παράγοντες TFAP2C ,TFAP2A και οι E2F1,E2F4 και E2F6 που είναι μέλη της ίδιας οικογένειας και έχουν σημαντικό ρόλο στον έλεγχο του κυτταρικού κύκλου και της δράσης ογκοκατασταλτικών γονιδίων.

Στη τέταρτη ομάδα ανήκουν 10 μεταγραφικοί παράγοντες που ξεχώρισαν από τα δεδομένα της HeLa κυτταρικής σειράς οι οποίοι ρυθμίζουν 267 γονίδια. Σε αυτούς ανήκουν οι SMARCC1,SMARCC2 και SMARCA4 μέλη του SWI/SNF συμπλόκου. Επιπλέον εντοπίστηκαν και οι παράγοντες BRF2 και BDP1 που σχετίζονται με τη δράση της RNA pol III.



Εικόνα 44: Θερμικός χάρτης με ομαδοποιημένους όλους τους μεταγραφικούς παράγοντες από τις 3 κυτταρικές σειρές ως προς την ύπαρξη θέσης πρόσδεσης σε απόσταση έως 5000bp ανοδικά ή καθοδικά του TSS των 348 γονιδίων.

Τέλος για την κυτταρική σειρά HeLa από την οποία υπάρχουν δεδομένα για τους περισσότερους μεταγραφικούς παράγοντες και ιστονικές τροποποιήσεις στο ENCODE δημιουργήθηκε ένας θερμικός χάρτης βασισμένος στα κοινά σημεία πρόσδεσης αυτών των παραγόντων (Εικόνα 45). Έτσι βρέθηκε 1 ρυθμιστικό σύμπλοκο (module) μεταγραφικών παραγόντων και ιστονικών τροποποιήσεων οι οποίοι συνεντοπίζονται στο γονιδίωμα. Στον Πίνακα 29 φαίνονται οι μεταγραφικοί παράγοντες και οι ιστονικές τροποποιήσεις που συνεντοπίζονται. Με κόκκινο χρώμα παρουσιάζονται οι παράγοντες MAX, MYC και η ιστονική τροποποίηση H3K4me3. Οι παράγοντες αυτοί σύμφωνα με τους Orkin SH et al. 2010 ανήκουν στο MYC ρυθμιστικό σύμπλοκο (module), το οποίο ρυθμίζει την έκφραση συγκεκριμένων γονιδίων στα ES κύτταρα. Στην ίδια εργασία αποδεικνύεται επίσης ότι υπάρχουν κοινοί μηχανισμοί ρύθμισης των γονιδίων σε ES και καρκινικά κύτταρα.



Εικόνα 45: Θερμικός χάρτης με ομαδοποιημένους τους μεταγραφικούς παράγοντες και τις ιστονικές τροποποιήσεις ως προς τις κοινές θέσεις πρόσδεσης τους στο DNA.

MODULE 1	
POL2	GTF3C2
TBP	CTCF
ELK1	RAD21
RFX5	SMC3
MAX	H3K79me2
RCOR1	H3K36me3
H3K9ac	PRDM1
H3K4me3	STAT1
CHD2	FOS
GTF2F1	JUN
MAZ	JUND
MXI1	EP300
BRCA1	STAT3
ZKSCAN1	TCF7L2
ZNF143	CEBPB
H3K36me3	H2AFZ
USF2	H3K4me2
H3K27ac	H3K4me1
H2AFZ	USF2
H3K4me2	H3K27ac
H3K4me1	SMARCC2
CEBP	SMARCC1
MAFK	MYC
SMC3	ZKSCAN1
RAD21	ELK1
CTCF	RFX5
H3K79me2	RCOR1
	H3K9ac
	MAFK
	H3K4me3
	MAZ
	TBP
	MXI1
	GTF2F1
	CHD2
	MAX
	BRCA1
	ZNF143

Πίνακας 29: Μεταγραφικοί παράγοντες και οι ιστονικές τροποποιήσεις που συνεντοπίζονται.

ΣΥΖΗΤΗΣΗ

Τα κύτταρα αποκρίνονται στη μόλυνση με ένα δίκτυο ελέγχου της μεταγραφικής τους δραστηριότητας στο οποίο συμμετέχουν διάφορα γονίδια που εμπλέκονται στην άμυνα του οργανισμού. Οι παθογόνοι μικροοργανισμοί μπορούν να παρεμβαίνουν στην έκφραση αυτών των γονιδίων για να εντείνουν την παθογόνο δράση τους. Για να αντιμετωπίσουν τα κύτταρα-ξενιστές την απειλή αυτή έχουν αναπτύξει περίπλοκους αμυντικούς μηχανισμούς. Αρκετοί από αυτούς τους μηχανισμούς είναι κοινοί σε διαφορετικές κυτταρικές σειρές και περιλαμβάνουν την συντονισμένη δράση χιλιάδων γονιδίων μεταξύ αυτών και πολλών μεταγραφικών παραγόντων προκειμένου να επιβιώσει το κύτταρο.

Οι εναλλακτικοί μηχανισμοί απόκρισης των κυττάρων στους παθογόνους μικροοργανισμούς μπορούν να μελετηθούν με την αξιοποίηση δεδομένων από (high throughput) πειράματα μελέτης της γονιδιακής έκφρασης και διερεύνησης της ρύθμισής της. Έτσι μπορούν να εντοπιστούν οι κοινοί μηχανισμοί που ενεργοποιούνται σε διαφορετικές κυτταρικές σειρές κατόπιν μόλυνσης των κυττάρων με ιούς ή βακτήρια.

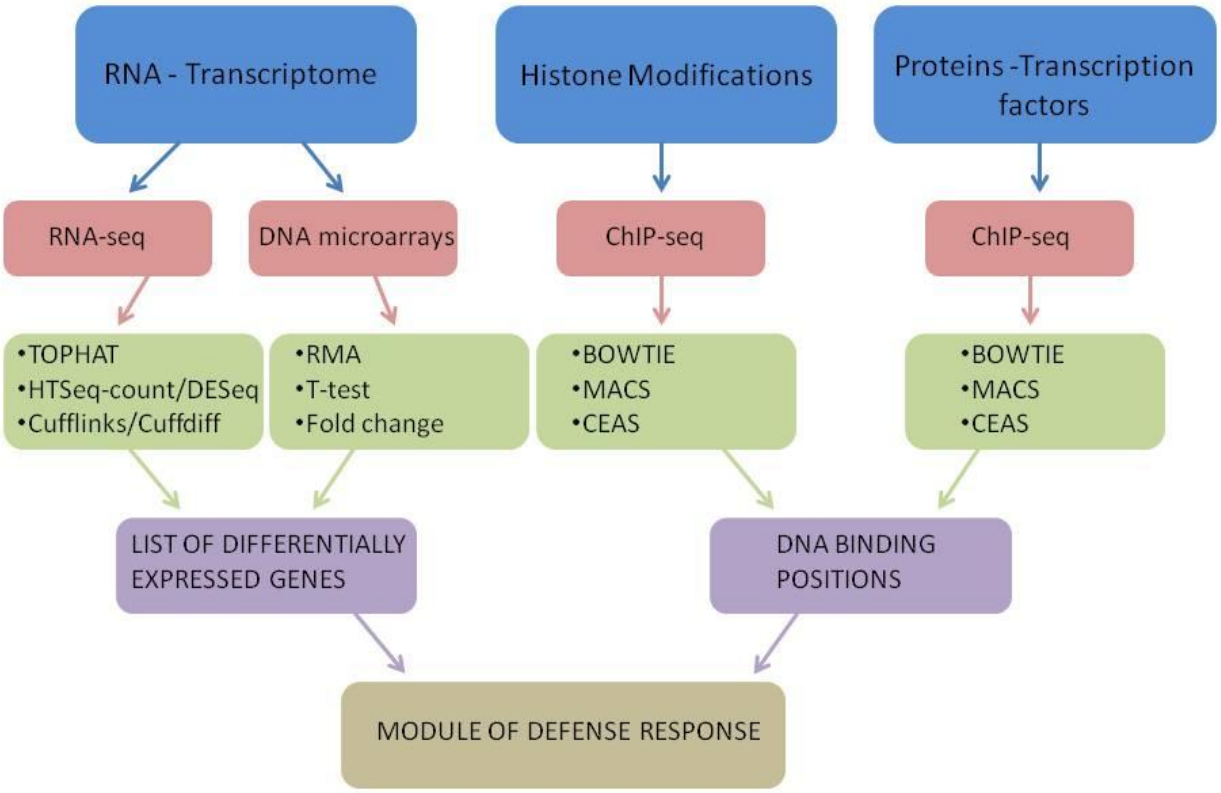
Στην παρούσα εργασία, δεδομένα διαφορικής γονιδιακής έκφρασης από πειράματα DNA μικροσυστοιχιών και RNA-seq που πραγματοποιήθηκαν σε ανθρώπινες κυτταρικές σειρές μετά από μόλυνση με διάφορους ιούς ή βακτήρια συσχετίστηκαν με την ύπαρξη θέσεων πρόσδεσης μεταγραφικών παραγόντων που εξήχθησαν από την ανάλυση ChIP-seq πειραμάτων. Δημιουργήθηκε έτσι μια βάση δεδομένων στην οποία εντοπίστηκαν 348 γονίδια των οποίων η έκφραση αυξάνεται σε όλες τις κυτταρικές σειρές που μελετήθηκαν κατόπιν μόλυνσης. Τα γονίδια αυτά αποτελούν τον πυρήνα ενός κοινού μηχανισμού απόκρισης ανεξαρτήτως της κυτταρικής σειράς ή της μόλυνσης. Από το σύνολο των μελετών, γονίδια που επάγονται συχνότερα μετά από την επίδραση του μολυσματικού παράγοντα σε όλες τις κυτταρικές σειρές είναι η IFNB1, CCL5, IL7, IRF7, JAK1 και CASP1.

Από την ανάλυση της γονιδιακής οντολογίας των γονιδίων αυτών προέκυψε ότι συμμετέχουν στις βιολογικές λειτουργίες της ανοσολογικής και αμυντικής απόκρισης, της απόκρισης σε ιούς καθώς επίσης της απόπτωσης και του κυτταρικού θανάτου. Επιπλέον πραγματοποιήθηκε ανάλυση για τα μονοπάτια μεταγωγής σήματος (από τη βάση δεδομένων KEGG) στα οποία συμμετέχουν αυτά τα γονίδια. Διαπιστώθηκε ότι συμμετέχουν στα σηματοδοτικά μονοπάτια των Jak-STAT, των Toll υποδοχέων των RIG-I υποδοχέων και στην απόπτωση.

Πίσω από αυτό τον πυρήνα γονιδίων υπάρχει ένα σύνολο μεταγραφικών παραγόντων και ιστονικών τροποποιήσεων που ρυθμίζουν την έκφρασή τους και οι οποίοι εντοπίστηκαν με τη χρήση ChIP-seq πειραμάτων. Πιο συγκεκριμένα μεταγραφικοί παράγοντες όπως οι MAX, FOS, MYC, BRCA1, JUN, STAT1 και οι ιστονικές τροποποιήσεις H3K79me2, H3K4me3, H3K36me3 φαίνεται να ρυθμίζουν τα γονίδια αυτά καθώς συνεντοπίζονται στις ρυθμιστικές περιοχές των περισσότερων γονιδίων στην περιοχή (-/+ 5000 σε σχέση με το TSS).

Στο πλαίσιο της εργασίας αυτής οι επιμέρους μελέτες χρησιμοποιήθηκαν προκειμένου να αναπτυχθεί μια καινοτόμος βιοπληροφορική και στατιστική μεθοδολογία μετα-ανάλυσης των πειραματικών δεδομένων με τρόπο ώστε να ελαχιστοποιούνται οι αποκλίσεις στα αποτελέσματα διαφορετικών πειραμάτων (Εικόνα 46).

Επιπροσθέτως, κατασκευάστηκε μια βάση δεδομένων που παρέχει τη δυνατότητα γρήγορης αναζήτησης της σχετικής αλλαγής της έκφρασης οποιουδήποτε γονιδίου σε 16 ανθρώπινες κυτταρικές σειρές οι οποίες είχαν μολυνθεί με ιούς. Επιπλέον προσφέρει τη δυνατότητα μελέτης ομάδων γονιδίων που ενδιαφέρουν τον ερευνητή. Τέλος αυτή η βάση δεδομένων μπορεί να επεκταθεί μελλοντικά ενσωματώνοντας δεδομένα από νέα πειράματα DNA μικροσυστοιχιών, RNA-seq και ChIP-seq εμπλουτίζοντας έτσι τον διαθέσιμο όγκο πληροφοριών και μειώνοντας τον βιολογικό θόρυβο από τα διάφορα πειράματα.



Εικόνα 46: Διαγραμματική απεικόνιση της μεθοδολογίας ανάλυσης των πειραμάτων που μελετήθηκαν στη παρούσα εργασία.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Akey C.W. and Luger K. (2003). "Histone chaperones and nucleosome assembly". Curr Opin Struct Biol 13, 6-14.

Ambekar C.S., Cheung B., Lee J., Chan L.C., Liang R., Kumana C.R. "Metabolism of chloramphenicol succinate in human bone marrow"(2000). European Journal of Clinical Pharmacology, 56 (5), pp. 405-409.

Analysis of Microarray Data, Genestat.

Anders, S., Huber, W. "Differential expression analysis for sequence count data"(2010). Genome Biology, 11 (10), art. no. R106.

Antonaki A., Demetriades C., Polyzos A., Banos A., Vatsellas G., Lavigne M.D., Apostolou E., Mantouvalou E., Papadopoulou D., Mosialos G., Thanos D. "Genomic analysis reveals a novel nuclear factor-κB (NF-κB)-binding site in Alu-repetitive elements"(2011). Journal of Biological Chemistry, 286 (44), pp. 38768-38782.

Börgeling Y, Schmolke M, Viemann D, Nordhoff C et al. "Inhibition of p38 mitogen-activated protein kinase impairs influenza virus-induced primary and secondary host gene responses and protects mice from lethal H5N1 infection". J Biol Chem 2014 Jan 3;289(1):13-27.

Bardet AF, He Q, Zeitlinger J, Stark A. "A computational pipeline for comparative ChIP-seq analyses"(2011). Nat Protoc 7: 45-61.

Blackham S, Baillie A, Al-Hababi F, Remlinger K et al. "Gene expression profiling indicates the roles of host oxidative stress, apoptosis, lipid metabolism, and intracellular transport genes in the replication of hepatitis C virus"(2010). J Virol May;84(10):5404-14.

Burrows M, Wheeler DJ. "A Block Sorting Lossless Data Compression Algorithm". Technical Report 124 Palo Alto, CA: Digital Equipment Corporation 1994.

Butchar J.P., Cremer T.J., Clay C.D., Gavrilin M.A., Wewers M.D., Marsh C.B., Schlesinger L.S., Tridandapani S. "Microarray analysis of human monocytes infected with Francisella tularensis identifies new targets of host response subversion"(2008) PLoS ONE, 3 (8), art. no. e2924.

Chang ST, Sova P, Peng X, Weiss J et al. "Next-generation sequencing reveals HIV-1-mediated suppression of T cell activation and RNA processing and regulation of noncoding RNA expression in a CD4+ T cell line"(2011). MBio;2(5).

Chen Y, Negre N, Li Q, Mieczkowska JO, Slattery M, et al. "Systematic evaluation of factors influencing ChIP-seq fidelity" (2012). Nat Methods 9: 609-614.

Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM:. "Stem cell transcriptome profiling via massive-scale mRNA sequencing". Nat Methods 2008, 5:613-619.

Dalrymple NA, Mackow ER. "Endothelial cells elicit immune-enhancing responses to dengue virus infection"(2012). J Virol Jun;86(12):6408-15.

Dimos J.T., Rodolfa, K.T., Niakan K.K., Weisenthal L.M., Mitsumoto H., Chung W., Croft G.F., Saphier G., Leibel R., Goland R. et al. 2008. "Induced pluripotent stem cells gener ated from patients with ALS can be differentiated into motorneurons". Science 321: 1218-1221.

Evans V.C., Barker G., Heesom K.J., Fan J., Bessant C., Matthews D.A. "De novo derivation of proteomes from transcriptomes for transcript and protein identification" (2012). Nature Methods, 9 (12), pp. 1207-1211.

Ewing B, Green P. "Base-calling of automated sequencer traces using phred". Error probabilities. Genome Res 1998,8:186-194.

Ferragina, Paolo, Manzini, Giovanni. "Opportunistic data structures with applications"(2000). Annual Symposium on Foundations of Computer Science - Proceedings, pp. 390-398.

Georges S.A., Giebler H.A., Cole P.A., Luger K., Laybourn P.J., and Nyborg J.K. (2003). "Tax recruitment of CBP/p300, via the KIX domain, reveals a potent requirement for acetyltransferase activity that is chromatin dependent and histone tail independent". Mol Cell Biol 23, 3392-3404.

Gerlach RL, Camp JV, Chu YK, Jonsson CB. "Early host responses of seasonal and pandemic influenza A viruses in primary well-differentiated human lung epithelial cells". PLoS One 2013;8(11):e78912.

Guan K., Nayernia K., Maier L.S., Wagner S., Dressel R., Lee J.H., Nolte J., Wolf F., Li M., Engel W., and Hasenfuss G. 2006. "Pluripotency of spermatogonial stem cells from adult mouse testis". Nature 440: 1199-1203.

Hach F, Hormozdiari F, Alkan C, Birol I, Eichler EE, Sahinalp SC. "mrsFAST: a cache-oblivious algorithm for short-read mapping". Nat Methods 2010, 7:576-577.
Hanna J, B.W. Carey and R. "Jaenisch Reprogramming of Somatic Cell Identity". Cold Spring Harb Symp Quant Biol 2008 73: 147-155.

Haushalter K.A., and Kadonaga J.T. (2003). "Chromatin assembly by DNA-translocating motors". Nat Rev Mol Cell Biol 4, 613-620.

Hill A.V., Whitehouse D.B., Bowden D.K., Hopkinson D.A., Draper C.C., Peto T.E., Clegg J.B. and Weatherall D.J. (1987) "Ahaptoglobinaemia in Melanesia: DNA and malarial antibody studies". Trans. R. Soc. Trop. Med. Hyg., 81, 573-577.

Hillier L.W., Marth G.T., Quinlan A.R., Dooling D., Fewell G., Barnett D., Fox P., Glasscock J.I., Hickenbotham M., Huang W., Magrini V.J., Richt R.J., Sander S.N., Stewart D.A., Stromberg M., Tsung E.F., Wylie T., Schedl T., Wilson R.K., Mardis E.R. "Whole-genome sequencing and variant discovery in C. elegans"(2008). Nature Methods, 5 (2), pp. 183-188.

Hu C. M., S. Y. Jang, J. C. Fanzo and A. B. Pernis. 2002. "Modulation of T cell cytokine production by interferon regulatory factor-4". The Journal of Biological Chemistry 277:49238-49246.

Irizarry R.A., Hobbs B., Collin F., Beazer-Barclay Y.D., Antonellis K.J., Scherf U., Speed T.P. "Exploration, normalization, and summaries of high density oligonucleotide array probe level data"(2003). Biostatistics (Oxford, England), 4 (2), pp. 249-264.

Jaenisch R. 2004. "Human cloning-The science and ethics of nuclear transplantation". N. Engl. J. Med. 351: 2787-2791.

Jepson A.P., Banya W.A., Sisay-Joof F., Hassan-King M., Bennett S. and Whittle H.C. (1995) "Genetic regulation of fever in Plasmodium falciparum malaria in Gambian twin children". J. Infect. Dis., 172, 316-319.

Jubier-Maurin V., R. A. Boigegrain, A. Cloeckaert, A. Gross, M. T. Alvarez-Martinez, A. Terraza, J. Liautard, S. Kohler, B. Rouot, J. Dornand and J. P. Liautard. 2001. "Major Outer Membrane Protein Omp25 of Brucella suis is involved in inhibition of tumor necrosis factor alpha production during infection of human macrophages". Infection and Immunity 69:4823-4830.

Kim J., Woo A.J., Chu J., Snow J.W., Fujiwara Y., Kim C.G., Cantor A.B., Orkin S.H. "A Myc Network Accounts for Similarities between Embryonic Stem and Cancer Cell Transcription Programs" (2010). Cell, 143 (2), pp. 313-324.

Kohler S., V. Foulongne, S. Ouahrani-Bettache, G. Bourg, J. Teyssier, M. Ramuz and J. P. Liautard. 2002. "The analysis of the intramacrophagic virulome of Brucella suis deciphers the environment encountered by the pathogen inside the macrophage host cell". PNAS 99:15711-15716.

Kulesh DA, Clive DR, Zarlenga DS, Greene JJ (1987). "Identification of interferon-modulated proliferation-related cDNA sequences". Proc. Natl. Acad. Sci. U.S.A. 84 (23): 8453-8457.

Langmead B., Trapnell C., Pop M., Salzberg S.L. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome" (2009). Genome Biology, 10 (3), art. no. R25.

Lapaque N., I. Moriyon, E. Moreno and J. P. Gorvel. 2005. "Brucella lipopolysaccharide acts as a virulence factor". Current Opinion in Microbiology 8:60-66.

Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, Brown PO, Davis RW (1997). "Yeast microarrays for genome wide parallel genetic and gene expression analysis". Proc. Natl. Acad. Sci. U.S.A. 94 (24): 13057-13062.

Lazzaro B.P., Sceurman B.K. and Clark A.G. (2004). "Genetic basis of natural variation in D. melanogaster antibacterial immunity". Science, 303, 1873-1876.

Lestrate P., A. Dricot, R. M. Delrue, C. Lambert, V. Martinelli, X. De Bolle, J. J. Letesson and A. Tibor. 2003. "Attenuated signature-tagged mutagenesis mutants of Brucella melitensis identified during the acute phase of infection in mice". Infection and Immunity 71:7053-7060.

Li H., Ruan J., Durbin R. "Mapping short DNA sequencing reads and calling variants using mapping quality scores"(2008). Genome Research, 18 (11), pp. 1851-1858.

Li Q, Pène V, Krishnamurthy S, Cha H et al. "Hepatitis C virus infection activates an innate pathway involving IKK-α in lipogenesis and viral assembly"(2013). Nat Med Jun;19(6):722-9

Luger K. et al. (1997). "Crystal structure of the nucleosome core particle at 2.8 A resolution". Nature 389(6648): 251-60.

Marko N. F., B. Frank, J. Quackenbush and N. H. Lee. 2005. "A robust method for the amplification of RNA in the sense orientation". BMC Genomics 6:27. marrow. Eur. J. Clin. Pharmacol. 56:405 - 409, 2000.

McBryant S.J., Park Y.J., Abernathy S.M., Laybourn P.J., Nyborg J.K. and Luger K. (2003). "Preferential binding of the histone (H3-H4)2 tetramer by NAP1 is mediated by the amino-terminal histone tails". J Biol Chem 278, 44574-44583.

Mortazavi A, Williams BA, Kenneth McCue, Lorian Schaeffer and Barbara Wold. "Mapping and quantifying mammalian transcriptomes by RNA-Seq". Nature Methods, Volume 5, 621 - 628 (2008) .

Nagalakshmi U., Wang Z., Waern K., Shou C., Raha D., Gerstein M., Snyder M. "The transcriptional landscape of the yeast genome defined by RNA sequencing"(2008). Science, 320 (5881), pp. 1344-1349.

Newport M.J., Huxley, C.M., Huston, S., Hawrylowicz, C.M., Oostra, B.A., Williamson R. and Levin M. (1996) "A mutation in the interferon-gamma-receptor gene and susceptibility to mycobacterial infection". N. Engl. J. Med., 335, 1941-1949.

Nissim O, Melis M, Diaz G, Kleiner DE et al. "Liver regeneration signature in hepatitis B virus (HBV)-associated acute liver failure identified by gene expression profiling"(2012). PLoS One 2012;7(11):e49611.

Njau F., Geffers R., Thalmann J., Haller H., Wagner A.D. "Restriction of Chlamydia pneumoniae replication in human dendritic cell by activation of indoleamine 2,3-dioxygenase"(2009). Microbes and Infection, 11 (13), pp. 1002-1010.

Pepper S.D., Saunders E.K., Edwards L.E., Wilson C.L., Miller C.J."The utility of MAS5 expression summary and detection call algorithms" (2007) BMC Bioinformatics, 8, art. no. 273.

Quinlan A.R., Hall I.M. "BEDTools: A flexible suite of utilities for comparing genomic features"(2010). Bioinformatics, 26 (6), art. no. btq033, pp. 841-842.

Samarajiwa S.A., Forster S., Auchettl K., Hertzog P.J. "INTERFEROME: The database of interferon regulated genes"(2009). Nucleic Acids Research, 37 (SUPPL. 1), pp. D852-D857.

Schena M, Shalon D, Davis RW, Brown PO (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray". Science 270 (5235): 467-470.

Schmolke M, Viemann D, Roth J, Ludwig S. "Essential impact of NF-kappaB signaling on the H5N1 influenza A virus-induced transcriptome"(2009). J Immunol Oct 15;183(8):5180-9.

Shin H., Liu T., Manrai A.K., Liu S.X. "CEAS: Cis-regulatory element annotation system"(2009). Bioinformatics, 25 (19), pp. 2605-2606 .

Statistical Algorithms Description Document 2002. Affymetrix.

Statistical Algorithms Reference Guide. Affymetrix.

Stephen G. Landt, Georgi K. Marinov, Anshul Kundaje, et al 2012."ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia". Genome Res. 2012 22: 1813-1831.

Sutejo R, Yeo DS, Myaing MZ, Hui C et al. "Activation of type I and III interferon signalling pathways occurs in lung epithelial cells infected with low pathogenic avian influenza viruses". PLoS One 2012;7(3):e33732.

The ENCODE Project Consortium. "An integrated encyclopedia of DNA elements in the human genome"(2012). Nature 489, 57-74.

Trapnell C, Hendrickson D,Sauvageau S, Goff L, Rinn JL, Pachter L. "Differential analysis of gene regulation at transcript resolution with RNA-seq"(2013). Nature Biotechnology doi:10.1038/nbt.2450.

Trapnell C., Williams B.A., Pertea G., Mortazavi A., Kwan G., Van Baren M.J., et al. "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation"(2010). Nature Biotechnology, 28 (5), pp. 511-515.

Trapnell, C., Pachter, L., Salzberg, S.L. "TopHat: Discovering splice junctions with RNA-Seq"(2009). Bioinformatics, 25 (9), pp. 1105-1111.

Viemann D., Schmolke M., Lueken A., Boergeling Y., Friesenhagen J., Wittkowski H., Ludwig S., Roth J. "H5N1 virus activates signaling pathways in human endothelial cells resulting in a specific imbalanced inflammatory response"(2011). Journal of Immunology, 186 (1), pp. 164-173.

Wan Z, Zhi N, Wong S, Keyvanfar K et al. "Human parvovirus B19 causes cell cycle arrest of human erythroid progenitors via deregulation of the E2F family of transcription factors(2010). J Clin Invest Oct;120(10):3530-44.

Wang ET, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. "Alternative isoform regulation in human tissue transcriptomes". Nature, Volume 456, 470 - 476 (2008).

Wang J., Nikrad M.P., Travanty E.A., Zhou B.,et al. "Innate immune response of human alveolar macrophages during influenza a infection"(2012). PLoS ONE, 7 (3), art. no. e29879.

Yoshikawa T, Hill TE, Yoshikawa N, Popov VL et al. "Dynamic innate immune responses of human bronchial epithelial cells to severe acute respiratory syndrome-associated coronavirus infection"(2010). PLoS One Jan 15;5.

Zhang Y., Liu T., Meyer C.A., Eeckhoute J., Johnson D.S., Bernstein B.E., Nussbaum C., Myers R.M., Brown M., Li W., Shirley X.S. "Model-based analysis of ChIP-Seq (MACS)"(2008). Genome Biology, 9 (9), art. no. R137.

Zhang Y., Liu T., Meyer C.A., Eeckhoute J., Johnson D.S., Bernstein B.E., Nussbaum C., Myers R.M., Brown M., Li, W., Shirley X.S. "Model-based analysis of ChIP-Seq (MACS)" (2008). Genome Biology, 9 (9), art. no. R137.