ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
Σχολή Ηλεκτρονικών Μηχανικών και Μηχανικών Υπολογιστών

Διπλωματική Εργασία

Αναγνώριση και Κατηγοριοποίηση των κυκλοφορούντων καρκινικών κυττάρων μέσω χρήσης μεταγραφικών δεδομένων (Identification and classification of circulating tumor cells using transcriptome data)

Κοτρωνιά Μαρία

Επιβλέπων Καθηγητής :Καθηγητής Ζερβάκης Μιχάλης

Εξεταστική Επιτροπή: Καθηγητής Ζερβάκης Μιχάλης

Καθηγητής Μπάλας Κωνσταντίνος

Καθηγητής Γαροφαλάκης Μίνως

Χανιά, Ιούλιος 2015

Ευχαριστίες

Θα ήθελα να ευχαριστήσω

Τον καθηγητή κύριο Ζερβάκη Μιχάλη, για την βοήθεια και αμέριστη στήριξη του καθόλη τη διάρκεια της διπλωματικής μου εργασίας.

Τους καθηγητές κύριο Γαρφαλάκη Μίνω και κύριο Μπάλα Κωνσταντίνο για τη συνεισφορά τους ως μέλη της εξεταστικής επιτροπής.

Τον κ. Καφετζόπουλο Δημήτρη για την παροχή των πειραματικών δεδομένων.

Την υποψήφια διδάκτορα Καλαντζάκη Καλλιόπη για τις χρήσιμες επιστημονικές συμβουλές της , την συνεργασία της και την αμέριστη στήριξή της.

Τον μετάδιδάκτορα ερευνητή Αρχοντάκη Στέλιο για την πολύτιμη βοήθεια του και συνεισφορά του , αλλά και για την στήριξη που μου παρείχε κατά την διάρκεια της εκπόνησης της διπλωματική μου.

Την οικογένεια μου και όλους μου τους φίλους που με στήριξαν πολύ καθόλη τη διάρκεια της φοιτητικής μου ζωής

Περιεχόμενα

Περίληψη 9
Abstract 10
Κεφάλαιο 1 - ΕΙΣΑΓΩΓΗ 12
1.1 Ανάλυση του προβλήματος 12
1.2 Περιοχή της εργασίας 13
1.3 State of the art 14
1.4 Καινοτομία της εργασίας 18
Κεφάλαιο 2 - ΒΙΟΛΟΓΙΚΟ ΥΠΟΒΑΘΡΟ 19
2.1 Μοριακή Βιολογία 19
2.1.1 Το κύτταρο και η δομή του 19
2.1.2 Κυτταρικές Σειρές και κυτταρικές καλλιέργειες 21
2.1.3 Εισαγωγή στη Μοριακή Βιολογία 22
2.1.4 Το γονίδιο 24
2.1.5 Γονιδιακή Έκφραση στο Κύτταρο 25
2.2 Τι είναι ο καρκίνος 26
2.3 Κυκλοφορούντα καρκινικά κύτταρα (CTCs- Circulating tumor cells) 29
2.4 Βιοπληροφορική 31
2.5 Μικροσυστοιχίες 32
2.5.1 Μικροσυστοιχίες και γονιδιακή έκφραση 32
2.5.3 Μικροσυστοιχίες DNA 34
2.5.5 GeneChips U133 & Prime View 39
Κεφάλαιο 3 - ΑΝΑΛΥΤΙΚΟ ΥΠΟΒΑΘΡΟ 40
3.4 Ανάλυση Μεθοδολογιών 43
3.4.1 Ο Αλγόριθμος RMA 43
3.4.2 Η μέθοδος Κανονικοποίησης Lowess 45
3.4.3 MvA plots 46
3.4.4 Συντελεστής Συσχέτισης (Correlations Coefficient) 47
3.4.5 Global Normalization 47
3.4.6 Clustering (Ομαδοποιήση δεδομένων) 51
3.4.6.1 Hierarchical clustering 52
3.4.6.2 Heat Map 53

Κεφάλαιο 4 - ΠΡΟΤΕΙΝΟΜΕΝΗ ΜΕΘΟΔΟΛΟΓΙΑ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ.....54
4.1 Δεδομένα προς επεξεργασία.....54
4.2 Block Diagram.....57
4.3 Αντιστοίχιση Genes Symbols - Probes στα GeneChips U133 Plus 2.0 και Prime View.....59
4.4 Εφαρμογή RMA.....60
4.5 Εφαρμογή μεθόδου Averaging.....61
4.6 Εφαρμογή της μεθόδου κανονικοποίησης Lowess.....63
4.7 Επιβεβαίωση αποτελεσμάτων μέσω των MvA plots.....66
4.8Επιβεβαίωση αποτελεσμάτων μέσω του συντελεστή συσχέτισης (Correlations Coefficient).....67
4.9 Εφαρμογή της μεθόδου Global Normalization.....70
4.10 Φιλτράρισμα γονιδίων για την κατηγοριοποίηση των δειγμάτων με χρήση διακύμανσης.....73
4.11Εφαρμογή μεθόδων κατηγοριοποίησης και οπτικοποίησης αποτελεσμάτων.....76
Κεφάλαιο 5 - ΥΛΟΠΟΙΗΣΗ ΑΛΓΟΡΙΘΜΩΝ.....84
5.1 Πλατφόρμες λογισμικού επεξεργασίας βιολογικών δεδομένων.....84
5.2 Επεξεργασία μεγάλου όγκου βιολογικών δεδομένων.....85
5.3 Μεθοδολογία ανάπτυξης αλγορίθμων.....89
Κεφάλαιο 6 - ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΟΝΤΙΚΕς ΕΠΕΚΤΑΣΕΙΣ.....93
6.1 Συμπεράσματα.....93
6.2 Μελλοντικές επεκτάσεις.....94
ΒΙΒΛΙΟΓΡΑΦΙΑ.....95
ΠΑΡΑΡΤΗΜΑΤΑ.....102
Παράρτημα Α.....103
Α.1 - Οι γραφικές παραστάσεις των MvA plots για όλα τα δείγματα των κυτταρικών σειρών, πριν και μετά την κανονικοποίηση LOWESS.....103
Α.2 - Οι πίνακες με τους συντελεστές συσχέτισης για όλα τα δείγματα των κυτταρικών σειρών, πριν και μετά την κανονικοποίηση LOWESS.....120
Α.3 – Οι γραφικές παραστάσεις όλων των δειγμάτων της Κυτταρικής σειράς 4 πριν και μετά την Lowess.....122
Παράρτημα Β.....129

Λίστα Εικόνων

Εικόνα 1. Διάγραμμα ροής εργασιών για την ανάλυση δεδομένων μικροσυστοιχιών της εταιρείας Affymetrix.....15
Εικόνα 2. Τυπικό ευκαριωτικό κύτταρο με τα οργανίδιά του [E.1].....20
Εικόνα 3. Η διαδικασία μιας in vitro κυτταρικής καλλιέργειας [E.2].....22
Εικόνα 4. Το μόριο του DNA [E.4].....23
Εικόνα 5. Το Κεντρικό Δόγμα της Μοριακής Βιολογίας σήμερα.....24
Εικόνα 6. Απεικόνιση γονιδίου [E.3].....25
Εικόνα 7. Διαίρεση των κυττάρων: Α-κανονική διαίρεση κυττάρων, Β- καρκινική διαίρεση κυττάρων 1-απόπτωση, 2-κατεστραμμένο κύτταρο [E.5].....27
Εικόνα 8. Η μετάσταση από τα Κυκλοφορούντα καρκινικά κύτταρα ή τις μικροεμβολές [E.6].....29
Εικόνα 9. Slide Array [E.7].....33
Εικόνα 10. Τα βήματα ενός πειράματος με μικροσυστοιχίες cDNA [E.8].....36
Εικόνα 11. GeneChip Affymetrix [E.9].....37
Εικόνα 12. Τα βήματα κατά την εκτέλεση ενός πειράματος με το GeneChip της Affymetrix [E.8].....38
Εικόνα 13. Fitting με την μέθοδο lowess για διάφορους παράγοντες εξομάλυνσης [E.10].....45
Εικόνα 14. Διάγραμμα ροής (flowchart) όλων των τύπων κανονικοποίησης.....49
Εικόνα 15. Παράδειγμα - Διαφορετικοί τύποι Λευχαιμίας. Το Clustering βασίστηκε σε 150 γονίδια με την highest variance κατά μήκος των δειγμάτων.....52
Εικόνα 16. Αναπαράσταση Heat map , με δεδομένα από DNA microarray [E.10].....53
Εικόνα 17. Η μεθοδολογία μας διαγραμματικά.....57
Εικόνα 18. Το διάγραμμα με όλα τα στάδια της μεθόδου που ακολουθήσαμε αναλυτικά.....58
Εικόνα 19. Συσχετισμός 18034 κοινών γονιδίων από τα Chips U133 και PrimeView της εταιρίας Affymetrix.....59
Εικόνα 20. Η διαδικασία όπου τα Cell Files επεξεργάζονται με τον RMA.....61
Εικόνα 21. Η αντιστοίχιση των Probes στα αντίστοιχα Gene Symbols με την μέθοδο Average.....62
Εικόνα 22. Εφαρμογή κανονικοποίησης Lowess μόνο στα δείγματα που έχουν reference και η δημιουργία Expression Universe με Reference Dataset και Reference plus Query Dataset.....64
Εικόνα 23. MvA plot δείγματος 3cell_line_1c της Cell Line 3 (Κυτταρικής Σειράς 3) με παράθυρο κανονικοποίησης στη Lowess 0,06.....65
Εικόνα 24. MvA plot δείγματος 3cell_line_1c της Cell Line 3 (Κυτταρικής Σειράς 3) με παράθυρο κανονικοποίησης στη Lowess 0,05.....65
Εικόνα 25. Παράδειγμα ma plot Κυτταρικής Σειράς 4.....67
Εικόνα 26. Plot δείγματος 5000_Eb-opt , με το αντίστοιχο bulk (MCF7) , πριν την Lowess.....69

Εικόνα 27. Plot δείγματος 5000_Eb-opt , με το αντίστοιχο bulk (MCF7) , μετά την Lowess 69
Εικόνα 28. Regression lines για $y=g_x(x)$ [κόκκινο] και $x=g_y(y)$ [μπλέ] [E.11] 70
Εικόνα 29. Διαδικασία Global Normalization 71
Εικόνα 30. Το στάδιο της Global Normalization στην προτεινόμενη μεθοδολογία ... 72
Εικόνα 31. Η μέθοδος με τις Variances για το expression universe 1 74
Εικόνα 32. Η κανονικοποίηση με χρήση της Διακύμανσης και το φιλτράρισμα των γονιδίων 75
Εικόνα 33. Το τελευταίο στάδιο της μεθοδολογίας μας, Hierarchical clustering , Heat maps 77
Εικόνα 34. Heat map για το Expression Universe 1 με 18034 γονίδια 78
Εικόνα 35. Στιγμιότυπο του Heat map για το Expression Universe 1 με 18034 γονίδια 78
Εικόνα 36. Heat map για το Expression Universe 1 με 11000 γονίδια 79
Εικόνα 37. Στιγμιότυπο του Heat map για το Expression Universe 1 με 11000 γονίδια 79
Εικόνα 38. Στιγμιότυπο του Heat map για το Expression Universe 1 με 11000 γονίδια 80
Εικόνα 39. Heat map για το Expression Universe 1 με 1400 γονίδια , με highest Variance 81
Εικόνα 40. Στιγμιότυπα του Heat map για το Expression Universe με 1400 γονίδια , με highest Variance. Κατηγοριοποίηση Cell Line 2,3 81
Εικόνα 41. Στιγμιότυπο του Heat map για το Expression Universe με 1400 γονίδια , με highest Variance. Απεικόνιση κατηγοριοποίησης του δείγματος Cell line 5 bulk 82
Εικόνα 42. Στιγμιότυπο του Heat map για το Expression Universe με 1400 γονίδια , με highest Variance. Κατηγοριοποίηση Cell Line 1 82
Εικόνα 43. Στιγμιότυπο του Heat map για το Expression Universe με 1400 γονίδια , με highest Variance. Κατηγοριοποίηση Cell Line 4 με δείγματα μαστού 83
Εικόνα 44. αύξηση απόδοσης μέσω task priority settings 86
Εικόνα 45. Εικόνες αύξησης της απόδοσης μέσω της χρήσης πολλών πυρήνων. 88

Λίστα Πινάκων

Πίνακας 1. Βασικές ιδιότητες μεθόδων ταξινόμησης που εφαρμόζονται στα δεδομένα γονιδιακής έκφρασης για την εκτίμηση προτύπων. [88] 17
Πίνακας 2. Λίστα με τις διάφορες μεθόδους κανονικοποίησης 50
Πίνακας 3. Οι Κυτταρικές Σειρές που επεξεργαστήκαμε στην παρούσα εργασία 54
Πίνακας 4. Στοιχεία των Βάσεων Δεδομένων που χρησιμοποιήθηκαν για την ταυτοποίηση των Κυτταρικών Σειρών. 56
Πίνακας 5. Στοιχεία για Probes U133 και Prime View 60
Πίνακας 6. Παράδειγμα συντελεστές συσχέτισης για την μέθοδο αντιστοίχισης των Probes με Average και Max. 63
Πίνακας 7. Οι τιμές του συντελεστή συσχέτισης για την Κυτταρική Σειρά 1 68

Περίληψη

Παρά την τεχνολογική πρόοδο που έχει επιτευχθεί για την ανάλυση της γονιδιακής έκφρασης με την χρήση Μικροσυστοιχιών DNA και την επιτυχή εξαγωγή βιολογικών συμπερασμάτων, υπάρχει σημαντική έλλειψη τυποποίησης σε μεθόδους κανονικοποίησης για την αφαίρεση του συστηματικού θορύβου. Ο τελευταίος, υφισέρχεται στα δεδομένα κατά τη διάρκεια της πειραματικής διαδικασίας, ή της έλλειψης τυποποίησης των μεθόδων κατηγοριοποίησης δεδομένων, καθώς και της δυσκολίας στην διαχείριση μεγάλου όγκου δεδομένων που παράγεται από τη χρήση Μικροσυστοιχιών. Για τις ανάγκες αυτής της διπλωματικής εργασίας αναπτύχθηκε ένα πακέτο λογισμικού στην πλατφόρμα Matlab, ικανό να διαχειριστεί έναν μεγάλο όγκο βιολογικών δεδομένων από έναν προσωπικό υπολογιστή. Το λογισμικό προσφέρει την δυνατότητα στο χρήστη να επιλέξει τη μέθοδο που επιθυμεί για την ανάλυση των βιολογικών δεδομένων μέσα από μία ποικιλία αλγορίθμων κανονικοποίησης και κατηγοριοποίησης, καθώς και ανάπτυξης ιδιοκατασκευασμένων μεθόδων. Τα δεδομένα αποθηκεύονται και διαχειρίζονται από ένα εξειδικευμένο σύστημα διαχείρισης βάσης δεδομένων, και δοκιμάστηκαν στην ανάλυση πληθώρας βιολογικών δειγμάτων με ιδιαίτερη επιτυχία. Η παρούσα διπλωματική εργασία αποτελεί τη βάση για τη δημιουργία ενός καινοτόμου προγραμματιστικού εργαλείου διαχείρισης και επεξεργασίας βιολογικών δεδομένων, που θα χρησιμοποιηθεί σε βιολογικές έρευνες, με σκοπό αφενός την παραγωγή αξιόπιστων και συγκρίσιμων αποτελεσμάτων από διαφορετικές πλατφόρμες μικροσυστοιχιών και αφετέρου την ελευθερία επιλογής ανάμεσα από μία πλήρη γκάμα στατιστικών μεθόδων ανάλυσης.

Abstract

Despite recent advances in Microarray technology towards gene expression analysis and extraction of biological significance indices, the successful use of this technology is still elusive for many researchers. This is mainly because there is no standardization yet in methods used for the normalization of systematic noise which occurs during experimental procedures or in methods used for the biological data classification. Also, the analysis of large amount of data produced by such experiments remains a significant challenge. The objective of this diploma thesis was the development of a software platform in Matlab, able of managing and analyzing large amounts of biological data through a single personal computer. The software offers the potential to the user to choose the desired method for data analysis through a variety of normalization and classification algorithms, as well as to develop and integrate custom methods. Data is managed and stored by a specialized database management system. The software platform was tested in a plethora of biological samples with great success. This thesis is intended to serve in biological research as the basis of an innovative and complete software tool for the management and processing of large amounts of biological data in order both to produce reliable and comparable results from different microarray experiments and also to offer potential of choosing between a complete range statistical analysis methods.

Κεφάλαιο 1 - Εισαγωγή

1.1 Ανάλυση του προβλήματος

Το αντικείμενο της παρούσας διπλωματικής εργασίας αφορά την ανάλυση δεδομένων γονιδιακής έκφρασης που προκύπτουν μέσω της χρήσης μικροσυστοιχιών. Πέραν της θεωρητικής προσέγγισης του ζητήματος, υλοποιήθηκε μία σειρά επεξεργασιών με χρήση πραγματικών δεδομένων , εφαρμόστηκαν δοκιμασμένες αλλά και πρωτότυπες μέθοδοι στατιστικής ανάλυσης και υλοποιήθηκε μία προγραμματιστική πλατφόρμα που εμπεριέχει τόσο τους στατιστικούς αλγορίθμους ανάλυσης, όσο και τους προγραμματιστικούς αλγορίθμους διαχείρισης μεγάλου όγκου βιολογικών δεδομένων γονιδιακής έκφρασης από μικροσυστοιχίες.

Η χρήση των μικροσυστοιχιών επιτρέπει την εξέταση των επιπέδων έκφρασης mRNA δεκάδων χιλιάδων γονιδίων ταυτόχρονα σε έναν ιστό ή μια συνθήκη. Αυτή η πρόοδος στον τομέα της βιοϊατρικής έχει δώσει την ευκαιρία μελέτης ολόκληρου του γονιδιώματος με ένα πείραμα, όπου νωρίτερα ήταν δυνατή η μελέτη της έκφρασης ελαχίστων γονιδίων μόνο. Αρχικά, παράγοντες όπως η χαμηλή ποιότητα των δεδομένων και το υψηλό κόστος εξοπλισμού περιόρισαν την έρευνα πάνω στις μικροσυστοιχίες σε ελάχιστα ερευνητικά κέντρα. Όμως, οι βελτιώσεις που έγιναν σταδιακά στο υλισμικό (hardware) και στη συναφή τεχνολογία, έχουν αυξήσει όχι μόνο την ποιότητα και την επαναληψιμότητα των δεδομένων, αλλά πέτυχαν και σημαντική μείωση του κόστους, καθιστώντας την έρευνα στις μικροσυστοιχίες προσιτή στους περισσότερους ερευνητές.

Παρότι όμως η τεχνολογική πρόοδος κατέστησε τις μικροσυστοιχίες ως αξιόπιστο και σχετικά προσιτό ερευνητικό τομέα, πολλά "σημεία συμφόρησης" εξακολουθούν να υπάρχουν ανάμεσα στις πειραματικές διαδικασίες που ακολουθούν οι ερευνητές, καθώς και στην ανάλυση δεδομένων που προκύπτουν από αυτού του τύπου την έρευνα. Αρχικά, υπάρχει ελλιπής εκπαίδευση προσωπικού στα προγράμματα λογισμικού ανάλυσης δεδομένων. Επίσης συχνά το κόστος αυτών των προγραμμάτων είναι ιδιαίτερα υψηλό. Επιπρόσθετα, η ανάλυση δεδομένων μικροσυστοιχιών αντιμετωπίζεται ως ένα σύνολο ξεχωριστών σταδίων και οι ερευνητές συχνά χρησιμοποιούν πολλαπλά πακέτα λογισμικού για να ολοκληρώσουν την ανάλυση των βιολογικών δεδομένων. Ακόμα, έχει δοθεί από την επιστημονική κοινότητα μεγάλη έμφαση στην στατιστική ανάλυση και την οπτικοποίηση των δεδομένων, ενώ το σημαντικότερο κομμάτι θα έπρεπε να είναι η εξαγωγή βιολογικών παραμέτρων ή συμπερασμάτων. Εκτός αυτού, πέραν της στατιστικής ανάλυσης, η επαναληψιμότητα των αποτελεσμάτων αποτελεί ακόμα ανοιχτό θέμα συζήτησης ανάμεσα στους ερευνητές, καθώς τα στάδια της πειραματικής διαδικασίας και της προετοιμασίας των προς ανάλυση δειγμάτων εμπεριέχει υποκειμενικούς παράγοντες. Στην πραγματικότητα, θα έπρεπε η διαδικασία της ανάλυσης των δεδομένων μικροσυστοιχιών να επεκταθεί ώστε να συμπεριλαμβάνει πρόσθετες συσχετίσεις και αλληλεξαρτώμενες πτυχές, όπως ο πειραματικός σχεδιασμός, η διαχείριση των

δεδομένων, η προσβασιμότητα και η επιλογή πλατφόρμας λογισμικού. Οι συσχετίσεις αυτές, αν και δε θεωρούνται παραδοσιακά ως αναπόσπαστο μέρος της διαδικασίας ανάλυσης δεδομένων, αλλά παρόλα αυτά μπορούν να έχουν σοβαρές επιπτώσεις στη στατιστική ανάλυση.

1.2 Περιοχή της εργασίας

Η περιοχή της εργασίας εστιάζεται στη διαχείριση μεγάλου όγκου βιολογικών δεδομένων γονιδιακής έκφρασης, στην κανονικοποίηση (normalization) και στην κατηγοριοποίηση τους (classification). Ειδικότερα, αναπτύχθηκε μια ενιαία προγραμματιστική πλατφόρμα για την αποτελεσματική διαχείριση και ανάλυση χιλιάδων πειραμάτων από μικροσυστοιχίες και η εφαρμογή καινοτόμων αλγορίθμων στατιστικής επεξεργασίας γονιδιακής έκφρασης. Επίσης αναπτύχθηκαν εξειδικευμένα προγραμματιστικά εργαλεία για τη διαχείριση των βιολογικών δεδομένων που προκύπτουν από κάθε πείραμα μικροσυστοιχία, ανεξαρτήτου όγκου. Τα δεδομένα εισάγονται, επεξεργάζονται και αποθηκεύονται τμηματικά. Όλες οι διαδικασίες γίνονται με τη χρήση ενός και μόνο υπολογιστή, άσχετα από τον όγκο των δεδομένων. Παράλληλα, μέσω της πλατφόρμας Matlab, αναπτύχθηκε ένα πακέτο λογισμικού ικανό να επεξεργαστεί τα βιολογικά δεδομένα σε όλα τα απαιτούμενα στάδια, αντικαθιστώντας έτσι την ανάγκη χρήσης πολλαπλών πακέτων λογισμικού. Ακόμα, εφαρμόστηκαν καινοτόμοι αλγόριθμοι προεπεξεργασίας, με σημαντικά αποτελέσματα.

Η διπλωματική αυτή αποτελεί τη βάση για τη δημιουργία ενός καινοτόμου προγραμματιστικού εργαλείου διαχείρισης και επεξεργασίας βιολογικών δεδομένων, που θα χρησιμοποιηθεί από τους ερευνητές γονιδιακής έκφρασης, με σκοπό αφενός την παραγωγή αξιόπιστων και συγκρίσιμων αποτελεσμάτων και αφετέρου την ελευθερία επιλογής ανάμεσα από μία πλήρη γκάμα στατιστικών μεθόδων. Επιπλέον, ένας ερευνητής θα έχει και την ελευθερία να παραμετροποιήσει ή να προσαρμόσει στις ανάγκες του οποιοδήποτε στάδιο της ανάλυσης.

Η δομή της εργασίας χωρίζεται σε 5 ενότητες. Στην πρώτη ενότητα περιγράφονται οι γενικές αρχές της Μοριακής Βιολογίας και οι αρχές λειτουργίας των Μικροσυστοιχιών. Στη δεύτερη ενότητα αναλύονται τα δεδομένα εξόδου που προκύπτουν από τα πειράματα μικροσυστοιχιών και οι παράγοντες θορύβου που υφισέρχονται στα διάφορα στάδια της πειραματικής διαδικασίας. Στην τρίτη ενότητα αποσαφηνίζονται οι μέθοδοι ανάλυσης των δεδομένων γονιδιακής έκφρασης, τόσο για την εξάλειψη του θορύβου, όσο και για την εξόρυξη πληροφοριών βιολογικής σημασίας. Παράλληλα θίγονται οι ορθές προγραμματιστικές τεχνικές για την ανάλυση αυτών των δεδομένων, με έμφαση στην επεξεργασία μεγάλου όγκου πληροφοριών. Τέλος, στην πέμπτη ενότητα, παρατίθενται τα αποτελέσματα των μεθόδων, εμπλουτισμένα με σχολιασμό, συμπεράσματα και μελλοντικές επεκτάσεις.

1.3 State of the art

Η επεξεργασία βιολογικών δεδομένων περιλαμβάνει 4 στάδια:

- Σχεδιασμός πειραματικής διαδικασίας
- Διαχείριση όγκου δεδομένων
- Προ-επεξεργασία (Pre-processing)
- Εξαγωγή βιολογικών συμπερασμάτων

Για κάθε ένα από αυτά τα στάδια καταγράφονται οι κυριότερες και πιο σύγχρονες προσεγγίσεις από την επιστημονική κοινότητα.

Σχεδιασμός πειραματικής διαδικασίας

Ο σχεδιασμός της πειραματικής διαδικασίας είναι καθοριστικός για την επιτυχημένη χρήση μικροσυστοιχιών και την εξαγωγή αξιόπιστων αποτελεσμάτων. Αυτό το στάδιο συμπεριλαμβάνει τόσο τα βήματα της πειραματικής διαδικασίας, όσο και τον αριθμό των επαναλήψεων των πειραμάτων.

Ο τύπος του πειράματος επηρεάζει καθοριστικά τον τρόπο προσέγγισης της ανάλυσης των δεδομένων. Αν για παράδειγμα χρειάζεται να συγκριθούν δύο καταστάσεις τύπου "control" και "experiment", απαιτείται ένα τεστ 2 ομάδων, όπως είναι το t-test. Αν χρειάζεται να συγκριθούν πολλαπλές καταστάσεις, όπως για παράδειγμα η εξέλιξη ενός φαινομένου κατά τη διάρκεια μιας συγκεκριμένης χρονικής περιόδου(time –course experiment), τότε χρειάζεται ένα test πολλαπλών δειγμάτων, όπως το ANOVA (analysis of variance), προσαρμοσμένο στον αριθμό των προς εξέταση παραγόντων [89]. Συνήθως οι παράγοντες που εξετάζονται είναι η κατάσταση της ασθένειας, το μοντέλο του οργανισμού, φαρμακολογική επίδραση ή παρακολούθηση μετά τη θεραπεία.

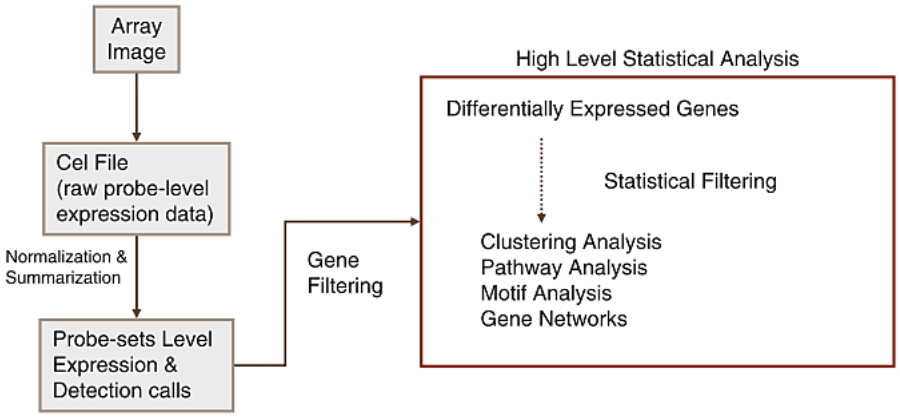
Η επανάληψη των πειραμάτων είναι απαραίτητη για τη διασφάλιση της αντικειμενικότητας των μετρήσεων και τον περιορισμό του (τυχαίου) θορύβου, μέσω της σύγκρισης των αποτελεσμάτων των επαναλήψεων των ίδιων βημάτων. Γενικά υιοθετούνται δύο τύποι επαναλήψεων, τεχνικού και βιολογικού τύπου. Οι τεχνικές επαναλήψεις εστιάζονται στις διακυμάνσεις που εμφανίζονται ως προς την τεχνική που ακολουθείται. Για παράδειγμα, ένα είδος τεχνικών επαναλήψεων θα ήταν η υβριδοποίηση του ίδιου δείγματος σε 3 διαφορετικές μικροσυστοιχίες. Οι βιολογικές επαναλήψεις εστιάζονται στις διακυμάνσεις βιολογικού τύπου, οι οποίες μπορεί να οφείλονται σε γενετικούς ή περιβαλλοντικούς παράγοντες. Ένα παράδειγμα βιολογικών επαναλήψεων είναι η εξαγωγή RNA από τους εγκεφάλους τριών ποντικιών – πειραματόζωων και η υβριδοποίηση τους σε ξεχωριστές συστοιχίες. Γενικότερα, ο αριθμός των επαναλήψεων είναι σχετικός. Συνήθως τρεις επαναλήψεις θεωρούνται ο ελάχιστος αριθμός, ενώ επαρκείς θεωρούνται από πέντε και πάνω επαναλήψεις [90].

Διαχείριση όγκου δεδομένων

Τα πειράματα με τη χρήση μικροσυστοιχιών παράγουν τεράστιο όγκο ακατέργαστων δεδομένων (raw data). Συνήθως, ένα μόνο πείραμα παράγει εκατομμύρια πειραματικά σημεία. Παράλληλα με τα δεδομένα εξόδου, απαιτείται πληροφορία για τον τύπο των δειγμάτων και τις διαδικασίες προετοιμασίας τους. Η διαχείριση αυτού του μεγάλου όγκου ανεπεξέργαστων δεδομένων είναι μία πολύ σημαντική – συχνά καθοριστική – διαδικασία. Συχνά είναι αδύνατον να ολοκληρωθεί η επεξεργασία των δεδομένων από έναν προσωπικό υπολογιστή και απαιτούνται ειδικοί υπερυπολογιστές. Επίσης προς αυτή την κατεύθυνση, η χρήση βάσεων δεδομένων διευκολύνει την οργάνωση, την αποθήκευση και την ανάκτηση πληροφοριών. Ιδανικά, τα δεδομένα θα έπρεπε να διαχειρίζονται με τέτοιο τρόπο ώστε να είναι δυνατή η ανάκτηση υποσυνόλων βασισμένη σε συγκεκριμένες τιμές έκφρασης, ειδικευμένα στατιστικά αποτελέσματα, σημειολογία (annotation) των δειγμάτων, κατ' επιλογήν ομάδες γονιδίων ή με βάση την πειραματική διαδικασία που ακολουθήθηκε. Μέχρι σήμερα, δεν έχει κατασκευαστεί μία πλατφόρμα ικανή να διαχειριστεί τα δεδομένα με αυτόν τον τρόπο, σε όλα τα στάδια.

Τα πακέτα λογισμικού που υπάρχουν στην αγορά για την επεξεργασία δεδομένων από μικροσυστοιχίες είναι πάρα πολλά. Μάλιστα, πολύ συχνά χρησιμοποιούνται διαφορετικά προγράμματα για κάθε βήμα. Το γεγονός αυτό, από τη μία πλευρά έχει καταστήσει την έρευνα πάνω σε αυτού του τύπου τα δεδομένα προσβάσιμη σε περισσότερους ερευνητές, αλλά από την άλλη εμποδίζει τη σύγκριση των αποτελεσμάτων μεταξύ των ερευνητών. Παράλληλα, υπάρχουν και διαφορετικοί τύποι προγραμμάτων όσον αφορά τον τρόπο λειτουργίας τους. Δηλαδή υπάρχουν προγράμματα τα οποία λειτουργούν σε ένα desktop PC, προγράμματα τα οποία προορίζονται για server-based συστήματα και προγράμματα που λειτουργούν online, μέσω internet ή intranet, καθένα με τα πλεονεκτήματα και τα μειονεκτήματα του.

Ανάλυση δεδομένων



Εικόνα 1. Διάγραμμα ροής εργασιών για την ανάλυση δεδομένων μικροσυστοιχιών της εταιρείας Affymetrix

Προ-επεξεργασία (Pre-processing)

Από το στάδιο των ανεπεξέργαστων δεδομένων (raw data) στην εξαγωγή συμπερασμάτων βιολογικής σημασίας, απαιτείται η στατιστική επεξεργασία των δεδομένων. Αυτό γίνεται σε δύο στάδια. Το πρώτο στάδιο, γνωστό στη βιβλιογραφία ως "Microarray Data Pre-processing", χρησιμοποιείται γενικά για την εξομάλυνση των διαφόρων τύπων θορύβου και τη διαφοροποίηση της χρήσιμης πληροφορίας από τον θόρυβο. Επίσης σε αυτό το στάδιο επεξεργάζονται τα δεδομένα από την επανάληψη του ίδιου πειράματος, ώστε να καταστούν αξιόπιστα και αντικειμενικά. Τα στάδια της προ-επεξεργασίας είναι τα εξής:

- Επεξεργασία εικόνας (Image Analysis)
- Διόρθωση θορύβου υποβάθρου (Background adjustment)
- Φιλτράρισμα (Filtering)
- Κανονικοποίηση (Normalization)

Το στάδιο της επεξεργασίας εικόνας πραγματοποιείται εντός της συσκευής της Μικροσυστοιχίας (τουλάχιστον για τα συστήματα της Affymetrix) και διορθώνει σφάλματα που προέκυψαν κατά την καταγραφή των αποτελεσμάτων. Για το σκοπό αυτό υιοθετούνται τεχνικές signal averaging,ανιχνευτές ελέγχου(control probes), κλπ. Στη συνέχεια, για τη διόρθωση του θορύβου υποβάθρου χρησιμοποιούνται ειδικοί αλγόριθμοι, με πιο διαδεδομένο τον αλγόριθμο RMA [91]. Έπειτα στο σύνολο των δεδομένων εφαρμόζονται τεχνικές κανονικοποίησης (Normalization), για την επιπλέον διόρθωση των ενδο-εργαστηριακών και δια-εργαστηριακών διαφορών και για την αφαίρεση πειραματικού σφάλματος από τα δεδομένα μας (κεφ. 4.6 και 4.9). Ακόμα, πραγματοποιείται μία διαδικασία φιλτραρίσματος, για την αφαίρεση των προβληματικών δεδομένων , που εμποδίζουν τη σωστή κατηγοριοποίηση των δεδομένων (κεφ. 4.10).Οι διαδικασίες αυτές είναι αναμφισβήτητα απαραίτητες, αλλά οι ακριβείς μέθοδοι ή ο βαθμός στον οποίο χρησιμοποιούνται δεν είναι καθορισμένα γενικώς. Διαφορετικοί ερευνητές, διαφορετικές τεχνολογίες, διαφορετικές πειραματικές διαδικασίες και διαφορετικές τεχνικές γενικότερα, καθιστούν το στάδιο της προ-επεξεργασίας αντί για μια τυποποιημένη διαδικασία, μια ειδική διαδικασία σε κάθε περίπτωση.

Βιολογική σημασία

Η εξαγωγή χρήσιμων βιολογικών συμπερασμάτων πραγματοποιείται μέσα από την αναζήτηση προτύπων (patterns) ανάμεσα στα δεδομένα. Η ανάλυση βάσει ταξινομητών (cluster analysis) είναι μια διερευνητική τεχνική που χρησιμοποιείται για να «αποκαλύψει» κλάσεις ή ομάδες γονιδίων ή δειγμάτων που λειτουργούν συνδετικά ή ομαδικά κατά τη διάρκεια μιας βιολογικής διεργασίας. Η (αν)ομοιότητα καθορίζεται από το αποτέλεσμα μιας μετρικής απόστασης μεταξύ των διανυσμάτων κάθε ζεύγους γονιδίων – δειγμάτων. Με βάση αυτή την απόσταση, στη συνέχεια, μέσω της εφαρμογής στατιστικών αλγορίθμων ομαδοποίησης (clustering) τα αντίστοιχα ζεύγη κατηγοριοποιούνται στους ίδιους clusters, εφόσον το προφίλ έκφρασης είναι παρόμοιο. Υπάρχει πληθώρα αλγορίθμων κατηγοριοποίησης, με πολλά περιθώρια

παραμετροποίησης. Η επιλογή διαφορετικού αλγορίθμου ή παραμετροποίησης του ιδίου, δύναται να οδηγήσει σε διαφορετικά αποτελέσματα κατηγοριοποίησης.

Υπάρχει πληθώρα ταξινομητών που μπορεί να χρησιμοποιήσει κανείς. Κάθε προσέγγιση διαφέρει ως προς την πολυπλοκότητα, την υπολογιστική ισχύ, την αναγκαιότητα ή όχι αρχικής γνώσης για τα δεδομένα (a priori knowledge). Η πιο άμεση και συνηθισμένη προσέγγιση είναι αυτή των unsupervised ταξινομητών βάσει κάποιου κριτηρίου ομοιότητας, συνήθως μια μετρική απόστασης. Άλλες προσεγγίσεις αφορούν την αναζήτηση κεντροειδών ανάμεσα στις κλάσεις ταξινόμησης, την εφαρμογή μεθόδων νευρωνικών δικτύων, τη διαχωριστική ανάλυση του Fischer (Fisher's Linear Discriminant Analysis - FLDA) και άλλες. Στον πίνακα 1 παρατίθεται μία συγκριτική ανάλυση των μεθόδων ταξινόμησης για δεδομένα γονιδιακής έκφρασης.

Πίνακας 1. Βασικές ιδιότητες μεθόδων ταξινόμησης που εφαρμόζονται στα δεδομένα γονιδιακής έκφρασης για την εκτίμηση προτύπων. [88]

Method	Important Properties	Comput. Cost	Transparency	Built-in FS?	Can report PS?	Generalizability
k-NN	Based on simple concept of similarity, as such metric dependent. Once metric is chosen, implementation independent. Very robust, frequently used as a benchmark for other classifiers.	Low	Low	No	No	Good
SVM	The complexity of the classifier is based on the number of support vectors rather than the dimensionality of the feature space. This makes the algorithm less prone to over-fitting.	High	Low	Yes	No	Good
CART	Decisions/splits at nodes are binary, hence decision boundaries are parallel to the feature axes, as such they are intrinsically suboptimal. CART is also frequently used as a benchmark for other classifiers.	Low	High	Yes	No	Can easily be over-trained (pruning may alleviate the problem)
NSC (PAM)	A variation of the nearest-mean classifier.	Low	High	Yes	Yes	Good
Neural Networks	Not transparent. Black box approach.	High	Low	Yes (MLP)	Yes (MLP)	Can easily be over-trained
FLDA	Collapses/projects all the features onto optimal axes, on which class separation (defined by BCV/WCV) is maximum	Low	Low	Yes	Yes	Bad, if n is too small compared to p, i.e., for $n << p$
DLDA	Assumes that the features are independent/uncorrelated. Very sensitive to low n. All the classes assumed to have the same covariance matrix, hence the decision boundary is a hyper-plane. Cannot be designed when $p > n$, so the initial p should be reduced using FS, before this approach can be applied	Low	Low	No	Yes	Bad, if n is too small compared to p, i.e., for $n << p$
DQDA	It differs from DLDA only in that different classes assumed to have different covariance matrices, hence the decision surface is a hyper-quadratic.	Low	Low	No	Yes	Bad, if n is too small compared to p, i.e., for $n << p$

1.4 Καινοτομία της εργασίας

Όπως είναι κατανοητό, παρά την αναμφισβήτητη πρόοδο, ο τομέας της επεξεργασίας βιολογικών δεδομένων από μικροσυστοιχίες έχει ακόμα πολλά ανοιχτά θέματα, στα οποία η επιστημονική κοινότητα καλείται να δώσει απαντήσεις, αλλά και να εγκαθιδρύσει διεθνή πρότυπα.

Προς αυτή την κατεύθυνση, η παρούσα διπλωματική εργασία κλήθηκε να συνεισφέρει στο μεγάλο αυτό πρόβλημα. Η εργασία είναι εστιασμένη στο κομμάτι της διαχείρισης των δεδομένων και της ανάλυσης τους. Συγκεκριμένα αναπτύχθηκε μία ενιαία προγραμματιστική πλατφόρμα που διαχειρίζεται τα αρχεία εξόδου των μικροσυστοιχιών, ανεξαρτήτου όγκου, στην αρχική, ανεπεξέργαστη μορφή τους και τα επεξεργάζεται μέχρι και το τελικό στάδιο της εξαγωγής, κατηγοριοποίησης και αναγνώρισης των clusters και των heatmaps.

Η καινοτομία της εργασίας εστιάζεται σε τρία σημεία. Πρώτον, έγινε δυνατή η επεξεργασία αρχείων μικροσυστοιχιών σε έναν προσωπικό υπολογιστή, ανεξάρτητα από τον όγκο των δεδομένων. Μέχρι τώρα, ένας προσωπικός υπολογιστής, χρησιμοποιώντας μια σειρά ειδικών προγράμματων λογισμικού που επεξεργάζονται τέτοιου είδους δεδομένα, μπορούσε να διαχειριστεί περιορισμένο όγκο δεδομένων δειγμάτων. Για μεγαλύτερο αριθμό δειγμάτων έπρεπε να χρησιμοποιηθούν ειδικοί υπερυπολογιστές, οι οποίοι όχι μόνο είναι ακριβοί και δυσεύρετοι, αλλά απαιτούν και ιδιαίτερα εξειδικευμένες προγραμματιστικές τεχνικές και τεχνογνωσίες (π.χ Hadoop). Εμείς, επεξεργαστήκαμε με επιτυχία πειραματικά δεδομένα έως και 10000 δειγμάτων με τη χρήση ενός φορητού υπολογιστή μεσαίων επιδόσεων, μέσα σε 20 ώρες.

Δεύτερον, η προγραμματιστική πλατφόρμα είναι δομημένη ούτως ώστε ένας ερευνητής να μπορεί υλοποιήσει όλα τα στάδια προεπεξεργασίας και ανάλυσης των δεδομένων μέσα από ένα πακέτο λογισμικού. Επίσης, μέσα από αυτήν την πλατφόρμα λογισμικού είναι δυνατή η παραμετροποίηση κάθε βήματος με μεγάλη ελευθερία. Δηλαδή, για παράδειγμα ο ερευνητής είναι ελεύθερος να επιλέξει όποια μέθοδο κανονικοποίησης ή clustering επιθυμεί, ή ακόμα να υλοποιήσει μία δική του ιδέα σε οποιοδήποτε στάδιο της διαδικασίας.

Τρίτον, στο στάδιο της προ – επεξεργασίας υλοποιήθηκαν δύο καινούργιες προσεγγίσεις στις περιοχές της κανονικοποίησης και φιλτραρίσματος γονιδίων, για μια επιτυχή αναγνώριση και κατηγοριοποίηση δειγμάτων.

Κεφάλαιο 2 - Βιολογικό Υπόβαθρο

2.1 Μοριακή Βιολογία

2.1.1 Το κύτταρο και η δομή του

Στην παρούσα εργασία επεξεργαστήκαμε κυτταρικά δεδομένα. Για αυτό είναι σημαντικό ξεκινώντας να αναφέρουμε τι είναι το κύτταρο και ποια είναι η δομή του.

Το κύτταρο αποτελεί τη βασική δομική και λειτουργική μονάδα κάθε οργανισμού. Ως κύτταρο νοείται το μικρότερο δομικό συστατικό της έμβιας ύλης, που αποτελείται από μια συστηματικά οργανωμένη ομάδα μορίων, που βρίσκονται σε δυναμική αλληλεπίδραση μεταξύ τους. Διαθέτει μορφολογική, φυσική και χημική οργάνωση και την ικανότητα της αφομοίωσης, της ανάπτυξης και της αναπαραγωγής. Το κύτταρο είναι μια μονάδα της ζωής ανεξάρτητη ως προς την αυτορρύθμιση και την προσαρμοστικότητά του σε σχέση με το περιβάλλον. Αναπτύσσεται, αναπαράγεται και πολλαπλασιάζεται αλλά και πεθαίνει με το μηχανισμό της απόπτωσης.

Μεγάλες ομάδες ομοειδών κυττάρων, χαρακτηρίζονται ως ιστοί, (π.χ. μυϊκός ιστός), οι οποίοι και αποτελούν την μονάδα δεύτερης τάξης στον ανθρώπινο οργανισμό, μετά τα κύτταρα. Οι οργανισμοί διακρίνονται σε μονοκύτταρους και πολυκύτταρους.

Τα κύτταρα ποικίλουν στο μέγεθος και στις διαστάσεις τους. Συγκεκριμένα η διάμετρός ξεκινάει από δέκατα του μικρομέτρου (ή χιλιοστά του χιλιοστομέτρου), όπως για παράδειγμα στα βακτήρια, έως και μερικά εκατοστόμετρα, όπως για παράδειγμα συμβαίνει στα αυγά πτηνών. Τα ανθρώπινα κύτταρα τα συναντάμε σε τάξη μεγέθους από 5 χιλιοστών του χιλιοστομέτρου μέχρι 1,5 χιλιοστόμετρο. Το ανθρώπινο σώμα αποτελείται από εκατό τρισεκατομμύρια κύτταρα.

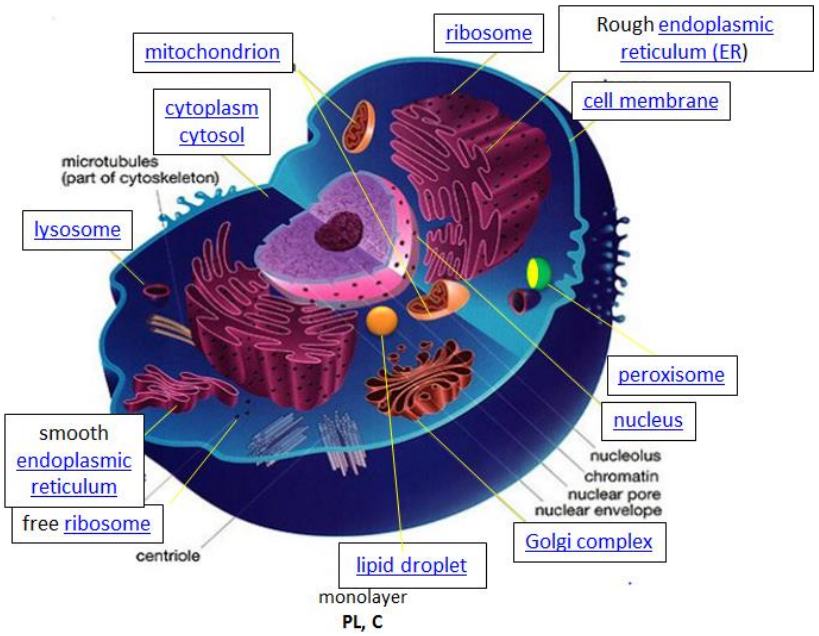
Ως οργανισμός, το κύτταρο διαθέτει την ικανότητα να ζει ακόμη και χωρίς την ύπαρξη άλλων κυττάρων. Αυτό όμως προϋποθέτει την ύπαρξη μιας μεταβολικής μηχανής που μπορεί να αντλήσει ενέργεια από το περιβάλλον και να τη χρησιμοποιήσει σε ουσιώδεις βιοχημικές διεργασίες. Εκτός από τη μεταβολική μηχανή του το κύτταρο διαθέτει ομάδες γονιδίων που καθορίζουν τη σύνθεση ουσιών και μια διακριτή δομή την κυτταρική ή πλασματική μεμβράνη που τα απομονώνει από το εξωτερικό περιβάλλον. Προκειμένου να είναι βιώσιμο ένα κύτταρο, αρκούν 400 γονίδια ή και λιγότερα, ωστόσο τα περισσότερα κύτταρα περιέχουν αρκετά περισσότερα.

Υπάρχουν δύο είδη κυττάρων: τα κύτταρα των προκαρυωτικών και τα κύτταρα των ευκαρυωτικών οργανισμών. Οι προκαρυωτικοί, όπως τα βακτήρια και τα κυανοφύκη είναι οργανισμοί πιο πρωτόγονοι από τους ευκαρυωτικούς. Η βασική διαφορά των δύο ειδών εστιάζεται στο γεγονός ότι τα κύτταρα των προκαρυωτικών οργανισμών δεν έχουν σχηματισμένο πυρήνα στα κύτταρα τους σε αντίθεση με αυτά των ευκαρυωτικών που είναι πιο πολύπλοκα και έχουν ξεκάθαρα σχηματισμένο πυρήνα. Ο άνθρωπος

προφανώς είναι ένας πολυκύτταρος ευκαρυωτικός οργανισμός στον οποίο υπάρχουν πολλά είδη κυττάρων τα οποία και είναι διαφοροποιημένα μεταξύ τους και επιτελούν διαφορετική λειτουργία. Έτσι, έχουμε για παράδειγμα τα μυϊκά κύτταρα ή τα νευρικά κύτταρα που έχουν διαφορετική μορφή.

Στην εικόνα 2 απεικονίζεται η γενική μορφή ενός ευκαριωτικού κυττάρου με τα οργανίδιά του. τα οποία και περιγράφουμε παρακάτω. Τα βασικά συστατικά ενός ευκαριωτικού κυττάρου είναι:

1. Την **κυτταρική ή πλασματική μεμβράνη**. Η κυτταρική μεμβράνη είναι μία βιολογική μεμβράνη που χωρίζει το εσωτερικό όλων των κυττάρων από το εξωτερικό περιβάλλον.
2. Το **κυτταρόπλασμα**, μία παχύρευστη και ομοιογενής ύλη εξαιρετικά πολύπλοκή καθώς μέσα της σχηματίζεται ένα ολόκληρο δίκτυο από κανάλια, το **ενδοπλασματικό δίκτυο.** Μέρος του δικτύου επικοινωνεί με την εξωτερική επιφάνεια. Μέσα στο κυτταρόπλασμα υπάρχουν τα υπόλοιπα οργανίδια του κυττάρου.
3. Τα **μιτοχόνδρια**, τα οποία μπορεί να είναι στρογγυλά ή σε σχήμα μπαστουνιού
4. Το **σύμπλεγμα Golgi,** το οποίο περιέχει οργανίδια που συνδέονται στη λειτουργία τους με το ενδοπλασματικό δίκτυο. Τροποποιούν μερικές πρωτεΐνες, ορισμένες από τις οποίες εκκρίνονται από το κύτταρο. Επίσης βοηθούν και στην παραγωγή κυτταρικών μεμβρανών.
5. Τα **λυσσοσώματα**, που είναι κύστες που έχουν μέσα τους αποθηκευμένα ένζυμα τα οποία χρησιμεύουν στην πέψη ουσιών που «τρώει» το κύτταρο.
6. Τον **πυρήνα.** Είναι το πιο σημαντικό οργανίδιο του κυττάρου καθώς περιέχει το γενετικό υλικό (DNA). Διακρίνουμε την **πυρηνική μεμβράνη** και τον **πυρηνίσκο** του. Όταν ένα κύτταρο διαιρείται φαίνονται καθαρά και τα χρωμοσώματα που περιέχει.



Εικόνα 2. Τυπικό ευκαριωτικό κύτταρο με τα οργανίδιά του [Ε.1]

2.1.2 Κυτταρικές Σειρές και κυτταρικές καλλιέργειες

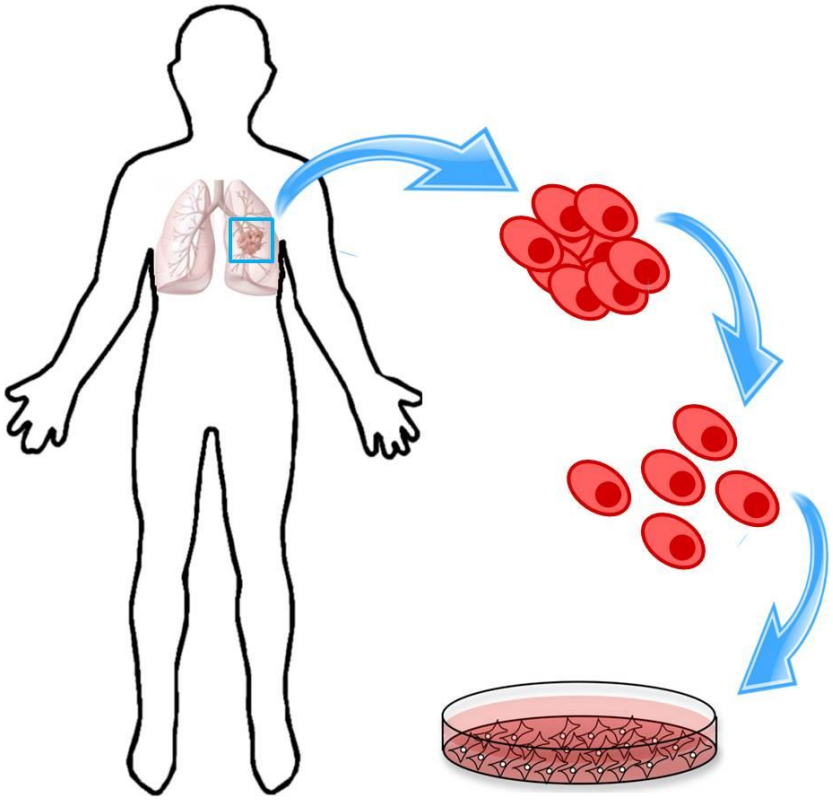
Κυτταρική καλλιέργεια είναι η διαδικασία μέσω της οποίας τα κύτταρα πολλαπλασιάζονται κάτω από ελεγχόμενες συνθήκες εκτός του φυσικού τους περιβάλλοντος. Πρακτικά ο όρος κυτταρική καλλιέργεια παραπέμπει στην καλλιέργεια κυττάρων που προέρχονται από πολυκύτταρους ευκαριωτικούς οργανισμούς και κυρίως ανθρώπων ή ζώων. Ωστόσο υπάρχουν καλλιέργειες από φυτά, μύκητες, μικρόβια, ιούς, βακτήρια και πρώτιστα. Ιστορικά η ανάπτυξη μεθόδων κυτταρικών καλλιεργειών είναι συνδεδεμένη με την καλλιέργεια ιστών και οργάνων.

Οι κυτταρικές σειρές αποτελούν κυτταρικές καλλιέργειες προερχόμενες από ένα μεμονωμένο κύτταρο που περιέχουν την ίδια γενετική σύσταση. Το αρχικό κύτταρο έχει απομονωθεί σε ζωντανή μορφή από τον ιστό. Η ανάπτυξη κυτταρικών σειρών είναι μία καθαρά εργαστηριακή τεχνική.

Τα κύτταρα μπορούν να απομονωθούν από τους ιστούς για καλλιέργεια τύπου ex vivo με διάφορους τρόπους. Εύκολα μπορούν να απομονωθούν από το αίμα ωστόσο μόνο τα λευκά αιμοσφαίρια είναι ικανά για ανάπτυξη σε καλλιέργεια. Τα μονοπύρηνα κύτταρα μπορούν να απελευθερώνονται από τους μαλακούς ιστούς με ενζυματική χώνεψη, με ένζυμα όπως η κολλαγιανάση, θρυψίνη ή προνάση.

Τα κύτταρα που καλλιεργούνται απευθείας από ένα υποκείμενο είναι γνωστά ως πρωτογενή κύτταρα. Με την εξαίρεση κάποιων που προέρχονται από όγκους οι περισσότερες πρωτογενείς καλλιέργειες κυττάρων έχουν περιορισμένη διάρκεια ζωής. Μία καθιερωμένη ή αθανατοποιημένη κυτταρική σειρά έχει αποκτήσει την ικανότητα να πολλαπλασιάζεται επ' αόριστόν είτε μέσω τυχαίας μετάλλαξης ή σκόπιμης τροποποίησης όπως τεχνητή έκφραση του γονιδίου τελομεράσης. Πλέον υπάρχουν πολυάριθμες κυτταρικές σειρές οι οποίες έχουν καθιερωθεί ως αντιπροσωπευτικές για διάφορους τύπους κυττάρων. [5]

Αντίστοιχα έχουν αναπτυχθεί κυτταρικές σειρές προερχόμενες από ανθρώπους με σκοπό την μελέτη ασθενειών και την ανακάλυψη θεραπειών. Το συγκεκριμένο θέμα έχει εγείρει διάφορους προβληματισμούς βιοηθικής αλλά το συγκεκριμένο θέμα δεν αποτελεί αντικείμενο μελέτης της παρούσας διπλωματικής.



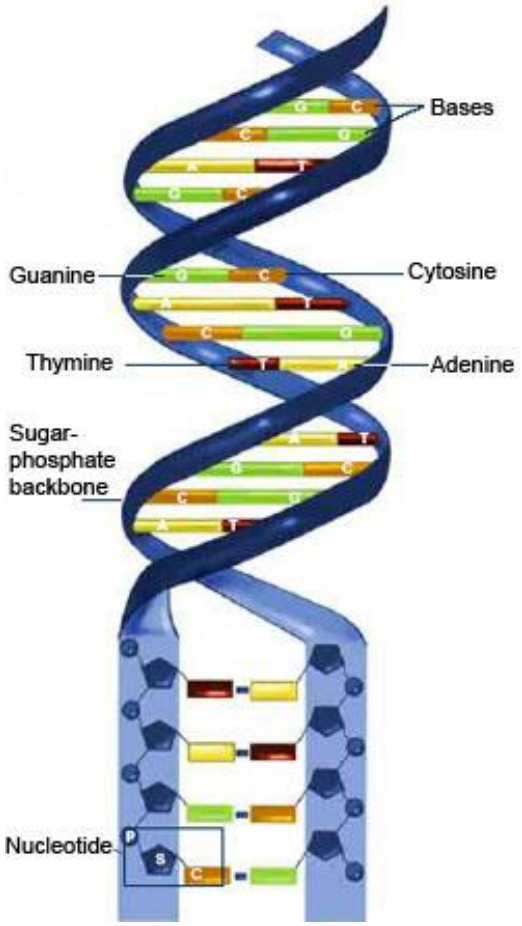
Εικόνα 3. Η διαδικασία μιας in vitro κυτταρικής καλλιέργειας [Ε.2].

2.1.3 Εισαγωγή στη Μοριακή Βιολογία

Όλα τα κύτταρα ενός οργανισμού περιέχουν DNA (δε(σ)οξυριβο(ζο)νουκλεϊ(νι)κό, οξύ Deoxyribonucleic acid), το κληρονομικό υλικό των οργανισμών. Το DNA είναι νουκλεϊκό οξύ που περιέχει τη γενετική πληροφορία η οποία καθορίζει τη λειτουργία του κυττάρου, επομένως και του οργανισμού αλλά και την βιολογική ανάπτυξη των ιών. Αποτελείται από δύο επιμήκης πολυνουκλεοτιδικές αλυσίδες, που σχηματίζουν διπλή δεξιόστροφη έλικα, και αποτελεί συνδυασμό 4 αζωτούχων βάσεων:

- ✓ Κυτοσίνη C
- ✓ Γουανίνη G
- ✓ Αδενίνη A
- ✓ Θυμίνη T

Μία από τις σημαντικότερες ιδιότητες του DNA είναι η συμπληρωματικότητα των τεσσάρων προαναφερθέντων βάσεων. Συγκεκριμένα, η Γουανίνη ενώνεται πάντα με την Κυτοσίνη και η Αδενίνη πάντα με την Θυμίνη.



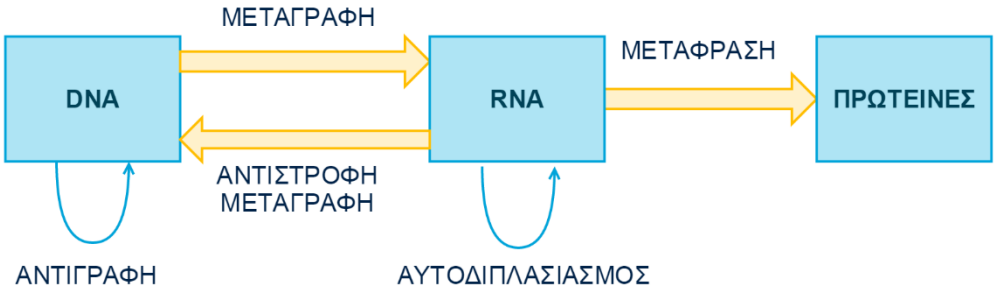
Εικόνα 4. Το μόριο του DNA [E.4]

Στο γονιδίωμα παρέχονται οδηγίες για την παρασκευή και άλλων συστατικών του κυττάρου, όπως για παράδειγμα το RNA. Το RNA (ριβοζονουκλεϊκό οξύ, Ribonucieic acid) είναι μία από τις δύο κατηγορίες των νουκλεϊκών οξέων και είναι μια μονή αλυσίδα που βασίζεται στον συνδυασμό τεσσάρων βάσεων. Η διαφορά ότι αντί για Θυμίνη περιέχει Ουρακίλη (U). Δηλαδή η Αδενίνη έχει ως συμπληρωματική της την Ουρακίλη. Η αρχή της συμπληρωματικότητας ισχύει και για το μόριο του RNA. Το RNA μαζί με το DNA αποτελούν το γενετικό υλικό των οργανισμών.

Η μέθοδος που εκφράζεται ένα γονίδιο είναι μέσω της παραγωγής πρωτεϊνών, των δομικών στοιχείων της ζωής. Κάθε γονίδιο συνήθως κωδικοποιεί μια συγκεκριμένη πρωτεΐνη, και σε δεδομένη στιγμή της ζωής του το κύτταρο παράγει διαφορετικές πρωτεΐνες. Ο τρόπος με τον οποίο ένας οργανισμός αντιδρά στις περιβαλλοντικές και βιολογικές αλλαγές, αλλά και στα διάφορα στάδια ανάπτυξης του, είναι μέσω της ενεργοποίησης ή απενεργοποίησης της παραγωγής συγκεκριμένων πρωτεϊνών.

Τα γονίδια παράγουν πρωτεΐνες με έναν πολύπλοκο μηχανισμό που υπόκειται σε διάφορα επίπεδα ρύθμισης. Το πρώτο στάδιο είναι η μεταγραφή ενός γονιδίου από το DNA σε ένα μόριο RNA, όπου ονομάζεται mRNA (messenger RNA). Στο δεύτερο στάδιο, που είναι η μετάφραση, ο κυτταρικός μηχανισμός παράγει μια πρωτεΐνη

χρησιμοποιώντας το μόριο του mRNA σαν προσχέδιο. Τα στάδια της μεταγραφής και της μετάφρασης, μαζί με την αντιγραφή του DNA, αποτελούν το Κεντρικό Δόγμα της Μοριακής Βιολογίας, το οποίο διατυπώθηκε το 1958 από τον Φράνσις Κρικ. Σύμφωνα με αυτό το δόγμα η γενετική πληροφορία ρέει από τα νουκλεϊκά οξέα (το DNA και RNA) προς τις πρωτεΐνες. Σήμερα έχουν επινοηθεί και άλλοι τρόποι ροής της γενετικής πληροφορίας.

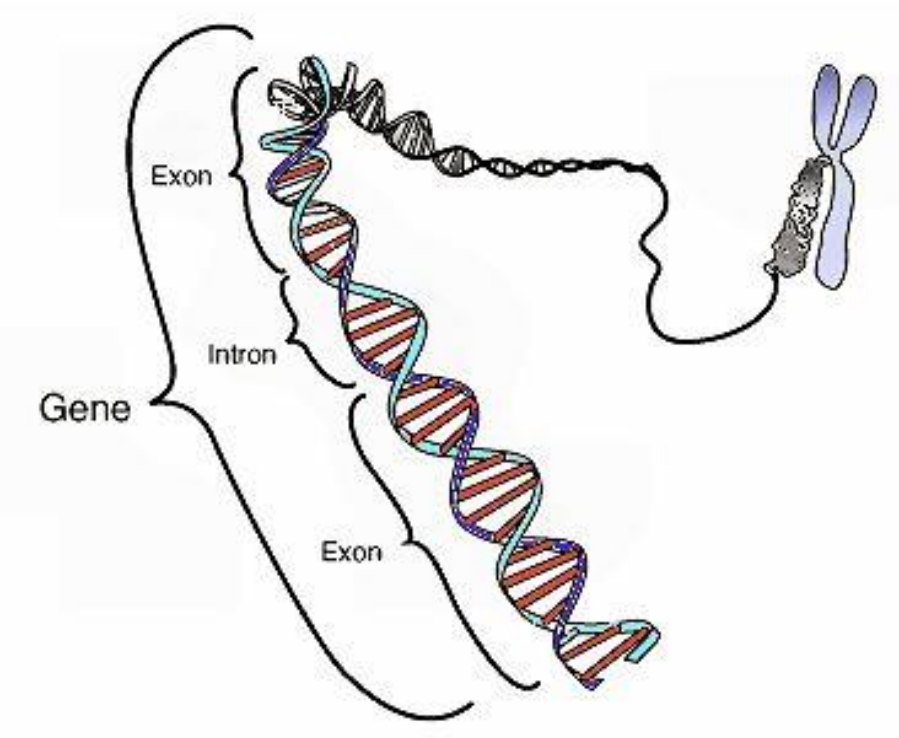


Εικόνα 5. Το Κεντρικό Δόγμα της Μοριακής Βιολογίας σήμερα

2.1.4 Το γονίδιο

Τα γονίδια είναι συγκεκριμένες αλληλουχίες βάσεων του DNA, όπου περιέχουν αποθηκευμένη μία συγκεκριμένη γενετική πληροφορία. Ένας πολύπλοκος μηχανισμός διαβάζει αυτή την πληροφορία και μεταφράζει την νουκλεοτιδική ακολουθία σε αμινοξική. Με τον τρόπο αυτό παράγεται ένα συγκεκριμένο είδος πρωτεΐνης. Τα γονίδια καθορίζουν τη σειρά ή την αλληλουχία των αμινοξέων σε μία πρωτεΐνη.

Τα γονίδια των ευκαρυωτικών οργανισμών αποτελούνται από *εξόνια* και *ιντρόνια*. Τα εξόνια είναι τμήματα του γονιδίου που κωδικοποιούν πρωτεΐνη. Τα ιντρόνια είναι οι μη κωδικές περιοχές. Τα γονίδια μπορούν να ελέγχουν κάθε κυτταρική δραστηριότητα και κατευθύνουν τη φυσική ανάπτυξη και συμπεριφορά του οργανισμού. Τα περισσότερα γονίδια κωδικοποιούν πρωτεΐνες, οι οποίες είναι βιολογικά μακρομόρια αποτελούμενα από γραμμικές αλυσίδες αμινοξέων και μπορεί να ελέγχουν τις βιοχημικές αντιδράσεις που πραγματοποιούνται στα κύτταρα, ενώ άλλες πρωτεΐνες έχουν άλλους ρόλους. Έτσι λοιπόν τα γονίδια ρυθμίζουν την παραγωγή των πρωτεϊνών, καθώς και το πότε και σε τι ποσότητα θα παραχθούν αυτές. Μερικά γονίδια όμως δεν κωδικοποιούν πρωτεΐνες. Ο ρόλος τους σε αυτήν την περίπτωση είναι να ρυθμίζουν την έκφραση άλλων γονιδίων Το σύνολο των γονιδίων ενός οργανισμού αποτελεί το γονιδίωμα του και βρίσκετε οργανωμένο σε χρωμοσώματα, που είναι στον πυρήνα των κυττάρων.



Εικόνα 6. Απεικόνιση γονιδίου [Ε.3]

2.1.5 Γονιδιακή Έκφραση στο Κύτταρο

Στη Γενετική με τον όρο γονιδιακή έκφραση ή έκφραση γονιδίων, (gene expression), χαρακτηρίζεται η διαδικασία εκείνη που προκαλεί τη μεταφορά κωδικοποιημένων πληροφοριών (του γονιδίου) στο λειτουργικό προϊόν του γονιδίου (πρωτεΐνη ή RNA).Γενικά η έκφραση γονιδίων εξισώνεται με τη διαδικασία της μεταγραφής και της μετάφρασης. Στη περίπτωση όμως που το προϊόν είναι μόνο RNA τότε εμπλέκεται η μεταγραφή. Έτσι όταν λέμε ότι ένα "γονίδιο εκφράζεται" αυτό σημαίνει πως πρόκειται για ενεργό γονίδιο. [12]

Η έκφραση γονιδίων είναι μια περίπλοκη και αυστηρά ελεγχόμενη διαδικασία που επιτρέπει σε ένα κύτταρο να απαντά στα περιβαλλοντικά ερεθίσματα καθώς και στις ανάγκες του. Αυτός ο μηχανισμός λειτουργεί ως διακόπτης ανοικτού-κλειστού, ο οποίος ελέγχει ποια γονίδια εκφράζονται στο κάθε κύτταρο. Ακόμα δρα σαν διακόπτης «ελέγχου ροής», δηλαδή αν κριθεί απαραίτητο αυξάνει ή ελαττώνει αντίστοιχα το επίπεδο έκφρασης κάποιων γονιδίων.

Για να ξεχωρίσουμε τις παθολογικές από τις φυσιολογικές βιολογικές λειτουργίες , πρέπει να εντοπιστούν τα γονίδια που είναι υπεύθυνα για το συγκεκριμένο φαινότυπο. Το ποσοστό του mRNA είναι ένας δείκτης που μας υποδεικνύει κατά πόσο

και αν εκφράστηκε ένα γονίδιο. Ο φαινότυπος του κυττάρου σχετίζεται με την ποσότητα μετάφρασης των γονιδίων, με το πόσο mRNA υπάρχει στο κύτταρο από κάθε γονίδιο καθώς και με το ποια γονίδια μεταφράστηκαν.

Επομένως, η μελέτη της γενετικής πληροφορίας απαιτεί:

- ✓ την εξέταση της μορφής των κυττάρων,
- ✓ την αλληλεπίδραση των γονιδίων μεταξύ τους και
- ✓ το πόσο ενεργά (active) είναι διάφορα γονίδια κάτω από συγκεκριμένες συνθήκες.

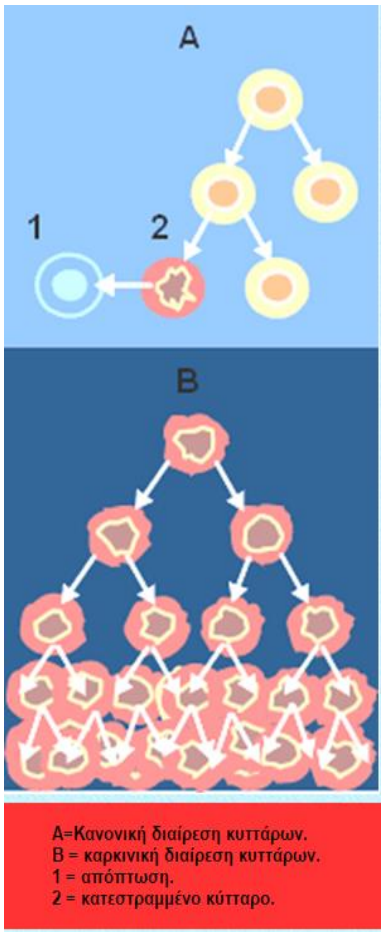
Πρέπει να ερευνηθούν πολλά γεγονότα του κυττάρου παράλληλα για να φανεί ποια είναι εκείνα τα γονίδια, τα οποία έχουν μεγαλύτερη γονιδιακή έκφραση καθώς και σε ποια κατάσταση συμβαίνει αυτό. Ένας τρόπος μελέτης για να φανούν αυτά είναι η «φωτογράφηση» των κυττάρων. Ένα ερώτημα που είναι πολύ σημαντικό να απαντηθεί είναι τι ποσό από κάθε mRNA βρίσκεται στο κύτταρο. Τα microarrays (μικροσυστοιχίες) μας επιτρέπουν να μετράμε την ποσότητα του mRNA για χιλιάδες γονίδια ταυτόχρονα. [13]

2.2 Τι είναι ο καρκίνος

Στις μέρες μας ένα από τα σοβαρότερα προβλήματα υγείας που καλείται να αντιμετωπίσουν οι επιστήμονες είναι ο καρκίνος. Οι στατιστικές μελέτες δείχνουν ότι είναι η δεύτερη αιτία θανάτου μετά τις καρδιοπάθειες. Η ασθένεια αυτή μπορεί να προσβάλλει ανθρώπους όλων των ηλικιών αν και τα συχνότερα κρούσματα εμφανίζονται σε άτομα μεγάλης ηλικίας.

Σε έναν άνθρωπο υγιή, εκατομμύρια κύτταρα καθημερινά αυξάνονται διαιρούνται και πεθαίνουν με έναν αυστηρά ελεγχόμενο τρόπο. Τα κύτταρα που υφίστανται βλάβες αντικαθίστανται από νέα υγιή κύτταρα έτοιμα να κάνουν και αυτά το κύκλο τους. Αυτός ο μηχανισμός δεν σταματάει ποτέ και δίνει την ευκαιρία στους ιστούς και στα όργανα του ανθρώπινου σώματος να επιδιορθωθούν και να αναζωογονηθούν. Ο ασταμάτητος αυτός μηχανισμός ελέγχεται αυστηρά από το γενετικό κώδικα που περιέχεται στον πυρήνα των κυττάρων.

Τα καρκινικά κύτταρα διαφέρουν στο γεγονός ότι δεν υπακούνε σε αυτές τις λειτουργίες συνεχίζοντας να διαιρούνται ανεξέλεγκτα, χωρίς να υπόκεινται στις διαδικασίες της απόπτωσης ή άλλης μορφής προγραμματισμένου κυτταρικού θανάτου. Επομένως η ανάπτυξη του καρκίνου στον άνθρωπο είναι αποτέλεσμα του ανεξέλεγκτου πολλαπλασιασμού των ανώμαλων , παθογόνων κυττάρων. (Εικόνα 7) Αυτό έχει ως αποτέλεσμα την ανάπτυξη μιας μάζας κυττάρων, που ονομάζεται όγκος. Οι όγκοι μπορεί να είναι καλοήθεις ή κακοήθεις. Οι καλοήθεις (μη-καρκινοειδείς) όγκοι δεν εξαπλώνονται σε άλλα μέρη του σώματος (μεταστατικοί) και είναι πολύ σπάνια απειλητικοί για τη ζωή [14].



Εικόνα 7. Διαίρεση των κυττάρων: Α-κανονική διαίρεση κυττάρων, Β-καρκινική διαίρεση κυττάρων 1-απόπτωση, 2-κατεστραμμένο κύτταρο [Ε.5].

Η αιτία έναρξης ενός ανεξέλεγκτου πολλαπλασιασμού καρκινικών κυττάρων μπορεί να οφείλεται σε αλλοίωση ή μετάλλαξη όπου υπέστη το DNA. Ο καρκίνος με άλλα λόγια θα λέγαμε ότι είναι η αποτυχία ρύθμισης της ανάπτυξης ενός ιστού. Συγκεκριμένα για να μετατραπεί ένα φυσιολογικό κύτταρο σε καρκινικό πρέπει να λάβουν χώρα μία ή περισσότερες μεταλλάξεις σε γονίδια. Τα γονίδια αυτά χωρίζονται σε δύο κατηγορίες [17] , τα ογκογονίδια, που προωθούν την ανάπτυξη του ιστού, και τα ογκοκατασταλτικά γονίδια, που είναι υπεύθυνα για τη διακοπή του κυτταρικού πολλαπλασιασμού. Συνήθως για την εμφάνιση ενός καρκίνου απαιτούνται μεταλλάξεις σε πολλά γονίδια. Οι μεταλλάξεις αυτές μπορεί να οφείλονται σε πολλούς λόγους. Ενδεικτικά αναφέρουμε τα λάθη που συμβαίνουν κατά τη μίτωση, όπως η απώλεια ή ο διπλασιασμός της περιοχής ενός χρωμοσώματος.

Τα καρκινικά κύτταρα ταξιδεύουν συχνά μέσω της κυκλοφορίας του αίματος ή της λέμφου σε άλλα σημεία του σώματος. Μετάσταση είναι η εξάπλωση των καρκινικών κύτταρων τα οποία εισβάλουν στους φυσιολογικούς ιστούς και σιγά σιγά

τους αντικαθιστούν. Το ποιο όργανο του σώματος έχει μολυνθεί από την ασθένεια του καρκίνου οφείλεται στην αρχική ανάπτυξη της κακοήθειας σε αυτό το όργανο (πρωτοπαθής) ή και από το φαινόμενο της μετάστασης. Με την μετάσταση μπορούν να σχηματιστούν δευτεροπαθείς όγκοι. [15]

Η ασθένεια του καρκίνου μπορεί να προκαλέσει πολλά διαφορετικά συμπτώματα, ανάλογα με το χαρακτήρα της κακοήθειας. Κάθε καρκίνος (π.χ. καρκίνος του πνεύμονα, της μήτρας, του προστάτη κτλ.) έχει διαφορετικά συμπτώματα, διαφορετική εξέλιξη και επομένως αποτελεί διαφορετική ασθένεια Μια οριστική διάγνωση απαιτεί συνήθως ιστολογική εξέταση από έναν παθολόγο. Αυτός ο ιστός λαμβάνεται από βιοψία ή χειρουργική επέμβαση. Μόλις εντοπιστεί, ο καρκίνος θεραπεύεται συνήθως με χειρουργική επέμβαση, χημειοθεραπεία, ή ακτινοθεραπεία. [16]

Άλλες μέθοδοι διάγνωσης νεοπλασιών που εφαρμόζονται σήμερα περιλαμβάνουν την ενδοσκοπική εξέταση και απεικονιστικές τεχνολογίες, όπως ακτίνες Χ, αξονική / μαγνητική τομογραφία και υπερηχογράφημα. Το πρόβλημα σε όλες τις παραπάνω τεχνικές είναι ότι είναι αποτελεσματικές όταν η νεοπλασία είναι εμφανής, και σε προχωρημένο στάδιο. Επίσης η έγκαιρη διάγνωση εξαρτάται από την εμπειρία του ιατρού. Γι' αυτό αναπτύσσονται νέες διαγνωστικές μέθοδοι, περισσότερο ευαίσθητες και αντικειμενικές. Ορισμένες από αυτές αναφέρονται παρακάτω.

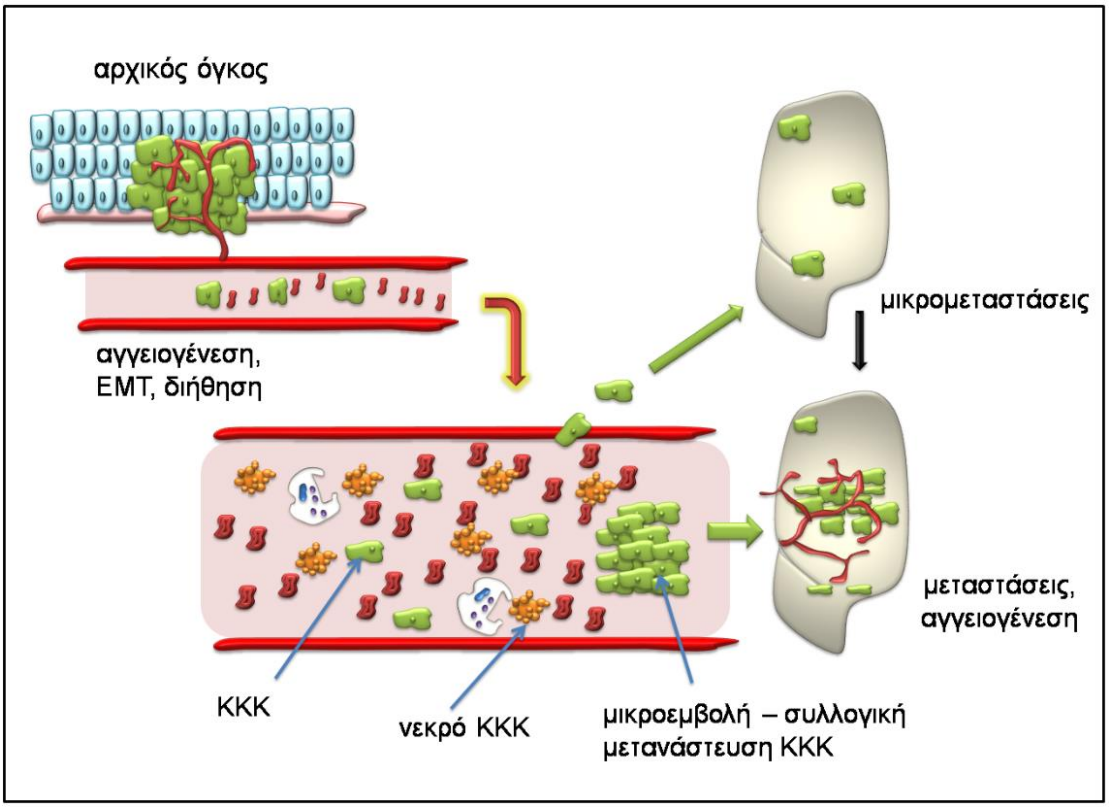
Πέραν της κλασικής βιοψίας, όπου τμήμα του ιστού αφαιρείται από την υπό εξέταση περιοχή και αναλύεται αργότερα στο εργαστήριο, την τελευταία δεκαετία έχει ανακαλυφθεί και μία πρωτοποριακή, μη επεμβατική μέθοδος, η «οπτική» βιοψία. Η συγκεκριμένη μέθοδος εξετάζει τους ιστούς in vivo και διενεργεί βιοχημική ανάλυση σε πραγματικό χρόνο, με μη επεμβατικό τρόπο, κάνοντας χρήση φασματικής απεικόνισης με μη ιονίζουσα ακτινοβολία [18]. Η μέθοδος αυτή ανιχνεύει νεοπλασίες από τα πρώτα στάδια ανάπτυξης και δεν εξαρτάται από την εμπειρία του εξεταστή.

Το 'στοίχημα' σήμερα είναι η έγκαιρη διάγνωση του καρκίνου μέσω απλών αιματολογικών εξετάσεων. Ενώ υπάρχουν αρκετοί καρκινικοί δείκτες, η εν λόγω εξέταση δεν ανιχνεύει όλους τους τύπους καρκίνου, ούτε τη νεοπλασία στο αρχικό στάδιο και πολλές φορές τα αποτελέσματα είναι αμφιλεγόμενα.

Το μεγαλύτερο κενό προς αυτή την κατεύθυνση βρίσκεται στην τεχνολογία που χρησιμοποιείται. Για παράδειγμα, μία «νεογέννητη» νεοπλασία μπορεί να απελευθερώσει έναν απειροελάχιστο αριθμό CTCs στην κυκλοφορία του αίματος. Οι σημερινές μέθοδοι όμως δεν μπορούν να ανιχνεύσουν τέτοιες μικροσυγκεντρώσεις.

2.3 Κυκλοφορούντα καρκινικά κύτταρα (CTCs- Circulating tumor cells)

Κυκλοφορούντα καρκινικά κύτταρα (CTCs circulating tumor cells) ονομάζονται τα καρκινικά κύτταρα που μπορεί να εντοπιστούν στην κυκλοφορία του αίματος ή τα λεμφαγγεία και προέρχονται από συμπαγείς όγκους. Τα CTCs μπορεί να κυκλοφορούν είτε ως μοναδικά κύτταρα, είτε ως συσσωματώματα (μικροεμβολές). Τα CTCs είχαν περιγράφει από τον Αυστραλό ιατρό Thomas Ashworth το 1869. Τις τελευταίες δεκαετίες δίνεται μεγάλη σημασία στα κυκλοφορούντα κύτταρα του όγκου (circulating tumor cells - CTCs) στο αίμα, ως δείκτη για τη δυναμική ανάπτυξης μεταστάσεων και συνεπακόλουθα καλής ή κακής πρόγνωσης.



Εικόνα 8. Η μετάσταση από τα Κυκλοφορούντα καρκινικά κύτταρα ή τις μικροεμβολές [Ε.6]

Υπάρχουν πολλοί λόγοι που ένα καρκινικό κύτταρο μπορεί να εγκαταλείψει την αρχική μάζα του όγκου και να μετατραπεί σε CTC. Ένας από τους πιο πιθανούς λόγους είναι οι ικανότητες διήθησης που αποκτά το καρκινικό κύτταρο στην διάρκεια του χρόνου. Η αποκόλληση μπορεί να προκληθεί είτε τυχαία ως απόρροια των τυπικών κυτταρικών επεκτάσεων είτε λόγω μηχανικών πιέσεων στην συγκεκριμένη περιοχή του όγκου. Όποιος και αν είναι ο λόγος που προκλήθηκε η αποκόλληση του καρκινικού κυττάρου από τον αρχικό όγκο, δεν παύει να είναι μια εγκατάλειψη της αρχικής θέσης. Η εγκατάλειψη αυτή υποδηλώνει μεταστατική συμπεριφορά. [19]

Τα μεμονωμένα καρκινικά κύτταρα είναι πιο εύκολο να πεθάνουν από ότι τα συσσωματώματα, αφού αποτελούνται από μεγαλύτερο αριθμό κυττάρων. Έτσι οι μικροεμβολές ή αλλιώς συσσωματώματα είναι πιο πιθανόν να δημιουργήσουν μεταστάσεις. Καθημερινά περίπου 1Χ106 κύτταρα ανά γραμμάριο όγκου απελευθερώνονται στην κυκλοφορία. Μελέτες σε ζωικά μοντέλα που εισήγαγαν καρκινικά κύτταρα κατ' ευθείαν στην κυκλοφορία έδειξαν ότι μόνο το 0,01% αυτών των κυττάρων κατάφεραν να σχηματίσουν μεταστάσεις. Η μεταστατική «ανεπάρκεια» που παρατηρείται καθορίζεται κυρίως από την μεγάλη ευαισθησία που παρουσιάζουν τα CTCs στην απόπτωση.[23]

Η ανίχνευση και καταμέτρηση τους ως ρουτίνα στην κλινική πράξη ήταν (και εξακολουθεί να είναι) δύσκολη και οι τεχνικές που χρησιμοποιούνται ως σήμερα είναι είτε δύσκολα εφαρμόσιμες, αναξιόπιστες στη απόδοση των ευρημάτων τους. Η δυσκολία του εντοπισμού τους, έγκειται στο γεγονός ότι σε ασθενείς με μεταστατική νόσο ανευρίσκεται ένα απ' αυτά σε κάθε 105–107 περιφερικά μονονοπύρηνα κύτταρα στο αίμα (λευκοκύτταρα). Σε ασθενείς με εντοπισμένο καρκίνο αυτή η αναλογία είναι ακόμα μικρότερη, 1 κυκλοφορούν καρκινικό κύτταρο σε 108 άλλα κύτταρα. Ως σήμερα έχουν χρησιμοποιηθεί τεχνικές με βάση την ανοσοϊστοχημεία, την κυτταρομετρία σάρωσης με laser, την κυτταρομετρία ροής κ.α [22]

Από τις διάφορες τεχνικές που έχουν χρησιμοποιηθεί ως σήμερα έχει ξεχωρίσει μία με την ονομασία CellSearch (Veridex), η οποία είναι και η μόνη με έγκριση από την FDA. Η συγκεκριμένη μέθοδος, που μπορεί να ανιχνεύσει 1 κύτταρο ανάμεσα σε 107 , παρουσιάζει τα εξής πλεονεκτήματα [20],[21],[22] :

- Υψηλή ευαισθησία και ειδικότητα,
- Αυτοματοποιημένη, ποσοτική,
- Σε υψηλό ποσοστό αναπαραγώγιμη,
- Μεσαία ποσότητα δείγματος είναι αναγκαία,
- Αναγνώριση βιώσιμων και μη βιώσιμων κυττάρων,
- Εμπορικά διαθέσιμη,
- είναι η μόνη εξέταση που έχει την έγκριση της FDA.

Ωστόσο έχει και τα παρακάτω μειονεκτήματα:

- Περιορισμένες παράμετροι ανάλυσης,
- η χρήση του EpCam για να συλληφθούν τα CTCs μπορεί να χάσει κάποια κύτταρα όγκου,
- Τα πολλαπλά βήματα εμπλουτισμού και επεξεργασίας μπορεί να έχουν ως αποτέλεσμα να χαθούν CTCs,
- Εν μέρει υποκειμενική ανάγνωση δεδομένων

Μία πρόσφατη μεγάλη προοπτική μελέτη με ασθενείς που έπασχαν από μεταστατική νόσο και υποβάλλονταν σε πρώτης γραμμής χημειοθεραπεία, έδειξε ότι η προγνωστική και προβλεπτική αξία των CTCs είναι ανεξάρτητη από τους καρκινικούς

δείκτες και πως η συγκεκριμένη μέτρηση θα μπορούσε να χρησιμοποιηθεί για την παρακολούθηση τυχόν ωφελημάτων απότοκων της θεραπείας.

Η ανίχνευση, καταμέτρηση και ο μοριακός χαρακτηρισμός των κυκλοφορούντων καρκινικών κυττάρων (Circulating Tumor Cells, CTCs), η ανάλυση μεταλλάξεων συγκεκριμένων γονιδίων και γενετικών αλλαγών στο ελεύθερο κυκλοφορούν καρκινικό DNA (circulating tumor DNA, ctDNA) καθώς και η ανίχνευση συγκεκριμένων κυκλοφορούντων microRNAs (circulating miRNAs) στο περιφερικό αίμα αποτελούν μία σημαντική μη-επεμβατική πηγή πληροφορίας για το προφίλ των όγκων ασθενών με καρκίνο. Όλα αυτά οδηγούν στην αναγκαιότητα εύρεσης μίας νέας πιο ακριβής και αξιόπιστης μεθόδου ανίχνευσης των κυκλοφορούντων καρκινικών κυττάρων.

2.4 Βιοπληροφορική

Βιοπληροφορική ονομάστηκε ο επιστημονικός κλάδος ο οποίος αναδύθηκε από τον συγκερασμό της μοριακής βιολογίας και της πληροφορικής. Είναι ένας τομέας ο οποίος παρέχει μεθόδους και εργαλεία τα οποία υποστηρίζουν την ανάγκη για εκμετάλλευση μεγάλης υπολογιστικής ισχύος για εξαγωγή γνώσης από βιολογικά δεδομένα. Έτσι οι επιστήμονες χρησιμοποιώντας μεθοδολογίες από την επιστήμη των υπολογιστών, κατάφεραν να επιλύσουν προβλήματα που προκύπτουν από τη Βιολογία. Συγκεκριμένα θεωρώντας τα βιολογικά δεδομένα(DNA, RNA, πρωτεΐνες) ως ψηφιακή πληροφορία, εφαρμόζονται αλγόριθμοι για την επεξεργασία τους και την παραγωγή χρήσιμων συμπερασμάτων με αποδοτικό τρόπο. Συχνά προκειμένου να επιτευχθούν τα παραπάνω αξιοποιούνται μέθοδοι κλάδων της τεχνητής νοημοσύνης όπως η εξόρυξη δεδομένων (π.χ νευρωνικά δίκτυα κ.α) αλλά και ο εξελικτικός υπολογισμός (π.χ γενετικοί αλγόριθμοι). [1]

Οι στόχοι της Βιοπληροφορικής μπορούν να ταξινομηθούν σε τρεις ομάδες [2]:

- ✓ Σε πρώτο επίπεδο η Βιοπληροφορική επιτρέπει την αποδοτική οργάνωση των δεδομένων, ώστε να είναι δυνατή η αποθήκευση, ανάκτηση και ενημέρωσή τους (π.χ. η βάση δεδομένων της δομής τρισδιάστατων μορίων Protein Data Bank).
- ✓ Σε δεύτερο επίπεδο η Βιοπληροφορική περιλαμβάνει εργαλεία τα οποία επιτρέπουν την ανάλυση βιολογικών δεδομένων.
- ✓ Σε τρίτο επίπεδο θέτει ως στόχο την ανάπτυξη εργαλείων με σκοπό την ερμηνεία αποτελεσμάτων βιολογικής σημασίας.

Οι κυριότερες εφαρμογές της Βιοπληροφορικής σήμερα αφορούν [4] :

- ➢ Υλοποίηση και σχεδιασμό υπολογιστικών εργαλείων για αυτόματη ανάκτηση γνώσης από Βάσεις Βιολογικών Δεδομένων
- ➢ Ανάλυση ακολουθιών Βιολογικών Δεδομένων (Στοίχιση ανά ζεύγη, αναζήτηση ομοιοτήτων σε βάσεις δεδομένων , πολλαπλή στοίχιση , φυλογενετική ανάλυση)
- ➢ Κατηγοριοποίηση Βιολογικών Δεδομένων

- Δομική Βιοπληροφορική (στοίχιση δομών στο χώρο, πρόγνωση δευτερογενούς και τριτογενούς δομής).
- Μοριακή μοντελοποίηση (Προσομοιώσεις, Σχέση δράσης –δομής)
- Ανάλυση πρωτεϊνών
- Γονιδιακή έκφραση (Μελέτη ρυθμιστικών στοιχείων, ανάλυση δεδομένων μικροσυστοιχιών)
- Συγκριτική ανάλυση Γονιδιωμάτων
- Πρόγνωση γονιδίων
- Σχεδιασμό φαρμάκων με βοήθεια ηλεκτρονικού υπολογιστή

Οι τεχνικές από τον χώρο της πληροφορικής που χρησιμοποιούνται είναι η τεχνολογία Βάσεων Δεδομένων για την οργάνωση, αποθήκευση και ανάκτηση βιολογικών δεδομένων, τεχνικές επεξεργασίας συμβολοσειρών (string manipulation techniques) για την ανάλυση ακολουθιών, τεχνικές μηχανικής μάθησης και εξόρυξης δεδομένων (data mining) στην ανακάλυψη προτύπων και τέλος αλγόριθμοι τρισδιάστατων συγκρίσεων στην ανάλυση τρισδιάστατης δομής βιολογικών μορίων. Στις περισσότερες περιπτώσεις οι υπολογιστικές τεχνικές ενσωματώνουν και στατιστικούς ελέγχους των αποτελεσμάτων. [3]

2.5 Μικροσυστοιχίες

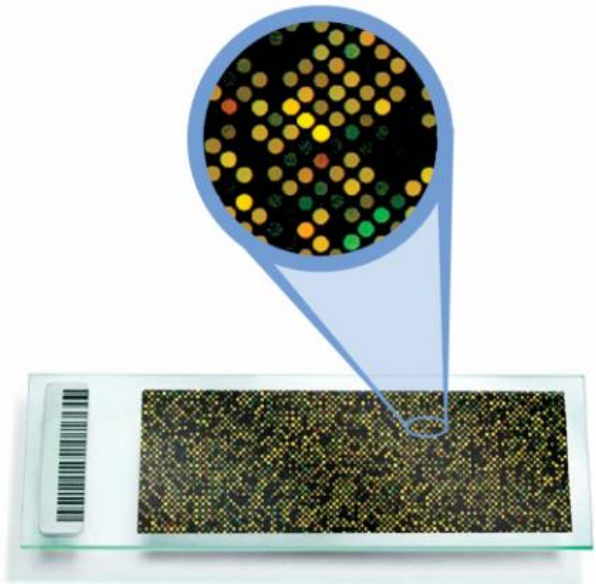
2.5.1 Μικροσυστοιχίες και γονιδιακή έκφραση

Για να μελετηθεί η γονιδιακή έκφραση , πρέπει πρώτα να βρεθεί η ποσότητα mRNA ή πρωτεϊνών, που παράγονται από ένα κύτταρο κάποια χρονική στιγμή. Τεχνικά εμπόδια και ζητήματα κόστους έκαναν απαγορευτική τη μελέτη πολλών γονιδίων ταυτόχρονα, μέχρι και τις αρχές της δεκαετίας του 1990 [26]. Οι επιστήμονες επιλύσαν αυτό το πρόβλημα με την ανάπτυξη μιας επαναστατικής μεθόδου τις μικροσυστοιχίες DNA ή αλλιώς DNA chip. Οι πρώτες μικροσυστοιχίες (cDNA microarrays) αναπτύχθηκαν στο πανεπιστήμιο του Stanford και περιεγράφηκαν για πρώτη φορά από τους Shena et al. (1995) για τον οργανισμό-μοντέλο Arabidopsis thaliana. Η τεχνολογία των μικροσυστοιχιών αποτελεί μετεξέλιξη της τεχνικής Southern, κατά την οποία αναλύεται ένας μικρός αριθμός μορίων [25].

Οι μικροσυστοιχίες χρησιμοποιούν μεθόδους ανάλυσης υψηλής απόδοσης. Αποτελούν μια τεχνολογία πειραματικής διαδικασίας σχεδιασμένη με σκοπό να εκτελεί παράλληλα πολλά πειράματα βιολογικού ενδιαφέροντος σε ένα μόνο κύκλο εκτέλεσης της διαδικασίας. Η τεχνολογία αυτή συγκαταλέγεται στην κατηγορία των «Lab-On a-Chip» συσκευών, δηλαδή συσκευές που ενσωματώνουν μία ή περισσότερες εργαστηριακές λειτουργίες σε ένα μόνο τσιπ ή πλακίδιο μεγέθους μερικών τετραγωνικών χιλιοστών ή εκατοστών. Η βασική ιδέα της μεθόδου στηρίζεται στην αρχή της συμπληρωματικότητας των βάσεων και υπολογίζει το ποσό μορίων mRNA

σε ένα "κελί", ανιχνευτικό σημείο της μικροσυστοιχίας. Έτσι για κάθε σημείο επιτρέπεται η μέτρηση των επιπέδων έκφρασης των γονιδίων για δεδομένο mRNA [25].

Οι μικροσυστοιχίες υπόσχονται την ταυτόχρονη μελέτη χιλιάδων γονιδίων και την ανάλυση της έκφρασής τους γρήγορα και αποτελεσματικά. Μελετώντας την έκφραση όλων των γονιδίων μια δεδομένη χρονική στιγμή, οι επιστήμονες δημιούργησαν μια νέα θεώρηση του πως λειτουργούν τα κύτταρα και απαντούν σε διαφορετικά ερεθίσματα, όπως είναι η αλλαγή του περιβάλλοντος, η εύρεση γονιδιακών ρυθμιστικών δικτύων, η έλλειψη θρεπτικών υλικών ή ακόμα και στη εύρεση μιας νέας μορφής ασθένειας όπως ο καρκίνος.



Εικόνα 9. Slide Array [E.7]

Οι μικροσυστοιχίες γονιδίων είναι μία διάταξη μικροσκοπικών σημείων που αντιπροσωπεύουν μοναδικά γονίδια και ακινητοποιούνται με ομοιοπολικούς δεσμούς σε μία στερεή επιφάνεια (συνήθως γυάλινη– array slide) [24]. Οι μικροσυστοιχίες περιλαμβάνουν το βιολογικό υλικό που προορίζεται για εξέταση. Χιλιάδες δείγματα βιολογικού υλικού, λόγω του μικρού τους μεγέθους, μπορούν να τοποθετούνται μαζί πάνω σε αυτό το στερεό υπόστρωμα και να εξετάζονται ταυτόχρονα από κάποιον ειδικό. Συγκεκριμένα πάνω στην γυάλινη επιφάνεια τοποθετούνται αλληλουχίες νουκλεοτιδίων (probes) σε προκαθορισμένες θέσεις (spots). Οι αλληλουχίες αυτές μπορεί να είναι είτε cDNAs προερχόμενες από κλώνους (ESTs) είτε ολιγονουκλεοτίδια, εξαρτάται από τη μέθοδο κάθε φορά. Η κάθε αλληλουχία τοποθετείτε σε προκαθορισμένη θέση της μικροσυστοιχίας και είναι συμπληρωματική έναντι ενός ειδικού μετάγραφου του γονιδιώματος. Η αλληλουχία αυτή αποτελεί τον ανιχνευτή του μετάγραφου του προς στόχευση γονιδίου στο βιολογικό δείγμα που εξετάζεται. Έτσι μπορεί να προσδιοριστεί η έκφραση του εν λόγω γονιδίου για κάποια θέση της μικροσυστοιχίας [27].

2.5.2 Κατηγορίες μικροσυστοιχιών

Η τεχνολογία των μικροσυστοιχιών περιλαμβάνει τις εξής κατηγορίες:

- Μικροσυστοιχίες DNA: σε αυτή την κατηγορία συγκαταλέγονται οι μικροσυστοιχίες cDNA, οι μικροσυστοιχίες ολιγονουκλεοτιδίων και οι μικροσυστοιχίες απλών πολυμορφισμών νουκλεοτιδίων (SNPs).
- MMChips: για την παρακολούθηση των πληθυσμών microRNA
- Μικροσυστοιχίες πρωτεϊνών
- Μικροσυστοιχίες πεπτιδίων: για την πραγματοποίηση λεπτομερών αναλύσεων σχετικά με τα σημεία αλληλεπιδράσεων των πρωτεϊνών μεταξύ τους
- Μικροσυστοιχίες ιστών
- Κυτταρικές μικροσυστοιχίες ή μικροσυστοιχίες επιμόλυνσης
- Μικροσυστοιχίες χημικών ενώσεων
- Μικροσυστοιχίες αντισωμάτων
- Συστοιχίες υδατανθράκων (glycoarrays)
- Μικροσυστοιχίες φαινοτύπων

2.5.3 Μικροσυστοιχίες DNA

Η γενική ιδέα πίσω από τα microarrays είναι η μέτρηση του επιπέδου έκφρασης πολλών γονιδίων, μέσω του υπολογισμού της ποσότητας του mRNA (και όχι της πρωτεΐνης) που είναι παρούσα στο κύτταρο για κάθε γονίδιο, σε διάφορες καταστάσεις και διάφορες χρονικές στιγμές. Επειδή το mRNA αποσυντίθεται στο κύτταρο πολύ γρήγορα, τα επίπεδά του αντικατοπτρίζουν αρκετά καλά και τον ρυθμό με τον οποίο το κύτταρο παράγει την αντίστοιχη πρωτεΐνη [25].

Οι πλέον διαδεδομένοι τύποι, που χρησιμοποιούνται σήμερα, είναι οι μικροσυστοιχίες cDNA και οι ολιγονουκλεοτιδικές μικροσυστοιχίες (τύπου Affymetrix).Στην συγκεκριμένη έρευνα χρησιμοποιήθηκε η τεχνολογία των DNA μικροσυστοιχιών και πιο συγκεκριμένα της εταιρίας Affymetrix.

Ένα πείραμα με μικροσυστοιχίες DNA έχει τα παρακάτω βήματα [28]:

1. Τοποθέτηση DNA στην μικροσυστοιχία:

Από γονίδια που μας ενδιαφέρουν συλλέγονται δείγματα, απομονώνονται και αυξάνονται με την μέθοδο PCR (Αλυσιδωτή Αντίδραση της Πολυμεράσης). Έτσι προκύπτει ένας αρκετά μεγάλος αριθμός cDNA probes και έπειτα καθαρίζονται, και ελέγχεται η ποιότητά τους. Στη συνέχεια τοποθετείται ποσότητα 5nl από κάθε probe σε συγκεκριμένες θέσεις (spots) πάνω στην επιφάνεια εναπόθεσης, με τη βοήθεια ενός ρομποτικού μηχανήματος πολύ μεγάλης ακρίβειας.

2. Απομόνωση mRNA από τα δείγματα:

Οι επιστήμονες ανάλογα με την κάθε περίπτωση που μελετάνε, απομονώνουν κάποια ποσότητα mRNA από το πειραματικό δείγμα (δείγμα εξέτασης - *experimental*) και από το δείγμα αναφοράς (δείγμα ελέγχου – *control*) και αυτό αντιστοιχεί στα γονίδια που εκφράζονται στο κύτταρο . Τα δείγματα μπορεί να είναι, για παράδειγμα, ένα φυσιολογικό (control) και ένα παθολογικό (καρκινικό κύτταρο). Το φυσιολογικό δείγμα αναφέρεται σαν δείγμα αναφοράς. Δηλαδή με βάση αυτό εντοπίζουν διαφορές στην έκφραση των γονιδίων σε ένα παθολογικό (καρκινικό κύτταρο).

3. Μετατροπή του mRNA σε cDNA και σήμανση με χρωστική ουσία:

Από το δείγμα εξέτασης και ελέγχου οι επιστήμονες λοιπόν απομονώνουν κάποια ποσότητα mRNA και το υποβάλλουν σε αντίστροφη μεταγραφή για να μετατραπεί σε cDNA. Μετά σημαίνονται με δύο φθορίζουσες ουσίες, με την Cye3-dUTP και την Cye5-dUTP. Η ουσία Cye3-dUTP έχει πράσινο φθορισμό (Cyanine-3) και η ουσία Cye5-dUTP κόκκινο (Cyanine-5) φθορισμό. Πιο συγκεκριμένα έχουν διαφορετική συχνότητα έκφρασης , η πρώτη έχει στα 550nm και δεύτερη στα 650nm.

4. Περίχυση του cDNA στη μικροσυστοιχία και Υβριδοποίηση:

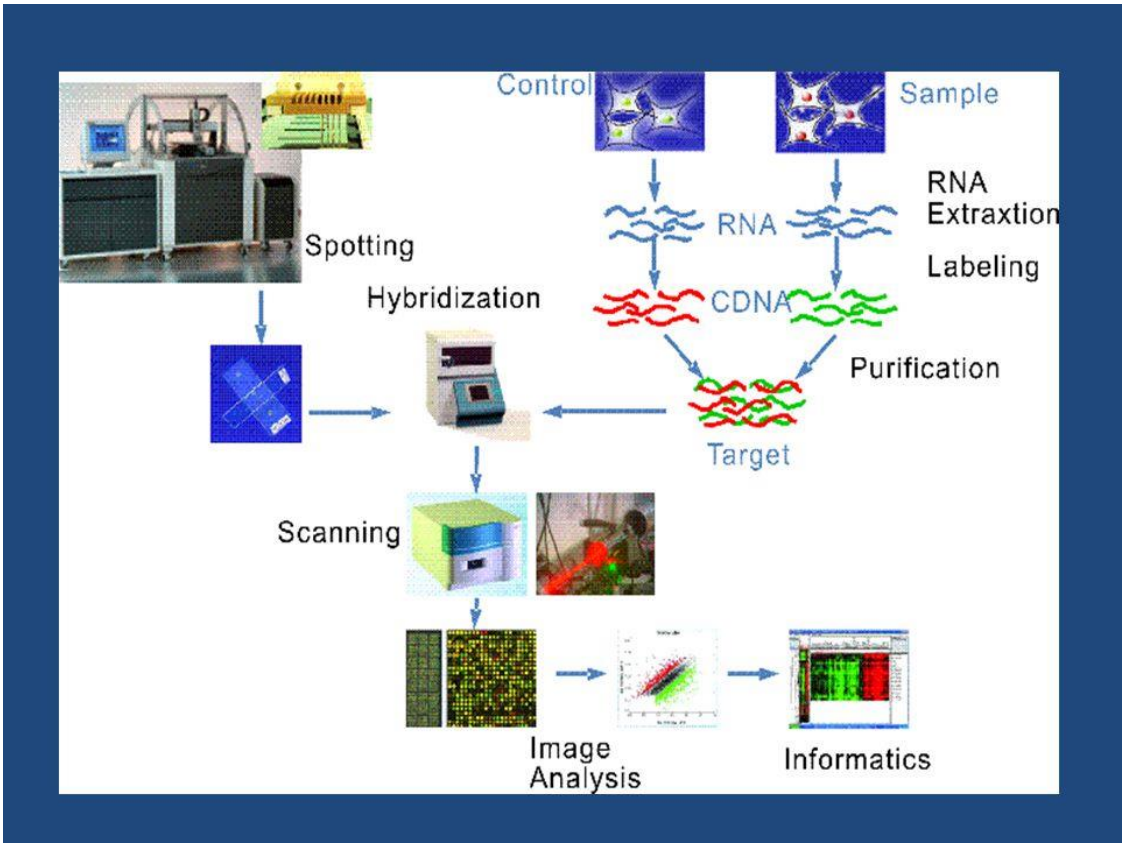
Το cDNA που σημάνθηκε, είναι ο στόχος και περιχύνεται στην μικροσυστοιχία. Έτσι υβριδοποιείται με αντιδράσεις ταιριάσματος βάσεων, σύμφωνα με την αρχή της συμπληρωματικότητας. Όταν ο στόχος εκτεθεί για αρκετό διάστημα στην μικροσυστοιχία (12 με 16 ώρες περίπου) ξεπλένουνε την μικροσυστοιχία.

5. Ανίχνευση μετά από διέγερση με laser:

Ειδικοί σαρωτές εικόνας (scanners), με τη βοήθεια περίπλοκων υπολογιστικών προγραμμάτων, αναλύουν την εικόνα που προκύπτει από την υβριδοποίηση υπολογίζοντας τα επίπεδα έκφρασης καθενός από τα γονίδια στον υπό εξέταση πληθυσμό RNA.

Αρχικά γίνεται σάρωση της συστοιχίας στα 550nm, όπου ανιχνεύεται ο πράσινος ιχνηθέτης (Cyanine-3), και έπειτα στα 650nm, όπου ανιχνεύεται ο κόκκινος ιχνηθέτης (Cyanine-5). Τις εικόνες που παίρνουμε τις συνενώνουμε και τις κανονικοποιούμε. Οι κηλίδες πάνω στη συστοιχία έχουν τα παρακάτω χρώματα :

- Κόκκινο: σε σημεία που υβριδοποιείται το cDNA με τον κόκκινο ιχνηθέτη (Cyanine-5).
- Πράσινο: σε σημεία που υβριδοποιείται το cDNA με τον πράσινο ιχνηθέτη (Cyanine-3).
- Κίτρινο : σε σημεία που υβριδοποιείται το cDNA και με τους δύο παραπάνω ιχνηθέτες (Cyanine-3 & Cyanine-5) .
- Μαύρο: σε σημεία που δε γίνεται υβριδοποίηση



Εικόνα 10. Τα βήματα ενός πειράματος με μικροσυστοιχίες cDNA [E.8]

Η τεχνολογία των μικροσυστοιχιών επιτρέπει στους ειδικούς να δώσουν απαντήσεις σε τρία βασικά ερωτήματα σε σχέση με την έκφραση του γονιδίου στο κύτταρο. Συγκεκριμένα, ο επιστήμονας μπορεί να καταλήξει σε υψηλής σημασίας συμπεράσματα για το πόσο ενεργά είναι κάποια γονίδια όταν βρίσκονται σε διαφορετικά κύτταρα ή ιστούς οργάνων, για το πώς αλλάζει η ενεργητικότητα των γονιδίων στα διάφορα στάδια μιας βιολογικής διαδικασίας (του κύκλου ζωής του κυττάρου), ή σε περιβαλλοντικές συνθήκες και ασθένειες και για το ποια γονίδια εκφράζονται παρόμοια και συνεργάζονται [26].

Η μέθοδος των μικροσυστοιχιών της μοριακής βιολογίας, λοιπόν έχει το πλεονέκτημα ότι δύναται να εξετάζει ταυτόχρονα την έκφραση χιλιάδων γονιδίων, και ενδείκνυται για συγκριτικές μελέτες γονιδιωμάτων. Τα μειονεκτήματα της μεθόδου έγκειται στο υψηλό κόστος και στη συχνή ανακρίβεια των αποτελεσμάτων λόγω τεχνικών προβλημάτων όπως η μη ειδική υβριδοποίηση φθοριζουσών χρωστικών σε λάθος γονίδια [24].

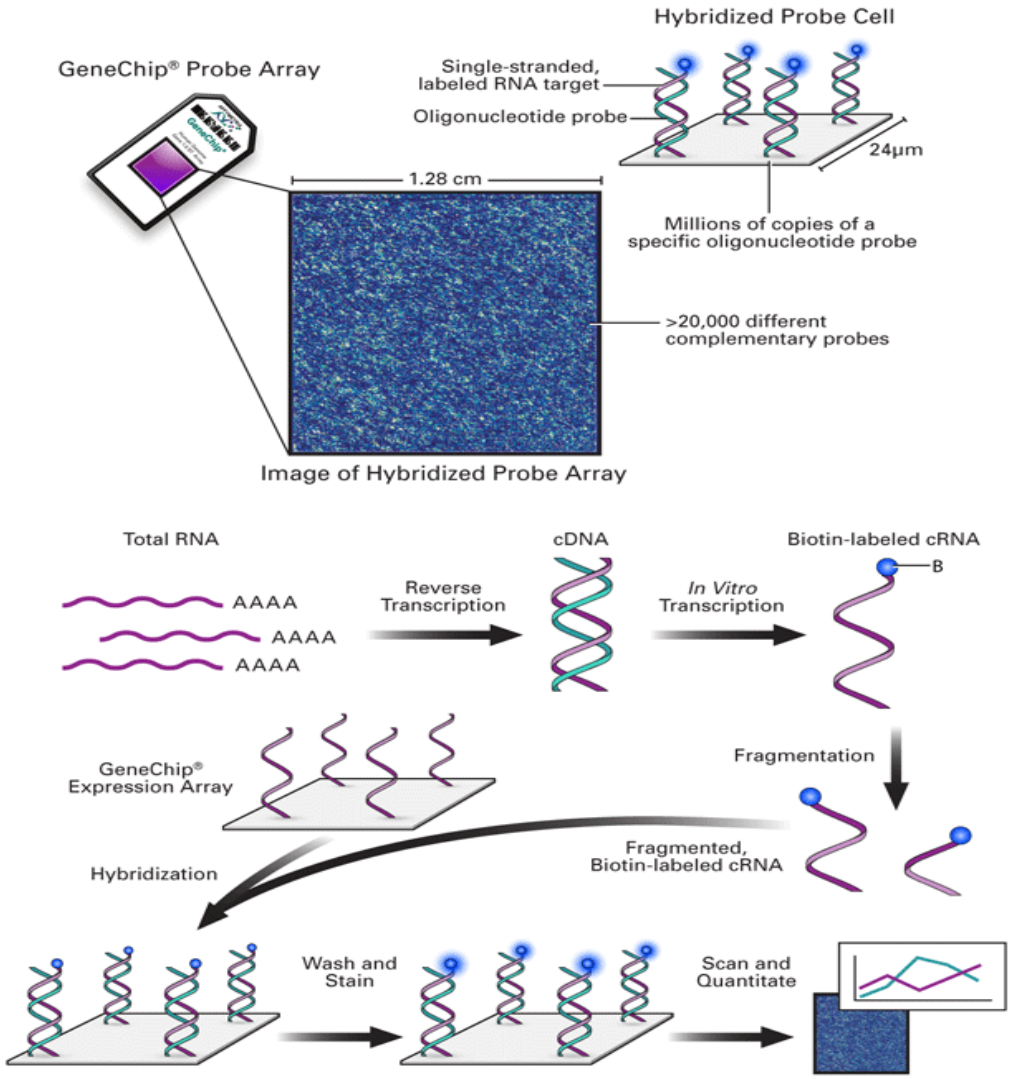
2.5.4 Μικροσυστοιχίες Affymetrix

Το GeneChip είναι το εμπορικό σήμα της Affymetrix Inc. και αντιπροσωπεύει ένα ολοκληρωμένο σύστημα προϊόντων και υπηρεσιών, συμπεριλαμβανομένων τις υψηλής πυκνότητας μικροσυστοιχίες.

Εικόνα 11. GeneChip Affymetrix [E.9]

Τα GeneChips της Affymetrix βοηθούν τον ερευνητή να πιστοποιήσει γρήγορα την έκφραση συγκεκριμένων γονιδίων σε ένα βιολογικό δείγμα. Αυτού του τύπου οι μικροσυστοιχίες ονομάζονται και μικροσυστοιχίες ολιγονουκλεοτιδίων. Η πιστοποίηση των γονιδίων από το βιολογικό δείγμα γίνεται μέσω της ανίχνευσης συγκεκριμένων τμημάτων mRNA. Ένα τσιπ μπορεί να χρησιμοποιηθεί για την ανάλυση χιλιάδων γονιδίων με μία πειραματική δοκιμασία. Τα τσιπ όμως είναι μιας χρήσης.

Η τεχνολογία κατασκευής αυτών των chip μοιάζει αρκετά με την κατασκευή των chip πυριτίου που χρησιμοποιούνται στους μικροεπεξεργαστές. Αυτή η μέθοδος επιτρέπει την μαζική παραγωγή μονάδων. Η βασική τεχνική που χρησιμοποιείται είναι φωτολιθογραφία μέσω της χρήσης ειδικών μασκών. Η σύνθεση των ολιγονουκλεοτιδίων στα socket του chip γίνεται μέσω μιας τροποποιημένης εκδοχής της μεθόδου phosphoramidite. Η συγκεκριμένη διαδικασία μπορεί να παράξει μεγάλο αριθμό ολιγονουκλεοτιδίων μήκους μέχρι και 25 ολιγονουκλεοτίδια και έχει καθιερωθεί ως η πιο χρήσιμη διαδικασία.



Εικόνα 12. Τα βήματα κατά την εκτέλεση ενός πειράματος με το GeneChip της Affymetrix [E.8]

Η ανάλυση της γονιδιακής έκφρασης γίνεται μέσω της χρήσης συμπληρωματικών τμημάτων του mRNA για το αντίστοιχο γονίδιο. Για το κάθε γονίδιο μπορούν να χρησιμοποιηθούν μέχρι και 40 ολιγονουκλεοτίδια. Η Affymetrix χρησιμοποιεί για την ανίχνευση συγκεκριμένα τμήματα του γονιδίου (πιθανότατα τμήματα που εμφανίζουν την μικρότερη δυνατή ομοιότητα με τα άλλα γονίδια). Ορισμένα γονίδια χρησιμοποιούνται για perfect match(PM) ενώ άλλα ολιγονουκλεοτίδια είναι τροποποιημένα για mismatch (MM). Τα mismatch ολιγονουκλεοτίδια είναι ίδια με τα perfect match εκτός από την θέση 13 (κέντρο) όπου το αντίστοιχο νουκλεοτίδιο έχει αντικατασταθεί με το συμπληρωματικό του.

Η εταιρεία μαζί με τα μαζικά GeneChips, σχεδιασμένα (εκ κατασκευής) για να ανιχνεύουν τμήματα υψηλής διαγνωστικής σημασίας ανθρώπινου και ζωικού γονιδιώματος , παρέχει και για την ταχεία ανάλυση microarray δεδομένων. Εκτός των συστημάτων της Affymetrix, ανταγωνιστικές εταιρείες κατασκευής μικροσυστοιχιών

είναι και οι Illumina, GE Healthcare, Applied Biosystems, Beckman Coulter, Eppendorf Biochip Systems και η Agilent. Τα τελευταία χρόνια κατασκευάζονται και εξειδικευμένα αντίστοιχα Chip από πλαστικό, από μικρές εταιρείες και εργαστήρια ανά τον κόσμο. Πιστεύεται ότι στο μέλλον η μαζική παραγωγή των Chip θα γίνεται εξ' ολοκλήρου από πλαστικό.

2.5.5 GeneChips U133 Plus 2.0 & Prime View

Για τη μελέτη του ανθρώπινου γονιδιώματος, η εταιρεία Affymetrix παρέχει διάφορους τύπους GeneChips. Στην παρούσα εργασία αναλύθηκαν δεδομένα από τα Human Genome Expression Arrays των σειρών U133 plus 2.0 και PrimeView. Τα γονίδια και οι ακολουθίες νουκλεοτιδίων που χρησιμοποιούνται, επιλέγονται μέσα από επιστημονικά καθιερωμένες βάσεις δεδομένων, αποδεκτές από την επιστημονική κοινότητα, ούτως ώστε τα αποτελέσματα των ερευνών να είναι αντικειμενικά και επαναλήψιμα. Πρέπει να τονιστεί πως μέχρι τώρα, οι μελέτες είναι ακόμα πειραματικές και εξυπηρετούν ερευνητικούς σκοπούς, όχι διαγνωστικούς.

Η σειρά U133 plus 2.0 αποτελείται από δύο συστοιχίες GeneChip. Περιέχει περίπου 45000 ανιχνευτές (probes) που αντιστοιχούν σε περισσότερες από 39000 μεταγραφές, οι οποίες προέρχονται από 33000, ανθρώπινα γονίδια. Οι ακολουθίες που χρησιμοποιεί αυτό το σετ προέρχονται από τις βάσεις δεδομένων GenBank, dbEST και RefSeq. Οι clusters των ακολουθιών δημιουργήθηκαν από τη βάση δεδομένων UniGene database (Build 133, April 20, 2001). Συγκρίθηκαν αναλυτικά με μία σειρά από άλλες βάσεις δεδομένων, ανοιχτές προς το κοινό, συμπεριλαμβανομένων των Washington University EST trace repository και University of California, Santa Cruz Golden Path human genome database (April 2001 release).

Αντίστοιχα, η σειρά PrimeView χρησιμοποιεί ακολουθίες επιλεγμένες από τις βάσεις RefSeq version 36, UniGene database 219, καθώς και πλήρους μήκους ανθρώπινο mRNA από την GenBank. Δύναται να μετρήσει περισσότερες από 36000 μεταγραφές ανά δείγμα, με 53000 ανιχνευτές (probes), που αντιπροσωπεύουν πάνω από 20000 γονίδια. Οι αλληλουχίες EST και mRNA που χρησιμοποιήθηκαν στο σχεδιασμό των τσιπ συνενώθηκαν για να δημιουργήσουν και εναλλακτικές μορφές συρραφής, με ειδικές αναλύσεις στον προσανατολισμό τους. Επίσης ο τρόπος που έχει σχεδιαστεί η συγκεκριμένη σειρά είναι συμπληρωματικός ως προς τις διαθέσιμες βάσεις βιολογικών δεδομένων. Για παράδειγμα, περισσότεροι από 1000 ανιχνευτές αντιπροσωπεύουν μεταγραφές οι οποίες δεν έχουν gene symbol στη βάση UniGene, αλλά βασίζονται σε προβλεπόμενες ακολουθίες της RefSeq.

Κεφάλαιο 3. Αναλυτικό υπόβαθρο

3.1 Μεθοδολογία Βιβλιογραφικής Ανασκόπησης

Η διαρκώς αναπτυσσόμενη τεχνολογία στον τομέα του γονιδιακής έκφρασης (gene expression) αναμένεται να συνεισφέρει σημαντικά στην κατανόηση των κυτταρικών λειτουργιών που σχετίζονται με την καρκινογέννεση. Επίσης, τα δεδομένα από τον τομέα της γονιδιακής έκφρασης εκτιμάται ότι θα παίξουν καταλυτικό ρόλο στην αποτελεσματική διάγνωση του καρκίνου και τη δημιουργία βάσεων δεδομένων ειδικής κατηγοριοποίησης και διαφοροποίησης. Πέραν των καινοτομιών και των βελτιώσεων που συμβαίνουν σε επίπεδο συσκευών και υλικών (hardware) στον τομέα της Μοριακής Βιολογίας, σπουδαίες ανακαλύψεις εμφανίζονται και στον τομέα της στατιστικής επεξεργασίας των βιολογικών δεδομένων.

Ο σκοπός της παρούσας βιβλιογραφικής έρευνας είναι αφενός μεν να κάνει μία επισκόπηση όλων των στατιστικών μεθόδων που έχουν γίνει μέχρι σήμερα με τις μικροσυστοιχίες της Affymetrix και αφετέρου να υποδείξει ποια είναι η αιχμή της τεχνολογίας στον τομέα αυτό ή αλλιώς το *"state of the art"*.

Δεν αποτελεί αντικείμενο αυτής της μελέτης η περιγραφή όλων των στατιστικών μεθόδων που χρησιμοποιούνται στις μικροσυστοιχίες άλλων εταιρειών πέραν της Affymetrix. Ο στόχος είναι να επισημανθούν τα κατάλληλα εργαλεία που αξίζει να χρησιμοποιηθούν σήμερα στον τομέα του Affymetrix Microarray Data Processing (βάσει της εμπειρίας και της έρευνας προηγούμενων ερευνητών) αλλά και να εξεταστούν βελτιώσεις στις αντίστοιχες μεθόδους.

3.2 Βιβλιογραφική Ανασκόπηση

Από τη δεκαετία του 1990 κι έπειτα, οι μικροσυστοιχίες (γνωστές και ως gene chip ή DNA chip) χρησιμοποιούνται ευρέως για την παρακολούθηση της έκφρασης χιλιάδων γονιδίων ταυτόχρονα, χρησιμεύοντας ως αναμφισβήτητα πολύτιμο εργαλείο για τις βιολογικές μελέτες. Οι μικροσυστοιχίες χρησιμοποιούνται για ποικίλους ερευνητικούς σκοπούς, όπως την εύρεση

- γονιδίων με διαφοροποιημένη έκφραση / διαφοροποιημένα γονίδια[36],
- γονιδιωματικών μοτίβων [37-41],
- "συνεκφραζόμενων γονίδίων¨" ή αλλιώς "συν-ρυθμιζόμενων"[42,43],
- Κατηγοριοποίησης καρκινικών υπο-ομάδων [44],
- Αλληλούχησης (genotyping and sequencing)[47].

Υπάρχουν δύο μορφές μικροσυστοιχιών, με συστοιχίες cDNA και με συστοιχίες ολιγονουκλεοτιδίων. Η συστοιχία cDNA περιέχει τμήμα DNA ακινητοποιημένο πάνω σε μια γυάλινη επιφάνεια (ή σε άλλη στερεή επιφάνεια) και εκτίθεται σε επισημασμένους (labeled) ανιχνευτές (probes). Η δεύτερου τύπου συστοιχία έχει ανιχνευτές ολιγονουκλεοτιδίων συντεθιμένους σε ειδικά chips [48, 49].

Οι μικροσυστοιχίες της Affymetrix ανήκουν στους ανιχνευτές ολιγονουκλεοτιδίων [50-53]. Η επισήμανση του (τμήματος) του γενετικού υλικού γίνεται με τη χρήση δύο χρωμάτων, κατά συνέπεια και δύο καναλιών, για τη συγκριτική υβριδοποίηση [38]. Τα δεδομένα που εξάγονται από αυτό το είδος συστήματος περιέχουν τη μετρούμενη ένταση δύο συγκεκριμένων χρωστικών / βαφών, της cy5 και cy3, από το ίδιο σημείο της συστοιχίας.

Έχουν προταθεί αρκετές μέθοδοι για την ανάλυση των πειραματικών αποτελεσμάτων. Γενικά, ομαδοποιούνται σε δύο βασικές κατηγορίες. Η πρώτη κατηγορία αφορά τον εντοπισμό διαφορετικά εκφρασμένων γονιδίων (differentially expressed genes) σε ένα χρονικό σημείο, όπως η model based ανάλυση [54-56], ANOVA [57-60] ή η εμπειρική ανάλυση σύμφωνα με την προσέγγιση Bayes [61]. Η δεύτερη ομάδα αφορά την ομαδοποίηση των γονιδίων σε clusters, βασισμένη σε πρότυπα έκφρασης που εμφανίζονται ανά τακτές χρονικές στιγμές κατά την πειραματική διαδικασία, όπως το hierarchical clustering [39], self-organizing map (SOM) [62, 63], quality cluster algorithm [64], neighborhood analysis [44], k-means clustering [65], singular value decomposition (SVD) [66], support vector machines (SVM) [67] and single-pulse model (SPM) [68].

Άλλα σημαντικά στατιστικά ζητήματα έχουν επίσης τεθεί και απαντηθεί, όπως το normalization [69,70], replication [71], statistical experimental design για cDNA microarray [57, 58], and reliability clustering results [72]. Υπάρχουν ακόμα βιβλιογραφικές αναφορές σχετικές με την επισκόπηση των μικροσυστοιχιών [65, 73, 74], και εισαγωγής στις μεθόδους της Affymetrix [75]. Η παρούσα εργασία επικεντρώνεται στις μικροσυστοιχίες ολιγονουκλεοτιδίων της Affymetrix.

3.3 Στατιστική προσέγγιση δεδομένων από μικροσυστοιχίες

Η χρήση μεταγραφικών δεδομένων σε μικροσυστοιχίες χρησιμοποιείται ευρέως για τη μελέτη πολλών βιολογικών λειτουργιών. Ένα ευρύ φάσμα προσεγγίσεων είναι διαθέσιμο για την ταξινόμηση των στοιχείων που λαμβάνονται από τις πειραματικές μετρήσεις. Η κατάλληλη επιλογή της τεχνικής ανάλυσης των δεδομένων εξαρτάται τόσο από τα δεδομένα, όσο και από τους στόχους του πειράματος.

Η ανάλυση ταξινόμησης (cluster analysis) είναι μια διερευνητική τεχνική που χρησιμοποιείται για να «αποκαλύψει» κλάσεις ή ομάδες γονιδίων ή δειγμάτων που λειτουργούν συνδετικά ή ομαδικά κατά τη διάρκεια μιας βιολογικής διεργασίας. Η (αν)ομοιότητα καθορίζεται από το αποτέλεσμα μιας μετρικής απόστασης μεταξύ των διανυσμάτων κάθε ζεύγους γονιδίων – δειγμάτων. Με βάση αυτή την απόσταση, στη

συνέχεια, μέσω της εφαρμογής στατιστικών αλγορίθμων clustering τα αντίστοιχα ζεύγη κατηγοριοποιούνται στους ίδιους clusters, εφόσον το προφίλ έκφρασης είναι παρόμοιο. Υπάρχει πληθώρα αλγορίθμων κατηγοριοποίησης, με πολλά περιθώρια παραμετροποίησης. Η επιλογή διαφορετικού αλγορίθμου ή παραμετροποίησης του ιδίου, δύναται να οδηγήσει σε διαφορετικά αποτελέσματα κατηγοριοποίησης.

Η επαλήθευση και αξιολόγηση της ταξινόμησης (classification) γίνεται σε δύο μέρη. Το πρώτο αφορά τη στατιστική συνοχή (consistency - stability) των παραγόμενων clusters. Το δεύτερο αφορά τη βιολογική λειτουργική αντιστοιχία των παραγόμενων ομαδοποιήσεων. Όσον αφορά το unsupervised classification, δεν υπάρχει ένας μοναδικός αλγόριθμος που να υλοποιεί καλύτερα την ομαδοποίηση των γονιδίων ή των δειγμάτων σε λειτουργικές ομάδες μέσω των προφίλ έκφρασης των γονιδίων. Η επιλογή αυτή εξαρτάται (ακόμα) από τη φύση του συνόλου των δεδομένων και από το «expert's opinion» για την αξιολόγηση των αποτελεσμάτων.

Άλλοι παράγοντες επίσης εκτός από την ακρίβεια συμβάλλουν στην εξέταση της αξίας ενός δεδομένου ταξινομητή. Αυτές περιλαμβάνουν την απλότητα της μεθόδου, το υπολογιστικό κόστος και την επίγνωση που αποκτήθηκε στην προγνωστική δομή των δεδομένων (διαγνωστική σημασία).

Όπως προαναφέρθηκε, στην εξόρυξη χρήσιμων πληροφοριών από δεδομένα μικροσυστοιχιών, η τεχνική που χρησιμοποιείται ευρέως είναι η ανάλυση ταξινόμησης (cluster analysis), μια διερευνητική τεχνική που χρησιμοποιείται για να «αποκαλύψει» κλάσεις ή ομάδες γονιδίων ή δειγμάτων που λειτουργούν συνδετικά ή ομαδικά κατά τη διάρκεια μιας βιολογικής διεργασίας. Οι αλγόριθμοι κατηγοριοποίησης (clustering) ομαδοποιούνται στους αλγόριθμους Ιεραρχικής ταξινόμησης (Hierarchical clustering) και Διαχωρισμού (Partitional Clustering).

Στην ιεραρχική ταξινόμηση [76-81] ακολουθούνται οι προσεγγίσεις του agglomerative clustering και linkage clustering. Στην ταξινόμηση διαχωρισμού χρησιμοποιούνται τεχνικές Minimum Spanning Tree algorithm (MST), Squared Error Clustering algorithm, K-Means clustering, Nearest Neighbor algorithm, Partitioning Around Medoids algorithm (Clustering large applications (CLARA) and Clustering large applications based upon randomized search (CLARANS), Bond Energy algorithm (BEA), Clustering with Genetic algorithms, Clustering with Neural Networks.

Η βιβλιογραφία αναφέρει χρήση όλων των παραπάνω αλγορίθμων. Οι συγκριτικές αναλύσεις δεν μπορούν να καταλήξουν στο ποιος είναι ο πιο ακριβής αλγόριθμος για κάθε δείγμα. Πάντα τα αποτελέσματα εξαρτώνται από τη φύση των δεδομένων και τις μικρορυθμίσεις (fine tuning) των ίδιων των αλγορίθμων [76-81]. Σε όλες όμως τις περιπτώσεις, ως reference χρησιμοποιείται το hierarchical clustering, όπως επίσης είναι και η βασική μέθοδος ταξινόμησης που προτείνει η ίδια η Affymetrix και η πλατφόρμα Matlab, μέσω του Bioinformatics toolbox.

Η προσέγγιση του Hierarchical clustering ακολουθήθηκε και σε αυτή την εργασία, με αρκετές βέβαια τροποποιήσεις και βελτιώσεις, οι οποίες αναλύονται επεξηγηματικά στα επόμενα κεφάλαια.

3.4 Ανάλυση Μεθοδολογιών

Στόχος αυτής της ενότητας είναι να περιγράψουμε το θεωρητικό υπόβαθρο κάποιων αλγορίθμων και τεχνικών που χρησιμοποιήθηκαν στην παρούσα εργασία. Οι τεχνικές και αλγόριθμοι που υλοποιήθηκαν , επιλέχτηκαν κατόπιν αναλυτικής έρευνας των υπαρχουσών τεχνικών. Σε περιπτώσεις όπου δεν εξάγονταν τα επιθυμητά αποτελέσματα δημιουργούσαμε ιδιοκατασκευασμένες (custom) τεχνικές και αλγορίθμους προκειμένου να εξάγουμε καλύτερα αποτελέσματα. Στα υποκεφάλαια που ακολουθούν περιγράφονται τεχνικές που αντλήσαμε από την βιβλιογραφία και συνίστανται για επεξεργασία δεδομένων μικροσυστοιχιών.

3.4.1 Ο Αλγόριθμος RMA

Ο αλγόριθμος RMA (Robust Multi-array Average) χρησιμοποιείται για τη διόρθωση θορύβου που εισάγεται κατά τη διαδικασία παραγωγής των δεδομένων εξόδου από τις μικροσυστοιχίες της Affymentrix. Η διόρθωση αυτή ονομάζεται και προεπεξεργασία των δεδομένων (preprocessing).

Τα δεδομένα εξόδου περιέχουν τις pixel-level intensities της μικροσυστοιχίας σε ένα αρχείο DAT. Στη συνέχεια, για τιμές των pixels που αντιστοιχούν σε ένα συγκεκριμένο probe υπολογίζεται ο μέσος όρος τους και έτσι εξάγεται η τιμή του εκάστοτε probe, οι οποίες βρίσκονται στα αρχεία τύπου CEL. Οι εντάσεις των probes που υπολογίστηκαν συνδυάζονται ανά τα σύνολα των probes που αντιστοιχούν στο κάθε γονίδιο. Έτσι τελικά υπολογίζεται η γονιδιακή έκφραση.

Η διαδικασία της προεπεξεργασίας των δεδομένων περιλαμβάνει 3 στάδια:

- Διόρθωση υποβάθρου (Background correction).
    - Αφαίρεση τοπικών σφαλμάτων και θορύβου, ούτως ώστε οι μετρήσεις να μην επηρεάζονται από τις γειτονικές τιμές.
- Κανονικοποίηση (Normalization)
    - Διόρθωση του θορύβου που εισάγεται από τη μικροσυστοιχία, ώστε οι τιμές που λαμβάνονται από διαφορετικές μετρήσεις στην ίδια συσκευή να είναι συγκρίσιμες.
- Συνόψιση (Summarization)
    - Συνδυασμός των τιμών των ίδιων probes που λήφθηκαν από διαφορετικές μετρήσεις, ώστε να υπάρξει μια πιο αντικειμενική εικόνα για τις τιμές της γονιδιακής έκφρασης.

Η χρήση του αλγορίθμου RMA λαμβάνει υπόψη μόνο τις τιμές έντασης φθορισμού του υβριδισμού από τις περιπτώσεις με perfect match (PM), όπου τα ολιγονουκλεοτίδια – στόχος είναι συμπληρωματικά με τα ολιγονουκλεοτίδια – ανιχνευτές.

Για τη διόρθωση του θορύβου του υποβάθρου (background correction) στα GeneChips της Affymetrix χρησιμοποιούνται μη γραμμικές διαδικασίες, οι οποίες βασίζονται στις τιμές που λαμβάνονται από το σήμα φθορισμού 17000 probes περίπου, οι οποίοι χρησιμοποιούνται ως τιμές αναφοράς. Ο θόρυβος (E) που υπεισέρχεται στο σήμα (S) ακολουθεί κανονική κατανομή και το λαμβανόμενο σήμα (S) είναι μίξη των προηγούμενων δύο σημάτων. Για την απομόνωση θορύβου στον αλγορίθμο RMA, ο θόρυβος υπολογίζεται, μέσω της μεθόδου εκτίμησης πυρήνα, ως ο μέσος όρος του καθαρού σήματος (S), υπό τη συνθήκη των τιμών των perfect match probes (O), όπως φαίνεται στον ακόλουθο τύπο [85],[86],[87]:

$$E\,(S|O=o) = \alpha + b\,\frac{\varphi\left(\frac{a}{b}\right) - \varphi\left(\frac{O-\alpha}{\beta}\right)}{\Phi\left(\frac{a}{b}\right) - \Phi\left(o-\frac{\alpha}{\beta}\right) - 1}$$

$$\text{Όπου } \alpha = o - \mu - \sigma^2\alpha \text{ , } b = \alpha$$

Η κανονικοποίηση των τιμών γίνεται μέσω της μεθόδου “quantile normalization” (προσαρτημόρια κανονικοποίηση). Η διαδικασία αυτή έχει ως στόχο να μετατρέψει διαφορετικές κατανομές σε πανομοιότυπες, ως προς τα στατιστικά τους χαρακτηριστικά, συνήθως βάση κάποιας κατανομής αναφοράς. Στην περίπτωση του RMA, όπου δεν υπάρχει κατανομή αναφοράς, επιλέγεται ο αριθμητικός μέσος των τιμών των κατανομών. Με αυτόν τον τρόπο, η υψηλότερη τιμή θα είναι ο αριθμητικός μέσος των υψηλότερων τιμών, η δεύτερη υψηλότερη τιμή θα είναι ο αριθμητικός μέσος των αμέσως επόμενων υψηλότερων τιμών, και ούτω καθεξής.

Στη συνέχεια, η συνόψιση των δεδομένων (summarization) περιλαμβάνει τις τιμές από τη διόρθωση του υποβάθρου, τις “quantile” κανονικοποιημένες τιμές και το λογάριθμο των perfect match τιμών. Οι παράμετροι του τελικού μοντέλου εκτιμώνται με τη χρήση του “median polish” ως εξής [85],[86],[87]:

$$Y_{ijk} = \mu_{ik} + \alpha_{jk} + \varepsilon_{ijk}$$

α_{jk} → Probe affinity effect for each k , $\sum_j (a_{jk}) = 0$

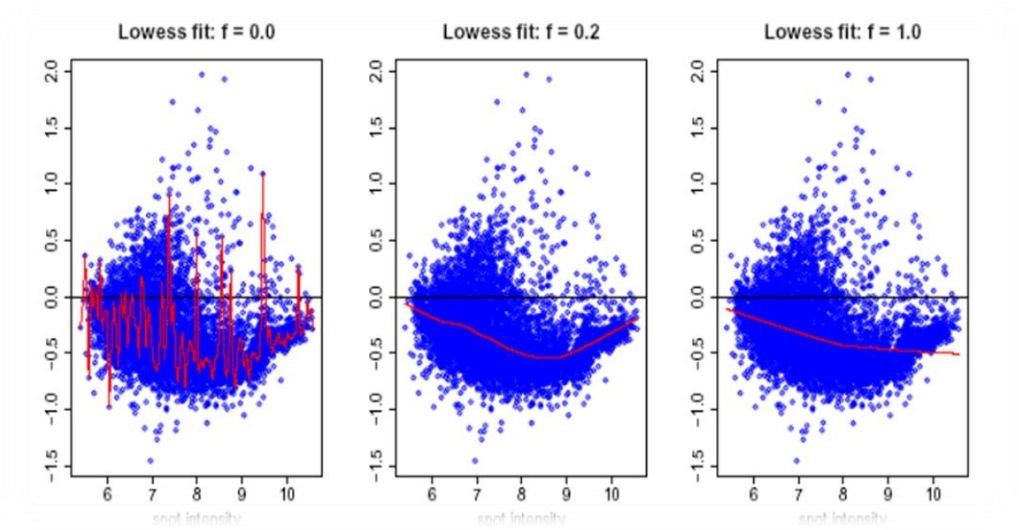
μ_{ik} → Log scale expression level for gene k on array i

3.4.2 Η μέθοδος Κανονικοποίησης Lowess

Η μέθοδος κανονικοποίησης lowess (locally weighted polynomial regression) μας επιτρέπει να κάνουμε συγκρίσιμα τα μονοκυτταρικά δείγματα με non umplified bulk. Πιο συγκεκριμένα είναι μια μη παραμετρική μέθοδος παλινδρόμησης (regression) που διορθώνει τις παρασιτικές μεταβολές που παρατηρούνται ανά σημεία στα δεδομένα προς ανάλυση. Η βασική ιδέα είναι ότι η πρόβλεψη για ένα σημείο x της συνάρτησης παλινδρόμησης G(x) μπορεί να προσεγγιστεί τοπικά. Η τοπική προσέγγιση της παλινδρόμησης , λαμβάνεται μέσω της εφαρμογής μιας μάσκας παλινδρόμησης (fitting regression surface) στα δεδομένα που βρίσκονται γύρω από μία επιλεγμένη γειτονιά του σημείου x (neighborhood). Πιο συγκεκριμένα στη μέθοδο lowess εφαρμόζεται βεβαρυμμένη έκδοση των ελαχίστων τετραγώνων στο κέντρο της «γειτονιάς». Η ακτίνα της γειτονιάς επιλέγεται έτσι ώστε να περιλαμβάνει ένα συγκεκριμένα ποσοστό των δεδομένων. Στην άμεση εφαρμογή της μεθόδου το fitting γίνεται για κάθε σημείο της συνάρτησης παλινδρόμησης. Εναλλακτικά μία απλοποιημένη έκδοση της μεθόδου (η οποία καταναλώνει πολύ λιγότερους υπολογιστικούς πόρους) είναι να εφαρμοστεί ένα τοπικό fitting σε ένα δείγμα σημείων και στη συνέχεια να εφαρμόσει παντού αυτά τα τοπικά πολυώνυμα για να παραχθεί η καμπύλη παλινδρόμησης.

Η μέθοδος lowess προκειμένου να εφαρμοστεί κάνει τις εξής δύο υποθέσεις:

- Τα περισσότερα γονίδια της μικροσυστοιχίας δεν είναι εκφρασμένα μεταξύ των δειγμάτων.
- Ο αριθμός των υπέρ-εκφρασμένων και των υπό-εκφρασμένων των γονιδίων σε κάθε επίπεδο έντασης είναι περίπου ο ίδιος σε κάθε πλακίδιο (Xiong et al. 2008).



Εικόνα 13. Fitting με την μέθοδο lowess για διάφορους παράγοντες εξομάλυνσης [E.10]

Πλεονεκτήματα Lowess

Το βασικό πλεονέκτημα της μεθόδου είναι ότι σαν είσοδο χρειάζεται μόνο να προσδιοριστεί ο παράγοντας εξομάλυνσης και ο βαθμός του τοπικού πολυωνύμου. Επίσης, λόγω του τοπικού χαρακτήρα της μοντελοποιεί πολύπλοκα σύνολα δεδομένων για τα οποία μπορεί να μην υπάρχει κανένα θεωρητικό μοντέλο. Ακόμα παρόλο που είναι φαινομενικά απλή έχει μια πολύπλοκη ντετερμινιστική δομή.

Μειονεκτήματα Lowess

Ένα βασικό μειονέκτημα της μεθόδου είναι ότι για να εφαρμοστεί σωστά απαιτεί πολυπληθή και συμπυκνωμένα σύνολα δεδομένων για να παράγει ικανοποιητικά μοντέλα. Αυτό βέβαια είναι λογικό να συμβαίνει αφού για να εκτελεστεί σωστά η τοπική πρόβλεψη πρέπει να έχει καλά θεμελιωμένη πληροφορία. Στην περίπτωση των δεδομένων από μικροσυστοιχίες, η μέθοδος αυτή ενδείκνυται να εφαρμοστεί. Επίσης το τίμημα της καλής μοντελοποίησης είναι ένας πολύπλοκος μαθηματικός τύπος , γεγονός το οποίο δυσκολεύει την μεταφορά των αποτελεσμάτων μεταξύ των ερευνητών. Τέλος η πολυπλοκότητα της μεθόδου είναι υπολογιστικά αυξημένη και αρκετά επίπονη. Η υπολογιστική πολυπλοκότητα μπορεί να αυξηθεί ακόμα περισσότερο αν χρησιμοποιηθεί η επαναληπτική έκδοση της μεθόδου για την μείωση της ευαισθησίας στις ακραίες τιμές.

3.4.3 MvA plots

Τα δεδομένα των μικροσυστοιχιών κανονικοποιούνται με βάση το unamplified control , προκειμένου να εξαλειφθούν τα systematic biases που μπορεί να οφείλονται στην χρωστική ουσία κατά την διαδικασία της υβριδοποίησης τους ή και σε άλλους παράγοντες Για τον έλεγχο της ανάγκης της κανονικοποίησης, καθώς και για την επιβεβαίωσή της, χρησιμοποιείται η μέθοδος "MA plot", η οποία αναπαριστά την κατανομή των λόγων κόκκινων / πράσινων σημάτων φθορισμού ως προς τη μέση ένταση. Ο όρος "M" αποτελεί το δυαδικό λογάριθμο της έντασης του λόγου κόκκινο/πράσινο και ο όρος "A" αποτελεί το λογάριθμο της μέσης έντασης για κάθε στοιχείο. Η γραφική αναπαράσταση αυτού, με τους όρους "M" και "A" στους άξονες "y" και "x" αντίστοιχα, δίνει μία εικόνα για την κατανομή των δεδομένων, την εξάρτηση από τον θόρυβο που εισάγουν οι διαφορές στα σήματα φθορισμού και τα αποτελέσματα πριν και μετά την κανονικοποίηση. Το MA plot , επομένως είναι ένα σχεδιάγραμμα κατανομής εντάσεων αναλογίας (ratio) ‘M’ , και μέσου όρου εντάσεων (average) ‘A’. Το M, A ορίζονται παρακάτω [84]:

$$A = \tfrac{1}{2} \log_2 (RG) = \tfrac{1}{2} (\log_2 (R) + \log_2 (G))$$

$$M = \log_2 (R /G) = \tfrac{1}{2} (\log_2 (R) - \log_2 (G))$$

3.4.4 Συντελεστής Συσχέτισης (Correlations Coefficient)

Correlation Coefficient ή συντελεστής συσχέτισης (ή πολλές φορές αναφέρεται και ως PPMCC ή PCC ή Pearson's r) , στις στατιστικές μελέτες είναι ένα μέτρο της γραμμικής εξάρτησης (συσχέτισης) , μεταξύ των δύο μεταβλητών X και Y , δίνοντας μια τιμή μεταξύ του 1 και του -1 , όπου 1 είναι συνολικά θετική συσχέτιση, 0 είναι καμία συσχέτιση και −1 είναι συνολικά αρνητική συσχέτιση. Χρησιμοποιείται ευρέως στις επιστήμες ως μέτρο του βαθμού της γραμμικής εξάρτησης μεταξύ των δύο μεταβλητών.

Ο συντελεστή συσχέτισης, συμβολίζεται με το γράμμα r και είναι η συνδιακύμανση των δύο μεταβλητών διαιρούμενη δια του γινομένου των τυπικών τους αποκλίσεων. Αν έχουμε ένα dataset {x 1,..., xn} που περιέχει n τιμές και ένα άλλο σύνολο δεδομένων {Y1,..., yn} που περιέχουν n τιμές , ο τύπος του r είναι:

$$r = r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Όπου ,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$, παρόμοια ορίζεται και το $\bar{y}$

Ένας άλλος τύπος που εκφράζει τον συντελεστή συσχέτισης είναι:

$$r = r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}$$

Όπου , s_x η τυπική απόκλιση και ορίζεται

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$, παρόμοια ορίζεται και το s_y

3.4.5 Global Normalization

Στις μικροσυστοιχίες cDNA, ο στόχος της κανονικοποίησης είναι η εξισορρόπηση των εντάσεων σημάτων φθορισμού των δύο χρωστικών (πράσινη Cy3 και κόκκινη Cy5) και η εξασφάλιση συγκρίσιμων αποτελεσμάτων από διαφορετικές πειραματικές μετρήσεις.

Η διαδικασία της βαφής των δειγμάτων εισάγει πειραματικό θόρυβο και μάλιστα πολυπαραγοντικό. Το γεγονός αυτό αποδεικνύεται προφανέστατα στην

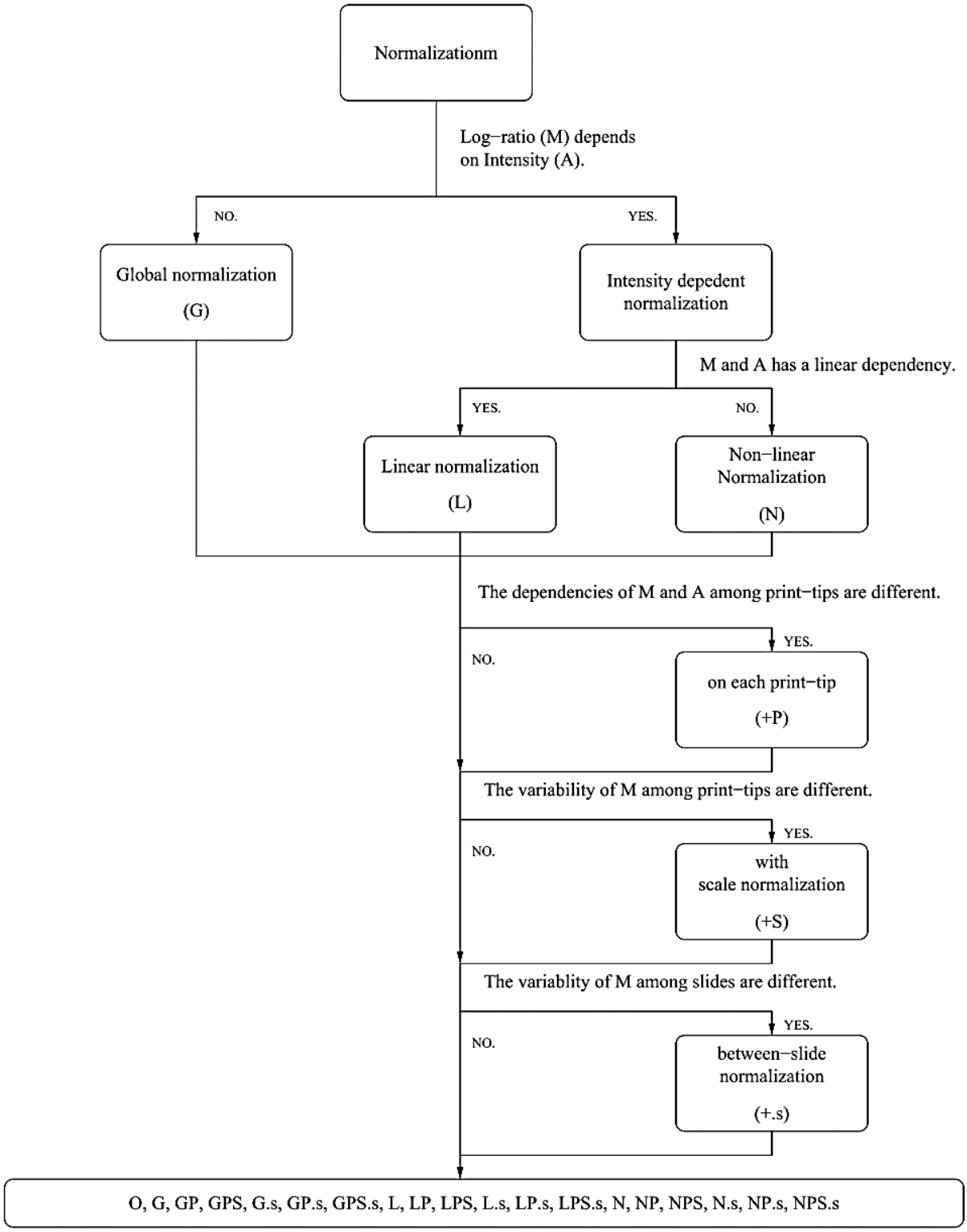
περίπτωση όπου δύο ίδια δείγματα mRNA βάφουν με δύο διαφορετικές χρωστικές και στη συνέχεια υβριδοποιηθούν στο ίδιο πλακίδιο / GeneChip. Σε αυτήν την περίπτωση δε θα έχουμε ποτέ την ίδια μέση ένταση στο σήμα φθορισμού. Αντίθετα, στις περισσότερες περιπτώσεις θα είχαμε μεγαλύτερη ένταση για την πράσινη χρωστική. Το σφάλμα αυτό πηγάζει από μία ποικιλία παραγόντων, συμπεριλαμβανομένων των φυσικοχημικών ιδιοτήτων των χρωστικών ουσιών, ευαισθησία στο φως και τη θερμοκρασία, χρόνοι ημιζωής (relative half-life), αποτελεσματικότητα στην ενσωμάτωση της βαφής, πειραματική μεταβλητότητα στην σύζευξη των probes και στα στάδια επεξεργασίας δεδομένων, ρυθμίσεις της συσκευής κατά τη συλλογή δεδομένων, κλπ. Πολλοί από αυτούς τους παράγοντες, οι οποίοι σχετίζονται ή δε σχετίζονται με τη φύση του δείγματος, παρουσιάζουν μοναδικές (για κάθε πειραματική διαδικασία) δυσκολίες για μία ορθή ολική κανονικοποίηση (global normalization).

Επιπλέον, τα σχετικά επίπεδα έκφρασης των γονιδίων (μετρημένα ως προς λογαριθμικές αναλογίες), κατά την επανάληψη όμοιων πειραμάτων, εμφανίζουν αποκλίσεις, λόγω του ότι η πειραματική διαδικασία δεν μπορεί ποτέ να επαναληφθεί πανομοιότυπα. Γι' αυτό, απαιτείται μία προσαρμογή κλίμακας (scale adjustment) στις τιμές των δεδομένων, ούτως ώστε οι ακραίες τιμές, που συνεπάγονται θόρυβο, να μην κυριαρχούν κατά τον υπολογισμό των μέσω όρων των σχετικών εκφράσεων γονιδίων για επανειλημμένα πειράματα.

Ακόμα, μια σειρά παραγόντων που επηρεάζουν τις πειραματικές μετρήσεις αφορούν το ίδιο το βιολογικό δείγμα ή τη διαδικασία προετοιμασίας του για την εισαγωγή του στη μικροσυστοιχία.

- ✓ Κυτταρική ετερογένεια (Heterogeneity of cells)
- ✓ Ενίσχυση του RNA (RNA amplification)
- ✓ Υποβάθμιση του RNA (Time-dependent RNA degradation during IVT)

Για τη διόρθωση των παραπάνω σφαλμάτων και την εξαγωγή ασφαλών πειραματικών αποτελεσμάτων και βιολογικών συμπερασμάτων, έχουν προταθεί μία σειρά από τεχνικές. Μία κατηγορία τεχνικών αφορά καθαρά τη βελτίωση και την τυποποίηση της πειραματικής διαδικασίας, ενώ μία άλλη κατηγορία αφορά την συνεισφορά υπολογιστικών μεθόδων για την εξάλειψη ή το μετριασμό των πηγών θορύβου. Όσον αφορά τις υπολογιστικές μεθόδους, υπάρχει τεράστια βιβλιογραφία και συγκριτικές μελέτες .Στην εικόνα 14 και στον πίνακα 2 συνοψίζονται οι βασικότερες προσεγγίσεις.



Εικόνα 14. Διάγραμμα ροής (flowchart) όλων των τύπων κανονικοποίησης

Πίνακας 2. Λίστα με τις διάφορες μεθόδους κανονικοποίησης

Method	Notation	Description
	O	Original Data
Global	G	Global median normalization
	GP	Global median normalization on each print-tip
	GPS	Global median normalization on each print-tip with scale normalization
	G.s	Global median normalization and between-slide scale normalization
	GP.s	Global median normalization on each print-tip and between-slide scale normalization
	GPS.s	Global median normalization on print-tip with scale normalization and between-slide scale
Linear	L	Intensity dependent linear regression normalization
	LP	Intensity dependent linear regression normalization on each print-tip
	LPS	Intensity dependent linear regression normalization on each print-tip with scale normalization
	L.s	Intensity dependent linear regression normalization and between-slide scale normalization
	LP.s	Intensity dependent linear regression normalization on each print-tip and between-slide scale normalization
	LPS.s	Intensity dependent linear regression normalization on each print-tip with scale normalization and between-slide scale normalization
Nonlinear	N	Intensity dependent nonlinear regression normalization (LOWESS)
	NP	Intensity dependent nonlinear regression normalization (LOWESS) on each print-tip
	NPS	Intensity dependent nonlinear regression normalization (LOWESS) on each print-tip with scale normalization
	N.s	Intensity dependent nonlinear regression normalization (LOWESS) and between-slide scale normalization
	NP.s	Intensity dependent nonlinear regression normalization (LOWESS) on each print-tip and between slide scale normalization
	NPS.s	Intensity dependent nonlinear regression normalization (LOWESS) on each print-tip with scale normalization and between-slide scale normalization

3.4.6 Clustering (Ομαδοποίηση δεδομένων)

Το clustering των δεδομένων γίνεται με σκοπό την ομαδοποίηση των δεδομένων με βάση κάποια ομοιότητα. Υπάρχουν πολλοί μέθοδοι που clustering. Χωρίζονται σε 2 βασικές κατηγορίες. Supervised, όπου υπάρχει κάποια αρχική γνώση για τον τρόπο που θα κατηγοριοποηθούν τα δεδομένα και unsupervised όπου δεν υπάρχει καμία γνώση για τον τρόπο με τον οποίο θα κατηγοριοποιηθούν τα δεδομένα, όπως είναι η δική μας περίπτωση.

Τα microarray data αναπαριστώνται κυρίως σε έναν δισδιάστατο πίνακα, όπου οι γραμμές είναι τα γονίδια και οι στήλες τα samples / μετρήσεις / πειράματα. Το clustering ως προς τις γραμμές αναμένεται να δώσει πληροφορίες για χρονικά ή τοπικά patterns σχετικά με την έκφραση των γονιδίων, δηλαδή με ποια χρονικά σειρά εκφράζονται ή από ποια περιοχή του γονιδιώματος. Το clustering ως προς τις στήλες αναμένεται να δώσει πληροφορίες ως τον έλεγχο ποιότητας της πειραματικής διαδικασίας (αν έγινε σωστά το πείραμα), την κατηγοριοποίηση των samples με βάση κάποιο χαρακτηριστικό (π.χ. από το ίδιο όργανο) ή την αναγνώριση και ταυτοποίηση νέων κατηγοριών, π.χ. καρκινικών.

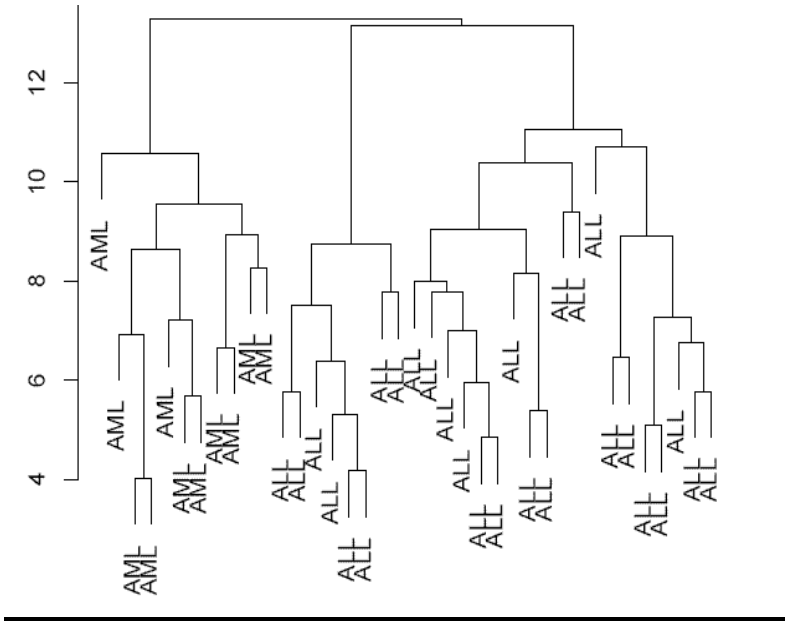
Το clustering εφαρμόζεται με βάση 2 παραμέτρους, την απόσταση και τον αλγόριθμο. Η απόσταση είναι μια μετρική η οποία έχει ως σκοπό να ποσοτικοποιήσει την ομοιότητα ανάμεσα στα δεδομένα. Ο αλγόριθμος είναι η λογική διαδικασία που ακολουθείται για την ομαδοποίηση των δεδομένων, με τελικό σκοπό η απόσταση μεταξύ των δεδομένων να είναι η μικρότερη δυνατή και η απόσταση μεταξύ των κλάσεων η μεγαλύτερη δυνατή.

Η επιλογή της κατάλληλης μετρικής / μεθόδου για την απόσταση εξαρτάται από τα δεδομένα μας και τι ακριβώς ομοιότητες / συσχετίσεις ψάχνουμε να βρούμε.

Η Ευκλείδεια απόσταση μετράει τις απόλυτες διαφορές. Είναι ιδανική για περιπτώσεις όπου στο κάθε cluster υπάρχει κάποιο κοινό centroid, π.χ. ένας μέσος όρος. Τα δεδομένα κατηγοριοποιούνται στους clusters με βάση την απόσταση από το κεντροειδές ενώ οι clusters διαχωρίζονται με βάση των απόσταση των κεντροειδών τους. Σε όλα τα δεδομένα αποδίδονται τα ίδια βάρη. Δηλαδή αναζητά το κεντροειδές της κλάσης με βάση την αναζήτηση του μέσου όρου των δεδομένων.

3.4.6.1 Hierarchical clustering

Με αυτήν την προσέγγιση οι ομοιότητες των δεδομένων αναπαρίστανται με δενδροδιαγράμματα. Αρχικά όλα τα δεδομένα θεωρούνται ως ξεχωριστά clusters. Σε κάθε βήμα δεδομένα με την ελάχιστη απόσταση μεταξύ τους ομαδοποιούνται σε clusters. Από το δεύτερο βήμα και μετά ομαδοποιούνται οι κόμβοι του δένδρου, έως ότου να φτάσουμε σε έναν μοναδικό cluster και θα βρίσκεται στο μέγιστο ύψος του δένδρου. Ανάλογα με το ύψος του δένδρου (δηλαδή σε κάθε επανάληψη του αλγορίθμου) καταλήγουμε και σε διαφορετικό ύψος.



Εικόνα 15. Παράδειγμα - Διαφορετικοί τύποι Λευχαιμίας. Το Clustering βασίστηκε σε 150 γονίδια με την highest variance κατά μήκος των δειγμάτων

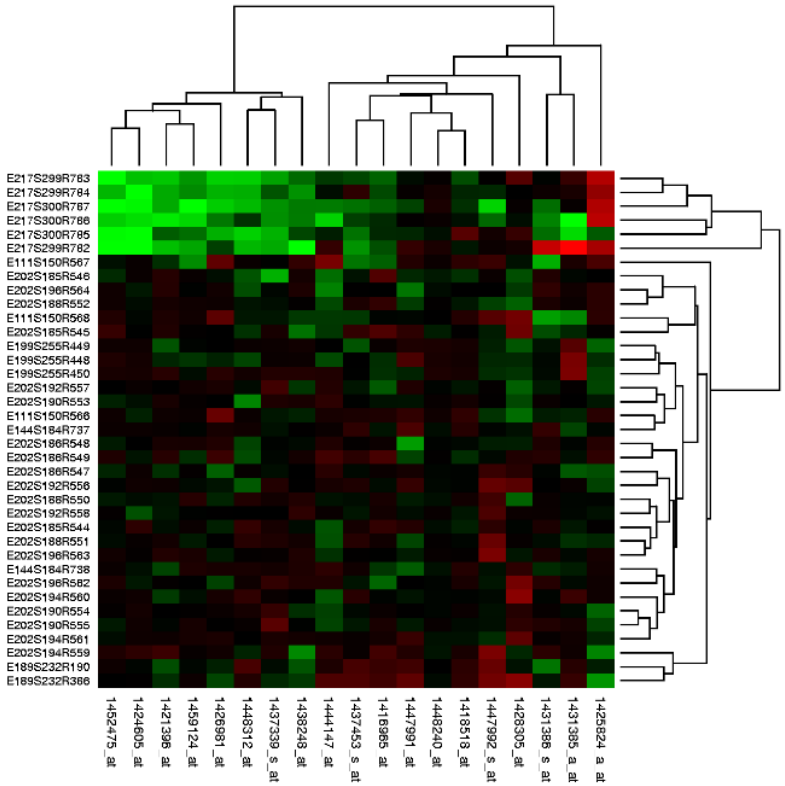
Υπάρχουν διάφορες μέθοδοι με τις οποίες μπορεί να υλοποιηθεί το Clustering των δεδομένων. Ανάλογα με το ποια κατηγοριοποιεί τα αποτελέσματα καλύτερα καθώς και τι ερευνάται ο επιστήμονας καλείται να διαλέξει. Τέτοιες μέθοδοι είναι οι:

- Η 'centroid' μετρική αρχικά θεωρεί όλα τα δεδομένα ως υποψήφια κεντροειδή και έπειτα υπολογίζει τις μακρύτερες αποστάσεις μεταξύ τους.
- Η ‘complete’ μετρική αναζητά απ’ευθείας τις μέγιστες αποστάσεις ανάμεσα στα δεδομένα, ενώ η ‘shortest’ αναζητά απ’ευθείας τις ελάχιστες αποστάσεις ανάμεσα στα δεδομένα.
- Οι μέθοδοι 'weighted' και 'median' είναι αντίστοιχες των 'average' και 'centroid', αλλά δεν αντιμετωπίζουν τα δεδομένα ως ίσα. Σε κάθε βήμα αλλάζουν τα βάρη των δεδομένων.
- Η μέθοδος 'ward' υπολογίζει τις αποστάσεις μέσω βεβαρυμένων αθροισμάτων τετραγώνων ανάμεσα στα κεντροειδή.

3.4.6.2 Heat Map

Το Heat Map είναι ένας τρόπος γραφικής αναπαράστασης δεδομένων, όπου οι τιμές των στοιχείων ενός πίνακα αναπαρίστανται με διαβαθμισμένα χρώματα ανά pixel σε μία δισδιάστατη εικόνα. Συχνά συμπληρώνονται από fractal maps και tree maps για να απεικονίσουν τις εξαρτήσεις ιεραρχίας μεταξύ των δεδομένων. Στη μοριακή Βιολογία τα heatmaps χρησιμοποιούνται για τη κατάδειξη του επιπέδου έκφρασης πολλών γονιδίων σε μία σειρά από δείγματα, όπως για παράδειγμα ομάδες κυττάρων σε διαφορετικά επίπεδα ανάπτυξης από διαφορετικούς ασθενείς. Συνήθως, οι στήλες αναπαριστούν τα δείγματα και οι γραμμές τα γονίδια.

Η αναπαράσταση των δεδομένων με Heat Maps συνήθως είναι το τελικό στάδιο επεξεργασίας των δεδομένων, αφού έχουν εφαρμοστεί οι τεχνικές κανονικοποίησης και ομαδοποιήσης - clustering. Οπότε, τα δεδομένα είναι ομαδοποιημένα και γίνονται εμφανείς οι εξαρτήσεις μεταξύ των δειγμάτων. Δηλαδή, δείγματα τα οποία βρίσκονται σε διαδοχικές στήλες ομαδοποιήθηκαν με βάση κάποιο χαρακτηριστικό, π.χ. ανήκουν στον ίδιο ιστό. Επίσης δείγματα τα οποία για κάποιο λόγο δεν ταιριάζουν με τα υπόλοιπα απομονώνεται εμφανώς (έλεγχος της πειραματικής διαδικασίας). Αντίστοιχα, γονίδια που ομαδοποιήθηκαν σε διαδοχικές γραμμές αποκαλύπτουν πληροφορίες για το ποιες περιοχές του γονιδιώματος εκφράστηκαν και με ποια σειρά.



Εικόνα 16. Αναπαράσταση Heat map , με δεδομένα από DNA microarray [E.10]

Κεφάλαιο 4 - Προτεινόμενη μεθοδολογία και Αποτελέσματα

4.1 Δεδομένα προς επεξεργασία

Τα δεδομένα που επεξεργαστήκαν σε αυτή την εργασία αποτελούνται από:

- 5 κυτταρικές σειρές, όπου η καθεμία περιέχει διαφορετικό αριθμό δειγμάτων και διαφορετικό αριθμό κυττάρων(single cells, 40,1000,5000 cells)
- 5 bulks, με τον όρο αυτό εννοείται το unamplified δείγμα, δηλαδή εκείνο το δείγμα όπου δεν υπέστη την πειραματική διαδικασία του amplification
- 2 Κυκλοφορούντα Καρκινικά Κύτταρα
- 3 μεγάλες βάσεις δεδομένων με διαφορετικό αριθμό και τύπο δειγμάτων η κάθε μια.

Συγκεκριμένα επεξεργαστήκαμε τις Κυτταρικές Σειρές που παρουσιάζονται στον παρακάτω πίνακα. Αυτές οι Κυτταρικές Σειρές, μαζί με τα δύο Κυκλοφορούντα Καρκινικά Κύτταρα αποτελούν το Query dataset. Όλες οι Κυτταρικές Σειρές περιέχουν δείγματα της σειράς PrimeView εκτός από κάποια Bulks (unamplified RNA) που είναι της σειράς U133 και αναγράφονται με αντίστοιχη ένδειξη.

Πίνακας 3. Οι Κυτταρικές Σειρές που επεξεργαστήκαμε στην παρούσα εργασία

Όνομα Κυτταρικής Σειράς	**Αριθμός δειγμάτων (και bulks αν υπάρχουν)**	**Ονόματα δειγμάτων που περιέχει κάθε κυτταρική σειρά**
Cell Line 1	3 samples και το αντίστοιχο bulk	1. 1Cell_line1_1c 2. 2Cell_line1_1c 3. 3Cell_line1_1c **4. Cell_line1_bulk_U133Plus**
Cell Line 2	3 samples και το αντίστοιχο bulk	1. 1Cell_line2_1c 2. 2Cell_line2_1c 3. 3Cell_line2_1c 4. **Cell_line2_bulk_U133Plus**
ell Line 3	3 samples και το αντίστοιχο bulk	1. 1Cell_line3_1c 2. 2Cell_line3_1c 3. 3Cell_line3_1c **4. Cell_line3_bulk_U133Plus**
Cell Line 4 **MCF7**	13 samples και το αντίστοιχο bulk –mcf7	1. 1Cell_line4_1c 2. 2Cell_line4_1c 3. Cell_line4_1000c 4. 5000c_Eb-opt 5. 5000c K-OPT

		6. 1cB-Eb-opt 7. 1cC_K-opt 8. 1c Eb-Optimized Eberwine 9. 1c_Eb-RNA-opt_Eberwine 10. 1c_K-RNA-opt-Prime_A2_RNA 11. 1c_K-RNA-opt_D2_A2_RNA 12. 1c_K-opt_A2_Sm_1cell 13. 1c_K_ExTaq_SSIII_1cell 14. **MCF7 _bulk_PrimeView**
Cell Line 5	10 samples και το αντίστοιχο bulk	1. 1CellLine5_1c_Eb 2. 1CellLine5_1c_K 3. 2_CellLine5_1c_Eb 4. 2_CellLine5_1c_K 5. 3_CellLine5_1c_Eb 6. 3_CellLine5_1c_K 7. CellLine5_1000c_Eb 8. CellLine5_1000c_K 9. CellLine5_40c_Eb 10. CellLine5_40c_K **11. Cell_line5_bulk_U133Plus**
CTCs	2 samples , δεν είχαμε reference	1. A_CTC 2. 13_CTC

Επομένως συνολικά επεξεργαστήκαμε 5 κυτταρικές σειρές , όπου αποτελούνται από 33 δείγματα . Ακόμα επεξεργαστήκαν 5 bulks και 2 Κυκλοφορούντα καρκινικά κύτταρα. Είναι σημαντικό να αναφερθεί ότι δεν γνωρίζουμε την ταυτότητα αυτών των δειγμάτων, εκτός από αυτά που έχουν την ένδειξη MCF7 , όπου υποδηλώνουν καρκίνο του Μαστού. Δηλαδή από τα 40 δείγματα είναι γνωστά μόνο τα 14 (Cell line 4) και το bulk MCF7. Όλες οι παραπάνω Κυτταρικές Καλλιέργειες και δείγματα , προήλθαν από το ΙΤΕ (Ίδρυμα Τεχνολογίας & Έρευνας) Ηρακλείου, από το Ινστιτούτο Μοριακής Βιολογίας και Βιοτεχνολογίας.

Εκτός από τα παραπάνω δείγματα όπως προαναφέρθηκε χρησιμοποιήθηκαν και τρείς βάσεις δεδομένων για την ταυτοποίηση των παραπάνω Κυτταρικών Σειρών και των Κυκλοφορούντων Καρκινικών Κυττάρων. Η πρώτη βάση δεδομένων που χρησιμοποιήθηκε περιείχε 230 δείγματα, 677 η δεύτερη και 2081 η τρίτη. Παρόλο που με την πρώτη βάση εξήχθησαν κάποια επιθυμητά αποτελέσματα, χρησιμοποιήθηκαν και οι άλλες δύο βάσεις δεδομένων , που περιείχαν μεγαλύτερο αριθμό δειγμάτων , επεκτείνοντας τις δυνατότητες των αλγορίθμων και των προγραμμάτων που παρήγαμε ώστε να αυξηθεί η πιθανότητα ταυτοποίησης με ιστούς και κυτταρικές σειρές μελλοντικά.

Οι βάσεις δεδομένων που χρησιμοποιήθηκαν αποτελούν το reference dataset για την έρευνά μας , καταγράφονται στον επόμενο πίνακα και είναι τύπου U133 Plus 2.0 Array :

Πίνακας 4. Στοιχεία των Βάσεων Δεδομένων που χρησιμοποιήθηκαν για την ταυτοποίηση των Κυτταρικών Σειρών.

Βάσεις Δειγμάτων	**# Δειγμάτων**	**Πηγή**
1st Database	230 (Κυτταρικές Σειρές)	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE34211
2nd Database	677 (Κυτταρικές Σειρές)	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7307
3rd Database	2081 (Ιστοί)	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2109

Οι βάσεις δεδομένων που χρησιμοποιήθηκαν , αρχικά είχαν ένα μεγαλύτερο αριθμό δειγμάτων από αυτόν που αξιοποιήθηκε εν τέλει. Για παράδειγμα η πρώτη βάση δεδομένων περιείχε 402 δείγματα , όμως μόνο τα 230 δείγματα ανήκαν στην κατηγορία U133 Plus2 , η οποία και μας ενδιέφερε. Ένα ακόμα παράδειγμα αποτελεί η τρίτη βάση δεδομένων , όπου δύο δείγματα (το GSM277703 και GSM277707), δεν συμπεριλήφθηκαν τελικά στις αναλύσεις μας διότι δεν μας δώσανε expression values. Αυτό συνέβη διότι τα αρχικά αρχεία τους ήταν κατεστραμμένα (corrupted). Έτσι λοιπόν από τις βάσεις δεδομένων επιλέχθηκαν εν τέλει μόνο εκείνα τα δείγματα (2988 δείγματα) τα οποία ήταν συμβατά με τα υπόλοιπα δεδομένα μας και δεν ήταν προβληματικά.

Στόχος είναι να δημιουργηθεί ένα μεγάλο σετ δεδομένων , από μίξη φυσιολογικών και παθολογικών δειγμάτων, προκειμένου:

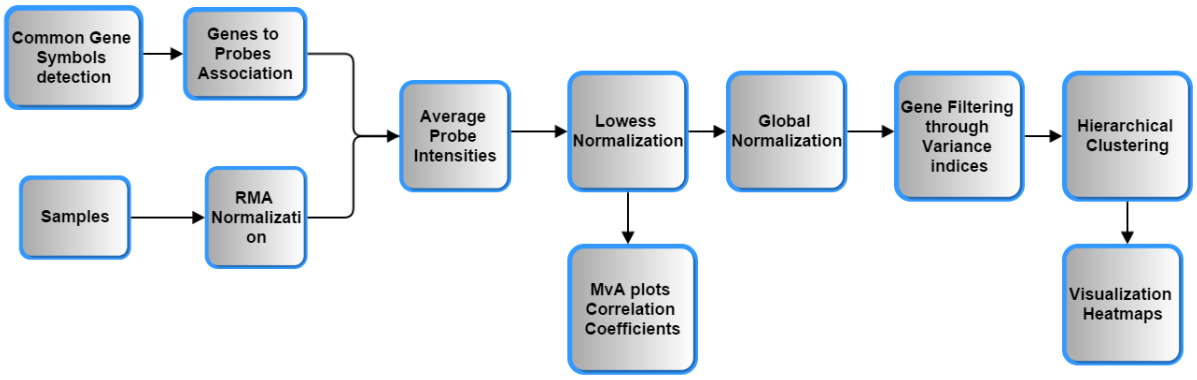
- ✓ να ταυτοποιηθούν δείγματα από τις κυτταρικές σειρές (Cell line 1, Cell line 2, Cell line 3, Cell line 4, Cell line 5)
- ✓ Και των δύο CTCs

4.2 Block Diagram

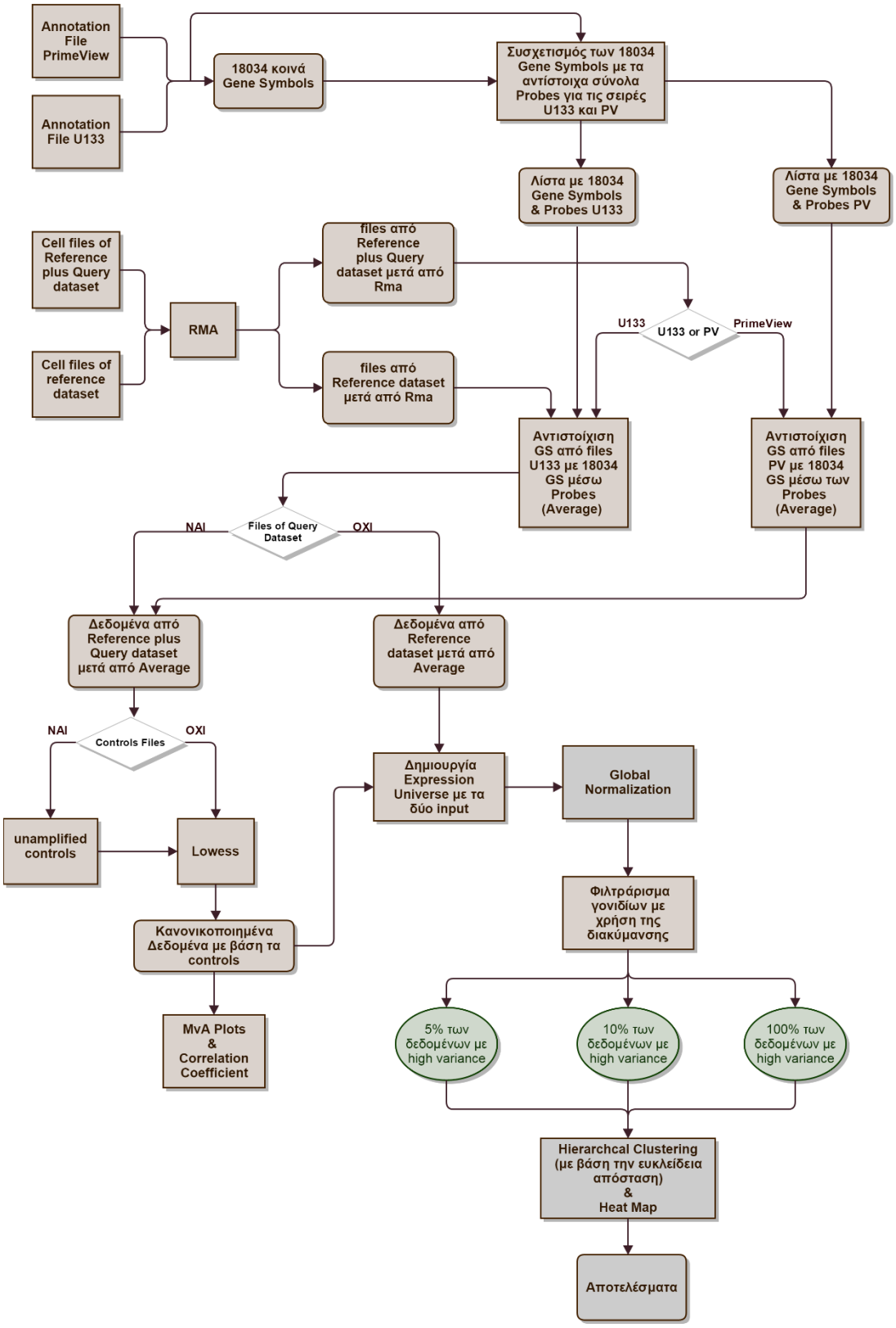
Στόχος αυτής της υποενότητας είναι να περιγραφούν διαγραμματικά όλα τα στάδια της μεθόδου που ακολουθήσαμε. Στην επόμενη εικόνα (Εικόνα 17)φαίνονται συνοπτικά τα βήματα που ακολουθήσαμε για την επεξεργασία και την κατηγοριοποίηση των δεδομένων μας. Στην εικόνα 18 απεικονίζονται , μέσω ενός διαγράμματος πιο περιγραφικά τα στάδια που ακολουθήσαμε. Στις επόμενες υποενότητες επεξηγούνται αναλυτικότερα όλα τα στάδια βήμα βήμα.

Η μεθοδολογία μας , πολύ περιληπτικά περιλαμβάνει τα εξής στάδια:

- Βρίσκουμε τα κοινά γονίδια των σειρών U133 plus 2.0 και PrimeView.
- Συνδέουμε τους ανιχνευτές των σειρών με τα κοινά γονίδια που βρήκαμε στο παραπάνω στάδιο
- Εφαρμόζουμε στα δείγματα την μέθοδο επεξεργασίας Rma
- Βρίσκουμε από τις τιμές των ανιχνευτών που αντιστοιχίστηκαν με τα γονίδια τον μέσο όρο τους.
- Κανονικοποιούμε με την μέθοδο Lowess όσα δείγματα αντιστοιχούν με κάποιο unamplified control.
- Εξετάζουμε την Lowess μέσω των MvA plots και του συντελεστή συσχέτισης των δειγμάτων με τα αντίστοιχα controls.
- Εφαρμόζουμε την μέθοδο Global Normalization.
- Φιλτράρουμε τα γονίδια προκειμένου να κατηγοριοποιήσουμε τα δείγματα κάνοντας χρήση της διακύμανσής τους.
- Ομαδοποιούμε τα δείγματα με Hierarchical clustering
- Εξετάζουμε τα αποτελέσματα της ομαδοποίησης μέσω οπτικοποίησης της κατηγοριοποίησης που επιτεύχθηκε.



Εικόνα 17. Η μεθοδολογία μας διαγραμματικά



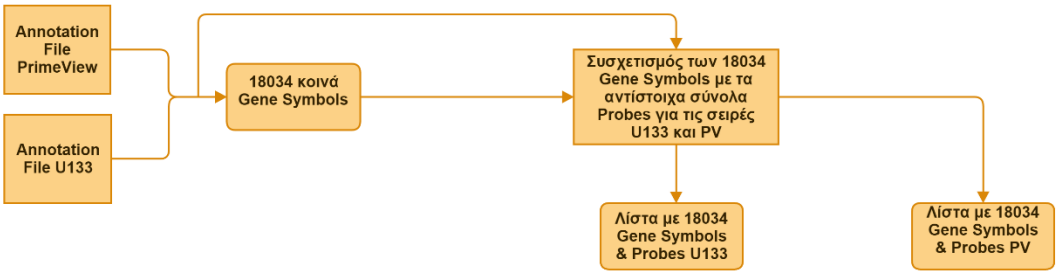
Εικόνα 18. Το διάγραμμα με όλα τα στάδια της μεθόδου που ακολουθήσαμε αναλυτικά.

4.3 Αντιστοίχιση Genes Symbols - Probes στα GeneChips U133 & Prime View

Όπως έχει αναφερθεί τα δεδομένα που χρησιμοποιήθηκαν στην παρούσα εργασία, προέρχονται από δυο Gene Chips της Affymetrix:

- GeneChip® Human Genome U133 Plus 2.0 Array
- GeneChip® PrimeView™ Human Gene Expression Array

Από αυτές τις δύο σειρές επιλέχτηκαν τα μεταξύ τους κοινά γονίδια. Αυτό έγινε με σκοπό να κατασκευαστεί μια βάση δεδομένων μέσω της οποίας θα γίνουν ερευνητικές μελέτες από δεδομένα που πάρθηκαν και από τους δύο τύπους μικροσυστοιχιών. Πιο αναλυτικά αυτή η συσχέτιση ήταν μείζονος σημασίας προκειμένου, για παράδειγμα δεδομένα όπως η Κυτταρική Σειρά 5 (Cell Line 5), όπου τα δείγματά της είναι όλα της σειράς PrimeView να μπορέσουν να κανονικοποιηθούν με βάση το αντίστοιχο unamplified control τους , που είναι του τύπου U133.



Εικόνα 19. Συσχετισμός 18034 κοινών γονιδίων από τα Chips U133 και PrimeView της εταιρίας Affymetrix

Μέσω των annotation αρχείων, που περιέχουν πληροφορίες για τις Probes και τα Gene Symbols και δίνονται από τον κατασκευαστή, βρέθηκαν συνολικά 18034 κοινά γονίδια, τα οποία συσχετίστηκαν με τις 46337 probes της Prime View και 38793 της σειράς U133 (Πίνακας 4). Πρέπει να σημειωθεί ότι σε ένα Gene Symbol μπορεί να αντιστοιχούν περισσότεροι από ένα ανιχνευτές. Το συνολικό νούμερο ανιχνευτών της PrimeView είναι 49372 , ενώ αυτά της U133 54675. Από αυτά 481 Probes της Prime View δεν αξιοποιήθηκαν επειδή στο αρχικό annotation δεν ήταν αντιστοιχισμένα με κάποιο Gene Symbol. Συγκεκριμένα στο πεδίο του gene symbol είχαν την ένδειξη " --- " (τρείς παύλες). Για τον ίδιο λόγο δεν επιλέχτηκαν 9564 Probes της U133 για περαιτέρω αναλύσεις. Ακόμα 2554 Probes της Prime View δεν συμπεριλήφθηκαν στο annotation που δημιουργήσαμε διότι δεν είχαν κοινό Gene Symbol με τα Probes της U133. Επίσης 6318 Probes της U133 δεν συμπεριλήφθηκαν για τον ίδιο λόγο.

Στον παρακάτω πίνακα αναγράφεται ο αριθμός των Probes πρίν και μετά την επεξεργασία συγχώνευσης των δύο microarray PrimeView και U133.

Πίνακας 5. Στοιχεία για Probes U133 και Prime View

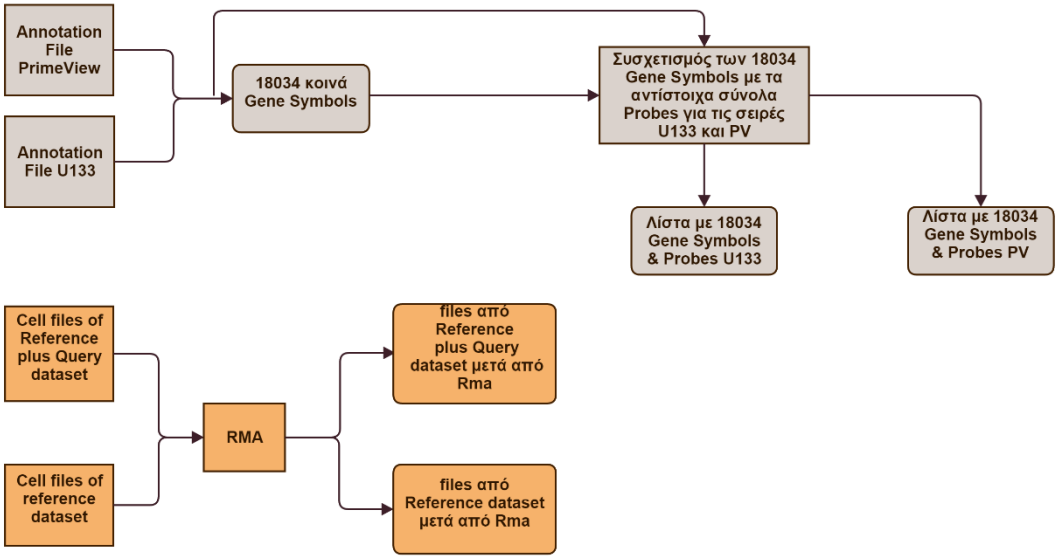
	PrimeView	**UI133**
Αρχικά Probes	49372	54675
Probes που αντιστοιχίστηκαν	46337	38793
Probes που δεν ήταν αντιστοιχισμένα με κάποιο Gene Symbol (“---“) , από το annotation file	481	9564
Probes που δεν αντιστοιχίστηκαν σε κοινό Gene Symbol	2554	6318
Common Gene Symbols	18034	18034

Το αποτέλεσμα που εξήγαμε ήταν δύο λίστες όπου η μία ήταν για το Chip PrimeView και η άλλη για το Chip U133. Η κάθε λίστα είχε δυο στήλες. Η μία στήλη είχε τα Gene Symbols (μόνο τα κοινά τους) και η δεύτερη στήλη τις Probes. Κάθε Probe που αντιστοιχούσε στο ίδιο Gene Symbol με μία άλλη είχε διαφορετική τιμή έντασης. Το πρόβλημα ήταν ότι δεν μπορούσαμε να γνωρίζουμε ποια ήταν η πιο αντικειμενική τιμή που θα μπορούσε να μας οδηγήσει σε αξιόπιστα αποτελέσματα. Στην ενότητα 4.5 περιγράφεται πως λύθηκε αυτό το πρόβλημα.

4.4 Εφαρμογή RMA

Όλα τα δείγματα μας αρχικά επεξεργάστηκαν με τον αλγόριθμο RMA , ανεξάρτητα από το αν ανήκαν στο Reference Dataset ή στο Reference plus Query Dataset ή στην σειρά U133 ή PrimeView. Ο αλγόριθμος RMA, επιλέχτηκε διότι διορθώνει τον θόρυβο που εισάγεται κατά τη διαδικασία παραγωγής των δεδομένων εξόδου από τις μικροσυστοιχίες της Affymentrix και θεωρείται ιδανικός για δεδομένα αυτής της εταιρίας. Περισσότερες πληροφορίες δίνονται στο ενότητα 3.4.1.

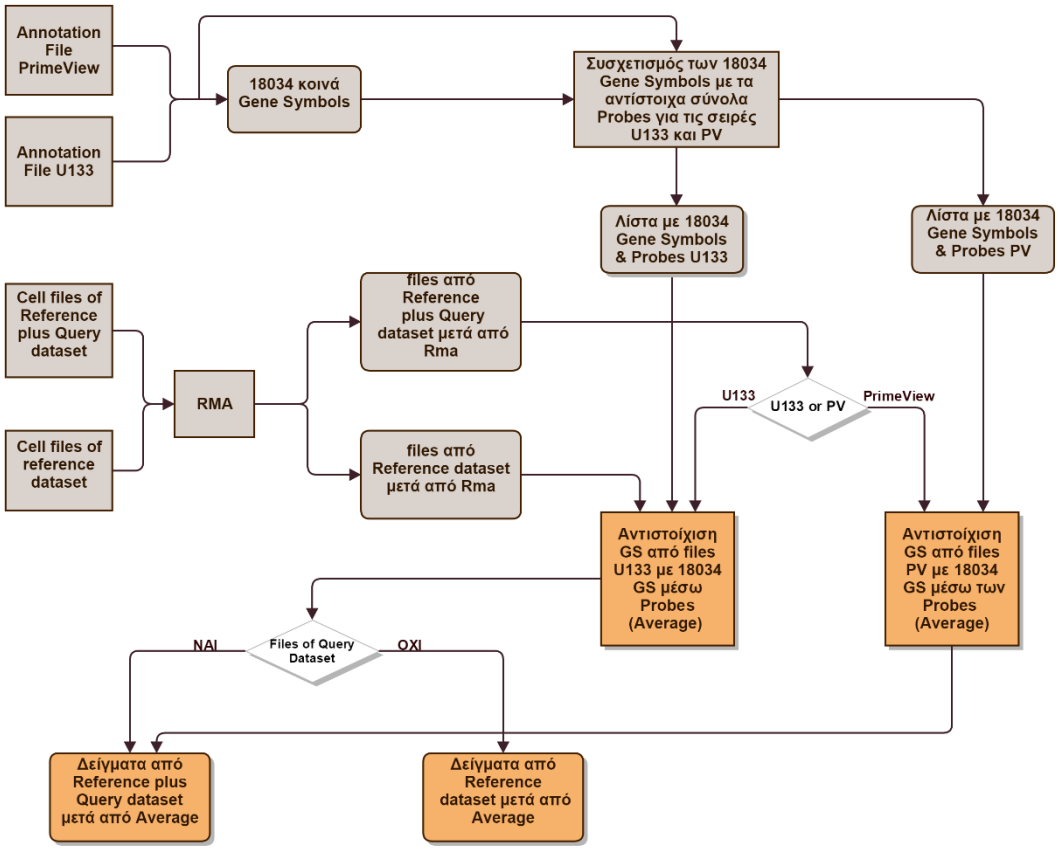
Κάθε δείγμα περνούσε από την διαδικασία της διόρθωσης ξεχωριστά και έπειτα όλα μαζί ενώνονταν σε ένα κοινό αρχείο όπου μπορούσαμε να δούμε την γονιδιακή έκφραση. Συγκεκριμένα το αρχείο αυτό περιείχε τις τιμές των Probes του δείγματός μετά την διόρθωση, τις Probes αντιστοιχισμένες με τις τιμές τους και με τα αντίστοιχα Gene Symbols. Περισσότερες πληροφορίες για το προγραμματιστικό μέρος δίνονται στην ενότητα 5.3.



Εικόνα 20. Η διαδικασία όπου τα Cell Files επεξεργάζονται με τον RMA

1.5 Εφαρμογή μεθόδου Averaging

Όπως προαναφέρθηκε διαφορετικά Probes αντιστοιχούν στο ίδιο Gene Symbol.Προκειμένου να ληφθούν αντικειμενικές τιμές για τις εντάσεις φθορισμού των Gene Symbols , υπολογίστηκε ο μέσος όρος όλων των Probes που έχουν κοινό Gene Symbol. Με αυτή την προσέγγιση εξομαλύνονται οι ακραίες τιμές και μειώνεται ο θόρυβος. Η επιλογή της λίστας που δημιουργήσαμε σε προηγούμενο βήμα και με την οποία γίνεται η αντιστοίχιση των Probes των δειγμάτων , γίνεται με βάση τη σειρά στην οποία ανήκει το εκάστοτε δείγμα. Αν το δείγμα για παράδειγμα είναι της σειράς PrimeView θα χρησιμοποιήσουμε την λίστα της PrimeView και αντίστοιχα για την U133. Τα δείγματα του Reference Dataset είναι αποκλειστικά τύπου U133. Αυτό όμως δεν ισχύει και για το Query Dataset , όπου τα δείγματα είναι και της σειράς Prime View αλλά και της σειράς U133.Ο αλγόριθμος, με τον οποίο εκτελέστηκε έχει ιδιαίτερο ενδιαφέρον, διότι λόγω του μεγάλου όγκου δεδομένων καθώς και της αντιστοίχισης χιλιάδων Probes , έπρεπε να υλοποιηθεί μια ιδιοκατασκευασμένη βάση δεδομένων. Η μέθοδος περιγράφεται στην ενότητα 5.3 .



Εικόνα 21. Η αντιστοίχιση των Probes στα αντίστοιχα Gene Symbols με την μέθοδο Average.

Εκτός από την συσχέτιση των Probes με το Gene Symbol , με βάση τον μέσο όρο των εντάσεων τους, δοκιμάσαμε άλλη μία μεθοδολογία προκειμένου να εξετάσουμε ποια είχε τα βέλτιστα αποτελέσματα. Πιο συγκεκριμένα αντιστοιχίσαμε σε κάθε Gene Symbol την μεγαλύτερη τιμή από όλα τα Probes που αντιστοιχούσανε σε ένα συγκεκριμένο Gene Symbol. Στην περίπτωση που είχαμε ένα Gene Symbol να αντιστοιχεί σε ένα μόνο Probe , κρατούσαμε την τιμή αυτού του Probe. Εν τέλει αποδείχτηκε πως με την πρώτη μεθοδολογία, αυτή του μέσου όρου, προκύπτανε καλύτερα αποτελέσματα, αφού οι ακραίες τιμές ενδέχεται να είναι θόρυβος. Αυτό είναι κάτι που επίσης διαπιστώθηκε από τα correlation coefficient (συντελεστής συσχέτισης). Για όλα τα δείγματα που επεξεργαστήκαμε οι τιμές του συντελεστή συσχέτισης με την μέθοδο του μέσου όρου των Probes ήταν αρκετά υψηλότερες. Ενδεικτικά αναφέρονται οι παρακάτω τιμές , όπου με την μέθοδο του μέσου όρου (Average) των Probes που αντιστοιχούν σε ένα Gene Symbol, είναι εμφανές ότι ο συντελεστής συσχέτισης είναι μεγαλύτερος από ότι με την μέθοδο του μεγίστης (Max) τιμής μιας Probe ανά Gene Symbol.

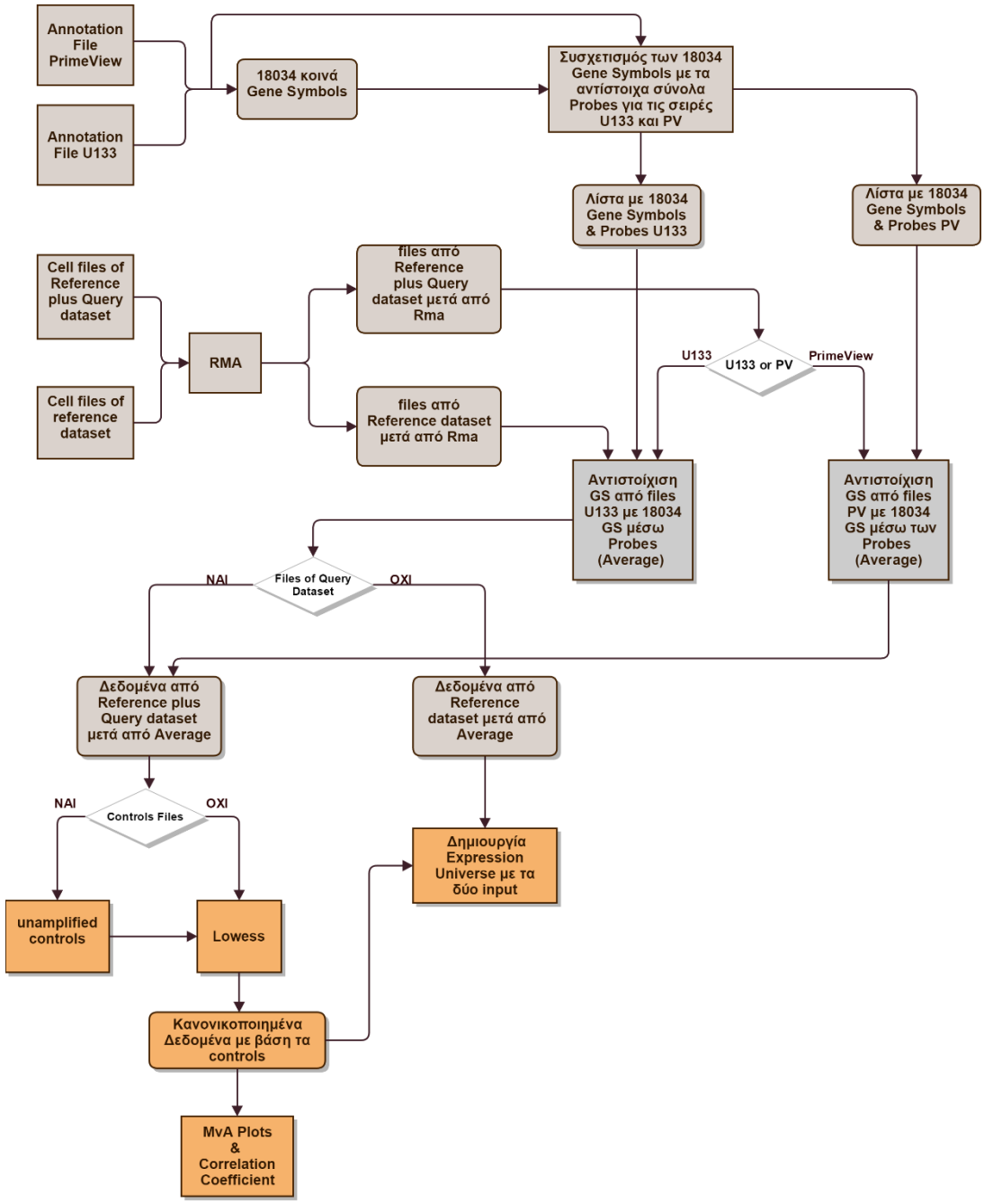
Πίνακας 6. Παράδειγμα συντελεστές συσχέτισης για την μέθοδο αντιστοίχισης των Probes με Average και Max.

Δείγμα	Τιμή Correlation Coefficient πρίν από Lowess - Average	Τιμή Correlation Coefficient μετά από Lowess , με παράθυρο 5% - Average	Τιμή Correlation Coefficient μετά από Lowess , με παράθυρο 5% - Max
1CellLine5_1c_Eb	0,6959	0,7057	0,4020
1CellLine5_1c_K	0,6976	0,6947	0,5378
2_CellLine5_1c_Eb	0,6933	0,7068	0,0922

4.6 Εφαρμογή της μεθόδου κανονικοποίησης Lowess

Ένας από τους βασικότερους λόγους που χρησιμοποιήθηκε η lowess ήταν για να διορθωθεί το σφάλμα που εισάγεται κατά την πειραματική διαδικασία της amplification Eberwine και Kurimoto. Για να επιτευχθεί αυτό, αξιοποιήθηκε το unamplified control. Έτσι κανονικοποιήθηκαν όλα τα samples που είχαν reference . Αυτό έγινε διότι το unamplified reference δεν έχει πειραματικό σφάλμα, γιατί δεν έχει περάσει από amplification και αυτό το καθιστά πιο αξιόπιστο βιολογικά. Κατά συνέπεια μόνο 33 δείγματα κανονικοποιήθηκαν με την μέθοδο Lowess, με βάση τα αντίστοιχα bulks (unamplified).

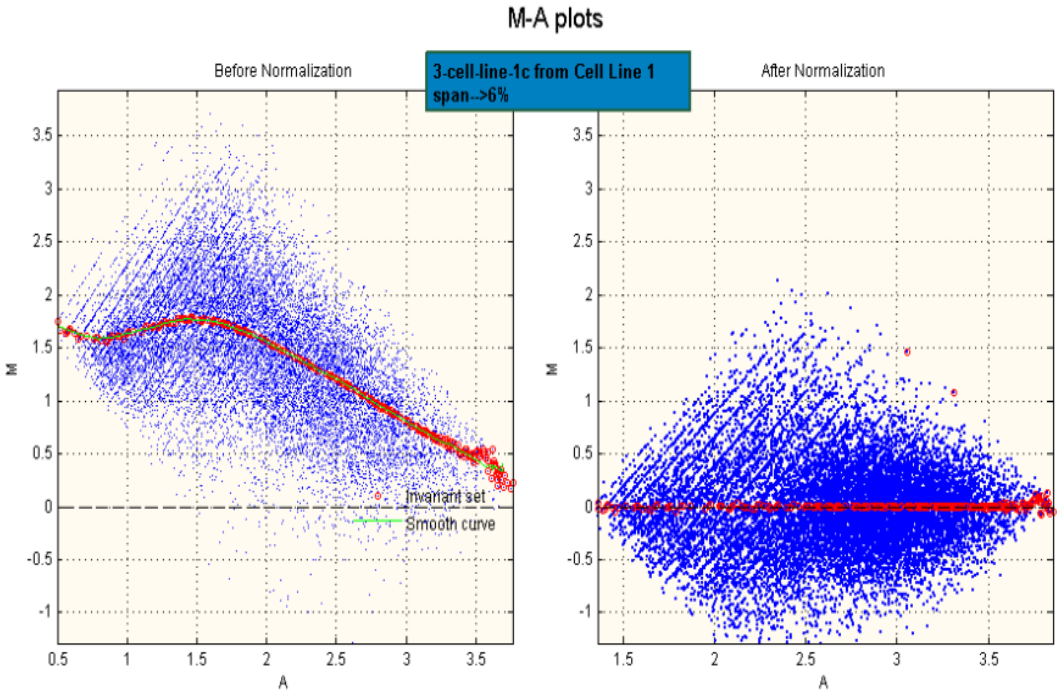
Η lowess, όπως αναφέρθηκε χρησιμοποιείται για smoothing και regression των δεδομένων. Ο τρόπος που εφαρμόζεται είναι με τη χρήση ενός τοπικού παραθύρου στα δεδομένα. Το μέγεθος του παραθύρου είναι καθοριστικός παράγοντας για το smoothing. Ένα μεγάλο μέγεθος παραθύρου οδηγεί σε πιο ομαλό Smoothing αλλά χάνει ορισμένες μικρομεταβολές στην καμπύλη των δεδομένων, ενώ ένα μικρό μέγεθος παραθύρου μπορεί να μοντελοποιήσει σωστά όλες τις μικρομεταβολές των δεδομένων αλλά μέσα σε αυτό συμπεριλαμβάνει και το θόρυβο. Οπότε το βέλτιστο μέγεθος παραθύρου, όπως αναφέρεται και στη βιβλιογραφία το επιλέγει ο ερευνητής μέσω της διαδικασίας δοκιμής και σφάλματος (trial and error), κρίνοντας από τα αποτελέσματα. Δηλαδή το σωστό smoothing είναι κατά κάποιο τρόπο και «τέχνη», δηλαδή εξαρτάται από την εμπειρία του ερευνητή, πέρα από κάποιες αντικειμενικές παραμέτρους. [82]



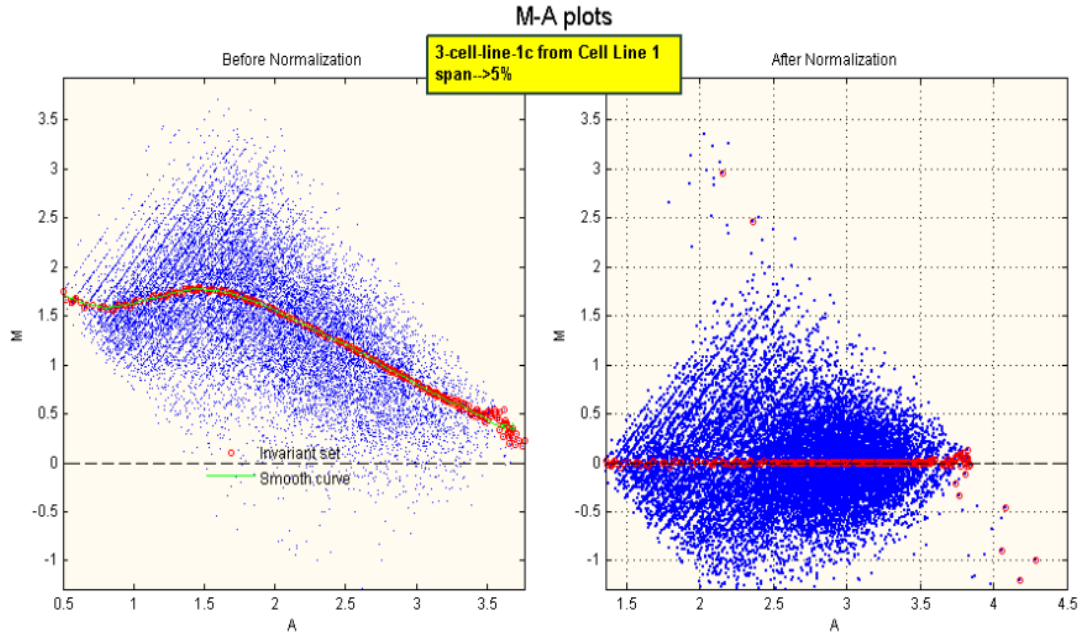
Εικόνα 22. Εφαρμογή κανονικοποίησης Lowess μόνο στα δείγματα που έχουν reference και η δημιουργία Expression Universe με Reference Dataset και Reference plus Query Dataset.

Όλα τα δείγματα κανονικοποιήθηκαν με παράθυρο 5% , εκτός από ένα δείγμα , το δείγμα 3Cell_line1_1c της Κυτταρικής Σειράς 1 , όπου ύστερα από πειράματα και δοκιμές κανονικοποιήθηκε εν τέλει με παράθυρο 6%. Υπάρχουν περιπτώσεις όπου η κανονικοποίηση Lowess θα πρέπει να προσαρμόζεται με βάση τα χαρακτηριστικά του πειραματικού σφάλματος που πρόκειται να κανονικοποιηθεί. Αυτό συμβαίνει επειδή πολλές φορές ένα μεγάλο παράθυρο μπορεί να οδηγήσει σε πιο ομαλό smoothing.

Παρακάτω φαίνεται το διάγραμμα πριν και μετά την κανονικοποίηση Lowess με παράθυρο 0,06 (6%) και 0,05 (5%) αντίστοιχα.



Εικόνα 23. MvA plot δείγματος 3cell_line_1c της Cell Line 3 (Κυτταρικής Σειράς 3) με παράθυρο κανονικοποίησης στη Lowess 0,06



Εικόνα 24. MvA plot δείγματος 3cell_line_1c της Cell Line 3 (Κυτταρικής Σειράς 3) με παράθυρο κανονικοποίησης στη Lowess 0,05

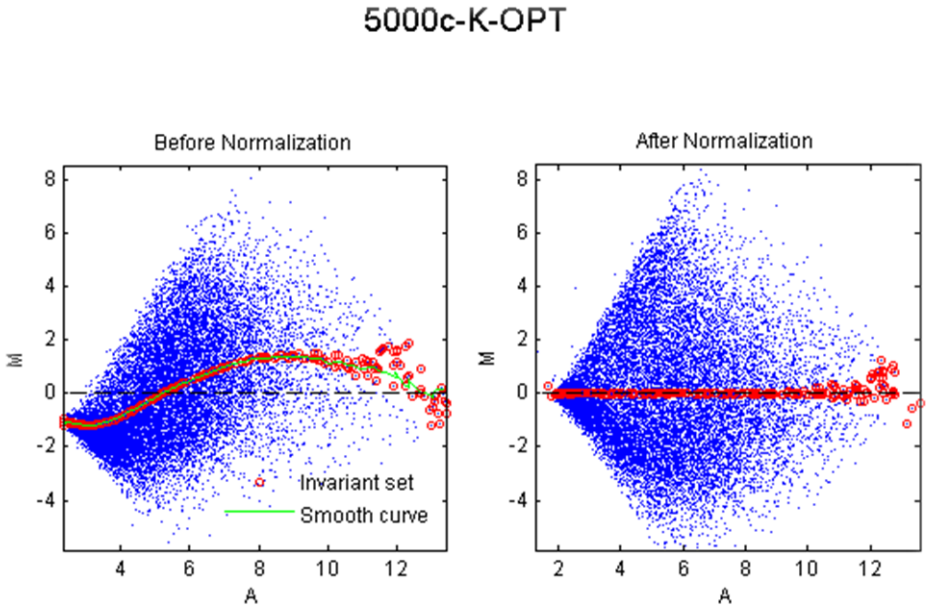
Από τα δύο σχεδιαγράμματα φαίνεται ότι η κανονικοποίηση στο συγκεκριμένο δείγμα με παράθυρο 6% πετυχαίνει καλύτερα από ότι με παράθυρο 5%. Συγκεκριμένα στο δεύτερο σχεδιάγραμμα, στον κατακόρυφο άξονα, φαίνονται κάποιες τιμές να ξεφεύγουν και να υπερβαίνουν την τιμή 3 κατά την κανονικοποίηση τους , ενώ το εύρος τιμών που έχουν κανονικοποιηθεί όλες οι τιμές του δείγματος με παράθυρο 6% κυμαίνεται από [1,5 – 2,2]. Στον οριζόντιο άξονα με παράθυρο 5% η κανονικοποίηση δεν έχει τόσο καλά αποτελέσματα όσο με παράθυρο 6%, αφού στο δεύτερο σχεδιάγραμμα φαίνονται κάποιες τιμές του δείγματος να υπερβαίνουν την τιμή 4. Με το συγκεκριμένο δείγμα φαίνεται καθαρά λοιπόν ότι, με ένα μεγαλύτερο παράθυρο είναι δυνατόν να επιτευχθεί ένα πιο ομαλό smoothing και έτσι η κανονικοποίηση που θα προκύψει να δίνει καλύτερα αποτελέσματα από ότι μία κανονικοποίηση με μικρότερο παράθυρο. Αυτό , όμως δεν σημαίνει σε καμία περίπτωση ότι σε όλα τα δείγματα πρέπει να εφαρμόζονται μεγάλα παράθυρα κατά την κανονικοποίηση τους. Το κάθε δείγμα θα πρέπει να αντιμετωπίζετε διαφορετικά και με έρευνα , μέσω της μέθοδού δοκιμής και σφάλματος, θα πρέπει να αποφασίζεται κατά περίπτωση το βέλτιστο μέγεθος παραθύρου για την κανονικοποίηση. Ωστόσο μια καλή αρχή στην έρευνα παραθύρου είναι η δοκιμή 0,05 μέγεθος παραθύρου, όπου από την βιβλιογραφία αναφέρεται ως το πιο σύνηθες μέγεθος που χρησιμοποιείται από τους ερευνητές, σε περιπτώσεις κανονικοποίησης δειγμάτων από μικροσυστοιχίες, με την μέθοδο Lowess.

Κριτήρια ως προς την επιτυχία ή όχι της Lowess , πέρα από τις τιμές που δίνουν τα δείγματα, αποτελούν τα MvA plots καθώς και οι Correlations Coefficient , που αναλύονται στις επόμενες υπόενότητες. Μετά την κανονικοποίηση των δεδομένων με την μέθοδο Lowess, δημιουργήσαμε μια βάση δεδομένων με όλα τα κανονικοποιημένα και μη κανονικοποιημένα δείγματα προκειμένου να προχωρήσουμε με την μεθοδολογία μας.

4.7 Επιβεβαίωση αποτελεσμάτων μέσω των MvA plots

Παρατηρώντας ένα MvA plot, τα σωστά δεδομένα (συνεπώς αυτά για τα οποία δεν απαιτείται κανονικοποίηση ή η κανονικοποίηση έγινε επιτυχημένα) θα συγκεντρώνονται πολύ κοντά στο 0, αφού δε θα εμφανίζουν διαφορές στο σήμα τους. Η κατανομή γονιδίων από μικρή προς μεγάλη ένταση θα μας δείξει αν υπάρχει συσχέτιση στο control και το δείγμα. Η κανονικοποίηση αποτυγχάνει στην περίπτωση όπου ο λόγος (M) του unamplified control με το δείγμα έχει συσχέτιση με το γινόμενο τους (A).

Στις Κυτταρικές Σειρές που επεξεργαστήκαμε , παρατηρούμε ότι σε όλα τα δείγματα ο λόγος M με το A δεν έχουν κάποια συσχέτιση. Επομένως η κανονικοποίηση έγινε επιτυχημένα. Στην εικόνα που ακολουθεί βλέπουμε ένα τέτοιο παράδειγμα.



Εικόνα 25. Παράδειγμα ma plot Κυτταρικής Σειράς 4

Στην συγκεκριμένη εργασία δημιουργήθηκαν τα αντίστοιχα MvA plots για κάθε ένα από τα δείγματα των κυτταρικών σειρών Cell line 1,2,3,4,5. Όλα τα σχεδιαγράμματα παρουσιάζονται στο Παράρτημα Α – Α.1

4.8 Επιβεβαίωση αποτελεσμάτων μέσω του συντελεστή συσχέτισης (Correlations Coefficient)

Ο συντελεστής συσχέτισης αποτέλεσε έναν αρχικό δείκτη ομοιότητας στην γονιδιακή έκφραση των δειγμάτων, όπου χρησιμοποιήθηκε για την αξιοπιστία των αποτελεσμάτων . Αν τα δεδομένα δεν έχουν κάποια συσχέτιση τότε αυτός ο δείκτης θα τείνει στο 0 και δεν θα έχει σύγκλιση , αν τείνει στα άκρα , δηλαδή στο 1 ή στο -1 τότε τα δεδομένα παρουσιάζουν ισχυρή συσχέτιση. Για να προκύψει ο συντελεστής συσχέτισης βρήκαμε την γραμμική εξάρτηση που έχει το κάθε δείγμα με το αντίστοιχο unamplified control του (bulk). Επί της ουσίας δηλαδή εξετάστηκε τι απόκλιση ή σύγκλιση έχει ένα δείγμα με το αντίστοιχο δείγμα όπου δεν έχει πειραματικό σφάλμα. Ο συντελεστής συσχέτισης βρέθηκε πριν τα δεδομένα κανονικοποιηθούν με την μέθοδο Lowess και αφότου τα δεδομένα κανονικοποιήθηκαν. Οι τιμές που προέκυψαν

συγκρίθηκαν μεταξύ τους για να διερευνηθεί ποια πλησιάζει στα άκρα ή στο μηδέν και κυρίως ποια τιμή τείνει στο 1.

Παρακάτω παρουσιάζεται ο Πίνακας με τις τιμές των Correlation Coefficient για την κυτταρική σειρά 1:

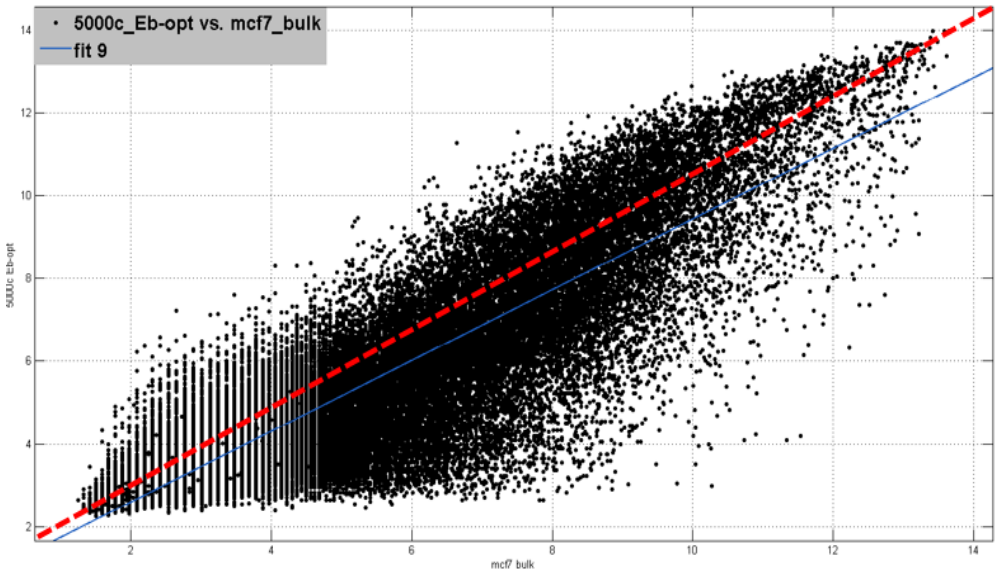
Πίνακας 7. Οι τιμές του συντελεστή συσχέτισης για την Κυτταρική Σειρά 1

Δείγμα	Τιμή Correlation Coefficient πρίν την Lowess	Τιμή Correlation Coefficient μετά την Lowess , με παράθυρο 5%	Τιμή Correlation Coefficient μετά την Lowess , με παράθυρο 6%
1Cell_line1_1c	0.637	0.6442	0.6482
2Cell_line1_1c	0.6734	0.703	0.7032
3Cell_line1_1c	**0.6603**	**0.0706**	**0.679**

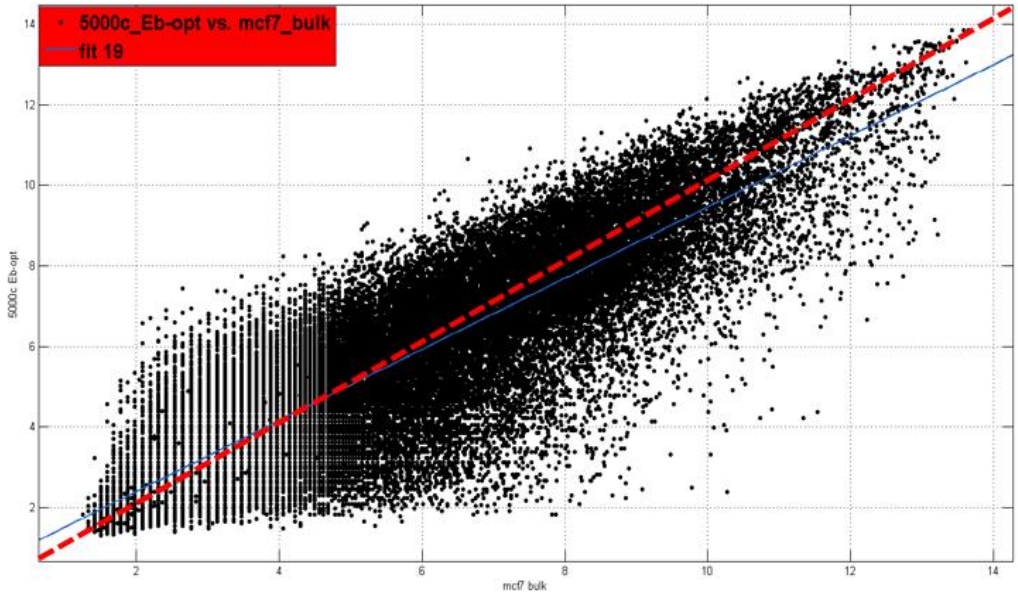
Όλοι οι συντελεστές συσχέτισης βρέθηκαν μέσω του matlab. Στο Παράρτημα Α - Α.2 παρουσιάζονται αναλυτικά όλοι οι πίνακες με τους συντελεστές συσχέτισης για όλα τα δείγματα.

Αξίζει να σημειωθεί ότι , στο δείγμα 3Cell_line1_1c , αφού του εφαρμόστηκε η μέθοδος Lowess ,ο συντελεστής συσχέτισης, του δείγματος με το unamplified control από 0,6603 που ήταν πριν την κανονικοποίηση έγινε 0,0706. Αυτό ουσιαστικά αποτέλεσε μια πρώτη ένδειξη ότι το εν λόγω δείγμα ήθελε ένα άλλο παράθυρο κανονικοποίησης. Πράγματι με παράθυρο 6% , μετά την κανονικοποίηση, η τιμή του συντελεστή συσχέτισης ανεβαίνει και το δείγμα με το unamplified control , δίνει μια καλύτερη γραμμική εξάρτηση. Αυτό πρακτικά σημαίνει ότι το δείγμα διορθώθηκε με βάση το control.

Στα παρακάτω διαγράμματα, φαίνεται η γραμμική συσχέτιση του δείγματος 5000_Eb-opt , της κυτταρικής σειράς 4, με το αντίστοιχο bulk του , που είναι το MCF7 control, πριν την κανονικοποίηση Lowess και στο δεύτερο σχεδιάγραμμα μετά την Lowess:



Εικόνα 26. Plot δείγματος 5000_Eb-opt , με το αντίστοιχο bulk (MCF7) , πριν την Lowess

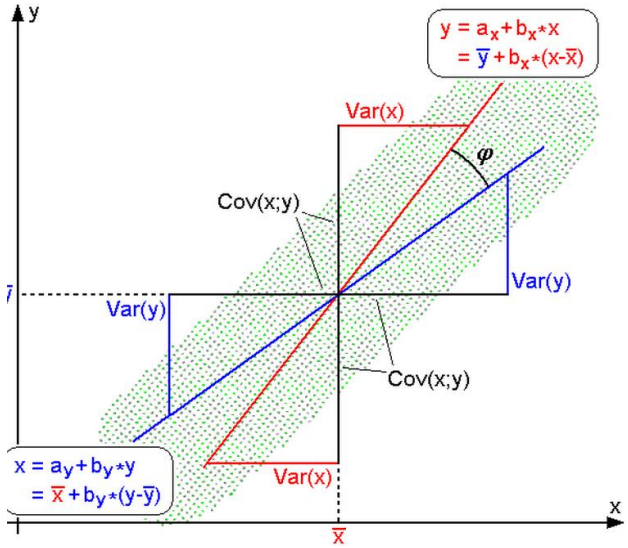


Εικόνα 27. Plot δείγματος 5000_Eb-opt , με το αντίστοιχο bulk (MCF7) , μετά την Lowess

Στο Παράρτημα Α - Α.3 παρουσιάζονται όλα τα σχεδιαγράμματα της Κυτταρικής σειράς 4,πριν και μετά την κανονικοποίηση με την μέθοδο της Lowess.

Στο συγκεκριμένο δείγμα ο δείκτης συσχέτισης πριν την κανονικοποίηση είναι 0.916 και μετά την κανονικοποίηση Lowess ,παίρνει την τιμή 0.9105. Το δείγμα αυτό

και πρίν και μετά την κανονικοποίηση του έχει πολύ υψηλό δείκτη συσχέτισης. Από τις παραπάνω γραφικές διακρίνεται ότι η γραμμική συσχέτιση του δείγματος με το αντίστοιχο unamplified control είναι πολύ κοντά στην διαγώνιο του γραφήματος. Αυτό πρακτικά σημαίνει ότι όσο πιο υψηλό συντελεστή συσχέτισης έχει ένα δείγμα τόσο η γραμμική εξάρτησή του από το bulk πλησιάζει την διαγώνιο του γραφήματος. Το συμπέρασμα αυτό αποδεικνύεται και από το παρακάτω σχεδιάγραμμα, δεδομένου ότι το X μετατρέπεται σε a + bX και αντίστοιχα το Y σε c + dY, όπου a, b, c, and d είναι σταθερές με b, d > 0 και δεδομένου ότι η εξίσωση aX+b, περνάει από την αρχή των αξόνων και αποτελεί την διαγώνιο.



Εικόνα 28. Regression lines για $y=g_x(x)$ [κόκκινο] και $x=g_y(y)$ [μπλέ] [E.11]

Ο συντελεστής συσχέτισης είναι: $r = \sec\varphi - \tan\varphi$ [83]

4.9 Εφαρμογή της μεθόδου Global Normalization

Η συνοχή των αποτελεσμάτων ανάμεσα στις επαναλήψεις του ίδιου πειράματος σε διαφορετικά ή ακόμα και στο ίδιο εργαστήριο αποτελεί ένα σημαντικό κριτήριο για την αξιολόγηση των αποτελεσμάτων. Η ολική κανονικοποίηση εξαλείφει πειραματικές διαφορές που οφείλονται σε παράγοντες όπως inter experimental και inter /intra laboratory variability.

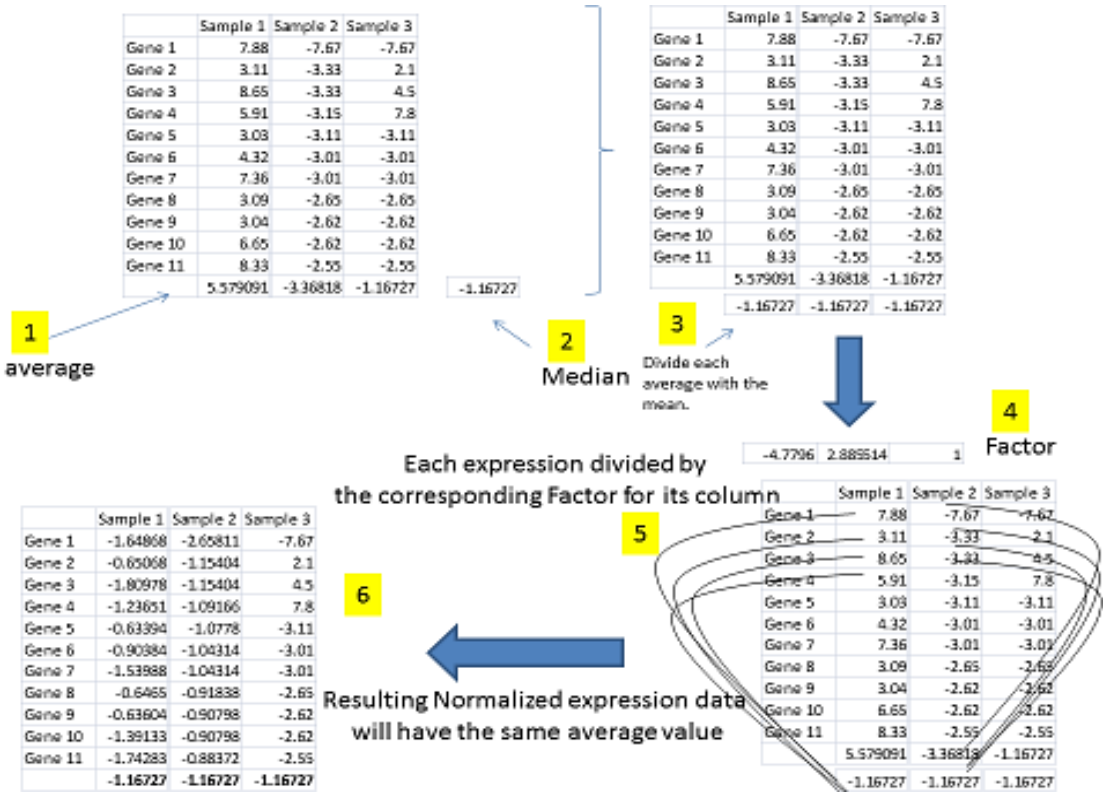
Η εφαρμογή μιας καθολικής (global) κανονικοποίησης στο σύνολο των δεδομένων γίνεται με βάση την υπόθεση ότι τα περισσότερα γονίδια εκφράστηκαν ισοδύναμα στα Cy-3 και Cy-5 κανάλια και ότι το συνολικό mRNA παρέμεινε σταθερό. Γι' αυτό θα πρέπει για κάθε δείγμα, το άθροισμα των τιμών των εντάσεων να έχει κάποια ίδια στατιστικά χαρακτηριστικά. Οποιεσδήποτε διακυμάνσεις που προκλήθηκαν από γονίδια που δεν εκφράστηκαν ισοδύναμα θα πρέπει εξομαλυνθούν. Προς αυτήν την

κατεύθυνση υπολογίζεται ένας σταθερός συντελεστής (factor) και πολλαπλασιάζεται / διαρείται με όλες τις τιμές των εντάσεων των δειγμάτων για την αναπροσαρμογή (rescaling) των τιμών τους.

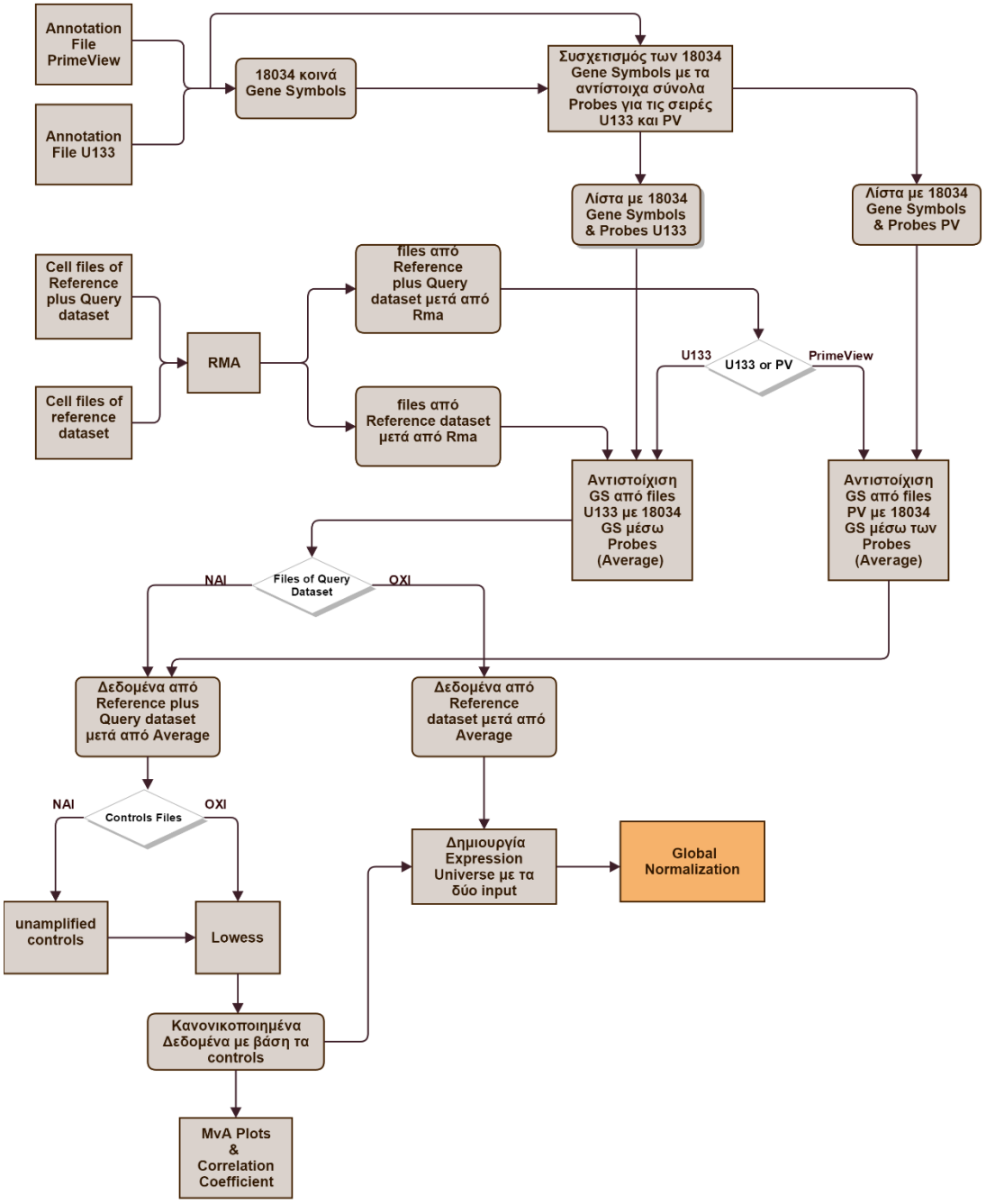
Γενικότερα όμως, η κανονικοποίηση δεδομένων μικροσυστοιχιών είναι μια πολύπλοκη και πολυπαραγοντική διαδικασία και πιθανότατα δεν πρόκειται ποτέ να τυποποιηθεί. Ακόμα και σήμερα, η "κοινή λογική" και η βαθιά γνώση και εμπειρία πάνω στα βιολογικά δεδομένα και τις πειραματικές διαδικασίες λειτουργούν ως ανεκτίμητης αξίας οδηγοί στη διαδικασία της κανονικοποίησης

Στη συγκεκριμένη διπλωματική εργασία ακολουθήθηκε μία προσέγγιση ολικής κανονικοποίησης, βασισμένη στα παρακάτω βήματα:

1. Υπολογισμός του μέσου όρου για κάθε δείγμα.
2. Εύρεση του αριθμητικού μέσου ανάμεσα στους μέσους όρους που υπολογίστηκαν στο βήμα 1.
3. Διαίρεση καθενός μέσου όρου που υπολογίστηκε στο βήμα 1 με τον (κοινό) αριθμητικό μέσο που υπολογίστηκε στο βήμα 2.
4. Τον καινούργιο όρο που προέκυψε στο βήμα 3 τον ονομάζουμε Factor.
5. Διαίρεση όλων των τιμών των δειγμάτων με τον αντίστοιχο Factor που υπολογίστηκε στο βήμα 4. Για τις καινούργιες normalized expression values θα προκύπτει ό ίδιος μέσος όρος σε όλα τα δείγματα.



Εικόνα 29. Διαδικασία Global Normalization



Εικόνα 30. Το στάδιο της Global Normalization στην προτεινόμενη μεθοδολογία

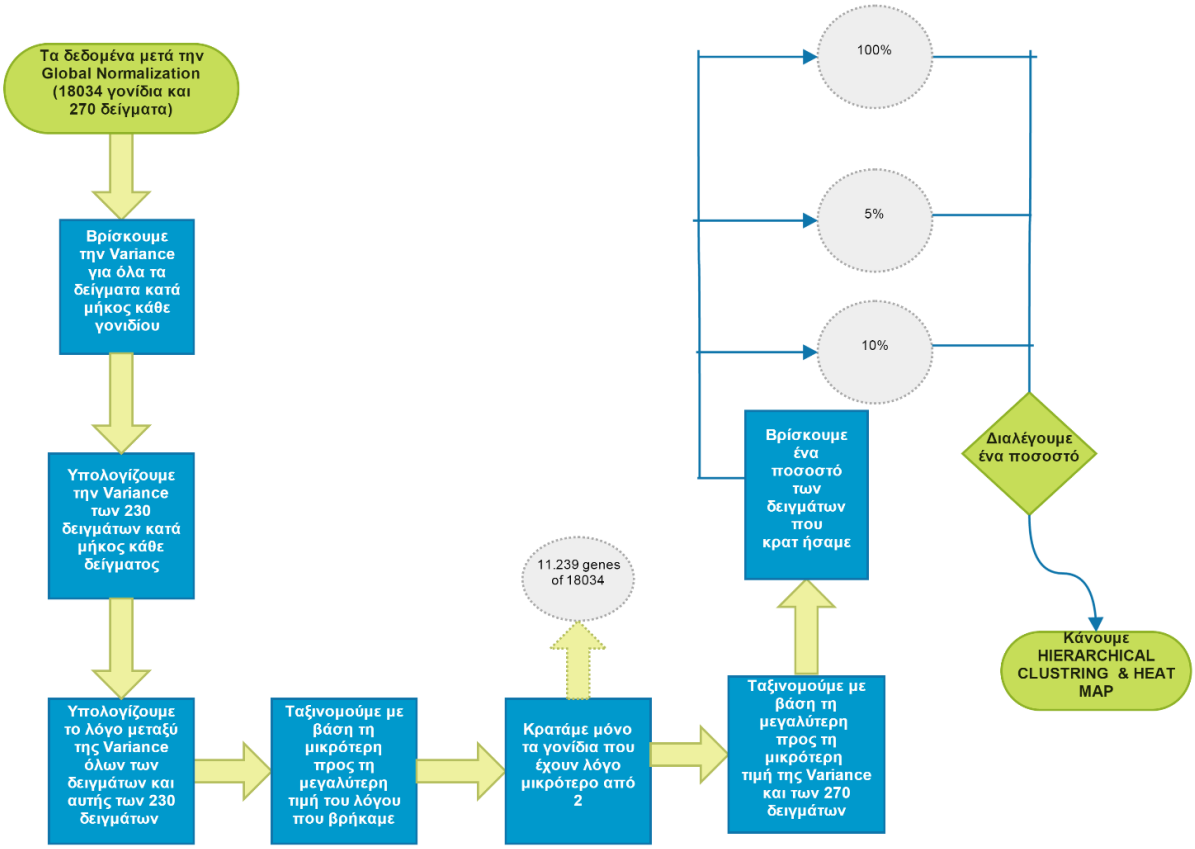
1.10 Φιλτράρισμα γονίδιών για την κατηγοριοποίηση δειγμάτων με χρήση διακύμανσης (Variance)

Η Variance (διακύμανση) , μπορεί να υποδείξει την διαφοροποίηση , την μεταβολή της έκφρασης ενός γονιδίου , για διαφορετικά δείγματα. Η ανάγκη ελέγχου των διακυμάνσεων των γονιδίων κατά μήκος των δειγμάτων είναι πολύ μεγάλη , διότι αν κάποιο γονίδιο έχει πολύ μεγάλες αποκλίσεις , αυτό σημαίνει ότι μπορεί να μην έχει κανονικοποιηθεί επιτυχημένα . Κάτι τέτοιο υποδηλώνει ότι οι διάφοροι τύποι θορύβου που αρχικά περιείχε, δεν γνωρίζουμε αν τελικά εξαλείφθηκαν. Έτσι το συγκεκριμένο γονίδιο μπορεί τελικά να μην μας δίνει χρήσιμη βιολογική πληροφορία και για αυτό το λόγο θα πρέπει να απομονωθεί.

Για αυτό το λόγο, εφαρμόσαμε μία μέθοδο προκειμένου να απομονώσουμε τέτοια γονίδια. Αρχικά βρήκαμε την expression variance κάθε γονιδίου στο reference dataset και στο Reference plus Query dataset. Με αυτόν τον τρόπο εντοπίσαμε κάποια γονίδια τα οποία , μελετώντας το προφίλ τους παρατηρήσαμε ότι είχαν πολύ μεγάλες αποκλίσεις. Εφαρμόσαμε μια μέθοδο προκειμένου να μειώσουμε τον αριθμό των δειγμάτων που θα κατηγοριοποιηθούν και να περιοριστούμε στα γονίδια που έχουν χρήσιμη βιολογική πληροφορία και κατ' επέκταση υψηλή βιολογική variance. Παρατηρήσαμε ότι, παρά τις διαδικασίες κανονικοποίησης που εφαρμόσαμε, κάποια γονίδια χαρακτηρίζονταν ακόμα από experimentally-related variance (πειραματική σχετική διακύμανση) συγκρίνοντάς τα με εκείνα του reference dataset. Η μέθοδος που εφαρμόσαμε περιγράφεται από τα εξής στάδια:

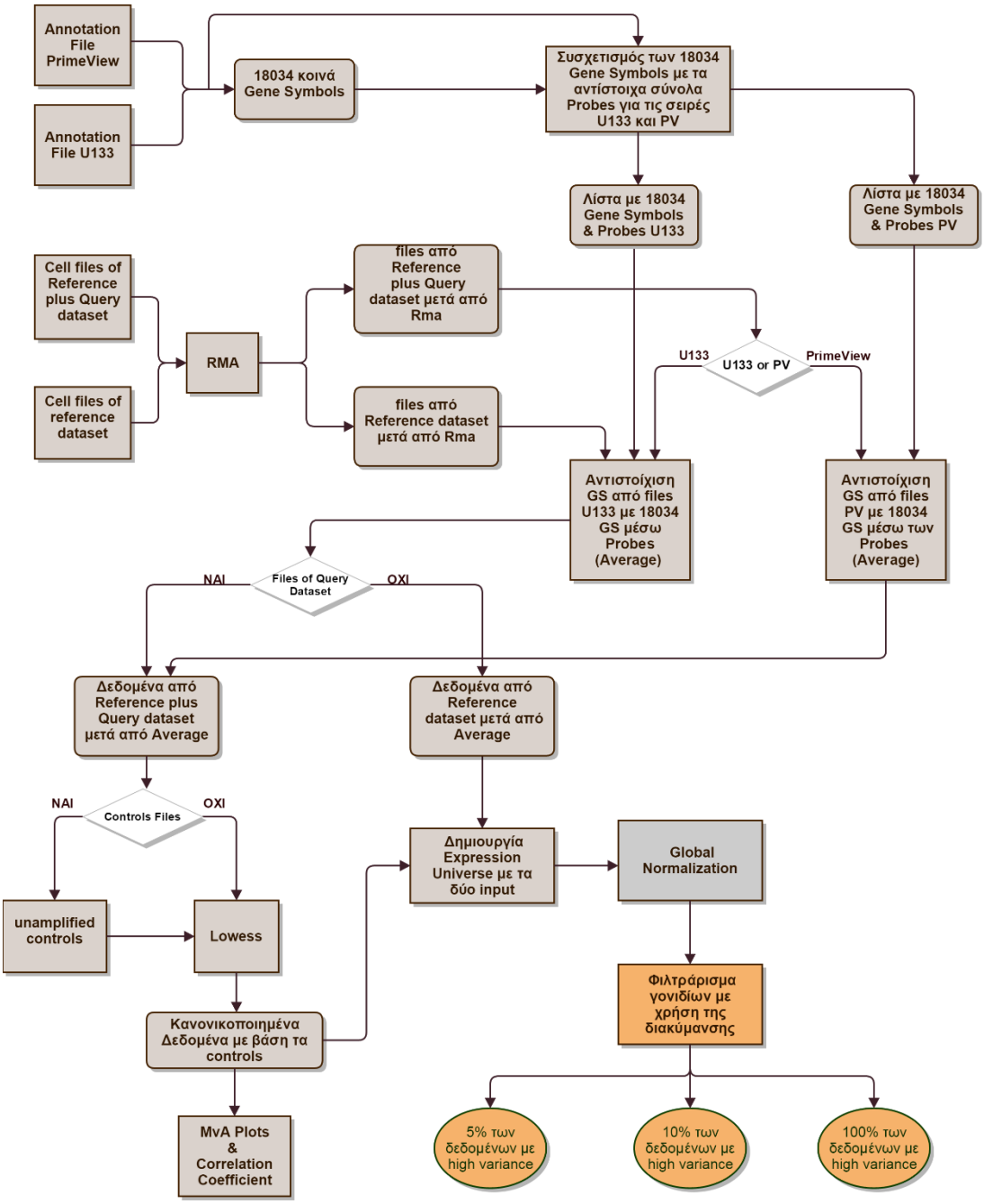
- Βρίσκουμε την variance κατά μήκος ενός γονιδίου, για όλα τα δείγματα (reference dataset & Reference plus Query dataset).
- Υπολογίζουμε τη variance μόνο του reference dataset.
- Υπολογίζουμε τον λόγο της Variance όλων των δειγμάτων και της Variance των δειγμάτων του reference dataset.
- Κρατάμε μόνο τα γονίδια που ο λόγος της παραπάνω πράξης προκύπτει μικρότερος του 2 και μεγαλύτερος του 0,5. Αυτό συμβαίνει επειδή ένας λόγος υψηλότερος του 2 ή χαμηλότερος του 0,5, συνεπάγεται σημαντικές διαφορές στην διακύμανση, αντανακλώντας κυρίως μη βιολογικά αίτια. Δηλαδή η μεταβολή του σήματος έκφρασης οφείλεται σε πειραματικές διαδικασίες .
- Τέλος, από τα γονίδια που κρατήσαμε, τα ταξινομούμε κατά αύξουσα τιμής της διακύμανσής τους στο Reference plus Query dataset .

Παρακάτω, στην εικόνα που ακολουθεί φαίνονται σχεδιαγραμματικά τα βήματα που ακολουθήσαμε για αυτήν την μέθοδο για το Expression Universe 1. Να σημειωθεί ότι δεν είχαμε τιμές για τους λόγους των variances κάτω από 0,5 και ο αριθμός των γονίδιων τελικά μειώθηκε κατά 6795. Κάποια βήματα τις παρακάτω μεθόδου αναλύονται μετέπειτα.



Εικόνα 31. Η μέθοδος με τις Variances για το expression universe 1

Στην επόμενη εικόνα , φαίνεται σε ποιο στάδιο της μεθοδολογίας μας βρισκόμαστε. Βλέπουμε ότι παίρνουμε το αρχείο που παρήγαγε η Global Normalization, όπως το έχουμε περιγράψει παραπάνω και ξανά κανονικοποιούμε τα δεδομένα μας με την χρήση της διακύμανσης. Τέλος δημιουργούμε αρχεία για όλα τα δείγματα που χρησιμοποιήσαμε αλλά κρατάμε κάποιο ποσοστό των γονιδίων με υψηλή διακύμανση της έκφρασής τους , σύμφωνα με την μέθοδο που περιγράψαμε παραπάνω. Τα αρχεία αυτά στο επόμενο στάδιο της μεθοδολογίας που έχουμε ακολουθήσει , θα περάσουν από την διαδικασία του Clustering. Πιο αναλυτικά τα αρχεία αυτά , που αποτελούν την βάση δεδομένων μας όπου μέσα έχουμε αποθηκευμένη πληροφορία για τα γονίδια (γραμμές) και τα δείγματα (στήλες), θα κατηγοριοποιηθούν με βάση τα δείγματα. Αυτός είναι και ο τελικός στόχος αυτής της εργασίας.



Εικόνα 32. Η κανονικοποίηση με χρήση της Διακύμανσης και το φιλτράρισμα των γονιδίων

4.11 Εφαρμογή μεθόδων κατηγοριοποίησης και οπτικοποίησης αποτελεσμάτων

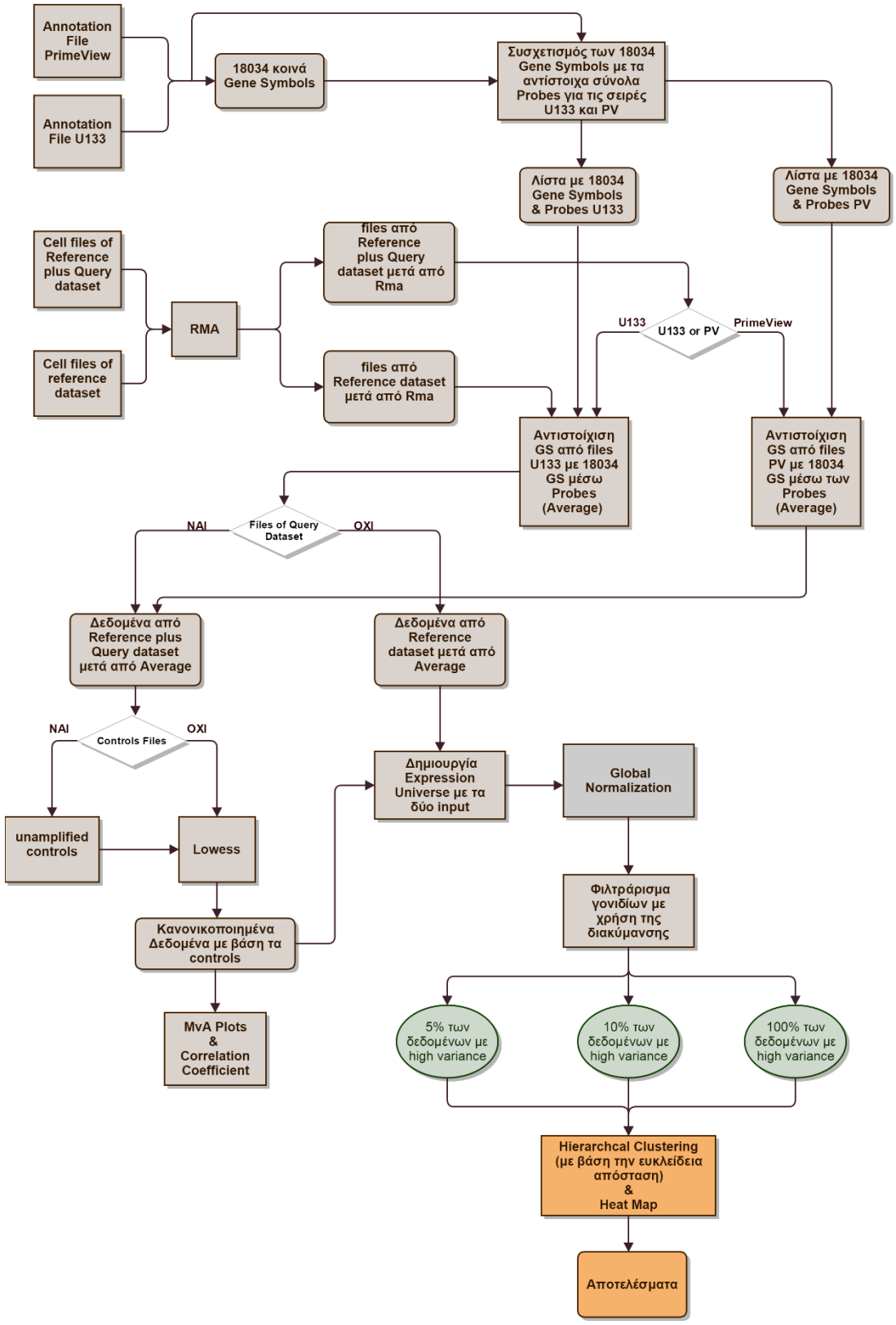
Το επόμενο στάδιο της παρούσας εργασίας ήταν η ομαδοποίηση (clustering) των δεδομένων. Αυτό έγινε με σκοπό να ομαδοποιήσουμε το Reference Dataset και κατ΄επέκταση να κατηγοριοποιήσουμε το Reference Plus Query Dataset. Όπως έχει ήδη αναφερθεί δεν υπήρχε κάποια αρχική γνώση για τον τρόπο με τον οποίο θα κατηγοριοποιηθούν τα δεδομένα (unsupervised). Με βάση αυτό για να επιτευχθεί η απαιτούμενη ομαδοποίηση και για να γίνει επιτυχημένα κάναμε μία σειρά από δοκιμές όπου περιγράφονται παρακάτω.

Αρχικά σαν είσοδο είχαμε μία βάση η οποία περιείχε όλα τα δείγματα του Reference και του Query Dataset , τα οποία είχανε κανονικοποιηθεί κατάλληλα με τις μεθόδους που περιγράψαμε παραπάνω. Όλα τα δείγματα είχαν κοινά Gene Symbols (18034). Αυτό το αρχείο προσπαθήσαμε να το κατηγοριοποιήσουμε και να το ομαδοποιήσουμε με βάση το βαθμό ομοιότητας των δειγμάτων. Αφού λοιπόν στόχος ήταν να κατηγοριοποιήσουμε και να ταυτοποιήσουμε τα δείγματα μας ,το clustering υλοποιήθηκε ως προς τις στήλες. Όπως έχει αναφερθεί τα microarray data , αναπαρίστανται σε έναν δισδιάστατο πίνακα , όπου οι γραμμές είναι τα γονίδια και οι στήλες τα samples / μετρήσεις / πειράματα.

Το clustering το εφαρμόσαμε με βάση 2 παραμέτρους. Η πρώτη παράμετρος είναι η απόσταση ενώ η δεύτερη ο αλγόριθμος. Η απόσταση που επιλέξαμε είναι η Ευκλείδεια απόσταση. Ενώ ο αλγόριθμος είναι το Hierarchical clustering. Πέρα από όλες αυτές της μεθόδους όπως έχει αναφερθεί και στην μελέτη της βιβλιογραφίας , υπάρχει μία πληθώρα άλλων τεχνικών που μπορούν να μελετηθούν. Ωστόσο η εταιρία της Affymetrix, από όπου προέρχονται και τα Chips των microarrays που χρησιμοποιήσαμε , συστήνει να χρησιμοποιούνται αυτές οι μέθοδοι σε περιπτώσεις όπως η δική μας.

Με το Hierarchical clustering που υλοποιήσαμε οι ομοιότητες των δεδομένων αναπαρίστανται με δενδροδιαγράμματα. Στην αρχή όλα τα δεδομένα θεωρούνται ως ξεχωριστά clusters. Σε κάθε βήμα δεδομένα με την ελάχιστη απόσταση μεταξύ τους ομαδοποιούνται σε clusters. Από το δεύτερο βήμα και μετά ομαδοποιούνται οι κόμβοι του δένδρου, έως ότου να φτάσουμε σε έναν μοναδικό cluster και θα βρίσκεται στο μέγιστο ύψος του δένδρου.

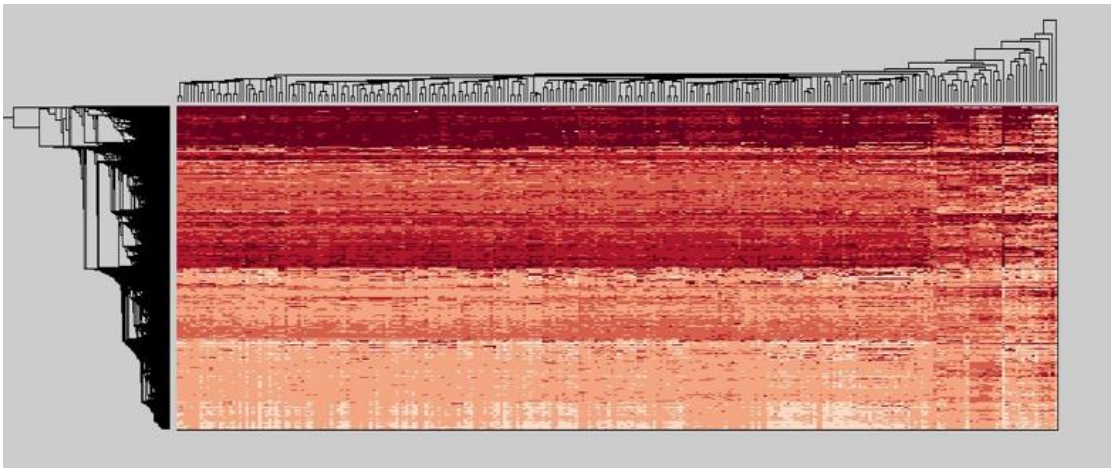
Το τελευταίο στάδιο της εργασίας είναι το Heat Map. Με αυτόν τον τρόπο καταφέραμε να γίνουν εμφανείς οι εξαρτήσεις μεταξύ των δειγμάτων. Συγκεκριμένα , τα δείγματα που βρίσκονται σε διαδοχικές στήλες ομαδοποιήθηκαν και κατηγοριοποιήθηκαν μεταξύ τους με κριτήριο αν ανήκαν στον ίδιο ιστό ή επρόκειτο για πανομοιότυπη Κυτταρική Σειρά. Επίσης δείγματα τα οποία για κάποιο λόγο δεν ταιριάζουν με τα υπόλοιπα απομονώνονται εμφανώς (έλεγχος της πειραματικής διαδικασίας). Αντίστοιχα, φαίνονται τα γονίδια που ομαδοποιήθηκαν.



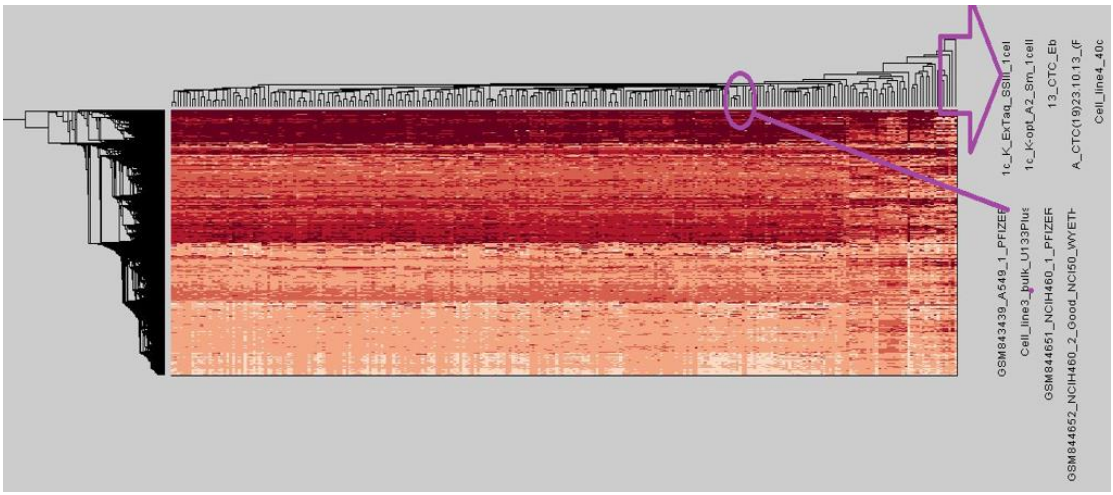
Εικόνα 33. Το τελευταίο στάδιο της μεθοδολογίας μας, Hierarchical clustering , Heat maps

Τα Heat maps, που δημιουργήσαμε δεν είναι εύκολο να απεικονιστούν , κάνοντας αντιληπτές όλες τις εξαρτήσεις των δειγμάτων ή των γονίδιων , χωρίς να έχουμε ανοιχτό το πρόγραμμα από τα οποίο τα δημιουργήσαμε. Λόγω του μεγάλου όγκου δεδομένων (18034 γονίδια και μέχρι 2100 δείγματα) , δεν είναι δυνατόν να αναπαρασταθεί η κάθε εξάρτηση. Ωστόσο παρακάτω θα παρουσιάσουμε κάποιες εικόνες από τα Heat maps που δημιουργήσαμε καθώς και τα σημεία όπου φαίνονται οι επιτυχημένες κατηγοριοποιήσεις που επιτεύχθηκαν.

Στις επόμενες εικόνες απεικονίζεται , μέσω Heat map , το Clustering του Expression Universe 1, με 18034 γονίδια.



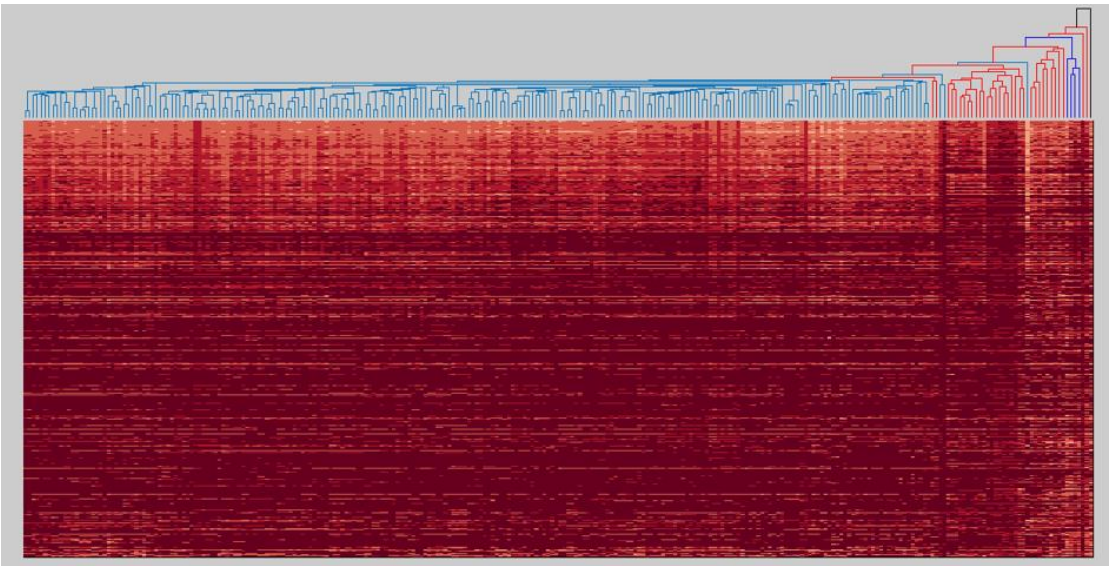
Εικόνα 34. Heat map για το Expression Universe 1 με 18034 γονίδια



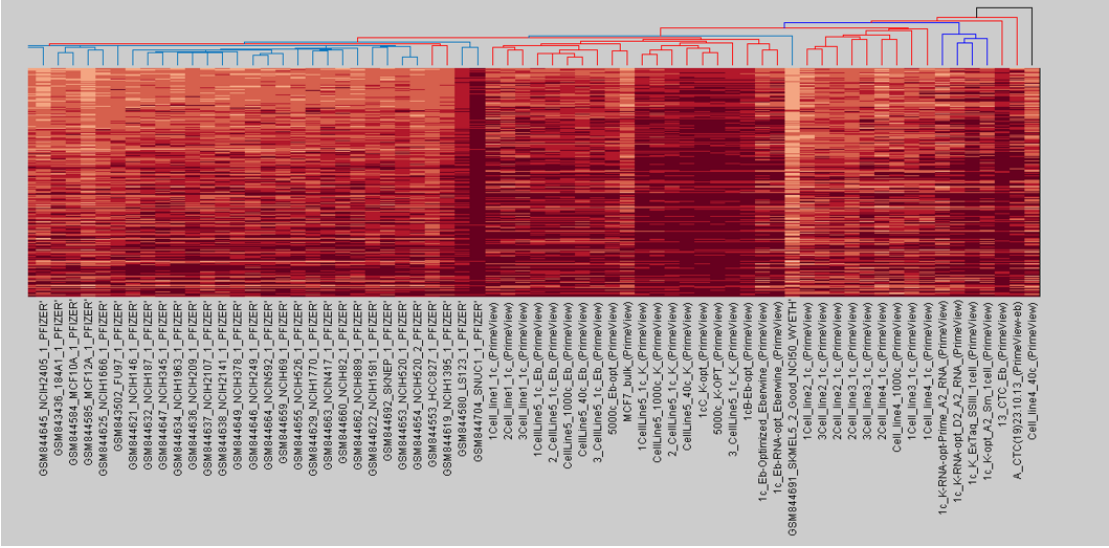
Εικόνα 35. Στιγμιότυπο του Heat map για το Expression Universe 1 με 18034 γονίδια

Παρατηρήσαμε ότι δεν προκύπτει κάποιο αξιόλογο αποτέλεσμα κατηγοριοποίησης, χρησιμοποιώντας όλο το σύνολο των γονιδίων του dataset, πέραν της σωστής ομαδοποίησης των δειγμάτων του Reference Dataset. Όπως αναφέρθηκε

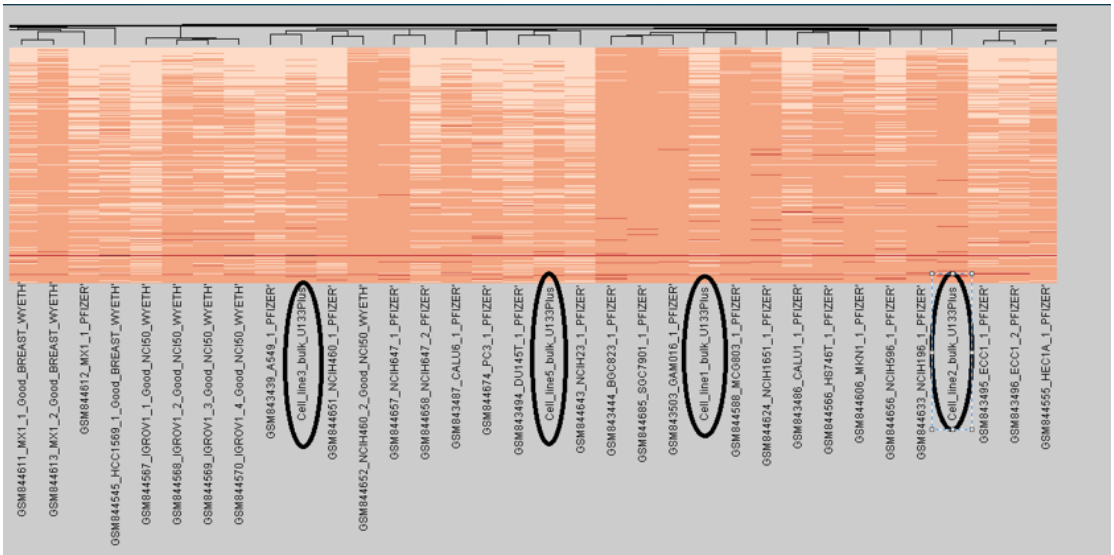
στην ενότητα 4.10 ορισμένες τιμές των Gene Symbols παρουσιάζουν experimentally related variance και για να τα απομονώσουμε εφαρμόζουμε τον αλγόριθμο που αναφέρετε στην ίδια υποενότητα. Μετά την εφαρμογή του αλγορίθμου περίπου 7000 Gene Symbols δεν χρησιμοποιήθηκαν και ένα καινούριο heat map προέκυψε για περίπου 11.000 γονίδια .



Εικόνα 36. Heat map για το Expression Universe 1 με 11000 γονίδια



Εικόνα 37. Στιγμιότυπο του Heat map για το Expression Universe 1 με 11000 γονίδια



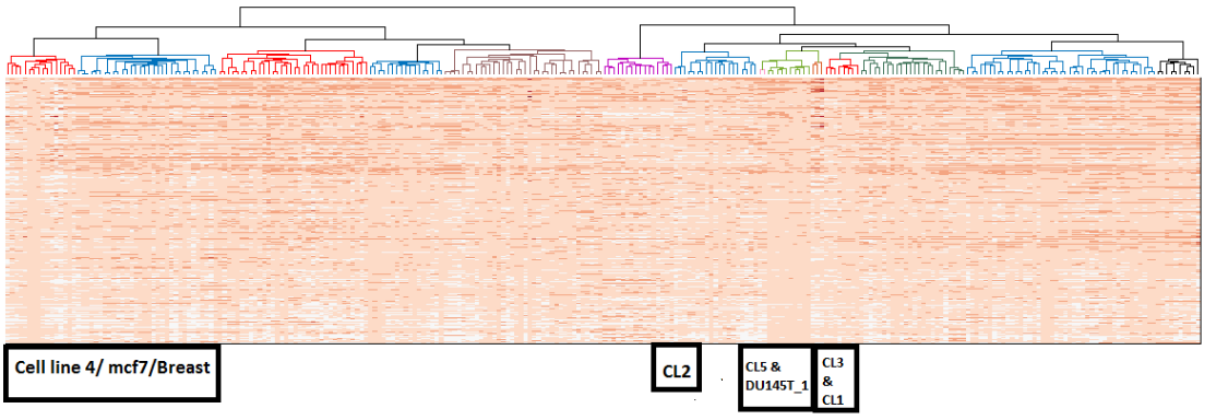
Εικόνα 38. Στιγμιότυπο του Heat map για το Expression Universe 1 με 11000 γονίδια

Στην εικόνα 38 βλέπουμε την πρώτη επιθυμητή κατηγοριοποίηση. Το δείγμα Cell line 5 bulk , κατηγοριοποιείται με το δείγμα DU145T_1_PFIZER. Ωστόσο, στην εικόνα 37 ,παρατηρήσαμε ότι δείγματα της Κυτταρικής Σειράς 4 δεν κατηγοριοποιούνται καθόλου , με δείγματα μαστού . Σε αυτό το σημείο θα πρέπει να τονιστεί , όπως έχει αναφερθεί , ότι δεν γνωρίζουμε για κανένα δείγμα του Query Dataset , τι δείγμα είναι εκτός από τα δείγματα της Κυτταρικής Σειράς 4.

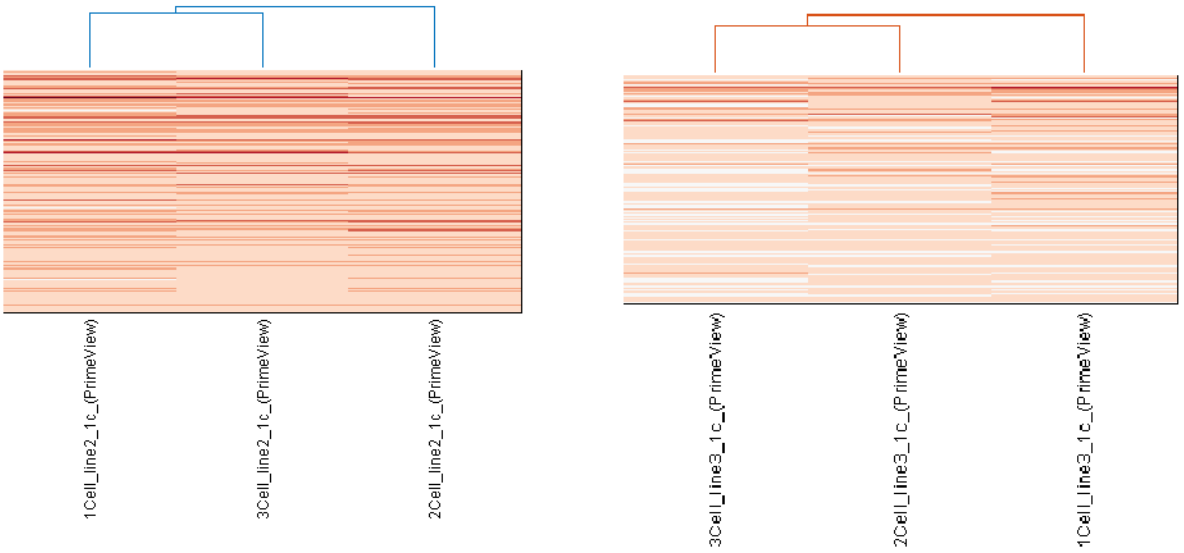
Η κατηγοριοποίηση αυτή δεν μας έδωσε επομένως κάποιο ολοκληρωμένο αποτέλεσμα και έτσι δοκιμάσαμε πειραματικά να πάρουμε κάποια ποσοστά αυτών των γονιδίων. Δοκιμάσαμε να υλοποιήσουμε το heat map με το 5% και το 10% των γονιδίων που είχαν high variance καθώς και με τα 50 και 100 γονίδια που είχαν την υψηλότερη διακύμανση, σύμφωνα πάντα με την μέθοδο που περιγράψαμε στην ενότητα 4.10. Τα αποτελέσματα που προκύπτουν παρουσιάζονται στο Παράρτημα Β.

Στα στιγμιότυπα της κατηγοριοποίησης των 50 ,100 και 900 περίπου γονιδίων , παρατηρούμε ότι και πάλι δεν είχαμε κάποιο επιθυμητό αποτέλεσμα. Οι Κυτταρικές Σειρές , του Query Dataset κατηγοριοποιούνται μεταξύ τους, χωρίς να ταυτοποιούνται με άλλες σειρές από το Reference Dataset , πλην του δείγματος της Cell line 5 bulk , , όπου κατηγοριοποιείται με το σωστό δείγμα. Έτσι λοιπόν υλοποιήσαμε μία σειρά από πειραματικές διαδικασίες και δοκιμές προκειμένου να καταφέρουμε να πάρουμε το καλύτερο δυνατό αποτέλεσμα. Δοκιμάσαμε να ταυτοποιήσουμε τις Κυτταρικές Σειρές κάνοντας χρήση κάποιου μεγαλύτερου ποσοστού γονιδίων. Στο Παράρτημα Β φαίνονται τα Heat maps που υλοποιήσαμε.

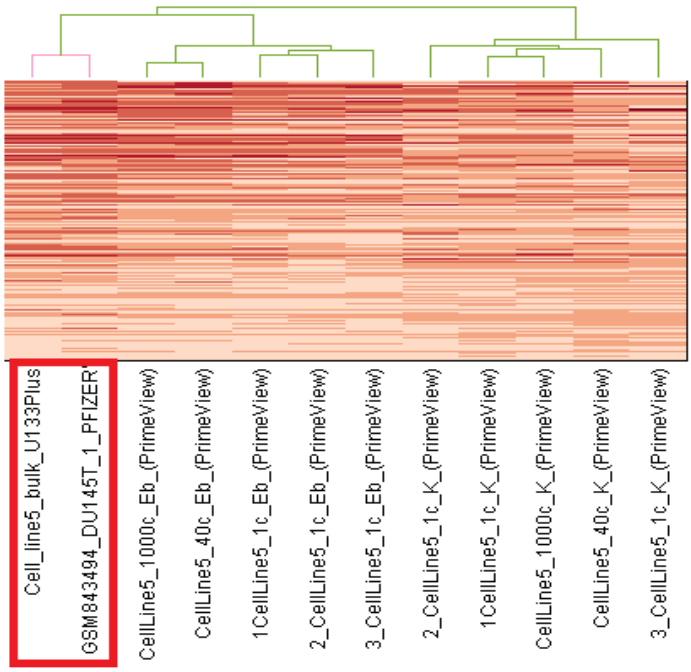
Στις επόμενες εικόνες φαίνονται τα στιγμιότυπα του Heat map για 1400 γονίδια , τα οποία επιλέχτηκαν με βάση του κλάσματος μεταξύ της διακύμανσης για όλα τα δείγματα και αυτής μόνο του reference (τιμή από 1,9 -0,9. Στην εικόνα 39 φαίνονται στα μαύρα κουτάκια σε σχέση με όλο το Heat map , που βρίσκονται οι κυτταρικές σειρές.



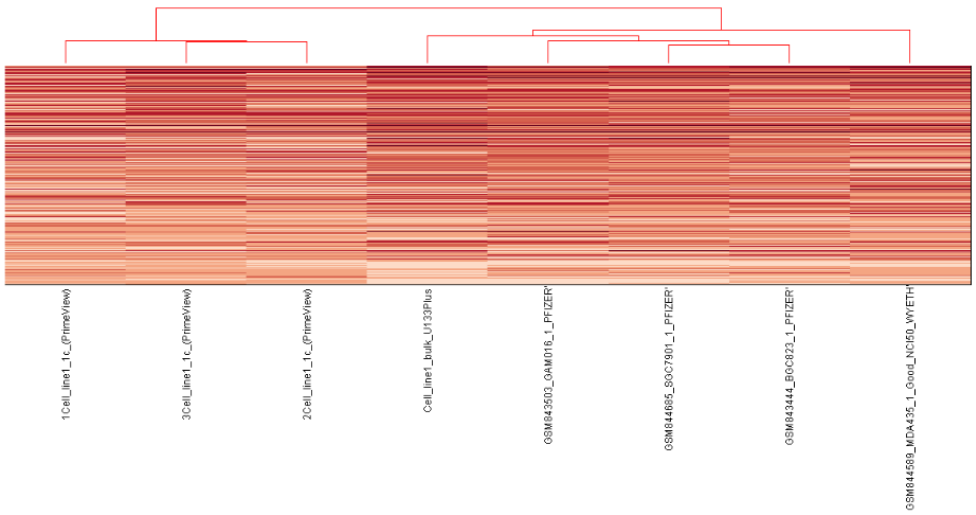
Εικόνα 39. Heat map για το Expression Universe 1 με 1400 γονίδια , με highest Variance



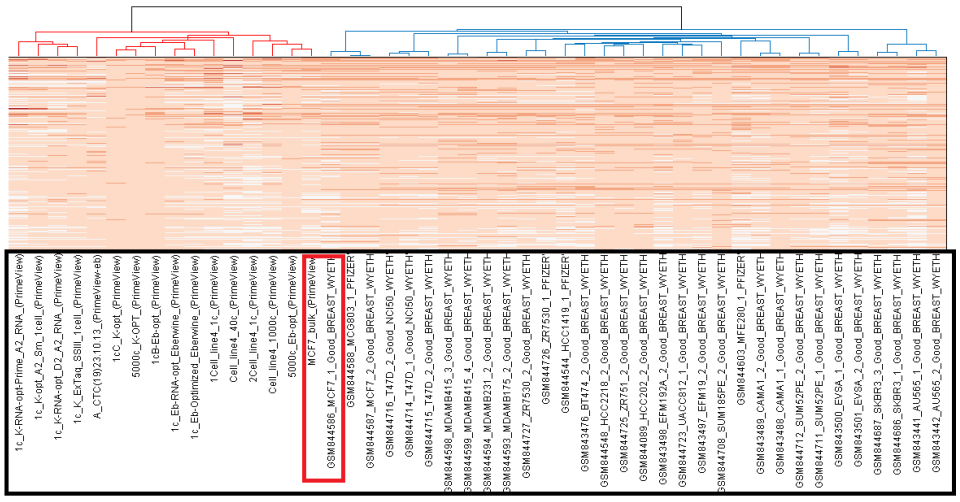
Εικόνα 40. Στιγμιότυπα του Heat map για το Expression Universe με 1400 γονίδια , με highest Variance. Κατηγοριοποίηση Cell Line 2,3



Εικόνα 41. Στιγμιότυπο του Heat map για το Expression Universe με 1400 γονίδια , με highest Variance. Απεικόνιση κατηγοριοποίησης του δείγματος Cell line 5 bulk



Εικόνα 42. Στιγμιότυπο του Heat map για το Expression Universe με 1400 γονίδια , με highest Variance. Κατηγοριοποίηση Cell Line 1



Εικόνα 43. Στιγμιότυπο του Heat map για το Expression Universe με 1400 γονίδια , με highest Variance. Κατηγοριοποίηση Cell Line 4 με δείγματα μαστού

Το αποτέλεσμα αυτής της έρευνας , το πήραμε με την κατηγοριοποίηση των 1400 γονιδίων. Στα στιγμιότυπα που φαίνονται παραπάνω δείχνουμε ότι με βάση το Hierachical clustering , που υλοποιήσαμε για το Expression Universe 1 και όλες τις μεθόδους που εκτελέσαμε , καταφέραμε:

- ✓ Να κατηγοριοποιήσουμε την Κυτταρική Σειρά Cell line 4 με δείγματα μαστού.
- ✓ Κατηγοριοποιήσαμε το δείγμα Cell line 5 bulk , της Κυτταρικής Σειράς 5, με το αντίστοιχο δείγμα.
- ✓ Ομαδοποιήθηκαν οι όμοιες Κυτταρικές Σειρές και Ιστοί του Reference Dataset.
- ✓ Οι Κυτταρικές Σειρές του Query Dataset , που δεν κατηγοριοποιήθηκαν , δεν υπήρχαν εν τέλει στα Reference Datasets που χρησιμοποιήσαμε , ωστόσο κατηγοριοποιήθηκαν μεταξύ τους. Κάτι που υποστηρίζει ότι η κατηγοριοποίηση των δεδομένων έγινε επιτυχημένα.

Για την εξαγωγή αυτών των αποτελεσμάτων χρειάστηκαν να γίνουν μία σειρά από πειραματικές δοκιμές. Τα πιο ενδεικτικά Heat maps που υλοποιήσαμε με τις διάφορες τεχνικές που τους εφαρμόσαμε , φαίνονται στο Παράρτημα Β. Ωστόσο για την εξέταση τέτοιου είδους αποτελεσμάτων , δεν ενδείκνυνται από τα στιγμιότυπα, αλλά η εξέταση και ο έλεγχος μέσα από το ίδιο το πρόγραμμα που τα παράγει. Παρ'όλα αυτά τα στιγμιότυπα αυτά αποτελούν μια ένδειξη της δουλείας μας και των έγκαιρων αποτελεσμάτων μας.

Κεφάλαιο 5 Υλοποίηση αλγορίθμων

5.1 Πλατφόρμες λογισμικού επεξεργασίας βιολογικών δεδομένων

Η ανάγκη της χρήσης υπολογιστών στη Μοριακή Βιολογία δημιουργήθηκε από τη δεκαετία του 1950, αμέσως μετά την ανακάλυψη της πρώτης ακολουθίας βιομορίου, της ινσουλίνης, από τον Frederick Sanger. Έκτοτε, η αναγκαιότητα αυτή αυξάνεται διαρκώς. Σήμερα υπάρχουν στην αγορά πληθώρα πακέτων λογισμικού για τη στατιστική επεξεργασία βιολογικών δεδομένων, τα οποία προέρχονται τόσο από ιδιωτικούς, όσο και από δημόσιους, κυρίως ακαδημαϊκούς, φορείς.

Δύο από τις πιο διαδεδομένες πλατφόρμες λογισμικού Βιοπληροφορικής, ιδίως στις κοινότητες των ακαδημαϊκών και των μηχανικών, είναι η Matlab της εταιρείας Mathworks και η open source γλώσσα προγραμματισμού R, διαθέσιμη δωρεάν υπό την GNU General Public License. Η μεν Matlab πλαισιώνεται και από δεκάδες άλλα εργαλεία, πέραν της Βιοπληροφορικής και της στατιστικής, ενώ η R είναι εξειδικευμένη σε αυτές τις δύο κατηγορίες.

Εξαιρώντας το γεγονός ότι η R διατίθεται δωρεάν, αντίθετα με την Matlab, συγκρίνοντας τις δύο αυτές γλώσσες ως προς την απόδοση, παρατηρούμε παρόμοια χαρακτηριστικά. Επίσης, εξετάζοντας την ύπαρξη βιβλιοθηκών με έτοιμες συναρτήσεις αλγορίθμων Βιοπληροφορικής και γραφικών αναπαραστάσεων αυτών, πάλι θα βρεθούν περισσότερες ομοιότητες, παρά διαφορές. Η σύγκριση αποκτά νόημα όταν εξετάζεται η ευκολία υλοποίησης, η διαθέσιμη βιβλιογραφία, η τεχνική υποστήριξη και η πλαισίωση με εργαλεία επεξεργασίας δεδομένων πέραν αυτών της Βιοπληροφορικής.

Η πλατφόρμα της Matlab, καθότι αναπτύχθηκε για την εξυπηρέτηση ιδιωτικών συμφερόντων, διαθέτει ένα καλύτερο προγραμματιστικό περιβάλλον σε σχέση με την R. Αυτό συμπεριλαμβάνει πληρέστερα τεκμηριωμένο και ευκολότερα ερευνήσιμο documentation, αποδοτικότερα συστήματα εντοπισμού σφαλμάτων (debuggers), καθώς και φιλικότερο σύστημα περιήγησης αντικειμένων (object browser). Επίσης, η Matlab εξελίσσεται ταχύτερα προς την παραλληλοποίηση των μεταγλωττιστών της, αφού η παραλληλοποίηση αφορά όλα τα εργαλεία λογισμικού της, όχι μόνο του Bioinformatics toolbox.

Ακόμα, η Matlab προσφέρει μία μεγάλη γκάμα προγραμματιστικών εργαλείων (toolboxes) πέραν της Βιοπληροφορικής, καλύπτοντας αξιοσημείωτα ευρύ φάσμα των φυσικών επιστημών. Για παράδειγμα, ο χρήστης μπορεί να εμπλουτίσει την έρευνα του πάνω στα βιολογικά δεδομένα με εργαλεία από το χώρο του Control Engineering, Machine Learning, Signal Processing, Database management ή του Control Engineering. Ταυτόχρονα, τα εκτελέσιμα αρχεία που παράγει μπορούν είτε να λειτουργήσουν αυτόνομα σε υπολογιστικά συστήματα, χωρίς την ανάγκη η πλατφόρμα Matlab να είναι εγκατεστημένη στο σύστημα (independent executable files), είτε να μεταγλωττίσει τον κώδικα του σε γλώσσα C, VHDL κ.ά. Τέλος, η Matlab παρέχει τη

δυνατότητα για επικοινωνία σε πραγματικό χρόνο με πληθώρα συσκευών, όπως π.χ. μικροελεγκτές, κάμερες, DSPs, FPGAs κλπ. Όλα τα παραπάνω δεν είναι εφικτά με τη χρήση της R, ή τουλάχιστον απαιτούν πολύ κόπο και εξειδικευμένες γνώσεις. Αυτοί είναι και οι λόγοι για τους οποίους επιλέχθηκε η πλατφόρμα της Matlab για την υλοποίηση της παρούσας εργασίας.

5.2 Επεξεργασία μεγάλου όγκου βιολογικών δεδομένων

Η ανάγκη για την επεξεργασία βιολογικών δεδομένων διαρκώς αυξάνεται, καθώς προοδεύει η έρευνα στην Μοριακή Βιολογία και τη Στατιστική. Νέα βιολογικά δεδομένα ανακαλύπτονται διαρκώς, όπως και νέες μέθοδοι επεξεργασίας. Συχνά, ο όγκος των δεδομένων είναι τόσο μεγάλος ή αντίστοιχα η πολυπλοκότητα των αλγορίθμων είναι τόσο υψηλή, όπου μπορούν εύκολα να παραλύσουν έναν απλό προσωπικό υπολογιστή.

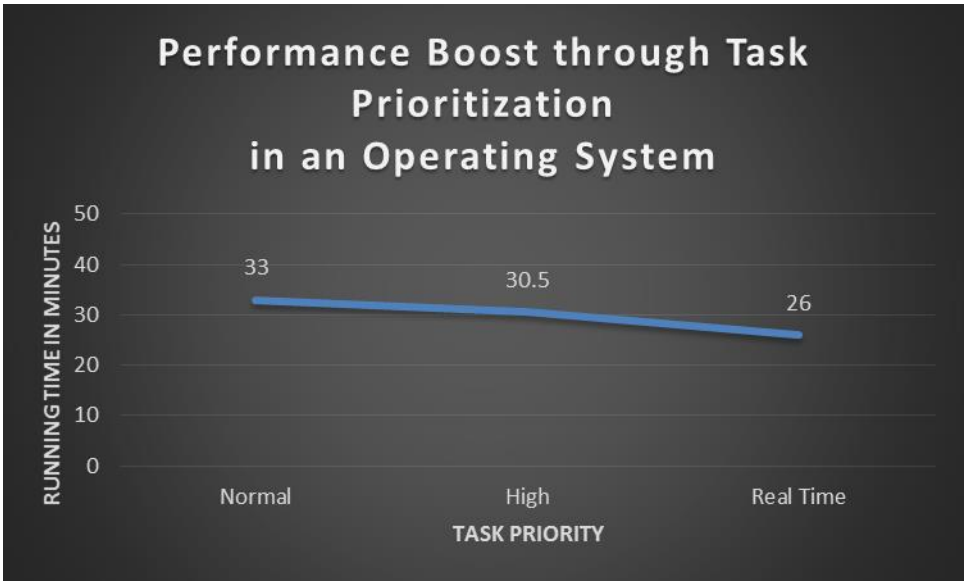
Οι ανάγκες της παρούσας εργασίας αποτελούνταν από έναν μεγάλο όγκο δεδομένων (πάνω από 3000 CEL files), τα οποία έπρεπε να επεξεργαστούν σε έναν προσωπικό υπολογιστή. Χρησιμοποιώντας κανείς τις έτοιμες συναρτήσεις από την πλατφόρμα Matlab / R παρατηρεί πως από την εισαγωγή 500 και πάνω αρχείων τα συστήματα αρχίζουν να παραλύουν.

Το πρώτο και βασικό πρόβλημα που έπρεπε να λυθεί ήταν πως μπορεί ένας σύγχρονος προσωπικός υπολογιστής μεσαίων επιδόσεων (για παράδειγμα ένα laptop με επεξεργαστή Intel i5, usb 3.0 και μνήμη RAM 4Gb και λειτουργικό σύστημα windows 7) να καταστεί ικανός να επεξεργαστεί gene expression data ανεξαρτήτου όγκου. Το πρόβλημα αυτό λύθηκε μέσω μιας σειράς προγραμματιστικών μεθόδων για τη βέλτιστη, αποδοτική και ασφαλή χρήση του εκάστοτε συστήματος. Οι τεχνικές που υιοθετήθηκαν για τη βελτίωση της απόδοσης αναλύονται παρακάτω.

Επιλογή και παραμετροποίηση λειτουργικού συστήματος

Κατ' αρχάς είναι σωστό να επιλεχθεί λειτουργικό σύστημα κωδικοποίησης 64bit. Η επιλογή αυτή οδηγεί σε ταχύτερη εκτέλεση πολλών εντολών αλλά και στην ικανότητα διευθυνσιοδότησης περισσότερης μνήμης RAM. Για παράδειγμα, η έκδοση του Matlab για Windows 32bit μπορεί να χρησιμοποιήσει μέχρι 2 GB RAM, ενώ για Windows 64 bit 500GB, ή ακόμα και 4 TB εάν πρόκειται για την έκδοση Windows Server 2012.

Επίσης το ίδιο το λειτουργικό σύστημα μπορεί να παραμετροποιηθεί για τη βελτίωση της απόδοσης του προγράμματος, με δύο απλούς σχετικά τρόπους. Ο πρώτος είναι μέσα από τη Διαχείριση Εργασιών (εάν πρόκειται για λειτουργικό σύστημα WIndows) να οριστεί το επίπεδο προτεραιότητας της εκτελούμενης διεργασίας (στη δεδομένη περίπτωση του Matlab). Εάν η προτεραιότητα αυξηθεί από «κανονική» σε «υψηλή» ή σε «πραγματικό χρόνο», τότε η απόδοση του προγράμματος αυξάνεται, όπως φαίνεται στο παρακάτω διάγραμμα.



Εικόνα 44. αύξηση απόδοσης μέσω task priority settings

Η ορθή χρήση της μνήμης RAM μπορεί επίσης να βελτιώσει την απόδοση. Ειδικότερα στην περίπτωση όπου οι απαιτήσεις των προγραμμάτων χρησιμοποιούν ολόκληρη τη μνήμη RAM ή την υπερβαίνουν. Γι' αυτό το λόγο πρέπει την ώρα που εκτελείται το πρόγραμμα να εκτελούνται παράλληλα μόνο οι απολύτως απαραίτητες διεργασίες του λειτουργικού συστήματος, ούτως ώστε να είναι ελεύθερη η περισσότερη κατά τα το δυνατόν ποσότητα μνήμης RAM. Στη συνέχεια, για να συνεχίσει η χρήση της μνήμης να είναι στο επιθυμητό επίπεδο, θα πρέπει να τρέχουν και ειδικές ρουτίνες (scripts) καθαρισμού της μνήμης cache του συστήματος από τις

μη χρησιμοποιούμενες σελίδες. Η διαδικασία αυτή είναι ακόμα περισσότερο απαραίτητη όταν ένας υπολογιστής λειτουργεί για πολλές ώρες ή και μέρες.

Ακόμα, άλλο ένα θέμα σχετικό με τη χρήση της μνήμης RAM είναι η εικονική μνήμη (virtual RAM). Δηλαδή, όταν τα πρόγραμμα που εκτελείται από το λειτουργικό σύστημα απαιτεί περισσότερη μνήμη από τη διαθέσιμη RAM του συστήματος, τότε το λειτουργικό σύστημα χρησιμοποιεί μνήμη από το σκληρό δίσκο του συστήματος για να εξυπηρετήσει αυτές τις ανάγκες. Ο σκληρός δίσκος όμως είναι τάξεις μεγέθους πιο αργός από την RAM, οπότε η απόδοση του συστήματος πέφτει κατακόρυφα. Ένας τρόπος για να προσπελαστεί αυτό το εμπόδιο είναι η χρήση ενός εξωτερικού USB3 flash disk («στικάκι»). Η προσαρμογή αυτή, μπορεί να μην είναι εξίσου αποτελεσματική με την εισαγωγή περισσότερης RAM για παράδειγμα, αλλά είναι κοντά σε αυτή την επιλογή και η πτώση της απόδοσης του συστήματος είναι πολύ μικρή.

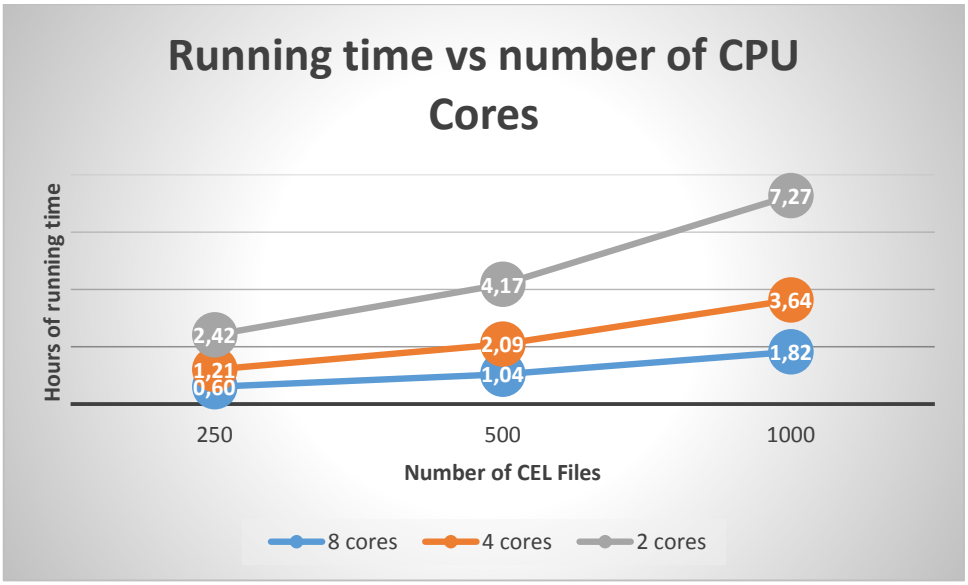
Block processing

Για την καλύτερη διαχείριση των μεγάλων αρχείων της εφαρμογής που αναπτύχθηκε, χρησιμοποιήθηκαν τεχνικές block processing και map indexing. Μέσα από ειδικές διαδικασίες είναι δυνατή η τμηματοποίση μεγάλων αρχείων, ούτως ώστε να φορτώνονται και να επεξεργάζονται τμηματικά (block processing). Η διαδικασία αυτή χρησιμοποιείται τόσο για την αύξηση της απόδοσης, όσο και για την ορθότερη χρήση της μνήμης σε περιπτώσεις όπου τα αρχεία είναι μεγαλύτερα από τη διαθέσιμη μνήμη RAM. Παράλληλα, με κατάλληλη διευθυνσιοδότηση (indexing), οι διαδικασίες αυτές βελτιστοποιούνται ακόμα περισσότερο.

Παραλληλοποίηση

Η τεχνική της παράλληλης επεξεργασίας (parallel computing) εκμεταλλεύεται στο έπακρον τις δυνατότητες των σύγχρονων επεξεργαστών. Πλέον οι επεξεργαστές διαθέτουν περισσότερους του ενός πυρήνες (συνήθως 2, 4 ή 8 ανά επεξεργαστή). Έτσι, επιτυγχάνεται η παράλληλη διεκπεραίωση διεργασιών. Επίσης, οι ίδιοι οι πυρήνες υλοποιούν τεχνικές «υπερνημάτωσης» (hyperthreading). Η τεχνολογία αυτή επιτρέπει την παράλληλη επεξεργασία δεδομένων σε κάθε πυρήνα, διπλασιάζοντας έτσι το throughput τους, αφού το λειτουργικό σύστημα αναγνωρίζει τον κάθε πυρήνα σαν δύο (στην πραγματικότητα πρόκειται για έναν πραγματικό και έναν λογικό).

Η πλατφόρμα Matlab είναι βελτιστοποιημένη ως προς τις παραπάνω δύο τεχνικές. Πολλές συναρτήσεις, κυρίως μαθηματικές, είναι “intrinsic multicore”, επιτρέποντας την ταυτόχρονη χρήση πολλών νημάτων (multithreaded). Ακόμα, μέσω του Parallel Computing Toolbox, επιτυγχάνεται η διεκπεραίωση ενός βρόγχου (for loop) σε πολλούς πυρήνες.



Εικόνα 45. Εικόνες αύξησης της απόδοσης μέσω της χρήσης πολλών πυρήνων.

Παρατηρήσεις

Εκτός από τις προαναφερθείσες τεχνικές, υπάρχουν και άλλες που χρησιμοποιούνται ευρέως για την αύξηση της απόδοσης, αλλά δεν ήταν κατάλληλες στην συγκεκριμένη εργασία. Ορισμένες από αυτές είναι η "GPU computing", η χρήση δηλαδή και των επεξεργαστών των καρτών γραφικών για την επεξεργασία δεδομένων, οι οποίοι είναι ταχύτεροι από τους συμβατικούς στον υπολογισμό ορισμένων πράξεων, κυρίως γραμμικής άλγεβρας. Όμως οι συμβατικές κάρτες γραφικών έχουν μικρή μνήμη, συνήθως 1 GB, κάτι το οποίο δεν είναι αρκετό στη δική μας περίπτωση και κάρτες γραφικών με περισσότερη μνήμη είναι «επαγγελματικού επιπέδου» και απαντώνται σε πολύ ακριβά υπολογιστικά συστήματα.

Άλλες τεχνικές είναι αυτές του Cluster Computing και του Hadoop. Σε αυτές τις περιπτώσεις χρησιμοποιούνται συστοιχίες από επεξεργαστές (υπερυπολογιστές) για την παράλληλη επεξεργασία των δεδομένων και συστοιχίες από σκληρούς δίσκους για την παράλληλη αποθήκευση και ανάγνωση δεδομένων. Όπως είναι αντιληπτό, δεν μπορούν να εφαρμοστούν σε έναν υπολογιστή, όπως είναι και το ζητούμενο.

5.3 Μεθοδολογία ανάπτυξης αλγορίθμων

Στο υποκεφάλαιο αυτό περιγράφεται η μεθοδολογία που χρησιμοποιήθηκε για τη στατιστική επεξεργασία των δεδομένων, μαζί με σχόλια πάνω στις προγραμματιστικές τεχνικές για τη βελτίωση της απόδοσης του προγράμματος.

Η είσοδος του προγράμματος είναι τα CEL files, τα οποία είναι το προϊόν εξόδου από τα συστήματα μικροσυστοιχιών της Affymetrix GeneChip. Στα αρχεία τύπου CEL αποθηκεύονται τα αποτελέσματα των υπολογισμών των τιμών εντάσεων των εικονοστοιχείων των αρχείων DAT, δηλαδή των τιμών των αρχικών, ανεπεξέργαστων δεδομένων (raw data). Στο αρχείο συμπεριλαμβάνονται επίσης η τιμή της τυπικής απόκλισης για κάθε τιμή έντασης, ο αριθμός των εικονοστοιχείων που χρησιμοποιήθηκαν για τον υπολογισμό της έντασης, ένα σήμα τύπου "flag" που υποδυκνείει τις ακραίες τιμές και ένα ακόμα σήμα τύπου "flag" ορισμένο από το χρήστη, που υποδεικνύει χαρακτηριστικά που πρέπει να εξαιρεθούν από την περαιτέρω ανάλυση των δεδομένων. Όλες οι πληροφορίες αποθηκεύονται με αντιστοίχηση σε συγκεκριμένο probe της μικροσυστοιχίας.

Τα αρχεία τύπου CDF περιέχουν τις κατάλληλες πληροφορίες για τη διάταξη μιας μικροσυστοιχίας Affymetrix GeneChip και λειτουργούν ως αρχεία βιβλιοθήκης (library files). Η διάταξη μιας μικροσυστοιχίας μπορεί να συμπεριλαμβάνει σύνολα από probes τύπου Expression, Genotyping, CustomSeq, Copy Number ή/και Tag. Όλα τα ονόματα των συνόλων των probes μιας μικροσυστοιχίας είναι μοναδικά. Μέσα στο ίδιο αρχείο τύπου CDF μπορεί να συμπεριλαμβάνονται αντίγραφα συνόλων probes, με επίσης μοναδικά ονόματα.

Παρακάτω αναλύονται προγραμματιστικές τεχνικές με τις οποίες αναπτύξαμε τους αλγορίθμους μας , με ιδιαίτερη έμφαση στον τομέα της απόδοσής και της ορθής διαχείρισης της μνήμης.

Preprocessing - Εφαρμογή αλγορίθμου RMA στα CEL files

Στην εκκίνηση του προγράμματος, ζητείται από το χρήστη να επιλέξει το φάκελο με τα αρχεία CEL. Όλα τα αρχεία πρέπει να βρίσκονται στον ίδιο φάκελο. Στη συνέχεια ζητείται από το χρήστη να επιλέξει το κατάλληλο αρχείο βιβλιοθήκης, τύπου CDF (Chip Description FIle), το οποίο περιέχει τις πληροφορίες για την αποκωδικοποίηση των δεδομένων που περιέχονται στα CEL files.

Ουσιαστικά το πρόγραμμα μέσω ενός βρόγχου, εφαρμόζει τον αλγόριθμο RMA για κάθε αρχείο τύπου CEL. Η συνάρτηση υλοποίησης παρέχεται από το Bioinformatics Toolbox του Matlab και είναι η affyrma(). Η συγκεκριμένη συνάρτηση είναι ήδη βελτιστοποιημένη, οπότε δεν αφήνει περιθώρια επέμβασης. Η αύξηση της απόδοσης του συστήματος σε αυτό το βήμα γίνεται με την παραλληλοποίηση του βρόγχου. Δηλαδή η τροποποίηση με τέτοιο τρόπο ούτως ώστε ο βρόγχος να τρέχει παράλληλα σε όλους τους πυρήνες του επεξεργαστή. Ο εκτιμώμενος χρόνος εκτέλεσης τότε μειώνεται περίπου γραμμικά, όπως φαίνεται στην εικόνα 45, ανάλογα με τον αριθμό των διαθέσιμων φυσικών πυρήνων του επεξεργαστή.

Η έξοδος του βήματος αυτού είναι ένα αρχείο κειμένου (txt file) για κάθε αρχείο CEL, με τον αντίστοιχο τίτλο, το οποίο περιέχει τα αποτελέσματα του αλγορίθμου RMA στο συγκεκριμένο αρχείο. Ο λόγος που επιλέχθηκε η παραγωγή αυτών των αρχείων και όχι η προσωρινή αποθήκευση τους στη μνήμη είναι αφενός μεν για την αποσυμφόρηση της μνήμης RAM και αφετέρου για να εξασφαλιστεί ότι αν κανείς επιθυμεί να τροποποιήσει κάποιο επόμενο στάδιο του προγράμματος, να μη χρειαστεί να επαναλάβει αυτό το χρονοβόρο βήμα. Επίσης τα αρχεία κειμένου καταλαμβάνουν ελάχιστο χώρο στη μνήμη και ο λόγος συμπίεσης τους είναι πολύ υψηλός.

Φόρτωση των αποτελεσμάτων της εφαρμογής του αλγορίθμου RMA στη μνήμη

Αφού ολοκληρωθεί το προηγούμενο βήμα, πραγματοποιείται η φόρτωση των αποτελεσμάτων της εφαρμογής του αλγορίθμου RMA στη μνήμη. Για να εξασφαλιστεί η ευστάθεια του συστήματος, τα αρχεία διαβάζονται σειριακά και κατασκευάζεται επίσης σειριακά ένας συγκεντρωτικός πίνακας με τα αποτελέσματα.

Σε αυτό το βήμα μπορούν να γίνουν σημαντικές βελτιστοποιήσεις όσον αφορά τη διαχείριση της μνήμης. Αρχικά χρησιμοποιούνται οι κατάλληλοι δείκτες (pointers) για τον εντοπισμό τον αρχείων και μόλις ολοκληρωθεί η ανάγνωση ενός αρχείου οι δείκτες αυτοί διαγράφονται, ούτως ώστε κάθε στιγμή να είναι μόνο ένα αρχείο ανοικτό. Έτσι αποφεύγονται πιθανά "conflicts" που μπορούν να παρουσιαστούν όταν πάρα πολλά αρχεία είναι ανοικτά ταυτόχρονα.

Στη συνέχεια, καθώς κατασκευάζεται ο συγκεντρωτικός πίνακας με τα αποτελέσματα, ιδιαίτερη προσοχή πρέπει να δοθεί στον τύπο δεδομένων του πίνακα αυτού. Επειδή τα δεδομένα είναι «μεικτού τύπου» (mixed type), δηλαδή πρόκειται και για αριθμούς και για κείμενο, π.χ. τίτλοι αρχείων κλπ, αν χρησιμοποιήσουμε mixed type πίνακα, όπως είναι ο πίνακας τύπου cell στη Matlab, τότε ο όγκος του παραγόμενου πίνακα θα είναι τεράστιος, ιδιαίτερα όταν έχουμε μερικές χιλιάδες αρχεία εισόδου.

Γι' αυτό το λόγο, η ορθή πρακτική έγκειται στη δημιουργία δύο πινάκων, έναν μόνο για τα αριθμητικά στοιχεία κι έναν μόνο με τα στοιχεία κειμένου. Εννοείται πως πρέπει να εξασφαλιστεί μία αντιστοιχία με κάποιον τρόπο (π.χ. αντιστοιχία με τις θέσεις των στοιχείων ή κάποιο index log) ανάμεσα στους δύο αυτούς πίνακες. Με αυτόν τον τρόπο αποσυμφορίζεται η μνήμη και καθίσταται δυνατή η επεξεργασία χιλιάδων αρχείων σε έναν υπολογιστή χωρίς προβλήματα.

Αντιστοίχιση probes - Genes

Εδώ αντιστοιχίζονται τα probes των αρχείων εισόδου με τα αντίστοιχα γονίδια. Η λίστα αντιστοίχισης έχει κατασκευαστεί σε ξεχωριστό βήμα, το οποίο περιγράφεται αργότερα.

Η διαδικασία αυτή εμπεριέχει διαδικασίες παρόμοιες με αυτές που συναντώνται στα queries των βάσεων δεδομένων, δηλαδή αντιστοίχιση στοιχείων από δύο πίνακες με βάση κάποια συνθήκη, κάτι το οποίο γίνεται ταχύτατα από αντίστοιχες πλατφόρμες λογισμικού βάσεων δεδομένων, οι οποίες διαθέτουν τα κατάλληλα συστήματα indexing. Στη συγκεκριμένη περίπτωση όμως η βελτιστοποιήσεις γίνονται με ιδιοκατασκευές αλγορίθμων.

Το πρόβλημα στην απόδοση εμφανίζεται όταν οι πίνακες έχουν μεγάλο μέγεθος. Αναζητώντας ένα στοιχείο σε έναν μεγάλο πίνακα και στη συνέχεια αναζητώντας το αντίστοιχο του σε έναν επίσης μεγάλο πίνακα μπορεί να οδηγήσει σε ιδιαίτερα χρονοβόρα διαδικασία. Ευτυχώς, για τη βελτιστοποίηση αυτών των διεργασιών υπάρχει τεράστια βιβλιογραφία στον τομέα των αλγορίθμων της επιστήμης υπολογιστών.

Η ορθή προγραμματιστική προσέγγιση σε αυτό το βήμα έγκειται καταρχάς στην ταξινόμηση των πινάκων με κάποιο κριτήριο, π.χ. αλφαβητική ταξινόμηση ή δυαδικών δένδρων (binary trees), ούτως ώστε η πολυπλοκότητα να μειωθεί από εκθετική (αν δεν κάναμε καμία βελτιστοποίηση) σε λογαριθμική (με βελτιστοποιήσεις). Επίσης, σε περιπτώσεις που οι πίνακες έχουν πολύ μεγάλο μέγεθος σε μνήμη και σε πεδία, τότε μπορούμε να δημιουργήσουμε έναν ιδιοκατασκευασμένο τύπο διευθυνσιοδότησης (indexing), ούτως ώστε να μη συνωστίζεται η μνήμη και οι αναζητήσεις να γίνονται γρηγορότερα.

Υπολογισμός μέσων όρων τιμών των probes που ανήκουν στο ίδιο γονίδιο

Σε συνέχεια του βήματος 3, υπολογίζονται οι μέσοι όροι των ομάδων probes που αντιπροσωπεύουν το ίδιο γονίδιο και παράγεται ένας καινούργιος πίνακας, ο οποίος θα περιέχει σε κάθε γραμμή τα εξής:

- Πεδίο 1: Gene Symbol
- Πεδίο 2: Μία συμβολοσειρά (string) με όλα τα ονόματα των probes που αντιστοιχούν στο εν λόγω gene symbol
- Πεδίο 3 έως N: Η μέση τιμή που προέκυψε από το σύνολο των probes που αντιστοιχούν στο εν λόγω gene symbol

Στο πεδίο 2, όπου όλα τα ονόματα των probes περιέχονται σε μία συμβολοσειρά, διαχωρίζονται μεταξύ τους με το σύμβολο "#". Το σύμβολο αυτό επιλέχθηκε επειδή δεν περιέχεται στα ονόματα των probes, σε αντίθεση π.χ. με το σύμβολο "/", το οποίο περιέχεται.

Επίσης, σε αυτό το σημείο χρειάζεται ιδιαίτερη προσοχή στη μνήμη του συστήματος. Ο πίνακας που φορτώθηκε στο βήμα 3 από το βήμα 2, δε χρειάζεται να παραμείνει στη μνήμη για το βήμα 4. Επίσης, στο βήμα 4 υπάρχει ο πίνακας του βήματος 3 και ταυτόχρονα δημιουργείται ένας ακόμα πίνακας που περιέχει μόνο τους μέσους όρους των τιμών των γονιδίων. Επειδή οι πίνακες ενδέχεται να έχουν τεράστιο μέγεθος (ανάλογα φυσικά με τον αριθμό των δειγμάτων που εισάγονται στο μοντέλο), αν ακολουθήσει κανείς τις οδηγίες για τον τύπο των πινάκων που περιεγράφηκαν στο βήμα 2, τότε δεν θα υπάρξει πρόβλημα. Σε περίπτωση όμως όπου ο αριθμός των δειγμάτων ξεπερνά τα 15000, τότε καλό θα ήταν ο πίνακας που δημιουργείται στο βήμα 4 να μην δημιουργείται στη μνήμη RAM, αλλά απευθείας στο δίσκο, βήμα-βήμα, με την τεχνική "append text to an existing txt file".

Από πλευράς υπολογιστικών πόρων, το βήμα 4, αν και δεν παραλληλοποιείται λόγω της φύσης των υπολογισμών, εκτελείται γρήγορα και δεν αναμένεται να δημιουργήσει προβλήματα στο σύστημα. Αφού ολοκληρωθεί, όλες οι μεταβλητές του

μοντέλου εκτός του πίνακα του βήματος 4, δεν χρειάζονται και πρέπει να διαγραφούν από τη μνήμη του συστήματος.

Lowess

Την κανονικοποίηση Lowess την εκτελέσαμε στην Matlab, με την συνάρτηση mainvarsetnorm. Η συγκεκριμένη συνάρτηση δέχεται πολλά ορίσματα και παραμέτρους. Τα δύο βασικά ορίσματα που απαιτεί οπωσδήποτε , είναι οι τιμές του δείγματος που πρόκειται να κανονικοποιηθεί και οι τιμές του bulk ή unamplified control,δηλαδή του δείγματος που δεν περιέχει πειραματικό σφάλμα και με βάση αυτό πρόκειται να κανονικοποιηθεί το εκάστοτε δείγμα που μας ενδιαφέρει. Εκτός από αυτά τα δύο ορίσματα η συγκεκριμένη συνάρτηση μπορεί να παραμετροποιηθεί και να κανονικοποιήσει τα δεδομένα που τις εισάγονται με το μέγεθος παραθύρου (Span) που επιθυμεί ο κάθε ερευνητής . Το πρόγραμμα matlab , έχει σαν default επιλογή το μέγεθος παραθύρου 5%, μιας και είναι το πιο σύνηθες. Τέλος με την σωστή παραμετροποίηση (Showplot), ο ερευνητής μπορεί να εξάγει τα MvA plots , όπου είναι σχεδιαγράμματα που στην συγκεκριμένη περίπτωση μπορούν να χρησιμοποιηθούν για υποδείξουν κατά πόσο πέτυχε ή όχι η κανονικοποίηση.

Global Normalization

Η εφαρμογή της μεθόδου Global Normalization, όπως περιεγράφηκε στην υποενότητα 4.9 ,υπολογιστικά απαιτεί ορισμένες πράξεις γραμμικής άλγεβρας, για τις οποίες η πλατφόρμα Matlab είναι ήδη βελτιστοποιημένη. Οι πράξεις είναι αρκετά απλές και δεν απαιτούν ιδιαίτερη επεξεργαστική ισχύ.

Το σημείο προσοχής σε αυτό το βήμα είναι η διαχείριση της μνήμης, καθώς οι καινούργιοι πίνακες που δημιουργούνται ενδέχεται να έχουν ιδιαίτερα αυξημένο μέγεθος, ανάλογα με τον αριθμό των δειγμάτων. Και εδώ, η ορθή προσέγγιση στη διαχείριση της μνήμης ακολουθεί τις οδηγίες που δόθηκαν στο βήμα 2 – «Φόρτωση των αποτελεσμάτων της εφαρμογής του αλγορίθμου RMA στη μνήμη».

Επίσης, ο τελικός normalized πίνακας, ως αποτέλεσμα της εφαρμογής της μεθόδου, πρέπει να αποθηκευτεί στη μνήμη του συστήματος. Ο πίνακας αυτός θα έχει ένα μέγεθος της τάξης μερικών εκατοντάδων MB έως αρκετά GB (π.χ. για 2000 αρχεία, το μέγεθος του πίνακα είναι περίπου 1.5 GB). Η διαδικασία εγγραφής επιταχύνεται αν αποφευχθεί η αποθήκευση στο σκληρό δίσκο και πραγματοποιηθεί σε μία μονάδα δίσκου flash memory USB 3.0.

Κανονικοποίηση με χρήση της Διακύμανσης

Η μεθοδολογία που ακολουθήσαμε για τον υπολογισμό των variances των δεδομένων, όπως περιεγράφηκε σε προηγούμενο κεφάλαιο αποτελεί εφαρμογή τύπων ,υλοποιείται εξίσου εύκολα στο matlab και στο excel και υπολογίζεται μέσα σε ελάχιστα δευτερόλεπτα.

Κεφάλαιο 6 - Συμπεράσματα και μελλοντικές επεκτάσεις

6.1 Συμπεράσματα

Η εφαρμογή κλασικών μεθόδων κανονικοποίησης και επεξεργασίας πειραματικών δεδομένων από μικροσυστοιχίες σε δείγματα από μεμονωμένα κύτταρα (single cells) δεν αρκεί για την επιτυχή ταυτοποίηση τους, αρκεί όμως για την μερική διόρθωση του πειραματικού σφάλματος με χρήση unamplified control δειγμάτων. Το επιπλέον βήμα απομόνωσης και χρήσης γονιδίων βασισμένο στη διακύμανση της γονιδιακής τους έκφρασης αποτέλεσε κομβικό σημείο για την επιπλέον αφαίρεση πειραματικού σφάλματος. Το παραπάνω επέτρεψε την επιτυχή ταυτοποίηση δειγμάτων όπου αυτή ήταν δυνατή σε ένα σύνολο από δείγματα κυτταρικών σειρών. Συγκεκριμένα με την χρήση αυτής της μεθόδου επιτεύχθηκε:

- Η κατηγοριοποίηση της κυτταρικής σειράς 4 με δείγματα μαστού.
- Η κατηγοριοποίηση του unamplified control της κυτταρικής σειράς 5 με το αντίστοιχο δείγμα καθώς και στον ίδιο cluster και τα υπόλοιπά δείγματα από μεμονωμένα κύτταρα της ίδιας κυτταρικής σειράς
- Η ομαδοποίηση των δειγμάτων όμοιων κυτταρικών σειρών (Cell lines 1 ,2,3).

Οι Κυτταρικές Σειρές του Query Dataset ,Cell lines 1 ,2 και 3, που δεν κατηγοριοποιήθηκαν , δεν υπήρχαν εν τέλει στα δείγματα των Reference Datasets που χρησιμοποιήσαμε. Ωστόσο κατηγοριοποιήθηκαν μεταξύ τους. Τέλος τα CTCs καθώς δεν είχαμε στα δεδομένα ιστούς, δε μπορούμε να αναγνωρίσουμε από που προέρχονται, αλλά καταλήξαμε στο γεγονός ότι δεν είναι όμοια μεταξύ τους.

Η επεξεργασία των παραπάνω δεδομένων καθώς και το τελικό αποτέλεσμα στηρίχθηκε στην ex novo δημιουργία και συνεχή βελτίωση ενός λογισμικού πακέτου. Το τελευταίο επιτρέπει την επέκταση και επεξεργασία των δεδομένων σε μελλοντικές έρευνες. Οι αλλαγές που μπορεί να γίνουν είναι τόσο με μεθόδους κανονικοποίησης όσο χρήσης νέων δεδομένων από τις συγκεκριμένες μικροσυστοιχίες Επίσης επιτρέπει την ολοκλήρωση της ανάλυσης των βιολογικών δεδομένων από έναν υπολογιστή, με βελτιστοποιημένες τεχνικές στη διαχείριση της μνήμης και του χρόνου επεξεργασίας.

6.2 Μελλοντικές επεκτάσεις

Η παρούσα εργασία αποτελεί τη βάση για τη δημιουργία ενός καινοτόμου προγραμματιστικού εργαλείου διαχείρισης και επεξεργασίας βιολογικών δεδομένων, που θα χρησιμοποιηθεί από τους ερευνητές γονιδιακής έκφρασης, με σκοπό αφενός την παραγωγή αξιόπιστων και συγκρίσιμων αποτελεσμάτων και αφετέρου την ελευθερία επιλογής ανάμεσα από μία πλήρη γκάμα στατιστικών μεθόδων. Επιπλέον, ένας ερευνητής θα πρέπει να έχει και την ελευθερία να παραμετροποιήσει ή να ιδιοκατασκευάσει οποιοδήποτε στάδιο της ανάλυσης. Προς αυτή την κατεύθυνση, πρέπει να γίνουν ακόμα οι ακόλουθες ενέργειες:

1. Ανάπτυξη γραφικού περιβάλλοντος

Προς το παρόν, όλες οι εντολές για την επεξεργασία των δεδομένων γίνονται μέσω της κονσόλας του Matlab. Το γεγονός αυτό προϋποθέτει αφενός εξοικείωση με το περιβάλλον της Matlab και αφετέρου προγραμματιστικές γνώσεις. Αντίθετα, θα πρέπει η εφαρμογή να ενσωματωθεί σε ένα κατάλληλο γραφικό περιβάλλον, όπου ο χρήστης θα καθοδηγείται μέσω βημάτων (wizard) και ειδικών αναδυόμενων μενού (pop-up menu) για την επιλογή και την παραμετροποίηση των μεθόδων επεξεργασίας.

2. Εμπλουτισμός λειτουργικότητας βάσης δεδομένων

Τα αποτελέσματα που παράγονται σε κάθε στάδιο της επεξεργασίας θα πρέπει να αποθηκεύονται και να διαχειρίζονται από μία ειδική βάση δεδομένων. Προς το παρόν αυτή η βάση είναι σε πρωταρχικό στάδιο και παρέχει βασική λειτουργικότητα. Σε ένα μελλοντικό στάδιο θα πρέπει να εμπλουτιστεί με επιπλέον λειτουργικότητες ώστε να είναι δυνατή η ανάκτηση υποσυνόλων βασισμένη σε συγκεκριμένες τιμές έκφρασης, ειδικευμένα στατιστικά αποτελέσματα, σημειολογία (annotation) των δειγμάτων, κατ' επιλογήν ομάδες γονιδίων ή με βάση την πειραματική διαδικασία που ακολουθήθηκε.

3. Συγκριτική μελέτη μεθόδων Κανονικοποίησης και Κατηγοριοποίησης

Όπως προαναφέρθηκε το ζήτημα της κανονικοποίησης και κατηγοριοποίησης των δεδομένων παραμένει ακόμα ανοιχτό. Ένας λόγος είναι πως δεν υπάρχουν συγκριτικές μελέτες που να αφορούν πληθώρα αλγορίθμων με ταυτόχρονη συγκριτική μελέτη πάνω στις μικρορυθμίσεις του εκάστοτε αλγορίθμου κανονικοποίησης / κατηγοριοποίησης, ούτως ώστε να βρεθεί η βέλτιστη κατά το δυνατόν παραμετροποίηση τους. Με το παρόν εργαλείο, μία τέτοια μελέτη είναι δυνατή και απαραίτητη.

Βιβλιογραφία

[1] Wikipedia, Βιοπληροφορική, available at: http://el.wikipedia.org/wiki/%CE%92%CE%B9%CE%BF%CF%80%CE%BB%CE%B7%CF%81%CE%BF%CF%86%CE%BF%CF%81%CE%B9%CE%BA%CE%AE[2] Αικατερίνη Γ. Περδικούρη & Αθανάσιος Κ. Τσακαλίδης, "Εισαγωγή στην Βιοπληροφορική", Μάρτιος 2004, Διπλωματική εργασία, Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής, Πανεπιστήμιο Πατρών

[2] Αικατερίνη Γ. Περδικούρη & Αθανάσιος Κ. Τσακαλίδης, "Εισαγωγή στην Βιοπληροφορική", Μάρτιος 2004, Διπλωματική εργασία, Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής, Πανεπιστήμιο Πατρών

[3] Συμεωνίδη Κάτια, "Βελτιστοποίηση υποστρώματος γονιδιακής ανάλυσης με χρήση γενετικών αλγορίθμων", 2010, Διπλωματική εργασία, Πολυτεχνείο Χανιά

[4] Τ. Θηραίος, Εισαγωγή στην Βιοπληροφορική, Διάλεξη, Γεωπονικό Πανεπιστήμιο Αθηνών, Τμήμα Βιοτεχνολογίας , Εργαστήριο Γενετικής, Available at : http://biotech.aua.gr/BSC_COURSES/bioinf/lecture1_ada.pdf

[5] Wikipedia, cell culture (κυτταρικές καλλιέργειες), Available at: http://en.wikipedia.org/wiki/Cell_culture

[6] Eisenberg E, and Levanon EY (July 2003). "Human housekeeping genes are compact". *TRENDS in Genetics* 19 (7): 362–365. doi:10.1016/S0168-9525(03)00140-9. PMID 12850439.

[7] kon Butte, AJ. et al. (2001). "Further defining housekeeping, or "maintenance," genes focus on 'a compendium of gene expression in normal human tissues'.". *Physiol.Genomics* 7 (2): 95–96. PMID 11773595.

[8] Zhu, J. et al. (2008). "On the nature of human housekeeping genes" , *Trends in genetics* 24 (10): 481–484. doi:10.1016/j.tig.2008.08.004.PMID 18786740.

[9] Quiagen. "RT2 Profiler PCR Array (96-Well Format and 384-Well Format". *Qiagen catalog no. 330231 PAHS-00ZA.*

[10] Greer S, Honeywell R, Geletu M, Arulanandam R, Raptis L (Feb 19, 2010). "Housekeeping genes; expression levels may change with density of cultured cells." *J Immunol Methods* 355 (1–2): 76–9. doi:10.1016/j.jim.2010.02.006. PMID 20171969.

[11] Tan SC, Carr CA, Yeoh KK, Schofield CJ, Davies KE, Clarke K. (Nov 2011). "Identification of valid housekeeping genes for quantitative RT-PCR analysis of cardiosphere-derived cells preconditioned under hypoxia or with prolyl-4-hydroxylase inhibitors.". *Mol Biol Rep* 39 (4): 4857–67. doi:10.1007/s11033-011-1281-5. PMC 3294216. PMID 22065248.

[12] wikipedia, Γονδιακή έκφραση, available at: http://el.wikipedia.org/wiki/%CE%93%CE%BF%CE%BD%CE%B9%CE%B4%CE%B9%CE%B1%CE%BA%CE%AE_%CE%AD%CE%BA%CF%86%CF%81%CE%B1%CF%83%CE%B7

[13] "Μικροσυστοιχίες και ανάλυση δεδομένων", Διάλεξη, University of Cyprus, Department of Computer Science, available at: http://www.cs.ucy.ac.cy/courses/EPL450/lectures/microarrays.pdf

[14] ΑΙΚΑΤΕΡΙΝΗ Ε. ΣΚΟΥΤΑ, " Ανάλυση γονιδιακής έκφρασης μικροσυστοιχιών σε σειρές λευχαιμίας ", Ιούλιος 2009, Διπλωματική Εργασία, Εθνικό Μετσόβιο Πολυτεχνείο, , available at: http://artemis-new.cslab.ece.ntua.gr:8080/jspui/bitstream/123456789/4901/1/DT2009-0115.pdf

[15]Wikipedia, Καρκίνος, available at: http://el.wikipedia.org/wiki/%CE%9A%CE%B1%CF%81%CE%BA%CE%AF%CE%BD%CE%BF%CF%82

[16] Αηδονόπουλος Ορφέας, "Εξόρυξη γνώσης από κυτταρικά δεδομένα ιστικών μικροσυστοιχιών", Διπλωματική Εργασία τμήματος «Μηχανικών Η/Υ και Πληροφορικής» της Πολυτεχνικής σχολής του Πανεπιστημίου Πατρών.

[17] Πέτρος Γ. Δέδες, "Κυτταρική και μοριακή μελέτη της επίδρασης του ζολενδρονικού οξέος σε κυτταρικές σειρές από καρκίνο του μαστού και σε οστικά κύτταρα", 2009, Διδακτορική Διατριβή, Πανεπιστήμιο Πατρών, Τμήμα Χημείας Τομέας Οργανικής Χημείας, Βιοχημείας & Φυσικών Προϊόντων, available at : http://nemertes.lis.upatras.gr/jspui/bitstream/10889/3882/1/petr%20pdf.pdf

[18] Lui, Harvey, et al. "Real-time Raman spectroscopy for in vivo skin cancer diagnosis." Cancer research 72.10 (2012): 2491-2500.

[19] Schwab, Manfred. Encyclopedia of Cancer (2nd edition), Springer. 2008.

[20] Filder IJ. The pathogenesis of cancer metastasis: the seed and soil hypothesis revised. Nat. Rev. Cancer 3:453-458. 2003.

[21] Journal of Oncology, doi:10.1155/2010/426218

[22] Ε.Φιλόπουλος, "Τα κυκλοφορούντα καρκινικά κύτταρα (circulating tumor cells - CTCs) ως προγνωστικός και προβλεπτικός δείκτης σε μεταστατικό καρκίνο",Άρθρο, 2011, available at: http://perimastologias.blogspot.gr/2011/06/circulating-tumor-cells-ctcs.html

[23] Κυκλοφορούντα καρκινικά κύτταρα, available at : http://www.bio-scientist.com/news/%CE%BA%CF%85%CE%BA%CE%BB%CE%BF%CF%86%CE%BF%CF%81%CE%BF%CF%8D%CE%BD%CF%84%CE%B1-%CE%BA%CE%B1%CF%81%CE%BA%CE%B9%CE%BD%CE%B9%CE%BA%CE%AC-%CE%BA%CF%8D%CF%84%CF%84%CE%B1%CF%81%CE%B1/

[24] Wikipedia, Μικροσυστοιχίες γονιδίων, available at : http://el.wikipedia.org/wiki/%CE%9C%CE%B9%CE%BA%CF%81%CE%BF%CF%83%CF%85%CF%83%CF%84%CE%BF%CE%B9%CF%87%CE%AF%CE%B5%CF%82_%CE%B3%CE%BF%CE%BD%CE%B9%CE%B4%CE%AF%CF%89%CE%BD

[25] Διονυσία Συμεωνίδη, "Βιοπληροφορική ανάλυση και χαρακτηρισμός γονιδίων που εμπλέκονται στη φαινοτυπική πλαστικότητα του zebrafish (Danio rerio, Hamilton 1822) ", 2011, Μεταπτυχιακό Δίπλωμα, Πανεπιστήμιο Πατρών

[26] http://www.embl.it/

[27] Παπαδημητρίου Χρήστος, "Μελέτη της διαφορικής έκφρασης γονιδίων βλαστικών/προγονικών κυττάρων του αιμοποιητικού συστήματος σε γενετικά τροποποιημένους μύες με τη χρήση μικροσυστοιχιών DNA", 2013, ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΑΤΡΙΒΗ, Πανεπιστήμιο Πατρών

[28] Moran, G., et al., Comparative genomics using Candida albicans DNA microarrays reveals absence and divergence of virulence-associated genes in Candida dubliniensis. Microbiology, 2004. 150(Pt 10): p. 3363-82.

[29] Adomas, A., et al., Comparative analysis of transcript abundance in Pinus sylvestris after challenge with a saprotrophic, pathogenic or mutualistic fungus. Tree Physiol, 2008. 28(6): p. 885-97.

[30] Pollack, J.R., et al., Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet, 1999. 23(1): p. 41-6.

[31] Moran, G., et al., Comparative genomics using Candida albicans DNA microarrays reveals absence and divergence of virulence-associated genes in Candida dubliniensis. Microbiology, 2004. 150(Pt 10): p. 3363-82.

[32] Aparicio, O., J.V. Geisberg, and K. Struhl, Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. Curr Protoc Cell Biol, 2004. Chapter 17: p. Unit 17 7.

[33] Van Steensel, B. and S. Henikoff, Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. Nat Biotechnol, 2000. 18(4): p. 424-8.

[34] Hacia, J.G., et al., Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. Nat Genet, 1999. 22(2): p. 164-

[35] Yazaki, J., B.D. Gregory, and J.R. Ecker, Mapping the genome landscape using tiling array technology. Curr Opin Plant Biol, 2007. 10(5): p. 534-42.

[36] Der, S.D., et al., ***Identification of genes differentially regulated by interferon alpha, beta, or gamma using oligonucleotide arrays***. Proc Natl Acad Sci U S A, 1998. 95(26): p. 15623-8.

[37] Iyer, V.R., et al., ***The transcriptional program in the response of human fibroblasts to serum.*** Science, 1999. 283(5398): p. 83-7.

[38] Brown, P.O. and D. Botstein, ***Exploring the new world of the genome with DNA microarrays.*** Nat Genet, 1999. 21(1 Suppl): p. 33-7.

[39] Eisen, M.B., et al., ***Cluster analysis and display of genome-wide expression patterns.*** Proc Natl Acad Sci U S A, 1998. 95(25): p. 14863-8.

[40] Lockhart, D.J., et al., ***Expression monitoring by hybridization to high-density oligonucleotide arrays.*** Nat Biotechnol, 1996. 14(13): p. 1675-80.

[41] Wodicka, L., et al., ***Genome-wide expression monitoring in Saccharomyces cerevisiae.*** Nat Biotechnol, 1997. 15(13): p. 1359-67.

[42] DeRisi, J.L., V.R. Iyer, and P.O. Brown, ***Exploring the metabolic and genetic control of gene expression on a genomic scale.*** Science, 1997. 278(5338): p. 680-6.

[43] Chiang, L.W., et al., ***An orchestrated gene expression component of neuronal programmed cell death revealed by cDNA array analysis***. Proc Natl Acad Sci U S A, 2001. 98(5): p. 2814-9.

[44] Golub, T.R., et al., ***Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.*** Science, 1999. 286(5439): p. 531-7.

[45] Scherf, U., et al., ***A gene expression database for the molecular pharmacology of cancer***. Nat Genet, 2000. 24(3): p. 236-44.

[46] Weinstein, J.N., et al., ***An information-intensive approach to the molecular pharmacology of cancer.*** Science, 1997. 275(5298): p. 343-9.

[47] Chee, M., et al., ***Accessing genetic information with high-density DNA arrays.*** Science, 1996. 274(5287): p. 610-4.

[48] Ramsay, G., ***DNA chips: state-of-the art.*** Nat Biotechnol, 1998. 16(1): p. 40-4.

[49] Gerhold, D., T. Rushmore, and C.T. Caskey, ***DNA chips: promising toys have become powerful tools***. Trends Biochem Sci, 1999. 24(5): p. 168-73.

[50] Schena, M., et al., ***Quantitative monitoring of gene expression patterns with a complementary DNA microarray.*** Science, 1995. 270(5235): p. 467-70.

[51] Alon, U., et al., ***Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.*** Proc Natl Acad Sci U S A, 1999. 96(12): p. 6745-50.

[52] Lipshutz, R.J., et al., ***High density synthetic oligonucleotide arrays.*** Nat Genet, 1999. 21(1 Suppl): p. 20-4.

[53] Lipshutz, R.J., et al., ***Using oligonucleotide probe arrays to access genetic diversity.*** Biotechniques, 1995. 19(3): p. 442-7.

[54] Li, C. and W. Hung Wong, ***Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application***. Genome Biol, 2001. 2(8).

[55] Li, C. and W.H. Wong, ***Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection***. Proc Natl Acad Sci U S A, 2001. 98(1): p. 31-6.

[56] Ideker, T., et al., ***Testing for differentially expressed genes by maximum likelihood analysis of microarray data***. J Comput Biol, 2000. 7(6): p. 805-17.

[57] Kerr, M.K., M. Martin, and G.A. Churchill, ***Analysis of variance for gene expression microarray data***. J Comput Biol, 2000. 7(6): p. 819-37.

[58] Kerr, M.a.C., GA, ***Experimental design for gene expression microarrays.*** Biostatistics, 2001. 2(2): p. 183-201.

[59] Thomas, J.G., et al., ***An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles***. Genome Res, 2001. 11(7): p. 1227-36.

[60] Arfin, S.M., et al., ***Global gene expression profiling in Escherichia coli K12. The effects of integration host factor***. J Biol Chem, 2000. 275(38): p. 29672-84.

[61] Newton, M.A., et al., ***On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data***. J Comput Biol, 2001. 8(1): p. 37-52.

[62] Tamayo, P., et al., ***Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation***. Proc Natl Acad Sci U S A, 1999. 96(6): p. 2907-2912.

[63] Törönen, P., et al., ***Analysis of gene expression data using self-organizing maps***. FEBS Lett, 1999. 451(2): p. 142-6.

[64] Heyer, L.J., S. Kruglyak, and S. Yooseph, ***Exploring Expression Data: Identification and Analysis of Coexpressed Genes.*** Genome Res, 1999. 9(11): p. 1106-1115.

[65] Sherlock, G., ***Analysis of large-scale gene expression data***. Curr Opin Immunol, 2000. 12(2): p. 201-5.

[66] Alter, O., P.O. Brown, and D. Botstein, ***Singular value decomposition for genome-wide expression data processing and modeling***. Proc Natl Acad Sci U S A, 2000. 97(18): p. 10101-6.

[67] Brown, M.P.S., et al., ***Knowledge-based analysis of microarray gene expression data by using support vector machines***. PNAS, 2000. 97(1): p. 262-267.

[68] Zhao, L.P., R. Prentice, and L. Breeden, ***Statistical modeling of large microarray data sets to identify stimulus- response profiles***. Proc Natl Acad Sci U S A, 2001. 98(10): p. 5631-6.

[69] Schuchhardt, J., et al., ***Normalization strategies for cDNA microarrays.*** Nucleic Acids Res, 2000. 28(10): p. E47.

[70] Tseng, G.C., et al., ***Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects***. Nucleic Acids Res, 2001. 29(12): p. 2549-57.

[71] Lee, M.L., et al., ***Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations***. Proc Natl Acad Sci U S A, 2000. 97(18): p. 9834-9.

[72] Kerr, M.K. and G.A. Churchill, ***Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments***. Proc Natl Acad Sci U S A, 2001. 98(16): p. 8961-5.

[73] Brazma, A. and J. Vilo, ***Gene expression data analysis***. FEBS Lett, 2000. 480(1): p. 17-24.

[74] Rockett, J.C. and D.J. Dix, ***DNA arrays: technology, options and toxicological applications***. Xenobiotica, 2000. 30(2): p. 155-177.

[75] Schadt, E.E., et al., ***Analyzing high-density oligonucleotide gene expression array data***. J Cell Biochem, 2000. 80(2): p. 192-202.

[76] T. Deepika, R. Porkodi, ***A Survey on Microarray Gene Expression Data sets in Clustering and Visualization Plots***, International Journal of Emerging Research in Management &Technology, ISSN: 2278-9359 (Volume-4, Issue-3), March 2015

[77] Rahila H. Sheikh, M. M. Raghuwanshi, Anil N. Jaiswal, ***Genetic Algorithm Based Clustering: A Survey***, IEEE Computer Society, DOI 10.1109/ICETET.2008.48.

[78] IllhoiYoo, Patricia Alafaireet, MiroslavMarinov, ***Data Mining in Healthcare and Biomedicine: A Survey of the Literature***, Springer Science-Business Media, LLC 2011.

[79] Chloé-Agathe Azencott and Karsten Borgwardt, ***Data Mining in Bioinformatics, Day 7: Clustering in Bioinformatics, Clustering Gene Expression*** Data, February 18 to March 1, 2013, Machine Learning & Computational Biology Research Group, Max Planck Institutes Tübingen and Eberhard Karls Universität Tübingen.

[80] Harun Pirim, BurakEks-io˘glu, Andy D. Perkins, Cetin Yuceer, ***Clustering of high throughput gene expression data***, Computers & Operations Research 39 (2012)3046–3061, journal home page: www.elsevier.com/locate/caor.

[81] Susmita Datta and Somnath Datta, ***Comparisons and validation of statistical clustering techniques for microarray gene expression data***, Vol. 19 no. 4, 2003, pages 459–466, DOI: 10.1093/bioinformatics/btg025.

[82]http://polisci.msu.edu/jacoby/icpsr/regress3/lectures/week4/15.Loess.pdf

[83] Schmid Jr., John (December 1947). "The Relationship between the Coefficient of Correlation and the Angle Included between Regression Lines". *The Journal of Educational Research* **41** (4)

[84] Dudoit, S, Yang, YH, Callow, MJ, Speed, TP. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat. Sin.* 12:1 111-139

[85] Bolstad, B.M., Irizarry R. A., Astrand M., and Speed, T.P. (2003), A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. *Bioinformatics* 19(2):185-193

[86] Rafael. A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs and Terence P. Speed (2003), Summaries of Affymetrix GeneChip probe level data *Nucleic Acids Research* 31(4):e15

[87] Irizarry, RA, Hobbs, B, Collin, F, Beazer-Barclay, YD, Antonellis, KJ, Scherf, U, Speed, TP (2003) Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data.*Biostatistics* .Vol. 4, Number 2: 249-264

[88] Musa H. Asyali, Dilek Colak, Omer Demirkaya and Mehmet S. Inan, Gene Expression Profile Classification: A Review, Current Bioinformatics, 2006, 1, 55-73

[89] Pavlidis P. Using ANOVA for gene selection from microarray studies of the nervous system. Methods 31:282–289, 2003.

[90] Pavlidis P, Li Q, Noble WS. The effect of replication on gene expression microarray experiments. Bioinformatics 19:1620–1627, 2003.

[91] Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., Speed, T. P. (2003). Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. Accepted for publication in Biostatistics.

[92] Cheng W-C, Shu W-Y, Li C-Y, Tsai M-L, Chang C-W, Chen C-R, et al. (2012) Intra- and Inter-Individual Variance of Gene Expression in Clinical Studies. PLoS ONE 7(6): e38650. doi:10.1371/journal.pone.0038650

[93] Huixia Wang, Xuming He1, Mark Band, Carole Wilson and Lei Liu, A study of inter-lab and inter-platform agreement of DNA microarray data, BMC Genomics 2005, 6:71 doi:10.1186/1471-2164-6-71

Πηγές εικόνων

[E.1] http://sociallyconsciousbird.com/storage/hlbio/cellbionotes.htm

[E.2] http://www.signalsblog.ca/cell-lines-patient-samples-and-cultures-oh-my

[E.3] https://en.wiki2.org/wiki/Intron

[E.4] http://www.nucleotidenutrition.com/nucell-im-featured-in-the-article-nucleotides-the-building-blocks-of-life/

[E.5]https://el.wikipedia.org/wiki/%CE%9A%CE%B1%CF%81%CE%BA%CE%AF%CE%BD%CE%BF%CF%82

[E.6]http://www.bioscientist.com/news/%CE%BA%CF%85%CE%BA%CE%BB%CE%BF%CF%86%CE%BF%CF%81%CE%BF%CF%8D%CE%BD%CF%84%CE%B1-%CE%BA%CE%B1%CF%81%CE%BA%CE%B9%CE%BD%CE%B9%CE%BA%CE%AC-%CE%BA%CF%8D%CF%84%CF%84%CE%B1%CF%81%CE%B1/

[E.7] http://core.montana.edu/index58ad.html?page=versarray

[E.8] http://www.tolpa.com/infographics.php

[E.9] http://seedgenenetwork.net/annotate

[E.10] https://en.wikipedia.org/wiki/Heat_map

Παραρτήματα

Το παράρτημα στην παρούσα εργασία χωρίζεται σε δύο ενότητες στο Παράρτημα Α και το Παράρτημα Β. Το παράρτημα Α , αφορά τα αποτελέσματα των δεδομένων ως προς τις κανονικοποιήσεις που εφαρμόστηκαν σε αυτά. Ενώ το Παράρτημα Β αφορά κυρίως την κατηγοριοποίηση των Δεδομένων.

Πιο αναλυτικά το Παράρτημα Α , περιλαμβάνει 3 υποενότητες ,το Α.1, Α.2 , Α.3.

- Στο παράρτημα Α.1 φαίνονται οι γραφικές παραστάσεις των MvA plots για όλα τα δείγματα των Κυτταρικών Σειρών (Reference plus Query Dataset) πρίν και μετά την κανονικοποίηση LOWESS. Υπάρχουν 31 MvA plots , και για το κάθε ένα αναφέρεται σε ποιο δείγμα και σε ποια κυτταρική σειρά ανήκει.
- Στο παράρτημα Α.2 υπάρχουν οι πίνακες με τους συντελεστές συσχέτισης για όλα τα δείγματα των κυτταρικών σειρών, πριν και μετά την κανονικοποίηση LOWESS. Συνολικά φαίνονται 5 πίνακας , ένας πίνακας για κάθε Κυτταρική Σειρά.
- Στο παράρτημα Α.3 βλέπουμε 20 γραφικές παραστάσεις. Είναι Οι γραφικές παραστάσεις 10 δειγμάτων της Κυτταρικής σειράς 4 πριν και μετά την Lowess. Έτσι λοιπόν βλέπουμε 10 γραφήματα πριν την κανονικοποίηση Lowess , για τα 10 δείγματα της Σειράς και 10 δείγματα μετά την κανονικοποίηση Lowess , πάλι για τα αντίστοιχα δείγματα. Σε κάθε διάγραμμα αναγράφεται το δείγμα και το bulk . Στην συγκεκριμένη Σειρά για όλα τα δείγματα το unamplified control είναι το MCF7 (καρκίνος του μαστού)

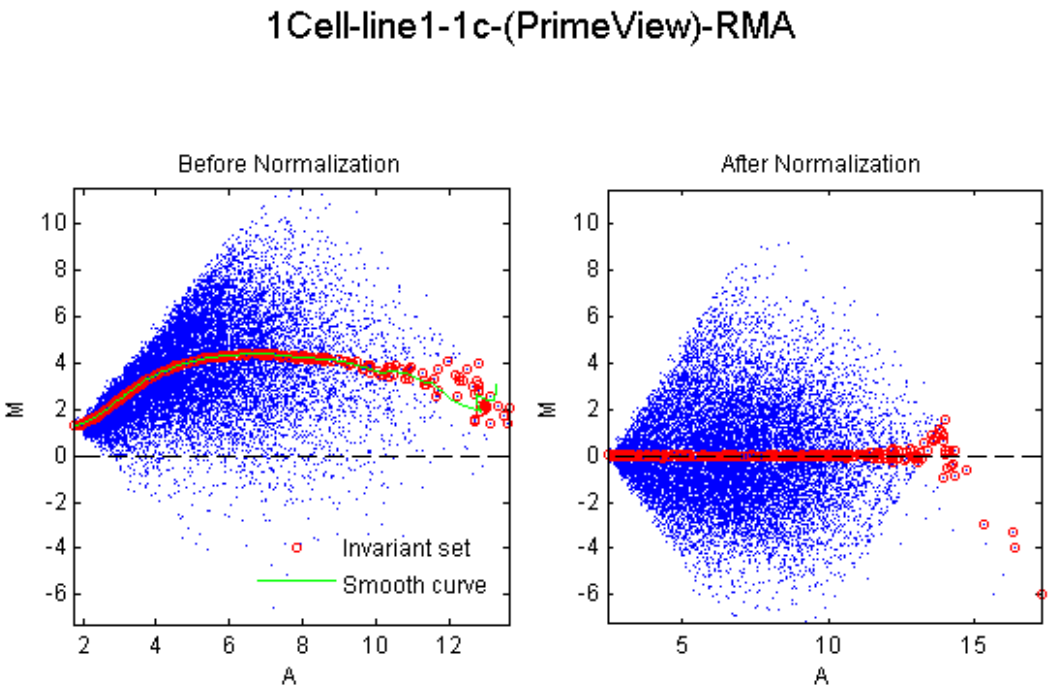
Το Παράτημα Β περιλαμβάνει κάποιες από τις δοκιμές που κάναμε κατά την διαδικασία του Clustering των δεδομένων μας. Τα αποτελέσματα αυτών των δοκιμών απεικονίζονται μέσω των Heat Maps. Είναι σημαντικό να τονιστεί ότι δεν είναι εύκολο να αναπαρασταθεί το Clustering τόσων εκατοντάδων δειγμάτων και γονιδίων που χρησιμοποιήσαμε. Για αυτό το λόγο για κάθε δοκιμή παραθέτουμε μία εικόνα που περιλαμβάνει όλο το Heat map και μετά στιγμιότυπα με σημεία όπου βλέπουμε το clustering των δεδομένων. Σε κάθε περίπτωση οι εικόνες δεν μπορούν να αποτελέσουν κριτήριο για το αν ένα clustering , δίνει σωστά αποτελέσματα. Ο σωστός έλεγχος της κατηγοριοποίησης μπορεί να γίνει μόνο μέσα από το πρόγραμμα όπου εξάγονται τα Heat maps. Ωστόσο οι απεικονίσεις αυτές είναι απαραίτητες προκειμένου να αποδειχτεί η ευστάθεια των αποτελεσμάτων μας.

Παράρτημα Α

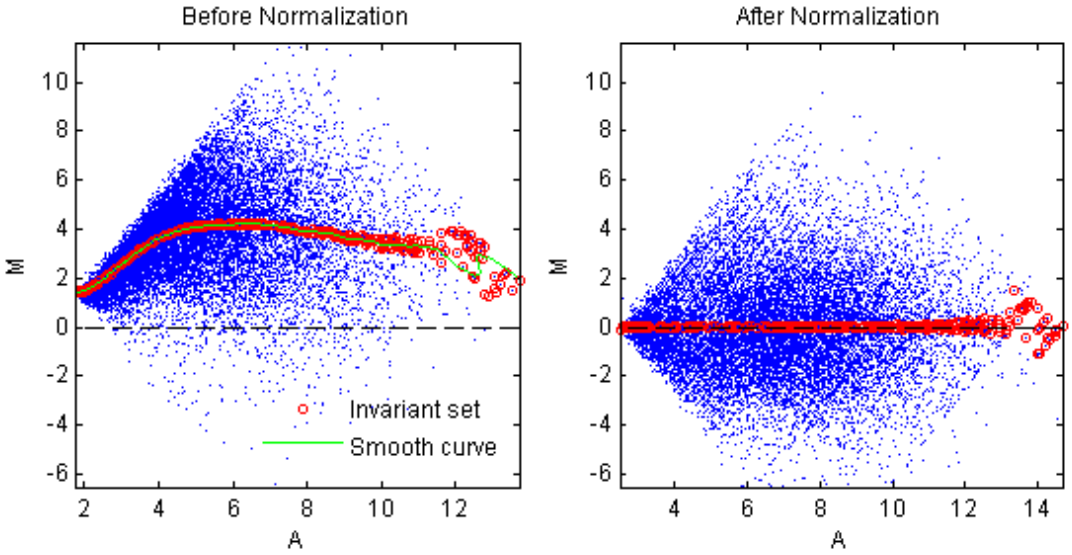
Α.1 - Οι γραφικές παραστάσεις των MvA plots για όλα τα δείγματα των κυτταρικών σειρών, πριν και μετά την κανονικοποίηση LOWESS.

Σε κάθε γράφημα αναφέρεται το δείγμα, η κατάληξη (PrimeView)– Rma αν αναγράφεται , υποδηλώνει ότι είναι της σειράς PrimeView και το δείγμα έχει περάσει από τον αλγόριθμο Rma. Σε περιπτώσεις που δεν αναγράφεται στον τίτλο και πάλι να θεωρείται ότι είναι της σειράς PrimeView και έχει περάσει από Rma.

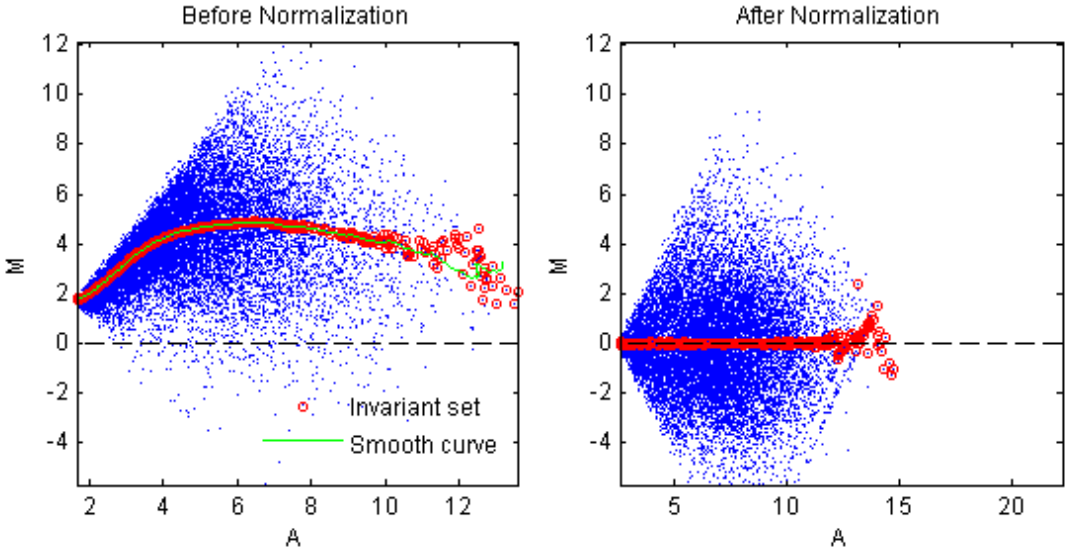
Για την Cell line 1:



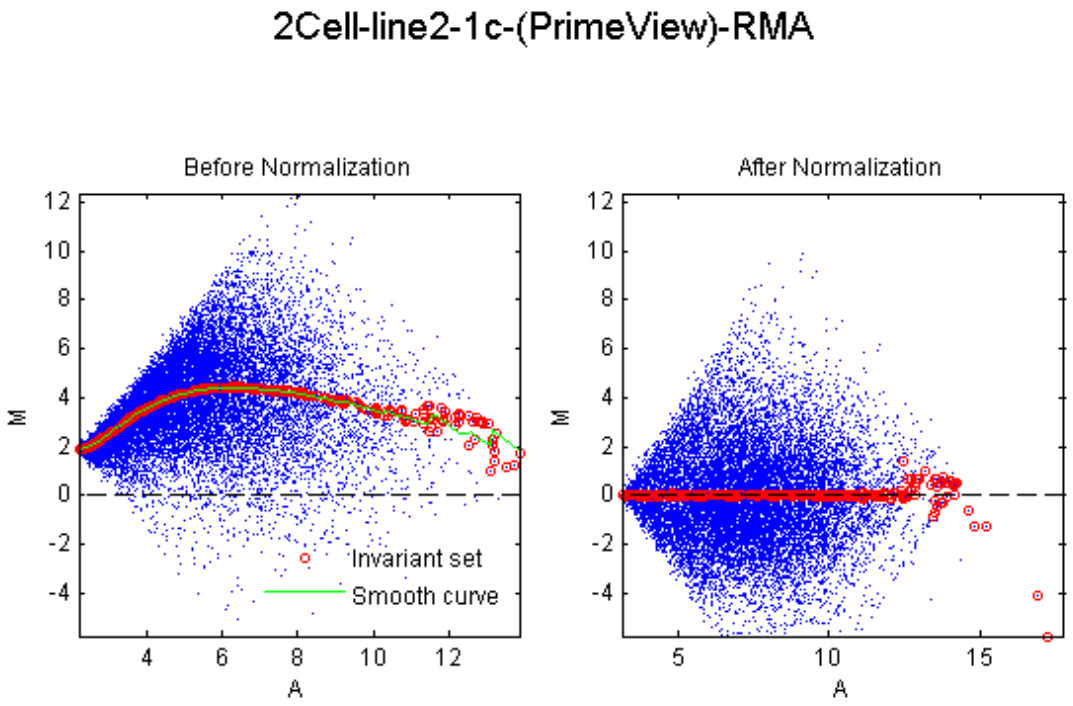
2Cell-line1-1c-(PrimeView)-RMA

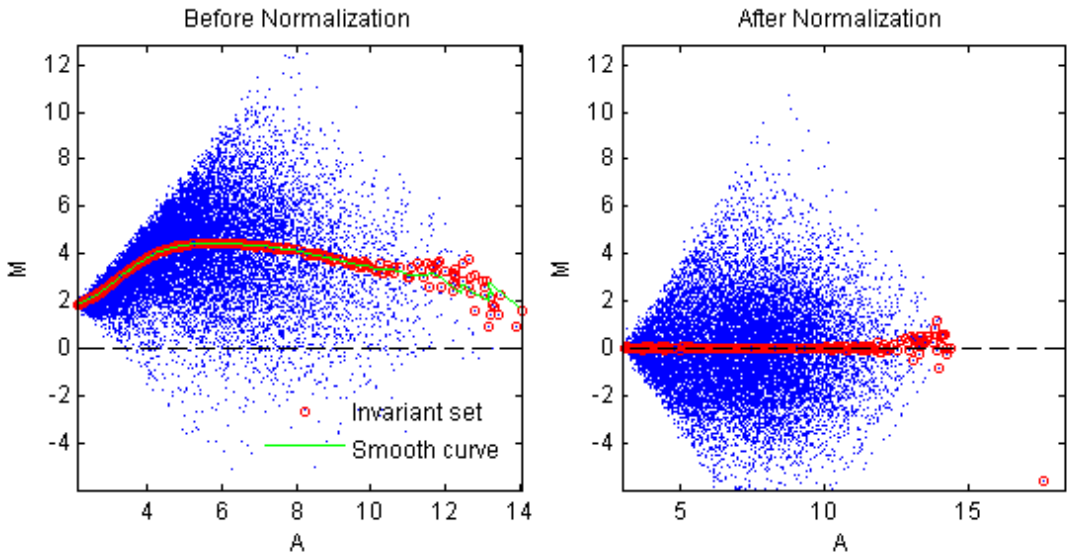


3Cell-line1-1c-(PrimeView)-RMA

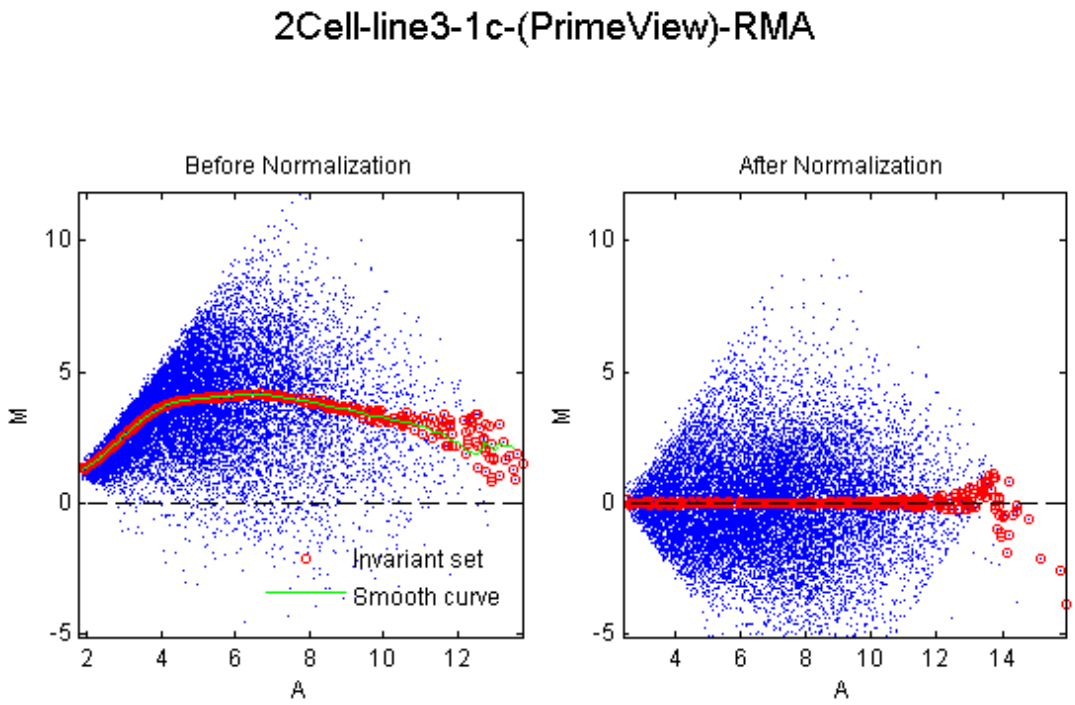


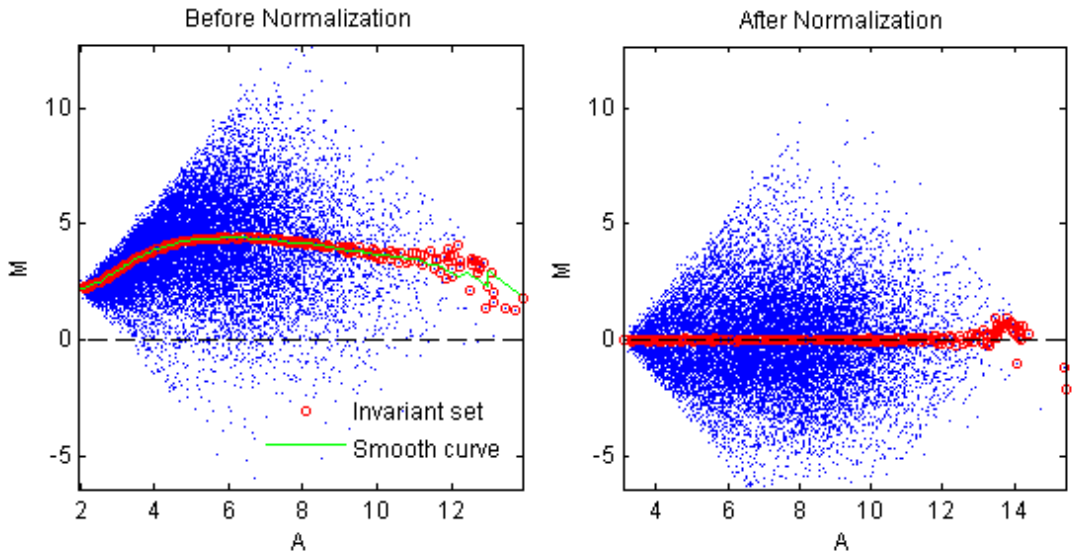
Για την Cell line 2:



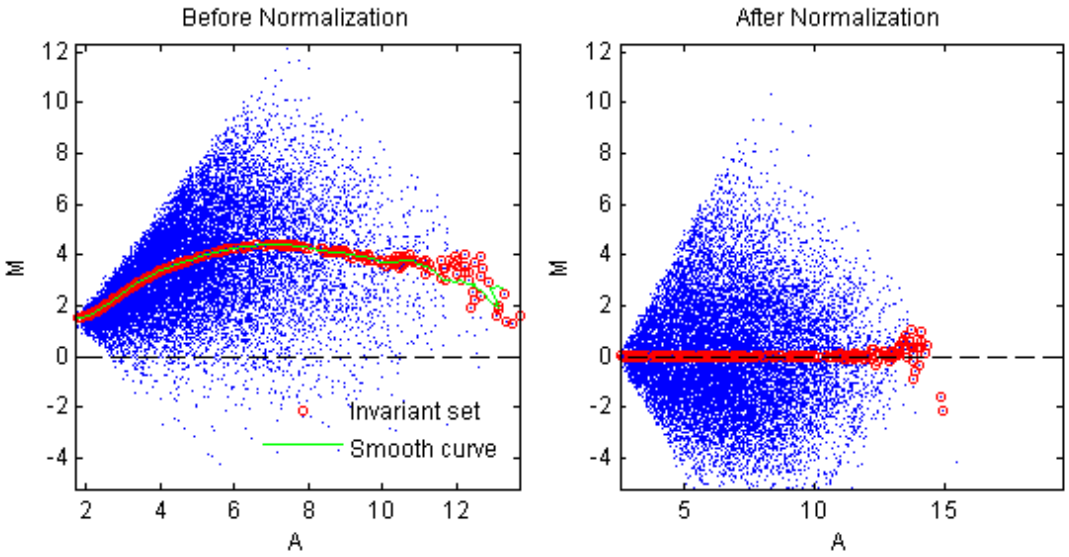


Για την Cell line 3:

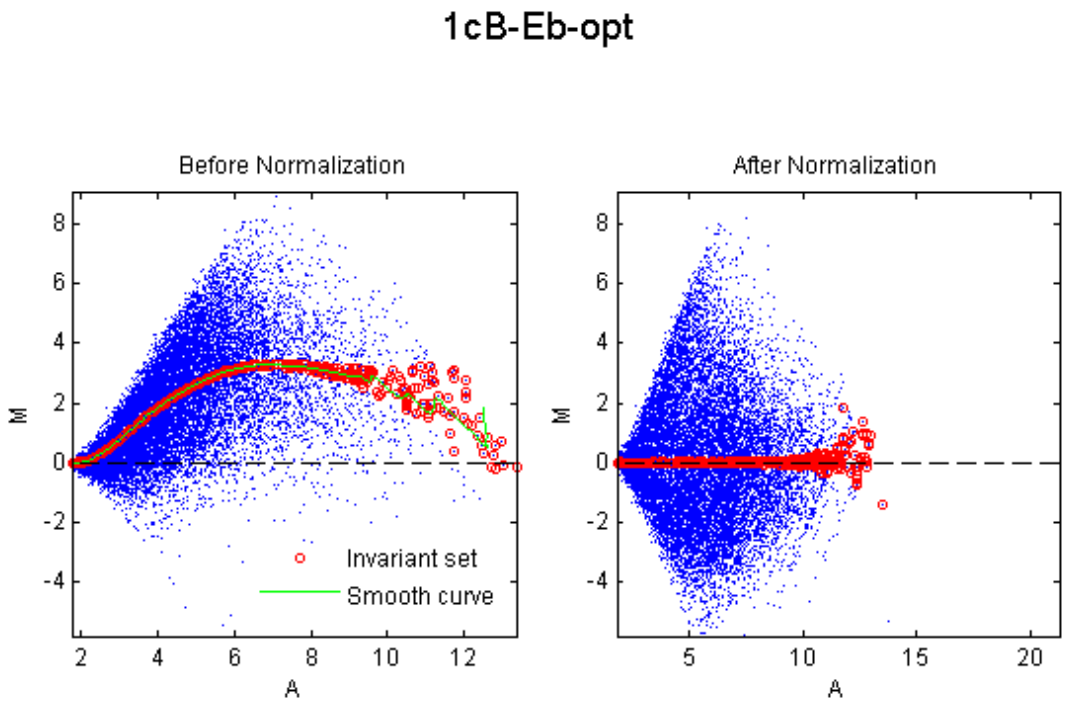


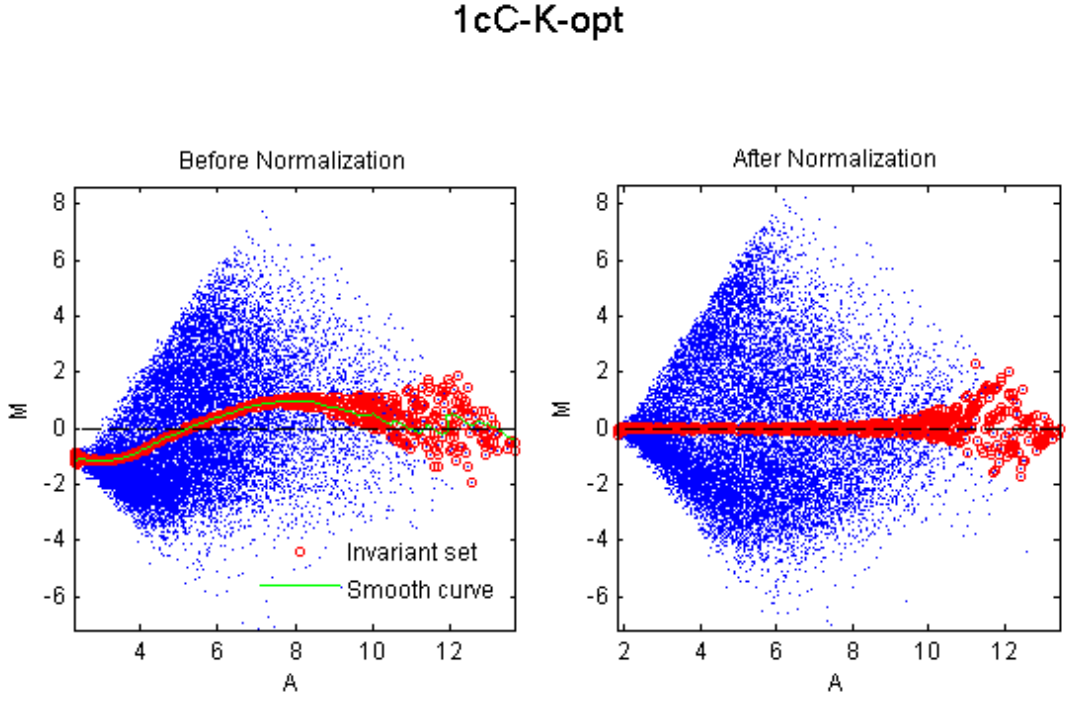


3Cell-line3-1c-(PrimeView)-RMA

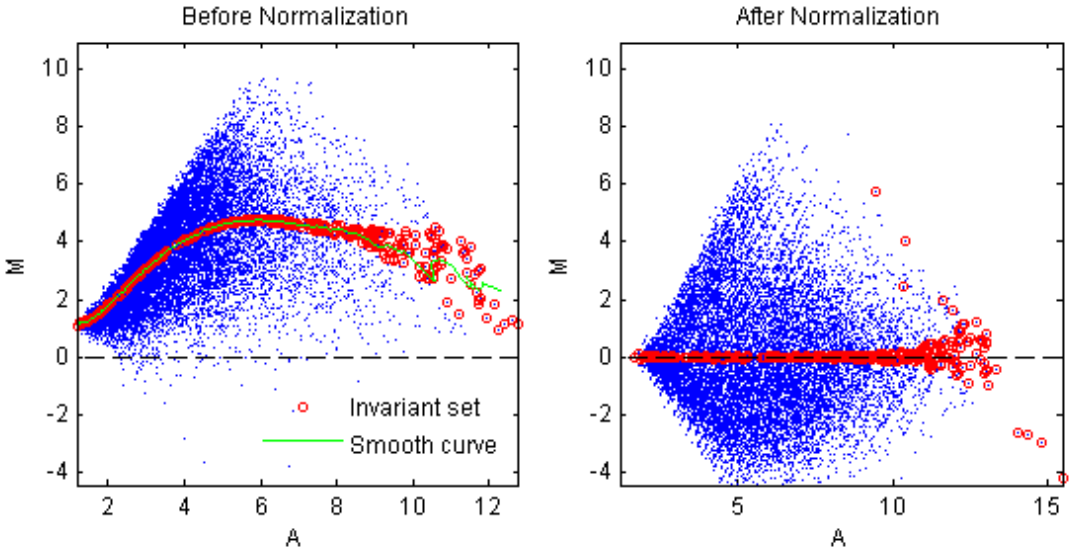


Για την Cell line 4 :

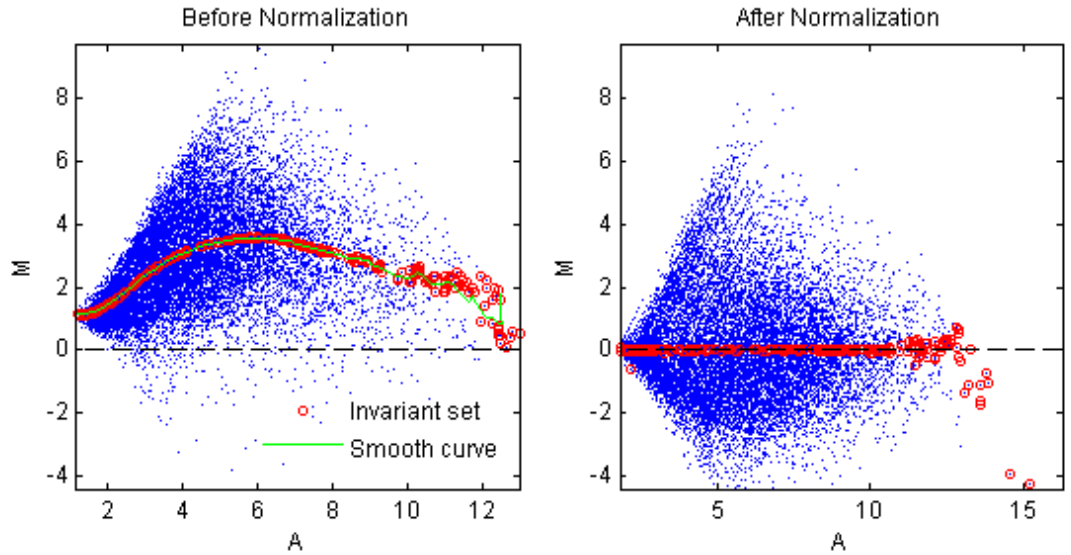




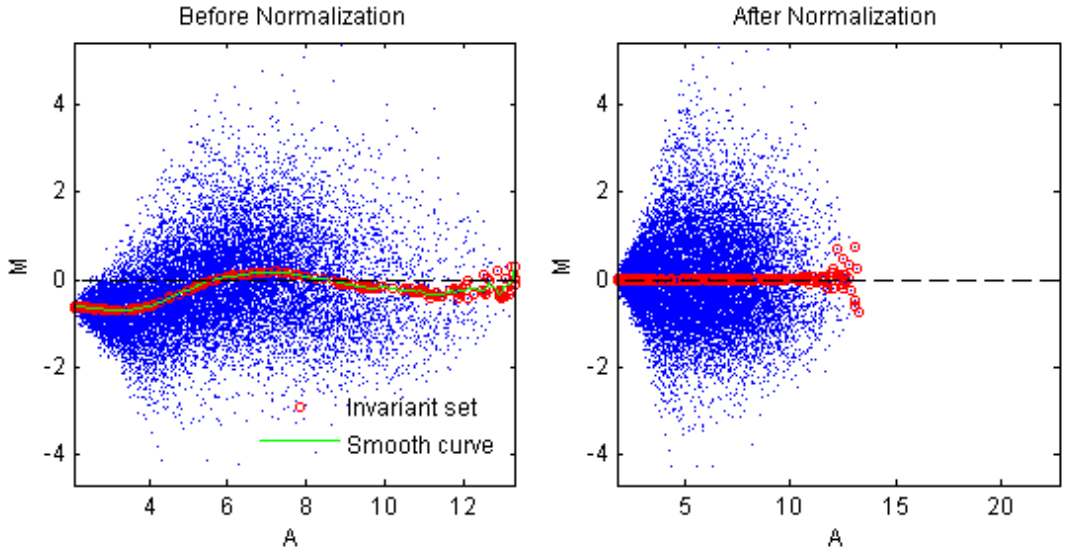
1Cell-line4-1c



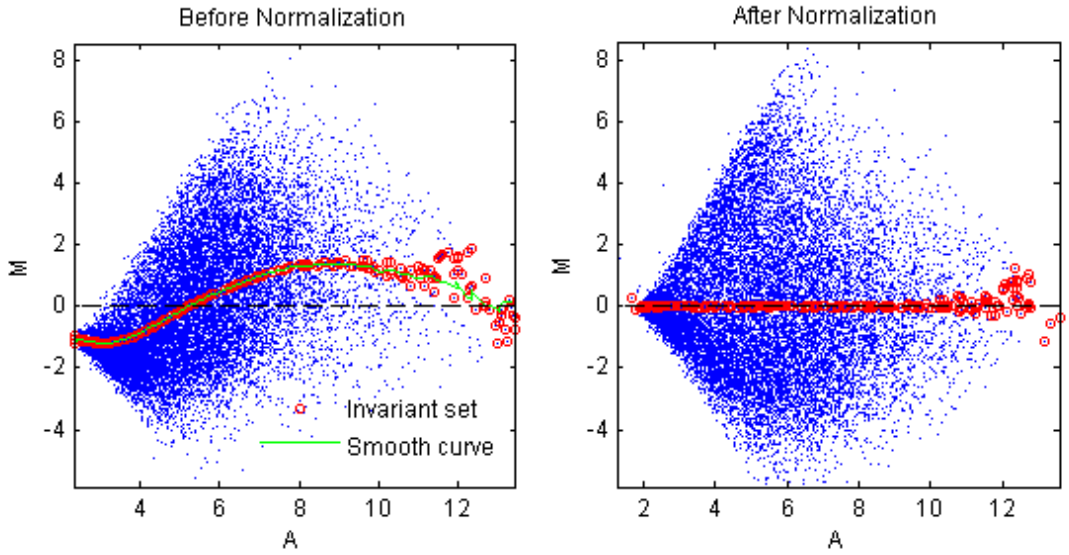
2Cell-line4-1c



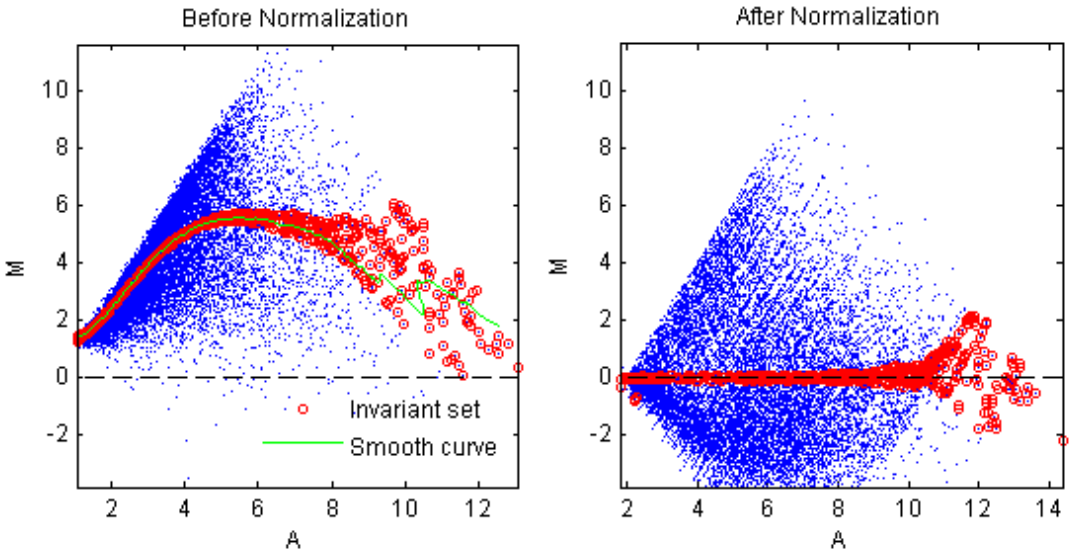
5000c-Eb-opt



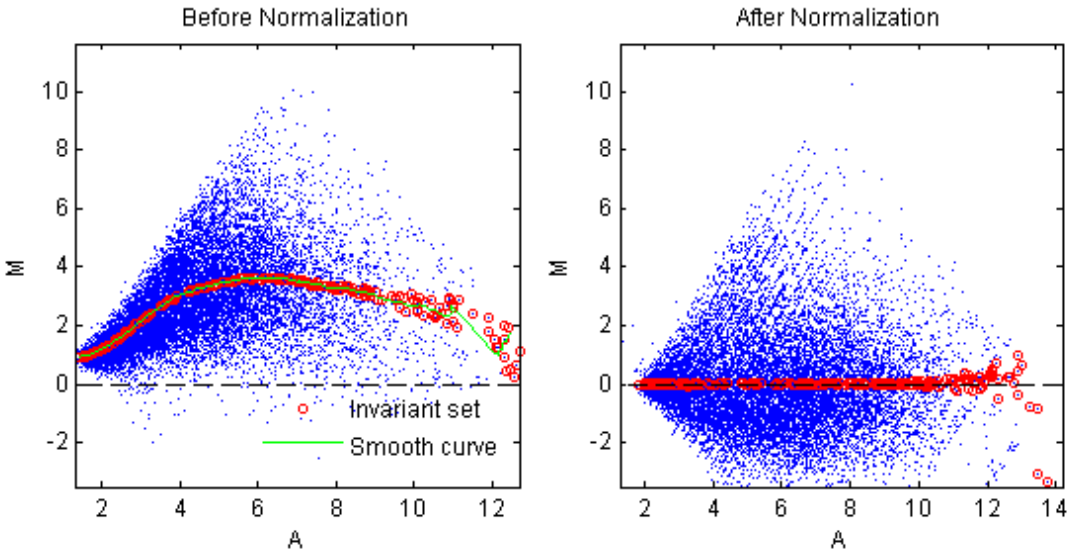
5000c-K-OPT



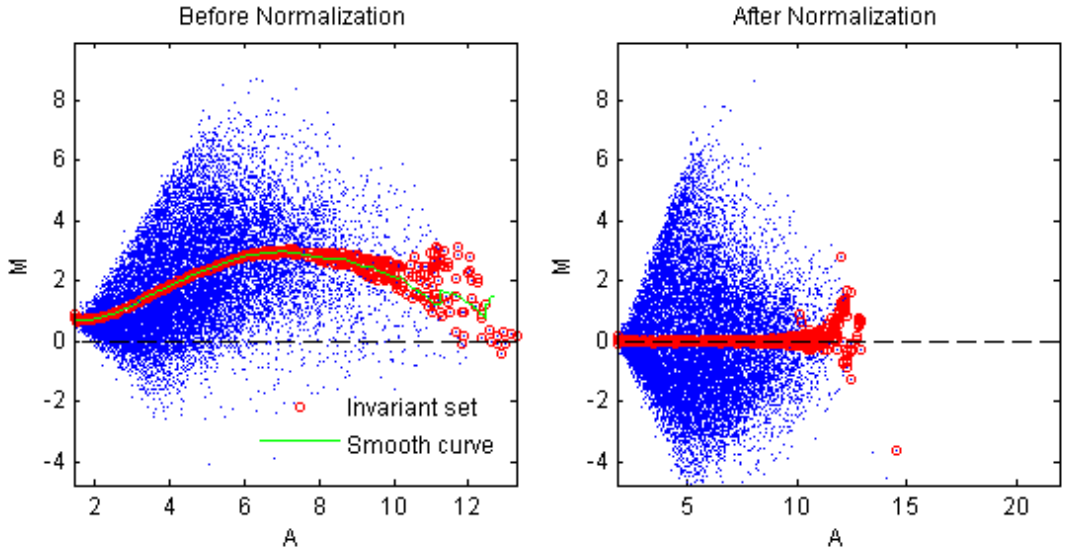
Cell-line4-40c



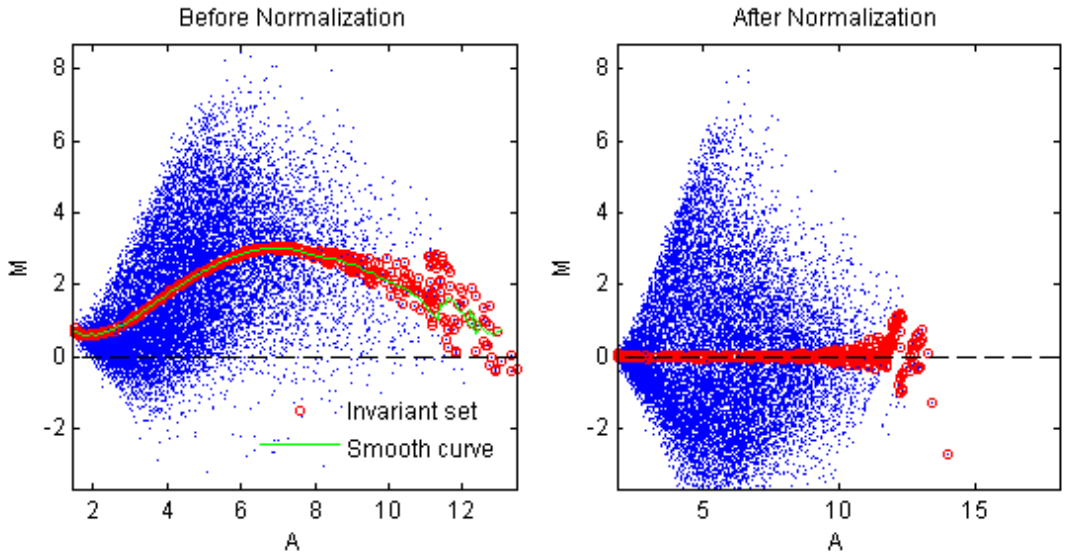
Cell-4line4-1000c



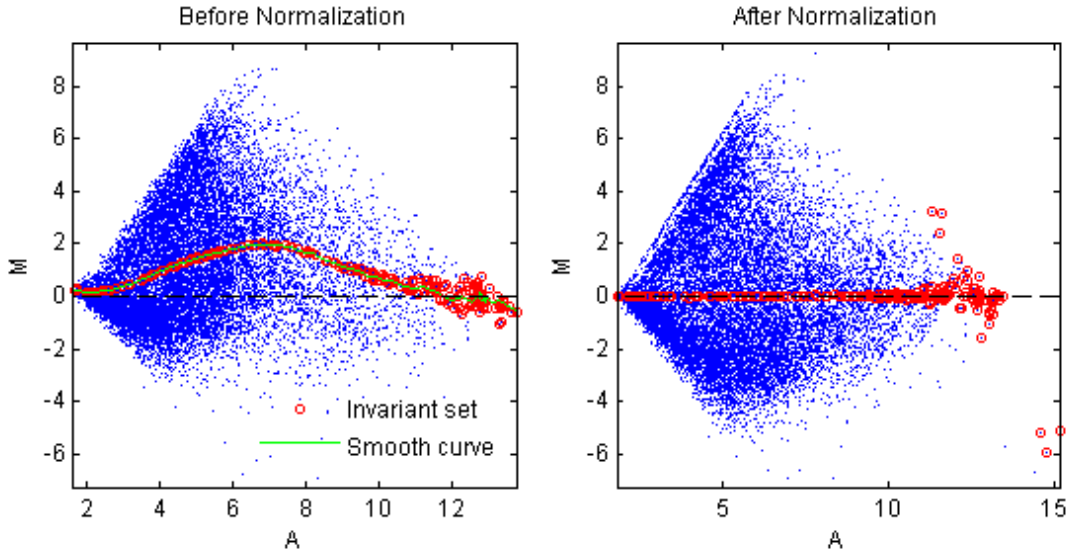
1c-Eb-Optimized-Eberwine



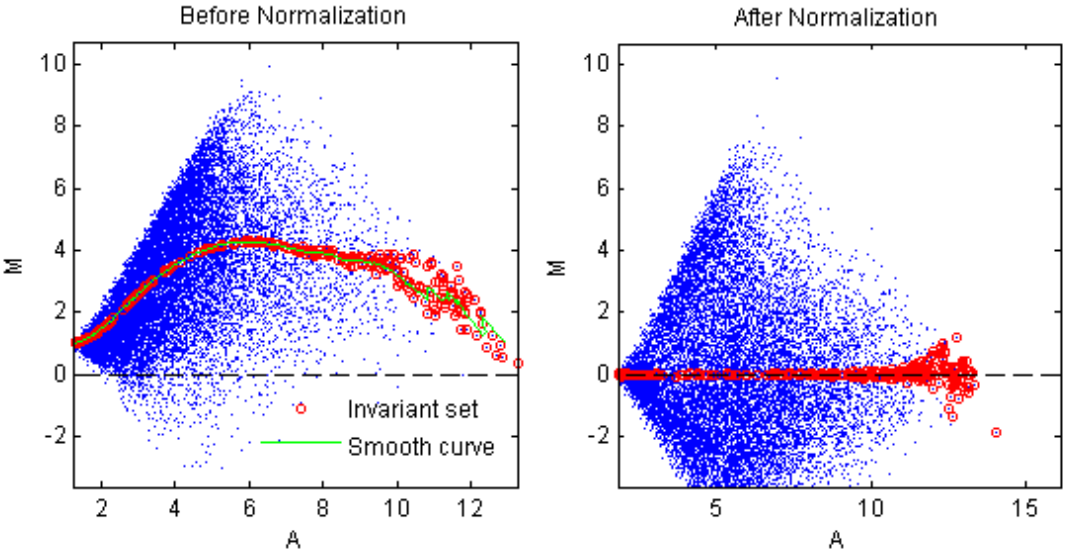
1c_Eb-RNA-opt_Eberwine_



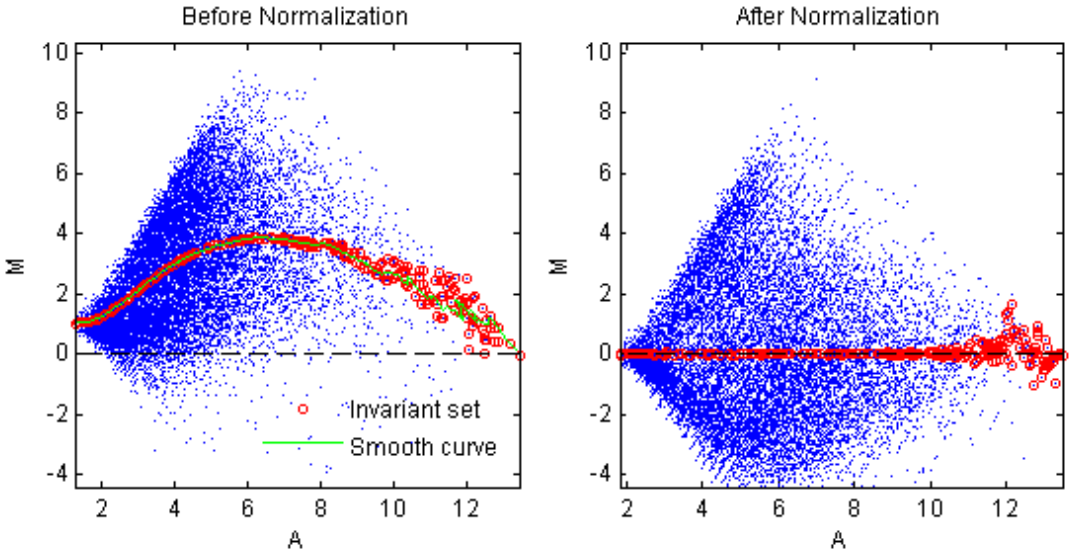
1c-K-ExTaq-SSIII-1cell



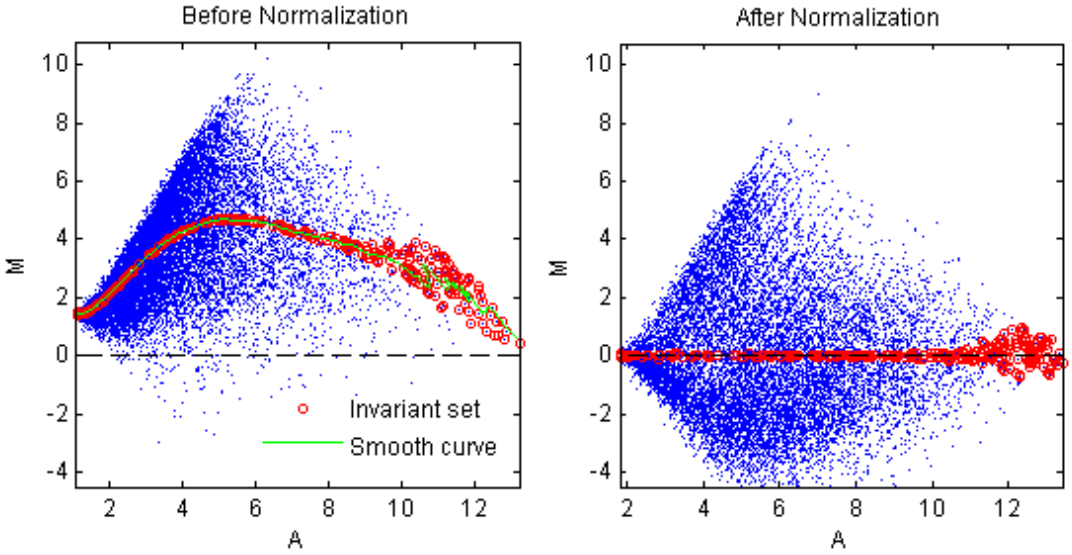
1c-K-opt-A2-Sm-1cell



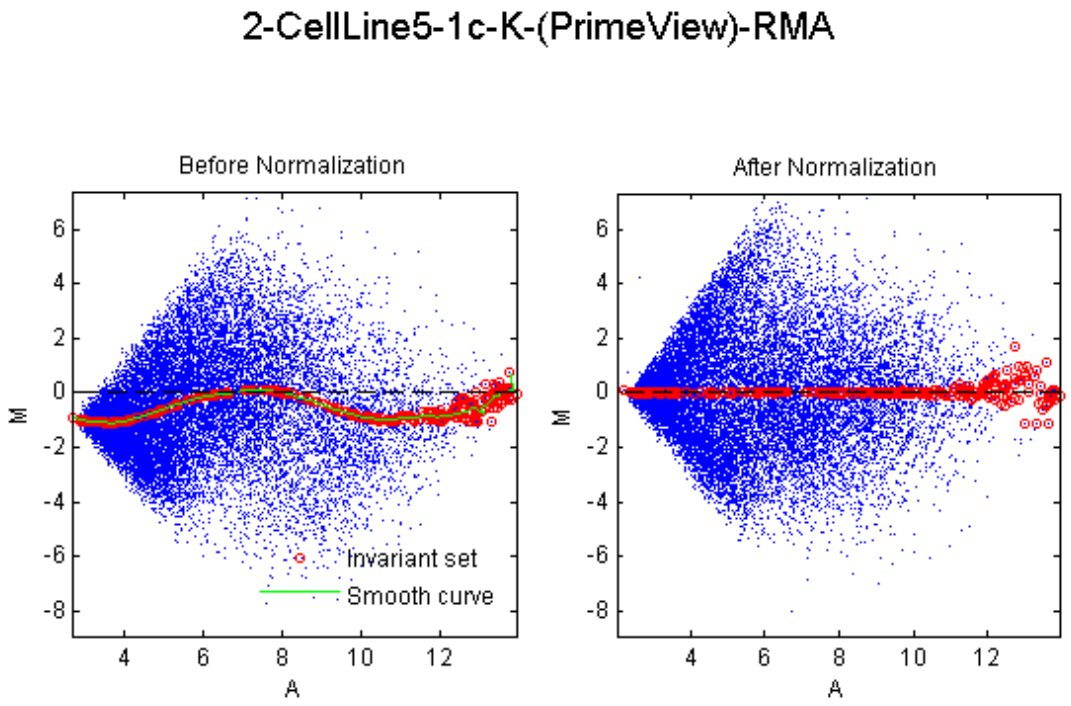
1c-K-RNA-opt-D2-A2-RNA

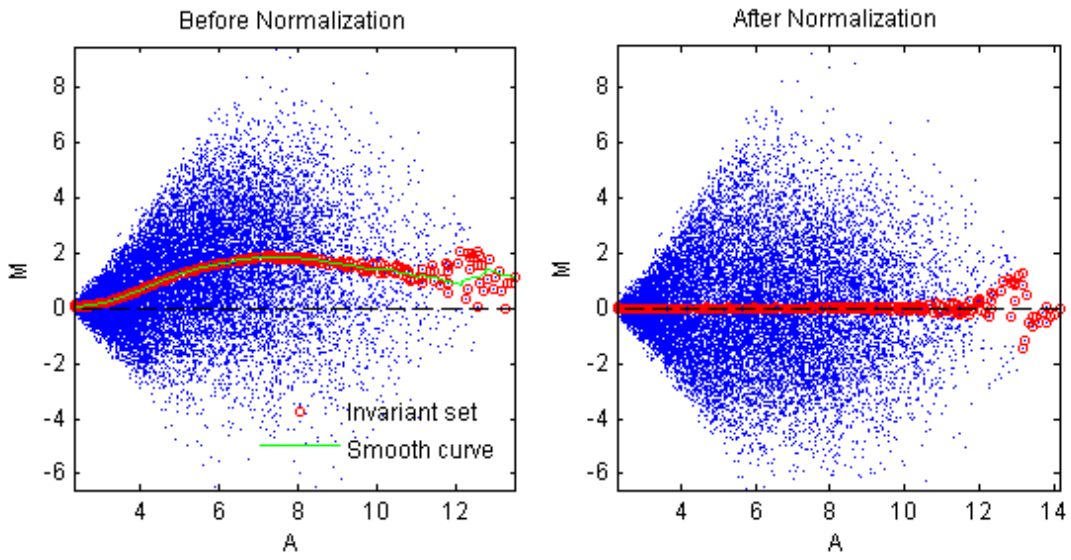


1c-K-RNA-opt-Prime-A2-RNA

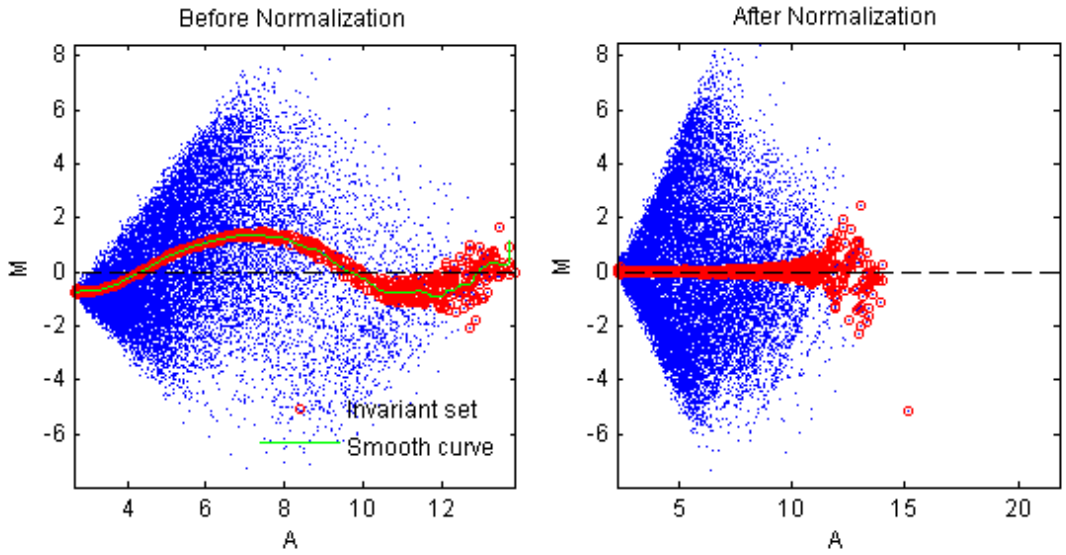


Για την Cell line 5:

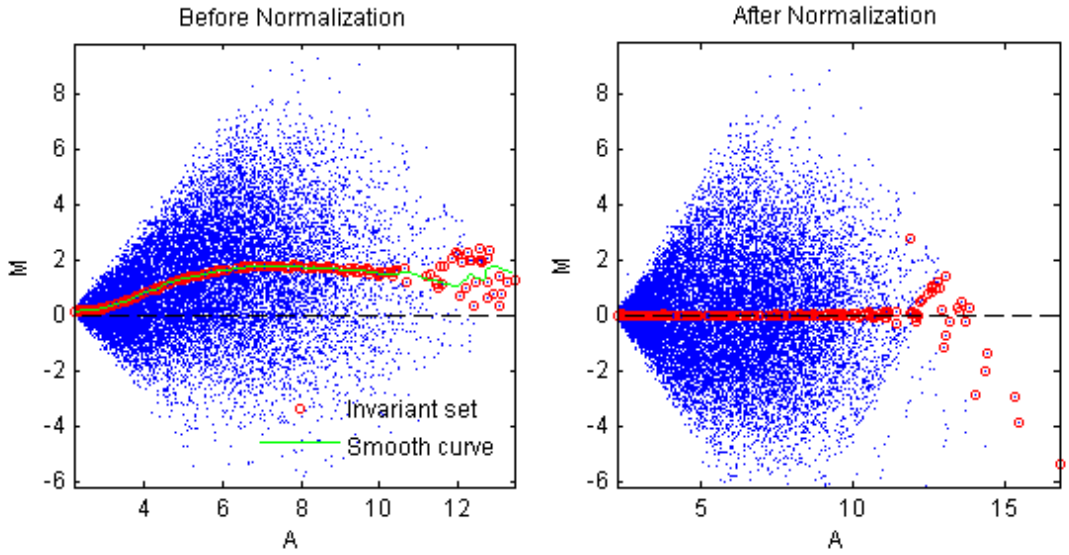




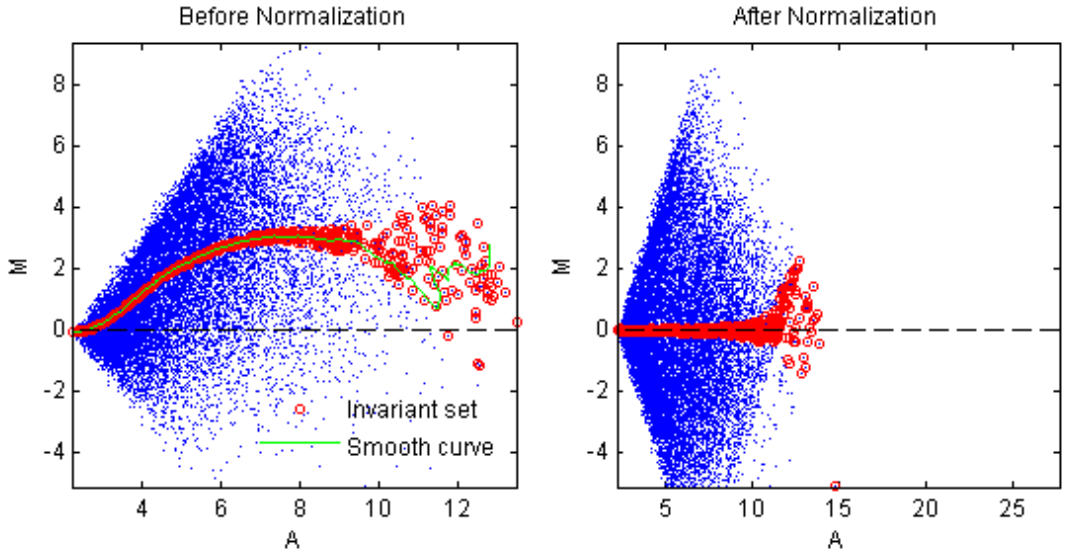
3-CellLine5-1c-K-(PrimeView)-RMA



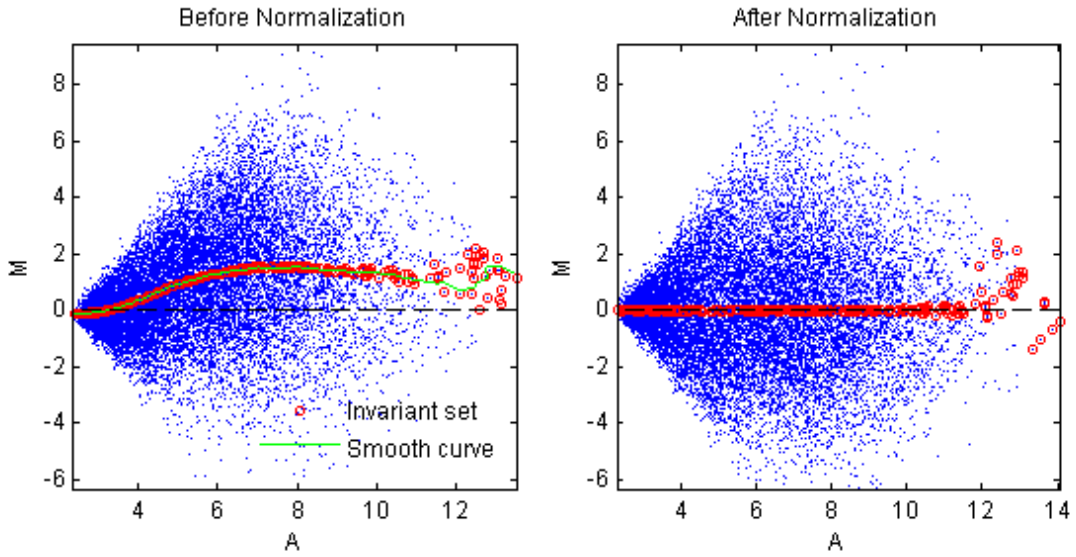
CellLine5-40c-Eb-(PrimeView)-RMA



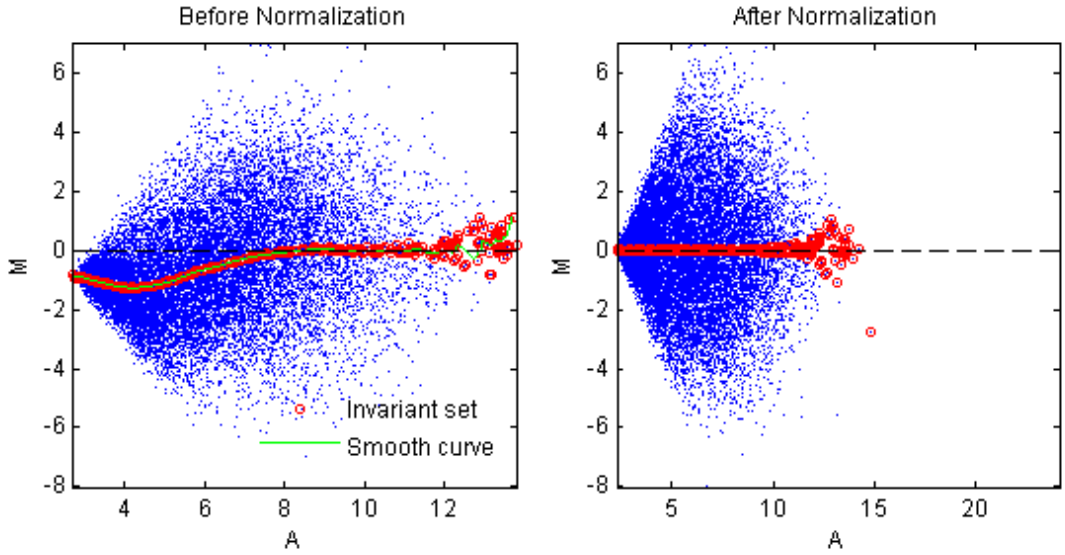
CellLine5-40c-K-(PrimeView)-RMA



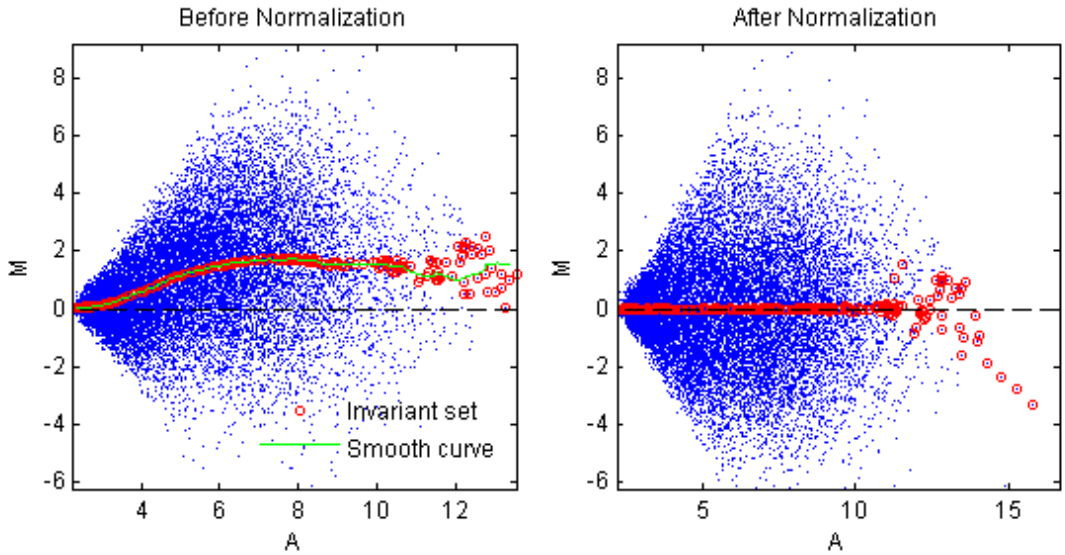
CellLine5-1000c-Eb-(PrimeView)-RMA



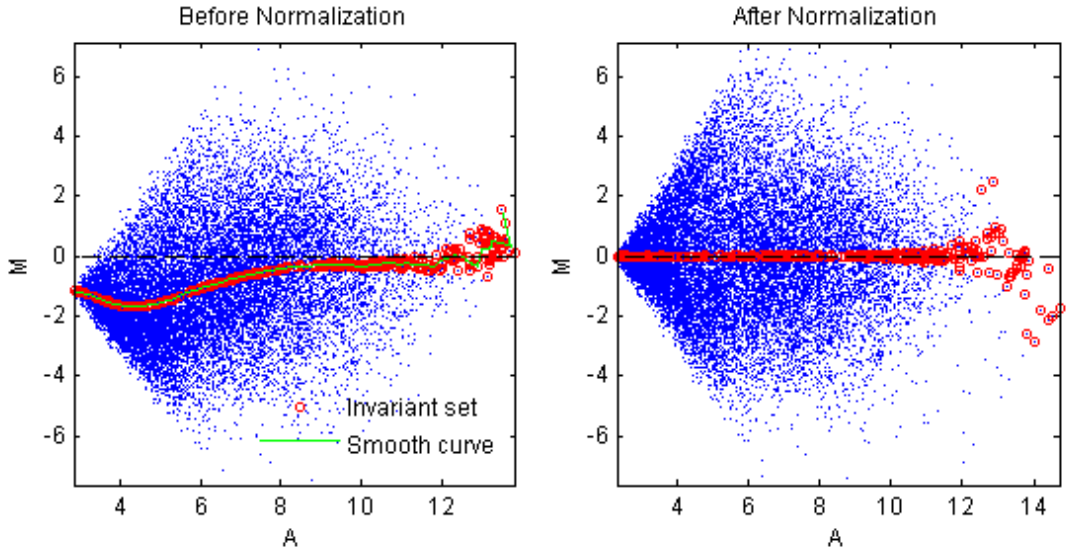
CellLine5-1000c-K-(PrimeView)-RMA



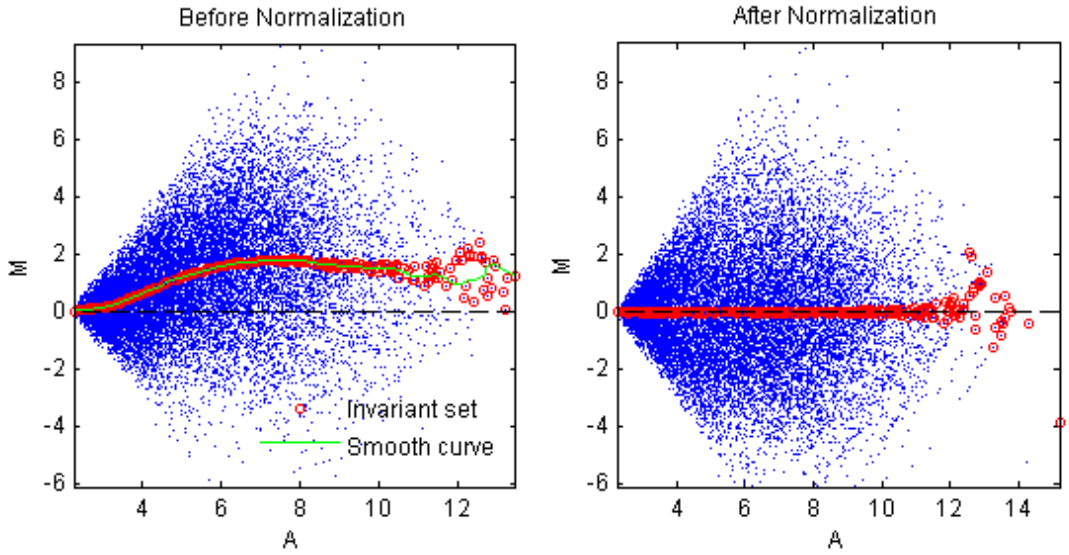
1CellLine5-1c-Eb-(PrimeView)-RMA



1CellLine5-1c-K-(PrimeView)-RMA



2-CellLine5-1c-Eb-(PrimeView)-RMA



Α.2 - Οι πίνακες με τους συντελεστές συσχέτισης για όλα τα δείγματα των κυτταρικών σειρών, πριν και μετά την κανονικοποίηση LOWESS.

Cell line 1:

Δείγμα	Τιμή Correlation Coefficient πρίν την Lowess	Τιμή Correlation Coefficient μετά την Lowess , με παράθυρο 5%	Τιμή Correlation Coefficient μετά την Lowess , με παράθυρο 6%
1Cell_line1_1c	0.637	0.6442	0.6482
2Cell_line1_1c	0.6734	0.703	0.7032
3Cell_line1_1c	**0.6603**	**0.0706**	**0.679**

Cell line 2:

Δείγμα	Τιμή Correlation Coefficient πρίν την Lowess	Τιμή Correlation Coefficient μετά την Lowess , με παράθυρο 5%	Τιμή Correlation Coefficient μετά την Lowess , με παράθυρο 6%
1Cell_line2_1c	0.6271	0.6116	0.6446
2Cell_line2_1c	0.623	0.6346	0.6398
3Cell_line2_1c	0.6236	0.6539	0.0818

Cell line 3:

Δείγμα	Τιμή Correlation Coefficient πρίν την Lowess	Τιμή Correlation Coefficient μετά την Lowess , με παράθυρο 5%
1Cell_line3_1c	**0.7052**	**0.0491**
2Cell_line3_1c	0.6028	0.7375
3Cell_line3_1c	0.6351	0.6342

Για την Cell line 5:

Δείγμα	Τιμή Correlation Coefficient πρίν την Lowess	Τιμή Correlation Coefficient μετά την Lowess , με παράθυρο 5%
1CellLine5_1c_Eb	0,695986042684879	0,705759276190332
1CellLine5_1c_K	0,697611684512344	0.694701426645003
2_CellLine5_1c_Eb	0,693352169655251	0,706827914132275
2_CellLine5_1c_K	0,662406785275256	0,650507036410551
3_CellLine5_1c_Eb	0,677711226193218	0,688789291002639
3_CellLine5_1c_K	0,543876012508878	0,498937114518333
CellLine5_1000c_Eb	0,677828056453185	0,691098063457023
CellLine5_1000c_K	0,715047492280746	0,699295538246883
CellLine5_40c_Eb	0,681462934227343	0,692851467860049
CellLine5_40c_K	0,500905378987856	0,463422790787202

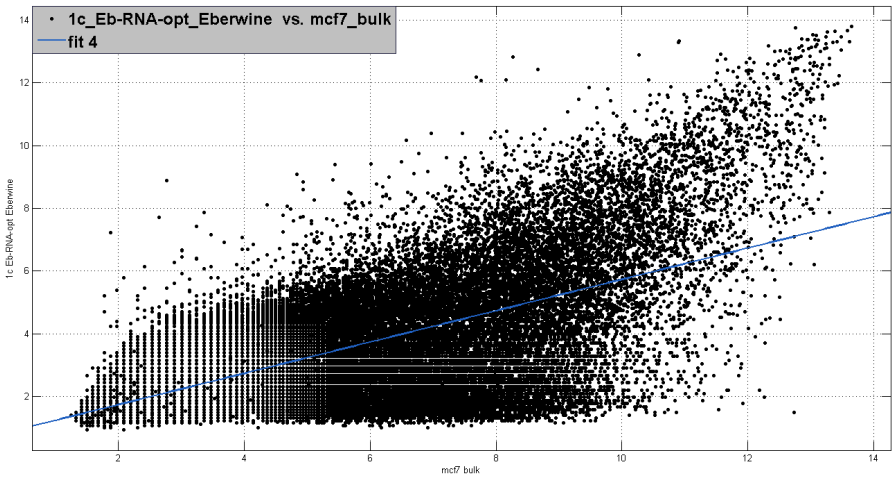
Για την Cell Line 4:

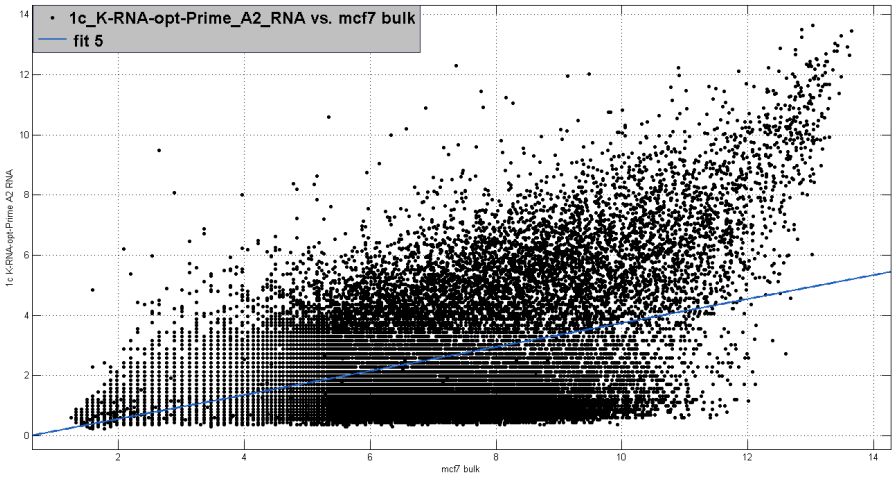
Δείγμα	Τιμή Correlation Coefficient πρίν την Lowess	Τιμή Correlation Coefficient μετά την Lowess , με παράθυρο 5%
1Cell_line4_1c	0.6618	0.5735
1cB-Eb-opt	0.6541	0.6377
1cC_K-opt	0.5532	-0.011
1c_Eb-Optimized_Eberwine	0.72	0.0755
1c_Eb-RNA-opt_Eberwine	0.703	0.6912
1c_K-RNA-opt-Prime_A2_RNA	0.6129	0.5841
1c_K-RNA-opt_D2_A2_RNA	0.583	0.5541
1c_K-opt_A2_Sm_1cell	0.5688	0.5322
1c_K_ExTaq_SSIII_1cell	0.5561	0.514
2Cell_line4_1c	0.7624	0.4542
5000c_Eb-opt	0.916	0.9105
5000c_K-OPT	0.5797	0.553
Cell_line4_1000c	0.7726	0.7933
Cell_line4_40c	0.5473	0.2872

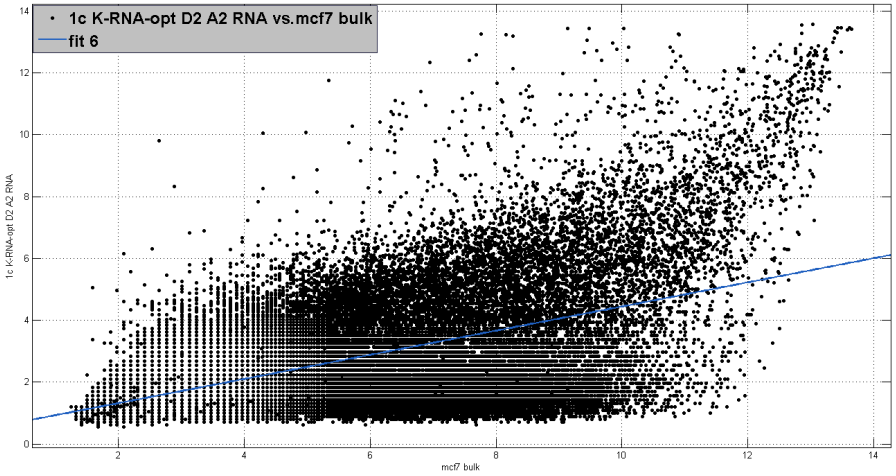
Α.3 – Οι γραφικές παραστάσεις 10 δειγμάτων της Κυτταρικής σειράς 4 πριν και μετά την Lowess.

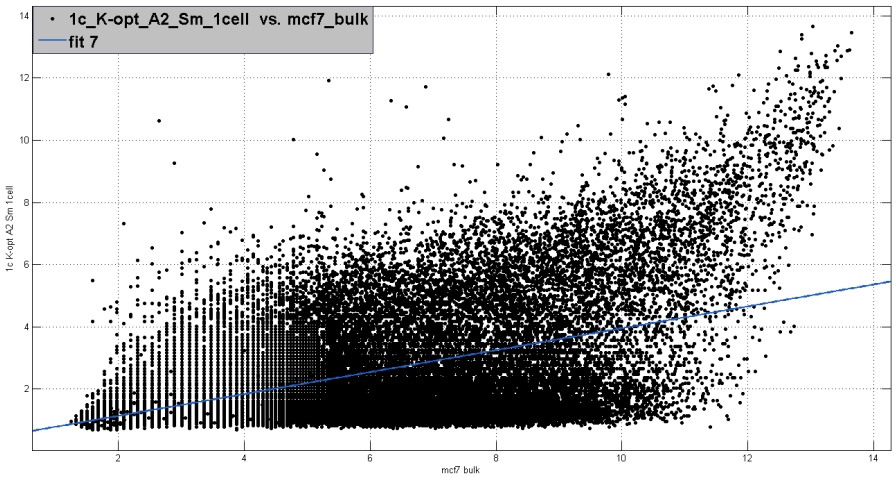
Πριν την κανονικοποίηση Lowess:

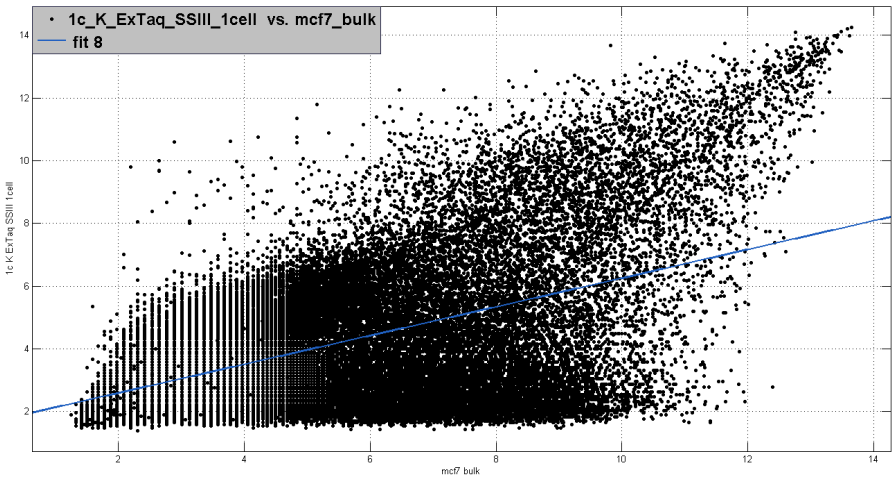
Σε κάθε διάγραμμα αναγράφεται το δείγμα και το bulk , για όλα είναι το MCF7.

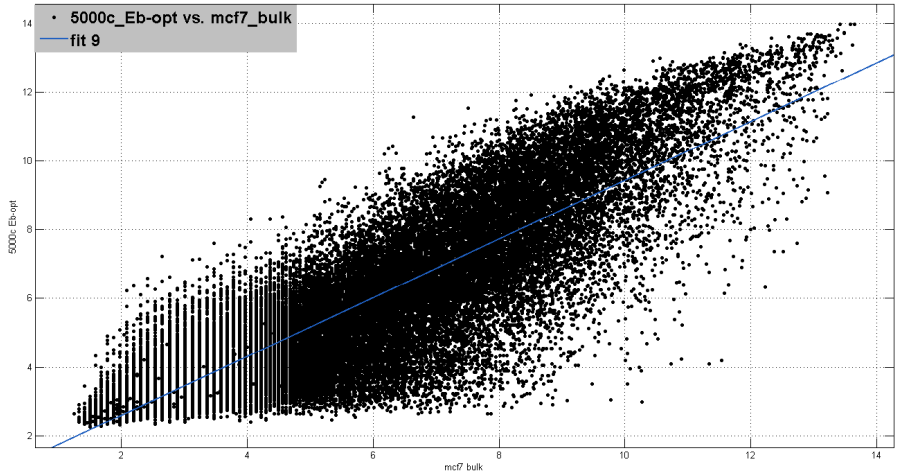


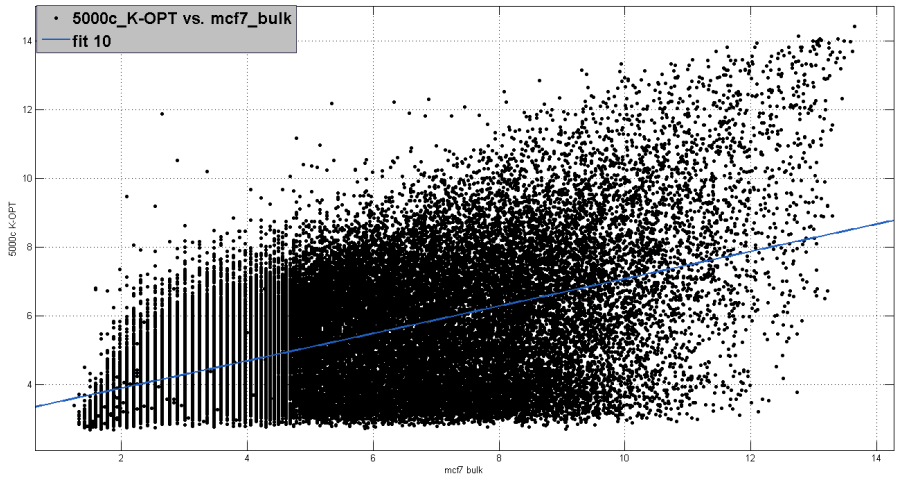


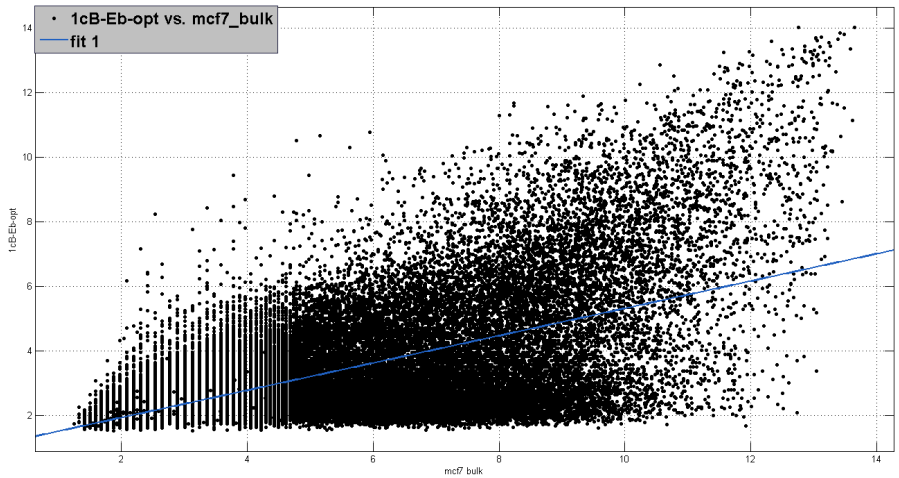

1c K-RNA-opt D2 A2 RNA vs.mcf7 bulk
fit 6
1c K-RNA-opt D2 A2 RNA
mcf7 bulk

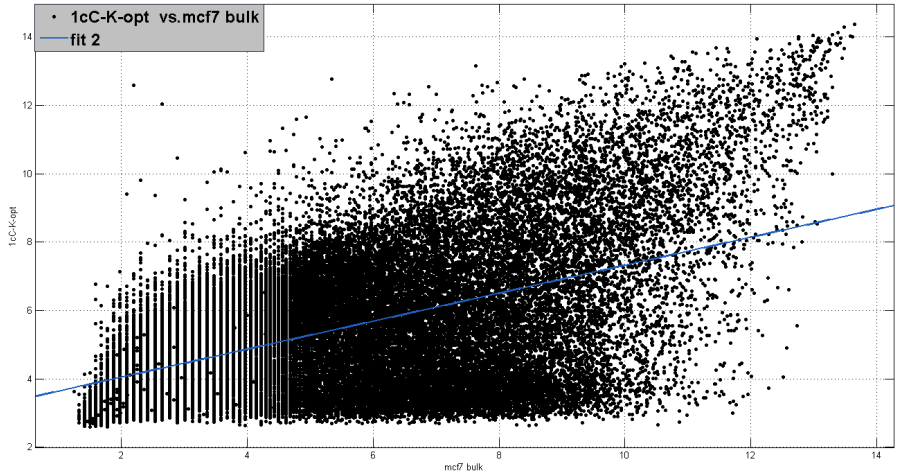

1c_K-opt_A2_Sm_1cell vs. mcf7_bulk
fit 7
1c K-opt A2 Sm 1cell
mcf7 bulk

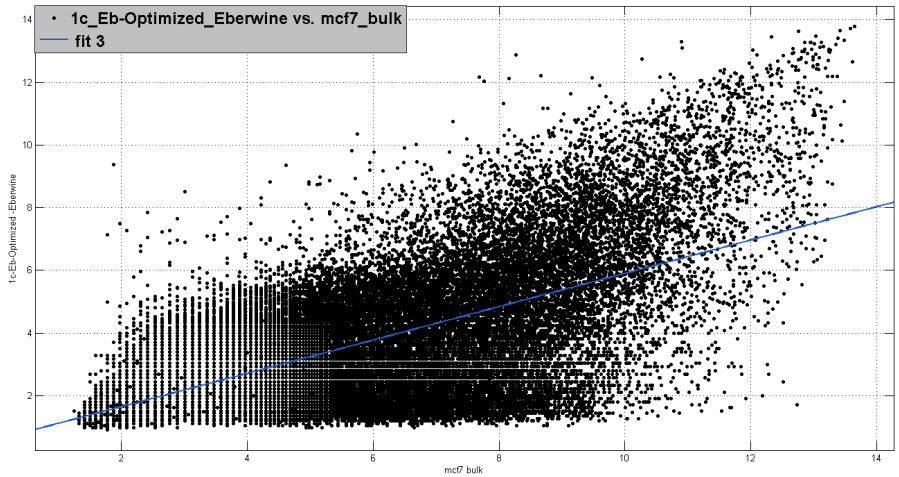

1c_K_ExTaq_SSIII_1cell vs. mcf7_bulk
fit 8
1c K ExTaq SSIII 1cell
mcf7 bulk


5000c_Eb-opt vs. mcf7_bulk
fit 9
5000c Eb-opt
mcf7 bulk

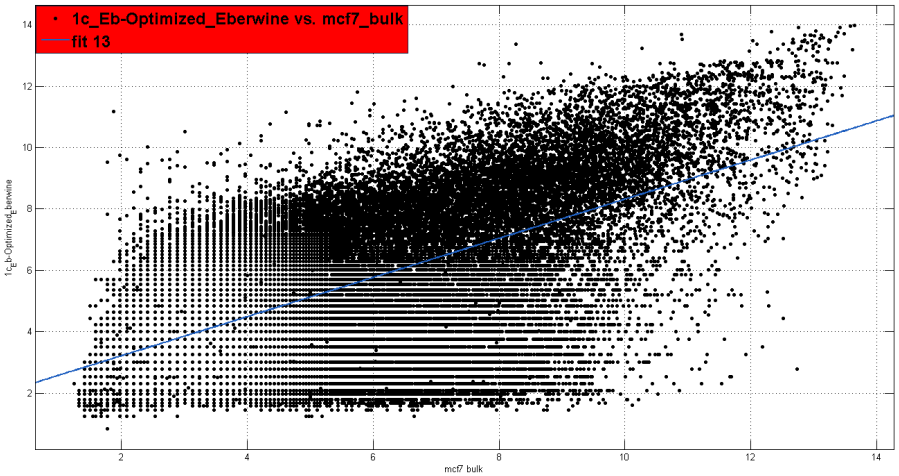

5000c_K-OPT vs. mcf7_bulk
fit 10
5000c K-OPT
mcf7 bulk

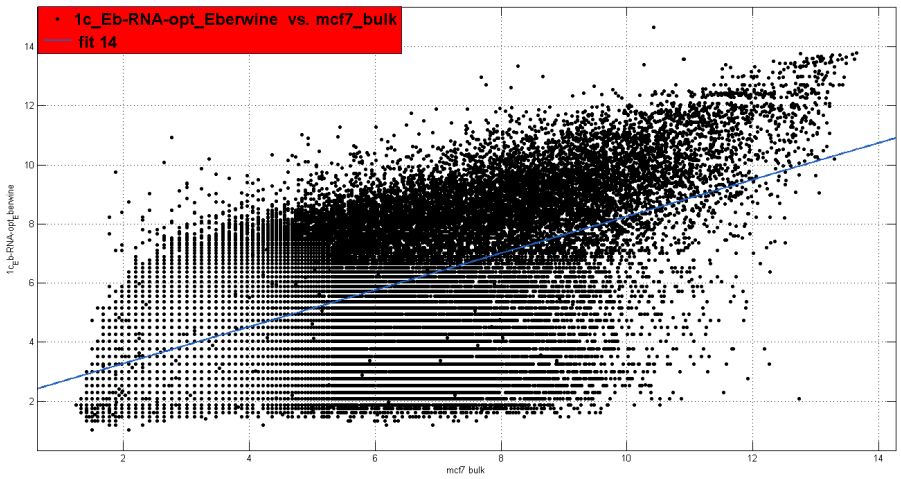

1cB-Eb-opt vs. mcf7_bulk
fit 1
1cB-Eb-opt
mcf7 bulk

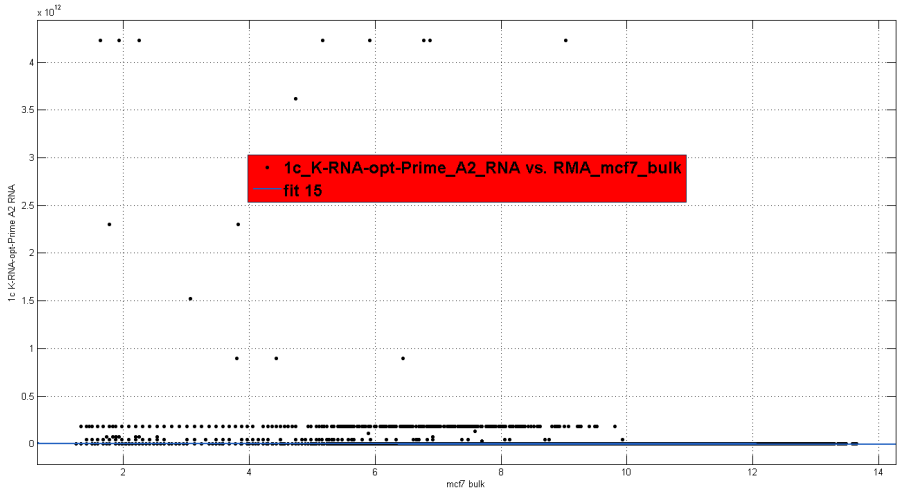


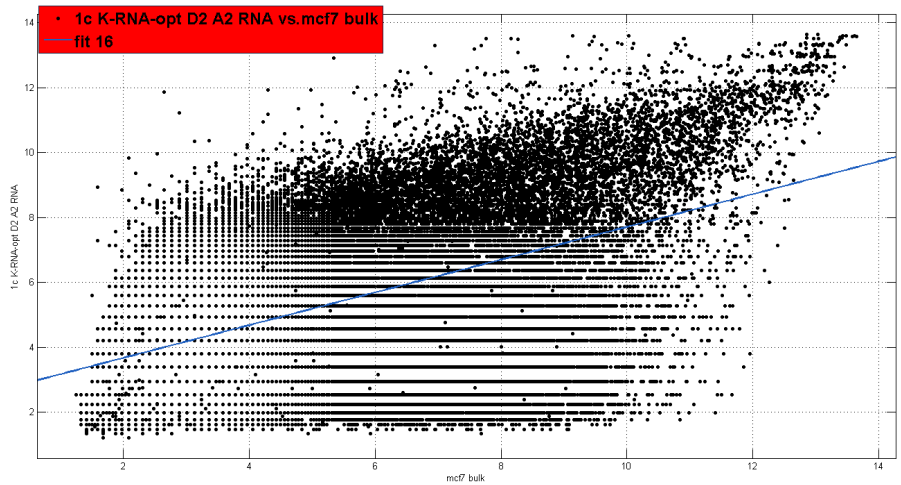


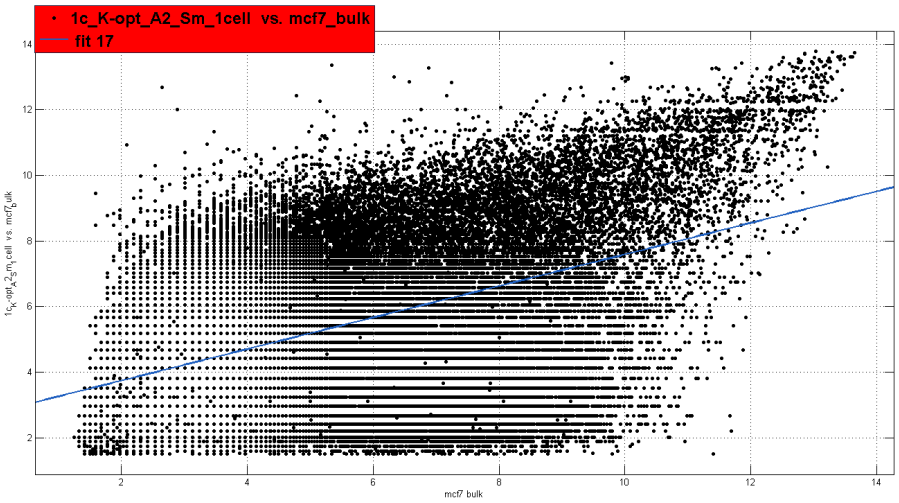
Μετά την κανονικοποίηση Lowess:

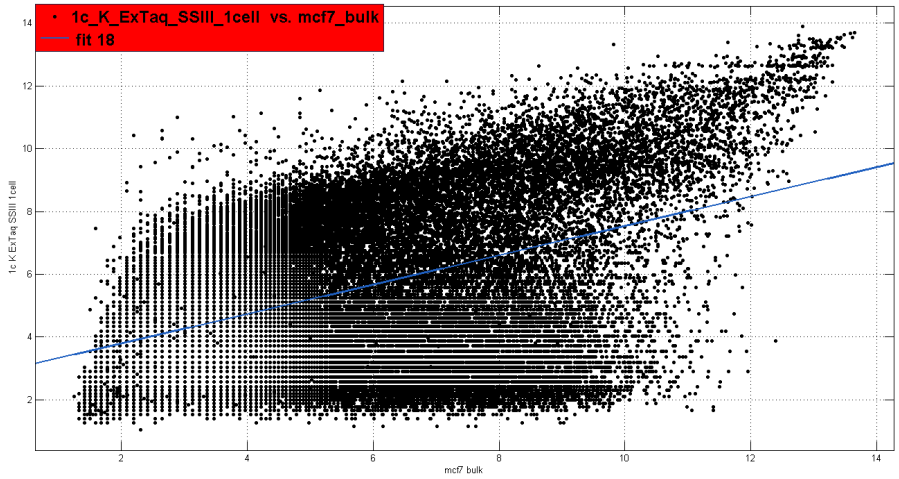



1c_Eb-RNA-opt_Eberwine vs. mcf7_bulk
fit 14
1c_Eb-RNA-opt_Eberwine
mcf7 bulk

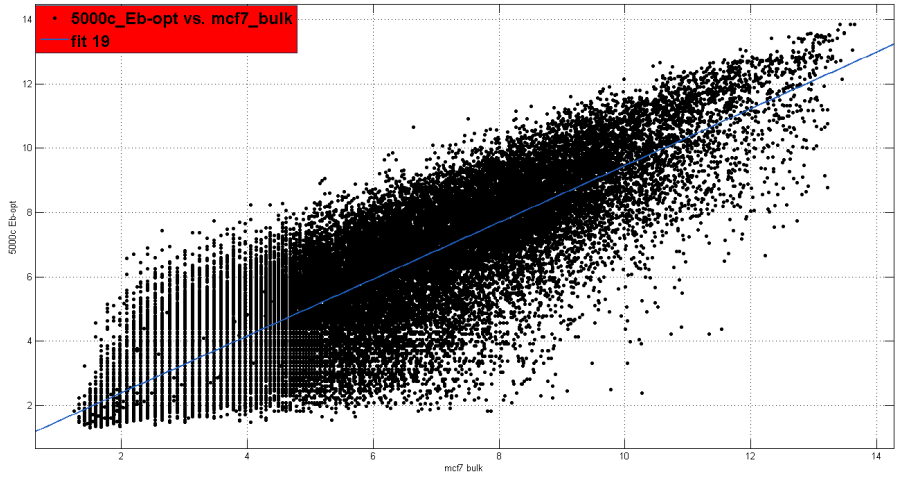

x 10^12
1c_K-RNA-opt-Prime_A2_RNA vs. RMA_mcf7_bulk
fit 15
1c K-RNA-opt-Prime A2 RNA
mcf7 bulk

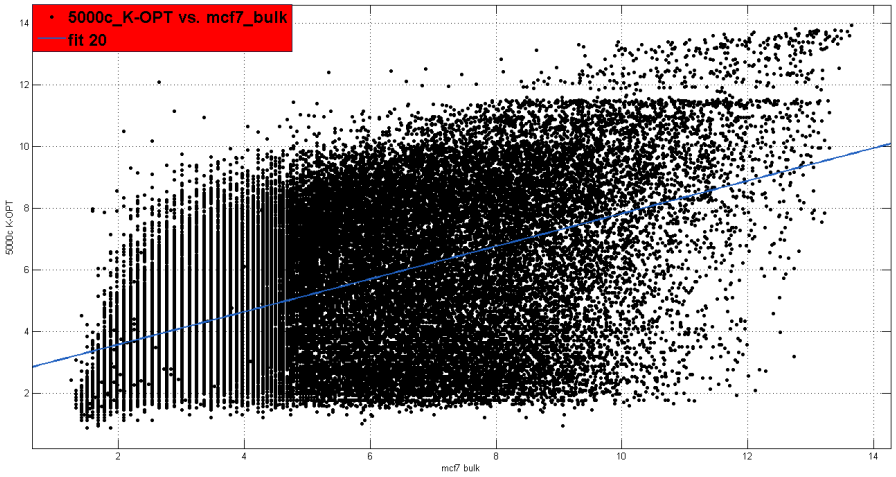

1c K-RNA-opt D2 A2 RNA vs.mcf7 bulk
fit 16
1c K-RNA-opt D2 A2 RNA
mcf7 bulk

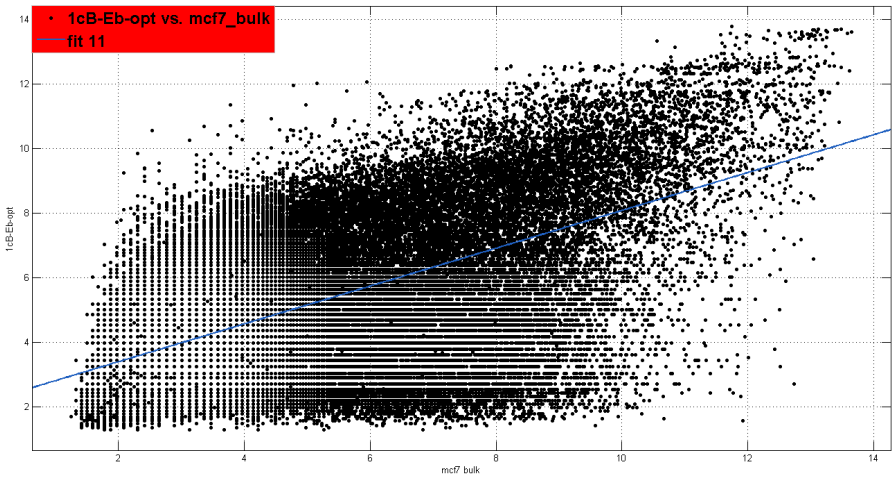

1c_K-opt_A2_Sm_1cell vs. mcf7_bulk
fit 17
mcf7 bulk

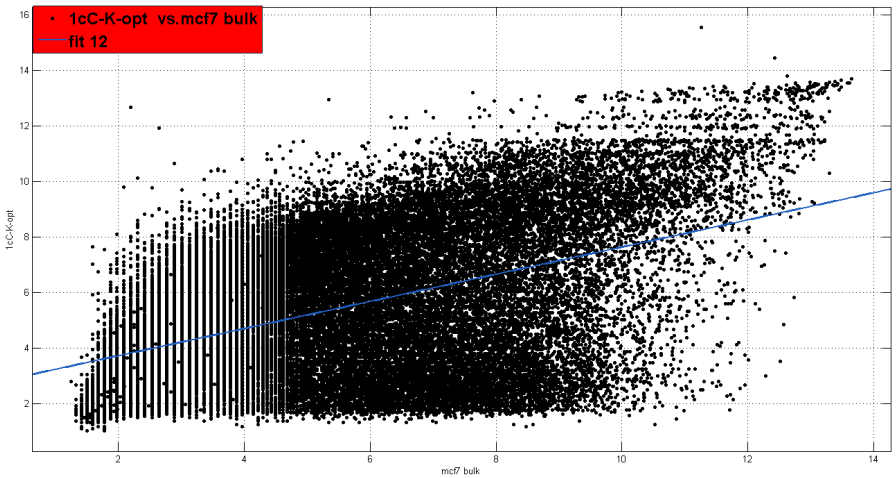

1c_K_ExTaq_SSIII_1cell vs. mcf7_bulk
fit 18
1c K ExTaq SSIII 1cell
mcf7 bulk

5


5000c_Eb-opt vs. mcf7_bulk
fit 19
5000c Eb-opt
mcf7 bulk


5000c_K-OPT vs. mcf7_bulk
fit 20
5000c K-OPT
mcf7 bulk


1cB-Eb-opt vs. mcf7_bulk
fit 11
1cB-Eb-opt
mcf7 bulk

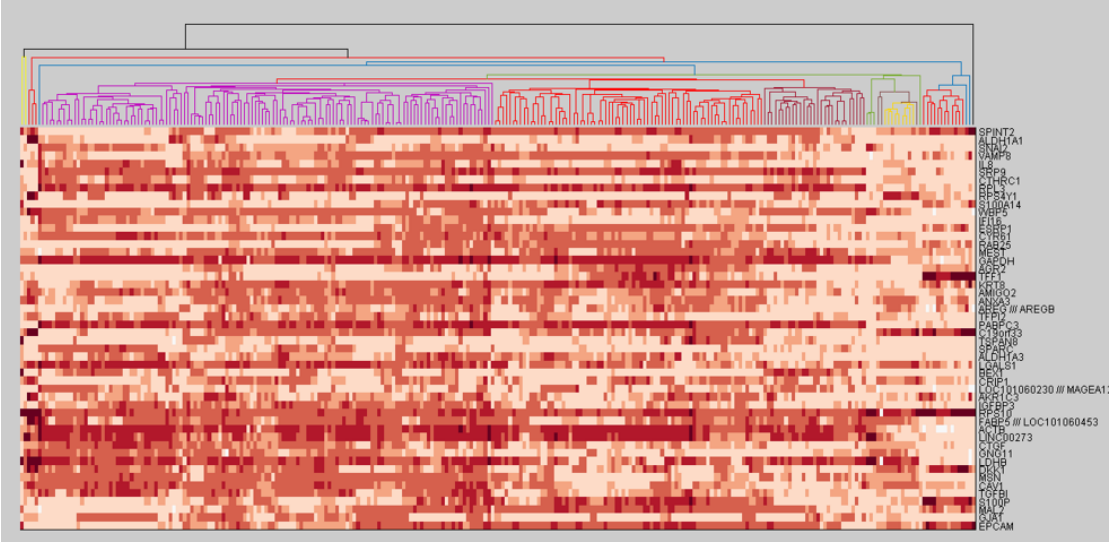

1cC-K-opt vs.mcf7 bulk
fit 12
1cC-K-opt
mcf7 bulk

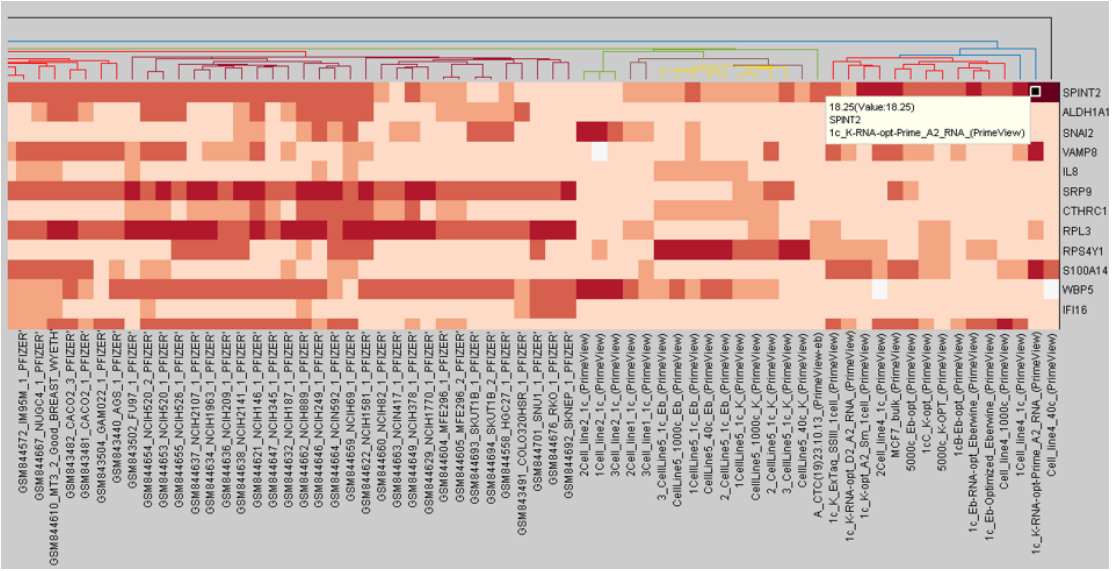
Παράρτημα Β

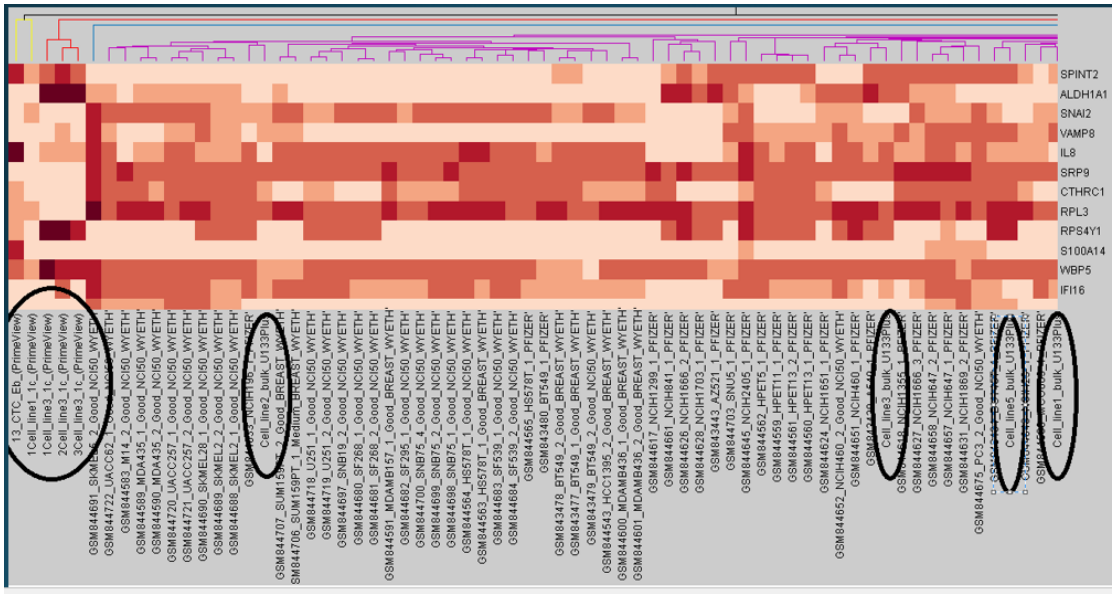
Στο Παράρτημα Β παρουσιάζονται διάφορες δοκιμές που έγιναν με το Clustering των δεδομένων και το αποτέλεσμα απεικονίζεται γραφικά μέσω των Heat Maps.

Expression Universe 1 - 50 γονίδια με highest Variance

Παρακάτω φαίνονται το Heat map, και διάφορα στιγμιότυπα του , που περιέχουν ενδεικτικές πληροφορίες. Το Heat map που απεικονίζεται , προέρχεται από το Clustering 270 samples (Expression Universe 1) , κατά μήκος 50 γονιδίων , με την highest Variance.

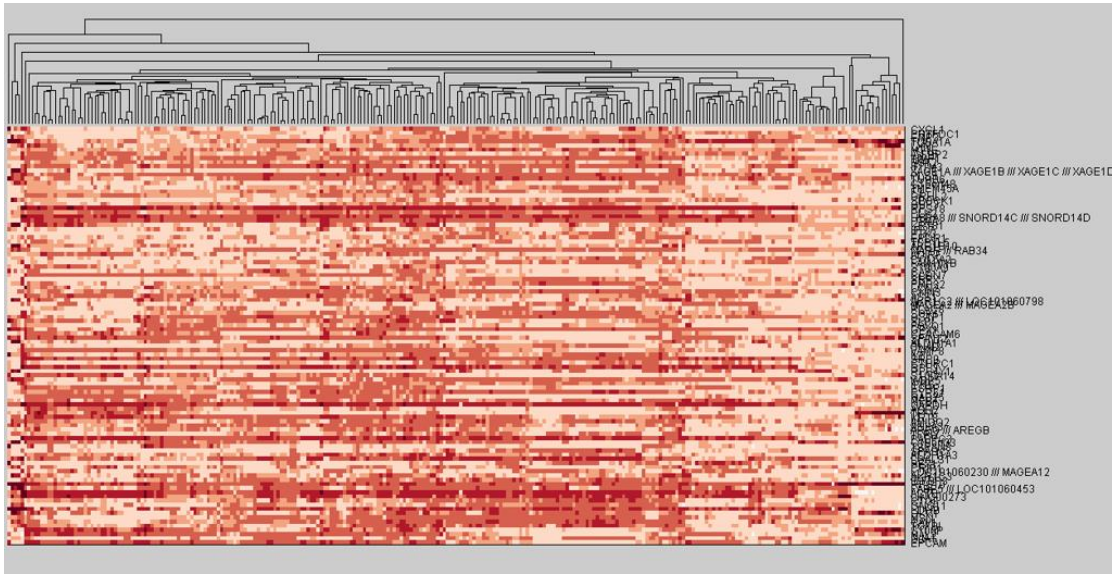


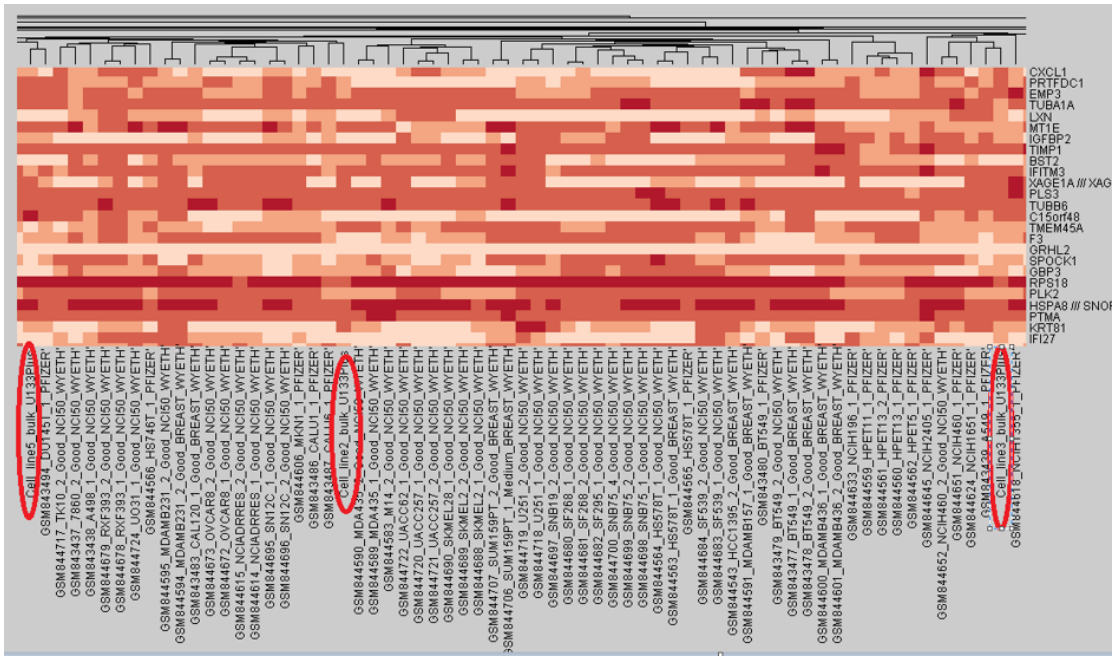




Expression Universe 1 - 100 γονίδια με highest Variance

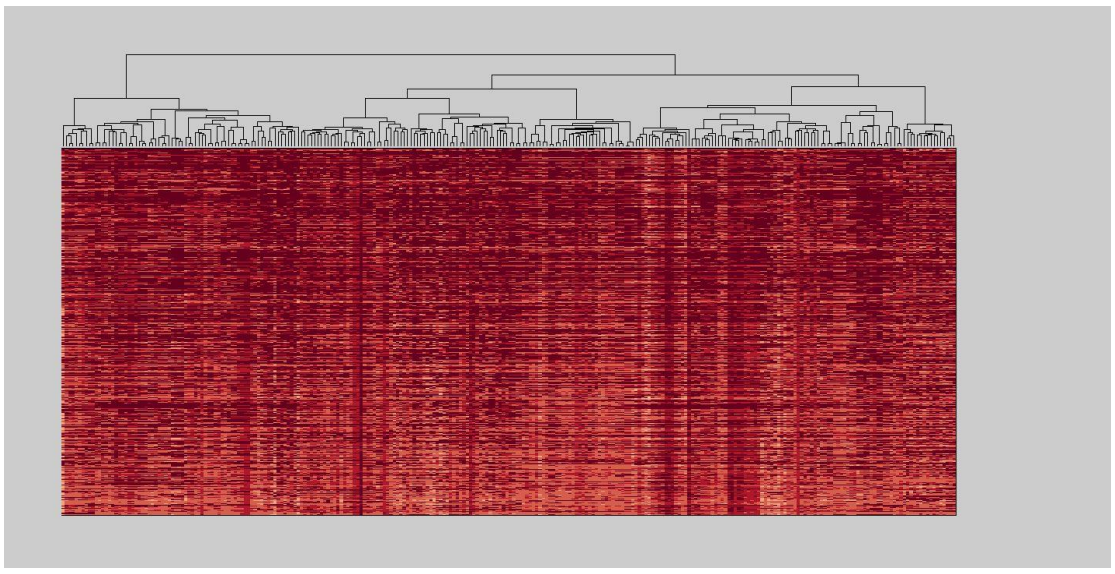
Παρακάτω φαίνονται το Heat map, και διάφορα στιγμιότυπα του , που περιέχουν ενδεικτικές πληροφορίες. Το Heat map που απεικονίζεται , προέρχεται από το Clustering 270 samples (Expression Universe 1) , κατά μήκος 100 γονιδίων , με την highest Variance.

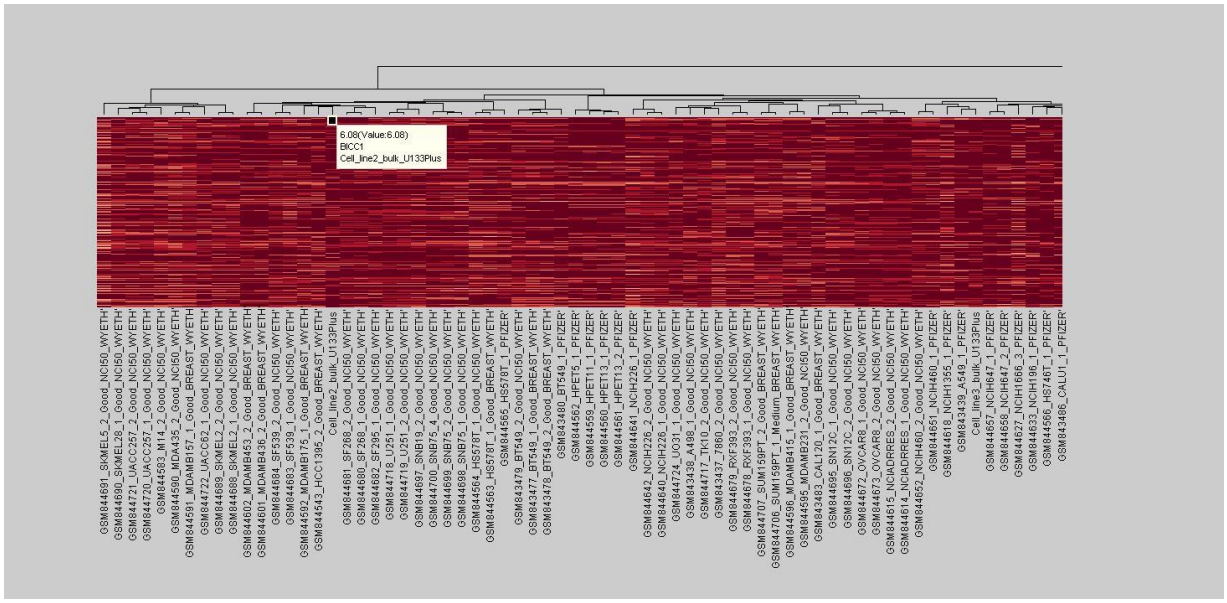


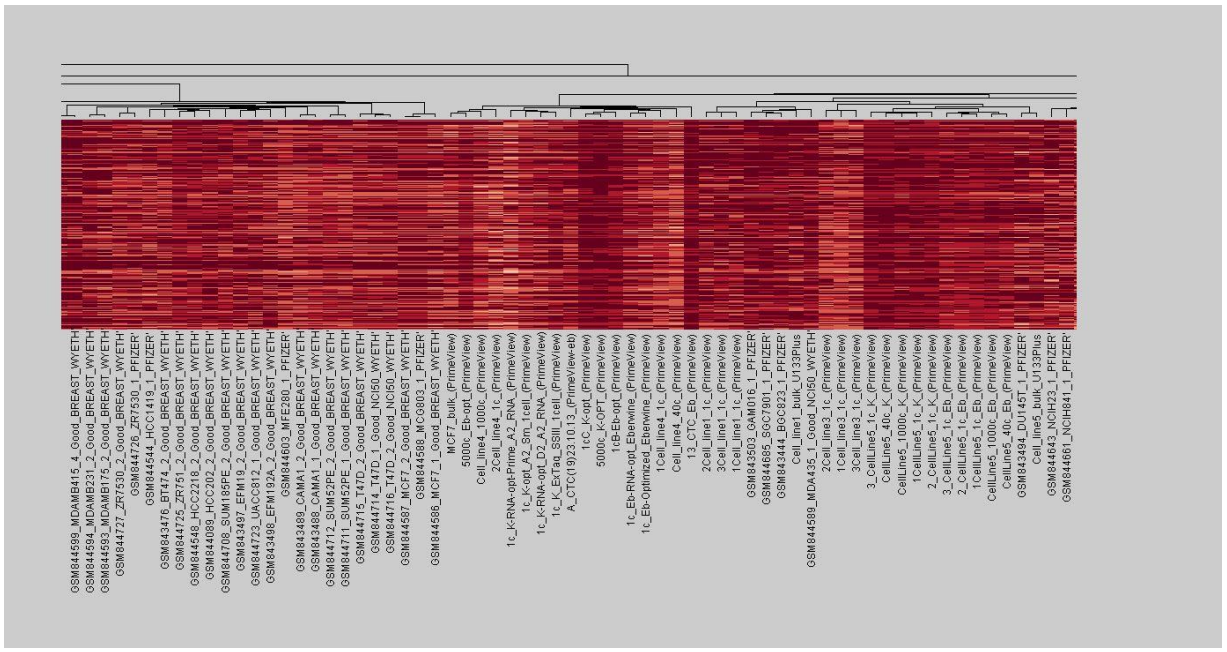


Expression Universe 1 - 900 γονίδια με highest Variance

Παρακάτω φαίνονται το Heat map, και διάφορα στιγμιότυπα του , που περιέχουν ενδεικτικές πληροφορίες. Το Heat map που απεικονίζεται , προέρχεται από το Clustering 270 samples (Expression Universe 1) , κατά μήκος 900 γονιδίων , με την highest Variance.

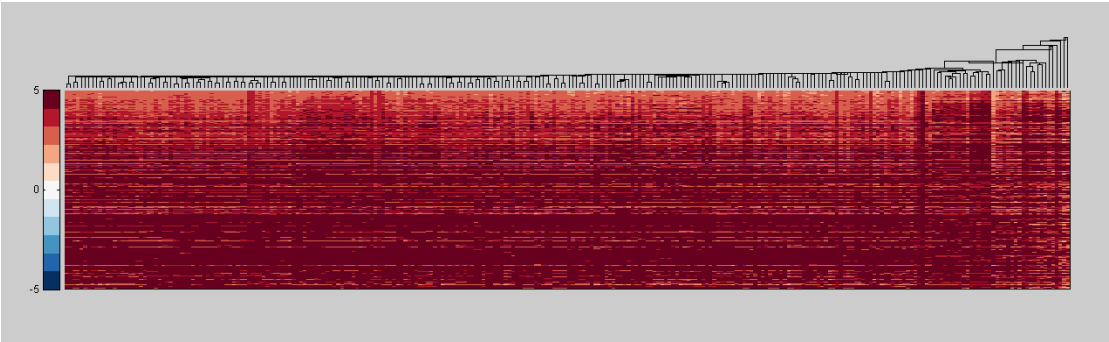


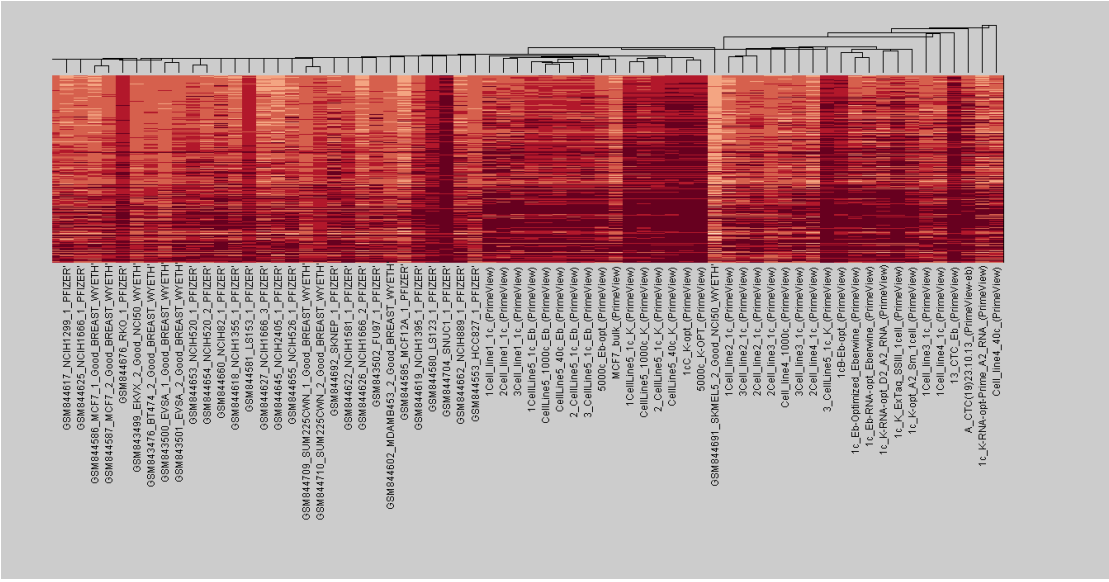

6.08(Value:6.08)
BICC1
Cell_line2_bulk_U133Plus



Expression Universe 1 - 18034 γονίδια με highest Variance - Μέθοδος - Centroid

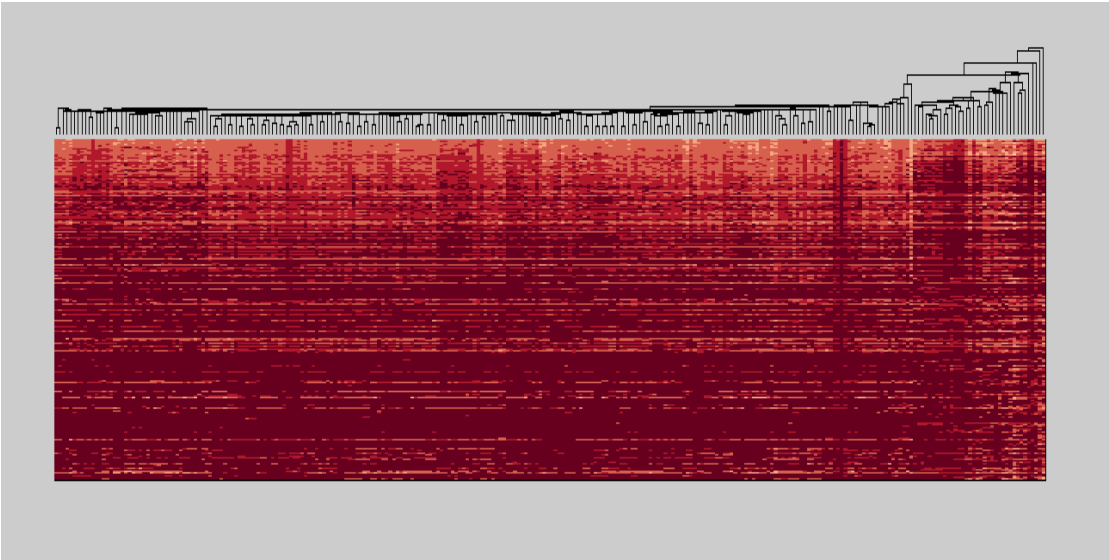
Παρακάτω φαίνονται ολόκληρο το Heat map, και διάφορα στιγμιότυπα του , που περιέχουν ενδεικτικές πληροφορίες. Το Heat map που απεικονίζεται , προέρχεται από το Clustering 270 samples (Expression Universe 1) , κατά μήκος 18034 γονιδίων , με την highest Variance και εκτελέστηκε με την μέθοδο Centroid.

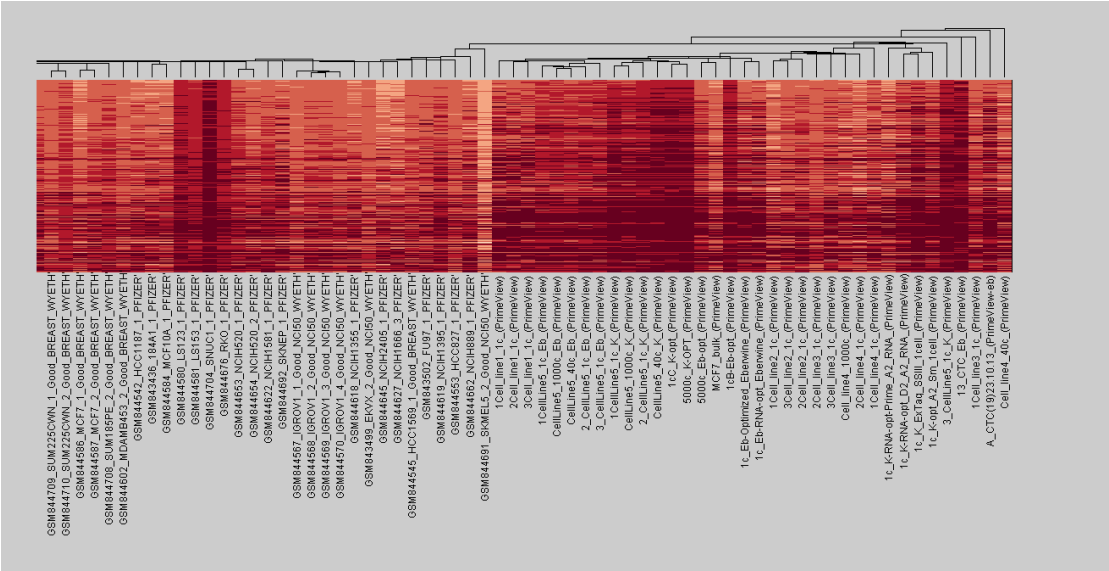




Expression Universe 1 - 18034 γονίδια με highest Variance - Μέθοδος - Median

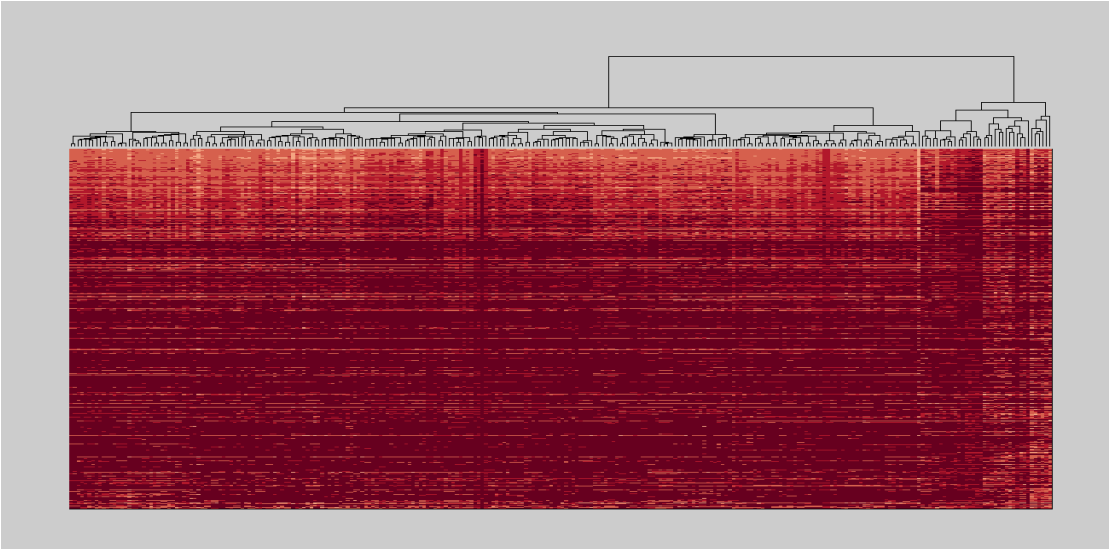
Παρακάτω φαίνονται ολόκληρο το Heat map, και διάφορα στιγμιότυπα του , που περιέχουν ενδεικτικές πληροφορίες. Το Heat map που απεικονίζεται , προέρχεται από το Clustering 270 samples (Expression Universe 1) , κατά μήκος 18034 γονιδίων , με την highest Variance και εκτελέστηκε με την μέθοδο Median.





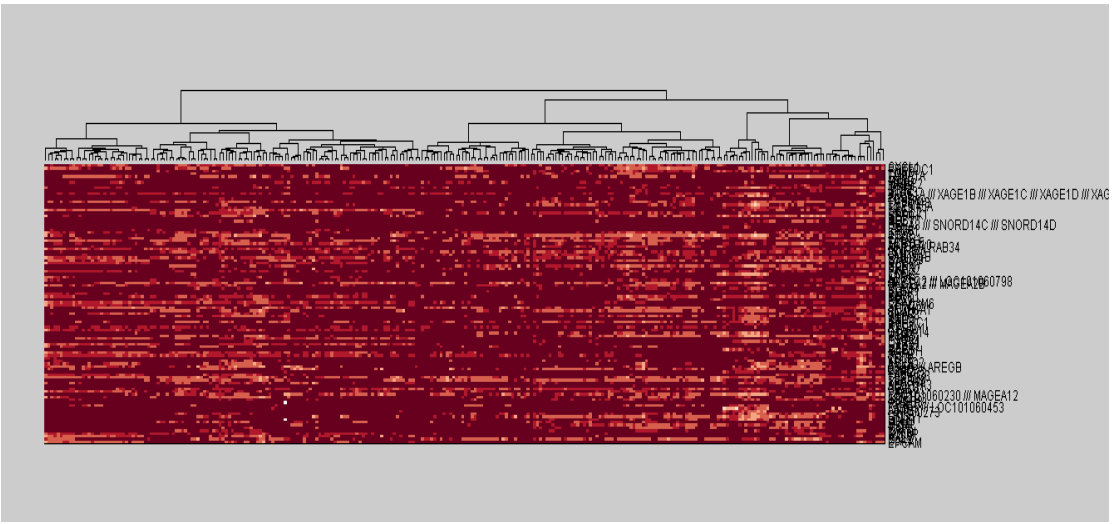
Expression Universe 1 - 18034 γονίδια με highest Variance - Μέθοδος - Ward

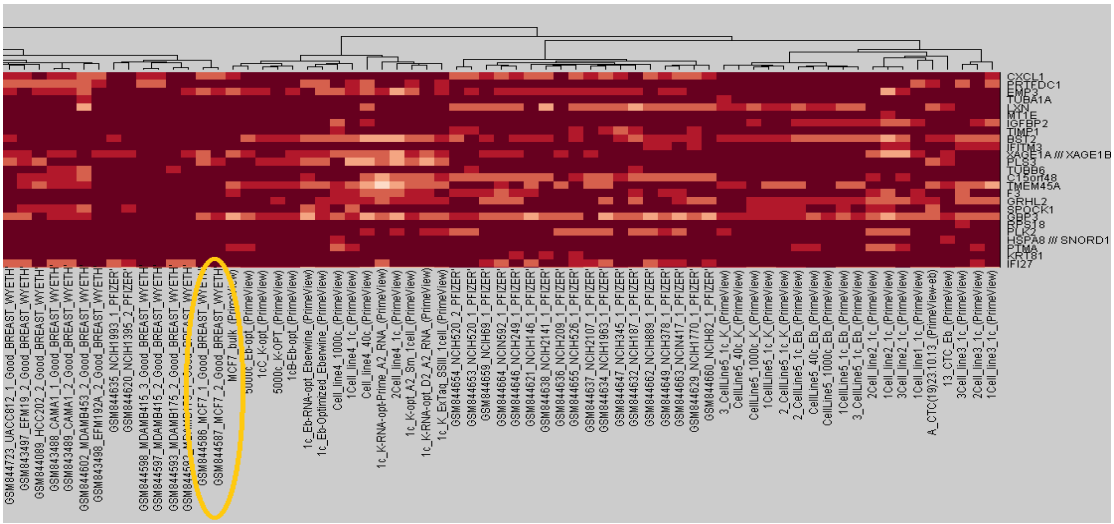
Παρακάτω φαίνονται ολόκληρο το Heat map. Το Heat map που απεικονίζεται , προέρχεται από το Clustering 270 samples (Expression Universe 1) , κατά μήκος 18034 γονιδίων , με την highest Variance και εκτελέστηκε με την μέθοδο Ward.



Expression Universe 1 - 100 γονίδια με highest Variance - Μέθοδος - ward

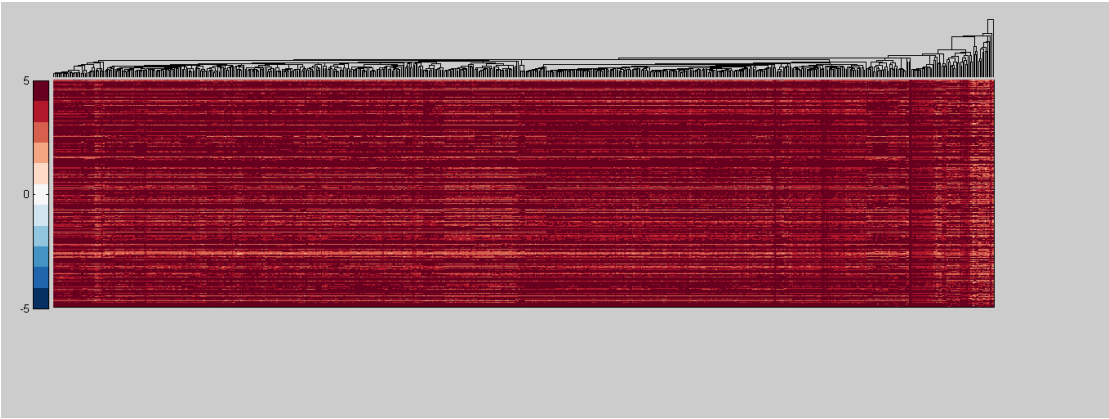
Παρακάτω φαίνονται ολόκληρο το Heat map, και διάφορα στιγμιότυπα του , που περιέχουν ενδεικτικές πληροφορίες. Το Heat map που απεικονίζεται , προέρχεται από το Clustering 270 samples (Expression Universe 1) , κατά μήκος 100 γονιδίων , με την highest Variance και εκτελέστηκε με την μέθοδοWard.

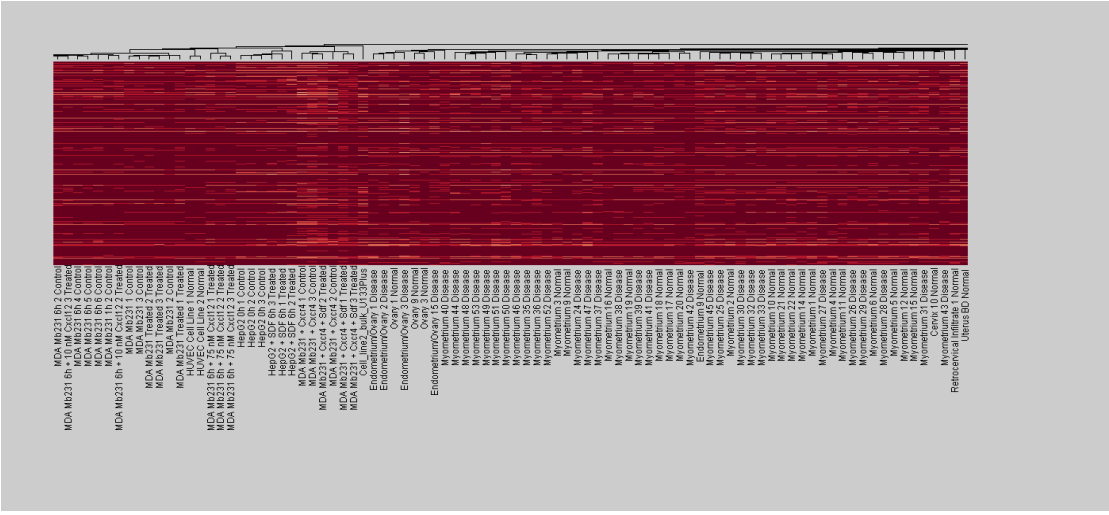


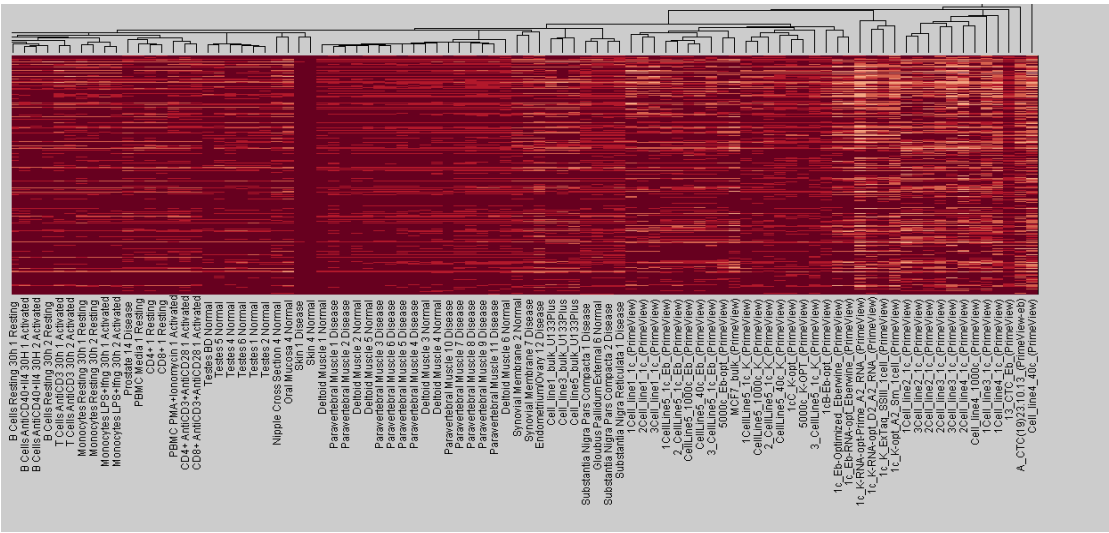


Expression Universe 2 - 18034 γονίδια με highest Variance - Μέθοδος - Average

Παρακάτω φαίνονται ολόκληρο το Heat map, και διάφορα στιγμιότυπα του , που περιέχουν ενδεικτικές πληροφορίες. Το Heat map που απεικονίζεται , προέρχεται από το Clustering 717 samples (Expression Universe 2) , κατά μήκος 18034 γονιδίων , με την highest Variance και εκτελέστηκε με την μέθοδο Average.

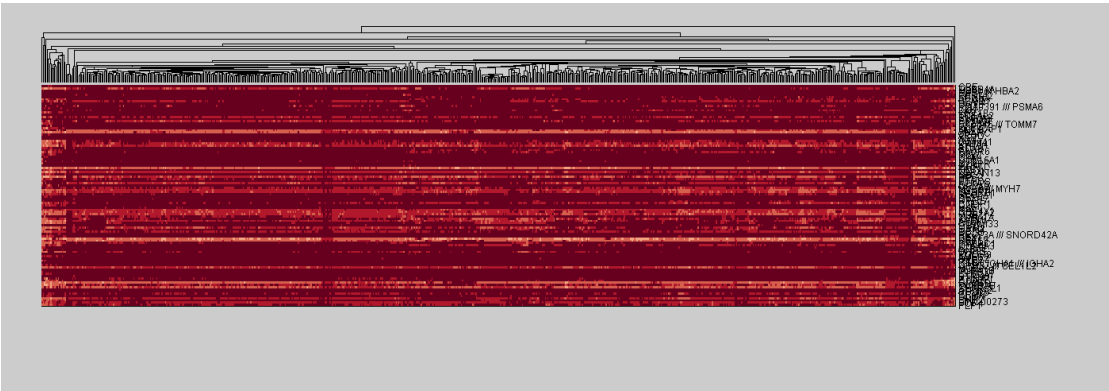

5
0
-5

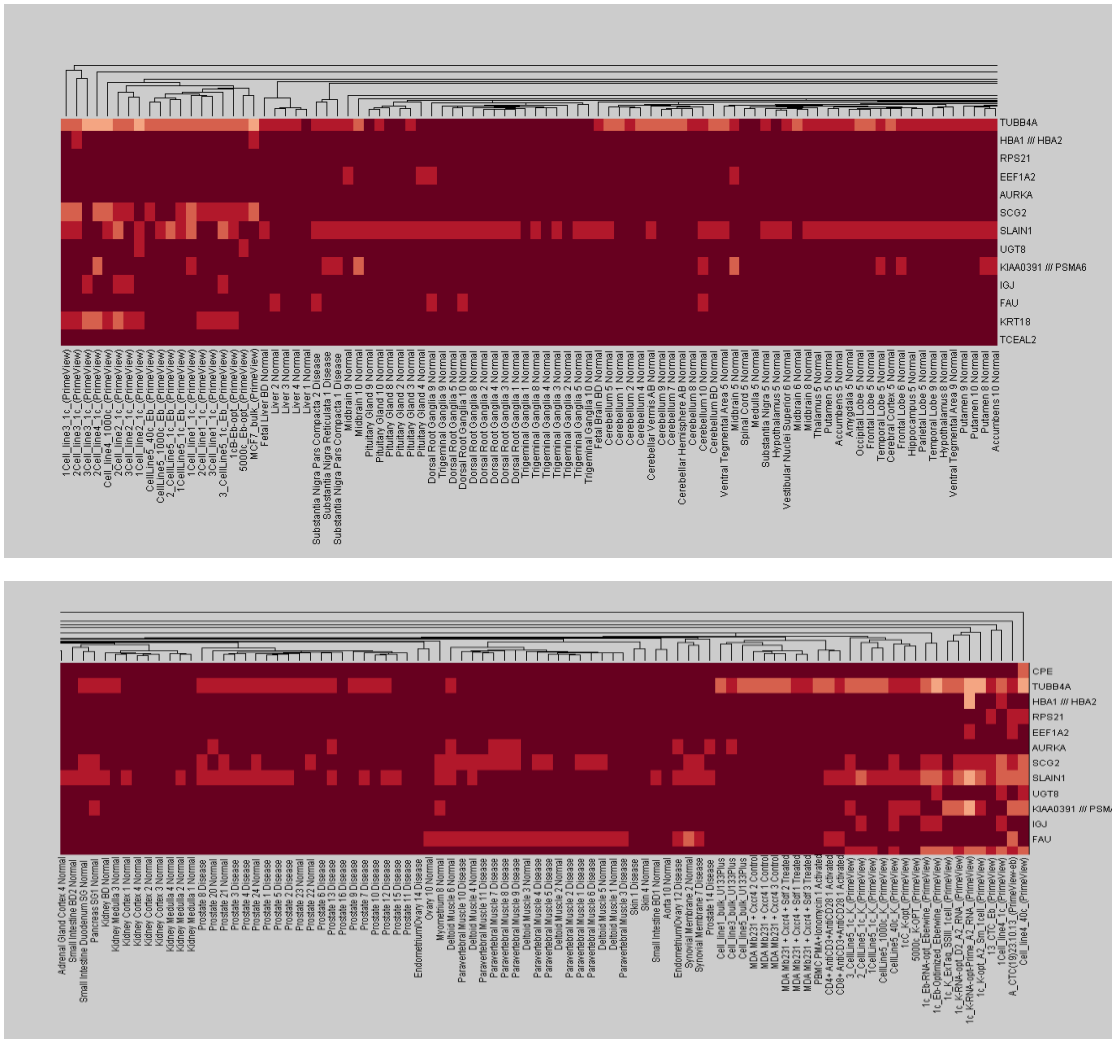

MDA Mb231 6h 2 Control
MDA Mb231 6h 3 Control
MDA Mb231 6h 4 Control
MDA Mb231 6h 5 Control
MDA Mb231 6h 6 Control
MDA Mb231 6h 1 Control
MDA Mb231 6h + 10 nM Cxcl12 2 Treated
MDA Mb231 1 Control
MDA Mb231 2 Control
MDA Mb231 Treated 2 Treated
MDA Mb231 Treated 3 Treated
MDA Mb231 2 Control
MDA Mb231 Treated 1 Treated
HUVEC Cell Line 1 Normal
HUVEC Cell Line 2 Normal
MDA Mb231 6h + 75 nM Cxcl12 2 Treated
MDA Mb231 6h + 75 nM Cxcl12 1 Treated
MDA Mb231 6h + 75 nM Cxcl12 3 Treated
HepG2 0h 1 Control
HepG2 0h 2 Control
HepG2 0h 3 Control
HepG2 + SDF 6h 3 Treated
HepG2 + SDF 6h 1 Treated
HepG2 + SDF 6h 2 Treated
MDA Mb231 + Cxcr4 1 Control
MDA Mb231 + Cxcr4 3 Control
MDA Mb231 + Cxcr4 + Sdf 2 Treated
MDA Mb231 + Cxcr4 + Sdf 3 Treated
MDA Mb231 + Cxcr4 + Sdf 1 Treated
Endometrium/Ovary 1 Disease
Endometrium/Ovary 2 Disease
Ovary 1 Normal
Endometrium/Ovary 3 Disease
Ovary 9 Normal
Ovary 3 Normal
Endometrium/Ovary 15 Disease
Myometrium 44 Disease
Myometrium 48 Disease
Myometrium 49 Disease
Myometrium 43 Disease
Myometrium 51 Disease
Myometrium 50 Disease
Myometrium 40 Disease
Myometrium 35 Disease
Myometrium 36 Disease
Myometrium 52 Disease
Myometrium 33 Normal
Myometrium 9 Normal
Myometrium 24 Disease
Myometrium 47 Disease
Myometrium 37 Disease
Myometrium 16 Normal
Myometrium 38 Disease
Myometrium 45 Normal
Myometrium 39 Disease
Myometrium 41 Disease
Myometrium 18 Normal
Myometrium 19 Normal
Myometrium 20 Normal
Endometrium 42 Disease
Endometrium 1 Normal
Myometrium 25 Disease
Myometrium 2 Normal
Myometrium 32 Disease
Myometrium 30 Disease
Myometrium 33 Disease
Myometrium 10 Normal
Myometrium 23 Normal
Myometrium 22 Normal
Myometrium 14 Normal
Myometrium 27 Normal
Myometrium 4 Normal
Myometrium 11 Normal
Myometrium 7 Normal
Myometrium 29 Disease
Myometrium 6 Normal
Myometrium 28 Disease
Myometrium 8 Normal
Myometrium 12 Normal
Myometrium 15 Normal
Myometrium 31 Disease
Cervix 1 Normal
Myometrium 43 Disease
Retrocervical Infiltrate 1 Normal
Uterus BD Normal



Expression Universe 2 - 100 γονίδια με highest Variance - Μέθοδος - Average

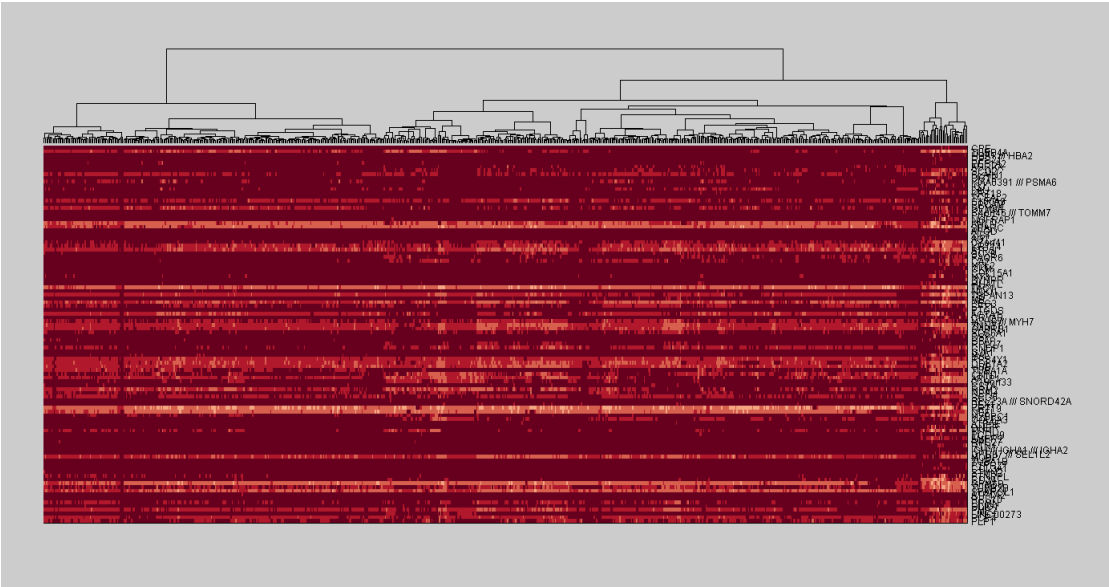
Παρακάτω φαίνονται ολόκληρο το Heat map, και διάφορα στιγμιότυπα του , που περιέχουν ενδεικτικές πληροφορίες. Το Heat map που απεικονίζεται , προέρχεται από το Clustering 717 samples (Expression Universe 2) , κατά μήκος 100 γονιδίων , με την highest Variance και εκτελέστηκε με την μέθοδο Average.

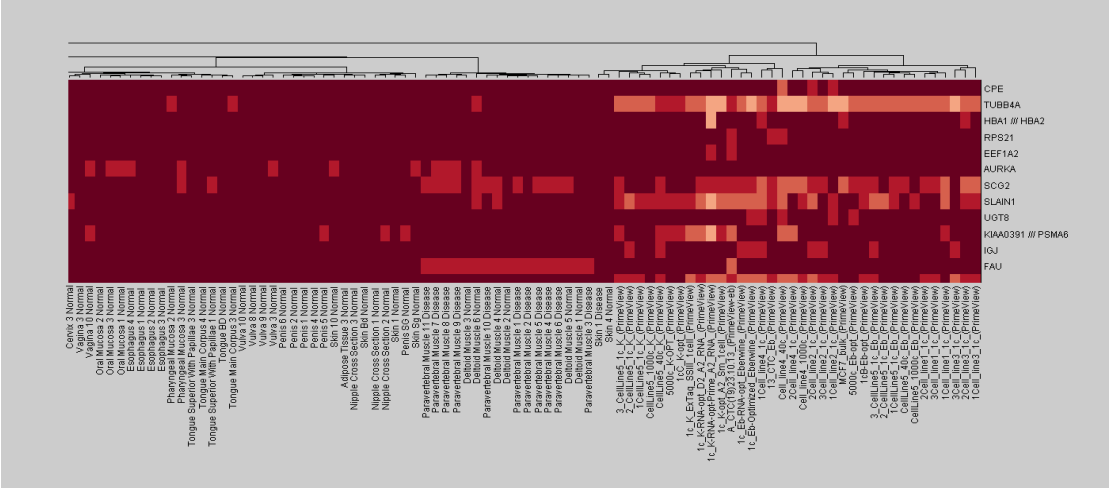


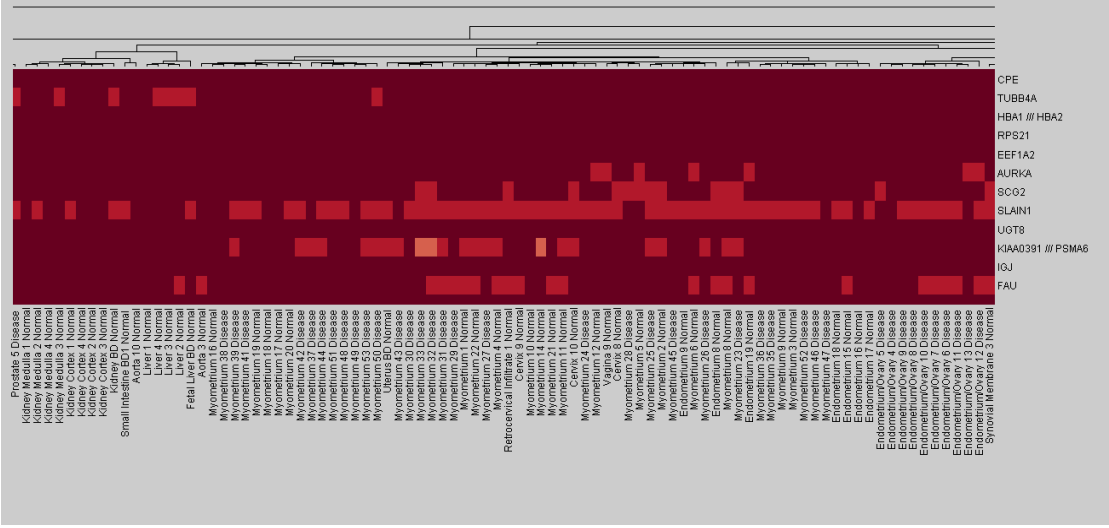


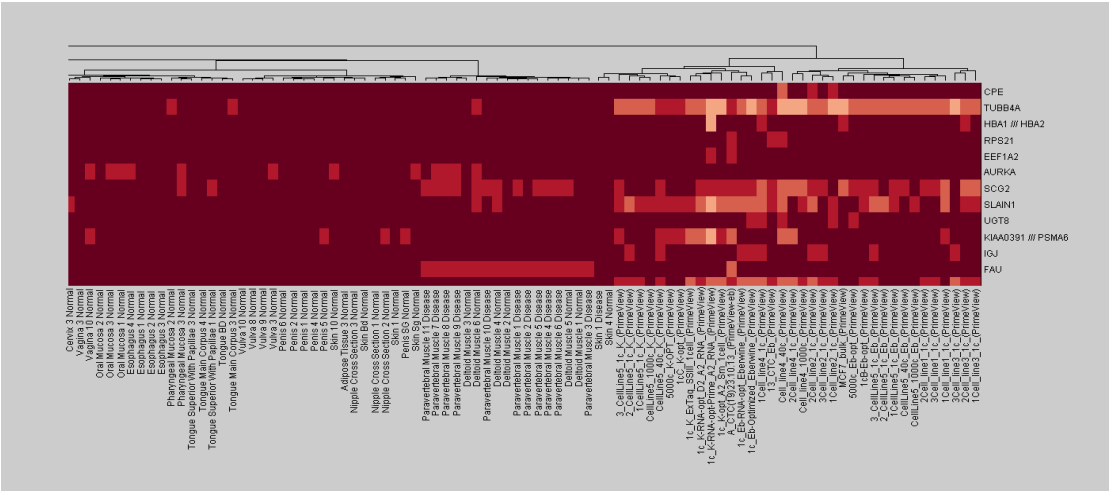
Expression Universe 2 - 100 γονίδια με highest Variance - Μέθοδος - Ward

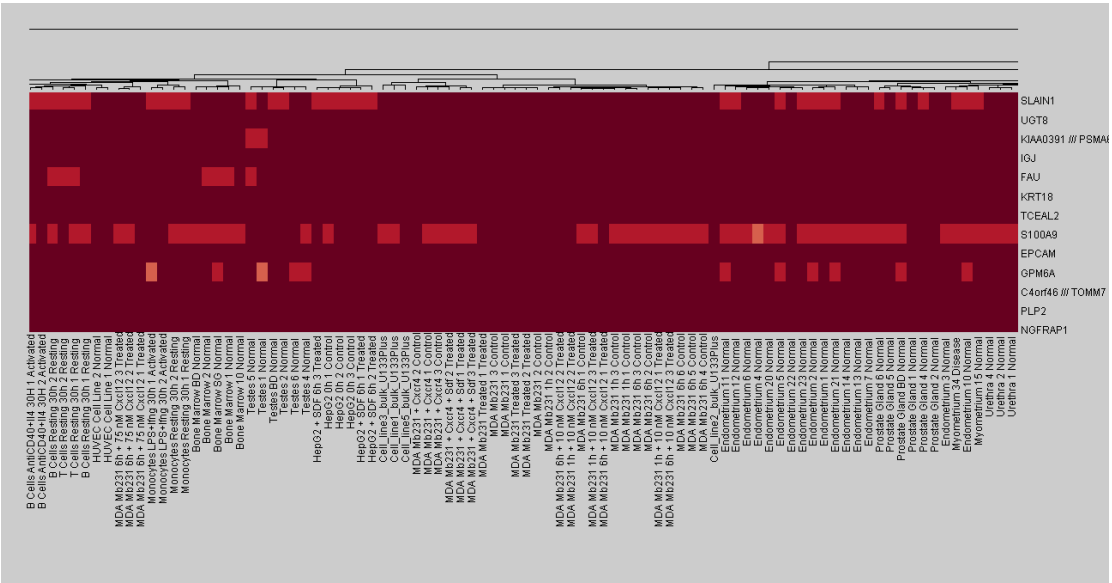
Παρακάτω φαίνονται ολόκληρο το Heat map, και διάφορα στιγμιότυπα του , που περιέχουν ενδεικτικές πληροφορίες. Το Heat map που απεικονίζεται , προέρχεται από το Clustering 717 samples (Expression Universe 2) , κατά μήκος 100 γονιδίων , με την highest Variance και εκτελέστηκε με την μέθοδο Ward.




CPE
TUBB4A
HBA1 /// HBA2
RPS21
EEF1A2
AURKA
SCG2
SLAIN1
UGT8
KIAA0391 /// PSMA6
IGJ
FAU


CPE
TUBB4A
HBA1 /// HBA2
RPS21
EEF1A2
AURKA
SCG2
SLAIN1
UGT8
KIAA0391 /// PSMA6
IGJ
FAU


CPE
TUBB4A
HBA1 /// HBA2
RPS21
EEF1A2
AURKA
SCG2
SLAIN1
UGT8
KIAA0391 /// PSMA6
IGJ
FAU


SLAIN1
UGT8
KIAA0391 /// PSMA6
IGJ
FAU
KRT18
TCEAL2
S100A9
EPCAM
GPM6A
C4orf46 /// TOMM7
PLP2
NGFRAP1