Ατομική Διπλωματική Εργασία

ΚΑΤΑΝΕΜΗΜΕΝΟΣ ΑΛΓΟΡΙΘΜΟΣ ΑΝΑΛΥΣΗΣ ΤΗΣ ΛΕΙΤΟΥΡΓΙΚΗΣ ΕΠΙΔΡΑΣΗΣ ΤΩΝ ΣΗΜΕΙΑΚΩΝ ΝΟΥΚΛΕΟΤΙΔΙΚΩΝ ΠΟΛΥΜΟΡΦΙΣΜΩΝ, ΣΤΑ ΕΠΙΠΕΔΑ ΕΚΦΡΑΣΗΣ ΤΩΝ ΑΛΛΗΛΟΥΧΙΩΝ ΤΟΥ mRNA ΚΑΙ ΤΩΝ ΠΡΩΤΕΪΝΩΝ

Ιωάννα Κάλβαρη

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάιος 2009

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Κατανεμημένος Αλγόριθμος Ανάλυσης της Λειτουργικής Επίδρασης των Σημειακών Νουκλεοτιδικών Πολυμορφισμών στα Επίπεδα Έκφρασης των Αλληλουχιών του mRNA και των Πρωτεϊνών

Ιωάννα Κάλβαρη

Επιβλέπων Καθηγητής
Κωσταντίνος Παττίχης

Η Ατομική Διπλωματική Εργασία υποβλήθηκε προς μερική εκπλήρωση των απαιτήσεων απόκτησης του πτυχίου Πληροφορικής του Τμήματος Πληροφορικής του Πανεπιστημίου Κύπρου

Μάιος 2009

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον κύριο Άθω Αντωνιάδη για την πολύτιμη υποστήριξη και καθοδήγηση που προσέφερε καθ' όλη τη διάρκεια του project. Για την κατατόπιση γύρω από το θέμα και την κατανόηση βιολογικών εννοιών που σχετίζονται με το θέμα και την υποστήριξη του σε θέματα πληροφορικής και την παροχή πρόσβασης στο Grid, που βρισκόταν διαθέσιμο για την ικανοποίηση των αναγκών της μελέτης από την Glaxo Smith Kline (GSK).

Επίσης θα ήθελα να ευχαριστήσω θερμά τον καθηγητή κύριο Παττίχη για τη συνεχή επίβλεψη και υποστήριξη του για την ομαλή διεξαγωγή της μελέτης.

Επιπλέον θα ήθελα να δώσω τις θερμές μου ευχαριστίες στους επιστήμονες από την GSK για την σημαντική συμβολή της στην διεξαγωγή του project, παρέχοντας τα δεδομένα καθώς επίσης και την περιγραφή του προβλήματος.

Ευχαριστίες επίσης για τους κυρίους Enrico Domenichi για τη χρήσιμη καθοδήγηση του γύρω από τον τομέα των proteomics αναφορικά με το mRNA και τις πρωτεΐνες και τον κύριο Pierandrea Muglia για την πολύτιμη προσφορά γύρω από θέματα που σχετίζονται με την γενετική και για την παροχή των δεδομένων DNA.

Περίληψη

Η υφιστάμενη μελέτη έχει ως θέμα της την υλοποίηση ενός κατανεμημένου αλγόριθμου ανάλυσης, της λειτουργικής επίδρασης των σημειακών νουκλεοτιδικών πολυμορφισμών (SNP), στα επίπεδα έκφρασης των αλληλουχιών mRNA και πρωτεϊνών αντίστοιχα.

Έχοντας ως σημείο αφετηρίας ένα αρχείο με συνολικά 550 χιλιάδες δεδομένα για SNPs καθώς επίσης για 56 χιλιάδες δεδομένα έκφρασης mRNA και τέλος ένα με 89 πρωτεΐνες, ακολουθήθηκε μία πορεία διαδικασίας ανάλυσης των δεδομένων σε ένα κατανεμημένο σύστημα.

Η όλη διαδικασία διεξήχθη με την βοήθεια ενός πλέγματος υπολογιστών μεγέθους 200 επεξεργαστών. Για τη χρήση του Grid καθώς επίσης και την διαχείριση των δεδομένων, υλοποιήθηκαν κατάλληλοι αλγόριθμοι για την προσαρμογή τους στο grid προς εκτέλεση. Επίσης υλοποιήθηκε κώδικας για περεταίρω διαχείριση των αποτελεσμάτων που παράχθηκαν κατά την ανάλυση και την μετέπειτα προσαρμογή τους στην εφαρμογή Spotfire, για την αποδοτική απεικόνιση και παρουσίαση των αποτελεσμάτων.

Η ανάλυση των δεδομένων έγινε με την εφαρμογή στατιστικών μεθόδων που προσφέρονταν ήδη από μία εφαρμογή ανοικτού κώδικα το Plink. Συγκεκριμένα εφαρμόστηκαν οι διεργασίες μεταλλαγής και ποσοτική ανάλυση στα δεδομένα για την ανεύρεση τυχόν συσχετίσεων μεταξύ των SNPs και των μετάγραφων mRNA.

Τα δεδομένα όσο και το πλέγμα υπολογιστών καθώς επίσης και η εφαρμογή Spotfire που χρησιμοποιήθηκε σε τελευταίο στάδιο, για την απεικόνιση και την παρουσίαση των αποτελεσμάτων, παρέχονταν για ικανοποίηση των απαιτήσεων της μελέτης, από την φαρμακευτική εταιρεία GSK – GlaxoSmithKline.

Στα κεφάλαια που ακολουθούν αναλύεται εις βάθος η μεθοδολογία που χρησιμοποιήθηκε για την παραλληλοποίηση, οι αλγόριθμοι που εφαρμόστηκαν, καθώς επίσης και οι εφαρμογές που χρησιμοποιήθηκαν για την διεξαγωγή της μελέτης.

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

1.1 Εισαγωγή

Το μεγάλο ενδιαφέρον για την κατανόηση του τεράστιου όγκου πληροφοριών που προέρχονται από την μελέτη του γονιδιώματος των οργανισμών, οδήγησε στην ανάγκη για μελέτη των πρωτεϊνών και πιο συγκεκριμένα την δομή τους και τις λειτουργίες που παρέχουν σε ένα οργανισμό (proteomics: large – scale study of proteins particularly their structures and functions). Η πολυπλοκότητα των πρωτεϊνών καθιστά πολύ δύσκολη την αναγνώριση και κατανόηση της λειτουργίας τους. Η δυσκολία αυτή που παρατηρείται στην αναγνώριση και την κατανόηση της λειτουργίας των πρωτεϊνών και κατά συνέπεια των γονιδίων που τις κωδικοποιούν, συντείνει στην ανάπτυξη νέων τεχνολογιών για συστηματική και περιεκτική ανάλυση της δομής και της λειτουργίας των πρωτεϊνών, που τον τελευταίο καιρό αποτελεί πρόκληση στον τομέα της έρευνας [12].

Το έναυσμα που οδήγησε στην πρόσφατη πρόοδο που παρατηρήθηκε στις βιολογικές επιστήμες, είναι η ολοκλήρωση της αλληλουχίας του ανθρώπινου γονιδιώματος που οδήγησε στην αναγνώριση περίπου 35 χιλιάδων γονιδίων [12].

Το δύσκολο έργο ανάθεσης λειτουργιών σε κάθε ένα από αυτά τα 35 χιλιάδες γονίδια, μόλις που άρχισε να εξελίσσεται και αποτελεί τον πρωταρχικό στόχο της μελέτης της

λειτουργίας του ανθρώπινου γονιδιώματος (human functional genomics). Η λειτουργία ενός γονιδίου, καθορίζεται από το προϊόν της πρωτεΐνης που κωδικοποιεί [12].

Επομένως βάσει των πιο πάνω μπορεί να γίνει αντιληπτό, πως η μελέτη του ανθρώπινου γονιδιώματος, αποτελεί και θέμα μέγιστου ενδιαφέροντος στον τομέα της έρευνας. Αποτέλεσμα αυτής της πρόκλησης που αποτελεί η μελέτη των πρωτεϊνών, οδηγεί στην ανάπτυξη μηχανισμών και ειδικού εξοπλισμού που θα υποβοηθούν και θα επιταχύνουν την έκφραση του μεγάλου αριθμού ανθρώπινων γονιδίων και των προϊόντων που κωδικοποιούν, έτσι ώστε να επιτυγχάνεται μια συστηματική και περιεκτική ανάλυση της δομής και της λειτουργίας των πρωτεϊνών [12].

Η έκφραση των γνωρισμάτων των γονιδίων στα λεμφοκύτταρα είναι κληρονομική και οι γενετικές διαφορές παρατηρήθηκε πως συμβάλλουν στην μεταβλητότητα της έκφρασης των γονιδίων, στα περιφερειακά κύτταρα. Πρόσφατα, έχει γίνει αναφορά για συγκεκριμένη έκφραση των αλληλόμορφων σε ένα ευρύ φάσμα του ανθρώπινου γονιδιώματος, στα κύτταρα και τους εγκεφαλικούς ιστούς. Αυτό έχει ως αποτέλεσμα την μεγάλη επίδραση της γενετικής μεταβλητότητας στους μηχανισμούς μεταγραφής σε διαφορετικούς ιστούς [9].

Η βάση δεδομένων που χρησιμοποιήθηκε προήλθε από μία έρευνα πάνω στην ασθένεια Μονοπολική Κατάθλιψη (Unipolar Depression) με ασθενή και υγιή άτομα. Για την ασθένεια αυτή έγινε σκιαγράφηση της έκφρασης του mRNA, στα δείγματα αίματος των ατόμων που συμμετείχαν στη διαδικασία. Αυτό σε μία προσπάθεια αναγνώρισης βιολογικών δεικτών που σχετίζονται με τη συγκεκριμένη ασθένεια [9].

Σε αυτή την έρευνα αφαιρέθηκε η παράμετρος της ασθένειας της Μονοπολικής Κατάθλιψης έτσι ώστε να μελετηθούν μόνο οι συσχετίσεις μεταξύ της έκφρασης mRNA, πρωτεϊνών και πολυμορφισμών στο γονιδίωμα.

Η ενσωμάτωση των γενετικών δεδομένων με τις πληροφορίες που παράχθηκαν κατά τη σκιαγράφηση της έκφρασης των γονιδίων, έγινε με την διαδικασία σάρωσης των συσχετίσεων σε ολόκληρο το γονιδίωμα, για ανίχνευση των εκφρασμένων γνωρισμάτων από περίπου 56 χιλιάδες probeSets έναντι 550 χιλιάδων SNP δεικτών.

Στο πιο κάτω κείμενο περιγράφεται η διαδικασία επεξεργασίας και ανάλυσης των δεδομένων, καθώς επίσης και τα αποτελέσματα που παράχθηκαν από την πειραματική μελέτη που διεξάγει πάνω σε 190 ανθρώπινα αιματολογικά δείγματα. Τα δείγματα προέρχονται από ένα σύνολο ασθενών με Μονοπολική Κατάθλιψη (Unipolar Depression), 126 σε αριθμό ασθενείς (cases) και 64 φυσιολογικά δείγματα (controls).

Η μελέτη αυτή έχει ως σκοπό να προσδιορίσει όλες τις συσχετίσεις μεταξύ γενετικών πολυμορφισμών σε ολόκληρο το γονιδίωμα και την έκφραση γονιδίων μέσω του επίπεδου έκφρασης του mRNA τους.

Τα δεδομένα που παράχθηκαν χρησιμοποιώντας αυτή την προσέγγιση μπορούν να εφαρμοστούν σε οποιαδήποτε περιοχή ασθενειών ή οποιονδήποτε γενετικών χαρακτηριστικών.

1.2 Παρουσίαση Προβλήματος

Γενετικοί πολυμορφισμοί έχουν βρεθεί να επηρεάζουν την έκφραση γονιδίων σε όλα τα είδη κυττάρων. Έχοντας στη διάθεση μας πειραματικά δεδομένα από 190 δείγματα περιφερικού αίματος στα οποία έγινε ανάλυση έκφρασης mRNA και ορισμένων πρωτεϊνών καθώς επίσης και ανάλυση κάποιων γονοτύπων.

Λόγω του γεγονότος ότι τα 190 δείγματα έχουν συλλεχθεί από ένα σύνολο 126 ασθενών με Μονοπολική Κατάθλιψη (Unipolar Depression) και 64 φυσιολογικά δείγματα (control samples), ήταν αναγκαίο να αφαιρεθεί η παράμετρος της ασθένειας έτσι ώστε όταν γινόταν έλεγχος συσχέτισης μεταξύ έκφρασης mRNA ή πρωτεϊνών και σημειακών νουκλεοτιδικών πολυμορφισμών ώστε να μην μετρούνται τυχών επίπεδα συσχετίσεων που έχουν να κάνουν με την Μονοπολική Κατάθλιψη, αφού τόσο τα SNPs όσο και η έκφραση mRNA και πρωτεϊνών, έχουν ήδη μελετηθεί σε προηγούμενες μελέτες.

Η συλλογή των δεδομένων έκφρασης mRNA έγινε με το affimetrix HU133 plus V2 genechips το οποίο καλύπτει περίπου 56 χιλιάδες περιοχές mRNA. Συλλέχθηκαν επίσης

επίπεδα έκφρασης 80 συγκεκριμένων πρωτεϊνών. Από τα ίδια δείγματα έγινε συλλογή γενετικών δεδομένων με την πλατφόρμα illumina 550K η οποία εξετάζει 550 χιλιάδες SNP από ολόκληρο το γονιδίωμα καλύπτοντας έτσι το μεγαλύτερο μέρος των πιθανών γενετικών παραλλαγών.

Στόχος είναι η ανακάλυψη γενετικών πολυμορφισμών οι οποίοι σχετίζονται με την έκφραση συγκεκριμένων ακολουθιών mRNA ή και πρωτεϊνών.

Η ανάλυση των δεδομένων, λόγω του μεγάλου αριθμού των ελέγχων που εφαρμόστηκαν στα δεδομένα (550 χιλιάδες SNPs και 56 χιλιάδες δεδομένα έκφρασης mRNA), αποτελούσε μια υπολογιστικά ακριβή μέθοδο που εκτός από μεγάλες απαιτήσεις σε χρόνο εκτέλεσης, υπήρχε και μεγάλη ανάγκη σε αποθηκευτικό χώρο. Η μέθοδος επίλυσης για τα προβλήματα που μόλις αναφέρθηκαν έγινε με την εφαρμογή κατανεμημένου υπολογισμού που αναλύεται σε μεγαλύτερο βάθος στην συνέχεια.

Όσον αφορά τα αποτελέσματα που παράχθηκαν κατά την ανάλυση, λόγω του μεγάλου αριθμού των δεδομένων και όλων των δυνατών συσχετίσεων που εφαρμόστηκαν μεταξύ των 550 χιλιάδων SNPs και των 56 χιλιάδων δεδομένων έκφρασης μορίων mRNA, ήταν αναγκαία η χρήση ειδικής εφαρμογής, που επιτρέπει την αποδοτική διαχείριση αποτελεσμάτων μεγάλου όγκου, όπως και αυτών που παράχθηκαν σε αυτή τη μελέτη. Η εφαρμογή που χρησιμοποιήθηκε για την διαχείριση και την γραφική απεικόνιση των αποτελεσμάτων παρουσιάζεται αναλυτικά σε μετέπειτα στάδιο [12].

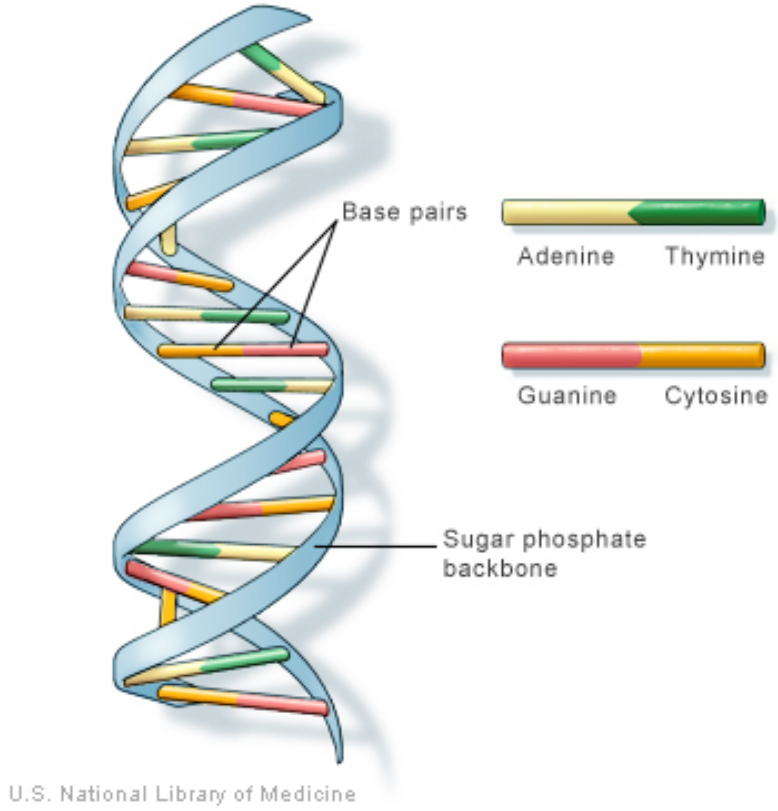
Κεφάλαιο 2

Βασικές Έννοιες Μοριακές Βιολογίας

2.1 Δισόξυριβοζονουκλεϊκό οξύ

DNA δισοξυριβοζονουκλεϊκό οξύ (Deoxyribonucleic acid) περιέχει όλες τις γενετικές πληροφορίες των περισσότερων οργανισμών. Βρίσκεται στον πυρήνα των κυττάρων και ο κύριος ρόλος του είναι η μακροχρόνια αποθήκευση πληροφοριών και οδηγιών για την παραγωγή άλλων συστατικών των κυττάρων, όπως είναι μόρια RNA και πρωτεΐνες. Σύμφωνα με την περιγραφή της δομής του DNA από τους Watson & Crick το 1953, το

DNA από χημικής πλευράς, αποτελεί ένα δίκλωνο μόριο με μορφή έλικας. Κάθε κλώνος του DNA αποτελεί μία πολυνουκλεοτιδική αλυσίδα τα στοιχεία της οποίας αποτελούνται από 4 είδη νουκλεοτιδίων A,T,C και G. Κάθε κλώνος έχει αρχή και τέλος, με την αρχή να συμβολίζεται ως το 5΄ άκρο και το τέλος ως 3΄ άκρο. Ο τρόπος με τον οποίο συνδέονται οι δύο κλώνοι είναι από το άκρο 5΄ προς το άκρο 3΄ με τέτοιο τρόπο ώστε να είναι αντιπαράλληλοι, δηλαδή να είναι ενωμένοι με τέτοιο τρόπο που απέναντι από το 5΄ άκρο του ενός να βρίσκεται το 3΄ άκρο του άλλου κλώνου. Τα νουκλεοτίδια συνδέονται ομοιοπολικά μεταξύ τους με φωσφοδιεστερικούς δεσμούς και με τέτοιο τρόπο προσδίδοντας έτσι χημική πολικότητα στον κάθε κλώνο DNA. Μεταξύ των αζωτούχων βάσεων των νουκλεοτιδίων, που βρίσκονται το ένα απέναντι από το άλλο, δημιουργούνται δεσμοί υδρογόνου που συγκρατούν τους δύο κλώνους ενωμένους. Οι βάσεις βρίσκονται πάντοτε στο εσωτερικό της έλικας ενώ ο σάκχαρο-φωσφορικός σκελετός στο εξωτερικό της έλικας. Αυτή η δομή του DNA με τις βάσεις στο εσωτερικό της έλικας προσδίδουν προστασία στις γενετικές πληροφορίες, αφού φροντίζει ώστε να παραμένουν αναλλοίωτες. Επιπλέον συνδέονται πάντοτε σύμφωνα με τον κανόνα συμπληρωματικότητας των Watson & Crick, με την αδενίνη (A) να ενώνεται πάντοτε με τη θυμίνη (T) με δύο δεσμούς υδρογόνου και την κυτοσίνη (C) να ενώνεται πάντοτε με τη γουανίνη (G) με τρείς δεσμούς υδρογόνου. Η αλληλουχία των νουκλεοτιδίων είναι υπεύθυνη για την κωδικοποίηση και την αποθήκευση της πληροφορίας όπου κάθε μία από τις βάσεις μπορεί να θεωρηθεί ως ένα γράμμα από ένα αλφάβητο τεσσάρων γραμμάτων που χρησιμοποιείται για την αποθήκευση της χημικής πληροφορίας, για το πως εκφράζεται μία πρωτεΐνη. Η διαδικασία με την οποία αναπαράγεται το DNA ονομάζεται αντιγραφή [2].

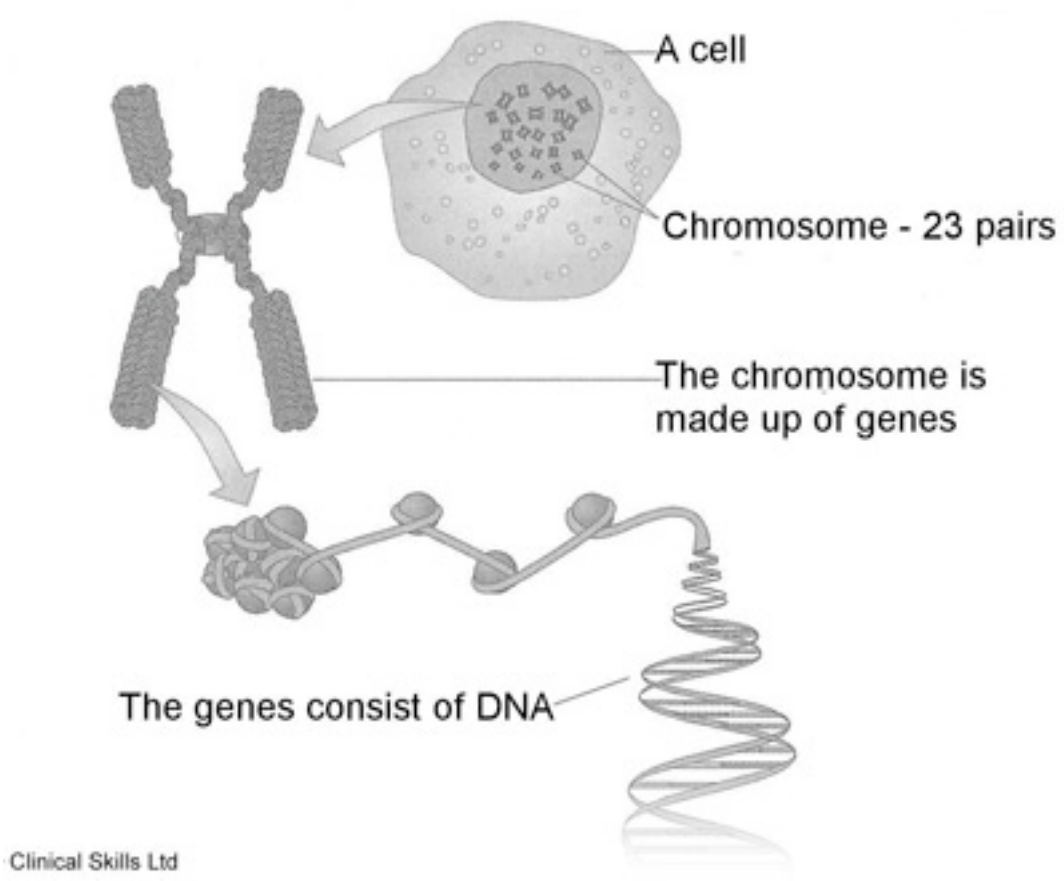


Εικόνα 2.1 Η δομή του DNA παρθέν από 'U.S. National Library of Medicine' (http://ghr.nlm.nih.gov/handbook/illustrations/dnastructure.jpg)

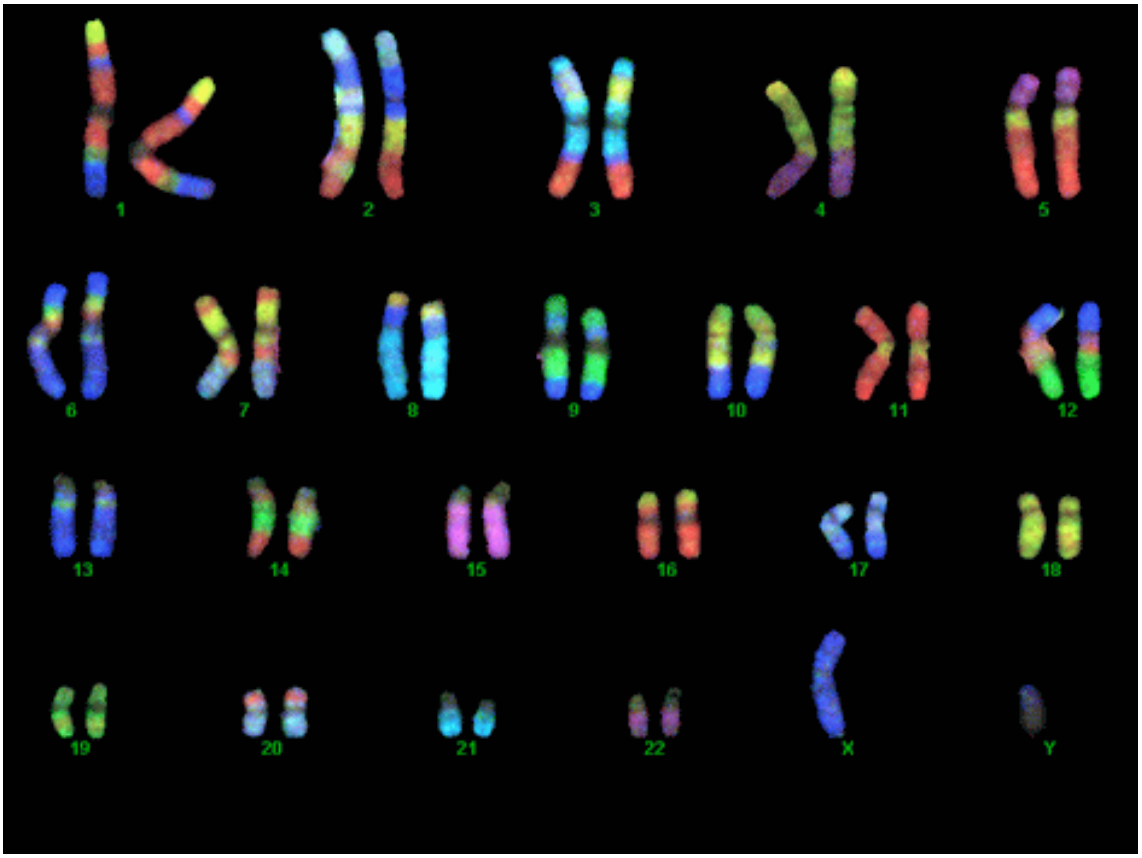
2.2 Χρωμοσώματα

Τα χρωμοσώματα αποτελούν οργανωμένες δομές DNA ή διαφορετικά, δομές που σχηματίζουν πακέτα DNA. Κάθε ένα από αυτά περιέχει ένα πάρα πολύ μακρύ μόριο DNA, που είναι πακεταρισμένο με τέτοιο τρόπο ώστε να έχει 50000 φορές μικρότερο μήκος. Στη διαδικασία δημιουργίας των χρωμοσωμάτων λαμβάνουν μέρος κάποιες ειδικές πρωτεΐνες που συνδέονται μαζί με το μακρομόριο του DNA σχηματίζοντας ένα σύμπλοκο που ονομάζεται χρωματίνη. Το βασικό στοιχείο από το οποίο αποτελείται η χρωματίνη είναι το νουκλεόσωμα, ένα πρωτεϊνικό οκταμερές που αποτελείται από τέσσερις διαφορετικούς τύπους πρωτεϊνών που ονομάζονται ιστόνες. Στον άνθρωπο κάθε σωματικό κύτταρο περιέχει δύο αντίγραφα από κάθε χρωμόσωμα από τα οποία το ένα προέρχεται από τον πατέρα και το άλλο από τη μητέρα. Τα δύο αυτά αντίγραφα σχηματίζουν ζεύγη χρωμοσωμάτων που ονομάζονται ομόλογα λόγω της ίδιας δομής και

οργάνωσης που παρουσιάζουν. Υπάρχουν 23 τέτοια ζεύγη ομόλογων χρωμοσωμάτων στον άνθρωπο, κάθε ένα από τα οποία είναι υπεύθυνο για την έκφραση ενός διαφορετικού χαρακτηριστικού του ανθρώπινου οργανισμού και ένα από αυτά είναι υπεύθυνο για το φύλο του ανθρώπου. Συγκεκριμένα το χρωμόσωμα που είναι υπεύθυνο για το φύλο είναι το χρωμόσωμα 23. Οφείλουμε να αναφέρουμε πως στην περίπτωση του αρσενικού ατόμου το 23^{ο} ζεύγος δεν αποτελεί ομόλογα χρωμοσώματα γιατί το ένα αποτελεί το X και το άλλο το Y. Σε περίπτωση που τα δύο αυτά χρωμοσώματα είναι ομόλογα δηλαδή XX τότε το άτομο αυτό είναι θηλυκό. Διαφορετικά όπως αναφέραμε και προηγούμενος με το χρωμόσωμα 23 να είναι XY τότε το άτομο είναι αρσενικό [2].



Εικόνα 2.2 Το χρωμόσωμα παρθέν από 'EuroGentest' (http://www.eurogentest.org/content/images/unit6/patientLeaflets/english/genesChromosomesDna.jpg)



Εικόνα 2.3 Τα 23 ζεύγη ομόλογων χρωμοσωμάτων παρθέν από 'EMERGENCE' (http://www.homodiscens.com/home/core_content/who_knows/astonishing_predicament/contingency/humilis/dangerous_algorithm/nature_self_aware/karyotype.gif)

2.3 Γονίδια

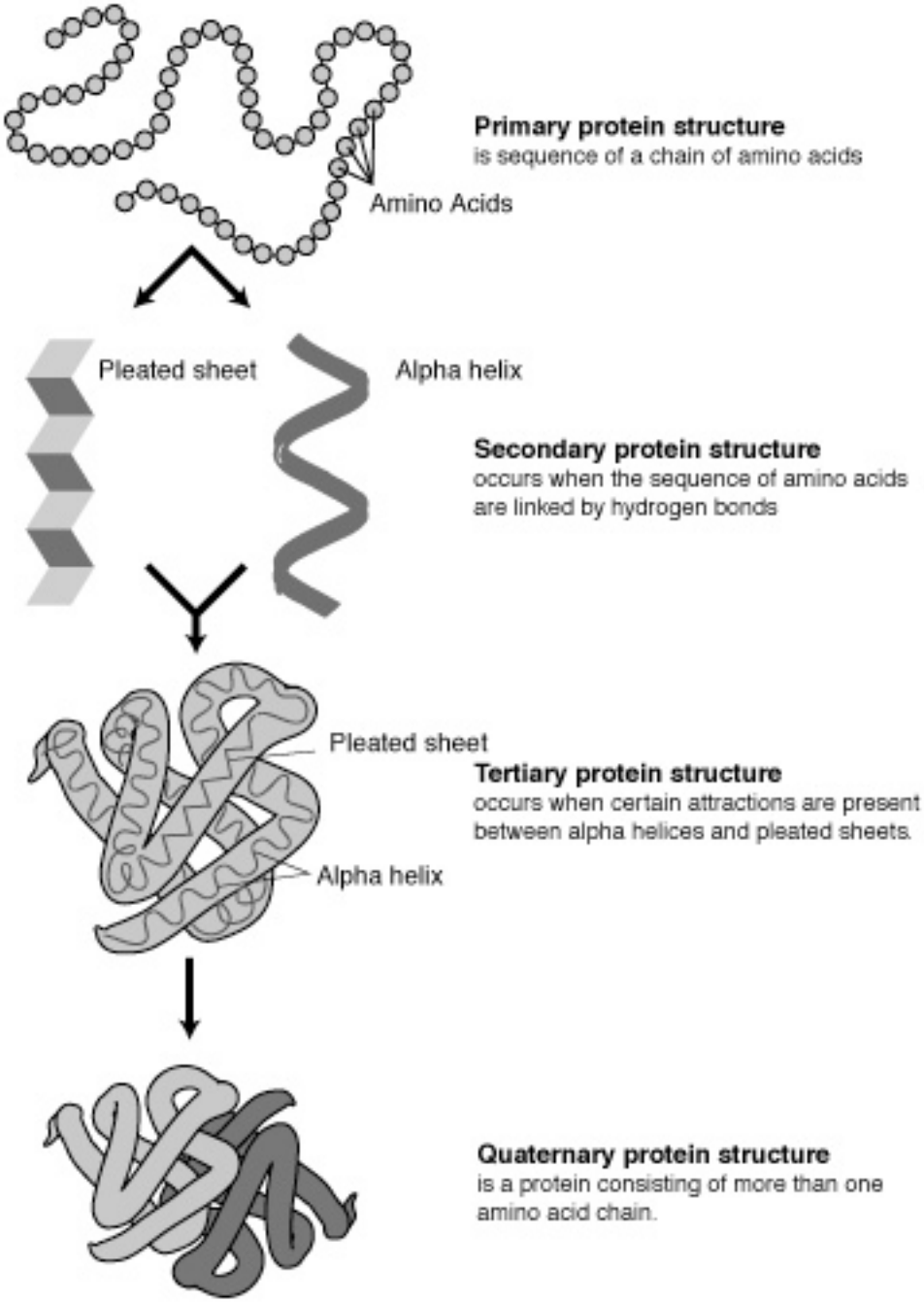
Τα γονίδια αποτελούν τμήματα DNA που μπορούν να μεταφραστούν σε ένα προϊόν πρωτεΐνης. Τα τμήματα αυτά του DNA που δεν μεταφράζονται δηλαδή που δεν παράγουν κάποιο προϊόν ονομάζονται ιντρόνια ενώ τα τμήματα που μεταφράζονται παράγοντας κάποιο προϊόν ονομάζονται εξώνια. Τα γονίδια αποτελούν εξώνια γιατί κάθε μετάφραση γονιδίου παράγει κάποιο προϊόν πρωτεΐνης [2].

Εικόνα 2.4 Το γονίδιο παρθέν από ‘Walgreens’
(http://www.walgreens.com/library/contents.html?docid=000437&doctype=10)

2.4 Πρωτεΐνες

Οι πρωτεΐνες αποτελούν βιολογικά μακρομόρια που παράγονται χρησιμοποιώντας τις πληροφορίες που βρίσκονται αποθηκευμένες στο γενετικό υλικό (DNA). Αποτελούν τα εργαλεία που κατασκευάζει ο οργανισμός για να επιβιώσει καθώς αποτελούν τα κύρια δομικά στοιχεία του οργανισμού. Το ανθρώπινο γενετικό υλικό περιέχει πληροφορίες για να κωδικοποιήσει 20-25 χιλιάδες διαφορετικές πρωτεΐνες. Το κύριο δομικό στοιχείο των πρωτεϊνών είναι τα αμινοξέα που αποτελούν απλά μόρια και υπάρχουν 20 διαφορετικά τέτοια μόρια. Ο αριθμός των μορίων αυτών σε μία πρωτεΐνη καθώς επίσης η σειρά με την οποία εμφανίζονται καθώς επίσης και οι μεταξύ τους αλληλεπιδράσεις, καθορίζουν και τη λειτουργία μίας πρωτεΐνης. Η ένωση των αμινοξέων σε μια πρωτεΐνη σχηματίζει έναν πεπτιδικό δεσμό. Καθώς μία πρωτεΐνη αποτελείται από πολλά

αμινοξέα. Ο αριθμός πεπτιδικών δεσμών σε μία πρωτεΐνη είναι μεγάλος καθώς αυτή μπορεί να αποτελείται από πολλά μόρια, για το λόγο αυτό οι πρωτεΐνες ονομάζονται και πολυπεπτίδια, λόγω του μεγάλου αριθμού πεπτιδικών δεσμών που μπορούν να περιέχουν. Υπάρχουν 4 διαφορετικά επίπεδα στα οποία μπορούμε να χωρίσουμε τις πρωτεΐνες όσον αφορά την οργάνωση τους. Οι κατηγορίες αυτές είναι η πρωτοταγής, δευτεροταγής, τριτοταγής και τεταρτοταγής δομή. Η πρωτοταγής δομή αποτελεί το πρώτο και βασικό επίπεδο στο οποίο έχουμε την αμινοξική αλληλουχία σε μία πρωτεΐνη. Στη συνέχεια η δευτεροταγής δομή των πρωτεϊνών καθορίζει τα τμήματα εκείνα της πολυπεπτιδικής αλυσίδας που σχηματίζουν γνωστά δομικά μοτίβα όπως είναι οι α-έλικες και τα β-πτυχωτά φύλλα. Η τριτοταγής δομή καθορίζει και το τρισδιάστατο σχήμα που έχει μία πρωτεΐνη αν αυτή αποτελείται από μία και μόνο πολυπεπτιδική αλυσίδα, ενώ η τεταρτοταγής δομή καθορίζει το τρισδιάστατο σχήμα μιας πρωτεΐνης που αποτελεί ένα σύμπλοκο, δηλαδή που αποτελείται από περισσότερες από μία πολυπεπτιδικές αλυσίδες. Το τελικό σχήμα της πρωτεΐνης είναι πολύ σημαντικό γιατί αυτό θα καθορίσει και την λειτουργία της πρωτεΐνης. Η λειτουργία παραγωγής των πρωτεϊνών ονομάζεται μετάφραση [2].



Εικόνα 2.5 Η δομή των πρωτεϊνών παρθέν από 'The Matc biotechnology project' (http://matcmadison.edu/biotech/resources/proteins/labManual/images/220_04_114.png)

2.5 Αμινοξέα

Τα αμινοξέα αποτελούν το κύριο δομικό στοιχείο των πρωτεϊνών. Υπάρχουν 20 διαφορετικά τέτοια στοιχεία. Όλα ανεξαρτήτως τα αμινοξέα έχουν την ίδια γενική δομή. Ένα κεντρικό άτομο άνθρακα (α-άνθρακας) που συνδέεται με ένα υδρογόνο, μια αμινομάδα, μια καρβοξυλομάδα και μία πλευρική αλυσίδα. Η αλυσίδα αυτή είναι που διαφοροποιεί τα αμινοξέα αλλάζοντας τις φυσικές και χημικές τους ιδιότητες, αλλάζοντας με αυτό τον τρόπο και τη λειτουργικότητα κάθε αμινοξέως. Βάση του φορτίου της πλευρικής τους αλυσίδας τα αμινοξέα διαχωρίζονται σε 4 διαφορετικές κατηγορίες: όξινα (αρνητικό φορτίο), βασικά (θετικό φορτίο), καθώς επίσης τα πολικά αμινοξέα που δεν έχουν φορτίο, όμως η πλευρική τους αλυσίδα έχει 2 περιοχές με διαφορετικό φορτίο, αλλά και τα μη πολικά αμινοξέα που δεν έχουν φορτίο. Σε φυσιολογικές συνθήκες pH που επικρατούν μέσα στο κύτταρο τα αμινοξέα έχουν την αμινομάδα τους φορτισμένη θετικά και την καρβοξυλομάδα τους φορτισμένη αρνητικά. Το γεγονός αυτό είναι που τα κάνει να συμπεριφέρονται σαν δίπολα. Η συμπεριφορά αυτή ευκολύνει τη σύνδεση των αμινοξέων μεταξύ τους, σχηματίζοντας πεπτιδικούς δεσμούς και τον σχηματισμό των πρωτεϊνών. Η σύνδεση των αμινοξέων γίνεται με τη σύνδεση του καρβοξυτελικού άκρου του αμινοξέως που προηγείται με την αμινομάδα του αμινοξέως που ακολουθεί, σχηματίζοντας με αυτό τον τρόπο έναν πεπτιδικό δεσμό. Κατά την ένωση δύο αμινοξέων και το σχηματισμό ενός πεπτιδικού δεσμού έχουμε ταυτόχρονα και την αποβολή ενός μορίου νερού. Τα άτομα που συμμετέχουν στο σχηματισμό του πεπτιδικού δεσμού έχουν τον περιορισμό ότι πρέπει όλα να βρίσκονται στο ίδιο επίπεδο, περιορίζοντας το εύρος κινήσεων των αμινοξέων που εμφανίζονται σε μία πρωτεΐνη, καθώς δεν μπορούν να περιστραφούν γύρω από τον εαυτό τους. Αφού τα αμινοξέα ενσωματωθούν στην πρωτεΐνη, το μόνο φορτίο που τους μένει είναι αυτό της πλευρικής τους αλυσίδας. Ωστόσο το αμινοτελικό και καρβοξυτελικό άκρο μιας πολυπεπτιδικής αλυσίδας (πρωτεΐνης) παραμένουν θετικά και αρνητικά φορτισμένα αντίστοιχα [2].

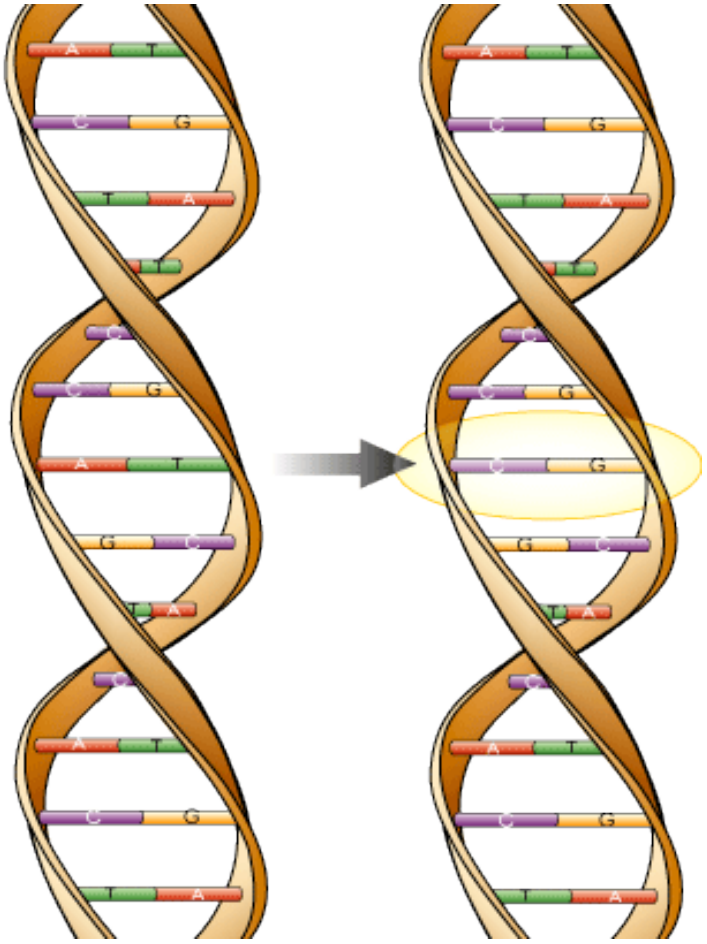
Amino Acid	SLC	DNA codons
Isoleucine	I	ATT, ATC, ATA
Leucine	L	CTT, CTC, CTA, CTG, TTA, TTG
Valine	V	GTT, GTC, GTA, GTG
Phenylalanine	F	TTT, TTC
Methionine	M	ATG
Cysteine	C	TGT, TGC
Alanine	A	GCT, GCC, GCA, GCG
Glycine	G	GGT, GGC, GGA, GGG
Proline	P	CCT, CCC, CCA, CCG
Threonine	T	ACT, ACC, ACA, ACG
Serine	S	TCT, TCC, TCA, TCG, AGT, AGC
Tyrosine	Y	TAT, TAC
Tryptophan	W	TGG
Glutamine	Q	CAA, CAG
Asparagine	N	AAT, AAC
Histidine	H	CAT, CAC
Glutamic acid	E	GAA, GAG
Aspartic acid	D	GAT, GAC
Lysine	K	AAA, AAG
Arginine	R	CGT, CGC, CGA, CGG, AGA, AGG
Stop codons	Stop	TAA, TAG, TGA

Πίνακας 2.1 Τα αμινοξέα και τα κωδικόνια τους

2.6 Σημειακοί Νουκλεοτιδικοί Πολυμορφισμοί (SNP)

Σημειακοί νουκλεοτιδικοί πολυμορφισμοί (single nucleotide polymorphism SNP), είναι παραλλαγές που παρατηρούνται στις ακολουθίες του DNA και οι οποίες εμφανίζονται όταν ένα νουκλεοτίδιο (A,T,C ή G) στην ακολουθία του γονιδιώματος αλλαχθεί. Για παράδειγμα ένα SNP θα μπορούσε να αλλάξει μια ακολουθία DNA από AAGGCTAA σε ATGGCTAA. Τα SNPs καλύπτουν 90% των γενετικών παραλλαγών που παρατηρούνται στον άνθρωπο. Ένας τέτοιος πολυμορφισμός παρατηρείται κάθε 100 μέχρι και 300 βάσεις κατά μήκος των 3 δισεκατομμυρίων βάσεων του ανθρώπινου γονιδιώματος. Στο παράδειγμα που προαναφέρθηκε, τα δύο διαφορετικά νουκλεοτίδια που παρουσιάζονται στην ίδια θέση στις δύο νουκλεοτιδικές ακολουθίες, θεωρούνται

αλληλόμορφα. Τα περισσότερα κοινά SNPs αποτελούνται από 2 μόνο αλληλόμορφα. Οι τέσσερεις διαφορετικές περιπτώσεις SNPs είναι ΑΑ, Αα, αα και η τέταρτη περίπτωση αυτή στην οποία παρατηρούνται missing data δηλαδή έλλειψη δεδομένων, διαφορετικά έλλειψη νουκλεοτιδίων σε κάποιες θέσεις των ακολουθιών DNA. Τα SNPs μπορούν να εμφανιστούν και στα εξώνια αλλά και στα ιντρόνια. Δηλαδή μπορούν να εμφανιστούν σε περιοχές του DNA που κωδικοποιούν κάποιο προϊόν αλλά και σε αυτές τις περιοχές που δεν κωδικοποιούν κάποια πρωτεΐνη ή κάποιο μόριο mRNA. Πολλά από τα SNPs δεν επηρεάζουν την λειτουργία του κυττάρου όμως πιστεύεται πως κάποια άλλα θα μπορούσαν να δημιουργήσουν την προδιάθεση στον άνθρωπο να ασθενήσει ή ακόμη να επηρεάσουν την αντίδραση του οργανισμού τους σε κάποιο φάρμακο.



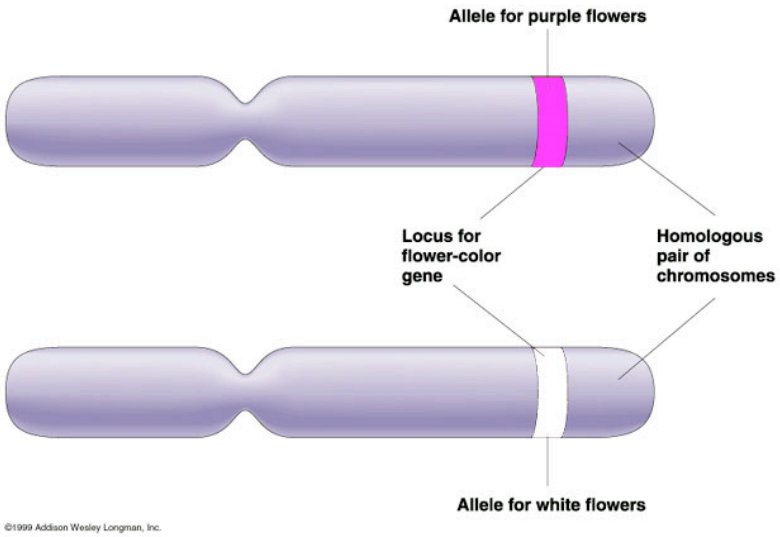
Εικόνα 2.6 Σημεικός Νουκλεοτιδικός Πολυμορφισμός (SNP) παρθέν από 'The Science Creative Quarterly' (http://www.scq.ubc.ca/wp-content/uploads/2006/07/dna1.gif)

2.7 Linkage Disequilibrium (LD)

Ανισορροπία συνδέσμων (linkage disequilibrium - LD) η κατάσταση στην οποία κάποιοι συνδυασμοί αλληλόμορφων γονιδίων ή γενετικών δεικτών παρατηρούνται σε μεγαλύτερο ή λιγότερο συχνά σε ένα πληθυσμό, απ' ότι θα αναμενόταν από τον τυχαίο σχηματισμό των απλότυπων (haplotypes) των αλληλόμορφων γονιδίων, που βασίζεται στη συχνότητα τους. Αποτελούν μη τυχαίες συσχετίσεις μεταξύ πολυμορφισμών που παρατηρούνται σε διαφορετικούς γεωμετρικούς τόπους. Αιτία αυτής της κατάστασης είναι η παρουσία γενετικών συνδέσμων, δηλαδή γενετικές περιοχές στις οποίες το ποσοστό ανασυνδυασμού που μπορεί να προκύψει παρατηρείται να είναι σταθερό σε ένα πληθυσμό και διαφορετικό από το τι θα αναμενόταν αν οι συνδυασμοί ήταν τυχαίοι [1].

2.8 Αλληλόμορφα Γονίδια

Ένα αλληλόμορφο γονίδιο είναι μια συγκεκριμένη ακολουθία νουκλεοτιδίων που μπορεί να έχει ένα γονίδιο από ένα σύνολο ν γνωστών πιθανών ακολουθιών. Σαν παράδειγμα, ας θεωρήσουμε ότι μονό ένα γονίδιο ευθυνόταν για το χρώμα των ματιών. Τότε διαφορετικά αλληλόμορφα αυτού του γονιδίου θα ευθύνονταν για το κάθε πιθανό χρώμα ματιών. Τα αλληλόμορφα γονίδια μπορεί να είναι τόσο σε περιοχές DNA οι οποίες κωδικοποιούν μια ακολουθία mRNA (εξώνια), , αλλά υπάρχουν περιπτώσεις στις οποίες μπορούν να αποτελούν και περιοχές του DNA που να μην κωδικοποιούν μια ακολουθία mRNA (ιντρόνια).



Εικόνα 2.7 Τα αλληλόμορφα γονίδια παρθέν από 'Science of Heredity' (http://porpax.bio.miami.edu/~cmallery/150/mendel/allele.jpg)

2.9 Έκφραση Γονιδίων (Gene Expression)

Είναι η διαδικασία στην οποία μια ακολουθία νουκλεοτιδίων DNA αντιγράφεται και μετατρέπεται σε ένα λειτουργικό γονιδιακό προϊόν όπως μία πρωτεΐνη ή ένα μόριο RNA κατά τις λειτουργίες της μετάφρασης και μεταγραφής αντίστοιχα [13,11,7].

2.10 Έκφραση Πρωτεϊνών (Protein Expression)

Έκφραση πρωτεϊνών αποτελεί ένα τμήμα της διαδικασίας έκφρασης γονιδίων. Περιλαμβάνει τα στάδια στα οποία το DNA έχει ήδη μεταφραστεί σε αμινοξικές αλυσίδες, που στη συνέχεια θα αναδιπλωθούν στο χώρο σχηματίζοντας την δευτεροταγή, τριτοταγή ή τεταρτοταγή δομή και τον τελικό σχηματισμό της πρωτεΐνης.

2.11 Κωδικόνια

Αποτελούν τριάδες βάσεων συγκεκριμένων ακολουθιών που κάθε ένα από αυτά αντιπροσωπεύει κάποιο συγκεκριμένο αμινοξύ. Ένα αμινοξύ είναι δυνατό να αντιπροσωπεύεται από περισσότερα από ένα κωδικόνια, έτσι διαφορετικές ακολουθίες βάσεων του μορίου του mRNA μπορούν να μεταφράζονται στο ίδιο προϊόν πρωτεΐνης.

Υπάρχουν 64 διαφορετικά κωδικόνια από τα οποία 3 αποτελούν κωδικόνια λήξης, που προσδιορίζουν και το τέλος μιας μεταφραζόμενης περιοχής και ένα κωδικόνιο που προσδιορίζει την αρχή μιας μεταφραζόμενης περιοχής (κωδικόνιο έναρξης). Το κωδικόνιο έναρξης είναι το AUG, ενώ τα κωδικόνια λήξης είναι τα UAG, UAA και UGA. Για παράδειγμα το κωδικόνιο CGU αποτελεί ένα από τα τέσσερα διαφορετικά κωδικόνια που κωδικοποιούν το αμινοξύ αργινίνη [2].

GCA GCC GCG GCU	AGA AGG CGA CGC CGG CGU	GAC GAU	AAC AAU	UGC UGU	GAA GAG	CAA CAG	GGA GGC GGG GGU	CAC CAU	AUA AUC AUU	UUA UUG CUA CUC CUG CUU	AAA AAG	AUG	UUC UUU	CCA CCC CCG CCU	AGC AGU UCA UCC UCG UCU	ACA ACC ACG ACU	UGG	UAC UAU	GUA GUC GUG GUU
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V

Εικόνα 2.8 Τα 64 κωδικόνια παρθέν από ‘Center of BioMolecular Modeling - CBM’ (http://www.rpc.msoe.edu/cbm2/images/gfp/gfp3-2.jpg)

2.12 mRNA

Αποτελεί το μόριο RNA που μεταφέρει τις γενετικές πληροφορίες από το DNA στο ριβόσωμα όπου εκεί θα γίνει η μετάφραση του στο προϊόν πρωτεΐνης. Το mRNA κωδικοποιεί τις αντίστοιχες ακολουθίες με νουκλεοτίδια όπως ακριβός και το DNA απλώς κάθε θέση του νουκλεοτιδίου της θιμίνης (T) αντικαθιστάται με την ουρακίλη (U) και αντίθετα με το DNA δεν είναι δίκλωνο αλλά μονόκλωνο [2].

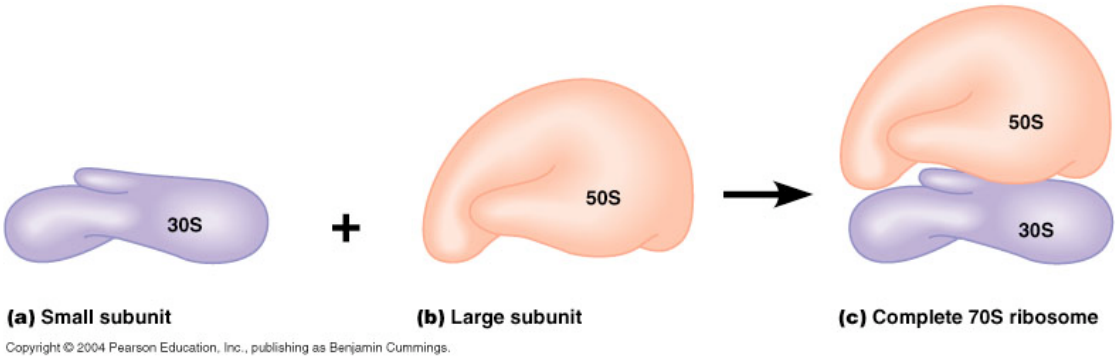
2.13 tRNA

Transfer RNA (tRNA) αποτελεί ένα μικρό μόριο RNA μήκους περίπου 74-95 νουκλεοτίδια, που μεταφέρει κάθε φορά ένα συγκεκριμένο αμινοξύ στο ριβόσωμα, για να ενσωματωθεί στην σχηματιζόμενη πολυπεπτιδική αλυσίδα, κατά την πρωτεϊνοσύνθεση όταν βρίσκεται σε εξέλιξη η διαδικασία της μετάφρασης του RNA. Διαθέτει ένα 3΄άκρο στο οποίο συνδέεται το αμινοξύ. Επίσης περιέχει μια περιοχή μήκους τριών βάσεων που ονομάζεται αντικωδικόνιο, που ζευγαρώνει με την αντίστοιχη περιοχή μήκους τριών βάσεων που βρίσκεται πάνω στο μόριο του mRNA. Κάθε τύπος μορίου tRNA μπορεί να προσδεθεί σε ένα μόνο τύπο αμινοξέως. Λόγω του

γεγονότος ότι ο γενετικός κώδικας περιέχει πολλαπλά κωδικόνια που καθορίζουν το ίδιο αμινοξύ, τα μόρια tRNA με διαφορετικά αντικωδικόνια, που όμως αντιστοιχούν στο ίδιο αμινοξύ, μπορούν να μεταφέρουν το ίδιο αμινοξύ στο οποίο και αντιστοιχούν. Υπάρχουν 31 διαφορετικά tRNAs [2].

2.14 Ριβόσωμα

Αποτελεί το εργοστάσιο σύνθεσης πρωτεϊνών στο κύτταρο. Βρίσκεται στο κυτόπλασμα και αποτελείται από 65% rRNA και 35% πρωτεΐνες. Το οργανίδιο αυτό είναι ένας από τους πιο σύνθετους μοριακούς μηχανισμούς του κυττάρου και αποτελεί ένα μικρό μόνο μέρος του συνολικού δικτύου μηχανισμών που απαιτούνται ώστε να γίνει με επιτυχία η πρωτεϊνοσύνθεση. Αποτελείται από δύο υπομονάδες (τη μικρή και τη μεγάλη υπομονάδα) κάθε μία από τις οποίες είναι ένα τεράστιο σύμπλοκο πρωτεϊνών και RNA. Όταν οι δύο αυτές υπομονάδες είναι ενωμένες τότε σχηματίζουν το πλήρες ριβόσωμα που φέρει τέσσερεις θέσεις σύνδεσης με το RNA. Οι τρείς από τις θέσεις αυτές συνδέονται με tRNAs και η τέταρτη με το mRNA που μεταφράζεται [2].



Εικόνα 2.9 Το ριβόσωμα παρθέν από 'Hunter College of The City University New York' (http://diverge.hunter.cuny.edu/~weigang/Images/04-19_ribosome_1.jpg)

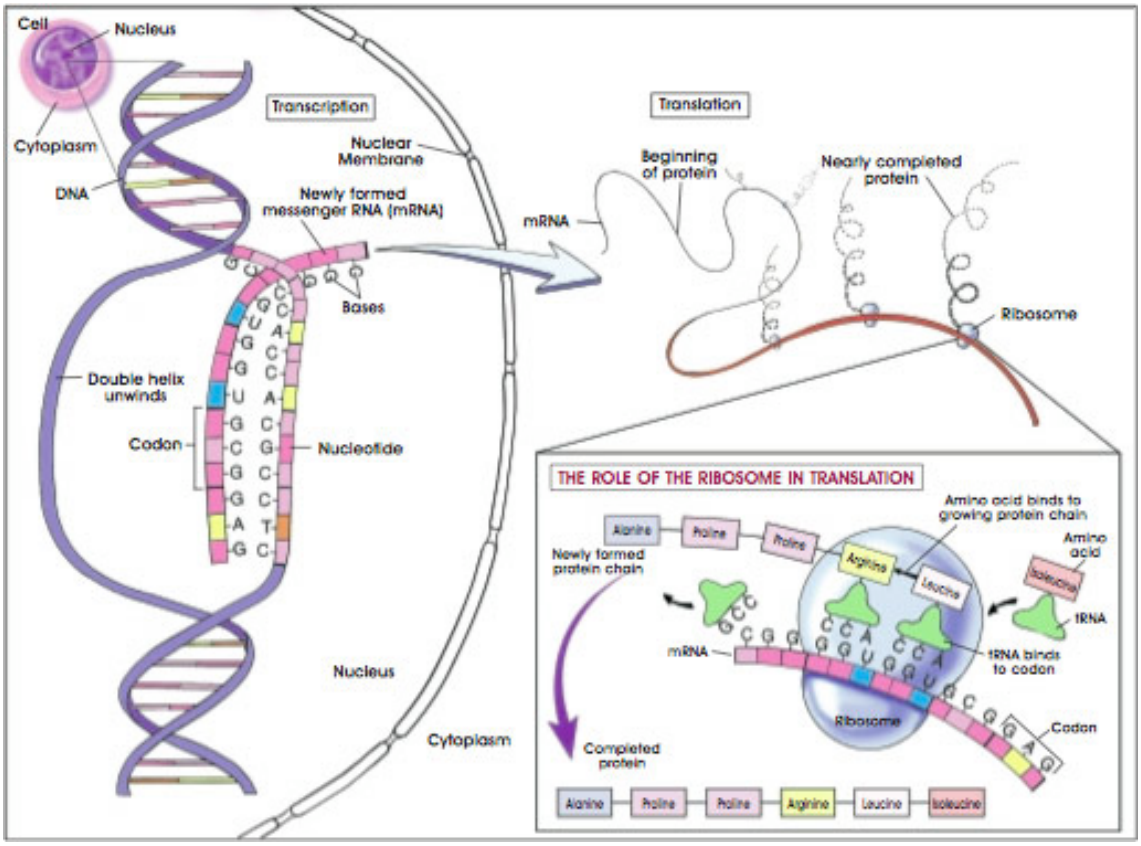
2.15 Μεταγραφή

Μεταγραφή είναι η διαδικασία στην οποία η πληροφορία που περιέχεται στο μόριο του DNA, μεταγράφεται σε ένα μόριο mRNA που αυτό με τη σειρά του θα καθορίσει την ακολουθία αμινοξέων της πρωτεϊνικής δομής. Η διαδικασία αυτή ξεκινά με τον διπλασιασμό του DNA όπου το μόριο του DNA διαχωρίζεται σε δύο ξεχωριστούς

κλώνους των οποίων οι βάσεις είναι συμπληρωματικές. Πιο συγκεκριμένα η διαδικασία αυτή μπορεί να ονομαστεί και σύνθεση RNA καθώς η νουκλεοτιδική ακολουθία DNA μεταγράφεται – τροποποιείται σε RNA πληροφορία. Και των δύο μορίων οι νουκλεοτιδικές ακολουθίες χρησιμοποιούν συμπληρωματική γλώσσα, έτσι αυτός είναι και ο λόγος που η πληροφορία απλά μεταγράφεται ή πιο εύκολα αντιγράφεται από το ένα μόριο στο άλλο. Η διαδικασία μεταγραφής του DNA είναι παρόμοια όπως και η αντιγραφή του με τη διαφορά ότι σε αυτή την περίπτωση συμμετέχουν διαφορετικά ένζυμα και το ταίριασμα των βάσεων αδενίνης που βρίσκονται στο DNA με ουρακίλη U που είναι μία από τις βασικές διαφορές του μορίου του DNA από το μόριο του RNA. Έτσι το μόριο mRNA που παράγεται είναι το ίδιο με τον συμπληρωματικό κλώνο DNA με τη μόνη διαφορά ότι όπου στο DNA υπήρχαν θυμίνες T στο μόριο του mRNA θα υπάρχου ουρακίλες U. Το mRNA μεταγράφεται από την RNA πολυμεράση, ένα ένζυμο που προσδένεται στο ένα μονόκλωνο DNA, για να δημιουργήσει με αυτόν τον τρόπο το συμπληρωματικό κλώνο RNA που ονομάζεται messenger RNA – mRNA. Η ονομασία του προέρχεται από το γεγονός ότι μεταφέρει ένα γενετικό μήνυμα ή διαφορετικά τη γενετική πληροφορία από το DNA στο μηχανισμό πρωτεϊνοσύνθεσης του κυττάρου, το ριβόσωμα. Η διαδικασία αυτή είναι το πρώτο στάδιο που οδηγεί στην έκφραση γονιδίων (gene expression) με την παραγωγή του ενδιάμεσου μορίου mRNA, που αποτελεί ένα πιστό μετάγραφο της πληροφορίας του γονιδίου που κωδικοποιεί τη σύνθεση κάποιας πρωτεΐνης. Το συγκεκριμένο τμήμα DNA που μεταγράφεται σε mRNA ονομάζεται μετάγραφο. Αυτό το συγκεκριμένο τμήμα, το μετάγραφο, περιέχει αλληλουχίες νουκλεοτιδικών βάσεων που όχι μόνο κωδικοποιούν την αλληλουχία που μεταφράζεται αλλά επιπλέον κατευθύνει και ρυθμίζει την πρωτεΐνοσύνθεση. Αυτό γίνεται καθώς ο αριθμός πρωτεϊνών που θα παραχθούν εξαρτάται από τον αριθμό των μορίων mRNA που έχει στη διάθεση του το κύτταρο για μετάφραση. Η μεταγραφή γίνεται χρησιμοποιώντας τον 3΄- 5΄ κλώνο του DNA έτσι ώστε το μόριο mRNA που θα πάρουμε να έχει κατεύθυνση 5΄- 3΄ για να μπορέσει να χρησιμοποιηθεί στη συνέχεια στη διαδικασία της μετάφρασης, στην οποία θα γίνει και η πρωτεϊνοσύνθεση [2].

2.16 Μετάφραση

Μετάφραση είναι η διαδικασία κατά την οποία ένα μόριο ή περισσότερα μόρια mRNA μεταφράζονται σε ένα η περισσότερα μόρια πρωτεΐνης, ανάλογα με την ακολουθία των βάσεων που μεταφέρει το μόριο αυτό. Κάθε τρείς από τις βάσεις του μορίου του mRNA αντιπροσωπεύουν ένα αμινοξύ. Τα περισσότερα από αυτά εκφράζονται από περισσότερα από ένα κωδικόνια. Αυτός είναι και ο κύριος λόγος για τον οποίο διαφορετικές ακολουθίες βάσεων μπορούν οδηγήσουν στην παραγωγή της ίδιας πρωτεΐνης. Σε αυτή τη διαδικασία λαμβάνουν μέρος δύο διαφορετικά είδη RNA, το tRNA που μεταφέρει το αμινοξύ στην παραγόμενη πολυπεπτιδική αλυσίδα, και το mRNA που μεταφέρει τις πληροφορίες που θα μεταφραστούν ώστε να παραχθεί το νέο προϊόν. Η διαδικασία εκτελείται στην περιοχή του κυτοπλάσματος όπου και βρίσκονται τα ριβοσώματα. Τα ριβοσώματα αποτελούνται από τη μεγάλη και τη μικρή υπομονάδα και στο τέλος όταν θα αρχίσει η διαδικασία, ολόκληρο το ριβόσωμα πλαισιώνει το μόριο του mRNA. Η διαδικασία ξεκινά με το εναρκτήριο tRNA που φέρει τη μεθειονίνη που συνδέεται με τη μικρή ριβοσωμική υπομονάδα και έπειτα το σύμπλοκο αυτό, ενώνεται στο 5΄άκρο του mRNA που στην συνέχεια θα ενωθεί με μία σειρά από πρωτεΐνες που ονομάζονται παράγοντες έναρξης που βοηθούν στην έναρξη της διαδικασίας. Στη συνέχεια όλο το σύμπλοκο αρχίζει να κινείται προς το 3΄άκρο του mRNA, μέχρι να συναντήσει το πρώτο AUG που αποτελεί το κωδικόνιο έναρξης, οπότε αποδεσμεύονται οι παράγοντες έναρξης και συγχρόνως ενώνεται η μικρή ριβοσωμική υπομονάδα με τη μεγάλη πλαισιώνοντας έτσι το μόριο του mRNA. Ακολούθως όλο το σύμπλοκο – ριβόσωμα κινείται προς το 3΄άκρο του mRNA, διαβάζοντας τα κωδικόνια και προσθέτοντας τα κατάλληλα αμινοξέα στην συνεχώς αυξανόμενη πολυπεπτιδική αλυσίδα, μέχρι να συναντήσει τα κωδικόνια λήξης. Υπάρχουν τρία διαφορετικά κωδικόνια που υποδεικνύουν τη λήξη μιας περιοχής που κωδικοποιεί ένα προϊόν και αυτά είναι τα UAG, UAA και UGA. Τα κωδικόνια λήξης δεν αναγνωρίζονται από κάποιο tRNA αλλά ειδοποιούν το ριβόσωμα ότι πρέπει να σταματήσει την πρωτεϊνοσύνθεση και να διασπαστεί. Η πεπτιδυλοτρανσφεράση είναι το ένζυμο που αποσυνδέει τα αμινοξέα από το tRNA και τα συνδέει στο καρβοξυτελικό άκρο της νεοσυντιθέμενης πολυπεπτιδικής αλυσίδας. Όταν η πρωτεϊνοσύνθεση φτάσει στο τέλος, η πρωτεΐνη ελευθερώνεται στο κυτόπλασμα ώστε να χρησιμοποιηθεί από το κύτταρο [2].



Εικόνα 2.10 Οι λειτουργίες της Μεταγραφής και της Μετάφρασης παρθέν από 'National Institutes of Health – Stem Cell Information (http://stemcells.nih.gov/StaticResources/info/scireport/images/figurea6.jpg)

2.17 Cis – Ενεργοποιητικά στοιχεία (acting elements)

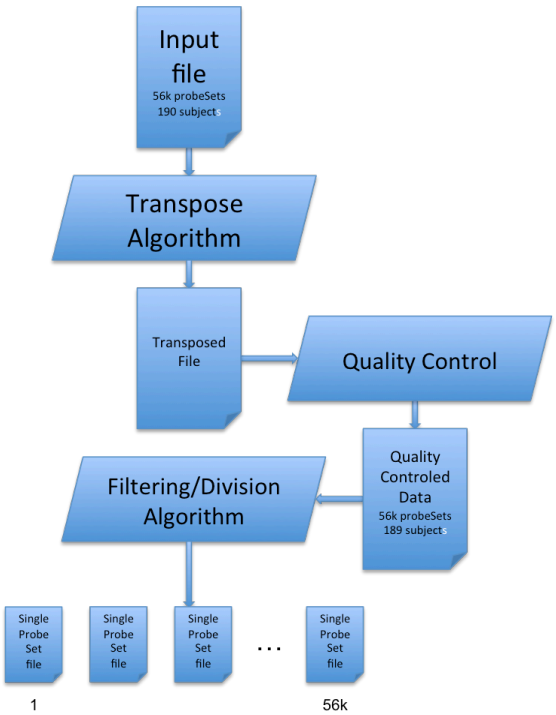
Τα στοιχεία ενεργοποιητές, αποτελούν συνήθως ακολουθίες DNA που εμφανίζονται στο δομικό μέρος ενός γονιδίου και είναι απαραίτητα για την έκφραση του γονιδίου αυτού. Πάνω σε αυτές τις ακολουθίες προσδένονται οι ενεργοποιητικοί παράγοντες για να υποβοηθήσουν την λειτουργία της έκφρασης ενός γονιδίου.

2.18 Trans – Ενεργοποιητικοί παράγοντες (acting factors)

Οι ενεργοποιητικοί παράγοντες συνήθως αποτελούν πρωτεΐνες που προσδένονται πάνω στις cis ενεργοποιητικές ακολουθίες DNA και υποβοηθούν και ελέγχουν την έκφραση κάποιου γονιδίου.

Κεφάλαιο 3

Προεπεξεργασία [6]



Διάγραμμα 3.1 Προεπεξεργασία

3.1 Δεδομένα Έκφρασης mRNA

Τα δεδομένα έκφρασης mRNA παράχθηκαν για το σύνολο των 190 δειγμάτων εκ των οποίων τα 64 ήταν φυσιολογικά δείγματα ενώ τα υπόλοιπα 126 δείγματα προέρχονταν από ασθενείς με Μονοπολική Κατάθλιψη. Η μέθοδος που χρησιμοποιήθηκε για την παραγωγή των δεδομένων έκφρασης mRNA είναι η μέθοδος Affymetrix HU133 v.2 GeneChips. Σε προηγούμενη έρευνα, probeSets που παρουσίασαν σημαντικές διαφορές στην έκφραση μεταξύ των δειγμάτων που προέρχονταν από υγιή άτομα και αυτών που προέρχονταν από ασθενείς, όσο αφορά τα στατιστικά στοιχεία που μελετήθηκαν, έγινε ανάλυση των διαφορών που βρέθηκαν (analysis of variance).

3.2 Δεδομένα Έκφρασης Πρωτεϊνών

Η πηγή που χρησιμοποιήθηκε για τη συλλογή των δεδομένων έκφρασης των πρωτεϊνών είναι η ιστοσελίδα του Rules Based Medicine στο διαδίκτυο που αποτελεί ένα εργαστήριο βιολογικών δεικτών με μία συλλογή από αναπαραγωγίσιμα, ποσοτικά δεδομένα. Συγκεκριμένα τα δεδομένα συλλέχτηκαν από τη βάση Human Map 1.5 που αποτελείται από ένα σύνολο 89 πρωτεϊνικών εκφράσεων, για τις οποίες διατίθεται το όνομα, η περιγραφή του κάθε διαφορετικού δείγματος καθώς επίσης και ένας μοναδικός αριθμός που καθορίζει την κάθε μία από αυτές. Η λίστα δειγμάτων των πρωτεϊνικών εκφράσεων παρουσιάζεται πιο κάτω:

Α/Α	Πρωτεϊνη	Α/Α	Πρωτεϊνη
1	Adiponectin	46	Interleukin-3
2	Alpha-1 Antitrypsin	47	Interleukin-4
3	Alpha-Fetoprotein	48	Interleukin-5
4	Alpha-2 Macroglobulin	49	Interleukin-6
5	Apolipoprotein A-1	50	Interleukin-7
6	Apolipoprotein C-III	51	Interleukin-8
7	Apolipoprotein H	52	Interleukin-10
8	Beta-2 Microglobulin	53	Interleukin-12 p40
9	BDNF	54	Interleukin-12 p70
10	C-Reactive Protein	55	Interleukin-13
11	Calcitonin	56	Interleukin-15
12	Cancer Antigen 19-9	57	Interleukin-16
13	Cancer Antigen 125	58	Leptin
14	Carcinoembryonic Antigen	59	Lipoprotein (a)
15	CD40	60	Lymphotactin
16	CD40 Ligand	61	MDC
17	Complement 3	62	MIP-1 alpha
18	CK-MB	63	MIP-1 beta
19	Endothelin-1	64	MMP-2
20	Eotaxin	65	MMP-3
21	Epidermal Growth Factor	66	MMP-9
22	ENA-78	67	MCP-1
23	Erythropoietin	68	Myeloperoxidase
24	ENRAGE	69	Myoglobin
25	Factor VII	70	PAI-1
26	Fatty Acid Binding Protein	71	PAPP-A
27	Ferritin	72	PSA, Free
28	Fibrinogen	73	Prostatic Acid Phosphatase
29	FGF-basic	74	RANTES
30	GST	75	Serum Amyloid P
31	G-CSF	76	SGOT
32	GM-CSF	77	Sex Hormone Binding Globulin
33	Growth Hormone	78	Stem Cell Factor
34	Haptoglobin	79	Thrombopoietin
35	Immunoglobulin A	80	Thyroid Binding Globulin
36	Immunoglobulin E	81	Thyroid Stimulating Hormone
37	Immunoglobulin M	82	Tissue Factor
38	Insulin	83	TIMP-1
39	IGF-1	84	Tumor Necrosis Factor-alpha
40	ICAM-1	85	Tumor Necrosis Factor-beta
41	Interferon-gamma	86	Tumor Necrosis Factor RII
42	Interleukin-1 alpha	87	VCAM-1
43	Interleukin-1 beta	88	VEGF
44	Interleukin-1 ra	89	von Willebrand Factor
45	Interleukin-2		

Πίνακας 3.1 Οι 89 Πρωτεΐνες

3.3 Δεδομένα Έκφρασης DNA

Οι γονότυποι (δεδομένα έκφρασης DNA) συλλέχτηκαν με την μέθοδο του illumina 550K. Η μέθοδος αυτή χρησιμοποιήθηκε έτσι ώστε να είναι εφικτή η συλλογή δεδομένων από όλο το γονιδίωμα. Εφαρμόστηκε πάνω σε 2000 υποψηφίους, 1000 ασθενείς (cases) και 1000 υγιείς (controls), εκ των οποίων και τα 190 δείγματα που αναλύθηκαν σε mRNA μεταφράζονται σε κάποιο πρωτεϊνικό προϊόν.

3.3.1 Προεπεξεργασία Δεδομένων Έκφρασης DNA

Στο συγκεκριμένο σημείο χρησιμοποιήθηκαν συγκεκριμένα περίπου 550 χιλιάδες SNPs που όπως ήδη αναφέρθηκε αρκετές φορές, προήλθαν από 190 δείγματα περιφερικού αίματος, τα οποία συλλέχτηκαν από μία βάση που περιέχει 126 ασθενείς με κατάθλιψη (Unipolar Depression) και 64 φυσιολογικά δείγματα (control samples). Από τα δείγματα αυτά τα 43 προέρχονται από άντρες ενώ τα υπόλοιπα 133 από γυναίκες.

Έγινε έλεγχος σε όλα τα δείγματα έτσι ώστε να διερευνηθεί, αν μεγαλύτερο ποσοστό από το 10% των δειγμάτων αυτών παρουσίαζαν έλλειψη δεδομένων για να αφαιρεθούν. Κανένα από τα SNPs δεν παρουσίαζε κάποια έλλειψη όσον αφορά τα δεδομένα. Επιπλέον έγινε έλεγχος της συχνότητας των αλληλόμορφων γονιδίων ώστε να είναι το μέγιστο 1%, διαφορετικά αυτά να αφαιρούνταν από το σύνολο. Κατά τη διαδικασία αυτή παρουσιάστηκαν περίπου 10 χιλιάδες SNPs που απέτυχαν τον έλεγχο αυτό, με αποτέλεσμα να αφαιρεθούν [9,14].

Ως αποτέλεσμα των πιο πάνω ελέγχων ήταν ο περιορισμός των δεδομένων από τα 550 χιλιάδες SNPs που ήταν διαθέσιμα αρχικά, σε έναν μικρότερο αριθμό αυτών [9].

3.4 Δεδομένα Εισόδου

Όπως έχει ήδη προαναφερθεί, τα δεδομένα προήλθαν από την ανάλυση αιματολογικών δειγμάτων, 190 ανθρώπων από τα οποία τα 124 προέρχονταν από ασθενείς με Μονοπολική Κατάθλιψη (Unipolar Depression) οι cases ενώ τα υπόλοιπα 64 αποτελούσαν υγιή δείγματα (controls).

Συγκεκριμένα τα 190 δείγματα χρησιμοποιήθηκαν για να ληφθούν δεδομένα για περίπου 56 χιλιάδες διαφορετικά probeSets. Τα δεδομένα για κάθε ένα probeSet και δείγμα, αποτελούσαν μία κανονικοποιημένη τιμή η οποία ήταν από το 0 μέχρι και το ∞.

Το μεγαλύτερο βάρος δίνεται στις μικρότερες τιμές καθώς μια τιμή p-value αντιστοιχεί στην πιθανότητα το αποτέλεσμα να είναι false positive, δηλαδή να έχει τη μεγαλύτερη πιθανότητα ώστε να είναι λανθασμένο. Όσο μικρότερη η τιμή/πιθανότητα αυτή τόσο και πιο μεγάλη η σημασία του δείγματος στα δεδομένα. Αργότερα δίνεται και ο ακριβής ορισμός της πιθανότητας σημαντικότητας (p-value) όπως ορίζεται στη στατιστική.

3.5 Δομή Αρχείου Δεδομένων

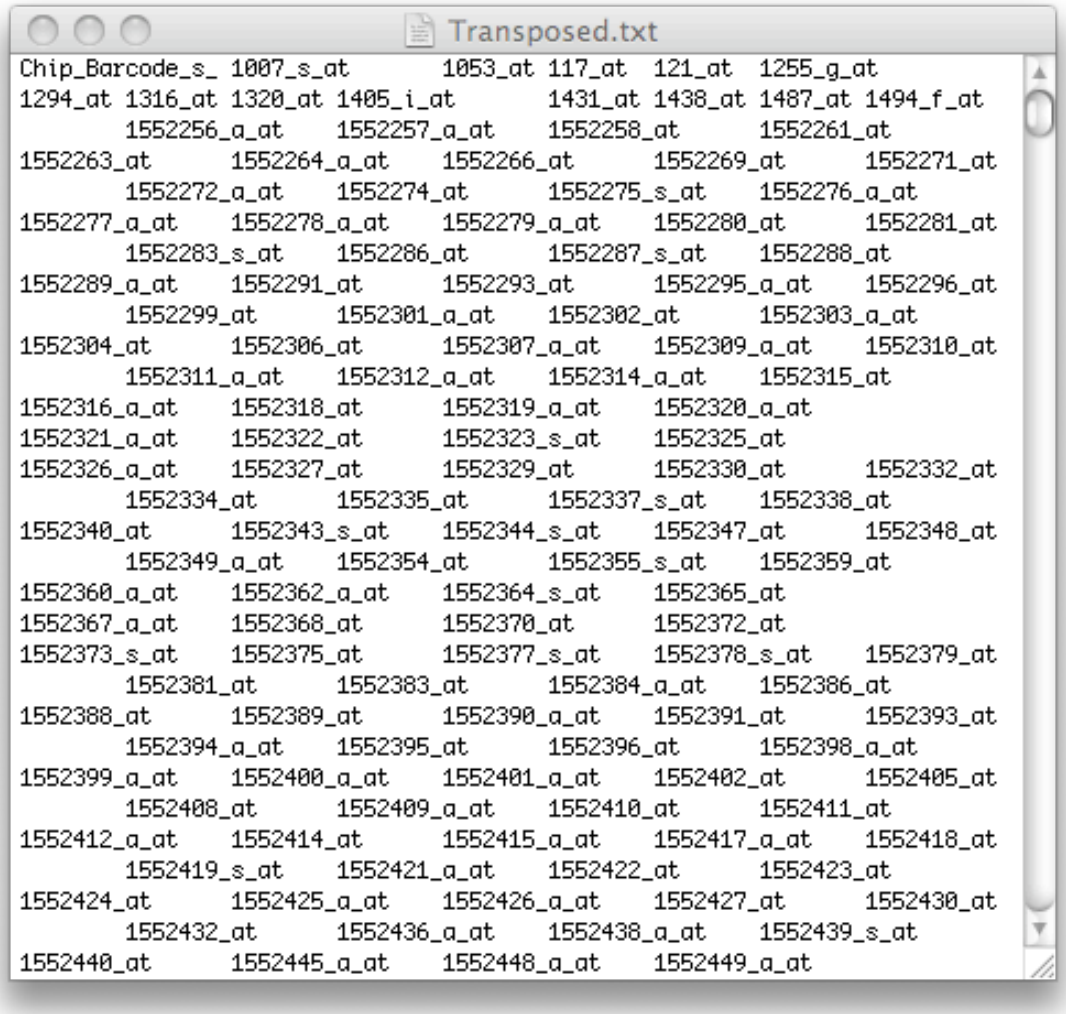
Σε αρχικό στάδιο τα δεδομένα περιέχονταν σε ένα αρχείο με περίπου 56 χιλιάδες στήλες, όσος και ο αριθμός των probeSets που μελετείται και 190 γραμμές που περιέχουν τα δεδομένα για κάθε άνθρωπο που συμμετείχε στο πείραμα. Για κάθε ένα από τα διαφορετικά probeSets υπήρχε μία τιμή διαφορετική για τον κάθε άνθρωπο ξεχωριστά.

3.6 Μετάθεση Αρχείου Δεδομένων (File Transpose)

Για την ανάλυση των δεδομένων του αρχείου αναγκαία ήταν η μετάθεση των περιεχομένων του, δηλαδή η μετατροπή των δεδομένων του από γραμμές σε στήλες και αντίστροφα. Δημιουργία δηλαδή ενός νέου αρχείου με τα ίδια δεδομένα όπως και το αρχικό, με τη διαφορά ότι βρίσκεται σε μετατεθειμένη μορφή (transposed).

Λόγω του μεγάλου όγκου των δεδομένων του αρχείου, ο πρώτος και πιο σημαντικός στόχος της μελέτης ήταν η διάσπαση του σε μικρότερα, ώστε η ανάλυση των δεδομένων να γίνει πιο εύκολη και πιο γρήγορη. Για να μειωθεί ο χρόνος εκτέλεσης χρησιμοποιήθηκε ένα πλέγμα υπολογιστών (grid) του οποίου η χρήση αποσκοπούσε στην παράλληλη εκτέλεση των αλγόριθμων που υλοποιήθηκαν. Με τον τρόπο αυτό

επιτεύχθηκε μείωση του συνολικού χρόνο που ήταν απαραίτητος για την ανάλυση και την επεξεργασία των δεδομένων.



Εικόνα 3.1 Το Μετατεθημένο Αρχείο

3.7 Διαδικασία Διάσπασης Δεδομένων

Λόγω του μεγάλου όγκου του αρχείου με τα δεδομένα, απαραίτητη ήταν η διάσπαση του σε άλλα αρχεία μικρότερου μεγέθους. Οι μέθοδοι που χρησιμοποιήθηκαν για την ανάλυση των δεδομένων, απαιτούσαν τεράστια ποσά χρόνου για την εκτέλεση τους. Επομένως η διάσπαση σε μικρότερα αρχεία, αποσκοπούσε κυρίως στη μείωση του χρόνου εκτέλεσης και διευκόλυνση της ανάλυσης των δεδομένων.

3.7.1 Ποιοτικός Έλεγχος Δεδομένων (Quality Control)

Ως μέρος της προετοιμασίας των δεδομένων έγινε έλεγχος ποιότητας (Quality Control) στα δεδομένα, ώστε να αφαιρεθούν τα δείγματα που παρουσίαζαν έλλειψη δεδομένων. Βρέθηκε ένα μόνο δείγμα με ελλειπή δεδομένα, το οποίο μετά από εισήγηση των ειδικών βιολόγων της ομάδας, αφαιρέθηκε από το σύνολο. Αυτή η διαδικασία εκτελέστηκε έτσι ώστε να είναι βέβαιο ότι τα δεδομένα είναι ακέραια και τα αποτελέσματα που θα προκύψουν από την ανάλυση θα είναι όσο το δυνατό πιο έγκυρα.

3.7.2 Διάσπαση Αρχείου Δεδομένων

Σαν δεύτερο στάδιο της επεξεργασίας των δεδομένων, ήταν η διάσπαση του μεγάλου αρχείου σε μικρότερα, διευκολύνοντας με αυτό τον τρόπο τη διαδικασία ανάλυσης των δεδομένων. Ο αριθμός των probeSets ανά αρχείο παραμετροποιήθηκε αφού αυτός όσο πιο μικρός ήταν, τόσο θα αυξανόταν το overhead της ανάλυσης, ενώ αντίθετα όσο πιο μεγάλος θα ήταν, θα αυξανόταν το μέγεθος των ενδιάμεσων δεδομένων που θα παράγονταν. Η διάσπαση έγινε αρχικά σε αρχεία μεγέθους των 10 probeSets ανά αρχείο. Με την ενέργεια αυτή δεν επιτεύχθηκε η μείωση του όγκου των ενδιάμεσων δεδομένων, με αποτέλεσμα να εξακολουθεί να είναι μεγαλύτερος από τον διαθέσιμο αποθηκευτικό χώρο στη μνήμη. Ακολούθησε περεταίρω διάσπαση σε ακόμη μικρότερα αρχεία της τάξεως του ενός probeSet ανά αρχείο. Το βήμα αυτό οδήγησε στη δημιουργία 56 χιλιάδων διαφορετικών αρχείων, κάθε ένα από τα οποία αντιστοιχεί σε ένα probeSet. Τα αρχεία που δημιουργήθηκαν περιέχουν τις πληροφορίες προς ανάλυση των δειγμάτων. Επιπλέον, αυτός ο τρόπος διάσπασης των δεδομένων διευκόλυνε την μετέπειτα ανάλυση και επεξεργασία τους.

Ο μεγάλος αριθμός αρχείων, καθιστούσε χρονοβόρα την σειριακή ανάλυση τους. Η χρήση διανεμημένου υπολογισμού, αποσκοπούσε στην μείωση του συνολικού χρόνου εκτέλεσης των μεθόδων ανάλυσης των δεδομένων. Η μείωση επιτεύχθηκε με παραλληλοποίηση της ανάλυσης των δεδομένων, με την χρήση όλων των διαθέσιμων πόρων που μπορούσε να παρέχει το grid, βάσει των παραμέτρων προτεραιότητας με τις οποίες είχαν αρχικά καταχωρηθεί στο grid τα δεδομένα προς επεξεργασία. Το συγκεκριμένο grid αποτελείτο από ένα σύνολο 200 υπολογιστών περίπου των οποίων η

χρήση καθορίζεται από τις παραμέτρους που καθόριζαν την σειρά προτεραιότητας και σημαντικότητας των διεργασιών όσον αφορά τη χρήση του grid.

3.8 Φιλτράρισμα Δεδομένων

Το σημαντικότερο μέρος από το οποίο αποτελείται η επεξεργασία και η προετοιμασία των δεδομένων, καθώς επίσης και των αποτελεσμάτων που παράχθηκαν κατά την ανάλυση, είναι η απομόνωση των πιο σημαντικών από αυτά, αλλά εξίσου σημαντικός είναι και ο περιορισμός του όγκου δεδομένων. Για την επίτευξη των δύο αυτών στόχων, εφαρμόστηκε φιλτράρισμα στα δεδομένα, με τη χρήση κάποιου κατωφλίου, βάσει του οποίου έγινε η επιλογή των σημαντικότερων δειγμάτων για τη φάση των ακατέργαστων δεδομένων και αργότερα για τη συλλογή των σημαντικότερων από τα αποτελέσματα.

Το φιλτράρισμα εφαρμόστηκε μετά τη διαδικασία διάσπασης των δεδομένων σε ένα probeSet ανά αρχείο, που καθιστούσε και τη διαδικασία συλλογής των σημαντικότερων δεδομένων από αυτά για κάθε συγκεκριμένο probeSet. Στην δεύτερη περίπτωση το φιλτράρισμα εφαρμόστηκε στο στάδιο της μετά επεξεργασίας των δεδομένων κατά τη συγχώνευση των αρχείων και τη συλλογή των αποτελεσμάτων υψίστης σημασίας.

Με τον τρόπο αυτό απομονώθηκαν τα σημαντικότερα δεδομένα και πληροφορίες για κάθε στάδιο αντίστοιχα, αυτά δηλαδή με τις μικρότερες τιμές των πιθανοτήτων σημαντικότητας (p-values).

3.8.1 Φιλτράρισμα για περιορισμό του όγκου των Δεδομένων

Ο περιορισμός του όγκου δεδομένων αποτέλεσε σημαντικό παράγοντα για δημιουργία και εφαρμογή κώδικα φιλτραρίσματος στα δεδομένα. Λόγω του μεγάλου όγκου δεδομένων, η δημιουργία και εφαρμογή κώδικα που φιλτράρει τα δεδομένα με τη χρήση ενός κατωφλίου κρίθηκε απαραίτητη. Το κατώφλι (threshold) στον κώδικα εισάγεται από τον προγραμματιστή ως παράμετρος, από την γραμμή εντολών κατά την εκτέλεση. Με τον περιορισμό του όγκου των δεδομένων στο αρχικό στάδιο επιτεύχθηκε περιορισμός των δεδομένων στα σημαντικότερα από αυτά.

Όπως προαναφέρθηκε, σημαντικότερα δεδομένα αποτελούσαν τα δείγματα με τις μικρότερες τιμές στην πιθανότητα σημαντικότητας. Περιορίζοντας όσο το δυνατό περισσότερο το κατώφλι σύγκρισης των δεδομένων για το φιλτράρισμα τους, απομόνωνε τα δεδομένα στα περισσότερο σημαντικά.

Η σημασία αυτής της διαδικασίας κατά την ανάλυση και επεξεργασία των δεδομένων αποδείχθηκε να είναι μεγάλη, καθώς παρατηρήθηκαν συγκριτικά μικρότεροι χρόνοι εκτέλεσης των αλγορίθμων ανάλυσης. Επιπλέον τα αποτελέσματα περιορίστηκαν στα σημαντικότερα και αργότερα κατά την γραφική απεικόνιση τους, τα αποτελέσματα θα ήταν πιο ευδιάκριτα στο μάτι και θα οδηγούσαν στη διεξαγωγή πιο ξεκάθαρων συμπερασμάτων.

Αυτού του είδους φιλτράρισμα εφαρμόστηκε αμέσως μετά το στάδιο της διάσπασης των δεδομένων και ένα στάδιο πριν την ανάλυσή τους και συγκεκριμένα υλοποιήθηκε κατάλληλος κώδικας που επεξεργαζόταν τα δεδομένα προτού περάσουν στο στάδιο της ανάλυσης.

Κεφάλαιο 4

Αλγόριθμοι Ανάλυσης Δεδομένων

4.1 Ανάλυση Δεδομένων και Διαχείριση Αρχείων

Η ανάλυση των δεδομένων αποτελεί το σημαντικότερο μέρος της όλης μελέτης. Όπως αναφέρθηκε και σε προηγούμενα υποκεφάλαια, ο αριθμός ελέγχων που έπρεπε να διεξαχθούν ήταν τεράστιος καθιστώντας την επεξεργασία με τη χρήση ενός και μόνο υπολογιστικού συστήματος πολύ χρονοβόρα και πολύπλοκη διαδικασία. Για το λόγο αυτό είναι που χρησιμοποιήθηκε και το κατανεμημένο σύστημα (grid) που ήταν διαθέσιμο και οι αλγόριθμοι που υλοποιήθηκαν, αποσκοπούσαν στη βέλτιστη χρήση του.

Ένας πολύ σημαντικός παράγοντας που περιόριζε την χρήση του grid, ήταν ο περιορισμένος αποθηκευτικός χώρος στη μνήμη, που ήταν διαθέσιμος από το σύστημα. Ο όγκος των αποτελεσμάτων υπολογίστηκε να ξεπερνά τα 400TB ενώ ο διαθέσιμος αποθηκευτικός χώρος έφτανε μόλις τα 200GB.

Το πρόβλημα αυτό αντιμετωπίστηκε χρησιμοποιώντας αλγόριθμους συμπίεσης/αποσυμπίεσης κατά την εκτέλεση των μεθόδων ανάλυσης πάνω στα δεδομένα, σε συνδυασμό με κώδικα φιλτραρίσματος για μείωση του όγκου των πληροφοριών, απομονώνοντας τις σημαντικότερες από αυτές.

Συγκεκριμένα κατά την ανάλυση των δεδομένων ενός αρχείου, μετά την παραγωγή κάποιου αποτελέσματος, και πριν από την καταγραφή του σε ένα νέο αρχείο εξόδου, εφαρμοζόταν σε αυτό κώδικας φιλτραρίσματος, που επέλεγε τις πληροφορίες βάσει ενός προκαθορισμένου κατωφλίου, καθώς επίσης και με την ενδιαφέρουσα στατιστική παράμετρο. Στη συνέχεια οι πληροφορίες που ικανοποιούσαν όλες τις συνθήκες επιλογής, καταγράφονταν στο αρχείο και συμπιέζονταν μειώνοντας έτσι στο μεγαλύτερο δυνατό βαθμό το μέγεθος του αρχείου.

Το όνομα κάθε αρχείου καθοριζόταν από το όνομα του probeSet για το οποίο περιείχε τα δεδομένα. Η ονομασία με αυτό τον τρόπο έπαιξε σημαντικό ρόλο για το τελικό στάδιο εφαρμογής περεταίρω φιλτραρίσματος. Η διαχείριση του grid και των αρχείων όπως επίσης και των αλγορίθμων ανάλυσης, του κώδικα συμπίεσης των αρχείων καθώς επίσης και του κώδικα φιλτραρίσματος, έγινε με τη βοήθεια script files και κώδικα γραμμένο σε C++ που αυτοματοποιούσαν την όλη διαδικασία.

Αποτέλεσμα της ανάλυσης των δεδομένων ήταν η παραγωγή 56 χιλιάδων συμπιεσμένων αρχείων, ένα για κάθε διαφορετικό probeSet. Στα αρχεία αυτά υπάρχουν οι τιμές των πιθανοτήτων σημαντικότητας για τα SNPs με τα οποία το probeSet ξεπερνά την τιμή κατωφλίου.

4.2 Μέθοδοι Ανάλυσης

Για την ανάλυση των δεδομένων εφαρμόστηκαν τρείς διαφορετικοί αλγόριθμοι που ήταν διαθέσιμοι από το εργαλείο plink. Οι τρείς αυτοί αλγόριθμοι ανάλυσης είναι η Ποσοτική Ανάλυση (Quantitative Trait Analysis), οι Διεργασίες Μεταλλαγής (Permutation Procedures) και η εφαρμογή Γραμμικών και Λογιστικών Μοντέλων ανάλυσης (Linear and Logistic Models), που αναλύονται με μεγαλύτερη λεπτομέρεια στη συνέχεια.

4.2.1 Ποσοτική Ανάλυση (Quantitative Trait Analysis)

Ποσοτική ανάλυση γνωρισμάτων (Quantitative Trait Analysis) είναι η μέθοδος με την οποία επιτυγχάνεται ο προσδιορισμός του επιπέδου σημασίας μιας μεμονωμένης ακολουθίας DNA. Ο προσδιορισμός του επιπέδου σημασίας μιας ακολουθίας DNA, γίνεται σε σχέση με τις υπόλοιπες γενετικές και μη επιδράσεις που παρατηρούνται πάνω σε ένα συγκεκριμένο γνώρισμα. Στη συγκεκριμένη μελέτη τα δεδομένα που μελετήθηκαν ήταν ένα σύνολο από 550 χιλιάδες SNPs και ένα σύνολο από 56 χιλιάδες μετάγραφα mRNA τα οποία συσχετίζονταν μεταξύ τους ως ζεύγη καρτεσιανού γινομένου, για κάθε ένα από τα οποία παραγόταν και μία τιμή σημαντικότητας (p-value).

Αυτό το είδος ανάλυσης οδήγησε στην παραγωγή ενός πολύ μεγάλου όγκου αποτελεσμάτων, συνολικά περίπου 550 χιλιάδες επί 56 χιλιάδες για κάθε διαφορετική συσχέτιση.

Η διαχείριση των αποτελεσμάτων επεξηγείται πιο αναλυτικά κατά την ανάλυση της διαδικασίας μετά επεξεργασίας των αποτελεσμάτων [8].

4.2.2 Διεργασίες Μεταλλαγής (Permutation Procedures)

Οι διεργασίες μεταλλαγής αποτελούν μια εντατικά υπολογιστική προσέγγιση για την παραγωγή επιπέδων σημασίας με εμπειρικό τρόπο. Οι τιμές που παράγονται με αυτόν τον τρόπο έχουν κάποιες ιδιότητες, όπως για παράδειγμα η χαλάρωση των υποθέσεων

όσον αφορά την κανονικοποίηση συνεχόμενων φαινοτύπων και η αρχή των Hardy-Weinberg που ασχολείται με τη διαχείριση σπάνιων αλληλόμορφων και δηλώνει πως οι συχνότητες των αλληλόμορφων όπως επίσης και των γονοτύπων παραμένουν σταθερές, βρίσκονται δηλαδή σε ισορροπία. Επιπλέον ασχολείται και με δείγματα μικρού μεγέθους, παρέχοντας με αυτό τον τρόπο ένα πλαίσιο εργασίας για διόρθωση, όσον αφορά τους πολλαπλούς ελέγχους [5] καθώς επίσης του ελέγχου των αναγνωρισμένων υποδομών ή άλλων συγγενικών συσχετίσεων, εφαρμόζοντας την διαδικασία μεταλλαγής μόνο σε μία ομάδα.

Οι διεργασίες μεταλλαγής προσφέρονται για μια πληθώρα δοκιμών και χωρίζονται σε δύο κατηγορίες ανάλογα με τους τομείς στους οποίους εφαρμόζονται.
Οι δύο κατηγορίες είναι:

1. Label – swapping έναντι gene dropping
2. Adaptive έναντι max(T)

Για τους σκοπούς αυτής της μελέτης χρησιμοποιήθηκε ο προσαρμοστικός αλγόριθμος μεταλλαγής (adaptive permutation) καταλήγοντας όμως στο τέλος να είναι ίσος με τον αλγόριθμο μεγίστου κατωφλίου (max(T)), το οποίο αναλύεται περιληπτικά στη συνέχεια .

Ακολουθώντας την προσαρμοστική προσέγγιση, οι μεταλλαγές που εφαρμόζονται στα SNPs τερματίζονται όταν τα αποτελέσματα που παράγονται είναι χαμηλού επιπέδου σημασίας (non significant) απο το αρχικό ακόμη στάδιο, απ’ ότι αν αυτά είναι σημαντικά. Για παράδειγμα, εάν μετά από 10 μεταλλαγές παρατηρηθεί πως για 9 από τα στατιστικά αποτελέσματα που έχουν που έχουν παραχθεί για ένα συγκεκριμένο SNP, είναι μεγαλύτερα από τα ήδη γνωστά αποτελέσματα, τότε δεν υπάρχει λόγος για περαιτέρω επεξεργασία του συγκεκριμένου SNP, καθώς δεν είναι πιθανό να οδηγήσει στη εξαγωγή κάποιου αποτελέσματος υψίστης σημασίας (όσο μεγαλύτερα αποτελέσματα τόσο μικρότερη η σημασία τους). Με τον τρόπο αυτό επιταχύνεται η διεργασία μεταλλαγής. Η επιτάχυνση αυτή επιτυγχάνεται λόγω του γεγονότος ότι τα περισσότερα από τα SNPs που δεν θεωρούνται σημαντικά θα απορριφθούν αρκετά σύντομα, έτσι ώστε να είναι δυνατός ο ορθός υπολογισμός της σημαντικότητας μιας μικρότερης ομάδας SNPs, που απαιτούν εκατομμύρια μεταλλαγές να υπολογιστούν.

Ως συνήθως η ακρίβεια με την οποία γίνεται ο υπολογισμός της σημαντικότητας ενός p-value που σχετίζεται με τον αριθμό των μεταλλαγών που εκτελέστηκαν (permuted), αποτελεί την ίδια την τιμή σημαντικότητας. Για τους περισσότερους όμως σκοπούς χρήσης των p-values, αυτό ακριβώς θα είναι και το επιθυμητό αποτέλεσμα καθώς αποτελούν μικρού ενδιαφέροντος αποτελέσματα, ενώ ένα καθαρά μη συσχετίσιμο SNP έχει στην πραγματικότητα τιμή σημαντικότητας (p-value) ίση με 0.78 ή 0.87.

Λόγω της τεράστιας σημαντικότητας των αποτελεσμάτων που χρησιμοποιήθηκαν στις διεργασίες μεταλλαγής, ο αριθμός των μεταλλαγών που εφαρμόστηκε στα περισσότερα από αυτά, έφτανε τον μέγιστο αριθμό που μπορούσαν να εκτελεστούν.

Για τον λόγο αυτό, από προσαρμοστική μέθοδος (adaptive) από την οποία ξεκίνησε αρχικά η ανάλυση, κατέληξε στην περίπτωση διεργασιών μεταλλαγής μέγιστου κατωφλίου (max(T)). Σε αυτή την περίπτωση αντίθετα με την προσαρμοστική μέθοδο, κανένα από τα SNPs δεν απορρίφθηκε καθ' όλη τη διάρκεια της διαδικασίας. Αυτό είχε ως αποτέλεσμα για κάθε SNP να εκτελείται ο μέγιστος αριθμός permutations που είχαν αρχικά καθοριστεί. Το προτέρημα αυτής της μεθόδου σε αντίθεση με την προσαρμοστική μέθοδο είναι ότι μπορούν να υπολογιστούν δύο διαφορετικά σύνολα εμπειρικών, σημαντικών τιμών. Δηλαδή ο υπολογισμός μιας τιμής σημαντικότητας για κάθε SNP ξεχωριστά, αλλά και μίας άλλης τιμής που ελέγχει το γεγονός ότι ένας μεγάλος αριθμός επιπλέον SNPs έχουν ελεγχθεί. Αυτό επιτυγχάνεται συγκρίνοντας κάθε στατιστική τιμή ελέγχου που είναι ήδη γνωστή έναντι της μέγιστης τιμής από όλες τις στατιστικές τιμές που παράχθηκαν κατά την εφαρμογή των διεργασιών μεταλλαγής σε όλα τα SNPs, για κάθε ένα από τα αντίγραφα. Με άλλα λόγια η τιμή p-value σε αυτή την περίπτωση ελέγχει το σχετικό ποσοστό σφάλματος, καθώς το p-value απεικονίζει την πιθανότητα παρατήρησης ενός στατιστικού πειράματος αυτού του μεγέθους, έχοντας ως δεδομένο πως εξετάστηκαν όλα τα διαθέσιμα δεδομένα που υπήρχαν.

Η μέθοδος Bonferroni λειτουργεί κάτω από την υπόθεση πως όλες οι δοκιμές είναι ανεξάρτητες μεταξύ τους. Αυτό έρχεται σε αντίθεση με τις διεργασίες μεταλλαγής οι οποίες διατηρούν μια συσχετιστική δομή μεταξύ των SNPs, παρέχοντας με αυτό τον τρόπο ευχέρεια διόρθωσης των πολλαπλών δοκιμών. Ακριβώς επειδή η τιμή

ενδιαφέροντος όταν εφαρμόζονται οι διεργασίες μεταλλαγής, είναι η διορθωμένη τιμή p-value, έχει ξεπεραστεί το πρόβλημα των πολλαπλών δοκιμών.

4.2.2.1 Ο ρόλος των Διεργασιών Μεταλλαγής στη Μελέτη

Οι διεργασίες μεταλλαγής χρησιμοποιήθηκαν στην μελέτη για την εξακρίβωση της εγκυρότητας των αποτελεσμάτων που παράχθηκαν από μεθόδους ανάλυσης που εφαρμόστηκαν νωρίτερα για τους σκοπούς της συγκεκριμένης μελέτης. Όπως προαναφέρθηκε η μέθοδος μεταλλαγής είναι πολύ χρονοβόρα διαδικασία και πόσο μάλλον με τόσο μεγάλο αριθμό συσχετίσεων. Όπως προαναφέρθηκε η χρήση των διεργασιών μεταλλαγής αποσκοπούσαν στην εξακρίβωση και επιβεβαίωση της εγκυρότητας των δεδομένων και όχι για την ανάλυση τους. Για τον λόγο αυτό δεν ήταν απαραίτητο να εφαρμοστεί η μέθοδος σε όλο το σύνολο τον δεδομένων ανεξαίρετα και εφόσον αποτελεί μια χρονοβόρα μέθοδο η εφαρμογή της σε ένα υποσύνολο των δεδομένων αυτών, ήταν αρκετή για να βγουν κάποια ενδεικτικά αποτελέσματα που θα ήταν σε θέση να οδηγήσουν στην διεξαγωγή ορθών συμπερασμάτων. Έτσι από το σύνολο των στατιστικά σημαντικών αποτελεσμάτων επιλέγηκαν τυχαία με Bonferroni Correction μόνο χίλια πεντακόσια από αυτά για την εφαρμογή της μεθόδου μεταλλαγής. Ο αριθμός των βρόγχων εκτέλεσης των μεταλλαγών τέθηκαν σε 10^6.

Ο όρος p-value ορίζει την πιθανότητα σημαντικότητας (significance probability) που αποτελεί την μικρότερη τιμή του επιπέδου σημαντικότητας a, για την οποία η μηδενική υπόθεση H_0 ενός ελέγχου απορρίπτεται. Όσο πιο μικρό είναι το επίπεδο σημαντικότητας τόσο πιο δύσκολο είναι να απορριφθεί η μηδενική υπόθεση [15].

4.2.3 Γραμμικά και Λογιστικά Μοντέλα (Linear & Logistic Models)

Τα γραμμικά και λογιστικά μοντέλα ανάλυσης (Linear & Logistic Models) αποτελούν ένα είδος ποσοτικής ανάλυσης που επιτρέπει την ανάλυση μεταξύ ποσοτικών και ποιοτικών μεταβλητών όπως στην περίπτωση μας τα μετάγραφα mRNA και SNPs αντίστοιχα, που χρησιμοποιήθηκαν και στην διαδικασία συσχέτισης ποσοτικών μεταβλητών κατά την ποσοτική ανάλυση που περιγράφηκε προηγουμένως. Η διαφορά αυτών των μοντέλων από την ποσοτική ανάλυση, είναι ότι επιτρέπουν την χρήση

συμμεταβλητών (covariates) για μελέτη των αλληλεπιδράσεων αυτών των συμμεταβλητών (covariates) με τις διάφορες συσχετίσεις των μετάγραφων mRNA και των σημειακών νουκλεοτιδικών πολυμορφισμών (SNPs) που προκύπτουν κατά την ανάλυση. Δηλαδή, το πως επηρεάζουν οι συμμεταβλητές αυτές στη συσχέτιση των SNPs με την έκφραση του mRNA.

Επομένως τα μοντέλα αυτά σχηματίζουν όλες τις πιθανές συσχετίσεις μεταξύ των SNPs και των μετάγραφων mRNA ελέγχοντας κάθε φορά κατά πόσο οι συμμεταβλητές (covariates) που θέτονται επηρεάζουν με κάποιο τρόπο στις συσχετίσεις αυτές. Υπάρχουν δύο διαφορετικά είδη τύπων συμμεταβλητών που είναι οι συνεχόμενες και οι δυαδικές (continuous and binary).

Τα covariates που ελέχθησαν στη μελέτη αυτή είναι η ασθένεια (disease), το φύλο (gender), το batch (το χρονικό πλαίσιο στο οποίο έγινε η ανάλυση), που είναι δυαδικού τύπου και η ηλικία (age) που είναι συνεχόμενου τύπου (continuous).

Η ανάλυση εκτελέστηκε σε δύο φάσεις όπου στην πρώτη αναλύθηκε το πρώτο μισό από τα δεδομένα και στην δεύτερη τα υπόλοιπα. Το batch καθορίζει σε ποιά από τις δύο αυτές φάσεις αναλύθηκαν τα δεδομένα.

4.3 Εργαλεία που χρησιμοποιήθηκαν

Για την διεξαγωγή της μελέτης απαραίτητη ήταν η χρήση ορισμένων εργαλείων. Αυτά αποτελούσαν κάποιες εφαρμογές που διατίθενται δωρεάν όπως η εφαρμογή plink που χρησιμοποιήθηκε για την ανάλυση των δεδομένων, αλλά και άλλες, όπως είναι η εφαρμογή Spotfire για την παρουσίαση των αποτελεσμάτων, καθώς επίσης και κάποιοι διαθέσιμοι πόροι σε υλικό που επίσης ήταν διαθέσιμη από την εταιρεία για ικανοποίηση των αναγκών κατά τη διεξαγωγή της μελέτης.

Η προσαρμογή του plink στο grid έγινε με τη χρήση λογισμικού που υλοποιήθηκε ειδικά για τη συγκεκριμένη μελέτη όπως επίσης και για την προσαρμογή των αποτελεσμάτων της ανάλυσης στην εφαρμογή Spotfire, για την τελική τους παρουσίαση.

4.3.1 Εφαρμογή Plink

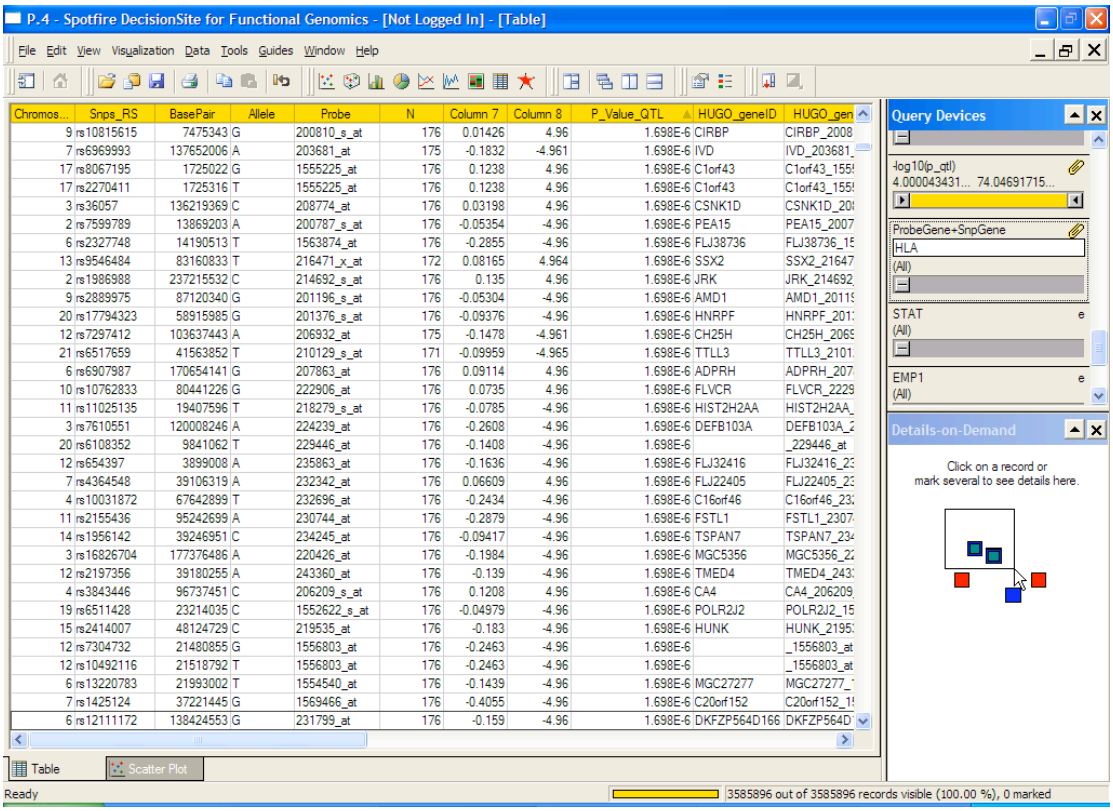
Το plink αποτελεί μια εφαρμογή που προσφέρεται ως ανοικτός κώδικας και χρησιμοποιείται για ανάλυση δεδομένων και συγκεκριμένα ως ένα εργαλείο που μπορεί να αναλύσει τις συσχετίσεις που μπορεί να υπάρχουν σε ολόκληρο το γονιδίωμα. Επιπλέον είναι σχεδιασμένο για να εκτελεί μια σειρά από αναλύσεις ευρείας κλίμακας με αποτελεσματικό και αποδοτικό τρόπο όσον αφορά την μεθοδολογία υπολογισμού.

4.3.2 Εφαρμογή Spotfire

Το spotfire αποτελεί ένα εργαλείο του οποίου η διαδραστική απεικόνιση της πληροφορίας καθώς επίσης και οι αναλυτικές λύσεις που προσφέρει στους χρήστες, τους χαρίζει μια αξιόλογη εμπειρία όσον αφορά την γρήγορη και εύκολη αναζήτηση σε βάσεις δεδομένων αλλά και αναφορά αποτελεσμάτων που είναι χρήσιμα για ψηλότερου επιπέδου επιστημονικών απαιτήσεων. Επιτρέπει την γραφική απεικόνιση αποτελεσμάτων και αποτελεί ενα ευκολόχρηστο εργαλείο δια τη διαχείριση δεδομένων αφου μέσω αυτού μπορούν να εφαρμοστούν διάφορα φίλτρα και ερωτήματα επιλογής (queries) για την απομόνωση δεδομένων.

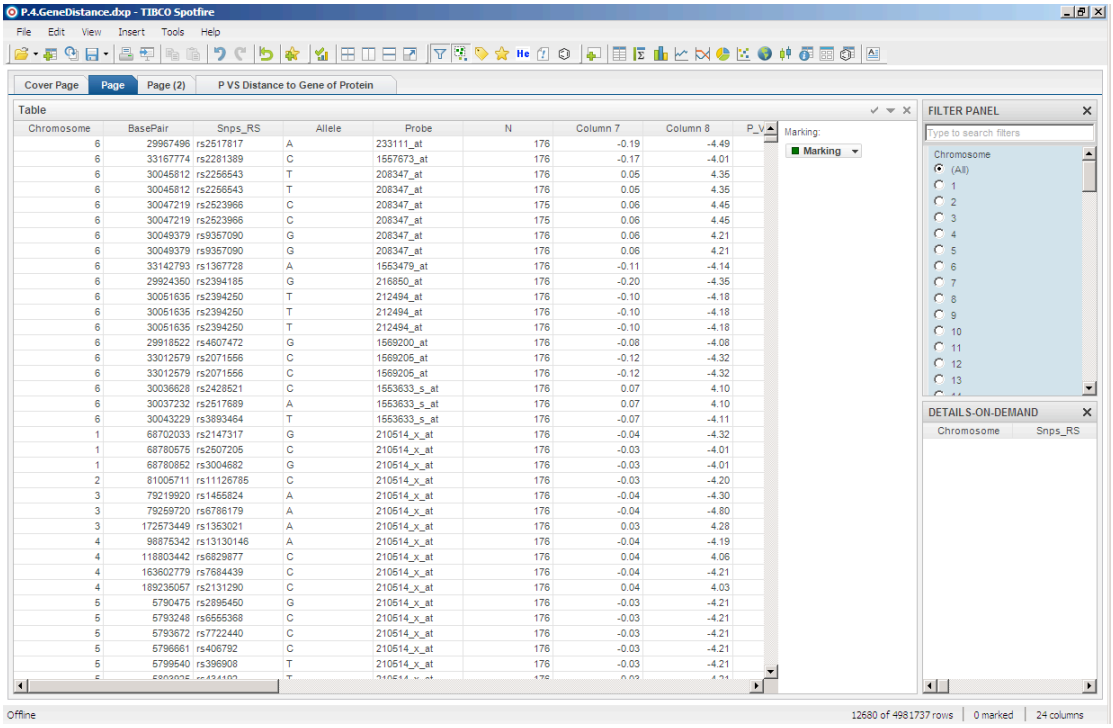
Ο ρόλος της εφαρμογής στη συγκεκριμένη μελέτη ήταν για διευκόλυνση της διαχείρισης των δεδομένων καθώς επίσης και γραφική απεικόνιση και παρουσίαση των αποτελεσμάτων.

4.3.2.1 Παραδείγματα Χρήσης της Εφαρμογής Spotfire



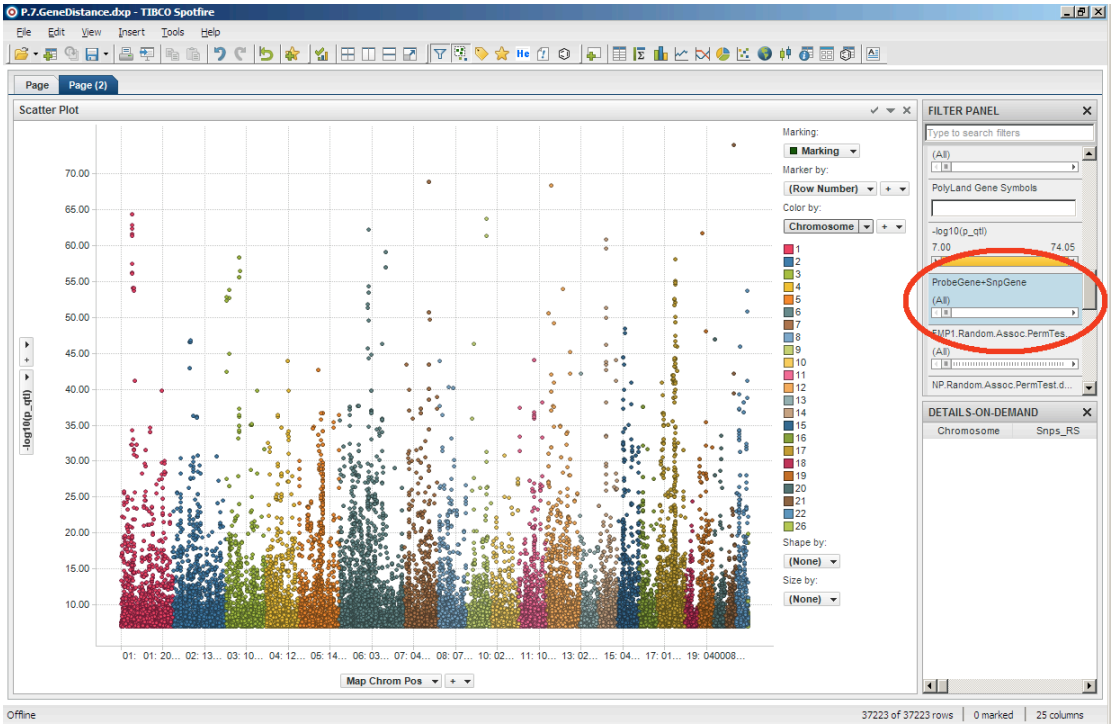
Εικόνα 4.1 Υπηρεσίες Αναζήτησης του Spotfire

Η εικόνα 4.1 αποτελεί ένα δείγμα της εισαγωγής των αποτελεσμάτων υπο μορφή πίνακα στην εφαρμογή Spotfire. Δεξιά μπορούν να παρατηρηθούν τα υπηρεσίες αναζήτησης, με την βοήθεια των οποίων έχει γίνει αναζήτηση της επιλογής των αποτελεσμάτων όπου το mRNA ή το SNP ανήκει σε ένα από τα γονίδια της οικογένειας HLA.



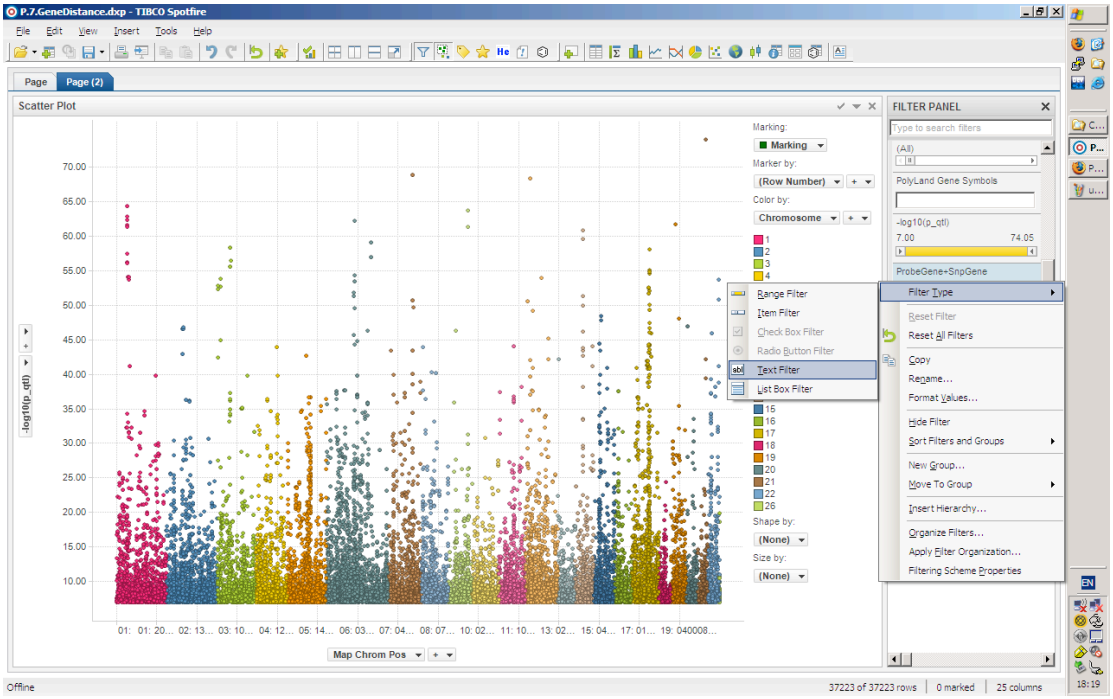
Εικόνα 4.2 Φιλτράρισμα βάσει Χρωμοσώματος στο Spotfire

Η εικόνα 4.2 αποτελεί ένα δείγμα από τα στοιχεία που περιέχει το αρχείο με τις πληροφορίες που παράχθηκαν κατά την εφαρμογή του φίλτρου με κατώφλι 10^{-4} και έχουν εισαχθεί στην εφαρμογή Spotifire. Η προσοχή εστιάζεται πάνω δεξιά στο filter panel από όπου μπορεί να επιλεγεί κάποιο συγκεκριμένο χρωμόσωμα.



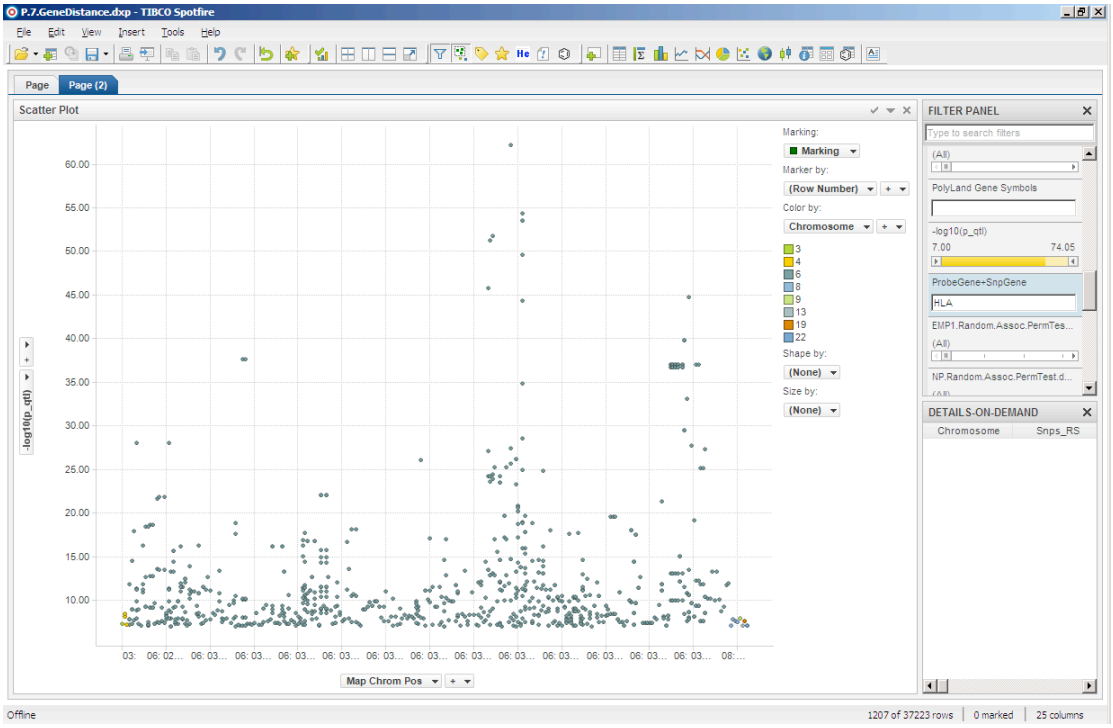
Εικόνα 4.3 Εφαρμογή Φίλτρου μέσω Spotfire

Στην εικόνα 4.3 παρουσιάζεται η διαδικασία εφαρμογής φίλτρου στα αποτελέσματα χρησιμοποιώντας την εφαρμογή Spotfire. Μπορεί να παρατηρηθεί πως η διαδικασία είναι σχετικά απλή και αρκεί μόνο να επιλεγεί το κατάλληλο πεδίο στο οποίο θα εφαρμοστεί το επιθυμητό φίλτρο.



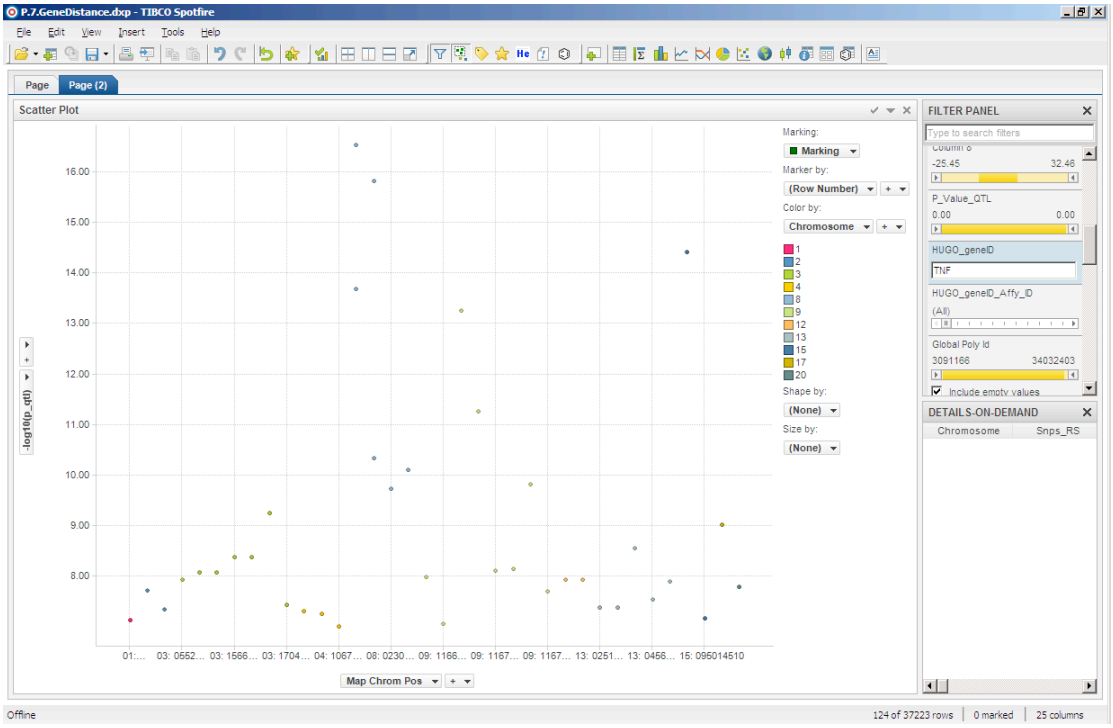
Εικόνα 4.4 Διαδικασία Εφαρμογής Νέου Φίλτρου μέσω Spotfire

Στην εικόνα 4.4 παρουσιάζεται ο τρόπος με τον οποίο επιλέγεται ο τύπος του φίλτρου που θα εφαρμοστεί. Όπως μπορεί να παρατηρηθεί υπάρχουν αρκετοί διαφορετικοί φίλτρου. Αυτό που καθορίζει τον τύπο του φίλτρου που θα χρησιμοποιηθεί είναι οι επιθυμητές γραφικές παραστάσεις που θα παραχθούν μέσω της επιλογής του φίλτρου και επιπλέον την ακρίβεια με την οποία θα γίνει το φιλτράρισμα των αποτελεσμάτων. Όπως και προηγουμένως μπορεί να παρατηρηθεί πως η διαδικασία που ακολουθείται είναι πολύ απλή και εύκολη.



Εικόνα 4.5 Αποτελέσματα Φιλτραρίσματος της οικογένειας γονιδίων HLA

Στην εικόνα 4.5 μπορούν να παρατηρηθούν τα αποτελέσματα της γραφικής που παράγεται με την εφαρμογή φίλτρου σε μία συγκεκριμένη οικογένεια γονιδίων. Στην περίπτωση αυτή το φιλτράρισμα εφαρμόστηκε στην οικογένεια γονιδίων HLA.



Εικόνα 4.6 Αποτελέσματα Φιλτραρίσματος της οικογένειας γονιδίων TNF

Στην εικόνα 4.6 μπορούν να παρατηρηθούν τα αποτελέσματα της γραφικής που παράγεται με την εφαρμογή φίλτρου σε ένα συγκεκριμένο γονίδιο. Στην περίπτωση αυτή το φιλτράρισμα εφαρμόστηκε για το γονίδιο TNF. Επιπλέον παρατηρείται πως τα αποτελέσματα σε σχέση με την εφαρμογή φίλτρου στο γονίδιο HLA που παρουσιάστηκε νωρίτερα, είναι διαφορετικά και η διαδικασία εξαγωγής αυτής της διαφορετικής γραφικής είναι πολύ απλή και καθόλου χρονοβόρα. Ακολουθήθηκε η ίδια διαδικασία και για το φιλτράρισμα του γονιδίου BRCA1 τα αποτελέσματα του οποίου επίσης θα σχολιαστούν στη συνέχεια.

4.3.3 Grid

Το πλέγμα υπολογιστών (grid) αποτελούσε το κύριο υλικό μέσο που χρησιμοποιήθηκε για την ανάλυση των δεδομένων. Το πλέγμα υπολογιστών διέθετε 200 επεξεργαστές και χάρη στην υλοποίηση ειδικού κώδικα για τους σκοπούς βελτιστοποίησης των συνθηκών εκτέλεσης, επιτεύχθηκε η επιτάχυνση της ανάλυσης, περίπου κατά 200 φορές. Η χρήση του διευκόλυνε αλλά και επιτάχυνε κατά μεγάλο βαθμό τη διαδικασία ανάλυσης των δεδομένων, λόγω και του μεγάλου αριθμού αρχείων που προέκυψαν

κατά τη διαδικασία της διάσπασης, των οποίων η διαχείριση και ανάλυση αποτελούσε πολύπλοκη και χρονοβόρα διαδικασία χωρίς τη χρήση του grid.

4.3.3.1 Αποθηκευτικός Χώρος

Ένας από τους περιορισμούς που έπρεπε να αντιμετωπιστούν όσον αφορά τη χρήση του grid ήταν το μέγεθος της διαθέσιμης αποθηκευτικής μνήμης που παρείχε το σύστημα. Ο διαθέσιμος αποθηκευτικός χώρος ανερχόταν περίπου στα 200 GB ενώ ο πραγματικός χώρος που ήταν αναγκαίος για αποθήκευση των αποτελεσμάτων, ήταν της τάξεως των 400TB. Κατά την παραγωγή των αποτελεσμάτων χρησιμοποιήθηκε αλγόριθμός φιλτραρίσματος και συμπίεσης των αρχείων έτσι ώστε να περιοριστεί το ποσό της μνήμης που ήταν αναγκαίο για την αποθήκευση τους, χωρίς όμως να χαθούν σημαντικές πληροφορίες. Στη συνέχεια με τη χρήση παραμετροποιημένου αλγορίθμου αποσυμπίεσης αλλά και συγχώνευσης, επιτεύχθηκε η συλλογή των αποτελεσμάτων σε ένα κοινό αρχείο, που καταλάμβανε χώρο ανάλογο του αριθμού των κορυφαίων αποτελεσμάτων που προσδιορίστηκαν, καθώς επίσης και του επιπέδου λεπτομέρειας της περιγραφής τους.

4.3.3.2 Διαθεσιμότητα Συστήματος

Οι πόροι του υφιστάμενου συστήματος δεν ήταν αφοσιωμένοι στις υπολογιστικές ανάγκες της μελέτης αυτής ανά πάσα στιγμή, καθώς ήταν προσβάσιμοι και από δεκάδες άλλες ομάδες ερευνητών. Η επιτυχής εξυπηρέτηση όλων των διαφορετικών χρηστών που επιχειρούν να χρησιμοποιήσουν το grid, γίνεται με τη βοήθεια την πλατφόρμας L.S.F. Η πλατφόρμα L.S.F δίνει προτεραιότητα στις διεργασίες που έχουν τον ψηλότερο δείκτη προτεραιότητας. Ο δείκτης προτεραιότητας για μία διεργασία καθορίζεται από τον εκάστοτε χρήστη. Συγκεκριμένα ο LSF αποτελεί έναν αλγόριθμο που αποσκοπεί στην δίκαιη κατανομή πόρων μεταξύ διεργασιών, και εξυπηρέτηση των χρηστών που επιθυμούν να χρησιμοποιήσουν το σύστημα, αυξάνοντας με αυτό τον τρόπο την αποδοτικότητα του συστήματος, εξυπηρετώντας κάθε χρήστη ανεξαίρετα. Αυτή η δίκαιη κατανομή των πόρων έχει ως στόχο εκτός από τη δίκαιη κατανομή τόσο των επεξεργαστών του συστήματος, όσο και της συνολικής διαθέσιμης μνήμης του συστήματος. Οι διεργασίες διαχωρίζονται σε κάποιες κατηγορίες ανάλογα με τον απαιτούμενο χρόνο εκτέλεσης. Υπάρχουν τέσσερις διαφορετικές κατηγορίες κάθε μια

από τις οποίες υλοποιείται ως μια ουρά που κρατά τις υποψήφιες διεργασίες προς εκτέλεση. Οι τέσσερις κατηγορίες ως προς το χρόνο εκτέλεσης είναι short, medium, long και very long. Η ουρά στην οποία θα υποβληθεί μια διεργασία μέχρι τη σειρά της προς εκτέλεση, επιλέγεται από τον ίδιο τον χρήστη. Μία διεργασία που υποβάλλεται στη ουρά short έχει ως μέγιστο χρόνο εκτέλεσης μέχρι δεκαπέντε λεπτά, ενώ διεργασίες που υποβάλλονται στην medium ουρά έχουν μέγιστο διαθέσιμο χρόνο εκτέλεσης μία ώρα, σαράντα οκτώ ώρες διαθέσιμου χρόνου εκτέλεσης για την ουρά long και απεριόριστο χρόνο εκτέλεση για τις διεργασίες που υποβάλλονται στην ουρά very long. Συνήθως η ουρά medium είναι η προκαθορισμένη ουρά από το σύστημα, εκτός κι αν επιλεγεί κάποια άλλη ουρά από τον χρήστη ανάλογα με τις απαιτήσεις των διεργασιών. Επιπλέον μπορεί να δοθεί κάποιος βαθμός προτεραιότητας στην διεργασία καθορίζοντας με αυτό τον τρόπο τη σειρά που λαμβάνει στην ουρά η συγκεκριμένη διεργασία. Ο βαθμός προτεραιότητας μιας διεργασίας είναι επίσης επιλογή του χρήστη και όσο μεγαλύτερος αυτός ο βαθμός, τόσο πιο σημαντική είναι η διεργασία, θα τοποθετηθεί στις πρώτες θέσεις ώστε να εκτελεστεί το συντομότερο δυνατό. Με τον τρόπο αυτό επιτυγχάνεται η ταξινόμηση των διεργασιών στην ουρά. Αν θεωρηθεί απαραίτητο ή αναγκαίο ο βαθμός προτεραιότητας μπορεί να αλλαχτεί.

Όσον αφορά τα αποτελέσματα που παράγονται από την εκτέλεση των διεργασιών, αυτά μπορούν να στέλνονται κατευθείαν στην προσωπική ηλεκτρονική διεύθυνση του χρήστη ή σε κάποιο αρχείο εξόδου που θα επιλέξει. Σε γενικές γραμμές τα αποτελέσματα αποθηκεύονται στον προσωπικό χώρο του χρήστη, για το λόγο αυτό είναι απαραίτητο να υπάρχει ο κατάλληλος αποθηκευτικός χώρος διαθέσιμος, ή σε αντίθετη περίπτωση να ληφθούν τα κατάλληλα μέτρα ώστε να τα αποτελέσματα να προσαρμόζονται στην διαθέσιμη μνήμη.

Για τους σκοπούς αυτής της μελέτης ο δείκτης προτεραιότητας των διεργασιών που εκκρεμούσαν ως προς τη χρήση του grid, ήταν ο χαμηλότερος δυνατός και η ουρά στην οποία καταχωρούνταν οι διεργασίες αυτή με τον απεριόριστο χρόνο εκτέλεσης έτσι ώστε να μπορούν να προχωρούν προς εκτέλεση οποιαδήποτε στιγμή υπάρχουν διαθέσιμοι πόροι, χωρίς να επηρεάζει οποιεσδήποτε άλλες εκκρεμότητες της εταιρείας.

4.4 Αλγόριθμος Χρήσης της Εφαρμογής Plink μεσω χρήσης του Grid

Για σκοπούς ανάλυσης των δεδομένων χρησιμοποιώντας τους αλγόριθμους ανάλυσης που ήταν διαθέσιμοι από την εφαρμογή plink, εγκαταστάθηκε σε κάθε μηχανή που ήταν ενωμένη στο grid η εφαρμογή plink. Για να μπορέσει να αξιοποιηθεί η εφαρμογή για την χρήση των αλγορίθμων, απαραίτητη ήταν η υλοποίηση κώδικα, που βάσει των δεδομένων που περιείχε κάθε αρχείο, δημιουργούσε scripts ώστε να είναι δυνατή η αυτόματη καταχώρηση τους στο σύστημα προς εκτέλεση. Συγκεκριμένα ο κώδικας σαρώνει τα δεδομένα που περιέχει κάθε αρχείο, διαβάζοντας τα SNPs που περιέχει κάθε αρχείο που αντιστοιχεί σε ένα probeSet. Αφού διαβαστούν τα περιεχόμενα του αρχείου το πρόγραμμα δημιουργεί ένα script file στο οποίο αποθηκεύεται η εντολή με την οποία θα καλείται η κατάλληλη μέθοδος από το plink, με τις σωστές παραμέτρους. Η μέθοδος που καλείτο με τη χρήση αυτών των αρχείων ήταν η διεργασία μεταλλαγής (permutation procedure) για την οποία ήταν αναγκαία η καταχώρηση των SNPs που λάμβαναν μέρος στην διαδικασία και αποτελούσαν και περιεχόμενα του αρχείου με τα δεδομένα. Η καταχώρηση των αρχείων προς επεξεργασία ένα προς ένα θα ήταν πολύ χρονοβόρα κι έτσι με τον τρόπο αυτό επιτεύχθηκε η χρήση του plink και του grid ταυτόχρονα κάνοντας την όλη διαδικασία ανάλυσης των δεδομένων αποδοτικότερη.

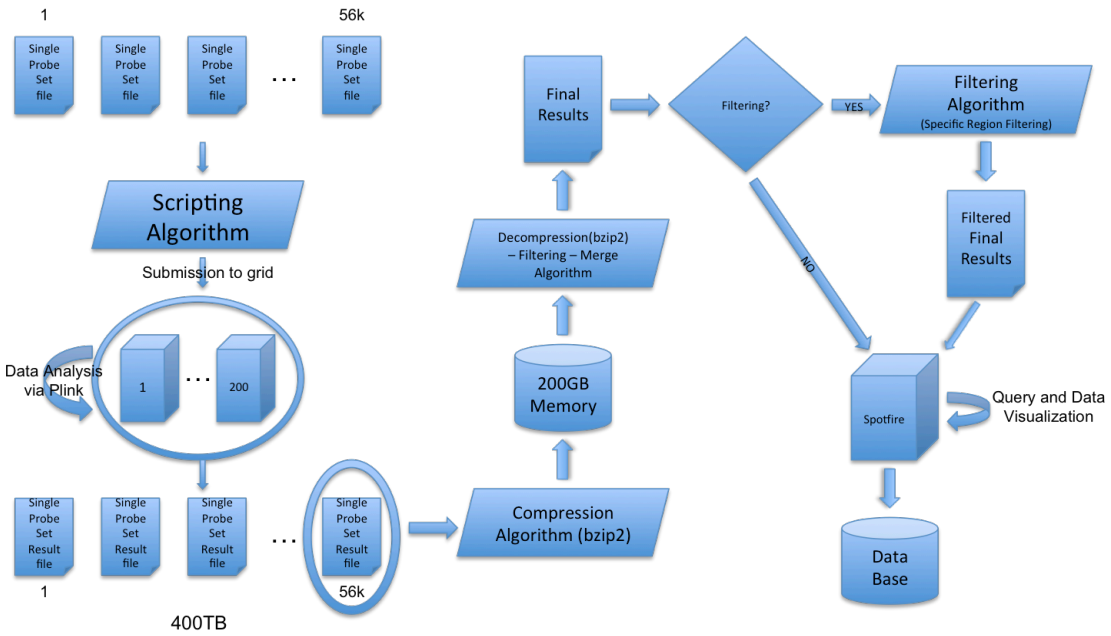
4.5 Αλγόριθμος Προσαρμογής των Αποτελεσμάτων στην Εφαρμογή Spotfire

Ο ουσιαστικός κώδικας για τη μεταφορά των αποτελεσμάτων από το plink στην εφαρμογή Spotfire για γραφική απεικόνιση των αποτελεσμάτων, ήταν αυτός που εκτελούσε την συγχώνευση των αρχείων με τα αποτελέσματα σε ένα και ταυτόχρονα το φιλτράρισμα τους για την απομάκρυνση των άχρηστων πληροφοριών που παράγονται κατά την ανάλυση καθώς επίσης και φιλτράρισμα για την απομόνωση των σημαντικότερων από αυτά. Το τελικό αρχείο όμως εξακολουθούσε να είναι μεγάλο σε σχέση με τον όγκο δεδομένων που μπορούσε να εξυπηρετήσει το Spotfire αποτελεσματικά. Για τον λόγο αυτό υλοποιήθηκε επιπλέον κώδικας που εφάρμοζε περεταίρω φιλτράρισμα πάνω στα δεδομένα και πάλι βάσει ενός κατωφλίου που ορίζεται από τον χρήστη. Με τον τρόπο αυτό επιτυγχάνεται περεταίρω μείωση του όγκου των δεδομένων στα πιο σημαντικά, αλλά και region specific filtering, περιορίζοντας τα δεδομένα για μια συγκεκριμένη περιοχή η οποία επιθυμείται να μελετηθεί. Επιπλέον παρέχεται η δυνατότητα μελέτης διαφόρων περιοχών ανάλογα με

τα ενδιαφέροντα αφού τα αρχικά δεδομένα δεν χάνονται, αλλά παράγονται νέα αρχεία με τα φιλτραρισμένα δεδομένα και τις σημαντικότερες τιμές. Επιπλέον τα μικρότερου όγκου αρχεία μπορούν να εξυπηρετηθούν αποδοτικότερα από το Spotfire χωρίς μεγάλους χρόνους καθυστέρησης στην απόκριση του συστήματος. Αυτός είναι και ο λόγος που στην μελέτη αυτή εφαρμόστηκε φιλτράρισμα αρχικά με κατώφλι $1x10^{-4}$ που παρουσίαζε μεγάλη καθυστέρηση όσον αφορά τον χρόνο απόκρισης της εφαρμογής Spotfire και για αυτό εφαρμόστηκε περεταίρω φιλτράρισμα στα δεδομένα με κατώφλι $1x10^{-7}$ που δεν επηρέαζε την σημαντικότητα των αποτελεσμάτων, απλά μόνο επιτάχυνε την επεξεργασία τους μέσω του Spotfire που χρησιμοποιήθηκε για την γραφική απεικόνιση και την παρουσίαση τους.

Κεφάλαιο 5

Μεταεπεξεργασία



Διάγραμμα 5.1 Μεταεπεξεργασία

5.1 Περιγραφή Διαδικασίας

Ως μέρος της μετά επεξεργασίας των αποτελεσμάτων ήταν η συγχώνευση τους σε ένα ενιαίο αρχείο ώστε να είναι εφικτή στη συνέχεια η απεικόνιση και παρουσίαση των αποτελεσμάτων. Η απεικόνιση των αποτελεσμάτων υπό μορφή γραφικών αναπαραστάσεων έγινε με τη χρήση του εργαλείου spotfire που βρισκόταν διαθέσιμο από την εταιρεία.

Η συγχώνευση των αρχείων έγινε με υλοποίηση κώδικα ειδικά σχεδιασμένου ώστε να μπορεί να ανταποκριθεί στις απαιτήσεις του συστήματος. Για τη συγχώνευση των αποτελεσμάτων απαραίτητη ήταν η αποσυμπίεση των αρχείων με τα αποτελέσματα, που παράχθηκαν κατά το στάδιο της προ επεξεργασίας. Πολλά από τα αποτελέσματα που παράχθηκαν κατά την ανάλυση των δεδομένων ήταν περιττά και για το λόγο αυτό ο περιορισμός τους στα άκρως απαραίτητα, μείωσε την ανάγκη για αποθηκευτικό χώρο σε πολύ μεγάλο βαθμό, αλλά επίσης περιόρισε τις πληροφορίες στις άκρως ενδιαφέρουσες για τη μελέτη. Η συλλογή αυτή των σημαντικότερων από τα αποτελέσματα έγινε επίσης με τη βοήθεια κώδικα που υλοποιήθηκε ακριβώς για να φιλτράρει τις σημαντικότερες από τις πληροφορίες βάση ενός κατωφλίου που καθορίζεται από τον χρήστη.

Η αποσυμπίεση όπως επίσης και η συμπίεση των παραγόμενων αρχείων έγινε βάσει του αλγορίθμου συμπίεσης αποσυμπίεσης bzip2. Η διαδικασία αρχίζει με το στάδιο της αποσυμπίεσης των αρχείων και έπειτα το φιλτράρισμα τους, προτού προχωρήσει στο στάδιο της συγχώνευσης τους. Με τον τρόπο αυτό επιτεύχθηκε η απομόνωση των σημαντικότερων αποτελεσμάτων.

Επιπλέον κώδικας υλοποιήθηκε για περεταίρω φιλτράρισμα του αρχείου των αποτελεσμάτων. Για το φιλτράρισμα έγινε θεσμός ανώτερων κατωφλίων για τα μεγάλα αρχεία (περίπου μεγέθους 200MB) , με p-value να ισούται με $1x10^{-4}$ ενώ για τα μικρότερα αρχεία το κατώφλι περιορίστηκε στο $1x10^{-7}$. Η διαδικασία ου φιλτραρίσματος σε αυτό το σημείο της επεξεργασίας των δεδομένων έγινε για να απομονωθούν μόνο οι σημαντικότερες από τις τιμές που παράχθηκαν κατά τη διαδικασία της ανάλυσης των δεδομένων. Επίσης για μελέτη των τιμών γύρω από μια

συγκεκριμένη περιοχή ενδιαφέροντος καθώς επίσης και για αποδοτικότερη χρήση του εργαλείου Spotfire για παρουσίαση και απεικόνιση των αποτελεσμάτων. Για σκοπούς αυτής της μελέτης το μέγεθος των αρχείων καθορίστηκε από το κατώφλι που ορίστηκε σε $1x10^{-7}$. Σημαντικό είναι να αναφερθεί πως το μικρότερο μέγεθος των αρχείων που προέκυψε μετά από την εφαρμογή του κώδικα φιλτραρίσματος με την μικρότερη τιμή κατωφλίου, δεν αλλοίωνε το επίπεδο σημασίας των αποτελεσμάτων, αλλά τα περιόριζε στα σημαντικότερα από αυτά.

Αποτέλεσμα της μετά επεξεργασίας των δεδομένων ήταν το τελικό αρχείο με τα σημαντικότερα αποτελέσματα, που χρησιμοποιήθηκαν στο εργαλείο Spotfire για την γραφική απεικόνιση τους και τη διεξαγωγή συμπερασμάτων. Επιπλέον προστέθηκαν στις πληροφορίες περιγραφικές λεπτομέρειες για κάθε σημείο mRNA και SNP για τα οποία μπορεί να γίνεται και πιο συγκεκριμένο φιλτράρισμα για κάθε διαφορετικό SNP ξεχωριστά. Για παράδειγμα το γονίδιο στο οποίο ανήκει το αντίστοιχο mRNA και SNP.

5.2 Φιλτράρισμα

Όπως και στο στάδιο της προεπεξεργασίας έτσι και εδώ κρίθηκε αναγκαία η εφαρμογή η υλοποίηση και εφαρμογή κώδικα φιλτραρίσματος που περιγράφεται αναλυτικότερα στη συνέχεια.

5.2.1 Φιλτράρισμα για απομόνωση των χρήσιμων πληροφοριών

Μία δεύτερη χρήσιμη εφαρμογή φιλτραρίσματος των δεδομένων κρίθηκε να είναι η περίπτωση φιλτραρίσματος των αποτελεσμάτων που παράχθηκαν μετά την ανάλυση των δεδομένων. Στην περίπτωση αυτή όπως και σε αυτή που αναφέρθηκε στο στάδιο της προ επεξεργασίας των δεδομένων, υλοποιήθηκε κατάλληλος κώδικας που ανταποκρινόταν στις νέες απαιτήσεις.

Αυτού του είδους φιλτράρισμα αποσκοπούσε στο να βοηθήσει τον ερευνητή, να προσαρμόσει τις πληροφορίες ανάλογα με τις απαιτήσεις της έρευνας και το βάρος που χρειάζεται να δώσει κάθε φορά στις πληροφορίες. Επίσης και σε αυτή την περίπτωση η επιλογή των πληροφοριών γίνεται με τη χρήση κάποιου κατωφλίου το οποίο και πάλι ορίζεται από τον χρήστη. Ισχύουν οι ίδιοι περιορισμοί όσο αφορά τις τιμές, δίνοντας μεγαλύτερο βάρος στις μικρότερες τιμές (p-values) και μικρότερο στις μεγαλύτερες τιμές από αυτές.

Το φιλτράρισμα σε αυτό το σημείο έγινε όπως προαναφέρθηκε κατά τη συγχώνευση των αρχείων και αποσκοπούσε στο να διατηρήσει ένα σύνολο από τα σημαντικότερα αποτελέσματα που παράχθηκαν κατά την ανάλυση των δεδομένων.

Το φιλτράρισμα έγινε θέτοντας ανώτερα κατώφλια για τα μεγάλα σε μέγεθος αρχεία (περίπου μεγέθους 200MB) , με p-value να ισούται με $1x10^{-4}$ ενώ για τα μικρότερα αρχεία το κατώφλι περιορίστηκε στο $1x10^{-7}$. Η διαδικασία ου φιλτραρίσματος σε αυτό το σημείο της επεξεργασίας των δεδομένων έγινε για να απομονωθούν μόνο οι σημαντικότερες από τις τιμές που παράχθηκαν κατά τη διαδικασία της ανάλυσης των δεδομένων.

Επιπλέον φιλτράρισμα στο στάδιο αυτό ήταν εφικτό και μετά την συγχώνευση των αποτελεσμάτων σε ένα αρχείο. Για την περίπτωση αυτή υλοποιήθηκε κώδικας που φίλτραρε τα δεδομένα βάσει κάποιου κατωφλίου που και σε αυτή την περίπτωση δινόταν ως παράμετρος από τον χρήστη. Σκοπός του φιλτραρίσματος σε αυτό το στάδιο ήταν ο περιορισμός των δεδομένων στα άκρως σημαντικότερα χωρίς να αλλοιώνει τα τελικά αποτελέσματα που παράχθηκαν και η δημιουργία πολύ μικρότερων αρχείων που θα κάνει αποδοτικότερη την επεξεργασία τους με τη βοήθεια της εφαρμογής Spotifire. Επίσης επιτρέπει τον περιορισμό των δεδομένων γύρω από κάποια συγκεκριμένη περιοχή ενδιαφέροντος που μπορεί να προκύψει σε οποιαδήποτε μελέτη. Αυτό επιτυγχάνεται χωρίς να χαθεί το αρχείο με όλο το σύνολο δεδομένων που παράχθηκαν μετά τη συγχώνευση τους και καλύπτουν ένα ευρύτερο φάσμα.

Με αυτό τον τρόπο επιτεύχθηκε ο περιορισμός των αποτελεσμάτων στα περισσότερο σημαντικά μόνο διευκολύνοντας την διεξαγωγή συμπερασμάτων και την πιο καθαρή απεικόνιση τους.

5.3 Συγχώνευση Αρχείων

Τελικό στάδιο της επεξεργασίας των δεδομένων και της παραγωγής των αποτελεσμάτων αποτέλεσε η συγχώνευση των αρχείων που παράχθηκαν στο στάδιο της ανάλυσης των δεδομένων. Όπως και στην ανάλυση των δεδομένων έτσι και εδώ ο έλεγχος των διαδικασιών και της ανάθεσης διεργασιών στο grid, έγινε μέσω της χρήσης scripts.

Πιο αναλυτικά σε αυτή τη φάση έγινε η συγχώνευση των αποτελεσμάτων σε ένα μεγαλύτερο αρχείο, ώστε να είναι δυνατή η απεικόνιση και παρουσίαση των αποτελεσμάτων με τη χρήση ειδικών εφαρμογών.

Για τη συγχώνευση των αρχείων απαραίτητη, ήταν η χρήση αλγόριθμου αποσυμπίεσης των αρχείων και κώδικα που έγραφε τα δεδομένα κάθε μικρού αρχείου σε ένα νέο ενιαίο και μεγαλύτερο αρχείο που θα περιείχε τα τελικά αποτελέσματα.
Μετά την καταγραφή των πληροφοριών από τα μικρότερα αρχεία στο τελικό, κάθε ένα από αυτά διαγραφόταν από τη μνήμη ώστε ο διαθέσιμος αποθηκευτικός χώρος να παραμένει σταθερός και προς αποφυγή προβλημάτων υπερχείλισης της μνήμης.

Με την καταγραφή των αποτελεσμάτων σημαντική ήταν και η προσθήκη μιας επιπλέον στήλης στο τελικό αρχείο με τα αποτελέσματα, που περιείχε το όνομα του αρχείου από το οποίο προήλθαν και συγκεκριμένα το όνομα αυτό ήταν το όνομα του probeSet. Με τον τρόπο αυτό μπορούσαν να είναι γνωστά τα αποτελέσματα που παράχθηκαν για κάθε διαφορετικό probeSet, ώστε να μπορούν να χρησιμοποιηθούν και ξεχωριστά αν αυτό θεωρηθεί απαραίτητο.

Η συγχώνευση των αποτελεσμάτων σε ένα αρχείο έγινε σε συνδυασμό με φιλτράρισμα των δεδομένων, ώστε να απομονωθούν οι σημαντικότερες από τις πληροφορίες που παράχθηκαν, ως αποτέλεσμα της ανάλυσης των δεδομένων και αποτελούν το επίκεντρο της προσοχής της μελέτης.

Όπως προαναφέρθηκε μετά την διαδικασία παραγωγής του τελικού αρχείου με τα αποτελέσματα, κώδικας επιπλέον φιλτραρίσματος μπορούσε να εφαρμοστεί στις πληροφορίες προσαρμόζοντας τις ανάλογα με τις απαιτήσεις της έρευνας.

5.4 Προσθήκη Πληροφοριών

Η προσθήκη πληροφοριών έγινε για την διευκόλυνση της ανάγνωσης των αποτελεσμάτων, καθώς επίσης και την παραγωγή γραφικών για προβολή των αποτελεσμάτων της ανάλυσης. Η προβολή γίνεται με τέτοιο τρόπο ώστε η πληροφορία που υποδηλώνουν τα αποτελέσματα να παρουσιάζεται οπτικά.

5.4.1 Επιπρόσθετες Πληροφορίες για τα SNPs

Οι πληροφορίες που προστέθηκαν επιπλέον για τα SNPs είναι οι ακόλουθες:

Χρωμόσωμα(Chromosome): αποτελεί έναν ακέραιο αριθμό που υποδείχνει τον αριθμό του χρωμοσώματος πάνω στο οποίο βρίσκεται ένα συγκεκριμένο SNP. Για τα χρωμοσώματα X και Y καθώς επίσης και για τα μιτοχονδριακά SNPs ανατέθηκαν ξεχωριστοί αριθμοί για το καθένα.

Θέση χρωμοσώματος (Chromosome location): αποτελεί ένα ακέραιο αριθμό που υποδείχνει τη θέση στην οποία βρίσκεται το SNP στο συγκεκριμένο χρωμόσωμα. Πιο συγκεκριμένα τον αριθμό του νουκλεοτιδίου πάνω στο χρωμόσωμα όπου εμφανίζεται το SNP, όπως αυτό χαρακτηρίζεται στη βάση dbSNP στο hapmap.

Γονίδιο (Gene): αποτελεί μια συμβολοσειρά που υποδείχνει το όνομα του γονιδίου στο οποίο εμφανίζεται το SNP.

Γονίδιο αποστάσεως ±20 νουκλεοτιδίων (Gene within ±20kb): οι πληροφορίες περιλαμβάνουν τα ονόματα των γονιδίων που απέχουν κατά ± 20 νουκλεοτίδια από το SNP. Σε περίπτωση που κανένα γονίδιο δεν βρίσκεται στο εύρος αυτό των ± 20 νουκλεοτιδίων, τότε το πεδίο αυτό παραμένει κενό. Σε περίπτωση που υπάρχουν περισσότερα από ένα τότε συμπεριλαμβάνονται όλα. Η απόσταση από το SNP μπορεί να καθορίζεται όση θεωρείται απαραίτητη κάθε φορά έχοντας $\pm X$ νουκλεοτίδια όπου το X ορίζει μία παράμετρο.

5.4.2 Επιπρόσθετες Πληροφορίες για τα ProbeSets

Οι πληροφορίες που προστέθηκαν επιπλέον για τα probeSets είναι οι ακόλουθες:

Χρωμόσωμα(Chromosome): αποτελεί έναν ακέραιο αριθμό που υποδείχνει τον αριθμό του χρωμοσώματος από το οποίο προέρχεται ένα συγκεκριμένο probeSet.

Θέση χρωμοσώματος (Chromosome location): αποτελεί δύο ακέραιους αριθμούς που καθορίζουν την αρχή και το τέλος του probeSet στο συγκεκριμένο χρωμόσωμα, δηλαδή από που αρχίζει και που τελειώνει ο συγκεκριμένο αριθμό των νουκλεοτιδίων που αποτελούν το probeSet.

Γονίδιο (Gene): αποτελεί μια συμβολοσειρά που υποδείχνει το όνομα του γονιδίου στο οποίο εμφανίζεται το probeSet.
mRNA ID – HUGO ID: αποτελεί μια συμβολοσειρά που υποδείχνει το όνομα της ακολουθίας mRNA την οποία ανιχνεύτηκε το συγκεκριμένο probeSet.

5.4.3 Επιπρόσθετες Πληροφορίες για τις Πρωτεΐνες

Οι πληροφορίες που προστέθηκαν για τις πρωτεΐνες είναι οι ίδιες με την περίπτωση των probeSets:

Χρωμόσωμα(Chromosome): αποτελεί δύο ακέραιους αριθμούς που καθορίζουν την αρχή και το τέλος του γονιδίου στο συγκεκριμένο χρωμόσωμα του οποίου η έκφραση παράγει την πρωτεΐνη, δηλαδή από που αρχίζει και που τελειώνει, τον συγκεκριμένο αριθμό των νουκλεοτιδίων.

Θέση χρωμοσώματος (Chromosome location): αποτελεί δύο ακέραιους αριθμούς που καθορίζουν την αρχή και το τέλος του γονιδίου που με την έκφραση του παράγεται η συγκεκριμένη πρωτεΐνη, δηλαδή από που αρχίζει και που τελειώνει, τον συγκεκριμένο αριθμό των νουκλεοτιδίων.

Γονίδιο (Gene): αποτελεί μια συμβολοσειρά που υποδείχνει το όνομα του γονιδίου του οποίου η έκφραση παράγει τη συγκεκριμένη πρωτεΐνη.

mRNA ID – HUGO ID: αποτελεί μια συμβολοσειρά που υποδείχνει το όνομα της ακολουθίας mRNA του οποίου η μετάφραση παράγει την πρωτεΐνη.

5.4.4 Επιπρόσθετες Πληροφορίες για τις Κορυφαίες Συσχετίσεις Μεταξύ SNPs και ProbeSets

Στο στάδιο αυτό της μετά επεξεργασίας των δεδομένων, αναλύθηκε σε μεγαλύτερο βάθος η συσχέτιση μεταξύ SNP και probeSets. Συγκεκριμένα η συσχέτιση αυτή χωρίστηκε σε τέσσερις διαφορετικές περιπτώσεις:

- Περίπτωση 1 – Το SNP προηγείται του mRNA
- Περίπτωση 2 – Το SNP έπεται του mRNA
- Περίπτωση 3 – Το SNP και το mRNA δεν έχουν καμία απολύτως συσχέτιση
- Περίπτωση 4 – Το SNP βρίσκεται πάνω στο mRNA

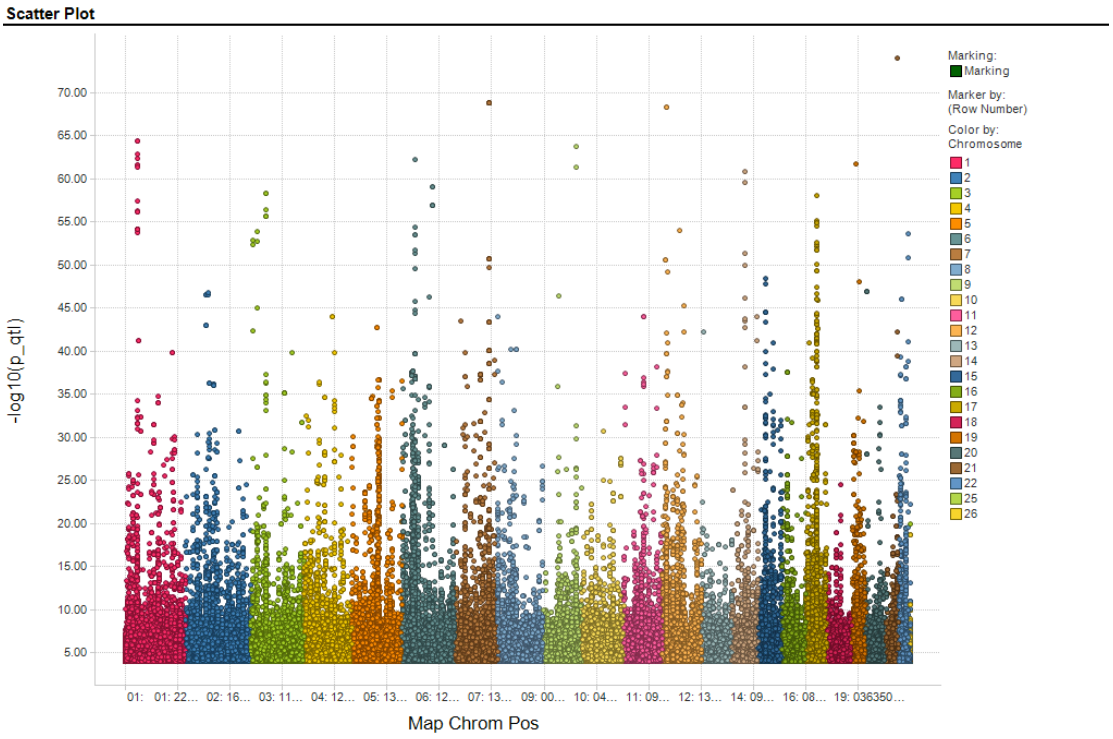
Για τις δύο πρώτες περιπτώσεις η τιμή που χρησιμοποιήθηκε είναι η ελάχιστη απόσταση σε αριθμό νουκλεοτιδίων, μεταξύ του mRNA και του SNP. Για την τρίτη περίπτωση όπου το SNP και το mRNA δεν έχουν καμία απολύτως συσχέτιση η τιμή που χρησιμοποιήθηκε είναι 900000000 ως sentinel value για να μπορεί να είναι εμφανής η διαφορά κατά την προβολή των αποτελεσμάτων. Για την τέταρτη περίπτωση στην οποία το SNP βρίσκεται στην περιοχή του mRNA θεωρούμε πως η ελάχιστη απόσταση μεταξύ τους είναι μηδέν.

Κεφάλαιο 6

Αποτελέσματα

6.1 Αποτελέσματα Κατωφλίου Φιλτραρίσματος 10^{-4}

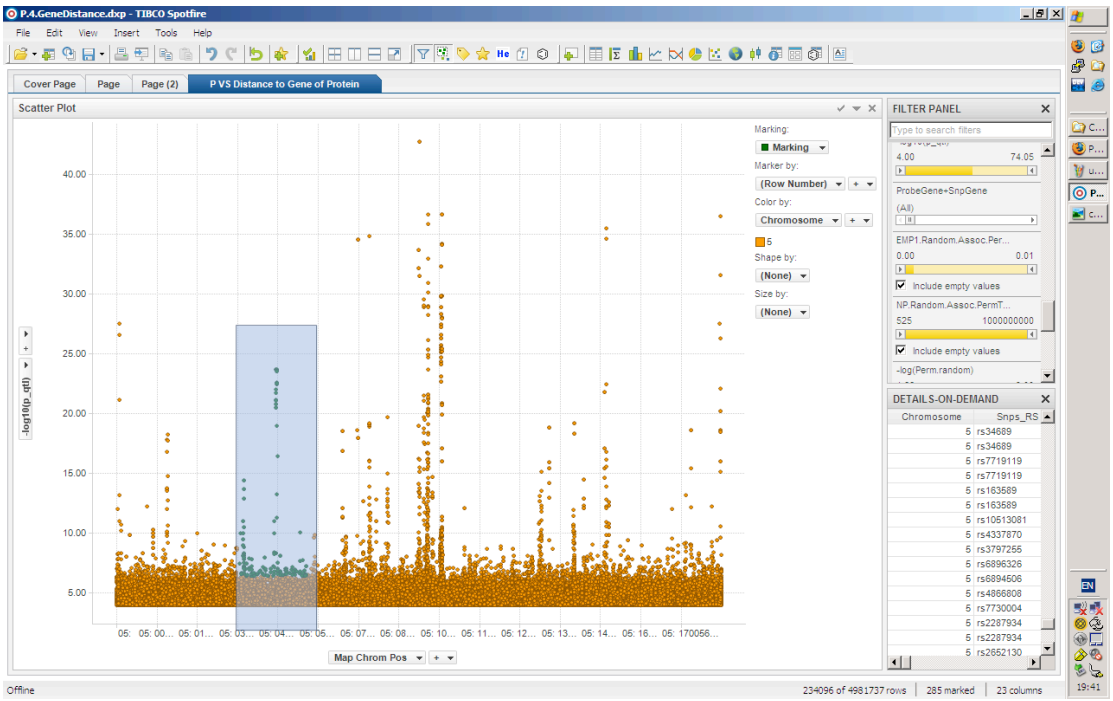
Τα αποτελέσματα που αναλύονται και σχολιάζονται πιο κάτω είναι αυτά που παράχθηκαν κατά στην εφαρμογή της ποσοτικής ανάλυσης (Quantitative Trait Analysis) στα δεδομένα και αργότερα του φιλτραρίσματος τους χρησιμοποιώντας ως κατώφλι την τιμή 10^{-4} για το φιλτράρισμα των πιθανοτήτων σημαντικότητας (p-values). Όπως θα παρατηρηθεί και από τις γραφικές η συγκέντρωση των τιμών πιθανότητας σημαντικότητας (p-values) είναι μεγαλύτερη από αυτή στην περίπτωση χρήσης κατωφλίου 10^{-7} για το φιλτράρισμα. Οι αρνητικοί λογάριθμοι των πιθανοτήτων σημαντικότητας που παρουσιάζονται στους X άξονες των γραφικών παραστάσεων, έχουν ως βάση το 10.



Σχήμα 6.1 Εφαρμογή Κατωφλίου 10^{-4} σε όλα τα αποτελέσματα

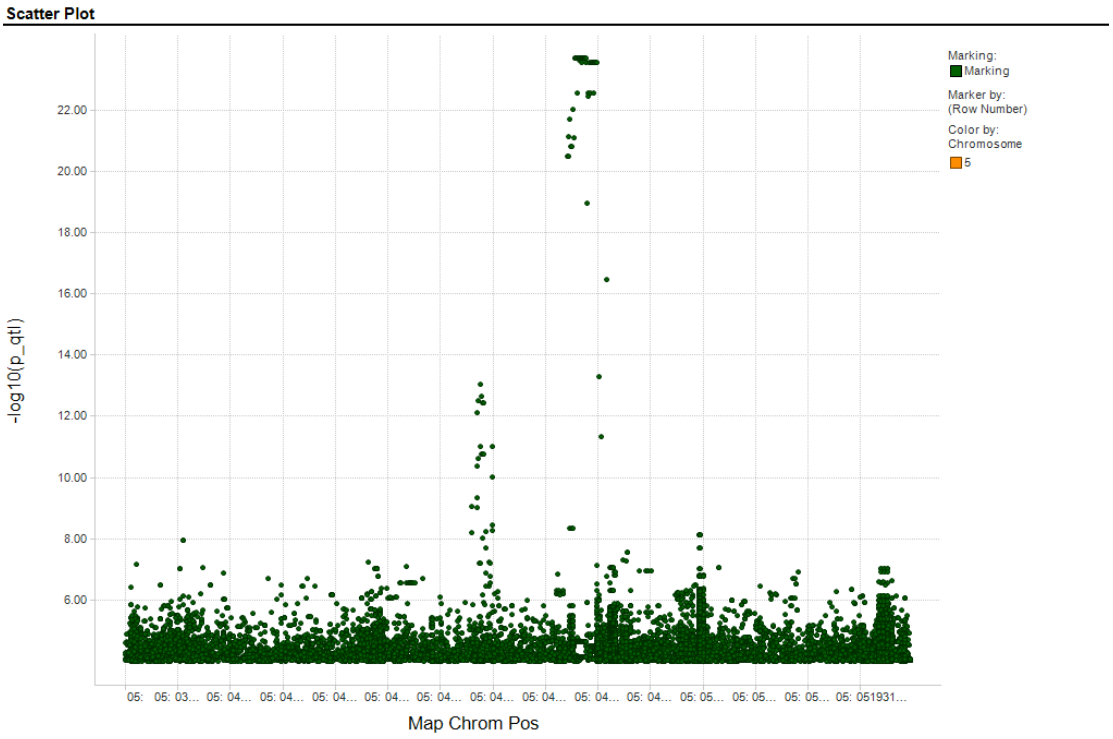
Στο σχήμα 6.1 παρατηρούνται οι τιμές της πιθανότητας σημαντικότητας (p-values) που παράχθηκαν κατά την ποσοτική ανάλυση (QT), για κάθε διαφορετικό από τα 550 χιλιάδες SNPs, σε σχέση με την θέση στην οποία βρίσκονται πάνω σε ένα χρωμόσωμα. Στον άξονα των X βρίσκονται διατεταγμένα τα SNPs με βάση την θέση τους στον γενετικό χάρτη. Δηλαδή για κάθε SNP η θέση του χαρακτηρίζεται από το χρωμόσωμα στο όποιο ανήκει και ακολούθως από τον αριθμό νουκλεοτιδίων από την αρχή του συγκεκριμένου χρωμοσώματος μέχρι το συγκεκριμένο SNP. Για παράδειγμα, το SNP 2:1123 θα βρίσκεται στο 1123^{o} νουκλεοτίδιο από την αρχή του χρωμοσώματος 2. Στον άξονα Ψ διατάσσονται οι αρνητικοί λογάριθμοι των πιθανοτήτων σημαντικότητας για κάθε SNP που μελετήθηκε. Οι πιθανότητες σημαντικότητας είναι προσαρμοσμένες βάσει του λογάριθμου με βάση το 10, αυτό γίνεται γιατί στον τομέα τις γενετικής, οι γραφικές παρουσιάζονται πάντα με αυτό τον τρόπο. Χάρη στην κανονικοποίηση τον αποτελεσμάτων με τον αρνητικό λογάριθμο με βάση το 10, βλέποντας την γραφική είναι δυνατή η εύκολη αναγνώριση του επιπέδου σημαντικότητας, αφού για κάθε αριθμητική μείωση του εκθέτη στην πιθανότητα σημαντικότητας (p-value) σε scientific format, η κανονικοποιημένη τιμή παίρνει μια τιμή ίση με τον εκθέτη της τιμής

πιθανότητας σημαντικότητας (p-value) σε scientific format. Για παράδειγμα p-value e^{-2} δίνει log(0.01)=2. Οι σημαντικότερες από τις τιμές αυτές είναι εκείνες που παρουσιάζονται να έχουν την μεγαλύτερη τιμή στον άξονα των Ψ, με τους αρνητικούς λογάριθμους των πιθανοτήτων σημαντικότητας. Αυτό μπορεί να επεξηγηθεί σύμφωνα με τον ορισμό των πιθανοτήτων σημαντικότητας (p-values), σύμφωνα με τον οποίο οι σημαντικότερες από αυτές, είναι οι μικρότερες που μπορούν να ληφθούν από τα δείγματα. Όσο πιο μικρές είναι οι τιμές τόσο πιο δύσκολο είναι να απορριφθεί η μηδενική υπόθεση. Επομένως αφού στον άξονα των Ψ παρουσιάζονται οι αρνητικοί λογάριθμοι αυτών των τιμών, όπου οι σημαντικότερες από αυτές θα είναι και οι μεγαλύτερες. Τα δείγματα που παρουσιάζονται στη συγκεκριμένη γραφική είναι πυκνότερα στην αντίστοιχη που θα παρουσιαστεί αργότερα για τα αποτελέσματα που παράχθηκαν εφαρμόζοντας κατώφλι 10^{-7}.



Εικόνα 6.1 Επιλογή Συγκεκριμένου Τμήματος Χρωμοσώματος

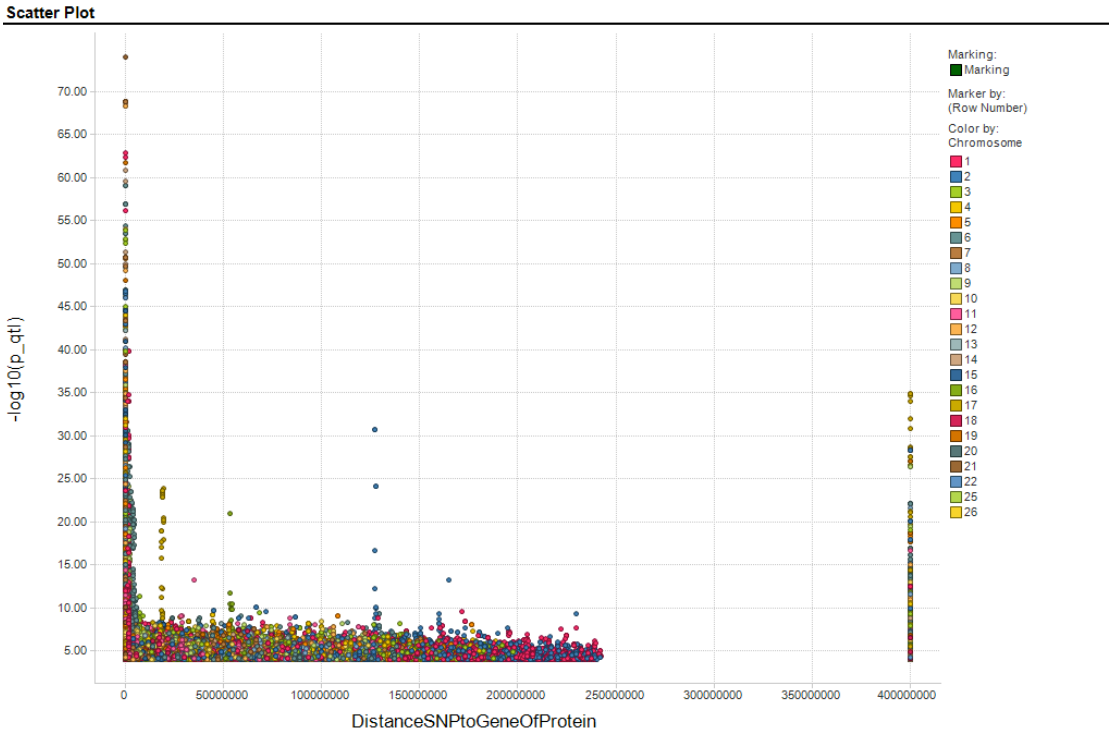
Πιο πάνω παρουσιάζεται μια από τις μεθόδους με τις οποίες γίνεται η επιλογή κάποιας συγκεκριμένης τοποθεσίας σε ένα χρωμόσωμα, που επιθυμείται να μελετηθεί με μεγαλύτερη λεπτομέρια.



Σχήμα 6.2 Αποτελέσματα συγκεκριμένου τμήματος χρωμοσώματος - Κατώφλι 10^{-4}

Στην γραφική παράσταση του σχήματος 6.2 παρατηρούνται οι τιμές της πιθανότητας σημαντικότητας (p-values) της συγκεκριμένης περιοχής που παρουσιάζεται στην εικόνα 6.1. Αποτελεί μεγέθυνση των αποτελεσμάτων και περιορισμό τους σε ένα συγκεκριμένο χρωμόσωμα και πιο συγκεκριμένα στο χρωμόσωμα 5. Στον άξονα των Χ βρίσκονται διατεταγμένες οι διάφορες περιοχές του γονιδίου που έχει επιλεχθεί, ενώ στον άξονα των Ψ διατάσσονται οι αρνητικοί λογάριθμοι των πιθανοτήτων σημαντικότητας για κάθε SNP που μελετήθηκε. Οι σημαντικότερες από τις τιμές αυτές, είναι εκείνες που παρουσιάζονται να έχουν την μεγαλύτερη τιμή στον άξονα των Ψ με τους αρνητικούς λογάριθμους των πιθανοτήτων σημαντικότητας. Αυτό μπορεί να επεξηγηθεί σύμφωνα με τον ορισμό των πιθανοτήτων σημαντικότητας (p-values), που κατά τον οποίο οι σημαντικότερες από αυτές είναι οι μικρότερες που μπορούν να ληφθούν από τα δείγματα, καθώς όσο πιο μικρές είναι οι τιμές τόσο πιο δύσκολο είναι να απορριφθεί η μηδενική υπόθεση. Επομένως αφού στον άξονα των Ψ παρουσιάζονται οι αρνητικοί λογάριθμοι αυτών των τιμών, οι σημαντικότερες από αυτές θα είναι και οι μεγαλύτερες. Επίσης εδώ μπορεί να παρατηρηθεί η μεγαλύτερη συγκέντρωση τιμών πιθανοτήτων σημαντικότητας (pvalues) για τα διάφορα SNPs που

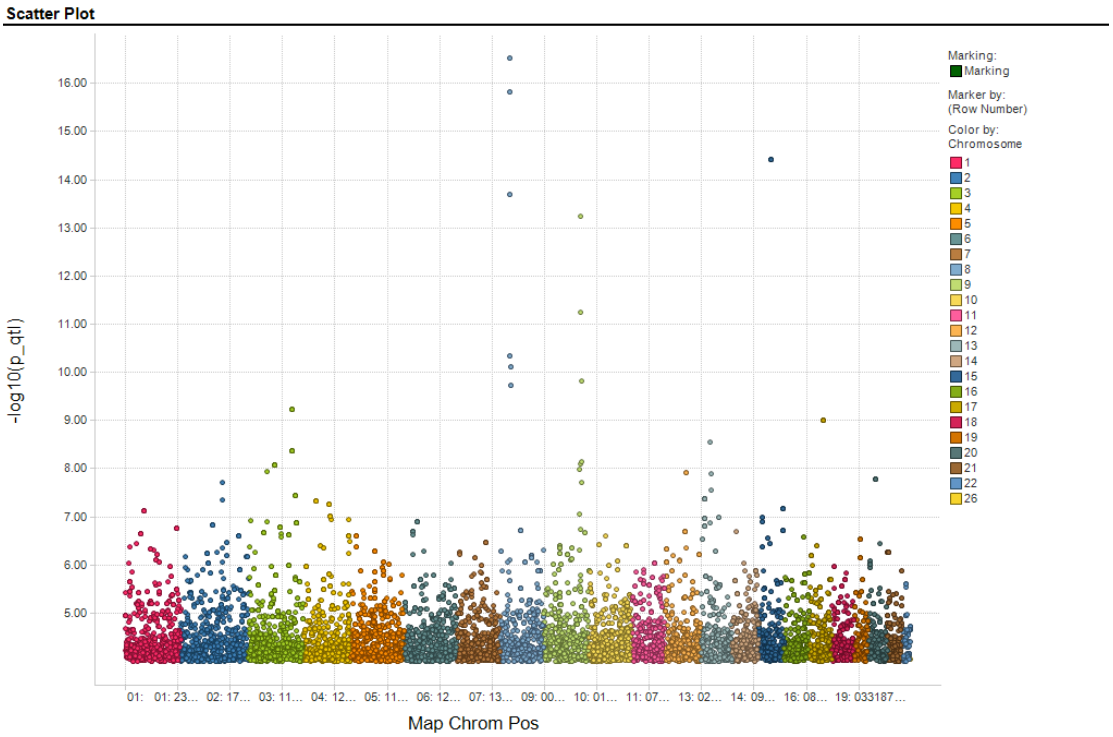
χρησιμοποιήθηκαν στην ανάλυση, λόγω της μεγαλύτερης τιμής του κατωφλίου που δόθηκε σαν είσοδος, κατά τη διαδικασία του φιλτραρίσματος των αποτελεσμάτων.



Σχήμα 6.3 Ελάχιστη απόσταση του SNP μεταξύ μετάγραφου mRNA

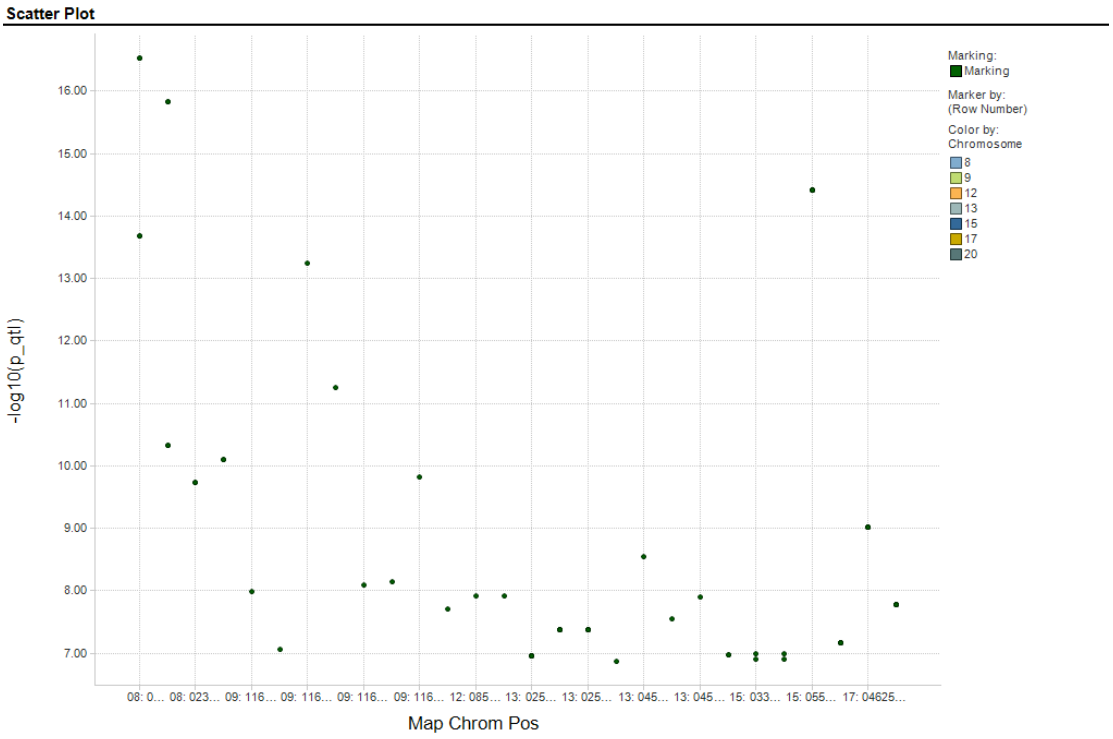
Στην γραφική παράσταση του σχήματος 6.3 παρουσιάζεται η ελαχίστη απόσταση του SNP από τη γενετική θέση από την οποία μεταγράφεται το μόριο του mRNA. Για κάθε αποτέλεσμα στο όποιο τόσο το mRNA όσο και το SNP βρίσκονται στο ίδιο χρωμόσωμα και για όλες τις περιπτώσεις όπου τα συσχετισμένα SNP και mRNA βρίσκονται σε διαφορετικά χρωμοσώματα, δίνεται ο σημαφόρος με τιμή 400000000. Στον άξονα Ψ όπως και στα σχήματα 6.1 και 6.2, διατάσσονται οι αρνητικοί λογάριθμοι των πιθανοτήτων σημαντικότητας (p-value) για κάθε SNP που μελετήθηκε. Παρατηρείται πως η απόσταση για τα περισσότερα από τα SNPs τείνει να μηδενιστεί, δηλαδή έχουν μεγαλύτερη πιθανότητα να είναι γειτονικά με τα μόρια mRNA και να βρίσκονται πάνω στο ίδιο χρωμόσωμα. Αυτό που συμβαίνει όταν η έκφραση του μορίου του mRNA συσχετίζεται από κάποιο SNP που βρίσκεται στο ίδιο χρωμόσωμα και σε μικρή απόσταση από αυτό, στην βιολογία ονομάζεται cis και αποτελεί μια μικρή

αλληλουχία DNA, πάνω στην οποία προσδένονται οι παράγοντες ενεργοποίησης, για να υποβοηθήσουν την έναρξη ή τη λήξη της διαδικασίας της μετάφρασης του σε κάποιο προϊόν πρωτεΐνης. Η γνώση για τον τύπο της συσχέτισης μεταξύ SNP και mRNA αν είναι δηλαδή cis ή trans, είναι εξαιρετικής σημασίας για τη διεξαγωγή νέων πειραμάτων. Ο λόγος της μεγάλης σημασίας της γνώσης αυτής είναι γιατί μπορεί να χρησιμοποιηθεί στη διεξαγωγή βιολογικών πειραμάτων που θα μπορέσουν να μελετήσουν σε βάθος τη λειτουργία των βιολογικών μηχανισμών. Τα SNPs που παρουσιάζονται να απέχουν κατά τη μεγαλύτερη απόσταση από κάποιο μόριο mRNA και που ουσιαστικά δεν βρίσκονται πάνω στο ίδιο χρωμόσωμα είναι πιο πιθανόν να αποτελούν παράγοντες ενεργοποίησης (trans). Η έκφραση του μορίου του mRNA μπορεί να επηρεάζεται έμμεσα από κάποιο SNP που δεν βρίσκεται στο ίδιο χρωμόσωμα με αυτό ή ακόμη και να μην επηρεάζεται καθόλου. Από την γραφική παράσταση μπορεί επίσης να παρατηρηθεί πως η μέγιστη απόσταση των SNPs από κάποια περιοχή έκφρασης ενός μορίου mRNA είναι 400000000. Η συγκέντρωση των τιμών πιθανοτήτων σημαντικότητας είναι εμφανές πως είναι πολύ μεγαλύτερη, λόγω και του μεγαλύτερου όγκου αποτελεσμάτων που χρησιμοποιήθηκαν για την παραγωγή των γραφικών σε αυτό το στάδιο.



Σχήμα 6.4 Αποτελέσματα μετά από εφαρμογή φίλτρου στην οικογένεια γονιδίων TNF

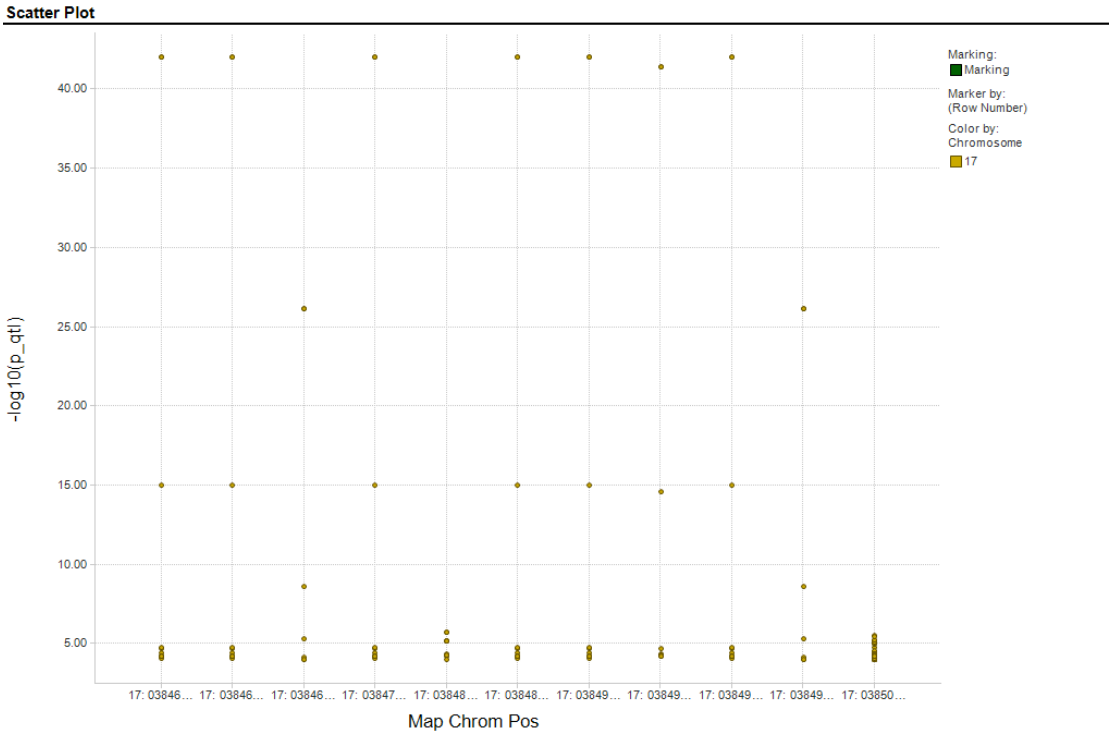
Η γραφική παράσταση του σχήματος 6.4 παρουσιάζει τα αποτελέσματα μετά από την εφαρμογή φίλτρου στην οικογένεια γονιδίων TNF . Στον άξονα των X διατάσσονται οι θέσεις των SNPs πάνω στα διάφορα χρωμοσώματα, ενώ στον άξονα των Ψ τους αρνητικούς λογάριθμους για τις τιμές των πιθανοτήτων σημαντικότητας. Στη δεξιά μεριά του σχήματος 6.4 φαίνονται αριθμημένα τα διάφορα χρωμοσώματα για τα οποία υπάρχουν αποτελέσματα για το συγκεκριμένο γονίδιο. Επιπλέον όπως αναφέρθηκε και νωρίτερα οι σημαντικότερες από τις τιμές που παρουσιάζονται στη γραφική είναι αυτές με την μεγαλύτερη τιμή στον άξονα των Ψ. Δηλαδή οι τιμές με τον μεγαλύτερο αρνητικό λογάριθμο είναι και οι σημαντικότερες βάσει του ορισμού των πιθανοτήτων σημαντικότητας. Τα αποτελέσματα που εμφανίζονται στο συγκεκριμένο σχήμα είναι αποτελέσματα που παράχθηκαν και για τα 26 διαφορετικά χρωμοσώματα του ανθρώπινου γονιδιώματος.



Σχήμα 6.5 Αποτελέσματα της οικογένειας του γονιδίου TNF μετά από την εφαρμογή φιλτραρίσματος

Η γραφική παράσταση του σχήματος 6.5 παρουσιάζει τα αποτελέσματα μετά από την εφαρμογή φίλτρου στην οικογένεια γονιδίου TNF . Στον άξονα των X διατάσσονται οι θέσεις που κατέχουν τα SNPs πάνω στα διάφορα χρωμοσώματα, ενώ στον άξονα των Ψ τις τιμές των πιθανοτήτων σημαντικότητας προσαρμοσμένες βάσει του αρνητικού λογάριθμου με βάση το 10. Στη δεξιά μεριά της γραφικής στο πάνω μέρος, βρίσκονται αριθμημένα τα διάφορα χρωμοσώματα για τα οποία υπάρχουν αποτελέσματα σε αυτή τη γραφική, για το συγκεκριμένο γονίδιο. Από τις τιμές των αρνητικών λογαρίθμων των πιθανοτήτων σημαντικότητας που παρατηρούνται στο σχήμα 6.5, μπορεί να εξαχθεί το συμπέρασμα πως είναι στατιστικώς σημαντικά τα συγκεκριμένα αποτελέσματα, καθώς οποιαδήποτε τιμή πιθανότητας σημαντικότητας που είναι μικρότερη από 0.05 αποτελεί μία στατιστικά σημαντική τιμή, βάσει του ορισμού των πιθανοτήτων σημαντικότητας που θέτει σημαντική οποιαδήποτε τιμή είναι μικρότερη από το 0.05. Επιπλέον παρατηρείται από τη γραφική πως τα περισσότερα από τα αποτελέσματα είναι αρκετά μικρότερα από το 0.05 με τον αρνητικό λογάριθμο να ξεκινά από το 7 και πάνω. Επομένως, οι παρατηρήσεις οδηγούν στο συμπέρασμα πως οι τιμές των πιθανοτήτων

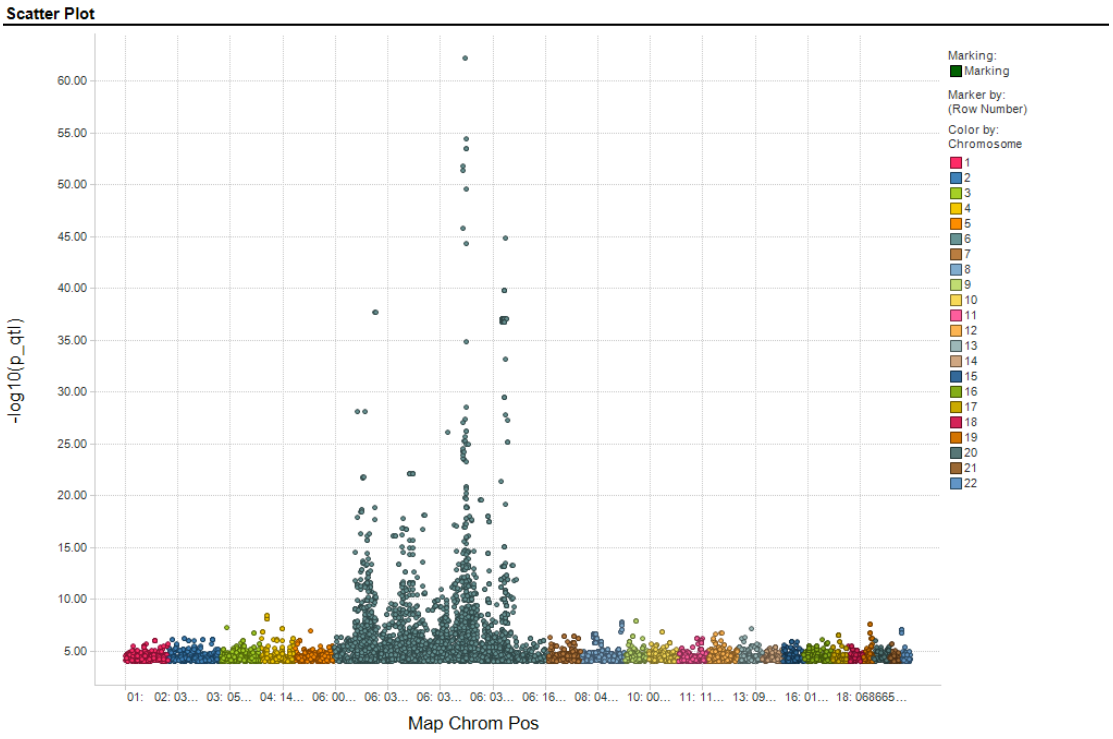
σημαντικότητας (p-values) για την οικογένεια γονιδίου TNF, είναι εξαιρετικής σημασίας με τα SNPs να σχετίζονται άμεσα με τις περιοχές έκφρασης του mRNA πάνω στο γονίδιο. Η γραφική αυτή παράσταση παράχθηκε μετά από την μεγέθυνση των αποτελεσμάτων επιλογής της συγκεκριμένης οικογένειας του γονιδίου TNF.



Σχήμα 6.6 Αποτελέσματα της οικογένειας του γονιδίου BRCA1 μετά από την εφαρμογή φιλτραρίσματος

Η γραφική παράσταση του σχήματος 6.5 παρουσιάζει τα αποτελέσματα μετά από την εφαρμογή φίλτρου στην οικογένεια γονιδίου BRCA1. Το φίλτρο που εφαρμόστηκε σε αυτή την περίπτωση είναι διαφορετικό σε σύγκριση με το φιλτράρισμα που εφαρμόστηκε πάνω στο γονίδιο TNF (σχήμα 6.4). Οι άξονες εντούτοις παρουσιάζουν τις ίδιες μονάδες. Ο διαφορά είναι ότι το φίλτρο προσαρμόζεται σε κάθε διαφορετική περίπτωση ξεχωριστά. Στον άξονα των X διατάσσονται οι θέσεις των SNPs πάνω στα διαφορετικά χρωμοσώματα, ενώ στον άξονα των Ψ τις τιμές των πιθανοτήτων σημαντικότητας. Στο πλάι φαίνονται αριθμημένα τα διάφορα χρωμοσώματα για τα οποία υπάρχουν αποτελέσματα για το συγκεκριμένο γονίδιο. Επιπλέον σε αυτή τη

γραφική όπως και σε όλες τις γραφικές των σχημάτων που προηγούνται, οι σημαντικότερες από τις τιμές που παρουσιάζονται στον άξονα των Ψ είναι αυτές με την μεγαλύτερη τιμή. Δηλαδή οι τιμές με τον μεγαλύτερο αρνητικό λογάριθμο είναι και οι σημαντικότερες βάσει του ορισμού των πιθανοτήτων σημαντικότητας. Μπορεί να παρατηρηθεί ότι τα αποτελέσματα για το συγκεκριμένο γονίδιο αναφέρονται μόνο σε ένα μόνο χρωμόσωμα, το χρωμόσωμα 17. Δηλαδή τα SNPs που βρίσκονται στο συγκεκριμένο γονίδιο είναι όλα συγκεντρωμένα σε ένα και μόνο χρωμόσωμα, το χρωμόσωμα 17. Επιπλέον οι λογαριθμικές τιμές των πιθανοτήτων σημαντικότητας (p-value) φαίνονται να είναι υψίστης σημασίας σε σχέση με προηγούμενα αποτελέσματα καθώς είναι και συγκριτικά μεγαλύτερες. Λαμβάνοντας υπόψη και τον ορισμό των πιθανοτήτων σημαντικότητας όπως έχει ορισθεί στη στατιστική, θέτει ως σημαντικές τις τιμές αυτές που είναι μικρότερες από το 0.05. Στο γονίδιο αυτό παρατηρείται πως οι τιμές για τις πιθανότητες σημαντικότητας δεν είναι εξαιρετικής σημασίας καθώς η μεγαλύτερη συγκέντρωση των τιμών αυτών εμφανίζονται κοντά στην τιμή 5 του άξονα Χ. Αναφερόμενοι στον ορισμό των πιθανοτήτων σημαντικότητας (p-value) που θέτει ώς σημαντικές τις τιμές που είναι μικρότερες από 0.05, μπορεί να παρατηρηθεί πως σε αυτή την περίπτωση τα αποτελέσματα δεν είναι τόσο μεγάλης σημασίας όπως αυτά του σχήματος 6.5, εντούτοις υπάρχουν κάποιες τιμές από αυτές που φαίνονται να είναι τεράστιας σημασίας καθώς φτάνουν και μέχρι την τιμή 40 του αρνητικού λογάριθμου στον άξονα των Χ και κάποιες άλλες μέχρι το 16. Η μεγαλύτερη συγκέντρωση τιμών όμως παρατηρείται κοντά στο 5 που δεν αποτελούν σημαντικές τιμές καθώς βρίσκονται πάνω στο όριο που θέτει ο ορισμός των τιμών των πιθανοτήτων σημαντικότητας .

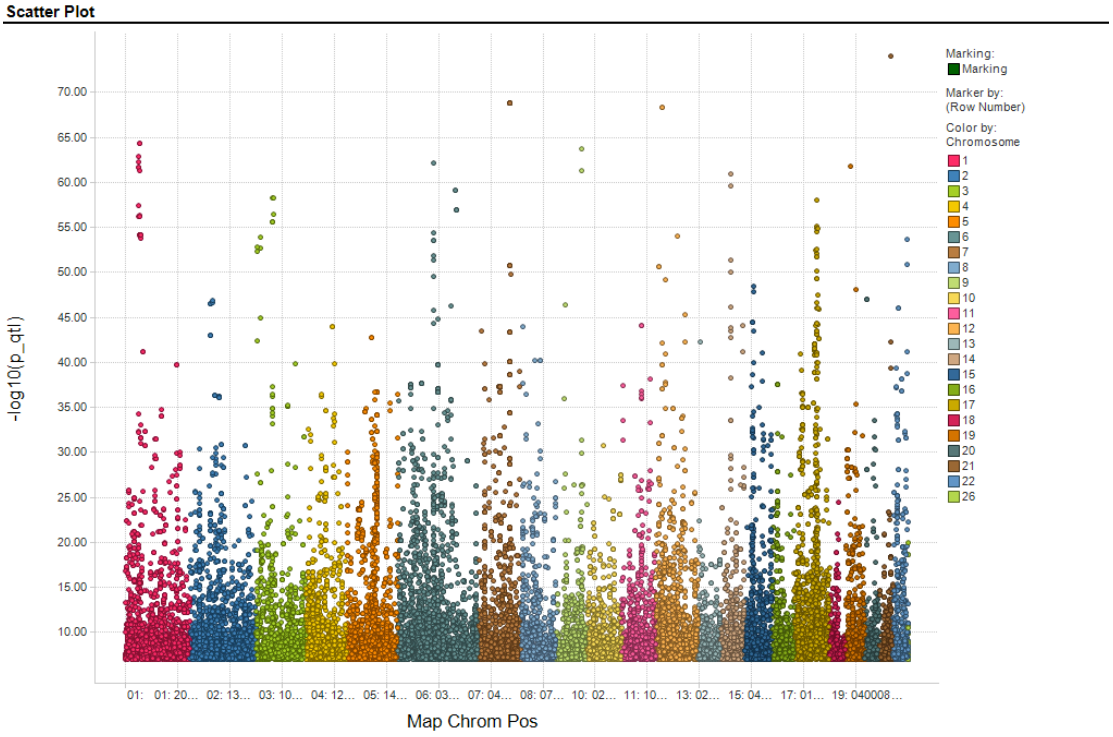


Σχήμα 6.7 Αποτελέσματα της οικογένειας του γονιδίου HLA μετά από την εφαρμογή φιλτραρίσματος

Το σχήμα 6.7 παρουσιάζει τα αποτελέσματα μετά από την εφαρμογή φίλτρου στην οικογένεια γονιδίων HLA. Στον άξονα των X διατάσσονται οι θέσεις των SNPs πάνω στα διάφορα χρωμοσώματα, ενώ στον άξονα των Ψ οι τιμές των πιθανοτήτων σημαντικότητας προσαρμοσμένες βάσει του αρνητικού λογάριθμου με βάση το 10. Στο πλάι αριθμημένα φαίνονται τα διάφορα χρωμοσώματα για τα οποία υπάρχουν αποτελέσματα σε αυτή τη γραφική για το συγκεκριμένο γονίδιο. Όπως είναι ήδη γνωστό και από τον ορισμό των πιθανοτήτων σημαντικότητας οι σημαντικότερες από τις τιμές, από αυτές που παρουσιάζονται στη γραφική, είναι αυτές με την μεγαλύτερη τιμή στον άξονα των Ψ. Μπορεί να παρατηρηθεί πως σε σχέση με τα αποτελέσματα που παρουσιάστηκαν για τις υπόλοιπες οικογένειες γονιδίων, οι τιμές των πιθανοτήτων σημαντικότητας, είναι οι μεγαλύτερες που έχουν παρατηρηθεί και επομένως μπορούν να ληφθούν υπόψη και ως υψίστης σημασίας με τις μικρότερες τιμές των πιθανοτήτων σημαντικότητας (p-values).

6.2 Αποτελέσματα Κατωφλίου Φιλτραρίσματος 10^{-7}

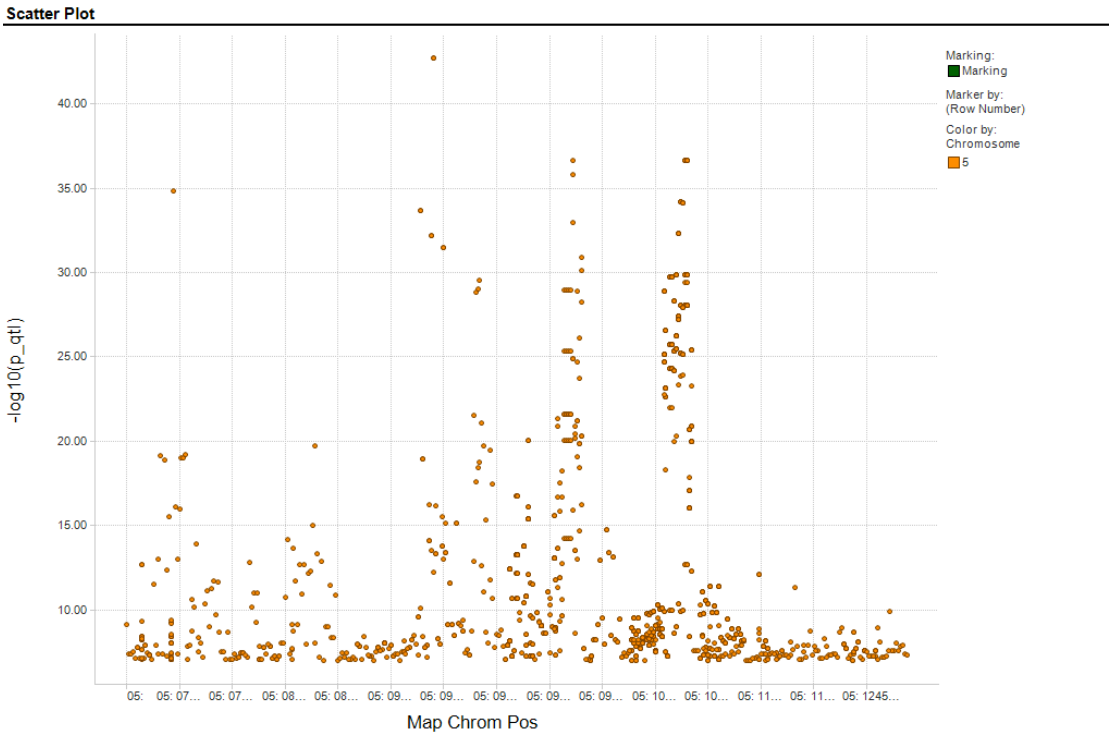
Τα πιο κάτω αποτελέσματα που αναλύονται και σχολιάζονται είναι αυτά που παράχθηκαν κατά στην εφαρμογή της ποσοτικής ανάλυσης (Quantitative Trait Analysis) στα δεδομένα και αργότερα του φιλτραρίσματος τους χρησιμοποιώντας ως κατώφλι την τιμή 10^{-7} για το φιλτράρισμα των πιθανοτήτων σημαντικότητας (p-values). Οι αρνητικοί λογάριθμοι των πιθανοτήτων σημαντικότητας που παρουσιάζονται στους Χ άξονες των γραφικών παραστάσεων, έχουν ως βάση το 10. Η εφαρμογή του κατωφλίου αυτού περιορίζει περεταίρω τις τιμές παρατήρησης αυτού του υποκεφαλαίου στις πιο σημαντικές. Αυτό γίνεται γιατί εφαρμόζεται ένα συγκριτικά πολύ μικρότερο κατώφλι από το 10^{-4} τα αποτελέσματα του οποίου μελετήθηκαν στο υποκεφάλαιο 6.1.



Σχήμα 6.8 Εφαρμογή κατωφλίου 10^{-7} σε όλα τα αποτελέσματα

Στην γραφική παράσταση του σχήματος 6.8, παρατηρούνται οι τιμές των πιθανοτήτων σημαντικότητας (p-values) μετά από την εφαρμογή της ποσοτικής ανάλυσης (QT) στα δεδομένα. Οι τιμές που παρουσιάζονται σε αυτή τη γραφική αναφέρονται στα 550 χιλιάδες διαφορετικά SNPs που χρησιμοποιήθηκαν για την ανάλυση, και σχετίζονται με την θέση στην οποία βρίσκονται πάνω σε ένα χρωμόσωμα. Στον άξονα των Χ βρίσκονται διατεταγμένα τα διάφορα χρωμοσώματα του ανθρώπινου οργανισμού, ενώ στον άξονα των Ψ διατάσσονται οι αρνητικοί λογάριθμοι των πιθανοτήτων σημαντικότητας για κάθε SNP που μελετήθηκε. Οι σημαντικότερες από τις τιμές αυτές είναι εκείνες που παρουσιάζονται να έχουν την μεγαλύτερη τιμή στον άξονα των Ψ, στον οποία φαίνονται οι αρνητικοί λογάριθοι των πιθανοτήτων σημαντικότητας. Αυτό μπορεί να επεξηγηθεί σύμφωνα με τον ορισμό των πιθανοτήτων σημαντικότητας (p-values), που κατά τον οποίο οι σημαντικότερες από αυτές είναι οι μικρότερες που μπορούν να ληφθούν από τα δείγματα, καθώς όσο πιο μικρές είναι οι τιμές τόσο πιο δύσκολο είναι να απορριφθεί η μηδενική υπόθεση. Σύμφωνα με τον ορισμό μία πιθανότητα σημαντικότητας μπορεί να θεωρηθεί ως σημαντική εάν αυτή είναι μικρότερη από το κατώφλι του 0.05. Αφού στον άξονα των Ψ παρουσιάζονται οι

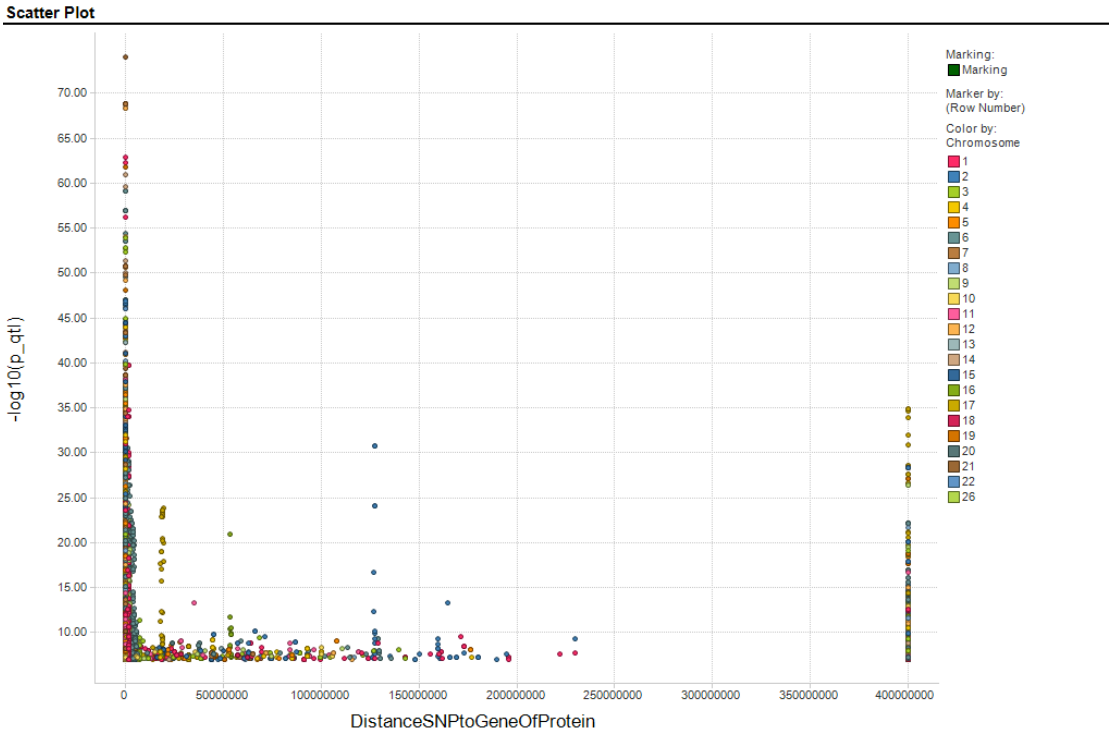
αρνητικοί λογάριθμοι αυτών των τιμών, οι σημαντικότερες από αυτές θα είναι και οι μεγαλύτερες.



Σχήμα 6.9 Αποτελέσματα συγκεκριμένου τμήματος χρωμοσώματος - Κατώφλι 10^{-7}

Στην γραφική παράσταση του σχήματος 6.9, παρατηρούνται οι τιμές της πιθανότητας σημαντικότητας (p-values) για τα διάφορα SNPs σε σχέση με τη θέση που κατέχουν σε κάποιο χρωμόσωμα. Αποτελεί μεγέθυνση των αποτελεσμάτων και περιορισμό τους σε ένα συγκεκριμένο χρωμόσωμα και πιο συγκεκριμένα στο χρωμόσωμα 6. Στον άξονα των Χ βρίσκονται διατεταγμένες οι διάφορες περιοχές του γονιδίου που έχει επιλεχθεί, ενώ στον άξονα των Ψ διατάσσονται οι αρνητικοί λογάριθμοι των πιθανοτήτων σημαντικότητας για κάθε SNP που μελετήθηκε. Οι σημαντικότερες από τις τιμές αυτές είναι εκείνες που παρουσιάζονται να έχουν την μεγαλύτερη τιμή στον άξονα των Ψ με τους αρνητικούς λογάριθμους των πιθανοτήτων σημαντικότητας. Παρατηρώντας την γραφική μπορεί να γίνει αντιληπτό, πως τα αποτελέσματα είναι εξαιρετικής σημασίας καθώς οι τιμές που παίρνουν σε σχέση με τον άξονα των Χ (αρνητικός λογάριθμος των πιθανοτήτων σημαντικότητας), ξεκινούν από την τιμή 15. Άν ανατρέξουμε στον ορισμό των πιθανοτήτων σημαντικότητας που θέτει ώς σημαντική κάθε μία από αυτές τις τιμές

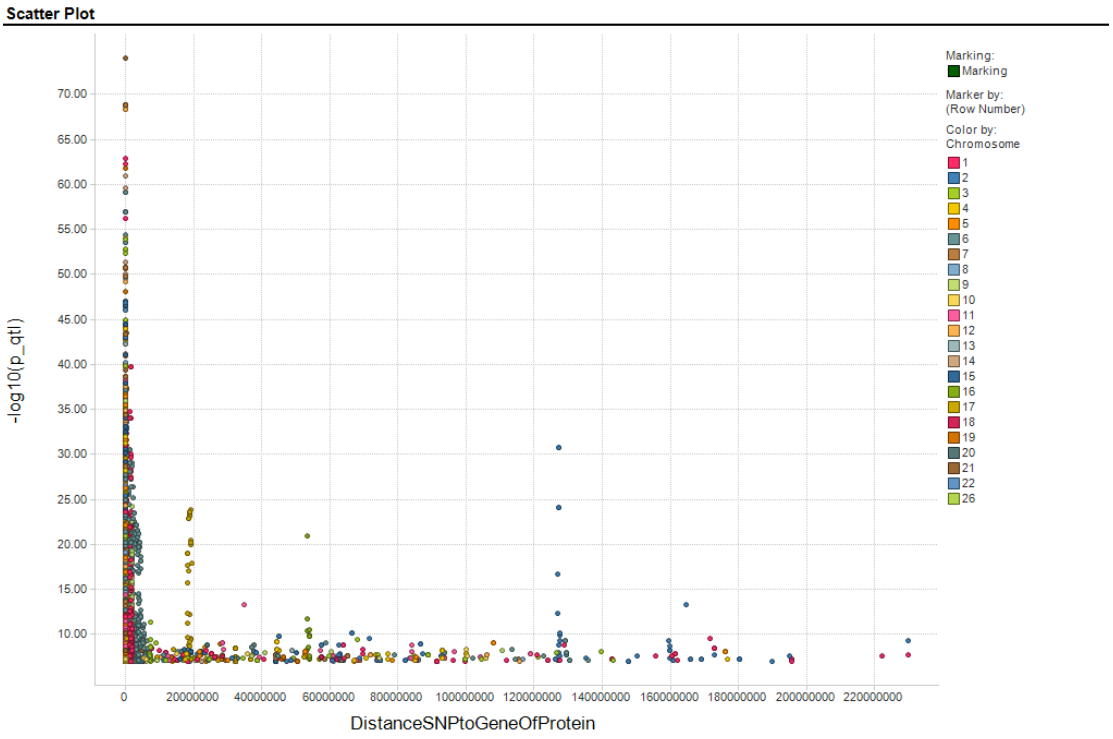
που είναι μικρότερη από το 0.05 (5 βάσει του αρνητικού λογαρίθμου), παρατηρείται πως για αυτό το χρωμόσωμα οι τιμές που παράχθηκαν είναι τεράστιας σημασίας καθώς απέχουν κατα πολύ από το κατώφλι του ορισμού των πιθανοτήτων σημαντικότητας. Επομένως τα SNPs των οποίων η τοποθεσία βρίσκεται πάνω στο χρωμόσωμα 6 (HLA) είναι τεράστιας σημασίας και είναι πολύ πιθανόν να επηρεάζουν την έκφραση κάποιου μορίου mRNA που μεταγράφεται από κάποια περιοχή αυτού του χρωμοσώματος.



Σχήμα 6.10 Ελάχιστη απόσταση των SNPs από τα μετάγραφα mRNA

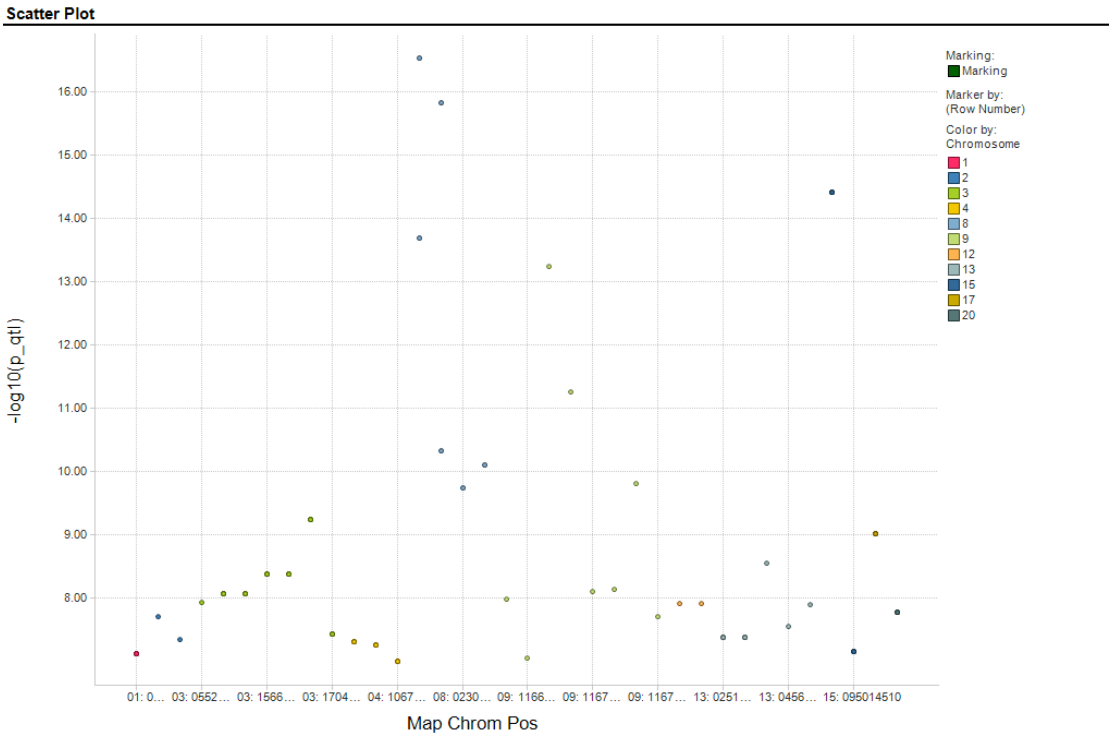
Στο σχήμα 6.10, παρουσιάζεται η απόσταση των SNPs από τη γενετική θέση στην οποία βρίσκονται τα μόρια mRNA πάνω στα χρωμοσώματα. Στον άξονα των X βρίσκονται διατεταγμένες οι διάφορες αποστάσεις των SNPs από τα μόρια mRNA, ενώ στον άξονα των Ψ διατάσσονται οι αρνητικοί λογάριθμοι των πιθανοτήτων σημαντικότητας (p-value) για κάθε SNP που μελετήθηκε. Παρατηρείται πως η απόσταση για τα περισσότερα από τα SNPs τείνει να μηδενιστεί, δηλαδή έχουν μεγαλύτερη πιθανότητα να είναι γειτονικά με τα μόρια mRNA και να βρίσκονται πάνω στο ίδιο χρωμόσωμα. Αυτό σημαίνει πως η έκφραση των μορίων του mRNA είναι πολύ πιθανό να επηρεάζεται άμεσα από κάποιο SNP, που κατέχει όσο το δυνατό κοντινότερη

θέση με αυτό, πάνω στο ίδιο χρωμόσωμα. Οι τιμές αυτές που υποδεικνύουν SNPs που κατέχουν γειτονική θέση στο χρωμόσωμα με κάποιο mRNA αποτελούν στοιχεία ενεργοποίησης της έκφρασης του συγκεκριμένου μορίου mRNA που εμφανίζεται να είνι γειτονικό με αυτό (cis). SNPs που παρουσιάζονται να απέχουν κατά τη μεγαλύτερη απόσταση από κάποιο μόριο mRNA και που ουσιαστικά δεν βρίσκονται πάνω στο ίδιο χρωμόσωμα, είναι πιθανόν να αποτελούν παράγοντες ενεργοποίησης της έκφρασης των μορίων mRNA (trans) ή σε περίπτωση που η τιμή της πιθανότητας σημαντικότητας τους είναι μεγαλύτερη από το 0.05 που θέτει ο ορισμός των πιθανοτήτων σημαντικότητας, είναι πιθανόν να μην επηρεάζουν καθόλου την έκφραση του mRNA. Από την γραφική παράσταση μπορούμε να παρατηρήσουμε πως η μέγιστη αυτή απόσταση των SNPs από κάποιο μόριο mRNA είναι 400000000. Για αυτή την περίπτωση SNPs μπορεί να γίνει η υπόθεση πως κατα μεγάλη πιθανότητα αποτελούν παράγοντες ενεργοποίησης (trans), ενώ όσο πλησιάζουν οι τιμές το 0 η πιθανότητα να αποτελούν παράγοντες ενεργοποίησης τα SNPs μικραίνει ενώ αυξάνεται η πιθανότητα τα αποτελούν στοιχεία cis. Τα cis στοιχεία φαίνονται στις τιμές που τείνουν να πλησιάσουν πιο κοντά στο 0 σε σχέση με τον άξονα των Ψ που καθορίζει την απόσταση των SNPs πάνω στο χρωμόσωμα σε σχέση με τη θέση από την οποία μεταγράφεται το mRNA. Για τις ενδιάμεσες τιμές αυτές που παρατηρείται πως απομακρύνονται από το 0 και έχουν μικρότερη απόσταση από 400000000, δεν μπορεί να προσδιοριστεί με σιγουριά αν αυτές αποτελούν περιπτώσεις cis ή trans και οι τιμές των αρνητικών λογαρίθμων των πιθανοτήτων σημαντικότητας παρατηρείται να μειώνονται όπως επίσης και η σημασία των αποτελεσμάτων αυτών.



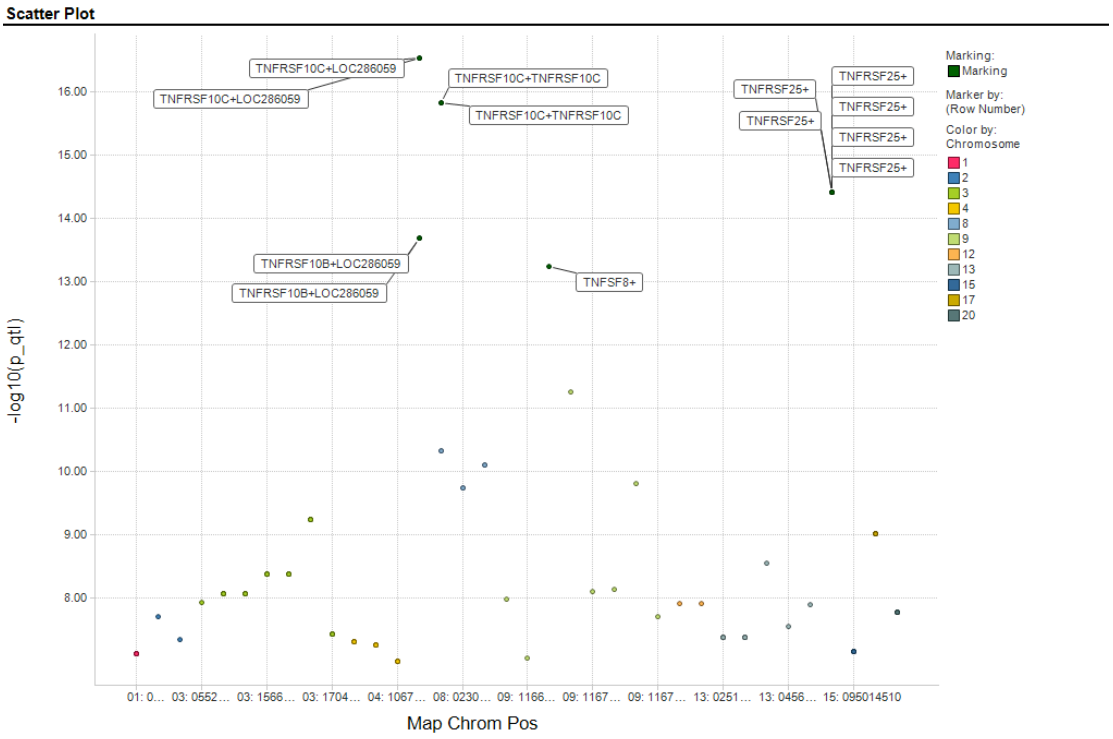
Σχήμα 6.11 Ελάχιστη απόσταση του SNP μεταξύ μετάγραφο mRNA

Η γραφική παράσταση του σχήματος 6.11 όπως και αυτή του σχήματος 6.10 παρουσιάζει, την απόσταση των SNPs, από τη γενετική θέση στην οποία βρίσκονται τα μόρια mRNA πάνω στα χρωμοσώματα. Σε αυτή την περίπτωση, γίνεται μια μεγένθυνση των αποτελεσμάτων έτσι ώστε να μπορεί να μελετηθεί με μεγαλύτερη λεπτομέρια άν η απόσταση των SNPs από την θέση απ' όπου μεταγράφονται τα μόρια mRNA, ώστε να μπορεί να αποφασιστεί ποιά από αυτά αποτελούν περιπτωσεις cis. Παρατηρώντας τη συγκέντρωση των σημείων κοντά στο 0, φαίνεται πως η μεγαλύτερη συγκέντρωση αυτών, αναφέρεται στα χρωμοσώματα 1 και 20 (βλέπε διάταξη χρωμοσωμάτων στα δεξιά). Επομένως τα SNPs που βρίσκονται πάνω σε αυτά τα χρωμοσώματα αποτελούν κατά μεγάλη πιθανότητα περιπτώσεις cis. Επιπλέον και οι τιμές των αρνητικών λογάριθμων των πιθανοτήτων σημαντιότητας παρατηρούνται να είναι πολύ μεγαλύτερες από αυτές των άλλων χρωμοσωμάτων με ορισμένες εξαιρέσεις.



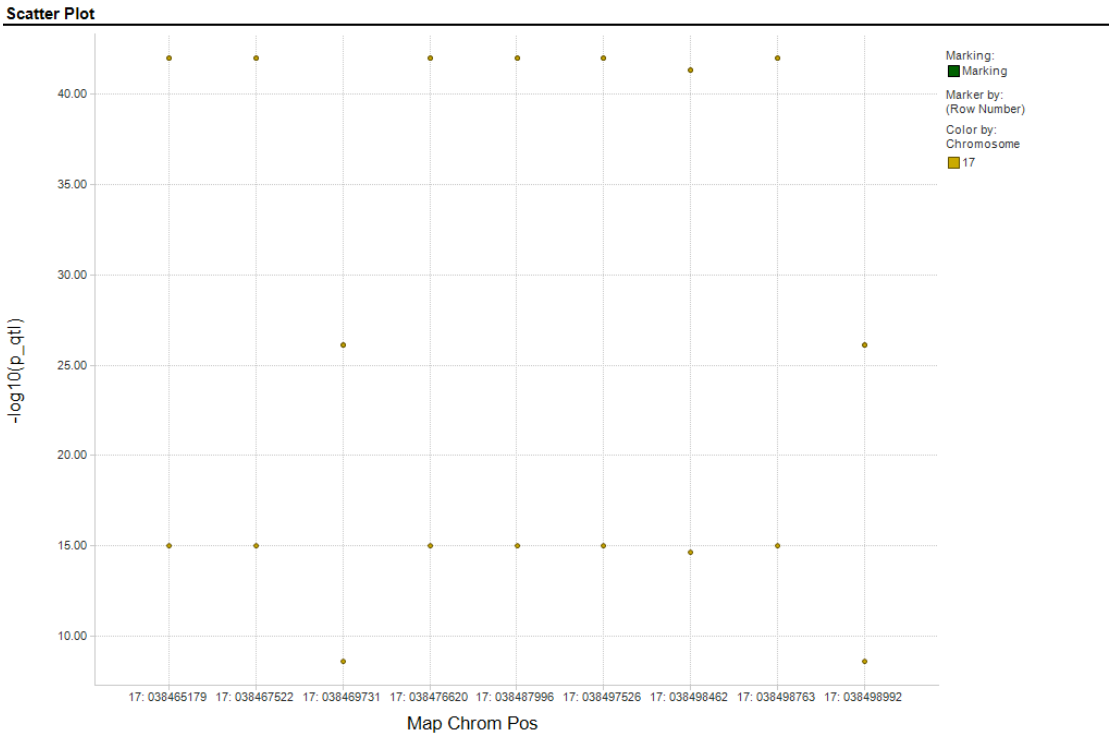
Σχήμα 6.12 Αποτελέσματα της οικογένειας του γονιδίου TNF μετά από την εφαρμογή φιλτραρίσματος

Το σχήμα 6.12 παρουσιάζει τα αποτελέσματα μετά από την εφαρμογή φίλτρου στην οικογένεια γονίδιων TNF. Στον άξονα των X διατάσσονται οι θέσεις των SNPs πάνω στα διάφορα χρωμοσώματα, ενώ στον άξονα των Ψ τις τιμές των πιθανοτήτων σημαντικότητας. Στο πλάι δεξιά φαίνονται αριθμημένα τα διάφορα χρωμοσώματα για τα οποία υπάρχουν αποτελέσματα για τη συγκεκριμένη οικογένεια γονιδίου. Οι περισσότερες απο τις τιμές αυτές παρουσιάζονται να φτάνουν μέχρι το 8 συγκρίνοντας βάσει του άξονα των X, με ορισμένες από αυτές να βρίσκονται και κάτω από το 5, καθιστώντας τις μη σημαντικές βάσει του στατιστικού ορισμού των πιθανοτήτων σημαντικότητας (p-values), ενώ άλλες ξεπερνούν κατά πολύ το όριο αυτό καθιστώντας τα αποτελέσματα σημαντικά σε σχέση με τα υπόλοιπα. Γενικά μπορεί να βγεί το συμπέρασμα πως η σημασία των αποτελεσμάτων για αυτή τη συγκεκριμένη οικογένεια γονιδίου δεν είναι μεγάλη έστω κι αν κάποια για κάποια από αυτά παρατηρούνται στατιστικά σημαντικές τιμές στον άξονα των X.



Σχήμα 6.13 Αποτελέσματα της οικογένειας του γονιδίου TNF μετά από την εφαρμογή φιλτραρίσματος και επιλογή των σημαντικότερων αποτελεσμάτων

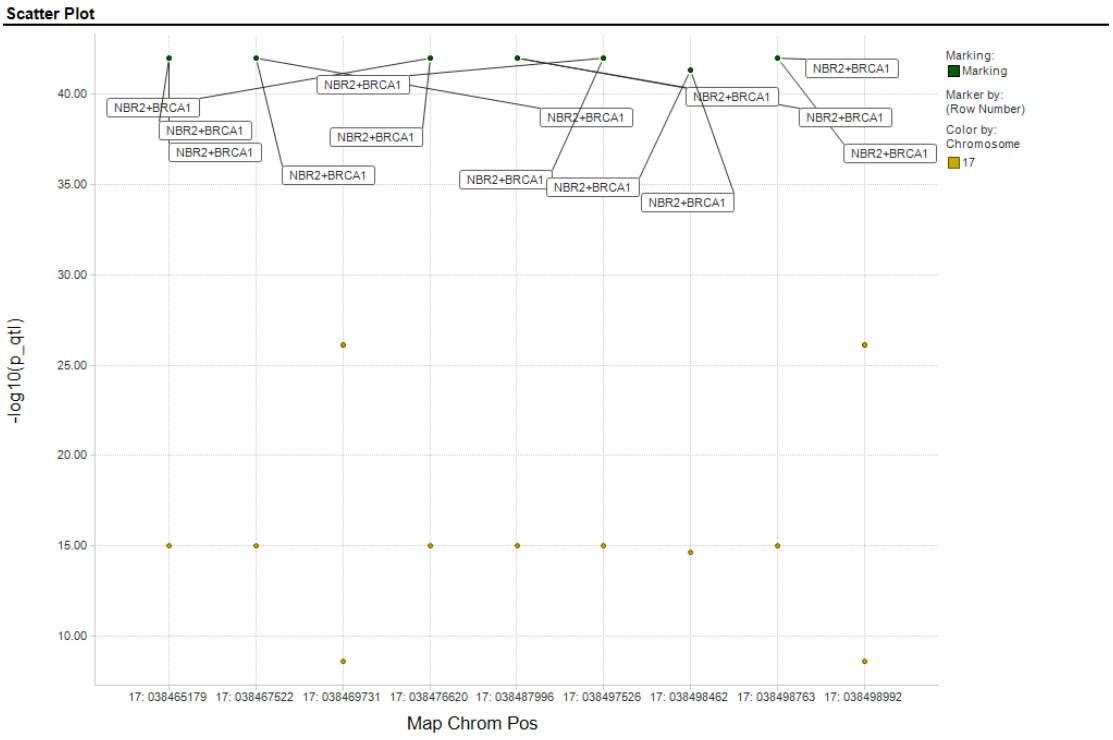
Το σχήμα 6.13 παρουσιάζει ακριβώς τα ίδια αποτελέσματα με το σχήμα 6.12 δηλαδή, με την εφαρμογή φίλτρου πάνω στο γονίδιο TNF. Η διαφορά είναι πως σε αυτή την περίπτωση έχουν επιλεγεί οι σημαντικότερες από τις λογαριθμικές τιμές των πιθανοτήτων σημαντικότητας. Η επιλογή αυτή έγινε για την προβολή περεταίρω πληροφοριών των σημαντικότερων από τις τιμές του γονιδίου αυτού και για επίδειξη των υπηρεσιών που μπορεί να μας παρέχει η συγκεκριμένη εφαρμογή.



Σχήμα 6.14 Αποτελέσματα της οικογένειας του γονιδίου BRCA1 μετά από την εφαρμογή φιλτραρίσματος

Το σχήμα 6.14 παρουσιάζει τα αποτελέσματα μετά από την εφαρμογή φίλτρου στην οικογένεια γονιδίου BRCA1. Το φίλτρο που εφαρμόστηκε σε αυτή την περίπτωση είναι διαφορετικό σε σύγκριση με το φιλτράρισμα που εφαρμόστηκε πάνω στο γονίδιο TNF. Στους άξονες εντούτοις παρουσιάζονται οι ίδιες μονάδες. Ο μόνη διαφορά είναι ότι το φίλτρο προσαρμόζεται σε κάθε διαφορετική περίπτωση ξεχωριστά. Στον άξονα των X διατάσσονται οι θέσεις των SNPs πάνω στα διαφορετικά χρωμοσώματα, ενώ στον άξονα των Ψ οι τιμές των πιθανοτήτων σημαντικότητας προσαρμοσμένες βάσει του αρνητικού λογάριθμου. Δεξιά στο πλάι φαίνονται αριθμημένα τα διάφορα χρωμοσώματα για τα οποία υπάρχουν αποτελέσματα για το συγκεκριμένο γονίδιο. Επιπλέον όπως αναφέρθηκε και νωρίτερα οι σημαντικότερες από τις τιμές είναι αυτές που παρουσιάζονται στη γραφική είναι αυτές με την μεγαλύτερη τιμή στον άξονα των Ψ. Δηλαδή οι τιμές με τον μεγαλύτερο λογάριθμο είναι και οι σημαντικότερες βάσει του ορισμού των πιθανοτήτων σημαντικότητας [15]. Μπορεί να παρατηρηθεί ότι τα αποτελέσματα για το συγκεκριμένο γονίδιο αναφέρονται μόνο σε ένα μόνο χρωμόσωμα, το χρωμόσωμα 17. Δηλαδή τα SNPs που βρίσκονται στο συγκεκριμένο

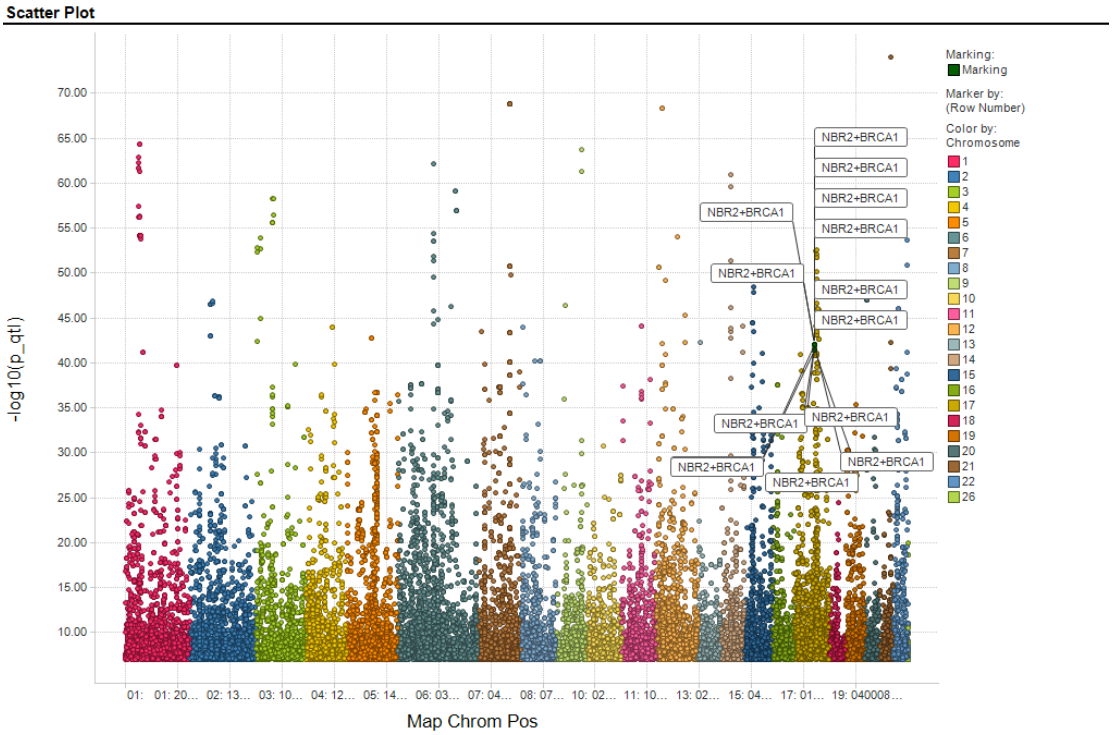
γονίδιο είναι όλα συγκεντρωμένα σε ένα και μόνο χρωμόσωμα, το χρωμόσωμα 17. Επιπλέον οι λογαριθμικές τιμές των πιθανοτήτων σημαντικότητας (p-value) φαίνονται να είναι υψίστης σημασίας σε σχέση με προηγούμενα αποτελέσματα καθώς είναι και συγκριτικά μεγαλύτερες. Τα λιγότερο σημαντικά αποτελέσματα είναι περίπου στο μισό του 15 όπως πολύ καθαρά μπορεί να παρατηρηθεί στο σχήμα 6.14. Τα υπόλοιπα ξεκινούν από την τιμή 16 όσον αφορά τον αρνητικό λογάριθμο που παρουσιάζεται στον άξονα των Ψ. Σημαίνει πως ξεπερνούν κατά πολύ το ανώτατο όριο που θέτει ο ορισμός των πιθανοτήτων σημαντικότητας 0.05 [15], καθώς οι τιμές των αποτελεσμάτων ανέρχοναι γύρω στο 0.016 όπως μπορεί να παρατηρηθεί και από τη γραφική παράσταση. Σύμφωνα με τις παρατηρήσεις αυτές, οι τιμές των συσχετίσεων για την οικογένεια γονιδίων BRCA1 αποτελούν τιμές μεγάλης σημαντικότητας σε σχέση με αυτές που αναλύθηκαν στο σχήμα 6.13 για την οικογένεια γονιδίων TNF.



Σχήμα 6.15 Αποτελέσματα της οικογένειας του γονιδίου BRCA1 μετά από την εφαρμογή φιλτραρίσματος και επιλογή σημαντικότερων αποτελεσμάτων

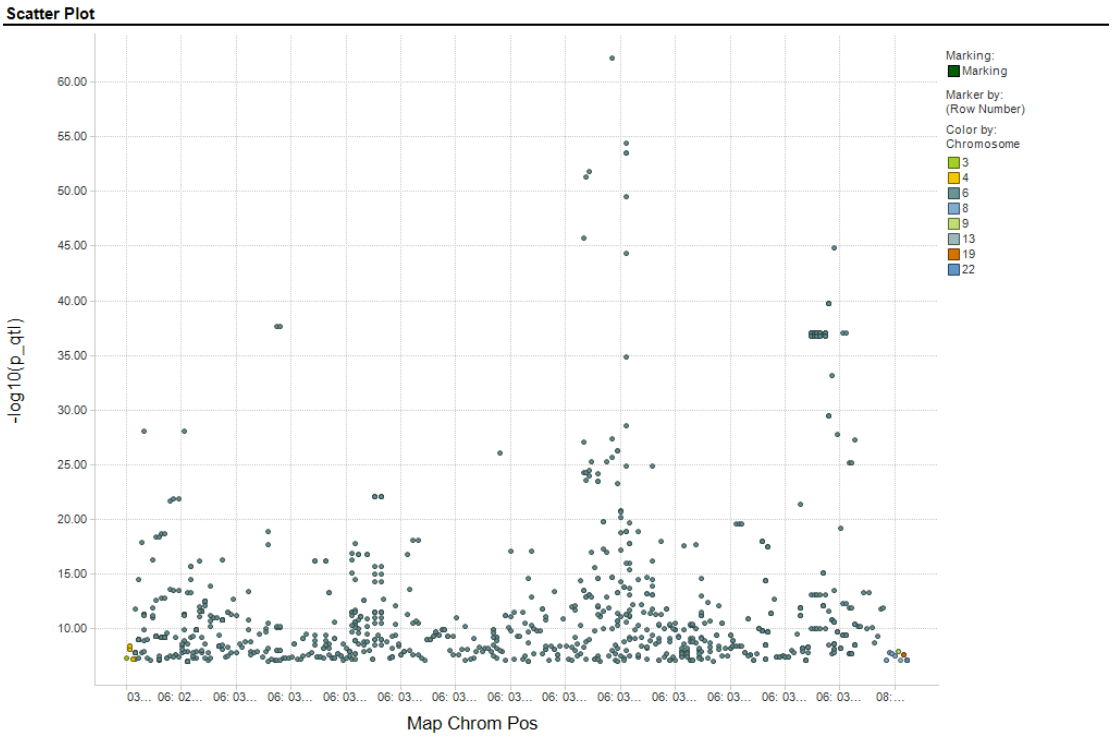
Το σχήμα 6.15 παρουσιάζει ακριβώς τα ιδία αποτελέσματα που σχολιάστηκαν και στο σχήμα 6.14 μόνο που σε αυτή την περίπτωση έχουν επιλεγεί οι μεγαλύτερες από τις

τιμές σε σχέση με τον άξονα των X για παρουσίαση περεταίρω πληροφοριών. Επίσης και για προβολή των υπηρεσιών που μπορεί να μας παρέχει η εφαρμογή Spotfire και τον ρόλο που έπαιξε όσον αφορά την διευκόλυνση στην παρουσίαση και απεικόνιση των αποτελεσμάτων.



Σχήμα 6.16 Προβολή των αποτελεσμάτων του φιλτραρίσματος στο γονίδιο BRCA1 σε σχέση με τα υπόλοιπα αποτελέσματα

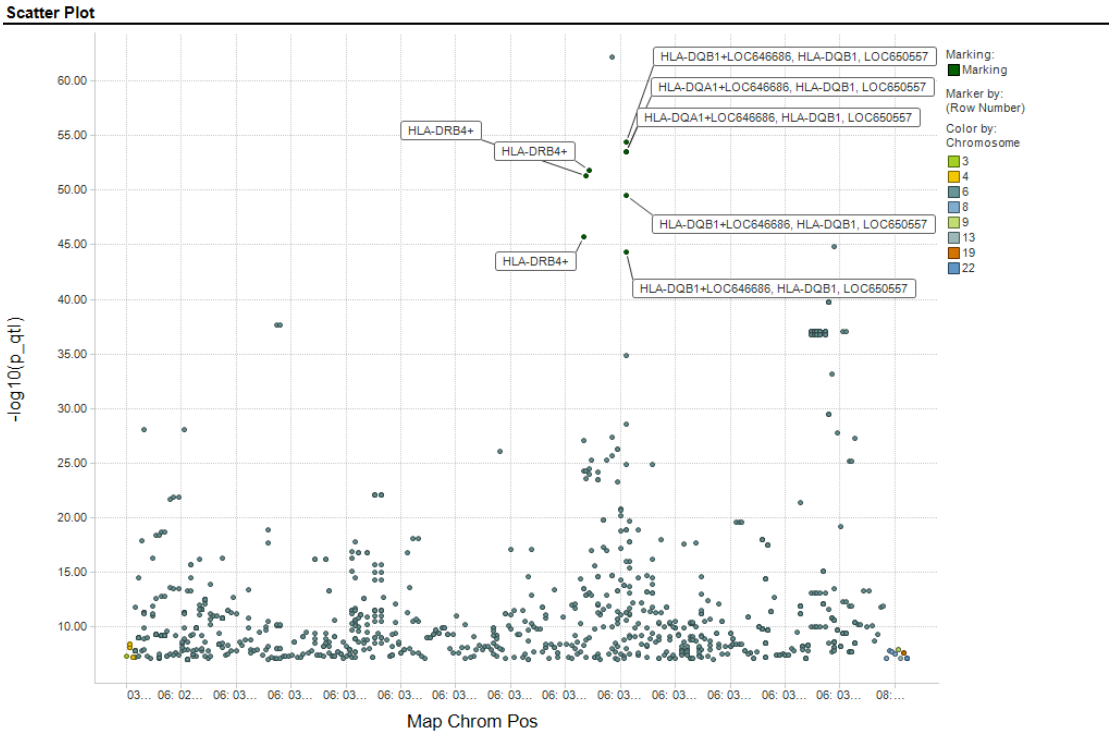
Στο σχήμα 6.16 παρουσιάζονται οι τιμές του γονιδίου BRCA1 που επιλέχθηκαν εφαρμόζοντας το φίλτρο που αναφέρθηκε και στα σχήματα προηγούμενες γραφικές με τη διαφορά ότι σε αυτή την περίπτωση τα αποτελέσματα παρουσιάζονται σε σχέση με όλα τα υπόλοιπα χρωμοσώματα.



Σχήμα 6.17 Αποτελέσματα της οικογένειας του γονιδίου HLA μετά από την εφαρμογή φιλτραρίσματος

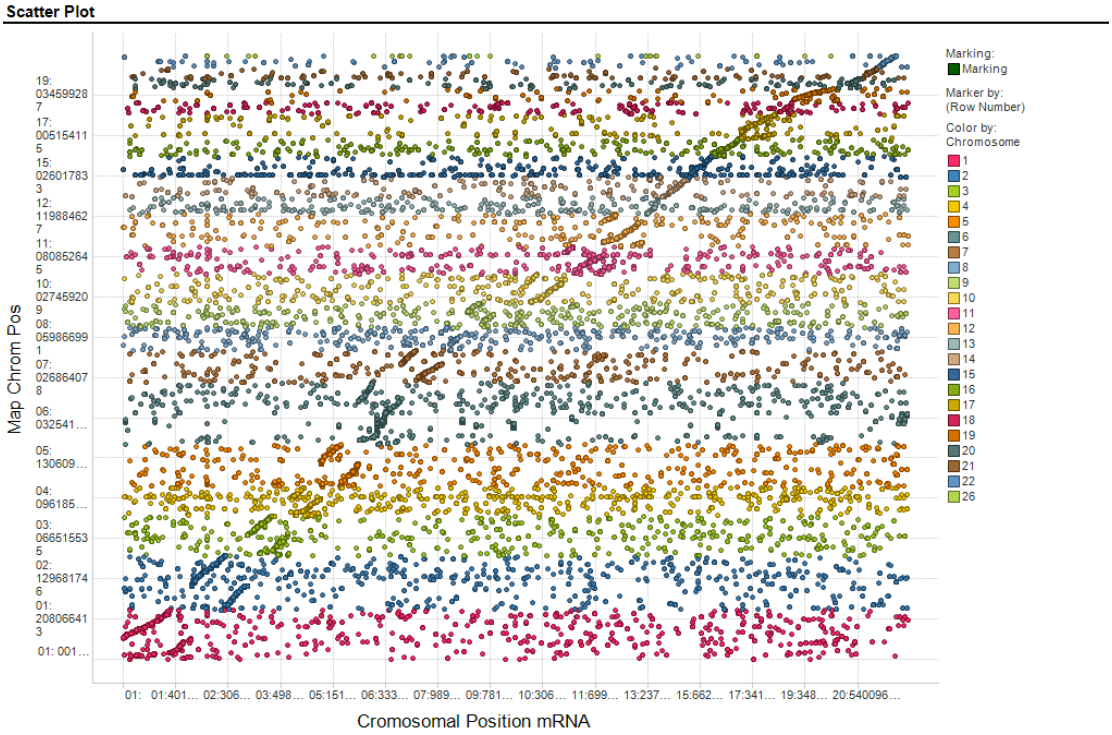
Το σχήμα 6.17 παρουσιάζει τα αποτελέσματα μετά από την εφαρμογή φίλτρου στο γονίδιο HLA . Στον άξονα των X διατάσσονται οι θέσεις των SNPs πάνω στα διάφορα χρωμοσώματα, ενώ στον άξονα των Ψ τις τιμές των πιθανοτήτων σημαντικότητας προσαρμοσμένες βάσει του αρνητικού λογάριθμου με βάση το 10. Στο πλάι αριθμημένα φαίνονται τα διάφορα χρωμοσώματα για τα οποία υπάρχουν αποτελέσματα σε αυτή τη γραφική για το συγκεκριμένο γονίδιο. Επιπλέον όπως αναφέρθηκε και στα προηγούμενα σχήματα, οι σημαντικότερες από τις τιμές που παρουσιάζονται στη γραφική είναι αυτές με την μεγαλύτερη τιμή στον άξονα των Ψ. Δηλαδή οι τιμές με τον μεγαλύτερο λογάριθμο είναι και οι σημαντικότερες βάσει του ορισμού των πιθανοτήτων σημαντικότητας [15] καθώς σημαντικότερες απο αυτές τις τιμές είναι οι μικρότερες, ενώ με τον αρνητικό λογάριθμο σημαντικότερες από αυτές είναι οι μεγαλύτερες. Παρατηρείται πως έστω και αν υπάρχει κάποια τάση συγκέντρωσης των τιμών των πιθανοτήτων σημαντικότητας για κάποιες αυτές γύρω στο 5, οι περισσότερες απο τις τιμές σημαντικότητας είναι μεγαλύτερες από το 5. Αυτό τις καθιστά στατιστικά σημαντικές σύμφωνα με τον ορισμό των πιθανοτήτων σημαντικότητας [15].

Συγκρίνοντας όμως τα αποτελέσματα σε σχέση με αυτά των σχημάτων 6.12 και 6.14 η συγκέντρωση των αποτελεσμάτων είναι μεγαλύτερη απο τις υπόλοιπες περιπτώσεις, και οι τιμές αποτελούν στατιστικά σημαντικές τιμές, επομένως τα SNPs που βρίσκονται σε αυτό το γονίδιο είναι πολύ πιθανόν να επηρεάζουν την έκφραση κάποιου μορίου mRNA του οποίου η περιοχή μεταγραφής είναι γειτονική με αυτή των SNPs στο συγκεκριμένο γονίδιο.



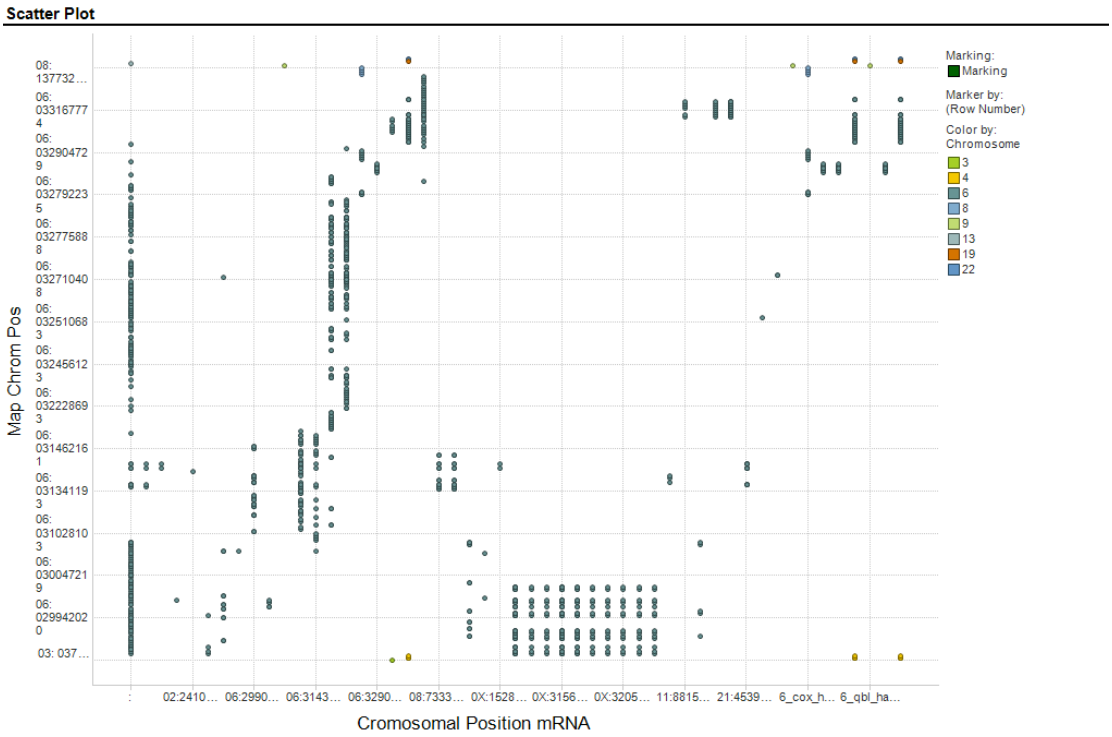
Σχήμα 6.18 Αποτελέσματα της οικογένειας του γονιδίου HLA μετά από την εφαρμογή φιλτραρίσματος και επιλογή σημαντικότερων αποτελεσμάτων

Η γραφική του σχήματος 6.18 παρουσιάζει ακριβώς τα ίδια αποτελέσματα που έχουν σχολιαστεί και στη γραφική παράσταση του σχήματος 6.17. Επιλέγηκαν οι σημαντικότερες από τις τιμές που παρατηρήθηκαν στην γραφική του σχήματος 6.17 για προβολή περεταίρω πληροφοριών που μπορεί να μας παρέχει η εφαρμογή που χρησιμοποιήθηκε για την απεικόνιση και παρουσίαση των αποτελεσμάτων. Επίσης ο μεγάλος αριθμός των τιμών που παρατηρούνται να συγκεντρώνονται στο χρωμόσωμα 6 υποδεικνύουν και τη μεγάλη συγκέντρωση SNPs στο χρωμόσωμα αυτό.



Σχήμα 6.19 Μηδενική απόσταση μεταξύ SNP τοποθεσίας και mRNA μετάγραφου

Η γραφική παράσταση του σχήματος 6.19 παρουσιάζει τις περιπτώσεις cis, δηλαδή αυτές στις οποίες τα SNPs και τα μόρια mRNA βρίσκονται στις ίδιες θέσεις ή σε γειτονικές πάνω στα διάφορα χρωμοσώματα. Ο άξονας των X παρουσιάζει την διάταξη των θέσεων των μορίων του mRNA πάνω σε κάποιο χρωμόσωμα και ο άξονας των Ψ το αντίστοιχο για τα SNPs. Το cis φαίνεται από την συγκέντρωση των τιμών πάνω στην κεντρική διαγώνιο σχηματίζοντας την. Οι τιμές πιθανοτήτων σημαντικότητας για τα SNPs και mRNA σύμφωνα με την πιο πάνω γραφική υποδεικνύουν τη συσχέτιση τους και την κοινή θέση στο χρωμόσωμα που κατέχουν. Αυτό καθορίζει πως τα SNPs που βρίσκονται στην ίδια ή γειτονική θέση με το μόριο του mRNA πάνω στο χρωμόσωμα, δρούν ως στοιχεία ενεργοποίησης (cis) επηρεάζοντας την έκφραση του.



6.20 Κοινές θέσεις SNPs και mRNA σε συγκεκριμένα χρωμοσώματα

Η γραφική παράσταση του σχήματος 6.20 παρουσιάζει ακριβώς ποιά SNPs και ποιά μόρια mRNA κατέχουν κοινές θέσεις και πάνω σε ποια χρωμοσώματα. Όπως μπορεί να παρατηρηθεί οι κοινές θέσεις των SNPs και μορίων mRNA συγκεντρώνονται κυρίως στο χρωμόσωμα 6 του γονιδίου HLA, υποδεικνύοντας καθαρά όλες τις περιοχές Cis πάνω στο γονίδιο. Δηλαδή τις περιοχές όπου τα SNPs βρίσκονται κοντά ή και μέσα στην περιοχή του μορίου του mRNA με αποτέλεσμα να επηρεάζουν και την έκφραση του.

6.3 Γενικά Συμπεράσματα

Τα γενικά συμπεράσματα που μπορούν να προκύψουν από την μελέτη των αποτελεσμάτων των γραφικών παραστάσεων είναι ότι οι παρατηρήσεις και τα αποτελέσματα που διεξάγονται από τη μελέτη των γραφικών είναι τα ίδια και για τις δύο περιπτώσεις εφαρμογής φιλτραρίσματος με διαφορετικό κατώφλι, οδηγώντας στα ίδια συμπεράσματα. Επομένως το περεταίρω φιλτράρισμα μπορεί να φανεί πολύ χρήσιμο για αποδοτική λειτουργία των μετέπειτα εργαλείων που θα χρησιμοποιούνται για την παρουσίαση και προβολή των αποτελεσμάτων.

Κεφάλαιο 7

Συζήτηση

7.1 Γενική Συζήτηση γύρω από το Θέμα

Τα αποτελέσματα που παράχθηκαν από τη συγκεκριμένη μελέτη είναι πάρα πολύ ψηλής σημαντικότητας.

Συνήθως τόσο σε γενετικές μελέτες με τον ίδιο αριθμό SNPs, όσο και σε μελέτες με τον αντίστοιχο αριθμό mRNA που στόχος είναι να μελετηθεί η συσχέτιση με ένα γενετικό χαρακτηριστικό (πχ μια ασθένεια με γενετική προδιάθεση) των mRNA ή SNPs, δεν παρουσιάζονται πολύ σημαντικές τιμές p-value. Εντούτοις σε αυτή την έρευνα τα κορυφαία αποτελέσματα ήταν πέραν τον 10^{60} φορές πιο σημαντικά από ότι τα αναμενόμενα από τις αντίστοιχες άλλες μελέτες [3].

Από πλευράς στατιστικής το πρόβλημα των πολλαπλών ελέγχων μπορεί να ευθύνεται για μια παρατηρημένη αύξηση στο επίπεδο σημαντικότητας, αλλά σύμφωνα με την μέθοδο Bonferroni για την διόρθωση του προβλήματος πολλαπλών ελέγχων, [5] η διαφορά δεν πρέπει να είναι πέραν του 10^4 για την περίπτωση των SNPs και 10^5 για την περίπτωση των μελετών με mRNA δεδομένα. Έχοντας χρησιμοποιήσει και τον έλεγχο σε ένα τυχαίο υποσύνολο των κορυφαίων τιμών της μεθόδου διεργασίας μεταλλαγής, η οποία προσφέρει ακριβή διόρθωση των p-values, είχαμε διαπιστώσει ότι όντος η διαφορά στις τιμές σημαντικότητας, λόγω του προβλήματος πολλαπλών ελέγχων είναι ακόμα μικρότερες και από αυτές που υποδεικνύει η ευρηστική μέθοδος Bonferroni. Επομένως η μόνη άλλη εξήγηση για την τεράστια διαφορά στα επίπεδα σημαντικότητας είναι ότι επειδή όλα τα δεδομένα προέρχονται από μετρήσεις απευθείας από τους βιολογικούς μηχανισμούς, αντί των κλασσικών αναλύσεων όπου η μια μεταβλητή σε

όλους τους ελέγχους είναι η διάγνωση ενός δείγματος για κάποια ασθένεια, είμαστε σε θέση να βρίσκουμε συσχετίσεις όπου σε αυτές εμπεριέχεται πολύ πιο λίγος θόρυβος.

Ένα ακόμα εντυπωσιακό αποτέλεσμα ήταν ο βαθμός με τον οποίο ξεχώριζαν οι συσχετίσεις cis από τις trans στις γραφικές που παρουσιάζονται στα σχήματα 6.3, 6.10, 6.11 και 6.19. Οι διαφορές ειχαν παρατηρηθεί και σε άλλη μελέτη με πολύ παρόμοια δεδομένα [3] αλλά δεν ήταν τόσο αισθητές. Πιστεύεται ότι αυτό είναι ένδειξη της ανώτερης ποιότητας δεδομένων που ήταν διαθέσιμα για αυτή την μελέτη.

Όταν τα αποτελέσματα παρουσιάστηκαν για πρώτη φορά σε συνεδρία με γενετιστές και ιατρούς από διάφορες ειδικότητες, μια από τις σημαντικότερες παρατηρήσεις ήταν ότι ανάμεσα στις κορυφαίες συσχετίσεις υπήρχαν αρκετές που ενέπλεκαν περιοχές γνωστές ως υπεύθυνες για συχνές, και πολύπλοκες ασθένειες. Αυτό αυξάνει τις ελπίδες ότι τα αποτελέσματα από αυτή την έρευνα θα εφαρμοστούν και σε νέα βιολογικά πειράματα.

Επομένως εκτός από τα αποτελέσματα υψίστης σημασίας η μελέτη αυτή διευρύνει τους ορίζοντες για την διεξαγωγή νέων ερευνών γύρω από το θέμα αυτό και περεταίρω μελλοντικές εργασίες.

Επιπλέον σημαντικά είναι και τα αποτελέσματα που επιτεύχθηκαν με τον τρόπο υλοποίησης του κώδικα που χρησιμοποιήθηκε για την ανάλυση και διαχείριση των δεδομένων. Ιδιαίτερα σημαντικό ρόλο έπαιξε στην μείωση των χρόνων εκτέλεσης τόσο η υλοποιίηση κώδικα όσο και η χρήση του grid. Η βελτιστοποίηση στους χρόνους εκτέλεσης ήταν αισθητή καθώς επιτεύχθηκε μείωση των χρόνων εκτέλεσης απο τις 400 εβδομάδες που υπολογίστηκε ο αναμενόμενος απαιτούμενος χρόνος σειριακής επεξεργασίας των δεδομένων σε μόνο 2 εβδομάδες, βάσει του σχεδιασμού και της υλοποίησης του κώδικα που υλοποιήθηκε στην συγκεκριμένη εργασία. Αυτό είχε ώς αποτέλεσμα την βελτιστοποίηση των χρόνων εκτέλεσης κατά 98%.

Από τα αποτελέσματα αυτά είναι αντιλυπτό πως η σημασία των αποτελεσμάτων δεν περιορίζεται μόνο στα αποτελέσματα που παράχθηκαν από την ανάλυση των δεδομένων αλλά επίσης σημαντική και αξιοσημείωτη είναι η αύξηση της απόδοσης του

συστήματος, βάσει του σχεδίου επεξεργασίας των δεδομένων πάνω στο οποίο διεξήχθηκε η όλη διαδικασία.

Επιπλέον ο τρόπος με τον οποίο έγινε η υλοποίηση του κώδικα διευκολύνει την μετέπειτα εξέλιξη του για την εφαρμογή του σε ακόμη πιο πολύπλοκη ανάλυση δεδομένων καθώς επίσης και την φορητότητα του σε οποιοδήποτε λειτουργικό σύστημα καθώς είναι υλοποιημένο σε γλώσσα C++ και μπορεί να χρησιμοποιηθεί σε περιβάλλοντα Unix, Windows, Mac OS X, Linux αρκεί να μεταγλωττιστούν στο κατάλληλο λειτουργικό σύστημα. Επιπλέον η υλοποίηση του διευκολύνει την συστήρηση του που είναι ένας από τους σημαντικότερους παράγοντες για τη ζωή ενός προγράμματος, όπως επίσης και για την επιδιόρθωση διάφορων σφαλμάτων που πιθανόν να προκύψουν κατά την εκτέλεση αλλά και για την τροποποίηση του ώστε να μπορεί να προσαρμοστεί εύκολα και σε μελέτες με παρόμοιου τύπου δεδομένα.

Κλείνοντας , τα αποτελέσματα είναι τα βέλτιστα δυνατά τόσο στον τομέα της βιολογίας όσο και στον τομέα της πληροφορικής καθώς ικανοποιούνται σε μεγάλο βαθμό όλες οι απαιτήσεις από ένα σύστημα αυτού του είδους. Τα μεγάλης σημασίας αποτελέσματα που επιτεύχθηκαν και στους δύο τομείς ανεβάζουν το επίπεδο της μελέτης και αυξάνουν την χρησιμότητα της σε άλλες μελέτες όμοιου τύπου στις οποίες μπορεί να φανεί χρήσιμη σε σχέση με άλλες μελέτες που διεξήχθηκαν γύρω από το ίδιο θέμα.

Κεφάλαιο 8

Συμπεράσματα

8.1 Γενικά Συμπεράσματα

1. Έχει επιτευχθεί η μελέτη μεταξύ γονότυπων και των εκφραζόμενων φαινοτύπων σε ολόκληρο το ανθρώπινο γονιδίωμα

2. Έχει δημιουργηθεί μια βάση δεδομένων με προοπτικές έρευνας για αξιολόγηση της συσχέτισης μεταξύ των SNPs και της έκφρασης του mRNA

3. Η συγκεκριμένη βάση δεδομένων μπορεί να βοηθήσει στην ανακάλυψη νέων φαρμάκων σε όλους τους θεραπευτικούς τομείς. Η ανακάλυψη νέων φαρμάκων μπορεί να γίνει αξιολογώντας τα λειτουργικά αποτελέσματα που μπορούν να έχουν κάποιοι παράγοντες κινδύνου ή κάνοντας εισηγήσεις για γενετικά βασιζόμενους (genetic-based) βιολογικούς δείκτες (biomarkers). Επίσης τονίζοντας τα SNPs που επηρεάζουν την έκφραση γονιδίων που αποτελούν στόχους, μπορούν να ληφθούν υπόψη για μελέτες φαρμακογενετικής (PGx)

4. Όσον αφορά την μεθοδολογία διάσπασης και επεξεργασίας των δεδομένων, ο τρόπος με τον οποίο γράφτηκε ο κώδικας τον καθιστά ικανό να χρησιμοποιηθεί μελλοντικά για επεξεργασία παρόμοιου τύπου δεδομένων μιας μελέτης όπου υπάρχουν για τα ίδια δείγματα, διαθέσιμα δεδομένα DNA (SNPs) και οποιουδήποτε άλλου είδους φαινοτύπου. Επιπλέον η μεθοδολογία μπορεί να είναι εφαρμόσιμη για παρόμοιες μελέτες για δεδομένα μικρότερου ή και

μεγαλύτερου όγκου από αυτά που χρησιμοποιήθηκαν στην συγκεκριμένη μελέτη

5. Η χρήση κώδικα που υλοποιήθηκε σε αυτή τη μελέτη, έπαιξε καθοριστικό ρόλο στην επεξεργασία των δεδομένων καθώς με τη χρήση του, έγινε εφικτή η μείωση του όγκου των ενδιάμεσων αποτελεσμάτων, προσαρμόζοντας τα στα ποσά διαθέσιμης μνήμης που παρείχε το υφιστάμενο σύστημα. Επιπλέον όσον αφορά τη χρήση του grid από πλευράς απόδοσης και ταχύτητας της επεξεργασίας των δεδομένων, υπολογίζοντας τον απαιτούμενο χρόνο για εκτέλεση χωρίς τη χρήση του grid ανέρχεται περίπου γύρω στις 400 περίπου εβδομάδες, δηλαδή 5 χρόνια, ενώ η επεξεργασία εφαρμόζοντας τη δική μας μέθοδο, διήρκεσε μόνο 2 εβδομάδες. Με αυτό τον τρόπο επιτεύχθηκε 200 φορές μεγαλύτερη ταχύτητα στην επεξεργασία των δεδομένων αυξάνοντας έτσι και την απόδοση του συστήματος

6. Η χρήση του κώδικα φιλτραρίσματος για περιορισμό του όγκου των δεδομένων του αρχείου με τα τελικά αποτελέσματα, επιτρέπει την απομόνωση δεδομένων γύρω από μία συγκεκριμένη περιοχή ενδιαφέροντος (region specific), ενός υποσυνόλου mRNA ή ενός υποσυνόλου γονιδίων ή και συνδυασμού των δύο. Για παράδειγμα, μπορεί να δημιουργηθεί μια λίστα με γονίδια, mRNA και πρωτεΐνες που ανήκουν σε ένα pathway, και η εφαρμογή φίλτρου στα αποτελέσματα του mRNA, πρωτεΐνης ή γονιδίου που εμπλάκηκε στο pathway. Αυτό επιτρέπει την προσαρμογή των δεδομένων για διάφορους σκοπούς μελέτης καθιστώντας τα χρήσιμα και για μελλοντική χρήση σε άλλες μελέτες. Επιπλέον αυτό μπορεί να γίνει χωρίς να χαθούν τα αρχικά αποτελέσματα, που προέκυψαν μετά την εφαρμογή κώδικα φιλτραρίσματος κατά τη συγχώνευση

8.2 Μελλοντική Εργασία

Στο πεδίο της συγκεκριμένης μελέτης υπάρχουν ορισμένοι τομείς οι οποίοι μπορούν να προσφερθούν για μελλοντική εργασία.

Μια σημαντική περίπτωση για μελλοντική εργασία πάνω στο συγκεκριμένο θέμα, είναι η επιβεβαιωτική επανάληψη (Replication Testing). Η μέθοδος αυτή μέσω της χρήσης ανάλογων δεδομένων από άλλη μελέτη αποσκοπεί στην εξακρίβωση της εγκυρότητας των αποτελεσμάτων.

Επιπλέον η δημοσιοποίηση των αποτελεσμάτων και διαθεσιμότητα τους στο διαδίκτυο μέσω μίας ιστοσελίδας που θα υποστηρίζει και τις λειτουργίες που προσφέρει το Spotfire, για απεικόνιση και παρουσίαση των αποτελεσμάτων. Αυτό θα διεύρυνε την δυνατότητα χρήσης των αποτελεσμάτων από επιστήμονες που ασχολούνται με βιολογικά ερωτήματα, όπου η γνώση συσχετίσεων μεταξύ γονότυπων και έκφρασης mRNA ή πρωτεϊνών θα τους βοηθούσε. Να σημειωθεί ότι ανάγκη για την συγκεκριμένη γνώση υπάρχει σε παρά πολλά είδη πειραμάτων, και η γνώση μπορεί να χρησιμοποιηθεί ασχέτως της ασθένειας η άλλου γενετικού χαρακτηριστικού που πιθανόν να ενδιαφέρει τους ερευνητές.

Η εφαρμογή των αποτελεσμάτων σε μελέτες φαρμακογενετικής (PGx) δηλαδή μελέτες που αποσκοπούν στην ανακάλυψη νέων φαρμάκων, αξιολογώντας τα λειτουργικά αποτελέσματα που μπορούν να έχουν κάποιοι παράγοντες κινδύνου ή κάνοντας εισηγήσεις για γενετικά βασιζόμενους (genetic-based) βιολογικούς δείκτες (biomarkers), που θα μπορούσαν να χρησιμοποιηθούν σε μελέτες ελέγχου αποδοτικότητας νέων φαρμακευτικών ουσιών. Η βάση δεδομένων είναι υψίστης σημασίας καθώς μπορεί να εφαρμοστεί σε όλους τους θεραπευτικούς τομείς [10].

Όσο αφορά τον κώδικα που υλοποιήθηκε για τη διαχείριση και επεξεργασία των δεδομένων σε όλα τα στάδια της μελέτης, έγινε με τέτοιο τρόπο ώστε να μπορεί να εφαρμοστεί σε παρόμοιου τύπου δεδομένα, ανεξαρτήτως όγκου και περιεχομένων που μπορούν να ακολουθήσουν την ίδια διαδικασία ανάλυσης. Η παραμετροποίηση των

δεδομένων εισόδου καθιστούν τον κώδικα εύκολα εφαρμόσιμο και χρησιμοποιήσιμο σε μελλοντικές μελέτες.

Θα μπορούσε να κωδικοποιηθεί η γνώση που παράχθηκε χάρη σε αυτή την μελέτη, σε κανόνες οι οποίοι χαρακτηρίζουν τις συσχετίσεις αύξησης ή μείωσης της έκφρασης του mRNA ή της πρωτεΐνης ενός γονίδιου, με βάση των γονότυπο ενός SNP. Σε αυτούς τους κανόνες θα μπορούσε να προστεθεί ήδη υπάρχουσα γνώση όσον αφόρα την λειτουργία των γονίδιων στα οποία ανήκει είτε η γενετική περιοχή ενός SNP, ή το mRNA από μια συσχέτιση, ειδικά αν αυτή είναι σε μορφή βιολογικών μονοπατιών (pathway). Ακόμη θα μπορούσαν να προστεθούν και αποτελέσματα από αλλά πειράματα, που να πρόσφεραν γνώση όσον αφορά την συσχέτιση μεταξύ αλληλόμορφων και συγκεκριμένων γενετικών χαρακτηριστικών, όπως πολύπλοκες ασθένειες. Το τελικό σύστημα θα μπορούσε να συμπληρώσει κενά στις γνώσεις μας για τα βιολογικά μονοπάτια, ή ακόμη και να απαντούσε σε ερωτήματα τα οποία χωρίς την χρήση του θα ήταν πολύ πολύπλοκο να απαντηθούν. Για παράδειγμα, πώς θα επηρεαζόταν ο κωδικοποιημένος βιολογικός μηχανισμός αν χορηγείτο στο δείγμα ένα φάρμακο που αύξανε την έκφραση ενός mRNA.

8.3 Επίλογος

Με το πέρας της μελέτης αυτής όσο αφορά τον τομέα της πληροφορικής, επιτεύχθηκε η υλοποίηση κατάλληλα διαμορφωμένου κώδικα που υποστηρίζει την χρήση grid για την επεξεργασία και ανάλυση δεδομένων.

Η χρήση του grid επιτρέπει την παραλληλοποίηση των αλγόριθμων ανάλυσης και επεξεργασίας των δεδομένων καθώς επίσης και την ελαχιστοποίηση των χρόνων εκτέλεσης αλλά και την αυτοματοποίηση των διαδικασιών υποβολής διεργασιών προς εκτέλεσης καθώς υποβάλλονται μια φορά σε κάποια ουρά στο σύστημα με κάποια σειρά προτεραιότητας και το σύστημα είναι αυτό υπεύθυνο στη συνέχεια να υποβάλει μια διεργασία προς εκτέλεση σε περίπτωση που υπάρχουν διαθέσιμοι πόροι.

Όσον αφορά τα δεδομένα δημιουργηθεί μια βάση δεδομένων με προοπτικές έρευνας για αξιολόγηση της συσχέτισης μεταξύ ενός μεγάλου αριθμού SNPs και μετάγραφων mRNA που μπορεί να βοηθήσει στην ανακάλυψη νέων φαρμάκων σε όλους τους θεραπευτικούς τομείς. Η ανακάλυψη νέων φαρμάκων μπορεί να γίνει αξιολογώντας τα λειτουργικά αποτελέσματα που μπορούν να έχουν κάποιοι παράγοντες κινδύνου ή κάνοντας εισηγήσεις για γενετικά βασιζόμενους (genetic-based) βιολογικούς δείκτες (biomarkers).

Γενικά η διεξαγωγή της μελέτης αυτής μπορεί να αποτελέσει πηγή αναφοράς για τη διεξαγωγή παρόμοιων μελετών και τα αποτελέσματα της ως δεδομένα στη φαρμακοβιομηχανία για την παραγωγή νέων φαρμάκων.

Επίσης η γενική προσέγγιση των αλγόριθμων που υλοποιήθηκαν για την διαχείριση και επεξεργασία των δεδομένων, τους καθιστά εύχρηστους και εφαρμόσιμους σε άλλες μελέτες που απαιτούν την επεξεργασία παρόμοιων τύπων δεδομένων όπως αυτών που χρησιμοποιήθηκαν.

Βιβλιογραφία

[1] Abecasis, G.R., Cookson, W.O. & Cardon, L.R, "Selection Strategies for disequilibrium mapping of quantitative traits in nuclear families," Am. J. Hum. Genet., vol. 65, pp. A245, 1999.

[2] Alberts, B., Bray, D., Johnson, A., Lewis, J., Raff, M., Roberts, K., et. al., "Βασικές Αρχές Κυτταρικής Βιολογίας: Εισαγωγή στη Μοριακή Βιολογία του Κυττάρου," 2000

[3] Anna L Dixon et. al., "A genome-wide association study of global gene expression," Nature Genetics, doi:10.1038/ng2109, 2007.

[4] Ashburner, M. et. al., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," Nature Genetics, vol. 25, pp. 25-29, 2000.

[5] Benjamini, Y. & Hochberg, "Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing," J. R. Statist. Soc. Ser. B, vol. 57, pp. 289-300, 1995.

[6] Bolstad, B.M., Irizzary, R.A., Astrand, M. & Speed, T.P., "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," Bioinformatics, vol. 19, pp. 185-193, 2003.

[7] Cheung, V.G et. al., "Natural variation in human gene expression assessed in lymphoblastoid cell," Nature Genetics, vol. 33, pp. 422-425, 2003.

[8] Devlin, B., Roeder, K. and Wasserman, "L. Genomic Control, a new approachto genetic-based association studies," Theor. Popul. Biol., vol. 60, pp. 155-166, 2001.

[9] Enrico Domenici et. al., "Allelic expression in human blood from a depression case/control collection: a database to assess functional impact of SNPs on a genome-scale that enables the identification of novel genetic-driven biomarkers," Scinovation, September 2008.

[10] Harris M.A. et al., "The Gene Ontology (GO) database and informatics resource," Nucleic Acids Res., vol. 32, pp. D258-D261, 2004.

[11] Morley, M et. al., "Genetic analysis of genome-wide variation in human gene expression," Nature, vol. 430, pp. 743-747, 2004.

[12] Scott A. Lesley, "High-Throughput Proteomics: Protein Expression and Purification in the Postgenomic World", Genomics Institute, Novartis Research Foundation, 3115 Merryfield Row, San Diego, California 92121, June 14, 2001.

[13] Shadt, E.E. et. al., "Genetics of gene expression surveyed in maize, mouse and man," Nature, vol. 422, pp. 297-302, 2003.

[14] Yan, H., Yuan, W., Velculescu, V.E., Vogelstein, B. & Kinzler, K.W. "Allelic variation in human gene expression," Science, vol. 297, pp. 1143, 2002.

[15] Ιλία Βόντα, "Εισαγωγή στις πιθανότητες και στατιστική", 2005.

Ατομική Διπλωματική Εργασία

ΚΑΤΑΝΕΜΗΜΕΝΟΣ ΑΛΓΟΡΙΘΜΟΣ ΑΝΑΛΥΣΗΣ ΤΗΣ ΛΕΙΤΟΥΡΓΙΚΗΣ ΕΠΙΔΡΑΣΗΣ ΤΩΝ ΣΗΜΕΙΑΚΩΝ ΝΟΥΚΛΕΟΤΙΔΙΚΩΝ ΠΟΛΥΜΟΡΦΙΣΜΩΝ, ΣΤΑ ΕΠΙΠΕΔΑ ΕΚΦΡΑΣΗΣ ΤΩΝ ΑΛΛΗΛΟΥΧΙΩΝ ΤΟΥ mRNA ΚΑΙ ΤΩΝ ΠΡΩΤΕΪΝΩΝ

Ιωάννα Κάλβαρη

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάιος 2009

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Κατανεμημένος Αλγόριθμος Ανάλυσης της Λειτουργικής Επίδρασης των Σημειακών Νουκλεοτιδικών Πολυμορφισμών στα Επίπεδα Έκφρασης των Αλληλουχιών του mRNA και των Πρωτεϊνών

Ιωάννα Κάλβαρη

Επιβλέπων Καθηγητής
Κωσταντίνος Παττίχης

Η Ατομική Διπλωματική Εργασία υποβλήθηκε προς μερική εκπλήρωση των απαιτήσεων απόκτησης του πτυχίου Πληροφορικής του Τμήματος Πληροφορικής του Πανεπιστημίου Κύπρου

Μάιος 2009

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον κύριο Άθω Αντωνιάδη για την πολύτιμη υποστήριξη και καθοδήγηση που προσέφερε καθ' όλη τη διάρκεια του project. Για την κατατόπιση γύρω από το θέμα και την κατανόηση βιολογικών εννοιών που σχετίζονται με το θέμα και την υποστήριξη του σε θέματα πληροφορικής και την παροχή πρόσβασης στο Grid, που βρισκόταν διαθέσιμο για την ικανοποίηση των αναγκών της μελέτης από την Glaxo Smith Kline (GSK).

Επίσης θα ήθελα να ευχαριστήσω θερμά τον καθηγητή κύριο Παττίχη για τη συνεχή επίβλεψη και υποστήριξη του για την ομαλή διεξαγωγή της μελέτης.

Επιπλέον θα ήθελα να δώσω τις θερμές μου ευχαριστίες στους επιστήμονες από την GSK για την σημαντική συμβολή της στην διεξαγωγή του project, παρέχοντας τα δεδομένα καθώς επίσης και την περιγραφή του προβλήματος.

Ευχαριστίες επίσης για τους κυρίους Enrico Domenichi για τη χρήσιμη καθοδήγηση του γύρω από τον τομέα των proteomics αναφορικά με το mRNA και τις πρωτεΐνες και τον κύριο Pierandrea Muglia για την πολύτιμη προσφορά γύρω από θέματα που σχετίζονται με την γενετική και για την παροχή των δεδομένων DNA.

Περίληψη

Η υφιστάμενη μελέτη έχει ως θέμα της την υλοποίηση ενός κατανεμημένου αλγόριθμου ανάλυσης, της λειτουργικής επίδρασης των σημειακών νουκλεοτιδικών πολυμορφισμών (SNP), στα επίπεδα έκφρασης των αλληλουχιών mRNA και πρωτεϊνών αντίστοιχα.

Έχοντας ως σημείο αφετηρίας ένα αρχείο με συνολικά 550 χιλιάδες δεδομένα για SNPs καθώς επίσης για 56 χιλιάδες δεδομένα έκφρασης mRNA και τέλος ένα με 89 πρωτεΐνες, ακολουθήθηκε μία πορεία διαδικασίας ανάλυσης των δεδομένων σε ένα κατανεμημένο σύστημα.

Η όλη διαδικασία διεξήχθη με την βοήθεια ενός πλέγματος υπολογιστών μεγέθους 200 επεξεργαστών. Για τη χρήση του Grid καθώς επίσης και την διαχείριση των δεδομένων, υλοποιήθηκαν κατάλληλοι αλγόριθμοι για την προσαρμογή τους στο grid προς εκτέλεση. Επίσης υλοποιήθηκε κώδικας για περεταίρω διαχείριση των αποτελεσμάτων που παράχθηκαν κατά την ανάλυση και την μετέπειτα προσαρμογή τους στην εφαρμογή Spotfire, για την αποδοτική απεικόνιση και παρουσίαση των αποτελεσμάτων.

Η ανάλυση των δεδομένων έγινε με την εφαρμογή στατιστικών μεθόδων που προσφέρονταν ήδη από μία εφαρμογή ανοικτού κώδικα το Plink. Συγκεκριμένα εφαρμόστηκαν οι διεργασίες μεταλλαγής και ποσοτική ανάλυση στα δεδομένα για την ανεύρεση τυχόν συσχετίσεων μεταξύ των SNPs και των μετάγραφων mRNA.

Τα δεδομένα όσο και το πλέγμα υπολογιστών καθώς επίσης και η εφαρμογή Spotfire που χρησιμοποιήθηκε σε τελευταίο στάδιο, για την απεικόνιση και την παρουσίαση των αποτελεσμάτων, παρέχονταν για ικανοποίηση των απαιτήσεων της μελέτης, από την φαρμακευτική εταιρεία GSK – GlaxoSmithKline.

Στα κεφάλαια που ακολουθούν αναλύεται εις βάθος η μεθοδολογία που χρησιμοποιήθηκε για την παραλληλοποίηση, οι αλγόριθμοι που εφαρμόστηκαν, καθώς επίσης και οι εφαρμογές που χρησιμοποιήθηκαν για την διεξαγωγή της μελέτης.

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

1.1 Εισαγωγή

Το μεγάλο ενδιαφέρον για την κατανόηση του τεράστιου όγκου πληροφοριών που προέρχονται από την μελέτη του γονιδιώματος των οργανισμών, οδήγησε στην ανάγκη για μελέτη των πρωτεϊνών και πιο συγκεκριμένα την δομή τους και τις λειτουργίες που παρέχουν σε ένα οργανισμό (proteomics: large – scale study of proteins particularly their structures and functions). Η πολυπλοκότητα των πρωτεϊνών καθιστά πολύ δύσκολη την αναγνώριση και κατανόηση της λειτουργίας τους. Η δυσκολία αυτή που παρατηρείται στην αναγνώριση και την κατανόηση της λειτουργίας των πρωτεϊνών και κατά συνέπεια των γονιδίων που τις κωδικοποιούν, συντείνει στην ανάπτυξη νέων τεχνολογιών για συστηματική και περιεκτική ανάλυση της δομής και της λειτουργίας των πρωτεϊνών, που τον τελευταίο καιρό αποτελεί πρόκληση στον τομέα της έρευνας [12].

Το έναυσμα που οδήγησε στην πρόσφατη πρόοδο που παρατηρήθηκε στις βιολογικές επιστήμες, είναι η ολοκλήρωση της αλληλουχίας του ανθρώπινου γονιδιώματος που οδήγησε στην αναγνώριση περίπου 35 χιλιάδων γονιδίων [12].

Το δύσκολο έργο ανάθεσης λειτουργιών σε κάθε ένα από αυτά τα 35 χιλιάδες γονίδια, μόλις που άρχισε να εξελίσσεται και αποτελεί τον πρωταρχικό στόχο της μελέτης της

λειτουργίας του ανθρώπινου γονιδιώματος (human functional genomics). Η λειτουργία ενός γονιδίου, καθορίζεται από το προϊόν της πρωτεΐνης που κωδικοποιεί [12].

Επομένως βάσει των πιο πάνω μπορεί να γίνει αντιληπτό, πως η μελέτη του ανθρώπινου γονιδιώματος, αποτελεί και θέμα μέγιστου ενδιαφέροντος στον τομέα της έρευνας. Αποτέλεσμα αυτής της πρόκλησης που αποτελεί η μελέτη των πρωτεϊνών, οδηγεί στην ανάπτυξη μηχανισμών και ειδικού εξοπλισμού που θα υποβοηθούν και θα επιταχύνουν την έκφραση του μεγάλου αριθμού ανθρώπινων γονιδίων και των προϊόντων που κωδικοποιούν, έτσι ώστε να επιτυγχάνεται μια συστηματική και περιεκτική ανάλυση της δομής και της λειτουργίας των πρωτεϊνών [12].

Η έκφραση των γνωρισμάτων των γονιδίων στα λεμφοκύτταρα είναι κληρονομική και οι γενετικές διαφορές παρατηρήθηκε πως συμβάλλουν στην μεταβλητότητα της έκφρασης των γονιδίων, στα περιφερειακά κύτταρα. Πρόσφατα, έχει γίνει αναφορά για συγκεκριμένη έκφραση των αλληλόμορφων σε ένα ευρύ φάσμα του ανθρώπινου γονιδιώματος, στα κύτταρα και τους εγκεφαλικούς ιστούς. Αυτό έχει ως αποτέλεσμα την μεγάλη επίδραση της γενετικής μεταβλητότητας στους μηχανισμούς μεταγραφής σε διαφορετικούς ιστούς [9].

Η βάση δεδομένων που χρησιμοποιήθηκε προήλθε από μία έρευνα πάνω στην ασθένεια Μονοπολική Κατάθλιψη (Unipolar Depression) με ασθενή και υγιή άτομα. Για την ασθένεια αυτή έγινε σκιαγράφηση της έκφρασης του mRNA, στα δείγματα αίματος των ατόμων που συμμετείχαν στη διαδικασία. Αυτό σε μία προσπάθεια αναγνώρισης βιολογικών δεικτών που σχετίζονται με τη συγκεκριμένη ασθένεια [9].

Σε αυτή την έρευνα αφαιρέθηκε η παράμετρος της ασθένειας της Μονοπολικής Κατάθλιψης έτσι ώστε να μελετηθούν μόνο οι συσχετίσεις μεταξύ της έκφρασης mRNA, πρωτεϊνών και πολυμορφισμών στο γονιδίωμα.

Η ενσωμάτωση των γενετικών δεδομένων με τις πληροφορίες που παράχθηκαν κατά τη σκιαγράφηση της έκφρασης των γονιδίων, έγινε με την διαδικασία σάρωσης των συσχετίσεων σε ολόκληρο το γονιδίωμα, για ανίχνευση των εκφρασμένων γνωρισμάτων από περίπου 56 χιλιάδες probeSets έναντι 550 χιλιάδων SNP δεικτών.

Στο πιο κάτω κείμενο περιγράφεται η διαδικασία επεξεργασίας και ανάλυσης των δεδομένων, καθώς επίσης και τα αποτελέσματα που παράχθηκαν από την πειραματική μελέτη που διεξάγει πάνω σε 190 ανθρώπινα αιματολογικά δείγματα. Τα δείγματα προέρχονται από ένα σύνολο ασθενών με Μονοπολική Κατάθλιψη (Unipolar Depression), 126 σε αριθμό ασθενείς (cases) και 64 φυσιολογικά δείγματα (controls).

Η μελέτη αυτή έχει ως σκοπό να προσδιορίσει όλες τις συσχετίσεις μεταξύ γενετικών πολυμορφισμών σε ολόκληρο το γονιδίωμα και την έκφραση γονιδίων μέσω του επίπεδου έκφρασης του mRNA τους.

Τα δεδομένα που παράχθηκαν χρησιμοποιώντας αυτή την προσέγγιση μπορούν να εφαρμοστούν σε οποιαδήποτε περιοχή ασθενειών ή οποιονδήποτε γενετικών χαρακτηριστικών.

1.2 Παρουσίαση Προβλήματος

Γενετικοί πολυμορφισμοί έχουν βρεθεί να επηρεάζουν την έκφραση γονιδίων σε όλα τα είδη κυττάρων. Έχοντας στη διάθεση μας πειραματικά δεδομένα από 190 δείγματα περιφερικού αίματος στα οποία έγινε ανάλυση έκφρασης mRNA και ορισμένων πρωτεϊνών καθώς επίσης και ανάλυση κάποιων γονοτύπων.

Λόγω του γεγονότος ότι τα 190 δείγματα έχουν συλλεχθεί από ένα σύνολο 126 ασθενών με Μονοπολική Κατάθλιψη (Unipolar Depression) και 64 φυσιολογικά δείγματα (control samples), ήταν αναγκαίο να αφαιρεθεί η παράμετρος της ασθένειας έτσι ώστε όταν γινόταν έλεγχος συσχέτισης μεταξύ έκφρασης mRNA ή πρωτεϊνών και σημειακών νουκλεοτιδικών πολυμορφισμών ώστε να μην μετρούνται τυχών επίπεδα συσχετίσεων που έχουν να κάνουν με την Μονοπολική Κατάθλιψη, αφού τόσο τα SNPs όσο και η έκφραση mRNA και πρωτεϊνών, έχουν ήδη μελετηθεί σε προηγούμενες μελέτες.

Η συλλογή των δεδομένων έκφρασης mRNA έγινε με το affimetrix HU133 plus V2 genechips το οποίο καλύπτει περίπου 56 χιλιάδες περιοχές mRNA. Συλλέχθηκαν επίσης

επίπεδα έκφρασης 80 συγκεκριμένων πρωτεϊνών. Από τα ίδια δείγματα έγινε συλλογή γενετικών δεδομένων με την πλατφόρμα illumina 550K η οποία εξετάζει 550 χιλιάδες SNP από ολόκληρο το γονιδίωμα καλύπτοντας έτσι το μεγαλύτερο μέρος των πιθανών γενετικών παραλλαγών.

Στόχος είναι η ανακάλυψη γενετικών πολυμορφισμών οι οποίοι σχετίζονται με την έκφραση συγκεκριμένων ακολουθιών mRNA ή και πρωτεϊνών.

Η ανάλυση των δεδομένων, λόγω του μεγάλου αριθμού των ελέγχων που εφαρμόστηκαν στα δεδομένα (550 χιλιάδες SNPs και 56 χιλιάδες δεδομένα έκφρασης mRNA), αποτελούσε μια υπολογιστικά ακριβή μέθοδο που εκτός από μεγάλες απαιτήσεις σε χρόνο εκτέλεσης, υπήρχε και μεγάλη ανάγκη σε αποθηκευτικό χώρο. Η μέθοδος επίλυσης για τα προβλήματα που μόλις αναφέρθηκαν έγινε με την εφαρμογή κατανεμημένου υπολογισμού που αναλύεται σε μεγαλύτερο βάθος στην συνέχεια.

Όσον αφορά τα αποτελέσματα που παράχθηκαν κατά την ανάλυση, λόγω του μεγάλου αριθμού των δεδομένων και όλων των δυνατών συσχετίσεων που εφαρμόστηκαν μεταξύ των 550 χιλιάδων SNPs και των 56 χιλιάδων δεδομένων έκφρασης μορίων mRNA, ήταν αναγκαία η χρήση ειδικής εφαρμογής, που επιτρέπει την αποδοτική διαχείριση αποτελεσμάτων μεγάλου όγκου, όπως και αυτών που παράχθηκαν σε αυτή τη μελέτη. Η εφαρμογή που χρησιμοποιήθηκε για την διαχείριση και την γραφική απεικόνιση των αποτελεσμάτων παρουσιάζεται αναλυτικά σε μετέπειτα στάδιο [12].

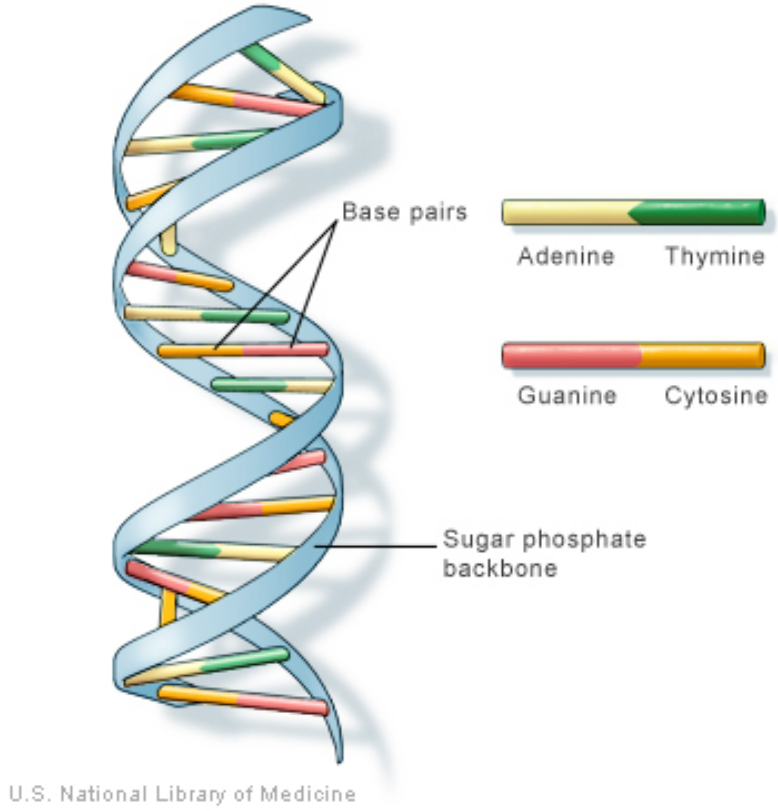
Κεφάλαιο 2

Βασικές Έννοιες Μοριακές Βιολογίας

2.1 Δισόξυριβοζονουκλεϊκό οξύ

DNA δισοξυριβοζονουκλεϊκό οξύ (Deoxyribonucleic acid) περιέχει όλες τις γενετικές πληροφορίες των περισσότερων οργανισμών. Βρίσκεται στον πυρήνα των κυττάρων και ο κύριος ρόλος του είναι η μακροχρόνια αποθήκευση πληροφοριών και οδηγιών για την παραγωγή άλλων συστατικών των κυττάρων, όπως είναι μόρια RNA και πρωτεΐνες. Σύμφωνα με την περιγραφή της δομής του DNA από τους Watson & Crick το 1953, το

DNA από χημικής πλευράς, αποτελεί ένα δίκλωνο μόριο με μορφή έλικας. Κάθε κλώνος του DNA αποτελεί μία πολυνουκλεοτιδική αλυσίδα τα στοιχεία της οποίας αποτελούνται από 4 είδη νουκλεοτιδίων A,T,C και G. Κάθε κλώνος έχει αρχή και τέλος, με την αρχή να συμβολίζεται ως το 5΄ άκρο και το τέλος ως 3΄ άκρο. Ο τρόπος με τον οποίο συνδέονται οι δύο κλώνοι είναι από το άκρο 5΄ προς το άκρο 3΄ με τέτοιο τρόπο ώστε να είναι αντιπαράλληλοι, δηλαδή να είναι ενωμένοι με τέτοιο τρόπο που απέναντι από το 5΄ άκρο του ενός να βρίσκεται το 3΄ άκρο του άλλου κλώνου. Τα νουκλεοτίδια συνδέονται ομοιοπολικά μεταξύ τους με φωσφοδιεστερικούς δεσμούς και με τέτοιο τρόπο προσδίδοντας έτσι χημική πολικότητα στον κάθε κλώνο DNA. Μεταξύ των αζωτούχων βάσεων των νουκλεοτιδίων, που βρίσκονται το ένα απέναντι από το άλλο, δημιουργούνται δεσμοί υδρογόνου που συγκρατούν τους δύο κλώνους ενωμένους. Οι βάσεις βρίσκονται πάντοτε στο εσωτερικό της έλικας ενώ ο σάκχαρο-φωσφορικός σκελετός στο εξωτερικό της έλικας. Αυτή η δομή του DNA με τις βάσεις στο εσωτερικό της έλικας προσδίδουν προστασία στις γενετικές πληροφορίες, αφού φροντίζει ώστε να παραμένουν αναλλοίωτες. Επιπλέον συνδέονται πάντοτε σύμφωνα με τον κανόνα συμπληρωματικότητας των Watson & Crick, με την αδενίνη (A) να ενώνεται πάντοτε με τη θυμίνη (T) με δύο δεσμούς υδρογόνου και την κυτοσίνη (C) να ενώνεται πάντοτε με τη γουανίνη (G) με τρείς δεσμούς υδρογόνου. Η αλληλουχία των νουκλεοτιδίων είναι υπεύθυνη για την κωδικοποίηση και την αποθήκευση της πληροφορίας όπου κάθε μία από τις βάσεις μπορεί να θεωρηθεί ως ένα γράμμα από ένα αλφάβητο τεσσάρων γραμμάτων που χρησιμοποιείται για την αποθήκευση της χημικής πληροφορίας, για το πως εκφράζεται μία πρωτεΐνη. Η διαδικασία με την οποία αναπαράγεται το DNA ονομάζεται αντιγραφή [2].

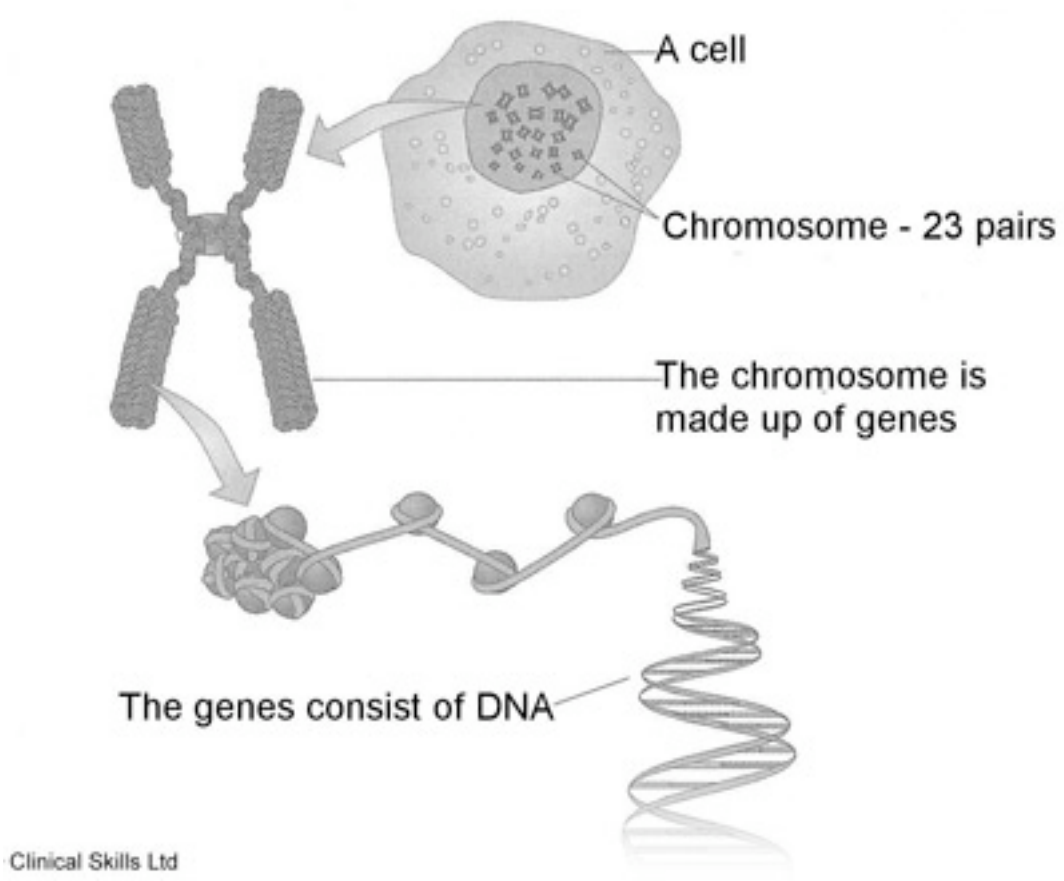


Εικόνα 2.1 Η δομή του DNA παρθέν από 'U.S. National Library of Medicine' (http://ghr.nlm.nih.gov/handbook/illustrations/dnastructure.jpg)

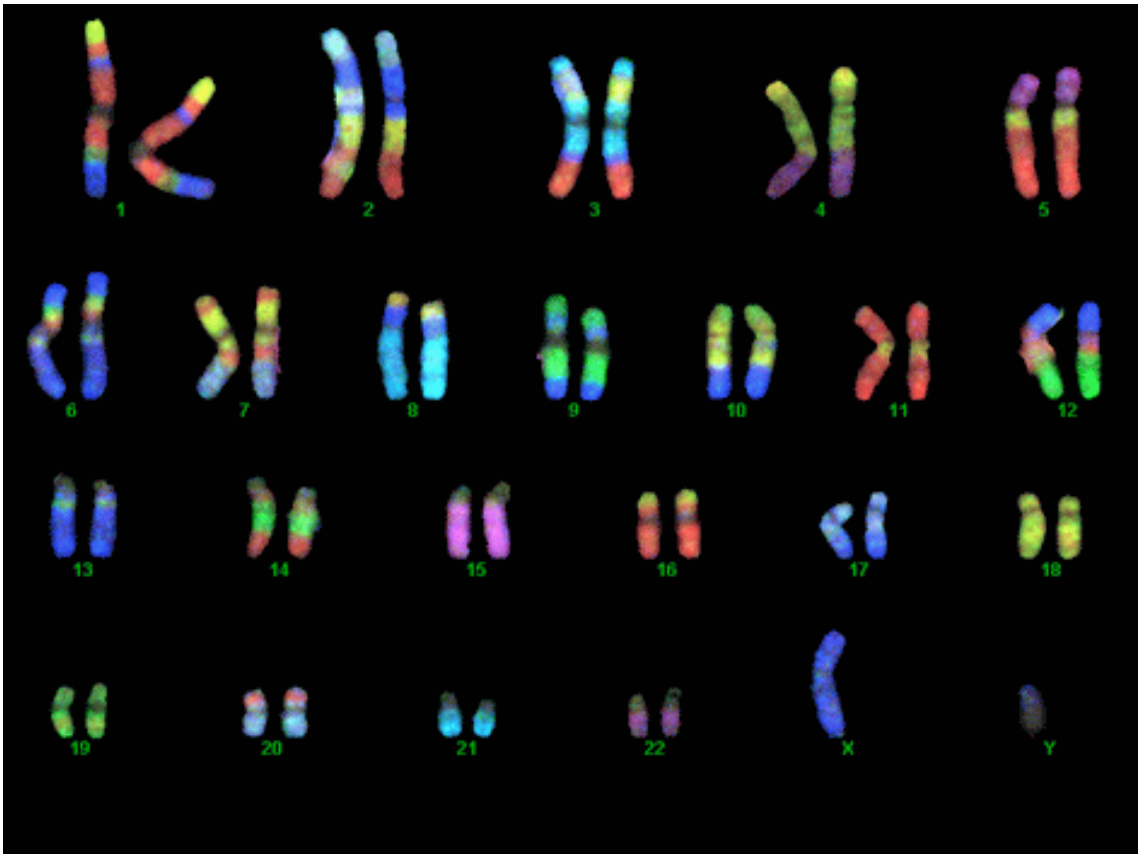
2.2 Χρωμοσώματα

Τα χρωμοσώματα αποτελούν οργανωμένες δομές DNA ή διαφορετικά, δομές που σχηματίζουν πακέτα DNA. Κάθε ένα από αυτά περιέχει ένα πάρα πολύ μακρύ μόριο DNA, που είναι πακεταρισμένο με τέτοιο τρόπο ώστε να έχει 50000 φορές μικρότερο μήκος. Στη διαδικασία δημιουργίας των χρωμοσωμάτων λαμβάνουν μέρος κάποιες ειδικές πρωτεΐνες που συνδέονται μαζί με το μακρομόριο του DNA σχηματίζοντας ένα σύμπλοκο που ονομάζεται χρωματίνη. Το βασικό στοιχείο από το οποίο αποτελείται η χρωματίνη είναι το νουκλεόσωμα, ένα πρωτεϊνικό οκταμερές που αποτελείται από τέσσερις διαφορετικούς τύπους πρωτεϊνών που ονομάζονται ιστόνες. Στον άνθρωπο κάθε σωματικό κύτταρο περιέχει δύο αντίγραφα από κάθε χρωμόσωμα από τα οποία το ένα προέρχεται από τον πατέρα και το άλλο από τη μητέρα. Τα δύο αυτά αντίγραφα σχηματίζουν ζεύγη χρωμοσωμάτων που ονομάζονται ομόλογα λόγω της ίδιας δομής και

οργάνωσης που παρουσιάζουν. Υπάρχουν 23 τέτοια ζεύγη ομόλογων χρωμοσωμάτων στον άνθρωπο, κάθε ένα από τα οποία είναι υπεύθυνο για την έκφραση ενός διαφορετικού χαρακτηριστικού του ανθρώπινου οργανισμού και ένα από αυτά είναι υπεύθυνο για το φύλο του ανθρώπου. Συγκεκριμένα το χρωμόσωμα που είναι υπεύθυνο για το φύλο είναι το χρωμόσωμα 23. Οφείλουμε να αναφέρουμε πως στην περίπτωση του αρσενικού ατόμου το 23^{ο} ζεύγος δεν αποτελεί ομόλογα χρωμοσώματα γιατί το ένα αποτελεί το X και το άλλο το Y. Σε περίπτωση που τα δύο αυτά χρωμοσώματα είναι ομόλογα δηλαδή XX τότε το άτομο αυτό είναι θηλυκό. Διαφορετικά όπως αναφέραμε και προηγούμενος με το χρωμόσωμα 23 να είναι XY τότε το άτομο είναι αρσενικό [2].



Εικόνα 2.2 Το χρωμόσωμα παρθέν από 'EuroGentest' (http://www.eurogentest.org/content/images/unit6/patientLeaflets/english/genesChromosomesDna.jpg)



Εικόνα 2.3 Τα 23 ζεύγη ομόλογων χρωμοσωμάτων παρθέν από 'EMERGENCE' (http://www.homodiscens.com/home/core_content/who_knows/astonishing_predicament/contingency/humilis/dangerous_algorithm/nature_self_aware/karyotype.gif)

2.3 Γονίδια

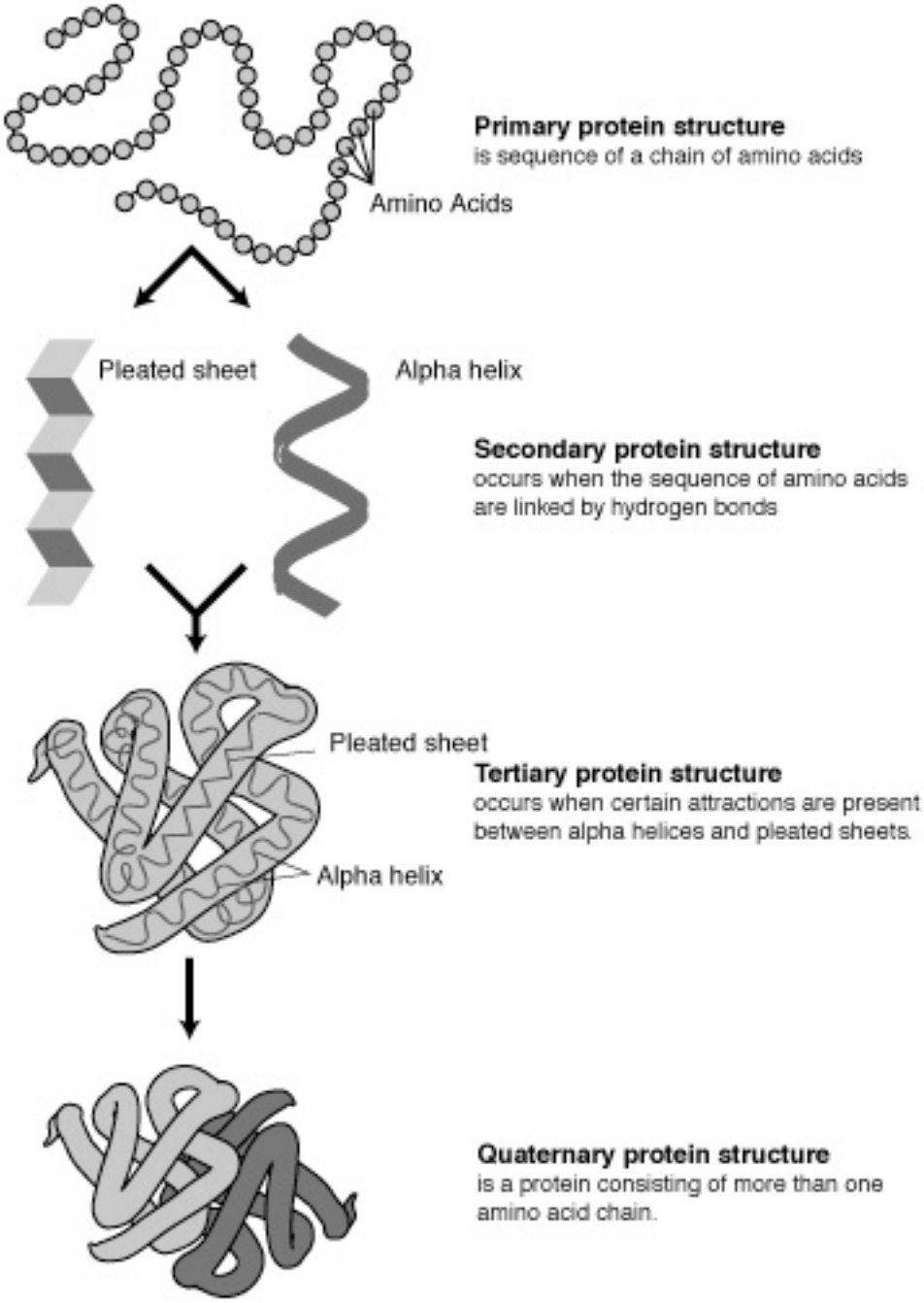
Τα γονίδια αποτελούν τμήματα DNA που μπορούν να μεταφραστούν σε ένα προϊόν πρωτεΐνης. Τα τμήματα αυτά του DNA που δεν μεταφράζονται δηλαδή που δεν παράγουν κάποιο προϊόν ονομάζονται ιντρόνια ενώ τα τμήματα που μεταφράζονται παράγοντας κάποιο προϊόν ονομάζονται εξώνια. Τα γονίδια αποτελούν εξώνια γιατί κάθε μετάφραση γονιδίου παράγει κάποιο προϊόν πρωτεΐνης [2].

Εικόνα 2.4 Το γονίδιο παρθέν από ‘Walgreens’
(http://www.walgreens.com/library/contents.html?docid=000437&doctype=10)

2.4 Πρωτεΐνες

Οι πρωτεΐνες αποτελούν βιολογικά μακρομόρια που παράγονται χρησιμοποιώντας τις πληροφορίες που βρίσκονται αποθηκευμένες στο γενετικό υλικό (DNA). Αποτελούν τα εργαλεία που κατασκευάζει ο οργανισμός για να επιβιώσει καθώς αποτελούν τα κύρια δομικά στοιχεία του οργανισμού. Το ανθρώπινο γενετικό υλικό περιέχει πληροφορίες για να κωδικοποιήσει 20-25 χιλιάδες διαφορετικές πρωτεΐνες. Το κύριο δομικό στοιχείο των πρωτεϊνών είναι τα αμινοξέα που αποτελούν απλά μόρια και υπάρχουν 20 διαφορετικά τέτοια μόρια. Ο αριθμός των μορίων αυτών σε μία πρωτεΐνη καθώς επίσης η σειρά με την οποία εμφανίζονται καθώς επίσης και οι μεταξύ τους αλληλεπιδράσεις, καθορίζουν και τη λειτουργία μίας πρωτεΐνης. Η ένωση των αμινοξέων σε μια πρωτεΐνη σχηματίζει έναν πεπτιδικό δεσμό. Καθώς μία πρωτεΐνη αποτελείται από πολλά

αμινοξέα. Ο αριθμός πεπτιδικών δεσμών σε μία πρωτεΐνη είναι μεγάλος καθώς αυτή μπορεί να αποτελείται από πολλά μόρια, για το λόγο αυτό οι πρωτεΐνες ονομάζονται και πολυπεπτίδια, λόγω του μεγάλου αριθμού πεπτιδικών δεσμών που μπορούν να περιέχουν. Υπάρχουν 4 διαφορετικά επίπεδα στα οποία μπορούμε να χωρίσουμε τις πρωτεΐνες όσον αφορά την οργάνωση τους. Οι κατηγορίες αυτές είναι η πρωτοταγής, δευτεροταγής, τριτοταγής και τεταρτοταγής δομή. Η πρωτοταγής δομή αποτελεί το πρώτο και βασικό επίπεδο στο οποίο έχουμε την αμινοξική αλληλουχία σε μία πρωτεΐνη. Στη συνέχεια η δευτεροταγής δομή των πρωτεϊνών καθορίζει τα τμήματα εκείνα της πολυπεπτιδικής αλυσίδας που σχηματίζουν γνωστά δομικά μοτίβα όπως είναι οι α-έλικες και τα β-πτυχωτά φύλλα. Η τριτοταγής δομή καθορίζει και το τρισδιάστατο σχήμα που έχει μία πρωτεΐνη αν αυτή αποτελείται από μία και μόνο πολυπεπτιδική αλυσίδα, ενώ η τεταρτοταγής δομή καθορίζει το τρισδιάστατο σχήμα μιας πρωτεΐνης που αποτελεί ένα σύμπλοκο, δηλαδή που αποτελείται από περισσότερες από μία πολυπεπτιδικές αλυσίδες. Το τελικό σχήμα της πρωτεΐνης είναι πολύ σημαντικό γιατί αυτό θα καθορίσει και την λειτουργία της πρωτεΐνης. Η λειτουργία παραγωγής των πρωτεϊνών ονομάζεται μετάφραση [2].



Εικόνα 2.5 Η δομή των πρωτεϊνών παρθέν από 'The Matc biotechnology project' (http://matcmadison.edu/biotech/resources/proteins/labManual/images/220_04_114.png)

2.5 Αμινοξέα

Τα αμινοξέα αποτελούν το κύριο δομικό στοιχείο των πρωτεϊνών. Υπάρχουν 20 διαφορετικά τέτοια στοιχεία. Όλα ανεξαρτήτως τα αμινοξέα έχουν την ίδια γενική δομή. Ένα κεντρικό άτομο άνθρακα (α-άνθρακας) που συνδέεται με ένα υδρογόνο, μια αμινομάδα, μια καρβοξυλομάδα και μία πλευρική αλυσίδα. Η αλυσίδα αυτή είναι που διαφοροποιεί τα αμινοξέα αλλάζοντας τις φυσικές και χημικές τους ιδιότητες, αλλάζοντας με αυτό τον τρόπο και τη λειτουργικότητα κάθε αμινοξέως. Βάση του φορτίου της πλευρικής τους αλυσίδας τα αμινοξέα διαχωρίζονται σε 4 διαφορετικές κατηγορίες: όξινα (αρνητικό φορτίο), βασικά (θετικό φορτίο), καθώς επίσης τα πολικά αμινοξέα που δεν έχουν φορτίο, όμως η πλευρική τους αλυσίδα έχει 2 περιοχές με διαφορετικό φορτίο, αλλά και τα μη πολικά αμινοξέα που δεν έχουν φορτίο. Σε φυσιολογικές συνθήκες pH που επικρατούν μέσα στο κύτταρο τα αμινοξέα έχουν την αμινομάδα τους φορτισμένη θετικά και την καρβοξυλομάδα τους φορτισμένη αρνητικά. Το γεγονός αυτό είναι που τα κάνει να συμπεριφέρονται σαν δίπολα. Η συμπεριφορά αυτή ευκολύνει τη σύνδεση των αμινοξέων μεταξύ τους, σχηματίζοντας πεπτιδικούς δεσμούς και τον σχηματισμό των πρωτεϊνών. Η σύνδεση των αμινοξέων γίνεται με τη σύνδεση του καρβοξυτελικού άκρου του αμινοξέως που προηγείται με την αμινομάδα του αμινοξέως που ακολουθεί, σχηματίζοντας με αυτό τον τρόπο έναν πεπτιδικό δεσμό. Κατά την ένωση δύο αμινοξέων και το σχηματισμό ενός πεπτιδικού δεσμού έχουμε ταυτόχρονα και την αποβολή ενός μορίου νερού. Τα άτομα που συμμετέχουν στο σχηματισμό του πεπτιδικού δεσμού έχουν τον περιορισμό ότι πρέπει όλα να βρίσκονται στο ίδιο επίπεδο, περιορίζοντας το εύρος κινήσεων των αμινοξέων που εμφανίζονται σε μία πρωτεΐνη, καθώς δεν μπορούν να περιστραφούν γύρω από τον εαυτό τους. Αφού τα αμινοξέα ενσωματωθούν στην πρωτεΐνη, το μόνο φορτίο που τους μένει είναι αυτό της πλευρικής τους αλυσίδας. Ωστόσο το αμινοτελικό και καρβοξυτελικό άκρο μιας πολυπεπτιδικής αλυσίδας (πρωτεΐνης) παραμένουν θετικά και αρνητικά φορτισμένα αντίστοιχα [2].

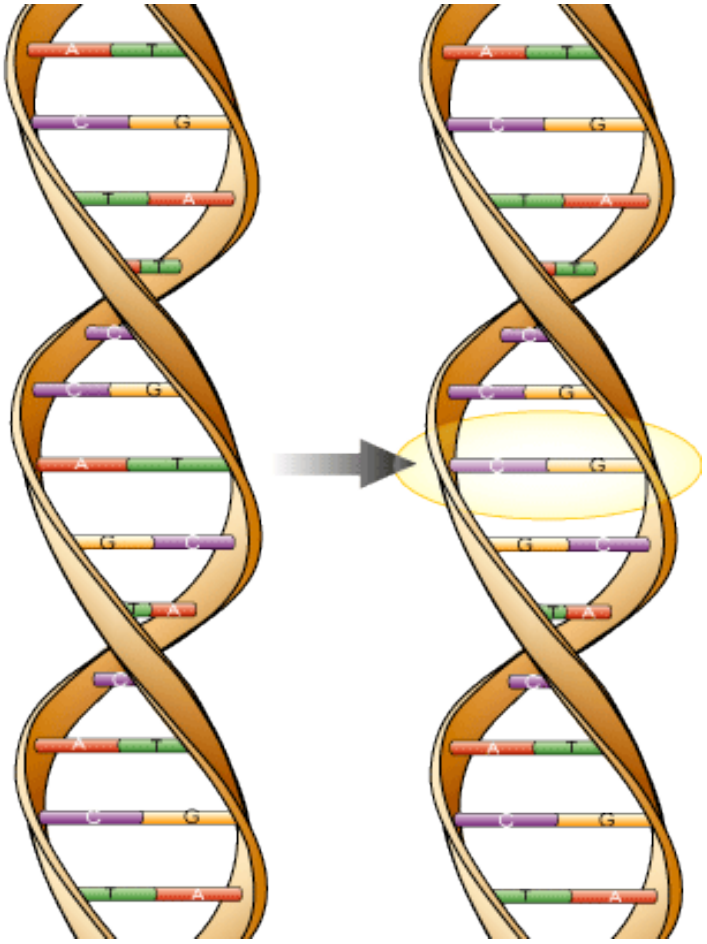
Amino Acid	SLC	DNA codons
Isoleucine	I	ATT, ATC, ATA
Leucine	L	CTT, CTC, CTA, CTG, TTA, TTG
Valine	V	GTT, GTC, GTA, GTG
Phenylalanine	F	TTT, TTC
Methionine	M	ATG
Cysteine	C	TGT, TGC
Alanine	A	GCT, GCC, GCA, GCG
Glycine	G	GGT, GGC, GGA, GGG
Proline	P	CCT, CCC, CCA, CCG
Threonine	T	ACT, ACC, ACA, ACG
Serine	S	TCT, TCC, TCA, TCG, AGT, AGC
Tyrosine	Y	TAT, TAC
Tryptophan	W	TGG
Glutamine	Q	CAA, CAG
Asparagine	N	AAT, AAC
Histidine	H	CAT, CAC
Glutamic acid	E	GAA, GAG
Aspartic acid	D	GAT, GAC
Lysine	K	AAA, AAG
Arginine	R	CGT, CGC, CGA, CGG, AGA, AGG
Stop codons	Stop	TAA, TAG, TGA

Πίνακας 2.1 Τα αμινοξέα και τα κωδικόνια τους

2.6 Σημειακοί Νουκλεοτιδικοί Πολυμορφισμοί (SNP)

Σημειακοί νουκλεοτιδικοί πολυμορφισμοί (single nucleotide polymorphism SNP), είναι παραλλαγές που παρατηρούνται στις ακολουθίες του DNA και οι οποίες εμφανίζονται όταν ένα νουκλεοτίδιο (A,T,C ή G) στην ακολουθία του γονιδιώματος αλλαχθεί. Για παράδειγμα ένα SNP θα μπορούσε να αλλάξει μια ακολουθία DNA από AAGGCTAA σε ATGGCTAA. Τα SNPs καλύπτουν 90% των γενετικών παραλλαγών που παρατηρούνται στον άνθρωπο. Ένας τέτοιος πολυμορφισμός παρατηρείται κάθε 100 μέχρι και 300 βάσεις κατά μήκος των 3 δισεκατομμυρίων βάσεων του ανθρώπινου γονιδιώματος. Στο παράδειγμα που προαναφέρθηκε, τα δύο διαφορετικά νουκλεοτίδια που παρουσιάζονται στην ίδια θέση στις δύο νουκλεοτιδικές ακολουθίες, θεωρούνται

αλληλόμορφα. Τα περισσότερα κοινά SNPs αποτελούνται από 2 μόνο αλληλόμορφα. Οι τέσσερεις διαφορετικές περιπτώσεις SNPs είναι ΑΑ, Αα, αα και η τέταρτη περίπτωση αυτή στην οποία παρατηρούνται missing data δηλαδή έλλειψη δεδομένων, διαφορετικά έλλειψη νουκλεοτιδίων σε κάποιες θέσεις των ακολουθιών DNA. Τα SNPs μπορούν να εμφανιστούν και στα εξώνια αλλά και στα ιντρόνια. Δηλαδή μπορούν να εμφανιστούν σε περιοχές του DNA που κωδικοποιούν κάποιο προϊόν αλλά και σε αυτές τις περιοχές που δεν κωδικοποιούν κάποια πρωτεΐνη ή κάποιο μόριο mRNA. Πολλά από τα SNPs δεν επηρεάζουν την λειτουργία του κυττάρου όμως πιστεύεται πως κάποια άλλα θα μπορούσαν να δημιουργήσουν την προδιάθεση στον άνθρωπο να ασθενήσει ή ακόμη να επηρεάσουν την αντίδραση του οργανισμού τους σε κάποιο φάρμακο.



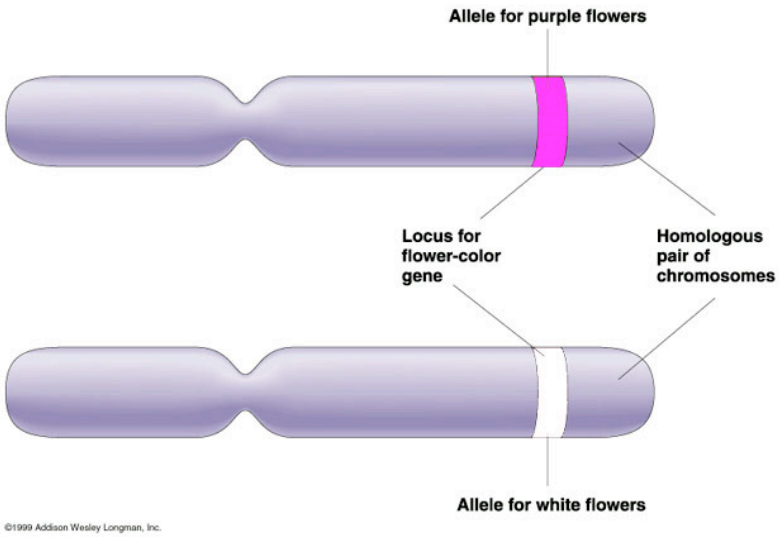
Εικόνα 2.6 Σημεικός Νουκλεοτιδικός Πολυμορφισμός (SNP) παρθέν από 'The Science Creative Quarterly' (http://www.scq.ubc.ca/wp-content/uploads/2006/07/dna1.gif)

2.7 Linkage Disequilibrium (LD)

Ανισορροπία συνδέσμων (linkage disequilibrium - LD) η κατάσταση στην οποία κάποιοι συνδυασμοί αλληλόμορφων γονιδίων ή γενετικών δεικτών παρατηρούνται σε μεγαλύτερο ή λιγότερο συχνά σε ένα πληθυσμό, απ' ότι θα αναμενόταν από τον τυχαίο σχηματισμό των απλότυπων (haplotypes) των αλληλόμορφων γονιδίων, που βασίζεται στη συχνότητα τους. Αποτελούν μη τυχαίες συσχετίσεις μεταξύ πολυμορφισμών που παρατηρούνται σε διαφορετικούς γεωμετρικούς τόπους. Αιτία αυτής της κατάστασης είναι η παρουσία γενετικών συνδέσμων, δηλαδή γενετικές περιοχές στις οποίες το ποσοστό ανασυνδυασμού που μπορεί να προκύψει παρατηρείται να είναι σταθερό σε ένα πληθυσμό και διαφορετικό από το τι θα αναμενόταν αν οι συνδυασμοί ήταν τυχαίοι [1].

2.8 Αλληλόμορφα Γονίδια

Ένα αλληλόμορφο γονίδιο είναι μια συγκεκριμένη ακολουθία νουκλεοτιδίων που μπορεί να έχει ένα γονίδιο από ένα σύνολο ν γνωστών πιθανών ακολουθιών. Σαν παράδειγμα, ας θεωρήσουμε ότι μονό ένα γονίδιο ευθυνόταν για το χρώμα των ματιών. Τότε διαφορετικά αλληλόμορφα αυτού του γονιδίου θα ευθύνονταν για το κάθε πιθανό χρώμα ματιών. Τα αλληλόμορφα γονίδια μπορεί να είναι τόσο σε περιοχές DNA οι οποίες κωδικοποιούν μια ακολουθία mRNA (εξώνια), , αλλά υπάρχουν περιπτώσεις στις οποίες μπορούν να αποτελούν και περιοχές του DNA που να μην κωδικοποιούν μια ακολουθία mRNA (ιντρόνια).



Εικόνα 2.7 Τα αλληλόμορφα γονίδια παρθέν από 'Science of Heredity' (http://porpax.bio.miami.edu/~cmallery/150/mendel/allele.jpg)

2.9 Έκφραση Γονιδίων (Gene Expression)

Είναι η διαδικασία στην οποία μια ακολουθία νουκλεοτιδίων DNA αντιγράφεται και μετατρέπεται σε ένα λειτουργικό γονιδιακό προϊόν όπως μία πρωτεΐνη ή ένα μόριο RNA κατά τις λειτουργίες της μετάφρασης και μεταγραφής αντίστοιχα [13,11,7].

2.10 Έκφραση Πρωτεϊνών (Protein Expression)

Έκφραση πρωτεϊνών αποτελεί ένα τμήμα της διαδικασίας έκφρασης γονιδίων. Περιλαμβάνει τα στάδια στα οποία το DNA έχει ήδη μεταφραστεί σε αμινοξικές αλυσίδες, που στη συνέχεια θα αναδιπλωθούν στο χώρο σχηματίζοντας την δευτεροταγή, τριτοταγή ή τεταρτοταγή δομή και τον τελικό σχηματισμό της πρωτεΐνης.

2.11 Κωδικόνια

Αποτελούν τριάδες βάσεων συγκεκριμένων ακολουθιών που κάθε ένα από αυτά αντιπροσωπεύει κάποιο συγκεκριμένο αμινοξύ. Ένα αμινοξύ είναι δυνατό να αντιπροσωπεύεται από περισσότερα από ένα κωδικόνια, έτσι διαφορετικές ακολουθίες βάσεων του μορίου του mRNA μπορούν να μεταφράζονται στο ίδιο προϊόν πρωτεΐνης.

Υπάρχουν 64 διαφορετικά κωδικόνια από τα οποία 3 αποτελούν κωδικόνια λήξης, που προσδιορίζουν και το τέλος μιας μεταφραζόμενης περιοχής και ένα κωδικόνιο που προσδιορίζει την αρχή μιας μεταφραζόμενης περιοχής (κωδικόνιο έναρξης). Το κωδικόνιο έναρξης είναι το AUG, ενώ τα κωδικόνια λήξης είναι τα UAG, UAA και UGA. Για παράδειγμα το κωδικόνιο CGU αποτελεί ένα από τα τέσσερα διαφορετικά κωδικόνια που κωδικοποιούν το αμινοξύ αργινίνη [2].

GCA GCC GCG GCU	AGA AGG CGA CGC CGG CGU	GAC GAU	AAC AAU	UGC UGU	GAA GAG	CAA CAG	GGA GGC GGG GGU	CAC CAU	AUA AUC AUU	UUA UUG CUA CUC CUG CUU	AAA AAG	AUG	UUC UUU	CCA CCC CCG CCU	AGC AGU UCA UCC UCG UCU	ACA ACC ACG ACU	UGG	UAC UAU	GUA GUC GUG GUU
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V

Εικόνα 2.8 Τα 64 κωδικόνια παρθέν από ‘Center of BioMolecular Modeling - CBM’ (http://www.rpc.msoe.edu/cbm2/images/gfp/gfp3-2.jpg)

2.12 mRNA

Αποτελεί το μόριο RNA που μεταφέρει τις γενετικές πληροφορίες από το DNA στο ριβόσωμα όπου εκεί θα γίνει η μετάφραση του στο προϊόν πρωτεΐνης. Το mRNA κωδικοποιεί τις αντίστοιχες ακολουθίες με νουκλεοτίδια όπως ακριβός και το DNA απλώς κάθε θέση του νουκλεοτιδίου της θιμίνης (T) αντικαθιστάται με την ουρακίλη (U) και αντίθετα με το DNA δεν είναι δίκλωνο αλλά μονόκλωνο [2].

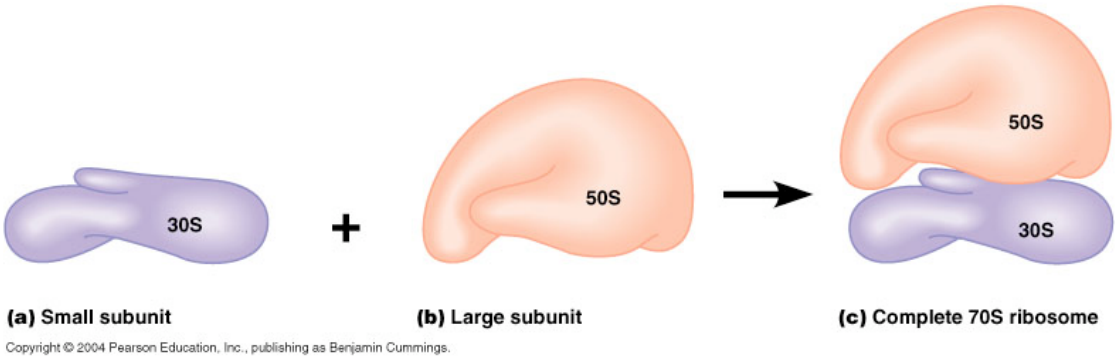
2.13 tRNA

Transfer RNA (tRNA) αποτελεί ένα μικρό μόριο RNA μήκους περίπου 74-95 νουκλεοτίδια, που μεταφέρει κάθε φορά ένα συγκεκριμένο αμινοξύ στο ριβόσωμα, για να ενσωματωθεί στην σχηματιζόμενη πολυπεπτιδική αλυσίδα, κατά την πρωτεϊνοσύνθεση όταν βρίσκεται σε εξέλιξη η διαδικασία της μετάφρασης του RNA. Διαθέτει ένα 3΄άκρο στο οποίο συνδέεται το αμινοξύ. Επίσης περιέχει μια περιοχή μήκους τριών βάσεων που ονομάζεται αντικωδικόνιο, που ζευγαρώνει με την αντίστοιχη περιοχή μήκους τριών βάσεων που βρίσκεται πάνω στο μόριο του mRNA. Κάθε τύπος μορίου tRNA μπορεί να προσδεθεί σε ένα μόνο τύπο αμινοξέως. Λόγω του

γεγονότος ότι ο γενετικός κώδικας περιέχει πολλαπλά κωδικόνια που καθορίζουν το ίδιο αμινοξύ, τα μόρια tRNA με διαφορετικά αντικωδικόνια, που όμως αντιστοιχούν στο ίδιο αμινοξύ, μπορούν να μεταφέρουν το ίδιο αμινοξύ στο οποίο και αντιστοιχούν. Υπάρχουν 31 διαφορετικά tRNAs [2].

2.14 Ριβόσωμα

Αποτελεί το εργοστάσιο σύνθεσης πρωτεϊνών στο κύτταρο. Βρίσκεται στο κυτόπλασμα και αποτελείται από 65% rRNA και 35% πρωτεΐνες. Το οργανίδιο αυτό είναι ένας από τους πιο σύνθετους μοριακούς μηχανισμούς του κυττάρου και αποτελεί ένα μικρό μόνο μέρος του συνολικού δικτύου μηχανισμών που απαιτούνται ώστε να γίνει με επιτυχία η πρωτεϊνοσύνθεση. Αποτελείται από δύο υπομονάδες (τη μικρή και τη μεγάλη υπομονάδα) κάθε μία από τις οποίες είναι ένα τεράστιο σύμπλοκο πρωτεϊνών και RNA. Όταν οι δύο αυτές υπομονάδες είναι ενωμένες τότε σχηματίζουν το πλήρες ριβόσωμα που φέρει τέσσερεις θέσεις σύνδεσης με το RNA. Οι τρείς από τις θέσεις αυτές συνδέονται με tRNAs και η τέταρτη με το mRNA που μεταφράζεται [2].



Εικόνα 2.9 Το ριβόσωμα παρθέν από 'Hunter College of The City University New York' (http://diverge.hunter.cuny.edu/~weigang/Images/04-19_ribosome_1.jpg)

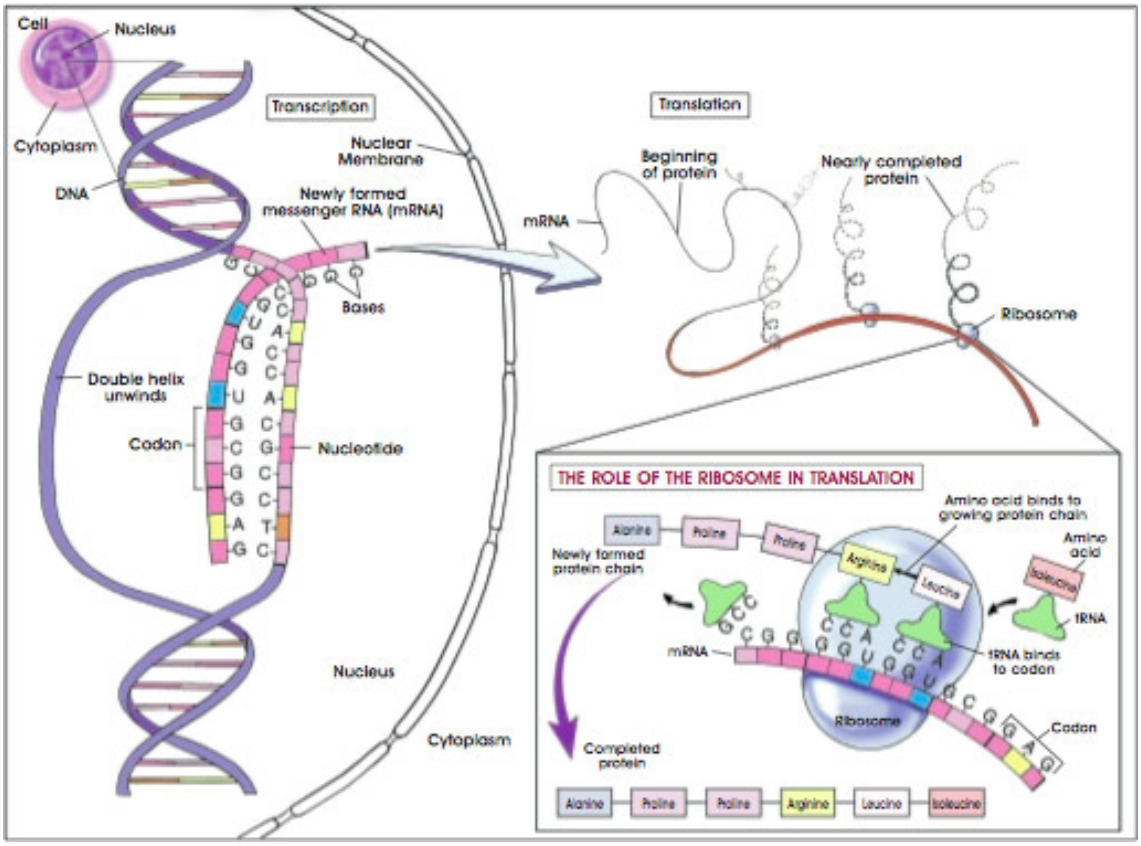
2.15 Μεταγραφή

Μεταγραφή είναι η διαδικασία στην οποία η πληροφορία που περιέχεται στο μόριο του DNA, μεταγράφεται σε ένα μόριο mRNA που αυτό με τη σειρά του θα καθορίσει την ακολουθία αμινοξέων της πρωτεϊνικής δομής. Η διαδικασία αυτή ξεκινά με τον διπλασιασμό του DNA όπου το μόριο του DNA διαχωρίζεται σε δύο ξεχωριστούς

κλώνους των οποίων οι βάσεις είναι συμπληρωματικές. Πιο συγκεκριμένα η διαδικασία αυτή μπορεί να ονομαστεί και σύνθεση RNA καθώς η νουκλεοτιδική ακολουθία DNA μεταγράφεται – τροποποιείται σε RNA πληροφορία. Και των δύο μορίων οι νουκλεοτιδικές ακολουθίες χρησιμοποιούν συμπληρωματική γλώσσα, έτσι αυτός είναι και ο λόγος που η πληροφορία απλά μεταγράφεται ή πιο εύκολα αντιγράφεται από το ένα μόριο στο άλλο. Η διαδικασία μεταγραφής του DNA είναι παρόμοια όπως και η αντιγραφή του με τη διαφορά ότι σε αυτή την περίπτωση συμμετέχουν διαφορετικά ένζυμα και το ταίριασμα των βάσεων αδενίνης που βρίσκονται στο DNA με ουρακίλη U που είναι μία από τις βασικές διαφορές του μορίου του DNA από το μόριο του RNA. Έτσι το μόριο mRNA που παράγεται είναι το ίδιο με τον συμπληρωματικό κλώνο DNA με τη μόνη διαφορά ότι όπου στο DNA υπήρχαν θυμίνες T στο μόριο του mRNA θα υπάρχου ουρακίλες U. Το mRNA μεταγράφεται από την RNA πολυμεράση, ένα ένζυμο που προσδένεται στο ένα μονόκλωνο DNA, για να δημιουργήσει με αυτόν τον τρόπο το συμπληρωματικό κλώνο RNA που ονομάζεται messenger RNA – mRNA. Η ονομασία του προέρχεται από το γεγονός ότι μεταφέρει ένα γενετικό μήνυμα ή διαφορετικά τη γενετική πληροφορία από το DNA στο μηχανισμό πρωτεϊνοσύνθεσης του κυττάρου, το ριβόσωμα. Η διαδικασία αυτή είναι το πρώτο στάδιο που οδηγεί στην έκφραση γονιδίων (gene expression) με την παραγωγή του ενδιάμεσου μορίου mRNA, που αποτελεί ένα πιστό μετάγραφο της πληροφορίας του γονιδίου που κωδικοποιεί τη σύνθεση κάποιας πρωτεΐνης. Το συγκεκριμένο τμήμα DNA που μεταγράφεται σε mRNA ονομάζεται μετάγραφο. Αυτό το συγκεκριμένο τμήμα, το μετάγραφο, περιέχει αλληλουχίες νουκλεοτιδικών βάσεων που όχι μόνο κωδικοποιούν την αλληλουχία που μεταφράζεται αλλά επιπλέον κατευθύνει και ρυθμίζει την πρωτεΐνοσύνθεση. Αυτό γίνεται καθώς ο αριθμός πρωτεϊνών που θα παραχθούν εξαρτάται από τον αριθμό των μορίων mRNA που έχει στη διάθεση του το κύτταρο για μετάφραση. Η μεταγραφή γίνεται χρησιμοποιώντας τον 3΄- 5΄ κλώνο του DNA έτσι ώστε το μόριο mRNA που θα πάρουμε να έχει κατεύθυνση 5΄- 3΄ για να μπορέσει να χρησιμοποιηθεί στη συνέχεια στη διαδικασία της μετάφρασης, στην οποία θα γίνει και η πρωτεϊνοσύνθεση [2].

2.16 Μετάφραση

Μετάφραση είναι η διαδικασία κατά την οποία ένα μόριο ή περισσότερα μόρια mRNA μεταφράζονται σε ένα η περισσότερα μόρια πρωτεΐνης, ανάλογα με την ακολουθία των βάσεων που μεταφέρει το μόριο αυτό. Κάθε τρείς από τις βάσεις του μορίου του mRNA αντιπροσωπεύουν ένα αμινοξύ. Τα περισσότερα από αυτά εκφράζονται από περισσότερα από ένα κωδικόνια. Αυτός είναι και ο κύριος λόγος για τον οποίο διαφορετικές ακολουθίες βάσεων μπορούν οδηγήσουν στην παραγωγή της ίδιας πρωτεΐνης. Σε αυτή τη διαδικασία λαμβάνουν μέρος δύο διαφορετικά είδη RNA, το tRNA που μεταφέρει το αμινοξύ στην παραγόμενη πολυπεπτιδική αλυσίδα, και το mRNA που μεταφέρει τις πληροφορίες που θα μεταφραστούν ώστε να παραχθεί το νέο προϊόν. Η διαδικασία εκτελείται στην περιοχή του κυτοπλάσματος όπου και βρίσκονται τα ριβοσώματα. Τα ριβοσώματα αποτελούνται από τη μεγάλη και τη μικρή υπομονάδα και στο τέλος όταν θα αρχίσει η διαδικασία, ολόκληρο το ριβόσωμα πλαισιώνει το μόριο του mRNA. Η διαδικασία ξεκινά με το εναρκτήριο tRNA που φέρει τη μεθειονίνη που συνδέεται με τη μικρή ριβοσωμική υπομονάδα και έπειτα το σύμπλοκο αυτό, ενώνεται στο 5΄άκρο του mRNA που στην συνέχεια θα ενωθεί με μία σειρά από πρωτεΐνες που ονομάζονται παράγοντες έναρξης που βοηθούν στην έναρξη της διαδικασίας. Στη συνέχεια όλο το σύμπλοκο αρχίζει να κινείται προς το 3΄άκρο του mRNA, μέχρι να συναντήσει το πρώτο AUG που αποτελεί το κωδικόνιο έναρξης, οπότε αποδεσμεύονται οι παράγοντες έναρξης και συγχρόνως ενώνεται η μικρή ριβοσωμική υπομονάδα με τη μεγάλη πλαισιώνοντας έτσι το μόριο του mRNA. Ακολούθως όλο το σύμπλοκο – ριβόσωμα κινείται προς το 3΄άκρο του mRNA, διαβάζοντας τα κωδικόνια και προσθέτοντας τα κατάλληλα αμινοξέα στην συνεχώς αυξανόμενη πολυπεπτιδική αλυσίδα, μέχρι να συναντήσει τα κωδικόνια λήξης. Υπάρχουν τρία διαφορετικά κωδικόνια που υποδεικνύουν τη λήξη μιας περιοχής που κωδικοποιεί ένα προϊόν και αυτά είναι τα UAG, UAA και UGA. Τα κωδικόνια λήξης δεν αναγνωρίζονται από κάποιο tRNA αλλά ειδοποιούν το ριβόσωμα ότι πρέπει να σταματήσει την πρωτεϊνοσύνθεση και να διασπαστεί. Η πεπτιδυλοτρανσφεράση είναι το ένζυμο που αποσυνδέει τα αμινοξέα από το tRNA και τα συνδέει στο καρβοξυτελικό άκρο της νεοσυντιθέμενης πολυπεπτιδικής αλυσίδας. Όταν η πρωτεϊνοσύνθεση φτάσει στο τέλος, η πρωτεΐνη ελευθερώνεται στο κυτόπλασμα ώστε να χρησιμοποιηθεί από το κύτταρο [2].



Εικόνα 2.10 Οι λειτουργίες της Μεταγραφής και της Μετάφρασης παρθέν από 'National Institutes of Health – Stem Cell Information (http://stemcells.nih.gov/StaticResources/info/scireport/images/figurea6.jpg)

2.17 Cis – Ενεργοποιητικά στοιχεία (acting elements)

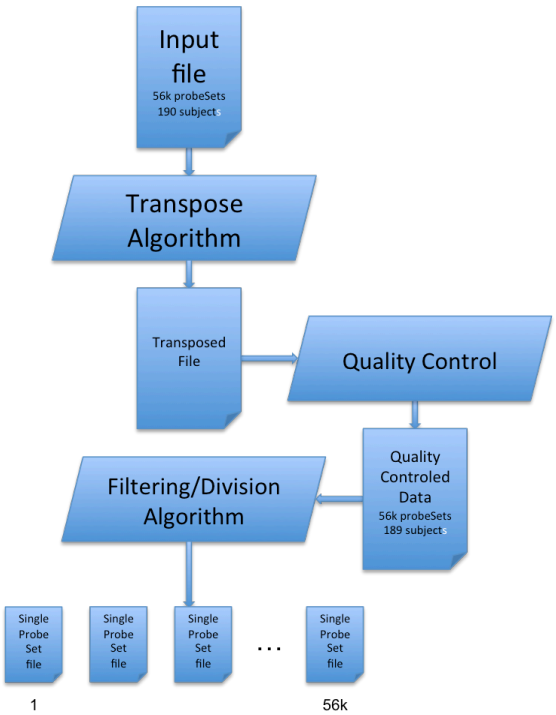
Τα στοιχεία ενεργοποιητές, αποτελούν συνήθως ακολουθίες DNA που εμφανίζονται στο δομικό μέρος ενός γονιδίου και είναι απαραίτητα για την έκφραση του γονιδίου αυτού. Πάνω σε αυτές τις ακολουθίες προσδένονται οι ενεργοποιητικοί παράγοντες για να υποβοηθήσουν την λειτουργία της έκφρασης ενός γονιδίου.

2.18 Trans – Ενεργοποιητικοί παράγοντες (acting factors)

Οι ενεργοποιητικοί παράγοντες συνήθως αποτελούν πρωτεΐνες που προσδένονται πάνω στις cis ενεργοποιητικές ακολουθίες DNA και υποβοηθούν και ελέγχουν την έκφραση κάποιου γονιδίου.

Κεφάλαιο 3

Προεπεξεργασία [6]



Διάγραμμα 3.1 Προεπεξεργασία

3.1 Δεδομένα Έκφρασης mRNA

Τα δεδομένα έκφρασης mRNA παράχθηκαν για το σύνολο των 190 δειγμάτων εκ των οποίων τα 64 ήταν φυσιολογικά δείγματα ενώ τα υπόλοιπα 126 δείγματα προέρχονταν από ασθενείς με Μονοπολική Κατάθλιψη. Η μέθοδος που χρησιμοποιήθηκε για την παραγωγή των δεδομένων έκφρασης mRNA είναι η μέθοδος Affymetrix HU133 v.2 GeneChips. Σε προηγούμενη έρευνα, probeSets που παρουσίασαν σημαντικές διαφορές στην έκφραση μεταξύ των δειγμάτων που προέρχονταν από υγιή άτομα και αυτών που προέρχονταν από ασθενείς, όσο αφορά τα στατιστικά στοιχεία που μελετήθηκαν, έγινε ανάλυση των διαφορών που βρέθηκαν (analysis of variance).

3.2 Δεδομένα Έκφρασης Πρωτεϊνών

Η πηγή που χρησιμοποιήθηκε για τη συλλογή των δεδομένων έκφρασης των πρωτεϊνών είναι η ιστοσελίδα του Rules Based Medicine στο διαδίκτυο που αποτελεί ένα εργαστήριο βιολογικών δεικτών με μία συλλογή από αναπαραγωγίσιμα, ποσοτικά δεδομένα. Συγκεκριμένα τα δεδομένα συλλέχτηκαν από τη βάση Human Map 1.5 που αποτελείται από ένα σύνολο 89 πρωτεϊνικών εκφράσεων, για τις οποίες διατίθεται το όνομα, η περιγραφή του κάθε διαφορετικού δείγματος καθώς επίσης και ένας μοναδικός αριθμός που καθορίζει την κάθε μία από αυτές. Η λίστα δειγμάτων των πρωτεϊνικών εκφράσεων παρουσιάζεται πιο κάτω:

Α/Α	Πρωτεϊνη	Α/Α	Πρωτεϊνη
1	Adiponectin	46	Interleukin-3
2	Alpha-1 Antitrypsin	47	Interleukin-4
3	Alpha-Fetoprotein	48	Interleukin-5
4	Alpha-2 Macroglobulin	49	Interleukin-6
5	Apolipoprotein A-1	50	Interleukin-7
6	Apolipoprotein C-III	51	Interleukin-8
7	Apolipoprotein H	52	Interleukin-10
8	Beta-2 Microglobulin	53	Interleukin-12 p40
9	BDNF	54	Interleukin-12 p70
10	C-Reactive Protein	55	Interleukin-13
11	Calcitonin	56	Interleukin-15
12	Cancer Antigen 19-9	57	Interleukin-16
13	Cancer Antigen 125	58	Leptin
14	Carcinoembryonic Antigen	59	Lipoprotein (a)
15	CD40	60	Lymphotactin
16	CD40 Ligand	61	MDC
17	Complement 3	62	MIP-1 alpha
18	CK-MB	63	MIP-1 beta
19	Endothelin-1	64	MMP-2
20	Eotaxin	65	MMP-3
21	Epidermal Growth Factor	66	MMP-9
22	ENA-78	67	MCP-1
23	Erythropoietin	68	Myeloperoxidase
24	ENRAGE	69	Myoglobin
25	Factor VII	70	PAI-1
26	Fatty Acid Binding Protein	71	PAPP-A
27	Ferritin	72	PSA, Free
28	Fibrinogen	73	Prostatic Acid Phosphatase
29	FGF-basic	74	RANTES
30	GST	75	Serum Amyloid P
31	G-CSF	76	SGOT
32	GM-CSF	77	Sex Hormone Binding Globulin
33	Growth Hormone	78	Stem Cell Factor
34	Haptoglobin	79	Thrombopoietin
35	Immunoglobulin A	80	Thyroid Binding Globulin
36	Immunoglobulin E	81	Thyroid Stimulating Hormone
37	Immunoglobulin M	82	Tissue Factor
38	Insulin	83	TIMP-1
39	IGF-1	84	Tumor Necrosis Factor-alpha
40	ICAM-1	85	Tumor Necrosis Factor-beta
41	Interferon-gamma	86	Tumor Necrosis Factor RII
42	Interleukin-1 alpha	87	VCAM-1
43	Interleukin-1 beta	88	VEGF
44	Interleukin-1 ra	89	von Willebrand Factor
45	Interleukin-2		

Πίνακας 3.1 Οι 89 Πρωτεΐνες

3.3 Δεδομένα Έκφρασης DNA

Οι γονότυποι (δεδομένα έκφρασης DNA) συλλέχτηκαν με την μέθοδο του illumina 550K. Η μέθοδος αυτή χρησιμοποιήθηκε έτσι ώστε να είναι εφικτή η συλλογή δεδομένων από όλο το γονιδίωμα. Εφαρμόστηκε πάνω σε 2000 υποψηφίους, 1000 ασθενείς (cases) και 1000 υγιείς (controls), εκ των οποίων και τα 190 δείγματα που αναλύθηκαν σε mRNA μεταφράζονται σε κάποιο πρωτεϊνικό προϊόν.

3.3.1 Προεπεξεργασία Δεδομένων Έκφρασης DNA

Στο συγκεκριμένο σημείο χρησιμοποιήθηκαν συγκεκριμένα περίπου 550 χιλιάδες SNPs που όπως ήδη αναφέρθηκε αρκετές φορές, προήλθαν από 190 δείγματα περιφερικού αίματος, τα οποία συλλέχτηκαν από μία βάση που περιέχει 126 ασθενείς με κατάθλιψη (Unipolar Depression) και 64 φυσιολογικά δείγματα (control samples). Από τα δείγματα αυτά τα 43 προέρχονται από άντρες ενώ τα υπόλοιπα 133 από γυναίκες.

Έγινε έλεγχος σε όλα τα δείγματα έτσι ώστε να διερευνηθεί, αν μεγαλύτερο ποσοστό από το 10% των δειγμάτων αυτών παρουσίαζαν έλλειψη δεδομένων για να αφαιρεθούν. Κανένα από τα SNPs δεν παρουσίαζε κάποια έλλειψη όσον αφορά τα δεδομένα. Επιπλέον έγινε έλεγχος της συχνότητας των αλληλόμορφων γονιδίων ώστε να είναι το μέγιστο 1%, διαφορετικά αυτά να αφαιρούνταν από το σύνολο. Κατά τη διαδικασία αυτή παρουσιάστηκαν περίπου 10 χιλιάδες SNPs που απέτυχαν τον έλεγχο αυτό, με αποτέλεσμα να αφαιρεθούν [9,14].

Ως αποτέλεσμα των πιο πάνω ελέγχων ήταν ο περιορισμός των δεδομένων από τα 550 χιλιάδες SNPs που ήταν διαθέσιμα αρχικά, σε έναν μικρότερο αριθμό αυτών [9].

3.4 Δεδομένα Εισόδου

Όπως έχει ήδη προαναφερθεί, τα δεδομένα προήλθαν από την ανάλυση αιματολογικών δειγμάτων, 190 ανθρώπων από τα οποία τα 124 προέρχονταν από ασθενείς με Μονοπολική Κατάθλιψη (Unipolar Depression) οι cases ενώ τα υπόλοιπα 64 αποτελούσαν υγιή δείγματα (controls).

Συγκεκριμένα τα 190 δείγματα χρησιμοποιήθηκαν για να ληφθούν δεδομένα για περίπου 56 χιλιάδες διαφορετικά probeSets. Τα δεδομένα για κάθε ένα probeSet και δείγμα, αποτελούσαν μία κανονικοποιημένη τιμή η οποία ήταν από το 0 μέχρι και το ∞.

Το μεγαλύτερο βάρος δίνεται στις μικρότερες τιμές καθώς μια τιμή p-value αντιστοιχεί στην πιθανότητα το αποτέλεσμα να είναι false positive, δηλαδή να έχει τη μεγαλύτερη πιθανότητα ώστε να είναι λανθασμένο. Όσο μικρότερη η τιμή/πιθανότητα αυτή τόσο και πιο μεγάλη η σημασία του δείγματος στα δεδομένα. Αργότερα δίνεται και ο ακριβής ορισμός της πιθανότητας σημαντικότητας (p-value) όπως ορίζεται στη στατιστική.

3.5 Δομή Αρχείου Δεδομένων

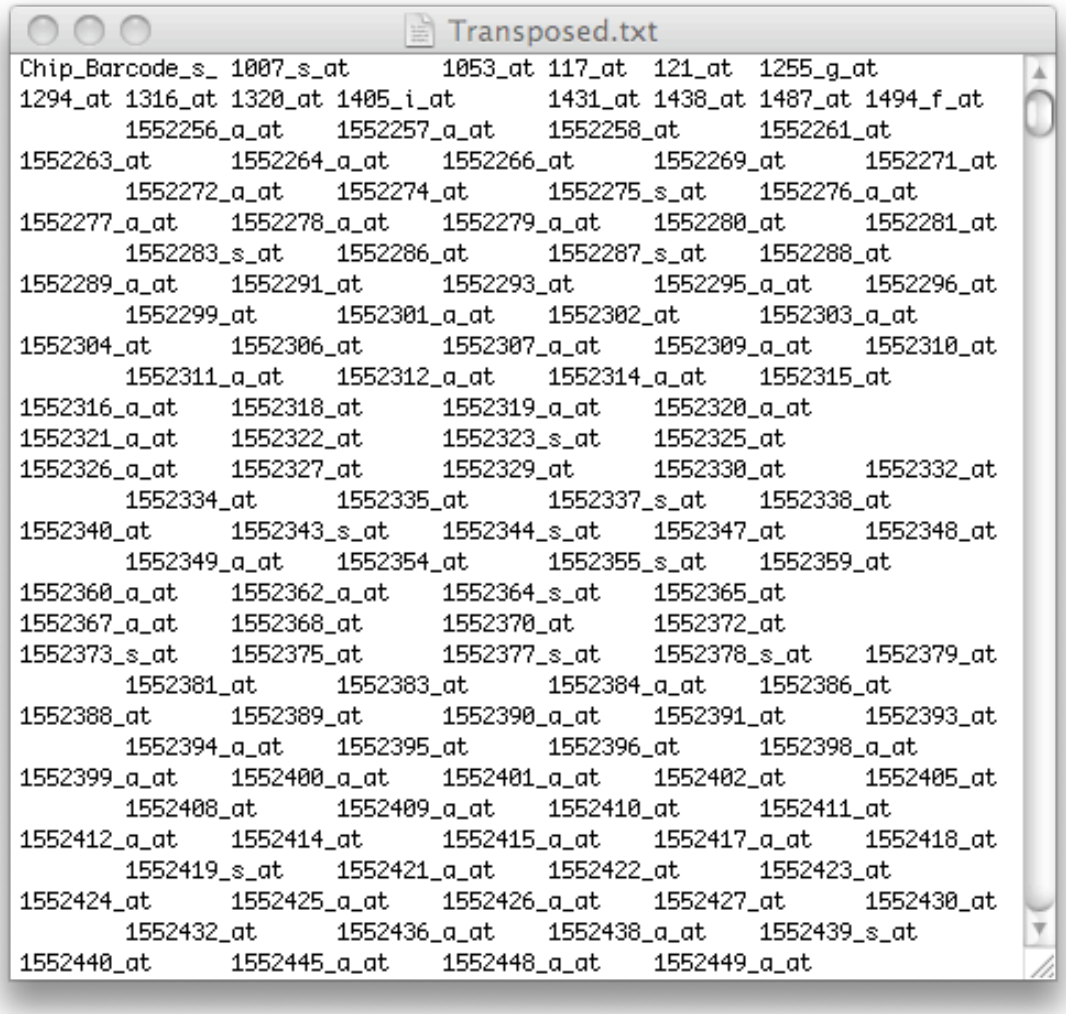
Σε αρχικό στάδιο τα δεδομένα περιέχονταν σε ένα αρχείο με περίπου 56 χιλιάδες στήλες, όσος και ο αριθμός των probeSets που μελετείται και 190 γραμμές που περιέχουν τα δεδομένα για κάθε άνθρωπο που συμμετείχε στο πείραμα. Για κάθε ένα από τα διαφορετικά probeSets υπήρχε μία τιμή διαφορετική για τον κάθε άνθρωπο ξεχωριστά.

3.6 Μετάθεση Αρχείου Δεδομένων (File Transpose)

Για την ανάλυση των δεδομένων του αρχείου αναγκαία ήταν η μετάθεση των περιεχομένων του, δηλαδή η μετατροπή των δεδομένων του από γραμμές σε στήλες και αντίστροφα. Δημιουργία δηλαδή ενός νέου αρχείου με τα ίδια δεδομένα όπως και το αρχικό, με τη διαφορά ότι βρίσκεται σε μετατεθειμένη μορφή (transposed).

Λόγω του μεγάλου όγκου των δεδομένων του αρχείου, ο πρώτος και πιο σημαντικός στόχος της μελέτης ήταν η διάσπαση του σε μικρότερα, ώστε η ανάλυση των δεδομένων να γίνει πιο εύκολη και πιο γρήγορη. Για να μειωθεί ο χρόνος εκτέλεσης χρησιμοποιήθηκε ένα πλέγμα υπολογιστών (grid) του οποίου η χρήση αποσκοπούσε στην παράλληλη εκτέλεση των αλγόριθμων που υλοποιήθηκαν. Με τον τρόπο αυτό

επιτεύχθηκε μείωση του συνολικού χρόνο που ήταν απαραίτητος για την ανάλυση και την επεξεργασία των δεδομένων.



Εικόνα 3.1 Το Μετατεθημένο Αρχείο

3.7 Διαδικασία Διάσπασης Δεδομένων

Λόγω του μεγάλου όγκου του αρχείου με τα δεδομένα, απαραίτητη ήταν η διάσπαση του σε άλλα αρχεία μικρότερου μεγέθους. Οι μέθοδοι που χρησιμοποιήθηκαν για την ανάλυση των δεδομένων, απαιτούσαν τεράστια ποσά χρόνου για την εκτέλεση τους. Επομένως η διάσπαση σε μικρότερα αρχεία, αποσκοπούσε κυρίως στη μείωση του χρόνου εκτέλεσης και διευκόλυνση της ανάλυσης των δεδομένων.

3.7.1 Ποιοτικός Έλεγχος Δεδομένων (Quality Control)

Ως μέρος της προετοιμασίας των δεδομένων έγινε έλεγχος ποιότητας (Quality Control) στα δεδομένα, ώστε να αφαιρεθούν τα δείγματα που παρουσίαζαν έλλειψη δεδομένων. Βρέθηκε ένα μόνο δείγμα με ελλειπή δεδομένα, το οποίο μετά από εισήγηση των ειδικών βιολόγων της ομάδας, αφαιρέθηκε από το σύνολο. Αυτή η διαδικασία εκτελέστηκε έτσι ώστε να είναι βέβαιο ότι τα δεδομένα είναι ακέραια και τα αποτελέσματα που θα προκύψουν από την ανάλυση θα είναι όσο το δυνατό πιο έγκυρα.

3.7.2 Διάσπαση Αρχείου Δεδομένων

Σαν δεύτερο στάδιο της επεξεργασίας των δεδομένων, ήταν η διάσπαση του μεγάλου αρχείου σε μικρότερα, διευκολύνοντας με αυτό τον τρόπο τη διαδικασία ανάλυσης των δεδομένων. Ο αριθμός των probeSets ανά αρχείο παραμετροποιήθηκε αφού αυτός όσο πιο μικρός ήταν, τόσο θα αυξανόταν το overhead της ανάλυσης, ενώ αντίθετα όσο πιο μεγάλος θα ήταν, θα αυξανόταν το μέγεθος των ενδιάμεσων δεδομένων που θα παράγονταν. Η διάσπαση έγινε αρχικά σε αρχεία μεγέθους των 10 probeSets ανά αρχείο. Με την ενέργεια αυτή δεν επιτεύχθηκε η μείωση του όγκου των ενδιάμεσων δεδομένων, με αποτέλεσμα να εξακολουθεί να είναι μεγαλύτερος από τον διαθέσιμο αποθηκευτικό χώρο στη μνήμη. Ακολούθησε περεταίρω διάσπαση σε ακόμη μικρότερα αρχεία της τάξεως του ενός probeSet ανά αρχείο. Το βήμα αυτό οδήγησε στη δημιουργία 56 χιλιάδων διαφορετικών αρχείων, κάθε ένα από τα οποία αντιστοιχεί σε ένα probeSet. Τα αρχεία που δημιουργήθηκαν περιέχουν τις πληροφορίες προς ανάλυση των δειγμάτων. Επιπλέον, αυτός ο τρόπος διάσπασης των δεδομένων διευκόλυνε την μετέπειτα ανάλυση και επεξεργασία τους.

Ο μεγάλος αριθμός αρχείων, καθιστούσε χρονοβόρα την σειριακή ανάλυση τους. Η χρήση διανεμημένου υπολογισμού, αποσκοπούσε στην μείωση του συνολικού χρόνου εκτέλεσης των μεθόδων ανάλυσης των δεδομένων. Η μείωση επιτεύχθηκε με παραλληλοποίηση της ανάλυσης των δεδομένων, με την χρήση όλων των διαθέσιμων πόρων που μπορούσε να παρέχει το grid, βάσει των παραμέτρων προτεραιότητας με τις οποίες είχαν αρχικά καταχωρηθεί στο grid τα δεδομένα προς επεξεργασία. Το συγκεκριμένο grid αποτελείτο από ένα σύνολο 200 υπολογιστών περίπου των οποίων η

χρήση καθορίζεται από τις παραμέτρους που καθόριζαν την σειρά προτεραιότητας και σημαντικότητας των διεργασιών όσον αφορά τη χρήση του grid.

3.8 Φιλτράρισμα Δεδομένων

Το σημαντικότερο μέρος από το οποίο αποτελείται η επεξεργασία και η προετοιμασία των δεδομένων, καθώς επίσης και των αποτελεσμάτων που παράχθηκαν κατά την ανάλυση, είναι η απομόνωση των πιο σημαντικών από αυτά, αλλά εξίσου σημαντικός είναι και ο περιορισμός του όγκου δεδομένων. Για την επίτευξη των δύο αυτών στόχων, εφαρμόστηκε φιλτράρισμα στα δεδομένα, με τη χρήση κάποιου κατωφλίου, βάσει του οποίου έγινε η επιλογή των σημαντικότερων δειγμάτων για τη φάση των ακατέργαστων δεδομένων και αργότερα για τη συλλογή των σημαντικότερων από τα αποτελέσματα.

Το φιλτράρισμα εφαρμόστηκε μετά τη διαδικασία διάσπασης των δεδομένων σε ένα probeSet ανά αρχείο, που καθιστούσε και τη διαδικασία συλλογής των σημαντικότερων δεδομένων από αυτά για κάθε συγκεκριμένο probeSet. Στην δεύτερη περίπτωση το φιλτράρισμα εφαρμόστηκε στο στάδιο της μετά επεξεργασίας των δεδομένων κατά τη συγχώνευση των αρχείων και τη συλλογή των αποτελεσμάτων υψίστης σημασίας.

Με τον τρόπο αυτό απομονώθηκαν τα σημαντικότερα δεδομένα και πληροφορίες για κάθε στάδιο αντίστοιχα, αυτά δηλαδή με τις μικρότερες τιμές των πιθανοτήτων σημαντικότητας (p-values).

3.8.1 Φιλτράρισμα για περιορισμό του όγκου των Δεδομένων

Ο περιορισμός του όγκου δεδομένων αποτέλεσε σημαντικό παράγοντα για δημιουργία και εφαρμογή κώδικα φιλτραρίσματος στα δεδομένα. Λόγω του μεγάλου όγκου δεδομένων, η δημιουργία και εφαρμογή κώδικα που φιλτράρει τα δεδομένα με τη χρήση ενός κατωφλίου κρίθηκε απαραίτητη. Το κατώφλι (threshold) στον κώδικα εισάγεται από τον προγραμματιστή ως παράμετρος, από την γραμμή εντολών κατά την εκτέλεση. Με τον περιορισμό του όγκου των δεδομένων στο αρχικό στάδιο επιτεύχθηκε περιορισμός των δεδομένων στα σημαντικότερα από αυτά.

Όπως προαναφέρθηκε, σημαντικότερα δεδομένα αποτελούσαν τα δείγματα με τις μικρότερες τιμές στην πιθανότητα σημαντικότητας. Περιορίζοντας όσο το δυνατό περισσότερο το κατώφλι σύγκρισης των δεδομένων για το φιλτράρισμα τους, απομόνωνε τα δεδομένα στα περισσότερο σημαντικά.

Η σημασία αυτής της διαδικασίας κατά την ανάλυση και επεξεργασία των δεδομένων αποδείχθηκε να είναι μεγάλη, καθώς παρατηρήθηκαν συγκριτικά μικρότεροι χρόνοι εκτέλεσης των αλγορίθμων ανάλυσης. Επιπλέον τα αποτελέσματα περιορίστηκαν στα σημαντικότερα και αργότερα κατά την γραφική απεικόνιση τους, τα αποτελέσματα θα ήταν πιο ευδιάκριτα στο μάτι και θα οδηγούσαν στη διεξαγωγή πιο ξεκάθαρων συμπερασμάτων.

Αυτού του είδους φιλτράρισμα εφαρμόστηκε αμέσως μετά το στάδιο της διάσπασης των δεδομένων και ένα στάδιο πριν την ανάλυσή τους και συγκεκριμένα υλοποιήθηκε κατάλληλος κώδικας που επεξεργαζόταν τα δεδομένα προτού περάσουν στο στάδιο της ανάλυσης.

Κεφάλαιο 4

Αλγόριθμοι Ανάλυσης Δεδομένων

4.1 Ανάλυση Δεδομένων και Διαχείριση Αρχείων

Η ανάλυση των δεδομένων αποτελεί το σημαντικότερο μέρος της όλης μελέτης. Όπως αναφέρθηκε και σε προηγούμενα υποκεφάλαια, ο αριθμός ελέγχων που έπρεπε να διεξαχθούν ήταν τεράστιος καθιστώντας την επεξεργασία με τη χρήση ενός και μόνο υπολογιστικού συστήματος πολύ χρονοβόρα και πολύπλοκη διαδικασία. Για το λόγο αυτό είναι που χρησιμοποιήθηκε και το κατανεμημένο σύστημα (grid) που ήταν διαθέσιμο και οι αλγόριθμοι που υλοποιήθηκαν, αποσκοπούσαν στη βέλτιστη χρήση του.

Ένας πολύ σημαντικός παράγοντας που περιόριζε την χρήση του grid, ήταν ο περιορισμένος αποθηκευτικός χώρος στη μνήμη, που ήταν διαθέσιμος από το σύστημα. Ο όγκος των αποτελεσμάτων υπολογίστηκε να ξεπερνά τα 400TB ενώ ο διαθέσιμος αποθηκευτικός χώρος έφτανε μόλις τα 200GB.

Το πρόβλημα αυτό αντιμετωπίστηκε χρησιμοποιώντας αλγόριθμους συμπίεσης/αποσυμπίεσης κατά την εκτέλεση των μεθόδων ανάλυσης πάνω στα δεδομένα, σε συνδυασμό με κώδικα φιλτραρίσματος για μείωση του όγκου των πληροφοριών, απομονώνοντας τις σημαντικότερες από αυτές.

Συγκεκριμένα κατά την ανάλυση των δεδομένων ενός αρχείου, μετά την παραγωγή κάποιου αποτελέσματος, και πριν από την καταγραφή του σε ένα νέο αρχείο εξόδου, εφαρμοζόταν σε αυτό κώδικας φιλτραρίσματος, που επέλεγε τις πληροφορίες βάσει ενός προκαθορισμένου κατωφλίου, καθώς επίσης και με την ενδιαφέρουσα στατιστική παράμετρο. Στη συνέχεια οι πληροφορίες που ικανοποιούσαν όλες τις συνθήκες επιλογής, καταγράφονταν στο αρχείο και συμπιέζονταν μειώνοντας έτσι στο μεγαλύτερο δυνατό βαθμό το μέγεθος του αρχείου.

Το όνομα κάθε αρχείου καθοριζόταν από το όνομα του probeSet για το οποίο περιείχε τα δεδομένα. Η ονομασία με αυτό τον τρόπο έπαιξε σημαντικό ρόλο για το τελικό στάδιο εφαρμογής περεταίρω φιλτραρίσματος. Η διαχείριση του grid και των αρχείων όπως επίσης και των αλγορίθμων ανάλυσης, του κώδικα συμπίεσης των αρχείων καθώς επίσης και του κώδικα φιλτραρίσματος, έγινε με τη βοήθεια script files και κώδικα γραμμένο σε C++ που αυτοματοποιούσαν την όλη διαδικασία.

Αποτέλεσμα της ανάλυσης των δεδομένων ήταν η παραγωγή 56 χιλιάδων συμπιεσμένων αρχείων, ένα για κάθε διαφορετικό probeSet. Στα αρχεία αυτά υπάρχουν οι τιμές των πιθανοτήτων σημαντικότητας για τα SNPs με τα οποία το probeSet ξεπερνά την τιμή κατωφλίου.

4.2 Μέθοδοι Ανάλυσης

Για την ανάλυση των δεδομένων εφαρμόστηκαν τρείς διαφορετικοί αλγόριθμοι που ήταν διαθέσιμοι από το εργαλείο plink. Οι τρείς αυτοί αλγόριθμοι ανάλυσης είναι η Ποσοτική Ανάλυση (Quantitative Trait Analysis), οι Διεργασίες Μεταλλαγής (Permutation Procedures) και η εφαρμογή Γραμμικών και Λογιστικών Μοντέλων ανάλυσης (Linear and Logistic Models), που αναλύονται με μεγαλύτερη λεπτομέρεια στη συνέχεια.

4.2.1 Ποσοτική Ανάλυση (Quantitative Trait Analysis)

Ποσοτική ανάλυση γνωρισμάτων (Quantitative Trait Analysis) είναι η μέθοδος με την οποία επιτυγχάνεται ο προσδιορισμός του επιπέδου σημασίας μιας μεμονωμένης ακολουθίας DNA. Ο προσδιορισμός του επιπέδου σημασίας μιας ακολουθίας DNA, γίνεται σε σχέση με τις υπόλοιπες γενετικές και μη επιδράσεις που παρατηρούνται πάνω σε ένα συγκεκριμένο γνώρισμα. Στη συγκεκριμένη μελέτη τα δεδομένα που μελετήθηκαν ήταν ένα σύνολο από 550 χιλιάδες SNPs και ένα σύνολο από 56 χιλιάδες μετάγραφα mRNA τα οποία συσχετίζονταν μεταξύ τους ως ζεύγη καρτεσιανού γινομένου, για κάθε ένα από τα οποία παραγόταν και μία τιμή σημαντικότητας (p-value).

Αυτό το είδος ανάλυσης οδήγησε στην παραγωγή ενός πολύ μεγάλου όγκου αποτελεσμάτων, συνολικά περίπου 550 χιλιάδες επί 56 χιλιάδες για κάθε διαφορετική συσχέτιση.

Η διαχείριση των αποτελεσμάτων επεξηγείται πιο αναλυτικά κατά την ανάλυση της διαδικασίας μετά επεξεργασίας των αποτελεσμάτων [8].

4.2.2 Διεργασίες Μεταλλαγής (Permutation Procedures)

Οι διεργασίες μεταλλαγής αποτελούν μια εντατικά υπολογιστική προσέγγιση για την παραγωγή επιπέδων σημασίας με εμπειρικό τρόπο. Οι τιμές που παράγονται με αυτόν τον τρόπο έχουν κάποιες ιδιότητες, όπως για παράδειγμα η χαλάρωση των υποθέσεων

όσον αφορά την κανονικοποίηση συνεχόμενων φαινοτύπων και η αρχή των Hardy-Weinberg που ασχολείται με τη διαχείριση σπάνιων αλληλόμορφων και δηλώνει πως οι συχνότητες των αλληλόμορφων όπως επίσης και των γονοτύπων παραμένουν σταθερές, βρίσκονται δηλαδή σε ισορροπία. Επιπλέον ασχολείται και με δείγματα μικρού μεγέθους, παρέχοντας με αυτό τον τρόπο ένα πλαίσιο εργασίας για διόρθωση, όσον αφορά τους πολλαπλούς ελέγχους [5] καθώς επίσης του ελέγχου των αναγνωρισμένων υποδομών ή άλλων συγγενικών συσχετίσεων, εφαρμόζοντας την διαδικασία μεταλλαγής μόνο σε μία ομάδα.

Οι διεργασίες μεταλλαγής προσφέρονται για μια πληθώρα δοκιμών και χωρίζονται σε δύο κατηγορίες ανάλογα με τους τομείς στους οποίους εφαρμόζονται.
Οι δύο κατηγορίες είναι:

1. Label – swapping έναντι gene dropping
2. Adaptive έναντι max(T)

Για τους σκοπούς αυτής της μελέτης χρησιμοποιήθηκε ο προσαρμοστικός αλγόριθμος μεταλλαγής (adaptive permutation) καταλήγοντας όμως στο τέλος να είναι ίσος με τον αλγόριθμο μεγίστου κατωφλίου (max(T)), το οποίο αναλύεται περιληπτικά στη συνέχεια .

Ακολουθώντας την προσαρμοστική προσέγγιση, οι μεταλλαγές που εφαρμόζονται στα SNPs τερματίζονται όταν τα αποτελέσματα που παράγονται είναι χαμηλού επιπέδου σημασίας (non significant) απο το αρχικό ακόμη στάδιο, απ’ ότι αν αυτά είναι σημαντικά. Για παράδειγμα, εάν μετά από 10 μεταλλαγές παρατηρηθεί πως για 9 από τα στατιστικά αποτελέσματα που έχουν που έχουν παραχθεί για ένα συγκεκριμένο SNP, είναι μεγαλύτερα από τα ήδη γνωστά αποτελέσματα, τότε δεν υπάρχει λόγος για περαιτέρω επεξεργασία του συγκεκριμένου SNP, καθώς δεν είναι πιθανό να οδηγήσει στη εξαγωγή κάποιου αποτελέσματος υψίστης σημασίας (όσο μεγαλύτερα αποτελέσματα τόσο μικρότερη η σημασία τους). Με τον τρόπο αυτό επιταχύνεται η διεργασία μεταλλαγής. Η επιτάχυνση αυτή επιτυγχάνεται λόγω του γεγονότος ότι τα περισσότερα από τα SNPs που δεν θεωρούνται σημαντικά θα απορριφθούν αρκετά σύντομα, έτσι ώστε να είναι δυνατός ο ορθός υπολογισμός της σημαντικότητας μιας μικρότερης ομάδας SNPs, που απαιτούν εκατομμύρια μεταλλαγές να υπολογιστούν.

Ως συνήθως η ακρίβεια με την οποία γίνεται ο υπολογισμός της σημαντικότητας ενός p-value που σχετίζεται με τον αριθμό των μεταλλαγών που εκτελέστηκαν (permuted), αποτελεί την ίδια την τιμή σημαντικότητας. Για τους περισσότερους όμως σκοπούς χρήσης των p-values, αυτό ακριβώς θα είναι και το επιθυμητό αποτέλεσμα καθώς αποτελούν μικρού ενδιαφέροντος αποτελέσματα, ενώ ένα καθαρά μη συσχετίσιμο SNP έχει στην πραγματικότητα τιμή σημαντικότητας (p-value) ίση με 0.78 ή 0.87.

Λόγω της τεράστιας σημαντικότητας των αποτελεσμάτων που χρησιμοποιήθηκαν στις διεργασίες μεταλλαγής, ο αριθμός των μεταλλαγών που εφαρμόστηκε στα περισσότερα από αυτά, έφτανε τον μέγιστο αριθμό που μπορούσαν να εκτελεστούν.

Για τον λόγο αυτό, από προσαρμοστική μέθοδος (adaptive) από την οποία ξεκίνησε αρχικά η ανάλυση, κατέληξε στην περίπτωση διεργασιών μεταλλαγής μέγιστου κατωφλίου (max(T)). Σε αυτή την περίπτωση αντίθετα με την προσαρμοστική μέθοδο, κανένα από τα SNPs δεν απορρίφθηκε καθ' όλη τη διάρκεια της διαδικασίας. Αυτό είχε ως αποτέλεσμα για κάθε SNP να εκτελείται ο μέγιστος αριθμός permutations που είχαν αρχικά καθοριστεί. Το προτέρημα αυτής της μεθόδου σε αντίθεση με την προσαρμοστική μέθοδο είναι ότι μπορούν να υπολογιστούν δύο διαφορετικά σύνολα εμπειρικών, σημαντικών τιμών. Δηλαδή ο υπολογισμός μιας τιμής σημαντικότητας για κάθε SNP ξεχωριστά, αλλά και μίας άλλης τιμής που ελέγχει το γεγονός ότι ένας μεγάλος αριθμός επιπλέον SNPs έχουν ελεγχθεί. Αυτό επιτυγχάνεται συγκρίνοντας κάθε στατιστική τιμή ελέγχου που είναι ήδη γνωστή έναντι της μέγιστης τιμής από όλες τις στατιστικές τιμές που παράχθηκαν κατά την εφαρμογή των διεργασιών μεταλλαγής σε όλα τα SNPs, για κάθε ένα από τα αντίγραφα. Με άλλα λόγια η τιμή p-value σε αυτή την περίπτωση ελέγχει το σχετικό ποσοστό σφάλματος, καθώς το p-value απεικονίζει την πιθανότητα παρατήρησης ενός στατιστικού πειράματος αυτού του μεγέθους, έχοντας ως δεδομένο πως εξετάστηκαν όλα τα διαθέσιμα δεδομένα που υπήρχαν.

Η μέθοδος Bonferroni λειτουργεί κάτω από την υπόθεση πως όλες οι δοκιμές είναι ανεξάρτητες μεταξύ τους. Αυτό έρχεται σε αντίθεση με τις διεργασίες μεταλλαγής οι οποίες διατηρούν μια συσχετιστική δομή μεταξύ των SNPs, παρέχοντας με αυτό τον τρόπο ευχέρεια διόρθωσης των πολλαπλών δοκιμών. Ακριβώς επειδή η τιμή

ενδιαφέροντος όταν εφαρμόζονται οι διεργασίες μεταλλαγής, είναι η διορθωμένη τιμή p-value, έχει ξεπεραστεί το πρόβλημα των πολλαπλών δοκιμών.

4.2.2.1 Ο ρόλος των Διεργασιών Μεταλλαγής στη Μελέτη

Οι διεργασίες μεταλλαγής χρησιμοποιήθηκαν στην μελέτη για την εξακρίβωση της εγκυρότητας των αποτελεσμάτων που παράχθηκαν από μεθόδους ανάλυσης που εφαρμόστηκαν νωρίτερα για τους σκοπούς της συγκεκριμένης μελέτης. Όπως προαναφέρθηκε η μέθοδος μεταλλαγής είναι πολύ χρονοβόρα διαδικασία και πόσο μάλλον με τόσο μεγάλο αριθμό συσχετίσεων. Όπως προαναφέρθηκε η χρήση των διεργασιών μεταλλαγής αποσκοπούσαν στην εξακρίβωση και επιβεβαίωση της εγκυρότητας των δεδομένων και όχι για την ανάλυση τους. Για τον λόγο αυτό δεν ήταν απαραίτητο να εφαρμοστεί η μέθοδος σε όλο το σύνολο τον δεδομένων ανεξαίρετα και εφόσον αποτελεί μια χρονοβόρα μέθοδο η εφαρμογή της σε ένα υποσύνολο των δεδομένων αυτών, ήταν αρκετή για να βγουν κάποια ενδεικτικά αποτελέσματα που θα ήταν σε θέση να οδηγήσουν στην διεξαγωγή ορθών συμπερασμάτων. Έτσι από το σύνολο των στατιστικά σημαντικών αποτελεσμάτων επιλέγηκαν τυχαία με Bonferroni Correction μόνο χίλια πεντακόσια από αυτά για την εφαρμογή της μεθόδου μεταλλαγής. Ο αριθμός των βρόγχων εκτέλεσης των μεταλλαγών τέθηκαν σε 10^6.

Ο όρος p-value ορίζει την πιθανότητα σημαντικότητας (significance probability) που αποτελεί την μικρότερη τιμή του επιπέδου σημαντικότητας a, για την οποία η μηδενική υπόθεση H_0 ενός ελέγχου απορρίπτεται. Όσο πιο μικρό είναι το επίπεδο σημαντικότητας τόσο πιο δύσκολο είναι να απορριφθεί η μηδενική υπόθεση [15].

4.2.3 Γραμμικά και Λογιστικά Μοντέλα (Linear & Logistic Models)

Τα γραμμικά και λογιστικά μοντέλα ανάλυσης (Linear & Logistic Models) αποτελούν ένα είδος ποσοτικής ανάλυσης που επιτρέπει την ανάλυση μεταξύ ποσοτικών και ποιοτικών μεταβλητών όπως στην περίπτωση μας τα μετάγραφα mRNA και SNPs αντίστοιχα, που χρησιμοποιήθηκαν και στην διαδικασία συσχέτισης ποσοτικών μεταβλητών κατά την ποσοτική ανάλυση που περιγράφηκε προηγουμένως. Η διαφορά αυτών των μοντέλων από την ποσοτική ανάλυση, είναι ότι επιτρέπουν την χρήση

συμμεταβλητών (covariates) για μελέτη των αλληλεπιδράσεων αυτών των συμμεταβλητών (covariates) με τις διάφορες συσχετίσεις των μετάγραφων mRNA και των σημειακών νουκλεοτιδικών πολυμορφισμών (SNPs) που προκύπτουν κατά την ανάλυση. Δηλαδή, το πως επηρεάζουν οι συμμεταβλητές αυτές στη συσχέτιση των SNPs με την έκφραση του mRNA.

Επομένως τα μοντέλα αυτά σχηματίζουν όλες τις πιθανές συσχετίσεις μεταξύ των SNPs και των μετάγραφων mRNA ελέγχοντας κάθε φορά κατά πόσο οι συμμεταβλητές (covariates) που θέτονται επηρεάζουν με κάποιο τρόπο στις συσχετίσεις αυτές. Υπάρχουν δύο διαφορετικά είδη τύπων συμμεταβλητών που είναι οι συνεχόμενες και οι δυαδικές (continuous and binary).

Τα covariates που ελέχθησαν στη μελέτη αυτή είναι η ασθένεια (disease), το φύλο (gender), το batch (το χρονικό πλαίσιο στο οποίο έγινε η ανάλυση), που είναι δυαδικού τύπου και η ηλικία (age) που είναι συνεχόμενου τύπου (continuous).

Η ανάλυση εκτελέστηκε σε δύο φάσεις όπου στην πρώτη αναλύθηκε το πρώτο μισό από τα δεδομένα και στην δεύτερη τα υπόλοιπα. Το batch καθορίζει σε ποιά από τις δύο αυτές φάσεις αναλύθηκαν τα δεδομένα.

4.3 Εργαλεία που χρησιμοποιήθηκαν

Για την διεξαγωγή της μελέτης απαραίτητη ήταν η χρήση ορισμένων εργαλείων. Αυτά αποτελούσαν κάποιες εφαρμογές που διατίθενται δωρεάν όπως η εφαρμογή plink που χρησιμοποιήθηκε για την ανάλυση των δεδομένων, αλλά και άλλες, όπως είναι η εφαρμογή Spotfire για την παρουσίαση των αποτελεσμάτων, καθώς επίσης και κάποιοι διαθέσιμοι πόροι σε υλικό που επίσης ήταν διαθέσιμη από την εταιρεία για ικανοποίηση των αναγκών κατά τη διεξαγωγή της μελέτης.

Η προσαρμογή του plink στο grid έγινε με τη χρήση λογισμικού που υλοποιήθηκε ειδικά για τη συγκεκριμένη μελέτη όπως επίσης και για την προσαρμογή των αποτελεσμάτων της ανάλυσης στην εφαρμογή Spotfire, για την τελική τους παρουσίαση.

4.3.1 Εφαρμογή Plink

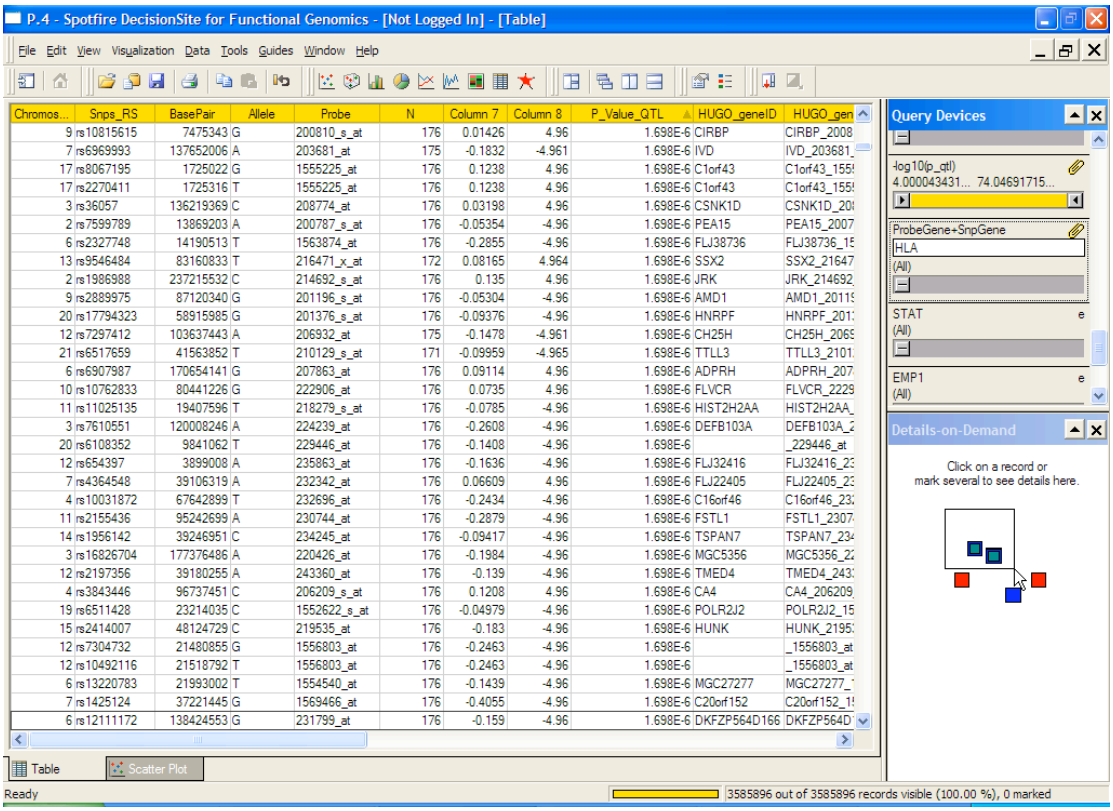
Το plink αποτελεί μια εφαρμογή που προσφέρεται ως ανοικτός κώδικας και χρησιμοποιείται για ανάλυση δεδομένων και συγκεκριμένα ως ένα εργαλείο που μπορεί να αναλύσει τις συσχετίσεις που μπορεί να υπάρχουν σε ολόκληρο το γονιδίωμα. Επιπλέον είναι σχεδιασμένο για να εκτελεί μια σειρά από αναλύσεις ευρείας κλίμακας με αποτελεσματικό και αποδοτικό τρόπο όσον αφορά την μεθοδολογία υπολογισμού.

4.3.2 Εφαρμογή Spotfire

Το spotfire αποτελεί ένα εργαλείο του οποίου η διαδραστική απεικόνιση της πληροφορίας καθώς επίσης και οι αναλυτικές λύσεις που προσφέρει στους χρήστες, τους χαρίζει μια αξιόλογη εμπειρία όσον αφορά την γρήγορη και εύκολη αναζήτηση σε βάσεις δεδομένων αλλά και αναφορά αποτελεσμάτων που είναι χρήσιμα για ψηλότερου επιπέδου επιστημονικών απαιτήσεων. Επιτρέπει την γραφική απεικόνιση αποτελεσμάτων και αποτελεί ενα ευκολόχρηστο εργαλείο δια τη διαχείριση δεδομένων αφου μέσω αυτού μπορούν να εφαρμοστούν διάφορα φίλτρα και ερωτήματα επιλογής (queries) για την απομόνωση δεδομένων.

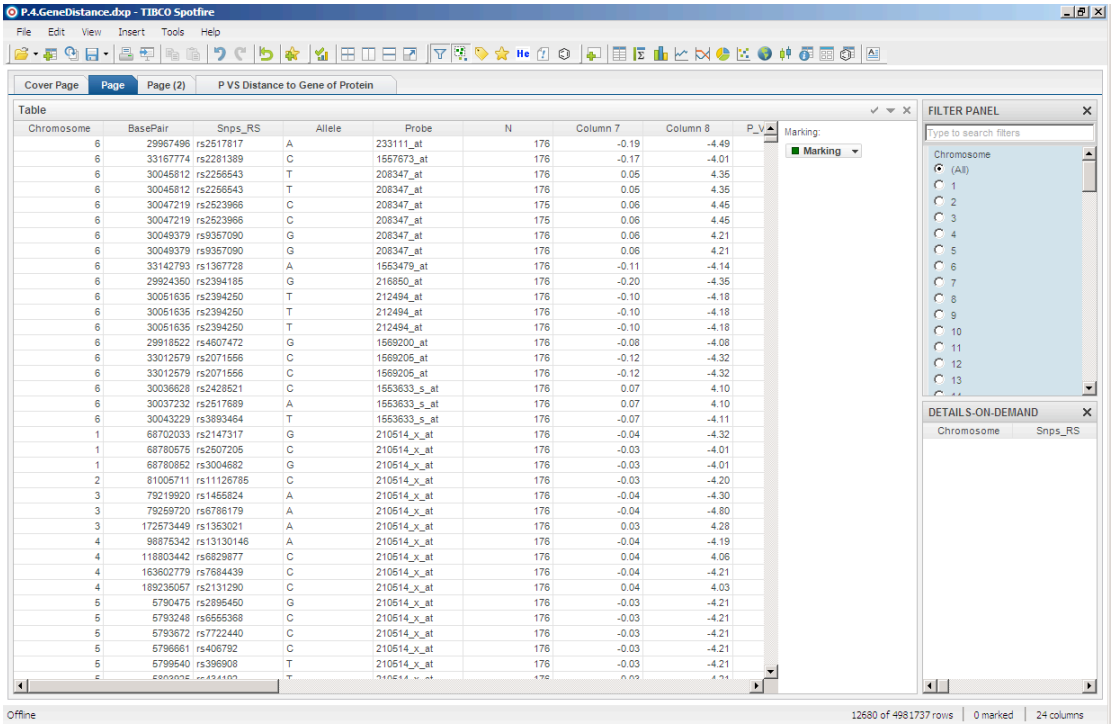
Ο ρόλος της εφαρμογής στη συγκεκριμένη μελέτη ήταν για διευκόλυνση της διαχείρισης των δεδομένων καθώς επίσης και γραφική απεικόνιση και παρουσίαση των αποτελεσμάτων.

4.3.2.1 Παραδείγματα Χρήσης της Εφαρμογής Spotfire



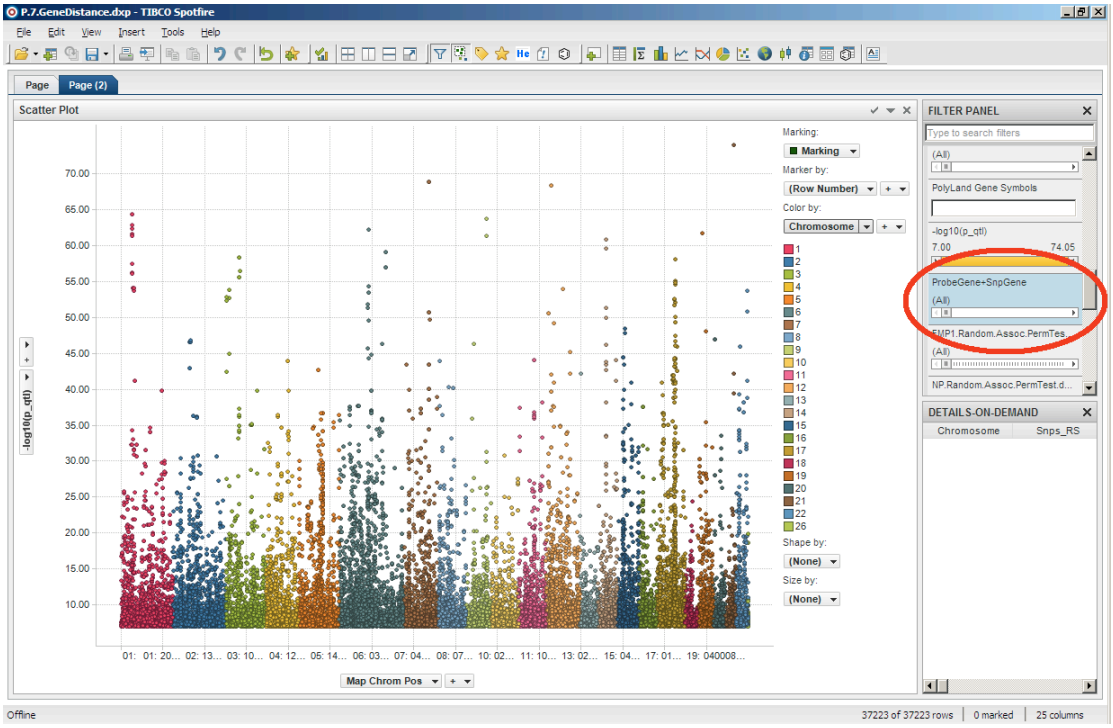
Εικόνα 4.1 Υπηρεσίες Αναζήτησης του Spotfire

Η εικόνα 4.1 αποτελεί ένα δείγμα της εισαγωγής των αποτελεσμάτων υπο μορφή πίνακα στην εφαρμογή Spotfire. Δεξιά μπορούν να παρατηρηθούν τα υπηρεσίες αναζήτησης, με την βοήθεια των οποίων έχει γίνει αναζήτηση της επιλογής των αποτελεσμάτων όπου το mRNA ή το SNP ανήκει σε ένα από τα γονίδια της οικογένειας HLA.



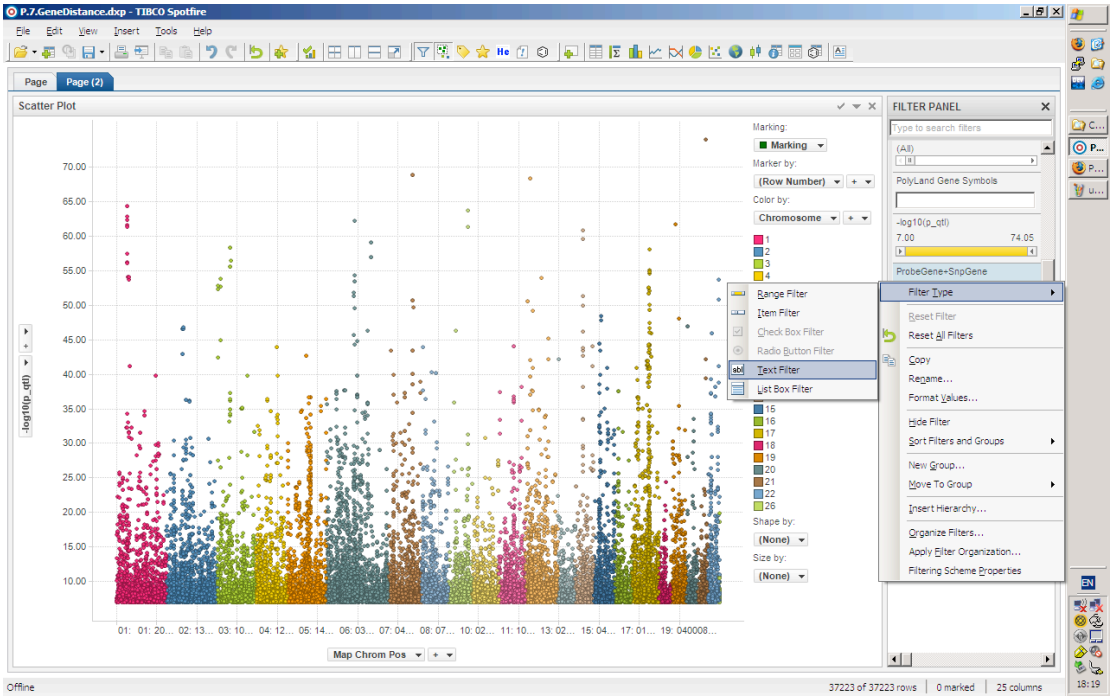
Εικόνα 4.2 Φιλτράρισμα βάσει Χρωμοσώματος στο Spotfire

Η εικόνα 4.2 αποτελεί ένα δείγμα από τα στοιχεία που περιέχει το αρχείο με τις πληροφορίες που παράχθηκαν κατά την εφαρμογή του φίλτρου με κατώφλι 10^{-4} και έχουν εισαχθεί στην εφαρμογή Spotifire. Η προσοχή εστιάζεται πάνω δεξιά στο filter panel από όπου μπορεί να επιλεγεί κάποιο συγκεκριμένο χρωμόσωμα.



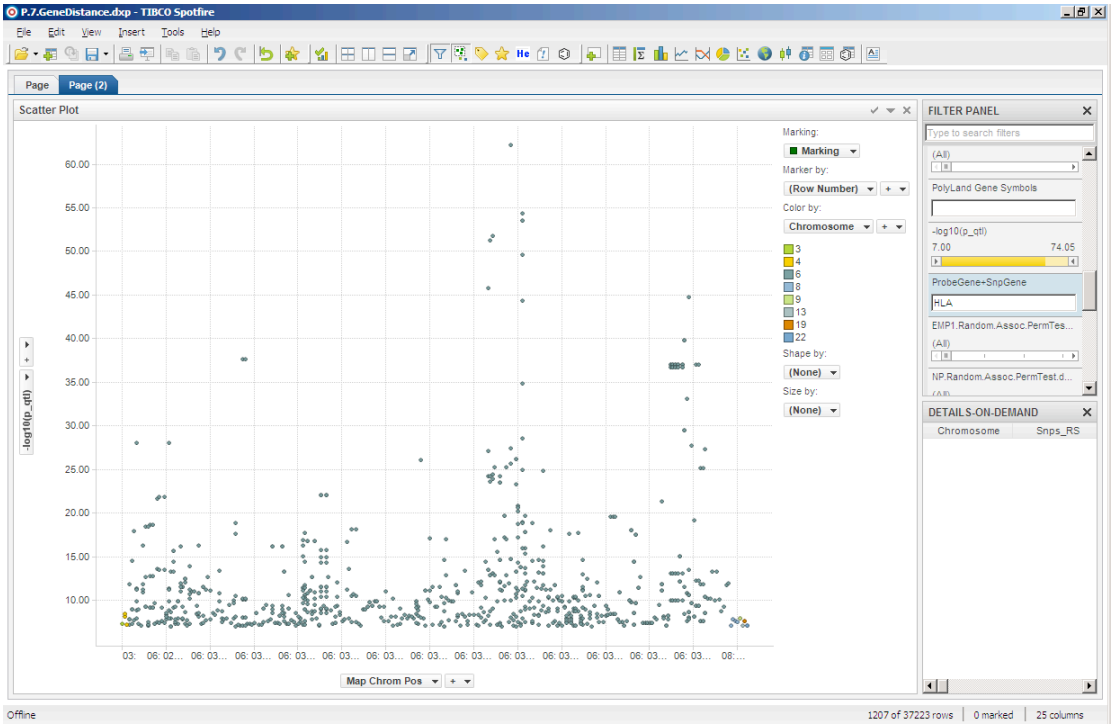
Εικόνα 4.3 Εφαρμογή Φίλτρου μέσω Spotfire

Στην εικόνα 4.3 παρουσιάζεται η διαδικασία εφαρμογής φίλτρου στα αποτελέσματα χρησιμοποιώντας την εφαρμογή Spotfire. Μπορεί να παρατηρηθεί πως η διαδικασία είναι σχετικά απλή και αρκεί μόνο να επιλεγεί το κατάλληλο πεδίο στο οποίο θα εφαρμοστεί το επιθυμητό φίλτρο.



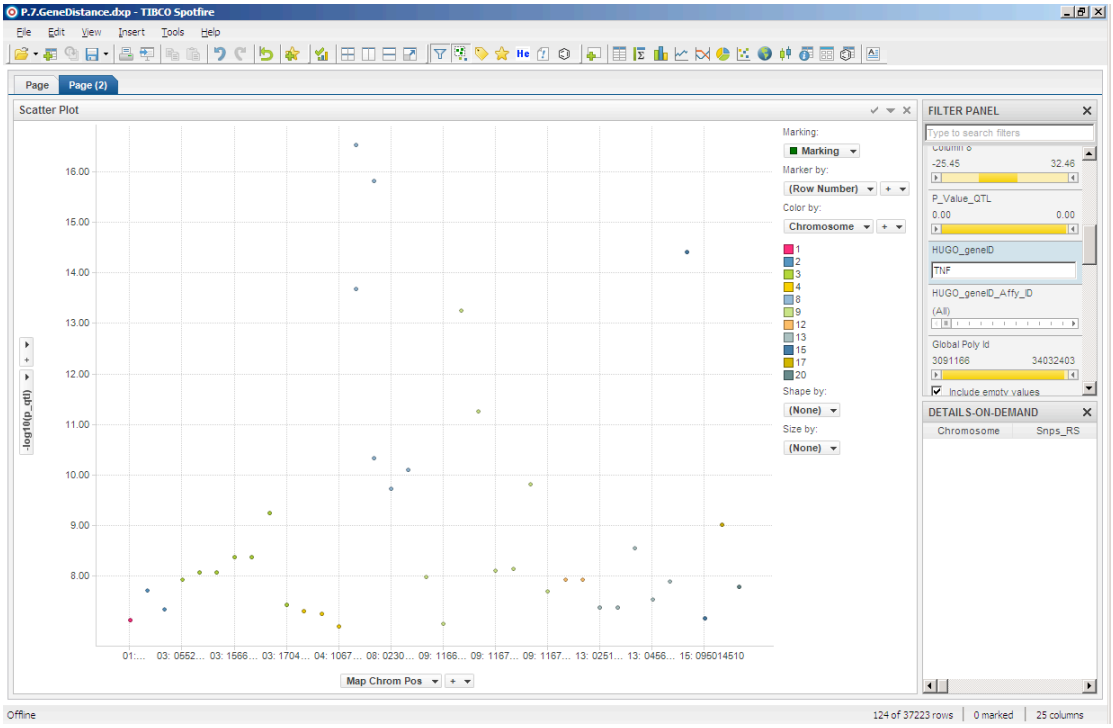
Εικόνα 4.4 Διαδικασία Εφαρμογής Νέου Φίλτρου μέσω Spotfire

Στην εικόνα 4.4 παρουσιάζεται ο τρόπος με τον οποίο επιλέγεται ο τύπος του φίλτρου που θα εφαρμοστεί. Όπως μπορεί να παρατηρηθεί υπάρχουν αρκετοί διαφορετικοί φίλτρου. Αυτό που καθορίζει τον τύπο του φίλτρου που θα χρησιμοποιηθεί είναι οι επιθυμητές γραφικές παραστάσεις που θα παραχθούν μέσω της επιλογής του φίλτρου και επιπλέον την ακρίβεια με την οποία θα γίνει το φιλτράρισμα των αποτελεσμάτων. Όπως και προηγουμένως μπορεί να παρατηρηθεί πως η διαδικασία που ακολουθείται είναι πολύ απλή και εύκολη.



Εικόνα 4.5 Αποτελέσματα Φιλτραρίσματος της οικογένειας γονιδίων HLA

Στην εικόνα 4.5 μπορούν να παρατηρηθούν τα αποτελέσματα της γραφικής που παράγεται με την εφαρμογή φίλτρου σε μία συγκεκριμένη οικογένεια γονιδίων. Στην περίπτωση αυτή το φιλτράρισμα εφαρμόστηκε στην οικογένεια γονιδίων HLA.



Εικόνα 4.6 Αποτελέσματα Φιλτραρίσματος της οικογένειας γονιδίων TNF

Στην εικόνα 4.6 μπορούν να παρατηρηθούν τα αποτελέσματα της γραφικής που παράγεται με την εφαρμογή φίλτρου σε ένα συγκεκριμένο γονίδιο. Στην περίπτωση αυτή το φιλτράρισμα εφαρμόστηκε για το γονίδιο TNF. Επιπλέον παρατηρείται πως τα αποτελέσματα σε σχέση με την εφαρμογή φίλτρου στο γονίδιο HLA που παρουσιάστηκε νωρίτερα, είναι διαφορετικά και η διαδικασία εξαγωγής αυτής της διαφορετικής γραφικής είναι πολύ απλή και καθόλου χρονοβόρα. Ακολουθήθηκε η ίδια διαδικασία και για το φιλτράρισμα του γονιδίου BRCA1 τα αποτελέσματα του οποίου επίσης θα σχολιαστούν στη συνέχεια.

4.3.3 Grid

Το πλέγμα υπολογιστών (grid) αποτελούσε το κύριο υλικό μέσο που χρησιμοποιήθηκε για την ανάλυση των δεδομένων. Το πλέγμα υπολογιστών διέθετε 200 επεξεργαστές και χάρη στην υλοποίηση ειδικού κώδικα για τους σκοπούς βελτιστοποίησης των συνθηκών εκτέλεσης, επιτεύχθηκε η επιτάχυνση της ανάλυσης, περίπου κατά 200 φορές. Η χρήση του διευκόλυνε αλλά και επιτάχυνε κατά μεγάλο βαθμό τη διαδικασία ανάλυσης των δεδομένων, λόγω και του μεγάλου αριθμού αρχείων που προέκυψαν

κατά τη διαδικασία της διάσπασης, των οποίων η διαχείριση και ανάλυση αποτελούσε πολύπλοκη και χρονοβόρα διαδικασία χωρίς τη χρήση του grid.

4.3.3.1 Αποθηκευτικός Χώρος

Ένας από τους περιορισμούς που έπρεπε να αντιμετωπιστούν όσον αφορά τη χρήση του grid ήταν το μέγεθος της διαθέσιμης αποθηκευτικής μνήμης που παρείχε το σύστημα. Ο διαθέσιμος αποθηκευτικός χώρος ανερχόταν περίπου στα 200 GB ενώ ο πραγματικός χώρος που ήταν αναγκαίος για αποθήκευση των αποτελεσμάτων, ήταν της τάξεως των 400TB. Κατά την παραγωγή των αποτελεσμάτων χρησιμοποιήθηκε αλγόριθμός φιλτραρίσματος και συμπίεσης των αρχείων έτσι ώστε να περιοριστεί το ποσό της μνήμης που ήταν αναγκαίο για την αποθήκευση τους, χωρίς όμως να χαθούν σημαντικές πληροφορίες. Στη συνέχεια με τη χρήση παραμετροποιημένου αλγορίθμου αποσυμπίεσης αλλά και συγχώνευσης, επιτεύχθηκε η συλλογή των αποτελεσμάτων σε ένα κοινό αρχείο, που καταλάμβανε χώρο ανάλογο του αριθμού των κορυφαίων αποτελεσμάτων που προσδιορίστηκαν, καθώς επίσης και του επιπέδου λεπτομέρειας της περιγραφής τους.

4.3.3.2 Διαθεσιμότητα Συστήματος

Οι πόροι του υφιστάμενου συστήματος δεν ήταν αφοσιωμένοι στις υπολογιστικές ανάγκες της μελέτης αυτής ανά πάσα στιγμή, καθώς ήταν προσβάσιμοι και από δεκάδες άλλες ομάδες ερευνητών. Η επιτυχής εξυπηρέτηση όλων των διαφορετικών χρηστών που επιχειρούν να χρησιμοποιήσουν το grid, γίνεται με τη βοήθεια την πλατφόρμας L.S.F. Η πλατφόρμα L.S.F δίνει προτεραιότητα στις διεργασίες που έχουν τον ψηλότερο δείκτη προτεραιότητας. Ο δείκτης προτεραιότητας για μία διεργασία καθορίζεται από τον εκάστοτε χρήστη. Συγκεκριμένα ο LSF αποτελεί έναν αλγόριθμο που αποσκοπεί στην δίκαιη κατανομή πόρων μεταξύ διεργασιών, και εξυπηρέτηση των χρηστών που επιθυμούν να χρησιμοποιήσουν το σύστημα, αυξάνοντας με αυτό τον τρόπο την αποδοτικότητα του συστήματος, εξυπηρετώντας κάθε χρήστη ανεξαίρετα. Αυτή η δίκαιη κατανομή των πόρων έχει ως στόχο εκτός από τη δίκαιη κατανομή τόσο των επεξεργαστών του συστήματος, όσο και της συνολικής διαθέσιμης μνήμης του συστήματος. Οι διεργασίες διαχωρίζονται σε κάποιες κατηγορίες ανάλογα με τον απαιτούμενο χρόνο εκτέλεσης. Υπάρχουν τέσσερις διαφορετικές κατηγορίες κάθε μια

από τις οποίες υλοποιείται ως μια ουρά που κρατά τις υποψήφιες διεργασίες προς εκτέλεση. Οι τέσσερις κατηγορίες ως προς το χρόνο εκτέλεσης είναι short, medium, long και very long. Η ουρά στην οποία θα υποβληθεί μια διεργασία μέχρι τη σειρά της προς εκτέλεση, επιλέγεται από τον ίδιο τον χρήστη. Μία διεργασία που υποβάλλεται στη ουρά short έχει ως μέγιστο χρόνο εκτέλεσης μέχρι δεκαπέντε λεπτά, ενώ διεργασίες που υποβάλλονται στην medium ουρά έχουν μέγιστο διαθέσιμο χρόνο εκτέλεσης μία ώρα, σαράντα οκτώ ώρες διαθέσιμου χρόνου εκτέλεσης για την ουρά long και απεριόριστο χρόνο εκτέλεση για τις διεργασίες που υποβάλλονται στην ουρά very long. Συνήθως η ουρά medium είναι η προκαθορισμένη ουρά από το σύστημα, εκτός κι αν επιλεγεί κάποια άλλη ουρά από τον χρήστη ανάλογα με τις απαιτήσεις των διεργασιών. Επιπλέον μπορεί να δοθεί κάποιος βαθμός προτεραιότητας στην διεργασία καθορίζοντας με αυτό τον τρόπο τη σειρά που λαμβάνει στην ουρά η συγκεκριμένη διεργασία. Ο βαθμός προτεραιότητας μιας διεργασίας είναι επίσης επιλογή του χρήστη και όσο μεγαλύτερος αυτός ο βαθμός, τόσο πιο σημαντική είναι η διεργασία, θα τοποθετηθεί στις πρώτες θέσεις ώστε να εκτελεστεί το συντομότερο δυνατό. Με τον τρόπο αυτό επιτυγχάνεται η ταξινόμηση των διεργασιών στην ουρά. Αν θεωρηθεί απαραίτητο ή αναγκαίο ο βαθμός προτεραιότητας μπορεί να αλλαχτεί.

Όσον αφορά τα αποτελέσματα που παράγονται από την εκτέλεση των διεργασιών, αυτά μπορούν να στέλνονται κατευθείαν στην προσωπική ηλεκτρονική διεύθυνση του χρήστη ή σε κάποιο αρχείο εξόδου που θα επιλέξει. Σε γενικές γραμμές τα αποτελέσματα αποθηκεύονται στον προσωπικό χώρο του χρήστη, για το λόγο αυτό είναι απαραίτητο να υπάρχει ο κατάλληλος αποθηκευτικός χώρος διαθέσιμος, ή σε αντίθετη περίπτωση να ληφθούν τα κατάλληλα μέτρα ώστε να τα αποτελέσματα να προσαρμόζονται στην διαθέσιμη μνήμη.

Για τους σκοπούς αυτής της μελέτης ο δείκτης προτεραιότητας των διεργασιών που εκκρεμούσαν ως προς τη χρήση του grid, ήταν ο χαμηλότερος δυνατός και η ουρά στην οποία καταχωρούνταν οι διεργασίες αυτή με τον απεριόριστο χρόνο εκτέλεσης έτσι ώστε να μπορούν να προχωρούν προς εκτέλεση οποιαδήποτε στιγμή υπάρχουν διαθέσιμοι πόροι, χωρίς να επηρεάζει οποιεσδήποτε άλλες εκκρεμότητες της εταιρείας.

4.4 Αλγόριθμος Χρήσης της Εφαρμογής Plink μεσω χρήσης του Grid

Για σκοπούς ανάλυσης των δεδομένων χρησιμοποιώντας τους αλγόριθμους ανάλυσης που ήταν διαθέσιμοι από την εφαρμογή plink, εγκαταστάθηκε σε κάθε μηχανή που ήταν ενωμένη στο grid η εφαρμογή plink. Για να μπορέσει να αξιοποιηθεί η εφαρμογή για την χρήση των αλγορίθμων, απαραίτητη ήταν η υλοποίηση κώδικα, που βάσει των δεδομένων που περιείχε κάθε αρχείο, δημιουργούσε scripts ώστε να είναι δυνατή η αυτόματη καταχώρηση τους στο σύστημα προς εκτέλεση. Συγκεκριμένα ο κώδικας σαρώνει τα δεδομένα που περιέχει κάθε αρχείο, διαβάζοντας τα SNPs που περιέχει κάθε αρχείο που αντιστοιχεί σε ένα probeSet. Αφού διαβαστούν τα περιεχόμενα του αρχείου το πρόγραμμα δημιουργεί ένα script file στο οποίο αποθηκεύεται η εντολή με την οποία θα καλείται η κατάλληλη μέθοδος από το plink, με τις σωστές παραμέτρους. Η μέθοδος που καλείτο με τη χρήση αυτών των αρχείων ήταν η διεργασία μεταλλαγής (permutation procedure) για την οποία ήταν αναγκαία η καταχώρηση των SNPs που λάμβαναν μέρος στην διαδικασία και αποτελούσαν και περιεχόμενα του αρχείου με τα δεδομένα. Η καταχώρηση των αρχείων προς επεξεργασία ένα προς ένα θα ήταν πολύ χρονοβόρα κι έτσι με τον τρόπο αυτό επιτεύχθηκε η χρήση του plink και του grid ταυτόχρονα κάνοντας την όλη διαδικασία ανάλυσης των δεδομένων αποδοτικότερη.

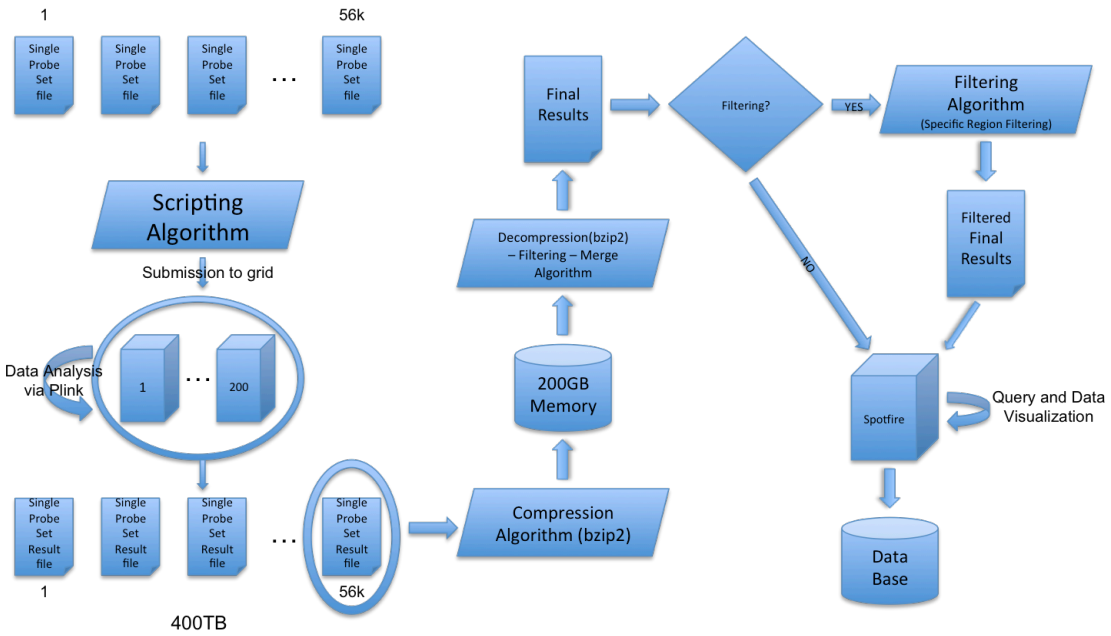
4.5 Αλγόριθμος Προσαρμογής των Αποτελεσμάτων στην Εφαρμογή Spotfire

Ο ουσιαστικός κώδικας για τη μεταφορά των αποτελεσμάτων από το plink στην εφαρμογή Spotfire για γραφική απεικόνιση των αποτελεσμάτων, ήταν αυτός που εκτελούσε την συγχώνευση των αρχείων με τα αποτελέσματα σε ένα και ταυτόχρονα το φιλτράρισμα τους για την απομάκρυνση των άχρηστων πληροφοριών που παράγονται κατά την ανάλυση καθώς επίσης και φιλτράρισμα για την απομόνωση των σημαντικότερων από αυτά. Το τελικό αρχείο όμως εξακολουθούσε να είναι μεγάλο σε σχέση με τον όγκο δεδομένων που μπορούσε να εξυπηρετήσει το Spotfire αποτελεσματικά. Για τον λόγο αυτό υλοποιήθηκε επιπλέον κώδικας που εφάρμοζε περεταίρω φιλτράρισμα πάνω στα δεδομένα και πάλι βάσει ενός κατωφλίου που ορίζεται από τον χρήστη. Με τον τρόπο αυτό επιτυγχάνεται περεταίρω μείωση του όγκου των δεδομένων στα πιο σημαντικά, αλλά και region specific filtering, περιορίζοντας τα δεδομένα για μια συγκεκριμένη περιοχή η οποία επιθυμείται να μελετηθεί. Επιπλέον παρέχεται η δυνατότητα μελέτης διαφόρων περιοχών ανάλογα με

τα ενδιαφέροντα αφού τα αρχικά δεδομένα δεν χάνονται, αλλά παράγονται νέα αρχεία με τα φιλτραρισμένα δεδομένα και τις σημαντικότερες τιμές. Επιπλέον τα μικρότερου όγκου αρχεία μπορούν να εξυπηρετηθούν αποδοτικότερα από το Spotfire χωρίς μεγάλους χρόνους καθυστέρησης στην απόκριση του συστήματος. Αυτός είναι και ο λόγος που στην μελέτη αυτή εφαρμόστηκε φιλτράρισμα αρχικά με κατώφλι $1x10^{-4}$ που παρουσίαζε μεγάλη καθυστέρηση όσον αφορά τον χρόνο απόκρισης της εφαρμογής Spotfire και για αυτό εφαρμόστηκε περεταίρω φιλτράρισμα στα δεδομένα με κατώφλι $1x10^{-7}$ που δεν επηρέαζε την σημαντικότητα των αποτελεσμάτων, απλά μόνο επιτάχυνε την επεξεργασία τους μέσω του Spotfire που χρησιμοποιήθηκε για την γραφική απεικόνιση και την παρουσίαση τους.

Κεφάλαιο 5

Μεταεπεξεργασία



Διάγραμμα 5.1 Μεταεπεξεργασία

5.1 Περιγραφή Διαδικασίας

Ως μέρος της μετά επεξεργασίας των αποτελεσμάτων ήταν η συγχώνευση τους σε ένα ενιαίο αρχείο ώστε να είναι εφικτή στη συνέχεια η απεικόνιση και παρουσίαση των αποτελεσμάτων. Η απεικόνιση των αποτελεσμάτων υπό μορφή γραφικών αναπαραστάσεων έγινε με τη χρήση του εργαλείου spotfire που βρισκόταν διαθέσιμο από την εταιρεία.

Η συγχώνευση των αρχείων έγινε με υλοποίηση κώδικα ειδικά σχεδιασμένου ώστε να μπορεί να ανταποκριθεί στις απαιτήσεις του συστήματος. Για τη συγχώνευση των αποτελεσμάτων απαραίτητη ήταν η αποσυμπίεση των αρχείων με τα αποτελέσματα, που παράχθηκαν κατά το στάδιο της προ επεξεργασίας. Πολλά από τα αποτελέσματα που παράχθηκαν κατά την ανάλυση των δεδομένων ήταν περιττά και για το λόγο αυτό ο περιορισμός τους στα άκρως απαραίτητα, μείωσε την ανάγκη για αποθηκευτικό χώρο σε πολύ μεγάλο βαθμό, αλλά επίσης περιόρισε τις πληροφορίες στις άκρως ενδιαφέρουσες για τη μελέτη. Η συλλογή αυτή των σημαντικότερων από τα αποτελέσματα έγινε επίσης με τη βοήθεια κώδικα που υλοποιήθηκε ακριβώς για να φιλτράρει τις σημαντικότερες από τις πληροφορίες βάση ενός κατωφλίου που καθορίζεται από τον χρήστη.

Η αποσυμπίεση όπως επίσης και η συμπίεση των παραγόμενων αρχείων έγινε βάσει του αλγορίθμου συμπίεσης αποσυμπίεσης bzip2. Η διαδικασία αρχίζει με το στάδιο της αποσυμπίεσης των αρχείων και έπειτα το φιλτράρισμα τους, προτού προχωρήσει στο στάδιο της συγχώνευσης τους. Με τον τρόπο αυτό επιτεύχθηκε η απομόνωση των σημαντικότερων αποτελεσμάτων.

Επιπλέον κώδικας υλοποιήθηκε για περεταίρω φιλτράρισμα του αρχείου των αποτελεσμάτων. Για το φιλτράρισμα έγινε θεσμός ανώτερων κατωφλίων για τα μεγάλα αρχεία (περίπου μεγέθους 200MB) , με p-value να ισούται με $1x10^{-4}$ ενώ για τα μικρότερα αρχεία το κατώφλι περιορίστηκε στο $1x10^{-7}$. Η διαδικασία ου φιλτραρίσματος σε αυτό το σημείο της επεξεργασίας των δεδομένων έγινε για να απομονωθούν μόνο οι σημαντικότερες από τις τιμές που παράχθηκαν κατά τη διαδικασία της ανάλυσης των δεδομένων. Επίσης για μελέτη των τιμών γύρω από μια

συγκεκριμένη περιοχή ενδιαφέροντος καθώς επίσης και για αποδοτικότερη χρήση του εργαλείου Spotfire για παρουσίαση και απεικόνιση των αποτελεσμάτων. Για σκοπούς αυτής της μελέτης το μέγεθος των αρχείων καθορίστηκε από το κατώφλι που ορίστηκε σε $1x10^{-7}$. Σημαντικό είναι να αναφερθεί πως το μικρότερο μέγεθος των αρχείων που προέκυψε μετά από την εφαρμογή του κώδικα φιλτραρίσματος με την μικρότερη τιμή κατωφλίου, δεν αλλοίωνε το επίπεδο σημασίας των αποτελεσμάτων, αλλά τα περιόριζε στα σημαντικότερα από αυτά.

Αποτέλεσμα της μετά επεξεργασίας των δεδομένων ήταν το τελικό αρχείο με τα σημαντικότερα αποτελέσματα, που χρησιμοποιήθηκαν στο εργαλείο Spotfire για την γραφική απεικόνιση τους και τη διεξαγωγή συμπερασμάτων. Επιπλέον προστέθηκαν στις πληροφορίες περιγραφικές λεπτομέρειες για κάθε σημείο mRNA και SNP για τα οποία μπορεί να γίνεται και πιο συγκεκριμένο φιλτράρισμα για κάθε διαφορετικό SNP ξεχωριστά. Για παράδειγμα το γονίδιο στο οποίο ανήκει το αντίστοιχο mRNA και SNP.

5.2 Φιλτράρισμα

Όπως και στο στάδιο της προεπεξεργασίας έτσι και εδώ κρίθηκε αναγκαία η εφαρμογή η υλοποίηση και εφαρμογή κώδικα φιλτραρίσματος που περιγράφεται αναλυτικότερα στη συνέχεια.

5.2.1 Φιλτράρισμα για απομόνωση των χρήσιμων πληροφοριών

Μία δεύτερη χρήσιμη εφαρμογή φιλτραρίσματος των δεδομένων κρίθηκε να είναι η περίπτωση φιλτραρίσματος των αποτελεσμάτων που παράχθηκαν μετά την ανάλυση των δεδομένων. Στην περίπτωση αυτή όπως και σε αυτή που αναφέρθηκε στο στάδιο της προ επεξεργασίας των δεδομένων, υλοποιήθηκε κατάλληλος κώδικας που ανταποκρινόταν στις νέες απαιτήσεις.

Αυτού του είδους φιλτράρισμα αποσκοπούσε στο να βοηθήσει τον ερευνητή, να προσαρμόσει τις πληροφορίες ανάλογα με τις απαιτήσεις της έρευνας και το βάρος που χρειάζεται να δώσει κάθε φορά στις πληροφορίες. Επίσης και σε αυτή την περίπτωση η επιλογή των πληροφοριών γίνεται με τη χρήση κάποιου κατωφλίου το οποίο και πάλι ορίζεται από τον χρήστη. Ισχύουν οι ίδιοι περιορισμοί όσο αφορά τις τιμές, δίνοντας μεγαλύτερο βάρος στις μικρότερες τιμές (p-values) και μικρότερο στις μεγαλύτερες τιμές από αυτές.

Το φιλτράρισμα σε αυτό το σημείο έγινε όπως προαναφέρθηκε κατά τη συγχώνευση των αρχείων και αποσκοπούσε στο να διατηρήσει ένα σύνολο από τα σημαντικότερα αποτελέσματα που παράχθηκαν κατά την ανάλυση των δεδομένων.

Το φιλτράρισμα έγινε θέτοντας ανώτερα κατώφλια για τα μεγάλα σε μέγεθος αρχεία (περίπου μεγέθους 200MB) , με p-value να ισούται με $1x10^{-4}$ ενώ για τα μικρότερα αρχεία το κατώφλι περιορίστηκε στο $1x10^{-7}$. Η διαδικασία ου φιλτραρίσματος σε αυτό το σημείο της επεξεργασίας των δεδομένων έγινε για να απομονωθούν μόνο οι σημαντικότερες από τις τιμές που παράχθηκαν κατά τη διαδικασία της ανάλυσης των δεδομένων.

Επιπλέον φιλτράρισμα στο στάδιο αυτό ήταν εφικτό και μετά την συγχώνευση των αποτελεσμάτων σε ένα αρχείο. Για την περίπτωση αυτή υλοποιήθηκε κώδικας που φίλτραρε τα δεδομένα βάσει κάποιου κατωφλίου που και σε αυτή την περίπτωση δινόταν ως παράμετρος από τον χρήστη. Σκοπός του φιλτραρίσματος σε αυτό το στάδιο ήταν ο περιορισμός των δεδομένων στα άκρως σημαντικότερα χωρίς να αλλοιώνει τα τελικά αποτελέσματα που παράχθηκαν και η δημιουργία πολύ μικρότερων αρχείων που θα κάνει αποδοτικότερη την επεξεργασία τους με τη βοήθεια της εφαρμογής Spotifire. Επίσης επιτρέπει τον περιορισμό των δεδομένων γύρω από κάποια συγκεκριμένη περιοχή ενδιαφέροντος που μπορεί να προκύψει σε οποιαδήποτε μελέτη. Αυτό επιτυγχάνεται χωρίς να χαθεί το αρχείο με όλο το σύνολο δεδομένων που παράχθηκαν μετά τη συγχώνευση τους και καλύπτουν ένα ευρύτερο φάσμα.

Με αυτό τον τρόπο επιτεύχθηκε ο περιορισμός των αποτελεσμάτων στα περισσότερο σημαντικά μόνο διευκολύνοντας την διεξαγωγή συμπερασμάτων και την πιο καθαρή απεικόνιση τους.

5.3 Συγχώνευση Αρχείων

Τελικό στάδιο της επεξεργασίας των δεδομένων και της παραγωγής των αποτελεσμάτων αποτέλεσε η συγχώνευση των αρχείων που παράχθηκαν στο στάδιο της ανάλυσης των δεδομένων. Όπως και στην ανάλυση των δεδομένων έτσι και εδώ ο έλεγχος των διαδικασιών και της ανάθεσης διεργασιών στο grid, έγινε μέσω της χρήσης scripts.

Πιο αναλυτικά σε αυτή τη φάση έγινε η συγχώνευση των αποτελεσμάτων σε ένα μεγαλύτερο αρχείο, ώστε να είναι δυνατή η απεικόνιση και παρουσίαση των αποτελεσμάτων με τη χρήση ειδικών εφαρμογών.

Για τη συγχώνευση των αρχείων απαραίτητη, ήταν η χρήση αλγόριθμου αποσυμπίεσης των αρχείων και κώδικα που έγραφε τα δεδομένα κάθε μικρού αρχείου σε ένα νέο ενιαίο και μεγαλύτερο αρχείο που θα περιείχε τα τελικά αποτελέσματα.
Μετά την καταγραφή των πληροφοριών από τα μικρότερα αρχεία στο τελικό, κάθε ένα από αυτά διαγραφόταν από τη μνήμη ώστε ο διαθέσιμος αποθηκευτικός χώρος να παραμένει σταθερός και προς αποφυγή προβλημάτων υπερχείλισης της μνήμης.

Με την καταγραφή των αποτελεσμάτων σημαντική ήταν και η προσθήκη μιας επιπλέον στήλης στο τελικό αρχείο με τα αποτελέσματα, που περιείχε το όνομα του αρχείου από το οποίο προήλθαν και συγκεκριμένα το όνομα αυτό ήταν το όνομα του probeSet. Με τον τρόπο αυτό μπορούσαν να είναι γνωστά τα αποτελέσματα που παράχθηκαν για κάθε διαφορετικό probeSet, ώστε να μπορούν να χρησιμοποιηθούν και ξεχωριστά αν αυτό θεωρηθεί απαραίτητο.

Η συγχώνευση των αποτελεσμάτων σε ένα αρχείο έγινε σε συνδυασμό με φιλτράρισμα των δεδομένων, ώστε να απομονωθούν οι σημαντικότερες από τις πληροφορίες που παράχθηκαν, ως αποτέλεσμα της ανάλυσης των δεδομένων και αποτελούν το επίκεντρο της προσοχής της μελέτης.

Όπως προαναφέρθηκε μετά την διαδικασία παραγωγής του τελικού αρχείου με τα αποτελέσματα, κώδικας επιπλέον φιλτραρίσματος μπορούσε να εφαρμοστεί στις πληροφορίες προσαρμόζοντας τις ανάλογα με τις απαιτήσεις της έρευνας.

5.4 Προσθήκη Πληροφοριών

Η προσθήκη πληροφοριών έγινε για την διευκόλυνση της ανάγνωσης των αποτελεσμάτων, καθώς επίσης και την παραγωγή γραφικών για προβολή των αποτελεσμάτων της ανάλυσης. Η προβολή γίνεται με τέτοιο τρόπο ώστε η πληροφορία που υποδηλώνουν τα αποτελέσματα να παρουσιάζεται οπτικά.

5.4.1 Επιπρόσθετες Πληροφορίες για τα SNPs

Οι πληροφορίες που προστέθηκαν επιπλέον για τα SNPs είναι οι ακόλουθες:

Χρωμόσωμα(Chromosome): αποτελεί έναν ακέραιο αριθμό που υποδείχνει τον αριθμό του χρωμοσώματος πάνω στο οποίο βρίσκεται ένα συγκεκριμένο SNP. Για τα χρωμοσώματα X και Y καθώς επίσης και για τα μιτοχονδριακά SNPs ανατέθηκαν ξεχωριστοί αριθμοί για το καθένα.

Θέση χρωμοσώματος (Chromosome location): αποτελεί ένα ακέραιο αριθμό που υποδείχνει τη θέση στην οποία βρίσκεται το SNP στο συγκεκριμένο χρωμόσωμα. Πιο συγκεκριμένα τον αριθμό του νουκλεοτιδίου πάνω στο χρωμόσωμα όπου εμφανίζεται το SNP, όπως αυτό χαρακτηρίζεται στη βάση dbSNP στο hapmap.

Γονίδιο (Gene): αποτελεί μια συμβολοσειρά που υποδείχνει το όνομα του γονιδίου στο οποίο εμφανίζεται το SNP.

Γονίδιο αποστάσεως ±20 νουκλεοτιδίων (Gene within ±20kb): οι πληροφορίες περιλαμβάνουν τα ονόματα των γονιδίων που απέχουν κατά ± 20 νουκλεοτίδια από το SNP. Σε περίπτωση που κανένα γονίδιο δεν βρίσκεται στο εύρος αυτό των ± 20 νουκλεοτιδίων, τότε το πεδίο αυτό παραμένει κενό. Σε περίπτωση που υπάρχουν περισσότερα από ένα τότε συμπεριλαμβάνονται όλα. Η απόσταση από το SNP μπορεί να καθορίζεται όση θεωρείται απαραίτητη κάθε φορά έχοντας $\pm X$ νουκλεοτίδια όπου το X ορίζει μία παράμετρο.

5.4.2 Επιπρόσθετες Πληροφορίες για τα ProbeSets

Οι πληροφορίες που προστέθηκαν επιπλέον για τα probeSets είναι οι ακόλουθες:

Χρωμόσωμα(Chromosome): αποτελεί έναν ακέραιο αριθμό που υποδείχνει τον αριθμό του χρωμοσώματος από το οποίο προέρχεται ένα συγκεκριμένο probeSet.

Θέση χρωμοσώματος (Chromosome location): αποτελεί δύο ακέραιους αριθμούς που καθορίζουν την αρχή και το τέλος του probeSet στο συγκεκριμένο χρωμόσωμα, δηλαδή από που αρχίζει και που τελειώνει ο συγκεκριμένο αριθμό των νουκλεοτιδίων που αποτελούν το probeSet.

Γονίδιο (Gene): αποτελεί μια συμβολοσειρά που υποδείχνει το όνομα του γονιδίου στο οποίο εμφανίζεται το probeSet.
mRNA ID – HUGO ID: αποτελεί μια συμβολοσειρά που υποδείχνει το όνομα της ακολουθίας mRNA την οποία ανιχνεύτηκε το συγκεκριμένο probeSet.

5.4.3 Επιπρόσθετες Πληροφορίες για τις Πρωτεΐνες

Οι πληροφορίες που προστέθηκαν για τις πρωτεΐνες είναι οι ίδιες με την περίπτωση των probeSets:

Χρωμόσωμα(Chromosome): αποτελεί δύο ακέραιους αριθμούς που καθορίζουν την αρχή και το τέλος του γονιδίου στο συγκεκριμένο χρωμόσωμα του οποίου η έκφραση παράγει την πρωτεΐνη, δηλαδή από που αρχίζει και που τελειώνει, τον συγκεκριμένο αριθμό των νουκλεοτιδίων.

Θέση χρωμοσώματος (Chromosome location): αποτελεί δύο ακέραιους αριθμούς που καθορίζουν την αρχή και το τέλος του γονιδίου που με την έκφραση του παράγεται η συγκεκριμένη πρωτεΐνη, δηλαδή από που αρχίζει και που τελειώνει, τον συγκεκριμένο αριθμό των νουκλεοτιδίων.

Γονίδιο (Gene): αποτελεί μια συμβολοσειρά που υποδείχνει το όνομα του γονιδίου του οποίου η έκφραση παράγει τη συγκεκριμένη πρωτεΐνη.

mRNA ID – HUGO ID: αποτελεί μια συμβολοσειρά που υποδείχνει το όνομα της ακολουθίας mRNA του οποίου η μετάφραση παράγει την πρωτεΐνη.

5.4.4 Επιπρόσθετες Πληροφορίες για τις Κορυφαίες Συσχετίσεις Μεταξύ SNPs και ProbeSets

Στο στάδιο αυτό της μετά επεξεργασίας των δεδομένων, αναλύθηκε σε μεγαλύτερο βάθος η συσχέτιση μεταξύ SNP και probeSets. Συγκεκριμένα η συσχέτιση αυτή χωρίστηκε σε τέσσερις διαφορετικές περιπτώσεις:

- Περίπτωση 1 – Το SNP προηγείται του mRNA
- Περίπτωση 2 – Το SNP έπεται του mRNA
- Περίπτωση 3 – Το SNP και το mRNA δεν έχουν καμία απολύτως συσχέτιση
- Περίπτωση 4 – Το SNP βρίσκεται πάνω στο mRNA

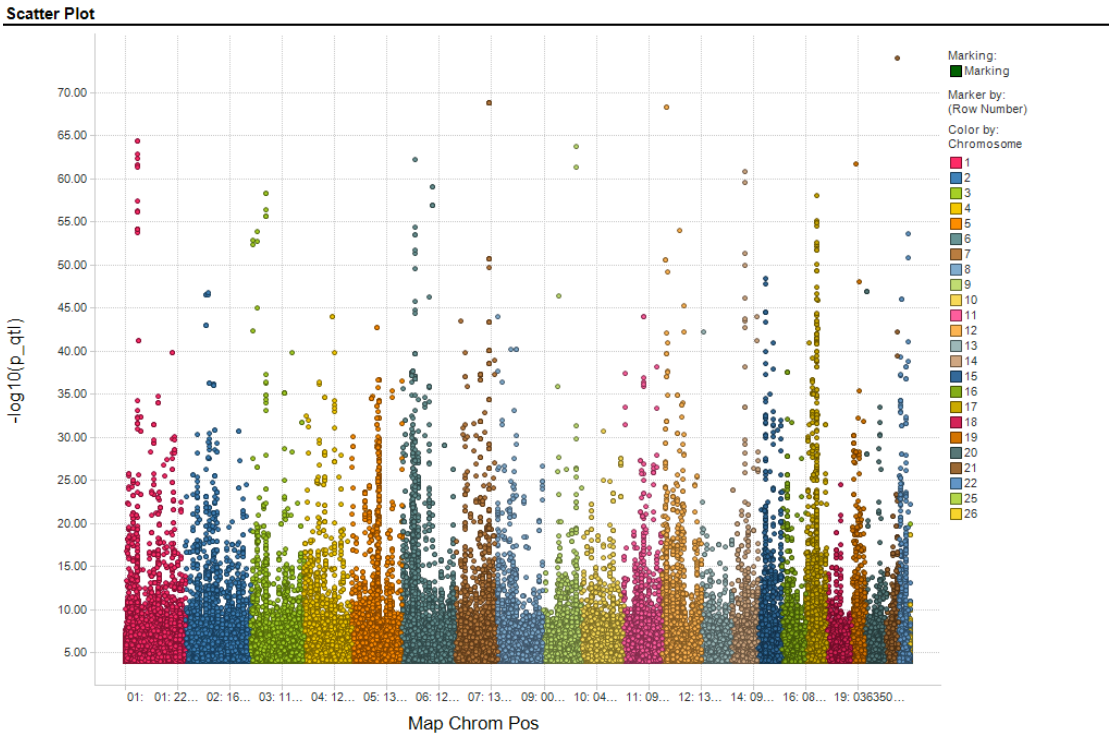
Για τις δύο πρώτες περιπτώσεις η τιμή που χρησιμοποιήθηκε είναι η ελάχιστη απόσταση σε αριθμό νουκλεοτιδίων, μεταξύ του mRNA και του SNP. Για την τρίτη περίπτωση όπου το SNP και το mRNA δεν έχουν καμία απολύτως συσχέτιση η τιμή που χρησιμοποιήθηκε είναι 900000000 ως sentinel value για να μπορεί να είναι εμφανής η διαφορά κατά την προβολή των αποτελεσμάτων. Για την τέταρτη περίπτωση στην οποία το SNP βρίσκεται στην περιοχή του mRNA θεωρούμε πως η ελάχιστη απόσταση μεταξύ τους είναι μηδέν.

Κεφάλαιο 6

Αποτελέσματα

6.1 Αποτελέσματα Κατωφλίου Φιλτραρίσματος 10^{-4}

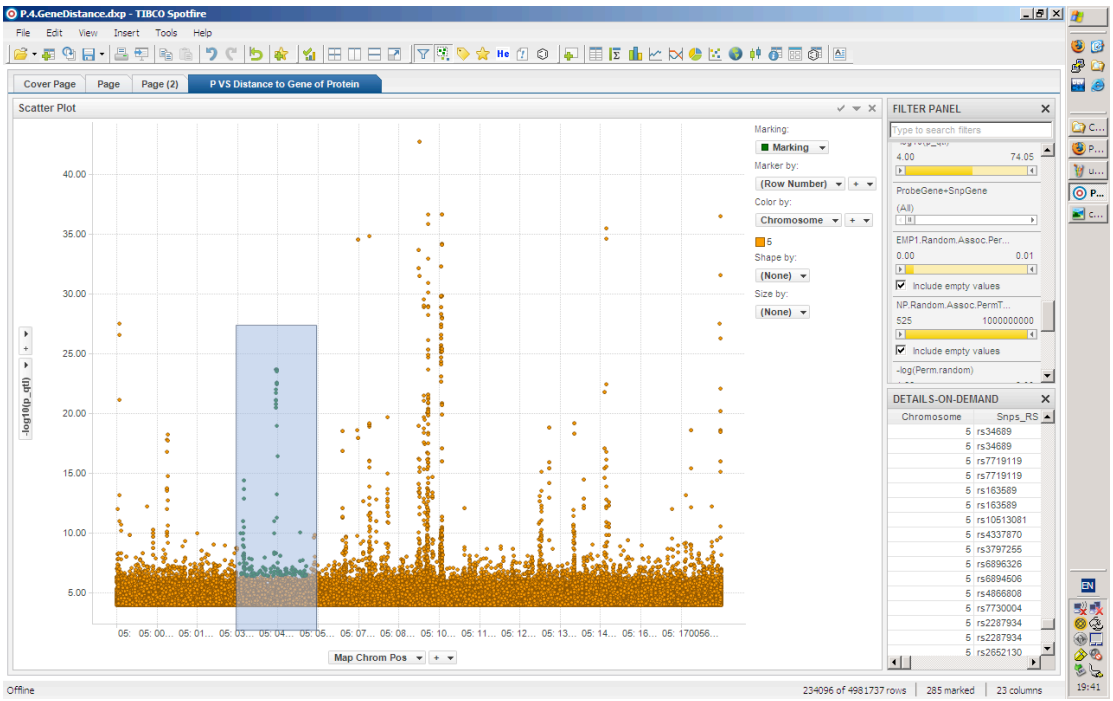
Τα αποτελέσματα που αναλύονται και σχολιάζονται πιο κάτω είναι αυτά που παράχθηκαν κατά στην εφαρμογή της ποσοτικής ανάλυσης (Quantitative Trait Analysis) στα δεδομένα και αργότερα του φιλτραρίσματος τους χρησιμοποιώντας ως κατώφλι την τιμή 10^{-4} για το φιλτράρισμα των πιθανοτήτων σημαντικότητας (p-values). Όπως θα παρατηρηθεί και από τις γραφικές η συγκέντρωση των τιμών πιθανότητας σημαντικότητας (p-values) είναι μεγαλύτερη από αυτή στην περίπτωση χρήσης κατωφλίου 10^{-7} για το φιλτράρισμα. Οι αρνητικοί λογάριθμοι των πιθανοτήτων σημαντικότητας που παρουσιάζονται στους X άξονες των γραφικών παραστάσεων, έχουν ως βάση το 10.



Σχήμα 6.1 Εφαρμογή Κατωφλίου 10^{-4} σε όλα τα αποτελέσματα

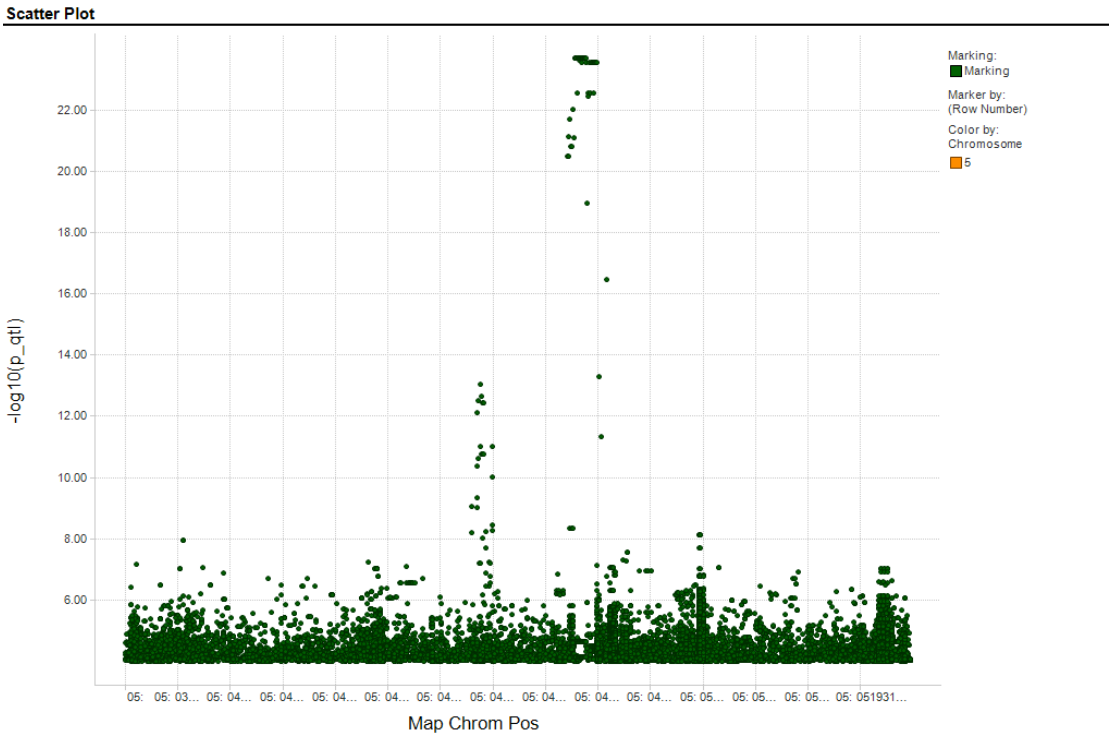
Στο σχήμα 6.1 παρατηρούνται οι τιμές της πιθανότητας σημαντικότητας (p-values) που παράχθηκαν κατά την ποσοτική ανάλυση (QT), για κάθε διαφορετικό από τα 550 χιλιάδες SNPs, σε σχέση με την θέση στην οποία βρίσκονται πάνω σε ένα χρωμόσωμα. Στον άξονα των X βρίσκονται διατεταγμένα τα SNPs με βάση την θέση τους στον γενετικό χάρτη. Δηλαδή για κάθε SNP η θέση του χαρακτηρίζεται από το χρωμόσωμα στο όποιο ανήκει και ακολούθως από τον αριθμό νουκλεοτιδίων από την αρχή του συγκεκριμένου χρωμοσώματος μέχρι το συγκεκριμένο SNP. Για παράδειγμα, το SNP 2:1123 θα βρίσκεται στο 1123^{o} νουκλεοτίδιο από την αρχή του χρωμοσώματος 2. Στον άξονα Ψ διατάσσονται οι αρνητικοί λογάριθμοι των πιθανοτήτων σημαντικότητας για κάθε SNP που μελετήθηκε. Οι πιθανότητες σημαντικότητας είναι προσαρμοσμένες βάσει του λογάριθμου με βάση το 10, αυτό γίνεται γιατί στον τομέα τις γενετικής, οι γραφικές παρουσιάζονται πάντα με αυτό τον τρόπο. Χάρη στην κανονικοποίηση τον αποτελεσμάτων με τον αρνητικό λογάριθμο με βάση το 10, βλέποντας την γραφική είναι δυνατή η εύκολη αναγνώριση του επιπέδου σημαντικότητας, αφού για κάθε αριθμητική μείωση του εκθέτη στην πιθανότητα σημαντικότητας (p-value) σε scientific format, η κανονικοποιημένη τιμή παίρνει μια τιμή ίση με τον εκθέτη της τιμής

πιθανότητας σημαντικότητας (p-value) σε scientific format. Για παράδειγμα p-value e^{-2} δίνει log(0.01)=2. Οι σημαντικότερες από τις τιμές αυτές είναι εκείνες που παρουσιάζονται να έχουν την μεγαλύτερη τιμή στον άξονα των Ψ, με τους αρνητικούς λογάριθμους των πιθανοτήτων σημαντικότητας. Αυτό μπορεί να επεξηγηθεί σύμφωνα με τον ορισμό των πιθανοτήτων σημαντικότητας (p-values), σύμφωνα με τον οποίο οι σημαντικότερες από αυτές, είναι οι μικρότερες που μπορούν να ληφθούν από τα δείγματα. Όσο πιο μικρές είναι οι τιμές τόσο πιο δύσκολο είναι να απορριφθεί η μηδενική υπόθεση. Επομένως αφού στον άξονα των Ψ παρουσιάζονται οι αρνητικοί λογάριθμοι αυτών των τιμών, όπου οι σημαντικότερες από αυτές θα είναι και οι μεγαλύτερες. Τα δείγματα που παρουσιάζονται στη συγκεκριμένη γραφική είναι πυκνότερα στην αντίστοιχη που θα παρουσιαστεί αργότερα για τα αποτελέσματα που παράχθηκαν εφαρμόζοντας κατώφλι 10^{-7}.



Εικόνα 6.1 Επιλογή Συγκεκριμένου Τμήματος Χρωμοσώματος

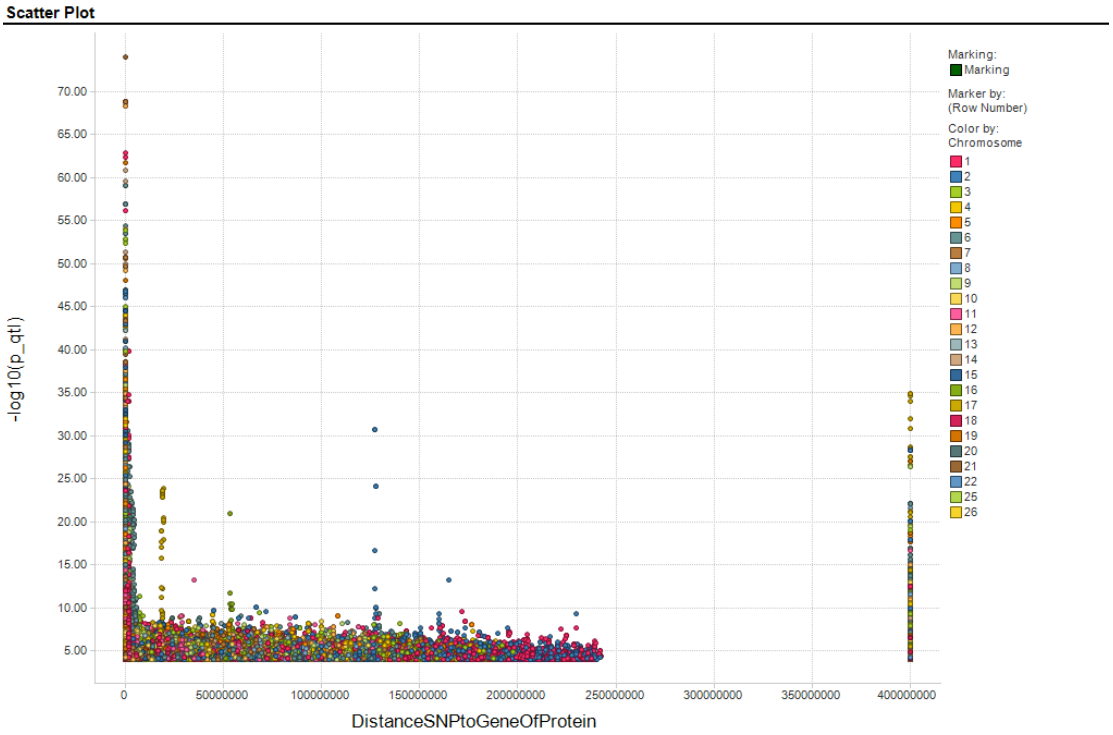
Πιο πάνω παρουσιάζεται μια από τις μεθόδους με τις οποίες γίνεται η επιλογή κάποιας συγκεκριμένης τοποθεσίας σε ένα χρωμόσωμα, που επιθυμείται να μελετηθεί με μεγαλύτερη λεπτομέρια.



Σχήμα 6.2 Αποτελέσματα συγκεκριμένου τμήματος χρωμοσώματος - Κατώφλι 10^{-4}

Στην γραφική παράσταση του σχήματος 6.2 παρατηρούνται οι τιμές της πιθανότητας σημαντικότητας (p-values) της συγκεκριμένης περιοχής που παρουσιάζεται στην εικόνα 6.1. Αποτελεί μεγέθυνση των αποτελεσμάτων και περιορισμό τους σε ένα συγκεκριμένο χρωμόσωμα και πιο συγκεκριμένα στο χρωμόσωμα 5. Στον άξονα των Χ βρίσκονται διατεταγμένες οι διάφορες περιοχές του γονιδίου που έχει επιλεχθεί, ενώ στον άξονα των Ψ διατάσσονται οι αρνητικοί λογάριθμοι των πιθανοτήτων σημαντικότητας για κάθε SNP που μελετήθηκε. Οι σημαντικότερες από τις τιμές αυτές, είναι εκείνες που παρουσιάζονται να έχουν την μεγαλύτερη τιμή στον άξονα των Ψ με τους αρνητικούς λογάριθμους των πιθανοτήτων σημαντικότητας. Αυτό μπορεί να επεξηγηθεί σύμφωνα με τον ορισμό των πιθανοτήτων σημαντικότητας (p-values), που κατά τον οποίο οι σημαντικότερες από αυτές είναι οι μικρότερες που μπορούν να ληφθούν από τα δείγματα, καθώς όσο πιο μικρές είναι οι τιμές τόσο πιο δύσκολο είναι να απορριφθεί η μηδενική υπόθεση. Επομένως αφού στον άξονα των Ψ παρουσιάζονται οι αρνητικοί λογάριθμοι αυτών των τιμών, οι σημαντικότερες από αυτές θα είναι και οι μεγαλύτερες. Επίσης εδώ μπορεί να παρατηρηθεί η μεγαλύτερη συγκέντρωση τιμών πιθανοτήτων σημαντικότητας (pvalues) για τα διάφορα SNPs που

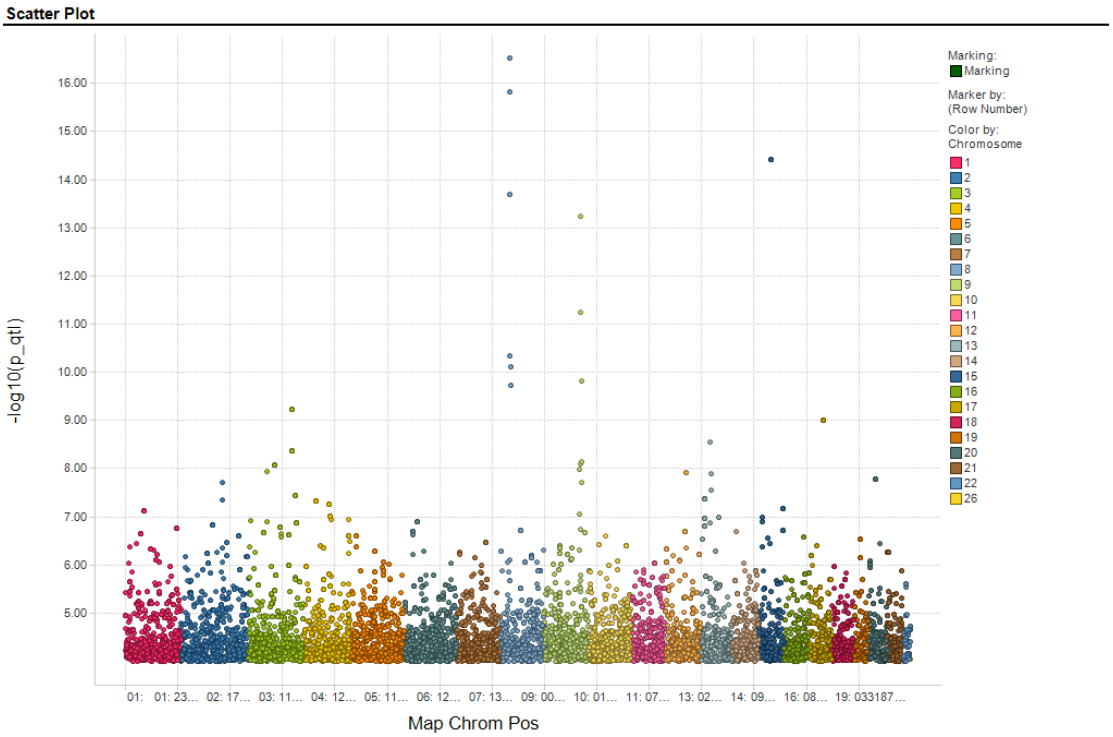
χρησιμοποιήθηκαν στην ανάλυση, λόγω της μεγαλύτερης τιμής του κατωφλίου που δόθηκε σαν είσοδος, κατά τη διαδικασία του φιλτραρίσματος των αποτελεσμάτων.



Σχήμα 6.3 Ελάχιστη απόσταση του SNP μεταξύ μετάγραφου mRNA

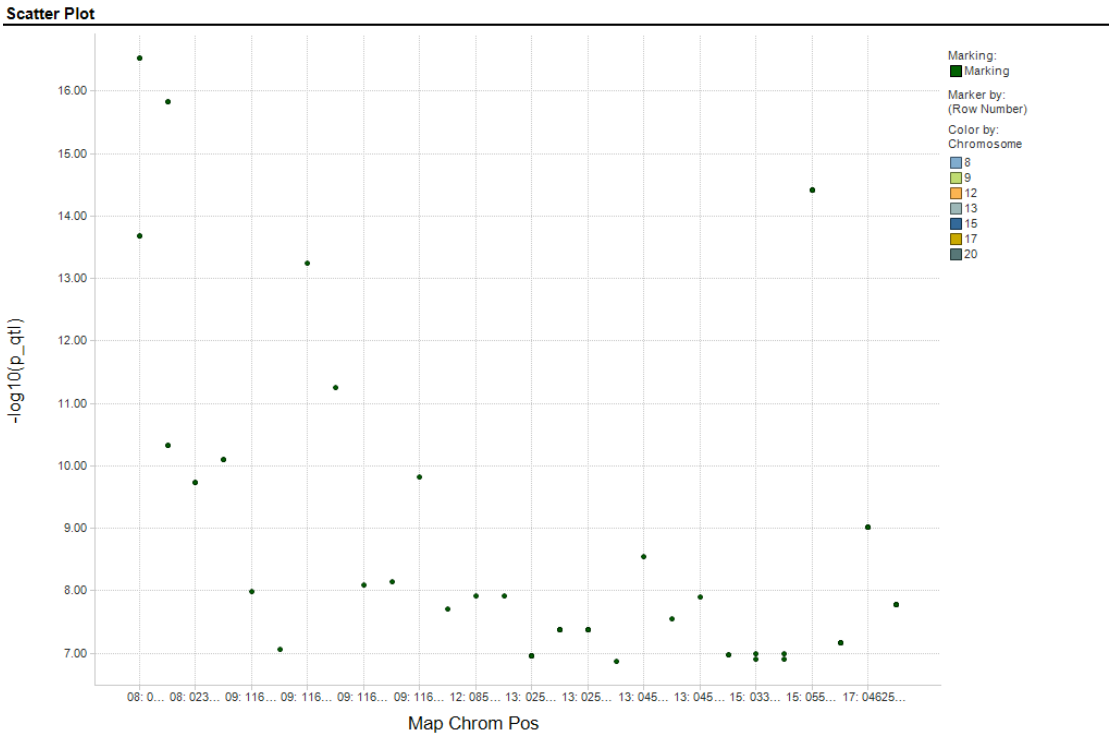
Στην γραφική παράσταση του σχήματος 6.3 παρουσιάζεται η ελαχίστη απόσταση του SNP από τη γενετική θέση από την οποία μεταγράφεται το μόριο του mRNA. Για κάθε αποτέλεσμα στο όποιο τόσο το mRNA όσο και το SNP βρίσκονται στο ίδιο χρωμόσωμα και για όλες τις περιπτώσεις όπου τα συσχετισμένα SNP και mRNA βρίσκονται σε διαφορετικά χρωμοσώματα, δίνεται ο σημαφόρος με τιμή 400000000. Στον άξονα Ψ όπως και στα σχήματα 6.1 και 6.2, διατάσσονται οι αρνητικοί λογάριθμοι των πιθανοτήτων σημαντικότητας (p-value) για κάθε SNP που μελετήθηκε. Παρατηρείται πως η απόσταση για τα περισσότερα από τα SNPs τείνει να μηδενιστεί, δηλαδή έχουν μεγαλύτερη πιθανότητα να είναι γειτονικά με τα μόρια mRNA και να βρίσκονται πάνω στο ίδιο χρωμόσωμα. Αυτό που συμβαίνει όταν η έκφραση του μορίου του mRNA συσχετίζεται από κάποιο SNP που βρίσκεται στο ίδιο χρωμόσωμα και σε μικρή απόσταση από αυτό, στην βιολογία ονομάζεται cis και αποτελεί μια μικρή

αλληλουχία DNA, πάνω στην οποία προσδένονται οι παράγοντες ενεργοποίησης, για να υποβοηθήσουν την έναρξη ή τη λήξη της διαδικασίας της μετάφρασης του σε κάποιο προϊόν πρωτεΐνης. Η γνώση για τον τύπο της συσχέτισης μεταξύ SNP και mRNA αν είναι δηλαδή cis ή trans, είναι εξαιρετικής σημασίας για τη διεξαγωγή νέων πειραμάτων. Ο λόγος της μεγάλης σημασίας της γνώσης αυτής είναι γιατί μπορεί να χρησιμοποιηθεί στη διεξαγωγή βιολογικών πειραμάτων που θα μπορέσουν να μελετήσουν σε βάθος τη λειτουργία των βιολογικών μηχανισμών. Τα SNPs που παρουσιάζονται να απέχουν κατά τη μεγαλύτερη απόσταση από κάποιο μόριο mRNA και που ουσιαστικά δεν βρίσκονται πάνω στο ίδιο χρωμόσωμα είναι πιο πιθανόν να αποτελούν παράγοντες ενεργοποίησης (trans). Η έκφραση του μορίου του mRNA μπορεί να επηρεάζεται έμμεσα από κάποιο SNP που δεν βρίσκεται στο ίδιο χρωμόσωμα με αυτό ή ακόμη και να μην επηρεάζεται καθόλου. Από την γραφική παράσταση μπορεί επίσης να παρατηρηθεί πως η μέγιστη απόσταση των SNPs από κάποια περιοχή έκφρασης ενός μορίου mRNA είναι 400000000. Η συγκέντρωση των τιμών πιθανοτήτων σημαντικότητας είναι εμφανές πως είναι πολύ μεγαλύτερη, λόγω και του μεγαλύτερου όγκου αποτελεσμάτων που χρησιμοποιήθηκαν για την παραγωγή των γραφικών σε αυτό το στάδιο.



Σχήμα 6.4 Αποτελέσματα μετά από εφαρμογή φίλτρου στην οικογένεια γονιδίων TNF

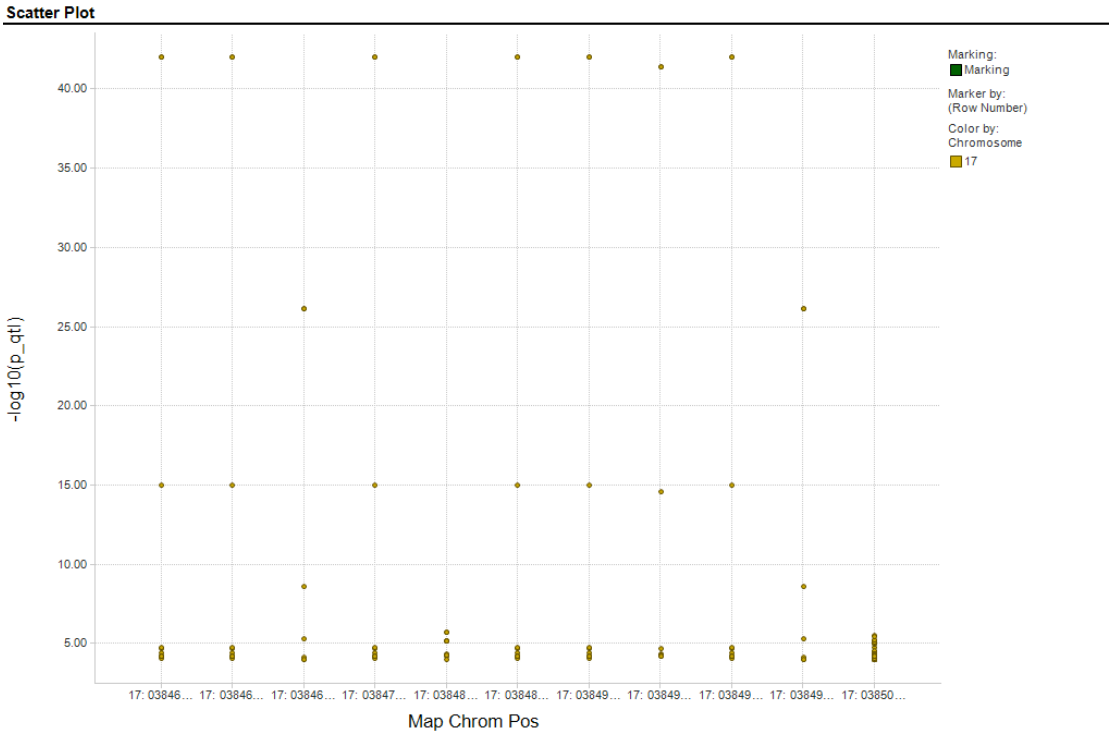
Η γραφική παράσταση του σχήματος 6.4 παρουσιάζει τα αποτελέσματα μετά από την εφαρμογή φίλτρου στην οικογένεια γονιδίων TNF . Στον άξονα των X διατάσσονται οι θέσεις των SNPs πάνω στα διάφορα χρωμοσώματα, ενώ στον άξονα των Ψ τους αρνητικούς λογάριθμους για τις τιμές των πιθανοτήτων σημαντικότητας. Στη δεξιά μεριά του σχήματος 6.4 φαίνονται αριθμημένα τα διάφορα χρωμοσώματα για τα οποία υπάρχουν αποτελέσματα για το συγκεκριμένο γονίδιο. Επιπλέον όπως αναφέρθηκε και νωρίτερα οι σημαντικότερες από τις τιμές που παρουσιάζονται στη γραφική είναι αυτές με την μεγαλύτερη τιμή στον άξονα των Ψ. Δηλαδή οι τιμές με τον μεγαλύτερο αρνητικό λογάριθμο είναι και οι σημαντικότερες βάσει του ορισμού των πιθανοτήτων σημαντικότητας. Τα αποτελέσματα που εμφανίζονται στο συγκεκριμένο σχήμα είναι αποτελέσματα που παράχθηκαν και για τα 26 διαφορετικά χρωμοσώματα του ανθρώπινου γονιδιώματος.



Σχήμα 6.5 Αποτελέσματα της οικογένειας του γονιδίου TNF μετά από την εφαρμογή φιλτραρίσματος

Η γραφική παράσταση του σχήματος 6.5 παρουσιάζει τα αποτελέσματα μετά από την εφαρμογή φίλτρου στην οικογένεια γονιδίου TNF . Στον άξονα των X διατάσσονται οι θέσεις που κατέχουν τα SNPs πάνω στα διάφορα χρωμοσώματα, ενώ στον άξονα των Ψ τις τιμές των πιθανοτήτων σημαντικότητας προσαρμοσμένες βάσει του αρνητικού λογάριθμου με βάση το 10. Στη δεξιά μεριά της γραφικής στο πάνω μέρος, βρίσκονται αριθμημένα τα διάφορα χρωμοσώματα για τα οποία υπάρχουν αποτελέσματα σε αυτή τη γραφική, για το συγκεκριμένο γονίδιο. Από τις τιμές των αρνητικών λογαρίθμων των πιθανοτήτων σημαντικότητας που παρατηρούνται στο σχήμα 6.5, μπορεί να εξαχθεί το συμπέρασμα πως είναι στατιστικώς σημαντικά τα συγκεκριμένα αποτελέσματα, καθώς οποιαδήποτε τιμή πιθανότητας σημαντικότητας που είναι μικρότερη από 0.05 αποτελεί μία στατιστικά σημαντική τιμή, βάσει του ορισμού των πιθανοτήτων σημαντικότητας που θέτει σημαντική οποιαδήποτε τιμή είναι μικρότερη από το 0.05. Επιπλέον παρατηρείται από τη γραφική πως τα περισσότερα από τα αποτελέσματα είναι αρκετά μικρότερα από το 0.05 με τον αρνητικό λογάριθμο να ξεκινά από το 7 και πάνω. Επομένως, οι παρατηρήσεις οδηγούν στο συμπέρασμα πως οι τιμές των πιθανοτήτων

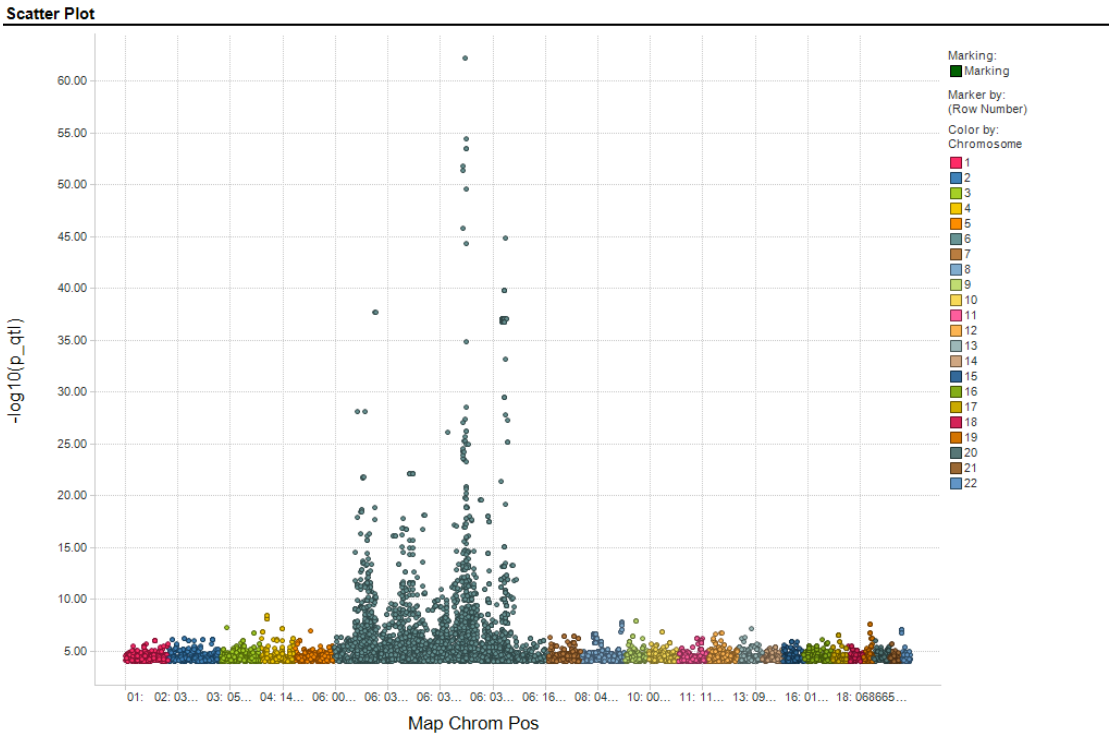
σημαντικότητας (p-values) για την οικογένεια γονιδίου TNF, είναι εξαιρετικής σημασίας με τα SNPs να σχετίζονται άμεσα με τις περιοχές έκφρασης του mRNA πάνω στο γονίδιο. Η γραφική αυτή παράσταση παράχθηκε μετά από την μεγέθυνση των αποτελεσμάτων επιλογής της συγκεκριμένης οικογένειας του γονιδίου TNF.



Σχήμα 6.6 Αποτελέσματα της οικογένειας του γονιδίου BRCA1 μετά από την εφαρμογή φιλτραρίσματος

Η γραφική παράσταση του σχήματος 6.5 παρουσιάζει τα αποτελέσματα μετά από την εφαρμογή φίλτρου στην οικογένεια γονιδίου BRCA1. Το φίλτρο που εφαρμόστηκε σε αυτή την περίπτωση είναι διαφορετικό σε σύγκριση με το φιλτράρισμα που εφαρμόστηκε πάνω στο γονίδιο TNF (σχήμα 6.4). Οι άξονες εντούτοις παρουσιάζουν τις ίδιες μονάδες. Ο διαφορά είναι ότι το φίλτρο προσαρμόζεται σε κάθε διαφορετική περίπτωση ξεχωριστά. Στον άξονα των X διατάσσονται οι θέσεις των SNPs πάνω στα διαφορετικά χρωμοσώματα, ενώ στον άξονα των Ψ τις τιμές των πιθανοτήτων σημαντικότητας. Στο πλάι φαίνονται αριθμημένα τα διάφορα χρωμοσώματα για τα οποία υπάρχουν αποτελέσματα για το συγκεκριμένο γονίδιο. Επιπλέον σε αυτή τη

γραφική όπως και σε όλες τις γραφικές των σχημάτων που προηγούνται, οι σημαντικότερες από τις τιμές που παρουσιάζονται στον άξονα των Ψ είναι αυτές με την μεγαλύτερη τιμή. Δηλαδή οι τιμές με τον μεγαλύτερο αρνητικό λογάριθμο είναι και οι σημαντικότερες βάσει του ορισμού των πιθανοτήτων σημαντικότητας. Μπορεί να παρατηρηθεί ότι τα αποτελέσματα για το συγκεκριμένο γονίδιο αναφέρονται μόνο σε ένα μόνο χρωμόσωμα, το χρωμόσωμα 17. Δηλαδή τα SNPs που βρίσκονται στο συγκεκριμένο γονίδιο είναι όλα συγκεντρωμένα σε ένα και μόνο χρωμόσωμα, το χρωμόσωμα 17. Επιπλέον οι λογαριθμικές τιμές των πιθανοτήτων σημαντικότητας (p-value) φαίνονται να είναι υψίστης σημασίας σε σχέση με προηγούμενα αποτελέσματα καθώς είναι και συγκριτικά μεγαλύτερες. Λαμβάνοντας υπόψη και τον ορισμό των πιθανοτήτων σημαντικότητας όπως έχει ορισθεί στη στατιστική, θέτει ως σημαντικές τις τιμές αυτές που είναι μικρότερες από το 0.05. Στο γονίδιο αυτό παρατηρείται πως οι τιμές για τις πιθανότητες σημαντικότητας δεν είναι εξαιρετικής σημασίας καθώς η μεγαλύτερη συγκέντρωση των τιμών αυτών εμφανίζονται κοντά στην τιμή 5 του άξονα Χ. Αναφερόμενοι στον ορισμό των πιθανοτήτων σημαντικότητας (p-value) που θέτει ώς σημαντικές τις τιμές που είναι μικρότερες από 0.05, μπορεί να παρατηρηθεί πως σε αυτή την περίπτωση τα αποτελέσματα δεν είναι τόσο μεγάλης σημασίας όπως αυτά του σχήματος 6.5, εντούτοις υπάρχουν κάποιες τιμές από αυτές που φαίνονται να είναι τεράστιας σημασίας καθώς φτάνουν και μέχρι την τιμή 40 του αρνητικού λογάριθμου στον άξονα των Χ και κάποιες άλλες μέχρι το 16. Η μεγαλύτερη συγκέντρωση τιμών όμως παρατηρείται κοντά στο 5 που δεν αποτελούν σημαντικές τιμές καθώς βρίσκονται πάνω στο όριο που θέτει ο ορισμός των τιμών των πιθανοτήτων σημαντικότητας .

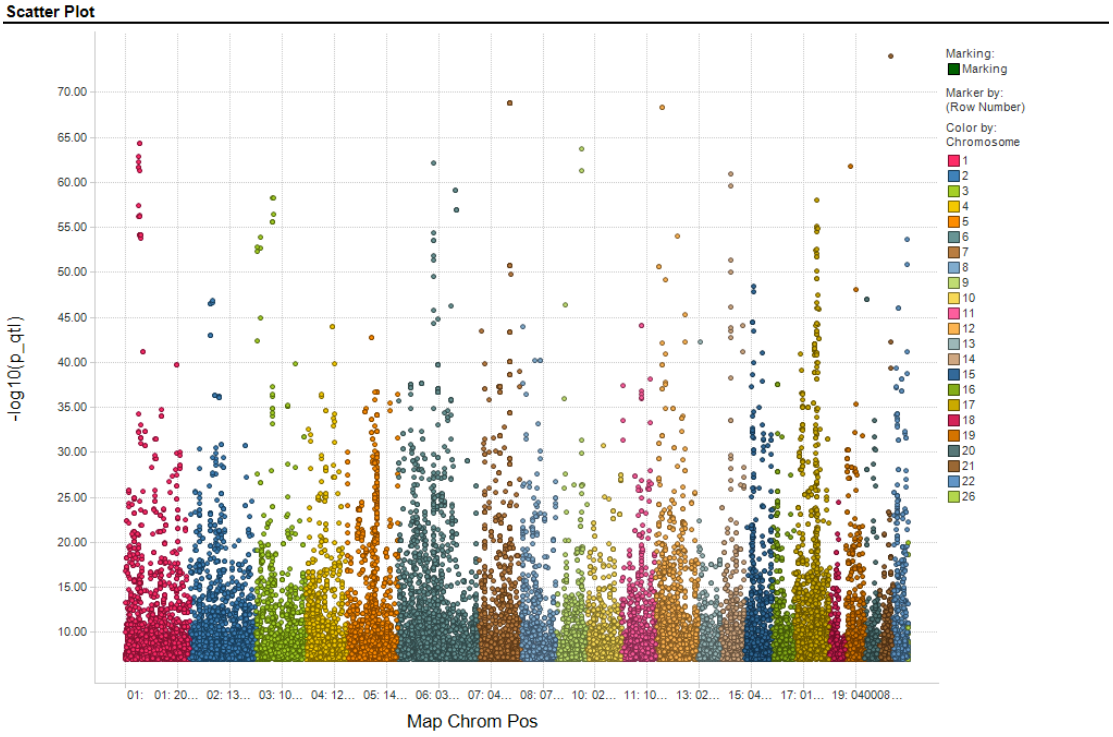


Σχήμα 6.7 Αποτελέσματα της οικογένειας του γονιδίου HLA μετά από την εφαρμογή φιλτραρίσματος

Το σχήμα 6.7 παρουσιάζει τα αποτελέσματα μετά από την εφαρμογή φίλτρου στην οικογένεια γονιδίων HLA. Στον άξονα των X διατάσσονται οι θέσεις των SNPs πάνω στα διάφορα χρωμοσώματα, ενώ στον άξονα των Ψ οι τιμές των πιθανοτήτων σημαντικότητας προσαρμοσμένες βάσει του αρνητικού λογάριθμου με βάση το 10. Στο πλάι αριθμημένα φαίνονται τα διάφορα χρωμοσώματα για τα οποία υπάρχουν αποτελέσματα σε αυτή τη γραφική για το συγκεκριμένο γονίδιο. Όπως είναι ήδη γνωστό και από τον ορισμό των πιθανοτήτων σημαντικότητας οι σημαντικότερες από τις τιμές, από αυτές που παρουσιάζονται στη γραφική, είναι αυτές με την μεγαλύτερη τιμή στον άξονα των Ψ. Μπορεί να παρατηρηθεί πως σε σχέση με τα αποτελέσματα που παρουσιάστηκαν για τις υπόλοιπες οικογένειες γονιδίων, οι τιμές των πιθανοτήτων σημαντικότητας, είναι οι μεγαλύτερες που έχουν παρατηρηθεί και επομένως μπορούν να ληφθούν υπόψη και ως υψίστης σημασίας με τις μικρότερες τιμές των πιθανοτήτων σημαντικότητας (p-values).

6.2 Αποτελέσματα Κατωφλίου Φιλτραρίσματος 10^{-7}

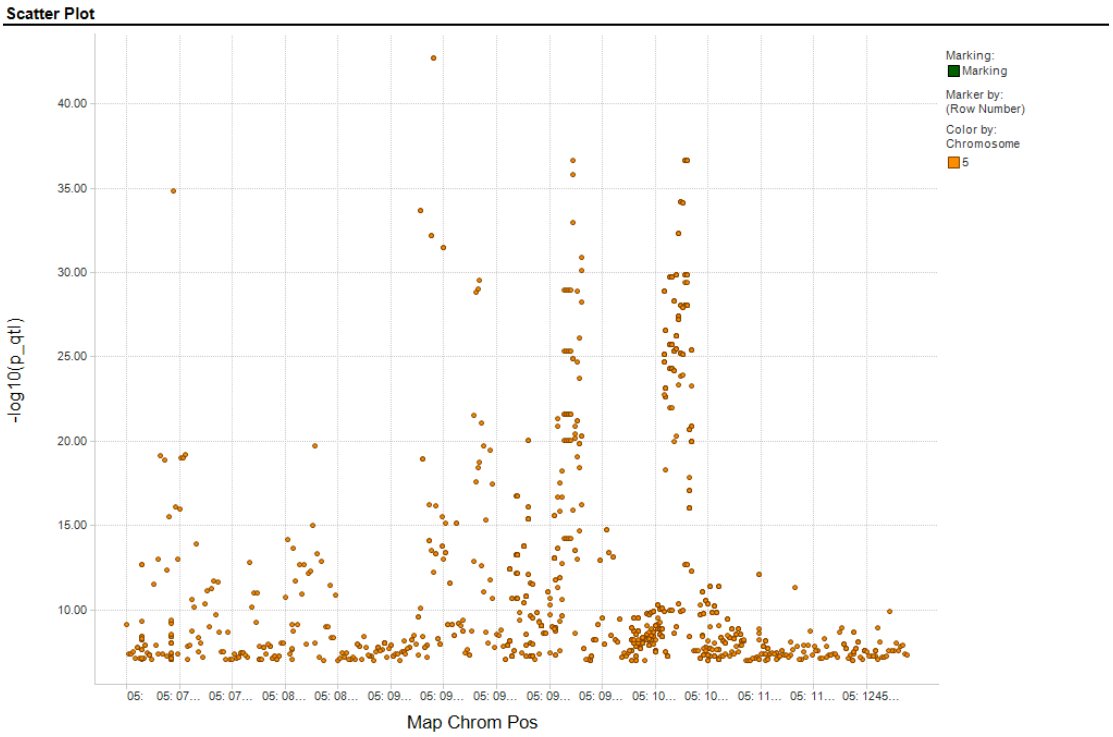
Τα πιο κάτω αποτελέσματα που αναλύονται και σχολιάζονται είναι αυτά που παράχθηκαν κατά στην εφαρμογή της ποσοτικής ανάλυσης (Quantitative Trait Analysis) στα δεδομένα και αργότερα του φιλτραρίσματος τους χρησιμοποιώντας ως κατώφλι την τιμή 10^{-7} για το φιλτράρισμα των πιθανοτήτων σημαντικότητας (p-values). Οι αρνητικοί λογάριθμοι των πιθανοτήτων σημαντικότητας που παρουσιάζονται στους Χ άξονες των γραφικών παραστάσεων, έχουν ως βάση το 10. Η εφαρμογή του κατωφλίου αυτού περιορίζει περεταίρω τις τιμές παρατήρησης αυτού του υποκεφαλαίου στις πιο σημαντικές. Αυτό γίνεται γιατί εφαρμόζεται ένα συγκριτικά πολύ μικρότερο κατώφλι από το 10^{-4} τα αποτελέσματα του οποίου μελετήθηκαν στο υποκεφάλαιο 6.1.



Σχήμα 6.8 Εφαρμογή κατωφλίου 10^{-7} σε όλα τα αποτελέσματα

Στην γραφική παράσταση του σχήματος 6.8, παρατηρούνται οι τιμές των πιθανοτήτων σημαντικότητας (p-values) μετά από την εφαρμογή της ποσοτικής ανάλυσης (QT) στα δεδομένα. Οι τιμές που παρουσιάζονται σε αυτή τη γραφική αναφέρονται στα 550 χιλιάδες διαφορετικά SNPs που χρησιμοποιήθηκαν για την ανάλυση, και σχετίζονται με την θέση στην οποία βρίσκονται πάνω σε ένα χρωμόσωμα. Στον άξονα των Χ βρίσκονται διατεταγμένα τα διάφορα χρωμοσώματα του ανθρώπινου οργανισμού, ενώ στον άξονα των Ψ διατάσσονται οι αρνητικοί λογάριθμοι των πιθανοτήτων σημαντικότητας για κάθε SNP που μελετήθηκε. Οι σημαντικότερες από τις τιμές αυτές είναι εκείνες που παρουσιάζονται να έχουν την μεγαλύτερη τιμή στον άξονα των Ψ, στον οποία φαίνονται οι αρνητικοί λογάριθοι των πιθανοτήτων σημαντικότητας. Αυτό μπορεί να επεξηγηθεί σύμφωνα με τον ορισμό των πιθανοτήτων σημαντικότητας (p-values), που κατά τον οποίο οι σημαντικότερες από αυτές είναι οι μικρότερες που μπορούν να ληφθούν από τα δείγματα, καθώς όσο πιο μικρές είναι οι τιμές τόσο πιο δύσκολο είναι να απορριφθεί η μηδενική υπόθεση. Σύμφωνα με τον ορισμό μία πιθανότητα σημαντικότητας μπορεί να θεωρηθεί ως σημαντική εάν αυτή είναι μικρότερη από το κατώφλι του 0.05. Αφού στον άξονα των Ψ παρουσιάζονται οι

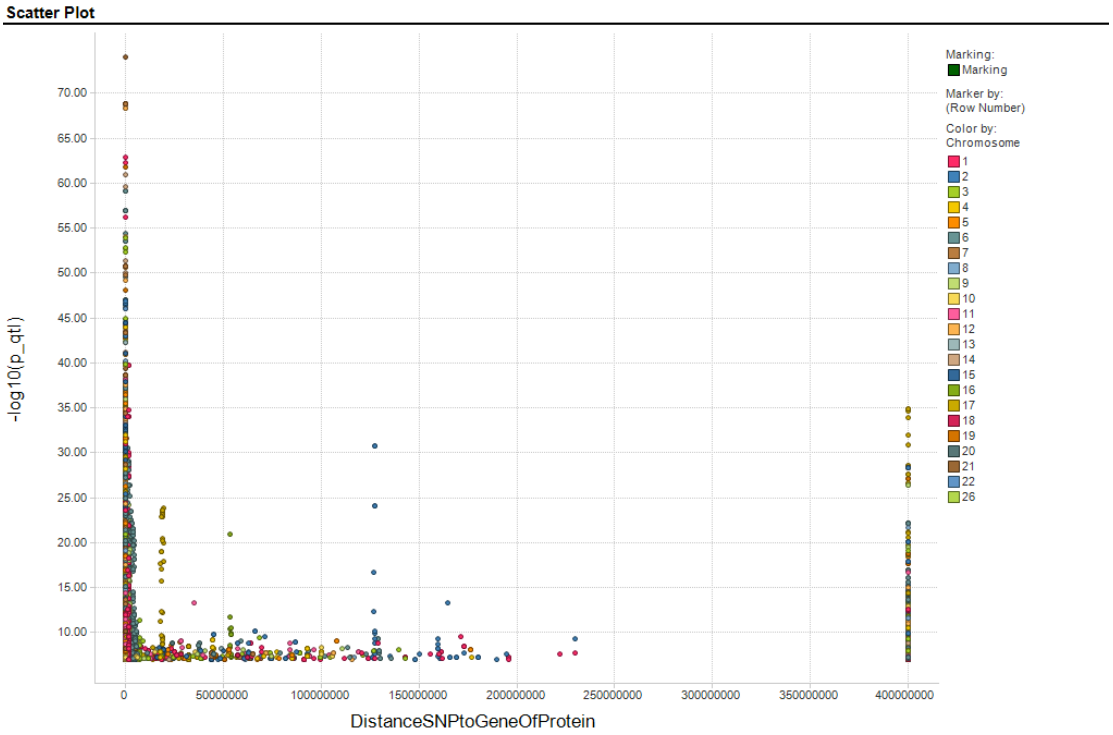
αρνητικοί λογάριθμοι αυτών των τιμών, οι σημαντικότερες από αυτές θα είναι και οι μεγαλύτερες.



Σχήμα 6.9 Αποτελέσματα συγκεκριμένου τμήματος χρωμοσώματος - Κατώφλι 10^{-7}

Στην γραφική παράσταση του σχήματος 6.9, παρατηρούνται οι τιμές της πιθανότητας σημαντικότητας (p-values) για τα διάφορα SNPs σε σχέση με τη θέση που κατέχουν σε κάποιο χρωμόσωμα. Αποτελεί μεγέθυνση των αποτελεσμάτων και περιορισμό τους σε ένα συγκεκριμένο χρωμόσωμα και πιο συγκεκριμένα στο χρωμόσωμα 6. Στον άξονα των Χ βρίσκονται διατεταγμένες οι διάφορες περιοχές του γονιδίου που έχει επιλεχθεί, ενώ στον άξονα των Ψ διατάσσονται οι αρνητικοί λογάριθμοι των πιθανοτήτων σημαντικότητας για κάθε SNP που μελετήθηκε. Οι σημαντικότερες από τις τιμές αυτές είναι εκείνες που παρουσιάζονται να έχουν την μεγαλύτερη τιμή στον άξονα των Ψ με τους αρνητικούς λογάριθμους των πιθανοτήτων σημαντικότητας. Παρατηρώντας την γραφική μπορεί να γίνει αντιληπτό, πως τα αποτελέσματα είναι εξαιρετικής σημασίας καθώς οι τιμές που παίρνουν σε σχέση με τον άξονα των Χ (αρνητικός λογάριθμος των πιθανοτήτων σημαντικότητας), ξεκινούν από την τιμή 15. Άν ανατρέξουμε στον ορισμό των πιθανοτήτων σημαντικότητας που θέτει ώς σημαντική κάθε μία από αυτές τις τιμές

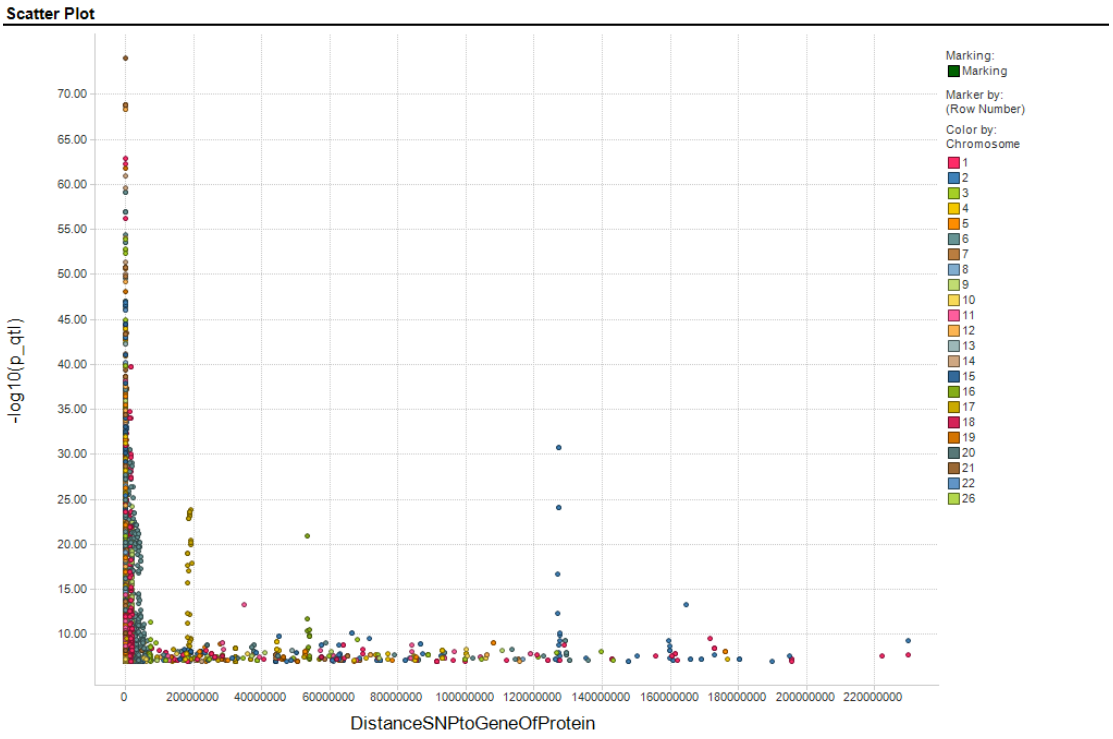
που είναι μικρότερη από το 0.05 (5 βάσει του αρνητικού λογαρίθμου), παρατηρείται πως για αυτό το χρωμόσωμα οι τιμές που παράχθηκαν είναι τεράστιας σημασίας καθώς απέχουν κατα πολύ από το κατώφλι του ορισμού των πιθανοτήτων σημαντικότητας. Επομένως τα SNPs των οποίων η τοποθεσία βρίσκεται πάνω στο χρωμόσωμα 6 (HLA) είναι τεράστιας σημασίας και είναι πολύ πιθανόν να επηρεάζουν την έκφραση κάποιου μορίου mRNA που μεταγράφεται από κάποια περιοχή αυτού του χρωμοσώματος.



Σχήμα 6.10 Ελάχιστη απόσταση των SNPs από τα μετάγραφα mRNA

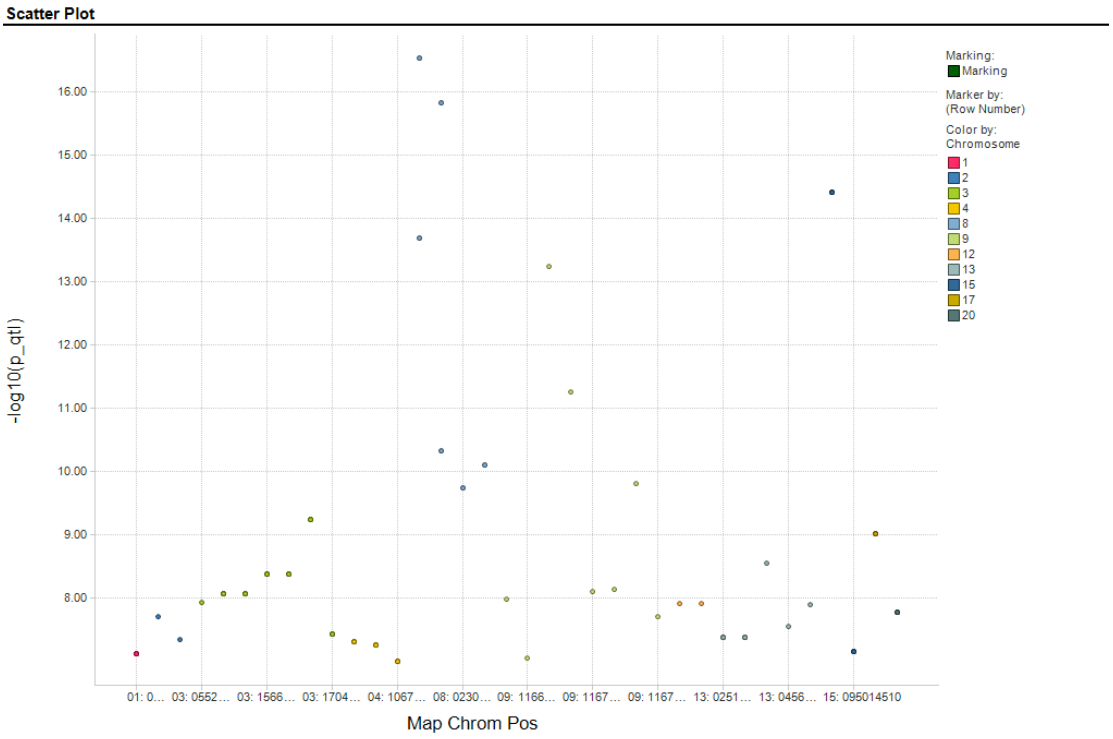
Στο σχήμα 6.10, παρουσιάζεται η απόσταση των SNPs από τη γενετική θέση στην οποία βρίσκονται τα μόρια mRNA πάνω στα χρωμοσώματα. Στον άξονα των X βρίσκονται διατεταγμένες οι διάφορες αποστάσεις των SNPs από τα μόρια mRNA, ενώ στον άξονα των Ψ διατάσσονται οι αρνητικοί λογάριθμοι των πιθανοτήτων σημαντικότητας (p-value) για κάθε SNP που μελετήθηκε. Παρατηρείται πως η απόσταση για τα περισσότερα από τα SNPs τείνει να μηδενιστεί, δηλαδή έχουν μεγαλύτερη πιθανότητα να είναι γειτονικά με τα μόρια mRNA και να βρίσκονται πάνω στο ίδιο χρωμόσωμα. Αυτό σημαίνει πως η έκφραση των μορίων του mRNA είναι πολύ πιθανό να επηρεάζεται άμεσα από κάποιο SNP, που κατέχει όσο το δυνατό κοντινότερη

θέση με αυτό, πάνω στο ίδιο χρωμόσωμα. Οι τιμές αυτές που υποδεικνύουν SNPs που κατέχουν γειτονική θέση στο χρωμόσωμα με κάποιο mRNA αποτελούν στοιχεία ενεργοποίησης της έκφρασης του συγκεκριμένου μορίου mRNA που εμφανίζεται να είνι γειτονικό με αυτό (cis). SNPs που παρουσιάζονται να απέχουν κατά τη μεγαλύτερη απόσταση από κάποιο μόριο mRNA και που ουσιαστικά δεν βρίσκονται πάνω στο ίδιο χρωμόσωμα, είναι πιθανόν να αποτελούν παράγοντες ενεργοποίησης της έκφρασης των μορίων mRNA (trans) ή σε περίπτωση που η τιμή της πιθανότητας σημαντικότητας τους είναι μεγαλύτερη από το 0.05 που θέτει ο ορισμός των πιθανοτήτων σημαντικότητας, είναι πιθανόν να μην επηρεάζουν καθόλου την έκφραση του mRNA. Από την γραφική παράσταση μπορούμε να παρατηρήσουμε πως η μέγιστη αυτή απόσταση των SNPs από κάποιο μόριο mRNA είναι 400000000. Για αυτή την περίπτωση SNPs μπορεί να γίνει η υπόθεση πως κατα μεγάλη πιθανότητα αποτελούν παράγοντες ενεργοποίησης (trans), ενώ όσο πλησιάζουν οι τιμές το 0 η πιθανότητα να αποτελούν παράγοντες ενεργοποίησης τα SNPs μικραίνει ενώ αυξάνεται η πιθανότητα τα αποτελούν στοιχεία cis. Τα cis στοιχεία φαίνονται στις τιμές που τείνουν να πλησιάσουν πιο κοντά στο 0 σε σχέση με τον άξονα των Ψ που καθορίζει την απόσταση των SNPs πάνω στο χρωμόσωμα σε σχέση με τη θέση από την οποία μεταγράφεται το mRNA. Για τις ενδιάμεσες τιμές αυτές που παρατηρείται πως απομακρύνονται από το 0 και έχουν μικρότερη απόσταση από 400000000, δεν μπορεί να προσδιοριστεί με σιγουριά αν αυτές αποτελούν περιπτώσεις cis ή trans και οι τιμές των αρνητικών λογαρίθμων των πιθανοτήτων σημαντικότητας παρατηρείται να μειώνονται όπως επίσης και η σημασία των αποτελεσμάτων αυτών.



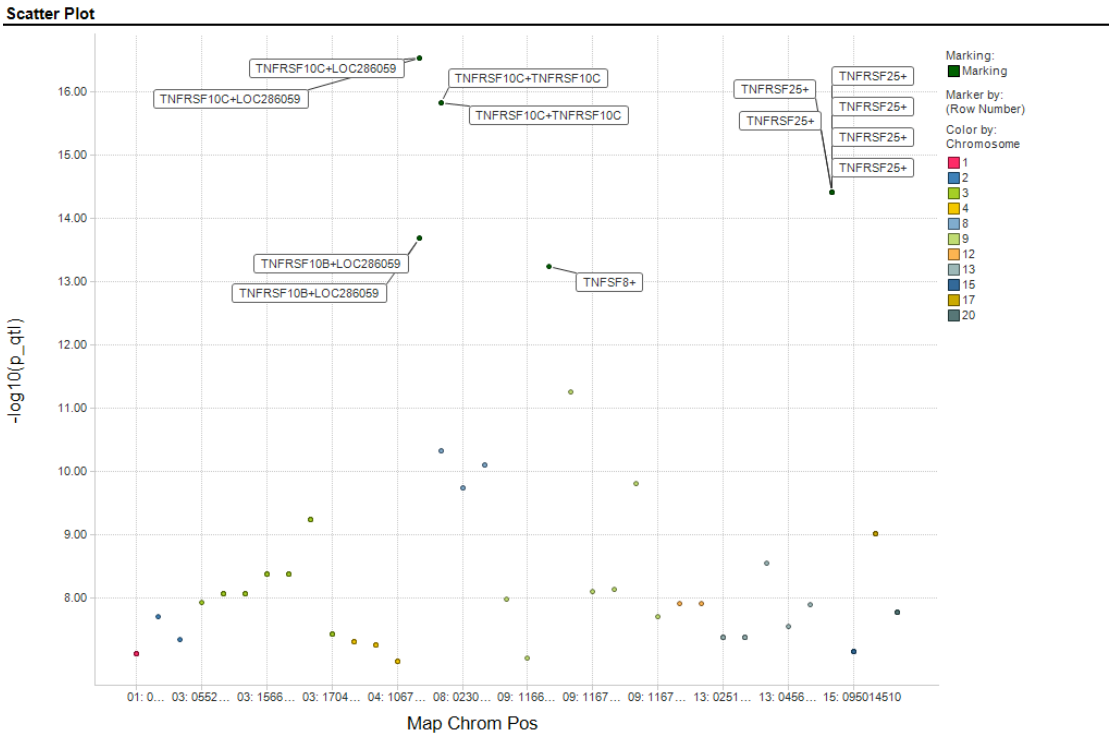
Σχήμα 6.11 Ελάχιστη απόσταση του SNP μεταξύ μετάγραφο mRNA

Η γραφική παράσταση του σχήματος 6.11 όπως και αυτή του σχήματος 6.10 παρουσιάζει, την απόσταση των SNPs, από τη γενετική θέση στην οποία βρίσκονται τα μόρια mRNA πάνω στα χρωμοσώματα. Σε αυτή την περίπτωση, γίνεται μια μεγένθυνση των αποτελεσμάτων έτσι ώστε να μπορεί να μελετηθεί με μεγαλύτερη λεπτομέρια άν η απόσταση των SNPs από την θέση απ' όπου μεταγράφονται τα μόρια mRNA, ώστε να μπορεί να αποφασιστεί ποιά από αυτά αποτελούν περιπτωσεις cis. Παρατηρώντας τη συγκέντρωση των σημείων κοντά στο 0, φαίνεται πως η μεγαλύτερη συγκέντρωση αυτών, αναφέρεται στα χρωμοσώματα 1 και 20 (βλέπε διάταξη χρωμοσωμάτων στα δεξιά). Επομένως τα SNPs που βρίσκονται πάνω σε αυτά τα χρωμοσώματα αποτελούν κατά μεγάλη πιθανότητα περιπτώσεις cis. Επιπλέον και οι τιμές των αρνητικών λογάριθμων των πιθανοτήτων σημαντιότητας παρατηρούνται να είναι πολύ μεγαλύτερες από αυτές των άλλων χρωμοσωμάτων με ορισμένες εξαιρέσεις.



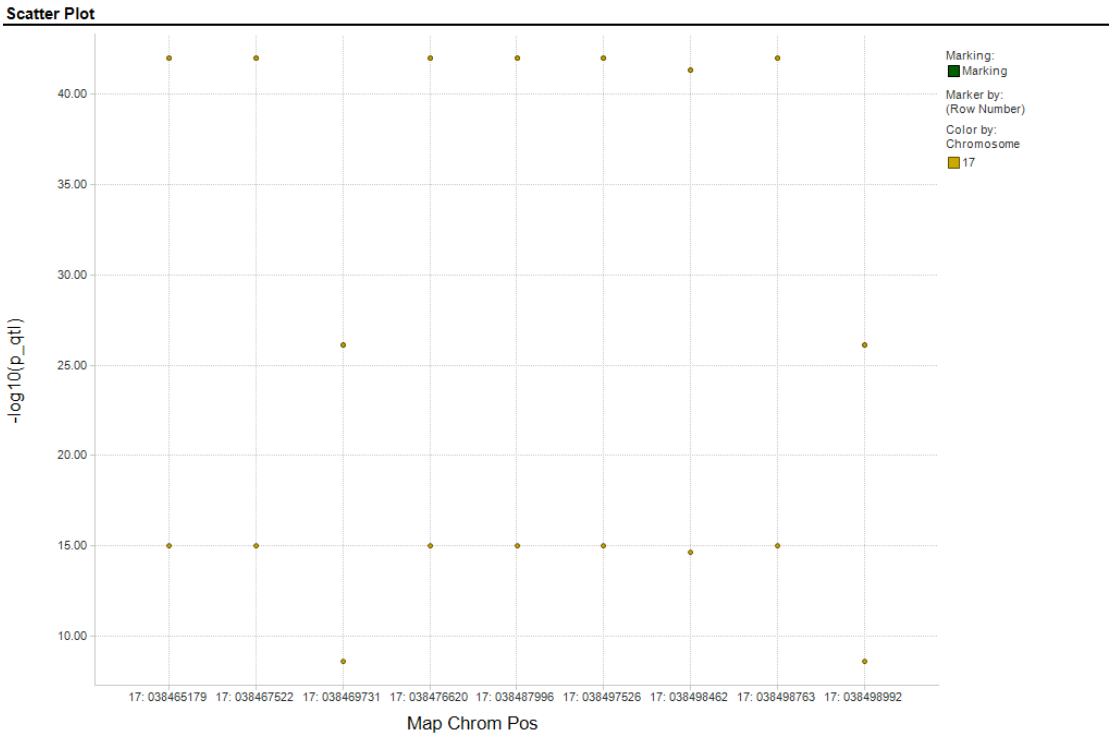
Σχήμα 6.12 Αποτελέσματα της οικογένειας του γονιδίου TNF μετά από την εφαρμογή φιλτραρίσματος

Το σχήμα 6.12 παρουσιάζει τα αποτελέσματα μετά από την εφαρμογή φίλτρου στην οικογένεια γονίδιων TNF. Στον άξονα των X διατάσσονται οι θέσεις των SNPs πάνω στα διάφορα χρωμοσώματα, ενώ στον άξονα των Ψ τις τιμές των πιθανοτήτων σημαντικότητας. Στο πλάι δεξιά φαίνονται αριθμημένα τα διάφορα χρωμοσώματα για τα οποία υπάρχουν αποτελέσματα για τη συγκεκριμένη οικογένεια γονιδίου. Οι περισσότερες απο τις τιμές αυτές παρουσιάζονται να φτάνουν μέχρι το 8 συγκρίνοντας βάσει του άξονα των X, με ορισμένες από αυτές να βρίσκονται και κάτω από το 5, καθιστώντας τις μη σημαντικές βάσει του στατιστικού ορισμού των πιθανοτήτων σημαντικότητας (p-values), ενώ άλλες ξεπερνούν κατά πολύ το όριο αυτό καθιστώντας τα αποτελέσματα σημαντικά σε σχέση με τα υπόλοιπα. Γενικά μπορεί να βγεί το συμπέρασμα πως η σημασία των αποτελεσμάτων για αυτή τη συγκεκριμένη οικογένεια γονιδίου δεν είναι μεγάλη έστω κι αν κάποια για κάποια από αυτά παρατηρούνται στατιστικά σημαντικές τιμές στον άξονα των X.



Σχήμα 6.13 Αποτελέσματα της οικογένειας του γονιδίου TNF μετά από την εφαρμογή φιλτραρίσματος και επιλογή των σημαντικότερων αποτελεσμάτων

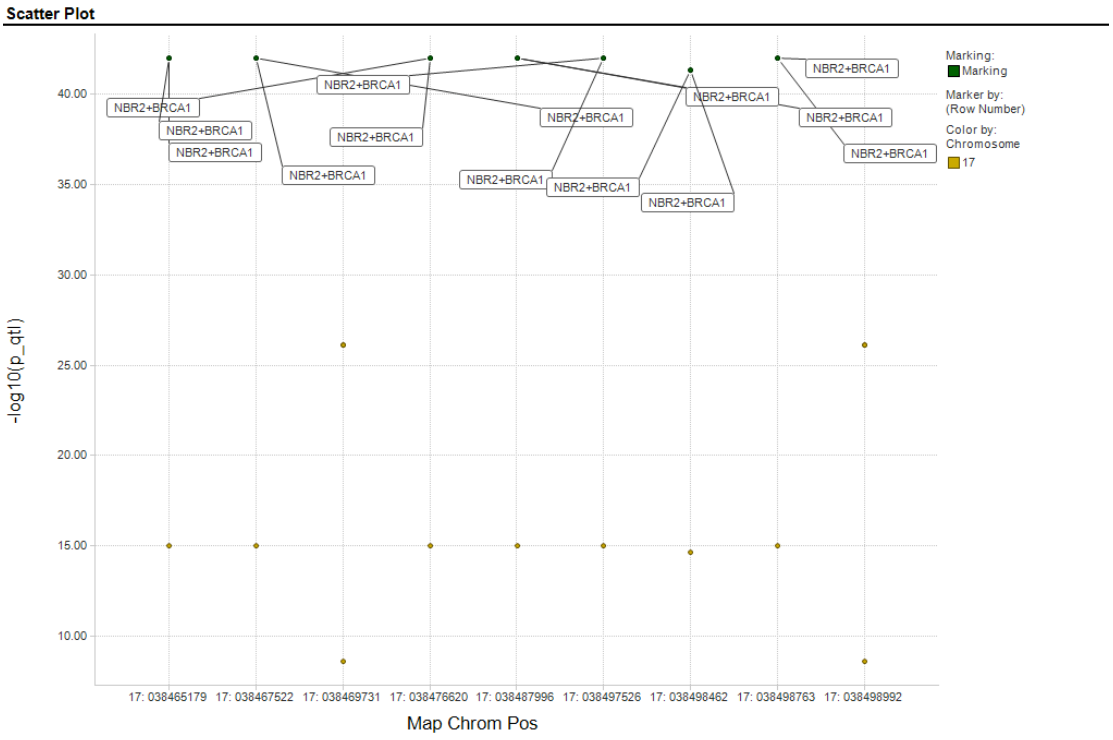
Το σχήμα 6.13 παρουσιάζει ακριβώς τα ίδια αποτελέσματα με το σχήμα 6.12 δηλαδή, με την εφαρμογή φίλτρου πάνω στο γονίδιο TNF. Η διαφορά είναι πως σε αυτή την περίπτωση έχουν επιλεγεί οι σημαντικότερες από τις λογαριθμικές τιμές των πιθανοτήτων σημαντικότητας. Η επιλογή αυτή έγινε για την προβολή περεταίρω πληροφοριών των σημαντικότερων από τις τιμές του γονιδίου αυτού και για επίδειξη των υπηρεσιών που μπορεί να μας παρέχει η συγκεκριμένη εφαρμογή.



Σχήμα 6.14 Αποτελέσματα της οικογένειας του γονιδίου BRCA1 μετά από την εφαρμογή φιλτραρίσματος

Το σχήμα 6.14 παρουσιάζει τα αποτελέσματα μετά από την εφαρμογή φίλτρου στην οικογένεια γονιδίου BRCA1. Το φίλτρο που εφαρμόστηκε σε αυτή την περίπτωση είναι διαφορετικό σε σύγκριση με το φιλτράρισμα που εφαρμόστηκε πάνω στο γονίδιο TNF. Στους άξονες εντούτοις παρουσιάζονται οι ίδιες μονάδες. Ο μόνη διαφορά είναι ότι το φίλτρο προσαρμόζεται σε κάθε διαφορετική περίπτωση ξεχωριστά. Στον άξονα των X διατάσσονται οι θέσεις των SNPs πάνω στα διαφορετικά χρωμοσώματα, ενώ στον άξονα των Ψ οι τιμές των πιθανοτήτων σημαντικότητας προσαρμοσμένες βάσει του αρνητικού λογάριθμου. Δεξιά στο πλάι φαίνονται αριθμημένα τα διάφορα χρωμοσώματα για τα οποία υπάρχουν αποτελέσματα για το συγκεκριμένο γονίδιο. Επιπλέον όπως αναφέρθηκε και νωρίτερα οι σημαντικότερες από τις τιμές είναι αυτές που παρουσιάζονται στη γραφική είναι αυτές με την μεγαλύτερη τιμή στον άξονα των Ψ. Δηλαδή οι τιμές με τον μεγαλύτερο λογάριθμο είναι και οι σημαντικότερες βάσει του ορισμού των πιθανοτήτων σημαντικότητας [15]. Μπορεί να παρατηρηθεί ότι τα αποτελέσματα για το συγκεκριμένο γονίδιο αναφέρονται μόνο σε ένα μόνο χρωμόσωμα, το χρωμόσωμα 17. Δηλαδή τα SNPs που βρίσκονται στο συγκεκριμένο

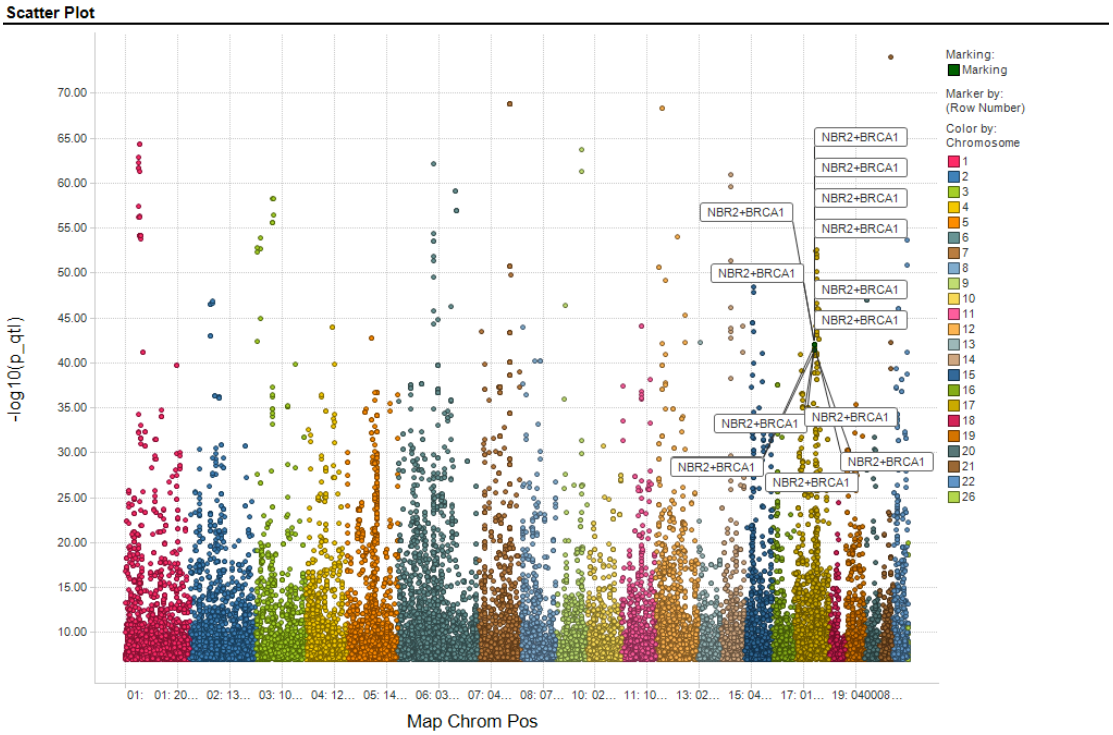
γονίδιο είναι όλα συγκεντρωμένα σε ένα και μόνο χρωμόσωμα, το χρωμόσωμα 17. Επιπλέον οι λογαριθμικές τιμές των πιθανοτήτων σημαντικότητας (p-value) φαίνονται να είναι υψίστης σημασίας σε σχέση με προηγούμενα αποτελέσματα καθώς είναι και συγκριτικά μεγαλύτερες. Τα λιγότερο σημαντικά αποτελέσματα είναι περίπου στο μισό του 15 όπως πολύ καθαρά μπορεί να παρατηρηθεί στο σχήμα 6.14. Τα υπόλοιπα ξεκινούν από την τιμή 16 όσον αφορά τον αρνητικό λογάριθμο που παρουσιάζεται στον άξονα των Ψ. Σημαίνει πως ξεπερνούν κατά πολύ το ανώτατο όριο που θέτει ο ορισμός των πιθανοτήτων σημαντικότητας 0.05 [15], καθώς οι τιμές των αποτελεσμάτων ανέρχοναι γύρω στο 0.016 όπως μπορεί να παρατηρηθεί και από τη γραφική παράσταση. Σύμφωνα με τις παρατηρήσεις αυτές, οι τιμές των συσχετίσεων για την οικογένεια γονιδίων BRCA1 αποτελούν τιμές μεγάλης σημαντικότητας σε σχέση με αυτές που αναλύθηκαν στο σχήμα 6.13 για την οικογένεια γονιδίων TNF.



Σχήμα 6.15 Αποτελέσματα της οικογένειας του γονιδίου BRCA1 μετά από την εφαρμογή φιλτραρίσματος και επιλογή σημαντικότερων αποτελεσμάτων

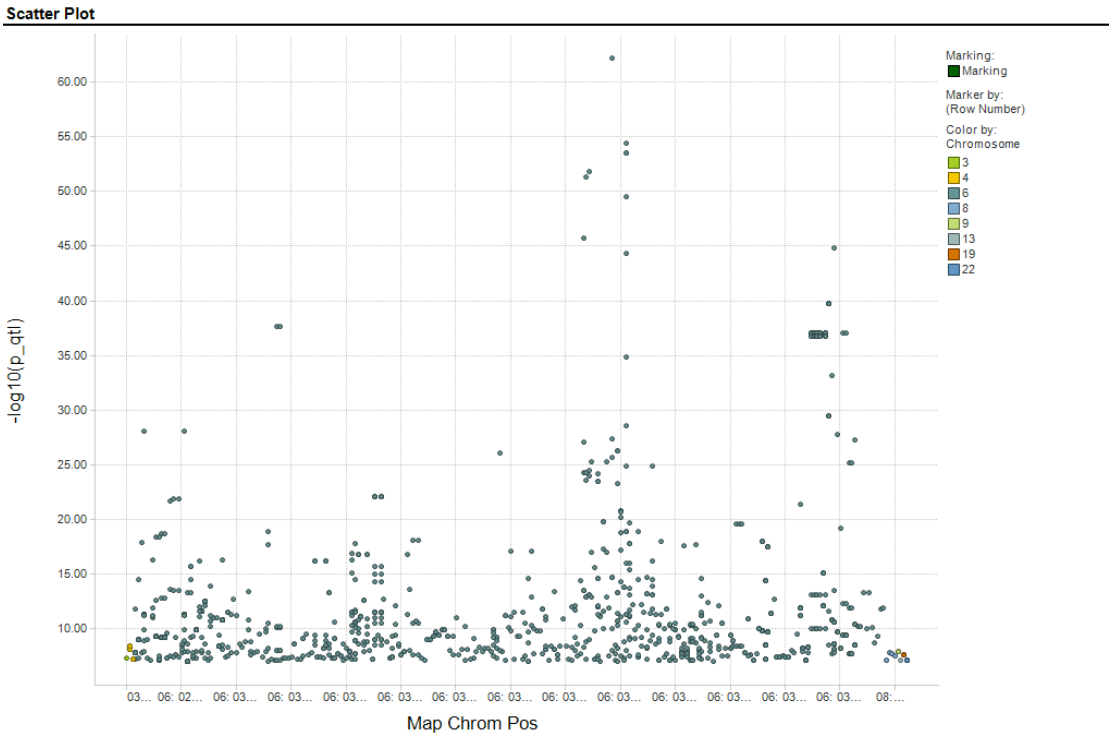
Το σχήμα 6.15 παρουσιάζει ακριβώς τα ιδία αποτελέσματα που σχολιάστηκαν και στο σχήμα 6.14 μόνο που σε αυτή την περίπτωση έχουν επιλεγεί οι μεγαλύτερες από τις

τιμές σε σχέση με τον άξονα των X για παρουσίαση περεταίρω πληροφοριών. Επίσης και για προβολή των υπηρεσιών που μπορεί να μας παρέχει η εφαρμογή Spotfire και τον ρόλο που έπαιξε όσον αφορά την διευκόλυνση στην παρουσίαση και απεικόνιση των αποτελεσμάτων.



Σχήμα 6.16 Προβολή των αποτελεσμάτων του φιλτραρίσματος στο γονίδιο BRCA1 σε σχέση με τα υπόλοιπα αποτελέσματα

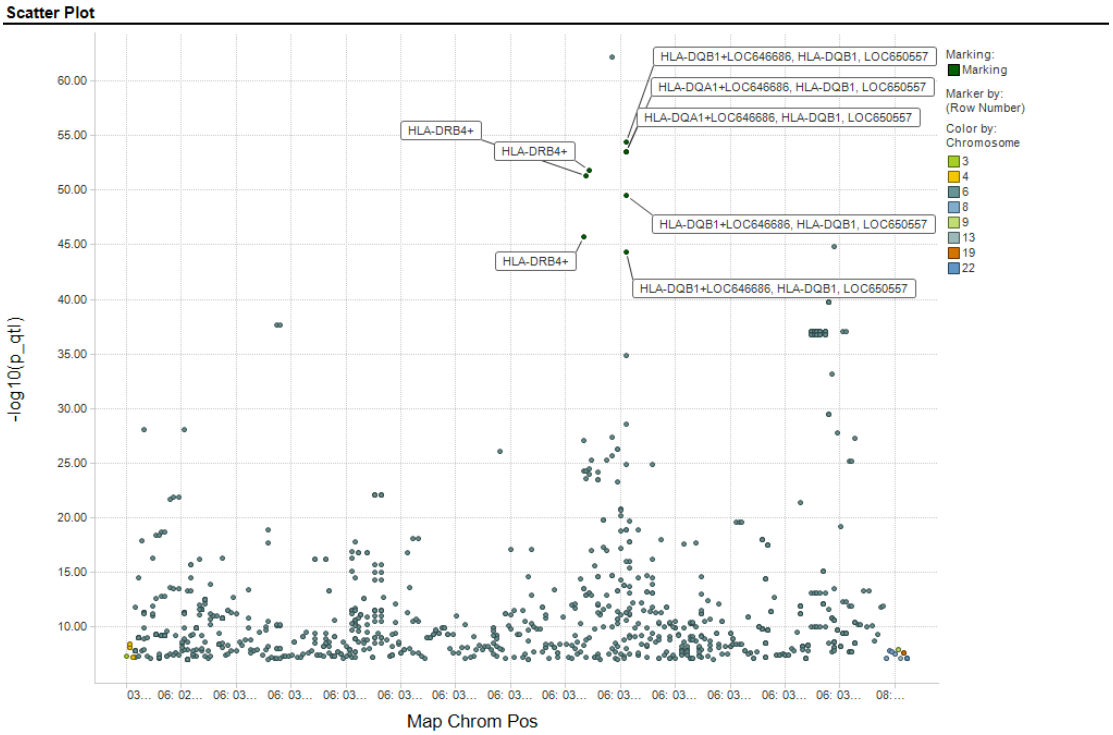
Στο σχήμα 6.16 παρουσιάζονται οι τιμές του γονιδίου BRCA1 που επιλέχθηκαν εφαρμόζοντας το φίλτρο που αναφέρθηκε και στα σχήματα προηγούμενες γραφικές με τη διαφορά ότι σε αυτή την περίπτωση τα αποτελέσματα παρουσιάζονται σε σχέση με όλα τα υπόλοιπα χρωμοσώματα.



Σχήμα 6.17 Αποτελέσματα της οικογένειας του γονιδίου HLA μετά από την εφαρμογή φιλτραρίσματος

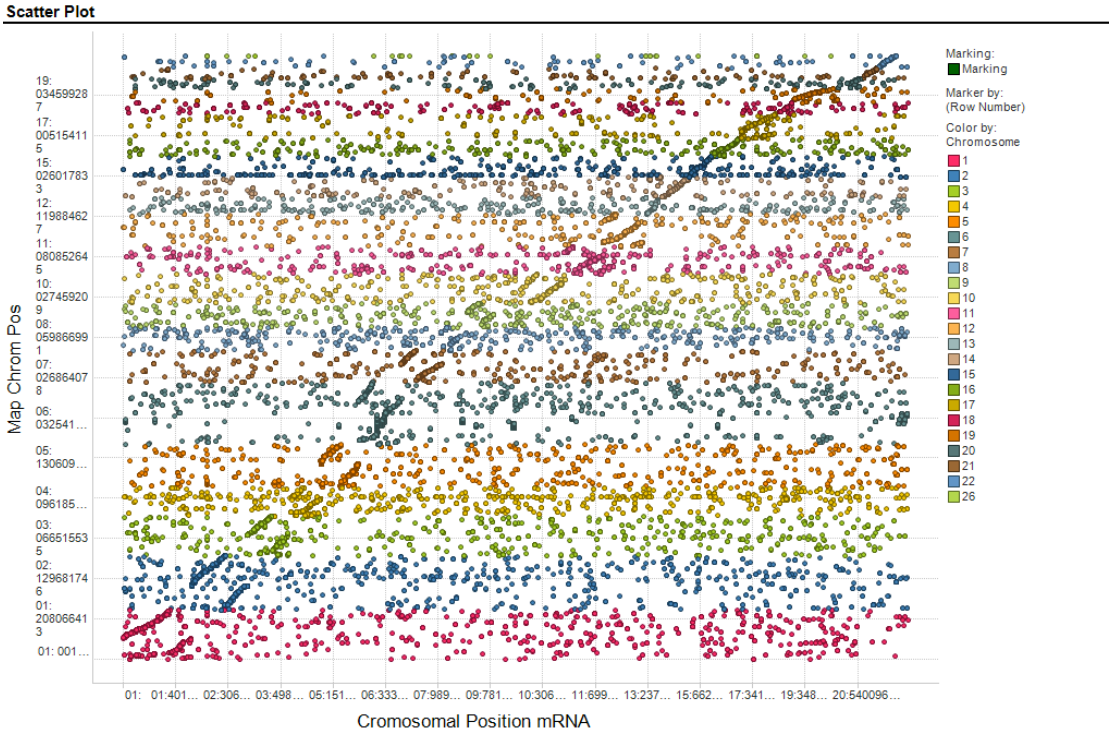
Το σχήμα 6.17 παρουσιάζει τα αποτελέσματα μετά από την εφαρμογή φίλτρου στο γονίδιο HLA . Στον άξονα των X διατάσσονται οι θέσεις των SNPs πάνω στα διάφορα χρωμοσώματα, ενώ στον άξονα των Ψ τις τιμές των πιθανοτήτων σημαντικότητας προσαρμοσμένες βάσει του αρνητικού λογάριθμου με βάση το 10. Στο πλάι αριθμημένα φαίνονται τα διάφορα χρωμοσώματα για τα οποία υπάρχουν αποτελέσματα σε αυτή τη γραφική για το συγκεκριμένο γονίδιο. Επιπλέον όπως αναφέρθηκε και στα προηγούμενα σχήματα, οι σημαντικότερες από τις τιμές που παρουσιάζονται στη γραφική είναι αυτές με την μεγαλύτερη τιμή στον άξονα των Ψ. Δηλαδή οι τιμές με τον μεγαλύτερο λογάριθμο είναι και οι σημαντικότερες βάσει του ορισμού των πιθανοτήτων σημαντικότητας [15] καθώς σημαντικότερες απο αυτές τις τιμές είναι οι μικρότερες, ενώ με τον αρνητικό λογάριθμο σημαντικότερες από αυτές είναι οι μεγαλύτερες. Παρατηρείται πως έστω και αν υπάρχει κάποια τάση συγκέντρωσης των τιμών των πιθανοτήτων σημαντικότητας για κάποιες αυτές γύρω στο 5, οι περισσότερες απο τις τιμές σημαντικότητας είναι μεγαλύτερες από το 5. Αυτό τις καθιστά στατιστικά σημαντικές σύμφωνα με τον ορισμό των πιθανοτήτων σημαντικότητας [15].

Συγκρίνοντας όμως τα αποτελέσματα σε σχέση με αυτά των σχημάτων 6.12 και 6.14 η συγκέντρωση των αποτελεσμάτων είναι μεγαλύτερη απο τις υπόλοιπες περιπτώσεις, και οι τιμές αποτελούν στατιστικά σημαντικές τιμές, επομένως τα SNPs που βρίσκονται σε αυτό το γονίδιο είναι πολύ πιθανόν να επηρεάζουν την έκφραση κάποιου μορίου mRNA του οποίου η περιοχή μεταγραφής είναι γειτονική με αυτή των SNPs στο συγκεκριμένο γονίδιο.



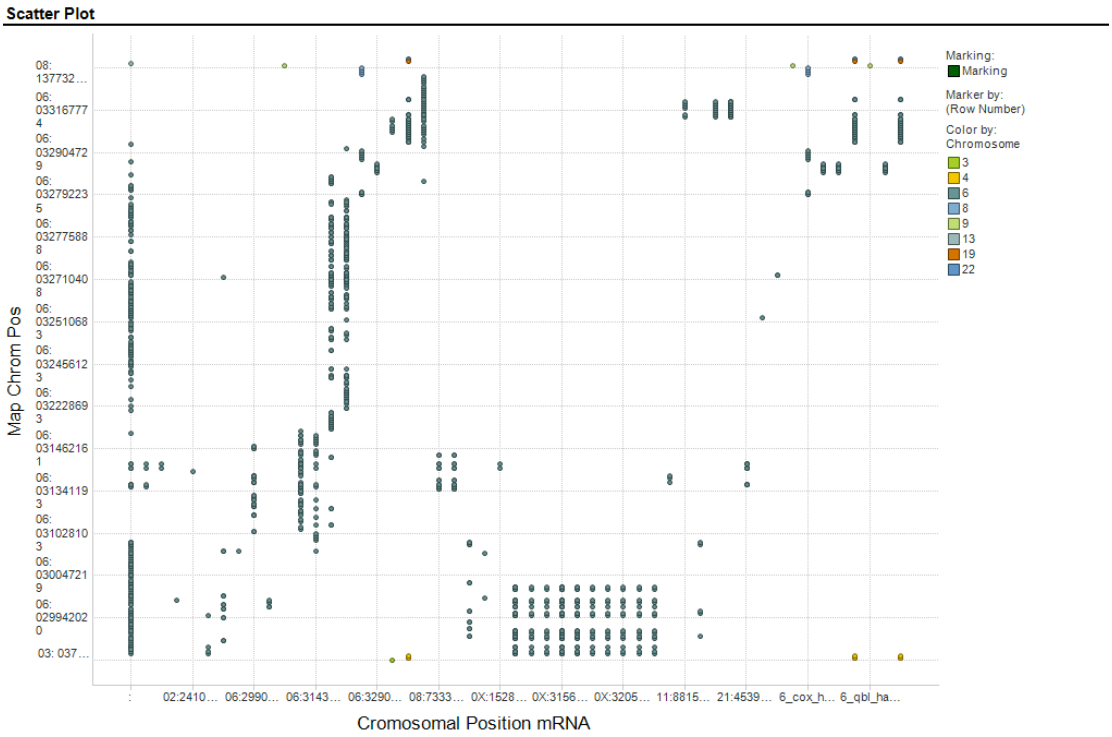
Σχήμα 6.18 Αποτελέσματα της οικογένειας του γονιδίου HLA μετά από την εφαρμογή φιλτραρίσματος και επιλογή σημαντικότερων αποτελεσμάτων

Η γραφική του σχήματος 6.18 παρουσιάζει ακριβώς τα ίδια αποτελέσματα που έχουν σχολιαστεί και στη γραφική παράσταση του σχήματος 6.17. Επιλέγηκαν οι σημαντικότερες από τις τιμές που παρατηρήθηκαν στην γραφική του σχήματος 6.17 για προβολή περεταίρω πληροφοριών που μπορεί να μας παρέχει η εφαρμογή που χρησιμοποιήθηκε για την απεικόνιση και παρουσίαση των αποτελεσμάτων. Επίσης ο μεγάλος αριθμός των τιμών που παρατηρούνται να συγκεντρώνονται στο χρωμόσωμα 6 υποδεικνύουν και τη μεγάλη συγκέντρωση SNPs στο χρωμόσωμα αυτό.



Σχήμα 6.19 Μηδενική απόσταση μεταξύ SNP τοποθεσίας και mRNA μετάγραφου

Η γραφική παράσταση του σχήματος 6.19 παρουσιάζει τις περιπτώσεις cis, δηλαδή αυτές στις οποίες τα SNPs και τα μόρια mRNA βρίσκονται στις ίδιες θέσεις ή σε γειτονικές πάνω στα διάφορα χρωμοσώματα. Ο άξονας των X παρουσιάζει την διάταξη των θέσεων των μορίων του mRNA πάνω σε κάποιο χρωμόσωμα και ο άξονας των Ψ το αντίστοιχο για τα SNPs. Το cis φαίνεται από την συγκέντρωση των τιμών πάνω στην κεντρική διαγώνιο σχηματίζοντας την. Οι τιμές πιθανοτήτων σημαντικότητας για τα SNPs και mRNA σύμφωνα με την πιο πάνω γραφική υποδεικνύουν τη συσχέτιση τους και την κοινή θέση στο χρωμόσωμα που κατέχουν. Αυτό καθορίζει πως τα SNPs που βρίσκονται στην ίδια ή γειτονική θέση με το μόριο του mRNA πάνω στο χρωμόσωμα, δρούν ως στοιχεία ενεργοποίησης (cis) επηρεάζοντας την έκφραση του.



6.20 Κοινές θέσεις SNPs και mRNA σε συγκεκριμένα χρωμοσώματα

Η γραφική παράσταση του σχήματος 6.20 παρουσιάζει ακριβώς ποιά SNPs και ποιά μόρια mRNA κατέχουν κοινές θέσεις και πάνω σε ποια χρωμοσώματα. Όπως μπορεί να παρατηρηθεί οι κοινές θέσεις των SNPs και μορίων mRNA συγκεντρώνονται κυρίως στο χρωμόσωμα 6 του γονιδίου HLA, υποδεικνύοντας καθαρά όλες τις περιοχές Cis πάνω στο γονίδιο. Δηλαδή τις περιοχές όπου τα SNPs βρίσκονται κοντά ή και μέσα στην περιοχή του μορίου του mRNA με αποτέλεσμα να επηρεάζουν και την έκφραση του.

6.3 Γενικά Συμπεράσματα

Τα γενικά συμπεράσματα που μπορούν να προκύψουν από την μελέτη των αποτελεσμάτων των γραφικών παραστάσεων είναι ότι οι παρατηρήσεις και τα αποτελέσματα που διεξάγονται από τη μελέτη των γραφικών είναι τα ίδια και για τις δύο περιπτώσεις εφαρμογής φιλτραρίσματος με διαφορετικό κατώφλι, οδηγώντας στα ίδια συμπεράσματα. Επομένως το περεταίρω φιλτράρισμα μπορεί να φανεί πολύ χρήσιμο για αποδοτική λειτουργία των μετέπειτα εργαλείων που θα χρησιμοποιούνται για την παρουσίαση και προβολή των αποτελεσμάτων.

Κεφάλαιο 7

Συζήτηση

7.1 Γενική Συζήτηση γύρω από το Θέμα

Τα αποτελέσματα που παράχθηκαν από τη συγκεκριμένη μελέτη είναι πάρα πολύ ψηλής σημαντικότητας.

Συνήθως τόσο σε γενετικές μελέτες με τον ίδιο αριθμό SNPs, όσο και σε μελέτες με τον αντίστοιχο αριθμό mRNA που στόχος είναι να μελετηθεί η συσχέτιση με ένα γενετικό χαρακτηριστικό (πχ μια ασθένεια με γενετική προδιάθεση) των mRNA ή SNPs, δεν παρουσιάζονται πολύ σημαντικές τιμές p-value. Εντούτοις σε αυτή την έρευνα τα κορυφαία αποτελέσματα ήταν πέραν τον 10^{60} φορές πιο σημαντικά από ότι τα αναμενόμενα από τις αντίστοιχες άλλες μελέτες [3].

Από πλευράς στατιστικής το πρόβλημα των πολλαπλών ελέγχων μπορεί να ευθύνεται για μια παρατηρημένη αύξηση στο επίπεδο σημαντικότητας, αλλά σύμφωνα με την μέθοδο Bonferroni για την διόρθωση του προβλήματος πολλαπλών ελέγχων, [5] η διαφορά δεν πρέπει να είναι πέραν του 10^4 για την περίπτωση των SNPs και 10^5 για την περίπτωση των μελετών με mRNA δεδομένα. Έχοντας χρησιμοποιήσει και τον έλεγχο σε ένα τυχαίο υποσύνολο των κορυφαίων τιμών της μεθόδου διεργασίας μεταλλαγής, η οποία προσφέρει ακριβή διόρθωση των p-values, είχαμε διαπιστώσει ότι όντος η διαφορά στις τιμές σημαντικότητας, λόγω του προβλήματος πολλαπλών ελέγχων είναι ακόμα μικρότερες και από αυτές που υποδεικνύει η ευρηστική μέθοδος Bonferroni. Επομένως η μόνη άλλη εξήγηση για την τεράστια διαφορά στα επίπεδα σημαντικότητας είναι ότι επειδή όλα τα δεδομένα προέρχονται από μετρήσεις απευθείας από τους βιολογικούς μηχανισμούς, αντί των κλασσικών αναλύσεων όπου η μια μεταβλητή σε

όλους τους ελέγχους είναι η διάγνωση ενός δείγματος για κάποια ασθένεια, είμαστε σε θέση να βρίσκουμε συσχετίσεις όπου σε αυτές εμπεριέχεται πολύ πιο λίγος θόρυβος.

Ένα ακόμα εντυπωσιακό αποτέλεσμα ήταν ο βαθμός με τον οποίο ξεχώριζαν οι συσχετίσεις cis από τις trans στις γραφικές που παρουσιάζονται στα σχήματα 6.3, 6.10, 6.11 και 6.19. Οι διαφορές ειχαν παρατηρηθεί και σε άλλη μελέτη με πολύ παρόμοια δεδομένα [3] αλλά δεν ήταν τόσο αισθητές. Πιστεύεται ότι αυτό είναι ένδειξη της ανώτερης ποιότητας δεδομένων που ήταν διαθέσιμα για αυτή την μελέτη.

Όταν τα αποτελέσματα παρουσιάστηκαν για πρώτη φορά σε συνεδρία με γενετιστές και ιατρούς από διάφορες ειδικότητες, μια από τις σημαντικότερες παρατηρήσεις ήταν ότι ανάμεσα στις κορυφαίες συσχετίσεις υπήρχαν αρκετές που ενέπλεκαν περιοχές γνωστές ως υπεύθυνες για συχνές, και πολύπλοκες ασθένειες. Αυτό αυξάνει τις ελπίδες ότι τα αποτελέσματα από αυτή την έρευνα θα εφαρμοστούν και σε νέα βιολογικά πειράματα.

Επομένως εκτός από τα αποτελέσματα υψίστης σημασίας η μελέτη αυτή διευρύνει τους ορίζοντες για την διεξαγωγή νέων ερευνών γύρω από το θέμα αυτό και περεταίρω μελλοντικές εργασίες.

Επιπλέον σημαντικά είναι και τα αποτελέσματα που επιτεύχθηκαν με τον τρόπο υλοποίησης του κώδικα που χρησιμοποιήθηκε για την ανάλυση και διαχείριση των δεδομένων. Ιδιαίτερα σημαντικό ρόλο έπαιξε στην μείωση των χρόνων εκτέλεσης τόσο η υλοποιίηση κώδικα όσο και η χρήση του grid. Η βελτιστοποίηση στους χρόνους εκτέλεσης ήταν αισθητή καθώς επιτεύχθηκε μείωση των χρόνων εκτέλεσης απο τις 400 εβδομάδες που υπολογίστηκε ο αναμενόμενος απαιτούμενος χρόνος σειριακής επεξεργασίας των δεδομένων σε μόνο 2 εβδομάδες, βάσει του σχεδιασμού και της υλοποίησης του κώδικα που υλοποιήθηκε στην συγκεκριμένη εργασία. Αυτό είχε ώς αποτέλεσμα την βελτιστοποίηση των χρόνων εκτέλεσης κατά 98%.

Από τα αποτελέσματα αυτά είναι αντιλυπτό πως η σημασία των αποτελεσμάτων δεν περιορίζεται μόνο στα αποτελέσματα που παράχθηκαν από την ανάλυση των δεδομένων αλλά επίσης σημαντική και αξιοσημείωτη είναι η αύξηση της απόδοσης του

συστήματος, βάσει του σχεδίου επεξεργασίας των δεδομένων πάνω στο οποίο διεξήχθηκε η όλη διαδικασία.

Επιπλέον ο τρόπος με τον οποίο έγινε η υλοποίηση του κώδικα διευκολύνει την μετέπειτα εξέλιξη του για την εφαρμογή του σε ακόμη πιο πολύπλοκη ανάλυση δεδομένων καθώς επίσης και την φορητότητα του σε οποιοδήποτε λειτουργικό σύστημα καθώς είναι υλοποιημένο σε γλώσσα C++ και μπορεί να χρησιμοποιηθεί σε περιβάλλοντα Unix, Windows, Mac OS X, Linux αρκεί να μεταγλωττιστούν στο κατάλληλο λειτουργικό σύστημα. Επιπλέον η υλοποίηση του διευκολύνει την συστήρηση του που είναι ένας από τους σημαντικότερους παράγοντες για τη ζωή ενός προγράμματος, όπως επίσης και για την επιδιόρθωση διάφορων σφαλμάτων που πιθανόν να προκύψουν κατά την εκτέλεση αλλά και για την τροποποίηση του ώστε να μπορεί να προσαρμοστεί εύκολα και σε μελέτες με παρόμοιου τύπου δεδομένα.

Κλείνοντας , τα αποτελέσματα είναι τα βέλτιστα δυνατά τόσο στον τομέα της βιολογίας όσο και στον τομέα της πληροφορικής καθώς ικανοποιούνται σε μεγάλο βαθμό όλες οι απαιτήσεις από ένα σύστημα αυτού του είδους. Τα μεγάλης σημασίας αποτελέσματα που επιτεύχθηκαν και στους δύο τομείς ανεβάζουν το επίπεδο της μελέτης και αυξάνουν την χρησιμότητα της σε άλλες μελέτες όμοιου τύπου στις οποίες μπορεί να φανεί χρήσιμη σε σχέση με άλλες μελέτες που διεξήχθηκαν γύρω από το ίδιο θέμα.

Κεφάλαιο 8

Συμπεράσματα

8.1 Γενικά Συμπεράσματα

1. Έχει επιτευχθεί η μελέτη μεταξύ γονότυπων και των εκφραζόμενων φαινοτύπων σε ολόκληρο το ανθρώπινο γονιδίωμα

2. Έχει δημιουργηθεί μια βάση δεδομένων με προοπτικές έρευνας για αξιολόγηση της συσχέτισης μεταξύ των SNPs και της έκφρασης του mRNA

3. Η συγκεκριμένη βάση δεδομένων μπορεί να βοηθήσει στην ανακάλυψη νέων φαρμάκων σε όλους τους θεραπευτικούς τομείς. Η ανακάλυψη νέων φαρμάκων μπορεί να γίνει αξιολογώντας τα λειτουργικά αποτελέσματα που μπορούν να έχουν κάποιοι παράγοντες κινδύνου ή κάνοντας εισηγήσεις για γενετικά βασιζόμενους (genetic-based) βιολογικούς δείκτες (biomarkers). Επίσης τονίζοντας τα SNPs που επηρεάζουν την έκφραση γονιδίων που αποτελούν στόχους, μπορούν να ληφθούν υπόψη για μελέτες φαρμακογενετικής (PGx)

4. Όσον αφορά την μεθοδολογία διάσπασης και επεξεργασίας των δεδομένων, ο τρόπος με τον οποίο γράφτηκε ο κώδικας τον καθιστά ικανό να χρησιμοποιηθεί μελλοντικά για επεξεργασία παρόμοιου τύπου δεδομένων μιας μελέτης όπου υπάρχουν για τα ίδια δείγματα, διαθέσιμα δεδομένα DNA (SNPs) και οποιουδήποτε άλλου είδους φαινοτύπου. Επιπλέον η μεθοδολογία μπορεί να είναι εφαρμόσιμη για παρόμοιες μελέτες για δεδομένα μικρότερου ή και

μεγαλύτερου όγκου από αυτά που χρησιμοποιήθηκαν στην συγκεκριμένη μελέτη

5. Η χρήση κώδικα που υλοποιήθηκε σε αυτή τη μελέτη, έπαιξε καθοριστικό ρόλο στην επεξεργασία των δεδομένων καθώς με τη χρήση του, έγινε εφικτή η μείωση του όγκου των ενδιάμεσων αποτελεσμάτων, προσαρμόζοντας τα στα ποσά διαθέσιμης μνήμης που παρείχε το υφιστάμενο σύστημα. Επιπλέον όσον αφορά τη χρήση του grid από πλευράς απόδοσης και ταχύτητας της επεξεργασίας των δεδομένων, υπολογίζοντας τον απαιτούμενο χρόνο για εκτέλεση χωρίς τη χρήση του grid ανέρχεται περίπου γύρω στις 400 περίπου εβδομάδες, δηλαδή 5 χρόνια, ενώ η επεξεργασία εφαρμόζοντας τη δική μας μέθοδο, διήρκεσε μόνο 2 εβδομάδες. Με αυτό τον τρόπο επιτεύχθηκε 200 φορές μεγαλύτερη ταχύτητα στην επεξεργασία των δεδομένων αυξάνοντας έτσι και την απόδοση του συστήματος

6. Η χρήση του κώδικα φιλτραρίσματος για περιορισμό του όγκου των δεδομένων του αρχείου με τα τελικά αποτελέσματα, επιτρέπει την απομόνωση δεδομένων γύρω από μία συγκεκριμένη περιοχή ενδιαφέροντος (region specific), ενός υποσυνόλου mRNA ή ενός υποσυνόλου γονιδίων ή και συνδυασμού των δύο. Για παράδειγμα, μπορεί να δημιουργηθεί μια λίστα με γονίδια, mRNA και πρωτεΐνες που ανήκουν σε ένα pathway, και η εφαρμογή φίλτρου στα αποτελέσματα του mRNA, πρωτεΐνης ή γονιδίου που εμπλάκηκε στο pathway. Αυτό επιτρέπει την προσαρμογή των δεδομένων για διάφορους σκοπούς μελέτης καθιστώντας τα χρήσιμα και για μελλοντική χρήση σε άλλες μελέτες. Επιπλέον αυτό μπορεί να γίνει χωρίς να χαθούν τα αρχικά αποτελέσματα, που προέκυψαν μετά την εφαρμογή κώδικα φιλτραρίσματος κατά τη συγχώνευση

8.2 Μελλοντική Εργασία

Στο πεδίο της συγκεκριμένης μελέτης υπάρχουν ορισμένοι τομείς οι οποίοι μπορούν να προσφερθούν για μελλοντική εργασία.

Μια σημαντική περίπτωση για μελλοντική εργασία πάνω στο συγκεκριμένο θέμα, είναι η επιβεβαιωτική επανάληψη (Replication Testing). Η μέθοδος αυτή μέσω της χρήσης ανάλογων δεδομένων από άλλη μελέτη αποσκοπεί στην εξακρίβωση της εγκυρότητας των αποτελεσμάτων.

Επιπλέον η δημοσιοποίηση των αποτελεσμάτων και διαθεσιμότητα τους στο διαδίκτυο μέσω μίας ιστοσελίδας που θα υποστηρίζει και τις λειτουργίες που προσφέρει το Spotfire, για απεικόνιση και παρουσίαση των αποτελεσμάτων. Αυτό θα διεύρυνε την δυνατότητα χρήσης των αποτελεσμάτων από επιστήμονες που ασχολούνται με βιολογικά ερωτήματα, όπου η γνώση συσχετίσεων μεταξύ γονότυπων και έκφρασης mRNA ή πρωτεϊνών θα τους βοηθούσε. Να σημειωθεί ότι ανάγκη για την συγκεκριμένη γνώση υπάρχει σε παρά πολλά είδη πειραμάτων, και η γνώση μπορεί να χρησιμοποιηθεί ασχέτως της ασθένειας η άλλου γενετικού χαρακτηριστικού που πιθανόν να ενδιαφέρει τους ερευνητές.

Η εφαρμογή των αποτελεσμάτων σε μελέτες φαρμακογενετικής (PGx) δηλαδή μελέτες που αποσκοπούν στην ανακάλυψη νέων φαρμάκων, αξιολογώντας τα λειτουργικά αποτελέσματα που μπορούν να έχουν κάποιοι παράγοντες κινδύνου ή κάνοντας εισηγήσεις για γενετικά βασιζόμενους (genetic-based) βιολογικούς δείκτες (biomarkers), που θα μπορούσαν να χρησιμοποιηθούν σε μελέτες ελέγχου αποδοτικότητας νέων φαρμακευτικών ουσιών. Η βάση δεδομένων είναι υψίστης σημασίας καθώς μπορεί να εφαρμοστεί σε όλους τους θεραπευτικούς τομείς [10].

Όσο αφορά τον κώδικα που υλοποιήθηκε για τη διαχείριση και επεξεργασία των δεδομένων σε όλα τα στάδια της μελέτης, έγινε με τέτοιο τρόπο ώστε να μπορεί να εφαρμοστεί σε παρόμοιου τύπου δεδομένα, ανεξαρτήτως όγκου και περιεχομένων που μπορούν να ακολουθήσουν την ίδια διαδικασία ανάλυσης. Η παραμετροποίηση των

δεδομένων εισόδου καθιστούν τον κώδικα εύκολα εφαρμόσιμο και χρησιμοποιήσιμο σε μελλοντικές μελέτες.

Θα μπορούσε να κωδικοποιηθεί η γνώση που παράχθηκε χάρη σε αυτή την μελέτη, σε κανόνες οι οποίοι χαρακτηρίζουν τις συσχετίσεις αύξησης ή μείωσης της έκφρασης του mRNA ή της πρωτεΐνης ενός γονίδιου, με βάση των γονότυπο ενός SNP. Σε αυτούς τους κανόνες θα μπορούσε να προστεθεί ήδη υπάρχουσα γνώση όσον αφόρα την λειτουργία των γονίδιων στα οποία ανήκει είτε η γενετική περιοχή ενός SNP, ή το mRNA από μια συσχέτιση, ειδικά αν αυτή είναι σε μορφή βιολογικών μονοπατιών (pathway). Ακόμη θα μπορούσαν να προστεθούν και αποτελέσματα από αλλά πειράματα, που να πρόσφεραν γνώση όσον αφορά την συσχέτιση μεταξύ αλληλόμορφων και συγκεκριμένων γενετικών χαρακτηριστικών, όπως πολύπλοκες ασθένειες. Το τελικό σύστημα θα μπορούσε να συμπληρώσει κενά στις γνώσεις μας για τα βιολογικά μονοπάτια, ή ακόμη και να απαντούσε σε ερωτήματα τα οποία χωρίς την χρήση του θα ήταν πολύ πολύπλοκο να απαντηθούν. Για παράδειγμα, πώς θα επηρεαζόταν ο κωδικοποιημένος βιολογικός μηχανισμός αν χορηγείτο στο δείγμα ένα φάρμακο που αύξανε την έκφραση ενός mRNA.

8.3 Επίλογος

Με το πέρας της μελέτης αυτής όσο αφορά τον τομέα της πληροφορικής, επιτεύχθηκε η υλοποίηση κατάλληλα διαμορφωμένου κώδικα που υποστηρίζει την χρήση grid για την επεξεργασία και ανάλυση δεδομένων.

Η χρήση του grid επιτρέπει την παραλληλοποίηση των αλγόριθμων ανάλυσης και επεξεργασίας των δεδομένων καθώς επίσης και την ελαχιστοποίηση των χρόνων εκτέλεσης αλλά και την αυτοματοποίηση των διαδικασιών υποβολής διεργασιών προς εκτέλεσης καθώς υποβάλλονται μια φορά σε κάποια ουρά στο σύστημα με κάποια σειρά προτεραιότητας και το σύστημα είναι αυτό υπεύθυνο στη συνέχεια να υποβάλει μια διεργασία προς εκτέλεση σε περίπτωση που υπάρχουν διαθέσιμοι πόροι.

Όσον αφορά τα δεδομένα δημιουργηθεί μια βάση δεδομένων με προοπτικές έρευνας για αξιολόγηση της συσχέτισης μεταξύ ενός μεγάλου αριθμού SNPs και μετάγραφων mRNA που μπορεί να βοηθήσει στην ανακάλυψη νέων φαρμάκων σε όλους τους θεραπευτικούς τομείς. Η ανακάλυψη νέων φαρμάκων μπορεί να γίνει αξιολογώντας τα λειτουργικά αποτελέσματα που μπορούν να έχουν κάποιοι παράγοντες κινδύνου ή κάνοντας εισηγήσεις για γενετικά βασιζόμενους (genetic-based) βιολογικούς δείκτες (biomarkers).

Γενικά η διεξαγωγή της μελέτης αυτής μπορεί να αποτελέσει πηγή αναφοράς για τη διεξαγωγή παρόμοιων μελετών και τα αποτελέσματα της ως δεδομένα στη φαρμακοβιομηχανία για την παραγωγή νέων φαρμάκων.

Επίσης η γενική προσέγγιση των αλγόριθμων που υλοποιήθηκαν για την διαχείριση και επεξεργασία των δεδομένων, τους καθιστά εύχρηστους και εφαρμόσιμους σε άλλες μελέτες που απαιτούν την επεξεργασία παρόμοιων τύπων δεδομένων όπως αυτών που χρησιμοποιήθηκαν.

Βιβλιογραφία

[1] Abecasis, G.R., Cookson, W.O. & Cardon, L.R, "Selection Strategies for disequilibrium mapping of quantitative traits in nuclear families," Am. J. Hum. Genet., vol. 65, pp. A245, 1999.

[2] Alberts, B., Bray, D., Johnson, A., Lewis, J., Raff, M., Roberts, K., et. al., "Βασικές Αρχές Κυτταρικής Βιολογίας: Εισαγωγή στη Μοριακή Βιολογία του Κυττάρου," 2000

[3] Anna L Dixon et. al., "A genome-wide association study of global gene expression," Nature Genetics, doi:10.1038/ng2109, 2007.

[4] Ashburner, M. et. al., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," Nature Genetics, vol. 25, pp. 25-29, 2000.

[5] Benjamini, Y. & Hochberg, "Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing," J. R. Statist. Soc. Ser. B, vol. 57, pp. 289-300, 1995.

[6] Bolstad, B.M., Irizzary, R.A., Astrand, M. & Speed, T.P., "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," Bioinformatics, vol. 19, pp. 185-193, 2003.

[7] Cheung, V.G et. al., "Natural variation in human gene expression assessed in lymphoblastoid cell," Nature Genetics, vol. 33, pp. 422-425, 2003.

[8] Devlin, B., Roeder, K. and Wasserman, "L. Genomic Control, a new approachto genetic-based association studies," Theor. Popul. Biol., vol. 60, pp. 155-166, 2001.

[9] Enrico Domenici et. al., "Allelic expression in human blood from a depression case/control collection: a database to assess functional impact of SNPs on a genome-scale that enables the identification of novel genetic-driven biomarkers," Scinovation, September 2008.

[10] Harris M.A. et al., "The Gene Ontology (GO) database and informatics resource," Nucleic Acids Res., vol. 32, pp. D258-D261, 2004.

[11] Morley, M et. al., "Genetic analysis of genome-wide variation in human gene expression," Nature, vol. 430, pp. 743-747, 2004.

[12] Scott A. Lesley, "High-Throughput Proteomics: Protein Expression and Purification in the Postgenomic World", Genomics Institute, Novartis Research Foundation, 3115 Merryfield Row, San Diego, California 92121, June 14, 2001.

[13] Shadt, E.E. et. al., "Genetics of gene expression surveyed in maize, mouse and man," Nature, vol. 422, pp. 297-302, 2003.

[14] Yan, H., Yuan, W., Velculescu, V.E., Vogelstein, B. & Kinzler, K.W. "Allelic variation in human gene expression," Science, vol. 297, pp. 1143, 2002.

[15] Ιλία Βόντα, "Εισαγωγή στις πιθανότητες και στατιστική", 2005.