Γονιδιωματική τμηματοποίηση. *Διερεύνηση μεθοδολογιών για την τμηματοποίηση ευκαρυωτικών γονιδιωμάτων με συγκεκριμένα χαρακτηριστικά*

Μεταπτυχιακή Εργασία
Μαρία Λυδία Καλαϊτζάκη

Πρόγραμμα Μεταπτυχιακών Σπουδών “Βιοπληροφική”

Τμήμα Ιατρικής, Πανεπιστήμιο Κρήτης

Foundation for Research and Technology Hellas (FORTH)

Επιβλέπων καθηγητής: Αν. Καθηγητής Χριστόφορος Νικολάου,
Τμήμα Βιολογίας, Πανεπιστήμιο

Genome Segmentation. *Investigation of methodologies for the delination of chromosomal domains with specific characteristics in eukaryotic genomes.*

Master thesis

Maria Lydia Kalaitzaki

M.Sc programme “Bioinformatics”

School of Medicine, University of Crete

Foundation for Research and Technology Hellas (FORTH)

Supervisor: Ass. Prof. Christoforos Nikolaou

Περίληψη

Οι επεκτεινόμενες τεχνολογίες αλληλούχισης επόμενης γενιάς (Next-Generation Sequencing) και η βελτιωμένη ακρίβειά τους επιτρέπουν μια νέα προσέγγιση στον τομέα, η οποία είναι η τμηματοποίηση των δεδομένων που προκύπτουν από RNA-Seq πειράματα, ώστε να προσδιοριστούν ομάδες γονιδίων, όρια γονιδίων και γονίδια που ακολοθούν κάποιο μοτίβο διαφορικής έκφρασης. Στην παρούσα εργασία, εφαρμόστηκε η μέθοδος τμηματοποίησης iSeg (Girimurugan, S.B., et al., 2018), καθώς και η μέθοδος DFOE, η οποία αναπτύχθηκε από την ομάδα μας, με στόχο την τμηματοποίηση γονιδιωμάτων με βάση δεδομένα γονιδιακής έκφρασης και τον εντοπισμό ευρύτερων περιοχών αυξημένης έκφρασης σε μία διάσταση, δηλαδή σε γραμμικά χρωμοσώματα. Πραγματοποιήθηκε RNA-seq ανάλυση σε δείγματα ολικού αίματος 200 ατόμων. Τα 142 άτομα ήταν άτομα τα οποία είχαν διαγνωσθεί στο παρελθόν με ΣΛΕ και τα 58 άτομα ήταν υγιή. Το πρωταρχικό συμπέρασμα, που προέκυψε από την εφαρμογή της μεθόδου DFOE στο σύνολο δεδομένων ασθενών με ΣΕΛ, ήταν η σημαντικά μεγαλύτερη κάλυψη του γονιδιώματος από περιοχές αυξημένης έκφρασης στους ασθενείς σε σχέση με τους υγιείς μάρτυρες (controls). Η κάλυψη αυτή ήταν επιπλέον θετικά σχετιζόμενη με την ενεργότητα της ασθένειας, κάτι που υποδεικνύει ότι εκτός από αυξημένα επίπεδα έκφρασης, οι ασθενείς με ΣΕΛ τείνουν να εμφανίζουν υπερ-έκφραση σε εστιασμένες περιοχές του γονιδιώματος. Κατά την εκτέλεση της μεθόδου εύρεσης στατιστικά σημαντικών γονιδιωματικών τμημάτων που προέκυψαν από την εκτέλεση του αλγορίθμου DFOE, καταφέραμε να απομονώσουμε σημαντικές περιοχές τόσο στους υγιείς όσο και στους ασθενείς. Από την περαιτέρω ανάλυση με GprofileR, παρατηρήθηκε ότι ένα μεγάλο μέρος των λειτουργίων, που προέκυψαν να σχετίζονται με τα γονίδια των ευρεθέντων τμημάτων, σχετίζονται με την ασθένεια ΣΛΕ. Τα αποτελέσματα που προέκυψαν από την παραπάνω εργασία φαίνεται να χαρακτηρίζονται από αφθονία πληροφοριών. Περαιτέρω ανάλυση των παραπάνω δεδομένων θα μπορούσαν να οδηγήσουν τόσο στην αναγνώριση σημάτων που σχετίζονται με ασθένειες όσο και στην ανακάλυψη νέων βιοδεικτών.

Abstract

Expanded Next-Generation Sequencing technologies and their improved precision allow for a new approach in the field, which is segmentation of data derived from RNA-Seq experiments, to identify gene clusters, gene boundaries and genes follow a pattern of differential expression. In the present work, the iSeg segmentation method (Girimurugan, SB, et al., 2018), as well as the DFOE method, developed by our team, were used to segment genomes based on gene expression data and identify wider regions. increased expression in one dimension, i.e. in linear chromosomes. RNA-seq analysis was performed on whole blood samples of 200 individuals. 142 individuals were previously diagnosed with TFE and 58 were healthy. The primary conclusion, which came from the application of the DFOE method to the SLE patient data set, was the significantly greater genome coverage of areas of increased expression in patients compared to healthy controls. This coverage was also positively correlated with disease activity, indicating that in addition to elevated expression levels, patients with SLE tend to be over-expressing in targeted regions of the genome. When performing the method of finding statistically significant genomic segments resulting from the DFOE algorithm, we were able to isolate significant regions in both healthy and patients. From further analysis with GprofileR, it was observed that a large proportion of the functions found to be related to the genes of the found fragments are related to TFE disease. The results obtained from the above work appear to be characterized by an abundance of information. Further analysis of the above data could lead to both the identification of disease-related signals and the discovery of new biomarkers.

Ευχαριστίες

Αρχικά, θα ήθελα να εκφράσω τις ευχαριστίες μου στον Χριστόφορο Νικολάου, για την εμπιστοσύνη που μου έδειξε κατά την διάρκεια του μεταπτυχιακού αλλά και ως προπτυχιακή φοιτήτρια, για τις γνώσεις που μου μετέδωσε τόσο στην κλάδο της Υπολογιστικής Βιολογίας, όσο και για ατάκες ταινιών αλλά και για την ιστορία της Liverpool.

Επίσης θα ήθελα να ευχαριστήσω ιδιαιτέρως τους καθηγητές και τους ερευνητές που συμμετείχαν στο πρόγραμμα και ιδιαιτέρως τον Παύλο Παυλίδη για την εμπιστοσύνη του και την στήριξη του σε εμένα.

Ευχαριστώ ιδιαιτέρως τα παιδιά από το εργαστήριο, τον Αντώνη που ήταν δίπλα μου σε ότι τον χρειαζόμουν, τον Αιμίλιο που μας έκανε να χαμογελάμε κάθε μέρα, τον Στέλιο, την Μυρσίνη, τον Κωνσταντίνο , την Ελένη, την Ίλια, την Σοφία και την Σοφία.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου, Ρένα, Γιάννη, Μάνο, Κάσσυ, Υακίνθη, που είναι δίπλα μου σε ότι απόφαση πάρω και τους φίλους μου, Μηνά, Φαίδωνα, Ειρήνη, Κατερίνα, Μαρία, Ανδρέα, Τίτο, Μάρια, Κώστα, Δημήτρη, Αλέξανδρο, Άρια, Ναταλία, Μαίρη, Εύα, Παναγίωτα, που με ανέχονται και με ηρεμούν.

Εισαγωγή

Οι γονιδιωματικές δοκιμές υψηλής απόδοσης, όπως οι μικροσυστοιχίες (Microarrays) και η αλληλούχιση επόμενης γενιάς (Next-generation sequencing), είναι ισχυρά εργαλεία για τη μελέτη γενετικών και επιγενετικών λειτουργικών στοιχείων σε κλίμακα γονιδιώματος (Consortium EP. et al. , 2012). Η εμφάνιση γονιδιωματικών δεδομένων υψηλής πυκνότητας και μεγάλου όγκου δημιούργησε την ανάγκη εύρεσης εργαλειών για τη σύνοψη μεγάλων συνόλων δεδομένων και την αποτελεσματική ανάλυση τους. Η ανάλυση και ο σχολιασμός του γονιδιώματος είναι ένα σημαντικό ζήτημα στον κλάδο της Βιολογίας, το οποίο έχει αντιμετωπιστεί με μεθόδους πρόβλεψης γονιδίων και χειρωνακτικά πειράματα που απαιτούν βιολογική γνώση και εμπειρία. Οι επεκτεινόμενες τεχνολογίες αλληλούχισης επόμενης γενιάς (Next-Generation Sequencing) και η βελτιωμένη ακρίβειά τους επιτρέπουν μια νέα προσέγγιση στον τομέα, η οποία είναι η τμηματοποίηση των δεδομένων που προκύπτουν από RNA-Seq πειράματα, ώστε να προσδιοριστούν ομάδες γονιδίων, όρια γονιδίων και γονίδια που ακολοθούν κάποιο μοτίβο διαφορικής έκφρασης. Υπάρχουσες προσεγγίσεις στο πρόβλημα της τμηματοποίησης (segmentation) έχουν αναπτυχθεί με σκοπό να αναλύσουν δεδομένα μεγαλύτερης διακριτικής ικανότητας που προέρχονται από πειράματα ChIPSeq ή άλλων πρωτοκόλων που παράγουν δεδομένα σε μεγαλύτερη έκταση. Οι περισσότερες δημοσιευμένες μέθοδοι αποσκοπούν στον προσδιορισμό περιοχών της χρωματίνης με συγκεκριμένα χαρακτηριστικά όπως προκύπτουν από επιγενετικά δεδομένα (Ernst J, et. al, 2017)(Hon G, et. al, 2008)(Hoffman MM, et. al ,2012).

Στην παρούσα εργασία, εφαρμόστηκε η μέθοδος τμηματοποίησης iSeg (Girimurugan, S.B., et al., 2018), καθώς και η μέθοδος DFOE, η οποία αναπτύχθηκε από την ομάδα μας, με στόχο την τμηματοποίηση γονιδιωμάτων με βάση δεδομένα γονιδιακής έκφρασης και τον εντοπισμό ευρύτερων περιοχών αυξημένης έκφρασης σε μία διάσταση, δηλαδή σε γραμμικά χρωμοσώματα. Επιπρόσθετα, εργαστήκαμε απόκλειστικά στην ανάλυση ενός πρόσφατα δημοσιευμένου, εκτεταμένου και υψηλής ποιότητας συνόλου δεδομένων που έχει προκύψει από το μεταγραφικό profiling ασθενών με Συστηματικό Ερυθηματώδη Λύκο (ΣΕΛ) (Panousis et al, 2019). Το συγκεκριμένο σύνολο δεδομένων μας παρείχε τα πλεονεκτήματα του στατιστικού μεγέθους (~200 δείγματα) καθώς και της δυνατότητας ερμηνείας των αποτελεσμάτων μας στο πλαίσιο μιας καλά μελετημένης παθολογικής κατάστασης. Αυτό γίνεται σαφές από το τελευταίο μέρος της παρούσας εργασίας, όπου οι εξαχθείσες υπερ-εκφραζόμενες περιοχές μπόρεσαν να αναλυθούν σε επίπεδο λειτουργίας και να αντιπαρατεθούν με κλινικά και γενετικά δεδομένα τα οποία είναι διαθέσιμα για τη συγκεκριμένη ασθένεια σε μεγάλο βαθμό.

Η αδυναμία αυτοανοχής είναι μία από τις κύριες αιτίες εκδήλωσης των αυτοάνοσων νοσημάτων.Οι αυτοάνοσες ασθένειες είναι μια από τις βασικές αιτίες νοσηρότητας και θνησιμότητας σε όλο τον κόσμο. Τα αυτοάνοσα νοσήματα θεωρείται ότι προκύπτουν ως συνδυασμός γονιδιακής προδιάθεσης σε πολλαπλούς γενετικούς τόπους, περιβαλλοντικών παραγόντων, όπως το κάπνισμα, η έκθεση σε παθογόνα και τα ορμονικά επίπεδα και άλλων στοχαστικών γεγονότων και έχουν έως αποτέλεσμα την ανοσολογική απόκριση του οργανισμού σε εαυτά αντιγόνα (John D. Rioux et al., 2005).

Μία από τις σοβαρότερες αυτοάνοσες ασθένειες είναι ο Συστηματικός Ερυθηματώδης Λύκος (ΣΛΕ). Ο ΣΛΕ είναι μία πολυπαραγοντική αυτοάνοση διαταραχή, της οποίας η αιτιολογία δεν έχει κατανοηθεί ακόμα πλήρως. Οι αιτίες της νόσου παραμένουν μέχρι σήμερα άγνωστες. Τα συσσωρευμένα στοιχεία τα τελευταία χρόνια έχουν δείξει ότι προκαλούνται από ένα συνδυασμό γενετικών και περιβαλλοντικών παραγόντων καθώς επίσης και από τον τρόπο ζωής και την εθνικότητα (Mayami Sengupta et al., 2011). Ο ΣΛΕ μπορεί να επηρεάσει οποιοδήποτε μέρος του σώματος αλλά το συνηθέστερο είναι να προκαλεί βλάβες στο δέρμα, την καρδιά, τις αρθρώσεις, στο αιμοποιητικό και το νευρικό σύστημα. Η εκδήλωση της ασθένειας μπορεί να εμφανιστεί σε οποιαδήποτε ηλικία αλλά είναι

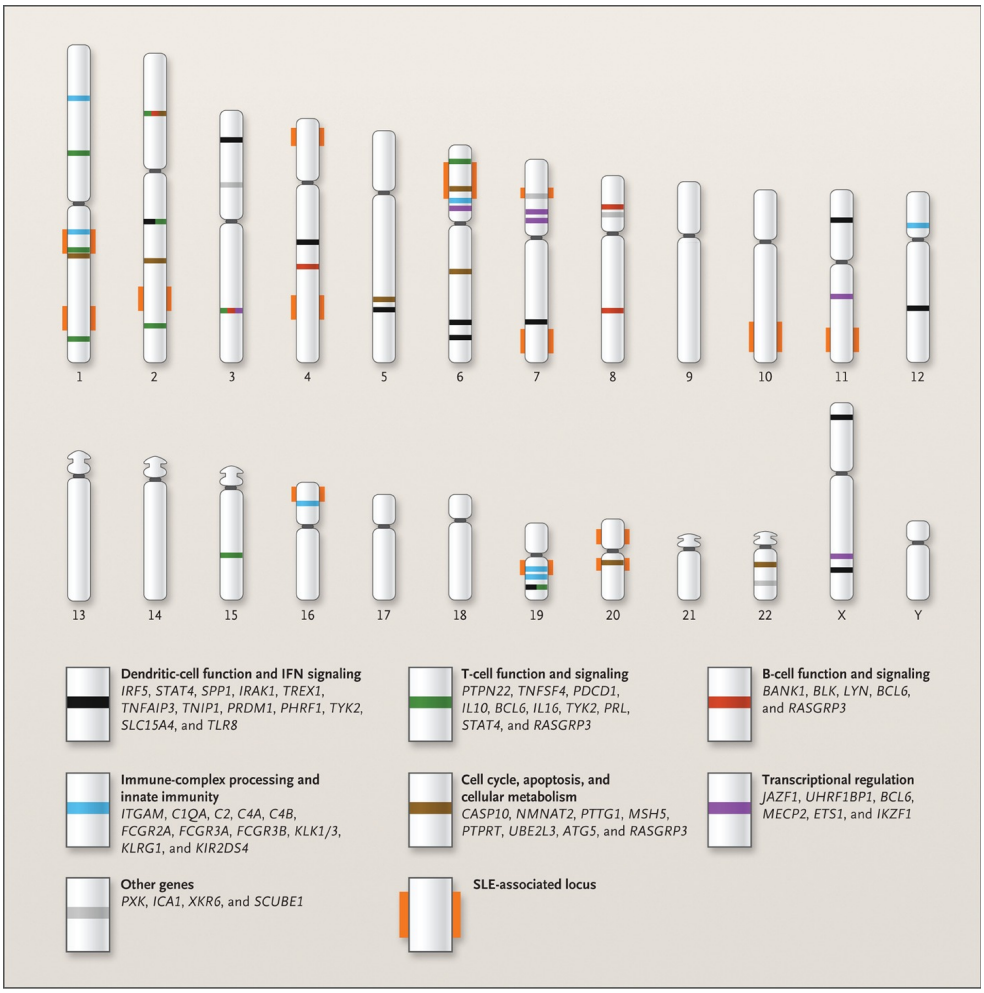
συχνότερη στις γυναίκες ηλικίας 20-45 ετών με αναλογία εμφάνισης γυναίκες προς άντρες περίπου 9:1, υποδηλώνοντας ότι οι ορμόνες είναι κύρια αιτία ανάπτυξης της νόσου (Danchenko N et. Al, 2006)

Βασικό χαρακτηριστικό του ΣΛΕ είναι ότι το ανοσοποιητικό σύστημα του πάσχοντα οργανισμού είναι «υπερδραστήριο» και έτσι παράγει μεταξύ των άλλων αντισώματα εναντίον της διπλής έλικας του DNA του. Τα αντισώματα αυτά είναι αυτο-αντισώματα, κυρίως αντι-πυρηνικά, τα οποία αναγνωρίζονται ως ξένα στοιχεία του ίδιου του εαυτού και είναι γνωστά ως “ANAs”(anti-nuclear antibodies). Τα “ANAs” εντοπίζονται και σε υγιή άτομα, γεγονός το οποίο υποδηλώνει την παρουσία τους και σε άλλους παράγοντες που προκαλούν κάποια ασθένεια. Η παραγωγή των αυτοαντισωμάτων καθώς και άλλων ουσιών οδηγεί στην «επίθεση» του ανοσοποιητικού συστήματος στα διάφορα όργανα του σώματος και με αυτόν τον τρόπο δημιουργείται η φλεγμονή. Πρέπει να σημειωθεί ότι μια ποικιλία των αυτοαντισωματών είναι μοναδικά στα άτομα με Συστηματικό Ερυθηματώδη Λύκο. Μια ελκυστική εικασία είναι να θεωρήσουμε αυτά τα «μοναδικά» αντισώματα ως οδηγούς ασθενειών. (Tsokos, 2011) (Kaul et al., 2016).

Η νόσος έχει χαρακτηριστεί ως συστηματική επειδή η μη φυσιολογική ανοσοαπόκριση που στοχεύει τους ιστούς κάθε ατόμου είναι πολύ ευρεία και επηρεάζει σχεδόν όλα τα όργανα. Μεταξύ άλλων, οι νεφροί, οι πνεύμονες, η καρδιά, ο εγκέφαλος και το δέρμα είναι υποψήφιοι ιστοί για σοβαρή βλάβη. Μερικοί ασθενείς με ΣΛΕ αναπτύσσουν νεφρίτιδα, νευρογνωστικά ελαττώματα και άλλα ελαττώματα. Η ετερογένεια της νόσου είναι εμφανής ακόμη και σε βλάβες των οργάνων που αναπτύσσονται για κάθε ασθενή. Μέχρι σήμερα είναι αδύνατο να προβλεφθεί η ευαισθησία κάθε οργάνου σε κάθε ασθενή Lupus. Η σοβαρότητα της νόσου είναι διαφορετική μεταξύ των ατόμων. Ένα κοινό έδαφος είναι η ύπαρξη δύο φάσεων της νόσου: μια ενεργή φάση ή «φλεγμονή» όπου εμφανίζονται τα περισσότερα συμπτώματα και μια αδρανή φάση (Tsokos, G. C., 2011) (Kaul et al., 2016). Σε ακραίες περιπτώσεις μια μεμονωμένη γονιδιακή μετάλλαξη μπορεί να οδηγήσει στην ανάπτυξη της νόσου του Lupus.

Οι περισσότεροι πολυμορφισμοί ενός νουκλεοτιδίου (SNPs) που σχετίζονται με το ΣΛΕ εμπίπτουν σε μη κωδικές περιοχές DNA γονιδίων που σχετίζονται με την ανοσολογική απόκριση(Harley JB, et. al, 1998). Ορισμένα γονίδια έχουν συσχετιστεί με διάφορες αυτοάνοσες νόσους (π.χ. STAT4 και PTPN22 με ρευματοειδή αρθρίτιδα και διαβήτη) και άλλα γονίδια φαίνεται να αυξάνουν τον κίνδυνο για ΣΛΕ. Ορισμένα SNPs που συνδέονται με ΣΛΕ έχουν ταυτοποιηθεί για γονίδια των οποίων τα προϊόντα μπορεί να συνεισφέρουν στην μη φυσιολογική λειτουργία Τ-κυττάρων στο ΣΛΕ (CD3-ζ και PP2Ac). Μια πρόσφατη μελέτη επιβεβαίωσε μερικούς από αυτούς τους συνδυασμούς και αναγνώρισε τα TNIP1, PRDM1, JAZF1, UHRF1BP1 και IL10 ως σημεία κινδύνου για ΣΛΕ (Gateva V, et. al, 2009). Αν και αυτά τα ευρήματα είναι ελπιδοφόρα, οι τόποι που εντοπίστηκαν μέχρι στιγμής μπορούν να αντιπροσωπεύουν μόνο περίπου το 15% την κληρονομικότητα του ΣΛΕ (Manolio TA, et. al, 2009). Επιπλέον, ένας αλλοιωμένος αριθμός αντιγράφων ορισμένων γονιδίων, όπως τα C4, FCGR3B, και TLR7, έχει συνδεθεί με την έκφραση της νόσου.

Στην παρούσα εργασία, πραγματοποιήθηκε RNA-seq ανάλυση σε δείγματα ολικού αίματος 200 ατόμων. Τα 142 άτομα ήταν άτομα τα οποία είχαν διαγνωσθεί στο παρελθόν με ΣΛΕ και τα 58 άτομα ήταν υγιή. Οι ασθενείς με ΣΛΕ είχαν μεγάλη ποικιλία κλινικών συμπτωμάτων. Εκτός από τα δεδομένα που προέκυψαν απο την ανάλυση του γονιδιακού προφίλ έκφρασης του κάθε ασθενή, υπήρχαν στην διαθεσή μας και επιπρόσθετα δεδομένα (π.χ παρουσία ρινικών προβλημάτων, μετρήσεις κυτταρικών πληθυσμών στο αίμα).



Chromosome Loci and Genes Associated with SLE.
(source:Tsokos et. al,2011)

3. Υλικά και Μέθοδοι

3.1 Συλλογή δειγμάτων και RNA sequencing

Η συλλογή των δειγμάτων, το RNA sequencing και η χαρτογράφιση είχαν ήδη εκτελεστεί. Περαιτέρω πληροφορίες για τα στοιχεία των ασθενών, την συλλογή των δειγμάτων και την χαρτογράφιση περιγράφονται από Panousis et al.(Panousis et al.,2018). Το εναρκτήριο υλικό για αυτή την μελέτη ήταν τα bam alignment αρχεία που προέκυψαν από την διαδικασία της χαρτογράφησης.

3.2 Ποσοτικοποίηση αριθμού αναγνώσεων ανά μετάγραφο(Fragment summarization)

Για την εξαγωγή των αρχικών/ακατέργαστων μετρήσεων (raw counts) και την ποσοτικοποίηση των επιπέδων έκφρασης ενός σετ από γονίδια στον άνθρωπο, χρησιμοποιήθηκε το FeatureCounts25, με βάση τον τελευταίο χαρακτηρισμό GENCODE v1526. Ένα κομμάτι μετρήθηκε σε περίπτωση κάποιας επικάλυψης με κάποιο εξώνιο και οι μετρήσεις (counts) ομαδοποιήθηκαν με βάση το χαρακτηριστικό ‘gene_name’ από το annotation file. Τα κομμάτια που χρησιμοποιήθηκαν για την σύνοψη (summatization) ήταν εκείνα των οποιών και τα δύο τους άκρα χαρτογραφήθηκαν (mapped) με επιτυχία. Τα κομμάτια τα οποία ήταν χιμαιρικά, είχαν επικάλυψη με γονίδια, δεν ήταν μοναδικά χαρτογραφημένα, ή δεν είχαν κάποιες “αναγνώσεις αλληλουχιών” που είχαν σημειωθεί ως διπλότυπα απορρίφθηκαν.

3.3 Τμηματοποίηση γονιδιώματος (Genome Segmentation)

3.3.1 iSeg

Το iSeg είναι ένα αποτελεσματικό εργαλείο γονιδιωματικής τμηματοποίησης (ή κατάτμησης) το οποίο έχει ως στόχο την εύρεση σημαντικών περιοχών σε προφίλ γονιδιακής έκφρασης. Το iSeg προσφέρει υπολογιστική αποτελεσματικότητα, η οποία είναι χρήσιμη στην ανάλυση προφίλ γονιδίων υψηλής απόδοσης (Girimurugan S.B. et al., 2018).

Το πρώτο βήμα του αλγορίθμου iSeg περιλαμβάνει το σάρωμα του προφίλ της γονιδιακής έκφρασης χρησιμοποιώντας ένα παράθυρο του οποίου το μήκος καθορίζεται από τον χρήστη έτσι ώστε να βρεθούν περιοχές οι οποίες είναι σημαντικές. Στην συγκεκριμένη εργασία το μικρότερο μέγεθος (Wmin) του παραθύρου ήταν 1 και το μεγαλύτερο ήταν 500 (Wmax) . Έπειτα, χρησιμοποιείται ένα αρχικό όριο σημαντικότητας (initial significance level) για τον προσδιορισμό της σημαντικότητας των περιοχών.

Για την αρχική σάρωση είναι απαραίτητο να οριστεί μία τιμή p-value κάτω από το αρχικό όριο σημαντικότητας. Έπειτα από κάθε σάρωση τα παράθυρα αυξάνονται εκθετικά χρησιμοποιώντας έναν συντελεστή ισχύος (power factor) και οι περιοχές κατατάσσονται από την μικρότερη στην μεγαλύτερη με βάση τα p-values και την βοήθεια του αλγορίθμου Balanced Binary Trees (BBT) (Figure 2). Στην συνέχεια οι περιοχές συγχωνεύονται/συρρικνώνονται σε περίπτωση που το p-value μιας περιοχής είναι

μικρότερο απο το p-value των αρχικών περιοχών. Τέλος, χρησιμοποιείται το ποσοστό ψευδώς θετικών αποτελεσμάτων (FDR : False Discovery Rate) μέσω της διαδικασίας Benjamini-Hochberg, ώστε να γίνει μια εκκαθάριση στις περιοχές που απομονώθηκαν, επιλέγοντας τις περιοχές που είναι οι περισσότερο σημαντικές.

Ο τύπος αρχείου που δέχεται ο αλγόριθμος iSeg είναι bedGraph, το οποίο αποτελείται από τέσσερις κολώνες (χρωμόσωμα, αρχή τμήματος, τέλος τμήματος, score).

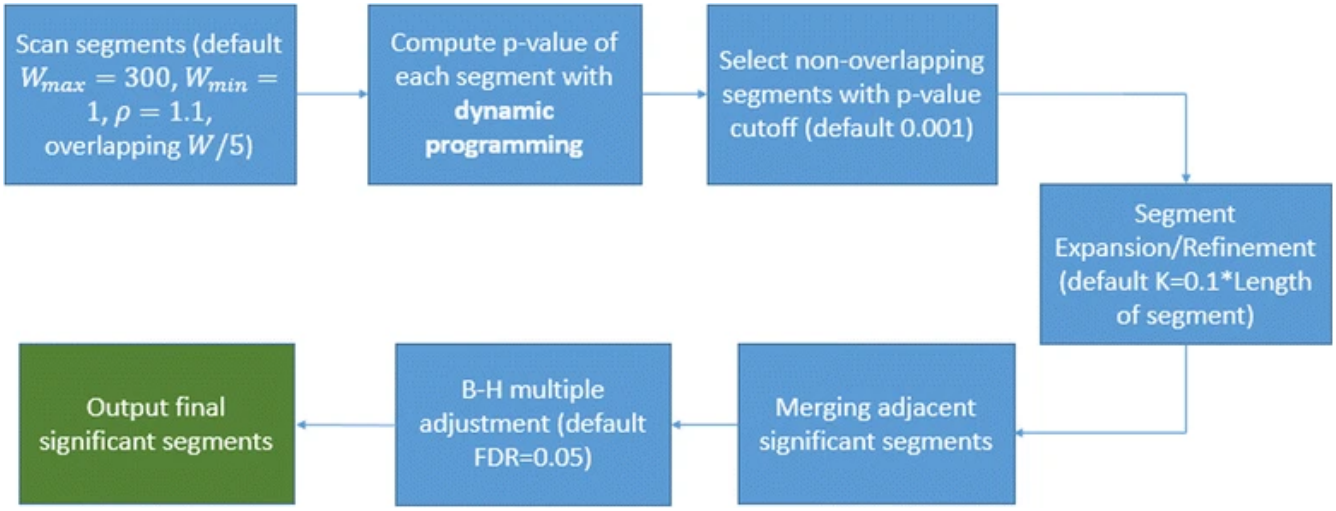


Figure 1:A schematic illustration of the workflow of iSeg.(Figure source: Girimurugan S.B. et al.,2018)

```
procedure SelectSignificantSegments
initialize BBT2 // BBT2 is empty at the beginning
while(BBT1 not empty)
        S = top ranked segment in BBT1 (smallest p-value among all segments in BBT1)
        delete S from BBT1
        l = left boundary of S
        r= right boundary of S
        if(checkoverlap (BBT2, l, r) == FALSE) // no overlapping
                insert pair(l, r) into BBT2
                insert S to set SS
```

Figure 2: Detecting overlapping segments and updating significant segments using coupled balanced binary trees.(Figure source: Girimurugan S.B. et al.,2018)

3.3.2 Ο αλγόριθμος DFOE

Για την δημιουργία του αλγορίθμου DFOE εφαρμόστηκε μια μέθοδος βασισμένη σε αμέρολητο αναδρομικό διαχωρισμό όπως περιγράφεται Hothorn,T. et. al (Hothorn,T. et. al, 2006). Τα δεδομένα μέτρησης έκφρασης γονιδίων (ως log2FC) χρησιμοποιήθηκαν ως τιμές και οι γονιδιωματικές συντεταγμένες τους ως διακριτή μεταβλητή "χρονικού τύπου". Ένας προσαρμοσμένα αλγόριθμος στην R γράφτηκε με τη χρήση της συνάρτησης "breakpoints" στην R από το πακέτο "strucchange"(Zeileis,A. et. al, 2001). Η συνάρτηση εκτελεί τμηματοποίηση του γονιδιώματος βάσει ενός F-test (Chow Test) που εφαρμόζεται σε συνεχή γραμμικά μοντέλα. Μόλις καθοριστούν τα σημεία διακοπής (breakpoints), ο αλγόριθμος δημιουργεί μια πλήρη τμηματοποίηση του γονιδιώματος σε διακεκριμένες περιοχές, οι οποίες περιγράφονται από α) τον αριθμό των γονιδίων που περιέχουν και β) από τη μέση βαθμολογία έκφρασής τους. Ένα αυθαίρετο κριτήριο ενός αριθμού ολικών γονιδίων μεγαλύτερου ή ίσο του 50 χρησιμοποιήθηκε για να εντοπίσει σημαντικές περιοχές εστιακής υπερ-έκφρασης (DFOE). Οι παραπάνω περιοχές χρησιμοποιήθηκαν για τη δημιουργία διμερών δικτύων. Η διαδικασία τμηματοποίησης από τον αλγόριθμο περιγράφεται στο παρακάτω διάγραμμα.

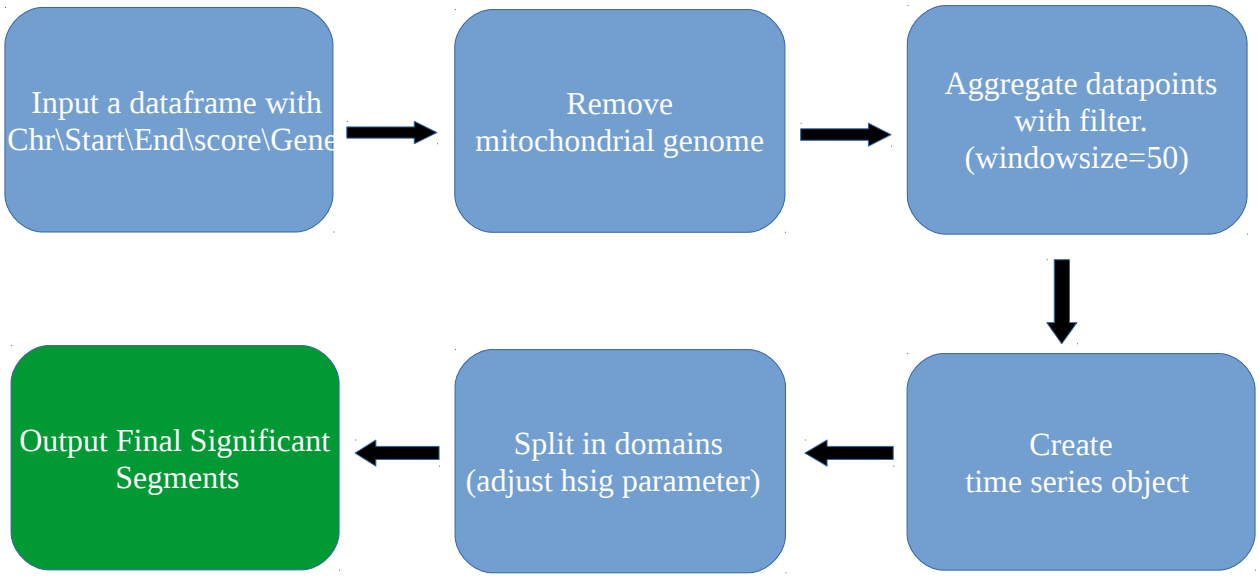


Figure : A schematic illustration of the workflow of DFOE algorithm.

3.4 Κατάταξη ανά χρωμόσωμα

Η κατάταξη κατα αύξουσα σειρά συντεταγμένων ανά χρωμόσωμα (sorting) έγινε μέσω της χρήσης της εντολής “sortBed” των bedtools.

3.5 Εύρεση συμπληρωματικών τμημάτων γονιδιώματος

Η εύρεση των συμπληρωματικών τμημάτων (complement) γονιδιώματος έγινε εξίσου με την χρήση bedtools και πιο συγκεκριμένα με την εντολή “complement”.

3.6 (Ταυτοποίηση) γονιδίων στα ευρεθέντα τμήματα και στα συμπληρωματικά

Η εύρεση των γονιδίων των ευρεθέντων περιοχών και των συμπληρωματικών τους έγινε με την χρήση bedtools και την εντολή intersect.

3.7 Ανάλυση διαφορικής έκφρασης γονιδίων

Τα διαφορικώς εκφραζόμενα γονίδια (DEGs) βρέθηκαν/ταυτοποιήθηκαν με την χρήση του πακέτου MDSeq (Ran et al.,2017). Αρχικά, κανονικοποιήθηκαν οι αρχικές/ακατέργαστες μετρήσεις (raw counts) χρησιμοποιώντας relative log expression (RLE)(Anders et al.,2010). Στην συνέχεια κατασκευάστηκε ένας πίνακας (design matrix) βασισμένος στις ομάδες που συγκρίθηκαν. Για την ανάλυση, τα υγιή άτομα τέθηκαν ως το control group και τα ΣΛΕ άτομα αποτέλεσαν το test group. Τέλος, για οποιαδήποτε αξιολόγηση σημαντικότητας χρησιμοποιήθηκαν η διορθωμένη τιμή p-value (q-value) και ο λογάριθμος βάσης-2 του fold change (log2FC). Τα γονίδια που εκτιμήθηκαν ως DEGs ήταν αυτά που είχαν q-value μικρότερο ή ίσο με 0.05 και απόλυτη τιμή log2FC μεγαλύτερη ή ίση με 0.5.

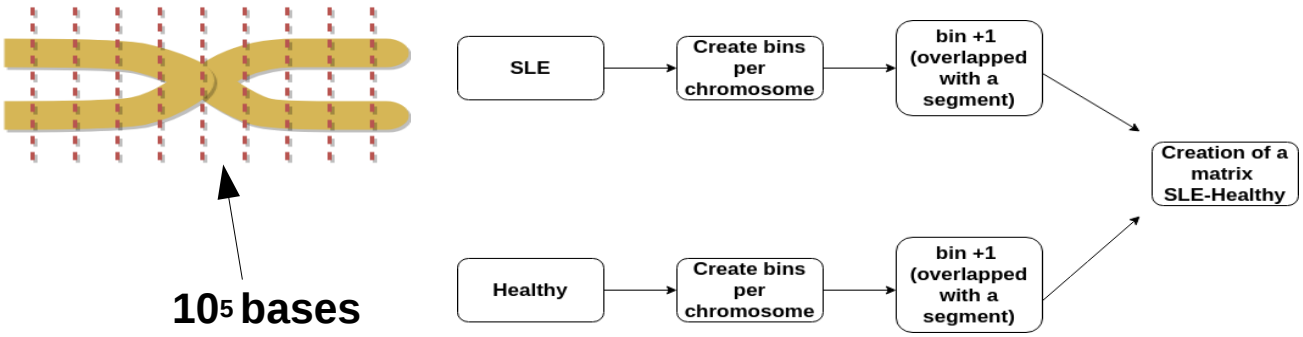
3.8 Εύρεση στατιστικά σημαντικών γονιδιωματικών τμημάτων

Δημιουργήθηκε μια λίστα, η οποία περιέχει 24 υπολίστες όπου η κάθε υπολίστα αντιστοιχεί σε ένα χρωμόσωμα και το μήκος της κάθες υπολίστας είναι ίσο με το μήκος του αντίστοιχου χρωμοσώματος προς 10^5. Κάθε σημείο (bin) της υπολίστας αντιστοιχεί σε μια συγκεκριμένη θέση στο χρωμόσωμα μήκους 10^5 βάσεις. Για κάθε bin υπολογίστηκε η επικάλυψη για κάθε τμήμα που προέκυψε από την μέθοδο τμηματοποίησης με τον αλγόριθμο DFOE. Σε περίπτωση αλληλεπικάλυψης προστίθεται στο bin +1. Με τον παραπάνω τρόπο κατασκευάζονται δύο λίστες, μια για τα υγιή άτομα και μια για τα άτομα με ΣΛΕ. Κάθε λίστα διαιρείται με τον αριθμό των ατόμων της κάθε κατηγορίας. Στην συνέχεια δημιουργήθηκε ένας πίνακας για κάθε λίστα όπου κάθε κολώνα του πίνακα αντιστοιχεί σε ένα χρωμόσωμα και οι σειρές αντιστοιχούν στα bins που είχαν δημιουργηθεί στις λίστες. Έπειτα δημιουργήθηκε ένας τελικός πίνακας από την αφαίρεση των δυο αρχικών πινάκων (SLE-Healthy). Το επόμενο βήμα ήταν η εύρεση των 5% και 95% ποσοστιμορίων για κάθε χρωμόσωμα μέσω της εντολής quantile στην R. Η γενική συνάρτηση quantile παράγει δείγμα ποσοτήτων που αντιστοιχούν στις δεδομένες πιθανότητες.

```
> qs
  chr1       chr2      chr3       chr4       chr5        chr6       chr7        chr8        chr9 chr10      chr11
5%     0 0.00000000 0.4310345 -0.7557994 -0.3037618 -0.02257053 -0.7413793 -0.68965517 -0.25862069     0 0.00000000
50%    0 0.01724138 0.6724138 -0.1739812  0.5307210  0.34482759  0.1551724 -0.28087774 -0.07774295     0 0.00000000
95%    0 0.17241379 0.8448276  0.3965517  0.5636364  0.46551724  0.3275862 -0.00815047  0.48589342     0 0.05454545
     chr12      chr13      chr14      chr15      chr16      chr17       chr18     chr19      chr20      chr21
5%  -0.01912226 -0.9137931 -0.2156740 -0.3717868 -0.3583072 0.01724138 -0.98275862 0.2586207 -0.9310345 -0.6015674
50%  0.56363636 -0.4752351  0.7021944 -0.0338558  0.5000000 0.20689655 -0.66959248 0.7241379 -0.1721003 -0.6015674
95%  0.92727273 -0.2463950  0.8094044  0.4542320  0.6542320 0.43103448  0.05454545 0.7586207  0.1724138 -0.3103448
     chr22       chrX chrY
5%  -0.04796238 -0.3266458    0
50%  0.44294671 -0.1830721    0
95%  0.55203762  0.4341693    0
```

Παραπάνω βλέπουμε το αποτέλεσμα που προκύπτει από την εντολή quantile. Σκοπός αυτης τη εντολής ήταν να βρεθούν οι περιοχές που είχαν τις μεγαλύτερες διαφορές μεταξύ υγιών και ασθενών. Αναμένουμε τα bins με τιμές μεγαλύτερες από το 95% ποσοστιμόριο να είναι περιοχές που βρίσκονται σε τμήματα πιο συχνά σε ασθενείς από ότι σε υγιείς. Αντίστοιχα τα bins με τιμές μικρότερες από το 5% ποσοστιμόριο αναμένουμε την ανάποδη τάση. Σε περίπτωση που ένα χρωμόσωμα τμηματοποιείται πιο έντονα είτε στους ασθενείς είτε στους υγιείς στο σύνολο του, το στατιστικό θα είναι ομόσημο για όλα τα bins του εκάστοτε χρωμοσώματος : Σε τέτοιες περιπτώσεις κρατάμε μόνο το ένα άκρο της κατανομής των τιμών όπως φαίνεται παρακάτω.

5%	95%	Significant Healthy	Significant SLE
-	+	<=5%	>=95%
+	+		>=95%
-	-	<=5%	



Διαδικασία εύρεσης στατιστικά σημαντικών γονιδιωματικών τμημάτων

3.9 Λειτουργική ανάλυση με GprofileR

Η λειτουργική ανάλυση πραγματοποιήθηκε με τη βοήθεια του εργαλείου gProgileR, το οποίο έχει προσβάση σε δεδομένα από διαφορετικές βάσεις δεδομένων και πραγματοποιεί υπεργεωμετρικό τεστ και διορθώσεις για πολλαπλά τεστ με αποτέλεσμα την εύρεση στατιστικά σημαντικών εμπλουτισμένων λειτουργικών οντολογιών της παρεχόμενης λίστας γονιδίων. Στην παρούσα εργασία η ανάλυση πραγματοποιήθηκε στα εξής μονοπάτια : GO:KEGG, GO:MF(Molecular Function).

4. Αποτελέσματα

Μέθοδος τμηματοποίησης iSeg

Από την εκτέλεση του εργαλείου γονιδιωματικής τμηματοποίησης, iSeg, παράχθηκαν 200 νέα bedGraph αρχεία, όπου κάθε αρχείο αντιστοιχεί σε ένα από τα άτομα που συμμετείχαν στην έρευνα. Όπως φαίνεται στα Figures 2,3 τα τμήματα που προέκυψαν από αυτή την τμηματοποίηση ήταν πολλά σε αριθμό και μικρά σε μήκος τόσο στους υγιείς όσο και στους ασθενείς. Και στα δύο διαγράμματα (Figures 2,3) φαίνεται ότι οι ασθενείς έχουν μικρότερα και λιγότερα σε αριθμό τμήματα από τους υγιείς, χωρίς όμως να υπάρχει κάποια σημαντική διαφορά, φαινόμενο το οποίο μπορεί να εξηγηθεί από τον τρόπο που γίνεται η τμηματοποίηση με το εργαλείο iSeg. Το παραπάνω οδηγεί στο συμπέρασμα ότι καταλήγουμε με μεγάλο αριθμό τμημάτων τα οποία ουσιαστικά δεν δημιουργούν ομαδοποίηση των γονιδίων, γεγονός που επιβεβαιώνεται στο Figure 3. Το παραπάνω φαινόμενο συμπληρώνει το γεγόνος ότι τα γονίδια που βρίσκονται στα τμήματα που προέκυψαν είναι σχέδον ίσα σε αριθμό και όμοια με τα γονίδια που βρίσκονται στα συμπληρωματικά τμήματα (Figure 1). Πιο συγκεκριμένα, τα τμήματα που προέκυψαν από την μέθοδο τμηματοποίησης έχουν γονίδια με μετρήσεις απολύτως συγκρίσιμες με τις μετρήσεις των γονιδίων που βρίσκονται στα συμπληρωματικά τμήματα. Το γεγονός αυτό οδηγεί στο συμπέρασμα ότι δεν έχουν δημιουργηθεί τμήματα που να έχουν προφίλ υψηλής ή χαμηλής έκφρασης γονιδιων. Παράλληλα όταν ανάλυθηκε το μήκος των τμημάτων άνα χρωμόσωμα, παρατηρήθηκε ότι στα πρισσότερα χρωμοσώματα τα τμήματα των υγιών ατόμων ήταν μεγαλύτερα σε μήκος από τα τμήματα των ατόμων με ΣΛΕ.

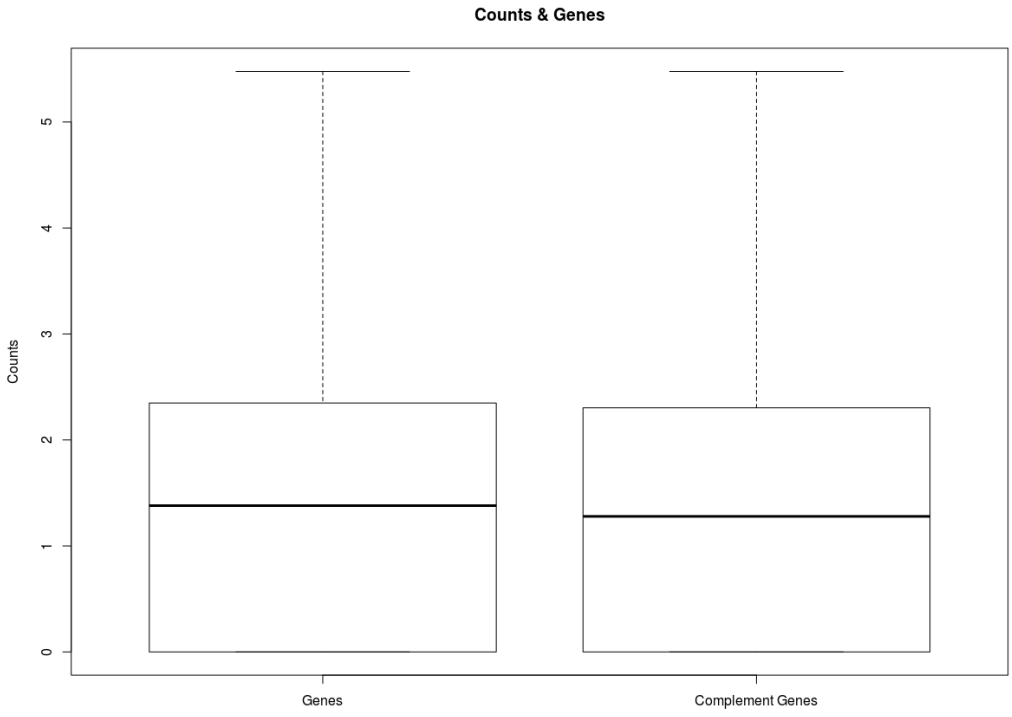


Figure 1

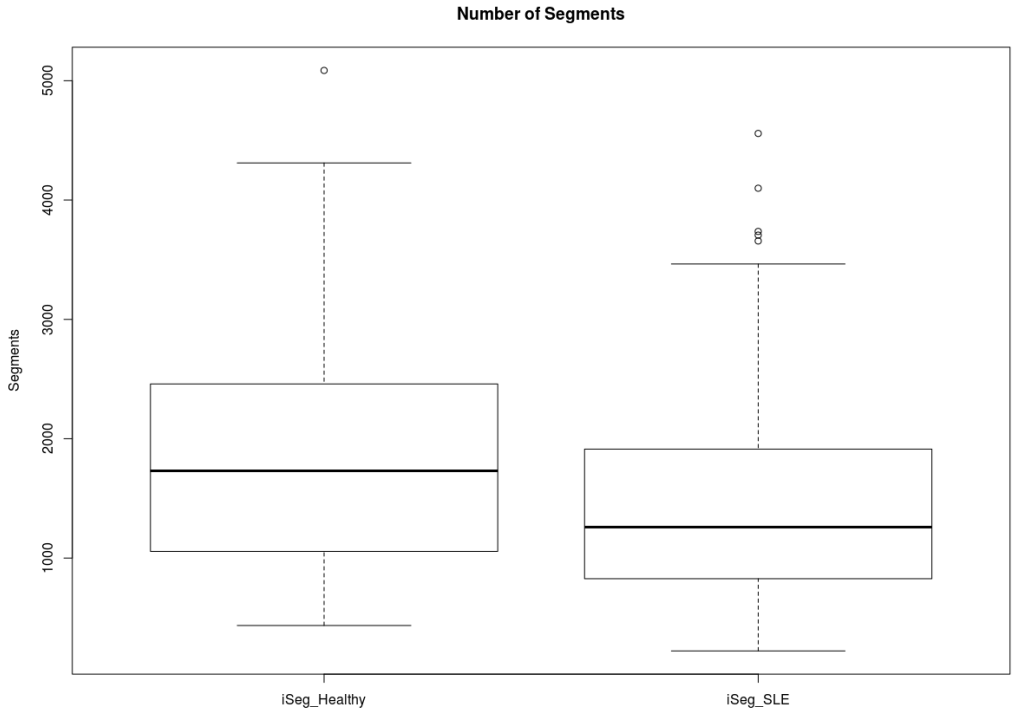


Figure 2

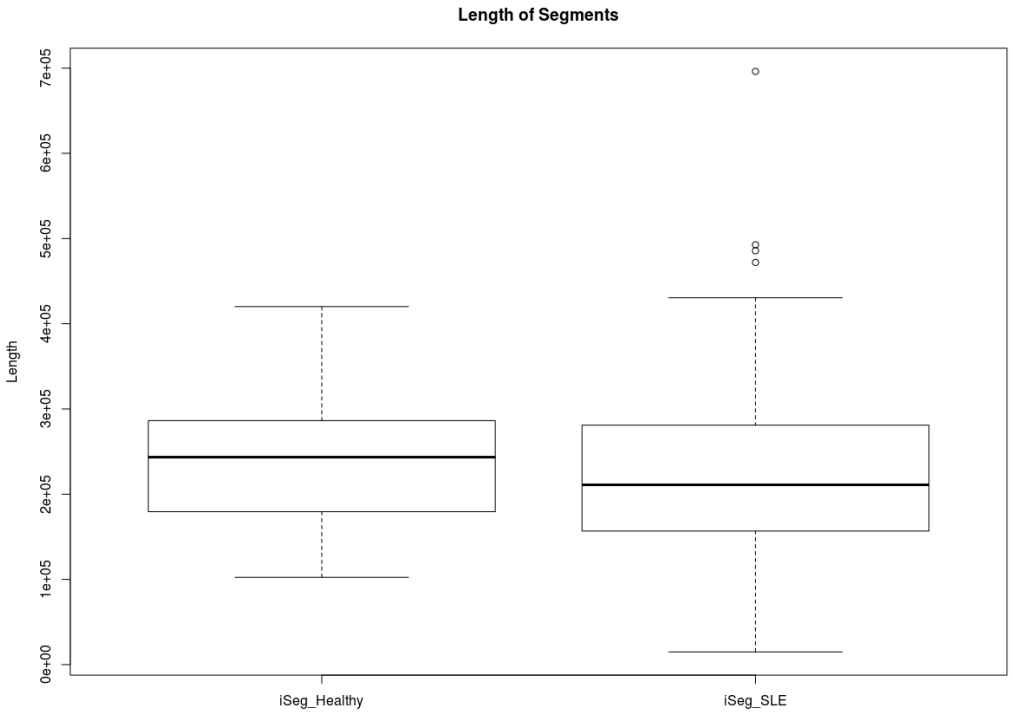
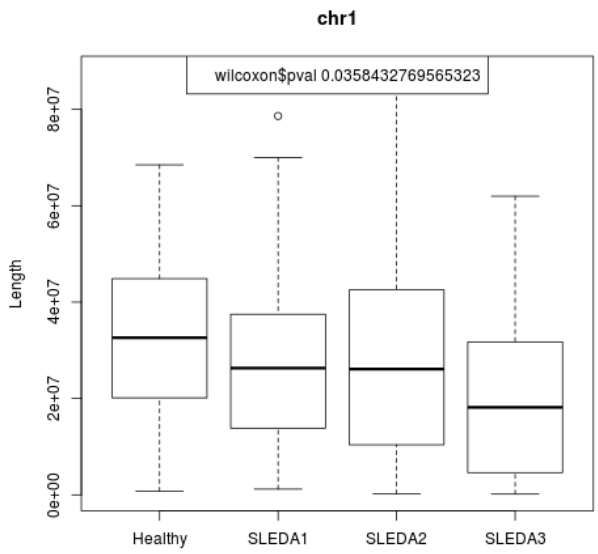
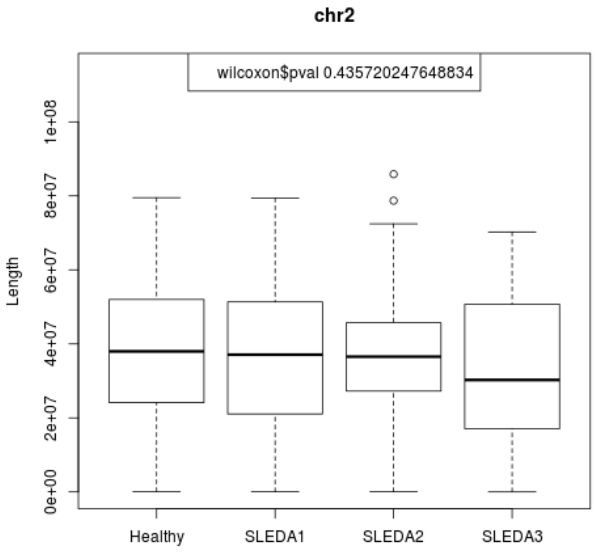
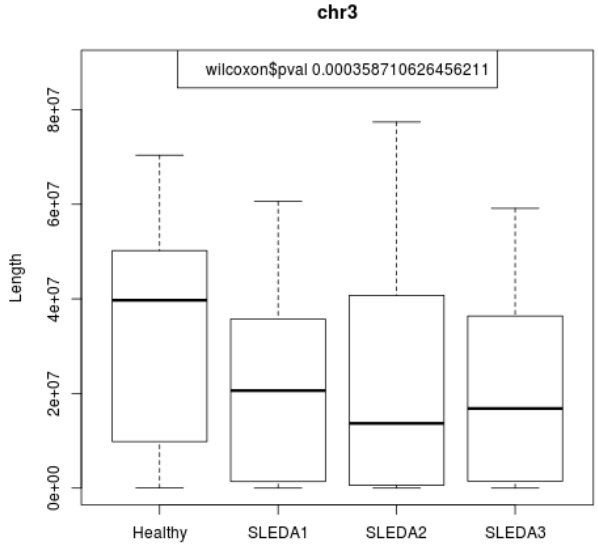
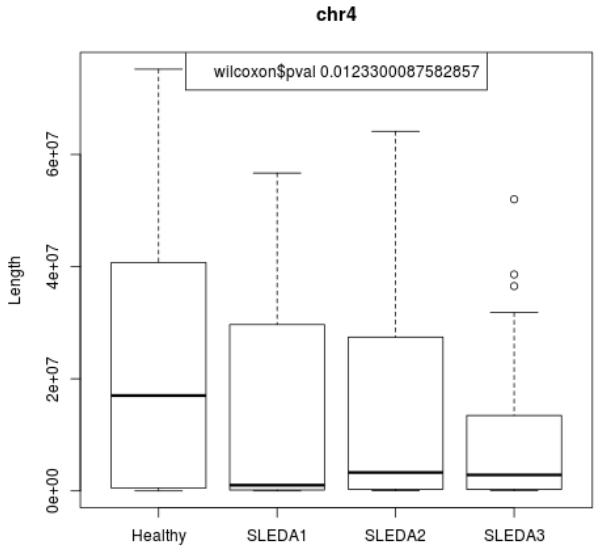


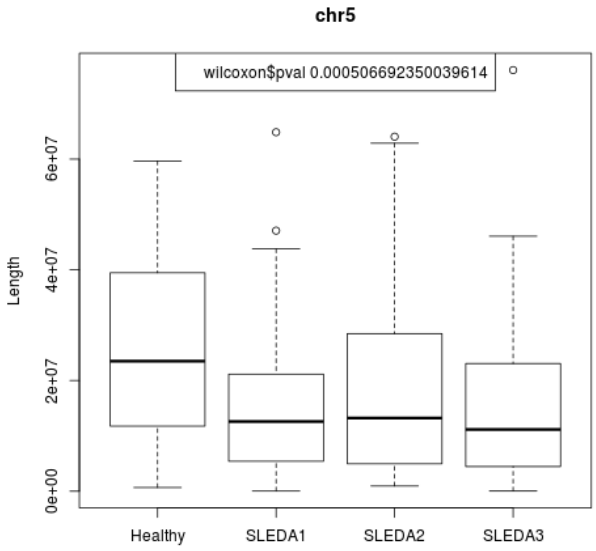
Figure 3

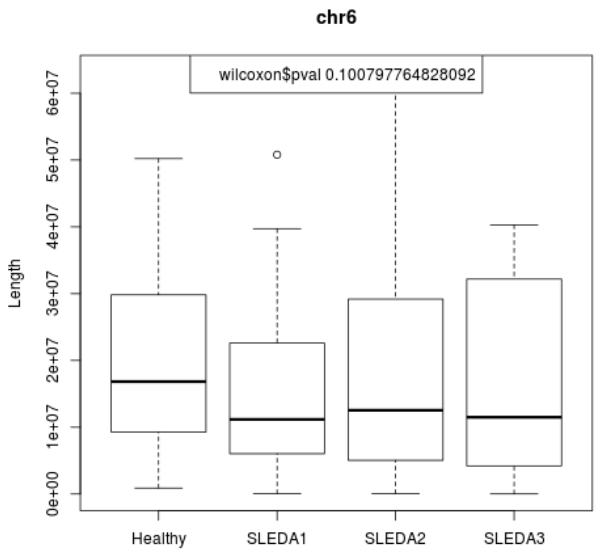

chr1
wilcoxon$pval 0.0358432769565323
Length
0e+00
2e+07
4e+07
6e+07
8e+07
Healthy
SLEDA1
SLEDA2
SLEDA3

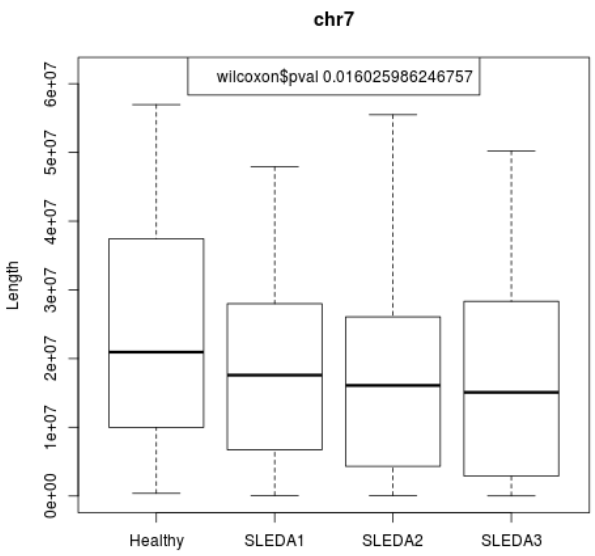

chr2
wilcoxon$pval 0.435720247648834
Length
0e+00
2e+07
4e+07
6e+07
8e+07
1e+08
Healthy
SLEDA1
SLEDA2
SLEDA3

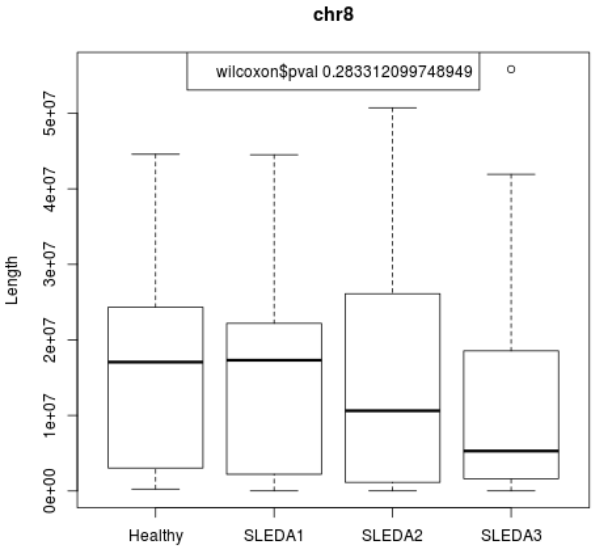

chr3
wilcoxon$pval 0.000358710626456211
Length
0e+00
2e+07
4e+07
6e+07
8e+07
Healthy
SLEDA1
SLEDA2
SLEDA3

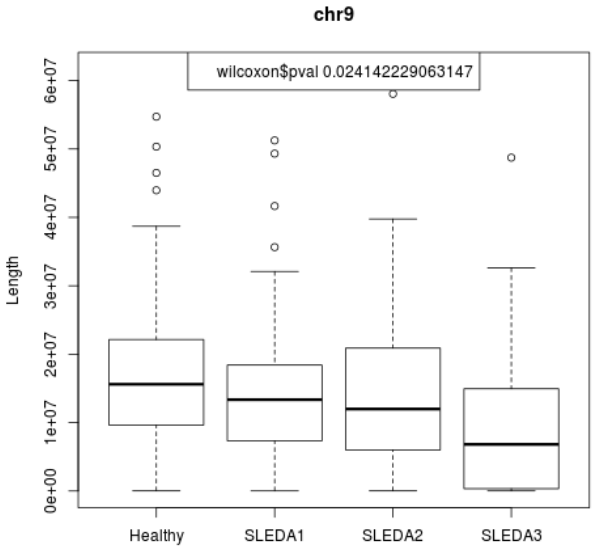

chr4
wilcoxon$pval 0.0123300087582857
Length
0e+00
2e+07
4e+07
6e+07
Healthy
SLEDA1
SLEDA2
SLEDA3

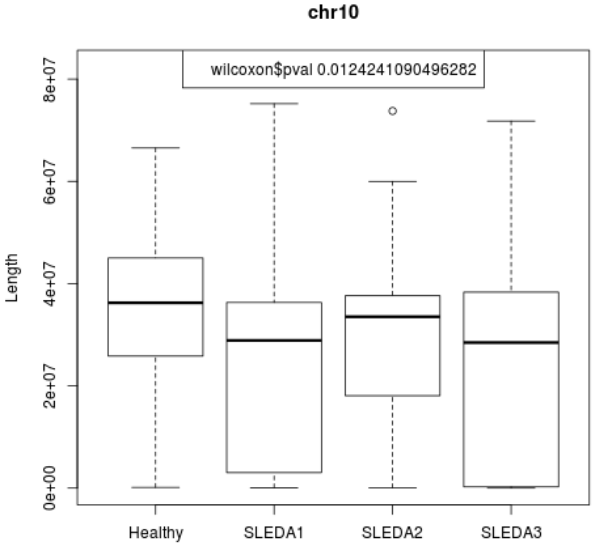

chr5
wilcoxon$pval 0.000506692350039614
Length
0e+00
2e+07
4e+07
6e+07
Healthy
SLEDA1
SLEDA2
SLEDA3

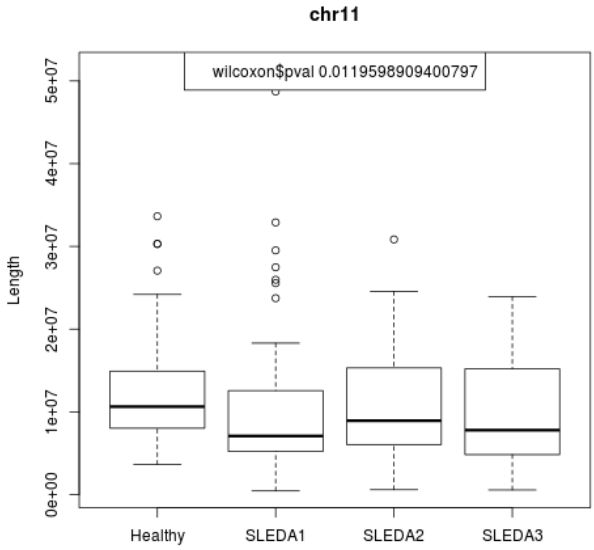

chr6
wilcoxon$pval 0.100797764828092
Length
0e+00
1e+07
2e+07
3e+07
4e+07
5e+07
6e+07
Healthy
SLEDA1
SLEDA2
SLEDA3

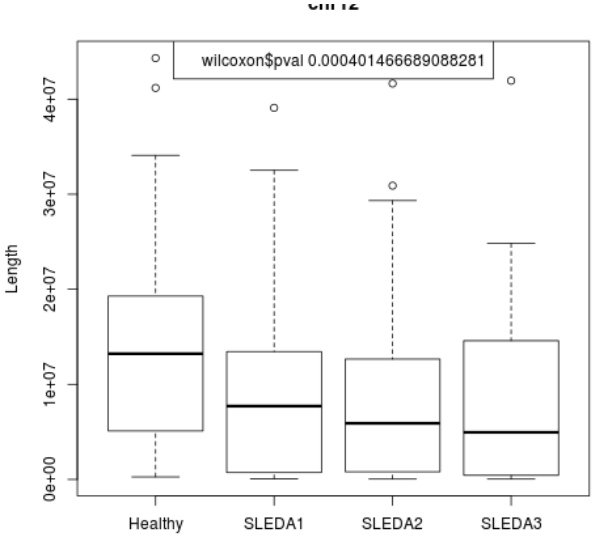

chr7
wilcoxon$pval 0.016025986246757
Length
0e+00
1e+07
2e+07
3e+07
4e+07
5e+07
6e+07
Healthy
SLEDA1
SLEDA2
SLEDA3

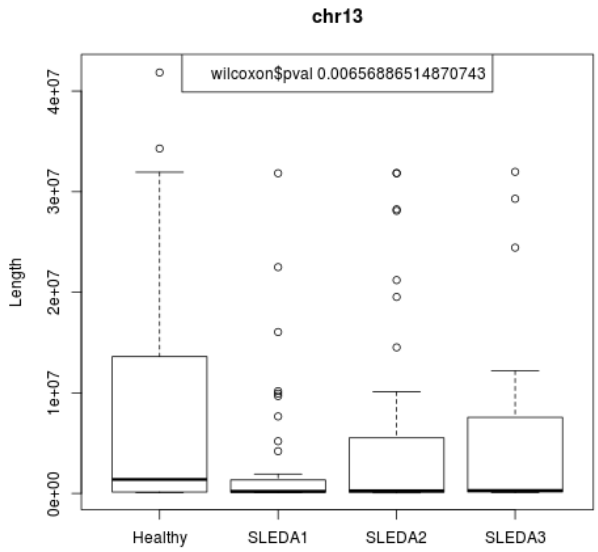

chr8
wilcoxon$pval 0.283312099748949
Length
0e+00
1e+07
2e+07
3e+07
4e+07
5e+07
Healthy
SLEDA1
SLEDA2
SLEDA3

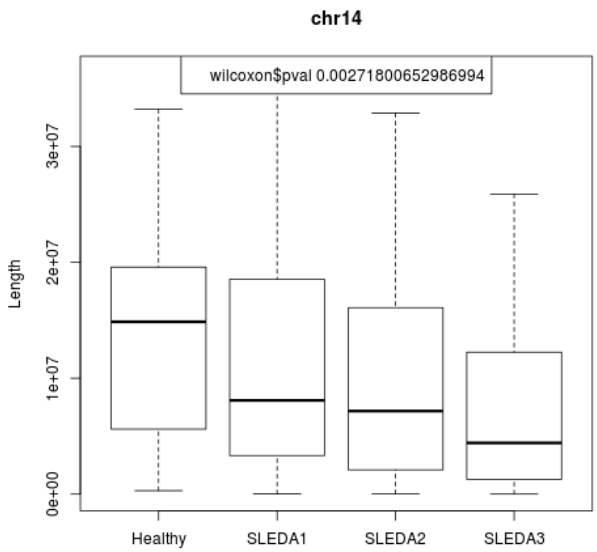

chr9
wilcoxon$pval 0.024142229063147
Length
0e+00
1e+07
2e+07
3e+07
4e+07
5e+07
6e+07
Healthy
SLEDA1
SLEDA2
SLEDA3

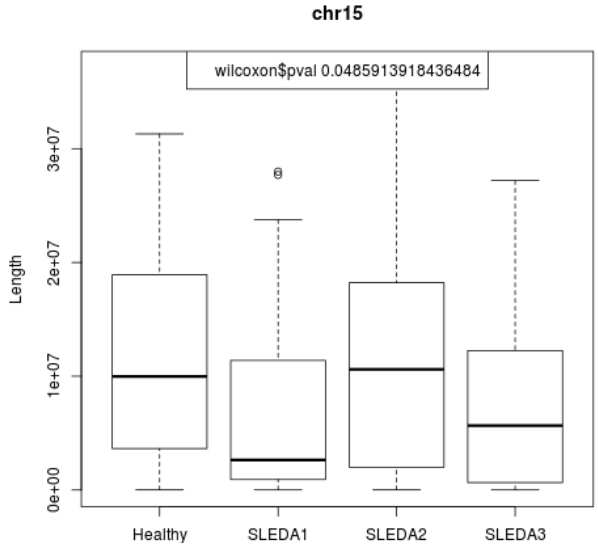

chr10
wilcoxon$pval 0.0124241090496282
Length
0e+00
2e+07
4e+07
6e+07
8e+07
Healthy
SLEDA1
SLEDA2
SLEDA3

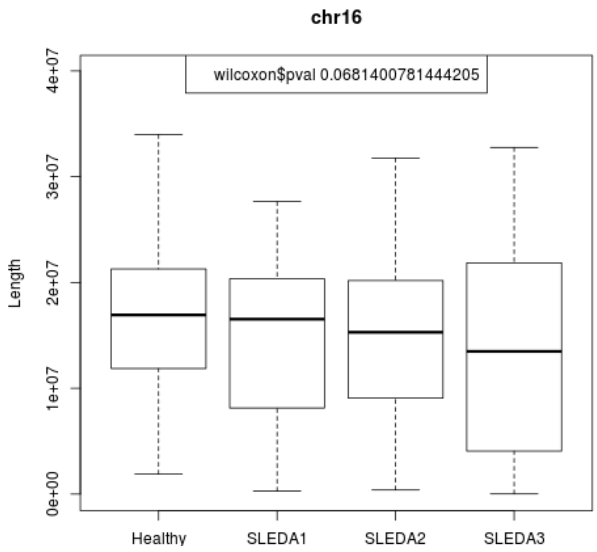

chr11
wilcoxon$pval 0.0119598909400797
Length
0e+00
1e+07
2e+07
3e+07
4e+07
5e+07
Healthy
SLEDA1
SLEDA2
SLEDA3

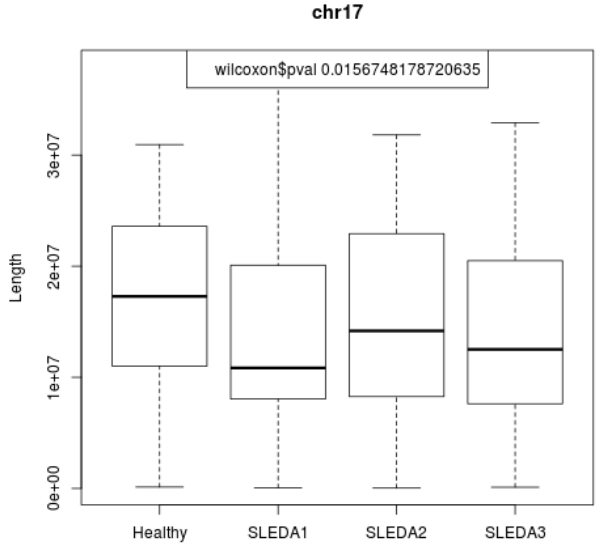

wilcoxon$pval 0.000401466689088281
Length
0e+00
1e+07
2e+07
3e+07
4e+07
Healthy
SLEDA1
SLEDA2
SLEDA3

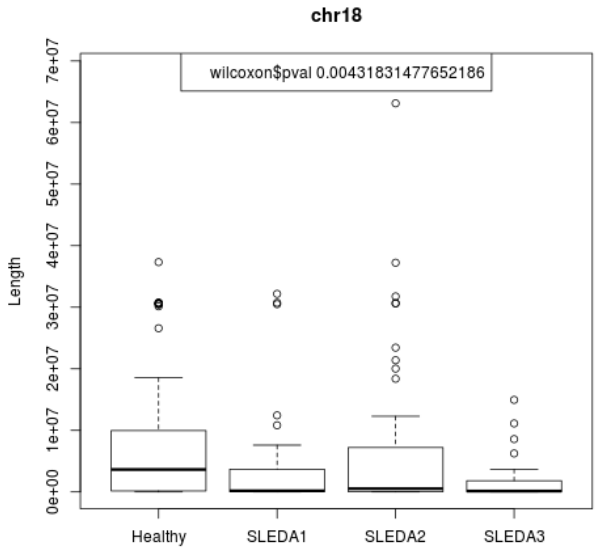

chr13
wilcoxon$pval 0.00656886514870743
Length
0e+00
1e+07
2e+07
3e+07
4e+07
Healthy
SLEDA1
SLEDA2
SLEDA3


chr14
wilcoxon$pval 0.00271800652986994
Length
0e+00
1e+07
2e+07
3e+07
Healthy
SLEDA1
SLEDA2
SLEDA3


chr15
wilcoxon$pval 0.0485913918436484
Length
0e+00
1e+07
2e+07
3e+07
Healthy
SLEDA1
SLEDA2
SLEDA3


chr16
wilcoxon$pval 0.0681400781444205
Length
0e+00
1e+07
2e+07
3e+07
4e+07
Healthy
SLEDA1
SLEDA2
SLEDA3


chr17
wilcoxon$pval 0.0156748178720635
Length
0e+00
1e+07
2e+07
3e+07
Healthy
SLEDA1
SLEDA2
SLEDA3


chr18
wilcoxon$pval 0.00431831477652186
Length
0e+00
1e+07
2e+07
3e+07
4e+07
5e+07
6e+07
7e+07
Healthy
SLEDA1
SLEDA2
SLEDA3

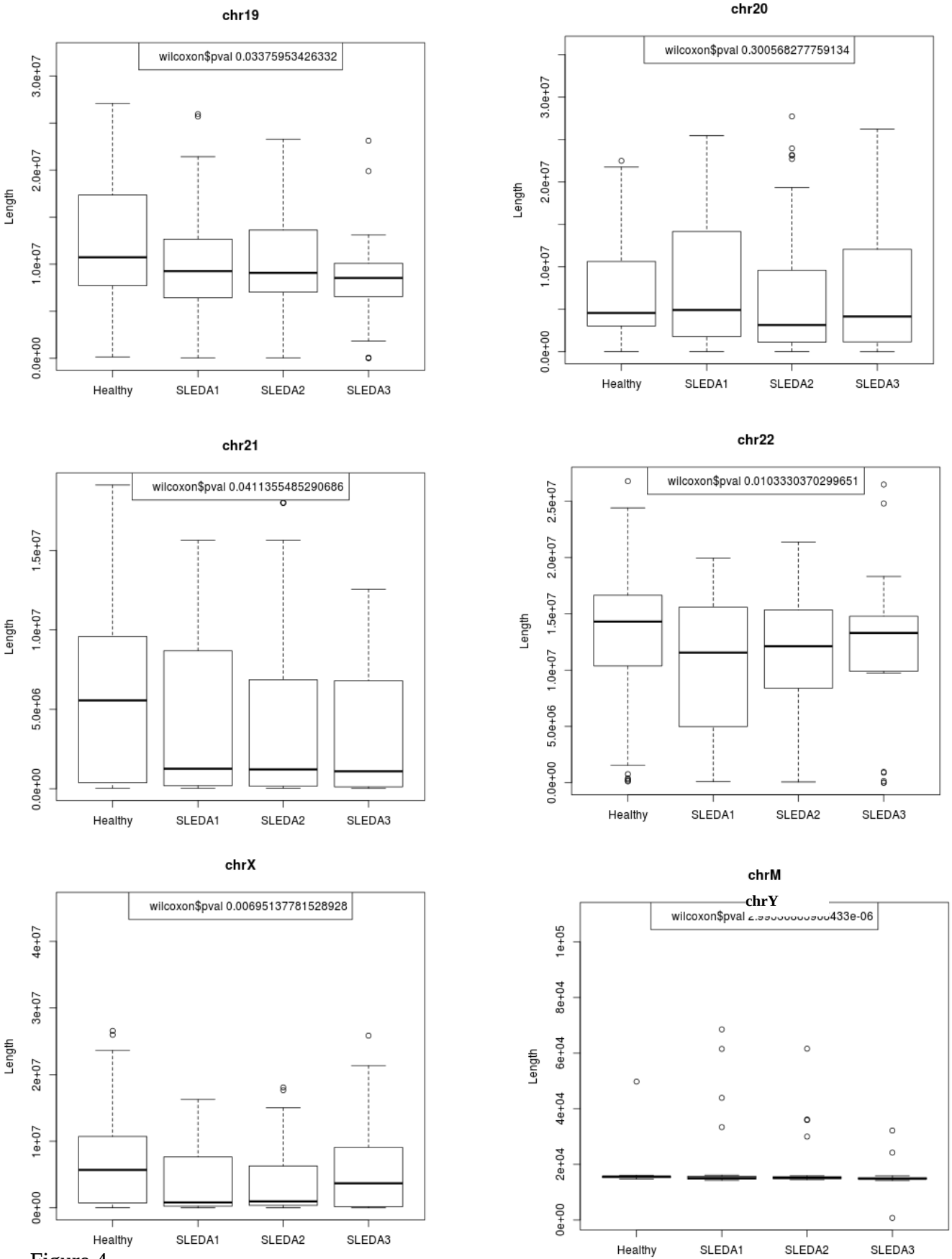

chr19
wilcoxon$pval 0.03375953426332
Length
0.0e+00
1.0e+07
2.0e+07
3.0e+07
Healthy
SLEDA1
SLEDA2
SLEDA3
chr20
wilcoxon$pval 0.300568277759134
chr21
wilcoxon$pval 0.0411355485290686
5.0e+06
1.5e+07
chr22
wilcoxon$pval 0.0103330370299651
2.5e+07
chrX
wilcoxon$pval 0.00695137781528928
0e+00
2e+07
3e+07
4e+07
chrM
chrY
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05

Figure 4

Συσχέτιση

Μελετήθηκε η συσχέτιση μεταξύ όλων των συνδυασμών των κλινικών τιμών και των δεδομένων που παράχθηκαν μετά την εκτέλεση του εργαλείου iSeg. Ο συνδυασμός που παρατηρήθηκε να έχει σημαντική αρνητική συσχέτιση (cor= -0.24) ήταν μεταξύ των διαφορετικών τύπων εμφάνισης της ασθένειας (SLEDA1,SLEDA2,SLEDA3) και του μέσου μήκους των περιοχών που προέκυψαν από την μέθοδο τμηματοποίησης. Με βάση αυτη την αρνητική συσχέτιση(Figure 5) ,όσο πιο μεγάλη είναι η ενεργότητα της ασθένειας τόσο μικρότερα τμήματα δημιουργούνται από την μέθοδο τμηματοποίησης iSeg.

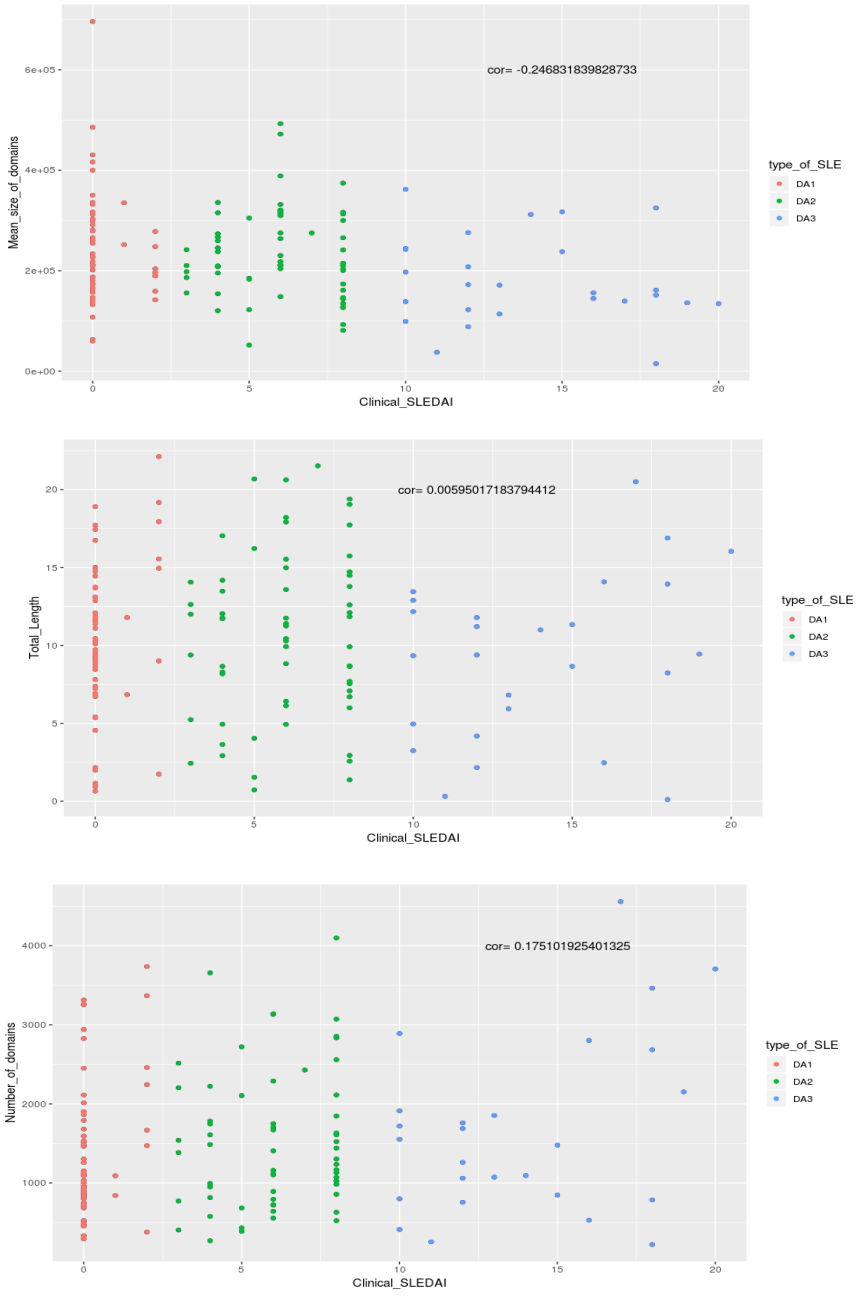


Figure 5

Δημιουργήθηκαν δύο heatmaps (Figures 6,7) με βάση το ποσοστό του χρωμοσώματος που καλύπτεται από τα τμήματα του κάθε ατόμου που προέκυψαν από την μέθοδο τμηματοποίησης iSeg. Δεν παρατηρήθηκε κάποια διαφορά στα χρωμοσώματα με βάση την κατάσταση (Υγιείς ή ΣΛΕ) των ατόμων που μελετήσαμε. Ωστόσο, παρατηρήθηκε ότι κάποια χρωμοσώματα ,όπως το χρωμόσωμα 22 και το χρωμόσωμα 10, είχαν μεγαλύτερη κάλυψη από τα τμήματα ανεξαρτήτως την κατάσταση υγείας των ατόμων. Το φαίνομενο αυτό μπορεί να εξηγηθεί από το γεγόνος ότι τα συγκεκριμένα χρωμοσώματα είναι μεγαλύτερα σε μέγεθος από τα υπόλοιπα.

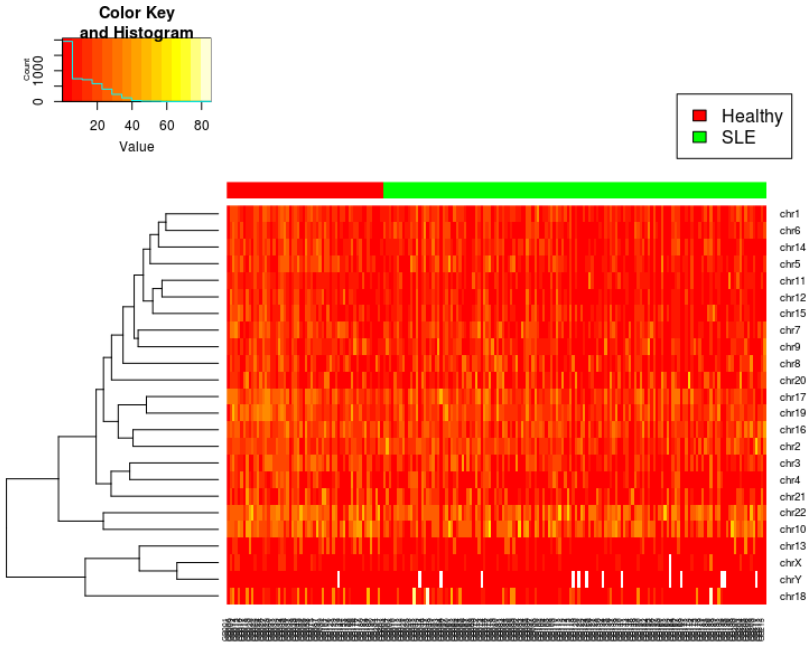


Figure 6

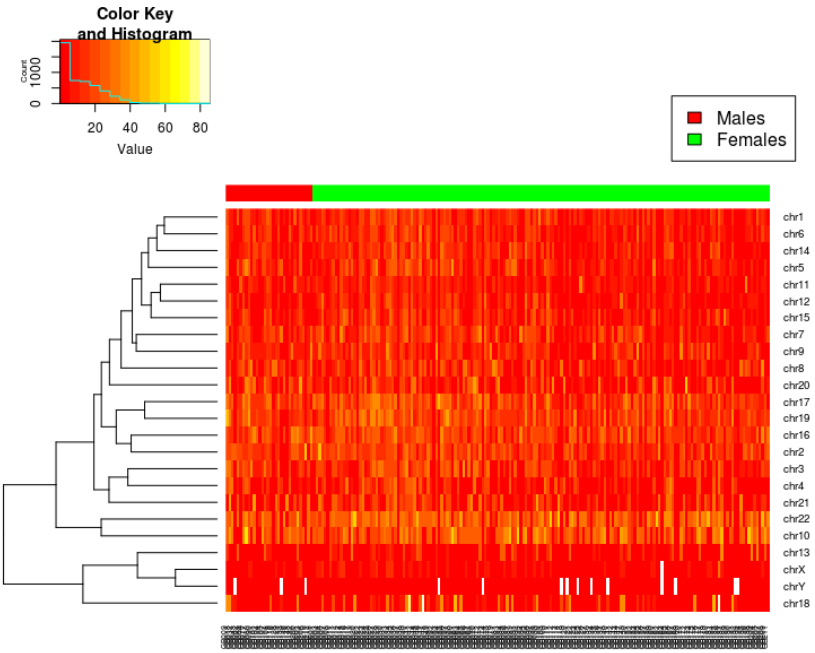


Figure 7

Λόγω των αποτελεσμάτων που προέκυψαν από το Heatmap (Figure7) έγινε προσπάθεια σύγκρισης του βαθμού κάλυψης του κάθε χρωμοσώματος από τα τμήματα σε σχέση με την πυκνότητα των στοιχείων που περιέχει και φάνηκε ότι υπάρχει μια μικρή θετική συσχέτιση. Δημιουργήθηκε ένα boxplot (Figure 8) όπου φαίνεται η διακύμανση και η τιμή της επικάλυψης των χρωμοσωμάτων από τα τμήματα που προέκυψαν από την μέθοδο τμηματοποίησης iSeg. Η κόκκινη γραμμή δείχνει την πυκνότητα των γονιδίων ανά χρωμόσωμα. Συμπερασματικά, φαίνεται ότι το εργαλείο iSeg είναι ευαίσθητο στην πυκνότητα γονιδίων στο χρωμόσωμα. Είναι αντιληπτό ότι σε κάποια από τα χρωμοσώματα που έχουν μεγάλη πυκνότητα γονιδίων, τα τμήματα που δημιουργούνται είναι μεγαλύτερα ένω σε κάποια από τα χρωμοσώματα που έχουν μικρή πυκνότητα γονιδίων, τα τμήματα που δημιουργούνται είναι μικρότερα και για αυτό τον λόγο δεν δημιουργούνται ομάδες γονιδίων (clusters). Το παραπάνω είναι ενδεικτικό ότι το εργαλείο iSeg εξαρτάται από πολλές παραμέτρους που δεν σχετίζονται τόσο με πείραμα γονιδιακής έκφρασης αλλά με τη δομή του γονιδιώματος.

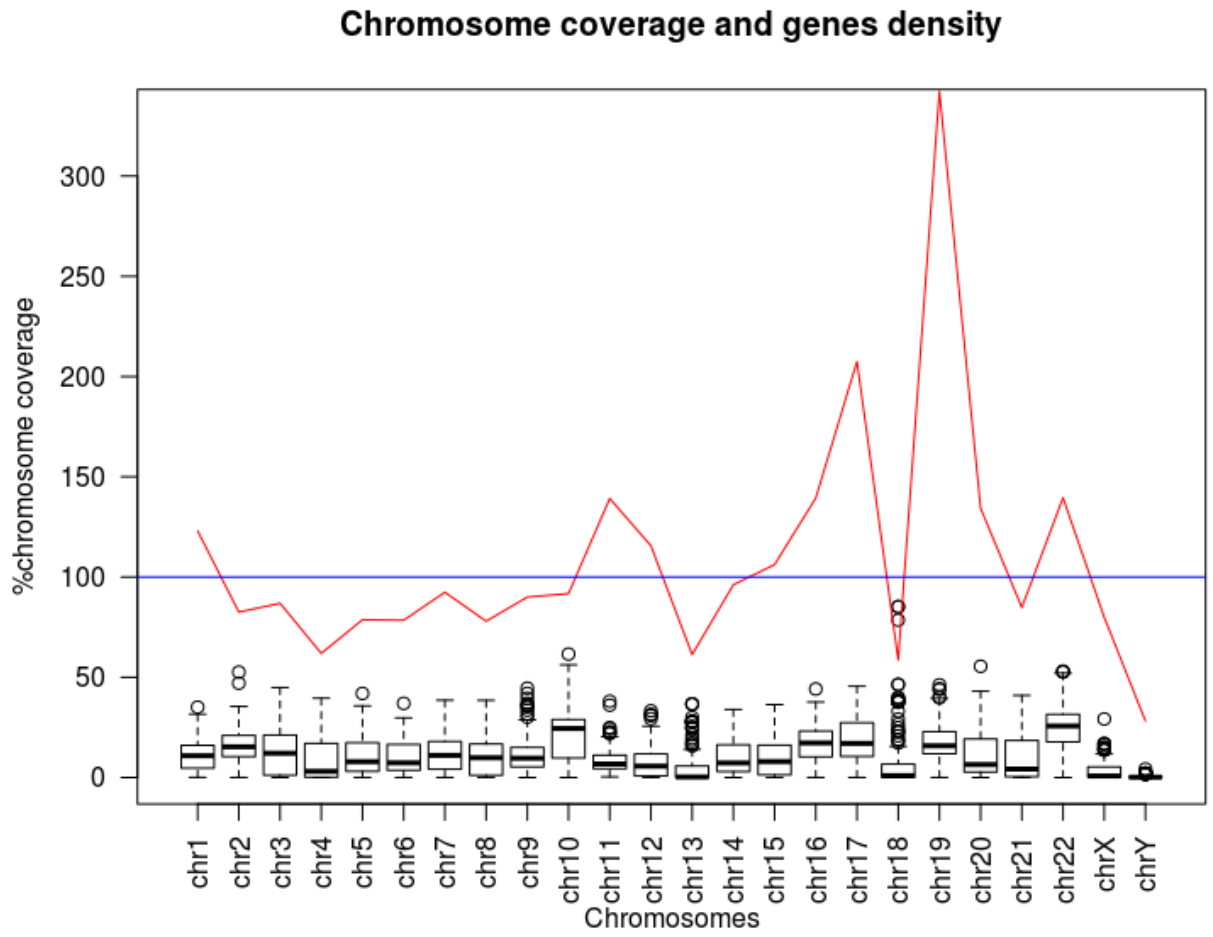


Figure 8

Domains of Focal Over-Expression (DFOE)

Από την εκτέλεση του αλγορίθμου γονιδιωματικής τμηματοποίησης, Domains of Focal Over-Expression (DFOE), παράχθηκαν 200 νέα bedGraph αρχεία. Τα γονίδια τα οποία είχαν λιγότερα από πέντε μετρήσεις σε όλα τα άτομα αφαιρέθηκαν. Όπως παρατηρήθηκε στα Figures 9,10 τα τμήματα που προέκυψαν από την τμηματοποίηση με τον αλγόριθμο DFOE ήταν λίγα σε αριθμό και μεγάλα σε μήκος τόσο στους υγιείς όσο και στους ασθενείς. Ωστόσο και στα δύο διαγράμματα φαίνεται ότι οι ασθενείς έχουν μεγαλύτερα και λιγότερα σε αριθμό τμήματα από τους υγιείς, χωρίς όμως να υπάρχει κάποια σημαντική διαφορά, φαινόμενο το οποίο μπορεί να εξηγηθεί από τον τρόπο που γίνεται η τμηματοποίηση με τον αλγόριθμο DFOE (βλ. Υλικά και Μέθοδοι 3.3.2). Από τα τμημάτα που προέκυψαν αφαιρέθηκαν εκείνα που είχαν λιγότερα από πενήντα γονίδια (Figures 11,12), διότι από τον αλγόριθμο DFOE προκύπτουν τμήματα τα οποία καλύπτουν όλο το γονιδίωμα, κάτι που δεν είναι χρήσιμο για την ερευνά μας. Όσο αυξάνεται η ενεργότητα της ασθένειας τόσο μεγαλύτερη είναι η τάση να υπάρχουν υπερεκφραζόμενα γονίδια κάτι που είναι ήδη γνωστό από δεδομένα γονιδιακής έκφρασης , όπου τα στατιστικά σημαντικά υπερεκφραζόμενα γονίδια αυξάνονται με την αύξηση της ενεργότητας της ασθένειας. Πιο συγκεκριμένα όσο πιο μεγάλη η ενεργότητα της ασθένειας τόσο μεγαλύτερη τάση υπάρχει να υπάρχουν περιοχές με υπερέκφραση, οι οποίες δημιουργούν clusters που περιέχουν πάνω απο 50 γονίδια. Το Figure 11 μπορεί να δικαιολογήσει την απόφαση μας να ασχοληθούμε παραπάνω με τον αλγόριθμο DFOE συγκριτικά με το αντίστοιχο Figure 8 από το εργαλείο iSeg , διότι ο στόχος μας ήταν να δημιουργήσουμε μεγάλου μεγέθους clusters που να έχουν μια κάλυψη του γονιδιώματος ύψους 60%, ώστε να καταλήξουμε με clusters που περιέχουν μεγάλο αριθμό γονιδίων και επί πρόσθετα να ενωθούν περιοχές που μπορεί να είναι μακριά στο γονιδίωμα αλλά να υπάρχει μια συνέχεια στις μεταξύ του περιοχές.

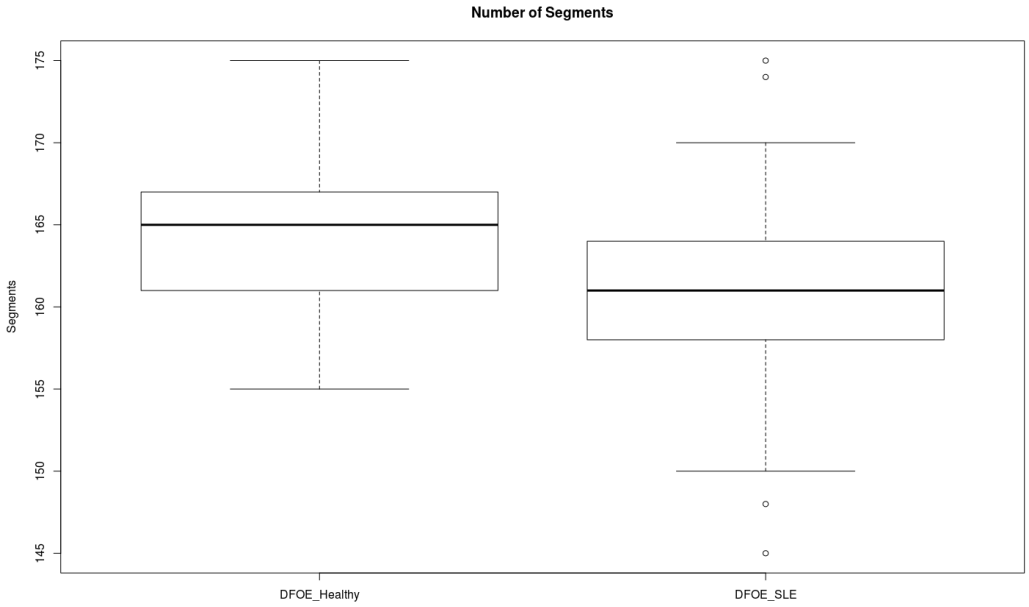


Figure 9

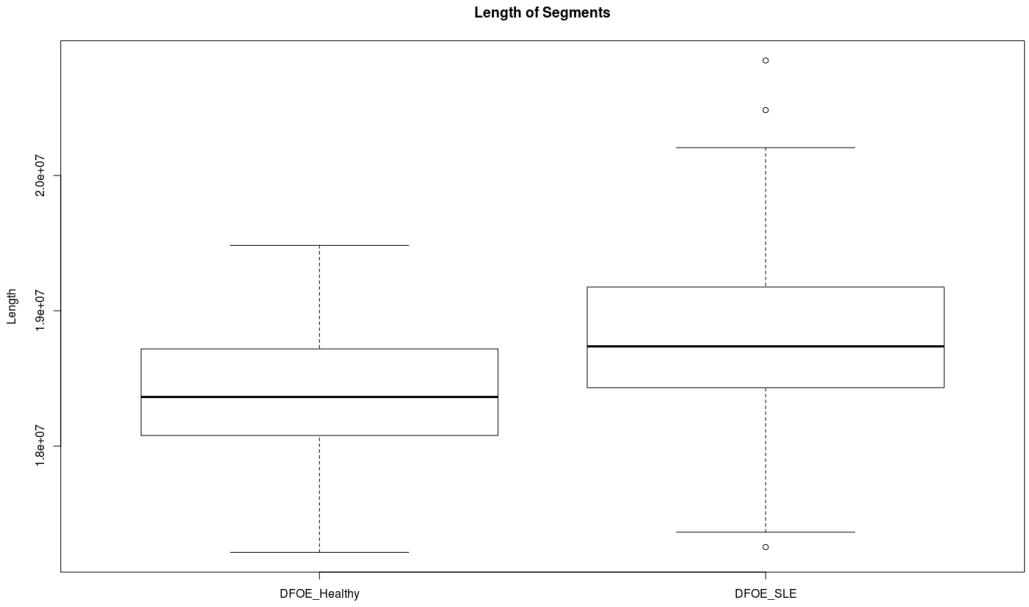


Figure 10

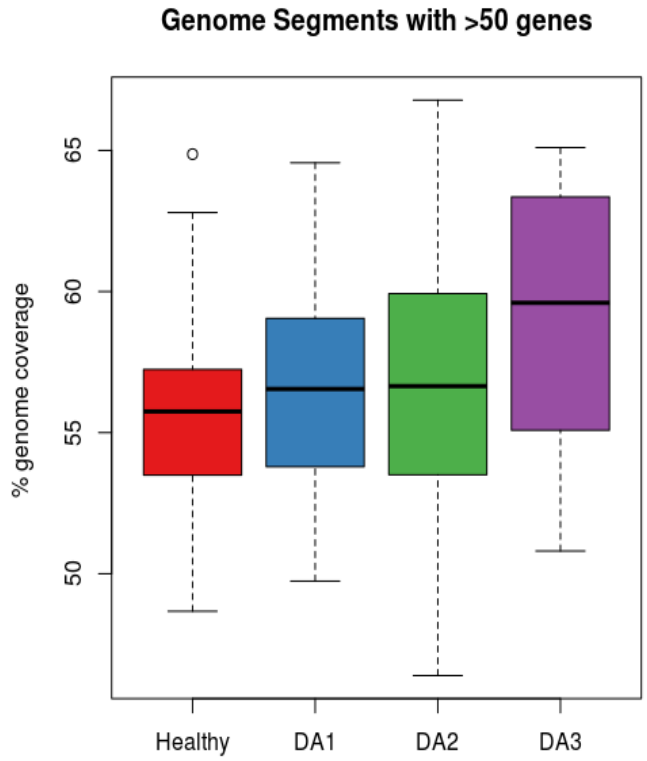


Figure 11

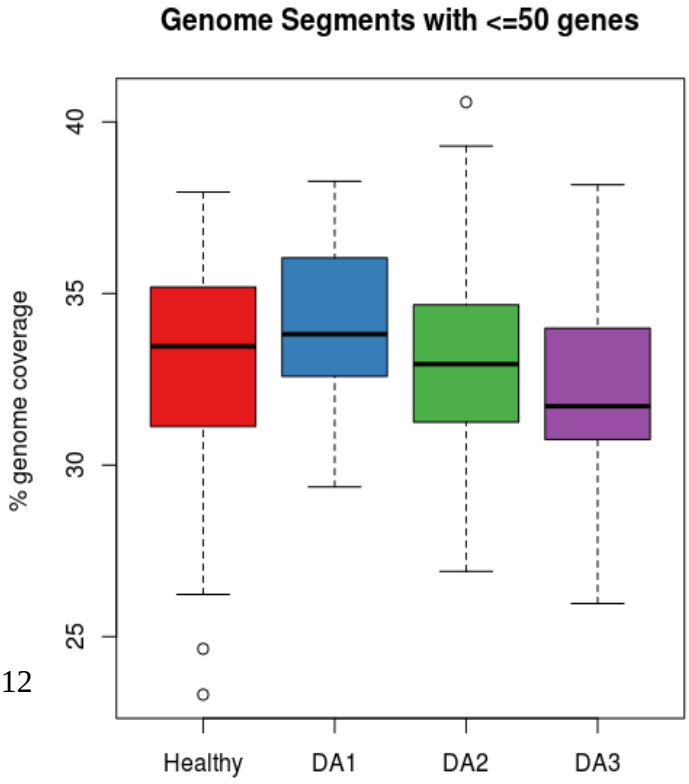


Figure 12

Από τον αλγόριθμο εύρεσης στατιστικά σημαντικών γονιδιωματικών τμημάτων (βλ. Υλικά και Μέθοδοι 3.8) φάνηκε ότι υπάρχουν κάποιες σημαντικές διαφορές των τμημάτων που προέκυψαν με την μέθοδο τμηματοποίησης DFOE σε συγκεκριμένα χρωμοσώματα μεταξύ υγιών και ασθενών ατόμων (Figure 14). Με κόκκινο χρώμα εμφανίζονται τα υγιή άτομα ενώ με μπλε χρώμα τα άτομα με ΣΛΕ. Επιπλέον, από τον πίνακα (SLE-Healthy) παράχθηκε το Figure 13 όπου φαίνεται η διακύμανση των 5%, 50% και 95% ανά χρωμόσωμα. Στην συνέχεια, δημιουργήθηκαν δύο νέα bed αρχεία τα οποία περιείχαν τις θέσεις όπου βρέθηκαν να είναι στατιστικά σημαντικές στους ασθενείς και στους υγιείς. Πραγματοποιήθηκε λειτουργική ανάλυση αυτών των τμημάτων με τη βοήθεια του εργαλείου gProgileR (Figures 15,16,17,18).

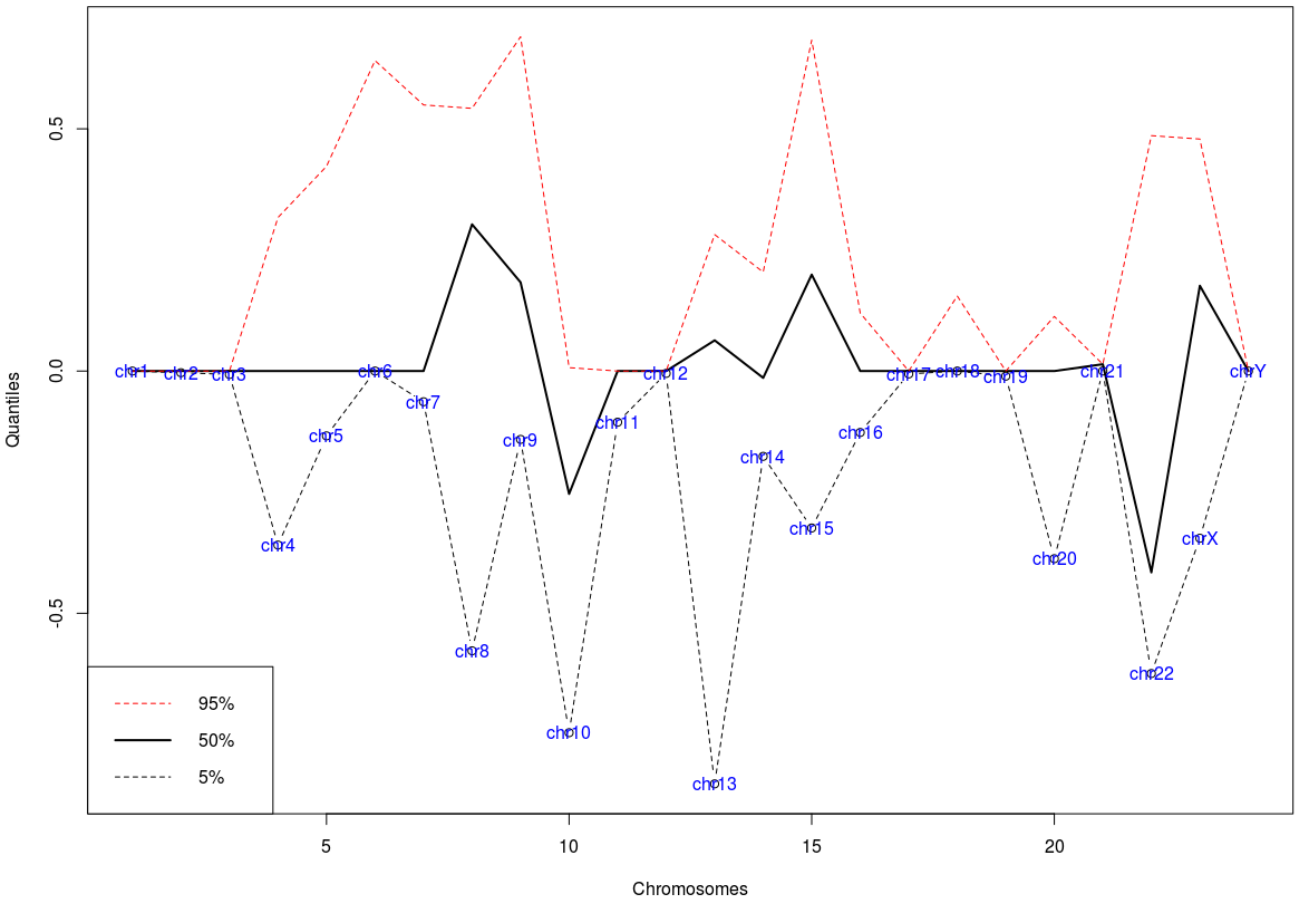
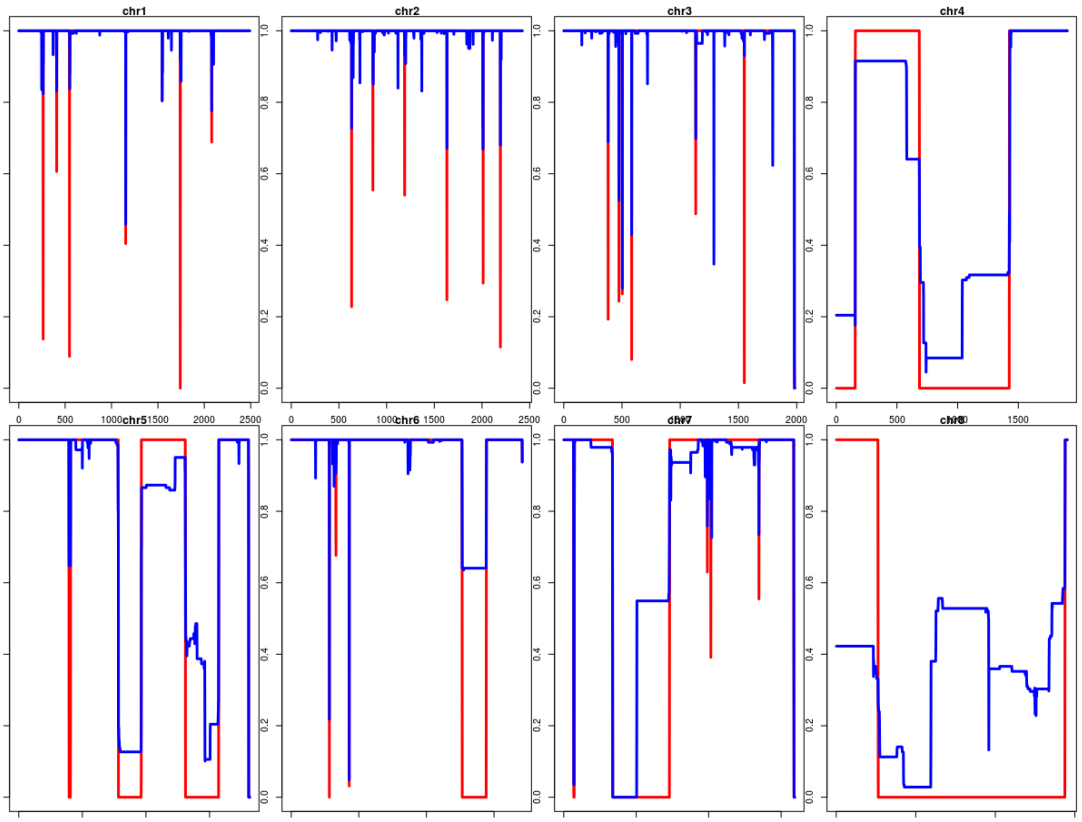


Figure 13


chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8

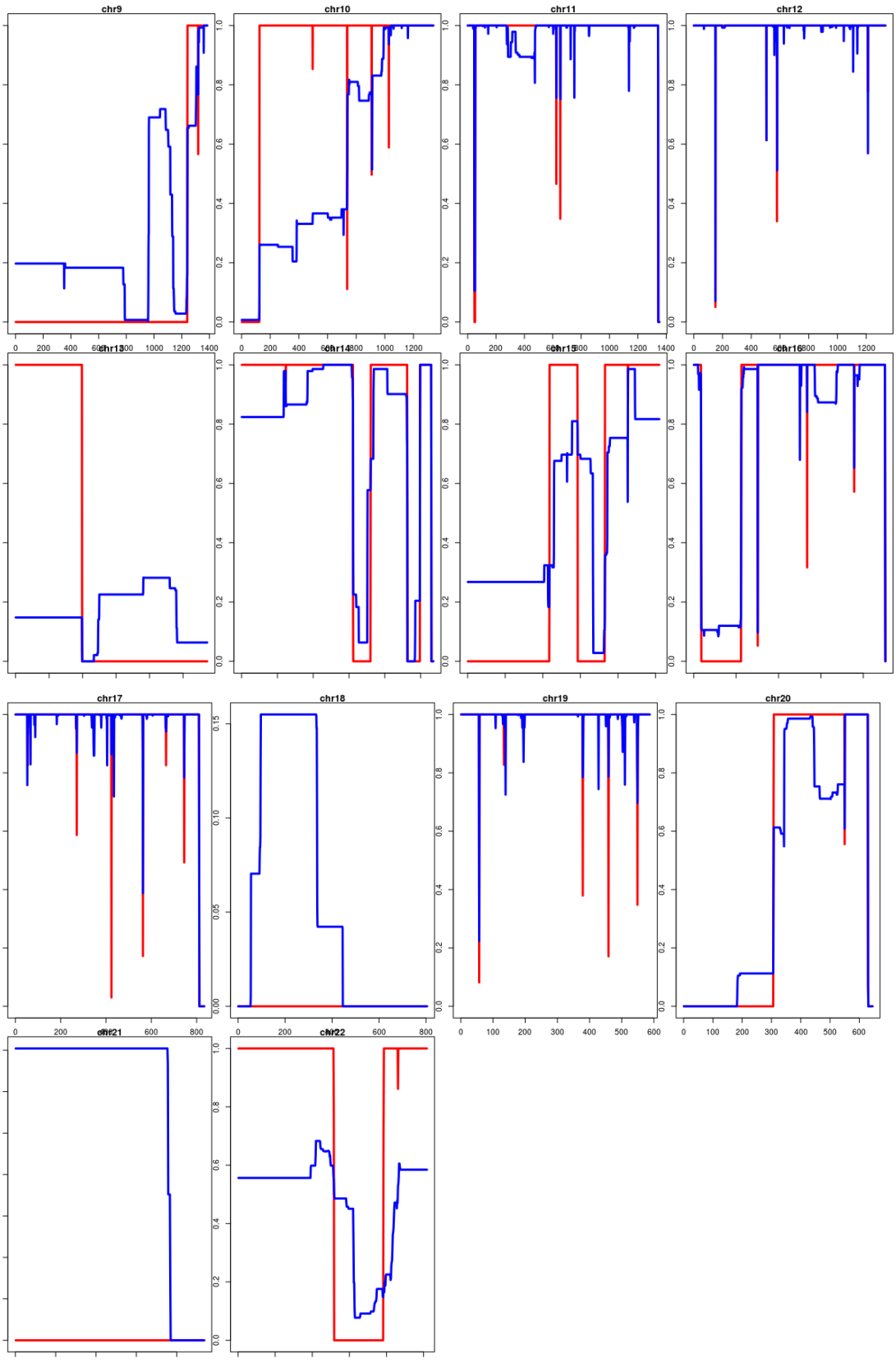

chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22

Figure 14

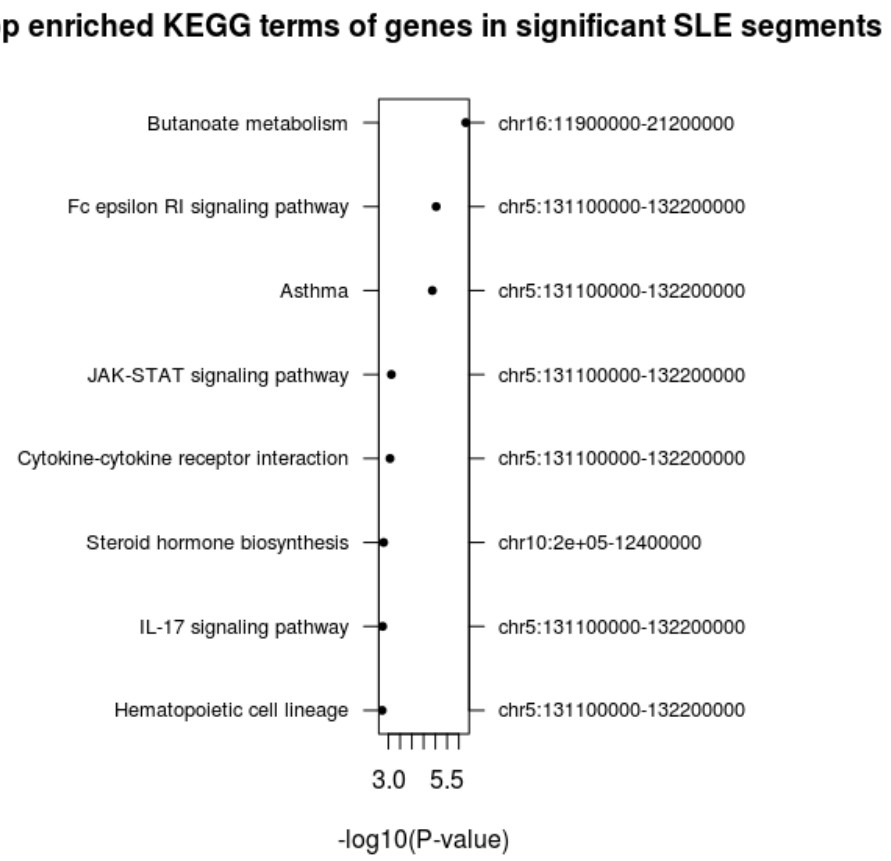


Figure 15

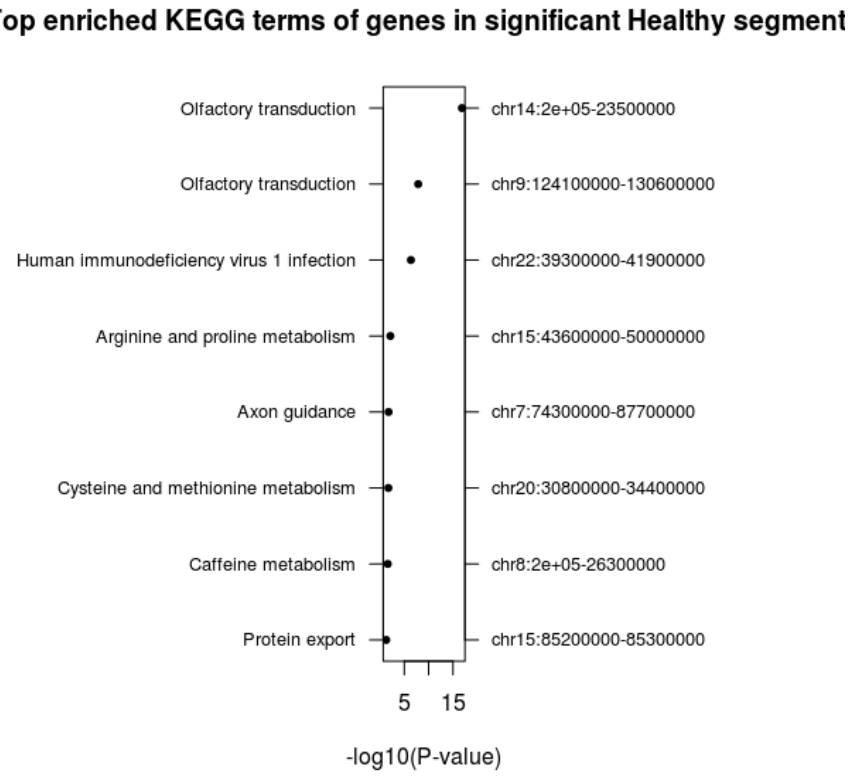


Figure 16

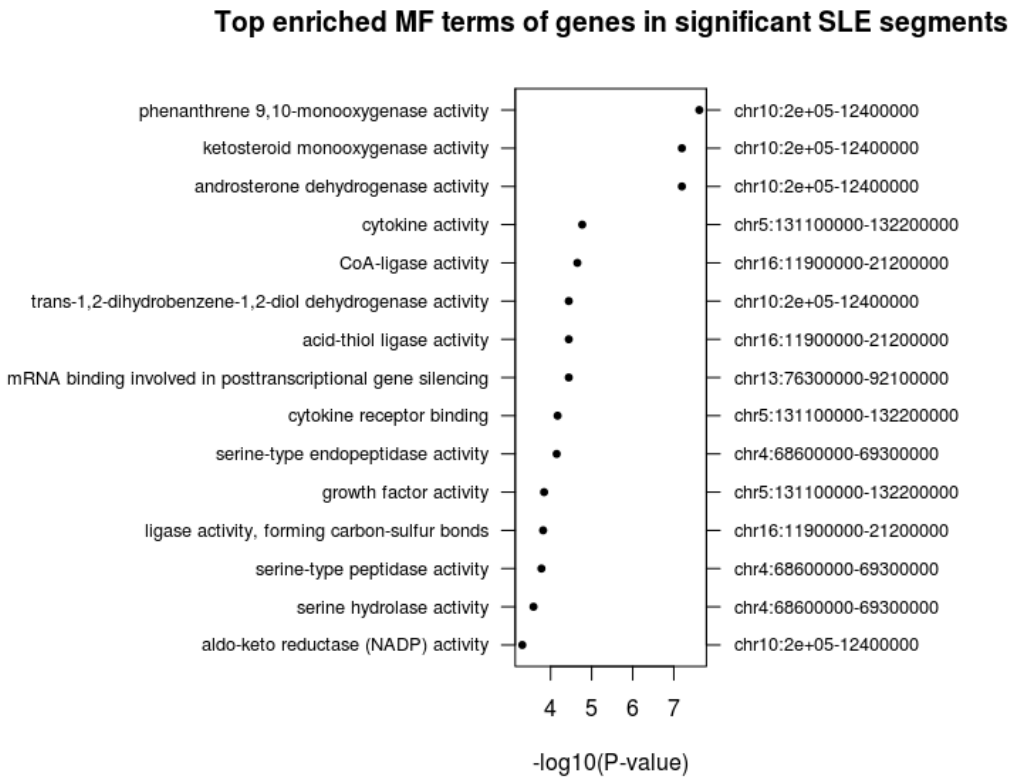


Figure 17

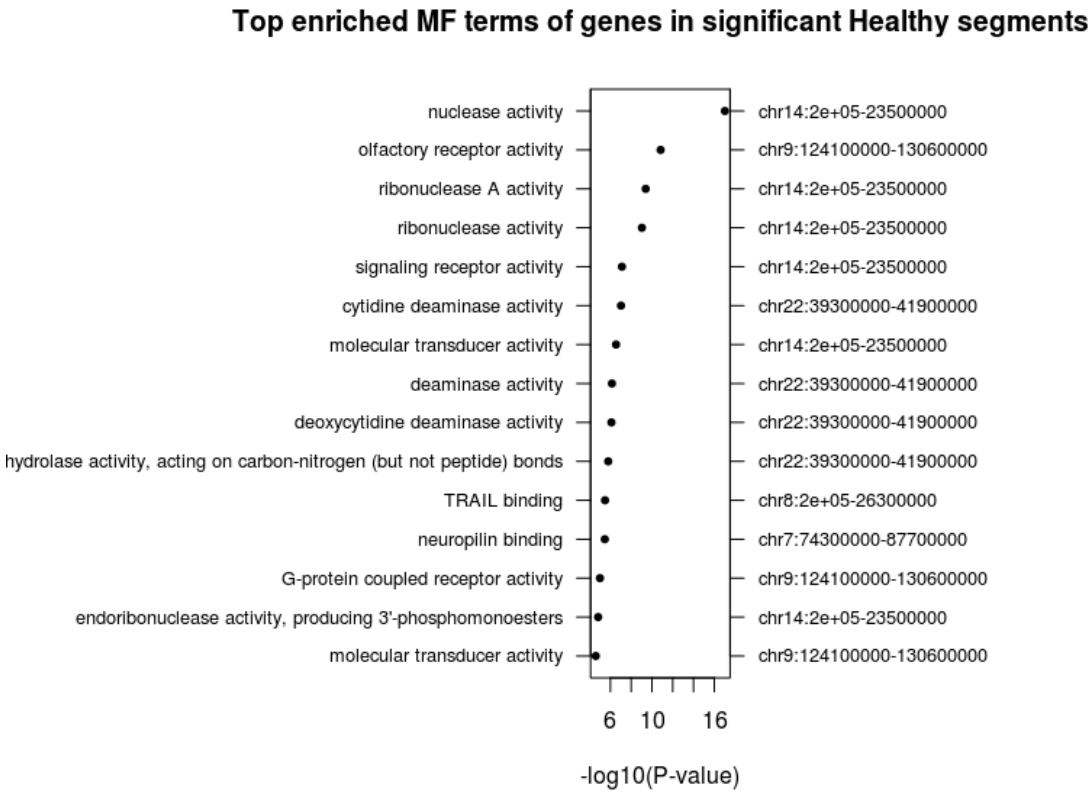


Figure 18

Από τα διαγράμματα που προέκυψαν από το GprofileR παρατηρήθηκε ότι οι χρωμοσωμικές περιοχές στις οποίες αντιστοιχούν οι στατιστικά σημαντικές λειτουργίες επαναλαμβάνονται. Για αυτό το λόγο σχεδιάστηκαν τα παρακάτω δίκτυα (Figures 17,18,19,20) όπου φαίνεται γύρω από ποιες χρωμοσωμικές συντεταγμένες συγκεντρώνονται οι λειτουργίες που προέκυψαν απο το GprofileR.

Figure 19

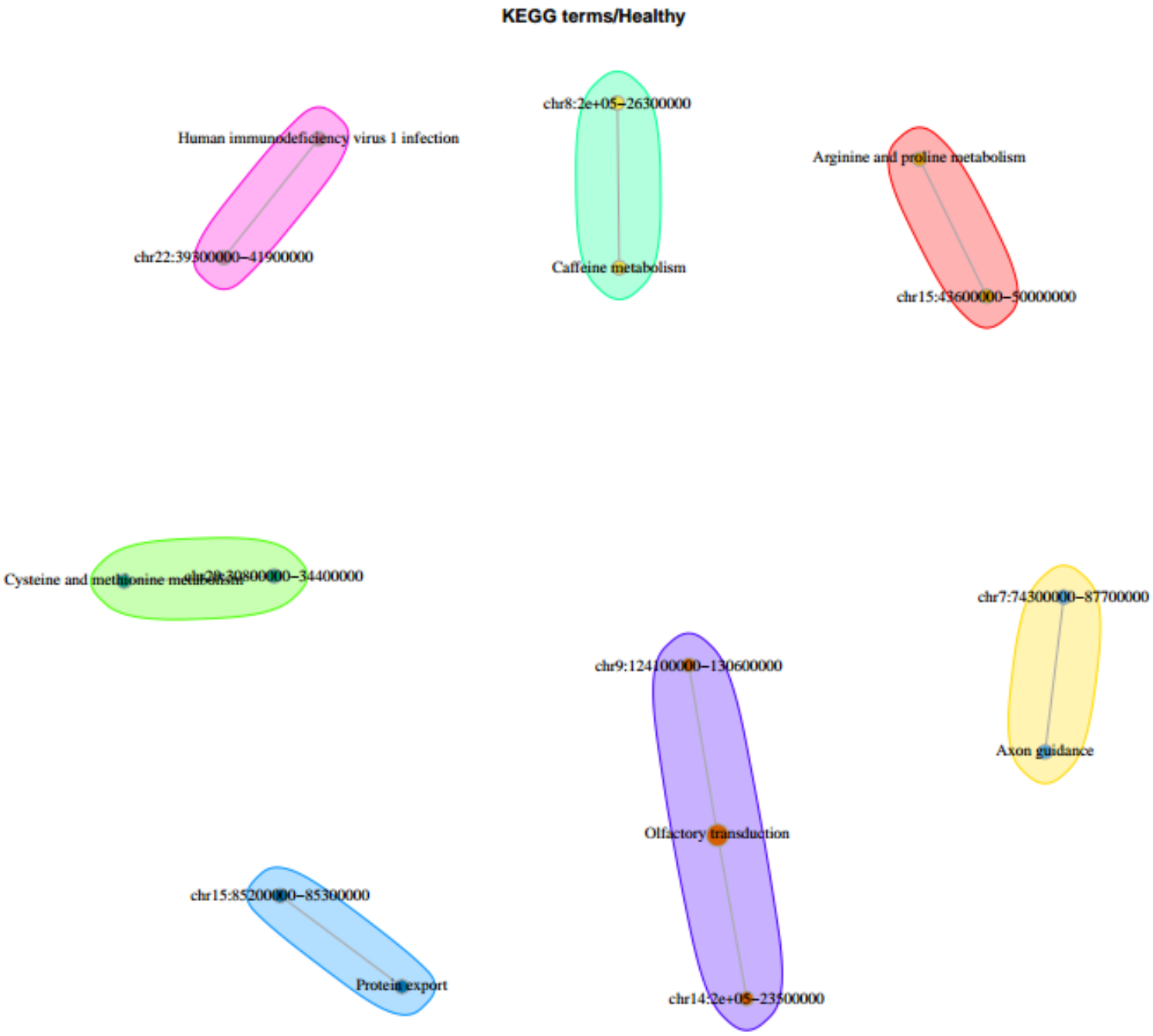

KEGG terms/Healthy
Human immunodeficiency virus 1 infection
chr22:39300000–41900000
chr8:2e+05–26300000
Caffeine metabolism
Arginine and proline metabolism
chr15:43600000–50000000
Cysteine and methionine metabolism
chr20:30800000–34400000
chr7:74300000–87700000
Axon guidance
chr9:124100000–130600000
Olfactory transduction
chr15:85200000–85300000
Protein export
chr14:2e+05–23500000

Figure 20


MF terms/SLE
tRNA−4−demethylwyosine synthase activity
fructose 1,6−bisphosphate 1−phosphatase activity
chr9:96400000−108400000
chr7:50500000−73000000
chr13:76300000−92100000
chr5:132400000−140200000
histidine−tRNA ligase activity
butyrate−CoA ligase activity
CoA−ligase activity
C−acyltransferase activity
acyl−CoA ligase activity
chr16:11900000−21200000
fatty−acyl−CoA synthase activit
fatty acid ligase activity
alcohol dehydrogenase (NADP+) activity
ketosteroid monooxygenase activity
aldo−keto reductase (NADP) activity
phenanthrene 9,10−monooxygenase activity
androsterone dehydrogenase activity
trans−1,2−dihydrobenzene−1,2−diol dehydrogenase activity
chr10:2e+05−12400000
oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor
indanol dehydrogenase (NADP+) activity
reductase activity, acting on the CH−OH group of donors, NAD or NADP as acceptor
quaternary ammonium group transmembrane transporter activity
growth factor activity
cytokine receptor binding
chr5:131100000−132200000
receptor ligand activity
receptor regulator activity
growth factor receptor binding
histone kinase activity (H3−S10 specific)
chr14:97100000−99700000
peptidase activity, acting on L−amino acid peptides
peptidase activity
endopeptidase activity
serine−type peptidase activity
carboxylic ester hydrolase activity

Figure 21

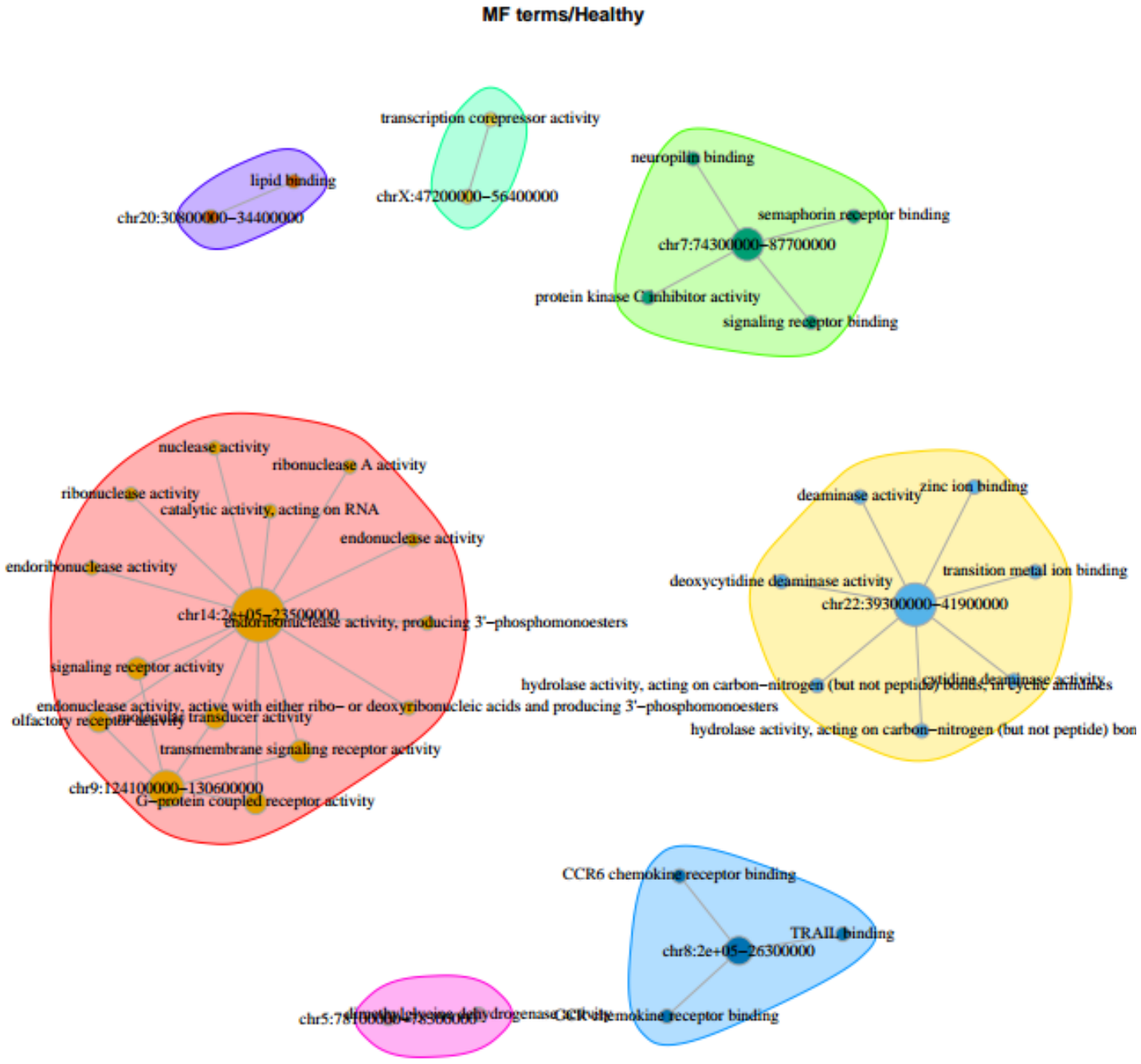

MF terms/Healthy
transcription corepressor activity
chrX:47200000−56400000
lipid binding
chr20:30800000−34400000
neuropilin binding
semaphorin receptor binding
chr7:74300000−87700000
protein kinase C inhibitor activity
signaling receptor binding
nuclease activity
ribonuclease A activity
ribonuclease activity
catalytic activity, acting on RNA
endonuclease activity
endoribonuclease activity
chr14:2e+05−23500000
endoribonuclease activity, producing 3'−phosphomonoesters
signaling receptor activity
endonuclease activity, active with either ribo− or deoxyribonucleic acids and producing 3'−phosphomonoesters
olfactory receptor activity
transmembrane signaling receptor activity
chr9:124100000−130600000
G−protein coupled receptor activity
deaminase activity
zinc ion binding
deoxycytidine deaminase activity
transition metal ion binding
chr22:39300000−41900000
hydrolase activity, acting on carbon−nitrogen (but not peptide) bonds, in cyclic amidines
hydrolase activity, acting on carbon−nitrogen (but not peptide) bon
CCR6 chemokine receptor binding
TRAIL binding
chr8:2e+05−26300000
CCR6 chemokine receptor binding

Figure 22

Συζήτηση

Πρόκληση της συγκεκριμένης εργασίας αποτέλεσε η διερεύνηση των δυνατοτήτων τμηματοποίησης του γονιδιώματος για δεδομένα μεταγραφής γονιδίων, τα οποία είναι μικρότερης διακριτικής ικανότητας και καλύπτουν ένα περιορισμένο μέρος του γονιδιώματος. Θα πρέπει να σημειωθεί ότι, στα πλαίσια της προσπάθειάς μας επιλέξαμε να επικεντρωθούμε στο τμήμα της πληροφορίας που αντιστοιχεί σε σχολιασμένα στοιχεία του μετραγραφώματος παρότι είναι γνωστό ότι η μεταγραφική ενεργότητα εκτείνεται σε μεγάλες περιοχές εκτός γνωστών γονιδίων (Clark MB, et al, 2011). Οι βασικοί λόγοι πίσω από αυτήν την επιλογή ήταν αφενός η δυσκολία απόδοσης λειτουργίας και κατά συνέπεια ερμηνείας της μεταγραφικής ενεργότητας εκτός γονιδίων και αφετέρου η δυσκολία ποσοτικοποίησης της "υποκείμενης" (pervasive) μεταγραφής ακόμα και με προσεγγίσεις NGS.

Σε ό,τι αφορά το μεθοδολογικό κομμάτι, το βασικό συμπέρασμά μας είναι ότι στην παρούσα φάση δεν υπάρχουν διαθέσιμα εργαλεία για την ανάλυση μεταγραφωμικών δεδομένων σε επίπεδο τμηματοποίησης του γονιδιώματος. Μια βιβλιογραφική αναζήτηση υπέδειξε το iSeg (Girimurugan, S.B., et al., 2018) ως την πιο υποσχόμενη μεθοδολογία για κάτι τέτοιο. Ωστόσο, η εφαρμογή του εργαλείου iSeg, ανέδειξε σημαντικούς περιορισμούς. Η μικρή διακριτική ικανότητα των δεδομένων που δόθηκαν ως input (gene counts) οδηγεί σε πολύ μικρά τμήματα τα οποία πρακτικά δεν διακρίνουν μεταξύ επιπέδων έκφρασης. Παρατηρήθηκε ότι τα επίπεδα έκφρασης των γονιδίων που βρέθηκαν μέσα στα τμήματα που προέκυψαν δεν διαφέρουν σημαντικά με τα επίπεδα έκφρασης των γονιδιών που βρέθηκαν στα συμπληρωματικά τμήματα (Figure 1). Αυτό μπορεί να εξηγηθεί από το γεγονός ότι τα τμήματα που προκύπτουν από το εργαλείο iSeg έχουν μικρό μήκος και είναι πολλά σε αριθμό (Figures 2,3). Οι περιορισμοί της εφαρμογής του iSeg είναι επίσης εμφανείς στις πολύ σημαντικές διαφορές, σε ό,τι αφορά το επίπεδο κάλυψης του γονιδιώματος από τμήματα υψηλότερης έκφρασης, μεταξύ των χρωμοσωμάτων, διαφορές που κατά πάσα πιθανότητα δεν σχετίζονται με τα ποσοτικά επίπεδα έκφρασης αλλά με εγγενείς ιδιότητες του γονιδιώματος όπως είναι η διαφορετική γονιδιακή πυκνότητα μεταξύ χρωμοσωμάτων. Συμπερασματικά μπορούμε να πούμε ότι, η προσαρμογή μιας δημοσιευμένης μεθοδολογίας σε μεταγραφωματικά δεδομένα δεν είναι αποδοτική καθώς η μέθοδος είναι εξαιρετικά ευαίσθητη σε δομικά χαρακτηριστικά του μεταγραφώματος.

Προκειμένου να ξεπεράσουμε τα εμπόδια που συνοψίσαμε παραπάνω, προσαρμόσαμε μια μεθοδολογία ανάλυσης σήματος στα πλαίσια των δεδομένων μας. Με σημείο εκκίνησης έναν αλγόριθμο που προέρχεται από το πεδίο της ταξινόμησης (Unbiased Recursive Partitioning) και συγκεκριμένα την προσαρμογή του στην ανάλυση χρονοσειρών (breakpoints), δημιουργήσαμε μια σειρά εντολών στην R που α) μπορεί να χειριστεί δεδομένα διαφορετικής διακριτικής ικανότητας μέσω λήψης μέσων τιμών σε μεταβλητό παράθυρο (running averages) β) να εκτιμήσει σημεία στατιστικά σημαντικών μεταβολών στο εξομαλυμένο (smoothed) σήμα μέσω της εφαρμογής του Chow test για τη σύγκριση συντελεστών παλινδρόμησης (βλ. Μέθοδοι). Η διαδικασία αυτή οδηγεί σε τμηματοποίηση σε περιοχές (Domains of Focal Over-Expression, DFOE) που έχουν τα επιθυμητά χαρακτηριστικά έκτασης και κάλυψης του γονιδιώματος. Το πρωταρχικό συμπέρασμα, που προκύπτει από την εφαρμογή της DFOE στο σύνολο δεδομένων ασθενών με ΣΕΛ, ήταν η σημαντικά μεγαλύτερη κάλυψη του γονιδιώματος από περιοχές αυξημένης έκφρασης στους ασθενείς σε σχέση με τους υγιείς μάρτυρες (controls). Η κάλυψη αυτή ήταν επιπλέον θετικά σχετιζόμενη με την ενεργότητα της ασθένειας, κάτι που υποδεικνύει ότι εκτός από αυξημένα επίπεδα έκφρασης, οι ασθενείς με ΣΕΛ τείνουν να εμφανίζουν υπερ-έκφραση σε εστιασμένες περιοχές του γονιδιώματος. Η παρατήρηση αυτή είναι σημαντική, ιδίως σε σχέση με γνωστές εκτεταμένες περιοχές που συνδέονται γενετικά με την νόσο (Tsokos,G.C. et al., 2011). Κατά την εκτέλεση της μεθόδου εύρεσης στατιστικά σημαντικών γονιδιωματικών τμημάτων που προέκυψαν από την εκτέλεση του αλγορίθμου DFOE, καταφέραμε να απομονώσουμε σημαντικές περιοχές τόσο στους υγιείς όσο και στους ασθενείς. Από την περαιτέρω

ανάλυση με GprofileR, παρατηρήθηκε ότι ένα μεγάλο μέρος των λειτουργίων, που προέκυψαν να σχετίζονται με τα γονίδια των ευρεθέντων τμημάτων, σχετίζονται με την ασθένεια ΣΛΕ. Ένα παράδειγμα είναι το σηματοδοτικό μονοπάτι JAK-STAT, όπου πρόσφατη έρευνα του σήματος μεταγωγής της κινάσης Janus (JAK) και του μονοπατιού σηματοδότησης του ενεργοποιητή μεταγραφής (STAT) αποκάλυψε παρεκκλίνουσα σηματοδότηση STAT σε φλεγμονώδεις καταστάσεις και αυτοάνοσες ασθένειες, συμπεριλαμβανομένου του ΣΛΕ (Alunno A, et. al,2019). Ένα ακόμα παράδειγμα αποτελεί η Th1 και Th2 κυτταρική διαφοροποίηση, όπου σε παλαιότερη εργασία έχει δειχθεί ανισορροπία των παραγόντων μεταγραφής Th1 / Th2 σε ασθενείς με νεφρίτιδα του λύκου (R. W.-Y. Chan, et. Al, 2006).

Συνοψίζοντας, σε αυτή την εργασία επιτεύχθηκε η δημιουργία ενός αποτελεσματικού εργαλείου γονιδιακής τμηματοποίησης (DFOE) και επιπλέον θεωρήσαμε ότι η μέθοδος τμηματοποίησης DFOE προσαρμόστηκε καλύτερα και αποτελεσματικότερα στα δεδόμενα μας σε αντίθεση με το εργαλείο iSeg, η εφαρμογή του οποίου ήταν, εκτός των άλλων, και αρκετά χρονοβόρα. Τα αποτελέσματα που προέκυψαν από την παραπάνω εργασία φαίνεται να χαρακτηρίζονται από αφθονία πληροφοριών. Περαιτέρω ανάλυση των παραπάνω δεδομένων θα μπορούσαν να οδηγήσουν τόσο στην αναγνώριση σημάτων που σχετίζονται με ασθένειες όσο και στην ανακάλυψη νέων βιοδεικτών.

6 Βιβλιογραφία

Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74.

Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc*. 2017;12(12):2478–2492. doi:10.1038/nprot.2017.124

Hon G, Ren B, Wang W. ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. *PLoS Comput Biol*. 2008;4(10):e1000201. doi:10.1371/journal.pcbi.1000201

Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods*. 2012;9(5):473–476. Published 2012 Mar 18. doi:10.1038/nmeth.1937

John D. Rioux , Abul K. Abbas: Paths to understanding the genetic basis of autoimmune disease. Nature 435, 584-589 (2 June 2005)

Mayami Sengupta, Laurence Morel. Lupus at the molecular level. Protein cell 2011, 2(12): 941-943

Danchenko N, Satia JA, Anthony MS. Epidemiology of systemic lupus erythematosus: a comparison of worldwide disease burden. Lupus. 2006; 15(5): 308-18

Panousis N. et al. Genomic dissection of Systemic Lupus Erythematosus: Distinct Susceptibility, Activity and Severity Signatures. BioRxiv 255109 (2018).

Tsokos, G. C. Systemic Lupus Erythematosus. N. Engl. J. Med. 365, 2110–2121 (2011).

Tsokos, G. C., Lo, M. S., Reis, P. C. & Sullivan, K. E. New insights into the immunopathogenesis of systemic lupus erythematosus. Nat. Rev. Rheumatol.12, 716–730 (2016).

Cleynen, A., Dudoit, S. & Robin, S. JABES (2014) 19: 101. https://doi.org/10.1007/s13253-013-0159-5

Kaul A,Gordon C,Crow MK,Touma Z,Urowitz MB,van Vollenhoven R,Ruiz-Irastorza G,Hughes G. Systemic lupus erythematosus. Nat Rev Dis Primers. 2016 Jun 16;2:16039. doi: 10.1038/nrdp.2016.39.

Girimurugan, S.B., Liu, Y., Lung, P. *et al*. iSeg: an efficient algorithm for segmentation of genomic and epigenomic data. *BMC Bioinformatics* **19,** 131 (2018) doi:10.1186/s12859-018-2140-3

Ran D. & Daye Z. J. Gene expression variability and the analysis of largescale RNA-seq studies with the MDSeq. Nucleic Acids Res. 45, (2017).

Anders S. & Huber W. Differential expression analysis for sequence count data. Genome Biol. 11, R106 (2010).

Alunno A, Padjen I, Fanouriakis A, Boumpas DT. Pathogenic and Therapeutic Relevance of JAK/STAT Signaling in Systemic Lupus Erythematosus: Integration of Distinct Inflammatory Pathways and the Prospect of Their Inhibition with an Oral Agent. *Cells*. 2019;8(8):898. Published 2019 Aug 15. doi:10.3390/cells8080898

R. W.-Y. Chan, F. M.-M. Lai, E. K.-M. Li, L.-S. Tam, K.-M. Chow, P. K.-T. Li, C.-C. Szeto, Imbalance of Th1/Th2 transcription factors in patients with lupus nephritis, *Rheumatology*, Volume 45, Issue 8, August 2006, Pages 951–957,https://doi.org/10.1093/rheumatology/kel029

Harley JB, Moser KL, Gaffney PM, Behrens TW. The genetics of human systemic lupus erythematosus. Curr Opin Immunol 1998;10:690-696

Gateva V, Sandling JK, Hom G et al. A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus. Nat Genet 2009;41:1228-1233

Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature2009;461:747-753

Hothorn,T., Hornik,K. and Zeileis,A. (2006) Unbiased Recursive Partitioning: A Conditional Inference Framework. *J. Comput. Graph. Stat.*, **15**, 651–674.

Zeileis,A., Leisch,F., Hornik,K., Kleiber,C., Zeileis,A., Leisch,F., Hornik,K. and Kleiber,C. (2001) Strucchange: An R package for testing for structural change in linear regression models.

Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, et al. (2011) The Reality of Pervasive Transcription. PLoS Biol 9(7): e1000625. https://doi.org/10.1371/journal.pbio.1000625