Άσκηση 6

Βιοπληροφορική / Υπολογιστική Βιολογία

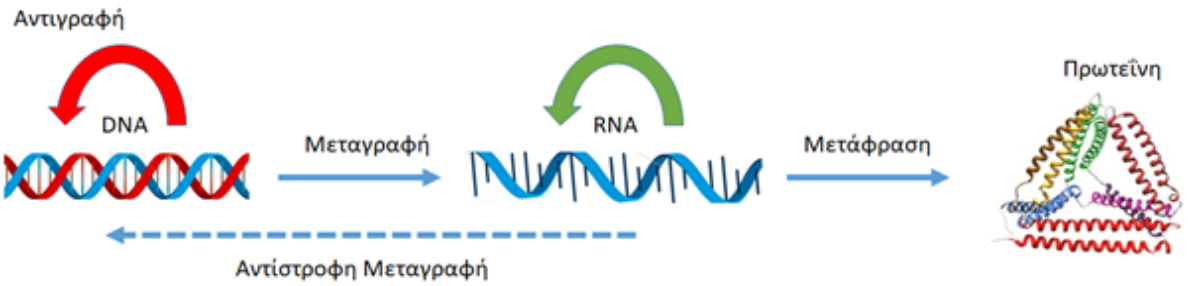
Γονίδια και μεταβολικές ασθένειες – Εύρεση θέσεων αναπαραγωγής του DNA για τη γονιδιακή θεραπεία μεταβολικών ασθενειών - Ανάλυση δεδομένων γονιδιακής έκφρασης από μικροσυστοιχίες DNA και πρακτική εξάσκηση στο πεδίο της Διατροφογενωμικής

1. Σκοπός Άσκησης

Σκοπός της άσκησης είναι να εξοικειώσει τους φοιτητές με έννοιες της Βιοπληροφορικής και της Υπολογιστικής Βιολογίας, καθώς και να τους ενημερώσει για τις βασικές τεχνικές ανάλυσης που χρησιμοποιούνται. Το θεωρητικό μέρος της άσκησης περιλαμβάνει εισαγωγικά στοιχεία, περιγραφή του ρόλου συγκεκριμένων γονιδίων στην εμφάνιση μεταβολικών ασθενειών, θεωρητικά στοιχεία για την εύρεση θέσεων αναπαραγωγής του DNA, περιγραφή τεχνολογιών για τη μέτρηση γονιδιακής έκφρασης (gene expression profiling) με έμφαση στις μικροσυστοιχίες DNA, καθώς και βασικά βήματα της ανάλυσης δεδομένων από πειράματα με μικροσυστοιχίες. Ένας από τους βασικούς στόχους της ανάλυσης είναι η εύρεση γονιδίων που παρουσιάζουν τη λεγόμενη διαφορική έκφραση (differential expression) μεταξύ καταστάσεων του πειράματος. Περιγράφονται τεχνικές κανονικοποίησης των δεδομένων, στατιστικών ελέγχων που εφαρμόζονται στα δεδομένα, καθώς και ανάλυσης των δεδομένων με ομαδοποίηση (clustering). Στο πρακτικό μέρος της άσκησης, οι φοιτητές έχουν τη δυνατότητα να εξασκηθούν σε βασικά βήματα της ανάλυσης βιολογικών δεδομένων. Η ανάλυση πραγματοποιείται με το περιβάλλον προγραμματισμού MATLAB καθώς και με τη γλώσσα R, η οποία είναι ιδιαίτερα διαδεδομένη στο πεδίο της Επιστήμης Δεδομένων (Data Science), καθώς και με κατάλληλα διαδικτυακά αναλυτικά εργαλεία και βάσεις δεδομένων, που προσφέρονται από οργανισμούς όπως το National Center for Biotechnology Information (NCBI).

2. Προετοιμασία Άσκησης

Για την κατανόηση της άσκησης είναι απαραίτητες βασικές γνώσεις Βιολογίας, σχετικά με την οργάνωση της γενετικής πληροφορίας στο κύτταρο και τη μετάδοσή της. Μπορείτε να ανατρέξετε σε κατάλληλες πηγές στο διαδίκτυο σχετικά με όρους όπως DNA, RNA, πρωτεΐνες, μεταγραφή, μετάφραση, κεντρικό δόγμα της Μοριακής Βιολογίας, γονιδίωμα, μεταγράφωμα κ.ά. Όπως φαίνεται στην Εικόνα 1, με βάση το κεντρικό δόγμα της Μοριακής Βιολογίας, η κωδικοποιημένη πληροφορία στα γονίδια μεταφέρεται στις πρωτεΐνες με την ακόλουθη ροή: DNA -> RNA -> πρωτεΐνες. Η σύνθεση ενός RNA αντιγράφου από το DNA ονομάζεται μεταγραφή, ενώ η σύνθεση πρωτεΐνης με βάση το RNA ονομάζεται μετάφραση.



Εικόνα 1: Το κεντρικό δόγμα της Μοριακής Βιολογίας

3. Γονίδια και μεταβολικές ασθένειες

3.1 Εισαγωγικά

Το DNA περιέχεται στον πυρήνα των κυττάρων, περιέχει γονίδια και αποτελεί το πρότυπο για τη σύνθεση των πρωτεϊνών καθώς και το μέσο μεταφοράς της κληρονομούμενης από γενιά σε γενιά πληροφορίας. Το DNA είναι ένα μακρομόριο, που αποτελείται από νουκλεοτίδια. Κάθε νουκλεοτίδιο αποτελείται από ένα μόριο σακχάρου, τη δεοξυριβόζη, ενωμένη με μία φωσφορική ομάδα και μία αζωτούχο βάση. Στα νουκλεοτίδια του DNA η αζωτούχος βάση μπορεί να είναι μία από τις: αδενίνη (A), γουανίνη (G), κυτοσίνη (C) και θυμίνη (T). Κάθε γονίδιο αποτελεί μία υποακολουθία του DNA και αποτελείται από κωδικοποιούσες περιοχές, τα εξώνια, και μη κωδικοποιούσες περιοχές, τα εσώνια. Μικρές μεταβολές στο DNA, όπως η μετάλλαξη ενός μόνο νουκλεοτιδίου, ή και μεγάλες μεταβολές, όπως η ταυτόχρονη μετάλλαξη πολλών γονιδίων ή ακόμα και ο μη φυσιολογικός αριθμός των χρωμοσωμάτων, σε συνδυασμό με περιβαλλοντικούς παράγοντες και παράγοντες που αφορούν τον τρόπο ζωής ενός ατόμου μπορούν να αποτελέσουν προδιαθεσικό παράγοντα ή και να προκαλέσουν την εμφάνιση μίας ή περισσοτέρων ασθενειών.

3.2 Γονίδια του βιολογικού ρολογιού και μεταβολικές ασθένειες

Το εσωτερικό βιολογικό ρολόι καθορίζει κιρκαδιανές μεταβολές στη συμπεριφορά και τη φυσιολογία, και ρυθμίζεται ανάλογα με το φως ή το σκοτάδι στο περιβάλλον, διατροφικές διαταραχές καθώς και άλλα ορμονικά σήματα. Ο όρος κιρκαδιανός προέρχεται από τις λατινικές λέξεις «circa» και «dies», που σημαίνουν «περίπου» και «ημέρα», αντίστοιχα. Οι κιρκαδιανοί ρυθμοί υπάρχουν στους περισσότερους ζωντανούς οργανισμούς, από τους μονοκύτταρους ως τα θηλαστικά, είναι εικοσιτετράωροι και επιτρέπουν στους ζωντανούς οργανισμούς να συγχρονίζονται με τον εξωτερικό κύκλο φωτός-σκοταδιού. Αρκετές μελέτες έχουν δείξει ότι το φως, το οποίο γίνεται αντιληπτό από ειδικά φωτοευαίσθητα γαγγλιακά κύτταρα του αμφιβληστροειδούς που προσεκβάλλουν στον υπερχιασματικό πυρήνα, είναι ο κύριος συντονιστής του κιρκαδιανού ρυθμού στον άνθρωπο [1]. Το κιρκαδιανό σύστημα εκτός από τον κύκλο ύπνου/εγρήγορσης και τους κύκλους λήψης γεύματος/νηστείας ρυθμίζει και φυσιολογικές διαδικασίες όπως είναι ο μεταβολισμός λιπιδίων και γλυκόζης, η θερμοκρασία του σώματος και η έκκριση ορμονών, επιτρέποντας τη βελτιστοποίηση της λήψης, χρήσης και αποθήκευσης της ενέργειας κατά τη διάρκεια της ημέρας.
Πολλές παράμετροι που σχετίζονται με το μεταβολισμό της γλυκόζης, όπως η ανοχή στη γλυκόζη, η ευαισθησία της ινσουλίνης, καθώς και τα επίπεδα γλυκόζης, γλυκαγόνης και ινσουλίνης στο πλάσμα παρουσιάζουν κιρκαδιανές μεταβολές. Στον άνθρωπο, το μέγιστο της έκκρισης της ινσουλίνης παρατηρείται κατά τη διάρκεια της ημέρας, ενώ κατά τη διάρκεια της νύχτας παρατηρείται μείωση της έκκρισης της ινσουλίνης και αύξηση της παραγωγής της γλυκόζης. Στα τρωκτικά, αυτό το μοτίβο έκκρισης της ινσουλίνης μετατοπίζεται χρονικά κατά 12 ώρες σύμφωνα με τη νυχτερινή τους δραστηριότητα.
Ο κιρκαδιανός ρυθμός μπορεί να απορρυθμιστεί εξαιτίας της μη φυσιολογικής λειτουργίας του βιολογικού ρολογιού ή εξαιτίας του αποσυγχρονισμού ανάμεσα στον υπερχιασματικό πυρήνα και το εξωτερικό περιβάλλον ή εξαιτίας του αποσυγχρονισμού ανάμεσα στον υπερχιασματικό πυρήνα και τα περιφερικά ρολόγια. Η αποδιοργάνωση του κιρκαδιανού ρυθμού μπορεί να οδηγήσει στην εμφάνιση ασθενειών όπως το Μεταβολικό Σύνδρομο, η Παχυσαρκία και ο Σακχαρώδης Διαβήτης Τύπου 2. Ο σημερινός τρόπος ζωής και συνήθειες, όπως η εργασία και η λήψη γεύματος κατά τη διάρκεια της νύχτας, η έκθεση σε τεχνητό φως κατά τη διάρκεια της νύχτας, και τροποποιημένα ωράρια ύπνου αποτελούν τους σημαντικότερους παράγοντες απορύθμισης του κιρκαδιανού ρυθμού. Στους εργαζόμενους σε βάρδιες, οι οποίοι αποτελούν και το κατεξοχήν παράδειγμα απορύθμισης του κιρκαδιανού ρυθμού, παρατηρούνται μεταβολές στην απόκριση των β-κυττάρων του παγκρέατος και στο μεταβολισμό των λιπιδίων και της γλυκόζης, καθώς και αυξημένος κίνδυνος εμφάνισης Μεταβολικού Συνδρόμου, Καρδιαγγειακών Παθήσεων, Καρκίνου, Παχυσαρκίας και Σακχαρώδους Διαβήτη Τύπου 2. Άτομα με τροποποιημένα ή μειωμένα ωράρια ύπνου εμφανίζουν αυξημένο Δείκτη Μάζας Σώματος, μειωμένη ανοχή στη γλυκόζη και αυξημένη αντίσταση στην ινσουλίνη. Σε παχύσαρκα άτομα ή άτομα με Σακχαρώδη Διαβήτη Τύπου 1 ή Τύπου 2 εμφανίζονται διαταραχές στο κιρκαδιανό ρυθμό, στην έκκριση της ινσουλίνης καθώς και στην ανοχή στη γλυκόζη. Ορισμένες από

αυτές τις διαταραχές έχουν παρατηρηθεί σε άτομα με Σακχαρώδη Διαβήτη που έχουν στερηθεί τον ύπνο τους.
Πειράματα που έχουν πραγματοποιηθεί σε πειραματόζωα με τροποποιημένα τα γονίδια του βιολογικού ρολογιού έχουν αποκαλύψει ότι ο κιρκαδιανός ρυθμός παίζει κεντρικό ρυθμιστικό ρόλο στο μεταβολισμό και την ομοιόσταση της γλυκόζης, και πιο συγκεκριμένα ότι η μη φυσιολογική λειτουργία των γονιδίων του βιολογικού ρολογιού μπορεί να οδηγήσει σε προδιαβήτη ή διαβήτη. Μελέτες της αλληλεπίδρασης των γονιδίων του βιολογικού ρολογιού με το μεταβολισμό της γλυκόζης στον άνθρωπο έδειξαν ότι γενετικές παραλλαγές του γονιδίου CLOCK συσχετίζονται με αυξημένη ευαισθησία (susceptibility) στην Παχυσαρκία και το Μεταβολικό Σύνδρομο. Γενετικές παραλλαγές (Genetic variants) του γονιδίου BMAL1 συσχετίζονται με την Υπέρταση, το Διαβήτη Κύησης και τον ΣΔΤ2, ενώ μονονουκλεοτιδικοί πολυμορφισμοί του γονιδίου PER2 έχουν συσχετιστεί με υψηλά επίπεδα γλυκόζης νηστείας και κοιλιακής παχυσαρκίας. Γενετικές παραλλαγές (Genetic variants) του γονιδίου CRY2 συσχετίζονται με το ΣΔΤ2. Ένας γενετικός πολυμορφισμός του υποδοχέα της μελατονίνης, μίας ουσίας η οποία παίζει καθοριστικό ρόλο στον καθορισμό των κιρκαδιανών ρυθμών, επίσης συσχετίζεται με μειωμένη έκκριση ινσουλίνης, διαβήτη κύησης και ΣΔΤ2. Συνοψίζοντας, υπάρχουν πολλές μελέτες οι οποίες υποδεικνύουν την ύπαρξη στενού δεσμού μεταξύ της δυσλειτουργίας των γονιδίων του βιολογικού ρολογιού και μεταβολικών ασθενειών, όπως το Μεταβολικό Σύνδρομο, η Παχυσαρκία και ο ΣΔΤ2 [2].

3.3 Φυλετικά γονίδια και Σακχαρώδης Διαβήτης

Ο Σακχαρώδης Διαβήτης (ΣΔ) αποτελεί σημαντικό πρόβλημα υγείας παγκοσμίως, με τον Παγκόσμιο Οργανισμό Υγείας να προειδοποιεί ότι περισσότεροι από 400 εκατομμύρια άνθρωποι θα νοσούν από ΣΔ το 2030. Ο ΣΔ Τύπου 2 (ΣΔΤ2), που αποτελεί την επικρατέστερη μορφή του ΣΔ, χαρακτηρίζεται από υψηλές τιμές συγκέντρωσης της γλυκόζης στο αίμα, οι οποίες οφείλονται είτε στην ελλιπή έκκριση της ορμόνης ινσουλίνης από το πάγκρεας είτε στην ινσουλινοαντίσταση των ιστών. Ο ΣΔΤ2 είναι αποτέλεσμα της πολύπλοκης αλληλεπίδρασης παραγόντων που αφορούν τον τρόπο ζωής του ατόμου και του γονιδιώματος που έχει κληρονομήσει.
Διαβητογενείς περιβαλλοντικοί παράγοντες επιδρούν διαφορετικά σε άτομα με διαφορετικό γενετικό υπόβαθρο. Έτσι, για τη μεγαλύτερη επιτυχία της πρόληψης και της θεραπείας του ΣΔΤ2 είναι απαραίτητη η κατανόηση των μηχανισμών αλληλεπίδρασης γονιδίου με γονίδιο και γονιδίου με το περιβάλλον.
Η μελέτη της εμφάνισης και εξέλιξης του ΣΔΤ2 στους Ινδιάνους της φυλής Πίμα στην Αριζόνα των ΗΠΑ έχει συμβάλει σημαντικά στην κατανόησή μας για το διαβήτη. Οι Ινδιάνοι της φυλής Πίμα, όπως πολλοί άλλοι γηγενείς πληθυσμοί, εκτοπίστηκαν από τα εδάφη τους με το Νόμο του 1902 (Reclamation Act). Η μεταβολή από τον αγροτικό τρόπο ζωής στον φτωχότερο οικονομικά και διατροφικά αστικό τρόπο ζωής οδήγησε στην έξαρση της εμφάνισης της παχυσαρκίας και του ΣΔΤ2 στον πληθυσμό αυτό. Τα πολύ ψηλά ποσοστά εμφάνισης του ΣΔΤ2 (>50%) και η πολύ πρώιμη έναρξή του, ακόμα και στην εφηβεία, στους Ινδιάνους της φυλής Πίμα σε αντίθεση με τους άλλους πληθυσμούς στην περιοχή με Ευρωπαϊκή καταγωγή και με παρόμοιο τρόπο ζωής, υποδεικνύουν την καταλυτική επίδραση των γονιδίων στην εμφάνιση του ΣΔΤ2 [3].

3.4 Πρακτικό μέρος

Ζητούμενο 3.1.: Βρείτε πληροφορίες για τα γονίδια INS και CLOCK χρησιμοποιώντας κατάλληλες διαδικτυακές πηγές (http://www.genenames.org/ και http://omim.org/) και επιβεβαιώστε ότι ο Σακχαρώδης Διαβήτης Τύπου 2 συνδέεται με το γονίδιο INS.

Ζητούμενο 3.2.: Βρείτε σε ποιο χρωμόσωμα και σε ποια γονιδιωματική περιοχή βρίσκεται το γονίδιο CLOCK χρησιμοποιώντας κατάλληλες διαδικτυακές πηγές (http://genome.ucsc.edu/).

Ζητούμενο 3.3.: Ανακτήστε την αλληλουχία του DNA του γονιδίου INS και της πρωτεΐνης που κωδικοποιεί, καθώς επίσης και των ορθόλογων γονιδίων στον χιμπαντζή και τον ποντικό.

Στην αναφορά σας θα πρέπει να απαντήσετε σε όλα τα παραπάνω Ζητούμενα 3.1. έως 3.3., καθώς και να συμπεριλάβετε τα απαραίτητα γραφήματα (screenshots) που αιτιολογούν τις απαντήσεις σας.

4. Εύρεση θέσεων αναπαραγωγής του DNA για τη γονιδιακή θεραπεία μεταβολικών ασθενειών

4.1 Εισαγωγικά

Η αναπαραγωγή του γονιδιώματος αποτελεί μία από τις πιο σημαντικές λειτουργίες του κυττάρου. Πριν τη διαίρεση του κυττάρου, επιτελείται η αναπαραγωγή του γονιδιώματος, ώστε κάθε ένα από τα θυγατρικά κύτταρα να κληρονομήσει το δικό του αντίτυπο του γονιδιώματος. Οι Watson και Crick στη μελέτη τους το 1953 περιέγραψαν τη διαδικασία της αναπαραγωγής του γονιδιώματος ως εξής: Οι δύο έλικες του αρχικού μορίου DNA ξετυλίγονται και η κάθε αλυσίδα αποτελεί πρότυπο για τη σύνθεση του νέου μορίου για κάθε ένα από τα θυγατρικά κύτταρα.
Η αναπαραγωγή του DNA ξεκινάει σε συγκεκριμένες γονιδιωματικές περιοχές που ονομάζονται περιοχές αναπαραγωγής (replication origin – ori) και πραγματοποιούνται από μοριακές μηχανές αντιγράφου (molecular copy machines) που ονομάζονται DNA πολυμεράσες.

4.2 Εύρεση θέσεων αναπαραγωγής του DNA

Η εύρεση των περιοχών αναπαραγωγής είναι ιδιαίτερης σημασίας όχι μόνο για την κατανόηση της λειτουργίας της αναπαραγωγής των κυττάρων αλλά και για την επίλυση σημαντικών βιοϊατρικών προβλημάτων με μεθόδους γονιδιωματικής θεραπείας. Οι μέθοδοι γονιδιωματικής θεραπείας χρησιμοποιούν γενετικά τροποποιημένα μινι-γονιδώματα, που ονομάζονται διανύσματα ιών (viral vectors) και μπορούν να διαπεράσουν τις κυτταρικές μεμβράνες, όπως οι πραγματικοί ιοί. Το 1990, η γονιδιωματική θεραπεία εφαρμόστηκε πρώτη φορά σε άνθρωπο για να αντιμετωπιστεί η Βαριά Συνδυασμένη Ανοσοανεπάρκεια σε ένα τετράχρονο κοριτσάκι.
Η κεντρική ιδέα της γονιδιωματικής θεραπείας είναι η μόλυνση ενός ατόμου με έλλειψη σε ένα κρίσιμο γονίδιο με ένα διάνυσμα ιών που περιέχει ένα τεχνητό γονίδιο το οποίο κωδικοποιεί τη θεραπευτική πρωτεΐνη για την υπό μελέτη ασθένεια. Για να βεβαιωθούν οι βιολόγοι ότι το διάνυσμα των ιών αναπαράγεται σωστά μέσα στο κύτταρο, πρέπει να γνωρίζουν τις ακριβείς περιοχές αναπαραγωγής του DNA.

4.3 Πρακτικό μέρος

Στη συνέχεια της άσκησης θα επικεντρωθούμε στο πρόβλημα της εύρεσης περιοχών αναπαραγωγής στο γονιδίωμα βακτηρίων. Θα πρέπει να υπάρχει κάποιο «κρυφό μήνυμα» στην περιοχή αναπαραγωγής που να καθορίζει ότι εκεί πρέπει να ξεκινήσει η διαδικασία της αναπαραγωγής. Πράγματι γνωρίζουμε ότι η έναρξη της αναπαραγωγής υλοποιείται με τη βοήθεια της DnaA, μίας πρωτεΐνης που συνδέεται σε ένα σύντομο τμήμα μέσα στην περιοχή αναπαραγωγής που ονομάζεται DnaA box.
Βασιζόμενοι στην υπόθεση ότι το DNA είναι μία γλώσσα, θα ψάξουμε για συχνές «λέξεις» μέσα στην περιοχή αναπαραγωγής του γονιδώματος του βακτηρίου. Θεωρούμε ότι αυτές οι συχνές «λέξεις» μπορεί να αποτελούν τις θέσεις που συνδέεται η πρωτεΐνη DnaA. Για παράδειγμα η «λέξη» ACTAT εμφανίζεται πολύ συχνά στην ακόλουθη συμβολοακολουθία:

ACA**ACTAT**GCAT**ACTAT**CGGGA**ACTAT**CCT

Ζητούμενο 4.1.: Υλοποιήστε συνάρτηση στο περιβάλλον προγραμματισμού MATLAB που να βρίσκει πόσες φορές εμφανίζεται μία δοσμένη «λέξη» σε ένα τμήμα γονιδιώματος και δοκιμάστε στη συνέχεια με τη βοήθεια αυτής της συνάρτησης να υπολογίσετε πόσες φορές εμφανίζεται η «λέξη» ATC στο ακόλουθο τμήμα γονιδιώματος.

```
atcaatgatcaacgtaagcttctaagcatgatcaaggtgctcacacagtttatccacaac
ctgagtggatgacatcaagataggtcgttgtatctccttcctctcgtactctcatgacca
cggaaagatgatcaagagaggatgatttcttggccatatcgcaatgaatacttgtgactt
gtgcttccaattgacatcttcagcgccatattgcgctggccaaggtgacggagcgggatt
acgaaagcatgatcatggctgttgttctgtttatcttgttttgactgagacttgttagga
tagacggtttttcatcactgactagccaaagccttactctgcctgacatcgaccgtaaat
tgataatgaatttacatgcttccgcgacgatttacctcttgatcatcgatccgattgaag
```

```
atcttcaattgttaattctcttgcctcgactcatagccatgatgagctcttgatcatgtt
tccttaaccctctatttttacggaagaatgatcaagctgctgctcttgatcatcgtttc
```

Ζητούμενο 4.2.: Υλοποιήστε συνάρτηση στο περιβάλλον προγραμματισμού MATLAB που να υπολογίζει ποιες είναι οι συχνότερα εμφανιζόμενες «λέξεις» μήκους k χαρακτήρων σε ένα τμήμα γονιδιώματος και έπειτα δοκιμάστε να βρείτε τις συχνότερα εμφανιζόμενες «λέξεις» μήκους 9 χαρακτήρων στο παραπάνω τμήμα του γονιδιώματος του βακτηρίου.

Στην αναφορά σας θα πρέπει να συμπεριλάβετε τις απαντήσεις σας στα παραπάνω Ζητούμενα 4.1. και 4.2., καθώς και τον κώδικα σε MATLAB που υλοποιήσατε.

5. Ανάλυση δεδομένων γονιδιακής έκφρασης από μικροσυστοιχίες DNA και πρακτική εξάσκηση στο πεδίο της Διατροφογενωμικής

5.1 Εισαγωγικά

Η μελέτη του γονιδιώματος και του τρόπου κωδικοποίησης και οργάνωσής του αποτελούν αντικείμενο της Γονιδιωματικής (Genomics). Στο πεδίο των λεγόμενων "omics" περιλαμβάνεται επίσης η Μεταγραφωμική (Transcriptomics), που μελετά τα μόρια τα οποία προκύπτουν από τη μεταγραφή του γονιδιώματος, όπως το mRNA. Η μελέτη των πρωτεϊνών, ως λειτουργικών μορίων που παράγονται με βάση το RNA, αποτελεί αντικείμενο του πεδίου της Πρωτεωμικής (Proteomics).

Τα γονίδια που περιλαμβάνονται στο γονιδίωμα ενός οργανισμού δεν είναι διαρκώς ενεργοποιημένα. Κάθε κύτταρο του οργανισμού, σε κάθε στιγμή και υπό συγκεκριμένες συνθήκες, παράγει διαφορετικούς συνδυασμούς και ποσότητες πρωτεϊνών, μέσω της έκφρασης (μεταγραφής) των σχετικών γονιδίων. Είναι πρακτικό και ταυτόχρονα αξιόπιστο να γίνεται ποσοτική μέτρηση του επιπέδου του mRNA που εκφράζεται στα κύτταρα, αντί των πρωτεϊνών. Για τη μελέτη του μεταγραφώματος (transcriptome), χρησιμοποιούνται πειράματα μεγάλης κλίμακας, στα οποία μελετάται η έκφραση των γονιδίων, πιθανές μεταβολές σε αυτήν (ενεργοποίηση/καταστολή γονιδίων) και η βιολογική σημασία των μεταβολών αυτών.

5.2 Τεχνολογίες για την καταγραφή της γονιδιακής έκφρασης

Οι κύριες τεχνολογίες που χρησιμοποιούνται για τη μέτρηση του εκφραζόμενου mRNA σε μεγάλη κλίμακα (genome-wide survey of the transcriptome) είναι κυρίως οι μικροσυστοιχίες DNA και η αλληλούχιση RNA (RNA sequencing). Κάθε μια από αυτές παρουσιάζει πλεονεκτήματα και μειονεκτήματα. Η αλληλούχιση RNA φαίνεται πως έχει πλέον επικρατήσει των μικροσυστοιχιών DNA, ωστόσο οι μικροσυστοιχίες DNA είναι μία ώριμη τεχνολογία για την οποία υπάρχει μεγάλη εμπειρία και γνώση ως προς την ανάλυση δεδομένων. Επιπλέον, υπάρχει πληθώρα διαθέσιμων δεδομένων προς ανάλυση σε δημόσια αποθετήρια. Η αλληλούχιση RNA είναι ακόμη μια ακριβή τεχνολογία και η ανάλυση των δεδομένων που προκύπτουν είναι απαιτητική. Από την άλλη, η αλληλούχιση RNA παρέχει μεγάλη ευελιξία ως προς τον εντοπισμό ακολουθιών. Αξίζει να σημειωθεί ότι παρά τον διαφορετικό τύπο δεδομένων που παράγουν οι δυο τεχνολογίες, η προσέγγιση της ανάλυσης για την εξαγωγή βιολογικών συμπερασμάτων εμφανίζει κοινά χαρακτηριστικά.

5.3 Σύντομη εισαγωγή στην τεχνολογία των μικροσυστοιχιών

Η τεχνολογία των μικροσυστοιχιών DNA χρησιμοποιείται για την παράλληλη μέτρηση ενός μεγάλου αριθμού μεταγράφων (transcripts) RNA σε ένα κυτταρικό δείγμα. Στη συνέχεια θα εξηγηθεί αδρά πώς γίνεται η ποσοτικοποίηση της έκφρασης των γονιδίων. Η παρουσίαση περισσότερων τεχνικών λεπτομερειών ξεφεύγει από τους σκοπούς της άσκησης. Η λειτουργία των μικροσυστοιχιών βασίζεται στην ιδιότητα του DNA να υβριδοποιείται, δηλαδή μονόκλωνα νουκλεϊκά οξέα μπορούν να αλληλεπιδράσουν με συμπληρωματικές σε αυτά ακολουθίες, προς δημιουργία δίκλωνου συμπλόκου. Οι μικροσυστοιχίες είναι ουσιαστικά μικρο-πλακίδια στα οποία έχουν τοποθετηθεί, σε διάταξη πλέγματος, ορισμένα μόρια νουκλεϊκών οξέων. Το κάθε πλακίδιο σχεδιάζεται έτσι ώστε να

περιλαμβάνει ακολουθίες από χαρακτηριστικά γονίδια του μελετώμενου οργανισμού. Αυτά τα ακινητοποιημένα πάνω στο πλακίδιο τμήματα DNA ονομάζονται ανιχνευτές (probes). Για τη μέτρηση του εκφραζόμενου mRNA ακολουθούνται τα εξής: πρώτα απομονώνεται το mRNA από το κυτταρικό δείγμα και έπειτα δημιουργείται, μέσω αντίστροφης μεταγραφής, το συμπληρωματικό DNA (cDNA), το οποίο σημαίνεται με κατάλληλη φθορίζουσα χρωστική ουσία. Στη συνέχεια, διάλυμα με το cDNA περνάει πάνω από το πλακίδιο και υβριδοποιείται με τους αντίστοιχους συμπληρωματικούς ανιχνευτές πάνω στη μικροσυστοιχία. Όταν έχει γίνει η υβριδοποίηση, ενεργοποιείται ο φθορισμός και μέσω της μέτρησης του επιπέδου φθορισμού μπορούμε να ποσοτικοποιήσουμε το εκφραζόμενο mRNA [4].

Τα αποτελέσματα των μετρήσεων σε πείραμα γονιδιακής έκφρασης με μικροσυστοιχίες έχουν τη μορφή δισδιάστατου πίνακα, όπως φαίνεται στην Εικόνα 2. Οι γραμμές του πίνακα αντιστοιχούν στους διαφορετικούς ανιχνευτές της μικροσυστοιχίας, ενώ οι στήλες του πίνακα αντιστοιχούν στα διαφορετικά δείγματα που έχουν καταμετρηθεί με τον ίδιο τύπο μικροσυστοιχίας στο πείραμα. Κάθε κελί του πίνακα περιλαμβάνει τη μέτρηση που έχει καταγραφεί για τον αντίστοιχο ανιχνευτή (γονίδιο) και για το αντίστοιχο δείγμα του πειράματος.

ID_REF	GSM162954	GSM162956	GSM162957	GSM162958	GSM162959
1007_s_at	218,27	255,46	137,36	254,60	121,07
1053_at	92,81	121,42	51,46	119,53	85,78
117_at	53,72	64,33	49,31	59,87	42,92
121_at	225,12	208,08	279,43	208,61	246,80
1255_g_at	11,71	11,62	12,69	12,15	13,14
1294_at	243,61	277,53	170,93	312,98	316,07
1316_at	52,19	52,38	54,02	49,59	51,79
1320_at	50,09	51,02	43,81	44,33	51,06
1405_i_at	112,19	87,33	114,67	405,02	136,76
1431_at	12325,61	11801,22	6845,59	10051,77	11614,58
1438_at	55,71	48,14	69,34	47,41	50,05
1487_at	337,21	441,02	222,22	382,95	328,81
1494_f_at	8941,16	7506,00	3975,62	6156,13	8837,92
1552256_a_at	954,65	1291,16	717,29	1354,82	1614,45
1552257_a_at	133,61	243,35	127,39	148,38	187,68

Εικόνα 2: Μέρος μετρήσεων πειράματος γονιδιακής έκφρασης με μικροσυστοιχίες DNA (μετά από σχετική προ-επεξεργασία των δεδομένων)

5.4 Ανάλυση ενός πειράματος γονιδιακής έκφρασης με μικροσυστοιχίες DNA

Σε αδρές γραμμές, τα στάδια που ακολουθούνται σε ένα πείραμα με μικροσυστοιχίες DNA περιλαμβάνουν: α) τον ορισμό του βιολογικού ερωτήματος που εξετάζεται, β) τον σχεδιασμό κατάλληλου πειράματος ώστε να διερευνηθεί το βιολογικό ερώτημα και τον προσδιορισμό του τύπου μικροσυστοιχίας που θα χρειαστεί για τις μετρήσεις, γ) την κατάλληλη επεξεργασία των δειγμάτων και την υβριδοποίηση, δ) την ανάγνωση των μετρήσεων από τη μικροσυστοιχία, ώστε να εξαχθεί το επίπεδο έκφρασης με βάση τις τιμές φθορισμού, και τέλος, ε) την επεξεργασία και ανάλυση δεδομένων με κατάλληλες μεθόδους για την εξαγωγή συμπερασμάτων. Στο πλαίσιο αυτής της εργαστηριακής άσκησης το ενδιαφέρον εστιάζεται στο στάδιο της ανάλυσης των δεδομένων.

Για τον σκοπό αυτό, οι ποσοτικοποιημένες μετρήσεις που έχουν προκύψει από τη μικροσυστοιχία πρέπει να υποστούν κατάλληλη προ-επεξεργασία, που περιλαμβάνει «καθαρισμό δεδομένων», έλεγχο ποιότητας και κανονικοποίηση. Στη συνέχεια, μπορούν να εφαρμοστούν βασικές τεχνικές ανάλυσης

δεδομένων μικροσυστοιχιών, όπως ο προσδιορισμός γονιδίων που εμφανίζουν διαφορική έκφραση και ο εντοπισμός ομάδων γονιδίων με κοινά πρότυπα έκφρασης.

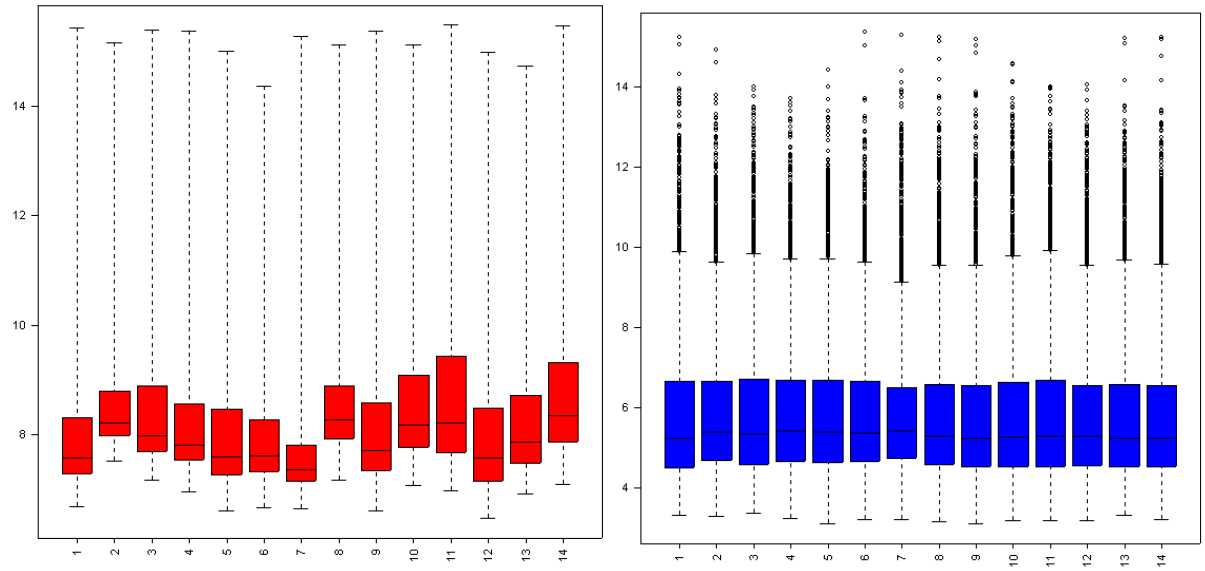
Κανονικοποίηση δεδομένων

Η κανονικοποίηση των δεδομένων μικροσυστοιχιών είναι πολύ σημαντική. Με την κανονικοποίηση μπορούμε να μετατρέψουμε τα δεδομένα από όλα τα δείγματα ενός πειράματος σε κοινή κλίμακα, ώστε να είναι δυνατή η σύγκρισή τους. Με αυτόν τον τρόπο αφαιρούνται συστηματικά σφάλματα που μπορεί να έχουν προκύψει από τεχνικούς λόγους και μη. Οι αρχικές μετρήσεις από ένα πείραμα με μικροσυστοιχία (μετρήσεις φθορισμού) μετασχηματίζονται με την εφαρμογή λογαρίθμου (συνήθως λογαρίθμου με βάση το 2). Οι αρχικές τιμές έντασης φθορισμού παρουσιάζουν μεγάλη διασπορά, ενώ με την κανονικοποίηση οι τιμές κατανέμονται ομοιόμορφα και έτσι μπορούν να συγκριθούν τα διάφορα δείγματα του πειράματος μεταξύ τους. Σε περίπτωση που θέλουμε να συγκρίνουμε δεδομένα από διαφορετικά πειράματα, τότε πρέπει να χρησιμοποιηθούν πιο «εξελιγμένες» μορφές κανονικοποίησης [4].

Θηκογράμματα (Boxplots)

Τα θηκογράμματα χρησιμοποιούνται στη στατιστική για να παρουσιάσουν με γραφικό τρόπο τα περιληπτικά μέτρα μια κατανομής και βοηθούν στη σύγκριση κατανομών μεταξύ τους. Στα πειράματα γονιδιακής έκφρασης χρησιμοποιούμε τα θηκογράμματα για να διαπιστώσουμε αν τα δεδομένα χρειάζονται κανονικοποίηση ή αν έχουν κανονικοποιηθεί με επιτυχία. Για κάθε δείγμα του πειράματος σχεδιάζεται το θηκόγραμμα της κατανομής των δεδομένων που έχουν μετρηθεί για όλα τα γονίδια στο δείγμα. Αυτό που αναμένεται είναι ότι για κανονικοποιημένα δεδομένα τα θηκογράμματα θα είναι συγκρίσιμα για όλα τα δείγματα. Μεγάλη απόκλιση στα θηκογράμματα αποτελεί ένδειξη ότι τα δεδομένα δεν είναι ισορροπημένα.

Στην Εικόνα 3 φαίνονται τα θηκογράμματα που έχουν προκύψει από ένα πείραμα γονιδιακής έκφρασης πριν και μετά την κανονικοποίηση των δεδομένων. Σε κάθε θηκόγραμμα σημειώνεται ένα ορθογώνιο, η κάτω πλευρά του οποίου δείχνει το πρώτο τεταρτημόριο και η άνω πλευρά δείχνει το τρίτο τεταρτημόριο της κατανομής. Η οριζόντια γραμμή εντός του ορθογωνίου δείχνει τη διάμεσο. Επιπλέον, είναι σημειωμένες η ελάχιστη και η μέγιστη τιμή της κατανομής (που υπολογίζονται αφού εξαιρεθούν οι ακραίες τιμές – outliers). Εφόσον το επιλέξουμε, τα θηκογράμματα μπορούν να εμφανίζουν και τις ακραίες τιμές. Αν οι τιμές της διαμέσου των κατανομών σε όλα τα δείγματα είναι στο ίδιο επίπεδο, αυτό είναι ένδειξη ότι τα δεδομένα είναι κανονικοποιημένα και κατά συνέπεια τα δείγματα μπορούν να συγκριθούν μεταξύ τους.



Εικόνα 3: Θηκογράμματα για κάθε δείγμα του πειράματος πριν την κανονικοποίηση (αριστερή εικόνα) και μετά την κανονικοποίηση (δεξιά εικόνα) των μετρήσεων [5]

Προσδιορισμός γονιδίων με διαφορική έκφραση και έλεγχος της στατιστικής σημαντικότητας

Στη συνέχεια εξετάζεται πώς ορίζονται τα γονίδια που εμφανίζουν διαφορική έκφραση μεταξύ καταστάσεων ενός πειράματος και πώς χρησιμοποιείται η στατιστική ανάλυση για να ελεγχθεί αν μια τέτοια διαφορική έκφραση είναι στατιστικά σημαντική. Ως διαφορετικές καταστάσεις σε ένα πείραμα μπορούν να οριστούν, για παράδειγμα: δείγματα φυσιολογικών κυττάρων έναντι παθολογικών, κυτταρικά δείγματα πριν και μετά την επίδραση κάποιου παράγοντα (π.χ. σε μελέτες Διατροφογενωμικής, αυτός ο παράγοντας μπορεί να είναι κάποιο διατροφικό στοιχείο), διαφορετικές φάσεις του κυτταρικού κύκλου του ίδιου δείγματος κυττάρων κ.ά. Ο προσδιορισμός των διαφορικά εκφραζόμενων γονιδίων είναι ιδιαίτερα χρήσιμος. Για παράδειγμα, σε πειράματα στα οποία οι μελετώμενες καταστάσεις είναι «φυσιολογικός ιστός» και «καρκινικός ιστός», η εύρεση στατιστικά σημαντικής διαφορικής έκφρασης μεταξύ των καταστάσεων μπορεί να υποδείξει γονίδια ως χαρακτηριστικά (features), που είναι δυνατό να χρησιμοποιηθούν διαγνωστικά με χρήση τεχνικών ταξινόμησης. Να σημειωθεί ότι ο στατιστικός έλεγχος υποθέσεων αποτελεί κατάλληλη προσέγγιση ανάλυσης για να αντιμετωπιστούν ιδιαιτερότητες των δεδομένων μικροσυστοιχιών, όπως ο τυχαίος θόρυβος. Όταν μια διαφορά είναι στατιστικά σημαντική, αυτό σημαίνει ότι δεν οφείλεται στην τυχαιότητα, αλλά έχει βιολογικό υπόβαθρο.

Για τον υπολογισμό της διαφορικής έκφρασης ενός γονιδίου μεταξύ δύο καταστάσεων (συνθηκών) του πειράματος, μπορεί να χρησιμοποιηθεί ο λογάριθμος (με βάση το 2) του λόγου των τιμών έκφρασης του γονιδίου στις δύο συνθήκες που συγκρίνονται (γνωστό και ως logFC – log fold change). Χρησιμοποιώντας όρους στατιστικής, η μία από τις συνθήκες ονομάζεται συνθήκη μελέτης (test), και η τιμή έκφρασης σε αυτήν τη συνθήκη τοποθετείται στον αριθμητή του λόγου, ενώ η άλλη είναι η συνθήκη ελέγχου (control), και η τιμή έκφρασης σε αυτή τη συνθήκη τοποθετείται στον παρονομαστή. Ανάλογα με την τιμή του λογαρίθμου προσδιορίζουμε τις σχετικές μεταβολές στην έκφραση π.χ. ενεργοποίηση, καταστολή έκφρασης γονιδίου [4].

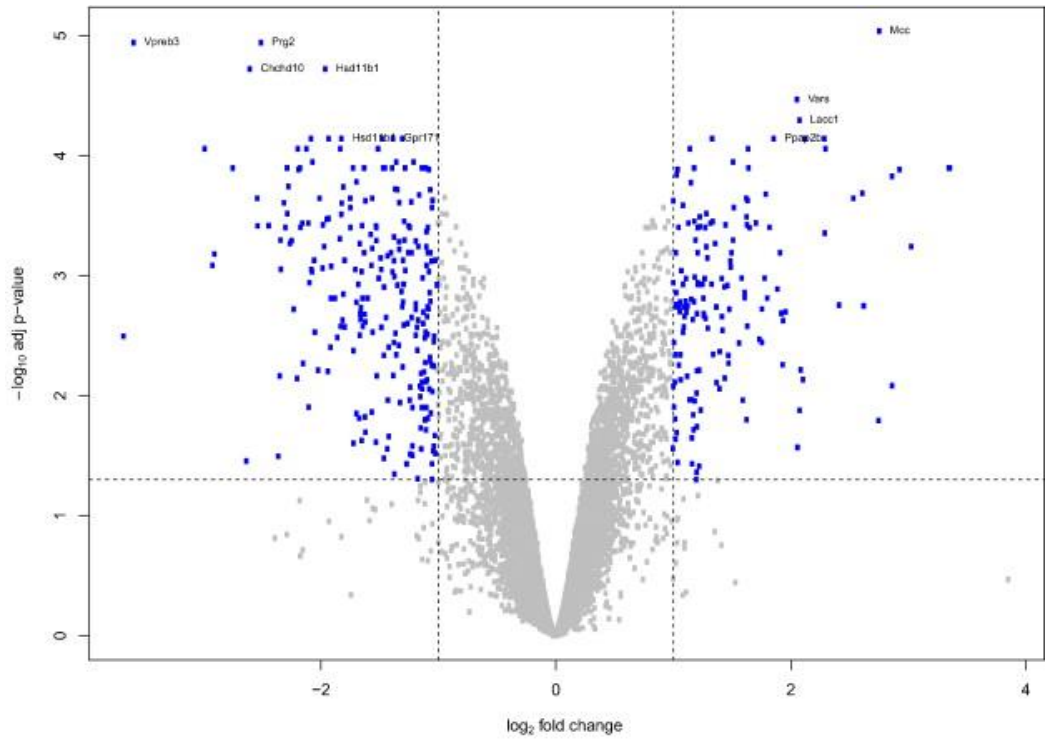
Υπάρχουν διάφορες τεχνικές στατιστικής ανάλυσης που χρησιμοποιούνται για την εύρεση των γονιδίων με στατιστικά σημαντική διαφορική έκφραση ανάμεσα στις καταστάσεις του πειράματος. Προκειμένου να προσδιορίσουμε τα γονίδια που παρουσιάζουν διαφορική έκφραση μεταξύ δύο συνθηκών ενός πειράματος έκφρασης (συνθήκη μελέτης/συνθήκη ελέγχου) μπορεί να χρησιμοποιηθεί ο έλεγχος στατιστικών υποθέσεων, με χρήση του ελέγχου t (Student's t-test ή απλά t-test). Για παράδειγμα, με t-test μπορεί να αναλυθεί ένα πείραμα με το οποίο μελετάμε τη γονιδιακή έκφραση στον ιστό πριν και μετά τη λήψη συγκεκριμένου διατροφικού στοιχείου. Για περισσότερες από δύο συγκρινόμενες συνθήκες στο πείραμα, μπορεί να χρησιμοποιηθεί η Ανάλυση Διακύμανσης (ANOVA – Analysis of Variance).

Η διαδικασία για τον έλεγχο στατιστικής υπόθεσης με t-test περιλαμβάνει: τον ορισμό της στατιστικής υπόθεσης, με τον προσδιορισμό της μηδενικής και της εναλλακτικής υπόθεσης, τον υπολογισμό της στατιστικής ελέγχου t και τέλος, την απόφαση για την απόρριψη ή όχι της μηδενικής υπόθεσης για το επίπεδο σημαντικότητας που έχει επιλεγεί. Στον έλεγχο υποθέσεων, ως μηδενική υπόθεση επιλέγουμε την υπόθεση που αντιπροσωπεύει ότι δεν υπάρχει αλλαγή στην παράμετρο που μελετάμε. Για παράδειγμα, στην ανάλυση πειραμάτων γονιδιακής έκφρασης, μπορούμε να θεωρήσουμε ως μηδενική υπόθεση (H_0) την υπόθεση ότι η μέση τιμή της έκφρασης του γονιδίου στις δύο συγκρινόμενες συνθήκες δεν αλλάζει. Η εναλλακτική υπόθεση (Ha) από την άλλη πλευρά, είναι το λογικό αντίθετο της μηδενικής υπόθεσης, ότι δηλαδή η μέση τιμή έκφρασης του γονιδίου στις δύο συνθήκες είναι διαφορετική. Όταν η εναλλακτική υπόθεση διατυπώνεται με αυτόν τον τρόπο, ο έλεγχος στατιστικής υπόθεσης καλείται δίπλευρος έλεγχος. Αν η εναλλακτική υπόθεση οριστεί ως η υπόθεση η μέση τιμή έκφρασης του γονιδίου στη συνθήκη μελέτης να είναι μόνο μεγαλύτερη ή μόνο μικρότερη της μέσης τιμής έκφρασης του γονιδίου στη συνθήκη ελέγχου, ο έλεγχος καλείται μονόπλευρος.

Στη συνέχεια της διαδικασίας ελέγχου στατιστικής υπόθεσης, εφαρμόζεται η στατιστική ελέγχου t-test και υπολογίζεται η τιμή του μεγέθους t, με βάση τα δείγματα που έχουμε και θεωρώντας ως δεδομένο ότι η μηδενική υπόθεση ισχύει. Με βάση την τιμή της στατιστικής ελέγχου t που υπολογίστηκε, μπορεί να βρεθεί από κατάλληλους στατιστικούς πίνακες μια πιθανότητα p

(probability), που είναι γνωστή ως p-τιμή (p-value). Η p-τιμή ορίζεται ως η πιθανότητα η τιμή της στατιστικής ελέγχου t να είναι ίση ή πιο «ακραία» από την τιμή που υπολογίστηκε με βάση την υπόθεση ότι ισχύει η μηδενική υπόθεση (ο ορισμός του τι αποτελεί ακραία τιμή εξαρτάται από το αν έλεγχος υπόθεσης είναι μονόπλευρος ή δίπλευρος). Σε περίπτωση που η p-τιμή είναι μικρότερη ή ίση από το επιλεγμένο επίπεδο σημαντικότητας, αυτό υποδηλώνει ότι είναι μικρή η πιθανότητα τα δεδομένα που παρατηρήθηκαν στα δείγματα να είναι σε συμφωνία με τη μηδενική υπόθεση και έτσι, η μηδενική υπόθεση μπορεί να απορριφθεί. Συνεπώς, δεχόμαστε την εναλλακτική υπόθεση (στο παράδειγμα μας, δεχόμαστε ότι υπάρχει διαφορά στη μέση τιμή έκφρασης του γονιδίου μεταξύ των δύο μελετώμενων συνθηκών). Διαφορετικά, η μηδενική υπόθεση δεν μπορεί να απορριφθεί.

Στα πειράματα στα οποία μελετάται η διαφορική έκφραση μεταξύ δύο καταστάσεων του πειράματος, χρησιμοποιούνται συχνά τα διαγράμματα «κρατήρα ηφαιστείου» (volcano plots), τα οποία ονομάζονται έτσι λόγω του χαρακτηριστικού σχήματός τους. Σε αυτά τα διαγράμματα, κάθε σημείο αναφέρεται σε ένα συγκεκριμένο γονίδιο και έχει συντεταγμένες, στον οριζόντιο άξονα, το λογάριθμο του λόγου έκφρασης που αναφέρθηκε πιο πάνω ($\log_2$FC), και στον κατακόρυφο άξονα, τον αρνητικό δεκαδικό λογάριθμο του p-value ($-\log_{10}$p-value). Όσο μεγαλύτερη είναι η (απόλυτη) τιμή της τετμημένης ενός σημείου, τόσο μεγαλύτερη διαφορική έκφραση εμφανίζει το γονίδιο αυτό, ενώ όσο μεγαλύτερη είναι η τιμή της τεταγμένης, τόσο πιο στατιστικά σημαντική είναι αυτή η διαφορική έκφραση. Με βάση τη βιβλιογραφία, υπάρχουν όρια που θεωρούνται αποδεκτά ως όρια διαφορικής έκφρασης και στατιστικής σημαντικότητας (για παράδειγμα, αποδεκτά όρια μπορεί να είναι τα $|\log_2 FC|>1,5$ και p-value<=0,05). Μπορούμε έτσι να θεωρήσουμε ως διαφορικά εκφραζόμενα (differentially expressed genes) τα γονίδια που πληρούν και τις δύο προϋποθέσεις ως προς αυτά τα όρια [4].

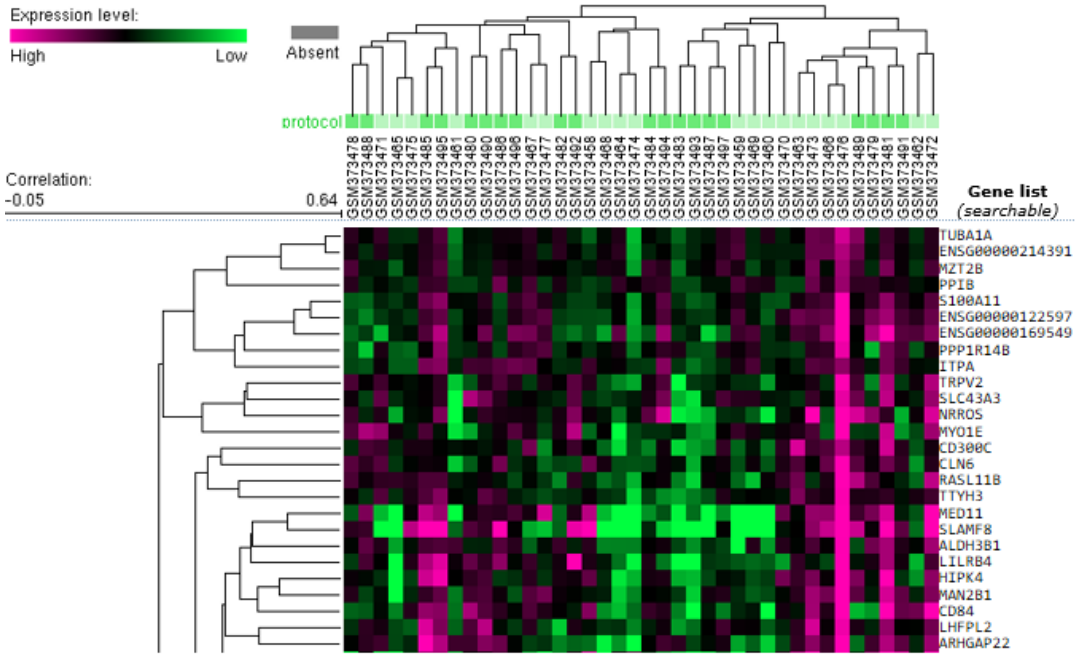


Εικόνα 4: Διάγραμμα «κρατήρα ηφαιστείου» από δεδομένα μικροσυστοιχίας. Τα όρια που έχουν επιλεγεί ως προς τη διαφορική έκφραση και τη στατιστική σημαντικότητα φαίνονται με τις διακεκομμένες γραμμές. Ως σημαντικά προς ανάλυση θεωρούνται τα γονίδια που εμφανίζονται με μπλε σημεία στο γράφημα. Για τα δέκα πιο σημαντικά γονίδια απεικονίζεται και το σύμβολο του γονιδίου. [6]

Ανάλυση ομαδοποίησης (cluster analysis)

Η ομαδοποίηση των δεδομένων γονιδιακής έκφρασης που βρίσκονται στη μορφή που παρουσιάστηκε στην Εικόνα 2, χρησιμεύει στον εντοπισμό προτύπων στα δεδομένα. Για παράδειγμα, η ομαδοποίηση μπορεί να γίνει στο επίπεδο των δειγμάτων (στήλες του πίνακα) και να μας επιτρέψει να εντοπίσουμε άγνωστες/καινούργιες κατηγορίες που υπάρχουν στα δείγματά μας (π.χ. αν τα δείγματα είναι από καρκινικούς ιστούς, να εντοπίσουμε διαφορετικές κατηγορίες τύπων καρκίνου στα δεδομένα γονιδιακής έκφρασης). Ομαδοποίηση μπορεί επίσης να γίνει στο επίπεδο των γονιδίων (γραμμές του πίνακα). Αυτό μπορεί να έχει εφαρμογή στην επιλογή γονιδίων ως χαρακτηριστικών (feature selection), που μπορούν έπειτα να χρησιμοποιηθούν σε μοντέλα πρόβλεψης. Ακόμη, μία τέτοια ομαδοποίηση μπορεί να χρησιμοποιηθεί για τον εντοπισμό ομάδων γονιδίων που εμφανίζουν συν-έκφραση [7]. Ως συνέκφραση χαρακτηρίζουμε τη συμμεταβολή της έκφρασης των γονιδίων υπό διαφορετικές πειραματικές καταστάσεις. Η εύρεση ομάδων γονιδίων που συνεκφράζονται είναι ιδιαίτερα σημαντική, γιατί μπορεί να θεωρηθεί ως ένδειξη κοινής λειτουργικότητας των γονιδίων. Μπορούμε έτσι να εκτιμήσουμε ποια είναι η άγνωστη λειτουργία ενός γονιδίου, παρατηρώντας με ποια άλλα γονίδια αυτό συνεκφράζεται, εφόσον γνωρίζουμε ήδη τη λειτουργία των άλλων γονιδίων (το επονομαζόμενο και «guilt by association»).

Πολύ χρήσιμο στην ανάλυση ομαδοποίησης είναι το γράφημα που ονομάζεται θερμικός χάρτης (clustered heatmap), γνωστό και ως διπλό δενδρόγραμμα, το οποίο μπορεί να προκύψει και με ιεραρχική ομαδοποίηση (hierarchical clustering). Οι θερμικοί χάρτες χρησιμοποιούνται συχνά στην ανάλυση και οπτικοποίηση πολυδιάστατων δεδομένων και η χρήση τους είναι συχνή στα πειράματα μεγάλης κλίμακας με μικροσυστοιχίες, καθώς διευκολύνουν την εύρεση τάσεων και προτύπων. Πιο συγκεκριμένα, ο θερμικός χάρτης είναι ένας πίνακας που σε κάθε κελί του απεικονίζεται το επίπεδο έκφρασης του αντίστοιχου γονιδίου στο αντίστοιχο δείγμα, με χρήση χρωματικής κλίμακας. Επιπλέον, οι στήλες και οι γραμμές του πίνακα είναι διατεταγμένες έτσι ώστε τα γονίδια και τα δείγματα που παρουσιάζουν παρόμοια πρότυπα να είναι ομαδοποιημένα το ένα δίπλα στο άλλο. Με αυτόν τον οπτικό τρόπο μπορούμε να εντοπίσουμε πιθανούς συσχετισμούς μεταξύ πειραματικών καταστάσεων και προτύπων έκφρασης, καθώς και ομάδες γονιδίων με κοινά πρότυπα έκφρασης.



Εικόνα 5: Λεπτομέρεια θερμικού χάρτη πειράματος μικροσυστοιχίας, που έχει προκύψει με χρήση της διαδικτυακής εφαρμογής GDSbrowser του NCBI.

5.5 Διατροφογενετική (Nutrigenetics) /Διατροφογενωμική (Nutrigenomics) και ανάλυση δεδομένων γονιδιακής έκφρασης

Ως Διατροφογενετική (Nutrigenetics) ορίζεται το ερευνητικό πεδίο που μελετά πώς τα διαφορετικά γενετικά χαρακτηριστικά των ατόμων μπορούν να οδηγήσουν σε διαφορετική απόκριση στο διαιτολόγιο και κατ' επέκταση σε διαφορετικό φαινότυπο. Για παράδειγμα, άτομα που ακολουθούν την ίδια διατροφή μπορεί, λόγω διαφορετικού γενετικού υλικού, να ανταποκριθούν διαφορετικά ως προς τα επίπεδα χοληστερόλης ορού [8].

Ο όρος Διατροφογενωμική (Nutrigenomics) από την άλλη, αναφέρεται στο ερευνητικό πεδίο που «ασχολείται με τον χαρακτηρισμό όλων των προϊόντων γονιδίων που επηρεάζονται από διατροφικά στοιχεία και τις μεταβολικές τους συνέπειες» [9]. Με άλλα λόγια, ασχολείται με το πώς τα προσλαμβανόμενα διατροφικά στοιχεία επηρεάζουν τη γονιδιακή απόκριση: τη μεταγραφή σε RNA, τις κωδικοποιούμενες πρωτεΐνες κ.ά.

Η Διατροφογενετική και η Διατροφογενωμική (Nutrigenetics/Nutrigenomics) μπορούν να χρησιμοποιηθούν στο πλαίσιο της λεγόμενης εξατομικευμένης διατροφής, όπου το προσωπικό διαιτολόγιο επιλέγεται με βάση το γονιδίωμα του ατόμου, με στόχο την προώθηση της υγείας και την πρόληψη/διαχείριση χρόνιων ασθενειών [8]. Για τον σκοπό αυτό, χρειάζεται κατανόηση του τρόπου με τον οποίο οι πολυάριθμες αλληλεπιδράσεις μεταξύ των διατροφικών στοιχείων, γονιδίων, πρωτεϊνών και μεταβολικών μονοπατιών, επηρεάζουν τα βιολογικά μονοπάτια ασθενειών (disease pathways).

Στο πλαίσιο της Διατροφογενωμικής μπορούν να σχεδιαστούν κατάλληλες μελέτες με πειράματα που εξετάζουν την επίδραση διατροφικών παραγόντων στη γονιδιακή έκφραση. Η μέτρηση της γονιδιακής έκφρασης μπορεί να γίνει με μικροσυστοιχίες DNA και η ανάλυση να ακολουθήσει τα στάδια που περιγράφηκαν στην παράγραφο 3.4 της άσκησης.

5.6 Πρακτικό μέρος

Στο πλαίσιο της εργαστηριακής άσκησης θα αναλυθούν δεδομένα από μικροσυστοιχίες DNA, που έχουν ανακτηθεί από το δημόσιο αποθετήριο Gene Expression Omnibus (GEO) του NCBI (www.ncbi.nlm.nih.gov/geo). Η ανάλυση πραγματοποιείται με χρήση της γλώσσας προγραμματισμού R και συγκεκριμένα μέσω του γραφικού περιβάλλοντος χρήστη (graphical user interface) RGui. Επιπλέον, χρησιμοποιούνται ορισμένα εργαλεία που είναι διαθέσιμα στον διαδικτυακό τόπο του NCBI.

Προκειμένου να απαντήσετε στα ζητούμενα που αναφέρονται στη συνέχεια, θα χρειαστεί να συμβουλευτείτε και τον κώδικα/σχόλια που δίνονται στο αρχείο Exercise6.R, που βρίσκεται στη σελίδα της άσκησης στο MyCourses.

Η γλώσσα προγραμματισμού R

Η γλώσσα προγραμματισμού R είναι ελεύθερο λογισμικό/λογισμικό ανοικτού κώδικα που χρησιμοποιείται ευρέως στο πεδίο της Βιοπληροφορικής/Υπολογιστικής Βιολογίας, καθώς υπάρχει πληθώρα διαθέσιμων πακέτων/βιβλιοθηκών στην R ειδικά για την επεξεργασία βιολογικών δεδομένων. Επίσης, η χρήση της R είναι πολύ διαδεδομένη γενικότερα στο πεδίο της Επιστήμης Δεδομένων (Data Science).

Μπορείτε να «κατεβάσετε» την R από τη σελίδα www.r-project.org/ . Η πιο πρόσφατη έκδοση της R είναι η 3.5.1. Στο διαδίκτυο υπάρχει πληθώρα οδηγών για την R, καθώς και πληροφορίες που μπορείτε να βρείτε από εγχειρίδια χρήσης για κάθε πακέτο και συνάρτηση. Ιδιαίτερα χρήσιμος είναι ο σύνδεσμος www.rdocumentation.org . Μπορείτε επίσης να συμβουλευτείτε τον εισαγωγικό οδηγό από την επίσημη σελίδα, στον σύνδεσμο cran.r-project.org/doc/manuals/r-release/R-intro.pdf . Είναι χρήσιμο να εξοικειωθείτε με το γραφικό περιβάλλον και με λειτουργίες όπως:

- ορισμός του working directory με την *setwd()*,
- εγκατάσταση και «φόρτωση» πακέτων (packages). Υπάρχουν πολλά διαθέσιμα πακέτα στο CRAN - Comprehensive R Archive Network. Η εγκατάσταση ενός πακέτου μπορεί να πραγματοποιηθεί με την εντολή *install.packages("thepackagename")* . Για να χρησιμοποιηθεί

το πακέτο μετά την εγκατάστασή του, πρέπει να φορτωθεί με χρήση της εντολής *library("thepackagename"),*

- διαχείριση διανυσμάτων και πινάκων στην R,
- χρήση δομών επανάληψης (βρόχοι – loops) στην R.

Ανάλυση δεδομένων πειράματος γονιδιακής έκφρασης

Τα δεδομένα που θα αναλυθούν προέρχονται από πείραμα γονιδιακής έκφρασης με μικροσυστοιχία DNA. Πρόκειται για το πείραμα με GEO id: **GSE7117** και τίτλο «Γονιδιακή έκφραση στο ήπαρ μετά από υποθερμιδική δίαιτα με χαμηλά λιπαρά, σε παχύσαρκες γυναίκες και σε μάρτυρες (controls)». Το πείραμα αυτό εντάσσεται στο πεδίο της Διατροφογενωμικής (Nutrigenomics), καθώς ασχολείται με το πώς μία συγκεκριμένη δίαιτα μπορεί να επηρεάσει το μεταβολικό προφίλ και την ηπατική γονιδιακή έκφραση στον άνθρωπο.

Ανάκτηση και επισκόπηση δεδομένων

Τα δεδομένα για το πείραμα GSE7117 που θα αναλυθούν είναι αποθηκευμένα με τη μορφή αρχείου δεδομένων μικροσυστοιχιών που λέγεται «Series Matrix File» (.txt αρχείο). Αυτή η μορφή είναι ιδιαίτερα εύχρηστη, καθώς τα δεδομένα έχουν ήδη υποστεί προ-επεξεργασία και κανονικοποίηση και είναι έτοιμα για περαιτέρω ανάλυση. Βέβαια, για κάθε πείραμα στο GEO υπάρχουν διαθέσιμα πειραματικά δεδομένα σε διάφορες μορφές, όπως και τα αρχικά ανεπεξέργαστα (raw) δεδομένα.

Ζητούμενο 5.1.: Να αναζητήσετε στην ηλεκτρονική σελίδα του Gene Expression Omnibus του NCBI, στην υπηρεσία Accession Display, το πείραμα με GEO accession id: GSE7117 (στον σύνδεσμο www.ncbi.nlm.nih.gov/geo/query/acc.cgi). Αφού διαβάσετε τη σύνοψη του πειράματος (Summary) να αναφέρετε πόσα δείγματα μετρήθηκαν, ποιες είναι οι συγκρινόμενες καταστάσεις στο πείραμα και πόσα δείγματα ανήκουν σε κάθε κατάσταση. Να αναφέρετε επίσης τον τύπο της πλατφόρμας μικροσυστοιχίας DNA με την οποία έγιναν οι μετρήσεις.

Ζητούμενο 5.2.: Να «κατεβάσετε» από τη σελίδα της άσκησης στο MyCourses το αρχείο GSE7117_series_matrix.txt. Όπως αναφέρθηκε, τα δεδομένα που είναι διαθέσιμα στη μορφή Series Matrix File έχουν ήδη υποστεί κανονικοποίηση. Να ανοίξετε το .txt αρχείο με ένα πρόγραμμα επεξεργασίας .txt αρχείων και να παραθέσετε τις ακόλουθες πληροφορίες για το πείραμα:

i. Ποιος είναι ο τίτλος κάθε δείγματος και ποιο το Sample_geo_accession id του;
ii. Από τι είδους οργανισμό προέρχονται τα δείγματα;
iii. Σε ποια από τις δύο συγκρινόμενες συνθήκες του πειράματος ανήκει κάθε δείγμα;

Να χρησιμοποιήσετε τη συνάρτηση read.table() που παρατίθεται στο αρχείο Excercise6.R για το Ζητούμενο 2, προκειμένου να εκχωρήσετε το τμήμα του .txt αρχείου που περιέχει τις πειραματικές μετρήσεις στο data frame (matrix-like object) με όνομα **x**.

iv. Να εξηγήσετε τι σημαίνουν τα ορίσματα που έχουν δοθεί στον κώδικα για την read.table().
v. Να χρησιμοποιήσετε κατάλληλη εντολή και να παραθέσετε ποιες είναι οι διαστάσεις του x.
vi. Να χρησιμοποιήσετε κατάλληλη εντολή και να παραθέσετε τα ονόματα κάθε στήλης του x. Σε τι αντιστοιχούν οι διαφορετικές στήλες;
vii. Να χρησιμοποιήσετε κατάλληλη εντολή και να παραθέσετε τα ονόματα των 15 πρώτων γραμμών του x. Σε τι αντιστοιχούν οι διαφορετικές γραμμές του x;
viii. Ποια είναι η τιμή που έχει καταγραφεί για το στοιχείο x[3,5] και τι περιγράφει η τιμή αυτή;
ix. Τι πληροφορίες μας δίνει η 200η γραμμή και τι η 7η στήλη του πίνακα x;
x. Να χρησιμοποιήσετε κατάλληλη συνάρτηση της R για να σχεδιάσετε το ιστόγραμμα συχνοτήτων για τις μετρήσεις που έχουν καταγραφεί για το πρώτο και το τρίτο δείγμα του πειράματος.

Κανονικοποίηση δεδομένων

Ζητούμενο 5.3.:

i. Να εφαρμόσετε την εντολή που αναφέρεται στον κώδικα για το Ζητούμενο 3-i και να παραθέσετε το αποτέλεσμα. Τι δείχνει το αποτέλεσμα; Να εξηγήσετε τις συναρτήσεις και τα ορίσματα που χρησιμοποιούνται σε αυτή τη γραμμή κώδικα.
ii. Να ελέγξετε αν τα δεδομένα είναι όντως κανονικοποιημένα, δημιουργώντας θηκογράμματα (boxplots) για τις μετρήσεις κάθε δείγματος. Συμβουλευτείτε τα σχόλια που παρατίθενται στον κώδικα γι' αυτό το ζητούμενο. Να παραθέσετε το διάγραμμα που προέκυψε και να σχολιάσετε.

Προσδιορισμός διαφορικά εκφραζόμενων γονιδίων

Ζητούμενο 5.4.: Στο ζητούμενο αυτό θα κάνετε έλεγχο στατιστικής υπόθεσης με t-test για τον προσδιορισμό των διαφορικά εκφραζόμενων γονιδίων μεταξύ των δειγμάτων χωρίς διαιτητική παρέμβαση (controls) και με διαιτητική παρέμβαση (diet intervention), συμπληρώνοντας κατάλληλα τον κώδικα για το ζητούμενο.

i. Να συμπληρώσετε κατάλληλα τη συνάρτηση c(), η οποία εκχωρεί στο διάνυσμα (vector) "xsamplelabels" τιμές που δίνουν την πειραματική συνθήκη για κάθε δείγμα. Χρησιμοποιήστε το 0 για τα δείγματα ελέγχου και το 1 για τα δείγματα με διαιτητική παρέμβαση.
ii. Να χρησιμοποιήσετε τη συνάρτηση t.test() της R και να γράψετε κώδικα για να κάνετε έλεγχο υπόθεσης t-test για κάθε ανιχνευτή για τις δύο πειραματικές συνθήκες που αναφέρθηκαν. Να υπολογίσετε την τιμή του p-value για κάθε ανιχνευτή (γονίδιο) και να παραθέσετε στην αναφορά σας τις τιμές του p-value για τους 15 πρώτους ανιχνευτές της μικροσυστοιχίας.
iii. Να γράψετε κώδικα και να παραθέσετε τους ανιχνευτές για τους οποίους το p-value που υπολογίσατε με το t-test ήταν μικρότερο από την τιμή 0,001. Επίσης, να βρείτε τον ανιχνευτή με το μικρότερο p-value.

Ζητούμενο 5.5.: Να κάνετε ανάλυση διαφορικής έκφρασης χρησιμοποιώντας αυτή τη φορά την υπηρεσία GEO2R του NCBI στον σύνδεσμο www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE7117. Μπορείτε να βρείτε τις οδηγίες χρήσης στη σελίδα του GEO2R (σύνδεσμος www.youtube.com/watch?v=EUPmGWS8ik0). Να ορίσετε πρώτα ποια δείγματα αντιστοιχούν στις δύο καταστάσεις του πειράματος: χωρίς διαιτητική παρέμβαση (control group)/με διαιτητική παρέμβαση (diet intervention group), και να προσδιορίσετε τα 250 πρώτα πιο σημαντικά γονίδια που εκφράζονται διαφορικά, χρησιμοποιώντας τη σχετική υπηρεσία που προσφέρεται.

i. Να παραθέσετε όλα τα αποτελέσματα που εξάγονται για τα δυο πρώτα πιο σημαντικά γονίδια που προκύπτουν. Τι έχετε να σχολιάσετε ως προς τις τιμές p-value και logFC που έχουν, με βάση και τα αναφερόμενα στο θεωρητικό μέρος; Η έκφραση των δύο γονιδίων καταστέλλεται ή ενισχύεται; Επιλέξτε πάνω στη γραμμή για τον πρώτο ανιχνευτή (γονίδιο) και παραθέστε το γράφημα που εικονίζεται. Τι ακριβώς δείχνει;
ii. Να παραθέσετε ορισμένες συνοπτικές πληροφορίες για το γονίδιο το οποίο βρήκατε ότι εκφράζεται διαφορικά σε στατιστικά σημαντικότερο βαθμό με το GEO2R, κάνοντας αναζήτηση στην υπηρεσία Gene του NCBI, στον σύνδεσμο https://www.ncbi.nlm.nih.gov/gene. Αναζητήστε πληροφορίες σχετικές με το γονίδιο στον άνθρωπο (Homo Sapiens).

Ομαδοποίηση δεδομένων

Ζητούμενο 5.6.: Να δημιουργήσετε θερμικό χάρτη (heatmap) με τα δεδομένα του πειράματος GSE7117, χρησιμοποιώντας την υπηρεσία Dataset Browser που προσφέρεται από το NCBI στον σύνδεσμο www.ncbi.nlm.nih.gov/sites/GDSbrowser (επιλογή Cluster Analysis).

i. Πόσες γραμμές και πόσες στήλες έχει ο πλήρης θερμικός χάρτης;

Να εστιάσετε στο πάνω μέρος του θερμικού χάρτη, ώστε να φαίνονται οι λεπτομέρειες, ακολουθώντας τις οδηγίες που αναφέρονται στην ιστοσελίδα.

ii. Να παραθέσετε τμήμα της λεπτομέρειας του θερμικού χάρτη που προέκυψε, ώστε να φαίνονται οι βασικές πληροφορίες που αυτός περιλαμβάνει. Τι απεικονίζεται σε κάθε γραμμή και στήλη και τι επιπλέον πληροφορίες δίνονται;

iii. Με βάση τον χρωματικό κώδικα που δίνεται, πώς θα χαρακτηρίζατε την τιμή έκφρασης στο πρώτο γονίδιο στη λίστα για το πρώτο δείγμα;

iv. Με βάση την ομαδοποίηση που προέκυψε, μπορείτε να κάνετε κάποιες παρατηρήσεις σχετικά με πιθανούς συσχετισμούς μεταξύ των καταστάσεων του πειράματος και των προτύπων έκφρασης ομάδων γονιδίων; Να αιτιολογήσετε το σκεπτικό σας.

Στην αναφορά σας θα πρέπει να συμπεριλάβετε τις απαντήσεις σε όλα τα παραπάνω Ζητούμενα 5.1. έως 5.6. και τα υποερωτήματά τους, τα σχετικά γραφήματα που ζητούνται, καθώς και τον συμπληρωμένο κώδικα του αρχείου Exercise6.R.

Βιβλιογραφία

[1] Τσαούσογλου, Μ., Μπερή, Δ., Βγόντζας, Α., Χρούσος, Γ., «Μοριακοί μηχανισμοί κιρκάδιων ρυθμών: μελέτη σε πειραματόζωα και οι πρώτες ενδείξεις στον άνθρωπο», Δελτία Α' Παιδιατρικής Κλινικής Πανεπιστημίου Αθηνών, νο. 53, 2006.

[2] Vieira, E., Burris, T., Quesada, I., "Clock genes, pancreatic function, and diabetes", Trends in Molecular Medicine, vol. 20, no. 12, 2014.

[3] Franks, P., Merino, J., "Gene-lifestyle interplay in type 2 diabetes", Current Opinion in Genetics & Development, no. 50, pp.35-40, 2018.

[4] Νικολάου, Χ., Χουβαρδάς, Π., "Υπολογιστική βιολογία", Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, 2015. [ηλεκτρ. βιβλ.] Διαθέσιμο στο: http://hdl.handle.net/11419/1577

[5] www.bioconductor.org, "MalariaLifeCycle - Statistical Microarray Analysis using affylmGUI" Available from: www.bioconductor.org/packages/devel/bioc/vignettes/affylmGUI/inst/doc/LifeCycle/MalariaLifeCycle.html#RawIntensityBoxPlot

[6] Xie, Ping & Moore, Carissa & R. Swerdel, Mavis & Hart, Ronald, "Transcriptomic profiling of splenic B lymphomas spontaneously developed in B cell-specific TRAF3-deficient mice», Genom Data. 2. 386-388. 10.1016/j.gdata.2014.10.017, 2014.

[7] Cancer Research UK Cambridge Institute, "Microarray-analysis, Materials on the analysis of microarray expression data; focus on re-analysis of public data", 2018. [online] Available from: http://bioinformatics-core-shared-training.github.io/microarray-analysis/

[8] Simopoulos A. P., "Nutrigenetics/Nutrigenomics," Annu. Rev. Public Health, vol. 31, no. 1, pp. 53–68, 2010.

[9] Ordovas J. M., Ferguson L. R., Tai E. S., Mathers J. C., "Personalised nutrition and health," BMJ, vol. 361, pp. 1–7, 2018.

[10] Universitat de Barcelona, “Introduction to Microarray and Next Generation Sequencing Data Analysis, Practicals”, 2018. [online] Available from: http://www.ub.edu/stat/docencia/bioinformatica/microarrays/ADM/