Κεφάλαιο 7

Ανάλυση της Γονιδιακής Έκφρασης

Σύνοψη

Στο δεύτερο αυτό μέρος του βιβλίου, η θεματολογία περνά από την ανάλυση των γονιδιωματικών αλληλουχιών, στη μελέτη ανώτερης τάξης βιολογικών δεδομένων. Βασικό αντικείμενο αυτού του κεφαλαίου θα είναι η μελέτη και ανάλυση δεδομένων γονιδιακής έκφρασης. στο πλαίσιο της παρουσίασης του φαινομένου και τους είδους των πειραματικών δεδομένων που καλούμαστε να αναλύσουμε, θα γίνει αρχικά μια εισαγωγή στις σχετικές μεθοδολογίες αιχμής με έμφαση σε αυτές των μικρο-συστοιχειών και της αλληλούχισης RNA (RNASeq). Σημείο εκκίνησης θα είναι το πραγματικό βιολογικό πρόβλημα της εξαγωγής γονιδίων με διαφορική έκφραση από πειράματα σε γονιδιωματική κλίμακα. Αφού παρουσιαστεί λεπτομερώς, η διαδοχή των βημάτων για μια πλήρη ανάλυση της γονιδιακής έκφρασης, θα συζητηθούν τεχνικές κανονικοποίησης των πρωτογενών δεδομένων. Σε μαθηματικό επίπεδο θα παρουσιαστούν, στη συνέχεια, οι τεχνικές σύγκρισης μεταξύ των τιμών δειγμάτων και οι έλεγχοι υποθέσεων που οδηγούν στην εξαγωγή καταλόγων διαφορικά εκφραζόμενων γονιδίων μεταξύ δύο διαφορετικών συνθηκών, καθώς και τρόποι αναπαράστασης των δεδομένων. Στο δεύτερο μέρος του κεφαλαίου, και με σκοπό την ανάδειξη συστάδων γονιδίων με κοινά πρότυπα έκφρασης, θα παρουσιαστούν βασικές μεθοδολογίες ομαδοποίησης δεδομένων (clustering) όπως η ανάλυση κύριων συνιστωσών (PCA), η ιεραρχική ομαδοποίηση και η ομαδοποίηση κ-μέσων (k-means).

Στο τέλος του Κεφαλαίου θα πρέπει να μπορείτε:

- *Να χειριστείτε αρχεία από ένα πείραμα γονιδιακής έκφρασης και να εξάγετε καταλόγους κανονικοποιημένων τιμών έκφρασης.*
- *Να πραγματοποιήσετε τους απαραίτητους στατιστικούς ελέγχους για την εξαγωγή διαφορικά εκφραζόμενων γονιδίων.*
- *Να κατανοήσετε γραφικές αναπαραστάσεις αποτελεσμάτων από πειράματα έκφρασης, όπως τα διαγράμματα "κρατήρα ηφαιστείου" (volcano plots) και οι θερμικοί χάρτες (heatmaps).*
- *Να ομαδοποιήσετε γονίδια με βάση τα πρότυπα έκφρασής τους σε διαφορετικές συνθήκες με διάφορες μεθόδους.*

Εισαγωγή

Μέχρι στιγμής έχουμε συζητήσει για προβλήματα που σχετίζονται με τις βιολογικές (γονιδιωματικές και πρωτεϊνικές) αλληλουχίες. Στα κεφάλαια που απομένουν συνολικά θα ασχοληθούμε κυρίως με προβλήματα που σχετίζονται με τα προϊόντα αυτών των αλληλουχιών. Σ' αυτό το δεύτερο μισό του βιβλίου, το ενδιαφέρον μας μετατοπίζεται έτσι από το πεδίο της κωδικοποίησης και της οργάνωσης του γονιδιώματος στις λειτουργικά χαρακτηριστικά βιολογικών οντοτήτων όπως είναι τα μόρια του mRNA και οι πρωτεΐνες.

Οι αλληλουχίες αποτελούν τη βασική "σταθερά" όλων των ζωντανών οργανισμών με την έννοια ότι είναι αυτές που καθορίζουν τη φύση τους, όμως τα πιο εντυπωσιακά από τα χαρακτηριστικά της ζωής προέρχονται ακριβώς από την ευελιξία που επιδεικνύουν οι ίδιοι οι οργανισμοί στην αξιοποίηση των δυνατοτήτων που τους "παρέχει" η γονιδιωματική τους αλληλουχία. Μ' αυτόν τον τρόπο, η οργάνωση διαφορετικών λειτουργιών στο χωρό και τον χρόνο, η δυνατότητα απόκρισης σε εξωτερικά ερεθίσματα, η διατήρηση ενεργειακών ισοζυγίων και χημικών ισορροπιών εντός του κυττάρου, η προγραμματισμένη διαίρεσή του, η επικοινωνία με άλλα κύτταρα και (στους πιο πολύπλοκους πολυκύτταρους οργανισμούς) η εξειδίκευση λειτουργιών σε ιστούς και όργανα είναι διαφορετικές εκφάνσεις της πολυπλοκότητας που προκύπτει από τη συντονισμένη διαχείριση της γονιδιωματικής πληροφορίας. Ένα δεδομένο κύτταρο "επιλέγει" να χρησιμοποιήσει αυτήν την πληροφορία με διαφορετικό τρόπο κάτω από διαφορετικές συνθήκες μέσω της παραγωγής συγκεκριμένων συνδυασμών πρωτεϊνικών μορίων σε συγκεκριμένες ποσότητες, ενεργοποιώντας ή καταστέλλοντας τη μεταγραφή των αντίστοιχων γονίδιων τους[1]. Λέμε τότε ότι το κύτταρο "εκφράζει" τα γονίδια του με συγκεκριμένο τρόπο. Η διαδικασία αυτή της έκφρασης είναι εξαιρετικά συντονισμένη και ενέχει χαρακτηριστικά προγράμματος (μιλάμε συχνά για "προγράμματα έκφρασης") .

Στα επόμενα δύο κεφάλαια θα εξετάσουμε ερωτήματα που σχετίζονται με τρόπους μελέτης της έκφρασης γονιδίων σε ό,τι αφορά τόσο τις μεθοδολογίες για την μέτρηση και την ανάλυση του βαθμού ενεργοποίησης και καταστολής της μεταγραφής γονιδίων, όσο και τις προσεγγίσεις για τη βιολογική ερμηνεία των αποτελεσμάτων τέτοιων πειραμάτων που είναι συνήθως μεγάλης κλίμακας.

Το βιολογικό πρόβλημα

Το γονιδίωμα ενός σύνθετου πολυκύτταρου οργανισμού, όπως του ανθρώπου, περιέχει περίπου 22000 γονίδια (Lander et al. 2001). Τα περισσότερα απ' αυτά δεν είναι ενεργοποιημένα σε καθέναν από τους ~200 κυτταρικούς τύπους αλλά, αντίθετα εκφράζονται σε διαφορετικό χρονικά σημεία και με διαφορετική ένταση ανάλογα με το είδος της λειτουργίας που επιτελείται από το κάθε κύτταρο (Lukk et al. 2010). Το γονίδιο της αιμοσφαιρίνης π.χ. που κωδικοποιεί την πρωτεΐνη που

[1] *Στην πραγματικότητα, η ρύθμιση των επιπέδων των πρωτεϊνικών μορίων γίνεται σε πολλά ακόμα επίπεδα εκτός από αυτό της μεταγραφής. Η παραγωγή των πρωτεϊνών εξαρτάται από το ρυθμό με τον οποίο γίνεται η επεξεργασία και η αποικοδόμηση των mRNA, την ταχύτητα μετάφρασής τους από τα ριβοσώματα κ.ά. Η μεταγραφή αποτελεί ένα μόνο στάδιο της διαδικασίας παραγωγής τους. Βλ. και "Μέθοδοι για τη μελέτη της γονιδιακής έκφρασης: Ανάλυση Μεταγραφώματος".*

δεσμεύει το σίδηρο και μεταφέρει το οξυγόνο στα ερυθροκύτταρα δεν εκφράζεται στα κύτταρα του εγκεφάλου ενώ το γονίδιο της ινσουλίνης αντίστοιχα εκφράζεται μόνο στα πανγκρεατικά κύτταρα. Η έκφραση κάποιων γονιδίων μπορεί να είναι επίσης συνάρτηση του χρόνου. Για παράδειγμα, μια σειρά από γονίδια που κωδικοποιούν για συγκεκριμένες κινάσες (ένζυμα των οποίων ο ρόλος είναι η ενεργοποίηση άλλων πρωτεϊνών μέσω προσθήκης ιόντων φωσφορικού οξέος) εκφράζονται σ' όλα τα κύτταρα αλλά μόνο κατά τη διαδικασία της κυτταρικής διαίρεσης, ενώ είναι κατεσταλμένα κατά το μεγαλύτερο μέρος της ζωής του κυττάρου. Ειδικά στους πολυκύτταρους οργανισμούς, η εναλλακτική, ή "διαφορική" (differential) όπως αποκαλείται, έκφραση των γονιδίων αποτελεί την αιτιακή βάση για την αρχική διαφοροποίηση των κυττάρων κατά την ανάπτυξη του εμβρύου και την εξειδίκευση της κυτταρικής λειτουργίας, την ομοιόστασή τους και την απόκρισή τους σε εσωτερικά και εξωτερικά ερεθίσματα. Γνωρίζουμε επίσης ότι σημαντικές διαφορές στο πρόγραμμα έκφρασης γονιδίων συμβαίνουν κατά την εκδήλωση παθολογικών καταστάσεων.

Το βιολογικό πρόβλημα με το οποίο θα ασχοληθούμε σ' αυτό το κεφάλαιο σχετίζεται με την ανάλυση της γονιδιακής έκφρασης, η μελέτη της οποίας είναι πρωταρχικής σημασίας στη σύγχρονη μοριακή βιολογία. Πιο συγκεκριμένα, τα ερωτήματα που θα προσπαθήσουμε να απαντήσουμε είναι:

1. *Με ποιους τρόπους μπορούμε να ποσοτικοποιήσουμε την γονιδιακή έκφραση; Σε ποιο επίπεδο της διαδικασίας παραγωγής πρωτεϊνών είναι προτιμότερο να κάνουμε τις μετρήσεις μας; Ποιες είναι οι πιο κατάλληλες μέθοδοι για να το κάνουμε;*
2. *Πώς προσδιορίζουμε ποια γονίδια είναι ενεργοποιημένα και ποια όχι από το σύνολο των γονιδίων ενός οργανισμού; Πώς ποσοτικοποιούμε τη μεταβολή των επιπέδων έκφρασης ώστε να ορίσουμε γονίδια που είναι "διαφορικά εκφραζόμενα" (differentially expressed) μεταξύ δύο καταστάσεων; Σε ποιο βαθμό μπορούμε να είμαστε σίγουροι ότι ένα γονίδιο είναι ενεργοποιημένο ή κατεσταλμένο;*
3. *Πώς μπορούμε να ορίσουμε ομάδες-υποσύνολα γονιδίων που έχουν κοινά χαρακτηριστικά στα πρότυπα έκφρασής τους; Πώς μπορούμε δηλαδή να εντοπίσουμε γονίδια που ενεργοποιούνται ή/και καταστέλλονται κάτω από τις ίδιες συνθήκες;*

Στη συνέχεια θα εξετάσουμε με τη σειρά τα τρία αυτά βασικά βιολογικά προβλήματα ξεκινώντας από τη μεθοδολογία (1), περνώντας στην αρχική επεξεργασία των δεδομένων για την εξαγωγή των διαφορικά εκφραζόμενων γονιδίων (2), για να καταλήξουμε στο τελευταίο μέρος του κεφαλαίου στις μεθόδους ομαδοποίησης και πώς αυτές χρησιμοποιούνται στην ανάλυση δεδομένων έκφρασης γονιδίων (3).

Μέθοδοι για τη μελέτη της γονιδιακής έκφρασης

Ανάλυση μεταγραφώματος (transcriptome analysis)

Πώς μπορούμε να ποσοτικοποιήσουμε τη γονιδιακή έκφραση; Δεδομένου ότι οι κυτταρικές λειτουργίες επιτελούνται από πρωτεΐνες το πιο λογικό θα ήταν να προσπαθήσουμε να μετρήσουμε τις συγκεντρώσεις όλων των πρωτεϊνών σ' ένα κύτταρο. Τέτοιου είδους προσεγγίσεις υπάρχουν και

κατηγοριοποιούνται κάτω από το γενικό όρο “πρωτεωμική ανάλυση”. Σύγχρονες μεθοδολογίες μας επιτρέπουν να ποσοτικοποιήσουμε 8000-10000 πρωτεΐνες σ' ένα κυτταρικό δείγμα. Παρ' όλ' αυτά, οι πρωτεωμικές αναλύσεις δεν είναι η μέθοδος που επιλέγουμε κατά προτίμηση όταν θέλουμε να αναλύσουμε τη γονιδιακή έκφραση, για μια σειρά από λόγους. Ο πρώτος λόγος έχει να κάνει με τεχνικά ζητήματα. Οι συγκεντρώσεις πρωτεϊνών σ' έναν κυτταρικό πληθυσμό διαφέρουν μεταξύ τους σε τρομακτικό βαθμό που καλύπτει πολλές τάξεις μεγέθους. Κάποιες υπάρχουν σε εκατομμύρια αντίγραφα, ενώ κάποιες άλλες σε μόλις μερικές δεκάδες. Δεύτερον, η ποικιλομορφία στις λειτουργίες που επιτελούν αντανακλάται και στην κυτταρική τους χωροθέτηση αλλά και στις χημικές ιδιότητες. Τα ένζυμα είναι διαλυτά και εύκολα απομονώνονται από το κυτταροδιάλυμα (cytosol), αλλά οι διαμεμβρανικές πρωτεΐνες και οι μεταφορείς βρίσκονται εγκολπωμένες στις λιπιδικές μεμβράνες. Σε γενικές γραμμές, δεν διαθέτουμε μια ιδανική πειραματική διάταξη που να διαθέτει διακριτική ικανότητα τέτοια, που να επιτρέπει την ταυτοποίηση και ποσοτικοποίηση πρωτεϊνών σε ίχνη, αλλά και να μπορεί να παίρνει ικανοποιητικό δείγμα από όλα τα σημεία του κυττάρου.

Ξεπερνώντας τα τεχνικά θέματα, θα πρέπει να σκεφτούμε ότι η παραγωγή των πρωτεϊνών είναι μια πολύπλοκη διαδικασία που περιλαμβάνει πολλά επιμέρους βήματα, ξεκινώντας από την ενεργοποίηση της μεταγραφής των γονιδίων τους, περνώντας στην επεξεργασία του mRNA και τη μετάφρασή του από τα ριβοσώματα, στις μετα-μεταφραστικές τροποποιήσεις και την τελική διαδικασία της αναδίπλωσης των αμινοξικών αλυσίδων σε πλήρως λειτουργική πρωτεΐνη. Από τα παραπάνω βήματα, το πρώτο χρονικά, η μεταγραφή των mRNA είναι αυτό που συγκεντρώνει και τα περισσότερα πλεονεκτήματα για αξιόπιστες και αποτελεσματικές μετρήσεις. Το mRNA απομονώνεται εύκολα και σε ποσότητες ικανές για αποτελεσματική ποσοτικοποίηση, οι μεθοδολογίες που υπάρχουν μπορούν να ενισχύσουν το δείγμα και να ταυτοποιήσουν έτσι ακόμα και μόρια mRNA που βρίσκονται σε πολύ μικρές συγκεντρώσεις, ενώ, επιπλέον, απ' όσο γνωρίζουμε η μεταγραφή είναι το στάδιο στη διαδικασία της έκφρασης γονιδίων, που υπόκειται στην πιο αυστηρή ρύθμιση. Με βάση τα παραπάνω, οι μετρήσεις σε επίπεδο mRNA, εξασφαλίζουν ομοιογένεια του δείγματος, είναι πιο εύκολες πειραματικά και μπορούν να είναι ποσοτικά ακριβείς. Επιπλέον, τα επίπεδα του mRNA αντανακλούν πιο άμεσα τις αλλαγές που επισυμβαίνουν στην έκφραση των γονιδίων, καθώς ο χρόνος που μεσολαβεί από τη στιγμή που ένα γονίδιο ενεργοποιείται ως την παραγωγή της πρωτεΐνης που κωδικοποιεί μπορεί συχνά να είναι απαγορευτικός για την μελέτη ταχείας απόκρισης σε ερεθίσματα.

Τέλος, μια σειρά από μελέτες έχουν δείξει ότι σε σταθερές συνθήκες (steady state) τα επίπεδα των mRNA συσχετίζονται σε μεγάλο βαθμό με τα αντίστοιχα των πρωτεϊνών (Vogel and Marcotte 2012) και παρά το γεγονός ότι ο βαθμός συσχέτισης είναι μειωμένος για τα mRNA που βρίσκονται σε χαμηλή συγκέντρωση (Maier, Güell, and Serrano 2009), η ποσοτικοποίηση των mRNA θεωρείται ο πιο διαδεδομένος, άμεσος και συστηματικός τρόπος για τη μελέτη της γονιδιακής έκφρασης. Στη συνέχεια αυτής της ενότητας θα δούμε τις βασικές μεθοδολογίες για την πειραματική μελέτη των επιπέδων mRNA. Καθώς το αντικείμενο του κεφαλαίου (αλλά και του βιβλίου γενικότερα) είναι η υπολογιστική ανάλυση, η συζήτηση δε θα επικεντρωθεί στις λεπτομέρειες των πειραματικών διατάξεων και των βιοχημικών διαδικασιών αλλά στο είδος των δεδομένων που παράγουν οι δύο κυριότερες από αυτές τις μεθοδολογίες, οι μικροσυστοιχίες DNA

και η αλληλούχιση RNA.

Μικροσυστοιχίες DNA (DNA microarrays)

Οι μικροσυστοιχίες DNA ή DNA microarrays (Duggan et al. 1999) ήταν μέχρι πρότινος η πιο διαδεδομένη μέθοδος ανάλυσης της γονιδιακής έκφρασης. Η ευκολία στο χειρισμό, η μεγάλη διάχυση της τεχνολογίας και το σχετικά χαμηλό κόστος της είναι οι κύριοι λόγοι για τη μεγάλη τους εξάπλωση. Πάρα την ολοένα αυξανόμενη χρήση μεθόδων αλληλούχισης νέας γενιάς (βλ. Αλληλούχιση RNA παρακάτω), η μεγάλη πλειοψηφία των πειραμάτων έκφρασης που είναι διαθέσιμα έχουν διενεργηθεί με τη χρήση μικροσυστοιχειών DNA. Η αρχή της μεθόδου στηρίζεται στο συνδυασμό της φυσικής ιδιότητας της υβριδοποίησης του DNA και στην πρόοδο της νανοτεχνόλογιας που επιτρέπει την ακινητοποίηση ενός μεγάλου αριθμού μορίων σε μικρο-πλακίδια με εξαιρετικά μεγάλη ακρίβεια. Μια μικροσυστοιχία DNA περιέχει έτσι έναν πολύ μεγάλο αριθμό τμημάτων μονόκλωνου γονιδιωματικού DNA που έχουν ακινητοποιηθεί σε ένα στερεό υπόστρωμα. Τα τμήματα αυτά έχουν τη φυσική τάση να ζευγαρώνουν με τα συμπληρωματικά τους (κατά Watson-Crick) μονόκλωνα μόρια εφόσον αυτά εντοπίζονται σ' ένα διάλυμα που διέρχεται πάνω από το υπόστρωμα (πλακίδιο ή chip), μέσω της ιδιότητας που ονομάζουμε υβριδοποίηση.

Σε ένα πείραμα γονιδιακής έκφρασης σε μικροσυστοιχία DNA, το πλακίδιο σχεδιάζεται με τέτοιο τρόπο ώστε να περιέχει χαρακτηριστικά τμήματα DNA απ' όσο το δυνατόν περισσότερα γονίδια του υπό μελέτη οργανισμού. Τα τμήματα αυτά ονομάζονται ανιχνευτές (probes) και μπορούν έτσι να υβριδοποιηθούν με συμπληρωματικά τμήματα cDNA των αντίστοιχων γονιδίων[2]. Τόσο η (ποιοτική) ανίχνευση όσο και η ποσοτικοποίηση του mRNA γίνεται μέσω της μέτρησης φθορισμού που εκπέμπεται από τη στιγμή που τα δύο συμπληρωματικά μόρια υβριδοποιηθούν. Ο φθορισμός προκύπτει καθώς το δείγμα, πριν περάσει πάνω από το πλακίδιο, έχει σημανθεί με μια συγκεκριμένη φθορίζουσα χρωστική, η οποία ενεργοποιείται και εκπέμπει μόνο σε δίκλωνη διαμόρφωση. Ένα πείραμα έκφρασης σε μικροσυστοιχία συνολικά περιλαμβάνει τα εξής στάδια:

- Απομόνωση του mRNA από το δείγμα.
- Δημιουργία συμπληρωματικού DNA (cDNA) μέσω αντίστροφης μεταγραφής.
- Σήμανση του cDNA με μια φθορίζουσα χρωστική ουσία.
- Υβριδοποίηση του cDNA στη μικροσυστοιχία και μέτρηση του φθορισμού.

Χωρίς να μπούμε σε τεχνικές λεπτομέρειες, αρκεί να αναφέρουμε ότι το τελικό αποτέλεσμα ενός πειράματος σε μικροσυστοιχία DNA συνίσταται από μια (μακριά) σειρά από μετρήσεις φθορισμού που αντιστοιχούν στη σχετική αφθονία mRNA μορίων στο δείγμα μας και που ταυτοποιούνται με βάση τη συμπληρωματικότητά τους για συγκεκριμένους ανιχνευτές. Το σύνολο των τιμών που καταγράφονται εξαρτάται από το είδος του οργανισμού που μελετάται, τον αριθμό των γονιδίων, των οποίων την έκφραση επιθυμούμε να ποσοτικοποιήσουμε και τη διακριτική

[2] *Τα περισσότερα chip του εμπορίου περιέχουν πλέον όχι μόνο ανιχνευτές που αντιστοιχούν στο σύνολο των γνωστών γονιδίων αλλά σε πολλές περιπτώσεις συμπεριλαμβάνουν στοιχεία που αντιστοιχούν σε διαφορετικά μετάγραφα, χρήση εναλλακτικών σημείων συρραφής (alternative splice sites) κλπ (Kapranov et al. 2005).*

ικανότητα της μικροσυστοιχίας. Σε κάθε περίπτωση, πρόκειται για πειράματα μεγάλης κλίμακας, που μεταφράζεται σε μερικές χιλιάδες ή δεκάδες χιλιάδες τιμές. Σχηματικά ένα μέρος μόνο από τα αποτελέσματα ενός τέτοιου πειράματος φαίνεται στην Εικόνα 7.1. Σε επόμενη ενότητα θα δούμε πώς χειριζόμαστε αυτό το είδος των δεδομένων πριν περάσουμε στην κατά κύριο λόγο ανάλυση της διαφορικής έκφρασης.

ID_REF	GSM183695	GSM185526	GSM185527	GSM185528	GSM185529	GSM185530	GSM185531
1000_at	1569.51	1585.62	1099.23	1527.75	1013.3	1341.91	2235.19
1001_at	55.4826	37.9262	20.7475	35.6907	9.18595	35.4699	20.4733
1002_f_at	10.7225	7.08931	6.55284	4.34082	7.502	10.8898	5.8394
1003_s_at	42.8653	18.7231	19.788	23.6005	24.8676	27.5205	30.4685
1004_at	82.4252	72.2625	63.43	71.3506	110.458	129.447	62.5745
1005_at	3927.36	1561.68	2143.34	1368.22	652.855	1126.38	1891.47
1006_at	22.3963	8.03122	20.5788	3.55786	1.25394	66.1442	2.03623
1007_s_at	976.181	1018.13	842.372	483.802	455.1	1094.53	551.697
1008_f_at	3328.22	2417.84	1404.77	1571.02	1838.4	2340.35	2206.38
1009_at	3412.83	4165.01	2486.12	3378.94	2875.03	3835.5	3408.27
100_g_at	458.13	659.593	414.027	339.647	429.243	619.421	573.235
1010_at	51.471	17.9678	9.93612	24.4365	26.0201	9.68313	7.18712
1011_s_at	1358.13	1050.57	848.434	840.406	811.129	965.555	1196.79
1012_at	92.6114	56.6347	57.6028	49.9186	31.0457	54.3793	80.8679

Εικόνα 7.1: *Οι πρώτες γραμμές του αποτελέσματος ενός πειράματος έκφρασης σε μικροσυστοιχία DNA. Η πρώτη στήλη περιέχει τον κωδικό αριθμό του ανιχνευτή (probe) που μπορεί να αντιστοιχηθεί σε ένα συγκεκριμένο γονίδιο. Οι τιμές που ακολουθούν στις στήλες 2-8 αντιστοιχούν στη μέτρηση φθορισμού για το δεδομένο ανιχνευτή για καθένα από επτά διαφορετικά δείγματα.*

Αλληλούχιση RNA (RNA Sequencing)

Η υψηλή ζήτηση για αλληλούχιση χαμηλού κόστους έχει οδηγήσει την τελευταία δεκαετία στην ανάπτυξη της αλληλούχισης υψηλής απόδοσης (ή αλληλούχισης επόμενης γενιάς, next generation sequencing, NGS). Πρόκειται για τη βιολογική τεχνολογία της “παραλληλοποίησης” της διαδικασίας αλληλούχισης, με τρόπο που να καθιστά δυνατό τον προσδιορισμό της πρωτοταγούς διαδοχής βάσεων εκατομμυρίων ή δεκάδων εκατομμυρίων αλληλουχιών ταυτόχρονα. Η επεκτασιμότητα, η ταχύτητα αλλά κυρίως η σχέση κόστους-απόδοσης των NGS εφαρμογών επιτρέπουν στους ερευνητές να μελετήσουν τα βιολογικά συστήματα σε επίπεδο που δεν ήταν δυνατό μέχρι πρότινος. Περισσότερα για την τεχνολογία και τις σύγχρονες μεθόδους αλληλούχισης επόμενης γενιάς (NGS) θα συζητήσουμε στο Κεφάλαιο 11, που θέμα του έχει τη βιολογία στην εποχή των μεγάλων συνόλων δεδομένων. Για την ώρα αρκεί να περιγράψουμε περιληπτικά τη διαδικασία αποκομιδής των δεδομένων για ένα πείραμα έκφρασης γονιδίων με τη χρήση NGS και -κυριότερα- τη μορφή που αυτά έχουν.

Η NGS ποσοτικοποίηση της γονιδιακής έκφρασης γίνεται με τη μαζική αλληλούχιση mRNA

που αρχικά απομονώνεται από το δείγμα και στη συνέχεια μετατρέπεται σε cDNA όπως στην περίπτωση των μικροσυστοιχιών. Οι διαφορές της μεθόδου αλληλούχισης RNA, ή RNASeq όπως είναι ευρύτερα γνωστή, ξεκινούν εδώ: Αρχικά το cDNA δείγμα υπόκειται σε ένα στάδιο κλασμάτωσης πριν την αλληλούχιση καθώς οι υπάρχουσες τεχνολογίες, στη συντριπτική τους πλειοψηφία, αποδίδουν αξιόπιστα αποτελέσματα για αλληλουχίες όχι μεγαλύτερες από 300-500 βάσεις. Στη συνέχεια και ανάλογα με την τεχνολογία που εφαρμόζεται, το cDNA ενισχύεται μέσω αλυσιδωτής αντίδρασης πολυμεράσης (PCR) (η οποία μπορεί να διενεργηθεί με διαφορετικούς τρόπους) και αλληλουχείται μαζικά συνήθως "μέσω σύνθεσης". Αυτό σημαίνει ότι νέοι κλώνοι cDNA συντίθενται πάνω στο εκμαγείο των κλώνων του δείγματος και η διαδικασία της σύνθεσης καταγράφεται νουκλεοτίδιο-νουκλεοτίδιο (Wang, Gerstein, and Snyder 2009). Το αποτέλεσμα είναι ένα αρχείο που περιέχει έναν πολύ μεγάλο αριθμό (της τάξης των δεκάδων εκατομμυρίων) αλληλουχιών μικρού μήκους (μεταξύ 100 και 500 βάσεων). Μια ακόμα βασική διαφορά με τις μεθοδολογίες που βασίζονται στην υβριδοποίηση έχει να κάνει με την ποσοτικοποίηση των αποτελεσμάτων. Αυτή γίνεται για ένα πείραμα αλληλούχισης επόμενης γενιάς μέσω των εξής βημάτων (Mortazavi et al. 2008):

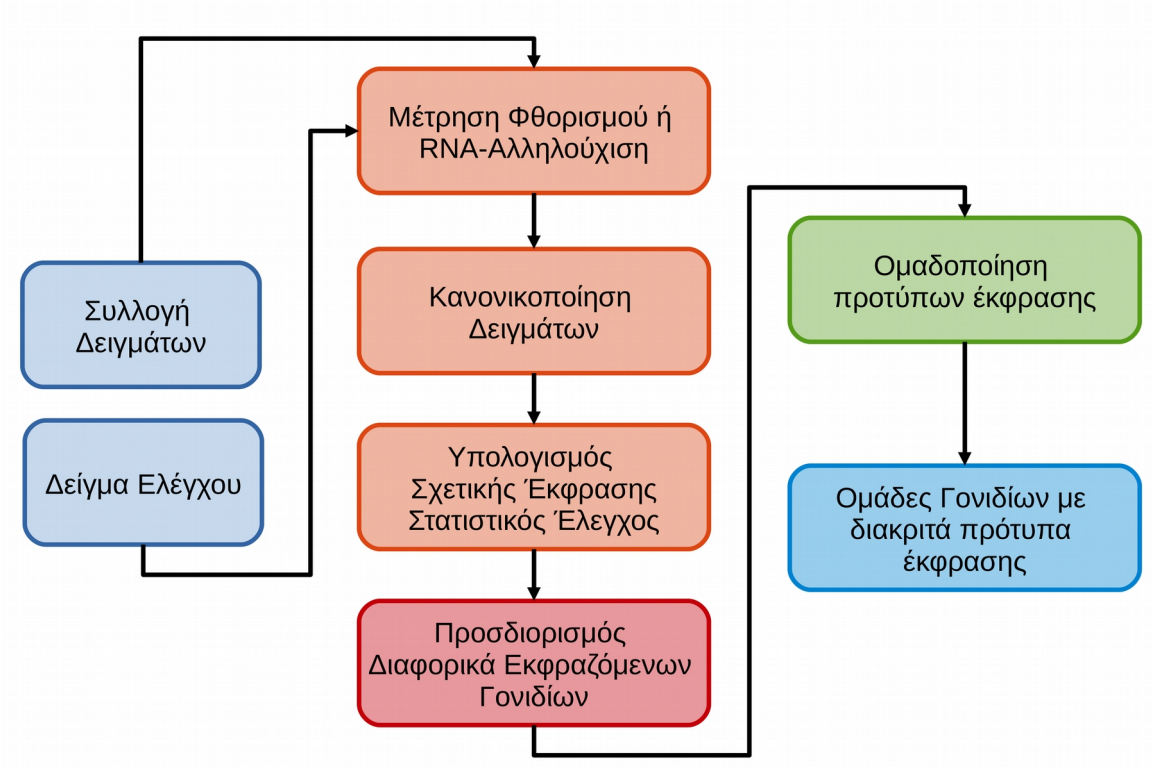
1. Έλεγχος ποιότητας των αλληλουχιών και αποκλεισμός αυτών που δεν ικανοποιούν συγκεκριμένα κριτήρια αξιοπιστίας. Στο πρώτο αυτό στάδιο απορρίπτονται οι αλληλουχίες που δεν πληρούν τις προϋποθέσεις ποιότητας που είναι γενικώς αποδεκτές. Ποσοστά απόρριψης μεταξύ 5 και 30% είναι φυσιολογικά, ανάλογα με το πείραμα.
2. Χαρτογράφηση των αλληλουχιών στο γονιδίωμα αναφοράς. Πρόκειται για τη βασική διαδικασία μέσω της οποίας ποσοτικοποιούνται τα αποτελέσματα. Για καθεμία από τα εκατομμύρια των μικρών αλληλουχιών (που ονομάζονται και "αναγνώσεις αλληλουχιών", sequence reads ή απλά reads) εντοπίζεται η θέση του γονιδιώματος από την οποία προέρχεται (το οποίο ονομάζουμε γονιδίωμα αναφοράς, reference genome). Μετά τη χαρτογράφηση του συνόλου των reads, μπορούμε να γνωρίζουμε με ακρίβεια πόσες φορές "διαβάστηκε" κάθε νουκλεοτίδιο του υπο μελέτη γονιδιώματος[3].
3. Ποσοτικοποίηση αριθμού αναγνώσεων ανά μετάγραφο. Παίρνοντας ως δεδομένες τις θέσεις των γονιδίων/μεταγράφων στο γονιδίωμα αναφοράς μπορούμε να υπολογίσουμε τον αριθμό των αναγνώσεων που επικαλύπτονται με κάθε γενετικό τόπο ξεχωριστά. Λαμβάνοντας υπόψη το μήκος της αντίστοιχης γονιδιωματικής περιοχής, (ή για την ακρίβεια των μεταγράφων mRNA που προκύπτουν από αυτήν), αλλά και το συνολικό αριθμό των αναγνώσεων που προέκυψαν από το πείραμα μπορούμε να καταλήξουμε σε μια αριθμητική τιμή που είναι έτσι δηλωτική της ποσότητας mRNA που υπήρχε στο αρχικό δείγμα από το συγκεκριμένο γενετικό τόπο.

***Ερώτηση**: Με ποιον τρόπο περιμένετε να επηρεάζει τη μέτρηση σε ένα RNASeq πείραμα, το μήκος του υπό εξέταση γονιδίου;*

[3] *Τη διαδικασία της χαρτογράφησης θα συζητήσουμε αναλυτικά στο Κεφάλαιο 11.*

Στάδια ανάλυσης ενός πειράματος γονιδιακής έκφρασης

Ανεξάρτητα από τη μέθοδο που χρησιμοποιούμε για την μελέτη της γονιδιακής έκφρασης, τα στάδια της ανάλυσης είναι λίγο πολύ τα ίδια και περιλαμβάνουν: α) την απομόνωση του mRNA β) την ποσοτικοποίηση του γ) τον προσδιορισμό των διαφορικά εκφραζόμενων γονιδίων και δ) την ομαδοποίηση γονιδίων ανάλογα με τα πρότυπα έκφρασής τους. Σχηματικά αυτή η διαδικασία φαίνεται στην Εικόνα 7.2.



***Εικόνα 7.2**: Σχηματική αναπαράσταση των σταδίων ενός πειράματος γονιδιακής έκφρασης από την αποκομιδή των πρωτογενών δεδομένων ως τη δημιουργία ομάδων γονιδίων με χαρακτηριστικά πρότυπα έκφρασης.*

Αποκομιδή και χειρισμός πρωτογενών δεδομένων

Μέτρησεις

Όπως είδαμε παραπάνω τα πρωτογενή δεδομένα εξαρτώνται από τη μεθοδολογία που χρησιμοποιήθηκε για την ανάλυση. Στην περίπτωση των μικροσυστοιχιών το μετρούμενο μεγέθος είναι η ένταση φθορισμού που προκύπτει από την υβριδοποίηση συμπληρωματικών στους ανιχνευτές αλληλουχιών. Στην περίπτωση του RNASeq οι μετρήσεις αφορούν καθαρά το πλήθος

των συντιθέμενων αλληλουχιών που προέρχονται από ένα συγκεκριμένο mRNA μόριο. Και στις δύο περιπτώσεις είναι απαραίτητη μια σειρά χειρισμών των δεδομένων έτσι όπως λαμβάνονται από την πειραματική διάταξη και είναι βασικό να προηγηθεί μια διαδικασία κανονικοποίησης των πρωτογενών δεδομένων πριν περάσουμε στην περαιτέρω ανάλυσή τους. Αξίζει να θυμηθούμε εδώ ότι η κανονικοποίηση είναι η διαδικασία με την οποία μετατρέπουμε δεδομένα που έχουν προκύψει με διαφορετικούς τρόπους σε μια κλίμακα που να τα καθιστά άμεσα συγκρίσιμα. Μέσω της κανονικοποίησης, αφαιρούμε συστηματικά σφάλματα που μπορεί να προέρχονται από τους χειρισμούς των πειραματιστών, την πειραματική διάταξη ή άλλους αστάθμητους παράγοντες. Μ' αυτόν τον τρόπο δεδομένα από διαφορετικά πειράματα μπορούν να συγκριθούν στη βάση των διαφορών τους που αφορούν το βιολογικό υπόβαθρο, ελαχιστοποιώντας την επίδραση τεχνικών σφαλμάτων και εξωγενών παραγόντων.

Μικροσυστοιχίες DNA και λογαρίθμιση έντασης φθορισμού

Οι μετρήσεις φθορισμού που προκύπτουν από ένα πείραμα μικροσυστοιχιών διατηρούν κάποια χαρακτηριστικά που δυσχαιρένουν την περαιτέρω ανάλυση. Πιο συγκεκριμένα, οι ανεπεξέργαστες τιμές φθορισμού εμφανίζουν μεγάλη διασπορά σε ό,τι αφορά το εύρος της έντασης, με πολλές τιμές να είναι εξαιρετικά μικρές και λίγες να είναι πολύ μεγάλες (βλ. Εικόνα 7.3α). Κάτι τέτοιο είναι προβληματικό καθώς ιδανικά, και προκειμένου τα πειράματα να είναι συγκρίσιμα, θα πρέπει η διασπορά των τιμών να είναι ανεξάρτητη της έντασης φθορισμού. Οι τιμές δηλαδή θα πρέπει να κατανέμονται με όσο το δυνατόν πιο ομοιόμορφο τρόπο. Προκειμένου να μειωθεί η επίδραση αυτής της τάσης για διασπορά που σχετίζεται με την ένταση, οι αρχικές τιμές ενός πειράματος μικροσυστοιχιών μετασχηματίζονται με τη λήψη λογαρίθμου. Η λογαρίθμηση των τιμών φθορισμού οδηγεί σε μια προσεγγιστικά κανονική κατανομή των τιμών που κάνει τον περαιτέρω χειρισμό τους πιο εύκολο. Αυτό συμβαίνει καθώς (επίσης προσεγγιστικά) η αρχική κατανομή των τιμών φθορισμού ακολουθεί τη λογαριθμοκανονική κατανομή (log-normal distribution). Η βάση που επιλέγουμε για το λογάριθμο είναι τις περισσότερες φορές το 2 για λόγους που έχουν να κάνουν με την εύκολη ερμηνεία των αποδιδόμενων τιμών σε περιπτώσεις σύγκρισης δύο δειγμάτων, όπως θα δούμε πιο αναλυτικά παρακάτω (βλ. Υπολογισμός βαθμού διαφοράς έκφρασης).

RNA Sequencing και αριθμός αναγνώσεων ανά 1000 βάσεις (RPKM)

Όπως συζητήσαμε και νωρίτερα, στα πειράματα RNASeq η ποσοτικοποίηση της έκφρασης γίνεται στη βάση του αριθμού των reads που βρίσκονται να επικαλύπτονται με μια συγκεκριμένη γονιδιωματική περιοχή, που αντιστοιχεί σ' ένα δεδομένο μετάγραφο (transcript) ή γονίδιο. Ωστόσο, η απόλυτη τιμή αυτής της ποσότητας δεν μπορεί να χρησιμοποιηθεί αυτούσια για δύο βασικούς λόγους. Ο πρώτος είναι ότι εξαρτάται από το μήκος του μεταγράφου. Μεγάλα γονίδια που εκτείνονται για πολλές χιλιάδες βάσεις (κάποιες φορές και εκατοντάδες χιλιάδες) θα συγκεντρώνουν μεγαλύτερο αριθμό αναγνώσεων απλώς και μόνο λόγω μεγέθους. Ο δεύτερος

λόγος είναι ότι πειράματα που συνολικά παράγουν μεγαλύτερο αριθμό αναγνώσεων, επειδή το δείγμα εμπλουτίστηκε περισσότερο, επειδή η ποιότητα της αλληλούχισης ήταν καλύτερη και οδήγησε στην απόρριψη μικρότερου αριθμού αναγνώσεων ή απλώς επειδή το δείγμα αλληλουχήθηκε σε μεγαλύτερο "βάθος", θα δίνουν συστηματικά μεγάλυτερες τιμές αναγνώσεων ανά μετάγραφο.

Για τους δύο αυτούς λόγους, για την εκτίμηση του βαθμού έκφρασης μεταγράφων υπολογίζουμε μια διορθωμένη τιμή που ονομάζεται RPKM ή FPKM, από τα αρχικά "Reads/Fragments per Kilobase of gene per Million". Η τιμή αυτή αντιστοιχεί σε μια διπλή κανονικοποίηση του αριθμού των αναγνώσεων ανά μετάγραφο, αρχικά ως προς το μήκος του (ανά χιλιάδα βάσεων) κι έπειτα ως προς το σύνολο των παραχθεισών αλληλουχιών (ανά εκατομμύριο αλληλουχιών)[4]. Αν λοιπόν το μετάγραφο *t* με μήκος *l* επικαλύπτεται με *r* αναγνώσεις σε σύνολο *N* χαρτογραφημένων αλληλουχιών, τότε η τιμή RPKM θα είναι ίση με:

$$RPKM(t) = \frac{r}{\frac{l}{10^3}\frac{N}{10^6}} = \frac{r\,10^9}{lN} \qquad 7.1$$

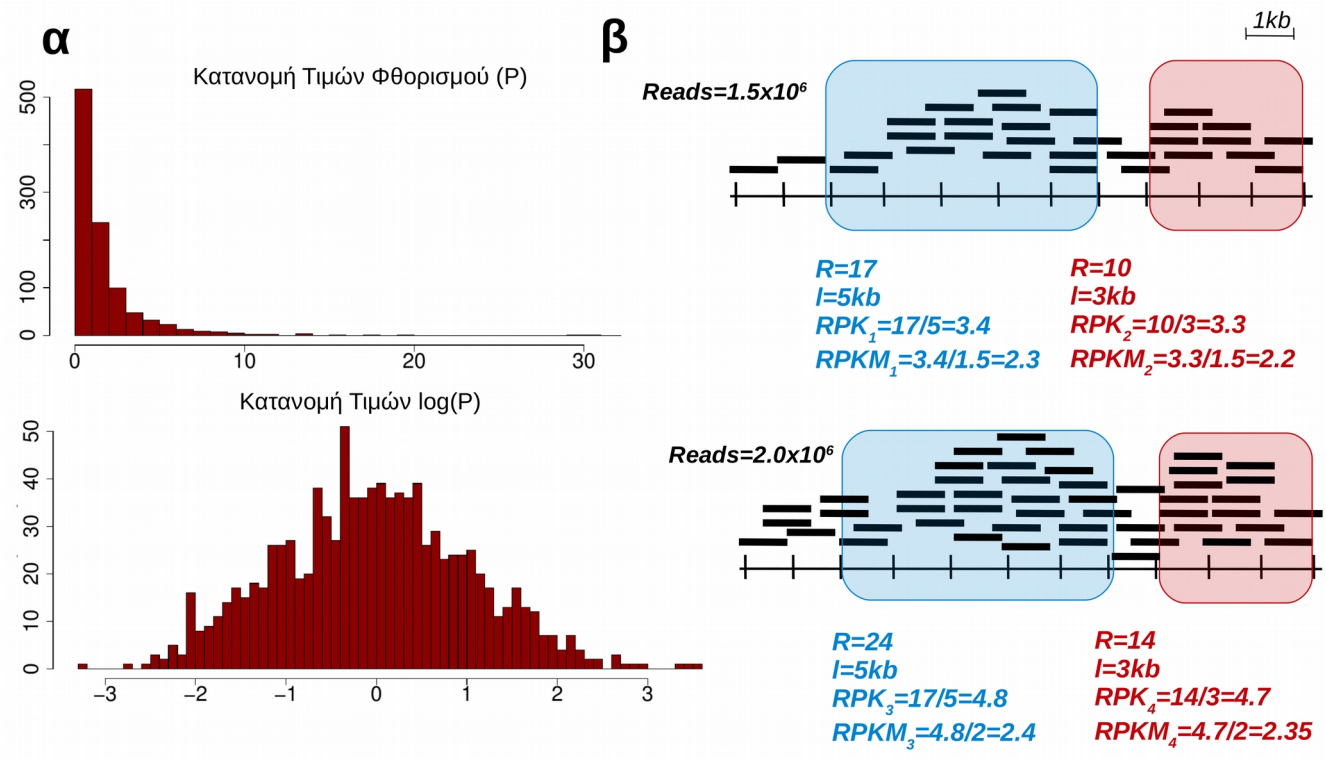
Η παραπάνω μετατροπή εξασφαλίζει ότι η τιμή στην οποία αναφέρεται ο βαθμός έκφρασης ενός μεταγράφου θα είναι ανεξάρτητη τόσο του μήκους του όσο και του βάθους της αλληλούχισης.

***Ερώτηση**: Μπορείτε να σκεφτείτε άλλα χαρακτηριστικά των γονιδιωματικών αλληλουχιών που να επιφέρουν ανωμαλίες στις μετρήσεις RNASeq και να προτείνετε τρόπους κανονικοποίησής τους;*

Μαθηματικό Ιντερμέδιο Ι. Κανονικοποίηση

Στις αμέσως προηγούμενες ενότητες συζητήσαμε τους πιο βασικούς τρόπους για μια πρώτη επεξεργασία των δεδομένων, η οποία αφορά μια εσωτερική κανονικοποίηση. Τι συμβαίνει όμως όταν θέλουμε να συγκρίνουμε μεταξύ διαφορετικών πειραμάτων, μεταξύ επαναλήψεων του ίδιου πειράματος, ή του ίδιου πειράματος κάτω από ελαφρώς διαφορετικές συνθήκες; Σ' όλες τις παραπάνω περίπτωσεις είναι απαραίτητο να καταφύγουμε σε πιο εκλεπτυσμένες μεθόδους για την κανονικοποίηση των τιμών. Ανάμεσα στις διάφορες μεθόδους κανονικοποίησης δεδομένων γονιδιακής έκφρασης (Bolstad et al. 2003), θα αναφερθούμε στη συνέχεια σ' αυτές με το μεγαλύτερο ενδιαφέρον τόσο από θεωρητικής όσο και από πρακτικής άποψης.

[4] *Σημειώνεται εδώ ότι η τιμή συνολικών αναγνώσεων με την οποία διαιρούμε στην RPKM κανονικοποίηση αντιστοιχεί στις αναγνώσεις εκείνες που χαρτογραφήθηκαν τελικά στο γονιδίωμα αναφοράς και όχι στο αρχικό σύνολο που περιλαμβάνει αναπόφευκτα και τις αναγνώσεις χαμηλής ποιότητας.*



Εικόνα 7.3: *α) Ιστόγραμμα τιμών φθορισμού που ακολουθούν λογαριθμοκανονική κατανομή (επάνω) και μετά από λήψη των λογαρίθμων τους (κάτω) που μετατρέπει την κατανομή τους σε κανονική β) Γραφική αναπαράσταση του υπολογισμού της τιμής RPKM από δύο πειράματα αλληλούχισης RNA. Οι δύο χρωματισμένες περιοχές περιέχουν διαφορετικό αριθμό μικρο-αναγνώσεων (reads) όμως αυτό είναι αποτέλεσμα του διαφορετικού τους μήκους (5kb έναντι 3kb). Διαίρεση με το μήκος (RPK) δίνει παραπλήσιες τιμές για τις δύο περιοχές στο ίδιο πείραμα. Μεταξύ δύο πειραμάτων με διαφορετικό συνολικό αριθμό αναγνώσεων χρειάζεται μια ακόμα διόρθωση ως προς το συνολικό αριθμό των reads. Ετσι οι τιμές RPKM είναι πολύ παρόμοιες για τις δύο περιοχές και μεταξύ των δύο πειραμάτων.*

Ολική κανονικοποίηση γύρω από σταθερή τιμή

Αυτή μπορεί να γίνει με διάφορες μεθόδους, το κοινό χαρακτηριστικό των οποίων είναι ότι εφαρμόζονται συνολικά στα δεδομένα με τον ίδιο τρόπο. Μια μέθοδος είναι η διαίρεση όλων των τιμών με τις αντίστοιχες χαρακτηριστικές τιμές έκφρασης συστατικών (housekeeping) γονιδίων, η οποία θεωρείται σταθερή ανεξάρτητα από τις συνθήκες ή τον κυτταρικό τύπο. Για την όσο το δυνατόν καλύτερη εφαρμογή της μεθόδου, ορίζεται αρχικά ένα σύνολο από συστατικά γονίδια και η τιμή "διόρθωσης" αποτελεί έναν σταθμισμένο μέσο της έκφρασής τους. Έτσι αν *E(g)* είναι η αρχική τιμή για το γονίδιο *g*, τότε η κανονικοποιημένη τιμή του θα είναι:

$$E'(g)=kE(g) \quad 7.2$$

όπου *k* μια σταθερά για το δεδομένο πείραμα που θα αντιστοιχεί στη συνδυασμένη έκφραση ενός συνόλου συστατικών γονιδίων.

Με απολύτως ανάλογο τρόπο μπορούμε να σκεφτούμε τρόπους κανονικοποίησης με

εναλλακτικές σταθερές *k*. Στη θέση της τιμής *k* μπορούμε έτσι να χρησιμοποιήσουμε τη μέση τιμή των τιμών έκφρασης ή αντίστοιχα τη μέγιστη ή ελάχιστη τιμή του κάθε δείγματος. Όλες οι παραπάνω μέθοδοι είναι ουσιαστικά μέθοδοι που μεταβάλλουν την κλίμακα των τιμών μέσω ενός παράγοντα επανακλιμάκωσης (rescaling factor) που είναι η τιμή *k*.

Τυποποίηση με z-score (z-standardization)

Συναντήσαμε αυτόν τον τύπο κανονικοποίησης στο Κεφάλαιο 1. Πρόκειται για ένα είδος σταθμισμένης επανακλιμάκωσης, όπου εκτός από την μέση τιμή λαμβάνεται υπόψη και η τυπική απόκλιση του δείγματος. Έτσι αν *E(g)* είναι η τιμή έκφρασης του γονιδίου *g* από ένα δείγμα με μέση τιμή *μ* και τυπική απόκλιση *σ*, τότε η κανονικοποιημένη τιμή έκφρασής του *g* θα είναι:

$$Z(g)=\frac{E(g)-\mu}{\sigma} \quad 7.3$$

Η εξίσωση 7.3 εκφράζει ουσιαστικά την απόσταση της τιμής έκφρασης *E(g)* από τη μέση τιμή του δείγματος σε μονάδες τυπικής απόκλισης. Εφαρμογή αυτού του μετασχηματισμού σε διαφορετικά πειράματα τα κάνει συγκρίσιμα όχι μόνο σε ό,τι αφορά τη μέση τιμή τους (που είναι τώρα το 0) αλλά και σε ό,τι αφορά τη διασπορά τους. Ωστόσο, η τυποποίηση είναι σωστό να εφαρμόζεται μόνο σε κατανομές που είναι κανονικές ή προσεγγιστικά κανονικές κι έτσι π.χ. σε ένα πείραμα μικροσυστοιχιών θα πρέπει πάντα να έχει προηγηθεί λογαριθμική τροποποίηση (Irizarry et al. 2003). Σε πειράματα RNASeq αντίστοιχα, καθώς η κανονικότητα δεν εξασφαλίζεται θα πρέπει να σκεφτούμε διαφορετικούς τρόπους που είναι μη-παραμετρικοί.

Κανονικοποίηση ποσοστημορίων (quantile normalization)

Μια μη-παραμετρική μέθοδος κανονικοποίησης που βρίσκει ευρεία εφαρμογή στην ανάλυση δεδομένων γονιδιακής έκφρασης, είναι η κανονικοποίηση ποσοστημορίων (quantile normalization) (Hansen, Irizarry, and Wu 2012). Η συγκεκριμένη μέθοδος οδηγεί σε συγκρίσιμες κατανομές τιμών ακόμα και στην περίπτωση που οι αρχικές κατανομές δεν είναι κανονικές. Η διαδικασία βασίζεται ουσιαστικά στη σύγκριση της κατάταξης των τιμών και για το λόγο αυτό είναι ανεξάρτητη των ροπών (μέση τιμή, διασπορά κλπ). Δεδομένου ενός πίνακα *N* τιμών από *M* διαφορετικά πειράματα/δείγματα, υπολογίζεται αρχικά η κατάταξη (κατά αύξουσα σειρά) των τιμών εντός του κάθε δείγματος σε έναν νέο πίνακα *R[N,M]*. Στη συνέχεια δημιουργείται ένα μοναδικό διάνυσμα μέσων τιμών που λαμβάνουν υπόψη τους την κατάταξη. Έτσι η τιμή Q*[1]* είναι η μέση τιμή των πρώτων στην κατάταξη (μικρότερων) τιμών των *M* δειγμάτων, η τιμή *Q[2]* είναι η μέση τιμή των δεύτερων στην κατάταξη τιμών κ.ο.κ. Οι τιμές *Q* χρησιμοποιούνται στη συνέχεια, στη θέση των τιμών με την αντίστοιχη κατάταξη στον πίνακα *R*. Έτσι η χαμηλότερη τιμή για το πρώτο δείγμα εξισώνεται με την *Q[1]*, το ίδιο και η χαμηλότερη τιμή για το δεύτερο, το τρίτο κ.ο.κ. Το τελικό αποτέλεσμα είναι ένας πίνακας που περιέχει τις ίδιες ακριβώς τιμές σε κάθε στήλη με διαφορετική

ωστόσο κατάταξη. Αυτός ο μετασχηματισμός, εξασφαλίζει ότι οι κατανομές είναι απολύτως συγκρίσιμες χωρίς όμως να χάνεται η εσωτερική τους δομή.

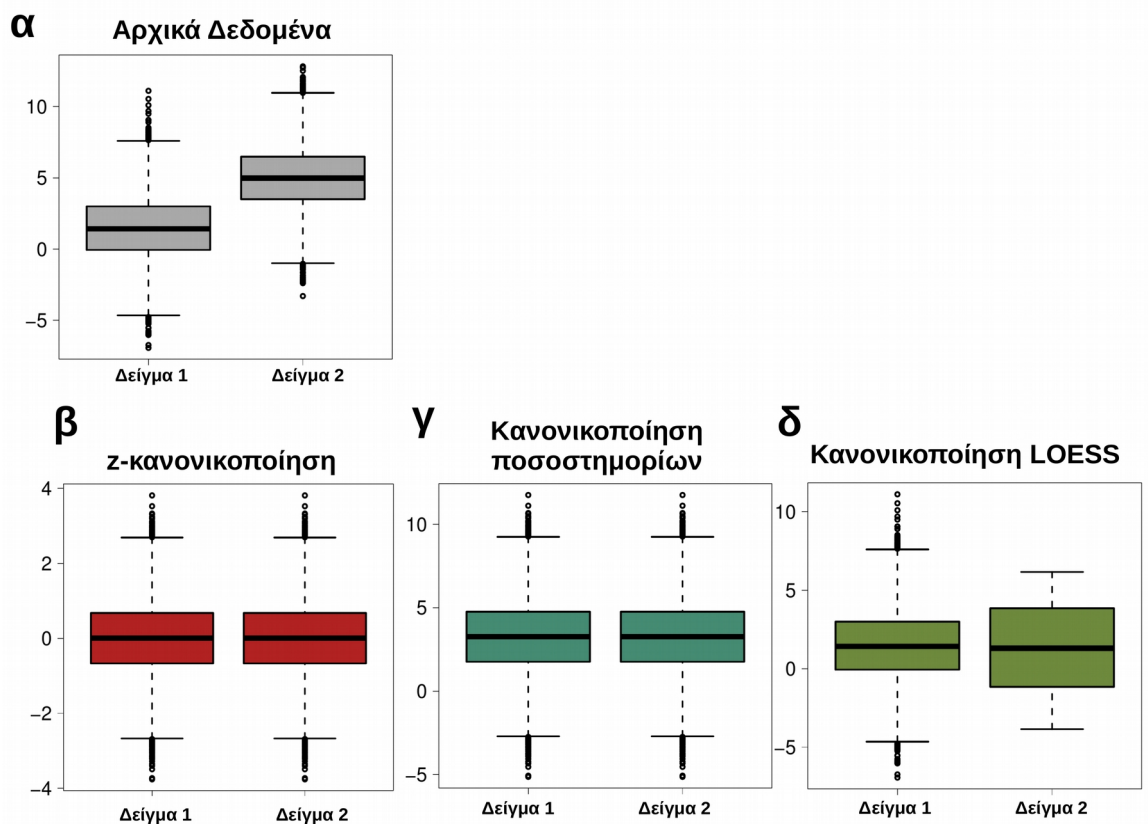
Μια πιο τυπική περιγραφή της μεθοδολογίας μπορεί να γίνει μέσω του παρακάτω αλγορίθμου:

Αλγόριθμος :: QuantileNorm
Δεδομένα Εισόδου: Ένας πίνακας E[N,M] N τιμών έκφρασης από M πειράματα
Για i=1 έως i=M:
    Υπολόγισε την κατάταξη r των στοιχείων της στήλης i του E
    Πρόσθεσε την r σε έναν πίνακα κατατάξεων R[N,M]
Δημιούργησε έναν πίνακα Q[N,M]
Κατάταξε τις τιμές κάθε στήλης του E[N,M] με αύξουσα σειρά → Q[N,M]
Για i=1 έως i=N:
    Υπολόγισε τη μέση τιμή των στοιχείων Q[i,M] → q[i]
Στον πίνακα R[N,M]:
    Αντικατάστησε την κάθε τιμή R[i,M] με την αντίστοιχη q[i]
Απόδωσε αποτέλεσμα: Πίνακας R[N,M]
Τερματισμός

Είναι προφανές ότι για την εφαρμογή της μεθόδου απαιτείται ένας πίνακας με τον ίδιο αριθμό μετρήσεων N για όλα τα πειράματα. Η κανονικοποίηση ποσοστημορίων αποτελεί τη βάση μιας ευρύτατα διαδεδομένης εφαρμογής για την ανάλυση δεδομένων μικροσυστοιχιών που ονομάζεται RMA (Robust Multi-array Analysis). Η RMA αποτελεί στην ουσία έναν συνδυασμό της κανονικοποίησης ποσοστημορίων ακολουθούμενης από μια τυποποίηση (βλ. παραπάνω) με τη χρήση διάμεσων τιμών.

Τοπική κανονικοποίηση μέσω σταθμισμένης εξομάλυνσης. (Locally weighted scatterplot smoothing, LOESS)

Η τοπική κανονικοποίηση διαφέρει από την ολική ακριβώς ως προς το γεγονός ότι εφαρμόζεται με διαφορετικό τρόπο σε υποσύνολα των δεδομένων. Μια μέθοδος που χρησιμοποιείται σε μεγάλη έκταση στην ανάλυση δεδομένων μικροσυστοιχιών είναι η LOESS που μπορεί να γίνει κατανοητή και ως μια μέθοδος τοπικής παλιδρόμησης (LOcal RegrESSion) (Ballman et al. 2004). Εφαρμόζεται σε ζευγαρωμένα δείγματα, δηλαδή σε δείγματα που προέρχονται από δύο διαφορετικές συνθήκες ή επαναλήψεις του ίδιου πειράματος με σκοπό να εξομαλύνει τόσο συστηματικά όσο και τυχαία σφάλματα. Η μέθοδος είναι εξαιρετικά απαιτητική υπολογιστικά καθώς περιλαμβάνει τα εξής βήματα:



Εικόνα 7.4: *Κανονικοποίηση δύο συνόλων τιμών έκφρασης από δύο δείγματα (α) με β) z-κανονικοποίηση που μετατρέπει την κλίμακα σε νέα κλίμακα με κέντρο το 0 γ) κανονικοποίηση ποσοστημορίων που μετατρέπει την κλίμακα σε μια σταθμισμένη κλίμακα με βάση την κατανομή ποσοστημορίων. Τόσο η β) όσο και η γ) διατηρούν τη διασπορά του δείγματος. Η κανονικοποίηση LOESS (δ) αλλάζει τις τιμές στο ένα μόνο δείγμα (εδώ Δείγμα 2) ανάλογα με το πού εφαρμόζεται το μοντέλο. Η πλήρης κανονικοποίηση περιλαμβάνει και την αντίστροφη διαδικασία (κανονικοποίηση του Δείγματος 1 με βάση το 2).*

- Εφαρμόζει παλινδρόμηση σταθμισμένων ελαχίστων τετραγώνων για διαδοχικά υποσύνολα των τιμών. Το μέγεθος των υποσυνόλων προκύπτει από μια παράμετρο εξομάλυνσης (smoothing parameter)
- Υπολογίζει μια καμπύλη τοπικής παλινδρόμησης (LOESS) η οποία είναι πολυωνυμική αλλά με μεταβλητές παραμέτρους που καθορίζονται από τα σημεία κάθε υποσυνόλου. Στο σημείο αυτό έγκεινται τόσο η τοπικότητα της μεθόδου, όσο και οι αυξημένες υπολογιστικές απαιτήσεις της.
- Επανυπολογίζει τις τιμές με βάση την απόστασή τους από την καμπύλη LOESS.

Στην Εικόνα 7.4δ φαίνεται μια χαρακτηριστική εικόνα σημείων (ζευγών τιμών έκφρασης) πριν και μετά την εφαρμογή της LOESS μεθόδου. Βασικό πλεονέκτημα της μεθόδου είναι ότι δεν προϋποθέτει τη γνώση των παραμέτρων του πολυωνύμου με βάση το οποίο γίνεται η παλινδρόμηση. Στα μειονεκτήματά της ωστόσο, πέρα από τις υπολογιστικές απαιτήσεις θα πρέπει να προσμετρηθεί το γεγονός ότι χρειάζεται μεγάλο αριθμό τιμών και μικρή διασπορά μεταξύ τους.

Διαφορική έκφραση γονιδίων

Έχοντας δει τα βασικά βήματα για την αρχική ανάλυση των δεδομένων έκφρασης περνάμε τώρα στη διαδικασία με την οποία θα εκτιμήσουμε τη διαφορική έκφραση γονιδίων από ένα πείραμα που διενεργείται σε γονιδιωματική κλίμακα. Το ερώτημα είναι:

Δεδομένων τιμών έκφρασης για Ν διαφορετικά γονίδια του ίδιου οργανισμού για δύο διαφορετικές συνθήκες, πώς θα προσδιορίσουμε ποια γονίδια είναι ενεργοποιημένα, ποια κατεσταλμένα και ποια εκείνα των οποίων η έκφραση δε μεταβάλλεται;

Στην περίπτωση πειραμάτων μεγάλης κλίμακας ο αριθμός Ν μπορεί να βρίσκεται μεταξύ 3000 (για ένα απλό προκαρυωτικό γονιδίωμα) και 50000 (για ένα σύνθετο ευκαρυωτικό γονιδίωμα στο οποίο εξετάζουμε και εναλλακτικές μορφές μεταγράφων). Οι διαφορετικές συνθήκες μπορεί να είναι στάδια στην ανάπτυξη ενός οργανισμού, διαφορετικοί κυτταρικοί τύποι, ο ίδιος κυτταρικός τύπος πριν και μετά την επίδραση μιας ουσίας ή ενός τροποποιητικού παράγοντα ή ένας πληθυσμός υγιών έναντι παθολογικών δειγμάτων. Σε κάθε περίπτωση θα πρέπει να σημειώσουμε ότι η σύγκριση γίνεται μεταξύ δύο καταστάσεων, καθώς η διαφορική έκφραση βασίζεται στην έννοια της μεταβολής. Στη συνέχεια θα δούμε πώς υπολογίζουμε την μεταβολή αυτή ως λογάριθμο λόγων έκφρασης και πώς αξιολογούμε αυτήν την τιμή στατιστικά.

Προσδιορισμός διαφορικά εκφραζόμενων γονιδίων

Ο υπολογισμός του βαθμού της μεταβολής της έκφρασης του ίδιου γονιδίου μεταξύ δύο διαφορετικών συνθηκών γίνεται με τη χρήση του λογαρίθμου του λόγου των τιμών έκφρασης στη συνθήκη μελέτης (test) προς τη συνθήκη ελέγχου (control). Η απόδοση των συνθηκών γίνεται από τον πειραματιστή και βασίζεται στο βιολογικό ερώτημα. Έτσι π.χ. αν αναζητούμε τις μεταβολές της έκφρασης σε μια παθολογική κατάσταση είναι λογικό οι τιμές των παθολογικών δειγμάτων να είναι η συνθήκη μελέτης και αυτές των φυσιολογικών να είναι η συνθήκη ελέγχου. Υπολογίζουμε το λογάριθμο του λόγου τους ως εξής:

$$\log_2 FC(g) = \log 2 \frac{E(g)_{test}}{E(g)_{control}} \quad 7.4$$

Η χρήση του λόγου των τιμών έκφρασης είναι προφανής. Τιμές του λόγου >1 θα είναι ενδεικτικές μεγαλύτερης έκφρασης στη συνθήκη μελέτης και συνεπώς ενεργοποίησης του γονιδίου ενώ τιμές <1 θα είναι ενδεικτικές καταστολής του. Η εφαρμογή του λογαρίθμου γίνεται για δύο λόγους. Αρχικά, για να μειώσει τη διασπορά των τιμών λόγων έκφρασης, με τον ίδιο τρόπο που είδαμε παραπάνω για τις καθαρές τιμές έκφρασης. Κατά δεύτερο λόγο, για να μετατρέψει το "ουδέτερο" σημείο της ποσότητας από το 1 στο 0. Οι τιμές $\log_2$FC είναι θετικές στην περίπτωση της

ενεργοποίησης του γονιδίου *g*, της αύξησης δηλαδή των επιπέδων έκφρασής του στη συνθήκη μελέτης σε σχέση με τη συνθήκη ελέγχου και αρνητικές στην περίπτωση καταστολής. Μηδενικές μεταβολές αντιστοιχίζονται στην τιμή 0. Επιπλέον, η χρήση του δυαδικού λογάριθμου επιτρέπει μια απευθείας ανάγνωση του βαθμού της διαφορικής έκφρασης. Μια τιμή $\log_2$FC=1 σημαίνει διπλάσια έκφραση σε σχέση με τη συνθήκη ελέγχου, ενώ μια τιμή $\log_2$FC=-1 υποδηλώνει μείωση στο μισό. Η χρήση της λογαριθμικής κλίμακας επιτρέπει γενικότερα ευκολότερη ερμηνεία των αποτελεσμάτων. Στην περίπτωση που η έκφραση του κάθε γονιδίου έχει μετρηθεί περισσότερες από μία φορές σε επαναλήψεις του ίδιου πειράματος, η εξίσωση 7.4 δεν αλλάζει αλλά στη θέση των $E(g)_{test}$ και $E(g)_{control}$ χρησιμοποιούνται οι αντίστοιχες μέσες τιμές των επαναλήψεων. Η σημασία της πραγματοποίησης επαναλήψεων του ίδιου πειράματος είναι πολύ μεγάλη για λόγους στατιστικής αξιολόγησης που θα συζητηθούν στη συνέχεια.

Σημειώνεται εδώ ότι η ζευγαρωτή σχέση συνθήκης μελέτης/ελέγχου δε σημαίνει ότι σε κάθε πείραμα υπάρχει μόνο μια συνθήκη μελέτης. Κρατώντας σταθερή τη συνθήκη ελέγχου κανείς μπορεί να υπολογίσει τη σχετική έκφραση σε μια σειρά από καταστάσεις. Έτσι π.χ. μπορεί κανείς να συγκρίνει παθολογικά δείγματα με δείγματα στα οποία οι ασθενείς υποβάλλονται σε διαφορετικές θεραπείες ή να μελετήσει τη διαφορική έκφραση σε διαφορετικά στάδια μιας διαδικασίας εφ' όσον όλα συγκρίνονται με ένα αρχικό χρονικό σημείο t=0 κλπ.

Μαθηματικό Ιντερμέδιο ΙΙ. Στατιστική ανάλυση διαφορικής έκφρασης

Είδαμε παραπάνω ότι η σχετική έκφραση ποσοτικοποιείται σε μια λογαριθμική κλίμακα λόγων. Πώς όμως μπορούμε να προσδιορίσουμε κατά πόσο μια τιμή $\log_2$FC=1 είναι αρκετά μεγάλη ώστε να κατατάξουμε το γονίδιο στο οποίο αντιστοιχεί στα ενεργοποιημένα; Θα χρειαστούμε ένα στατιστικό έλεγχο της τιμής αυτής ώστε να αξιολογήσουμε σε ποιο βαθμό η διπλάσια έκφραση αντανακλά ένα υπαρκτό βιολογικό φαινόμενο ή μια τυχαία διακύμανση στα επίπεδα του mRNA. Οπωσδήποτε, ακραίες τιμές $\log_2$FC είναι καταρχάς ενδεικτικές πραγματικών διαφορών, ωστόσο ένα απλό διαισθητικά ερώτημα σχετίζεται με το κατά πόσο μια τιμή είναι επαναλήψιμη. Πιο απλά, αν θα παίρναμε την ίδια (ή σχεδόν την ίδια) τιμή $\log_2$FC αν επαναλαμβάναμε το πείραμα. Η έννοια της επαναληψιμότητας είναι βασική για το στατιστικό (και ουσιαστικό) έλεγχο κάθε πειράματος, αλλά στην περίπτωση των πειραμάτων έκφρασης έχει ιδιαίτερα κεντρικό ρόλο. Είναι πρακτικά αδύνατο να αξιολογήσουμε τη διαφορική έκφραση μ' ένα μοναδικό πείραμα. Κι αυτό γιατί ο στατιστικός έλεγχος απαιτεί να συγκρίνουμε όχι ένα ζεύγος τιμών για κάθε γονίδιο αλλά ένα ζεύγος κατανομών των τιμών αυτών. Απ' αυτήν τη σκοπιά είναι σημαντικό να γνωρίζουμε ότι κάθε πείραμα έκφρασης θα πρέπει να διενεργείται τουλάχιστο σε τρεις επαναλήψεις, που σημαίνει ότι για κάθε συνθήκη και για κάθε γονίδιο θα πρέπει να έχουμε τουλάχιστον τρεις μετρήσεις έκφρασης (Lee et al. 2000). Ο λόγος γι' αυτόν τον περιορισμό είναι ότι για να αξιολογήσουμε τη διαφορική έκφραση στατιστικά, θα πρέπει να υπολογίσουμε μια πιθανότητα που να εκτιμά τη διαφορά μεταξύ των μέσων τιμών των μετρήσεων για κάθε συνθήκη (και για μια μέση τιμή καλό είναι να έχουμε τουλάχιστον τρεις μετρήσεις). Σε πειράματα έκφρασης, όπου εξετάζουμε μια συνθήκη μελέτης με μια συνθήκη ελέγχου, ο στατιστικός έλεγχος γίνεται με τη χρήση του ελέγχου t (Student's t-test ή πιο απλά t-test)

που αποτελεί ένα στατιστικό έλεγχο υποθέσεων. Η υπόθεση αυτή που στη στατιστική ονομάζεται και "μηδενική υπόθεση" (null hypothesis) και συμβολίζεται με H0 είναι, στην περίπτωση του t-test, ότι οι μέσες τιμές δύο συνόλων τιμών είναι ταυτόσημες[5].

Πιο αναλυτικά, δεδομένων δύο συνόλων τιμών X_1 και X_2, το t-test υπολογίζει ένα μέγεθος t[6] το οποίο είναι μικρό αν οι μέσες τιμές μ_1 και μ_2 είναι παραπλήσιες. Όσο μεγαλύτερο είναι το t τόσο μικρότερη είναι η πιθανότητα οι δύο μέσες τιμές να ταυτίζονται. Η διενέργεια του t-test γίνεται πολύ γρήγορα σε όλα τα διαθέσιμα στατιστικά προγράμματα (R, SPSS, Matlab κλπ) και σε κάθε περίπτωση η αποδιδόμενη τιμή είναι η πιθανότητα ταύτισης των δύο μέσων τιμών, που αντιστοιχίζεται σε μια τιμή p (p-value). Όσο μικρότερη είναι η τιμή p-value τόσο μικρότερη είναι η πιθανότητα οι δύο μέσες τιμές να είναι ίδιες και συνεπώς τα δύο δείγματα X_1 και X_2 να προέρχονται από την ίδια κατανομή. Κατ' αυτόν τον τρόπο, μικρές τιμές p-value σε πειράματα έκφρασης είναι ισχυρή ένδειξη ότι οι διαφορές που παρατηρούνται στα επίπεδα έκφρασης ενός γονιδίου είναι βιολογικά σημαντικές.

Στην Εικόνα 7.5 μπορούμε να δούμε πώς οι δύο αυτές τιμές μπορούν να αναπαρασταθούν γραφικά για να περιγράψουν τη γενικότερη εικόνα ενός πειράματος διαφορικής έκφρασης που συγκρίνει δύο συνθήκες. Η Εικόνα 7.5 αναπαριστά ένα "διάγραμμα ηφαιστείου" ή volcano plot όπως είναι γνωστό. Πρόκειται ουσιαστικά για ένα διάγραμμα σκέδασης όπου κάθε σημείο αντιστοιχεί σε ένα γονίδιο. Οι καρτεσιανές συντεταγμένες του κάθε γονιδίου είναι η τιμή $\log_2$FC στον οριζόντιο άξονα και ο αρνητικός δεκαδικός λογάριθμος του p-value στον κάθετο. Η εγγενής, αναμενόμενη τάση που έχουν οι ακραίες τιμές $\log_2$FC να αντιστοιχούν σε μικρές τιμές p-value οδηγεί στο χαρακτηριστικό σχήμα που προσομοιάζει έναν κρατήρα ηφαιστείου. Μ' αυτόν τον τρόπο, όσο ψηλότερα στον κάθετο άξονα βρίσκεται ένα σημείο (μεγάλη τιμή αρνητικού λογαρίθμου του p-value) τόσο πιο σημαντική στατιστικά είναι η διαφορική του έκφραση, ενώ όσο πιο μακριά από το σημείο 0 στον οριζόντιο άξονα, τόσο πιο μεγάλη είναι η έντασή της.

Πώς μπορούμε να χρησιμοποιήσουμε αυτήν την αναπαράσταση για να εκτιμήσουμε τη διαφορική έκφραση; Στην πράξη, ο προσδιορισμός των γονιδίων με διαφορική έκφραση γίνεται ορίζοντας κάποια όρια τιμών $\log_2$FC και p-value. Στη βιβλιογραφία θα συναντήσουμε συχνά την τιμή p-value<=0.05 ως όριο σημαντικότητας και την αντίστοιχη απόλυτη τιμή $|\log_2 FC| > 1.5$ ως όριο διαφορικής έκφρασης. Αυτό σημαίνει ότι γονίδια με $\log_2 FC > 1.5$ ή $\log_2 FC$ και < -1.5 που ταυτόχρονα έχουν p-value<=0.05 προσδιορίζονται ως ενεργοποιημένα και κατεσταλμένα αντίστοιχα. Ονομάζουμε αυτά τα γονίδια διαφορικά εκφραζόμενα (diffentially expressed genes, DEG). Τα συγκεκριμένα όρια είναι αυτά που έχουν χρησιμοποιηθεί στην Εικόνα 7.5 για να χρωματίσουν με διαφορετικό τρόπο τα ενεργοποιημένα (κόκκινα) από τα κατεσταλμένα (πράσινα) γονίδια. Φυσικά τα όρια αυτά είναι αυθαίρετα και όχι σπάνια, μπορεί κανείς να συναντήσει διαφορετικά (περισσότερο ή λιγότερο αυστηρές τιμές κατωφλίων), η γενική λογική όμως είναι ότι για να προσδιορίσουμε ένα γονίδιο ως διαφορικά εκφραζόμενο χρειάζεται μια επαρκώς μεγάλη (σε απόλυτη τιμή) $\log_2$FC και μια αντίστοιχη τιμή p-value όχι μεγαλύτερη από 0.05.

---

[5] *Στην περίπτωση που συγκρίνουμε περισσότερα των δύο δείγματα ο στατιστικός έλεγχος γίνεται με τη χρήση της ανάλυσης διακύμανσης (Analysis of Variance, ANOVA), την οποία δε θα συζητήσουμε σε αυτό το σημείο, καθώς μας ενδιαφέρει περισσότερο να γίνει κατανοητή η λογική μιας ζευγαρωτής σύγκρισης.*

[6] *Ο μαθηματικός τύπος για τον υπολογισμό του t παραλείπεται καθώς είναι αρκετά πολύπλοκος χωρίς να προσφέρει στην κατανόηση της σχετικής θεωρίας. Αρκεί να αναφέρουμε ότι τόσο ο υπολογισμός του, όσο και η διενέργεια του t-test γίνεται με μοναδικά δεδομένα τα δύο σύνολα τιμών X_1 και X_2.*

***Ερώτηση**: Τι συμπεράσματα θα βγάζατε για γονίδια που σε ένα volcano plot θα αντιστοιχούσαν σε σημεία που θα βρίσκονταν ψηλά στον κάθετο y άξονα αλλά κοντά στο 0 στον αντίστοιχο οριζόντιο x άξονα;*

***Εικόνα 7.5:** Διάγραμμα "κρατήρα ηφαιστείου", (volcano plot) από ένα πείραμα μέτρησης διαφορικής γονιδιακής έκφρασης. Κάθε σημείο αντιστοιχεί σε ένα γονίδιο με τη θέση στον οριζόντιο άξονα να αντιστοιχεί στο δυαδικό λογάριθμο του λόγου διαφορικής έκφρασης και τη θέση στον κάθετο άξονα να αντιστοιχεί στον αρνητικό δεκαδικό λογάριθμο της τιμής p-value. Με πράσινο και κόκκινο φαίνονται τα στατιστικά σημαντικά υπο- και υπερ-εκφραζόμενα γονίδια (για τιμές κατωφλίων |log2FC|>=1.5 και p-value<=0.05).*

Ακόμα όμως και χαμηλές τιμές p-value θα πρέπει να αντιμετωπίζονται με προσοχή. Στην περίπτωση που υπολογίζουμε μεγάλο αριθμό από p-values (όπως στην περίπτωση ενός πειράματος γονιδιακής έκφρασης) αυτό που κάνουμε είναι να διενεργήσουμε τον ίδιο στατιστικό έλεγχο πολλές

φορές, να κάνουμε δηλαδή αυτό που στη στατιστική ονομάζεται "έλεγχος πολλαπλών υποθέσεων". Αυτό που συμβαίνει σε αυτές τις περιπτώσεις είναι ότι για καθαρά στατιστικούς λόγους κάποιες τιμές p-values, που είναι αρκετά μικρές ώστε να χαρακτηρίσουν στατιστικά σημαντικές αλλαγές στα επίπεδα έκφρασης, μπορούν να έχουν προκύψει τυχαία. Αυτές θα είναι περισσότερες όσο μεγαλύτερος είναι ο αριθμός των πολλαπλών υποθέσεων (Dudoit, Shaffer, and Boldrick 2003). Περισσότερα όμως για αυτήν τη σημαντική πτυχή της στατιστικής ανάλυσης καθώς και για τις απαραίτητες διορθώσεις θα συζητήσουμε στο αμέσως επόμενο κεφάλαιο.

Ανάλυση γονιδίων με κοινά χαρακτηριστικά πρότυπα έκφρασης

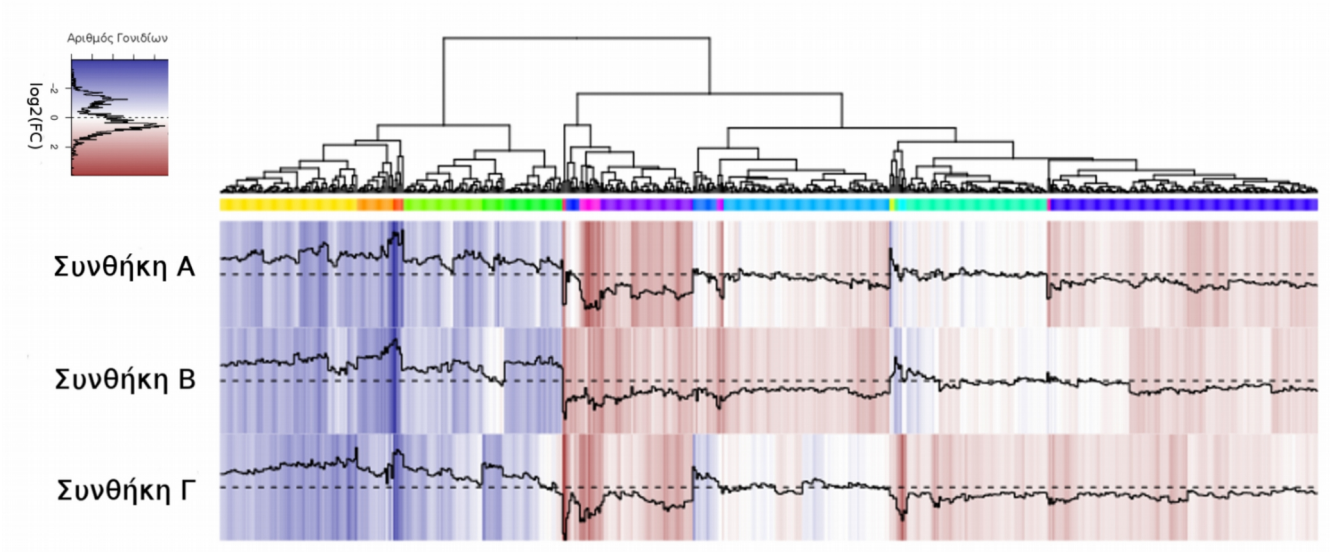
Έχουμε ως τώρα απαντήσει στα δύο από τα τρία ερωτήματα που θέσαμε αρχικά, σχετικά με την ποσοτικοποίηση των ρυθμών έκφρασης και τον προσδιορισμό διαφορικά εκφραζόμενων γονιδίων. Στο τελευταίο αυτό μέρος του κεφαλαίου, θα προσπαθήσουμε να απαντήσουμε στο τρίτο και ίσως σημαντικότερο από τα ερωτήματα που έχει να κάνει με την ερμηνεία ενός πειράματος έκφρασης. Το ερώτημα αυτό θυμηθείτε ήταν:

Πώς μπορούμε να ορίσουμε ομάδες-υποσύνολα γονιδίων που έχουν κοινά χαρακτηριστικά στα πρότυπα έκφρασής τους και να εντοπίσουμε γονίδια που ενεργοποιούνται ή/και καταστέλλονται κάτω από τις ίδιες συνθήκες;

Το πρόβλημα έχει ιδιαίτερο ενδιαφέρον καθώς δε σχετίζεται μόνο με την ύπαρξη ή όχι ενεργοποιημένων και κατεσταλμένων γονιδίων αλλά και με τις πιθανές υποκείμενες σχέσεις γονιδιακής ρύθμισης που υπάρχουν μεταξύ διαφορετικών γονιδίων και διαφορετικών συνθηκών. Συγκρίνοντας την έκφραση γονιδίων σε περισσότερες από δύο συνθήκες, μπορούμε να προσδιορίσουμε γονίδια που συμμεταβάλλουν την έκφρασή τους, γονίδια δηλαδή που τείνουν να ενεργοποιούνται μαζί και αντίστοιχα να καταστέλλονται μαζί, ή αντίθετα γονίδια που φαίνεται να ακολουθούν αντίρροπες τάσεις σε ό,τι αφορά την έκφρασή τους. Τέτοιου είδους σχέσεις μπορούν αρχικά να αναπαρασταθούν γραφικά με τη μορφή θερμικών χαρτών (heatmaps) οι οποίοι είναι γραφήματα τριών διαστάσεων που πρακτικά συμπιέζονται σε δύο, μέσω της απόδοσης της τρίτης διάστασης σε ένα χρωματικό κώδικα (Wilkinson and Friendly 2012). Οι θερμικοί χάρτες χρησιμοποιούνται ευρύτατα στην παρουσίαση αποτελεσμάτων πειραμάτων γονιδιακής έκφρασης (θυμηθείτε ότι τους συζητήσαμε ήδη από το εισαγωγικό κεφάλαιο) και είναι καλό να εξοικειωθούμε με την επισκόπηση και την ερμηνεία τους.

Ένας θερμικός χάρτης ενός πειράματος έκφρασης γονιδίων φαίνεται στην Εικόνα 7.6. Στο διάγραμμα αυτό, κάθε γραμμή αντιστοιχεί σε μια διαφορετική πειραματική συνθήκη και κάθε στήλη σε ένα γονίδιο. Στις δύο διαστάσεις αναπαρίστανται οι σχετικές τιμές έκφρασης (συνήθως σε $\log_2$FC) σε σχέση με τη συνθήκη ελέγχου. Κάθε στοιχείο του πίνακα (που έχει διαστάσεις $N_{[αριθμός\ γονιδίων]}$ x $M_{[αριθμός\ συνθηκών]}$) χρωματίζεται ανάλογα με την τιμή $\log_2$FC με μια σχετική διαβάθμιση στην ένταση μεταξύ αρνητικών (εδώ γαλάζιων) και θετικών (εδώ κόκκινων) τιμών.

Το βασικό χαρακτηριστικό πλεονέκτημα των θερμικών χαρτών είναι ότι η σειρά των γονιδίων δεν είναι τυχαία αλλά μπορεί να διαμορφωθεί με τρόπο που αντανακλά τις σχέσεις των τιμών έκφρασης μεταξύ γονιδίων και συνθηκών. Οι σχέσεις αυτές συχνά αναπαρίστανται μ' ένα δενδρόγραμμα που συνοδεύει το θερμικό χάρτη και ουσιαστικά αντιστοιχεί σε μια ομαδοποίηση των γονιδίων με βάση την ομοιότητα των προτύπων έκφρασής τους. Πώς όμως θα υπολογίσουμε τις ομοιότητες μεταξύ των προτύπων αυτών; Στην επόμενη ενότητα θα ορίσουμε επαρκώς την έννοια της ομοιότητας και τους τρόπους με τους οποίους μπορούμε να την ποσοτικοποιήσουμε στην περίπτωση των πειραμάτων γονιδιακής έκφρασης.



Εικόνα 7.6: *Θερμικός χάρτης που αναπαριστά τις σχετικές τιμές έκφρασης 650 γονιδίων όπως αυτές μετρήθηκαν σε τρεις διαφορετικές συνθήκες (Α, Β και Γ). Το γαλάζιο αντιστοιχεί σε χαμηλότερη και το κόκκινο σε υψηλότερη έκφραση σε σχέση με την κατάσταση ελέγχου, καθώς στο θερμικό χάρτη εμφανίζονται μόνο σχετικές τιμές έκφρασης. Ο χάρτης συνοδεύεται από ιεραρχική ομαδοποίηση (βλ. Παρακάτω) των γονιδίων με βάση τα πρότυπα έκφρασής τους στις τρεις συνθήκες. Γονίδια που βρίσκονται στον ίδιο κλάδο του δέντρου εμφανίζουν μεγαλύτερη ομοιότητα σε ό,τι αφορά την αυξομείωση των επιπέδων έκφρασης μεταξύ των συνθηκών.*

Ερώτηση*: Σε μια προσπάθεια ερμηνείας των αποτελεσμάτων του θερμικού χάρτη της Εικόνας 7.6, μπορείτε να περιγράψετε σε τι διαφέρουν τα πρότυπα έκφρασης των γονιδίων που αντιστοιχούν στην ανοιχτή πράσινη και στη γαλάζια ομαδα; (τα χρώματα αναφέρονται στη ζώνη που συνοδεύει το δενδρόγραμμα).*

Μαθηματικό Ιντερμέδιο ΙΙΙ. Μέτρα Απόστασης και Συντελεστές Συσχέτισης

Έχοντας στα χέρια μας ένα προφίλ γονιδιακής έκφρασης με διαστάσεις $N_{[αριθμός\ γονιδίων]}$ x $M_{[αριθμός}$

συνθηκών] αυτό που μας ενδιαφέρει είναι να εντοπίσουμε γονίδια (ή εναλλακτικά συνθήκες) των οποίων τα πρότυπα έκφρασης μοιάζουν μεταξύ τους. Χρειάζεται έτσι να ποσοτικοποιήσουμε μια έννοια ομοιότητας μεταξύ αριθμητικών τιμών. Μέτρα ποσοτικοποίησης της ομοιότητας έχουμε ήδη συζητήσει σε αρκετά από τα προηγούμενα κεφάλαια κι έτσι οι έννοιες της ομοιότητας και της απόστασης και η μεταξύ τους σχέση δε θα πρέπει να μας ξενίζει. Στην περίπτωση των πειραμάτων γονιδιακής έκφρασης η φύση των δεδομένων και ο αριθμός των διαστάσεών τους είναι τέτοια που επιβάλουν η έννοια της απόστασης να οριστεί πιο αυστηρά. Έτσι, για τη μέτρηση της ομοιότητας δεδομένων γονιδιακής έκφρασης χρησιμοποιούμε ποσότητες που ορίζονται με βάση την μαθηματική έννοια του **μέτρου** και έχουν συγκεκριμένες ιδιότητες. Μια ποσότητα έχει χαρακτηριστικά μέτρου απόστασης όταν:

- Η απόσταση μεταξύ δύο οποιωνδήποτε αντικειμένων είναι μεγαλύτερη ή ίση με το μηδέν.
- Η απόσταση μεταξύ ενός αντικειμένου και του εαυτού του είναι πάντα ίση με το μηδέν (και αντίστροφα, όταν δύο αντικείμενα έχουν μηδενική απόσταση τότε ταυτίζονται).
- Η απόσταση μεταξύ των Α και Β είναι ίση με την απόσταση μεταξύ των Β και Α, δηλαδή η ποσότητα είναι συμμετρική (αντιμεταθετικότητα).
- Η απόσταση μεταξύ των αντικειμένων Α και Γ θα πρέπει να είναι μικρότερη από ή ίση με το άθροισμα των αποστάσεων μεταξύ Α και Β, και Β και Γ (τριγωνική ανισότητα).

Έχοντας υπόψη τα παραπάνω ως απαραίτητες προϋποθέσεις, μπορούμε να ορίσουμε ποσότητες που θα μετρούν την απόσταση μεταξύ προτύπων έκφρασης. Φανταστείτε καταρχάς την πιο απλή, διαισθητικά, έκφραση της απόστασης, αυτήν της γεωμετρικής απόστασης μεταξύ δύο σημείων (α,β) σε ένα σύστημα συντεταγμένων με συντεταγμένες $a(x_1,y_1)$ και $\beta(x_2,y_2)$. Η απόστασή τους δίνεται από τον τύπο:

$$d(\alpha,\beta)=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2} \quad 7.5$$

Η 7.5 είναι ουσιαστικά ο τύπος που υπολογίζει το μήκος της υποτείνουσας του ορθογωνίου τριγώνου που ορίζουν τα σημεία α, β και το σημείο τομής των κάθετων προβολών τους και για το λόγο αυτό ονομάζεται **Ευκλείδεια Απόσταση** των α και β. Στην απλούστερή της μορφή η Ευκλείδεια απόσταση υπολογίζει την απόσταση δύο σημείων σε δύο διαστάσεις. Τι συμβαίνει όμως όταν θέλουμε να υπολογίσουμε την απόσταση δύο γονιδίων g_1 και g_2, των οποίων η έκφραση έχει μετρηθεί σε περισσότερες από δύο συνθήκες; Η Ευκλείδεια Απόσταση μπορεί να υπολογιστεί σε Ν διαστάσεις χωρίς ο τύπος να μεταβληθεί σημαντικά. Συγκεκριμένα, η εξίσωση που δίνει αυτήν την απόσταση είναι:

$$dEuclidean(g_1,g_2)=\sqrt{\sum_{i=1}^{N}(\log_2 FC_{i,1}-\log 2\, FC_{i,2})^2} \quad 7.6$$

υπολογίζοντας δηλαδή την τετραγωνική ρίζα του αθροίσματος των επιμέρους τετραγώνων των διαφορών για κάθε συνθήκη *i*.

Η Ευκλείδεια Απόσταση αποτελεί την πιο ευρέως χρησιμοποιούμενη μέθοδο απόστασης λόγω της εύκολης γεωμετρικής της ερμηνείας αλλά δεν είναι η μόνη. Άλλα μέτρα απόστασης είναι η απόσταση Manhattan που δίνεται από τον τύπο:

$$dManhattan(g_{1,}g_2)=\sum_{i=1}^{N}|(\log_2 FC_{i,1}-\log 2\, FC_{i,2})| \quad 7.7$$

η απόσταση Camberra:

$$dCamberra(g_{1,}g_2)=\sum_{i=1}^{N}\frac{|(\log_2 FC_{i,1}-\log 2\, FC_{i,2})|}{|(\log_2 FC_{i,1}+\log 2\, FC_{i,2})|} \quad 7.8$$

και η γενικευμένη απόσταση Minkowski:

$$dMinkowski(g_{1,}g_2)=\left(\sum_{i=1}^{N}(\log_2 FC_{i,1}-\log 2\, FC_{i,2})^r\right)^{\frac{1}{r}} \quad 7.9$$

Στην πραγματικότητα, οι τελευταίες χρησιμοποιούνται πολύ σπάνια για την ανάλυση της απόστασης μεταξύ προτύπων γονιδιακής έκφρασης, είναι όμως χρήσιμο να αναλογιστούμε ότι οι αποστάσεις μπορούν να υπολογιστούν με διαφορετικούς τρόπους. Ένα χαρακτηριστικό που έχουν όλα τα μέτρα απόστασης (άλλα περισσότερο και άλλα λιγότερο) είναι ότι δεν είναι ανεξάρτητα από την κλίμακα. Αυτό σημαίνει ότι δύο γονίδια μπορεί να μοιάζουν ποιοτικά σε ό,τι αφορά την αυξομοίωση της έκφρασής τους αλλά επειδή οι απόλυτες τιμές της έκφρασής τους διαφέρουν σημαντικά αυτό να οδηγεί και σε μεγάλη απόσταση. Ακόμα πιο σημαντικό είναι το αντίθετο φαινόμενο, η ανεξαρτησία δηλαδή των μέτρων απόστασης από τη συνδιακύμανση των τιμών έκφρασης δύο γονιδίων. Μπορούμε να φανταστούμε δύο γονίδια των οποίων οι τιμές έκφρασης διαφέρουν ποιοτικά σε μεγάλο αριθμό συνθηκών, με το ένα να εμφανίζει ανεβασμένα επίπεδα εκεί που το άλλο εμφανίζει μειωμένα και αντίστροφα. Αν ωστόσο οι απόλυτες τιμές έκφρασης είναι κοντά, τότε τα μέτρα απόστασης θα δίνουν μικρές αποστάσεις και κατά συνέπεια μια εσφαλμένη εκτίμηση της ομοιότητας μεταξύ των γονιδιακών προτύπων έκφρασης.

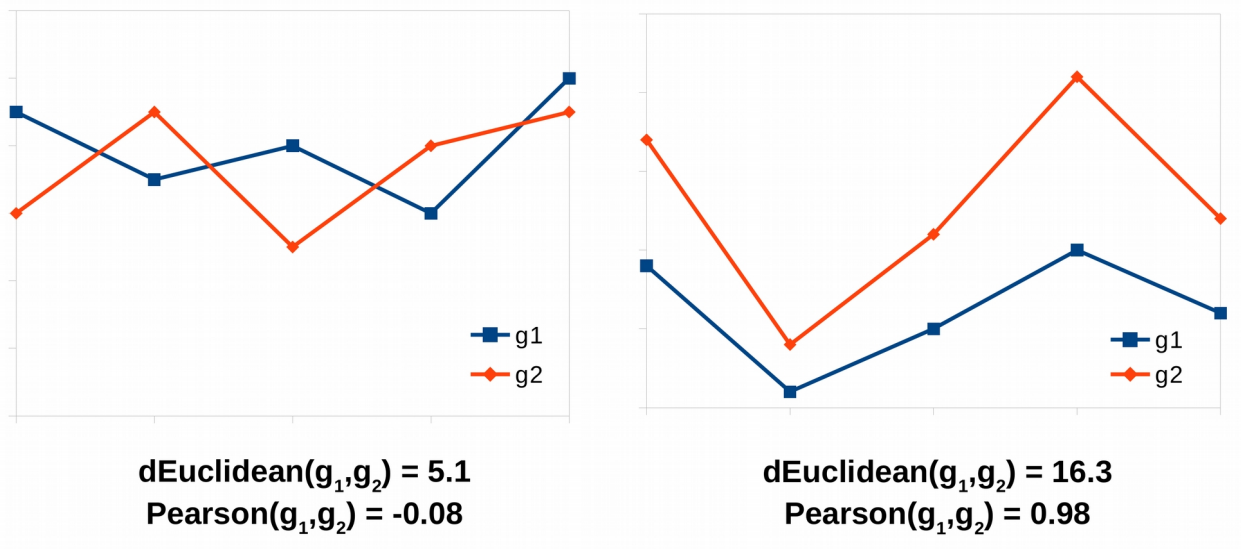
Για την αποφυγή τέτοιων εκτιμήσεων, στην πράξη, πέρα από την Ευκλείδεια Απόσταση, για τη σύγκριση προτύπων γονιδιακής έκφρασης χρησιμοποιούνται συχνότερα συντελεστές στατιστικής συσχέτισης. Οι συντελεστές συσχέτισης είναι στατιστικές ποσότητες που ποσοτικοποιούν το βαθμό συνδιακύμανσης δύο μεγεθών (στην περίπτωση των δεδομένων γονιδιακής έκφρασης, αυτές είναι οι τιμές έκφρασης δύο γονιδίων για Ν συνθήκες). Οι συντελεστές συσχέτισης δεν έχουν χαρακτηριστικά απόστασης αλλά αποδίδουν ασφαλέστερα ποιοτικά χαρακτηριστικά όπως είναι ο βαθμός στον οποίον δύο μεγέθη τείνουν να "ανεβοκατεβαίνουν" με τον ίδιο τρόπο. Ο πιο κοινός συντελεστής συσχέτισης είναι ο συντελεστής Γραμμικής Συσχέτισης Pearson που δίνεται από τον (κάπως πολύπλοκο) τύπο:

$$rPearson(g_{1,}g_2)=\frac{N\sum_{i=1}^{N}g_{i,1}g_{i,2}-\sum_{i=1}^{N}g_{i,1}\sum_{i=1}^{N}g_{i,2}}{\sqrt{\left(N\sum_{i=1}^{N}g_{i,1}^2-\left(\sum_{i=1}^{N}g_{i,1}\right)^2\right)\left(N\sum_{i=1}^{N}g_{i,2}^2-\left(\sum_{i=1}^{N}g_{i,2}\right)^2\right)}} \quad 7.10$$

Οι συντελεστές συσχέτισης είναι θετικοί όταν τα δύο μεγέθη διακυμαίνονται με τον ίδιο τρόπο (ακολουθούν το ένα τα "σκαμπανεβάσματα" του άλλου) και αρνητικοί όταν η διακύμανσή τους έχει αντίθετη φορά. Τα όρια των τιμών συσχέτισης είναι από -1 (τέλεια αντίθετη διακύμανση)

έως 1 (τέλεια συνδιακύμανση) με την τιμή 0 να αντιστοιχεί σε μεγέθη που διακυμαίνονται μάλλον τυχαία μεταξύ τους. Στην Εικόνα 7.7 φαίνεται οι διαφορές μεταξύ της χρήσης ενός μέτρου απόστασης και ενός συντελεστή συσχέτισης για δύο διαφορετικά ζεύγη προτύπων έκφρασης. Στην πρώτη περίπτωση η συνδιακύμανση είναι αρνητική αλλά οι απόλυτες τιμές έκφρασης είναι παραπλήσιες. Έτσι η Ευκλείδεια Απόσταση είναι μικρή και ο συντελεστής Pearson ελαφρά αρνητικός. Στη δεύτερη περίπτωση οι διαφορές μεταξύ των απόλυτων τιμών έκφρασης αποδίδουν μεγάλη απόσταση αλλά η συνδιακύμανση είναι ισχυρά θετική. Η Ευκλείδεια Απόσταση αποτυγχάνει εδώ να αποδώσει την ποιοτική σχέση μεταξύ των δύο προτύπων που κανείς μπορεί να διακρίνει ακόμα και με το μάτι, όμως ο συντελεστής Pearson συλλαμβάνει αυτό το χαρακτηριστικό δίνοντας μια ισχυρά θετική τιμή (0.98).

Ο συντελεστής συσχέτισης Pearson είναι ένας μόνο από τους πολλούς τρόπους υπολογισμού συσχετίσεων όμως επειδή η ανάλυση στατιστικών συσχετίσεων είναι πολύ χρήσιμη για την εξαγωγή σχέσεων μεταξύ δεδομένων, πιο αναλυτική συζήτησή της αφήνεται για επόμενα κεφάλαια.



Εικόνα 7.7: *Η διαφορά μεταξύ ενός μέτρου απόστασης (Ευκλείδεια Απόσταση) κι ενός συντελεστή συσχέτισης (Συντελεστής γραμμικής συσχέτισης Pearson) στην ποσοτικοποίηση δύο προτύπων έκφρασης δύο γονιδίων για πέντε διαφορετικές συνθήκες.*

Ομαδοποίηση

Έχοντας εξετάσει τρόπους για την ποσοτικοποίηση των ομοιοτήτων μεταξύ διαφορετικών προτύπων έκφρασης γονιδίων, περνάμε τώρα στη συζήτηση επιμέρους μεθοδολογιών ομαδοποίησης τους. Με τον όρο ομαδοποίηση αναφερόμαστε γενικώς στη διαδικασία οργάνωσης ενός συνόλου αντικειμένων σε υποσύνολα με τέτοιο τρόπο ώστε οι ομοιότητες των αντικειμένων εντός των υποσυνόλων αυτών να είναι μεγαλύτερες από ότι μεταξύ διαφορετικών υποσυνόλων. Από θεωρητική άποψη, η ομαδοποίηση είναι μια από τις πιο διαδεδομένες τεχνικές μηχανικής

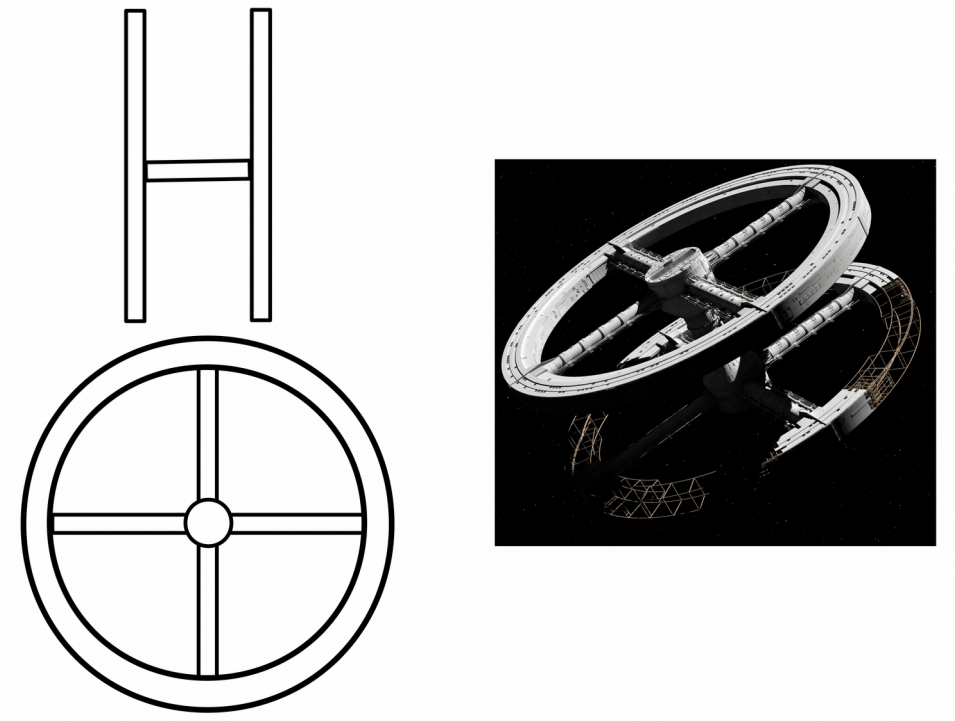
μάθησης για τη διερεύνηση ομοιοτήτων (και διαφορών) μεταξύ δεδομένων (Everitt et al. 2011). Πρακτικά αποτελεί μία βασική μέθοδο στατιστικής ανάλυσης δεδομένων που αποσκοπεί στο να αναδείξει επιμέρους χαρακτηριστικά της οργάνωσής τους.

Στην περίπτωση των πειραμάτων γονιδιακής έκφρασης η ανάλυση ομαδοποίησης μπορεί να εντοπίσει υποσύνολα γονιδίων των οποίων τα πρότυπα έκφρασης μοιάζουν και που κατά συνέπεια, θα είναι πολύ πιθανό να εμπλέκονται σε κοινές ή παρόμοιες κυτταρικές λειτουργίες. Η ανάλυση ομαδοποίησης (cluster analysis) αποτελεί τη μεθοδολογική βάση για την μελέτη προτύπων έκφρασης. Στις επόμενες ενότητες θα περιγράψουμε τις βασικότερες μεθόδους ομαδοποίησης που χρησιμοποιούνται στην ανάλυση της γονιδιακής έκφρασης.

Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis, PCA)

Το βασικότερο ποιοτικό χαρακτηριστικό των δεδομένων γονιδιακής έκφρασης είναι η πολυπλοκότητά τους σε ότι αφορά τις διαστάσεις τους. Με το όρο “διαστάσεις” ή καλύτερα “διαστασιμότητα” (dimensionality) αναφερόμαστε στη δομή των δεδομένων που περιλαμβάνουν πολύ συχνά τιμές έκφρασης για χιλιάδες ή δεκάδες χιλιάδες γονιδίων και για μεγάλο αριθμό διαφορετικών συνθηκών. Στην προσπάθειά μας να διακρίνουμε ομάδες, υποσύνολα γονιδίων που συμπεριφέρονται με παρόμοιο τρόπο, μια πρώτη προσέγγιση που θα πρέπει να αναλογιστούμε είναι να προσπαθήσουμε να μειώσουμε την πολυπλοκότητα στο χώρο των διαστάσεων. Τεχνικές που αποσκοπούν σ' αυτό ονομάζονται τεχνικές “μείωσης διαστασιμότητας” (dimensionality reduction techniques). Η πιο χαρακτηριστική από αυτές είναι η Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis, στο εξής PCA) (Abdi and Williams 2010). Η ευρύτερα χρησιμοποιούμενη μέθοδος μείωσης διαστασιμότητας, η PCA έχει ως βασικό στόχο ν' αναδείξει τις σημαντικότερες διαφορές μεταξύ των στοιχείων μιας δομής δεδομένων (στην περίπτωσή μας, τα πρότυπα έκφρασης χιλιάδων γονιδίων) κρατώντας και παρουσιάζοντας εκείνα τα χαρακτηριστικά τους που ευθύνονται για τη μεγαλύτερη ποικιλομορφία στο δείγμα μας.

Για να καταλάβετε καλύτερα την αρχή της μεθόδου, αναλογιστείτε το παρακάτω διαισθητικό παράδειγμα. Φανταστείτε ότι προσπαθείτε να διακρίνετε τη μορφή ενός αντικειμένου στις τρεις διαστάσεις. Κοιτάζοντάς το από μια συγκεκριμένη οπτική γωνία βλέπετε ένα σχήμα που μοιάζει με το γράμμα “ήτα” σαν αυτό που φαίνεται στην Εικόνα 7.8 (πάνω αριστερά). Το ίδιο αντικείμενο ειδωμένο από μια άλλη γωνία και αφότου έχει περιστραφεί κατά 90 μοίρες προς τον κάθετο άξονά του φαίνεται να έχει το σχήμα που μοιάζει με τιμόνι στην Εικόνα 7.8 (κάτω αριστερά). Από μια τρίτη γωνία τέλος, στην Εικονά 7.8 (δεξιά), το αντικείμενο φαίνεται να είναι ένα μοντέλο διαστημικού σταθμού (στην πραγματικότητα, ένα μοντέλο που ο Stanley Kubrick χρησιμοποίησε στην ταινία “2001: A Space Odyssey” του 1968). Τι καταλαβαίνουμε από το παραπάνω νοητικό πείραμα; Σε γενικές γραμμές, ότι η πληροφορία που προσλαμβάνουμε εξαρτάται από την οπτική γωνία της παρατήρησής μας. Ανάλογα με το σημείο από το οποίο κοιτάζουμε ένα αντικείμενο μπορούμε να αποκομίσουμε περισσότερες ή λιγότερες στοιχεία για τη μορφή του.



Εικόνα 7.8: *PCA για το διαστημικό σταθμό του 2001 A Space Odyssey (Εικόνα από NASA Commons, CC0, πηγή: Flickr).*

Φανταστείτε τώρα ότι αντί για ένα αντικείμενο σε τρεις διαστάσεις καλούμαστε να διακρίνουμε τις μορφολογικές ιδιότητες ενός συνόλου από χιλιάδες σημεία σε περισσότερες διαστάσεις. Τα σημεία είναι τα γονίδια και οι διαστάσεις είναι ο αριθμός των συνθηκών στις οποίες έχουμε μετρήσει την έκφρασή του. Η ερώτηση που καλούμαστε να απαντήσουμε είναι:

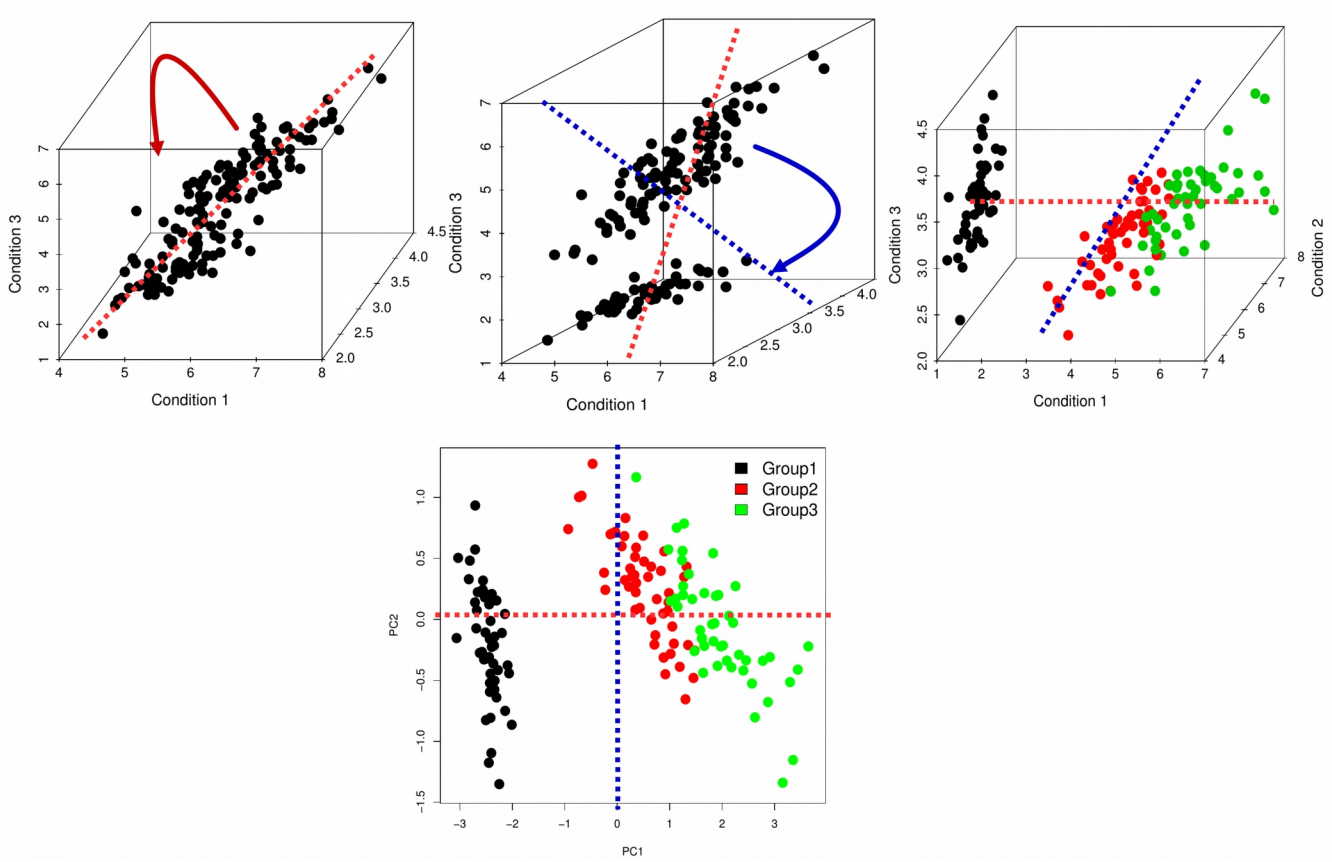
Σε ποιο βαθμό το "νέφος" των σημείων μπορεί να οργανωθεί σε υποσύνολα και πώς θα τα διακρίνουμε;

Η PCA μπορεί να δώσει την απάντηση και στα δύο αυτά ερωτήματα. Η μέθοδος βασίζεται ουσιαστικά σε μια σταδιακή μεγιστοποίηση της διασποράς των επιμέρους τιμών που χαρακτηρίζουν τα στοιχεία. Θα χρησιμοποιήσουμε ένα απλό παράδειγμα μ' ένα πείραμα έκφρασης 150 γονιδίων σε τρεις διαφορετικές συνθήκες. Κάθε γονίδιο μπορεί έτσι να αντιστοιχηθεί σ' ένα σημείο στο χώρο των τριών διαστάσεων που ορίζουν οι τρεις επιμέρους τιμές έκφρασης που έχουμε υπολογίσει γι' αυτό. Επιπλέον, μπορούμε να οπτικοποιήσουμε τα γονίδια αυτά σ' ένα "νέφος" τιμών σε τρεις διαστάσεις[7].

Η PCA προχωρά μετασχηματίζοντας τις διαστάσεις αυτές σε νέες διαστάσεις (που ονομάζονται κύριες συνιστώσες) με μοναδικό κριτήριο τη μεγιστοποίηση της διασποράς των τιμών σε καθεμιά από αυτές. Ο μετασχηματισμός αυτός αντιστοιχεί ουσιαστικά σε ορθογώνια προβολή των σημείων σ' έναν νέο άξονα που ορίζεται ως αυτός που μεγιστοποιεί τη διασπορά, με τον ίδιο

[7] *Οι τρεις διαστάσεις επιλέχθηκαν στο παράδειγμα για λόγους αναπαράστασης και μόνο. Η διαδικασία δεν αλλάζει για μεγαλύτερο αριθμό διαστάσεων.*

τρόπο που περιστρέφοντας το βλέμμα μας γύρω από το διαστημικό σταθμό της "Οδύσσειας" θα "παγώναμε" την περιστροφή στο σημείο εκείνο που η οπτική πληροφορία θα ήταν μεγαλύτερη.



Εικόνα 7.9: *Σχηματική αναπαράσταση της διαδικασίας της PCA σε ένα σύνολο τιμών έκφρασης 150 γονιδίων σε τρεις διαφορετικές συνθήκες. Ενώ αρχικά δεν φαίνεται να υπάρχουν διακριτά υποσύνολα, περιστροφή του "νέφους" γύρω από τους άξονες των κύριων συνιστωσών (κόκκινη και γαλάζια διακεκομμένη γραμμή) αναδεικνύει τρία διακριτά υποσύνολα.*

Η διαδικασία φαίνεται σχηματικά στην Εικόνα 7.9. Το δυσδιάκριτο νέφος προβάλλεται αρχικά ορθογώνια στην κόκκινη διακεκομμένη γραμμή που είναι η διάσταση με τη μεγαλύτερη διασπορά (7.9α). Στη συνέχεια, κρατώντας σταθερή τη διάσταση αυτή (που αποτελεί την πρώτη κύρια συνιστώσα) η PCA περιστρέφει το νέφος κάθετα ως προς αυτήν για να εντοπίσει την επόμενη γωνία στην οποία η διασπορά είναι τώρα η μέγιστη (η γαλάζια διακεκομμένη γραμμή, 7.9β). Θέτοντας τις δύο αυτές διαστάσεις ως το νέο σύστημα συντεταγμενων μπορεί κανείς να διακρίνει τρία υποσύνολα γονιδίων 7.9γ. Στην Εικόνα 7.9δ αναπαρίστανται τα 150 γονίδια σε δύο νέες διαστάσεις. Οι διαστάσεις αυτές δεν είναι οι αρχικές τιμές έκφρασής τους στις πρώτες συνθήκες του πειράματος αλλά οι σχετικές συντεταγμένες τους ως προς τις δύο πρώτες κύριες συνιστώσες PC1 και PC2 (κόκκινη και γαλάζια γραμμή). Βλέπουμε, με αυτόν τον τρόπο, πώς το αρχικό συγκεχυμένο νέφος σημείων, ειδωμένο μέσα από το "πρίσμα" των κύριων συνιστωσών του, αναδεικνύει τρία

διακριτά υποσύνολα (ομάδες).

Η παρατήρηση της Εικόνας 7.9 είναι ενδεικτική της λειτουργίας της μεθόδου. Βλέπουμε ότι το μεγαλύτερο μέρος της διακριτικής ικανότητας βασίζεται στη διάταξη των σημείων κατά τον οριζόντιο άξονα, την πρώτη δηλαδή κύρια συνιστώσα. Η διάταξη κατά τον κάθετο άξονα προσδίδει μια ελαφριά βελτίωση στη διάκριση των δύο ομάδων (2 και 3) που επικαλύπτονται. Αυτό είναι αποτέλεσμα του τρόπου με τον οποίον υπολογίζονται οι συνιστώσες. Η πρώτη κύρια συνιστώσα θα είναι πάντα αυτή με τη μεγαλύτερη διακριτική ικανότητα, η δεύτερη θα προσθέτει ένα κλάσμα αυτής της ικανότητας διάκρισης, η τρίτη ακόμα λιγότερο κ.ο.κ. Ο αριθμός των κύριων συνιστωσών που μπορούν να υπολογιστούν είναι θεωρητικά N-1 όπου N ο αριθμός των αρχικών διαστάσεων, αλλά πολύ συχνά αρκούν οι πρώτες 2-3 για να φτάσουμε σε ικανοποιητικά συμπεράσματα. Στο συγκεκριμένο παράδειγμα η τυπική απόκλιση που προέκυψε για την PC1 ήταν 1.9 και για την PC2 μόλις 0.49 (25% δηλαδή της PC1). Κατά τη διενέργεια μιας PCA ανάλυσης, ένα από τα χαρακτηριστικά που υπολογίζεται είναι η συνεισφορά, ή βάρος (loading) της κάθε κύριας συνιστώσας. Τα βάρη αυτά αντιστοιχούν ουσιαστικά στο βαθμό στον οποίο η κάθε συνιστώσα μπορεί να “εξηγήσει” τη διασπορά του δείγματος.

***Ερώτηση**: Δύο πειράματα αναλύονται με PCA. Για το πρώτο το 90% της συνολικής διασποράς αντιστοιχεί στα βάρη των δύο πρώτων κύριων συνιστωσών, ενώ για το δεύτερο το ποσοστό αυτό μοιράζεται μεταξύ των πρώτων τεσσάρων. Ποιο από τα δύο σύνολα τιμών πιστεύετε ότι είναι το λιγότερο πολύπλοκο σε ό,τι αφορά τη δομή του;*

Ως μέθοδος, η PCA αποτελεί μία από τις πρώτες επιλογές μας για μια διερευνητική ανάλυση των δεδομένων μας. Μεταξύ των πλεονεκτημάτων της, είναι η ταχύτητα και η σταθερότητα των υπολογισμών, ωστόσο παρότι μπορεί να παρουσιάσει εποπτικά την ποικιλομορφία των δεδομένων δεν μπορεί να αποδώσει ομάδες, καθώς στην ουσία δεν είναι καθαρή μέθοδος ομαδοποίησης. Στο παράδειγμα, η διάκριση των τριών ομάδων έγινε με βάση την εκ των προτέρων γνώση μας για το πείραμα. Προκειμένου όχι μόνο να διακρίνουμε την ύπαρξη διαφορετικών υποσυνόλων, αλλά και να διακρίνουμε πόσα είναι αυτά και από ποια στοιχεία απαρτίζονται, θα πρέπει να καταφύγουμε σε πιο εκλεπτυσμένες μεθόδους ομαδοποιήσης όπως αυτές που θα συζητήσουμε στη συνέχεια.

Ιεραρχική Ομαδοποίηση (Hierarchical Clustering)

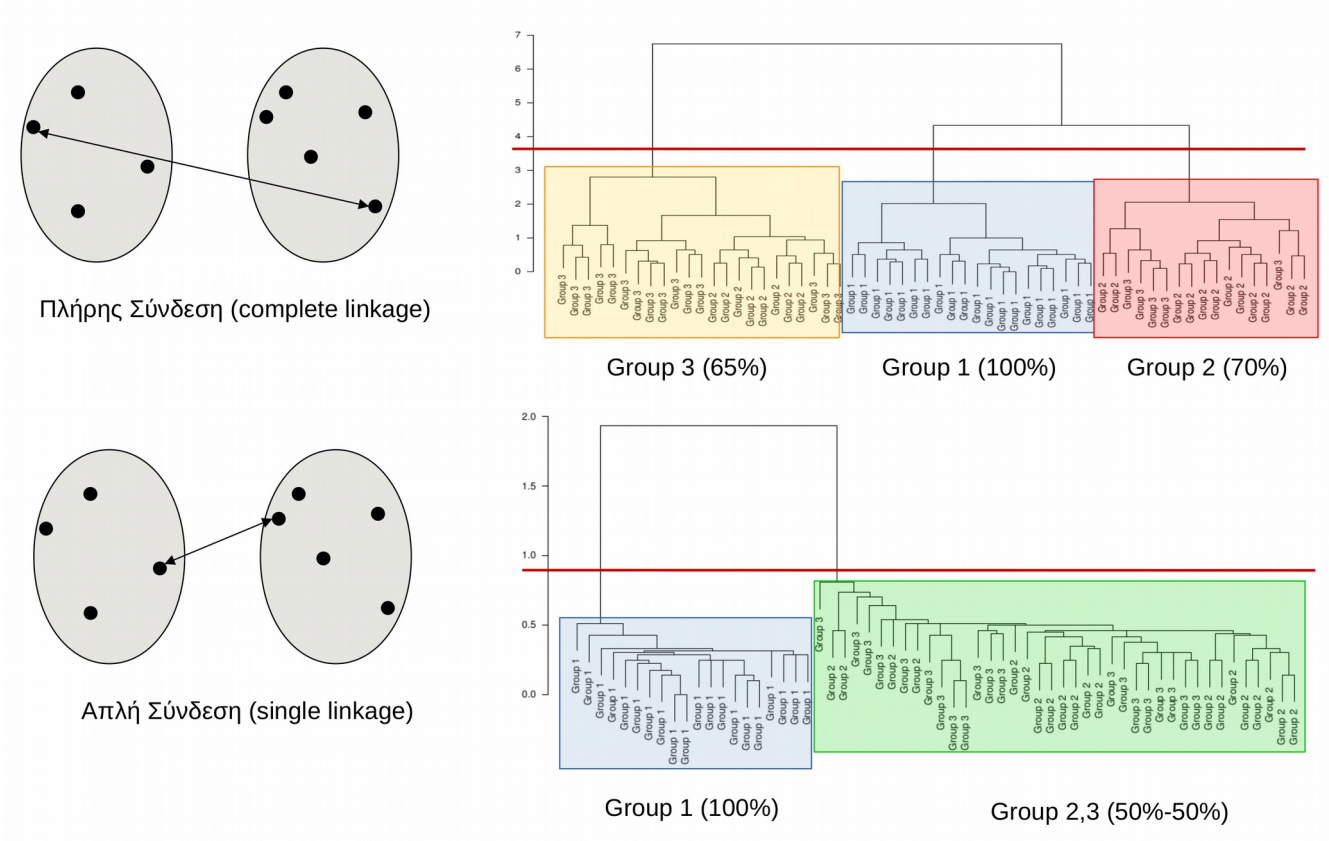
Η μέθοδος της ιεραρχικής ομαδοποίησης έχει πολλά κοινά στοιχεία με τις μεθόδους αποστάσεων που συζητήσαμε στο προηγούμενο κεφάλαιο για τη Φυλογενετική Ανάλυση. Η μέθοδος ουσιαστικά συνίσταται στη διαδοχική σύνδεση στοιχείων με βάση τη διαβαθμισμένη μεταξύ τους ομοιότητα, που έχει ποσοτικοποιηθεί με μέτρα απόστασης όπως αυτά που συζητήσαμε σε προηγούμενη ενότητα. Η διαδικασία έχει ως αποτέλεσμα ένα δέντρο αποστάσεων ανάλογο με αυτό που προκύπτει από την μέθοδο UPGMA κατά τη φυλογενετική ανάλυση αλληλουχιών. Σημείο εκκίνησης είναι ο υπολογισμός ενός Πίνακα Αποστάσεων όλων των στοιχείων μεταξύ τους, πάνω στον οποίο εφαρμόζεται στη συνέχεια ο παρακάτω αλγόριθμος:

Αλγόριθμος :: HClust
Δεδομένου ενός Πίνακα Αποστάσεων D[m^2]:
Εντόπισε από τα στοιχεία του D το στοιχείο i,j για το οποίο ισχύει D[i,j]=d_{min}
Ένωσε τα στοιχεία i,j σε ένα νέο στοιχείο ij.
Κατάγραψε την απόσταση d[i,j]
m=m-1
Υπολόγισε έναν νέο πίνακα D[m^2] υπολογίζοντας νέες αποστάσεις ως εξής:
    Για κάθε k=1 έως k=m
    D[k,ij]=D[ij,k]=func(D[i,k]+D[j,k])
Συνέχισε μέχρι m=1
Απόδωσε αποτέλεσμα: Το δέντρο που αντιστοιχεί στην ιεραρχική σειρά των ενώσεων
Τερματισμός

Όπως φαίνεται παραπάνω, ο αλγόριθμος μοιάζει πολύ με τον UPGMA. Η σειρά σύνδεσης των "φύλλων" του δέντρου γίνεται προοδευτικά με κριτήριο την ελάχιστη απόσταση και οι κλάδοι των φύλλων είναι ίσου μήκους. Η βασική διαφορά έχει να κάνει με τον τρόπο επανυπολογισμού των αποστάσεων, τον οποίον σκόπιμα συμβολίσαμε με *func* στον παραπάνω αλγόριθμο. Κι αυτό γιατί, σε αντίθεση με την UPGMA όπου οι νέες αποστάσεις υπολογίζονται πάντα στη βάση των μέσων τιμών των στοιχείων που απαρτίζουν την κάθε ομάδα, στην περίπτωση του ιεραχικού clustering μπορούμε να επιλέξουμε ανάμεσα σε διαφορετικούς τρόπους για τον υπολογισμό αυτό (Ward 1963). Έτσι, εκτός από την μέση απόσταση (που είναι πολύ κοντά στο ανάλογο της UPGMA) μπορούμε να υπολογίσουμε την απόσταση μεταξύ δύο ομάδων/υποσυνόλων ως την ελάχιστη απόσταση, των πλησιέστερων στοιχείων τους (και που ονομάζουμε απλή σύνδεση ή single linkage) ή αντίθετα ως τη μέγιστη απόσταση των πιο απομακρυσμένων από αυτά (που ονομάζουμε πλήρη σύνδεση ή complete linkage). Οι διαφορές που προκύπτουν μεταξύ των διαφορετικών μεθόδων υπολογισμού αποστάσεων δεν είναι αμελητέες, όπως φαίνεται και γραφικά στην Εικόνα 7.10. Η χρήση της απλής σύνδεσης δίνει γενικώς χειρότερα αποτελέσματα σε σύγκριση με τη χρήση της πλήρους ή της μέσης σύνδεσης κι αυτό γιατί συχνά μπορεί να ενώσει εσφαλμένα ομάδες που είναι πρακτικά μακριά η μία από την άλλη, απλώς και μόνο επειδή ένα ζεύγος από στοιχεία τυχαίνει να βρίσκονται κοντύτερα μεταξύ τους.

Στην Εικόνα 7.10 έχουμε χρησιμοποιήσει ιεραχική ομαδοποίηση στο ίδιο σύνολο 150 γονιδίων που αναλύσαμε στην περίπτωση της PCA. Τα 150 αυτά γονίδια κατανέμονται ιδανικά σε τρεις διαφορετικές ομάδες, με μια από αυτές να είναι αρκετά ξεκάθαρη (η "μαύρη ομάδα" στο παράδειγμα της Εικόνας 7.9) και τις άλλες δύο να έχουν συνολικά μικρότερες μεταξύ τους αποστάσεις. Στο πάνω μέρος της Εικόνας 7.10 βλέπουμε ότι χρήση της πλήρους σύνδεσης οδηγεί σε ομαδοποίηση που διακρίνει ικανονοποιητικά τις τρεις ομάδες, δημιουργώντας τρία cluster. Ένα από αυτά περιέχει αμιγώς όλα τα στοιχεία της ομάδας 1, ενώ τα άλλα δύο εμφανίζουν μια σημαντική υπερ-εκπροσώπηση στις ομάδες 2 και 3 αντίστοιχα, παρόλο που ο βαθμός συνεκτικότητας (το κατά πόσο δηλαδή το cluster περιέχει στοιχεία μόνο από μία ομάδα) δεν είναι το ίδιο υψηλός με αυτόν της ομάδας 1. Σε κάθε περίπτωση, η πλήρης σύνδεση λειτουργεί καλύτερα σε σχέση με την απλή σύνδεση (Εικόνα 7.10 κάτω μέρος), η οποία οδηγεί στο σχηματισμό δύο μόνο clusters, με το

δεύτερο να περιέχει τα στοιχεία τόσο της ομάδας 2 όσο και της ομάδας 3.



Εικόνα 7.10: *Ιεραρχική ομαδοποίηση για 60 από τα 150 γονίδια που αναλύθηκαν με PCA στην προηγούμενη ενότητα με 20 γονίδια να ανήκουν στο καθένα από τα 3 υποσύνολα. Επάνω: Υπολογισμός των αποστάσεων με πλήρη σύνδεση αποδίδει τρεις ομάδες με πολύ καλή συμφωνία με την (εκ των προτέρων γνωστή) αρχική ομαδοποίηση. Κάτω: Υπολογισμός των αποστάσεων με απλή σύνδεση οδηγεί στο σχηματισμό δύο ομάδων χωρίς να μπορεί να διακρίνει μεταξύ των Ομάδων 2 και 3.*

Η μέθοδος της ιεραρχικής ομαδοποίησης έχει κάποια σημαντικά πλεονεκτήματα σε σχέση με ανάλογες τεχνικές. Είναι ταχύτατη και αρκετά σταθερή (τα αποτελέσματα δεν αλλάζουν εφόσον οι παράμετροι που χρησιμοποιούνται δεν μεταβάλλονται), είναι ιδανική για να αναπαριστά ιεραρχικές σχέσεις και παρέχει ευελιξία σε ό,τι έχει να κάνει με τον καθορισμό των ομάδων. Καθώς δεν αποδίδει ομάδες αλλά ένα ιεραρχικό δέντρο, οι ομάδες δημιουργούνται με βάση ένα αυθαίρετο όριο απόστασης που ορίζεται από τον πειραματιστή. Στην Εικόνα 7.10 το όριο αυτό αντιστοιχεί στην κόκκινη γραμμή που διατρέχει εγκάρσια τους κλάδους του δέντρου και καθορίζει την αντιστοιχία φύλλων σε ομάδες.

Ερώτηση*: Μπορείτε να προσπαθήσετε να μετατοπίσετε το όριο απόστασης στο δέντρο της Εικόνας 7.10 για την απλή απόσταση με τρόπο που να οδηγεί σε καλύτερη διάκριση μεταξύ των ομάδων*

2 και 3; Ποιο θα ήταν το κόστος μιας τέτοιας προσπάθειας στη συνολική ομαδοποίηση;

Η ευελιξία αυτή μπορεί να είναι ωστόσο και μειονέκτημα της μεθόδου καθώς οδηγεί σ' έναν αυθαίρετο ορισμό του αριθμού των ομάδων χωρίς να παρέχει ένα αντικειμενικό κριτήριο για το ποιος θα πρέπει να είναι αυτός. Στη συνέχεια θα συζητήσουμε μια μέθοδο που παρέχει ένα τέτοιο κριτήριο για την επιλογή του αριθμού των ομάδων.

Ομαδοποίηση k-μέσων (k-means Clustering)

Η τελευταία μέθοδος ομαδοποίησης που θα συζητήσουμε σ' αυτό το κεφάλαιο είναι η μέθοδος των k-μέσων (k-means) (Hartigan and Wong 1979). Η k-means αποτελεί ένα πολύ καλό παράδειγμα μεθόδου βελτιστοποίησης κι από αυτήν την άποψη έχει κοινά χαρακτηριστικά με τη μέθοδο δειγματοληψίας Gibbs για τον εντοπισμό μοτίβων σε αλληλουχίες (Κεφάλαιο 3). Όπως και η δειγματοληψία Gibbs, ξεκινά από ένα αυθαίρετα ορισμένο σημείο εκκίνησης και προχωρά μέσω διαδοχικών επαναλήψεων συγκλίνοντας προς ένα βέλτιστο που ορίζεται είτε αντικειμενικά (θέτοντας ένα όριο σύγκλισης) είτε αριθμητικά (ορίζοντας το μέγιστο αριθμό επαναλήψεων). Σε αντίθεση με τις μεθόδους που έχουμε συναντήσει ως τώρα, για την k-means χρειάζεται να οριστεί εκ των προτέρων ο αριθμός των ομάδων που επιθυμούμε να σχηματίσουμε. Ο αριθμός αυτός, αντιστοιχεί στο k του ονόματος της μεθόδου και είναι το μόνο που χρειάζεται ως τιμή εισόδου για τον αλγόριθμο, καθώς η μέθοδος δεν είναι ιεραρχική και κατά συνέπεια δεν απαιτεί έναν πίνακα αποστάσεων.

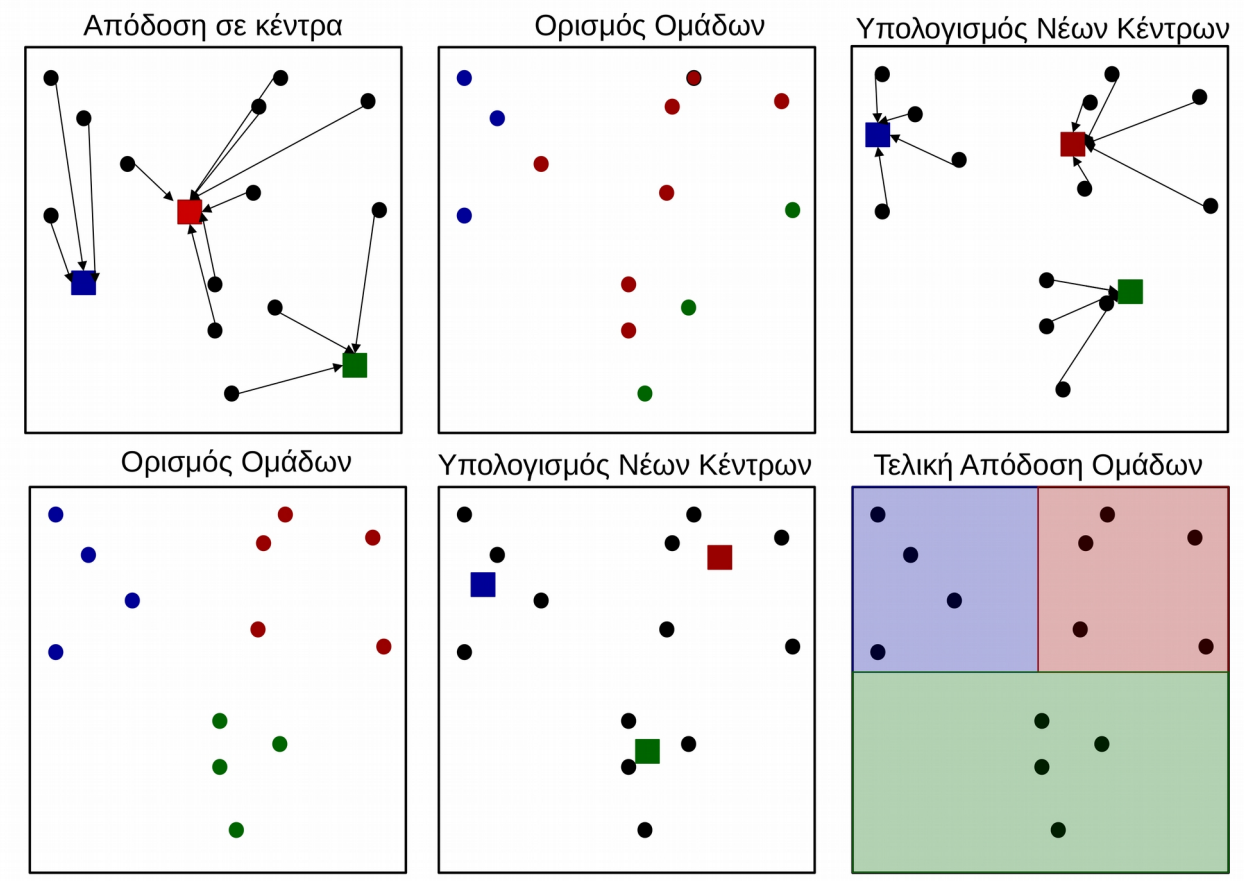
Ο κυρίως αλγόριθμος περιγράφεται σχηματικά στην Εικόνα 7.11, όπου δίνεται ένα απλό παράδειγμα σε χώρο δύο διαστάσεων. Η διαδικασία, ωστόσο, δεν αλλάζει σε περισσότερες διαστάσεις και ουσιαστικά αποτελεί τη διαδοχική επανάληψη τριών βημάτων. Σαν πρώτο βήμα, ορίζονται k σημεία στο χώρο που ορίζουν τα N στοιχεία των δεδομένων. Στο παράδειγμα για k=3 ορίζουμε τρία τέτοια σημεία που φαίνονται με τα χρωματιστά τετράγωνα. Η επαναληπτική διαδικασία ξεκινά απ' αυτό το σημείο και περιλαμβάνει τρία βήματα:

Βήμα 1: Κάθε στοιχείο αποδίδεται σ' ένα κέντρο με γνώμονα την ελάχιστη απόσταση. Στο τέλος αυτής της διαδικασίας, κάθε στοιχείο έχει συνδεθεί με ακριβώς ένα κέντρο.

Βήμα 2: Ορίζονται k ομάδες με βάση την προηγούμενη απόδοση των στοιχείων στα κέντρα, τα οποία στη συνέχεια απαλείφονται.

Βήμα 3: Με βάση τη σύσταση των k ομάδων υπολογίζονται νέα κέντρα, των οποίων οι συντεταγμένες ορίζονται με βάση τη μέση τιμή των συντεταγμένων των στοιχείων που απαρτίζουν την ομάδα.

Η διαδικασία επαναλαμβάνεται από το Βήμα 1 έως ότου η σύσταση των ομάδων να μη μεταβάλλεται περαιτέρω ή εναλλακτικά να ολοκληρωθεί ένας δεδομένος αριθμός επαναλήψεων που έχει τεθεί ως όριο εξαρχής. Παρόλο που για απλά και περιορισμένου μεγέθους σύνολα δεδομένων, η απόλυτη βελτιστοποίηση είναι εφικτή και ο αλγόριθμος συγκλίνει σχετικά γρήγορα, ο ορισμός ενός ορίου επαναλήψεων είναι απαραίτητος για δεδομένα μεγάλων διαστάσεων.



Εικόνα 7.11: *Σχηματική αναπαράσταση του αλγορίθμου της ομαδοποίησης k-μέσων.*

Η διαδικασία μπορεί να περιγραφεί και τυπικά στον παρακάτω αλγόριθμο:

Αλγόριθμος :: K-means
Δεδομένου ενός συνόλου M τιμών σε ν-διαστάσεις
Δεδομένου ενός αριθμού κέντρων k
Αρχικοποίηση: Όρισε τυχαία k σημεία σε ν-διάστατο χώρο
Επανάληψη:
1. Για κάθε στοιχείο m του M και για κάθε k υπολόγισε τις αποστάσεις d[m,k]
2. Υπολόγισε την ελάχιστη d_{min}[m,k] → Απόδωσε το m στο αντίστοιχο k
3. Δημιούργησε k ομάδες με βάση τις d_{min}
4. Όρισε νέα k ως τα κέντρα βάρους των ομάδων σε ν-διαστάσεις
5. Πήγαινε στο 1.

Τερματισμός Επανάληψης:
α) Αν η σύσταση ομάδων k δεν αλλάζει
β) Αν έχουν ολοκληρωθεί Α κύκλοι επανάληψεων
Απόδωσε αποτέλεσμα: Ομάδες k
Απόδωσε αποτέλεσμα: Συντεταγμένες κέντρων k

Απόδωσε αποτέλεσμα: Άθροισμα αποστάσεων των στοιχείων κάθε ομάδας από το κέντρο k
Τερματισμός

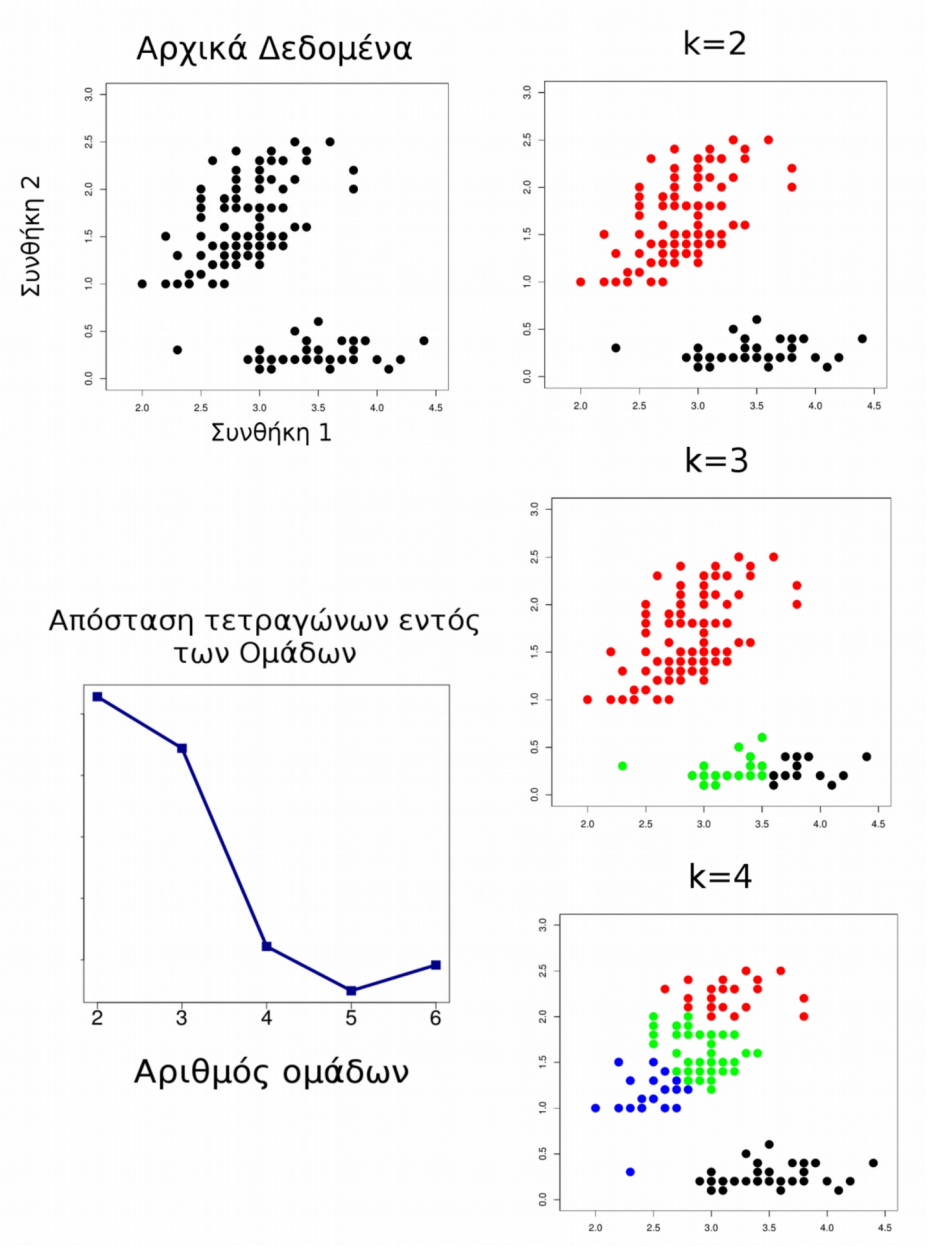
Ως τώρα είδαμε πώς μπορούμε να εφαρμόσουμε την k-means ομαδοποίηση για τη δημιουργία ομάδων με συγκεκριμένο αριθμό. Το ερώτημα όμως που παραμένει είναι πώς θα αποφασίσουμε ποιος θα είναι ο αριθμός των ομάδων που θα δημιουργήσουμε. Κάποιοι απλοί αυθαίρετοι κανόνες ορίζουν το βέλτιστο αριθμό k_{opt} με τον αριθμό N των στοιχείων που αναλύονται, με την πιο συχνά απαντώμενη σχέση να είναι $k_{opt}=\sqrt{N/2}$. Ωστόσο μια πιο αναλυτικά ασφαλής μέθοδος προσφέρει ένα αντικειμενικό κριτήριο για την επιλογή του βέλτιστου k. Το κριτήριο αυτό ορίζεται στη βάση της συμπεριφοράς μιας ποσότητας που σχετίζεται με το πόσο "συμπαγείς" είναι οι δημιουργούμενες ομάδες (Thorndike 1953). Η ποσότητα που χρησιμοποιούμε για να ποσοτικοποιήσουμε αυτό το βαθμό συνεκτικότητας, ισούται με το άθροισμα των τετραγώνων των αποστάσεων των στοιχείων από τα k κέντρα στα οποία έχουν αντιστοιχηθεί και ορίζεται ως εξής:

$$WSS=\sum_{i=1}^{N} D\left[i,k_i\right]^2 \quad 7.11$$

όπου k_i είναι το κέντρο στο οποίο αντιστοιχεί το στοιχείο *i* και *D* είναι η απόστασή τους στο χώρο των v-διαστάσεων. Το άθροισμα WSS (Within-Sum-of-Squares) μειώνεται καθώς αυξάνεται ο αριθμός των ομάδων k, κάτι που είναι αναμενόμενο, αφού όσο μεγαλώνει ο αριθμός των ομάδων, τόσο πιο μικρές αναμένεται να είναι οι αποστάσεις τους από τα στοιχεία που τους αποδίδονται. Η μείωση αυτή ωστόσο, δεν είναι γίνεται με σταθερό ρυθμό και ένας καλός τρόπος να επιλέξει κανείς σε ποιο σημείο θα σταματήσει να αυξάνει τον αριθμό των ομάδων είναι να παρατηρήσει τη γραφική παράσταση του WSS συναρτήσει του k. Καθώς από ένα σημείο και μετά, η προσθήκη νέων ομάδων απλώς θρυμματίζει περαιτέρω το σύνολο των στοιχείων σε ολοένα και μικρότερες ομάδες, το σημείο που η καμπύλη εμφανίζει την πιο απότομη μείωση είναι ενδεικτικό του βέλτιστου k. Στην Εικόνα 7.12 παρουσιάζεται γραφικά η ανάλυση με k-means των 150 γονιδίων του παραδείγματος που συζητήσαμε στην ενότητα της PCA. Η ανάλυση με 2, 3 και 4 ομάδες δίνει αρκετά διαφορετικά αποτελέσματα όμως με βάση το κριτήριο της μείωσης της τιμής WSS καταλήγουμε στο συμπέρασμα ότι η καλύτερη ομαδοποίηση επιτυγχάνεται για k=4. Για k=4 διατηρείται αυτούσιο το Group1 που συναντήσαμε στις περιπτώσεις των PCA και ιεραρχικής ομαδοποίησης, ενώ τα Group2, 3 διαχωρίζονται επιμέρους σε τρεις νέες ομάδες. Σημειώστε ότι με βάση τον απλό κανόνα του $k_{opt}=\sqrt{N/2}$ ο ιδανικός αριθμός ομάδων για τα 150 γονίδια θα ήταν μεταξύ 8 και 9, αριθμός αρκετά μεγάλος για να περιγράψει με τον καλύτερο τρόπο το δείγμα μας.

Σε γενικές γραμμές, όλες οι μέθοδοι ομαδοποίησης θα πρέπει να εφαρμόζονται με γνώμονα τη διατήρηση της ισορροπίας μεταξύ της όσο το δυνατόν καλύτερης περιγραφής του συστήματος, χωρίς όμως να χάνεται η δυνατότητα γενίκευσης των συμπερασμάτων. Κάθε ομαδοποίηση είναι στην ουσία δημιουργία ενός μοντέλου που σκοπό έχει να ερμηνεύσει τα δεδομένα με τρόπο που να είναι σταθερός και επαναλήψιμος. Από αυτήν την άποψη, η δημιουργία μεγάλου αριθμού ομάδων δεν είναι καλή πρακτική καθώς μπορεί μεν να εξηγεί επαρκώς τα συγκεκριμένα δεδομένα, πάνω στα οποία εφαρμόστηκε, αλλά όχι ένα εναλλακτικό σύνολο δεδομένων από το ίδιο ή παραπλήσιο σύστημα. Στο φαινόμενο αυτό, που ονομάζεται "υπερπροσαρμογή" (over-fitting) και είναι ένας

βασικός κίνδυνος για όλες τις μεθόδους μοντελοποίησης, θα επανέλθουμε και σε επόμενα κεφάλαια.



Εικόνα 7.12: *Ανάλυση των 150 γονιδίων του παραδείγματος της PCA για δύο συνθήκες με τη χρήση της ομαδοποίησης k-μέσων. Ομαδοποίηση με 2, 3 και 4 ομάδες αποδίδει διαφορετικά αποτελέσματα, αλλά το κριτήριο των αποστάσεων τετραγώνων εντός των ομάδων (WSS) συνηγορεί υπέρ της δημιουργίας τεσσάρων (4) ομάδων.*

Συμπεράσματα

Το κεφάλαιο αυτό αγγίζει μια σειρά από εξαιρετικά τεχνικά θέματα που άπτονται της ανάλυσης δεδομένων μεγάλης κλίμακας. Τόσο εδώ, όσο και σε αρκετά από τα επόμενα κεφάλαια, ο στόχος μας είναι κυρίως να παρουσιάσουμε τα προβλήματα και τις θεωρητικές προσεγγίσεις για τη λύση τους και όχι να παράσχουμε τεχνικές "συνταγές" για τη διενέργεια αναλύσεων. Ο λόγος δεν είναι ότι τέτοιου τύπου "πρωτόκολλα" βιοπληροφορικής ανάλυσης δεν είναι χρήσιμα, αλλά γιατί η παράθεσή τους χωρίς το αναγκαίο θεωρητικό υπόβαθρο δε θα είχε κανένα νόημα. Επιπλέον, μια σειρά από αναλύσεις που περιγράφηκαν σ' αυτό το κεφάλαιο, όπως οι διαδικασίες κανονικοποίησης, η εξαγωγή καταλόγων διαφορικά εκφραζόμενων γονιδίων αλλά και οι διάφορες τεχνικές ομαδοποίησης, μπορούν να διενεργηθούν με τη χρήση διαφορετικών υπολογιστικών εργαλείων, εφαρμογών που είναι διαθέσιμες στο διαδίκτυο ή ακόμα και προγραμμάτων που μπορούν να συνταχθούν από τον αναγνώστη (βλ. Ερωτήσεις/Ασκήσεις).

Στην καθημερινή πρακτική, τόσο η ανάλυση των πρωτογενών δεδομένων, όσο και η κανονικοποίησή τους διενεργείται, τις περισσότερες φορές, ως μέρος της τυπικής, προκαθορισμένης γραμμής επεξεργασίας των δεδομένων (data processing pipeline) από την τεχνική υπηρεσία που κάνει το πείραμα. Είναι ωστόσο σημαντικό να γνωρίζουμε το θεωρητικό υπόβαθρο και να μπορούμε ενδεχομένως να παρέμβουμε σε οποιοδήποτε στάδιο της ανάλυσης ακόμα κι αν αυτό θεωρείται απολύτως τυποποιημένο. Σε αντίθεση με την πρωτογενή ανάλυση, η διαδικασία της εξαγωγής των καταλόγων διαφορικά εκφραζόμενων γονιδίων (DEG) αποτελεί συχνότερα μέρος της ερευνητικής εργασίας ενός βιολόγου. Για το λόγο αυτό, ο τρόπος με τον οποίο αξιολογούμε ένα volcano plot, ή τα όρια σημαντικότητας και έντασης διαφορικής έκφρασης που θέτουμε για την αποκομιδή των DEG, έχουν ιδιαίτερη σημασία. Ακόμα σημαντικότερη είναι η έννοια του ελέγχου πολλαπλών υποθέσεων και η ανάγκη χρήσης διορθωμένων τιμών p-value, που συζητήσαμε μόνο επιγραμματικά σ' αυτό το κεφάλαιο και στις οποίες θα επανέλθουμε στο αμέσως επόμενο.

Το πεδίο της εφαρμογής μεθόδων ομαδοποίησης, τέλος, είναι αυτό που προσφέρει τη μεγαλύτερη ελευθερία για πρωτότυπες και οξυδερκείς αναλύσεις, που μπορούν να εξαγάγουν κρυμμένη πληροφορία από τα δεδομένα μας. Οι προσεγγίσεις μεγάλης κλίμακας όπως αυτές που συζητήθηκαν σ' αυτό το κεφάλαιο (και στις οποίες θα επανέλθουμε σ' όλα σχεδόν τα επόμενα) απαιτούν τον έξυπνο συνδυασμό υπολογιστικών τεχνικών που να μπορούν ν' αναδείξουν στοιχεία του συστήματος που δεν είναι προφανή σε πρώτο επίπεδο αλλά αναδύονται μέσα από γενικευμένες θεωρήσεις. Ο τρόπος με τον οποίο η ομαδοποίηση δεδομένων έκφρασης δημιουργεί υποσύνολα γονιδίων που εμφανίζουν κοινά πρότυπα έκφρασης είναι χαρακτηριστικό παράδειγμα τέτοιων προσεγγίσεων. Στο αμέσως επόμενο κεφάλαιο, θα δούμε πώς η πληροφορία που εξάγεται από πειράματα μεγάλης κλίμακας μπορεί να συνδυαστεί με προ-υπάρχουσα γνώση για τα βιολογικά συστήματα με τρόπο που να οδηγεί στην εξαγωγή ακόμα πιο συνεκτικών συμπερασμάτων.

Ερωτήσεις/Ασκήσεις

1. Ανάμεσα στις πιο χαρακτηριστικές πηγές ανωμαλιών (bias) σε μετρήσεις RNASeq είναι, όπως έχουμε ήδη δει, το μήκος του γονιδίου αλλά και το ποσοστό GC% (Risso et al. 2011). Μπορείτε να προτείνετε τρόπους για να ενσωματωθεί ο έλεγχος και για τις δύο αυτές ποσότητες σ' έναν τύπο κανονικοποίησης;

2. Ποιος θεωρείτε ότι είναι ο λόγος για τον οποίον οι τιμές φθορισμού σ' ένα πείραμα μικροσυστοιχίας κατανέμονται λογαριθμοκανονικά; Ποια είναι η φυσική σημασία της κατανομής στο πάνω μέρος της Εικόνας 7.3α;

3. Τι σημαίνει για ένα πείραμα ανάλυσης διαφορικής έκφρασης ένα πιο “ανοιχτό” διάγραμμα “κρατήρα ηφαιστείου” σε σχέση μ' ένα άλλο; Σε ποιο φυσικό φαινόμενο θα πρέπει να αποδοθεί ο πιο φαρδύς “κρατήρας”;

4. Στην ενότητα όπου συζητήσαμε την ιεραρχική ομαδοποίηση είδαμε πως έχει σημαντικές ομοιότητες με την UPGMA μέθοδο για τη δημιουργία φυλογενετικών δέντρων. Να εξηγήσετε γιατί δε θα ήταν καλή ιδέα να προσπαθήσουμε να εξάγουμε φυλογενετικές σχέσεις με τη χρήση της μεθόδου των κ-μέσων.

5. Κατά την υλοποίηση του αλγορίθμου των κ-μέσων που περιγράφεται στην αντίστοιχη ενότητα δεν εξασφαλίζεται ότι η τελική κατανομή των Ν στοιχείων θα γίνει όντως σε κ ομάδες. Αυτό συμβαίνει γιατί υπάρχει η πιθανότητα “σύντηξης” δύο (ή και περισσότερων) ομάδων κατά τη διαδικασία των επαναλήψεων. α) Κάτω από ποιες συνθήκες αυτό είναι πλεονέκτημα ή μειονέκτημα της μεθόδου. β) Να προτείνετε έναν εναλλακτικό αλγόριθμο που να εξασφαλίζει ότι θα υπάρξουν στο τέλος της διαδικασίας ακριβώς κ ομάδες.

Διαβάστε Περισσότερα

Για τις Μεθόδους Ανάλυσης Γονιδιακής έκφρασης:
Μια καλή εισαγωγή στις μικροσυστοιχίες DNA μπορεί να βρει κανείς στο *DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling (Baldi and Hatfield 2002)*. Σχετικά με την αλληλούχιση RNA, μια πρόσφατη επισκόπηση της βιβλιογραφίας δίνεται από τους (Wang, Gerstein, and Snyder 2009).

Για την κανονικοποίηση δεδομένων μικροσυστοιχειών:
Μια πλήρης περιγραφή της διαδικασίας δίνεται στο *Microarray Bioinformatics* (Stekel 2003) και κυρίως στα Κεφάλαια 2-5. Γενικότερα στοιχεία για τις μεθοδολογίες που περιγράφονται στις ενότητες της z-τυποποίησης και της κανονικοποίησης ποσοστημορίων μπορούν να βρεθούν σε εγχειρίδια βιοστατιστικής όπως το *Biostatistics* (Daniel and Cross 2012) ή το *Intuitive Biostatistics* (Motulsky 2010).

Για την πρωτογενή ανάλυση δεδομένων από πειράματα RNASeq:
Διάφορα άρθρα περιγράφουν τις διαδικασίες ανάλυσης των δεδομένων. Ξεχωρίζουν η εργασία των (Mortazavi et al. 2008), που είναι μια από τις πρώτες εφαρμογές της μεθοδολογίας, αλλά και οι εργασίες της ομάδας του Lior Pachter (Trapnell, Pachter, and Salzberg 2009; Trapnell et al. 2010) όπου περιγράφονται τόσο οι μεθοδολογίες κανονικοποίησης αλλά και ανάλυσης της διαφορικής γονιδιακής έκφρασης.

Για τις μεθόδους Ομαδοποίησης γενικά:
Μια κλασσική εισαγωγή αρκετά "φιλική" για φοιτητές θετικών επιστημών είναι το *Cluster Analysis* (Everitt et al. 2011).

Για την Ανάλυση Κύριων Συνιστωσών (PCA):
Τα περισσότερα εγχειρίδια που αναφέρονται σε τεχνικές μηχανικής μάθησης θα περιέχουν αναπόφευκτα μια ενότητα αφιερωμένη στην PCA. Ο αναγνώστης που επιθυμεί να εμβαθύνει μπορεί να ανατρέξει στο (Witten and Frank 2005). Ωστόσο, μια πιο σύντομη και αρκετά εύληπτη περιγραφή της μεθοδολογίας (και του μαθηματικού υποβάθρου) μπορεί να βρει κανείς στο εκπαιδευτικό άρθρο επισκόπησης των (Abdi and Williams 2010).

Για την Ιεραρχική Ομαδοποίηση:
Στο Κεφάλαιο 4 του *Cluster Analysis* (Everitt et al. 2011) περιέχει μια εξαιρετική περιγραφή του αλγορίθμου αλλά και λεπτομέρειες για την εφαρμογή του.

Για την Ομαδοποίηση κ-μέσων:
Η κλασσική αναφορά είναι αυτή που παρατίθεται στο κείμενο (Hartigan and Wong 1979). Μια πιο προσιτή, από τεχνική άποψη περιγραφή, προσαρμοσμένη στο πρόβλημα της γονιδιακής έκφρασης,

δίνεται στο Κεφάλαιο 8 του *Computational Genome Analysis: An Introduction* (Deonier, Tavaré, and Waterman 2007)

Βιβλιογραφία

Abdi, Hervé, and Lynne J. Williams. 2010. “Principal Component Analysis.” *Wiley Interdisciplinary Reviews: Computational Statistics* 2 (4): 433–59. doi:10.1002/wics.101.

Baldi, Pierre, and G. Wesley Hatfield. 2002. *DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling*. Cambridge University Press.

Ballman, K. V., D. E. Grill, A. L. Oberg, and T. M. Therneau. 2004. “Faster Cyclic Loess: Normalizing RNA Arrays via Linear Models.” *Bioinformatics* 20 (16): 2778–86. doi:10.1093/bioinformatics/bth327.

Bolstad, B.M., R.A Irizarry, M. Astrand, and T.P. Speed. 2003. “A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias.” *Bioinformatics* 19 (2): 185–93. doi:10.1093/bioinformatics/19.2.185.

Daniel, Wayne W., and Chad L. Cross. 2012. *Biostatistics: A Foundation for Analysis in the Health Sciences, 10th Edition: A Foundation for Analysis in the Health Sciences*. Wiley Global Education.

Deonier, Richard C., Simon Tavaré, and Michael Waterman. 2007. *Computational Genome Analysis: An Introduction*. Springer Science & Business Media.

Dudoit, Sandrine, Juliet Popper Shaffer, and Jennifer C. Boldrick. 2003. “Multiple Hypothesis Testing in Microarray Experiments.” *Statistical Science* 18 (1). Institute of Mathematical Statistics: 71–103.

Duggan, David J, Michael Bittner, Yidong Chen, Paul Meltzer, and Jeffrey M. Trent. 1999. “Expression Profiling Using cDNA Microarrays.” *Nat Genet* 21: 10–14. doi:10.1038/4434.

Everitt, Brian S., Sabine Landau, Morven Leese, and Daniel Stahl. 2011. *Cluster Analysis*. John Wiley & Sons.

Hansen, Kasper D, Rafael A Irizarry, and Zhijin Wu. 2012. “Removing Technical Variability in RNA-Seq Data Using Conditional Quantile Normalization.” *Biostatistics (Oxford, England)* 13 (2): 204–16. doi:10.1093/biostatistics/kxr054.

Hartigan, J. A., and M. A. Wong. 1979. “Algorithm AS 136: A K-Means Clustering Algorithm.” *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28 (1): 100–108. doi:10.2307/2346830.

Irizarry, Rafael A, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. 2003. “Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data.” *Biostatistics (Oxford, England)* 4 (2): 249–64.

doi:10.1093/biostatistics/4.2.249.

Kapranov, Philipp, Jorg Drenkow, Jill Cheng, Jeffrey Long, Gregg Helt, Sujit Dike, and Thomas R Gingeras. 2005. "Examples of the Complex Architecture of the Human Transcriptome Revealed by RACE and High-Density Tiling Arrays." *Genome Research* 15 (7): 987–97. doi:10.1101/gr.3455305.

Lander, E S, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, et al. 2001. "Initial Sequencing and Analysis of the Human Genome." *Nature* 409 (6822). Macmillian Magazines Ltd.: 860–921. doi:10.1038/35057062.

Lee, M.-L. T., F. C. Kuo, G. A. Whitmore, and J. Sklar. 2000. "Importance of Replication in Microarray Gene Expression Studies: Statistical Methods and Evidence from Repetitive cDNA Hybridizations." *Proceedings of the National Academy of Sciences* 97 (18): 9834–39. doi:10.1073/pnas.97.18.9834.

Lukk, Margus, Misha Kapushesky, Janne Nikkilä, Helen Parkinson, Angela Goncalves, Wolfgang Huber, Esko Ukkonen, and Alvis Brazma. 2010. "A Global Map of Human Gene Expression." *Nature Biotechnology* 28 (4): 322–24. doi:10.1038/nbt0410-322.

Maier, Tobias, Marc Güell, and Luis Serrano. 2009. "Correlation of mRNA and Protein in Complex Biological Samples." *FEBS Letters* 583 (24): 3966–73. doi:10.1016/j.febslet.2009.10.036.

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq." *Nature Methods* 5 (7). Nature Publishing Group: 621–28. doi:10.1038/nmeth.1226.

Motulsky, Harvey. 2010. *Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking*. Oxford University Press.

Risso, Davide, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. 2011. "GC-Content Normalization for RNA-Seq Data." *BMC Bioinformatics* 12 (1): 480. doi:10.1186/1471-2105-12-480.

Stekel, Dov. 2003. *Microarray Bioinformatics*. Cambridge University Press.

Thorndike, Robert L. 1953. "Who Belongs in the Family?" *Psychometrika* 18 (4): 267–76. doi:10.1007/BF02289263.

Trapnell, Cole, Lior Pachter, and Steven L Salzberg. 2009. "TopHat: Discovering Splice Junctions with RNA-Seq." *Bioinformatics (Oxford, England)* 25 (9): 1105–11. doi:10.1093/bioinformatics/btp120.

Trapnell, Cole, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. 2010. “Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation.” *Nature Biotechnology* 28 (5). Nature Publishing Group: 511–15. doi:10.1038/nbt.1621.

Vogel, Christine, and Edward M Marcotte. 2012. “Insights into the Regulation of Protein Abundance from Proteomic and Transcriptomic Analyses.” *Nature Reviews. Genetics* 13 (4). Nature Publishing Group: 227–32. doi:10.1038/nrg3185.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. “RNA-Seq: A Revolutionary Tool for Transcriptomics.” *Nature Reviews. Genetics* 10 (1). Nature Publishing Group: 57–63. doi:10.1038/nrg2484.

Ward, J H. 1963. “Hierarchical Grouping to Optimize an Objective Function.” *Journal of the American Statistical Association* 58 (301). Taylor & Francis Group: 236–44. doi:10.1080/01621459.1963.10500845.

Wilkinson, Leland, and Michael Friendly. 2012. “The History of the Cluster Heat Map.” *The American Statistician*, January. Taylor & Francis.

Witten, Ian H., and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann.