ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΠΕΡΙΦΕΡΕΙΑ ΗΠΕΙΡΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ Ε.Π.
ΠΕΡΙΦΕΡΕΙΑΣ ΗΠΕΙΡΟΥ

Έργο : «Βιοπληροφορική Ανάλυση Μεγάλου Όγκου Γενωμικών Δεδομένων Ασθενών με Όγκο Εγκεφάλου και Ανάπτυξη Εξελιγμένων Αλγορίθμων για την Εύρεση Βιοδεικτών και την Πρόβλεψη της Πιθανότητας Κυτταρικών Μεταλλάξεων – BIOPREDICTOR»

Αρ. Σύμβασης : ΗΠ1ΑΒ-00128

Παραδοτέο 1.2: «Τρέχουσα τεχνολογία σε αλγόριθμους για την επεξεργασία βιολογικών δεδομένων μεγάλου όγκου»

Ιωάννινα, 31/03/2019

ΠΑΚΕΤΟ ΕΡΓΑΣΙΑΣ 1

Παραδοτέο 1.2

Τίτλος Παραδοτέου 1.2.: Τρέχουσα τεχνολογία σε αλγόριθμους για την επεξεργασία βιολογικών δεδομένων μεγάλου όγκου.

Είδος Παραδοτέου: Έκθεση

Πακέτο εργασίας: Ανασκόπηση της πρόσφατης βιβλιογραφίας

Υπεύθυνος: COMITECH

Παράδοση: 31 Μαρτίου 2019

ΛΙΣΤΑ ΑΤΟΜΩΝ ΠΟΥ ΣΥΝΕΙΣΕΦΕΡΑΝ

- ΓΛΑΡΟΣ ΚΩΝΣΤΑΝΤΙΝΟΣ-Comitech ΑΕ
- ΠΑΠΑΝΙΚΟΛΑΟΥ ΚΩΝΣΤΑΝΤΙΝΟΣ- Comitech ΑΕ
- ΣΑΒΒΟΣ ΕΥΑΓΓΕΛΟΣ- Comitech ΑΕ
- ΠΟΜΩΝΗΣ ΙΩΑΝΝΗΣ- Comitech ΑΕ

Περιεχόμενα

Παραδοτέο 1.2.: Τρέχουσα τεχνολογία σε αλγόριθμους για την επεξεργασία βιολογικών δεδομένων μεγάλου όγκου..... 4

1.2.1 Μέθοδοι επεξεργασίας γενετικών δεδομένων..... 4

1.2.2 Προγνωστικοί μηχανισμοί στη γενετική 20

1.2.3. Επεξεργασία δεδομένων γλοιοβλαστώματος..... 38

Παραδοτέο 1.2.: Τρέχουσα τεχνολογία σε αλγόριθμους για την επεξεργασία βιολογικών δεδομένων μεγάλου όγκου

1.2.1 Μέθοδοι επεξεργασίας γενετικών δεδομένων

Η εποχή της γενωμικής και των μεγάλων δεδομένων έχει φέρει την ανάγκη συνεργασίας και ανταλλαγής δεδομένων, προκειμένου να αξιοποιηθεί αποτελεσματικά αυτή η νέα γνώση. Με την ανάπτυξη τεχνολογιών προσδιορισμού αλληλουχίας υψηλής απόδοσης, τα τελευταία χρόνια συσσωρεύτηκαν αποδείξεις ότι ο καρκίνος είναι μια ασθένεια του γονιδιώματος και παρατηρήθηκε μεγάλη έκρηξη δεδομένων και συστηματική μελέτη του γονιδιώματος του καρκίνου. Για πρώτη φορά, διατίθενται δεδομένα για τις πλήρεις αλληλουχίες γονιδιώματος που περιλαμβάνουν σημειακές μεταλλάξεις και δομικές εναλλαγές για μεγάλο αριθμό τύπων καρκίνου, επιτρέποντας τη διαφοροποίηση υποτύπων καρκίνου. Παρακάτω περιγράφονται τα διαδικτυακά μέσα για την έρευνα και τη γενωμική του καρκίνου που περιλαμβάνουν τα αρχεία καταγραφής δεδομένων και τα εργαλεία ανάλυσης και αξιολογούνται με βάση την ποικιλία των τύπων καρκίνου, το μέγεθος του δείγματος, την πληρότητα των δεδομένων και την εμπειρία των χρηστών.

European Genome-phenome Archive

Η European Genome-phenome Archive (EGA) είναι ένα κέντρο δεδομένων για όλους τους τύπους πειραμάτων προσδιορισμού αλληλουχίας και γονότυπου. Σχεδόν το 58% όλων των μελετών στην EGA σχετίζεται με τον καρκίνο, συμπεριλαμβανομένων των δεδομένων που παράγονται από το (International Cancer Genome Foundation, ICGC). Από την ίδρυσή της το 2008, η ICGC έχει παραγάγει terabytes δεδομένων από δωρητές και πάνω από 50 προγράμματα καρκίνου (https://dcc.icgc.org/repository/release_17). Τα δεδομένα σωματικών παραλλαγών (somatic variant data) είναι ανοιχτά προσβάσιμα στην ICGC Data Portal (https://dcc.icgc.org/repository), ενώ τα δεδομένα ακατέργαστης ακολουθίας, οι γενετικές μεταλλάξεις και τα κλινικά δεδομένα διατηρούνται στην EGA με ελεγχόμενη πρόσβαση.

Catalog Of Somatic Mutations In Cancer

Η Catalog Of Somatic Mutations In Cancer (COSMIC) είναι η μεγαλύτερη βάση δεδομένων για σωματικές μεταλλάξεις και τις επιπτώσεις τους στον καρκίνο του ανθρώπου (Forbes SA *et al*., 2015). Η βάση δεδομένων επιμελείται ατομικά από τη δημοσιευμένη βιβλιογραφία και επιτρέπει πολύ ακριβείς ορισμούς των τύπων ασθενειών και των λεπτομερειών του ασθενούς. Η βάση δεδομένων ενημερώνεται κάθε 2 μήνες και τα δεδομένα μπορούν να αναζητηθούν με λέξεις-κλειδιά και να ληφθούν από εγγεγραμμένους χρήστες. Το τεράστιο,

τακτικά ενημερωμένο σύνολο δεδομένων της COSMIC την καθιστά ανεκτίμητο πόρο για μελέτες για τον καρκίνο.

Cancer Program Resource Gateway

Το Broad Institute είναι ένα από τα πιο διάσημα ακαδημαϊκά κέντρα για τη μελέτη του καρκίνου. Το πρόγραμμά του για τον καρκίνο στοχεύει στη διερεύνηση των βασικών μηχανισμών του καρκίνου και της έρευνας από την ανακάλυψη έως τις κλινικές εφαρμογές. Το Πρόγραμμα δίνει πολλά σύνολα δεδομένων και εργαλεία για την επιστημονική έρευνα, τα οποία είναι διαθέσιμα στην Cancer Program Resource Gateway (CPRG). Ένα από τα πιο σημαντικά μέλη του CPRG είναι το Broad's Genome Data Analysis Center (GDAC) που περιγράφεται παρακάτω.

GDAC του Broad

Είναι σημαντικό, αλλά γενικά χρονοβόρο ή μερικές φορές ακόμη και αδύνατο για τα περισσότερα εργαστήρια να αναλύουν terabytes δεδομένων αλληλουχίας. Ωστόσο, το σύστημα Firehose από το GDAC της Broad αλλάζει αυτή την κατάσταση. Το GDAC αναλύει συστηματικά και επεξεργάζεται δεδομένα από την TCGA (The Cancer Genome Atlas) κάθε 2 εβδομάδες και τα καθιστά διαθέσιμα στη συνέχεια (Marx V *et al*., 2013). Το Firehose περιέχει σειρά τυποποιημένων αγωγών για γονιδιωματική ανάλυση και το περιβάλλον πληροφορικής είναι προσιτό στο κοινό έτσι ώστε οι χρήστες να μπορούν να εγκαταστήσουν και να χρησιμοποιήσουν τα δικά τους εργαλεία για την ανάλυση δεδομένων. Αξιοποιώντας το ισχυρό υπολογιστικό περιβάλλον στο Broad, το Firehose παρέχει συνεχώς ενημερωμένα δεδομένα και αποτελέσματα σε διαφορετικές βαθμίδες, συμπεριλαμβανομένων αποτελεσμάτων ανάλυσης και φιλικών προς τους βιολόγους αναφορών.

SNP500Cancer

Η βάση δεδομένων SNP500Cancer, Cancer Genome Anatomy Project (CGAP) (http://cgap.nci.nih.gov/Tools), αποτελεί τη βάση που εναποτίθενται οι επαληθεύσεις γονοτύπου και αλληλουχίας των απλών νουκλεοτιδικών πολυμορφισμών (Single Nucleotide Polymorphisms, SNPs) στον καρκίνο και σε άλλες πολύπλοκες ασθένειες (Packer BR *et al*., 2006). Ο κύριος στόχος του έργου SNP500Cancer είναι η επανάληψη της εύρεσης ακολουθίας δειγμάτων αναφοράς για την εύρεση γνωστών ή νέων SNPs για μελέτες μοριακής επιδημιολογίας στον καρκίνο. Η βάση δεδομένων παρέχει πληροφορίες αλληλουχίας για δείγματα ανώνυμου DNA και μπορεί να αναζητηθεί από το γονίδιο, τη γονιδιακή οντολογία, το χρωμόσωμα, την βάση δεδομένων SNP Database (dbSNP), ή την SNP500Cancer SNP ID. Η SNP500Cancer είναι ένα μέσο για τους ερευνητές να επιλέγουν SNPs για περαιτέρω ανάλυση, παρόλα αυτά ο όγκος των δεδομένων τους είναι περιορισμένος.

canEvolve

Με την ταχεία ανάπτυξη των βιολογικών τεχνολογιών, η μελέτη του προφίλ του όγκου σε ολόκληρο το γονιδίωμα έχει αυξηθεί δραστικά σε κλίμακα και διαθεσιμότητα. Το canEvolve πληροί την ανάγκη για ενσωμάτωση και ερμηνεία δεδομένων και περιέχει ολοκληρωμένα δεδομένα από μελέτες στις οποίες συμμετείχαν περισσότεροι από 10.000 ασθενείς. Όλα τα δεδομένα μπορούν να αναζητηθούν και να απεικονιστούν σε μορφή πίνακα ή γραφήματος. Το canEvolve είναι μια ολοκληρωμένη λειτουργική πλατφόρμα γενωμικής και είναι πλήρως προσβάσιμη στο κοινό (Samur MK *et al*., 2013).

MethyCancer

Η μεθυλίωση του DNA παίζει σημαντικό ρόλο στην ανάπτυξη του καρκίνου και σχετίζεται με την ενεργοποίηση των ογκογονιδίων, την αστάθεια των χρωμοσωμάτων και τη σίγαση του γονιδίου καταστολής του όγκου. Το MethyCancer έχει σχεδιαστεί για να ερμηνεύει τη σχέση μεταξύ της μεθυλίωσης DNA, της γονιδιακής έκφρασης και του καρκίνου. Το MethyCancer φιλοξενεί στοιχεία για (1) τη μεθυλίωση του DNA, (2) πληροφορίες σχετικά με το CNV και τον καρκίνο, (3) κλώνους νήσων CpG και (4) τους συσχετισμούς μεταξύ αυτών των συνόλων δεδομένων. Η βάση δεδομένων μπορεί εύκολα να αναζητηθεί με την φιλική προς το χρήστη οθόνη MethyView (He X *et al*., 2008).

SomamiR

Η παραλλαγή της αλληλουχίας του miRNA συμβάλλει σε μια ποικιλία καρκίνων. Το SomamiR είναι μια βάση δεδομένων που δημιουργήθηκε για να διερευνήσει τη συσχέτιση σωματικών και γενετικών μεταλλάξεων με τη λειτουργία του miRNA στον καρκίνο. Το SomamiR περιέχει επίσης ένα σύνολο πειραματικά επικυρωμένων σωματικών και γενετικών μεταλλάξεων που διακόπτουν τη λειτουργία miRNA που σχετίζεται με τον καρκίνο. Η ενσωμάτωση των δεδομένων μετάλλαξης στο SomamiR θα επιτρέψει στους επιστήμονες να προβλέψουν με μεγαλύτερη ακρίβεια εάν οι μεταλλάξεις θα επηρεάσουν τη δέσμευση του miRNA και κατά συνέπεια τη λειτουργική ρύθμιση (Bhattacharya A *et al*., 2013).

cBioPortal

Το cBioPortal για τη γενωμική του καρκίνου είναι μια προσβάσιμη πύλη που επιτρέπει στους ερευνητές να διερευνήσουν, να απεικονίσουν και να αναλύσουν τα πολυδιάστατα δεδομένα γενωμικής του καρκίνου (Gao J *et al*., 2013). Η πύλη περιέχει σύνολα δεδομένων από πολλές δημοσιευμένες μελέτες καρκίνου και επεξεργάζεται πρωτότυπα δεδομένα μοριακού προφίλ από καρκινικούς ιστούς και κυτταρικές σειρές σε μικρότερα σύνολα δεδομένων. Οι

ερευνητές μπορούν να διερευνήσουν με διαδραστικό τρόπο τα πρότυπα των γενετικών αλλαγών σε όλα τα δείγματα σε μια μεμονωμένη μελέτη, να συγκρίνουν τις συχνότητες αλλαγής γονιδίων σε πολλαπλές μελέτες ή να συνοψίσουν όλες τις σχετικές γονιδιωματικές παραλλαγές σε ένα μεμονωμένο δείγμα όγκου μέσω της διεπαφής ερωτημάτων ιστού cBioPortal. Υποστηρίζει επίσης τη διερεύνηση της βιολογικής οδού, την ανάλυση επιβίωσης (survival analysis) και την υπηρεσία λήψης δεδομένων.

UCSC Καρκίνος Genomics Browser

Το UCSC Cancer Genomics Browser είναι ένα online εργαλείο ανάλυσης για τη φιλοξενία, την οπτικοποίηση και την ανάλυση πληροφοριών σχετικά με τη γενωμική του καρκίνου και την κλινική έρευνα (Goldman M *et al*., 2013). Ο περιηγητής παρέχει στους χρήστες μετρήσεις από ένα μόνο πείραμα και τις σχετικές κλινικές πληροφορίες για πολλαπλά δείγματα. Επιπλέον, μπορούν να προβληθούν ταυτόχρονα δύο ή περισσότερα σύνολα δεδομένων για να επιτρέψουν συγκρίσεις της γονιδιακής έκφρασης σε διαφορετικά δεδομένα και τύπους καρκίνου.

Cancer Genome Work Bench

Διευκολύνει τη συστηματική διερεύνηση των γενετικών αλλαγών στους όγκους. Το CGWB αποτελείται από τρία συστατικά μέρη: (1) ένα αυτοματοποιημένο αγωγό ανάλυσης που ανιχνεύει και καταγράφει τις αλλαγές υποκατάστασης και εισαγωγής/διαγραφής (2) ένα σύστημα διαχείρισης βάσεων δεδομένων για τη διαχείριση έργου, δείγματος, αρχείου παρακολούθησης αλληλουχίας και δεδομένων γενετικής παραλλαγής (3) ένα εργαλείο εφαρμογής, Cancer Genome Browser προσαρμοσμένο για την ολοκληρωμένη ανάλυση παραλλαγών του όγκου. Το λογισμικό μπορεί να είναι χρήσιμο τόσο για τη λεπτομερή ανάλυση των προφίλ μεταλλάξεων σε ένα μεμονωμένο ερευνητικό εργαστήριο όσο και για την ανάλυση μεγάλης κλίμακας δεδομένων υψηλής απόδοσης (Zhang J *et al*., 2007).

Genomics of Drug Sensitivity in Cancer (GDSC)

Η (GDSC) είναι η μεγαλύτερη ανοικτή βάση γνώσεων σχετικά με την ευαισθησία και την ανταπόκριση των φαρμάκων στα καρκινικά κύτταρα (Yang W *et al*., 2013). Περιγράφει την ανταπόκριση σε σχεδόν 200 αντικαρκινικά θεραπευτικά μέσα σε περισσότερες από 1000 κυτταρικές γραμμές καρκίνου. Στην ιστοσελίδα του GDSC, η ευαισθησία του φαρμάκου μπορεί να διερευνηθεί από το γονίδιο του καρκίνου, την κυτταρική γραμμή του καρκίνου και την ένωση. Τα ανακτηθέντα δεδομένα παρουσιάζονται σε μια ποικιλία γραφικών παραστάσεων για προβολή και λήψη. Η μεγάλη συλλογή κυτταρικών σειρών, η ευαισθησία φαρμάκων και τα σύνολα δεδομένων γονιδιώματος έχουν ενισχύσει την κατανόηση της γονιδιωματικής ετερογένειας των καρκινικών κυττάρων και διευκόλυναν την ανακάλυψη νέων ειδικών για τον ασθενή θεραπειών.

canSAR

Το canSAR είναι μια ανοιχτή βάση διεπιστημονικών γνώσεων με επίκεντρο τον καρκίνο που (Bulusu KC *et al*., 2014). Οι ερευνητές μπορούν να αποκτήσουν πληροφορίες σχετικά με τις τρέχουσες πληροφορίες για την πρωτεΐνη ή το φάρμακο, όπως η έκφραση ή η μετάλλαξη της πρωτεΐνης στον καρκίνο, τα προφίλ κυτταρικής ευαισθησίας και οι ειδικές πρωτεΐνες πρόσδεσης του φαρμάκου. Μια μεγάλη συλλογή ανθρώπινων πρωτεομικών δεδομένων στεγάζεται στο canSAR αυτή τη στιγμή καθώς επίσης και περιλήψεις δομών 3D πρωτεϊνών.

NONCODE

Τα μη κωδικοποιημένα RNAs (ncRNAs) λειτουργούν σε ποικίλους τύπους καρκίνου. Το NONCODE είναι μια – τακτικής ενημέρωσης – βάση δεδομένων ncRNA (εκτός από tRNAs και rRNAs) από διάφορα είδη όπως ο άνθρωπος και ο ποντικός (Liu C *et al*., 2005). Περισσότερο από το 80% των δεδομένων ncRNA στο NONCODE παράγονται πειραματικά, παρέχοντας έτσι στους χρήστες έναν εξαιρετικά αξιόπιστο πόρο. Εκτός από την παροχή προφίλ έκφρασης του lncRNA σε όλους τους ιστούς, το NONCODE παρέχει διαδικτυακά (online) έναν αγωγό που ονομάζεται iLncRNA που βοηθάει του χρήστες να αναλύουν προσαρμοσμένα δεδομένα RNA-seq. Έχει επίσης μια λειτουργία μετατροπής που RefSeq ή Ensembl IDs σε NONCODE IDs (Fang S *et al*., 2017).

Οι παραπάνω διαδικτυακές πλατφόρμες (Πίνακας 1), μπορούν να χωριστούν σε πέντε κατηγορίες. Πρώτον, πλατφόρμες όπως το EGA, το COSMIC και το cBioPortal χρησιμεύουν ως εγκυκλοπαίδεια ειδών καρκίνου και τύπων δεδομένων omics. Άλλες πλατφόρμες προσφέρουν εργαλεία για την ανάλυση και ενσωμάτωση δεδομένων (π.χ. CPRG, GDAC και canEvolve). Η τρίτη κατηγορία χρησιμοποιείται κυρίως για οπτικοποίηση (π.χ., πρόγραμμα περιήγησης καρκίνου UCSC και CGWB). Η τέταρτη τάξη περιλαμβάνει βάσεις δεδομένων που εστιάζουν στην εμπλοκή των ειδικών βιολογικών χαρακτηριστικών σε καρκίνο (π.χ. MethyCancer, SomamiR και NONCODE). Τέλος, βάσεις δεδομένων όπως η GDSC και η CanSAR υποστηρίζουν την εφαρμογή της γενωμικής στην ανακάλυψη φαρμάκων. Υπάρχουν και άλλες χρήσιμες βάσεις δεδομένων και συνέχεια προστίθενται και νέες στον αγώνα εναντίον του καρκίνου. Ένα από τα σημαντικότερα εμπόδια παραμένει η ανομοιογένεια των καρκινικών κυττάρων και η μεταβλητότητα στην ανταπόκριση στα αντικαρκινικά φάρμακα μεταξύ ασθενών με παρόμοια συμπτώματα. Για να αντιμετωπιστεί αυτό το ζήτημα, τα προγράμματα προσωπικής γενωμικής πρέπει να ξεκινήσουν σε μεγαλύτερο πληθυσμό και να αναπτυχθούν προηγμένες υπολογιστικές και στατιστικές μεθοδολογίες με στόχο τη διαχείριση μεγάλων δεδομένων, ώστε να διαπιστωθεί η κλινική σημασία της ανίχνευσης του γονιδιώματος του καρκίνου.

ΟΝΟΜΑ	ΗΛΕΚΤΡΟΝΙΚΗ ΔΙΕΥΘΥΝΣΗ
EGA	https://www.ebi.ac.uk/ega
COSMIC	https://cancer.sanger.ac.uk/cosmic
CPRG	https://www.broadinstitute.org/software/cprg
GDAC	https://gdac.broadinstitute.org/
SNP500Cancer	http://cgap.nci.nih.gov/Tools
CanEVOLVE	http://www.canevolve.org
MethyCancer	http://methycancer.psych.ac.cn/
SomamiR	http://compbio.uthsc.edu/SomamiR/
cBioPortal	https://www.cbioportal.org/
USCS Cancer Genomics Browser	https://xena.ucsc.edu/welcome-to-ucsc-xena/
CGWB	https://omictools.com/cgwb-tool
GDSC	https://www.cancerrxgene.org/
canSAR	https://cansarblack.icr.ac.uk/
NONCODE	http://www.noncode.org/

Πίνακας 1. Διαδικτυακές πλατφόρμες για την έρευνα, επεξεργασία και ανάλυση του καρκίνου.

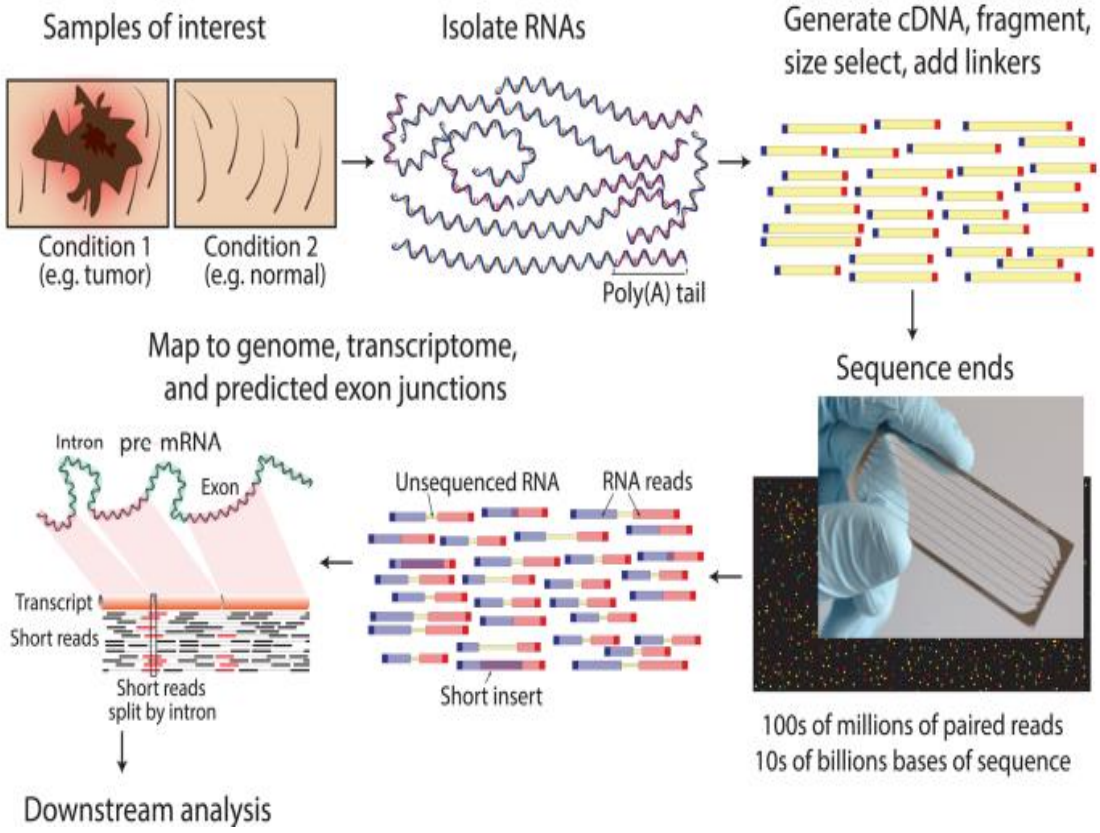
Πληροφορική για την εύρεση αλληλουχίας του RNA. Ένας ιστοχώρος για ανάλυση στο icloud

Η μαζικά παράλληλη ανάλυση αλληλουχίας RNA (RNA-seq) έχει γίνει γρήγορα η δοκιμασία επιλογής για την εξέταση της αφθονίας και της ποικιλομορφίας του μεταγραφικού RNA.. Διαθέσιμα RNA-seq tutorials με ανοιχτή πρόσβαση που καλύπτουν το cloud computing, την εγκατάσταση εργαλείων, τις σχετικές μορφές αρχείων, τα γονιδιώματα αναφοράς, τις σχολιασμένες μεταγραφές, τις στρατηγικές ελέγχου ποιότητας, την έκφραση, τη διαφορική έκφραση και τις εναλλακτικές μεθόδους ανάλυσης ματίσματος διατίθενται ελεύθερα στη διεύθυνση www.rnaseq.wiki

Εισαγωγή στην αλληλουχία RNA

Η γονιδιακή έκφραση είναι μια ευρέως μελετημένη διαδικασία και μια κύρια περιοχή εστίασης για τη λειτουργική γενωμική (Cheung VG *et al*., 2009). Η γονιδιακή έκφραση αφορά τη ροή γενετικής πληροφορίας από το γενωμικό DNA πρότυπο σε λειτουργικά πρωτεϊνικά προϊόντα. Η μαζικά παράλληλη ανάλυση αλληλουχίας RNA (RNA-seq) έχει καταστεί πρότυπη δοκιμασία έκφρασης γονιδίου, ιδιαίτερα για τη μέτρηση της σχετικής αφθονίας και ποικιλίας των αντιγράφων. Αρκετές μελέτες επιβεβαίωσαν ότι η ακρίβεια μέτρησης της RNA-seq ανταγωνίζεται εκείνες άλλων καλά εδραιωμένων μεθόδων όπως οι μικροσυστοιχίες (microarrays) και η ποσοτική αλυσιδωτή αντίδραση πολυμεράσης (quantitative Polymerace Chain Reaction, qPCR) (Consortium SM-I, doi: 10.1038/nbt.2957 PMID: 25150838). Οι

πειραματικές παράμετροι σχεδιασμού του RNA-seq παραμένουν μια περιοχή ανάπτυξης και μπορεί να έχουν σημαντικές επιπτώσεις στη στρατηγική ανάλυσης (van Dijk EL et al., 2014). . Τα πρωτόκολλα κατασκευής βιβλιοθήκης RNA-seq διαφέρουν ευρέως, και αυτές οι διαφορές έχουν σημαντικές συνέπειες για την ερμηνεία και ανάλυση δεδομένων (σχήμα 3).



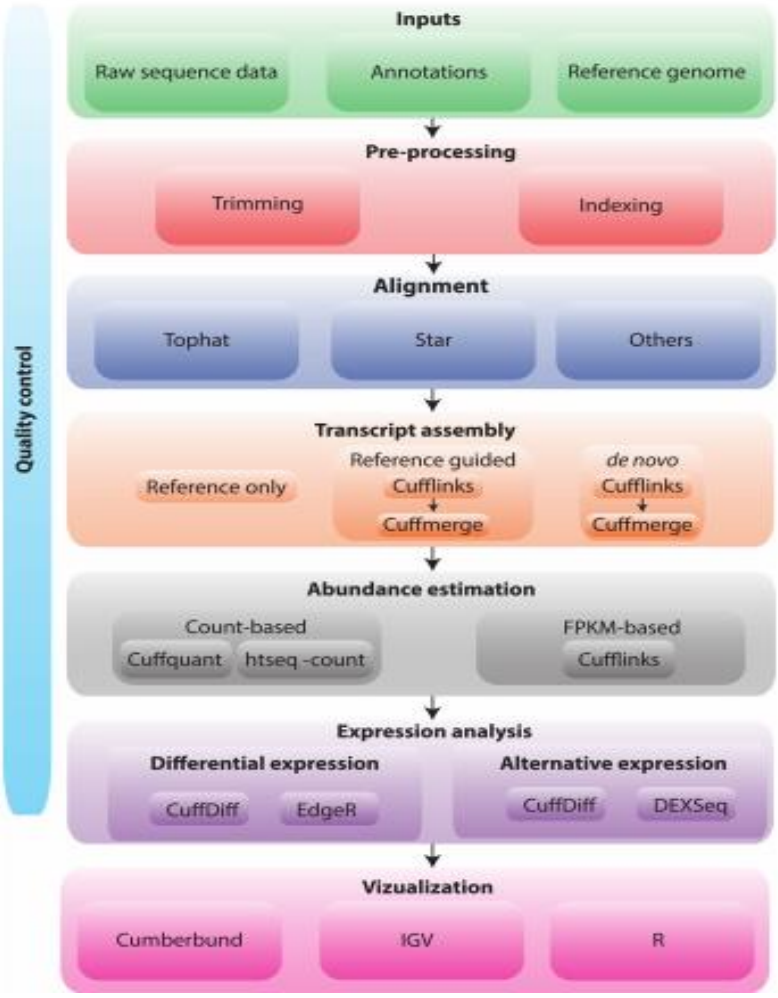
Σχήμα 3. Η μέθοδος RNA-seq τυπικά αποτελείται από ταυτοποίηση κατάλληλων βιολογικών δειγμάτων, απομόνωση ολικού RNA, εμπλουτισμό μη-ριβοσωματικών RNAs, μετατροπή RNA σε cDNA, κατασκευή μιας βιβλιοθήκης θραυσμάτων, προσδιορισμό αλληλουχίας σε μια πλατφόρμα προσδιορισμού αλληλουχίας υψηλής απόδοσης, δημιουργία "single or paired-end" αναγνώσεις (reads) μήκους 30-300 ζευγών βάσεων, ευθυγράμμιση ή συναρμολόγηση τους, και "downstream" ανάλυση (Σχήμα 3) (Nagalakshmi U *et al*., 2008, Wang Z *et al*., 2009). Το RNA-seq είναι κατάλληλο για διάφορα είδη ανάλυσης που περιλαμβάνουν: την ανακάλυψη μεταγράφων (transcript discovery) (Grabherr MG *et al*., 2011, Li B *et al*., 2014), σχολιασμό και καταγραφή γονιδιώματος (genome annotation) (Garber M *et al*., 2011), τη μελέτη των μηχανισμών της γονιδιακής ρύθμισης (Trapnell C *et al*., 2013), ανάλυση διαφορετικής γονιδιακής έκφρασης (Tarazona S *et al*., 2011), ανάλυση εναλλακτικής έκφρασης (Griffith M *et al*., 2010), ανάλυση έκφρασης συγκεκριμένου αλληλόμορφου (allele specific) (Pastinen T *et al*., 2010) ανίχνευση επεξεργασίας RNA (Peng Z *et al*., 2012), ανίχνευση ιών (Capobianchi MR *et al*., 2013, Khoury JD *et al*., 2013), ανίχνευση σύντηξης γονιδίων (Carrara M *et al*., 2013, Yoshihara K *et al*., 2014) και άλλους τύπους ανίχνευσης πολυμορφισμών (Quinn EM *et al*., 2013, Piskol R *et al*., 2013). Εκτός από αυτές τις συγκεκριμένες εφαρμογές, το RNA seq οδήγησε σε σημαντικές ανακαλύψεις σε πολλαπλά ερευνητικά πεδία. Οι ανακαλύψεις αυτές περιλαμβάνουν ανακαλύψεις σύντηξης στον καρκίνο (Honeyman JN *et al*., 2014), καλύτερη κατανόηση του επιπολασμού, των μηχανισμών και της ρύθμισης του εναλλακτικού ματίσματος (Sultan M *et al*., 2008, de Klerk E *et al*., 2015), βελτιωμένη κατανόηση του επιπολασμού και της λειτουργικής σημασίας των μη κωδικοποιήσιμων γονιδίων RNA (Mercer TR *et al*., 2012), μια αυξημένη (αλλά αμφιλεγόμενη)

εκτίμηση του επιπολασμού της επεξεργασίας του RNA (Peng *et al*., 2012) και πολλά άλλα. Το RNA-seq μεταφέρεται επίσης ενεργά σε κλινικές εφαρμογές σε πολλές ανθρώπινες ασθένειες (Kalari KR et al., 2014, Van Keuren-Jensen K et al., 2014).

Μορφές δεδομένων RNA-Seq, Έλεγχος ποιότητας, Περικοπή, Ευθυγράμμιση και Οπτικοποίηση

Για να κατανοήσουμε τα ακατέργαστα δεδομένα και τις ευθυγραμμίσεις RNA-seq, πρέπει να έχουμε υπόψιν μας τις μορφές των αρχείων καθώς και τα υποκείμενα μοντέλα δεδομένων που αντιπροσωπεύουν. Αυτές περιλαμβάνουν τη μορφή FASTA για την αποθήκευση δεδομένων γονιδιώματος (Pearson WR *et al*., 1988), τη μορφή GTF (Gene Transfer Format) για την αποθήκευση των σχολιασμών για γονίδια/αντίγραφα, τη μορφή FASTQ για την αποθήκευση πρωτογενών δεδομένων ανάγνωσης (Cock PJ *et al*., 2010), το χάρτη ευθυγράμμισης αλληλουχιών (SAM/BAM) (Li H *et al*., 2009), τις σημαίες SAM/BAM για την αποτελεσματική ταξινόμηση ορισμένων χαρακτηριστικών των ευθυγραμμίσεων ανάγνωσης και τις συμβολοσειρές CIGAR (Compact Idiosyncratic Gapped Alignment Report) για την αναπαράσταση συγκεκριμένων γραμμικών ευθυγραμμίσεων (Li H *et al*., 2009). Τα πρώτα βήματα των ροών εργασίας RNA-seq συχνά περιλαμβάνουν κάποιο αρχικό έλεγχο ποιότητας [(QC), Quality Control] των πρωτογενών δεδομένων σε αρχεία FASTQ (Σχήμα 4). Χωρίς ευθυγραμμίσεις, αυτό συνήθως περιλαμβάνει ανάλυση k-mer για τον εντοπισμό δυνητικών προβλημάτων, όπως μόλυνση από προσαρμογέα (adapter contamination), αναποτελεσματική απομάκρυνση ριβοσωμικών αλληλουχιών ή αφθονία θραυσμάτων μικρότερων από το μήκος αναγνώσεως στόχου (target read length). Επιπλέον μετρήσεις QC που αποκτήθηκαν σε αυτό το στάδιο μπορεί να προσδιορίσουν μη αποδεκτά ποιοτικά προφίλ, προβληματικούς κύκλους που μπορεί να έχουν συμβεί κατά τη διάρκεια της αλληλουχίας ή υπερβολικά πολλές διφορούμενες βάσεις (που υποδεικνύονται ως 'N' στα αρχεία FASTQ). Ανάλογα με τη στρατηγική κατασκευής της βιβλιοθήκης RNA-seq, κάποια μορφή ανάγνωσης μπορεί να είναι χρήσιμη πριν από την ευθυγράμμιση των δεδομένων του RNA seq.

Δύο κοινές στρατηγικές περικοπής περιλαμβάνουν το "κόψιμο του προσαρμογέα" και την "ποιοτική αποκοπή" (quality trimming). Η κοπή του προσαρμογέα περιλαμβάνει την αφαίρεση της ακολουθίας προσαρμογέα καλύπτοντας συγκεκριμένες ακολουθίες που χρησιμοποιούνται κατά την κατασκευή της βιβλιοθήκης. Η ποιοτική αποκοπή γενικά αφαιρεί τα άκρα των ακολουθιών όπου οι βαθμολογίες ποιότητας της βάσης έχουν πέσει σε ένα τέτοιο επίπεδο ώστε τα σφάλματα ακολουθίας και οι προκύπτουσες αναντιστοιχίες να αποτρέπουν τις αναγνώσεις από την ευθυγράμμιση. Τα εργαλεία όπως το skewer (Jiang H *et al*., 2014) και το trimmomatic (Bolger AM *et al*., 2014) συνδυάζουν μαζί διάφορους αλγόριθμους για την προσαρμογή των δεδομένων των ακατέργαστων RNA-seq και την αξιολόγηση της ποιότητας των δεδομένων ανάγνωσης πριν από την ευθυγράμμιση. (συμπληρωματικές πληροφορίες και διαδικασίες στη διεύθυνση www.rnaseq.wiki).



Σχήμα 4. Ένα παράδειγμα ροής εργασίας για ανάλυση RNA-seq για τυπική έκφραση γονιδίου και ανάλυση διαφορικής έκφρασης. Αυτές οι ροές εργασίας έχουν αρκετά κοινά θέματα σε διαφορετικά σύνολα εργαλείων και στόχους ανάλυσης RNA-seq. Η ανάλυση RNA-seq βασικά βασίζεται σε εισόδους όπως αλληλουχίες γονιδιώματος αναφοράς, παρατηρήσεις γονιδίων και δεδομένα ακατέργαστης ακολουθίας. Η εργασία με αυτές τις εισόδους απαιτεί εξοικείωση με αρκετές τυποποιημένες μορφές αρχείων όπως FASTA (.fa), FASTQ και [GTF, (Gene Transfer Format)]. Οι τυπικές ροές εργασίας ανάλυσης RNA-seq ξεκινούν με τον έλεγχο ποιότητας ακατέργαστων δεδομένων [QC, (Quality Control)] και στη συνέχεια εκτελούν ανάγνωση, ευθυγράμμιση ή συναρμολόγηση ανάγνωσης, εφαρμόζουν προσαρμοσμένους αλγορίθμους για συγκεκριμένο στόχο ανάλυσης (π.χ. Cufflinks και Cuffdiff για ανάλυση γονιδιακής έκφρασης) σύνοψη και απεικόνιση των αποτελεσμάτων. Για κάθε βήμα παρουσιάζονται εναλλακτικά και αντιπροσωπευτικά εργαλεία και στρατηγικές.

Αφού ολοκληρωθεί η ανάγνωση, το επόμενο βήμα στις περισσότερες εφαρμογές RNA-seq είναι η ευθυγράμμιση (Engstrom PG *et al*., 2013) ή η συναρμολόγηση (Martin JA *et al*., 2011). Η συναρμολόγηση RNA-seq είναι η προσπάθεια συγχώνευσης σε μεγαλύτερες συνεχόμενες αλληλουχίες (contigs) που βασίζονται μόνο στην ομοιότητα αλληλουχίας μεταξύ τους με την

ελπίδα να παράγουν ένα contig ανά μετάγραφο (transcript). Αυτή η τεχνική δεν βασίζεται σε προηγούμενες ακολουθίες για αναφορά.

Η ευθυγράμμιση RNA-seq περιλαμβάνει σύγκριση κάθε ανάγνωσης με μια προηγουμένως συναρμολογημένη αλληλουχία γονιδιώματος αναφοράς ή βάση δεδομένων μεταγραφικών αλληλουχιών. Μόλις ολοκληρωθούν οι ευθυγραμμίσεις, συχνά απαιτείται QC και ερμηνεία πριν από την εφαρμογή των διάφορων προγραμμάτων ανάλυσης της ακολουθίας. Τα δεδομένα ευθυγράμμισης RNA-seq είναι πολύπλοκα, μεγάλα και αφηρημένα. Αυτές οι ιδιότητες δημιουργούν μια ανάγκη για εργαλεία και εφαρμογές οπτικοποίησης που συνθέτουν, συνοψίζουν και προβάλλουν τα ακατέργαστα δεδομένα σε ένα ειδικό interface (διεπαφή). Ο περιηγητής γονιδιώματος (Genome Browser), όπως ο IGV (Integrative Genomics Viewer) (Thorvaldsdottir H *et al*., 2013), ο Savant (Fiume M *et al*., 2012) και ο IGB (Integrated Genome Browser), (Nicol JW *et al*., 2009) είναι σε θέση να εμφανίζουν αρχεία ευθυγράμμισης RNA-seq και έχουν προσαρμοστεί ώστε να αντιπροσωπεύουν μοναδικά χαρακτηριστικά των αλληλουχιών RNA όπως τα όρια εξονίου-εσωνίου, τις θέσεις ματίσματος κλπ. Ορισμένοι από αυτούς τους περιηγητές έχουν ενσωματώσει προσθήκες (plugins) που αφορούν συγκεκριμένες εφαρμογές RNA-seq. Για παράδειγμα, οι γραφικές παραστάσεις "sashimi" (Katz Y *et al*., 2015) του IGV επιτρέπουν την ερμηνεία των σύνθετων μοτίβων ματίσματος RNA που υποδεικνύονται από τα πρότυπα κάλυψης (coverage patterns) σε ένα σύνολο δεδομένων RNA-seq.

Έκφραση και διαφορική έκφραση

Μία από τις πιο ευρέως χρησιμοποιούμενες εφαρμογές του RNA-seq είναι η εκτίμηση της αφθονίας (abundance) γονιδίου ή μετάγραφου (transcript) και η σύγκριση αυτής της αφθονίας σε βιολογικές συνθήκες (Εικ. 4). Η εκτίμηση της αφθονίας των γονιδίων επιχειρεί να μετρήσει τη μεταγραφική έξοδο για μια φυσική θέση στο γονιδίωμα. Η εκτίμηση αφθονίας του μετάγραφου ασχολείται με το πιο πολύπλοκο πρόβλημα της προσπάθειας να προβλέψει και να μετρήσει την αφθονία συγκεκριμένων ισομορφών μεταγράφου RNA από κάθε τόπο (Rapaport F *et al*., 2013, Seyednasrollah F *et al*., 2015). Η ερμηνεία, η σύνοψη και η οπτικοποίηση των αποτελεσμάτων έκφρασης και διαφορικής έκφρασης μπορούν να είναι εξίσου περίπλοκα με τη δημιουργία αυτών των αποτελεσμάτων (Εικ. 4). Υπάρχουν πολλές πρόσθετες εφαρμογές και πλατφόρμες για μεταγενέστερη ερμηνεία της βιολογικής σημασίας της έκφρασης και των αποτελεσμάτων της διαφορικής έκφρασης. Το Weka (Frank E *et al*., 2015) είναι μια συλλογή αλγορίθμων μηχανικής μάθησης για εργασίες εξόρυξης δεδομένων. Περιέχει εργαλεία για την προετοιμασία των δεδομένων, την ταξινόμηση, την παλινδρόμηση, την ομαδοποίηση, την εξόρυξη κανόνων σύνδεσης και την απεικόνιση. Χρήσιμα εργαλεία επίσης είναι και το SeqGSEA (Sequence Based Gene-Set Enrichment Analysis) (Wang *X et al*., 2014) για ανάλυση γονιδίων και οδών, το GAGE (Generally Applicable Gene-set Enrichment for pathway analysis) (Luo W *et al*., 2009), το PathView (Luo W *et al*., 2013), το GoSeq (Gene Ontology analysis for RNA-seq) (Young MD *et al*., 2010) and το Cytoscape (Saito R *et al*., 2010). Το τελευταίο είναι μια ελεύθερης πρόσβασης πλατφόρμα

λογισμικού για την απεικόνιση δικτύων μοριακής αλληλεπίδρασης και την ενσωμάτωση με προφίλ γονιδιακής έκφρασης.

Γενικά, η τεχνολογία RNA-Seq είναι πολύ χρήσιμη για την ανάλυση διαφορικής έκφρασης και περιλαμβάνει κάποιες συγκεκριμένες συνθήκες, στις οποίες υιοθετούνται πέντε στάδια. Πρώτον, τα δείγματα RNA κατακερματίζονται σε μικρές συμπληρωματικές αλληλουχίες DNA (cDNA, complementary DNA) και στη συνέχεια προσδιορίζεται η αλληλουχία τους από πλατφόρμα υψηλής απόδοσης. Δεύτερον, οι μικρές παραγόμενες αλληλουχίες χαρτογραφούνται σε ένα γονιδίωμα. Τρίτον, για κάθε γονίδιο (ή ισόμορφα) υπολογίζονται τα επίπεδα έκφρασης. Τέταρτον, τα χαρτογραφημένα δεδομένα κανονικοποιούνται και με τη χρήση μεθόδων στατιστικής και μηχανικής μάθησης, προσδιορίζονται τα διαφοροποιημένα γονίδια (DEG, Differentially Expressed Genes). Τέλος, η συνάφεια των παραγόμενων δεδομένων αξιολογείται τελικά από ένα βιολογικό πλαίσιο (Li P *et al*., 2015). Με την αυξανόμενη δημοτικότητα της τεχνολογίας RNA-Seq, πολλά λογισμικά και αγωγοί αναπτύχθηκαν για ανάλυση διαφορικής γονιδιακής έκφρασης από αυτά τα δεδομένα.

Οι μέθοδοι για ανάλυση διαφορικής γονιδιακής έκφρασης από RNA-Seq μπορούν να ομαδοποιηθούν σε δύο κύριες υπομονάδες: παραμετρικές και μη παραμετρικές. Οι παραμετρικές μέθοδοι συλλαμβάνουν όλες τις πληροφορίες σχετικά με τα δεδομένα μέσα στις παραμέτρους. Σε αυτές τις περιπτώσεις, είναι δυνατόν να προβλεφθεί η αξία των άγνωστων δεδομένων από την παρατήρηση του υιοθετημένου μοντέλου και των παραμέτρων του. Όταν εφαρμόζονται παραμετρικές μέθοδοι σε διαφορική γονιδιακή έκφραση, υποθέτουμε ότι, συνήθως μετά από μια κανονικοποίηση, κάθε τιμή έκφρασης για ένα συγκεκριμένο γονίδιο χαρτογραφείται σε μια συγκεκριμένη κατανομή, όπως η Poisson (Hardcastle TJ *et al*., 2010) ή αρνητική διωνυμική (Robinson MD *et al*., 2010). Από την άλλη πλευρά, οι μη παραμετρικές μέθοδοι μπορούν να καταγράψουν περισσότερες λεπτομέρειες σχετικά με τη διανομή των δεδομένων, δηλαδή να μην επιβάλλουν την τοποθέτηση ενός άκαμπτου μοντέλου. Αυτό συμβαίνει επειδή τα μη παραμετρικά μοντέλα λαμβάνουν υπόψη ότι η κατανομή των δεδομένων δεν μπορεί να οριστεί από ένα πεπερασμένο σύνολο παραμέτρων, με αποτέλεσμα η ποσότητα πληροφοριών σχετικών με τα δεδομένα να αυξάνεται.

Ανακάλυψη Ισομορφών και Εναλλακτική Έκφραση

Ενώ πολλά πειράματα RNA-seq επικεντρώνονται στην εκτίμηση της αφθονίας και στην ανάλυση της διαφορικής έκφρασης γνωστών γονιδίων, η σχετικά αμερόληπτη δειγματοληψία "shotgun" του RNA-seq επιτρέπει επίσης την ανακάλυψη νέων ισομορφών μεταγράφων, την ανίχνευση διαφορετικών μοτίβων ματίσματος και την ανίχνευση χιμαιρικών γονιδίων σύντηξης . Ωστόσο, αυτές οι εφαρμογές περιορίζονται από τις σημαντικές προκλήσεις που σχετίζονται με την εξαγωγή μεταγράφων πλήρους μήκους από σχετικά μικρά θραύσματα RNA-seq. Το μέσο μετάγραφο κωδικοποίησης ανθρώπινης πρωτεΐνης έχει περίπου 8-10 εξόνια και είναι μήκους περίπου 2.000 bp. Εν τούτοις, μία

βιβλιοθήκη RNA-seq αποτελείται από θραύσματα cDNA μήκους περίπου 200-400 bp στα οποία η ακολουθία έχει προσδιορισθεί μερικώς από κάθε άκρο. Επίσης, ο κλώνος από τον οποίο μεταγράφηκε η αρχική αλληλουχία mRNA είναι άγνωστος σε πολλές στρατηγικές προετοιμασίας της βιβλιοθήκης γεγονός που κάνει το βασικό πρόβλημα που πολύ περίπλοκο. Όσο μεγαλύτερα τα μετάγραφα και όσο περισσότερα μετάγραφα εκφράζονται από έναν και μοναδικό τόπο (locus), τόσο πιο δύσκολο είναι να προσδιοριστούν οι μεταγραφικές ακολουθίες πλήρους μήκους και η αφθονία τους. Κάθε ισομορφή μεταγράφου έχει πολύ λίγα (αν υπάρχουν) εξόνια και συνδέσεις εξονίου-εξονίου που είναι μοναδικές σε αυτή τη ισομορφή. Ο χρυσός κανόνας για την ανάλυση της εναλλακτικής έκφρασης του τρανσκριπτώματος (transcriptome) παραμένει ο προσδιορισμός της αλληλουχίας πλήρους μήκους μιας μεγάλης ομάδας κλωνοποιημένων cDNAs που παράγονται με RT-PCR (Reverse Transcription PCR) (Team MGCP *et al*., 2009).

Ειδικές προκλήσεις για το RNA-Seq

Υπάρχουν διάφορες προκλήσεις αναφορικά με την ανάλυση RNA-seq (Williams AG *et al*., 2014) σε σύγκριση με την ανάλυση του DNA. Οι πιο σημαντικές είναι αυτές που σχετίζονται με την καθαρότητα του δείγματος, την ποιότητα και την ποσότητα. Το RNA είναι ασταθές και επιρρεπές σε υποβάθμιση, απαιτώντας πολλές εξειδικευμένες στρατηγικές διαχείρισης, δειγματοληψίας και ανάλυσης. Η κατασκευή των βιβλιοθηκών RNA-seq για τον προσδιορισμό της ακολουθίας έχει αλλάξει ραγδαία και οι σχετικές παραλλαγές στην προετοιμασία της βιβλιοθήκης μπορούν να επηρεάσουν το σχεδιασμό, την ανάλυση και την ερμηνεία της μελέτης RNA. Αυτό περιλαμβάνει διαφοροποιήσεις στις μεθόδους απομόνωσης και αποθήκευσης RNA, στις στρατηγικές εμπλουτισμού RNA, στις μεθόδους θρυμματισμού και επιλογής μεγέθους, στη χρήση της ενίσχυσης (amplification) και πολλά άλλα. Σε σύγκριση με την ανάλυση αλληλουχίας DNA, το στάδιο ευθυγράμμισης ανάγνωσης του RNA-seq είναι σημαντικά πιο δύσκολο (Trapnell C *et al*., 2009). Στους ευκαρυώτες η ανάγκη να επιλυθεί η δομή εξονίου/εσωνίου από σχετικά μικρές αναγνώσεις περιπλέκει τα στάδια ευθυγράμμισης και ανάλυσης. Τα εξόνια μπορούν να διαχωριστούν από μεγάλα ιντρόνια έτσι ώστε μια ευθυγράμμιση ανάγνωσης μιας μοναδικής αλληλουχίας να μπορεί να εκτείνεται σε εκατοντάδες χιλιάδες βάσεις σε δύο ή περισσότερα κενά που αντιστοιχούν στις θέσεις ματίσματος εσωνίων. Επίσης, σε σύγκριση με τον προσδιορισμό της ακολουθίας του γονιδιώματος, η αναμενόμενη σχετική αφθονία των RNA ποικίλει ευρέως, με δημοσιευμένες εκτιμήσεις να κάνουν λόγο ότι αναμένονται τάξεις μεγέθου τουλάχιστον 10^5-10^7 μεταξύ των γονιδίων με τη χαμηλότερη και την υψηλότερη έκφραση [85,86]. Εφόσον το RNA-seq λειτουργεί με τυχαία δειγματοληψία, ένα μικρό ποσοστό γονιδίων με υψηλή έκφραση μπορεί να καταναλώσει την πλειονότητα των αναγνώσεων. Μία συνέπεια αυτού του γεγονότος είναι ότι για να πάρουμε ένα πολύ μικρό τμήμα (snapshot) του τρανσκριπτόματος που περιλαμβάνει γονίδια χαμηλής έκφρασης, μια βιβλιοθήκη RNA-seq πρέπει να είναι πολύ βαθύτερη από αυτή που μπορεί κανείς να περιμένει με βάση την αναλογία βάσεων σε ένα γονιδίωμα που εμφανίζονται ως εκφρασμένες. Τα ριβοσωματικά και μιτοχονδριακά γονίδια εκφράζονται ιδιαίτερα σε πολλούς ιστούς και σημαντικά βήματα στις στρατηγικές

προετοιμασίας και ανάλυσης της βιβλιοθήκης RNA-seq αφορούν την απομάκρυνση τους. Ένα άλλο ξεχωριστό χαρακτηριστικό των μορίων του RNA που επηρεάζει την ανάλυση είναι ότι εμφανίζονται σε ένα ευρύ φάσμα μεγεθών. Πολύ μικρά RNAs (<100bp) όπως τα μικρά RNAs (miRNA) πρέπει γενικά να συλλαμβάνονται και να προσδιορίζεται η αλληλουχία τους με μια ανεξάρτητη στρατηγική, καθώς οι στρατηγικές με βάση το μέγεθος λογικά τα αποκλείουν και αυτά τα πολύ μικρά RNAs (Malone C *et al*., 2012).

Οι στόχοι ανάλυσης των πειραμάτων RNA-seq είναι ποικίλοι. Κάθε ένας από αυτούς τους στόχους ανάλυσης έχει ξεχωριστές απαιτήσεις και προκλήσεις. Ωστόσο, μια κοινή ροή εργασιών περιλαμβάνει γενικά τη λήψη πρώτων δεδομένων, την προεπεξεργασία αυτών των δεδομένων τη βασική εκτίμηση της ποιότητας που οδηγεί είτε σε ευθυγράμμιση είτε σε συναρμολόγηση των αναγνώσεων, την επεξεργασία των ευθυγραμμίσεων που προκύπτουν με ένα εργαλείο ειδικό για ανάλυση και την περίληψη και οπτικοποίηση των αποτελεσμάτων.

Βιβλιογραφία

Bhattacharya A, Ziebarth JD, Cui Y. SomamiR: a database for somatic mutations impacting microRNA function in cancer. Nucleic Acids Res 2013;41:D977–82.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30(15):2114–20. doi: 10.1093/bioinformatics/btu170 PMID: 24695404

Bulusu KC, Tym JE, Coker EA, Schierz AC, Al-Lazikani B. CanSAR: updated cancer research and drug discovery knowledgebase. Nucleic Acids Res 2014;42:D1040–7.

Capobianchi MR, Giombini E, Rozera G. Next-generation sequencing technology in clinical virology. Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases. 2013; 19(1):15–22.

Carrara M, Beccuti M, Cavallo F, Donatelli S, Lazzarato F, Cordero F, et al. State of art fusion-finder algorithms are suitable to detect transcription-induced chimeras in normal tissues? BMC bioinformatics. 2013; 14 Suppl 7:S2. doi: 10.1186/1471-2105-14-S7-S2 PMID: 23815381

Cheung VG, Spielman RS. Genetics of human gene expression: mapping DNA variants that influence gene expression. Nature reviews Genetics. 2009; 10(9):595–604. doi: 10.1038/nrg2630 PMID: 19636342

Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic acids research. 2010; 38(6):1767–71. doi: 10.1093/nar/gkp1137 PMID: 20015970

Consortium SM-I. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nature biotechnology. 2014; 32(9):903–14. doi: 10.1038/nbt.2957 PMID: 25150838

de Klerk E, t Hoen PA. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. Trends in genetics: TIG. 2015; 31(3):128–39. doi: 10.1016/j.tig.2015.01.001 PMID: 25648499

Engstrom PG, Steijger T, Sipos B, Grant GR, Kahles A, Ratsch G, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. Nature methods. 2013; 10(12):1185–91. doi: 10.1038/nmeth.2722 PMID: 24185836

Fang, S., Zhang, L., Guo, J., et al. (2017) NONCODEV5: a comprehensive annotation database for long non-coding RNAs. Nucleic Acids Res.

Fiume M, Smith EJ, Brook A, Strbenac D, Turner B, Mezlini AM, et al. Savant Genome Browser 2: visualization and analysis for population-scale genomics. Nucleic acids research. 2012; 40(Web Server issue):W615–21. doi: 10.1093/nar/gks427 PMID: 22638571

Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using Weka. Bioinformatics. 2004; 20(15):2479–81. PMID: 15073010

Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013;6:l1.

Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. Nature methods. 2011; 8(6):469–77 doi: 10.1038/nmeth.1613 PMID: 21623353

Goldman M, Craft B, Swatloski T, Ellrott K, Cline M, Diekhans M, et al. The UCSC cancer genomics browser: update 2013. Nucleic Acids Res 2013;41:D949–54

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology. 2011; 29(7):644–52. doi: 10.1038/nbt.1883 PMID: 21572440

Griffith M, Griffith OL, Mwenifumbo J, Goya R, Morrissy AS, Morin RD, et al. Alternative expression analysis by RNA sequencing. Nature methods. 2010; 7(10):843–7. doi: 10.1038/nmeth.1503 PMID: 20835245

Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC bioinformatics. 2010;11(1):422. pmid:20698981

He X, Chang S, Zhang J, Zhao Q, Xiang H, Kusonmano K, et al. MethyCancer: the database of human DNA methylation and cancer. Nucleic Acids Res 2008;36:D836–41.

Honeyman JN, Simon EP, Robine N, Chiaroni-Clarke R, Darcy DG, Lim II, et al. Detection of a recurrent DNAJB1-PRKACA chimeric transcript in fibrolamellar hepatocellular carcinoma. Science. 2014; 343 (6174):1010–4. doi: 10.1126/science.1249484 PMID: 24578576

Jiang H, Lei R, Ding SW, Zhu S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC bioinformatics. 2014; 15:182. doi: 10.1186/1471-2105-15-182 PMID: 24925680

Kalari KR, Nair AA, Bhavsar JD, O'Brien DR, Davila JI, Bockol MA, et al. MAP-RSeq: Mayo Analysis Pipeline for RNA sequencing. BMC bioinformatics. 2014; 15:224. doi: 10.1186/1471-2105-15-224 PMID: 24972667

Katz Y, Wang ET, Silterra J, Schwartz S, Wong B, Thorvaldsdottir H, et al. Quantitative visualization of alternative exon expression from RNA-seq data. Bioinformatics. 2015; 31: 2400–2402. doi: 10.1093/bioinformatics/btv034 PMID: 25617416

Khoury JD, Tannir NM, Williams MD, Chen Y, Yao H, Zhang J, et al. Landscape of DNA virus associations across human malignant cancers: analysis of 3,775 cases using RNA-Seq. Journal of virology. 2013; 87(16):8916–26. doi: 10.1128/JVI.00340-13 PMID: 23740984

Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, et al. Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome biology. 2014; 15(12):553. doi: 10.1186/s13059-014-0553-5 PMID: 25608678

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16):2078–9. doi: 10.1093/bioinformatics/btp352 PMID: 19505943

Li P, Piao Y, Shon HS, Ryu KH. Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. BMC bioinformatics. 2015;16(1):347. pmid:26511205

Liu C, Bai B, Skogerbo G, Cai L, Deng W, Zhang Y, et al. NONCODE: an integrated knowledge database of non-coding RNAs. Nucleic Acids Res 2005;33:D112–5.

Luo W, Brouwer C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics. 2013; 29(14):1830–1. doi: 10.1093/bioinformatics/btt285 PMID: 23740750

Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. GAGE: generally applicable gene set enrichment for pathway analysis. BMC bioinformatics. 2009; 10:161. doi: 10.1186/1471-2105-10-161 PMID: 19473525

Malone C, Brennecke J, Czech B, Aravin A, Hannon GJ. Preparation of small RNA libraries for highthroughput sequencing. Cold Spring Harbor protocols. 2012; 2012(10):1067–77. doi: 10.1101/pdb.prot071431 PMID: 23028068

Martin JA, Wang Z. Next-generation transcriptome assembly. Nature reviews Genetics. 2011; 12 (10):671–82. doi: 10.1038/nrg3068 PMID: 21897427

Marx V. Drilling into big cancer-genome data. Nat Methods 2013;10:293–7.

Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, et al. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nature biotechnology. 2012; 30 (1):99–104.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008; 320(5881):1344–9. doi: 10.1126/science.1158441 PMID: 18451266

Nicol JW, Helt GA, Blanchard SG Jr., Raja A, Loraine AE. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics. 2009; 25(20):2730–1. doi: 10.1093/bioinformatics/btp472 PMID: 19654113

Packer BR, Yeager M, Burdett L, Welch R, Beerman M, Qi L, et al. SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes. Nucleic Acids Res 2006;34:D617–21.

Pastinen T. Genome-wide allele-specific analysis: insights into regulatory variation. Nature reviews Genetics. 2010; 11(8):533–8. doi: 10.1038/nrg2815 PMID: 20567245

Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences of the United States of America. 1988; 85(8):2444–8. PMID: 3162770

Peng Z, Cheng Y, Tan BC, Kang L, Tian Z, Zhu Y, et al. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. Nature biotechnology. 2012; 30(3):253–60. doi: 10.1038/nbt.2122 PMID: 22327324

Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-seq data. American journal of human genetics. 2013; 93(4):641–51. doi: 10.1016/j.ajhg.2013.08.008 PMID: 24075185

Quinn EM, Cormican P, Kenny EM, Hill M, Anney R, Gill M, et al. Development of strategies for SNP detection in RNA-seq data: application to lymphoblastoid cell lines and evaluation using 1000 Genomes data. PloS one. 2013; 8(3):e58815. doi: 10.1371/journal.pone.0058815 PMID: 23555596

Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome biology. 2013; 14(9):R95. PMID: 24020486.

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–140. pmid:19910308

Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, Lotia S, et al. A travel guide to Cytoscape plugins. Nature methods. 2012; 9(11):1069–76. doi: 10.1038/nmeth.2212 PMID: 23132118

Samur MK, Yan Z, Wang X, Cao Q, Munshi NC, Li C, et al. CanEvolve: a web portal for integrative oncogenomics. PLoS One 2013;8:e56228.

Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. Briefings in bioinformatics. 2015; 16(1):59–70. doi: 10.1093/bib/bbt086 PMID: 24300110

Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science. 2008; 321 (5891):956–60. doi: 10.1126/science.1160342 PMID: 18599741

Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: a matter of depth. Genome research. 2011; 21(12):2213–23 doi: 10.1101/gr.124321.111 PMID: 21903743

Team MGCP, Temple G, Gerhard DS, Rasooly R, Feingold EA, Good PJ, et al. The completion of the Mammalian Gene Collection (MGC). Genome research. 2009; 19(12):2324–33. doi: 10.1101/gr. 095976.109 PMID: 19767417

Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics. 2013; 14(2):178–92. doi: 10. 1093/bib/bbs017 PMID: 22517427

Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25(9):1105–11. doi: 10.1093/bioinformatics/btp120 PMID: 19289445

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols. 2012; 7(3):562–78. doi: 10.1038/nprot.2012.016 PMID: 22383036

van Dijk EL, Jaszczyszyn Y, Thermes C. Library preparation methods for next-generation sequencing: tone down the bias. Experimental cell research. 2014; 322(1):12–20. doi: 10.1016/j.yexcr.2014.01.008 PMID: 24440557

Van Keuren-Jensen K, Keats JJ, Craig DW. Bringing RNA-seq closer to the clinic. Nature biotechnology. 2014; 32(9):884–5. doi: 10.1038/nbt.3017 PMID: 25203037

Wang X, Cairns MJ. SeqGSEA: a Bioconductor package for gene set enrichment analysis of RNA-Seq data integrating differential expression and splicing. Bioinformatics. 2014; 30(12):1777–9. doi: 10.1093/ bioinformatics/btu090 PMID: 24535097

Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews Genetics. 2009; 10(1):57–63. doi: 10.1038/nrg2484 PMID: 19015660

Wilks C, Cline MS, Weiler E, Diehkans M, Craft B, Martin C, et al. The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. Database (Oxford) 2014;2014:bau093

Williams AG, Thomas S, Wyman SK, Holloway AK. RNA-seq Data: Challenges in and Recommendations for Experimental Design and Analysis. Current protocols in human genetics / editorial board, Haines Jonathan L [et al]. 2014; 83:11 3 1–3 20.

Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res 2013;41:D955–61.

Yoshihara K, Wang Q, Torres-Garcia W, Zheng S, Vegesna R, Kim H, et al. The landscape and therapeutic relevance of cancer-associated transcript fusions. Oncogene. 2014. doi: 10.1038/onc.2014.406 E-pub ahead of print.

Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome biology. 2010; 11(2):R14. doi: 10.1186/gb-2010-11-2-r14 PMID: 20132535

Zhang J, Finney RP, Rowe W, Edmonson M, Yang SH, Dracheva T, et al. Systematic analysis of genetic alterations in tumors using Cancer Genome WorkBench (CGWB). Genome Res 2007;17:1111–7.

1.2.2 Προγνωστικοί μηχανισμοί στη γενετική

Η βαθιά μάθηση είναι συλλογή μεθόδων και κατάλληλων αλγορίθμων με τις οποίες βελτιώνεται η αποδοτικότητα μιας μηχανής (βλ. υπολογιστή) στο έργο της για την εκτέλεση «ευφυών» εργασιών.

Η βαθιά μάθηση (Deep Learning) παίζει επίσης σημαντικό ρόλο στον προσδιορισμό της αλληλουχίας του γονιδιώματος και στις αναλύσεις γονιδιακής έκφρασης. Για να εξαχθούν τα προφίλ έκφρασης των γονιδίων -στόχων που βασίζονται σε περίπου 1000 γονίδια ορόσημα από το πρόγραμμα Integrated Network Based Cellular Signatures (LINCS) του National Institute of Health (NIH), οι Chen *et al*., παρουσίασαν τη D-GEX, μια μέθοδο βαθιάς μάθησης η οποία υπερέβαινε σημαντικά την γραμμική παλινδρόμηση [Linear Regression, (LR)] όσον αφορά την ακρίβεια της πρόβλεψης τόσο στα δεδομένα μικροσυστοιχειών (microarray) όσο και σε δεδομένα RNA-seq. Με την εφαρμογή ενός πολυτροπικού δικτύου βαθιάς μάθησης [Deep Belief Network, (DBN)] για τη μοντελοποίηση των δομικών δεσμευτικών προτιμήσεων και για την πρόβλεψη θέσεων πρόσδεσης πρωτεϊνών σύνδεσης RNA [RNA Binding Proteins (RBPs)] χρησιμοποιώντας την πρωτογενή αλληλουχία καθώς και τα δευτερεύοντα και τριτοταγή δομικά προφίλ, οι Zhang *et al*., πέτυχαν μια “Area Under the Curve” (AUC)0,98 για ορισμένες πρωτεΐνες. Για την πρόβλεψη θέσεων δέσμευσης πρωτεϊνών σύνδεσης DNA και RNA, οι Alipanahi *et al*. ανέπτυξαν τη DeepBind, μια μέθοδο βασισμένη στο [Convolutional Neural Network, Συνελικτικό Νευρωνικό Δίκτυο (CNN)], η οποία ξεπέρασε άλλες μεθόδους τελευταίας τεχνολογίας, ακόμα και όταν εκπαιδεύτηκε με *in vitro* δεδομένα και δοκιμάστηκε με δεδομένα *in vivo*.

Η είσοδος αυτών των βαθιών CNNs είναι κωδικοποιημένοι χαρακτήρες αλληλουχίας που λαμβάνονται μέσω μικροσυστοιχιών δέσμευσης πρωτεϊνών και η έξοδος είναι μια πραγματική τιμή που υποδεικνύει εάν η αλληλουχία είναι μια θέση δέσμευσης ή όχι. Το βαθύτερο μοντέλο μπορεί να κάνει πιο ακριβή ταξινόμηση, εξάγοντας χαρακτηριστικά υψηλότερου επιπέδου από τις πρώτες αλληλουχίες νουκλεοτιδίων. Επιπροσθέτως, οι Kelley *et al*., παρουσίασαν το Basset, ένα πακέτο ανοιχτού κώδικα για την εφαρμογή βαθιών CNN για να μάθει τον κώδικα πρόσβασης της χρωματίνης, επιτρέποντας το σχολιασμό (annotation) και την ερμηνεία του μη κωδικοποιούμενου γονιδιώματος.

Η γενετική διαφοροποίηση μπορεί να επηρεάσει τη μεταγραφή του DNA και τη μετάφραση του mRNA (Guigo R *et al*., 2015). Η κατανόηση των επιδράσεων των παραλλαγών της αλληλουχίας στο προ-mRNA μάτισμα διευκολύνει όχι μόνο τον πλήρη σχολιασμό του γονιδιώματος αλλά και την κατανόηση της λειτουργίας του γονιδιώματος. Για να προβλέψουνε τη σύνδεση ματίσματος (splice junction) σε επίπεδο DNA, ο Yoon και οι συνεργάτες του ανέπτυξαν μια νέα μέθοδο βασισμένη στο DBN που εκπαιδεύτηκε στα [Restricted Boltzmann Machines, (RBMs), Hinton GE *et al*., 2007]. Η μέθοδος τους όχι μόνο πέτυχε μεγαλύτερη ακρίβεια αλλά και ανακάλυψε τα λεπτά μη-κανονικά μοτίβα ματίσματος. Επιπλέον, εκμεταλλευόμενοι τα επαναλαμβανόμενα νευρωνικά δίκτυα [Recurrent Neural Networks, (RNNs)] για να μοντελοποιήσουν και να ανιχνεύσουν τις συνδέσεις ματίσματος από αλληλουχίες DNA, οι ίδιοι συγγραφείς πέτυχαν επίσης καλύτερη απόδοση από την προηγούμενη μέθοδο που βασίστηκε στο DBN.

Οι Frey *et al*., διατύπωσαν τη συναρμολόγηση ενός κώδικα ματίσματος ως πρόβλημα στατιστικής επίτευξης και πρότειναν μια Bayesian μέθοδο για την πρόβλεψη του ματίσματος που ρυθμίζεται από ιστούς χρησιμοποιώντας αλληλουχίες RNA και κυτταρικό πλαίσιο. Στη συνέχεια, ανέπτυξαν ένα μοντέλο DNN με εγκατάλειψη για να μάθουν και να προβλέψουν το εναλλακτικό μάτισμα [Alternative Splicing (AS)] (Leung MKK *et al*., 2014). Αυτό το μοντέλο πήρε ως εισροές τόσο τα γενωμικά χαρακτηριστικά όσο και το πλαίσιο των ιστών και προέβλεψε μοτίβα ματίσματος σε μεμονωμένους ιστούς και διαφορές στα μοτίβα ματίσματος σε όλους τους ιστούς. Έδειξαν ότι η μέθοδος τους ξεπέρασε τις προηγούμενες Bayesian μεθόδους και άλλους κοινούς αλγόριθμους μηχανικής μάθησης, όπως η Multinomial Logistic Regression (MLR) και οι Support Vector Machines (SVMs), από την άποψη της πρόβλεψης AS. Επιπλέον, δημιούργησαν ένα υπολογιστικό μοντέλο χρησιμοποιώντας έναν Bayesian deep αλγόριθμο μάθησης για να προβλέψουν τις επιπτώσεις των γενετικών παραλλαγών στο AS (Xiong HY *et al*., 2014).

Αυτό το μοντέλο έλαβε μόνο τις αλληλουχίες DNA ως εισροές χωρίς τη χρήση σχολιασμών για τις ασθένειες ή δεδομένων πληθυσμού και έπειτα βαθμολόγησε τις επιδράσεις που είχαν οι παραλλαγές στο AS παρέχοντας πολύτιμες γνώσεις στους γενετικούς καθοριστικούς παράγοντες της σπονδυλικής μυϊκής ατροφίας, του καρκίνου του παχέος εντέρου και της διαταραχής του φάσματος του αυτισμού.

Η βαθιά μάθηση είναι χρήσιμη στην ερμηνεία μεγάλου όγκου δεδομένων και βοηθάει στην πρόγνωση ασθενειών, την πρόληψη, τη διάγνωση και τη θεραπεία. Είναι σίγουρο ότι στο

άμεσο μέλλον θα υπάρξουν πιο βαθιές εφαρμογές μάθησης στην πρόβλεψη των επιδημιών, στην πρόληψη των ασθενειών και στην κλινική λήψη αποφάσεων.

Υπολογιστική νοημοσύνη – Computer Intelligence (CI)

Η υπολογιστική νοημοσύνη [Computer Intelligence, (CI)] έχει πολλά από τα χαρακτηριστικά των βιολογικών υπολογιστών και είναι ικανή να εκτελέσει μια ποικιλία εργασιών που είναι δύσκολο να γίνουν με τη χρήση συμβατικών τεχνικών. Πρόκειται για μια μεθοδολογία που περιλαμβάνει μηχανισμούς προσαρμογής και/ή ικανότητα μάθησης που διευκολύνει την έξυπνη συμπεριφορά σε περίπλοκα και μεταβαλλόμενα περιβάλλοντα, έτσι ώστε το σύστημα να θεωρείται ότι διαθέτει ένα ή περισσότερα χαρακτηριστικά λογικής, όπως γενίκευση, ανακάλυψη, συσχέτιση και αφαίρεση. Παρακάτω θα δείξουμε πώς μπορούν να εφαρμοστούν τεχνικές CI για την επίλυση προβλημάτων βιοπληροφορικής και πώς τα δεδομένα βιοπληροφορικής μπορούν να αναλυθούν, να υποστούν επεξεργασία και να χαρακτηριστούν από το CI.

Οι σύγχρονές τεχνικές του CI περιλαμβάνουν τα τεχνητά νευρωνικά δίκτυα (Artificial Neural Networks ANN, τις περιορισμένες μηχανές Boltzmann (Restricted Boltzmann machines (RBM)], τα βαθιά δίκτυα πεποιθήσεων [Deep Belief Network (DBN)], τις μηχανές διανυσματικής υποστήριξης (Support Vector Machines SVMs), την ασαφή λογική (fuzzy logic), τα ακατέργαστα σύνολα (rough sets), τους εξελικτικούς αλγόριθμους (Evolutionary Algorithms, EA), τους γενετικούς αλγόριθμους (GAs), το τεχνητό ανοσοποιητικό σύστημα,

Τεχνητά Νευρωνικά Δίκτυα

Τα τεχνητά νευρωνικά δίκτυα [Artificial Neural Networks, (ANN)] έχουν αναπτυχθεί ως γενικεύσεις μαθηματικών μοντέλων βιολογικών νευρικών συστημάτων. Σε ένα απλοποιημένο μαθηματικό μοντέλο ενός νευρώνα, τα αποτελέσματα των συνάψεων αντιπροσωπεύονται από βάρη σύνδεσης που διαμορφώνουν το αποτέλεσμα των συναφών σημάτων εισόδου, ενώ το μη γραμμικό χαρακτηριστικό που παρουσιάζεται από τους νευρώνες αντιπροσωπεύεται από μια λειτουργία μεταφοράς (Motsinger FA *et al*., 2006).

Περιορισμένες Μηχανές Boltzmann

Η περιορισμένη μηχανή Boltzmann [Restricted Boltzmann machine (RBM)] είναι ένα μη-προσανατολισμένο παραγωγικό (generative) μοντέλο βασισμένο στην ενέργεια που χρησιμοποιεί ένα στρώμα κρυφών μεταβλητών για να μοντελοποιήσει τη διανομή πάνω από ορατές μεταβλητές (Larochelle H *et al*., 2008). Το μοντέλο με βάση την ενέργεια σημαίνει ότι η κατανομή πιθανότητας πάνω από τις μεταβλητές ενδιαφέροντος ορίζεται μέσω μιας συνάρτησης. Συντίθεται από ένα σύνολο παρατηρήσιμων μεταβλητών V = {vi} και ένα σύνολο

κρυφών μεταβλητών H = {hj}, κόμβος i στο ορατό στρώμα, κόμβος j στο κρυφό στρώμα. Είναι περιορισμένη υπό την έννοια ότι δεν υπάρχουν ορατές-ορατές ή κρυφές-κρυφές συνδέσεις

Βαθιά δίκτυα πεποιθήσεων

Στη μηχανική μάθηση, ένα δίκτυο βαθιών πεποιθήσεων [Deep Belief Network (DBN)] είναι ένα παραγωγικό (generative) γραφικό μοντέλο ή εναλλακτικά ένα βαθύ νευρωνικό δίκτυο που αποτελείται από πολλαπλά στρώματα λανθανουσών μεταβλητών ("κρυμμένες μονάδες"), με συνδέσεις μεταξύ των στρωμάτων αλλά όχι μεταξύ μονάδων κάθε στρώματος (Hinton J 2009). Όταν εκπαιδεύεται σε ένα σύνολο παραδειγμάτων χωρίς επίβλεψη, ένα DBN μπορεί να μάθει να ανασυνθέτει τις εισροές του. Τα στρώματα λειτουργούν ως ανιχνευτές χαρακτηριστικών. Μετά από αυτό το βήμα μάθησης, ένα DBN μπορεί να εκπαιδευτεί περαιτέρω με επίβλεψη για να κάνει την ταξινόμηση.

Μηχανές Διανυσματικής Υποστήριξης

Τα SVM είναι μια ομάδα εποπτευόμενων μεθόδων μάθησης με συναφείς αλγόριθμους μάθησης που αναλύουν δεδομένα που χρησιμοποιούνται για ανάλυση ταξινόμησης και παλινδρόμησης. Η ταξινόμηση επιτυγχάνεται με γραμμική ή μη γραμμική επιφάνεια διαχωρισμού στον χώρο εισόδου του συνόλου δεδομένων. Το SVM παρέχει κορυφαίες επιδόσεις σε πραγματικές εφαρμογές, όπως κατηγοριοποίηση κειμένου, χειρόγραφη αναγνώριση χαρακτήρων, ταξινόμηση εικόνων, ανάλυση βιολογικών αλληλουχιών κλπ. και έχει πλέον καθιερωθεί ως ένα από τα βασικά εργαλεία για CI και εξόρυξη δεδομένων. Χρησιμοποιεί πυρήνες (kernels) για να μετατρέψει τα δεδομένα εισόδου σε ένα πιο διακριτό χώρο χαρακτηριστικών διαστάσεων, σιωπηρά στην οποία τα δεδομένα καθίστανται γραμμικά διαχωρίσιμα. Το όριο της γραμμικής απόφασης σχεδιάζεται με τέτοιο τρόπο ώστε να μεγιστοποιείται το περιθώριο, η ελάχιστη απόσταση μεταξύ των παραδειγμάτων εκπαίδευσης και των ορίων. Σε περίπτωση που τα χαρτογραφημένα σημεία δεδομένων είναι γραμμικά αδιαχώριστα, συμπεριλαμβάνεται ένα κόστος για την καταχώρηση των εσφαλμένα ταξινομημένων παραδειγμάτων και το περιθώριο μεγιστοποιείται μαζί με την ελαχιστοποίηση του κόστους (Burges CJC 1998).

Ασαφής Λογική

Η ασαφής λογική ξεκινά με την έννοια των ασαφών συνόλων. Ένα ασαφές σύνολο είναι ένα σετ χωρίς σαφώς καθορισμένο όριο και μπορεί να περιέχει μόνο στοιχεία με μερικό βαθμό συμμετοχής.. Σε αντίθεση με τα συμβατικά σύνολα, το ασαφές σύνολο εκφράζει τον βαθμό στον οποίο ένα στοιχείο ανήκει σε ένα σετ. Επομένως, η συνάρτηση μέλους ενός ασαφούς συνόλου μπορεί να πάρει μια τιμή μεταξύ 0 και 1, δηλώνοντας τον βαθμό συμμετοχής ενός στοιχείου σε ένα δεδομένο σετ (Nguyen, H.T *et al*., 1999).

Rough Sets (Ακατέργαστα Σύνολα)

Η θεωρία των ακατέργαστων συνόλων είναι μια αρκετά νέα έξυπνη τεχνική για τη διαχείριση της αβεβαιότητας που μπορεί να χρησιμοποιηθεί για την εξεύρεση εξαρτήσεων δεδομένων, την αξιολόγηση της σημασίας των χαρακτηριστικών, την ανακάλυψη προτύπων στα δεδομένα, τη μείωση των χαρακτηριστικών και την εξαγωγή κανόνων από τις βάσεις δεδομένων. Αυτοί οι κανόνες έχουν τη δυνατότητα να αποκαλύψουν νέα πρότυπα στα δεδομένα και μπορούν επίσης να λειτουργήσουν συλλογικά ως ταξινομητές για αόρατα σύνολα δεδομένων. Σε αντίθεση με άλλες τεχνικές υπολογιστικής νοημοσύνης, η ανάλυση της ακατέργαστης δέσμης δεν απαιτεί εξωτερικές παραμέτρους και χρησιμοποιεί μόνο τις πληροφορίες που υπάρχουν στα διαθέσιμα δεδομένα. Ένα από τα ενδιαφέροντα χαρακτηριστικά της θεωρίας των ακατέργαστων συνόλων είναι ότι μπορεί να πει αν τα δεδομένα είναι πλήρη ή δεν βασίζονται στα ίδια τα δεδομένα. Εάν τα δεδομένα είναι ελλιπή, προτείνονται περισσότερες πληροφορίες σχετικά με τα αντικείμενα προς συλλογή προκειμένου να δημιουργηθεί ένα καλό μοντέλο ταξινόμησης. Από την άλλη πλευρά, εάν τα δεδομένα είναι πλήρη, τα ακατέργαστα σύνολα μπορούν να καθορίσουν τα ελάχιστα δεδομένα που είναι απαραίτητα για ταξινόμηση (Pawlak Z *et al*., 1995, Polkowski L 2003).

Εξελικτικοί αλγόριθμοι

Οι εξελικτικοί αλγόριθμοι [Evolutionary Algorithms, (EA)] είναι προσαρμοστικές μέθοδοι, οι οποίες μπορεί να χρησιμοποιηθούν για την επίλυση προβλημάτων αναζήτησης και βελτιστοποίησης, με βάση γενετικές διαδικασίες των βιολογικών οργανισμών. Σε πολλές γενιές, οι φυσικοί πληθυσμοί εξελίσσονται σύμφωνα με τις αρχές της φυσικής επιλογής και "επιβίωσης του ικανότερου". Με τη μίμηση αυτής της διαδικασίας, οι εξελικτικοί αλγόριθμοι μπορούν να "εξελίξουν" λύσεις στα προβλήματα του πραγματικού κόσμου, αν έχουν κωδικοποιηθεί κατάλληλα (Fogel DB *et al*., 1999). Συνήθως, υπό τον όρο εξελικτικοί αλγόριθμοι βρίσκουμε τους τομείς των γενετικών αλγορίθμων, τις στρατηγικές εξέλιξης που ομαδοποιούνται υπό τον όρο εξελικτικοί αλγόριθμοι ή εξελικτικοί υπολογισμό, βρίσκουμε τους τομείς των γενετικών αλγορίθμων (Ολλανδία, 1975), τις στρατηγικές εξέλιξης (Back T *et al*., 1996), τον εξελικτικό προγραμματισμό, το γενετικό προγραμματισμό (Smolinski TG *et al*., 2008). Όλοι μοιράζονται μια κοινή εννοιολογική βάση μέσω διαδικασιών επιλογής, μετάλλαξης και αναπαραγωγής. ΑΗ ΕΑ ασχολούνται με παράμετρους πεπερασμένου μήκους, οι οποίες κωδικοποιούνται με πεπερασμένο αλφάβητο, αντί να χειρίζονται άμεσα τις ίδιες τις παραμέτρους.

Στην τεχνητή νοημοσύνη, ένας εξελικτικός αλγόριθμος (ΕΑ) είναι ένα υποσύνολο του εξελικτικού υπολογισμού, ένας γενικός πληθυσμιακός αλγόριθμος μεταευρετικής (metaheuristic) βελτιστοποίησης. Η μεταευρετική είναι μια διαδικασία υψηλότερου επιπέδου που έχει σχεδιαστεί για να βρει, να δημιουργήσει ή να επιλέξει έναν ευρετικό (μερικό αλγόριθμο αναζήτησης) που μπορεί να προσφέρει μια επαρκώς καλή λύση σε ένα

πρόβλημα βελτιστοποίησης, ειδικά με ελλιπείς ή ατελείς πληροφορίες ή περιορισμένη ικανότητα υπολογισμού. Ένας ΕΑ χρησιμοποιεί μηχανισμούς εμπνευσμένους από τη βιολογική εξέλιξη, όπως η αναπαραγωγή, η μετάλλαξη, ο ανασυνδυασμός και η επιλογή. Οι υποψήφιες λύσεις στο πρόβλημα βελτιστοποίησης παίζουν το ρόλο των ατόμων σε έναν πληθυσμό και η λειτουργία φυσικής κατάστασης καθορίζει την ποιότητα των λύσεων (βλ. Επίσης τη λειτουργία απώλειας). Η εξέλιξη του πληθυσμού πραγματοποιείται στη συνέχεια μετά την επανειλημμένη εφαρμογή των παραπάνω φορέων εκμετάλλευσης.

Γενετικός Αλγόριθμος

Ο γενετικός αλγόριθμος [Genetic Algorithm (GA)] είναι αρχικά μια τεχνική προσαρμοστικής αναζήτησης που εισήχθη από την Ολλανδία (Mohamed AR *et al*., 2009). Είναι μια στοχαστική μέθοδος αναζήτησης για την επίλυση βέλτιστων λύσεων μέσα σε μεγάλους και περίπλοκους χώρους αναζήτησης. Ο γενετικός αλγόριθμος λειτουργεί σε ένα σύνολο ατόμων που ονομάζονται πληθυσμός, όπου κάθε άτομο είναι μια κωδικοποίηση των δεδομένων εισόδου του προβλήματος και καλούνται χρωμοσώματα. Κάθε χρωμόσωμα αποτελείται από γονίδια, το καθένα από τα οποία έχει δυαδική αξία που δείχνει την παρουσία ή όχι ενός συγκεκριμένου στοιχείου του σετ. Η αναζήτηση της καλύτερης λύσης καθοδηγείται από μια αντικειμενική λειτουργία που ονομάζεται λειτουργία φυσικής κατάστασης. Οι επιλεγμένες λύσεις υψηλότερης ικανότητας είναι αυτές που έχουν μεγαλύτερη ικανότητα παραγωγής νέων λύσεων. Η λειτουργία Fitness ελέγχει την επιλογή της καλύτερης λύσης και παρέχει κριτήρια για την αξιολόγηση των υποψήφιων ατόμων. Γενικά, η GA περιλαμβάνει τρεις θεμελιώδεις φορείς: επιλογή, διασταύρωση και μετάλλαξη εντός των χρωμοσωμάτων. Ένας πληθυσμός δημιουργείται από μια ομάδα τυχαίων ατόμων. Δύο άτομα επιλέγονται για την επόμενη γενιά βάσει της ικανότητάς τους. Crossover είναι μια διαδικασία που έχει σαν αποτέλεσμα τον ανασυνδυασμό των συμβολοσειρών bit μέσω μια ανταλλαγής τμημάτων μεταξύ ζευγών χρωμοσωμάτων για τη δημιουργία νέων ατόμων. Τέλος, η μετάλλαξη έχει ως αποτέλεσμα την εξασφάλιση ότι όλα τα πιθανά χρωμοσώματα είναι προσβάσιμα ή ότι ένας ορισμένος αριθμός γενεών έχουν περάσει. Πρέπει να σημειώσουμε ότι τα GA είναι υπολογιστικό μοντέλο που σχεδιάστηκε για να προσομοιώνει τις εξελικτικές διαδικασίες στη φύση (Hinton GE *et al*., 2006).

Τεχνητό ανοσοποιητικό σύστημα

Τα τεχνητά ανοσοποιητικά συστήματα [Artificial Immune Systems (AISs)] εμπνεύστηκαν από το ανθρώπινο ανοσοποιητικό σύστημα [Human Immune System, HIS)], το οποίο είναι ανθεκτικό, αποκεντρωμένο, ανεκτικό σε σφάλματα και προσαρμοστικό. Το HIS έχει διαφορετικά κελιά με τόσα διαφορετικά καθήκοντα, έτσι ώστε οι προκύπτοντες αλγόριθμοι δίνουν διαφορετικά επίπεδα πολυπλοκότητας και μπορούν να ολοκληρώσουν ένα φάσμα εργασιών. Εμφανίστηκε στη δεκαετία του 1990 ως ένας νέος υπολογιστικός ερευνητικός χώρος και συνδέει διάφορα αναδυόμενα υπολογιστικά πεδία εμπνευσμένα από βιολογική

συμπεριφορά, όπως τεχνητά νευρωνικά δίκτυα και τεχνητή ζωή. Υπάρχουν διάφορα μοντέλα AIS που χρησιμοποιούνται στην εξερεύνηση σχεδιασμού αλληλουχιών συναρμολόγησης, στην πρόβλεψη του κυτταρικού εντοπισμού της πρωτεΐνης και σε διάφορες άλλες εφαρμογές στον τομέα της βιοπληροφορικής.

CI στην ομαδοποίηση των γονιδιακών εκφράσεων

Οι γονιδιακές εκφράσεις περιλαμβάνουν μια διαδικασία μέσω της οποίας η κωδικοποιημένη πληροφορία ενός γονιδίου μετατρέπεται σε δομές που λειτουργούν στο κύτταρο. Παρέχει τα φυσικά στοιχεία ότι ένα γονίδιο έχει ενεργοποιηθεί. Τα εκφρασμένα γονίδια περιλαμβάνουν αυτά που μεταγράφονται σε m-RNA και στη συνέχεια μεταφράζονται σε πρωτεΐνη και εκείνα που μεταγράφονται σε RNA αλλά δεν μεταφράζονται σε πρωτεΐνη (Luscombe et al., 2001). Τα επίπεδα έκφρασης χιλιάδων γονιδίων μπορούν να μετρηθούν ταυτόχρονα χρησιμοποιώντας τη σύγχρονη τεχνολογία των μικροσυστοιχειών (microarrays) (Aaron MN *et al*., 2010). Οι πρόσφατα αναπτυγμένες CI προσεγγίσεις για τις εκφράσεις γονιδίων περιλαμβάνουν τα νευρωνικά δίκτυα (Yuhui Y *et al*., 2002), τα ασαφή σύνολα (Arima *et al*., 2008), τα ακατέργαστα σύνολα (rough sets) (Wroblewski J, *et al*., 2007) και διάφορες άλλες τεχνικές (Roberto A *et al*., 2009).

Οι Arima *et al*., 2008) ανέπτυξαν ένα νευρικό δίκτυο συσσωμάτωσης στην ανάλυση δεδομένων γονιδιακής έκφρασης. Η προσέγγιση αυτή αξιολογεί τις ομοιότητες μεταξύ δύο οποιωνδήποτε γονιδίων μέσω των αλληλεπιδράσεων μιας ομάδας δειγμάτων γονιδίων. Παρέχει πιο ισχυρή απόδοση σε σύγκριση με τις ομοιότητες που εκτιμώνται από τις άμεσες αποστάσεις. Η απόδοση της αναπτυγμένης προσέγγισης έχει δοκιμαστεί στο σύνολο δεδομένων της λευχαιμίας και τα αποτελέσματα καταδεικνύουν ότι τα συσσωματωμένα νευρικά δίκτυα μπορούν να επιτύχουν καλές επιδόσεις σε δεδομένα μεγάλης διαστάσεως. Ανέφεραν επίσης ότι η απόδοση μπορεί να ενισχυθεί περαιτέρω όταν ενσωματωθούν ορισμένες χρήσιμες μεθοδολογίες επιλογής χαρακτηριστικών. Οι Mahanta MS *et al*., περιγράφουν με λεπτομέρεια διάφορες τεχνικές εξαγωγής χαρακτηριστικών.

Στην βιοπληροφορική, έχουν προταθεί διάφορες προσεγγίσεις ομαδοποίησης για την ανάλυση δεδομένων γονιδιακής έκφρασης (Yu Z *et al*., 2007, Roberto A *et al*., 2009). Οι Roberto *et al*., (2009), πρότειναν νέες στρατηγικές και ανάπτυξη μεθόδων ομαδοποίησης που συνδυάζουν την ακρίβεια και την αποτελεσματικότητα των τεχνικών ομαδοποίησης συμπλεγμάτων με βάση τις τυχαίες προβολές, με την εκφραστική ικανότητα των ασαφών συνόλων για να αποκτήσουν αξιόπιστους αλγόριθμους ομαδοποίησης. Τα πειραματικά αποτελέσματα δείχνουν ότι η προτεινόμενη προσέγγιση ασαφούς συνόλου είναι ανταγωνιστική με άλλες μεθόδους συνόλων και μπορεί να εφαρμοστεί με επιτυχία στην ανάλυση δεδομένων γονιδιακής έκφρασης.

CI στην επιλογή γονιδίων

Η επιλογή γονιδίων που παρέχουν πληροφορίες από τα τεράστια δεδομένα γονιδιακής έκφρασης των μικροσυστοιχιών είναι ένα σημαντικό και δύσκολο ερευνητικό θέμα βιοπληροφορικής (Banerjee M *et al*., 2007, Li D *et al*., 2006). Οι Fernando D *et al*., καταδεικνύουν πώς ένας αλγόριθμος ελεγχόμενου ασαφούς προτύπου μπορεί να χρησιμοποιηθεί για τη μείωση των δεδομένων μικροσυστοιχίας του DNA σε πραγματικές τιμές. Τα οφέλη της μεθόδου αυτής μπορούν να χρησιμοποιηθούν για να βρεθούν βιολογικά σημαντικές πληροφορίες σχετικές με σημαντικά γονίδια, προκειμένου να βελτιωθούν οι προηγούμενες επιτυχημένες τεχνικές. Πειραματικά αποτελέσματα σε διάγνωση οξείας μυελογενούς λευχαιμίας δείχνουν την αποτελεσματικότητα της προτεινόμενης προσέγγισης.

Οι Khan J *et al*., ανέπτυξαν μια μέθοδο τεχνητών νευρωνικών δικτύων για την ταξινόμηση καρκίνων σε συγκεκριμένες διαγνωστικές κατηγορίες με βάση τις γονιδιακές τους εκφράσεις. Στη συνέχεια, εκπαίδευσαν τα δίκτυα χρησιμοποιώντας μικρούς, στρογγυλεμένους κυανοκυτταρικούς όγκους ως μοντέλο. Αυτοί οι καρκίνοι ανήκουν σε τέσσερις ξεχωριστές διαγνωστικές κατηγορίες και συχνά παρουσιάζουν διαγνωστικά διλήμματα στην κλινική πρακτική. Το νευρωνικό δίκτυο ταξινόμησε σωστά όλα τα δείγματα και ταυτοποίησε τα γονίδια που σχετίζονται περισσότερο με την ταξινόμηση. Αυτή η μελέτη υπογραμμίζει τις πιθανές εφαρμογές της αναπτυγμένης μεθόδου για τη διάγνωση του όγκου και την ταυτοποίηση των υποψήφιων στόχων για θεραπεία.

Επίσης πολλές πετυχημένες μελέτες αναφέρονται στο πρόβλημα της εξόρυξης και επιλογής γονιδίων (Banerjee M *et al*., 2007, Li D *et al*., 2006)

CI σε ευθυγράμμιση πολλών ακολουθιών

Η ευθυγράμμιση της αλληλουχίας αναφέρεται στη διαδικασία τακτοποίησης των πρωτογενών αλληλουχιών του DNA, του RNA ή της πρωτεΐνης για τον προσδιορισμό περιοχών ομοιότητας που μπορεί να είναι συνέπεια λειτουργικών, δομικών ή εξελικτικών σχέσεων μεταξύ των αλληλουχιών. Λαμβάνοντας υπόψη δύο αλληλουχίες X και Y, μια ευθυγράμμιση με ζεύγη δείχνει θέσεις κάθε ακολουθίας που θεωρούνται ότι είναι λειτουργικά ή εξελικτικά σχετικές. Από μια οικογένεια S = (S0, S1, ..., S_{N-1}) των ακολουθιών N, θα θέλαμε να μάθουμε τα κοινά πρότυπα αυτής της οικογένειας. Δεδομένου ότι η ευθυγράμμιση κάθε ζεύγους ακολουθιών από το S ξεχωριστά συχνά δεν αποκαλύπτει τις κοινές πληροφορίες, είναι απαραίτητο να εκτελεστεί πολλαπλή ευθυγράμμιση ακολουθίας [Multiple Sequence Alignment (MSA)]. Σε γενικές γραμμές, το σύνολο των εισερχόμενων αλληλουχιών θεωρείται ότι έχουν μια εξελικτική σχέση με την οποία μοιράζονται μια σύνδεση και προέρχονται από έναν κοινό πρόγονο (Chen Y *et al*., 2006). Για να αξιολογηθεί η ποιότητα μιας ευθυγράμμισης, μια δημοφιλής επιλογή είναι να χρησιμοποιηθεί το άθροισμα των ζευγών (SP) σαν μέθοδος βαθμολόγησης. Η βαθμολογία SP συνθέτει βασικά τα αποτελέσματα υποκατάστασης όλων των πιθανών συνδυασμών ζεύγους των χαρακτήρων της αλληλουχίας σε μία στήλη μιας ευθυγράμμισης πολλαπλών ακολουθιών.

Η προοδευτική ευθυγράμμιση είναι μια ευρέως χρησιμοποιούμενη ευριστική μέθοδος MSA που δεν εγγυάται κανένα επίπεδο βελτιστοποίησης. Η Clustal W (Thompson JD *et al*., (1994), είναι ένα άλλο ευρέως δημοφιλές πρόγραμμα με κύριο μειονέκτημα όμως ότι μόλις ευθυγραμμιστεί μια ακολουθία, αυτή η ευθυγράμμιση δεν μπορεί ποτέ να τροποποιηθεί ακόμα και αν έρχεται σε σύγκρουση με τις ακολουθίες που προστέθηκαν αργότερα.

Οι Fasheng X *et al*., σχεδίασαν ένα βελτιωμένο PSO (Particle Swarm Optimization) για την επίλυση του MSA. Στον αλγόριθμο, κάθε σωματίδιο αντιπροσωπεύει μια ευθυγράμμιση και πηγαίνει στο σωματίδιο το οποίο έχει την καλύτερη λύση, προκειμένου να επεκταθεί η ποικιλομορφία του αλγορίθμου και να ενισχυθεί η δυνατότητα εύρεσης της βέλτιστης λύσης. Η προτεινόμενη προσέγγιση δοκιμάστηκε χρησιμοποιώντας κάποιες οικογένειες πρωτεϊνών και συγκρίθηκε με τις ευθυγραμμίσεις που δημιουργήθηκαν από τον αλγόριθμο Clustal X 2.0.

Συμπεράσματα και μελλοντικές κατευθύνσεις

Ένας συνδυασμός διάφορων τεχνολογιών υπολογιστικής νοημοσύνης στη βιοπληροφορική έχει γίνει ένας από τους πιο ελπιδοφόρους τρόπους διερεύνησης της βιοπληροφορικής. Από την οπτική γωνία του CI, είναι αναγκαίες περαιτέρω διερευνήσεις πιθανών υβριδοποιήσεων των τεχνολογιών CI για να δοθεί μια πληρέστερη εικόνα των εφαρμογών που βασίζονται σε CI στη βιοπληροφορική, όπως οι νευρωνικά-ασαφείς (neural-fuzzy), νευρωνικά-ακατέργαστες (neural-rough) EAs, GAs και άλλες υβριδοποιήσεις διαφορετικών τεχνολογιών CI.

Στατιστική επεξεργασία στη γενετική

Οι αρχές και οι μέθοδοι στατιστικής επεξεργασίας έχουν αποδειχθεί ιδιαίτερα χρήσιμες στο χώρο της γενετικής, παρέχοντας το υπόβαθρο για την εξαγωγή συμπερασμάτων από τον τεράστιο όγκο των γενετικών δεδομένων. Στις μελέτες γενομικής εμπλέκονται αρκετοί **τύποι δεδομένων**, που μπορούν να κατηγοριοποιηθούν ως συνεχείς, διακριτές και κατηγορικές μεταβλητές. Οι συνεχείς μεταβλητές, όπως η ηλικία, θεωρητικά μπορούν να πάρουν οποιαδήποτε πραγματική τιμή, αν και σε πολλές εφαρμογές παίρνουν οποιεσδήποτε τιμές εντός ενός ορισμένου διαστήματος. Οι διακριτές μεταβλητές παίρνουν ακέραιες τιμές, με παράδειγμα το πλήθος των φαρμάκων υπό διερεύνηση. Οι κατηγορικές, ή αλλιώς ποιοτικές, μεταβλητές παίρνουν ποιοτικές τιμές υπό τη μορφή κατηγοριών, αν και συχνά ανατίθεται μία αριθμητική ψευδοτιμή ανά κατηγορία για τις ανάγκες της ανάλυσης. Χαρακτηρίζονται ως διατάξιμες αν οι τιμές τους ακολουθούν μία εγγενή διάταξη ή ακολουθία, όπως για παράδειγμα ο χαρακτηρισμός της υγείας ενός ασθενή, ή μη διατάξιμες στην αντίθετη περίπτωση, όπως για παράδειγμα το φύλο και η φυλή. (Kapur, 2017)

Η στατιστική σημαντικότητα ενός συμπεράσματος εξετάζεται με βάση την απόδειξη ότι το συμπέρασμα αυτό δε θα μπορούσε να προκύψει από τύχη ή λάθος, για την υλοποίηση της οποίας επιστρατεύεται ένα πιθανοτικό μοντέλο. Στα πλαίσια της απόδειξης ενός στατιστικού

συμπεράσματος, κεντρικό ρόλο έχει η έννοια της μηδενικής υπόθεσης (null hypothesis, H_0), που μπορεί να οριστεί ως η γενική τοποθέτηση ότι δεν υπάρχει συσχέτιση μεταξύ δύο μετρούμενων φαινομένων ή δύο ομάδων, καθώς και η έννοια της p-value, που μπορεί να οριστεί ως η πιθανότητα εμφάνισης των παρατηρούμενων δεδομένων ή σχέσης στην περίπτωση που ισχύει η μηδενική υπόθεση. Αν η p-value προκύπτει αρκετά μικρή, το συμπέρασμα υπό διερεύνηση θεωρείται στατιστικά σημαντικό, με το κατώφλι 0.05 να είναι πλέον κοινώς αποδεκτό για τον έλεγχο της μηδενικής υπόθεσης. Επιπλέον, συμπληρωματική της μηδενικής υπόθεσης είναι η έννοια της εναλλακτικής υπόθεσης (alternative hypothesis, H_1), η οποία γίνεται δεκτή στην περίπτωση που απορριφθεί η μηδενική υπόθεση. (Smith, 2005; Kapur, 2017)

Μία πιο εύκολα υλοποιήσιμη στρατηγική είναι η δειγματοθέτηση (bootstrapping), το πλαίσιο εφαρμογής της οποίας μπορεί να περιγραφεί ως η ύπαρξη δύο μελών ενός πληθυσμού με μικρή απόσταση μεταξύ τους, των οποίων η συσχέτιση πρέπει να διαπιστωθεί. Σύμφωνα με τη μέθοδο δειγματοθέτησης, οι τιμές δεδομένων του πληθυσμού θεωρούνται ότι προκύπτουν τυχαία. Για να εκτιμηθεί η στατιστική σημαντικότητα του εν λόγω συμπεράσματος, οι τιμές ανακατατάσσονται τυχαία αρκετές φορές, κάθε φορά υπολογίζοντας την απόσταση μεταξύ των πληθυσμών υπό διερεύνηση και σημειώνοντας το ποσοστό των φορών που είναι μικρότερη ή ίση με την απόσταση υπό διερεύνηση, το οποίο συνιστά την τιμή p-value. (Smith, 2005)

Η απόδοση των στατιστικών ελέγχων υποθέσεων αξιολογείται με βάση τα σφάλματα τύπου I και II. Το σφάλμα τύπου I, ή αλλιώς τύπου α, ορίζεται ως η πιθανότητα απόρριψης της μηδενικής υπόθεσης στην περίπτωση που είναι αληθής, και είναι η p-value. Αν, δηλαδή, τα δείγματα υπό μελέτη προήλθαν από την κατανομή της μηδενικής υπόθεσης, το σφάλμα τύπου I είναι η πιθανότητα του εσφαλμένου συμπεράσματος ότι τα δείγματα προήλθαν από την εναλλακτική υπόθεση. Αντίθετα, το σφάλμα τύπου II, ή αλλιώς τύπου β, είναι η πιθανότητα αποδοχής της μηδενικής υπόθεσης στην περίπτωση που η εναλλακτική υπόθεση είναι αληθής. Με άλλα λόγια, αν τα δεδομένα προέρχονται από την κατανομή της εναλλακτικής υπόθεσης, το σφάλμα τύπου II είναι η πιθανότητα του εσφαλμένου συμπεράσματος ότι τα δείγματα αποκτήθηκαν από τη μηδενική υπόθεση. Το σφάλμα τύπου I αξιοποιείται κυρίως στον στατιστικό έλεγχο υποθέσεων με κατώφλι που συνήθως ορίζεται στην τιμή 0.05, ενώ το σφάλμα τύπου II χρησιμοποιείται περισσότερο για την επιλογή του μεγέθους του δείγματος με κατώφλι που συνήθως ορίζεται στην πιο ελαστική τιμή 0.2. (Kapur, 2017)

Οι μελέτες με αντικείμενο τη στατιστική ανάλυση γενετικών δεδομένων μπορούν να κατηγοριοποιηθούν με κριτήριο τη στρατηγική τους, με βασικές κατηγορίες να είναι οι μέθοδοι που βασίζονται στον υπολογισμό αποστάσεων και οι μέθοδοι που βασίζονται στην εφαρμογή μοντέλων. Οι τεχνικές υπολογισμού αποστάσεων έχουν ως βασική μεθοδολογία τον υπολογισμό των αποστάσεων μεταξύ όλων των ζευγών ατόμων του πληθυσμού υπό εξέταση και τη διαμόρφωση ενός πίνακα κατά ζεύγη αποστάσεων. Παραδείγματα τέτοιων τεχνικών είναι η πολυδιάστατη κλιμακοποίηση και η ανάλυση κατά ομάδες. Η μέθοδος της πολυδιάστατης κλιμακοποίησης (Multidimensional scaling, MDS) έχει ως στόχο την εύρεση μίας διάταξης σημείων λιγότερων διαστάσεων, των οποίων οι κατά ζεύγη αποστάσεις είναι οι κοντινότερες δυνατές στις αποστάσεις των αρχικών σημείων περισσότερων διαστάσεων.

Οι διαφορετικοί τρόποι μέτρησης της εγγύτητας αυτής διαμορφώνουν τις διαφορετικές παραλλαγές της MDS. Η ανάλυση κατά ομάδες (Cluster Analysis) είναι μία μέθοδος με στόχο την κατάταξη των υπαρχουσών παρατηρήσεων σε ομάδες αξιοποιώντας την πληροφορία που εμπεριέχεται σε συγκεκριμένες μεταβλητές. Παράγει μία δενδρική δομή που αντανακλά την εγγενή οργάνωση των δεδομένων, με τα δεδομένα να θεωρούνται ως τα φύλλα του δέντρου και τα κλαδιά να αντιπροσωπεύουν τις σχέσεις μεταξύ των δεδομένων. Το ύψος κάθε κλαδιού αποτελεί έναν τρόπο μέτρησης της απόστασης μεταξύ των εμπλεκόμενων σε αυτό δεδομένων, και η επιλογή του τρόπου υπολογισμού της απόστασης εξαρτάται από την επιλεγμένη μέθοδο και τον τύπο των δεδομένων. (Smith, 2005; Tsetsos *et al*., 2018)

Δύο πολύ διαδεδομένες μέθοδοι που βασίζονται στην εφαρμογή μοντέλων είναι η ανάλυση κύριων συνιστωσών και η παραγοντική ανάλυση. Η ανάλυση κύριων συνιστωσών (Principal Component Analysis, PCA) έχει ως στόχο την απλοποίηση των δεδομένων πολλών διαστάσεων διατηρώντας παράλληλα τα μοτίβα που τα διέπουν. Ο στόχος αυτός επιτυγχάνεται με το μετασχηματισμό των αρχικών δεδομένων σε δεδομένα λιγότερων διαστάσεων, τα οποία λειτουργούν ως συνοπτικά χαρακτηριστικά των αρχικών. Συγκεκριμένα, τα αρχικά δεδομένα προβάλλονται γεωμετρικά σε λιγότερες διαστάσεις, οι οποίες ονομάζονται κύριες συνιστώσες, με σκοπό την εύρεση της καλύτερης επιτομής των δεδομένων χρησιμοποιώντας έναν περιορισμένο αριθμό κύριων συνιστωσών. Η πρώτη κύρια συνιστώσα επιλέγεται με γνώμονα την ελαχιστοποίηση της ολικής απόστασης μεταξύ των δεδομένων και της προβολής τους στην κύρια συνιστώσα. Οι επόμενες κύριες συνιστώσες επιλέγονται με την ίδια λογική, με την επιπλέον απαίτηση να μην υπάρχει συσχέτιση με όλες τις προηγούμενες. (Lever *et al*., 2017)

Η παραγοντική ανάλυση (Factor Analysis) είναι μία μέθοδος που μπορεί να θεωρηθεί ως προέκταση της ανάλυσης κύριων συνιστωσών. Στόχος της είναι η περιγραφή των σχέσεων μεταξύ ενός συνόλου παρατηρούμενων μεταβλητών με ένα μικρότερο πλήθος μη παρατηρήσιμων μεταβλητών, που ονομάζονται παράγοντες, καθώς και η ομαδοποίηση των αρχικών δεδομένων με βάση τους εξαγόμενους αυτούς παράγοντες. Η παραγοντική ανάλυση στηρίζεται στην ιδέα ότι η συσχέτιση μεταξύ δύο μεταβλητών μπορεί να οφείλεται σε μία τρίτη μεταβλητή που σχετίζεται και με τις δύο, παρά σε μία πραγματική αιτία. Σε αυτή την περίπτωση, η παρουσία αυτής της μεταβλητής προκαλεί μία ψευδή συσχέτιση μεταξύ των δύο πρώτων. Οι μέθοδοι παραγοντικής ανάλυσης διακρίνονται στη διερευνητική παραγοντική ανάλυση (exploratory factor analysis, EFA) και την επιβεβαιωτική παραγοντική ανάλυση (confirmatory factor analysis, CFA). Η πρώτη χρησιμοποιείται όταν δεν υπάρχει εκ των προτέρων γνώση σχετικά με τη δομή των δεδομένων υπό διερεύνηση, ή όταν δεν απαιτείται να προκύψει κάποια δομή από τους αναλυτικούς υπολογισμούς. Η επιβεβαιωτική ανάλυση προϋποθέτει ο ερευνητής να έχει γνώση ή θεωρία σχετικά με τη δομή των παραγόντων, η οποία χρησιμοποιείται ως είσοδος στην ανάλυση για τον περιορισμό των παραμέτρων προς υπολογισμό. (Alkharkhi *et al*., 2019; Thompson, 2007; Bartholomew, 2015)

Οι έννοιες της παλινδρόμησης και των γραμμικών μοντέλων είναι έννοιες με ιδιαίτερη σημασία όσον αφορά τη στατιστική επεξεργασία των γενετικών δεδομένων. Η παλινδρόμηση έχει ως βασικό αντικείμενο την πρόβλεψη της τιμής ή τάξης μίας εξαρτώμενης μεταβλητής y, ή αλλιώς μεταβλητής απόκρισης, από τις τιμές ενός συνόλου ανεξάρτητων μεταβλητών X, ή αλλιώς μεταβλητών πρόβλεψης, με βάση μία συνάρτηση $\mathrm{y} = \mathrm{f}(\mathrm{X})$. Περιλαμβάνει αρκετές

μεθόδους, με πιο συνήθεις τη γραμμική και τη λογιστική παλινδρόμηση, και χρησιμοποιείται σε ένα μεγάλο εύρος επιστημονικών εφαρμογών, ανάμεσα στις οποίες είναι και οι εφαρμογές γενετικής. Στο πλαίσιο της εφαρμογής της παλινδρόμησης, ιδιαίτερη σημασία έχει η επιλογή των χαρακτηριστικών των δεδομένων που θα συμπεριληφθούν στο μοντέλο. Στις μελέτες με γενετικά δεδομένα, ένα χαρακτηριστικό μπορεί να έχει ποιοτικό χαρακτήρα, αναφερόμενο στην κατάσταση της ασθένειας, ή ποσοτικό χαρακτήρα, όπως ένας ενδιάμεσος βιοδείκτης που σχετίζεται με την ασθένεια, για παράδειγμα το επίπεδο της LDL χοληστερόλης. Η επιλογή του χαρακτηριστικού επηρεάζει άμεσα την επιλογή της κατάλληλης μεθόδου στατιστικής ανάλυσης, με τη λογιστική παλινδρόμηση να εφαρμόζεται στην περίπτωση των ποιοτικών χαρακτηριστικών και τη γραμμική παλινδρόμηση στην περίπτωση των ποσοτικών χαρακτηριστικών. Έχει παρατηρηθεί ότι τα ποσοτικά χαρακτηριστικά οδηγούν σε πιο αξιόπιστα αποτελέσματα. (Kotu, 2019; Fitzmaurice, 2016; Wang *et al*., 2018)
Η γραμμική παλινδρόμηση περιγράφει πώς οι μέσες ποσοτικές τιμές της μεταβλητής απόκρισης Y προκύπτουν ως γραμμική συνάρτηση ενός συνόλου μεταβλητών πρόβλεψης X. Το μοντέλο γραμμικής παλινδρόμησης μπορεί να αναπαρασταθεί μαθηματικά ως

$$\mathrm{Y_i} = \beta_0 + \beta_1 \mathrm{X_{1i}} + \beta_2 \mathrm{X_{2i}} + \cdots + \beta_\mathrm{P} \mathrm{X_{pi}} + \mathrm{e_i}, \mathrm{i} = 1, \ldots, \mathrm{n} \Leftrightarrow \mathrm{Y} = \mathrm{X}\beta + \mathrm{e}$$

όπου n το πλήθος των ατόμων και p το πλήθος των μεταβλητών πρόβλεψης X. Οι συντελεστές β εκφράζουν την εξάρτηση της μεταβλητής απόκρισης από τις μεταβλητές πρόβλεψης. Οι όροι e αντιπροσωπεύουν το τυχαίο σφάλμα, και θεωρείται ότι ακολουθούν κανονική κατανομή με μέση τιμή 0 και διακύμανση σ^2. Αντίθετα, η λογιστική παλινδρόμηση περιγράφει πώς η τιμή μίας δυαδικής μεταβλητής απόκρισης προκύπτει ως συνάρτηση ενός συνόλου μεταβλητών πρόβλεψης X, και αναπαριστάται ως

$$\mathrm{Y_i} = \log\left(\frac{\mathrm{P}(\mathrm{z_i} = 1)}{\mathrm{P}(\mathrm{z_i} = 0)}\right) = \beta_0 + \beta_1 \mathrm{X_{1i}} + \beta_2 \mathrm{X_{2i}} + \cdots + \beta_\mathrm{P} \mathrm{X_{pi}} + \mathrm{e_i}, \mathrm{i} = 1, \ldots, \mathrm{n}$$

όπου $\mathrm{z_i}$ η δυαδική τιμή της παρατηρούμενης εξόδου του ατόμου i, και $\mathrm{P}(\mathrm{z_i} = 1), \mathrm{P}(\mathrm{z_i} = 0)$ η πιθανότητα επιτυχίας ή αποτυχίας, αντίστοιχα. (Fitzmaurice, 2016; Kapur, 2017)
Στα πλαίσια των γενετικών μελετών συχνά αξιοποιούνται παραλλαγές των παραπάνω μοντέλων. Μία από αυτές είναι το γενικευμένο γραμμικό μοντέλο (generalized linear model, GLM) που μπορεί να αναπαρασταθεί ως:

$$\mathrm{g}(\mu) = \sum_\mathrm{j} \beta_\mathrm{j} \mathrm{X_j} + \mathrm{uG} + \sum_\mathrm{k} \gamma_\mathrm{k} \mathrm{PC_k}$$

όπου $\mu = \mathrm{E}(\mathrm{Y})$ και g η συνάρτηση σύνδεσης που πραγματοποιεί έναν μονότονο μετασχηματισμό στη μέση τιμή της μεταβλητής απόκρισης. Το δεξί μέρος της εξίσωσης αποτελεί γραμμικό συνδυασμό των ανεξάρτητων μεταβλητών, με το $\mathrm{X_j}$ να είναι η j-οστή συμμεταβλητή που αντιπροσωπεύει έναν κλινικό ή περιβαλλοντικό παράγοντα κινδύνου, β_j ο συντελεστής της $\mathrm{X_j}$, G ο γενότυπος της γενετικής παραλλαγής υπό διερεύνηση με συντελεστή u, $\mathrm{PC_k}$ η k-οστή κύρια συνιστώσα όπως υπολογίζεται από τον πίνακα γενοτύπων, και γ_k ο συντελεστής της $\mathrm{PC_k}$. Στην περίπτωση που το Y είναι μία συνεχής τυχαία μεταβλητή, το μοντέλο ανάγεται σε κλασική γραμμική παλινδρόμηση, ενώ αν το Y είναι δυαδική μεταβλητή, μπορεί να εφαρμοστεί η συνάρτηση logit και το παραπάνω μοντέλο ανάγεται σε λογιστική παλινδρόμηση. Το γενικευμένο γραμμικό μοντέλο μπορεί να εφαρμοστεί όταν

πρέπει να ελεγχθούν μεταβλητές όπως η ηλικία, το γένος και η θεραπευτική αγωγή εκτός από την πληθυσμιακή δομή. (Wang *et al*., 2018)

Το γραμμικό μικτό μοντέλο (linear mixed model, LMM) είναι ένας διαδεδομένος τρόπος μελέτης της συμμεταβολής που προκύπτει από σύνθετες δομές συσχέτισης. Το γραμμικό μικτό μοντέλο αντιμετωπίζει τη γενετική παραλλαγή υπό εξέταση και τις κλινικές και περιβαλλοντικές συμμεταβλητές ως μεταβλητές σταθερών επιδράσεων, και τους υπόλοιπους γενότυπους ως μεταβλητές τυχαίας επίδρασης. Η γενική μορφή αναπαράστασης του γραμμικού μικτού μοντέλου είναι η

$$\Upsilon = X\beta + Zu + e,$$

όπου n το πλήθος ατόμων, Υ είναι ο πίνακας $(nx1)$ των χαρακτηριστικών ή φαινοτύπων υπό μελέτη, X ο πίνακας (nxp) των συμμεταβλητών σταθερής επίδρασης που μπορούν να επηρεάσουν τις τιμές του Υ, και β το διάνυσμα $(px1)$ συντελεστών σταθερής επίδρασης. Z είναι ο πίνακας (nxq) συμμεταβλητών τυχαίας επίδρασης, u το διάνυσμα $(qx1)$ παραμέτρων τυχαίας επίδρασης και e το διάνυσμα $(nx1)$ σφαλμάτων. Οι παράμετροι u, e θεωρούνται ότι ακολουθούν κανονική κατανομή $(u \sim N(0, \tau I_q), e \sim N(0, \sigma^2 I_n))$ και οι παράμετροι του μοντέλου προς υπολογισμό είναι οι σ, τ και β. Στις γενετικές μελέτες, οι συμμεταβλητές σταθερής επίδρασης X αντιστοιχούν σε κλινικές ή περιβαλλοντικές παραμέτρους, όπως η ηλικία, το φύλο, και η έκθεση σε κάποιον εξωτερικό παράγοντα, ή σε γενότυπους ενός πολυμορφισμού μονού νουκλεοτιδίου που εν δυνάμει σχετίζονται με τους φαινότυπους, ενώ οι συμμεταβλητές τυχαίας επίδρασης Z αντιπροσωπεύουν τους γενότυπους σε q γενετικούς τόπους. Με δεδομένο ότι το γραμμικό μικτό μοντέλο μπορεί να λάβει υπόψιν την όποια συνδιακύμανση προκύπτει από τη συσχέτιση μεταξύ των υποκειμένων υπό μελέτη, μπορεί να χρησιμοποιηθεί για τον έλεγχο συσχέτισης σε οικογενειακά δεδομένα ή δεδομένα στρωματοποιημένου πληθυσμού. (Dandine-Rouland & Perdry, 2015; Wang *et al*., 2018)

Μία δημοφιλής εφαρμογή του γραμμικού μικτού μοντέλου στη γενετική αφορά τον έλεγχο της συσχέτισης μεταξύ ενός φαινοτύπου και ενός συνόλου πολυμορφισμών μονού νουκλεοτιδίου (Single Nucleotide Polymorphisms, SNPs), ο οποίος είναι ιδιαίτερα χρήσιμος στο πλαίσιο διερεύνησης σπάνιων γενετικών παραλλαγών. Σε αυτή την περίπτωση, το γραμμικό μικτό μοντέλο διαμορφώνεται ως

$$\Upsilon = Xa + GWu + e$$

όπου n ο αριθμός ατόμων, Υ ο $(nx1)$ πίνακας φαινοτύπων τους, X ο (nxp) πίνακας συμμεταβλητών και W ο διαγώνιος (qxq) πίνακας συντελεστών, ο ρόλος του οποίου είναι να αποδώσει βάρος σε κάθε γενετική παραλλαγή. Στόχος της μελέτης αυτού του είδους είναι ο έλεγχος της συσχέτισης q σπάνιων γενετικών παραλλαγών με τους φαινότυπους Υ. Ο (nxq) πίνακας G περιλαμβάνει τους γενότυπους $g_{ij}, i = 1, \dots, n, j = 1, \dots, q$, οι οποίοι παίρνουν τιμές 0, 1 ή 2 ανάλογα με το πλήθος των αλληλόμορφων γονιδίων. Η παράμετρος τ εκφράζει τη συσχέτιση μεταξύ των πινάκων Υ, G, και η μηδενική υπόθεση είναι ότι δεν υπάρχει συσχέτιση μεταξύ τους, δηλαδή ότι η επίδραση των γενότυπων G είναι τυχαία $(\tau = 0)$. (Dandine-Rouland & Perdry, 2015)

Μία επιπλέον παραλλαγή του γραμμικού μικτού μοντέλου έχει αξιοποιηθεί στα πλαίσια της θεωρίας γενομικής κληρονομικότητας με την ακόλουθη μορφή

$$\Upsilon = 1_n\beta + Zu + e,$$

όπου n το πλήθος των (μη συγγενικών) ατόμων και q οι αυτοσωμικοί πολυμορφισμοί μονού νουκλεοτιδίου υπό διερεύνηση. Ο πίνακας Y αναφέρεται στους φαινότυπους των ατόμων και ο πίνακας Z περιλαμβάνει τους τυποποιημένους γενότυπους $z_{ij} = (g_{ij} - 2p_j)/\sqrt{2p_j(1-p_j)}$, με δεδομένο ότι οι γενότυποι g_{ij} κωδικοποιούνται με τις τιμές 0, 1, ή 2. Με βάση τα παραπάνω, η κληρονομικότητα υπολογίζεται σύμφωνα με τον τύπο $h^2 = \hat{\tau}/(\hat{\tau} + \hat{\sigma}^2)$. (Dandine-Rouland & Perdry, 2015)

Το γραμμικό μικτό μοντέλο έχει ακόμη επιστρατευθεί στις μελέτες συσχέτισης ολόκληρου του γονιδιώματος (Genome-Wide Association Studies, GWAS) με τη μορφή

$$Y = \sum_j jX_j + \sum_k v_k Z_k + uG + e,$$

όπου ο πίνακας G περιλαμβάνει τους πολυμορφισμούς μονού νουκλεοτιδίου (SNPs) υπό έλεγχο, ο πίνακας X τις μεταβλητές σταθερής επίδρασης, ο πίνακας Z τους γενότυπους τυχαίας επίδρασης και το διάνυσμα e τους όρους σφάλματος. Ο δείκτης j αντιστοιχεί στις συμμεταβλητές σταθερής επίδρασης και ο δείκτης k στους γενότυπους τυχαίας επίδρασης. Όπως προηγουμένως, οι όροι v, e θεωρούνται ότι ακολουθούν κανονική κατανομή. (Wang *et al*., 2018)

Το γραμμικό μικτό μοντέλο έχει προταθεί ως παραλλαγή της μεθόδου των κύριων συνιστωσών στις μελέτες GWAS όταν είναι απαραίτητη η ενσωμάτωση ενός μεγάλου αριθμού δειγμάτων για την κατάκτηση ικανοποιητικής στατιστικής ισχύος μέσω της ανάλυσης του πίνακα Z σε ιδιάζουσες συνιστώσες. Σε αυτή την περίπτωση, ισχύει $Z = U\Sigma L'$, όπου U ένας ορθογώνιος πίνακας nxn, Σ ένας διαγώνιος nxn πίνακας και L ένας qxn πίνακας με $L'L = I_n$. Επομένως, αν θεωρηθεί ότι $w = (w_1, \dots, w_n)' = L'u$, το γραμμικό μικτό μοντέλο μπορεί να πάρει τη μορφή

$$Y = X\beta + Zu + \varepsilon = X\beta + (U\Sigma)(L'u) + \varepsilon = X\beta + PC_1 w_1 + \cdots + PC_n w_n + \varepsilon,$$

όπου οι παράγοντες $w_1, \dots, w_n$ ακολουθούν κανονική κατανομή. Η παραπάνω εκδοχή είναι αρκετά αποτελεσματική, αφού αφ' ενός έχει την ικανότητα μοντελοποίησης του πληθυσμού μέσα από ένα μεγάλο αριθμό κύριων συνιστωσών, αφ' ετέρου αποφεύγεται το φαινόμενο σφαλματώδους υπεραρμογής (overfitting) που θα παρατηρούνταν με έναν υπερβολικά μεγάλο αριθμό k. (Dandine-Rouland & Perdry, 2015)

Τα γραμμικά μοντέλα αξιοποιούνται και στην παραγοντική ανάλυση, κύριος στόχος της οποίας είναι η περιγραφή των σχέσεων μεταξύ ενός συνόλου k παρατηρούμενων μεταβλητών με ένα μικρότερο πλήθος μη παρατηρήσιμων μεταβλητών που ονομάζονται παράγοντες, καθώς και η ομαδοποίηση των αρχικών μεταβλητών με βάση τους εξαγόμενους αυτούς παράγοντες. Αν υποτεθεί ένα τυχαίο δείγμα με k παρατηρούμενες μεταβλητές $Y_i, i = 1, \dots, k$, και m μη παρατηρήσιμες μεταβλητές $f_j, j = 1, \dots, m$, σύμφωνα με τη μέθοδο της παραγοντικής ανάλυσης κάθε παρατηρούμενη μεταβλητή Y_i μπορεί να εκφραστεί ως γραμμικός συνδυασμός των παραγόντων f_j και του σφάλματος e_i, όπως περιγράφεται από την εξίσωση

$$Y_i = \sum_{j=1}^{m} d_{ij} f_j + e_i, i = 1, \dots, k \Leftrightarrow Y = Df + e$$

Με δεδομένο ότι ένας στόχος της παραγοντικής ανάλυσης είναι η μείωση δεδομένων, πρέπει να ισχύει $m < k$. Οι συντελεστές d_{ij} εκφράζουν τη συνεισφορά ενός κοινού παράγοντα f_j σε

μία παρατηρούμενη μεταβλητή Y_i, ή αλλιώς την εξάρτηση της μεταβλητής Y_i από τον παράγοντα f_j. (Alkharkhi *et al*., 2019)

Οι προαναφερθείσες στατιστικές μέθοδοι μπορούν να εντοπιστούν στις στρατηγικές αρκετών γενετικών ερευνών. Ένα παράδειγμα είναι η έρευνα των Lane *et al*., (2016), οι οποίοι πραγματοποίησαν μελέτη συσχέτισης ολόκληρου του γονιδιώματος (GWAS) με επίκεντρο τη διάρκεια ύπνου, τα συμπτώματα αϋπνίας και την υπερβολική υπνηλία κατά τη διάρκεια της ημέρας, χρησιμοποιώντας γραμμική και λογιστική παρεμβολή. Αποτέλεσμα της μελέτης αυτής ήταν η ταυτοποίηση 9 στατιστικά σημαντικών ($P < 5 \times 10^{-8}$) γενετικών τόπων και 14 υποδηλωτικών ($5 \times 10^{-8} < P < 5 \times 10^{-7}$) γενετικών τόπων που συνδέονται με τα προβλήματα ύπνου. (Lane *et al*., 2016)

Η εφαρμογή των γραμμικών μοντέλων μπορεί να σημειωθεί στην έρευνα των Fanous *et al*., (2012), οι οποίοι πραγματοποίησαν έλεγχο συσχέτισης αλληλόμορφων γονιδίων όσον αφορά τρία χαρακτηριστικά συμπτωμάτων σε ασθενείς με σχιζοφρένεια. Συγκεκριμένα, χρησιμοποίησαν γραμμική παρεμβολή όπως υλοποιείται στο λογισμικό PLINK (Purcell *et al*., 2007) για να ελέγξουν τις επιδράσεις των αλληλόμορφων γονιδίων στις τιμές των τριών αυτών χαρακτηριστικών, με συμμεταβλητές που συμπεριλαμβάνουν την τοποθεσία της μελέτης, την ηλικία και το φύλο. (Fanous *et al.*, 2012)

Επιπλέον, αξίζει να σημειωθεί η εργασία των Bani-Fatemi *et al.,* (2016), αντικείμενο της οποίας ήταν η αναζήτηση πιθανών γενετικών δεικτών που σχετίζονται με το ενδεχόμενο απόπειρας αυτοκτονίας στους ασθενείς με σχιζοφρένεια. Μετά από τη συλλογή δειγμάτων αίματος, την εξαγωγή του γενετικού υλικού από αυτά και την ανάλυση γονιδιακής έκφρασης με το σύστημα Illumina Omnia 2.5 (Illumina, San Diego, CA, USA), οι συγγραφείς εφάρμοσαν ανάλυση κύριων συνιστωσών, οι κύριες συνιστώσες που προέκυψαν από την οποία συμπεριλήφθηκαν σε διαδικασία λογιστικής παρεμβολής. Οι 1205383 πολυμορφισμοί μονού νουκλεοτιδίου που προέκυψαν υποβλήθηκαν σε επεξεργασία με το λογισμικό PLINK (Purcell et al., 2007). Τέλος, τα δημογραφικά και κλινικά χαρακτηριστικά μεταξύ της ομάδας ασθενών που έκαναν απόπειρα αυτοκτονίας και της ομάδας ασθενών που δεν έκαναν απόπειρα συγκρίθηκαν με λογιστική παρεμβολή. (Bani-Fatemi *et al*., 2016)

Η έρευνα των Kustra *et al*. (2006) αποτελεί μία περίπτωση εφαρμογής της μεθόδου παραγοντικής ανάλυσης για την πρόβλεψη των λειτουργιών μη χαρακτηρισμένων γονιδίων, χρησιμοποιώντας δεδομένα μικροσυστοιχιών από ζυμομύκητες και λειτουργικές κατηγορίες γενετικών οντολογιών και βιολογικών διαδικασιών. Οι συγγραφείς χρησιμοποίησαν το συμβολισμό $g_i^{(j)} = \sum_{l=1}^{L} \lambda_{il} f_l^{(j)} + e_i^{(j)}$, όπου $g_i^{(j)}, i = 1, \dots, p$ είναι το i-οστό γονίδιο του j-οστού πειράματος μικροσυστοιχίας, $f_l^{(j)}, l = 1, \dots, L$ είναι οι παράγοντες που επηρεάζουν το επίπεδο έκφρασης του $g_i^{(j)}$, λ_{il} οι συντελεστές των παραγόντων, που εκτιμήθηκαν με ανάλυση κύριων παραγόντων, και $e_i^{(j)}$ ο όρος σφάλματος. (Kustra *et al*., 2006)

Ένα πιο πρόσφατο παράδειγμα είναι η μελέτη των Brown et al. (2015), στην οποία η παραγοντική ανάλυση αξιοποιήθηκε για την εξερεύνηση συσχετίσεων μεταξύ της γενετικής έκφρασης, της κληρονομικότητας και της γήρανσης. Συγκεκριμένα, οι συγγραφείς χρησιμοποίησαν δείγματα δερματικού ιστού προερχόμενα από δίδυμα άτομα. Τα δείγματα υποβλήθηκαν σε ανάλυση γονιδιακής έκφρασης με το σύστημα Illumina Human HT-12 V3 BeadChips και ακολούθως σε ποιοτικό έλεγχο, καταλήγοντας στα δεδομένα 647 ατόμων. Τα

δεδομένα γενετικής έκφρασης υποβλήθηκαν σε παραγοντική ανάλυση με το λογισμικό PEER (Parts *et al*., 2011) καταλήγοντας σε πέντε κοινούς συνολικούς παράγοντες, οι οποίοι αφαιρέθηκαν από τα δεδομένα. Στη συνέχεια, οι συγγραφείς ομαδοποίησαν τα γονίδια με βάση 186 μεταβολικά μονοπάτια και εξήγαγαν με παραγοντική ανάλυση νέους φαινοτυπικούς παράγοντες, εξειδικευμένους για κάθε μονοπάτι. Ακολούθως, πραγματοποιήθηκαν έλεγχοι συσχέτισης μεταξύ κάθε φαινοτυπικού παράγοντα και της ηλικίας, καθώς και μεταξύ μοναδικών γονιδίων και της ηλικίας χρησιμοποιώντας το λογισμικό lme4 (Bates *et al*. 2014). Αποτέλεσμα αυτής της διαδικασίας ήταν να προκύψουν 69 συσχετισμοί από 57 μονοπάτια. (Brown *et al*., 2015)

Τέλος, σύμφωνα με την εργασία των Iacob et al. (2016), προτάθηκε η εφαρμογή ενός μοντέλου διερευνητικής παραγοντικής ανάλυσης σε γενετικά δεδομένα με στόχο τη διερεύνηση πιθανών συσχετίσεων μεταξύ 34 υποψήφιων γονιδίων και της ινομυαλγίας, του συνδρόμου χρόνιας κόπωσης και της κατάθλιψης. Τα γενετικά δεδομένα που εξετάστηκαν προήλθαν από mRNA λευκοκυττάρων 261 ατόμων. Τα 34 υποψήφια γονίδια υποβλήθηκαν σε διερευνητική παραγοντική ανάλυση με το λογισμικό STATA 13.0 (StataCorp, 2013) αναδεικνύοντας 4 ανεξάρτητους παράγοντες που χαρακτηρίστηκαν με βάση το λειτουργικό υπόβαθρό τους. Στη συνέχεια εφαρμόστηκε γραμμική παρεμβολή για τη διερεύνηση της συσχέτισης αυτών των παραγόντων με τις ασθένειες υπό εξέταση, από την οποία προέκυψε ότι δύο από τους τέσσερις παράγοντες παρουσιάζουν θετική συσχέτιση με το σύνδρομο χρόνιας κόπωσης και αρνητική συσχέτιση με τη σοβαρότητα κατάθλιψης, αλλά όχι με τη ινομυαλγία. (Iacob *et al*., 2016)

Βιβλιογραφία

Aaron, M.N., James, B.C., 2010. AutoSOME: a clustering method for identifying gene expression modules without prior knowledge of cluster number. BMC Bioinformatics, 11–117

Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol 2015;33:1–9.

Alkarkhi, A. F. M., Alqaraghuli, W. A. A., Alkarkhi, A. F. M., & Alqaraghuli, W. A. A. (2019). Factor Analysis. Easy Statistics for Food Science with R, 143–159. http://doi.org/10.1016/B978-0-12-814262-2.00009-1

Arima, Chinatsu, Hakamada, Kazumi, Okamoto, Masahiro, Hanai, Taizo, 2008 March. Modified Fuzzy Gap statistic for estimating preferable number of clusters in Fuzzy k-means clustering. Journal of Bioscience and Bioengineering 105 (3), 273–281.

Back, T., 1996. Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic algorithms. Oxford University Press, New York

Banerjee, M., Mitra, S., Banka, H., 2007. Evolutionary rough feature selection in gene expression data. IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews 37 (4), 622–632.

Bani-Fatemi, A., Graff, A., Zai, C., Strauss, J., & De Luca, V. (2016). GWAS analysis of suicide attempt in schizophrenia: Main genetic effect and interaction with early life trauma. Neuroscience Letters, 622, 102–106. http://doi.org/10.1016/J.NEULET.2016.04.043

Bartholomew, D. J. (2015). Factor Analysis and Latent Structure Analysis: Overview. International Encyclopedia of the Social & Behavioral Sciences, 691–697. http://doi.org/10.1016/B978-0-08-097086-8.42043-X

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. Retrieved from http://arxiv.org/abs/1406.5823

Brown, A. A., Ding, Z., Viñuela, A., Glass, D., Parts, L., Spector, T., … Durbin, R. (2015). Pathway-based factor analysis of gene expression data produces highly heritable phenotypes that associate with age. G3 (Bethesda, Md.), 5(5), 839–47. http://doi.org/10.1534/g3.114.011411

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2 (2), 121–167.

Chen Y, Li Y, Narayan R, Subramanian A, Xie X. Gene expression inference with deep learning. Bioinformatics 2016;32:1832–9.

Chen, Y., Pan, Y., Chen, L., Chen, J., 2006. Partitioned optimization algorithms for multiple sequence alignment. In: Proc. of the 20th International Conference on Advanced Information Networking andApplications – (AINA'06), IEEE Computer Society Press, Washington, DC, pp. 618–622

Dandine-Roulland, C., & Perdry, H. (2015). The Use of the Linear Mixed Model in Human Genetics. Human Heredity, 80(4), 196–206. http://doi.org/10.1159/000447634

Fanous, A. H., Zhou, B., Aggen, S. H., Bergen, S. E., Amdur, R. L., Duan, J., … Levinson, D. F. (2012). Genome-Wide Association Study of Clinical Dimensions of Schizophrenia: Polygenic Effect on Disorganized Symptoms. American Journal of Psychiatry, 169(12), 1309–1317. http://doi.org/10.1176/appi.ajp.2012.12020218

Fasheng, X., Yuehui, C., 2009. A method for multiple sequence alignment based on particle swarm optimization. Lecture Notes in Computer Science 5755, 965–973.

Fernando, D., Fdez-Riverola, F., Glez-Pea, D., Corchado, J.M., 2006. Using Fuzzy patterns for gene selection and data reduction on microarray data. Lecture Notes in Computer Science (Bioinformatics and Bio-inspired Models) 4224, 1087–1094.

Fitzmaurice, G. M. (2016). Regression. Diagnostic Histopathology, 22(7), 271–278. http://doi.org/10.1016/J.MPDHP.2016.06.004

Fogel, D.B., 1999. Evolutionary Computation: Toward a New Philosophy of Machine Intelligence, 2nd ed. IEEE Press, Piscataway, NJ.

Frey BJ, Xiong HY, Barash Y. Bayesian prediction of tissue regulated splicing using RNA sequence and cellular context. Bioinformatics 2011;27:2554–62.

Guigo R, Valcarcel J. Prescribing splicing. Science 2015;347:124–5.

*Hinton G (2009). "Deep belief networks". Scholarpedia. **4** (5): 5947.*

Iacob, E., Light, A. R., Donaldson, G. W., Okifuji, A., Hughen, R. W., White, A. T., & Light, K. C. (2016). Gene Expression Factor Analysis to Differentiate Pathways Linked to Fibromyalgia, Chronic Fatigue Syndrome, and Depression in a Diverse Patient Sample. Arthritis Care & Research, 68(1), 132–140. http://doi.org/10.1002/acr.22639

Kapur, K. (2017). Principles of Biostatistics. Clinical and Translational Science, 243–260. http://doi.org/10.1016/B978-0-12-802101-9.00014-4

Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res 2016;26:990–9.

Khan, J., Wei, J.S., Ringnïer, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., Meltzer, P.S., 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine 7 (6), 673–679.

Kotu, V., Deshpande, B., Kotu, V., & Deshpande, B. (2019). Regression Methods. Data Science, 165–197. http://doi.org/10.1016/B978-0-12-814761-0.00005-8

Kustra, R., Shioda, R., & Zhu, M. (2006). A factor analysis model for functional genomics. BMC Bioinformatics, 7(1), 216. http://doi.org/10.1186/1471-2105-7-216

Lane, J. M., Liang, J., Vlasac, I., Anderson, S. G., Bechtold, D. A., Bowden, J., … Saxena, R. (2017). Genome-wide association analyses of sleep disturbance traits identify new loci and highlight shared genetics with neuropsychiatric and metabolic traits. Nature Genetics, 49(2), 274–281. http://doi.org/10.1038/ng.3749

Larochelle, H., Bengio, Y., 2008. Classification using discriminative restricted Boltzmann machines. In: Proceedings of the 25th International Conference on Machine Learning, vol. 307, pp. 536–543.

Leung MKK, Xiong HY, Lee LJ, Frey BJ. Deep learning of the tissue-regulated splicing code. Bioinformatics 2014;30:i121–9.

Lever, J., Krzywinski, M., & Altman, N. (2017). Points of Significance: Principal component analysis. Nature Methods, 14(7), 641–642. http://doi.org/10.1038/nmeth.4346

Li, D., Zhang, W., 2006. Gene selection using rough set theory. In: Lecture Notes in Computer Science, pp. 778–785.

Mahanta, M.S., Aghaei, A.S., Plataniotis, K.N., Pasupathy, S., 2012. Heteroscedastic linear feature extraction based on sufficiency conditions. Pattern Recognition 45 (2), 821–830.

Mohamed, A.R., Dahl, G., Hinton, G.E., 2009. Deep belief networks for phone recognition. NIPS 22 Workshop on Deep Learning for Speech Recognition

Motsinger, F.A., Dudek, F.M., Hahn, F.W., Ritchie, M.D., 2006. Comparison of neural network optimization approaches for studies of human genetics. EvoWorkshops 2006, 103–114.

Nguyen, H.T., Walker, E.A., 1999. A First Course in Fuzzy Logic. CRC Press

Parts, L., Stegle, O., Winn, J., & Durbin, R. (2011). Joint Genetic Analysis of Gene Expression Data with Inferred Cellular Phenotypes. PLoS Genetics, 7(1), e1001276. http://doi.org/10.1371/journal.pgen.1001276

Pawlak, Z., Grzymala-Busse, J., Slowinski, R., Ziarko, W., 1995. Rough sets. Communications of the ACM 38 (11), 88–95.

Polkowski, L., 2003. Rough Sets: Mathematical Foundations. Physica-Verlag.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics, 81(3), 559–75. http://doi.org/10.1086/519795

Richardson, S., Tseng, G. C., & Sun, W. (2016). Statistical Methods in Integrative Genomics. Annual Review of Statistics and Its Application, 3, 181–209. http://doi.org/10.1146/annurev-statistics-041715-033506

Roberto, A., Giorgio, V., 2009. Fuzzy ensemble clustering based on random projections for DNA microarray data analysis. Artificial Intelligence in Medicine 45 (February–March (2–3)), 173–183.

Smith, L. (2005). Exploratory Genomic Data Analysis. In Medical Informatics (pp. 573–592). Boston: Kluwer Academic Publishers. http://doi.org/10.1007/0-387-25739-X_20

Smolinski, T.G., Milanova, M.M., Hassanien, A.E., 2008a. Applications of Computational Intelligence in Biology: Current Trends and Open Problems, Studies in Computational Intelligence. Springer, pp. 122.

StataCorp. 2013. Stata Statistical Software: Release 13. College Station, TX: StataCorp LP

Thompson, B. (2007). Factor Analysis. In The Blackwell Encyclopedia of Sociology. Oxford, UK: John Wiley & Sons, Ltd. http://doi.org/10.1002/9781405165518.wbeosf003

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Research 22 (22), 4673–4680.

Tsetsos, F., Drineas, P., & Paschou, P. (2019). Genetics and Population Analysis. Encyclopedia of Bioinformatics and Computational Biology, 363–378. http://doi.org/10.1016/B978-0-12-809633-8.20114-3

Wang, M. H., Cordell, H. J., & Van Steen, K. (2018). Statistical methods for genome-wide association studies. Seminars in Cancer Biology. http://doi.org/10.1016/J.SEMCANCER.2018.04.008

Wroblewski, J., Slezak, D., 2007. Roughfication of numeric decision tables: the case study of gene expression data. RSKT, 316–323.

Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, et al. The human splicing code reveals new insights into the genetic determinants of disease. Science 2014;347:1254806

Yoon S, Lee T. Boosted categorical restricted boltzmann machine for computational prediction of splice junctions. Proc Int Conf Mach Learn 2015;37

Yu, Z., Wong, H., Wang, H., 2007. Graph based consensus clustering for class discovery from gene expression data. Bioinformatics 23 (21), 2888–2896.

Yuhui, Y., Lihui, C., Goh, A., Wong, A., 2002. Clustering gene data via associative clustering neural network neural. In: Proceedings of the 9th International Conference on Information Processing, ICONIP02, vol. 5, 18–22 November, pp. 2228–2232.

Zhang S, Zhou J, Hu H, Gong H, Chen L, Cheng C, et al. A deep learning framework for modeling structural features of RNA-binding protein targets. Nucleic Acids Res 2016;44:e32

Zhou, Y., Wang, P., Wang, X., Zhu, J., & Song, P. X.-K. (2017). Sparse multivariate factor analysis regression models and its applications to integrative genomics analysis. Genetic Epidemiology, 41(1), 70–80. http://doi.org/10.1002/gepi.22018

1.2.3. Επεξεργασία δεδομένων γλοιοβλαστώματος

Τεχνολογίες omics

Η πρόοδος στις τεχνολογίες omics - όπως η γενωμική, η μεταγραφική, η πρωτεϊνωματική και η μεταβολική (genomics, transcriptomics, proteomics and metabolomics) έχουν αρχίσει να επιτρέπουν την εξατομικευμένη ιατρική σε ένα εξαιρετικά λεπτομερές μοριακό επίπεδο. Μεμονωμένα, αυτές οι τεχνολογίες συνέβαλαν στην ιατρική πρόοδο που άρχισε να εισέρχεται στην κλινική πρακτική. Ωστόσο, κάθε τεχνολογία ξεχωριστά δεν μπορεί να συλλάβει ολόκληρη τη βιολογική πολυπλοκότητα των περισσότερων ανθρώπινων ασθενειών. Η ενσωμάτωση πολλαπλών τεχνολογιών προέκυψε ως μια προσέγγιση για την παροχή μιας πληρέστερης άποψης για τη βιολογία και τις ασθένειες. Στην ιδανική περίπτωση, οι διαφορετικές τεχνολογίες θα συνδυάζονται τόσο για τη διάγνωση της νόσου όσο και για τη δημιουργία μιας ολιστικής εικόνας των ανθρώπινων φαινοτύπων και ασθενειών. Ωστόσο, η εφαρμογή των δεδομένων πολλαπλών "omics" τεχνολογιών εισάγει νέες προκλήσεις πληροφορικής και ερμηνείας. Συγκεκριμένα, απαιτούνται καινοτόμες αναλυτικές και στατιστικές μέθοδοι για τον συνδυασμό διαφορετικών συνόλων δεδομένων, καθώς και τυποποιημένες μετρήσεις ελέγχου ποιότητας. Επιπλέον θα πρέπει οι καινούριες αυτές μέθοδοι να είναι σε θέση να αντιμετωπίσουν τις προκλήσεις στην ερμηνεία των μοριακών συμβάντων και να μπορούν να καθοδηγήσουν τη θεραπευτική και την κλινική φροντίδα.

Αναλύσεις δικτύου. Η ενσωμάτωση πολλαπλών τύπων δεδομένων μπορεί να χρησιμοποιηθεί για τον περιορισμό του χώρου αναζήτησης για τα γονίδια της νόσου και τον εντοπισμό μηχανισμών αιτιολογίας της νόσου. Συγκεκριμένα, τα μοντέλα δικτύων, συμπεριλαμβανομένης της αλληλεπίδρασης πρωτεΐνης-πρωτεΐνης, των δικτύων ρύθμισης και συν-έκφρασης, αποδείχθηκαν πολύτιμοι πόροι για την ιεράρχηση και τον εντοπισμό γονιδίων και οδών ασθενείας (Cowen L *et al*., 2017). Αυτά τα δίκτυα μπορούν να χρησιμοποιηθούν με οποιοδήποτε σύνολο δεδομένων κλίμακας γονιδιώματος, συμπεριλαμβανομένου των απλών νουκλεοτιδικών πολυμορφισμών [Single Nucleotide Polymorphisms, (SNPs)] ή των δεδομένων γονιδιακής έκφρασης, για να διερευνήσουν τις τοπολογικές ιδιότητες των πιο σημαντικών γονιδίων που σχετίζονται με τη νόσο. Στην περίπτωση δεδομένων γενετικών παραλλαγών, υπάρχει πρόκληση στη χαρτογράφηση των SNPs στο προσβεβλημένο γονίδιο: σε ορισμένες περιπτώσεις, η επίδραση της παραλλαγής είναι εμφανής - όπως ή παραλλαγή μετατόπισης πλαισίου σε ένα γονίδιο που σχετίζεται με την ανοσολογική απάντηση, NOD2, στη νόσο του Crohn - αλλά συχνότερα, το προσβεβλημένο γονίδιο για μια παραλλαγή μπορεί να είναι διφορούμενο (Huang, H. *et al*., 2017).

Αναλύσεις "omics" στον καρκίνο

Ένας τομέας όπου οι πολλαπλές αναλύσεις "omics" είχαν και θα συνεχίσουν να έχουν τεράστιο αντίκτυπο είναι στον προσδιορισμό του καρκίνου, τη διάγνωση και τη θεραπεία. Πράγματι, πολλές από τις παραπάνω αναλύσεις δικτύου είναι αποτελεσματικές στον εντοπισμό γενετικών μηχανισμών καρκίνου. Ωστόσο, υπάρχουν εννοιολογικές διαφορές στους καρκίνους που περιπλέκουν τις αναλύσεις τους και απαιτούν ειδικό χειρισμό. Εκτός από τις τεχνικές προκλήσεις που τίθενται για την πρόσκληση σωματικών παραλλαγών, η

πλειονότητα των γενετικών αλλαγών που εμφανίζονται σε περιπτώσεις καρκίνου είναι καλοήθεις και δεν οδηγούν στην ανάπτυξη καρκινικών κυττάρων. Επομένως, ο προσδιορισμός των οδηγών μεταλλάξεων και των μονοπατιών εξακολουθεί να αποτελεί σημαντική πρόκληση. Επιπρόσθετα, παρόλο που μερικοί τύποι καρκίνου μοιράζονται γενετικές υπογραφές μεταξύ ατόμων, εξακολουθεί να υπάρχει υψηλό επίπεδο διαφοροποίησης μεταξύ των οδηγών μεταλλάξεων, γεγονός που μπορεί να οδηγήσει σε διαφορές στην πρόγνωση και τη θεραπευτική.

Αναγνώριση των οδηγών μεταλλάξεων. Μια τυπική διαδικασία για την αναγνώριση των οδηγών μεταλλάξεων περιλαμβάνει WGS (προσδιορισμός ακολουθίας ολόκληρου του γονιδιώματος, Whole Genome Sequencing) πολλαπλών όγκων για τον εντοπισμό επαναλαμβανόμενων μεταλλαγμένων γονιδίων. Η επικάλυψη λειτουργικών δεδομένων μπορεί να βοηθήσει στην ιεράρχηση αυτών των πληροφοριών, καθώς οι οδηγοί μεταλλάξεις είναι πιο πιθανό να είναι σε γονίδια που εκφράζονται σε έναν δεδομένο καρκίνο. Για παράδειγμα, σε μια ανάλυση οδηγών μεταλλάξεων που προσδιορίστηκαν χρησιμοποιώντας WES (Whole Exome Sequencing) σε συνδυασμό με δεδομένα μικροσυστοιχείας, τα δεδομένα RNAseq χρησιμοποιήθηκαν για να αναγνωρίσουν μια εκφρασμένη γονιδιακή σύντηξη του EGFR-SEPT14, η οποία ήταν λειτουργικά επικυρωμένη για να επηρεάσει την ανάπτυξη του γλοιώματος (Frattini, V. *et al*., 2013). Σε μια διαφορετική ανάλυση που χρησιμοποιεί παρόμοιες τεχνολογίες, οι οδηγοί μεταλλάξεις και οι διεργασίες που κρύβουν πολλαπλές μεταστάσεις σε ένα άτομο έδειξαν ότι είναι σε μεγάλο βαθμό παρόμοιες σε όλες τις μεταστάσεις, γεγονός που υποδηλώνει ότι μια μεμονωμένη μετάσταση είναι επαρκής για ανάλυση (Kumar, A. *et al*., 2016). Με αυτόν τον τρόπο, η χρήση πρόσθετων δεδομένων omics συμπληρώνει τα γενετικά δεδομένα, παρέχοντας ένα μηχανισμό για να φιλτράρει τις γενετικές παραλλαγές σε λειτουργικά συναφείς παραλλαγές.

Μοριακές υπογραφές καρκίνου. Εκτός από την αναγνώριση των οδηγών μεταλλάξεων οδηγών, πολλαπλοί τύποι δεδομένων omics μπορούν να αποκαλύψουν γενικές βιοχημικές οδούς που είναι ενεργές σε μεμονωμένους καρκίνους και τους ταξινομούν σε υποτύπους. Ως εκ τούτου, αυτό μπορεί να είναι ένα πολύτιμο εργαλείο για να εξακριβωθεί ποιες βιοχημικές οδοί θα πρέπει να στοχευθούν, ακόμη και αν δεν ανιχνεύονται ισχυρές υποψήφιες μεταλλάξεις σε αυτές τις οδούς. Για παράδειγμα, πρότυπα (patterns) μεθυλίωσης DNA και συστάδες μεταγράφων έχουν χρησιμοποιηθεί για την αναγνώριση υποτύπων καρκίνων, οι οποίοι έχουν διάφορες προγνώσεις επιβίωσης (Sato, Y. *et al*., 2013).

Πρόσφατα, τρεις μελέτες της CPTAC (Clinical Proteomic Tumour Analysis Consortium) χρησιμοποίησαν πρωτεομικές προσεγγίσεις για τον εντοπισμό υποτύπων καρκίνων για το παχύ έντερο, τις ωοθήκες και το μαστό βασισμένες σε συγκεκριμένες πρωτεϊνικές εκφράσεις (Mertins, P. *et al*., 2016). Είναι σημαντικό ότι τα δεδομένα της πρωτεομικής έρευνας αποκάλυψαν αλληλεπικαλυπτόμενη αλλά όχι ταυτόσημη συσχέτιση με τα μεταγραφικά και γενετικά δεδομένα, υποδεικνύοντας ότι οι διαφορετικοί τύποι δεδομένων δείχνουν διαφορετικούς τύπους πληροφοριών. Αυτές οι μελέτες κατέδειξαν τις ξεχωριστές γενετικές και μεταγραφικές διαδικασίες που μεταφράζονται σε πρωτεοματικές αλλοιώσεις.

Τέλος, η ενσωμάτωση των πληροφοριών απεικόνισης με πληροφορίες omics αναμένεται να είναι πολύτιμη για τη διάγνωση και την πρόγνωση του καρκίνου (Yu, K.-H *et al*., 2016). Οι πρόσφατες εξελίξεις στο χαρακτηρισμό των μη κωδικοποιημένων περιοχών που ρυθμίζουν την γονιδιακή έκφραση έχουν γίνει όλο και πιο πολύτιμες για την κατανόηση του ρυθμιστικού τοπίου του καρκίνου. Μελέτες που ενσωματώνουν σύνολα δεδομένων με πληροφορίες σχετιζόμενες με τη ρύθμιση με WGS δεδομένα από το TCGA (The Cancer Genome Atlas) αποκάλυψαν μια σειρά ρυθμιστικών περιοχών που εμπλουτίζονται για μεταλλάξεις σε ασθενείς με καρκίνο (Araya, C. L. *et al*., 2015). Σε αυτές τις περιπτώσεις, εξακολουθεί να είναι δύσκολο να εντοπιστεί το αίτιο της γενετικής ποικιλομορφίας σε αυτές τις μη κωδικοποιήσιμες περιοχές, υπογραμμίζοντας τη συνεχιζόμενη ανάγκη έρευνας, παρόλα αυτά η διαδικτυακή κοινοποίηση της τοπολογίας σε άτομα με τον ίδιο καρκίνο μπορεί να ενημερώσει για υποτύπους καρκίνου που έχουν διαφορετικές προγνώσεις και θεραπευτικές στρατηγικές. Τέλος, δεδομένης της έντονης εξάρτησης της καρκινικής ανάπτυξης από τις μεταβολικές αλλαγές, είναι πιθανό ότι και η μεταβολομική (metabolomics) θα διαδραματίσει σημαντικό ρόλο στην διάγνωση ή πρόγνωση του καρκίνου στο μέλλον.

Παρ 'όλα αυτά, το “omics” προφίλ μπορεί να είναι ένας αποτελεσματικός τρόπος ανίχνευσης μεταβολών σε επίπεδο μεταβολικού μονοπατιού καθώς και μεταβολών μεγάλης κλίμακας, φθηνότερα και πιο ολοκληρωμένα από την εκτέλεση χιλιάδων μεμονωμένων εξετάσεων. Παρά το γεγονός ότι εξακολουθούν να υπάρχουν προκλήσεις στην καθιέρωση κλινικών κατευθυντήριων γραμμών, πολλές από τις έννοιες που περιβάλλουν την ερμηνεία των γενετικών παραλλαγών (ιδιαίτερα σπάνιες ή καινοφανείς παραλλαγές) μπορούν να εφαρμοστούν σε ένα γενικό μοριακό γεγονός (όπως διαφορικά εκφρασμένο γονίδιο, νέα φωσφορυλίωση πρωτεΐνης ή μοναδική υπογραφή μεταβολόματος) όσο η κατανόησή μας για τη βιολογία και τις σχετικές βάσεις δεδομένων γίνεται και πιο ώριμη.

Ανάλυση Εμπλουτισμού Μονοπατιών (Pathway Enrichment Analysis)

Η ανάλυση εμπλουτισμού μονοπατιού (Pathway Enrichment Analysis) βοηθά τους ερευνητές να αποκτήσουν μηχανιστική γνώση σε λίστες γονιδίων που παράγονται από πειράματα “omics” σε κλίμακα γονιδιώματος. Αυτή η μέθοδος αναγνωρίζει βιολογικά μονοπάτια που εμπλουτίζονται σε μια λίστα γονιδίων περισσότερο από ότι θα περίμενε κανείς σε μια τυχαία περίπτωση και συνοψίζει τη μεγάλη λίστα γονιδίων ως μικρότερη λίστα με πιο εύκολα ερμηνεύσιμα μονοπάτια. Τα μονοπάτια εξετάζονται στατιστικά για υπερβολική αναπαράσταση στον κατάλογο πειραματικών γονιδίων σε σχέση με αυτό που αναμένεται τυχαία, χρησιμοποιώντας αρκετές κοινές στατιστικές δοκιμές που λαμβάνουν υπόψη τον αριθμό των γονιδίων που ανιχνεύθηκαν στο πείραμα, τη σχετική κατάταξή τους και τον αριθμό των γονιδίων που σημειώνονται στο βιολογικό μονοπάτι υπό μελέτη. Για παράδειγμα, πειραματικά δεδομένα που περιέχουν 40% γονίδια κυτταρικού κύκλου είναι πάρα πολύ εμπλουτισμένα, δεδομένου ότι μόνο 8% των γονιδίων που κωδικοποιούν ανθρώπινες πρωτεΐνες εμπλέκονται σε αυτή τη διαδικασία.

Επισκόπηση της διαδικασίας

Η ανάλυση εμπλουτισμού μονοπατιών περιλαμβάνει τρία κύρια στάδια:

1. Ορισμός ενός καταλόγου ενδιαφέροντος γονιδίων που χρησιμοποιεί δεδομένα omics. Ένα πείραμα omics μετράει εκτενώς τη δραστηριότητα των γονιδίων σε ένα πειραματικό πλαίσιο. Το ακατέργαστο σύνολο δεδομένων που προκύπτει λαμβάνοντας υπόψη τον πειραματικό σχεδιασμό απαιτεί γενικά υπολογιστική επεξεργασία για τον εντοπισμό των γονιδίων ενδιαφέροντος όπως κανονικοποίηση και βαθμολόγηση. Για παράδειγμα, ένας κατάλογος γονιδίων διαφορικά εκφρασμένων μεταξύ δύο ομάδων δειγμάτων μπορεί να προέρχεται από δεδομένα αλληλουχίας RNA (Anders S *et al*., 2013) Σε αυτό το πρωτόκολλο μπορούν να χρησιμοποιηθούν λίστες γονιδίων που προέρχονται από άλλους τύπους πειραμάτων omics, όπως μικροσυστοιχίες γονιδιακής έκφρασης (Ritchie M E *et al*., 2015) ποσοτική πρωτεομική, (Schubert OT *et al*., 2017) εύρεση αλληλουχίας γονδιώματος (Yang H *et al*., 2015), δοκιμασίες μεθυλίωσης DNA (Assenov, Y. *et al*., 2017), ωστόσο κάθε τύπος δεδομένων ενδέχεται να απαιτεί συγκεκριμένα βήματα προεπεξεργασίας (βλ. ενότητα «Σύγκριση με εναλλακτικές μεθόδους»).

2. Ανάλυση εμπλουτισμού μονοπατιών.

Χρησιμοποιείται μια στατιστική μέθοδος για τον εντοπισμό μονοπατιών εμπλουτισμένων στη λίστα γονιδίων από το στάδιο 1, σε σχέση με αυτό που αναμένεται τυχαία. Όλες οι διαδρομές σε μια δεδομένη βάση δεδομένων δοκιμάζονται για εμπλουτισμό στη λίστα γονιδίων. Είναι διαθέσιμες αρκετές καθιερωμένες μέθοδοι ανάλυσης εμπλουτισμού και η επιλογή της χρήσης εξαρτάται από τον τύπο της λίστας γονιδίων.

3. Οπτικοποίηση και ερμηνεία των αποτελεσμάτων της ανάλυσης εμπλουτισμού μονοπατιών. Πολλά εμπλουτισμένα μονοπάτια μπορούν να ταυτοποιηθούν στο στάδιο 2, που συχνά περιλαμβάνουν σχετικές παραλλαγές του ίδιου μονοπατιού. Η απεικόνιση μπορεί να βοηθήσει στην ταυτοποίηση των κύριων βιολογικών θεμάτων και των σχέσεών τους για εις βάθος μελέτη και πειραματική αξιολόγηση.

Πακέτα λογισμικού και βάσεις δεδομένων

Στο παρόν έργο πολλά από τα δεδομένα που έχουμε ύστερα από την εξαγωγή RNA και την εύρεση της αλληλουχίας είναι σε μορφή FASTQ. Η μορφή FASTQ είναι μια μορφή που βασίζεται σε κείμενο για την αποθήκευση τόσο μιας βιολογικής αλληλουχίας (συνήθως αλληλουχίας νουκλεοτιδίων) όσο και της αντίστοιχης βαθμολογίας ποιότητας. Τόσο το γράμμα αλληλουχίας όσο και η βαθμολογία ποιότητας κωδικοποιούνται με έναν μόνο χαρακτήρα ASCII για συντομία. Ένα αρχείο FASTQ χρησιμοποιεί κανονικά τέσσερις γραμμές ανά ακολουθία.

• Η Γραμμή 1 ξεκινάει με ένα χαρακτήρα '@' και ακολουθεί ένα αναγνωριστικό ακολουθίας και μια προαιρετική περιγραφή (όπως μια γραμμή τίτλου FASTA).

• Η Γραμμή 2 είναι τα γράμματα της πρώτης ακολουθίας.

• Η Γραμμή 3 αρχίζει με έναν χαρακτήρα '+' και ακολουθεί προαιρετικά το ίδιο αναγνωριστικό ακολουθίας (και οποιαδήποτε περιγραφή) ξανά.

• Η γραμμή 4 κωδικοποιεί τις τιμές ποιότητας για την ακολουθία στη γραμμή 2 και πρέπει να περιέχει τον ίδιο αριθμό συμβόλων με τα γράμματα στην ακολουθία.

Ένα αρχείο FASTQ που περιέχει μια μεμονωμένη ακολουθία μπορεί να φαίνεται ως εξής:

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Το byte που αντιπροσωπεύει ποιότητα κυμαίνεται από 0x21 (χαμηλότερη ποιότητα, '!' στο ASCII) έως 0x7e (υψηλότερη ποιότητα, '~' στην ASCII). Ακολουθούν οι χαρακτήρες των τιμών ποιότητας από τα αριστερά προς τα δεξιά με αυξανόμενη σειρά ποιότητας

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

Διάφορα πακέτα λογισμικού χρησιμοποιούνται για την ανάλυση εμπλουτισμού μονοπατιών, τα παρακάτω όμως είναι τα πιο συνηθισμένα και είναι εύκολα στη χρήση, έχουν δωρεάν πρόσβαση, προηγμένα χαρακτηριστικά, εκτενή τεκμηρίωση ενημερωμένες βάσεις δεδομένων και είναι όλα διαθέσιμα στο διαδίκτυο:

● g:Profiler (https://biit.cs.ut.ee/gprofiler/) (Reimand J *et al*., 2016)

● GSEA (http://software.broadinstitute.org/gsea/) (Subramanian A *et al*., 2005)

● Cytoscape (http://www.cytoscape.org/) (Shannon P *et al.*, 2003)

● EnrichmentMap (http://www.baderlab.org/Software/EnrichmentMap) (Merico D et al., 2010)

Οι παρακάτω βάσεις δεδομένων διατηρούνται από μια ομάδα επιμελητών που συλλέγουν λεπτομερείς πληροφορίες για τα μονοπάτια, συμπεριλαμβανομένων των βιοχημικών αντιδράσεων, των ρυθμιστικών συμβάντων γονιδίων και άλλων αλληλεπιδράσεων γονιδίων. Οι πληροφορίες μπορούν να εξάγονται ή να μετατρέπονται σε μορφή γονιδίων.

● Reactome: Η πιο ενημερωμένη δημόσια βάση δεδομένων γενικού σκοπού για τα ανθρώπινα μονοπάτια (http://www. reactive.org) (Faberge A *et al*., 2016)

● Panther: Ανθρώπινα μονοπάτια σηματοδότησης (Human signaling pathways) (http://pantherdb.org/pathway) (Mi H *et al*., 2013)

● NetPath: Ανθρώπινα μονοπάτια σηματοδότησης με εστίαση στον καρκίνο και την ανοσολογία (http://www.netpath.org/) (Kandasamy K *et al*., 2010)

● HumanCyc: Ανθρώπινα μεταβολικά μονοπάτια (http://humancyc.org/) (Caspi, R. *et al*., 2016)
● National Cancer Institute (NCI) (www.cancer.gov)

● KEGG83: Η βάση δεδομένων KEGG (Kanehisa, M *et al*., 2017) είναι πολύ χρήσιμη για ενστικτώδη διαγράμματα μονοπατιών. Περιέχει πολλούς τύπους μονοπατιών μερικά από τα οποία δεν είναι φυσιολογικά αλλά είναι μάλλον γονιδιακά σύνολα που σχετίζονται με νόσο όπως για παράδειγμα “μονοπάτια στον καρκίνο”. (http://www.genome.jp/kegg/).

Στο δικό μας έργο, τα δεδομένα επεξεργάστηκαν περαιτέρω με το πακέτο HTSeq καθώς και τα πακέτα λογισμικού EnrichR, g:Profiler και Reactome. Αυτά τα προγράμματα μαζί με κάποια άλλα ευρέως χρησιμοποιούμενα που θα χρησιμοποιηθούν για την λεπτομερή βιοπληροφορική ανάλυση περιγράφονται παρακάτω:

HTSeq

Η HTSeq είναι μια βιβλιοθήκη Python με σκοπό τη διευκόλυνση της ταχείας ανάπτυξης σεναρίων για την επεξεργασία και την ανάλυση των δεδομένων HTS (High-Throughput Sequencing). Το HTSeq περιλαμβάνει συνδρομητές για κοινές μορφές αρχείων για ποικίλους τύπους δεδομένων εισόδου και είναι κατάλληλο ως γενική πλατφόρμα για ένα ευρύ φάσμα εργασιών. Ένα βασικό συστατικό του HTSeq είναι ότι απλοποιεί τη συνεργασία με δεδομένα που σχετίζονται με γονιδιωματικές συντεταγμένες, δηλαδή τιμές που αποδίδονται σε θέσεις γονιδιώματος (π.χ., κάλυψη ανάγνωσης) ή σε γονιδιωματικά διαστήματα (π.χ., γονιδιωματικά χαρακτηριστικά όπως εξόνια ή γονίδια). Δύο αυτόνομες εφαρμογές που αναπτύχθηκαν με HTSeq είναι η htseq-qa για την αξιολόγηση της ποιότητας ανάγνωσης και η htseq-count για προεπεξεργασία RNA-Seq ευθυγραμμίσεων για μελέτη διαφορικής έκφρασης.

EnrichR

Η ανάλυση εμπλουτισμού είναι μια υπολογιστική μέθοδος για την εξαγωγή γνώσεων σχετικά με ένα γονίδιο εισόδου που έχει καθοριστεί, συγκρίνοντάς το με τα σημειωμένα σύνολα γονιδίων που αντιπροσωπεύουν προηγούμενη βιολογική γνώση. Η ανάλυση εμπλουτισμού ελέγχει αν ένα σύνολο εισροών γονιδίων επικαλύπτει σημαντικά τα σύνολα γονιδίων με σχολιασμό.

Ένα από τα κύρια εργαλεία στον τομέα αυτό ονομάζεται Enrichr (http://amp.pharm.mssm.edu/Enrichr/help#basics&q=0). Ο Enrichr περιλαμβάνει αυτήν τη στιγμή μια μεγάλη συλλογή από ποικίλες βιβλιοθήκες που είναι διαθέσιμες για ανάλυση και λήψη. Έχουν προστεθεί νέες λειτουργίες στο Enrichr, συμπεριλαμβανομένης της δυνατότητας υποβολής ασαφών συνόλων, μεταφόρτωσης αρχείων BED, βελτιωμένης διασύνδεσης προγραμματισμού εφαρμογών και απεικόνισης των αποτελεσμάτων ως clustergrams. Συνολικά, το Enrichr είναι μια περιεκτική πηγή για τα καθορισμένα σύνολα γονιδίων και μια μηχανή αναζήτησης που συσσωρεύει βιολογική γνώση για περαιτέρω βιολογικές ανακαλύψεις.

Ο Enrichr εφαρμόζει τέσσερις βαθμολογίες για να αναφέρει τα αποτελέσματα εμπλουτισμού: p-value, q-value, rank (Z-score) και συνδυασμένη βαθμολογία.

Η τιμή p υπολογίζεται χρησιμοποιώντας μια τυποποιημένη στατιστική μέθοδο που χρησιμοποιείται από τα περισσότερα εργαλεία ανάλυσης εμπλουτισμού: η ακριβής δοκιμή Fisher ή η υπεργομετρική δοκιμή. Αυτή είναι μια δοκιμασία διωνυμικής αναλογίας που προϋποθέτει διωνυμική κατανομή και ανεξαρτησία για πιθανότητα οποιουδήποτε γονιδίου που ανήκει σε οποιοδήποτε σετ.

Η τιμή q είναι μια προσαρμοσμένη τιμή p χρησιμοποιώντας τη μέθοδο Benjamini-Hochberg για διόρθωση για δοκιμές πολλαπλών υποθέσεων (Benjamini Y et al., 1995).

Ο βαθμός βαθμολογίας ή η βαθμολογία z υπολογίζονται με τη χρήση μιας τροποποίησης στην ακριβή δοκιμασία του Fisher στην οποία υπολογίζουμε ένα z-score για απόκλιση από την αναμενόμενη κατάταξη.

Τέλος, η συνδυασμένη βαθμολογία είναι ένας συνδυασμός της τιμής p και της βαθμολογίας z που υπολογίζεται πολλαπλασιάζοντας τις δύο βαθμολογίες.

TopHat2

Το TopHat2 είναι ένα δημοφιλές πρόγραμμα για μάτισμα σε πειράματα αλληλουχίας RNA (RNA-seq). Το TopHat2 μπορεί να ευθυγραμμίσει τις αναγνώσεις διαφόρων μηκών που παράγονται από τις τελευταίες τεχνολογίες προσδιορισμού αλληλουχίας, όπως επίσης και τις αναγνώσεις σε θραύσματα σύντηξης τα οποία μπορεί να εμφανιστούν μετά από γονιδιωματικές μετατοπίσεις. Το TopHat2 συνδυάζει την ικανότητα εντοπισμού νέων θέσεων ματίσματος με άμεση χαρτογράφηση σε γνωστές μετάγραφα, παράγοντας ευαίσθητες και ακριβείς ευθυγραμμίσεις, ακόμη και για γονίδια μεγάλης επαναληπτικότητας ή παρουσία ψευδογονιδίων. Το TopHat2 είναι διαθέσιμο στη διεύθυνση http://ccb.jhu.edu/software/tophat.

g:Profiler

Το g:Profiler (https://biit.cs.ut.ee/gprofiler) χωρίζεται στα παρακάτω μέρη:

g:GOSt. Η g:GOSt εκτελεί στη λίστα γονιδίων εισόδου ανάλυση λειτουργικού εμπλουτισμού, γνωστή και ως ανάλυση εμπλουτισμού γονιδίων (GSEA). Χαρτογραφεί γονίδια σε γνωστές πηγές λειτουργικών πληροφοριών και ανιχνεύει στατιστικά σημαντικά εμπλουτισμένους όρους. Ανανεώνει τα δεδομένα από τη βάση δεδομένων Ensembl Genomes και από ειδικά δεδομένα για παράσιτα από το WormBase ParaSite. Εκτός από την οντολογία των γονιδίων, συμπεριλαμβάνει μεταβολικά μονοπάτια από το KEGG Reactome και το WikiPathways, miRNA στόχους από την miRTarBase, ρυθμιστικά μοτίβα από το TRANSFAC, εξειδίκευση ιστού από το Human Protein Atlas, πρωτεϊνικά σύμπλοκα από το CORUM και φαινότυπους ανθρώπινης νόσου από το Human Phenotype Oncology. Το g:GOSt υποστηρίζει περίπου 500 οργανισμούς και δέχεται εκατοντάδες τύπους αναγνωριστικών.

g:Convert. Η g:Convert επιτρέπει τη μετατροπή μεταξύ διαφόρων γονιδίων, πρωτεϊνών, ανιχνευτών μικροσυστοιχιών και πολλών άλλων τύπων. Παρέχει τουλάχιστον 40 τύπους αναγνωριστικών για περισσότερα από 60 είδη. Οι 98 διαφορετικοί χώροι ονομάτων που υποστηρίζονται για τον άνθρωπο περιλαμβάνουν τα αναγνωριστικά Ensembl, Refseq, Illumina, Entrezgene και Uniprot. Όλοι οι χώροι ονομάτων αποκτώνται μέσω της αντιστοίχισής τους μέσω αναγνωριστικών γονιδίων Ensembl ως αναφοράς.

g:Orth. Η g:Orth μεταφράζει αναγνωριστικά γονιδίων μεταξύ οργανισμών. Παρέχει ορθολογικές χαρτογραφήσεις γονιδίων με βάση τις πληροφορίες που ανακτώνται από τη βάση δεδομένων Ensembl.

g:SNPense. Η g:SNPense χαρτογραφεί έναν κατάλογο ανθρώπινων SNP rs-codes (π.χ. rs7961894) σε ονόματα γονιδίων και λαμβάνει χρωμοσωμικές συντεταγμένες και προβλεπόμενα αποτελέσματα παραλλαγής. Η χαρτογράφηση είναι ενεργοποιημένη μόνο για παραλλαγές που επικαλύπτονται με τουλάχιστον ένα γονίδιο Ensembl που κωδικοποιεί πρωτεΐνη. Όλα τα υποκείμενα δεδομένα ανακτώνται από τα δεδομένα της βάσης Ensembl.

Reactome

Το Reactome (https://reactome.org/user/guide) είναι μια επιμελημένη βάση δεδομένων για τα μονοπάτια και τις αντιδράσεις στην ανθρώπινη βιολογία. Οι αντιδράσεις μπορούν να θεωρηθούν ως «βήματα» της πορείας. Το Reactome ορίζει μια «αντίδραση» ως οποιοδήποτε γεγονός στη βιολογία που αλλάζει την κατάσταση ενός βιολογικού μορίου. Η δέσμευση, η ενεργοποίηση, η μετατόπιση, η υποβάθμιση και τα κλασσικά βιοχημικά συμβάντα που περιλαμβάνουν έναν καταλύτη είναι όλες αντιδράσεις. Οι πληροφορίες στη βάση δεδομένων συντάσσονται από εμπειρογνώμονες βιολόγους, οι οποίοι εισήλθαν και διατηρήθηκαν από την ομάδα επιμελητών και συντακτικού προσωπικού του Reactome. Το περιεχόμενο Reactome συχνά παραπέμπει σε άλλους πόρους, π.χ. NCBI, Ensembl, UniProt, KEGG (Gene and Compound), ChEBI, PubMed και GO.

Ο ακρογωνιαίος λίθος του Reactome είναι μια ελεύθερα διαθέσιμη, ανοιχτού κώδικα βάση δεδομένων μεταβολικών μορίων και οι σχέσεις των μορίων αυτών οργανωμένες σε βιολογικά μονοπάτια και διαδικασίες. Η βασική μονάδα του μοντέλου δεδομένων Reactome είναι η αντίδραση. Οι οντότητες (νουκλεϊκά οξέα, πρωτεΐνες, σύμπλοκα, εμβόλια, αντικαρκινικά θεραπευτικά και μικρά μόρια) που συμμετέχουν σε αντιδράσεις σχηματίζουν ένα δίκτυο βιολογικών αλληλεπιδράσεων και ομαδοποιούνται σε μονοπάτια. Παραδείγματα βιολογικών μονοπατιών στο Reactome περιλαμβάνουν τον κλασσικό ενδιάμεσο μεταβολισμό, τη σηματοδότηση, τη ρύθμιση μεταγραφής, την απόπτωση και την ασθένεια. Στόχος του REACTOME είναι η παροχή εργαλείων βιοπληροφορικής για την απεικόνιση, ερμηνεία και ανάλυση της γνώσης των βιολογικών μονοπατιών για τη στήριξη της βασικής και της κλινικής έρευνας, της ανάλυσης του γονιδιώματος, της μοντελοποίησης, της βιολογίας συστημάτων και της εκπαίδευσης.

Network analyst, MASER

Η network analyst (https://www.networkanalyst.ca/faces/home.xhtml) είναι μια Web εφαρμογή που επιτρέπει πολύπλοκες μετα-ανάλυση και απεικόνιση. Είναι σχεδιασμένη από εξειδικευμένους βιοπληροφορικούς έτσι ώστε να είναι προσβάσιμο στους βιολόγους. Ενσωματώνει προηγμένες στατιστικές μεθόδους και καινοτόμο οπτικοποίηση δεδομένων για αποτελεσματικές συγκρίσεις δεδομένων, βιολογική ερμηνεία και δημιουργία υποθέσεων

Ο σκοπός της network analyst είναι να παρέχει ένα ολοκληρωμένο “web-based” πλαίσιο για την επεξεργασία των δεδομένων, τη λειτουργική ανάλυση και την οπτικοποίηση των διαφόρων μορφών των δεδομένων της γονιδιακής έκφρασης [λίστα (ες) γονιδίων, στοιχεία έκφρασης γονιδίων (ακατέργαστα και επεξεργασμένα), πολλαπλά σύνολα δεδομένων έκφρασης γονιδίων, αρχεία δικτύου]. Χρησιμοποιεί προηγμένες στατιστικές μεθόδους συνδυασμένες με οπτικοποίηση διαδραστικών δεδομένων και έχει όριο 500MB για τη μεταφόρτωση δεδομένων.

Μια τελευταία πλατφόρμα ανάλυσης δεδομένων βασισμένη σε cloud ονομάζεται MASER (Management and Analysis System for Enormous Reads) (Sonoco K *et al*., 2018), και επιτρέπει στους χρήστες να διαχειρίζονται μέχρι 2 terabytes δεδομένων για να διεξάγουν αναλύσεις με εύκολες γραφικές διεπαφές χρήστη και προσφέρει αγωγούς ανάλυσης στους οποίους πολλά μεμονωμένα εργαλεία συνδυάζονται ως ένας αγωγός για πολύ συνηθισμένες και τυποποιημένες αναλύσεις. Με αυτή τη λειτουργία, οι αγωγοί Maser μπορούν να εμφανίσουν γραφικά την έξοδο των αποτελεσμάτων από όλα τα ενσωματωμένα εργαλεία και τα αποτελέσματα χαρτογράφησης σε ένα πρόγραμμα περιήγησης ιστού.

Τα NGS (Next Generation Sequencing) δεδομένα δημιούργησαν μια σημαντική πρόκληση στον τρόπο χειρισμού των τεράστιων ποσοτήτων δεδομένων και ποικιλίας εργαλείων NGS καθώς και στην απεικόνιση των αποτελεσμάτων που προκύπτουν. Παραπάνω παρουσιάσαμε τα κυριότερα προγράμματα, βάσεις δεδομένων και τεχνικές που ακολουθούνται και θα χρησιμοποιήσουμε και εμείς με σκοπό την εκτεταμένη

βιοπληροφορική ανάλυση μεγάλου όγκου γενωμικών δεδομένων από RNA-seq για την ανάπτυξη αλγορίθμων για την εύρεση βιοδεικτών και την συγκέντρωση πληροφοριών που θα εξηγήσουν τις βιοχημικές αλλαγές που συμβαίνουν κατά τη χορήγηση συγκεκριμένων φαρμάκων σε συγκεκριμένα κυτταρικά μοντέλα με γλοιοβλάστωμα εγκεφάλου.

Βιβλιογραφία

Anders, S. *et al*. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat. Protoc. 8, 1765–1786 (2013).

Araya, C. L. *et al*. Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations. Nat. Genet. 48, 117–125 (2015).

Assenov, Y. *et al*. Comprehensive analysis of DNA methylation data with RnBeads. Nat. Methods 11, 1138–1140 (2014)

Caspi, R. *et al*. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 44, D471–D480 (2016).

Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. Nat. Rev. Genet. 18, 551–562 (2017).

Fabregat, A. *et al*. The Reactome pathway Knowledgebase. Nucleic Acids Res. 44, D481–D487 (2016).

Frattini, V. *et al*. The integrated landscape of driver genomic alterations in glioblastoma. Nat. Genet. 45, 1141–1149 (2013).

Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009 Jan; (1):1-13. doi: 10.1093/nar/gkn923. Epub 2008 Nov 25

Huang, H. *et al*. Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature 547, 173–178 (2017).

Kandasamy, K. *et al*. NetPath: a public resource of curated signal transduction pathways. Genome Biol. 11, R3 (2010)

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 45, D353–D361 (2017).

Kumar, A. *et al*. Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate cancer. Nat. Med. 22, 369–378 (2016).

Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PLoS ONE 5, e13984 (2010).

Mertins, P. *et al*. Proteogenomics connects somatic mutations to signaling in breast cancer. Nature 534, 55–62 (2016)

Mi, H., Muruganujan, A. & Thomas, P. D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res. 41, D377–D386 (2013).

Reimand, J. *et al*. g:Profiler-a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res. 44, W83–89 (2016).

Ritchie, M. E. *et al*. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47 (2015).

Sato, Y. *et al*. Integrated molecular analysis of clearcell renal cell carcinoma. Nat. Genet. 45, 860–867 (2013).

Schubert, O. T., Rost, H. L., Collins, B. C., Rosenberger, G. & Aebersold, R. Quantitative proteomics: challenges and opportunities in basic and applied research. Nat. Protoc. 12, 1289–1294 (2017)

Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–2504 (2003).

Sonoco Kinko, Norikazou Monma, Sadahiko Misu, Norikazu Kitamura, Junichi Imoto, Kazutoshi Yoshitake, Takashi Gojobori and Kazuho Ikeo. Maser: one=stop platform for NGS big data from analysis to visualization. Database (Oxford). 2018; 2018: bay027. Published online 2018 Apr 13. doi: 10.1093/database/bay027 PMCID: PMC5905357 PMID: 29688385

Subramanian, A. *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genomewide expression profiles. Proc. Natl. Acad. Sci. USA 102, 15545–15550 (2005)

Yang, H. & Wang, K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat. Protoc. 10, 1556–1566 (2015)

Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Poweful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*. Vol. 57, No. 1 (1995), pp. 289-300

Yu, K.-H. & Snyder, M. Omics profiling in precision oncology. Mol. Cell. Proteom. 15, 2525–2536 (2016)