

«Big data management and processing. Proteomic and proteogenomic analyses of lung cancer patients using bioinformatic tools.»

ΟΛΥΜΠΙΑ ΕΙΡΗΝΗ ΜΠΟΥΛΙΟΓΛΟΥ

ΕΘΝΙΚΟ & ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ

ΙΑΤΡΙΚΗ ΣΧΟΛΗ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΑΝΑΠΤΥΞΗ ΝΕΩΝ ΦΑΡΜΑΚΩΝ: ΕΡΕΥΝΑ, ΚΥΚΛΟΦΟΡΙΑ ΚΑΙ ΠΡΟΣΒΑΣΗ»

ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

ΕΠΙΒΛΕΠΩΝ: Αθανάσιος Αναγνωστόπουλος, Ειδικός Λειτουργικός
Επιστήμονας Β', Υπεύθυνος Μονάδας Πρωτεωγενωμικής ΙΙΒΕΑΑ

ΜΕΛΟΣ: Ιωάννης Γκιάζος, Ακαδημαϊκός Υπότροφος, Ιατρική Σχολή, ΕΚΠΑ

ΜΕΛΟΣ: Ιωάννης Βαμβακάρης, PhD Παθολογοανατόμος, ΓΝΝΘΑ «Η
Σωτηρία»

Πίνακας περιεχομένων

1. Περίληψη	5
2. Summary	7
3. Εισαγωγή	9
Πίνακας 1: Σύνοψη τεχνολογιών παραγωγής δεδομένων γενομικής.	17
Πίνακας 2: Σύνοψη τεχνολογιών παραγωγής δεδομένων πρωτεομικής.	18
4. Κυριότερες κατηγορίες εργαλείων & Βάσεις Δεδομένων	19
Εργαλεία ευθυγράμμισης	19
Εργαλεία καταμέτρησης χαρακτηριστικών	21
Εργαλεία ταυτοποίησης πρωτεϊνών – Proteome Discoverer	22
Στατιστικά εργαλεία	24
Εργαλεία εμπλουτισμού	25
Εργαλεία οπτικοποίησης δεδομένων	27
Εικόνα 1: Διάγραμμα ροής για οπτικοποίηση κάθε τύπου δεδομένων	29
Βάσεις Δεδομένων	30
Πίνακας 3: Βάσεις δεδομένων πρωτεϊνών.	32
5. Σκοπός της εργασίας	33
6. Υλικά και Μέθοδοι	33
Εικόνα 2: PRISMA flow chart στο οποίο αναλύεται η διαδικασία εύρεσης και ελέγχου βιβλιογραφίας γι αυτή την εργασία.	34
7. Εργαλεία Βιοπληροφορικής	35
TopHat	35
featureCounts	43
DESeq2	52
Geneset Enrichment Analysis (GSEA)	62
DAVID (Database for Annotation, Visualization and Integrated Discovery)	71
clusterProfiler	82
Metascape	86
ggplot2	98
Cytoscape	102
StringDB	108
Εικόνα 3: Παράδειγμα pipeline ανάλυσης δεδομένων πρωτεομικής και πρωτεογενομικής	115
8. Συζήτηση	116
Εικόνα 4: Παράδειγμα ροής εργασιών σε μελέτη multi-omics	118

9. Βιβλιογραφία	128
10. Σύνοψη.....	153

1. Περίληψη

Οι τεχνολογίες multi-omics αποτελούν αδιάσπαστο τμήμα της βασικής και κλινικής έρευνας. Πολύπλοκες ασθένειες, όπως για παράδειγμα ο καρκίνος του πνεύμονα, έχουν ως ατιολογικό υπόβαθρο πολλαπλούς γενετικούς και περιβαλλοντικούς παράγοντες η ολιστική μελέτη των οποίων αποτελεί ιδιαίτερα απαιτητική διαδικασία. Η επιστήμη της βιοπληροφορικής συμβάλλει σημαντικά στην επεξεργασία και ανάλυση δεδομένων omics, χρησιμοποιώντας τα εργαλεία της πληροφορικής για να απαντήσει πολύπλοκα βιολογικά ερωτήματα. Ωστόσο, παρά την διαθεσιμότητα εργαλείων για την ανάλυση τέτοιου είδους δεδομένων δεν καθίσταται πάντα εύκολη η σωστή διεξαγωγή βιοπληροφορικών αναλύσεων. Στην παρούσα Διπλωματική Εργασία (ΔΕ) παρατίθενται όλες οι σύγχρονες προσεγγίσεις για τη βιοπληροφορική ανάλυση, καθώς και παραδείγματα για τον σχεδιασμό μιας πρωτεογενομικής μελέτης. Ειδικότερα, περιγράφονται οι αρχές μεθόδου για την προεπεξεργασία, στατιστική ανάλυση και σύγκριση των αποτελεσμάτων με την προϋπάρχουσα βιβλιογραφία. Επίπλέον, αναφέρονται τρόποι οπτικοποίησης των δεδομένων μέσω γλώσσας προγραμματισμού ή μέσω διαδικτυακών πλατφορμών (π.χ. StringDB). Σε αυτά τα πλαίσια, παρατίθενται βάσεις δεδομένων πρωτεομικής που έχουν προκύψει από κλινικές και βιοχημικές μελέτες οι οποίες αποτελούν σημεία αναφοράς για τη συνεχιζόμενη έρευνα των πρωτεϊνών. Επιπροσθέτως, σημειώνονται συνοπτικά παραδείγματα μελετών multi-omics καθώς και η χρήση των εργαλείων βιοπληροφορικής που οδήγησαν σε κλινικά συμπεράσματα. Συμπερασματικά, αυτές οι εξελισσόμενες τεχνολογίες αποτελούν αναπόσπαστο κομμάτι της κλινικής έρευνας. Παρόλα αυτά αντιμετωπίζουν προκλήσεις σχετικές με την ευστοχία τους και τις απαιτήσεις

τους σε υπολογιστική υποδομή. Στη μοντέρνα κλινική έρευνα, πολλές από τις προαναφερθείσες τεχνολογίες έχουν πλέον κάνει εφικτή την ανάλυση μοναδικών κυττάρων (single-cell omics) ή ιστών (spatial omics). Με αυτόν τον τρόπο, δύναται να εξαχθούν συμπεράσματα, όχι μόνο για τον μέσο φαινότυπο ενός κυτταρικού πληθυσμού, αλλά και για την εσωτερική διακύμανσή του. Στόχος της παρούσας ΔΕ είναι η δημιουργία ενός εγχειρίδιου βιοπληροφορικής ανάλυσης παρέχοντας τα βασικά βήματα βιοπληροφορικής επεξεργασίας λεπτομερής παράθεση βιοπληροφορικών εργαλείων προεπεξεργασίας, στατιστικής, εμπλουτισμού και οπτικοποίησης δεδομένων omics.

2. Summary

Multi-omics technologies are indispensable for basic and clinical research. Complex diseases, such as lung cancer, can be attributed to multiple genetic and environmental factors that were previously impossible to study comprehensively. Bioinformatics is a field of science that combines informatics and biological reasoning, allowing thus the analysis of complex omics data. However, despite the availability of tools to analyze this data, it is often strenuous to follow the best practices. This work lists the most common approaches for bioinformatic analysis, as well as presents examples for designing a proteogenomics research project and more specifically the principles of method for the pre-processing, statistical analysis and comparison of the results with the existing literature (enrichment) in mind. Furthermore it exhibits ways of visualizing data, either with a programming language or with online platforms, *e.g.* StringDB. In these contexts, proteomic databases derived from clinical and biochemical studies are also listed, as they are reference points for ongoing protein research. Included are brief examples of multi-omics studies and the use of bioinformatic tools. Generally, these advancing technologies are an integral part of clinical research, still they face challenges related to their precision and computing infrastructure requirements. In regards to modern clinical research, many of the aforementioned technologies have achieved single-cell resolution. Conclusions can be drawn in this way, not only about the average phenotype of a cell population, but also about its cell-to-cell variance. The scope of this thesis is to meticulously describe the bioinformatics tools used for pre-processing, statistics, enrichment and visualization of omics

data. However, it may also serve the reader as a manual for the principal steps of a bioinformatic analysis.

3. Εισαγωγή

Οι οργανισμοί υγειονομικής περίθαλψης παγκόσμια χρησιμοποιούν διάφορα μοντέλα πληροφοριακών συστημάτων, προκειμένου να παρέχουν τις καλύτερες δυνατές υπηρεσίες στους ασθενείς. Τα μοντέλα αυτά, όντας χρήσιμα στον τομέα της εξατομικευμένης ιατρικής, βασίζονται στη χρήση ηλεκτρονικών αρχείων υγείας και τεράστιων ποσοτήτων πολύπλοκων βιοϊατρικών δεδομένων (1). Οι σύγχρονες τεχνολογίες γονιδιωματικής και μεταγονιδιωματικής, παράγουν μεγάλου όγκου ακατέργαστα δεδομένα που σχετίζονται με πολύπλοκες βιοχημικές και ρυθμιστικές διεργασίες που αφορούν τον άνθρωπο, όπως και τους υπόλοιπους ζωντανούς οργανισμούς (2). Αυτά τα δεδομένα είναι ετερογενή και πολύ συχνά αποθηκεύονται με διαφορετικές μορφοποιήσεις.

Τα *"Big Data"* («Μεγάλα Δεδομένα») στην υγειονομική περίθαλψη και την ιατρική αναφέρονται σε αυτά τα διάφορα μεγάλα και πολύπλοκα δεδομένα. Εξαιτίας της πολυπλοκότητας αυτής, είναι δύσκολη η διαχείρισή τους, καθώς και η ανάλυσή τους από τα υπάρχοντα διαθέσιμα λογισμικά (3,4). Η διαχείριση τους περιλαμβάνει αρχικά την ενοποίηση ετερογενών δεδομένων, στη συνέχεια τον ποιοτικό τους έλεγχο, την ανάλυσή τους, τη μοντελοποίησή τους και, τέλος, την ερμηνεία τους και την επικύρωσή τους (5). Η καινοτομία που προσφέρει η ανάλυση των Big Data επιτρέπει την εξαγωγή ολιστικών συμπερασμάτων η οποία οφείλεται στον τεράστιο διαθέσιμο όγκο τους.

Πιο συγκεκριμένα, η ανάλυση των Big Data επιτρέπει στους επιστήμονες της υγείας:

α) Να διαχειριστούν και να επεξεργαστούν μεγάλα σύνολα δεδομένων που προκύπτουν από χιλιάδες ασθενείς.

β. Να ομαδοποιήσουν αυτά τα σύνολα δεδομένων (clustering) και στη συνέχεια να τα συσχετίσουν μεταξύ τους.

γ. Να αναπτύξουν προγνωστικά μοντέλα μέσω της χρήσης τεχνικών εξόρυξης δεδομένων (2,6).

Ο όρος Big Data συνοδεύεται από τα ακόλουθα χαρακτηριστικά: τιμή, όγκος, ταχύτητα, ποικιλία, ακρίβεια και μεταβλητότητα (7). Ο όγκος αναφέρεται στην ποσότητα των δεδομένων, ενώ η ταχύτητα σχετίζεται με τη συχνότητα δημιουργίας, επεξεργασίας και ανάλυσης δεδομένων. Η πολυπλοκότητα και η ετερογένεια των πολλαπλών συνόλων δεδομένων (τα οποία μπορούν να είναι δομημένα, ημιδομημένα και μη) αναφέρονται στην ποικιλία. Τέλος, η ακρίβεια αναφέρεται στην ποιότητα των δεδομένων, καθώς και στην αξιοπιστία και την προγνωστική τους αξία (8), ενώ η μεταβλητότητα αφορά την αξιοπιστία των δεδομένων σε βάθος χρόνου.

Η ανάλυση των Big Data επιτρέπει στους επιστήμονες που εργάζονται στον κλάδο της υγείας να βελτιώσουν τις υπηρεσίες φροντίδας των ασθενών. Ειδικότερα, μέσω αυτών μπορούν πλέον να ανιχνεύουν νωρίτερα τις μεταδιδόμενες ασθένειες και ταυτόχρονα να δημιουργούν νέες γνώσεις για τους μηχανισμούς των ασθενειών αυτών. Επίσης ο συνδυασμός των δεδομένων αυτών με τα ηλεκτρονικά αρχεία υγείας, διευκολύνει τους υγειονομικούς να παρακολουθούν την ποιότητα των ιατρικών ιδρυμάτων και τους επιτρέπει να παρέχουν καλύτερες μεθόδους θεραπείας (9–11). Αυτό μπορεί να επιτευχθεί με τη χρήση του διαδικτύου, των μέσων κοινωνικής δικτύωσης, ακόμα και των κινητών τηλεφώνων (12,13).

Η ανάλυση και επεξεργασία των Big Data προϋποθέτει την ορθή συλλογή και διαχείρισή τους. Όσον αφορά τη συλλογή τους, θα πρέπει να

ληφθούν υπόψη ορισμένα ζητήματα. Το κύριο ζήτημα που επηρεάζει την επιλογή μεθόδων υψηλής απόδοσης (high-throughput omics), έναντι παραδοσιακών πειραμάτων είναι η διαθέσιμη χρηματοδότηση για την εκάστοτε μελέτη. Επιπλέον η ετερογένεια των πηγών δεδομένων, ο «θόρυβος» (noise) των δεδομένων omics, η ποικιλία των πειραματικών τεχνικών, οι περιβαλλοντικές συνθήκες και η βιολογική φύση θα πρέπει να ληφθούν υπόψη πριν από την ενσωμάτωσή τους και πριν από την εφαρμογή μεθόδων εξόρυξης δεδομένων. Μερικές από αυτές τις μεθόδους περιλαμβάνουν διαδικασίες όπως: ανίχνευση ανωμαλιών, ομαδοποίηση, ταξινόμηση, κανόνες συσχέτισης καθώς και περίληψη και απεικόνιση αυτών των μεγάλων συνόλων δεδομένων (14).

Για την παραγωγή δεδομένων omics απαιτούνται μηχανήματα υψηλής απόδοσης (high-throughput) που μπορούν να διαχειριστούν την ανάγνωση και καταγραφή της μοριακής πληροφορίας, δηλαδή από κάθε δείγμα να παράξουν ένα αρχείο ακατέργαστων δεδομένων (.raw data file). Πιο συγκεκριμένα, για την παραγωγή δεδομένων γενομικής, επιγενομικής και μεταγραφομικής χρησιμοποιούνται αλληλουχητές (sequencers), οι οποίοι διαβάζουν τα νουκλεοτίδια από τις εισαγόμενες αλληλουχίες DNA/RNA. Οι πιο γνωστές πλατφόρμες – μηχανήματα αλληλούχισης περιγράφονται στον Πίνακα 1.

Αντίστοιχα για την πρωτεομική, δεν υπάρχουν αλληλουχητές που να διαβάζουν την ακολουθία αμινοξέων της εισαγόμενης πρωτεΐνης. Παρόλα αυτά, η ανάπτυξη της τεχνολογίας φασματοσκοπίας μάζας μάς επιτρέπει να αντιστοιχίσουμε την αναλογία μάζας – φορτίου ενός ιονισμένου θραύσματος με την πρωτεΐνη προέλευσής του. Αυτή η ανάλυση γίνεται σε μηχανήματα

φασματοσκοπίας μάζας (mass spectrometers), μερικά από τα οποία αναλύονται στον Πίνακα 2.

Μετά τη συλλογή των δεδομένων λαμβάνει χώρα η προεπεξεργασία τους. Αυτή συνήθως αφορά τον χειρισμό «θορυβωδών» δεδομένων, ακραίων τιμών, τιμών που λείπουν, μετασχηματισμού δεδομένων και κανονικοποίηση. Έτσι η ποιότητα και τα αποτελέσματα της ανάλυσης των Big Data είναι δυνατόν να βελτιωθούν και να οδηγήσουν στην ανακάλυψη νέας γνώσης. Αυτή η νέα γνώση θα πρέπει να έχει πάντα ως απώτερο σκοπό τη βελτίωση της υγειονομικής περίθαλψης που εφαρμόζεται στους ασθενείς καθώς και την προηγμένη λήψη αποφάσεων από τους υπεύθυνούς της (14).

Τα Big Data συχνά συνδέονται με μακροσυστήματα μεγάλης κλίμακας με κατανεμημένη επεξεργασία δεδομένων, που συχνά υπερβαίνει τις δυνατότητες των τοπικών επιτραπέζιων υπολογιστών και του παραδοσιακού λογισμικού λόγω των περιορισμών που επιβάλλονται από την ταχύτητα και τον όγκο της επεξεργασίας. Η ποσότητα και η ποιότητά τους επηρεάζονται από στατιστικές επαναλήψεις σε μεγάλης κλίμακας επιστημονικά πειράματα και υπόκεινται σε πολλές γνωσιολογικές αβεβαιότητες (15).

Ένα παράδειγμα εφαρμογής των Big Data είναι η ιατρική ακριβείας. Οι αντίστοιχες κλινικές δοκιμές βασίζονται στην επιλογή ασθενών σύμφωνα με το γενετικό προφίλ, το οποίο παρέχει βιοδείκτες για στοχευμένη θεραπεία, αντί για μια τυπική προσέγγιση που χρησιμοποιείται για ολόκληρο τον πληθυσμό (16). Η επιλογή ατόμων που μοιράζονται την ίδια γενετική ανωμαλία για κλινικές δοκιμές οδηγεί σε ακριβέστερη ανάπτυξη φαρμάκων και πιο ακριβείς θεραπείες με ήδη υπάρχοντα φάρμακα (16,17). Πρόσθετα πλεονεκτήματα αποτελούν οι

μειώσεις στο μέγεθος του δείγματος και η μειωμένη μεταβλητότητα στις κλινικές δοκιμές — χωρίς την ανάγκη για εικονικό φάρμακο (placebo). Υπάρχει επίσης η δυνατότητα ο προσδιορισμός της αλληλουχίας του γονιδιώματος να παράγει μεγαλύτερη διαγνωστική ευαισθησία και πιο ακριβή θεραπεία (18). Μελέτες πληθυσμού μεγάλης κλίμακας και αναλύσεις Big Data είναι εφικτό να χρησιμοποιηθούν για ακριβέστερα γονοτυπικά και φαινοτυπικά δεδομένα για την υποστήριξη ερευνών αιτιότητας.

Ο συνδυασμός των Big Data με την συνεχώς αναπτυσσόμενη χρήση της τεχνητής νοημοσύνης, δύναται να παράξει αποτελέσματα που δημιουργούν νέους ορίζοντες στον ιατρικό κλάδο (15). Η τεχνητή νοημοσύνη (artificial intelligence, AI) είναι ένας γενικός όρος για μια μηχανή (ή διαδικασία) που ανταποκρίνεται σε περιβαλλοντικά ερεθίσματα (ή νέα δεδομένα) και στη συνέχεια τροποποιεί τη λειτουργία της για να μεγιστοποιήσει έναν δείκτη απόδοσης (19,20). Στην πράξη, η διαδικασία εκμάθησης υλοποιείται με χρήση μαθηματικών, στατιστικών, λογικής και προγραμματισμού. Αυτή επιτρέπει στο μοντέλο AI να εκπαιδεύεται σε δεδομένα σε μια επαναληπτική διαδικασία που αξιοποιεί προσαρμογή παραμέτρων με δοκιμή και σφάλμα, χρησιμοποιώντας κανόνες ενίσχυσης. Ο δείκτης απόδοσης μπορεί να είναι η ελαχιστοποίηση της διαφοράς μεταξύ προβλέψεων μοντέλων και πειραματικών δεδομένων.

Μόλις εκπαιδευτεί, το μοντέλο AI μπορεί να εφαρμοστεί σε νέα δεδομένα για τη λήψη αποφάσεων, η οποία πιθανόν να περιλαμβάνει εντοπισμό, διάκριση και ταξινόμηση. Ο πιθανός ρόλος της τεχνητής νοημοσύνης στην επεξεργασία των Big Data στην ιατρική ακριβείας έχει προταθεί λόγω των τεράστιων πόρων που απαιτούνται για την προγνωστική ανάλυση (21).

Μια σημαντική πρόκληση στη μελλοντική ανάπτυξη της τεχνητής νοημοσύνης είναι ο τρόπος εκμετάλλευσης των Big Data για την παραγωγή abstractions υψηλού επιπέδου που μπορούν να προσομοιώσουν μια ανθρώπινη υποκειμενική απόκριση. Σε κλινικό πλαίσιο, υπάρχει δυνατότητα για την ανάπτυξη ενός ψηφιακού εμπειρογνώμονα που διευκολύνει την αυτόματη μετατροπή των αποτελεσμάτων της παθολογίας σε γραπτή αναφορά ή προφορική εξήγηση.

Παραδείγματα προσεγγίσεων τεχνητής νοημοσύνης περιλαμβάνουν τη μηχανική εκμάθηση (machine learning), την αναγνώριση προτύπων (pattern recognition), τα έμπειρα συστήματα (expert systems) και τη ασαφή λογική (fuzzy logic). Πολλά μοντέλα τεχνητής νοημοσύνης έχουν ισχυρές στατιστικές βάσεις και κανόνες μάθησης που βασίζονται σε διαδικασίες μη γραμμικής βελτιστοποίησης (20,22).

Όπως προαναφέρθηκε, η μηχανική εκμάθηση ή *machine learning* είναι ένας κλάδος της τεχνητής νοημοσύνης. Ειδικότερα, εστιάζει στη μαθησιακή πτυχή της τεχνητής νοημοσύνης αναπτύσσοντας αλγόριθμους που αντιπροσωπεύουν καλύτερα ένα σύνολο δεδομένων. Σε αντίθεση με τον κλασικό προγραμματισμό, στον οποίο ένας αλγόριθμος μπορεί να κωδικοποιηθεί ρητά χρησιμοποιώντας γνωστά χαρακτηριστικά, η μηχανική εκμάθηση χρησιμοποιεί υποσύνολα δεδομένων, για να δημιουργήσει έναν αλγόριθμο που μπορεί να χρησιμοποιεί νέους ή διαφορετικούς συνδυασμούς χαρακτηριστικών και βαρών από αυτούς που μπορούν να προκύψουν από τις πρώτες αρχές (23,24). Στη μηχανική εκμάθηση, υπάρχουν τέσσερις κοινά χρησιμοποιούμενες μέθοδοι εκπαίδευσης, εκ των οποίων καθεμία είναι χρήσιμη για την επίλυση διαφορετικών προβλημάτων: εποπτευόμενη

(supervised), χωρίς επίβλεψη (unsupervised), ημιοπτευόμενη (semisupervised) και ενισχυτική μάθηση (reinforcement learning) (23–25).

Η εποπτευόμενη μάθηση στοχεύει στη μοντελοποίηση των εισροών και των αντίστοιχων γνωστών εκροών. Στην υπολογιστική μονάδα δίνονται μερικά παραδείγματα εισαγωγής (ερωτήσεις) με την αντίστοιχη έξοδο (οι σωστές απαντήσεις). Στη συνέχεια αυτή μαθαίνει μια συνάρτηση, ώστε να αντιστοιχίζει ερωτήσεις σε απαντήσεις, συχνά μέσω ταξινόμησης και παλινδρόμησης. Οι περισσότερες εφαρμογές χρησιμοποιούν εποπτευόμενη μάθηση. Αντίθετα, ο στόχος της μηχανικής μάθησης χωρίς επίβλεψη (unsupervised machine learning) είναι να βρει μοτίβα στα δεδομένα εισόδου χωρίς ανάδραση (feedback) από τον χρήστη. Τεχνικές όπως η ομαδοποίηση και η συσχέτιση αποτελούν τη βάση της μάθησης χωρίς επίβλεψη (19). Στην ιατρική, η μάθηση χωρίς επίβλεψη έχει τη δυνατότητα να αποκαλύψει παλαιότερα άγνωστους παθοφυσιολογικούς μηχανισμούς, οι οποίοι θα μπορούσαν με τη σειρά τους να οδηγήσουν σε νέα θεραπευτικά σχήματα (26).

Η πρόσφατη ανάπτυξη τεχνολογιών απεικόνισης και αλληλούχισης επιτρέπει συστηματική πρόοδο στην κλινική μελέτη του καρκίνου του πνεύμονα. Οι προσεγγίσεις που βασίζονται στο machine learning διαδραματίζουν κρίσιμο ρόλο στην ενσωμάτωση και ανάλυση αυτών των μεγάλων και πολύπλοκων συνόλων δεδομένων. Αυτά έχουν χαρακτηρίσει εκτενώς τον καρκίνο του πνεύμονα μέσω της διαφορετικής οπτικής που προσφέρουν στην αντιμετώπιση αυτής της ασθένειας (27).

Ο καρκίνος του πνεύμονα είναι ένας από τους πιο συχνά διεγνωσμένους καρκίνους και η κύρια αιτία θανάτων από καρκίνο παγκοσμίως. Περίπου 2,2

εκατομμύρια νέοι ασθενείς διαγιγνώσκονται με καρκίνο του πνεύμονα κάθε χρόνο (28) και το 75% από αυτούς πεθαίνουν μέσα σε πέντε χρόνια από τη διάγνωση (29). Η υψηλή ενδοκαρκινική ετερογένεια και η πολυπλοκότητα των καρκινικών κυττάρων που προκαλούν αντοχή στα φάρμακα καθιστούν τη θεραπεία του καρκίνου πιο δύσκολη (30).

Τις τελευταίες δεκαετίες, η συνεχής εξέλιξη των τεχνολογιών στην έρευνα για τον καρκίνο έχει συμβάλει σε πολλά μεγάλα συλλογικά έργα, τα οποία έχουν δημιουργήσει πολυάριθμες βάσεις δεδομένων κλινικής, ιατρικής απεικόνισης και αλληλουχίας (31–33). Αυτές οι βάσεις δεδομένων διευκολύνουν τους ερευνητές στη διερεύνηση ολοκληρωμένων προτύπων καρκίνου του πνεύμονα από τη διάγνωση, τη θεραπεία και τις ανταποκρίσεις στα κλινικά αποτελέσματα (34). Ειδικότερα, οι τρέχουσες μελέτες σχετικά με τις αναλύσεις omics, όπως η γονιδιωματική (genomics), η μεταγραφομική (transcriptomics), η πρωτεομική (proteomics) και η μεταβολομική (metabolomics), έχουν επεκτείνει τα εργαλεία και τις δυνατότητές μας για έρευνα (27).

Η παρούσα Διπλωματική Εργασία (ΔΕ) εστιάζει στην περιγραφή των εργαλείων που καθιστούν δυνατές τις ανωτέρω ερευνητικές προσεγγίσεις. Παρέχουν, δηλαδή, στον ερευνητή τη δυνατότητα να προσεγγίσει ένα ιατρικό ή βιολογικό ερώτημα με τρόπο ολιστικό που λαμβάνει υπόψη του πολλές παραμέτρους του υπό μελέτη συστήματος. Σε αυτό τα πλαίσιο διακρίνεται μια εξέλιξη στην έρευνα, δηλαδή από την πεποίθηση ότι κάθε ασθένεια μπορεί να αποδοθεί σε ένα ελαττωματικό γονίδιο, στην αντίληψη της κάθε ασθένειας ως αποτέλεσμα της αλληλεπίδρασης του περιβάλλοντος με το γονιδίωμα.

Εταιρεία	Πλατφόρμα	Ενίσχυση	Αλληλούχηση	Μήκος Ανάγνωσης	Παραγωγικότητα/ χρόνος ανά τρέξιμο	Κύριος Τύπος Σφάλματος	Συνικό Ποσοστό σφάλματος
Roche/454 Life Sciences	GS FLX Titanium XL+ GS FLX Titanium XLR70 GS Junior HiSeq 2000	PCR σε γαλάκτωμα	Πυροαλληλούχηση	Έως 1 kb Έως 600 bp ~400 bp 36 - 100 bp	700 Mb/23 ώρες 450 Mb/10 ώρες 35 Mb/10 ώρες 105-600 Gb/2- 11 ημέρες	Εισαγωγή/ Διαγραφή	0.5%
Illumina	Genome Analyzer Ix MiSeq	PCR γέφυρας	Αλληλούχηση μέσω σύνθεσης	35 - 150 bp 36 - 250 bp	10-95 Gb/2- 14 ημέρες 540 Mb-8.5 Gb/4- 39 ώρες	Αντικατάσταση	0.2%
Life Technologies/ Applied Biosystems	5500xl SOLiD™ system SOLiD™ 4 system	PCR σε γαλάκτωμα	Αλληλούχηση λήψης	35 - 75 bp 25 - 50 bp	10-15 Gb/ημέρα 25-100 Gb/3.5- 16 ημέρες	Αντικατάσταση	0.1%
Life Technologies/Ion Torrent	Ion Proton™ sequencer (Proton I chip) Ion PGM™ sequencer (318 chip)	PCR σε γαλάκτωμα	Αλληλούχηση ημιαγωγού	Έως 200 bp 35 - 200 bp	Έως 10 Gb/2-4 ώρες 300 Mb-1 Gb/0.9- 4.5 ώρες	Εισαγωγή/ Διαγραφή	1%
Helicos BioSciences	Heliscope™ single molecule sequencing	Καμία	Αλληλούχηση μοναδικού μορίου	25 - 55 bp	21-35 Gb/8 ημέρες	Διαγραφή	5%
Pacific Biosciences	PacBio RS	Καμία	Αλληλούχηση μοναδικού μορίου	250 bp - 10 kb	NA	Εισαγωγή/ Διαγραφή	15%

Πίνακας 1: Σύνοψη τεχνολογιών παραγωγής δεδομένων γενομικής.

Αναλυτής Μάζας	Ανάλυση μάζας	Ακρίβεια μάζας (ppm)	Ευαισθησία	Εύρος m/z	Δυνατότητα MS/MS	Πηγή ιόντων
IT	1.000 - 1.500	100 - 1.000	Picomole	50 - 2.000; 200 - 4.000	MS ⁿ	ESI
LTQ (LIT)	2.000	100 - 500	Femtomole	50 - 2.000; 200 - 4.000	MS ⁿ	ESI
Q-q-Q	1.000	100 - 1.500	Attomole	50 - 4.000	MS/MS	ESI
TOF	10.000 - 20.000	5 - 50	Femtomole	Χωρίς ανώτατο όριο	n/a	MALDI
TOF-TOF	10.000 - 40.000	5 - 50	Femtomole	Χωρίς ανώτατο όριο	MS/MS	MALDI
Q-q-TOF	10.000 - 40.000	5 - 50	Attomole	Χωρίς ανώτατο όριο	MS/MS	ESI; MALDI
LTQ-FTICR	50.000 - 800.000	1 - 2	Femtomole	50 - 2.000; 200 - 4.000	MS ⁿ	ESI; MALDI
LTQ-Orbitrap	50.000 - 500.000	<5	Femtomole	50 - 2.000; 200 - 4.000	MS ⁿ	ESI; MALDI

Πίνακας 2: Σύνοψη τεχνολογιών παραγωγής δεδομένων πρωτεομικής.

4. Κυριότερες κατηγορίες εργαλείων & Βάσεις Δεδομένων

Εργαλεία ευθυγράμμισης

Στη βιοπληροφορική, η ευθυγράμμιση αλληλουχίας είναι ένας τρόπος διάταξης των αλληλουχιών DNA, RNA ή πρωτεΐνης για τον εντοπισμό περιοχών ομοιότητας που μπορεί να είναι συνέπεια λειτουργικών, δομικών ή εξελικτικών σχέσεων μεταξύ των αλληλουχιών (35,36). Οι ευθυγραμμισμένες αλληλουχίες νουκλεοτιδίων ή αμινοξέων αντιπροσωπεύονται τυπικά ως σειρές μέσα σε ένα πίνακα (matrix). Εισάγονται κενά μεταξύ των υπολειμμάτων έτσι ώστε πανομοιότυποι ή παρόμοιοι χαρακτήρες να ευθυγραμμίζονται σε διαδοχικές στήλες (37).

Πολύ σύντομες ή πολύ παρόμοιες ακολουθίες μπορούν να ευθυγραμμιστούν με το χέρι. Ωστόσο, τα περισσότερα ενδιαφέροντα προβλήματα απαιτούν την ευθυγράμμιση μακρών, εξαιρετικά μεταβλητών ή εξαιρετικά πολυάριθμων ακολουθιών που δεν μπορούν να ευθυγραμμιστούν αποκλειστικά με ανθρώπινη προσπάθεια. Αντίθετα, η ανθρώπινη γνώση εφαρμόζεται στην κατασκευή αλγορίθμων για την παραγωγή ευθυγραμμίσεων ακολουθιών υψηλής ποιότητας και περιστασιακά στην προσαρμογή των τελικών αποτελεσμάτων ώστε να αντικατοπτρίζουν μοτίβα που είναι δύσκολο να αναπαραστηθούν αλγοριθμικά (ειδικά στην περίπτωση αλληλουχιών νουκλεοτιδίων). Οι υπολογιστικές προσεγγίσεις για τη στοίχιση ακολουθιών εμπίπτουν γενικά σε δύο κατηγορίες: καθολική ευθυγράμμιση (global alignment) και τοπική ευθυγράμμιση (local alignment). Ο υπολογισμός μιας καθολικής ευθυγράμμισης είναι μια μορφή καθολικής βελτιστοποίησης που

"αναγκάζει" τη στοίχιση να εκτείνεται σε όλο το μήκος όλων των αλληλουχιών αναφοράς. Αντίθετα, οι τοπικές ευθυγραμμίσεις προσδιορίζουν περιοχές ομοιότητας μέσα σε μακριές ακολουθίες που συχνά είναι ευρέως αποκλίνουσες συνολικά. Οι τοπικές ευθυγραμμίσεις είναι συχνά προτιμότερες, αλλά μπορεί να είναι πιο δύσκολο να υπολογιστούν λόγω της πρόσθετης πρόκλησης του προσδιορισμού των περιοχών ομοιότητας (38). Στο πρόβλημα της ευθυγράμμισης ακολουθιών έχουν εφαρμοστεί μια ποικιλία υπολογιστικών αλγορίθμων. Αυτές περιλαμβάνουν αργές αλλά τυπικά σωστές μεθόδους όπως ο δυναμικός προγραμματισμός. Αυτά περιλαμβάνουν επίσης αποτελεσματικούς, ευρετικούς αλγόριθμους ή πιθανολογικές μεθόδους που έχουν σχεδιαστεί για αναζήτηση βάσεων δεδομένων μεγάλης κλίμακας, που δεν εγγυώνται την εύρεση των καλύτερων αντιστοιχίσεων (39). Μερικά παραδείγματα γνωστών προγραμμάτων ευθυγράμμισης είναι τα: Bowtie, TopHat και HISAT.

Οι ευθυγραμμίσεις ακολουθιών μπορούν να αποθηκευτούν σε μια μεγάλη ποικιλία μορφών αρχείων κειμένου, πολλές από τις οποίες αναπτύχθηκαν αρχικά σε συνδυασμό με ένα συγκεκριμένο πρόγραμμα ή υλοποίηση ευθυγράμμισης. Τα περισσότερα διαδικτυακά εργαλεία επιτρέπουν περιορισμένο αριθμό μορφών εισόδου και εξόδου, όπως η μορφή FASTA και η μορφή GenBank και η έξοδος δεν είναι εύκολα επεξεργάσιμη. Διατίθενται αρκετά προγράμματα μετατροπής που παρέχουν γραφικά διαδραστικά περιβάλλοντα ή/και γραμμής εντολών, όπως το READSEQ και το EMBOSS. Υπάρχουν επίσης πολλά πακέτα προγραμματισμού που παρέχουν αυτή τη λειτουργία μετατροπής, όπως τα BioPython, BioRuby και BioPerl. Τα αρχεία SAM/BAM χρησιμοποιούν τη μορφή συμβολοσειράς CIGAR (Compact

Idiosyncratic Gapped Alignment Report) για να αναπαραστήσουν μια στοίχιση μιας ακολουθίας σε μια αναφορά κωδικοποιώντας μια ακολουθία γεγονότων (π.χ. αντιστοίχιση/αναντιστοιχία, εισαγωγές, διαγραφές) (40).

Εργαλεία καταμέτρησης χαρακτηριστικών

Τα εργαλεία καταμέτρησης χαρακτηριστικών (feature counting) είναι προγράμματα περίληψης ανάγνωσης γενικής χρήσης που μετρούν χαρτογραφημένες αναγνώσεις για γονιδιωματικά χαρακτηριστικά όπως γονίδια, εξόνια, υποκινητές, σώματα γονιδίων, γονιδιωματικά δοχεία και χρωμοσωμικές θέσεις. Μπορεί να χρησιμοποιηθούν για τη μέτρηση αναγνώσεων RNA-seq και γονιδιωματικού DNA-seq.

Αυτά λαμβάνουν ως είσοδο αρχεία ευθυγράμμισης SAM ή BAM και ένα αρχείο σχολιασμού που περιλαμβάνει χρωμοσωμικές συντεταγμένες χαρακτηριστικών. Εξάγουν αριθμούς αναγνώσεων που έχουν εκχωρηθεί σε χαρακτηριστικά (ή μετα-χαρακτηριστικά). Εξάγουν επίσης πληροφορίες στατιστικών στοιχείων για τα συνολικά αποτελέσματα σύνοψης, συμπεριλαμβανομένου του αριθμού των αναγνώσεων που ανατέθηκαν με επιτυχία και του αριθμού των αναγνώσεων που απέτυχαν να εκχωρηθούν για διάφορους λόγους. Το αρχείο σχολιασμού θα πρέπει να είναι είτε σε μορφή GTF είτε σε μορφή απλοποιημένου σχολιασμού (SAF).

Κάθε καταχώρηση στο παρεχόμενο αρχείο σχολιασμού λαμβάνεται ως χαρακτηριστικό (π.χ. ένα εξόνιο). Ένα μετα-χαρακτηριστικό είναι η συνάθροιση ενός συνόλου χαρακτηριστικών (π.χ. ενός γονιδίου). Για παράδειγμα, το πρόγραμμα featureCounts χρησιμοποιεί το χαρακτηριστικό gene_id που είναι

διαθέσιμο στον σχολιασμό μορφής GTF (ή στη στήλη GeneID στον σχολιασμό μορφής SAF) για να ομαδοποιήσει τα χαρακτηριστικά σε μετα-χαρακτηριστικά, π.χ. Τα χαρακτηριστικά που ανήκουν στο ίδιο μετα-χαρακτηριστικό έχουν το ίδιο αναγνωριστικό γονιδίου.

Το featureCounts μπορεί να μετρήσει αναγνώσεις είτε σε επίπεδο χαρακτηριστικών είτε σε επίπεδο μετα-χαρακτηριστικών. Κατά τη σύνοψη των αναγνώσεων σε επίπεδο μετα-δυνατοτήτων, θα προστεθούν οι μετρήσεις ανάγνωσης που λαμβάνονται για τα χαρακτηριστικά που περιλαμβάνονται στην ίδια μετα-λειτουργία, ώστε να προκύψει ο αριθμός αναγνωσμένων για το αντίστοιχο μετα-χαρακτηριστικό (41). Ένα εναλλακτικό πρόγραμμα feature counting είναι επίσης το htseq-count.

Εργαλεία ταυτοποίησης πρωτεϊνών – Proteome Discoverer

Οι ερευνητές που ασχολούνται με την πρωτεομική αντιμετωπίζουν μια ενδιαφέρουσα πρόκληση: πώς να επιλέξουν ανάμεσα στα δεκάδες pipelines επεξεργασίας και ανάλυσης δεδομένων που είναι διαθέσιμα για τη μετατροπή των διαδοχικών αρχείων φασματομετρίας μάζας σε ταυτοποιήσεις πρωτεϊνών. Σήμερα, υπάρχουν πάνω από 1000 διαφορετικά εργαλεία για την επεξεργασία δεδομένων proteomics (42). Αυτός ο αριθμός περιλαμβάνει τόσο το ενεργό λογισμικό όσο και τα εργαλεία που έχουν περιθωριοποιηθεί λόγω των ραγδαίων εξελίξεων στον τομέα της ανάλυσης πρωτεϊνών. Τα κύρια εργαλεία αναζήτησης που χρησιμοποιούνται από τους περισσότερους ερευνητές σήμερα είναι πολύ λιγότερα. Οι περισσότεροι χρήστες είναι πιθανό να προτιμούν εργαλεία κυρίως μεταξύ των MaxQuant (43,44), Mascot (45),

SpectroNaut (46), Proteome Discoverer και PEAKS (47). Οι υπόλοιποι χρήστες προτιμούν είτε άλλα ισχυρά εμπορικά εργαλεία, ή εργαλεία ανοιχτού κώδικα (open source).

Το Proteome Discoverer (PD) είναι ένα εμπορικό προϊόν της Thermo Fisher Scientific που κυκλοφόρησε για πρώτη φορά το 2007, αντικαθιστώντας το παλαιότερο BioWorks. Σήμερα, το PD προσφέρει στον χρήστη ένα ισχυρό και ευέλικτο περιβάλλον με εμπορικά και ελεύθερα διαθέσιμα εργαλεία. Με αυτό μπορούν να πραγματοποιηθούν ροές εργασιών (workflows) πρωτεομικής, όπως η ποσοτικοποίηση σεσημασμένων (labeled) ή μη (label-free) δεδομένων, η γλυκοπρωτεομική και άλλα (48,49).

Τα τελευταία 14 χρόνια, το λογισμικό έχει εξελιχθεί για να ανταποκρίνεται στις προκλήσεις που αντιμετωπίζουν οι επιστήμονες πρωτεομικής. Ωστόσο, τέσσερα κοινά χαρακτηριστικά ήταν θεμελιώδη σε όλες τις εκδόσεις του λογισμικού:

- *Ευέλικτη εισαγωγή δεδομένων:* Ενώ το PD είναι προϊόν της Thermo Fisher Scientific και προορίζεται για την επεξεργασία αρχείων από αυτόν τον προμηθευτή, το PD αποδέχεται δεδομένα και από άλλες πλατφόρμες προμηθευτών εφόσον τα αρχεία έχουν μετατραπεί στην κατάλληλη μορφοποίηση (Mascot Generic Format, mzML και mzXML).
- *Ο συνδυασμός των αποτελεσμάτων αναζήτησης:* Ένα πρώιμο χαρακτηριστικό που ξεχώρισε το PD από άλλα εργαλεία ήταν η δυνατότητα συνδυασμού αποτελεσμάτων από πολλές μηχανές αναζήτησης.

- *Άμεση πρόσβαση στα αρχικά φάσματα MS/MS που οδήγησαν στην εκάστοτε ταυτοποίηση:* Το κύριο πλεονέκτημα του PD σε σύγκριση με άλλα εργαλεία αναζήτησης είναι η συνεχής ευκολία πρόσβασης στα αρχικά δεδομένα MS/MS.

Ταχεία σύνδεση ανακάλυψης και ταυτοποίησης: Τα αποτελέσματα μπορούν να εξαχθούν απευθείας σε μορφή λίστας για όλα τα όργανα Orbitrap.

Στατιστικά εργαλεία

Η επιλογή της κατάλληλης στατιστικής μεθόδου είναι πολύ σημαντικό βήμα στην ανάλυση των βιοϊατρικών δεδομένων. Μια λανθασμένη επιλογή της στατιστικής μεθόδου όχι μόνο δημιουργεί κάποιο σοβαρό πρόβλημα κατά την ερμηνεία των ευρημάτων αλλά επηρεάζει και το συμπέρασμα της μελέτης. Στη στατιστική, για κάθε συγκεκριμένη κατάσταση, διατίθενται στατιστικές μέθοδοι για ανάλυση και ερμηνεία των δεδομένων. Για να επιλεγεί η κατάλληλη στατιστική μέθοδος, πρέπει να γνωρίζει κανείς την υπόθεση και τις συνθήκες των στατιστικών μεθόδων, ώστε να μπορεί να επιλεγεί η κατάλληλη στατιστική μέθοδος για ανάλυση δεδομένων (50).

Πέρα από τη γνώση των στατιστικών μεθόδων, μια άλλη πολύ σημαντική πτυχή είναι η φύση και το είδος των δεδομένων που συλλέγονται και ο στόχος της μελέτης, διότι σύμφωνα με τον στόχο επιλέγονται οι αντίστοιχες στατιστικές μέθοδοι που είναι κατάλληλες για δεδομένα. Η πρακτική λανθασμένης ή ακατάλληλης στατιστικής μεθόδου είναι σύνηθες φαινόμενο στα δημοσιευμένα άρθρα στη βιοϊατρική έρευνα. Εσφαλμένες στατιστικές μέθοδοι μπορούν να παρατηρηθούν σε πολλές συνθήκες, όπως η χρήση μη ζευγαρωμένου t-test σε

ζευγαρωμένα δεδομένα ή η χρήση παραμετρικού τεστ για τα δεδομένα που δεν ακολουθούν την κανονική κατανομή κ.λπ.

Επί του παρόντος, πολλά στατιστικά λογισμικά όπως SPSS, R, Stata και SAS είναι διαθέσιμα και χρησιμοποιώντας αυτά τα λογισμικά, μπορεί κανείς να πραγματοποιήσει εύκολα τη στατιστική ανάλυση, αλλά η επιλογή του κατάλληλου στατιστικού τεστ εξακολουθεί να είναι μια δύσκολη εργασία για τους βιοϊατρικούς ερευνητές, ειδικά αυτούς με μη στατιστικό υπόβαθρο (51). Στην ανάλυση δεδομένων χρησιμοποιούνται δύο κύριες στατιστικές μέθοδοι: η περιγραφική στατιστική, η οποία συνοψίζει δεδομένα χρησιμοποιώντας δείκτες όπως ο μέσος όρος, η διάμεσος, η τυπική απόκλιση και μια άλλη είναι η συμπερασματική στατιστική, η οποία εξάγει συμπεράσματα από δεδομένα χρησιμοποιώντας στατιστικά τεστ όπως το student's t-test, το ANOVA test. και τα λοιπά (52).

Για παράδειγμα, στα πλαίσια της R, έχουν ανακύψει αξιοσημείωτα εργαλεία διαφορικής ανάλυσης για πιο εξειδικευμένες αναλύσεις, όπως το DESeq2, το limma και το GSEA. Ενώ και τα τρία εργαλεία αυτά αρχικά δημιουργήθηκαν για την ανάλυση δεδομένων μικροσυστοιχιών RNA, σήμερα βρίσκουν ευρεία εφαρμογή στον κλάδο της μεταγραφομικής και της πρωτεογενομικής.

Εργαλεία εμπλουτισμού

Οι τεχνικές εμπλουτισμού για δεδομένα «omics» είναι βασικά εργαλεία για την κατανόηση πολύπλοκων βιολογικών συστημάτων. Αυτά τα εργαλεία μειώνουν την πολυπλοκότητα των δεδομένων, βελτιώνουν την ερμηνεία και την

κατανόηση των βιολογικών συστημάτων και βοηθούν στη δημιουργία υποθέσεων (53,54). Τα εργαλεία εμπλουτισμού υποδηλώνουν οποιαδήποτε αναλυτική τεχνική που επωφελείται από πληροφορίες μοριακών μονοπατιών ή δικτύων για την απόκτηση εικόνας ενός βιολογικού συστήματος (55).

Η πιο ευρέως χρησιμοποιούμενη μεθοδολογία για την εκτέλεση μιας τέτοιας ανάλυσης ονομάζεται ανάλυση λειτουργικού εμπλουτισμού ή υπερ-αναπαράστασης (over-representation analysis ή ORA) (56). Αυτή η ανάλυση αναζητά λέξεις-κλειδιά ή περιγραφείς (descriptors) του συνόλου των μορίων που παρουσιάζουν ενδιαφέρον (π.χ. αυτά που εκφράζονται υπερβολικά) σε σχέση με ένα σύνολο αναφοράς (π.χ. ολόκληρο το γονιδίωμα, μεταγράψωμα, πρωτέωμα, μεταβόλωμα ή το σύνολο μορίων που ανιχνεύονται από την τεχνολογία που χρησιμοποιείται) (57). Οι κλασικές αναλύσεις εμπλουτισμού χρησιμοποιούν το ακριβές τεστ του Fisher (Fisher's exact test), αλλά πολλές άλλες μέθοδοι εμπλουτισμού έχουν προκύψει από αυτό, π.χ. υπεργεωμετρική, στατιστικές δοκιμές Kolmogorov–Smirnov ή Wilcoxon (54,56).

Η ανάλυση μοριακών μετρήσεων υψηλής απόδοσης σε λειτουργικό επίπεδο είναι πολύ ελκυστική για δύο λόγους. Αρχικά, η ομαδοποίηση χιλιάδων γονιδίων, πρωτεϊνών και/ή άλλων βιολογικών μορίων με βάση τα μονοπάτια στα οποία εμπλέκονται μειώνει την πολυπλοκότητα σε μερικές μόνο εκατοντάδες μονοπάτια για το εκάστοτε πείραμα. Δεύτερον, ο εντοπισμός ενεργών οδών που διαφέρουν μεταξύ δύο συνθηκών μπορεί να έχει περισσότερη επεξηγηματική δύναμη από μια απλή λίστα διαφορετικών γονιδίων ή πρωτεϊνών (58).

Μέχρι στιγμής υπάρχει μια πληθώρα εργαλείων εμπλουτισμού, εκ των οποίων μερικά γνωστά είναι το DAVID, το PANTHER, το Metascape, το GSEA και το clusterProfiler.

Εργαλεία οπτικοποίησης δεδομένων

Η οπτικοποίηση ήταν από καιρό το κλειδί για να βοηθήσει στην κατανόηση των βιολογικών συστημάτων, όπως ο μεταβολισμός (59), η σηματοδότηση (60) και η ρύθμιση της γονιδιακής έκφρασης (61). Τα τελευταία χρόνια, η μελέτη τέτοιων συστημάτων έχει επηρεαστεί βαθιά από την ανάπτυξη ενός ευρέος φάσματος πειραματικών μεθόδων υψηλής απόδοσης, με αποτέλεσμα έναν πολύ αυξημένο όγκο πολύπλοκων, διασυνδεδεμένων δεδομένων.

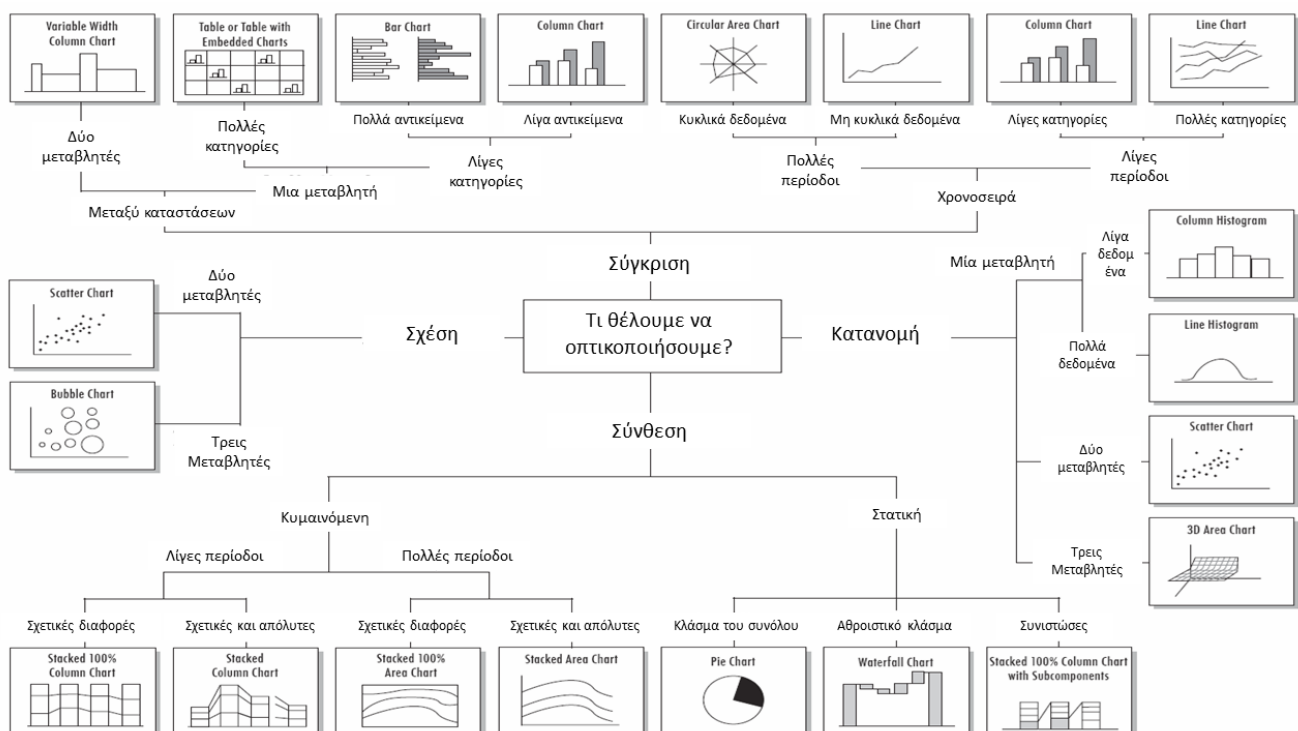
Καθώς οι πειραματικές μέθοδοι υψηλής απόδοσης έχουν γίνει πιο συνήθεις, πολλοί επιστήμονες χρησιμοποιούν οπτικοποίηση δικτύων και μονοπατιών για να καταγράψουν και να κοινοποιήσουν τα ευρήματά τους. Υπάρχουν τώρα πάνω από 300 διαδικτυακοί πόροι που παρέχουν πρόσβαση σε πολλές χιλιάδες μονοπάτια και δίκτυα που τεκμηριώνουν εκατομμύρια αλληλεπιδράσεις μεταξύ πρωτεϊνών, γονιδίων και μικρών μορίων (62).

Υπήρξε αντίστοιχη αύξηση στην ανάπτυξη εργαλείων οπτικοποίησης για δεδομένα βιολογίας συστημάτων (63–65). Αυτά τα εργαλεία είναι πολύ διαφορετικά, αλλά μπορούν να χωριστούν ευρέως σε δύο εν μέρει επικαλυπτόμενες κατηγορίες. Η πρώτη αποτελείται από εργαλεία που επικεντρώνονται σε αυτοματοποιημένες μεθόδους ερμηνείας και εξερεύνησης μεγάλων βιολογικών δικτύων και η δεύτερη αποτελείται από εργαλεία που

επικεντρώνονται στη συναρμολόγηση και την επιμέλεια των μονοπατιών. Πολλά από αυτά τα εργαλεία είναι στενά ενσωματωμένα με δημόσιες βάσεις δεδομένων, επιτρέποντας έτσι στους χρήστες να οπτικοποιούν και να ερμηνεύουν τα δικά τους δεδομένα στο πλαίσιο της ήδη υπάρχουσας γνώσης.

Για τους χρήστες και τους προγραμματιστές αυτών των εργαλείων οπτικοποίησης, μία από τις βασικές προκλήσεις είναι πώς να επωφεληθούν από τον υπερμεγέθη όγκο στα δεδομένα βιολογίας συστημάτων χωρίς να κατακλύζονται από αυτά—ή, πρακτικά, πώς να παρουσιάζονται τα δεδομένα στο σωστό επίπεδο λεπτομέρειας, με έναν συνεκτικό και διορατικό τρόπο. Σαφώς, οι απαντήσεις εξαρτώνται από το πλαίσιο της εκάστοτε έρευνας (66).

Συνήθως τα εργαλεία οπτικοποίησης που χρησιμοποιούνται είναι διαθέσιμα στις ίδιες γλώσσες προγραμματισμού που χειρίζονται και την ανάλυση δεδομένων, όπως η Python και η R. Μερικά από τα πιο γνωστά εργαλεία οπτικοποίησης είναι το plotly για την Python και το ggplot2 για την R. Επιπλέον, υπάρχουν και εργαλεία τα οποία οπτικοποιούν δεδομένα ταυτόχρονα με την ανάλυσή τους, όπως το Cytoscape, το Metascape, το GSEA και η StringDB.



Εικόνα 1: Διάγραμμα ροής για οπτικοποίηση κάθε τύπου δεδομένων.

Στο διάγραμμα εμφανίζονται πολλαπλοί τρόποι οπτικοποίησης δεδομένων. Ανάλογα με τους τύπους δεδομένων που χρειάζεται να οπτικοποιηθούν, ο ερευνητής μπορεί να ακολουθήσει αυτό το διάγραμμα ροής προκειμένου να επιλέξει το γράφημα που αρμόζει στα δεδομένα του. Για παράδειγμα, αν ο ερευνητής θέλει να οπτικοποιήσει τη σχέση δύο μεταβλητών, μπορεί να το κάνει με ένα γράφημα διασποράς (scatter chart).

Βάσεις Δεδομένων

Στην πληροφορική, μια βάση δεδομένων είναι μια οργανωμένη συλλογή δεδομένων ή ένας τύπος αποθήκευσης δεδομένων που βασίζεται στη χρήση ενός συστήματος διαχείρισης βάσεων δεδομένων (DBMS), του λογισμικού που αλληλεπιδρά με τους τελικούς χρήστες, τις εφαρμογές και την ίδια τη βάση δεδομένων για τη συλλογή και ανάλυση των δεδομένων. Το DBMS περιλαμβάνει επιπλέον τις βασικές εγκαταστάσεις που παρέχονται για τη διαχείριση της βάσης δεδομένων. Το σύνολο της βάσης δεδομένων, του DBMS και των σχετικών εφαρμογών μπορεί να αναφέρεται ως σύστημα βάσης δεδομένων. Συχνά ο όρος αυτός χρησιμοποιείται επίσης χαλαρά για να αναφέρεται σε οποιοδήποτε από τα DBMS, το σύστημα βάσης δεδομένων ή μια εφαρμογή που σχετίζεται με τη βάση δεδομένων.

UniProt

Η Universal Protein Resource (UniProt) (67) είναι μια ολοκληρωμένη πηγή για δεδομένα αλληλουχίας πρωτεϊνών και σχολιασμού. Η UniProt αποτελείται από τη γνωσιακή βάση UniProt (UniProtKB), τα συμπλέγματα αναφοράς UniProt (UniRef) και το αρχείο UniProt (UniParc). Η κοινοπραξία της UniProt και των ιδρυμάτων EMBL-EBI, SIB και PIR έχουν δεσμευτεί για τη μακροπρόθεσμη διατήρηση της βάσης δεδομένων UniProt. Η UniProt είναι μια συνεργασία μεταξύ του Ευρωπαϊκού Ινστιτούτου Βιοπληροφορικής (EMBL-EBI), του SIB Swiss Institute of Bioinformatics (SIB) και του Protein Information Resource (PIR). Σε αυτά τα τρία ινστιτούτα, περισσότερα από 100 άτομα

εμπλέκονται σε διαφορετικές εργασίες, όπως η επιμέλεια των βάσεων δεδομένων, η ανάπτυξη λογισμικού και η τεχνική υποστήριξη.

Σε αυτή τη βάση δεδομένων, μπορεί κάποιος να βρει τις εξής πληροφορίες για κάθε καταχώρηση-πρωτεΐνη:

1. Όνομα πρωτεΐνης – γονιδίου
2. Λειτουργία
3. Ενεργότητα κατάλυσης, συμπαράγοντες, καταλυτικά κατάλοιπα
4. Ενδοκυτταρική τοπολογία
5. Αλληλεπιδράσεις με άλλες πρωτεΐνες
6. Προφίλ έκφρασης
7. Ρόλοι δομικών περιοχών (domains) της πρωτεΐνης
8. Τοποθεσία πρόσδεσης ιόντων, υποστρωμάτων και συμπαράγοντων και
9. Εναλλακτικές μορφές της πρωτεΐνης

Παρότι η UniProt είναι η πιο ευρέως γνωστή βάση δεδομένων, υπάρχουν και άλλες που μπορούν να λειτουργήσουν συμπληρωματικά με αυτή. Μερικές από αυτές είναι η PRIDE (68), η PeptideAtlas (69), η MaxQB (70) και η Human Protein Atlas (71).

Όνομα Βάσης Δεδομένων	Περιγραφή περιεχομένου	Τύποι Δεδομένων	Επιλογές Αναζήτησης	Διεπαφή Χρήστη
UniProt	Αλληλουχίες πρωτεϊνών και λειτουργικές πληροφορίες	Αλληλουχίες πρωτεϊνών Χαρακτηρισμοί Μονοπάτια	Λέξεις-Κλειδιά BLAST Προηγμένη αναζήτηση	Διαδίκτυακή πλατφόρμα API
PRIDE	Δεδομένα φασματοσκοπίας μάζας Πειράματα πρωτεομικής	Ταυτοποιήσεις πεπτιδίων και πρωτεϊνών Πειραματικά μεταδεδομένα	Λέξεις-Κλειδιά Επιλογές Φιλτραρίσματος	Διαδίκτυακή πλατφόρμα API
PeptideAtlas	Ταυτοποιήσεις πεπτιδίων και πρωτεϊνών Δεδομένα ποσοτικοποίησης	Αλληλουχίες πεπτιδίων Ταυτοποιήσεις πρωτεϊνών Τιμές ποσοτικοποίησης	Λέξεις-Κλειδιά Προηγμένη αναζήτηση	Διαδίκτυακή πλατφόρμα
MaxQB	Δεδομένα ποσοτικοποίησης πρωτεομικής Αφθονία πρωτεϊνών	Επίπεδα έκφρασης πρωτεϊνών Μετα-μεταφραστικές τροποποιήσεις	Όνομα πρωτεΐνης Επιλογές Φιλτραρίσματος	Διαδίκτυακή πλατφόρμα
Human Protein Atlas	Έκφραση πρωτεϊνών σε ανθρώπινα κύτταρα και ιστούς	Προφίλ έκφρασης πρωτεϊνών Ενδοκυτταρική τοπολογία	Αναζήτηση ιστού ή κυτταρικής σειράς	Διαδίκτυακή πλατφόρμα API

Πίνακας 3: Βάσεις δεδομένων πρωτεϊνών.

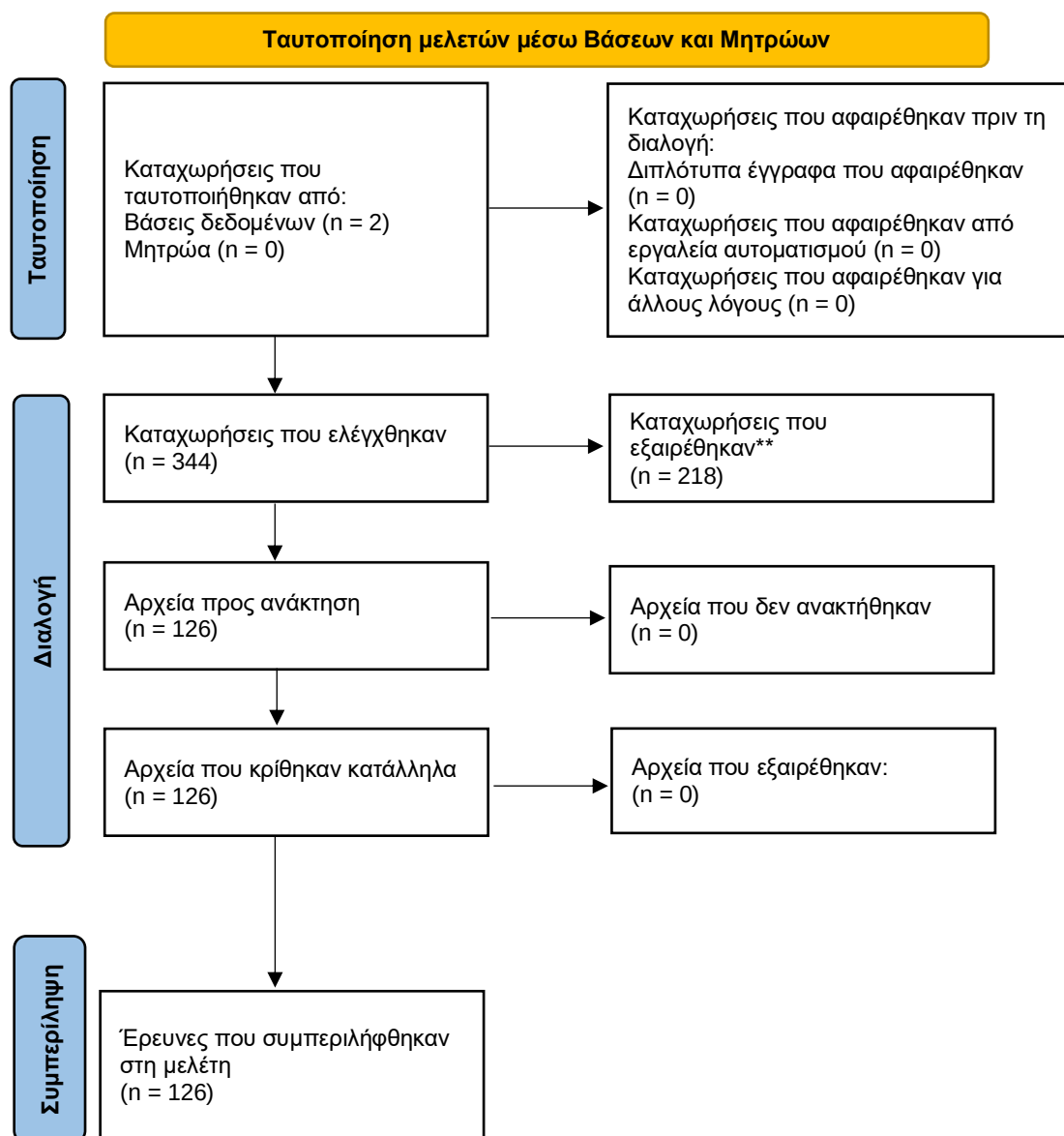
5. Σκοπός της εργασίας

Σκοπός της παρούσας εργασίας είναι η εύρεση και καταγραφή των κυριότερων κατηγοριών βιοπληροφορικών εργαλείων που μπορούν να χρησιμοποιηθούν για έρευνες πρωτομικής και πρωτεογενομικής. Επιπροσθέτως δύναται να χρησιμεύσει ως εγχειρίδιο για τον ερευνητή που επιθυμεί να εισαχθεί στον κόσμο της βιοπληροφορικής ανάλυσης. Παρότι τα περισσότερα από αυτά τα εργαλεία μπορούν να χρησιμοποιηθούν ευρέως, αυτή η εργασία εστιάζει στη χρήση τους στα πλαίσια της έρευνας για τον καρκίνο του πνεύμονα.

6. Υλικά και Μέθοδοι

Προκειμένου να εξαχθεί το βιβλιογραφικό υλικό για τη συγγραφή της εργασίας, χρησιμοποιήθηκαν ευριστικά εργαλεία (heuristic tools) όπως το Google Scholar, η PubMed καθώς και εγχειρίδια (vignettes) για τη χρήση του εκάστοτε εργαλείου. Κατά την αναζήτηση βιβλιογραφίας χρησιμοποιήθηκαν οι λέξεις-κλειδιά: *genome analysis*, *omics* και *lung cancer* για την αρχική εύρεση πηγών. Στη συνέχεια οι ανασκοπήσεις (reviews) που χρησιμοποιήθηκαν σε αυτή τη μελέτη ελέγχθηκαν ως προς τη βιβλιογραφία τους για περεταίρω πηγές.

Συγκεκριμένα, το σύνολο των μελετών που συγκεντρώθηκαν ελέγχθηκαν ως προς τη συσχέτισή τους με εργαλεία omics και την εφαρμογή τους με τον καρκίνο του πνεύμονα. Λεπτομέρειες σχετικότερα με τη διαδικασία εύρεσης βιβλιογραφίας εξηγούνται στο παρακάτω PRISMA flow chart (72) (Εικόνα 2).



Εικόνα 2: PRISMA flow chart στο οποίο αναλύεται η διαδικασία εύρεσης και ελέγχου βιβλιογραφίας για αυτή την εργασία.

7. Εργαλεία Βιοπληροφορικής

TopHat

Το TopHat (73) είναι ένα πακέτο λογισμικού που προσδιορίζει θέσεις ματίσματος (splice sites) εξαρχής με χαρτογράφηση μεγάλης κλίμακας των αναγνώσεων RNA-Seq. Το TopHat ευθυγραμμίζει αναγνώσεις (mapping reads) για να συνδυάσει θέσεις σε ένα γονιδίωμα θηλαστικού με ρυθμό ~2,2 εκατομμύρια αναγνώσεις ανά ώρα CPU. Αντί να φιλτράρει πιθανές τοποθεσίες σύνδεσης με ένα σχήμα βαθμολόγησης, το TopHat ευθυγραμμίζει όλους τους ιστότοπους, βασιζόμενος σε μια αποτελεσματική κωδικοποίηση 2 bit ανά βάση και σε μια διάταξη δεδομένων που χρησιμοποιεί αποτελεσματικά τη μνήμη cache σε σύγχρονους επεξεργαστές. Αυτή η στρατηγική λειτουργεί καλά στην πράξη, επειδή το TopHat αρχικά χαρτογραφεί αναγνώσεις χωρίς διασταύρωση (αυτές που περιέχονται στα εξόνια) χρησιμοποιώντας το Bowtie [1], ένα εξαιρετικά γρήγορο πρόγραμμα χαρτογράφησης σύντομης ανάγνωσης. Το Bowtie δημιουργεί ένα ευρετήριο με βάση το γονιδίωμα αναφοράς (indexing) χρησιμοποιώντας μια τεχνική δανεισμένη από τη συμπίεση δεδομένων, τον μετασχηματισμό Burrows-Wheeler. Αυτή η αποδοτική σε μνήμη δομή δεδομένων επιτρέπει στο Bowtie να σαρώνει τις αναγνώσεις σε ένα γονιδίωμα θηλαστικού χρησιμοποιώντας περίπου 2 GB.

Το TopHat βρίσκει διασταυρώσεις αντιστοιχίζοντας τις αναγνώσεις στην αναφορά σε δύο φάσεις. Στην πρώτη φάση, η ακολουθία εργασιών (pipeline) χαρτογραφεί όλες τις αναγνώσεις στο γονιδίωμα αναφοράς χρησιμοποιώντας το Bowtie. Όλες οι αναγνώσεις που δεν αντιστοιχούν στο γονιδίωμα παραμερίζονται ως «αρχικά μη χαρτογραφημένες αναγνώσεις» ή ως

αναγνώσεις IUM (initially unread maps). Το Bowtie αναφέρει, για κάθε ανάγνωση, μία ή περισσότερες ευθυγραμμίσεις (alignments) που δεν περιέχουν περισσότερες από μερικές αναντιστοιχίες (mismatches, δύο από προεπιλογή) στα άκρα 5' της ανάγνωσης με μήκος s (όπου s : παράμετρος που ορίζεται από τον χρήστη). Το υπόλοιπο τμήμα της ανάγνωσης στο άκρο 3' ενδέχεται να έχει πρόσθετες αναντιστοιχίες, υπό την προϋπόθεση ότι η σταθμισμένη απόσταση Hamming (Phred-quality-weighted Hamming distance) είναι μικρότερη από ένα καθορισμένο όριο (70 από προεπιλογή). Αυτή η πολιτική βασίζεται στην εμπειρική παρατήρηση ότι το 5' άκρο μιας ανάγνωσης περιέχει λιγότερα σφάλματα αλληλουχίας από το 3' άκρο. Το TopHat επιτρέπει στο Bowtie να αναφέρει περισσότερες από μία ευθυγραμμίσεις για μια ανάγνωση (10 από προεπιλογή) και καταστέλλει όλες τις στοιχίσεις για αναγνώσεις που έχουν περισσότερες από αυτόν τον αριθμό. Αυτή η πολιτική επιτρέπει την αναφορά των λεγόμενων «πολλαπλών αναγνώσεων» από γονίδια με πολλαπλά αντίγραφα, αλλά αποκλείει ευθυγραμμίσεις σε ακολουθίες χαμηλής πολυπλοκότητας, με τις οποίες συχνά ευθυγραμμίζονται οι αποτυχημένες αναγνώσεις. Οι αναγνώσεις χαμηλής πολυπλοκότητας δεν περιλαμβάνονται στο σύνολο αναγνώσεων IUM, απλά απορρίπτονται.

Στη συνέχεια, το TopHat συγκεντρώνει τις αντιστοιχισμένες αναγνώσεις χρησιμοποιώντας τη μονάδα συναρμολόγησης στο Maq. Το TopHat εξάγει τις ακολουθίες για τις προκύπτουσες νησίδες (islands) συνεχόμενης ακολουθίας από την αραιή ομοφωνία (sparse consensus), συνάγοντας ότι είναι υποτιθέμενα εξόνια. Για να δημιουργήσει τις ακολουθίες νησίδων, το TopHat επικαλείται την υποεντολή συναρμολόγησης Maq (παράμετρος *flag -s*) η οποία παράγει ένα συμπαγές αρχείο *consensus* που περιέχει τις καλεσμένες βάσεις

και τις αντίστοιχες βάσεις αναφοράς. Επειδή αυτό το αρχείο μπορεί να περιλαμβάνει λανθασμένα καλεσμένες βάσεις λόγω σφαλμάτων αλληλουχίας σε περιοχές χαμηλής κάλυψης, τέτοιες νησίδες μπορεί να αποτελούν «ψευδοομοφωνία». Για τυχόν θέσεις χαμηλής κάλυψης ή χαμηλής ποιότητας, το TopHat χρησιμοποιεί το γονιδίωμα αναφοράς για να καλέσει τη βάση. Επειδή οι περισσότερες αναγνώσεις που καλύπτουν τα άκρα των εξονίων θα εκτείνονται επίσης σε συνδέσεις ματίσματος, τα άκρα των εξονίων στην ψευδοομοφωνία θα καλύπτονται αρχικά από λίγες αναγνώσεις και ως αποτέλεσμα, από την ψευδοομοφωνία ενός εξονίου πιθανότατα θα λείπει μια μικρή ποσότητα ακολουθίας σε κάθε άκρο. Προκειμένου να καταγραφεί αυτή η ακολουθία μαζί με τις τοποθεσίες δότη και δέκτη από πλευρικά εσώνια, το TopHat περιλαμβάνει μια μικρή ποσότητα πλευρικής ακολουθίας από την αναφορά και στις δύο πλευρές κάθε νησίδας (από προεπιλογή 45 bp).

Επειδή τα γονίδια που μεταγράφονται σε χαμηλά επίπεδα θα αλληλουχούνται σε χαμηλή κάλυψη, τα εξόνια σε αυτά τα γονίδια μπορεί να έχουν κενά. Το TopHat έχει μια παράμετρο που ελέγχει πότε δύο διαφορετικά αλλά κοντινά εξόνια πρέπει να συγχωνεύονται σε ένα μόνο εξόνιο. Αυτή η παράμετρος καθορίζει το μήκος του μεγαλύτερου επιτρεπόμενου κενού κάλυψης σε μια μόνο νησίδα. Επειδή τα εσώνια μικρότερα από 70 bp είναι σπάνια σε γονιδιώματα θηλαστικών όπως το ποντίκι, οποιαδήποτε τιμή μικρότερη από 70 bp για αυτήν την παράμετρο είναι λογική. Για να είμαστε συντηρητικοί, η προεπιλογή του TopHat είναι 6 bp.

Για να αντιστοιχίσει τις αναγνώσεις σε συνδέσμους συναρμολόγησης (splice junctions), το TopHat απαριθμεί πρώτα όλες τις κανονικές τοποθεσίες δότη και δέκτη εντός των αλληλουχιών νησίδων (καθώς και τα αντίστροφα

συμπληρώματά τους). Στη συνέχεια, εξετάζει όλα τα ζεύγη αυτών των τοποθεσιών που θα μπορούσαν να σχηματίσουν κανονικά (GT–AG) εσώνια μεταξύ γειτονικών (αλλά όχι απαραίτητα εφαιπτόμενων) νησίδων. Κάθε πιθανό εσώνιο ελέγχεται έναντι των ενδείξεων IUM για αναγνώσεις που εκτείνονται στη διασταύρωση ματίσματος. Από προεπιλογή, το TopHat εξετάζει μόνο πιθανά εσώνια μεγαλύτερα από 70 bp και μικρότερα από 20.000 bp, αλλά αυτά τα προεπιλεγμένα ελάχιστα και μέγιστα μήκη εσωνίων μπορούν να προσαρμοστούν από τον χρήστη. Αυτές οι τιμές περιγράφουν τη συντριπτική πλειοψηφία των γνωστών ευκαρυωτικών εσωνίων. Για παράδειγμα, περισσότερο από το 93% των εσωνίων ποντικού στο γνωστό σύνολο γονιδίων UCSC εμπίπτουν σε αυτό το εύρος. Ωστόσο, οι χρήστες που είναι πρόθυμοι να κάνουν μια μικρή «θυσιά» στην ευαισθησία (sensitivity) θα δουν σημαντικά χαμηλότερο χρόνο λειτουργίας μειώνοντας το μέγιστο μήκος εσωνίου. Για να βελτιωθούν οι χρόνοι εκτέλεσης και να αποφευχθεί η αναφορά ψευδώς θετικών (false positives), το πρόγραμμα αποκλείει ζεύγη δότη-αποδέκτη που εμπίπτουν εξ ολοκλήρου σε μία μόνο νησίδα, εκτός εάν αυτή έχει υψηλή κάλυψη (deeply sequenced).

Προκειμένου να ανιχνεύσει τέτοιες διασταυρώσεις (junctions) χωρίς να θυσιάζει την απόδοση και την ειδίκευση, το TopHat αναζητά εσώνια με υψηλή κάλυψη μέσα σε νησίδες. Κατά τη φάση εξεύρεσης νησίδας, ο αλγόριθμος υπολογίζει τα ακόλουθα στατιστικά στοιχεία για κάθε νησίδα που εκτείνεται στις συντεταγμένες i έως j στον χάρτη:

$$D_{ij} = \frac{\sum_{m=i}^j d_m}{j - i} \cdot \frac{1}{\sum_{m=0}^n d_m}$$

όπου dm είναι το βάθος κάλυψης στη συντεταγμένη m στον χάρτη Bowtie και n είναι το μήκος του γονιδιώματος αναφοράς. Όταν κλιμακώνεται στο εύρος $[0, 1000]$, αυτή η τιμή αντιπροσωπεύει το κανονικοποιημένο βάθος κάλυψης για μια νησίδα. Παρατηρήθηκε ότι οι θέσεις ματίσματος μιας νησίδας τείνουν να εμπίπτουν σε νησίδες με υψηλό D . Έτσι, το TopHat αναζητά διασταυρώσεις που περιέχονται σε νησίδες με $D \geq 300$, αν και αυτή η παράμετρος μπορεί να αλλάξει από τον χρήστη. Μια υψηλή τιμή D θα εμποδίσει το TopHat να αναζητήσει διασταυρώσεις σε μεμονωμένες νησίδες, κάτι που θα βελτιώσει τον χρόνο λειτουργίας. Μια χαμηλή τιμή D θα αναγκάσει το TopHat να κοιτάξει μέσα σε πολλά νησιά, επιβραδύνοντας την ακολουθία εργασιών, αλλά πιθανώς να βρει περισσότερες διασταυρώσεις.

Για κάθε θέση ματίσματος, το TopHat αναζητά τις αναγνώσεις IUM για να βρει αναγνώσεις που εκτείνονται σε διασταυρώσεις χρησιμοποιώντας μια στρατηγική *seed-and-extend*. Το pipeline δημιουργεί ευρετήριο για τις αναγνώσεις IUM χρησιμοποιώντας έναν απλό πίνακα αναζήτησης για να αποσβέσει το κόστος αναζήτησης μιας ματισμένης στοίχισης (spliced alignment) σε πολλές αναγνώσεις. Το TopHat βρίσκει οποιοσδήποτε αναγνώσεις που εκτείνουν τις διασταυρώσεις κατά τουλάχιστον k βάσεις σε κάθε πλευρά (όπου $k = 5$ bp από προεπιλογή), επομένως ο πίνακας πληκτρολογείται με «2k-mers», όπου κάθε «2k-mer» σχετίζεται με αναγνώσεις που περιέχουν αυτό το «2k-mer». Για κάθε ανάγνωση, ο πίνακας περιέχει $(s-2k+1)$ εγγραφές που αντιστοιχούν σε πιθανές θέσεις όπου ένα μάτισμα μπορεί να εμπίπτει σε μια ανάγνωση, όπου s είναι το μήκος της περιοχής υψηλής ποιότητας στο άκρο 5' (από προεπιλογή 28 bp). Οι χρήστες με μεγαλύτερες αναγνώσεις μπορεί να επιθυμούν να αυξήσουν το s για να

βελτιώσουν την ευαισθησία. Η μείωση του s θα βελτιώσει τον χρόνο λειτουργίας, αλλά μπορεί να μειώσει την ευαισθησία. Η αύξηση του k θα βελτιώσει τον χρόνο εκτέλεσης, αλλά μπορεί να περιορίσει το TopHat στην εύρεση συνδέσεων μόνο σε υψηλής έκφρασης (και επομένως βαθιά καλυμμένα) γονίδια. Η μείωσή του θα αυξήσει δραματικά τον χρόνο εκτέλεσης και ενώ η ευαισθησία θα βελτιωθεί, το πρόγραμμα μπορεί να αναφέρει περισσότερα ψευδώς θετικά. Μετά το TopHat παίρνει κάθε πιθανή θέση ματίσματος (splice junction) και δημιουργεί ένα εναρκτήριο (seed) «2k-mer» για αυτό συνενώνοντας τις k βάσεις κάτωθεν του δέκτη με τις k βάσεις άνωθεν του δότη. Ο δείκτης ανάγνωσης IUM τίθεται στη συνέχεια με αυτό το «2k-mer» για να βρεθούν όλες οι αναγνώσεις που περιέχουν το seed. Αυτή η ακριβής αντιστοίχιση «2k-mer» επεκτείνεται για να βρει όλες τις αναγνώσεις που εκτείνονται στη διασταύρωση ματίσματος. Για να επεκτείνει την ακριβή αντιστοίχιση για την περιοχή του seed, το TopHat ευθυγραμμίζει τα τμήματα της ανάγνωσης στα αριστερά και δεξιά του seed με την αριστερή νησίδα και τη δεξιά νησίδα, αντίστοιχα, επιτρέποντας έναν αριθμό αναντιστοιχιών που καθορίζεται από τον χρήστη. Το TopHat θα χάσει ματισμένες ευθυγραμμίσεις σε αναγνώσεις με αναντιστοιχίες στην περιοχή του seed κοντά στη θέση ματίσματος, αλλά αναμένουμε ότι αυτή η αντιστάθμιση μεταξύ ταχύτητας και ευαισθησίας θα είναι ευνοϊκή για τους περισσότερους χρήστες.

Ο αλγόριθμος αναφέρει όλες τις ματισμένες ευθυγραμμίσεις που βρίσκει και, στη συνέχεια, δημιουργεί ένα σύνολο μη περιττών συνδέσεων ματίσματος χρησιμοποιώντας αυτές τις ευθυγραμμίσεις. Ωστόσο, ορισμένες συνδεδεμένες ευθυγραμμίσεις απορρίπτονται πριν από την αναφορά διασταυρώσεων, προκειμένου να αποφευχθεί η αναφορά ψευδών διασταυρώσεων. Για κάθε

διασταύρωση, το μέσο βάθος κάλυψης ανάγνωσης υπολογίζεται για τις αριστερές και δεξιές πλευρικές περιοχές της διασταύρωσης ξεχωριστά. Ο αριθμός των ευθυγραμμίσεων που διασχίζουν τη διασταύρωση διαιρείται με την κάλυψη της πιο βαθιάς καλυμμένης πλευράς για να ληφθεί μια εκτίμηση της δευτερεύουσας ισόμορφης συχνότητας (minor isoform frequency). Εάν το TopHat εκτιμήσει ότι η διασταύρωση ματίσματος εμφανίζεται στο 15% του βάθους κάλυψης των εξονίων που την πλευρίζουν, η διασταύρωση δεν αναφέρεται. Η ελάχιστη δευτερεύουσα παράμετρος συχνότητας ισομορφής προσαρμόζεται από το χρήστη και μπορεί να απενεργοποιηθεί πλήρως.

TopHat2

Το TopHat2 (74) παρέχει σημαντικές βελτιώσεις ακρίβειας σε σχέση με προηγούμενες εκδόσεις και σε σχέση με άλλα εργαλεία χαρτογράφησης RNA-seq. Επειδή το TopHat2 είναι χτισμένο γύρω από το Bowtie2 [1], μπορεί τώρα να ευθυγραμμίσει τις αναγνώσεις σε μικρά indel με υψηλή ακρίβεια, ένα χαρακτηριστικό κρίσιμο για μελέτες που αξιολογούν τις επιδράσεις των γενετικών μεταλλάξεων στην έκφραση γονιδίων και μεταγράφων. Το TopHat2 έχει σχεδιαστεί για να λειτουργεί καλά με ένα ευρύ φάσμα πειραματικών σχεδίων RNA-seq και είναι βελτιστοποιημένο για τις ευρέως διαθέσιμες αναγνώσεις μακράς ζεύξης. Αυτές οι αναγνώσεις θέτουν νέες προκλήσεις επειδή μπορούν να εκτείνονται σε πολλαπλές τοποθεσίες σύνδεσης και όχι μόνο σε μία ή δύο. Υπολογίζεται ότι σχεδόν οι μισές αναγνώσεις μήκους 150 bp θα καλύπτουν δύο ή περισσότερα ανθρώπινα εξόνια. Οι αλγοριθμικές βελτιώσεις στο TopHat2 αντιμετωπίζουν αυτήν την πρόκληση, διατηρώντας

τόσο την ακρίβεια όσο και την ταχύτητα. Άλλες βελτιώσεις στον αλγόριθμο αυξάνουν την ακρίβεια για τις αναγνώσεις που εκτείνονται σε μια διασταύρωση μόνο με μια μικρή προεξοχή (≤ 10 bp), μειώνοντας τα σφάλματα στη συναρμολόγηση μεταγραφής κατάντη χρησιμοποιώντας εργαλεία όπως το Cufflinks. Το TopHat2 κάνει επίσης ισχυρή χρήση των διαθέσιμων γονιδιακών σχολιασμών, που του επιτρέπουν να αποφεύγει τη λανθασμένη αντιστοίχιση αναγνώσεων σε ψευδογονίδια και γενικά να βελτιώνει τη συνολική ακρίβεια ευθυγράμμισης. Ο σχολιασμός επιτρέπει επίσης στο TopHat2 να ευθυγραμμίζει καλύτερα τις αναγνώσεις που καλύπτουν μικροεξόνια, μη κανονικές θέσεις ματίσματος και άλλα «ασυνήθιστα» χαρακτηριστικά ευκαρυωτικών μεταγραφωμάτων. Συμπερασματικά, το TopHat2 αναφέρει πιο ακριβείς ευθυγραμμίσεις από ανταγωνιστικά εργαλεία, χρησιμοποιώντας λιγότερους υπολογιστικούς πόρους.

Παράδειγμα χρήσης του TopHat

Η κεντρική υπόθεση της μελέτης των Sharpnack et al. (75) είναι ότι οι μετρήσεις πρωτεϊνών και mRNA των όγκων αδеноκαρκινώματος του πνεύμονα περιλαμβάνουν ανεξάρτητες πληροφορίες που μπορούν να αξιοποιηθούν για την ανακάλυψη νέων απορρυθμισμένων γονιδίων και ολοκληρωμένων κλινικών βιοδεικτών. Συγκεκριμένα διερευνήθηκε η διαφορική συσχέτιση mRNA-πρωτεΐνης μεταξύ υποτροπιαζόντων και μη υποτροπιζόντων όγκων αδеноκαρκινώματος πνεύμονα. Στη συνέχεια, αξιοποιήθηκε αυτή τη διαφορά, σε συνδυασμό με διαφορικές αφθονίες mRNA και πρωτεϊνών για να προβλεφθεί η υποτροπή του πνευμονικού αδеноκαρκινώματος με ταιριαστά

μεταγραφομικά και πρωτεομικά δεδομένα χρησιμοποιώντας έναν νέο εποπτευόμενο αλγόριθμο ταξινόμησης.

Τα ακατέργαστα αρχεία FASTQ ευθυγραμμίστηκαν με το γονιδίωμα αναφοράς του ανθρώπινου γονιδιώματος 19 (hg19) του Πανεπιστημίου της Καλιφόρνια–Σάντα Κρουζ (UCSC) με το TopHat2 χρησιμοποιώντας προεπιλεγμένες παραμέτρους. Οι μετρήσεις ανάγνωσης (reads) δημιουργήθηκαν με το featureCounts (που περιγράφεται στη συνέχεια) χρησιμοποιώντας ορισμούς γονιδίου RefSeq.

Συνολικά, 5.482 και 6.581 πρωτεϊνικές ομάδες αναγνωρίστηκαν στις κοόρτες Vanderbilt και WashU, αντίστοιχα. Πέντε χιλιάδες διακόσιες ογδόντα τέσσερις και 5.253 από αυτές τις πρωτεΐνες αντιστοιχίστηκαν με σύμβολο γονιδίου με το αντίστοιχο mRNA τους στις κοόρτες Vanderbilt και WashU, αντίστοιχα. Συνολικά μετρήθηκαν 6.577 γονίδια σε τουλάχιστον μία μελέτη και 3.960 γονίδια μετρήθηκαν ποσοτικά και στις δύο μελέτες.

Συνοψίζοντας οι ερευνητές χρησιμοποίησαν το TopHat2, ώστε να συστοιχίσουν δεδομένα μεταγράφων mRNA στο ανθρώπινο γονιδίωμα και τα αποτελέσματά αυτά υποδηλώνουν ότι τα δεδομένα mRNA και πρωτεϊνών έχουν ανεξάρτητη βιολογική και κλινική σημασία, η οποία μπορεί να αξιοποιηθεί για την εύρεση βιοδεικτών έκφρασης υψηλότερης ισχύος.

featureCounts

Το featureCounts (76) είναι ένα εξαιρετικά βελτιστοποιημένο πρόγραμμα καταμέτρησης αναγνώσεων. Το featureCounts μπορεί να χρησιμοποιηθεί για την ποσοτικοποίηση των αναγνώσεων που παράγονται είτε από τεχνολογίες

προσδιορισμού αλληλούχησης RNA είτε DNA από την άποψη οποιουδήποτε τύπου γονιδιωματικού χαρακτηριστικού. Εφαρμόζει κατακερματισμό χρωμοσωμάτων (chromosome hashing), αποκλεισμό χαρακτηριστικών (feature blocking) και άλλες στρατηγικές για την ανάθεση αναγνώσεων σε χαρακτηριστικά με υψηλή απόδοση. Υποστηρίζει multithreading, το οποίο παρέχει περαιτέρω βελτιώσεις ταχύτητας σε μεγάλα προβλήματα δεδομένων.

Εισαγόμενα δεδομένα (Input data)

Η εισαγωγή δεδομένων στο featureCounts αποτελείται από (i) ένα ή περισσότερα αρχεία στοιχισμένων αναγνώσεων είτε σε μορφή Sequence Alignment/Map (SAM) είτε Binary Alignment/Map (BAM) και (ii) μια λίστα γονιδιωματικών χαρακτηριστικών είτε σε μορφή γενικών χαρακτηριστικών (general feature format) (GFF) ή απλοποιημένη μορφή σχολιασμού (simplified annotation format) (SAF). Η μορφή εισόδου ανάγνωσης (SAM ή BAM) ανιχνεύεται αυτόματα και επομένως δεν χρειάζεται να καθοριστεί από το χρήστη. Τόσο η ευθυγράμμιση ανάγνωσης όσο και ο σχολιασμός χαρακτηριστικών θα πρέπει να αντιστοιχούν στο ίδιο γονιδίωμα αναφοράς, το οποίο είναι ένα σύνολο ακολουθιών αναφοράς που αντιπροσωπεύουν χρωμοσώματα ή συναρμολογημένες περιοχές του γονιδιώματος (contigs). Για κάθε ανάγνωση, το αρχείο SAM ή BAM δίνει το όνομα του χρωμοσώματος ή του συνόλου αναφοράς στο οποίο αντιστοιχίστηκε η ανάγνωση, τη θέση έναρξης της ανάγνωσης στο χρωμόσωμα ή το contig και τη λεγόμενη συμβολοσειρά Concise Idiosyncratic Gapped Alignment Report (CIGAR)

δίνοντας τις λεπτομερείς πληροφορίες ευθυγράμμισης, συμπεριλαμβανομένων των εισαγωγών και διαγραφών και ούτω καθεξής σε σχέση με τη θέση έναρξης.

Τα γονιδιωματικά χαρακτηριστικά μπορούν να καθοριστούν είτε σε μορφή GFF είτε σε μορφή SAF. Η μορφή SAF είναι η απλούστερη και περιλαμβάνει μόνο πέντε απαιτούμενες στήλες για κάθε χαρακτηριστικό: αναγνωριστικό χαρακτηριστικού (feature identifier), όνομα χρωμοσώματος, θέση έναρξης, θέση τερματισμού και κλώνο (strand). Αυτές οι πέντε στήλες παρέχουν τις ελάχιστες επαρκείς πληροφορίες για σκοπούς ανάγνωσης ποσοτικού προσδιορισμού. Και στις δύο μορφές, τα αναγνωριστικά χαρακτηριστικών θεωρείται ότι είναι μοναδικά, σύμφωνα με τη βελτιστοποίηση του GFF στο Μορφή Μεταφοράς Γονιδίων (GTF) που χρησιμοποιείται συνήθως.

Ο αριθμός των ακολουθιών αναφοράς μπορεί να είναι μικρός ή μεγάλος ανάλογα με την εφαρμογή. Για καλά εδραιωμένα γονιδιώματα, ο αριθμός των αλληλουχιών αναφοράς είναι ίσος ή κοντά στον αριθμό των χρωμοσωμάτων. Ο αριθμός των αλληλουχιών αναφοράς μπορεί, ωστόσο, να είναι πολύ μεγαλύτερος για γονιδιώματα με ατελείς ή χαμηλής ποιότητας συναρμολογήσεις, επειδή κάθε contig γίνεται μια αλληλουχία αναφοράς. Οι αναγνώσεις RNA-seq μερικές φορές ευθυγραμμίζονται με το μεταγράφημα αντί για το γονιδίωμα. Σε αυτήν την περίπτωση, μπορεί να υπάρχουν εκατοντάδες χιλιάδες μεταγραφές και κάθε μεταγραφή γίνεται μια ακολουθία αναφοράς.

Το featureCounts υποστηρίζει μέτρηση αναγνώσεων για συγκεκριμένο κλώνο, εάν παρέχονται πληροφορίες για τον συγκεκριμένο κλώνο. Τα αποτελέσματα ανάγνωσης χαρτογράφησης συνήθως περιλαμβάνουν

βαθμολογίες ποιότητας χαρτογράφησης για χαρτογραφημένες αναγνώσεις. Οι χρήστες μπορούν προαιρετικά να καθορίσουν μια ελάχιστη βαθμολογία ποιότητας χαρτογράφησης που πρέπει να ικανοποιούν οι εκχωρημένες αναγνώσεις.

Χαρακτηριστικά και μετα-χαρακτηριστικά (features and meta-features)

Κάθε χαρακτηριστικό είναι ένα διάστημα (εύρος θέσεων) σε μία από τις ακολουθίες αναφοράς. Ορίζουμε επίσης ένα μετα-χαρακτηριστικό ως ένα σύνολο χαρακτηριστικών που αντιπροσωπεύουν ένα βιολογικό κατασκεύασμα ενδιαφέροντος. Για παράδειγμα, τα χαρακτηριστικά αντιστοιχούν συχνά σε εξόνια και τα μετα-χαρακτηριστικά σε γονίδια. Τα features που μοιράζονται το ίδιο αναγνωριστικό χαρακτηριστικών στον σχολιασμό GFF ή SAF θεωρείται ότι ανήκουν στο ίδιο μετα-χαρακτηριστικό. Το featureCounts μπορεί να συνοψίσει τις αναγνώσεις είτε σε επίπεδο χαρακτηριστικών είτε σε επίπεδο μετα-χαρακτηριστικών.

Επικάλυψη αναγνώσεων με χαρακτηριστικά

Το featureCounts εκτελεί ακριβή ανάθεση ανάγνωσης συγκρίνοντας τη θέση αντιστοίχισης κάθε βάσης στην ανάγνωση ή το θραύσμα με τη γονιδιωματική περιοχή που εκτείνεται από κάθε χαρακτηριστικό. Λαμβάνει υπόψη τυχόν κενά (εισαγωγές, διαγραφές, διασταυρώσεις εξωνίου-εξωνίου ή συντήξεις) που εντοπίζονται στην ανάγνωση. Καλεί ένα «hit» εάν βρεθεί κάποια επικάλυψη (1 bp ή περισσότερο) μεταξύ της ανάγνωσης ή του τμήματος και ενός χαρακτηριστικού. Ένα hit καλείται για ένα μετα-χαρακτηριστικό, εάν η

ανάγνωση ή το τμήμα επικαλύπτει οποιοδήποτε χαρακτηριστικό στοιχείο του μετα-χαρακτηριστικού.

Πολλαπλές επικαλύψεις

Μια ανάγνωση ή θραύσμα πολλαπλής επικάλυψης είναι αυτό που επικαλύπτει περισσότερα από ένα χαρακτηριστικά ή περισσότερα από ένα μετα-χαρακτηριστικά κατά τη σύνοψη σε επίπεδο μετα-δυνατοτήτων. Το featureCounts παρέχει στους χρήστες την επιλογή είτε να εξαιρέσουν αναγνώσεις πολλαπλής επικάλυψης είτε να τις μετρήσουν για κάθε χαρακτηριστικό που επικαλύπτεται. Η απόφαση εάν θα μετρηθούν αυτές οι αναγνώσεις καθορίζεται συχνά από τον τύπο του πειράματος. Συνίσταται να μην μετρώνται αναγνώσεις ή θραύσματα που επικαλύπτουν περισσότερα από ένα γονίδια για πειράματα αλληλουχίας RNA, επειδή οποιοδήποτε μεμονωμένο θραύσμα πρέπει να προέρχεται από ένα μόνο από τα γονίδια-στόχους, αλλά η ταυτότητα του αληθινού γονιδίου στόχου δεν μπορεί να προσδιοριστεί με βεβαιότητα. Από την άλλη πλευρά, προτείνεται να μετρώνται αναγνώσεις πολλαπλών επικαλύψεων ή θραύσματα για τα περισσότερα πειράματα ChIP-seq, επειδή οι επιγενετικές τροποποιήσεις που συνάγονται από αυτές τις αναγνώσεις ενδέχεται να ρυθμίζουν τις βιολογικές λειτουργίες όλων των επικαλυπτόμενων γονιδίων τους.

Σημειώνεται ότι, κατά την καταμέτρηση σε επίπεδο μετα-χαρακτηριστικών, διαβάζεται ότι τα πολλαπλά επικαλυπτόμενα χαρακτηριστικά του ίδιου μετα-χαρακτηριστικού μετρώνται πάντα ακριβώς μία φορά για αυτό το μετα-χαρακτηριστικό, υπό την προϋπόθεση ότι δεν υπάρχει επικάλυψη με

οποιοδήποτε άλλο μετα-χαρακτηριστικό. Για παράδειγμα, μια ανάγνωση που εκτείνεται σε εξόνιο θα μετρηθεί μόνο μία φορά για το αντίστοιχο γονίδιο, ακόμη κι αν επικαλύπτεται με περισσότερα από ένα εξόνια.

Κατακερματισμός χρωμοσωμάτων (Chromosome hashing)

Το πρώτο βήμα του αλγόριθμου featureCounts είναι η δημιουργία ενός πίνακα κατακερματισμού για τα ονόματα των ακολουθιών αναφοράς. Αυτό επιτρέπει τη γρήγορη αντιστοίχιση των ονομάτων ακολουθιών αναφοράς που βρίσκονται στα αρχεία SAM και στον σχολιασμό GFF. Αυτό είναι ιδιαίτερα χρήσιμο όταν υπάρχει μεγάλος αριθμός αλληλουχιών αναφοράς. Μετά την αντιστοίχιση αναγνώσεων και χαρακτηριστικών κατά ακολουθία αναφοράς, η επακόλουθη ανάλυση μπορεί να προχωρήσει για κάθε ακολουθία αναφοράς ξεχωριστά.

«Genome bins» και «feature blocks»

Μετά τον κατακερματισμό των ονομάτων της ακολουθίας αναφοράς, τα χαρακτηριστικά σε κάθε ακολουθία αναφοράς ταξινομούνται με βάση τις θέσεις έναρξης (αριστερές θέσεις βάσης). Στη συνέχεια δημιουργείται μια ιεραρχία δύο επιπέδων για κάθε ακολουθία αναφοράς. Πρώτον, η ακολουθία αναφοράς χωρίζεται σε μη επικαλυπτόμενα bins («καλάθια») 128 kb και τα χαρακτηριστικά εκχωρούνται σε bins σύμφωνα με τις θέσεις εκκίνησης τους. Μέσα σε κάθε bin, ίσος αριθμός διαδοχικών χαρακτηριστικών ομαδοποιούνται σε blocks. Ο αριθμός των μπλοκ σε ένα bin είναι η τετραγωνική ρίζα του αριθμού των χαρακτηριστικών σε αυτόν το bin (στρογγυλοποιημένος στον

επόμενο ακέραιο αριθμό). Αυτό διασφαλίζει ότι ο αριθμός των χαρακτηριστικών σε ένα μπλοκ είναι σχεδόν ίσος με τον αριθμό των μπλοκ σε ένα bin, μια βέλτιστη ρύθμιση για μια ιεραρχική αναζήτηση.

Η χρήση μιας ιεραρχικής δομής δεδομένων (χαρακτηριστικά μέσα σε μπλοκ μέσα σε bins) είναι ένα βασικό στοιχείο του αλγορίθμου `featureCounts`. Διευκολύνει την ανάθεση ταχείας ανάγνωσης περιορίζοντας γρήγορα τη γονιδιωματική περιοχή που θα μπορούσε να περιέχει χαρακτηριστικά που επικαλύπτονται με το ερώτημα που τέθηκε. Η ανάγνωση του ερωτήματος συγκρίνεται πρώτα με τα πεδία γονιδιωματικής, μετά με τα μπλοκ χαρακτηριστικών εντός οποιωνδήποτε επικαλυπτόμενων δοχείων και μετά με τα χαρακτηριστικά σε τυχόν επικαλυπτόμενα μπλοκ. Αντί να χρησιμοποιεί πολλαπλά επίπεδα δοχείων, ο αλγόριθμος χρησιμοποιεί μόνο ένα επίπεδο bins σε συνδυασμό με τα μπλοκ χαρακτηριστικών. Τέλος, ο αλγόριθμος αποφασίζει πώς να εκχωρήσει την ανάγνωση σύμφωνα με το επίπεδο σύνοψης που εκτελείται (επίπεδο χαρακτηριστικών ή μετα-χαρακτηριστικό επίπεδο) και εάν η ανάγνωση επιτρέπεται να επικαλύπτεται με περισσότερους από έναν στόχους σε αυτό το επίπεδο.

Πλεονεκτήματα του `featureCounts`

Το `featureCounts` παρέχει ένα ευρύ φάσμα επιλογών που επιτρέπει στους χρήστες να ελέγχουν πλήρως τον τρόπο με τον οποίο μπορούν να συνοψιστούν καλύτερα τα δεδομένα ανάγνωσής τους. Οι χρήστες μπορούν να επιλέξουν αν θα μετρήσουν τις αναγνώσεις που επικαλύπτονται με περισσότερες από μία λειτουργίες ή μετα-χαρακτηριστικά. Αυτή η επιλογή

συχνά καθορίζεται από τον τύπο του πειράματος. Οι αναγνώσεις που επικαλύπτονται με περισσότερα από ένα γονίδια (ένα μετα-χαρακτηριστικό) δεν θα πρέπει να υπολογίζονται σε ένα πείραμα RNA-seq, επειδή τέτοιες αναγνώσεις μπορούν να προέρχονται μόνο από ένα γονίδιο, αλλά συνήθως θα πρέπει να μετρώνται σε ένα πείραμα gDNA-seq, όπως ένα πείραμα ChIP-seq ιστόνης. Αυτό το πρόγραμμα επιτρέπει επίσης στους χρήστες να φιλτράρουν τις αναγνώσεις πριν από τη σύνοψη χρησιμοποιώντας μια σειρά από μετρήσεις, όπως βαθμολογίες ποιότητας χαρτογράφησης, αντιστοίχιση θραυσμάτων (ανεξάρτητα από το αν δύο άκρα από το ίδιο τμήμα έχουν χαρτογραφηθεί επιτυχώς ή όχι), το μήκος του θραύσματος, κατεύθυνση (strandness), χιμαιρισμούς και ούτω καθεξής. Μπορεί να ανιχνεύσει αυτόματα την είσοδο ανάγνωσης σε μορφή SAM ή BAM και να ταξινομήσει τις αναγνώσεις με βάση το όνομα εάν οι ζευγαρωμένες αναγνώσεις δεν βρίσκονται σε διαδοχικές θέσεις στην είσοδο. Επιτρέπει επίσης στους χρήστες να προσδιορίσουν εάν αυτές οι αναγνώσεις που αναφέρθηκαν με περισσότερες από μία τοποθεσίες αντιστοίχισης (πολλαπλή αντιστοίχιση) θα πρέπει να καταμετρηθούν ή όχι. Πολλές από αυτές τις χρήσιμες λειτουργίες δεν υποστηρίζονται από άλλα προγράμματα. Λόγω της υψηλής απόδοσης και ακρίβειάς του, το πρόγραμμα featureCounts καθίσταται ως ένα χρήσιμο εργαλείο στην εργαλειοθήκη βιοπληροφορικής για την ανάλυση δεδομένων αλληλουχίας επόμενης γενιάς.

Παράδειγμα χρήσης του featureCounts

Ο στόχος της μελέτης των Djureinovic et al. (77) ήταν να διερευνήσει το ολοκληρωμένο τοπίο των CTAs (Cancer Testis Antigens ή Αντιγόνα Καρκίνου Όρχεων) στο μη μικροκυτταρικό καρκίνο του πνεύμονα (Non-Small Cell Lung Cancer ή NSCLC). Η βάση για αυτήν την ανάλυση ήταν δεδομένα αλληλούχησης RNA (RNAseq) από 199 ασθενείς με NSCLC και 142 δείγματα από 32 διαφορετικούς τύπους φυσιολογικού ανθρώπινου ιστού.

Τα ακατέργαστα δεδομένα αλληλούχησης χαρτογραφήθηκαν στο ανθρώπινο γονιδίωμα αναφοράς (GRCh37) και στον σχολιασμό γονιδίων του Ensembl version 73, χρησιμοποιώντας το TopHat2. Οι μετρήσεις ακατέργαστων αναγνωσμένων υπολογίστηκαν χρησιμοποιώντας το featureCounts από το πακέτο Subread.

Συνοπτικά, η μελέτη αυτή παρέχει προφίλ έκφρασης RNAseq ολόκληρου του τοπίου των CTA στο NSCLC, συμπεριλαμβανομένης μιας ανάλυσης προηγουμένως προτεινόμενων CTA, καθώς και την αναγνώριση νέων CTA. Χάρη στο featureCounts μπόρεσαν να κάνουν περεταίρω ανάλυση του μεταγραφώματος και όρισαν ένα σύνολο 90 αξιόπιστων CTA, συμπεριλαμβανομένων πληροφοριών για την έκφραση πρωτεϊνών, τη μεθυλίωση και τη συσχέτιση επιβίωσης. Ο λεπτομερής κατάλογος RNAseq μπορεί να καθοδηγήσει τις μελέτες βιοδεικτών και τις προσπάθειες εντοπισμού στόχων για ανοσοθεραπευτικές στρατηγικές.

DESeq2

Το DESeq2 (78) είναι μια μέθοδος διαφορικής ανάλυσης για δεδομένα καταμέτρησης (count data), η οποία εκτιμώντας τη συρρίκνωση της διασποράς (shrinkage dispersion) και της μεταβολής (fold change) βελτιώνει τη σταθερότητα και την ερμηνεία των εκτιμήσεων. Αυτό επιτρέπει μια ποσοτική ανάλυση η οποία βασίζεται κυρίως στην ισχύ παρά στην απλή παρουσία της διαφορικής ανάλυσης. Συγκεκριμένα το DESeq2 εντάσσει τα ήδη υπάρχοντα μεθοδολογικά δεδομένα μαζί με αρκετά καινοτόμα χαρακτηριστικά για να διευκολύνει την ποσοτική ανάλυση συγκριτικών αποτελεσμάτων, χρησιμοποιώντας εκτιμητές μείωσης (shrinkage estimators) για τη διασπορά και τη μεταβολή. Επιπλέον, λαμβάνει υπόψη την εγγενή μεταβλητότητα (inherent variability) στα δεδομένα αλληλουχίας και τη διορθώνει, παρέχοντας πιο ακριβή αποτελέσματα. Δημιουργήθηκε ώστε να βελτιώσει το το αρχικό πακέτο DESeq και προσφέρει μια πιο κατανοητή οπτική στη γονιδιακή έκφραση, σε αντίθεση με την τεχνολογία μικροσυστοιχιών (microarray technology).

Το DESeq2 χρησιμοποιεί ένα στατιστικό μοντέλο για να υπολογίσει τη διαφορά στη γονιδιακή έκφραση μεταξύ δύο ή περισσότερων ομάδων δειγμάτων. Αρχικά υπολογίζει τη διασπορά των επιπέδων την γονιδιακής έκφρασης και στη συνέχεια προσαρμόζει αρνητική διωνυμική κατανομή σε κάθε γονίδιο. Η αρνητική διωνυμική κατανομή λαμβάνει υπόψη την υπερβολική διασπορά των δεδομένων αλληλουχίας, η οποία μπορεί να οδηγήσει σε πιο ακριβείς εκτιμήσεις του περιθώριου σφάλματος (p-value) και του ψευδούς ρυθμού ανακάλυψης (False Discovery Rate). Μόλις το στατιστικό μοντέλο είναι κατάλληλο, το DESeq2 το χρησιμοποιεί για να υπολογίσει το περιθώριο

σφάλματος για κάθε γονίδιο, αντιπροσωπεύοντας την πιθανότητα ότι η παρατηρούμενη διαφορά στην έκφραση οφείλεται σε τυχαία πιθανότητα. Στη συνέχεια εφαρμόζεται ένα κατώφλι για το περιθώριο σφάλματος (p -value threshold) για τον εντοπισμό διαφορικά εκφραζόμενων γονιδίων, με ένα μικρότερο περιθώριο σφάλματος που υποδεικνύει ισχυρότερα στοιχεία για διαφορική έκφραση.

Συνήθη βήματα διαφορικής ανάλυσης με τη χρήση του DESeq2

Η χρήση του DESeq2 σε μια ανάλυση αλληλούχισης RNA (RNA-seq) τυπικά περιλαμβάνει τα ακόλουθα βήματα:

- 1) *Προεπεξεργασία ακατέργαστων δεδομένων RNA-Seq*: Αυτό το βήμα περιλαμβάνει τον καθαρισμό των ακατέργαστων δεδομένων αλληλουχίας, την αφαίρεση των αναγνώσεων χαμηλής ποιότητας και τη χαρτογράφηση των αναγνώσεων σε ένα γονιδίωμα αναφοράς.
- 2) *Ποιοτικός έλεγχος*: Αυτό το βήμα περιλαμβάνει τον έλεγχο της ποιότητας των επεξεργασμένων δεδομένων αλληλουχίας και το φιλτράρισμα τυχόν δειγμάτων που δεν πληρούν ορισμένα κριτήρια, όπως το χαμηλό βάθος αλληλουχίας (low sequencing depth).
- 3) *Δημιουργία καταμετρήσεων (Count Generation)*: Σε αυτό το βήμα, οι χαρτογραφημένες αναγνώσεις μετρώνται για κάθε γονίδιο, παρέχοντας ένα μέτρο των επιπέδων έκφρασης για κάθε γονίδιο σε κάθε δείγμα.
- 4) *Στατιστική ανάλυση*: Αυτό είναι το κύριο βήμα όπου χρησιμοποιείται το DESeq2. Το DESeq2 λαμβάνει τα ακατέργαστα δεδομένα μέτρησης (raw count

data) και προσαρμόζει ένα στατιστικό μοντέλο σε αυτό. Το DESeq2 βασίζεται σε βιολογικά αντίγραφα (biological replicates) για τον υπολογισμό της διακύμανσης, επομένως, απαιτούνται βιολογικά αντίγραφα σε κάθε συνθήκη. Αυτό στη συνέχεια χρησιμοποιείται για να παρέχει εκτιμήσεις της διαφορικής έκφρασης μεταξύ δύο ή περισσότερων ομάδων δειγμάτων.

5) *Οπτικοποίηση δεδομένων*: Αυτό το βήμα περιλαμβάνει την οπτικοποίηση των αποτελεσμάτων της στατιστικής ανάλυσης, χρησιμοποιώντας συνήθως εργαλεία όπως γραφικές παραστάσεις, heatmaps ή volcano plots.

6) *Ανάλυση βιολογικών μονοπατιών (Pathway Analysis)*: Αυτό το βήμα περιλαμβάνει τη χρήση των αποτελεσμάτων της ανάλυσης διαφορικής έκφρασης για τον εντοπισμό μονοπατιών ή βιολογικών διεργασιών που εμπλουτίζονται για διαφορεικά εκφραζόμενα γονίδια. Η ανάλυση εμπλουτισμού συνόλου γονιδίων (GSEA) ή η ανάλυση υπερ-αναπαράστασης είναι μέθοδοι που χρησιμοποιούνται συνήθως για την εκτέλεση αυτού του τύπου ανάλυσης. Βάσεις δεδομένων όπως το MSigDB χρησιμοποιούνται συχνά σε συνδυασμό με πακέτα R.

7) *Ερμηνεία*: Τέλος, τα αποτελέσματα της ανάλυσης πρέπει να ερμηνεύονται στο πλαίσιο του πειράματος και της βιολογίας που μελετάται. Αυτό θα μπορούσε να περιλαμβάνει πρόσθετη ανάλυση δεδομένων ή βιβλιογραφική έρευνα για την κατανόηση των βιολογικών επιπτώσεων των αποτελεσμάτων.

Αυτά τα βήματα παρέχουν μια γενική επισκόπηση μιας τυπικής ακολουθίας εργασιών (pipeline) κατά τη χρήση του DESeq2. Σημειώνεται ότι τα βήματα 1-3 εκτελούνται πριν από τη χρήση του DESeq2.

Εισαγόμενα δεδομένα (Input Data)

Ως εισαγόμενα δεδομένα (input), το πακέτο DESeq2 αναμένει δεδομένα καταμέτρησης όπως λαμβάνονται, π.χ. από RNA-seq ή άλλο πείραμα προσδιορισμού αλληλουχίας υψηλής απόδοσης, με τη μορφή πίνακα (matrix) ακεραίων τιμών. Η τιμή στην *i*-η σειρά και στην *j*-η στήλη του πίνακα αναφέρει πόσες αναγνώσεις μπορούν να αντιστοιχιστούν στο γονίδιο *i* στο δείγμα *j*. Ανάλογα, για άλλους τύπους δοκιμασιών, οι σειρές του πίνακα μπορεί να αντιστοιχούν π.χ. σε περιοχές σύνδεσης (με ChIP-Seq) ή πεπτιδικές αλληλουχίες (με ποσοτική φασματομετρία μάζας). Οι τιμές στον πίνακα θα πρέπει να είναι μη κανονικοποιημένες μετρήσεις ή εκτιμώμενες μετρήσεις αναγνώσεων αλληλουχίας (για αλληλουχία RNA μονού άκρου) (single-end RNA-seq) ή θραύσματα (για αλληλουχία RNA ζευγαρωμένου άκρου) (paired-end RNA-seq). Είναι σημαντικό να παρέχονται πίνακες καταμέτρησης ως εισαγόμενα δεδομένα για να διατηρείται το στατιστικό μοντέλο του DESeq2, καθώς μόνο οι τιμές μέτρησης επιτρέπουν τη σωστή αξιολόγηση της ακρίβειας της μέτρησης. Το μοντέλο του DESeq2 διορθώνει εσωτερικά για το μέγεθος της βιβλιοθήκης, επομένως οι μετασχηματισμένες ή κανονικοποιημένες τιμές όπως οι μετρήσεις που κλιμακώνονται βάσει του μεγέθους της βιβλιοθήκης δεν θα πρέπει να χρησιμοποιούνται ως εισαγόμενα δεδομένα.

Η κλάση αντικειμένου που χρησιμοποιείται από το πακέτο DESeq2 για την αποθήκευση των μετρήσεων ανάγνωσης και των ενδιάμεσων εκτιμώμενων ποσοτήτων κατά τη στατιστική ανάλυση είναι το `DESeqDataSet`, το οποίο συνήθως αναπαρίσταται στον κώδικα εδώ ως αντικείμενο `dds`. Ένα `DESeqDataSet` πρέπει να έχει συσχετισμένο τύπο σχεδίασης. Ο τύπος σχεδιασμού εκφράζει τις μεταβλητές που θα χρησιμοποιηθούν στη

μοντελοποίηση. Ο τύπος θα πρέπει να είναι μια περισπωμένη (~) ακολουθούμενη από τις μεταβλητές με τα σύμβολα συν μεταξύ τους (θα εξαναγκαστεί σε έναν τύπο αν δεν είναι ήδη). Ο σχεδιασμός μπορεί να αλλάξει αργότερα, ωστόσο στη συνέχεια θα πρέπει να επαναληφθούν όλα τα βήματα διαφορικής ανάλυσης, καθώς ο τύπος σχεδιασμού χρησιμοποιείται για την εκτίμηση των διασπορών και για την εκτίμηση των log2 μεταβολής (log2 fold change) του μοντέλου (σημειώνεται πως για να χρησιμοποιηθούν οι προεπιλεγμένες ρυθμίσεις του πακέτου, θα πρέπει η μεταβλητή ενδιαφέροντος να τοποθετηθεί στο τέλος του τύπου και το επίπεδο ελέγχου να είναι το πρώτο επίπεδο). Υπάρχουν 4 τρόποι κατασκευής ενός συνόλου δεδομένων για το DESeq2 (DESeqDataSet), ανάλογα με την ακολουθία εργασιών (pipeline) που χρησιμοποιήθηκε πριν από το DESeq2 για δημιουργημένες μετρήσεις ή εκτιμώμενες μετρήσεις, οι οποίες είναι από μεταγραφικά αρχεία αφθονίας (transcript abundance files) και tximport, από έναν πίνακα μέτρησης (count matrix), από αρχεία μέτρησης htseq (htseq-count files) και από αρχεία SummarizedExperiment (SummarizedExperiment object), εκ των οποίων ο πιο συνήθης είναι η δημιουργία πίνακα μέτρησης (count matrix).

Η συνάρτηση DESeqDataSetFromMatrix μπορεί να χρησιμοποιηθεί εάν έχει ήδη προετοιμαστεί ένας πίνακας μετρήσεων ανάγνωσης από άλλη πηγή. Για να χρησιμοποιηθεί το DESeqDataSetFromMatrix, ο χρήστης θα πρέπει να παρέχει τον πίνακα καταμέτρησης, τις πληροφορίες σχετικά με τα δείγματα (τις στήλες του πίνακα μέτρησης) ως *DataFrame* ή *data.frame* και τον τύπο σχεδίασης. Είναι απολύτως κρίσιμο οι στήλες του πίνακα καταμέτρησης και οι σειρές των δεδομένων στηλών (πληροφορίες για δείγματα) να είναι με την ίδια σειρά. Το DESeq2 δεν θα κάνει εικασίες για το ποια στήλη του πίνακα

καταμέτρησης ανήκει σε ποια σειρά των δεδομένων της στήλης, αυτά πρέπει να παρέχονται στο DESeq2 ήδη με τη σωστή σειρά. Σε περίπτωση που δεν έχουν τη σωστή σειρά, πρέπει να τακτοποιηθούν εκ νέου το ένα ή το άλλο έτσι ώστε να είναι συνεπή όσον αφορά τη σειρά δειγμάτων (εάν δεν γίνει, οι μεταγενέστερες συναρτήσεις θα προκαλέσουν σφάλμα). Ο πίνακας μέτρησης και τα δεδομένα στηλών μπορούν συνήθως να διαβαστούν σε R από επίπεδα αρχεία (flat files) χρησιμοποιώντας βασικές συναρτήσεις της R όπως *read.csv* ή *read.delim*. Εάν υπάρχουν πρόσθετα δεδομένα χαρακτηριστικών, μπορούν να προστεθούν στο DESeqDataSet προσθέτοντας στις στήλες μεταδεδομένων ενός αντικειμένου που κατασκευάστηκε πρόσφατα.

Προεργασία (Pre-filtering)

Αν και δεν είναι απαραίτητο να προεπεξεργαστούν τα γονίδια χαμηλού αριθμού αναγνώσεων πριν να εκτελεστούν οι λειτουργίες του DESeq2, υπάρχουν δύο λόγοι που κάνουν την προεργασία χρήσιμη: αφαιρώντας σειρές στις οποίες υπάρχουν πολύ λίγες αναγνώσεις, μειώνουμε το μέγεθος της μνήμης του αντικειμένου δεδομένων (count object) *dds* και αυξάνουμε την ταχύτητα της μοντελοποίησης της μέτρησης στο DESeq2. Μπορεί επίσης να βελτιώσει τις απεικονίσεις, καθώς τα χαρακτηριστικά χωρίς πληροφορίες για διαφορική έκφραση δεν σχεδιάζονται σε διαγράμματα διασποράς ή γραφικά MA (MA plots). Το πλήθος του 10 είναι μια λογική επιλογή για μαζικό RNA-seq. Μια σύσταση για τον ελάχιστο αριθμό δειγμάτων είναι να καθοριστεί το μικρότερο μέγεθος ομάδας. Εάν δεν υπάρχουν διακριτές ομάδες, μπορεί κανείς

να χρησιμοποιήσει τον ελάχιστο αριθμό δειγμάτων όπου οι μη μηδενικές μετρήσεις θα θεωρούνταν ενδιαφέρουσες.

Ανάλυση διαφορικής έκφρασης (Differential Expression Analysis)

Τα τυπικά βήματα ανάλυσης διαφορικής έκφρασης συμπεριλαμβάνονται σε μια ενιαία συνάρτηση, την *DESeq*. Οι πίνακες αποτελεσμάτων δημιουργούνται χρησιμοποιώντας τα αποτελέσματα της συνάρτησης, τα οποία εξάγουν έναν πίνακα αποτελεσμάτων με το $\log_2$ μεταβολής ($\log_2$ fold change), τις τιμές του περιθωρίου σφάλματος (p-value) και τις προσαρμοσμένες τιμές του περιθωρίου σφάλματος (p-adjusted). Χωρίς πρόσθετα επιχειρήματα (arguments) στην εντολή *results*, το $\log_2$ μεταβολής ($\log_2$ fold change) και η τιμή του περιθωρίου σφάλματος (p-value) στη διαδικασία Wald (Wald test) θα είναι για την τελευταία μεταβλητή (last variable) στον τύπο σχεδίασης και εάν αυτός είναι ένας παράγοντας, η σύγκριση θα είναι το τελευταίο επίπεδο (last level) αυτής της μεταβλητής σε σχέση με το επίπεδο αναφοράς (reference level). Ωστόσο, η σειρά των μεταβλητών του σχεδίου δεν έχει σημασία, εφόσον ο χρήστης καθορίζει τη σύγκριση για τη δημιουργία ενός πίνακα αποτελεσμάτων, χρησιμοποιώντας τα ορίσματα *name* ή *contrast* των αποτελεσμάτων. Λεπτομέρειες σχετικά με τη σύγκριση εκτυπώνονται στην κονσόλα (console), ακριβώς πάνω από τον πίνακα αποτελεσμάτων.

Το μοντέλο του DESeq2

Η ανάλυση διαφορικής έκφρασης στο DESeq2 χρησιμοποιεί ένα γενικευμένο γραμμικό μοντέλο της μορφής:

$$K_{ij} \sim NB(\mu_{ij}, \alpha_i)$$

$$\mu_{ij} = s_j q_{ij}$$

$$\log_2(q_{ij}) = x_j \beta_i$$

όπου οι καταμετρήσεις K_{ij} για το γονίδιο i , το δείγμα j μοντελοποιούνται χρησιμοποιώντας αρνητική διωνυμική κατανομή με προσαρμοσμένο μέσο (fitted mean) μ_{ij} και μια ειδική για γονίδιο παράμετρο διασποράς α_i . Ο προσαρμοσμένος μέσος όρος αποτελείται από έναν συντελεστή μεγέθους συγκεκριμένου δείγματος s_j και μια παράμετρο q_{ij} ανάλογη με την αναμενόμενη πραγματική συγκέντρωση θραυσμάτων για το δείγμα j . Οι συντελεστές β_i δίνουν τις στο $\log_2$ μεταβολές ($\log_2$ fold change) για το γονίδιο i για κάθε στήλη του μοντέλου μήτρας X . Σημειώνεται ότι το μοντέλο μπορεί να γενικευτεί για τη χρήση παραγόντων κανονικοποίησης που εξαρτώνται από το δείγμα και τα γονίδια s_{ij} . Η παράμετρος διασποράς α_i ορίζει τη σχέση μεταξύ της διακύμανσης του παρατηρούμενου αριθμού και της μέσης τιμής του. Με άλλα λόγια, πόσο περιμέναμε ότι θα είναι η παρατηρούμενη μέτρηση από τη μέση τιμή, η οποία εξαρτάται τόσο από τον παράγοντα μεγέθους (size factor) s_j και το εξαρτώμενο από τη συνμεταβλητή μέρος (covariate-dependent part) q_{ij} όπως ορίζεται παραπάνω.

$$Var(K_{ij}) = E[(K_{ij} - \mu_{ij})^2] = \mu_{ij} + \alpha_i \mu_{ij}^2$$

Μια επιλογή στο DESeq2 είναι να παρέχει μέγιστες εκ των υστέρων εκτιμήσεις των $\log_2$ μεταβολών ($\log_2$ fold change) στο β_i μετά την ενσωμάτωση ενός μηδενικού κέντρου Normal prior (betaPrior). Ενώ προηγουμένως, αυτές οι μετριάσμενες ή συρρικνωμένες εκτιμήσεις δημιουργούνταν από το DESeq ή

από τη συνάρτηση *nbinomWaldTest*, τώρα παράγονται από τη συνάρτηση *lfcShrink*. Οι διασπορές εκτιμώνται χρησιμοποιώντας αναμενόμενες μέσες τιμές από την εκτίμηση μέγιστης πιθανότητας των log2 μεταβολών (log2 fold change) και βελτιστοποιώντας την πιθανότητα προσαρμοσμένου προφίλ Cox-Reid. Τα βήματα που εκτελούνται από τη συνάρτηση *DESeq* είναι τα εξής: εκτίμηση συντελεστών μεγέθους sj από την εντολή *estimateSizeFactors*, εκτίμηση της διασποράς α_i με την εντολή *estimateDispersions* και τέλος αρνητική διωνυμική προσαρμογή GLM για β_i και στατιστικά Wald από το *nbinomWaldTest*.

Πλεονεκτήματα του DESeq2 ως εργαλείο διαφορικής ανάλυσης

Υπάρχουν διάφοροι λόγοι για τους οποίους το DESeq2 είναι μια δημοφιλής επιλογή για ανάλυση RNA-Seq. Αρχικά, παρέχει ακριβή διόρθωση της μεταβλητότητας καθώς το DESeq2 χρησιμοποιεί ένα στατιστικό μοντέλο για τη διόρθωση της μεταβλητότητας στα δεδομένα αλληλουχίας, οδηγώντας σε πιο ακριβή αποτελέσματα. Επιπροσθέτως, μπορεί να χειριστεί πολύπλοκα πειραματικά σχέδια, όπως πολλαπλές ομάδες με πολλαπλά αντίγραφα. Τέλος, το DESeq2 είναι ένα ευρέως χρησιμοποιούμενο εργαλείο στην κοινότητα της βιοπληροφορικής, με μεγάλη βάση χρηστών και καλά τεκμηριωμένο εγχειρίδιο χρήστη.

Το DESeq2, όπως και το limma (79), δημιουργήθηκε με βάση τεχνολογίες μεταγραφομικής, αλλά έχει πρόσφατα βρει εφαρμογή στη διαφορική ανάλυση του πρωτεώματος (80,81). Μάλιστα, το στατιστικό μοντέλο του DESeq2 αποδίδει με μεγαλύτερη ακρίβεια και ευαισθησία για label-free

proteomics σε σχέση με άλλα εργαλεία διαφορικής ανάλυσης που χρησιμοποιούνται στην πρωτεομική (82,83).

Παράδειγμα χρήσης του DESeq2

Δεδομένης της σημασίας της επιγενετικής τροποποίησης και της ανοσίας στην ογκογένεση και ανάπτυξη, οι Ren et al. (84) προσπάθησαν να συνδυάσουν γονίδια που σχετίζονται με την παράτυπη μεθυλίωση (aberrant methylation) και τη διήθηση του όγκου από CD8 T- λεφοκύτταρα. Έτσι δημιουργούν ένα προγνωστικό μοντέλο και να διερευνούν τους βασικούς βιοδείκτες του πρώιμου σταδίου LUAD.

Τα διαφορικά εκφρασμένα mRNA μεταξύ καρκίνου και φυσιολογικών δειγμάτων ταυτοποιήθηκαν με το πακέτο DEseq2 σε γλώσσα R. Γονίδια με $|FC| \geq 1,5$ και προσαρμοσμένο $P \leq 0,05$ αναγνωρίστηκαν ως διαφορικά εκφρασμένα. Προσδιορίστηκαν 4501 γονίδια με αυξημένη έκφραση και 2688 με μειωμένη έκφραση σε καρκινικούς ιστούς σε σύγκριση με φυσιολογικούς ιστούς.

Με βάση τα αποτελέσματα του DESeq2, δημιουργήθηκε ένα μοντέλο πρόβλεψης χρησιμοποιώντας 17 γονίδια οδηγούς μεθυλίωσης και τέσσερα γονίδια που σχετίζονται με τα κύτταρα CD8. Ένα νομογράφημα που συνδυάζει τη βαθμολογία κινδύνου και κλινικοπαθολογικούς παράγοντες μπορεί να προβλέψει την πιθανότητα επιβίωσης. Η μελέτη τους παρέχει ένα ισχυρό μοντέλο και υποψήφιους βιοδείκτες για εξατομικευμένη θεραπεία ασθενών με LUAD πρώιμου σταδίου.

Geneset Enrichment Analysis (GSEA)

Η Ανάλυση Εμπλουτισμού Συνόλου Γονιδίων (Geneset Enrichment Analysis ή GSEA) (85,86) είναι μια αναλυτική μέθοδος για την ερμηνεία δεδομένων γονιδιακής έκφρασης. Η συγκεκριμένη μέθοδος εστιάζει σε σύνολα γονιδίων, δηλαδή σε ομάδες γονιδίων που έχουν κοινή βιολογική λειτουργία, χρωμοσωμική θέση ή ρύθμιση. Συγκεκριμένα, όπου η ανάλυση ενός γονιδίου βρίσκει μικρή ομοιότητα μεταξύ δύο ανεξάρτητων μελετών, η GSEA αποκαλύπτει πολλά κοινά βιολογικά μονοπάτια. Η GSEA αξιολογεί δεδομένα μικροσυστοιχιών σε επίπεδο συνόλων γονιδίων. Τα σύνολα γονιδίων ορίζονται με βάση προηγούμενες βιολογικές γνώσεις, π.χ. δημοσιευμένες πληροφορίες σχετικά με βιοχημικές οδούς ή συνέκφραση σε προηγούμενα πειράματα. Ο στόχος του GSEA είναι να προσδιορίσει εάν τα μέλη ενός συνόλου γονιδίων *S* τείνουν να εμφανίζονται προς την κορυφή (ή το κάτω μέρος) της λίστας *L*, οπότε το σύνολο γονιδίων συσχετίζεται με τη διάκριση φαινοτυπικής κατηγορίας.

Επισκόπηση της μεθόδου

Η GSEA εξετάζει πειράματα με προφίλ έκφρασης σε ολόκληρο το γονιδίωμα από δείγματα που ανήκουν σε δύο κατηγορίες, με την επισήμανση 1 ή 2. Τα γονίδια ταξινομούνται με βάση τη συσχέτιση μεταξύ της έκφρασής τους και της διάκρισης κατηγορίας χρησιμοποιώντας οποιαδήποτε κατάλληλη μέτρηση. Με δεδομένο ένα εκ των προτέρων καθορισμένο σύνολο γονιδίων *S* (π.χ. γονίδια που κωδικοποιούν προϊόντα σε μια μεταβολική οδό, που βρίσκονται στην ίδια κυτταρογενετική ζώνη ή μοιράζονται την ίδια κατηγορία GO), ο στόχος του GSEA είναι να προσδιορίσει εάν τα γονίδια του συνόλου *S*

κατανέμονται τυχαία σε όλο τη λίστα L ή κυρίως στο επάνω ή στο κάτω μέρος. Αναμένουμε ότι σύνολα που σχετίζονται με τη φαινοτυπική διάκριση θα τείνουν να δείχνουν την τελευταία κατανομή.

Υπάρχουν τρία βασικά στοιχεία της μεθόδου GSEA:

1) *Υπολογισμός Βαθμού Εμπλουτισμού (Enrichment Score)*: Υπολογίζουμε ένα βαθμό εμπλουτισμού (ES) που αντικατοπτρίζει τον βαθμό στον οποίο ένα σύνολο S υπερεκπροσωπείται στα άκρα (πάνω ή κάτω) ολόκληρης της λίστας κατάταξης L. Ο βαθμός υπολογίζεται περπατώντας στη λίστα L, αυξάνοντας ένα στατιστικό αθροίσματος όταν συναντάμε ένα γονίδιο στο S και μειώνοντάς το όταν συναντάμε γονίδια όχι στο S. Το μέγεθος της αύξησης (increment) εξαρτάται από τη συσχέτιση του γονιδίου με τον φαινότυπο. Ο βαθμός εμπλουτισμού (Enrichment Score) είναι η μέγιστη απόκλιση από το μηδέν που συναντάται στον τυχαίο walk: αντιστοιχεί σε μια σταθμισμένη στατιστική που μοιάζει με Kolmogorov–Smirnov.

2) *Εκτίμηση Επιπέδου Σημαντικότητας Βαθμού Εμπλουτισμού (Significance Level of Enrichment Score)*: Υπολογίζουμε τη στατιστική σημασία (ονομαστική τιμή περιθωρίου σφάλματος) του βαθμού εμπλουτισμού χρησιμοποιώντας μια εμπειρική διαδικασία δοκιμής μετάθεσης βασισμένη σε φαινότυπο που διατηρεί τη σύνθετη δομή συσχέτισης των δεδομένων έκφρασης γονιδίου. Συγκεκριμένα, μεταθέτουμε τις ετικέτες φαινοτύπου (phenotype labels) και υπολογίζουμε εκ νέου το βαθμό εμπλουτισμού (enrichment score) του συνόλου γονιδίων (gene set) για τα μετατεθέντα δεδομένα, το οποίο δημιουργεί μια μηδενική κατανομή για το βαθμό εμπλουτισμού (enrichment score). Η εμπειρική, ονομαστική τιμή περιθωρίου σφάλματος (p-value) του

παρατηρούμενου βαθμού εμπλουτισμού στη συνέχεια υπολογίζεται σε σχέση με αυτήν τη μηδενική κατανομή. Είναι σημαντικό ότι η μεταλλαγή των ετικετών κλάσης (class labels) διατηρεί τις συσχετίσεις γονιδίου-γονιδίου και, ως εκ τούτου, παρέχει μια βιολογικά λογική εκτίμηση της σημαντικότητας από ό,τι θα μπορούσε να επιτευχθεί με τη μεταλλαγή γονιδίων.

3) *Προσαρμογή για τον έλεγχο πολλαπλών υποθέσεων (Multiple Hypothesis Testing)*: Όταν αξιολογείται μια ολόκληρη βάση δεδομένων συνόλων γονιδίων, προσαρμόζουμε το εκτιμώμενο επίπεδο σημαντικότητας για να ληφθεί υπόψη ο έλεγχος πολλαπλών υποθέσεων. Αρχικά κανονικοποιούμε το βαθμό εμπλουτισμού (enrichment score) για κάθε σύνολο γονιδίων ώστε να λαμβάνεται υπόψη το μέγεθος του συνόλου, δίνοντας ένα κανονικοποιημένο βαθμό εμπλουτισμού (normalized enrichment score). Στη συνέχεια ελέγχουμε την αναλογία των ψευδώς θετικών υπολογίζοντας το ποσοστό ψευδούς ανακάλυψης (false discovery rate) που αντιστοιχεί σε κάθε κανονικοποιημένο βαθμό εμπλουτισμού. Το ποσοστό ψευδούς εύρεσης (false discovery rate) είναι η εκτιμώμενη πιθανότητα ότι ένα σύνολο με δεδομένο κανονικοποιημένο βαθμό εμπλουτισμού αντιπροσωπεύει ένα ψευδώς θετικό εύρημα. Υπολογίζεται συγκρίνοντας τις ουρές (tails) των παρατηρούμενων και μηδενικών κατανομών για τον κανονικοποιημένο βαθμό εμπλουτισμού.

Το υποσύνολο αιχμής

Τα σύνολα γονιδίων μπορούν να οριστούν χρησιμοποιώντας μια ποικιλία μεθόδων, αλλά δεν θα συμμετέχουν όλα τα μέλη ενός συνόλου γονιδίων σε μια βιολογική διαδικασία. Συχνά είναι χρήσιμο να εξαχθούν τα

πυρηνικά μέλη των συνόλων γονιδίων υψηλής βαθμολογίας που συμβάλλουν στο βαθμό εμπλουτισμού. Ορίζουμε ως υποσύνολο αιχμής εκείνα τα γονίδια στο σύνολο γονιδίων S που εμφανίζονται στη λίστα κατάταξης L στο ή πριν από το σημείο όπου το τρέχον άθροισμα φτάνει τη μέγιστη απόκλιση από το μηδέν. Το υποσύνολο αιχμής μπορεί να ερμηνευτεί ως ο πυρήνας ενός συνόλου γονιδίων που αντιπροσωπεύει το σήμα εμπλουτισμού. Η εξέταση του υποσυνόλου αιχμής μπορεί να αποκαλύψει ένα βιολογικά σημαντικό υποσύνολο μέσα σε ένα σύνολο γονιδίων. Αυτή η προσέγγιση είναι ιδιαίτερα χρήσιμη με manually επιμελημένα σύνολα γονιδίων, τα οποία μπορεί να αντιπροσωπεύουν μια συγχώνευση διαδικασιών αλληλεπίδρασης.

Τα σύνολα γονιδίων υψηλής βαθμολογίας μπορούν να ομαδοποιηθούν με βάση τα κορυφαία υποσύνολα γονιδίων που μοιράζονται. Τέτοιες ομαδοποιήσεις μπορούν να αποκαλύψουν ποια από αυτά τα σύνολα γονιδίων αντιστοιχούν στις ίδιες βιολογικές διεργασίες και ποια αντιπροσωπεύουν ξεχωριστές διαδικασίες.

Παραλλαγές της μεθόδου GSEA

Όπως προαναφέρθηκε, η GSEA χρησιμοποιείται για την ανάλυση μιας ταξινομημένης λίστας γονιδίων που αντικατοπτρίζει τη διαφορική έκφραση μεταξύ δύο τάξεων, καθεμία από τις οποίες αντιπροσωπεύεται από μεγάλο αριθμό δειγμάτων. Τα γονίδια μπορούν να ταξινομηθούν με βάση τις διαφορές που παρατηρούνται σε ένα μικρό σύνολο δεδομένων, με πολύ λίγα δείγματα για να επιτραπεί η αυστηρή αξιολόγηση των επιπέδων σημαντικότητας με τη μετάθεση των ετικετών κλάσης. Σε αυτές τις περιπτώσεις, μια τιμή περιθωρίου

σφάλματος μπορεί να εκτιμηθεί μεταθέτοντας τα γονίδια, με αποτέλεσμα τα γονίδια να εκχωρούνται τυχαία στα σύνολα διατηρώντας το μέγεθός τους. Αυτή η προσέγγιση δεν είναι αυστηρά ακριβής: επειδή αγνοεί τις συσχετίσεις γονιδίων-γονιδίων, θα υπερεκτιμήσει τα επίπεδα σημαντικότητας και μπορεί να οδηγήσει σε ψευδώς θετικά αποτελέσματα. Ωστόσο, μπορεί να είναι χρήσιμο για τη δημιουργία υποθέσεων.

Τα γονίδια μπορούν επίσης να ταξινομηθούν με βάση το πόσο καλά η έκφρασή τους συσχετίζεται με ένα δεδομένο πρότυπο στόχο (όπως το πρότυπο έκφρασης ενός συγκεκριμένου γονιδίου). Και πάλι, οι κατά προσέγγιση τιμές περιθωρίου σφάλματος μπορούν να εκτιμηθούν με μετάθεση γονιδίων.

Η GSEA αξιολογεί ένα σύνολο δεδομένων μικροσυστοιχίας ερωτημάτων χρησιμοποιώντας μια συλλογή συνόλων γονιδίων. Ως εκ τούτου, δημιουργήθηκε ένας αρχικός κατάλογος 1.325 συνόλων γονιδίων, που ονομάζεται MSigDB 1.0 που αποτελείται από τέσσερις τύπους συνόλων.

Κυτταρογενετικά σύνολα (C1, 319 σύνολα γονιδίων): Αυτός ο κατάλογος περιλαμβάνει 24 σύνολα, ένα για καθένα από τα 24 ανθρώπινα χρωμοσώματα και 295 σύνολα που αντιστοιχούν σε κυτταρογενετικές ζώνες. Αυτά τα σύνολα είναι χρήσιμα για τον εντοπισμό επιδράσεων που σχετίζονται με χρωμοσωμικές διαγραφές (chromosomal deletions) ή ενισχύσεις (chromosomal amplifications), αντιστάθμιση δόσης (dosage compensation), επιγενετική σίγαση (epigenetic silencing) και άλλες τοπικές επιδράσεις.

Λειτουργικά σύνολα (C2, 522 σύνολα γονιδίων): Αυτός ο κατάλογος περιλαμβάνει 472 σύνολα που περιέχουν γονίδια των οποίων τα προϊόντα

εμπλέκονται σε συγκεκριμένα μεταβολικά και σηματοδοτικά μονοπάτια, όπως αναφέρονται σε οκτώ δημόσια διαθέσιμες, χειροκίνητα επιμελημένες βάσεις δεδομένων, και 50 σετ που περιέχουν γονίδια που συσχετίζονται ως απόκριση σε γενετικές και χημικές διαταραχές, όπως αναφέρεται σε διάφορα πειραματικά άρθρα.

Σύνολα ρυθμιστικών μοτίβων (C3, 57 σύνολα γονιδίων): Αυτός ο κατάλογος αναφέρει 57 συνήθως διατηρημένα ρυθμιστικά μοτίβα στις περιοχές προαγωγέα των ανθρώπινων γονιδίων και καθιστά δυνατή τη σύνδεση των αλλαγών σε ένα πείραμα μικροσυστοιχίας με ένα διατηρημένο, υποτιθέμενο ρυθμιστικό στοιχείο cis (cis-regulatory element).

Γειτονικά σύνολα (C4, 427 σύνολα γονιδίων): Αυτά τα σύνολα ορίζονται από γειτονιές έκφρασης (expression neighborhoods) που επικεντρώνονται σε γονίδια που σχετίζονται με τον καρκίνο.

Λογισμικό GSEA-P και σύνολα γονιδίων MsigDB

Για τη διευκόλυνση της χρήσης του GSEA, έχουν αναπτυχθεί πόροι που διατίθενται δωρεάν από το Broad Institute κατόπιν αιτήματος. Αυτοί οι πόροι περιλαμβάνουν το λογισμικό GSEA-P, το MSigDB 1.0 και τη συνοδευτική τεκμηρίωση. Το λογισμικό είναι διαθέσιμο ως (i) μια εφαρμογή επιτραπέζιου υπολογιστή ανεξάρτητη από πλατφόρμα με γραφική διεπαφή χρήστη (ii) προγράμματα σε R και σε java που οι προχωρημένοι χρήστες μπορούν να ενσωματώσουν στις δικές τους αναλύσεις ή περιβάλλοντα λογισμικού (iii) μια αναλυτική ενότητα στο πακέτο ανάλυσης μικροσυστοιχίας γενετικών προτύπων (iv) ένας μελλοντικός διακομιστής GSEA που θα βασίζεται στον ιστό που θα

επιτρέπει στους χρήστες να εκτελούν τη δική τους ανάλυση απευθείας στον ιστότοπο.

Πλεονεκτήματα της μεθόδου GSEA

Οι παραδοσιακές στρατηγικές για την ανάλυση γονιδιακής έκφρασης έχουν επικεντρωθεί στον εντοπισμό μεμονωμένων γονιδίων που παρουσιάζουν διαφορές μεταξύ δύο καταστάσεων ενδιαφέροντος. Αν και είναι χρήσιμα, αποτυγχάνουν να ανιχνεύσουν βιολογικές διεργασίες, όπως μεταβολικά μονοπάτια, μεταγραφικά προγράμματα και αποκρίσεις στρες, που κατανέμονται σε ένα ολόκληρο δίκτυο γονιδίων στο επίπεδο μεμονωμένων γονιδίων.

Η GSEA είχε αρχικά αναπτυχθεί για την ανάλυση τέτοιων δεδομένων σε επίπεδο συνόλων γονιδίων. Η μέθοδος χρησιμοποιήθηκε για την ανακάλυψη μεταλλαγμένων μεταβολικών οδών στον ανθρώπινο διαβήτη και στη συνέχεια εφαρμόστηκε για την ανακάλυψη διεργασιών που εμπλέκονται στο διάχυτο λέμφωμα μεγάλων Β κυττάρων, οδούς ανίχνευσης θρεπτικών ουσιών που εμπλέκονται στον καρκίνο του προστάτη και στη σύγκριση των προφίλ έκφρασης του ποντικίου σε αυτά των ανθρώπων. Πλέον, η αρχική προσέγγιση έχει βελτιωθεί σε μια ευαίσθητη, ισχυρή αναλυτική μέθοδο και εργαλείο με πολύ ευρύτερη εφαρμογή μαζί με μια μεγάλη βάση δεδομένων συνόλων γονιδίων. Η GSEA μπορεί σαφώς να εφαρμοστεί και σε άλλα σύνολα δεδομένων, όπως πρωτεομικά δεδομένα ορού, πληροφορίες γονότυπου ή προφίλ μεταβολικών.

Η GSEA διαθέτει μια σειρά από πλεονεκτήματα σε σύγκριση με τις μεθόδους ενός γονιδίου. Πρώτον, διευκολύνει την ερμηνεία ενός πειράματος μεγάλης κλίμακας εντοπίζοντας μονοπάτια και διαδικασίες. Αντί να επικεντρωθούν σε γονίδια υψηλής βαθμολογίας (high scoring genes), τα οποία μπορεί να σχολιαστούν ελάχιστα και μπορεί να μην είναι αναπαραγώγιμα, οι ερευνητές μπορούν να επικεντρωθούν σε σύνολα γονιδίων, τα οποία τείνουν να είναι πιο αναπαραγώγιμα και πιο ερμηνεύσιμα. Δεύτερον, όταν τα μέλη ενός συνόλου γονιδίων παρουσιάζουν ισχυρή διασταυρούμενη συσχέτιση (cross-correlation), η GSEA μπορεί να ενισχύσει την αναλογία σήματος προς θόρυβο (signal-to-noise ratio) και να καταστήσει δυνατό τον εντοπισμό μέτρων αλλαγών σε μεμονωμένα γονίδια. Τρίτον, η ανάλυση αιχμής (leading-edge analysis) μπορεί να βοηθήσει στον καθορισμό υποσυνόλων γονιδίων για την αποσαφήνιση των αποτελεσμάτων.

Πολλά άλλα εργαλεία αναπτύχθηκαν πρόσφατα για την ανάλυση της γονιδιακής έκφρασης χρησιμοποιώντας πληροφορίες μονοπατιού (pathway) ή οντολογίας (ontology). Τα περισσότερα καθορίζουν εάν μια ομάδα διαφορικά εκφραζόμενων γονιδίων εμπλουτίζεται για ένα μονοπάτι ή έναν όρο οντολογίας χρησιμοποιώντας στατιστικές επικάλυψης όπως η αθροιστική υπεργεωμετρική κατανομή (cumulative hypergeometric distribution). Η GSEA διαφέρει σε δύο σημαντικά σημεία. Πρώτον, η GSEA εξετάζει όλα τα γονίδια σε ένα πείραμα, όχι μόνο αυτά που βρίσκονται πάνω από ένα αυθαίρετο διαχωρισμό (arbitrary cutoff) όσον αφορά τη μεταβολή (fold change) ή τη σημασία. Δεύτερον, η GSEA αξιολογεί τη σημασία μεταθέτοντας τις ετικέτες κλάσης, οι οποίες διατηρούν τις συσχετίσεις γονιδίων-γονιδίων και, ως εκ τούτου, παρέχει ένα πιο ακριβές μηδενικό μοντέλο (null model).

Η πραγματική δύναμη της GSEA, ωστόσο, έγκειται στην ευελιξία της. Ταυτόχρονα με αυτή, έχει δημιουργηθεί μια αρχική βάση δεδομένων μοριακών υπογραφών (molecular signature database) που αποτελείται από 1.325 σύνολα γονιδίων, συμπεριλαμβανομένων αυτών που βασίζονται σε βιολογικά μονοπάτια, χρωμοσωμική θέση, μοτίβα ανάντη cis (upstream cis motifs), αποκρίσεις σε μια φαρμακευτική θεραπεία ή προφίλ έκφρασης σε σύνολα δεδομένων μικροσυστοιχιών που δημιουργήθηκαν προηγουμένως. Μπορούν να δημιουργηθούν περαιτέρω σύνολα μέσω γενετικών και χημικών διαταραχών (perturbation), υπολογιστικής ανάλυσης γονιδιωματικών πληροφοριών και πρόσθετου βιολογικού σχολιασμού. Επιπλέον, η ίδια η GSEA θα μπορούσε να χρησιμοποιηθεί για να βελτιώσει χειροκίνητα επιμελημένα μονοπάτια και σύνολα, προσδιορίζοντας τα σύνολα αιχμής (leading-edge sets) που μοιράζονται σε διάφορα πειραματικά σύνολα δεδομένων. Καθώς προστίθενται τέτοια σύνολα, εργαλεία όπως το GSEA θα βοηθήσουν στη σύνδεση της προηγούμενης γνώσης με τα πρόσφατα δημιουργούμενα δεδομένα και ως εκ τούτου θα βοηθήσουν στην αποκάλυψη της συλλογικής συμπεριφοράς των γονιδίων σε καταστάσεις υγείας και ασθένειας.

Παράδειγμα χρήσης του GSEA

Οι Zhang et al. (87) διεξήγαγαν multi-omic μελέτη τόσο σε προεπεμβατικά όσο και σε επεμβατικά στάδια του αδενοκαρκινώματος του πνεύμονα. Οι έρευνά τους διευκρινίζει τη μοριακή αρχιτεκτονική που συνδέεται με την εξέλιξη του αδενοκαρκινώματος του πνεύμονα και αποκαλύπτουν τους πρωτεομικούς υποτύπους.

Πραγματοποιήθηκε GSEA χρησιμοποιώντας τα σύνολα γονιδίων συμπεριλαμβανομένων των KEGG, Reactome και HALLMARK από τη βάση δεδομένων Molecular Signatures Data Base (MSigDB) ως βάση δεδομένων αναφοράς (reference database). Ως αποτέλεσμα, βρέθηκε ότι οι φυσιολογικοί ιστοί των πνευμόνων συσχετίστηκαν κυρίως με την κυτταρική προσκόλληση, τη μεταφορά οξυγόνου και τη μεταβολική διαδικασία ρετινοειδών, η οποία θα μπορούσε να εξηγηθεί εν μέρει από τις φυσιολογικές λειτουργίες των πνευμόνων, όπως η πνευμονική ανταλλαγή αερίων.

Συνοπτικά, η μελέτη τους απεικονίζει τον περιεκτικό γενομικό, πρωτεομικό και φωσφοπρωτεομικό χάρτη από τη νεοπλασία του πνεύμονα έως το IAC (invasive adenocarcinoma ή διηθητικό αδеноκαρκίνωμα). Σε αυτά τα πλαίσια, χάρη στο GSEA σκιαγραφήθηκε η αρχιτεκτονική και τα χαρακτηριστικά του αδеноκαρκινώματος του πνεύμονα σε πρώιμο στάδιο. Τα δεδομένα από αυτό τη μελέτη χρησιμεύουν ως πολύτιμη πηγή για την επιστημονική κοινότητα για τη διερεύνηση της ογκογένεσης και της εξέλιξης του αδеноκαρκινώματος του πνεύμονα, οδηγώντας προς μια βαθύτερη κατανόηση αυτής της θανατηφόρας νόσου.

DAVID (Database for Annotation, Visualization and Integrated Discovery)

Το DAVID bioinformatics resources (88) αποτελείται από μια ολοκληρωμένη βάση βιολογικών γνώσεων και αναλυτικά εργαλεία που στοχεύουν στη συστηματική εξαγωγή βιολογικού νοήματος από μεγάλες λίστες γονιδίων ή πρωτεϊνών. Συγκεκριμένα είναι ένα υψηλής απόδοσης και

ολοκληρωμένο περιβάλλον εξόρυξης δεδομένων (data-mining environment), για την ανάλυση λιστών γονιδίων που προέρχονται από γονιδιωματικά πειράματα υψηλής απόδοσης. Πρώτα η διαδικασία απαιτεί τη μεταφόρτωση μιας λίστας γονιδίων που περιέχει οποιονδήποτε αριθμό κοινών αναγνωριστικών γονιδίων ακολουθούμενη από ανάλυση που χρησιμοποιεί ένα ή περισσότερα κείμενα και εργαλεία εξόρυξης μονοπατιών όπως γονιδιακή λειτουργική ταξινόμηση, γράφημα ή ομαδοποίηση λειτουργικών σχολιασμών και πίνακα functional annotation. Χρησιμοποιώντας το DAVID, οι ερευνητές είναι σε θέση να αποκτήσουν μια εις βάθος κατανόηση των βιολογικών θεμάτων σε λίστες γονιδίων που εμπλουτίζονται σε μελέτες κλίμακας γονιδιώματος.

Το DAVID είναι σε θέση να εξάγει βιολογικά χαρακτηριστικά και νόημα που σχετίζονται με μεγάλες λίστες γονιδίων, ενώ επιπλέον μπορεί να χειρίζεται οποιοδήποτε τύπο λίστας γονιδίων, ανεξάρτητα από το ποια πλατφόρμα γονιδιώματος ή πακέτο λογισμικού που τα δημιούργησε. Σε σύγκριση με άλλες παρόμοια εργαλεία, το DAVID παρέχει ορισμένα μοναδικά χαρακτηριστικά και δυνατότητες, όπως μία ολοκληρωμένη και διευρυμένη βάση δεδομένων σχολιασμού, προηγμένοι modular αλγόριθμοι εμπλουτισμού και ισχυρή εξερευνητική ικανότητα σε ένα ολοκληρωμένο περιβάλλον εξόρυξης δεδομένων.

Αρχή του «υποβάθρου του γονιδιακού πληθυσμού» στην ανάλυση εμπλουτισμού

Το βασικό θεμέλιο της ανάλυσης εμπλουτισμού είναι ότι εάν μια βιολογική διαδικασία είναι ανώμαλη (abnormal) σε μια δεδομένη μελέτη, τα

συνεπικουρικά λειτουργικά γονίδια (co-functioning genes) θα πρέπει να έχουν υψηλότερη πιθανότητα, δηλαδή να είναι εμπλουτισμένα (enriched), για να επιλεγούν ως μια σχετική ομάδα από τεχνολογίες ελέγχου υψηλής απόδοσης. Για να αποφασιστεί ο βαθμός εμπλουτισμού, πρέπει να τεθεί ένα συγκεκριμένο υπόβαθρο ώστε να πραγματοποιηθεί η σύγκριση. Ο εμπλουτισμός μπορεί να μετρηθεί ποσοτικά με κάποιες κοινές και γνωστές στατιστικές μεθόδους, συμπεριλαμβανομένης αυτής του χ^2 , του τεστ ακριβείας Fisher (Fisher's exact test), της διωνυμικής πιθανότητας (binomial probability) και της υπεργεωμετρικής κατανομής (hypergeometric distribution). Ωστόσο, δε δύναται να βγει ασφαλές συμπέρασμα χωρίς να υπάρξει σύγκριση με το βασικό υπόβαθρο.

Ένα από τα πλεονεκτήματα του DAVID είναι η ευελιξία στη ρύθμιση διαφορετικών πληθυσμιακών υπόβαθρων για την αντιμετώπιση διαφορετικών καταστάσεων. Το DAVID έχει μια αυτοματοποιημένη διαδικασία για να προβλέπει το υπόβαθρο ως το παγκόσμιο σύνολο γονιδίων στο γονιδίωμα με βάση την ανεβασμένη λίστα γονιδίων από τον χρήστη. Έτσι, σε μια συνήθη κατάσταση, οι χρήστες δεν είναι απαραίτητο να δημιουργήσουν από μόνοι τους ένα πληθυσμιακό υπόβαθρο. Η παραπάνω από ικανοποιητική λειτουργία του φάνηκε από το γεγονός ότι οι περισσότερες από τις μελέτες που αναλύθηκαν από το DAVID είναι μελέτες σε όλο το γονιδίωμα (genome-wide studies) ή σχεδόν σε όλο το γονιδίωμα (close to genome-wide studies). Επιπλέον, άλλες επιλογές είναι επίσης διαθέσιμες για τις προτιμήσεις του χρήστη, συμπεριλαμβανομένων όλων των γονιδίων στο γονιδίωμα που μελετήθηκε, γονίδια σε διάφορα τσιπ μικροσυστοιχιών (microarray chips) και, πιο σημαντικό, οποιοδήποτε σύνολο γονιδίων που ορίζουν και ανεβάζουν οι

χρήστες. Το τελευταίο χαρακτηριστικό απαιτεί ιδιαίτερα μεγάλη υπολογιστική ισχύ, έτσι ώστε σπάνια βρίσκεται σε παρόμοιες διαδικτυακές εφαρμογές. Συνοψίζοντας, διάφορες ρυθμίσεις και επιλογές για υπόβαθρα πληθυσμού μπορούν να καλύπτουν το φάσμα των αναγκών των γενικών έως εκείνων των απαιτητικών χρηστών.

DAVID gene name batch viewer

Οι ταυτότητες των γονιδίων (Gene IDs), όπως το Entrez Gene 3558, συνήθως δεν μεταφέρουν βιολογικό νόημα από μόνες τους. Το gene name batch viewer είναι σε θέση να προσδώσει γρήγορα νόημα σε μια λίστα ταυτοτήτων γονιδίων μεταφράζοντάς τα στα αντίστοιχα γονιδιακά τους ονόματα. Έτσι, πριν προχωρήσουμε σε ανάλυση με άλλα πιο ολοκληρωμένα αναλυτικά εργαλεία, οι ερευνητές μπορούν να ρίξουν γρήγορα μια ματιά στα ονόματα των γονιδίων για να αποκτήσουν περαιτέρω πληροφορίες σχετικά με τη μελέτη τους και να απαντήσουν σε ερωτήσεις όπως αν περιέχει μια λίστα γονιδίων σημαντικά γονίδια που σχετίζονται με την εκάστοτε μελέτη. Επιπλέον, παρέχεται ένα σύνολο υπερσυνδέσμων για κάθε καταχώρηση γονιδίου, επιτρέποντας στους χρήστες να εξερευνήσουν περαιτέρω πρόσθετες λειτουργικές πληροφορίες για κάθε γονίδιο.

Γονιδιακή λειτουργική ταξινόμηση από το DAVID (DAVID gene functional classification)

Καθώς η ανάλυση προχωρά, η γονιδιακή λειτουργική ταξινόμηση (gene functional classification) παρέχει τη διακριτή ικανότητα των ερευνητών να

εξερευνούν και να βλέπουν λειτουργικά σχετιζόμενα γονίδια μαζί, ως μονάδα, για να επικεντρωθούν στο μεγαλύτερο βιολογικό δίκτυο και όχι στο επίπεδο ενός μεμονωμένου γονιδίου. Στην πραγματικότητα, η πλειονότητα των συνλειτουργικών γονιδίων μπορεί να έχει διαφοροποιημένα ονόματα, έτσι ώστε τα γονίδια να μην μπορούν απλά να ταξινομηθούν σε λειτουργικές ομάδες σύμφωνα με το όνομά τους. Ωστόσο, η λειτουργική ταξινόμηση γονιδίων, που ολοκληρώνεται με ένα σύνολο νέων τεχνικών ασαφούς ομαδοποίησης, είναι σε θέση να ταξινομήσει τα γονίδια εισόδου σε λειτουργικά σχετιζόμενες γονιδιακές ομάδες (ή τάξεις) με βάση τον όρο σχολιασμού (annotation) συν-εμφάνιση (co-occurrence) και όχι με ονόματα γονιδίων. Η συμπύκνωση μεγάλων λιστών γονιδίων σε βιολογικά σημαντικές ενότητες βελτιώνει σημαντικά την ικανότητα του ατόμου να αφομοιώνει μεγάλες ποσότητες πληροφοριών και έτσι αλλάζει τη λειτουργική ανάλυση σχολιασμού από μια γενετική ανάλυση σε μια ανάλυση βιολογικής ενότητας (biological module-centric module). Λαμβάνοντας υπόψη τη λειτουργία «drill-down» που σχετίζεται με κάθε βιολογική ενότητα και τις οπτικοποιήσεις για την προβολή των σχέσεων μεταξύ των συσχετίσεων πολλών γονιδίων σε πολλούς όρους, οι ερευνητές είναι σε θέση να κατανοήσουν πληρέστερα πώς συνδέονται τα γονίδια μεταξύ τους και με το λειτουργικό σχολιασμό.

Λειτουργικό γράφημα σχολιασμού DAVID (DAVID functional annotation chart)

Το γράφημα λειτουργικών σχολιασμών παρέχει τυπική ανάλυση εμπλουτισμού γονιδιακού όρου (υπερεκπροσωπούμενη), η οποία παρέχεται επίσης από άλλα παρόμοια εργαλεία, για τον εντοπισμό των πιο σχετικών

(υπερεκπροσωπούμενων) βιολογικών όρων που σχετίζονται με μια δεδομένη λίστα γονιδίων. Σε σύγκριση με άλλα παρόμοια εργαλεία ανάλυσης εμπλουτισμού, η αξιοσημείωτη διαφορά αυτής της λειτουργίας που παρέχεται από το DAVID είναι η εκτεταμένη κάλυψη σχολιασμών (annotations), η οποία αυξάνεται από μόνο GO στην αρχική έκδοση του DAVID σε πάνω από 40 κατηγορίες σχολιασμού επί του παρόντος, συμπεριλαμβανομένων όρων GO, αλληλεπιδράσεων πρωτεΐνης-πρωτεΐνης, λειτουργικές περιοχές πρωτεϊνών, συσχετίσεις ασθενειών, βιο-μονοπάτια, χαρακτηριστικά αλληλουχίας, ομολογία, περιλήψεις λειτουργικών γονιδίων, έκφραση γονιδιακού ιστού και βιβλιογραφία. Οι κατηγορίες σχολιασμών μπορούν να συμπεριληφθούν ή να εξαιρεθούν ευέλικτα από την ανάλυση με βάση τις επιλογές του χρήστη. Η βελτιωμένη κάλυψη σχολιασμού από μόνη της αυξάνει την αναλυτική ισχύ επιτρέποντας στους ερευνητές να αναλύσουν τα γονίδια τους από πολλές διαφορετικές βιολογικές πτυχές σε έναν ενιαίο χώρο. Επιπλέον, για να αξιοποιηθούν πλήρως τα γνωστά μονοπάτια KEGG και BioCarta, το πρόγραμμα προβολής μονοπατιών DAVID μπορεί να εμφανίσει γονίδια από τη λίστα ενός χρήστη σε χάρτες μονοπατιών για να διευκολύνει τη βιολογική ερμηνεία σε ένα πλαίσιο δικτύου. Τέλος, η επιλογή προκατασκευασμένων ή καθορισμένων από τον χρήστη υποβάθρων γονιδιακού πληθυσμού παρέχει στον χρήστη τη δυνατότητα να προσαρμόσει την ανάλυση εμπλουτισμού ώστε να ανταποκρίνεται στη συγκεκριμένη αναλυτική κατάσταση του χρήστη.

Ομαδοποίηση λειτουργικών σχολιασμών από το DAVID (DAVID functional annotation clustering)

Η ομαδοποίηση λειτουργικών σχολιασμών (functional annotation clustering) χρησιμοποιεί μια παρόμοια έννοια ασαφούς ομαδοποίησης ως

λειτουργική ταξινόμηση μετρώντας τις σχέσεις μεταξύ των όρων σχολιασμού (annotations) με βάση τον βαθμό συσχέτισής τους με γονίδια στη λίστα του χρήστη για να ομαδοποιήσει κάπως ετερογενή, αλλά πολύ παρόμοιο σχολιασμό σε ομάδες λειτουργικών σχολιασμών (functional annotation groups). Αυτό μειώνει το βάρος της συσχέτισης διαφορετικών όρων που σχετίζονται με την παρόμοια βιολογική διαδικασία, επιτρέποντας έτσι τη βιολογική ερμηνεία να εστιάζεται περισσότερο στο επίπεδο της «βιολογικής ενότητας». Το εργαλείο δισδιάστατης προβολής (2D view tool) παρέχεται επίσης για την εξέταση των εσωτερικών σχέσεων μεταξύ των ομαδοποιημένων όρων και γονιδίων. Αυτός ο τύπος ομαδοποίησης λειτουργικών σχολιασμών μπορεί να δώσει μια πιο διορατική άποψη των σχέσεων μεταξύ κατηγοριών και όρων σχολιασμού σε σύγκριση με την παραδοσιακή γραμμική λίστα εμπλουτισμένων όρων, καθώς οι πολύ σχετικοί ή περιττοί όροι σχολιασμού μπορεί να διασκορπίζονται σε εκατοντάδες, αν όχι σε χιλιάδες, άλλων όρων.

Λειτουργικός πίνακας σχολιασμών DAVID (DAVID functional annotation table)

Ο πίνακας λειτουργικών σχολιασμών (functional annotation table) είναι μια μηχανή ερωτημάτων (query engine) για τη βάση δεδομένων DAVID, χωρίς στατιστικούς υπολογισμούς. Για μια δεδομένη λίστα γονιδίων, το εργαλείο μπορεί να αναζητήσει γρήγορα τον αντίστοιχο σχολιασμό για κάθε γονίδιο και να τους παρουσιάσει σε μορφή πίνακα. Έτσι, οι χρήστες μπορούν να εξερευνήσουν τον σχολιασμό με τρόπο γονίδιο προς γονίδιο. Αυτή είναι μια χρήσιμη αναλυτική ενότητα, ιδιαίτερα όταν οι χρήστες θέλουν να δουν προσεκτικά τον σχολιασμό των εξαιρετικά ενδιαφέροντων γονιδίων.

Εκσυγχρονισμός του DAVID

Σε πρόσφατη ενημέρωση (89), δημιουργήθηκε ξανά η βάση δεδομένων DAVID, ώστε να επεκταθεί η κάλυψη σχολιασμού και ταξινόμησης. Συγκεκριμένα, το DAVID ενημερώθηκε για όλους τους τύπους σχολιασμών με βάση το νέο σύστημα γονιδίων, εάν υπήρχαν διαθέσιμες ενημερώσεις της βάσης δεδομένων πηγής. Από τους 78 ενημερωμένους τύπους σχολιασμών, οι περισσότεροι έχουν αυξήσει σημαντικά τον αριθμό των εγγραφών γονιδιακών όρων με μόνο πέντε τύπους σχολιασμών να μειώνονται στον αριθμό των εγγραφών γονιδιακών όρων, συμπεριλαμβανομένων των τύπων σχολιασμού MINT, DIP και CGNC με τις μεγαλύτερες μειώσεις. Η ανάλυση δείχνει ότι ορισμένα από τις ταυτότητες γονιδίων στο DAVID v6.8 για αυτούς τους τρεις τύπους σχολιασμού έχουν καταστεί άκυρες λόγω επιμέλειας, με αποτέλεσμα τη μείωση του αριθμού των εγγραφών. Στην πραγματικότητα, το CGNC έχει 49,45% αύξηση των έγκυρων αναγνωριστικών στην ενημερωμένη βάση σε σύγκριση με το DAVID v6.8.

Επιπλέον, προστέθηκαν νέοι τύποι σχολιασμού, συμπεριλαμβανομένων των αλληλεπιδράσεων μικρού μορίου-γονιδίου (PubChem) και των αλληλεπιδράσεων φαρμάκου-γονιδίου (DrugBank), επιτρέποντας στον χρήστη να εντοπίσει μικρά μόρια και φάρμακα που αλληλεπιδρούν με τα γονίδια στη μελέτη του και θα μπορούσαν να διευκολύνουν την ανακάλυψη φαρμάκων μέσω της βιοπληροφορικής ανάλυσης.

Επιπροσθέτως, έχει βελτιστοποιηθεί η σελίδα που αφορά τα αποτελέσματα σύνοψης σχολιασμών (annotation summary results): προσθέτοντας τους νέους τύπους σχολιασμού και αφαιρώντας περιττούς ή

χαμηλής αξίας. Είναι σημαντικό ότι έχει χωριστεί ο τύπος σχολιασμού λέξης-κλειδιού Uniprot (Uniprot Key-word) στο "Λειτουργικοί σχολιασμοί" σε πιο συγκεκριμένους υποτύπους για να επιτραπεί μεγαλύτερη εξειδίκευση στις αναλύσεις λειτουργικού εμπλουτισμού. Η σελίδα αναφοράς γονιδίου έχει εμπλουτιστεί στην ενημέρωση με περισσότερες διαθέσιμες πληροφορίες σχετικά με γονίδια στο DAVID, συμπεριλαμβανομένων των νέων δεδομένων αλληλεπίδρασης μικρού μορίου-γονιδίου και φαρμάκου-γονιδίου. Τέλος, υπάρχουν νέοι μηχανισμοί για τη διευκόλυνση της αποτελεσματικής ενημέρωσης του νέου συστήματος γονιδίων στο DAVID και του συνδεδεμένου σχολιασμού στη βάση.

Συμπερασματικά, αυτή η ενημέρωση επεκτείνει σημαντικά τη Γνωσιακή Βάση και ενισχύει τη δύναμη ανακάλυψης του DAVID, ενώ ελπίζουμε ότι βελτιώνει την εμπειρία του χρήστη. Η προσθήκη μικρών αλληλεπιδράσεων μορίου-γονιδίου και φαρμάκου-γονιδίου στο DAVID ανοίγει την πόρτα για πρόσθετες ανακαλύψεις και οι νέοι μηχανισμοί ενημέρωσης που αναπτύχθηκαν επιτρέπουν στο DAVID να παραμείνει ενημερωμένο και να επεκταθεί στο μέλλον.

Πλεονεκτήματα του DAVID

Συλλογικά, όλες οι αναλυτικές ενότητες του DAVID στοχεύουν στην εξαγωγή βιολογικού νοήματος από τη δεδομένη λίστα γονιδίων από διαφορετικές βιολογικές γωνίες με εξαιρετικά συνεπή και αναμενόμενα αποτελέσματα για μια δεδομένη μελέτη. Η ενσωμάτωση των αποτελεσμάτων από τις διαφορετικές αναλυτικές ενότητες θα επωφεληθεί από τη διαφορετική

εστίαση και δύναμη κάθε ενότητας, για να κάνει τη συνολική βιολογική εικόνα που συναρμολογείται με βάση τη λίστα γονιδίων, πιο ολοκληρωμένη και λεπτομερή.

Για μια δεδομένη λίστα γονιδίων, το DAVID bioinformatics resources είναι σε θέση να βοηθήσει τους χρήστες:

- 1) να μετατρέψουν τις ταυτότητες των γονιδίων (gene IDs) από τον έναν τύπο στον άλλον.
- 2) να διαγνώσουν και να διορθώσουν προβλήματα σχετικά με τις ταυτότητες των γονιδίων.
- 3) να εξερευνήσουν τα ονόματα των γονιδίων κατά ομάδες.
- 4) να ανακαλύψουν λειτουργικά εμπλουτισμένες ομάδες γονιδίων (functionally enriched gene groups).
- 5) να εμφανίσουν τη σχέση πολλών γονιδίων με πολλούς όρους σε δισδιάστατη προβολή.
- 6) να έχουν μια πρώτη ματιά στις κύριες βιολογικές λειτουργίες που σχετίζονται με τη λίστα γονιδίων.
- 7) να εντοπίσει εμπλουτισμένους (enriched) όρους σχολιασμού (annotation terms).
- 8) να οπτικοποιήσουν τα γονίδια σε χάρτες μονοπατιών (pathway maps) KEGG και BioCarta.
- 9) να κάνουν σύνδεση συσχετισμού μεταξύ γονιδίων και ασθενειών.
- 10) να επισημάνουν τις λειτουργικές περιοχές και τα μοτίβα των πρωτεϊνών.
- 11) να ανακατευθυνθούν σε σχετική βιβλιογραφία.
- 12) να απαριθμήσουν πρωτεΐνες που αλληλεπιδρούν.

- 13) να ομαδοποιούν περιπτώσεις και ετερόζυγους όρους σχολιασμού (annotations).
- 14) να αναζητήσουν άλλα παρόμοια λειτουργικά γονίδια στο γονιδίωμα, τα οποία δε βρίσκονται στη λίστα που έχουν υποβάλει.
- 15) να αναζητήσουν άλλους λειτουργικά παρόμοιους σχολιασμούς με αυτούς που τους ενδιαφέρουν.
- 16) να διαβάσουν όλα τα περιεχόμενα σχετικά με τον σχολιασμό που αφορούν ένα γονίδιο.

Συνολικά, το DAVID παρέχει μια τακτική αναλυτική ροή για να ξεκινήσουν οι νέοι χρήστες, καθώς και την ευελιξία για έμπειρους χρήστες να χρησιμοποιούν, τις ενότητες σε διαφορετικούς συνδυασμούς για να εξισορροπούν τις διαφορετικές εστίες και τα δυνατά σημεία κάθε ενότητας για να ανταποκρίνονται καλύτερα σε συγκεκριμένες αναλυτικές ερωτήσεις.

Παράδειγμα χρήσης του DAVID

Η μελέτη των Wu et al. (90) διερευνά τις διαφορές μεταξύ NSCLC και SCLC με ολοκληρωμένες αναλύσεις πρωτεομικής και μεταγραφομικής. Με τη σύγκριση δεδομένων πρωτεομικής και γενομικής, εντοπίστηκαν αρκετά υποψήφια γονίδια/πρωτεΐνες με πιθανή βιολογική σημασία, εκ των οποίων συζητήθηκαν περαιτέρω οι πιθανές λειτουργίες στον καρκίνο του πνεύμονα.

Η GO ανάλυση πραγματοποιήθηκε για τα σημαντικά διαφορικά εκφραζόμενα γονίδια και πρωτεΐνες προκειμένου να βρεθούν οι βιολογικές διαδικασίες, τα κυτταρικά συστατικά και οι μοριακές λειτουργίες. Η ανάλυση εμπλουτισμού GO πραγματοποιήθηκε από το DAVID ακολουθούμενη από το

πακέτο γλώσσας R ggplot2. Όσον αφορά τις βιολογικές διεργασίες, οι upregulated πρωτεΐνες εμπλέκονταν κυρίως στη διαδικασία ματίσματος mRNA, και εξαγωγής RNA από τον πυρήνα. Οι downregulated πρωτεΐνες σε κυτταρικές σειρές SCLC συμμετείχαν στις βιολογικές διαδικασίες της συσσώρευσης αιμοπεταλίων και στην προσκόλληση κυτάρου-κυτάρου.

Συμπερασματικά, η βιοπληροφορική ανάλυση έδειξε ότι αυτές οι πρωτεΐνες εμπλέκονται κυρίως στη μετανάστευση, τον πολλαπλασιασμό και την εισβολή των κυττάρων και πολλές από αυτές έχουν αναφερθεί ότι σχετίζονται με καρκίνους. Αυτή η έρευνα αποκάλυψε σημαντικές πρωτεογενομικές διαφορές μεταξύ των κυττάρων NSCLC και SLCL, οι οποίες μπορούν να είναι συμπληρωματικές με τα υπάρχοντα γενομικά και πρωτεομικά δεδομένα που σχετίζονται με καρκίνους του πνεύμονα και θα είναι ζωτικής σημασίας για την κατανόηση του μοριακού μηχανισμού των καρκίνων του πνεύμονα σε επίπεδο βιολογίας.

clusterProfiler

Η αύξηση των ποσοτικών δεδομένων που παράγονται από τη μεταγραφομική (transcriptomics) και την πρωτεομική (proteomics) απαιτεί ολοκληρωμένες στρατηγικές για ανάλυση. Το clusterProfiler (91) είναι ένα εργαλείο της R το οποίο αυτοματοποιεί τη διαδικασία ταξινόμησης βιολογικών όρων και την ανάλυση εμπλουτισμού των συστάδων γονιδίων (gene clusters). Επί του παρόντος, το clusterProfiler υποστηρίζει τρία είδη, συμπεριλαμβανομένων των ανθρώπων, των ποντικών και της ζύμης. Οι

μέθοδοι που παρέχονται σε αυτό το πακέτο μπορούν εύκολα να επεκταθούν σε άλλα είδη και οντολογίες (ontologies).

Η ανάλυση ομαδοποίησης (clustering analysis) χρησιμοποιείται ευρέως για να αποκαλύψει τα κρυφά μοτίβα σε επίπεδο συστημάτων, ιδίως για την αναζήτηση κοινών προαγωγέων (promoters) ή ρυθμιστών (regulators), για την ταξινόμηση βιολογικών διεργασιών, για την πρόβλεψη κακώς χαρακτηρισμένων ή νέων γονιδίων σε συνέκφραση με γνωστές λειτουργίες και για την ανίχνευση πρωτεϊνικών κοινοτήτων (protein communities).

Ένας άλλος κοινός τρόπος για την αναζήτηση κοινών λειτουργιών μεταξύ των γονιδίων είναι η ενσωμάτωση της βιολογικής γνώσης που παρέχεται από τις βιολογικές οντολογίες (ontologies). Για παράδειγμα, η Gene Ontology (GO) χαρακτηρίζει γονίδια σε βιολογικές διεργασίες, μοριακές λειτουργίες και κυτταρικά συστατικά σε μια κατευθυνόμενη ακυκλική δομή γραφήματος και η Kyoto Encyclopedia of Genes and Genomes (KEGG) annotates γονίδια σε μονοπάτια. Τα περισσότερα εργαλεία μπορούν να υπολογίσουν αυτόματα τις στατιστικά σημαντικές κατηγορίες, το manual βήμα για την επιλογή ενδιαφέροντων συστάδων (clusters) που ακολουθείται από ανάλυση εμπλουτισμού (enrichment analysis) σε κάθε επιλεγμένο σύμπλεγμα εξακολουθεί να είναι αργό και κουραστικό. Το clusterProfiler πραγματοποιεί στατιστική ανάλυση του GO και του KEGG, επιτρέποντας τη σύγκριση βιολογικών θεμάτων μεταξύ των συστάδων γονιδίων.

Χρήση του clusterProfiler

Το clusterProfiler χρησιμοποιείται κυρίως για σύγκριση συστάδων γονιδίων. Δεν περιορίζεται σε ομάδες γονιδίων που λαμβάνονται από δεδομένα γονιδιακής έκφρασης, αλλά μπορεί επίσης να εφαρμοστεί σε ομάδες γονιδίων από άλλες προσεγγίσεις, όπως αναλύσεις αλληλεπιδράσεων μεταξύ πρωτεϊνών (protein-protein interactions, PPI) και γονίδια-στόχους miRNA. Συγκεκριμένα προσφέρει τρεις μεθόδους: groupGO, enrichGO και enrichKEGG, για γονιδιακή ταξινόμηση και αναλύσεις εμπλουτισμού. Είναι σημαντικό όμως να αναφερθεί το γεγονός ότι το clusterProfiler εφαρμόζει βιολογική ταξινόμηση όρων και αναλύσεις εμπλουτισμού στη σύγκριση συστάδων γονιδίων, βοηθώντας στην καλύτερη κατανόηση των λειτουργιών ανώτερης τάξης του βιολογικού συστήματος. Υιοθετήθηκε επίσης ένα σχέδιο μοντέλου με μεγαλύτερη ευελιξία. Η λειτουργία σύγκρισης του clusterProfiler έχει σχεδιαστεί ως ένα γενικό πλαίσιο για τη σύγκριση κάθε είδους βιολογικών και βιοϊατρικών οντολογιών. Επιπλέον, το clusterProfiler παρέχει μια ενότητα οπτικοποίησης για την εμφάνιση των αποτελεσμάτων ανάλυσης.

Αυτό το πακέτο είναι ένα απλό στη χρήση εργαλείο, ειδικά σχεδιασμένο για βιολόγους που επιθυμούν να αναλύσουν δεδομένα υψηλής απόδοσης που λαμβάνονται από τη μεταγραφομική (transcriptomics) ή πρωτεομική (proteomics). Μπορεί εύκολα να επεκταθεί για την υποστήριξη νέων οργανισμών και να ενσωματωθεί σε μια ακολουθία εργασιών (pipeline) για ανάλυση δεδομένων υψηλής απόδοσης, ειδικά σε συνεργασία με άλλα πακέτα από το Bioconductor.

Παράδειγμα χρήσης του clusterProfiler

Οι Wu et al. (92) χρησιμοποίησαν multi-omics μεθόδους έρευνας για τη διερεύνηση γονιδίων που σχετίζονται με την κυτταρολυτική δραστηριότητα μέσω των ανοσορυθμιστικών παραγόντων HAVCR2 (TIM3) που επηρεάζουν την επιβίωση και την πρόγνωση του αδενοκαρκινώματος του πνεύμονα.

Το clusterProfiler χρησιμοποιήθηκε για να χαρακτηρίσει λειτουργικά τους διαφορετικούς παράγοντες και να διερευνήσει ολοκληρωμένα τη λειτουργική συσχέτιση αυτών των προγνωστικών γονιδίων. Η Gene Ontology (GO) και η Εγκυκλοπαίδεια Γονιδίων και Γονιδιωμάτων του Κιότο (KEGG) χρησιμοποιήθηκαν για την αξιολόγηση σχετικών λειτουργικών κατηγοριών. Τα μονοπάτια εμπλουτισμού GO και KEGG με τιμές p και Q μικρότερες από 0,05 θεωρήθηκαν ως σημαντικές κατηγορίες. Η ανάλυση εμπλουτισμού GO και KEGG έδειξε ότι αυτά τα διαφορετικά εκφραζόμενα γονίδια εμπλουτίστηκαν σημαντικά σε μεγάλο αριθμό μονοπατιών. Για παράδειγμα, υπάρχει upregulation της κυτταρικής ενεργοποίησης και της δέσμευσης πεπτιδικού αντιγόνου. Ένας μεγάλος αριθμός γονιδίων εμπλουτίστηκε στην ενεργοποίηση των λεμφοκυττάρων και σε άλλες οδούς. Στη διαδικασία του εμπλουτισμού KEGG, βρέθηκαν διάφορες οδοί, πρωτογενείς ανοσοανεπάρκειες και μεταναστεύσεις λευκοκυττάρων. Ταυτόχρονα, ένας μεγάλος αριθμός γονιδίων εμπλουτίστηκε σε μονοπάτια που σχετίζονται με το μεταβολισμό.

Ο ανοσορυθμιστικός παράγοντας TIM3 μπορεί να ρυθμίζει τον πολλαπλασιασμό και την εισβολή των κυττάρων του αδενοκαρκινώματος του πνεύμονα μέσω σηματοδοτικών μονοπατιών (όπως η κυτταρολυτική δραστηριότητα), να ρυθμίζει περαιτέρω το ανοσοποιητικό μικροπεριβάλλον του

αδενοκαρκινώματος του πνεύμονα και έχει μια ορισμένη συσχέτιση κινδύνου με την αντίσταση των καρκινικών κυττάρων στο φάρμακο, η οποία επηρεάζει έμμεσα την επιβίωση και πρόγνωση ασθενών με αδενοκαρκίνωμα πνεύμονα σε κάποιο βαθμό. Το clusterProfiler συνέβαλε σημαντικά στην εξαγωγή συμπερασμάτων μέσω λειτουργικών χαρακτηρισμών των δεδομένων αυτής της μελέτης.

Metascape

Το Metascape (93) είναι μια διαδικτυακή πύλη που έχει σχεδιαστεί για να παρέχει μια ολοκληρωμένη πηγή σχολιασμού και ανάλυσης λίστας γονιδίων για πειραματικούς βιολόγους. Όσον αφορά τα χαρακτηριστικά σχεδιασμού, το Metascape συνδυάζει λειτουργικό εμπλουτισμό (functional enrichment), ανάλυση αλληλεπίδρασης (interactome analysis), σχολιασμό γονιδίων (gene annotation) και αναζήτηση μελών (membership search) για να αξιοποιήσει περισσότερες από 40 ανεξάρτητες βάσεις γνώσης σε μια ολοκληρωμένη πύλη. Επιπλέον, διευκολύνει τις συγκριτικές αναλύσεις των συνόλων δεδομένων σε πολλαπλά ανεξάρτητα και ορθογώνια (orthogonal) πειράματα. Το Metascape παρέχει μια σημαντικά απλοποιημένη εμπειρία χρήστη μέσω μιας διεπαφής Express Analysis με ένα κλικ για τη δημιουργία ερμηνεύσιμων αποτελεσμάτων. Συνολικά, το Metascape είναι ένα αποτελεσματικό και αποδοτικό εργαλείο για πειραματικούς βιολόγους να αναλύουν και να ερμηνεύουν διεξοδικά τις μελέτες OMICs στην εποχή των μεγάλων δεδομένων.

Το Metascape αναπτύχθηκε για να αξιοποιήσει τις βέλτιστες πρακτικές αναλύσεων δεδομένων OMICs που έχουν εμφανιστεί την τελευταία δεκαετία σε

μια ολοκληρωμένη πύλη. Τα χαρακτηριστικά αυτής της πλατφόρμας παρέχουν βελτιωμένη απόδοση δεδομένων σε επίπεδο συστήματος, που έχουν σχεδιαστεί για να προάγουν την ανάπτυξη υποθέσεων με δυνατότητα δράσης. Αυτές περιλαμβάνουν λύσεις μηχανικής δεδομένων (data-engineering solutions) για περισσότερες από 40 βάσεις γνώσεων, καθώς και ολοκληρωμένες και απλοποιημένες ροές εργασίας για σχολιασμό γονιδίων, αναζήτηση μελών και συγκριτική ανάλυση πολλαπλών καταλόγων. Επιπλέον, παρέχονται μοναδικά γραφικά αποτελέσματα που ενισχύουν τις αναφορές και ενισχύουν την κατανόηση δεδομένων. Συνολικά, το Metascape παρέχει ένα ολοκληρωμένο και φιλικό προς τον χρήστη εργαλείο web που έχει σχεδιαστεί για να διευκολύνει την ανάλυση και την ερμηνεία δεδομένων OMIC πολλαπλών πλατφορμών για την πειραματική ερευνητική κοινότητα.

Επισκόπηση της μεθόδου

Η κύρια εμπειρία χρήστη Metascape παρέχεται μέσω μιας διεπαφής ανάλυσης με ένα κλικ. Ο χρήστης παρέχει πρώτα μία ή πολλές λίστες γονιδίων εισόδου και, στη συνέχεια, ξεκινά μια αυτοματοποιημένη ροή εργασιών ανάλυσης που αποτελείται από τέσσερα κύρια στοιχεία: μετατροπή αναγνωριστικού (identifier conversion), σχολιασμός γονιδίου (gene annotation), αναζήτηση μέλους (membership search) και ανάλυση εμπλουτισμού (enrichment analysis) (CAME process). Παράγεται μια αναφορά ανάλυσης που συνοψίζει τα βασικά αποτελέσματα, συνοδευόμενη από ένα βιβλίο εργασίας του Excel, μια παρουσίαση PowerPoint και ένα πακέτο Zip που περιέχει όλα τα υποστηρικτικά (supporting) αρχεία δεδομένων.

Η πλατφόρμα Metascape σχεδιάστηκε για να περιορίζει τις λειτουργίες του χρήστη σε έναν ελάχιστο αριθμό αλληλεπιδράσεων. Τα δεδομένα εισόδου υποβάλλονται σε αυτόματη επεξεργασία για την αναγνώριση τύπων αναγνωριστικών, όπως Entrez Gene ID, σύμβολο γονιδίου, αριθμός RefSeq, Ensembl ID κ.λπ., χωρίς προδιαγραφές χρήστη. Η εισαγωγή αρχείου υποστηρίζει πολλαπλές ευέλικτες μορφές και μπορεί να ενσωματώσει μια προαιρετική λίστα γονιδίων υπόβαθρου για την προσαρμογή της ανάλυσης εμπλουτισμού. Τα αναγνωριστικά γονιδίων από δέκα δημοφιλείς οργανισμούς μοντέλων αναγνωρίζονται από το Metascape (βλ. Μέθοδοι) και η αυτόματη χαρτογράφηση ορθολόγου επιτρέπει την ανάλυση λιστών γονιδίων στο πλαίσιο βάσεων δεδομένων για συγκεκριμένο είδος σε οποιονδήποτε από αυτούς τους δέκα οργανισμούς. Η ανάλυση στόχου ενός είδους είναι ο άνθρωπος από προεπιλογή, καθώς οι ανθρωποκεντρικές βάσεις δεδομένων είναι συχνά οι πιο ολοκληρωμένες.

Η ιδέα πίσω από την Express Analysis είναι να ενσωματώσει τις πιο ευρέως χρησιμοποιούμενες επιλογές ανάλυσης και τις βέλτιστες πρακτικές ανάλυσης σε μια βελτιωμένη προσέγγιση. Για παράδειγμα, η Express Analysis περιλαμβάνει αυτόματα μεταδεδομένα σχολιασμού γονιδίων που είναι πιθανό να παρέχουν πολύτιμο πλαίσιο για την ερμηνεία των υποψηφίων στις περισσότερες περιπτώσεις, όπως η σύντομη περιγραφή τους, οι βασικές βιολογικές διεργασίες που εμπλέκονται, οι πρωτεϊνικές λειτουργίες, οι υποκυτταρικές θέσεις και οι ρόλοι σε κανονικές οδούς. Ενσωματώνει ένα βασικό σύνολο προεπιλεγμένων οντολογιών για ανάλυση εμπλουτισμού, συμπεριλαμβανομένων διεργασιών GO, μονοπατιών KEGG, συνόλων γονιδίων Reactome, κανονικών οδών και συμπλεγμάτων CORUM. Η Express

Analysis ορίζει επίσης αυτόματα στατιστικά φίλτρα για να αφαιρέσει όρους οντολογίας και δίκτυα αλληλεπίδρασης που δεν πληρούν ελάχιστες στατιστικές απαιτήσεις. Εναλλακτικά στην Express Analysis, οι προχωρημένοι χρήστες μπορούν να ελέγξουν τα τέσσερα μεμονωμένα βήματα της ροής εργασίας CAME (CAME process) και να αντικαταστήσουν τις προεπιλεγμένες ρυθμίσεις μέσω της ροής εργασίας Custom Analysis.

Ανάλυση μετατροπής αναγνωριστικού (Identifier conversion analysis)

Το πρώτο βήμα της ροής εργασίας ανάλυσης Metascape είναι η μετατροπή αναγνωριστικού γονιδίου. Το Metascape αναγνωρίζει αυτόματα τους δημοφιλείς τύπους αναγνωριστικών γονιδίων που περιγράφονται παραπάνω, καθώς και τα ονόματα των πρωταρχικών θέσεων από διάφορες βάσεις δεδομένων μοντέλων οργανισμών, όπως SGD, FlyBase, WormBase κ.λπ. Τα αναγνωριστικά πηγής (source identifiers) αντιστοιχίζονται σε μια μοναδική λίστα Entrez Gene ID πριν από τις αναλύσεις, καθώς ένας σημαντικός αριθμός βάσεων γνώσεων βιοπληροφορικής στις οποίες βασίζεται το Metascape βασίζεται στα Entrez Gene ID ως τα κύρια κλειδιά τους. Τα καταργημένα αναγνωριστικά γονιδίων Entrez αναγνωρίζονται επίσης και αντικαθίστανται από αναγνωριστικά διαδόχου όταν συναντώνται.

Πολλές βάσεις δεδομένων σχολιασμού γονιδίων, μονοπατιών και αλληλεπίδρασης πρωτεϊνών καταρτίζονται κυρίως για ανθρώπινα γονίδια ή πρωτεΐνες. Ως εκ τούτου, μπορεί να είναι ωφέλιμο να τοποθετούνται υποψήφια γονίδια που λαμβάνονται σε οργανισμούς-μοντέλα στα ανθρώπινα ορθόλογά τους πριν από την ανάλυση. Η χαρτογράφηση Ortholog (ortholog mapping)

είναι μια ενσωματωμένη δυνατότητα και μπορεί να ενεργοποιηθεί επιλέγοντας ρητά το είδος ανάλυσης στόχου, το οποίο θα καθοδηγήσει τα επόμενα βήματα ανάλυσης για τη χρήση πόρων για το συγκεκριμένο είδος.

Γονιδιακός σχολιασμός (Gene annotation)

Λεπτομερείς σχολιασμοί για γονίδια παρέχουν βιολογικό πλαίσιο που θα χρησιμεύσει ως σημαντικά κριτήρια επιλογής κατά την επιλογή υποψηφίων για μετέπειτα μελέτες. Το Metascape ενσωματώνει πληροφορίες σχολιασμού από περισσότερες από 40 βάσεις γνώσεων, που περιλαμβάνουν περιγραφές γονιδίων (gene descriptions), περιλήψεις γονιδίων (gene summaries), επιπτώσεις ασθενειών, γονιδιωματικές παραλλαγές (genome variants), υποκυτταρικούς εντοπισμούς (subcellular localizations), έκφραση ιστού (tissue expression), διαθεσιμότητα χημικών ανιχνευτών (chemical probes) κ.λπ. Η Express Analysis παρέχει εννέα περιγραφικά πεδία από προεπιλογή, ωστόσο, η δυνατότητα προσαρμοσμένης ανάλυσης επιτρέπει στο χρήστη να προσαρτήσσει έως και 47 στήλες μεταδεδομένων σε ένα υπολογιστικό φύλλο σχολιασμού γονιδίου.

Για να αφομοιώσει αυτό το επίπεδο πληροφοριών χρησιμοποιώντας τρέχοντα εργαλεία, ο ερευνητής πρέπει πρώτα να εντοπίσει τις κατάλληλες βάσεις γνώσεων που περιέχουν τις απαιτούμενες πληροφορίες. Δεδομένου ότι πολλές από αυτές τις πύλες δεν ενσωματώνουν δυνατότητες μεταφόρτωσης λίστας γονιδίων για σχολιασμό παρτίδας, οι χρήστες πρέπει, σε ορισμένες περιπτώσεις, να αναζητούν μεμονωμένα υποψήφια γονίδια ένα κάθε φορά. Για παράδειγμα, η αναγνώριση όλων των εξωκυτταρικών πρωτεϊνών μέσω της

UniProt ή όλων των χημικών ανιχνευτών μέσω του DrugBank γίνεται επαχθής, χρησιμοποιώντας μια λίστα με εκατοντάδες υποψήφια γονίδια. Έτσι, η δυνατότητα αυτοματοποιημένου σχολιασμού γονιδίων του Metascape παρέχει στους χρήστες λεπτομερείς πληροφορίες σχολιασμού που δεν είναι εύκολα προσβάσιμες κατά τα άλλα, βοηθώντας στην ερμηνεία δεδομένων μεγάλης κλίμακας συνόλων δεδομένων και παρέχοντας κριτήρια ιεράρχησης για τα υποψήφια γονίδια.

Ομαδοποίηση εμπλουτισμού (Enrichment clustering)

Η ανάλυση εμπλουτισμού περιλαμβάνει τον πυρήνα των περισσότερων υπαρχουσών πυλών σχολιασμού γονιδίων. Κατά τη διάρκεια της ανάλυσης εμπλουτισμού, η λίστα γονιδίων εισόδου συγκρίνεται με χιλιάδες σύνολα γονιδίων που ορίζονται από τη συμμετοχή τους σε συγκεκριμένες βιολογικές διεργασίες, τον εντοπισμό πρωτεϊνών, την ενζυματική λειτουργία, τη συμμετοχή στο μονοπάτι ή άλλα χαρακτηριστικά. Τα σύνολα γονιδίων των οποίων τα μέλη υπερεκπροσωπούνται σημαντικά στη λίστα γονιδίων εισόδου αναφέρονται στους χρήστες για να χρησιμεύσουν ως υποτιθέμενες βιολογικές γνώσεις στη μελέτη τους. Ένα πρόβλημα που παραμελείται στις περισσότερες τρέχουσες αναλύσεις είναι ότι οι πλεονασμοί σε περιγραφείς και οντολογίες μπορούν συχνά να περιπλέξουν την ερμηνεία του αποτελέσματος. Για παράδειγμα, οι όροι οντολογίας που βρίσκονται στο GO σχηματίζουν μια ιεραρχική δομή αυξανόμενης ευαισθησίας, καθιστώντας τους όρους εγγενώς περιττούς. Οι όροι σε διαφορετικές πηγές οντολογίας, όπως GO, KEGG και MSigDB, κ.λπ. μπορούν επίσης να σχετίζονται στενά. Ως αποτέλεσμα, η ανάλυση λειτουργικού

εμπλουτισμού μπορεί να εντοπίσει επικαλυπτόμενους ή σχετικούς όρους, καθιστώντας δύσκολη την εξαγωγή μη περιττών και αντιπροσωπευτικών διαδικασιών για αναφορά στο αποτέλεσμα της ανάλυσης. Με την απορρόφηση των περισσότερων πλεονασμάτων σε αντιπροσωπευτικά συμπλέγματα, η ομαδοποίηση εμπλουτισμού εξαλείφει τα μπερδεμένα ζητήματα ερμηνείας δεδομένων που μπορεί να προκύψουν από την αναφορά πολλαπλών οντολογιών.

Ανάλυση δικτύου πρωτεϊνών (Protein Network Analysis)

Η ανάλυση λιστών γονιδίων στο πλαίσιο των πρωτεϊνικών αλληλεπιδράσεων μπορεί να βοηθήσει στην εύρεση βιοχημικών συμπλεγμάτων ή στοιχείων μεταγωγής σήματος (transduction components) που διέπουν τις βιολογικές εκροές. Ωστόσο, μια τυπική λίστα γονιδίων με μερικές εκατοντάδες ανθρώπινες πρωτεΐνες μπορεί να σχηματίσει εξαιρετικά περίπλοκες διατάξεις δικτύου, καθιστώντας δύσκολη την ερμηνεία. Για να συναγάγει πιο βιολογικά ερμηνεύσιμα αποτελέσματα, το Metascape εφαρμόζει έναν αλγόριθμο αναγνώρισης που ονομάζεται MCODE για να εξάγει αυτόματα συμπλέγματα πρωτεϊνών που είναι ενσωματωμένα στο μεγάλο δίκτυο. Εκμεταλλευόμενοι τη δυνατότητα ανάλυσης λειτουργικού εμπλουτισμού του Metascape, οι τρεις πιο σημαντικά εμπλουτισμένοι όροι οντολογίας συνδυάζονται για να σχολιάσουν υποτιθέμενους βιολογικούς ρόλους για κάθε σύμπλεγμα MCODE. Τα αποτελέσματα του Metascape είναι σχετικά απλά στην οπτικοποίηση και την ερμηνεία.

Μετα-ανάλυση καταλόγου πολλαπλών γονιδίων (Multi-gene-list meta-analysis)

Ο πολλαπλασιασμός και η γενική προσαρμογή των πλατφορμών OMICs, συμπεριλαμβανομένης της πρωτεωμικής, της γονιδιωματικής, των λειτουργικών γενετικών οθονών (functional genetics screens) και της μεταβολομικής, έχει καταστήσει τη μετα-ανάλυση πολλαπλών καταλόγων γονιδίων μια κρίσιμη και σε μεγάλο βαθμό αδιευκρίνιστη ανάγκη. Οι συνήθεις συγκριτικές προσεγγίσεις περιλαμβάνουν τη χρήση διαγραμμάτων Venn για τον εντοπισμό γονιδίων που είναι κοινά ή μοναδικά σε ορισμένες λίστες. Έχει αναφερθεί προηγουμένως ότι μια επικάλυψη μεταξύ των συνόλων δεδομένων omics είναι πιο εύκολα εμφανής σε επίπεδο μονοπατιών ή συμπλεγμάτων πρωτεϊνών που μπορεί να είναι πολύ πιο δύσκολο να εφαρμοστούν χωρίς εκτεταμένη υπολογιστική τεχνογνωσία.

Το Metascape έχει σχεδιαστεί για να επιτρέπει τη διασταυρούμενη σύγκριση ενός αυθαίρετου αριθμού λιστών γονιδίων τόσο σε ταυτότητες γονιδίων όσο και σε οντολογίες. Συγκεκριμένα, με την υποβολή πολλαπλών λιστών γονιδίων, όλοι οι υποψήφιοι συγχωνεύονται σε μία λίστα, ενώ κωδικοποιούνται οι αρχικές λίστες ως πρόσθετες δυαδικές στήλες μελών. Αυτή η περιστρεφόμενη διάταξη απλοποιεί σημαντικά την αναπαράσταση των αποτελεσμάτων της ανάλυσης, καθώς τόσο ο σχολιασμός γονιδίου όσο και η αναζήτηση μελών μπορούν να εφαρμοστούν στη συνδυασμένη λίστα με τον ίδιο τρόπο όπως σε μια εισαγωγή λίστας γονιδίων. Η χρήση ενός ενιαίου υπολογιστικού φύλλου εξόδου για την ενσωμάτωση στοιχείων που συλλέγονται από κάθε λίστα γονιδίων επιτρέπει έτσι αποτελεσματικό παγκόσμιο φιλτράρισμα γονιδίων και ιεράρχηση προτεραιοτήτων. Η επικάλυψη μεταξύ των λιστών γονιδίων οπτικοποιείται χρησιμοποιώντας μια γραφική παράσταση

Circos, η οποία είναι μια πιο διαισθητική και κλιμακούμενη αναπαράσταση σε σύγκριση με ένα διάγραμμα Venn.

Για να διευκολυνθεί η κατανόηση των μονοπατιών (και των συστάδων μονοπατιών) που μοιράζονται ή αποδίδονται επιλεκτικά σε συγκεκριμένες λίστες γονιδίων, αναπτύχθηκαν πρόσθετες απεικονίσεις για το Metascape. Πρώτον, το Metascape απεικονίζει τα κορυφαία εμπλουτισμένα συμπλέγματα και τα μοτίβα εμπλουτισμού τους σε πολλές λίστες γονιδίων ως συγκεντρωμένο θερμικό χάρτη (heatmap). Ο χάρτης θερμότητας συμπληρώνεται από ένα δίκτυο εμπλουτισμού όπου κάθε κόμβος δικτύου αντιπροσωπεύεται από ένα γράφημα πίτας, όπου το μέγεθος του τομέα είναι ανάλογο με τον αριθμό των γονιδίων που προέρχονται από κάθε λίστα γονιδίων. Αυτή η αναπαράσταση έχει σχεδιαστεί για να είναι μια βιολογικά διαισθητική απεικόνιση επιλεκτικών ή κοινών συστάδων μονοπατιών. Είναι σημαντικό ότι αυτή η προσέγγιση μπορεί επίσης να εφαρμοστεί για την οπτικοποίηση δικτύων αλληλεπίδρασης πρωτεΐνης-πρωτεΐνης για να καταστεί δυνατή η αποσαφήνιση κοινών ή επιλεκτικών πολύπλοκων συστατικών που εμπλουτίζονται σε σύνολα δεδομένων OMICS.

Πλεονεκτήματα χρήσης του Metascape

Το Metascape σχεδιάστηκε για να παρέχει πιο ερμηνεύσιμα και εφαρμόσιμα αποτελέσματα στους ερευνητές. Ο πλεονασμός μεταξύ των όρων οντολογίας παρουσιάζει ένα σημαντικό εμπόδιο στην ερμηνεία των αποτελεσμάτων. Η κύρια στρατηγική που υιοθετείται από υπάρχουσες πύλες, όπως φαίνεται στα g:Profiler, WebGestalt, Babelomics και άλλα, είναι η χρήση

κατευθυνόμενου ακυκλικού γραφήματος (DAG) ή δενδρικών δομών ενσωματωμένων σε οντολογίες όπως το GO. Το WebGestalt παρέχει μια επιλογή μόνο χρησιμοποιώντας ένα υποσύνολο όπως το GOSlim για τη μείωση του πλεονασμού. Ωστόσο, αυτές οι στρατηγικές δεν είναι εφαρμόσιμες για τη μείωση των απολύσεων σε διαφορετικές πηγές οντολογίας. Τα Metascape και DAVID, αντίθετα, υιοθετούν μια δυναμική λύση ομαδοποίησης που αποφεύγει τις προαναφερθείσες ελλείψεις. Το Metascape απλοποιεί περαιτέρω την παρουσίαση των αποτελεσμάτων μέσω ραβδογραμμάτων και απεικονίσεων θερμικών χαρτών, σε σύγκριση με την παρουσίαση σε πίνακα του DAVID. Αν και η τιμή περιθωρίου σφάλματος (p-value) χρησιμοποιείται ως η προεπιλεγμένη μέτρηση αναπαράστασης για την κατάταξη εμπλουτισμένων όρων, ωστόσο έχει τους περιορισμούς της. Οι χρήστες μπορεί να χρειαστεί να εξετάσουν το ενδεχόμενο χρήσης εναλλακτικών μετρήσεων, όπως η διορθωμένη τιμή q πολλαπλών δοκιμών ή ο παράγοντας εμπλουτισμού με ευαισθησία στο βιολογικό περιβάλλον. Όλες οι δημοφιλείς μετρήσεις εμπλουτισμού, συμπεριλαμβανομένων των λεπτομερών μετρήσεων γονιδίων, υπολογίζονται και διατίθενται για αναλύσεις εκτός σύνδεσης. Το Metascape ενισχύει περαιτέρω την ερμηνευσιμότητα των αποτελεσμάτων της ανάλυσης μέσω προηγμένων γραφικών και υπολογιστικών φύλλων, μιας συνοπτικής παρουσίασης PowerPoint και μιας διαδικτυακής αναφοράς ανάλυσης σε στυλ περιοδικού που περιέχει πρόσθετες περιγραφές μεθόδων και αναφορές.

Συμπερασματικά, το Metascape έχει σχεδιαστεί για πειραματιστές (βιολόγους πάγκου) ώστε να εφαρμόζουν ισχυρές ακολουθίες εργασιών υπολογιστικής ανάλυσης για την ανάλυση και την ερμηνεία συνόλων δεδομένων μεγάλης κλίμακας. Για την αντιμετώπιση των κενών στα τρέχοντα

εργαλεία ανάλυσης, το Metascape εισάγει μια ροή εργασίας που ενσωματώνει σχολιασμούς γονιδίων, ανάλυση μελών και δυνατότητες μετα-ανάλυσης λίστας πολλαπλών γονιδίων. Το πλούσιο σύνολο εργαλείων ανάλυσης του είναι προσβάσιμο μέσω μιας βολικής σύντομης ανάλυσης (Express Analysis) με ένα κλικ και τα αποτελέσματα κοινοποιούνται μέσω μιας αναφοράς. Η ευρύτερη υιοθέτηση του Metascape θα ενισχύσει σημαντικά τη βιολογική ερμηνεία των μελετών OMICS, επιτρέποντας στους ερευνητές να αναλύσουν απευθείας τα δεδομένα τους για να εντοπίσουν νέους θεραπευτικούς στόχους, μηχανισμούς δράσης ή μοριακή εικόνα της νόσου.

Παράδειγμα χρήσης του Metascape

Στη μελέτη των Qi et al. (94) διερευνήθηκαν τα πρότυπα έκφρασης και οι προγνωστικές αξίες των PLODs (Procollagen-lysine, 2-oxoglutarate 5-dioxygenases) σε ασθενείς με αδενοκαρκίνωμα πνεύμονα (Lung Adenocarcinoma, LUAD) και ακανθοκυτταρικό καρκίνωμα πνεύμονα (Lung Squamous Cell Carcinoma, LUSC). Συγκεκριμένα, πραγματοποιήθηκαν βιοπληροφορικές αναλύσεις δημόσιων βάσεων δεδομένων για τη διερεύνηση της διαφορικής έκφρασης, της μετάλλαξης και της προγνωστικής αξίας των μελών της οικογένειας των PLOD. Αναλύθηκαν οι πιθανές βιολογικές λειτουργίες των PLODs και διερευνήθηκαν περαιτέρω οι συσχετίσεις μεταξύ PLODs και ανοσολογικών δεικτών, συμπεριλαμβανομένων των μορίων που σχετίζονται με το ανοσοποιητικό και των ανοσοδιηθημένων κυττάρων.

Το Metascape χρησιμοποιήθηκε για την ανάλυση κυτταρικών μονοπατιών και εμπλουτισμού των γονιδίων PLOD1/2/3 και των

συνεκφραζόμενων γονιδίων που εξήχθησαν από το cBioPortal (www.cbioportal.org) []. Οι αναλύσεις εμπλουτισμού που πραγματοποιήθηκαν περιελάμβαναν αναλύσεις κυτταρικών μονοπατιών Gene Ontology (GO) και Kyoto Encyclopedia of Genes and Genomes (KEGG). Οι αναλύσεις GO αποτελούνταν από βιολογικές διεργασίες, κυτταρικά συστατικά και μοριακές λειτουργίες.

Τα συνεκφραζόμενα γονίδια με συντελεστές συσχέτισης Spearman $>0,4$ επιλέχθηκαν για τη διεξαγωγή αναλύσεων μονοπατιών GO και KEGG στο Metascape. Βιολογικές διεργασίες, συμπεριλαμβανομένης της κυτταρικής διαίρεσης, της θετικής ρύθμισης της διαδικασίας του κυτταρικού κύκλου και της απόκρισης στην υποξία ρυθμίστηκαν σε σημαντικό βαθμό από τα γονίδια συνέκφρασης των PLODs στο LUAD. Στη συνέχεια, πραγματοποιήθηκε περαιτέρω ανάλυση μονοπατιών KEGG και βρέθηκε ότι οι λειτουργίες αυτών των συνεκφραζόμενων γονιδίων εμπλέκονται σε πολλαπλά κυτταρικά μονοπάτια, συμπεριλαμβανομένων του κυτταρικού κύκλου, της αντιγραφής του DNA, της γλυκόλυσης/γλυκονεογένεσης, του ομόλογου ανασυνδυασμού, του κεντρικού μεταβολισμού του άνθρακα στον καρκίνο, της οδού σηματοδότησης p53, μονοπάτια στον καρκίνο και μονοπάτι σηματοδότησης PI3K-Akt.

Επιπλέον, πραγματοποιήθηκαν αναλύσεις εμπλουτισμού των συνεκφραζόμενων γονιδίων των PLODs στο LUSC. Όσον αφορά τις βιολογικές διεργασίες, τα συνεκφραζόμενα γονίδια των PLODs από το LUSC εμπλουτίστηκαν σε οργάνωση εξωκυτταρικής δομής, μετα-μεταφραστική τροποποίηση πρωτεϊνών και απόκριση στην υποξία. Για την ανάλυση KEGG αυτών των γονιδίων από το LUSC, βρέθηκαν τρία σχετικά μονοπάτια: αλληλεπίδραση ECM-υποδοχέα, επεξεργασία πρωτεϊνών στο ενδοπλασματικό

δίκτυο και βιοσύνθεση glycosaminoglycan-chondroitin sulfate/dermatan sulfate.

Τέλος στην έρευνα συμπεραίνεται ότι τα PLODs εκφράζονται σε μεγάλο βαθμό στον καρκίνο του πνεύμονα και μπορεί να είναι κατάλληλοι προγνωστικοί δείκτες. Στη διαπίστωση αυτών των συμπερασμάτων βοήθησε και το Metascape, οδηγώντας τους από τα βασικά δεδομένα στην εξαγωγή αναλύσεων εμπλουτισμού και κυτταρικών μονοπατιών.

ggplot2

Το ggplot2 (95) είναι ένα πακέτο της R για την παραγωγή στατιστικών ή γραφημάτων δεδομένων. Σε αντίθεση με τα περισσότερα άλλα πακέτα γραφημάτων, το ggplot2 έχει μια υποκείμενη γραμματική, βασισμένη στο *Grammar of Graphics* (Wilkinson 2005), που επιτρέπει τη σύνθεση γραφημάτων συνδυάζοντας ανεξάρτητα στοιχεία, γεγονός το οποίο κάνει το ggplot2 ισχυρό. Οι χρήστες δεν περιορίζονται σε σύνολα προκαθορισμένων γραφικών και μπορούν να δημιουργήσουν νέα γραφικά που είναι προσαρμοσμένα στο εκάστοτε πρόβλημά. Το ggplot2 είναι στην πραγματικότητα εύκολο στο χειρισμό, καθώς αποτελείται από ένα απλό σύνολο βασικών αρχών και υπάρχουν πολύ λίγες ειδικές περιπτώσεις.

Το ggplot2 παρέχει όμορφα διαγράμματα, εύκολα στη δημιουργία, που περιλαμβάνουν επιπρόσθετες λεπτομέρειες όπως η σχεδίαση λεζάντων. Στην πραγματικότητα, οι προσεκτικά επιλεγμένες προεπιλογές του σημαίνουν ότι μπορούν να παραχθούν γραφήματα υψηλής ποιότητας σε δευτερόλεπτα. Ωστόσο, εάν υπάρχουν ειδικές απαιτήσεις μορφοποίησης, το ολοκληρωμένο

σύστημα θεμάτων του ggplot2 τις διευκολύνει. Σε τελική ανάλυση, αυτό σημαίνει ότι αντί να ξοδεύει τον χρόνο του ο χρήστης κάνοντας το γράφημα να φαίνεται όμορφο, μπορεί να εστιάσει στη δημιουργία του γραφήματος που αποκαλύπτει καλύτερα το μήνυμα στα δεδομένα.

Το ggplot2 έχει σχεδιαστεί για να λειτουργεί επαναληπτικά. Η αρχή γίνεται με ένα επίπεδο που δείχνει τα ανεπεξέργαστα δεδομένα. Στη συνέχεια, προστίθενται επίπεδα σχολιασμών και στατιστικών περιλήψεων. Αυτό επιτρέπει στους χρήστες να παράγουν γραφήματα χρησιμοποιώντας την ίδια δομημένη σκέψη που θα χρησιμοποιούσαν για να σχεδιαστεί μια ανάλυση. Επιπλέον μειώνεται η «απόσταση» μεταξύ ενός αρχικού σχεδίου στη σκέψη με αυτού που θα τυπωθεί τελικά στη σελίδα. Αυτό είναι ιδιαίτερα χρήσιμο για τους αρχάριους χρήστες που δεν έχουν ακόμη αναπτύξει τη δομημένη προσέγγιση ανάλυσης που χρησιμοποιείται από ειδικούς.

Η εκμάθηση της γραμματικής του εργαλείου αυτού όχι μόνο βοηθάει στη δημιουργία γραφημάτων με τα οποία ο χρήστης είναι εξοικειωμένος, αλλά θα βοηθάει επίσης στη δημιουργία νεότερων και βελτιωμένων γραφημάτων. Χωρίς γραμματική, δεν υπάρχει υποκείμενη θεωρία, επομένως τα περισσότερα πακέτα γραφημάτων είναι απλώς μια μεγάλη συλλογή από ειδικές περιπτώσεις. Για παράδειγμα, στη βασική R ένα νέο γραφικό αποτελείται από ακατέργαστα στοιχεία γραφικής παράστασης, όπως γραμμές και σημεία, επομένως είναι δύσκολο να σχεδιαστούν νέα στοιχεία που συνδυάζονται με ήδη υπάρχοντα σχέδια. Στο ggplot2, οι εκφράσεις που χρησιμοποιούνται για τη δημιουργία ενός νέου γραφήματος αποτελούνται από στοιχεία υψηλότερου επιπέδου, όπως αναπαραστάσεις ακατέργαστων δεδομένων και στατιστικούς

μετασχηματισμούς, που μπορούν εύκολα να συνδυαστούν με νέα σύνολα δεδομένων και άλλες γραφικές παραστάσεις.

Ο Wilkinson (2005) δημιούργησε το *Grammar of Graphics* για να περιγράψει τα θεμελιώδη χαρακτηριστικά που βρίσκονται κάτω από όλα τα στατιστικά γραφήματα. Αυτό το βιβλίο είναι μια απάντηση στο ερώτημα τι είναι ένα στατιστικό γράφημα. Το ggplot2 βασίζεται στη γραμματική του Wilkinson εστιάζοντας στην υπεροχή των επιπέδων και προσαρμόζοντάς το για χρήση στην R. Εν συντομία, η γραμματική μας λέει ότι ένα γραφικό αντιστοιχίζει τα δεδομένα με τα αισθητικά χαρακτηριστικά (χρώμα, σχήμα, μέγεθος) γεωμετρικών αντικειμένων (σημεία, γραμμές, ράβδους). Η γραφική παράσταση μπορεί επίσης να περιλαμβάνει στατιστικούς μετασχηματισμούς των δεδομένων και πληροφοριών σχετικά με το σύστημα συντεταγμένων του γραφήματος. Η πτυχοποίηση (faceting) μπορεί να χρησιμοποιηθεί για τη γραφική παράσταση για διαφορετικά υποσύνολα δεδομένων. Ο συνδυασμός αυτών των ανεξάρτητων στοιχείων είναι που συνθέτουν μια γραφική παράσταση.

Όλα τα διαγράμματα αποτελούνται από τα δεδομένα, τις πληροφορίες που οπτικοποιούνται και μια αντιστοίχιση, την περιγραφή του τρόπου με τον οποίο οι μεταβλητές των δεδομένων αντιστοιχίζονται σε αισθητικά χαρακτηριστικά. Υπάρχουν πέντε στοιχεία χαρτογράφησης:

- 1) Το *layer* είναι μια συλλογή από γεωμετρικά στοιχεία και στατιστικούς μετασχηματισμούς. Τα γεωμετρικά στοιχεία, τα *geom* για συντομία, αντιπροσωπεύουν αυτό που πραγματικά βλέπετε στη γραφική παράσταση: σημεία, γραμμές, πολύγωνα κ.λπ. Οι στατιστικοί

μετασχηματισμοί, τα *stat* για συντομία, συνοψίζουν τα δεδομένα: για παράδειγμα, δέσμευση και καταμέτρηση παρατηρήσεων για τη δημιουργία ενός ιστογράμματος ή προσαρμογή ενός γραμμικού μοντέλου.

- 2) Το *scale* χαρτογραφεί τις τιμές αντιστοίχισης στο χώρο δεδομένων σε τιμές στον αισθητικό χώρο. Αυτό περιλαμβάνει τη χρήση χρώματος, σχήματος ή μεγέθους. Το *scale* σχεδιάζει επίσης το υπόμνημα και τους άξονες, που καθιστούν δυνατή την ανάγνωση των αρχικών τιμών δεδομένων από την γραφική παράσταση (μια αντίστροφη χαρτογράφηση).
- 3) Το *coord*, ή αλλιώς ένα σύστημα συντεταγμένων, περιγράφει τον τρόπο με τον οποίο οι συντεταγμένες δεδομένων αντιστοιχίζονται στο επίπεδο του γραφήματος. Παρέχει επίσης άξονες και γραμμές πλέγματος που βοηθούν στην ανάγνωση του γραφήματος. Συνήθως χρησιμοποιείται το καρτεσιανό σύστημα συντεταγμένων, αλλά υπάρχουν και άλλα διαθέσιμα, συμπεριλαμβανομένων των πολικών συντεταγμένων και των προβολών χαρτών.
- 4) Το *facet* (πτυχή) καθορίζει τον τρόπο διαχωρισμού και εμφάνισης υποσυνόλων δεδομένων ως μικρά πολλαπλάσια. Αυτό είναι επίσης γνωστό ως προετοιμασία ή πλέγμα.
- 5) Το *theme* ελέγχει τα πιο λεπτά σημεία εμφάνισης, όπως το μέγεθος της γραμματοσειράς και το χρώμα του φόντου.

Το *ggplot2* είναι μια προσπάθεια να ληφθούν τα θετικά στοιχεία από τα βασικά γραφήματα της R και από αυτά του *lattice graphics* και να τα βελτιωθούν με ένα ισχυρό υποκείμενο μοντέλο που υποστηρίζει την παραγωγή κάθε είδους

στατιστικών γραφικών, με βάση τις αρχές που περιγράφονται παραπάνω. Το συμπαγές υποκείμενο μοντέλο του ggplot2 καθιστά εύκολη την περιγραφή ενός ευρέος φάσματος γραφικών με συμπαγή σύνταξη και ανεξάρτητα στοιχεία που κάνουν εύκολες τις επεκτάσεις. Με σχεδόν 20 χρόνια λειτουργίας, το ggplot2 χρησιμοποιείται πλέον από εκατοντάδες χιλιάδες χρήστες βοηθώντας τους να δημιουργήσουν εκατομμύρια γραφικές παραστάσεις υψηλής ποιότητας, οι οποίες μπορεί να πλασιώνουν εργασίες, παρουσιάσεις μέχρι και επιστημονικά άρθρα.

Cytoscape

Το Cytoscape (96) είναι ένα λογισμικό ανοιχτού κώδικα για την ενοποίηση δικτύων βιομοριακής αλληλεπίδρασης με δεδομένα έκφρασης υψηλής απόδοσης και άλλες μοριακές καταστάσεις σε ένα ενοποιημένο εννοιολογικό πλαίσιο. Αν και εφαρμόζεται σε οποιοδήποτε σύστημα μοριακών συστατικών και αλληλεπιδράσεων, το Cytoscape είναι πιο ισχυρό όταν χρησιμοποιείται σε συνδυασμό με μεγάλες βάσεις δεδομένων πρωτεΐνης-πρωτεΐνης, πρωτεΐνης-DNA και γενετικών αλληλεπιδράσεων που είναι ολοένα και πιο διαθέσιμες για ανθρώπους και μοντέλα οργανισμούς. Το λογισμικό Core του Cytoscape παρέχει βασικές λειτουργίες για τη διάταξη και την αναζήτηση του δικτύου: να ενσωματώσει οπτικά το δίκτυο με προφίλ έκφρασης, φαινοτύπους και άλλες μοριακές καταστάσεις και να συνδέσει το δίκτυο με βάσεις δεδομένων λειτουργικών σχολιασμών. Το Core είναι επεκτάσιμο μέσω μιας απλής αρχιτεκτονικής plug-in, επιτρέποντας την ταχεία ανάπτυξη πρόσθετων υπολογιστικών αναλύσεων και χαρακτηριστικών.

Δομή και κατασκευή του Cytoscape

Η μεταφορά της κεντρικής οργάνωσης (central organizing metaphor) του Cytoscape είναι ένα γράφημα δικτύου, με μοριακά είδη που αναπαρίστανται ως κόμβοι και διαμοριακές αλληλεπιδράσεις ως σύνδεσμοι, δηλαδή ακμές, μεταξύ κόμβων. Το λογισμικό Core του Cytoscape παρέχει βασική λειτουργικότητα για την ενσωμάτωση αυθαίρετων δεδομένων στο γράφημα, οπτική αναπαράσταση του γραφήματος και ενσωματωμένων δεδομένων, εργαλεία επιλογής και φιλτραρίσματος και μια διεπαφή με εξωτερικές μεθόδους που υλοποιούνται ως πρόσθετα.

Ενσωμάτωση δεδομένων

Τα δεδομένα ενσωματώνονται με το μοντέλο γραφήματος χρησιμοποιώντας το *Attributes*. Αυτά είναι ζεύγη (όνομα, τιμή) που αντιστοιχίζουν ονόματα κόμβων ή ακμών σε συγκεκριμένες τιμές δεδομένων. Οι τιμές του *Attributes* μπορούν να λάβουν οποιονδήποτε τύπο (π.χ. συμβολοσειρές κειμένου, διακριτούς ή συνεχείς αριθμούς, διευθύνσεις URL ή λίστες) και είτε φορτώνονται από ένα χώρο αποθήκευσης δεδομένων είτε δημιουργούνται δυναμικά μέσα σε μια περίοδο λειτουργίας. Τα προγράμματα περιήγησης γραφικών επιτρέπουν στο χρήστη να εξετάσει όλα τα χαρακτηριστικά στους επιλεγμένους κόμβους και ακμές.

Μεταφορά σχολιασμών

Ενώ ένα χαρακτηριστικό είναι ένα ενιαίο κατηγορημα ενός κόμβου ή ακμής, ένα *Annotation* αντιπροσωπεύει μια ιεραρχική ταξινόμηση (δηλαδή, μια οντολογία, τυπικά, ένα κατευθυνόμενο άκυκλο γράφημα) προοδευτικά πιο ειδικών περιγραφών ομάδων κόμβων ή ακμών. Οι σχολιασμοί αντιστοιχούν συνήθως σε μια υπάρχουσα αποθήκη γνώσης που είναι μεγάλη, πολύπλοκη και σχετικά στατική, όπως η βάση δεδομένων Gene Ontology. Για παράδειγμα, ο όρος «μεταβολισμός εξόζης», που ορίζεται στο επίπεδο 6 της οντολογίας GO Biological Process Ontology, καλύπτει πολλές πιο συγκεκριμένες διαδικασίες στο επίπεδο 7, όπως ο μεταβολισμός της γαλακτόζης και ο μεταβολισμός της γλυκόζης. Το Cytoscape ενσωματώνει σχολιασμούς με άλλους τύπους δεδομένων δικτύου μεταφέροντας τα επιθυμητά επίπεδα σχολιασμού σε χαρακτηριστικά κόμβου ή ακμής. Χρησιμοποιώντας τον ελεγκτή *Annotation*, είναι δυνατό να υπάρχουν πολλά επίπεδα σχολιασμού ενεργά και να εμφανίζονται ταυτόχρονα, το καθένα ως διαφορετικό χαρακτηριστικό στους κόμβους ή τις άκρες που παρουσιάζουν ενδιαφέρον.

Διάταξη γραφήματος

Ένα από τα πιο θεμελιώδη εργαλεία για την ερμηνεία των δεδομένων μοριακής αλληλεπίδρασης είναι η οπτικοποίηση των κόμβων και των ακμών ως ένα δισδιάστατο δίκτυο. Το Cytoscape υποστηρίζει μια ποικιλία αυτοματοποιημένων αλγορίθμων διάταξης δικτύου, όπως διάταξη ενσωματωμένη σε ελατήριο, ιεραρχική διάταξη και κυκλική διάταξη. Μεταξύ αυτών, το *spring embedder* είναι η πιο ευρέως χρησιμοποιούμενη μέθοδος για

τη διάταξη γενικών δισδιάστατων γραφημάτων. Μοντελοποιεί ένα μηχανικό σύστημα στο οποίο οι άκρες του γραφήματος αντιστοιχούν σε ελατήρια, δημιουργώντας μια ελκτική δύναμη μεταξύ των κόμβων που απέχουν πολύ μεταξύ τους και μια απωστική δύναμη μεταξύ των κόμβων που βρίσκονται κοντά.

Attribute-to-Visual Mapping

Ενώ η διάταξη καθορίζει τη θέση των κόμβων και των άκρων στο παράθυρο, μια αντιστοίχιση χαρακτηριστικών σε γράφημα επιτρέπει στα χαρακτηριστικά δεδομένων να ελέγχουν την εμφάνιση των σχετικών κόμβων και ακμών τους. Το Cytoscape υποστηρίζει μια μεγάλη ποικιλία οπτικών ιδιοτήτων, όπως το χρώμα, το σχήμα και το μέγεθος του κόμβου, χρώμα και πάχος περιγράμματος κόμβου, και χρώμα, πάχος και στυλ άκρων. Ένα χαρακτηριστικό δεδομένων αντιστοιχίζεται σε μια οπτική ιδιότητα χρησιμοποιώντας είτε πίνακα αναζήτησης είτε παρεμβολή, ανάλογα με το αν το χαρακτηριστικό έχει διακριτή τιμή ή συνεχές. Με την οπτική υπέρθεση μοριακών καταστάσεων στα μονοπάτια αλληλεπίδρασης που υποτίθεται ότι ρυθμίζουν αυτές τις καταστάσεις, οι αντιστοιχίσεις χαρακτηριστικών προς οπτική συνδέουν άμεσα τα παρατηρούμενα δεδομένα με ένα υποκείμενο μοντέλο.

Επιλογή γραφήματος και φιλτράρισμα

Για να μειωθεί η πολυπλοκότητα ενός μεγάλου δικτύου μοριακής αλληλεπίδρασης, είναι απαραίτητο να εμφανιστούν επιλεκτικά υποσύνολα

κόμβων και ακμών. Οι κόμβοι και οι ακμές μπορούν να επιλεγούν σύμφωνα με μια μεγάλη ποικιλία κριτηρίων, συμπεριλαμβανομένης της επιλογής με βάση το όνομα, από μια λίστα ονομάτων ή με βάση ένα χαρακτηριστικό. Πιο πολύπλοκα ερωτήματα επιλογής δικτύου υποστηρίζονται από μια εργαλειοθήκη φιλτραρίσματος που περιλαμβάνει ένα *Minimum Neighbors* φίλτρο, το οποίο επιλέγει κόμβους που έχουν έναν ελάχιστο αριθμό γειτόνων σε μια καθορισμένη απόσταση στο δίκτυο, ένα φίλτρο τοπικής απόστασης, το οποίο επιλέγει κόμβους σε μια καθορισμένη απόσταση από μια ομάδα προεπιλεγμένων κόμβων, ένα φίλτρο διαφορικής έκφρασης, το οποίο επιλέγει κόμβους σύμφωνα με τα σχετικά δεδομένα έκφρασης τους και ένα φίλτρο συνδυασμού, το οποίο επιλέγει κόμβους με αυθαίρετα ή/και συνδυασμούς άλλων φίλτρων.

Προσαρμογή Cytoscape μέσω προσθηκών

Τα *plug-in modules* παρέχουν ένα ισχυρό μέσο επέκτασης του Core για την εφαρμογή νέων αλγορίθμων, πρόσθετων αναλύσεων δικτύου ή/και βιολογικής σημασιολογίας. Οι προσθήκες έχουν πρόσβαση στο μοντέλο του βασικού δικτύου και μπορούν επίσης να ελέγχουν την εμφάνιση του δικτύου.

Παράδειγμα χρήσης του Cytoscape

Οι Sultana et al. (97) χρησιμοποίησαν υπολογιστικά μοντέλα για την ανάλυση δεδομένων single-cell RNA-seq ώστε να αποκαλύψουν την ετερογένεια του όγκου των ιστών μη μικροκυτταρικού καρκίνου του πνεύμονα (NSCLC). Συγκεκριμένα, προσδιόρισαν διαφορετικά εκφρασμένα γονίδια (DEGs), τις σχετικά κυτταρικά μονοπάτια τους και το δίκτυο PPI (protein-protein

interactions) για να εξετάσουν τα βασικά γονίδια (key genes ή KG), τα βασικά miRNA και τα βασικούς μεταγραφικούς παράγοντες (transcription factors, TF) για εξατομικευμένη διάγνωση και πρόγνωση του NSCLC και πραγματοποίησαν στη συνέχεια περαιτέρω ανάλυση για να επικυρώσουν το αποτέλεσμα.

Σε αυτή τη μελέτη, χρησιμοποιήθηκε η StringDB για την κατασκευή δικτύου PPI των DEGs. Το Cytoscape (έκδοση 3.7.1) χρησιμοποιήθηκε για τη διάκριση των δικτύων PPI μεταξύ των DEG. Προσδιόρισαν 12 KG χρησιμοποιώντας δύο τοπολογικές μεθόδους Betweenness and Stress στην προσθήκη του Cytoscape cytoHubba. Με βάση τη STRING, χρησιμοποιήθηκαν συνολικά 158 DEGs στο δίκτυο PPI, που αφορούσαν 156 κόμβους και 3.282 άκρες με μέσο βαθμό κόμβου (node degree) 42,2 και τιμή P εμπλουτισμού PPI $< 1,0e-16$. Στη συνέχεια εξήγαγαν 12 κοινά γονίδια από τα κορυφαία 15 γονίδια των δύο μεθόδων και τα θεώρησαν ως βασικά γονίδια (KG), όπως τα MS4A1, CCL5, GZMB, HBA1, TYROBP, CD74, LTB, EEF1A1, RPLP1, RPS18, GAPDH και ACTB. Επιπλέον, αξιολόγησαν τη σημασία των 12 KG μέσω ανάλυσης επιβίωσης, μοντέλου πρόβλεψης καρκίνου και ανάλυσης ρυθμιστικού δικτύου αλληλεπίδρασης.

Συνοψίζοντας, με τη χρήση του Cytoscape των δικτύων PPI, αλλά και οι αναλύσεις δικτύων GO, KEGG και PPI οδήγησαν στην ταυτοποίηση 12 βασικών γονιδίων. Ο λειτουργικός εμπλουτισμός και η ανάλυση μονοπατιών, η ανάλυση επιβίωσης, η ανάλυση ROC (receiver operating characteristics) και η ανάλυση ρυθμιστικού δικτύου τόνισαν τους κρίσιμους ρόλους των βασικών γονιδίων. Τα ευρήματά τους ενδέχεται να διαδραματίσουν σημαντικό ρόλο ως υποψήφιοι βιοδείκτες στη διάγνωση και την πρόγνωση του NSCLC.

StringDB

Η βάση δεδομένων STRING (98) (<https://string-db.org/>) συλλέγει συστηματικά και ενσωματώνει αλληλεπιδράσεις πρωτεΐνης-πρωτεΐνης, τόσο τις φυσικές αλληλεπιδράσεις όσο και τις λειτουργικές συσχετίσεις. Τα δεδομένα προέρχονται από διάφορες πηγές: αυτοματοποιημένη εξόρυξη (mining) κειμένου της επιστημονικής βιβλιογραφίας, προβλέψεις υπολογιστικής αλληλεπίδρασης από συνέκφραση, διατηρημένο γονιδιωματικό πλαίσιο, βάσεις δεδομένων πειραμάτων αλληλεπίδρασης και γνωστά σύμπλοκα/μονοπάτια από επιμελημένες πηγές. Όλες αυτές οι αλληλεπιδράσεις αξιολογούνται κριτικά, βαθμολογούνται και στη συνέχεια μεταφέρονται αυτόματα σε λιγότερο καλά μελετημένους οργανισμούς χρησιμοποιώντας πληροφορίες ιεραρχικής ορθολογίας. Η πρόσβαση στα δεδομένα είναι δυνατή μέσω της ιστοσελίδας, αλλά και μέσω προγραμματισμού και μέσω μαζικών λήψεων. Οι πιο πρόσφατες εξελίξεις στο STRING (έκδοση 12.0) είναι: (i) είναι πλέον δυνατή η δημιουργία, η περιήγηση και η ανάλυση ενός πλήρους δικτύου αλληλεπίδρασης για οποιοδήποτε νέο γονιδίωμα ενδιαφέροντος, υποβάλλοντας το συμπλήρωμα κωδικοποιημένων πρωτεϊνών του, (ii) το κανάλι συνέκφρασης χρησιμοποιεί τώρα μεταβλητούς αυτόματους κωδικοποιητές για την πρόβλεψη αλληλεπιδράσεων και καλύπτει δύο νέες πηγές, single-cell RNA-seq και πειραματικά δεδομένα πρωτεομικής και (iii) η εμπιστοσύνη σε κάθε πειραματικά προερχόμενη αλληλεπίδραση υπολογίζεται τώρα με βάση τη μέθοδο ανίχνευσης που χρησιμοποιείται και κοινοποιείται στον χρήστη στη διεπαφή ιστού. Επιπλέον, η STRING συνεχίζει να βελτιώνει τις εγκαταστάσεις της για ανάλυση λειτουργικού εμπλουτισμού, οι οποίες είναι πλέον πλήρως διαθέσιμες και για γονιδιώματα που υποβάλλονται από τους χρήστες.

Οι ερευνητές χρησιμοποιούν μετα-πόρους δικτύου πρωτεϊνών για μια μεγάλη ποικιλία σκοπών, που εμπίπτουν ευρέως σε τρεις κατηγορίες: (i) διευκόλυνση μεμονωμένων, μικρής κλίμακας μοριακών ανακαλύψεων, (ii) διευκόλυνση της ανάλυσης δεδομένων μεγάλης κλίμακας και (iii) συμβολή σε νέες μεθοδολογίες και ροές εργασίας (workflows). Όσον αφορά το STRING, έχει, για παράδειγμα, αποδειχθεί χρήσιμο στην ερμηνεία και τη μείωση των νεοαποκτηθέντων συνόλων δεδομένων γενετικού ελέγχου. Αυτά είναι μερικές φορές θορυβώδη και δυσκίνητα και το STRING μπορεί να χρησιμοποιηθεί για την απόσταση τέτοιων δεδομένων σε πιο διαχειρίσιμα σύνολα παρατηρήσεων και υποθέσεων.

Περιεχόμενο της βάσης δεδομένων

Ο σκοπός των συνδέσεων πρωτεΐνης-πρωτεΐνης στο STRING είναι αυτό της «λειτουργικής συσχέτισης». Οι πρωτεΐνες θεωρείται ότι συνδέονται όταν υπάρχουν στοιχεία που υποδηλώνουν μια εξελιγμένη, συγκεκριμένη λειτουργική συνεργασία μεταξύ των δύο. Αυτός ο ορισμός περιλαμβάνει πρωτεΐνες που συνδέονται φυσικά μεταξύ τους σε ένα πρωτεϊνικό σύμπλεγμα ή σε μια παροδική αλληλεπίδραση, αλλά και πρωτεΐνες που συνδέονται πιο έμμεσα: Μπορούν να εργαστούν προς έναν κοινό στόχο σε ένα μεταβολικό ή σηματοδοτικό μονοπάτι, μπορεί να ρυθμίζουν το ένα το άλλο μέσω ενδιάμεσων ή μπορεί να συμβάλλουν από κοινού σε μια κοινή κυτταρική δομή. Η ευαισθησία του τι συνιστά μια «κοινή λειτουργία» δεν ορίζεται επίσημα. Ωστόσο, δεν θα πρέπει να γίνει πολύ ευρέως κατανοητό, και λειτουργικά αντιστοιχεί χονδρικά στην ιδιαιτερότητα των «χαρτών» ή «διαγραμμάτων»

μονοπατιών σε βάσεις γνώσεων όπως το KEGG ή το Reactome. Θα πρέπει να σημειωθεί ότι ο ορισμός μιας λειτουργικής συσχέτισης μπορεί να περιλαμβάνει πρωτεΐνες που δρουν ανταγωνιστικά μεταξύ τους, αν και στην ίδια συνολική οδό.

Όλα τα στοιχεία συσχέτισης πρωτεΐνης-πρωτεΐνης στο STRING αξιολογούνται και ποσοτικοποιούνται και η αντιστοιχία του με τον παραπάνω ορισμό συγκρίνεται με την κοινή συμμετοχή σε χάρτες οδών KEGG (εξαιρουμένων των χαρτών που βασίζονται σε μεγάλο βαθμό στην ομολογία, όπως οι «μεταφορείς ABC»). Το αποτέλεσμα αυτής της συγκριτικής αξιολόγησης/βαθμονόμησης είναι μια «βαθμολογία εμπιστοσύνης» STRING για κάθε συσχέτιση. Οι βαθμολογίες εμπιστοσύνης κλιμακώνονται μεταξύ μηδέν και ενός και αντιστοιχούν στην εκτιμώμενη πιθανότητα να είναι αληθής μια δεδομένη συσχέτιση, δεδομένων των υποκείμενων στοιχείων. Οι βαθμολογίες εμπιστοσύνης υπολογίζονται πρώτα ξεχωριστά ανά τύπο αποδεικτικών στοιχείων και στη συνέχεια ενσωματώνονται σε μια τελική, «συνδυασμένη» βαθμολογία εμπιστοσύνης. Όλα τα στοιχεία που συλλέγονται για ένα δεδομένο ζεύγος πρωτεϊνών συμβάλλουν στη βαθμολογία, ανεξάρτητα από την ακριβή φύση αυτών των πρωτεϊνών όσον αφορά τις εναλλακτικές ισομορφές ματίσματος ή τις μετα-μεταφραστικές τροποποιήσεις. Αντίστοιχα, η μονάδα αλληλεπίδρασης στα δίκτυα STRING είναι ολόκληρος ο τρόπος κωδικοποίησης πρωτεΐνης, που αντιπροσωπεύεται από το πιο κανονικό πρωτεϊνικό προϊόν του.

Όλες οι βαθμολογίες εμπιστοσύνης στο STRING είναι προ-υπολογισμένες και διαθέσιμες δωρεάν για λήψη, με άδεια Creative Commons Attribution (CC BY 4.0). Οι διάφοροι τύποι αποδεικτικών στοιχείων

ομαδοποιούνται σε επτά διακριτά «κανάλια αποδεικτικών στοιχείων», με ξεχωριστές υποβαθμίσεις διαθέσιμες για κάθε κανάλι. Αυτά τα κανάλια μπορούν επίσης να προβληθούν μεμονωμένα στη διεπαφή χρήστη STRING, μαζί με τα υποκείμενα στοιχεία τους, και μπορούν να ενεργοποιηθούν ή να απενεργοποιηθούν ξεχωριστά όπως επιθυμεί ο κάθε χρήστης. Από τα επτά κανάλια, τα τρία πρώτα («γειτονιά», «συγχώνευση» και «συνύπαρξη») ασχολούνται με στοιχεία συσχέτισης που μπορούν να συλλεχθούν μόνο από τις αλληλουχίες του γονιδιώματος. Αυτά τα λεγόμενα κανάλια «γονιδιωματικού πλαισίου» βασίζονται στην ανίχνευση εξελικτικών περιορισμών που προκύπτουν από συνεργασίες λειτουργικών γονιδίων-γονιδίων και είναι καλύτερα εφαρμόσιμα σε προκαρυωτικά γονιδιώματα. Ένα άλλο κανάλι («συνέκφραση») ασχολείται με μετρήσεις λειτουργικής γονιδιωματικής (μεταγραφές ή πρωτεΐνες) σε μια πληθώρα συνθηκών, αναζητώντας στοιχεία κοινής ρύθμισης της έκφρασης. Στη συνέχεια, το κανάλι «πειράματα» ασχολείται με εργαστηριακά πειράματα που διεξήχθησαν με εκφρασμένο στόχο να αποκαλυφθούν στοιχεία συσχέτισης πρωτεΐνης-πρωτεΐνης. Εισάγονται από τα κύρια αποθετήρια βάσεων δεδομένων: BioGRID, DIP, PDB, καθώς και από την IntAct και τις βάσεις δεδομένων συνεργατών της στην κοινοπραξία IMEx. Τα δύο τελευταία κανάλια αποδεικτικών στοιχείων σχετίζονται με συσχετίσεις πρωτεΐνης-πρωτεΐνης που είναι ήδη γνωστές. Το κανάλι «βάσης δεδομένων» εισάγει καθιερωμένη γνώση («γνώση σχολικών βιβλίων») σχετικά με συμπλέγματα πρωτεϊνών, μονοπάτια και άλλες λειτουργικές συνδέσεις από ειδικούς πόρους γνώσης: KEGG, Reactome, MetaCyc, EBI Complex Portal και Gene Ontology Complexes. Τέλος, το κανάλι «textmining» είναι το αποτέλεσμα ανάλυσης άρθρων πλήρους κειμένου από το υποσύνολο ανοιχτής πρόσβασης

PMC (μέχρι τον Απρίλιο του 2022), περιλήψεων του PubMed (έως τον Αύγουστο 2022), καθώς και περιληπτικών κειμένων από το OMIM και το SGD περιγραφές εισόδου. Όλα αυτά τα κείμενα αναλύονται για συναναφορές ζευγών πρωτεϊνών και αξιολογούνται με βάση τις συχνότητες όλων των ξεχωριστών αναφορών των αντίστοιχων πρωτεϊνών. Εκτός από τις λειτουργικές συσχετίσεις, τα τρία τελευταία κανάλια (π.χ. πειράματα, βάση δεδομένων και εξόρυξη κειμένου) παρέχουν επίσης τις αλληλεπιδράσεις για το φυσικό υποδίκτυο του STRING. Ο υπολογισμός των βαθμολογιών εμπιστοσύνης για τα ζεύγη πρωτεϊνών σε αυτό το δίκτυο διαφέρει από αυτόν του δικτύου λειτουργικής συσχέτισης.

Όλες οι συσχετίσεις πρωτεϊνής-πρωτεϊνής που συναρμολογούνται για τη βάση δεδομένων STRING μεταφέρονται στη συνέχεια μεταξύ των οργανισμών, όπου ισχύει, με βάση τις σχέσεις ορθολογίας με την υπόθεση ότι τα ορθόλογα των σχετικών πρωτεϊνών συνδέονται επίσης («interologs»). Για αυτό, οι σχέσεις ιεραρχικής ορθολογίας, σε διάφορα επίπεδα ταξινομικής ανάλυσης, εισάγονται από τη βάση δεδομένων eggNOG. Μετά τη μεταφορά των interolog και την ολοκλήρωση της τελικής συνδυασμένης βαθμολογίας, τα προκύπτοντα δίκτυα συσχέτισης πρωτεϊνών μπορούν να έχουν πρόσβαση με διάφορους τρόπους. Πρώτον, ο διαδραστικός ιστότοπος του STRING επιτρέπει την περιήγηση και την αναζήτηση, συμπεριλαμβανομένης της επιθεώρησης στοιχείων μέσω αποκλειστικών χρηστών. Οι χρήστες μπορούν επίσης να υποβάλουν μεγαλύτερα ερωτήματα εκεί, επιτρέποντας την κατασκευή αποκλειστικών δικτύων και για στατιστικές αναζητήσεις για λειτουργικούς εμπλουτισμούς. Εκτός από τον ιστότοπο, οι επιστήμονες μπορούν να έχουν πρόσβαση στο STRING μέσω μιας αποκλειστικής προσθήκης Cytoscape, καθώς και μέσω

ενός πακέτου Bioconductor και μιας ειδικής διεπαφής προγραμματισμού εφαρμογών (REST).

Προοπτικές της STRING

Η τελευταία έκδοση της STRING καλύπτει μια φυλογενετικά ποικιλόμορφη συλλογή 12.535 γονιδιωμάτων υψηλής ποιότητας. Πέρα από αυτά, το σύστημα θα καταγράφει ποια γονιδιώματα υποβάλλονται συχνά από τους χρήστες — και στη συνέχεια θα δοθεί προτεραιότητα σε αυτά για συμπερίληψη σε επόμενες εκδόσεις. Επιπλέον, θα αναλυθούν τα αποτελέσματα μιας συνεχιζόμενης διαδικτυακής έρευνας χρηστών. Με αυτόν τον τρόπο, το STRING θα συνεχίσει να συγκεντρώνει τους πόρους του σε εκείνους τους τομείς που ενδιαφέρουν περισσότερο τους χρήστες του.

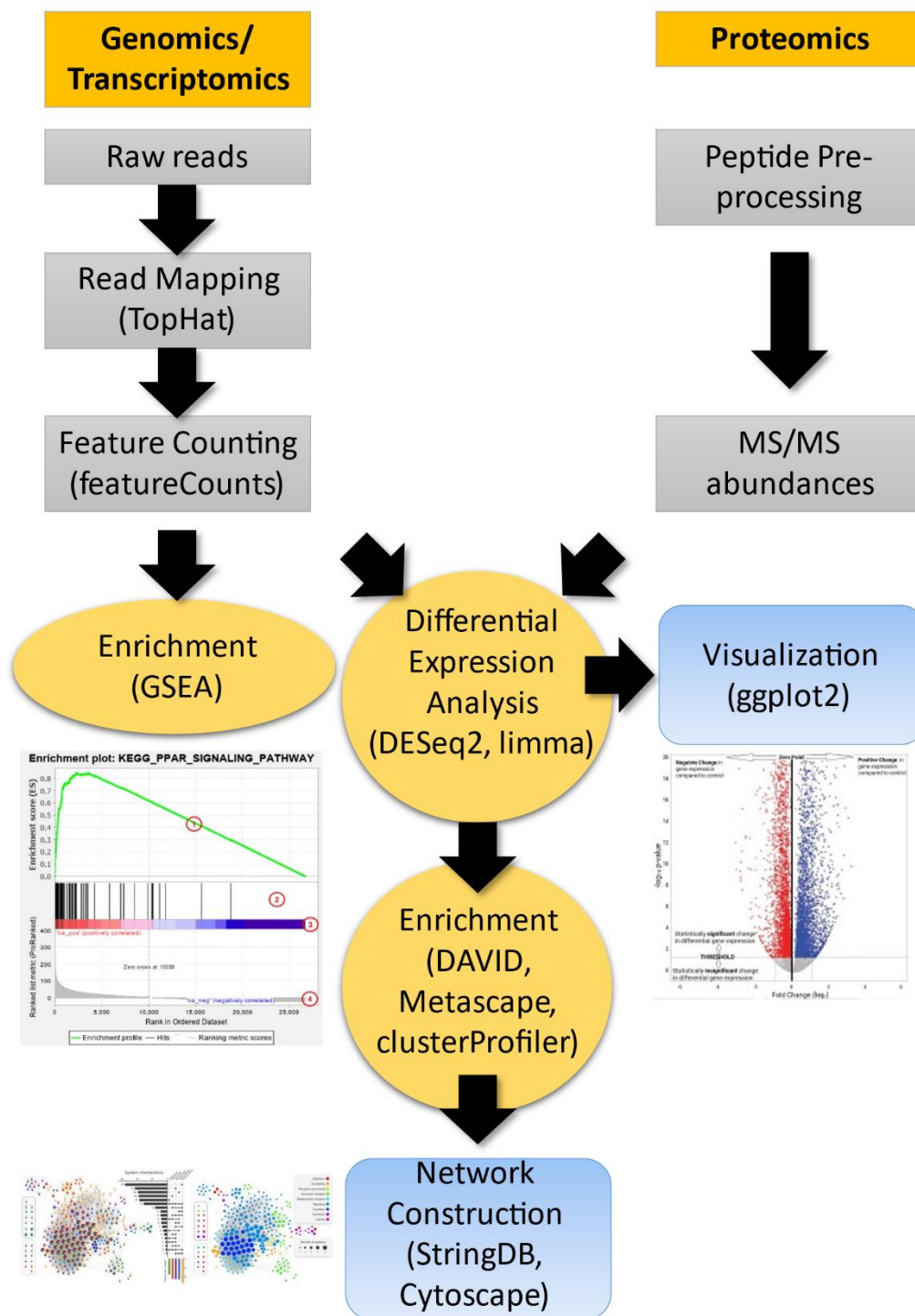
Παράδειγμα χρήσης της StringDB

Η μελέτη των Wang et al. (99) στοχεύει να διερευνήσει εάν ενδέχεται να υπάρχουν ορισμένα διαφορετικά εκφραζόμενα γονίδια που συνδέονται με την αντίσταση τόσο στην 1η όσο και στη 2η γενιά των EGFR-TKIs (EGFR-tyrosine kinase inhibitor ή υποδοχέα EGFR-κινάσης τυροσίνης), τη διήθηση των ανοσοκυττάρων και τον μετασχηματισμό μικροκυτταρικού καρκίνου του πνεύμονα (SCLC) μέσω βιοπληροφορικής ανάλυσης και δειγμάτων κλινικού ελέγχου.

Δημιουργήθηκε ένα δίκτυο για να εμφανίζει σχέσεις μεταξύ των πρωτεϊνών από τη βάση δεδομένων STRING. Αυτό χρησιμοποιήθηκε για να

κατασκευαστεί ένα δίκτυο PPI (protein-protein interactions ή αλληλεπιδράσεις πρωτεΐνης-πρωτεΐνης) για DEGs, χρησιμοποιήθηκε το Cytoscape για την οπτικοποίηση του δικτύου πρωτεϊνών και εξήχθησαν οι βασικές μονάδες στο δίκτυο PPI με το MCODE. Μέσω αυτών βρέθηκαν 11 πολύ συνδεδεμένα γονίδια (hub genes): RAB25, CD44, SPP1, ITGB4, ESRP1, FYN, CDH1, ERBB2, CLDN7, ZEB1 και MUC1. Η έκφραση των ZEB1 και SPP1 ρυθμίστηκε προς τα πάνω σε τρία προφίλ. Από την άλλη πλευρά, τα RAB25, ERBB2, ESRP1, ITGB4, MUC1, CLDN7 και CDH1 ρυθμίστηκαν προς τα κάτω σε τρία προφίλ. Το CD44 και το FYN ήταν αμφότερα μειωμένα στα προφίλ ανθεκτικότητας erlotinib- και gefitinib- και αυξήθηκαν στο προφίλ ανθεκτικότητας afatinib-.

Συμπερασματικά, η αντοχή στον EGFR-TKI είναι ένα κρίσιμο πρόβλημα μετά την έναρξη της στοχευμένης θεραπείας, καθώς αρκετοί μοριακοί μηχανισμοί συνδέονται με την αντίσταση. Η βιοπληροφορική ανάλυση και η κατασκευή των δικτύων PPI μέσω της StringDB οδήγησε στην εξαγωγή αρκετών συμπερασμάτων. Μέσω της μελέτης αυτής, οι Wang et al. πρότειναν ότι η ανοδική ρύθμιση του SPP1 μπορεί να προκαλέσει αντίσταση στον EGFR-TKI 1ης και 2ης γενιάς και να επηρεάσει την ανοσοδιήθηση του όγκου, με αποτέλεσμα τη μικρή πιθανότητα επιβίωσης.



Εικόνα 3: Παράδειγμα pipeline ανάλυσης δεδομένων πρωτεομικής και πρωτεογενομικής.

8. Συζήτηση

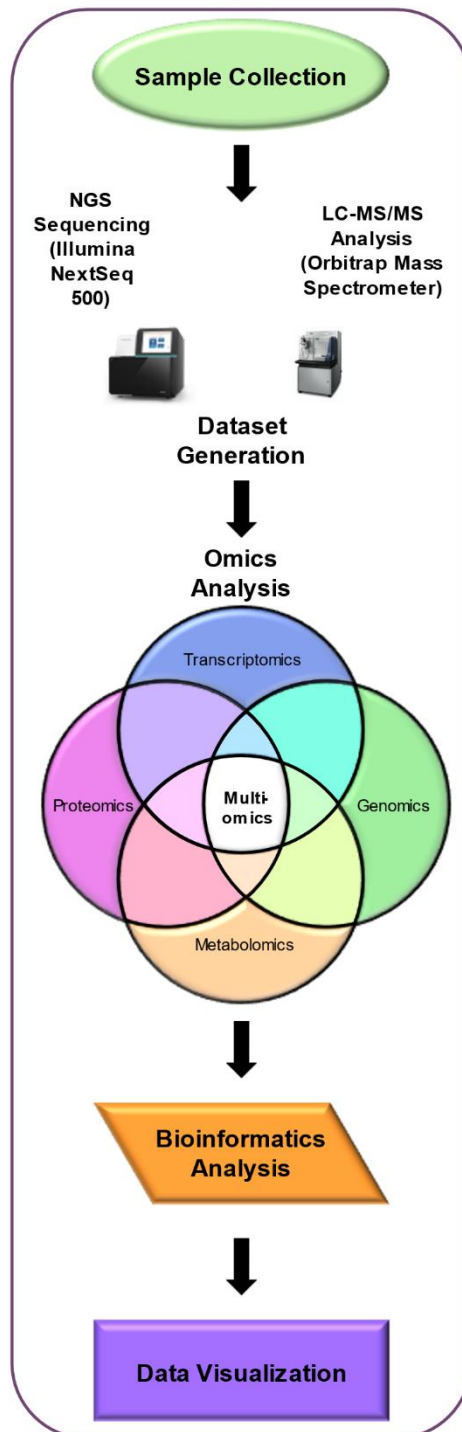
Τρέχουσες προσεγγίσεις

Με βάση τα εξαιρετικά πλεονεκτήματα των multi-omics στην ανακάλυψη των αιτιολογικών μηχανισμών νόσων, η προσανατολισμένη στην ασθένεια έρευνα απαιτεί ολοένα και περισσότερο τη συμμετοχή τεχνικών multi-omics και υπολογιστικών αλγορίθμων. Οι βιοδείκτες δεν μπορούν μόνο να εξερευνήσουν την παθογένεια σε μοριακό επίπεδο, αλλά έχουν επίσης μοναδικά πλεονεκτήματα στην εύστοχη και ευαίσθητη πρόβλεψη ασθενείων σε πρώιμα στάδια. Έτσι παρέχουν έγκαιρη προειδοποίηση, ανάλυση προγνωστικής αποτελεσματικότητας και ακριβή σταδιοποίηση και τυποποίηση των ασθενειών. Για τις κλινικές ασθένειες, η ανακάλυψη βιοδεικτών νόσου παίζει καθοδηγητικό ρόλο στη διάγνωση και την πρόγνωση των ασθενειών (100). Η διαλογή βιοδεικτών συνήθως απαιτεί τη χρήση μεθόδων omics υψηλής απόδοσης για τη μέτρηση κλινικών δειγμάτων μεγάλης κλίμακας, ακολουθούμενη από έλεγχο σε στατιστικά σημαντικά διαφορικά μόρια (γονίδια, πρωτεΐνες ή μεταβολίτες) και, τέλος, μια σειρά σύνθετων αναλύσεων βιοπληροφορικής για την ανάδειξη των βιοδεικτών-στόχων.

Αν και πολλοί βιοδείκτες που σχετίζονται με τον καρκίνο έχουν ταυτοποιηθεί με μεμονωμένες τεχνικές omics, τα multi-omics εμφανίζουν πλεονεκτήματα στην έρευνα για τον καρκίνο (101). Για παράδειγμα, οι Tran et al. (102), πραγματοποιώντας μια multi-omics έρευνα κατέληξαν στο συμπέρασμα ότι η μείωση του ALDH2 (μέλος της υπεροικογένειας του γονιδίου της αφυδρογονάσης της αλδεϋδης), η μη φυσιολογική μεθυλίωση του DNA και

την επακόλουθη απορρύθμιση σηματοδοτικών μονοπατιών βλαστοκυττάρων (stemness signaling pathway) είναι σχετικοί προγνωστικοί και θεραπευτικοί δείκτες στο αδενοκαρκίνωμα του πνεύμονα.

Επιπλέον, η ενσωμάτωση διαφόρων πληροφοριών omics αναμένεται να είναι πολύτιμη για την καθοδήγηση στοχευμένης θεραπείας του καρκίνου (103). Οι Li et al. (104) πραγματοποιώντας μελέτη multi-omics συμπέραναν ότι η στοχευμένη στα μιτοχόνδρια θεραπεία συνδυασμού φαρμάκων μπορεί να αναστείλει αποτελεσματικά την εμφάνιση και την ανάπτυξη όγκων, μέσω του επαναπρογραμματισμού του ενεργειακού μεταβολισμού και της αύξησης των κυττάρων του ανοσοποιητικού στους ιστούς του όγκου. Ένα παράδειγμα για τον σχεδιασμό μελέτης multi-omics δίνεται στην Εικόνα 3.



Εικόνα 4: Παράδειγμα ροής εργασιών σε μελέτη multi-omics.

Προκλήσεις και ευκαιρίες

Μια πρώιμη προσέγγιση στην ανάλυση omics ήταν η χωριστή ανάλυση διαφορετικών τύπων δεδομένων και στη συνέχεια συνδυασμός των αποτελεσμάτων για την απόκτηση ενός ολοκληρωμένου δικτύου μοριακών αλληλεπιδράσεων. Με σημαντικές εξελίξεις σε αυτόν τον τομέα, τα πλαίσια και οι μέθοδοι αλγοριθμικής μετα-ανάλυσης έχουν γίνει το κύριο μέσο μιας ολοκληρωμένης ανάλυσης δεδομένων multi-omics (105,106). Ωστόσο, η ανάλυση ενσωμάτωσης multi-omics αντιμετωπίζει επίσης ορισμένες προκλήσεις, όπως ο χειρισμός των ελλειπών τιμών (missing values), η ετερογένεια μεταξύ διαφορετικών omics, οι δυσκολίες στην ερμηνεία των μοντέλων multi-omics και τα προβλήματα στον σχολιασμό δεδομένων, την αποθήκευση και τους υπολογιστικούς πόρους (107).

Όσον αφορά τις ελλειπείς τιμές στα δεδομένα omics, η αντιμετώπιση τους είναι ένα αρκετά κρίσιμο ζήτημα στην ολοκληρωμένη ανάλυση δεδομένων multi-omics. Οι ελλειπείς τιμές μπορεί να οδηγήσουν σε μη ανιχνεύσιμα χαρακτηριστικά σε ορισμένα δείγματα, για παράδειγμα, σε πρωτεομικές και μεταβολομικές αναλύσεις. Οι ολοκληρωμένες μετρήσεις δεν είναι διαθέσιμες για ορισμένα δείγματα σε multi-omics, γεγονός που μπορεί να οδηγήσει σε περισσότερες τιμές που λείπουν. Για παράδειγμα, σε μελέτες κοόρτης, δεν μπορούν όλα τα άτομα να συλλέξουν όλους τους τύπους δεδομένων omics, επομένως ο αριθμός των ατόμων με πλήρεις εγγραφές είναι συνήθως πολύ μικρότερος από το συνολικό μέγεθος του δείγματος (108). Οι ελλειπείς τιμές μπορούν να επηρεάσουν την ανάλυση συσχέτισης μεταξύ διαφορετικών δεδομένων omics και να διακόψουν την ανάλυση ενσωμάτωσής τους. Ωστόσο, αυτές μπορούν να επιλυθούν χρησιμοποιώντας μια μέθοδο πολλαπλού

καταλογισμού (multiple imputation) σε ένα πλαίσιο ανάλυσης πολλαπλών παραγόντων (109). Σε ορισμένα δείγματα οι ελλειπείς τιμές μπορεί να οφείλονται σε διαδικασίες ποιοτικού ελέγχου, οι οποίες συνήθως μπορούν να επιλυθούν με τη χρήση της σταθμισμένης μεθόδου καταλογισμού k-πλησιέστερου γείτονα (k-nearest neighbor weighted imputation method) για τα δεδομένα που λείπουν από μπλοκ trans-omics (110), όπως το MOFA, το οποίο μπορεί να εκτελέσει αναγνώριση υποομάδας δειγμάτων, καταλογισμό δεδομένων και ανίχνευση ανώμαλου δείγματος για multi-omics δεδομένα (111). Ωστόσο, οι μέθοδοι καταλογισμού τιμών που λείπουν μπορούν να μειώσουν την αξιοπιστία του προκύπτοντος συνόλου δεδομένων (112) και να δημιουργήσουν δομές δεδομένων που παραβιάζουν τις υποθέσεις ανεξαρτησίας (independence assumptions) που απαιτούνται από πολλά στατιστικά πλαίσια. Ως εκ τούτου, σε ολοκληρωμένες αναλύσεις δεδομένων multi-omics, πραγματοποιείται ανάλυση ευαισθησίας σε καταλογισμένες ελλειπείς τιμές για να εκτιμηθεί ο αντίκτυπός τους στην επακόλουθη ανάλυση. Επιπλέον, για μη ειδικούς επεξεργαστές δεδομένων, ο ερευνητής συνήθως θα πρέπει να προσπαθεί να αποφύγει τα ενδοομαδικά ή διαομαδικά σφάλματα σε δείγματα από πειραματικούς χειρισμούς και να ελαχιστοποιήσει τους χρόνους δοκιμής των οργάνων omics για να αποφύγει την απώλεια τιμών λόγω παρτίδων, ειδικά για μελέτες omics μεγάλων κοορτών. Τέλος, οι ρυθμίσεις των παραμέτρων του οργάνου πρέπει να διορθωθούν εγκαίρως για να αποφευχθεί η απώλεια δεδομένων που προκαλείται από το θόρυβο του οργάνου.

Η ετερογένεια μεταξύ των διαφορετικών τεχνικών omics είναι επίσης ένας σημαντικός παράγοντας που επηρεάζει την ενσωμάτωση των δεδομένων. Διαφορετικές τεχνολογίες έχουν διαφορετικά επίπεδα ακρίβειας και η αναλογία

σήματος προς θόρυβο στις μετρήσεις συχνά επηρεάζει την ενσωμάτωση των δεδομένων. Για παράδειγμα, η πρωτεομική ευνοεί την ανίχνευση πρωτεϊνών που εκφράζονται σε αφθονία, η οποία σε μεγάλο βαθμό απουσιάζει στη μεταγραφομική. Λόγω αυτής της διαφοράς σήματος, η ανάλυση της σχέσης μεταξύ έκφρασης γονιδίου και πρωτεΐνης επηρεάζεται από τα πρωτεομικά δεδομένα (113). Έχουν αναπτυχθεί αλγόριθμοι για την εκτίμηση του βέλτιστου μεγέθους δείγματος που απαιτείται για διαφορετικά omics για την επίτευξη μιας δεδομένης στατιστικής ισχύος (113). Ωστόσο, αυτή η επεξεργασία μπορεί να οδηγήσει σε διαφορετικούς όγκους δειγματοληπτικών δεδομένων που συλλέγονται από διαφορετικούς τύπους omics, με αποτέλεσμα πολλές μέθοδοι στατιστικής ολοκλήρωσης multi-omics να μην είναι εφαρμόσιμες. Ως εκ τούτου, υπάρχει ανάγκη να βρεθούν νέες αναλυτικές προσεγγίσεις που επιτρέπουν ομοιογενείς δυνατότητες σε δεδομένα multi-omics. Επιπλέον, η ερμηνευσιμότητα των μοντέλων αυτών έχει εξαιρετικά πλεονεκτήματα για την κατανόηση της πολυπλοκότητας της ζωής σε μοριακό επίπεδο. Στις περισσότερες περιπτώσεις, ωστόσο, η ερμηνευσιμότητα αναφέρεται στην αναγνώριση των βιοδεικτών και των βιολογικών διεργασιών στις οποίες εμπλέκονται. Η ανάλυση λειτουργικού εμπλουτισμού χρησιμοποιείται συχνά στην ερμηνεία των δεδομένων multi-omics (114,115) αλλά αυτές οι μέθοδοι έχουν επίσης εγγενείς περιορισμούς στην παροχή μηχανιστικών γνώσεων στα μοριακά στρώματα. Τα διαγράμματα μονοπατιών έχουν επίσης χρησιμοποιηθεί για την ερμηνεία δεδομένων (116) αλλά αυτή η προσέγγιση δεν έχει την ευελιξία να αποκαλύψει νέες ρυθμιστικές σχέσεις. Επιπλέον, τα δίκτυα (όπως το Cytoscape) που συνδυάζουν δεδομένα multi-omics, αν και πιο γενικά, μπορούν εύκολα να γίνουν πολύ μεγάλα, πολύπλοκα και δύσκολα στην ερμηνεία (117).

Στο μέλλον, είναι απαραίτητο να αναπτυχθεί μια μέθοδος υπολογισμού που θα συνδυάζει τον μαθηματικό τύπο της προσαρμογής πολλαπλών επιπέδων με τη διαδραστική απεικόνιση.

Ο σχολιασμός και η αποθήκευση δεδομένων είναι επίσης μια πρόκληση για την ενσωμάτωση δεδομένων multi-omics. Ο ενσωματωμένος σχολιασμός δεδομένων συχνά απαιτεί λογισμικό που παρέχεται από παροχείς τεχνολογίας. Ωστόσο, αυτό το λογισμικό ενδέχεται να μην κοινοποιείται δημόσια. Παρά την αφθονία των διαθέσιμων στο κοινό μελετών και βάσεων δεδομένων, η ανάκτηση δεδομένων multi-omics για συγκεκριμένες βιολογικές οντότητες παραμένει μια πρόκληση. Είναι δύσκολο να συνδεθούν διαφορετικοί τύποι δεδομένων omics στο ίδιο ερευνητικό θέμα σε διάφορες πλατφόρμες δεδομένων. Συνήθως, διάφορα σύνολα δεδομένων omics φιλοξενούνται σε συγκεκριμένους τύπους αποθετηρίων, όπως το European Genome Archive (EGA, μια βάση δεδομένων αλληλουχίας) (118), το MetaboLights (μια βάση δεδομένων μεταβολιτών) (119), και το ProteomeXchange (μια βάση δεδομένων πρωτεϊνών) (120). Επομένως, οι υπερσύνδεσμοι είναι η μόνη επιλογή για τη σύνδεση της ίδιας μελέτης σε διαφορετικές βάσεις δεδομένων omics. Το Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής (European Bioinformatics Institute) έχει αναπτύξει ένα ενοποιημένο σύστημα ερωτημάτων που μπορεί να βρει γρήγορα σύνολα δεδομένων omics που ανταποκρίνονται στην ίδια λέξη-κλειδί σε όλα τα αποθετήρια (121). Στο μέλλον, τα πλαίσια περιγραφής μεταδεδομένων που ορίζουν παρτίδες, πειραματικές συνθήκες, προεπεξεργασία και αναπαραγωγή θα πρέπει να επιτρέπουν τη σύνδεση δειγμάτων στα αρχεία δεδομένων omics και να επιτρέπουν την αποτελεσματική επαναχρησιμοποίηση πολλαπλών δεδομένων omics.

Στο μέλλον, με την επέκταση των δειγμάτων και των χαρακτηριστικών, τα δεδομένα multi-omics θα γίνουν πιο περίπλοκα και η ανάπτυξη της ιατρικής ακριβείας απαιτεί την ενσωμάτωση συνόλων δεδομένων από μεγάλες ομάδες ασθενών (122). Η συμπερίληψη άλλων τύπων δεδομένων, όπως η ραδιομική και τα single-cell omics στην έρευνα multi-omics στοιχείων, θα αυξήσει περαιτέρω την πολυπλοκότητα της επεξεργασίας δεδομένων. Αξίζει να σημειωθεί ότι οι διακομιστές cloud έχουν αναπτυχθεί για να βοηθήσουν την αποθήκευση δεδομένων μεγάλης κλίμακας και την κοινή χρήση δεδομένων, εξοικονομώντας το κόστος λειτουργίας και συντήρησης της υπολογιστικής υποδομής (123). Επιπλέον, η επεξεργασία μεγάλων ποσοτήτων δεδομένων απαιτεί επίσης υψηλή υπολογιστική ισχύ και η υποδομή υπολογιστικού νέφους (cloud computing infrastructure) πρέπει να χρησιμοποιεί επαρκώς ισχυρές υπολογιστικές μονάδες για την επίτευξη γρήγορης επεξεργασίας δεδομένων multi-omics, όπως η συχνά χρησιμοποιούμενη μονάδα επεξεργασίας γραφικών (GPU) (124). Οι GPUs μπορούν επίσης να καλύψουν τη ζήτηση της επεξεργασίας deep learning, η οποία χρησιμοποιείται συχνά στην ανάλυση δεδομένων multi-omics (125). Συμπερασματικά, η ανάπτυξη αποδοτικού λογισμικού υπολογιστών και τεχνολογίας επεξεργασίας που βασίζεται σε υπηρεσίες cloud θα διευκολύνει την ενσωμάτωση δεδομένων multi-omics.

Single-cell omics: Το μέλλον των αναλύσεων omics

Τα single-cell multi-omics είναι η πιο ελπιδοφόρα κατεύθυνση ανάπτυξης στον τομέα της πολυομικής, η οποία μπορεί να συναγάγει ρυθμιστικά μοντέλα μεταξύ πολυεπίπεδων μορίων με καλύτερη ακρίβεια. Η

λογιστική για την ετερογένεια των τύπων κυττάρων θα διευκολύνει τη μοντελοποίηση πολύπλοκων βιολογικών διεργασιών (π.χ. όγκων (126)) και την κατανόηση της λειτουργίας σε εξαιρετικά διαφοροποιημένα όργανα (π.χ. εγκέφαλος (127)). Πολλαπλές μελέτες omics στο ίδιο κύτταρο ήταν επιτυχείς, όπως ο συνδυασμός RNA-sequencing, μεθυλίωσης DNA ή/και αφθονίας πρωτεϊνών σε κυτταρικό επίπεδο (128,129). Λαμβάνοντας υπόψη τα πολλαπλά δεδομένα omics που συλλέγονται στο ίδιο κύτταρο, καθώς έχουν τις ίδιες κυτταρικές πληροφορίες, μπορούν να αντιστοιχιστούν τα χαρακτηριστικά των διαφορετικών τρόπων (130) που είναι ευνοϊκά για την επακόλουθη ολοκληρωμένη ανάλυση, βελτιώνοντας έτσι το επίπεδο της έρευνας multi-omics.

Τα κύρια συστατικά μίας single-cell multi-omics ανάλυσης περιλαμβάνουν τεχνικές απομόνωσης μεμονωμένων κυττάρων, σήμανσης (barcoding) και αλληλούχισης για τη μέτρηση πολλαπλών τύπων μορίων από το ίδιο κύτταρο, καθώς και μια ολοκληρωμένη ανάλυση μορίων που μετρώνται σε κυτταρικό επίπεδο για τον εντοπισμό κυτταρικών τύπων και τις λειτουργίες τους που σχετίζονται με παθοφυσιολογικές διεργασίες με βάση τα μοριακά χαρακτηριστικά. Τα single-cell omics συνήθως επικεντρώνονται στα δεδομένα του γονιδιώματος-mRNA, της μεθυλίωσης του mRNA-DNA, της προσβασιμότητας mRNA-χρωματίνης και της πρωτεΐνης-mRNA (129).

Για την ολοκληρωμένη ανάλυση του γονιδιώματος ενός κυττάρου και των δεδομένων μεταγραφώματος, υπάρχουν διάφορες προσεγγίσεις για την επίτευξη αυτής της ανάλυσης, όπως η αλληλούχιση γονιδιώματος και μεταγραφώματος (G&T-seq) (131), αλληλούχιση gDNA-mRNA (DR-seq) (132), και ούτω καθεξής. Οι προσεγγίσεις για την αλληλούχιση ολόκληρου του

γονιδιώματος ενός κυττάρου (scWGS) περιέχουν πολλαπλή ενίσχυση μετατόπισης (multiple displacement amplification) (133), PicoPLEX (Rubicon Genomics PicoPLEX Kit) και ούτω καθεξής, ενώ οι μέθοδοι για scRNA-seq περιλαμβάνουν μηχανισμό μεταγωγής στο 5' άκρο του μετάγραφου RNA (Smart-seq) (134) και έκφραση κυττάρων με γραμμική ενίσχυση και αλληλούχιση (CEL-seq) (135).

Ωστόσο, παρόλο που τα δεδομένα single-cell multi-omics παρέχουν πλήθος βιολογικών πληροφοριών, η ολοκληρωμένη ανάλυσή τους αντιμετωπίζει τους ίδιους τύπους προκλήσεων όπως αυτών των υπόλοιπων multi-omics τεχνικών. Επιπλέον, υπάρχει έλλειψη τεχνολογίας ικανής να παράγει πολυτροπικά δεδομένα, με αποτέλεσμα περιεκτικές μελέτες που συχνά βασίζονται στην ενσωμάτωση τρόπων από διαφορετικά σύνολα δεδομένων. Σημειωτέον, η αντιστοίχιση cross-omics τύπων κυττάρων με τη χρήση δημοσίως διαθέσιμων δεδομένων από διαφορετικές μεθόδους μπορεί να αυξήσει σημαντικά τον αριθμό των μελετών single-cell multi-omics, όπως το MATCHER, μια συγκεντρωτική μέθοδος αφιερωμένη σε συγκεκριμένα single-cell δεδομένα (136). Η ενσωμάτωση δεδομένων single-cell multi-omics εμποδίζει επίσης τη χρήση μεθόδων επεξεργασίας δεδομένων παρτίδας λόγω συγκεκριμένων τεχνικών περιορισμών, όπως ο θόρυβος σε επίπεδο κυττάρου. Ως εκ τούτου, στο μέλλον, πρέπει επίσης να αναπτυχθούν σχετικές μέθοδοι για την ερμηνεία της βιολογικής σημασίας των δεδομένων που παράγονται από έναν μόνο τύπο κυττάρου.

Την τελευταία δεκαετία, οι single-cell τεχνολογίες έχουν αποκαλύψει την ετερογένεια του μικροπεριβάλλοντος ανοσοποιητικού-όγκου σε γονιδιωματικό, μεταγραφομικό και πρωτεομικό επίπεδο και έχουν βελτιώσει την κατανόησή

μας για τους μηχανισμούς ανάπτυξης όγκων. Οι τεχνολογίες αυτές έχουν επίσης χρησιμοποιηθεί για τον εντοπισμό πιθανών βιοδεικτών. Ωστόσο, σε αυτές τις προσεγγίσεις χάνονται χωρικές πληροφορίες σχετικά με το μικροπεριβάλλον ανοσοποιητικού-όγκου, όπως οι θέσεις των κυττάρων και οι αλληλεπιδράσεις κυττάρου-κυττάρου. Πρόσφατα, χωρικές τεχνολογίες multi-omics (spatial multi-omics) έχουν χρησιμοποιηθεί για τη μελέτη μεταγραφωμάτων, πρωτεϊνών και μεταβολισμών μικροπεριβαλλόντων ανοσοποιητικών όγκων σε διάφορους τύπους καρκίνου και τα δεδομένα που προέκυψαν από αυτές τις μεθόδους συνδυάστηκαν με ανοσοϊστοχημεία και πολυπαραμετρική ανάλυση για να δώσουν δείκτες εξέλιξης του καρκίνου (137).

Τα multi-omics σαν κομμάτι της βιολογικής και κλινικής έρευνας

Όπως αποδεικνύεται από την παγκόσμια προσπάθεια κατανόησης και εξεύρεσης τρόπων αντιμετώπισης του Covid-19, οι νέες τεχνολογίες και τεχνικές για την εξέταση της ανθρώπινης βιολογίας σε κυτταρικό επίπεδο είναι από τις πιο σημαντικές της εποχής μας. Οι αναλύσεις single-cell έχουν ανθίσει και περιλαμβάνουν μια σειρά πεδίων, συμπεριλαμβανομένης της γονιδιωματικής, της μεταγραφομικής, της πρωτεομικής, της επιγονιδιωματικής και της μεταβολομικής, καθένα από τα οποία παρέχει μοναδικές και κρίσιμες γνώσεις για την εσωτερική λειτουργία των ανθρώπινων κυττάρων.

Αυτές οι τεχνολογίες όταν χρησιμοποιούνται συνδυαστικά, έχουν αποδειχθεί ότι είναι ακόμη πιο ισχυρές. Τα multi-omics συνδυάζουν τις διάφορες προσεγγίσεις για την διερεύνηση σε επίπεδο κυττάρου, ώστε να αξιολογήσουν πληρέστερα την υγεία ή την ασθένεια ενός μεμονωμένου

κυττάρου. Είναι χρήσιμο να το σκεφτεί κάποιος, όπως όταν ένα άτομο αισθάνεται άρρωστο και επισκέπτεται έναν γιατρό. Για να καταλήξει στη σωστή διάγνωση και να καθορίσει τη σωστή θεραπεία, ο γιατρός μπορεί να εκτελέσει διάφορες εξετάσεις: από βασικές εξετάσεις, όπως θερμοκρασία, αρτηριακή πίεση και καρδιακό παλμό, έως αιμοληψίες, ακτινογραφίες και αξονική τομογραφία.

Μπορεί να φανταστεί κανείς ότι θα ήταν δυνατό να συγχωνευθούν όλα τα διαγνωστικά εργαλεία που έχει στη διάθεση ενός γιατρού σε ένα μόνο τεστ που έλεγε τα πάντα ταυτόχρονα. Θα εξοικονομούσε χρόνο, επιτρέποντας στον γιατρό να λάβει μια απάντηση για το τι ταλαιπωρεί τον ασθενή πιο γρήγορα, και χρήματα, καθώς θα ήταν μόνο μία εξέταση. Αυτή είναι μια μεγάλη αναλογία με αυτό που κάνουν τα multi-omics για τους ερευνητές και τις φαρμακευτικές εταιρείες, δίνοντας μια πιο γρήγορη, πληρέστερη εικόνα της απόκρισης των ασθενειών και των φαρμάκων, ειδικά για τις πιο προηγμένες κυτταρικές και γονιδιακές θεραπείες που υπόσχονται την εξάλειψη καταστροφικών ασθενειών όπως ο καρκίνος και άλλες γενετικές ανωμαλίες που έχουν αποφύγει τη θεραπεία μέχρι τώρα.

9. Βιβλιογραφία

1. Yang C, Li C, Wang Q, Chung D, Zhao H. Implications of pleiotropy: Challenges and opportunities for mining Big Data in biomedicine. *Front Genet* [Internet]. 2015 Jun 30 [cited 2024 Jun 2];6(JUN):140159. Available from: www.frontiersin.org
2. Viceconti M, Hunter P, Hose R. Big data, big knowledge: big data for personalized healthcare. *IEEE J Biomed Health Inform* [Internet]. 2015 Jul 1 [cited 2024 Jun 2];19(4):1209–15. Available from: <https://pubmed.ncbi.nlm.nih.gov/26218867/>
3. Kankanhalli A, Hahn J, Tan S, Gao G. Big data and analytics in healthcare: Introduction to the special section. *Information Systems Frontiers* [Internet]. 2016 Apr 1 [cited 2024 Jun 2];18(2):233–5. Available from: <https://link.springer.com/article/10.1007/s10796-016-9641-2>
4. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst* [Internet]. 2014 Feb 7 [cited 2024 Jun 2];2(1). Available from: <https://pubmed.ncbi.nlm.nih.gov/25825667/>
5. Wu PY, Cheng CW, Kaddi CD, Venugopalan J, Hoffman R, Wang MD. - Omic and Electronic Health Record Big Data Analytics for Precision Medicine. *IEEE Trans Biomed Eng* [Internet]. 2017 Feb 1 [cited 2024 Jun 2];64(2):263–73. Available from: <https://pubmed.ncbi.nlm.nih.gov/27740470/>
6. Wang Y, Kung LA, Wang WYC, Cegielski CG. An integrated big data analytics-enabled transformation model: Application to health care. *Information & Management*. 2018 Jan 1;55(1):64–79.

7. Andreu-Perez J, Poon CCY, Merrifield RD, Wong STC, Yang GZ. Big data for health. *IEEE J Biomed Health Inform* [Internet]. 2015 Jul 1 [cited 2024 Jun 2];19(4):1193–208. Available from: <https://pubmed.ncbi.nlm.nih.gov/26173222/>
8. Archenaa J, Anita EAM. A Survey of Big Data Analytics in Healthcare and Government. *Procedia Comput Sci*. 2015 Jan 1;50:408–13.
9. Jignesh Chowdary G, Punns NS, Sonbhadra SK, Agarwal S. Face Mask Detection Using Transfer Learning of InceptionV3. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* [Internet]. 2020 [cited 2024 Jun 2];12581 LNCS:81–90. Available from: https://link.springer.com/chapter/10.1007/978-3-030-66665-1_6
10. He KY, Ge D, He MM. Big Data Analytics for Genomic Medicine. *Int J Mol Sci* [Internet]. 2017 Feb 15 [cited 2024 Jun 2];18(2). Available from: <https://pubmed.ncbi.nlm.nih.gov/28212287/>
11. Tan SSL, Gao G, Koch S. Big Data and Analytics in Healthcare. *Methods Inf Med* [Internet]. 2015 [cited 2024 Jun 2];54(6):546–7. Available from: <https://pubmed.ncbi.nlm.nih.gov/26577624/>
12. Dinov ID, Heavner B, Tang M, Glusman G, Chard K, Darcy M, et al. Predictive Big Data Analytics: A Study of Parkinson's Disease Using Large, Complex, Heterogeneous, Incongruent, Multi-Source and Incomplete Observations. *PLoS One* [Internet]. 2016 Aug 1 [cited 2024 Jun 18];11(8). Available from: <https://pubmed.ncbi.nlm.nih.gov/27494614/>

13. Archenaa J, Anita EAM. A Survey of Big Data Analytics in Healthcare and Government. *Procedia Comput Sci*. 2015 Jan 1;50:408–13.
14. Ristevski B, Chen M. Big Data Analytics in Medicine and Healthcare. *J Integr Bioinform* [Internet]. 2018 May 10 [cited 2024 Jun 2];15(3). Available from: <https://pubmed.ncbi.nlm.nih.gov/29746254/>
15. Benke K, Benke G. Artificial Intelligence and Big Data in Public Health. *Int J Environ Res Public Health* [Internet]. 2018 Dec 1 [cited 2024 Jun 2];15(12). Available from: <https://pubmed.ncbi.nlm.nih.gov/30544648/>
16. Rubin R. A Precision Medicine Approach to Clinical Trials. *JAMA* [Internet]. 2016 Nov 15 [cited 2024 Jun 2];316(19):1953–5. Available from: <https://pubmed.ncbi.nlm.nih.gov/27760248/>
17. Rubin R. Precision medicine: the future or simply politics? *JAMA* [Internet]. 2015 Mar 17 [cited 2024 Jun 2];313(11):1089–91. Available from: <https://pubmed.ncbi.nlm.nih.gov/25781428/>
18. Ashley EA. Towards precision medicine. *Nat Rev Genet* [Internet]. 2016 Sep 1 [cited 2024 Jun 2];17(9):507–22. Available from: <https://pubmed.ncbi.nlm.nih.gov/27528417/>
19. Russell SJ, Norvig P, Davis E, Edwards DD, Forsyth D, Hay NJ, et al. Artificial intelligence: a modern approach. Pearson Education Limited [Internet]. 2016 [cited 2024 Jun 2];3:1151. Available from: <https://thuvienso.hoasen.edu.vn/handle/123456789/8967>
20. Nilsson NJ. Review of Nils Nilsson PRINCIPLES OF ARTIFICIAL INTELLIGENCE. Springer Science & Business Media [Internet]. 1982

- [cited 2024 Jun 2];3. Available from: https://books.google.com/books/about/Principles_of_Artificial_Intelligence.html?id=4JwN5DTpcqkC
21. Mesko B. The role of artificial intelligence in precision medicine. *Expert Rev Precis Med Drug Dev* [Internet]. 2017 [cited 2024 Jun 2];2(5):239–41. Available from: <https://www.tandfonline.com/doi/abs/10.1080/23808993.2017.1380516>
 22. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* [Internet]. 2016 Dec 13 [cited 2024 Jun 2];316(22):2402–10. Available from: <https://pubmed.ncbi.nlm.nih.gov/27898976/>
 23. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. 2021 [cited 2024 Jun 2]; Available from: <https://link.springer.com/10.1007/978-1-0716-1418-1>
 24. Hastie T, Friedman J, Tibshirani R. *The Elements of Statistical Learning*. 2001 [cited 2024 Jun 2]; Available from: <http://link.springer.com/10.1007/978-0-387-21606-5>
 25. Difference Between Supervised, Unsupervised, & Reinforcement Learning | NVIDIA Blog [Internet]. [cited 2024 Jun 2]. Available from: <https://blogs.nvidia.com/blog/supervised-unsupervised-learning/>
 26. Ting Sim JZ, Fong QW, Huang W, Tan CH. Machine learning in medicine: what clinicians should know. *Singapore Med J* [Internet]. 2023 Feb 1

- [cited 2024 Jun 2];64(2):91–7. Available from: <https://pubmed.ncbi.nlm.nih.gov/34005847/>
27. Li Y, Wu X, Yang P, Jiang G, Luo Y. Machine Learning for Lung Cancer Diagnosis, Treatment, and Prognosis. *Genomics Proteomics Bioinformatics* [Internet]. 2022 Oct 1 [cited 2024 Jun 2];20(5):850–66. Available from: <https://pubmed.ncbi.nlm.nih.gov/36462630/>
 28. Thai AA, Solomon BJ, Sequist L V., Gainor JF, Heist RS. Lung cancer. *Lancet* [Internet]. 2021 Aug 7 [cited 2024 Jun 2];398(10299):535–54. Available from: <https://pubmed.ncbi.nlm.nih.gov/34273294/>
 29. Svoboda E. Artificial intelligence is improving the detection of lung cancer. *Nature* [Internet]. 2020 Nov 1 [cited 2024 Jun 2];587(7834):S20–2. Available from: <https://pubmed.ncbi.nlm.nih.gov/33208974/>
 30. Ling S, Hu Z, Yang Z, Yang F, Li Y, Lin P, et al. Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution. *Proc Natl Acad Sci U S A* [Internet]. 2015 Nov 24 [cited 2024 Jun 2];112(47):E6496–505. Available from: <https://pubmed.ncbi.nlm.nih.gov/26561581/>
 31. Hudson TJ, Anderson W, Aretz A, Barker AD, Bell C, Bernabé RR, et al. International network of cancer genome projects. *Nature*. 2010 Apr 15;464(7291):993–8.
 32. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* [Internet]. 2013 Oct 1 [cited 2024 Jun 2];45(10):1113–20. Available from: <https://pubmed.ncbi.nlm.nih.gov/24071849/>

33. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging [Internet]. 2013 Dec [cited 2024 Jun 2];26(6):1045–57. Available from: <https://pubmed.ncbi.nlm.nih.gov/23884657/>
34. Pavlopoulou A, Spandidos DA, Michalopoulos I. Human cancer databases (review). Oncol Rep [Internet]. 2015 Jan 1 [cited 2024 Jun 2];33(1):3–18. Available from: <https://pubmed.ncbi.nlm.nih.gov/25369839/>
35. Bioinformatics: Sequence and Genome Analysis, Second Edition [Internet]. [cited 2024 Jun 2]. Available from: https://www.cshlpress.com/default.tpl?cart=171735500789975852&fromlink=T&linkaction=full&linksortby=oop_title&--eqSKUdataarg=466
36. Algorithms in Bioinformatics: Theory and Implementation - Paul A. Gagniuc - Google Books [Internet]. [cited 2024 Jun 2]. Available from: [https://books.google.gr/books?hl=en&lr=&id=y1I5EAAAQBAJ&oi=fnd&pg=PP18&dq=2.%09Gagniuc,+Paul+A.+\(2021\).+Algorithms+in+bioinformatics:+theory+and+implementation+\(1st+ed.\).+Hoboken,+NJ:+John+Wiley+%26+Sons.+pp.+1%E2%80%93528.+3.%09Ng+PC,+Henikoff+S.+Predicting+deleterious+amino+acid+substitutions&ots=cZFFYutyzh&sig=shiyH-fm8f6vjAXMRZQkD1L8rDo&redir_esc=y#v=onepage&q&f=false](https://books.google.gr/books?hl=en&lr=&id=y1I5EAAAQBAJ&oi=fnd&pg=PP18&dq=2.%09Gagniuc,+Paul+A.+(2021).+Algorithms+in+bioinformatics:+theory+and+implementation+(1st+ed.).+Hoboken,+NJ:+John+Wiley+%26+Sons.+pp.+1%E2%80%93528.+3.%09Ng+PC,+Henikoff+S.+Predicting+deleterious+amino+acid+substitutions&ots=cZFFYutyzh&sig=shiyH-fm8f6vjAXMRZQkD1L8rDo&redir_esc=y#v=onepage&q&f=false)

37. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Res* [Internet]. 2001 May [cited 2024 Jun 2];11(5):863–74. Available from: <https://pubmed.ncbi.nlm.nih.gov/11337480/>
38. Polyanovsky VO, Roytberg MA, Tumanyan VG. Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences. *Algorithms Mol Biol* [Internet]. 2011 Oct 27 [cited 2024 Jun 2];6(1). Available from: <https://pubmed.ncbi.nlm.nih.gov/22032267/>
39. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* [Internet]. 1990 Oct 25 [cited 2024 Jun 2];18(20):6097–100. Available from: <https://pubmed.ncbi.nlm.nih.gov/2172928/>
40. Sequence Alignment/Map Format Specification. 2023 [cited 2024 Jun 2]; Available from: <https://github.com/samtools/hts-specs>.
41. Shi W, Liao Y. Rsubread/Subread Users Guide. 2023;
42. Tsiamis V, Ienasescu HI, Gabrielaitis D, Palmblad M, Schwämmle V, Ison J. One Thousand and One Software for Proteomics: Tales of the Toolmakers of Science. *J Proteome Res* [Internet]. 2019 Oct 4 [cited 2024 Jun 16];18(10):3580–5. Available from: <https://pubmed.ncbi.nlm.nih.gov/31429284/>
43. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* [Internet]. 2008 Dec [cited 2024 Jun 16];26(12):1367–72. Available from: <https://pubmed.ncbi.nlm.nih.gov/19029910/>

44. Tyanova S, Temu T, Cox J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc* [Internet]. 2016 Dec 1 [cited 2024 Jun 16];11(12):2301–19. Available from: <https://pubmed.ncbi.nlm.nih.gov/27809316/>
45. Koenig T, Menze BH, Kirchner M, Monigatti F, Parker KC, Patterson T, et al. Robust prediction of the MASCOT score for an improved quality assessment in mass spectrometric proteomics. *J Proteome Res* [Internet]. 2008 Sep [cited 2024 Jun 16];7(9):3708–17. Available from: <https://pubmed.ncbi.nlm.nih.gov/18707158/>
46. Bruderer R, Bernhardt OM, Gandhi T, Miladinović SM, Cheng LY, Messner S, et al. Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol Cell Proteomics* [Internet]. 2015 May 1 [cited 2024 Jun 16];14(5):1400–10. Available from: <https://pubmed.ncbi.nlm.nih.gov/25724911/>
47. Zhang J, Xin L, Shan B, Chen W, Xie M, Yuen D, et al. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol Cell Proteomics* [Internet]. 2012 Apr [cited 2024 Jun 16];11(4). Available from: <https://pubmed.ncbi.nlm.nih.gov/22186715/>
48. Kou Q, Wu S, Liu X. Systematic Evaluation of Protein Sequence Filtering Algorithms for Proteoform Identification Using Top-Down Mass Spectrometry. *Proteomics* [Internet]. 2018 Feb 1 [cited 2024 Jun 16];18(3–4). Available from: <https://pubmed.ncbi.nlm.nih.gov/29327814/>

49. Cheon DH, Yang EG, Lee C, Lee JE. Low-Molecular-Weight Plasma Proteome Analysis Using Top-Down Mass Spectrometry. *Methods Mol Biol* [Internet]. 2017 [cited 2024 Jun 16];1619:103–17. Available from: <https://pubmed.ncbi.nlm.nih.gov/28674880/>
50. Nayak BK, Hazra A. How to choose the right statistical test? *Indian J Ophthalmol* [Internet]. 2011 Mar [cited 2024 Jun 2];59(2):85–6. Available from: <https://pubmed.ncbi.nlm.nih.gov/21350275/>
51. Jaykaran. How to select appropriate statistical test? *J Pharm Negat Results* [Internet]. 2010 Jun 7 [cited 2024 Jun 2];1(2):61–3. Available from: <https://www.pnrjournal.com/index.php/home/article/view/18>
52. Mishra P, Mayilvaganan S, Agarwal A. Statistical Methods in Endocrine Surgery Journal Club. *Statistical Methods in Endocrine Surgery Journal Club World Journal of Endocrine Surgery*. 7(1):21–3.
53. Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, Kitano H, et al. Visualization of omics data for systems biology. *Nat Methods* [Internet]. 2010 [cited 2024 Jun 2];7(3 Suppl):S56–68. Available from: <https://pubmed.ncbi.nlm.nih.gov/20195258/>
54. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol* [Internet]. 2012 [cited 2024 Jun 2];8(2). Available from: <https://pubmed.ncbi.nlm.nih.gov/22383865/>
55. Creixell P, Reimand J, Haider S, Wu G, Shibata T, Vazquez M, et al. Pathway and network analysis of cancer genomes. *Nat Methods*

- [Internet]. 2015 Jun 30 [cited 2024 Jun 2];12(7):615–21. Available from: <https://pubmed.ncbi.nlm.nih.gov/26125594/>
56. Khatri P, Drăghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* [Internet]. 2005 Sep 15 [cited 2024 Jun 2];21(18):3587–95. Available from: <https://pubmed.ncbi.nlm.nih.gov/15994189/>
 57. Chagoyen M, Pazos F. MBRole: enrichment analysis of metabolomic data. *Bioinformatics* [Internet]. 2011 Mar [cited 2024 Jun 2];27(5):730–1. Available from: <https://pubmed.ncbi.nlm.nih.gov/21208985/>
 58. Glazko G V., Emmert-Streib F. Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets. *Bioinformatics* [Internet]. 2009 Sep [cited 2024 Jun 2];25(18):2348–54. Available from: <https://pubmed.ncbi.nlm.nih.gov/19574285/>
 59. Biochemical Pathways: An Atlas of Biochemistry and Molecular Biology - Google Books [Internet]. [cited 2024 Jun 2]. Available from: [https://books.google.gr/books?hl=en&lr=&id=CXHihxlyWfEC&oi=fnd&pg=PR10&dq=Michal,+G.,+%26+Schomburg,+D.+\(Eds.\).+\(2012\).+Biochemical+pathways:+an+atlas+of+biochemistry+and+molecular+biology.+John+Wiley+%26+Sons&ots=5nLa79U4GJ&sig=JyuDzXrVi7TO30IVul87buWVKds&redir_esc=y#v=onepage&q=Michal%2C%20G.%2C%20%26%20Schomburg%2C%20D.%20\(Eds.\).%20\(2012\).%20Biochemical%20pathways%3A%20an%20atlas%20of%20biochemistry%20and%20molecular%20biology.%20John%20Wiley%20%26%20Sons&f=false](https://books.google.gr/books?hl=en&lr=&id=CXHihxlyWfEC&oi=fnd&pg=PR10&dq=Michal,+G.,+%26+Schomburg,+D.+(Eds.).+(2012).+Biochemical+pathways:+an+atlas+of+biochemistry+and+molecular+biology.+John+Wiley+%26+Sons&ots=5nLa79U4GJ&sig=JyuDzXrVi7TO30IVul87buWVKds&redir_esc=y#v=onepage&q=Michal%2C%20G.%2C%20%26%20Schomburg%2C%20D.%20(Eds.).%20(2012).%20Biochemical%20pathways%3A%20an%20atlas%20of%20biochemistry%20and%20molecular%20biology.%20John%20Wiley%20%26%20Sons&f=false)

60. Nishizuka Y. The role of protein kinase C in cell surface signal transduction and tumour promotion. *Nature* 1984 308:5961 [Internet]. 1984 [cited 2024 Jun 2];308(5961):693–8. Available from: <https://www.nature.com/articles/308693a0>

61. Levine M, Davidson EH. Gene regulatory networks for development. *Proceedings of the National Academy of Sciences* [Internet]. 2005 Apr 5 [cited 2024 Jun 2];102(14):4936–42. Available from: <https://www.pnas.org/doi/abs/10.1073/pnas.0408031102>

62. Bader GD, Cary MP, Sander C. Pathguide: a Pathway Resource List. *Nucleic Acids Res* [Internet]. 2006 Jan 1 [cited 2024 Jun 2];34(suppl_1):D504–6. Available from: <https://dx.doi.org/10.1093/nar/gkj126>

63. Saraiya P, North C, Duca K. Visualizing Biological Pathways: Requirements Analysis, Systems Evaluation and Research Agenda. <http://dx.doi.org/101057/palgrave.ivs9500102> [Internet]. 2005 Sep 1 [cited 2024 Jun 2];4(3):191–205. Available from: <https://journals.sagepub.com/doi/abs/10.1057/palgrave.ivs.9500102>

64. Suderman M, Hallett M. Tools for visually exploring biological networks. *Bioinformatics* [Internet]. 2007 Oct 15 [cited 2024 Jun 2];23(20):2651–9. Available from: <https://pubmed.ncbi.nlm.nih.gov/17720984/>

65. Pavlopoulos GA, Wegener AL, Schneider R. A survey of visualization tools for biological network analysis. *BioData Mining* 2008 1:1 [Internet]. 2008 Nov 28 [cited 2024 Jun 2];1(1):1–11. Available from: <https://link.springer.com/articles/10.1186/1756-0381-1-12>

66. Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, Kitano H, et al. Visualization of omics data for systems biology. *Nat Methods* [Internet]. 2010 [cited 2024 Jun 2];7(3 Suppl):S56–68. Available from: <https://pubmed.ncbi.nlm.nih.gov/20195258/>
67. Bateman A, Martin MJ, Orchard S, Magrane M, Ahmad S, Alpi E, et al. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* [Internet]. 2023 Jan 6 [cited 2024 Jun 2];51(D1):D523–31. Available from: <https://dx.doi.org/10.1093/nar/gkac1052>
68. Perez-Riverol Y, Bai J, Bandla C, García-Seisdedos D, Hewapathirana S, Kamatchinathan S, et al. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res* [Internet]. 2022 Jan 7 [cited 2024 Jun 2];50(D1):D543–52. Available from: <https://dx.doi.org/10.1093/nar/gkab1038>
69. Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, et al. The PeptideAtlas project. *Nucleic Acids Res* [Internet]. 2006 Jan 1 [cited 2024 Jun 2];34(suppl_1):D655–8. Available from: <https://dx.doi.org/10.1093/nar/gkj040>
70. Schaab C, Geiger T, Stoehr G, Cox J, Mann M. Analysis of high accuracy, quantitative proteomics data in the MaxQB database. *Mol Cell Proteomics* [Internet]. 2012 [cited 2024 Jun 2];11(3). Available from: <https://pubmed.ncbi.nlm.nih.gov/22301388/>
71. Omenn GS, Lane L, Overall CM, Pineau C, Packer NH, Cristea IM, et al. The 2022 Report on the Human Proteome from the HUPO Human Proteome Project. *J Proteome Res* [Internet]. 2023 Apr 7 [cited 2024 Jun

- 2];22(4):1024–42. Available from:
<https://pubmed.ncbi.nlm.nih.gov/36318223/>
72. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* [Internet]. 2021 Mar 29 [cited 2024 Jun 6];372. Available from: <https://www.bmj.com/content/372/bmj.n71>
73. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* [Internet]. 2009 [cited 2024 Jun 3];25(9):1105–11. Available from:
<https://pubmed.ncbi.nlm.nih.gov/19289445/>
74. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* [Internet]. 2013 Apr 25 [cited 2024 Jun 3];14(4). Available from:
<https://pubmed.ncbi.nlm.nih.gov/23618408/>
75. Sharpnack MF, Ranbaduge N, Srivastava A, Cerciello F, Codreanu SG, Liebler DC, et al. Proteogenomic Analysis of Surgically Resected Lung Adenocarcinoma. *J Thorac Oncol* [Internet]. 2018 Oct 1 [cited 2024 Jun 3];13(10):1519–29. Available from:
<https://pubmed.ncbi.nlm.nih.gov/30017829/>
76. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* [Internet]. 2014 Apr 1 [cited 2024 Jun 3];30(7):923–30. Available from: <https://dx.doi.org/10.1093/bioinformatics/btt656>

77. Djureinovic D, Hallström BM, Horie M, Margareta Mattsson JS, Fleur L La, Fagerberg L, et al. Profiling cancer testis antigens in non-small-cell lung cancer. JCI Insight [Internet]. 2016 Jul 7 [cited 2024 Jun 3];1(10). Available from: <https://pubmed.ncbi.nlm.nih.gov/27699219/>
78. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol [Internet]. 2014 Dec 5 [cited 2024 Jun 3];15(12):1–21. Available from: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8>
79. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res [Internet]. 2015 Jan 6 [cited 2024 Jun 3];43(7):e47. Available from: <https://pubmed.ncbi.nlm.nih.gov/25605792/>
80. Yohannes E, Kazanjian AA, Lindsay ME, Fujii DT, Ieronimakis N, Chow GE, et al. The human tubal lavage proteome reveals biological processes that may govern the pathology of hydrosalpinx. Scientific Reports 2019 9:1 [Internet]. 2019 Jun 20 [cited 2024 Jun 3];9(1):1–13. Available from: <https://www.nature.com/articles/s41598-019-44962-1>
81. Narayanan M, Walker WT, Kenia P, Robson EA, Lucas JS, Callaghan CO, et al. S24 Proteomics reveals distinct drivers of cystic fibrosis lung function and quality of life outcomes: post-hoc analysis of the AZTEC-CF study. Thorax [Internet]. 2023 Nov 1 [cited 2024 Jun 3];78(Suppl 4):A21–2. Available from: https://thorax.bmj.com/content/78/Suppl_4/A21.2

82. Lin MH, Wu PS, Wong TH, Lin IY, Lin J, Cox J, et al. Benchmarking differential expression, imputation and quantification methods for proteomics data. *Brief Bioinform* [Internet]. 2022 May 13 [cited 2024 Jun 3];23(3):1–13. Available from: <https://dx.doi.org/10.1093/bib/bbac138>
83. Langlely SR, Mayr M. Comparative analysis of statistical methods used for detecting differential expression in label-free mass spectrometry proteomics. *J Proteomics* [Internet]. 2015 Mar 9 [cited 2024 Jun 3];129:83–92. Available from: <https://pubmed.ncbi.nlm.nih.gov/26193490/>
84. Ren J, Yang Y, Li C, Xie L, Hu R, Qin X, et al. A Novel Prognostic Model of Early-Stage Lung Adenocarcinoma Integrating Methylation and Immune Biomarkers. *Front Genet* [Internet]. 2021 Jan 21 [cited 2024 Jun 3];11. Available from: <https://pubmed.ncbi.nlm.nih.gov/33552145/>
85. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* [Internet]. 2005 Oct 25 [cited 2024 Jun 3];102(43):15545–50. Available from: <https://pubmed.ncbi.nlm.nih.gov/16199517/>
86. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* [Internet]. 2003 Jul 1 [cited 2024 Jun 3];34(3):267–73. Available from: <https://pubmed.ncbi.nlm.nih.gov/12808457/>

87. Zhang Y, Fu F, Zhang Q, Li L, Liu H, Deng C, et al. Evolutionary proteogenomic landscape from pre-invasive to invasive lung adenocarcinoma. *Cell Rep Med* [Internet]. 2024 Jan 16 [cited 2024 Jun 3];5(1). Available from: <https://pubmed.ncbi.nlm.nih.gov/38183982/>
88. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* [Internet]. 2003 [cited 2024 Jun 3];4(5). Available from: <https://pubmed.ncbi.nlm.nih.gov/12734009/>
89. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res* [Internet]. 2022 Jul 5 [cited 2024 Jun 3];50(W1):W216–21. Available from: <https://pubmed.ncbi.nlm.nih.gov/35325185/>
90. Wu J, Hao Z, Ma C, Li P, Dang L, Sun S. Comparative proteogenomics profiling of non-small and small lung carcinoma cell lines using mass spectrometry. *PeerJ* [Internet]. 2020 [cited 2024 Jun 3];8(4). Available from: <https://pubmed.ncbi.nlm.nih.gov/32351780/>
91. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* [Internet]. 2012 May 1 [cited 2024 Jun 3];16(5):284–7. Available from: <https://pubmed.ncbi.nlm.nih.gov/22455463/>
92. Wu L, Zhong Y, Wu D, Xu P, Ruan X, Yan J, et al. Immunomodulatory Factor TIM3 of Cytolytic Active Genes Affected the Survival and Prognosis of Lung Adenocarcinoma Patients by Multi-Omics Analysis.

- Biomedicines [Internet]. 2022 Sep 1 [cited 2024 Jun 3];10(9). Available from: <https://pubmed.ncbi.nlm.nih.gov/36140350/>
93. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun [Internet]. 2019 Dec 1 [cited 2024 Jun 3];10(1). Available from: <https://pubmed.ncbi.nlm.nih.gov/30944313/>
94. Qi Q, Huang W, Zhang H, Zhang B, Sun X, Ma J, et al. Bioinformatic analysis of PLOD family member expression and prognostic value in non-small cell lung cancer. Transl Cancer Res [Internet]. 2021 Jun 1 [cited 2024 Jun 3];10(6):2707–24. Available from: <https://pubmed.ncbi.nlm.nih.gov/35116582/>
95. ggplot2: Elegant Graphics for Data Analysis (3e) [Internet]. [cited 2024 Jun 3]. Available from: <https://ggplot2-book.org/>
96. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res [Internet]. 2003 Nov [cited 2024 Jun 3];13(11):2498–504. Available from: <https://pubmed.ncbi.nlm.nih.gov/14597658/>
97. Sultana A, Alam MS, Liu X, Sharma R, Singla RK, Gundamaraju R, et al. Single-cell RNA-seq analysis to identify potential biomarkers for diagnosis, and prognosis of non-small cell lung cancer by using comprehensive bioinformatics approaches. Transl Oncol [Internet]. 2023

- Jan 1 [cited 2024 Jun 3];27. Available from: <https://pubmed.ncbi.nlm.nih.gov/36401966/>
98. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* [Internet]. 2023 Jan 6 [cited 2024 Jun 3];51(D1):D638–46. Available from: <https://pubmed.ncbi.nlm.nih.gov/36370105/>
 99. Wang Z, Zhang L, Xu W, Li J, Liu Y, Zeng X, et al. The Multi-Omics Analysis of Key Genes Regulating EGFR-TKI Resistance, Immune Infiltration, SCLC Transformation in EGFR-Mutant NSCLC. *J Inflamm Res* [Internet]. 2022 [cited 2024 Jun 3];15:649–67. Available from: <https://pubmed.ncbi.nlm.nih.gov/35140497/>
 100. Atkinson AJ, Colburn WA, DeGruttola VG, DeMets DL, Downing GJ, Hoth DF, et al. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther* [Internet]. 2001 Jan 1 [cited 2024 Jun 3];69(3):89–95. Available from: <https://pubmed.ncbi.nlm.nih.gov/11240971/>
 101. Chen C, Wang J, Pan D, Wang X, Xu Y, Yan J, et al. Applications of multi-omics analysis in human diseases. *MedComm (Beijing)* [Internet]. 2023 Aug 1 [cited 2024 Jun 3];4(4). Available from: <https://pubmed.ncbi.nlm.nih.gov/37533767/>
 102. Tran TO, Vo TH, Lam LHT, Le NQK. ALDH2 as a potential stem cell-related biomarker in lung adenocarcinoma: Comprehensive multi-omics

- analysis. *Comput Struct Biotechnol J* [Internet]. 2023 Jan 1 [cited 2024 Jun 3];21:1921–9. Available from: <https://pubmed.ncbi.nlm.nih.gov/36936815/>
103. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, et al. Proteogenomic characterization of human colon and rectal cancer. *Nature* [Internet]. 2014 Sep 18 [cited 2024 Jun 3];513(7518):382–7. Available from: <https://pubmed.ncbi.nlm.nih.gov/25043054/>
 104. Li C, Zhang Y, Xia Q, Hao B, Hong Y, Yue L, et al. Multi-omics analysis revealed the mitochondrial-targeted drug combination to suppress the development of lung cancer. *J Cancer Res Clin Oncol* [Internet]. 2023 Dec 1 [cited 2024 Jun 3];149(19):17159–74. Available from: <https://pubmed.ncbi.nlm.nih.gov/37783930/>
 105. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* [Internet]. 2016 Jan 20 [cited 2024 Jun 3];17 Suppl 2(Suppl 2). Available from: <https://pubmed.ncbi.nlm.nih.gov/26821531/>
 106. Pierre-Jean M, Deleuze JF, Le Floch E, Mauger F. Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration. *Brief Bioinform* [Internet]. 2020 Nov 1 [cited 2024 Jun 3];21(6):2011–30. Available from: <https://pubmed.ncbi.nlm.nih.gov/31792509/>
 107. Tarazona S, Arzalluz-Luque A, Conesa A. Undisclosed, unmet and neglected challenges in multi-omics studies. *Nature Computational*

- Science 2021 1:6 [Internet]. 2021 Jun 21 [cited 2024 Jun 3];1(6):395–402. Available from: <https://www.nature.com/articles/s43588-021-00086-z>
108. Conesa A, Beck S. Making multi-omics data accessible to researchers. Sci Data [Internet]. 2019 Dec 1 [cited 2024 Jun 3];6(1). Available from: <https://pubmed.ncbi.nlm.nih.gov/31672978/>
 109. Voillet V, Besse P, Liaubet L, San Cristobal M, González I. Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework. BMC Bioinformatics [Internet]. 2016 Oct 3 [cited 2024 Jun 3];17(1). Available from: <https://pubmed.ncbi.nlm.nih.gov/27716030/>
 110. Dong X, Lin L, Zhang R, Zhao Y, Christiani DC, Wei Y, et al. TOBML: trans-omics block missing data imputation using a k-nearest neighbor weighted approach. Bioinformatics [Internet]. 2019 Apr 15 [cited 2024 Jun 3];35(8):1278–83. Available from: <https://pubmed.ncbi.nlm.nih.gov/30202885/>
 111. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol [Internet]. 2018 Jun [cited 2024 Jun 3];14(6). Available from: <https://pubmed.ncbi.nlm.nih.gov/29925568/>
 112. Messer K, Vaida F, Hogan C. Robust analysis of biomarker data with informative missingness using a two-stage hypothesis test in an HIV treatment interruption trial: AIEDRP AIN503/ACTG A5217. Contemp Clin

- Trials [Internet]. 2006 Dec [cited 2024 Jun 3];27(6):506–17. Available from: <https://pubmed.ncbi.nlm.nih.gov/16962381/>
113. Tarazona S, Balzano-Nogueira L, Gómez-Cabrero D, Schmidt A, Imhof A, Hankemeier T, et al. Harmonization of quality metrics and power calculation in multi-omic studies. *Nat Commun* [Internet]. 2020 Dec 1 [cited 2024 Jun 3];11(1). Available from: <https://pubmed.ncbi.nlm.nih.gov/32555183/>
114. Canzler S, Hackermüller J. multiGSEA: a GSEA-based pathway enrichment analysis for multi-omics data. *BMC Bioinformatics* [Internet]. 2020 Dec 1 [cited 2024 Jun 3];21(1). Available from: <https://pubmed.ncbi.nlm.nih.gov/33287694/>
115. Arneson D, Bhattacharya A, Shu L, Mäkinen VP, Yang X. Mergeomics: a web server for identifying pathological pathways, networks, and key regulators via multidimensional data integration. *BMC Genomics* [Internet]. 2016 Sep 9 [cited 2024 Jun 3];17(1). Available from: <https://pubmed.ncbi.nlm.nih.gov/27612452/>
116. Long Y, Lu M, Cheng T, Zhan X, Zhan X. Multiomics-Based Signaling Pathway Network Alterations in Human Non-functional Pituitary Adenomas. *Front Endocrinol (Lausanne)* [Internet]. 2019 Dec 17 [cited 2024 Jun 3];10. Available from: <https://pubmed.ncbi.nlm.nih.gov/31920959/>
117. Su G, Morris JH, Demchak B, Bader GD. Biological network exploration with Cytoscape 3. *Curr Protoc Bioinformatics* [Internet]. 2014 [cited 2024

- Jun 3];47:8.13.1-8.13.24. Available from:
<https://pubmed.ncbi.nlm.nih.gov/25199793/>
118. Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, Ur-Rehman S, et al. The European Genome-phenome Archive of human data consented for biomedical research. *Nature Genetics* 2015 47:7 [Internet]. 2015 Jun 26 [cited 2024 Jun 3];47(7):692–5. Available from:
<https://www.nature.com/articles/ng.3312>
 119. Haug K, Cochrane K, Nainala VC, Williams M, Chang J, Jayaseelan KV, et al. MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic Acids Res* [Internet]. 2020 Jan 8 [cited 2024 Jun 3];48(D1):D440–4. Available from:
<https://dx.doi.org/10.1093/nar/gkz1019>
 120. Deutsch EW, Csordas A, Sun Z, Jarnuczak A, Perez-Riverol Y, Ternent T, et al. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res* [Internet]. 2017 Jan 4 [cited 2024 Jun 3];45(D1):D1100–6. Available from:
<https://dx.doi.org/10.1093/nar/gkw936>
 121. Perez-Riverol Y, Bai M, Da Veiga Leprevost F, Squizzato S, Park YM, Haug K, et al. Discovering and linking public omics data sets using the Omics Discovery Index. *Nature Biotechnology* 2017 35:5 [Internet]. 2017 May 9 [cited 2024 Jun 3];35(5):406–9. Available from:
<https://www.nature.com/articles/nbt.3790>
 122. Angione C. Human Systems Biology and Metabolic Modelling: A Review-From Disease Metabolism to Precision Medicine. *Biomed Res Int*

- [Internet]. 2019 [cited 2024 Jun 3];2019. Available from: <https://pubmed.ncbi.nlm.nih.gov/31281846/>
123. Oh M, Park S, Kim S, Chae H. Machine learning-based analysis of multi-omics data on the cloud for investigating gene regulations. *Brief Bioinform* [Internet]. 2021 Jan 1 [cited 2024 Jun 3];22(1):66–76. Available from: <https://pubmed.ncbi.nlm.nih.gov/32227074/>
124. Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Med Genomics* [Internet]. 2015 Jun 27 [cited 2024 Jun 3];8(1). Available from: <https://pubmed.ncbi.nlm.nih.gov/26112054/>
125. Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. *Genome Med* [Internet]. 2019 Nov 19 [cited 2024 Jun 3];11(1). Available from: <https://pubmed.ncbi.nlm.nih.gov/31744524/>
126. Fan J, Slowikowski K, Zhang F. Single-cell transcriptomics in cancer: computational challenges and opportunities. *Exp Mol Med* [Internet]. 2020 Sep 1 [cited 2024 Jun 3];52(9):1452–65. Available from: <https://pubmed.ncbi.nlm.nih.gov/32929226/>
127. Armand EJ, Li J, Xie F, Luo C, Mukamel EA. Single-Cell Sequencing of Brain Cell Transcriptomes and Epigenomes. *Neuron* [Internet]. 2021 Jan 6 [cited 2024 Jun 3];109(1):11–26. Available from: <https://pubmed.ncbi.nlm.nih.gov/33412093/>
128. Zhu C, Preissl S, Ren B. Single-cell multimodal omics: the power of many. *Nat Methods* [Internet]. 2020 Jan 1 [cited 2024 Jun 3];17(1):11–4. Available from: <https://pubmed.ncbi.nlm.nih.gov/31907462/>

129. Lee J, Hyeon DY, Hwang D. Single-cell multiomics: technologies and data analysis methods. *Exp Mol Med* [Internet]. 2020 Sep 1 [cited 2024 Jun 3];52(9):1428–42. Available from: <https://pubmed.ncbi.nlm.nih.gov/32929225/>
130. Forcato M, Romano O, Bicciato S. Computational methods for the integrative analysis of single-cell data. *Brief Bioinform* [Internet]. 2021 Jan 1 [cited 2024 Jun 3];22(1):20–9. Available from: <https://pubmed.ncbi.nlm.nih.gov/32363378/>
131. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat Methods* [Internet]. 2015 May 28 [cited 2024 Jun 3];12(6):519–22. Available from: <https://pubmed.ncbi.nlm.nih.gov/25915121/>
132. Dey SS, Kester L, Spanjaard B, Bienko M, Van Oudenaarden A. Integrated genome and transcriptome sequencing of the same cell. *Nat Biotechnol* [Internet]. 2015 [cited 2024 Jun 3];33(3):285–9. Available from: <https://pubmed.ncbi.nlm.nih.gov/25599178/>
133. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, et al. Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A* [Internet]. 2002 Apr 16 [cited 2024 Jun 3];99(8):5261–6. Available from: <https://pubmed.ncbi.nlm.nih.gov/11959976/>
134. Ramsköld D, Luo S, Wang YC, Li R, Deng Q, Faridani OR, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* [Internet]. 2012 [cited 2024 Jun

- 3];30(8):777–82. Available from:
<https://pubmed.ncbi.nlm.nih.gov/22820318/>
135. Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep* [Internet]. 2012 Sep 27 [cited 2024 Jun 3];2(3):666–73. Available from:
<https://pubmed.ncbi.nlm.nih.gov/22939981/>
136. Welch JD, Hartemink AJ, Prins JF. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biol* [Internet]. 2017 Jul 24 [cited 2024 Jun 3];18(1). Available from: <https://pubmed.ncbi.nlm.nih.gov/28738873/>
137. Hsieh WC, Budiarto BR, Wang YF, Lin CY, Gwo MC, So DK, et al. Spatial multi-omics analyses of the tumor immune microenvironment. *Journal of Biomedical Science* 2022 29:1 [Internet]. 2022 Nov 14 [cited 2024 Jun 3];29(1):1–17. Available from:
<https://link.springer.com/articles/10.1186/s12929-022-00879-y>

10. Σύνοψη

Τα *"Big Data"* («Μεγάλα Δεδομένα») στην υγειονομική περίθαλψη και την ιατρική αναφέρονται σε διάφορα μεγάλα και πολύπλοκα δεδομένα. Εξαιτίας της πολυπλοκότητας αυτής, είναι δύσκολη η διαχείρισή τους, καθώς και η ανάλυσή τους από τα υπάρχοντα διαθέσιμα λογισμικά. Η διαχείριση τους περιλαμβάνει αρχικά την ενοποίηση ετερογενών δεδομένων, στη συνέχεια τον ποιοτικό τους έλεγχο, την ανάλυσή τους, τη μοντελοποίησή τους και, τέλος, την ερμηνεία τους και την επικύρωσή τους. Η καινοτομία που προσφέρει η ανάλυση των Big Data επιτρέπει την εξαγωγή ολιστικών συμπερασμάτων η οποία οφείλεται στον τεράστιο διαθέσιμο όγκο τους.

Για την παραγωγή δεδομένων omics απαιτούνται μηχανήματα υψηλής απόδοσης (high-throughput) που μπορούν να διαχειριστούν την ανάγνωση και καταγραφή της μοριακής πληροφορίας, δηλαδή από κάθε δείγμα να παράξουν ένα αρχείο ακατέργαστων δεδομένων (.raw data file). Πιο συγκεκριμένα, για την παραγωγή δεδομένων γενομικής, επιγενομικής και μεταγραφομικής χρησιμοποιούνται αλληλουχητές (sequencers), οι οποίοι διαβάζουν τα νουκλεοτίδια από τις εισαγόμενες αλληλουχίες DNA/RNA. Αντίστοιχα για την πρωτεομική, δεν υπάρχουν αλληλουχητές που να διαβάζουν την ακολουθία αμινοξέων της εισαγόμενης πρωτεΐνης. Οι αναλύσεις αυτές γίνονται σε μηχανήματα φασματοσκοπίας μάζας (mass spectrometers).

Μετά τη συλλογή των δεδομένων λαμβάνει χώρα η προεπεξεργασία τους. Αυτή συνήθως αφορά τον χειρισμό «θορυβωδών» δεδομένων, ακραίων τιμών, τιμών που λείπουν, μετασχηματισμού δεδομένων και κανονικοποίηση. Έτσι η ποιότητα και τα αποτελέσματα της ανάλυσης των Big Data είναι δυνατόν να βελτιωθούν και να οδηγήσουν στην ανακάλυψη νέας γνώσης.

Ο συνδυασμός των Big Data με την συνεχώς αναπτυσσόμενη χρήση της τεχνητής νοημοσύνης, δύναται να παράξει αποτελέσματα που δημιουργούν νέους ορίζοντες στον ιατρικό κλάδο. Η τεχνητή νοημοσύνη (artificial intelligence, AI) είναι ένας γενικός όρος για μια μηχανή (ή διαδικασία) που ανταποκρίνεται σε περιβαλλοντικά ερεθίσματα (ή νέα δεδομένα) και στη συνέχεια τροποποιεί τη λειτουργία της για να μεγιστοποιήσει έναν δείκτη απόδοσης. Στην πράξη, η διαδικασία εκμάθησης υλοποιείται με χρήση μαθηματικών, στατιστικών, λογικής και προγραμματισμού.

Η πρόσφατη ανάπτυξη τεχνολογιών απεικόνισης και αλληλούχισης επιτρέπει συστηματική πρόοδο στην κλινική μελέτη του καρκίνου του πνεύμονα. Ο ανθρώπινος νους είναι περιορισμένος στον αποτελεσματικό χειρισμό και την πλήρη χρήση της συσσώρευσης τέτοιων τεράστιων ποσοτήτων δεδομένων. Οι προσεγγίσεις που βασίζονται στο machine learning διαδραματίζουν κρίσιμο ρόλο στην ενσωμάτωση και ανάλυση αυτών των μεγάλων και πολύπλοκων συνόλων δεδομένων, τα οποία έχουν χαρακτηρίσει εκτενώς τον καρκίνο του πνεύμονα μέσω της χρήσης διαφορετικών προοπτικών.

Ο καρκίνος του πνεύμονα είναι ένας από τους πιο συχνά διαγνωσμένους καρκίνους και η κύρια αιτία θανάτων από καρκίνο παγκοσμίως. Περίπου 2,2 εκατομμύρια νέοι ασθενείς διαγιγνώσκονται με καρκίνο του πνεύμονα κάθε χρόνο και το 75% από αυτούς πεθαίνουν μέσα σε πέντε χρόνια από τη διάγνωση. Τις τελευταίες δεκαετίες, η συνεχής εξέλιξη των τεχνολογιών στην έρευνα για τον καρκίνο έχει συμβάλει σε πολλά μεγάλα συλλογικά έργα, τα οποία έχουν δημιουργήσει πολυάριθμες βάσεις δεδομένων κλινικής, ιατρικής απεικόνισης και αλληλουχίας. Ειδικότερα, οι τρέχουσες μελέτες σχετικά με τις

αναλύσεις omics, όπως η γονιδιωματική (genomics), η μεταγραφομική (transcriptomics), η πρωτεομική (proteomics) και η μεταβολομική (metabolomics), έχουν επεκτείνει τα εργαλεία και τις δυνατότητές μας για έρευνα.

Η παρούσα Διπλωματική Εργασία (ΔΕ) εστιάζει στην περιγραφή των εργαλείων που καθιστούν δυνατές τις ανωτέρω ερευνητικές προσεγγίσεις. Παρέχουν, δηλαδή, στον ερευνητή τη δυνατότητα να προσεγγίσει ένα ιατρικό ή βιολογικό ερώτημα με τρόπο ολιστικό που λαμβάνει υπόψη του πολλές παραμέτρους του υπό μελέτη συστήματος. Σε αυτό τα πλαίσιο διακρίνεται μια εξέλιξη στην έρευνα, δηλαδή από την πεποίθηση ότι κάθε ασθένεια μπορεί να αποδοθεί σε ένα ελαττωματικό γονίδιο, στην αντίληψη της κάθε ασθένειας ως αποτέλεσμα της αλληλεπίδρασης του περιβάλλοντος με το γονιδίωμα.

Στη βιοπληροφορική, η ευθυγράμμιση αλληλουχίας είναι ένας τρόπος διάταξης των αλληλουχιών DNA, RNA ή πρωτεΐνης για τον εντοπισμό περιοχών ομοιότητας που μπορεί να είναι συνέπεια λειτουργικών, δομικών ή εξελικτικών σχέσεων μεταξύ των αλληλουχιών. Οι ευθυγραμμισμένες αλληλουχίες νουκλεοτιδίων ή αμινοξέων αντιπροσωπεύονται τυπικά ως σειρές μέσα σε ένα πίνακα (matrix). Εισάγονται κενά μεταξύ των υπολειμμάτων έτσι ώστε πανομοιότυποι ή παρόμοιοι χαρακτήρες να ευθυγραμμίζονται σε διαδοχικές στήλες. Μερικά παραδείγματα γνωστών προγραμμάτων ευθυγράμμισης είναι τα: Bowtie, TopHat και HISAT.

Τα εργαλεία καταμέτρησης χαρακτηριστικών (feature counting) είναι προγράμματα περίληψης ανάγνωσης γενικής χρήσης που μετρούν χαρτογραφημένες αναγνώσεις για γονιδιωματικά χαρακτηριστικά όπως γονίδια,

εξόνια, υποκινητής, σώματα γονιδίων, γονιδιωματικά δοχεία και χρωμοσωμικές θέσεις. Μπορεί να χρησιμοποιηθούν για τη μέτρηση αναγνώσεων RNA-seq και γονιδιωματικού DNA-seq. Αυτά λαμβάνουν ως είσοδο αρχεία ευθυγράμμισης SAM ή BAM και ένα αρχείο σχολιασμού που περιλαμβάνει χρωμοσωμικές συντεταγμένες χαρακτηριστικών. Εξάγουν αριθμούς αναγνώσεων που έχουν εκχωρηθεί σε χαρακτηριστικά (ή μετα-χαρακτηριστικά). Εξάγουν επίσης πληροφορίες στατιστικών στοιχείων για τα συνολικά αποτελέσματα σύνοψης, συμπεριλαμβανομένου του αριθμού των αναγνώσεων που ανατέθηκαν με επιτυχία και του αριθμού των αναγνώσεων που απέτυχαν να εκχωρηθούν για διάφορους λόγους. Το αρχείο σχολιασμού θα πρέπει να είναι είτε σε μορφή GTF είτε σε μορφή απλοποιημένου σχολιασμού (SAF). Τα πιο συνηθισμένα εργαλεία feature counting είναι το featureCounts και htseq-count.

Οι ερευνητές που ασχολούνται με την πρωτεομική αντιμετωπίζουν μια ενδιαφέρουσα πρόκληση: πώς να επιλέξουν ανάμεσα στα δεκάδες pipelines επεξεργασίας και ανάλυσης δεδομένων που είναι διαθέσιμα για τη μετατροπή των διαδοχικών αρχείων φασματομετρίας μάζας σε ταυτοποιήσεις πρωτεϊνών. Σήμερα, υπάρχουν πάνω από 1000 διαφορετικά εργαλεία για την επεξεργασία δεδομένων proteomics (42). Αυτός ο αριθμός περιλαμβάνει τόσο το ενεργό λογισμικό όσο και τα εργαλεία που έχουν περιθωριοποιηθεί λόγω των ραγδαίων εξελίξεων στον τομέα της ανάλυσης πρωτεϊνών. Τα κύρια εργαλεία αναζήτησης που χρησιμοποιούνται από τους περισσότερους ερευνητές σήμερα είναι πολύ λιγότερα. Οι περισσότεροι χρήστες είναι πιθανό να προτιμούν εργαλεία κυρίως μεταξύ των MaxQuant (43,44), Mascot (45), SpectroNaut (46), Proteome Discoverer και PEAKS (47). Οι υπόλοιποι χρήστες

προτιμούν είτε άλλα ισχυρά εμπορικά εργαλεία, ή εργαλεία ανοιχτού κώδικα (open source).

Η επιλογή της κατάλληλης στατιστικής μεθόδου είναι πολύ σημαντικό βήμα στην ανάλυση των βιοϊατρικών δεδομένων. Μια λανθασμένη επιλογή της στατιστικής μεθόδου όχι μόνο δημιουργεί κάποιο σοβαρό πρόβλημα κατά την ερμηνεία των ευρημάτων αλλά επηρεάζει και το συμπέρασμα της μελέτης. Στη στατιστική, για κάθε συγκεκριμένη κατάσταση, διατίθενται στατιστικές μέθοδοι για ανάλυση και ερμηνεία των δεδομένων. Για να επιλεγεί η κατάλληλη στατιστική μέθοδος, πρέπει να γνωρίζει κανείς την υπόθεση και τις συνθήκες των στατιστικών μεθόδων, ώστε να μπορεί να επιλεγεί η κατάλληλη στατιστική μέθοδος για ανάλυση δεδομένων. Στα πλαίσια της R, έχουν ανακύψει αξιοσημείωτα εργαλεία διαφορεικής ανάλυσης για πιο εξειδικευμένες αναλύσεις, όπως το DESeq2, το limma και το GSEA.

Οι τεχνικές εμπλουτισμού για δεδομένα «omics» είναι βασικά εργαλεία για την κατανόηση πολύπλοκων βιολογικών συστημάτων. Αυτά τα εργαλεία μειώνουν την πολυπλοκότητα των δεδομένων, βελτιώνουν την ερμηνεία και την κατανόηση των βιολογικών συστημάτων και βοηθούν στη δημιουργία υποθέσεων. Τα εργαλεία εμπλουτισμού υποδηλώνουν οποιαδήποτε αναλυτική τεχνική που επωφελείται από πληροφορίες μοριακών μονοπατιών ή δικτύων για την απόκτηση εικόνας ενός βιολογικού συστήματος. Μέχρι στιγμής υπάρχει μια πληθώρα εργαλείων εμπλουτισμού, εκ των οποίων μερικά γνωστά είναι το DAVID, το PANTHER, το Metascape, το GSEA και το clusterProfiler.

Η οπτικοποίηση ήταν από καιρό το κλειδί για να βοηθήσει στην κατανόηση των βιολογικών συστημάτων, όπως ο μεταβολισμός, η

σηματοδότηση και η ρύθμιση της γονιδιακής έκφρασης. Τα τελευταία χρόνια, η μελέτη τέτοιων συστημάτων έχει επηρεαστεί βαθιά από την ανάπτυξη ενός ευρέος φάσματος πειραματικών μεθόδων υψηλής απόδοσης, με αποτέλεσμα έναν πολύ αυξημένο όγκο πολύπλοκων, διασυνδεδεμένων δεδομένων. Υπήρξε αντίστοιχη αύξηση στην ανάπτυξη εργαλείων οπτικοποίησης για δεδομένα βιολογίας συστημάτων. Αυτά τα εργαλεία είναι πολύ διαφορετικά, αλλά μπορούν να χωριστούν ευρέως σε δύο εν μέρει επικαλυπτόμενες κατηγορίες. Η πρώτη αποτελείται από εργαλεία που επικεντρώνονται σε αυτοματοποιημένες μεθόδους ερμηνείας και εξερεύνησης μεγάλων βιολογικών δικτύων και η δεύτερη αποτελείται από εργαλεία που επικεντρώνονται στη συναρμολόγηση και την επιμέλεια των μονοπατιών. Πολλά από αυτά τα εργαλεία είναι στενά ενσωματωμένα με δημόσιες βάσεις δεδομένων, επιτρέποντας έτσι στους χρήστες να οπτικοποιούν και να ερμηνεύουν τα δικά τους δεδομένα στο πλαίσιο της ήδη υπάρχουσας γνώσης. Μερικά από τα πιο γνωστά εργαλεία οπτικοποίησης είναι το plotly για την Python και το ggplot2 για την R. Επιπλέον, υπάρχουν και εργαλεία τα οποία οπτικοποιούν δεδομένα ταυτόχρονα με την ανάλυσή τους, όπως το Cytoscape, το Metascape, το GSEA και η StringDB.

Στην πληροφορική, μια βάση δεδομένων είναι μια οργανωμένη συλλογή δεδομένων ή ένας τύπος αποθήκευσης δεδομένων που βασίζεται στη χρήση ενός συστήματος διαχείρισης βάσεων δεδομένων (DBMS), του λογισμικού που αλληλεπιδρά με τους τελικούς χρήστες, τις εφαρμογές και την ίδια τη βάση δεδομένων για τη συλλογή και ανάλυση των δεδομένων. Το DBMS περιλαμβάνει επιπλέον τις βασικές εγκαταστάσεις που παρέχονται για τη διαχείριση της βάσης δεδομένων. Το σύνολο της βάσης δεδομένων, του DBMS

και των σχετικών εφαρμογών μπορεί να αναφέρεται ως σύστημα βάσης δεδομένων. Συχνά ο όρος αυτός χρησιμοποιείται επίσης χαλαρά για να αναφέρεται σε οποιοδήποτε από τα DBMS, το σύστημα βάσης δεδομένων ή μια εφαρμογή που σχετίζεται με τη βάση δεδομένων. Παρότι η UniProt είναι η πιο ευρέως γνωστή βάση δεδομένων, υπάρχουν και άλλες που μπορούν να λειτουργήσουν συμπληρωματικά με αυτή. Μερικές από αυτές είναι η PRIDE, η PeptideAtlas, η MaxQB και η Human Protein Atlas.

Με βάση τα εξαιρετικά πλεονεκτήματα των multi-omics στην ανακάλυψη των αιτιολογικών μηχανισμών νόσων, η προσανατολισμένη στην ασθένεια έρευνα απαιτεί ολοένα και περισσότερο τη συμμετοχή τεχνικών multi-omics και υπολογιστικών αλγορίθμων. Οι βιοδείκτες δεν μπορούν μόνο να εξερευνήσουν την παθογένεια σε μοριακό επίπεδο, αλλά έχουν επίσης μοναδικά πλεονεκτήματα στην εύστοχη και ευαίσθητη αξιολόγηση της πρώιμης «ζημιάς» χαμηλού επιπέδου. Έτσι παρέχουν έγκαιρη προειδοποίηση, ανάλυση προγνωστικής αποτελεσματικότητας και ακριβή σταδιοποίηση και τυποποίηση των ασθενειών.

Αν και πολλοί βιοδείκτες που σχετίζονται με τον καρκίνο έχουν ταυτοποιηθεί με μεμονωμένες τεχνικές omics, τα multi-omics εμφανίζουν πλεονεκτήματα στην έρευνα για τον καρκίνο. Επιπλέον, η ενσωμάτωση διαφόρων πληροφοριών omics αναμένεται να είναι πολύτιμη για την καθοδήγηση στοχευμένης θεραπείας του καρκίνου.

Μια πρώιμη προσέγγιση στην πολυομική ανάλυση ήταν η χωριστή ανάλυση διαφορετικών τύπων δεδομένων και στη συνέχεια συνδυασμός των αποτελεσμάτων για την απόκτηση ενός ολοκληρωμένου δικτύου μοριακών

αλληλεπιδράσεων. Με σημαντικές εξελίξεις σε αυτόν τον τομέα, τα πλαίσια και οι μέθοδοι αλγοριθμικής μετα-ανάλυσης έχουν γίνει το κύριο μέσο μιας ολοκληρωμένης ανάλυσης δεδομένων multi-omics. Ωστόσο, η ανάλυση ενσωμάτωσης multi-omics αντιμετωπίζει επίσης ορισμένες προκλήσεις, όπως ο χειρισμός των ελλειπών τιμών (missing values), η ετερογένεια μεταξύ διαφορετικών omics, οι δυσκολίες στην ερμηνεία των μοντέλων multi-omics και τα προβλήματα στον σχολιασμό δεδομένων, την αποθήκευση και τους υπολογιστικούς πόρους.

Στο μέλλον, με την επέκταση των δειγμάτων και των χαρακτηριστικών, τα δεδομένα multi-omics θα γίνουν πιο περίπλοκα και η ανάπτυξη της ιατρικής ακριβείας απαιτεί την ενσωμάτωση συνόλων δεδομένων από μεγάλες ομάδες ασθενών. Η συμπερίληψη άλλων τύπων δεδομένων, όπως η ραδιομική και τα single-cell omics στην έρευνα multi-omics στοιχείων, θα αυξήσει περαιτέρω την πολυπλοκότητα της επεξεργασίας δεδομένων.

Τα single-cell multi-omics είναι η πιο ελπιδοφόρα κατεύθυνση ανάπτυξης στον τομέα της πολυομικής, η οποία μπορεί να συναγάγει ρυθμιστικά μοντέλα μεταξύ πολυεπίπεδων μορίων με καλύτερη ακρίβεια. Η λογιστική για την ετερογένεια των τύπων κυττάρων θα διευκολύνει τη μοντελοποίηση πολύπλοκων βιολογικών διεργασιών (π.χ. όγκων) και την κατανόηση της λειτουργίας σε εξαιρετικά διαφοροποιημένα όργανα (π.χ. εγκέφαλος). Πολλαπλές μελέτες omics στο ίδιο κύτταρο ήταν επιτυχείς, όπως ο συνδυασμός RNA-sequencing, μεθυλίωσης DNA ή/και αφθονίας πρωτεϊνών σε κυτταρικό επίπεδο.

Τα κύρια συστατικά μίας single-cell multi-omics ανάλυσης περιλαμβάνουν τεχνικές απομόνωσης μεμονωμένων κυττάρων, σήμανσης (barcoding) και αλληλούχισης για τη μέτρηση πολλαπλών τύπων μορίων από το ίδιο κύτταρο, καθώς και μια ολοκληρωμένη ανάλυση μορίων που μετρώνται σε κυτταρικό επίπεδο για τον εντοπισμό κυτταρικών τύπων και τις λειτουργίες τους που σχετίζονται με παθοφυσιολογικές διεργασίες με βάση τα μοριακά χαρακτηριστικά. Τα single-cell omics συνήθως επικεντρώνονται στα δεδομένα του γονιδιώματος-mRNA, της μεθυλίωσης του mRNA-DNA, της προσβασιμότητας mRNA-χρωματίνης και της πρωτεΐνης-mRNA.

Ωστόσο, παρόλο που τα δεδομένα single-cell multi-omics παρέχουν πλήθος βιολογικών πληροφοριών, η ολοκληρωμένη ανάλυσή τους αντιμετωπίζει τους ίδιους τύπους προκλήσεων όπως αυτών των υπόλοιπων multi-omics τεχνικών. Επιπλέον, υπάρχει έλλειψη τεχνολογίας ικανής να παράγει πολυτροπικά δεδομένα, με αποτέλεσμα περιεκτικές μελέτες που συχνά βασίζονται στην ενσωμάτωση τρόπων από διαφορετικά σύνολα δεδομένων.

Την τελευταία δεκαετία, οι single-cell τεχνολογίες έχουν αποκαλύψει την ετερογένεια του μικροπεριβάλλοντος ανοσοποιητικού-όγκου σε γονιδιωματικό, μεταγραφομικό και πρωτεομικό επίπεδο και έχουν βελτιώσει την κατανόησή μας για τους μηχανισμούς ανάπτυξης όγκων. Οι τεχνολογίες αυτές έχουν επίσης χρησιμοποιηθεί για τον εντοπισμό πιθανών βιοδεικτών. Πρόσφατα, χωρικές τεχνολογίες multi-omics (spatial multi-omics) έχουν χρησιμοποιηθεί για τη μελέτη μεταγραφωμάτων, πρωτεϊνών και μεταβολισμών μικροπεριβαλλόντων ανοσοποιητικών όγκων σε διάφορους τύπους καρκίνου και τα δεδομένα που προέκυψαν από αυτές τις μεθόδους συνδυάστηκαν με

ανοσοϊστοχημεία και πολυπαραμετρική ανάλυση για να δώσουν δείκτες εξέλιξης του καρκίνου.

Τα multi-omics συνδυάζουν τις διάφορες προσεγγίσεις για την διερεύνηση σε επίπεδο κυττάρου, ώστε να αξιολογήσουν πληρέστερα την υγεία ή την ασθένεια ενός μεμονωμένου κυττάρου. Οι αναλύσεις single-cell έχουν ανθίσει και περιλαμβάνουν μια σειρά πεδίων, συμπεριλαμβανομένης της γονιδιωματικής, της μεταγραφομικής, της πρωτεομικής, της επιγονιδιωματικής και της μεταβολομικής, καθένα από τα οποία παρέχει μοναδικές και κρίσιμες γνώσεις για την εσωτερική λειτουργία των ανθρώπινων κυττάρων.

Μπορεί να φανταστεί κανείς ότι θα ήταν δυνατό να συγχωνευθούν όλα τα διαγνωστικά εργαλεία που έχει στη διάθεση ενός γιατρού σε ένα μόνο τεστ που έλεγγχε τα πάντα ταυτόχρονα. Θα εξοικονομούσε χρόνο, επιτρέποντας στον γιατρό να λάβει μια απάντηση για το τι ταλαιπωρεί τον ασθενή πιο γρήγορα, και χρήματα, καθώς θα ήταν μόνο μία εξέταση. Αυτή είναι μια μεγάλη αναλογία με αυτό που κάνουν τα multi-omics για τους ερευνητές και τις φαρμακευτικές εταιρείες, δίνοντας μια πιο γρήγορη, πληρέστερη εικόνα της απόκρισης των ασθενειών και των φαρμάκων, ειδικά για τις πιο προηγμένες κυτταρικές και γονιδιακές θεραπείες που υπόσχονται την εξάλειψη καταστροφικών ασθενειών όπως ο καρκίνος και άλλες γενετικές ανωμαλίες που έχουν αποφύγει τη θεραπεία μέχρι τώρα.