ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Big Data στη Γενωμική- Μέθοδοι, Αλγόριθμοι, Μαθηματική Προσέγγιση: Μία Συστηματική Ανασκόπηση

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μαρία Δελαπόρτα

Επιβλέπων: Διονύσιος-Δημήτριος Κουτσούρης, Καθηγητής Ε.Μ.Π.
Συνεπιβλέπουσα: Ουρανία Πετροπούλου, Μέλος Ε.Δ.Ι.Π. Ε.Μ.Π.

Αθήνα, Οκτώβριος 2021

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Big Data στη Γενωμική- Μέθοδοι, Αλγόριθμοι, Μαθηματική Προσέγγιση: Μία Συστηματική Ανασκόπηση

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μαρία Δελαπόρτα

Επιβλέπων: Διονύσιος-Δημήτριος Κουτσούρης, Καθηγητής Ε.Μ.Π.
Συνεπιβλέπουσα: Ουρανία Πετροπούλου, Μέλος Ε.Δ.Ι.Π. Ε.Μ.Π.

Αθήνα, Οκτώβριος 2021

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Big Data στη Γενωμική- Μέθοδοι, Αλγόριθμοι, Μαθηματική Προσέγγιση: Μία Συστηματική Ανασκόπηση

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μαρία Δελαπόρτα

Επιβλέπων: Διονύσιος-Δημήτριος Κουτσούρης, Καθηγητής Ε.Μ.Π.
Συνεπιβλέπουσα: Ουρανία Πετροπούλου, Μέλος Ε.Δ.Ι.Π. Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 12η Οκτωβρίου 2021.

Διονύσιος-Δημήτριος Κουτσούρης **Καθηγητής Ε.Μ.Π.**	**Γεώργιος Ματσόπουλος** **Καθηγητής Ε.Μ.Π.**	**Παναγιώτης Τσανάκας** **Καθηγητής Ε.Μ.Π.**

Αθήνα, Οκτώβριος 2021

……

Μαρία Δελαπόρτα

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η χρήση μεθοδολογιών εξαιρετικά υψηλής απόδοσης (high throughput) στη βιολογία έχει αυξηθεί πολύ στις τελευταίες δεκαετίες. Οι δύο κύριες τεχνολογίες, δηλαδή οι Μικροσυστοιχίες και η Αλληλουχοποίηση Νέας Γενιάς (NGS), προσφέρουν απαραίτητες και δημοφιλείς μεθοδολογίες για την ανάλυση των δεδομένων της γενωμικής. Αυτό σημαίνει ότι πλέον απαιτείται η ανάπτυξη και εφαρμογή νέων ή βελτιωμένων βιοστατιστικών μεθόδων και εργαλείων βιοπληροφορικής για την ανάλυσή τους, είτε στην ανάλυση δεδομένων μικροσυστυχιών είτε στην ανάλυση δεδομένων αλληλουχοποίησης. Αυτές οι μέθοδοι χρησιμοποιούνται στις Ωμικές Επιστήμες και αναφέρονται σε ολοκληρωμένες βιολογικές μελέτες. Η παρούσα εργασία έχει ως στόχο την συγκέντρωση των σχετικών πληροφοριών μέσω ανασκόπησης της βιβλιογραφίας για τις μαθηματικές και υπολογιστικές μεθόδους που χρησιμοποιούνται στην ανάλυση γενωμικών δεδομένων, ειδικότερα στις προαναφερθείσες μεθόδους. Η ανάλυση Μικροσυστοιχιών έχει εμφανιστεί πολύ νωρίτερα και τα βιολογικά δεδομένα, που έχουν παραχθεί και συνεχίζουν να παράγονται από αυτή είναι αμέτρητα. Από την άλλη πλευρά, η ανάλυση NGS μπορεί να γίνει σε άγνωστο βιολογικό δείγμα, με την εκ νέου αλληλούχιση, και γι'αυτό τον λόγο είναι πιο ευρέως εφαρμόσιμη, την τελευταία δεκαετία υπάρχει ραγδαία αύξηση των δεδομένων αλληλουχοποίησης σε βάσεις δεδομένων με ανοιχτή πρόσβαση από διάφορα μέρη του κόσμου. Οι δύο προαναφερθείσες μέθοδοι ανάλυσης χρησιμοποιούνται ευρέως και ανά περίπτωση, ανάλογα με τον σκοπό του ερευνητή. Η ανασκόπηση αναφέρεται την αρχή λειτουργίας και την περιγραφή της πειραματικής διαδικασίας τους, καθώς και σε βασικές έννοιες της βιολογίας, για να γίνει κατανοητή η ορολογία που χρησιμοποιείται. Στην συνέχεια, γίνεται συστηματική ανασκόπηση στατιστικών ελέγχων και υπολογιστικών μεθόδων για την επεξεργασία των διαφορετικών δεδομένων, έτσι ώστε τα αποτελέσματα να είναι σωστά και αναπαράξιμα. Ακόμα, διερευνάται η διαδικασία χαρακτηρισμού γονιδίων και, παράλληλα, αναδεικνύονται εφαρμογές στην γονιδιακή οντολογία και την ανάλυση σηματοδοτικών μονοπατιών. Εμφανίζονται, ακόμα, μερικές εφαρμογές μαθηματικής προτυποποίησης. Το θέμα ολοκληρώνεται με τη μελέτη περιπτώσεων στη χρήση Μεγάλων Δεδομένων στη γενωμική και την σημασία που έχουν η εφαρμογές και το προγραμματιστικό περιβάλλον τους, αφού θα φέρουν την νέα επιστημονική επανάσταση σε πολλούς τομείς, συμπεριλαβανομένου του τομέα της φαρμακολογίας, της γεωπονικής και της προσωποποιημένης ιατρικής.

Λέξεις-κλειδιά: DNA, RNA, miRNA, μεθυλίωση, Μικροσυστοιχίες, Αλληλουχοποίηση Νέας Γενιάς, πείραμα, βιοστατιστική, βιοπληροφορική, κανονικοποίηση, διαφορικά

εκφραζόμενα γονίδια, εξόρυξη δεδομένων, χαρακτηρισμός γονιδίων, γονιδιακή οντολογία, ανάλυση σηματοδοτικών μονοπατιών, Μεγάλα Δεδομένα.

Abstract

The use of high throughput techniques in the life sciences has grown substantially over the past few decades. There is an plethora of popular analysis methods that derive from the two leading technologies, Microarrays and Next-Generation Sequencing (NGS). This means that there is a need for the development of new or improved biostatistics and bioinformatics methods and tools to analyze the biological data, whether we analyze DNA microarray data or RNA-Seq data. Those methods are used in the Omics sciences and refer to biological analysis in a comprehensive way. The goal of this thesis is to gather the pertinent information via a systematic literature review focusing on the mathematical and computational methods in the analysis of genomic data, mainly in those two leading technologies. Microarrays analysis has been around longer and there is no end to the data made available. However, NGS analysis does not require prior knowledge of the biological sample, i.e. de novo sequencing, and is a more thorough analysis than microarrays, therefore over the last decade there is an explosion of sequencing data made available in public repositories all over the world. Both analysis are used widely, always according to the Scientists' purpose. The review covers their principle and workflow, as well as basic concepts in biology, so that the terminology can be better understood. Following that, there is a systematic review of statistical tests as well as computational methods to process the diverse data so the results are accurate and reproducible. Furthermore, the process of gene annotation is investigated and at the same time Gene Ontology (GO) and Pathway Analysis is showcased. A few examples of Mathematical Modeling are shown. The thesis concludes with case studies in the use of Big Data in Genomics and the importance of the tools and environments that will constitute the next revolution in several fields of study, including pharmacology, agriculture and personalized medicine.

Keywords: DNA, RNA, miRNA, methylation, microarrays, NGS, wetlab, biostatistics, bioinformatics, normalization, DEGs, data mining, gene annotation, gene ontology, pathway analysis, Big Data.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τους επιβλέποντες της διπλωματικής μου, καθηγητή κ. Διονύσιο-Δημήτριο Κουτσούρη, διευθυντή του Εργαστηρίου Βιοϊατρικής Τεχνολογίας, και την κ. Ουρανία Πετροπούλου, μέλος Ε.ΔΙ.Π. του ίδιου εργαστηρίου της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνίου, που μου έδωσαν την ευκαιρία να εκπονήσω την παρούσα εργασία. Θα ήθελα να ευχαριστήσω και τον κ. Γεώργιο Λάμπρου, μέλος Ε.ΔΙ.Π. της Ιατρικής Σχολής του Ε.Κ.Π.Α. και ερευνητή του Χωρέμειου Ερευνητικού Εργαστηρίου Νοσοκομείου Παίδων «Αγία Σοφία», για την πολύτιμη βοήθεια του στην εκπόνησή της. Ακόμα, ένα μεγάλο ευχαριστώ στους φίλους μου για την ηθική συμπαράσταση. Τέλος, θα ήθελα να ευχαριστήσω τον σημαντικότερο άνθρωπο στην ζωή μου, που πάντα με υποστηρίζει και μου δίνει δύναμη, την μητέρα μου.

Στον πατέρα μου

Πίνακας Περιεχομένων

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Κατάλογος Εξισώσεων

Κατάλογος Χημικών Αντιδράσεων

1. Κεφάλαιο 1 Το Γονιδίωμα και οι Ιδιότητές του

Με την πρόοδο σε πολλούς κλάδους επιστημών που αφορούν την βιολογία, όπως την μοριακή βιολογία, την ιατρική, την μικροβιολογία, την βιοϊατρική μηχανική δημιουργήθηκαν ευνοϊκότεροι παράγοντες για την μελέτη ολόκληρου του γονιδιώματος. Οι τεχνικές ανάλυσης εξελίχθηκαν και έγιναν οικονομικά προσιτές, από εκατομμύρια σε μερικές χιλιάδες ευρώ. Η γενωμική είναι μια αναπτυσσόμενη επιστήμη για την διαλεύκανση μηχανισμών που οδηγούν στο γήρας, έτσι ώστε να επιβραδυνθεί όσο είναι δυνατόν στον άνθρωπο, και την ανάπτυξη εξατομικευμένων θεραπειών για την εξάλειψη παθογενειών και μη φυσιολογικών καταστάσεων, όπως είναι στον άνθρωπο ο καρκίνος και το Alzheimer. Το γένωμα ενός ανθρώπου αποτελείται από 39.109 γονίδια.

Omics. Οι ραγδαίες εξελίξεις στην επιστήμη της βιολογίας έφεραν στο φως τις τεχνολογίες με την ονομασία ωμική, δηλαδή την γενωμική (genomics), την επιγενωμική (epigenomics), την μεταγραφομική (transcriptomics), την πρωτεϊνομική (proteomics), την μεταβολική (metabolomics) και πολλές άλλες. Οι τεχνολογίες ωμικής (omics) έχουν σκοπό να δώσουν απαντήσεις για την διαλεύκανση βασικών μηχανισμών λειτουργίας του κυττάρου και των οργάνων. Δηλαδή αποτελούν έναν καινοτόμο τρόπο εξέτασης των βιολογικών δειγμάτων, αφού προσφέρουν μια πιο συνολική εικόνα για τους μηχανισμούς και τα μονοπάτια που σχετίζονται με αυτούς, αντί για μεμονωμένα στοιχεία που δεν περιλαμβάνουν πληροφορίες για το ευρύτερο εννοιολογικό πλαίσιο τους. Για παράδειγμα, στον ανθρώπινο εγκέφαλο αυτό έχει την μορφή διαλεύκανσης πολύπλοκων μηχανισμών σκέψης, νόησης και μνήμης. Το κεντρικό δόγμα της βιολογίας είναι η αρχή που ορίζει ότι οι γενετικές πληροφορίες ρέουν από το DNA στο RNA, μέσω μεταγραφής, από εκεί μεταφράζεται σε πρωτεΐνες για να φτάσουμε στον φαινότυπο. Δηλαδή, η ροή της πληροφορίας ρέει από το γένωμα στο μεταγράφομα και στο πρωτεΐνομα για να φτάσουμε στο φαινότυπο. Έτσι, εξηγείται πως οι ωμικές επιστήμες είναι αλληλένδετες, αλληλοεπηρεάζονται και αναπτύσσονται παράλληλα.

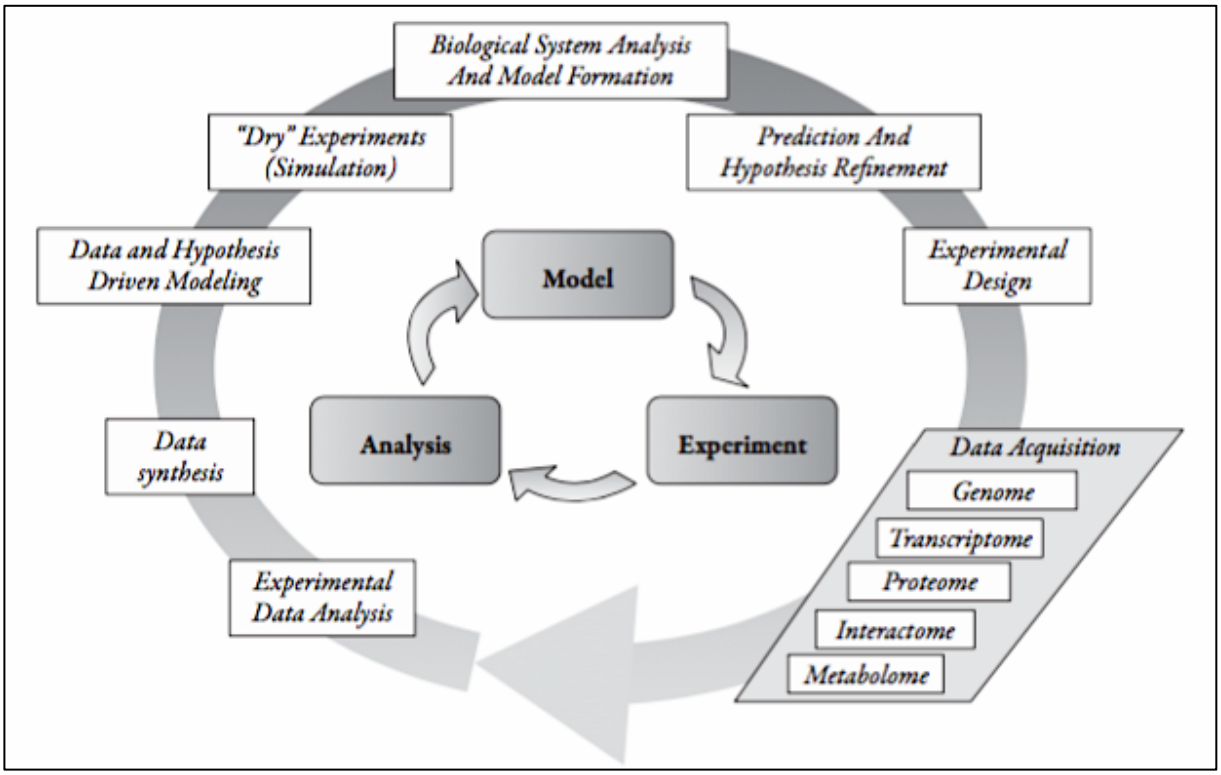
Ο κλάδος της γενωμικής είναι η μελέτη του συνόλου των γονιδίων ενός οργανισμού. Η επιγενωμική είναι η μελέτη ολόκληρης της γκάμας των επιγενετικών μεταλλάξεων σε όλο το γονιδίωμα του κυττάρου, ενώ η μεταβολομική είναι η μελέτη των χημικών διεργασιών που περιλαμβάνουν μεταβολίτες. Αντίστοιχα, η πρωτεϊνομική είναι η μελέτη της ανίχνευσης των πρωτεϊνών ενός οργανισμού και η μεταγραφική είναι η μελέτη ανίχνευσης mRNA (messenger RNA), δηλαδή των αγγελιοφόρων μικρών κινητών μορίων RNA (ribonucleic acids), δηλαδή ριβονουκλεϊνικών οξέων. Κάθε τμήμα DNA (deoxyribonucleic acids), δηλαδή

δεοξυριβονουκλεϊνικά οξέα, παράγει πολλά mRNA, καθένα από τα οποία μπορεί να συνθέσει πολλές πολυπεπτιδικές αλυσίδες.

Πιο αναλυτικά, η γενωμική αφορά την μελέτη των νουκλεϊνικών αλληλουχιών, δηλαδή της γραμμικής σειράς των υπομονάδων/νουκλεοτιδίων σε μια πολυμερή αλυσίδα, καθώς και των προϊόντων, της δομής και της βιολογικής λειτουργίας αυτής στο σύνολο του γονιδιώματος. Το γένωμα των περισσότερων οργανισμών είναι μεγάλο και για την αλληλούχιση του οι επιστήμονες πρώτα κατακερματίζουν τα νουκλεϊνικά οξέα σε θραύσματα και τα επανασυναρμολογούν, ελέγχοντας την ορθότητα του συναρμολογημένου γενώματος. Η γενωμική χωρίζεται σε υποκατηγορίες ανάλογα με το τι εξετάζει. Η λειτουργική γενωμική έχει να κάνει με τον ρόλο του κάθε γονιδίου ή ομάδας γονιδίων στις λειτουργίες ενός οργανισμού. Η συγκριτική (comparative) γενωμική χρησιμοποιεί πληροφορίες από την έρευνα ενός οργανισμού σε κάποιον άλλο.

Τα υπολογιστικά δίκτυα και η βιοπληροφορική είναι η ραχοκοκαλιά της έρευνας των επιστημών την ωμικής, αφού επιτρέπει την συστηματική ανάλυση των κλινικών μελετών και των επιστημονικών ερευνών για την χρησιμοποίηση σε περαιτέρω έρευνα πολλές φορές από διαφορετικές επιστημονικές ομάδες παράλληλα. Ο στόχος είναι η εφαρμογή της εκάστοτε τεχνολογίας και εκτός του ερευνητικού τομέα, αφού όμως έχουν διερευνηθεί οι κατευθυντήριες γραμμές της, βασισμένη σε στοιχεία από γενωμικά και επιστημονικά δεδομένα. Η **Εικόνα 1.** μας δίνει σχηματικά το πλαίσιο μοντελοποίησης και συστηματοποίησης της ερευνητικής διαδικασίας. Μετά την χρονοβόρα διαδικασία εξαγωγής δεδομένων, αναλύονται και δημιουργείται ένα υποθετικό μοντέλο. Στη συνέχεια, γίνονται προσομοιώσεις του στον υπολογιστή, έτσι ώστε να βελτιστοποιηθεί το μοντέλο και να γίνει το σχέδιο του πειράματος, για την εκ νέου εξαγωγή δεδομένων και επανάληψη των φάσεων με την ίδια σειρά μέχρι την ολοκληρωμένη μοντελοποίηση του πειράματος.

Για την καλύτερη κατανόηση του θέματος έχουν δοθεί ως υπόβαθρο, σε αυτό το κεφάλαιο, βασικές βιολογικές γνώσεις μοριακής βιολογίας καθώς και μια μικρή ιστορική αναδρομή των προσώπων και των ερευνών που κατέληξαν στην ανάπτυξη των αναλύσεων μικροσυστοιχιών και αλληλουχοποίησης νέας γενιάς, ειδικά αναλογιζόμενοι το αντίκτυπο που έχουν στην καθημερινότητά μας καθώς και η μελλοντική καθολική χρήση των αναλύσεων αυτών για την προσωποποιημένη υγεία των πολιτών.



Εικόνα 1. Πληροφορική ως κεντρικό δόγμα για τη Συστηματική Βιολογία και τις Γενωμικές Επιστήμες (1).

Έννοια του γονιδίου. Το γονιδίωμα ενός οργανισμού αναφέρεται στο σύνολο των γονιδίων του, η έκφραση των οποίων οδηγούν στην σύνθεση πρωτεϊνών. Αρχικά, λοιπόν πρέπει να αναφερθεί η πορεία της έννοιας γονίδιο, αφού κάθε καινούργια ανακάλυψη αναδιαμόρφωσε αυτή την έννοια. Ως αποτέλεσμα τις έρευνας του Mendel, γνωστός για τα επιτεύγματά του στην γενετική των φυτών, το γονίδιο έπαιζε τον ρόλο ενός διακεκριμένου στοιχείου, το οποίο κυβερνάει την εμφάνιση συγκεκριμένων γνωρισμάτων. Οι επιστήμονες Boveri, Weismann, Sutton και οι σύγχρονοί τους ανακάλυψαν ότι τα γονίδια έχουν μια φυσική παρουσία ως μέρος στα χρωμοσώματα. Οι Morgan, Sturtevant και οι συνάδελφοί τους έδειξαν ότι τα γονίδια έχουν συγκεκριμένες «διευθύνσεις», δηλαδή βρίσκονται σε συγκεκριμένες θέσεις και χρωμοσώματα, χωρίς να αλλάζουν από άτομο σε άτομο μέσα στον πληθυσμό ενός είδους. Από τους επιστήμονες Griffith, Avery, Hershey και Chase αποδείχτηκε ότι τα γονίδια αποτελούνται από DNA. Τέλος, οι Watson και Crick έφτιαξαν το μοντέλο DNA και απέδειξαν πως το μακρομόριο DNA μπορεί να κωδικοποιεί πληροφορίες για την κληρονομικότητα (2).

Ο ρόλος του γονιδίου στις κυτταρικές λειτουργίες και στην έκφρασή τους στα κληρονομικά γνωρίσματα έγινε το κέντρο πολλών ερευνών. Το 1908 ο Archibald Garrod, Σκοτσέζος

γιατρός, παρατήρησε ότι συμπτώματα που εκδηλώνονται λόγω σπάνιων κληρονομικών ασθενειών προκαλούνται από την απουσία συγκεκριμένων ενζύμων και τις αποκάλεσε «εσωτερικά λάθη του μεταβολισμού». Αυτή η ιδέα χρησιμοποιήθηκε σαν εναρκτήριο από τους ερευνητές Beadle και Tatum, οι οποίοι χρησιμοποίησαν σπόρους μούχλας Neurospora και απέδειξαν μέσω γενετικά μεταλλαγμένων σπόρων ότι αυτοί μεγάλωναν στο μέσο που είχε βιταμίνες και όχι σε αυτό που είχε αμινοξέα, δηλαδή υποδεικνύει ενζυμική έλλειψη για την παραγωγή μιας πρωτεΐνης, συγκεκριμένα το παντοθενικό οξύ. Το συμπέρασμα από το πείραμά τους είναι ότι ένα γονίδιο έχει τις πληροφορίες για την κατασκευή ενός πολυπεπτιδίου, από τα οποία αποτελούνται τα ένζυμα.

Αυτές οι πληροφορίες όμως δεν μπορούν να χρησιμοποιηθούν στις μεταβολικές διεργασίες άμεσα από το DNA. Η έκφραση των γονιδίων είναι ο τρόπος που παράγεται ένα λειτουργικό προϊόν, για τις μεταβολικές διεργασίες, χρησιμοποιώντας τις κωδικοποιημένες πληροφορίες στο γονίδιο. Οι κωδικοποιημένες πληροφορίες σε ένα τμήμα του DNA γίνονται προσβάσιμες στο κύτταρο μέσω την σύνθεση ενός συμπληρωματικού μορίου RNA, από το καλούπι DNA, αυτή η αντιγραφή των πληροφοριών ονομάζεται μεταγραφή (transcription). Οι δομικοί λίθοι των νουκλεϊνικών οξέων, DNA ή RNA, είναι τα νουκλεοτίδια (nt-nucleotides).

Η μοριακή δομή του DNA. Το DNA αποτελείται από επαναλαμβανόμενες υπομονάδες νουκλεοτιδίων. Κάθε νουκλεοτίδιο αποτελείται από ένα σάκχαρο με πέντε άτομα άνθρακα, που σχηματίζουν μια πεντόζη, συγκεκριμένου τύπου δεοξυριβόζης, στην οποία συνδέεται μια φωσφορική ομάδα και μια αζωτούχος βάση. Στο μόνο που διαφέρουν τα νουκλεοτίδια είναι η αζωτούχος βάση και υπάρχουν τέσσερις επιλογές:

• Αδενίνη (adenine, A) ανήκει στην οικογένεια των πουρινών.
• Γουανίνη (guanine, G) ανήκει στην οικογένεια των πουρινών.
• Κυτοσίνη (cytosine, C) ανήκει στις πυριμιδίνες και προέρχεται από τον εξαμελή δακτύλιο της πυριμιδίνης.
• Θυμίνη (thymine, T) αποκαλείται πυριμιδίνες και προέρχεται από τον εξαμελή δακτύλιο της πυριμιδίνης.

Οι δομικές διαφορές του DNA από το RNA. Το DNA διαφέρει δομικά από το RNA ως προς τον τύπο του σακχάρου που περιέχουν, δηλαδή το σάκχαρο είναι τύπου ριβόζης στο RNA, γι' αυτό και ονομάζεται ριβονουκλεϊνικό οξύ. Στο RNA αντικαθίσταται η αζωτούχος βάση θυμίνη (T), του DNA, από την αζωτούχο βάση ουρακίλη (U), οι δύο μοιάζουν χημικά. Το

RNA είναι συνήθως μονόκλωνο μόριο, ενώ το DNA το συναντάμε συνήθως ως δίκλωνο μόριο. Αυτές οι διαφορές επιτρέπουν στα ένζυμα να αναγνωρίσουν πότε το βιολογικό μόριο είναι DNA και πότε είναι RNA.

Σκοπός. Σε αυτή την εργασία θα εξεταστεί πως υπολογιστικά και μαθηματικά εργαλεία χρησιμοποιούνται για την εξερεύνηση των βιολογικών επιστημών, συγκεκριμένα για την γενωμική. Θα αναφερθούν και διαδικτυακές βάσεις δεδομένων, πηγές και διαδικτυακά εργαλεία της βιοπληροφορικής. Τα Μεγάλα Δεδομένα ή Big Data, αναφέρονται σε όλες τις πληροφορίες που έρχονται από μικρά και μεγάλα ερευνητικά κέντρα, βιολογικά δεδομένα ασθενών/πελατών από νοσοκομεία και ιδιωτικές κλινικές μαζί με το σύνολο των πληροφοριών από τις επιστήμες της υγείας πρέπει να ταξινομηθούν, να αναλυθούν και να χρησιμοποιηθούν για εξαγωγή χρήσιμων συμπερασμάτων για τις επιστήμες της υγείας.

1.1. Ιστορική αναδρομή της Γενωμικής

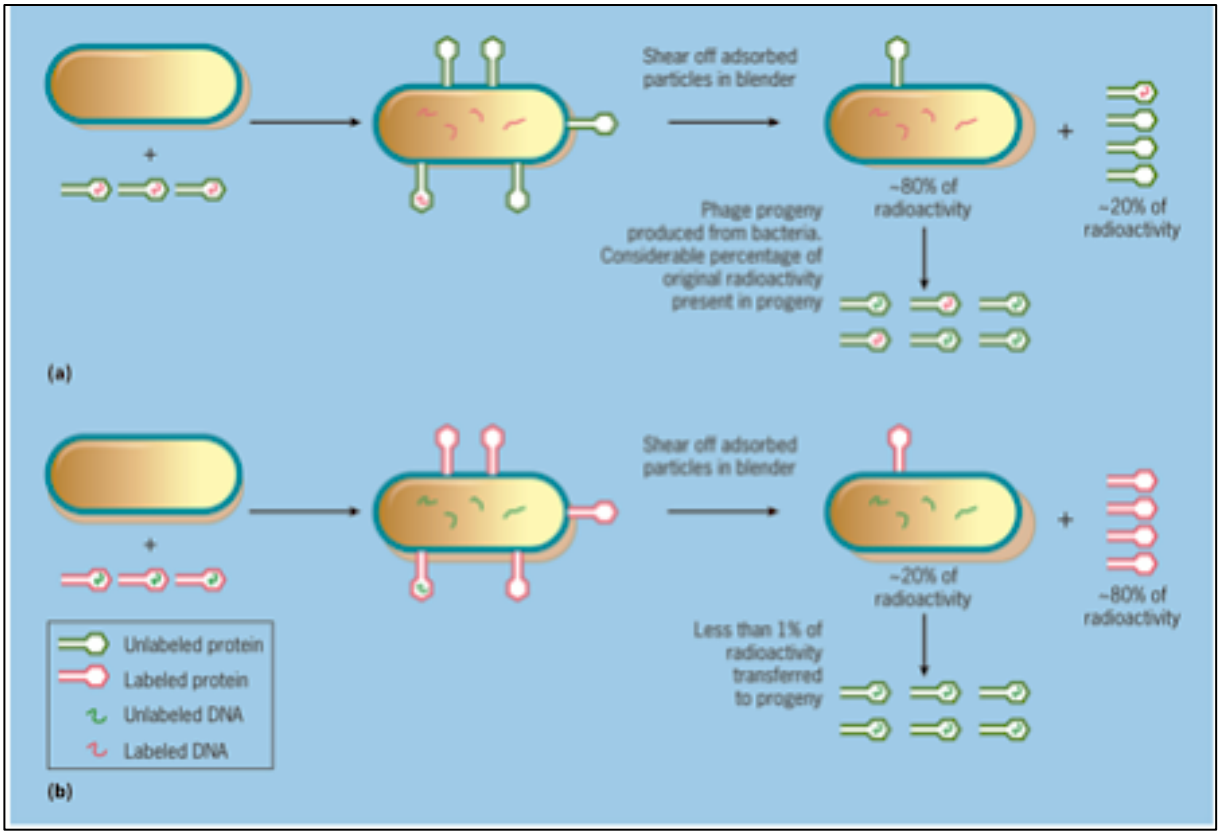
Η απόδειξη του ρόλου του DNA ως γενετικό υλικό. Όπως αναφέρθηκε στην προηγούμενη ενότητα, πολλοί επιστήμονες εργάστηκαν για να αποδείξουν ότι το DNA είναι το γενετικό υλικό. Ο Βρετανός γιατρός Fred Griffith μελέτησε τον Streptococcus pneumoniae, το βακτήριο που προκαλεί πνευμονία, ο οποίος όταν καλλιεργείται *in vitro*, δηλαδή διεργασία που πραγματοποιείται όχι σε έναν έμβιο οργανισμό αλλά σε απομονωμένο υποκυτταρικό εκχύλισμα, παίρνει δύο μορφές. Η μία είναι αθώα (S-smooth) και ασθενής, ενώ η άλλη προκαλεί νόσο (R-rough). Σε πειράματα που έκανε σε ποντίκια έδειξε ότι οι πνευμονιόκοκκοι που είχαν αδρανοποιηθεί με θέρμανση δεν ήταν πλέον παθογόνοι. Όταν πειραματίστηκε με την ανάμειξή τους με τα S βακτήρια, αυτά μεταμορφώθηκαν στην δεύτερη και θανατηφόρα R μορφή τους από τα αδρανοποιημένα και η αλλαγή ήταν μόνιμη. Κάτι που επιβεβαίωσε και ο Αμερικανός βακτηριολόγος Oswald Avery.

Ο Avery και οι συνεργάτες τους Colin Mac Leod και Maclyn McCarty, δημοσίευσαν το 1944 την μελέτη με την οποία απέδειξαν ότι το DNA είναι το γενετικό υλικό, η οποία ήταν λιγότερο πολύπλοκη γιατί έγινε in vitro. Οι επιστήμονες αδρανοποίησαν με θέρμανση ένα προς ένα τα βιοχημικά στοιχεία των S βακτηρίων και κάνανε έλεγχο της μεταμόρφωσης. Το πρώτο μέρος του πειράματος ήταν η υποβάθμιση του εξωτερικού καλύμματος του βακτηρίου S και η μεταμόρφωση παρέμεινε. Το δεύτερο μέρος του πειράματος ήταν η υποβάθμιση όλων των πρωτεϊνών του βακτηρίου S με ένα μείγμα ενζύμων που καταστρέφουν πρωτεΐνες, της θρυψίνης (trypsin) και της χυμοθρυψίνης (chymotrypsin) και η μεταμόρφωση παρέμεινε.

Αυτό τους εξέπληξε, αφού πολλοί σύγχρονοί τους επιστήμονες πίστευαν ότι οι πρωτεΐνες είναι το γενετικό υλικό. Το τρίτο μέρος του πειράματος ήταν η υποβάθμιση του ενζύμου RNάση (RNase), το οποίο αποσυνθέτει το μόριο RNA και μάλλον παίζει ρόλο στην σύνθεση πρωτεϊνών, και πάλι η μεταμόρφωση παρέμεινε. Όταν, όμως, χρησιμοποίησαν το ένζυμο DNάση (DNase), η οποία μπορεί να υδρολύσει εσωτερικούς φωσφοδιεστερικούς συνδέσμους, δηλαδή έχει την ιδιότητα να αλλοιώνει το δίκλωνο μόριο DNA του βακτηρίου S, η μεταμόρφωση δεν έγινε. Η απόδειξή τους χρειάστηκε λίγα ακόμα χρόνια να γίνει κοινά αποδεκτή από την επιστημονική κοινότητα.

Η ορθότητα του συμπεράσματός τους θεμελιώθηκε το 1952, με το πείραμα των Hershey και Chase **(**Εικόνα 2**),** αυτοί που πίστευαν ότι το γενετικό υλικό είναι οι πρωτεΐνες διαψεύστηκαν. Βασίστηκαν στη χημική διαφορετικότητα μεταξύ του DNA και των πρωτεϊνών, συγκεκριμένα το μόριο DNA περιέχει φώσφορο και η πρωτεΐνη περιέχει θείο, δηλαδή χημικά στοιχεία που δεν υπάρχουν και στις δύο ουσίες. Έκαναν το πείραμα σε δύο παρτίδες του ιού, τις οποίες άφησαν να μολύνουν το βακτήριο *Eschericia coli* (E. coli), που βρίσκεται στο παχύ έντερο, με το αντίστοιχο ραδιενεργό ισότοπο στην κάθε περίπτωση και μετά το μείγμα υποβλήθηκε σε φυγοκέντρηση. Στην πρώτη παρτίδα ραδιοσήμαναν με ραδιενεργά ισότοπα 35S τις πρωτεΐνες στο καψίδιο του ιού, όπου τα μόρια 35S-σημασμένων πρωτεϊνών είχαν παραμείνει στο εναιώρημα, μαζί με τα άδεια καψίδια των ιών, στο μεγαλύτερο τους ποσοστό (80%). Στην δεύτερη παρτίδα ραδιοσήμαναν το DNA με ραδιενεργά ισότοπα 32P, όπου μετά από ανάμειξή με το βακτήριο E. coli, βρέθηκαν τα μόρια 32P-σημασμένων μορίων DNA είχαν εισέλθει στα βακτηριακά κύτταρα, στο μεγαλύτερό τους ποσοστό (80%). Έτσι αποδείχθηκε μια και καλή ότι το DNA είναι το γενετικό υλικό.

Η ανακάλυψη της δομής του DNA. Ο Αυστριακός επιστήμονας Erwin Chargaff μετά από σύγκριση του DNA από διάφορους οργανισμούς ανακάλυψε ότι σε όλους κάθε μόριο DNA περιείχε τέσσερις αζωτούχες βάσεις, δηλαδή την γουανίνη, την αδενίνη, την κυτοσίνη και την θυμίνη. Στην Βρετανία οι επιστήμονες Rosalind Franklin και Maurice Wilkins χρησιμοποίησαν την επιστήμη της κρυσταλλογραφίας, δηλαδή με ακτίνες X-ray, για να ανακαλύψουν την δομή του μορίου DNA. Το αποτέλεσμα ήταν μια εικόνα στην οποία διαφαινόταν μια αχνή σκιά που έμοιαζε με μια περιστροφική σκάλα σε μορφή σπειροειδή έλικας. Μια άλλη ομάδα επιστημόνων, ο Αμερικανός James Watson και ο Βρετανός Francis Crick, την ίδια περίοδο έφτιαχναν μοντέλα για να καταλάβουν την δομή του DNA. Οι τελευταίοι εφάρμοσαν τις δύο παραπάνω μελέτες σε δύο νουκλεοτιδικές αλυσίδες που συνδέονται αντιπαράλληλα για να δημιουργήσουν μια διπλή έλικα. Για αυτή τους την ανακάλυψη κέρδισαν το 1962, μαζί με τον Wilkins, το Nobel Prize στην Ιατρική.



Εικόνα 2. Το πείραμα των Hershey και Chase (2).

Human Genome Project (HGP). Στο Πρόγραμμα Αλληλούχισης του Ανθρώπινου Γενώματος HGP χρησιμοποιήθηκε η γνωστή μέθοδος του Sanger από την δεκαετία του ’70, η διαφορά ήταν στην αυτοματοποίηση της αλληλουχοποίησης, που αναπτύχθηκε με ομαδική προσπάθεια στο εργαστήριο του Lee Hood στο αμερικάνικο πανεπιστήμιο Caltech, με την ερευνητική ομάδα του που αποτελούνταν από εκλεκτούς χημικούς, βιολόγους και μηχανικούς. Η αυτοματοποιημένη αλληλουχοποίηση ήταν η ιδέα δύο εκ αυτών του Lloyd Smith και του Mike Hunkapiller. Συγκεκριμένα ο Hunkapiller προσέλκυσε τον Smith για την χρήση βαφής 4 διαφορετικών χρωμάτων για την κάθε βάση, δηλαδή αντί για τέσσερις ξεχωριστές αντιδράσεις αλληλούχισης, σε διαφορετική λωρίδα γέλης, θα χρειαζόταν μόνο μια αντίδραση. Ο Smith ήταν ειδικός σε εφαρμογές λέιζερ, οπότε σκέφτηκε να χρησιμοποιήσει ειδικές βαφές που φθορίζουν όταν αυτές έρθουν σε επαφή με ακτίνες λέιζερ. Στη συνέχεια, σύμφωνα με την γνωστή μέθοδο Sanger, μια σειρά τμημάτων DNA τοποθετούνταν στην γέλη ανάλογα με το μέγεθός τους. Κάθε τμήμα ήταν σημασμένο με την φθορίζουσα βαφή που αντιστοιχεί στο διδεοξυριβονουκλεοτίδιο της, οπότε το χρώμα εκπομπής υποδεικνύει την βάση, μετά από σάρωση με λέιζερ και ανίχνευση με ένα «ηλεκτρικό μάτι». Αυτές οι πληροφορίες θα αποθηκεύονταν στον υπολογιστή.

Ο Hunkapiller το 1983 αποφασίζει να δουλέψει σε μια νέα, τότε, εταιρία κατασκευαστών μηχανημάτων, την Applied Biosystems, γνωστή ως ABI. Αυτή κατασκεύασε το πρώτο εμπορικό μηχάνημα αλληλουχοποίησης των Smith-Hunkapiller. Μέσω του HGP η αποδοτικότητα της διεργασίας βελτιώθηκε, η αργή και δύσχρηστη γέλη αντικαθίσταται από υψηλής απόδοσης τριχοειδή σύστημα (thin tubes) στο οποίο τα θραύσματα DNA ταχέως διαχωρίζονται ανάλογα με το μέγεθός τους. Οι νέες γενιές των αλληλουχοποιητών Sanger της συγκεκριμένης εταιρίας ήταν χιλιάδες φορές γρηγορότερες από το πρωτότυπο μηχάνημα. Με ελάχιστη ανθρώπινη παρέμβαση (15 λεπτά ανά εικοσιτετράωρο), αυτά τα μηχανήματα παρήγαγαν ως μισό εκατομμύριο ζεύγη βάσεων αλληλουχίας την ημέρα. Αυτή η τεχνολογία έκανε δυνατή την ανάλυση του γενώματος του ανθρώπου στο σύνολό της και την ολοκλήρωση του έργου το 2003, ενώ ήδη από τον Φεβρουάριο του 2001 είχε δημοσιευτεί ένα προσχέδιο του ανθρώπινου γενώματος.

Η γνώση ολόκληρου του γενώματος έδωσε περαιτέρω πληροφορίες για τα γονίδια, τα ρυθμιστικά στοιχεία καθώς και την δομή και οργάνωση των χρωμοσωμάτων. Οι τεχνικές αλληλουχοποίησης νέας γενιάς έκαναν δυνατή την αλληλούχιση ολόκληρων γονιδιωμάτων με μικρό κόστος και σε σύντομο χρονικό διάστημα. Η επιστήμη της γενωμικής ασχολείται με την ορθότητα του γενώματος (integrity), με τις γενετικές διαφοροποιήσεις στο ανθρώπινο γονιδίωμα και με την ανακάλυψη νέων προσωποποιημένων φαρμάκων. Η μεγάλη αύξηση των δεδομένων κρίνει απαραίτητη την εφαρμογή υπολογιστικών εργαλείων και την χρήση βάσεων δεδομένων για την λειτουργική ανάλυσή τους. Κάποιες από αυτές τις βάσεις δεδομένων είναι οι National Center for Biotechnology (NCBI), University of California Santa Cruz (UCSC) Genome Browser, the encyclopedia of DNA elements (ENCODE) και το roadmap epigenomics project (3, 4).

Το πρόγραμμα ENCODE. Το πρόγραμμα εγκυκλοπαίδειας στοιχείων DNA, γνωστό ως ENCODE (Encyclopedia of DNA Elements), με διάρκεια 5 χρόνια και σκοπό την ανάλυση του "junk DNA". Σε αυτό συμμετείχαν πάνω από 400 επιστήμονες από 32 χώρες, με χρήματα από το National Human Genome Institute, και το 2012 δημοσίευσαν 30 άρθρα σε επιστημονικά περιοδικά. Αυτά τα άρθρα, δέχτηκαν κριτική λόγω της θέσης τους ότι το 80% του ανθρώπινου γενώματος είναι λειτουργικό, δηλαδή έχουν κάποιο βιοχημικό ρόλο είτε ως θέση πρόσδεσης για μεταγραφικούς παράγοντες είτε ως πρωτεΐνες γονιδιακής ρύθμισης, που ενεργοποιούν και απενεργοποιούν τα γονίδια. Ο Ewan Birney, γνωστός και για τον ρόλο του στην κοινοπραξία ENCODE, παρατήρησε ότι "Αφού το 60% του γενώματος χαρακτηρίζεται ως εξόνιο ή εσόνιο, είναι λογικό το 20% να έχει λειτουργικό ρόλο". Το πρόγραμμα ENCODE παρέχει πληροφορίες του ανθρώπινου γενώματος αναφοράς και για τις κωδικές

και για τις μη κωδικές περιοχές, δηλαδή αφορά μια μεγάλη γκάμα μελετών. Για παράδειγμα, μελέτη πάνω στη μεθυλίωση DNA, μελέτη σε γονίδια που κωδικοποιούν πρωτεΐνες και μελέτη στα μετάγραφα RNA, τα οποία δεν μεταφράζονται σε πρωτεΐνες, αντιθέτως με το mRNA. Για το 8% του γενώματος υπήρχε επιβεβαιωμένη σχέση ανάμεσα στο DNA και τις πρωτεΐνες, ενώ για το υπόλοιπο 72% λειτουργικό DNA υπήρξε ανάγκη για περαιτέρω ανάλυση του ρόλου που παίζει στην γονιδιακή έκφραση (4).

Το πρόγραμμα HapMap. Ο αρχικός στόχος της διαχείρισης ενός τόσο μεγάλου εγχειρήματος ήταν ένα σχεδιάγραμμα ολόκληρου του γενώματος για την καθοδήγηση στη θέση κάθε αλληλουχίας, δηλαδή έπρεπε να χωριστεί σε διαχειρίσιμα κομμάτια, με στόχο αυτά να χαρτογραφηθούν. Το διεθνές απλότυπο πρόγραμμα χαρτογράφησης (International Haplotype Mapping Project or "HapMap") ξεκίνησε την διεργασία καταγραφής του πλήθους των διαφορών. Δεν ήταν η πρώτη προσπάθεια χαρτογράφησης και αναγνώρισης των γονιδίων, ειδικά αυτών υπεύθυνων για κληρονομικές ασθένειες, αλλά για να χαρτογραφηθούν οι γενετικοί παράγοντες που προκαλούν πολυπαραγοντικές ασθένειες, όπως διαβήτης, προβλήματα καρδιάς και καρκίνος, χρειάζεται συστηματική έρευνα των γενετικών διαφοροποιήσεων (4).

Ο στόχος του προγράμματος HapMap ήταν να αναγνωριστούν όλα τα σημεία στο ανθρώπινο γένωμα που έχουν ένα αποδεκτό ποσοστό (1%) διαφοροποίησης στον ανθρώπινο πληθυσμό, δηλαδή η αντικατάσταση μιας νουκλεοτιδικής βάσης με κάποιας άλλης ίσο-πιθανά εμφανιζόμενης βάσης.

SNPs. Οι ερευνητές με μοριακές τεχνικές ανιχνεύουν στην αλληλουχία βιολογικών μορίων DNA ή RNA αυτές τις παραλλαγές, δηλαδή τους μεμονωμένους νουκλεοτιδικούς πολυμορφισμούς (single nucleotide polymorphisms, SNPs, που διαβάζεται «snips») και εμφανίζονται στον κωδικοποιητή ή στις ρυθμιστικές περιοχές του γονιδίου.

GWAS. Οι πολυμορφισμοί αυτοί μπορεί είναι υπεύθυνοι για παθογένειες ακόμα και αν βρίσκονται σε μη κωδικές περιοχές. Οι επιστήμονες έχουν συλλέξει πολλούς τέτοιους βιολογικούς δείκτες για τον σχεδιασμό λεπτομερών χαρτών γενετικής σύνδεσης, από τους οποίους διαλέγουν τους πιο αντιπροσωπευτικούς δείκτες για την διεξαγωγή μελέτης γενετικής συσχέτισης σε επίπεδο γενώματος (genome-wide association studies ή GWAS). Με την φτηνή και εύκολη γονοτύπηση ανθρώπων για εκατοντάδες χιλιάδες ή και εκατομμύρια SNPs γίνεται η αναγνώριση των γονιδίων που εμπλέκονται σε αυτές τις ασθένειες. Γίνεται,

δηλαδή, η σύγκριση του απλοτύπου SNP φυσιολογικών ατόμων και αυτού από άτομα στα οποία εκφράζεται η εξεταζόμενη ασθένεια. Οι μη τυχαίοι συσχετισμοί μεταξύ των SNPs και της ασθένειας υποδεικνύουν ότι ένα ή περισσότερα γονίδια που σχετίζονται με την μη φυσιολογική κατάσταση είναι συνδεδεμένα με τους πολυμορφισμούς. Η μελέτες GWAS συνδέουν την εξεταζόμενη κατάσταση με μια περιοχή σε ένα χρωμόσωμα και στη συνέχεια εξετάζουν αυτή την περιοχή για τα γονίδια που μπορεί να είναι υπεύθυνα για αυτή την κατάσταση, αυτή είναι κάποιο χαρακτηριστικό ή μια ασθένεια.

Τα 270 άτομα που επιλέχθηκαν για το πρόγραμμα HapMap προέρχονταν από τέσσερις γεωγραφικά ποικίλους πληθυσμούς: Yoruba, Κινέζους (Han), Γιαπωνέζους, Αμερικανούς (Utah) ευρωπαϊκής καταγωγής. Στη γενωμική DNA ανάλυση των εθελοντών έγινε διαλογή τριών εκατομμυρίων SNPs. Το μεγαλύτερο κομμάτι της έρευνας έγινε από την εταιρία Perlegen, ένα παρακλάδι της εταιρίας Affymetrix, που υπήρξε πρωτοστάτης στην κατασκευή γονιδιακών τσιπς και εξαγοράστηκε το 2016 από μια άλλη εταιρία (Thermo Fisher Scientific Inc., Waltham, MA, USA) , δημιουργώντας ένα πλαίσιο για πάνω από 3 εκατομμύρια ίσο-κατανεμημένα SNPs, που λειτουργούν ως σηματοδότες για κάθε 900-1.000 βάσεις κατά μέσο όρο. Ο κατάλογος του ανθρώπινου SNPs τώρα υπερβαίνει τα 10 εκατομμύρια, κατανεμημένα ανά 300 βάσεις. Το πρόγραμμα που ακολούθησε ήταν το 1,000 Genomes Project και αυτό ολοκληρώθηκε το 2016 (4).

Insertion Deletion Variants (Indels). Οι μεταλλάξεις προσθηκών – αφαιρέσεων αναφέρονται στην προσθήκη ή/και αφαίρεση νουκλεοτιδίων στο γενωμικό DNA, με μέγεθος λιγότερο από 1kb. Χρησιμοποιούνται στις κλινικές μελέτες με τεχνικές νέας γενιάς (NGS), αφού τις συναντάμε σε πολλές μη φυσιολογικές καταστάσεις και στον καρκίνο. Η ανίχνευσή τους είναι λίγο πολύπλοκη, γιατί εξαρτάται από το μέγεθος των indels, από το πλαίσιο της αλληλουχίας και από τον χαρακτηρισμό της παραλλαγής. Τέτοιες μεταλλάξεις συμβολίζονται με «-» και λέγονται κενά (gaps).

Copy Number Variations (CNV). Οι αλλαγές αριθμού αντιγράφων, ή αλλιώς CNVs, είναι ένα από πολλά είδη υπομικροσκοπικών δομικών διαφοροποιήσεων που αποτελούνται από ελλείμματα ή διπλασιασμούς χρωμοσωμικού υλικού, με μέγεθος ως μερικά Mb. Στο ανθρώπινο γένωμα είναι πολύ λιγότερα από τα SNPs. Γι’ αυτό με σπάνιες μορφές των CNVs μελετώνται περίπλοκες ασθένειες, όπως το σύνδρομο Tourette. Οι τεχνικές μικροσυστοιχιών

DNA εφαρμόζονται σε πειράματα συγκριτικής υβριδοποίησης γενώματος (comparative genome hybridization-CGH) για την ποσοτικοποίηση των CNV στο γένωμα.

1.2. Ιδιότητες DNA/RNA

Αντιγραφή. Η πιο βασική βιολογική διεργασία για έναν οργανισμό είναι η αντιγραφή του DNA και η ιδιότητά της να επιδιορθώνει τα τυχόν λάθη στην αντιγραφόμενη αλληλουχία του. Ο διπλασιασμός αυτός της γενετικής πληροφορίας γίνεται γιατί θέλουμε να κρατήσουμε ένα αντίγραφο για το υπάρχον κύτταρο και ένα για το νέο κύτταρο. Σύμφωνα με το μοντέλο του Watson και Crick, κάθε κλώνος DNA λειτουργεί σαν εκμαγείο για να δημιουργήσει τον αντιπαράλληλο κλώνο DNA και κάθε κύτταρο θα αποτελείται από τον παλιό κλώνο DNA και τον νεο-συνταγμένο κλώνο DNA. Η αλληλουχία των βάσεων στον αντιπαράλληλο κλώνο καθορίζεται εύκολα σύμφωνα με τους κανόνες ζευγάρωσης των βάσεων. Συγκεκριμένα, όπου υπάρχει στο εκμαγείο το νουκλεοτίδιο A τοποθετείται το νουκλεοτίδιο T στον αντιπαράλληλο κλώνο, αυτή είναι η έννοια της συμπληρωματικότητας και ισχύει και αντιστρόφως. Τα νουκλεοτίδια G και T είναι επίσης συμπληρωματικά. Η διεργασία απαιτεί την παράλληλη λειτουργία αρκετών ενζύμων και άλλων πρωτεϊνών. Το ένζυμο DNA πολυμεράση έχει την δυνατότητα να ζευγαρώνει τα συμπληρωματικά νουκλεοτίδια στην αλυσίδα DNA. Αρχικά, το "πατρικό" δίκλωνο μόριο DNA ξετυλίγεται σε συγκεκριμένες θέσεις, που ονομάζονται αφετηρίες ή σημεία έναρξης της αντιγραφής. Η αντιγραφή είναι διπλής κατεύθυνσης, δηλαδή γίνεται παράλληλα και στις δύο διχάλες αντιγραφής. Η DNA πολυμεράση, μόλις εισέλθει το συμπληρωματικό στον "πατρικό" κλώνο νουκλεοτίδιο, το προσθέτει στο τέλος του αυξανόμενου "θυγατρικού" κλώνου. Η αντιγραφή ολοκληρώνεται όταν έχουν ξανατυλιχθεί όλες οι διχάλες αντιγραφής και σχηματιστεί τα δύο μόρια στην ίδια μορφή με το "πατρικό" δίκλωνο μόριο DNA.

Η διαδικασία της αντιγραφής γίνεται την ίδια στιγμή σε πολλές αφετηρίες ενώ είναι πολύ γρήγορη, προστίθενται περίπου 50 νουκλεοτίδια ανά δευτερόλεπτο. Μια αναπόσπαστη διεργασία της αντιγραφής είναι η "επιδιόρθωση δοκιμών", δηλαδή ο μηχανισμός ελέγχου της πιστότητας αντιγραφής, η DNA πολυμεράση πριν προστεθεί το καινούργιο νουκλεοτίδιο ελέγχει αν το προηγουμένως τοποθετημένο είναι σωστά ζευγαρωμένο. Αν είναι λάθος τοποθετημένο το αφαιρεί και επαναλαμβάνει το ζευγάρωμα. Η ακρίβεια είναι πολύ σημαντική για την αντιγραφή του γενετικού υλικού, αφού διασφαλίζει ότι όλα τα κύτταρα του οργανισμού έχουν την ίδια γενετική πληροφορία και ότι αυτή η πληροφορία θα περάσει με ακρίβεια σε τυχόν απογόνους (5).

Μεταγραφή. Στη διεργασία της μεταγραφής, τα ένζυμα RNA πολυμεράσες είναι υπεύθυνα για να καταλύουν την προσθήκη νουκλεοτιδίων, ένα την φορά, από τον εξεταζόμενο κλώνο DNA, σε ένα συμπληρωματικό κλώνο RNA. Οι ίδιες οι RNA πολυμεράσες δεν μπορούν να αναγνωρίσουν μια νουκλεοτιδική αλληλουχία που ονομάζεται υποκινητής ή προαγωγέας, αυτή είναι η δουλειά πρωτεϊνών που ονομάζονται μεταγραφικοί παράγοντες. Μόλις αναγνωριστεί ο υποκινητής και προσδεθεί ισχυρά η RNA πολυμεράση σε αυτόν, η διεργασία της μεταγραφής ξεκινάει. Ο υποκινητής παρέχει τις πληροφορίες για τη θέση έναρξης της για τη σύνθεση του μορίου RNA και ποιος από τους δύο κλώνους θα χρησιμοποιηθεί.

Από το καλούπι κλώνου DNA δημιουργείται το νέο μόριο RNA με κατεύθυνση 5΄ άκρη προς 3΄ άκρη, χρησιμοποιώντας ως πρώτη ύλη τριφωσφορικά νουκλεοτίδια. Η κατάλυση από την πολυμεράση γίνεται με την αντίδραση:

$RNA_n + NTP \rightarrow RNA_{n+1} + PP_i$ **Χημική Αντίδραση 1.** Μεταγραφή του DNA.

Στην οποία τα τριφωσφορικά ριβονουκλεοσίδια (nucleoside triphosphates ή NTPs), έτσι ονομάζεται το κομμάτι του ριβονουκλεοτίδιου που δεν περιέχει τις φωσφορικές ομάδες, χωρίζονται σε μονοφωσφορικά νουκλεοσίδια και πολυμερίζονται στην αλυσίδα. Με μια δεύτερη αντίδραση:

$PP_i \rightarrow 2P_i$ **Χημική Αντίδραση 2.** Πολυμερισμός φωσφορικών ριζών.

Η οποία, καταλύεται από το ένζυμο πυροφωσφατάση (pyrophosphatase), υδρολύεται (γίνεται διάσπαση του ομοιοπολικού δεσμού με προσθήκη νερού) σε ανόργανη φωσφατάση(P_i) (2).

Πίνακας 1. Οι πέντε κλάσεις ευκαρυωτικής πολυμεράσης RNA (6).

RNA πολυμεράση	Περιοχή	Σύνθεση RNA
I	Πυρηνίσκος	Όλα τα rRNAs, εκτός από το 5S rRNA.
II	Κυτταρόπλασμα	Το μεγαλύτερο μέρος των κωδικοποιών πυρηνικών pre-mRNAs (σε προκαρυωτικά mRNAs) και πολλά μη κωδικοποιά RNAs, όπως τα lncRNAs, τα περισσότερα miRNAs και τα μικρά πυρηνικά RNAs (snRNAs).
III	Κυτταρόπλασμα	Διάφορα μικρά μη κωδικοποιά RNAs όπως 5S rRNAs, tRNAs και snRNAs.
IV, V	Μιτοχόνδρια και Πλαστίδια.	Τα RNAs που μεταγράφονται από μιτοχονδρικό DNA και τα μικρά παρεμποδιστικά RNA (siRNAs).

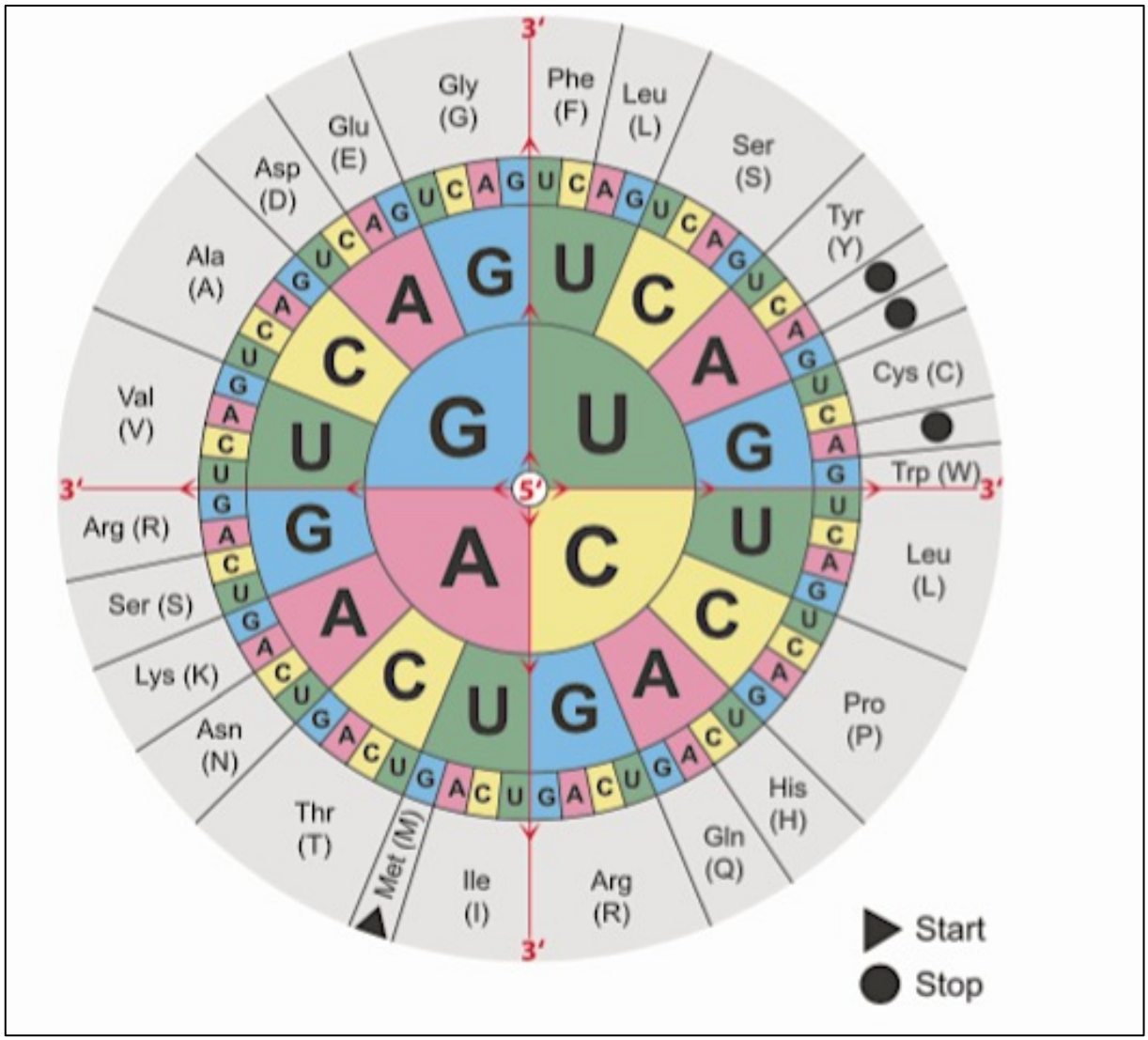
Από το σημείο έναρξης, η RNA πολυμεράση ξετυλίγει προσωρινά την διπλή έλικα DNA που βρίσκεται ακριβώς μπροστά της και χρησιμοποιεί τον ένα κλώνο του ως καλούπι και προσθέτει συμπληρωματικά ριβονουκλεοτίδια, κινούμενη με κατεύθυνση από το 3΄ άκρο

προς το 5΄ άκρο. Μόλις περάσει η πολυμεράση ένα κομμάτι του κλώνου DNA έρχεται στην αρχική του μορφή, δηλαδή του διπλού έλικα. Οι πολυμεράσες RNA μπορούν να ενσωματώσουν 20 με 50 νουκλεοτίδια το δευτερόλεπτο σε ένα μόριο RNA, ενώ πολλά γονίδια του κυττάρου μεταγράφονται παράλληλα από πάνω από 100 πολυμεράσες. Ο Πίνακας 1 παρουσιάζει τα διάφορα είδη του ενζύμου RNA πολυμεράσης και τα είδη RNA που συνθέτουν αυτές. Για την επιδιόρθωση σφαλμάτων (proofreading), το ένζυμο της RNA πολυμεράσης μπορεί να σταματήσει σε κάποιο σημείο ή να γυρίσει κάποια νουκλεοτίδια πίσω, έτσι ώστε να διορθώσει κάποιο λάθος τοποθετημένο νουκλεοτίδιο, σχετικά σπάνιο.

Το μόριο του αγγελιοφόρου RNA (mRNA) στα προκαρυωτικά κύτταρα μπορεί να ξεκινήσει άμεσα την μετάφραση, αφού βρίσκονται στο κυτταρόπλασμα, το οποίο περιέχει τα ριβοσώματα πάνω στα οποία γίνεται η πρωτεϊνοσύνθεση. Αντίθετα στα ευκαρυωτικά κύτταρα μετά την σύνθεσή του RNA στον πυρήνα από το DNA, επεξεργάζεται και μετά μεταφέρεται για πρωτεϊνοσύνθεση στο κυτταρόπλασμα. Το κυτταρόπλασμα είναι μια διαφανής ουσία που περικλείεται από την κυτταρική μεμβράνη και δεν υπάρχει στον πυρήνα των ευκαρυωτικών κυττάρων. Ένα είδος επεξεργασίας του βοηθάει τα ριβοσώματα, δηλαδή της μηχανής παρασκευής πρωτεϊνών, να διαφοροποιήσουν μεταξύ του RNA και του mRNA. Ο σχηματισμός καλύπτρας και η πολυαδενυλίωση (APA) είναι οι διεργασίες με τις οποίες γίνεται αυτή η επεξεργασία, δηλαδή με απλά λόγια είναι η προσθήκη νουκλεοτίδιων αντίστοιχα στην αρχή (5’ άκρο) και στο τέλος (3’ άκρο) του μορίου.

Ένα άλλο είδος επεξεργασίας, που είναι απαραίτητο στα ευκαρυωτικά κύτταρα, είναι η αφαίρεση των εσονίων, μη κωδικοποιών περιοχών μέσα στην αλληλουχία. Πριν την αποκοπή αυτή των εσονίων το μόριο ονομάζεται και pre-mRNA Οι κωδικοποιές περιοχές, από την άλλη πλευρά, ονομάζονται εξόνια και είναι οι περιοχές έκφρασης του γονιδίου. Η αποκοπή των εσονίων και, στη συνέχεια, η συρραφή των εξονίων σε μια συνεχόμενα κωδικοποιά αλληλουχία, είναι μια διεργασία δύο βημάτων και ονομάζεται συρραφή RNA (RNA splicing). Η διεργασία αυτή μας δίνει το “ώριμο” mRNA. Μια εξήγηση για την παρουσία των εσονίων είναι ότι καθιστά δυνατό ένα γονίδιο να κωδικοποιεί για διαφορετικές πρωτεΐνες, που ονομάζονται γι’ αυτό τον λόγο ισομορφές (isoform) πρωτεϊνών. Η εναλλακτική αυτή συρραφή γίνεται με την συρραφής των εξονίων σε διαφορετική σειρά και παράγει ισόμορφα πρωτεϊνών με διαφορετική λειτουργικότητα.

Μετάφραση. Το μεταφορικό μόριο RNA (tRNA) έχει την δυνατότητα να μετατρέπει την πληροφορία από την "γλώσσα" των νουκλεϊνικών οξέων, τα νουκλεοτίδια, στην "γλώσσα" των πρωτεϊνών, τα αμινοξέα. Επειδή τα νουκλεοτίδια είναι 4 και τα αμινοξέα είναι 20, οι επιστήμονες στην αρχή δεν καταλάβαιναν των μηχανισμό αυτής της μετάφρασης. Ο μηχανισμός αυτής της μετάφρασης αποτελεί τον γενετικό κώδικα. Πιο συγκεκριμένα, το μόριο tRNA μπορεί να συνδέει μια ομάδα τριών διαδοχικών νουκλεοτιδίων, δηλαδή ένα κωδικόνιο, με ένα αμινοξύ. Βέβαια το ίδιο αμινοξύ μπορεί να καθορίζεται από διαφορετικά κωδικόνια. Στην Εικόνα 3 φαίνεται το "λεξικό" του γενετικού κώδικα και ότι 61 τριπλέτες κωδικοποιούν για αμινοξέα.



Εικόνα 3. Τα αμινοξέα και τα νουκλεοτίδια από τα οποία αποτελούνται (7).

Μια από αυτές, η τριπλέτα AUG, όχι μόνο κωδικοποιεί για το αμινοξύ Μεθιονίνη, αλλά μπορεί να σηματοδοτεί και την αρχή της πολυπεπτιδικής αλυσίδας. Τρία κωδικόνια δεν

καθορίζουν κάποιο αμινοξύ αλλά σηματοδοτούν το τέλος της κωδικοποιάς αλληλουχίας. Ένα άλλο σημαντικό στοιχείο της μετάφρασης είναι το ανοιχτό πλαίσιο ανάγνωσης (Open Reading Frame, ORF), στο οποίο μια μετάλλαξη μπορεί να αποδειχθεί καταστροφική. Το πλαίσιο ανάγνωσης ασχολείται με τα συγκεκριμένα σημεία από τα οποία αρχίζει το διάβασμα των τριπλετών νουκλεοτιδικών βάσεων και συνεχίζει συνεχόμενα μέχρι το τέλος του πλαισίου. Με εισαγωγή ενός ζεύγους βάσης, δηλαδή, μπορεί να αχρηστευτεί το γενετικό μήνυμα. Με μια διορθωτική κίνηση αφαίρεσης ενός ζεύγους βάσης στην συνέχεια ο γενετικός κώδικας βρίσκει πάλι το νόημά του, από το σημείο της διόρθωσης και πέρα.

Η μετάφραση ξεκινάει, στα περισσότερα βακτήρια και σε όλα τα ευκαρυωτικά κύτταρα, όταν το κωδικόνιο AUG του μορίου mRNA συναντά το αντικωδικόνιο του μορίου tRNA, το οποίο είναι συμπληρωματικό ως προς το κωδικόνιο του μορίου mRNA, στην περιοχή P του ριβοσώματος. Τα μόρια tRNA έχουν τις δύο πολύ σημαντικές λειτουργίες να μεταφέρουν τα σωστά αμινοξέα και να αναγνωρίζουν τα σωστά κωδικόνια στο mRNA.

Πιο αναλυτικά, το ριβόσωμα είναι ένα μεγάλο σύμπλοκο που συντονίζει την λειτουργία των μορίων mRNA και tRNA και παράγει τα πολυπεπτίδια. Το οργανίδιο αυτό του κυτταροπλάσματος αποτελείται από δύο υποενότητες που αποτελούνται από μια πληθώρα από ριβοσωμικές πρωτεΐνες και ριβοσωμικό RNA (rRNA). Το ριβόσωμα έχει στη μικρή υποενότητα του την περιοχή πρόσδεσης του mRNA. Ενώ στην μεγάλη υποενότητα του έχει περιοχές πρόσδεσης των μορίων tRNA, την περιοχή P και την περιοχή A. Η περιοχή P έχει το tRNA με την αυξανόμενη πολυπεπτιδική αλυσίδα και η περιοχή A έχει το tRNA που θα προστεθεί στη συνέχεια στην πολυπεπτιδική αλυσίδα. Το αντικωδικόνιο του tRNA ζευγαρώνει με το κωδικόνιο στο mRNA, και η διαδικασία επαναλαμβάνεται μέχρι το κωδικόνιο τερματισμού να φτάσει στην περιοχή A του ριβοσώματος. Τότε το πολυπεπτίδιο ελευθερώνεται και το ριβόσωμα γυρνάει στην αρχική του κατάσταση (8).

1.3. Τα microRNAs ως Δευτερογενής (Επιγενετικός) Μηχανισμός Ελέγχου

Γενικά, τα μη κωδικοποιά RNAs παίζουν το ρόλο επιγενετικών τροποποιητών. Τα μικρά μη κωδικοποιά RNA (Small Non-coding RNAs, sncRNA), μόρια με μήκος μικρότερο των 200 νουκλεοτιδίων, είναι η οικογένεια μη κωδικοποιών μορίων στην οποία ανήκουν τα microRNAs (miRNAs ή miRs). Η άλλη οικογένεια μη κωδικών RNA είναι τα μακρά μη κωδικοποιά μόρια RNA (Long Non-coding RNAs, lncRNAs).

Τα microRNAs είναι ενδογενή μονόκλωνα μόρια RNA που αποτελούνται από περίπου 22 νουκλεοτίδια τα οποία βοηθάνε στην μετα-μεταγραφική διαμόρφωση της ρύθμισης της γονιδιακής έκφρασης σε σημαντικές λειτουργίες του κυττάρου, οπότε έχουν μεγάλες επιπτώσεις σε αρκετές ασθένειες, από καρκίνο μέχρι καρδιακή ανεπάρκεια. Το κάθε miRNA είναι δυνατό να ρυθμίσει αρκετές εκατοντάδες γονιδιακές εκφράσεις, ενώ κάθε mRNA μπορεί να διαθέτει περιοχές αναγνώρισης και πρόσδεσης πολλών διαφορετικών miRNAs. Για παράδειγμα, η απορρύθμιση των miRNAs στα επιδερμικά κύτταρα πιθανώς επιφέρει πρόωρη γήρανση του δέρματος, μελάνωμα και άλλες διαταραχές. Πιο συγκεκριμένα, τα miRNAs συνηθέστερα προσδένονται στην 3΄-αμετάφραστη περιοχή (3΄-UTR) γονιδίων στόχων, με αποτέλεσμα είτε την καταστολή της μετάφρασης είτε την υποβάθμιση του mRNA, μέσω του σχηματισμού ριβονουκλεοπρωτεϊνικών συμπλόκων. Οι τροποποιήσεις που επιφέρουν τα miRs προάγουν κάποιες διεργασίες βασικές για την ανάπτυξη, όπως αυτές του κυτταρικού πολλαπλασιασμού, της κυτταρικής διαφοροποίησης και της προγραμματισμένης κυτταρικής απόπτωσης.

Απομόνωση miRNAs. Από την ανακάλυψή τους το 1993 μέχρι και σήμερα, ένα μεγάλο εμπόδιο στην μελέτη των miRNAs με τεχνικές εξαιρετικά υψηλών αποδόσεων είναι η εξαγωγή και απομόνωση υψηλής επαρκώς ποιότητας δειγμάτων. Τα miRNAs που μπορεί να απομονωθεί από τα κύτταρα και τους ιστούς είναι μόλις μέχρι 1μg, ενώ τα miRNAs συνήθως εκπροσωπούν μόνο ένα μικρό κομμάτι του συνολικού RNA του κυττάρου. Ακόμα, μπορούν να εξετασθούν οι αλληλεπιδράσεις των miRNAs με άλλα μακρομόρια χρησιμοποιώντας τεχνικές ανοσοκαθίζησης (immune-precipitation, IP) (9).

Ο συμπληρωματικότητα των miRNAs. Τα miRNAs προσδένονται σε στόχους mRNA μέσω την συμπληρωματικής αλληλουχίας και αναστείλουν την μετάφραση αυτών των στόχων ή προωθούν την αποδόμηση τους. Έχει εκτιμηθεί υπολογιστικά ότι στον άνθρωπο τα miRNAs στοχεύουν περίπου 60% των κωδικοποιών γονιδίων μέσω συντηρημένης (conserved) ζευγάρωσης μεταξύ της 3'-αμετάφραστης περιοχής του mRNA και της 5' περιοχής του miRNA, αυτή η ονομάζεται περιοχή εκβλάστησης (seed region, SR). Σε αυτή την περιοχή γίνεται η υβριδοποίηση της αλληλουχίας εκβλάστησης, δηλαδή αλληλουχίας των 2-8 νουκλεοτιδίων του miRNA, η οποία αναγνωρίζει τις ταυτίσεις εκβλάστησης ανάλογα με τον αλγόριθμο που χρησιμοποιείται από ένα υπολογιστικό πρόγραμμα, όπως το MiRanda.

Εδώ είναι σημαντική η αξιόπιστη αναγνώριση των στόχων τους και ο χαρακτηρισμός των λειτουργιών των miRNAs. Στα φυτικά κύτταρα προσδένονται τα miRNAs στους στόχους mRNA με ολική συμπληρωματικότητα, στην οποία περίπτωση η αποικοδόμησή των στόχων γίνεται μέσω RNAi (RNA interference), ενώ στα ζωικά κύτταρα συνήθως προσδένονται στους στόχους mRNA με ατελή συμπληρωματικότητα. Γι' αυτόν τον λόγο, στα ζωικά κύτταρα, είναι δύσκολη η στόχευση με υψηλή ειδικότητα και η λειτουργία τους εξαρτάται από τον βαθμό συμπληρωματικότητας. Η ατελής συμπληρωματικότητα της πρόσδεσης οδηγεί σε καταστολή της μετάφρασης ή στην αποαδενυλίωση του mRNA στόχου, ενώ η ολική συμπληρωματικότητα στην αποικοδόμηση του μεταγράφου στόχου (10, 11).

Η βιοσύνθεση των miRNA. Η σύνθεση ενός τυπικού miRNA γίνεται με την μεταγραφή ενός εσονίου RNA μέσω της πολυμεράσης RNA II και σε λίγες περιπτώσεις μέσω τις πολυμεράσης RNA III. Το pri-miRNA (primary miRNA), δηλαδή το πρωταρχικό μετάγραφο, είναι ένα μονόκλωνο μόριο RNA με μήκος συνήθως αρκετών κιλοβάσεων (kb), που αποτελούνται από την καλύπτρα στο 5'-άκρο και την 3' πολυαδενυλιωμένη (APA) ουρά. Αυτό αναδιπλώνεται σε μια δευτεροταγή δομή με σχήμα μίσχου-θηλιάς (stem-loop) ή αλλιώς φουρκέτας (hairpin), μήκους 33 με 35 ζευγών βάσεων. Μετά από διάσπαση από το ένζυμο Drosha, που ανήκει στις RNάσες III, δημιουργείται ένα μικρότερου μήκους δίκλωνο pre-miRNA, το οποίο εκκρίνεται από τον πυρήνα στο κυτταρόπλασμα (12).

Ακόμα, υπάρχουν τα mirtrons, τα οποία είναι όμοια σε δομή με τα pre-miRNAs και προσπερνάνε την πρωταρχική επεξεργασία των κλασικών εσονίων miRNAs. Το miRNA διαμορφώνεται από το pre-miRNA, με την βοήθεια του ενζύμου Dicer, που ανήκει επίσης στις RNάσες III. Έτσι, από την μια πλευρά της φουρκέτας παίρνουμε ένα ευσταθή μονόκλωνο "ώριμο" μόριο miRNA. Το οποίο συνήθως είναι υψηλώς διατηρητέο σε ένα πληθυσμό. Τα "ώριμα" μόρια miRNA ενσωματώνονται μαζί με πρωτεΐνες στο σύμπλοκο RISC (RNA induced silencing complex), το οποίο "αποσιωπεί" τον στόχο mRNA. Πιο συγκεκριμένα, το miRNA του σύμπλοκου RISC οδηγεί τις σχετικές πρωτεΐνες Ago, που είναι υποομάδα των πρωτεϊνών Argonaute, σε κοντινή απόσταση στο προσδεδεμένο mRNA. Η πρωτεΐνη Argonaute προκαλεί είτε αποικοδόμηση του mRNA είτε καταστολή της μετάφρασής του. Ο μηχανισμός «αποσιώπισης» είναι πιθανόν να εξαρτάται από τον αριθμό, τον τύπο και την θέση των ασύζευκτων νουκλεοτιδίων μεταξύ των μορίων mRNA και miRNA.

Μηχανισμοί λειτουργίας των miRNAs. Τα γονίδια των miRNA έχουν αναγνωριστεί με διάφορους τρόπους, συνήθως όμως από βιοπληροφορική ανάλυση γενωμικών αλληλουχιών DNA, από την απομόνωση μεταλλάξεων και μικρών κυτταρικών RNAs. Είναι μάλλον σίγουρο ότι τα miRNAs παίζουν σημαντικό ρόλο στους πολυκύτταρους οργανισμούς για την διαφοροποίηση και την διατήρηση τους σε συγκεκριμένους τύπους κυττάρων, αφού γι' αυτό τον λόγο δεν υπάρχουν σε μονοκύτταρους οργανισμούς. Τα miRNAs παίζουν ρόλο ακόμα και στα πρωταρχικά στάδια της εμβρυονικής ανάπτυξης, αφού όταν λείπει το απαραίτητο ένζυμο Dicer σε αναπτυσσόμενο ζωικό οργανισμό αυτός δεν αναπτύσσεται πέρα από την γαστριδίωση. Αλλά και από όποιον ιστό λείπει το συγκεκριμένο ένζυμο, παρουσιάζει ανωμαλίες η ανάπτυξη. Ενώ η αστάθεια στα επίπεδα miRNA παίζει ρόλο σε πολλές κοινές ασθένειες. Ακόμα, μερικά miRNAs προσδένονται στα ORFs και στις 5΄UTR των mRNA στόχων και ενεργοποιούν ή καταστέλλουν τη μετάφρασή τους (13). Τα miRNAs είναι μάλλον σημαντικοί διαμεσολαβητές των αποκρίσεων στο στρες. Για παράδειγμα, μπορεί να διευκολύνει μια γρήγορη απόκριση στο στρες μέσω καταστολής της μετάφρασης συγκεκριμένων mRNAs, τα οποία παραμένουν σε P-σώματα στο κυτταρόπλασμα, μέχρι κάποιο εξωτερικό ερέθισμα να επιτρέψει στην μετάφραση να συνεχίσει (2). Ενώ στον πυρήνα, τα miRNAs μπορούν να προσδένονται στους υποκινητές των γονιδίων και να ρυθμίζουν τη γονιδιακή έκφραση.

Τα miRNAs, όπως και κάθε γενωμική αλληλουχία, είναι επιρρεπής σε πολυμορφισμούς SNPs. Τα miSNPs, όπως είναι γνωστά, μπορούν να δημιουργήσουν νέες ή να καταστρέψουν υπάρχουσες περιοχές πρόσδεσης των miRNAs, περιορίζοντας την ικανότητα τους για γονιδιακή ρύθμιση. Τα miSNPs έχουν χρησιμοποιηθεί εκτενώς για την διάγνωση βιοδεικτών στην φαινυλκετονουρία (PKU), αλλά βοηθάνε και γενικά στην κατανόηση των ρυθμιστικών δικτύων των miRNA. Γενικότερα, τα επίπεδα mRNA συσχετίζονται αντιστρόφως με τα επίπεδα του συμπληρωματικού του miRNAs, δηλαδή το πλήθος των mRNAs με τις θέσεις πρόσδεσης κατάλληλες για το συγκεκριμένο εισερχόμενο miRNA μειώνονται. Σε πείραμα μείωσης ενός συγκεκριμένου miRNA, το πλήθος mRNAs με τις θέσεις πρόσδεσης κατάλληλες για το συγκεκριμένο miRNA αυξάνεται. Ακόμα, πειράματα στα οποία κύτταρα εξαναγκάζονται να εκφράσουν συγκεκριμένα γονίδια miRNA θέλουν να δείξουν ότι αυτά μπορούν να μειώσουν τα ποσοστά έκφρασης συγκεκριμένων ομάδων mRNA (2).

Τα miRNAs δρούν ανασταλτικά στην μετάφραση των mRNAs, αφού παίζουν ρόλο στην ενεργοποίηση της. Όταν εκκρίνονται στον εξωκυττάριο χώρο λειτουργούν ως

διαμεσολαβητές για την κυτταρική επικοινωνία και για την ανοσολογική ρύθμιση (14, 15). Ακόμα, τα miRNAs μπορούν να δράσουν με την καταστολή της επιμήκυνσης της αλληλουχίας κατά την μετάφραση. Η υπερέκφραση ενός miRNA, που έχει ρόλο ογκογόνου, μπορεί να πάρει πολλές μορφές, δηλαδή μπορεί να ευθύνεται η παραγωγή αυξημένης ποσότητας του miRNA ή να παράγεται σε ιστό που φυσιολογικά δεν υπάρχει ή απλά να γίνεται σε λάθος χρονική στιγμή. Σε ογκοκατασταλτικά miRNAs έχουν βρεθεί αυξημένα επίπεδα μεθυλίωσης με αποτέλεσμα την υπερέκφραση των ογκογόνων στόχων τους (16). Γενικά, τα miRNAs στοχεύουν ένζυμα που εμπλέκονται στην επιγενετική ρύθμιση και επηρεάζουν ολόκληρο το γένωμα.

1.4. Μεθυλίωση και Επιγενετική

Επιγενετική. Ο όρος επιγενετική εισήχθηκε το 1939 από τον επιστήμονα Conrad Weddington, ενώ ο Andrian Bird έδωσε τον εξής ορισμό της επιγενετικής "Φαινόμενο που περιλαμβάνει κληρονομούμενες γενετικές αλλαγές στην γονιδιακή έκφραση όπου δεν μεταβάλλεται η νουκλεοτιδική αλληλουχία αλλά επιδρά στην διαμόρφωση του DNA".

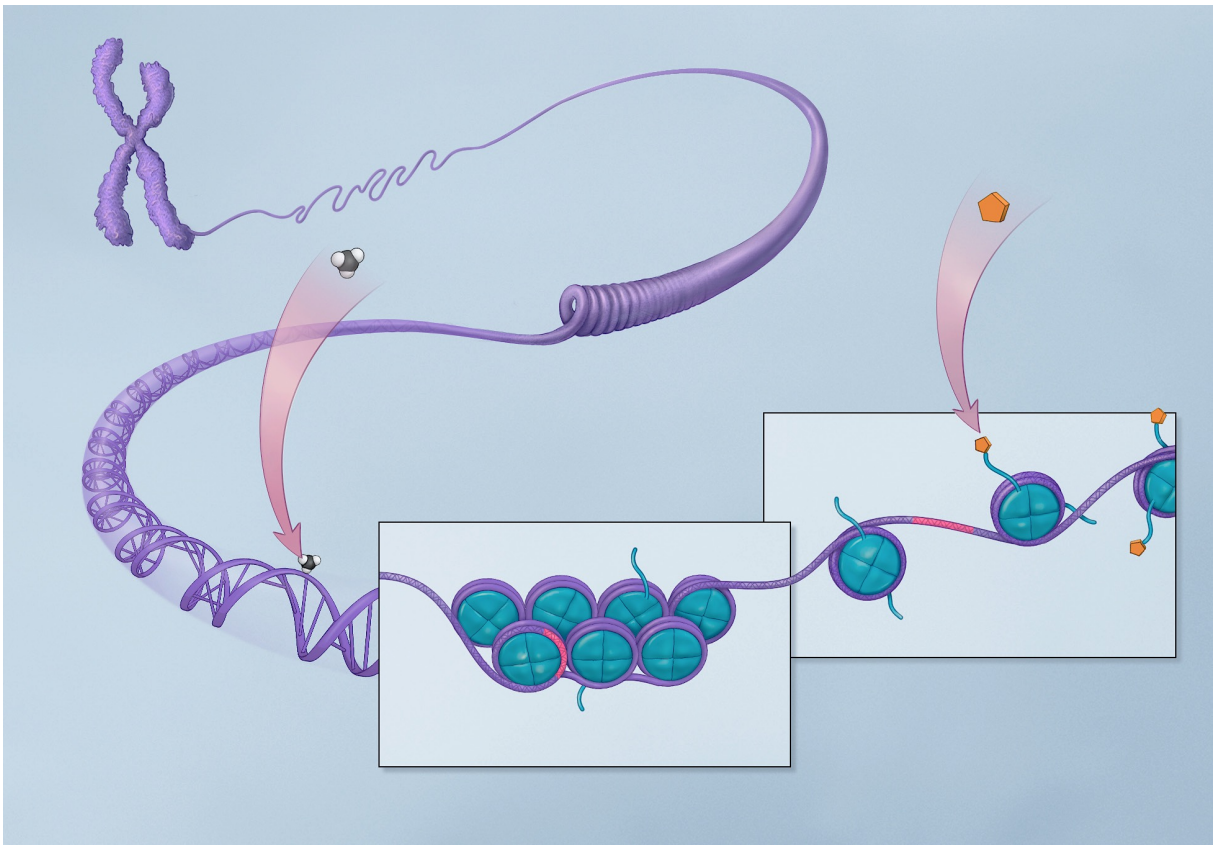
Οι επιγενετικές τροποποιήσεις κρύβουν πολλά από τα μυστικά της ζωής και οι κυριότερες είναι η μεθυλίωση του DNA, η διαφοροποιήσεις ιστονών και η δυσλειτουργία των miRNAs. Κάποιοι άλλοι μοριακοί μηχανισμοί, που εμπλέκονται στην επιγενετική είναι οι μετα-μεταγραφικές τροποποιήσεις (Posttranslational Modifications, PTM), τα ATP-εξαρτώμενα σύμπλοκα αναδιαμόρφωσης της χρωματίνης, polycomb/trithorax σύμπλοκα πρωτεϊνών, διάφορα μη-κωδικοποιά RNAs, για παράδειγμα siRNA. Οι PTM είναι τροποποιήσεις που παίζουν ρόλο σχεδόν σε κάθε λειτουργία σχετική με το DNA, π.χ. η δομή και οργάνωση του γενώματος, η γονιδιακή έκφραση, η αντιγραφή και επιδιόρθωση του DNA, ο κυτταρικός κύκλος και η απόπτωση.

Οι τροποποιήσεις στις ιστόνες έχουν δυναμικό χαρακτήρα, αφού μπορούν να προστεθούν ή να αφαιρεθούν από διάφορα ένζυμα, οπότε το επίπεδο πολυπλοκότητας των τροποποιήσεων είναι αυξημένο. Αυτές που δεν έχουν δομικό ρόλο αποτελούν σημεία πρόσδεσης στη μετάφραση των πρωτεϊνών και έμμεσα σηματοδοτούν τους λειτουργικούς ρόλους αυτών. Οι ποικιλόμορφοι επιγενετικοί μηχανισμοί συσχετίζονται και έχουν το ρόλο να σταθεροποιούν ο ένας τον άλλο για να εξασφαλιστεί η πιστότητα των επιγενετικών παραγόντων, ιδιαίτερα κατά την κυτταρική διαίρεση. Η παρουσία επιγενετικών τροποποιήσεων μπορεί να αυξήσει την πιθανότητα εκδήλωσης μεταλλάξεων. Οι επιγενετικές τροποποιήσεις στο ευκαρυωτικό

γονιδίωμα είναι πολύπλοκες και συνεχώς εναλλασσόμενες λόγω του πλήθους των ερεθισμάτων σε αυτά τα κύτταρα.

Οι επιγενετικοί παράγοντες και τα σχετικά ένζυμα έχουν πολύ σημαντικό ρόλο στην λειτουργία, την ταυτότητα και την ανάπτυξη των κυττάρων, αφού ρυθμίζουν τις μεταβολές μεταξύ των βημάτων διαφοροποίησης του κυττάρου (**Εικόνα 4**). Η επιγενετική μελετά τις αλλαγές στο φαινότυπο, όπως τα πρότυπα γονιδιακής έκφρασης συγκεκριμένου κυτταρικού τύπου που δεν προκαλούνται από αλλαγές στην κύρια αλληλουχία DNA. Αυτές οι αλλαγές κληρονομούνται μιτωτικά και, σε κάποιες περιπτώσεις, μειωτικά. Η επιγενετική ρύθμιση βοηθάει στην γενωμική προσαρμογή σε ένα περιβάλλον, αφού μιλάμε για στοχαστικές μεταβολές σύμφωνα με τα ενίοτε περιβαλλοντικά ερεθίσματα.

Επιγενωμική. Σχεδόν όλα τα κύτταρα του οργανισμού έχουν τα ίδια γενετικά υλικά αφού κωδικοποιούνται από την ίδια αλληλουχία DNA, αλλά εμφανίζουν πολλές διαφορές στην μορφολογία και στην λειτουργία τους. Η επιγενωμική αναφέρεται στην μελέτη αυτών των τροποποιήσεων στο σύνολο του γενώματος, οι οποίες είναι αποτέλεσμα χημικών τροποποιήσεων στο DNA ή σχετικών με αυτό ιστονών, όπως απεικονίζεται στην **Εικόνα 4**.



Εικόνα 4. Οι επιγενετικοί μηχανισμοί (17).

Πρόγραμμα χάρτη πορείας της επιγενωμικής (Roadmap Epigenomic Project). Η αλληλούχιση του γενώματος δεν είναι οι μόνες πληροφορίες που μας ενδιαφέρουν. Το έργο χάρτη πορείας προσπαθεί να ερευνήσει και τους επιγενετικούς μηχανισμούς για να έχουμε μια ολοκληρωμένη εικόνα της γονιδιακής έκφρασης, δηλαδή εξετάζει την αρχιτεκτονική του ανθρώπινου επιγενώματος (Πίνακας 2) (3).

WashU Epigenome Browser[1]. Για την οπτικοποίηση επιγενωμικών δεδομένων, π.χ. δεδομένων μεθυλίωσης DNA, μπορούμε να χρησιμοποιήσουμε τον φυλλομετρητή επιγενώματος. Είναι κατάλληλο λογισμικό για να αναλύονται βιοπληροφορικά ακόμα και από μη εξοικειωμένους σε αυτές τις αναλύσεις βιολόγους και προσφέρεται ως ελεύθερο λογισμικό από το Πανεπιστήμιο της Ουάσινγκτον στο Σεντ Λούι.

Πίνακας 2. Σύγκριση γενετικής και επιγενετικού μηχανισμού στην εξέλιξη (18).

Σύγκριση των δύο μηχανισμών	**Γενετική**	**Επιγενετική**
Η συχνότητα που συμβαίνουν	Πολύ σπάνια	Πολύ συχνά
Η επίδραση στο περιβάλλον	Αργή και πιθανώς τυχαία	Άμεση
Η κατεύθυνση των αλλαγών	Πιθανώς ουδέτερη	Κατευθυνόμενη
Η επιλεκτικότητα των αποκρίσεων	Μη επιλεκτική	Πολύ επιλεκτική
Η αντιστρεψιμότητα των αλλαγών	Σπάνια αντιστρέψιμη	Συνήθως αντιστρέψιμη
Οι απαιτήσεις μιας συνεχόμενης	Απαραίτητη για μια	Απαραίτητη για την
περιβαλλοντικής πίεσης	επιλογή	διατήρηση των αλλαγών
Το κόστος στον οργανισμό	Πολύ χαμηλό(εκτός των	Πολύ υψηλό
	βλαβερών μεταλλάξεων)	
Μακροχρόνιες λύσεις	Τυπικές	Προβληματικές

Με την Chip-seq μελετώνται οι μεταβολές στις ιστόνες. Η χρωματίνη (Chromatin) είναι ένα σύμπλοκο DNA και πρωτεϊνών που οργανώνεται σε σταθερά σύμπλοκα και μετά περνάει από ενζυματική κατεργασία ή ψαλίδισμα για να σπάσει η χρωματίνη σε μικρότερα κομμάτια. Κάποια τμήματα DNA προσδεδεμένα σε ιστόνες (Histones) επιλέγονται μέσω αντισωμάτων που στοχοποιούν συγκεκριμένες PTM. Αυτά τα τμήματα αλληλουχοποιούνται και χαρτογραφούνται σε ένα γένωμα αναφορά. Έτσι, δημιουργούν ένα χάρτη με τροποποιήσεις ιστονών στο γένωμα, οι οποίες χρησιμοποιούνται στην ρύθμιση των γονιδίων και την επιδιόρθωση του γενετικού υλικού.

[1] https://epigenomegateway.wustl.edu/

Η πιο μελετημένη επιγενετική τροποποίηση είναι η μεθυλίωση DNA, αφού από σχετικά νωρίς συνδέθηκε με το γήρας και τον καρκίνο, δύο τομείς που ενδιαφέρουν πάντα τους ερευνητές. Η αυξημένη γενωματική μεθυλίωση βοηθάει στην ευστάθεια και την εξέλιξη του γονιδιώματος (18).

Μεθυλίωση. Παρόλο που η μεθυλίωση του DNA είναι ένας σταθερός επιγενετικός μηχανισμός, η μεταφορά των επιγενετικών δεικτών από το πατρικό κύτταρο σε θυγατρικά εξαρτάται από την γονιδιακή ρύθμιση. Μεταξύ των σταδίων της γονιμοποίησης και των πρώτων κυτταρικών διαιρέσεων του γονιμοποιημένου ωαρίου λαμβάνει χώρα η πρώτη μεγάλη αλλαγή στα επίπεδα της επιγενετικής έκφρασης, με την απομάκρυνση των μαρκαρισμένων μεθυλιώσεων από το DNA. Η δεύτερη μεγάλη αλλαγή λαμβάνει χώρα όταν το έμβρυο μεταφερθεί στην μήτρα, όπου νέες μεθυλιώσεις διαδίδονται μέσα στα κύτταρα με αποτέλεσμα να εμφανιστεί ένας νέος «χάρτης» μεθυλιώσεων. Ακόμα δεν είναι γνωστός ο λόγος που μια συγκεκριμένη αλληλουχία γονιδίου μεθυλιώνεται ή όχι, όμως μη φυσιολογικά πρότυπα μεθυλίωσης έχουν παρατηρηθεί σε περίπτωση ασθένειας (2).

Η μεθυλίωση του DNA παίζει κεντρικό ρόλο σε αρκετές κριτικές αντιδράσεις στον οργανισμό, άρα μπορεί να μελετηθεί και με πολλούς διαφορετικούς τρόπους. Η χρήση της τεχνολογίας SMRT, από την εταιρία Pacific Biosciences, είναι η πιο ενδεδειγμένη λύση επειδή δεν χρησιμοποιεί την μέθοδο της κλωνοποίησης αλυσιδωτής αντίδρασης πολυμεράσης (PCR), που θα δούμε πιο αναλυτικά στην συνέχεια, οπότε δεν αλλοιώνονται οι τροποποιήσεις στο DNA και δεν χρησιμοποιεί γένωμα αναφοράς. Η μέθοδος του έχει να κάνει με το ποσοστό της ενσωμάτωσης βάσεων μέσω πολυμερισμού κατά την διάρκεια επιμήκυνσης της αλυσίδας. Το ποσοστό αυτό εξαρτάται από την χημική δομή των βάσεων που ενσωματώνονται και, ακόμα, έχει διαφορά με τις χημικά τροποποιημένες βάσεις, όπως τις μεθυλιωμένες αδενίνες (A) και κυτοσίνες (C).

Μια λιγότερο άμεση μέθοδος είναι η δισουλφιδική (bisulfite, διθειώδες) αλληλούχιση, που μελετά την μεθυλίωση σε κυτοσίνες. Το δισουλφιδικό μόριο μετατρέπει την κυτοσίνη σε ουρακίλη με το κατεργασμένο DNA, ενώ η μετατροπή δεν συμβαίνει σε μεθυλιωμένη κυτοσίνη. Η κλωνοποίηση PCR δισουλφιδικά κατεργασμένου DNA θα έχει ως αποτέλεσμα μη μεθυλιωμένες C, οι οποίες θα μετατραπούν σε T, ενώ οι μεθυλιωμένες C θα παραμείνουν C. Οι μικροαναγνώσεις (reads), δηλαδή η αλληλουχίες εξόδου που παίρνουμε από τις τεχνικές εξαιρετικά υψηλής διαλογής, μπορούν να χαρτογραφηθούν και να συγκριθούν με

ένα γένωμα αναφοράς, έτσι ώστε να ανιχνευθούν ποιες κυτοσίνες δεν μετατράπηκαν, μετά την κατεργασία με δισουλφίδια, σε άλλη βάση και έτσι να καταδείξουν ότι ήταν μεθυλιωμένες.

Πολλές φορές η ανάπτυξη καρκίνου οφείλεται στην αλλοιωμένη μεθυλίωση που οδηγεί σε αποσιώπηση ογκοκατασταλτικών γονιδίων. Το γενωμικό εντύπωμα είναι χαρακτηριστικό των θηλαστικών και αναφέρεται στην ικανότητα ενός γονιδίου να εκφράζεται σύμφωνα με το γένος του γονέα από τον οποίο προήλθε, αυτό γίνεται δυνατό μέσω μεθυλίωσης του DNA. Τα φυτά επιδεικνύουν πολύ μεγαλύτερα επίπεδα μεθυλίωσης στο DNA τους από ότι τα ζώα, αλλά και εκεί έχουν λειτουργία αποσιώπησης γονιδίων. Σε ένα χαρακτηριστικό πείραμα σε φυτά, τα οποία είχαν κατεργαστεί με ουσίες για ελαχιστοποίηση της μεθυλίωσης του DNA τους, είχαν αυξημένο αριθμό φυλλώματος και μορφολογικά διαφορετικά λουλούδια.

Η μεθυλίωση DNA ως το βιολογικό ρολόι του οργανισμού. Τα ομοζυγωτικά δίδυμα αποτελούσαν ειδικότερα στο παρελθόν την πλέον κατάλληλη ομάδα για να ελέγξουν οι επιστήμονες τις επιγενετικές τροποποιήσεις. Τέτοιες μελέτες έχουν δείξει ότι μόνο το 20% με 30% των διαφοροποιήσεων μακροζωίας εξαρτώνται από γενετικούς παράγοντες μεταξύ διαφορετικών ανθρώπων. Το υπόλοιπο 70% με 80 % των διαφοροποιήσεων εξαρτώνται από περιβαλλοντικούς παράγοντες (19).

Οι επιγενετικές τροποποιήσεις είναι απαραίτητες για την ανάπτυξη, ενώ όσο προχωράει το γήρας αυτές απορρυθμίζονται. Ένα μεγάλο μέρος της επιγενετικής επιστήμης γίνεται μέσω της ανάλυσης των τροποποιήσεων του DNA και των πρωτεϊνών ιστόνης, αυτοί είναι οι μηχανισμοί που συνήθως επηρεάζουν την δομή και οργάνωση της χρωματίνης. Η προχωρημένη ηλικία λειτουργεί ως επιβαρυντικός παράγοντας για την εμφάνιση καρκίνου, η αλλαγή είναι εμφανής και στα επίπεδα μεθυλίωσης DNA. Συγκεκριμένα, έχει παρατηρηθεί ότι όσο μεγαλώνει ο άνθρωπος υπάρχει μείωση στο ολικό μεθυλιωμένο DNA (20), εκτός τις περιοχής των νήσων CpG islands (CGIs), στην οποία συμβαίνει το αντίθετο. Η υπερμεθυλίωση των νήσων CGIs βρίσκεται κυρίως στην περιοχή υποκινητή γονιδίου βλαστικών κυττάρων και ακόμα στην περιοχή υποκινητή του ογκοκατασταλτικού γονιδίου (21, 22).

Η επιγενετική είναι βιοδείκτης ηλικίας αφού φανερώνει ότι όσο μεγαλώνουμε ελαχιστοποιούνται τα βλαστικά κύτταρα στον οργανισμό, λόγω της αυξημένης μεθυλίωσης στην περιοχή υποκινητή των γονιδίων βλαστικών κυττάρων και αυξάνεται η πιθανότητα εμφάνισης καρκίνου, λόγω της υπερμεθυλίωσης σε υποκινητές των ογκοκατασταλτικών γονιδίων (23). Για παράδειγμα, σε πειράματα με ηλικιωμένους αρουραίους, εμφανίστηκε μια αύξηση στις 5-μέθυλο-κυτοσίνες μέσα στα

σύμπλοκα ριβοσωμικού DNA (ribosomal DNA, rDNA) στο συκώτι τους, που μπορεί να εξηγήσει την μείωση στα επίπεδα ριβοσωμικού RNA (rRNA) σύμφωνα με το γήρας του αρουραίου (24).

Ένας άλλος ρόλος της επιγενετικής μπορεί να είναι και ως δείκτης για την ηλικία του οργανισμού, όπως έδειξε μελέτη σε διάφορους ανθρώπινους ιστούς (25, 26). Επιπλέον τα πρότυπα μεθυλίωσης DNA συνήθως διατηρούνται σε ένα πληθυσμό και υπό διάφορες συνθήκες (27-29). Ένα παράδειγμα που παρατηρείται σε πολλές μορφές καρκίνου είναι η υπομεθυλίωση ολικού DNA σε περιοχές που επαναλαμβάνονται συχνά (30), γιατί αυτή η κατάσταση συμβάλει στην κυτταρική μεταμόρφωση προωθώντας χρωμοσωμικές ανακατατάξεις και αυξάνοντας το πλήθος των μεταλλάξεων. Η περιοχή υποκινητή των γονιδίων που στοχοποιούνται για την υπερμεθυλίωση DNA κατά την προχωρημένη ηλικία πολλές φορές είναι τα ίδια με τις υπερμεθυλίωσης σε πολλούς καρκίνους (31). Για παράδειγμα, έχουν ανιχνευθεί αλλοιώσεις υπερμεθυλίωσης κατά το γήρας στους υποκινητές ογκοκατασταλτικών γονιδίων, όπως των LOX, RUNX3, TIG1, και είναι μια ένδειξη ότι η εμφάνιση συγκεκριμένων τύπων καρκίνου εξαρτώνται από την ηλικία (32).

Μετά από εξέταση του DNA θηλαστικών και άλλων σπονδυλωτών αποδεικνύεται ότι ακόμα και 1 στα 100 νουκλεοτίδια είναι μεθυλιωμένα. Η μεθυλομάδα είναι πάντα πάνω στο carbon 5 μιας κυτοσίνης. Οι μεθυλομάδες προστίθενται στο DNA από τα ένζυμα DNA μεθυλοτρανσφεράσες (DNA methyltransferases, DNMT). Αυτή η σχετικά απλή χημική τροποποίηση λειτουργεί ως επιγενετικό σημάδι, επιτρέποντας την αναγνώριση συγκεκριμένων περιοχών του DNA και την διαφοροποίηση της λειτουργίας τους από άλλες περιοχές. Το πρότυπο της μεθυλίωσης πρέπει να διατηρείται σε επαναλαμβανόμενες κυτταρικές διαιρέσεις, αυτό τον ρόλο παίζει το ένζυμο DNMT1. Συγκεκριμένα, κατά την διαδικασία της αντιγραφής μεθυλιώνει το θυγατρικό DNA, αντιγράφοντας το πρότυπο μεθυλίωσης των πατρικών κλώνων (2).

Η μεθυλίωση DNA παίζει ρόλο καταστολέα μεταγραφής συγκεκριμένων γονιδίων. Τα τελευταία χρόνια έχουν αναπτυχθεί τεχνικές για την αναγνώριση μεθυλιωμένων κατάλοιπων/δεικτών κυτοσίνης σε συγκεκριμένα γονίδια του γενώματος. Η μελέτη τους έχει αποδείξει ότι οι περιοχές υποκινητή ανενεργών γονιδίων έχουν αυξημένα επίπεδα μεθυλίωσης από αυτές των ενεργών γονιδίων και ότι τα πρότυπα μεθυλίωσης DNA διαφέρουν σε κάθε κυτταρικό τύπο, λογικό αφού εκτελούν διαφορετικές λειτουργίες σε κάθε ιστό. Κάποιες μελέτες σε θηλυκά θηλαστικά δείχνουν ότι η μεθυλίωση στους υποκινητές παίζει τον ρόλο να κρατάει ανενεργό το γονίδιο, η ίδια η απενεργοποίηση έχει προηγηθεί με άλλο μηχανισμό (2).

Η μεθυλίωση DNA συσχετίζεται με την μεθυλίωση ιστονών. Στη μεθυλίωση των ιστονών οι μεθυλομάδες μεταφέρονται στα αμινοξέα των πρωτεϊνών ιστόνης που βρίσκονται στα νουκλεοσώματα. Η μεθυλίωση ιστονών μπορεί να ενεργοποιήσει την καταστολή της μεταγραφής ενός προτύπου τροποποιήσεων ιστονών, δηλαδή ανάλογα με το ποια αμινοξέα μεθυλιώνονται και πόσες μεθυλομάδες προστίθενται. Τα κατάλοιπα μεθυλιωμένων κυτοσίνων μπορούν να χρησιμοποιηθούν ως σημεία πρόσδεσης για να προστεθούν και άλλα ένζυμα τροποποίησης των ιστονών, τα οποία καταστέλλουν περισσότερο την χρωματίνη του υποκινητή.

Τεχνική CRISPR (clustered regularly interspaced short palindromic repeats). Τα νέα φάρμακα, για την καταπολέμηση του καρκίνου και άλλων ασθενειών, που βασίζονται στην επιγενετική μελέτη δεν είναι ακόμα πολύ αποτελεσματικά. Κάποια από αυτά αλλάζουν το ποσοστό της ολικής μεθυλίωσης DNA ή ακετυλίωσης ιστονών στο κύτταρο, η τελευταία σχετίζεται με ενεργοποίηση γονιδίου, ενώ άλλα μπορούν να αναστείλουν τις πρωτεΐνες, οι οποίες ρυθμίζουν επιγενετικά πολλαπλά γονίδια-στόχους. Η ικανότητα της να ρυθμίζει επιγενετικά γονίδια με πολλαπλούς στόχους. Παρόλα αυτά, η ικανότητα επιλεκτικής ρύθμισης ενός συγκεκριμένου γονιδίου είναι κάτι που δεν μπορεί ακόμα να παρέχει η κλασική φαρμακοβιομηχανία.

Αρχίζει να αναπτύσσεται μια εναλλακτική μέθοδος με γενετικά παραγόμενα υβρίδια πρωτεϊνών που μπορούν να επανενεργοποιήσουν ή να καταστείλουν ένα συγκεκριμένο γονίδιο. Τα υβρίδια παράγονται με την συγκόλληση μέρους των μεταγραφικών παραγόντων, το οποίο μπορεί να ενσωματωθεί σε συγκεκριμένες αλληλουχίες του γενώματος, σε μια πρωτεΐνη που μπορεί να προσθέσει ή να αφαιρέσει την μεθυλίωση DNA ή τις τροποποιήσεις ιστονών. Οι νέες τεχνικές επιγενετικής μεταβολής της διάταξης των γονιδίων μπορούν να μετατρέψουν τον τρόπο που διαβάζεται και μεταφράζεται η αλληλουχία DNA, με τελικό στόχο την αντιστροφή ζημιογόνων προτύπων επιγενετικών τροποποιήσεων που είναι κοινά στον καρκίνο και άλλες ασθένειες. Μια νέα τεχνική μεταβολής της διάταξης αλληλουχίας DNA ονομάζεται CRISPR και μπορεί να μεταβληθεί, έτσι ώστε να μεταβάλεις την διάταξη στις επιγενετικές τροποποιήσεις αντί να μεταβάλλεις απευθείας την διάταξη της αλληλουχίας DNA. Μέρος της Cas9 πρωτεΐνης χρησιμοποιείται σαν περιοχή ενσωμάτωσης αντί για την περιοχή των μεταγραφικών παραγόντων στην αλληλουχία DNA. Η πρωτεΐνη Cas9 ενσωματώνεται σε ένα «οδηγό» κλώνο RNA (sgRNA, single-guide RNA), που αναζητά συμπληρωματικές αλληλουχίες DNA, δηλαδή οι επιστήμονες μπορούν να εισάγουν

οποιαδήποτε αλληλουχία βάσεων μέσα σε αυτόν τον κλώνο RNA. Η ρυθμιστικές υβριδικές πρωτεΐνες επιγενετικής περιέχουν περιοχές Cas9, που μπορούν να κατευθυνθούν προς την υπό εξέταση αλληλουχία DNA, επιτρέποντας την ενεργοποίηση ή την καταστολή ενός συγκεκριμένου γονιδίου. Η επιγενετική μεταβολή της διάταξης των γονιδίων έχει γίνει πειραματικά μόνο σε κυτταροκαλλιέργειες μέχρι στιγμής. Η τεχνική θέλει πολύ μελέτη μέχρι την ασφαλή χρησιμοποίηση της σε κλινικά πειράματα, αλλά έχει σημαντικές πιθανότητες να χρησιμοποιηθεί στην πρόγνωση και την αντιμετώπιση του καρκίνου και σε άλλες ασθένειες που αφορούν μετατροπές στα μοτίβα γονιδιακής ενεργοποίησης (33).

1.5. Ποια η Έννοια της Γενωμικής

Το μέγεθος του γονιδιώματος εξαρτάται από το είδος που εξετάζεται. Από μετρήσεις σε ανθρώπινα κύτταρα, η ποσότητα του DNA στο καθένα, περιέχει 3,2 δισεκατομμύρια ζεύγη βάσεων. Κάθε ανθρώπινο κύτταρο περιέχει γονιδίωμα που κληρονομεί και από τους δύο γονείς, δηλαδή έχει δύο αντίγραφα του κάθε χρωμοσώματος και του κάθε γονιδίου. Η αλληλούχισή του έδωσε απαντήσεις για τον ρόλο των γονιδίων όχι μόνο σε ασθένειες που προέρχονται από μεταλλάξεις, όπως η κυστική ίνωση και τον σύνδρομο Down, αλλά και γενετική προδιάθεση για καρκίνο ή καρδιακά προβλήματα, οι οποίες εμφανίζονται στο οικογενειακό ιστορικό. Ακόμα και μεταδοτικές ασθένειες, όπως η ιλαρά ή και το απλό κρυολόγημα, επηρεάζονται από το γονιδίωμα, αφού το ανοσοποιητικό μας σύστημα εξαρτάται από το DNA. Όπως είδαμε στην προηγούμενη ενότητα, όσο μεγαλώνουμε οι επιγενετικοί παράγοντες έχουν αυξημένη επίδραση στα κύτταρά μας , αφού περιλαμβάνουν και το σύνολο των γενετικών μεταλλάξεων που συμβαίνουν στην διάρκεια της ζωής μας. Από ένα απλό ανθρώπινο γονιμοποιημένο ωάριο γίνεται η μεταμόρφωση σε έναν πολύπλοκο ενήλικο οργανισμό, με σχεδόν 30 τρισεκατομμύρια κύτταρα. Στα μέσα του 1980, που άρχισαν οι συζητήσεις για την αλληλούχιση του ανθρώπινου γονιδίου, οι περισσότεροι αντιμετώπισαν την ιδέα με σκεπτικισμό.

Υπάρχουν διάφορες μέθοδοι που καταδεικνύουν τις σχέσεις μεταξύ RNA και πρωτεϊνών. Αυτές οι μέθοδοι, παρόμοιες με των τεχνικών ChIP, συμπεριλαμβάνουν την διασύνδεση του RNA με προσδεδεμένες πρωτεΐνες, οι οποίες στην συνέχεια επιλέγονται με συγκεκριμένα αντισώματα. Η κατανόηση αυτών των τεχνολογιών θα βοηθήσει στο να βρεθεί η σχέση μεταξύ γενετικών διαφοροποιήσεων και ασθένειας. Οι τεχνικές εξαιρετικά υψηλής απόδοσης προτιμώνται λόγω της συνεχής μείωσης στο κόστος και της ανάπτυξης νέων τεχνολογιών και μεθόδων. Βέβαια προϋπόθεση είναι η ανάλυση των μαζικών δεδομένων από σωστά

εκπαιδευμένους βιοπληροφορικούς επιστήμονες και η ασφαλής αποθήκευση και διαχείριση αυτών των δεδομένων.

Οι τεχνολογίες αλληλούχισης και τα πρωτόκολλα βελτιώνονται συνεχώς λόγω της ανταγωνιστικότητας ανάμεσα στις εταιρίες, και της ανάγκης τους να παραμένουν σύγχρονες. Νέες χημικές ανακαλύψεις εφαρμόζονται με μεγάλη συχνότητα που βελτιώνουν τον αριθμό των βάσεων που διαβάζονται και καλυτερεύουν την ποιότητα του. Ακόμα, υπάρχει μία τάση για απλοποίηση των τεχνικών για να είναι προσβάσιμη η τεχνολογία από το κάθε ερευνητικό κέντρο ή κλινικό εργαστήριο, μικρό και μεγάλο. Τα τελευταία χρόνια έχει δοθεί αγώνας να δημιουργηθούν μεθοδολογίες για την εφαρμογή γενωμικών και μεταγραφικών τεχνικών σε μονοκύτταρο επίπεδο. Η μεγαλύτερη δυσκολία είναι η ενίσχυση του ολικού γενώματος, χωρίς να ενισχυθούν τα σφάλματα, ενώ παράλληλα να μην γίνει επιμόλυνση από άλλα κύτταρα. Τεχνικές γνωστές για ενίσχυση όταν εξετάζουμε ολόκληρο το γονιδίωμα είναι η Multiple Displacement Amplification (MDA) και η αλυσιδωτή αντίδραση πολυμεράσης (Polymerase Chain Reaction, PCR). Η επιλογή μίας εκ των δύο τεχνικών χρησιμοποιούνται ανάλογα με των τύπο και τις αλληλουχίες που εξετάζονται. Η αλληλούχιση ενός κυττάρου θα επιτρέψει όχι μόνο την ανάλυση μεταξύ υγιών και ασθενών κυττάρων από τον ίδιο οργανισμό, αλλά και την γενωμική ετερογένεια μεταξύ ατόμων. Η γενωμική ενός κυττάρου είναι κατάλληλη για προγενετική εξέταση χρησιμοποιώντας απομονωμένα κύτταρα του μωρού από το αίμα της μητέρας (9).

PCR. Η PCR είναι το εργαλείο που χρησιμοποιούμε για να απομονώσουμε το γενετικό υλικό σε κατάλληλη ποσότητα και ποιότητα, έτσι ώστε να γίνει η σήμανση των μορίων για να κάνουμε την ανάλυση με το όργανο που προτιμάμε. Είναι χρήσιμη όταν αλληλουχοποιούμε το γένωμα και όταν το μελετάμε για SNPs, επιγενετικές τροποποιήσεις. Η PCR γίνεται στα μόρια DNA και, συνεπώς, και στα cDNA. Η επαύξηση αυτή του του DNA γίνεται με την βοήθεια εκκινητών, στις περιοχές που θέλουμε να αντιγραφούν, και των ενζύμων αντιγραφής του DNA. Η κλωνοποίηση έχει κατεύθυνση από το 5’-άκρο στο 3'-άκρο, αφορά στην αποδιάταξη του μορίου DNA από την δίκλωνη μορφή του στα σημεία που υποδεικνύουν οι εκκινητές , την υβριδοποίηση και την σύνθεση του νέου μορίου DNA. Αυτός ο κύκλος επαναλαμβάνεται πολλές φορές και τα εκατομμύρια αντίγραφα, δηλαδή τα προϊόντα της PCR, στο σύνολό τους ονομάζονται άμπλικονς. Ισχύει ακόμα ότι για να μην “χάνονται” νουκλεοτίδια στα άκρα της αλληλουχίας, το DNA κόβεται σε μικρότερα θραύσματα, που αντιγράφονται με μεγαλύτερη πιστότητα, όσο βελτιώνεται η τεχνική έχουν ακρίβεια στην

πιστότητα και μεγαλύτερα θραύσματα DNA. Σε κάθε κύκλο PCR κλωνοποιούνται αυτά τα θραύσματα DNA, έτσι ώστε να υπάρχει εκθετική αύξηση στην ποσότητα του DNA στο δείγμα. Για ολόκληρο το γένωμα χρησιμοποιούνται πολλαπλάσιοι εκκινητές για να κλωνοποιηθεί κάθε περιοχή του DNA σε κάθε κύκλο. Στο νεοσύνθετο DNA είναι σύνηθες τα νουκλεοτίδια να σημαίνονται, με αποτέλεσμα το μεγαλύτερο μέρος της αλληλουχίας DNA να είναι σημασμένο μετά από μερικούς κύκλους.

Από αυτήν την τεχνολογία έχουν προέλθει γρηγορότερες και ποιοτικά βελτιωμένες μορφές PCR, επειδή κατά την ενίσχυση μπορεί να εμφανιστούν σφάλματα, όσο περισσότερο επαυξάνεται το γενετικό υλικό, τόσο πιο πιθανά είναι τα σφάλματα. Μια τέτοια μορφή είναι η ποσοτικοποιημένη PCR (qPCR), η ποσοτικοποίηση είναι δυνατή αφού ο αριθμός των κλωνοποιήσεων ελέγχεται στην αντίδραση PCR. Έτσι, ένας εκκινητής επιλέγει μια μόνο περιοχή του DNA και σε κάθε κύκλο PCR, καταγράφεται το επίπεδο της σήμανσης στο δείγμα. Συνήθως γίνεται σε 96 «πηγαδάκια», προσφέρει λογαριθμική αύξηση στην ποσότητα του DNA στο δείγμα, ενώ η αρχική ποσότητα του μπορεί να βρεθεί. Για να χρησιμοποιηθεί αυτό το εργαλείο σε μόρια RNA κάνουμε αντίστροφη μεταγραφή PCR (RT-PCR), με την οποία συνθέτουμε τα cDNA και σε αυτά κάνουμε την φυσιολογική PCR. Η αντίστροφη μεταγραφάση (Reverse Transcriptase, RT) είναι ένα ένζυμο που χρησιμοποιείται για να παράγει συμπληρωματικό DNA από ένα καλούπι RNA. Το cDNA είναι πιο σταθερό μόριο από το RNA ενώ για να επιλέξουμε το RNA που μας ενδιαφέρει είναι σύνηθες να χρησιμοποιούμε mRNA με poly-A ουρές.

UCSC Genome Browser. Είναι προσβάσιμα στον γενωμικό φυλλομετρητή του Πανεπιστημίου της Καλιφόρνια στη Σαντα Κρούζ, δεδομένα από μεγάλα ερευνητικά προγράμματα, όπως από το ENCODE και το HGP (University of California, Santa Cruz, UCSC[2] Genome Browser). Ακόμα και δεκαετίες αργότερα, χρησιμοποιείται για να απεικονίσει γραφικά και να αναλύσει δεδομένα συναρμολόγησης του γενώματος που υπάρχουν ήδη στο σύστημα UCSC ή με δεδομένα που εισάγει ο χρήστης και καλύπτει την συναρμολόγηση του γενώματος πάνω από 93 διαφορετικών οργανισμών. Το πρόγραμμα περιήγησης UCSC προσφέρει πρόσβαση, εκτός από την βιβλιογραφία, στα ακόλουθα δεδομένα: χαρτογράφησης, αλληλουχοποίησης, γονοτύπου, φαινοτύπου, γονιδιακής έκφρασης, γονιδιακή ρύθμιση συγκριτική γενωμική, παραλλαγές SNPs και επαναλήψεων. Επιπροσθέτως, υπάρχουν διάφορα εργαλεία ανάλυσης στο πρόγραμμα περιήγησης γενώματος UCSC,

[2] https://genome.ucsc.edu

το κύριο λογισμικό που επιτρέπει την γραφική αναπαράσταση γενωμικών δεδομένων, το BLAST, το οποίο χρησιμοποιείται για στοίχιση αλληλουχιών, την PCR με προσομοίωση (in-silico), η οποία στοιχίζει αλληλουχίες εκκινητών στα γενώματα και το Gene Sorter, το οποίο κάνει δυνατή την αναζήτηση παρόμοιων γονιδίων μέσω στατιστικών (metrics) έκφρασης (3).

Integrated Genome Browser (IGB)[3]. Είναι ένας ενοποιημένος γενωμικός φυλλομετρητής με παρόμοιες δυνατότητες και λειτουργίες με τον UCSC Genome Browser, που δημιουργήθηκε αρχικά για ανάλυση δεδομένων ολιγονουκλεοτιδικών μικροσυστοιχιών της Affymetrix από την ίδια την εταιρία. Είναι πλέον ελεύθερο λογισμικό και κατάλληλο για ανάλυση δεδομένων μικροσυστοιχιών, δεδομένων από αλληλούχιση RNA-Seq/ ολόκληρου του γενώματος και άλλων δεδομένων.

Integrative Genomic Viewer (IGV)[4]. Η εφαρμογή γενωμικής αναπαράστασης έχει παρόμοιες δυνατότητες με τον UCSC Genome Browser. Παρέχει πρόσβαση σε δεδομένα από 1000 ανθρώπινα γενώματα και μπορεί να οπτικοποιήσει αρχεία BAM. Είναι πολύ καλή μέθοδος οπτικοποίησης για δεδομένα από αλληλούχιση SNP, αφού φαίνονται εύκολα οι ακραίες τιμές σε διάφορα διαγράμματα, από αλληλούχιση RNA-Seq και από αλληλούχιση ολόκληρου του γενώματος.

Η συλλογή βάσεων δεδομένων του National Center for Biotechnology Information (NCBI). Το εθνικό κέντρο για την βιοτεχνολογική πληροφορική NCBI είναι μέρος της εθνικής βιβλιοθήκης ιατρικής των Ηνωμένων Πολιτειών (National Library of Medicine, NLM), που με την σειρά του ανήκει στο εθνικό ινστιτούτο υγείας National Institutes of Health, που ιδρύθηκε το 1988. Ο αρχικός στόχος του NCBI ήταν η ανάπτυξη υπολογιστικών συστημάτων για την διαλογή, την μετάφραση και την παρουσίαση των συνεχώς αυξανόμενων δεδομένων στην ανθρώπινη μοριακή βιολογία, γενετική και βιοχημεία. Σήμερα όμως έχει επεκταθεί ο ρόλος του ως μια διεθνής πλατφόρμα για τον διαμοιρασμό δεδομένων και υπολογιστικών αναλύσεων, που παρέχει πρόσβαση σε πληροφορίες για κάθε οργανισμό. Η ιστοσελίδα της NCBI περιέχει ένα μεγάλο εύρος βιολογικών πληροφοριών στην οποία έχει πρόσβαση και μπορεί να κατεβάσει ο κάθε ενδιαφερόμενος. Η πλατφόρμα NCBI's Entrez προσφέρει πρόσβαση σε μια ευρεία κλίμακα βάσεων δεδομένων, όπως το Gene, μια βάση δεδομένων που αφορά γονιδιακά δεδομένα, προϊόντα γονιδίων, αλληλουχίες νουκλεϊνικών οξέων, γονιδιακούς χάρτες, χαρακτηρισμός διαφοροποιήσεων αλληλουχιών και γονιδιακά ομόλογα με επιπλέον συνδέσμους σε εξωτερικές βάσεις δεδομένων.

Η NCBI παρέχει πρόσβαση στη βάση δεδομένων για την Μεντελιανή κληρονομικότητα στον άνθρωπο OMIM (online Mendelian inheritance in man), είναι μια υπερπλήρης βάση δεδομένων με

[3] https://bioviz.org/download.html
[4] https://software.broadinstitute.org/software/igv/

πληροφορίες στις κληρονομικές ασθένειες και τα γονίδια που τις προκαλούν, τις περιοχές που βρίσκονται στα χρωμοσώματα, την σχέση μεταξύ φαινότυπου και γενότυπου και τα κλινικά χαρακτηριστικά τους. Μια άλλη χρήσιμη βάση δεδομένων η PubMed, που αποτελεί μέρος του οικοσυστήματος της NCBI, παρέχει πρόσβαση στο αρχείο αποσπασμάτων και άρθρων βιοϊατρικής βιβλιογραφίας, μέσω του MEDLINE και επιστημονικών περιοδικών. Στην παρούσα εργασία ήταν πολύ χρήσιμο εργαλείο για την βιβλιογραφική αναζήτηση. Εκτός από την βάση δεδομένων Pubmed χρησιμοποιήθηκαν εκτενώς και δύο άλλες πολύ γνωστές βάσεις δεδομένων, εκτός του οικοσυστήματος της NCBI, για την ανάκτηση βιβλιογραφικής πληροφορίας, αυτές είναι η Scopus και η ScienceDirect. Στην PMC βάση δεδομένων μπορούμε επίσης να βρούμε πολλά άρθρα δωρεάν πρόσβασης, όμως κάποια από αυτά οποία δεν είναι σε πρώιμο στάδιο και δεν έχουν ελεγχθεί από τρίτους αναγνωρισμένους επιστήμονες (peer reviewed). Η οργάνωση του NCBI επιτρέπει αλληλεπίδραση μεταξύ όλων αυτών των βάσεων δεδομένων, που είναι ενσωματωμένες σε αυτή, και για τις οποίες έχει εγχειρίδια χρήστη και άλλες χρήσιμες συμβουλές, στον παρακάτω πίνακα (Πίνακας 3) υπάρχουν κάποιες σχετικές βάσεις δεδομένων (3).

Πίνακας 3. Συγκεκριμένες βάσεις δεδομένων προσβάσιμες από τον ιστότοπο του NCBI (3).

Βάσεις Δεδομένων	Περιγραφή
ClinVar	Οργάνωση γενωμικών διαφοροποιήσεων και πως σχετίζονται με την ανθρώπινη υγεία.
dbVar	Υπομνηματισμός γενωμικών δομικών διαφοροποιήσεων.
Gene	Παρέχει αναλυτικές βιολογικές πληροφορίες για το κάθε γονίδιο.
Genome	Οργάνωση των γενωμικών δεδομένων σε κάθε είδος και πρόσβαση σε γενωμικές πηγές ανθρώπων, μικροβιακών γενωμάτων, organelles, ιών, προκαρυωτικών γενωμάτων αναφοράς.
MedGen	Παρέχει πληροφορίες για την ιατρική πλευρά της ανθρώπινης γενετικής.
Nucleotide	Οργάνωση νουκλεοτιδικών αλληλουχιών από πολλαπλές πηγές, παρέχει δηλαδή γενωμικές, γονιδιακές και μεταγραφικές αλληλουχίες.
OMIM	Παρέχει αναλυτικές πληροφορίες για κληρονομικές ασθένειες και την γενετική δομή τους.
Protein	Προσφέρει πρωτεϊνικές αλληλουχίες από πολλαπλές πηγές.
PMC	Παρέχει ελεύθερη πρόσβαση σε επιστημονικά περιοδικά βιοϊατρικής και βιολογικών επιστημών στα U.S. National Institutes of Health (NIH) και National Library of Medicine (NLM).
PubMed	Οργάνωση αναλυτικών αποσπασμάτων βιοϊατρικής έρευνας και βιβλιογραφίας από MEDLINE, επιστημονικά περιοδικά βιολογικών επιστημών και ψηφιακά βιβλία.
SNP	Οργάνωση πολυμορφισμών single nucleotide polymorphisms (SNPs) και διαφοροποιήσεις μικρής κλίμακας άλλου τύπου.

2. Κεφάλαιο 2 Περιγραφή Πειραματικών Μεθόδων Μελέτης Μικροσυστοιχιών και Αλληλουχοποίησης Νέας Γενιάς

2.1. Αρχή Λειτουργίας και Πειραματική Διαδικασία Μεθόδου Μικροσυστοιχιών (Microarrays)

Η μελέτη μικροσυστοιχιών είναι μια τεχνική εξαιρετικά υψηλής απόδοσης ανάλυσης γονιδίων, η οποία δίνει λεπτομερής πληροφορίες για την γονιδιακή έκφραση ενός οργανισμού και επιτρέπει την παράλληλη ανίχνευση μέχρι και 100.000 στοχευμένων γονιδίων, ενώ κάνει δυνατή την καταγραφή και την μελέτη της γονιδιακής έκφρασης. Βασίζεται στην διαδικασία του υβριδισμού, το στοχευόμενο DNA ή RNA καταδεικνύεται με χρήση ειδικών ανιχνευτών σηματοδότησης. Η ιδέα για τις μικροσυστοιχίες πρωτοεμφανίστηκε στα μέσα της δεκαετίας του ‘80 και βασίστηκε στην τεχνική εκτύπωσης κατά Southern, η οποία όμως επιτρέπει ανάλυση μόνο ενός μικρού αριθμού μορίων και οι θέσεις πρόσδεσης εμφανίζονται με αυτοραδιογραφία. Ακόμα, η μέθοδος φθορισμού που χρησιμοποιήθηκε στην εξελιγμένη τεχνική ήταν μία βελτίωση σε σχέση με τους ραδιοσημασμένους ανιχνευτές που χρησιμοποιείται στην τεχνική εκτύπωσης κατά Southern. Το 1990 εφευρέθηκε και το πρώτο ρομποτικό μηχάνημα για τις μικροσυστοιχίες το οποίο ήταν βασισμένο σε πείρους (pins).

Η τεχνική μικροσυστοιχιών έφερε μαζική παραγωγή πληροφοριών και, σε αυτή, συνέβαλλαν επιστήμονες και μηχανικοί διάφορων ειδικοτήτων. Υπάρχουν διάφοροι τύποι ανάλυσης συστοιχιών. Μπορείς να συγκρίνεις ιστούς από διαφορετικές περιοχές του σώματος. Μπορείς να συγκρίνεις κύτταρα της ίδιας περιοχής σε διαφορετικά στάδια της ανάπτυξης. Σε φαρμακολογικές μελέτες μπορείς να ελέγξεις την διάρκεια που ενεργεί ένα φάρμακο με την ανάλυση των δειγμάτων ανά τακτά χρονικά διαστήματα, μέσα σε ένα εικοσιτετράωρο (time course analysis), αλλά και την τυχόν τοξικότητά του. Στην φάση πριν από τις κλινικές μελέτες είναι ένα εργαλείο επιλογής των κατάλληλων συμμετεχόντων. Ακόμα είναι σύνηθες να χρησιμοποιείται μελέτες που συγκρίνουν φυσιολογικά κύτταρα με καρκινικά κύτταρα (two condition) σε δικάναλη πλατφόρμα ή και των δύο καταστάσεων σε μονοκάναλη πλατφόρμα. Μπορεί να είναι ένα πείραμα για τον έλεγχο των συνθηκών και τον σχεδιασμό του κυρίως πειράματος. Μια άλλη εφαρμογή θα μπορούσε να είναι η ανάλυση των μικροσυστοιχιών για των έλεγχο δυναμικών αποκρίσεων σε περιβαλλοντικά σήματα. Αλλά είναι και καλά διαγνωστικά εργαλεία, αφού μπορούν να ανιχνεύσουν πρότυπα γονιδιακής έκφρασης που σχετίζονται με καταστάσεις ασθένειας.

Οι μικροσυστοιχίες DNA (DNA microarray ή DNA chip ή βιοτσίπ) είναι μια διάταξη από ακινητοποιημένες αλληλουχίες νουκλεϊνικών οξέων ανίχνευσης (probes), δηλαδή είναι μια επιφάνεια από γυαλί, πυρίτιο κ.α. πάνω στην οποία τοποθετούνται σε συγκεκριμένες θέσεις, κηλίδες (spots), γνωστές και μοναδικές αλληλουχίες DNA ή RNA, συνήθως στην μορφή συμπληρωματικού DNA (complimentary DNA, cDNA). Ανάλογα με την ανάλυση που κάνουμε αυτές αντιστοιχούν σε ολόκληρα γονίδια ή γενωμικά θραύσματα ή σε ολιγονουκλεοτίδια. Παλιότερα αυτή η διαδικασία γινόταν μη αυτοματοποιημένα με το χάσιμο πολλών εργατοωρών στο εργαστήριο. Πλέον, η ροή εργασίας γίνεται με «κιτ» κατάλληλα για το όργανο της εταιρίας που χρησιμοποιείται για την ανάλυση. Εκτός από της μικροσυστοιχίες υπάρχουν και άλλες συσκευές bioMEMS, οι μικρορευστοϊκές διατάξεις (microfluidic). Ένα τέτοιο παράδειγμα είναι το μικρορευστοϊκό flow cell, που αποτελείται από τέσσερα διαφορετικά μέρη, και παράγεται από την GELifesciences, που είναι μέρος της εταιρίας General Electric.

Στα γενωμικά τσίπ είναι ακόμα πιο σημαντική η υψηλή ποιότητα και η επαρκή ποσότητα του γενετικού υλικού. Για παράδειγμα όταν θέλουμε να αναλύσουμε δείγματα που προέρχονται από βιοψία, όπου το γενετικό υλικό είναι ελλιπές, είναι σύνηθες το απομονωμένο DNA να μην πληροί τις προϋποθέσεις για να γίνει η γενωμική ανάλυση. Γι’αυτό είναι σημαντική η πιστότητα κατά την επαύξηση του γενωμικού DNA, η οποία μπορεί να είναι ειδικευμένη σε αλληλόμορφα, γονιδιακά ειδικευμένη ή σε ολόκληρο το γένωμα. Μια τεχνική κατάλληλη για την ανάλυση συστοιχιών διαλογής SNP, είναι η γονιδιακά ειδικευμένη T7-βασισμένη επαύξηση, η οποία χρησιμοποιεί μια πολυμεράση DNA με λειτουργίες 3-εξωνουκλεάσης. Αυτή μπορεί να εφαρμοστεί γα την ανάλυση οποιουδήποτε δείγματος, με την παράλληλη χρήση PCR και in vitro μεταγραφής.

Σχεδιάζονται γονιδιακά ειδικευμένοι εκκινητές και η αλληλουχία υποκινητών T7 (T7 primer), δηλαδή η (5′aaa cga cgg cca gtg aat tgt aat acg act cac tat agg cgc 3′), προσδένεται στο 5’-άκρο του μπροστινού εκκινητή για να κάνει την επαύξηση PCR του στοχευμένου γονιδίου σε πλήθος μονόκλωνων RNA, που τα λέμε και προϊόντα κλωνοποίησης PCR, και αυτά στην συνέχεια σημαίνονται με φθορίζουσα ή φθορίζουσες ουσίες και υβριδοποιούνται στην μικροσυστοιχία (34). Παρόλο που η PCR είναι πολύ χρήσιμη στις μικροσυστοιχίες, πολλές φορές ενισχύει και τυχόν σφάλματα. Γι’αυτό έχουν προέλθει από αυτήν διάφορες μοριακές τεχνικές επαύξησης του γενετικού υλικού αλλά και βελτιωμένες τεχνικές PCR.

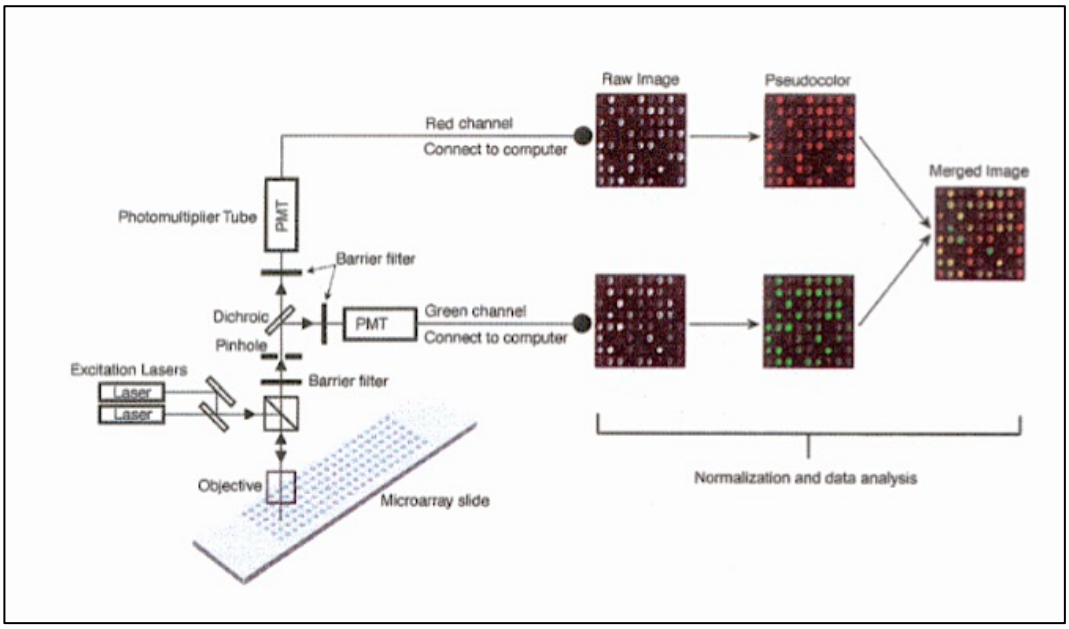
2.1.1. DNA και RNA ανάλυση μικροσυστοιχιών

Το κομμάτι πειραματικό κομμάτι (wetlab) της τεχνικής μικροσυστοιχιών αρχίζει με την συλλογή δείγματος ελέγχου (control sample), που χρησιμοποιείται ως αναφορά, και το δείγμα υπό εξέταση. Από το κάθε δείγμα απομονώνεται το γενετικό υλικό. Στη συνέχεια, κατακερματίζεται και σημαίνεται με έναν φθορίζοντα δείκτη. Οι φθορίζουσες ουσίες που χρησιμοποιούνται έχουν αρκετή ευαισθησία, μικρό κόστος και κάνουν δυνατή την χρήση δύο χρωστικών για την εξέταση δύο διαφορετικών δειγμάτων με την σάρωση μιας εικόνας. Αυτά τα κλάσματα DNA εφαρμόζονται στη μικροσυστοιχία και υβριδοποιούνται με τις συμπληρωματικές τους αλληλουχίες, παράλληλα γίνονται χιλιάδες αντιδράσεις υβριδισμού. Η υβριδοποίηση είναι η διαδικασία με την οποία δύο συμπληρωματικές μονόκλωνες αλληλουχίες, ακόμα και αν προέρχονται από διαφορετικές πηγές, θα δημιουργήσουν ένα δίκλωνο υβρίδιο. Τα δύο δείγματα αφήνονται στα spots, όπου γίνεται η επώαση και ο σχηματισμός των δεσμών μεταξύ των δίκλωνων μορίων. Υπάρχουν και άλλοι τρόποι να γίνει ο υβριδισμός in situ, δηλαδή πάνω στο τσίπ, όπως γίνεται στην τεχνική της φωτολιθογραφίας.

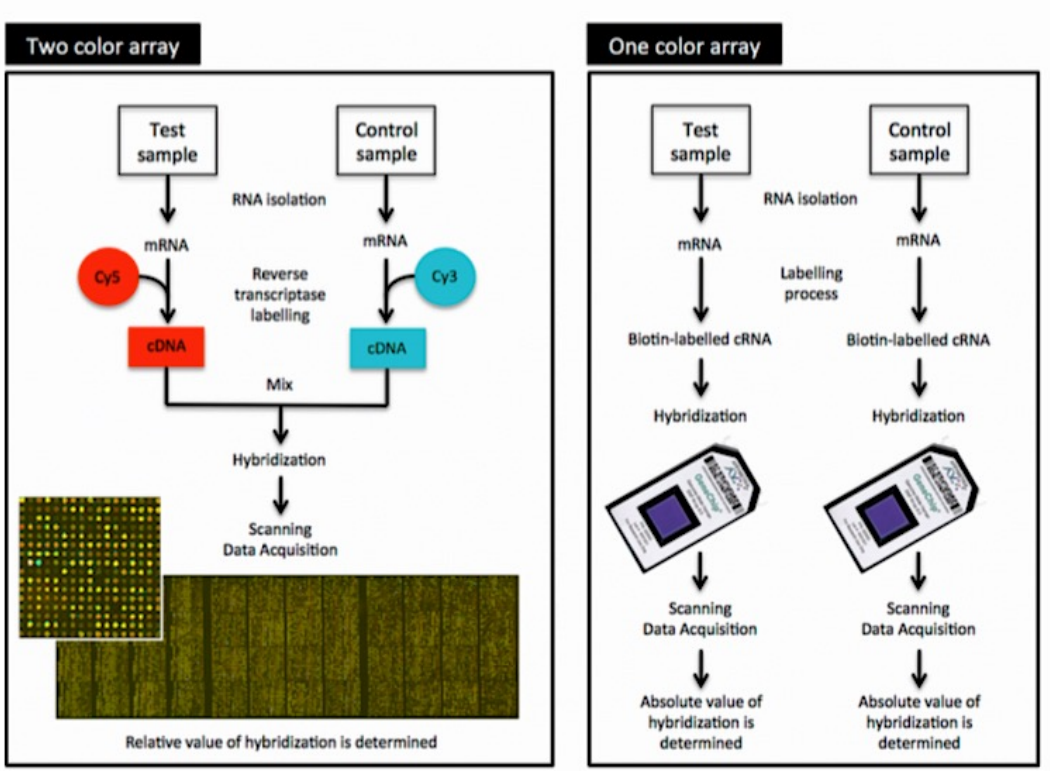
Οι μικροσυστοιχίες χρησιμοποιούν την αρχή συμπληρωματικότητας του γενετικού υλικού για την ανίχνευση του mRNA, στο υπό εξέταση δείγμα. Η ποσοτικοποίηση του mRNA μας οδηγεί στην εμφάνιση των διαφορικά εκφραζόμενων γονιδίων. Η αποκομιδή της εικόνας γίνεται αυτοματοποιημένα με ειδικούς σαρωτές μικροσυστοιχιών (Microarray Scanner), που έχουν τουλάχιστον δύο λέιζερ, ένα για κάθε χρωστική. Η αρχή λειτουργίας του ομοεστιακού λέιζερ φαίνεται στην **Εικόνα 5** και παρουσιάζει πως γίνεται αυτή η αποκομιδή των πρωτογενών δεδομένων. Τα αποτελέσματα έχουν διαφορετική σημασία ανάλογα με το «χρώμα» της φθορίζουσας ουσίας που υπερισχύει, αλλά και το πόσο έντονο είναι αυτό, και τον συνδυασμό ή την απουσία χρώματος. Όσο πιο έντονο είναι το χρώμα τόσο περισσότερα μετάγραφα περιέχει το spot. Η τεχνική έχει μεγάλη ευαισθησία, δηλαδή ένα mRNA μπορεί να ανιχνευθεί ακόμα και σε επίπεδο λιγότερο από ένα αντίγραφο ανά κύτταρο.

Μονοκάναλη και δικάναλη ανάλυση. Στην **Εικόνα 6**, παρουσιάζονται δύο περιπτώσεις αναλύσεων μικροσυστοιχιών την δικάναλη, η οποία δεν είναι πλέον η επικρατέστερη, και η μονοκάναλη. Η δικάναλη γίνεται με δύο χρωστικές, η μια για τον φθορισμό του φυσιολογικού δείγματος και η άλλη για τον φθορισμό του υπό εξέταση δείγματος. Τα αποτελέσματα με την μονοκάναλη μέθοδο έχουν εμφανίσει μεγαλύτερο βαθμό ακρίβειας, γι'αυτό τον λόγο προτιμάται πιο συχνά η πλατφόρμα ανάλυσης με μια χρωστική. Η

πειραματική διαδικασία μικροσυστοιχιών DNA δικάναλης πλατφόρμας, που απεικονίζεται αριστερά στην **Εικόνα 6**, για την σύγκριση μεταξύ δύο δειγμάτων με μικροσυστοιχίες DNA.



Εικόνα 5. Η λειτουργία του ομοεστιακού λέιζερ και πώς οπτικοποιείται η εικόνα στον υπολογιστή, μια ασπρόμαυρη εικόνα χρωματίζεται με επεξεργασία των πρωτογενών δεδομένων.



Εικόνα 6. Δικάναλη και Μονοκάναλη Ανάλυση Μικροσυστοιχιών που προέρχεται από την ιστιοσελίδα https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods.

Αρχικά, τα κύτταρα λαμβάνονται από τα δείγματα υπό εξέταση και ελέγχου, τα οποία μπορεί να είναι πολλαπλά. Μετά από κυτταρική καλλιέργεια θα χρησιμοποιηθεί το DNA ή το RNA ως εκμαγείο για την δημιουργία του cDNA (θα μπορούσε και του γενωμικού DNA, στην οποία περίπτωση ονομάζεται η μικροσυστοιχία DNA και genomic chip).

Η διάλυση των κυτταρικών ιστών γίνεται στον αναδευτήρα (Vortex Mixer), με εφαρμογή ειδικών διαλυτικών ουσιών, απομονώνεται το συνολικό RNA από τα υπόλοιπα συστατικά των κυττάρων και ο διαχωρισμός των μορίων RNA γίνεται στον φυγοκεντρητή (Microcentrifuge). Παραμένει μόνο το πλήθος των μορίων mRNA, 1-2% του συνολικού RNA, και αφαιρούνται τα υπόλοιπα είδη RNA. Στην πραγματικότητα προτιμάμε να γίνει η ανάλυση στο μόριο mRNA, που ονομάζεται αλλιώς μετάγραφο, αντί σε ολόκληρο το RNA, επειδή το mRNA προσφέρει λεπτομέρειες για την γονιδιακή έκφραση.

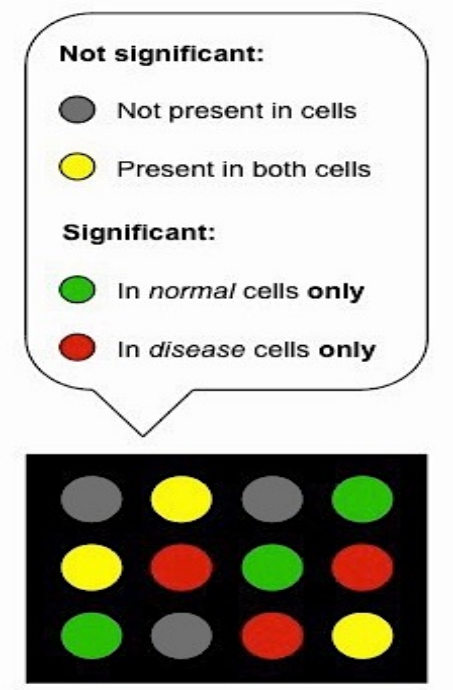
Όλα τα mRNA μετατρέπονται σε cDNA με την βοήθεια του ενζύμου αντίστροφη μεταγραφάση και παράλληλα ενσωματώνονται ενζυμικά τα φθοριούχα μόρια, που συνήθως είναι η κυανίνη 3 (Cy3) και η κυανίνη 5 (Cy5), ξεχωριστά σε κάθε μείγμα. Η αντίστροφη μεταγραφή (Reverse Transcription, RT) είναι η αντίστροφη διαδικασία από την μεταγραφή του DNA, στην οποία το DNA γίνεται RNA μέσω του ενζύμου πολυμεράση RNA. Με την αντίστροφη μεταγραφή (RT) τα μόρια mRNA θα γίνουν σημασμένα με φθορισμό μόρια cDNA.

Μετά την σήμανση αυτά αναμειγνύονται και υβριδοποιούνται στις μικροσυστοιχίες DNA. Η σύνθεση των μικροσυστοιχιών έχει ήδη ξεκινήσει με την εναπόθεση πλήθους πανομοιότυπου μονόκλωνου DNA στην κάθε κηλίδα του πλακιδίου. Αυτή η διαδικασία γίνεται με αυτοματοποιμένο τρόπο από ρομποτικά μηχανήματα στο εργαστήριο, αλλά είναι και διαθέσιμα στην αγορά από πολλές εταιρίες-κατασκευαστές βιοτσίπ, μειώνοντας τον κόπο αλλά και βελτιώνοντας την ποιότητα των μικροσυστοιχιών. Τα υβρίδια DNA-cDNA (που σε άλλη εφαρμογή μπορεί να είναι DNA-cRNA) δημιουργούνται λόγω της συμπληρωματικότητας των νουκλεοτιδικών βάσεων. Έτσι, το κάθε φθορισμένο μόριο cDNA αν βρεθεί συμπληρωματικό μονόκλωνο μόριο DNA σε κάποια κηλίδα της μικροσυστοιχίας DNA υβριδοποιείται σε αυτήν, με προσκόλληση των δύο κλώνων.

Μετά την υβριδοποίηση γίνεται πλύσιμο της μικροσυστοιχίας με ένα ειδικό υγρό και με αυτόν τον τρόπο αφαιρούνται όλα τα μονόκλωνα μόρια cDNA που δεν υβριδοποιήθηκαν. Με

τη βοήθεια ομοεστιακού λέιζερ γίνεται η κατάλληλη διέγερση της φθορίζουσας ουσίας και αποτυπώνεται ψηφιακά η εικόνα των spots στο μηχάνημα σάρωσης μικροσυστοιχιών και, μετά από κανονικοποίηση, έχουμε ποσοτικοποίηση των αποτελεσμάτων σύμφωνα με την ένταση των κηλίδων, που δίνεται αρχικά σε αποχρώσεις του γκρί. Η θέση της κηλίδας κάνει δυνατή την αναγνώριση του γονιδίου που ποσοτικοποιείται. Αυτά τα ελέγχουμε ως προς την ποιότητα των αποτελεσμάτων (QC) και μπορούμε να τα επεξεργαστούμε μαθηματικά και βιοπληροφορικά.

Για παράδειγμα, αν το δείγμα ελέγχου σημανθεί με πράσινη φθορίζουσα ουσία και το δείγμα υπό εξέταση σημανθεί με κόκκινη φθορίζουσα ουσία, τα γονίδια εκφράζονται περισσότερο στο δείγμα ελέγχου (upregulated) στις περιοχές που έχουν εντονότερη απόχρωση του πρασίνου, στην αντίθετη περίπτωση έχω εντονότερη την απόχρωση του κόκκινου, δηλαδή τα συγκεκριμένα γονίδια εκφράζονται περισσότερο στο υπό εξέταση δείγμα (downregulated) (**Εικόνα 7**).



Εικόνα 7. Υπερέκφραση και Υποέκφραση.

Ενώ, όπως απεικονίζεται στην **Εικόνα 7**, όταν τα γονίδια είναι εξίσου εκφρασμένα στο δείγμα ελέγχου και στο υπό εξέταση δείγμα έχουν κίτρινη απόχρωση, που στην περίπτωση του μονοκάναλου είναι η μίξη της σάρωσης από το ένα λέιζερ μόνο με κόκκινο φθορισμό και του άλλου λέιζερ μόνο με πράσινο φθορισμό και η εναποθέτηση της μιας εικόνας πάνω στην

άλλη, ενώ στην περίπτωση του δικάναλου έχω δύο χρωστικές και μια σάρωση αρκεί αλλά και περισσότερο «θόρυβο». Ακόμα, η απουσία χρώματος, δηλαδή μαύρα spots, δείχνει απουσία έκφρασης των γονιδίων στα δείγματα. Βέβαια, στην πραγματικότητα τα επίπεδα έντασης είναι όλα σε γκρί απόχρωση και χρωματίζονται κατάλληλα για να είναι ευδιάκριτα τα αποτελέσματα (5).

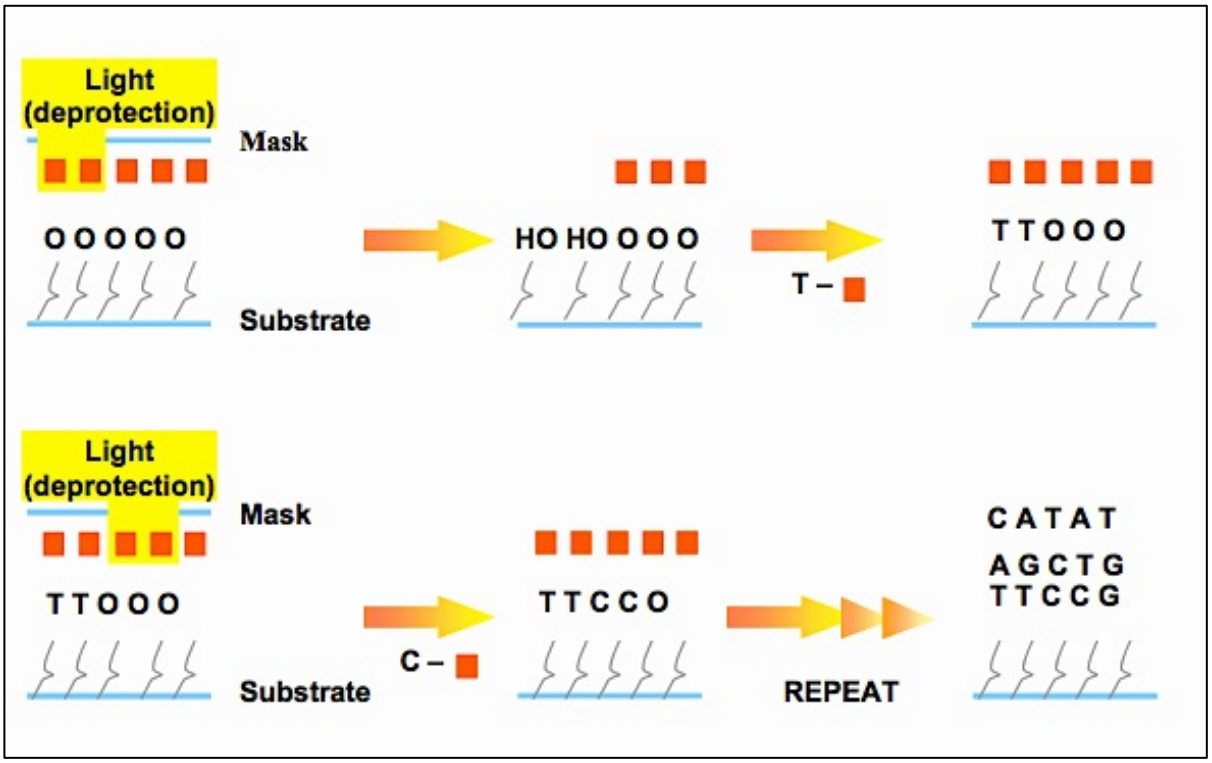
Οι τεχνικές εξαιρετικά υψηλών αποδόσεων έχουν ένα ευρύ φάσμα εφαρμογών. Για παράδειγμα, στην διεργασία αντιγραφής του DNA γίνεται δυνατός ο εντοπισμός των αφετηριών αντιγραφής, που δεν είναι προφανής στους ευκαρυωτικούς οργανισμούς όπως είναι στους προκαρυωτικούς οργανισμούς, τον χρόνο ενεργοποίησής τους και τον έλεγχό του για την σωστή μεταβίβαση της γενετικής πληροφορίας. Βεβαίως πάντα υπάρχει η πιθανότητα οι ανιχνευτές να δώσουν ψευδώς θετικό λόγω της δέσμευσης παραπάνω από μια αλληλουχιών στον ίδιο ανιχνευτή, ειδικά στα πολύπλοκα γενώματα, σε ομόλογα γονίδια και στα θηλαστικά σε οικογένειες γονιδίων με παραλλαγές ματίσματος. Ένα άλλο μειονέκτημα είναι ότι έχουν σχεδιαστεί να ανιχνεύουν συγκεκριμένες αλληλουχίες, κάτι που σημαίνει ότι είναι δυνατή η μελέτη μόνο γνωστών νουκλεοτιδικών αλληλουχιών.

Με παρόμοιο τρόπο χρησιμοποιούνται και τα γονιδιακά τσίπ (GeneChips®), που είναι εφεύρεση της εταιρίας Affymetrix, και φαίνονται στην **Εικόνα 8**, στην δεξιά πλευρά της εικόνας. Αυτά ανήκουν στην κατηγορία των ολιγονουκλεοτιδικών (oligos) μικροσυστοιχιών. Οι ολιγονουκλεοτιδικοί ανιχνευτές είναι μικρές αλληλουχίες ανιχνευτών που έχουν δημιουργηθεί για να ταιριάζουν σε μέρος γνωστών ή αναμενόμενων αλληλουχιών ORF. Η διαφορά στην ορολογία των ολίγο μικροσυστοιχιών, ενώ και στις μικροσυστοιχίες DNA χρησιμοποιούνται ολιγονουκλεοτίδια, είναι στην κατασκευή των τσίπ κυρίως, αλλά και στο ότι τα ολιγονουκλεοτίδια στις μικροσυστοιχίες DNA έχουν μέγεθος περίπου 100nt.

Οι ολιγονουκλεοτιδικές μικροσυστοιχίες σχεδιάζονται για να αναπαραστήσουν σε κάθε κηλίδα ένα μοναδικό γονίδιο και το τύπωμα των ολιγονουκλεοτιδίων να γίνει in situ, δηλαδή τα ολιγονουκλεοτίδια συντίθενται κατευθείαν πάνω στο τσίπ, αντί να τοποθετούνται ολόκληρες αλληλουχίες, όπως στις μικροσυστοιχίες DNA. Οι όλιγο αλληλουχίες μπορεί να είναι μακρά ολιγονουκλεοτίδια (π.χ. 60-μερές Agilent) ή μικρότερου μήκους (π.χ. 25-μερές Affymetrix), ανάλογα το είδος της ανάλυσης που μας ενδιαφέρει. Το μεγαλύτερο μήκος είναι κατάλληλο για να στοχεύουμε συγκεκριμένα γονίδια, ενώ το μικρότερο μήκος είναι

κατάλληλο για εμφάνιση των κηλίδων σε μεγάλη πυκνότητα και είναι πιο φτηνά από το 60-μερές.

Μια τεχνική κατασκευής τους είναι η σύνθεση με φωτολιθογραφία (Affymetrix, μέλος της εταιρίας Thermo Fisher Scientific Inc.) ένα νουκλεοτίδιο την φορά πάνω σε υπόστρωμα πυριτίου, με την βοήθεια του φωτός και φωτοευαίσθητων υλικών που αποκρύπτουν (masking) και προστατεύουν το υλικό από χημικές αλλαγές, όπως φαίνεται στην **Εικόνα 8**.



Εικόνα 8. Σύνθεση ολιγονουκλεοτιδικών μικροσυστοιχιών με φωτολιθογραφία (Affymetrix).

Αρχικά, το φως αφαιρεί την προστασία από την κηλίδα και χημικά τροποποιημένα νουκλεοσίδια προστίθενται, δημιουργούν χημικά σύμπλοκα (chemically coupled), ενώ ένας μηχανισμός επιβολής ανώτατου ορίου κάνει αποκλεισμό αυτών που δεν δημιουργούν σύμπλοκα. Αυτή η διεργασία επαναλαμβάνεται με διαφορετικές "μάσκες", μέχρι να γίνει η σύνθεση ολόκληρου του ανιχνευτή πάνω στην όλιγο μικροσυστοιχία. Συγκεκριμένα, για χρήση στα γονιδιακά τσίπ της Affymetrix, συνήθως 20-25nt, μεταγράφεται το cDNA σε cRNA, στο οποίο κάθε βάση ουρακίλης είναι πλέον σημασμένη με βιοτίνη. Για την σήμανση, χρησιμοποιούμε τις ουσίες στρεπταβιδίνη/φυκοερυθρίνη συζευγμένες με χρωστική Cy3/Cy5, που προσδένεται στην βιοτίνη, και τα μικρά κομμάτια cRNA υβριδοποιούνται στο γονιδιακό τσίπ. Μετά περνάει από πλύση για να αφαιρεθούν οι μη υβριδοποιημένες αλληλουχίες, η

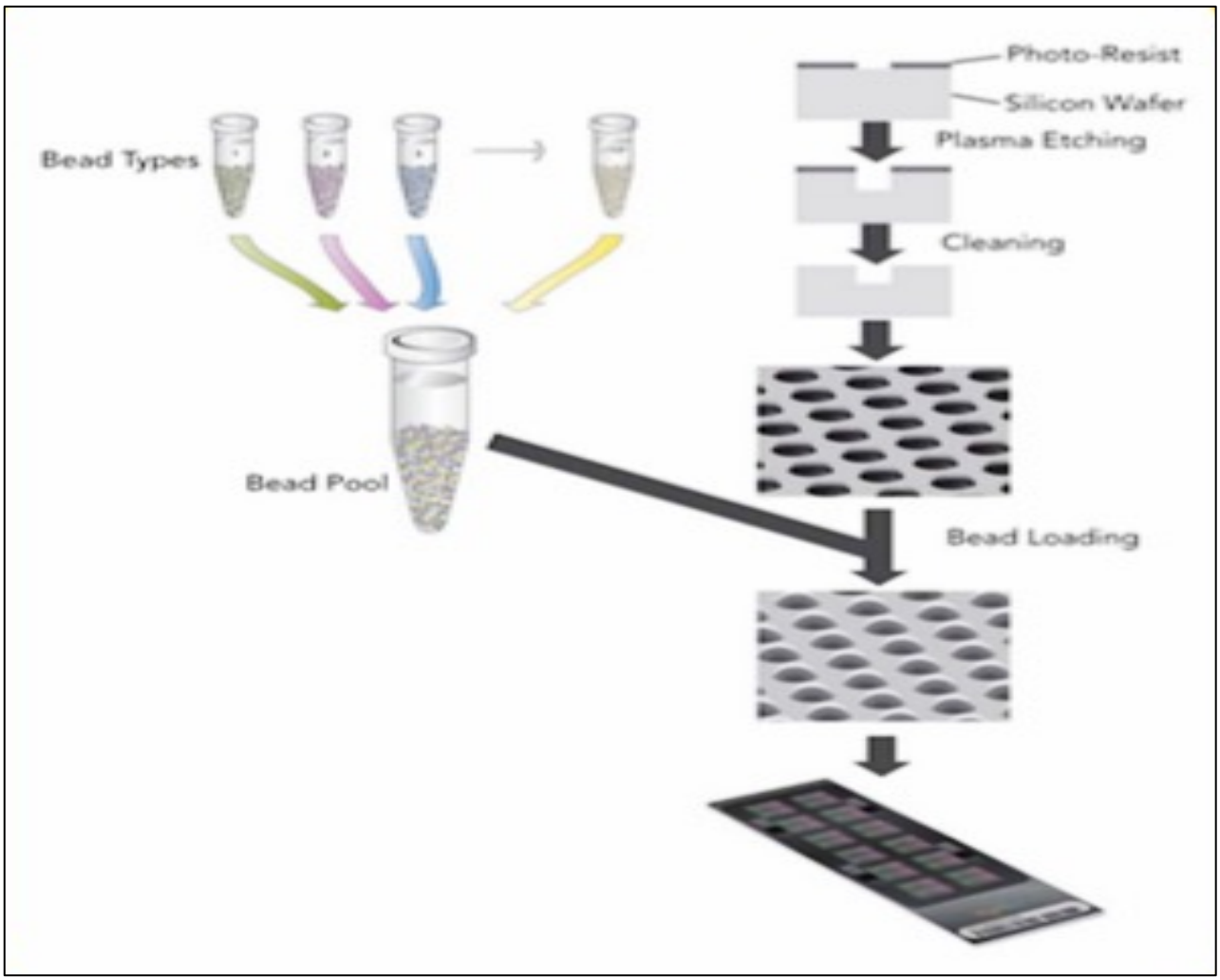
μικροσυστοιχία τοποθετείται σε έναν σαρωτή μικροσυστοιχιών και αποτυπώνεται ψηφιακά η γονιδιακή έκφραση (35). Η διαφορά είναι ότι εδώ μετράμε την γονιδιακή έκφραση μέσω της διαφοράς στην ένταση PM-MM, δηλαδή της τιμής κατά το ήμισυ απόλυτης ταύτισης του mRNA (half Perfectly Match mRNA, PM) και της τιμής κατά το ήμισυ μη ταύτισης (half have one Mismatch, MM).

Μια άλλη τεχνική κατασκευής τους είναι η inkjet εκτύπωση ψεκασμού με την πλατφόρμα SurePrint (25-μερές, 45-μερές, 60-μερές), μια πατέντα της εταιρίας Agilent, η οποία έχει πολλά είδη εκτυπωτικής πένας (nozzles) ψεκασμού. Είναι παρόμοια με τις πένες που χρησιμοποιήθηκαν στην τεχνολογία spotting με τα ρομποτικά μηχανήματα, με την διαφορά ότι αυτά χρειάζονται επαφή με το πλακίδιο για να εναποθέσουν τους ανιχνευτές. Αυτές οι μικροσυστοιχίες έχουν υψηλότερο ποσοστό ειδικότητας και την ευελιξίας στην ανάλυση, ενώ τα αποτελέσματα έχουν καλύτερη ποιότητα, κυρίως αν συγκριθούν με spotted μικροσυστοιχίες DNA που κατασκευάζονται με ρομποτικό μηχάνημα σε ένα εργαστήριο και μπορούν να περιέχουν σφάλματα στην κατασκευή τους.

Μια διαφορετική επιλογή για την κατασκευή των μακρομερών ολιγονουκλεοτιδικών μικροσυστοιχιών (50-μερές) είναι τα μικροσφαιρίδια τσίπ (BeadChips), που είναι τεχνολογία συστοιχιών μικροσφαιριδίων BeadArray της εταιρίας Illumina (San Diego, CA, USA). Είναι μια τεχνολογία που χρησιμοποιείται στην γενωμική επιστήμη για να αναλύσει την γονιδιακή έκφραση, την μεθυλίωση και τον γενότυπο των SNPs, εκτός των άλλων. Όπως φαίνεται στην παρακάτω εικόνα, πάνω στο τσίπ είναι μια συστοιχία μικροσφαιριδίων πυριτίου διαμέτρου 2μm, κάθε μια μπαίνει σε ένα "πηγαδάκι", και οι ολιγονουκλεοτιδικές αλληλουχίες προσδένονται με ομοιοπολικό δεσμό σε κάθε μικροσφαιρίδιο, ενώ ο κάθε τύπος μικροσφαιριδίου έχει μοναδική ολιγονουκλεοτιδική αλληλουχία (**Εικόνα 9**). Κάθε μικροσφαιρίδιο έχει περίπου 800.000 ολιγονουκλεοτίδια, ενώ τα μικροσφαιρίδια συναρμολογούνται τυχαία και αντιγράφονται πολλές φορές και σε κάθε τύπο μικροσφαιριδίων ανήκουν περίπου 15 μικροσφαιρίδια (36).

Η πειραματική διαδικασία, που απεικονίζεται στην **Εικόνα 10.**, για την σύγκριση μεταξύ δύο δειγμάτων με όλιγο μικροσυστοιχίες (Οι διαφορετικές επιλογές που έχει ο βιολόγος να κάνει το πείραμα είναι πολλές και εξαρτώνται όχι μόνο από τα μηχανήματα και τα αντιδραστήρια που χρησιμοποιεί αλλά και τι εξετάζει).

Αρχικά γίνεται η απομόνωση από ιστούς ή κύτταρα του ολικού RNA (total RNA) με την εφαρμογή ειδικών διαλυτικών ουσιών (αντιδραστηρίων), την φυγοκέντρηση και διαδοχικές πλύσεις. Μετά την απομόνωση του ολικού RNA θέλουμε να επιλέξουμε το mRNA, επειδή το μετάγραφο περιέχει πληροφορίες γονιδιακής έκφρασης, αυτό γίνεται με την βοήθεια των υποκινητών εκκινητών ολιγονουκλεοτιδίων δεοξυθυμιδίνης (deoxythymidine oligonucleotides, oligo-dT), που ενσωματώνεται στα mRNAs, στις μεγάλου μήκους ουρές poly-A, 50-200 βάσεων.



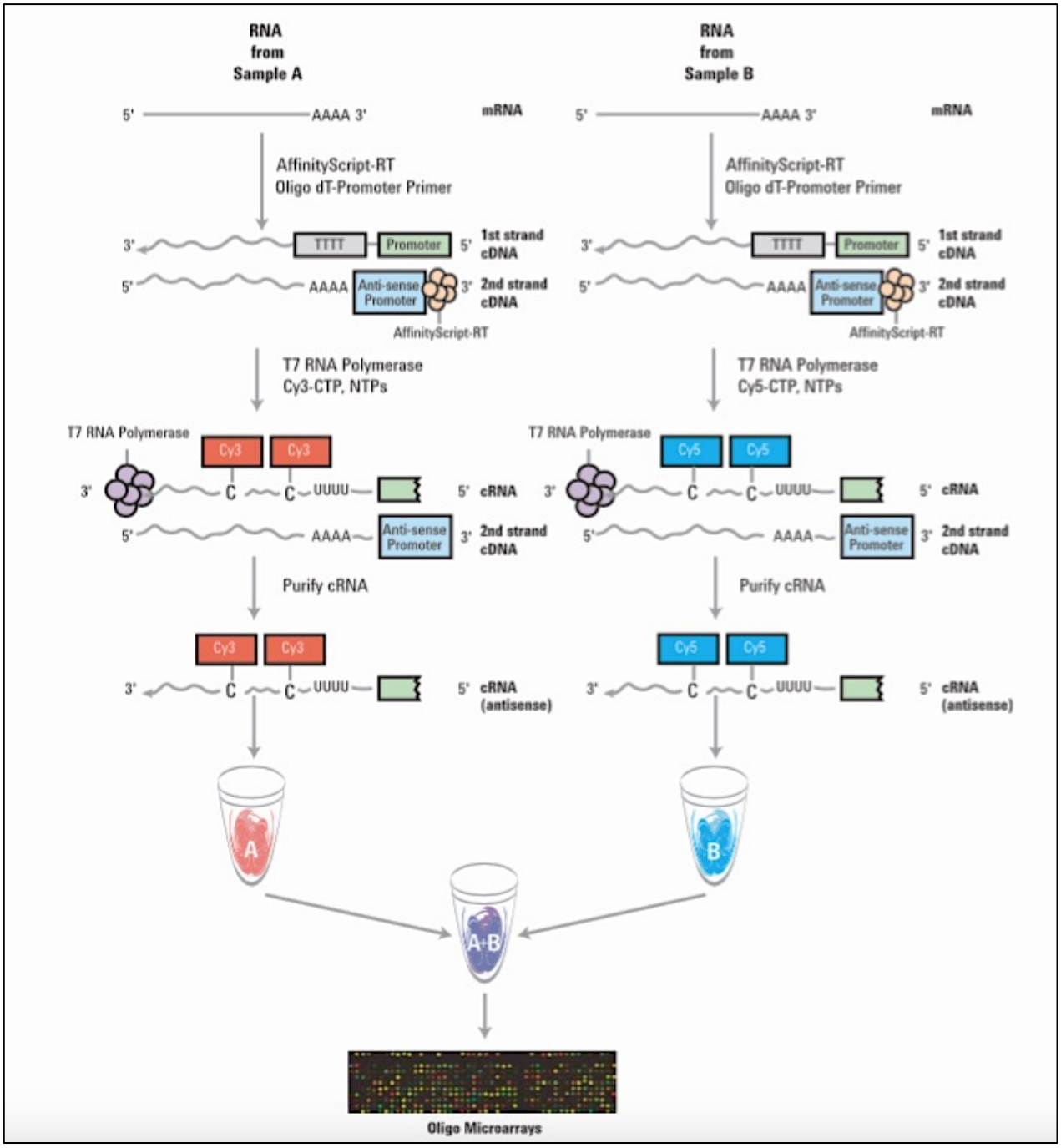
Εικόνα 9. Προετοιμασία μικροσφαιριδίων και τοποθέτηση γενετικού υλικού στα "πηγαδάκια" (Illumina).

Όλα τα mRNA μετατρέπονται σε cDNA με την βοήθεια του ενζύμου αντίστροφη μεταγραφάση, εδώ μπορεί να γίνει και η επαύξηση PCR (RT- PCR ή qPCR). Η αντίστροφη μεταγραφή (Reverse Transcription, RT) είναι η αντίστροφη διαδικασία από τον πολυμερισμό, στον οποίο το DNA γίνεται RNA, δηλαδή με την διαδικασία RT το RNA γίνεται DNA.

Γίνεται η μεταγραφή του cDNA σε cRNA μέσω της RNA πολυμεράσης και παράλληλα ενσωματώνονται ενζυμικά τα φθοριούχα μόρια, που συνήθως είναι τα Cy3 και Cy5, ξεχωριστά σε κάθε μείγμα.

Μετά την σήμανση αυτά αναμειγνύονται και υβριδοποιούνται στις μικροσυστοιχίες. Η σύνθεση των μικροσυστοιχιών γίνεται με την μέθοδο λιθογραφίας. Αυτή η διαδικασία γίνεται με αυτοματοποιημένο τρόπο από ειδικά μηχανήματα ενώ μπορείς να αγοράσεις έτοιμη την μικροσυστοιχία από διάφορες εταιρίες και να μην κάνεις χρονοβόρες διαδικασίες, ενώ τα αποτελέσματα του πειράματος θα μπορούν εύκολα να επαναληφθούν από κάποιο άλλο εργαστήριο. Τα υβρίδια DNA-cRNA δημιουργούνται λόγω της συμπληρωματικότητας των νουκλεοτιδικών βάσεων. Έτσι, το κάθε φθορισμένο μόριο cRNA αν βρεθεί συμπληρωματικό μονόκλωνο μόριο DNA σε κάποια κηλίδα της μικροσυστοιχίας DNA υβριδοποιείται σε αυτήν, με προσκόλληση των δύο κλώνων.

Μετά την υβριδοποίηση γίνεται πλύσιμο της μικροσυστοιχίας με ένα ειδικό υγρό και με αυτόν τον τρόπο αφαιρούνται όλα τα μονόκλωνα μόρια cRNA που δεν υβριδοποιήθηκαν.
Με την βοήθεια ομοεστιακού λέιζερ γίνεται η κατάλληλη διέγερση της φθορίζουσας ουσίας και αποτυπώνεται ψηφιακά η εικόνα των spots με το μηχάνημα σάρωσης μικροσυστοιχιών και, μετά από κανονικοποίηση, έχουμε ποσοτικοποίηση των αποτελεσμάτων σύμφωνα με την ένταση των κηλίδων, που δίνεται σε αποχρώσεις του γκρί. Η θέση της κηλίδας κάνει δυνατή την αναγνώριση του γονιδίου που ποσοτικοποιείται. Αυτά τα ελέγχουμε για την ποιότητα των αποτελεσμάτων (QC) και μπορούμε να τα επεξεργαστούμε μαθηματικά και βιοπληροφορικά.
Ένας σημαντικός σύμμαχος στην επαναχρησιμοποίηση και την αναπαραγωγή των αποτελεσμάτων στις μικροσυστοιχίες είναι τα μετα-δεδομένα (metadata), τα οποία είναι ενσωματωμένα στην εικόνα και αποθηκεύουν πληροφορίες για τις παραμέτρους αποκομιδής της εικόνας, που εφαρμόζονται στον σαρωτή των μικροσυστοιχιών. Η αποθήκευση αυτών των παραμέτρων κάνει δυνατή την επαναχρησιμοποίησή τους και σε άλλα δείγματα του πειράματος, έτσι ώστε να είναι συγκρίσιμα τα αποτελέσματα. Τα σωστά μετα-δεδομένα είναι σημαντικά και για την ακρίβεια της ποσοτικοποίησης και της στατιστικής ανάλυσης. Χρησιμοποιούνται για την ανάλυση της έντασης και την επεξεργασία των αποτελεσμάτων. Η σωστή σάρωση μπορεί να βοηθήσει στην αποφυγή τεχνικών σφαλμάτων (artifacts) ή στην διόρθωσή τους με την βοήθεια των μετα-δεδομένων.



Εικόνα 10. Παραγωγή cRNA για δικάναλη ανάλυση oligo μικροσυστοιχιών από την εταιρία Agilent Technologies. Για να χρησιμοποιηθεί με oligo μικροσυστοιχίες 60-μερές Agilent Gene Expression. Οι μικροσυστοιχίες κατασκευάζονται με την τεχνολογία Agilent SurePrint (37).

2.1.2. Μικροσυστοιχίες miRNA

Τα τελευταία χρόνια είναι ανανεωμένο το ενδιαφέρον για την ανίχνευση των miRNAs επειδή θεωρούνται σημαντικοί βιοδείκτες για τον καρκίνο όχι μόνο ως διαγνωστικά και προγνωστικά εργαλεία αλλά και στον θεραπευτικό ρόλο τους, αφού έχει ρόλο στην υποέκφραση και υπερέκφραση των γονιδίων. Τα miRNAs όμως είναι τα ρυθμιστικά μόρια

ολόκληρου του οργανισμού και υπάρχουν πάνω από 5.600 τέτοια μόρια στον ανθρώπινο οργανισμό, τα 2.000 από τα οποία εκφράζονται. Επομένως, έχουν καθοριστικό ρόλο σε πολλές ασθένειες εκτός του καρκίνου, όπως είναι ο διαβήτης και η ασθένεια του Κρον (Crohns' Disease).

Οι μικροσυστοιχίες είναι η συνήθης μέθοδος για την μέτρηση των επιπέδων miRNA αλλά και για την αναγνώριση νέων υπογραφών miRNA, ενώ καθιστά δυνατή την παράλληλη ανάλυση διαφορετικών δειγμάτων. Από την άλλη, η μέθοδος μικροσυστοιχιών δεν είναι κατάλληλη για την αναγνώριση νέων αλληλουχιών miRNA, ούτε μπορεί να ξεχωρίσει μεταξύ πρώιμου και ώριμου miRNA, όπου υπερτερούν οι μέθοδοι αλληλούχισης NGS, και ούτε για να ξεχωρίσει διαφοροποιήσεις μεταξύ είδη miRNA (38-40). Η πειραματική διαδικασία αρχίζει και σε αυτή την περίπτωση με την απομόνωση σταθερών μορίων συνολικού RNA, σε ποσότητα περίπου 100-150 ng, χρησιμοποιώντας αντιδραστήρια συντήρησης για να διασφαλιστεί η ακεραιότητα του RNA μετά από την λήψη του από τον ιστό ή τα κύτταρα. Στην περίπτωση της NGS χρειάζεται μεγαλύτερη ποσότητα συνολικού RNA για την ανάλυση miRNA.

Υπάρχουν στην αγορά ειδικές πλακέτες για την ανάλυση μικροσυστοιχιών miRNA, που χρησιμοποιούνται πλέον αρκετά χρόνια και είναι η προτιμότερη λύση από να φτιαχτεί στο εργαστήριο, και «κιτ» σήμανσης που είναι κατάλληλα για στόχευση miRNAs. Οι μικροσυστοιχίες για τα miRNAs κατασκευάζονται όπως αυτές που στοχεύουν μόρια DNA ή RNA, απλώς τα τοποθετημένα ολιγονουκλεοτίδια σε αυτή την περίπτωση εκτός από DNA/RNA μπορεί να είναι και άλλα νουκλεϊνικά οξέα, όπως τα LNA (locked nucleic acid, κλειδωμένο νουκλεϊνικό οξύ), τα οποία έχουν μεγαλύτερη ειδίκευση στα miRNA και έχουν πιο κανονικοποιημένα αποτελέσματα, κυρίως όταν η υβριδοποίηση γίνεται σε πολύ υψηλές θερμοκρασίες (41).

Ένα παράδειγμα είναι το miRCURY LNA microRNA Array (Exiqon, Denmark), η δανέζικη εταιρία κατασκευάζει μικροσυστοιχίες με τεχνολογία LNA, οι οποίες μπορούν παράλληλα να ποσοτικοποιήσουν τα miRNA σε άνθρωπο, ποντίκι και αρουραίο. Ένα άλλο παράδειγμα είναι τα TaqMan Array Human MicroRNA (Thermo Fisher Scientific Inc.,USA) χρησιμοποιούν επαύξηση PCR και λειτουργούν με έμμεση σήμανση, δηλαδή το σήμα φθορισμού TaqMan παράγεται όταν η πολυμεράση Taq εκτείνει τους εκκινητές και ελευθερώνει τη βαφή πάνω στον ανταποκριτή (reporter). Η DNA πολυμεράση Taq είναι θερμοσταθερή και προέρχεται από το βακτήριο *Thermus aquaticus*. Η εταιρία Agilent, από την άλλη, κατασκευάζει

μικροσυστοιχίες που μπορούν να ποσοτικοποιήσουν παράλληλα όλα τα ανθρώπινα miRNAs που αντιπροσωπεύονται από την πιο πρόσφατη έκδοση της βάσης δεδομένων miRBase (42).

Η σήμανση είναι περίπλοκο θέμα λόγω του μικρού μεγέθους των miRNAs, της μικρής ποσότητας τους στο βιολογικό δείγμα. Οπότε έχουν αναπτυχθεί διάφορες μέθοδοι έμμεσης και άμεσης σήμανσης του εξευγενισμένου μορίου miRNA (43). Το μόριο miRNA έχει σημανθεί με φθοριούχα χρωστική, με την βοήθεια της λιγκάσης T4 RNA. Ένα άλλο αρνητικό στην διαδικασία της υβριδοποίησης είναι ο υψηλός βαθμός ομοιότητας αλληλουχιών miRNA, που μπορεί να έχουν πολύ διαφορετική έκφραση ενώ διαφέρουν μόνο σε ένα νουκλεοτίδιο. Η υβριδοποίηση πάνω στην πλακέτα απαιτεί αρκετό χρόνο, δηλαδή κάμποσες ώρες ή ακόμα και μέρες, ενώ σε κάθε κηλίδα ολιγονουκλεοτιδίων προσδένεται ένα και μοναδικό είδος miRNA. Τα μόρια που δεν υβριδοποιηθούν αφαιρούνται μέσω διαδοχικών πλύσεων. Η μικροσυστοιχία μπαίνει στον σαρωτή και καταγράφεται η ένταση του φθορισμού. Στην συνέχεια γίνεται η επεξεργασία των αποτελεσμάτων (42).

Κάποια προβλήματα που μπορεί να προκύψουν στην διαδικασία του υβριδισμού είναι όταν κάποιες κηλίδες των μικροσυστοιχιών εκφράζουν άσχετα μεταξύ τους miRNA. Αυτό κάνει τα αποτελέσματα να διαφέρουν από πείραμα σε πείραμα και, ακόμα περισσότερο, ανάμεσα σε διαφορετικές πλατφόρμες, δηλαδή κάνει δύσκολη την αναπαραγωγή των αποτελεσμάτων, τα οποία απαιτούν ειδικές μέθοδοι κανονικοποίησης για miRNA (44). Συνεπώς, η ειδικότητα και η ευαισθησία είναι πολλές φορές ελλιπής στην ανάλυση miRNAs με μικροσυστοιχίες, αυτό σημαίνει ότι σε περιπτώσεις που αναλύεται miRNA χαμηλής συγκέντρωσης προτιμάται η NGS ανάλυση ή μέθοδοι ποσοτικοποιημένης επαύξησης PCR (42).

Ο Πίνακας 4 περιέχει εργαλεία για την αναγνώριση των μορίων miRNA που συγκρίνονται ταυτόχρονα ή επιβεβαιώνονται αργότερα με πειραματική διαδικασία μικροσυστοιχιών και τα πειραματικά δεδομένα γονιδιακής έκφρασης που προέρχονται από αυτά. Οι υπολογιστικές προσεγγίσεις αναπτύχθηκαν για να συμπληρώσουν τις πειραματικές μεθόδους της αναγνώρισης των γονιδίων miRNA, δηλαδή έχουν αναπτυχθεί και άλλες γενωμικές τεχνικές αναγνώρισης γονιδίων, οι οποίες δεν βασίζονται σε παραδοσιακές μεθόδους, όπως η ομολογία (homology) και η εγγύτητα (proximity) σε άλλα γνωστά γονίδια. Είναι σύνηθες, αυτές οι αναλύσεις να αναγνωρίζουν συντηρημένες (conserved) γενωμικές περιοχές, οι οποίες είναι έξω από τις προβλεπόμενες κωδικοποιές περιοχές και μπορεί πιθανώς να σχηματίζουν δομή μίσχου-θηλιάς. Αφού οι συντηρημένες (οι οποίες είναι υψηλού σκορ/ high

scoring ομοιότητας) γενωμικές περιοχές, όπου αναφερόμαστε στην σημαντικότητα των ίδιων γονιδίων σε όλα τα εξεταζόμενα είδη, είναι πιθανότερο να είναι λειτουργικές. Στη συνέχεια, βαθμολογούν αυτές τις υποψήφιες δομές miRNA για τα πρότυπα διατήρησης και ζευγάρωσης που χαρακτηρίζουν γνωστά μόρια miRNA. Για την αναγνώριση μη συντηρημένων γενωμικών περιοχών χρησιμοποιούνται μέθοδοι απαρχής (ab initio) αναγνώρισης, οι οποίες χρησιμοποιούν δομικά χαρακτηριστικά των miRNAs. Για να αναφέρουμε μόνο μερικές τεχνικές.

Πίνακας 4. Βιοπληροφορικά Εργαλεία για την αναγνώριση miRNA (45).

Εργαλείο	Οργανισμός	Τύπος Αλγορίθμου	Διάστημα Δημοσιεύσεων	Έκδοση	Λογισμικό ή Διαδικτύου	Σύνδεσμος και Βιβλιογραφία
miRscan	Ζώα	Εξελικτική διατήρηση	2003	2003	Δ	http://bartellab.wi.mit.edu/softwareDocs/MiRscan3/Introduction.html (46)
RNAz	Ζώα	Θερμοδυναμική Ισορροπία, Δομή	2005-2010	2011, v2.1	Λ+Δ	https://www.tbi.univie.ac.at/software/RNAz/ (47)
triplet-SVM	Φυτα, Ζώα	Μηχανική Μάθηση	2005	2005	Λ	http://bioinfo.au.tsinghua.edu.cn/mirnasvm/ (48)
MiPred	Φυτά, Ζώα	Μηχανική Μάθηση	2007	2016, v0.1	Δ	https://tools4mirs.org/software/precursor_prediction/mipred/ (49)
CID-miRNA	Ζώα	Μηχανική Μάθηση	2008	2008	Δ	https://github.com/tyagilab/cid-mirna (50)
MicroPC	Φυτά	Εξελικτική διατήρηση	2009	2009	Δ	http://www3a.biotec.or.th/micropc/ (51)
MatureBayes	Ζώα	Μηχανική Μάθηση	2010	2010	Λ+Δ	http://mirna.imbb.forth.gr/MatureBayes.html (52)
miRNAFold	Φυτά, Ζώα	Θερμοδυναμική Ισορροπία, Δομή	2012-2016	2016	Λ+Δ	https://evryrna.ibisc.univ-evry.fr/evryrna/mirnafold/mirnafold_home (53)
deepSOM	Φυτά, Ζώα	Μηχανική Μάθηση	2016	2016, v0.19	Λ+Δ	https://sinc.unl.edu.ar/web-demo/deepsom/ (54)

2.1.3 Ανάλυση Μεθυλίωσης με μικροσυστοιχίες

Ήδη ασχοληθήκαμε με την εφαρμογή των μικροσυστοιχιών στα μόρια miRNA, που μπορούν να αναστείλουν την μετάφραση, και, σε αυτό το σημείο της εργασίας, θα ασχοληθούμε με την εφαρμογή των μικροσυστοιχιών στον μηχανισμό μεθυλίωσης DNA, που μπορούν να αναστείλουν την μεταγραφή. Η μεθυλίωση DNA είναι από τις πιο εξονυχιστικά εξετασμένες επιγενετικές τροποποιήσεις, ο εν λόγω μηχανισμός χρησιμοποιεί τις DNA

μεθυλοτρανσφεράσες (DNMTs) για να μεταφέρουν μια μεθυλομάδα (– CH_3) από το ενεργό τους κέντρο σε ένα κατάλοιπο κυτοσίνης. Υπάρχουν πολλοί τύποι DNMTs, οι πιο σημαντικές είναι η DNMT1 για την διατήρηση της μεθυλίωσης και οι DNMT3a/DNMT3b για την καθιέρωση της μεθυλίωσης.

Εκτός των DNMTs, εμπλέκονται και άλλοι μηχανισμοί στη μεθυλίωση, όπως το TDG (Thymidine–DNA glycosylase) και το BER Pathway, για την αποκατάσταση της μεθυλίωσης. Η προσθήκη της μεθυλομάδας γίνεται, στους ευκαρυωτικούς οργανισμούς, κυρίως στον 5' άνθρακα του πυριμιδινικού δακτυλίου της κυτοσίνης και γι'αυτό ονομάζεται 5-μεθυλοκυτοσίνη (5-methylcytosine, 5mC). Ένας άλλος μηχανισμός που μας ενδιαφέρει είναι η απομεθυλίωση με τα ένζυμα TET (Ten Eleven Translocation), τα οποία οξειδώνουν τα 5mC σε 5-υδροξυμεθυλοκυτοσίνη (5-hydroxymethylcytosine, 5hmC) και με τα TDG μετατρέπεται σε απομεθυλιωμένη κυτοσίνη. Υπάρχουν και άλλα οξειδωτικά παράγοντα των 5mC, που χρησιμοποιούνται για καινοτόμες αναλύσεις στο γενωμικό DNA.

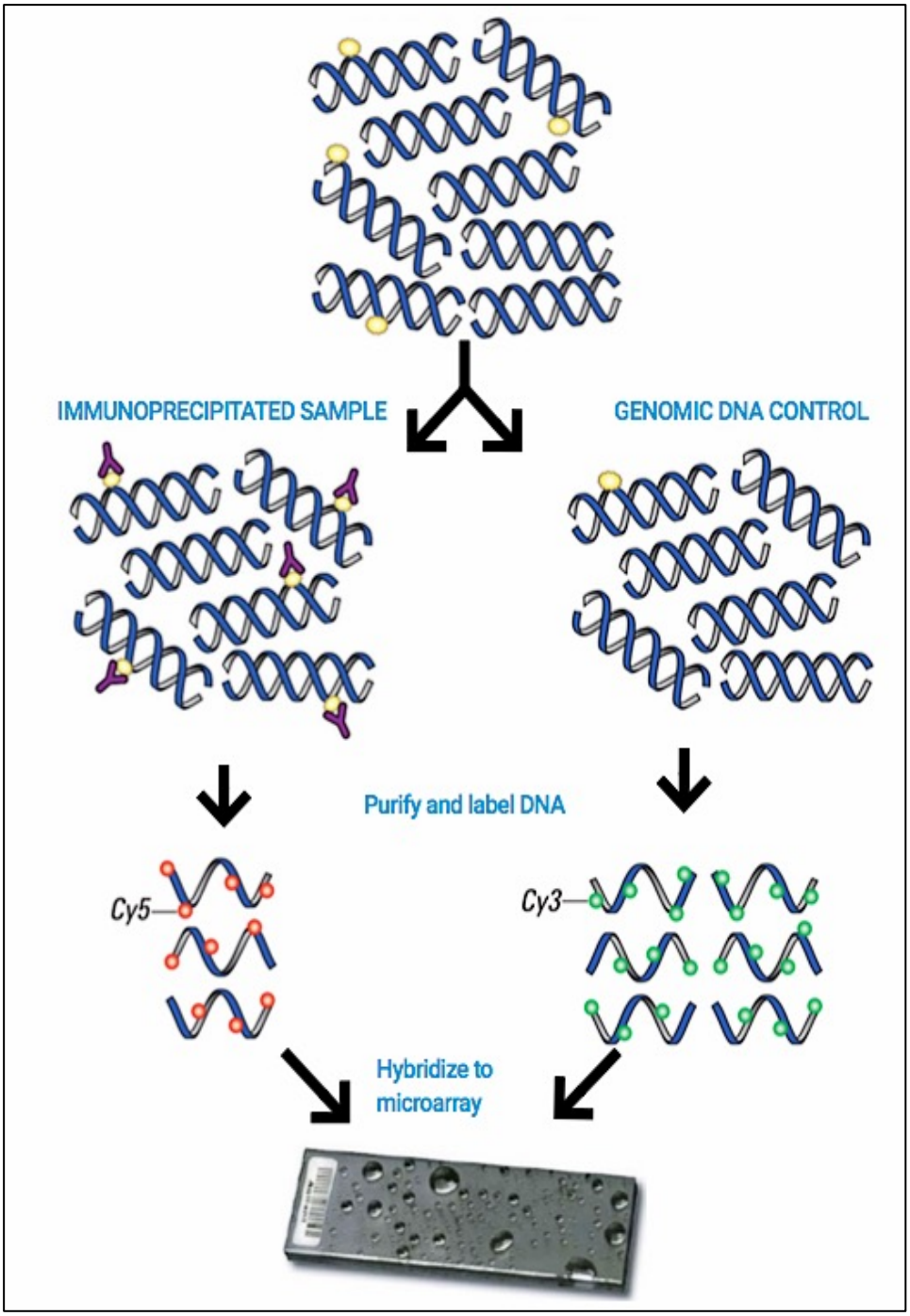
Από την διεξαγωγή της ευρείας μελέτης συσχέτισης πάνω από 1000 επιγενωμικών δεδομένων (Epigenome Wide Association Studies, EWAS) έχουν αναλυθεί ως προς την μεθυλίωση πάνω από 75.000 ανθρώπινα δείγματα (55, 56). Τα δεδομένα τέτοιων μελετών πρέπει να περάσουν από διάφορα στατιστικά τέστ και υποθέσεις. Στα θηλαστικά, η μεθυλίωση πέρνει χώρα κυρίως στις νησίδες CpG (Cytosine-phosphate-guanine islands, CGI), οι οποίες είναι περιοχές πλούσιες σε CpG, συνήθως στους υποκινητές και σε ρυθμιστικές περιοχές του γονιδίου. Παίζουν βασικό ρόλο στον μηχανισμό της γενωμικής αποτύπωσης (imprinting), που ρυθμίζει από ποιο χρωμόσωμα, το «πατρικό» ή το «μητρικό» δηλαδή, θα εκφραστούν τα γονίδια. Για παράδειγμα, το 70% με 80% των δινουκλεοτιδίων κυτοσίνης-θυμίνης (CpG dinucleotides) είναι μεθυλιωμένα στον άνθρωπο και το 60% γενικότερα στα θηλαστικά, στο υπόλοιπο ποσοστό επιτρέπεται η γονιδιακή έκφραση στον υποκινητή των γονιδίων των νησίδων CpG, δηλαδή η πρόσδεση μεταγραφικών παραγόντων στις θέσεις έναρξης μεταγραφής (TSS, transcription start site). Τα πρότυπα μεθυλίωσης στο κύτταρο συνήθως παίρνουν είτε την μορφή της ολικής υπομεθυλίωσης είτε την μορφή της υπερμεθυλίωσης των νησίδων CpG (57). Το 1999 ξεκίνησε το Πρόγραμμα Αλληλούχισης του Ανθρώπινου Επιγενώματος (Human Genome Project, HEP) για να καταγράψει αυτά τα πρότυπα μεθυλίωσης. Λόγω της εξονυχιστικής μελέτης της μεθυλίωσης στον άνθρωπο και της αντιστρεπτής φύσης της, ήταν δυνατό να αναπτυχθούν επιγενετικές θεραπείες, και μάλιστα με έγκριση του αμερικάνικου οργανισμού φαρμάκων (FDA), για διάφορες ασθένειες.

Οι μεθυλιωμένες περιοχές του γενώματος μπορούν να αναλυθούν με υβριδοποίηση σε μικροσυστοιχίες ή μικροσυστοιχίες μικροσφαιριδίων. Μια περίπτωση μικροσυστοιχιών είναι οι πλατφόρμες της Affymetrix, που είναι πλέον παρακλάδι της εταιρίας Thermo Fisher Scientific Inc. (Waltham, MA, USA), είναι το GeneChip® Human Promoter 1.0R Array (μέθοδος ChIP μόνο στους εκκινητές, όχι σε όλο το γένωμα), το GeneChip® Human Tiling 2.0R Array Set (μέθοδος ChIP που περιέχει όλες τις ταυτίσεις-PM από το 1.0R), το GeneChip® Human Tilling Array Set 1.0R και το πιο καινούργιο GeneChip® ENCODE01 1.0 Array, που είναι κατάλληλο και για de novo χαρτογράφηση. Αλλά πιο δημοφιλής είναι οι πλατφόρμες από την Illumina Infinium HumanMethylation450 BeadChip (450K), Methylation EPIC (850K). Μια άλλη πλατφόρμα είναι αυτή της Agilent Technologies με το κίτ Human CpG Island Microarray και η Human DNA Methylation Microarrays (MeDIP).

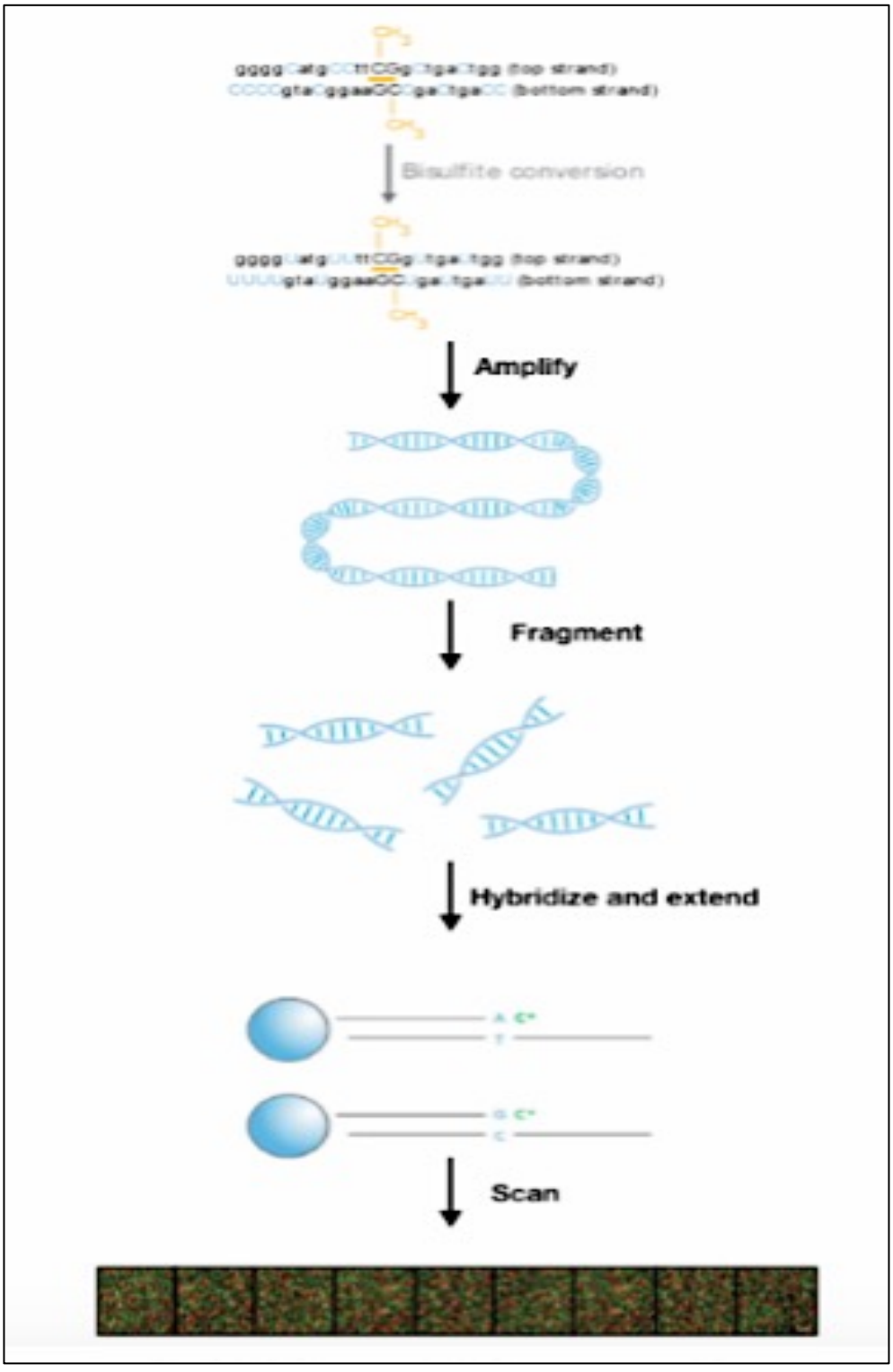
Αρχικά, οι βασικές τεχνολογίες ανάλυσης μεθυλίωσης σε ολόκληρο το γένωμα είναι ο εμπλουτισμός συνάφειας (affinity enrichment), η κατεργασία του δείγματος με δισουλφίδια και η κατεργασία περιορισμού . Η πρώτη μέθοδος βασίζεται στην υβριδοποίηση σε ολιγονουκλεοτιδικές μικροσυστοιχίες, που συνήθως αντιστοιχούν σε εκκινητές ή CGIs, με σημασμένα θραύσματα DNA, σε διαφορετικό κανάλι για τα εμπλουτισμένα για μεθυλίωση θραύσματα DNA και διαφορετικό κανάλι για τα θραύσματα ελέγχου. Η κύρια μορφή αυτής της μεθόδου είναι ο εμπλουτισμός των μεθυλιωμένων περιοχών για την ανοσοκαθίζηση γενωμικού DNA (Methylated DNA Immunoprecipitation, MeDIP), που χρησιμοποιεί συγκεκριμένα αντισώματα (αντί-5mC) για την μεθυλιωμένη κυτοσίνη, όπως φαίνεται στην **Εικόνα 11** (58). Μια άλλη μορφή εμλουτισμού είναι η καταγραφή μέθυλο-CpG πρωτεϊνικής καταγραφής (Methyl CpG binding protein capture, MBDCap). Τα επίπεδα της μεθυλίωσης καταγράφονται ως την ένταση της λογαριθμικής αναλογίας (log_2) μεταξύ των εμπλουτισμένων θέσεων και των θέσεων ελέγχου. Η δεύτερη μέθοδος είναι η δισουλφιδική μετατροπή του DNA ολόκληρου του γενώματος (WGBS), αν και προτιμάται η πιο στοχευμένη γενωμική ανάλυση π.χ. στο gDNA (genomic DNA) που αντιστοιχεί σε SNPs. Στην δεύτερη μέθοδο η υβριδοποίηση γίνεται σε συγκεκριμένα DNA ολιγομερή και η επέκταση μονών ζευγών (59). Η τρίτη μέθοδος είναι οι βασισμένες σε ένζυμα περιορισμού (νεοσχιζομερή, ισοσχιζομερή) και, πιο συγκεκριμένα, τις διαφοροποιημένες ιδιότητές τους κατά την πέψη. Ένα ζεύγος ισοσχιζομερών αναγνωρίζει την ίδια αλληλουχία και έχει το ίδιο σημείο διάσπασης, όμως εμφανίζει διαφορετική ευαισθησία στην μεθυλίωση DNA. Τα ευαίσθητα ως προς την μεθυλίωση (Methylation-sensitive restriction enzymes, MREs) ένζυμα

περιορισμού (π.χ. BstUI, HpaII, NotI, SmaI) διασπώνται μόνο στις αλληλουχίες-στόχους που δεν είναι μεθυλιωμένες, ενώ το μεθυλιωμένο μέρος της αλληλουχίας δεν αλλάζει. Η πέψη MRE χρησιμοποιείται για την εύρεση των επιπέδων μεθυλίωσης σε όλο το γένωμα (60, 61).

Infinium Methylation Arrays. Στην **Εικόνα 12** παρουσιάζεται η συστοιχία Infinium της εταιρίας Illumina, Inc. (San Diego, CA, USA), η οποία είναι μια σημαντική σειρά συστοιχιών μικροσφαιριδίων για την ανάλυση της μεθυλίωσης στο γένωμα. Προέρχεται από την πλατφόρμα Illumina GoldenGate για SNPs. Ανάλογα με τον τύπο Infinium περιέχει είτε 27.000 (27K BeadChip), είτε 485.512 (450K BeadChip), είτε 850.859 (850K EPIC BeadChip) θέσεις CpG που αναπαριστούν SNPs, από τις πιθανόν 28 εκατομμύρια θέσεις μεθυλιωμένων CpG στο ανθρώπινο γένωμα (62). Ο τύπος 27K έχει ξεπεραστεί πλέον καθώς δεν πληροί τις απαιτήσεις για κάποιες γενωμικές αναλύσεις ολόκληρου του γονιδιώματος. Τα Human BeadChips καλύπτουν την πλειονότητα των γονιδιακών υποκινητών, των μεταγραφικών παραγόντων και των UTRs, παρόλο που αντιστοιχούν γύρω στο 3% του συνόλου των GpGs (63).



Εικόνα 11. Μεθυλιωμένο gDNA με τεχνική MeDIP (Agilent) [https://www.cd-genomics.com/Microarray-Services.html].



Εικόνα 12. Μεθυλίωση της σειράς Infinium (Illumina) [https://www.cd-genomics.com/Microarray-Services.html].

Πιο συγκεκριμένα, όπως φαίνεται στην **Εικόνα 12**, το γενωμικό DNA που έχει υποστεί δισουλφιδική κατεργασία αναμειγνύεται με τα ολιγονουκλεοτίδια στις μικροσυστοιχίες μικροσφαιρίδιων. Κάποια νουκλεοτίδια είναι συμπληρωματικά στην ουρακίλη, που

προέρχεται από μη μεθυλιωμένη κυτοσίνη, και κάποια είναι συμπληρωματικά στην μεθυλιωμένη κυτοσίνη. Μετά την υβριδοποίηση στην μικροσυστοιχία, οι εκκινητές επεκτείνονται και περιδένονται στα ολιγονουκλεοτίδια συγκεκριμένων θέσεων CpG, έτσι ώστε να δημιουργήσουν εκμαγείο για την PCR. Οι σημασμένοι εκκινητές της PCR χρησιμοποιούνται για την ανίχνευση των μορίων και την καταγραφή της έντασης από το iS-can ή το NextSeq 550, σαρωτής μικροσυστοιχιών της Illumina, το επίπεδο μεθυλίωσης ποσοτικοποιείται σχετικώς μέσω της αναλογίας μεταξύ δύο τύπων μικροσφαιριδίων για κάθε θέση CpG, ενώ υπάρχει και η πιθανότητα οι ανιχνευτές να δώσουν ψευδώς θετικό λόγω της δέσμευσης παραπάνω από μια αλληλουχιών στον ίδιο ανιχνευτή.

Η σχετική ποσοτικοποίηση της μεθυλίωσης γίνεται με την εξέταση σε συγκεκριμένα σημεία και η δισουλφιδική μετατροπή του DNA είναι πολύ πιο αξιόπιστη αλλά και φτηνότερη από τεχνικές δισουλφιδικής μετατροπής ολόκληρου του γενώματος (WGBS), αφού αναλύεται μόνο ένα μικρό σημαντικό κομμάτι του γενώματος (64). Ακόμα, σε σχέση με την WGBS, μειώνουν την πολυπλοκότητα της ανάλυσης που προκύπτουν από την ανίχνευση των πολυμορφισμών. Στα αρνητικά της σειράς Infinium συγκαταλέγεται το ότι δεν διαχωρίζει ανάμεσα σε 5mC/5hmC και, με εξαίρεση το Infinium Mouse Methylation BeadChip, ειδικεύεται σε ανθρώπους (65).

Ο σκοπός των πειραμάτων μικροσυστοιχιών ή αλληλουχοποίησης νέας γενιάς είναι η ανάλυση των αποτελεσμάτων και η αξιολόγησή τους ως προς τα βιολογικά συμπεράσματα. Το λογισμικό σάρωσης θα δώσει ένα αρχείο με τα πρωτογενή δεδομένα, σε μορφή excel συνήθως, με πολλές στήλες δειγμάτων, συμπεριλαμβανομένου του δείγματος ελέγχου, δηλαδή από τον πίνακα γονιδιακής έκφρασης (γονίδιο/δείγμα) έχουμε λάβει πλέον τον λογαριθμισμένο αριθμό $e_{i,j}$ της έντασης φωτεινότητας για κάθε δείγμα σε περίπτωση μονοκάναλης πλατφόρμας, ενώ στην περίπτωση δικάναλης πλατφόρμας υπολογίζεται κατευθείαν η λογαριθμική ένταση φθορισμού του εξεταζόμενου δείγματος στο κόκκινο κανάλι και η λογαριθμική ένταση φθορισμού του δείγματος (ελέγχου) στο πράσινο κανάλι (**Εξίσωση 1**).

$$e_{i,j} = \log_2\left(\frac{R_{i,j}}{G_{i,j}}\right)$$

Εξίσωση 1. Τιμές φωτεινότητας του πίνακα γονιδιακής έκφρασης.

Η επεξεργασία των δεδομένων γίνεται με την χαρτογράφησή τους στο γένωμα, τον έλεγχο της ποιότητας τους, την στατιστική ανάλυσή τους και ίσως ένα πείραμα διαφορετικής μεθόδου για επιβεβαίωση των αποτελεσμάτων.

2.2. Αλληλουχοποίηση Νέας Γενιάς (Next Generation Sequencing (NGS))

Η αλληλουχοποίηση DNA είναι η διαδικασία με την οποία προσδιορίζεται η ακριβής σειρά των νουκλεοτιδίων μέσα στο μόριο του DNA. Καλύπτει οποιαδήποτε μέθοδο ή τεχνολογία που χρησιμοποιούνται για να υποδείξουν την σειρά των 4 βάσεων, σε ένα κλώνο του DNA. Παρόλο που η δομή του DNA είχε αναγνωριστεί ως διπλή έλικα το 1953, μόνο το 1970 κατόρθωσαν να αναλύσουν την αλληλουχία ενός τμήματος DNA στο εργαστήριο. Η πρώτη γενιά αλληλουχοποιητών κυκλοφόρησε στο εμπόριο το 1986 από την εταιρία Applied Biosystems (ABI), βασιζόταν στην μέθοδο αλληλουχοποίησης Sanger, που εφηύραν ο Frederick Sanger και οι συνεργάτες του το 1977. Περιλαμβάνει δημιουργία αλληλουχίας με επιλεκτική ενσωμάτωση διδεοξυριβονουκλεοτιδίων τερματισμού της αλυσίδας, είναι μια ολοκληρωτική ανάλυση και πολύ κατάλληλη για μεμονωμένα γονίδια. Ακόμα μπορεί να χρησιμοποιηθεί για στόχευση σε λιγότερο από 100 άμπλικονς.

Η δεύτερη γενιά των DNA αλληλουχοποιητών, γνωστοί και ως next-generation sequencers or NGS, έκανε δυνατή την καλύτερη μελέτη γενετικών, μεταγραφικών και επιγενετικών δεικτών σε επίπεδο γενώματος. Έτσι έγινε η ανάλυση πολλών οργανισμών, αφού αντίθετα με την πρώτη γενιά μπορεί να επεξεργάζεται ταυτόχρονα εκατομμύρια θραύσματα νουκλεϊνικών οξέων σε κάθε γύρο ανάλυσης. Η δεύτερη γενιά αλληλουχοποιητών αποτελείται από κάποιους πολύ δημοφιλής αλληλουχοποιητές όπως τις πλατφόρμες της Illumina, π.χ. MiSeq, που μπορούν να αναμετρώνται με αλληλουχοποιητές τρίτης γενιάς και ανήκουν στην τεχνολογία NGS. Η πλατφόρμες αυτές αποδίδουν από τα υψηλότερα ποσοστά ορθών βάσεων στην αλληλουχία με βαθμολογία Q30 (1 λάθος στις 1000 βάσεις, ακρίβεια 99,9%) στο δείκτη Phred, ελαχιστοποιώντας την ανίχνευση ψευδώς θετικών και ψευδώς αρνητικών. Σε πολλές εφαρμογές μεταγραφωμικής προτιμάται η NGS έναντι της γενωμικής ανάλυσης μικροσυστοιχιών, η οποία βασίζεται στον υβριδισμό, αφού η NGS είναι ψηφιακή και ποσοτικοποιημένη.

Οι πλατφόρμες της τρίτης γενιάς, με υψηλότερη απόδοση και μειωμένη τιμή, μπορούν να παράγουν μεγαλύτερου μήκους μικροαναγνώσεις (≤ 100 kb) ακόμα και από μονά μόρια DNA προερχόμενα από ελάχιστη ποσότητα γενωμικού. Κάποια δημοφιλή παραδείγματα είναι η

πλατφόρμα PacBio Sequel, η πλατφόρμα ONT MinION κ.α. . Το κόστος του προγράμματος ανθρώπινου γονιδιώματος υπολογίζεται, αν χρησιμοποιηθεί η μέθοδος πρώτης γενιάς αλληλουχοποίησης Sanger, στα 2,7 δισεκατομμύρια δολάρια, ενώ με την ίδια διαδικασία, αν χρησιμοποιηθεί η αλληλουχοποίηση νέας γενιάς, το κόστος μειώνεται στο ενάμιση εκατομμύριο (66).

Με την έλευση των τεχνικών εξαιρετικά υψηλής απόδοσης περνάμε από την αναλογική εποχή της αλληλούχισης στην ψηφιακή, ενώ οι μελέτες στις επιστήμες της ζωής προοδεύουν ραγδαία μέσω της γενωμικής και της συστηματικής βιολογίας. Η γενωμική επαναλληλούχιση, η γενωμική εκ νέου συναρμολόγηση, η μεταγραφομική, η επιγενωμική, η μεταγενωμική είναι κάποιες από τις τεχνικές που έχουν εφαρμοστεί στην ιατρική, την κλινική διάγνωση, την ανάπτυξη νέων φαρμάκων, τις βελτιώσεις στην γεωργία και την περιβαλλοντική προστασία. Οι τεχνικές εξαιρετικά υψηλών αποδόσεων μας δίνουν την ευκαιρία να κατανοήσουμε ολόκληρο το σύστημα και όχι μόνο τα επιμέρους κομμάτια του. Η αλληλουχοποίηση νέας γενιάς, βελτίωσε το κόστος ανάλυσης του γονιδιώματος και ελαχιστοποίησε τον χρόνο της ανάλυσης, αφού επέτρεψε την ταυτόχρονη διερεύνηση παραπάνω γενωμικών περιοχών και δειγμάτων.

Παράλληλα, προς αυτήν την κατεύθυνση συνεισέφερε και η ανάπτυξη περισσότερο εξελιγμένων αλγόριθμων βιοπληροφορικής ανάλυσης των δεδομένων. Η απουσία συγκεκριμένης αναλυτικής μεθοδολογίας, καθιστά μονόδρομο τη χρήση πακέτων βιοπληροφορικών αναλύσεων. Έτσι, η αναζήτηση για ακόμη πιο εξειδικευμένο προσωπικό σε συνδυασμό με την ανάγκη για αυξημένη υπολογιστική ισχύ, καθώς και μεγαλύτερους αποθηκευτικούς χώρους, λόγω του τεράστιου όγκου δεδομένων (big data), που απορρέουν από αυτές τις τεχνολογίες, αυξάνει εν μέρη το κόστος χρήσης της νέας τεχνολογίας.

Η τεχνολογία NGS απασχολεί πολλά διαφορετικά πεδία της βιολογίας, από έρευνες βακτηρίων και ιών μέχρι την έρευνα σε φυτά και ανθρώπινη αρρώστια. Οπότε χρειάζονται μεταξύ των χρηστών των δεδομένων NGS και αυτούς που σχεδιάζουν και ολοκληρώνουν τις διαδικασίες για την προετοιμασία των βιβλιοθηκών αλληλουχοποίησης, έτσι ώστε να απομονωθεί και να ενισχυθεί το σωστό στοιχείο. Εξίσου σημαντικοί είναι οι μηχανικοί που αναπτύσσουν τα διάφορα συστήματα μικροσυστοιχιών και τα αντίστοιχα μηχανήματα αλληλουχοποίησης ή οι βιοπληροφορικοί επιστήμονες που αναλύουν τα δεδομένα και τα μετατρέπουν σε χρήσιμες πληροφορίες (67).

Για ανακάλυψη των προτύπων γονιδιακής έκφρασης κάνουμε RNA αλληλούχιση, έτσι ώστε να διαλευκάνουμε τα μυστήρια των οδών ρύθμισης γονιδίων. Οι κυτταρικές διεργασίες είναι πολύπλοκες και οι ρυθμιστές μόνο συμβάλλουν σε αυτήν την πολυπλοκότητα. Η ανάλυση σε ένα βιολογικό δίκτυο αφορά διάφορα ρυθμιστικά στοιχεία και τον τρόπο με τον οποίο αυτά λειτουργούν στα κύτταρα. Με τις τεχνολογίες αλληλούχισης εξαιρετικά υψηλών αποδόσεων και τις μεθόδους ανάλυσης των βιολογικών δικτύων καθίσταται δυνατή η χαρτογράφηση κάθε γονιδίου σε σχέση με τον φαινότυπο του κάθε οργανισμού. Τα γενωμικά δεδομένα είναι όλο και πιο συχνά διαθέσιμα σε δημόσιες βάσεις δεδομένων. Πολλοί ερευνητές και ερευνητικά κέντρα γενωμικής, όπως το National Human Genome Research Institute (NHGRI), θεωρούν τις τεχνικές διαλογής εξαιρετικά υψηλής απόδοσης πολύ σημαντικές για το μέλλον της βιολογίας και έχουν σημαντική συνεισφορά και σε παράλληλες άλλες ωμικές αναλύσεις, για παράδειγμα την RNA αλληλούχιση (9).

Κάποιες τεχνολογίες που χρησιμοποιούνται στις τεχνικές εξαιρετικά υψηλών αποδόσεων: Υπάρχουν αρκετές μεγάλες εταιρίες που κατασκευάζουν αλληλουχοποιητές υψηλής απόδοσης: η τεχνολογία αλληλούχισης με σύνθεση (Sequencing By Synthesis, SBS) από την Illumina (San Diego, CA, USA), η τεχνολογία Ion Torrent από την Life Technologies, που είναι πλέον παρακλάδι της εταιρίας Thermo Fisher Scientific Inc. (Waltham, MA, USA), και η τεχνολογία Single molecule real time (SMRT) από την Pacific Biosciences (Menlo Park, CA, USA). Κάποιες από τις εταιρίες που δραστηριοποιούνται στην αλληλουχοποίηση νέας γενιάς DNA και RNA είναι η Illumina (NovaSeq, NextSeq), η Thermo Fisher Scientific (Ion Torrent: Ion GeneStudio S5, Ion Torrent Genexus, SOLiD), η Complete Genomics (DNA nanoball sequencing, DNB) και η Oxford Nanopore Technologies (MinION, GridION, PromethION). Αυτές οι εξελίξεις έγιναν δυνατές λόγω της καλύτερης κατανόησης της χημείας του DNA, και της υψηλής ευαισθησίας των καινοτόμων τεχνικών διαλογής εξαιρετικά υψηλής απόδοσης. Η ανάπτυξη υπολογιστικών μεθόδων και εργαλείων είναι απαραίτητη για την απόκτηση, αποθήκευση και ανάλυση δεδομένων μαζικής παραγωγής αλληλουχοποίησης DNA. Ο τομέας των τεχνολογιών αλληλουχοποίησης νέας γενιάς NGS μπορεί να έχει σημαντικές εφαρμογές όχι μόνο στην ζωή των ανθρώπων αλλά και σημαντικές βελτιώσεις στον τομέα της γεωργίας, της ζωολογίας και της προστασίας του περιβάλλοντος.

Κάθε τεχνολογία έχει τα πλεονεκτήματα της και τους περιορισμούς της, μερικά από τα οποία αναφέρονται στο επόμενο πίνακα (Πίνακας 5), έχοντας υπόψη μας βέβαια ότι η απόδοση εξαρτάται από την ποιότητα του δείγματος και από το μέγεθος του. Η απόφαση για το ποια

πλατφόρμα θα προτιμήσει να αγοράσει ένα εργαστήριο ή ποια πλατφόρμα θα χρησιμοποιήσει ένας ερευνητής που έχει πρόσβαση σε πάνω από ένα αλληλουχοποιητή θα εξαρτηθεί από την απαιτούμενη εφαρμογή. Επειδή η βιοτεχνολογία είναι ένας τομέας που εξελίσσεται διαρκώς, βγαίνουν συχνά στην αγορά βελτιωμένες πλατφόρμες από την κάθε εταιρία. Εκτός αυτού γίνονται συνεργασίες μεταξύ μεγάλων εταιριών και το αποτέλεσμα είναι νέες τεχνικές που αναπτύσσουν από κοινού και που, φυσικά, έχουν πλεονεκτήματα πάνω από μίας τεχνολογίας. Ένα τέτοιο παράδειγμα είναι η συνεργασία της Illumina με την Roche Clinical, η οποία υπόγραψε το 2020 ένα δεκαπενταετές συμβόλαιο με σκοπό την μικρότερη αναμονή για τα διαγνωστικά (in vitro diagnostic, IVD) τέστ, που είναι βασισμένα στην NGS των διαγνωστικών αλληλουχοποιητών της Illumina (Dx).

Είναι ενδιαφέρον ότι αυτή η ανακοίνωση έγινε μετά την αποτυχία εξαγοράς της εταιρίας Pacific Biosciences, ενός άλλου γίγαντα στις πλατφόρμες αλληλουχοποίησης, από την Illumina, επειδή θα είχαν μονοπώλιο στην αγορά (94%). Πολλά κράτη ακόμα κάνουν συνεργασίες για γενωμικό έλεγχο στους πολίτες. Συγκεκριμένα, η Αγγλία και το κρατικό σύστημα υγείας της (National Health Service, NHS) στοχεύει μέσω συνεργασίας της Illumina με την Genomics England, να έχει αλληλουχίσει, με πλατφόρμες αλληλουχοποίησης της Illumina, ανάμεσα σε 300.000 με 500.000 ασθενείς από το 2020 μέχρι το 2025 (68).

Η χρήση τεχνικών μαζικής παραγωγής δεδομένων έχει φέρει πολλές αλλαγές στην έρευνα από την ανθρώπινη βιολογία στην μικροβιολογία. Αυτές οι τεχνικές αναπτύχθηκαν για την ανακατασκευή μεγάλων γενωμάτων των πιο πολύπλοκων οργανισμών, όπως του ανθρώπινου, αλλά λόγω της αποτελεσματικότητάς τους χρησιμοποιούνται παντού. Η βασική εφαρμογή τους είναι η αλληλούχιση των βιολογικών δεδομένων ενός οργανισμού, για παράδειγμα στη μοριακή κλωνοποίηση, στη γονιδιακή συγκριτική αναγνώριση και στην εξελικτική επιστήμη. Κάποιες άλλες εφαρμογές είναι η αλληλούχιση εξόματος, δηλαδή η στοχευμένη αλληλούχιση των εξονίων ενός οργανισμού, η ανάλυση των διασυνδέσεων της χρωματίνης και του προτύπου των ριβοσωμάτων. Παράλληλα, τα τελευταία χρόνια έχουν ενσωματωθεί πολλές νέες τεχνολογίες στην λειτουργία του NGS. Η τεχνολογική πρόοδος και η αυξημένη αυτοματοποίηση στις τεχνικές διαλογής εξαιρετικά υψηλής απόδοσης μείωσαν το κόστος και την έκαναν εύκολα προσβάσιμη από όλους τους επιστήμονες (69). Συνεπώς, τα πειράματα γίνονται όλο και περισσότερο ακόμα και σε μικρά εργαστήρια και περισσότεροι ερευνητές έχουν πρόσβαση σε αυτά.

2.2.1. Ιστορική Αναδρομή και Αρχή Λειτουργίας της Τεχνικής NGS

Αλληλουχοποίηση πρώτης γενιάς. Η πρώτη γενιά των τεχνικών αλληλουχοποίησης DNA περιλάμβανε την μέθοδο χημικής αλληλούχισης που δημιουργήθηκε από τον Allan Maxam και τον Walter Gilbert το 1973 και την μέθοδο τερματισμού αλυσίδας που δημιουργήθηκε από τον Frederick Sanger το 1977. Αυτή η μέθοδος ονομάζεται κοινώς Sanger και είναι η βάση για την πρώτη ημιαυτόματη μηχανή αλληλουχοποίησης του DNA, και το 1978 αυτοματοποιήθηκε πλήρως από την εταιρία Applied Biosystems (ABI, που πλέον ανήκει στην Thermo Fisher). Στη συνέχεια, κατασκευάστηκαν βελτιωμένα μοντέλα, τα οποία χρησιμοποιήθηκαν στην αλληλούχιση του πρώτου βακτηριακού γενώματος το 1995, του πρώτου ευκαρυωτικού γενώματος το 1996 και της πρώτης δοκιμαστικής αλληλουχοποίησης του ανθρώπινου γονιδιώματος το 2001. Το 1996 εφευρέθηκε η μέθοδος πυροαλληλούχισης DNA, από τους Pål Nyrén και Mostafa Ronaghi. Ενώ το 2000 εφευρέθηκε η μαζικά παράλληλη αλληλουχοποίηση (massively parallel signature sequencing, MPSS) της Lynx Therapeutics, αλλά απέτυχαν να τις κυκλοφορήσουν στην αγορά.

Αλληλουχοποίηση Δεύτερης Γενιάς. Η δεύτερη γενιά των DNA αλληλουχοποιητών, ανήκουν στην κατηγορία των next-generation sequencers (NGS), έκανε δυνατή την καλύτερη μελέτη γενετικών, μεταγραφικών και επιγενετικών δεικτών σε επίπεδο γενώματος Η δεύτερη γενιά αλληλουχοποιητών DNA παράγει μαζικά βιολογικά δεδομένα. Το πρώτο συστήματα αλληλουχοποίησης νέας γενιάς, το μηχάνημα GS, βασισμένο στην πυροαλληλούχιση κυκλοφόρησε στο εμπόριο το 2005 από την εταιρία 454 Life Sciences με το όνομα όργανο GS και η εταιρία εξαγοράστηκε από την ελβετική Roche Diagnostics δύο χρόνια αργότερα. Σε σχέση με τους άλλους προχωρημένους αλληλουχοποιητές, που βασίζονται στην μέθοδο Sanger, η μηχανή της εταιρίας Lifesciences είχε ρίξει το κόστος στο ένα έκτο. Αργότερα όμως η τεχνολογία ξεπεράστηκε και η παραγωγή του σταμάτησε το 2015. Το 2006 κυκλοφόρησε στο εμπόριο η πλατφόρμα SOLiD (Sequencing by Oligo Ligation Detection) της ABI, που βασίζεται στην μέθοδο αλληλουχοποίησης ανίχνευσης ολιγονουκλοτιδίων λιγκασών. Στο αρχικό μοντέλο της SOLiD το μέγεθος των μικροαναγνώσεων ήταν 35 bp, ενώ ήταν βασισμένη σε τεχνική αλληλούχισης δύο βάσεων με με βαθμολογία Q30. Ένα επόμενο μοντέλο της SOLiD παρήγαγε 85 bp μικροαναγνώσεις με μεγαλύτερη ακρίβεια και υψηλότερης απόδοσης δεδομένα. Το 2007 μια γρήγορα αναπτυσσόμενη εταιρία γενωμικής, η Illumina, αγόρασε την εταιρία Solexa έναντι $650 εκατομμυρίων δολαρίων. Το ενδιαφέρον για τον πρότυπο αλληλουχοποιητή της Solexa ήταν ο λόγος για την εξαγορά, που είχε πρόσφατα εγκατασταθεί στα δύο μεγάλα κέντρα γενωμικής έρευνας, το Wellcome Trust

Sanger Institute στην Αγγλία και το Broad Institute στην Β. Αμερική. Αυτή η κίνηση τους έφερε ένα βήμα μπροστά από τους αντιπάλους τους, την εταιρία Affymetrix και την Life Technologies, οι οποίες πλέον ανήκουν στην Thermo Fisher Scientific Inc. (Waltham, MA, USA) (70).

Το 2010 η Illumina κυκλοφόρησε στην αγορά την πλατφόρμα αλληλουχοποίησης HiSeq 2000, που βασίζεται στην μέθοδο σύνθεσης. Τότε άρχισε η αλληλουχοποίηση ανθρώπινου γενώματος από διάφορους πληθυσμούς, έναντι εκατομμυρίων δολαρίων για το κάθε άτομο, σε πανεπιστήμια, ιδιωτικές εταιρίες και ινστιτούτα. Ο αλληλουχοποιητής HiSeq 2000 αποδίδει 600Gb ανά γύρο όταν είναι σε υψηλή απόδοση. Η εταιρία Illumina ισχυρίστηκε ότι στο μηχάνημα HiSeq 2000 ήταν φτηνότερη η αλληλούχιση σε σύγκριση με την πλατφόρμες 454 και την SOLiD. Το 2011 η Illumina κυκλοφόρησε στην αγορά την πλατφόρμα αλληλουχοποίησης MiSeq, επίσης βασισμένη στην αλληλουχοποίηση με σύνθεση, όμως είναι μηχάνημα μικρότερης κλίμακας. Το σύστημα MiSeq χρειάζεται μόνο μια ημέρα για να ολοκληρώσει την ανάλυση και αποδίδει μέχρι 15 Gb δεδομένων αλληλούχισης ανά γύρο ανάλυσης. Η πλατφόρμα MiSeq με αλληλούχιση προς την μία και μετά προς την αντίθετη κατεύθυνσή της (paired-end) μπορεί να παράξει το πολύ 250 bp μικροαναγνώσεις. Η επόμενη πλατφόρμα, η HiSeq 2500 μπορεί να αποδώσει 1 TB. Παρόλο που η σειρά HiSeq θεωρούνταν η καλύτερη τεχνική αλληλούχισης εξαιρετικά υψηλής απόδοσης το μήκος των μικροαναγνώσεων που αποδίδει ήταν το πολύ 250 bp, ενώ η πλατφόρμα HiSeq έχει ξεπεραστεί πλέον. Η απαιτούμενη PCR πριν την αλληλούχιση προκαλεί συχνά στατιστικά σφάλματα προκατάληψης (70).

Η πλατφόρμα Ion Torrent δημιουργήθηκε από την εταιρία PostLight Sequencing Technology το 2010 και μετά εξαγοράστηκε και διατέθηκε στην αγορά από την εταιρία Life Technologies Corp (η οποία αργότερα εξαγοράστηκε από την Thermo Fisher Scientific Inc.). Η τεχνολογία αλληλουχοποίησης Ion Torrent ουσιαστικά ελέγχει τα επίπεδα ιόντων υδρογόνου που εκλύονται σαν υποπροϊόντα κατά την ενσωμάτωση νουκλεοτιδίων (71). Οι αλληλουχίες με ομοιοπολυμερικές βάσεις είναι ένα μεγάλο μειονέκτημα της αλληλουχοποίησης Ion Torrent, αφού προκαλούνται αποκλίσεις ενθέσεων ή διαγραφών, και σε κάποιες περιπτώσεις, μπορεί να συμβεί αντικατάσταση βάσεων. Με την πλατφόρμα Proton, τα υψηλής πυκνότητας πηγαδάκια των Ion τσίπ μπορούν να παράγουν μέχρι 10 Gb σε κάθε ανάλυση. Με 200bp μικροανάγνωση, η πλατφόρμα Ion Proton μπορεί να εφαρμοστεί σε RNA-Seq και σε ταυτόχρονη αλληλουχοποίηση των άμπλικονς προς δύο αντίθετες κατευθύνσεις. Η δεύτερη

γενιά αλληλουχοποίησης έχει το μειονέκτημα ότι παράγει μικρού μήκους μικροαναγνώσεις και αυξημένο ποσοστό σφαλμάτων προκατάληψης (70).

Αλληλουχοποίηση Τρίτης Γενιάς. Τα κεντρικά χαρακτηριστικά της η τρίτης γενιάς αλληλουχοποίησης συμπεριλαμβάνουν τεχνικές που δεν χρειάζονται ενίσχυση και η ανίχνευση σημάτων γίνεται σε πραγματικό χρόνο κατά την διάρκεια της διαδικασίας αλληλούχισης (72, 73). Η εταιρεία Pacific BioSciences κατασκεύασε την πλατφόρμα αλληλουχοποίησης τρίτης γενιάς Single Molecule Real Time (SMRT) PacBio RS. Σε αυτή την τεχνολογία, κατά την διάρκεια ενζυματικών αντιδράσεων της ενσωμάτωσης νουκλεοτιδίων στον συμπληρωματικό κλώνο, η φθορίζουσα χρωστική του νουκλεοτιδίου που ενσωματώνεται, διαχωρίζεται και ανιχνεύεται αμέσως το σήμα (74). Υπερτερεί στο ότι δεν χρειάζεται ενίσχυση, συνεπώς μειώνεται το περιθώριο σφαλμάτων και ελαττώνεται ο απαιτούμενος χρόνος για την ανάλυση. Η πλατφόρμα PacBio RS είναι χαμηλής απόδοσης αλλά έχει μεγαλύτερου μήκους μικροαναγνώσεις (με μέση τιμή μικροαναγνώσεων μερικά kb) και χαμηλότερο βαθμό σφαλμάτων, αφού πρόκειται για τυχαία σφάλματα και όχι στατιστικές προκαταλήψεις (75, 76). Αυτή και η επόμενη εκδοχή αυτής της πλατφόρμας, η RSII, είναι πλέον ξεπερασμένες και η PacBio θα σταματήσει να τις υποστηρίζει μέχρι το τέλος του 2021. Οι μεγαλύτερες αλληλουχίες που αναπτύσσονται από τις διατασσόμενες μικροαναγνώσεις, λέγονται συνεχόμενες αλληλουχίες ή contigs. Το ποσοστό σφάλματος των μικροαναγνώσεων της αλληλούχισης μονού περάσματος (single pass) είναι μικρό, ενώ τα contigs έχουν υψηλή ακρίβεια (99.999%), δηλαδή δείκτη Phred Q50, όταν επιλέγεται η κυκλικά συναινετική αλληλούχιση (Circular Consesnus Sequencing, CCS), αφού το καλούπι των μορίων DNA αλληλουχοποιείται πολλές φορές όταν ενεργοποιείται αυτή και τα σφάλματα ελαχιστοποιούνται (77, 78). Οι περισσότερες αλληλουχοποιημένες μικροαναγνώσεις δεν έχουν καμία αστοχία ή ένθεση-διαγραφή επειδή έχουν μεγάλο μήκος τα contig, οπότε η πλατφόρμα έχει λιγότερες ασυνέχειες κατά την διεργασία συναρμολόγησης του γενώματος. Έτσι, είναι κατάλληλη και για την εκ νέου αλληλούχιση γενώματος (70, 79).

Μια άλλη πλατφόρμα τρίτης γενιάς, η Oxford Nanopore MinION, κάνει αλληλουχοποίηση με την βοήθεια των nanopores μιας συγκεκριμένης πρωτεΐνης, της alpha hemolysin (80). Είναι μία συσκευή στο μέγεθος ενός κινητού τηλεφώνου και σχετικά φτηνή. Η αλληλούχιση νανοπόρων έχει και το πλεονέκτημα ότι θέλει μειωμένο χρόνο για την προετοιμασία των δειγμάτων. Το μονόκλωνο DNA δεν εξετάζεται κατά την σύνθεση ή τον πολυμερισμό, αλλά κατά τον αποπολυμερισμό, οπού χρειάζεται μόνο η εξωνουκλεάση. Άρα δεν γίνεται

επαύξηση PCR, αλλά ούτε φθορίζουσα σήμανση, αφού η ανίχνευση γίνεται σύμφωνα με την αλλαγή τάσης στην νανοπορώδη επιφάνεια (nanopore), από δεοξυριβονουκλεοσίδια μονοφωσφατάσης διαφορετικών μεγεθών που ελευθερώνονται κατά τον αποπολυμερισμό του μορίου DNA (70).

Πίνακας 5. Σύγκριση Μεταξύ Κατασκευαστών Αλληλουχοποιητών (9, 81, 82).

Κατασκευαστής	Πλατφόρμα	Απόδοση	Μέγιστο read	Πλεονεκτήματα	Περιορισμοί
Illumina	HiSeq 2500	1 TB	250bp (SE), 250bp x2 (PE)	Υψηλής απόδοσης. Γνωστές μεθοδολογίες.Ακρίβειας πάνω από 99,9%. Η πιο δημοφιλής.	Μικρότερο μήκος μικροαναγνώσεων.
	MiSeq	15GB	300bp (SE)		
	NextSeq 550	120GB	150bp (SE), 150bp x2 (PE)		
	NovaSeq 6000	3000GB	250bp (SE), 250bp x2 (PE)		
	NextSeq 1000/2000	330GB	150bp (SE), 150bp x2 (PE)		
Thermo Fisher (IonTorrent)	PGM	2GB	400bp	Λιγότερο ακριβή τεχνολογία. Πιο γρήγορο τρέξιμο και καλύτερη κάλυψη.	Σφάλματα που σχετίζονται με ομοιοπολυμερή. Μικρό μήκος μικροαναγνώσεων.
	Proton	10GB	200bp		
	S5	15GB	200bp		
Pacific Biosciences	RS	1GB	40kbp	Μεγάλου μήκους μικροαναγνώσεις. Χωρίς σφάλματα λόγω PCR.Απευθείας ανίχνευση μεθυλιώματος. Ακρίβεια με την CLR 87-92%, ενώ με την CCS πάνω από 99%. Η τιμή με τις παλιότερες πλατφόρμες ήταν ακριβή, αλλά το κόστος έπεσε δραματικά με τις νέες πλατφόρμες (π.χ. SequellII έχει κόστος 13-26 $/GB).	Χαμηλότερης απόδοσης.
	RSII	1GB	>60 kbp		
	Sequel	20GB	>100 kbp		
	SequelI	7,6GB	>200 kbp CLR		
			>20 kbp CCS		
	SequelIIe	200GB	175kbp CLR		
		30GB	25kbp CCS		
Oxford Nanopore Technologies	MiniION/ GridION	30GB	>1000kbp Long	Χωρίς σφάλματα λόγω PCR. Μεγάλου μήκους μικροαναγνώσεις. Φτηνό.Μικρό μέγεθος πλατφόρμας MiniION.Απευθείας ανίχνευση μεθυλιώματος. Ακρίβεια 87-98%	Η GridION είναι x5 φορές πιο αποδοτική αλλά και πολύ πιο ακριβή από την MiniION. Απαιτητική υπολογιστική επεξεργασία των αποτελεσμάτων.
		2,5GB	>1500kbp Ultra-long		
	PromethION	180GB	>1000kbp Long		

Η κάλυψη της αλληλούχισης (depth of coverage) που αναφέρεται στον παραπάνω πίνακα, μας δίνει μια ένδειξη για το κόστος και την απόδοσης του πειράματος, και υπολογίζεται από την **Εξίσωση 2**, όπου η τιμή *O* είναι η μη συναρμολογημένη απόδοση της αλληλουχοποίησης, δηλαδή ο αριθμός των μικροαναγνώσεων (*R*) επί το μέσο μήκος κάθε μικροανάγνωσης (*L*), και η τιμή *I* είναι το εκτιμώμενο μήκος του εξεταζόμενου DNA ή το σύνολο του μήκους των contigs.

$$DepthofCoverage = \frac{O}{I} = \frac{R * L}{I}$$

Εξίσωση 2. Βάθος κάλυψης (9).

Με αυτήν την εξίσωση και την τιμή *I* μπορούμε να εξετάσουμε πώς θα καταφέρουμε το επιθυμητό βάθος κάλυψης, που μπορεί να είναι και 1000× βάθος για σωματικές μεταλλάξεις. Είναι σύνηθες, στις αναλύσεις που έχω περιορισμένο ποσοστό βιολογικής πληροφορίας π.χ. στην ανάλυση miRNA, να συμπεριληφθούν δείγματα από περισσότερες πηγές ή πειράματα, τα οποία αναμειγνύονται και διαβάζονται ταυτόχρονα (multiplexing) σε έναν κύκλο αλληλουχοποίησης. Αυτή η πολυπλεξία επιτυγχάνεται κατά την προετοιμασία της βιβλιοθήκης με την πρόσθεση μοναδικών σύνθετων μικρών αλληλουχιών, ή αλλιώς barcodes, σε κάθε δείγμα. Αυτό ονομάζεται διαδικασία ευρετηρίου (indexing) και αποτελεί ένα σημαντικό εργαλείο του βιοπληροφορικού για την αναγνώριση, δηλαδή την αποπολυπλεξία (demultiplexing), και την ομαδοποίηση των δειγμάτων με υπολογιστικές μεθόδους κατά την μετανάλυση (9). Εκτός από τα 6 με 8 από τα νουκλεοτίδια που μας επιτρέπουν να διαχωρίσουμε τα δείγματα και να τα ομαδοποιήσουμε κατάλληλα, το barcode χρησιμοποιείται στην βιοπληροφορική ανάλυση, δηλαδή για την ποσοτικοποίηση, για τα επίπεδα της ενζυμικής ενίσχυσης και για την πρόσδεση της βιβλιοθήκης στην επιφάνεια του flow cell, όπου θα πραγματοποιηθεί ο προσδιορισμός της τοποθέτησης των βάσεων.

Οι τεχνολογίες NGS έχουν το χαρακτηριστικό ότι πριν μπορέσει να αρχίσει η αλληλούχιση είναι απαραίτητη η κατασκευή βιβλιοθηκών. Ο σκοπός αυτού του βήματος είναι η ενσωμάτωση συνθετικών ολιγονουκλεοτιδικών προσαρμογέων στα άκρα των μορίων DNA/RNA. Για τις πλατφόρμες Illumina και Ion Torrent, αυτοί οι προσαρμογείς χρησιμοποιούνται για την κλωνική επαύξηση των ενθέσεων στην βιβλιοθήκη, έτσι ώστε να ενισχύσουν το ανιχνευόμενο σήμα κατά την αλληλούχιση. Γι' αυτό χρησιμοποιούνται βελτιωμένες τεχνικές PCR, οι οποίες είναι ακόμα πιο σημαντικές στην αλληλούχιση του γενετικού υλικού με NGS τεχνικές, που θα δούμε στην επόμενη ενότητα, ένα τέτοιο

παράδειγμα είναι η γέφυρα-PCR (bridge-PCR) που οδηγεί σε σύνθεση ομαδοποιημένων αντιγράφων. Βέβαια δεν χρησιμοποιείται παντού η PCR, ένα τέτοιο παράδειγμα είναι η πλατφόρμα NGS της εταιρίας Oxford Nanopore Technologies (ONT). Ακόμα, η πλατφόρμα PacBio έχει αρκετά υψηλή ευαισθησία για να μην χρειάζεται PCR. Οι προσαρμογείς είναι πολύ σημαντικοί για την εκκίνηση της αλληλουχοποίησης σε όλες τις πλατφόρμες.

Illumina SBS. Η τεχνολογία της βασίζεται στη μέθοδο μαζικής παράλληλης αλληλούχισης μέσω σύνθεσης (massive parallel sequencing by synthesis ή SBS), που χρησιμοποιεί την χημεία για αντιστρεπτό τερματισμό της αντιγραφής (Cyclic Reversible Termination, CRT) για να ταυτοποιήσει την αλληλουχία του μορίου DNA. Ξεχωρίζει λόγω της πληθώρας των παραγόμενων δεδομένων της παγκοσμίως αλλά και λόγω των πωλήσεων της εταιρίας σε μεγάλα κέντρα γενωμικής και ιδιωτικά εργαστήρια. Η αντιδράσεις αλληλούχισης στα μηχανήματα της Illumina πραγματοποιούνται σε μια μοναδικής χρήσης επιφάνεια πυριτίου (flow cell), στο οποίο συνήθως διαχωρίζουμε το κάθε δείγμα στην 1 από τις 8 διαδρομές (lanes). Αυτό εξαρτάται από το πείραμά μας αφού μπορούμε παράλληλα να κάνουμε πολλά πειράματα με ένα flow cell, ακόμα και στην ίδια διαδρομή, μέσω της διαδικασίας ευρετηρίου (indexing). Ανάλογα με την πλατφόρμα μπορεί να έχω πάνω από ένα index κάθε δείγμα της πλατφόρμας Illumina μπορεί να αποτελείται από το index1 και το index2. Προσδεδεμένα στην μοναδικής χρήσης επιφάνεια πυριτίου είναι ειδικά ολιγονουκλεοτίδια που είναι συμπληρωματικά στους προσαρμογείς που χρησιμοποιούνται για την προετοιμασία των βιβλιοθηκών και χρησιμοποιούνται για την καταγραφή των μονόκλωνων θραυσμάτων DNA. Χρησιμοποιώντας αυτές της συμπληρωματικές αλληλουχίες προσαρμογέα, αυτά τα θραύσματα είναι συστάδες κλώνων DNA από επαύξυση μέσω της bridge-PCR. Μέσα από αυτή την διεργασία, το επεκταμένο DNA δημιουργεί μια γέφυρα με τον συμπληρωματικό ακίνητο προσαρμογέα ο οποίος τώρα λειτουργεί ως εκκινητής για τον επόμενο κύκλο της σύνθεσης. Στο τελευταίο βήμα της διεργασίας κλωνικής επαύξησης, ο ένας κλώνος του DNA απομακρύνεται από το κάθε καλούπι μέσω της περιοριστικής περιοχής στον προσδεδεμένο στην επιφάνεια πυριτίου προσαρμογέα. Το τελικό αποτέλεσμα είναι μια συλλογή μονόκλωνων πολυμερισμένων αποικιών (polymerase+colony=polonies), δηλαδή συστάδες ακινητοποιημένων επαυξημένων μορίων DNA, που χρησιμοποιούνται σαν εκμαγείο για τη σύνθεση συμπληρωματικών αλληλουχιών, που επεκτείνονται μέσω ενζυμικής προσθήκης μονονουκλεοτιδίων, τα οποία φέρουν διαφορετικές φθορίζουσες ουσίες.

Σύμφωνα με αυτές, οι πλατφόρμες της Illumina χωρίζονται σε τρεις κατηγορίες, δηλαδή τις τετρακάναλες (τέσσερις χρωστικές), τις δικάναλες (δύο χρωστικές) και τις μονοκάναλες (μία χρωστική), που με την σειρά που αναφέρθηκαν είναι από την πιο παλιά στην πιο καινούργια τεχνολογία. Οι τετρακάναλες, π.χ. η πλατφόρμα MiSeq, είναι πολύ χρονοβόρες αφού πρέπει να οπτικοποιηθούν τέσσερις εικόνες της αλληλουχίας, μια για κάθε χρώμα. Οι δικάναλες, π.χ. οι πλατφόρμες MiniSeq, NextSeq 550, NovaSeq 6000, φέρουν δύο διαφορετικές φθορίζουσες ουσίες και αναλύονται στον μισό χρόνο. Άρα έχουμε τέσσερις βάσεις και δύο χρωστικές, οπότε οι αντιστοιχίες γίνονται π.χ. πράσινο στις θυμίνες, κόκκινο στις κυτοσίνες, κίτρινο (το μείγμα του κόκκινου-πράσινου) στις αδενίνες και η απουσία φθορισμού στις γουανίνες. Σε αυτές τις δύο τεχνικές, η λήψη του εκπεμπόμενου σήματος από το flow cell στην κάμερα γίνεται σε πραγματικό χρόνο. Έχουμε τέσσερις ή δύο, αντίστοιχα, πρωτογενής εικόνες, οι οποίες αναλύονται ως προς την ένταση του σήματος και αμέσως μετά την κανονικοποίηση σβήνονται λόγω του μεγάλου μεγέθους τους. Η καινούργια τεχνική είναι η σύμπτυξη της τεχνολογίας SBS με ολοκληρωμένα ψηφιακά κυκλώματα (CMOS) και νανοπηγαδάκια στην πλατφόρμα iSeq 100, η οποία γίνεται με μία χρωστική. Η διαφορά είναι ότι λαμβάνουμε τις δύο εικόνες και την ένταση του σήματος σε διαφορετικό χρόνο, δηλαδή μετά από διαφορετικό σημείο χημικής ανάλυσης. Ακόμα, μπορεί ο ερευνητής να επιλέξει αν η ανάλυση θα είναι προς μια κατεύθυνση (Single End, SE) ή μόλις τελειώσει αυτή θα γίνει ανάλυση και προς την αντίθετη κατεύθυνση (Paired End, PE). Όταν το θραύσμα είναι μεγαλύτερο από την μικροανάγνωση, η PE μπορεί να μας δώσει παραπάνω βιολογικές πληροφορίες, δηλαδή ξέρουμε ότι αυτό προήλθε από ένα μόνο μόριο. Ακόμα, με χρήση αλγορίθμων γίνεται η ευθυγράμμιση των εμπρόσθιων και των ανάστροφων αλληλουχιών. Όταν κάνεις PE λαμβάνεις δύο αρχεία (fastq, που είναι πολύ μεγάλα και συμπιέζονται σε fastq.gz) για κάθε δείγμα, αυτό προς την μια και αυτό προς την άλλη κατεύθυνση.

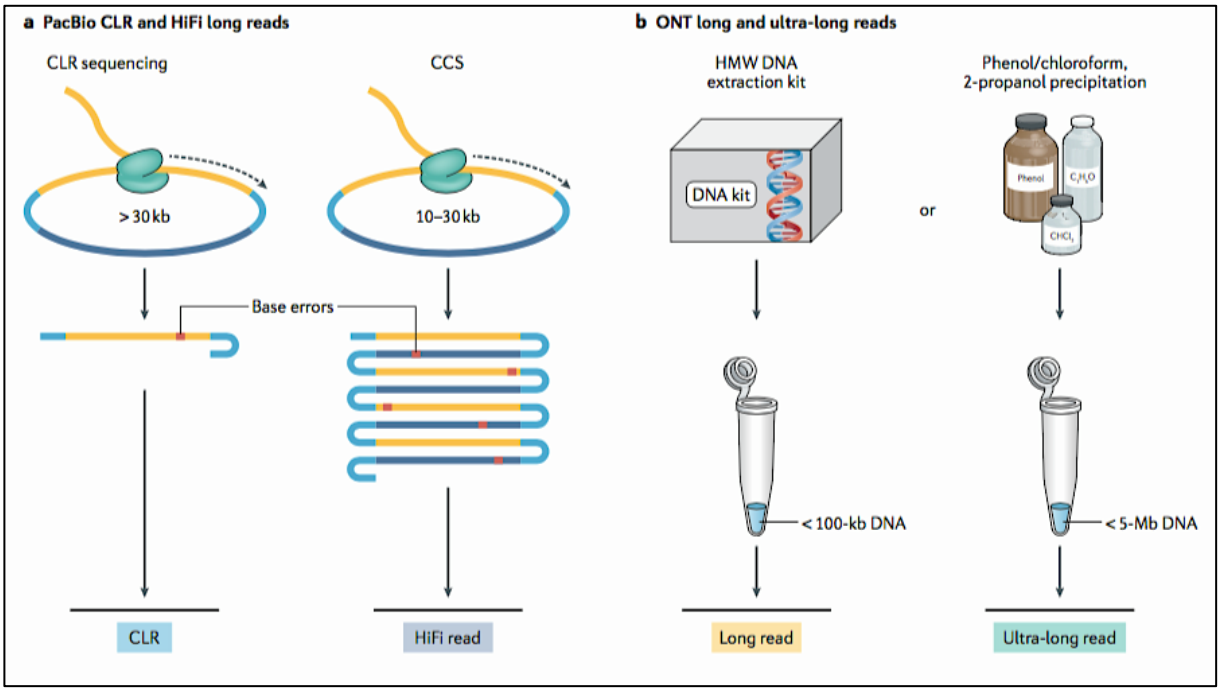
Ion Torrent Semiconductor Sequencing. Η τεχνολογία της πλατφόρμας IonTorrent από την τεχνολογία Life Technologies έχει πολλές ομοιότητες με την πλέον ξεπερασμένη πλατφόρμα 454 pyrosequencing, επειδή χρησιμοποιεί emPCR (emulsion-PCR, emPCR) για επαύξηση του δείγματος και επειδή σε κάθε κύκλο αλληλούχισης αναγνωρίζεται μόνο ένας τύπος νουκλεοτιδίου. Στην τεχνική emPCR, κλωνοποιούνται οι αλληλουχίες DNA στην επιφάνεια του σφαιριδίου μέσα σε μικρές υδάτινες φούσκες, που τοποθετούνται σε έλαιο-διάλυμα. Η τεχνολογία βασίζεται στην χρήση ημιαγωγών, δηλαδή ολοκληρωμένων ψηφιακών συστημάτων (complementary metal-oxide semiconductor, CMOS). Είναι η ίδια τεχνολογία με τα τσίπ των υπολογιστών. Η τεχνολογία αλληλουχοποίησης IonTorrent είναι η πρώτη

εμπορική πλατφόρμα, η οποία δεν χρησιμοποιεί φως ως το σήμα εξόδου, αλλά ανιχνεύει τα ιόντα του πρωτονίου που απελευθερώνονται με την πρόσδεση κάθε νουκλεοτιδίου και μπορεί να αποδίδει μικροαναγνώσεις μεγέθους 400 βάσεων με το ιοντικό τσίπ, που αναγνωρίζει τις αλλαγές στο pH όταν απελευθερώνονται ιόντα υδρονίου (hydronium, H_3O^+) ως παραπροϊόντα από την πρόσδεση νουκλεοτιδίων στην αλληλουχία του εκμαγείου. Η ένταση του σήματος είναι άμεσα εξαρτώμενη από των αριθμό των bp που ενσωματώνονται σε κάθε κύκλο αληλούχισης. Η αντίδραση γίνεται μέσα στα μικροπηγαδάκια του Ion τσίπ, το οποίο είναι ένα ειδικό ημιαγώγιμο τσίπ πυριτίου σχεδιασμένο συγκεκριμένα για να ανιχνεύσει τις αλλαγές στο pH χρησιμοποιώντας αισθητήρες ιόντων υδρογόνου στη βάση των μικροπηγαδιών. Με τον δίαυλο ιόντων απομακρύνεται η ανάγκη για φωτεινή πηγή, σάρωση και απόκτηση της εικόνας, που είναι συνήθως απαραίτητα για την ανίχνευση σημάτων στην αλληλουχοποίηση, οπότε είναι πολύ πιο γρήγορη η διαδικασία της αλληλούχισης. Στις αντιδράσεις αλληλούχισης της τεχνικής Ion Torrent, μη ραδιοσημασμένα εγγενή νουκλεοτίδια χρησιμοποιούνται για τον πολυμερισμό. Οπότε δεν δημιουργούνται σφάλματα που θα προκαλούνταν από τον φθορισμό και από τις ουσίες που αναστείλουν το αντιδρόν. Το μειονέκτημα της είναι ότι σε περιπτώσεις που η ίδια βάση επαναλαμβάνεται πολλές φορές, δηλαδή τα ομοιοπολυμερή, το σήμα εξόδου είναι υψηλό, κάτι που δυσκολεύει την ακριβή εκτίμηση του μεγέθους των επαναλήψεων μεταξύ ίσου μεγέθους ομοπολυμερή. Για παράδειγμα, ένα ομοιοπολυμερές που αποτελείται από εννιά κυτοσίνες μπορεί να έχει σήμα ίδιας έντασης με ένα ομοπολυμερές που αποτελείται από δέκα κυτοσίνες, άρα υπάρχει αυξημένο ποσοστό εσφαλμένων ενθέσεων και διαγραφών (indels) στα ομοπολυμερή. Με την IonTorrent τεχνική έχουν κατασκευαστεί διαφορετικές πλατφόρμες με άλλες δυνατότητες απόδοσης η καθεμία, π.χ. η προσωπική γενωμική μηχανή (Personal Genome Machine, PGM), η Proton κ.α. The PGM μπορούσε να αποδώσει μέχρι 2 Gb αλληλούχισης σε κάθε γύρο, ενώ το Proton μέχρι 10 GB. Μετά βελτιώθηκε η πλατφόρμα, Proton II, με απόδοση γύρω στο 32 GB και αναπτύχθηκε μια εναλλακτική μέθοδο στην emPCR, η οποία βασίζεται στην ισοθερμική επαύξηση, χωρίς όμως να αυξηθεί το κόστος της ανάλυσης. Όταν ξεπεράστηκαν οι παραπάνω πλατφόρμες, η Thermo Fisher τις αντικατέστησε με την πλατφόρμα S5 για καλύτερη απόδοση (15GB) και φτηνότερη ανάλυση (83). Μια νέα πλατφόρμα της εταιρίας, η Ion Torrent Genexus, υπόσχεται ένα ολοκληρωμένο αυτοματοποιημένο σύστημα, όπου οι χρήστες καλούνται να κάνουν ελάχιστη αλληλεπίδραση με την πλατφόρμα, μειώνοντας την ανάγκη για εξειδικευμένες βιοπληροφορικές γνώσεις και αυξάνοντας την επαναληψιμότητα των αποτελεσμάτων.

Pacific Biosciences SMRT (**Εικόνα 13**). Η εταιρία Pacific Biosciences (PacBio) με την μονομοριακή αλληλούχιση σε πραγματικό χρόνο (single molecule real-time, SMRT) κυκλοφόρησε πιο μετά στο εμπόριο της αλληλούχισης εξαιρετικά υψηλής απόδοσης. Η μέθοδος PacBio SMRT βασίζεται στα «πηγαδάκια» zero-mode κυματοδηγό (Zero-Mode Waveguide, ZMW), όπου το καθένα περιέχει ένα ένζυμο πολυμεράσης, το εκμαγείο DNA, τον εκκινητή της αλληλούχισης και σημασμένα με φθορισμό νουκλεοτίδια. Το σήμα φθορισμού που σχετίζεται με κάθε ενσωμάτωση νουκλεοτιδίου στον αυξανόμενο κλώνο DNA καταγράφεται σε πραγματικό χρόνο (84). Η τεχνική PacBIO SMRT έχει τα τρία παρακάτω πλεονεκτήματα. Αρχικά, η μεμονωμένη μοριακή (single molecule) ανίχνευση δεν βασίζεται στην PCR επαύξηση, το οποίο σημαίνει ότι τα δεδομένα δεν πλήττονται από οι στατιστικές προκαταλήψεις. Ένα άλλο πλεονέκτημα της μεθόδου είναι ότι παράγει αρκετά μεγαλύτερες μικροαναγνώσεις από τις πλατφόρμες Illumina και IonTorrent, το μήκος μικροαναγνώσεων φτάνει και περισσότερες από 60 kbp με την πλατφόρμα RSII (η οποία είναι παλιότερης τεχνολογίας και θα σταματήσει να υποστηρίζεται από την εταιρία PacBio μέσα στο 2021, θα υπάρχει ακόμα η δυνατότητα αλληλούχισης σε αυτά όσα χρόνια λειτουργούν τα ήδη υπάρχοντα μηχανήματα, που βελτιώνει την συναρμολόγηση γενωμικών περιοχών με μεγάλο ποσοστό επαναληπτικών αλληλουχιών. Ακόμα, η διάκριση μεταξύ ισόμορφων mRNA δεν είναι πρόβλημα λόγο του μεγαλύτερου μήκους μικροαναγνώσεων. Ένα ακόμα πλεονέκτημα, αφού η επέκταση της αλυσίδας παρατηρείται σε πραγματικό χρόνο, είναι η παρατήρηση των κινήσεων της πολυμεράσης. Το ποσοστό με το οποίο η πολυμεράση ενσωματώνει τα νουκλεοτίδια διαφοροποιείται μεταξύ τροποποιημένων και μη τροποποιημένων νουκλεοτιδίων. Συνεπώς, η διαφοροποίηση της ενσωμάτωσης νουκλεοτιδίων κατά την διεργασία σύνθεσης του DNA κάνει συνετή την άμεση μελέτη της μεθυλίωσης του DNA (85). Αυτές οι προχωρημένες τεχνολογίες καθιστούν δυνατή την ανακάλυψη ή ταυτοποίηση των miRNA, που είναι ρυθμιστές γονιδιακής έκφρασης, και των λειτουργιών τους, οι οποίες θα ήταν δύσκολο να αναγνωριστούν από κλασικές τεχνικές (9).

Oxford Nanopore Technologies (ONT) (**Εικόνα 13**). Το πρόβλημα για την εφαρμογή της αλληλουχοποίησης στην κλινική διάγνωση είναι ότι ο χρόνος της ιατρικής γνωμάτευσης επιμηκύνεται. Ο χρόνος αυτός μειώνεται με τις νέες τεχνολογίες. Το 2014 η εταιρία ONT ανέπτυξε μια πλατφόρμα, την MiniION, με την τεχνική νανοπόρων. Η αλληλουχοποίηση αυτή βασίζεται στις ηλεκτρικές διακυμάνσεις από το πέρασμα μονόκλωνων μορίων DNA (ssDNA) μέσα από τους νανοπόρους. Η αρχή λειτουργίας της είναι σχετικά απλή. Η διέλευση ενός μόριου από έναν πόρο γίνεται είτε μέσω ηλεκτρικού πεδίου είτε μέσω ενζύμου και η

καταγραφή γίνεται σε πραγματικό χρόνο. Με το πέρασμά τους διαταράσσουν τα ιόντα, που επίσης περνάνε από αυτούς τους ναονοπόρους, και μετρώνται οι αλλαγές στο ιονικό ρεύμα. Το σήμα που παράγεται έχει ιδιαίτερα χαρακτηριστικά, ενώ συγκεκριμένα χαρακτηριστικά υποδηλώνουν και άλλη βάση. Με αυτά τα δεδομένα γίνεται η αναγνώριση παρουσίας μιας συγκεκριμένης βάσης σε συγκεκριμένη θέση στην νεοσυντιθέμενη αλληλουχία (base calling), η οποία και καταγράφεται. Η τεχνική αλληλούχισης με νανοπόρους είχε δημοσιευθεί πολύ νωρίτερα βέβαια, στα μέσα της δεκαετίας του ’90. Η πλατφόρμα MiniION της ONT, που αποδίδει μέχρι 30Gb για κάθε flow cell, με πολύ υψηλή ακρίβεια. Τα δεδομένα είναι άμεσα προσβάσιμα σε πραγματικό χρόνο. Η GridION είναι ένα μηχάνημα που είναι σαν να περιέχει 5 MinION, αλλά είναι πολύ πιο ακριβό. Οι πλατφόρμες της ONT δίνουν κατά την ανίχνευση λιγότερα ψευδώς θετικά και ψευδώς αρνητικά, από πολλές πλατφόρμες άλλων εταιριών, ενώ είναι συμβατά με εφαρμογές κινητού. Η πλατφόρμα PromethION, που έχει την ίδια τεχνολογία με το MinION, είναι υψηλής απόδοσης και υψηλής πιστότητας αλληλούχιση νουκλεϊνικών οξέων. Με αυτήν μπορούν να γίνουν ταυτόχρονα πολλά πειράματα σε πραγματικό χρόνο, με μεγάλου μήκους μικροαναγνώσεις και χωρίς περιορισμό στο πόσο θα διαρκέσουν. Από την αρχή δεν χρειάστηκε επαύξηση PCR, χημική σήμανση και σάρωση της εικόνας, όπως στις τεχνικές άλλων εταιριών που περιγράφονται παραπάνω. Για αυτούς τους λόγους είναι τεχνική υψηλής απόδοσης και χαμηλού κόστους. Αυτή την τεχνολογία δίνει την δυνατότητα αλληλούχισης ακόμα και δίκλωνων μορίων DNA. Ένα μεγάλο πλεονέκτημα της πλατφόρμας MiniION είναι ότι μπορεί να χρησιμοποιηθεί σε περιοχές που δεν έχουν εργαστήρια τελευταίας τεχνολογίας, αφού μπορείς να το πάρεις μαζί σου όπου χρειάζεται. Για παράδειγμα, οι ερευνητές Joshua Quick και Nicholas Loman το 2015 ταξίδεψαν στην Γουινέα με την τότε νεοεισερχόμενη πλατφόρμα, για να αλληλουχίσουν τους παθογόνους μικροοργανισμούς του ιού Ebola, που εκείνη την περίοδο ήταν σε έξαρση στην χώρα. Η πλατφόρμα MiniION χρειάστηκε δύο μόνο μέρες από την απομόνωση του δείγματος για να αλληλουχιθεί ο ιός. Έτσι δίνεται η δυνατότητα να έχουν πρόσβαση στην NGS πολύ περισσότεροι επιστήμονες και η αλληλουχοποίηση ακόμα και κάτω από μη ιδανικές συνθήκες με καλύτερα αποτελέσματα και με μικρό σχετικά κόστος. Τα μειονεκτήματα αυτής της τεχνικής είναι ότι μπορεί να έχει φράξει ο πόρος ή τα ένζυμα να μην αλληλεπιδράσουν σωστά με τα νουκλεϊνικά οξέα ή να μην αναγνωριστούν σωστά οι βάσεις (83, 86).



Εικόνα 13. Πλατφόρμες 3ης γενιάς από τις εταιρίες (**a**) PacBio και (**b**) ONT. Στις μεγάλου μήκους μικροαναγνώσεις Hi-Fi η ακρίβεια είναι 99%. Στις μικροαναγνώσεις Ultra-long δεν έχουμε καλύτερη ακρίβεια από τις Long, απλώς έχουμε μεγάλου μήκους, πάνω από 1500kbp, μικροαναγνώσεις (82).

2.2.2. Διάκριση Μεταξύ Αναλύσεων NGS και Πειραματική Διαδικασία

Δύο πολύ δημοφιλείς μέθοδοι αλληλουχοποίησης του γενώματος είναι η Whole Genome Sequencing (WGS), που είναι η ανάλυση ολόκληρου του γενώματος ενός οργανισμού, και η Whole Exome Sequencing (WES), που αφορά στην ανάλυση μόνο ενός μικρού ποσοστού του γενώματος και συγκεκριμένα των τμημάτων που κωδικοποιούν για πρωτεΐνες. Η δεύτερη είναι μια μεταγραφομική ανάλυση, αφού αφορά στην αλληλούχιση ολόκληρου του μεταγραφόματος (mRNA) ενός οργανισμού. Η WES δεν αναλύει μόνο την περιοχή που κωδικοποιεί για πρωτεΐνες, αλλά και τα μη κωδικοποιά μόρια, όπως τα mi-RNAs. Τα πλεονεκτήματά της είναι ότι έχει καλύτερη κάλυψη και είναι φτηνότερη από την WGS. Όμως δεν μπορεί να ανιχνεύσει μεταλλάξεις και άλλες τροποποιήσεις στα εσόνια, δηλαδή το 98% του γενώματος. Η WES για να μπορεί να διαφοροποιήσει πολυμορφισμούς με την ανάλυση PE, στην πλατφόρμα της Illumina. Στην κλινική διάγνωση η WES είναι πιο συνήθης από την WGS, αφού οι γενετικές τροποποιήσεις αναλύονται πιο εύκολα στα εξόματα, αντί στις ενδογονιακές (intragenic) περιοχές και διαγονιδιακές (intergenic) περιοχές, ανάμεσα σε κωδικοποιά γονίδια, κυρίως όταν προκαλούν αλλαγές απευθείας στην πρωτεϊνική αλληλουχία. Με την αλληλουχοποίηση ολόκληρου του γενώματος (WGS) είναι δυνατός ο χαρακτηρισμός γενωμικών ελαττωμάτων, όπως τα CNVs, οι ενθέσεις-διαγραφές και οι περισσότερες μεταλλάξεις του γενώματος. Είναι μια ολοκληρωμένη ανάλυση της βιολογικής

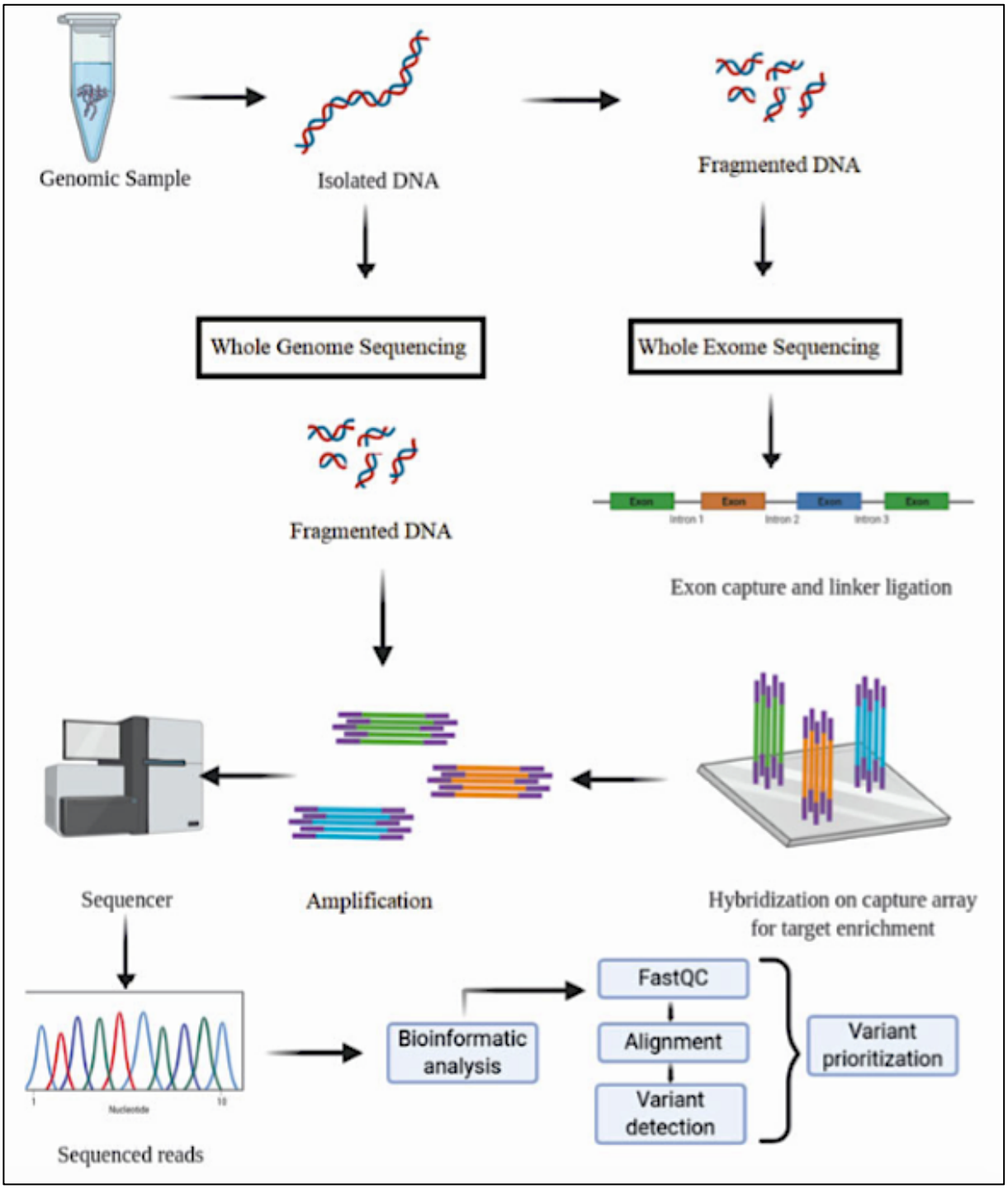
πληροφορίας, αλλά γι'αυτόν το λόγο χρειάζεται πολύπλοκη μετανάλυση, έχει αυξημένο κόστος και η αλληλουχοποίηση απαιτεί περισσότερο χρόνο. Η διαλεύκανση των μη κωδικοποιών περιοχών με τις τεχνικές NGS, όπως φαίνονται στην παραπάνω ενότητα, γίνονται όλο και πιο αξιόπιστες, ενώ τα βιοπληροφορικά εργαλεία και η εξοικείωση με αυτά συνεχώς βελτιώνεται.

2.2.2.1. DNA-Seq, Whole Genome Sequencing (WGS) και Whole Exome Sequencing (WES)

Πειραματική Διαδικασία αλληλούχισης DNA/RNA με την τεχνική SBS της Illumina (**Εικόνα 14**). Το πρώτο βήμα για την αλληλουχοποίηση είναι η προετοιμασία των βιβλιοθηκών νουκλεϊνικών οξέων και στην συνέχεια αυτά αλληλουχοποιούνται, με:

1. Απομόνωση των νουκλεϊνικών οξέων, DNA ή RNA, ανάλογα με την εφαρμογή που μας ενδιαφέρει.
2. Τον κατεκερματισμό των νουκλεϊνικών οξέων σε μικρά θραύσματα.
3. Την ενσωμάτωση των συνθετικών προσαρμογέων σε κάθε θραύσμα και την σήμανση αυτών με barcode, που είναι χρήσιμα σε περίπτωση πολυπλεξίας, δηλαδή αν η βιβλιοθήκη απαρτίζεται πάνω από ένα δείγμα.
4. Το υλικό υπό εξέταση προσδένεται στο flow cell, με την βοήθεια προσαρμογέων και αξιοποιώντας την αρχή της συμπληρωματικότητας, το οποίο τοποθετείται στον αλληλουχοποιητή.
5. Στην τεχνική SBS πραγματοποιείται κλωνική επαύξηση «γέφυρας», που οδηγεί σε σύνθεση συστάδων.
6. Στις συστάδες προσδένονται εκκινητές, μέσω των συνθετικών προσαρμογέων. Με την πολυμεράση λαμβάνουν την θέση τους τα νουκλεοτίδια, που περνάνε από ενζυμική ενσωμάτωση διαφορετικών φθοριζόντων μόριων (1,2, ή 4). Σε κάθε κύκλο υβριδισμού διεγείρεται η φθορίζουσα ουσία, όπου η καθεμία εκπέμπει φωτόνια σε διακριτό μήκος κύματος, και η βάση αναγνωρίζεται από ένα σύστημα ακτινοβολίας του αλληλουχοποιητή. Δηλαδή η εκπεμπόμενη ακτινοβολία ανιχνεύεται με μια κάμερα υψηλής ευκρίνειας (Charge-coupled device, CCD) και η κάθε βάση λαμβάνει την θέση της στην αλληλουχία. Στην συνέχεια, απομακρύνεται η φθοριούχα ουσία και αρχίζει νέος κύκλος υβριδισμού με άλλη φθορίζουσα ουσία (εκτός από την πλατφόρμα iSeq). Έτσι, γίνεται η ταυτοποίηση των βάσεων, η στοίχιση και η χαρτογράφηση όλων των συστάδων παράλληλα, ενώ για μεγαλύτερη αξιοπιστία προτιμάμε την ανάλυση PE, που μπορεί να χρησιμοποιηθεί στην

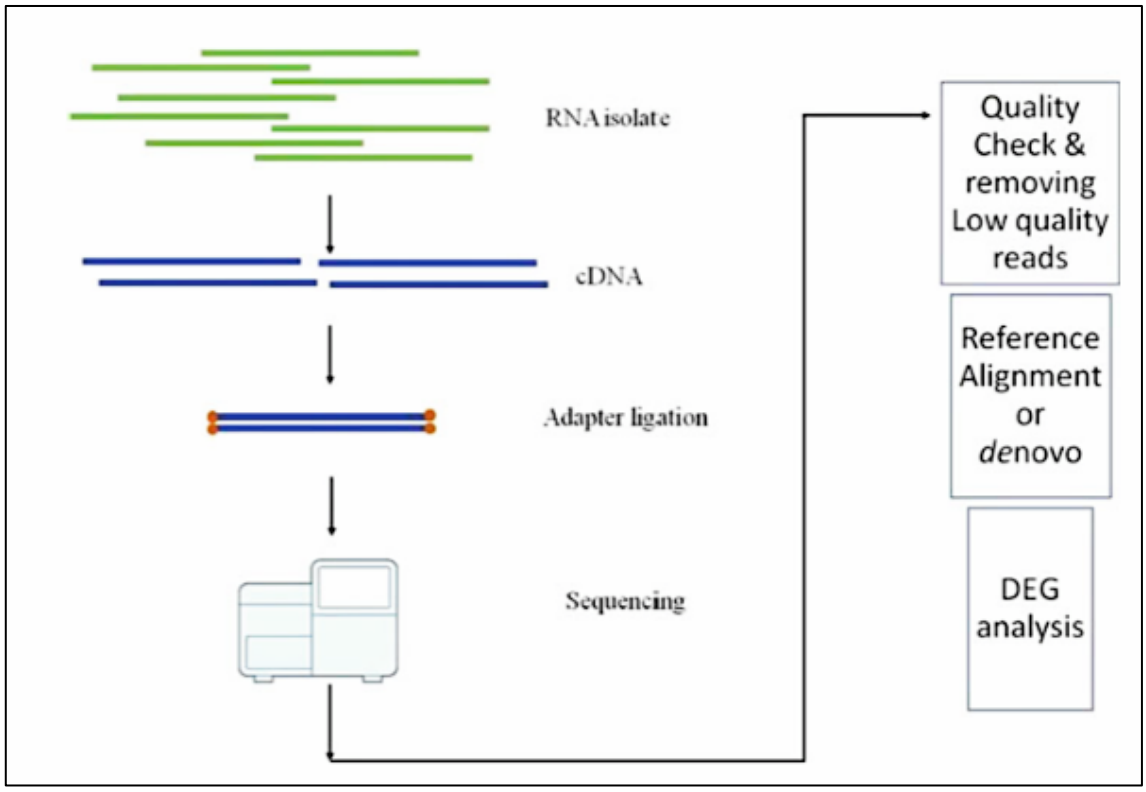
μετανάλυση για να βελτιώσει την ποιότητα των μικροαναγνώσεων. Η διάρκεια εκτέλεσης της αντίδρασης εξαρτάται από τον αριθμό των κύκλων που πραγματοποιούνται (88).



Εικόνα 14. Η ροή εργασίας για WGS και WES (87).

2.2.2.2. RNA-Seq

Είναι η αλληλούχιση και η ανάλυση όλων των μορίων RNA (rRNA, mRNA, miRNA κ.λ.π.) (**Εικόνα 15**), αφού περιέχει τις κωδικοποιούσες και μη κωδικοποιούσες περιοχές. Η αλληλούχιση αυτή επιτρέπει την διερεύνηση της γονιδιακής έκφρασης και μεταμεταγραφικών αλλαγών, όπως τα SNVs και η εναλλακτική συρραφή και των επιπέδων της βιολογικής πληροφορίας, που σχετίζεται με το βάθος κάλυψης της αλληλούχισης. Η RNA-seq είναι η πλέον κατάλληλη για την εκ νέου αλληλούχιση και έχει δυναμική μορφή (ολόκληρες αλληλουχίες αντί για spots), που συνήθως δεν γίνεται με τις μικροσυστοιχίες. Η βιοπληροφορική ανάλυση των δεδομένων RNA-seq με τα παλιά εργαλεία είναι περισσότερη περίπλοκη, αφού ήταν σχεδιασμένα για ανάλυση δεδομένων DNA.



Εικόνα 15. Η ροή εργασίας RNA-Seq (87).

Πειραματική διάταξη αλληλούχισης του γενώματος ενός ευκαρυωτικού οργανισμού.

1. Στην RNA-Seq, από το συνολικό RNA (total RNA) μέσω της αντίστοφης μεταγραφής συνθέτουμε το cDNA μας δίνεται η δυνατότητα να αλληλουχίσουμε νέα μετάγραφα ή και σπάνιες παραλλαγές.
2. Εμπλουτίζουμε την αλληλουχία με την επιλογή PolyA, που είναι τμήματα που περιέχουν επαναλαμβανόμενες βάσεις αδενίνης. Η πολυμεράση PolyA προσθέτει μονοφωσφορική αδενοσίνη (AMP), που προέρχεται από την τριφωσφορική αδενοσίνη (ATP), στο 3'-άκρο του mRNA.

3. Απαλείφουμε το ριβοσωμικό RNA (rRNA), αφού η ανάλυση μας επικεντρώνεται κυρίως στο mRNA ή στο miRNA (και το siRNA επίσης).
4. Το mRNA ή το miRNA, αντίστοιχα με την εφαρμογή που μας ενδιαφέρει, μετατρέπεται σε cDNA, μέσω της αντίστροφης μεταγραφής (RT), και κατακερματίζεται σε θραύσματα μικρού μεγέθους, ανάλογα με την πλατφόρμα αλληλουχοποίησης που χρησιμοποιείται.
5. Σε κάθε ένα από τα θραύσματα προστίθενται συνθετικοί προσαρμογείς στα άκρα τους, οι οποίοι περιέχουν απαραίτητα λειτουργικά στοιχεία για την αλληλούχιση, όπως τα σημεία που θα αρχίσει η αλληλούχιση, καθώς και τα σημεία πρόσδεσης των θραυσμάτων στον δίαυλο του αλληλουχοποιητή.
6. Ακολουθούν διαδοχικοί κύκλοι ενίσχυσης PCR.
7. Γίνεται ο εντοπισμός και η ταυτοποίηση θέση προς θέση των νουκλεοτιδίων (resequencing ή de novo sequencing), ο τρόπος της αλληλούχισης εξαρτάται από την πλατφόρμα που χρησιμοποιείται.
8. Η σύνθεση της αλληλουχίας παράγει τις μικροαναγνώσεις, καθεμία έχει την σήμανσή της, το όνομά της, την αλληλουχία των νουκλεοτιδίων της, και στο καθένα nt έχει χαρακτήρες που υποδηλώνουν την ποιότητα της αλληλούχισης. Σε κάθε δείγμα μπορούν να προκύψουν εκατομμύρια μικροαναγνώσεις, ενώ γονίδια μεγαλύτερα σε μήκος συγκεντρώνουν μεγαλύτερη ποσότητα μικροαναγνώσεων. Αυτές στοιχίζονται η μία κάτω από την άλλη σε μορφή contigs, χαρτογραφούνται και καταγράφονται σε αρχεία fastQ (89). Μετά την αντιστοίχιση βάσεων, λογισμικό εκτελεί το φιλτράρισμα και τον έλεγχο ποιότητας.

Η ακρίβεια στην ποσοτικοποίηση του μεταγράφου, δηλαδή των επιπέδων γονιδιακής έκφρασης είναι η βάση για όλες τις μετααναλύσεις. Προτάθηκε σε δημοσίευση του 2008 μια μέθοδος με βάση τον αριθμό των μικροαναγνώσεων ανά kb του μεταγράφου σε ένα εκατομμύριο χαρτογραφημένες μικροαναγνώσεις (reads per kilobase of a transcript per million mapped reads, RPKM) (90). Η τιμή RPKM λέγεται και κανονικοποίηση δυο σταδίων. Η μέθοδος RPKM λαμβάνει υπόψη το μήκος του μεταγράφου και το σύνολο των μικροαναγνώσεων που προσφέρονται από αυτό, αλλά δεν είναι αμερόληπτη και δεν λαμβάνει υπόψη τις διακυμάνσεις στις μικροαναγνώσεις RNA-Seq. Από στατιστικής πλευράς, η μέθοδος RPKM και άλλες τέτοιες μέθοδοι μετρητών (έτσι ονομάζεται το πλήθος των μικροαναγνώσεων που στοιχίζονται στο κάθε γονίδιο) ανά εκατομμύριο μικροαναγνώσεων (Counts Per Million reads, CPM), δηλαδή 1.000.000 μετρητές/μέγεθος βιβλιοθήκης, δεν προσφέρουν μεγάλη ακρίβεια στην ποσοτικοποίηση των δεδομένων. Αυτό γιατί σε αυτές

θεωρείται ότι οι μικροαναγνώσεις ακολουθούν την κατανομή Poisson για σταθερό ρυθμό έντασης και χρησιμοποιείται αυτός ο ρυθμός για τον υπολογισμό της γονιδιακής έκφρασης. Εκτός από την RPKM, που χρησιμοποιείται για την SE ανάλυση, υπάρχει και η FPKM, όταν εφαρμόζουμε PE ανάλυση (91). Αλλά για την εύρεση διαφορικά εκφραζόμενων γονιδίων, σε λογισμικά όπως το DESeq2 και το edgeR, είναι προϋπόθεση να εισάγουμε τα πρωτογενή δεδομένα και το ίδιο το λογισμικό πραγματοποιεί κανονικοποίηση των τιμών με τους αλγορίθμους που επιλέγουμε από αυτούς που προσφέρει το καθένα ή και χωρίς κανονικοποίηση, για σύγκριση και μεγαλύτερη αξιοπιστία στα αποτελέσματα (92, 93). Αντιθέτως από την DNA-Seq, όταν χαρτογραφούμε μικροαναγνώσεις που προέρχονται από RNA-Seq σε ένα γένωμα αναφοράς πρέπει να λάβουμε υπόψη στις ενώσεις μικροαναγνώσεων εξονίου με εξόνιο (exon junction).

$$RPK(readsperkilobase) = \frac{\mathrm{Re}\,ads}{Length}$$

$$RPKM = \frac{RPK}{Output}$$

$$RPKM_i = \frac{raw.counts}{gene.length \cdot seq.depth} = \frac{X_i}{\frac{l_i}{10^3} \cdot \frac{N}{10^6}}$$

Εξίσωση 3. Δύο σταδίων κανονικοποίηση RPKM, όπου ο όρος πρωτογενής μετρητής (raw counts) X είναι ο αριθμός των μικροαναγνώσεων τα οποία επικαλύπτονται με την ένωση των εξονίων ενός γονιδίου.

$$FPKM_i = \frac{raw.counts}{gene.length \cdot seq.depth} = \frac{X_i}{\frac{l_i}{10^3} \cdot \frac{N}{10^6}}$$

Εξίσωση 4. Για Paired End αλληλούχιση, όπου όπου ο όρος πρωτογενής μετρητής (raw counts) X είναι ο αριθμός των θραυσμάτων τα οποία επικαλύπτονται με την ένωση των εξονίων ενός γονιδίου, το σύμβολο l είναι το μήκος των μεταγράφων και το σύμβολο N είναι ο αριθμός των χαρτογραφημένων μικροαναγνώσεων.

2.2.2.3. miRNA-Seq

Η χρήση βιοπληροφοριακών εργαλείων (dry experiments) για την αναγνώριση των μορίων miRNA, και των στόχων αυτών των μορίων, ονομάζονται in silico και συνδυάζονται με πειράματα στο εργαστήριο in vivo και in vitro (wet experiments). Πολλές από τις μεθόδους ανάλυσης της RNA-Seq χρησιμοποιούνται για την miRNA –Seq ανάλυση αφού είναι και οι δύο μεταγραφομικές αναλύσεις, αλλά έχουν τις δικές τους βάσεις δεδομένων και τα δικά τους

λογισμικά ανάλυσης. Ο Πίνακας 7, περιέχει κάποια τέτοια χρήσιμα εργαλεία για την αναζήτηση των miRNA, π.χ. στην εκ νέου αλληλούχιση. Η miRNA-Seq γίνεται με αντίστοιχο τρόπο με την RNA-Seq, αλλά διαφέρουν στην προετοιμασία της βιβλιοθήκης. Για την δημιουργία της βιβλιοθήκης μετά την απομόνωση του RNA (και στην συνέχεια απομόνωση του μορίου miRNA με π.χ. πήκτωμα ηλεκτροφόρησης) ή του miRNA, γίνεται η περίδεση του προσαρμογέα στο 3'-άκρο και το 5'-άκρο με το ένζυμο λιγκάση. Στην συνέχεια, συντίθεται το cDNA μέσω της αντίστροφης μεταγραφής και επαυξάνεται με την PCR. Μετά γίνεται η αλληλούχιση και λαμβάνουμε τα πρωτογενή δεδομένα, τα οποία ελέγχονται ως προς την ποιότητα.

Ο αλληλουχοποιητής καταγράφει την ποιότητα (Q) για κάθε βάση κατά την ταυτοποίηση της παρουσίας του σε συγκεκριμένη θέση. Η ποιότητα υπολογίζεται από τον ποσοστό σφάλματος (E) : $Q = -10\log E$, δείκτης Phred. Οι μικροαναγνώσεις που έχουν χαμηλή ποιότητα, δηλαδή πάνω από 5 βάσεις κάτω από Q20, θα αφαιρεθούν στο φιλτράρισμα των δεδομένων. Ακόμα, στο φιλτράρισμα θα αφαιρεθούν και οι συνθετικοί προσαρμογείς που χρησιμοποιήθηκαν για την PCR, που είναι πολύ μεγάλοι σε σχέση με το μικρό μήκος των miRNAs. Μέσω του barcoding στους προσαρμογείς ξεχωρίζουν τα διάφορα δείγματα που μπορεί να υπάρχουν στην ίδια αντίδραση αλληλούχισης. Μετά ελέγχεται η στατιστική κατανομή του μήκους των αλληλουχιών για να είναι ανάλογη με αυτή του ώριμου μορίου miRNA.

Στα φίλτρα γίνεται ο χαρακτηρισμός των γονιδίων, δηλαδή στοιχίζονται με βοήθεια από τις βάσεις δεδομένων miRNA, μια από τις πιο γνωστές, η miRBase[5], περιέχει δεδομένα μορίων miRNA ανθρώπου, ποντικού κ.α. ειδών. Αν η αναζήτηση σε αυτές τις βάσεις δεδομένων αποδειχθεί άκαρπη, χρησιμοποιούνται βάσεις δεδομένων άλλων μικρών μη κωδικοποιών μορίων RNA, όπως την Rfam[6] (94, 95). Με την σύγκριση στις βάσεις δεδομένων των miRNAs και με μεθόδους πρόγνωσης, τα δεδομένα που δεν στοιχίστηκαν με τις βάσεις δεδομένων γνωστών μορίων miRNA μπορούν να χρησιμοποιηθούν εργαλεία για εκ νέου αλληλούχιση, όπως το miRDeep2 (96). Οι μικροαναγνώσεις χαρτογραφούνται στο γένωμα του υπό εξέταση δείγματος και η αλληλουχία του προδρόμου miRNA μπορεί να απομονωθεί από την χαρτογραφημένη περιοχή. Τα νέα miRNA συνήθως βρίσκονται στον μίσχο της δομής μίσχου-θηλιάς (97).

[5] https://www.mirbase.org/
[6] https://www.sanger.ac.uk/tool/rfam/

2.2.2.4. methyl-Seq

Bisulfite-Seq. Μια από τις πιο διαδεδομένες μεθόδους ανάλυσης της μεθυλίωσης DNA σε γενωμική κλίμακα είναι η δισουλφιδική αλληλουχούχιση (Bisulphite-Seq), δηλαδή η κατεργασία του δείγματος με θειώδες οξέος ή άλατος. Είναι σύνηθες η μονονουκλεοτιδικοί πολυμορφισμοί του DNA να προκαλούνται από την αυθόρμητη υδρολυτική απαμίνωση, δηλαδή την αφαίρεση αμινομάδας. Η εκμετάλλευση αυτής της ενδογενούς διεργασίας έκανε δυνατή την αλληλούχιση του μεθυλιώματος. Μετά την χημική απαμίνωση οι μεθυλιωμένες κυτοσίνες θα παραμείνουν κυτοσίνες, ενώ οι υπόλοιπες θα μετατραπούν σε ουρακίλες (**Εικόνα 16**). Κατά την επαύξηση συγκρίνεται με την ακατέργαστη (native) αλληλουχία και η εμφάνιση των θυμίνων, που είναι συμπληρωματικές προς την κυτοσίνη, μας βοηθάει να διακρίνουμε τις κυτοσίνες που δεν έχουν αλλάξει από την ακατέργαστη αλληλουχία, δηλαδή 5-μεθυλοκυτοσίνες. Στην δισουλφιδική αλληλούχιση σε όλο το γονιδίωμα (Whole Genome Bisoulfite Sequencing, WGBS), μπορούν να αναλυθούν με ακρίβεια πάνω από 90% των περιοχών CpG και 25% των εμβρυακών κυττάρων (98, 99). Η συγκεκριμένη αλληλούχιση είναι λίγο ακριβή, ενώ περιέχει και άσχετες πληροφορίες για το μεθυλίωμα, οπότε χρησιμοποιείται περισσότερο για την αλληλούχιση δειγμάτων αναφοράς (100). Άλλες BS-Seq είναι η Tet-assisted bisulfite sequencing (TAB-Seq) και η Oxidative bisulfite sequencing (oxBS-Seq).

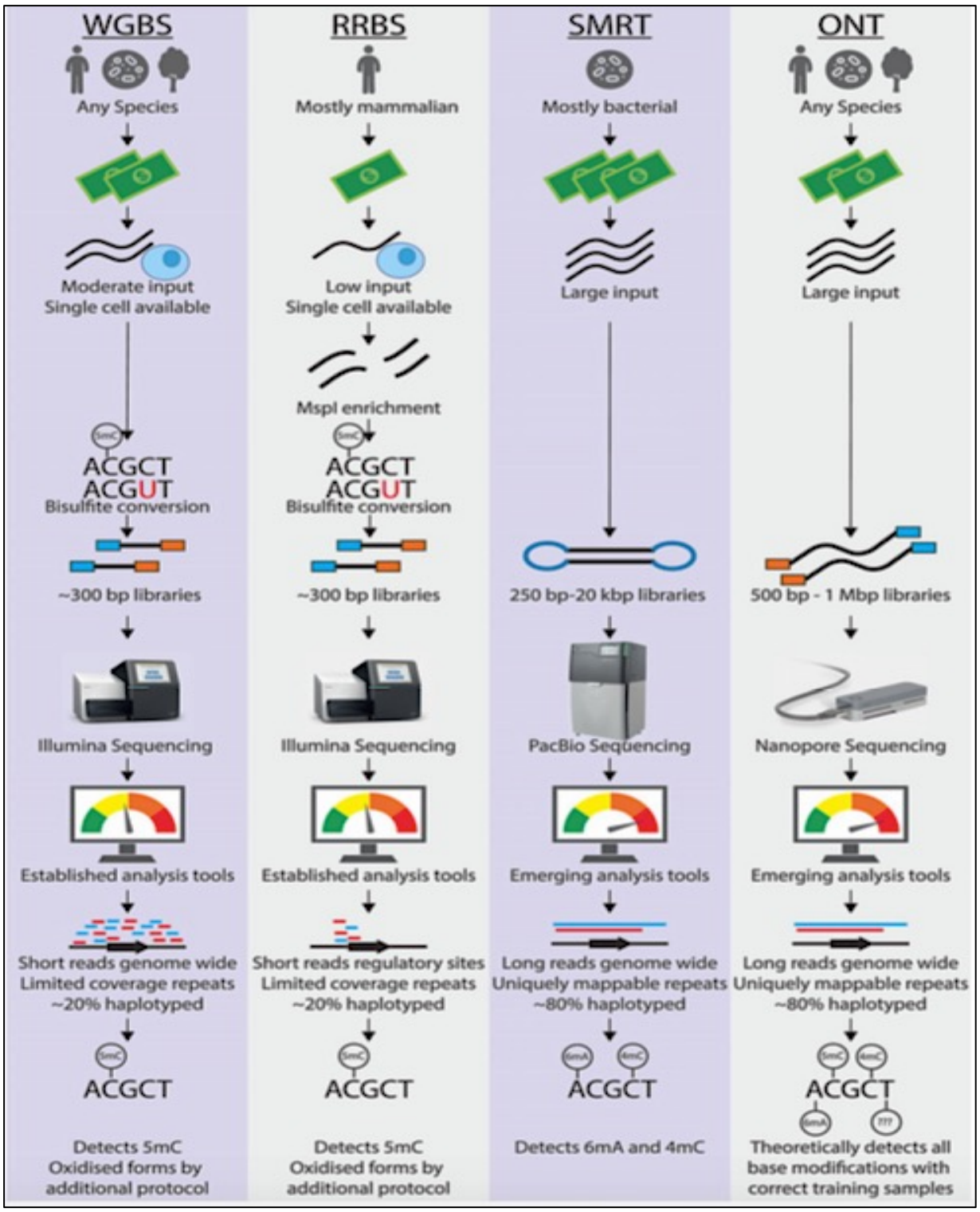
Μία πιο στοχευμένη τεχνική είναι η δισουλφιδική αλληλούχιση με ελαττωμένη αντιπροσώπευση (reduced representation bisulfite sequencing, rrBS), που βασίζεται στον περιορισμό πέψης του γενωμικού DNA, ενώ χρησιμοποιείται στην ανάλυση μεθυλίωσης στα CGIs που οδηγούν σε κακοήθεια (102, 103). Ο εμπλουτισμός διαφορετικών περιοχών του γονιδιώματος, π.χ. των CGIs, μπορεί να επιτευχθεί μέσω της ανάμειξης διαφορετικών περιοριστικών ενζύμων και, συνεπώς, η αλληλούχιση παρουσιάζει μεγαλύτερη κάλυψη των αλληλουχιών-στόχων ενώ παράλληλα μειώνεται και το κόστος της (100, 104). Η μετατροπή των δισουλφιδίων μπορεί να αυξήσει τα σφάλματα στην αλληλούχιση, κάτι που μετριάζεται με την δημιουργία συνθετικών βιβλιοθηκών, και χρειάζεται να εξαλειφθούν οι στατιστικές προκαταλήψεις, λόγω της επαύξησης PCR (99, 105). Ένα βιοπληροφορικό πρόβλημα είναι η πολυπλοκότητα της ανάλυσης του μεθυλιώματος, αφού γίνονται συγκρίσεις μεταξύ των τροποποιημένων και των native αλληλουχιών (100).

MRE-Seq (Methylated Restriction Enzyme-Sequencing). Με την μέθοδο πέψης MRE με αλληλουχοποίηση (MRE-seq), διασπούν τα περιοριστικά ένζυμα μόνο τις μη μεθυλιωμένες

περιοχές CpG του γενωμικού DNA και, στην συνέχεια, επιλέγονται ανάλογα με το μέγεθός τους τα θραύσματα DNA και αλληλουχοποιούνται. Από την ανάλυση εμφανίζονται οι περιοχές αναγνώρισης στις οποίες τα δινουκλεοτίδια δεν μεθυλιώνονται (106). Η MRE-seq μπορεί να εκτιμήσει τα σχετικά επίπεδα της μεθυλίωσης DNA αλλά έχει μικρό ποσοστό κάλυψης ολόκληρου του γενώματος γιατί είναι λίγες οι περιοχές αναγνώρισης που έχουν δινουκλεοτίδια (60).

MeDIP-Seq (Methylated DNA Immunoprecipitation-Sequencing, MeDIP). Είναι μια μέθοδος που χρησιμοποιείται για πάνω από μια δεκαετία για την αλληλούχιση του μεθυλιώματος του DNA στο γένωμα, βασίζεται στους ανιχνευτές αντί-5-μεθυλοκυτοσίνης (anti-5-mC) αντισώματα για να ταυτοποιήσει την παρουσία της 5-μεθυλοκυτοσίνης στα θραύσματα του γενωμικού DNA (107). Αρχίζει με τον κατακερματισμό και την αποδιάταξη του μορίου DNA για τον σχηματισμό μονόκλωνων θραυσμάτων και, στη συνέχεια, γίνεται η ανοσοκατακρήμνιση των μεθυλιωμένων θραυσμάτων DNA με τα αντισώματα anti-5-mC. Το εμπλουτισμένο μεθυλιωμένο θραύσμα DNA προσδένεται με τους προσαρμογείς της αλληλουχοποίησης για την προετοιμασία των βιβλιοθηκών. Με τον ίδιο τρόπο μπορεί να αναλυθεί η 5-hmC αντικαθιστώντας με τα αντισώματα 5-hmC (hMeDIP-seq), τέτοιες αναλύσεις μεθυλίωσης έχουν δημοσιευθεί για το γένωμα ανθρώπου και ποντικιού (108, 109). Μια άλλη τεχνική εμπλουτισμού για την ανάλυση της μεθυλίωσης DNA είναι η MBD-seq (Methyl-CpG binding domain-based capture and sequencing), που χρησιμοποιεί πρωτεΐνες για τον εμπλουτισμό του DNA. Μετά το εμπλουτισμένο DNA απομακρύνεται από το σύμπλοκο DNA και πρωτεϊνών. Η γενωμική βιβλιοθήκη προετοιμάζεται και το μεθυλίωμα αλληλουχοποιείται για να αναγνωριστούν οι περιοχές του.

SMRT-Seq. Η SMRT-seq χρησιμοποιεί τους ZMW για την αλληλούχιση του μεθυλιώματος με την ενσωμάτωση των σημασμένων νουκλεοτιδίων από την πολυμεράση DNA και την καταγραφή του χρονικού διαστήματος μεταξύ των ενσωματώσεων (84, 110). Σε σχέση με την Bisulphite-Seq, που καταγράφει κατευθείαν τα επίπεδα της μεθυλίωσης DNA από την αλληλουχία, στην SMRT-seq χρειάζονται στατιστικά τέστ για να αναγνωρίσουμε μικρά ποσοστά μεθυλίωσης και γι'αυτό δεν είναι η καλύτερη εφαρμογή για γενωμική ανάλυση μεθυλίωσης στα θηλαστικά (100, 111). Με την SMRT-Seq μπορεί να αναλύεται παράλληλα η μεθυλίωση σε όλα τα μέρη του γενώματος χωρίς να έχει προηγηθεί επαύξηση. Το υψηλότερο κόστος και τα υψηλότερα ποσοστά σφαλμάτων είναι κάποια από τα μειονεκτήματα αυτής της τεχνικής.



Εικόνα 16. Ανίχνευση μεθυλοκυτοσίνης (101).

Nanopore-Seq (ONT). Η σύγκριση των σημάτων σε περιοχές μεθυλιωμένου DNA και μη μεθυλιωμένου DNA, με διαφορετικά χαρακτηριστικά σύμφωνα με την διακύμανση των ιόντων, οδηγεί στην αναγνώριση του μεθυλιώματος. Η τιμή της γενωμικής Nanopore-seq είναι αντίστοιχη με της WGBS. Όμως, δεν γνωρίζουμε ακόμα όλες τις δυνατότητες και την

ευαισθησία της στην ταυτοποίηση της μεθυλίωσης DNA. Ακόμα, δεν υπάρχουν γενικευμένοι αλγόριθμοι που να κάνουν σωστή βιοπληροφορική ανάλυση σε όλους τους οργανισμούς. Ενώ κάθε φορά που αναβαθμίζεται η χημεία της πλατφόρμας αλληλουχοποίησης, αλλάζει και το ακατέργαστο σήμα, οπότε οι αλγόριθμοι πρέπει να τροποποιηθούν (101). Όπως και στην SMRT-seq, τα κενά στην αλληλούχιση καλύπτουν στατιστικά τέστ, κρυφά μοντέλα Markov και νευρωνικά δίκτυα, που μπορούν να βελτιώσουν την ακρίβεια σύμφωνα με τα περιβάλλοντα δεδομένα (context dependent) σε 80–85% (100).

Πίνακας 6. Δημοφιλείς μέθοδοι αλληλούχισης μεθυλιώματος (112, 113).

Τύπος	Τρόπος	Αρχή Λειτουργίας	Πλεονεκτήματα	Μειονεκτήματα
Δισουλφιδική	WGBS	Βασίζεται στην προετοιμασία βιβλιοθηκών (Προσαρμογείς, Εκκινητές βασισμένοι σε προσαρμογείς, Εκκινητές αλληλούχισης), στην δισουλφιδική μετατροπή και στην αλληλουχιση. Κάθε μεθυλιωμένη ή μη κυτοσίνη μπορεί να ανιχνευτεί στον συν ή στον πλήν κλώνο.	Μεμονωμένη διακριτική ικανότητα για τα δινουκλεοτίδια CpG. Δεν προϋποθέτει ομογενή μεθυλίωση. Επιτρέπει την ανακάλυψη νέων διαφορικά μεθυλιωμένων περιοχών. Ανιχνεύει και την μεθυλίωση που δεν βρίσκεται σε περιοχές CpG.	Ακριβό. Αποδίδει μεγάλη ποσότητα δεδομένων. Απαιτεί επαρκή ποσότητα ποιοτικού DNA. Η ανάλυση είναι πολύπλοκη και απαιτεί έμπειρο βιοπληροφορικό. Ένα μεγάλο ποσοστό των δεδομένων μπορεί να μην βγάζει νόημα. Δεν μπορεί να εφαρμοστεί σε μεγάλο μέγεθος δειγμάτων.
Δισουλφιδική	rrBS	Βασίζεται στην αλληλούχιση γενώματος μετά από πέψη μέσω του ενζύμου MspI, που δεν έχει ευαισθησία στην μεθυλίωση. Η μέθοδος είναι παρόμοια με την WGBS αλλά περιέχει πρόσθετα και τον εμπλουτισμό. Προσαρμογείς, εκκινητές βασισμένοι σε προσαρμογείς, εκκινητές αλληλούχισης για την προετοιμασία βιβλιοθηκών.	Μεμονωμένη διακριτική ικανότητα για τα δινουκλεοτίδια CpG (Single CpG resolution). Φτηνότερη από την WGBS. Δεν απαιτεί μεγάλη ποσότητα DNA. Για μεγάλο μέγεθος δειγμάτων είναι πιο πρακτική από την WGBS. Η βιοπληροφορική ανάλυση δεν είναι πολύ απαιτητική.	Η κάλυψη των περιοχών CpG είναι χαμηλότερη από την WGBS. Οριοθετείται από τις περιοριστικές περιοχές. Απαιτεί υπολογιστική ανάλυση πριν το πειραματικό κομμάτι για να ελεγχθεί αν οι υπο εξέταση περιοχές του γενώματος εκπροσωπούνται επαρκώς στο δείγμα μας.
Δισουλφιδική	Αλληλούχιση Νουκλεοσώματος και Μεθυλιώματος (NOMe-seq)	Βασίζεται στην χρήση εξογενούς μεθυλομεταγραφάσης για να ενσωματώσει μεθυλομάδες σε περιοχές CpG που δεν μεθυλιώνονται φυσιολογικά, χωρίς προστασία από νουκλεοσώματα και μεταγραφικούς παράγοντες. Στην συνέχεια χαρτογραφούνται οι πιθανές θέσεις	Μεμονωμένη διακριτική ικανότητα για τα δινουκλεοτίδια CpG. Σχετίζει την μεθυλίωση DNA με την παρουσία νουκλεοσωμάτων και μεταγραφικών παραγόντων.	Είναι κοπιαστική. Απαιτεί επαρκή ποσότητα ποιοτικού DNA. Εξαρτάται από την απόδοση της μεθυλομεταγραφής στα δινουκλεοτίδια (CpG) όταν ενσωματώνουν τις μεθυλομάδες. Τα τρινουκλεοτίδια GCG δεν αναλύονται με ακρίβεια αφού η μεθυλίωση μπορεί να είναι ενδογενής ή εξωγενής.

		νουκλεοσωμάτων, περιοχών πρόσδεσης μεταγραφικών παραγόντων και μεθυλίωσης DNA. Προσαρμογείς, εκκινητές βασισμένοι σε προσαρμογείς, εκκινητές αλληλούχισης για την προετοιμασία βιβλιοθηκών.		
Δισουλφιδική	Agilent SureSelect MethylSeq	Ένα κιτ για ανάλυση της μεθυλίωσης του DNA σε ένα εμπλουτισμένο υπόστρωμα του γενώματος (CGIs και ρυθμιστικά στοιχεία που είναι συνήθως διαφορετικά μεθυλιωμένα). Προσαρμογείς, εκκινητές βασισμένοι σε προσαρμογείς, εκκινητές αλληλούχισης, RNA ανιχνευτές για την προετοιμασία βιβλιοθηκών.	Μεμονωμένη διακριτική ικανότητα για τα δινουκλεοτίδια CpG. Φτηνότερη από την WGBS. Είναι χρήσιμη για μεγάλο μέγεθος δειγμάτων. Επικεντρώνεται στα CGIs και στις διαφορικά μεθυλιωμένες περιοχές (DMRs). Είναι κατάλληλη για την ανάλυση λιγότερο ποιοτικού DNA.	Μόνο ο ένας κλώνος αναλύεται. Οι επαναλαμβανόμενες αλληλουχίες δεν αναγνωρίζονται. Αυτή η μέθοδος έχει λιγότερη κάλυψη των περιοχών CpG από την μέθοδο WGBS.
Εμπλουτισμού	MeDIP-Seq	Βασίζεται στην προετοιμασία βιβλιοθηκών, στον εμπλουτισμό του μεθυλιωμένου DNA με την χρήση των αντισωμάτων anti-5mC και στην αλληλούχιση. Προσαρμογείς και εκκινητές για προετοιμασία βιβλιοθήκης και επιβεβαίωση της.	Ο θόρυβος του σήματος, δηλαδή ο δείκτης SNR, είναι χαμηλότερος σε σχέση με την MBD-Seq. Καλύτερη κάλυψη για τις περιοχές φτωχές σε CpG.	Με αυτή γίνεται η καταγραφή και η μετέπειτα ανάλυση των μεταθετών στοιχείων του ανθρώπινου γενώματος, ειδικά για τα 5mC. Είναι δύσκολη η απόλυτη ποσοτικοποίηση της μεθυλίωσης DNA.
Εμπλουτισμού	MBD-isolated genome sequencing (MiGS)	Η πρωτεΐνη MBD χρησιμοποιείται για τον εμπλουτισμό υψηλώς μεθυλιωμένων περιοχών του γενώματος για να μειώσει το κόστος της αλληλούχισης που ακολουθεί. Προσαρμογείς και εκκινητές για προετοιμασία βιβλιοθήκης και επιβεβαίωση της.	Επειδή αλληλουχεί μόνο υψηλώς μεθυλιωμένες περιοχές είναι προσιτό το κόστος. Χρησιμοποιείται για την εκτίμηση των επιπέδων μεθυλίωσης επαναλαμβανόμενων αλληλουχιών, κεντρομερών και υποτελομερών περιοχών.	Τα αποτελέσματα δίνουν μόνο σχετικά επίπεδα μεθυλίωσης. Έχει χαμηλή διακρίνεια. Μόνο οι υψηλά μεθυλιωμένες περιοχές παρουσιάζουν καθίζιση.
Εμπλουτισμού	SeqCap Epi CpGiant enrichment	Βασίζεται στην προετοιμασία βιβλιοθηκών από γενωμικά θραύσματα DNA, που έχουν περάσει από δισουλφιδική κατεργασία, προεπαυξηθεί και	Μπορεί να χρησιμοποιηθεί προσωποποιημένο σχέδιο. Δεν απαιτεί μεγάλη ποσότητα εισερχόμενου δείγματος.	Είναι σχετικά ακριβή.

		υβριδοποιηθεί σε ανιχνευτές καταγραφής, εκλουσθεί και επαυξηθεί πρίν από την αλληλουχοποίηση.		
Εμπλουτισμού	MBD-seq/ MethylCap- seq	Εμπλουτισμός μεθυλιωμένου DNA μέσω της πρωτεΐνης της περιοχής πρόσδεσης μεθυλομάδας, πριν ή μετά την προετοιμασία της βιβλιοθήκης και την αλληλούχιση που ακολουθεί.	Είναι κατάλληλη και για αναλύσεις από δείγματα φολμαλδεΰδης (formaldehyde-fixed paraffin-embedded, FFPE), υψηλής απόδοσης και μεγάλων πληθυσμών.	Είναι δύσκολη η απόλυτη ποσοτικοποίηση της μεθυλίωσης DNA και μόνο για τις 5mC.
Περιορισμού	MRE-seq	Βασίζεται στην πέψη γενωμικού DNA παράλληλα με διάφορα ένζυμα περιορισμού ειδικά για την μεθυλίωση και στην συνέχεια την προετοιμασία βιβλιοθήκης και την αλληλούχιση.	Χρησιμοποιείται κατα προτίμιση για την ανάλυση μη μεθυλιωμένων CpGs σε υψηλής περιεκτικότητας περιοχές CpG. Αν συνδυαστούν με την μέθοδο MeDIP υπάρχει ολική κάλυψη του γενώματος. Μεμονωμένη διακριτική ικανότητα για τα δινουκλεοτίδια CpG.	Οριοθετείται από τις περιοριστικές περιοχές. Χρησιμοποιείται κυρίως για να συμπληρώσει την MeDIP- seq.
Δισουλφιδική	Amplicon BS-Seq	Οι περιοχές ενδιαφέροντος επαυξάνονται με PCR από DNA με δισουλφιδική κατεργασία, ενώ οι βιβλιοθήκες προετοιμάζονται και η αλληλουχοποιούνται με NGS (MiSeq/IonTorrent).	Αποτελείται από κλωνικά θραύσματα και παρέχει υψηλή κάλυψη. Τα πολλαπλά άμπλικονς και οι μεμονωμένες αλληλουχίες μπορούν να αναλυθούν παράλληλα. Μπορεί να χρησιμοποιηθεί με μικρορευστονικές διατάξεις ή συστήματα σταγονιδίων για την προετοιμασία προϊόντων PCR.	Η ανάλυση είναι πολύπλοκη και απαιτεί έμπειρο βιοπληροφορικό. Είναι σχετικά ακριβή.

2.3. Ροή Εργασιών Γενωμικής Ανάλυσης

Γενικά βήματα που ακολουθούνται για την ανάλυση μικροσυστοιχιών και αλληλουχοποίησης.

Α) Δειγματοληψία.

Αποκομιδή των πρωτογενών δεδομένων, συνήθως του εξεταζόμενου δείγματος και ενός φυσιολογικού δείγματος (δείγμα ελέγχου).

Β) Αλληλούχιση.

Στις μικροσυστοιχίες γίνεται μέτρηση φθορισμού και στην αλληλουχοποίηση νέας γενιάς γίνεται αλληλούχιση των νουκλεοτιδίων και η συναρμολόγηση τους.

Γ) Μετά την σάρωση της πρωτογενούς εικόνας γίνεται εφαρμογή πλέγματος (gridding) και κατάτμηση (segmentation), με σκοπό την αποκομιδή των αποτελεσμάτων. Στις μικροσυστοιχίες γίνεται εντοπισμός των spots (στατική εικόνα spotted) και στην

αλληλουχοποίηση γίνεται εντοπισμός ολόκληρων μορίων (δυναμική η μορφή της πληροφορίας). Όμως δεν έχω χρωστικές σε όλες τις περιπτώσεις των αλληλουχοποιητών, π.χ. στις τεχνολογίες IonTorrent, SMRT, Nanopore δεν γίνεται σάρωση εικόνας, οπότε η ποσοτικοποίηση γίνεται με μειωμένα επίπεδα θορύβου.

Δ) Προεπεξεργασία πρωτογενών δεδομένων. Γίνεται η ψηφιακή επεξεργασία της εικόνας (εξαγωγή έντασης μέσω ποσοτικοποίησης του σήματος, διόρθωση υποβάθρου και φιλτράρισμα).

Σε αυτό το σημείο γίνεται η πιο σημαντική λειτουργία της ανάλυσης. Δηλαδή, η κανονικοποίηση και ο έλεγχος ποιότητας των δεδομένων. Αν έχουμε κακής ποιότητας δεδομένα πρέπει να αρχίσουμε το πείραμα από την αρχή.

Ε) Προσδιορισμός διαφορικά εκφραζόμενων γονιδίων μέσω στατιστικών μεθόδων επεξεργασίας των δεδομένων.

Ζ) Βιοπληροφορική ανάλυση (π.χ. αναγνώριση προτύπων και μελλοντικές εκτιμήσεις).

Η) Διαδικασία χαρακτηρισμού γονιδίων.

Θ) Μοντελοποίηση και Big Data.

3. Κεφάλαιο 3 Μέθοδος Ανάλυσης Μικροσυστοιχιών

Ανάλυση των βιολογικών σημάτων. Σε αυτό το κομμάτι της εργασίας έχουμε παρόμοια ανάλυση στις μικροσυστοιχίες και στην αλληλουχοποίηση νέας γενιάς με τις τεχνικές που βασίζονται στην σάρωση εικόνας, δηλαδή 2^{ης} γενιάς NGS (αντιπαράδειγμα η τεχνική IonTorrent). Το 2001 ο ερευνητής Brazma και οι συνεργάτες του δημοσίευσαν μια εργασία στην οποία πρότειναν ένα στάνταρ για την αποθήκευση γονιδιακών δεδομένων από πειράματα μικροσυστοιχιών MIAME (Minimum Information About a Microarray Experiment), το οποίο αποδείχτηκε χρήσιμο για την ποιοτική ανάλυση των δεδομένων (114). Τα στοιχεία που μας απασχολούν είναι τα πρωτογενή και τα κανονικοποιημένα δεδομένα για κάθε υβριδοποίηση, οι πληροφορίες για τα δείγματα και τις τιμές τους, την πειραματική διαδικασία και τους παράγοντές της, ο χαρακτηρισμός γονιδίων της μικροσυστοιχίας και τα βιοπληροφορικά λογισμικά ανάλυσης των δεδομένων.

Μετά το πειραματικό κομμάτι της μεθόδου μικροσυστοιχιών τα πρωτογενή δεδομένα, που μας δίνονται σε spots, ακολουθούνται από την προεπεξεργασία και την ποσοτικοποίηση των αποτελεσμάτων σε λίστες γονιδίων ανάλογα με την φωτεινότητα του κάθε spot, σε σύγκριση με την έκφραση του γονιδίου σε κάθε δείγμα. Για την εύρεση των διαφορικά εκφραζόμενων γονιδίων (DEG) εφαρμόζουμε στατιστικούς ελέγχους στις λίστες αυτές, συνήθως σε ένα μορφή αρχείου EXCEL (.xlsx ή .xls) της Microsoft.

Έστω ότι λαμβάνουμε τις λίστες με 1000 γονίδια σε δύο αρχεία .xls ένα για 10 δείγματα ελέγχου και ένα για 10 υπό εξέταση δείγματα. Μας ενδιαφέρει αν ο μέσος όρος των 10 δειγμάτων που εξετάστηκαν για το κάθε γονίδιο διαφέρει σημαντικά από τον μέσο όρο των 10 φυσιολογικών δειγμάτων. Έστω, επίσης, ότι από τα 1000 γονίδια, τα 100 έχουν σημαντικά διαφορετικούς μέσους όρους, αυτά ονομάζονται διαφορικά εκφραζόμενα γονίδια. Η σημαντικότητα της διαφορικής έκφρασης μεταξύ των δύο ομάδων είναι η βάση της αξιοπιστίας της ανάλυσης των δεδομένων μικροσυστοιχιών. Γι'αυτόν τον λόγο έχουν δημιουργηθεί πολλές στατιστικές μέθοδοι και εφαρμογές, που χρησιμοποιούνται ανάλογα με την περίσταση. Πιθανές στατιστικές καταχρήσεις μπορεί εύκολα να αλλοιώσουν την βιολογική πληροφορία.

Κάποια δημοφιλή πρόσθετα, ελεύθερης πρόσβασης, για την ανάλυση δεδομένων μικροσυστοιχιών στο EXCEL, είναι το SAM (Ανάλυση σημαντικότητας μικροσυστοιχιών, Significance Analysis of Microarrays)[7], PAM (Ανάλυση εκτίμησης για μικροσυστοιχίες, Prediction Analysis for Microarrays)[8] και το BRB[9]. Το ολοκληρωμένο πακέτο BRB ArrayTools είναι φτιαγμένο για την οπτικοποίηση και την στατιστική ανάλυση των δεδομένων έκφρασης μικροσυστοιχιών DNA από τον τομέα βιομετρικής έρευνας Biometrics Research Branch του NCI, παρόλο που χρησιμοποιείται στο Excel, τα εργαλεία του προγραμματίστηκαν σε R, C, Fortran και Java. Η εξοικείωση με το λογισμικό Microsoft EXCEL είναι βασική δεξιότητα, αφού το χρησιμοποιούν οι εργαζόμενοι στα περισσότερα γραφεία και διδάσκεται από το δημοτικό. Με αυτό μπορούν να εφαρμοστούν πολλές μέθοδοι για την στατιστική ανάλυση, αρχίζοντας με την εύρεση του μέσου όρου για κάθε γονίδιο ξεχωριστά, και, στην συνέχεια, εφαρμόζουμε στατιστικούς ελέγχους. Εκτός αυτού, τα μεγάλα αρχεία EXCEL είναι δύσχρηστα και καθίσταται δύσκολο να τα αναλύσουμε με εντολές .macro, αφού είναι χρονοβόρα διαδικασία και μπορεί το λογισμικό να μην αντέχει την απαιτητική υπολογιστική δύναμη που χρειαζόμαστε. Ακόμα, υπάρχουν πιο κατάλληλα λογισμικά για πολύπλοκη στατιστική ανάλυση, όπως το SAS, το STATA και το MATLAB (μαθηματικός προγραμματισμός).

Πολλοί ερευνητές έχουν ασχοληθεί και με το πόσες επαναλήψεις ενός πειράματος (replicates) δίνουν την μεγαλύτερη αξιοπιστία, χωρίς να γίνει τόσες φορές που θα είναι ακριβό. Ένα τέτοιο παράδειγμα επαναλαμβανόμενου πειράματος είναι η δικάναλη ανάλυση του ίδιου mRNA αλλά αντιστρέφοντας τις φθορίζουσες ουσίες σε κάθε ανεξάρτητο πείραμα. Βεβαίως, και το βιολογικό repilcate χρειάζεται να περάσει από τα ίδια βήματα της σάρωσης, της ποσοτικοποίησης και την κανονικοποίησης. Είναι σύνηθες να γίνονται τουλάχιστον τρεις επαναλήψεις, αλλά για ακόμα καλύτερα αποτελέσματα και για την εξάλειψη των ακραίων και των εσφαλμένως θετικών (False Positive, FP) τιμών χρειάζονται περισσότερες επαναλήψεις. Μπορούμε να ελέγξουμε πόσες επαναλήψεις θα χρειαστούν, για να είναι μην είναι ακριβή αλλά να είναι αξιόπιστη η ανάλυση, με το ποσοστό εσφαλμένων ανακαλύψεων (False Discovery Rate, FDR). Αυτός ο δείκτης αναφέρεται στο εκτιμώμενο ποσοστό όλων των δοκιμών που δηλώθηκαν σημαντικές και ελέγχει τον αριθμό των δοκιμών που δηλώθηκαν ψευδώς

[7] https://statweb.stanford.edu/~tibs/SAM
[8] https://statweb.stanford.edu/~tibs/PAM/
[9] https://brb.nci.nih.gov/BRB-ArrayTools/

σημαντικές. Ανάλογα με την ανάλυση μπορεί και ένας υψηλός FDR να είναι ικανοποιητικός αλλά και πιο αποδοτικός. Σε κάποιες εφαρμογές μας ενδιαφέρει να έχουμε υψηλή ευαισθησία, π.χ. στον έλεγχο διάγνωσης του COVID- 19 θέλουμε δηλαδή να αποφύγουμε τα εσφαλμένως αρνητικά (false negative, FN) γιατί μπορεί να κολλήσει κάποιον άλλο, ενώ σε κάποιες μας ενδιαφέρει η υψηλή ακρίβεια, π.χ. στον έλεγχο αντισωμάτων για τον COVID-19 θέλουμε να αποφύγουμε τα FP γιατί κάποιος μπορεί να θεωρεί εσφαλμένα ότι είναι προστατευμένος.

3.1. Προεπεξεργασία Δεδομένων

Μετά την αποκομιδή των πρωτογενών δεδομένων, ακολουθούνται τα διάφορα βήματα για τον καθαρισμό τους, ώστε να έχουμε όσο επαναλαμβανόμενα, αναπαράξιμα και συγκρίσιμα αποτελέσματα. Μας ενδιαφέρουν ακόμα η ευαισθησία και η ειδικότητα του πειράματος. Η ευαισθησία είναι η ικανότητα να ανιχνευτούν οι πραγματικές διαφορές (true-positive rate, TP). Η ειδικότητα είναι η ικανότητα να ανιχνευτούν οι διαφορές μόνο όταν υπάρχουν (true-negative rate, TN). Το έργο κριτηρίων ποιότητας (MAQC) έχει πολλές οργανωμένες δημόσιες βάσεις δεδομένων, που ασχολούνται με την επαναληψιμότητα ενός μεμονωμένου πλακιδίου, την αναπαραξιμότητα μεταξύ replicates και την συγκρισιμότητα μεταξύ διαφορετικών πλατφόρμων (115). Η συστηματική ποικιλομορφία μπορεί να αφαιρεθεί με την κανονικοποίηση, ενώ η στοχαστική ποικιλομορφία αφαιρείται με στατιστικούς ελέγχους.

3.1.1. Αφαίρεση Θορύβου (Noise Removal)

Η αφαίρεση του θορύβου αποτελείται κυρίως από την απομάκρυνση τεχνικών σφαλμάτων (artifacts) και την διόρθωση του υποβάθρου (background correction), δηλαδή θέλουμε να ελαχιστοποιείται ο θόρυβος. Το υπόβαθρο θεωρούμε ότι έχει ομοιόμορφη ένταση θορύβου. Θεωρητικά αυτή η διαδικασία αποδίδει μια αμερόληπτη εκτίμηση του πραγματικού σήματος, ωστόσο προκύπτουν διορθωμένες εντάσεις με ανεπιθύμητες στατιστικές ιδιότητες.

Επειδή η τεχνική μικροσυστοιχιών περιλαμβάνει πολλά στάδια, υπάρχει η πιθανότητα για τεχνικά σφάλματα σε κάθε ένα από αυτά. Για παράδειγμα, στο στάδιο της έκπλυσης και της σάρωσης και ο ανθρώπινος παράγοντας σε αυτά. Ακόμα, οι αλλαγές στην θερμοκρασία, στην φωτοευαισθησία και στην υγρασία του εργαστηρίου, μπορεί να επηρεάσουν την επαναληψιμότητα του πειράματος, αφού εμφανίζουν διαφορές στην αποδοτικότητα ενσωμάτωσης της φθορίζουσας ουσίας και γενικά στα αποτελέσματα του φθορισμού. Ακόμα, υπάρχει και η περίπτωση να έχει επιμολυνθεί το δείγμα και να υπάρχει διασταυρούμενη

υβριδοποίηση ή μη ειδική υβριδοποίηση, που είναι σφάλμα που δεν αφαιρείται. Το πλακίδιο να έχει τεχνικές ατέλειες, δηλαδή γρατζουνιές ή επικαθισμένη σκόνη ή αποτυπώματα, που μπορεί να θεωρηθεί κατά λάθος βιολογικής σημασίας. Ειδικά όταν τα πλακίδια είναι φτιαγμένα με ρομποτικά μηχανήματα στο κάθε εργαστήριο, οι πείροι του μπορεί να προσθέσουν συστηματική ποικιλομορφία στα αποτελέσματα λόγω της γεωμετρίας τους αλλά ακόμα και οι ίδιοι πείροι λόγω διαφορετικής ποσότητας και ακρίβειας εναπόθεσης γενετικού υλικού στην πλακέτα.

Τα πρωτογενή δεδομένα από εφαρμογές δικάναλων πειραμάτων με μικροσυστοιχίες DNA λαμβάνονται σε αρχεία εικόνας .tiff, ένα αρχείο 16-bit για κάθε κανάλι (φθορίζουσα ουσία), και με ένα πρόγραμμα ανάλυσης εικόνας εκτιμούν τα σημεία των spots, γίνεται η κατάτμηση και η εξαγωγή των πληροφοριών, δηλαδή των τιμών φωτεινότητας του σήματος και του υποβάθρου, καθώς και πληροφορίες για την ποιότητα του πειράματος. Στη συνέχεια, γίνεται η προεπεξεργασία για την διόρθωση υποβάθρου, όπου συνήθως αφαιρούμε τις τιμές υποβάθρου από τις τιμές φωτεινότητας του πραγματικού σήματος, όπως στην **Εξίσωση 5**. Ενώ υπάρχουν και πιο εξειδικευμένοι μέθοδοι, οι οποίοι βασίζονται σε ειδικούς αλγόριθμους, όπως το πακέτο limma (Linear Models of Microarrays, LIMMA) του λογισμικού BioConductor[10], το οποίο βασίζεται στην προγραμματιστική γλώσσα R.

$$R' = R - R_{background}$$
$$G' = G - G_{background}$$

Εξίσωση 5. Διόρθωση υπόβαθρου (background).

Τα λογισμικά ανάλυσης εικόνας κάνουν πλέον πιο αυτοματοποιημένη προεπεξεργασία δεδομένων, παρέχουν κριτήρια ποιότητας και παρέχουν στατιστικά εργαλεία για μεγαλύτερη αξιοπιστία, με την βοήθεια διαφόρων τιμών, όπως της μέσης τιμής (mean, συνήθως συμβολίζεται με «μ»), της διαμέσου τιμής (median) και της τυπικής απόκλισης (Standard Deviation, SD, συνήθως συμβολίζεται με «σ»). Το λογισμικό ανάλυσης εικόνας από την κάθε εταιρία κατασκευής των σαρωτών μικροσυστοιχιών είναι ένα ολοκληρωμένο πακέτο λογισμικού (suite), που περιέχει τα περισσότερα εργαλεία που χρειάζεται ο βιοπληροφορικός, χωρίς να καλύπτει όλες τις δυνατότητες των εργαλείων που υπάρχουν, και με φιλικό προς τον χρήστη περιβάλλον, συγκεκριμένα δίνει πολλές δυνατότητες οπτικοποίησης, δηλαδή

[10] https://www.bioconductor.org

καλύτερης κατανόησης, των αποτελεσμάτων. Είναι, όμως, δύσκολη η πρόσβαση σε αυτά για πολλά εργαστήρια , αφού συνήθως έχουν ακριβές ετήσιες συνδρομές. Εκτός αυτού, μπορεί να μην έχουν ευελιξία και να δημιουργήσουν σφάλματα συμβατότητας, που θα περιορίσουν την αποδοτικότητα της ανάλυσης των δεδομένων. Κάποια ελεύθερα λογισμικά είναι εξίσου καλά με τα εμπορικά, αφού παρέχουν πρόσβαση σε πολλές μεθόδους ανάλυσης, προϋπάρχουσες αλλά και την δυνατότητα να προσθέσει τον δικό του κώδικα ο χρήστης. Το πιο γνωστό ελεύθερο λογισμικό, που χρησιμοποιείται και στο κομμάτι της προεπεξεργασίας, είναι το λογισμικό στατιστικής R/BioConductor.

Μια εναλλακτική είναι το ελεύθερο λογισμικό M4 Suite, που βασίζεται στην γλώσσα Java και παρέχεται από το Ινστιτούτο Γενωμικής Έρευνας (The Institute for Genomic Research, TIGR)[11]. Το εργαστήριο Quackenbush ανέπτυξε ένα ολοκληρωμένο πακέτο ανάλυσης μικροσυστοιχιών με τέσσερις εφαρμογές για την ανάλυση δικάναλων μικροσυστοιχιών. Πιο συγκεκριμένα, το λογισμικό MADAM (MicroArray DAta Manager) κάνει εύκολη την προσθήκη των δεδομένων από την ανάλυση μικροσυστοιχιών σε βάση δεδομένων. Το λογισμικό εύρεσης κηλίδων (Spotfinder) αναλύει τις εικόνες .TIFF για να γίνει η εξαγωγή της φωτεινότητάς τους μέσω ενός δυναμικού αλγόριθμου κατωφλιού. Το λογισμικό MIDAS (MIcroarray Data Analysis System) χρησιμοποιείται για την κανονικοποίηση των δεδομένων. Τέλος, το λογισμικό αναπαράστασης MeV (Multiple Experiment Viewer) χρησιμοποιείται για την εξόρυξη των δεδομένων και την οπτικοποίηση τους (116).

Στα γονιδιακά τσίπ της Affymetrix είναι σύνηθες το μοντέλο ανάλυσης D-Chip[12]. Μπορεί ο βιοπληροφορικός να γράψει και τον δικό του κώδικα, συνήθως σε Python ή R, για να καλύψει πιο ειδικές ανάγκες ανάλυσης των μικροσυστοιχιών, αλλά σε αυτή την περίπτωση το εργαλείο παρέχει μόνο συγκεκριμένες δυνατότητες, πολλές φορές χωρίς την δυνατότητα οπτικοποίησης των αποτελεσμάτων. Ένα άλλο αρνητικό, είναι ότι προϋποθέτει πολύ καλές γνώσεις προγραμματισμού, κάτι που δεν χρειάζεται για να χρησιμοποιήσεις τα έτοιμα λογισμικά, ελεύθερης πρόσβασης ή εμπορικά.

Η μέθοδος κανονικοποίησης μπορεί να είναι τοπική ή ολική, έλεγχος αρνητικών τιμών, που μπορεί να γίνει με μοντέλα σταθεροποίησης της διασποράς (variance stabilizing models,

[11] https://mev.tm4.org/
[12] https://sites.google.com/site/dchipsoft/

VSM), και μορφολογική. Με την τοπική εξομάλυνση μπορούν να διορθωθούν οι εντάσεις ανάλογα με την μορφολογία της μικροσυστοιχίας. Για παράδειγμα, αφαιρούμε τις τιμές από τα υβριδοποιημένα και επαυξημένα μόρια που έχουν υβριδοποιηθεί το ένα πάνω στο άλλο. Οι μορφολογικές μέθοδοι είναι απαραίτητες για την σωστή ανάλυση περιοχών στις οποίες η ένταση στο εκτιμώμενο υπόβαθρο μπορεί να είναι υψηλότερη από την ένταση του πραγματικού σήματος. Αλλιώς το χαρακτηριστικό βιολογικού ενδιαφέροντος μετά την λογαρίθμιση δεν υπάρχει στην λίστα γονιδίων. Πιο συγκεκριμένα, δημιουργούμε ένα αντίγραφο της κάθε μεμονωμένης εικόνας, το οποίο φιλτράρεται για να κατασκευάσει μια εικόνα υποβάθρου. Τα δικάναλα πειράματα έχουν μεγαλύτερο σφάλμα προκατάληψης από τα μονοκάναλα, τα οποία χαρακτηρίζονται από μεγαλύτερη απλότητα και ευελιξία.

Οι πιο γνωστές εφαρμογές μονοκάναλων πειραμάτων είναι με τα ολιγονουκλεοτιδικά τσιπ της Affymetrix, τα οποία έχουν λίγο διαφορετική διόρθωση υποβάθρου από τις μικροσυστοιχίες DNA. Είναι τετράγωνα spots αντί για κυκλικά spots και σκανάρονται με την διαίρεση κάθε κυττάρου σε ένα τετραγωνικό πίνακα εικονοστοιχείων (pixels), και τα επίπεδα φωτεινότητας μετρώνται σε κάθε εικονοστοιχείο, αγνοούνται τα εικονοστοιχεία στις συνοριακές τιμές και αποθηκεύονται τα επίπεδα φωτεινότητας του spot, του ανιχνευτή κυττάρου, στα υπόλοιπα εικονοστοιχεία. Ο υπολογισμός του θορύβου υποβάθρου δίνεται με την **Εξίσωση 6** για κάθε πίνακα ανιχνευτή, όπου N είναι ο συνολικός αριθμός των κυττάρων στο υπόβαθρο, $stdev_i$ είναι η τυπική απόκλιση της φωτεινότητας των εικονοστοιχείων που απαρτίζουν τα κύτταρα στο υπόβαθρο, SF είναι ο συντελεστής κλίμακας και NF είναι ο συντελεστής κανονικοποίησης (117).

$$Q = \frac{1}{N * \sum_i \frac{stdev_i * SF * NF}{\sqrt{pixel_i}}}$$

Εξίσωση 6. Θόρυβος ολιγονουκλεοτιδικής μικροσυστοιχίας Affymetrix.

Στις ολιγονουκλεοτιδικές μικροσυστοιχίες της Affymetrix, πριν την κανονικοποίηση υπάρχει ένα επιπλέον στάδιο περίληψης (Summarizing), αφού οι γονιδιακή έκφραση αναπαρίσταται από πολλούς ανιχνευτές, ενώ εμείς θέλουμε να μετρήσουμε μια τιμή, δηλαδή το μετάγραφο. Στο λογισμικό R/Bioconductor εκτελείται από το πακέτο affy με την εντολή *summary*. Το αρχείο .CEL, που παράγεται από το λογισμικό της Affymetrix από ένα αρχείο πρωτογενών δεδομένων .DAT, περιέχει αυτά τα επίπεδα φωτεινότητας (118).

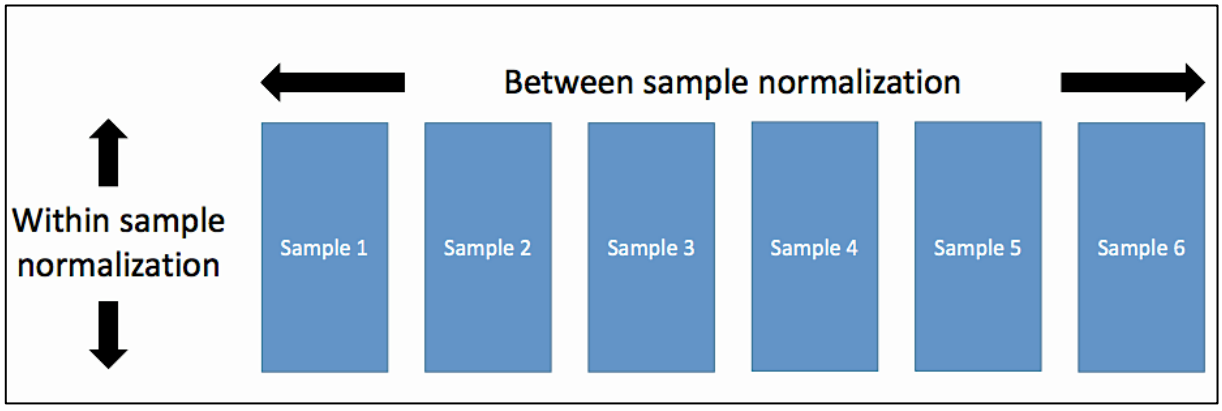
3.1.2. Φιλτράρισμα (Filtering)

Ουσιαστικά είναι η αφαίρεση των σημάτων που δεν ξεπερνάνε κάποιες τιμές (threshold). Φιλτράρονται οι τιμές φωτεινότητας που δεν είναι ανιχνεύσιμες και οι αρνητικές τιμές φωτεινότητας, δηλαδή προσαρμόζεται η εκτίμηση των τιμών έκφρασης στα κατάλληλα επίπεδα. Οι ανιχνευτές ελέγχου και οι αρνητικοί ανιχνευτές ελέγχου, οι οποίοι βρίσκονται σε τυχαία σημεία των μικροσυστοιχιών, είναι σημαντικό εργαλείο στο φιλτράρισμα τιμών για την αποφυγή θορύβου λόγω μη ειδικότητας κατά την υβριδοποίηση.

Το φιλτράρισμα βοηθάει να μειωθεί η πολυπλοκότητα των πολλαπλών ελέγχων και να αυξήσει την στατιστική δύναμη. Η προσοχή παραμένει στα ενεργά γονίδια, δηλαδή αυτά που εκφράζονται, οπότε χρειάζονται λιγότερες διορθώσεις. Ακόμα, θα υπάρχει μεγαλύτερο ποσοστό διαφορικά εκφραζόμενων γονιδίων. Όταν, όμως, αφαιρούνται τα δεδομένα χαμηλότερης ποιότητας μπορεί να χαθεί σημαντική βιολογική πληροφορία των σχετικών γονιδίων.

3.1.3. Κανονικοποίηση (Normalization)

Η κανονικοποίηση είναι η διαδικασία ελαχιστοποίησης της ανεπιθύμητης τεχνικής διακύμανσης μη βιολογικού ενδιαφέροντος στο εσωτερικό (within) του δείγματος ή μεταξύ (between) των δειγμάτων (**Εικόνα 17**).



Εικόνα 17. Εσωτερική και μεταξύ κανονικοποίηση (119).

Η εσωτερική ασχολείται με συστηματικά λάθη εντός των συγκρινόμενων δειγμάτων, ενώ η μεταξύ ασχολείται με συστηματικά λάθη ανάμεσα στα συγκρινόμενα δείγματα και στοχεύει στην αξιόπιστη σύγκριση επιπέδων έκφρασης πολλαπλών πειραμάτων. Η κανονικοποίηση είναι ένα απαραίτητο βήμα, πρέπει όμως πρώτα να κάνουμε κάποιες παραδοχές, δηλαδή ή ότι υπάρχουν λίγα DEGs ή ότι είναι ισάριθμα τα υποεκφρασμένα με τα υπερεκφρασμένα DEGs

ή και τα δύο. Η κανονικοποίηση στα μονοκάναλα πειράματα αφορά σε απόλυτες τιμές φωτεινότητας αντί για σχετικές τιμές φωτεινότητας. Πληροφορίες από πολλαπλά τσίπ μπορεί να χρησιμοποιηθούν για μια έρευνα, οπότε για να αφαιρέσουμε τις τεχνικές διακυμάνσεις λαμβάνουμε υπόψη την συγκεκριμένη ομάδα των δειγμάτων τα οποία εξετάστηκαν παράλληλα και, συνεπώς, υπό όμοιες συνθήκες, ώστε να κανονικοποιηθούν οι τιμές και επηρεαστούν όσο το δυνατόν λιγότερο οι στατιστικές δοκιμές, δηλαδή αφαιρούμε τον θόρυβο που προέρχεται από τις ταυτόσημες επιδράσεις μιας παρτίδας (batch effects). Πιο συγκεκριμένα, αν το εύρος και η διάμετρος των τιμών τους είναι παρόμοια για κάθε δείγμα, τότε έχει πετύχει η κανονικοποίηση. Όταν μας ενδιαφέρει να μετρήσουμε την διαφορική γονιδιακή έκφραση (Differential Gene Expression, DGE) στα δεδομένα υπάρχουν πολλοί μαθηματικοί μέθοδοι για τον μετασχηματισμό των δεδομένων, όπως η παρατηρούμενη τιμή ελέγχου στατιστικής σημαντικότητας (p-value), η λογαριθμισμένη διαφορά έκφρασης (logarithmic fold change, $\log_2$FC) και η απόλυτη διαφορά (absolute difference) (120).

Με την Fold Change καταφέρνουμε να μην λάβουμε υπόψη τις τιμές χαμηλές εκφράσεις, στις οποίες είναι δυσκολότερο να ελεγχθεί η σημαντικότητα και εμφανίζουν υψηλότερα επίπεδα θορύβου. Πρόκειται για μια απλή αφαίρεση των τιμών έκφρασης του γονιδίου ενός δείγματος υπό εξέταση με την έκφραση του ίδιου γονιδίου στο φυσιολογικό δείγμα (Wildtype, Wt), η σχετική έκφραση που παίρνουμε χρησιμοποιείται για περαιτέρω ελέγχους. Αν το γονίδιο έχει τιμή διαφοράς έκφρασης (FC) μέσα στο διάστημα $-1 \leq x \leq 1$ θεωρείται ότι δεν έχουμε διαφορά μεταξύ των δύο δειγμάτων, ενώ αν έχει τιμή fold $x<-1$ θεωρείται υπο-εκφρασμένο και αν έχει τιμή $x>1$ θεωρείται υπερ-εκφρασμένο. Συχνότερα, με την FC και την τιμή p-value μπορούμε να οπτικοποιήσουμε ποια γονίδια είναι υπερεκφρασμένα [logFC $\geq$ 1 & p-value $\leq$ 0,05] και υποεκφρασμένα [logFC $\leq$ -1 & p-value $\leq$ 0,05] με την βοήθεια ενός διαγράμματος κρατήρα ηφαιστείου (Volcano Plot), που εκτελείται με την εντολή *volcanoplot*, από το πακέτο limma του R/Bioconductor. Κάθε τελεία στο διάγραμμα ηφαιστείου αντιπροσωπεύει ένα γονίδιο διαφορικής έκφρασης, ο ένας άξονας είναι για τις λογαριθμισμένες τιμές FC και ο άλλος άξονας είναι η $-\log_{10}$ της p-value για το στατιστικό έλεγχο που χρησιμοποιούμε. Η FC είναι από τις πρώιμες μεθόδους να μετρήσουμε την DGE και δεν είναι λαμβάνει υπόψη την διασπορά, δεν είναι ευαίσθητη σε μικρές αλλαγές, δεν είναι αμερόληπτη στα υποεκφραζόμενα γονίδια, αγνοεί την διαφορετικότητα των γονιδιακών επιπέδων στα replicates και, συνήθως, δεν εκτιμάται ο δείκτης εσφαλμένης θετικότητας (FP). Αντί να χρησιμοποιήσουμε την υψηλή τιμή της FC, ελέγχουμε την σημαντικότητα με κατάλληλους στατιστικούς ελέγχους, με τους οποίους θα ασχοληθούμε στην επόμενη ενότητα. Στο

λογισμικό MATLAB η FC εκτελείται με την εντολή *mavolcanoplot* ενώ στο λογισμικό R/Bioconductor, και συγκεκριμένα στο πακέτο limma, γίνεται με την εντολή *topTabe*.

Η τιμή στατιστικής σημαντικότητας p-value είναι η τιμή του μικρότερου επιπέδου σημαντικότητας για την οποία το παρατηρούμενο στατιστικό δείγμα μας προτρέπει να απορρίψουμε την μηδενική υπόθεση H_0, δηλαδή όταν την συγκρίνουμε με μια τιμή κατωφλίου α και ισχύει ότι p-value <α, θεωρούμε τα γονίδια διαφορικά εκφραζόμενα. Το α ονομάζεται επίπεδο σημαντικότητας ή επιτρεπόμενο επίπεδο σφάλματος τύπου I (False Positive, FP), όπου η H_0 απορρίπτεται λανθασμένα, και το χρησιμοποιούμε στον έλεγχο υποθέσεων για να απορρίψουμε ή όχι την H_0. Μια συνήθης τιμή κατωφλιού είναι η 0,05 (αυθαίρετο κριτήριο), τότε για τιμές $p<0.05$ θεωρούμε ότι έχουμε στατιστική σημαντικότητα. Απορρίπτουμε την μηδενική υπόθεση και θεωρούμε ότι το συγκεκριμένο γονίδιο είναι διαφορικά εκφραζόμενο αν η p-value είναι πολύ μικρή (π.χ. $p < 0,01$), δηλαδή δεν είναι πολύ πιθανό ότι αυτή η διαφορά παρουσιάστηκε τυχαία. Το σφάλμα τύπου II (False Negative, FN), από την άλλη πλευρά, είναι η περίπτωση η μηδενική υπόθεση να μην απορρίπτεται, παρόλο που είναι ψευδής. Σε αντίθετη περίπτωση, δηλαδή αν p-value ≥ α, τότε δεν θα απορρίψουμε την μηδενική υπόθεση H_0 και θεωρούμε τα γονίδια φυσιολογικά.

Αν υπάρχουν replicates ελέγχονται οι τιμές για οποιαδήποτε σημαντική διαφορά με στατιστικά τέστ π.χ. αν έχουμε 3 replicates εξάγουμε για κάθε γονίδιο την μέση τιμή 3 επαναλήψεων για το δείγμα A, την μέση τιμή 3 επαναλήψεων για το δείγμα B και την μέση τιμή 3 επαναλήψεων φυσιολογικού δείγματος (Wt), και πάνω σε αυτές διενεργούμε στατιστικό έλεγχο t-test. Έτσι, μπορούμε να ελέγξουμε και για βιολογικά και για τεχνικά σφάλματα. Ακόμα, με την αφαίρεση spots κακής ποιότητας βελτιώνεται η περαιτέρω ανάλυση. Υπάρχουν πολλοί τρόποι κανονικοποίησης και ο καθένας χρησιμοποιεί τις δικές τους υποθέσεις, συνήθως θεωρείται ότι η διαφορική έκφραση συμβαίνει σε πολύ μικρό ποσοστό γονιδίων και ότι τα γονίδια που υποεκφράζονται είναι ισόποσα με τα γονίδια που υπερεκφράζονται. Η κανονικοποίηση κλίμακας (scaling) εφαρμόζεται μαζί με άλλες κανονικοποιήσεις και πρόκειται για διαίρεση ή πολλαπλασιασμό με μια σταθερά, συνήθως με το μέγιστο ή το ελάχιστο. Στο λογισμικό MATLAB η κανονικοποίηση κλίμακας και το κεντράρισμα εκτελείται με την εντολή *manorm* ενώ στο λογισμικό R/Bioconductor, και συγκεκριμένα στο πακέτο preprocessCore, γίνεται με την εντολή *normalize.constant*. Ανήκει και στις δύο κατηγορίες κανονικοποίησης, δηλαδή την within και την between, ενώ η c είναι μια σταθερά που, συνήθως, είναι η λογαριθμισμένη τιμή μέσου ή διάμεσου. Αυτή η

κανονικοποίηση εφαρμόζεται και ανάμεσα σε replicates. Την οπτικοποιούμε σε θηκογράμματα, που παρουσιάζουν πολύ βολικά τα αποτελέσματα.

$$\log_2\left(\frac{R_{i,j}}{G_{i,j}}\right) \rightarrow \log_2\left(\frac{R_{i,j}}{G_{i,j}}\right) - c = \log_2\left(\frac{R_{i,j}}{k \cdot G_{i,j}}\right)$$

Εξίσωση 7. Κανονικοποίηση κλίμακας.

Lowess (LOcally WEighted Scatterplot Smoothing) και Loess. Η lowess είναι μια γραμμική παλινδρόμηση τοπικών βαρών (121, 122), η οποία θεωρεί δεδομένο ότι η ποσότητα του αρχικού μεταγράφου είναι ίδια σε όλα τα δείγματα και ότι υπάρχει σχετική ισοεμφάνιση υπερεκφραζόμενων και υποεκφραζόμενων γονιδίων. Η κανονικοποίηση lowess αφορά σε αλλαγή της δομής των δεδομένων. Ενώ θεωρεί ότι υβριδοποιούνται ισάριθμα τα νουκλεϊνικά μόρια στα δείγματα και ότι η κάθε φθορίζουσα ουσία παρουσιάζει ίδια συνολική φωτεινότητα. Είναι μια μέθοδος για την ομαλοποίηση του διαγράμματος διασποράς (scatterplot) στις οποίες η διορθωμένη τιμή x_k είναι η τιμή μιας προσαρμοσμένης (fitted) γραμμής στα δεδομένα μέσω της μεθόδου των ελαχίστων τετραγώνων με βάρη, τα οποία για το σημείο (x_i , y_i) είναι μεγάλο αν το x_i βρίσκεται κοντά στο x_k και μικρό αν το x_i βρίσκεται κοντά στο x_k, ενώ οι ακραίες τιμές απορρίπτονται, δηλαδή αυτήν η κανονικοποίηση χρησιμοποιείται για να αλλάξουμε την δομή των δεδομένων. Η κανονικοποίηση lowess είναι πολύ παρόμοια με την loess και οι δύο χρησιμοποιούνται για την αφαίρεση των στατιστικών προκαταλήψεων λόγω φωτεινότητας, απλά η μια είναι γραμμική (y=a+bx) και η άλλη δευτεροβάθμια (y=a+bx +cx^2 quadratic), έτσι ώστε να αποφύγει τους επιπλέον χειρισμούς (123). Στο λογισμικό MATLAB εκτελείται με την εντολή *malowess*, ενώ στο λογισμικό R/Bioconductor εκτελείται από το πακέτο affy με τις εντολές *normalize.loess* και *normalize.AffyBatch.loess*.

M-A plot. Η κανονικοποίηση Lowess, για δικάναλο πείραμα μικροσυστοιχιών DNA ή για δύο πειράματα μονοκάναλων μικροσυστοιχιών, είναι σύνηθες να οπτικοποιείται στο διάγραμμα M-A plot, που ονομάζεται και R-I. Στον y-άξονα του διαγράμματος βρίσκεται η συνάρτηση M (**Εξίσωση 8**), δηλαδή τον λόγο της έντασης, που κανονικοποιείται με την αφαίρεση της αντίστοιχης τιμής από την καμπύλη της προσαρμοσμένης (fitted) lowess σε σχέση με τις τιμές της συνάρτησης A του x-άξονα, δηλαδή την μέση ένταση για αυτό το σημείο του διαγράμματος (**Εξίσωση 9**). Όταν οι τιμές του M είναι ομοιόμορφα κατανεμημένες γύρω από το 0 κατά μήκος όλων των A τιμών έντασης έχουμε καλά επίπεδα υβριδοποίησης. Στο λογισμικό MATLAB εκτελείται με την εντολή *mairplot*, ενώ στο λογισμικό R/Bioconductor εκτελείται από το πακέτο limma με την εντολή *plotMA*.

$$\mathrm{M} = \log_2\left(\frac{R_{i,j}}{G_{i,j}}\right) = \log_2(R_{i,j}) - \log_2(G_{i,j})$$

Εξίσωση 8. Η εξίσωση M για την οπτικοποίηση των τιμών φωτεινότητας σε MA Plot..

$$A = \frac{\log_2(R_{i,j} \cdot G_{i,j})}{2} = \frac{\log_2\left(R_{i,j}\right) + \log_2\left(G_{i,j}\right)}{2}$$

Εξίσωση 9. Η εξίσωση A για την οπτικοποίηση των τιμών φωτεινότητας σε MA Plot.

Rank invariant. Εφαρμόζουμε την μέθοδο σταθερής ταξινόμησης σε δικάναλα πειράματα, την οποία βρίσκουμε στο συμπληρωματικό υλικό της δημοσίευσης του Tseng και των συνεργατών του (124), αναζητούμε ποια γονίδια παραμένουν σταθερά και χρησιμοποιούμε αυτά στην καμπύλη κανονικοποίησης, δηλαδή αν έχουμε παρόμοια ταξινόμηση των κόκκινων και πράσινων εντάσεων ενός γονιδίου και αν η ταξινόμηση της μέσης έντασης των δύο καναλιών δεν είναι υπερβολικά υψηλή ή χαμηλή, τότε θεωρούμε ότι το γονίδιο δεν είναι διαφορικά εκφραζόμενο στα δύο δείγματα. Είναι πιο χρήσιμη όταν περιμένουμε υψηλή διαφορική έκφραση, π.χ. σε καρκινικούς ιστούς. Στο λογισμικό MATLAB εκτελείται με την εντολή *mainvarsetnorm*, ενώ στο λογισμικό R/Bioconductor εκτελείται από το πακέτο preprocessCore με την εντολή *normalize.invariantset*.

Quantile normalization. Στο λογισμικό MATLAB ο μετασχηματισμός ποσοστημορίων γίνεται με την εντολή quantilenorm ενώ στο λογισμικό R/Bioconductor, και συγκεκριμένα στο πακέτο preprocessCore, εκτελείται με τις εντολές *normalize.quantiles*, *normalize.quantiles.robust* και *normalize.AffyBatch.quantiles*. Ο μετασχηματισμός ποσοστημορίων στα μονοκάναλα πειράματα ακολουθεί τον εξής αλγόριθμο:

1) Πρώτα, ταξινομούμε τις λογαριθμισμένες τιμές φωτεινότητας κάθε συστοιχίας.

2) Στη συνέχεια, βρίσκουμε το μέσο όρο αυτών των τιμών για κάθε γονίδιο, έτσι ώστε να υπολογίσουμε την έκφρασης τους σε όλες τις συστοιχίες.

3) Στο τέλος, προσδιορίζουμε την θέση των εξεταζόμενων τιμών με βάση την βαθμονόμηση (ranking) του κάθε ανιχνευτή, από τον υψηλότερο στον χαμηλότερο.

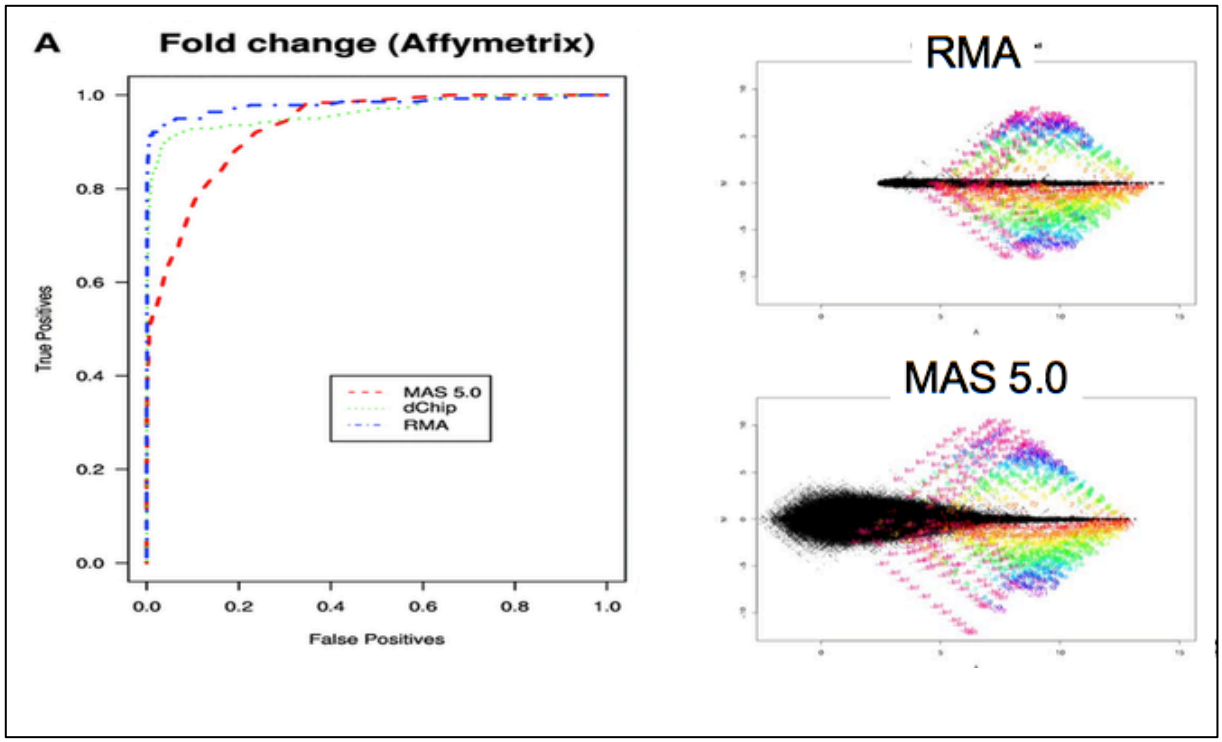
Στην δημοσίευση (125) παρουσιάζεται εκτενέστερη ανάλυση του μετασχηματισμού ποσοστημορίων.

Quantile global median. Ο ολικός μετασχηματισμός ποσοστημορίων διαμέσου τιμής ανήκει στις μεθόδους ολικής κανονικοποίησης πολλαπλών δειγμάτων και είναι μια ικανοποιητική και γρήγορη μέθοδος αφαίρεσης τεχνικού θορύβου από τα πρωτογενή δεδομένα. Ουσιαστικά

κανονικοποιούμε ώστε τα ποσοστημόρια να είναι ισόποσα σε κάθε τσιπ. Η διάμεσος τιμή προτιμάται από την μέση τιμή, αφού είναι λιγότερο επιρρεπές στην παρουσία ακραίων τιμών, και, μετά την εφαρμογή της μεθόδου έχουμε ίδια διάμεσο τιμή (ίση με 0) και ίδια τυπική απόκλιση, σε κάθε μικροσυστοιχία. Η ολική κανονικοποίηση των επιπέδων γονιδιακής έκφρασης αναφέρεται στο ότι αυτή επηρεάζεται από δύο παράγοντες, δηλαδή την γονιδιακή παραγωγή και την ειδικό θόρυβο σε κάθε πλακίδιο.

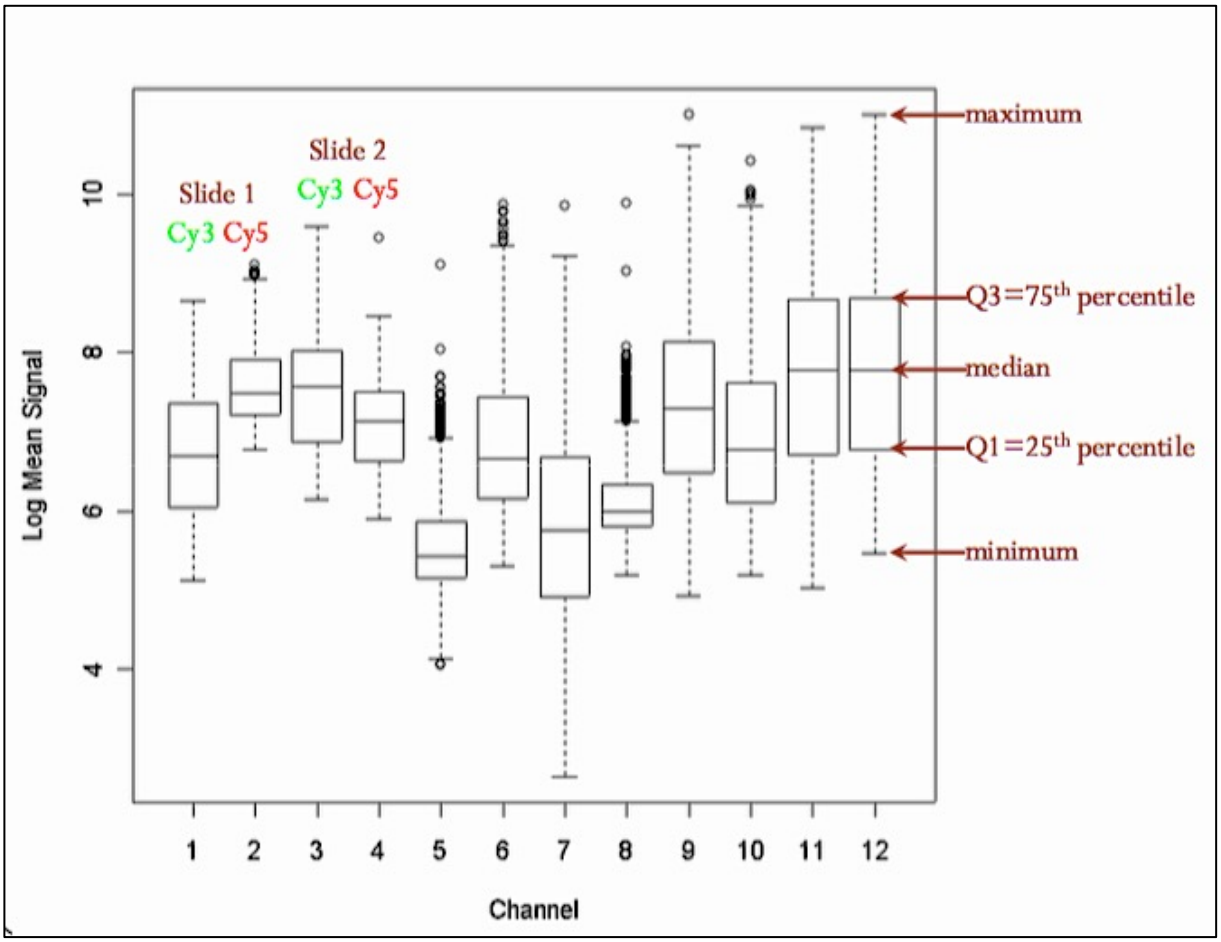
RMA (Robust Multichip Average). Η μέθοδος εύρωστης ανάλυσης του μέσου πολυ-συστοιχιών (RMA) βασίζεται στην συνέλιξη και την εφαρμόζουμε σε αρχεία (**Εικόνα 18**). CEL που προέρχονται από τις ολιγονουκλεοτιδικές μικροσυστοιχίες της Affymetrix και έχουμε τα κονονικοποιημένα αρχεία .CHP. Είναι αποτελεσματική για την μείωση της μεταβλητότητας των μεταγράφων χαμηλής συγκέντρωσης. Λειτουργεί με την ποσοτικοποίηση της φωτεινότητας των θέσεων ανιχνευτή PM, υπολογίζοντας το άθροισμα από τα εικονοστοιχεία ανεξάρτητης τυχαίας μεταβλητής για το υπόβαθρο και τον φθορισμό, δηλαδή μιας κανονικά κατανεμημένης και, αντίστοιχα, μιας εκθετικά κατανεμημένης μεταβλητής. Εχει καλή διακριτική ικανότητα ανάμεσα σε διαφορικά εκφραζόμενα γονίδια και μη. Στο λογισμικό MATLAB εκτελείται με τις εντολές *affyrma*, *rmabackadj* και *rmasummary* ενώ στο λογισμικό R/Bioconductor, και συγκεκριμένα στο πακέτο affy, γίνεται με την εντολή *rma*. Το λογισμικό RMA Express[13] χρησιμοποιείται για την διόρθωση υποβάθρου, την κανονικοποίηση ποσοστημορίου και την μέθοδο σύνοψης (summarization) των ολιγονουκλεοτιδικών τσίπ της Affymetrix (126).

[13] https://rmaexpress.bmbolstad.com/



Εικόνα 18. Η RMA έχει καλύτερα αποτελέσματα στην ανίχνευση διαφορικής έκφρασης στην πράξη από την MAS 5.0, ενώ με την RMA δεν παρατηρείται αυξημένος θόρυβος στις χαμηλές φωτεινότητες (127).

Microarray Suite 5 (MAS 5.0). Η συγκεκριμένη μέθοδος, έκδοσης 5.0, δημιουργήθηκε από την ίδια την εταιρία Affymetrix για την προεπεξεργασία δεδομένων από τα ολιγονουκλεοτιδικά της τσίπ. Στο λογισμικό R/Bioconductor, και συγκεκριμένα στο πακέτο affy, εκτελείται με την εντολή *mas5*. Σε κάποιες μελέτες θεωρείται καλύτερη μέθοδος από την RMA (128), όπως φαίνεται από την σύγκριση στην **Εικόνα 18**.



Εικόνα 19. Θηκόγραμμα πριν την εφαρμογή της κανονικοποίησης και το κεντράρισμα των κουτιών στο 0 (129).

Θηκόγραμμα. Για να είναι πιο κατανοητές οι πληροφορίες αυτών των μεθόδων κανονικοποίησης μπορούμε να τις οπτικοποιήσουμε και να τις συγκρίνουμε, αν έχουμε πάνω από ένα δείγμα, με την μέθοδο των θηκογραμμάτων (boxplots). Από την **Εικόνα 19** φαίνεται ότι η κεντρική γραμμή δείχνει τη διάμεσο (Q2=50%) και τα όρια του κουτιού τοποθετούνται στα δύο τεταρτημόρια (Q1=25%, Q3=75%). Οι πιό ακραίες τιμές παρουσιάζονται ως κουκκίδες και είναι μακριά από το κουτί. Με τα θηκογράμματα μπορούμε να ελέγξουμε εκτός από τις ακραίες τιμές και την συμμετρικότητα της κατανομής της διαμέσου και των ποσοστιαίων τεταρτημορίων. Αν η κεντρική γραμμή είναι κάτω από το μέσο του κουτιού, η κατανομή παρουσιάζει θετική ασυμμετρία και αν είναι πάνω από το μέσο του κουτιού, η κατανομή παρουσιάζει αρνητική ασυμμετρία.

Συμπλήρωση ελλιπών τιμών (Missing Values). Είναι σύνηθες να μην έχουν συμπεριληφθεί όλες οι τιμές βιολογικού ενδιαφέροντος στα εξαγόμενα σύνολα δεδομένων γονιδιακής έκφρασης. Υπάρχουν διάφοροι πιθανοί λόγοι αυτής της έλλειψης, όπως τα τεχνικά σφάλματα

των πλακιδίων της μικροσυστοιχίας, οι εκλιπούσες τιμές λόγω φιλτραρίσματος, τα σφάλματα κατά την εικονοληψία ή και λόγω κακής διακριτικής ικανότητας του συστήματος ανάλυσης. Οι πρώτες προσεγγίσεις είχαν σκοπό την αναγνώριση των περιοχών ελλιπών τιμών και, στην συνέχεια, η αντικατάστασή τους από μηδενικά ή με την μέση έκφραση σε σχέση με τις στήλες των δειγμάτων (130). Μια καλύτερη λύση είναι τα replicates, που μπορούν να χρησιμοποιηθούν για την συμπλήρωση ελλιπών τιμών. Ακόμα, στην δημοσίευση των ερευνητών (131), προτείνονται μέθοδοι βασισμένοι στις μηχανών υποστήριξης διανυσμάτων (Support vector machines, SVM) και η ταξινόμησης k-κοντινότερων γειτόνων (k-Nearest Neighbor, kNN), για την συμπλήρωση εκλιπών τιμών. Αυτές είναι τεχνικές ταξινόμησης που θα αναφερθούμε στην ενότητα 3.3.2.

3.2. Εξόρυξη Δεδομένων: Μαθηματικό Σκέλος

Η Εξόρυξη Δεδομένων (Data Mining) αποτελείται από διάφορα βήματα. Τα βασικά είναι ο εντοπισμός Διαφορικής Έκφρασης Γονιδίων (Differentially Expressed Genes, DEG), η Οπτικοποίηση (Visualisation) τους, η Ομαδοποίηση/Συσταδοποίηση (Clustering) τους και η Ταξινόμηση (Classification) τους.

3.2.1. Ο Εντοπισμός των Επιπέδων της Διαφορικής Έκφρασης με Στατιστικές Μεθόδους

Μετά το πειραματικό κομμάτι των μικροσυστοιχιών γίνεται καθαρισμός των δεδομένων και σύγκριση με το δείγμα αναφοράς (121). Αναγνωρίζεται η εικόνα και μετατρέπονται τα δεδομένα μικροσυστοιχιών σε αρχεία EXCEL που αποτελούνται από αριθμητικές λίστες. Έχουμε στην οριζόντια γραμμή του πίνακα τα διάφορα δείγματα υπό εξέταση, συμπεριλαμβανομένου του δείγματος αναφοράς, ενώ στην κατακόρυφη γραμμή του πίνακα τα χιλιάδες γονίδια που μας ενδιαφέρουν. Αρχικά, βρίσκουμε το μέσο όρο για κάθε γονίδιο ξεχωριστά. Στη συνέχεια, γίνεται η εκτίμηση των δεδομένων με διάφορους στατιστικούς ελέγχους και η επιλογή με τις τιμές p-value για επιβεβαίωση των επιπέδων διαφορικής έκφρασης, έτσι ώστε να επεξεργαστούμε και να αξιολογήσουμε σωστά τα δεδομένα. Οι παραμετρικοί έλεγχοι συνήθως εξετάζουν τυχαία δείγματα ενός φυσιολογικού πληθυσμού, ενώ στους μη παραμετρικούς ελέγχους χρησιμοποιείται μια μαθηματική εξίσωση για τον υπολογισμό της κατανομής υπό συγκεκριμένες συνθήκες. Οι στατιστικοί έλεγχοι είναι βασικό κομμάτι της μεταγενέστερης ανάλυσης (downstream analysis), όπου όσο μεγαλύτερο είναι το υπολογιστικό αποτέλεσμα της στατιστικής εξίσωσης, τόσο πιο πιθανό είναι να δεχτούμε την εναλλακτική υπόθεση (122).

Στην στατιστική διαμορφώνουμε υποθέσεις για να εξετάσουμε τα δεδομένα. Για να πάρουμε μια στατιστική απόφαση, αρχίζουμε πάντα με μια υπόθεση, η οποία ονομάζεται μηδενική υπόθεση και συμβολίζεται με H_0. Αν μια παραδοχή είναι ασυμβίβαστη με την H_0, τότε ονομάζουμε την υπόθεση εναλλακτική και την αριθμούμε αναλόγως, δηλαδή $H_1,...,H_m$. Για σύγκριση ανάμεσα σε δυο συνθήκες με άγνωστη όμοια διακύμανση μπορούμε να χρησιμοποιήσουμε τον έλεγχο του μαθητή (Student's t-Test), αλλά υπάρχουν πολλές άλλες εναλλακτικές ανάλογα με το ποιος στατιστικός έλεγχος ταιριάζει στις παρατηρήσεις υπό εξέταση. Πρώτα, βρίσκουμε την μέση τιμή και εφαρμόζουμε την επιλεγμένη στατιστική μέθοδο. Στην συνέχεια, συγκρίνουμε τα αποτελέσματα με την τιμή αναφοράς, p-value, και έτσι αποφασίζουμε αν θα απορριφθεί ή όχι η H_0. Συγκεκριμένα για το Student's t-Test υπολογίζουμε την πιθανότητα δύο κατανομών να προέρχονται από τον ίδιο δειγματικό χώρο και κατανομή. Αν έχουμε άνιση διακύμανση μεταξύ των τιμών της ομάδας φυσιολογικών δειγμάτων και των τιμών της ομάδας των υπό εξέταση δειγμάτων χρησιμοποιούμε μια άλλη εκδοχή του t-Test, το Welch's t-Test (133). Αν πρόκειται για ανάλυση δύο ομάδων προτιμάμε αυτά τα δύο t-Test, όμως για τρεις και παραπάνω ομάδες προτιμάμε τον έλεγχο ανάλυσης διακύμανσης ANOVA (Analysis of Variations) και τον Welch's ANOVA, οι οποίοι είναι επίσης παραμετρικοί έλεγχοι. Αν έχουμε πέντε και παραπάνω βιολογικά replicates μιας εξαρτημένης μεταβλητής, δηλαδή γονιδίου, τότε προτιμάμε μη παραμετρικές μεθόδους.

Μας ενδιαφέρει ο τύπος των παρατηρήσεων που εξετάζουμε, δηλαδή αν είναι paired (εξαρτημένων), όπως τα time course πειράματα, ή unpaired (ανεξάρτητων), όπως η σύγκριση ενός φυσιολογικού με άλλο υπό εξέταση δείγμα για εύρεση διαφορικά εκφρασμένου γονιδίου. Η ανεξαρτησία των παρατηρήσεων πρέπει να ισχύει και εντός μιας ομάδας. Στην περίπτωση των εξαρτημένων πειραμάτων δεν μπορούμε να τα αναλύσουμε ξεχωριστά, όπως ένα πείραμα που γίνεται πριν και μετά από μια θεραπεία.

3.2.1.1. Περιγραφική Στατιστική

Η στατιστική ανάλυση είναι σημαντική για να διερευνήσουμε πληροφορίες βιολογικής σημασίας, όπως την ποσοτική περιγραφή γενωμικών χαρακτηριστικών, γενικά μας ενδιαφέρει η ποσοτικοποίηση των τιμών των διαφορικά εκφραζόμενων γονιδίων και πως επηρεάζονται από διάφορους παράγοντες (83).

Κανονική Κατανομή (Normal Distribution). Στη γενωμική ανάλυση μικροσυστοιχιών μας απασχολεί πολύ οι κανονική κατανομή, η οποία χαρακτηρίζεται από την συνάρτηση του

Gauss (**Εξίσωση 10**) και, για συνεχείς μεταβλητές ή αν συνδέσεις όλα τα σημεία της κατανομής σχετικής συχνότητας ή ποσοστού, χαρακτηρίζεται από μια συμμετρική καμπύλη, που θυμίζει το σχήμα κουδουνιού (Bell Curve). Ένας τρόπος που η κατανομή μπορεί να διαφοροποιείται είναι η ασυμμετρία (skewness), η οποία μπορεί να είναι θετική ή αρνητική. Ένας άλλος τρόπος που η κατανομή μπορεί να διαφοροποιείται είναι το εύρος της κορυφή, δηλαδή ποια είναι η κύρτωσή (kurtosis) της.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)^2}$$

Εξίσωση 10. Η γκαουσιανή (Gauss) συνάρτηση, για x πραγματικό αριθμό.

Μέση τιμή (mean). Η μέση τιμή ή ο αριθμητικός μέσος του πληθυσμού μ (θεωρητική τιμή) και του δείγματος $\overline{x}$ (εμπειρική τιμή) χρησιμοποιείται συνήθως για να υπολογιστεί η κεντρική τάση συνεχών ή διακριτών τιμών. Αν η κατανομή είναι θεωρητική ονομάζεται αναμενόμενη *E*(*X*), δηλαδή είναι η θεωρητική μέση τιμή που θα περιμέναμε αν σχεδιάζαμε άπειρα σημεία από αυτή την κατανομή. Εκτός από την κανονική κατανομή *N*(0,1) (*E*(*X*)=0), ξέρουμε την αναμενόμενη τιμή και για άλλες γνωστές κατανομές, όπως η διωνυμική ($E(X)=N_p$), η υπεργεωμετρική (*N*,*M*,*n*) (*E*(*X*)=*n*(*M*/*N*)) και η εκθετική με ρυθμό λ(*E*(*X*)=1/λ).

$$\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Εξίσωση 11. Η μέση τιμή για *n* διαφορετικές τιμές, όπου το *μ* είναι ο μέσος του πληθυσμού και το $\overline{x}$ είναι ο μέσος του δείγματος.

$$\mathrm{E}(X) = \int_{-\infty}^{+\infty} x \cdot f(x)\,dx$$

Εξίσωση 12. Προσδοκώμενη τιμή για συνεχή τυχαία μεταβλητή *X*.

Τιμή διαμέσου (median). Μια άλλη τιμή υπολογισμού κεντρικής τάσης είναι η τιμή διαμέσου. Προτιμάμε να την χρησιμοποιούμε όταν έχουμε πολλές ακραίες τιμές στο δείγμα μας, αφού αντίθετα με την μέση τιμή, οι παρατηρήσεις είναι διατεταγμένες (ordinal) και είναι ισόποσες εκατέρωθεν της διαμέσου. Για δείγμα με περιττές παρατηρήσεις είναι η μεσαία τιμή και για δείγμα με άρτιες παρατηρήσεις είναι ο μέσος όρος των δύο μεσαίων τιμών.

$$Md = x_{\frac{n+1}{2}}$$ **Εξίσωση 13**. Η διάμεσος (M_d) για n περιττό.

$$Md = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$ **Εξίσωση 14**. Η διάμεσος (M_d) για n άρτιο.

Επικρατούσα τιμή (mode). Η παρατήρηση που εμφανίζεται πιο συχνά.

Διακύμανση ή Διασπορά (variance). Βρίσκουμε την πιο χρήσιμη στατιστική πληροφορία για τον υπολογισμό της διασποράς μιας κατανομής παίρνοντας κάθε σημείο και υπολογίζουμε την απόκλισή τους από το μέσο, το οποίο τετραγωνίζουμε και μετράμε τον μέσο τους. Αυτό το αποτέλεσμα είναι η διακύμανση και συμβολίζεται ως σ^2 (θεωρητική τιμή) ή s^2 (εμπειρική τιμή).

$$s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2$$ **Εξίσωση 15**. Υπολογισμός της διακύμανσης.

Συνδιακύμανση. Η συνδιακύμανση μεταξύ του X και Y συμβολίζεται με Cov(X,Y) ή S_{XY}. Είναι η τιμή Αν οι συνδιακυμάνσεις είναι θετικές, τότε οι δύο μεταβλητές είναι θετικά συσχετισμένες. Αν οι συνδιακυμάνσεις είναι αρνητικές, τότε οι δύο μεταβλητές είναι αρνητικά συσχετισμένες.

$$Cov(X,Y) = E\left[(X - \mu_X) \cdot (Y - \mu_Y)\right]$$ **Εξίσωση 16**. Υπολογισμός της συνδιακύμανσης.

Τυπική Απόκλιση (standard deviation, SD). Είναι η τετραγωνική ρίζα της διακύμανσης και δείχνει πόσο καλά αντιπροσωπεύεται το δείγμα στην καμπύλη, ενώ συμβολίζεται ως σ (εμπειρική τιμή) ή s (θεωρητική τιμή).

$$s = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2}$$ **Εξίσωση 17**. Η τυπική απόκλιση για μικρές τιμές n.

$$s = \sqrt{\frac{1}{n-1} \cdot \left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} \right]}$$

Εξίσωση 18. Η τυπική απόκλιση για μεγάλες τιμές n.

Βαθμοί ελευθερίας (degrees of freedom). Παρατηρούμε στις παραπάνω εξισώσεις για διακύμανση και τυπική απόκλιση το άθροισμα δεν διαιρείται με το n, αλλά με το n-1, δηλαδή τον αριθμό των παρατηρήσεων μείον μια. Ο αριθμός αυτός ονομάζεται βαθμός ελευθερίας (df) και αναπαριστά πόσες είναι οι ελάχιστες ανεξάρτητες πληροφορίες έτσι ώστε να προσδιοριστεί μια άλλη πληροφορία. Στις εξισώσεις παρουσιάζεται με το ελληνικό γράμμα ν. Είναι ο ίδιος για την σ^2 και την σ, αφού μόνο n-1 των τετραγωνισμένων αποκλίσεων μπορούν να μεταβάλλονται ελεύθερα, αλλά εξαρτάται από την τιμή που εξετάζουμε κάθε φορά.

IQR. Το εύρος ενδοποσοστημορίου (interquartile range, IQR) είναι η απόσταση μεταξύ του πρώτου και του τρίτου ποσοστημορίου και χρησιμοποιείται στην ανίχνευση ακραίων τιμών, ενώ είναι ένα μέτρο κεντρικότητας.

$$IQR = Q_3 - Q_1$$

Εξίσωση 19. Εύρος ενδοποσοστημορίου.

3.2.1.2. Στατιστικοί Έλεγχοι

Οι στατιστικοί έλεγχοι t-Test χρησιμοποιούνται με την παραδοχή ότι έχουμε φυσιολογικά δείγματα. Ακόμα και αν δεν έχουμε κανονική κατανομή, μπορούμε σε πολλές περιπτώσεις να κάνουμε αυτή την παραδοχή για μεσαίο ή μεγάλο πλήθος παρατηρήσεων, αφού τα σφάλματα τύπου I είναι σε αποδεκτά επίπεδα. Αυτό δεν ισχύει στην περίπτωση που έχουμε πολλές ακραίες τιμές, οπότε θα χρησιμοποιηθούν μη παραμετρικοί έλεγχοι. Οι μη παραμετρικοί έλεγχοι, όπως ο Mann-Whitney U και ο Kruskal-Wallis H, έχουν το πλεονέκτημα ότι είναι πιο εύκολα κατανοητοί, πιο γρήγοροι και πιο εύκολα εφαρμόσιμοι σε μικρά δείγματα.

Student's t-Test με 2 ανεξάρτητα δεδομένα. Ο στατιστικός έλεγχος του μαθητή (134) (ψευδώνυμο του W. S. Gosset σε αυτή την δημοσίευσή του) εφαρμόζεται συνήθως στην μέση ή την διάμεσο τιμή με την παραδοχή ότι ακολουθούμε κανονική κατανομή $N(\mu,\sigma^2)$, δηλαδή όταν συγκρίνουμε δύο δείγματα θεωρούμε ότι το δείγμα 1 έχει κατανομή $X_{1i} \sim N(\mu_1,\sigma^2)$ και το δείγμα 2 έχει κατανομή $X_{2i} \sim N(\mu_2,\sigma^2)$. Η σύγκριση των τιμών X_1 και X_2 μεταξύ των δυο

δειγμάτων εξετάζει το επίπεδο της ομοιότητας στις κατανομές των παρατηρήσεων ενός γονιδίου σε αυτά, δηλαδή να επικαλύπτονται για μια δεδομένη τιμή, ενώ όσο πιο μικρός είναι ο αριθμός που δίνει η εξίσωση *t*-Test, τόσο πιο κοντά είναι η τιμή μ_1 στην τιμή μ_2. Στην αντίθετη περίπτωση έχουμε διαφορική έκφραση. Μια άλλη παραδοχή του *t*-Test, είναι ότι οι δύο ομάδες έχουν σχετικά τον ίδιο αριθμό παρατηρήσεων. Παρόλο που το *t*-Test σχεδιάστηκε για εξέταση πληθυσμών μιας ομάδας δειγμάτων μέχρι 30 παρατηρήσεων, στην πράξη αποδεικνύεται ανθεκτικό και για εξέταση μεγαλύτερων πληθυσμών, αλλά δεν λειτουργεί καλά για ανάλυση πάνω από δύο ομάδων. Το *t*-Test θεωρείται ένα εύρωστο μέτρο ελέγχου, ακόμα και για μικρές αποκλίσεις έχει καλά αποτελέσματα. Υπάρχουν διάφορα μέτρα ποιότητας ελέγχου για την κανονικότητα, που χρησιμοποιούμε όταν θέλουμε να επιλέξουμε έναν έλεγχο. Κάποιοι τέτοιοι έλεγχοι είναι ο Anderson–Darling, ο Cramer–von Mises, ο Lilliefors (Kolmogorov–Smirnov), ο έλεγχος κανονικότητας Pearson chi-square και ο Shapiro–Francia (135).

Ακολουθεί ένα τυπικό μοντέλο στατιστικής σύγκρισης *t*-Test, όπως παρουσιάζεται στο βιβλίο των (136), για την σύγκριση των μέσων όρων δύο δειγμάτων, που αποτελούνται από ίσες παρατηρήσεις. Πρώτα θεωρούμε την μηδενική υπόθεση H_0 ότι $\mu_1=\mu_2$ και την εναλλακτική υπόθεση H_1 ότι $\mu_1 \neq \mu_2$.

$$SS_1 = \sum_i^{n_1} \left(X_{1i} - \mu \right)^2$$

Εξίσωση 20. Το άθροισμα τετραγώνων για το δείγμα 1, όπου το $1i \in (1,...,n_1)$ για τις τιμές X_{1i} .

$$SS_2 = \sum_i^{n_2} \left(X_{2i} - \mu \right)^2$$

Εξίσωση 21. Το άθροισμα τετραγώνων για το δείγμα 2, όπου το $2i \in (1,..., n_2)$ για τις τιμές X_{2i}.

$$s_p{}^2 = \frac{SS_1 + SS_2}{df_1 + df_2} = \frac{SS_1 + SS_2}{n_1 - 1 + n_2 - 1}$$

Εξίσωση 22. Η κοινή διακύμανση.

$$SEM = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

Εξίσωση 23. Το τυπικό σφάλμα του μέσου.

$$\mathrm{T} = \frac{\bar{X}_1 - \bar{X}_2}{SEM}$$

Εξίσωση 24. Υπολογίζουμε το στατιστικό *T* για να ανιχνεύσουμε DEGs.

Για να υπολογιστούν οι κρίσιμές τιμές και οι τιμές p-value θεωρούμε ότι για την H_0 ισχύει $T \sim t_{n1\text{-}n2\text{-}2}$ με α=0,05. Μετά γίνεται η σύγκριση της κρίσιμης τιμής t_{crit}, που βρήκαμε από τον πίνακα του μαθητή, με την τιμή που υπολογίσαμε στατιστικά t_{stat}. Αν η στατιστική τιμή είναι μεγαλύτερη από την κρίσιμη, τότε απορρίπτουμε την H_0 και θεωρούμε ότι έχουμε διαφορική έκφραση γονιδίου.

Welch's t-Test με 2 ανεξάρτητα δεδομένα. Αυτή η περίπτωση είναι παρόμοια με τον έλεγχο του μαθητή, με την διαφορά ότι δεν έχουμε ίδια διακύμανση, δηλαδή θεωρούμε ότι το δείγμα 1 έχει κατανομή $X_{1i} \sim N(\mu_1,\sigma_1^2)$ και το δείγμα 2 έχει κατανομή $X_{2i} \sim N(\mu_2,\sigma_2^2)$. Αυτή η διαφορά αλλάζει η εξίσωση των βαθμών ελευθερίας (d.f.) και η στατιστική εξίσωση t-Test, η οποία δεν έχει πλέον μια ακριβή κατανομή t, αλλά κατά προσέγγιση έχουμε καλά αποτελέσματα με την εξίσωση Welch–Satterthwaite (137, 138).

$$\mathrm{T} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Εξίσωση 25. Welch's *t*-Test.

$$s_1 = \frac{1}{n_1 - 1} \cdot \sum_i^{n_1} \left(X_{1i} - \bar{X}_1 \right)^2$$

Εξίσωση 26. Η αμερόληπτη εξίσωση της διακύμανσης για το δείγμα 1.

$$s_2 = \frac{1}{n_2 - 1} \cdot \sum_i^{n_2} \left(X_{2i} - \bar{X}_2 \right)^2$$

Εξίσωση 27. Η αμερόληπτη εξίσωση της διακύμανσης για το δείγμα 2.

$$\nu \approx \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{s_1^4}{n_1^2 \cdot (n_1 - 1)} + \frac{s_2^4}{n_2^2 \cdot (n_2 - 1)}}$$

Εξίσωση 28. Οι d.f. της εξίσωσης Welch–Satterthwaite.

Student's t-Test με εξαρτημένα/συσχετισμένα δεδομένα (Paired/Matched t-Test). Συχνά δύο ή περισσότεροι πληθυσμοί που εξετάζουμε μπορεί να μην είναι ανεξάρτητοι, π.χ. σε ένα time course experiment. Κάποια παραδείγματα τέτοιων ελέγχων είναι όταν ελέγχουμε πώς ένα φάρμακο επιδρά σε μια μετρήσιμη τιμή ή πως δυο διαφορετικά φάρμακα επιδρούν στον ίδιο οργανισμό ή ακόμα και να εξετάσουμε δείγματα από διαφορετικά άτομα που έχουν ακολουθήσει άλλη φαρμακευτική αγωγή και να τους συγκρίνουμε σύμφωνα με ηλικία ή άλλα χαρακτηριστικά. Έτσι, θα έχουμε ζευγάρια μετρήσεων, αφού η μια μέτρηση εξαρτάται από την άλλη. θεωρούμε ότι το δείγμα 1 έχει μετρήσεις $X_1,\ldots,X_n$ με μέσο μ_1 και το δείγμα 2 έχει μετρήσεις $Y_1,\ldots,Y_n$ με μέσο μ_2. Αφού δεν έχουμε ανεξάρτητες μεταβλητές δεν μπορούμε να χρησιμοποιήσουμε το κλασικό *t*-Test, οπότε θέτουμε μια μεταβλητή $U_i = Y_i - X_i$ με μέσο πληθυσμού μ_u. Τότε, θεωρούμε την μηδενική υπόθεση $H_0: \mu_u = 0$ και την εναλλακτική υπόθεση $H_1: \mu_u < 0$, η οποία μπορεί να εξεταστεί με το κλασικό *t*-test για ένα πληθυσμό, με μετρήσεις $U_1,\ldots,U_n$ (136).

$$T = \frac{\sqrt{n} \cdot \left(\bar{U} - 0\right)}{s}$$

Εξίσωση 29. Η στατιστική εξίσωση *t* ενός πληθυσμού, για μέσο 0.

$$s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(U_i - \bar{U}\right)$$

Εξίσωση 30. Η διακύμανση του πληθυσμού με βαθμούς ελευθερίας d.f.=*n*-1.

Μετά γίνεται η σύγκριση της κρίσιμης τιμής t_{crit} , που βρήκαμε από τον πίνακα του μαθητή, με την τιμή που υπολογίσαμε στατιστικά t_{stat}. Αν η στατιστική τιμή είναι μεγαλύτερη από την κρίσιμη, τότε απορρίπτουμε την H_0 και θεωρούμε ότι έχουμε διαφορική έκφραση γονιδίου.

Ανάλυση Διακύμανσης ή Διασποράς (ANalysis Of VAriations, ANOVA). Η ανάλυση διακύμανσης μπορεί να χρησιμοποιηθεί για την ανίχνευση DEGs ανάμεσα στις παρατηρήσεις

πάνω από δύο δειγμάτων. Είναι ένας παραμετρικός έλεγχος, οπότε μια βασική παραδοχή είναι ότι ακολουθούμε κανονική κατανομή και ότι κάθε πληθυσμός αποτελείται από ίσο πλήθος παρατηρήσεων. Προτιμάμε τον έλεγχο ANOVA για σύγκριση σε τρία δείγματα ή περισσότερα από το να κάνουμε πολλαπλά διαδοχικά *t*-Test, τα οποία αυξάνουν το ποσοστό των σφαλμάτων. Η ανάλυση διακύμανσης αφορά την εκτίμηση ποσοτικοποίησης των επιμέρους συνιστωσών της ολικής μεταβλητότητας, που υπολογίζεται σαν συνολικό άθροισμα τετραγώνων (Total Sum of Squares, SS_T), όπως είναι η εσωτερική διακύμανση κάθε πειράματος (Within groups Sum of Squares, SS_W), η διακύμανση που υπάρχει μεταξύ των δειγμάτων (Between groups Sum of Squares, SS_B) και η διακύμανση λόγω σφαλμάτων (Sum of Squared Errors, SS_E). Η εσωτερική διακύμανση κάθε πειράματος SS_W αναλύει την μεταβλητότητα η οποία δεν εξηγείται από τις διαφορές μεταξύ των μέσων των δειγμάτων. Η ANOVA μπορεί να είναι σύγκριση ανάμεσα στις μέσες τιμές τριών ή περισσότερων δειγμάτων κατά ένα παράγοντα (One way), κατά δύο παράγοντες (Two way), κατά τρεις παράγοντες (Three way) κ.λ.π. . Ο έλεγχος ANOVA κατά ένα παράγοντα χρησιμοποιείται για να αναγνωρίσει τις στατιστικά σημαντικές διαφορές για ένα χαρακτηριστικό ανάμεσα σε τρία ή περισσότερα δείγματα. Το μειονέκτημα της ANOVA είναι ότι δεν μπορεί να κατονομάσει ποια δείγματα εμφανίζουν στατιστική σημαντικά διαφορά, αυτό μπορεί να γίνει με έναν έλεγχο post hoc, απλώς ότι τουλάχιστον δύο δείγματα την εμφανίζουν. Ο έλεγχος ANOVA κατά δύο παράγοντες χρησιμοποιείται για να αναγνωρίσει τις στατιστικά σημαντικές διαφορές παράλληλα για δύο χαρακτηριστικά και την σχέση μεταξύ τους. Το πλεονέκτημα του είναι ότι αναπαριστά καλύτερα ένα πραγματικό ερευνητικό πείραμα, αφού οι μεταβλητές συνήθως επηρεάζονται από πάνω από ένα χαρακτηριστικό, οπότε χρησιμοποιείται στη γενωμική πιο συχνά από τον έλεγχο ANOVA κατά ένα παράγοντα (135).

Ο έλεγχος ANOVA του δείγματος σε διάφορες χρονικές στιγμές (time course), όπου για την κάθε μια υπάρχει και η αντίστοιχη κατάσταση αναφοράς, δηλαδή έχουμε ζευγάρια λογαριθμισμένων τιμών κατάστασης με αναφοράς. Το βασικό μέρος της ανάλυσης διακύμανσης είναι ο έλεγχος *F*.

$$F = \frac{\frac{SS_B}{\nu_B}}{\frac{SS_W}{\nu_B}} = \frac{\frac{SS_B}{k-1}}{\frac{SS_W}{n_1+n_2-k}}$$

Εξίσωση 31. Ο έλεγχος *F* για ανίχνευση DEGs, όπου k είναι το πλήθος των δειγμάτων, ενώ n_1 και n_2 είναι ο αριθμός των παρατηρήσεων της ομάδας 1 και 2, αντίστοιχα.

Ακολουθεί ένα τυπικό μοντέλο ανάλυσης διακύμανσης ANOVA κατά ένα παράγοντα, όπως παρουσιάζεται στο βιβλίο των (139), για την σύγκριση των μέσων όρων δύο δειγμάτων, που αποτελούνται από ίσες παρατηρήσεις. Αν έχουμε μεγάλη διαφορά στο SS_B σε σχέση με το SS_W, τότε υπάρχει διαφορική έκφραση γονιδίων, και αντιστρόφως.

$$s^2 = \frac{\sum_{i=1}^{n_1}(x_{i1}-\overline{x}_1)^2 + \sum_{i=1}^{n_2}(x_{i2}-\overline{x}_2)^2}{n_1+n_2-2} = \frac{SS_E}{n_1+n_2-2}$$

Εξίσωση 32. Εσωτερική διακύμανση (Within groups variation, SSW) κοινή για τις δύο ομάδες, όπου x_{i1} και x_{i2} είναι η *i*-οστή παρατήρηση στην ομάδα 1 και στην ομάδα 2, αντίστοιχα.

$$SS_B = n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2$$

Εξίσωση 33. Η διακύμανση που υπάρχει μεταξύ των δύο ομάδων, όπου $\overline{x}_1$ και $\overline{x}_2$ είναι η μέση τιμή των παρατηρήσεων του κάθε δείγματος (Between groups variation, SSB).

$$SS_E = \sum_{n=1}^{n}\left(x_i - \hat{x}_i \right)$$

Εξίσωση 34. Η διακύμανση λόγω σφαλμάτων, που ακολουθεί κανονική κατανομή με μέση τιμή 0.

$$SS_T = SS_W + SS_B = \sum_{i=1}^{n}(x_i - \overline{x})^2$$

Εξίσωση 35. Η ολική διακύμανση.

$$F = \frac{\frac{SS_B}{SS_W}}{n_1+n_2-2} = \frac{n_1(\overline{x}_1-\overline{x})^2 + n_2(\overline{x}_2-\overline{x})^2}{\frac{[\sum_{i=1}^{n}(x_i-\overline{x})^2] - n_1(\overline{x}_1-\overline{x})^2 - n_2(\overline{x}_2-\overline{x})^2}{n_1+n_2-2}}$$

Εξίσωση 36. Ο έλεγχος F για ανίχνευση DEGs.

Οι παράγοντες λέγονται σταθεροί (fixed), αν σε κάθε πείραμα έχουμε τα ίδια επίπεδα έκφρασής του, και τυχαίοι (random), αν έχουμε ένα δείγμα του πληθυσμού από τα πιθανά επίπεδα έκφρασης. Αυτό σημαίνει ότι για τυχαίους παράγοντες ενσωματώνεται ένα επιπλέον επίπεδο θορύβου στην διακύμανση του δείγματος. Χρησιμοποιούμε τυχαίο παράγοντα για να

μπορούμε να κάνουμε πιο επιτυχημένη γενίκευση, αφού σε αντίθετη περίπτωση, δηλαδή αν μια συστοιχία αναλύεται σαν σταθερού παράγοντα, θεωρούμε ότι έχουμε μια σταθερά θορύβου σε κάθε πείραμα, κάτι που δεν μπορεί να συμβεί στην πραγματικότητα σε κάθε παρτίδα (batch). Κατά τον υπολογισμό των σφαλμάτων όταν εκτελούμε στατιστικό έλεγχο ANOVA σε δεδομένα μικροσυστοιχιών, θεωρείται ότι τα σφάλματα είναι ομογενή. Αυτό είναι αδύνατο σε ένα πραγματικό πείραμα, αφού δεν μπορούμε να έχουμε το ίδιο σφάλμα μέτρησης σε όλους τους ανιχνευτές. Ακόμα, οι μετρήσεις καταλήγουν να βασίζονται στην FC, η οποία δεν έχει τόσο καλά αποτελέσματα. Αν από την άλλη, δεχόμασταν την ετερογένειά τους και ότι κάθε γονίδιο είχε το δικό του σφάλμα, τότε δεν θα είχαμε μια στατιστική μέθοδο με πολύ περιορισμένη ικανότητα να ανιχνεύσει σωστά τα DEGs. Χρησιμοποιούμε Μπεϋσιανή (Bayes) στατιστική για να εκτιμήσουμε την μετριασμένη διακύμανση, αφού μας επιτρέπει να εφαρμόσουμε συγκεκριμένες σχέσεις στις διακυμάνσεις (140).

Ακολουθεί ένα τυπικό μοντέλο ανάλυσης διακύμανσης ANOVA κατά δύο παράγοντες, όπως παρουσιάζεται στο βιβλίο των (139), για την σύγκριση των μέσων όρων δύο δειγμάτων, που αποτελούνται από ίσες παρατηρήσεις. Οι υπολογισμοί της ANOVA μπορεί να γίνουν αναλύοντας το πειραματικό μοντέλο με παλινδρόμηση μέσω υπολογιστή.

$$CM = \frac{\frac{(\sum_{i=1}^{n} x_i)^2}{n}}{} = \frac{(\sum_{i=1}^{n} x_i)^2}{a*b*r}$$

Εξίσωση 37. Διόρθωση για τον μέσο, όπου n το σύνολο των παρατηρήσεων, α είναι το πλήθος των επιπέδων του ανεξάρτητου παράγοντα 1, b είναι το πλήθος των επιπέδων του ανεξάρτητου παράγοντα 2 και r είναι το πλήθος των παρατηρήσεων για κάθε ζευγάρι επιπέδων ανεξάρτητων παραγόντων 1 και 2.

$$SS_T = \sum_{i=1}^{n} x_i^2 - CM = SS(A) + SS(B) + SS(AB) + SSE$$

Εξίσωση 38. Η ολική διακύμανση.

$$SS(A) = \frac{\sum_{i=1}^{a} A_i^2}{br} - CM$$

Εξίσωση 39. Άθροισμα τετραγώνων για τον παράγοντα της ανεξάρτητης μεταβλητής 1, όπου A_i είναι οι συνολικές παρατηρήσεις της ανεξάρτητης μεταβλητής 1 στο επίπεδο $i \in (1,2,...,\alpha)$.

$$SS(B) = \frac{\sum_{j=1}^{b} B_j^2}{ar} - CM$$

Εξίσωση 40. Άθροισμα τετραγώνων για τον παράγοντα της ανεξάρτητης μεταβλητής 2, όπου B_j είναι οι συνολικές παρατηρήσεις της ανεξάρτητης μεταβλητής 2 στο επίπεδο $j \in (1,2,...,b)$.

$$SS(AB) = \frac{\sum_{j=1}^{b}\sum_{i=1}^{a} AB_{ij}^{\ 2}}{r} - SS(A) - SS(B) - CM$$

Εξίσωση 41. Άθροισμα τετραγώνων για την αλληλεπίδραση των παρατηρήσεων της μεταβλητής A με της μεταβλητής B, όπου το AB_{ij} είναι οι συνολικές παρατηρήσεις στο επίπεδο $i \in (1,2,...,\alpha)$ της ανεξάρτητης μεταβλητής 1 και στο επίπεδο $j \in (1,2,...,b)$ της ανεξάρτητης μεταβλητής 2.

$$MSE = \frac{SS(AB) + SS_E}{n - a - b + 1}$$

Εξίσωση 42. Το μέσο τετραγωνικό σφάλμα (Mean Squared Error, MSE).

$$MS(AB) = \frac{SS(AB)}{(a-1)\cdot(b-1)}$$

Εξίσωση 43. Μέσο άθροισμα τετραγώνων, MS, για την αλληλεπίδραση των παρατηρήσεων της μεταβλητής A με της μεταβλητής B.

$$F = \frac{MS(AB)}{MSE}$$

Εξίσωση 44. Ο έλεγχος F για ανίχνευση DEGs.

Διαφορές παραμετρικών και μη παραμετρικών ελέγχων. Ο έλεγχος *t*-Test και ο έλεγχος ANOVA θεωρούν ότι οι μεταβλητές ακολουθούν κανονική κατανομή. Σε κάποιες περιπτώσεις οι παρατηρήσεις έχουν μεγάλες αποκλίσεις από την κανονική κατανομή και εφαρμόζονται οι μη παραμετρικές μέθοδοι. Ακόμα, με τους μη παραμετρικούς ελέγχους δεν απαιτούνται εκτιμήσεις των παραμέτρων των κατανομών. Η κλίμακα μέτρησης των παρατηρήσεων πρέπει να είναι διατεταγμένη (ordinal). Ο έλεγχος Mann-Whitney U είναι ένα μη παραμετρικό t-test για δύο ανεξάρτητα δείγματα. Ο έλεγχος Kruskal-Wallis H

αντιπροσωπεύει ένα μη παραμετρικό έλεγχο ANOVA με περισσότερα από δύο ανεξάρτητα δείγματα.

Mann-Whitney U Test ή Wilcoxon Rank Sum Test ή Mann-Whitney-Wilcoxon (MWW) με 2 ανεξάρτητα δείγματα. Ο έλεγχος Mann-Whitney U, αντιστοιχεί σε μη παραμετρική εκδοχή του *t*-Test, δηλαδή δεν είναι προϋπόθεση να έχουν κανονική κατανομή τα δείγματα, όταν έχω δύο ανεξάρτητα δείγματα. Η διαφορά είναι ότι ελέγχει την σχέση τάξεων μεταξύ των τιμών και όχι τις πραγματικές τιμές, όπως το *t*-Test. Η καλύτερη περίπτωση για την ταξινόμηση των επιπέδων έκφρασης δύο συνθηκών, είναι όλες οι αξίες μιας συνθήκης να ταξινομούνται υψηλότερα από την άλλη συνθήκη. Αφού όμως ο έλεγχος MWW δεν εξετάζει την διασπορά, η σημασία του αποτελέσματος περιορίζεται μόνο από τον αριθμό των επαναλήψεων για τις δύο συνθήκες. Για αυτό τον λόγο ο έλεγχος MWW, δεν δίνει σωστά αποτελέσματα για μικρό αριθμό επαναλήψεων και η κατανομή των p-value είναι σχετικά διάσπαρτη (141, 142). Η MWW εμφάνισε καλύτερα αποτελέσματα, όταν συγκρίθηκε, στην ακόλουθη δημοσίευση (132), με τους στατιστικούς ελέγχους *t*-Test 2 ανεξάρτητων δειγμάτων, t-Test 2 εξαρτημένων δειγμάτων και ANOVA. Η μέθοδος στατιστικής ανάλυσης με τον έλεγχο Mann-Whitney U ακολουθεί τον εξής αλγόριθμο:

1) Θεωρούμε μια μηδενική υπόθεση H_0, δηλαδή $\mu_1=\mu_2$, και μια εναλλακτική υπόθεση H_1, δηλαδή μ_1 διαφορετικό από το μ_2.

2) Ιεραρχούμε σε πίνακες όλες τις τιμές, από την μικρότερη στην μεγαλύτερη τιμή, και τις βαθμονομούμε, δηλαδή τους δίνουμε μια βαθμίδα.

3) Αθροίζουμε τις τιμές των βαθμίδων για κάθε δείγμα.

4)Υπολογίζουμε την διαφορά μεταξύ του αθροίσματος των βαθμίδων με την ελεγχοσυνάρτηση *U*, όπου θεωρούμε ότι έχουμε συμμετική κατανομή σε κάθε δείγμα.

5) Θέτουμε ένα κατώφλι, συνήθως το α=0,05 και σύμφωνα με αυτό βρίσκουμε στον σχετικό πίνακα την κρίσιμη τιμή U_{crit} και ελέγχουμε αν είναι πάνω ή όχι από την στατιστική τιμή που βρήκαμε U_{stat}. Αν αυτή η τιμή είναι μεγαλύτερη από την κρίσιμη τιμή, τότε απορρίπτουμε την μηδενική υπόθεση.

$$\mathrm{U}_1 = \mathrm{n}_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

Εξίσωση 45. Εφαρμογή του στατιστικού ελέγχου MWW στο δείγμα *A*, όπου R_1 είναι ο αριθμός της βαθμίδας και n_1 είναι το μέγεθος του δείγματος *A*.

$$\mathrm{U}_2 = \mathrm{n}_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = \mathrm{n}_1 n_2 - U_1$$

Εξίσωση 46. Εφαρμογή του στατιστικού ελέγχου MWW στο δείγμα *B*, όπου R_2 είναι ο αριθμός της βαθμίδας και n_2 είναι το μέγεθος του δείγματος *B*.

$$U = \min(U_1, U_2)$$ **Εξίσωση 47**. Ελεγχοσυνάρτηση U.

$$\mu_u = \frac{n_1 n_2}{2}$$ **Εξίσωση 48**. Υπολογίζουμε την μέση τιμή.

$$\sigma_u^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$ **Εξίσωση 49**. Υπολογίζουμε την διακύμανση.

Kruskal-Wallis H Test με πάνω από 2 ανεξάρτητα δείγματα. Ένας άλλος στατιστικός μη παραμετρικός έλεγχος είναι ο Kruskal-Wallis *H*, που αντιστοιχεί σε μη παραμετρική εκδοχή του ελέγχου ANOVA κατά ένα παράγοντα και εφαρμόζεται στην σύγκριση τριών ή παραπάνω πληθυσμών. Όμως, σε σχέση με τον έλεγχο ANOVA, έχει μικρότερη ισχύ και χρησιμοποιεί λιγότερη πληροφορία. Ο έλεγχος Kruskal–Wallis *H* εφαρμόζεται σε πλήρως τυχαιοποιημένα πειραματικά σχέδια και έχει καλύτερα αποτελέσματα όταν εξετάζουμε πέντε ή περισσότερα replicates. Η μέθοδος στατιστικής ανάλυσης για k ανεξάρτητα δείγματα (συνήθως τρία ή περισσότερα) με τον έλεγχο Kruskal–Wallis *H* ακολουθεί τον εξής αλγόριθμο:

1) Θεωρούμε μια μηδενική υπόθεση H_0 ότι όλα τα δείγματα προέρχονται από τον ίδιο πληθυσμό και μια εναλλακτική υπόθεση H_1 ότι τουλάχιστον ένα δείγμα προέρχεται από διαφορετικό πληθυσμό.
2) Ιεραρχούμε σε πίνακες όλες τις τιμές, από την μικρότερη στην μεγαλύτερη τιμή (ordinal), και τους δίνουμε βαθμίδες π.χ. από 1 ως *n*, αγνοώντας από ποιό δείγμα προέρχονται.
3) Αν υπάρχουν ίδιες τιμές (tie), τότε αρχικά τους δίνουμε τον αριθμό που θα είχαν αν δεν βρίσκονταν στην ίδια βαθμίδα και τελικά χρησιμοποιούμε τον μέσο των δύο δοσμένων αριθμών ως την βαθμίδα του.
4) Υπολογίζουμε την ελεγχοσυνάρτηση *H*, όπου θεωρούμε ότι έχουμε συμμετρική κατανομή μέσα σε κάθε δείγμα.
5) Θέτουμε ένα κατώφλι, συνήθως το α=0,05, η οποία είναι η κατάλληλη προσέγγιση όταν έχουμε τουλάχιστον πέντε παρατηρήσεις και σύμφωνα με τους πίνακες κατανομών

πιθανότητας χ^2 (*chi*-squared) και το d.f.=*k*-1, βρίσκουμε την κρίσιμη τιμή H_{crit} και την συγκρίνουμε με την στατιστική τιμή που υπολογίσαμε H_{stat}. Αν αυτή η τιμή είναι μεγαλύτερη από από την κρίσιμη τιμή, τότε απορρίπτουμε την μηδενική υπόθεση.

$$H = \frac{12}{n \cdot (n+1)} \cdot \sum \frac{R_i^2}{n_i} - 3 \cdot (n+1)$$

Εξίσωση 50. Η στατιστική εξίσωση Kruskal Wallis H, όπου R_i είναι το άθροισμα των βαθμίδων (ranks) των αντίστοιχων παρατηρήσεων (n_i), για κάθε δείγμα.

Ανάλυση Πολυμεταβλητής Διακύμανσης (Multivariate analysis of variance, MANOVA).
Πρόκειται για μια στατιστική μέθοδο ανάλυσης κατά δύο ή περισσότερους παράγοντες και δύο ή περισσότερες εξαρτημένες μεταβλητές, και εξετάζει κατά πόσο οι παράγοντες τις επηρεάζουν. Για τον έλεγχο της στατιστικής σημαντικότητας χρησιμοποιούμε έναν έλεγχο F, ο οποίος αποτελεί μία γενίκευση του ελέγχου t-Test, συγκρίνοντας τους μέσους όρους των μεταβλητών. Αν είναι σημαντικοί, τότε πραγματοποιούμε μονομεταβλητό έλεγχο (univariate) ANOVA, δηλαδή ένα για κάθε ξεχωριστή μεταβλητή.

Συσχέτιση (Correlation). Οι συσχετίσεις είναι στατιστικοί συνδυασμοί, έτσι ώστε να διερευνηθεί το πόσο κοντά είναι οι δύο μεταβλητές και να βρεθεί η γραμμική σχέση μεταξύ τους. Στην προγνωστική ανάλυση, μπορούν να βρεθούν είναι πιο σχετικές μεταβλητές σε αυτή που μας ενδιαφέρει και, έτσι, να μειώσουμε τις μεταβλητές. Δεν είναι μια τιμή που αφορά την αιτιατότητα (causation), δηλαδή πώς η αλλαγή σε μια μεταβλητή φέρνει αλλαγή σε άλλη μεταβλητή. Η θετική συσχέτιση, δηλαδή $0 < r_{xy} \leq 1$, σημαίνει ότι αν μια μεταβλητή αυξάνεται, τότε αυξάνεται και η άλλη μεταβλητή. Η αρνητική συσχέτιση, δηλαδή $-1 \leq r_{xy} < 0$, σημαίνει ότι αν μια μεταβλητή αυξάνεται, τότε η άλλη μεταβλητή μειώνεται. Όσο πιο μεγάλη είναι η τιμή κοντά στο 1 ή όσο πιο μικρή είναι η τιμή κοντά στο -1, τόσο πιο καλή είναι η σχέση.

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

Εξίσωση 51. Η εξίσωση συσχέτισης του Pearson με εύρος -1 μέχρι 1, με την τιμή 0 να δείχνει απουσία σχέσης.

Πολλαπλοί έλεγχοι (multiple testing). Όταν οι στατιστικοί έλεγχοι δείχνουν διαφορική έκφραση υπάρχουν διάφοροι τρόποι επιβεβαίωσής τους. Η p-value (probability-value), όπως είδαμε παραπάνω, είναι η πιθανότητα να της εμφάνισης μιας παρατηρούμενης στατιστικής

τιμής, αν θεωρήσουμε ότι η μηδενική υπόθεση ισχύει. Η q-value, από την άλλη πλευρά, είναι η τιμή σημαντικότητας όσο αφορά το ποσοστό εσφαλμένων ανακαλύψεων FDR (false discovery rate). Όσο μεγαλύτερη είναι η τιμή του στατιστικού αποτελέσματος τόσο μικρότερη είναι η p-value. Στο λογισμικό R εκτελείται η προσαρμογή της p-value μέσω της εντολής *p.adjust*, ενώ στο λογισμικό R/Bioconductor από το πακέτο limma με την εντολή *topTable*. Παρόλο που ο έλεγχος υποθέσεων παράγει μια τιμή p-value για κάθε γονίδιο, η επιλογή ενός κατωφλιού στατιστικής σημαντικότητας είναι δύσκολη στην περίπτωση των μικροσυστοιχιών αφού αφορά πολλαπλούς ελέγχους. Εφόσον εξετάζουμε χιλιάδες γονίδια, χρειάζεται προσαρμογή για να αφαιρέσουμε το στατιστικό σφάλμα από την απλή μέθοδο κατωφλιού, δηλαδή εφαρμόζουμε μια διαδικασία διόρθωσης πολλαπλών ελέγχων (Multiple Testing Procedure, MTP). Αυτή η προσαρμογή της p-value συνήθως γίνεται με ένα ολοκληρωμένο λογισμικό πακέτο λειτουργικού χαρακτηρισμού γονιδίων, όπως η εφαρμογή Galaxy[14], η εφαρμογή DAVID[15] (143, 144) και η εφαρμογή WebGestalt[16] (145). Έχουμε στη διάθεσή μας διάφορους μεθόδους για να αντιμετωπίσουμε τα σφάλματα. Για παράδειγμα, η PCER (Per-comparison error rate) είναι η αναμενόμενη τιμή (expected, E(X)) των σφαλμάτων τύπου I, ως προς τον αριθμό των υποθέσεων (m). Η FDR, που έχουμε ήδη αναφέρει και ονομάζεται επίσης Benjamini–Hochberg, μας δίνει την δυνατότητα να ελαχιστοποιήσουμε την αναλογία των σφαλμάτων τύπου I (146). Μια πιο συντηρητική μέθοδος είναι η FWER (Family Wise Error Rate), η οποία ονομάζεται και διόρθωση Bonferroni, αναφέρεται στην πιθανότητα (probability, P) να έχουμε τουλάχιστον ένα ψευδώς θετικό (V), δηλαδή ένα σφάλμα τύπου II. Πολλές φορές με αυτή την μέθοδο βιολογικά ενδιαφέροντα γονίδια δεν περνάνε το νέο κατώφλι p-value, τότε επιλέγονται τα γονίδια με την χαμηλότερη p-value.

$$PCER = \frac{E(V)}{m}$$

$$FDR = E\left(\frac{V}{R}, R > 0\right) = E\left(\frac{V}{R} \mid R > 0\right) P(R > 0)$$

$$FWER = \Pr(V \geq 1)$$

Εξίσωση 52. Διόρθωση σφαλμάτων τύπου I, όπου R είναι ο αριθμός των απορριφθέντων υποθέσεων.

Η διόρθωση του Bonferroni-Holm ακολουθεί συγκεκριμένα βήματα, σύμφωνα με την δημοσίευση του Holm (147). Η τιμή *p*-value κάθε γονιδίου ταξινομείται από την χαμηλότερη στην υψηλότερη τιμή. Στη συνέχεια, η πρώτη τιμή p-value πολλαπλασιάζεται με τον αριθμό

[14] https://usegalaxy.org
[15] https://david.ncifcrf.gov/
[16] http://webgestalt.org/

των γονιδίων που υπάρχουν στην γονιδιακή λίστα, αν η τελική τιμή είναι κάτω από 0.05, το συγκεκριμένο γονίδιο θεωρείται σημαντικό: Προσαρμοσμένη *p*-value=*p*-value×*n*<0.05. Μετά, η δεύτερη τιμή *p*-value πολλαπλασιάζεται με τον αριθμό των γονιδίων που υπάρχουν στην γονιδιακή λίστα, μειωμένη κατά μια μονάδα: Προσαρμοσμένη *p*-value=*p*-value×*n*-1<0.05. Το παρατηρούμενο επίπεδο ελέγχου σημαντικότητας (p-value) μπορεί ακόμα να υπολογιστεί και με την μέθοδο προσομοίωσης Monte-Carlo, με το λογισμικό στατιστικής SPSS και τον ακριβή έλεγχο του Fisher (Fisher's Exact Tests), που εκτελείται και από το λογισμικό R με την εντολή *fisher.test* και από το λογισμικό MATLAB με την εντολή *hygecdf*.

Θα πρέπει πάντα να εφαρμόζεται διόρθωση για τους πολλαπλούς ελέγχους, δηλαδή να ελέγχεται η στατιστική ανάλυση και να μειώνεται αρκετά το κατώφλι σημαντικότητας στο επίπεδο του p-value για να αποφύγουμε τις εσφαλμένες προβλέψεις.

Ανάλυση Σημαντικότητας Μικροσυστοιχιών (Significance Analysis of Microarrays, SAM). Significance Analysis of Microarrays (SAM), που δημοσιεύτηκε από την Virginia Tusher and και τους συνεργάτες της το 2001 (148), είναι μια στατιστική μέθοδος αναγνώρισης στατιστικά σημαντικών διαφορών. Το SAM[17] ανιχνεύει τα DEGs κάνοντας συγκεκριμένους ελέγχους *t*-Tests και υπολογίζει ένα στατιστικό αριθμό d_j για κάθε γονίδιο *j*, ο οποίος δίνει μέτρηση για την σχέση μεταξύ της γονιδιακής έκφρασης και μιας εξαρτημένης μεταβλητής. Πιο συγκεκριμένα, το SAM μπορεί να χρησιμοποιηθεί σαν πρόσθετο του Microsoft Excel. Προσαρμόζουμε την παράμετρο d_j, για να βρούμε το πλήθος των DEGs, τον FDR και να εκτιμήσουμε το μέγεθος του δείγματος. Έτσι, λαμβάνουμε την λίστα με τα διαφορικά εκφρασμένα γονίδια.

$$d_j = \frac{\bar{X}_{j1} - \bar{X}_{j2}}{s_j + s_0}$$

Εξίσωση 53. Το SAM d-test, όπου το $s_0=s^{\alpha}$.

Οπτικοποίηση. Μετά τις κανονικοποιήσεις στη μέσο ή την διάμεσο τιμή γίνεται η οπτικοποίηση των αποτελεσμάτων σε διάφορους τύπους διαγραμμάτων, όπως τα διαγράμματα διασποράς, τα volcano plots, τα θηκογράμματα και τα MA plots, που συναντήσαμε σε προηγούμενες παραγράφους, αλλά και ιστογράμματα (histograms),

[17] http://www-stat.stanford.edu/tibs/SAM/

διαγράμματα Venn, δεδρογράμματα, θερμικούς χάρτες, διαγράμματα πολυδιάστατης διαβάθμισης (MDS plot) κ.α. άμεσες οπτικοποιήσεις σε μορφή σύγκρισης δειγμάτων διαφορετικής φωτεινότητας. Οι τεχνικές υψηλής απόδοσης στις μικροσυστοιχίες είναι υψηλής πυκνότητας, δηλαδή αποτελούνται από πολλά σημεία δεδομένων. Η οπτικοποίηση βοηθάει να ελέγξουμε γρήγορα ότι έχουμε κάνει καλό «καθάρισμα» και κανονικοποίηση των δεδομένων, δηλαδή είναι ένας καλός ποιοτικός έλεγχος (Q.C.).

Θερμικός Χάρτης (Heatmap). Στους θερμικούς χάρτες χρησιμοποιούμε χρώματα για να οπτικοποιήσουμε τα αποτελέσματα από την ανάλυση της γονιδιακής έκφρασης σε σχέση με τα δείγματα που εξετάζουμε, είναι σύνηθες το κόκκινο χρώμα να υποδηλώνει την υπερέκφραση των γονιδίων, το πράσινο ή μπλέ την υποέκφραση των γονιδίων και το μαύρο την μη μεταβολή της. Στην γραμμή είναι τα ξεχωριστά γονίδια και στην στήλη είναι διαφορετικά δείγματα ή βιολογικά replicates.

Bayes. Ένα μοντέλο για εκ των υστέρων (posterior) μπεϋσιανή ανάλυση σε ένα κανονικό πληθυσμό $Y\sim(\mu,\sigma^2)$ και θεωρώντας ότι η πιθανότητα για την μεταβλητή ακολουθεί κανονική κατανομή $Pr(\mu)\sim N(\kappa,\varphi^2)$ (149).

$$\Pr\langle\theta|y\rangle:\ N\left(\frac{\frac{\kappa}{\sigma^2}+\frac{\sum_i^n y_i}{\sigma^2}}{\frac{1}{\varphi^2}+\frac{n}{\sigma^2}},\left(\frac{1}{\varphi^2}+\frac{n}{\sigma^2}\right)^{-1}\right)$$

Εξίσωση 54. Η posterior *Bayes* πιθανότητα.

Limma. Στο πακέτο limma του R/Bioconductor μπορούμε να εφαρμόσουμε και την εμπειρική Μπεϋσιανή προσέγγιση (empirical Bayes). Μας δίνει την δυνατότητα να αναλύσουμε δεδομένα γενετικής έκφρασης υψηλής απόδοσης, αφού δίνει τη δυνατότητα μείωσης της διαστασιμότητας. Συγκεκριμένα, στην εμπειρική Μπεϋσιανή προσέγγιση, όπως στην δημοσίευση των (150), θεωρούμε προγενέστερη γνώση του άγνωστου γονιδίου και των χρησιμοποιούμε την αντίστροφη *Γάμμα* κατανομή (Γ^{-1}) για να υπολογίσουμε το προσαρμοσμένο *t* για την limma. Θα δούμε ένα τέτοιο μοντέλο από την δημοσίευση των (151).

$$\sigma^2 \sim \Gamma^{-1}\left(\frac{\nu_0}{2},\frac{\nu_0 s_0^2}{2}\right)$$

Εξίσωση 55. Η αντίστροφη κατανομή Γάμμα της σ^2, όπου ν είναι η d.f., η s^2 και η s_0^2 αντιστοίχως είναι η εκ των υστέρων διακύμανση και η πρώιμη διακύμανση της κατανομής.

$$t = \frac{1}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \cdot \frac{\widehat{\beta}}{\tilde{s}}$$

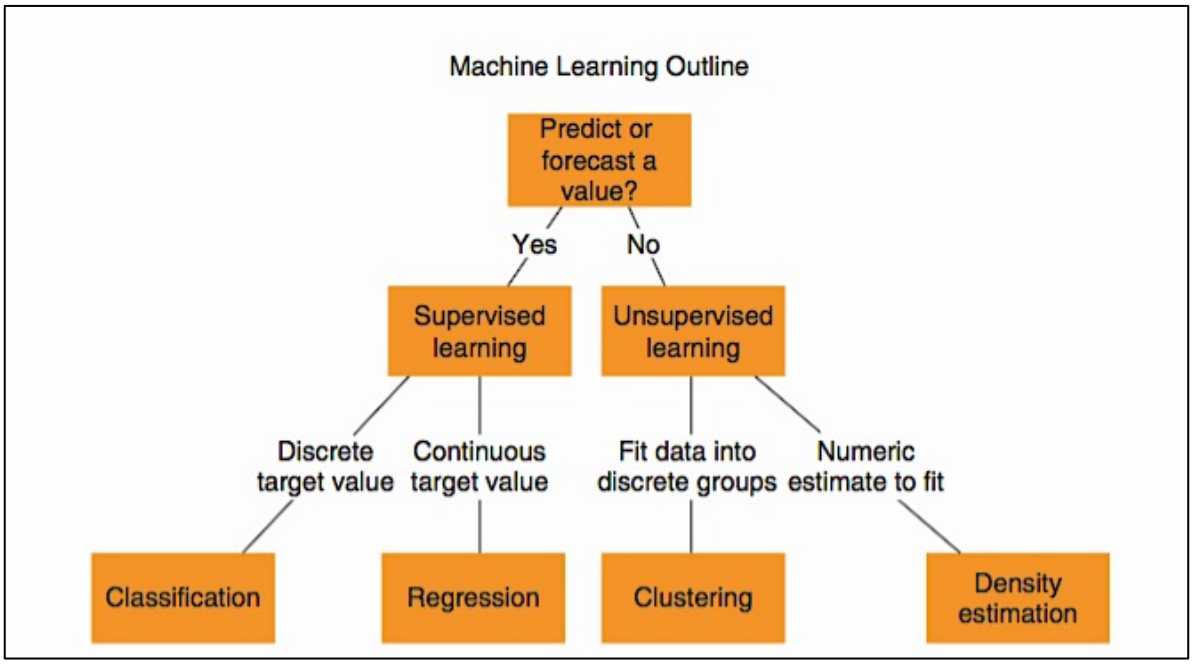
Εξίσωση 56. Η στατιστική t της limma, όπου η $\widehat{\beta}$ ακολουθεί κανονική κατανομή και μπορεί να οριστεί ως η log*FC* των δυο πληθυσμών και η $\tilde{s}$ ακολουθεί κατανομή *chi*-squared (χ^2).

$$\tilde{s}^2 = \frac{\nu_0 s_0^2 + \nu s^2}{\nu_0 + \nu}$$

Εξίσωση 57. Η εκ των υστέρων διακύμανση του πληθυσμού.

3.3. Υπολογιστικές Μέθοδοι

Μηχανική Μάθηση (Machine Learning, ML). Η Τεχνητή Νοημοσύνη είναι η επιστήμη που μαθαίνει των υπολογιστή να επεξεργάζεται στοιχεία και να παίρνει αποφάσεις με παρόμοιο τρόπο που σκέφτεται ο άνθρωπος. Ένας επιμέρους τομέας της Τεχνητής Νοημοσύνης ονομάζεται Μηχανική Μάθηση και μελετάει σύνολα δεδομένων, μέσω στατιστικών και υπολογιστικών μεθόδων (**Εικόνα 20**). Είναι ένας τομέας που συνδέεται στενά με τις έννοιες Εξόρυξη Δεδομένων και ανάλυση Μεγάλων Δεδομένων (Big Data). Ο βασικός στόχος της μηχανικής μάθησης είναι η εκπαίδευση αλγορίθμων για να ανακαλύπτουν ορθολογικά νέα στοιχεία, που δεν γνωρίζαμε πριν, αλλά και συσχετισμούς ανάμεσα στα δεδομένα. Με τις τεχνικές εξόρυξης δεδομένων να καλυτερεύουν παράγονται όλο και πιο σωστά συμπεράσματα και αποτελέσματα από την κάθε ανάλυση. Η έννοια της ανάλυσης Μεγάλων Δεδομένων είναι πολύ σημαντική για την διενέργεια μεγάλων ερευνητικών προγραμμάτων στους τομείς της Βιολογίας, της Ιατρικής και της Φαρμακολογίας. Ένα απλό παράδειγμα είναι η ανακάλυψη νέων φαρμάκων μεγάλων εταιριών, όπου συλλέγονται και αποθηκεύονται πληροφορίες από ομάδες επιστημόνων διαφορετικών ειδικοτήτων π.χ. μοριακών βιολόγων, γιατρών κ.α., σε αρχεία διαφορετικών μορφών και αυτές οι πληροφορίες επεξεργάζονται ίσως από μια άλλη ομάδα βιοπληροφορικών. Η διαχείριση, η επεξεργασία και η ερμηνεία των μεγάλων σε όγκο και πολυπλοκότητας δεδομένων πραγματοποιείται ολοκληρωμένα, από μια βάση δεδομένων, και έχουμε βελτιωμένες δυνατότητες ανακάλυψης νέας πληροφορίας αλλά και επαναχρησιμοποίησης των δεδομένων π.χ. αποθηκεύοντας τα σε μια δομημένη βάση, όπως την Γονιδιακή Οντολογία (Gene Ontologies, GO).



Εικόνα 20. Το διάγραμμα κατηγοριών του ML (152).

Επιβλεψημότητα. Η ML ανάλογα με το αν εκτελείται πρόβλεψη μιας τιμής ή όχι χωρίζεται σε δύο κατηγορίες, αν έχουμε πρόβλεψη ανήκει στην κατηγορία της επιβλεπόμενης μάθησης (Supervised) και εισάγεται η έννοια Αναγνώρισης Προτύπων (Pattern Recognision), αλλιώς ανήκει στην κατηγορία μη επιβλεπόμενης μάθησης (Unsupervised), αφού ο αλγόριθμος δεν χρειάζεται ανθρώπινη επίβλεψη. Θα χρησιμοποιήσουμε ένα παράδειγμα για να γίνει πιο κατανοητό, στην πρώτη περίπτωση έχουμε έναν αλγόριθμο που εξετάζει δείγματα από άτομα που ξέρουμε ότι έχουν καρκίνο και από άτομα που ξέρουμε ότι δεν ασθενούν και να καταλάβουμε ποιος παράγοντας και με ποιό τρόπο επηρέασε την υγεία του ατόμου, ενώ στην δεύτερη περίπτωση έχουμε έναν αλγόριθμο που εξετάζει δεδομένα από έναν φυσιολογικό κυτταρικό πληθυσμό σε αντιπαράθεση με ύποπτα κύτταρα για την ανίχνευση διαφορικής έκφρασης σε αυτά και την ύπαρξη κυτταρικών κυττάρων. Ακόμα, υπάρχει και η ημι-επιβλεπόμενη μάθηση, στην οποία μαζί με τα δεδομένα χωρίς πληροφορίες περιέχεται και μια μικρή ομάδα με γνωστές πληροφορίες. Η τελευταία περίπτωση είναι λιγότερο απαιτητική υπολογιστικά από την επιβλεπόμενη μάθηση, αλλά πιο προβληματική στην εφαρμογή της, αφού έχουμε δύο τύπων δεδομένα.

Η ταξινόμηση/κατηγοριοποίηση ανήκει στην κατηγορία επιβλεπόμενης γνώσης και την χρησιμοποιούμε για πρόβλεψη διακριτών τιμών, ενώ για την πρόβλεψη συνεχούς τιμής μπορούμε να πραγματοποιήσουμε και ανάλυση παλινδρόμησης (Regression). Η ομαδοποίηση/συσταδοποίηση ανήκει στην κατηγορία της μη επιβλεπόμενης γνώσης και αφορά εφαρμογή τοποθέτησης των δεδομένων σε ξεχωριστές ομάδες, ανάλογα με τα πρότυπα έκφρασης των γονιδίων και την λειτουργία τους. Αν θέλουμε μια αριθμητική εκτίμηση στην περίπτωση μη επιβλεπόμενης γνώσης τότε κάνουμε εκτίμηση πυκνότητας (Density Estimation), μια τέτοια μέθοδος είναι τα ιστογράμματα και οι εκτιμήτριες πυκνότητας πυρήνα (Kernel Density Estimation, KDE). Ένα παράδειγμα των μεθόδων KDE είναι ο αλγόριθμος k-κοντινότερων γειτόνων (k-Nearest Neighbors, kNN), στον οποίο θεωρούμε ότι παρόμοια γονίδια ή δείγματα βρίσκονται σε κοντινή απόσταση μεταξύ τους, δηλαδή θα εντοπίσουμε τα k κοντινότερα γονίδια στο υπό εξέταση σημείο και σύμφωνα με αυτά θα βγάλουμε το συμπέρασμά μας για αυτό (152).

3.3.1. Ομαδοποίηση ή Συσταδοποίηση (Clustering)

Ο στόχος της ομαδοποίησης είναι η αναγνώριση προτύπων, η οπτικοποίηση και ερμηνεία των πληροφοριών από την πειραματική διαδικασία και ο κατακερματισμός της γενετικής πληροφορίας σε μικρότερες ομογενείς ομάδες, έτσι ώστε να επιτευχθεί η αναγνώριση μιας

ομάδας γονιδίων, τα οποία ανταποκρίνονται με τον ίδιο τρόπο μεταξύ διαφορετικών καταστάσεων. Η ομαδοποίηση είναι μια πρώτη ένδειξη της βιολογικής λειτουργίας των ομαδοποιημένων λειτουργιών, αλλά δεν αποδεικνύεται η ανομοιότητα σε όλες τις περιπτώσεις βιολογικής σημασίας. Η ομαδοποίηση είναι μια ευαίσθητη διαδικασία και ο ερευνητής καλείται να χρησιμοποιήσει τους κατάλληλους αλγορίθμους ομαδοποίησης για τα εξεταζόμενα δεδομένα και την επιθυμητή ανάλυση. Ακόμα και αν έχει γίνει η σωστή επιλογή του αλγορίθμου, τα αποτελέσματα της ομαδοποίησης μπορούν εύκολα να διακυμανθούν με μικρές διαφορές στα επιλεγμένα γονίδια από τις επιλογές που ακολουθήσαμε στην προεπεξεργασία. Οπότε, είναι απαραίτητη η βιολογική ερμηνεία και επικύρωση των αλγοριθμικών υπολογισμών μέσω διασταύρωσης τους με προϋπάρχουσα γνώση. Γι'αυτό, σε επόμενες ενότητες, θα αναλύσουμε την διαδικασία χαρακτηρισμού γονιδίων και την μοντελοποίηση των προτύπων γονιδιακής έκφρασης. Παρόλο που συνήθως χρησιμοποιούμε μη επιβλεπόμενες μεθόδους γνώσης για να επεξεργαστούμε λίστες DEGs χωρίς να γνωρίζουμε τα φαινοτυπικά χαρακτηριστικά τους, σε κάποιες εφαρμογές τις χρησιμοποιούμε και όταν υπάρχουν διαθέσιμες φαινοτυπικές πληροφορίες. Ο στόχος μας σε αυτή την περίπτωση είναι να συγκρίνουμε τα αποτελέσματα ομαδοποίησης με τον γνωστό φαινότυπο. Οι πιο καθιερωμένοι τρόποι ομαδοποίησης είναι η ιεραρχική ομαδοποίηση, η ομαδοποίηση k-μέσων και η ανάλυση διαστασιμότητας, π.χ. η ανάλυση κυρίων συνιστωσών (Principal Component Analysis, PCA). Άλλοι τρόποι ομαδοποίησης είναι η ομαδοποίηση ασαφών προσεγγίσεων (Fuzzy Approaches Clustering), όπως ο ασαφής c-μέσων (Fuzzy c-means, FCM) (153), και οι νευρωνικοί μέθοδοι ομαδοποίησης (Neural Network-based Clustering Methods), όπως είναι οι αυτο-οργανωμένοι χάρτες (Self-Organizing Maps, SOMs). Ο αλγόριθμος μέγιστης πιθανοφάνειας των εκτιμήσεων παραμέτρων ενός μοντέλου (Expectation-Maximization, EM) επίσης χρησιμοποιείται στην ομαδοποίηση δεδομένων μικροσυστοιχιών, όπου συμβαίνει η σύγκλιση EM το δείγμα έχει την μέγιστη πιθανότητα υπό συνθήκη (conditional) (154, 155). Η EM μπορεί να χρησιμοποιηθεί και όταν κάποια από τα δείγματα έχουν ελλείψεις στις τιμές τους (156).

Ιεραρχική Ομαδοποίηση (Hierarchical Clustering). Ένας βασικός τρόπος ομαδοποίησης είναι η ιεραρχική, όπου κάθε γονίδιο συγκρίνεται με κάθε άλλο γονίδιο και παράγεται μια τιμή ομοιότητας ή συγγένειας για το γονίδιο, δηλαδή αφορά τα πρότυπα έκφρασης των γονιδίων μεταξύ δύο δειγμάτων. Η ομαδοποίηση εφαρμόζεται μέσω δεικτών απόστασης και μέσω δείκτη συσχέτισης (157).

Αν x_i είναι η λογαριθμισμένη τιμή για το γονίδιο x, δηλαδή την τιμή $\log_2 FC_x$, την χρονική στιγμή i και y_i είναι η λογαριθμισμένη τιμή για το γονίδιο y την χρονική στιγμή i, τότε για δυο γονίδια x και y, τα οποία αποτελούνται από n τιμές έκφρασης, υπολογίζεται η τιμή ομοιότητας/απόστασης, σύμφωνα με την σχετική δημοσίευση του D'Haeseleer (158). Στα δενδρογράμματα , η ομοιότητα απεικονίζεται μέσω του ύψους που δύο ομάδες ενώνονται σε αυτά, δηλαδή για μικρό ύψος έχουμε μεγάλη ομοιότητα.

Μέτρα Ομοιότητας/Απόστασης. Μερικά παραδείγματα μέτρων απόστασης είναι η απόσταση του Ευκλείδη (Euclidean distance), η οποία ομαδοποιεί αντικείμενα που έχουν όμοιες απόλυτες εκφράσεις, η απόσταση Μανχάταν (Manhattan distance), η απόσταση Minkowski και η απόσταση Mahalanobis (159).

$$d_{xy} = \sqrt{\sum_{i=1}^{n} \left(x_i - y_i \right)^2}$$

Εξίσωση 58. Ευκλείδια Απόσταση, που μετράει την γεωμετρική απόσταση μεταξύ δύο γονιδίων.

$$d_{xy} = \sum_{i=1}^{n} \left| x_i - y_i \right|$$

Εξίσωση 59. Απόσταση Μανχάταν.

$$d_{xy} = \sum_{i=1}^{n} \frac{\left| x_i - y_i \right|}{\left| x_i + y_i \right|}$$

Εξίσωση 60. Απόσταση Canberra.

$$d_{xy} = \left(\sum_{i=1}^{n} \left(x_i - y_i \right)^{\lambda} \right)^{\frac{1}{\lambda}}$$

Εξίσωση 61. Απόσταση Minkowski, όπου για λ=1 έχω την απόσταση Μανχάταν και για λ=2 την Ευκλείδια απόσταση.

$$d^2_{xy} = \left(\overline{x} - \overline{y} \right)' \sum_{i=1}^{n} {}^{-1} \left(\overline{x} - \overline{y} \right)$$

Εξίσωση 62. Απόσταση Mahalanobis d^2, όπου Σ είναι πίνακας συνδιακύμανσης.

Μέτρα Ομοιότητας/Συσχέτισης. Τα μέτρα συσχέτισης μπορούν να διακρίνουν καλά την ομοιότητα στο σχήμα αλλά είναι ευαίσθητα στις ακραίες τιμές. Μερικά παραδείγματα μέτρων συσχέτισης είναι ο συντελεστής συσχέτισης Pearson, που είναι η κανονικοποιημένη μορφή της συνδιακύμανσης και ομαδοποιεί αντικείμενα που έχουν παρόμοια πρότυπα διαφορών (160), ο συντελεστής συσχέτισης τάξεων μεγέθους του Spearman και της συνημιτονικής γωνίας (cosine angle). Τον συντελεστή Pearson τον χρησιμοποιούμε για δεδομένα που ακολουθούν κανονική κατανομή, σε αντίθετη περίπτωση προτιμούμε τον Spearman. Μπορούμε να συγκρίνουμε και γονίδια από διαφορετικά πειράματα ή διαφορετικές χρονικές στιγμές. Ακολουθούν γνωστές στατιστικές εξισώσεις συσχέτησης, όπως παρουσιάζονται στη δημοσίευση των (161), για την σύγκριση ομοιότητας δύο γονιδίων.

Συντελεστής συσχέτισης Pearson (Pearson correlation coefficient r_{xy}). Υπολογίζει την δύναμη της γραμμικής συσχέτισης μεταξύ των επιπέδων έκφρασης των γονιδίων. Όταν ο συντελεστής συσχέτισης παίρνει την τιμή 0, τότε δεν έχουμε γραμμική σχέση μεταξύ των συγκρινόμενων γονιδίων. Από την άλλη όταν παίρνει τις ακραίες τιμές του, δηλαδή την -1 και την 1, έχουμε τον μεγαλύτερο βαθμό συσχέτισης (χρησιμοποιούμε απόλυτο ώστε να έχουμε μόνο την θετική τιμή σε πολλές εφαρμογές). Για μη γραμμικές σχέσεις ο συντελεστής συσχέτισης Pearson δεν αποδεικνύει ομοιότητα και είναι ευαίσθητος στον θόρυβο.

Υπολογίζουμε τις τιμές από όλους τους συνδυασμούς γονιδίων και βρίσκουμε τα ζευγάρια γονιδίων με την μεγαλύτερη συσχέτιση σε μια από πάνω διαγώνια μήτρα τιμών (scoring). Βρίσκουμε το ζευγάρι με την υψηλότερη τιμή και τον μέσο αυτών. Αυτός ο κόμβος προσμετράται με τα υπόλοιπα στοιχεία και η μήτρα ανανεώνεται με τις νέες τιμές, αυτή η διαδικασία επαναλαμβάνεται μέχρι να μείνει ένα μόνο γονίδιο (157).

$$d_{xy} = 1 - r_{xy} = 1 - \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

Εξίσωση 63. Η απόσταση του συντελεστή συσχέτισης Pearson.

$$d_{xy} = 1 - \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$$

Εξίσωση 64. Συνημιτονική γωνία.

$$\rho = 1 - \frac{6\sum_{i=1}^{n} \left(x_i - y_i \right)}{n\left(n^2 - 1 \right)}$$

Εξίσωση 65. Ο συντελεστής συσχέτισης Spearman για διακριτές τιμές.

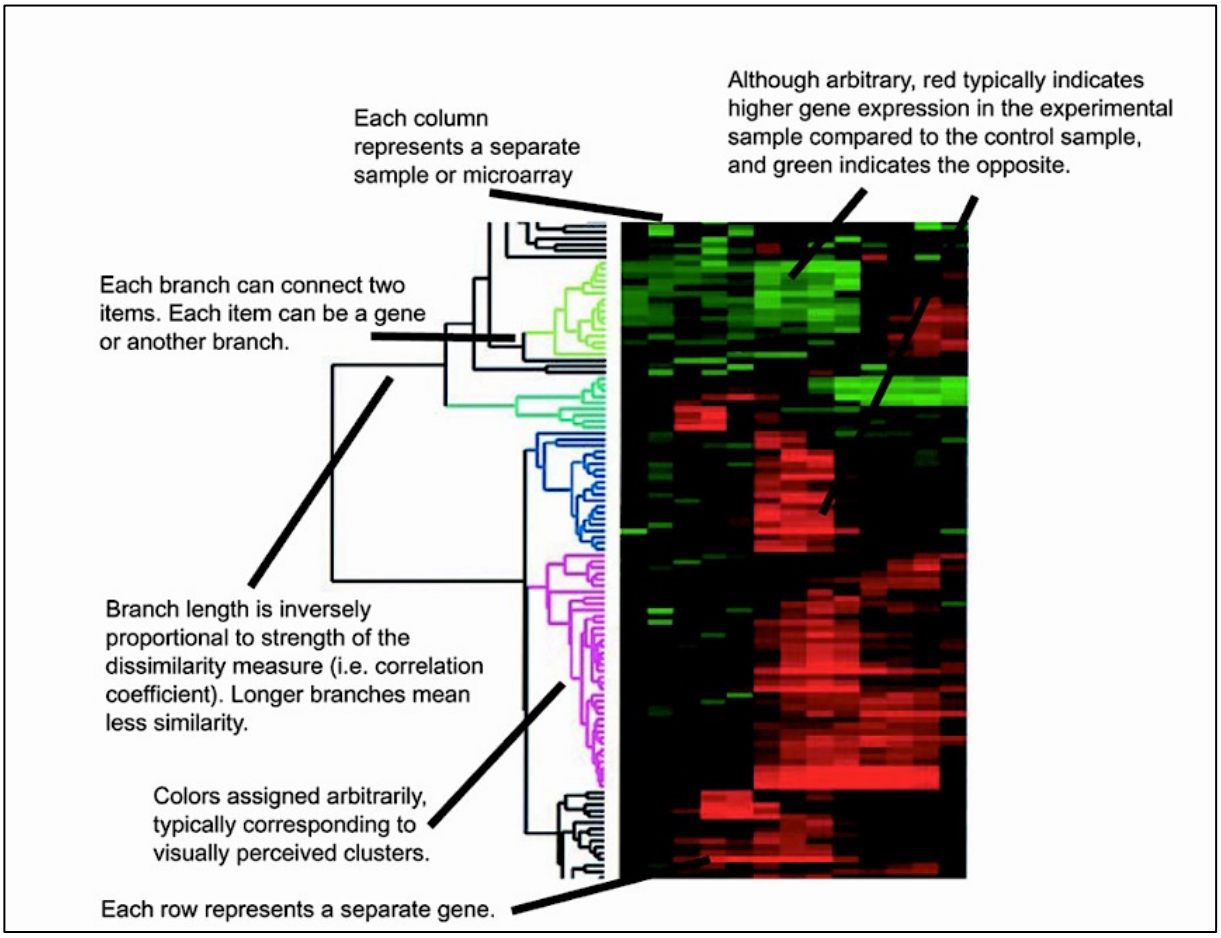
$$d_{xy} = 1 - r_{xy} = 1 - \frac{\sum_{i=1}^{n} (x_i' - \bar{x}')(y_i' - \bar{y}')}{\sqrt{\sum_{i=1}^{n} (x_i' - \bar{x}')^2 (y_i' - \bar{y}')^2}}$$

Εξίσωση 66. Η απόσταση του συντελεστή δειγματικής συσχέτισης Spearman, όπου x_i'=rank(x_i) και y_i'=rank(y_i).

Δενδρόγραμμα. Είναι σύνηθες η ανάλυση ιεραρχικής ομαδοποίησης (Hierarchical cluster analysis, HCA) να αναπαρίστανται σε δενδρόγραμμα ή αλλιώς φυλογενετικό δέντρο, όπου τα πιο παρόμοια δείγματα ομαδοποιούνται μαζί και οι πιο παρόμοιες ομάδες επίσης. Τα χρώματα στο δενδρόγραμμα αναπαρίστανται με κόκκινο για θετική λογαριθμική τιμή (μέχρι την τιμή 3, όπου έχω το πιο έντονο κόκκινο), μαύρο για μηδενική τιμή και κόκκινο για αρνητική λογαριθμική τιμή (από την τιμή -3, όπου έχω το πιο έντονο πράσινο), όπως φαίνεται στην **Εικόνα 21**. Το δεδρόγραμμα μπορούμε να το οπτικοποιήσουμε και δισδιάστατα, με τα γονίδια (y-άξονας) και τα δείγματα (*x*-άξονας) να ομαδοποιούνται ξεχωριστά. Όταν μια ομάδα γονιδίων είναι ρυθμιζόμενη από κοινού, οι περιοχές εκκινητών αυτών των γονιδίων ελέγχονται τα διατηρητέα πρότυπα που αναπαριστούν την αλληλεπίδραση με συγκεκριμένους μεταγραφικούς παράγοντες. Η μέθοδος ομαδοποίησης που θα χρησιμοποιηθεί σε κάθε πείραμα μικροσυστοιχιών εξαρτάται από την αρχική υπόθεση μας.

Συσσωρευτικοί και Διαχωριστικοί αλγόριθμοι. Η ιεραρχική ομαδοποίηση χωρίζεται σε δύο κατηγορίες ανάλογα με το ποια κατεύθυνση ακολουθεί η ομαδοποίηση. Με τον

συσσωρευτικό (agglomerative) αλγόριθμο θεωρούμε ότι κάθε στοιχείο ανήκει στον εαυτό του και οι ομάδες ξεκινάνε άδειες . Στο πρώτο βήμα τα δύο πιο όμοια στοιχεία γίνονται μια ομάδα και αυτό συνεχίζεται μέχρι όλα τα γονίδια να συγκεντρώνονται σε μια ομάδα. Στους συσσωρευτικούς αλγορίθμους χτίζουμε τις διακλαδώσεις του δενδρογράμματος αρχίζοντας με τις δύο πιο παρόμοιες ομάδες, όπου μια ομάδα ορίζεται σαν μια υποδιακλάδωση σε ένα φυλογενετικού τύπου δέντρο, που δημιουργείται χρησιμοποιώντας ένα μέτρο ζευγαρωτής ομοιότητας, π.χ. ο συντελεστής συσχέτισης Pearson.



Εικόνα 21. Μία κλασική αναπαράσταση δενδρογράμματος (117).

Η ιεράρχιση γίνεται αντίστροφα στους διαχωριστικούς (divisive ή partitional) αλγορίθμους, όπου αρχικά έχουμε όλα τα γονίδια σε μια ομάδα και τα χωρίζουμε σε υποομάδες. Χτίζουμε το δενδρόγραμμα αρχίζοντας με τις πιο ανόμοιες ομάδες, δηλαδή ψάχνουμε το καλύτερο τρόπο διαχωρισμού των γονιδίων, έτσι ώστε να έχουμε στο τέλος τις πιο ομογενής ομάδες γονιδίων με την μεγαλύτερη εμφάνιση ανομοιότητα μεταξύ διαφορετικών ομάδων, μέχρι η κάθε ομάδα να αποτελείται από ένα μόνο γονίδιο. Οι διαχωριστικοί αλγόριθμοι έχουν υψηλότερη υπολογιστική πολυπλοκότητα, γι'αυτό είναι πιο σύνηθες να χρησιμοποιούμε τους

συσσωρευτικούς αλγορίθμους. Οι συσσωρευτικοί αλγόριθμοι εφαρμόζονται με το πακέτο DIANA (DIvisive ANAlysis Clustering)[18] και οι διαχωριστικοί με το πακέτο AGNES (AGglomerative NESting)[19] στο λογισμικό R/Bioconductor (162).

Συνδεσιμότητα. Οι μέθοδοι ομαδοποίησης με τους αλγόριθμους απόστασης χρησιμοποιούνται για να υπολογίσουν διακεκριμένες τιμές δημιουργώντας μια ιεραρχία ομάδων. Η συνδεσιμότητα είναι το δεύτερο μέτρο που μας ενδιαφέρει, την οποία υπολογίζει ο αλγόριθμος κάνοντας σύγκριση του μέτρου της απόστασης μιας νεοσύστατης με μια ομάδα που δημιουργήθηκε σε προηγούμενο βήμα (162). Η διαφορά είναι ότι ενώ το πρώτο μέτρο πραγματοποιεί σύγκριση ομοιότητας ανάμεσα σε γονίδια, το δεύτερο μέτρο πραγματοποιεί σύγκριση απόστασης ανάμεσα σε δύο ομάδες. Η συνδεσιμότητα μεταξύ των διαφορετικών καταστάσεων μπορεί να πραγματοποιηθεί μέσω διαφορετικών τύπων σύνδεσης (linkage), ανάλογα με τον τρόπο που υπολογίζουμε την απόσταση ανάμεσα σε δύο ομάδες.

Στην απλή σύνδεση (single linkage) υπολογίζουμε το ζεύγος με την μικρότερη τιμή του μέτρου απόστασης που υπάρχει ανάμεσα στα δύο σύνολα. Με αυτήν υπάρχει ο κίνδυνος οι δύο ομάδες να μην είναι όμοιες στην πραγματικότητα, δηλαδή τα ζεύγος στοιχείων που επιλέχθηκαν να μην είναι καθόλου αντιπροσωπευτικά της κάθε ομάδας, τότε έχουμε το φαινόμενο μιας μακριάς και στενής αλυσίδας (chaining). Στην πλήρη σύνδεση (complete/compact linkage) υπολογίζουμε το ζεύγος με την μεγαλύτερη τιμή του μέτρου απόστασης που υπάρχει ανάμεσα στα δύο σύνολα. Με αυτήν οι δύο ομάδες είναι πολύ συνεκτικές, με όλα τα μέλη της ομάδας να έχουν σχετικά την ίδια απόσταση μεταξύ τους, αλλά εξαρτώνται πολύ από τα στοιχεία που χρησιμοποιήθηκαν για να ξεκινήσουν την ομαδοποίηση (crowding). Στην σύνδεση μέσου (average linkage) υπολογίζουμε την μέση τιμή των αποστάσεων όλων των στοιχείων σε μια ομάδα i με τα στοιχεία στην ομάδα *j*, μεσω ενός αλγορίθμου όπως την μέθοδο μέσου UPGMA (Unweighted Pair Group Method with Arithmetic mean), η οποία υπολογίζεται στο MATLAB με την εντολή *average*, και την μέθοδο σταθμισμένων μέσων WPGMA (Weighted Pair Group Method with Arithmetic mean), η οποία υπολογίζεται στο MATLAB με την εντολή *weighted,* παίρνοντας υπόψη όλα τα ζεύγη (163). Με αυτήν οι δύο ομάδες είναι ικανοποιητικά συνεκτικές, λιγότερο από την πλήρη σύνδεση. Παρόλο που είναι πιο απαιτητικές μέθοδοι από την απλή και πλήρη σύνδεση

[18] https://www.rdocumentation.org/packages/cluster/versions/2.1.2/topics/diana
[19] https://www.rdocumentation.org/packages/cluster/versions/2.1.2/topics/agnes

χρησιμοποιούνται πιο συχνά από αυτές, αφού συνδυάζουν με μέτρο τα πλεονεκτήματα και τα μειονεκτήματα αυτών. Στην σύνδεση κεντροειδούς (centroid linkage), δηλαδή κέντρου βάρους, υπολογίζουμε την απόσταση των κεντροειδών, δηλαδή των στοιχείων στην ομάδα i και στην ομάδα j που βρίσκονται στο κεντρικό σημείο της κάθε ομάδας, που έχουν την μικρότερη απόσταση μεταξύ τους, μεσω αλγορίθμων όπως την μέθοδο κεντροειδούς UPGMC (Unweighted Pair Group Method with Centroid), η οποία υπολογίζεται στο MATLAB με την εντολή *centroid,* και την μέθοδο σταθμισμένου κέντρου βάρους WPGMC (Weighted Pair Group Method with Centroid), η οποία υπολογίζεται στο MATLAB με την εντολή *median*. Τέλος, στην σύνδεση του Ward (Ward's linkage) εφαρμόζουμε την ομαδοποίηση σύμφωνα με ανάλυση διακύμανσης, αρχίζοντας από πολλές διακριτές ομάδες-στοιχεία n μέχρι όλα τα στοιχεία να έχουν ομαδοποιηθεί σε μια ομάδα.

Η συνδεσιμότητα μπορεί να υπολογιστεί, όπου p ο αριθμός πειραμάτων, *n* ο αριθμός γονιδίων και οι ομάδες (clusters) συμβολίζονται αντιστοίχως με C_k και C_l, σύμφωνα με τους (157, 164, 165) με τον εξής αλγόριθμο:

1) Για *v=n* αρχίζοντας με το καλύτερα διαχωρισμένο τμήμα (partition).

2) Υπολογίζουμε το νέο διαχωρισμένο τμήμα, ενώνοντας τις δύο ομάδες που ελαχιστοποιούν την απόσταση, σύμφωνα με τις εξισώσεις που ακολουθούν και πια επιλογή σύνδεσης θα έχει τα επιθυμητά αποτελέσματα.

3)Ανανεώνουμε τα στοιχεία αποστάσεων των υπολειπόμενων ομάδων με το νέο διαχωρισμένο τμήμα-ομάδα.

4) Επαναλαμβάνουμε το βρόχο μέχρι που το *v*=1, αφού τότε όλα τα δεδομένα είναι πλέον σε μία ομάδα.

$$d\left(C_k^{(v)}, C_l^{(v)}\right) = \min_{x_i \in C_k^{(v)}; x_j \in C_l^{(v)}} d\left(C_k^{(v)} - C_l^{(v)}\right)$$ **Εξίσωση 67**. Σε απλή σύνδεση.

$$d\left(C_k^{(v)}, C_l^{(v)}\right) = \max_{x_i \in C_k^{(v)}; x_j \in C_l^{(v)}} d\left(C_k^{(v)} - C_l^{(v)}\right)$$ **Εξίσωση 68**. Σε πλήρη σύνδεση.

$$d\left(C_k^{(v)}, C_l^{(v)}\right) = \frac{1}{\left|C_k^{(v)}\right|\left|C_l^{(v)}\right|} \sum_{x_i \in C_k^{(v)}; x_j \in C_l^{(v)}} d\left(x_i - x_j\right)$$ **Εξίσωση 69**. Σε σύνδεση μέσου.

Μείωση Διαστάσεων. Με την μείωση διαστάσεων μειώνεται η πολυπλοκότητα ενός προβλήματος, όσο αφορά τα γονίδια ή τα δείγματα. Δίνει μια λύση στο πρόβλημα διαχείρισης

δεδομένων πολλών διαστάσεων (*d*). Χρησιμοποιείται στην ανάλυση δεδομένων γονιδιακής έκφρασης, ως προκαταρτικό βήμα για την ομαδοποίηση δεδομένων υψηλής απόδοσης. Η ανάλυση κύριων συνιστωσών (Principal Components Analysis, PCA) και η πολυδιαστασημότητα κλίμακας (Multidimensional Scaling, MDS) ανήκουν στις μεθοδολογίες μείωσης διαστάσεων, όπου θέλουμε να μειώσουμε τα χαρακτηριστικά των δεδομένων μας για να κάνουμε περαιτέρω ανάλυση και για να έχουμε βελτιωμένη οπτικοποίηση των αποτελεσμάτων. Το σημαντικό είναι ο μετασχηματισμός αυτός να διατηρήσει τη βασική δομή των χαρακτηριστικών τους και γι'αυτό τον λόγο μας ενδιαφέρουν περιοχές με υψηλά επίπεδα διακύμανσης.

Ανάλυση Κύριων Συνιστωσών (Principal Components Analysis, PCA). Η ανάλυση κύριων συνιστωσών, ονομάζεται και καθωσπρέπει ορθογωνική αναδιάταξη (proper orthogonal decomposition, POD), είναι ένας τρόπος πρότερης (a priori) περιγραφής των δεδομένων, δηλαδή χωρίς να ψάχνουμε ειδικά για ένα προκαθορισμένο πρότυπο (166). Η PCA δημιουργεί συνιστώσες που είναι σταθμισμένοι μέσοι των αρχικών χαρακτηριστικών και επιλέγονται οι άξονες στους οποίους παρατηρείται η μέγιστη διακύμανση των δεδομένων (167). Προσπαθούμε να υπολογίσουμε ένα νέο συνδυασμό χαρακτηριστικών, χωρίς να μειώνονται οι βασικές διαστάσεις. Στην γενωμική, συγκεκριμένα, εξετάζει αν η γονιδιακή έκφραση είναι συνεπής σε όλα τα δείγματα μιας πειραματικής ομάδας, δηλαδή οι συντεταγμένες των σημείων στον νέο χώρο αντιπροσωπεύουν γραμμικό συνδυασμό των αρχικών γονιδίων και χάνονται ελάχιστες χρήσιμες πληροφορίες (133). Το πλεονέκτημα της είναι ότι οι μη σχετικές πληροφορίες αναπαρίστανται σε μια συνιστώσα, από την άλλη πλευρά οι συνιστώσες μπορεί να μην παρουσιάζουν προφανή βιολογική ερμηνεία. Κάποιες φορές χρησιμοποιούμε την τεχνική μοναδικής τιμής αναδιάταξης (Singular Value Decomposition, SVD) για την εύρεση των κυρίων συνιστωσών των δεδομένων.

Πολυδιαστασημότητα Κλίμακας (Multidimensional Scaling, MDS). Η MDS προσπαθεί να μετασχηματίσει πίνακες αποστάσεων μεταξύ αντικειμένων σε μικρότερη κλίμακα, χωρίς να χάσουν τα αντικείμενα τις αρχικές αποστάσεις τους. Ειδικότερα, όταν έχουμε αντικείμενα σε τρισδιάστατο χώρο μπορούμε να τα μετασχηματίζουμε όσο πιο πιστά γίνεται σε δισδιάστατο χώρο (34). Το αποτέλεσμα της ομαδοποίησης μπορεί να οπτικοποιηθεί σε κατάλληλα διαγράμματα συνιστωσών MDS (MDS plot), που περιγράφουν τις διακυμάνσεις στην γονιδιακή έκφραση (168, 169). Στο λογισμικό R δουλεύουμε με τις εντολές *sammon* και *isoMDS*, στο πακέτο MASS και την κλασική εντολή *cmdscale*, στο πακέτο mva (170).

Ομαδοποίηση με αλγόριθμους κοντινότερων γειτόνων. Η ομαδοποίηση k-μέσων και οι αυτο-οργανωμένοι χάρτες ανήκουν στην κατηγορία ομαδοποίησης κοντινότερων γειτόνων, όπου πρώτα αποφασίζεται με απαιτητικό υπολογιστικό κόστος πόσες ομάδες k θα έχουμε, στην συνέχεια υπολογίζονται οι ομάδες και κάθε γονίδιο τοποθετείται σε μια ομάδα. Από την μια πλευρά, αυτές οι δύο παρουσιάζουν τα αποτελέσματα με εύκολα κατανοητό τρόπο, π.χ. με διαφορετικά χρώματα, και μπορείς αμέσως να εντοπίσεις τις παραλλαγές στα πρότυπα γονιδιακής έκφρασης. Από την άλλη πλευρά, δεν γνωρίζουμε ακριβώς πως θα δημιουργήσουμε την αρχική γεωμετρία του χάρτη ή του διαγράμματος. Γι'αυτό τον λόγο, πραγματοποιούμε στοχαστική ανάλυση και είναι σύνηθες να επικυρώνουμε τα αποτελέσματα. Αλλά και στην περίπτωση των replicates, αν δεν κάνουμε σωστή ομαδοποίηση μπορεί να εμφανιστούν υπερεκτιμημένες τιμές του μέτρου απόστασης, λόγω της ομοιότητας των μετρήσεων. Μπορεί ακόμα και μια τιμή γονιδίου αν λείπει, να έχουμε εσφαλμένα αποτελέσματα ομαδοποίησης, αν ήταν σημαντικού βιολογικού ενδιαφέροντος. Ακόμα, οι αλγόριθμοι αυτοί δεν είναι κατάλληλοι στην περίπτωση αρνητικών τιμών, π.χ για έλεγχο ογκοκατασταλτικών γονιδίων.

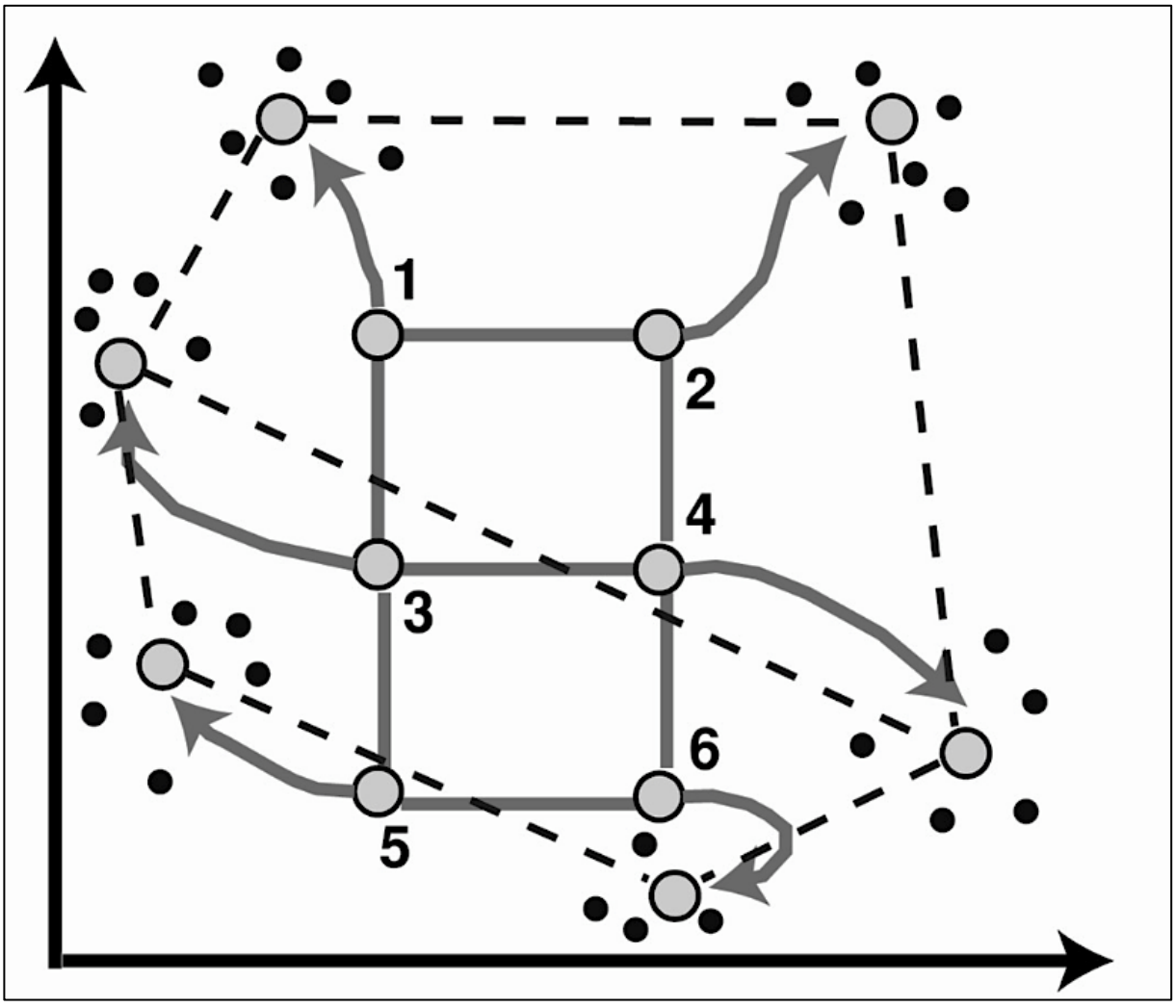
Ομαδοποίηση k-μέσων (k-means clustering). Στην ομαδοποίηση k-μέσων έχουμε ταξινόμηση και όχι ιεράρχιση. Είναι από τους πιο γνωστούς τρόπους ομαδοποίησης (171). Με αυτόν δεν υπολογίζουμε τις αποστάσεις μεταξύ κάθε γονιδίου, που είναι υπολογιστικά απαιτητικό, αλλά χρησιμοποιείται ένας γρήγορος αλγόριθμος με ένα πολύ έξυπνο τρόπο να πηγαίνει από διαχωρισμένο τμήμα σε διαχωρισμένο τμήμα. Η ομαδοποίηση k-μέσων διαχωρίζει τα γονίδια σε ομάδες που έχουν εσωτερικά μικρή διακύμανση και έχουν στο ενδιάμεσο μεταξύ τους μεγάλη διακύμανση (172, 173). Οι μετρήσεις μπορεί να αφορούν δύο ή παραπάνω χαρακτηριστικά. Ως πρώτο βήμα, αρχικοποιούμε τον αλγόριθμο, δηλαδή ο χρήστης ορίζει απαρχής τον αριθμό των *k* ομάδων, πολλές φορές με την βοήθεια ειδικών τεχνικών π.χ. PCA, στις οποίες θα διαχωριστούν τα δεδομένα γονιδιακής έκφρασης. Στην συνέχεια, ο αλγόριθμος ορίζει τυχαία τα *k* αυτά κεντροειδή σημεία, με διάφορους τρόπους π.χ. τελείως τυχαία στον χώρο μας ή τυχαία στις πιο πυκνές περιοχές ή με βάση επιλεγμένα πρότυπα (174). Μετά υπολογίζουμε τις αποστάσεις ανάμεσα τους χρησιμοποιώντας κάποιο από τα προαναφερόμενα μέτρα ομοιότητας, και αποδίδουμε σε κάθε κέντρο τα σημεία με την ελάχιστη απόσταση από αυτό, ενώ το κάθε σημείο αποδίδεται σε μια μόνο ομάδα. Αφού εκτιμήσαμε τις ομάδες δεν μας ενδιαφέρουν πλέον τα κέντρα αυτά και επικεντρωνόμαστε στα νέα κεντροειδή, που ορίζονται ως οι μέσες συντεταγμένες όλων των σημείων της κάθε ομάδας. Επαναλαμβάνουμε αυτή την διαδικασία, εφόσον χρειάζεται. Έτσι, συνεχίζουμε τις

επαναταξινομήσεις των στοιχείων σε διαδοχικούς κύκλους επανάληψης του αλγορίθμου μέχρι να μην μπορούμε να κάνουμε περαιτέρω βελτιώσεις (όπως είναι η ελαχιστοποίηση της απόστασης κάθε γονιδίου από τα εκτιμώμενα σημεία κεντροειδών της κάθε ομάδας), έτσι ώστε να μην αλλάζει πλέον η ταξινόμηση αυτών των στοιχείων. Τότε, ο αλγόριθμος επιστρέφει τις τελικές ομάδες (175, 176). Αυτή η διαδικασία ονομάζεται και *k*-κεντροειδών (*k*-centroids) ή μέσων βαρών. Στην πράξη, δοκιμάζουμε διάφορες αρχικές θέσεις *k* έτσι ώστε να έχουμε τα καλύτερα αποτελέσματα ομαδοποίησης. Κάποια λογισμικά εσκεμμένα επαναλαμβάνουν πολλές φορές τον αλγόριθμο για να πραγματοποιήσουν συναινετική ομαδοποίηση.

Από την άλλη πλευρά, αν έχουμε μεγάλο αριθμό *k* μπορεί να να μην τερματίζει ο αλγόριθμος, με κάποια σημεία να συνεχίζουν να παλινδρομούν από την μια ομάδα στην άλλη. Γι'αυτό είναι καλό να ορίσουμε ένα άνω όριο επαναλήψεων του αλγορίθμου και, ίσως, κάποια συνάρτηση για να καταλήξει αυτό το σημείο στην πιο σωστή ομάδα. Αντιστοίχως λειτουργούμε και στην διαδικασία ομαδοποίησης με την διάμεσο τιμή, η οποία ονομάζεται *k*-μεδοειδών (*k*-medoids) (177). Παρόλο που δεν είναι ο πιο κατάλληλος τρόπος ομαδοποίησης των δεδομένων γονιδιακής έκφρασης, είναι σύνηθες να χρησιμοποιείται ενδεικτικά και να ακολουθείται από άλλους τρόπους ομαδοποίησης.

Αυτο-οργανωμένοι χάρτες (Self-Organizing Maps, SOMs). Οι SOM είναι παρόμοια μέθοδος με την k-μέσων, αλλά πρόκειται για μια διαδικασία εκμάθησης ενός επιπέδου τεχνητών νευρωνικών δικτύων (Artificial Neural Networks, ANN) (**Εικόνα 22**), το πρώτο από τα οποία σχεδιάστηκε το 1958 από τον Frank Rosenblatt. Μια σημαντική διαφορά είναι ότι οι γειτονικές ομάδες στους χάρτες έχουν μεγαλύτερη σχέση από αυτές που είναι πιο απομακρυσμένες, δηλαδή αν κάποιες ομάδες γονιδιών είναι γειτονικές, τότε είναι πολύ πιθανό αυτά να έχουν παρόμοια λειτουργία. Ο χάρτης που παράγεται είναι συνήθως ένα δισδιάστατο πλέγμα κόμβων ή νευρώνων (*n*), οι οποίοι περιγράφονται από το βάρος w_{ij}, και ο στόχος είναι να διατηρηθεί η τοπολογία κατά την μείωση διαστασημότητας. Είναι σύνηθες να χρησιμοποιούμε το μέτρο του Ευκλείδη για να εντοπίσουμε την απόσταση. Αρχικά, ο χρήστης ορίζει τον αριθμό των ομάδων και από τον αλγόριθμο SOM ορίζονται τα κεντροειδή σε αυθαίρετες θέσεις x_i και κάθε νευρώνας σχετίζεται με ένα τυχαίο πρότυπο βάρους w_{ij}, που ονομάζεται και διάνυσμα αναφοράς (178). Κατά την διάρκεια της εκμάθησης, για κάθε επανάληψη (t=1,2,3^{η} επανάληψη κ.ο.κ.) τα γονίδια συνεχίζουν να χαρτογραφούνται, με κάθε κεντροειδές να κινείται προς ένα τυχαία επιλεγμένο γονίδιο, δηλαδή γίνεται εύρεση του *n* που

είναι πιο κοντά στο διάνυσμα αναφοράς και προσαρμόζουμε το βάρη w_{ij} του n και τον κοντινότερων σε αυτόν νεωρώνων, έτσι ώστε να πλησιάζουν ακόμα περισσότερο στο x_i. Είναι ενδιαφέρον ότι όσο υψηλότερος είναι ο αριθμός επανάληψης του αλγορίθμου που εφαρμόζουμε και όσο πιο μακριά είναι το κεντροειδές από το αυθαίρετα επιλεγμένο σημείο, τόσο μικρότερη απόσταση διανύεται. Ο κοντινότερος n στο διάνυσμα αναφοράς "κερδίζει" την θέση και το διάνυσμα αναφοράς ενημερώνεται. Μετά από την πλήρη εφαρμογή της εκμάθησης, οι n αναπαριστούν τις ομάδες (117). Οπότε κάθε κεντροειδές θα βρίσκεται στο κέντρο μιας ομάδας, το οποίο είναι ένα θεωρούμενο πρότυπο γονιδιακής έκφρασης (179). Στην πράξη επεκτείνεται σταδιακά η SOM μέχρι να αναγνωρίζονται εμφανώς τα ξεχωριστά πρότυπα έκφρασης. Είναι μια μέθοδος κατάλληλη για θορυβώδη δεδομένα έχει και συνήθως δεν επηρεάζεται πολύ από τις ακραίες τιμές, βέβαια σε κάποιες περιπτώσεις μπορεί να πειράξει μια βέλτιστη ομάδα (180). Ακόμα, είναι μια εύρωστη μέθοδος ομαδοποίησης δεδομένων υψηλής απόδοσης, με καλά αποτελέσματα οπτικοποίησης (181).



Εικόνα 22. Αρχή Λειτουργίας των SOM (117, 179).

Είναι σύνηθες να χρησιμοποιείται η SOM και για έλεγχο της ποιότητας μικροσυστοιχιών. Κάποια πολύ γνωστά ελεύθερα λογισμικά για την εύρεση των SOM είναι: GeneCluster[20] (Whitehead Institute, MIT), Xcluster[21] (Sherlock, Stanford University), Cluster and TreeView[22] (Eisen, Stanford University/Berkeley), J-Express[23] (Molmine). Στο λογισμικό MATLAB χρησιμοποιούμε την εντολή *newsom*, *train* και *sim*. Στο λογισμικό R με την εντολή *som*, από τα πακέτα kohonen ή som ή wccsom (182). Στο λογισμικό Waikato Environment for Knowledge Analysis (WEKA) με την εντολή *SelfOrganizingMap* (183). Το WEKA είναι ένα ελεύθερης πρόσβασης λογισμικό με φιλικό προς τον χρήστη περιβάλλον γραμμένο σε Java που περιέχει εφαρμογή διαφόρων γνωστών μεθόδων ομαδοποίησης, με το πακέτο Weka.cluster, και ταξινόμησης, με το πακέτο Weka.classifier.

SOTA (Self-Organizing Tree Algorithm). Ο αυτοοργανώμενος αλγόριθμος δέντρου βασίζεται στην SOM, όμως είναι μια ιεραρχική μέθοδος ομαδοποίησης, την οποία μπορούμε να εφαρμόσουμε στα δεδομένα γονιδιακής έκφρασης. Η βασική διαφορά είναι ότι μετά την σύγκλιση των τιμών, ο νεωρώνας με την υψηλότερη ανομοιότητα γενετικών προτύπων έκφρασης, διαμερίζεται σε αδερφικούς νεωρώνες, με δομή δυαδικού δέντρου το οποίο ολοκληρώνεται μέσω στατιστικών μεθόδων. Το αρνητικό είναι ότι δεν προσφέρει μέθοδο επιβεβαίωσης της βιολογικής σημασία αυτού.

Ομαδοποίηση προσεγγίσεων ασάφειας (Fuzzy Approaches). Το πλεονέκτημα της ομαδοποίησης με ασαφείς προσεγγίσεις είναι ότι ένα γονίδιο μπορεί να συμπεριληφθεί σε πάνω από μια ομάδες, δηλαδή είναι μια ασαφής τοποθέτηση, η οποία είναι πιο χρήσιμη όταν θέλουμε να διαλευκάνουμε τις σχέσεις ενός γονιδίου με πολλαπλές βιολογικές λειτουργίες. Έτσι, εισάγεται και η έννοια του βαθμού συμμετοχής (membership), μέσω της οποίας υπολογίζεται το κατά πόσο ένα χαρακτηριστικό ανήκει σε διάφορες ομάδες. Πρόκειται για μια τιμή που μπορεί να αναδείξει σχέσεις μεταξύ δεδομένων και των ομάδων τους (184). Προτιμάμε την ασαφή προσέγγιση στις μικροσυστοιχίες όταν εξετάζουμε γονίδια που εμπλέκονται παράλληλα σε πολλές βιολογικές λειτουργίες. Ειδικότερα, οι προσεγγίσεις ασάφειας είναι εύρωστες και κατάλληλες για γενωμικά δεδομένα, που παρουσιάζουν υψηλά επίπεδα θορύβου.

---

[20] http://www.genome.wi.mit.edu/MPR
[21] https://web.stanford.edu/group/sherlocklab/cluster.html
[22] http://diyhyl.us/~bryan/irc/protocol.cache/EisenSoftware.htm
[23] http://jexpress.bioinfo.no/site/

FCM. Ο αλγόριθμος ασάφειας μέσου διαχωρίζει αντικείμενα σε διαφορετικές ομάδες, ενώ ένα αντικείμενο μπορεί να ανήκει σε παραπάνω ομάδες. Ο υπολογισμός της με την γλώσσα R γίνεται με την εντολή cmeans και με το λογισμικό MATLAB με την εντολή fcm.

Η μέθοδος FCM, όπως αναπτύχθηκε και μετέπειτα εξελίχθηκε, υπολογίζεται από τις παρακάτω εξισώσεις (153, 185, 186):

$$J_m = \sum_{i=1}^{n}\sum_{j=1}^{C} u_{ij}^{m}(\left\|X_i - C_j\right\|)^2$$

Εξίσωση 70. Η εξίσωση υπολογισμού του ασαφούς μέσου, όπου το m>1.

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\left\|X_i - C_j\right\|}{\left\|x_i - C_k\right\|}\right)^{\frac{2}{m-1}}}$$

Εξίσωση 71. Ο βαθμός συμμετοχής u_{ij} (membership) του x_i στο C_j, όπου το είναι το μετρούμενο στοιχείο στην i-ιοστή διάσταση (n) και το X_i είναι η επιλεγόμενη ομάδα.

$$C_j = \frac{\sum_{i-1}^{n} u_{ij}^{m} \cdot x_i}{\sum_{i-1}^{n} u_{ij}^{m}}$$

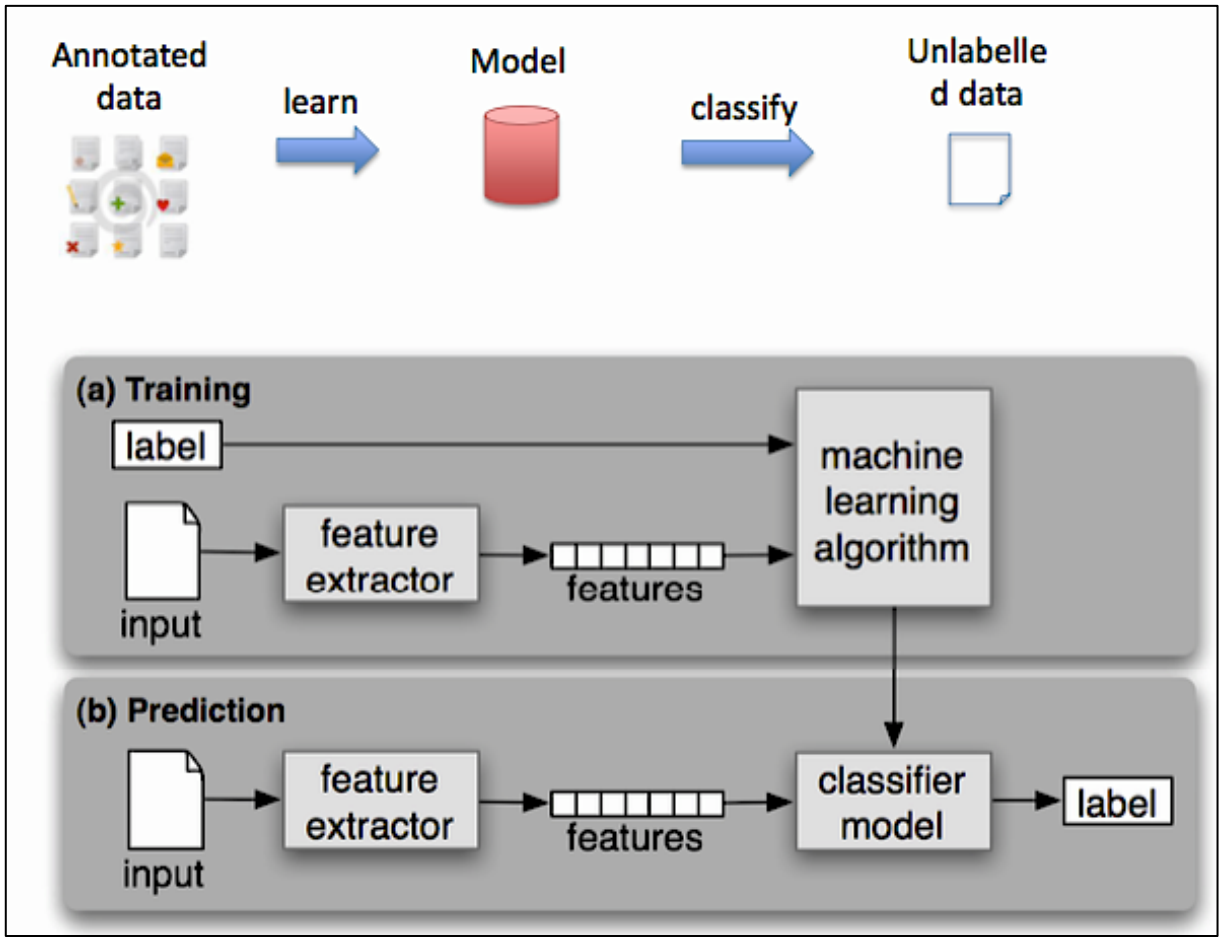
Εξίσωση 72. Το C_j είναι το κέντρο της ομάδας.

Δυο-Ομαδοποίηση (BiClustering). Τα γονίδια μπορεί να ανήκουν σε διαφορετικές ομάδες ανάλογα με τις συνθήκες του πειράματος και το κάθε δείγμα. Με την δυο-ομαδοποίηση μπορούμε να ανακαλύψουμε παράλληλα υπο-ομάδες γονιδίων και δειγμάτων, οι οποίες παρουσιάζουν μεγάλη ομοιότητα. Μια βασική λειτουργία αυτής της ομαδοποίησης είναι όταν απεικονίζουμε μαζί γονίδια και δείγματα, τα οποία όμως έχουν ομαδοποιηθεί ξεχωριστά, έτσι ώστε να ανακαλυφθούν πιθανές ομοιότητες στα πρότυπα γονιδιακής έκφρασης. Διαχωρίζονται κυρίως στους μεταευριστικούς ή στοχαστικούς αλγορίθμους και τους συστηματικούς αλγορίθμους. Για τους συστηματικούς αλγορίθμους υπάρχουν οι εξής βασικές προσεγγίσεις: διαίρει και βασίλευε (Divide-And-Conquer, DAC), άπληστοι επαναληπτικοί αλγόριθμοι (Greedy Iterative Search, GIS) και αλγόριθμοι απαρίθμησης δυο-ομάδων (Biclusters Enumeration, BE). Για τους μεταευριστικούς αλγορίθμους υπάρχουν διάφορες προσεγγίσεις: ευρετικοί αλγόριθμοι γειτονιάς (Neighborhood Search, NS), εξελικτικοί υπολογιστικοί αλγόριθμοι (Evolutionary Computation, EC) και οι υβριδικοί αλγόριθμοι (Hybrid, H), που είναι ο συνδυασμός των δυο παραπάνω προσεγγίσεων (187).

Επικύρωση και Ερμηνεία των αποτελεσμάτων. Στο τέλος της ομαδοποίησης χρειάζεται να γίνει επικύρωση των δημιουργημένων ομάδων, έτσι ώστε να ελεγχθούν αυτές ως πρός την ορθότητά τους μέσω κάποιου μέτρου ανάλυσης. Ακολουθεί η ερμηνεία των αποτελεσμάτων μέσω του χαρακτηρισμού των ομάδων.

3.3.2. Ταξινόμηση ή Κατηγοριοποίηση (Classification)

Η ταξινόμηση ανήκει στις μεθόδους επιβλεπόμενης μάθησης. Οι μέθοδοι επιβλεπόμενης αναγνώρισης προτύπων (supervised pattern recognition) γονιδιακής έκφρασης είναι μέθοδοι αυτόματης δημιουργίας μοντέλων μέσω γενίκευσης από ένα μεγάλο αριθμό δεδομένων εκμάθησης, έτσι ώστε να ελεγχθούν μελλοντικά δεδομένα. Η ταξινόμηση, πρόκειται για την είσοδο ενός υποσυνόλου δεδομένων εκπαίδευσης ή εκμάθησης (training ή learning), τα οποία περιέχουν τα επισημειωμένα ή με επιγραφή (labeled) πρότυπα γονιδιακής έκφρασης, έτσι ώστε να ταξινομηθούν στη συνέχεια με αξιοπιστία δεδομένα αξιολόγησης (validation) σε γνωστές κατηγορίες βιολογικών λειτουργιών (**Εικόνα 23**).



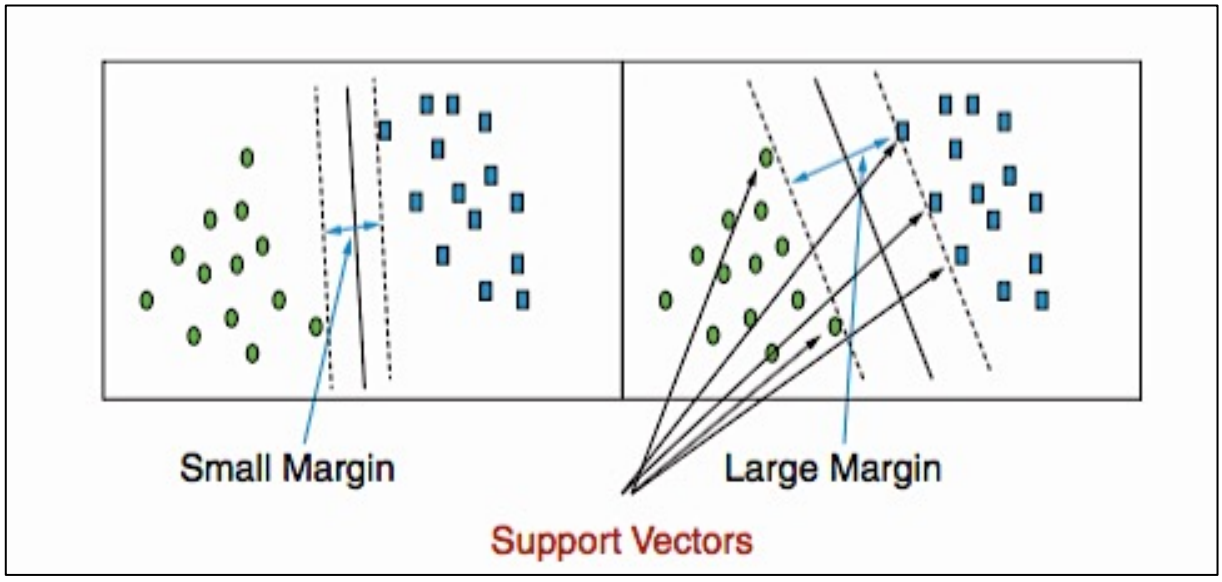
Εικόνα 23. Σχεδιάγραμμα Ταξινόμησης (188).

Μετά από τα στάδια της εκπαίδευσης και της αξιολόγησης το μοντέλο εφαρμόζεται στα δεδομένα δοκιμών (test) και ταξινομεί τα δεδομένα γονιδιακής έκφρασης όπως εκπαιδεύτηκε. Στις μικροσυστοιχίες είναι σύνηθες η επιγραφή να έχει την μορφή του αν προέρχονται από φυσιολογικά δείγματα ή όχι. Το αποτέλεσμα είναι η ταξινόμηση των γονιδίων ή δειγμάτων με κοινούς κανόνες συμπεριφοράς και η δημιουργία ενός μοντέλου πρόβλεψης, μέσω της εκτίμησης των παραμέτρων του κατηγοριοποιητή (classifier) για τα γονίδια ή δείγματα υπό εξέταση, γι'αυτό ονομάζεται και πρόβλεψη κατηγορίας (class prediction). Έτσι είναι δυνατή η εκτίμηση για την πρόβλεψη των τιμών ενός ή παραπάνω παραμέτρων που σχετίζονται με το άγνωστο δείγμα. Υπάρχουν πολλοί μέθοδοι για να καταφέρουμε την σωστή αναγνώριση και πρόβλεψη των γονιδίων. Θα διαλέξουμε την πιο ενδεδειγμένη μέθοδο, η οποία ταιριάζει στα δεδομένα μας και την απαιτούμενη ανάλυση, και αυτή που δίνει τα πιο σίγουρα αποτελέσματα, σε ακρίβεια ή σε ευαισθησία. Θα αναφέρουμε κάποιες από αυτές τις μεθόδους σε αυτή την ενότητα.

Ταξινόμηση k-κοντινότερων γειτόνων (k-Nearest Neighbor, kNN). Η μέθοδος ταξινόμησης kNN είναι ένας απλός αλγόριθμος (189, 190). Συγκαταλέγεται στους lazy αλγορίθμους, αφού δεν δημιουργείται ένα μοντέλο από εκμάθηση όπως στους άλλους κατηγοριοποιητές, αλλά γίνεται σύγκριση του μέτρου ομοιότητας του υπό εξέταση δεδομένου με όλα τα αποθηκευμένα δεδομένα εκμάθησης με κάθε νέο δείγμα. Πρέπει, αρχικά, να διαλέξουμε έναν αριθμό k για την ανεύρεση κοντινότερων γειτόνων, ο οποίος καλύτερα να είναι περιττός και όχι πολλαπλάσιο του αριθμού των κατηγοριών, και το μέτρο απόστασης που θα χρησιμοποιήσουμε. Ο αριθμός k πρέπει να είναι αρκετά μεγάλος για να γίνει σωστά η ταξινόμηση, αλλά, ταυτόχρονα, αρκετά μικρός ώστε τα σημεία να είναι αρκετά κοντά στο υπό εξέταση σημείο για να γίνει σωστή εκτίμηση της κατηγορίας του και να μην είναι πολύ χρονοβόρα η ανάλυση. Ο αλγόριθμος αυτός αποθηκεύει τα δεδομένα εκμάθησης και τις κατηγορίες τους, ενώ όταν έχουμε νέα δεδομένα προς εξέταση υπολογίζουμε τις αποστάσεις αυτών με τις ήδη αποθηκευμένες στην μνήμη. Στη συνέχεια, οι αποστάσεις ταξινομούνται από την μικρότερη τιμή προς την μεγαλύτερη τιμή. Αν ο αριθμός ήδη ταξινομημένων σημείων είναι ικανοποιητικός, υπολογίζουμε την πλειοψηφία μεταξύ των κατηγοριών από k παραδείγματα, που είναι τα πιο κοντινά στο υπό εξέταση σημείο. Ο kNN δεν έχει πολύ καλά αποτελέσματα σε δεδομένα μικροσυστοιχιών υψηλής απόδοσης χωρίς καλή πρότερη μείωση διαστασιμότητας, αφού τότε δεν υπάρχει ακρίβεια στις προβλέψεις και την ταξινόμηση των δεδομένων (191).

Γραμμική ανάλυση διακρίσεως LDA (Linear Discriminant Analysis). Η γραμμική ανάλυση διακρίσεως είναι μια από τις αρχικές μεθόδους ταξινόμησης πολυπαραγοντικών δεδομένων και είναι, ουσιαστικά, μια από τις μεθόδους μείωσης της διαστασημότητας. Βασίζεται στην εύρεση των γραμμικών προβολών των δεδομένων με τρόπο που διαχωρίζονται κατάλληλα οι k-κατηγορίες, έτσι ώστε να πραγματοποιήσουν πρόβλεψη του φαινοτύπου που μας ενδιαφέρει (192). Σε σύγκριση με άλλες μεθόδους δεν παρουσίασε καλά αποτελέσματα χωρίς να έχει γίνει πρώτα φιλτράρισμα των γονιδίων (190). Υπάρχουν και άλλες εκδοχές της ανάλυσης διακρίσεως, όπως η ευέλικτη ανάλυση διακρίσεως (Flexible Discriminant Analysis, FDA), η ανάμειξη αναλύσεων διακρίσεως (Mixture Discriminant Analysis, MDA) και η δευτεροβάθμια ανάλυση διακρίσεως (Quadratic Discriminant Analysis, QDA) (193).

Μηχανές υποστήριξης διανυσμάτων (Support vector machines, SVM). Η SVM είναι μια εύρωστη δυαδική μέθοδος ταξινόμησης, η οποία διαχωρίζει ένα σύνολο δεδομένων γονιδιακής έκφρασης από ένα άλλο εκατέρωθεν ενός υπερεπιπέδου (hyperplane) και χρησιμοποείται συχνά όταν θέλουμε να ανιχνεύσουμε διαφορές σε timecourse πειράματα ή σε πειράματα σύγκρισης μεταξύ διαφορετικών δειγμάτων (**Εικόνα 24**) (194, 195).



Εικόνα 24. Πιθανά υπερεπίπεδα και τα περιθώριά τους (196).

Πιο συγκεκριμένα, μεταμορφώνουμε την λίστα των DEGs σε μεγαλύτερες διαστάσεις μέσω μιας διαδικασίας που λέγεται kerneling, στην οποία χρησιμοποιείται μια συνάρτηση πυρήνα (kernel) έτσι ώστε να δημιουργηθεί ένα επίπεδο με τον συνδυασμό αυτών των γονιδίων. Μια βασική διαφορά με άλλες κάπως παρόμοιες μεθόδους π.χ. τις Μπεϋσιανές Μεθόδους είναι

στον τρόπο μείωσης των σφαλμάτων στο μοντέλο εκμάθησης, αφού στην SVM προσπαθούμε να μειώσουμε τα δομικά ελαττώματα. Πρακτικά, αυτό σημαίνει ότι βρίσκουμε ένα υπερεπίπεδο που να διαχωρίζει τα δείγματα/γονίδια στο πολυδιάστατο χώρο. Το υπερεπίπεδο πρέπει να έχει ίδια απόσταση από τα κοντινότερα διανύσματα στην καθεμία από τις δύο κατηγορίες, έτσι ώστε να μεγιστοποιήσουμε τα περιθώρια εκατέρωθεν του υπερεπιπέδου. Λαμβάνουμε υπόψη μόνο τα διανύσματα που είναι στα όρια αυτών και τα ονομάζουμε διανύσματα υποστήριξης. Η SVM κάνει σχετικά καλή γενίκευση στις μικροσυστοιχίες, αφού αποφεύγεται κατά κανόνα το πρόβλημα της υπερταύτισης (overfitting), που δημιουργείται από την υπερβολική εφαρμογή στις τιμές συνόλου δεδομένων εκμάθησης που περιέχουν θορυβώδη γονίδια. Ένα ακόμα πλεονέκτημα είναι ότι έχουν χαμηλό υπολογιστικό κόστος, ακόμη και στην περίπτωση μη γραμμικότητας, όπως και έχουμε συνήθως στις μικροσυστοιχίες και πρέπει να ρυθμίσουμε τα θέματα της υπερταύτισης με την βοήθεια της υπερπαραμέτρου C. Χρησιμοποιούνται δεδομένα ελέγχου για να υπολογιστεί η ακρίβεια του μοντέλου κατηγοριοποίησης και αν αυτή είναι ικανοποιητική, τότε είναι αποδεκτό το μοντέλο ταξινόμησης (181).

Δέντρα Αποφάσεων (Decision trees, DT). Με την κατασκευή ενός δέντρου αποφάσεως χρησιμοποιούμε ένα σύνολο εκμάθησης γονιδίων/δειγμάτων για να ταξινομήσουμε αντίστοιχα κάποια σύνολα αξιολόγησης. Αρχίζουμε με έναν κόμβο με όλα τα δεδομένα και μετά με τον διαιρετικό αλγόριθμο τα διαχωρίζουμε σε όλο και μικρότερα υποσύνολα, τα οποία οπτικοποιούμε στο αυξανόμενο δέντρο. Τα δέντρα αποτελείται από φύλλα, που εκφράζουν κάποια κατηγορία, και από τα κλαδιά, που «αποφασίζουν» για τον διαχωρισμό σύμφωνα με κάποια χαρακτηριστικά. Στο τέλος, για να λυθούν τυχόν προβλήματα υπερταύτισης, που είναι σύνηθες φαινόμενο στα δέντρα αποφάσεων, γίνεται «ψαλίδισμα» του δέντρου (tree prunning). Τα δέντρα αποφάσεων είναι εύρωστα και παρατηρείται γρήγορη ταξινόμηση ακόμα και δεδομένων υψηλής απόδοσης. Επίσης, τα αποτελέσματα είναι εύκολα κατανοητά από τον βιοπληροφορικό, κάτι που παίζει σημαντικό ρόλο για την δημοτικότητα των DTs. Ένας γνωστός DT αλγόριθμος είναι ο CART (Classification And Regression Trees), στον οποίο χρησιμοποιείται ένα υποψήφιο σύνολο ερωτήσεων με μια υποψήφια τιμή αξιολόγησης μεταξύ κάθε ζεύγους σημείων, έτσι ώστε να δημιουργηθεί ένα μοντέλο πρόβλεψης και χαρτογράφησης των δεδομένων (197). Άλλοι αλγόριθμοι που χρησιμοποιούνται είναι ο ID3 (Iterative Dichotomiser 3), ο CHAID (Chi-squared Automatic Interaction Detection) κ.α.

Τυχαίο Δάσος (Random Forest, RF). Χρησιμοποιούμε τον αλγόριθμο του τυχαίου δάσους για να αποφύγουμε το πρόβλημα της υπερταύτισης που εμφανίζεται στα DTs. Πρόκειται για την κατασκευή πολλαπλών DTs στο σκέλος της εκμάθησης, κάθε δέντρο δίνει την δική του ταξινόμηση και επιλέγεται η κατηγορία με τις περισσότερες «ψήφους», δηλαδή συγκλίνουν στο μέσο όρο των προβλέψεων. Το τυχαίο δάσος είναι ένας αλγόριθμος ακρίβειας και πραγματοποιεί πολύ καλή γενίκευση (198). Ένας παρόμοιος αλγόριθμος με του τυχαίου δάσους είναι ο Bootstrapped Aggregation (Bagging) και ο Gradient Boosted Regression Trees (GBRT).

Ανάλυση εκτίμησης για μικροσυστοιχίες (Prediction Analysis of Microarrays, PAM). Το εργαλείο PAM λειτουργεί παρομοίως με το εργαλείο SAM, που είδαμε σε προηγούμενο κεφάλαιο. Το PAM χρησιμοποιεί έναν βελτιωμένο, ως προς την ακρίβεια, αλγόριθμο συρρικνωμένου κεντροειδή (nearest shrunken centroids) για την ταξινόμηση με κατηγοριοποιητή κοντινότερου κεντροειδούς (199). Ο κατηγοριοποιητής αυτός αναγνωρίζει υποσύνολα γονιδίων που χαρακτηρίζουν καλύτερα μια κατηγορία, η συρρίκνωση, η οποία υπολογίζεται με την παράμετρο δέλτα (δ), ελέγχεται από ένα όριο, κάτω από το οποίο οι διαφοροποιήσεις θεωρούνται θόρυβος. Τα υψηλά όρια, για μεγάλο δ, έχουν συνήθως μεγάλο ποσοστό σφαλμάτων ταξινόμησης και όσο το όριο αυτό μειώνεται τόσο μειώνεται και αυτό το ποσοστό, τουλάχιστον μέχρι να παρουσιαστεί υπερταύτιση (34). Το εργαλείο αναγνωρίζει το δ που αντιστοιχεί στο μικρότερο σταυρωτά επικυρωμένο (cross-validated) σφάλμα πρόβλεψης και παρέχει μια λίστα με τα γονίδια που περιέχονται στον κατηγοριοποιητή για την συγκεκριμένη τιμή δ (200).

Μπεϋσιανά Δίκτυα (Bayesian Networks). Τα Μπεϋσιανά Δίκτυα ανήκουν στις Πιθανοτικές Μεθόδους Ταξινόμησης (Probabilistic Classification Methods). Ο κατηγοριοποιητής Μπέϋς βασίζεται στο θεώρημα του Bayes για την πρόβλεψη της κατηγορίας, όπου μεγιστοποιείται η εκ των υστέρων πιθανότητα. Έχουν ακρίβεια και καλή ταχύτητα για δεδομένα υψηλής απόδοσης. Επιπροσθέτως, κάνει καλή πρόβλεψη των αποτελεσμάτων ακόμα και αν παρέχονται μόνο ελλιπείς πληροφορίες. Ένα μειονέκτημα αυτού του κατηγοριοποιητή είναι ότι δεν υπάρχουν αρκετά δεδομένα για την αξιόπιστη εκτίμηση της από κοινού συνάρτηση πυκνότητας πιθανότητας.

Ένας άλλος γνωστός κατηγοριοποιητής, που αποτελεί μια απλοποίηση του κατηγοριοποιητή Μπέϋς, ονομάζεται Αφελούς Μπέϋς (Naive Bayes). Αυτά τα δίκτυα αποτελούν μια σχετικά

απλή προσέγγιση ταξινόμησης, με χαμηλό υπολογιστικό κόστος και εύκολη κατασκευή. Είναι αρκετά αποτελεσματικά για δεδομένα με πολλά ασθενώς συσχετισμένα χαρακτηριστικά. Από την άλλη, αν υπάρχουν ισχυρές ανεξάρτητες υποθέσεις, τότε περιορίζεται η απόδοση του αλγορίθμου, ειδικά όταν πολλά από τα χαρακτηριστικά έχουν ισχυρή συσχέτιση (201).

Λογιστική Ανάλυση Παλινδρόμησης (Logistic Regression). Η λογιστική ανάλυση παλινδρόμησης είναι ένα μη παραμετρικό μοντέλο εκτίμησης του πόσο μια ανεξάρτητη παράμετρος επηρεάζει μια εξαρτημένη παράμετρο, η οποία έχει μόνο δύο αποτελέσματα. Αυτό το μοντέλο χρησιμοποιείται για την εύρεση της κλίσης και της τέμνουσας μορφής μιας γραμμής και είναι ο αλγόριθμος που στην εξόρυξη δεδομένων χρησιμοποιείται ευρύτερα για την πρόβλεψη της πιθανότητας. Απαιτεί, ακόμα, την ανάλυση όλων των περιπτώσεων. Σε αντίθετη περίπτωση οποιαδήποτε παρατήρηση απορρίπτεται από την ανάλυση. Μια πιο απλή εκδοχή της είναι η γραμμική ανάλυση παλινδρόμησης, που εφαρμόζεται σε παραμετρικά μοντέλα, η οποία όμως δεν είναι συνήθως κατάλληλη για τα υψηλής απόδοσης δεδομένα μικροσυστοιχιών. Τα μοντέλα παλινδρόμησης παράγουν καλές εκτιμήσεις στα περισσότερα προβλήματα.

Νευρωνικά Δίκτυα (Neural Networks, NN). Το NN ή ANN είναι ένα υπολογιστικό μοντέλο που προήλθε σαν ιδέα από τον τρόπο που συνδέονται οι νευρώνες του εγκεφάλου και τις δυνάμεις μεταξύ τους. Πρωτοεμφανίστηκε την δεκαετία του πενήντα από τον Bernard Widrow του πανεπιστημίου Στάνφορντ. Το μοντέλο αυτό μπορεί να εκπαιδευτεί για να ταξινομεί περίπλοκα πρότυπα εμπειρικά. Το δίκτυο εκπαιδεύεται με την ρύθμιση και τον προσδιορισμό των βαρών των ακμών του κάθε νευρώνα, δηλαδή δεν πρόκειται για ένα μοντέλο βασισμένο σε κάποιο κανόνα ή σε ένα μη αλγοριθμικό μοντέλο, αλλά εκπαιδεύεται σε ένα ευρύτερο εννοιολογικό πλαίσιο, όπου και προσομοιώνεται η ανθρώπινη σκέψη. Η εκπαίδευση σταματάει όταν συγκλίνουν οι τιμές στα βάρη των ακμών. Η ικανότητα για γενικεύσεις παρέχεται από τα κρυμμένα στρώματα (hidden layers) και τους κρυφούς νευρώνες που περιέχουν. Στο δίκτυο υπάρχουν οι υπολογιστικοί νευρώνες, οι οποίοι πολλαπλασιάζουν τις εισόδους τους με τα συναπτικά βάρη και υπολογίζουν το άθροισμα του γινομένου, και οι νεωρώνες εισόδου, οι οποίοι βρίσκονται ανάμεσα στους πρώτους και στις εισόδους του δικτύου, οι οποίες συνήθως αριθμούν όσο και οι ανεξάρτητες μεταβλητές. Αυτό το άθροισμα είναι το όρισμα της συνάρτησης μεταφοράς, που μπορεί να είναι βηματική, γραμμική ή μη και στοχαστική. Ένα από τα πολλά παραδείγματα ανάλυσης δεδομένων

μικροσυστοιχιών με ANN είναι η μια δημοσίευση κλινικής έρευνας για την πρόβλεψη των αποτελεσμάτων ασθενών με νευροβλαστώμα (NB), όπου σε 25.000 γονίδια ο κατηγοριοποιητής του ANN αναγνώρισε ένα σύνολο δεκαεννιά γονιδίων με κοντινά συσχετισμένα επίπεδα γονιδιακής έκφρασης (202, 203). Γνωστοί αλγόριθμοι για τεχνητά νευρωνικά δίκτυα (ANN) είναι οι ακόλουθοι: Perceptron, Back-Propagation, Stochastic Gradient Descent, Hopfield Network, Radial Basis Function Network (RBFN). Για παράδειγμα, η οπισθοδιάδοση (Back-propagation) είναι πολύ χρήσιμη στην ανάλυση ιατρικών εικόνων και την εύρεση καρκινικών κυττάρων σε αυτές, αφού εκπαιδεύεται με την επαναλαμβανόμενη επεξεργασία των δεδομένων εκμάθησης και την σύγκριση των προβλέψεων του δικτύου με τις πραγματικές τιμές.

Η ύπαρξη πάνω από ενός κρυμμένου στρώματος υποδεικνύει δίκτυο βαθιάς μάθησης (deep learning). Οι αλγόριθμοι αυτοί χρησιμοποιούνται συχνά σε πιο πολύπλοκα δίκτυα και για την ανάλυση δεδομένων υψηλής απόδοσης. Σε αντιδιαστολή με τα θετικά των δικτύων βαθιάς μάθησης, η αύξηση του αριθμού των κρυφών στρωμάτων σε αυτά αυξάνει επίσης το χρόνο υπολογισμού και τον κίνδυνο υπερταύτισης, που οδηγεί σε κακές προβλέψεις. Ένας γνωστός αλγόριθμος deep learning είναι τα συνελεκτικά νευρωνικά δίκτυα (Convolution Neural Network, CNN). Άλλοι γνωστοί αλγόριθμοι για δίκτυα βαθιάς μάθησης είναι οι ακόλουθοι: Recurrent Neural Networks (RNNs), Long Short-Term Memory Networks (LSTMs), Deep Boltzmann Machine (DBM), Deep Belief Networks (DBN) (204).

Τα νευρωνικά δίκτυα μπορούν να δώσουν την λύση όταν άλλοι τρόποι ταξινόμησης δεν λειτουργούν σωστά και υπάρχουν πολλά ελεύθερα και μη λογισμικά για την επεξεργασία αυτών των δικτύων, αλλά η ερμηνεία των υπολογισμών στα βάρη των ακμών είναι δύσκολη.

Για την σωστή κατηγοριοποίηση των δεδομένων γονιδιακής έκφρασης πραγματοποιείται και επικύρωση ή σταυρωτή επικύρωση (cross validation) των αποτελεσμάτων και στο στάδιο των δοκιμών (testing). Η επικύρωση (εξωτερική) αφορά ετερογενή δεδομένα, δηλαδή αποτελείται από διαφορετικά δείγματα, τα οποία λήφθησαν ίσως σε διαφορετικές χρονικές στιγμές και με διαφορετικό πρωτόκολλο. Η σταυρωτή ή εσωτερική επικύρωση, γίνεται σε ένα υποσύνολο από το ίδιο σύνολο δεδομένων. Ο συντελεστής συσχέτισης Μάθιου (Mathew correlation coefficient, MCC) χρησιμοποιείται για μη ισορροπημένες τάξεις. Για παράδειγμα η σταυρωτή επικύρωση Κ-πτυχών (K-fold) χρησιμοποιείται συχνά για τον έλεγχο της λειτουργίας πρόγνωσης του ταξινομητή.

$$|MCC| = \sqrt{\frac{\chi^2}{n}}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Εξίσωση 73. Ο συντελεστής που μας ενδιαφέρει στην επικύρωση μη ισορροπημένων τάξεων.

4. Κεφάλαιο 4 Μέθοδος Ανάλυσης Αλληλουχοποίησης Νέας Γενιάς (Next Generation Sequencing (NGS))

Σε αυτό το κεφάλαιο, θα αναφερθούμε στη μελέτη και στην ποσοτικοποίηση της γονιδιακής έκφρασης καθώς και το μερίδιο της μεθυλίωσης σε συγκεκριμένες ή σχετιζόμενες αλληλουχίες με μεθόδους NGS σε πληθυσμούς συνήθως παρόμοιων δειγμάτων, δηλαδή μας ενδιαφέρει ο εντοπισμός της διαφορικής γονιδιακής έκφρασης και όχι οι αποκλίσεις από variants, SNPs κ.λ.π. Στο δεύτερο μέρος του κεφαλαίου θα δούμε τεχνικές ομαδοποίησης και ταξινόμησης, κάποιες από τις οποίες υπάρχουν στην ανασκόπηση του προηγούμενου κεφαλαίου. Παρόλο το κόστος της NGS έχει πτωτική τάση, το κόστος της ανάλυσης αυτών των δεδομένων μπορεί να είναι πολύ ακριβό.

4.1. Βάσεις Δεδομένων Αλληλούχισης και Ολοκληρωμένες Πλατφόρμες Βιοπληροφορικής

Η τεχνολογία αλληλούχισης νέας γενιάς όπως είδαμε θα απασχολήσει σύσσωμη την ιατρική κοινότητα, αφού γίνονται προσπάθειες για την χρήση προσωποποιημένης ιατρικής σε όλους τους τομείς και όχι μόνο στην έρευνα και στις κλινικές μελέτες. Η συστηματική ιατρική και βιολογία θα ξετυλίξει την πολυπλοκότητα του ανθρώπινου οργανισμού και θα αναπτυχθούν καλύτεροι δείκτες για την πρόγνωση, διάγνωση και θεραπεία ασθενειών. Με πλατφόρμες όπως οι διαγνωστικοί αλληλουχοποιητές και η τεχνολογία των big data, η προσωποποιημένη ιατρική είναι έτοιμη να καθιερωθεί και να εφαρμοστούν κοινά πρότυπα και ροές εργασίας, καθώς και τρόποι διασφάλισης της πρόσβασης στα ιατρικά δεδομένα των πολιτών μόνο εξουσιοδοτημένων ατόμων.

Οι επιστήμονες είναι υποχρεωμένοι να μοιράζονται τα αποτελέσματα τους σε βιολογικές βάσεις δεδομένων (Databases, DB), όταν δημοσιεύουν μια ερευνητική εργασία. Αυτή η μετάβαση σε μια ελεύθερη πρόσβαση στη γνώση ήταν ένας από τους βασικούς λόγους που η επιστημονική κοινότητα καταφέρνει να συνεργάζεται για ένα κοινό σκοπό και που ο τομέας της βιοπληροφορικής για την επεξεργασία αυτών των δεδομένων ευδοκιμεί. Η επεξεργασία τους σε πρωτογενές επίπεδο γίνεται συνήθως εξολοκλήρου στην πλατφόρμα αλληλούχισης, σε δευτερογενές επίπεδο υπάρχουν εφαρμογές εντός και εκτός της πλατφόρμας και σε τριτογενές επίπεδο έχουν δημιουργηθεί πολλές εφαρμογές εκτός της πλατφόρμας και μάλιστα παρέχουν εργαλεία για την οπτικοποίηση των συμπερασμάτων από τα αποτελέσματα της

έρευνας. Υπάρχουν, για τον σκοπό αυτό, ολοκληρωμένες σουίτες ανάλυσης των ωμικών δεδομένων, όπως οι πλατφόρμες DAVID και Galaxy.

Τα σχετιζόμενα δεδομένα στην πρωτογενή DB, δηλαδή αποθηκευμένες σε ψηφιακή μορφή αλληλουχίες από πειράματα NGS, μπορούν να τα κατατεθούν, να ανασυρθούν για ανάλυση και να ενημερωθούν από έναν μέσο χρήστη π.χ. έναν βιολόγο χωρίς ειδικές γνώσεις στην πληροφορική και την στατιστική. Μια από τις πιο σημαντικές έννοιες της βιοπληροφορικής και της βιολογίας είναι η σύγκριση δύο πρωτεϊνών ή αλληλουχιών νουκλεϊνικών οξέων. Πιο συγκεκριμένα, χρησιμοποιούμε εργαλεία, για την εύρεση μιας ομάδας συσχετιζόμενων ακολουθιών και την εύρεση του σκόρ ομοιότητας. Πραγματοποιούμε τη βιολογική στοίχιση για βραχύ-μικροανάγνωση (short reads) πλέον με νεότερα εργαλεία, όπως το Bowtie2[24] και το SOAP2[25], αφού είναι ένας από τους καλύτερους τρόπους να συγκρίνουμε τις αλληλουχίες αυτές με ταχύτητα και ευαισθησία, για να μπορούμε εντοπίσουμε σε ποιες θέσεις βρίσκονται πιθανώς DEGs.

Αν ξέρεις να διαχειρίζεσαι ένα εργαλείο αναζήτησης, μπορείς συνήθως να καταλάβεις και αποτελέσματα από άλλα εργαλεία αναζήτησης. Το μέγεθος των βάσεων δεδομένων αλληλούχισης αυξάνεται εκθετικά, όσο κατεβαίνει το κόστος των τεχνικών εξαιρετικά υψηλής απόδοσης. Το κάθε εργαλείο αναζήτησης έχει τις δικές του στατιστικές μορφές και τις μορφοποιήσεις των αποτελεσμάτων, κάτι που συμβάλει στην πολυπλοκότητα των αναλύσεων. Οπότε, πραγματοποιούμε υπολογισμό της ομοιότητας και την αναζήτηση των ομολόγων (homologues) αλληλουχιών, δηλαδή που προέρχονται από κοινούς προγόνους, σε κάποιες DB. Τα ομόλογα γονίδια διαχωρίζονται σε παράλογα γονίδια (paralogous genes), που εμφανίζονται σε διαφορετικά άτομα του ίδιου είδους, και σε ορθόλογα γονίδια (orthologous genes), που προκύπτουν από την εξέλιξη των οργανισμών και ανήκουν δηλαδή σε διαφορετικά είδη. Η ομολογία δηλαδή δείχνει ποιοτικά την ύπαρξη συσχέτισης ενώ η ομοιότητα δίνει αριθμητικά τον βαθμό της συσχέτισης. Οπότε ένας χρήστης της βιοπληροφορικής πλατφόρμας κάνει τις σχετικές αναζητήσεις στο γένωμα, οπτικοποιεί τα αποτελέσματα και εξάγει τα αντίστοιχα συμπεράσματα.

[24] http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
[25] http://soap.genomics.org.cn/soapaligner.html

Από την άλλη πλευρά, ο σκοπός του βιοπληροφορικού επιστήμονα είναι ο σχεδιασμός, ο προγραμματισμός και η εκπαίδευση των κατάλληλων αλγορίθμων για τα υπό ανάλυση δεδομένα. Αυτοί εφαρμόζονται στα δεδομένα δοκιμών (test) και ταξινομούνται τα δεδομένα γονιδιακής έκφρασης, όπως εκπαιδεύτηκαν οι αλγόριθμοι. Το έτοιμο μοντέλο τότε είναι κατάλληλο για χρήση ως βιοπληροφορικό εργαλείο ή web πλατφόρμα από την επιστημονική κοινότητα.

Παραγώμενα αρχεία αλληλούχισης. Τα αρχεία Fasta είναι ένας βασικός τύπος αρχείου που μπορεί να παράγει μια πλατφόρμα αλληλούχισης. Η FastQ είναι αρχείο αλληλούχισης Fasta μαζί με δεδομένα ποιότητας που παράγεται από την πλατφόρμα Illumina. Αντίστοιχα, η πλατφόρμα PacBio παράγει αρχεία αλληλούχισης BAM και η πλατφόρμα Nanopore (HDF5) παράγει αρχεία αλληλούχισης Fast5. Τα αρχεία αυτά είναι τα δεδομένα εισόδου σε λογισμικά ανάλυσης ποιότητας και γίνεται εξαγωγή μιας αναφορά σε HTML αρχείο, ακόμα και εκτός διαδικτύου. Τα εργαλεία αυτά μπορούν να χρησιμοποιηθούν κατευθείαν από το τερματικό (συνήθως σε περιβάλλον UNIX, αλλά κάποια εργαλεία δουλεύουν και σε περιβάλλον MAC ή WINDOWS) και μπορούν να αναλυθούν σε μια διαδικτυακή πλατφόρμα, όπως το Galaxy.

Πέναλτι κενών (Gap penalty). Το σκόρ ομοιότητας αφορά το άθροισμα του αριθμού ταυτίσεων ακριβείας και συντηρητικών (high scoring) αντικαταστάσεων στην στοίχιση, διαιρεμένο με το σύνολο των χαρακτήρων που στοιχίστηκαν, ενώ τα κενά μερικές φορές αγνοούνται (ungapped). Σε κάποιες άλλες περιπτώσεις λαμβάνουμε υπόψη τα κενά με γραμμικό ή affine τρόπο (linear vs affine gap penalty).

Τοπική Στοίχιση, Ολική Στοίχιση, Ημιολική Στοίχιση (Local, Global, Semiglobal alignment). Η τοπική στοίχιση γίνεται βάση της υψηλότερης πυκνότητας των ταυτίσεων. Η τοπική στοίχιση είναι gapped και δεν λαμβάνουμε υπόψη τα πέναλτι στις άκρες. Η τελευταία τιμή της στοίχισης είναι το υψηλότερο σκόρ στον πίνακα και μηδενίζουμε τις αρνητικές τιμές. Στις άλλες δύο περιπτώσεις δεν μηδενίζουμε τις αρνητικές τιμές. Στην ημιολική στοίχιση έχουμε καλύτερες στοιχίσεις σε ομόλογες αλληλουχίες διαφορετικού μήκους. Ακόμα, στην ημιολική δεν λαμβάνουμε υπόψη τα πέναλτι στις άκρες και η τελευταία τιμή της στοίχισης είναι το υψηλότερο σκόρ στην κατώτερη ή το υψηλότερο σκόρ στην δεξιότερη τιμή του πίνακα. Στην ολική στοίχιση πρέπει να ελεγχθούν για ταύτιση όλες οι θέσεις και στις δύο αλληλουχίες, άρα το χρησιμοποιούμε για αρκετά όμοιες και ίδιου μήκους αλληλουχίες. Ακόμα, σε αυτή

λαμβάνουμε υπόψη τα πέναλτι στις άκρες και η τελευταία τιμή της στοίχισης είναι η κατώτερη και δεξιότερη τιμή του πίνακα.

Αλγόριθμος Needleman-Wunsch (NW Algorithm). Λειτουργεί με την λογική του δυναμικού προγραμματισμού και τις έννοιες ταύτιση, αντικατάσταση και κενά. Είναι μια ολική στοίχιση μεταξύ των αλληλουχιών και έχει καλύτερα αποτελέσματα στοίχισης. Η διαδικασία είναι ντετερμινιστική και απολύτως επαληθεύσιμη.

Αλγόριθμος Smith-Waterman (SW Algorithm). Ο αλγόριθμος SW αφορά μόνο τα τμήματα με μικρή ομοιότητα μεταξύ τους και λειτουργεί με την λογική του δυναμικού προγραμματισμού για την εύρεση τοπικών στοιχίσεων μεταξύ αλληλουχιών. Καθορίζεται από την μικρότερη αλληλουχία και είναι μια παραλλαγή του Needleman-Wunsch. Η ιδιαιτερότητα του είναι ότι όλα τα αρνητικά σκόρ μηδενίζονται στον πίνακα για να αποφευχθούν κακοί υπολογισμοί στοίχισης και να είναι πιο εύκολο να αναγνωριστούν η αρχή και το τέλος των τοπικών στοιχίσεων σε οποιοδήποτε σημείο του πίνακα βρίσκονται.

BLAST. Ο ευρετικός αλγόριθμος για τοπική στοίχιση BLAST (Basic Local Alignment Search Tool) είναι ένα βασικό εργαλείο αναζήτησης ομοιότητας μιας ακολουθίας έναντι μιας βάσης αλληλουχιών και αφορά τοπικές στοιχίσεις, όπου αναζητούμε την ολική στοίχιση από μικρότερες τοπικές στοιχίσεις (205). Η αναγνώριση των αλληλουχιών στην βιολογία είναι μια από τις βασικές λειτουργίες της βιοπληροφορικής. Το BLAST δεν είναι ο πλέον ενδεδειγμένος τρόπος για στοίχιση, αφού υπάρχουν καλύτερα λογισμικά στοίχισης για γενωμικές αλληλουχίες DNA ή δεδομένα RNA-Seq, που προέρχονται από αλληλούχιση νέας γενιάς, λόγω του μεγάλου όγκου των δεδομένων, ενώ το παρόμοιο εργαλείο BLAT έχει λίγο καλύτερες επιδόσεις (206). Όμως είναι ένα χρήσιμο ευρετικό εργαλείο όταν έχουμε μια συγκεκριμένη αλληλουχία για την οποία ψάχνουμε ομόλογα ανάμεσα σε διαφορετικά είδη ή επιπλέον πρωτεΐνες με παρόμοιες δομικές περιοχές, στην διαδικτυακή πλατφόρμα Blast του NCBI, όπου τα αποθηκευμένα δεδομένα είναι πολυάριθμα. Τα εργαλεία αυτά διαχωρίζονται σε υπο-ομάδες ανάλογα με το τι αλληλουχίες εξετάζουν π.χ. για πρωτεΐνες υπάρχει το blastp και για νουκλεοτίδια το blastn κ.α.

Στοίχιση αλληλουχιών κατά ζεύγη (Pairwise Alignment). Έτσι ονομάζουμε την στοίχιση που πραγματοποιείται μεταξύ δύο αλληλουχιών. Για παράδειγμα, την στοίχιση κατά ζεύγη μπορούμε να την χρησιμοποιήσουμε για την μελέτη μεταξύ ενός DNA μεθυλώματος και ενός

μεταγράφου, δηλαδή για να βρούμε την σύμπραξη των DMGs-DEGs. Στην στοίχιση μεγάλων αλληλουχιών, όπως είναι τα ολόκληρα γενώματα, εκτός από τις αντικαταστάσεις (substitutions) και τα κενά (gaps), αναγνωρίζουμε επίσης τους διπλασιασμούς (duplications) και τις αναστροφές (reversals). Λογισμικά για στοίχιση κατά ζεύγη είναι το FASTA και το BLAST. The συστήματα PAM (Percent Accepted Mutation) και BLOSUM (BLOck Amino Acid SUbstitution Matrix) χρησιμοποιούνται για τον υπολογισμό του σκορ στην κατά ζεύγη στοίχιση (207, 208). Τα οποία έχουν τις εξής αντιστοιχίσεις με βάση τον ρόλο τους: το PAM250 με το BLOSUM45, το PAM160 με το BLOSUM62 και το PAM120 με το BLOSUM80.

Πολλαπλή στοίχιση (Multiple Alignment, MA). Έτσι ονομάζουμε την στοίχιση που πραγματοποιείται μεταξύ άνω των δύο αλληλουχιών. Κάποια λογισμικά για πολλαπλή στοίχιση ακολουθούν: MSA (Multiple Sequence Alignment) (209), Clustal W/W2/X (210), T-COFFEE (211). Αλλά και λογισμικά που δουλεύουν με μεθόδους ολικής επαναληπτικής στοίχισης όπως το HMMER (212), που χρησιμοποιεί Hidden Markov Models, και το SAGA (213), που χρησιμοποιεί γενετικούς αλγόριθμους.

Προσεγγίσεις στοίχισης. Έχουμε διαθέσιμα καταλληλότερα εργαλεία από το αργό στην στοίχιση BLAST για βραχύ-μικροανάγνωση, που είναι ο τύπος μικροαναγνώσεων σε πολλές πλατφόρμες NGS. Οι αλγόριθμοι στα NGS εργαλεία βασίζονται στην κατάταξη σε πίνακα (indexing) για την ταχύτερη χαρτογράφηση. Ένας τρόπος είναι μέσω της κατάταξης του γενώματος αναφοράς σε ένα πίνακα κατακερματισμού, π.χ. το εργαλείο SOAP και το εργαλείο MAQ[26] (Mapping and Assembly with Qualities) (214). Ένας άλλος είναι ο μετασχηματισμός Burrows-Wheeler, π.χ. το εργαλείο Bowtie, το εργαλείο SOAP2 και το εργαλείο BWA.

BWA[27] (Burrows-Wheeler Aligner). Ο μετασχηματισμός Burrows-Wheeler (Burrows-Wheeler Transform, BWT), που δημοσιεύτηκε το 1994, χρησιμοποιείται σε αυτό κ.α. εργαλεία για την ταχύτητα που πραγματοποιεί την διαδικασία στοίχισης (215). Τα εργαλεία που χρησιμοποιούν τον BWT δεν έχουν τόση ευαισθησία, που σημαίνει πιθανές αποκλίσεις κατά την στοίχιση, όσο αυτά που χρησιμοποιούν πίνακες κατακερματισμού.

[26] http://maq.sourceforge.net
[27] http://bio-bwa.sourceforge.net

Bowtie2. Είναι ένα από τα προτιμώμενα λογισμικά για NGS στοιχίσεις. Η κατάταξη στον πίνακα γίνεται με τον FM-δείκτη (Fulltext Minute-index) και παρέχει ταχύτατες στοιχίσεις, αφού κρατάει σε χαμηλά επίπεδα την χρήση της προσωρινής μνήμης με την επιτάχυνση SIMD. Ο αλγόριθμος αποτελείται από το στάδιο χωρίς κενά (ungapped) εύρεσης "σπόρων" (seeds) και από το στάδιο επέκτασης με κενά (gapped), το οποίο χρησιμοποιεί δυναμικό προγραμματισμό και βελτιώνει την ευαισθησία του αλγορίθμου (216).

SOAP2[28] *(Short Oligonucleotide Analysis Package)*. Το εργαλείο SOAP2 είναι η δεύτερη έκδοση του εργαλείου SOAP (217). Ένα βασικό πλεονέκτημα του εργαλείου είναι ότι αναφέρει και τις στοιχίσεις των μικροαναγνώσεων που έχουν χαρτογραφηθεί σε πολλαπλές θέσεις (218).

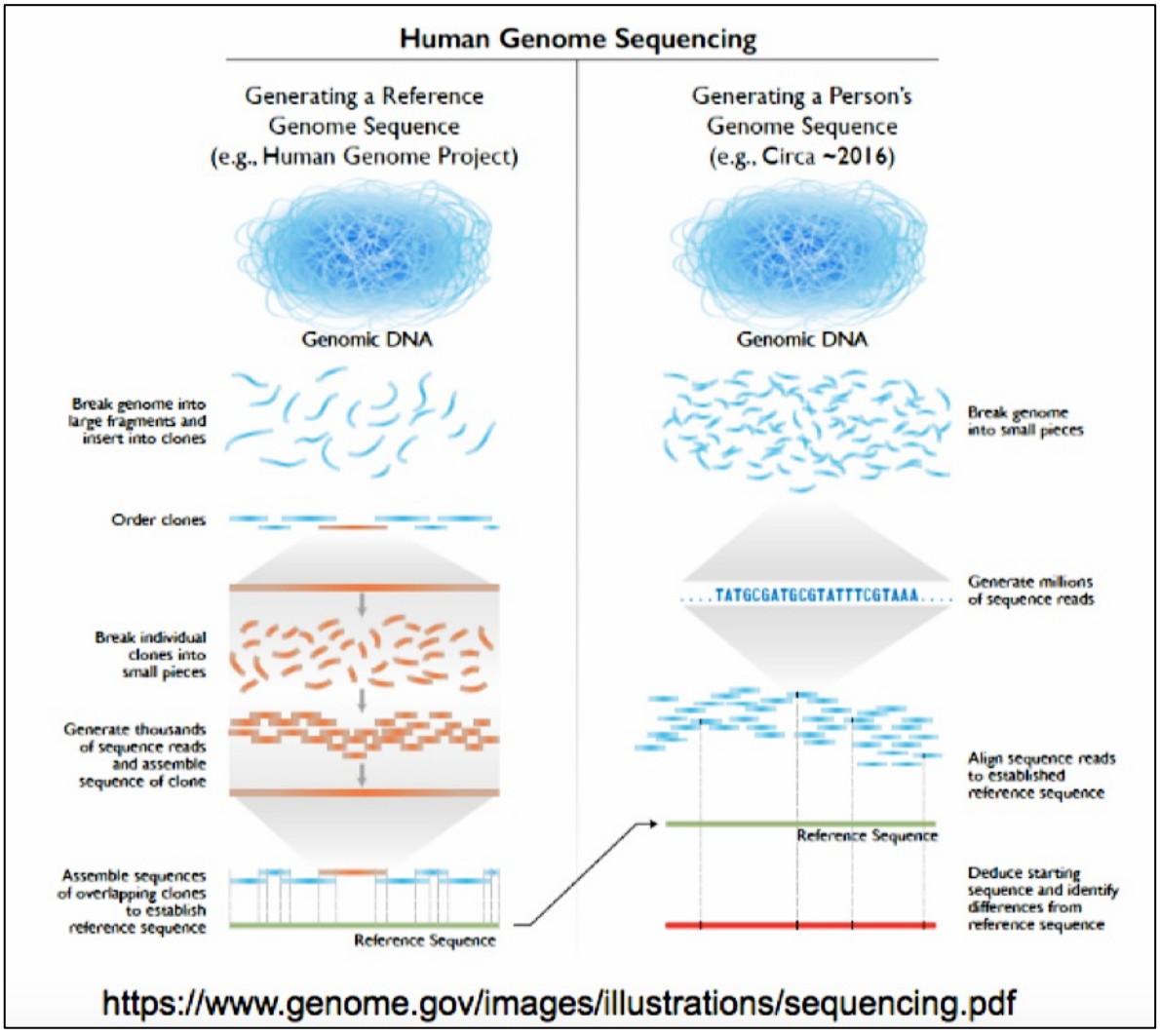
Η επιλογή ενός από τα πολυάριθμα εργαλεία που υπάρχουν έχει να κάνει με τις δυνατότητες του κάθε εργαλείου και την πλατφόρμα στην οποία κάνουμε την αλληλούχιση. Ακόμα εξαρτάται από το τι θέλουμε να αναλύσουμε, δηλαδή DNA ή RNA ή miRNA ή μεθυλίωση DNA.

Στην **Εικόνα 25**, παρατηρούμε ότι στην αριστερή πλευρά έχουμε αλληλουχίσει ένα γένωμα αναφοράς, πρόκειται για το ανθρώπινο γένωμα που έγινε διαθέσιμο πρώτη φορά με την ολοκλήρωση του HGP. Πρόκειται για μια de novo αλληλούχιση που πρέπει να πραγματοποιούμε για οργανισμούς που δεν έχουν γένωμα αναφοράς. Πιο συγκεκριμένα, πρόκειται για συναρμολόγιση του γενώματος, δηλαδή η δημιουργία ενός συναινετικού γενώματος από το σύνολο των ξεχωριστών μικροαναγνώσεων. Στην δεξιά πλευρά, έχουμε την χαρτογράφηση των αλληλουχιών στο γένωμα αναφοράς, δηλαδή πραγματοποιούμε αναλληλούχιση. Πιο συγκεκριμένα, παίρνουμε ως δεδομένες τις θέσεις των γονιδίων στο γένωμα αναφοράς και πραγματοποιούμε στοίχιση των μικροαναγνώσεων του υπο εξέταση γενώματος σε αυτό. Οπότε, μπορούμε στην συνέχεια να ποσοτικοποιήσουμε τα επίπεδα της διαφορικής γονιδιακής έκφρασης (90).

Παρομοίως, η τεχνολογία RNA-Seq βασίζεται στην αλληλούχιση μικρών θραυσμάτων συμπληρωματικού DNA (cDNA), το οποίο προέρχεται από RNA. Αντί να υβριδοποιούνται τα θραύσματα cDNA σε μικροσυστοιχίες DNA, τα θραύσματα αλληλουχοποιούνται

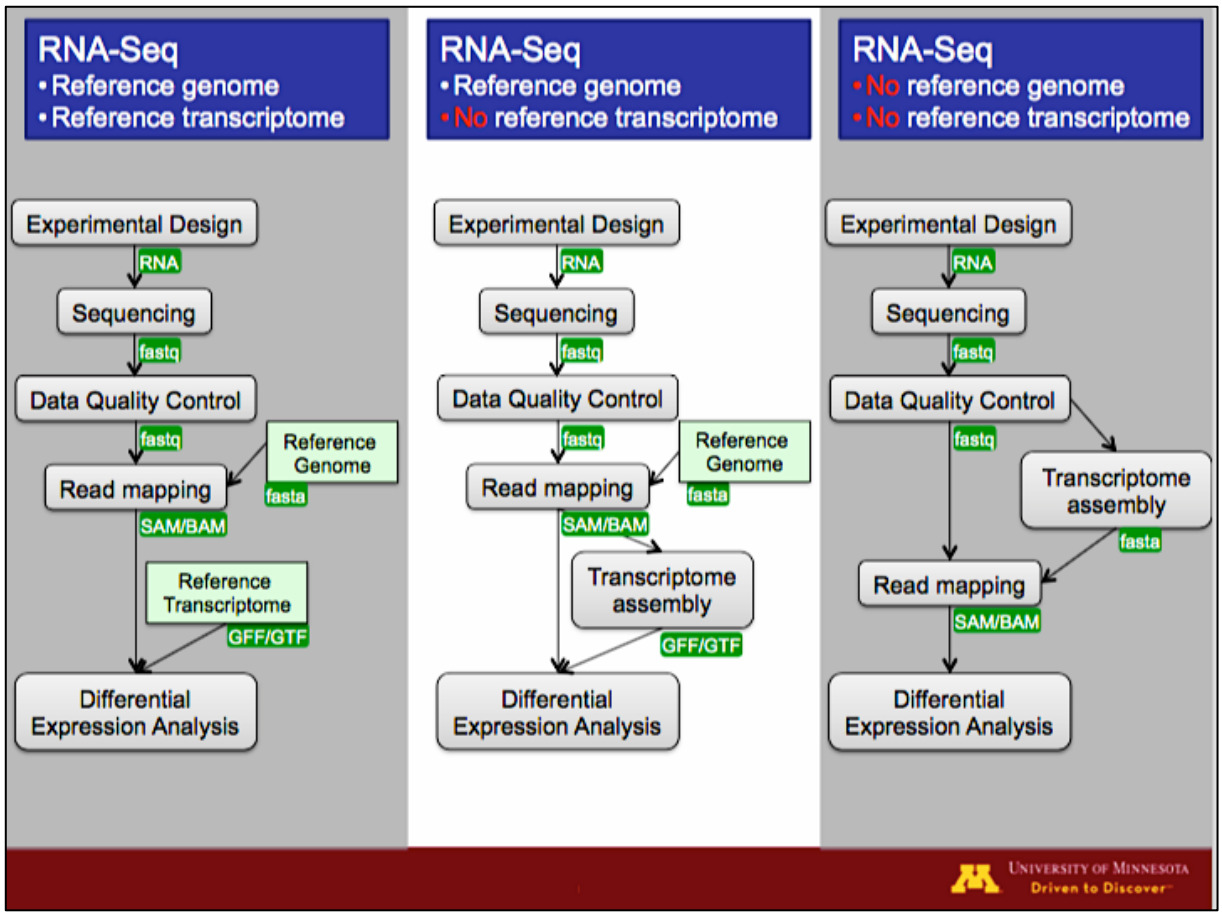
[28] http://soap.genomics.org.cn/soapaligner.html

απευθείας, χαρτογραφούνται στο γένωμα ή στο χαρακτηρισμένο μεταγράφωμα, όπου οι χαρακτηρισμένες αλληλουχίες αναφοράς (annotations) είναι διαθέσιμες για εξαγωγή από διάφορες διαδικτυακές βάσεις δεδομένων, όπως η Ensembl, η UCSC και η NCBI σε αρχεία .GTF (Gene Transfer Format). GFF (είναι το απλοποιημένο-flattened .GTF), με υπολογιστικές μεθόδους και υπολογίζονται οι επικαλυπτόμενες μικροαναγνώσεις. Έτσι, αναγνωρίζεται κάθε θραύσμα και επανασυναρμολογούνται τα θραύσματα σε ολόκληρα μετάγραφα με λογισμικά όπως το Cufflinks (219), το οποίο χρησιμοποιεί την προσέγγιση EM (expectation–maximization) και λαμβάνει υπόψη διάφορες αποκλίσεις, όπως την ανομοιόμορφη κατανομή των μικροαναγνώσεων κατά μήκος του γονιδίου, με στόχο την αξιόπιστη ποσοτικοποίηση των επιπέδων γονιδιακής έκφρασης των υπό εξέταση δειγμάτων. Το Cufflinks είναι ακόμα κατάλληλο για την βελτίωση ενός προϋπάρχοντος μεταγραφώματος αναφοράς, όμως η ανακατασκευή ενός μεταγραφώματος πρέπει να είναι η έσχατη λύση γιατί δεν έχει καλή ακρίβεια ούτε καλή ευαισθησία (220).



Εικόνα 25. Η αλληλούχιση DNA ενός γενώματος αναφοράς και η αναλληλούχιση.

Βάσεις δεδομένων νουκλεοτιδικών ακολουθιών (Nucleotide Sequence Database). Η INSDC[29] (International Nucleotide Sequence Database Collaboration) είναι η σύμπραξη τριών πολύ γνωστών βάσεων δεδομένων που χρησιμοποιούνται καθημερινά από χιλιάδες επιστήμονες είναι η Genbank[30], η EMBL-Bank[31] και η DDBJ[32]. Η GenBank ανήκει στην ομάδα του NCBI Entrez και περιέχει αποθηκευμένες αλληλουχίες DNA, RNA και πρωτεϊνών. Μπορεί να διαθέτει πολλαπλές αλληλουχίες τις ίδιας περιοχής, δηλαδή διαφορετικά δείγματα από την ίδια αλληλουχημένη περιοχή και μπορεί να χρησιμοποιηθεί για την αναγνώριση γενετικών διαφοροποιήσεων.



Εικόνα 26. Η RNA-Seq ροή έργου με γένωμα και μετάγραφο αναφοράς (221).

[29] https://www.insdc.org/
[30] https://www.ncbi.nlm.nih.gov/genbank/
[31] https://www.ebi.ac.uk/
[32] https://www.ddbj.nig.ac.jp/index-e.html

Αντιστοίχως, το αποθετήριο αρχείων SRA (Sequence Read Archive) παρέχει πρόσβαση σε δεδομένα που προέρχονται από DNA-Seq και RNA-Seq, τα οποία υπάρχουν στην Genbank, την DDBJ και την ENA[33] βάσεις δεδομένων.

Βάση δεδομένων RefSeq (Reference Sequence Database). Η RefSeq, επίσης μέρος της NCBI, έχει μόνο ένα δείγμα κάθε βιολογικού μορίου, κάποιες εγγραφές είναι αντίτυπα από την GenBank Records και άλλες προέρχονται από διαφορετικές πηγές (222), π.χ. FlyBase. Άλλες βάσεις δεδομένων είναι η βάση γενωμικών δεδομένων GOLD (Genome On Line Database[34]) και η βάση γενωμικών δεδομένων του NCBI[35], για να αναφέρουμε κάποιες.

GEO (Gene Expression Omnibus[36]). Η GEO περιέχει χιλιάδες δεδομένα από μελέτες και αποτελέσματα γονιδιακής έκφρασης (223).

R/Bioconductor. Το λογισμικό Bioconductor, που είναι γραμμένο στην στατιστική γλώσσα R, είναι ένα σημαντικό εργαλείο για την γενωμική και όχι μόνο. Περιέχει πολλές εκατοντάδες πακέτα, με τον δικό του ρόλο το καθένα. Με αυτό μπορούμε να εξερευνήσουμε γενωμικές βάσεις δεδομένων, να εξάγουμε δεδομένα από αυτές και επιτρέπουν την διαδραστικότητα μεταξύ γενωμικών φυλλομετρητών. Η ανάλυση και η οπτικοποίησή των γενωμικών δεδομένων γίνονται από κάποιο άλλο πακέτο, σε πολλά από τα οποία έχουμε ήδη αναφερθεί (224).

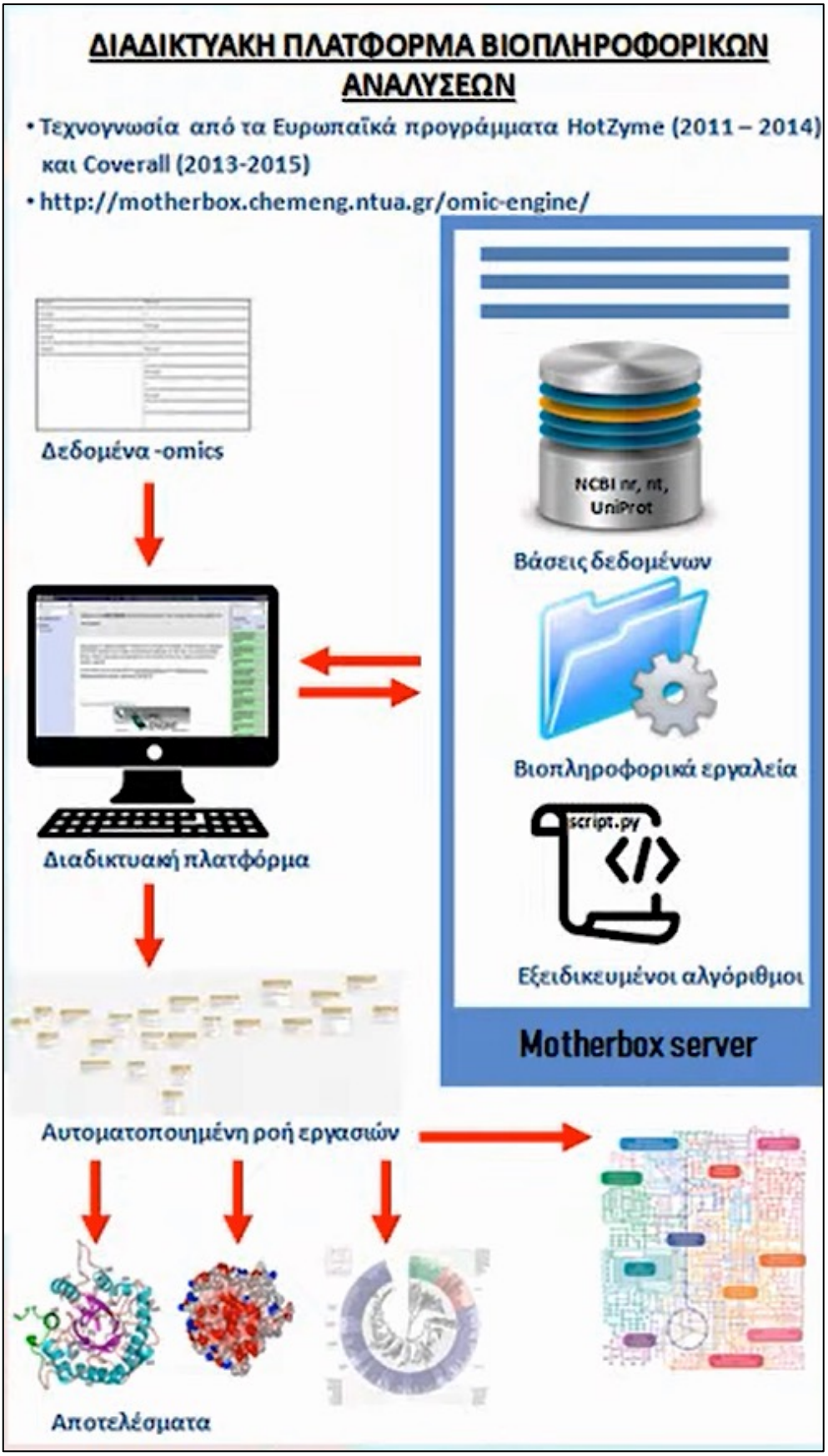
Galaxy[37]. Μέσω των εφαρμογών Galaxy (225), μπορεί ο βιοπληροφορικός να δημιουργήσει ένα διαδικτυακό γραφικό περιβάλλον, όπως στην **Εικόνα 27**. Πρόκειται για μια πλατφόρμα, διαδικτυακή και λογισμικό, ροής εργασιών στην οποία είναι δυνατή η ανάλυση σε γενωμικές βάσεις δεδομένων, ακόμα και με συμπλήρωση μέσω διαφορετικού τύπου δεδομένων, π.χ. δεδομένα που προέρχονται από μικροσυστοιχίες μαζί με δεδομένα από NGS. Τα δεδομένα εισάγονται με το εργαλείο Get Data, από τοπικό αρχείο ή με αυτοματοποιημένο τρόπο από κάποια βάση δεδομένων, π.χ. τον φυλλομετρητή UCSC Genome Browser, τον γενωμικό φυλλομετρητή Ensembl κ.α. . Στην συνέχεια γίνεται η προεπεξεργασία των δεδομένων, η

[33] https://www.ebi.ac.uk/ena/browser/home
[34] https://gold.jgi.doe.gov
[35] https://www.ncbi.nlm.nih.gov/genome
[36] https://www.ncbi.nlm.nih.gov/geo
[37] https://www.galaxyproject.org

στοίχιση με διάφορους αλγορίθμους, η στατιστική ανάλυση και η οπτικοποίησή τους. Αυτό μπορεί να γίνει σε παρτίδες για διάφορα πειράματα και αναλύσεις, έτσι ώστε τα αποτελέσματα της ανάλυσης να είναι αναπαράξιμα. Η χρήση της πλατφόρμας βελτιώνεται μέσω μιας μεγάλης κοινότητας υποστήριξης (224). Το πακέτο παραγωγής βιοπληροφορικών εργαλείων του Galaxy (Galaxy Tool Factory), διευκολύνει τις δοκιμές για την ανάπτυξη σε διάφορες προγραμματιστικές γλώσσες ενός νέου εργαλείου.

DAVID (*Database for Annotation, Visualization, and Integrated Discovery*) [38]. Η βάση δεδομένων DAVID κατασκευάστηκε για να γίνεται με ευκολία η ανάλυση μιας μεγάλης λίστας δεδομένων, αλλά και να διευκολύνει τον χαρακτηρισμό της λειτουργίας τους. Γι'αυτό παρέχει διάφορα εργαλεία εξόρυξης πληροφοριών με περιγραφικά διαγράμματα για την λειτουργία τους. Ακόμα, τα εργαλεία οπτικοποίησης, με δυνατότητα γονιδιακού χαρακτηρισμού, λειτουργικής ταξινόμησης και κατασκευή βιοχημικών μονοπατιών (227).

[38] http://david.ncifcrf.gov/



Εικόνα 27. Πλατφόρμα Galaxy (226).

4.2. Στατιστική Ανάλυση και Υπολογιστικές Μέθοδοι με Τεχνικές Χρήσης Χρωστικών (π.χ. Πλατφόρμα Illumina) και με άλλες Τεχνικές (π.χ. Τεχνική Ιόντων Υδρογόνου Ion Torrent)

4.2.1. Προεπεξεργασία Δεδομένων

Το RNA είναι δυναμικό και το DNA είναι στατικό, γι'αυτό τον λόγο έχουμε διαφορετικό τρόπο επεξεργασίας των αλληλουχιών υπο εξέταση, για παράδειγμα όταν αλληλουχούμε ένα γένωμα ή μεταγράφομα αναφοράς. Αυτό σημαίνει ότι έχουμε άλλη επεξεργασία όταν εξετάζουμε μια αλληλουχία DNA (π.χ. δεν έχουμε θέμα με την αλληλοεπικάλυψη εξονίων όπως στην ανάλυση mRNA), RNA, miRNA και μεθυλίωση αλληλουχίας DNA (Differentially Methylated Genes, DMGs), όπου στις δυο τελευταίες περιπτώσεις έχουμε παραπάνω βήματα επεξεργασίας. Γι'αυτό τον λόγο έχουν αναπτυχθεί πολλά λογισμικά που ειδικεύονται στην μια ή την άλλη κατηγορία αλληλουχιών υπό εξέταση. Θα δούμε σε αυτό το κεφάλαιο κάποιες από αυτές τις διαφορές.

Οι βάσεις δεδομένων για δεδομένα που παρήχθησαν με τεχνικές NGS επιτρέπουν σε βιοπληροφορικούς και βιολόγους πρόσβαση σε προϋπάρχουσες πληροφορίες και οργανωμένη προσθήκη νέων δεδομένων. Η βιοπληροφορική επιστήμη έχει τον βασικό ρόλο να αναπτύξει εφαρμογές και αλγορίθμους για την βέλτιστη ανάλυση των βιολογικών δεδομένων. Ο βασικός τύπος αρχείου για την αποθήκευση των παραγόμενων αλληλουχιών μικρού μήκους μικροαναγνώσεων είναι το αρχείο FastQ, το οποίο αποτελείται από την αλληλουχία και τις ASCII-κωδικοποιημένες ποιοτικά βαθμολογίες για κάθε νουκλεοτίδιο που αλληλουχοποιείται (228).

Έλεγχος Ποιότητας και Καθαρισμός Δεδομένων. Μετά την αλληλούχιση γίνεται καθαρισμός των δεδομένων και QC. Πριν οι αλληλουχοποιημένες μικροαναγνώσεις χρησιμοποιηθούν περαιτέρω σε κάποια αναλυτική ροή εργασίας της διαδικασίας (workflow), πρέπει να ελεγχθεί η ποιότητα των μικροαναγνώσεων και του κάθε νουκλεοτιδίου σε αυτό ξεχωριστά, δηλαδή γίνονται έλεγχοι στα πρωτογενή δεδομένα και όταν δεν τηρούνται συγκεκριμένες ποιοτικές συνθήκες επισημαίνονται (flagged) και αποκλείονται. Ο έλεγχος ποιότητας και ο καθαρισμός δεδομένων μπορούν να γίνουν πάνω από μια φορά, μέχρι να έχουμε μόνο καλής ποιότητας δεδομένα. Η ποιότητα, όμως, υπολογίζεται διαφορετικά σύμφωνα με τον

αλληλουχοποιητή και υπάρχουν πολλά εργαλεία QC για έλεγχο της ποιότητας των μικροαναγνώσεων. Αυτό είναι δυνατό με διάφορα εργαλεία, όπως το FastQC[39] (είναι γραμμένο σε Java), το Fastp, το MultiQC, το Ion Reporter (για πλατφόρμα Ion Torrent), το PRINSEQ[40] (229), το FastX-Toolkit[41] (που είναι κατάλληλο για βραχύ-μικροανάγνωση) και το TagCleaner (230). Για παράδειγμα, το πολύ διαδεδομένο εργαλείο FastQC διαθέτει πολλές απλές εφαρμογές για τον έλεγχο της ποιότητας των δεδομένων και στο τέλος παράγει μια αναφορά με τα αποτελέσματα με πίνακες σε ένα αρχείο HTML. Ακόμα, καταγράφονται στατιστικά στοιχεία των δεδομένων, όπως ο αριθμός και το μήκος των μικροαναγνώσεων, καθώς και η δυνατότητα να δούμε τα σκορ ποιότητας της αλληλουχίας σε κάθε βάση και διαγράμματα με την κατανομή των σκορ ποιότητας. Παρόλο που μπορούμε να χρησιμοποιήσουμε το εργαλείο FastQC για τις πλατφόρμες που λειτουργούν με την τεχνική Ion Torrent, υπάρχουν λογισμικά σχεδιασμένα για τις συγκεκριμένες πλατφόρμες που ίσως προσφέρουν καλύτερη αξιοπιστία, δηλαδή το εργαλείο Ion Reporter και το Torrent Suite, που παρέχονται από την ThermoFisher. Γενικά, ο έλεγχος ποιότητας γίνεται σε διάφορα στάδια της ανάλυσης των NGS δεδομένων.

Όταν η στάνταρ QC έλεγχοι ολοκληρώνονται, οι ερευνητές που μελετούν τα γενωμικά δείγματα, χρησιμοποιούν άλλη μια σειρά εργαλείων, όπως το DeconSeq, που χρησιμοποιεί φίλτρα προεπεξεργασίας για να εξαλείψει την επιμόλυνση από τις μικροαναγνώσεις γενωμικών αλληλουχιών. Οι επιστήμονες που ερευνάνε τον καρκίνο μπορούν να χρησιμοποιήσουν το εργαλείο ContEst για την εκτίμηση του ποσοστού της επιμόλυνσης διασταυρωμένων δειγμάτων στα δεδομένα αλληλούχισης (231). Το εργαλείο ContEst χρησιμοποιεί ένα Μπευσιανό πλαίσιο για την εκτίμηση των επιπέδων επιμόλυνσης από βασισμένους σε πίνακες γενότυπους και μικροαναγνώσεων αλληλούχισης (228).

Base calling. Στο στάδιο της προεπεξεργασίας, οι πλατφόρμες νέας γενιάς παράγουν διαφορετικής ποιότητας κλήσης βάσεων (base calling), που είναι η διαδικασία μετατροπής των εντάσεων για την πλατφόρμα Illumina ή την ανίχνευση των αλλαγών στο pH στα Ion Torrent και την αντίστοιχη ανίχνευση των μορίων με άλλες τεχνικές, σε ψηφιακές τιμές νουκλεοτιδικών σημάτων. Η κλήση βάσεων γίνεται από την πλατφόρμα Illumina με το

[39] https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
[40] https://edwards.sdsu.edu/cgi-bin/prinseq/prinseq.cgi
[41] http://hannonlab.cshl.edu/fastx_toolkit/

Illumina CASAVA (Consensus Assessment of Sequence And Variation), που είναι ένα εταιρικό εργαλείο και ανήκει στο λογισμικό ανάλυσης αλληλούχισης της Illumina (Illumina Genome Analyzer) και πραγματοποιεί στοίχισης της εξεταζόμενης αλληλουχίας σε ένα γένωμα αναφοράς και στην συνέχεια την ποσοτικοποίηση των μικροαναγνώσεων (read counting) (232). Επισημαίνεται ότι η τεχνική Ion Torrent έχει αρχικά πιο απλή διαδικασία από τις τεχνικές με χρωστικές, αφού το κάθε σήμα ψηφιοποιείται με την σειρά του χωρίς να χρειάζεται το βήμα λήψης της εικόνας. Αυτό όμως δεν σημαίνει ότι είναι απαραίτητα πιο εύκολη από τις τεχνικές χρωστικών, αφού η αποκωδικοποίηση του σήματος είναι τεχνικά απαιτητική (233). Ονομάζουμε βιβλιοθήκη τις εκατομμύρια μικροαναγνώσεις που παράγονται από ένα συγκεκριμένο δείγμα.

Ακόμα, ανάλογα με την χημεία τους (με χρωστικές ή με ημιαγωγούς κ.α.), οι πλατφόρμες εμφανίζουν διάφορα σφάλματα στην πρωτογενή και δευτερογενή ανάλυση, γι'αυτό είναι απαραίτητο να υπολογίζουμε την ποιότητα των βάσεων μέσω της ίδιας της πλατφόρμας και να αναφέρουμε μαζί με την κάθε βάση και το σκόρ της. Για παράδειγμα, με την τεχνική χρωστικών, π.χ. Illumina, μπορεί να έχουμε κλήση βάσεων κακής ποιότητας λόγω του φθορισμού στο δείγμα μας ή κακής ποιότητας μικροαναγνώσεις λόγω της PCR, που έχουμε αναφέρει ότι ενισχύει και τα σφάλματα. Στο ίδιο παράδειγμα πρέπει να πραγματοποιήσουμε καθαρισμό λόγω αναδιπλασιασμού PCR, έτσι ώστε να αποκλείσουμε παρόμοιες μικροαναγνώσεις και ζεύγη μικροαναγνώσεων που μπορεί να προέρχονται από την PCR/emPCR που χρησιμοποιείται στις περισσότερες πλατφόρμες και πειράματα που χρησιμοποιούν τεχνικές χρωστικών και προέρχονται από την ίδια περιοχή του γενώματος.

Φιλτράρισμα ή Ξάκρισμα (Trimming) Γονιδίων. Το φιλτράρισμα γονιδίων με χαμηλά ή μηδενικά επίπεδα έκφρασης και ο αποκλεισμός τους είναι χρήσιμο εργαλείο για την σωστότερη ερμηνεία των στατιστικών αποτελεσμάτων, δηλαδή απορρίπτουμε αυτά που δεν ικανοποιούν συγκεκριμένα κριτήρια αξιοπιστίας. Δεν υπάρχουν συγκεκριμένοι κανόνες αλλά είναι σύνηθες να απορριφθούν τα γονίδια με λιγότερο από 5 CPM σε όλα τα δείγματα. Αντίθετα με το φιλτράρισμα, στο ξάκρισμα γονιδίων, δηλαδή η υπολογιστική διεργασία αφαίρεσης των αλληλουχιών προσαρμογέα από τις μικροαναγνώσεις μετά το πέρας της αλληλούχισης, δεν απορρίπτουμε ολόκληρη την μικροανάγνωση, αλλά μόνο τις βάσεις της με κακή ποιότητα, εκτός αν όλες έχουν κακή ποιότητα. Οι συνθετικές ολιγονουκλεοτιδικές αλληλουχίες που επιμολύνουν τις μικροαναγνώσεις και οι κακής ποιότητας βάσεις που

βρίσκονται στα άκρα των αλληλουχιών είναι κάποιες από τις προβληματικές περιοχές της αλληλουχίας που αποκλείουμε σε αυτό το στάδιο.

Από την άλλη πλευρά το ξάκρισμα μειώνει το μήκος των μικροαναγνώσεων, το οποίο μπορεί να μην είναι ιδανικό για όλες τις εφαρμογές. Ακόμα πρέπει να προσέξουμε όταν κάνουμε ξάκρισμα σε ανάλυση PE, αφού πρέπει να προσέξουμε ότι οι μικροαναγνώσεις στα δύο εξαρτημένα αρχεία πρέπει να είναι αντίστοιχες, οπότε αν αφαιρέσουμε μια μικροανάγνωση από το ένα αρχείο πρέπει να αφαιρέσουμε την αντίστοιχη μικροανάγνωση και από το άλλο αρχείο, ακόμα και αν αυτό ήταν καλής ποιότητας. Ένα ακόμα βήμα που πρέπει να πραγματοποιήσουμε είναι η συγχώνευση των αλληλοεπικαλυπτόμενων ζευγών, για να δημιουργήσουμε μια συνδυασμένη αλληλουχία χωρίς ανώφελες πληροφορίες. Κάποια πολύ γνωστά εργαλεία για το ξάκρισμα είναι το Trimmomatic (234), το Flexbar (235), το Skewer, το PRINSEQ++ (236), το NGS QC Toolkit (είναι γραμμένο σε Perl) (237), PEAT (για ανάλυση PE) (238) και το TagCleaner. Το Trimmomatic και το Skewer λαμβάνουν υπόψη και τις αλληλουχίες προσαρμογέων και την ποιότητα των αλληλουχιών, επίσης είναι κατάλληλα και για ανάλυση PE.

$$P = 10^{\frac{-Q}{10}}$$

Εξίσωση 74. Πιθανότητα σφάλματος κλήσης βάσεων, όπου Q×2=ASCII.

Στοίχιση, Χαρτογράφηση και Συναρμολόγηση (Sequencing alignment, mapping and assembly). Για το στάδιο της στοίχισης, αρχικά πραγματοποιούμε την χαρτογράφηση σε ένα γένωμα αναφοράς και, στην συνέχεια, ταιριάζουμε τις στοιχισμένες μικροαναγνώσεις με τους γονιδιακούς χαρακτηρισμούς τους, έτσι ώστε να τις ποσοτικοποιήσουμε σε κάθε γονίδιο. Ενώ αν το εξεταζόμενο είδος έχει ποιοτικό χαρακτηρισμένο μεταγράφωμα (annotated transcriptome), τότε η στοίχιση γίνεται σε αυτό (239). Τα λογισμικά στοίχισης βραχέων μικροαναγνώσεων για NGS πλατφόρμες, που αναφέρθηκαν στην αρχή του κεφαλαίου, είναι πολλά και είναι δύσκολο πολλές φορές να επιλεγεί το σωστό λογισμικό στοίχισης από τον βιολόγο. Επιπροσθέτως υπάρχουν πολυάριθμες δημοσιεύσεις με χρήση του κάθε εργαλείου στις διάφορες πλατφόρμες και γίνεται ακόμα πιο εύκολη η επιλογή όταν χρησιμοποιείς πλατφόρμες όπως το DAVID και το Galaxy, όπου έχουν φτιαχτεί αυτοματοποιημένες ροές εργασιών για το κάθε είδος ανάλυσης. Ακόμα και σε αυτό το στάδιο είναι πολύ σημαντικό να πραγματοποιήσουμε έλεγχο της ποιότητας χαρτογράφησης.

Στην δημοσίευση των (240), αναφέρονται κάποια εργαλεία που ξεχώρισαν για ανάλυση WGS/WES. Στην πλατφόρμα Illumina είναι τα εργαλεία BWA, Bowtie2 και SOAP2. Ενώ στην πλατφόρμα SOLiD είναι τα εργαλεία BFAST (241) και SHRiMP (242), αλλά και η νεότερη έκδοσή του SHRiMP2[42] η οποία έχει δείξει καλά αποτελέσματα και στην στοίχιση του γενώματος δεδομένων miRNA (243).

Επιπροσθέτως, για αυτοματοποιημένες ροές εργασιών μεθυλίωσης DNA με ανάλυση BS-Seq ξεχωρίζουν τα εργαλεία Bismark (244), BS-seeker2 (245), Pash3 (246), BSMAP (247) και BRAT (248).

Κάποια εργαλεία στοίχισης λαμβάνουν υπόψη την εναλλακτική συρραφή (splicing) του RNA, που μας αφορά ιδιαιτέρως όταν στοιχίζουμε μικροαναγνώσεις από RNA-seq σε γένωματα αναφοράς, αφού αυτά περιέχουν εσόνια. Τέτοια είναι το STAR (Spliced Transcript Allignment to a Reference) (249), το Tophat2 (250), το Hisat2 (251), το GSNAP (252) και το Rsubread/Subread (253), το οποίο είναι διαθέσιμο στην R (239).

Ένα άλλο θέμα που πρέπει να λάβουμε υπόψη είναι ότι με την τεχνική Ion Torrent μπορεί να έχουμε πολύκλωνες μικροαναγνώσεις οι οποίες πρέπει να αφαιρεθούν γιατί αυτές έχουν μειωμένο σήμα με συχνότερη ανίχνευση αυτού. Κάποιες πλατφόρμες αλληλουχίζουν τα γενώματα με ιδιαίτερες τεχνικές και γι' αυτό στοιχίζονται οι μικροαναγνώσεις με ειδικούς αλγορίθμους, οι οποίοι είναι φτιαγμένοι για τις τεχνικές αυτές και τα πρότυπα σφάλματος που χαρακτηρίζουν την κάθε μια. Για παράδειγμα, η στοίχιση στην πλατφόρμα Ion Torrent πραγματοποιείται με το Tmap[43] (Torrent Mapping Alignment Program), που είναι εργαλείο του Torrent Suite[44] (λογισμικό ελεύθερης πρόσβασης από την εταιρεία ThermoFisher), και η στοίχιση στην πλατφόρμα SOLiD πραγματοποιείται με το LifeScope (εταιρικό λογισμικό που ανήκει στην ThermoFisher) (254). Ένα άλλο εργαλείο, κατάλληλο για την στοίχιση γενωμικών δεδομένων από την πλατφόρμα Ion Torrent, είναι το MIRA (255). Σε μία σύγκριση στην ανάλυση δεδομένων μεταξύ της πλατφόρμας Illumina και της πλατφόρμας Ion Torrent έδειξε παρόμοια αποτελέσματα στην ανάλυση με τα εργαλεία που χρησιμοποιήθηκαν, δηλαδή το GSNAP, το STAR, και ο συνδυασμός STAR+Bowtie2 (256).

[42] http://compbio.cs.toronto.edu/shrimp/
[43] https://github.com/iontorrent/TMAP
[44] https://github.com/iontorrent/TS

Τα δεδομένα από την πλατφόρμα SOLiD, που έχουν περισσότερα σφάλματα αντικατάστασης ανά μικροανάγνωση στον χρωματικό χώρο, γι'αυτό προτιμάμε εργαλεία με πιο ελαστικά κριτήρια μη ταύτισης, όπως το εργαλείο PerM. Αντίστοιχα, για μικροαναγνώσεις της πλατφόρμας PacBio, η οποία παρουσιάζει μεγάλα ποσοστά σφαλμάτων σε σχέση με άλλες πλατφόρμες, μπορούμε να χρησιμοποιήσουμε το εργαλείο BLASR, το οποίο είναι κατάλληλο για ανάλυση μακρύτερων μικροαναγνώσεων και λαμβάνει υπόψη τις αποκλίσεις indels. Εκτός από τις διαφορές στην διαδικασία κάθε πλατφόρμας και της ποιότητας των μετρήσεων κρίνεται απαραίτητη και η διόρθωση πιθανών σφαλμάτων κατά την διαδικασία κλήσης βάσεων.

Για την στοίχιση σε είδη που δεν υπάρχουν ποιοτικά γενώματα ή μεταγραφώματα αναφοράς μπορούμε να το συναρμολογήσουμε εκ νέου. Σε μια αυτοματοποιημένη ροή εργασιών (pipeline) της εκ νέου αλληλούχισης, η συναρμολόγηση πραγματοποιείται με αρκετά απαιτητικές υπολογιστικές προσεγγίσεις, όπως η κατασκευή γράφων k-μερή (τα k-mers είναι μια υπο-αλληλουχία με μήκος k) de Bruijn για τον τύπο βραχύ-μικροανάγνωση και η κατασκευή συναινετικού γενώματος με βάση τις επικαλύψεις (Overlap Layout Concensus, OLC) για μεγαλύτερου μήκους μικροαναγνώσεις, δηλαδή η κάθε μικροανάγνωση συγκρίνεται με τις υπόλοιπες μικροαναγνώσεις στο δείγμα για την δημιουργία του συναινετικού γενώματος επιλέγοντας την πιο πιθανή νουκλεοτιδική αλληλουχία για κάθε contig και τον ποιοτικό έλεγχο αυτού (257). Για την συναρμολόγηση του μεταγράφου εκ νέου χρησιμοποιούμε εργαλεία όπως το Trinity (258) και το Trans-ABySS (259) και να εκτιμήσουμε τα επίπεδα έκφρασης μέσω αυτής της συναρμολόγησης. Η δημοσίευση των (260), περιέχει έναν περιεκτικό πίνακα με εργαλεία συναρμολόγησης γενώματος φυτού βασισμένα και στις δύο αυτές προσεγγίσεις, συμπεριλαμβανομένου του από ποιες πλατφόρμες NGS μπορούν να προέρχονται οι ακατέργαστες μικροαναγνώσεις που εισάγονται στο κάθε λογισμικό.

Οι συναρμολογημένες μικροαναγνώσεις αποθηκεύονται σε αρχεία SAM (Sequence Alignment Map), τα οποία μετατρέπουμε στην συνέχεια στο δυαδικό αρχείο BAM, για να πετύχουμε καλύτερη αναζήτηση στο περιεχόμενό του και για να χρειαζόμαστε λιγότερο χώρο αποθήκευσης. Οι διαφορές στο μέγεθός τους είναι της τάξης: μέχρι τα 50Gb WES - 800Gb-1Tb WGS για το SAM και 2-10Gb WES -100-300Gb WGS για το BAM. Τα αρχεία BAM περιέχουν τις γενωμικές συντεταγμένες των ταυτίσεων με το γένωμα αναφοράς, καθώς και ανάλυση των όποιων διαφοροποιήσεων με την εξεταζόμενη αλληλουχία, δηλαδή μας

ενδιαφέρει να έχουμε μία γονιδιακή λίστα που να περιέχει πληροφορίες θέσης, όπως το χρωμόσωμα, τον κλώνο, την αρχή και το τέλος του. Δύο εργαλεία για QC μετά την στοίχιση είναι το RSeQC[45] (είναι γραμμένο στην Python, οπτικοποίηση στην R) (261) και το QoRTs (239). Σε αυτό το στάδιο ελέγχουμε τα ποσοστά των στοιχισμένων μικροαναγνώσεων, τα ποσοστά των εξονικών έναντι των εσονικών και διαγονιδιακών περιοχών, την γονιδιακή ποικιλομορφία και την κάλυψη του σώματος των γονιδίων (gene body coverage), δηλαδή την προκατάληψη στο 3'-άκρο με την αποδόμηση του RNA και τον εμπλουτισμό poly(A).

Picard[46]. Το Picard είναι μια σουίτα εργαλείων από το Broad Instutute γραμμένο στην Java. Πιο συγκεκριμένα, το εργαλείο του MarkDuplicates εισάγει αρχεία BAM, στα οποία επισημαίνει και μπορεί να αφαιρέσει τις διπλασιασμένες μικροαναγνώσεις. Αυτό συγκρίνει αλληλουχίες που έχουν τις ίδιες θέσεις 5'-άκρου, αν είναι ίδιες τότε επισημαίνονται ως διπλασιασμένες (262).

Ποσοτικοποίηση με την μέθοδο μετρητών (counts). Ποσοτικοποίηση είναι ουσιαστικά η διαδικασία του υπολογισμού των μικροαναγνώσεων που αλληλουχούνται από ένα γονίδιο ή μετάγραφο. Αντίθετα με την ποσοτικοποίηση στις μικροσυστοιχίες, που έχουμε συνεχείς τιμές, στην NGS έχουμε διακριτές τιμές σε μορφή μετρητών (counts). Στην πραγματικότητα αυτή δεν είναι μια εύκολη σύγκριση λόγω του φαινόμενου των επικαλύψεων. Αυτές μπορεί να είναι επικαλυπτόμενα χαρακτηρισμένα γονίδια ή επικαλυπτόμενα χαρακτηρισμένα εξόνια από διαφορετική ισομορφές του μεταγράφου ενός γονιδίου ή επικαλυπτόμενα χαρακτηρισμένα γονίδια από συμπληρωματικούς κλώνους DNA, αν δεν υπάρχει έλεγχος ποια κατεύθυνση έχει ο κάθε κλώνος. Σε περιπτώσεις επικαλύψεων, τα εργαλεία συνήθως παρέχουν την επιλογή στον χρήστη για το σε ποιες καταστάσεις θα προσμετρηθεί η μικροανάγνωση (239). Γενικότερα όταν πραγματοποιούμε ποσοτικοποίηση πρέπει να ξέρουμε πώς το εργαλείο που χρησιμοποιούμε χειρίζεται το επικαλυπτόμενο μέγεθος, δηλαδή αν καλύπτεται ολόκληρη ή μερικώς η μικροανάγνωση, οι μικροαναγνώσεις που βρίσκονται σε πολλές θέσεις του γενώματος (multi-mapping reads), οι μικροαναγνώσεις που επικαλύπτουν πολλαπλά γενωμικά γνωρίσματα του ίδιου είδους και οι μικροαναγνώσεις που επικαλύπτουν τις περιοχές εσονίων.

[45] http://reseqc.sourceforge.net/
[46] https://broadinstitute.github.io/picard/

Η μέθοδος μετρητών σε γονιδιακό επίπεδο παρουσιάζεται σε πίνακα έκφρασης μετρητών, όπως φαίνεται στην **Εξίσωση 75**, ο οποίος αναπαριστά την γονιδιακή έκφραση. Σε αυτή την εργασία μας ενδιαφέρει η εύρεση DEGs, οπότε παρόλο που υπάρχουν και άλλοι τρόποι να πραγματοποιήσουμε ποσοτικοποίηση του πειράματος, τα αποτελέσματα σε γονιδιακό επίπεδο είναι ο πιο κατάλληλος και εύκολα εφαρμόσιμος τρόπος, ειδικά αν το βάθος κάλυψης δεν είναι πολύ υψηλό στο πείραμά μας.

Tximport. Όταν έχουμε αποτελέσματα σε επίπεδο μεταγράφου χρησιμοποιούμε το εργαλείο Tximport, π.χ. αν έχουμε 3 διαφορετικές ισομορφές, τότε αυτές θεωρούνται 3 μετάγραφα του γονιδίου. Το εργαλείο αυτό είναι κατάλληλο για την μετατροπή των μη κανονικοποιημένων δεδομένων από αποτελέσματα πίνακα μετρητών σε επίπεδο μεταγράφου σε αποτελέσματα σε επίπεδο γονιδίου. Σε αυτή την περίπτωση από τα ποσοτικοποιημένα δεδομένα στον πίνακα εξάγουμε τις λίστες DGE μέσω στατιστικών μεθόδων.

$$c = \begin{pmatrix} c_{11} & \mathrm{K} & c_{1n} \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ c_{m1} & \mathrm{L} & c_{mn} \end{pmatrix}$$

Εξίσωση 75. Πίνακας μετρητών/Counts (read count table), όπου κάθε γραμμή είναι ένα από τα δεκάδες χιλιάδες γονίδια (1,2...,*m*), μια στήλη είναι ένα δείγμα (1,2,...*n*) και η τιμή κάθε εγγραφής είναι το παρατηρούμενο πλήθος των μικροαναγνώσεων που χαρτογραφούνται σε αυτό το γονίδιο του συγκεκριμένου δείγματος. Ακόμα, ο πίνακας μπορεί να περιέχει, στην πρώτη στήλη με μετάθεση των υπόλοιπων στοιχείων κατά μια στήλη, την λίστα γενωμικών χαρακτηριστικών (conditions ή genomic features), η οποία συμβολίζεται με $f=[f_1 \ldots f_m]^T$.

Υπάρχουν εργαλεία, όπως το STAR που αναφέρθηκε στο στάδιο της στοίχισης, τα οποία μπορούν, στη συνέχεια, να ολοκληρώσουν και το στάδιο της ποσοτικοποίησης. Ένα δημοφιλές εργαλείο για την ποσοτικοποίηση, γραμμένο στην python, είναι το HTSeq. Ένα άλλο δημοφιλές λογισμικό για την ποσοτικοποίηση είναι το featureCounts (263), το οποίο μπορεί να ποσοτικοποιήσει τις μικροαναγνώσεις που προέρχονται από RNA-Seq ή γενωμικό DNA-Seq. Χρησιμοποιεί διάφορους τόπους για να αντιστοιχίσει αξιόπιστα τις μικροαναγνώσεις σε γενωμικά χαρακτηριστικά (features), όπως τον κατακερματισμό χρωμοσώματος και τον έλεγχο του αριθμού των τετραγώνων σε σχέση με τα χαρακτηριστικά (feature blocking). Είναι εξαιρετικά γρήγορο όταν έχουμε μεγάλα δεδομένα και είναι προσβάσιμο είτε από το τερματικό είτε από το πακέτο Rsubread/Subread (264). Άλλα πακέτα του R/Bioconductor για να βρούμε τον πίνακα μετρητών C είναι το GenomicAlignments

(265) και το EasyRNASeq (266). Το εργαλείο DEXSeq χρησιμοποιείται για την ποσοτικοποίηση της WES ανάλυσης σε ένα πίνακα μετρητών εξονίων (exon counts).
Ενώ πρέπει να αναφερθούν και οι μέθοδοι ποσοτικοποίησης χωρίς χάρτη (map-free), δηλαδή λιγότερο υπολογιστικά απαιτητικές αναλύσεις που γίνονται δυνατές με ποιοτικά χαρακτηρισμένα γενώματα και μεταγραφώματα αναφοράς σε σύμπραξη με τα κατάλληλα βιοπληροφορικά εργαλεία. Τέτοιες μέθοδοι είναι η επιλεκτική στοίχιση, με το Salmon (267), η ψευδό-στοίχιση, με το Kallisto (268), και η μερική-χαρτογράφιση (quasi-mapping), με το Salmon και το Sailfish (269). Οι μέθοδοι χωρίς χάρτη προσδίδουν μειωμένο χρόνο ανάλυσης και απαιτούν λιγότερη χρήση προσωρινής μνήμης. Οι εκτιμήσεις των επιπέδων έκφρασης σε επίπεδο γονιδίου μπορούν να βελτιστοποιηθούν με την πρότερη ανάλυση σε επίπεδο μεταγράφου (270).

Κανονικοποίηση. Αφού έχουμε πραγματοποιήσει την προετοιμασία και αλληλούχιση των βιβλιοθηκών, πρέπει απλώς να μετρήσουμε τον αριθμό των μικροαναγνώσεων ανά γονίδιο για να ξεχωρίσουμε τα μεγαλύτερα από τα μικρότερα γονίδια, επειδή τα μεγαλύτερα είναι πιο πιθανό να έχουν μεγαλύτερο αριθμό μικροαναγνώσεων από τα μικρότερα. Στο κεφάλαιο 2, αναφερθήκαμε στις μεθόδους κανονικοποίησης RPKM και την FPKM, που είναι για PE ανάλυση και στην οποία παίρνουμε υπόψη ότι δύο μικροαναγνώσεις μπορεί να χαρτογραφηθούν στο ίδιο θραύσμα. Η κανονικοποίηση αφορά την σύνθεση και το μέγεθος της βιβλιοθήκης, αφού διαφοροποιείται όταν προέρχεται από άλλη διαδρομή του flow cell, αλλά και διόρθωση τιμών για πολλούς άλλους λόγους όπως η μη ομοιόμορφη κατανομή μεταγράφων, το διαφορετικό μέγεθος των μικροαναγνώσεων, το βάθος κάλυψης κ.α. Εκτός από βιολογικά replicates, που είδαμε σε προηγούμενη ενότητα, υπάρχουν στα περισσότερα πειράματα NGS και τεχνικά replicates, όπως όταν χωρίζεται μια βιβλιοθήκη σε δύο flow cells.

Μια γνωστή μέθοδος CPM είναι η κανονικοποίηση διαμέσου, όπου διαιρούμε τους μετρητές με την διάμεσο όλων των μετρητών. Θα δούμε και άλλες μεθόδους κανονικοποίησης για ανάλυση δεδομένων NGS, οι οποίες χρησιμοποιούνται ανάλογα με την ανάγκη που έχουμε κάθε φορά (271). Επιπροσθέτως, όπως έχουμε ήδη αναφέρει κάθε λογισμικό ανάλυσης χρησιμοποιεί το δικό του τρόπο προεπεξεργασίας για να μειώσει τα ποσοστά σφαλμάτων και εσφαλμένων συμπερασμάτων στην μεταγενέστερη επεξεργασία των δεδομένων. Μετά την κανονικοποίηση θα έχουμε ένα κανονικοποιημένο πίνακα μετρητών c (read count table).

Κανονικοποίηση στο λογισμικό EDA-Seq. Η συγκεκριμένη αφορά την σε γονιδιακό επίπεδο κανονικοποίηση εσωτερικά της διαδρομής (within lane) του περιεχομένου ποσοστού γουανίνης και κυτοσίνης (GC-content), με αυτή ρυθμίζει και το μέγεθος της βιβλιοθήκης και το μήκος των γονιδίων (90).

Κανονικοποίηση ποσοστημορίων υπό προυποθέσεις (Conditional quantile normalization, CQN). Ο αλγόριθμος CQN αποτελεί μια εύρωστη γενικευμένη παλινδρόμηση, με αυτή ρυθμίζει το μέγεθος της βιβλιοθήκης, το μήκος των γονιδίων και το περιεχόμενο ποσοστό γουανίνης και κυτοσίνης (272).

Μετάγραφα ανά εκατομμύριο (Transcripts Per Million, TPM). Η διαφορά με τις προαναφερόμενες μεθόδους είναι ότι κανονικοποιούμε τις τιμές πριν τον υπολογισμό της. Η TPM είναι πιο εύκολα ερμηνεύσιμη από την τιμή RPKM, αφού το άθροισμα των κανονικοποιημένων τιμών ισούται με 10^6 για κάθε βιβλιοθήκη, και δίνει καλύτερα αποτελέσματα στην ανάλυση μεταξύ διαφορετικών δειγμάτων (239).

$$TPM_i = \frac{\frac{X_i}{l_i} \cdot 10^6}{\sum_j \frac{X_j}{l_k}}$$

Εξίσωση 76. Κανονικοποίηση TPM, όπου διαιρούμε τον μετρητή μικροανάγνωσης γονιδίου ανα βάση με το σύνολο των γονιδιακών μετρητών για όλες τις γονιδιακές βάσεις.

Κανονικοποίηση ξακρισμένου μέσου των Μ-τιμών (Trimmed mean of M-values, TMM). Μια βελτιωμένη μέθοδος κανονικοποίησης, η TMM παίρνει υπόψη όχι μόνο το μήκος του γονιδίου αλλά και την ολική κατανομή στο γένωμα Αυτή εφαρμόζεται για την κανονικοποίηση πρωτογενών δεδομένων από το λογισμικό edgeR της R (273), αλλά και στο λογισμικό JMP Genomics της SAS.

Κανινικοποίηση χρησιμοποιώντας την διάμεσο του ποσοστού. Ένα άλλο πολύ γνωστό λογισμικό, το DESeq2, εφαρμόζει άλλη μέθοδο για την κανονικοποίηση πρωτογενών δεδομένων (92). Πιο συγκεκριμένα, βρίσκουμε το ποσοστό των λογαριθμισμένων τιμών ενός γονιδίου το οποίο διαιρούμε με τον μέσο των λογαριθμισμένων τιμών του συνόλου των γονιδίων στα δείγματα και παίρνουμε τη διάμεσο αυτών των τιμών για όλα τα γονίδια. Παίρνουμε την κανονικοποιημένη τιμή με την διαίρεση των πρωτογενή μετρητών του γονιδίου με αυτή την τιμή διαμέσου.

Κανονικοποίηση κλίμακας ανώτερου ποσοστημορίου (Upper-quartile Scaling, UQS). Η κανονικοποίηση του ανώτερου ποσοστημορίου είναι μέθοδος CPM και σε αυτή διαιρούνται οι μετρητές με τις τιμές μετρητών ανώτερου ποσοστημορίου (274).

Κανονικοποίηση για πιθανές ταυτόσημες ομαδικές επιδράσεις. Είναι εξίσου σημαντικό να αφαιρέσουμε τον θορύβο που προέρχεται από τις ταυτόσημες επιδράσεις μιας παρτίδας (batch effects), όπως για τα δεδομένα μικροσυστοιχιών, έτσι ώστε να ρυθμίσουμε καταλλήλως τις τιμές στο σύνολο των γονιδίων. Για αυτό τον λόγο χρειάζεται να κάνουμε PCA ανάλυση.

Το γενικευμένο γραμμικό μοντέλο (Generalized Linear Model, GLM). Η μέση τιμή των μετρητών μικροαναγνώσεων για το γονίδιο G στο δείγμα k μπορούν να μοντελοποιηθούν στο GLM.

$$\log\left(\mu_{G_k}\right) = \sum_{l=1}^{n} y_{kl} a_{Gl} + \log N_k$$

Εξίσωση 77. Το GLM, όπου y_{kl} είναι η τιμή δειγματικού συντελεστή του l (l=1,2,…n), a_{Gl} είναι η τιμή γονιδιακού συντελεστή του l και N_k είναι ο συνολικός αριθμός μικροαναγνώσεων της βιβλιοθήκης k.

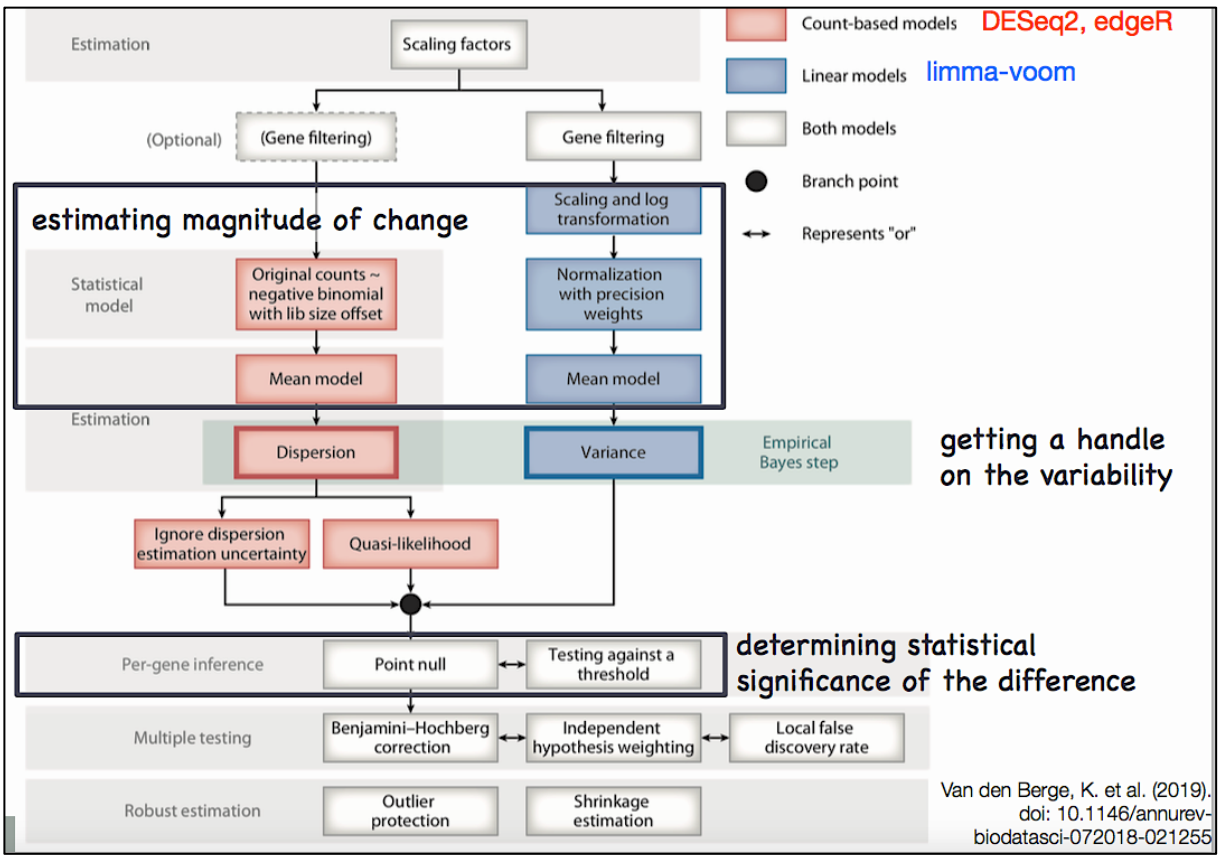
Κάθε συντελεστής y_{kl} και α_{Gl} υποδηλώνουν ένα χαρακτηριστικό του πειραματικού σχεδιασμού. Η μη μηδενική τιμή του συντελεστή y_{kl} υποδεικνύει την συμμετοχή του στην έκφραση του δείγματος k, ενώ ο συντελεστής α_{Gl} υποδεικνύει την επιρροή του στις τιμές έκφρασης του γονιδίου G στα σχετικά δείγματα (275).

4.2.2. Ανάλυση DEGs/DMGs

Μετά το στάδιο της κανονικοποίησης τα γονίδια συγκρίνονται για διαφορική έκφραση με λογισμικά, όπως το Cuffdiff (μέρος του Cufflinks) (276), το DESeq2 (92), το limma, το baySeq κ.α. Στην **Εικόνα 28**, φαίνεται η ροή εργασίας της ανάλυσης διαφορικής έκφρασης για δεδομένα από RNA-seq (277). Τα αποτελέσματα εξάγονται ως λίστες DEGs και με κάποιες διαφορές στην ροή οι λίστες DMGs. Το κάθε λογισμικό λειτουργεί με κάποιους στατιστικούς αλγορίθμους, πολλούς από τους οποίους έχουμε διερευνήσει σε προηγούμενη ενότητα, ο χρήστης έχει κάποιες επιλογές χωρίς να έχει πρόσβαση στην διαδικασία και το πρόγραμμα παρέχει τα αποτελέσματα. Το αν ένα γονίδιο θα αναγνωριστεί ως διαφορικά εκφραζόμενο εξαρτάται και από το ποιο λογισμικό χρησιμοποιούμε, γι'αυτό είναι σύνηθες να χρησιμοποιήσουμε πάνω από ένα λογισμικό και να συγκρίνουμε τα αποτελέσματα τους.

Κάποιοι πολύ γνωστοί στατιστικοί έλεγχοι είναι ο έλεγχος Chi-squared (χ^2), ο υπεργεωμετρικής κατανομής έλεγχος ακρίβειας του Fisher (274) και μέσω παλινδρόμησης Poisson. Ο έλεγχος chi-squared μπορεί να πραγματοποιήσει την ανάλυση του πίνακα μετρητών με την εντολή *chisq.test* στο R/Bioconductor. Επιπροσθέτως, η διαφορική ανάλυση μπορεί να πραγματοποιηθεί και με τον έλεγχο Poisson GLM, με το εργαλείο DEGseq (278).

Το πακέτο limma, στο οποίο αναφερθήκαμε και στην ανάλυση μικροσυστοιχιών, είναι κατάλληλο και για την ανάλυση διαφορικής έκφρασης σε RNA-Seq. Το εργαλείο limma + voom μπορεί να αναλύσει με μεγάλη ακρίβεια τους μετρητές μικροαναγνώσεων και με αυτό ελέγχουμε τα ποσοστά σφαλμάτων για τις πολλαπλές στοιχίσεις διαφόρων γονιδίων ταυτόχρονα (279).



Εικόνα 28. Διαφορές στο workflow μεταξύ μεθόδους μετρητών (GLM), όπως το edgeR και το DESeq2, και γραμμικές μεθόδους, όπως το limma+voom (277, 280).

Το μοντέλο για την διαφορική ανάλυση του εργαλείου DESeq, όπως δημοσιεύτηκε από τους (281). Τα δεδομένα ακατέργαστων μετρητών (Raw count) για γονίδιο i στο δείγμα j ακολουθούν την αρνητική διωνυμική κατανομή (negative binomial), δηλαδή $K_{ij} \sim NB(\mu_{ij}, \sigma_{ij}^2)$.

$$s_j = \underset{i}{Md} \frac{K_{ij}}{\left(\prod_{v=1}^{m} K_{iv} \right)^{\frac{1}{m}}}$$

Εξίσωση 78. Ο παράγοντας κλίμακας κανονικοποίησης για δεδομένα RNA-Seq με το DESeq, όπου *i* είναι το γονίδιο ή το ισόμορφο, *j* είναι το δείγμα, *m* είναι ο αριθμός των δειγμάτων, K_{ij} είναι ο αριθμός των μετρητών για το γονίδιο *i* στο πείραμα *j* και s_j είναι το βάθος κάλυψης για το πείραμα *j* (281).

$$\mu_{ij} = q_{ip(j)} s_j$$

$$\sigma_{ij}^2 = \mu_{ij} + s_j^2 v_p(q_{ip(j)})$$

$$q_{ip} = \frac{1}{replicates} \sum_j \frac{K_{ij}}{s_j}$$

Εξίσωση 79. Υπολογισμός της μέσης κλιμακωτής έκφρασης για το γονίδιο *i* στην κατάσταση *p*, όπου *p(j)* είναι η κατάσταση του δείγματος *j* (281).

Ακόμα, έστω ότι έχουμε την υπόθεση H_0 ότι δεν έχουμε διαφορική έκφραση ανάμεσα σε μια κατάσταση *A* και *B*. Τότε, έχουμε και την αντίθετη υπόθεση H_1 ότι υπάρχει διαφορική έκφραση. Ενώ ο έλεγχος αναλογίας πιθανότητας (Likelihood ratio) ακολουθεί την κατανομή χ^2 και μπορούμε να τον υπολογίσουμε (282).

$$T_i = 2\log \frac{P\left\langle K_{iA} \middle| H_1 \right\rangle \cdot P\left\langle K_{iB} \middle| H_1 \right\rangle}{P\left\langle K_{iA}, K_{iB} \middle| H_0 \right\rangle}$$

Εξίσωση 80. Η στατιστική τιμή *T* για τον έλεγχο αναλογίας πιθανότητας ακολουθεί ασυμπτωτικά την κατανομή *chi* squared (282).

$$P(H_0) \approx 1 - ChiSquaredCDF\left\langle T_i \middle| \nu \right\rangle$$

Εξίσωση 81. Η πιθανότητα H_0 (282).

Το μοντέλο που χρησιμοποιεί η DESeq2 για δεδομένα RNA-seq. Τα δεδομένα ακατέργαστων μετρητών (Raw count) για γονίδιο *i* στο δείγμα *j* ακολουθούν την αρνητική διωνυμική κατανομή (negative binomial), δηλαδή $K_{ij} \sim NB(\mu_{ij}, \alpha_i)$, όπου α είναι η ειδική γονιδιακή παράμετρος διασποράς (που προσαρμόζεται προς την μέση διασπορά). Όπου το πειραματικό σχέδιο συμπεριλαμβάνει πληροφορίες όπως αν είναι δείγμα αναφοράς ή ελέγχου, από ποια παρτίδα προέρχεται κ.α. Το εργαλείο DESeq2 αποκλείει αυτόματα ασθενώς εκφρασμένα γονίδια σε διαδικασία διόρθωσης πολλαπλών ελέγχων. Τα ασθενώς εκφρασμένα γονίδια προσδίδουν πολύ θόρυβο στο LFC και γι'αυτό συνήθως δεν αναγνωρίζονται ως διαφορικά

εκφραζόμενα, ενώ ο ελαττωμένος αριθμός ελέγχων αυξάνει την στατιστική δύναμη, δηλαδή έχουμε περισσότερες αναγνωρίσεις. Το σύνηθες όριο κατωφλιού FDR είναι η τιμή 0.1 (283).

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

Εξίσωση 82. Προσαρμοσμένος μέσος, όπου *q* είναι η εκτιμώμενη τιμή έκφρασης και *s* είναι ο παράγοντας του μεγέθους της βιβλιοθήκης (283).

$$\log_2(q_{ij}) = \sum_r x_{jr} \beta_{ir}$$

Εξίσωση 83. Η προσαρμοσμένη log*FC* για το γονίδιο i, όπου x_j είναι η στήλη του πίνακα πειραματικού μοντέλου για το δείγμα *j* και β_{ir} είναι ο συντελεστής του γενικευμένου γραμμικού μοντέλου (Generalized Linear Model, GLM), ένας για κάθε στήλη (283).

Η παλινδρόμηση Poisson είναι ένας άλλος στατιστικός έλεγχος που μπορούμε να πραγματοποιήσουμε. Θεωρούμε ότι η κατανομή του παρατηρούμενου αριθμού συνολικών μικροαναγνώσεων για το γονίδιο στο δείγμα i είναι R_i~Poisson($N_i p_i$), όπου N_i είναι ο συνολικός αριθμός των θραυσμάτων που μετρώνται στο δείγμα i και p_i είναι η πιθανότητα το θραύσμα να ταυτίζεται με το γονίδιο στο δείγμα i. Το στατιστικό μοντέλο ακολουθεί.

$$E[R_i] = Var[R_i] = N_i p_i$$

Εξίσωση 84. Παρατηρούμενος συνολικός αριθμός μικροαναγνώσεων που ταυτίζονται με το γονίδιο στο δείγμα *i*.

$$\log(p_i) = \beta_0 + \beta_1 T_i$$

Εξίσωση 85. Ο έλεγχος για την διαφορική έκφραση γίνεται με β_1=0 για την μηδενική υπόθεση. Το T_i είναι δείκτης 0 ή 1.

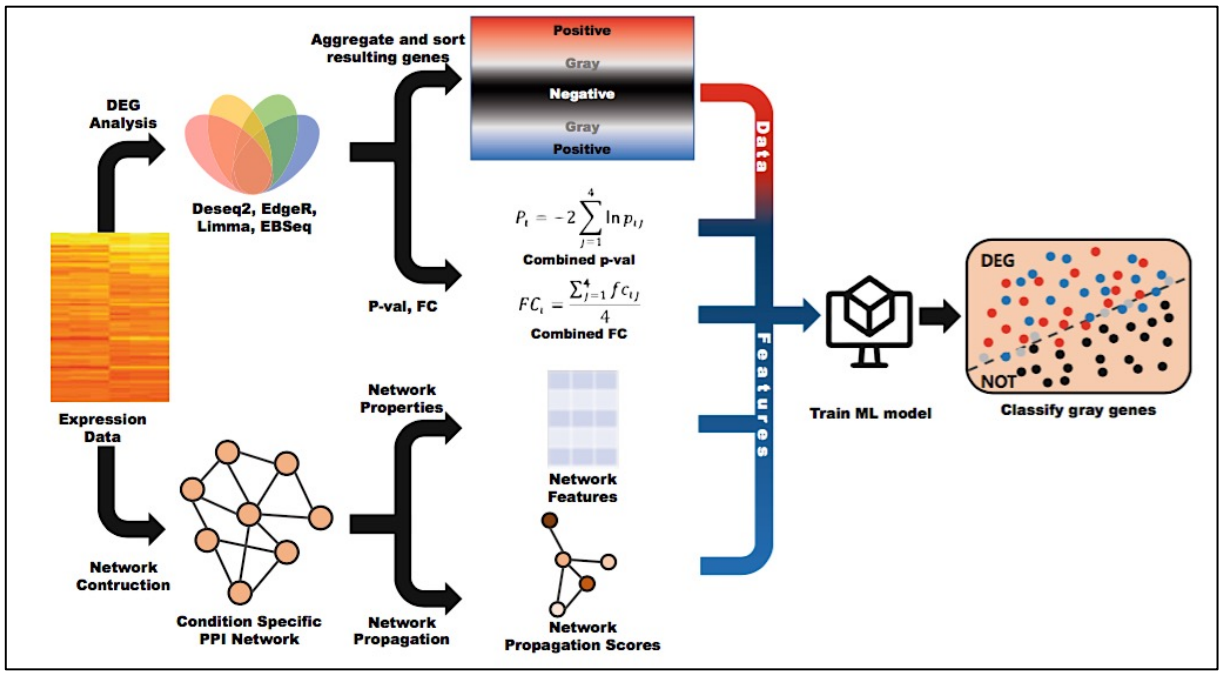
$$E[R_i] = N_i p_i = N_i e^{(\beta_0 + \beta_1 T_i)}$$
$$\log\left(E[R_i]\right) = \log N_i + \beta_0 + \beta_1 T_i$$

Εξίσωση 86. Η παλινδρόμηση Poisson, όπου για το $\beta_1 T_i$ χρησιμοποιούμε την επαναληπτική διαδικασία MLE και η logN_i δεν θεωρείται πολύ σημαντικη (284).

$$\chi^2_D = 2 \cdot \sum_i \left(R_i \cdot \log \frac{R_i}{\widehat{R_i}} \right)$$

Εξίσωση 87. Η παρέκκλιση με τον έλεγχο καλής προσαρμογής (goodness of fit) για την υπερδιασπορά (overdispersion) (284).

Στην **Εικόνα 29**, παρουσιάζεται μια προσέγγιση ML για την ταξινόμηση των DEGs με την εκπαίδευση μοντέλου λογιστικής παλινδρόμησης και τη βοήθεια δικτύων αλληλεπιδράσεων μεταξύ πρωτεϊνών (Protein-Protein Interactions, PPI) (285).

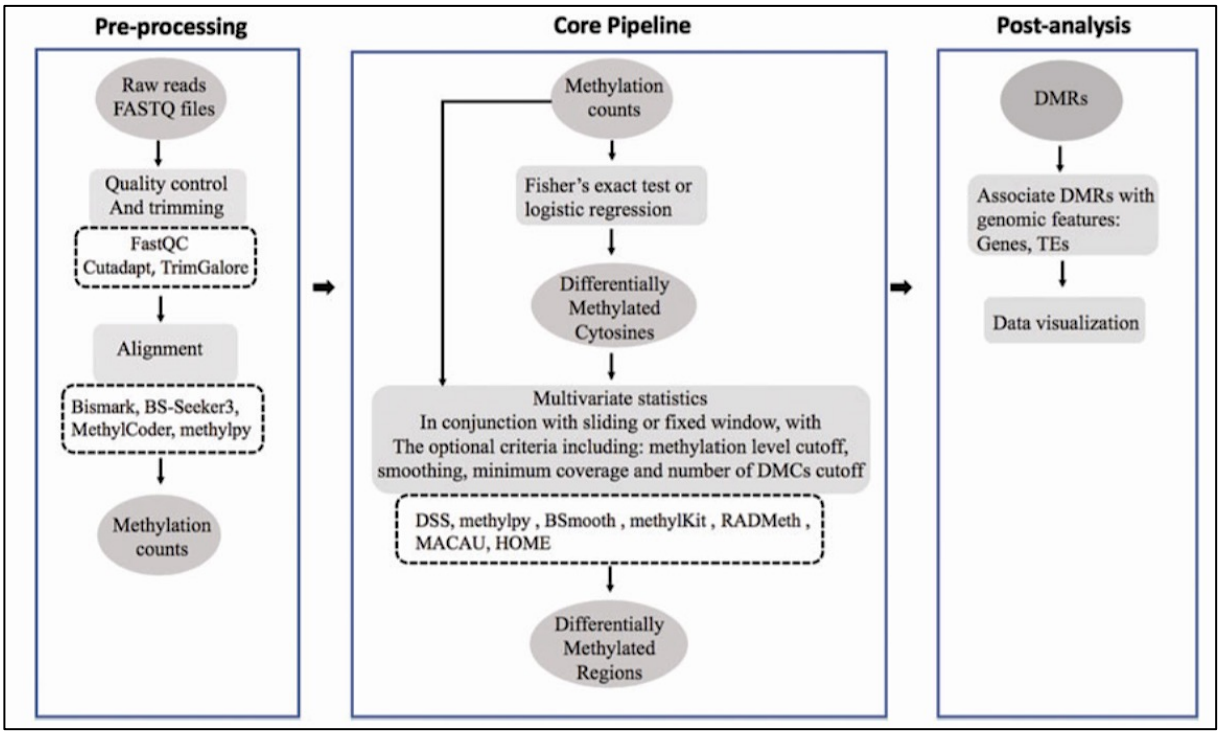


Εικόνα 29. Η ροή εργασιών MLDEG, όπου θετικά δεδομένα είναι τα DEGs και αρνητικά δεδομένα είναι τα μη διαφορικά εκφρασμένα (286).

Myrna. Ένα πολύ χρηστικό εργαλείο συννέφου (cloud) για την ανάλυση διαφορικής έκφρασης σε δεδομένα από RNA-seq είναι το Myrna. Η επεξεργασία μπορεί να γίνει στο σύννεφο μέσω του MapReduce (είναι ένα εργαλείο για την επεξεργασία Big Data μέσω παράλληλης εκτέλεσης σε μεγάλο αριθμό κόμβων) σε ένα cluster του Hadoop (είναι μια υποδομή που κάνει δυνατή την αποθήκευση και την επεξεργασία δεδομένων, ενώ αποτελείται από ονόματα κόμβων και κόμβους δεδομένων) ή σε έναν μεμονωμένο υπολογιστή με την αξιοποίηση πολλών CPUs (287).

Methyl-Seq. Ο σκοπός της ανάλυσης μεθυλίωσης είναι η εύρεση περιοχών μεθυλίωσης και η εξαγωγή μιας λίστας διαφορικά εκφραζόμενων (μεθυλιωμένων) γονιδίων DMGs, ενώ είναι συνήθης η σύγκριση τους με DEGs. Η βασικός τρόπος ανάλυσης μεθυλίωσης είναι η WGBS, σε σύγκριση με την MeDIP-Seq και τις μεθόδους ενζυμικού περιορισμού (π.χ. MRE-Seq). Ακόμα, σε κάποιες δημοσιεύσεις γίνεται και το επιπλέον βήμα εύρεσης των διαφορικά

μεθυλιωμένων γενωμικών περιοχών (DMGR), η οποία γίνεται σε εξαρτώμενα άτομα του ίδιου πληθυσμού και μπορεί να πραγματοποιηθεί με το υπολογιστικό πακέτο Methyl-IT[47], το οποίο είναι γραμμένο στην R. Στην δημοσίευση των (288), μπορούμε να δούμε την ροή εργασιών της λήψης σήματος και μιας προσέγγισης ML για την εύρεση των DMGRs.



Εικόνα 30. Η ροή εργασιών για την ανάλυση WGBS. Οι διακεκομμένες γραμμές αναπαριστούν πιθανά εργαλεία ανάλυσης, τα οβάλ είναι δεδομένα εισόδου/εξόδου και τα τετράγωνα είναι οι πρότυπες διεργασίες (289).

Η ανάλυση μεθυλιώματος από BS-Seq δεδομένα περιέχει περισσότερα βήματα από την DNA-Seq για την διαφορική ανάλυση, όπως φαίνεται στην δημοσίευση του (290). Ακόμα, προτιμώνται εργαλεία που περιέχουν και ειδικές συναρτήσεις για την ανάλυση της μεθυλίωσης. Τα βήματα της προεπεξεργασίας είναι όπως τα προαναφερθέντα, δηλαδή πραγματοποιείται ο ποιοτικός έλεγχος των μικροαναγνώσεων, το φιλτράρισμά και το ξάκρισμά τους, αναφερθήκαμε σε τέτοια εργαλεία στην ενότητα 4.2.1. Αλλά το Trim Galore, εκτός από εργαλείο διασφάλισης ποιότητας και αφαίρεσης προσαρμογέων, περιέχει ειδικές συναρτήσεις για την αφαίρεση της προκατάληψης των μεθυλιωμένων περιοχών σε μικροαναγνώσεις που προέρχονται από ανάλυση RRBS και μπορεί να εφαρμόσει αυτόματα FastQC στα ξακρισμένα δεδομένα, βελτιώνοντας την αποδοτικότητά του (290).

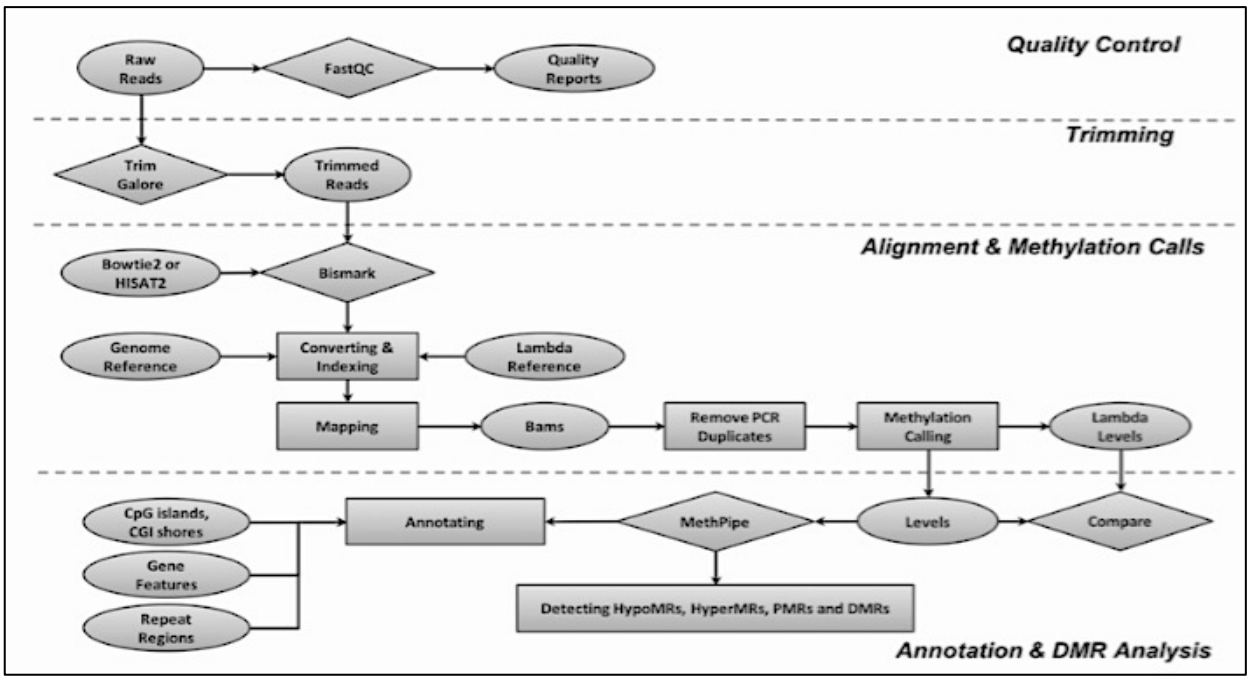
[47] https://github.com/genomaths/MethylIT

Στην συνέχεια στοιχίζονται οι μικροαναγνώσεις σε ένα γένωμα αναφοράς. Το πακέτο Bismark είναι ενδεδειγμένο για αυτές τις αναλύσεις (244), αφού είναι ικανό για την ανίχνευση των γενωμικών περιοχών μεθυλίωσης και για την παραγωγή γενωμικής κάλυψης των μεθυλιωμένων βάσεων. Για αυτό τον σκοπό, το Bismark κατασκευάζει το ευρετήριο και τις μετατροπές των γενωμικών αλληλουχιών, πριν από την στοίχιση στο γένωμα αναφοράς με ένα εργαλείο όπως το Bowtie2, που περιέχεται στο Bismark. Η αναφορά του, το bedGraph, παρουσιάζει τους μεθυλιωμένους και μη μεθυλιωμένους μετρητές και που βρίσκονται, για κάθε περιοχή CpG στο γένωμα (κλήση μεθυλίωσης) (290).

Η αυτοματοποιημένη ροή εργασιών MethPipe (291) είναι σχεδιασμένη για την ανάλυση, τον σχολιασμό και τον εντοπισμό της διαφορικής έκφρασης δισουλφιδικών δεδομένων αλληλούχισης, όπως φαίνεται στην **Εικόνα 31**. Για πολλαπλά replicates, χρησιμοποιείται η βήτα διωνυμική παλινδρόμηση και στην συνέχεια προσαρμόζουμε την p-value για να βρούμε σωστά τις διαφορικά εκφραζόμενες περιοχές (Differentially Methylated Regions, DMRs). Μια πιο απλή υπολογιστική αυτοματοποιημένη ροή εργασιών είναι η DMRfinder (292). Ακόμα, το πακέτο R περιέχει το εργαλείο methylAction (293), που είναι κατάλληλο και εύχρηστο για την ανίχνευση DMRs. Ένα άλλο εργαλείο για την ανάλυση δεδομένων από BS-Seq είναι το Methylkit[48] είναι γραμμένο στην R και χρησιμοποιείται στο Hadoop για την ανάλυση μεθυλίωσης DNA (294).

Οπτικοποίηση. Η οπτικοποίηση των αποτελεσμάτων είναι από τους ευκολότερους τρόπους, π.χ. οι θερμικοί χάρτες και οι κρατήρες ηφαιστείου, να κατανοήσουμε γρήγορα και να ερμηνεύσουμε σωστά τα πειραματικά δεδομένα. Το αρχείο SAM/ BAM /BED μπορεί να χρησιμοποιηθεί από ένα εργαλείο όπως το IGV, που μπορεί να χρησιμοποιηθεί και σε δεδομένα από μικροσυστοιχίες, και εξάγει τα αποτελέσματα σε αρχεία PDF κ.α. εύκολα χρησιμοποιούμενα αρχεία σε δημοσιεύσεις. Σε μια ανασκόπηση με εννιά από τα πιο δημοφιλή εργαλεία (295), για ανάλυση διαφορικής έκφρασης με δεδομένα NGS, το εργαλείο Cummerbund εμφανίζει τις πέντε από τις έξι βασικές λειτουργίες που επιλέχθηκαν (219). Αλλά υπάρχουν λογισμικά που περιέχουν πάνω από ένα από αυτά τα εργαλεία, οπότε, ουσιαστικά, μπορούμε να αξιοποιήσουμε όλες τις λειτουργίες.

[48] https://github.com/al2na/methylkit/



Εικόνα 31. Η ροή εργασιών BS-Seq και ανάλυση διαφορικής μεθυλίωσης με το εργαλείο MethPipe (290).

4.2.3. Ομαδοποίηση και Ταξινόμηση

Τα υπολογιστικά εργαλεία και η μαθηματική μοντελοποίηση, τα οποία έχουν περάσει πολλά στάδια ανάπτυξης, κάνουν πιο κατανοητές τις βιολογικές λειτουργίες και την λειτουργία πολύπλοκων συστημάτων. Ήδη από το προηγούμενο κεφάλαιο έχουμε αναφέρει αρκετές από τις μεθόδους ομαδοποίησης και ταξινόμησης για τα δεδομένα μικροσυστοιχιών, οι οποίες είναι εφαρμόσιμες και στα δεδομένα αλληλούχισης. Οπότε σε αυτό το κεφάλαιο θα αναφερθούμε και σε επιπλέον τεχνικές. Η μέθοδοι μείωσης διαστασιμότητας, π.χ. PCA, παίζουν βασικό ρόλο, όπως και στις μικροσυστοιχίες. Το ίδιο ισχύει και για τις άλλες μεθόδους στις οποίες αναφερθήκαμε στην ενότητα 3.3. Πιο συγκεκριμένα, μας ενδιαφέρουν και στην NGS οι προαναφερθείσες τεχνικές για την ομαδοποίηση, όπως η ιεραρχική ομαδοποίηση, οι SOMs και οι νευρωνικοί μέθοδοι ομαδοποίησης, αλλά και οι τεχνικές για την ταξινόμηση, όπως τα RF, οι kNN και οι SVM.

Η δημοσίευση των (296) περιέχει ανασκόπηση ωμικών και πολυ-ωμικών (δηλαδή που αποτελούνται από πάνω από μια ωμική επιστήμη π.χ. γενωμική και μεταγραφομική) μελετών, συμπεριλαμβανομένου δημοσιεύσεων που επικεντρώνονται στην ανάλυση γονιδιακής έκφρασης. Για παράδειγμα, στην μελέτη των (297), χρησιμοποιήθηκαν οι μέθοδοι ταξινόμησης SVM-RFE (recursive feature elimination), η οποία είναι η συνηθέστερα

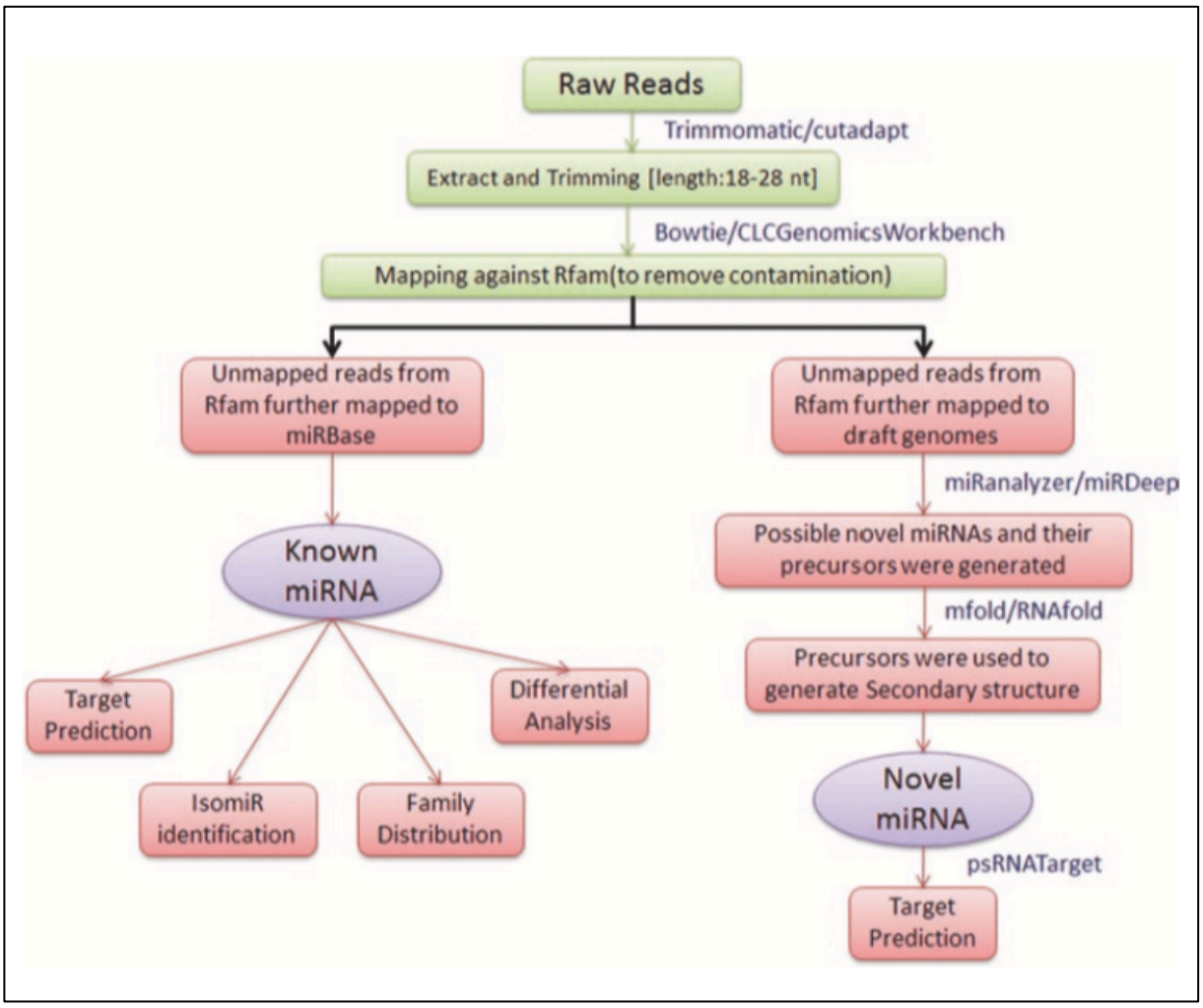
χρησιμοποιούμενη μέθοδος SVM, και οι kNN. Οι ερευνητές ανέπτυξαν 50 γονιδιακές υπογραφές (gene signatures) σε δεδομένα γονιδιακής έκφρασης με την LOOCV (leave-one-out cross validation) τεχνική αξιολόγησης του μοντέλου. Στην δημοσίευση των (298), όπου συγκρίθηκαν δεδομένα γονιδιακής έκφρασης με δεδομένα μεθυλίωσης DNA από ένα σύνολο δεδομένων, χρησιμοποιήθηκε ο κατηγοριοποιητής Αφελούς Μπέϋς, η ταξινόμηση SVM, RF και η λογιστική ανάλυση παλινδρόμησης. Η ωμική ανάλυση με τις μεθόδους μη επιβλεπόμενης αναγνώρισης προτύπων είχε καλύτερα αποτελέσματα σε σχέση με ένα μεμονωμένο σύνολο δεδομένων. Σε μια πρόσφατη δημοσίευση για ανάλυση ενός συνόλου γονιδιακής έκφρασης των (299), χρησιμοποιήθηκαν μέθοδοι ταξινόμησης όπως τα RF, τα NN, τα kNN και το SVM. Όμως καλύτερα αποτελέσματα στην διάγνωση του καρκίνου παρουσίασε η ταξινόμηση με RF. Επιπροσθέτως, στην δημοσίευση των (296) περιέχονται σε ένα πίνακα διάφορα εργαλεία που χρησιμοποιούνται αυτή την στιγμή στην ωμική ανάλυση.

4.2.4. Ειδικότερα για Αναλύσεις miRNA

miRNA-Seq. Η αναγνώριση miRNA από γενωμικές αλληλουχίες με υπολογιστικές μεθόδους είναι μια πολύ βοηθητική μέθοδος των πειραματικών μεθόδων, λαμβάνοντας υπόψη ότι οι πειραματικές μέθοδοι για την αλληλούχιση των miRNA είναι αρκετά προβληματικές λόγω της δυσκολίας εξαγωγής του κατάλληλου δείγματος (το μικρό μεγέθους και η χαμηλή έκφραση στα μόρια αυτά, καθώς και η ειδικότητας τους), του κόστους, και της αργής και χρονοβόρας διαδικασίας (**Εικόνα 32**). Οι υπολογιστικές (in silico) προγνώσεις των μορίων miRNA διαχωρίζονται σε γονιδιακή πρόγνωση miRNA (miRNA gene prediction), όπου μας αφορούν τα χαρακτηριστικά μορίων miRNA, και πρόγνωση στόχων miRNA (miRNA target prediction), όπου μας αφορούν τα χαρακτηριστικά φυσιολογίας. Υπάρχουν διάφοροι τρόποι να διαχωρίσουμε τα μόρια miRNA, οι βασικοί είναι: τα δομικά γνωρίσματά τους, όπως το μήκος του μίσχου-θηλιάς και η θερμοδυναμική σταθερότητα, τα γνωρίσματα στην αλληλουχία τους, όπως το περιεχόμενο σε νουκλεοτίδια και την θέση αυτών, και η εξελικτική διατήρηση τους.

Η γονιδιακή πρόγνωση βασίζεται στα ενημερωμένα και καλοδιατηρημένα σύνολα εκμάθησης που προέρχονται από διαδικασίες wetlab και έχουν ξεχωριστά γνωρίσματα από άλλα είδη μικρών RNAs. Οι μέθοδοι αυτοί είναι σημαντικό να έχουν την ίδια ή πολύ παρόμοια ροή εργασιών για τα σύνολα εκμάθησης, με αποτέλεσμα να υπάρχει συναίνεση στην ορολογία του miRNA και μια συνοχή στον χαρακτηρισμό των μορίων miRNA, αφού πολλές φορές βρίσκεις σχόλια αντίθετης απόψεως για την ίδια γενωμική περιοχή σε διαφορετικές βάσεις

δεδομένων (260). Κάποιες γνωστές μέθοδοι είναι η ανίχνευση γειτονικών/παρόμοιων δομών μίσχου-θηλιάς, η ανίχνευση γονιδίων μέσω συντηρημένων περιοχών, με χρήση εργαλείων όπως το MFold, και η εύρεση ομολογίας ήδη γνωστών miRNAs, με εργαλεία όπως το microHARVESTER για φυτικούς οργανισμούς. Η δημοσίευση των (300) πρόκειται για μια εφαρμογή μεθόδου πρόβλεψης ανθρώπινου miRNA χρησιμοποιώντας SVM. Με τη βοήθεια του λογισμικού DIANA-microH γίνεται η ταξινόμηση του miRNA ως εξής: Κάνουμε εκμάθηση σε ένα υποσύνολο των miRNAs, που βρίσκονται στη RFAM και μετά ελέγχουμε τα υπόλοιπα.



Εικόνα 32. Η miRNA-Seq ροή εργασιών για φυτά (260).

Χαρτογράφηση miRNA-seq με το εργαλείο Subread. Αν θέλουμε να χρησιμοποιήσουμε το Subread για την χαρτογράφηση των μικροαναγνώσεων από miRNA-seq, πρέπει πρώτα να κατασκευαστεί ένα πλήρες ευρετήριο, το οποίο αποτελείται από τα εξαγόμενα 16 bp-μερή από την κάθε περιοχή του γενώματος. Στο εργαλείο Subread δεν ασχολούμαστε με το βήμα

του ξακρίσματος, αφού έχει την δυνατότητα να αφαιρεί τα μέρη τις αλληλουχίας που δεν χαρτογραφούνται σωστά. Τα 16 bp-μερή, που εξάγονται από κάθε μικροανάγνωση, πρέπει να υπολογίζονται στο μήκος των μικροαναγνώσεων πλήν 15, έτσι ώστε να έχουμε υψηλή ευαισθησία στα πολύ μικρά αυτά μόρια. Ο αριθμός των συναινετικών υπο-αναγνώσεων (subreads) είναι από δύο μέχρι εφτά. Στις λιγότερες συναινετικές υπο-αναγνώσεις (2) έχουμε μεγαλύτερο σφάλμα χαρτογράφησης, αλλά μπορούμε να ανιχνεύσουμε ακόμα και αλληλουχίες miRNA πολύ μικρού μήκους (από 17 bp). Από την άλλη πλευρά, στις περισσότερες συναινετικές υπο-αναγνώσεις (7) μπορούμε να ανιχνεύσουμε μόνο αλληλουχίες miRNA μεγαλύτερου μήκους (από 22 bp). Το ποια τιμή θα διαλέξουμε σε αυτή την κλίμακα (2-7) εξαρτάται από το δείγμα που έχουμε. Μπορούμε στην συνέχεια να χρησιμοποιήσουμε ένα εργαλείο όπως το featureCounts για το στάδιο της ποσοτικοποίησης των μικροαναγνώσεων στα γονίδια miRNA και μια βάση δεδομένων miRNA, που περιέχει χαρακτηρισμό των γονιδίων αυτών (301).

Ανακάλυψη νέων μορίων miRNA. Οι περιοχές γενωμικού miRNA είναι διακεκριμένες σε σχέση με άλλους τύπους γονιδίων και κάποιες φορές βρίσκονται στα εσόνια. Τα δεδομένα αναλύονται στατιστικά και με τις προσεγγίσεις ML ως προς τα χαρακτηριστικά τους για την εκ νέου αλληλούχιση, έτσι ώστε να γίνει ο χαρακτηρισμός τους για την αναγνώριση νέων μορίων miRNA. Τα χαρακτηριστικά αυτά είναι κυρίως η δευτεροταγής δομή του πρόδρομου μορίου miRNA, η υψηλώς συντηρημένη αλληλουχία του ώριμου και του προδρόμου μορίου miRNA και τα χαρακτηρισμένα δεδομένα έκφρασης στις βιβλιοθήκες αλληλουχιών μικρού RNA. Η επιτυχία στην στοίχιση αλληλουχίας είναι περιορισμένη σε κάποια προδρομικά μόρια miRNA γιατί οι περιοχές του πρόδρομου μορίου συχνά δεν είναι συντηρητικές περιοχές. Ο πιο εύκολος τρόπος ανακάλυψης νέων μορίων miRNA είναι μέσω τοπικής ή ημι-ολικής ανάλυσης ομολογίας με παρόμοια είδη, σε λογισμικά όπως το gotohscan και το ssearch (302). Τα αποτελέσματα πρέπει να επιβεβαιωθούν για την στοίχιση, όπου στο ώριμο miRNA επιτρέπονται μόνο ελάχιστες αντικαταστάσεις, και για την πλήρη ομοιότητα στην δευτεροταγή δομή του μορίου. Το πρόδρομο miRNA οδηγεί σε ένα ή κάποιες φορές δύο μόρια ώριμου miRNA. Ένα καλό εργαλείο για την στοίχιση, αλληλουχίας και δομής, των miRNAs είναι το miRAlign. Ένας άλλος τρόπος είναι μια αυτοματοποιημένη ροή εργασιών που φιλτράρει και θα αποκόπτει κάποια miRNA σύμφωνα με τα όρια σε κάθε μέρος της ροής εργασιών, αυτός ο τρόπος είναι καλός για συγγενικά είδη. Η ανίχνευση με βάση το προφίλ ανίχνευσης των mRNAs είναι μια ακόμα επιλογή. Μετά τον χαρακτηρισμό ολόκληρο το γένωμα ελέγχεται με τις προσεγγίσεις ML, οι οποίες έχουν εκπαιδευμένους αλγορίθμους με

γνωστά miRNA, και οι πιθανές αλληλουχίες ταξινομούνται. Κάθε εργαλείο μπορεί να διαφέρει στους αλγορίθμους δομικής πρόγνωσης για την δομή μίσχου-θηλιάς, το πώς ορίζουμε τα χαρακτηριστικά για την ταξινόμηση και τις μεθόδους ML που χρησιμοποιούμε. Ένα τέτοιο παράδειγμα είναι το εργαλείο NOVOMIR (303), που χρησιμοποεί τεχνικές των κρυφών μοντέλων Μάρκοβ (Hidden Markov Models, HMM), πρώτα με το εργαλείο RNAlfold (304) για να ελέγξει την τοπική δευτεροταγή δομή, και στην συνέχεια εφαρμόζει το εργαλείο RNAshapes (305) για να ανακαλυφθούν οι περιοχές μίσχου θηλιάς. Τα πιθανά miRNAs ελέγχονται με τους υπάρχοντες χαρακτηρισμούς για να μειωθούν τα εσφαλμένως θετικά αποτελέσματα (306). Τέλος, ένα ολοκληρωμένο εργαλείο ανάλυσης miRNA σε διαδικτυακό διακομιστή (Web Server) είναι το MiRQuest (307).

Σε βραχύ-μικροανάγνωση αναλύσεις, απομονώνεται το miRNA, γίνεται η πρόσδεση των προσαρμογέων, αυτό επαυξάνεται και αλληλουχείται σε πλατφόρμες NGS, όπως η πλατφόρμα Illumina και η SOLiD. Η χαρτογράφηση γίνεται με τα εργαλεία που έχουμε ήδη αναφέρει, η διαφορά είναι ότι πρέπει να επιτρέψουμε τις μη ταυτίσεις και τις πολλαπλές χαρτογραφήσεις λόγω των αλλαγών στην αλληλουχία RNA (RNA editing), δηλαδή της δημιουργίας μίας παραλλαγής του miRNA (isomir) ή ακόμα και να επιδράσει στην βιογένεση του μορίου. Με την αλληλούχιση ξέρουμε τις μικροαναγνώσεις και την περιοχή στην οποία βρίσκονται τα miRNA, οπότε η αναγνώριση νέων miRNA γίνεται με αυτές τις πληροφορίες σε συνδυασμό με την δευτεροταγή δομή τους. Ένα χρήσιμο διαδικτυακό εργαλείο για την ανάλυση δεδομένων μη κωδικών μορίων με NGS είναι το DARIO (308). Χρησιμοποιούνται προφίλ χαρτογραφημένων δεδομένων υψηλής απόδοσης, οπότε δεν μας ενδιαφέρει από ποια πλατφόρμα προέρχονται, για να ταξινομηθούν τα miRNAs, στην περίπτωσή μας. Οι στοίβες των συνεχόμενων μικροαναγνώσεων οργανώνονται σε blocks, μετά από την ομαδοποίηση τους, αυτά τα τετράγωνα παίρνουν την μορφή συγκεκριμένων προτύπων για διαφορετικούς τύπους μη κωδικών μορίων RNA και με αυτά εκπαιδεύουμε έναν RF ταξινομητή (309). Έτσι μπορούμε να ξεχωρίσουμε ανάμεσα σε miRNAs, tRNAs και snoRNAs (306).

Πίνακας 7 Εργαλεία για την αναγνώριση miRNA από δεδομένα αλληλούχισης NGS (45).

Εργαλείο	Οργανισμός	Τύπος Αλγορίθμου	Διάστημα Δημοσιεύσεων	Έκδοση	Λογισμικό ή Διαδικτύου	Σύνδεσμος και Βιβλιογραφία
miRDeep2	Ζ	Μηχανική Μάθηση, NGS, Δομή	2008-2012	2016, v2.0.0.8	Λ	https://www.mdc-berlin.de/n-rajewsky#t-data,software&resources (310)
miRCat2	Ζ	NGS	2008-2017	2018, v4.5	Λ	http://srna-workbench.cmp.uea.ac.uk/mircat2/ (311)
miRanalyzer	Ζ+Φ	NGS, Ενοποίηση δεδομένων (ωμική)	2009-2010	2012, v0.3	Λ+Δ	https://bioinfo2.ugr.es/ceUGR/miranalyzer/ (312)
miRDeep-P2	Φ	Μηχανική Μάθηση, NGS, Δομή	2011	2019	Λ	https://sourceforge.net/projects/mirdp2 (313)
miRDeep*	Ζ+Φ	Ενοποίηση δεδομένων (ωμική), NGS, Δομή	2013	2016, v37	Λ	http://www.australianprostatecentre.org/research/software/mirdeep-star (314)
miReader	Ζ+Φ	Μηχανική Μάθηση, NGS	2013	2016	Λ	https://scbb.ihbt.res.in/2810-12/miReader.php (315)
miRPlex	Ζ+Φ	Μηχανική Μάθηση, NGS	2013	2013, v0.1	Λ	https://www.uea.ac.uk/computational-biology/software/mirplex (316)
miRdentify	Ζ	Θερμοδυναμική Ισορροπία, NGS	2014	2014, v1.0	Λ	https://www.ncrnalab.dk/#mirdentify/mirdentify.php (317)
miRPlant	Φ	NGS, Ενοποίηση δεδομένων (ωμική)	2014	2017, v5.1	Λ	https://sourceforge.net/projects/mirplant/ (318)
Mirnovo	Ζ+Φ	Μηχανική Μάθηση, NGS	2017	2018, v1.0	Λ+Δ	http://wwwdev.ebi.ac.uk/enright-dev/mirnovo/ (319)

Στόχοι miRNA. Η πρόγνωση στόχων miRNA, με μήκος pre-miRNA 60-80bp, έχει βασικό ρόλο την ρύθμιση της μετάφρασης και την διάγνωση του καρκίνου. Οι στόχοι miRNA συμπληρώνουν τα ώριμα miRNA, στο 5'-άκρο υπάρχει η περιοχή συντήρησης "σπόρων" μήκους 6-8 nt για την μείωση των εσφαλμένως θετικών (FP) προβλέψεων, ενώ το 3'-άκρο είναι το αντιστάθμισμα (compensatory) για την καλύτερη ευστάθεια και αποδοτικότητα του miRNA. Η εξεταζόμενη περιοχή μπορεί να είναι εσονική ή διαγονιακή και η ρύθμιση πραγματοποιείται συνήθως από το ένζυμο πολυμεράση ΙΙ. Για να βρούμε τους στόχους miRNA, που βρίσκονται στις περιοχές 3'-UTRs, ερευνάμε τα φυσικά χαρακτηριστικά και την δομή τους. Ενώ για τα φυτά έχουμε υψηλό ποσοστό συμπληρωματικότητας των αλληλουχιών, αυτό δεν ισχύει για τα ζώα. Εφόσον επιτρέπουμε τις μη ταυτίσεις θα έχουμε

και πολλά εσφαλμένως θετικά (FP) αποτελέσματα. Γι'αυτό τον λόγο πρέπει να βελτιώσουμε την ακρίβεια των προγνώσεων, π.χ. με καλύτερη συμπληρωματικότητα του 5'-άκρου.

Κάποια υπολογιστικά εργαλεία στόχων miRNA περιέχει ο Πίνακας 8, με αυτά ο ερευνητής κερδίζει σε χρόνο και κόστος. Αυτά μπορεί να αναλύουν την συμπληρωματικότητα (για παράδειγμα το 5'-άκρο του miRNA έχει περισσότερες συμπληρωματικές βάσεις στον στόχο του από το 3'-άκρο), τους υπολογισμούς ελεύθερης ενέργειας, εξελικτικά χαρακτηριστικά (για παράδειγμα οι περιοχές στόχων είναι λιγότερο συντηρημένες στα θηλαστικά) και η συνεργατικότητα της πρόσδεσης (πολλά miRNAs μπορούν να προσδεθούν σε ένα γονίδιο).

Ακόμα, έχουμε μεγάλη επανεμφάνιση των περιοχών για διαφορετικά miRNAs στον στόχο 3'-UTRs. Ενώ η παρουσία και η απουσία των περιοχών στόχων συσχετίζεται με την γονιδιακή λειτουργία. Μια καλή μέθοδος ταξινόμησης είναι η SVM, με βάση την συντηρητικότητα.

Η κατασκευή δικτύων πειραματικά χαρτογραφημένων miRNA γίνεται σε μικρά κομμάτια, με κάθε δημοσίευση να προσθέτει ένα λιθαράκι στο δίκτυο, δηλαδή νέα μοριακή αλληλεπίδραση, η οποία μπορεί να παίρνει μέρος σε πάνω από ένα διαφορετικό δίκτυο miRNA. Η υπολογιστική πρόγνωση για εύρεση μοριακών αλληλεπιδράσεων αποτελείται από μεθόδους εύρεσης στατιστικά μη πιθανών γνωρισμάτων, τα οποία χαρακτηρίζουν την αλληλεπίδραση αυτή. Ένα παράδειγμα είναι η αναζήτηση πιθανών ρυθμιστών miRNA, αυτό μπορεί να πραγματοποιηθεί με την αναζήτηση των περιοχών πρόσδεσης ρυθμιστικών παραγόντων (transcription factor binding sites, TFBS) στις περιοχές υποκινητή ή ενισχυτή (enhancer) του γονιδίου, ή ακόμα και σε πιθανές αλληλεπιδράσεις μεταξύ πιθανών στόχων miRNA, συμπληρωματικούς σπόρους πρόσδεσης στην 3'UTR περιοχή των mRNA. Για την εύρεση συντηρημένων χαρακτηριστικών έχουμε παράλληλη ανάλυση γενωμάτων διαφορετικών ειδών. Οι αλληλεπιδράσεις που βρίσκουμε από αυτές τις αναλύσεις προστίθενται στις διαδικτυακές βάσεις δεδομένων miRNA, αφού πρώτα επιβεβαιωθούν, και μπορούμε να τις χρησιμοποιήσουμε σε μεταναλύσεις.

Πίνακας 8. Εργαλεία πρόγνωσης στόχων μορίων miRNA (45).

Εργαλείο	Οργανισμός	Τύπος Αλγορίθμου	Διάστημα Δημοσιεύσεων	Έκδοση	Διαδικτυακό ή Λογισμικό	Σύνδεσμος και Βιβλιογραφία
miRanda	Ζ	Μηχανική Μάθηση, υβριδοποίηση αντιστάθμισης, συμπληρωματική ταύτιση, ταύτιση εκβλάστησης	2003-2010	2010, v3.3a	Δ+Λ	https://bioweb.pasteur.fr/packages/pack@miRanda@3.3a/ (320)
RNAhybrid	Ζ	Υβριδοποίηση αντιστάθμισης, ταύτιση εκβλάστησης	2004-2006	2006, v2.1.2	Δ+Λ	https://bibiserv.cebitec.uni-bielefeld.de/rnahybrid (321)
TargetScan	Ζ	εξελικτική συντήρηση, ταύτιση εκβλάστησης	2005-2015	2018, v7.2	Δ+Λ	http://www.targetscan.org (322)
PicTar	Ζ	εξελικτική συντήρηση, συμπληρωματική ταύτιση	2005-2006	2007	Δ	https://pictar.mdc-berlin.de/ (323)
TargetFinder	Φ	Συμπληρωματική ταύτιση	2005-2010	2015, v1.7	Λ	https://github.com/carringtonlab/TargetFinder (324)
TarBase	Ζ+Φ	Ενοποίηση δεδομένων (ωμική), Με επιμέλεια κατηγοριοποίησης	2006-2018	2017, v8	Δ	https://carolina.imis.athena-innovation.gr/diana_tools/web/index.php?r=tarbasev8%2Findex (325)
RNA22	Ζ	Συμπληρωματική ταύτιση, υβριδοποίηση αντιστάθμισης	2006-2012	2015, v2.0	Δ	https://cm.jefferson.edu/rna22/ (326)
PolymiRTS	Ζ	Ενοποίηση δεδομένων (ωμική), πολυμορφισμοί, ταύτιση εκβλάστησης	2007-2014	2014, v3.0	Δ	https://compbio.uthsc.edu/miRSNP/ (327)
miRDB	Ζ	Μηχανική Μάθηση	2008-2016	2016, v5.0	Δ	http://www.mirdb.org (328)
miRGator	Ζ	Ενοποίηση δεδομένων (ωμική), συσχέτιση δεδομένων έκφρασης	2008-2013	2013, v3.0	Δ	http://mirgator.kobic.re.kr/ (329)
miRecords	Ζ	Με επιμέλεια κατηγοριοποίησης	2009	2013, v4	Δ	http://c1.accurascience.com/miRecords/ (330)
miRWalk	Ζ	Ενοποίηση δεδομένων (ωμική), εξόρυξη κειμένου	2011-2018	2018, v3.0	Δ	http://mirwalk.umm.uni-heidelberg.de/ (331)
mirDIP	Ζ	Ενοποίηση δεδομένων (ωμική)	2011-2017	2018, v4.1	Δ	https://ophid.utoronto.ca/mirDIP/ (332)
miRTarBase	Ζ+Φ	Με επιμέλεια κατηγοριοποίησης	2011-2020	2021, v9.0	Δ	https://mirtarbase.cuhk.edu.cn/~miRTarBase_2022/php/index.php (333)
psRNATarget	Φ	Συμπληρωματική ταύτιση, ταύτιση εκβλάστησης	2011-2018	2019, v2	Δ	https://zhaolab.org/psRNATarget/ (334)
miRBShunter	Ζ	Ανισοκαθίζηση	2017	2017, v0.2	Λ	https://github.com/TrabucchiLab/miRBShunter (335)
miRTar2GO	Ζ	Ανισοκαθίζηση, υβριδοποίηση αντιστάθμισης, ταύτιση εκβλάστησης	2017	2017	Δ	http://www.mirtar2go.org (336)

5. Κεφάλαιο 5 Η Διαδικασία Χαρακτηρισμού των Γονιδίων

5.1. Γονιδιακή Οντολογία (Gene Ontology (GO))

Ο όρος Οντολογία αναφέρεται στην δομημένη περιγραφή ενός πεδίου και επισημοποιεί την ορολογία σε αυτό, σε αυτή την περίπτωση μελετάμε τα γονίδια, δηλαδή την Γονιδιακή Οντολογία (GO). Στην ηλεκτρονική πηγή δημοσιεύσεων Google Scholar υπάρχουν μέχρι αυτή την στιγμή, 305.000 δημοσιεύσεις που περιέχουν τον όρο "Gene Ontology". Οι μεθοδολογίες της ωμικής προσφέρουν λεπτομερή καταγραφή και χαρακτηρισμό σταδίων παθογένειας και ανάπτυξης, μοριακής διάταξης ολόκληρου ιστού, ανασκόπηση των σχέσεων των πρωτεϊνικών δικτύων κ.α. Αυτές οι περίπλοκες μοριακές σχέσεις εξερευνώνται στη γενωμική σε διαφόρους τύπους κυττάρων. Για να μπορούν να χρησιμοποιηθούν τα παραγόμενα δεδομένα σε περαιτέρω έρευνες οι επιστήμονες κάνουν προσπάθειες για την ενοποίηση των δεδομένων εξαιρετικά υψηλής απόδοσης με τα συμπεράσματα ερευνών γονιδιακής ή από άλλους κλάδους της ωμικής επιστήμης. Η τεράστια βάση δεδομένων GO περιέχει αντιστοίχιση λειτουργιών με το κάθε γονίδιο (337).

Το Gene Ontology Consortium[49] παρέχει περιγραφικό λεξικό λειτουργιών για τις φυσιολογικές μοριακές και βιολογικές διεργασίες του οργανισμού, καθώς και στις υποκυτταρικές περιοχές με τις οποίες συνδέονται (338). Οι συντελεστές αυτής της ομάδας και κύριοι προμηθευτές χαρακτηρισμένων αρχείων στην GO είναι η πρωτεϊνική βάση δεδομένων UniProt (339), η Mouse Genome Informatics (340), η *Saccharomyces* Genome Database (341), η Wormbase (342), η Flybase (343), η dictyBase (344) και η TAIR (345). Η GO διαχωρίζεται σε τρεις κατηγορίες, δηλαδή τις κυτταρικές περιοχές (cellular components, CC), τις βιολογικές διεργασίες (biological processes, BP) και τις μοριακές λειτουργίες (molecular functions, MF).

Τα γονίδια κωδικοποιούν για προϊόντα γονιδίων (gene products), δηλαδή οι πρωτεΐνες και τα μη κωδικά μόρια RNA, τα οποία μπορούν να λάβουν μέρος σε χημικές διαδικασίες και να επιτελέσουν κάποιες μοριακές διεργασίες σε συγκεκριμένες περιοχές του κυττάρου για να λάβουν μέρος σε μια απώτερη βιολογική λειτουργία σε συνεργασία με άλλες μοριακές διεργασίες (346). Οι όροι GO μπορούν να προστεθούν σε μεμονωμένα αρχεία γονιδίων και

[49] http://wiki.geneontology.org/index.php/Main_Page

πρωτεϊνών, που βρίσκονται σε βάσεις δεδομένων βιολογικών αλληλουχιών. Είναι μια πλατφόρμα που συνεχίζει και βελτιώνει την ομαλή ενοποίηση ξεχωριστών δεδομένων μέσω επισημειώσεων (curations).

Για να δημιουργηθούν οι όροι υπάρχουν επιστήμονες που οργανώνουν και έχουν την επιμέλεια χειροκίνητης κατηγοριοποίησης (manual curation) βιολογικών δεδομένων. Εκτός από αυτούς υπάρχει η δυνατότητα σε κάποια πανεπιστήμια, μέσω ημερίδων (workshops) και μαθημάτων βιοπληροφορικής, να γίνει η δουλειά του χαρακτηρισμού μέσω των συμμετεχόντων, ένα λαμπρό παράδειγμα αυτής της προσπάθειας είναι ο διαγωνισμός CACAO[50] (Community Assessment of Community Annotation with Ontologies). Αυτός δίνει την δυνατότητα στους φοιτητές για ένα τρίμηνο να μάθουν πως γίνονται οι χαρακτηρισμοί και πως να δουλεύουν σε μια ερευνητική ομάδα, ενώ προσφέρει νέους χαρακτηρισμούς στην επιστημονική κοινότητα (347). Η εξόρυξη κειμένου (text mining) είναι ένας βασικός τρόπος να γίνει η κατηγοριοποίηση των δεδομένων χαρακτηρισμού και αυτά που συνάγονται από τον επιστήμονα έχουν την τυποποίηση IC (Inferred by Currator).

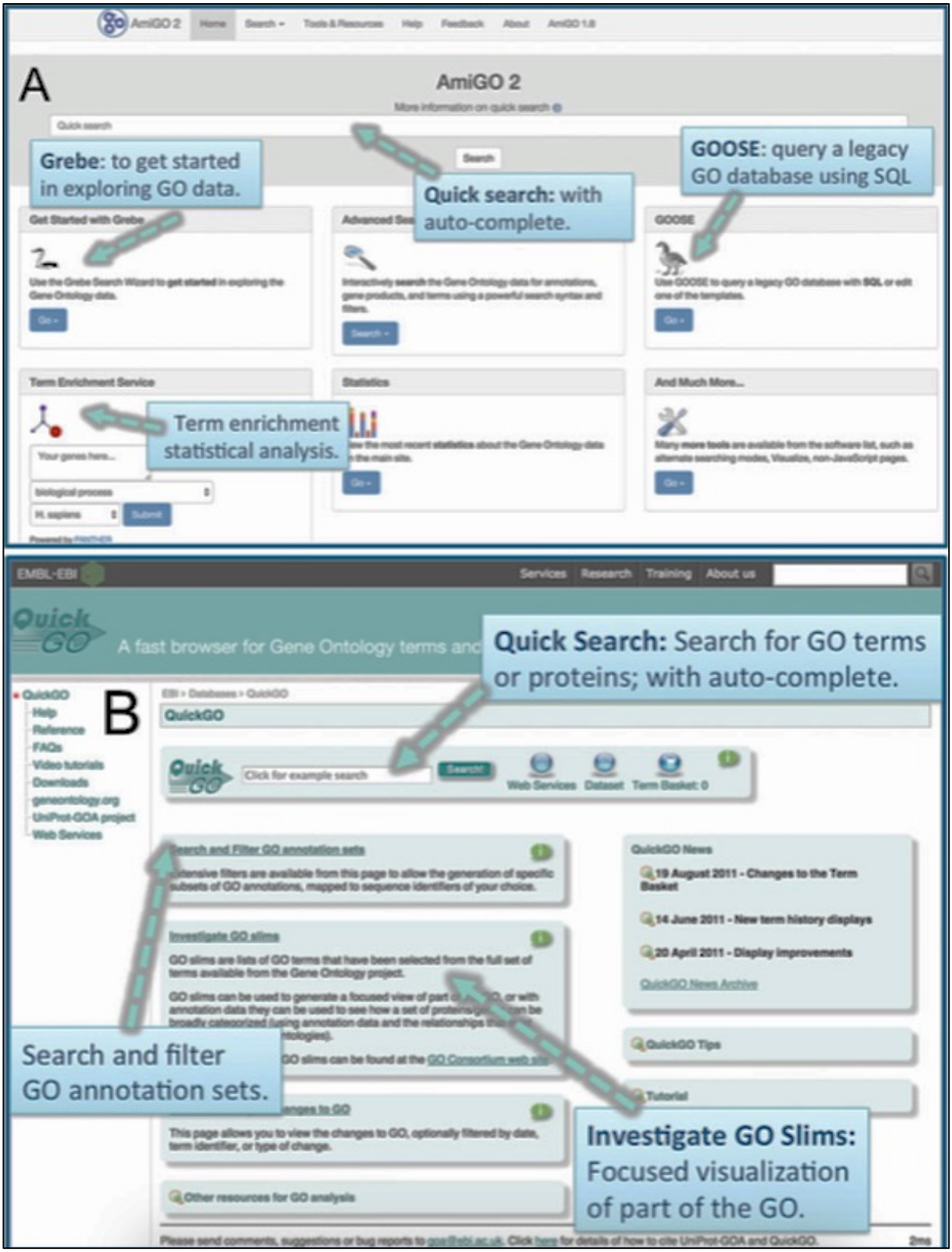
Οι όροι GO αποθηκεύονται σε γλώσσες που έχουν συγκεκριμένη μορφή, στις περισσότερες των περιπτώσεων. Κάθε όρος (γονίδιο, πρωτεΐνη, μη κωδικό RNA κ.λ.π.) ενώνεται με τα σχετικά αρχεία χαρακτηρισμού που αποθηκεύονται είτε σε μορφή GAF (Gene Association File), είτε σε μορφή GPAD (Gene Product Association Data). Οι γλώσσες της GO βελτιώνονται σε λογική έκφραση και πολυπλοκότητα, όπως και ήταν λογικό αφού βελτιώθηκε η υπολογιστική ικανότητα του λογισμικού και του υλικού εξοπλισμού. Η ανοιχτή βιοϊατρική γλώσσα οντολογίας (Open Biomedical Ontologies, OBO)[51], η οποία είναι κατανοήσιμη από τον άνθρωπο, σχεδιάστηκε για τις δομικές πληροφορίες και τα μεταδεδομένα (metadata) που σχετίζονται με τις γονιδιακές οντολογίες. Η επεξεργασία των οντολογιών γίνεται μέσω του εργαλείου OBO-Edit. Μια πιο καινούργια γλώσσα είναι η σημασιολογική πρότυπη διαδικτυακή γλώσσα οντολογίας (Web Ontology Language 2, OWL 2)[52], που χρησιμοποιεί το εργαλείο Protégé για την επεξεργασία των οντολογιών. Αυτή εδραιώθηκε λόγω της ευρείας αποδοχής από την επιστημονική κοινότητα και την καλή υποστήριξη των εργαλείων της (348), ακόμα καλύτερα υπάρχουν εργαλεία που μπορούν να

[50] http://gowiki.tamu.edu/wiki/index.php/Category:CACAO
[51] http://www.cs.man.ac.uk/~horrocks/obo/
[52] https://www.w3.org/TR/owl2-overview/

μεταφράσουν την πληροφορία από την μια γλώσσα στην άλλη (349). Αλλά υπάρχουν και οι μέθοδοι με την χρήση μιας διαδικτυακής πλατφόρμας για την σημασιολογική ανάλυση πολλαπλών οντολογιών, όπως είναι η BioPortal[53], η OLS και η OntoBee, (350-352).

Ο χρήστης μπορεί με ευκολία να αναζητήσει όρους χαρακτηρισμού γονιδιακών λιστών στην μηχανή αναζήτησης του AmiGO[54] (353) ή του QuickGO[55] (354), αφού οι πληροφορίες GO προέρχονται από την ίδια πηγή (**Εικόνα 33**). Για να εμφανιστούν τα αποτελέσματα πρέπει πρώτα να επιλέξεις τύπο αναζήτησης, να επιλέξεις γονιδιακό σύμβολο και να φιλτράρεις τα δεδομένα με βάση το είδος. Στην συνέχεια στέλνεις την ερώτηση (query) και μπορείς να χρησιμοποιείς τα αποτελέσματα για να κάνεις ομαδοποίηση με βάση το βιολογικό ενδιαφέρον. Μια δυνατότητα του GO είναι η ανάλυση εμπλουτισμού στα γονιδιακά σύνολα, αυτό γίνεται μέσω του εργαλείου PANTHER[56], αφού είναι συνδεδεμένο εργαλείο. Αλλά δεν είναι το μόνο εργαλείο για ανάλυση εμπλουτισμού που μπορεί να αλληλεπιδράσει με το GO και να παρουσιάσει ποια γονίδια είναι υπερεκφρασμένα ή υποεκφρασμένα. Για παράδειγμα, η πλατφόρμα DAVID έχει δεδομένα εισόδου λίστες μη ταξινομημένων γονιδίων, ενώ λαμβάνει υπόψη τις αλληλεπιδράσεις με την ορολογία GO. Εφαρμόζει μια παραλλαγή του ακριβή ελέγχου Fisher, που ονομάζεται σκόρ EASE, για να υπολογίσει τον εμπλουτισμό των όρων GO (143).

[53] https://bioportal.bioontology.org
[54] http://amigo.geneontology.org
[55] https://www.ebi.ac.uk/QuickGO/
[56] http://www.pantherdb.org/



Εικόνα 33. Η διεπαφή του χρήστη με την AmiGO και την QuickGO (355).

Οπτικοποίηση GO. Οι τρόποι που μπορείς να οπτικοποιήσεις τα δεδομένα των γονιδιακών οντολογιών είναι αρκετοί. Αυτοί συμπεριλαμβάνουν γράφους, επιφάνεια σημασιολογικής ομοιότητας, δενδρογράμματα και τα σύννεφα ορολογίας. Οι γράφοι GO έχουν τους όρους σε κόμβους και οι σχέσεις μεταξύ τους παρουσιάζονται με ακμές που έχουν κατευθυντικότητα. Υπάρχουν επιπλέον επιλογές επεξεργασίας του γράφου με το πρόσθετο Cytoscape[57]. Το εργαλείο και η πλατφόρμα Cytoscape σχεδιάστηκε για την ανάλυση και την οπτικοποίηση δικτύων, π.χ. κοινωνικών, οικονομικών, βιολογικών. Είναι ελεύθερης πρόσβασης, γι' αυτό χρησιμοποιείται σε πολλές άλλες εφαρμογές και ως πρόσθετο. Για την σημασιολογική ανάλυση ομοιότητας υπάρχουν πολλές αλγοριθμικές μέθοδοι (Πίνακας 9), τα αποτελέσματα παρουσιάζονται συνήθως με το εργαλείο REVIGO. Τα δενδρογράμματα είναι διαδραστικά, αφού μπορείς να επιλέξεις έναν όρο και να εμφανιστούν παραπάνω πληροφορίες για αυτόν. Τα σύννεφα ορολογίας είναι πολύχρωμα και μεγαλύτερα ή μικρότερα με βάση την σημαντικότητα για τον χρήστη, αυτό μπορεί να εφαρμοστεί με τα εργαλεία GOSummaries και REVIGO. Μπορούμε να εξάγουμε τα δεδομένα GO σε γνωστά λογισμικά οπτικοποίησης, όπως το πακέτο ggplot2 της R/Bioconductor και το εργαλείο gnuplot (356).

Πίνακας 9. Τα βέλτιστα μέτρα πληροφοριακού περιεχομένου (Information content, IC) που χρησιμοποιούνται στον υπολογισμό της σημασιολογικής ομοιότητας οντολογιών (357-364). Το Pfam[58] παρέχει επισημειωμένες ταξινομημένες αλληλουχίες με λειτουργική ανάλυση.

Σημασιολογική Ομοιότητα Οντολογιών	Μέτρα Ομοιότητας
Αλληλουχιών	SSDD, SimGIC, HRSS
Pfam	SORA, SSDD, SimGIC
Δεδομένα Έκφρασης	TCSS, SimGIC, SimIC, Best-Match-Avg (Resnik)
Αλληλεπίδραση πρωτεΐνης με πρωτεΐνη	TCSS, SimIC, Max (Resnik)

WebGestalt[59]. Οι αρχές Gestalt, δηλαδή η εγγύτητα, η ομοιότητα, το περίβλημα, η συμμετρία, το κλείσιμο, η συνέχεια και η σύνδεση, περιγράφουν πως οργανώνονται τα πρότυπα σε ομάδες. Αυτό το διαδικτυακό εργαλείο γονιδιακής ανάλυσης, που βασίζεται στις συγκεκριμένες αρχές, μπορεί να βοηθήσει στην οπτικοποίηση των αποτελεσμάτων και την επισήμανση της σημαντικότητας ή μη των δεδομένων. Η ανάλυση WebGestalt περιέχει την βάση δεδομένων WGDB, την γονιδιακή ή πρωτεϊνική λίστα ενδιαφέροντος (Gene Set Analysis Toolkit), και σε αυτήν μπορεί να γίνει η ανάλυση γονιδιακής οντολογίας και ο

[57] https://www.cytoscape.org/, http://cytoscapeweb.cytoscape.org
[58] https://pfam.xfam.org/
[59] http://www.webgestalt.org

διαχωρισμός σε διάφορες λειτουργικές κατηγορίες (145, 365). Αυτό το εργαλείο χρησιμοποιείται για την στατιστική ανάλυση των p-values από ανάλυση εμπλουτισμού στην GO.

Παρόλο που η ορολογία GO είναι πολύ χρήσιμη για την λειτουργική ταξινόμηση των γονιδίων, χρειάζονται περισσότερα δεδομένα συνδετικότητας και συσχετισμού για την δημιουργία σηματοδοτικών μονοπατιών (signaling pathways).

5.2. Ανάλυση Σηματοδοτικών Μονοπατιών (Pathway Analysis)

Η βιολογία των συστημάτων μελετάει ένα πολύπλοκο βιολογικό δίκτυο και όχι απλά ένα μεμονωμένο γονίδιο, επειδή παρόλο που μεγάλες διαφορές σε μεμονωμένα γονίδια μπορούν να έχουν σημαντικό βιολογικό ενδιαφέρον, αυτό ισχύει και για μικρότερες αλλά οργανωμένες αλλαγές. Γι'αυτό τον λόγο είναι σημαντική η σωστή και αποτελεσματική αποτύπωση νέων σηματοδοτικών μονοπατιών σε βάσεις δεδομένων που περιέχουν γνώση κυρίως για πρωτεΐνες με πληροφορίες για την ενεργοποίηση και την καταστολή αυτών. Αυτά τα βιολογικά μονοπάτια κατηγοριοποιούνται, με βάση τις πληροφορίες που περιέχουν, σε μεταβολικά, σηματοδοτικά, ρυθμιστικά κ.α. μονοπάτια. Για τα σηματοδοτικά μονοπάτια φτιάχνουμε γράφους, όπου ο κάθε κόμβος είναι μια πρωτεΐνη και πως αλληλεπιδρούν μεταξύ τους αυτές. Οι γράφοι περιέχουν κυρίως πρωτεΐνες που συνδέονται με βέλη για ενεργοποίηση και με κάθετη γραμμή για καταστολή. Μας ενδιαφέρει δηλαδή σε ποιες γνωστές σηματοδοτικές οδούς συμμετέχουν τα εξεταζόμενα γονίδια. Στην ρύθμιση γονιδιακής έκφρασης έχουμε πολλούς τρόπους να κατηγοριοποιήσουμε τις λειτουργίες των γονιδίων, οι οποίες είναι στόχος του ίδιου μεταγραφικού παράγοντα. Είναι επιθυμητό αυτά τα μονοπάτια να μπορούν να κατανοηθούν από ανθρώπους και μηχανές με σκοπό την ποσοτικοποιημένη υπολογιστική μοντελοποίηση. Υπάρχουν πολλοί τρόποι κατηγοριοποίησης των λειτουργιών των γονιδίων και οπτικοποίησης των αποτελεσμάτων σε δίκτυα πρωτεϊνικών αλληλεπιδράσεων. Θα αναφερθούμε και σε ανώτερου επιπέδου λειτουργικές σχέσεις π.χ. ασθενειών.

GSEA. Ένα πολύ χρήσιμο εργαλείο για την ανάλυση δεδομένων γονιδιακής έκφρασης, σε επίπεδο συνόλων γονιδίων, είναι το εργαλείο ανάλυσης γονιδιακού εμπλουτισμού (Gene Set Enrichment Analysis, GSEA[60]) (366). Με το GSEA μπορούμε να ελέγξουμε αν τα μέλη μιας προκαθορισμένης λειτουργικής κατηγορίας κατανέμονται τυχαία στην γονιδιακή λίστα ή αν

[60] https://www.gsea-msigdb.org/gsea/index.jsp

είναι ακραίες τιμές. Οπότε, μπορούμε να υπολογίσουμε τον βαθμό εμπλουτισμού (Estimation Score, ES), μέσω ενός διαφοροποιημένου ελέγχου KS (weighted Kolmogorov-Smirnov). Μπορούμε να πραγματοποιήσουμε εκτίμηση των επιπέδων σημαντικότητας του βαθμού εμπλουτισμού, μέσω ενός εμπειρικού φαινοτυπικού ελέγχου permutation. Ακόμα, πρέπει να εφαρμόσουμε μια διαδικασία διόρθωσης πολλαπλών ελέγχων, μέσω του FDR, σε αυτές τις εκτιμήσεις. Οπότε δεν χρειάζεται να έχεις σκληρό όριο μεταξύ διαφορικά εκφραζόμενων γονιδίων και μη, αλλά με την τιμή που υπολογίστηκε από τον έλεγχο t, ή άλλο έλεγχο, για τα γονίδια και ελέγχουμε αν μπορεί να εξηγηθεί από την συμμετοχή του στο γονιδιακό σύνολο. Γι' αυτό τον σκοπό μπορούμε να χρησιμοποιήσουμε οποιαδήποτε συλλογή γονιδιακού συνόλου, όμως είναι σύνηθες να χρησιμοποιούμε προκαθορισμένα γονιδιακά σύνολα, όπως αυτά που βρίσκονται στο GO ή στο KEGG (367). Με αυτήν την ανάλυση μπορούμε να ερευνήσουμε βιολογικά μονοπάτια και τις διεργασίες τους (368).

KEGG Pathways[61]. Η βάση δεδομένων για τα μονοπάτια της KEGG (Kyoto Encyclopedia of Genes and Genomes) αποτελείται από δίκτυα αλληλεπιδράσεων, όπου ο κάθε κωδικός στο δίκτυο αναπαριστά μια πρωτεΐνη/ ένζυμο και μπορούμε να δούμε ποια από τα γονίδια φαίνονται να έχουν χαρακτηριστικές ιδιότητες. Οι ενδιάμεσοι κόμβοι αναπαριστούν τους μεταβολίτες, δηλαδή τα προϊόντα και τα αντιδρώντα του μεταβολικού μονοπατιού. Η βάση δεδομένων KEGG περιέχει πληροφορίες γενωμικής λειτουργίας και δυνατότητα επεξεργασίας αυτών.

Εκτός από την KEGG υπάρχουν πολλές σημαντικές βάσεις δεδομένων μονοπατιών, που χρησιμοποιούνται κάθε μέρα από τους βιοπληροφορικούς. Αυτές συμπεριλαμβάνουν την Pathway Commons[62] (369), την Wikipathways[63] (370), την Panther Pathways, την Reactome Pathways[64] (371) και την PathwaySeq[65] κυρίως από δεδομένα RNA-Seq. Η βάση δεδομένων BioCyc/Metacyc περιέχει πάνω από 2,000 μονοπάτια προερχόμενα από πειράματα σε μόνο έναν οργανισμό το καθένα. Η SEED περιέχει υποσυστήματα που περιγράφουν τις μεταβολικές διεργασίες με προσεγμένες επισημειώσεις.

[61] https://www.genome.jp/kegg/pathway.html
[62] https://www.pathwaycommons.org/
[63] https://www.wikipathways.org/
[64] https://www.reactome.org
[65] https://rna-seqblog.com/pathwayseq-pathway-analysis-for-rna-seq-data/

ORA. Από μια λίστα DEGs μπορούμε να χρησιμοποιήσουμε την ανάλυση υπερ-αναπαράστασης (Over-Representation Analysis, ORA), για να εξετάσουμε την σημαντικότητα της σε κάποια λειτουργική κατηγορία με στατιστικά μέτρα όπως ο υπεργεωμετρικός έλεγχος.

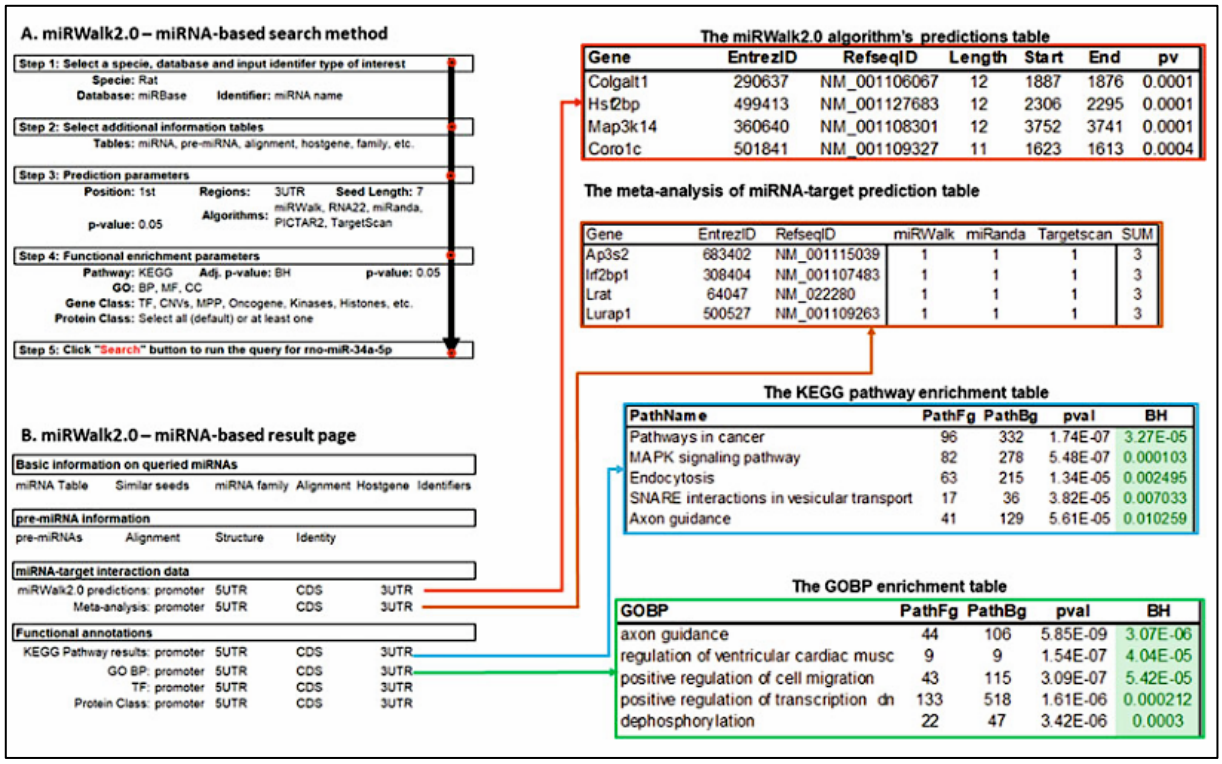
Σε μια δημοσίευση, των (372), πραγματοποιείται ανάλυση μονοπατιών με το εργαλείο Pathway Explorer (373), όπου τα γονίδια από την λίστα των DEGs χαρτογραφούνται σε γνωστά επισημειωμένα μονοπάτια. Αυτό το βήμα προσφέρει μια πρώτη εικόνα σε πόσα και ποια διαφορετικά μονοπάτια εμπλέκονται τα υπό εξέταση γονίδια, αλλά και τα ποσοστά της παρουσίας των γονιδίων σε αυτά. Στη συνέχεια, παρουσιάζεται το πιο πολύπλοκο στάδιο της μοντελοποίησης των χαρτογραφημένων γονιδίων στο επιλεγόμενο μονοπάτι, το οποίο πραγματοποιείται με το εργαλείο KEGG Converter (374).

miRWalk. Το εργαλείο miRWalk2.0 χρησιμοποιεί πειραματικώς επικυρωμένες αλληλεπιδράσεις των miRNA που σχετίζονται με γονίδια, ασθένειες, μονοπάτια, όργανα και πρωτεΐνες για την επεξεργασία των miRNA. Έχει δύο πρωτόκολλα, το VTM (Validated Target Module) για τους στόχους miRNA και το PTM για την πρόγνωση miRNA (Predcicted Target Module). Ο χρήστης μπορεί να εξάγει επικυρωμένα δεδομένα στόχων miRNA με την βοήθεια ευρετικών μεθόδων του πρωτοκόλλου VTM και της βάσης δεδομένων miR-TarBase. Τα αποτελέσματα μιας τέτοιας ανάλυσης παρουσιάζονται στην **Εικόνα 34**.

Τα αποτελέσματα αυτά μπορεί να χρησιμοποιηθούν για άλλα βήματα μετα-ανάλυσης με επιπλέον εργαλεία, όπως το DAVID και το IPA[66] (Ingenuity Pathway Analysis) (375), το οποίο είναι διαδικτυακό εργαλείο, με εμπορική διάθεση, και είναι κατάλληλο για σηματοδοτική ανάλυση δεδομένων προερχόμενων από διάφορες αναλύσεις, όπως η RNA-Seq, η miRNA-Seq και η miRNA από μικροσυστοιχίες (376).

Άλλα σημαντικά εργαλεία για ανάλυση μονοπατιών των μορίων miRNA είναι το miRPath του DIANA, το miRHrt, το miRnalyze, το miRNApath και το miRPathDB.

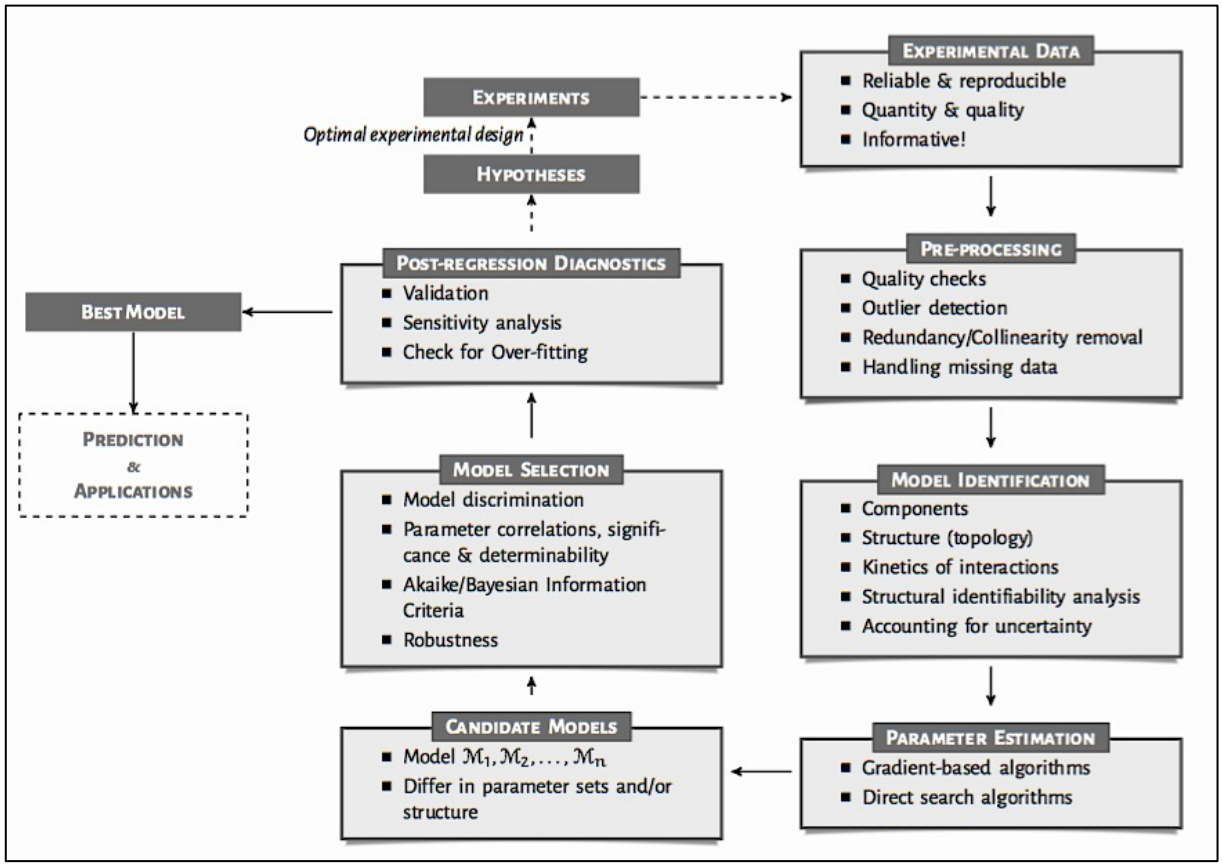
[66] https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/



Εικόνα 34. Το εργαλείο miRWalk και σηματοδοτικοί οδοί για τα μόρια miRNA (376).

6. Κεφάλαιο 6 Μαθηματική Προτυποποίηση

Η ανάλυση συσχέτισης παρουσιάζει καλύτερα αποτελέσματα όταν εξετάζονται μαζί δεδομένα από περισσότερους οργανισμούς, επειδή ακόμα και σε έναν από αυτούς να βρεθεί μια πρωτεΐνη με τις δυο περιοχές συσχετισμένες ή λειτουργικά συνδεδεμένες, αυτή μπορεί να αναγνωριστεί και στους υπόλοιπους, οπότε έχουμε περισσότερες συσχετίσεις στο υπό εξέταση γένωμα. Υπάρχει η δυνατότητα να συμπληρώσουμε τα κενά στην κατανόηση μας και την ερμηνεία της κάθε ανάλυσης μέσω μιας συμπληρωματικής ανάλυσης (**Εικόνα 35**). Ο όρος πολυσυγραμμικότητα (multicollinearity) χρησιμοποιείται όταν μια είσοδος των δεδομένων είναι πολύ συσχετισμένη με τουλάχιστον μια άλλη είσοδο και τότε παρατηρούνται αυξημένα ποσοστά σφάλματος στις εκτιμήσεις από τις βάσεις δεδομένων.



Εικόνα 35. Μια ροή έργων μηχανιστικής μοντελοποίησης (377).

Στη διαφορική ανάλυση μπορούμε να συγκρίνουμε δύο ή παραπάνω δίκτυα (378), έτσι ώστε να αναγνωριστούν οι συντηρημένες περιοχές, οι οποίες μπορεί να δείχνουν σημαντικότητα, και να αναγνωριστούν οι περιοχές που είναι διαφορετικές μεταξύ άλλων συνθηκών (π.χ. μεταξύ φυσιολογικών και καρκινικών κυττάρων), έτσι ώστε να βρεθεί η υποκείμενη αιτία της ασθένειας. Για την ανάλυση αυτών των δικτύων είναι εξαιρετικά χρήσιμη η εφαρμογή DyNet

Cytoscape , που είναι μια από τις εφαρμογές που χρησιμοποιούν το Cytoscape, και άλλα παρόμοια εργαλεία σύγκρισης (379).

Οπτικοποίηση. Εκτός από το πολύ χρήσιμο Cytoscape, στις διάφορες μορφές του, υπάρχουν και άλλα εργαλεία για την οπτικοποίηση μονοπατιών, π.χ. το λογισμικό Graph Visualization Software (GraphViz), το οποίο δέχεται ως είσοδο ένα αρχείο που περιέχει όλες τις αντιδράσεις του τελικού δικτύου.

Γονιδιακό ρυθμιστικό δίκτυο (Gene Regulatory Network, GRN). Τα GRN είναι ένας τύπος βιολογικών δικτύων, όπως τα δίκτυα αλληλεπιδράσεων μεταξύ πρωτεϊνών (PPI), τα δίκτυα συν-έκφρασης και τα δίκτυα ρύθμισης μεταγραφής (transcriptional regulatory network, TRN). Τα δίκτυα PPI παρουσιάζουν τις σχέσεις μεταξύ των πρωτεϊνών, τα δίκτυα συν-έκφρασης τις σχέσεις μεταξύ των γονιδίων και τα δίκτυα TRN αφορούν τους μεταγραφικούς παράγοντες που ρυθμίζουν την γονιδιακή έκφραση. Από την άλλη πλευρά, τα δίκτυα GRN παρουσιάζουν τις σχέσεις των γονιδίων με οποιοδήποτε άλλο στοιχείο που έχει ρόλο στην ρύθμισή του.

Παλινδρόμηση (Regression). Η παλινδρόμηση, που είδαμε και σε προηγούμενη ενότητα, μπορεί να αναδείξει τα δίκτυα συσχέτισης. Για παράδειγμα, σε ένα δίκτυο συν-έκφρασης, η μεθυλίωση γονιδίων μεταξύ των διαφορετικών δειγμάτων από δεδομένα έκφρασης γονιδίων. Η κατάλληλη ανάλυση παλινδρόμησης εξαρτάται από το ποιο φαινόμενο περιγράφουμε. Για παράδειγμα, η πολλαπλή γραμμική παλινδρόμηση, για συνεχείς τιμές, και η βηματική (stepwise) έχει καλή εφαρμογή στα μεγάλα δεδομένα, η οποία είναι μια μέθοδος συνδυασμού εισαγωγής και εξαγωγής δεδομένων. Σε περίπτωση που δεν έχουμε γραμμική συσχέτιση χρησιμοποιούμε τις πολυωνυμικές μεθόδους στο μοντέλο. Εκτός από τις stepwise υπάρχουν και οι πολύ διαδεδομένοι μέθοδοι εκτίμησης της συρρίκνωσης (shrinkage), που επιτρέπει μείωση των εκτιμήσεων του πλήρους μοντέλου, όπως ο συντελεστής συρρίκνωσης LASSO (Least Absolute Shrinkage and Selection Operator), ο κορυφής (ridge) και ο ελαστικού δικτύου (elastic net). Στην **Εικόνα 36**, παρουσιάζονται οι παράμετροι από διάφορα σχετικά σύγχρονα μοντέλα παλινδρόμησης, όπως εμφανίζονται στην δημοσίευση των ερευνητών (380). Οι ερευνητές συμπεριλαμβάνουν ακόμα, στην προαναφερθείσα δημοσίευση, μια βιβλιογραφική ανασκόπηση με το πόσες δημοσιεύσεις είχαν χρησιμοποιήσει το κάθε μοντέλο ανάλυσης μέχρι το 2012.

Model $p(\beta_j \mid \omega)$	Hyperparameters	Prior variance $\mathrm{Var}(\beta_j \mid \omega)$	Solution for scale/variance parameter
Bayesian ridge regression			
$N(\beta_j \mid 0, \sigma^2_\beta)$	σ^2_β	σ^2_β	$\sigma^2_\beta = \frac{h^2\sigma^2_p}{\mathrm{MS}_X}$
Bayesian LASSO			
$\mathrm{DE}(\beta_j \mid \sigma^2, \lambda^2)$	$\{\sigma^2, \lambda^2\}$	$2\frac{\sigma^2}{\lambda^2}$	$\lambda = \sqrt{2\frac{(1-h^2)}{h^2}\mathrm{MS}_X}$
BayesA			
$t(\beta_j \mid \mathrm{d.f.}_\beta, S_\beta)$	$\{\mathrm{d.f.}_\beta, S_\beta\}$	$\frac{\mathrm{d.f.}_\beta S^2_\beta}{\mathrm{d.f.}_\beta - 2}$	$S^2_\beta = \frac{(\mathrm{d.f.}_\beta - 2)}{\mathrm{d.f.}_\beta}\frac{h^2\sigma^2_p}{\mathrm{MS}_X}$
Spike–slab			
$\pi \times N\left(\beta_j \mid 0, \frac{\sigma^2_\beta}{\tau}\right) + (1-\pi)N(\beta_j \mid 0, \sigma^2_\beta)$, $(\tau > 1)$	$\{\pi, \sigma^2_\beta, \tau\}$	$\sigma^2_\beta \times \left[1 + \pi\frac{(1-\tau)}{\tau}\right]$	$\sigma^2_\beta = \left[\frac{\tau}{\tau + \pi(1-\tau)}\right]\frac{h^2\sigma^2_p}{\mathrm{MS}_X}$
BayesC			
$\pi \times 1(\beta_j = 0) + (1-\pi)N(\beta_j \mid 0, \sigma^2_\beta)$	$\{\pi, \sigma^2_\beta\}$	$\sigma^2_\beta \times (1-\pi)$	$\sigma^2_\beta = \frac{1}{(1-\pi)}\frac{h^2\sigma^2_p}{\mathrm{MS}_X}$
BayesB			
$\pi \times 1(\beta_j = 0) + (1-\pi)t(\beta_j \mid \mathrm{d.f.}_\beta, S_\beta)$	$\{\pi, \mathrm{d.f.}_\beta, S_\beta\}$	$(1-\pi)\frac{\mathrm{d.f.}_\beta S^2_\beta}{\mathrm{d.f.}_\beta - 2}$	$S^2_\beta = \frac{1}{(1-\pi)}\frac{(\mathrm{d.f.}_\beta - 2)}{\mathrm{d.f.}_\beta}\frac{h^2\sigma^2_p}{\mathrm{MS}_X}$

$MS_x = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - \bar{x}_j)^2$ where $x_{ij} \in (0, 1, 2)$ represents number of copies of the allele coded as one at the j^{th} ($j = 1,\ldots,p$) locus of the i^{th} ($i = 1,\ldots,n$) individual, and $\bar{x}_j$ is the average genotype at the j^{th} marker.

Εικόνα 36. Βασισμένη σε προηγούμενα δεδομένα πυκνότητα και διακύμανση των δεικτών επιδράσεως και οι προτεινόμενες εξισώσεις για την επιλογή τιμών υπερπαραμέτρων, σύμφωνα με το μοντέλο παλινδρόμησης (380).

6.1. Μοντελοποίηση των Σηματοδοτικών Οδών (Pathways)

Τα σηματοδοτικά μονοπάτια των πρωτεϊνών αποτελούν ένα πολύπλοκο σύστημα με συγχρονισμό λειτουργιών σε διάφορα επίπεδα. Οι επιστήμονες αρχικά εξετάζουν τα υποσυστήματα ενός κυττάρου και το αναλύουν σε σύστημα εισόδου – εξόδου. Στην συνέχεια, αναλύουν τις αλληλεπιδράσεις των λειτουργιών του και η κυτταρική συμπεριφορά ελέγχεται στο σύνολό της. Για την συστημική ανάλυση των σηματοδοτικών μονοπατιών έχουμε δυναμική συμπεριφορά και έτσι πιο περίπλοκη. Οι μέθοδοι θεωρίας συστημάτων χρησιμοποιούνται για αυτές τις αναλύσεις. Είναι μια δυναμική μοντελοποίηση των χημικών αντιδράσεων μεταξύ γονιδίων και γονιδιακών προϊόντων, οπότε είναι πολύπλοκη διαδικασία.

Η δυνατότητα ορθής εύρεσης των σχετικών αλληλεπιδράσεων των μονοπατιών από μια λίστα DEGs ή πρωτεϊνών είναι σημαντική για να μετατρέψουμε τα δεδομένα έκφρασης σε βιοϊατρικά συμπεράσματα. Για παράδειγμα, έστω ότι έχουμε 100 προβληματικά γονίδια, μεταφέρουμε την πληροφορία σε πρωτεΐνες και θέλουμε να αναλύσουμε τις αλληλεπιδράσεις του βιολογικού δικτύου, τότε για κάθε ακμή έχω μια αντίδραση, άρα μπορώ να φτιάξω, για την κάθε ακμή μια διαφορετική εξίσωση. Δηλαδή, μπορούμε να αναπαραστήσουμε το μαθηματικό μοντέλο ενός μονοπατιού μέσω ενός συστήματος απλών διαφορικών εξισώσεων, οι οποίες περιγράφουν τον ρυθμό όλων των αντιδράσεων σε ένα μονοπάτι. Με τις σωστές παραμέτρους, π.χ. τις αρχικές συγκεντρώσεις των μορίων και τις σταθερές της κίνησης

αντιδράσεων, ο υπολογιστής μπορεί με προσομοιώσεις να υπολογίσει τις συγκεντρώσεις μορίων από διαφορετικά είδη σε κάποιο συγκεκριμένο μονοπάτι.

Μία προσέγγιση είναι η ακόλουθη: η στατιστική επεξεργασία, η αναπαράσταση και ανάλυση του δικτύου μέσω αλγορίθμου, η δημιουργία μονοπατιού και, ίσως διασταύρωση με τα πειραματικά δεδομένα. Στην συνέχεια ανακαλύπτουμε τα αντικειμενικά μονοπάτια (normal pathways), τα αδύνατα μονοπάτια και τα εναλλακτικά μονοπάτια (alternative pathways). Έτσι, μπορούμε να ανακαλύψουμε τους κόμβους σύγκρουσης και να τις ελαχιστοποιήσουμε. Στην συνέχεια, μπορεί να γίνει η οπτικοποίηση με τις προαναφερθείσες μεθόδους.

Reactome. Η Reactome, περιέχει μια επισημειωμένη ανθρώπινη βάση δεδομένων μονοπατιών (π.χ. μεταβολικών, σηματοδοτικών και συσχέτισης ασθενειών) και συμπεριλαμβάνει μονοπάτια από άλλα 20 είδη που συσχετίζονται με τα ανθρώπινα μονοπάτια μέσω της αναγνώρισης ορθόλογων πρωτεϊνών. Έχει την δυνατότητα ανάλυσης εμπλουτισμού και καλή αναπαράσταση των μονοπατιών, αφού περιέχει πολλά σύμβολα για την αναπαράσταση διαφορετικών μορίων και της συσχέτισης μεταξύ τους (φυσικές αλληλεπιδράσεις, ρύθμιση κ.α.). Ο χρήστης μπορεί, ακόμα, να ενσωματώσει διάφορα δεδομένα, π.χ. δεδομένα έκφρασης, σε ένα υπάρχον μονοπάτι.

STRING. Το STRING προσφέρει μηχανή αναζήτησης σε βάση δεδομένων λειτουργικών αλληλεπιδράσεων με δεδομένα σε πάνω από 2000 γενώματα (381), η αναζήτηση σε αυτή μπορεί να γίνει μέσω του ονόματος του γονιδίου, μέσω της αλληλουχίας υπό εξέταση ή μέσω του αριθμού καταχώρησης στην βάση δεδομένων. Τα δεδομένα περιέχουν πληροφορίες για την γονιδιακή γειτονιά, φυλογενετικό προφίλ, συν-έκφρασης, όνομα πηγής, προσδιορισμένες αλληλεπιδράσεις μεταξύ πρωτεϊνών από πειραματική διεργασία κ.α. δεδομένα. Τα αποτελέσματα βαθμολογούνται με ένα συγκεκριμένο τρόπο σε σχέση με την επικύρωση των αλληλεπιδράσεων μεταξύ πρωτεϊνών και γνωστών μονοπατιών. Ακόμα, εκτός από το φιλικό προς το χρήστη περιβάλλον, το STRING παρέχει στον χρήστη την δυνατότητα να εξετάσει λεπτομερώς τα επιμέρους κομμάτια αυτών των αλληλεπιδράσεων. Ένα παρόμοιο εργαλείο είναι το GeneMANIA (382), το οποίο διαφέρει στο ότι εξάγει τα δεδομένα από διαφορετικές πηγές από το STRING και εξετάζει εννιά συγκεκριμένα μοντέλα οργανισμών (383).

SimBiology (add-in του MATLAB). Είναι πρόσθετο του MATLAB® για την δυναμική απεικόνιση των βιολογικών δεδομένων και τον εντοπισμό των διάφορων αλληλεπιδράσεων

για την κατασκευή/συμπλήρωση ενός μονοπατιού, ενώ ελέγχει και πολλές άλλες δυνατότητες. Οι διαφορικές εξισώσεις (Ordinary Differential Equations, ODEs) χρησιμοποιούνται για την προσομοίωση του μοντέλου. Ακόμα, υπάρχουν τρόποι να κατασκευαστούν κινητικά μοντέλα ενζύμων μέσω υπαρχόντων παραμέτρων κίνησης σε βάσεις δεδομένων ενζυμικών αντιδράσεων, όπως η BRENDA (384) και η SABIO-RK (385). Αυτά τα δεδομένα μπορούν να μεταφερθούν σε ένα μοντέλο του SimBiology.

CellDesigner[67] *(freeware).* Το πρόγραμμα CellDesigner έχει πολύ καλή διασύνδεση με την βάση δεδομένων Panther Pathway (386, 387). Ένα αρχείο CD του CellDesigner μπορεί να εισαχθεί στο πρόσθετο SimBiology του προγράμματος MATLAB και υπάρχουν δύο επιλογές, είτε να μπορείς να αλλάξεις μόνο την μορφοποίηση του μονοπατιού ή να μπορείς να χρησιμοποιήσεις τις εξισώσεις του MATLAB. Οι χάρτες σηματοδοτικών μονοπατιών χρησιμοποιούνται για την οπτικοποίηση γεγονότων που συμβαίνουν το ένα μετά το άλλο, αλλά δεν περιέχει δεδομένα χώρου. Για να έχουμε πρόσβαση σε περισσότερες πληροφορίες χρησιμοποιούμε την γλώσσα της συστημικής βιολογίας (Systems Biology Markup Language, SBML). Το πρόγραμμα CellDesigner και το LibSBGN υποστηρίζει την αυτόματη παραγωγή γραφικών αναπαραστάσεων σύμφωνα πρότυπο γραφιστικής συντόμευσης (Systems Biology Graphical Notation, SBGN) για την οπτική αναπαράσταση βιοχημικών και κυτταρικών διεργασιών (388), με σκοπό την κατανόηση των υποκείμενων μηχανισμών (389). Η γλώσσα SBML είναι βασισμένη στην XML και χρησιμοποιείται για την ανταλλαγή μαθηματικών μοντέλων μονοπατιών (390). Επιπροσθέτως, η βάση δεδομένων BioModels περιέχει μοντέλα SBML για πολλά μονοπάτια και υπάρχουν πολλά λογισμικά για την προσομοίωση των μοντέλων SBML στον υπολογιστή (391).

Άλλα εργαλεία κατασκευής μονοπατιών, όπως το Pathway Editor, το Knowledge Editor και το Map Editor, μπορούν να χρησιμοποιηθούν για επισημείωση από κάποιον επιστήμονα. Επιπροσθέτως για την κατασκευή μονοπατιών υπάρχουν και εργαλεία για εξόρυξη δεδομένων από δημοσιεύσεις, όπως το Pathway Studio, το Pathway Finder και το PubGene, όπου η συσχέτιση γίνεται μέσω τεχνικών επεξεργασίας φυσικής γλώσσας (Natural language processing, NLP).

[67] http://www.celldesigner.org/models.html

Το έργο E-Cell[68]. Το έργο E-Cell αφορά την δυναμική μοντελοποίηση και ανακατασκευή βιολογικών φαινομένων στον υπολογιστή (392). Με αυτό μπορεί ο χρήστης να προσομοιώσει, μέσω διαφορικών εξισώσεων, πολλές βιολογικές λειτουργίες, συμπεριλαμβανομένου δικτύων PPI και ρύθμισης γονιδιακής έκφρασης (393). Γι' αυτό τον σκοπό, υπάρχουν εκατοντάδες εξισώσεις αντιδράσεων για να χρησιμοποιηθούν σε μια προσομοίωση (394).

Εικόνα 37. Εφαρμογές στην κοινότητα του SBGN (395).

[68] https://www.e-cell.org

Model Systems	Parametrisation	Typical Predictions	Advantages	Disadvantages
Static Network Models				
• Protein interaction networks • Gene networks • Large- and small-scale networks	• Edges representing relationships • Edge weights can quantify strengths/ confidences	• Important nodes • Important edges, *i.e.* interactions • Important groups of nodes or sub-networks (clusters, modules, motifs)	• Nice synthesis of known information • Quick and easy	• No dynamics, typically
• Genome-scale metabolic networks	• Network topology	• Reachability	• Work with "draft" reconstructions	• No dynamics
Dynamic Kinetic Models				
• Small-scale biological processes • Signalling and regulatory networks	• Detailed kinetic parameters, *e.g.* v_{max}, K_M	• Concentrations of different species • Reaction rates • Sensitivity analyses	• Mechanistic insights • Capture dynamics—critical in biology	• Parameter estimation is challenging • Curse of dimensionality • Limited model size
Boolean Models				
• Small-scale biological processes • Signalling and regulatory networks	• Rules of interaction	• Regulatory states • Key interactions	• Mechanistic insights • Capture dynamics • Capture biological knowledge effectively	• Biological systems are seldom discrete
Constraint-based Models				
• Genome-scale metabolic networks	• Stoichiometry • Uptake rates, biomass composition	• Possible metabolic states • Key genes, proteins or reactions • Targets for manipulation	• No kinetics required • Enable true systems-level modelling • Mechanistic insights • Versatile	• Not easy to capture metabolite concentrations • More complex methods required to integrate/infer regulation/dynamics

Εικόνα 38. Διάφοροι τύποι μοντελοποίησης (377).

7. Κεφάλαιο 7 Η Γενωμική στην Εποχή των Big Data

Η ανάπτυξη της ML, η βελτίωση σε λογισμικό-υλικά συστήματα έκανα και η ζήτηση μεγάλων εταιριών και οργανισμών για αποθήκευση, οργάνωση και ανάλυση των δομημένων, ήμι-δομημένων και, συχνότερα, μη δομημένων δεδομένων. Τα λογισμικά βάσεων δεδομένων όπως το Hadoop, το Spark, το NoSQL κ.α. είναι πλαίσια ανοιχτού κώδικα για την ανάλυση μεγάλων συνόλων δεδομένων (Big Data). Η αρχιτεκτονική αυτών των συστημάτων είναι πολύ σημαντική για την σωστή λειτουργία του και για να αντεπεξέλθει το περιβάλλον στις ανάγκες του συστήματος, γι'αυτό για κάθε νέα υποδομή σχεδιάζεται ένα προσαρμοσμένο περιβάλλον από το φυσικό στρώμα (εξοπλισμός, υλικό), μέχρι και το υπολογιστικό στρώμα (εργαλεία για ανάλυση, οπτικοποίηση και σύστημα αναφορών), τα οποία πρέπει να είναι αλληλένδετα και άμεσα προσπελάσιμα από το δίκτυο και τους εξουσιοδοτημένους χρήστες. Γι'αυτό η σύνδεση στο δίκτυο γίνεται μέσω οπτικών ινών.

Ένας άλλος παράγοντας που μας ενδιαφέρει πολύ είναι η ασφάλεια των δεδομένων από κυβερνοεπιθέσεις και μη εξουσιοδοτημένες προσπελάσεις, ένας σκοπός που δυσχεραίνεται με τα Big Data. Η αποθήκευση των Big Data έχει ιδιαίτερο χαρακτήρα και διαφορετικές εφαρμογές, αφού τα δεδομένα πρέπει να εξαχθούν, να μετασχηματιστούν σε μια κοινή «γλώσσα» σύμφωνα με τους κανόνες του υπολογιστικού περιβάλλοντος, να επεξεργαστούν, να αναλυθούν, να επιβεβαιωθούν, να επικαιροποιηθούν και ότι άλλο χρειάζεται η συγκεκριμένη υλοποίηση. Ο διαδικτυακός χώρος αποθήκευσης, συνήθως διαχωρίζεται στους SAN (Storage Area Network), που συνήθως συνδέονται με οπτικές ίνες, και τους NAS (Network Attached Storage), που συνήθως συνδέονται με καλώδιο Ethernet. Συχνά χρησιμοποιείται η κατανεμημένη εφαρμογή με τον αλγόριθμο MapReduce για την επεξεργασία των δεδομένων και τη βελτιστοποίηση της οργάνωσης μεγάλων ροών δεδομένων.

Οι λειτουργικές βάσεις δεδομένων (ωμικές στην παρούσα εργασία) και οι αναλυτικές βάσεις δεδομένων, που υποστηρίζουν αναλυτικές πλατφόρμες για την εξόρυξη των δεδομένων, είναι η ραχοκοκαλιά των Big Data. Το Σύστημα Σχεσιακής Διαχείρισης Βάσης Δεδομένων (Relational Database Management System, RDBMS) έχει τον ρόλο της οργάνωσης των σημείων δεδομένων με προκαθορισμένες σχέσεις για γρήγορη προσπέλαση. Οι πιο γνωστές βάσεις δεδομένων για την προσπέλαση στα αποθηκευμένα δεδομένα είναι η IBM Db2, η Oracle Db, η HIVE (είναι ελεύθερης πρόσβασης και ειδικευμένη για μεγάλα κέντρα δεδομένων), η Microsoft SQL Server, η Microsoft Azure SQL Database, η MySQL και η

Teradata, όπου η SQL (Structured Query Language) είναι μια σχετικά εύκολα αναγνώσιμη και ερμηνεύσιμη γλώσσα προγραμματισμού.

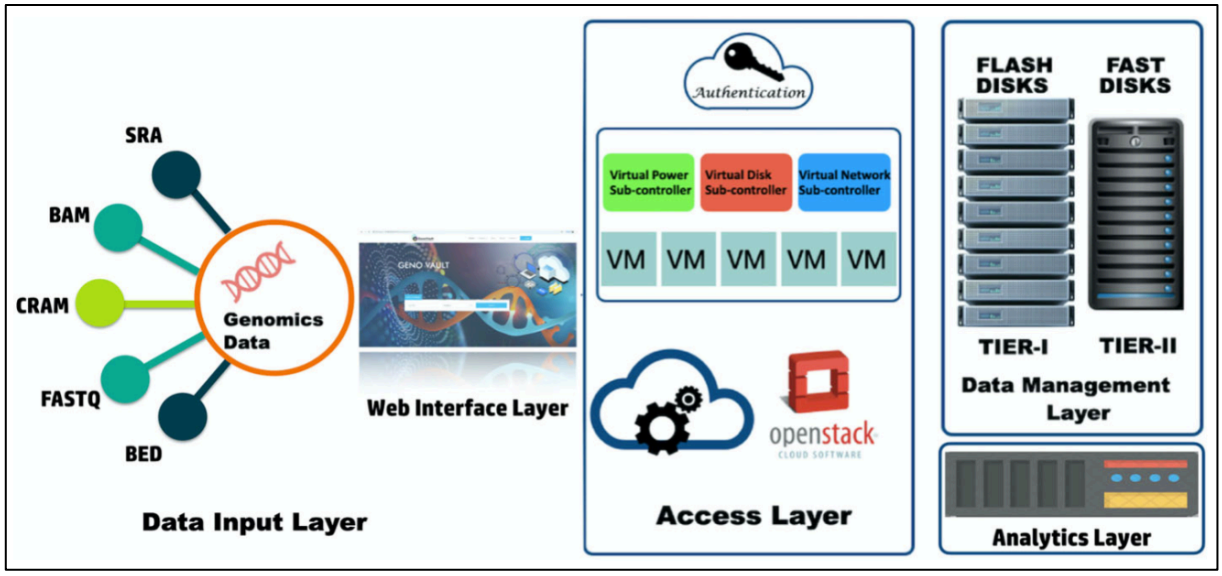
Τα εργαλεία και τα λογισμικά, καθώς και η δυνατότητα ταχείας ανάπτυξης νέων εργαλείων και διαχείρισης του όλου συστήματος είναι βασικό κομμάτι μιας καλής υλοποίησης αρχιτεκτονικής Big Data.

Σύννεφο (cloud). Το υπολογιστικό σύννεφο έχει επεκτείνει τις δυνατότητες των μεγάλων δεδομένων. Οι τεχνολογίες συννέφου με βιοπληροφορικές λύσεις για την ανάλυση δεδομένων από δεδομένα υψηλής απόδοσης απαιτούν σημαντικές υπολογιστικές και αποθηκευτικές ικανότητες. Η υποστήριξη αυτών των τεχνολογιών γίνεται από μεγάλες κεντρικές εγκαταστάσεις, οι οποίες αδυνατούν να αντεπεξέλθουν στην συνεχή και αυξανόμενη ροή των δεδομένων (396). Με τις τεχνολογίες συννέφου περισσότερες υπηρεσίες είναι προσβάσιμες πλέον από μικρότερα κέντρα ερευνών και μεμονωμένους ερευνητές (396, 397). Μια βιοπληροφορική πλατφόρμα συννέφου είναι η Cloud BioLinux[69] μπορεί να υποστηρίξει την κατασκευή και διάθεση εικονικών μηχανών (VM) και περιέχει μια μεγάλη γκάμα εργαλείων ανάλυσης αλληλούχισης που είναι συμβατό με το Amazon EC2 αλλά και προσωπικά και ακαδημαϊκά περιβάλλοντα συννέφου, όπου είναι σύνηθες να χρησιμοποιηθούν οι πλατφόρμες συννέφου Eucalyptus και OpenStack (389, 397, 398). Αλλά πολλές ακόμα εταιρίες δραστηριοποιούνται στην αξιοποίηση των δυνατοτήτων του συννέφου, όπως η Aspera (που ειδικεύεται στη γρήγορη μεταφορά δεδομένων μεταξύ δικτύων και συνεργάζεται με την Amazon για να προσφέρει υπηρεσίες διαχείρισης δεδομένων στο cloud), η FileCatalyst και η Data Expedition.

Genovault. Η Genovault είναι μια πολύ ενδιαφέρουσα εφαρμογή ιδιωτικού συννέφου αποθήκευσης γενωμικών δεδομένων βασισμένη στην OpenStack, με την χρήση της βάσης δεδομένων MySQL. Τα δεδομένα είτε προέρχονται από τις γνωστές γενωμικές βάσεις δεδομένων, είτε από ερευνητικά εργαστήρια και μπορεί ο χρήστης να ανεβάσει τα δεδομένα της αλληλούχισης με την διαδικτυακή διεπαφή ή με την διεπαφή JavaFX του Genovault. Το Picard είναι ένα από τα εργαλεία που χρησιμοποιούνται για την επικύρωση των αρχείων (399).

---

[69] http://cloudbiolinux.org



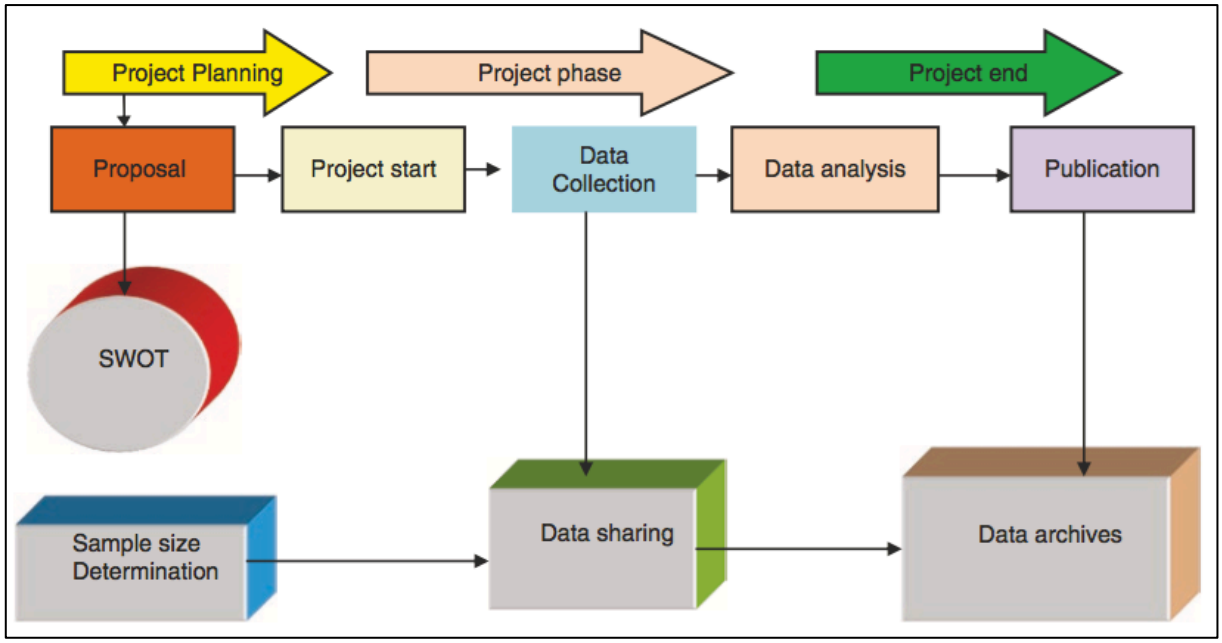
Εικόνα 39. Η αρχιτεκτονική της Genovault, όπου φαίνονται τα συστήματα και τα υποστρώματα (399).

Η περίπτωση της Novartis. Το Ινστιτούτο Βιοϊατρικών Ερευνών της φαρμακοβιομηχανίας Novartis (Novartis Institutes for Biomedical Research, NIBR) είναι μια καλή εφαρμογή χρήσης μεγάλων δεδομένων στην παγκόσμια βιοϊατρική έρευνα. Σε πρώτο στάδιο παραγωγής νέων φαρμάκων μεγάλο ρόλο ανάλυση της συμπεριφοράς της ασθένειας που θέλουμε να αντιμετωπίσουμε, έτσι ώστε να καθοριστεί η δομή της συγκεκριμένης έρευνας και οι παράγοντες που σχετίζονται με την συγκεκριμένη ασθένεια. Η έρευνα με δεδομένα NGS απαιτεί την συνεργασία μεταξύ επιστημών διαφορετικών εξειδικεύσεων και επαγγελμάτων αφού τα δεδομένα αυτά είναι ετερογενή και μπορούν να προέρχονται από κλινικά ή φαινοτυπικά ή πειραματικά κ.α. δεδομένα. Το σημαντικό για την εταιρία είναι η σύνδεση αυτών των δεδομένων με όσο γίνεται λιγότερο κόπο. Επειδή το κόστος της NGS όλο και πέφτει, τα δεδομένα που παράγονται από αυτή την τεχνική όλο και αυξάνονται. Σε αυτό το σημείο την λύση δίνει μια υψηλής ευλυγησίας οργανωτική δομή μεγάλων δεδομένων. Ο στόχος τους είναι να εισάγονται χωρίς πολύ κόπο στην πλατφόρμα τους τα πιο προχωρημένα αναλυτικά εργαλεία, οι πιο νέες τεχνικές και οι πιο ολοκληρωμένες βάσεις δεδομένων, στην προκειμένη περίπτωση επιλέχθηκε η πλατφόρμα Apache Hadoop. Αλλά δεν είναι μόνο η πληθώρα των δεδομένων αλληλουχοποίησης εξαιρετικά υψηλής απόδοσης NGS αλλά και το ότι δεν χρειάζεται βελτιστοποίησή τους, η οποία ήταν απαραίτητη για την χρήση σε κλασσικά υπολογιστικά συστήματα υψηλής απόδοσης (High-Performance Computing, HPC). Ακόμα, αρκετά από τα απαραίτητα εργαλεία που χρησιμοποιούν οι ερευνητές σε κλασικά υπολογιστικά συστήματα δεν λειτουργούν στο σύστημα κατανεμημένων αρχείων της

πλατφόρμας Hadoop (Hadoop Distributed File System, HDFS). Το νέο σύστημα παρέχει στους βιοπληροφορικούς επιστήμονες το σύστημα προσπέλασης αρχείων POSIX (MapR Hadoop, μια εμπορική έκδοση που χρησιμοποεί το σύστημα αρχείων MapR-FS) με το οποίο μπορούν να χρησιμοποιούν τα συνήθη εργαλεία τους. Το σύστημα ροής εργασίας παρουσιάζει έτσι καλή απόδοση και ισχύ, ενώ επιτρέπει τον συνδυασμό των πλεονεκτημάτων του κλασικού και του νέου Hadoop συστήματος, αφού επιτρέπει στους ερευνητές να σχεδιάσουν πολύπλοκες ροές εργασίας της διαδικασίας για αυτόν τον σκοπό. Στη συνέχεια με το Apache Spark ενσωματώνουν τα πολύ διαφορετικά σέτ δεδομένων. Η πρωτοπόρα μέθοδός τους για να αντιμετωπίσουν την ετερογένεια αυτή ήταν με την χρήση ενός πελώριου γραφήματος που αναπαραγάγει/παρουσιάζει αυτά τα δεδομένα (με τρισεκατομμύρια κορυφές ήδη), αποθηκεύεται σε HDFS και τροποποιείται με προσαρμοσμένο κώδικα Spark. Αυτό το γνωστικό γράφημα γνώσης χρησιμοποιείται από τους βιοπληροφοριακούς της εταιρίας για να μοντελοποιήσουν τα περίπλοκα βιολογικά δεδομένα και το πως συνδέονται μεταξύ τους, ενώ με το Spark μπορούν να αναλύσουν τα γραφήματα με αξιοπιστία και στη σωστή κλίμακα. Όσο αφορά την αναλυτική επιστήμη, οι ερευνητές έχουν πρόσβαση στα δεδομένα κατευθείαν από μια από τις καθορισμένες διεπαφές προγραμματισμού (API) του Spark ή από κάμποσες τελικού σημείου βάσεις δεδομένων με σχέδια φτιαγμένα ειδικά για τις συγκεκριμένες αναλυτικές τους ανάγκες. Η εργαλειοθήκη τους επιτρέπει ολόκληρα σχέδια με 100 δισεκατομμύρια σειρών να δημιουργηθεί γρήγορα από το γνωστικό γράφημα, τα οποία μετά εισάγονται στην επιθυμητή βάση δεδομένων του ερευνητή. Πλέον αυτή η μέθοδος χρησιμοποιείται και σε άλλους τομείς της εταιρίας, όπως στην πρωτεομική, την μεταγενωμική και την ανάλυση βίντεο. Η ενσωμάτωση της πλατφόρμας μεγάλων δεδομένων με τα γνωστά βιοπληροφορικά εργαλεία μπορεί να επιτευχθεί σε λίγες μέρες αντί για μήνες. Με τον συνδυασμό Spark και μιας ροής εργασιών βασισμένης στο Hadoop και στρώματα ενοποίησης (integration layers), οι ερευνητές της Novartis μπορούν να εκμεταλλευτούν δεκάδες χιλιάδες πειράματα που υπάρχουν σε δημόσιες βάσεις δεδομένων, που τους δίνει ένα σημαντικό πλεονέκτημα σε σχέση με τους ανταγωνιστές τους (400).

Η παραδειγματική γενωμική ανάλυση ενός λαού. Τα τελευταία χρόνια έχουν δημιουργηθεί τεράστιες συλλογές δεδομένων γονιδιακής έκφρασης, κυρίως με την μέθοδο αλληλουχοποίησης νουκλεινικών οξέων. Ένα τρανό παράδειγμα είναι αυτό της Ισλανδίας, μιας χώρας με μειωμένο πληθυσμό. Το 1998 το κοινοβούλιο της Ισλανδίας επέτρεψε στην βιοφαρμακευτική εταιρία "deCODE genetics" να συλλέξει και να αποθηκεύσει τα ιατρικά δεδομένα ολόκληρου του πληθυσμού της χώρας. Η συμβαλλόμενη εταιρία ήδη από το 2003

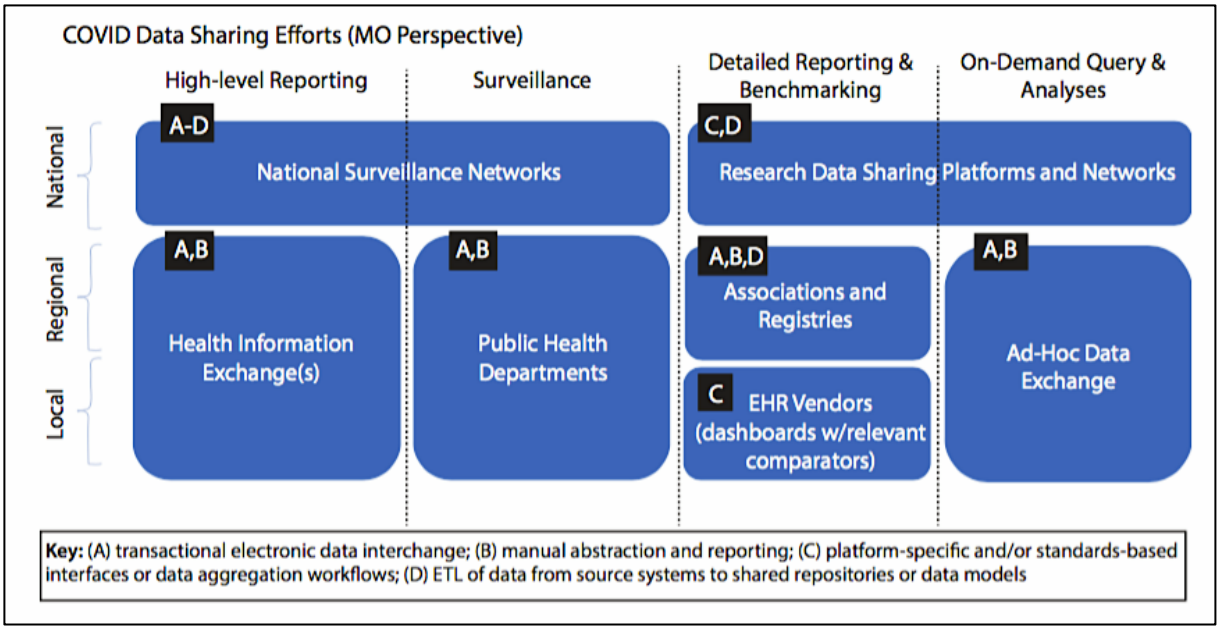
παρείχε διαδικτυακή πρόσβαση στις αλληλουχίες DNA του πληθυσμού στην βάση δεδομένων "Íslendingabók", που σημαίνει το βιβλίο των Ισλανδών. Όλοι οι πολίτες μπορούσαν να ερευνήσουν το γενεαλογικό τους δέντρο και να ανακαλύψουν τυχόν συγγένεια με άλλους πολίτες. Η βάση δεδομένων αριθμούσε το 2020 πάνω από 200.000 χρήστες και 900.000 προσθήκες αλληλουχιών DNA, καλύπτοντας ένα μεγάλο μέρος του πληθυσμού της χώρας. Έχει δημιουργηθεί μια από τις πιο σημαντικές και μεγάλες συλλογές δεδομένων γονιδιακής έκφρασης στον κόσμο. Η εξόρυξη χρήσιμων πληροφοριών από τα μοτίβα και τις αλληλεπιδράσεις μεταξύ τους κάνει δυνατή την στατιστική επεξεργασία τους. Λόγω της δυσκολίας αυτής της εξόρυξης οι εξεταζόμενες αλληλουχίες συγκρίνονται με τις ήδη υπάρχουσες και με την βοήθεια των αλγορίθμων ομαδοποίησης αναγνωρίζονται ομόλογες ομάδες με παρόμοια δεδομένα γονιδιακής έκφρασης, για παράδειγμα γονιδιακές λειτουργίες και κυτταρικές διεργασίες. Η ομολογία γονιδίων είναι πολύ σημαντική και για την ανάπτυξη εμβολίων. Η ομαδοποίηση των δεδομένων γονιδιακής έκφρασης εξαρτάται από το μέτρο ανομοιότητας, π.χ. με την απόσταση Levenshtein.



Εικόνα 40. SWOT Formulation and Frame Work of Research Project (186).

COVID-19. Η πανδημία του κορωνοϊού που ξέσπασε στο τέλος του 2019 στην Wuhan (Κίνα) είχε πρωτοφανή επίδραση στην ανάπτυξη συνεργασιών μεταξύ επιστημόνων, επιστημονικών εργαστηρίων και κρατών, για να βρεθούν σωτήριες λύσεις, δηλαδή καταλληλότητα υπαρχόντων φαρμάκων, εφεύρεση εμβολίων και παραγωγή νέων φαρμάκων, και για να υποστηριχθούν οι υπηρεσίες υγείας τοπικά, δηλαδή στα νοσοκομεία και ιδιωτικά

θεραπευτικά κέντρα της κάθε χώρας. Ο ρόλος των Big data είναι η αποθήκευση, ο μετασχηματισμός και η ενσωμάτωση νέων δεδομένων, που προέρχονται από αρχεία διαφορετικών ή/και νέων τύπων, η κατανόηση της πληροφορίας σε επίπεδο συστήματος και η εφαρμογή μεθόδων για την ανίχνευση των σημαντικών σημάτων ή προτύπων, σε συνθήκες εξαιρετικά μεγάλων ποσοτήτων δεδομένων (Volume), εξαιρετικά μεγάλης ταχύτητας δεδομένων (Velocity) και εξαιρετικά ευρείας ποικιλίας δεδομένων (Variety). Μια σημαντική προσπάθεια συγκέντρωσης των δεδομένων ασθενών που νόσησαν από το COVID είναι αυτή της αμερικανικής NIH με την βάση δεδομένων National Clinical COVID Collaboratory (N3C), που προέρχονται από CTSA χορηγούμενα ιδρύματα (**Εικόνα 41**). Αυτά τα δεδομένα είναι άμεσα προσβάσιμα για στατιστική ανάλυση και υπολογιστικές αναλύσεις (ML), αντί να πρέπει ο ερευνητής να ζητήσει ο ίδιος από κάθε νοσοκομείο τα δεδομένα (401).



Εικόνα 41. Τα πλαίσια ανταλλαγής πληροφοριών/δεδομένων (401).

Μελλοντικά. Στο τέλος αυτού του κεφαλαίου, αλλά και της διπλωματικής εργασίας μου, θα ήθελα να αναφερθώ σε μια πιθανή μελλοντική διερεύνηση της. Πιο συγκεκριμένα, αναφέρομαι στις διαφορές που υπάρχουν στις προαναφερόμενες μεθόδους και εργαλεία στην περίπτωση ανάλυσης μεμονωμένων κυττάρων (single cell, sc). Από το 2009 έχουν πραγματοποιηθεί αρκετές μελέτες σε μεμονωμένα κύτταρα και με την βελτίωση των αλληλουχοποιητών υπάρχει ακόμα πιο σημαντική αύξηση της επιλεκτικότητας και της ευκρίνειας στα πειράματα. Συνεπώς, ενδείκνυται μία μελλοντική ανασκόπηση βασισμένη σε μεμονωμένα κύτταρα.

Βιβλιογραφία

1. Kumar D, Eng C. Genomic medicine: principles and practice: Oxford University Press; 2014.

2. Karp G, Iwasa J, Marshall W. Karp's Cell and Molecular Biology: John Wiley & Sons; 2020.

3. Taneri B, Asilmaz E, Delikurt T, Savas P, Targen S, Esemen Y. Human Genetics and Genomics: A Practical Guide: Wiley; 2020.

4. Watson JD, Berry, A. , & Davies, K. DNA: The Story of the Genetic Revolution. : Knopf.; 2017.

5. Alberts B, Bray D, Hopkin K, Johnson A, Lewis J, Raff M, et al. Essential Cell Biology. 2nd ed: Garland Science; 2003.

6. Mokobi F. RNA polymerase- Definition, Types and Functions. 2020 [updated August 19, 2020. Available from: https://microbenotes.com/rna-polymerase/.

7. Dandekar T, Kunz M. Bioinformatik: Springer.

8. Simon EJ, Reece JB, Dickey J. Campbell essential biology with physiology: Benjamin Cummings; 2010.

9. Felekkis K, Voskarides K. Genomic Elements in Health, Disease and Evolution: Junk DNA: Springer New York; 2015.

10. Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009;136(2):215-33.

11. Djuranovic S, Nahvi A, Green R. A parsimonious model for gene regulation by miRNAs. Science (New York, NY). 2011;331(6017):550-3.

12. Akhtar MM, Micolucci L, Islam MS, Olivieri F, Procopio AD. A practical guide to miRNA target prediction. MicroRNA Target Identification: Springer; 2019. p. 1-13.

13. Cortez MA, Bueso-Ramos C, Ferdin J, Lopez-Berestein G, Sood AK, Calin GA. MicroRNAs in body fluids--the mix of hormones and biomarkers. Nat Rev Clin Oncol. 2011;8(8):467-77.

14. Valadi H, Ekström K, Bossios A, Sjöstrand M, Lee JJ, Lötvall JO. Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. Nature cell biology. 2007;9(6):654-9.

15. Vickers KC, Remaley AT. Lipid-based carriers of microRNAs and intercellular communication. Curr Opin Lipidol. 2012;23(2):91-7.

16. Weber B, Stresemann C, Brueckner B, Lyko F. Methylation of human microRNA genes in normal and neoplastic cells. Cell cycle. 2007;6(9):1001-5.

17. NIH. Mechanism of Epigenetics: The National Institutes of Health; [Available from: http://commonfund.nih.gov/epigenomics/figure.

18. Kovalchuk I, Kovalchuk O. Genome Stability: From Virus to Human Application: Elsevier Science; 2016.

19. Herskind AM, McGue M, Holm NV, Sörensen TI, Harvald B, Vaupel JW. The heritability of human longevity: a population-based study of 2872 Danish twin pairs born 1870–1900. Human genetics. 1996;97(3):319-23.

20. Singhal RP, Mays-Hoopes LL, Eichhorn GL. DNA methylation in aging of mice. Mechanisms of ageing and development. 1987;41(3):199-210.

21. Issa J-PJ, Ahuja N, Toyota M, Bronner MP, Brentnall TA. Accelerated age-related CpG island methylation in ulcerative colitis. Cancer research. 2001;61(9):3573-7.

22. Waki T, Tamura G, Sato M, Motoyama T. Age-related methylation of tumor suppressor and tumor-related genes: an analysis of autopsy samples. Oncogene. 2003;22(26):4128-33.

23. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. The Hallmarks of Aging. Cell. 2013;153(6):1194-217.

24. Oakes CC, Smiraglia DJ, Plass C, Trasler JM, Robaire B. Aging results in hypermethylation of ribosomal DNA in sperm and liver of male rats. Proceedings of the National Academy of Sciences. 2003;100(4):1775-80.

25. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14(10):3156.

26. Horvath S. Erratum to: DNA methylation age of human tissues and cell types. Genome Biol. 2015;16(1):96.

27. Sedivy JM, Banumathy G, Adams PD. Aging by epigenetics—A consequence of chromatin damage? Experimental Cell Research. 2008;314(9):1909-17.

28. Wood JG, Helfand SL. Chromatin structure and transposable elements in organismal aging. Front Genet. 2013;4:274-.

29. Zane L, Sharma V, Misteli T. Common features of chromatin in aging and cancer: cause or coincidence? Trends in Cell Biology. 2014;24(11):686-94.

30. Ehrlich M. DNA hypomethylation in cancer cells. Epigenomics. 2009;1(2):239-59.

31. McGarvey KM, Van Neste L, Cope L, Ohm JE, Herman JG, Van Criekinge W, et al. Defining a chromatin pattern that characterizes DNA-hypermethylated genes in colon cancer cells. Cancer research. 2008;68(14):5753-9.

32. So K, Tamura G, Honda T, Homma N, Waki T, Togawa N, et al. Multiple tumor suppressor genes are increasingly methylated with age in non-neoplastic gastric epithelia. Cancer Science. 2006;97(11):1155-8.

33. Ennis C, Pugh O. Introducing Epigenetics: A Graphic Guide: Icon Books Limited; 2017.

34. Nuber UA. DNA microarrays: Garland Science; 2005.

35. Moody SA. Principles of developmental genetics: Academic Press; 2014.

36. Illumina I. Illumina SNP Genotyping Technologies BeadArray and BeadChip Platform; BeadArrayUseCases.

37. Agilent. G4140-90050Gene Expression TwoColor Agilent; 2015 August.

38. Yin JQ, Zhao RC, Morris KV. Profiling microRNA expression with microarrays. Trends in biotechnology. 2008;26(2):70-6.

39. Li W, Ruan K. MicroRNA detection by microarray. Analytical and bioanalytical chemistry. 2009;394(4):1117-24.

40. Liu C-G, Calin GA, Volinia S, Croce CM. MicroRNA expression profiling using microarrays. Nature protocols. 2008;3(4):563.

41. Castoldi M, Schmidt S, Benes V, Noerholm M, Kulozik AE, Hentze MW, et al. A sensitive array for microRNA expression profiling (miChip) based on locked nucleic acids (LNA). RNA. 2006;12(5):913-20.

42. Kappel A, Keller A. miRNA assays in the clinical laboratory: workflow, detection technologies and automation aspects. Clinical Chemistry and Laboratory Medicine (CCLM). 2017;55(5):636-47.

43. Ying S-Y. Current perspectives in microRNAs (miRNA): Springer; 2008.

44. Wang B, Howel P, Bruheim S, Ju J, Owen LB, Fodstad O, et al. Systematic evaluation of three microRNA profiling platforms: microarray, beads array, and quantitative real-time PCR array. PloS one. 2011;6(2):e17167.

45. Chen L, Heikkinen L, Wang C, Yang Y, Sun H, Wong G. Trends in the development of miRNA bioinformatics tools. Brief Bioinform. 2019;20(5):1836-52.

46. Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, et al. The microRNAs of Caenorhabditis elegans. Genes & development. 2003;17(8):991-1008.

47. Gruber AR, Findeiß S, Washietl S, Hofacker IL, Stadler PF. RNAz 2.0: improved noncoding RNA detection. Biocomputing 2010: World Scientific; 2010. p. 69-79.

48. Xue C, Li F, He T, Liu G-P, Li Y, Zhang X. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. BMC Bioinformatics. 2005;6(1):1-7.

49. Jiang P, Wu H, Wang W, Ma W, Sun X, Lu Z. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic acids research. 2007;35(suppl_2):W339-W44.

50. Tyagi S, Vaz C, Gupta V, Bhatia R, Maheshwari S, Srinivasan A, et al. CID-miRNA: a web server for prediction of novel miRNA precursors in human genome. Biochemical and biophysical research communications. 2008;372(4):831-4.

51. Mhuantong W, Wichadakul D. MicroPC (μPC): A comprehensive resource for predicting and comparing plant microRNAs. BMC Genomics. 2009;10(1):1-8.

52. Gkirtzou K, Tsamardinos I, Tsakalides P, Poirazi P. MatureBayes: a probabilistic algorithm for identifying the mature miRNA within novel precursors. PloS one. 2010;5(8):e11843.

53. Tav C, Tempel S, Poligny L, Tahi F. miRNAFold: a web server for fast miRNA precursor prediction in genomes. Nucleic acids research. 2016;44(W1):W181-W4.

54. Stegmayer G, Yones C, Kamenetzky L, Milone DH. High class-imbalance in pre-miRNA prediction: a novel approach based on deepSOM. IEEE/ACM transactions on computational biology and bioinformatics. 2016;14(6):1316-26.

55. Xiong Z, Li M, Yang F, Ma Y, Sang J, Li R, et al. EWAS Data Hub: a resource of DNA methylation array data and metadata. Nucleic acids research. 2020;48(D1):D890-D5.

56. Li M, Zou D, Li Z, Gao R, Sang J, Zhang Y, et al. EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. Nucleic acids research. 2019;47(D1):D983-D8.

57. Reik W, Walter J. Genomic imprinting: parental influence on the genome. Nature Reviews Genetics. 2001;2(1):21-32.

58. Keshet I, Schlesinger Y, Farkash S, Rand E, Hecht M, Segal E, et al. Evidence for an instructive mechanism of de novo methylation in cancer cells. Nat Genet. 2006;38:149-53.

59. Rajendram R, Ferreira JC, Grafodatskaya D, Choufani S, Chiang T, Pu S, et al. Assessment of methylation level prediction accuracy in methyl-DNA immunoprecipitation and sodium bisulfite based microarray platforms. Epigenetics. 2011;6(4):410-5.

60. Yong W-S, Hsu F-M, Chen P-Y. Profiling genome-wide DNA methylation. Epigenetics & Chromatin. 2016;9(1):26.

61. Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature. 2010;466(7303):253-7.

62. Gatev E. DNA methylation microarray data reduction for co-methylation analysis: University of British Columbia; 2020.

63. Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. Epigenomics. 2016;8(3):389-99.

64. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011;98(4):288-95.

65. Lentini A, Nestor CE. Mapping DNA Methylation in Mammals: The State of the Art. DNA Modifications. 2021:37-50.

66. Morini E, Sangiuolo F, Caporossi D, Novelli G, Amati F. Application of Next Generation Sequencing for personalized medicine for sudden cardiac death. Front Genet. 2015;6(55).

67. Head SR, Ordoukhanian P, Salomon DR. Next Generation Sequencing: Methods and Protocols: Springer New York; 2017.

68. Philippidis A. Illumina Unveils New Sequencing Systems, Roche Clinical Dx Collaboration [updated 14/01/2020. Available from: www-genengnews-com-news-illumina-unveils-new-sequencing-systems-roche-clinical-d.pdf.

69. Wei DQ, Ma Y, Cho WCS, Xu Q, Zhou F. Translational Bioinformatics and Its Application: Springer Netherlands; 2017.

70. Kalapanulak S, Saithong T, Thammarongtham C. Networking Omic Data to Envisage Systems Biological Regulation. Network Biology. 2016:121-41.

71. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011;475(7356):348-52.

72. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Human Molecular Genetics. 2010;19(R2):R227-R40.

73. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, et al. The potential and challenges of nanopore sequencing. Nanoscience and technology: A collection of reviews from Nature Journals. 2010:261-8.

74. Timp W, Mirsaidov UM, Wang D, Comer J, Aksimentiev A, Timp G. Nanopore Sequencing: Electrical Measurements of the Code of Life. IEEE Transactions on Nanotechnology. 2010;9(3):281-94.

75. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012;13(1):1-13.

76. Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genomics. 2012;13(1):1-7.

77. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol. 2012;30(7):693-700.

78. Loomis EW, Eid JS, Peluso P, Yin J, Hickey L, Rank D, et al. Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. Genome Res. 2013;23(1):121-8.

79. Miyamoto M, Motooka D, Gotoh K, Imai T, Yoshitake K, Goto N, et al. Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes. BMC Genomics. 2014;15(1):699.

80. Cherf GM, Lieberman KR, Rashid H, Lam CE, Karplus K, Akeson M. Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. Nat Biotechnol. 2012;30(4):344-8.

81. Biocompare. The Buyer's Guide for Life Sciences. [cited 2021. Available from: https://www.biocompare.com/Molecular-Biology/23967-Whole-Genome-Sequencer-Whole-Genome-Analyzer/.

82. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. Nature Reviews Genetics. 2020;21(10):597-614.

83. Wilson K, Hofmann A, Walker JM, Clokie S. Wilson and Walker's principles and techniques of biochemistry and molecular biology: Cambridge University Press; 2018.

84. Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. Science. 2003;299(5607):682-6.

85. Schadt EE, Banerjee O, Fang G, Feng Z, Wong WH, Zhang X, et al. Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. Genome Res. 2013;23(1):129-41.

86. Chaitanya KV. Genome and Genomics: From Archaea to Eukaryotes: Springer Singapore; 2019.

87. Prasad A, Bhargava H, Gupta A, Shukla N, Rajagopal S, Gupta S, et al. Next Generation Sequencing. In: Singh V, Kumar A, editors. Advances in Bioinformatics. Singapore: Springer Singapore; 2021. p. 277-302.

88. Yohe S, Thyagarajan B. Review of Clinical Next-Generation Sequencing. Archives of Pathology & Laboratory Medicine. 2017;141(11):1544-57.

89. Χ. Νικολάου ΠΧ. Υπολογιστική Βιολογία. Ηράκλειο: Πανεπιστήμιο Κρήτης. ΣΕΑΒ.; 2015.

90. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):621-8.

91. SUN Z, WU H, QIN Z, ZHU Y. Model-Based Methods for Transcript Expression-Level Quantification in RNA-Seq. Advances in Statistical Bioinformatics: Models and Integrative Inference for High-Throughput Data. 2013:105.

92. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.

93. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2009;26(1):139-40.

94. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family database. Nucleic acids research. 2003;31(1):439-41.

95. Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Research. 2017;46(D1):D335-D42.

96. Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Research. 2011;40(1):37-52.

97. Hu Y, Lan W, Miller D. Next-generation sequencing for MicroRNA expression profile. Bioinformatics in MicroRNA Research: Springer; 2017. p. 169-77.

98. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462(7271):315-22.

99. Lister R, Ecker JR. Finding the fifth base: genome-wide sequencing of cytosine methylation. Genome Res. 2009;19(6):959-66.

100. Ruzov A, Gering M. DNA Modifications: Methods and Protocols: Springer; 2020.

101. Gouil Q, Keniry A. Latest techniques to study DNA methylation. Essays in biochemistry. 2019;63(6):639-48.

102. Landau Dan A, Clement K, Ziller Michael J, Boyle P, Fan J, Gu H, et al. Locally Disordered Methylation Forms the Basis of Intratumor Methylome Variation in Chronic Lymphocytic Leukemia. Cancer Cell. 2014;26(6):813-25.

103. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Research. 2005;33(18):5868-77.

104. Smith ZD, Gu H, Bock C, Gnirke A, Meissner A. High-throughput bisulfite sequencing in mammalian genomes. Methods. 2009;48(3):226-32.

105. Olova N, Krueger F, Andrews S, Oxley D, Berrens RV, Branco MR, et al. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. Genome Biol. 2018;19(1):1-19.

106. Li D, Zhang B, Xing X, Wang T. Combining MeDIP-seq and MRE-seq to investigate genome-wide CpG methylation. Methods. 2015;72:29-40.

107. Taiwo O, Wilson GA, Morris T, Seisenberger S, Reik W, Pearce D, et al. Methylome analysis using MeDIP-seq with low DNA concentrations. Nature protocols. 2012;7(4):617-36.

108. Ficz G, Branco MR, Seisenberger S, Santos F, Krueger F, Hore TA, et al. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. Nature. 2011;473(7347):398-402.

109. Wu H, D'Alessio AC, Ito S, Wang Z, Cui K, Zhao K, et al. Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. Genes & development. 2011;25(7):679-84.

110. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-Time DNA Sequencing from Single Polymerase Molecules. Science. 2009;323(5910):133-8.

111. Zhu S, Beaulaurier J, Deikus G, Wu TP, Strahl M, Hao Z, et al. Mapping and characterizing N6-methyladenine in eukaryotic genomes using single-molecule real-time sequencing. Genome Res. 2018;28(7):1067-78.

112. Awada Z, Akika R, Zgheib N. Methods for epigenomic analyses: DNA methylation. Genome Plasticity in Health and Disease: Elsevier; 2020. p. 27-45.

113. Tost J. Current and emerging technologies for the analysis of the genome-wide and locus-specific DNA methylation patterns. DNA Methyltransferases-Role and Function. 2016:343-430.

114. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet. 2001;29(4):365-71.

115. Patterson TA, Lobenhofer EK, Fulmer-Smentek SB, Collins PJ, Chu T-M, Bao W, et al. Performance comparison of one-color and two-color platforms within the Microarray Quality Control (MAQC) project. Nat Biotechnol. 2006;24(9):1140-50.

116. Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, et al. [9] TM4 Microarray Software Suite. Methods in Enzymology. 411: Academic Press; 2006. p. 134-93.

117. Kohane IS, Butte AJ, Kho A. Microarrays for an integrative genomics: MIT press; 2002.

118. Zhang H. Analyzing High-dimensional Gene Expression and DNA Methylation Data with R: CRC Press; 2020.

119. Erik Kristiansson JSID, David Lund Introduction to Bioinformatics. Lecture 7. MVE510. [Lecture Notes.]. Sweden: Chalmers tekniska hogskola; 2019.

120. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences. 2001;98(9):5116-21.

121. Cleveland WS. LOWESS: A program for smoothing scatterplots by robust locally weighted regression. American Statistician. 1981;35(1):54.

122. Cleveland WS. Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association. 1979;74(368):829-36.

123. Yang YH, Dudoit S, Luu P, Speed TP, editors. Normalization for cDNA microarry data. Microarrays: optical technologies and informatics; 2001: International Society for Optics and Photonics.

124. Tseng GC, Oh M-K, Rohlin L, Liao JC, Wong WH. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic acids research. 2001;29(12):2549-57.

125. Zhao Y, Wong L, Goh WWB. How to do quantile normalization correctly for gene expression data analyses. Scientific Reports. 2020;10(1):15534.

126. Bolstad BM, Irizarry RA, Åstrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19(2):185-93.

127. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic acids research. 2003;31(4):e15-e.

128. Furlotte NA, Xu L, Williams RW, Homayouni R, editors. Literature-based Evaluation of Microarray Normalization Procedures. 2011 IEEE International Conference on Bioinformatics and Biomedicine; 2011 12-15 Nov. 2011.

129. Xing E. Computational Genomics Class: Carnegie-Melon University, recitation 7; 2009.

130. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000;403(6769):503-11.

131. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17(6):520-5.

132. Vengatesan K, Mahajan SB, Sanjeevikumar P, Mangrule R, Kala V, Pragadeeswaran. Performance Analysis of Gene Expression Data Using Mann–Whitney U Test. Advances in Systems, Control and Automation: ETAEERE-2016. 2018:701-9.

133. Knudsen S. Guide to Analysis of DNA Microarray Data: Wiley; 2005.

134. STUDENT. THE PROBABLE ERROR OF A MEAN. Biometrika. 1908;6(1):1-25.

135. MacFarland TW, Yates JM. Using R for biostatistics. 2021.

136. Frommlet F, Bogdan M, Ramsey D. Phenotypes and genotypes: Springer; 2016.

137. Satterthwaite FE. An Approximate Distribution of Estimates of Variance Components. Biometrics Bulletin. 1946;2(6):110-4.

138. Welch BL. The Significance of the Difference Between Two Means when the Population Variances are Unequal. Biometrika. 1938;29(3/4):350-62.

139. Mendenhall W, Sincich T. ASecond COURSE IN STATISTICS REGRESSION ANALYSIS. 2012.

140. Kennedy RE, Cui X. Experimental designs and ANOVA for microarray data. Handbook of statistical bioinformatics: Springer; 2011. p. 151-69.

141. Knudsen S. A Biologist's Guide to Analysis of DNA Microarray Data: Wiley; 2011.

142. Barton B, Peat J. Medical Statistics: A Guide to SPSS, Data Analysis and Critical Appraisal: Wiley; 2014.

143. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols. 2009;4(1):44-57.

144. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research. 2009;37(1):1-13.

145. Wang J, Duncan D, Shi Z, Zhang B. WEB-based gene set analysis toolkit (WebGestalt): update 2013. Nucleic acids research. 2013;41(W1):W77-W83.

146. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological). 1995;57(1):289-300.

147. Holm S. A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics. 1979:65-70.

148. Chu G, Narasimhan B, Tibshirani R, Tusher V. SAM:"Significance Analysis of Microarrays" users guide and technical document. 2001.

149. Mezey J. Quantitative Genomics and Genetics class: Cornell University lecture notes; 2020.

150. Lonnstedt I, Speed T. REPLICATED MICROARRAY DATA. Statistica Sinica. 2002;12(1):31-46.

151. Bandyopadhyay S, Mallik S, Mukhopadhyay A. A Survey and Comparative Study of Statistical Tests for Identifying Differential Expression from Microarray Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2014;11(1):95-115.

152. Shah C. A hands-on introduction to data science: Cambridge University Press; 2020.

153. Dunn JC. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics. 1973;3(3):32-57.

154. Fraley C, Raftery AE. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. The Computer Journal. 1998;41(8):578-88.

155. McLachlan G, Krishnan T. The EM Algorithm and Extensions: Wiley; 1996.

156. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data Via the EM Algorithm. Journal of the Royal Statistical Society: Series B (Methodological). 1977;39(1):1-22.

157. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences. 1998;95(25):14863-8.

158. D'Haeseleer P. How does gene expression clustering work? Nat Biotechnol. 2005;23(12):1499-501.

159. Mahalanobis PC, editor On the generalized distance in statistics1936: National Institute of Science of India.

160. Pearson K. Notes on the History of Correlation. Biometrika. 1920;13(1):25-45.

161. Franco M, Vivo J-M. Cluster Analysis of Microarray Data. In: Bolón-Canedo V, Alonso-Betanzos A, editors. Microarray Bioinformatics. New York, NY: Springer New York; 2019. p. 153-83.

162. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis: John Wiley & Sons; 1990.

163. Michener CD, Sokal RR. A QUANTITATIVE APPROACH TO A PROBLEM IN CLASSIFICATION. Evolution. 1957;11(2):130-62.

164. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, et al. Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet. 2000;24(3):227-35.

165. Klipp E, Liebermeister W, Wierling C, Kowald A. Systems Biology: A Textbook: Wiley; 2016.

166. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 1901;2(11):559-72.

167. Yeung KY, Ruzzo WL. Principal component analysis for clustering gene expression data. Bioinformatics. 2001;17(9):763-74.

168. Khan J, Simon R, Bittner M, Chen Y, Leighton SB, Pohida T, et al. Gene Expression Profiling of Alveolar Rhabdomyosarcoma with cDNA Microarrays. Cancer Research. 1998;58(22):5009-13.

169. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature. 2000;406(6795):536-40.

170. Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S. Bioinformatics and Computational Biology Solutions Using R and Bioconductor: Springer New York; 2006.

171. Lloyd S. Least squares quantization in PCM. IEEE Transactions on Information Theory. 1982;28(2):129-37.

172. Hartigan JA, Wong MA. Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society Series C (Applied Statistics). 1979;28(1):100-8.

173. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat Genet. 1999;22(3):281-5.

174. Draghici S. Statistics and data analysis for microarrays using R and bioconductor: CRC Press; 2016.

175. Russell S, Meadows LA, Russell RR. Microarray Technology in Practice: Elsevier Science; 2008.

176. Causton H, Quackenbush J, Brazma A. Microarray Gene Expression Data Analysis: A Beginner's Guide: Wiley; 2003.

177. Reynolds AP, Richards G, de la Iglesia B, Rayward-Smith VJ. Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. Journal of Mathematical Modelling and Algorithms. 2006;5(4):475-504.

178. Kohonen T. Maps, Self-Organizing Berlin: Springer-Verlag; 1995.

179. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, et al. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proceedings of the National Academy of Sciences. 1999;96(6):2907-12.

180. Rhodius VA, Gross CA. Chapter four - Using DNA Microarrays to Assay Part Function. In: Voigt C, editor. Methods in Enzymology. 497: Academic Press; 2011. p. 75-113.

181. Bolón-Canedo V, Alonso-Betanzos A. Microarray Bioinformatics: Springer; 2019.

182. Abu-Jamous B, Fa R, Nandi AK. Integrative Cluster Analysis in Bioinformatics: Wiley; 2015.

183. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explor Newsl. 2009;11(1):10–8.

184. Rui X, Wunsch D. Survey of clustering algorithms. IEEE Transactions on Neural Networks. 2005;16(3):645-78.

185. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). The Annals of Statistics. 2000;28(2):337-407, 71.

186. Basavarajaiah DM, & Murthy, N. B. . Design of experiments and advanced statistical techniques in clinical research. : Springer.; 2020.

187. Ayadi W, Elloumi M. Biclustering of Microarray Data. Algorithms in Computational Molecular Biology2011. p. 651-63.

188. Widmer C, Leiva J, Altun Y, Rätsch G, editors. Leveraging Sequence Classification by Taxonomy-Based Multitask Learning2010; Berlin, Heidelberg: Springer Berlin Heidelberg.

189. Fix E, Hodges JL. Discriminatory analysis. Nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique. 1951;57(3):238-47.

190. Dudoit S, Fridlyand J, Speed TP. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. Journal of the American Statistical Association. 2002;97(457):77-87.

191. Hastie T, Tibshirani R, Friedman J. Overview of Supervised Learning. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY: Springer New York; 2009. p. 9-41.

192. FISHER RA. THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS. Annals of Eugenics. 1936;7(2):179-88.

193. Hastie T, Tibshirani R, Friedman J. Support Vector Machines and Flexible Discriminants. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY: Springer New York; 2009. p. 417-58.

194. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory; Pittsburgh, Pennsylvania, USA: Association for Computing Machinery; 1992. p. 144–52.

195. Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proceedings of the National Academy of Sciences. 2000;97(1):262-7.

196. Han J, Pei J, Kamber M. Data mining: concepts and techniques: Elsevier; 2011.

197. Breiman L, Friedman J, Olshen R, Stone C. Classification and regression trees. Wadsworth Int. Group. 1984;37(15):237-51.

198. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.

199. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences. 2002;99(10):6567-72.

200. Dr. R. Simon B-A, Dr. Richard Simon. BRB-ArrayTools Version 4.5 User's Manual: The EMMES Corporation; 2016.

201. Domingos P, Pazzani M. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. Machine Learning. 1997;29(2):103-30.

202. Cho HS, Kim TS, Wee JW, Jeon SM, Lee CH. cDNA microarray data based classification of cancers using neural networks and genetic algorithms. Nanotech. 2003;1:28-31.

203. Yip W-K, Amin SB, Li C. A survey of classification techniques for microarray data analysis. Handbook of Statistical Bioinformatics: Springer; 2011. p. 193-223.

204. Segall RS, Niu G. Open Source Software for Statistical Analysis of Big Data: Emerging Research and Opportunities. 2020.

205. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;215(3):403-10.

206. Kent WJ. BLAT—the BLAST-like alignment tool. Genome Res. 2002;12(4):656-64.

207. Dayhoff M, Schwartz R, Orcutt B. 22 a model of evolutionary change in proteins. Atlas of protein sequence and structure. 1978;5:345-52.

208. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences. 1992;89(22):10915-9.

209. Lipman DJ, Altschul SF, Kececioglu JD. A tool for multiple sequence alignment. Proceedings of the National Academy of Sciences. 1989;86(12):4412-5.

210. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research. 1994;22(22):4673-80.

211. Poirot O, O'Toole E, Notredame C. Tcoffee@igs: a web server for computing, evaluating and combining multiple sequence alignments. Nucleic Acids Research. 2003;31(13):3503-6.

212. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998;14(9):755-63.

213. Notredame C, Higgins DG. SAGA: Sequence Alignment by Genetic Algorithm. Nucleic Acids Research. 1996;24(8):1515-24.

214. Li H, Ruan J, Durbin R. Maq: Mapping and assembly with qualities. Version 06. 2008;3:508.

215. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25(14):1754-60.

216. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-9.

217. Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. Bioinformatics. 2008;24(5):713-4.

218. Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009;25(15):1966-7.

219. Goff LA, Trapnell C, Kelley D. CummeRbund: visualization and exploration of Cufflinks high-throughput sequencing data. R package version. 2012;2(0).

220. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols. 2012;7(3):562-78.

221. Garbe J. RNA-Seq Tutorial 1: University of Minnesota, Research Informatics Support Systems, lecture, tutorial 1; 2013.

222. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research. 2006;35(suppl_1):D61-D5.

223. Clough E, Barrett T. The Gene Expression Omnibus Database. In: Mathé E, Davis S, editors. Statistical Genomics: Methods and Protocols. New York, NY: Springer New York; 2016. p. 93-110.

224. Hutchins JRA. Genomic Database Searching. In: Keith JM, editor. Bioinformatics: Volume I: Data, Sequence Analysis, and Evolution. New York, NY: Springer New York; 2017. p. 225-69.

225. Blankenberg D, Kuster GV, Coraor N, Ananda G, Lazarus R, Mangan M, et al. Galaxy: a web−based genome analysis tool for experimentalists. Current protocols in molecular biology. 2010;89(1):19.0. 1-.0. 21.

226. Φραγκίσκος Κολίσης ΕΛ, Παναγιώτης Αγιουτάντης. Σεμινάριο εκμάθησης της πλατφόρμας βιοπληροφορικών αναλύσεων Συνθετικής Βιολογίας OMIC-ENGINE. Athens: OMIC-ENGINE; 2019.

227. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 2003;4(9):R60.

228. Panda B, Krishnan NM. BIOINFORMATICS, SYSTEMS BIOLOGY, AND SYSTEMS MEDICINE. Genomic Medicine: Principles and Practice. 2014:83.

229. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;27(6):863-4.

230. Schmieder R, Lim YW, Rohwer F, Edwards R. TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. BMC Bioinformatics. 2010;11(1):341.

231. Cibulskis K, McKenna A, Fennell T, Banks E, DePristo M, Getz G. ContEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics. 2011;27(18):2601-2.

232. Chu C-Y, Bhattacharya S, Zhou Z, Yee M, Lopez A, Lunger VA, et al. Effects of mapping algorithms on gene selection for RNA-Seq analysis: pulmonary response to acute neonatal hyperoxia. PeerJ Preprints. 2015;3:e833v1.

233. Cliften P. Chapter 7 - Base Calling, Read Mapping, and Coverage Analysis. In: Kulkarni S, Pfeifer J, editors. Clinical Genomics. Boston: Academic Press; 2015. p. 91-107.

234. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.

235. Dodt M, Roehr JT, Ahmed R, Dieterich C. FLEXBAR—Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. Biology. 2012;1(3):895-905.

236. Cantu VA, Sadural J, Edwards R. PRINSEQ++, a multi-threaded tool for fast and efficient quality control and preprocessing of sequencing datasets. PeerJ Preprints. 2019;7:e27553v1.

237. Patel RK, Jain M. NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. PLOS ONE. 2012;7(2):e30619.

238. Li Y-L, Weng J-C, Hsiao C-C, Chou M-T, Tseng C-W, Hung J-H. PEAT: an intelligent and efficient paired-end sequencing adapter trimming algorithm. BMC Bioinformatics. 2015;16(1):S2.

239. Akalin A. Computational Genomics with R: CRC Press; 2020.

240. Yu F, Coarfa C. Sequence Alignment, Analysis, and Bioinformatic Pipelines. Next Generation Sequencing: Springer; 2013. p. 59-77.

241. Homer N, Merriman B, Nelson SF. BFAST: An Alignment Tool for Large Scale Genome Resequencing. PLOS ONE. 2009;4(11):e7767.

242. Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M. SHRiMP: Accurate Mapping of Short Color-space Reads. PLOS Computational Biology. 2009;5(5):e1000386.

243. David M, Dzamba M, Lister D, Ilie L, Brudno M. SHRiMP2: Sensitive yet Practical Short Read Mapping. Bioinformatics. 2011;27(7):1011-2.

244. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27(11):1571-2.

245. Guo W, Fiziev P, Yan W, Cokus S, Sun X, Zhang MQ, et al. BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. BMC Genomics. 2013;14(1):774.

246. Coarfa C, Yu F, Miller CA, Chen Z, Harris RA, Milosavljevic A. Pash 3.0: A versatile software package for read mapping and integrative analysis of genomic and epigenomic variation using massively parallel DNA sequencing. BMC Bioinformatics. 2010;11(1):572.

247. Xi Y, Li W. BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics. 2009;10(1):232.

248. Harris EY, Ponts N, Le Roch KG, Lonardi S. BRAT-BW: efficient and accurate mapping of bisulfite-treated reads. Bioinformatics. 2012;28(13):1795-6.

249. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2012;29(1):15-21.

250. 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. 2013;14(4):R36.

251. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907-15.

252. Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ. GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. In: Mathé E, Davis S,

editors. Statistical Genomics: Methods and Protocols. New York, NY: Springer New York; 2016. p. 283-334.

253. Liao Y, Smyth GK, Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Research. 2013;41(10):e108-e.

254. Roy S, LaFramboise WA, Nikiforov YE, Nikiforova MN, Routbort MJ, Pfeifer J, et al. Next-Generation Sequencing Informatics: Challenges and Strategies for Implementation in a Clinical Environment. Archives of Pathology & Laboratory Medicine. 2016;140(9):958-75.

255. Guerrero-Sanchez VM, Maldonado-Alconada AM, Amil-Ruiz F, Verardi A, Jorrín-Novo JV, Rey M-D. Ion Torrent and lllumina, two complementary RNA-seq platforms for constructing the holm oak (Quercus ilex) transcriptome. PLOS ONE. 2019;14(1):e0210356.

256. Lahens NF, Ricciotti E, Smirnova O, Toorens E, Kim EJ, Baruzzo G, et al. A comparison of Illumina and Ion Torrent sequencing platforms in the context of differential gene expression. BMC Genomics. 2017;18(1):602.

257. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. Genome Res. 2012;22(3):549-56.

258. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols. 2013;8(8):1494-512.

259. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, et al. De novo assembly and analysis of RNA-seq data. Nat Methods. 2010;7(11):909-12.

260. Ghosh A, Mehta A. Concept, Development, and Application of Computational Methods for the Analysis and Integration of Omics Data. In: Hakeem KR, Malik A, Vardar-Sukan F, Ozturk M, editors. Plant Bioinformatics: Decoding the Phyta. Cham: Springer International Publishing; 2017. p. 241-66.

261. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics. 2012;28(16):2184-5.

262. Ebbert MTW, Wadsworth ME, Staley LA, Hoyt KL, Pickett B, Miller J, et al. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. BMC Bioinformatics. 2016;17(7):239.

263. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2013;30(7):923-30.

264. Lun ATL, Chen Y, Smyth GK. It's DE-licious: A Recipe for Differential Expression Analyses of RNA-seq Experiments Using Quasi-Likelihood Methods in edgeR. In: Mathé E, Davis S, editors. Statistical Genomics: Methods and Protocols. New York, NY: Springer New York; 2016. p. 391-416.

265. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for Computing and Annotating Genomic Ranges. PLOS Computational Biology. 2013;9(8):e1003118.

266. Delhomme N, Padioleau I, Furlong EE, Steinmetz LM. easyRNASeq: a bioconductor package for processing RNA-Seq data. Bioinformatics. 2012;28(19):2532-3.

267. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417-9.

268. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34(5):525-7.

269. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol. 2014;32(5):462-4.

270. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research. 2015;4.

271. González JR, Cáceres A. Omic Association Studies with R and Bioconductor: CRC Press; 2019.

272. Hansen KD, Irizarry RA, WU Z. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics. 2012;13(2):204-16.

273. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):R25.

274. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010;11(1):94.

275. Javed S. Differential Expression Analysis of RNA-Seq Data and Co-expression Networks. In: Pham TD, Yan H, Ashraf MW, Sjöberg F, editors. Advances in Artificial Intelligence, Computation, and Data Science: For Medicine and Life Science. Cham: Springer International Publishing; 2021. p. 29-76.

276. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28(5):511-5.

277. Van den Berge K HK, Soneson C, Tiberi S, Clement L, Love MI, Patro R, Robinson MD. RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis. 2019.

278. Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics. 2009;26(1):136-8.

279. 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 Research. 2015;43(7):e47-e.

280. Friederike Dündar LS, Paul Zumbo Introduction to differential gene expression analysis using RNA-seq2015-2020.

281. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106.

282. Christopher Burge DG, Ernest Fraenkel. 7.91J Foundations of Computational and Systems Biology. : Massachusetts Institute of Technology: MIT OpenCourseWare, License: Creative Commons BY-NC-SA.; 2014.

283. Jeff Glaubitz QS. Statistics of RNA-seq data analysis Bioinformatics Facility Institute of Biotechnology Cornell University, Workshop Session 22020

284. Stevens JR. Introduction to Next-Generation Sequencing Data and Analysis– STAT 5570: Statistical Bioinformatics Notes 6.3: Utah State University 2014.

285. Pak M, Jeong D, Moon JH, Ann H, Hur B, Lee S, et al. Network Propagation for the Analysis of Multi-omics Data. In: Yoon B-J, Qian X, editors. Recent Advances in Biological Network Analysis: Comparative Network Analysis and Network Module Detection. Cham: Springer International Publishing; 2021. p. 185-217.

286. Moon JH, Lee S, Pak M, Hur B, Kim S. MLDEG: A Machine Learning Approach to Identify Differentially Expressed Genes Using Network Property and Network Propagation. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2021:1-.

287. Langmead B, Hansen KD, Leek JT. Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biol. 2010;11(8):R83.

288. Sanchez R, Yang X, Maher T, Mackenzie SA. Discrimination of DNA Methylation Signal from Background Variation for Clinical Diagnostics. International Journal of Molecular Sciences. 2019;20(21):5343.

289. Yang X, Mackenzie SA. Approaches to Whole-Genome Methylome Analysis in Plants. In: Spillane C, McKeown P, editors. Plant Epigenetics and Epigenomics : Methods and Protocols. New York, NY: Springer US; 2020. p. 15-31.

290. Sang F. Bioinformatics Analysis of DNA Methylation Through Bisulfite Sequencing Data. In: Ruzov A, Gering M, editors. DNA Modifications: Methods and Protocols. New York, NY: Springer US; 2021. p. 441-50.

291. Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, et al. A Reference Methylome Database and Analysis Pipeline to Facilitate Integrative and Comparative Epigenomics. PLOS ONE. 2013;8(12):e81148.

292. Gaspar JM, Hart RP. DMRfinder: efficiently identifying differentially methylated regions from MethylC-seq data. BMC Bioinformatics. 2017;18(1):528.

293. Bhasin JM, Hu B, Ting AH. MethylAction: detecting differentially methylated regions that distinguish biological subtypes. Nucleic Acids Research. 2015;44(1):106-16.

294. Turan N, Ghalwash MF, Katari S, Coutifaris C, Obradovic Z, Sapienza C. DNA methylation differences at growth related genes correlate with birth weight: a molecular signature linked to developmental origins of adult disease? BMC Medical Genomics. 2012;5(1):10.

295. McDermaid A, Monier B, Zhao J, Liu B, Ma Q. Interpretation of differential gene expression results of RNA-seq data: review and integration. Brief Bioinform. 2018;20(6):2044-54.

296. Kaur P, Singh A, Chana I. Computational Techniques and Tools for Omics Data Analysis: State-of-the-Art, Challenges, and Future Directions. Archives of Computational Methods in Engineering. 2021.

297. Xu X, Zhang Y, Zou L, Wang M, Li A, editors. A gene signature for breast cancer prognosis using support vector machine. 2012 5th International Conference on BioMedical Engineering and Informatics; 2012 16-18 Oct. 2012.

298. Moon M, Nakai K. Integrative analysis of gene expression and DNA methylation using unsupervised feature extraction for detecting candidate cancer biomarkers. Journal of Bioinformatics and Computational Biology. 2018;16(02):1850006.

299. Ramroach S, Joshi A, John M. Optimisation of cancer classification by machine learning generates an enriched list of candidate drug targets and biomarkers. Molecular Omics. 2020;16(2):113-25.

300. Szafranski K, Megraw M, Reczko M, Hatzigeorgiou AG, editors. Support Vector Machines for Predicting microRNA Hairpins. BIOCOMP; 2006.

301. Wei Shi YL. Rsubread/Subread Users Guide. Rsubread v2.2.0/Subread v2.0.0 ed. Australia: Olivia Newton-John Cancer Research Institute Walter and Eliza Hall Institute of Medical Research University of Melbourne; 2020 17 April.

302. Hertel J, de Jong D, Marz M, Rose D, Tafer H, Tanzer A, et al. Non-coding RNA annotation of the genome of Trichoplax adhaerens. Nucleic Acids Research. 2009;37(5):1602-15.

303. Teune J-H, Steger G. <small class="sc">NOVO</small>MIR: De Novo Prediction of MicroRNA-Coding Regions in a Single Plant-Genome. Journal of Nucleic Acids. 2010;2010:495904.

304. Hofacker IL, Priwitzer B, Stadler PF. Prediction of locally stable RNA secondary structures for genome-wide surveys. Bioinformatics. 2004;20(2):186-90.

305. Giegerich R, Voß B, Rehmsmeier M. Abstract shapes of RNA. Nucleic Acids Research. 2004;32(16):4843-51.

306. Hertel J, Langenberger D, Stadler PF. Computational Prediction of MicroRNA Genes. In: Gorodkin J, Ruzzo WL, editors. RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods. Totowa, NJ: Humana Press; 2014. p. 437-56.

307. Aguiar R, Ambrosio L, Sepúlveda-Hermosilla G, Maracaja-Coutinho V, Paschoal A. miRQuest: integration of tools on a Web server for microRNA research. Genet Mol Res. 2016;15(1).

308. Fasold M, Langenberger D, Binder H, Stadler PF, Hoffmann S. DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments. Nucleic acids research. 2011;39(suppl_2):W112-W7.

309. LANGENBERGER D, BERMUDEZ-SANTANA CI, STADLER PF, HOFFMANN S. IDENTIFICATION AND CLASSIFICATION OF SMALL RNAS IN TRANSCRIPTOME SEQUENCE DATA. Biocomputing 2010. p. 80-7.

310. Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic acids research. 2012;40(1):37-52.

311. Paicu C, Mohorianu I, Stocks M, Xu P, Coince A, Billmeier M, et al. miRCat2: accurate prediction of plant and animal microRNAs from next-generation sequencing datasets. Bioinformatics. 2017;33(16):2446-54.

312. Hackenberg M, Rodríguez-Ezpeleta N, Aransay AM. miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. Nucleic acids research. 2011;39(suppl_2):W132-W8.

313. Kuang Z, Wang Y, Li L, Yang X. miRDeep-P2: accurate and fast analysis of the microRNA transcriptome in plants. Bioinformatics. 2019;35(14):2521-2.

314. An J, Lai J, Lehman ML, Nelson CC. miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. Nucleic acids research. 2013;41(2):727-37.

315. Jha A, Shankar R. miReader: Discovering novel miRNAs in species without sequenced genome. PloS one. 2013;8(6):e66857.

316. Mapleson D, Moxon S, Dalmay T, Moulton V. MirPlex: A Tool for Identifying miRNAs in High−Throughput sRNA Datasets Without a Genome. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution. 2013;320(1):47-56.

317. Hansen TB, Venø MT, Kjems J, Damgaard CK. miRdentify: high stringency miRNA predictor identifies several novel animal miRNAs. Nucleic acids research. 2014;42(16):e124-e.

318. An J, Lai J, Sajjanhar A, Lehman ML, Nelson CC. miRPlant: an integrated tool for identification of plant miRNA from RNA sequencing data. BMC Bioinformatics. 2014;15(1):1-4.

319. Vitsios DM, Kentepozidou E, Quintais L, Benito-Gutiérrez E, van Dongen S, Davis MP, et al. Mirnovo: genome-free prediction of microRNAs from small RNA sequencing data and single-cells using decision forests. Nucleic acids research. 2017;45(21):e177-e.

320. Betel D, Koppal A, Agius P, Sander C, Leslie C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol. 2010;11(8):1-14.

321. Krüger J, Rehmsmeier M. RNAhybrid: microRNA target prediction easy, fast and flexible. Nucleic acids research. 2006;34(suppl_2):W451-W4.

322. Agarwal V, Bell GW, Nam J-W, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. elife. 2015;4:e05005.

323. Lall S, Grün D, Krek A, Chen K, Wang Y-L, Dewey CN, et al. A genome-wide map of conserved microRNA targets in C. elegans. Current biology. 2006;16(5):460-71.

324. Fahlgren N, Carrington JC. miRNA target prediction in plants. Plant MicroRNAs: Springer; 2010. p. 51-7.

325. Karagkouni D, Paraskevopoulou MD, Chatzopoulos S, Vlachos IS, Tastsoglou S, Kanellos I, et al. DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions. Nucleic acids research. 2018;46(D1):D239-D45.

326. Loher P, Rigoutsos I. Interactive exploration of RNA22 microRNA target predictions. Bioinformatics. 2012;28(24):3322-3.

327. Bhattacharya A, Ziebarth JD, Cui Y. PolymiRTS Database 3.0: linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways. Nucleic acids research. 2014;42(D1):D86-D91.

328. Wang X. Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies. Bioinformatics. 2016;32(9):1316-22.

329. Cho S, Jang I, Jun Y, Yoon S, Ko M, Kwon Y, et al. MiRGator v3. 0: a microRNA portal for deep sequencing, expression profiling and mRNA targeting. Nucleic acids research. 2012;41(D1):D252-D7.

330. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA–target interactions. Nucleic acids research. 2009;37(suppl_1):D105-D10.

331. Sticht C, De La Torre C, Parveen A, Gretz N. miRWalk: An online resource for prediction of microRNA binding sites. PloS one. 2018;13(10):e0206239.

332. Tokar T, Pastrello C, Rossos AE, Abovsky M, Hauschild A-C, Tsay M, et al. mirDIP 4.1—integrative database of human microRNA target predictions. Nucleic acids research. 2018;46(D1):D360-D70.

333. Huang H-Y, Lin Y-C-D, Li J, Huang K-Y, Shrestha S, Hong H-C, et al. miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database. Nucleic acids research. 2020;48(D1):D148-D54.

334. Dai X, Zhuang Z, Zhao PX. psRNATarget: a plant small RNA target analysis server (2017 release). Nucleic acids research. 2018;46(W1):W49-W54.

335. Bottini S, Hamouda-Tekaya N, Tanasa B, Zaragosi L-E, Grandjean V, Repetto E, et al. From benchmarking HITS-CLIP peak detection programs to a new method for identification of miRNA-binding sites from Ago2-CLIP data. Nucleic acids research. 2017;45(9):e71-e.

336. Ahadi A, Sablok G, Hutvagner G. miRTar2GO: a novel rule-based model learning method for cell line specific microRNA target prediction that integrates Ago2 CLIP-Seq and validated microRNA–target interaction data. Nucleic acids research. 2017;45(6):e42-e.

337. Neumann EK. Knowledge-based bioinformatics. Knowledge−Based Bioinformatics: From Analysis to Interpretation. 2010:1-32.

338. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25-9.

339. Consortium TU. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Research. 2013;42(D1):D191-D8.

340. Drabkin HJ, Blake JA, Database ftMGI. Manual Gene Ontology annotation workflow at the Mouse Genome Informatics Database. Database. 2012;2012.

341. Hong EL, Balakrishnan R, Dong Q, Christie KR, Park J, Binkley G, et al. Gene Ontology annotations at SGD: new data sources and annotation methods. Nucleic acids research. 2007;36(suppl_1):D577-D81.

342. Grove C, Cain S, Chen WJ, Davis P, Harris T, Howe KL, et al. Using WormBase: A Genome Biology Resource for Caenorhabditis elegans and Related Nematodes. In: Kollmar M, editor. Eukaryotic Genomic Databases: Methods and Protocols. New York, NY: Springer New York; 2018. p. 399-470.

343. Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, et al. FlyBase: enhancing Drosophila Gene Ontology annotations. Nucleic acids research. 2008;37(suppl_1):D555-D9.

344. Chisholm RL, Gaudet P, Just EM, Pilcher KE, Fey P, Merchant SN, et al. dictyBase, the model organism database for Dictyostelium discoideum. Nucleic acids research. 2006;34(suppl_1):D423-D7.

345. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Research. 2011;40(D1):D1202-D10.

346. Thomas PD. The gene ontology and the meaning of biological function. The gene ontology handbook: Humana Press, New York, NY; 2017. p. 15-24.

347. Lovering RC. How does the scientific community contribute to gene ontology? The Gene Ontology Handbook. 2017:85-93.

348. Hastings J. Primer on ontologies. The gene ontology handbook: Humana Press, New York, NY; 2017. p. 3-13.

349. Tirmizi SH, Aitken S, Moreira DA, Mungall C, Sequeda J, Shah NH, et al. Mapping between the OBO and OWL ontology languages. Journal of Biomedical Semantics. 2011;2(1):S3.

350. Côté RG, Jones P, Apweiler R, Hermjakob H. The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. BMC Bioinformatics. 2006;7(1):97.

351. Xiang Z, Mungall C, Ruttenberg A, He Y, editors. Ontobee: A linked data server and browser for ontology terms. ICBO; 2011.

352. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, et al. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic acids research. 2011;39(suppl_2):W541-W5.

353. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, et al. AmiGO: online access to ontology and annotation data. Bioinformatics. 2008;25(2):288-9.

354. Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. QuickGO: a web-based tool for Gene Ontology searching. Bioinformatics. 2009;25(22):3045-6.

355. Munoz-Torres M, Carbon S. Get GO! retrieving GO data using AmiGO, QuickGO, API, files, and tools. The Gene Ontology Handbook: Humana Press, New York, NY; 2017. p. 149-60.

356. Supek F, Škunca N. Visualizing go annotations. The Gene Ontology Handbook: Humana Press, New York, NY; 2017. p. 207-20.

357. Pesquita C. Semantic similarity in the gene ontology. The gene ontology handbook: Humana Press, New York, NY; 2017. p. 161-73.

358. Guzzi PH, Mina M, Guerra C, Cannataro M. Semantic similarity analysis of protein data: assessment with biological features and issues. Brief Bioinform. 2011;13(5):569-85.

359. Wu X, Pang E, Lin K, Pei Z-M. Improving the Measurement of Semantic Similarity between Gene Ontology Terms and Gene Products: Insights from an Edge- and IC-Based Hybrid Method. PLOS ONE. 2013;8(5):e66745.

360. Teng Z, Guo M, Liu X, Dai Q, Wang C, Xuan P. Measuring gene functional similarity based on group-wise comparison of GO terms. Bioinformatics. 2013;29(11):1424-32.

361. Pesquita C, Pessoa D, Faria D, Couto F. CESSM: collaborative evaluation of semantic similarity measures. JB2009: challenges in bioinformatics. 2009;157:190.

362. Pesquita C, Faria D, Bastos H, Ferreira AEN, Falcão AO, Couto FM. Metrics for GO based protein semantic similarity: a systematic evaluation. BMC Bioinformatics. 2008;9(5):S4.

363. Jain S, Bader GD. An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. BMC Bioinformatics. 2010;11(1):562.

364. Xu Y GM, Shi W, Liu X, Wang C A novel insight into Gene Ontology semantic similarity. . Genomics 101(6); 2013.

365. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic acids research. 2019;47(W1):W199-W205.

366. 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. Proceedings of the National Academy of Sciences. 2005;102(43):15545-50.

367. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research. 2000;28(1):27-30.

368. Li S, Lu L. Computational Systems Biology Approaches for Deciphering Traditional Chinese Medicine. In: Jiang R, Zhang X, Zhang MQ, editors. Basics of Bioinformatics: Lecture Notes of the Graduate Summer School on Bioinformatics of China. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. p. 337-68.

369. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur Ö, Anwar N, et al. Pathway Commons, a web resource for biological pathway data. Nucleic acids research. 2010;39(suppl_1):D685-D90.

370. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Research. 2017;46(D1):D661-D7.

371. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Research. 2019;48(D1):D498-D503.

372. Lambrou GI, Adamaki M, Koultouki E, Moschovi M. Systems Biology Methodologies for the Understanding of Common Oncogenetic Mechanisms in Childhood Leukemic and Rhabdomyosarcoma Cells. Quality Assurance in Healthcare Service Delivery, Nursing and Personalized Medicine: Technologies and Processes: IGI Global; 2012. p. 111-68.

373. Mlecnik B, Scheideler M, Hackl H, Hartler J, Sanchez-Cabo F, Trajanoski Z. PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. Nucleic acids research. 2005;33(suppl_2):W633-W7.

374. Moutselos K, Kanaris I, Chatziioannou A, Maglogiannis I, Kolisis FN. KEGGconverter: a tool for the in-silico modelling of metabolic networks of the KEGG Pathways database. BMC Bioinformatics. 2009;10(1):324.

375. Krämer A, Green J, Pollard J, Jr, Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics. 2013;30(4):523-30.

376. Dweep H, Showe LC, Kossenkov AV. Functional Annotation of MicroRNAs Using Existing Resources. In: Allmer J, Yousef M, editors. miRNomics: MicroRNA Biology and Computational Analysis. New York, NY: Springer US; 2022. p. 57-77.

377. Raman K. An Introduction to Computational Systems Biology: Systems-Level Modelling of Cellular Networks: CRC Press; 2021.

378. Ideker T, Krogan NJ. Differential network biology. Molecular Systems Biology. 2012;8(1):565.

379. Goenawan IH, Bryan K, Lynn DJ. DyNet: visualization and analysis of dynamic molecular interaction networks. Bioinformatics. 2016;32(17):2713-5.

380. de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. Genetics. 2013;193(2):327-45.

381. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research. 2018;47(D1):D607-D13.

382. Franz M, Rodriguez H, Lopes C, Zuberi K, Montojo J, Bader GD, et al. GeneMANIA update 2018. Nucleic Acids Research. 2018;46(W1):W60-W4.

383. Baxevanis AD, Bader GD, Wishart DS. Bioinformatics: John Wiley & Sons; 2020.

384. Jeske L, Placzek S, Schomburg I, Chang A, Schomburg D. BRENDA in 2019: a European ELIXIR core data resource. Nucleic Acids Research. 2018;47(D1):D542-D9.

385. Wittig U, Kania R, Golebiewski M, Rey M, Shi L, Jong L, et al. SABIO-RK—database for biochemical reaction kinetics. Nucleic Acids Research. 2011;40(D1):D790-D6.

386. Funahashi A, Matsuoka Y, Jouraku A, Morohashi M, Kikuchi N, Kitano H. CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks. Proceedings of the IEEE. 2008;96(8):1254-65.

387. Funahashi A, Morohashi M, Kitano H, Tanimura N. CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. Biosilico. 2003;1(5):159-62.

388. van Iersel MP, Villéger AC, Czauderna T, Boyd SE, Bergmann FT, Luna A, et al. Software support for SBGN maps: SBGN-ML and LibSBGN. Bioinformatics. 2012;28(15):2016-21.

389. Nelson KE, Madupu R, Szpakowski S, Goll JB, Krampis K, Methé BA. Next-generation sequencing, metagenomes, and the human microbiome. Nextgeneration Sequencing: Current Technologies and Applications (ed J Xu) Caister Academic Press Norfolk, UK. 2014:141-55.

390. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics. 2003;19(4):524-31.

391. Le Novère N, Bornstein B, Broicher A, Courtot M, Donizelli M, Dharuri H, et al. BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic acids research. 2006;34(suppl_1):D689-D91.

392. Tomita M, Hashimoto K, Takahashi K, Shimizu TS, Matsuzaki Y, Miyoshi F, et al. E-CELL: software environment for whole-cell simulation. Bioinformatics (Oxford, England). 1999;15(1):72-84.

393. Takahashi K, Kaizu K, Hu B, Tomita M. A multi-algorithm, multi-timescale method for cell simulation. Bioinformatics. 2004;20(4):538-46.

394. Janowski SJ, Kaltschmidt B, Kaltschmidt C. Biological network modeling and analysis. Approaches in Integrative Bioinformatics: Springer; 2014. p. 203-44.

395. Funahashi A. Celldesigner Tutorial. Keio University: The Systems Biology Institute; 2007.

396. Stein LD. The case for cloud computing in genome informatics. Genome Biol. 2010;11(5):1-7.

397. Krampis K, Booth T, Chapman B, Tiwari B, Bicak M, Field D, et al. Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. BMC Bioinformatics. 2012;13(1):1-8.

398. Wilkening J, Wilke A, Desai N, Meyer F, editors. Using clouds for metagenomics: a case study. 2009 IEEE International Conference on Cluster Computing and Workshops; 2009: IEEE.

399. Jain S, Saxena A, Hesarur S, Bhadhadhara K, Bharti N, Kasibhatla SM, et al. GenoVault: a cloud based genomics repository. BioData Mining. 2021;14(1):36.

400. Krishnan K. 5 - Pharmacy industry applications and usage
. 2020. In: Building Big Data Applications [Internet]. Academic Press,
. Available from: (http://www.sciencedirect.com/science/article/pii/B9780128157466000053).

401. Payne PR, Embi PJ, Cimino JJ. Clinical research informatics. Biomedical Informatics: Springer; 2021. p. 913-40.