ΓΕΩΠΟΝΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΤΜΗΜΑ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΗΣ ΒΙΟΛΟΓΙΑΣ ΚΑΙ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΒΙΟΛΟΓΙΑ ΣΥΣΤΗΜΑΤΩΝ

Μεταπτυχιακή Διπλωματική Εργασία

Διερεύνηση υπολογιστικών μεθόδων ανάλυσης και ενσωμάτωσης δεδομένων scRNA σε κυπαρικούς άτλαντες

Μαρία Σ. Παξινού

Επιβλέπουσα Καθηγήτρια:
Δήμητρα Μηλιώνη, Αναπληρώτρια Καθηγήτρια

Αθήνα

2025

ΓΕΩΠΟΝΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΗΣ ΒΙΟΛΟΓΙΑΣ ΚΑΙ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΒΙΟΛΟΓΙΑ ΣΥΣΤΗΜΑΤΩΝ

Μεταπτυχιακή Διπλωματική Εργασία

Διερεύνηση υπολογιστικών μεθόδων ανάλυσης και ενσωμάτωσης δεδομένων scRNA σε κυτταρικούς άτλαντες

Exploration of computational methods for the analysis and integration of scRNA data into cellular atlases

Μαρία Σ. Παξινού

Εξεταστική Επιτροπή:
Δήμητρα Μηλιώνη, Αναπληρώτρια Καθηγήτρια ΓΠΑ (επιβλέπουσα)
Κωνσταντίνος Μπουγιούκος, Επίκουρος Καθηγητής Université Paris Cité
Γεράσιμος Δάρρας, Επίκουρος Καθηγητής ΓΠΑ

Διερεύνηση υπολογιστικών μεθόδων ανάλυσης και ενσωμάτωσης δεδομένων scRNA σε κυτταρικούς άτλαντες

ΠΜΣ Βιολογία Συστημάτων
Τμήμα Εφαρμοσμένης Βιολογίας και Βιοτεχνολογίας

Περίληψη

Η παραδοσιακή αλληλούχιση επόμενης γενιάς «next generation sequencing» (NGS), εξετάζει το γονιδίωμα ενός κυτταρικού πληθυσμού, όπως για παράδειγμα μιας κυτταρικής καλλιέργειας, ενός ιστού, ενός οργάνου ή και ενός ολόκληρου οργανισμού. Η αλληλούχιση αυτού του τύπου δίνει ως αποτέλεσμα το «μέσο γονιδίωμα» (average genome) του κυτταρικού πληθυσμού. Στις μέρες μας όμως οι νέες τεχνολογίες έχουν επιτρέψει την αλληλούχιση ενός κυττάρου (single cell sequencing), η οποία μετράει το γονιδίωμα μεμονωμένων κυττάρων από έναν κυτταρικό πληθυσμό. Το ενδιαφέρον σε αυτήν την τεχνική είναι πως αποκαλύπτει διαφορές στον κυτταρικό πληθυσμό και εξελικτικές σχέσεις μεταξύ των κυττάρων, πληροφορίες που μέχρι τώρα με τη μαζική αλληλούχιση δεν ήταν ορατές καθώς χανόταν κρίσιμη πληροφορία για την κυτταρική ετερογένεια. Όμως, η αλληλούχιση μεμονωμένων κυττάρων έρχεται και με έναν μεγάλο όγκο δεδομένων που καλούνται οι ερευνητές να αναλύσουν υπολογιστικά και να ερμηνεύσουν. Η έρευνα αυτή συνοπτικά περιλαμβάνει την προεπεξεργασία και τον καθαρισμό των πρωτογενών δεδομένων, την εφαρμογή σύγχρονων μεθόδων τεχνητής νοημοσύνης για την ανάλυση των δεδομένων και την ερμηνεία των κυτταρικών τύπων, που θα δώσουν έπειτα βιολογική σημασία στην υπολογιστική ανάλυση.

Στην παρούσα εργασία εξετάστηκαν διαφορετικές μέθοδοι υπολογιστικής ανάλυσης δεδομένων που παρήχθησαν από αλληλούχιση μεμονωμένων κυττάρων, με σκοπό την εύρεση και την εφαρμογή των βέλτιστων. Στη διαδικασία αυτή έγινε αντιληπτό πως οι υπολογιστικές μέθοδοι πάντα προσαρμόζονται στα εκάστοτε δεδομένα, χωρίς να υπάρχει πανάκια. Έπειτα, αναλύθηκαν δύο σύνολα δεδομένων σε συνεργασία με την ευρωπαϊκή κοινοπραξία Cost Action, με στόχο την ενσωμάτωσή τους στον πρώτο κυτταρικό άτλαντα μεμονωμένων κυττάρων για τον καρκίνο σε κεφάλι και λαιμό. Ο κυτταρικός άτλαντας δημοσιεύτηκε και είναι πλέον διαθέσιμος στην επιστημονική κοινότητα, ως μια πλατφόρμα για έρευνα που μπορεί να αποτελέσει εφαλτήριο στην μελέτη, τη θεραπεία και την πρόληψη του καρκίνου σε κεφάλι και λαιμό.

Επιστημονική Περιοχή: Βιοπληροφορική

Λέξεις Κλειδιά: Ανάλυση δεδομένων, scRNA seq, κυτταρικός άτλαντας

Exploration of computational methods for the analysis and integration of scRNA data into cellular atlases

MSc Systems Biology
Department of Applied Biology and Biotechnology

Abstract

Traditional next – generation sequencing (NGS) examines the genome of a cell population, such as a cell culture, tissue, organ, or even an entire organism. This type of sequencing results in the 'average genome' of the cell population. Today, however, new technologies enable the sequencing of individual cells (single-cell sequencing), which measures the genome of single cells within a population. This technique reveals differences within the cell population and evolutionary relationships between cells—information previously obscured by bulk sequencing, where important details about cellular heterogeneity were lost. Single-cell sequencing, however, produces a large volume of data, which researchers must analyse computationally and interpret. This research generally involves preprocessing and cleaning raw data, applying modern artificial intelligence techniques for data analysis, and interpreting cell types to give biological meaning to the computational analysis.

In this study, different computational analysis methods were tested on data from single cell sequencing to identify and apply the optimal approaches. Through this process, it became clear that computational methods must always be adapted to the specific dataset, as there is no one-size-fits-all solution. Afterwards, two datasets were analysed in collaboration with the European consortium Cost Action, aiming to incorporate them into the first single-cell atlas for head and neck cancer. The cell atlas has been published and is now available to the scientific community as a research platform that could serve as a foundation for studying, treating, and preventing head and neck cancer.

Scientific area: Bioinformatics

Key Words: Data analysis, scRNA seq, cell atlases

Περιεχόμενα

Ευχαριστίες

Η παρούσα εργασία εκπονήθηκε στα πλαίσια του μεταπτυχιακού προγράμματος «Βιολογία Συστημάτων» του τμήματος Βιοτεχνολογίας του Γεωπονικού Πανεπιστημίου Αθηνών υπό την επίβλεψη της αναπληρώτριας καθηγήτριας Δήμητρας Μηλιώνη.

Ξεκινώντας, οφείλω ένα μεγάλο ευχαριστώ στην κα Μηλιώνη που δέχθηκε να είναι η επιβλέπουσα της εργασίας μου, καθώς και στον κ. Μπουγιούκο που ήταν δίπλα μου σε αυτήν την εργασία ως καθοδηγητής και ως μέντορας. Η συνεργασία μας με βοήθησε να αποκτήσω γνώσεις και εμπειρία, χρήσιμες για την μετέπειτα πορεία μου. Επίσης, τους ευχαριστώ θερμά για όλη τη στήριξη και την εμπιστοσύνη που μου έδειξαν στα πλαίσια αυτής της προσπάθειας.

Στη συνέχεια, θα ήθελα να ευχαριστήσω την κα. Dragana Dudic από την ομάδα του Cost Action που με καθοδήγησε και με εμπιστεύθηκε να αναλάβω κάποια από τα σύνολα δεδομένων sc-RNA seq για τον κυτταρικό άτλαντα καρκίνου σε κεφάλι και λαιμό.

Επιπλέον, ευχαριστώ θερμά τα μέλη της τριμελούς επιτροπής, την Αναπληρώτρια Καθηγήτρια κα. Δήμητρα Μηλιώνη, τον Επίκουρο Καθηγητή κ. Γεράσιμο Δάρρα και τον Επίκουρο Καθηγητή κ. Κωνσταντίνο Μπουγιούκο για την υποστήριξη και την αξιολόγηση της προσπάθειάς μου.

Τέλος, ευχαριστώ από καρδιάς την οικογένειά μου που είναι πάντα αρωγοί στις προσπάθειες μου και τους φίλους μου που με στηρίζουν σε κάθε μου βήμα.

Με την άδειά μου, η παρούσα εργασία ελέγχθηκε από την Εξεταστική Επιτροπή μέσα από λογισμικό ανίχνευσης λογοκλοπής που διαθέτει το ΓΠΑ και διασταυρώθηκε η εγκυρότητα και η πρωτοτυπία της.

1. Εισαγωγή

Όλοι οι ζωντανοί οργανισμοί έχουν γραμμένη την ιστορία τους στο γονιδίωμα τους, δηλαδή το DNA τους. Το DNA που βρίσκεται στον πυρήνα των κυττάρων είναι σχεδόν ίδιο σε όλα τα κύτταρα του οργανισμού. Κι όμως ενώ τα κύτταρα ενός οργανισμού μοιράζονται το ίδιο DNA, διαφέρουν όσον αφορά τη βιολογική τους λειτουργία. Αυτό συμβαίνει καθώς η έκφραση του DNA του κάθε κυττάρου ρυθμίζεται διαφορετικά και συνεπώς δημιουργούνται διαφορές στην έκφραση της γενετικής πληροφορίας. Για παράδειγμα οι διαφοροποιήσεις μεταξύ των κυττάρων του ανοσοποιητικού συστήματος και των κυττάρων της καρδιάς οφείλονται στα γονίδια που ενεργοποιούνται ή απενεργοποιούνται σε αυτά τα κύτταρα. Στην πράξη όταν ένα γονίδιο ενεργοποιείται οδηγεί στη δημιουργία περισσότερων αντιγράφων RNA, το οποίο με τη σειρά του θα οδηγήσει στην παραγωγή πρωτεϊνών. Επομένως γίνεται πολύ ενδιαφέρουσα η μελέτη μεμονωμένων κυττάρων (single cell RNA sequencing) καθώς κάθε κύτταρο πρόκειται να έχει διαφορετική ποσότητα και τύπο παραγόμενων μορίων RNA (Hajji, 2024).

Τα δεδομένα που παράγονται από την μελέτη και την ανάλυση μεμονωμένων κυττάρων έχουν μεγάλο όγκο, χρειάζονται υπολογιστές με μεγάλη υπολογιστική ισχύ και για την ερμηνεία τους χρησιμοποιούνται υπολογιστικές αναλύσεις που περιλαμβάνουν την ενσωμάτωση αλγορίθμων μηχανικής μάθησης. Συνοπτικά τα βήματα της ανάλυσης που θα αναλυθούν ενδελεχώς παρακάτω ξεκινούν με την προεπεξεργασία των δεδομένων και τον καθαρισμό τους από τον θόρυβο που μπορεί να έχει προκύψει από τα διάφορα στάδια του πειράματος. Στη συνέχεια εφαρμόζονται αλγόριθμοι μηχανικής μάθησης για να εξαχθούν τα πρώτα αποτελέσματα της ανάλυσης με τον διαχωρισμό των κυττάρων σε ομάδες και τον χαρακτηρισμό τους όσον αφορά τον κυτταρικό τύπο ή την βιολογική κατάσταση.

1.1 Καρκίνος και single cell RNA sequencing

Ένα από τα πιο φλέγοντα θέματα στην ιατρική κοινότητα αποτελούσε διαχρονικά η ογκολογία. Είναι πλέον γνωστό ότι οι κακοήθεις όγκοι προκαλούνται από γενετικές μεταλλάξεις που προκύπτουν από την επίδραση εσωτερικών αλλά και περιβαλλοντικών παραγόντων. Η ανάπτυξη των όγκων είναι μια πολυεπίπεδη και πολύπλοκη διαδικασία κατά την οποία φυσιολογικά κύτταρα εξελίσσονται σε κακοήθεις όγκους, μέσω μιας σειράς πολλαπλών γονιδιακών μεταλλάξεων και συσσώρευσης σε σωματικά κύτταρα. Οι

μεταλλάξεις αναπαράγονται και συσσωρεύονται με επαναλαμβανόμενες διαδικασίες πολλαπλασιασμού και διαίρεσης, με αποτέλεσμα αλλαγές στους κυτταρικούς φαινοτύπους. Έχει φανεί πως κατά τη διαδικασία ανάπτυξης των όγκων οι μεταλλάξεις αυτές είναι κοινές. Βασικό μοριακό γεγονός για την ογκογένεση αποτελεί η «μετάλλαξη οδηγού» (driver mutation), καθώς επηρεάζει τον βαθμό κακοήθειας και την πρόγνωση των ασθενών. Επιπλέον, καθ' όλη τη διαδικασία πολλαπλασιασμού και διαιρέσεων των κυττάρων, δημιουργούνται βιολογικές ή γενετικές διαφορές εντός των καρκινικών κυττάρων, που έχουν ως αποτέλεσμα τον σχηματισμό της σύνθετης ετερογένειας του όγκου που παρατηρείται σε ασθενείς με καρκίνο. Παράλληλα οι καρκινικοί ιστοί διαφοροποιούνται σε διαφορετικούς κυτταρικούς τύπους και υποσύνολα και κατ' αυτόν τον τρόπο αναπτύσσουν πολλαπλά πλεονεκτήματα αντίστασης και πολλαπλασιασμού ανάλογα με το μικροπεριβάλλον τους. Είμαστε σε θέση να γνωρίζουμε πλέον ότι η ετερογένεια των όγκων είναι η κύρια αιτία της αντίστασης στα φάρμακα (Zhang Y, 2021).

1.1.1 Το Μικροπεριβάλλον του όγκου

Το μικροπεριβάλλον του όγκου διαδραματίζει σημαντικό ρόλο στην ανάπτυξη και την ετερογένεια του όγκου, καθώς αυτό θα καθορίζει τις φαινοτυπικές ιδιότητες και χαρακτηριστικά. Με άλλα λόγια, τα καρκινικά κύτταρα στην επιφάνεια του όγκου μπορούν να καθορίσουν την προαγωγή των όγκων και τη μετάσταση, όπως και τα κύτταρα στο εσωτερικό του είναι ικανά να μεγιστοποιήσουν τον πολλαπλασιασμό, ενισχύοντας τον μεταβολισμό. Παρατηρείται επίσης ότι στα μικροπεριβάλλοντα των διαφορετικών τύπων όγκων, διαφέρουν και η σύνθεση και ο βαθμός διείσδυσης των ανοσοκυττάρων. Στις περιπτώσεις όπου περισσότερα Τ κύτταρα καταφέρνουν και διεισδύουν στον ιστό του όγκου, αυτός παραμένει μικρότερος και παράλληλα οι ασθενείς έχουν καλύτερη πρόγνωση. Ταυτόχρονα, διάφορα άλλα κύτταρα του ανοσοποιητικού όπως τα μακροφάγα (macrophages) και τα ουδετερόφιλα (neutrophils), φαίνεται επίσης να ρυθμίζουν σε μεγάλο βαθμό το μικροπεριβάλλον του όγκου. Αυτό έχει ως άμεση συνέπεια, η ευαισθησία διαφορετικών ατόμων στην ανοσοθεραπεία να παρουσιάζει εκτεταμένη ετερογένεια. Τέλος, οι διαφορετικοί κυτταρικοί τύποι στο μικροπεριβάλλον του όγκου επικοινωνούν μεταξύ τους μέσω κυτταρικής επικοινωνίας, γεγονός που αυξάνει την πολυπλοκότητα της ανάπτυξής του. Επομένως, για την διαμόρφωση αποτελεσματικών στρατηγικών αντικαρκινικής ανοσοθεραπείας είναι πολύ σημαντική η κατανόηση αυτών των μηχανισμών επικοινωνίας (Zhang Y, 2021).

1.1.2 Η ετερογένεια του όγκου

Η ετερογένεια του όγκου παίζει επίσης πολύ σημαντικό ρόλο στην εξέλιξη του καρκίνου και είναι πολύ σημαντικό να κατανοήσουμε σε βάθος τα πρότυπα γονιδιακής έκφρασης των μεμονωμένων κυττάρων. Για αυτόν τον σκοπό έχουν αναπτυχθεί οι μέθοδοι αλληλούχισης επόμενης γενιάς (NGS), οι οποίες μπορούν να χρησιμοποιηθούν για την αξιολόγηση της ετερογένειας του όγκου, για την παρακολούθηση των αλλαγών και την αξιολόγηση της εξέλιξης των καρκινικών κυττάρων κατά τη διάρκεια της θεραπείας. Συνεπώς οι έρευνες στοχεύουν σε μια σαφέστερη κατανόηση των μοριακών μηχανισμών που προωθούν την εμφάνιση του όγκου και αποκαλύπτουν τις σωματικές μεταλλάξεις κατά τη διάρκεια της εξέλιξής του. Μεγάλο πλεονέκτημα αυτών των τεχνικών αυτών, και συγκεκριμένα της αλληλούχισης μεμονωμένων κυττάρων, είναι πως μπορεί να προσδιορίσει τις βασικές γονιδιακές μεταλλάξεις, αλλά και τη δυναμική αλλαγή της ετερογένειας του όγκου με την πάροδο του χρόνου. Επιπλέον, επιτρέπεται η παρακολούθηση σπάνιων κυτταρικών μεταλλάξεων και η ανίχνευση της μεταγραφικής δραστηριότητας του ανοσοποιητικού συστήματος. Συνοπτικά, η αλληλούχιση μεμονωμένων κυττάρων είναι ένα σημαντικό εργαλείο στην έρευνα των όγκων, καθώς η ευαισθησία και η ακρίβειά της αυξάνονται, παρέχοντας έτσι νέες ευκαιρίες και στρατηγικές προσεγγίσεις για την κλινική θεραπεία του καρκίνου (Zhang Y, 2021).

1.2 Τεχνολογίες αλληλούχισης επόμενης γενιάς (NGS)

Οι τεχνολογίες αλληλούχισης επόμενης γενιάς (NGS) αποτελούν ισχυρά εργαλεία για την μελέτη γονιδιωματικών χαρακτηριστικών και παρέχουν βασικές γνώσεις σε διάφορους ερευνητικούς και κλινικούς τομείς. Η αλληλούχιση του RNA (RNA seq) είναι η πιο διαδεδομένη τεχνική για την αποκρυπτογράφηση του πολύπλοκου μεταγραφικού τοπίου. Συνεπώς μέσα στα χρόνια έχουν αναπτυχθεί διάφορα πρωτόκολλα RNA seq με στόχο τη χαρτογράφηση της μεταγραφικής έκφρασης, με το κάθε πρωτόκολλο να εμφανίζει τα δικά του πλεονεκτήματα και μειονεκτήματα.

1.2.1 Μαζική αλληλούχιση (Bulk sequencing)

Στα χρόνια που πέρασαν οι ερευνητές σε όλον τον κόσμο έχουν χρησιμοποιήσει την λεγόμενη μαζική αλληλούχιση (bulk sequencing) για να αλληλουχίσουν RNA που έχουν εξάγει από έναν πληθυσμό κυττάρων με σκοπό τη μελέτη της γονιδιακής έκφρασης σε διάφορους ιστούς. Στην ουσία η μαζική αλληλούχιση βασίζεται στη μέση γονιδιακή έκφραση των κυττάρων για την ανάδειξη της παρουσίας και της αφθονίας του RNA στο

δείγμα. Όμως τα μεταγραφικά προφίλ των όγκων είναι εξαιρετικά ετερογενή και μεταξύ των διαφορετικών όγκων, αλλά και ανάμεσα στο μικροπεριβάλλον του ίδιου του όγκου ως αποτέλεσμα της διείσδυσης του στρώματος και άλλων κυτταρικών τύπων σε αυτόν. Συνεπώς είναι πολύ πιθανό τα πραγματικά σήματα που οδηγούν στην καρκινογένεση να αποκρύπτονται κατά τη μελέτη του μέσου προφίλ της γονιδιακής έκφρασης με το μαζικό RNA seq. Παράλληλα είναι δυνατό να κρυφτούν και μηχανισμοί θεραπευτικής αντίστασης από έναν σπάνιο κυτταρικό πληθυσμό ή τύπο κυττάρων στον όγκο. Αυτό το πρόβλημα ήταν καταλυτικής σημασίας για τη γέννηση της τεχνολογίας αλληλούχισης μεμονωμένων κυττάρων (single cell sequencing) (Jana-Charlotte Hegenbarth, 2022).

1.2.2 Αλληλούχιση ενός κυττάρου (Single Cell Sequencing)

Η τεχνολογία αλληλούχισης μεμονωμένων κυττάρων πήρε μεγάλη δημοσιότητα το 2013 όταν στο περιοδικό Nature Methods έγινε αναφορά σε αυτό ως την πολυαναμενόμενη τεχνολογία της χρονιάς. Λίγα χρόνια αργότερα το 2019 το ίδιο περιοδικό ανάδειξε την αλληλούχιση ενός κυττάρου ως την τεχνολογία της χρονιάς, χαρακτηρίζοντάς τη καθοριστική στον προσδιορισμό των κυτταρικών τύπων και λειτουργιών. Στα επόμενα χρόνια και μέχρι σήμερα η επίδρασή της στη γονιδιωματική έρευνα είναι αδιαμφισβήτητη, καθώς μπορούν να αποτυπώσουν την ατομική πολυπλοκότητα του κάθε κυττάρου και την ετερογένεια των ιστών. Σήμερα αυτή η τεχνολογία έχει επιτρέψει τη δημιουργία κυτταρικών ατλάντων, την ανάλυση εκατοντάδων χιλιάδων, ως και εκατομμυρίων κυττάρων παράλληλα, την ενσωμάτωση της χρωματίνης, αλλά και πολυμορφική ανάλυση (Jana-Charlotte Hegenbarth, 2022).

Η τεχνολογία αυτή μπορεί πλέον να εφαρμοστεί για τη μέτρηση του γονιδιώματος (scDNA-seq), του DNA μεθυλώματος ή του μεταγραφώματος (scRNA-seq) κάθε κυττάρου ενός πληθυσμού. Στην πράξη, μπορεί να χρησιμοποιηθεί για τον εντοπισμό νέων μεταλλάξεων σε καρκινικά κύτταρα, για τη διερεύνηση μεταβολών του επιγονιδιώματος που συμβαίνουν κατά την εμβρυϊκή ανάπτυξη και για την αξιολόγηση του τρόπου με τον οποίο ένας φαινομενικά ομοιογενής πληθυσμός κυττάρων εκφράζει συγκεκριμένα γονίδια (Vaga, 2022).

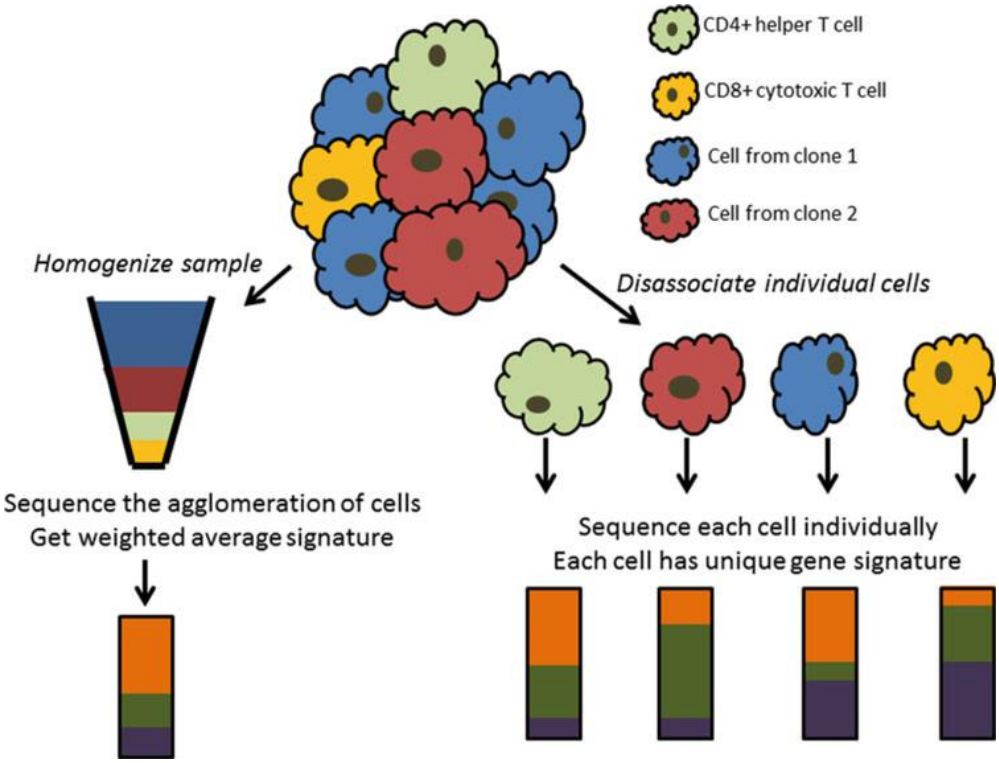
Στην παρούσα εργασία θα εξεταστεί το scRNA-seq. Μελετώντας τις μεθόδους υλοποίησης ανά τα χρόνια έχουν αναφερθεί πολλές «ειδικές» τεχνικές, με τις περισσότερες όμως να έχουν ακολουθήσει μια κοινή γενική μεθοδολογική διαδικασία. Το πρώτο και πιο σημαντικό βήμα για τη διεξαγωγή της scRNA-seq είναι η αποτελεσματική

απομόνωση ζωντανών μεμονωμένων κυττάρων από τον ιστό ενδιαφέροντος. Στη συνέχεια τα απομονωμένα μεμονωμένα κύτταρα λύονται για να επιτραπεί η περισυλλογή όσο το δυνατόν περισσότερων μορίων RNA. Προκειμένου να αναλυθούν τα ειδικά πολυαδενυλιωμένα μόρια mRNA και να αποφευχθεί η συλλογή ριβοσωμικών RNA, χρησιμοποιούνται συνήθως poly[T]-primers. Έπειτα το poly[T]- primed mRNA μετατρέπεται σε συμπληρωματικό DNA (cDNA) με μια αντίστροφη μεταγραφάση. Ανάλογα με το πρωτόκολλο scRNA-seq, στους εκκινητές της αντίστροφης μεταγραφής προστίθενται και άλλες νουκλεοτιδικές αλληλουχίες, όπως αλληλουχίες προσαρμογείς για την ανίχνευση σε πλατφόρμες NGS, μοναδικοί αναγνωριστές (Unique Molecular Identifiers, UMIs) για την αδιαμφισβήτητη σήμανση ενός μορίου mRNA, καθώς και αλληλουχίες για την διατήρηση πληροφοριών σχετικά με την κυτταρική προέλευση. Οι μικροσκοπικές ποσότητες cDNA ενισχύονται στη συνέχεια είτε με PCR είτε σε ορισμένες περιπτώσεις με in vitro μεταγραφή ακολουθούμενη από έναν ακόμα γύρο αντίστροφης μεταγραφής (ορισμένα πρωτόκολλα επιλέγουν τη σήμανση με barcodes σε αυτό το στάδιο για να διατηρηθούν οι πληροφορίες σχετικά με την κυτταρική προέλευση). Στη συνέχεια το ενισχυμένο και επισημασμένο cDNA από κάθε κύτταρο συγκεντρώνεται και αλληλουχίζεται με NGS, χρησιμοποιώντας τεχνικές προετοιμασίας βιβλιοθήκης, πλατφόρμες αλληλούχισης και εργαλεία γονιδιωματικής ευθυγράμμισης. Τέλος η ανάλυση και η ερμηνεία των δεδομένων αποτελούν από μόνες τους ένα ποικιλόμορφο και ταχέως αναπτυσσόμενο πεδίο (Haque, 2017) που θα συζητηθεί παρακάτω.

1.2.3 Bulk sequencing vs Single cell sequencing

Συνοπτικά συγκρίνοντας τις δύο τεχνολογίες που αναλύθηκαν παραπάνω η μαζική αλληλούχιση μετράει την μέση έκφραση ανάμεσα σε έναν κυτταρικό πληθυσμό, σε αντίθεση με την αλληλούχιση ενός κυττάρου που προσφέρει πιο λεπτομερείς πληροφορίες στα μεμονωμένα κύτταρα. Αυτό έχει ως συνέπεια η αλληλούχιση ενός κυττάρου να έχει καλύτερη ανάλυση και να φανερώνει την κυτταρική ετερογένεια και την έκφραση των υποπληθυσμών, ενώ στη μαζική αλληλούχιση αυτό συχνά καλύπτεται. Ωστόσο η μαζική αλληλούχιση είναι μια διαδικασία που δεν χρειάζεται πολλή εργαστηριακή δουλειά συγκριτικά με την αλληλούχιση ενός κυττάρου. Επιπλέον υπολογιστικά η ανάλυση των δεδομένων που έχουν προκύψει από τη μαζική αλληλούχιση είναι γρηγορότερη και λιγότερο πολύπλοκη. Τέλος η αλληλούχιση ενός κυττάρου είναι μια ακριβή διαδικασία συγκριτικά με τη μαζική αλληλούχιση, για αυτό και

θα πρέπει να επιλέγεται προσεκτικά από τον ερευνητή ποια από τις τεχνικές θα χρειαστεί για τη μελέτη του (BioLizard, 2021).



Εικόνα 1 Bulk sequencing vs single cell sequencing. Η εικόνα ανήκει στο άρθρο "Statistical and Bioinformatics Analysis of Data from Bulk and Single-Cell RNA Sequencing Experiments" των Xiaoqing Yu et al.

1.3 Η συνεισφορά της τεχνολογίας single cell analysis

Η σημαντικότητα της τεχνολογίας αλληλούχισης μεμονωμένων κυττάρων όπως αναφέρθηκε και παραπάνω έγκειται στην ανακάλυψη της πλούσιας κυτταρικής ετερογένειας, καθώς προσφέρει καθοριστικά μεγαλύτερη ανάλυση, μελετώντας ένα μεμονωμένο κύτταρο και όχι έναν κυτταρικό πληθυσμό μαζικά. Επομένως είναι δυνατό πλέον να εντοπιστούν και να χαρακτηριστούν σπάνιοι κυτταρικοί τύποι ή υποτύποι. Σήμερα, εφαρμογές αυτής της τεχνολογίας έχουν συνεισφέρει σε πολλούς τομείς της βιολογίας και της βιοϊατρικής έρευνας. Στην αναπτυξιακή βιολογία διευκολύνεται η μελέτη των δυναμικών αλλαγών στη γονιδιακή έκφραση και παρέχονται πληροφορίες για τις αναπτυξιακές διεργασίες, καθώς φαίνεται πως η μεταβλητότητα της έκφρασης μπορεί να καθορίσει τις επιλογές της κυτταρικής μοίρας στην πρώιμη ανάπτυξη (Jonathan A Griffiths, 2018). Επίσης, καταγράφει τις μεταβάσεις μεταξύ διαφορετικών κυτταρικών καταστάσεων, όπως η διαφοροποίηση, η ενεργοποίηση και η απόκριση σε

περιβαλλοντικά ερεθίσματα. Επιπλέον, συμβάλει στην ανάπτυξη της εξατομικευμένης ιατρικής, αποκαλύπτοντας μεμονωμένες κυτταρικές αποκρίσεις και παραλλαγές βοηθώντας στην εύρεση στοχευμένων θεραπειών. Με την ίδια λογική προσφέρει βαθύτερη κατανόηση των μηχανισμών των ασθενειών, ειδικά σε πολύπλοκες ασθένειες όπως είναι ο καρκίνος, οι νευροεκφυλιστικές διαταραχές αλλά και τα αυτοάνοσα νοσήματα. Πολύ σημαντική είναι η συνεισφορά αυτής της τεχνολογίας σε μολυσματικές ασθένειες όπως η Covid-19 μιας και χρησιμοποιήθηκε σε πολλούς ασθενείς για την κατανόηση του τοπίου της απόκρισης των ανοσοποιητικών κυττάρων (Dragomirka Jovic, 2022). Η αλληλούχιση ενός κυττάρου προσφέρει ακόμα πληροφορίες που διευκολύνουν την ανακάλυψη νέων βιοδεικτών βοηθώντας στην έγκαιρη διάγνωση και πρόγνωση των ασθενειών. Όσον αφορά τον τομέα της ανάπτυξης φαρμάκων μπορεί να εντοπίσει φαρμακευτικούς στόχους και παράλληλα να βοηθήσει στην αξιολόγηση της φαρμακευτικής ανταπόκρισης (Van de Sande, 2023).

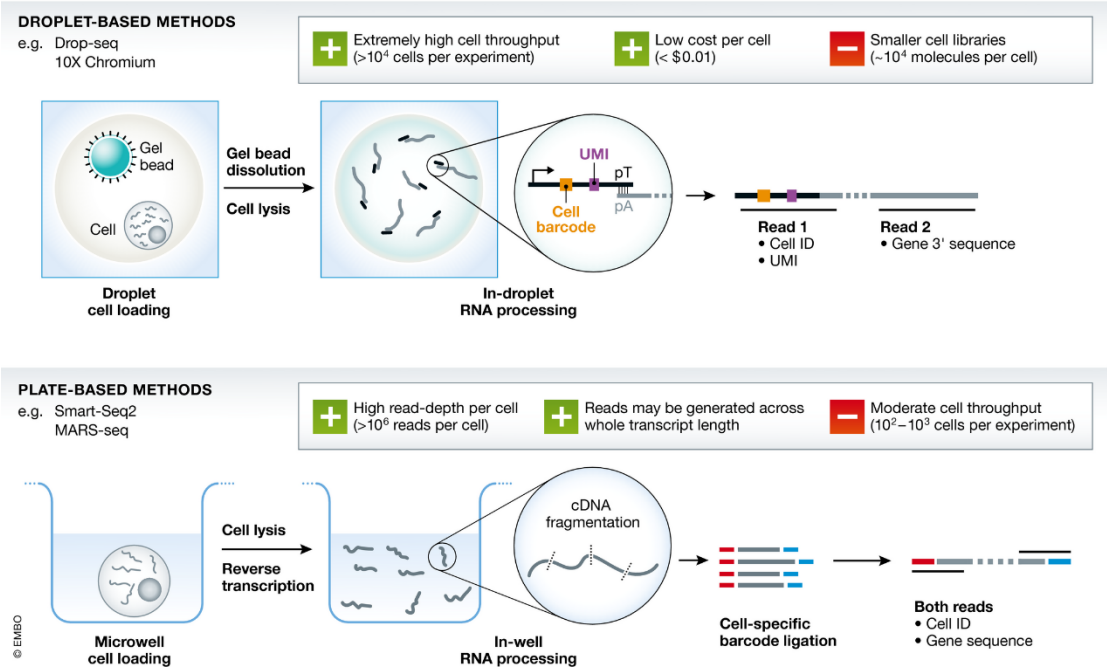
1.4 Droplet – based or Plate – based.

Για τη δημιουργία των δεδομένων single cell RNA sequencing υπάρχουν δύο βασικές μέθοδοι που χρησιμοποιούνται. Οι μέθοδοι που βασίζονται σε σταγονίδια (droplet based) και οι μέθοδοι που βασίζονται σε πλάκες (plate based) (Εικόνα 2).

Στις droplet based μεθόδους χρησιμοποιούνται μικρορευστομηχανές για τη συλλογή των μεμονωμένων κυττάρων σε μορφή σταγονιδίων μεγέθους νανολίτρων. Κάθε ένα από αυτά τα κύτταρα περιέχει αντιδραστήρια και μοναδικές ετικέτες αναγνώρισης (UMI). Μέσα σε αυτά τα σταγονίδια λαμβάνουν χώρα οι διαδικασίες της αντίστροφης μεταγραφής και η σήμανση των μεταγράφων. Στη συνέχεια το εναιώρημα των σταγονιδίων διασπάται για τη συγκέντρωση των κυτταρικών βιβλιοθηκών πριν από την αλληλούχιση. Στα πλεονεκτήματά της αυτή η μέθοδος προσφέρει υψηλή απόδοση καθώς μπορεί να επεξεργαστεί χιλιάδες έως και εκατομμύρια κύτταρα παράλληλα. Αυτό το γεγονός έχει ως άμεση συνέπεια πως είναι και η πιο οικονομικά αποδοτική οδός για τις μελέτες μεγάλης κλίμακας. Ωστόσο φαίνεται πως οι μέθοδοι σταγονιδίων αποτυπώνουν γενετικές πληροφορίες αποκλειστικά από το 3’ ή το 5’ άκρο του κάθε μετάγραφου σε αντίθεση με τις μεθόδου πλάκας που δεν έχουν αυτόν τον περιορισμό. Επιπλέον στα αρνητικά προστίθεται το ότι είναι πιθανό πολλά διαφορετικά κύτταρα να επισημανθούν με το ίδιο barcode, μιας και αυτή η μέθοδος δεν επιτρέπει αρκετή εποπτεία από τον ερευνητή (Jonathan A Griffiths, 2018).

Οι plate based μέθοδοι χρησιμοποιήθηκαν ευρέως στα αρχικά στάδια του single cell RNA sequencing. Εδώ, η προετοιμασία της βιβλιοθήκης πραγματοποιείται χειροκίνητα σε κύτταρα που πρώτα έχουν ταξινομηθεί και λυθεί σε μεμονωμένα τρυβλία μιας πλάκας μικροκυττάρων. Παράλληλα τα barcodes εισάγονται κατά την προετοιμασία της βιβλιοθήκης για να σημάνουν μοναδικά τα μετάγραφα του κάθε κυττάρου. Πλέον κάποιες από αυτές τις διαδικασίες έχουν αυτοματοποιηθεί καθώς έχουν δημιουργηθεί ρομποτικά και μικρορευστομηχανικά συστήματα που διευκολύνουν τους ερευνητές. Βασικό πλεονέκτημα των μεθόδων που βασίζονται σε πλάκες αποτελεί το ότι παρέχουν βιβλιοθήκες υψηλότερης ποιότητας, με το κόστος βέβαια της χαμηλής κυτταρικής απόδοσης μιας και μπορούν να επεξεργαστούν εκατοντάδες ή και χιλιάδες κύτταρα. Επιπλέον οι plate based μέθοδοι είναι περισσότερο αποτελεσματικές στον εντοπισμό σπάνιων τύπων κυττάρων. Τέλος οι μέθοδοι αυτές επιτρέπουν τη συλλογή μεταγραφικών πληροφοριών που σχετίζονται με παραλλαγές που συμβαίνουν κατά τη διάρκεια της ωρίμανσης του pre – mRNA (splice variants) και πληροφοριών σχετικών με τα αλληλόμορφα (Jonathan A Griffiths, 2018).

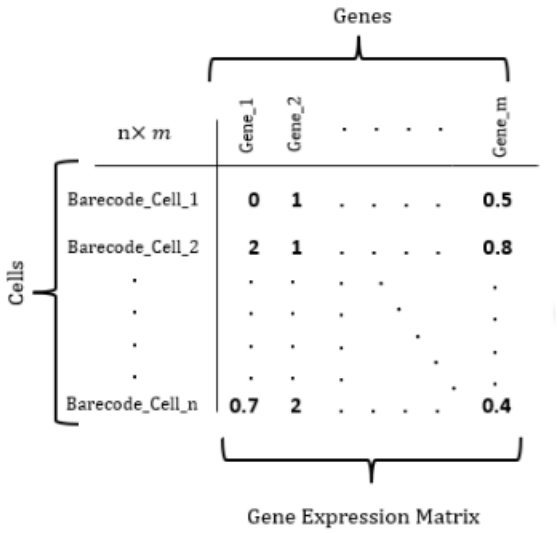
Συνοπτικά οι μέθοδοι που βασίζονται σε σταγονίδια προσφέρουν υψηλή απόδοση κυττάρων, ενώ οι μέθοδοι που βασίζονται σε πλάκες παρέχουν υψηλότερη ανάλυση σε κάθε μεμονωμένο κύτταρο. Συνεπώς καθώς και οι δύο μέθοδοι έχουν τα θετικά και τα αρνητικά τους, οι ερευνητές θα πρέπει να επιλέγουν τη μέθοδο που θα χρησιμοποιήσουν ανάλογα με τις συνθήκες του πειράματός τους.



Εικόνα 2 Η δημιουργία single cell RNA δεδομένων με τις droplet based και τις plate based μεθόδους. Η εικόνα ανήκει στο άρθρο «Using single cell genomics to understand developmental processes and cellular fate decisions» των Jonathan A Griffiths et al, 2018.

1.5 Δομές δεδομένων (Data Infrastructure)

Τα δεδομένα που έχουν προκύψει από single cell RNA sequencing για να είναι δυνατό να αναλυθούν υπολογιστικά πρέπει να αποκτήσουν μια δομή που να επιτρέπει στον υπολογιστή να τα αναλύσει με τη λιγότερη δυνατή πολυπλοκότητα και παράλληλα να τα καθιστά κατανοητά από τον χρήστη. Επομένως η δομή των δεδομένων παίρνει τη μορφή ενός πίνακα, όπου κάθε γραμμή αντιπροσωπεύει ένα κύτταρο το οποίο έχει επισημανθεί με ένα μοναδικό barcode και κάθε στήλη αντιστοιχεί σε ένα συγκεκριμένο γονίδιο. Ο πίνακας γεμίζει με αριθμητικές τιμές, οι οποίες αντιπροσωπεύουν τα επίπεδα έκφρασης του κάθε γονιδίου στο κάθε κύτταρο (Εικόνα 3). Με άλλα λόγια, οι τιμές του πίνακα φανερώνουν την ποσότητα του RNA που παράγεται από κάθε γονίδιο σε ένα συγκεκριμένο κύτταρο, δίνοντας έτσι πληροφορίες για τη δραστηριότητα των γονιδίων ανάμεσα στα διαφορετικά κύτταρα (Hajji, 2024).

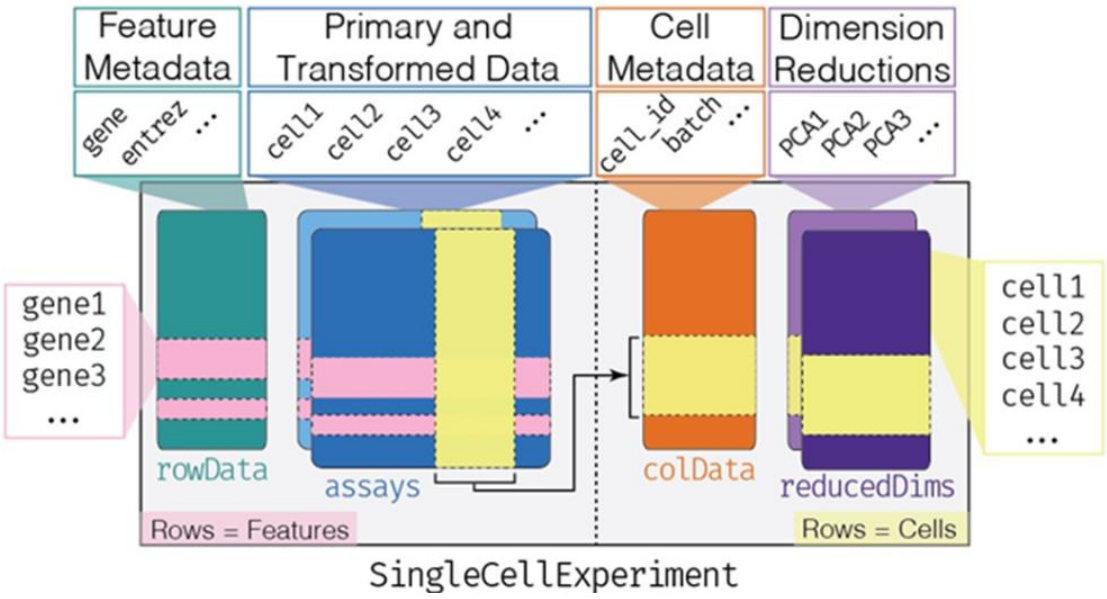


Εικόνα 3 Σχηματική αναπαράσταση του πίνακα έκφρασης των δεδομένων που έχουν προκύψει από single cell RNA sequencing. Η εικόνα ανήκει στην ιστοσελίδα "The Gradient", από το άρθρο της Fatima Zahra El Hajji, με τίτλο "Deep learning for single-cell sequencing: a microscope to see the diversity of cells"

Ένα από τα κύρια εργαλεία για την ανάλυση βιολογικών δεδομένων είναι το Bioconductor, το οποίο προσφέρει ελεύθερο λογισμικό στους ερευνητές ώστε να αναλύουν γρήγορα και εύκολα τα δεδομένα τους (Bioconductor, n.d.).
Το βασικό πλεονέκτημα αυτού του εργαλείου είναι πως έχει δομές που επιτρέπουν την ανάλυση σύνθετων και ιδιαίτερα αλληλεξαρτώμενων συνόλων δεδομένων. Τέτοιου είδους δομές μοιάζουν με τυποποιημένα «κοντέινερ» δεδομένων που επιτρέπουν την ανάπτυξη και την διαλειτουργικότητα διαφορετικών πακέτων. Πιο συγκεκριμένα το Bioconductor διατηρεί μια αντικειμενοστραφή δομή που ονομάζεται S4 και επιτρέπει την ενθυλάκωση πολλαπλών αντικειμένων μέσα σε ένα μόνο αντικείμενο. Αυτό είναι χρήσιμο σε μια βιολογική ανάλυση καθώς προσφέρει συνδέσεις μεταξύ των δεδομένων και διατηρεί μεταδεδομένα που θα χρειαστούν στα επερχόμενα βήματα της ανάλυσης (Amezquita, 2019).

Όσον αφορά τα δεδομένα από τα single cell RNA πειράματα, το Bioconductor διαθέτει την κλάση SingleCellExperiment η οποία αποτελεί ένα ισχυρό αποθετήριο των πρωτογενών δεδομένων αλλά και των μεταδεδομένων. Αυτά τα αποθετήρια οργανώνονται σε ενότητες. Τα πρωτογενή δεδομένα αποθηκεύονται στην ενότητα assays ως ένας ή περισσότεροι πίνακες έκφρασης. Οι ενότητες rawData και colData περιέχουν πληροφορίες σχετικές με τα κύτταρα και τα γονίδια αντίστοιχα. Επίσης μπορούν να αποθηκεύονται μεταδεδομένα που προκύπτουν από τα διαφορετικά στάδια

της ανάλυσης, όπως για παράδειγμα η μείωση των διαστάσεων των δεδομένων κ.α. (Εικόνα 4). Συνοπτικά με αυτό το αντικείμενο ο Bioconductor τυποποιεί την αποθήκευση των δεδομένων και των αποτελεσμάτων, ενισχύοντας έτσι τη διαλειτουργικότητα μεταξύ των πακέτων της ανάλυσης. Ενώ παράλληλα με αυτόν τον τρόπο διευκολύνει τις σύνθετες ροές εργασίας της ανάλυσης (Amezquita, 2019).



Εικόνα 4 Σχηματική αναπαράσταση του αντικειμένου SingleCellExperiment. Η Εικόνα ανήκει στο άρθρο των Amezquita et al. με τίτλο «Orchestrating Single Cell Analysis With Bioconductor», 2019.

1.6 Υπολογιστική ανάλυση των δεδομένων

Η υπολογιστική ανάλυση των δεδομένων που έχουν προκύψει από αλληλούχιση single cell RNA sequencing έχει απασχολήσει σε μεγάλο βαθμό τους ερευνητές με αποτέλεσμα να δημιουργούνται συνέχεια νέα βιοπληροφορικά εργαλεία που εξυπηρετούν τους σκοπούς της εκάστοτε ανάλυσης. Όμως ο βασικός κορμός της ανάλυσης παραμένει ίδιος σε όλες τις μελέτες. Όπως αναφέρθηκε νωρίτερα το πρώτο στάδιο αφορά την προεπεξεργασία των δεδομένων η οποία περιλαμβάνει τον έλεγχο ποιότητας όπου με τη βοήθεια κάποιων χαρακτηριστικών δεικτών ξεχωρίζονται κάποια κύτταρα που θεωρούνται κακής ποιότητας και αποκλείονται από τη μελέτη. Για να γίνει αυτό χρησιμοποιούνται κάποια κατώφλια για τις τιμές των δεικτών και τα δεδομένα φιλτράρονται. Σε αυτό το σημείο απαραίτητη είναι και η οπτικοποίηση των δεδομένων για να φανούν τυχόν αβλεψίες. Έπειτα, η κανονικοποίηση που αποτελεί βασικό βήμα για να

είναι τα δεδομένα συγκρίσιμα μεταξύ τους, ακολουθούμενη από την επιλογή των γονιδίων που παρουσιάζουν την μεγαλύτερη διακύμανση (Highly Variable Genes, HVGs) και είναι με άλλα λόγια τα πληροφοριακά γονίδια για την ανάλυση. Ένα υψηλής ποιότητας σύνολο HVGs θα πρέπει να περιλαμβάνει γονίδια που μπορούν να διακριθούν σε διαφορετικούς κυτταρικούς τύπους, μιας και η ποιότητά τους έχει καθοριστική επίδραση στην ακρίβεια της ομαδοποίησης που θα πραγματοποιηθεί αργότερα. Στη συνέχεια, σειρά έχει η εφαρμογή αλγορίθμων μηχανικής μάθησης για την μείωση των διαστάσεων των δεδομένων και την ομαδοποίησή τους. Η μείωση των διαστάσεων των δεδομένων είναι απαραίτητη σε τόσο μεγάλα σύνολα δεδομένων, τα οποία θα συμπυκνωθούν με τη λιγότερη δυνατή απώλεια πληροφορίας. Επιπλέον η ομαδοποίηση θα δώσει το πρώτο απτό αποτέλεσμα της ανάλυσης καθώς τα κύτταρα θα διαχωριστούν σε ομάδες με κοινά χαρακτηριστικά και θα δοθεί ερμηνεία για τον κυτταρικό τύπο ή την βιολογική κατάσταση στην οποία βρίσκονται. Τέλος ο κάθε ερευνητής οδηγεί την μελέτη του ανάλογα με το πού είναι προσανατολισμένη και συνεχίζει την ανάλυση με διαφορετικούς τρόπους για να φτάσει στα αποτελέσματα που επιθυμεί (Dragomirka Jovic, 2022).

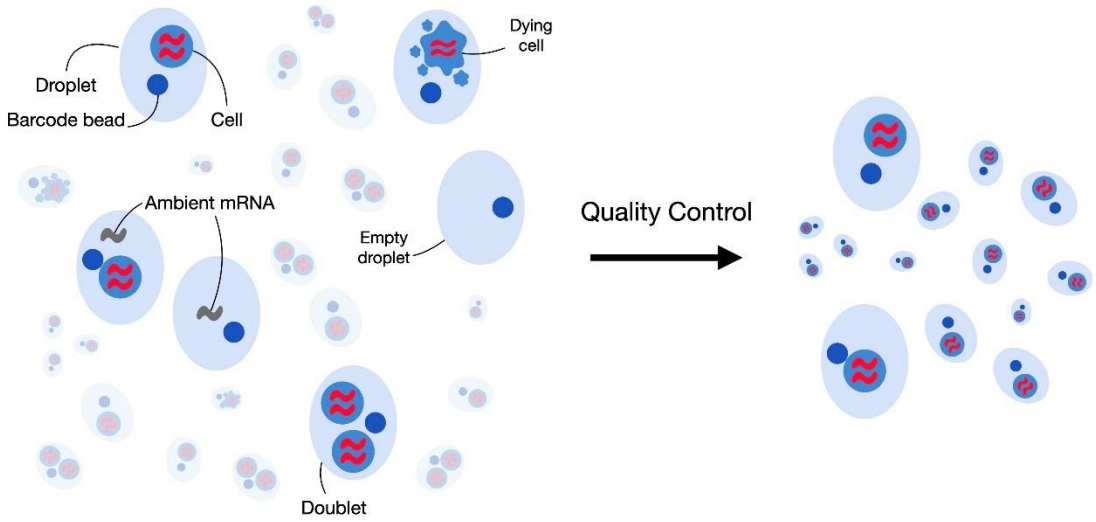
2 Μέθοδοι και εργαλεία

2.1 Έλεγχος ποιότητας (Quality Control)

Κατά την προετοιμασία της βιβλιοθήκης μπορούν να προκύψουν διάφορα προβλήματα όπως κάποια βλάβη στα κύτταρα ή να αποτύχουν οι διαδικασίες της αντίστροφης μεταγραφής και της PCR. Συνεπώς μπορεί να παρατηρηθούν κύτταρα χαμηλής ποιότητας, RNA χωρίς κύτταρα ή ακόμα και διπλότυπα (Εικόνα 5). Τέτοια προβλήματα οδηγούν σε βιβλιοθήκες χαμηλής ποιότητας στα δεδομένα scRNA-seq και συνήθως στην ανάλυση εμφανίζονται ως κύτταρα με χαμηλό αριθμό μετρήσεων (low total counts), λίγα εκφραζόμενα γονίδια και υψηλές αναλογίες μιτοχονδριακών ή spike in μεταγράφων. Συνεπώς είναι σημαντικό να γίνεται έλεγχος της ποιότητας των δεδομένων καθώς παραπλανητικά δεδομένα μπορούν να οδηγήσουν και σε παραπλανητικά αποτελέσματα στη συνέχεια της ανάλυσης.

Πιο συγκεκριμένα οι χαμηλής ποιότητας βιβλιοθήκες στην ανάλυση μπορεί να περιέχουν γονίδια που θα φαίνονται ψευδώς υψηλά εκφρασμένα (upregulated genes). Επίσης αυτά μπορούν να σχηματίσουν δικές τους ξεχωριστές ομάδες (clusters), περιπλέκοντας έτσι την ερμηνεία των αποτελεσμάτων. Επιπλέον παρεμποδίζουν τον χαρακτηρισμό της

ετερογένειας του κυτταρικού πληθυσμού κατά τον υπολογισμό της διακύμανσης και της ανάλυσης των κύριων συνιστωσών (Principal Component Analysis), καθώς οι πρώτες συνιστώσες θα αναπαριστούν διαφορές στην ποιότητα και όχι στην βιολογική πληροφορία (Aaron T.L. Lun D. J., 2016).



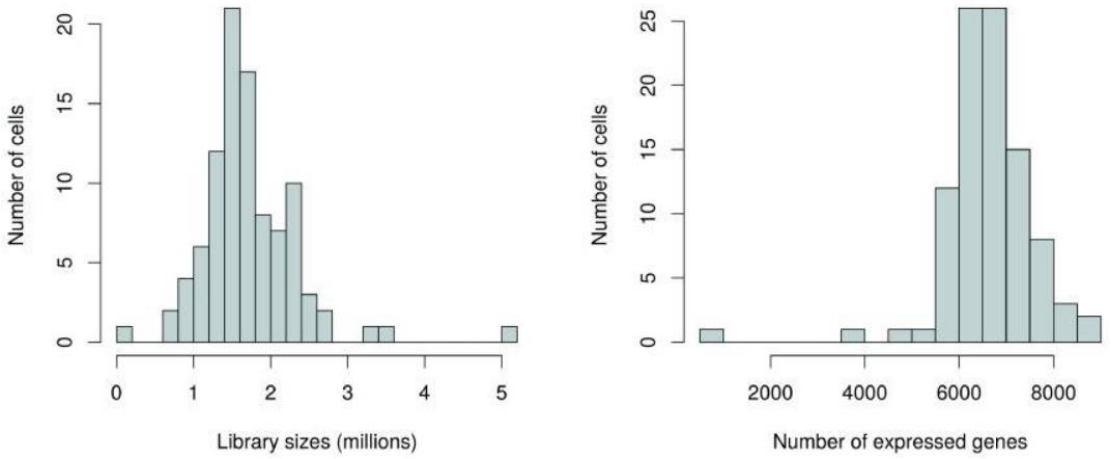
Εικόνα 5 Τα σύνολα δεδομένων που έχουν προκύψει από single cell RNA sequencing υφίστανται ποιοτικό έλεγχο έτσι ώστε να αφαιρεθούν κύτταρα χαμηλής ποιότητας και να διατηρηθεί η χρήσιμη βιολογική πληροφορία. Η εικόνα ανήκει στο βιβλίο «Single cell best practices» από Theis lab.

Για να εξαλειφθούν αυτά τα προβλήματα ή έστω να περιοριστούν, είναι αναγκαίο στην αρχή της ανάλυσης να καθαριστούν τα δεδομένα από τέτοιου είδους προβληματικά κύτταρα. Όμως αυτή η διαδικασία είναι πολύ κρίσιμη και πρέπει να γίνεται με μεγάλη προσοχή καθώς από τη φύση τους τα σύνολα δεδομένων που προκύπτουν από single cell RNA sequencing έχουν κάποιες ιδιαιτερότητες που ο ερευνητής πρέπει να λάβει υπόψιν του. Πρώτον τα δεδομένα αυτά περιλαμβάνουν υπερβολικά πολλά μηδενικά λόγω του περιορισμένου mRNA. Δεύτερον η βιολογική ερμηνεία δημιουργεί περιορισμούς όσον αφορά τη διόρθωση και τον ποιοτικό έλεγχο των δεδομένων καθώς με λάθος χειρισμούς και υποθέσεις μπορεί να χαθεί χρήσιμη βιολογική πληροφορία. Συνεπώς οι μέθοδοι που θα χρησιμοποιηθούν πρέπει να είναι κατάλληλες για τα συγκεκριμένα δεδομένα της εκάστοτε ανάλυσης (Heumos, 2023).

2.1.1 Δείκτες κατά τον έλεγχο ποιότητας

Προκειμένου να αντιμετωπιστούν τα παραπάνω προβλήματα πρέπει να προσδιοριστεί ο όρος βιβλιοθήκη χαμηλής ποιότητας. Ως μέγεθος βιβλιοθήκης (library size) ορίζεται το συνολικό άθροισμα των μετρήσεων σε όλα τα σχετικά χαρακτηριστικά κάθε κυττάρου, δηλαδή τα ενδογενή του γονίδια. Επομένως αυτό είναι ένα ισχυρό μετρικό καθώς κύτταρα με μικρό μέγεθος βιβλιοθήκης, θεωρούνται και χαμηλής ποιότητας. Σε αυτές τις περιπτώσεις το RNA έχει χαθεί κατά τη διαδικασία κατασκευής της βιβλιοθήκης είτε εξαιτίας της κυτταρικής λύσης είτε εξαιτίας μη αποτελεσματικής δημιουργίας της cDNA (Robert Amezquita, 2023).

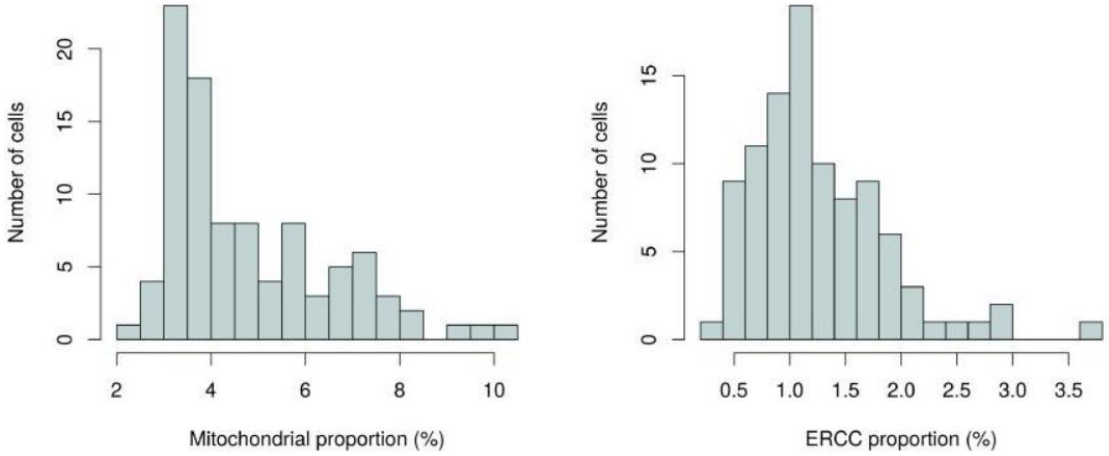
Ένας ακόμη δείκτης που πρέπει να εξετάζεται αφορά τον αριθμό των εκφραζόμενων χαρακτηριστικών σε κάθε κύτταρο, ο οποίος ορίζεται ως ο αριθμός των ενδογενών γονιδίων που έχουν μη μηδενικές μετρήσεις (non zero counts) στο κάθε κύτταρο. Κάθε κύτταρο με πολύ λίγα εκφραζόμενα γονίδια είναι πιθανό να είναι κακής ποιότητας, καθώς ο αριθμός μεταγράφων ίσως δεν έχει αποτυπωθεί επιτυχώς. Στην Εικόνα 6 φαίνονται οι κατανομές των δύο μετρικών.



Εικόνα 6 Το ιστόγραμμα του μεγέθους βιβλιοθήκης (αριστερά) και των εκφρασμένων γονιδίων (δεξιά) για όλα τα κύτταρα στο HSC dataset. Η εικόνα ανήκει στο άρθρο «A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor» των Aaron T.L Lun et al., 2016.

Ένας ακόμα δείκτης ποιότητας είναι η αναλογία των reads που αντιστοιχούν σε γονίδια στο μιτοχονδριακό γονιδίωμα. Τα υψηλά ποσοστά είναι δείκτης κακής ποιότητας κυττάρων, γεγονός που πιθανότατα οφείλεται σε αυξημένη απόπτωση ή και απώλεια κυτταροπλασματικού RNA από τα λυμένα κύτταρα.

Στην ίδια λογική στηρίζονται και τα reads που αντιστοιχούν σε spike in μετάγραφα. Τα spike ins είναι συνθετικά μόρια RNA που εισάγονται σε ένα βιολογικό δείγμα πριν από τη διαδικασία ανάλυσης RNA και σκοπός τους είναι να δίνουν μια εικόνα για την απόδοση, την ακρίβεια και την ευαισθησία της ανάλυσης RNA (Atlas M Sardoo, 2022). Γνωρίζοντας ότι η ίδια ποσότητα spike in RNA θα πρέπει να έχει προστεθεί σε κάθε κύτταρο, οποιαδήποτε αύξηση παρατηρηθεί στον αριθμό των spike in είναι ένδειξη απώλειας ενδογενούς RNA που οδηγεί ξανά στο συμπέρασμα ότι πρόκειται για κακής ποιότητας κύτταρα (Robert Amezquita, 2023). Στην Εικόνα 7 φαίνονται οι κατανομές των μιτοχονδριακών και των spike in αναλογιών σε όλα τα κύτταρα.



Εικόνα 7 Ιστόγραμμα της αναλογίας των reads που αντιστοιχούν σε μιτοχονδριακά γονίδια (αριστερά) και σε spike in μετάγραφα (δεξιά) σε όλα τα κύτταρα στο HSC dataset. Η εικόνα ανήκει στο άρθρο «A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor» των Aaron T.L Lun et al., 2016.

2.1.2 Αναγνώριση και αποκλεισμός κυττάρων χαμηλής ποιότητας

Ο βασικός τρόπος για τον αποκλεισμό των κυττάρων χαμηλής ποιότητας από την ανάλυση περιλαμβάνει την προσθήκη κάποιων κατώτατων ορίων για τους δείκτες που αναλύθηκαν παραπάνω και αφορά μεθόδους αποκλεισμού με σταθερά κατώφλια αλλά και με προσαρμοζόμενα, αναλόγως με τα δεδομένα της ανάλυσης.

2.1.2.1 Αποκλεισμός κυττάρων με χρήση σταθερών κατωφλιών

Η απλούστερη μέθοδος αποκλεισμού κυττάρων χαμηλής ποιότητας είναι βάζοντας στην ανάλυση εκ των προτέρων κάποια σταθερά κατώτατα κατώφλια για τα μετρικά που αναφέρθηκαν. Επομένως για παράδειγμα μπορούν να γίνονται δεκτά στην ανάλυση κύτταρα με μέγεθος βιβλιοθήκης μεγαλύτερο από 100.000 reads, με περισσότερα από 5000 εκφραζόμενα γονίδια και αναλογίες μιτοχονδριακών γονιδίων ή spike in κάτω του

10%. Αυτές οι τιμές είναι ενδεικτικές και αλλάζουν ανάλογα το πρωτόκολλο του κάθε πειράματος και το βιολογικό σύστημα. Δηλαδή δεδομένα count based και UMI based δεν μπορούν να έχουν ίδιες τιμές κατώτατων ορίων. Επίσης διαφορές στην μιτοχονδριακή δραστηριότητα ή στο συνολικό RNA απαιτούν συνεχή προσαρμογή των κατωφλιών των μιτοχονδριακών και των spike in για διαφορετικά βιολογικά συστήματα. Τέλος ακόμα και σε αναλύσεις με το ίδιο πρωτόκολλο και το ίδιο βιολογικό σύστημα, τα κατώτατα αυτά όρια πιθανό να χρειάζονται επαναπροσδιορισμό εξαιτίας της πολυπλοκότητας των βημάτων του πειράματος του single cell RNA sequencing (πχ τις ιδιοτροπίες κατά τη δημιουργία της cDNA) (Robert Amezquita, 2023).

2.1.2.2 Αποκλεισμός κυττάρων με χρήση προσαρμοζόμενων κατωφλιών

Στην περίπτωση του αποκλεισμού των κυττάρων με τη χρήση προσαρμοζόμενων κατωφλιών, γίνεται η παραδοχή ότι το μεγαλύτερο μέρος του συνόλου δεδομένων περιλαμβάνει κύτταρα υψηλής ποιότητας. Στη συνέχεια, εντοπίζονται τα κύτταρα που παρουσιάζουν ακραίες τιμές για τα μετρικά του ελέγχου ποιότητας, χρησιμοποιώντας τη διάμεση απόλυτη απόκλιση (Median Absolute Deviation, MAD) από τη διάμεση τιμή κάθε μετρικής σε όλα τα κύτταρα. Εξ' ορισμού τίθεται ως ακραία τιμή εάν απέχει 3 MADs από τη διάμεσο προς την «προβληματική κατεύθυνση», διότι με αυτόν τον τρόπο θα διατηρηθεί το 99% των μη ακραίων τιμών που ακολουθούν μια κανονική κατανομή (Robert Amezquita, 2023).

Στο πακέτο του Bioconductor στην R υπάρχει η συνάρτηση perCellQCFilters() η οποία υλοποιεί την παραπάνω τεχνική. Δηλαδή αναγνωρίζει χαμηλής ποιότητας κύτταρα ως κύτταρα με ακραίες τιμές και τα αποκλείει από την ανάλυση. Πιο συγκεκριμένα για κάθε στήλη που δίνεται στο όρισμα sub.fields εντοπίζονται μεγάλες ακραίες τιμές, γεγονός που αποσκοπεί στην αφαίρεση των κυττάρων που έχουν υψηλή περιεκτικότητα σε spike in ή σε μιτοχονδριακά γονίδια, τα οποία συνήθως αναπαριστούν κατεστραμμένα κύτταρα (rdrr.io, n.d.).

Με αυτήν την προσέγγιση, τα κατώφλια προσαρμόζονται τόσο στη θέση όσο και στη διασπορά της κατανομής των τιμών για ένα συγκεκριμένο μετρικό. Αυτό έχει ως αποτέλεσμα η διαδικασία ελέγχου ποιότητας να μπορεί να προσαρμόζεται σε αλλαγές στο βάθος της αλληλούχισης, στην αποτελεσματικότητα δημιουργίας της cDNA αλλά και στην αναλογία spike in κλπ. χωρίς να απαιτείται παρέμβαση από τον χρήστη.

2.1.2.3 Άλλες τεχνικές αποκλεισμού κυττάρων χαμηλής ποιότητας

Μια διαφορετική στρατηγική για την αναγνώριση κυττάρων χαμηλής ποιότητας περιλαμβάνει τον εντοπισμό των ακραίων τιμών στον πολυδιάστατο χώρο πάλι με βάσει τους δείκτες που συζητήθηκαν προηγουμένως για κάθε κύτταρο. Για τον σκοπό αυτόν στην R με μεθόδους από το πακέτο robustbase ποσοτικοποιείται το πόσο ακραίο είναι κάθε κύτταρο με βάση τα γνωστά μετρικά και με τη χρήση της συνάρτησης isOutlier() αναγνωρίζονται χαμηλής ποιότητας κύτταρα που εμφανίζουν ασυνήθιστα υψηλά επίπεδα ακραίων τιμών.

Στρατηγικές σαν αυτή βασίζονται στη λογική της PCA και φαίνεται να είναι πολύ αποτελεσματικές στη διάκριση των κυττάρων χαμηλής και υψηλής ποιότητας. Ωστόσο συμβαίνει συχνά να αποκλείουν από την ανάλυση κύτταρα χωρίς προφανείς λόγους και αυτό να δημιουργεί προβλήματα στην ερμηνεία των αποτελεσμάτων.

Τέλος ένας ακόμα τρόπος ανίχνευσης των ακραίων τιμών μπορεί να προέλθει από τα προφίλ γονιδιακής έκφρασης. Όμως αυτή είναι μια μέθοδος που έχει μεγάλο ρίσκο να χαθούν κύτταρα υψηλής ποιότητας σε σπάνιους πληθυσμούς (Robert Amezquita, 2023).

2.1.3 Έλεγχος με διαγνωστικά διαγράμματα

Είναι σημαντικό να επαναλάβουμε ότι οι αποφάσεις που θα παρθούν για κάθε βήμα και τα εργαλεία που θα χρησιμοποιηθούν επηρεάζουν καίρια την ανάλυση. Για παράδειγμα εάν αποκλειστούν πάρα πολλά κύτταρα αυτό θα μπορούσε να οδηγήσει στο να χαθούν σπάνιοι κυτταρικοί υποπληθυσμοί και κρίσιμη βιολογική πληροφορία. Στον αντίποδα εάν το φιλτράρισμα γίνει με μεγάλη επιείκεια και κρατηθούν στην ανάλυση κύτταρα χαμηλής ποιότητας, αυτό θα δυσκολέψει αργότερα τα βήματα της ερμηνείας και του σχολιασμού των αποτελεσμάτων (Heumos, 2023).

Μια καλή πρακτική σε αυτό το στάδιο της ανάλυσης είναι ο έλεγχος των κατανομών των μετρικών του ποιοτικού ελέγχου και του φιλτραρίσματος που έγινε στο προηγούμενο βήμα, έτσι ώστε να εντοπιστούν εγκαίρως πιθανά προβλήματα. Τα αναμενόμενα αποτελέσματα στην ιδανική περίπτωση είναι κανονικές κατανομές οι οποίες θα δικαιολογούσαν και το κατώφλι 3 MAD που αναλύθηκε στο προηγούμενο κεφάλαιο για την ανίχνευση των ακραίων τιμών. Στην περίπτωση που τα αποτελέσματα δείχνουν ένα μεγάλο ποσοστό των κυττάρων σε μια άλλη κατάσταση φανερώνει ότι τα μετρικά του ελέγχου ποιότητας είναι πιθανό να σχετίζονται με κάποια βιολογική κατάσταση. Αυτό θα μπορούσε να οδηγήσει στην απώλεια διακριτών τύπων κυττάρων στα επόμενα στάδια

της ανάλυσης. Επίσης ενδεχομένως να υποδηλώνει ότι υπήρξαν ασυνέπειες κατά τη διαδικασία προετοιμασίας της βιβλιοθήκης για ένα υποσύνολο κυττάρων, κάτι που συμβαίνει συχνότερα στα plate based πρωτόκολλα (Robert Amezquita, 2023).

Ένας άλλος τρόπος διαγνωστικού ελέγχου περιλαμβάνει την απεικόνιση του ποσοστού του μιτοχονδριακών μετρήσεων σε σχέση με τα υπόλοιπα μετρικά. Αυτό αποσκοπεί στο να φανεί ξεκάθαρα εάν υπάρχουν κύτταρα με μεγάλες συνολικές μετρήσεις (total counts) και μεγάλα ποσοστά μιτοχονδριακού γονιδιώματος, έτσι ώστε να γίνει σαφές πως δεν αφαιρούνται κατά λάθος υψηλής ποιότητας κύτταρα που τυχαίνει να είναι ιδιαίτερα μεταβολικά ενεργά.

Αναγνωριστικό μπορεί να είναι επίσης η σύγκριση των ποσοστών των spike in και του μιτοχονδριακού γονιδιώματος. Είναι πιθανό να βρεθούν κύτταρα με μικρά ποσοστά μιτοχονδριακού γονιδιώματος, μεγάλα ποσοστά spike in μεταγράφων και μικρά μεγέθη βιβλιοθήκης, τα οποία είναι κύτταρα κακής ποιότητας καθώς υποθέτουμε ότι αποτελούν απογυμνωμένους πυρήνες. Αυτό πρακτικά σημαίνει πως σε κάποιο στάδιο του πειράματος έπαθαν τόσο μεγάλη βλάβη που έχουν χάσει όλο το κυτταροπλασματικό τους περιεχόμενο. Από την άλλη πλευρά κύτταρα με υψηλά ποσοστά μιτοχονδριακού γονιδιώματος και χαμηλά ποσοστά spike in μεταγράφων είναι πιθανό να αντιπροσωπεύουν μη κατεστραμμένα κύτταρα τα οποία είναι μεταβολικά ενεργά (Robert Amezquita, 2023).

2.2 Κανονικοποίηση (Normalization)

Στα δεδομένα single cell RNA sequencing συχνά παρατηρούνται διαφορές στην κάλυψη αλληλούχισης μεταξύ των βιβλιοθηκών. Αυτές οι διαφορές προκύπτουν από τεχνικά προβλήματα στην αποτελεσματικότητα λήψης της cDNA ή στην ενίσχυση PCR. Η κανονικοποίηση αποσκοπεί στην εξάλειψη αυτών των διαφορών έτσι ώστε να μην παρεμβαίνουν στις συγκρίσεις των προφίλ έκφρασης μεταξύ των κυττάρων. Αυτό το βήμα είναι καθοριστικής σημασίας για την εξέλιξη της ανάλυσης καθώς εξασφαλίζεται ότι οποιαδήποτε παρατηρούμενη ετερογένεια ή διαφορετική έκφραση εντός του κυτταρικού πληθυσμού οφείλεται καθαρά στη βιολογία και όχι σε τεχνικές διακυμάνσεις.

Μια πολύ διαδεδομένη στρατηγική καλείται κανονικοποίηση κλιμάκωσης και περιλαμβάνει τη διαίρεση όλων των μετρήσεων για κάθε κύτταρο με έναν ειδικό για κάθε κύτταρο παράγοντα κλιμάκωσης, που συχνά ονομάζεται «παράγοντας μεγέθους» (size factor). Η υπόθεση εδώ είναι ότι οποιαδήποτε μεροληψία για συγκεκριμένο κύτταρο (πχ

στην αποτελεσματικότητα λήψης ή ενίσχυσης) επηρεάζει εξίσου όλα τα γονίδια μέσω της κλιμάκωσης του αναμενόμενου μέσου αριθμού για το συγκεκριμένο κύτταρο. Ο παράγοντας μεγέθους για κάθε κύτταρο αντιπροσωπεύει την εκτίμηση της σχετικής μεροληψίας σε αυτό το κύτταρο, οπότε η διαίρεση των μετρήσεων του με τον παράγοντα μεγέθους του θα πρέπει να αφαιρέσει την εν λόγω μεροληψία. Οι κανονικοποιημένες τιμές έκφρασης που θα προκύψουν μπορούν στην συνέχεια να χρησιμοποιηθούν για μεταγενέστερες αναλύσεις, όπως η ομαδοποίηση (clustering) και η μείωση της διαστατικότητας (dimensionality reduction) (Aaron T.L. Lun D. J., 2016).

2.3 Επιλογή των Γονιδίων (Feature Selection)

Το επόμενο στάδιο της ανάλυσης ονομάζεται επιλογή γονιδίων (Feature Selection) και είναι απαραίτητο καθώς η σύγκριση των κυττάρων που θα γίνει στα μετέπειτα βήματα βασίζεται στα προφίλ γονιδιακής έκφρασης έτσι ώστε να χαρακτηριστεί η ετερογένειά τους. Πιο συγκεκριμένα οι αλγόριθμοι της ομαδοποίησης (clustering) και της μείωσης των διαστάσεων των δεδομένων (dimensionality reduction) εμπλέκουν τα προφίλ γονιδιακής έκφρασης σε μια μέτρηση «ομοιότητας» μεταξύ των κυττάρων. Συνεπώς η επιλογή των γονιδίων δεν μπορεί να γίνει τυχαία. Χρειάζεται να διερευνηθεί η βιολογική πληροφορία που προσφέρουν και να αποφευχθούν εκείνα που περιέχουν τυχαίο θόρυβο. Έτσι λοιπόν διατηρείται η βιολογική σημασία, απομακρύνεται η διακύμανση που την καλύπτει και τα δεδομένα γίνονται λιγότερα εξυπηρετώντας και στην υπολογιστική αποδοτικότητα των βημάτων που θα ακολουθήσουν. Η απλούστερη προσέγγιση για την επιλογή των γονιδίων είναι η αναζήτηση των γονιδίων αυτών που έχουν την πιο μεταβλητή έκφραση ανάμεσα στον πληθυσμό. Φυσικά για αυτό γίνεται η παραδοχή ότι οι γνήσιες βιολογικές διαφορές θα εμφανιστούν ως αυξημένη διακύμανση στα επηρεαζόμενα γονίδια, σε σύγκριση με εκείνα που επηρεάζονται μόνο από τεχνικό θόρυβο (Amezquita, 2019).

2.3.1 Ποσοτικοποίηση της διακύμανσης ανά γονίδιο

Η διαδικασία της επιλογής των γονιδίων ξεκινάει με την ποσοτικοποίηση της μεταβλητότητας ανά γονίδιο. Ο απλούστερος τρόπος για να γίνει αυτό είναι με τον υπολογισμό της διακύμανσης της λογαριθμικής κανονικοποιημένης έκφρασης (log normalized expression values) του κάθε γονιδίου σε όλα τα κύτταρα. Η κύρια ευελιξία που προσφέρει αυτή η προσέγγιση είναι πως τα γονίδια που θα επιλεγούν βασίζονται στις ίδιες λογαριθμικές τιμές που θα χρησιμοποιηθούν στους αλγορίθμους των επόμενων βημάτων της ανάλυσης. Πιο συγκεκριμένα γονίδια με μεγαλύτερες

διακυμάνσεις στις λογαριθμικές τους τιμές θα συνεισφέρουν περισσότερο στις Ευκλείδειες αποστάσεις μεταξύ των κυττάρων κατά τις διαδικασίες της ομαδοποίησης και της μείωσης των διαστάσεων. Ένα ακόμα πλεονέκτημα αφορά τη συνάφεια της ανάλυσης καθώς χρησιμοποιώντας λογαριθμικές τιμές σε αυτό το στάδιο εξασφαλίζεται ότι ο ποσοτικός ορισμός της ετερογένειας των κυττάρων θα είναι σταθερός κατά τη διάρκεια όλης της ανάλυσης (Robert Amezquita, 2023).

Κάθε τεχνική έρχεται με πλεονεκτήματα αλλά και μειονεκτήματα. Στην προκειμένη το πρόβλημα είναι πως ο λογαριθμικός μετασχηματισμός στις περισσότερες περιπτώσεις δεν μπορεί από μόνος του να σταθεροποιήσει την μεταβλητότητα. Αυτό σημαίνει πως η συνολική μεταβλητότητα ενός γονιδίου επηρεάζεται περισσότερο από την αφθονία στην οποία θα βρεθεί, παρά από την υποκείμενη βιολογική του ετερογένεια. Για να ληφθεί υπόψη αυτό χρησιμοποιούνται διάφορες συναρτήσεις όπως η modelGeneVar() του πακέτου scran η οποία προσαρμόζει μια τάση στη διακύμανση σε σχέση με την αφθονία σε όλα τα γονίδια. Σε κάθε δεδομένη αφθονία υποθέτουμε ότι η διακύμανση της έκφρασης για τα περισσότερα γονίδια οφείλεται σε διαδικασίες που δεν αφορούν τη μελέτη, όπως για παράδειγμα ο θόρυβος δειγματοληψίας. Υπό αυτήν την παραδοχή η προσαρμοσμένη τιμή τάσης σε κάθε δεδομένη αφθονία γονιδίου αντιπροσωπεύει μια εκτίμηση της «μη ενδιαφέρουσας» μεταβολής της, η οποία καλείται τεχνική συνιστώσα. Στη συνέχεια ορίζεται η βιολογική (bio) συνιστώσα για κάθε γονίδιο ως η διαφορά μεταξύ της συνολικής (total) διακύμανσης και της τεχνικής (tech) συνιστώσας. Η επονομαζόμενη βιολογική συνιστώσα αντιπροσωπεύει την «ενδιαφέρουσα» διακύμανση για κάθε γονίδιο και μπορεί να χρησιμοποιηθεί για την επιλογή των γονιδίων υψηλής μεταβλητότητας (Highly Variable Genes, HVGs). Αξίζει να σημειωθεί ότι ορισμένα γονίδια έχουν αρνητικές τιμές στις βιολογικές συνιστώσες, οι οποίες δεν έχουν προφανή ερμηνεία και μπορούν να αγνοηθούν. Όμως κατά την προσαρμογή μιας τάσης στις διακυμάνσεις των γονιδίων, δεν μπορεί να παραληφθούν καθώς περίπου τα μισά γονίδια θα βρίσκονται κάτω από την τάση. Η ερμηνεία της προσαρμοσμένης τάσης ως τεχνική συνιστώσα προϋποθέτει ότι τα προφίλ έκφρασης των περισσότερων γονιδίων κυριαρχούνται από τυχαίο τεχνικό θόρυβο. Στην πραγματικότητα όλα τα εκφραζόμενα γονίδια θα παρουσιάζουν κάποιο μη μηδενικό επίπεδο βιολογικής μεταβλητότητας λόγω γεγονότων όπως η μεταγραφική έκρηξη (transcriptional bursting). Συνεπώς θα ήταν καταλληλότερο να θεωρηθούν αυτές οι εκτιμήσεις ως τεχνικός θόρυβος ή «μη ενδιαφέρουσα βιολογική μεταβολή», με την παραδοχή βέβαια ότι τα περισσότερα γονίδια δεν συμμετέχουν στις

διαδικασίες που οδηγούν σε ενδιαφέρουσα ετερογένεια στον πληθυσμό (Robert Amezquita, 2023).

2.3.2 Ποσοτικοποίηση του τεχνικού θορύβου

Είναι πιθανό η υπόθεση που αναλύθηκε στην προηγούμενη ενότητα να αντιμετωπίζει κάποιες δυσκολίες σε εξαιρετικές περιπτώσεις όπου πολλά γονίδια σε μια συγκεκριμένη ποσότητα επηρεάζονται από μια βιολογική διαδικασία. Το πιο προφανές παράδειγμα αφορά γονίδια που είναι εξειδικευμένα για έναν τύπο κυττάρων και συνεπώς υπερεκφράζονται, οδηγώντας έτσι σε εμπλουτισμό των γονιδίων υψηλής μεταβλητότητας (HVGs). Αυτό θα μπορούσε να δυσκολέψει τον εντοπισμό των HVGs, καθώς η προσαρμοσμένη τάση θα διογκωνόταν σε αυτό το εύρος ποσότητας. Για να δοθεί λύση στο παραπάνω πιθανό σενάριο γίνεται η παραδοχή ότι τα spike in δεν επηρεάζονται από τη βιολογική διακύμανση, και η τάση που θα προσαρμοστεί στη διακύμανσή τους γίνεται εξαρτώμενη από τον μέσο όρο (Robert Amezquita, 2023).

Στην περίπτωση όπου απουσιάζουν spike in μετάγραφα από την ανάλυση, η δημιουργία μιας γραμμής τάσης θα μπορούσε να γίνει με κάποιες υποθέσεις κατανομής με βάση τον θόρυβο. Δηλαδή θόρυβος που θα μπορούσε να έχει προκύψει από την προετοιμασία της βιβλιοθήκης και την αλληλούχιση μπορεί να χρησιμοποιηθεί σε σχέση με τις μετρήσεις των UMIs τα οποία συνήθως εμφανίζουν διακύμανση κοντά σε αυτήν της κατανομής Poison (Po-Yuan Tung, 2017).

Ενδιαφέρουσα είναι η παρατήρηση πως οι τάσεις που βασίζονται καθαρά στον τεχνικό θόρυβο τείνουν να αποδίδουν μεγάλες βιολογικές συνιστώσες για τα γονίδια που υπερεκφράζονται. Είναι συχνό σε αυτές τις περιπτώσεις οι αναλύσεις να περιλαμβάνουν τα λεγόμενα «house keeping» γονίδια, δηλαδή γονίδια που εμπλέκονται σε όλες τις βασικές κυτταρικές λειτουργίες, αλλά όμως θεωρούνται αδιάφορα στον χαρακτηρισμό της κυτταρικής ετερογένειας που ενδιαφέρει τη δεδομένη μελέτη. Το γεγονός αυτό είναι μια ισχυρή ένδειξη ότι ένα ακριβέστερο μοντέλο θορύβου δεν αποδίδει απαραίτητα και την καλύτερη επιλογή των HVGs (Robert Amezquita, 2023).

2.3.3 Τελική επιλογή των γονιδίων με την μεγαλύτερη μεταβλητότητα

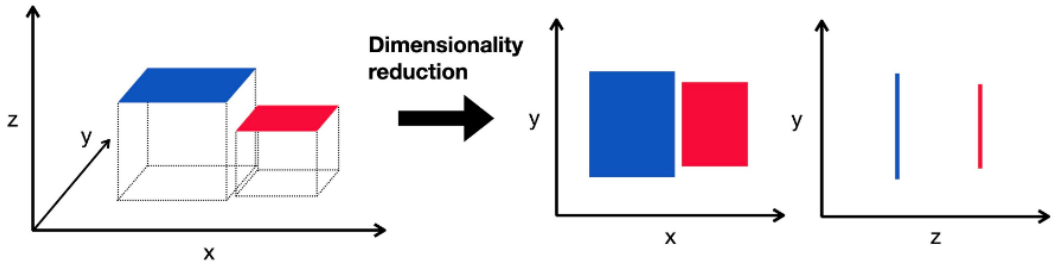
Αφότου ολοκληρωθεί η ποσοτικοποίηση της διακύμανσης ανά γονίδιο, το επόμενο βήμα είναι η επιλογή των HVGs που θα χρησιμοποιηθούν και στη συνέχεια της ανάλυσης. Η πιο προφανής στρατηγική περιλαμβάνει την επιλογή των κορυφαίων n γονιδίων με τις μεγαλύτερες τιμές για τη διακύμανση που μας ενδιαφέρει. Πλεονέκτημα αυτής της

μεθόδου είναι πως ο χρήστης μπορεί να κατευθύνει και να έχει έλεγχο στον αριθμό των γονιδίων που θα επιλεγούν, γεγονός που διασφαλίζει ότι η υπολογιστική πολυπλοκότητα στη συνέχεια της ανάλυσης θα είναι προβλέψιμη. Είναι δύσκολο να δοθεί ένα συγκεκριμένο πλαίσιο επιλογής καθώς κάτι που μπορεί να θεωρηθεί θόρυβος σε μια μελέτη, να είναι χρήσιμη πληροφορία σε μια άλλη. Πιο αναλυτικά παίρνοντας ένα μεγαλύτερο υποσύνολο γονιδίων ως HVGs μειώνεται ο κίνδυνος να χαθεί ενδιαφέρουσα βιολογική πληροφορία, αλλά παράλληλα αυξάνεται το ρίσκο να διατηρηθεί θόρυβος που θα εμποδίσει το σήμα της ενδιαφέρουσας πληροφορίας. Κάτι που πιθανώς θα οδηγούσε σε παραπλανητικά αποτελέσματα στην ερμηνεία της ανάλυσης (Robert Amezquita, 2023).

2.4 Μείωση των διαστάσεων των δεδομένων (Dimensionality Reduction)

Ένα ανθρώπινο σύνολο δεδομένων που έχει προκύψει από single cell RNA sequencing μπορεί να έχει τιμές έκφρασης μέχρι και για 25000 γονίδια. Όπως αναλύθηκε και στα προηγούμενα κεφάλαια πολλά από αυτά τα γονίδια δεν θα είναι πληροφοριακά και πολλά θα έχουν μηδενικές μετρήσεις. Συνεπώς το σύνολο δεδομένων θα φιλτραριστεί και θα καθαριστεί από αυτά. Όμως και πάλι ο αριθμός των γονιδίων θα κυμαίνεται ανάμεσα σε αρκετές χιλιάδες γονίδια, που συνεπάγεται και χώρο διαστάσεων για χιλιάδες χαρακτηριστικά. Επομένως χρειάζεται να εφαρμοστούν μέθοδοι για τη μείωση της διαστατικότητας των δεδομένων, για να μικρύνει με αυτόν τον τρόπο και η υπολογιστική πολυπλοκότητα της ανάλυσης (Theis, 2019) .

Η μείωση των διαστάσεων αποσκοπεί στη μείωση του αριθμού των ξεχωριστών διαστάσεων στα δεδομένα. Στην πράξη, αντί να αποθηκεύονται ξεχωριστές πληροφορίες για μεμονωμένα γονίδια, συμπτύσσονται πολλαπλά χαρακτηριστικά σε μία μόνο διάσταση. Αυτό είναι εφαρμόσιμο διότι διαφορετικά γονίδια φαίνεται να συσχετίζονται εάν αυτά επηρεάζονται από την ίδια βιολογική διαδικασία. Συνεπώς μετά από αυτό το βήμα τα δεδομένα αποτυπώνονται σε χαμηλότερες διαστάσεις που στοχεύουν στη διατήρηση των πιο ουσιαστικών δομών στο σύνολο δεδομένων (Εικόνα 8). Άμεση συνέπεια αυτού, είναι και η ελαχιστοποίηση του θορύβου με τη δημιουργία «μέσου όρου» στα πολλαπλά γονίδια έτσι ώστε να αναπαρασταθεί μια ακριβέστερη απεικόνιση των μοτίβων στα δεδομένα. Παράλληλα εξοικονομείται υπολογιστική ισχύς στα επόμενα βήματα της ανάλυσης καθώς οι υπολογισμοί θα πρέπει να εκτελούνται μόνο για λίγες διαστάσεις αντί για χιλιάδες γονίδια (Amezquita, 2019).



Εικόνα 8 Η μείωση της διαστατικότητας ενσωματώνει τα δεδομένα μεγάλης διάστασης, σε έναν χώρο χαμηλότερης διάστασης. Η εικόνα ανήκει στο βιβλίο «Single cell best practices» από Theis lab.

2.4.1 Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis)

Η πιο διαδεδομένη μέθοδος για τη μείωση των διαστάσεων είναι η ανάλυση κύριων συνιστωσών (PCA). Η PCA ανιχνεύει άξονες (κύριες συνιστώσες, PCs) στον χώρο υψηλών διαστάσεων που αποδίδουν το μεγαλύτερο ποσοστό διακύμανσης. Οι κορυφαίοι (πρώτοι) άξονες PCs αποτυπώνουν τους επικρατέστερους παράγοντες ετερογένειας στο σύνολο των δεδομένων και συνεπώς μπορούν να χρησιμοποιηθούν για τη μείωση της διαστατικότητας (Amezquita, 2019).

Με απλά λόγια μπορούμε να φανταστούμε κάθε άξονα ως μια γραμμή. Έστω ότι σχεδιάζουμε μια γραμμή οπουδήποτε και μετακινούμε κάθε κύτταρο στο σύνολο δεδομένων μας στην κοντινότερη θέση της γραμμής. Η διακύμανση που καταλαμβάνει αυτός ο άξονας ορίζεται ως η διακύμανση των θέσεων των κελιών κατά μήκος αυτής της γραμμής. Επομένως, ο πρώτος άξονας (πρώτη κύρια συνιστώσα, PC1) επιλέγεται έτσι ώστε να μεγιστοποιεί αυτή τη διακύμανση. Ο επόμενος άξονας (PC2) θα διαλεχθεί έτσι ώστε να είναι ορθογώνιος με τον πρώτο και να καταγράφει το μεγαλύτερο ποσοστό διακύμανσης που απομένει. Η διαδικασία συνεχίζεται με την ίδια λογική και για τις υπόλοιπες συνιστώσες (Robert Amezquita, 2023).

Στόχος της PCA είναι να αναγνωρίσει γραμμικούς συνδυασμούς των μεταβλητών (στην προκειμένη περίπτωση των γονιδίων) οι οποίοι μεγιστοποιούν τη διακύμανση που ερμηνεύεται με κάθε συνδυασμό. Όμως τα δεδομένα που έχουν προκύψει από single cell RNA sequencing δεν είναι γραμμικά. Αυτό έχει σαν αποτέλεσμα η PCA συχνά να αποτυγχάνει να αποτυπώσει αρκετές πληροφορίες μόλις σε 2 διαστάσεις, οι οποίες βοηθούν στην οπτικοποίηση και την κατανόηση των δεδομένων από τον ερευνητή. Παρ’

όλα αυτά θετικό αυτής της μεθόδου είναι πως διατηρεί δεκάδες άξονες/διαστάσεις οι οποίες είναι πολύ χρήσιμες σε μεταγενέστερα στάδια της ανάλυσης όπως η ομαδοποίηση (clustering) (Kevin R. Moon, 2017).

2.4.2 T – Distributed Stochastic Neighbor

H t – distributed stochastic neighbor (t-SNE) είναι μια ακόμα κοινή μέθοδος που χρησιμοποιείται για μείωση των διαστάσεων. Εδώ, η οπτικοποίηση διαμορφώνεται με την τοποθέτηση παρόμοιων κυττάρων -σύμφωνα με τις μετρήσεις ομοιότητας στις τιμές της γονιδιακής τους έκφρασης- κοντά το ένα με το άλλο και τοποθετώντας κύτταρα που δεν μοιάζουν μακριά το ένα με το άλλο στον χώρο μικρότερης διάστασης. Στον αλγόριθμο αυτό όμως δίνεται περισσότερη σημασία στα παρόμοια κύτταρα και συνεπώς σε ένα διάγραμμα t-SNE οι αποστάσεις (global distances) και οι σχετικές θέσεις των ομάδων δεν έχουν ιδιαίτερη σημασία (Kevin R. Moon, 2017).

Είναι ενδιαφέρον το ότι φαίνεται συχνά να λειτουργεί καλύτερα από την PCA για τον διαχωρισμό κυττάρων σε περισσότερο διαφορετικούς πληθυσμούς. Αυτό συμβαίνει διότι η t-SNE μπορεί να αποτυπώσει κατευθείαν τις μη γραμμικές σχέσεις σε έναν χώρο πολλών διαστάσεων, ενώ η PCA πρέπει να τις αναπαραστήσει ως γραμμικές συνιστώσες. Ωστόσο η t-SNE είναι μια στοχαστική μέθοδος γεγονός που σημαίνει πως ο χρήστης θα πρέπει να εκτελέσει αρκετές φορές τον αλγόριθμο ώστε να διασφαλίσει πως τα αποτελέσματα είναι αντιπροσωπευτικά και έπειτα να διασφαλίσει και το ότι είναι αναπαραγωγίσιμα. Επομένως όλα αυτά τα βήματα αυξάνουν την υπολογιστική πολυπλοκότητα αυτής της μεθόδου (Aaron T.L. Lun D. J., 2016).

Τέλος θα πρέπει να λαμβάνεται υπόψιν η περιπλοκότητα (perplexity), μια σημαντική παράμετρος που καθορίζει τη λεπτομέρεια απεικόνισης. Φαίνεται πως όταν ο χρήστης δίνει χαμηλές τιμές περιπλοκότητας ευνοείται η ανάλυση λεπτότερης δομής. Αυτό μπορεί να διακινδυνέψει η οπτικοποίηση να επηρεαστεί από τυχαίο θόρυβο. Συνεπώς συνιστάται να δοκιμάζονται διαφορετικές τιμές περιπλοκότητας για να εξασφαλιστεί ότι η επιλογή της δεν θα κατευθύνει την ερμηνεία του γραφήματος t-SNE (Robert Amezquita, 2023).

2.4.3 Uniform manifold approximation and projection

H Uniform manifold approximation and projection (UMAP) είναι κι αυτή μια μη γραμμική μέθοδος μείωσης των διαστάσεων που μοιάζει πολύ με την t-SNE που αναλύθηκε προηγουμένως. Προσπαθεί κι αυτή να βρει μια χαμηλής διάστασης αναπαράσταση που

να διατηρεί τις σχέσεις μεταξύ των γειτονικών κυττάρων στον χώρο υψηλής διάστασης. Διαφοροποιείται από την t-SNE όσον αφορά τις εξισώσεις στάθμισης στα γραφήματα, γεγονός που δίνει και διαφορετική απεικόνιση. Σε σύγκριση με την t-SNE η απεικόνιση που δίνει η UMAP έχει πιο συμπαγείς ομάδες με περισσότερο κενό χώρο ανάμεσά τους. Επιπλέον προσπαθεί να διατηρήσει τη σημασία των αποστάσεων μεταξύ των ομάδων. Από άποψη οικονομίας είναι μια γρήγορη μέθοδος, χαρακτηριστικό που την κάνει πολύ διαδεδομένη σε μεγάλα σύνολα δεδομένων. Επίσης και σε αυτήν την περίπτωση χρειάζεται προσοχή ώστε να μπορεί να είναι επαναλήψιμη, μιας και περιλαμβάνει μια σειρά από βήματα τυχαιοποίησης.

Διάφοροι παράμετροι επηρεάζουν την οπτικοποίηση κι εδώ, με τον αριθμό των γειτόνων και την ελάχιστη απόσταση μεταξύ των σημείων να έχουν την μεγαλύτερη επίδραση στα διαγράμματα. Στην περίπτωση που ο χρήστης δώσει πολύ χαμηλές τιμές σε αυτές τις παραμέτρους η ερμηνεία κινδυνεύει να αντιμετωπίσει τον τυχαίο θόρυβο εσφαλμένα ως μια δομή υψηλής ανάλυσης. Από την άλλη πλευρά στην περίπτωση που ο χρήστης δώσει πολύ υψηλές τιμές θα κερδηθεί μια ακριβής απόδοση ολόκληρου του συνόλου δεδομένων, με κόστος όμως μια λεπτομερέστερη δομή. Επομένως και πάλι χρειάζεται να γίνεται δοκιμή στις παραμέτρους για να διασφαλίζεται ότι δεν θέτουν σε κίνδυνο τα συμπεράσματα που προκύπτουν από ένα διάγραμμα UMAP (Robert Amezquita, 2023).

2.5 Ομαδοποίηση (Clustering)

Στα προηγούμενα κεφάλαια αναπτύχθηκαν οι τρόποι με τους οποίους γίνεται η προεπεξεργασία και η οπτικοποίηση των δεδομένων που προκύπτουν από το single cell RNA sequencing. Αυτές οι διαδικασίες βοήθησαν στον καθαρισμό, την περιγραφή αλλά και την μείωση της διαστατικότητας του συνόλου δεδομένων. Εντούτοις, εξακολουθούν να τα δεδομένα να παραμένουν αφηρημένα. Για αυτό σειρά έχει ο προσδιορισμός της κυτταρικής δομής και ετερογένειας στο σύνολο δεδομένων.

Πιο συγκεκριμένα σε μια ανάλυση single cell RNA sequencing ο προσδιορισμός της κυτταρικής δομής γίνεται με μια διαδικασία που ονομάζεται σχολιασμός της κυτταρικής ταυτότητας και περιλαμβάνει τον εντοπισμό κυτταρικών ταυτοτήτων, οι οποίες συνδέονται με γνωστές κυτταρικές καταστάσεις ή στάδια του κυτταρικού κύκλου. Επομένως πρώτο βήμα για να γίνει αυτό, είναι ο διαχωρισμός των κυττάρων σε ομάδες έτσι ώστε παρόμοια κύτταρα με κοινή ταυτότητα να έρθουν κοντά (Heumos, 2023).

Η ομαδοποίηση είναι μια διαδικασία μάθησης χωρίς επίβλεψη, ανήκει στον χώρο της μηχανικής μάθησης (Machine Learning) και σε αυτήν την περίπτωση χρησιμοποιείται για τον εμπειρικό καθορισμό ομάδων κυττάρων με παρόμοια προφίλ έκφρασης. Είναι βασικό στάδιο σε μια τέτοιου είδους ανάλυση, καθώς βοηθά στην παρουσίαση των πολύπλοκων δεδομένων από το single cell RNA sequencing, συνοψίζοντάς τα σε μια μορφή πιο φιλική για να δοθεί ανθρώπινη ερμηνεία. Πιο αναλυτικά, η πολυδιάστατη πολλαπλότητα στην οποία βρίσκονται στην πραγματικότητα τα κύτταρα, τροποποιείται έτσι ώστε να είναι εύκολα κατανοητή και να επιτρέπεται η περιγραφή της ετερογένειας του πληθυσμού με διακριτές ετικέτες. Μετά τον σχολιασμό με βάση κάποια γονίδια δείκτες (marker genes), οι ομάδες μπορούν πλέον να αντιμετωπιστούν ως αντιπροσωπευτικά για πιο αφηρημένες βιολογικές έννοιες όπως τύποι κυττάρων ή καταστάσεις (Robert Amezquita, 2023).

Μια παρομοίωση της ομαδοποίησης είναι αυτή με το μικροσκόπιο, αφού κατά την εξερεύνηση των δεδομένων μπορεί να γίνει μεγέθυνση ή σμίκρυνση αλλάζοντας τις παραμέτρους της ομαδοποίησης ή και αλλάζοντας τελείως μέθοδο ομαδοποίησης. Μάλιστα είναι πολύ σημαντικό να γίνεται πειραματισμός με τις διαφορετικές μεθόδους ομαδοποίησης που διατίθενται καθώς μπορούν να βρεθούν εναλλακτικές προοπτικές των δεδομένων. Φυσικά γεννούνται ερωτήματα έπειτα από αυτό όπως «Είναι ορθή η ομαδοποίηση που επιλέχθηκε;» ή «Ποιός είναι ο πραγματικός αριθμός των ομάδων;». Όμως τέτοιου είδους ερωτήματα δεν έχουν σημασία διότι μπορούν όλες οι ομαδοποιήσεις να είναι κατά μια βάση «ορθές». Αυτό συμβαίνει διότι κάθε ομαδοποίηση αντιπροσωπεύει τη δική της διαμερισματοποίηση του χώρου της υψηλής διάστασης των εκφράσεων και είναι τόσο «πραγματική» όσο και οποιαδήποτε άλλη.

Θα ήταν ωστόσο ωφέλιμο να αναρωτηθεί κανείς το κατά πόσο οι ομάδες που έχουν δημιουργηθεί προσεγγίζουν τους κυτταρικούς τύπους ή τις υπό μελέτη καταστάσεις. Η απάντηση για αυτό το ερώτημα είναι δύσκολο να δοθεί καθώς εξαρτάται από την υποκείμενη βιολογική ερμηνεία. Αυτό το ερώτημα πάντα εξαρτάται από την ανάλυση και τον ερευνητή, για παράδειγμα κάποιος ενδιαφέρεται για την εύρεση των κύριων κυτταρικών τύπων, ενώ κάποιος άλλος επιθυμεί την ανάλυση υποτύπων ή την εξερεύνηση διαφορετικών καταστάσεων εντός αυτών των υποτύπων, όπως η μεταβολική δραστηριότητα, το στρες κ.α. Συνεπώς προκύπτει δύο ομαδοποιήσεις να είναι εξαιρετικά ασυνεπείς αλλά να είναι και οι δύο έγκυρες, καθώς μπορεί να διαχωρίζουν τα κύτταρα με βάση διαφορετικές πτυχές της βιολογίας. Επομένως επιστρέφοντας στην παρομοίωση

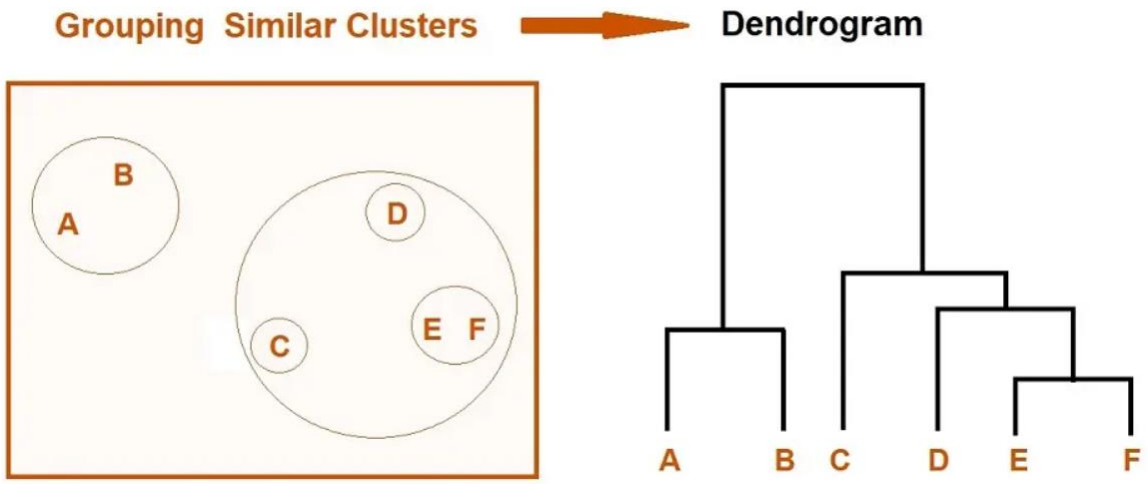
με το μικροσκόπιο, το να ζητά κανείς καλύτερη ομαδοποίηση είναι σαν να ζητά καλύτερη μεγέθυνση χωρίς όμως κάποιο πλαίσιο αναφοράς (Robert Amezquita, 2023).

2.5.1 Ιεραρχική Ομαδοποίηση (Hierarchical Clustering)

Μια ευρέως διαδεδομένη μέθοδος ομαδοποίησης είναι η ιεραρχική ομαδοποίηση. Πρόκειται για μια παλιά τεχνική που ταξινομεί τα δείγματα σε μια ιεραρχία με βάση τη σχετική τους ομοιότητα. Στην πράξη δημιουργείται ένα δενδρόγραμμα που φανερώνει ομάδες με προοδευτικά αυξανόμενη λεπτομέρεια. Αυτό γίνεται με ένωση των πιο όμοιων δειγμάτων σε μια ομάδα και έπειτα οι ομάδες που δημιουργήθηκαν ενώνονται μεταξύ τους ανάλογα με την ομοιότητά τους και δημιουργούνται μεγαλύτερες ομάδες κοκ, έως ότου στο τέλος όλα τα δείγματα να ανήκουν σε μια ενιαία ομάδα (Εικόνα 9). Κάθε μέθοδος ιεραρχικής ομαδοποίησης διαφέρει από τις υπόλοιπες όσον αφορά τον τρόπο με τον οποίο κατασκευάζουν τις ομαδοποιήσεις. Παραδείγματος χάρη η πλήρης σύνδεση στοχεύει στη συγχώνευση των ομάδων με την μεγαλύτερη απόσταση μεταξύ των στοιχείων τους. Ενώ η προσέγγιση με το όνομα Ward στοχεύει στην ελαχιστοποίηση της αύξησης διακύμανσης εντός της ομάδας (Robert Amezquita, 2023).

Η ιεραρχική ομαδοποίηση είναι μια πολύ χρήσιμη τεχνική στις αναλύσεις single cell RNA sequencing, διότι η δημιουργία του δενδρογράμματος βοηθάει στην καλύτερη συνοπτική αποτύπωση των σχέσεων μεταξύ των υποπληθυσμών. Επιπλέον παίζει καθοριστικό ρόλο στην ερμηνεία, καθώς σε ένα δενδρόγραμμα υψηλής ανάλυσης μπορούν να αναδειχθούν ομάδες που εμφωλεύουν σε εκείνες που φαίνονται σε μια χαμηλή ανάλυση. Αξιοσημείωτο είναι επίσης πως το δενδρόγραμμα αποτελεί μια φυσική αναπαράσταση των δεδομένων σε περιπτώσεις όπου τα κύτταρα έχουν προέλθει από έναν σχετικά πρόσφατο κοινό πρόγονο.

Βασικό μειονέκτημα αυτής της ομαδοποίησης αποτελεί ο χρόνος που χρειάζεται για να υλοποιηθεί. Συστήνεται επομένως να χρησιμοποιείται μόνο για μικρά σύνολα δεδομένων single cell RNA sequencing καθώς στις περισσότερες εφαρμογές της απαιτείται ένας πίνακας απόστασης (distance matrix) μεταξύ των κυττάρων, ο υπολογισμός του οποίου είναι υπολογιστικά δαπανηρός για έναν μεγάλο αριθμό κυττάρων. Επίσης η άπληστη ομαδοποίηση είναι πιθανό να οδηγήσει σε κακής ποιότητας διαμέριση στα υψηλότερα επίπεδα του δενδρογράμματος. Παρ’ όλα αυτά, η ιεραρχική ομαδοποίηση μπορεί να είναι χρήσιμη ακόμα και σε αυτήν την περίπτωση εάν συνδυαστεί με τεχνικές κβαντισμού διανυσμάτων όπως στην περίπτωση του k-means (Robert Amezquita, 2023).



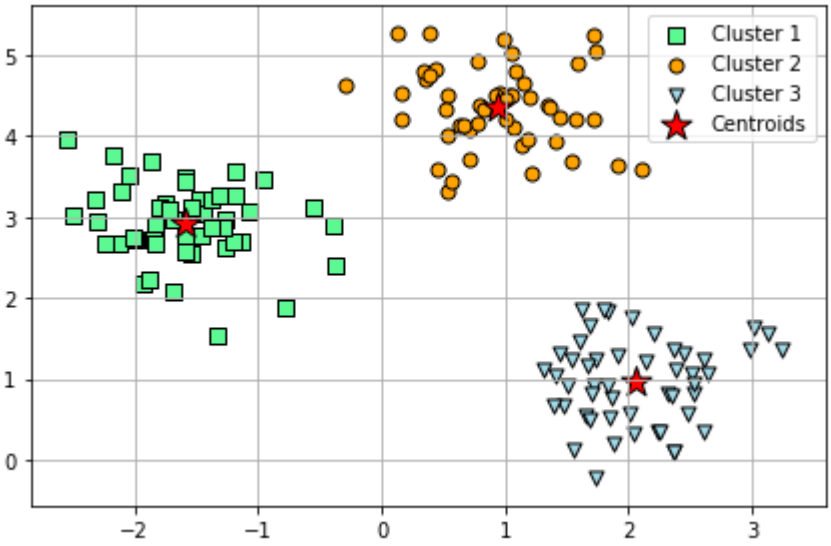
Εικόνα 9 Ιεραρχική Ομαδοποίηση κατά την οποία κατασκευάζεται το δενδρόγραμμα και από αυτό διαχωρίζονται και προκύπτουν οι ομάδες. Η εικόνα αντλήθηκε από τον διαδικτυακό ιστότοπο https://www.ejable.com/tech-corner/ai-machine-learning-and-deep-learning/hierarchical-clustering-in-machine-learning/.

2.5.2 Κβαντισμός διανυσμάτων με k-means (Vector quantization with k-means)

Μια άλλη μέθοδος ομαδοποίησης είναι ο κβαντισμός διανυσμάτων ο οποίος χωρίζει τις παρατηρήσεις σε ομάδες, οι οποίες με τη σειρά τους συνδέονται με ένα αντιπροσωπευτικό σημείο, δηλαδή ένα διάνυσμα στον χώρο των συντεταγμένων. Στην πράξη, αυτή η προσέγγιση συμπιέζει τα δεδομένα αντικαθιστώντας πολλά διαφορετικά σημεία με ένα μόνο αντιπροσωπευτικό. Τα αντιπροσωπευτικά στη συνέχεια της ανάλυσης θα αντιμετωπιστούν ως «δείγματα» μειώνοντας έτσι την υπολογιστική επεξεργασία των επόμενων βημάτων. Η προσέγγιση αυτή θα εξαλείψει τις διαφορές στην πυκνότητα των κυττάρων σε όλον τον χώρο έκφρασης, διασφαλίζοντας έτσι πως ο τύπος κυττάρων που βρίσκεται σε μεγαλύτερη αφθονία δεν θα κυριαρχήσει στα αποτελέσματα της μετέπειτα διαδικασίας (Robert Amezquita, 2023).

Η κλασική τεχνική αυτής της μεθόδου ονομάζεται k-means, πρόκειται για μια ημί – επιβλεπόμενη μέθοδο μηχανικής μάθησης (semi – supervised machine learning method) όπου χρησιμοποιεί έναν μικρό αριθμό σχολιασμένων δεδομένων με σκοπό να δοθεί μια κατεύθυνση στην ομαδοποίηση (Xin Wang, 2011). Στην πράξη κάθε κύτταρο ανατίθεται στην ομάδα με το πλησιέστερο κεντροειδές. Αυτό γίνεται με την ελαχιστοποίηση του αθροίσματος τετραγώνων εντός της ομάδας χρησιμοποιώντας μια τυχαία αρχική διαμόρφωση για το κεντροειδές (Εικόνα 10). Στις περισσότερες υλοποιήσεις συνηθίζεται να θέτουν μια μεγάλη τιμή, όπως η τετραγωνική ρίζα του αριθμού των κυττάρων, για να

παρατηρηθούν μικρές ομάδες. Με βάση τα προηγούμενα κύριο πλεονέκτημα αυτής της μεθόδου είναι η ταχύτητά της, δεδομένης της απλότητας και της ευκολίας εφαρμογής του αλγορίθμου. (Robert Amezquita, 2023)



Εικόνα 9 Ομαδοποίηση με κβαντισμό διανυσμάτων (k - means). Στην εικόνα παρατηρούνται τρεις διαφορετικές ομάδες που έχουν προκύψει με βάση τα κεντροειδή τους, τα οποία σημαίνονται με αστέρι. Η εικόνα αντλήθηκε από τον διαδικτυακό ιστότοπο https://vitalflux.com/k-means-clustering-explained-with-python-example/.

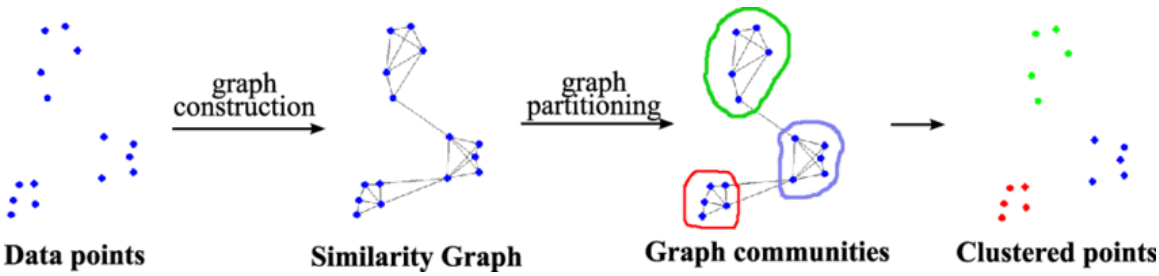
2.5.3 Ομαδοποίηση που βασίζεται σε γράφους (Graph – based Clustering)

Η ομαδοποίηση που βασίζεται σε γράφους χρησιμοποιείται συχνά σε μεγάλα σύνολα δεδομένων single cell RNA sequencing καθώς είναι μια ευέλικτη και κλιμακούμενη τεχνική. Πρώτο βήμα είναι η κατασκευή ενός γράφου όπου κάθε κόμβος αντιπροσωπεύει κάθε κύτταρο που συνδέεται με τους πλησιέστερους γείτονές του στον χώρο υψηλής διάστασης. Στη συνέχεια οι ακμές που συνδέουν τους κόμβους παίρνουν «βάρη» με βάση την ομοιότητα των εμπλεκόμενων κυττάρων. Τα μεγαλύτερα βάρη θα δίνονται στα κύτταρα που είναι πιο στενά συνδεδεμένα μεταξύ τους. Έπειτα εφαρμόζονται αλγόριθμοι για τον εντοπισμό «κοινοτήτων» κυττάρων, οι οποίες αντιπροσωπεύονται από ομάδες που μπορούν να χρησιμοποιηθούν για περαιτέρω ανάλυση και ερμηνεία (Εικόνα 11).

Το σημαντικότερο πλεονέκτημα αυτής της μεθόδου είναι η επεκτασιμότητά της, αφού απαιτεί μόνο μια αναζήτηση k κοντινότερου γείτονα που μπορεί να γίνει σε λογαριθμογραμμικό χρόνο. Σε αντίθεση με τις προσεγγίσεις ιεραρχικής ομαδοποίησης

που έχουν χρόνους εκτέλεσης τετραγωνικούς σε σχέση με τον αριθμό των κυττάρων. Η κατασκευή των γράφων αποφεύγει να κάνει ισχυρές υποθέσεις σχετικά με την κατανομή των κυττάρων μέσα σε κάθε ομάδα όπως συνήθως γίνεται στη k-means που ευνοούνται οι σφαιρικές ομάδες. Στους γράφους κάθε κύτταρο συνδέεται αναγκαστικά με έναν ελάχιστο αριθμό γειτονικών κυττάρων, γεγονός που μειώνει τον κίνδυνο δημιουργίας πολλών μη πληροφοριακών ομάδων που θα αποτελούνται μόλις από ένα ή δύο κύτταρα.

Αρνητικό των μεθόδων που βασίζονται σε γράφους αποτελεί το ότι, αφότου κατασκευαστεί ο γράφος, οποιαδήποτε πληροφορία σχετική με τις σχέσεις πέραν των γειτονικών κυττάρων ή των ομάδων χάνεται. Αυτό έχει σημαντικές απώλειες σε μελέτες που θέλουν να αναδείξουν διαφορές στην πυκνότητα των κυττάρων, καθώς απαιτούνται περισσότερα βήματα μέσω του γράφου για να διανυθεί η ίδια απόσταση σε μια περιοχή με μεγαλύτερη πυκνότητα κυττάρων. Επιπλέον αυτό το φαινόμενο επηρεάζει και τους αλγόριθμους ανίχνευσης κοινοτήτων διότι «διογκώνει» τις περιοχές υψηλής πυκνότητας, με συνέπεια οποιαδήποτε εσωτερική υποδομή ή θόρυβος να μπορεί να σχηματίσει υποομάδες. Επομένως η ανάλυση της ομαδοποίησης εξαρτάται από την πυκνότητα των κυττάρων η οποία μπορεί να είναι παραπλανητική εάν υπερτονίζει την ετερογένεια των δεδομένων (Robert Amezquita, 2023).



Εικόνα 10 Μέθοδος Ομαδοποίησης που βασίζεται σε γράφους. Η Εικόνα αντλήθηκε από το άρθρο «Graph-based data clustering via multiscale community detection» των Zijing Liu et al, DOI:10.1007/s41109-019-0248-7.

2.5.4 Αξιολόγηση της Ομαδοποίησης (Ranking Clustering)

Είναι εξαιρετικά ενδιαφέρον στην ανάλυση να ποσοτικοποιηθεί με διάφορες μετρήσεις η συμπεριφορά μιας μεθόδου ομαδοποίησης και να αξιολογηθεί με αυτόν τον τρόπο σε σύγκριση με τις υπόλοιπες μεθόδους ομαδοποίησης που θα δοκιμαστούν. Τέτοιου είδους μετρήσεις γίνονται με τεχνικές όπως το silhouette width ή το graph modularity όπου έχουν την ευελιξία να κάνουν μετρήσεις ξεχωριστά σε κάθε ομάδα έτσι ώστε να δοθεί προτεραιότητα σε ορισμένες ομάδες για πιο προσεκτική ανάλυση. Στην περίπτωση των ανεπαρκώς χωρισμένων ομάδων μπορούν αυτές να έχουν προτεραιότητα για πιο

προσεκτική ανάλυση με το χέρι, έτσι ώστε να προσδιοριστούν οι διαφορές τους με τις γειτονικές ομάδες. Παράλληλα είναι δυνατό οι ετερογενείς ομάδες να υποβληθούν σε περαιτέρω ομαδοποίηση ούτως ώστε να εντοπιστούν τυχόν εσωτερικές δομές (Robert Amezquita, 2023).

Επιπλέον μετρήσεις μπορούν να αφορούν τους διαφορετικούς αλγορίθμους και τις παραμέτρους τους, για να υπάρχει σαφέστερη εικόνα όσον αφορά για τον τρόπο με τον οποίο το σύνολο δεδομένων ανταποκρίνεται σε αλλαγές στη μέθοδο ομαδοποίησης. Αυτό θα βοηθούσε στη συνέχεια να επιλεγεί η ομαδοποίηση και οι παράμετροι εκείνες που θα βελτιστοποιούν αυτές τις μετρήσεις και συνεπώς είναι επαρκέστερες για το εκάστοτε σύνολο δεδομένων. Στην πράξη βέβαια η ομαδοποίηση εκείνη που παρέχει τον καλύτερο διαχωρισμό μεταξύ των ομάδων συχνά μπορεί να μην είναι αρκετά χρήσιμη καθώς περιγράφει προφανείς διαφορές μεταξύ των γνωστών κυτταρικών τύπων. Επομένως ιδιαίτερο ενδιαφέρον πιθανώς να έχουν ομάδες που δεν είναι καλά διαχωρισμένες, ειδικότερα εάν στόχος της μελέτης είναι να χαρακτηριστούν στενά συνδεδεμένοι υποτύποι ή καταστάσεις. Συμπερασματικά αυτές οι μετρήσεις αξιολόγησης των ομαδοποιήσεων χρησιμοποιούνται καλύτερα για την καθοδήγηση της ερμηνείας των αποτελεσμάτων, παρά για να φιλτράρουν εντελώς της «κακές» ομάδες. (Robert Amezquita, 2023)

2.5.4.1 Silhouette Width

Η πιο καθιερωμένη μέτρηση για τον αξιολόγηση του διαχωρισμού των ομάδων γίνεται με τα διαγράμματα που αποτυπώνουν το silhouette width. Εδώ, για κάθε κύτταρο υπολογίζεται η μέση απόσταση από όλα τα κύτταρα της ίδιας ομάδας. Επίσης υπολογίζεται η μέση απόσταση από όλα τα κύτταρα σε μια άλλη ομάδα, λαμβάνεται το ελάχιστο των μέσων όρων σε όλες τις άλλες ομάδες και τελικά το silhouette width ορίζεται ως η διαφορά μεταξύ αυτών των δύο τιμών διαιρούμενη με το μέγιστό τους. Στην πράξη τα κύτταρα με μεγάλα θετικά silhouette width είναι πιο κοντά σε άλλα κύτταρα της ίδιας ομάδας από ότι στις υπόλοιπες ομάδες. Άρα και ομάδες με μεγάλα θετικά silhouette width θεωρούνται καλά διαχωρισμένες.

Το silhouette width είναι ένα φυσικό διαγνωστικό στοιχείο στις προσεγγίσεις ιεραρχικής ομαδοποίησης μιας και υπάρχει ήδη διαθέσιμος ο πίνακας αποστάσεων (distance matrix). Ωστόσο όταν πρόκειται για πολύ μεγάλα σύνολα δεδομένων προτιμάται μια προσεγγιστική μέθοδος που χρησιμοποιεί τη ρίζα των μέσων τετραγωνικών

αποστάσεων, αντί της ίδιας της μέσης απόστασης, για να αποφεύγεται ο χρονοβόρος υπολογισμός των αποστάσεων ανά ζεύγη.

Μεγάλο πλεονέκτημα αυτής της μεθόδου είναι πως αποτυπώνει πληροφορία τόσο για το underclustering όσο και για το overclustering. Αυτό μπορεί να το εκμεταλλευτεί ο ερευνητής για να πάρει μια βολική αρχική επιλογή για τον αριθμό των ομάδων μεγιστοποιώντας έτσι το silhouette width (Robert Amezquita, 2023).

2.5.4.2 Cluster Purity

Ένας διαφορετικός τρόπος για την αξιολόγηση του διαχωρισμού των ομάδων είναι ο βαθμός στον οποίο κύτταρα από πολλές ομάδες αναμειγνύονται στον χώρο έκφρασης. Η «καθαρότητα της ομάδας» (cluster purity) ορίζεται για κάθε κύτταρο ως το ποσοστό των γειτονικών κυττάρων που βρίσκονται στην ίδια ομάδα, μετά από κάποια στάθμιση για την προσαρμογή των διαφορών στον αριθμό των κυττάρων μεταξύ των ομάδων. Επομένως αναμένεται να εμφανιστούν υψηλές τιμές καθαρότητας στα κύτταρα μέλη των καλά διαχωρισμένων ομάδων, αφού θα έχουν μικρή ανάμειξη μεταξύ των κυττάρων.

Η σύγκριση μεταξύ των δύο μεθόδων αξιολόγησης της ομαδοποίησης που αναφέρθηκαν μπορεί να είναι ωφέλιμη στην ανάδειξη των ετερογενών ομάδων, των καλά διαχωρισμένων ή και των δύο. Κύρια διαφορά τους αποτελεί το ότι το silhouette width αγνοεί τη διακύμανση εντός των ομάδων. Με αυτόν τον τρόπο παρέχει μια απλούστερη ερμηνεία του διαχωρισμού των ομάδων, ένα χαμηλό silhouette width είναι πιθανό να φανεί ακόμα και σε καλά διαχωρισμένες ομάδες εάν παρουσιάζουν υψηλή εσωτερική ετερογένεια, ενώ κάτι τέτοιο δεν θα μπορούσε να συμβεί με το cluster purity.

2.5.4.3 Jaccard Index

Ο δείκτης Jaccard είναι ένας ακόμα τρόπος αξιολόγησης της ομαδοποίησης, κατά τον οποίο μετριέται η αναλογία των αληθή θετικών (True Positives, TP) σε σχέση με τα αληθή θετικά, συν τα ψευδώς θετικά (False Positives, FP), συν τα ψευδώς αρνητικά (False Negatives).

Δεδομένου ότι αυτή η απλή μέτρηση εξετάζει την αναλογία όλων των σωστών προβλέψεων ως προς τα αληθή θετικά και όλες τις λανθασμένες προβλέψεις, μια υψηλότερη αναλογία υποδεικνύει μεγάλη ομοιότητα μεταξύ των τιμών ανάμεσα στις δυο υπό εξέταση ομάδες. Όσο υψηλότερος είναι ο δείκτης Jaccard τόσο καλύτερο είναι και το μοντέλο ομαδοποίησης.

Όμως είναι εύκολο να υπάρξουν παρερμηνείες καθώς ο δείκτης Jaccard μπορεί να μετρήσει επαρκώς την ομοιότητα μεταξύ δυο ομάδων, αλλά είναι ευάλωτος σε πιθανές παρερμηνείες της σχέσης μεταξύ δύο ομάδων. Δηλαδή εάν μια ομάδα περιέχονταν πλήρως μέσα σε μια άλλη, θα μπορούσε μέσω της μέτρησης να φαίνεται ότι διαχωρίζεται ξεκάθαρα σε δύο ξεχωριστές ομάδες (Evaluation Metrics for Machine Learning Models, 2022).

2.6 Σχολιασμός των κυτταρικών τύπων (Annotation)

Αφότου έχουν προηγηθεί οι υπολογιστικές μέθοδοι που έχουν καθαρίσει/φιλτράρει τα δεδομένα και αυτά έχουν διαιρεθεί σε ομάδες, έρχεται η ώρα για τον προσδιορισμό των κυτταρικών τύπων στους οποίους ανήκουν αυτές οι ομάδες. Σε αυτό το σημείο γεφυρώνεται το χάσμα μεταξύ των επιστημών της πληροφορικής και της βιολογίας (Robert Amezquita, 2023).

Η ερμηνεία των αποτελεσμάτων, είναι και η πιο απαιτητική διαδικασία κατά την ανάλυση δεδομένων μεμονωμένων κυττάρων, καθώς ο σχολιασμός εξαρτάται άμεσα από τα δεδομένα. Συνεπώς η επιστημονική κοινότητα δεν έχει κάποια πολύ συγκεκριμένη και τυποποιημένη ροή εργασίας για αυτό το στάδιο της ανάλυσης (Tran, 2022).

2.6.1 Εμπόδια κατά τον σχολιασμό των κυτταρικών τύπων

Τροχοπέδη στον σχολιασμό των δεδομένων αποτελεί η ίδια η φύση των βιολογικών δεδομένων, καθώς εδώ τίθεται το ερώτημα ως προς ποιο επίπεδο θα γίνει η διάκριση των κυτταρικών τύπων.

Δεν υπάρχει ομοφωνία στην επιστημονική κοινότητα όταν πρόκειται για τον ορισμό του «κυτταρικού τύπου» και αυτό έγκειται στο γεγονός ότι ένα κύτταρο με βάση διαφορετικά κριτήρια όπως ο ιστός ή ο μοριακός τύπος του, μπορεί να κατηγοριοποιηθεί και σε διαφορετικά επίπεδα. Με άλλα λόγια, ανάλογα με το βάθος και το πόσο συγκεκριμένα επιθυμεί ο ερευνητής να διαιρέσει τους κυτταρικούς πληθυσμούς, ο ορισμός του «κυτταρικού τύπου» ποικίλλει. Για παράδειγμα τα κύτταρα του ανοσοποιητικού συστήματος μπορούν να διαχωριστούν περαιτέρω σε T cells, B cells, natural killer cells κλπ. Και έπειτα κάθε ομάδα περιλαμβάνει δικούς της κυτταρικούς υποτύπους όπως τα T cells, που με βάση κάποιους δείκτες ή κάποιες ειδικές λειτουργίες χαρακτηρίζονται διαφορετικά σε επιπλέον υποτύπους όπως οι CD4+ T cells, οι CD8+ T cells, οι T regulatory κλπ. (Tran, 2022)

Ένας ακόμα λόγος που ο «κυτταρικός τύπος» δεν είναι σαφώς ορισμένος, είναι η ελλιπής κατανόηση της ταυτότητας των κυττάρων. Όπως αναφέρθηκε και προηγουμένως, αν εμβαθύνουμε περισσότερο στην κυτταρική ετερογένεια, ανακαλύπτονται νέοι κυτταρικοί υποτύποι. Συχνά βέβαια τα μεταγραφικά προφίλ αυτών των νέων υποτύπων δεν διαφέρουν τόσο με τους αρχικούς, ώστε να χρειαστεί να γίνει κάποια νέα οριοθέτηση. Όμως υπάρχουν και περιπτώσεις όπου αυτή η νέα οριοθέτηση είναι χρήσιμη καθώς βρίσκονται κύτταρα σε μια διαβάθμιση καταστάσεων και αυτά πρέπει να χαρακτηριστούν/σχολιαστούν διαφορετικά. Τέτοιου είδους περιπτώσεις συναντώνται συχνά στους αναπτυσσόμενους ιστούς (Tran, 2022).

2.6.2 Συσχετισμός της Ομαδοποίησης και του Σχολιασμού των κυτταρικών τύπων

Η Ομαδοποίηση και ο Σχολιασμός των «κυτταρικών τύπων» είναι δύο διαδικασίες αλληλένδετα συνδεδεμένες μεταξύ τους.

Το ερώτημα που συζητήθηκε παραπάνω είναι πολύ σημαντικό να έχει προηγηθεί και κατά τη διαδικασία της Ομαδοποίησης έτσι ώστε να υπάρχει μια πρόχειρη εκτίμηση για τον σωστό αριθμό των ομάδων. Ο ερευνητής θα πρέπει να έχει αποφασίσει ποιο επίπεδο της «κυτταρικής ταυτότητας» επιθυμεί να μελετήσει.

Στα πολύ ετερογενή σύνολα δεδομένων είναι πιθανό να προκύψουν και υπερβολικά πολλές ομάδες, οι οποίες θα κάνουν τον σχολιασμό μια ακόμα πιο περίπλοκη διαδικασία. Για αυτό κατά την Ομαδοποίηση ίσως είναι πιο ωφέλιμο να μην γίνεται υπερβολικός διαχωρισμός των ομάδων. Για καλύτερα αποτελέσματα προτείνεται η υπό-ομαδοποίηση, κατά την οποία οι χρήστες για αρχή θα ομαδοποιούν και θα σχολιάζουν σε ένα μέτριο επίπεδο και στη συνέχεια θα αφαιρέσουν κάποιες περισσότερο ενδιαφέρουσες ομάδες, τις οποίες και θα αντιμετωπίσουν ως ένα νέο σύνολο δεδομένων. Με αυτόν τον τρόπο θα αυξηθεί η εστίαση και η ανάλυση στις ομάδες που ενδιαφέρουν την εκάστοτε μελέτη και οι ενδιαφέροντες υποτύποι κυττάρων μπορούν να διερευνηθούν ευκολότερα (Tran, 2022).

2.6.3 Αναλογικός σχολιασμός των κυτταρικών τύπων

Ένας πολύ διαδεδομένος τρόπος για τον χαρακτηρισμό των κυτταρικών τύπων είναι ο αναλογικός σχολιασμός τους με το χέρι. Με αυτόν τον τρόπο η διαδικασία γίνεται εξαιρετικά ευέλικτη και προσαρμόζεται στο εκάστοτε σύνολο δεδομένων. Πρόκειται όμως

για μια πολύ χρονοβόρα τεχνική που απαιτεί ισχυρή βιολογική κατανόηση και είναι επιρρεπής σε μεροληψίες, καθώς εξαρτάται αποκλειστικά και μόνο από τους ερευνητές.

Βασικό πρώτο βήμα για τον αναλογικό προσδιορισμό αποτελεί η ταυτοποίηση των κυττάρων, μέσω κάποιων γονιδίων δεικτών μια διαδικασία που ονομάζεται marker gene detection. Διαφορετικά για τον σκοπό αυτόν μπορεί να είναι πολύ χρήσιμη και η ανάλυση της διαφορικής έκφρασης των γονιδίων.

Στην απλούστερη περίπτωση όταν τα περισσότερα γονίδια σε έναν κυτταρικό πληθυσμό υπερεκφράζουν τα κανονικά γονίδια δεικτών για έναν συγκεκριμένο κυτταρικό πληθυσμό, ο σχολιασμός γίνεται πολύ εύκολα. Όμως τις περισσότερες φορές τα κλασσικά γονίδια δείκτες δεν είναι αρκετά για την ταυτοποίηση, διότι υπάρχουν πολλοί σπάνιοι κυτταρικοί τύποι που πρέπει να εξεταστούν. Ακόμα υπάρχουν αναλύσεις όπου βρίσκονται ελάχιστα ή καθόλου γνωστά γονίδια δεικτών, σε αυτήν την περίπτωση θα χρειαζόταν μια ανάλυση εμπλουτισμού για να βρεθούν οι λειτουργίες που επιτελούν τα κύτταρα σε αυτές τις «άγνωστες» ομάδες. Τέτοιου είδους αναλύσεις μπορεί να έχουν προκύψει από σύνολα που περιέχουν κύτταρα «κακής ποιότητας» ή κύτταρα που ανήκουν σε κάποιον νέο κυτταρικό τύπο (Tran, 2022).

Τα γονίδια δείκτες που αναφέρονται εδώ μπορούν να βρεθούν σε άτλαντες μεμονωμένων κυττάρων. Καλή τεχνική είναι το δείγμα να προέρχεται από παρόμοιο βιολογικό πλαίσιο, δηλαδή ίδιο οργανισμό ή ιστό ή ασθένεια. Πολύ σημαντική είναι και η αναζήτηση της κατάλληλης βιβλιογραφίας, ειδικά εάν πρόκειται για σύνολο δεδομένων που περιλαμβάνει κυτταρικούς τύπους που δεν έχουν διερευνηθεί ιδιαίτερα στο παρελθόν.

2.6.4 Αυτόματος σχολιασμός των κυτταρικών τύπων

Καθώς ο σχολιασμός των κυτταρικών τύπων αναλογικά φαίνεται να είναι μια διαδικασία χρονοβόρα και με χαμηλή αξιοπιστία, οι επιστήμονες στράφηκαν σε αυτοματοποιημένες λύσεις με τη βοήθεια της Βιοπληροφορικής.

Τα τελευταία χρόνια με την ανάπτυξη της τεχνολογίας αλληλούχισης μεμονωμένων κυττάρων έχει γίνει διαθέσιμος ένας πολύ μεγάλος όγκος δεδομένων που μπορεί να χρησιμοποιηθεί για την επιβλεπόμενη εκπαίδευση μοντέλων μηχανικής μάθησης (supervised machine learning models), τα οποία θα σχολιάζουν/χαρακτηρίζουν νέα σύνολα δεδομένων (Xiangling Ji, 2023).

Απλουστευμένα λοιπόν κατά τον αυτόματο σχολιασμό των κυτταρικών τύπων γίνεται σύγκριση των προφίλ έκφρασης ενός κυττάρου με σύνολα δεδομένων αναφοράς, που έχουν ήδη σχολιαστεί στο παρελθόν. Με αυτόν τον τρόπο μπορούν να αποδοθούν ετικέτες σε κάθε κύτταρο με βάση το πιο παρόμοιο με αυτό κύτταρο από το σύνολο αναφοράς που έχει επιλεγεί. (Robert Amezquita, 2023).

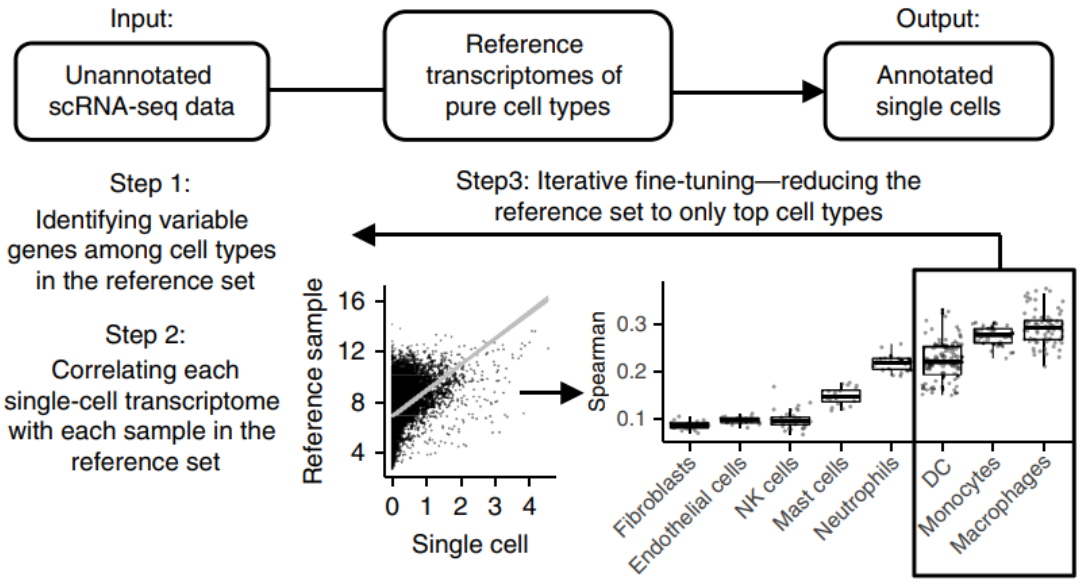
Οι αυτοματοποιημένες αυτές μέθοδοι μπορούν να εξασφαλίσουν ότι ο σχολιασμός των κυττάρων μπορεί να είναι αναπαραγώγιμος και αμερόληπτος. Όμως έχουν και κάποιους περιορισμούς, όπως είναι η απαίτηση μεγάλων συνόλων αναφοράς, τα οποία κατά συνέπεια αυξάνουν και τον υπολογιστικό χρόνο. Επίσης φαίνεται να μειονεκτούν στον εντοπισμό των καρκινικών κυττάρων, καθώς συχνά αποτυγχάνουν να τα διακρίνουν από τα φυσιολογικά (Bassel Ghaddar, 2023).

2.6.4.1 Το εργαλείο αυτόματου σχολιασμού SingleR

Στην παρούσα εργασία για τον σχολιασμό των κυτταρικών τύπων επιλέχθηκε το εργαλείο SingleR. Αυτή η μέθοδος δίνει ετικέτες στα κύτταρα βασισμένη σε ένα σύνολο δεδομένων αναφοράς, κάνοντας αντιστοίχιση με τις υψηλότερες βαθμολογίες του συντελεστή συσχέτισης κατά Spearman, χρησιμοποιώντας μόνο γονίδια δείκτες μεταξύ των ζευγών ετικετών για να εστιάσει στις σχετικές διαφορές μεταξύ των κυτταρικών τύπων (Dvir Aran, 2019).

Αυτά τα σύνολα δεδομένων αναφοράς προέρχονται από μια βιβλιοθήκη του πακέτου Bioconductor που ονομάζεται celldex και περιλαμβάνει συλλογές από επισημασμένους κυτταρικούς τύπους, που χρησιμοποιούνται για τον αυτόματο σχολιασμό στις αναλύσεις μεμονωμένων κυττάρων αλλά και για τα δεδομένα που έχουν προκύψει από τη μαζική αλληλούχιση (Dvir Aran, 2019).

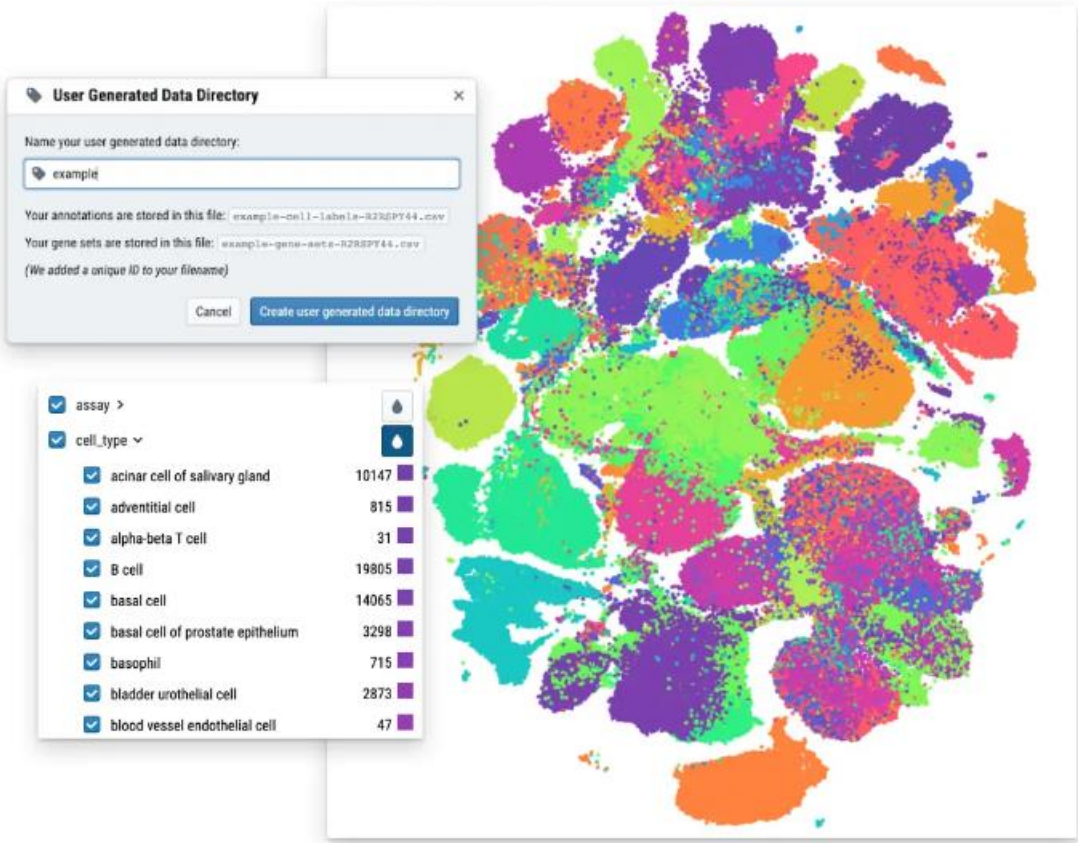
Επιπλέον, πραγματοποιεί μικρές αλλαγές σε κάθε κύτταρο για να τελειοποιήσει τις συσχετίσεις, οι οποίες υπολογίζονται εκ νέου μόνο με τα γονίδια δείκτες για τις ετικέτες με την υψηλότερη βαθμολογία (Εικόνα 12). Αυτό το βήμα αποσκοπεί στην απομάκρυνση του θορύβου και συνεπώς την εξάλειψη οποιασδήποτε ασάφειας μεταξύ των ετικετών (Dvir Aran, 2019).



Εικόνα 11 Σχηματική αναπαράσταση του αλγορίθμου του SingleR για τον σχολιασμό των κυτταρικών τύπων. Η εικόνα αντλήθηκε από το άρθρο «Reference – based analysis of lung single – cell sequencing reveals a transitional profibrotic macrophage» των Dvir Aran et al.

2.7 Cell by Gene

Το Cell by Gene είναι μία εφαρμογή οπτικοποίησης των κυτταρικών ατλάντων και αποτελείται από ένα σύνολο εργαλείων που εξυπηρετούν στην επιστημονική έρευνα. Οι ερευνητές μπορούν να βρίσκουν, να κατεβάζουν, να διερευνούν, να σχολιάζουν αλλά και να δημοσιεύουν σύνολα δεδομένων μεμονωμένων κυττάρων. Στην πράξη πρόκειται για έναν διαδραστικό περιηγητή δεδομένων (Εικόνα 13) ο οποίος βοηθάει στην οπτικοποίηση των δεδομένων μεμονωμένων κυττάρων και στην καλύτερη αναπαράστασή τους έτσι ώστε να είναι ευκολότερη η ερμηνεία των αποτελεσμάτων και η εξαγωγή συμπερασμάτων. Τα συμπεράσματα αυτά μπορεί να αφορούν γονίδια – δείκτες, κυτταρική προέλευση και σχετικά σύνολα δεδομένων, καθώς η πλατφόρμα συγκεντρώνει πάνω από επτακόσιους κυτταρικούς τύπους. Μεταξύ άλλων προσφέρει οπτικοποίηση της γονιδιακής έκφρασης, κατανόηση των δεδομένων μέσω διαδραστικών αναλύσεων και τέλος λειτουργεί ως εφαλτήριο για τον εντοπισμό νέων υποτύπων και καταστάσεων κυττάρων (CZ Cell x Gene, n.d.).



Εικόνα 12 Η Εικόνα αντλήθηκε από τον διαδικτυακό ιστότοπο του Cell x Gene (https://cellxgene.cziscience.com/)

2.8 scRAFIKI (ex SIMBA)

Το λογισμικό scRAFIKI (πρώην όνομα SIMBA) έχει σχεδιαστεί από μέλη της ομάδας Cost Action CA20117 με σκοπό τη ενσωμάτωση των δεδομένων που έχουν προκύψει από τις αναλύσεις μεμονωμένων κυττάρων στο cell by gene για την οπτικοποίηση και την περαιτέρω ανάλυση των δεδομένων. Επιπλέον με αυτό μπορούν να γίνουν διαδικασίες φιλτραρίσματος των δεδομένων, αλλά και απόδοσης ετικετών (github.com, 2024).

Για τη δημιουργία αυτού του λογισμικού οι ερευνητές βασίστηκαν στο άρθρο των Salcher et al. με τίτλο «High resolution single cell atlas reveals diversity and plasticity of tissue resident neutrophils in non-small cell lung cancer», το οποίο επικεντρώνεται στη δημιουργία ενός άτλαντα δεδομένων μεμονωμένων κυττάρων που αποσκοπεί στην αποκάλυψη της ποικιλότητας των ουδετερόφιλων που κατοικούν στους ιστούς και στον μη μικροκυτταρικό καρκίνο του πνεύμονα (NSCLC) (Stefan Salcher, 2022).

3 Δεδομένα

3.1 Σύνολο δεδομένων Lun 416B cell line

Το πρώτο σύνολο δεδομένων που αναλύθηκε στην παρούσα εργασία προέρχεται από το άρθρο των Aaron T. L. Lun et al. με τίτλο «Assessing the reliability of spike in normalization for analyses of single cell RNA sequencing data». Στο άρθρο αξιολογείται η αποτελεσματικότητα των spike in ελέγχων, οι οποίοι χρησιμοποιούνται σε πειράματα μεμονωμένων κυττάρων (single cell RNA sequencing), για την κανονικοποίηση δεδομένων μεταξύ διαφορετικών κυττάρων και πειραματικών συνθηκών. Οι έλεγχοι αυτοί είναι μεγάλης σημασίας για αυτές τις μελέτες καθώς μπορούν να επηρεάσουν όλα τα βήματα της ανάλυσης και κατ' επέκταση την ακρίβειά της.

Το σύνολο δεδομένων που χρησιμοποιήθηκε έχει τον κωδικό E-MTAB-5522 τη βάση βιολογικών δεδομένων EMBL – EBI και περιλαμβάνει δεδομένα μεμονωμένων κυττάρων στα οποία προστέθηκαν συνθετικά RNA «spike ins» σε κάθε δείγμα ως ένα μέσο για έλεγχο. Στην πράξη αυτά τα spike ins αποτελούν σημεία αναφοράς με βάση τα οποία μπορούν οι ερευνητές να αξιολογήσουν τη συνέπεια της ποσοτικοποίησης του RNA μεταξύ των δειγμάτων (Aaron T.L. Lun F. J.-N.-V., 2017).

3.2 Σύνολο Δεδομένων GSE212797

Το σύνολο δεδομένων GSE212797 αντλήθηκε από το άρθρο των Maha K. Rahim et al. με τίτλο «Dynamic CD8+ T cell responses to cancer immunotherapy in human regional lymph nodes are disrupted in metastatic lymph nodes». Οι ερευνητές σε αυτό το άρθρο εξετάζουν τις αποκρίσεις των CD8+ T κυττάρων σε ανοσοθεραπείες καρκίνου και πιο συγκεκριμένα σε ασθενείς με καρκίνο στους περιφερειακούς λεμφαδένες και διαπιστώνουν ότι οι αποκρίσεις αυτές επηρεάζονται αρνητικά από την παρουσία μεταστάσεων. Καταλήγουν στο συμπέρασμα ότι τα CD8+ T κύτταρα στους λεμφαδένες παίζουν εξαιρετικά σημαντικό ρόλο στη μάχη ενάντια στον καρκίνο, αλλά η παρουσία καρκινικών κυττάρων που έχουν εξαπλωθεί από τις διάφορες μεταστάσεις παρεμποδίζουν τη φυσιολογική τους λειτουργία. Τέλος η μελέτη αυτή εμπλουτίζει τις πληροφορίες που έχουμε σχετικά με το πως οι μεταστάσεις διαταράσσουν τις ανοσολογικές αποκρίσεις, βλάπτοντας με αυτόν τον τρόπο την αποτελεσματικότητα των ανοσοθεραπειών (Maha K. Rahim, 2023).

3.3 Σύνολο Δεδομένων GSE210963

Το τρίτο και τελευταίο σύνολο δεδομένων που χρησιμοποιήθηκε στην παρούσα εργασία μελετήθηκε στο άρθρο των Himanshu Savardekar et al. με τίτλο «Single cell RNA seq analysis of patient myeloid derived suppressor cells and the response to inhibition of Bruton's tyrosine kinase (BTK)». Οι ερευνητές χρησιμοποίησαν την τεχνολογία αλληλούχισης μεμονωμένων κυττάρων για να αναλύσουν Μυελοειδή Κατασταλτικά κύτταρα (MDSCs) ασθενών με καρκίνο. Η κατανόηση της λειτουργίας αυτών των κυττάρων φαίνεται να είναι πολύ κρίσιμη για τη βελτίωση της ανοσοθεραπείας, καθώς αυτού του είδους τα κύτταρα όταν βρίσκονται σε περιβάλλοντα όγκων καταστέλλουν τις ανοσολογικές αποκρίσεις. Πιο αναλυτικά η μελέτη αυτή εστιάζει στην επίδραση της αναστολής της κινάσης Bruton, μιας πρωτεΐνης που εμπλέκεται σε σηματοδοτικά μονοπάτια που επηρεάζουν την ανάπτυξη και τη λειτουργία των MDSCs. Η ανάλυση των μεμονωμένων κυττάρων σε αυτήν την περίπτωση βοήθησε στην λεπτομερή χαρτογράφηση της γονιδιακής έκφρασης στα MDSCs.

Τα αποτελέσματα της έρευνας αυτής υποδεικνύουν ότι η αναστολή της BTK μειώνει την κατασταλτική δράση των MDSCs, καθιστώντας τα λιγότερο ικανά να καταστείλουν τις αντικαρκινικές ανοσολογικές αποκρίσεις. Το γεγονός αυτό έχει άμεσες και σημαντικές συνέπειες στην ανάπτυξη νέων στρατηγικών ανοσοθεραπείας, καθώς η BTK προτείνεται ως ένας πιθανός θεραπευτικός στόχος που ενδεχομένως θα ενισχύσει την αποτελεσματικότητα των ανοσοθεραπειών κατά του καρκίνου (Himanshu Savardekar, 2024).

4 Αποτελέσματα

Στην παρούσα εργασία τα βήματα που αναλύθηκαν παραπάνω ήταν αναγκαία για την προεπεξεργασία και την ανάλυση ορισμένων συνόλων δεδομένων στα πλαίσια συνεργασίας με την Ευρωπαϊκή κοινοπραξία Cost Action CA20117.

4.1 Επαν-ανάλυση δημόσιων δεδομένων

Για τις ανάγκες της διερεύνησης των καλύτερων και πιο αξιόπιστων μεθόδων για την ανάλυση δεδομένων που έχουν προκύψει από την αλληλούχιση μεμονωμένων κυττάρων, πραγματοποιήθηκε μια επαν-ανάλυση των δημόσιων δεδομένων που έχουν δημοσιοποιηθεί στο άρθρο «Assessing the reliability of spike – in normalization for

analyses of single – cell RNA sequencing data» των Lun A. T. L. et al, όπου μελετάται η κυτταρική ετερογένεια σε κύτταρα 416B (μια σειρά μυελοειδών προγονικών κυττάρων ποντικού) (Lun A. T.-N.-V., 2017).

Η ανάλυση πραγματοποιήθηκε στο RStudio όπου τα πρωτογενή δεδομένα (raw data) φορτώθηκαν σε αντικείμενα τύπου SingleCellExperiment για να ακολουθήσει η προεπεξεργασία τους.

4.1.1 Έλεγχος ποιότητας και διαγνωστικά διαγράμματα του συνόλου δεδομένων 416 B.

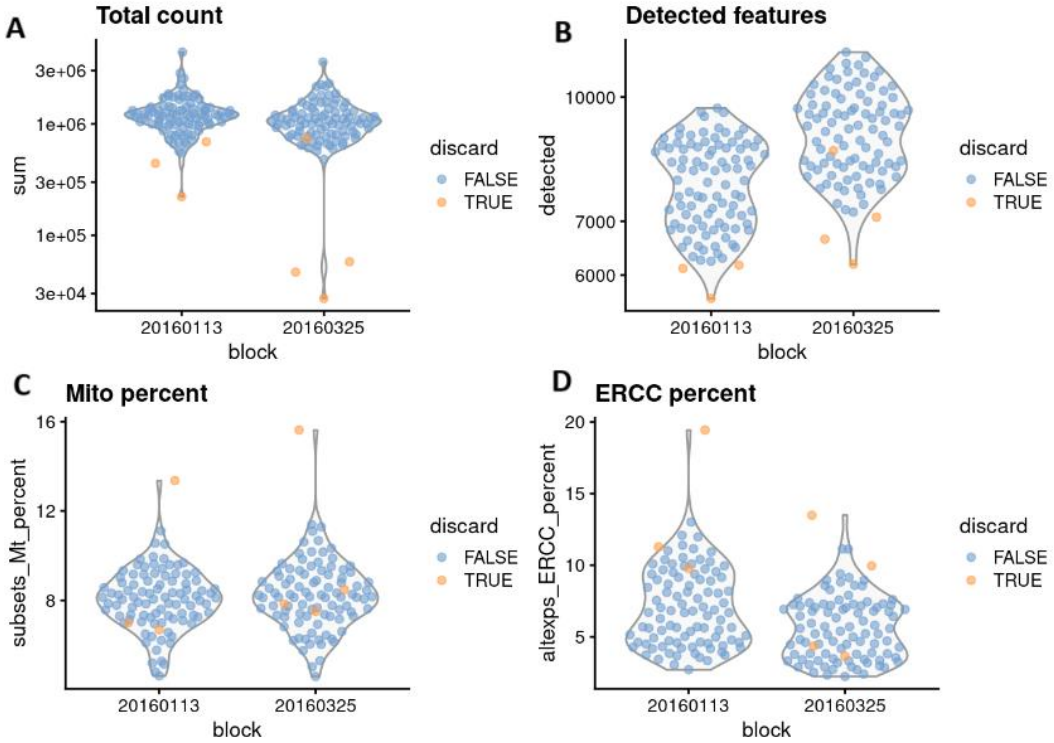
Πριν από την ανάλυση των δεδομένων πρέπει να διασφαλιστεί ότι όλα τα κύτταρα που έχουν επισημανθεί με barcodes αντιστοιχούν σε βιώσιμα κύτταρα. Αυτό εξασφαλίζεται από τον έλεγχο ποιότητας με βάση τρεις συνδιακυμάνσεις: τον αριθμό των μετρήσεων ανά barcode, τον αριθμό των γονιδίων ανά barcode και το κλάσμα των μετρήσεων από μιτοχονδριακά γονίδια ανά barcode. Οι κατανομές που θα προκύψουν από αυτές τις συνδιακυμάνσεις, θα εξεταστούν στη συνέχεια για ακραίες κορυφές, έτσι ώστε τα κύτταρα που τις αποτελούν να απορριφθούν από τη μελέτη με τη χρήση των κατάλληλων κατωφλιών (Theis, 2019).

Σε αυτήν την επαν-ανάλυση χρησιμοποιήθηκαν γνωστές συναρτήσεις από το πακέτο του Bioconductor. Αρχικά η συνάρτηση perCellQCMetrics του πακέτου scuttle, η οποία υπολογίζει και εντοπίζει χρήσιμες παραμέτρους για την αφαίρεση των προβληματικών κυττάρων. Οι παράμετροι αυτές αφορούν το μέγεθος της βιβλιοθήκης και τον αριθμό των εντοπισμένων γονιδίων, το ποσοστό των οποίων αποτελεί ένα ενδεικτικό μέτρο της πολυπλοκότητας της βιβλιοθήκης (Lun A. , Per - cell quality control metrics, 2024). Επίσης ακολούθησε η εφαρμογή της συνάρτησης quickPerCellQC η οποία αναγνωρίζει τα κύτταρα χαμηλής ποιότητας (Lun A. , quickPerCellQC: Quick cell level QC, 2024).

Είναι πολύ σημαντικό να αποτυπώνονται οι παραπάνω διαδικασίες με κάποια διαγνωστικά διαγράμματα που βοηθάνε στο να επιβεβαιωθεί ότι έχουν γίνει όσο το δυνατόν με μεγαλύτερη ακρίβεια. Όπως φαίνεται στην Εικόνα 14 το δείγμα που εξετάζεται αποτελείται από δύο πλάκες (plate based method) με ονόματα 20160113 και 20160325, με κάθε σημείο να αντιπροσωπεύει ένα κύτταρο και να είναι χρωματισμένο ανάλογα με το εάν αυτό το κύτταρο έχει αποκλειστεί ή όχι (με πορτοκαλί χρώμα σηματοδοτούνται τα κύτταρα που έχουν αποκλειστεί από τη μελέτη).
Στην Εικόνα 14 Α παρουσιάζεται η συνδιακύμανση του συνολικού αριθμού των

μετρήσεων ανά barcode, στην Εικόνα 14 Β βλέπουμε την συνδιακύμανση των γονιδίων που έχουν εντοπιστεί και τέλος στις Εικόνες 14 C και D αποτυπώνονται οι μιτοχονδριακές αναλογίες και οι αναλογίες των spike – ins αντίστοιχα.

Στην περίπτωση που υπήρχαν πολλά κύτταρα που έχουν αποκλειστεί σε μια συγκεκριμένη περιοχή αυτό θα μπορούσε να αποτελεί ισχυρή ένδειξη συσχετισμού αυτών με κάποια βιολογική κατάσταση την οποία αγνοούμε. Διαφορετικά θα μπορούσε να είναι και δείγμα ασυνεπειών κατά τη διαδικασία προετοιμασίας της βιβλιοθήκης για ένα υποσύνολο κυττάρων, γεγονός που δεν σπανίζει σε πρωτόκολλα που βασίζονται σε πλάκες (Amezquita, 2019).



Εικόνα 14 Οι κατανομές όλων των μετρήσεων του ελέγχου ποιότητας σε όλα τα κύτταρα που συνόλου δεδομένων 416B, χωρισμένα ανάλογα με την πλάκα από την οποία προήλθαν. Κάθε σημείο αντιπροσωπεύει ένα κύτταρο και είναι χρωματισμένο ανάλογα με το εάν έχει απορριφθεί από τη μελέτη. Η Εικόνα δημιουργήθηκε στο RStudio έπειτα από την επαν-ανάλυση του συνόλου δεδομένων Lun 416B cell line.

Επιπλέον, κοιτώντας πιο βαθιά είναι δυνατό να γνωρίζουμε τους λόγους απόρριψης των κυττάρων από την εν λόγω μελέτη. Στην Εικόνα 15 εμφανίζονται συνολικά τα αποτελέσματα των κυττάρων που απορρίφθηκαν με πέντε από αυτά να αφαιρέθηκαν εξ' αιτίας χαμηλής ποιότητας βιβλιοθήκης και δύο λόγω παρουσίας υψηλών ποσοστών μιτοχονδριακών γονιδίων ή spike – in μεταγράφων (μιτοχονδριακές αναλογίες και spike –

in μετάγραφα εξυπηρετούν τον ίδιο σκοπό στη συγκεκριμένη μελέτη, δηλαδή την ύπαρξη ενός συνόλου αναφοράς).

```
##              low_lib_size            low_n_features   high_subsets_Mt_percent
##                         5                         0                         2
## high_altexps_ERCC_percent                   discard
##                         2                         7
```

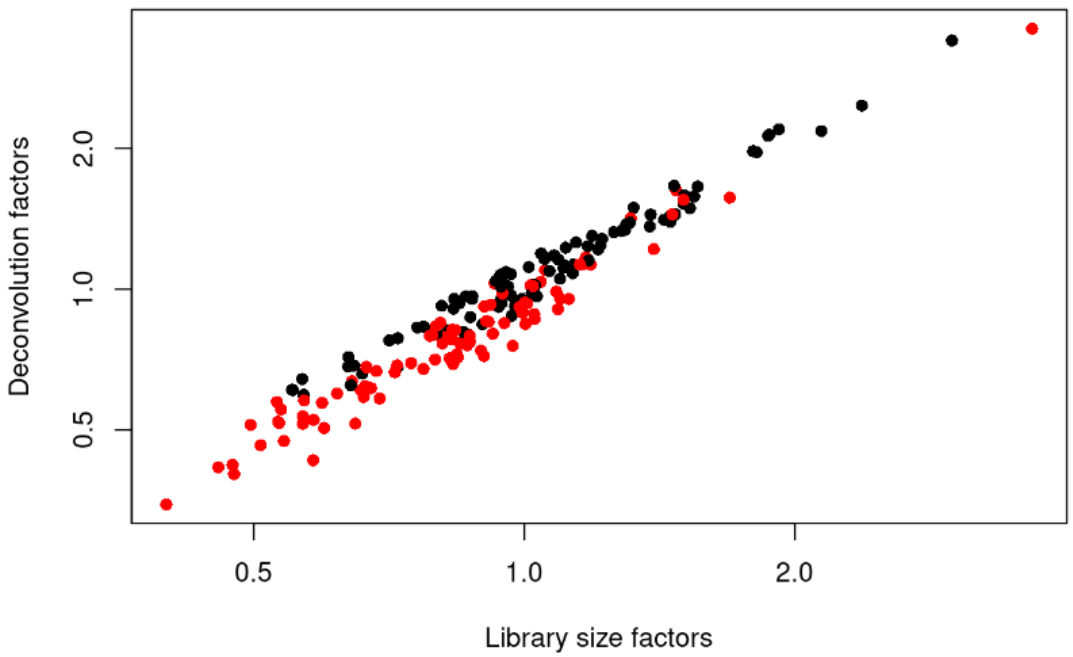
Εικόνα 13 Αιτίες απόρριψης κυττάρων από τη μελέτη. Η Εικόνα δημιουργήθηκε στο RStudio έπειτα από την επαν-ανάλυση του συνόλου δεδομένων Lun 416B cell line.

4.1.2 Κανονικοποίηση συνόλου δεδομένων 416B

Κατά την κανονικοποίηση του συνόλου δεδομένων 426B χρησιμοποιήθηκε η κλιμακούμενη κανονικοποίηση η οποία αναλύθηκε πρωτύτερα και εν ολίγοις εξυπηρετεί στην ανάδειξη των σωστών συσχετιζόμενων αφθονιών της γονιδιακής έκφρασης μεταξύ των κυττάρων.

Η συνάρτηση που χρησιμοποιήθηκε για αυτόν τον σκοπό ήταν η computeSumFactors από το πακέτο του Bioconductor που κλιμακώνει την κανονικοποίηση των δεδομένων RNA – seq με αποκλιμάκωση των παραγόντων μεγέθους από τα σύνολα των κυττάρων (Aaron Lun, 2021).

Επιπλέον χρειάστηκε να υπολογιστούν οι λογαριθμικά τροποποιημένες τιμές της κανονικοποίησης με τη συνάρτηση logNormCounts (Lun A. , logNormCounts: Compute log-normalized expression values, 2024).

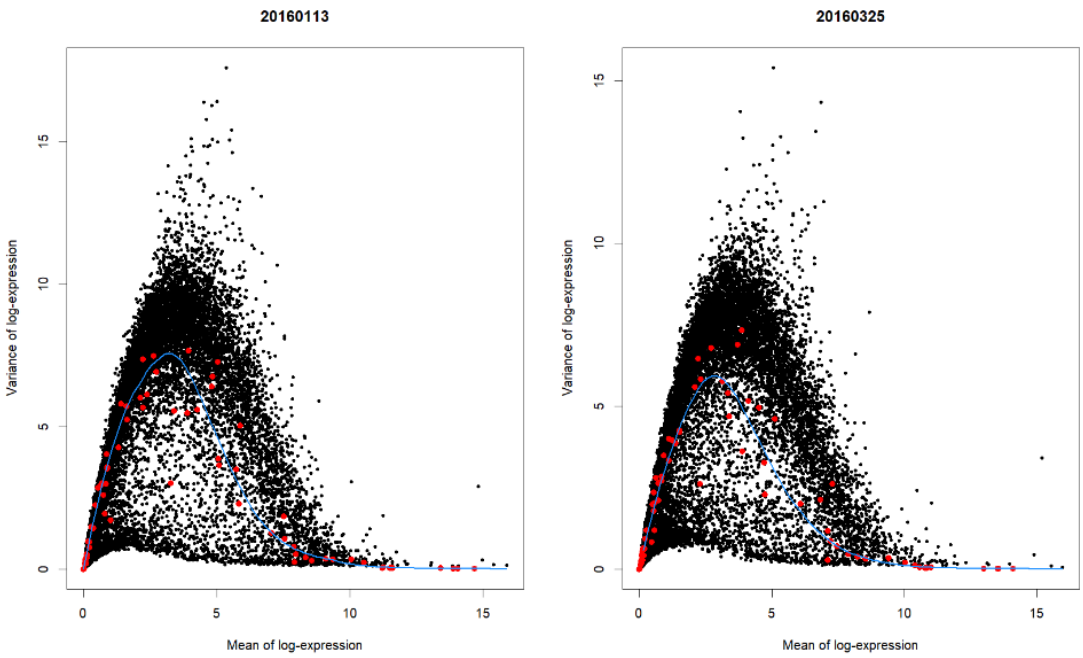


Εικόνα 14 Σχέση μεταξύ των παραγόντων μεγέθους της βιβλιοθήκης και των παραγόντων μεγέθους της αποκλιμάκωσης στο σύνολο δεδομένων 416B. Η Εικόνα δημιουργήθηκε στο RStudio έπειτα από την επανάλυση του συνόλου δεδομένων Lun 416B cell line.

Στην Εικόνα 16 παρατηρείται ότι τα επαγόμενα κύτταρα έχουν παράγοντες μεγέθους που μετατοπίζονται συστηματικά από τα μη επαγόμενα κύτταρα, γεγονός που υποδεικνύει την παρουσία ενός σφάλματος διάταξης. Το χρώμα στο διάγραμμα αντιπροσωπεύει την κατάσταση επαγωγής των ογκογονιδίων.

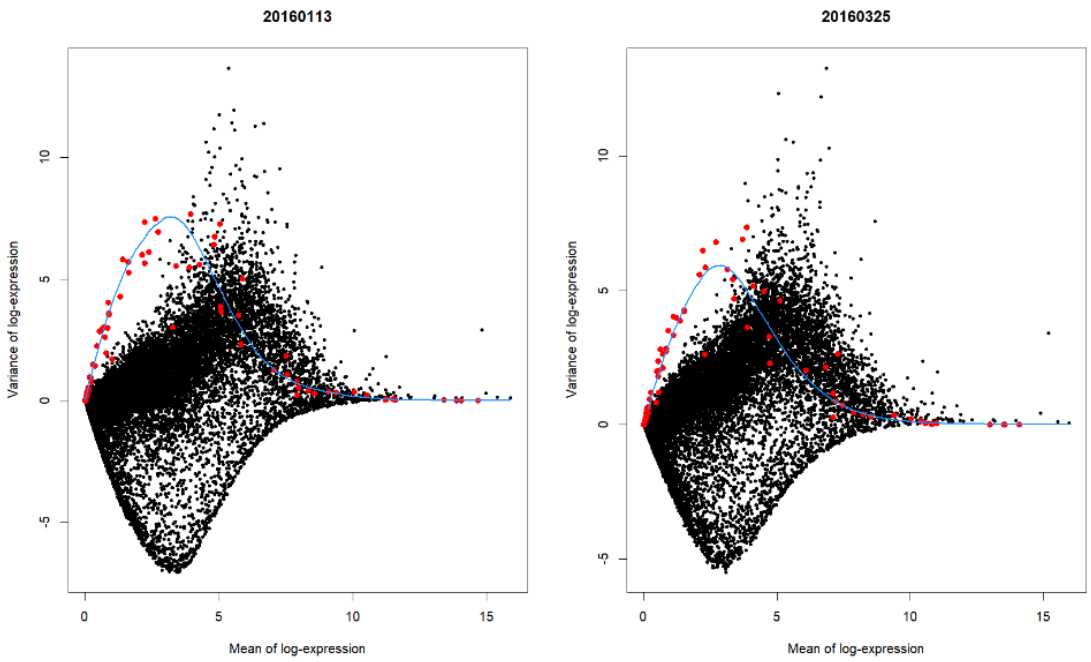
4.1.3 Επιλογή των γονιδίων για το σύνολο δεδομένων 416B

Για να γίνει η επιλογή των γονιδίων χρειάστηκε πρώτα να υπολογιστεί η μοντελοποίηση της διακύμανσης των προφίλ της λογαριθμικής έκφρασης του κάθε γονιδίου. Στη συνέχεια αυτή η διακύμανση διαχωρίστηκε σε τεχνικές και σε βιολογικές συνιστώσες με βάση μια τάση μέσης διακύμανσης, η οποία προσαρμόστηκε στην τάση μέσης διακύμανσης που είχε προκύψει από τις αναλογίες αναφοράς των spike – in μεταγράφων. Για να γίνει αυτό χρησιμοποιήθηκε η συνάρτηση modelGeneVarWithSpikes από το πακέτο του Bioconductor (Lun A. , modelGeneVarWithSpikes: Model the per-gene variance with spike-ins, 2024). Στην Εικόνα 17 φαίνεται η συνολική διακύμανση όπου αναπαρίστανται τα υπερεκφρασμένα και υποεκφρασμένα γονίδια (μαύρες κουκίδες).



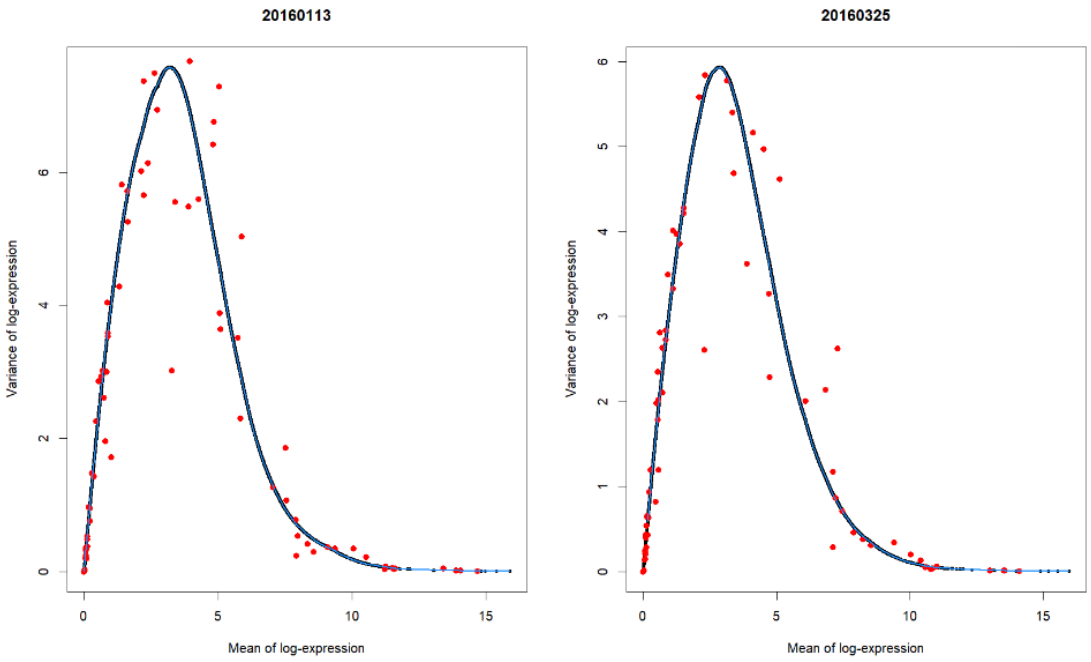
Εικόνα 15 Συνολική διακύμανση (total). Στη συνολική διακύμανση φαίνονται τα υπερεκφρασμένα και τα υποεκφρασμένα γονίδια. Με μαύρες κουκίδες αναπαρίστανται τα γονίδια, με κόκκινες κουκίδες τα spike in μετάγραφα και η μπλε γραμμή είναι η τάση μέσης διακύμανσης προσαρμοζόμενη στα spike in. Η εικόνα δημιουργήθηκε στο RStudio με τη χρήση του συνόλου δεδομένων «Lun 416B cell line».

Στην Εικόνα 18 φαίνεται η ενδιαφέρουσα βιολογική πληροφορία που αντιπροσωπεύεται από τη βιολογική συνιστώσα, δηλαδή η διαφορά μεταξύ της συνολικής διακύμανσης και της τεχνικής συνιστώσας.



Εικόνα 16 Βιολογική συνιστώσα (bio), δηλαδή η διαφορά μεταξύ της συνολικής (total) διακύμανσης και της τεχνικής (tech) συνιστώσας, είναι και η «ενδιαφέρουσα» πληροφορία για την επιλογή των γονιδίων. Με μαύρες κουκίδες αναπαρίστανται τα γονίδια, με κόκκινες κουκίδες τα spike in μετάγραφα και η μπλε γραμμή είναι η τάση μέσης διακύμανσης προσαρμοζόμενη στα spike in. Η εικόνα δημιουργήθηκε στο RStudio με τη χρήση του συνόλου δεδομένων «Lun 416B cell line».

Και τέλος στην Εικόνα 19 φαίνεται η τεχνική συνιστώσα που αποτελεί τη μη ενδιαφέρουσα πληροφορία και εφάπτεται απόλυτα με την τάσης της μέσης διακύμανσης, γεγονός αναμενόμενο αφού τα spike ins (κόκκινες κουκίδες) χρησιμοποιούνται στην ανάλυση ως επίπεδα αναφοράς της μεταγραφής, για να υφίστανται κάποιες συγκρίσιμες τιμές.



Εικόνα 17 Τεχνική συνιστώσα (tech), αντιπροσωπεύει την «μη ενδιαφέρουσα πληροφορία». Με μαύρες κουκίδες αναπαρίστανται τα γονίδια, με κόκκινες κουκίδες τα spike in μετάγραφα και η μπλε γραμμή είναι η τάση μέσης διακύμανσης προσαρμοζόμενη στα spike in. Το ενδιαφέρον εδώ είναι πως η τεχνική συνιστώσα εφάπτεται απόλυτα με την τάση μέσης διακύμανσης (μπλε γραμμή) (για αυτό και δεν φαίνονται οι μαύρες κουκίδες). Το φαινόμενο αυτό είναι λογικό καθώς τα spike ins χρησιμοποιούνται στην ανάλυση ως επίπεδα αναφοράς της μεταγραφής έτσι ώστε να είναι γνωστές κάποιες αναμενόμενες τιμές και να μπορεί να γίνει σύγκριση. Επομένως δεν δίνουν κάποια βιολογική πληροφορία για το δείγμα και θεωρούνται τεχνικός θόρυβος. Η εικόνα δημιουργήθηκε στο RStudio με τη χρήση του συνόλου δεδομένων «Lun 416B cell line».

Στη συνέχεια χρειάζεται να γίνει η επιλογή των γονιδίων εκείνων που παρουσιάζουν την μεγαλύτερη διακύμανση (HVGs). Για αυτόν τον σκοπό μέσω της συνάρτησης getTopHVGs ορίστηκε ένα σύνολο γονιδίων με υψηλή μεταβλητότητα με βάση τα στατιστικά που αντλήθηκαν από το προηγούμενο βήμα με την μοντελοποίηση της διακύμανσης (Lun A. , getTopHVGs: Identify HVGs, 2021). Συνήθως επιλέγεται ένα ποσοστό της τάξης του 10% των υψηλότερων μεταβλητών γονιδίων, αλλά σε αυτή τη μελέτη δοκιμάστηκαν αρκετά ποσοστά για να βρεθεί το ιδανικό για το συγκεκριμένο σύνολο δεδομένων.

4.1.4 Μείωση των διαστάσεων, Ομαδοποίηση και αξιολόγηση για το σύνολο δεδομένων 416B

Το επόμενο βήμα μετά την επιλογή των γονιδίων είναι η μείωση των διαστάσεων των δεδομένων, με τη βοήθεια των αλγορίθμων που συζητήθηκαν νωρίτερα. Αυτοί οι αλγόριθμοι ενσωματώνουν τον πίνακα έκφρασης σε έναν χώρο χαμηλών διαστάσεων

που είναι σχεδιασμένος ώστε να αποτυπώνει την ενυπάρχουσα δομή των δεδομένων σε όσο το δυνατών λιγότερες διαστάσεις. Αυτή η μέθοδος μπορεί να εφαρμοστεί σε τέτοιου είδους μελέτες, καθώς τα δεδομένα από τις αναλύσεις sc–RNA seq είναι από τη φύση τους χαμηλής διάστασης (Graham Heimberg, 2016). Με άλλα λόγια αυτό σημαίνει ότι η βιολογική πολλαπλότητα στην οποία βρίσκονται τα προφίλ κυτταρικής έκφρασης μπορεί να περιγράφει σε πολύ λιγότερες διαστάσεις από τον αριθμό των γονιδίων. Συνεπώς η μείωση των διαστάσεων αποσκοπεί στην εύρεση αυτών των διαστάσεων (Heumos, 2023).

Οι συναρτήσεις που χρησιμοποιήθηκαν για τον υπολογισμό των χαμηλότερων διαστάσεων προέρχονται από το πακέτο του Bioconductor και είναι οι runPCA, runTSNE runUMAP. Αυτές οι συναρτήσεις παίρνουν ως ορίσματα, μεταξύ άλλων, το σύνολο δεδομένων, τα επιλεγμένα γονίδια με τη μέγιστη μεταβλητότητα (hvgs) και τον αριθμό των διαστάσεων που επιθυμούμε να παραχθούν (ncomponents).

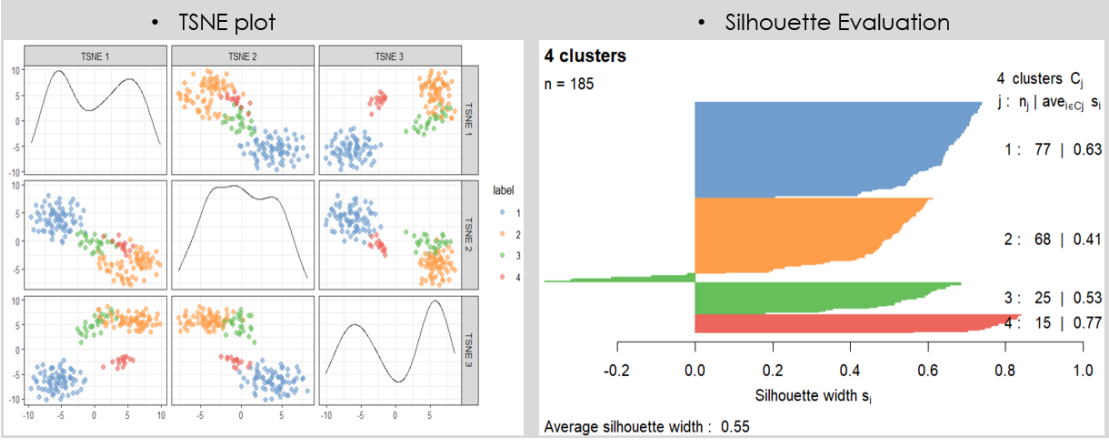
Για το σύνολο δεδομένων 416B δοκιμάστηκαν και οι τρεις μέθοδοι μείωσης διαστάσεων που αναλύθηκαν στο αντίστοιχο κεφάλαιο. Τα αποτελέσματα αυτών φορτώθηκαν έπειτα στους αλγορίθμους ομαδοποίησης, όπου με δοκιμές σε διαφορετικές παραμέτρους, παράχθηκε μια πρώτη εικόνα της ανάλυσης των δεδομένων, συμπεραίνοντας μέσω αυτής την ταυτότητα των κυττάρων. Στη συνέχεια πραγματοποιήθηκε αξιολόγηση των διαφορετικών μεθόδων και των παραμετροποιήσεων τους για να βρεθεί η καταλληλότερη για το συγκεκριμένο σύνολο δεδομένων.

Επομένως παρακάτω αναλύονται οι διαφορετικοί συνδυασμοί μεθόδων μείωσης διαστάσεων, ομαδοποίησης και οι αντίστοιχες παραμετροποιήσεις τους.

4.1.4.1 1η Δοκιμή: TSNE και Hierarchical clustering

Στην 1η Δοκιμή επιλέχθηκε σαν μέθοδος μείωσης διαστάσεων η TSNE με ορίσματα για τον αριθμό των γονιδίων το 10% το οποίο για τη δεδομένη ανάλυση έχει τιμή 1067 και αριθμό διαστάσεων το τρία. Η τεχνική ομαδοποίησης που εφαρμόστηκε ήταν το Hierarchical Clustering.

Όπως φαίνεται στην Εικόνα 20 αυτή η ανάλυση έδωσε τέσσερις διακριτές ομάδες και η αξιολόγηση με βάση το silhouette width είχε τιμή 0,55.

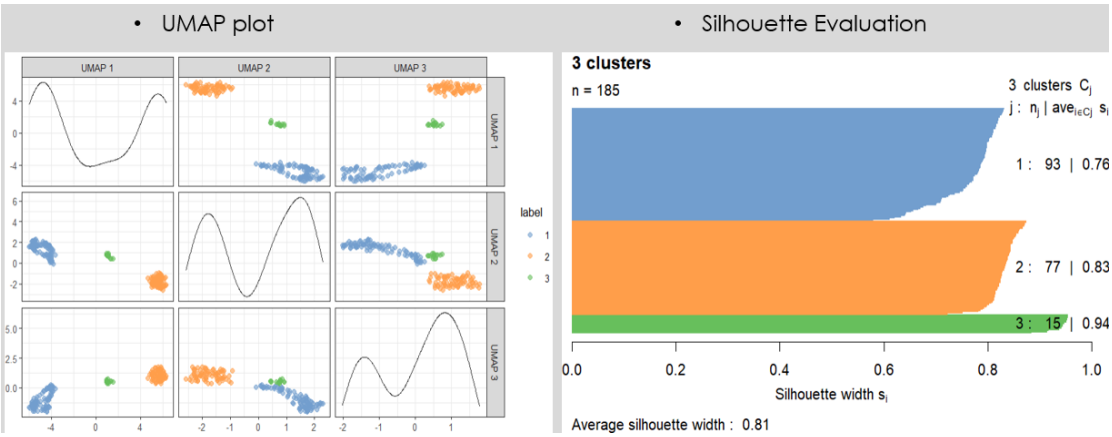


Εικόνα 20 Διάγραμμα από Hierarchical clustering για το σύνολο δεδομένων 416B με τη μέθοδο μείωσης διαστάσεων TSNE και το διάγραμμα του silhouette width για την αξιολόγηση. Η Εικόνα δημιουργήθηκε στο RStudio.

4.1.4.2 2η Δοκιμή: UMAP και Hierarchical clustering

Στην 2η Δοκιμή επιλέχθηκε σαν μέθοδος μείωσης διαστάσεων η UMAP με ορίσματα για τον αριθμό των γονιδίων το 10% το οποίο για τη δεδομένη ανάλυση έχει τιμή 1067 και αριθμό διαστάσεων το τρία. Η τεχνική ομαδοποίησης που εφαρμόστηκε ήταν το Hierarchical Clustering.

Όπως φαίνεται στην Εικόνα 21 αυτή η ανάλυση έδωσε τρεις διακριτές ομάδες και η αξιολόγηση με βάση το silhouette width είχε τιμή 0,81.

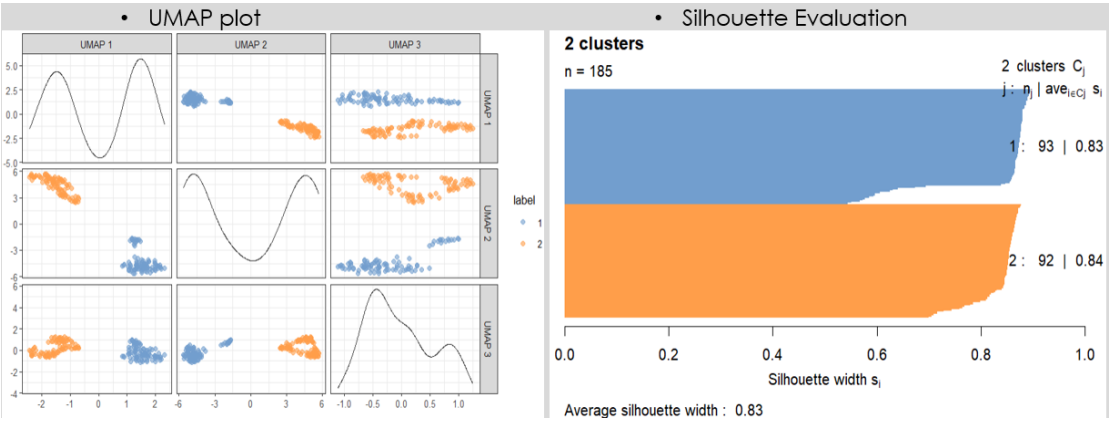


Εικόνα 21 Διάγραμμα από Hierarchical clustering για το σύνολο δεδομένων 416B με τη μέθοδο μείωσης διαστάσεων UMAP και το διάγραμμα του silhouette width για την αξιολόγηση. Η Εικόνα δημιουργήθηκε στο RStudio.

4.1.4.3 3η Δοκιμή: UMAP και Hierarchical clustering

Στην 3η Δοκιμή επιλέχθηκε σαν μέθοδος μείωσης διαστάσεων η UMAP με ορίσματα για τον αριθμό των γονιδίων τα 2000 γονίδια και αριθμό διαστάσεων το τρία. Η τεχνική ομαδοποίησης που εφαρμόστηκε ήταν το Hierarchical Clustering.

Όπως φαίνεται στην Εικόνα 22 αυτή η ανάλυση έδωσε δύο διακριτές ομάδες και η αξιολόγηση με βάση το silhouette width είχε τιμή 0,83.

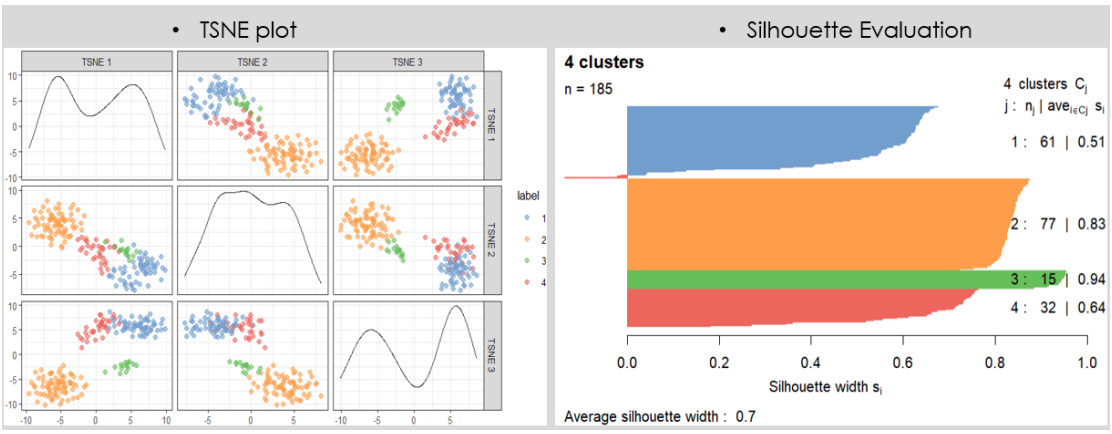


Εικόνα 18 Διάγραμμα από Hierarchical clustering για το σύνολο δεδομένων 416B με τη μέθοδο μείωσης διαστάσεων UMAP και το διάγραμμα του silhouette width για την αξιολόγηση. Η Εικόνα δημιουργήθηκε στο RStudio.

4.1.4.4 4η Δοκιμή: TSNE και K – means clustering

Στην 4η Δοκιμή επιλέχθηκε σαν μέθοδος μείωσης διαστάσεων η TSNE με ορίσματα για τον αριθμό των γονιδίων το 10% το οποίο για τη δεδομένη ανάλυση έχει τιμή 1067 και αριθμό διαστάσεων το τρία. Η τεχνική ομαδοποίησης που εφαρμόστηκε ήταν το K – means Clustering.

Όπως φαίνεται στην Εικόνα 23 αυτή η ανάλυση έδωσε τέσσερις διακριτές ομάδες και η αξιολόγηση με βάση το silhouette width είχε τιμή 0,7.

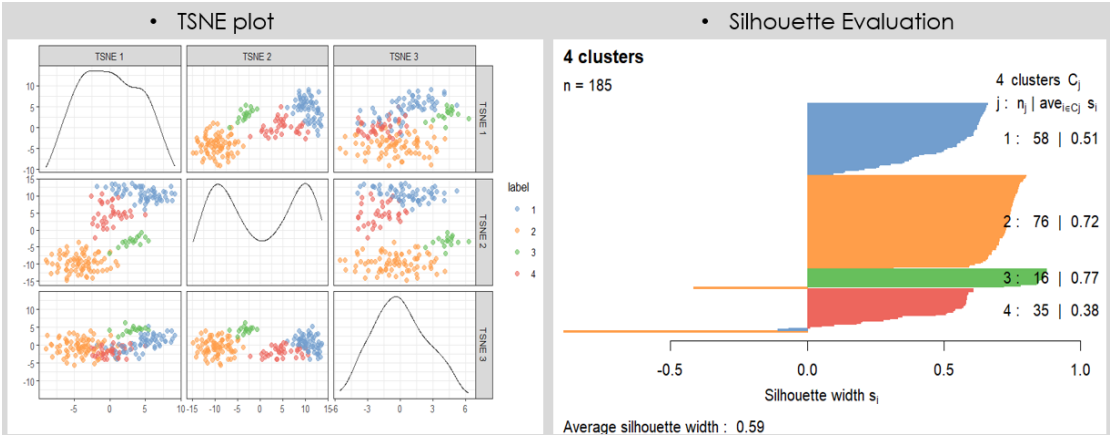


Εικόνα 19 Διάγραμμα από K – means clustering για το σύνολο δεδομένων 416B με τη μέθοδο μείωσης διαστάσεων TSNE και το διάγραμμα του silhouette width για την αξιολόγηση. Η Εικόνα δημιουργήθηκε στο RStudio.

4.1.4.5 5η Δοκιμή: TSNE και K – means clustering

Στην 5η Δοκιμή επιλέχθηκε σαν μέθοδος μείωσης διαστάσεων η TSNE με ορίσματα για τον αριθμό των γονιδίων τα 2000 γονίδια και αριθμό διαστάσεων το τρία. Η τεχνική ομαδοποίησης που εφαρμόστηκε ήταν το K – means Clustering.

Όπως φαίνεται στην Εικόνα 24 αυτή η ανάλυση έδωσε τέσσερις διακριτές ομάδες και η αξιολόγηση με βάση το silhouette width είχε τιμή 0,59.



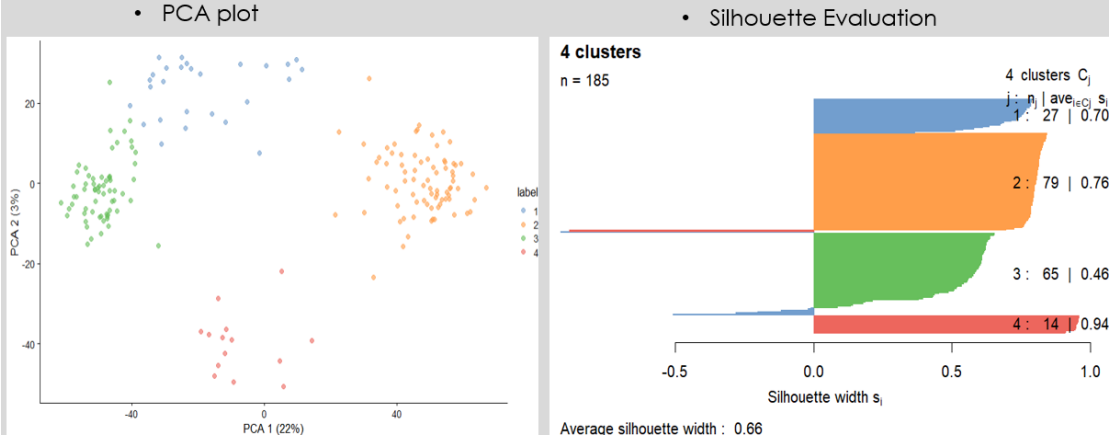
Εικόνα 24 Διάγραμμα από K – means clustering για το σύνολο δεδομένων 416B με τη μέθοδο μείωσης διαστάσεων TSNE και το διάγραμμα του silhouette width για την αξιολόγηση. Η Εικόνα δημιουργήθηκε στο RStudio.

4.1.4.6 6η Δοκιμή: PCA και K – means clustering

Στην 6η Δοκιμή επιλέχθηκε σαν μέθοδος μείωσης διαστάσεων η PCA με ορίσματα για τον αριθμό των γονιδίων το 10% το οποίο για τη δεδομένη ανάλυση έχει τιμή 1067 και

αριθμό διαστάσεων το δύο. Η τεχνική ομαδοποίησης που εφαρμόστηκε ήταν το K – means Clustering.

Όπως φαίνεται στην Εικόνα 25 αυτή η ανάλυση έδωσε τέσσερις διακριτές ομάδες και η αξιολόγηση με βάση το silhouette width είχε τιμή 0,66.

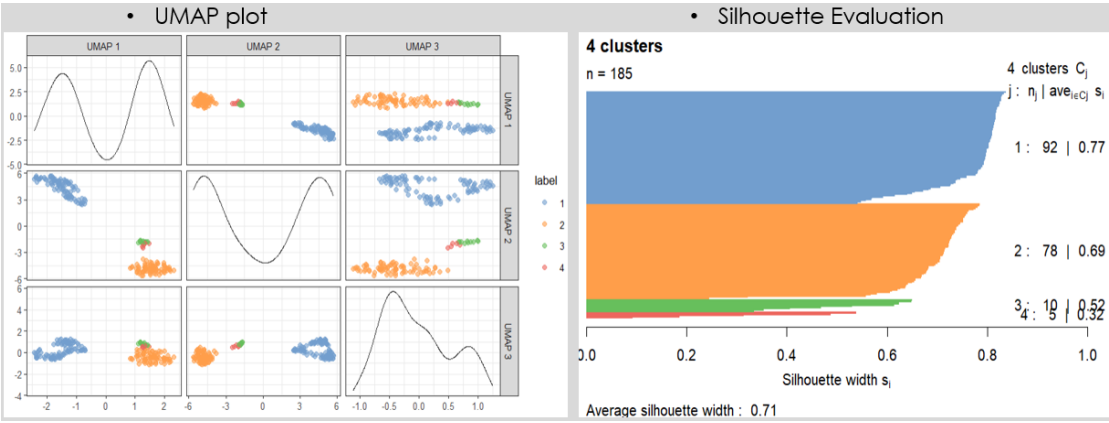


Εικόνα 25 Διάγραμμα από K – means clustering για το σύνολο δεδομένων 416B με τη μέθοδο μείωσης διαστάσεων PCA και το διάγραμμα του silhouette width για την αξιολόγηση. Η Εικόνα δημιουργήθηκε στο RStudio.

4.1.4.7 7η Δοκιμή: UMAP και K – means clustering

Στην 7η Δοκιμή επιλέχθηκε σαν μέθοδος μείωσης διαστάσεων η UMAP με ορίσματα για τον αριθμό των γονιδίων τα 2000 γονίδια και αριθμό διαστάσεων το τρία. Η τεχνική ομαδοποίησης που εφαρμόστηκε ήταν το K – means Clustering.

Όπως φαίνεται στην Εικόνα 26 αυτή η ανάλυση έδωσε τέσσερις διακριτές ομάδες και η αξιολόγηση με βάση το silhouette width είχε τιμή 0,71.

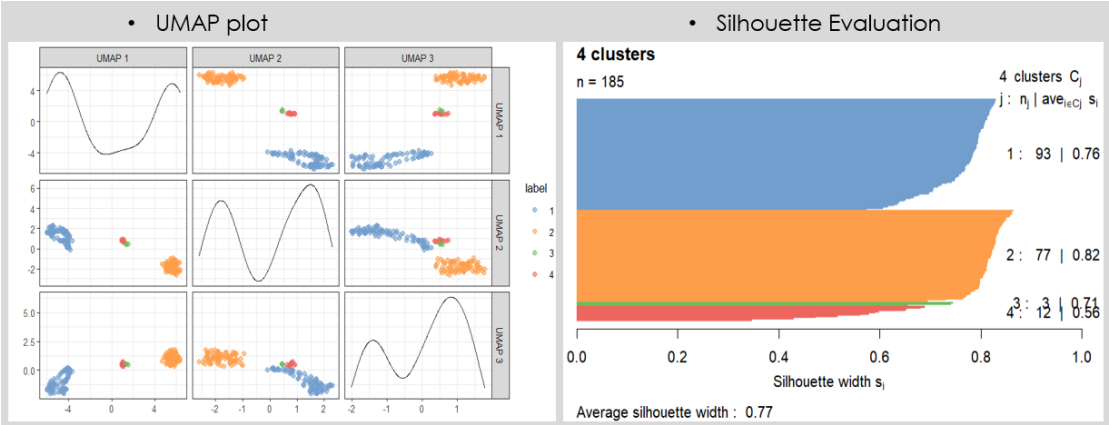


Εικόνα 26 Διάγραμμα από K – means clustering για το σύνολο δεδομένων 416B με τη μέθοδο μείωσης διαστάσεων UMAP και το διάγραμμα του silhouette width για την αξιολόγηση. Η Εικόνα δημιουργήθηκε στο RStudio.

4.1.4.8 8η Δοκιμή: UMAP και K – means clustering

Στην 8η Δοκιμή επιλέχθηκε σαν μέθοδος μείωσης διαστάσεων η UMAP με ορίσματα για τον αριθμό των γονιδίων το 10% το οποίο για τη δεδομένη ανάλυση έχει τιμή 1067 και αριθμό διαστάσεων το τρία. Η τεχνική ομαδοποίησης που εφαρμόστηκε ήταν το K – means Clustering.

Όπως φαίνεται στην Εικόνα 27 αυτή η ανάλυση έδωσε τέσσερις διακριτές ομάδες και η αξιολόγηση με βάση το silhouette width είχε τιμή 0,77.

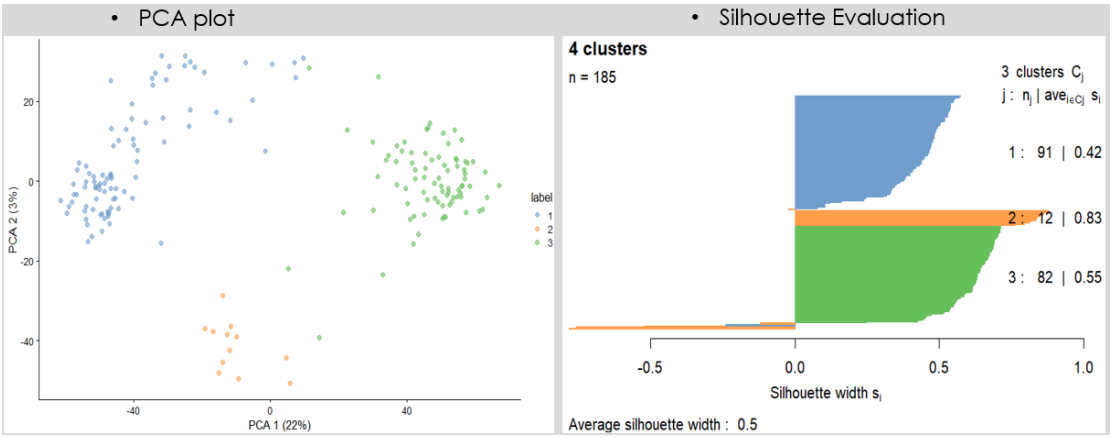


Εικόνα 27 Διάγραμμα από K – means clustering για το σύνολο δεδομένων 416B με τη μέθοδο μείωσης διαστάσεων UMAP και το διάγραμμα του silhouette width για την αξιολόγηση. Η Εικόνα δημιουργήθηκε στο RStudio.

4.1.4.9 9η Δοκιμή: PCA και Graph – based clustering

Στην 9η Δοκιμή επιλέχθηκε σαν μέθοδος μείωσης διαστάσεων η PCA με ορίσματα για τον αριθμό των γονιδίων το 10% το οποίο για τη δεδομένη ανάλυση έχει τιμή 1067 και αριθμό διαστάσεων το δύο. Η τεχνική ομαδοποίησης που εφαρμόστηκε ήταν το Graph – based Clustering.

Όπως φαίνεται στην Εικόνα 28 αυτή η ανάλυση έδωσε τέσσερις διακριτές ομάδες και η αξιολόγηση με βάση το silhouette width είχε τιμή 0,5.

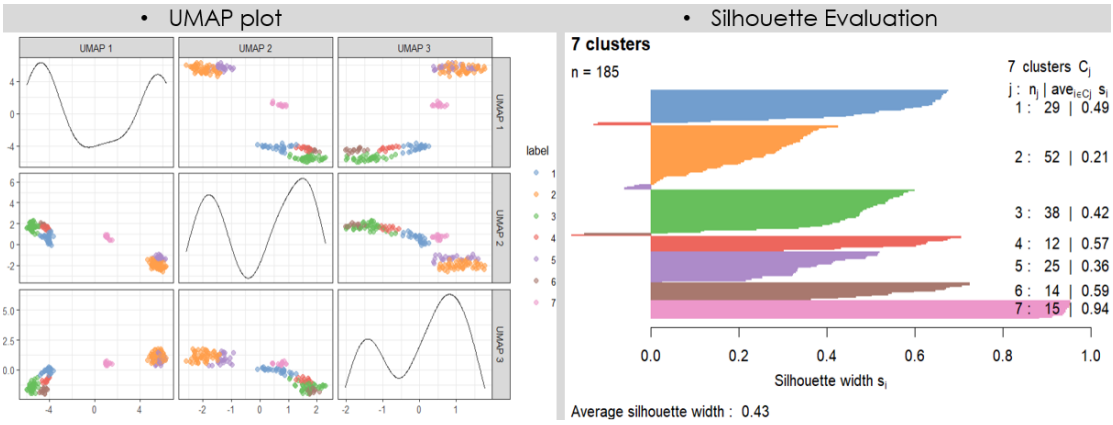


Εικόνα 28 Διάγραμμα από Graph – based clustering για το σύνολο δεδομένων 416B με τη μέθοδο μείωσης διαστάσεων PCA και το διάγραμμα του silhouette width για την αξιολόγηση. Η Εικόνα δημιουργήθηκε στο RStudio.

4.1.4.10 10η Δοκιμή: UMAP και Graph – based clustering

Στην 10η Δοκιμή επιλέχθηκε σαν μέθοδος μείωσης διαστάσεων η PCA με ορίσματα για τον αριθμό των γονιδίων το 10% το οποίο για τη δεδομένη ανάλυση έχει τιμή 1067 και αριθμό διαστάσεων το τρία. Η τεχνική ομαδοποίησης που εφαρμόστηκε ήταν το Graph – based Clustering.

Όπως φαίνεται στην Εικόνα 29 αυτή η ανάλυση έδωσε εφτά διακριτές ομάδες και η αξιολόγηση με βάση το silhouette width είχε τιμή 0,43.

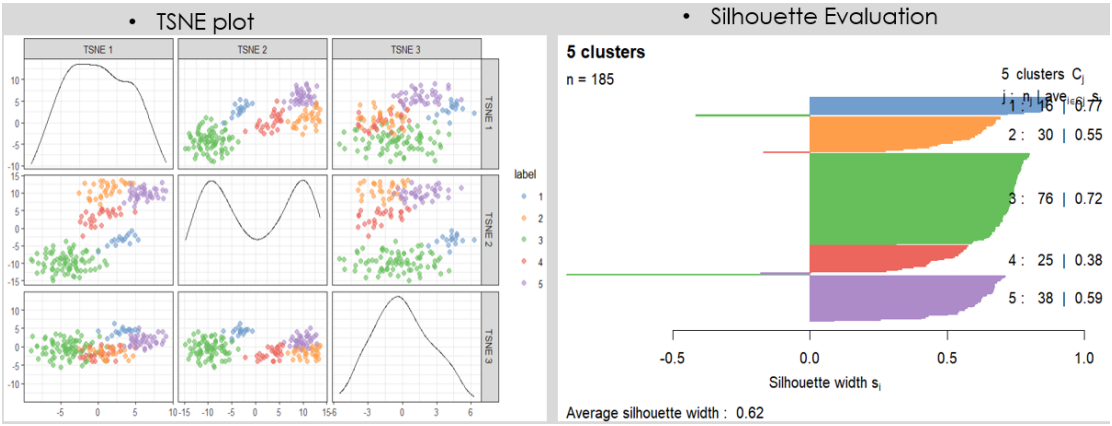


Εικόνα 29 Διάγραμμα από Graph – based clustering για το σύνολο δεδομένων 416B με τη μέθοδο μείωσης διαστάσεων UMAP και το διάγραμμα του silhouette width για την αξιολόγηση. Η Εικόνα δημιουργήθηκε στο RStudio.

4.1.4.11 11η Δοκιμή: PCA και Graph – based clustering

Στην 11η Δοκιμή επιλέχθηκε σαν μέθοδος μείωσης διαστάσεων η TSNE με ορίσματα για τον αριθμό των γονιδίων τα 2000 γονίδια και αριθμό διαστάσεων το τρία. Η τεχνική ομαδοποίησης που εφαρμόστηκε ήταν το Graph – based Clustering.

Όπως φαίνεται στην Εικόνα 30 αυτή η ανάλυση έδωσε πέντε διακριτές ομάδες και η αξιολόγηση με βάση το silhouette width είχε τιμή 0,62.



Εικόνα 30 Διάγραμμα από Graph – based clustering για το σύνολο δεδομένων 416B με τη μέθοδο μείωσης διαστάσεων TSNE και το διάγραμμα του silhouette width για την αξιολόγηση. Η Εικόνα δημιουργήθηκε στο RStudio.

4.1.4.12 Συμπεράσματα από τις δοκιμές για το σύνολο δεδομένων 416B

Από την παραπάνω ανάλυση φαίνεται πως για το συγκεκριμένο σύνολο δεδομένων η πιο κατάλληλη τεχνική για να ξεχωρίσει με ακρίβεια τις ομάδες μεταξύ τους είναι το Hierarchical clustering με τη μέθοδο μείωσης διαστάσεων UMAP, με τον αριθμό των γονιδίων μέγιστης μεταβλητότητας στα 2000 και τρείς διαστάσεις, καθώς έχει silhouette score ίσο με 0,83. Έπειτα ακολουθεί το Hierarchical clustering, ξανά με UMAP, αλλά με 1067 γονίδια και silhouette score ίσο με 0,81. Και τρίτη στη σειρά έρχεται η τεχνική του K – means clustering, με τη μέθοδο UMAP, 2000 γονίδια και silhouette score ίσο με 0,71. Την χειρότερη απόδοση φαίνεται να έχει η τεχνική του Graph – based clustering, με τη μέθοδο UMAP, 1067 γονίδια και silhouette score ίσο με 0,43. Στον Πίνακα 1 φαίνονται αναλυτικά οι μέθοδοι που εφαρμόστηκαν με τις παραμέτρους τους και τα αποτελέσματά τους σε silhouette score και σε αριθμό ομάδων.

Clustering Method	Dimensionality Reduction	Parameters (components, hvgs)	Silhouette Score	Clusters Num
Hierarchical	TSNE	N=3 , hvgs = 1067	0.55	4
Hierarchical	UMAP	N=3 , hvgs = 1067	0.81	3
Hierarchical	UMAP	N=3 , hvgs = 2000	0.83	2
K – means	TSNE	N=3 , hvgs = 1067	0.7	4
K – means	TSNE	N=3 , hvgs = 2000	0.59	4
K – means	PCA	hvgs = 1067	0.66	4
K – means	UMAP	N=3 , hvgs = 1067	0.77	4
K – means	UMAP	N=3 , hvgs = 2000	0.71	4
Graph - based	PCA	hvgs = 1064	0.5	4
Graph – based	TSNE	N=3 , hvgs = 2000	0.62	5
Graph - based	UMAP	N=3 , hvgs = 1067	0.43	7

Πίνακας 1 Συγκεντρωτικά αποτελέσματα εφαρμογής μεθόδων μείωσης διαστάσεων, ομαδοποίησης και αξιολόγησης

4.2 COST Action (European Cooperation in Science and Technology) Mye-Info-Bank

Τη σημερινή εποχή βλέπουμε ότι οι επιστήμες απαιτούν ολιστικές προσεγγίσεις για να δοθούν απαντήσεις στα ερωτήματα που θέτουν. Συνεπώς η συνεργασία μεταξύ

επιστημόνων που έχουν διαφορετικές γνώσεις, δεξιότητες και εμπειρίες είναι απαραίτητη για την προαγωγή της καινοτομίας και την επίλυση προβλημάτων.

Έτσι δημιουργήθηκε το Cost Action, μια Ευρωπαϊκή πρωτοβουλία που σκοπό έχει να δημιουργήσει ένα διεπιστημονικό ερευνητικό δίκτυο που θα φέρνει σε επαφή ερευνητές και ανθρώπους που πρωτοπορούν από όλες τις βαθμίδες της ακαδημαϊκής ιεραρχίας, για να μελετήσουν ένα θέμα της επιλογής τους. Τα θέματα αυτά περιλαμβάνουν πολλούς και διαφορετικούς τομείς της επιστήμης και της τεχνολογίας, ενισχύοντας την έρευνα και δημιουργώντας ευκαιρίες δικτύωσης, μιας και οι ερευνητές έχουν την ευκαιρία να συναντηθούν και να επικοινωνήσουν τις ιδέες τους για τα πολύπλοκα προβλήματα που απασχολούν την επιστημονική κοινότητα.

Το Cost φέρνει κοντά επιστήμονες από όλον τον κόσμο με σκοπό να αυξήσει τον πλούτο της γνώσης, να αποφευχθεί η απομόνωση των ερευνητικών κοινοτήτων, αλλά και να μειωθεί η επανάληψη στην έρευνα. Για να το κάνει αυτό διοργανώνει συνεχή επιστημονικά συνέδρια και εργαστήρια στα οποία οι ενδιαφερόμενοι παρουσιάζουν τις έρευνές τους αλλά και εκπαιδεύονται σε νέες τεχνολογίες χρήσιμες για τα θέματα που απασχολούν την ομάδα. Έπειτα οι συμμετέχοντες συνεργάζονται και μοιράζονται τις γνώσεις τους για την υλοποίηση των έργων της ομάδας CA20117 (COST, European Cooperation in Science and Technology, n.d.). https://www.mye-infobank.eu/home.html

Στα πλαίσια αυτής της διπλωματικής εργασίας έγινε συνεισφορά στο έργο του Cost Action που αφορά τη δημιουργία κυτταρικών ατλάντων στην εφαρμογή Cell x Gene και αναλύθηκαν δεδομένα για τύπους καρκίνου στο κεφάλι και στον λαιμό. Τα δεδομένα αναλύθηκαν, σχολιάστηκαν και ήρθαν στην κατάλληλη μορφή ώστε να ενσωματωθούν με το κατάλληλο λογισμικό (scRAFIKI) που έχει αναπτύξει η ομάδα του Cost ώστε να μπορέσουν να ενταχθούν στους ήδη υπάρχοντες άτλαντες του Cell x Gene.

4.3 Ενσωμάτωση δεδομένων στον sc-άτλαντα Καρκίνου του Κεφαλιού και Λαιμού (Head and Neck Cancer Atlas)

4.3.1 Επιλογή των συνόλων δεδομένων για την ενσωμάτωση στον sc – άτλαντα

Τα παρακάτω σύνολα δεδομένων επιλέχθηκαν από την ομάδα του Cost Action από τη βάση βιολογικών δεδομένων GEO (Gene Expression Omnibus), με σκοπό να ενσωματωθούν στον πρώτο κυτταρικό άτλαντα μεμονωμένων κυττάρων για καρκίνους

σε κεφάλι και λαιμό. Ύψιστος στόχος της ομάδας είναι να δημιουργηθεί μια ανοιχτή σε όλους πλατφόρμα για συνεισφορά στην έρευνα για την κατανόηση του καρκίνου του Κεφαλιού και Λαιμού.

Τα σύνολα δεδομένων περιέχουν πρωτογενή δεδομένα, για την ανάλυση των οποίων χρησιμοποιήθηκαν αντικείμενα τύπου Seurat. Το Seurat είναι ένα πολύ διαδεδομένο και ισχυρό εργαλείο ανάλυσης δεδομένων μεμονωμένων κυττάρων, καθώς χρησιμοποιείται ως αποθετήριο για τα πρωτογενή δεδομένα, αλλά και μεταδεδομένα των μελετών. Παράλληλα βοηθάει στην έρευνα περιέχοντας συναρτήσεις που θα χρησιμοποιηθούν για τον έλεγχο ποιότητας, την ανάλυση και τη διερεύνηση των δεδομένων και των αποτελεσμάτων (Hoffman P, 2024).

Για να μπορέσουν τα δεδομένα να ενσωματωθούν στον κυτταρικό άτλαντα, χρειάζεται να έρθουν στην κατάλληλη μορφή την οποία αναγνωρίζει το εργαλείο scRAFIKI, που ανέπτυξε η ομάδα του Cost Action και αναφέρθηκε πρωτύτερα. Συνεπώς εκτός από την ανάλυση, για την οποία ακολουθήθηκαν τα βήματα που έχουν περιγραφεί στα προηγούμενα κεφάλαια, έπρεπε να αντληθούν όλα τα μεταδεδομένα που χρειάζεται και αναγνωρίζει το λογισμικό του scRAFIKI και να αποκτήσουν την κατάλληλη δομή.

Τα μεταδεδομένα αυτά είναι πληροφορίες που αντιστοιχίζουν το κάθε κύτταρο με ορισμένα χαρακτηριστικά. Τα βασικά που χρειάστηκαν σε αυτή τη μελέτη ήταν το id του κάθε ασθενή, το φύλο του κάθε ασθενή, ο ιστός από τον οποίο προέρχεται κάθε κύτταρο, ο κυτταρικός τύπος, η τεχνολογία με βάση την οποία εκμαιεύτηκαν τα πρωτογενή δεδομένα και τέλος ο χαρακτηρισμός των κυττάρων όσον αφορά το εάν είναι θετικά ή αρνητικά στους ιούς HPV και EBV.

4.3.1.1 Ενσωμάτωση του συνόλου δεδομένων GSE212797

Ένα από τα σύνολα δεδομένων που ενσωματώθηκαν ήταν το GSE212797, το οποίο μετά από τις τεχνικές ελέγχου ποιότητας, κανονικοποίησης και φιλτραρίσματος, περιλαμβάνει 8245 πληροφοριακά κύτταρα, τα οποία έχουν προκύψει από πρωτόκολλα με χρήση της τεχνολογίας 10x genomics. Αυτά 8245 κύτταρα γνωρίζουμε ότι προέρχονται από πέντε ασθενείς, με τα 6376 κύτταρα να αντιστοιχούν σε γυναίκες, ενώ τα 1869 σε άντρες. Όσον αφορά τον ιστό τα 1498 κύτταρα προέρχονται από λεμφαδένες, ενώ τα 6747 από όγκους στην στοματική κοιλότητα. Πιο αναλυτικά, όπως φαίνεται και στον Πίνακα 2, για τους κυτταρικούς τύπους έχουμε 952 κύτταρα να αντιστοιχούν σε CD8_Effector, 287 κύτταρα σε CD8_Effector_GZMK, 1097 κύτταρα σε

CD8_Effector_IFNG_TNF, 2433 κύτταρα σε CD8_Exhausted, 2129 κύτταρα σε CD8_Naive/CM, 264 κύτταρα σε CD8_Prolif/Exhausted, 691 κύτταρα σε CD8_TCF7_Exhausted, 195 κύτταρα σε CD8_TEMRA και 197 κύτταρα σε CD8_TRM. Τέλος για τους ιούς HPV και EBV, κανένα κύτταρο δεν βρέθηκε θετικό για κάποιον από τους δύο ιούς.

Κυτταρικοί τύποι	Αριθμός κυττάρων
CD8_Effector	952
CD8_Effector_GZMK	287
CD8_Effector_IFNG_TNF	1097
CD8_Exhausted	2433
CD8_Naive/CM	2129
CD8_Prolif/Exhausted	264
CD8_TCF7_Exhausted	691
CD8_TEMRA	195
CD8_TRM	197

Πίνακας 2 Οι κυτταρικοί τύποι που περιλαμβάνονται στο σύνολο δεδομένων GSE212797

4.3.1.2 Ενσωμάτωση του συνόλου δεδομένων GSE210963

Το επόμενο σύνολο δεδομένων που ενσωματώθηκε ήταν το GSE210963, φορτώθηκε κι αυτό σε αντικείμενο τύπου Seurat και ακολουθήθηκε η ίδια διαδικασία όπως προηγουμένως για την ανάλυση και την άντληση των μεταδεδομένων που είναι απαραίτητα, με σκοπό την ενσωμάτωσή του με το εργαλείο scRAFIKI και τελικά την προσθήκη του στον κυτταρικό άτλαντα.

Επομένως για τα μεταδεδομένα βρέθηκε ότι περιλαμβάνει 4719 πληροφοριακά κύτταρα που ανήκουν όλα σε έναν ασθενή, του οποίου το φύλο δεν γνωρίζουμε. Τα κύτταρα αυτά βρέθηκαν στο αίμα του ασθενή και οι κυτταρικοί τύποι που βρέθηκαν έπειτα και από τη χρήση του εργαλείου Single R για τον αυτόματο σχολιασμό, παρουσιάζονται στον Πίνακα 3 και ήταν τα ακόλουθα: 57 κύτταρα χαρακτηρίστηκαν ως B – cells, 293 ως CD4+ T – cells , 289 ως CD8+ T – cells, 24 ως Eosinophils, 3 ως HSC, 3655 ως Monocytes, 40

ως Neutrophils και 358 ως NK – cells. Όπως και στο προηγούμενο σύνολο δεδομένων κανένα κύτταρο δεν βρέθηκε θετικό για κάποιον από τους δύο ιούς HPV και EBV. Τέλος, ισχύει και σε αυτήν την περίπτωση, ότι τα κύτταρα αυτά έχουν προκύψει από πρωτόκολλα με χρήση της τεχνολογίας 10x genomics.

Κυτταρικοί τύποι	Αριθμός κυττάρων
B – cells	57
CD4+ T – cells	293
CD8+ T – cells	289
Eosinophils	24
HSC	3
Monocytes	3655
Neutrophils	40
NK – cells	358

Πίνακας 3 Οι κυτταρικοί τύποι που περιλαμβάνονται στο σύνολο δεδομένων GSE210963

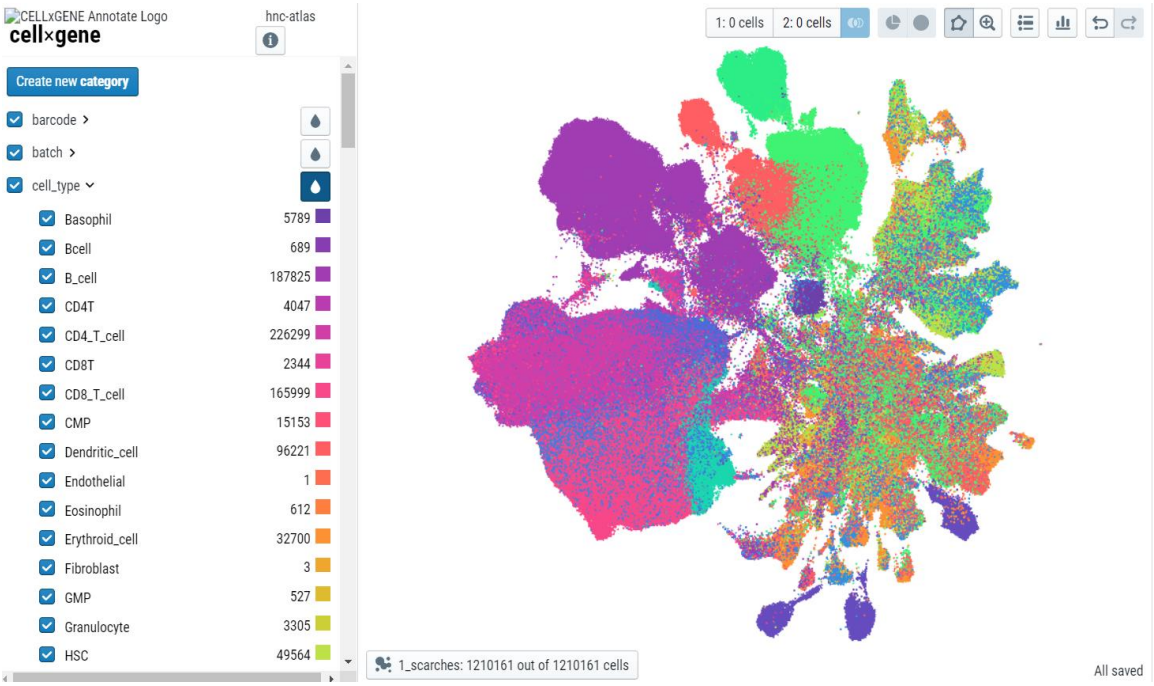
4.3.2 Ο sc - κυτταρικός άτλαντας Καρκίνου Κεφαλιού και Λαιμού

Ο στόχος της ομάδας να δημιουργήσει έναν μεγάλο κυτταρικό άτλαντα μεμονωμένων κυττάρων, που θα αποτελέσει μια πλατφόρμα για έρευνα για τον καρκίνο του κεφαλιού και του λαιμού επιτεύχθηκε, με την ομάδα να καταφέρνει να ενσωματώσει 1,2 εκατομμύρια κύτταρα.

Οι ενδιαφερόμενοι και οι ερευνητές μπορούν να βρουν αυτόν τον άτλαντα στην σελίδα του Cell x Gene (https://exbio.wzw.tum.de/cost/app/hnc-atlas-cellxgene) και να δοκιμάσουν τα φίλτρα του καταλόγου για να διερευνήσουν αυτά τα κύτταρα με βάση το σύνολο δεδομένων, τους ασθενείς, το φύλο τους, τον ιστό, τον κυτταρικό τύπο, ακόμα και το πρωτόκολλο που έχει χρησιμοποιηθεί κατά το βιολογικό πείραμα κ.α. Ανάλογα με το τί επιθυμεί να δει ο χρήστης μπορεί να χρωματίσει τα κύτταρα με βάση τις κατηγορίες που διατίθενται, όπως για παράδειγμα φαίνεται στην Εικόνα 31 όπου τα κύτταρα έχουν χρωματιστεί με βάση τον κυτταρικό τύπο. Επιπλέον, μια πολύ ενδιαφέρουσα λειτουργικότητα του λογισμικού αποτελεί η επιλογή τεχνικής Ομαδοποίησης που

χρησιμοποιήθηκε σε κάθε ανάλυση, έτσι ώστε οι χρήστες να μπορούν μόλις σε λίγο χρόνο να έχουν μια πλήρη εικόνα για την πολυπλοκότητα του άτλαντα και να μπορούν να δουν όλες τις διαφορετικές «όψεις του ίδιου νομίσματος».

Αυτό το μεγάλο αποθετήριο δεδομένων μεμονωμένων κυττάρων και των μεταδεδομένων τους, μπορεί να χρησιμοποιηθεί από τους ερευνητές για τις δικές τους μελέτες και να συνεισφέρει σημαντικά στην κατανόηση της βιολογίας, της ετερογένειας και του μικροπεριβάλλοντος των όγκων. Ως άμεση συνέπεια ο εντοπισμός διαφορετικών υποτύπων θα μπορέσει να βοηθήσει στην εύρεση εξατομικευμένων στρατηγικών θεραπείας του καρκίνου, αλλά και στον εντοπισμό νέων θεραπευτικών στόχων ή νέων βιοδεικτών για την πιο έγκαιρη διάγνωση της νόσου (Zhang Y, 2021).



Εικόνα 20 O sc - κυτταρικός άτλαντας καρκίνου κεφαλιού και λαιμού, χρωματισμένος με βάση τον κυτταρικό τύπο. Η Εικόνα αντλήθηκε από τη σελίδα του Cell x Gene https://exbio.wzw.tum.de/cost/app/hnc-atlas-cellxgene

Βιβλιογραφία

Aaron Lun, K. B. (2021). computeSumFactors: Normalization by deconvolution. Ανάκτηση από https://rdrr.io/bioc/scran/man/computeSumFactors.html

Aaron T.L. Lun, D. J. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research*. doi:10.12688/f1000research.9501.2

Aaron T.L. Lun, F. J.-N.-V. (2017). Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. *Genome Research*. doi:10.1101/gr.222877.117

Amezquita, R. L. (2019). Orchestrating Single-Cell Analysis with Bioconductor. *Nature Methods*. Ανάκτηση από https://doi.org/10.1038/s41592-019-0654-x

Atlas M Sardoo, S. Z. (2022). Decoding brain memory formation by single-cell RNA sequencing. *Briefings in Bioinformatics*. doi:10.1093/bib/bbac412

Bassel Ghaddar, S. D. (2023). Hierarchical and automated cell-type annotation and inference of cancer cell of origin with Census. *Bioinformatics*. Ανάκτηση από https://doi.org/10.1093/bioinformatics/btad714

Bioconductor. (χ.χ.). Ανάκτηση από https://www.bioconductor.org/

BioLizard. (2021). Ανάκτηση από Single-cell vs. bulk sequencing: which one to use when?: https://lizard.bio/knowledge-hub/single-cell-vs-bulk-sequencing

COST, European Cooperation in Science and Technology. (χ.χ.). Ανάκτηση από https://www.cost.eu/

CZ Cell x Gene. (χ.χ.). Ανάκτηση από https://cellxgene.cziscience.com/

Dragomirka Jovic, X. L. (2022). Single-cell RNA sequencing technologies and applications: A brief overview. *CIlinical and Translational Medicine*. doi:10.1002/ctm2.694

Dvir Aran, A. P. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Natute Immunology*.

github.com. (2024). Ανάκτηση από Mye-InfoBank-scRAFIKI: https://github.com/Mye-InfoBank/scRAFIKI?tab=readme-ov-file

Graham Heimberg, R. B.-S. (2016). Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. *Cell Syst.* doi:10.1016/j.cels.2016.04.001

Hajji, F. Z. (2024). *The Gradient*. Ανάκτηση από Deep learning for single-cell sequencing: a microscope to see the diversity of cells: https://thegradient.pub/deep-learning-for-single-cell-sequencing-a-microscope-to-uncover-the-rich-diversity-of-individual-cells/

Haque, A. E. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med*. Ανάκτηση από https://doi.org/10.1186/s13073-017-0467-4

Heumos, L. S. (2023). Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*. Ανάκτηση από https://doi.org/10.1038/s41576-023-00586-w

Himanshu Savardekar, C. A. (2024). Single-Cell RNA-Seq Analysis of Patient Myeloid-Derived Suppressor Cells and the Response to Inhibition of Bruton's Tyrosine Kinase. *Molecular Cancer Research*. Ανάκτηση από https://doi.org/10.1158/1541-7786.MCR-22-0572

Hoffman P, S. R. (2024). *SeuratObject: Data Structures for Single Cell Data.* Ανάκτηση από https://cloud.r-project.org/web/packages/SeuratObject/SeuratObject.pdf

Jana-Charlotte Hegenbarth, G. L. (2022). Perspectives on Bulk-Tissue RNA Sequencing and Single-Cell RNA Sequencing for Cardiac Transcriptomics. *Frontiers in Molecular Medicine*. Ανάκτηση από https://doi.org/10.3389/fmmed.2022.839338

Jonathan A Griffiths, A. S. (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. *Molecular Systems Biology*. Ανάκτηση από https://doi.org/10.15252/msb.20178046

Kevin R. Moon, J. S. (2017). Manifold Learning-based Methods for Analyzing Single-Cell RNA-Sequencing Data. *Current Opinion in Systems Biology*. doi:10.1016/j.coisb.2017.12.008

Lun, A. (2021). getTopHVGs: Identify HVGs. Ανάκτηση από https://rdrr.io/bioc/scran/man/getTopHVGs.html

Lun, A. (2024). logNormCounts: Compute log-normalized expression values. Ανάκτηση από https://rdrr.io/github/LTLA/scuttle/man/logNormCounts.html

Lun, A. (2024). modelGeneVarWithSpikes: Model the per-gene variance with spike-ins. Ανάκτηση από https://rdrr.io/github/MarioniLab/scran/man/modelGeneVarWithSpikes.html

Lun, A. (2024). Per - cell quality control metrics. Ανάκτηση από https://rdrr.io/github/LTLA/scuttle/man/perCellQCMetrics.html

Lun, A. (2024). quickPerCellQC: Quick cell level QC. Ανάκτηση από https://rdrr.io/github/LTLA/scuttle/man/quickPerCellQC.html

Lun, A. T.-N.-V. (2017). Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. *Genome Res*. doi:10.1101/gr.222877.117

Maha K. Rahim, T. L. (2023). Dynamic CD8+ T cell responses to cancer immunotherapy in human regional lymph nodes are disrupted in metastatic lymph nodes. *Cell*. Ανάκτηση από https://doi.org/10.1016/j.cell.2023.02.021

Po-Yuan Tung, J. D. (2017). Batch effects and the effective design of single-cell gene expression studies. *Scientific Reports*. doi:10.1038/srep39921

rdrr.io. (χ.χ.). Ανάκτηση από https://rdrr.io/github/LTLA/scuttle/man/perCellQCFilters.html

Robert Amezquita, A. L. (2023). *Basics of Single-Cell Analysis with Bioconductor.* Bioconductor.

Stefan Salcher, G. S. (2022). High-resolution single-cell atlas reveals diversity and plasticity of tissue-resident neutrophils in non-small cell lung cancer. *Cancer Cell*. doi:doi:10.1016/j.ccell.2022.10.008

Theis, M. D. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*. Ανάκτηση από https://doi.org/10.15252/msb.20188746

Tran, M.-H. (2022). *scRNA-Seq Cell Type Annotation: Common Approaches and Tools*. Ανάκτηση από BioTuring: https://blog.bioturing.com/2022/03/03/scrna-seq-cell-type-annotation-common-approaches-and-tools/#:~:text=Two%20Common%20Approaches%3A%20Manual%20or,profiles%20to%20assign%20cell%20identities.

Vaga, S. (2022). Understanding Single Cell Sequencing, How It Works and Its Applications. *Genomics Research*.

Van de Sande, B. L.-G. (2023). Applications of single-cell RNA sequencing in drug discovery and development. *Nature Reviews Drug Discovery*. Ανάκτηση από https://doi.org/10.1038/s41573-023-00688-4

Xiangling Ji, D. T. (2023). scAnnotate: an automated cell-type annotation tool for single-cell RNA-sequencing data. *Bioinformatics Advances*. Ανάκτηση από https://doi.org/10.1093/bioadv/vbad030

Xiaoning Tang, Y. H. (2019). The single-cell sequencing: new developments and medical applications. *Cell & Bioscience*. Ανάκτηση από https://doi.org/10.1186/s13578-019-0314-y

Xin Wang, C. W. (2011). Semi–supervised K-Means Clustering by Optimizing Initial Cluster Centers. *Lecture Notes in Computer Science*. Ανάκτηση από https://doi.org/10.1007/978-3-642-23982-3_23

Zhang Y, W. D. (2021). Single-cell RNA sequencing in cancer research. *J Exp Clin Cancer Res.* doi:10.1186/s13046-021-01874-1