ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

ΕΡΓΑΛΕΙΑ ΔΙΕΡΕΥΝΗΤΙΚΗΣ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΤΕΧΝΙΚΕΣ Single-Cell RNA-Sequencing

ΚΟΣΜΑΡΙΚΑ ΜΑΡΙΑ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων Υπεύθυνος: ΒΡΑΧΑΤΗΣ ΑΡΙΣΤΕΙΔΗΣ

Λαμία, 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

ΕΡΓΑΛΕΙΑ ΔΙΕΡΕΥΝΗΤΙΚΗΣ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΤΕΧΝΙΚΕΣ Single-Cell RNA-Sequencing

ΚΟΣΜΑΡΙΚΑ ΜΑΡΙΑ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων Υπεύθυνος: ΒΡΑΧΑΤΗΣ ΑΡΙΣΤΕΙΔΗΣ

Λαμία, 2022

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [1], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. *Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί* ***χωρίς να τα περικλείω σε εισαγωγικά*** *και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

2. *Δέχομαι ότι η αυτολεξεί* ***παράθεση χωρίς εισαγωγικά****, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. *Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια*

4. *Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.*

Ημερομηνία: ……/…../20……

Ο – Η Δηλ.

(Υπογραφή)

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

ΕΡΓΑΛΕΙΑ ΔΙΕΡΕΥΝΗΤΙΚΗΣ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΤΕΧΝΙΚΕΣ Single-Cell RNA-Sequencing

ΚΟΣΜΑΡΙΚΑ ΜΑΡΙΑ

Τριμελής Επιτροπή:

Ονοματεπώνυμο: ΒΡΑΧΑΤΗΣ ΑΡΙΣΤΕΙΔΗΣ

Ονοματεπώνυμο: ΤΑΣΟΥΛΗΣ ΣΩΤΗΡΙΟΣ

Ονοματεπώνυμο: ΠΛΑΓΙΑΝΑΚΟΣ ΒΑΣΙΛΕΙΟΣ

ΠΕΡΙΕΧΟΜΕΝΑ

Περιεχόμενα εικόνων και πινάκων
Περίληψη
Abstract
Ευχαριστίες

Περιεχόμενα

1 Εισαγωγικά στοιχεία……..12

1.1 **Δομή και λειτουργίες του DNA και του RNA……..**12

1.2 **Αντιγραφή……...**14

1.3 **Μεταγραφή και Μετάφραση**……14

1.4 **Μηχανισμοί και ένζυμα**……..15

1.5 **Ανάγκη εξειδικευμένης μελέτης**……15

2 Μέθοδοι μελέτης των αλληλουχιών DNA και RNA……17

3 Η διαδικασία single-cell-RNA-sequencing……...21

3.1 Πρώτο στάδιο……21

3.1.1 **Μέθοδοι συλλογής κυτταρικών δειγμάτων……..**21

3.1.2 **Δημιουργία cDNA βιβλιοθήκης και κλωνοποίηση**……..22

3.1.3 **Συνοπτική παρουσίαση των βημάτων κατά την εκτέλεση προετοιμασίας ανάλυσης κυττάρων από ασθενείς με COVID-19**……24

3.1.4 **Γιατί η single-cell-RNA-sequencing διαφοροποιείται κατά την μελέτη κυττάρων από καρδιακό ιστό;**……26

3.2 Δεύτερο στάδιο……...……28

3.2.1 **Πίνακας δεδομένων**……28

3.2.2 **Ποιοτικός έλεγχος**……28

3.2.3 **Κανονικοποίηση**……...29

3.2.4 **Μείωση διαστάσεων δεδομένων**……29

3.3 Τελευταίο στάδιο……30

3.3.1 **Συσταδοποίηση**……31

3.3.2 **Trajectory analysis**……..33

3.3.3 **Ταυτοποίηση κυτταρικών ομάδων**……33

3.3.4 **Differential expression analysis**……34

4 Προκλήσεις κατά την μελέτη δεδομένων single-cell-RNA-sequencing……...35

4.1 **Dropouts**……..35

4.2 **Αφαίρεση batch effect παραγόντων**……..37

4.3 **Άλλες προκλήσεις**……...38

4.3.1 **Χαρτογράφηση**……38

4.3.2 **Βελτίωση της ανάλυσης (trajectory)**……39

4.3.3 **Αλληλεπίδραση**……39

4.3.4 **Συνεχείς αξιολογήσεις**……40

5 Σύγκριση πρωτοκόλλων 10X Genomics και Smart-Seq2……..42

6 Υπολογιστικά εργαλεία και πλατφόρμες……45

6.1 **Scedar**……48

6.2 **CALISTA**……...51

6.3 **Seurat**……51

6.4 **Scanpy**……..53
6.5 **Gene Pattern Notebook**……..54
6.6 **scGEATool**……..55
6.7 **BBrowser**…….56

7 Ανάλυση και συμπεράσματα…….58
7.1 **Δεδομένα ανάλυσης**……58
7.2 **Πλατφόρμες ανάλυσης και συμπεράσματα**……..60
7.2.1 **Gene Pattern Notebook**……..60
7.2.2 **ScGEATool**……..61

8 Συμπεράσματα……88

Βιβλιογραφία

Περιεχόμενα εικόνων και πινάκων

Εικόνες:

Εικόνα 1

Εικόνα 2

Εικόνα 3

Εικόνα 4

Εικόνα 5

Εικόνα 6

Εικόνα 7

Εικόνα 8

Εικόνα 9

Εικόνα 10

Εικόνα 11

Εικόνα 12

Εικόνα 13

Εικόνα 14

Εικόνα 15

Εικόνα 16

Εικόνα 17

Εικόνα 18

Εικόνα 19

Εικόνα 20

Εικόνα 21

Πίνακες:

Πίνακας 1

Πίνακας 2

Περίληψη

Η διαδικασία single-cell RNA-sequencing έφερε την επανάσταση στην διερεύνηση πεδίων της επιστήμης της ιατρικής και της βιολογίας. Η μελέτη της συμπεριφοράς καθενός κυττάρου ξεχωριστά και των γονιδίων που εκφράζονται σε αυτό έδωσε νέα προοπτική στην ανακάλυψη πτυχών των ασθενειών που η κατά μέσο όρο προσέγγιση που υπήρξε πριν (RNA-sequencing) αδυνατούσε να δώσει λεπτομέρειες σε τέτοιο βάθος. Μαζί με τις νέες προοπτικές προέκυψαν και προκλήσεις που χρειάζεται να αντιμετωπιστούν. Το ενδιαφέρον των ερευνητών αναπτύσσεται ραγδαία δημιουργώντας συνεχώς νέα εργαλεία για μελέτη και απόκτηση περισσότερων γνώσεων και αντιμετώπιση των ζητημάτων.

Η παρούσα εργασία εκπονήθηκε με σκοπό την καταγραφή των βασικών στοιχείων της διαδικασίας single-cell-RNA-sequencing έπειτα από βιβλιογραφική αναζήτηση, την αναζήτηση και εύρεση μεταξύ άλλων του κατάλληλου υπολογιστικού εργαλείου που θα απευθύνεται σε χρήστες που δεν έχουν προηγούμενες προγραμματιστικές γνώσεις, την πραγματοποίηση διερευνητικής ανάλυσης δεδομένων, δηλαδή μιας πλήρους ανάλυσης δεδομένων τύπου single-cell RNA σε αυτό, και την αξιολόγηση του υπολογιστικού εργαλείου.

Λέξεις-κλειδιά: single-cell-RNA-sequencing, υπολογιστικά εργαλεία, διερευνητική ανάλυσης δεδομένων, αξιολόγηση

Abstract

The single-cell RNA-sequencing analysis has a huge impact in discovering new areas for study in medical field and biology. By measuring the activity of one cell at a time and the genes of this cell, this opens up more opportunities than the bulk RNA-sequencing offered in the past. Nevertheless, this new type of analysis has also some challenges to tackle. The scientists interest is increasing rapidly and a lot of new tools created in order to discover more information and to handle with all of the problems.
This thesis is written in the purpose of mentioning all of the characteristics of single-cell RNA-sequencing, the challenges of the single-cell RNA-sequencing process after completing research, and try finding out the best user friendly tool for analysis for users without any previous programming skills, accomplishing an exploratory single-cell-RNA data analysis and benchmarking the tool.

Keywords: single-cell-RNA-sequencing, bioinformatic tools, exploratory data analysis, benchmarking

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέπων καθηγητή μου κ. Βραχάτη Αριστείδη για την πολύτιμη καθοδήγηση του καθ'όλη την διάρκεια εκπόνησης της πτυχιακής εργασίας και την ευκαιρία που μου έδωσε να μελετήσω σε βάθος ένα τόσο σημαντικό θέμα στον κλάδο της Βιοπληροφορικής.

Τέλος θα ήθελα να ευχαριστήσω την οικογένειά μου που με στήριξε όλα αυτά τα χρόνια των σπουδών μου και τα ξαδέρφια μου Δημήτρη και Πέτρο και τις οικογένειές τους που ήταν στο πλάι μου σε μια δύσκολη περίοδο κατά την εκπόνηση της εργασίας αυτής.

1 Εισαγωγικά στοιχεία

Έχουμε αναρωτηθεί ποτέ, πως έφτασε το ανθρώπινο είδος να έχει ένα τόσο αξιόλογο βιοτικό επίπεδο ζωής; Τι έχει μεσολαβήσει έτσι ώστε η πρόληψη και οι σωστές μέθοδοι μελέτης και θεραπείας να μπορούν να βελτιώσουν την ποιότητα ζωής των ανθρώπων; Οι τελευταίες δεκαετίες χαρακτηρίζονται από ριζικές αλλαγές σε πολλούς τομείς και φυσικά δεν θα μπορούσαν να αφήσουν ανεπηρέαστο τον τομέα των βιοϊατρικών επιστημών. Η Βιοπληροφορική, ως φυσική εξέλιξη των επιστημών της Βιολογίας και της ραγδαίας ανάπτυξης της επιστήμης των Υπολογιστών περιλαμβάνει όλες εκείνες τις υπολογιστικές μεθόδους (μαθηματικές συναρτήσεις, υπολογιστικές τεχνικές, αλγόριθμοι κ.α.) και σε συνδυασμό με άλλες θετικές επιστήμες όπως η Μοριακή Βιολογία, η Βιοχημεία, η Φυσική, τα Μαθηματικά, η Στατιστική κ.α. που επεξεργάζονται και αναλύουν έναν τεράστιο όγκο μοριακών πληροφοριών που αφορούν τόσο το DNA όσο και το RNA δίνοντας πληροφορίες για τις ασθένειες και σχεδιάζοντας αποτελεσματικούς τρόπους βελτίωσης αυτών και θεραπείες. Χαρακτηριστικό παράδειγμα έκρηξης και ανάλυσης δεδομένων αποτελεί και το αντικείμενο μελέτης αυτής της εργασίας, η μέθοδος single-cell RNA-sequencing.

Χάρη στις μεθόδους εύρεσης και ανάλυσης της αλληλουχίας μορίων ζωντανών οργανισμών πραγματοποιήθηκαν ριζοσπαστικές αλλαγές στην επιστήμη της Βιολογίας.

Κάθε ζωντανός οργανισμός είναι μοναδικός. Τι κάνει κάθε οργανισμό μοναδικό; Η μοναδικότητα εντοπίζεται στο γενετικό υλικό του κάθε οργανισμού που καθορίζει τις λειτουργίες του.

1.1 Δομή και λειτουργίες του DNA και του RNA

Η συνεχής ύπαρξη ζωής και η βελτίωσή της οφείλεται στο DNA, δηλαδή στη δομή του και στις ιδιότητές του, στην ικανότητά του να διαφοροποιείται και να εξελίσσεται, να αποθηκεύσει όλες τις γενετικές πληροφορίες που καθορίζουν την μοναδικότητα του οργανισμού, να τις μεταβιβάζει αναλλοίωτες από κύτταρο σε κύτταρο και από οργανισμό σε οργανισμό και να δημιουργεί τα πρωτεϊνικά προϊόντα που συμβάλλουν στην εκτέλεση των διάφορων λειτουργιών του [**1**] .

Το DNA και το RNA είναι νουκλεϊκά οξέα που αποτελούνται από νουκλεοτίδια, τα δεοξυριβονουκλεοτίδια και τα ριβονουκλεοτίδια αντίστοιχα. Αυτά έχουν συγκεκριμένη δομή, για παράδειγμα ένα νουκλεοτίδιο DNA αποτελείται από τρία μέρη : μια πεντόζη που ονομάζεται δεοξυριβόζη, μια φωσφορική ομάδα και μια αζωτούχο βάση η οποία μπορεί να είναι μια εκ των τεσσάρων αδενίνη(A), θυμίνη (T), γουανίνη (G) και κυτοσίνη(C). Πολλά νουκλεοτίδια ενώνονται μεταξύ τους με την βοήθεια ενός φωσφοδιεστερικού δεσμού και σχηματίζουν την πολυνουκλεοτιδική αλυσίδα. Ο προσανατολισμός και η σύνδεση των νουκλεοτιδίων παίζουν σημαντικό ρόλο για την δομή της, ως διπλή έλικα αλλά και για την λειτουργία που επιτελεί. Η έλικα όπως περιγράφεται επιτελεί σπουδαίο ρόλο στην αντιγραφή, την μεταγραφή και την μετάφραση του γενετικού υλικού. Μάλιστα, η δομή της και οι λειτουργίες της αν και θεωρητικά χρειάζεται να μεταφέρονται αναλλοίωτες από κύτταρο σε κύτταρο μέσω των γονιδίων, εντούτοις σε αυτές συμβαίνουν διάφορες μεταλλάξεις σε περιοχές που είτε επηρεάζουν είτε δεν επηρεάζουν απαραίτητα την έκφραση του γενετικού υλικού, συμβάλλοντας έτσι στη διαφοροποίηση, την βελτίωση, και την προσαρμογή του γενετικού υλικού στις εκάστοτε συνθήκες βοηθώντας τον οργανισμό να επιβιώσει. Τελικά δηλαδή, αυτή η ποικιλομορφία που συναντάται στους οργανισμούς συμβάλλει στην εξέλιξη της ζωής [**1**] !

Το RNA συμμετέχει και αυτό καταλυτικά στην διαδικασία της γονιδιακής έκφρασης. Αντίστοιχα με την δομή των δεοξυριβονουκλεοτιδίων τα ριβονουκλεοτίδια αποτελούνται από μια ριβόζη, μια φωσφορική ομάδα και μια αζωζούχο βάση η οποία μπορεί να είναι αδενίνη (A), ουρακίλη (U), γουανίνη (G) ή κυτοσίνη (C) . Αν και δεν σχετίζεται με το γενετικό υλικό (εκτός κάποιων RNA ιών) είναι παρών στις διαδικασίες της μεταγραφής και της μετάφρασης που αποτελούν την γονιδιακή έκφραση, με τη μορφή κυρίως ενζύμων που καταλύουν, επιταχύνουν, διευκολύνουν την διαδικασία να πραγματοποιηθεί όσο το δυνατόν με ακρίβεια, ταχύτητα και ασφάλεια. Τι συμβαίνει κατά τις παραπάνω διαδικασίες [**1**] ;

Σύμφωνα με το κεντρικό δόγμα της Βιολογίας, η γενετική πληροφορία που είναι αποθηκευμένη στο DNA, -για την ακρίβεια στα γονίδια- αντιγράφεται, στην συνέχεια μεταγράφεται, δηλαδή αυτή η γενετική πληροφορία μεταφέρεται στο RNA και στη συνέχεια το RNA μεταφέρεται εκ νέου και μεταφράζεται σε πρωτεΐνες [**57**] . Οι πρωτεΐνες είναι υπεύθυνες για την δομή και την λειτουργία των οργανισμών [**1**] .

1.2 Αντιγραφή

Οι παραπάνω διεργασίες κάθε άλλο παρά απλές δεν χαρακτηρίζονται. Η πολυπλοκότητά τους μπορεί να περιγραφεί τόσο από το πλήθος των παραγόντων που συμβάλλουν σε αυτές όσο και για τον αριθμό των μορίων που συμμετέχουν, την ταχύτητα, την συχνότητα, την ακρίβεια και την δυνατότητα διόρθωσης λαθών. Η αντιγραφή ξεκινά με το ξετύλιγμα της διπλής έλικας και την συμμετοχή των πρωτεϊνών που εκκινούν την διαδικασία δηλαδή την δημιουργία ζευγαρώματος των συμπληρωματικών βάσεων A-T, G-C, T-A, C-G. Η DNA πολυμεράση, είναι ένα ένζυμο, που αναλαμβάνει να ταιριάξει τις αζωτούχες βάσεις του παλαιού κλώνου για την δημιουργία του νέου κλώνου, αντίγραφο του παλαιού. Τα λάθη που μπορεί να συμβούν διορθώνονται σχεδόν κατά 99% από την DNA πολυμεράση είτε συγχρόνως κατά την σύνθεση των νέων αλυσίδων είτε και σε επόμενα στάδια [**1**] .

1.3 Μεταγραφή και Μετάφραση

Πώς όμως όλη η γενετική πληροφορία που βρίσκεται ανάμεσα στην τυχαία σειρά A,T,G,C μπορεί να αποκωδικοποιηθεί; Η απάντηση δίνεται από τον γενετικό κώδικα. Μια τριπλέτα γνωστή και ως κωδικόνιο μεταξύ των A,U,G,C βάσεων μπορεί να αποτελεί το αμινοξύ έναρξης ή το αντίστοιχο της λήξης ή κάποιο από εκείνα που υπάρχουν στον γενετικό κώδικα. Μια αλληλουχία αμινοξέων συνθέτουν μια πρωτεΐνη που ανάλογα την αλληλουχία αυτών στην πολυπεπτιδική αλυσίδα, αυτή έχει διαφορετική δομή και λειτουργία. Για να συμβεί αυτό, ακολουθεί η διαδικασία της μεταγραφής. Τμήματα DNA που περιέχουν γονίδια, δηλαδή περιοχές με σημαντικό για την πρωτεϊνοσύνθεση ενδιαφέρον αντιγράφονται σε RNA. Η συμπληρωματικότητα είναι A-U, T-A, G-C, C-G. Τα ένζυμα που καταλύουν την διεργασία της μεταγραφής ονομάζονται RNA πολυμεράσες. Σε αυτό το σημείο είναι σημαντικό να αναφερθεί το πλήθος και ο ρόλος διάφορων ειδών RNA. Είναι το mRNA που κωδικοποιεί τις πρωτεΐνες, το rRNA που είναι συστατικό των ριβοσωμάτων και καταλύει την σύνθεση των πρωτεϊνών ,το miRNA που ρυθμίζει την γονιδιακή έκφραση ,το tRNA που χρησιμοποιείται στην πρωτεϊνοσύνθεση ως μόριο προσαρμογής μεταξύ του mRNA και των αμινοξέων κ.α. Η διαδικασία της μεταγραφής πραγματοποιείται στον πυρήνα όπου περιοχές με άγνωστες πληροφορίες (εσώνια) αποκόπτονται με την βοήθεια μικρο-πυρηνικού snRNA , και έτσι το ώριμο πλέον mRNA με όλη την γενετική πληροφορία που είναι απαραίτητη εξάγεται από τον πυρήνα με κατεύθυνση προς τα ριβοσωμάτια. Τα ριβοσωμάτια προσδένονται με το μόριο mRNA και σε κάθε βήμα των τριών βάσεων αντιστοιχίζεται στα κωδικόνια, όπου το κάθε κωδικόνιο αντιστοιχίζεται σε ένα

αμινοξύ. Κατά αυτή την διαδικασία της μετάφρασης, μόρια tRNA προσδένονται μεταφέροντας τα αντικωδικόνια στα αντίστοιχα κωδικόνια. Η έναρξη σηματοδοτείται μόνο όταν βρεθεί το κωδικόνιο έναρξης (AUG) και ολοκληρώνεται όταν η RNA πολυμεράση συναντήσει ένα από τα κωδικόνια τερματισμού (UAA, UAG, UGA). Τέλος σημαντικό είναι να αναφερθεί και η rRNA που ενσωματώνει τα αμινοξέα δημιουργώντας την πρωτεΐνη [**1**] .

1.4 Μηχανισμοί και ένζυμα

Στις παραπάνω διαδικασίες γίνεται προσπάθεια να εξασφαλιστεί η πραγματοποίηση λιγότερων λαθών. Για αυτό υπάρχουν τα ένζυμα, τα λεγόμενα επιδιορθωτικά ένζυμα που αναγνωρίζουν και επιδιορθώνουν βλάβες με πολύ υψηλά ποσοστά επιτυχίας. Εκτός από τα λάθη που συμβαίνουν κατά την διάρκεια της αντιγραφής, αλλαγές στην αλληλουχία του DNA μπορεί να συμβούν και από την δράση άλλων παραγόντων. Πολλές από αυτές σχετίζονται προϊόντα της καθημερινότητας π.χ. χημικές ουσίες ή παράγοντες του φυσικού περιβάλλοντος όπως η ακτινοβολία προκαλώντας μεταλλάξεις. Ως μετάλλαξη θεωρείται η μια αλλαγή που συμβαίνει οπουδήποτε στην αλληλουχία του DNA και αυτό το γεγονός επηρεάζει την γενετική πληροφορία και τα προϊόντα της γονιδιακής έκφρασης, τις πρωτεΐνες. Αξίζει να σημειωθεί πως τα κύτταρα διαθέτουν και επιπλέον μηχανισμούς για την παραγωγή των πρωτεϊνών. Μηχανισμοί που ελέγχουν το που, πότε, πώς, πόση ποσότητα και γιατί παράγεται η κάθε πρωτεΐνη. Οποιοδήποτε λάθος κατά την διαδικασία της πρωτεϊνοσύνθεσης μπορεί να προκαλέσει αρνητικές επιπτώσεις [**1**] .

1.5 Ανάγκη εξειδικευμένης μελέτης

Είναι γνωστό το γεγονός ότι όλοι οι κυτταρικοί πληθυσμοί ενός οργανισμού έχουν κοινά χαρακτηριστικά. Κάθε κύτταρο έχει όλο το γενετικό υλικό του οργανισμού. Αυτό είναι οργανωμένο σε μονάδες γνωστές και ως γονίδια τα οποία εντοπίζονται στα χρωμοσώματα. Η μεταγραφή και τελικά η μετάφραση, που αποτελούν την γονιδιακή έκφραση, πραγματοποιούνται σε όλων των ειδών τις κυτταρικές ομάδες όποτε, όπως και με όποιο τρόπο αυτό κρίνεται αναγκαίο ξεχωριστά για κάθε κύτταρο ανάλογα με τις ανάγκες που προκύπτουν. Η συχνότητα, η ποσότητα των πρωτεϊνών που παράγονται και πολλές ακόμη παράμετροι εξαρτώνται τοπικά από τις ανάγκες του κάθε κυττάρου. Για αυτό λοιπόν ήταν επιτακτική ανάγκη να ανακαλυφθεί μια μέθοδος που

θα μπορεί να μελετά και να καταγράφει σε βάθος τι συμβαίνει σε κάθε κύτταρο ξεχωριστά [**1**] .

2 Μέθοδοι μελέτης των αλληλουχιών DNA και RNA

Πολλές σοβαρές ασθένειες και ανωμαλίες που προκαλούνται είτε από τον ίδιο τον οργανισμό, την φύση, είτε λόγω εξωτερικών παραγόντων, είτε λόγω κληρονομικότητας έχουν ως κοινό παρονομαστή τις μικρές εκείνες αλλαγές στο γενετικό υλικό που χρειάζεται να αναλυθούν σε βάθος για να γίνει κατανοητός ο τρόπος που επιδρά αυτή η μικρή ή τελικά μεγάλη διαφοροποίηση, και ως ευρύτερη εικόνα η ασθένεια, ή ο δείκτης που εφιστά προσοχή στον οργανισμό και να σχεδιαστεί η θεραπεία εξασφαλίζοντας στον οργανισμό καλύτερο βιοτικό επίπεδο και ένα ευρύ φάσμα γνώσεων για την καλύτερη αντιμετώπιση και σε επόμενες γενιές. Την προσπάθεια των επιστημόνων για αναλυτικές πληροφορίες επιταχύνουν η ανάλυση DNA sequencing και η πιο εκτεταμένη ανάλυση η RNA sequencing. Τα τελευταία χρόνια ιδιαίτερα γνωστές έγιναν και αναπτύχθηκαν ραγδαία η single-cell RNA-sequencing και η single-nucleus RNA-sequencing. Χαρακτηριστικά παραδείγματα ασθενειών για τα οποία η μέθοδος που μελετάμε θεωρείται πανάκεια είναι μεταξύ άλλων ο καρκίνος, οι αιμοσφαιρινοπάθειες, οι καρδιαγγειακές ασθένειες, ο κορωνοϊός SARS-COVID-19 και η συμβολή της δεν σταματάει σε αυτές αλλά ερευνά και πολλές ακόμη.

2.1 DNA sequencing

Η πρώτη εκτέλεση ανάλυσης DNA sequencing έγινε πίσω στο μακρινό 1970 περίπου. Από τότε μέχρι σήμερα η DNA sequencing εφαρμόζεται με πολλές διαφορετικές μεθόδους σε πολλούς τομείς όπως των κλινικών μελετών, της βιοτεχνολογίας, της εγκληματολογίας κ.α. Ο τρόπος δράσης της αφορά την σύγκριση ενός υγιούς κατά τα άλλα οργανισμού με έναν ασθενή εντοπίζοντας εκείνα τα σημεία που φανερώνουν τα χαρακτηριστικά των δυσλειτουργιών, και στοχεύουν στην δημιουργία εξατομικευμένης θεραπείας για κάθε τι που έχει παρατηρηθεί [**2**] .

2.2 RNA sequencing

Από την άλλη η RNA sequencing είναι μια τεχνική ανάλυσης RNA σε ένα δείγμα μια συγκεκριμένη χρονική στιγμή που μελετά την κατά μέσο όρο έκφραση των γονιδίων μεταξύ όλων των κυττάρων του δείγματος. Μελετάται η διαφορά στην ποσότητα RNA που εντοπίζεται έπειτα από την απομόνωση από όλα τα κύτταρα όλων των

διαφορετικών κατηγοριών του που μελετήθηκαν παραπάνω (mRNA, tRNA, rRNA) και ο τρόπος κατά τον οποίο μεταβάλλονται όσο πραγματοποιούνται εσωτερικά των κυττάρων του δείγματος οι διάφορες διαδικασίες γονιδιακής έκφρασης. Η RNA sequencing έχει τη δυνατότητα να δώσει πληροφορίες για τα περισσότερα στάδια τα οποία ακολουθεί το γονίδιο ως την γονιδιακή έκφραση και τις διαφοροποιήσεις που ενδεχομένως συμβαίνουν, όπως και την καταστροφή των μορίων. Μεταξύ των παραπάνω βρίσκεται και η περίπτωση του single-cell RNA-sequencing [**3**] .

2.3 single-cell RNA-sequencing

Σκοπός της single-cell RNA-sequencing είναι η κατανόηση του ρόλου και της συμπεριφοράς καθενός μεμονωμένου κυττάρου χωριστά μέσα σε ένα δείγμα και η γνώση λεπτομερειών που σχετίζονται με την γονιδιακή έκφραση αυτών. Μέσω απόκτησης της παραπάνω γνώσης οι ερευνητές θα μπορέσουν να χαρτογραφήσουν κάθε κύτταρο και κάθε κυτταρικό τύπο. Κι αυτό κρίνεται αναγκαίο, καθώς η έλλειψη αυτών των λεπτομερειών δυσκολεύει την εύρεση πιο στοχευμένης θεραπείας σε περιπτώσεις όπως αυτή των διάφορων μορφών καρκίνου, των καρδιοπαθειών κ.α. Η RNA-sequencing μελετούσε χονδρικά τα δεδομένα. Η single-cell RNA-sequencing καλείται να αντιμετωπίσει αυτό το κενό στα αποτελέσματα της ανάλυσης [**3**] , [**5**] , [**7**] .

Χαρακτηριστικό παράδειγμα αποτελεί το μικροπεριβάλλον ενός καρκινικού όγκου, καθώς η μεγάλη ετερογένεια μεταξύ των κυτταρικών τύπων που συνθέτουν τον όγκο που οφείλεται στο μεταβολισμό, στην κινητικότητα, στον ρυθμό πολλαπλασιασμού, στην πιθανότητα μετάστασης, στην αλληλεπίδραση των κυττάρων μεταξύ τους κ.α. δεν θα ήταν δυνατόν να αναλυθεί σε βάθος με την RNA-sequencing. Μια κατά μέσο όρο προσέγγιση θα μπορούσε να αποκρύψει μια μικρή πληθυσμιακά κυτταρική ομάδα που παίζει ρόλο ενδεχομένως στην εκδήλωση της συμπεριφοράς που μετατρέπει το περιβάλλον των υγιών κατά τα άλλα κυτταρικών ιστών σε ένα καρκινικό όγκο. Για να εξαλειφθεί κάθε είδους ανακρίβεια χρειάζεται μια λεπτομερέστατη μελέτη, που η single-cell RNA-sequencing μπορεί να φέρει εις πέρας [**4**] .

Πρόκειται λοιπόν για μια πολύ σύγχρονη και πρωτοπόρα μέθοδο ανάλυσης δεδομένων που ήδη από το 2009 που πραγματοποιήθηκε για πρώτη φορά μελέτη σε εργαστήριο έχει κερδίσει έδαφος και η δημοφιλία της μπορεί να μετρηθεί και από το πλήθος των υπολογιστικών εργαλείων, 1314 σε πλήθος, που συνεχώς αυξάνεται , που εκτελούν είτε ολόκληρη την ανάλυση είτε επιλύουν μεμονωμένα κάποια από τα ζητήματα της

όπως την οπτική απεικόνιση των δεδομένων ή τον ποιοτικό έλεγχο κ.α. Δεν είναι λίγα τα περιοδικά που έχουν τιμήσει αυτή την μέθοδο της χρονιάς, όπως το περιοδικό Nature Publishing Group που το 2013 την χαρακτηρίζει ως ορόσημο σημαντικών και τεράστιων αλλαγών, ή και το περιοδικό Science το 2018 που την χαρακτηρίζει ως ιδιαίτερα σημαντική [**5-9**] .

2.4 single-nucleus RNA-sequencing

Η single-nucleus/nuclei RNA-sequencing είναι μια μέθοδος που συμπληρώνει την single-cell RNA-sequencing στις περιπτώσεις των κυττάρων που είναι περίπλοκο να απομονωθούν. Συγκεκριμένα προσπαθεί να απομονώσει το ολικό RNA των πυρήνων των κυττάρων του δείγματος. Χαρακτηριστικά παραδείγματα κυτταρικών πληθυσμών για τα οποία μέσω πειραμάτων αποδείχθηκε ότι η single-cell RNA-sequencing χρειάζεται την παράλληλη στήριξη της single-nucleus RNA-sequencing είναι τα νευρικά κύτταρα, κύτταρα στο συκώτι, στους πνεύμονες και κύτταρα της καρδιάς κ.α. Για αυτό παρακάτω θα γίνει λόγος για ευρήματα μελετών από την καρδιά και για νόσους του κεντρικού νευρικού συστήματος [**10**] , [**11**] , [**12**] , [**13**] .

Έπειτα από μελέτη σχετικών δημοσιευμάτων με τον καρδιακό ιστό, όπου περιγράφονται μια σειρά πειραμάτων και δοκιμών, οι επιστήμονες προσπαθούν να ανακαλύψουν λεπτομέρειες σχετικά με τον τρόπο που εκδηλώνονται διάφορες καρδιοαγγειακές ασθένειες όπως η στεφανιαία νόσος, οι αρρυθμίες, οι δυσλειτουργίες της αορτής κ.α. Τα διαθέσιμα ευρήματα δεν είναι επαρκή, όμως πιστεύεται πως θα έχουν αισιόδοξο αντίκτυπο σε μελλοντικές προσπάθειες. Τα προβλήματα κυρίως επικεντρώνονται στην κατανόηση των καρδιακών λειτουργιών, της ετερογένειας των κυτταρικών τύπων του καρδιοαγγειακού συστήματος και της μεταξύ τους αλληλεπίδρασης με αποτέλεσμα να παρεμποδίζεται η αποκρυπτογράφηση γνωστών και άγνωστων παθογενειών. Ακόμη, μηχανισμοί κυτταρικής διαφοροποίησης και ωρίμανσης της καρδιάς παραμένουν δυσνόητοι. Οι καρδιοαγγειοπάθειες και ο καρκίνος αποτελούν τις 2 πιο γνωστές αιτίες θανάτου παγκοσμίως. Αυτό και μόνο αποτελεί εφαλτήριο για την διεξαγωγή και την έως τώρα προσπάθεια και πρόοδο των ερευνών. Η single-cell RNA-sequencing δεν κατάφερε με επιτυχία να ανιχνεύσει την ύπαρξη ιεραρχίας κατά την γονιδιακή έκφραση, όμως ανακάλυψε κάποιες ρυθμιστικές περιοχές που ελέγχουν την κυτταρική διαίρεση. Τα καρδιακά κύτταρα είναι μεγάλα σε διάμετρο, γεγονός που δυσκολεύει την απομάκρυνση τους. Για αυτό οι επιστήμονες επικεντρώθηκαν στην απομάκρυνση εμβρυϊκών κυττάρων. Η single-nucleus RNA-

sequencing δεν κατάφερε να δώσει βέβαια πληροφορίες για τα ένζυμα και άλλα σημαντικά στοιχεία. Συνεπώς μόνο η συνεργασία αυτών των 2 μεθόδων μπορεί να δώσει πληροφορίες κατά τα στάδια ωρίμανσης της καρδιάς. Η ανατομία και τα περίπλοκα στοιχεία που αποτελούν την καρδιά (φλέβες, αγγεία, αρτηρίες) επίσης δυσκολεύουν την διερεύνηση της [**10**] , [**11**] .

Μια άλλη μελέτη, για την νόσο Parkinson, μας πληροφορεί πως η νόσος οφείλεται σε ένα σύνολο μεταλλάξεων των πρωτεϊνικών μορίων και των γονιδίων που οδηγεί στην καταστροφή των ντοπαμινεργικών νευρώνων. Η εφαρμογή της single-cell RNA-sequencing είναι πολύ περιορισμένη εξαιτίας και της χαμηλής ποιότητας του RNA των δειγμάτων. Ο συνδυασμός όμως της single-cell RNA-sequencing με την single-nucleus RNA-sequencing φέρει νέα δεδομένα καθώς αυτές αξιολογούν την ύπαρξη ετερογενών κυτταρικών πληθυσμών, χαρακτηρίζουν με επιτυχία διαφοροποιημένους κυτταρικούς πληθυσμούς και αναγνωρίζουν την συσχέτιση αυτών με το Parkinson. Το κλειδί των δυο αυτών μεθόδων είναι η μελέτη της ετερογένειας των πληθυσμών που θα βοηθήσει στην κατανόηση σε βάθος της βιολογίας των νευρώνων. Η μελέτη αυτών των δεδομένων έδειξε την συσχέτιση μεταξύ της απώλειας των ντοπαμινεργικών νευρώνων με τους πυρήνες των νευρικών κυττάρων της περιοχής του υποθαλάμου. Οι έρευνες αυτές είναι σημαντικές καθώς αποτελούν την αρχή για να τροποποιηθούν τα πρωτόκολλα καθώς απαιτείται καλύτερη κατανόηση των ντοπαμινεργικών κυττάρων, της νευρογέννεσης και της διαδικασίας της διαφοροποίησης των κυττάρων [**12**] .

Τέλος, μια ακόμη μελέτη σχετικά με τους νευρώνες και την νόσο Αλτσχάιμερ φανερώνει την σημασία της single-nucleus RNA-sequencing ως απαραίτητης συμπληρωματικής μεθόδου. Η νόσος Αλτσχάιμερ είναι μια νευροεκφυλιστική νόσος και έχει ως αποτέλεσμα την καταστροφή των νευρώνων μερικώς ή ολοσχερώς από κάποιες πρωτεΐνες (amyloid beta (Aβ), tau) και από φλεγμονές. Οι μέθοδοι single-cell RNA-sequencing και single-nucleus RNA-sequencing είναι εξαιρετικά χρήσιμες διότι εξετάσουν τις λειτουργίες και τις δυσλειτουργίες των ετερογενών κυτταρικών πληθυσμών στον εγκέφαλο και δίνουν απαντήσεις στο ερώτημα γιατί κάποια κύτταρα είναι ευάλωτα στην νόσο [**13**] .

3 Η διαδικασία single-cell RNA-sequencing

Η διαδικασία ανάλυσης single-cell RNA-sequencing αποτελείται από 3 στάδια. Το πρώτο στάδιο, η προετοιμασία ή προεπεξεργασία των δεδομένων αφορά την απομόνωση του δείγματος ενδιαφέροντος και την προετοιμασία του για ανάλυση. Έπειτα σε ένα δεύτερο στάδιο πραγματοποιείται ο ποιοτικός έλεγχος, η κανονικοποίηση, η μείωση της διάστασης των δεδομένων για απεικόνιση και πολλές ακόμη μορφές κατωφλίων και φίλτρων. Το τελευταίο στάδιο αφορά την συσταδοποίηση, την ταυτοποίηση των κυττάρων του δείγματος, την ανίχνευση διαφοροποιημένων κυττάρων, την οπτική απεικόνιση και πολλές ακόμη επιλογές. Παρακάτω θα γίνει διεξοδική επεξήγηση όλων των όρων αυτής της παραγράφου [**15**] .

3.1 **Πρώτο στάδιο**

3.1.1 **Μέθοδοι συλλογής κυτταρικών δειγμάτων**

Ιδιαίτερα δημοφιλής τρόποι απομόνωσης κυτταρικών δειγμάτων αποτελούν η κυτταρομετρία ροής ή αλλιώς γνωστή στην επιστημονική κοινότητα ως FACS (Fluorescence-activated cell sorting) και η Droplet barcoding ή αλλιώς γνωστή ως droplet-based.

Η κυτταρομετρία ροής (FACS) αφορά την μέτρηση και τον χαρακτηρισμό σωματιδίων μέσα σε υγρό περιβάλλον δηλαδή ξεχωρίζει, ταξινομεί τις ετερογενείς ομάδες με χαμηλό κόστος, μεγάλη ταχύτητα όμως έχει και μεγάλη πιθανότητα καταστροφής. Επιτρέπει την ταυτόχρονη ανάλυση πολλών μεταβλητών των κυττάρων που μελετώνται τα οποία ρέουν μέσω μιας συσκευής. Με λίγα λόγια, ένα σύνολο ανιχνευτών βρίσκεται στο σημείο όπου ρέουν τα σωματίδια και σκεδάζουν το φως που ρίχνεται από δέσμη φωτός. Αυτά τα ίχνη φωτός ανιχνεύονται από τους ανιχνευτές και δίνουν λεπτομέρειες σχετικά με την φυσική και χημική κατάσταση του κυττάρου, τον όγκο, το σχήμα του πυρήνα και πολλά ακόμη [**16**] .

Η μέθοδος Droplet barcoding βασίζεται στην διαίρεση της ροής (με μια συγκεκριμένη σταθερή ταχύτητα ροής) του δείγματος σε σταγόνες που επιτρέπει την μελέτη των

διάφορων γονιδίων και των μοριακών διεργασιών μεγάλου δείγματος κυττάρων που μπορεί να κυμαίνεται ακόμη και στα 10000-20000 κύτταρα [**17**] .

3.1.2 **Δημιουργία cDNA βιβλιοθήκης και κλωνοποίηση**

Για να πραγματοποιηθεί η single-cell RNA-sequencing διαδικασία είναι απαραίτητο το cDNA. Τι είναι το cDNA; Μια αλυσίδα cDNA δημιουργείται με βάση ένα ώριμο μόριο mRNA σύμφωνα με την συμπληρωματικότητα των βάσεων. Η αντίστροφη μεταγραφάση είναι το ένζυμο που πραγματοποιεί την σύνθεση της. Έτσι, δημιουργούνται υβριδικά μόρια cDNA-mRNA. Σε αυτό ακριβώς το σημείο θα γίνει λόγος για το κεντρικό δόγμα Βιολογίας, όχι όμως αυτό που αναφέρθηκε παραπάνω, το πρωταρχικό, αλλά αυτό για το οποίο ερευνητές ανακάλυψαν αργότερα πως η πορεία της γενετικής πληροφορίας δεν είναι μόνο μονόδρομη, αλλά μπορεί να είναι και αμφίδρομη, δηλαδή ότι και μερικοί ιοί που έχουν RNA ως γενετικό υλικό διαθέτουν το ένζυμο αντίστροφη μεταγραφάση που χρησιμοποιείται το RNA ως πρότυπο για να συνθέσει DNA. Αυτό αποτελεί το σύγχρονο κεντρικό δόγμα της Βιολογίας [**1**] , [**18**] .

Αναλυτικότερα, η διαδικασία single-cell RNA-sequencing απαιτεί την δημιουργία μιας cDNA βιβλιοθήκης. Για να κατασκευαστεί μια cDNA βιβλιοθήκη συμβαίνουν τα εξής:

- Αρχικά απομονώνεται το ολικό ώριμο mRNA του κυτταροπλάσματος των κυττάρων εκείνων που έχουν επιλεχθεί για μελέτη του εκάστοτε ζητήματος.
- Έπειτα το mRNA χρησιμοποιείται σαν πρότυπο για την δημιουργία του συμπληρωματικού DNA (cDNA) με την βοήθεια του ενζύμου αντίστροφη μεταγραφάση. Έτσι προκύπτουν τα υβριδικά μόρια cDNA-mRNA.
- Στη συνέχεια το mRNA αποδιατάσσεται με την χρήση διάφορων χημικών ουσιών ή την αξιοποίηση της θερμότητας.
- Έπειτα το cDNA κι αυτό με την σειρά του αποτελεί πρότυπο για την δημιουργία μιας συμπληρωματικής ως προς αυτό αλυσίδας DNA και με την βοήθεια του ενζύμου DNA πολυμεράση τελικά θα προκύψουν δίκλωνα μόρια DNA. Αυτά, εισάγονται μέσα σε πλασμίδια ή βακτηριοφάγους τα οποία τέμνονται σε συγκεκριμένα σημεία και μόνο μια φορά με την βοήθεια

των περιοριστικών ενδονουκλεασών (ενζύμων) και έπειτα μέσα σε αυτά ,τα μόρια κλωνοποιούνται. Ένζυμα αναλαμβάνουν να συνδέσουν τα πλασμίδια με τα δίκλωνα μόρια DNA . Οι κλώνοι των βακτηρίων που δημιουργούνται έχουν την ικανότητα να συνθέσουν το πρωτεϊνικό προϊόν που εκφράζεται από το συγκεκριμένο γονίδιο που βρίσκεται πλέον στο κύτταρο-ξενιστή. Στο σημείο αυτό χρειάζεται να αναλυθεί η διαδικασία της κλωνοποίησης μέσα στα πλασμίδια και τους βακτηριοφάγους [**18**] , [**19**] .

Τα πλασμίδια με τα δίκλωνα μόρια DNA είναι πλέον έτοιμα για κλωνοποίηση, δηλαδή να πολλαπλασιαστούν. Η κλωνοποίηση χαρακτηρίζεται ως μια διαδικασία κατασκευής πολλών πανομοιότυπων μορίων. Για την υλοποίηση μιας τέτοιας κατασκευής κρίνεται απαραίτητη η παρουσία ενός φορέα κλωνοποίησης (συνήθως χρησιμοποιούνται πλασμίδια ή βακτηριοφάγοι) τα οποία έχουν την ικανότητα να πολλαπλασιάζονται ανεξάρτητα από το κύτταρο-ξενιστή. Συνεπώς χάρη σε αυτή την ιδιότητα ο πολλαπλασιασμός οποιουδήποτε μορίου προς μελέτη καθίσταται ευκολότερος και αποτελεσματικός. Όσο περισσότερα τα αντίγραφα του μορίου τόσο πιο ποιοτικά αποτελέσματα μπορούν να εξαχθούν καθώς κατά την επεξεργασία του δείγματος πολλά μόρια καταστρέφονται. Οι φορείς κλωνοποίησης ενσωματώνονται με τα μόρια που μελετώνται αφού αυτά παρεμβάλλονται στο σημείο που έχει γίνει η τομή από το ένζυμο περιοριστική ενδονουκλεάση. Τα πλασμίδια μάλιστα αποτελούν ιδιαίτερα σημαντικούς φορείς κλωνοποίησης καθώς έχουν τη δυνατότητα να εισαχθούν με ευκολία σε ένα κύτταρο-ξενιστή. Αυτό συμβαίνει γιατί τα πλασμίδια είναι μικρά μόρια DNA με γονίδια ανθεκτικά σε περιβάλλον με παρουσία αντιβιοτικών, και μάλιστα περιέχουν και πρωτεΐνες που καταστρέφουν ή απενεργοποιούν τα αντιβιοτικά. Τέτοιες συνθήκες δυσκολεύονται να αντιμετωπίσουν με μεγάλη επιτυχία τα βακτήρια. Τα γονίδια ανθεκτικότητας των πλασμιδίων επιτρέπουν την επιβίωση μόνο των βακτηρίων που δέχτηκαν ανασυνδυασμένα πλασμίδια, δηλαδή πλασμίδια που φέρουν πλέον το μόριο που χρειάζεται να διπλασιαστεί. Κατά την διάρκεια του μετασχηματισμού (είσοδος ανασυνδυασμένου DNA, δηλαδή φορέας κλωνοποίησης μαζί με το μόριο εισέρχονται σε κύτταρο-ξενιστή) μπορεί τα κύτταρα- ξενιστές π.χ. βακτήρια να έχουν πλασμίδια με ανασυνδυασμένο DNA είτε πλασμίδια με μη ανασυνδυασμένο DNA είτε να μην δέχτηκαν καθόλου πλασμίδια. Μεταξύ όλων αυτών, σε μια καλλιέργεια σε θρεπτικό υλικό και με την παρουσία κάποιου αντιβιοτικού θα επιβιώσουν μόνο τα πλασμίδια με το ανασυνδυασμένο μόριο DNA ακριβώς επειδή διαθέτουν εκείνα τα

γονίδια που τα κάνουν ανθεκτικά στο αντιβιοτικό. Έτσι, θα εξασφαλιστεί η επιβίωση και ο αναδιπλασιασμός κυρίως των μορίων που χρειάζεται να μελετηθούν [**18**] , [**19**] .

3.1.3 Συνοπτική παρουσίαση των βημάτων κατά την εκτέλεση προετοιμασίας ανάλυσης κυττάρων από ασθενείς με COVID-19

Σύμφωνα με μια δημοσίευση στην NCBI παρουσιάζονται αναλυτικά όλα τα βήματα ανάλυσης και προετοιμασίας του κυτταρικού δείγματος πριν αυτό μεταφερθεί για την single-cell RNA-sequencing ανάλυση στις υπολογιστικές πλατφόρμες ανάλυσης και διερεύνησης των δεδομένων. Οι *Yao, C., Bora, S. A., Chen, P., Goodridge, H. S., & Gharib, S. A. (2021)* μελέτησαν δείγματα από ασθενείς με COVID-19 που εκδήλωσαν ήπια συμπτώματα, που νόσησαν βαριά και που έχουν πλέον αναρρώσει. Τα κύτταρα του αίματος είναι πλούσια σε αντισώματα καθώς και άλλα στοιχεία όπως ιοί που μπορεί να δυσκολέψουν την μελέτη. Οι προκλήσεις αυτές όπως και πολλές ακόμη που αναφέρονται εκτενέστερα παρακάτω σε ξεχωριστό κεφάλαιο χρειάζεται να ληφθούν σοβαρά υπόψη έτσι ώστε η έρευνα να δώσει ποιοτικά αποτελέσματα για περαιτέρω ανάλυση [**20**] .

Συνοπτικά, τα βήματα όπως παρουσιάζονται από τους *Yao, C., Bora, S. A., Chen, P., Goodridge, H. S., & Gharib, S. A. (2021)* είναι τα εξής:

1ο στάδιο: ΠΡΙΝ (διάρκεια περίπου 6-8 ώρες)

- Συλλογή αίματος
- Φυγοκέντριση (διαδικασία διαχωρισμού μιγμάτων)
- Συλλογή ποσότητας < 1% του αίματος που περιλαμβάνει τα λευκά αιμοσφαίρια και τα αιμοπετάλια
- Αποθήκευση στους -80 βαθμούς Κελσίου

Τα δείγματα από την στιγμή της λήψης από τον ασθενή πρέπει αμέσως και με προσοχή να ακολουθήσουν την παραπάνω διαδικασία έτσι ώστε να εξασφαλιστεί ότι αυτά σε επόμενα στάδια θα διατηρήσουν τα χαρακτηριστικά τους, θα είναι δηλαδή ζωντανά και δεν θα χαθεί κάποια πληροφορία [**20**] .

2ο στάδιο: ΤΑΞΙΝΟΜΗΣΗ (διάρκεια περίπου 30 λεπτά/ δείγμα)

Ο χρόνος ταξινόμησης ποικίλλει ανάλογα το χρόνο που χρειάζεται να ξεπαγώσει το δείγμα προς ταξινόμηση. Αυτό το στάδιο περιλαμβάνει απόψυξη, καθαρισμό του δείγματος και ταξινόμηση.

- Συλλογή παγωμένου δείγματος
- Απόψυξη κάθε δείγματος χωριστά στους 37 βαθμούς Κελσίου σε νερό μέχρι να μείνει ελάχιστος πάγος
- Φυγοκέντριση και καθαρισμός του δείγματος μετά
- Προσθήκη DAPI (φθορίζοντα στοιχεία)
- Ταξινόμηση

Όσα κύτταρα έχουν παραμείνει ζωντανά, είναι σε ποσότητα πολύ λιγότερα από όσα απομονώθηκαν στην αρχή στο 1ο στάδιο. Αρκετά από αυτά καταστρέφονται κατά την διάρκεια. Εν τέλη μετά την απόψυξη και την ταξινόμηση ο αριθμός των κυττάρων υπολογίζεται γύρω στις 200 – 250.000 [**20**] .

3ο στάδιο: 10X Genomics (διάρκεια περίπου 1 ώρα)

- Προετοιμασία του buffer (ρυθμιστικό διάλυμα) στους 4 βαθμούς Κελσίου
- Δεν πρέπει να υπάρξει καθυστέρηση για να μην καταστραφεί το RNA
- Είναι προτιμότερο να υπάρχουν περισσότερα δείγματα εξ' αρχής διότι έτσι διασφαλίζεται ένα περιθώριο μελέτης επαρκούς δείγματος ακόμη κι αν χαθούν περισσότερα κύτταρα από όσα υπολογίζονται
- Φυγοκέντριση, καθαρισμός και προσθήκη DBPS (διατήρηση των συνθηκών του περιβάλλοντος μέσα στο ρυθμιστικό διάλυμα)
- Ψύξη στους -20 βαθμούς Κελσίου για 30 λεπτά

4ο στάδιο: ΕΝΥΔΑΤΩΣΗ (διάρκεια περίπου 10-15 λεπτά)

- Προετοιμασία του 10X Genomics
- Ψύξη
- Φυγοκέντριση
- Καθαρισμός του δείγματος (φυλάσσεται)
- Μέτρηση των κυττάρων (πιθανότητα να χαθεί το 50%)

5ο στάδιο: single-cell RNA sequencing

- Φιλτράρισμα (επιλογή και απομάκρυνση κυττάρων με πολύ λίγα ή πολλά περισσότερα γονίδια από όσα πραγματικά χρειάζεται να αναλυθούν)
- Χρήση ειδικών υπολογιστικών εργαλείων για κανονικοποίηση, μετάβαση των δεδομένων σε άλλη διάσταση
- Clustering (=συσταδοποίηση)
- Ποσοτικοποίηση των ταυτοποιημένων δεδομένων με τον κωδικό που αποκτούν οι αλληλουχίες
- Με τη βοήθεια του λογισμικού που περιλαμβάνει βιβλιοθήκες γονιδίων και γενετικό υλικό του SARS-Cov-2 γίνεται το ταίριασμα.
- Οπτικοποίηση των δεδομένων που προέκυψαν από την ανάλυση"

Είναι σημαντικό να αναφερθεί πως τα παραπάνω βήματα, στάδια 1-5 έχουν αποτελέσει υλικό μελέτης και καταγραφής σε σχετική δημοσίευση από τους *Yao, C., Bora, S. A., Chen, P., Goodridge, H. S., & Gharib, S. A. (2021)* [**20**] .

3.1.4 Γιατί η single-cell-RNA-sequencing διαφοροποιείται κατά την μελέτη κυττάρων από καρδιακό ιστό;

Αν και τα βασικά βήματα της διαδικασίας single-cell-RNA-sequencing και μερικές τεχνικές που χρησιμοποιούνται παραμένουν οι ίδιες , όπως διεξοδικά καταγράφονται παραπάνω, για όλων των ειδών τα κύτταρα, εντούτοις υπάρχουν τεχνικές που διαφοροποιούνται για ομάδες κυττάρων όπως χαρακτηριστικά αναφέρονται ο καρδιακός ιστός, κύτταρα από το συκώτι, τους πνεύμονες και κύτταρα του νευρικού συστήματος, και χρειάζεται να αναλυθούν εκτενέστερα παρακάτω [**11**] . Οι *Wang, M., Gu, M., Liu, L., Liu, Y., & Tian, L. (2021)* μελέτησαν και παρουσίασαν κάποιες λεπτομέρειες που θα περιγραφούν σε αυτή την υποενότητα.

Όπως σε κάθε άλλη περίπτωση η αρχική σωστή προετοιμασία του δείγματος κάτω από τις κατάλληλες συνθήκες μπορεί να δώσει υψηλής ποιότητας αποτελέσματα. Τα πρωτόκολλα είναι ένα μέρος της μεθόδου single-cell RNA-sequencing που αλλάζει συνεχώς υπό το πρίσμα περισσότερων παραγόντων (όπως το σχήμα των κυττάρων, η βιωσιμότητα των κυττάρων κ.α.). Η απομόνωση κυττάρων με την μέθοδο FACS που χρησιμοποιείται από πολλά

υπολογιστικά εργαλεία όπως το Chronium (10X Genomics) “αφορά κύτταρα μικρού μεγέθους μόλις 30μm, για τα καρδιομυϊκά κύτταρα όμως που είναι μεγαλύτερα (πάνω από 130 μm) διαφοροποιείται το ακροφύσιο που χρησιμοποιείται για την μεταφορά των κυττάρων στη δοκιμαστική βάση” *Wang, M., Gu, M., Liu, L., Liu, Y., & Tian, L. (2021)* . Η FACS μέθοδος διευκολύνει την παράλληλη μελέτη πολλών κυττάρων διαφορετικών μεγεθών μεταξύ τους δίνοντας λεπτομέρειες για τα γονιδιακά προφίλ αυτών κάτι το οποίο ήταν αδύνατο για τα καρδιομυϊκά κύτταρα με την μέθοδο droplet-barcode-based που απελευθερώνει μέρος του δείγματος με συγκεκριμένο ρυθμό ροής. Η μέθοδος single-nucleus RNA-sequencing χρησιμοποιείται ως εναλλακτική και συμπληρωματική της single-cell RNA-sequencing διότι τα κύτταρα του καρδιακού ιστού είναι δύσκολο να απομονωθούν και με αυτή τη μέθοδο απομονώνονται οι πυρήνες τους και το RNA του. Στη συνέχεια ακολουθούν πολλές επαναλήψεις πολλαπλασιασμού των κυττάρων με PCR και κατασκευάζονται οι cDNA βιβλιοθήκες. Τα δεδομένα που παράγονται αποθηκεύονται σε πίνακες όπου οι γραμμές αναπαριστούν τα γονίδια και οι στήλες το κάθε κύτταρο και διαβάζονται από ειδικές πλατφόρμες. Αυτά αναπαρίστανται σε γράφημα που προκύπτει από την εφαρμογή του αλγορίθμων συσταδοποίησης και τέλος με την μέθοδο tSNE για την οποία γίνεται λόγος παρακάτω και οπτικοποιούνται. Συνεπώς οι ιδιαίτερες διαφορές της μελέτης του καρδιακού ιστού αφορούν κυρίως το αρχικό στάδιο απομόνωσης και αποθήκευσης των κυττάρων λόγω του μεγέθους τους και την συμμετοχή της μεθόδου single-nucleus RNA-sequencing για την απομόνωση του συνολικού RNA του πυρήνα. Παρά την αντικατάσταση με μεθόδους περισσότερο αποδοτικές που στοχεύουν στην ασφαλή διεξαγωγή των αρχικών βημάτων για να μην προκληθεί ζημιά στα κύτταρα και χαθούν πληροφορίες, εντούτοις οι προκλήσεις συνεχίζουν να υπάρχουν. “Η single-cell-RNA-sequencing δεν μπορεί να εντοπίσει μέρη του συνόλου του RNA όταν αυτά βρίσκονται σε μικρή ποσότητα στο δείγμα. Μόλις 10% του συνόλου του RNA μπορεί να ανιχνευτεί για κάθε κύτταρο και το ποσοστό RNA πληροφορίας που χάνεται (dropouts) αγγίζει το 60% ανά κύτταρο. Τα χαμηλά επίπεδα RNA που ανιχνεύονται δημιουργούν υψηλά επίπεδα θορύβου στα αποτελέσματα” *Wang, M., Gu, M., Liu, L., Liu, Y., & Tian, L. (2021)* .Τέλος, εκτός αυτών, η χαρτογράφηση κατά την γραμμική ανάπτυξη των διάφορων κυτταρικών τύπων της καρδιάς που θα

μπορούσε να δώσει βάθος στην μελέτη αυτών για το τι συμβαίνει αλλά είναι ακόμα σε εξέλιξη [**11**] .

3.2 Δεύτερο στάδιο

Στο στάδιο αυτό πραγματοποιείται ένα φιλτράρισμα των δεδομένων, όπου αντιμετωπίζονται και αφαιρούνται κάποια στοιχεία που μόνο προβληματικά μπορεί να αποδειχθούν για τα δεδομένα προκαλώντας θόρυβο. Ακόμη αφαιρούνται δεδομένα που δεν χρειάζεται να αναλυθούν σε κάθε περίπτωση. Κάθε πείραμα και μελέτη διεξάγεται έχοντας μια συγκεκριμένη υπόθεση και ένα πλαίσιο μέσα στο οποίο τίθενται όρια. Αυτά τα όρια/παράμετροι διευκολύνουν την εξαγωγή των αποτελεσμάτων.

3.2.1 Πίνακας δεδομένων

Αρχικά λοιπόν δημιουργείται ο πίνακας δεδομένων (συνήθως περιλαμβάνει τα ονόματα των αντίστοιχων γονιδίων του κάθε δείγματος και το πόσα από το καθένα γονίδια παρατηρήθηκε σε κάθε κύτταρο του δείγματος). Κάθε μόριο αποκτά ένα μοναδικό κωδικό που το χαρακτηρίζει γνωστό ως (Unique molecular identifier ή αλλιώς UMI) ο οποίος χρησιμεύει κατά την καταμέτρηση των γονιδίων σε κάθε κύτταρο. Με άλλα λόγια, τα κύτταρα έπειτα από τον πολλαπλασιασμό τους καταμετρώνται και ταξινομούνται σε ένα πίνακα. Εάν σε ένα κύτταρο, κάποιες από τις αλληλουχίες των μερών του RNA έχουν τον ίδιο κωδικό UMI αυτό σημαίνει ότι το γονίδιο προέρχεται από την ίδια αλληλουχία RNA και μετράται ως 1 γονίδιο. Από την άλλη εάν ένα γονίδιο του ίδιου κυττάρου παρουσιάζει 2 διαφορετικούς κωδικούς UMI τότε θα καταμετρηθούν ως 2 διαφορετικά στοιχεία που προήλθαν από διαφορετικά μόρια RNA και έτσι θα καταγραφούν στη θέση του πίνακα για το πλήθος των γονιδίων αντίστοιχα. Σε αυτό το στάδιο μπορεί να δημιουργηθούν διπλότυπα οδηγώντας σε λανθασμένα συμπεράσματα τον ερευνητή [**21-23**] .

3.2.2 Ποιοτικός έλεγχος

Στη συνέχεια πραγματοποιείται ο ποιοτικός έλεγχος (quality control). Επίσης ένα σημαντικό βήμα που ασχολείται κυρίως με 3 παραμέτρους. Αφορά τον συνολικό αριθμό UMIs που έχουν καταγραφεί ανά κύτταρο, τον αριθμό των γονιδίων που εκφράζονται ανά κύτταρο και το ποσοστό των μιτοχονδριακών γονιδίων που εκφράζονται. Υψηλά ποσοστά μιτοχονδριακών γονιδίων εκτός του ότι προκαλούν

σύγχυση μεταξύ των υπόλοιπων δεδομένων αποτελούν και δείκτη χαμηλής ποιότητας ή δηλώνουν πως τα κύτταρα βρίσκονται σε κατάσταση απόπτωσης, δηλαδή σε κατάσταση προγραμματισμένου θανάτου. Κατά τον ποιοτικό έλεγχο, τίθενται κατώφλια στις παραπάνω 3 παραμέτρους έτσι ώστε το δείγμα να είναι έτοιμο με όσο το δυνατόν χρήσιμη πληροφορία προς απεικόνιση [**15**] , [**24-25**] .

3.2.3 Κανονικοποίηση

Έπειτα πραγματοποιείται κανονικοποίηση (normalization). Η κανονικοποίηση είναι μια διαδικασία διόρθωσης των διαφορών που εντοπίζονται στις τιμές των παραμέτρων των γονιδίων και άλλων στοιχείων μεταξύ των κυττάρων του δείγματος. Τα δεδομένα κατά την κανονικοποίηση μετασχηματίζονται με τον λογάριθμο log(). Η κανονικοποίηση κρίνεται απαραίτητη έτσι ώστε οι τιμές των δεδομένων που αρχικά εμφανίζουν μεγάλες αποκλίσεις να είναι συγκρίσιμές μεταξύ τους [**15**] , [**26**] .

3.2.4 Μείωση διαστάσεων δεδομένων

Μια ακόμη διαδικασία ακολουθεί η οποία σχετίζεται με την μείωση των διαστάσεων των δεδομένων συνήθως από τις n διαστάσεις στις 2. Χαρακτηριστικός αλγόριθμος για μείωση της διάστασης είναι ο t-SNE(t-Distributed Stochastic Neighbor Embedding). Ο αλγόριθμος βασίζεται στην απόσταση των σημείων μεταξύ τους. Έστω, ένα παράδειγμα, ότι έχω ένα γράφημα 2D που απεικονίζει 2 συστάδες δεδομένων και χρειάζεται να μεταβώ σε ένα γράφημα 1D. Τα δεδομένα της ίδιας συστάδας που τείνουν να πλησιάζουν μεταξύ τους, θα πλησιάζουν και στο γράφημα μειωμένης διάστασης. Τα σημεία των συστάδων που είναι απομακρυσμένα θα τείνουν να είναι απομακρυσμένα μεταξύ τους στο γράφημα μειωμένης διάστασης. Τα βήματα του αλγορίθμου είναι τα εξής:

- Εύρεση της απόστασης μεταξύ των σημείων. Ο αλγόριθμος επαναλαμβάνεται ώστε για όλα τα σημεία να βρεθεί η μεταξύ τους απόσταση. Από τις μετρήσεις τα απομακρυσμένα σημεία έχουν χαμηλές ομοιότητες μεταξύ των τιμών των παραμέτρων ενώ αυτά σε κοντινή απόσταση εμφανίζουν μεγάλη ομοιότητα ως προς τις τιμές των παραμέτρων. Οι τιμές αποθηκεύονται σε έναν πίνακα τιμών.

- Το πρώτο βήμα επαναλαμβάνεται, μόνο που διαφοροποιείται η καμπύλη και αλλάζουν οι τιμές των αποστάσεων των σημείων και άρα αλλάζει ο πίνακας τιμών.
- Στόχος του τελευταίου βήματος είναι να τροποποιηθεί ο πίνακας του δεύτερου βήματος και να πάρει την μορφή αυτού στο πρώτο βήμα. Συνεπώς η αναπαράσταση των σημείων στο γράφημα με την μικρότερη διάσταση να αποτυπώνει τις κινήσεις των σημείων ώστε οι πίνακες τιμών να γίνουν όμοιοι [**27-29**] .

Σχετικά με την μέθοδο t-SNE, όταν αναφέρεται για δεδομένα μεγάλης διάστασης, αυτό σχετίζεται με το πλήθος των συστάδων στις οποίες έχουν μοιραστεί τα δεδομένα. Όταν πρόκειται για 3D και πάνω δεδομένα οι συνθήκες αναπαράστασης σε γράφημα γίνονται πιο περίπλοκες και χρειάζονται επιπλέον δεδομένα όπως για παράδειγμα η γωνία απεικόνισης. Με άλλα λόγια ο δείκτης της διάστασης δείχνει τα γνωρίσματα που έχει κάθε σημείο και τα οποία πρέπει να αναπαρασταθούν. Συνήθως όποια κι αν είναι η διάσταση των δεδομένων εισόδου, η απεικόνιση είναι προτιμότερο να γίνεται σε 2D διαστάσεων γράφημα. Η βασική ιδέα του αλγορίθμου είναι να διατηρήσει την απόσταση των δεδομένων μεταξύ τους [**27-29**] .

3.3 Τελευταίο στάδιο

Το τελευταίο στάδιο ανάλυσης δεδομένων σχετίζεται με τα υπολογιστικά εργαλεία πλατφόρμες, δηλαδή ανάλυση και εξαγωγή αποτελεσμάτων που αφορούν την ταξινόμηση των κυτταρικών τύπων σε ομάδες (clustering), την ταυτοποίηση (annotation) αυτών, την εύρεση των σταδίων ζωής των κυττάρων του δείγματος (trajectory analysis), την οπτικοποίηση (visualization) των αποτελεσμάτων σε ιστογράμματα (histograms), γραφήματα (graphs), violin plots, χαρτογράφηση (maps, heatmaps), εύρεση εκφράσεων των γονιδίων ανά το κυτταρικό δείγμα, εύρεση διαφοροποιημένων εκφράσεων (differential expression) και πολλές ακόμη συγκριτικές μεταξύ των τιμών και μη επιλογές που δίνουν σημαντικές πληροφορίες για την κατάσταση του δείγματος.

3.3.1 Συσταδοποίηση

Η συσταδοποίηση ή αλλιώς clustering είναι μια διαδικασία κατά την οποία ένα σύνολο από "αντικείμενα", στην προκειμένη περίπτωση ένα σύνολο κυττάρων, διαχωρίζεται σε ένα σύνολο από λογικές ομάδες/ τις συστάδες/ τους κυτταρικούς πληθυσμούς που χρειάζεται να μελετήσουν οι ερευνητές. Συχνά η συσταδοποίηση συναντάται στη βιβλιογραφία και ως μη επιβλεπόμενη μάθηση. Στη μη επιβλεπόμενη μάθηση ή αλλιώς μάθηση χωρίς επίβλεψη (unsupervised learning) οι ομάδες στις οποίες θα χωριστούν τα δεδομένα δεν είναι γνωστές από την αρχή και η ομοιότητα ή μη των αντικειμένων που θα διαχωριστούν εξαρτάται κάθε φορά από τις συνθήκες του εκάστοτε προβλήματος, λόγω χάρη τα χαρακτηριστικά του κάθε κυτταρικού τύπου, ο αριθμός των γονιδίων, ποιες και πόσες πρωτεΐνες εκφράζονται από τα γονίδια κ.α. αποτελούν κριτήριο διαχωρισμού στην single-cell RNA sequencing όπως διαπιστώθηκε από τα παραδείγματα που αναφέρονται παραπάνω [**30**] .

Τα βασικά βήματα της συσταδοποίησης είναι τα εξής:

- Η επιλογή χαρακτηριστικών γνωρισμάτων/ ιδιοτήτων με τα οποία η συσταδοποίηση θα επιτύχει την καλύτερη δυνατή ομοιογένεια κάθε συστάδας
- Η εφαρμογή των αλγορίθμων συσταδοποίησης στα δεδομένα στοχεύει στην επαναληπτική εφαρμογή βημάτων που συγκρίνει κάθε ένα από τα δεδομένα και τα εντάσσει στην αντίστοιχη συστάδα με την οποία εμφανίζει και την μεγαλύτερη ομοιότητα. Αρκετά γνωστός αλγόριθμος συσταδοποίησης είναι ο k-means
- Αξιολόγηση και επιβεβαίωση της ορθότητας των αποτελεσμάτων, δηλαδή κατά πόσο έχει επιτευχθεί η ομοιότητα όλων των αντικειμένων κάθε συστάδας [**30**] .

Μερικά παραδείγματα αλγορίθμων συσταδοποίησης που συναντώνται ανά τις πλατφόρμες ανάλυσης δεδομένων είναι ο Hierarchical Agglomerative Clustering, ο k-nearest-neighbor και ο k-means [**30**] .

Ο αλγόριθμος Hierarchical Agglomerative Clustering (HAC) που χρησιμοποιείται από την πλατφόρμα κατά το στάδιο της συσταδοποίησης αφορά την ιεραρχία (δείχνει δηλαδή πως είναι οργανωμένα τα δεδομένα) μεταξύ των συστάδων. Κάθε παρατήρηση ξεκινάει από τη δική της συστάδα και ένα ζευγάρι συστάδων. Η παράμετρος που χρησιμοποιεί είναι η απόσταση μεταξύ συστάδων. Αν τα δεδομένα χρήζουν μιας άλλη

προσέγγισης μπορεί ο ερευνητής να θέσει κάποιο κατώφλι (threshold). Τα βασικά βήματα του αλγορίθμου HAC είναι:

- Κάθε συστάδα αποτελείται από ένα σημείο,
- Βρες τα σημεία που έχουν την πιο μικρή απόσταση μεταξύ τους,
- Τα σημεία με την μικρότερη απόσταση αποτελούν μια συστάδα,
- Σε μικρότερης κλίμακας δεδομένα κατά την επανάληψη των παραπάνω βημάτων και την σύμπτυξη των συστάδων με άλλες συστάδες δημιουργείται ένα δεντρόγραμμα [**30**] .

Ο αλγόριθμος k-nearest-neighbor χρησιμοποιείται για την ταξινόμηση των σημείων που αναπαριστούν ομάδες κυττάρων σε συστάδες. Ο τρόπος λειτουργίας του είναι απλός και περιγράφεται ως εξής:

- Έστω ότι επιλέγω το κύτταρο/σημείο που χρειάζεται να ελέγξω σε ποια συστάδα ανήκει,
- Βρίσκω ποια κύτταρα/σημεία είναι πιο κοντά σε αυτό,
- Ανάλογα το πλήθος αυτών ο αλγόριθμος αποφασίζει σε ποια συστάδα ανήκει,
- Αποφασίζει ότι ανήκει σε αυτή την συστάδα με την οποία το κύτταρο/σημείο βρίσκεται κοντά με τα περισσότερα κύτταρα/σημεία της.

Σημαντικές παρατηρήσεις σχετικά με τον αλγόριθμο k-nearest-neighbor που πρέπει να αναφερθούν είναι ότι:

- Έστω κ ορίζω το πλήθος των στοιχείων,
- Το κ δεν πρέπει να έχει ίδια τιμή με το πλήθος των ομάδων που θα ταξινομηθούν τα στοιχεία [**30**] , [**31**] .

Το μειονέκτημα του k-nearest-neighbor είναι η πολυπλοκότητα του να ψάχνει το γειτονικό ή τα γειτονικά στοιχεία για κάθε σημείο/δείγμα. Στον συγκεκριμένο αλγόριθμο η παράμετρος της απόστασης αφορά την απόσταση σημείου με γειτονικά του σημεία [**30**] , [**31**] .

Από την άλλη ο αλγόριθμος k-means -που συχνά συγχέεται με τον k-nearest-neighbor όμως πρόκειται για δυο ξεχωριστούς αλγορίθμους- είναι μια μέθοδος κατηγοριοποίησης n αντικειμένων σε k συστάδες όπου κάθε σημείο/παρατήρηση ανήκει στην συστάδα (που έχει οριστεί) στην οποία βρίσκεται πιο κοντά στο κέντρο της από κάποιο άλλο κέντρο. Η απόσταση του σημείου από το κέντρο της συστάδας δεν αναπαριστά την Ευκλείδια απόσταση αυτή δηλαδή που σχετίζεται με το

Πυθαγόρειο θεώρημα. Ο αλγόριθμος ανήκει στην κατηγορία αλγορίθμων μηχανικής μάθησης αυτών χωρίς επίβλεψη [**30**] , [**31**] , [**32**] .

3.3.2 Trajectory analysis

Η καταγραφή της πορείας ζωής των κυττάρων (trajectory analysis), δίνει πιθανές αλλαγές στην πορεία αυτών δηλαδή ένα κύτταρο ξεφεύγει από τους φυσιολογικούς ρυθμούς ανάπτυξής του (λ.χ. καρκίνος, διαφοροποίηση) και αυτό σηματοδοτεί κάποια παθογένεια, νόσο κλπ. Μάλιστα αναφέρεται πως η κυτταρική διαφοροποίηση κάποιες φορές δεν είναι δυνατό να ανακαλυφθεί μόνο με την συσταδοποίηση καθώς η συσταδοποίηση αποτελεί χρονικό στιγμιότυπο ταξινόμησης των κυττάρων και άρα εκεί κρίνεται αναγκαία η εκτέλεση μιας περαιτέρω ανάλυσης που περιλαμβάνει λεπτομέρειες που σχετίζονται με την πορεία ανάπτυξης των οργάνων, ή μεταξύ των σταδίων κάποιας ασθένειας ή ακόμη και τοπολογικές λεπτομέρειες τοπικά στην περιοχή μελέτης. Η ετερογένεια μπορεί να εμφανιστεί σε οποιοδήποτε στάδιο ζωής. Τα δεδομένα αυτά καταγράφονται με βάση μια χρονολογική κλίμακα που χαρακτηρίζεται ως ψευδοχρόνος (pseudotime). Μια νεότερη μέθοδος, γνωστή ως RNA velocity αφορά την ταχύτητα των μορίων RNA και ανιχνεύει με βάση την ταχύτητα αν το mRNA μόριο έχει διαχωριστεί από τα εσώνια ή όχι. Μάλιστα ο δείκτης αυτός μπορεί να προβλέψει μελλοντικές καταστάσεις του μορίου [**15**] , [**23**] .

3.3.3 Ταυτοποίηση κυτταρικών ομάδων

Κατά την συσταδοποίηση, για κάθε τύπου κυττάρου του δείγματος υπάρχουν κάποια συγκεκριμένα γονίδια που χαρακτηρίζουν την κάθε κυτταρική ομάδα, τα οποία συχνά αναφέρονται ως marker genes. Σε αυτό το στάδιο είναι απαραίτητη η συμβολή βάσεων δεδομένων όπως το Human Cell Atlas [**15**] . “Το Human Cell Atlas είναι μια διεθνής ομάδα ερευνητών που με την εκμετάλλευση της προόδου των τεχνολογικών εξελίξεων είναι σε θέση πλέον να μελετούν και να χαρτογραφούν όλα τα κύτταρα ένα προς ένα. Καταγράφουν όλα τα χαρακτηριστικά τους, δηλαδή το σχήμα τους, την θέση τους, την αλληλεπίδραση με άλλων τύπων κυττάρων ή και με το περιβάλλον γύρω τους, τους μηχανισμούς τους και πάρα πολλά ακόμη. Η μελέτη αυτή έχει σοβαρό αντίκτυπο σε οτιδήποτε συμβαίνει γύρω από την κατανόηση των ασθενειών, την αποτελεσματική θεραπεία και την συνεισφορά στην φαρμακολογία κ.α. Συνεπώς η δημιουργία του Human Cell Atlas συγκεντρώνει όσα αναφέρθηκαν δημιουργώντας ένα

από τα πολλά σημεία εκκίνησης για τους ερευνητές. Χάρη σε αυτή και άλλες πλατφόρμες η σύγκριση των γονιδίων μεταξύ 2 συστάδων μπορεί να προσδιορίσει τον κυτταρικό τύπο" [**33**] .

3.3.4 Differential expression analysis

Η διαφοροποιημένη έκφραση των γονιδίων, ένα γνώρισμα που πολλά υπολογιστικά εργαλεία υπολογίζουν και μεμονωμένα και συνεισφέρει στην ταυτοποίηση τον κυτταρικών τύπων. Τα γονίδια, ιδιαίτερα όταν πρόκειται για περιπτώσεις που η διαφοροποίηση τους σηματοδοτεί την εκδήλωση μιας παθογένειας. Η differential analysis στοχεύει στον εντοπισμό και προσδιορισμό των γονιδίων με τις μεγαλύτερες διαφορές που μπορεί να εμφανίσουν κατά την γονιδιακή έκφραση. Συγκεκριμένα γονίδια συνήθως 2 συστάδων/κυτταρικών ομάδων ανιχνεύονται και ελέγχονται με εφαρμογή στατιστικών μεθόδων για το κατά πόσο διαφέρουν κατά την έκφραση τους. Οι υποθέσεις που τίθενται για να πραγματοποιηθεί ο στατιστικός έλεγχος ελέγχονται στο κατά πόσο το γράφημα της κατανομής τους είναι όμοιο. Ο έλεγχος της διαφοροποιημένης έκφρασης αποτελεί μείζων ζήτημα και για αυτό υπάρχουν υπολογιστικά εργαλεία που ασχολούνται μόνο με αυτό. Μερικά από αυτά είναι: edgeR, DESeq, DEXSeq, SAMseq, NOIseq και πολλά ακόμη [**15**] , [**34**] .

4 Προκλήσεις κατά την μελέτη δεδομένων single-cell-RNA-sequencing

Η τεχνολογική ανάπτυξη και πρόοδος σε συνδυασμό με τις προοπτικές που υπόσχεται η μέθοδος single-cell RNA-sequencing, έδωσε ώθηση στην πραγματοποίηση μελετών. Εκτός από όλα τα πλεονεκτήματα των μελετών, έχουν προκύψει και συνεχίζουν να προκύπτουν προκλήσεις που επηρεάζουν την ποιότητα των δεδομένων που εξάγονται. Αυτές όχι μόνο δεν μπορούν να παραβλεφθούν, αλλά έχουν δημιουργηθεί ξεχωριστά υπολογιστικά εργαλεία για την αναζήτηση και την αντιμετώπιση τους. Κάποιες από τις προκλήσεις αφορούν την ανίχνευση των επιπέδων των dropouts, την αφαίρεση των batch effects παραγόντων και πολλά ακόμη.

4.1 Dropouts

Οι περισσότερες δημοσιεύσεις που ασχολούνται με τα προβλήματα που προκύπτουν κατά την ανάλυση δεδομένων τύπου single-cell-RNA παρουσιάζουν τα dropouts ως ένα από αυτά. Με τον όρο dropouts γενικότερα οι επιστήμονες αναφέρονται στην ελάχιστη απεικόνιση της έκφρασης μορίων του RNA (γνωστό ως transcriptome) σε σχέση με την ποσότητα αυτού που υπάρχει στην πραγματικότητα σε κάθε κύτταρο που μελετάται ξεχωριστά ή δεν ανιχνεύονται ενώ αυτό είναι αναμενόμενο. Με άλλα λόγια, τα dropouts αφορούν τα γονίδια που εκφράζονται από κύτταρα ενώ ταυτόχρονα δεν εκφράζονται σε άλλα κύτταρα του ίδιου τύπου (συνήθως λέγονται true zeros). Ακόμη, παρατηρείται η κατηγορία των γνωστών ως dropout zeros που αφορούν τα δεδομένα που χάνονται λόγω χαμηλών επιπέδων RNA δεδομένων εισόδου [**35**] , [**36**] .

Τα dropouts μπορεί να συμβούν είτε εξαιτίας βιολογικών είτε εξαιτίας τεχνικών λόγων.

- Ως βιολογικοί λόγοι αναφέρονται: το γεγονός να εκραγεί το γενετικό υλικό κατά την διαδικασία της μεταγραφής του DNA σε RNA (γνωστό ως transcriptional bursting), ή η αποικοδόμηση του RNA.
- Ως τεχνικοί λόγοι αναφέρονται: ο ελάχιστος αρχικός αριθμός δειγμάτων που λήφθηκε μην αφήνοντας περιθώριο να συμβούν πιθανά λάθη, ή η έλλειψη αντιγράφων γονιδίων από το στάδιο της υβριδοποίησης γεγονός που οφείλεται στα γνωστά ως batch effects, δηλαδή λήψη μη ακριβή δεδομένων εξαιτίας των συνθηκών του εργαστηρίου που διεξάγεται η μελέτη, των επιπέδων όζοντος του εργαστηρίου, του χρόνου διεξαγωγής,

των εργαλείων και μηχανημάτων που χρησιμοποιήθηκαν κατά την μελέτη κ.α. [**35**] .

Σύμφωνα με μια επιστημονική δημοσίευση των *Gong, W., Kwak, IY., Pota, P. et al.(2018)*, τα γεγονότα dropouts είναι το αποτέλεσμα του συνδυασμού τόσο των γονιδίων που δεν εκφράζονται όσο και της μικρής (ανεπαρκούς) ποσότητας RNA των κυττάρων που αποτελούσαν τα δεδομένα εισόδου. Στην ίδια επιστημονική δημοσίευση γίνεται λόγος για την δημιουργία του DrImpute που φτιάχτηκε για να μετρά τα ποσοστά των dropouts και να ενισχύσει την λειτουργία των ήδη υφιστάμενων πλατφορμών ανάλυσης δεδομένων ως προς την καλύτερη ταυτοποίηση των κυτταρικών τύπων. Οι *Gong, W., Kwak, IY., Pota, P. et al.(2018)* θέλοντας να καταφέρουν αξιοσημείωτες μετρήσεις χρησιμοποίησαν 9 διαφορετικές πλατφόρμες και υπολογιστικά εργαλεία, αναπαράγοντας την μέθοδο που προτείνουν αρκετές φορές. Η προσέγγιση DrImpute έχει καλύτερη απόδοση κατά την μέτρηση των dropout zeros σε σχέση με την μέτρηση των true zeros [**35**] .

Σε αντίθεση με τα παραπάνω ο *Qiu, P.(2020)* δεν υποστηρίζει πως τα dropouts αποτελούν μόνο αρνητικά στοιχεία που πρέπει να αφαιρεθούν αλλά υποστηρίζει το όφελος που προκύπτει από την ποσοτικοποίηση και μέτρηση των dropouts. Αναφέρεται πως με την εφαρμογή του αλγορίθμου M3Drop που αποτελεί μια μέθοδο μοντελοποίησης των σχέσεων μεταξύ της μέσης τιμής έκφρασης του γονιδίου και του ρυθμού που συμβαίνει το dropout, τα αποτελέσματα έδειξαν ότι κάποια γονίδια είναι πιο ευαίσθητα ώστε να χαθούν και είναι χρήσιμα για χαρτογράφηση μεταξύ των υπόλοιπων πληροφοριών που αναγράφονται μετά την ανάλυση. Τα dropouts είναι ο λόγος που τα δεδομένα είναι διάσπαρτα, όμως αυτό δεν αναιρεί το όφελος από την μελέτη τους. Αρχικά παρουσιάζεται ένας αλγόριθμος συσταδοποίησης που συνυπολογίζει μεταξύ ενός ζεύγους γονιδίων αν αυτά θα ανιχνευτούν μαζί σε άλλα κύτταρα του ίδιου είδους. Τα κύτταρα στα οποία εντοπίζονται τα ίδια γονίδια χωρίζονται σε συστάδες, όπου η κάθε συστάδα συνοδεύεται από κάποιες μεταβλητές (μεταξύ των μεταβλητών είναι η αναλογία πληροφορίας/θορύβου, η μέση τιμή μεταξύ των συστάδων κ.α.) που μετρούν αν τα γονίδια κατά τα διάφορα στάδια έκφρασης τους παρουσιάζουν διαφορετική συμπεριφορά έκφρασης, δηλαδή διαφοροποιούνται. Για να υποστηριχθούν τα παραπάνω πραγματοποιήθηκε ανάλυση single-cell RNA-sequencing από τις πλατφόρμες Seurat και 10X Genomics σε δεδομένα που είναι διαθέσιμα από την 10X Genomics. Η σύγκριση των αποτελεσμάτων από τις δυο διαφορετικές πλατφόρμες έδειξε ότι ο αλγόριθμος M3Drop εντόπισε περισσότερα

γονίδια στην πρώτη πλατφόρμα από ότι στην δεύτερη. Αυτό σημαίνει ότι αρκετά από τα γονίδια για τα οποία γίνεται λόγος χάθηκαν κατά την εκτέλεση του αλγορίθμου συσταδοποίησης [**36**] .

Σε επόμενο στάδιο της δημοσίευσης, ο *Qiu, P.(2020)* διεξάγει έρευνα για να διαπιστωθούν τα dropout κατά τις διάφορες φάσης ζωής των κυττάρων και των γονιδίων. Τα single-cell RNA-sequencing δεδομένα που μελετήθηκαν προερχόταν από εμβρυϊκά κύτταρα ανθρώπου και μελετήθηκαν με την πλατφόρμα SMART-seq2. Η οπτικοποίηση των αποτελεσμάτων έδειξε ότι η πορεία όλων των γονιδίων και των συστάδων εμφάνισαν εντυπωσιακές διαφορές στα μοτίβα dropouts [**36**] .

Συμπερασματικά λοιπόν , ο τρόπος που χάνονται τα γονίδια ή αλλιώς ο τρόπος έκφρασης τους σε μια κυτταρική ομάδα βοηθάει στον καλύτερο διαχωρισμό των συστάδων(ομάδες κυττάρων). Φυσικά, δεν αποτελεί μια μέθοδο ορθού προσδιορισμού των κυτταρικών ομάδων, όμως μπορεί να χρησιμοποιηθεί ως εναλλακτική μέθοδος διασταύρωσης των υποθέσεων ή σύγκρισης με άλλα αποτελέσματα για τον χαρακτηρισμό των κυτταρικών τύπων [**36**] .

4.2 Αφαίρεση batch effect παραγόντων

Με τον όρο batch-effect, χαρακτηρίζονται όλοι εκείνοι οι παράγοντες που επηρεάζουν την διεξαγωγή των πειραμάτων και τα αποτελέσματα. Αναλυτικά αναφέρονται ως παράγοντες η θερμοκρασία, τα επίπεδα του όζοντος στο εργαστήριο, οι συνθήκες υγιεινής του εργαστηρίου, ο χρόνος διεξαγωγής του πειράματος, τα εργαλεία και τα μηχανήματα στα οποία συγκεντρώθηκαν και έγινε η ανάλυση του δείγματος καθώς και οι υπολογιστικές πλατφόρμες. Από την άλλη πειράματα σε single-cell δεδομένα υποφέρουν ταυτόχρονα από τα dropouts, αποτυχία καταγραφής του RNA ή της υβριδοποίησης για αυτό γίνονται προσπάθειες για δημιουργία πλατφορμών που θα αναλύουν δεδομένα που συνδυάζουν όλα τα παραπάνω χαρακτηριστικά. Αρκετά πακέτα δεδομένων, τα οποία έχουν συλλεχθεί έπειτα από διαφορετικές αναλύσεις του ίδιου ιστού, εμφανίζουν ποικιλία στις τιμές των παραμέτρων. Αυτό είναι αναπόφευκτο όταν πρόκειται για δεδομένα που λαμβάνονται από ανθρώπινους κυτταρικούς ιστούς καθώς μελετώνται σε διαφορετικές χρονικές στιγμές. Βέβαια, σε κάποια κυτταρική ομάδα μεταξύ των κυττάρων του δείγματος που έχουν αναλυθεί ταυτόχρονα υπάρχουν κάποια κύτταρα που είναι πιο ευάλωτα στα batch-effects [**37**] , [**38**] , [**39**] .

“Η ανάλυση των batch-effects που επηρεάζουν την κάθε κυτταρική ομάδα και όχι η γενικευμένη μορφή των batch-effects μπορεί να βελτιώσει την αφαίρεση αυτών κατά την ανάλυση των δειγμάτων” *Haghverdi, L., Lun, A. T., Morgan, M. D., & Marioni, J. C. (2018)* . Μέθοδοι όπως το Seurat μελετούν και συγκρίνουν ανά ζευγάρι τα δείγματα ενός κυτταρικού ιστού που έχουν ληφθεί διαφορετικές χρονικές στιγμές. Η μέθοδος DESC είναι μια μέθοδος μη επιβλεπόμενης μάθησης η οποία επιτυγχάνει σε επαναληπτική μορφή την σταδιακή αφαίρεση των batch-effects σε βαθμό ώστε να μην αλλοιωθούν όλες οι βιολογικές διαφορές που εντοπίζονται μεταξύ των κυττάρων παρά μόνο οι τεχνικές που προαναφέρθηκαν. Η σύγκριση του DESC με άλλα εργαλεία έδειξε την αποτελεσματικότητα του να αφαιρεί ακόμη και τους πιο περίεργους συνδυασμούς των batch-effects. Ανάμεσα στα πειράματα σύγκρισης αναφέρονται ότι η μέθοδος εφαρμόστηκε σε κύτταρα του αμφιβληστροειδούς χιτώνα από 4 μαϊμούδες, σε κύτταρα του παγκρεατικού ιστού όπου τα δεδομένα αναλύθηκαν με 4 διαφορετικά πρωτόκολλα ανάλυσης single-cell RNA-sequencing, σε περίπλοκα δεδομένα ανθρώπινων κυτταρικών ιστών όπου οι τεχνικοί παράγοντες συγχέονται με τους βιολογικούς κ.α. Εκτός αυτών, μελετήθηκε ο χρόνος και η μνήμη του υπολογιστή σε σχέση με τον αριθμό των κυττάρων και την πολυπλοκότητα των batch-effects [**37**] , [**38**] , [**39**] .

4.3 Άλλες προκλήσεις

4.3.1 Χαρτογράφηση

Η χαρτογράφηση των αποτελεσμάτων είναι μια μέθοδος οπτικής απεικόνισης. Πρόκειται για τα τελικά αποτελέσματα της ανάλυσης. Η χαρτογράφηση όμως και κατά την διάρκεια των διάφορων σταδίων της ανάλυσης θα ήταν ιδιαίτερα βοηθητική, εντούτοις αυτό ακόμη είναι δύσκολο να πραγματοποιηθεί ή αν συμβεί τα αποτελέσματα δεν θα είναι απόλυτα σωστά. Η χαρτογράφηση των κυττάρων παρουσιάζει τόσο την δομή και την οργάνωση του ιστού όσο και τα στάδια ανάπτυξης του, μετρά την συνεχόμενη εναλλαγή καταστάσεων του, ξεχωρίζει τους κυτταρικούς πληθυσμούς που ενδεχομένως προκύπτουν κατά τα στάδια ανάπτυξης και μετατρέπονται σε επιβλαβείς, επιτρέπει την επιλογή της χρονικής στιγμής μελέτης των κυττάρων, και περιλαμβάνει σχόλια που αφορούν την βιολογία, την λειτουργικότητα και λεπτομέρειες τεχνικής φύσεως. Κάτι τέτοιο θα πρόσφερε άλλου επιπέδου γνώση

και προσέγγιση της μελέτης. Ως τώρα έχει σχεδιαστεί ένα ιδανικό μοντέλο (δεν έχει υλοποιηθεί) για single-cell RNA δεδομένα για το πως η χαρτογράφηση διάφορων σταδίων θα μπορούσε να ενισχύσει τα δεδομένα σύμφωνα με όσα αναγράφονται και παραπάνω. Ιδιαίτερα αισιόδοξο είναι το αποτέλεσμα της μελέτης των *Zheng, K., Lin, L., Jiang, W., Chen, L., Zhang, X., Zhang, Q., ... & Hao, J. (2022)* όπου η ανάλυση μιας φλεγμονής του κεντρικού νευρικού συστήματος (εγκέφαλος και νωτιαίος μυελός) θα αποκαλύψει γονίδια για τα οποία δεν έχει γίνει κάποια αναφορά στο παρελθόν. Η μεγάλη σε βαθμό ετερογένεια μεταξύ των νευρικών κυττάρων στο εύρος του κεντρικού νευρικού συστήματος κρύβει πολλές προκλήσεις και έτσι δυσκολεύει τον προσδιορισμό του ρόλου των εγκεφαλικών κυττάρων. Όμως η single-cell RNA-sequencing είναι παρούσα για να ενισχύσει την προσπάθεια δημιουργίας ενός χάρτη κυτταρικών πληθυσμών από δείγματα ποντικιών. Τα microglia είναι ένας τύπος νευρογλοιακών κυττάρων τα οποία είναι διαφοροποιημένα μετά το ισχαιμικό εγκεφαλικό επεισόδιο. Επιπλέον ταυτοποιήθηκαν υποσυστάδες μεταξύ των εγκεφαλικών αγγειακών κυττάρων και άλλων ειδών νευρογλοιακών κυττάρων. Η single-cell RNA-sequencing κατάφερε να παρουσιάσει στο μικροσκόπιο την εξέλιξη της φλεγμονής του κεντρικού νευρικού συστήματος και να αποκαλύψει πολύτιμες πληροφορίες εξετάζοντας νέους τρόπους και μηχανισμούς φαρμακευτικής αγωγής [**40**] , [**41**] .

4.3.2 Βελτίωση της ανάλυσης (trajectory)

Μοντέλα μελέτης των διαφοποιήσεων που παρατηρούνται κατά την ανάλυση (trajectory analysis) σε επίπεδο πρωτεϊνικό, μεταβολικό και επιγενετικό βρίσκονται ακόμη σε πρόωρο στάδιο. Από την άλλη μοντέλα μελέτης του συνόλου του RNA σημειώνουν αρκετά καλύτερα αποτελέσματα. Ερωτήσεις όπως το πως μπορεί να καταγραφεί αυτή η νέα πορεία και ποιες είναι οι κατάλληλες παράμετροι να οριστούν για αυτή τη διαδικασία, ή πως είναι δυνατόν να συγκριθούν ξεχωριστά μονοπάτια από το ίδιο είδος δεδομένων ώστε να καταγραφούν ενδεχομένως κι άλλες πτυχές, παραμένουν αναπάντητες [**40**] .

4.3.3 Αλληλεπίδραση

Στην συγκεκριμένη πρόκληση αναφέρεται το παράδειγμα του καρκίνου και το γεγονός ότι μπορεί να συμβούν πολλά χωροταξικά μοτίβα και σχέσεις αλληλεπίδρασης που

αφορούν από ομάδες κυττάρων του ανοσοποιητικού συστήματος μέχρι ομάδες κυττάρων στο μικροπεριβάλλον του όγκου. Ως τώρα δεν έχουν δημιουργηθεί μέθοδοι που μπορούν να δώσουν πλήρη αναφορά για αυτό δηλαδή χωρικές συντεταγμένες και τις εκφράσεις του RNA ή γονιδίων από τις οποίες προέρχονται. Τα μοντέλα που υπολογίζουν την ικανότητα των καρκινικών κυττάρων να επηρεάσουν άλλα υγιή κύτταρα στο μικροπεριβάλλον του όγκου αντιμετωπίζουν προβλήματα που προέρχονται από την ολοένα αυξανόμενη ποσότητα κυττάρων που αναλύεται η αλληλουχία τους και τον αυξανόμενο ρυθμό εμφάνισης υπολογιστικών εργαλείων που ταυτοποιούν τις αλληλουχίες ανά γονιδίωμα. Αυτό οδηγεί σε μη ορθά συμπεράσματα καθώς η πληθώρα των παραπάνω προκαλεί αντιφάσεις στο τι είναι αποδεκτό και τι όχι. Ένα άλλο πρόβλημα σχετίζεται με γενετικές παραμέτρους της ετερογένειας του όγκου. Μοντέλα μηχανικής μάθησης έχουν δημιουργηθεί για το περιβάλλον του όγκου που αναγνωρίζουν δομές, μοτίβα και άλλους βιοδείκτες του όγκου παρόλα αυτά παραμένουν ασαφής και η εξέλιξη τους είναι ένα πιθανό σενάριο. Μερικές από αυτές τις αλληλεπιδράσεις είναι εξαιρετικά δύσκολο να ανιχνευθούν μέσα σε δεδομένα που εντοπίζονται με πολύ χαμηλή συχνότητα [**40**] .

Μια ακόμη πρόκληση της συγκεκριμένης κατηγορίας αφορά την ενοποίηση δεδομένων τύπου single-cell μεταξύ δειγμάτων, πειραμάτων και τύπων μετρήσεων. Οι βιολογικές διαδικασίες είναι περίπλοκες και δυναμικές διαδικασίες που διαφέρουν μεταξύ κυτταρικών ομάδων, ιστών, οργάνων, οργανισμών. Όταν χρειάζεται να ταυτοποιηθούν μοτίβα μεταξύ οργανισμών ή κυττάρων τα δείγματα που συγκρίνονται αφαιρούνται είτε σε διαφορετικό χρόνο είτε από διαφορετικό τοπολογικό σημείο του όγκου. Η σύγκρισή τους απαιτεί επίγνωση παραγόντων του τύπου batch effects [**40**] .

4.3.4 **Συνεχείς αξιολογήσεις**

Τέλος, προκλήσεις εντοπίζονται και σε εργαλεία συγκριτικών αξιολογήσεων για μετρήσεις single-cell δεδομένων. Συνεχώς προκύπτουν νέα μοντέλα μελέτης. Για αυτό και οι παράλληλες συγκριτικές αξιολογήσεις κρίνονται απαραίτητες σε τακτικά χρονικά διαστήματα σε ήδη γνωστά δεδομένα. Για να θεωρηθεί μια μέθοδος αξιόπιστη και χρήσιμη στα χέρια των ερευνητών πρέπει τουλάχιστον να εκτελούνται με μεγάλη ακρίβεια η συσταδοποίηση, η μελέτη διαφοροποιημένων κυττάρων, η διαχείρηση υψηλών επιπέδων θορύβου, η ανίχνευση αλληλόμορφων που καταστράφηκαν κ.α. Δυστυχώς παρατηρείται έλλειψη των συστημάτων προσομοίωσης των παραπάνω.

Υπάρχουν ελάχιστα ενσωματωμένα σε άλλα προγράμματα ως υποπρογράμματα και έτσι η αξιολόγηση δεν εγγυάται σωστά βιολογικά και τεχνικά αποτελέσματα [**40**] .

5 Σύγκριση πρωτοκόλλων 10X Genomics και Smart-Seq2

Υπάρχει μια μεγάλη ποικιλία υπολογιστικών εργαλείων που ασχολούνται με το πρώτο και δεύτερο μέρος της ανάλυσης, δηλαδή κυρίως την προεξεργασία των δεδομένων, τον ποιοτικό έλεγχο, την κανονικοποίηση κ.α. Μερικές από τις πιο δημοφιλής πλατφόρμες, των οποίων τα δεδομένα χρησιμοποιούνται για περαιτέρω διερεύνηση δεδομένων τύπου single-cell RNA sequencing είναι οι: Smart-Seq2, 10X Genomics Chromium, inDrop, CEL-seq κ.α.

Σύμφωνα με την αντίστοιχη δημοσίευση των *Wang, X., He, Y., Zhang, Q., Ren, X., & Zhang, Z. (2021)*, γίνεται σύγκριση των δυνατοτήτων δυο πλατφορμών των Smart-Seq2 και 10X Genomics Chromium αξιοποιώντας τα αντιγόνα CD45 που λήφθηκαν από ασθενείς με καρκίνο. Ο πρώτος ασθενής είχε διαγνωστεί με ηπατοκυτταρικό καρκίνωμα από τον οποίο αφαιρέθηκε όγκος από το ήπαρ και ιστός από υγιή γειτονική περιοχή του ήπατος και ο δεύτερος ασθενής είχε διαγνωστεί με καρκίνο στο ορθό (περιοχή ακριβώς πριν από τον πρωκτό) και έχοντας κάνει μετάσταση στο συκώτι από τον οποίο συλλέχθηκε και ο πρωταρχικός και ο μεταστατικός όγκος [**42**] .

Στο παρακάτω πίνακα αποτυπώνονται συνοπτικά ορισμένες διαφορές που παρατηρήθηκαν κατά την αξιολόγηση τους και παράθεση τους στην δημοσίευση των *Wang, X., He, Y., Zhang, Q., Ren, X., & Zhang, Z. (2021)*:

Smart-Seq2	10X Genomics
Πολύ υψηλό κόστος	Χαμηλό κόστος
Το mRNA χωρίζεται σε πολλούς δοκιμαστικούς σωλήνες	Το δείγμα μπορεί να αναλυθεί όλο μαζί σε ένα δοκιμαστικό σωλήνα
Ανιχνεύει περισσότερα γονίδια με μικρή αλληλουχία, ή γονίδια που έχουν εμφυτευτεί ή μιτοχονδριακά γονίδια	Ανιχνεύει καλύτερα τα σπάνια κύτταρα
Πιο επιτυχημένη πλατφόρμα	Πιο γνωστή πλατφόρμα
Διαβάζει περισσότερη ποσότητα μιτοχονδριακού γενετικού υλικού	Διαβάζει μικρότερη ποσότητα μιτοχονδριακού γενετικού υλικού
Ανιχνεύει περίπου 80,5%-92,6% lncRNAs	Ανιχνεύει περίπου 77,4%-99,2% lncRNAs

Ανιχνεύει περισσότερα γονίδια	Ανιχνεύει περισσότερες κυτταρικές ομάδες
5 συστάδες	11 συστάδες
Μικρότερη πιθανότητα καταστροφής κυττάρων	Μεγαλύτερη πιθανότητα καταστροφής κυττάρων
Καλύτερη απόδοση	Χαμηλή ικανότητα και χαμηλή ποιότητα απεικόνισης του mRNA εξαιτίας θορύβου

"©[(Data via Wang, X., He, Y., Zhang, Q., Ren, X., & Zhang, Z. (2021). Direct comparative analyses of 10X genomics chromium and smart-seq2. *Genomics, proteomics & bioinformatics*, *19*(2), 253-266.]"

Πίνακας 1: Σύγκριση πρωτοκόλλων και εύρεση διαφορών για τα πρωτόκολλα Smart-Seq2 και 10X Genomics Chromium

Μεταξύ των παραπάνω διαφορών των δυο πλατφορμών αναφέρονται και άλλα γνωρίσματα. *"*Και οι δυο πλατφόρμες ανιχνεύουν περίπου 10-30% των αλληλουχιών των μη κωδικών περιοχών" *Wang, X., He, Y., Zhang, Q., Ren, X., & Zhang, Z. (2021)* . Η Smart-Seq2 μέθοδος θεωρείται πιο αξιόπιστη καθώς λιγότερα κύτταρα χάνονται, και άρα στο ρυθμιστικό διάλυμα επιβιώνουν αρκετά ώστε να εξαχθούν ασφαλή συμπεράσματα. Εκτός αυτών η Smart-Seq2 δεν χρειάζεται ιδιαίτερο εξοπλισμό και μπορεί να αξιοποιηθεί από τους περισσότερους. Διαφορές παρατηρούνται και στις δυο πλατφόρμες σχετικά με την ανίχνευση υποομάδων του RNA. "Από όλα τα γονίδια που ανιχνεύτηκαν το 5,6% με την πλατφόρμα 10X Genomics έχουν σχόλιο που η πλατφόρμα τα χαρακτηρίζει, ενώ μόλις το 2,7% με την Smart-Seq2 έχει σχόλιο" *Wang, X., He, Y., Zhang, Q., Ren, X., & Zhang, Z. (2021)* . Παρόλα αυτά όμως η Smart-Seq2 μπορεί να ανιχνεύσει γονίδια με άγνωστες ως τώρα λειτουργίες. Επίσης ένα άλλο χαρακτηριστικό που ελέγχεται από τις πλατφόρμες είναι η επικοινωνία των κυτταρικών πληθυσμών στο μικροπεριβάλλον του όγκου ή προσπαθεί να προβλεφθεί ο τρόπος επικοινωνίας όπου δεν είναι γνωστός. Και σε αυτή την περίπτωση η τεχνική Smart-Seq2 φαίνεται να είναι προτιμότερη της 10X Genomics. Επιπρόσθετα στην 10X Genomics υπάρχει μεγαλύτερη πιθανότητα να καταστραφούν περισσότερα κύτταρα από ότι στη Smart-Seq2. Και στις δυο πλατφόρμες είναι προτιμότερο να χρησιμοποιούνται μερικές χιλιάδες κύτταρα ώστε στο τέλος να περισσέψει επαρκή ποσότητα αυτών για να μελετηθούν όσα περισσότερα είναι δυνατό. Η ποικιλομορφία

ενός ιστού μπορεί να προέρχεται από διαφορετικούς κυτταρικούς τύπους, για αυτό το λόγο μια λάθος καταμέτρηση ή η καταστροφή αρκετών σε ένα μικρό δείγμα θα οδηγήσει σε λανθασμένα αποτελέσματα [**42**] .

6 Υπολογιστικά εργαλεία και πλατφόρμες

Όπως έγινε λόγος και παραπάνω, η μελέτη των single-cell RNA δεδομένων έφερε νέες προοπτικές και για αυτό συνεχώς αναπτύσσονται πλατφόρμες πραγματοποιώντας ανάλυση ή επιλύοντας επιμέρους ζητήματα κατά την ανάλυση. Οι ανάγκες για ανάλυση δεδομένων μεγάλου όγκου, η ταχύτητα ανάλυσης, η αξιοπιστία των δεδομένων, η ανάλυση σε μια πλατφόρμα από την αρχή ως το τέλος για ένα συγκεκριμένο σύνολο δεδομένων, οι απαιτήσεις υπολογιστικών πόρων και η μετα-ανάλυση είναι κάποια από τα κριτήρια που λαμβάνονται υπόψη κατά την δημιουργία νέων υπολογιστικών εφαρμογών ή κατά την αξιολόγηση και βελτίωση των ήδη υφιστάμενων. Κατά την συγγραφή αυτής της εργασίας έγινε αναζήτηση για web εργαλεία που υποστηρίζουν online ανάλυση ή εφαρμογές που εγκαθίστανται τοπικά στον υπολογιστή που εμφανίζουν τα εξής στοιχεία:

- Να συμπεριλαμβάνονται όλα τα στάδια ανάλυσης όπως (quality control, normalization, clustering, visualization, dimensionality reduction, gene filtering, marker genes, highly variable genes). Ο βασικός λόγος αναζήτησης ενός εργαλείο που θα πραγματοποιεί μια ανάλυση από την αρχή μέχρι το τέλος είναι το γεγονός ότι κάθε παράμετρος/ κατώφλι ή οποιαδήποτε μορφή φιλτραρίσματος εφαρμόζεται παραμένει η ίδια καθ'όλη την διάρκεια της ανάλυσης. Η πραγματοποίηση όλων των παραπάνω διεργασιών ανάλυσης από διαφορετικά πακέτα ανάλυσης ενδεχομένως να προκαλέσει κενά, μη ποιοτικά αποτελέσματα καθώς κάθε πακέτο ανάλυσης έχει κάποιες σταθερές που διαφοροποιούνται μεταξύ των εργαλείων. Να σημειωθεί πως το στάδιο της προεπεξεργασίας (απομόνωση κυττάρων ενδιαφέροντος, φυγοκέντριση, καθαρισμός του δείγματος κ.α.) πραγματοποιούνται ξεχωριστά με την εφαρμογή πρωτοκόλλων όπως το 10X Genomics και τα δεδομένα που εξάγονται αναλύονται από τις εν λόγο υπολογιστικές πλατφόρμες.
- Να μην απαιτούν από το χρήστη να κατέχει προγραμματιστικές γνώσεις ή να χρειάζεται απλά να εκτελέσει κάποιο μέρος κώδικα με σύντομο και κατανοητό τρόπο.
- Να μπορεί να εγκαταστήσει γρήγορα και εύκολα το πακέτο ανάλυσης.

- Να μπορεί η ανάλυση να πραγματοποιείται στον ίδιο εικονικό χώρο χωρίς την χρήση επιπλέον πακέτων, γεγονός που δυσκολεύει και μπερδεύει τον χρήστη.
- Να μπορεί ο χρήστης να μοιραστεί, να ρωτήσει και να συζητήσει οποιαδήποτε απορία σχετικά με την ανάλυση του ή αναλύσεις άλλων ερευνητών σε ειδικά διαμορφωμένο περιβάλλον της πλατφόρμας.
- Να μπορεί η ανάλυση των δεδομένων να πραγματοποιείται σε εξωτερικούς server και όχι τοπικά στον υπολογιστή καθώς οι υπολογιστικές απαιτήσεις είναι τεράστιες.

Αναλυτικότερα λεπτομέρειες για τις πλατφόρμες θα καταγραφούν παρακάτω.

Με βάση τις παραπάνω απαιτήσεις βρέθηκαν τα παρακάτω 7 εργαλεία που εκτελούν τις εξής λειτουργίες και παρουσιάζονται συνοπτικά στον παρακάτω πίνακα:

	ScGEATool	Gene Pattern Notebook	BBrowser	Seurat	Scedar	SCANPY	CALISTA
Clustering	✓	✓	✓	✓	✓	✓	✓
Quality control	✓	✓	✓	✓	✗	✓	✓
Dimensionality reduction	✓	✓	✓	✓	✓	✓	✗
Differrential expression	✓	✗	✓	✓	✗	✓	✗
Marker genes	✓	✓	✓	✓	✓	✗	✗
Rare cells	✗	✗	✗	✗	✓	✗	✗
Normalization	✓	✓	✓	✓	✗	✓	✓
Visualization	✓	✓	✓	✓	✓	✓	✓
Gene filtering	✓	✗	✓	✗	✗	✓	✗
Highly Variable genes	✓	✗	✓	✓	✗	✓	✗
Gene Networks	✓	✗	✗	✗	✗	✗	✗
Gene Sets	✓	✗	✗	✗	✗	✗	✗
Integration	✓	✗	✗	✓	✗	✗	✗
Ordering	✓	✗	✗	✗	✗	✗	✓
Interactive	✗	✓	✗	✗	✗	✗	✗
Imputation	✗	✗	✗	✓	✓	✗	✗
Simulation	✗	✗	✗	✗	✗	✓	✗

“©[(Data via Zappia L, Phipson B, Oshlack A. "Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database", PLOS Computational Biology (2018), DOI: 10.1371/journal.pcbi.1006245)]”

Πίνακας 2: Στάδια ανάλυσης υπολογιστικών πλατφορμών

6.1 **Scedar**

Γλώσσα προγραμματισμού: Python

Λειτουργικό σύστημα: MacOS, Linux

Η σχετική δημοσίευση των *Zhang, Y., Kim, M. S., Reichenberger, E. R., Stear, B., & Taylor, D. M. (2020)* μας ενημερώνει για τα χαρακτηριστικά της πλατφόρμας Scedar. Η πλατφόρμα Scedar είναι ένα λογισμικό για την εκτέλεση της διερευνητικής ανάλυσης δεδομένων. Έχει σχεδιαστεί αντικειμενοστρεφώς σε γλώσσα Python η οποία χρησιμοποιείται και από τον χρήστη για να πραγματοποιήσει ανάλυση σε server παρέχοντας μια ισχυρή πλατφόρμα με όλες τις εργασίες που προσφέρονται που αναφέρονται στον παραπάνω πίνακα. Είναι ευέλικτο στη χρήση για ανάπτυξη μελέτης ανεξαρτήτως του όγκου δεδομένων καθώς προσαρμόζεται στις εκάστοτε απαιτήσεις. Γίνεται βελτιστοποίηση του χρόνου με τη μνήμη καθώς τρέχουν σε παράλληλες διεργασίες τα αποτελέσματα των οποίων αποθηκεύονται σε μια δομή δεδομένων για περαιτέρω ανάλυση. Αφορμή της δημιουργίας του αποτελεί το γεγονός ότι ενώ ο αριθμός των κυττάρων που εξετάζονται αυξάνεται, το βάθος της μελέτης και η ποιότητα ανάλυσης έχει μειωθεί. Βασικά πλεονεκτήματα της πλατφόρμας είναι η αύξηση της αποδοτικότητας και η μείωση της πολυπλοκότητας. Τα δεδομένα οπτικοποιούνται , προσφέρεται μέτρηση της απόδοσης ακόμη και των γονιδίων που καταστράφηκαν (dropouts) ή εμφάνισαν μεγάλες αποκλείσεις. Βοηθά στην ανακάλυψη σπάνιων γονιδίων που εκφράζονται αλλά και σπάνιων κυττάρων χρησιμοποιώντας τον αλγόριθμο k-nearest-neighbor ο οποίος συγκρίνει τα κύτταρα αυτά με αυτά που έχουν τα περισσότερα κοινά γνωρίσματα και έτσι μπορεί να ταυτοποιήσει το είδος των κυττάρων που διαφοροποιήθηκαν και να δώσει λεπτομέρειες. Τέλος καταγράφει λεπτομέρειες για το σύνολο του RNA (mRNAs, snRNAs, lnRNAs κ.α.). Πραγματοποιεί συσταδοποίηση (clustering) για δεδομένα μεγάλης κλίμακας [**6**] , [**43**] .

Επιπλέον σύμφωνα με τους *Zhang, Y., Kim, M. S., Reichenberger, E. R., Stear, B., & Taylor, D. M. (2020)* η ανάλυση δεδομένων απαιτεί τεράστιους προγραμματιστικούς πόρους και μεθόδους, εκμετάλλευση πολλών πυρήνων των ηλεκτρονικών υπολογιστών για την πραγματοποίηση των διεργασιών, αντιμετώπιση των σφαλμάτων, σωστή διαχείρηση των δεδομένων μεγάλης διάστασης και η μετατροπή αυτών σε δεδομένα χαμηλότερης διάστασης για να αναπαρασταθούν τα αποτελέσματα και χαμηλά επίπεδα θορύβου ή να πραγματοποιείται ο βέλτιστος περιορισμός του θορύβου. Αυτές οι

απαιτήσεις μπορούν να σταθούν εμπόδιο κατά τον σχεδιασμό των συστημάτων ανάλυσης. Για αυτό το λόγο, για να μην επιβαρυνθεί το σύστημα αναζητούνται τρόποι όπως η εκμετάλλευση νέων υπολογιστικών αρχιτεκτονικών και υψηλά αποδοτικών μεθόδων. Ένας ακόμη στόχος είναι η δημιουργία διερευνητικών μεθόδων οι οποίες θα μπορούν να εφαρμοστούν σε δεδομένα που έχουν παραχθεί από διαφορετικές τεχνολογίες και πρωτόκολλα όπως από τις πλατφόρμες SMARTer, 10x Genomics, Drop-seq. Μάλιστα μέσω μεθόδων μηχανικής μάθησης έχουν βρεθεί μέθοδοι που αναγνωρίζουν από ποια πλατφόρμα έχουν εξαχθεί τα δεδομένα και τελικά καθοδηγείται στο να επιλέξει αν αυτή ή κάποια άλλη είναι η πιο κατάλληλη για να ερμηνεύσει τα βιολογικά δεδομένα [**6**] , [**43**] .

Πριν την χρήση του Scedar για ανάλυση χρειάζεται να γίνει απαραίτητα μια προεπεξεργασία με την μέθοδο της κανονικοποίησης. Το Scedar δεν ειδικεύεται για να εκτελέσει αυτό το στάδιο της ανάλυσης δεδομένων το οποίο μπορεί να αποδειχθεί ότι είναι ιδιαίτερα σημαντικό για τα δεδομένα που θα παραχθούν. Ένα πρόβλημα που αντιμετωπίζει η πλατφόρμα κατά την διερευνητική ανάλυση δεδομένων είναι το να χαθούν κάποια δεδομένα, φαινόμενο γνωστό και ως dropouts. Όταν αρχικά τίθεται η υπόθεση ή το ερώτημα ώστε να διενεργηθεί διερευνητική ανάλυση δεδομένων ένα γονίδιο που ψάχνουν να εντοπίσουν εκτιμάται πως θα βρεθεί στα αποτελέσματα ενώ τελικά κάτι τέτοιο δεν συμβαίνει. Ένα ύψιστης σημασίας ζήτημα είναι η εύρεση σπάνιων κυτταρικών πληθυσμών ή γονιδιακών προφίλ. Ως σπάνια χαρακτηρίζονται εκείνα που διαφέρουν ελάχιστα από τα γειτονικά τους κύτταρα τα οποία ανιχνεύονται με την εκτέλεση του αλγορίθμου k-nearest-neighbor. "Ο k-nearest-neighbor χρειάζεται 25 μονάδες πυρήνων CPU για να τρέξει ο αλγόριθμος" *Zhang, Y., Kim, M. S., Reichenberger, E. R., Stear, B., & Taylor, D. M. (2020)*. Η εύρεση των σπάνιων γονιδιακών προφίλ διευκολύνει την καλύτερη διερεύνηση των δεδομένων χωρίς αυτά να αφαιρούνται [**6**] , [**43**] .

Συγκριτικές αξιολογήσεις πραγματοποιούνται κατά διαστήματα για να ελέγχεται κατά πόσο λειτουργούν οι πλατφόρμες ανάλυσης. Οι παράμετροι που τίθενται ορίζονται με τον ίδιο τρόπο και για κάθε στάδιο και εφαρμόζονται σε ήδη γνωστά δεδομένα. Η ομοιότητα των αποτελεσμάτων φανερώνει την σταθερότητα των μεθόδων μελέτης. Δυο θεωρούνται οι βασικότερες παράμετροι που εξετάζουν την ακρίβεια και την σταθερότητα ο Cluster Consistent Ratio (CCR) και ο Adjusted Rand Index (ARI). Ο CCR δείχνει την συχνότητα σημείων που βρίσκονται μεταξύ δυο συστάδων, δηλαδή ο αλγόριθμος δεν μπορεί να αποφασίσει σε ποια συστάδα ανήκει. Ο ARI αφορά την

αναμενόμενη ομοιότητα μεταξύ των συγκρίσεων όλων των συστάδων, οι οποίες έχουν προκύψει με οποιαδήποτε από τα μοντέλα συσταδοποίησης. Ο CCR είναι προτιμότερος του ARI κατά την χρήση του Scedar καθώς η τελευταία παράμετρος απαγορεύει την διαίρεση μιας μεγάλης συστάδας σε μικρότερες γεγονός που αποτρέπει την ανακάλυψη σπανιότερων κυτταρικών πληθυσμών ή γονιδίων [**6**] , [**43**] .

6.2 CALISTA

Γλώσσα προγραμματισμού: MATLAB / R

Λειτουργικό σύστημα: MacOS, Linux, Microsoft Windows

Το εργαλείο CALISTA προσφέρει την δυνατότητα ανάλυσης είτε με την χρήση του περιβάλλοντος του MATLAB είτε με την χρήση της R. Οι λειτουργίες που εκτελεί αναγράφονται στον παραπάνω πίνακα. Συγκριτικά με τα άλλα εργαλεία εκτελεί τις λιγότερες εργασίες ανάλυσης οι οποίες όμως είναι ιδιαίτερα απαραίτητες. Τα δεδομένα εισόδου μπορεί να είναι προ-επεξεργασμένα δεδομένα από τα πρωτόκολλα Smart-seq2 και scDrop-seq. Τα 4 βασικά στάδια ανάλυσης περιλαμβάνουν την συσταδοποίηση, λεπτομέρειες σχετικά με τα διαφοροποιημένα κύτταρα, ταυτοποίηση κυτταρικών ομάδων και διενέργεια trajectory ανάλυσης παρουσιάζοντας τα στάδια ζωής και διαφοποιήσεων. Το σύνολο αυτών των εργασιών θεωρείται αξιόπιστο καθώς όλα τα βήματα ακολουθούν τις ίδιες τιμές των παραμέτρων που τίθενται ή των στοιχείων που φιλτράρονται όπως τα dropouts, ο θόρυβος κ.α. Η διαδικασία συσταδοποίησης έχει υιοθετήσει μια προηγούμενη μέθοδο και το εργαλείο CALISTA ασχολήθηκε περισσότερο με την ανάπτυξη των υπόλοιπων τριών σταδίων. Κατά την εκτέλεση του CALISTA οι διεργασίες πραγματοποιούνται παράλληλα και με καλύτερη ταχύτητα. Λειτουργεί εξίσου αποτελεσματικά και με σύνολα δεδομένων μεγαλύτερου όγκου. Ακόμη ο χρήστης μπορεί να ξαναεπεξεργαστεί τα αποτελέσματα που έχουν εξαχθεί. Μερικές από τις λεπτομέρειες που μπορεί να λάβει ο ερευνητής είναι τα στάδιο ζωής (στάδιο κυτταρικού κύκλου) των κυττάρων, το σύνολο των UMIs που μετρήθηκαν, τα dropouts, γράφημα σχετικά με την εξέλιξη των διαφοροποιημένων εκφράσεων [**6**] , [**44**] .

6.3 Seurat

Γλώσσα προγραμματισμού: R

Λειτουργικό σύστημα: MacOS, Linux, Microsoft Windows

Το Seurat είναι ένα πακέτο πρωτοκόλλων και υπορουτινών (σύνολο προγραμμάτων που έχουν υλοποιηθεί σε R) και ο χρήστης χρησιμοποιεί την γλώσσα R και εκτελεί ανάλυση (πολλών επιπέδων) δεδομένων τύπου single-cell RNA. Για την πραγματοποίηση της ανάλυσης ο χρήστης χρειάζεται να κατεβάσει το πακέτο R, όπως

και το περιβάλλον RStudio για να εργαστεί σε αυτό όπως επιπρόσθετα πρέπει να κατεβάσει το πακέτο Seurat και τα δεδομένα τα οποία χρειάζονται για την ανάλυση. Για την ανάλυση δημιουργείται ένα αντικείμενο, στο οποίο αποθηκεύονται οι ανάλυσεις όλων των σταδίων που πραγματοποιεί ο χρήστης. Ένα σημαντικό πρόβλημα που εντοποίζεται τόσο στο Seurat όσο και στο Scanpy είναι πως εξαιτίας των γλωσσών ανάπτυξής του περιορίζονται οι δυνατότητες που μπορούν να εκτελέσουν για αυτό και υπάρχουν άλλα πακέτα όπως θα αναλυθούν παρακάτω. Υπάρχουν και άλλα προγράμματα όπως το Signac, SeuratData, SeuratWrappers, SeuratDisk, Azimuth κ.α. που επεκτείνουν το πακέτο Seurat που σημαίνει ότι κληρονομούν όλα τα προγράμματα και τις λειτουργίες/ εργασίες που προσφέρει. Συγκεκριμένα, το Signac χρειάζεται εγκατάσταση στο περιβάλλον της R και προσφέρει την εκτέλεση του ποιοτικού ελέγχου, της μείωσης των διαστάσεων δεδομένων, της ταυτοποίησης των δεδομένων, της συσταδοποίησης, της απεικόνισης κ.α. Το SeuratDisk χρειάζεται εγκατάσταση και αυτό ασχολείται με την ανάλυση και την αποθήκευση των δεδομένων σε μορφή .h5, .h5ad, .rds αρχείου. Το Azimuth είναι μια online εφαρμογή ή μπορεί και να εγκατασταθεί τοπικά στον υπολογιστή και ασχολείται με την κανονικοποίηση , την οπτική απεικόνιση, το data correction το οποίο αφορά τον έλεγχο δεδομένων που θεωρούνται ύποπτα ή υπεύθυνα να προκαλέσουν αλλοίωση της ποιότητας των αποτελεσμάτων, την ταυτοποίηση των κυτταρικών τύπων με την εκτέλεση της εύρεσης διαφοροποιημένων εκφράσεων μέσω του εντοπισμού των βιοδεικτών και των marker genes, την εύρεση των HVGs κ.α. Η εφαρμογή Azimuth προσφέρει 8 επιλογές μεταξύ των οποίων υπάρχει η επιλογή διερεύνησης κυττάρων ανθρώπου ή ποντικιού. Οι 8 επιλογές είναι οι : Human PBMC, Human Motor Cortex, Mouse Motor Cortex, Human Pancreas, Human Fetal Development, Human Kidney, Human Bone Marrow και Human Lung v2. Εάν τα δεδομένα που διαθέτει ο χρήστης ανήκουν σε κάποια κατηγορία μπορεί να ανεβάσει το αντίστοιχο αρχείο και να ξεκινήσει την ανάλυση του. Σε διαφορετική περίπτωση οι προγραμματιστές του Seurat ανάλογα με τις ανάγκες που παρατηρούνται κατά την διεθνή βιβλιογραφία και τα πειράματα που βρίσκονται σε εξέλιξη αναπτύσσουν νέα πακέτα ανάλυσης [**6**] , [**15**] , [**45-47**] .

Σύμφωνα με την βιβλιογραφία το πακέτο ανάλυσης Seurat αποτελεί εργαλείο προ-ανάλυσης των δεδομένων, και τα αποτελέσματα του εισάγονται σε άλλα εργαλεία για περαιτέρω διερευνητική ανάλυση [**6**] , [**15**] , [**45-47**] .

Αν και το Seurat αποτελεί εύκολο στη χρήση εργαλείο, εντούτοις ο χρόνος εγκατάστασης πολλών πακέτων εργασίας καθώς και οι υπολογιστικοί πόροι που

απαιτούνται για να καλύψουν αυτή την πληθώρα εργασιών δυσκολεύουν τον χρήστη [**6**] , [**15**] , [**45-47**] .

6.4 **Scanpy**

Γλώσσα προγραμματισμού: Python

Λειτουργικό σύστημα: MacOS, Linux, Microsoft Windows

Το Scanpy εργαλείο υλοποιείται και χρησιμοποιεί την γλώσσα Python για την πραγματοποίηση όλων των επιπέδων ανάλυσης που αναγράφονται στον παραπάνω πίνακα. Χαρακτηριστικό που διαφοροποιεί το Scanpy είναι η δυνατότητες ανάλυσης, δηλαδή να φέρει εις πέρας αναλύσεις δεδομένων μεγάλου όγκου. Για την εισαγωγή των δεδομένων χρησιμοποιεί το μοντέλο ANNDATA, δηλαδή ένα πίνακα με τα κύτταρα και τα γονίδια καθώς και το πλήθος κάθε γονιδίου που εντοπίζεται σε κάθε κύτταρο. Όπως και το CALISTA εκτελεί παράλληλη ανάλυση εξοικονομώντας χρόνο. Περιλαμβάνει το Jypyter Notebook το οποίο είναι ένα φιλικό προς τον χρήστη πακέτο που επιτρέπει την βηματική ανάλυση. Γραμμές κώδικα σε γλώσσα Python αντιστοιχούν στα βήματα της ανάλυσης τα οποία ο χρήστης εκτελεί λαμβάνοντας ως αποτελέσματα τα διάφορα γραφήματα. Τα αποτελέσματα της ανάλυσης συνήθως εξάγονται σε αρχεία δεδομένων του τύπου .h5, .h5ad, .rds [**6**] , [**15**] , [**48**] , [**49**] .

Αναλυτικότερα σχετικά με την εγκατάσταση του Scanpy, ο χρήστης πρέπει και σε αυτή την περίπτωση όπως και του Seurat να εγκαταστήσει και να ακολουθήσει μια σειρά βημάτων για να ετοιμάσει την επιφάνεια εργασίας της μελέτης του. Εκτός του ότι πρέπει να εγκαταστήσει κάποια από τις διαθέσιμες εκδόσεις Python (3.8, 3.9, 3.10) όπου συνήθως προτιμάται η εγκατάσταση μια παλαιότερης και όχι της πρόσφατα ανανεωμένης έκδοσης, ο χρήστης εγκαθιστά κάποιο εργαλείο καταγραφής κώδικα. Ακολουθεί η εγκατάσταση του Jypyter Notebook. Έπειτα γίνεται η εγκατάσταση του Scanpy και μια σειρά ακόμα πακέτων. Αυτή είναι μια συνοπτική εκδοχή εγκατάστασης και προετοιμασίας του περιβάλλον για ανάλυση δεδομένων μέσα στο Jypyter Notebook [**6**] , [**15**] , [**48**] , [**49**] .

Όπως και για το Seurat, έτσι και το Scanpy πακέτο ανάλυσης, συνήθως στην βιβλιογραφία συναντάται να πραγματοποιεί αναλύσεις δεδομένων στα αρχικά στάδια, και έπειτα τα αποτελέσματα εισάγονται σε άλλες πλατφόρμες για περαιτέρω διερεύνηση [**6**] , [**15**] , [**48**] , [**49**] .

6.5 Gene Pattern Notebook

Γλώσσα προγραμματισμού: Python

Λειτουργικό σύστημα: ONLINE (MacOS, Linux, Microsoft Windows)*

Το πιο σημαντικό χαρακτηριστικό της πλατφόρμας Gene Pattern Notebook που την διαφοροποιεί από τις υπόλοιπες που ασχολούνται με την μέθοδο single-cell RNA sequencing είναι το ότι το Gene Pattern Notebook είναι σχεδιασμένο με τέτοιο τρόπο ώστε να είναι φιλικό και εύκολο στην αλληλεπίδραση του χρήστη με το υπολογιστικό περιβάλλον και να μην χρειάζεται ο χρήστης να γνωρίζει κάποια γλώσσα προγραμματισμού συντάσσοντας κάποιο κώδικα για να εκκινήσει την διαδικασία. Το Gene Pattern Notebook επεκτείνει την λειτουργία του ενσωματώνοντας το Jypyter Notebook. Η πλατφόρμα εκτελεί την ανάλυση single-cell-RNA-sequencing δεδομένων προσφέροντας τις επιλογές που αναφέρονται στον παραπάνω πίνακα χωρίς να χρειάζεται ο χρήστης να κατεβάσει πακέτα αρχείων εγκατάστασης, παρά μόνο με την χρήση της ιστοσελίδας. Συνδέεται στον λογαριασμό που έχει δημιουργήσει από οποιαδήποτε μηχανή αναζήτησης χρησιμοποιεί και προτιμά και μεταφέρεται σε ένα εικονικό περιβάλλον, server όπου μπορεί να πραγματοποιήσει την ανάλυση του. Η ανάλυση είναι συγκεντρωμένη σε ένα και μόνο εικονικό χώρο, στην ίδια σελίδα της ιστοσελίδας. Τα αποτελέσματα της ανάλυσης είναι δυνατόν να διαμοιραστούν και σε άλλους χρήστες. Φυσικά, το περιβάλλον είναι διαμορφωμένο να φιλοξενήσει και χρήστες που θα ήθελαν να προσεγγίσουν την ανάλυση και από μια προγραμματιστική οπτική [**6**] , [**50**] , [**59**].

Ο χρήστης ανεβάζει στην πλατφόρμα ένα αρχείο τύπου .txt, .csv, .tsv, ή ένα σετ τριών αρχείων δεδομένων των οποίων η προεπεξεργασία έχει γίνει από το πρωτόκολλο 10X Genomics , διευκρινίζοντας την μορφή των δεδομένων του σε επιλογή της πλατφόρμας εάν οι σειρές στον πίνακα δεδομένων αναπαριστούν τα γονίδια και οι στήλες αναπαριστούν τα κύτταρα. Μόλις ολοκληρωθεί με επιτυχία το ανέβασμα των αρχείων, εμφανίζονται γραφήματα τα οποία αφορούν μερικούς παράγοντες που εξετάζονται . Χαρακτηριστικά αναφέρονται κάποιοι από αυτούς τους παράγοντες: ο αριθμός των γονιδίων κάθε κυττάρου, η τελική καταμέτρηση για κάθε κύτταρο, και ένα ποσοστό των καταμετρήσεων στις οποίες χαρτογραφήθηκαν μιτοχονδριακά γονίδια όποτε αυτό είναι διαθέσιμο. Κατά την οπτική απεικόνιση ο χρήστης και πάλι έχει τη δυνατότητα μέσω επιλογών να προσαρμόσει και να εμφανίσει τα δεδομένα του όπως χρειάζεται να

παρουσιαστούν. Συγκεκριμένα ο χρήστης έχει τη δυνατότητα της οπτικοποίησης διαφοροποιημένων γονιδίων ενός κυτταρικού πληθυσμού σε σχέση με κάποιο άλλο κυτταρικό πληθυσμό [**6**] , [**50**] , [**59**] .

Τέλος, ο τρόπος εξαγωγής των δεδομένων χωρίζεται σε δυο κατηγορίες. Ο πρώτος τρόπος είναι η εξαγωγή δεδομένων σε CSV αρχείο το οποίο συνοδεύεται με μια περιγραφή που αφορά τον πρώτο πίνακα των δεδομένων, τον χαρακτηρισμό των κυττάρων , τα αποτελέσματα οπτικοποίησης σε άλλη διάσταση και η σειρά κατάταξης των γονιδίων. Ο δεύτερος τρόπος αφορά την εξαγωγή σε H5AD αρχείο το οποίο μπορεί να επαναχρησιμοποιηθεί από το λογισμικό. Οι επιλογές και τα φίλτρα που έχει επιλέξει ο χρήστης να θέσει κατά την διάρκεια της ανάλυσης καταγράφονται [**6**] , [**50**] , [**59**] .

6.6 scGEAToolbox

Γλώσσα προγραμματισμού: MATLAB

Λειτουργικό σύστημα: Microsoft Windows(εφαρμογή)*, ONLINE

Το συγκεκριμένο υπολογιστικό εργαλείο βρίσκεται τόσο σε online μορφή στο προγραμματιστικό περιβάλλον της MATLAB, όπως επίσης δίνει την δυνατότητα εγκατάστασης τοπικά στον υπολογιστή χρησιμοποιώντας το MATLAB Runtime version 9.11 χωρίς να απαιτείται αγορά του λογισμικού. Η επιλογή αυτή προσφέρει ένα ιδιαίτερα προσβάσιμο, φιλικό προς τον χρήστη περιβάλλον χωρίς να απαιτεί καμία γνώση προγραμματισμού σε περιβάλλον της MATLAB και μόνο με την αλληλεπίδραση με τις πολλές επιλογές που προσφέρονται ο χρήστης μπορεί να διερευνήσει όλα αυτά που καταγράφονται στον παραπάνω πίνακα. Από την άλλη, η δεύτερη επιλογή (online ή και με την χρήση του MATLAB που έχει εγκατασταθεί τοπικά), η οποία θα χρησιμοποιηθεί παρακάτω απαιτεί την ικανότητα σκέψης, ανάλυσης και κατανόησης μερικών προγραμματιστικών εντολών σε περιβάλλον MATLAB, από την εγκατάσταση του έως και την υλοποίηση των αναλύσεων. Το περιβάλλον προσφέρει μια πληθώρα συναρτήσεων, αρκετές από τις οποίες ο χρήστης μπορεί και να τροποποιήσει δημιουργώντας μια άλλη προσέγγιση στα δεδομένα. Φυσικά, το περιβάλλον περιλαμβάνει και αλληλεπίδραση του χρήστη με κουμπιά και άλλες επιλογές προσφέροντας μια εμπειρία πλοήγησης μέτριας δυσκολίας. Περιλαμβάνονται πληθώρα συναρτήσεων επεξεργασίας και ανάλυσης των δεδομένων που αφορούν την κανονικοποίηση (normalization), την αφαίρεση και παρουσίαση των batch-effect παραγόντων, την μέθοδο της συσταδοποίησης (clustering), την

οπτικοποίηση των δεδομένων με την μέθοδο t-SNE (visualization) και πολλές ακόμη επιλογές που αναφέρονται στον πίνακα παραπάνω [**6**] , [**51**] , [**58**] .
Ιδιαίτερα σημαντική πληροφορία αφορά τα αρχεία που μπορεί να δεχτεί ως είσοδο. Τα αρχεία αυτά στην online μορφή μπορούν να αποθηκευτούν στο MATLAB Drive το οποίο διαθέτει χωρητικότητα 20GB και να απελευθερώσουν το σύστημα του υπολογιστή από έναν μεγάλο όγκο δεδομένων. Τα δεδομένα που μπορούν να εισαχθούν μπορεί να είναι από την πλατφόρμα 10X Genomics που απαιτεί τρία αρχεία (ένα πίνακα δεδομένων, ένα αρχείο με τα ονόματα των γονιδίων, και ένα με τα barcodes), να έχουν επεξεργαστεί πριν από τα πρωτόκολλα Seurat και Drop-seq ή να μην έχει γίνει κάποια επεξεργασία πριν και ο χρήστης να ανεβάσει τα δεδομένα του στην αρχική τους μορφή. Είναι σημαντικό να αναφερθεί πως προσφέρονται και άλλες επιλογές εισαγωγής των δεδομένων [**6**] , [**51**] , [**58**] .

6.7 BBrowser

Γλώσσα προγραμματισμού: -

Λειτουργικό σύστημα: ONLINE (MacOS, Microsoft Windows)*

Το BBrowser είναι ένα online εργαλείο που δεν εμφανίζει στον χρήστη καμιά επιλογή ή καταγραφή κώδικα, με ένα ιδιαίτερα φιλικό προς τον χρήστη, κατανοητό περιβάλλον πλοήγησης και διερεύνησης. Πρόκειται για μια εφαρμογή που την κατεβάζει ο χρήστης τοπικά στον υπολογιστή του. Ασχολείται με την διερευνητική ανάλυση ακόμη και των πιο απαιτητικών σε όγκο δεδομένων και αυτά οπτικοποιούνται με τις πιο σύγχρονες μεθόδους. Το online εργαλείο αποτελείται από 3 στοιχεία: μια βάση δεδομένων, μια αναλυτική ροή διερεύνησης και μια μέθοδο οπτικής απεικόνισης [**6**] , [**52**] .
Τα αρχεία των δεδομένων βρίσκονται στο υπολογιστικό σύστημα είτε αποθηκευμένα σε κάποιο server. Το BBrowser μπορεί να επεξεργαστεί και να αναλύσει έναν τεράστιο όγκο δεδομένων, να αναλύσει συγκεκριμένους κυτταρικούς πληθυσμούς χωριστά, να αποθηκεύσει τα μετα-δεδομένα και πολλές ακόμη επιλογές που αναφέρονται στον παραπάνω πίνακα [**6**] , [**52**] .
Τα αρχεία που εισάγονται αφορούν είτε προ επεξεργασμένα δεδομένα είτε και μη επεξεργασμένα. Τα συνηθισμένα πρωτόκολλα προ επεξεργασίας των δεδομένων είναι το 10X Genomics, Seurat, Scanpy κ.α. Τα δεδομένα μπορούν να είναι του τύπου .tsv, .csv, .mtx. Έπειτα από την ανάλυση τους μπορούν να διατηρηθούν ή και ο ερευνητής να τα μοιραστεί με την μέθοδο διαμοιρασμού με άλλους χρήστες της πλατφόρμας, ή

να κατεβάσει άλλα δεδομένα και να επικοινωνήσει με ερωτήσεις και σχόλια σχετικά με την χρήση της πλατφόρμας και άλλα δεδομένα και αναλύσεις [**6**] , [**52**] .

7 Ανάλυση και συμπεράσματα

Τα δεδομένα θα αναλυθούν σε 2 υπολογιστικές πλατφόρμες την Gene Pattern Notebook, το scGEATool. Σκοπός είναι η ανακάλυψη και αξιολόγηση των δυνατοτήτων των συγκεκριμένων πλατφορμών. Η επιλογή έγινε με βάση τα κριτήρια που περιγράφονται στην εισαγωγική παράγραφο του προηγούμενου κεφαλαίου. Συγκεκριμένα, η συγκεντρωτική ανάλυση σε μια εφαρμογή ή μια online ιστοσελίδα που θα μπορεί να υποστηρίξει μια πλήρη ανάλυση (ποιοτικό έλεγχο, κανονικοποίηση, συσταδοποίηση, έλεγχο διαφοροποιημένων γονιδίων, οπτική απεικόνιση, ταυτοποίηση, μείωση διάστασης των δεδομένων κ.α.) με την ελάχιστη έως καθόλου συγγραφή κώδικα και την φιλοξενία σε ένα περιβάλλον φιλικό προς τον χρήστη και εύκολο για περιήγηση με την ελάχιστη δέσμευση υπολογιστικών πόρων αποτελεί το ιδανικό εργαλείο για πραγματοποίηση διερευνητικής ανάλυσης δεδομένων τύπου single-cell RNA-sequencing.

7.1 Δεδομένα και τρόπος ανάλυσης

Τα πακέτα δεδομένων που θα αναλυθούν αφορούν σύνολα δεδομένων τύπου single-cell RNA τα οποία βρίσκονται στην βάση δεδομένων Gene Expression Omnibus (GEO) και ο συγκεκριμένος τύπος δεδομένων αφορά δεδομένα γονιδιακής έκφρασης για τα οποία διεξάγεται διερευνητική ανάλυση δεδομένων (exploratory data analysis). Με την αναζήτηση παρόμοιων καταχωρήσεων ο χρήστης μπορεί να βρει πληροφορίες για τα πακέτα δεδομένων ή και να αποθηκεύσει για πραγματοποίηση ανάλυσης δεδομένα έχοντας πρόσβαση και σε άλλες βάσεις δεδομένων. Εάν ο χρήστης επιθυμεί να αναζητήσει πακέτα δεδομένων σε single-cell RNA τύπο μπορεί να χρησιμοποιήσει και την βάση δεδομένων scRNASeqDB. Η βάση δεδομένων GEO προσφέρει αναζήτηση των καταχωρήσεων με τον χρήστη να γνωρίζει τον κωδικό πρόσβασης π.χ. 'GSE74923'. Η δεύτερη βάση δεδομένων δίνει περισσότερες πληροφορίες απευθείας στον χρήστη, δηλαδή μέσω μιας λίστας ο χρήστης βρίσκει λεπτομέρειες όπως τον κωδικό πρόσβασης που μπορεί να μεταφέρει απευθείας τον χρήστη στην πρώτη βάση δεδομένων, μια σύντομη περιγραφή των δεδομένων καθώς και τους τύπους των κυτταρικών ομάδων. Η παράλληλη αναζήτηση σε αυτές τις 2 βάσεις δεδομένων βοήθησε στην εύρεση των κυτταρικών δειγμάτων που θα αναλυθούν [**53**] , [**56**] .

Τα δεδομένα τύπου single-cell RNA sequencing που θα αναλυθούν είναι: 'GSE74923', 'GSE75688'. Και τα δυο πακέτα δεδομένων περιλαμβάνουν δεδομένα που έχουν υποστεί προ-επεξεργασία. Τι σημαίνει αυτό; Τα δεδομένα που συλλέχθηκαν έπειτα από την ανάλυση σε εργαστήριο κανονικοποιήθηκαν, επιλέχθηκαν και αφαιρέθηκαν τα δεδομένα που προκαλούν ζημιά (π.χ. θόρυβος), επιλέχθηκαν τα δεδομένα που χρειάζεται να αναλυθούν (π.χ. κάποια γονίδια) και μειώθηκε η διάσταση των δεδομένων [**15**] , [**54**] , [**55**] , [**56**] .

Αναλυτικότερα,

'GSE74923': Αυτό το πακέτο δεδομένων εξετάζει τις μεταλλάξεις και τον κυτταρικό κύκλο εξαρτημένων προφίλ 2 κυτταρικών τύπων του οργανισμού Mus musculus [**54**] .

'GSE75688': Πραγματοποιήθηκε single-cell-RNA sequencing ανάλυση δείγματος σε αρχικά στάδια καρκίνου του στήθους από 11 ασθενείς [**55**] .

Σχετικά με τον τρόπο ανάλυσης, τα εργαλεία πραγματοποιούν διερευνητική ανάλυση δεδομένων. Ως exploratory data analysis ορίζεται η ανάλυση και η διερεύνηση δεδομένων (πολλών γνωρισμάτων) που μελετήθηκαν των οποίων τα αποτελέσματα παρουσιάζονται με τις διάφορες μεθόδους οπτικής απεικόνισης (π.χ. t-SNE). Η διαδικασία αφορά την διατύπωση ερωτήσεων και υποθέσεων και την πραγματοποίηση αναλύσεων και διερεύνησης για την εύρεση απαντήσεων. Χρησιμοποιούνται στατιστικές μέθοδοι, μαθηματικά μοντέλα, βιολογικές μέθοδοι, αλγόριθμοι κ.α. για την αναζήτηση των συσχετίσεων και μοτίβων και σχέσεων αλληλεπίδρασης μεταξύ των τιμών των παραμέτρων οι οποίες σχέσεις φανερώνονται μέσω γραφημάτων, ιστογραμμάτων, box plots (minimum, maximum, median), charts, bubble charts , heatmaps κ.α. Μεταξύ των υπολογιστικών εργαλείων που θα χρησιμοποιηθούν για ανάλυση είναι η συσταδοποίηση (clustering), η μείωση των υψηλά διαστατικών δεδομένων σε χαμηλής διαστατικότητας για να επιτευχθεί η αναπαράσταση, η οπτικοποίηση (visualization) με στατιστικά στοιχεία, γραφήματα , ο υπολογισμός των σχέσεων μεταξύ των μεταβλητών, οι αλγόριθμοι όπως ο k-means clustering ή ο k-nearest-neighbor. Οι περισσότερες πλατφόρμες που αποτελούν τα εργαλεία ανάλυσης χρησιμοποιούν τις γλώσσες Python και R ενώ τελευταία έχει μπει στο προσκήνιο και το προγραμματιστικό περιβάλλον MATLAB, αν και τα τελευταία χρόνια αναπτύσσονται υπολογιστικά εργαλεία που δεν απαιτούν καμία γνώση

προγραμματισμού καθιστώντας έτσι ένα πιο φιλικό και εύκολο στο χρήστη περιβάλλον [**14**] .

7.2 Πλατφόρμες ανάλυσης και συμπεράσματα

7.2.1 Gene Pattern Notebook

Η πρώτη προσέγγιση αφορά το online εργαλείο Gene Pattern Notebook. Έπειτα από την εγγραφή, στο προφίλ του χρήστη υπάρχει μια δημόσια βιβλιοθήκη με εργασίες/πακέτα που μπορεί να ξαναχρησιμοποιήσει ο χρήστης για να εκτελέσει την δική του ανάλυση. Μερικά από τα σύνολα ολοκληρωμένων μεθόδων που προσφέρονται είναι και η επιλογή του Single-Cell RNA-seq Clustering Analysis Notebook που χρησιμοποιήθηκε για αυτή την μελέτη. Το περιβάλλον που πραγματοποιείται η ανάλυση είναι ένας server. Ο χρήστης ανεβάζει τα αρχεία του επιλέγοντας μια από τις 2 μορφές δεδομένων εισαγωγής που προσφέρονται. Η ροή της μελέτης πραγματοποιείται σειριακά στο ίδιο περιβάλλον (παράθυρο της ιστοσελίδας). Τα στάδια της ανάλυσης είναι τα εξής: 1) Setup Analysis (ανέβασμα του αρχείου δεδομένων), 2) Preprocess Counts (Quality control analysis normalization, HVGs, dimensional reduction), 3) Cluster cells (clustering), 4)Visualize Cluster Markers, 5) Export Analysis Data (εξαγωγή αποτελεσμάτων). Αν και η ροή και τα βήματα φαίνονται απλά και κατανοητά προς τον ερευνητή, και οι χρόνοι ολοκλήρωσης κάθε βήματος εκτιμώνται από 5-90 sec ανάλογα τον αριθμό των κυττάρων και των γονιδίων και την ταχύτητα σύνδεσης στο διαδίκτυο, εντούτοις τα πράγματα εξελίχθηκαν διαφορετικά [**59**] .

Η ανάλυση δοκιμάστηκε να πραγματοποιηθεί τόσο σε διαφορετικά λειτουργικά συστήματα (Microsoft Windows, Mac OS) όσο και με την χρήση διαφορετικών συνδέσεων στο διαδίκτυο με διαφορετικές ταχύτητες (upload, download) [**59**] .

Κατά την προσπάθεια ανάλυσης σε υπολογιστή με λειτουργικό σύστημα Microsoft Windows και σύνδεση σε οικιακό δίκτυο ο χρόνος ανεβάσματος του αρχείου ήταν τα 20-30 λεπτά. Το επόμενο βήμα δηλαδή το 2) όπως περιγράφεται στην παραπάνω παράγραφο δοκιμάστηκε με διαφορετικούς συνδυασμούς στις τιμές των παραμέτρων (minimum number of cells, minimum number of genes, maximum number of genes,

minimum total counts, maximum total counts, minimum & maximum % of mito genes, log normalize, perform regression). Τελικά κατά το βήμα 2) αναμένοντας έως και 3 ώρες ανάλυσης δεν προέκυψε κάποιο αποτέλεσμα. Η συνέχεια στο βήμα 3) απαιτεί την ολοκλήρωση του βήματος 2) για αυτό και δεν πραγματοποιήθηκε [**59**] .

Κατά την προσπάθεια ανάλυσης σε υπολογιστή με λειτουργικό σύστημα Mac OS και σύνδεση σε φορητό δίκτυο ο χρόνος ανεβάσματος του αρχείου ήταν μόλις τα 2-3 λεπτά. Δοκιμάστηκαν οι ίδιοι συνδυασμοί των τιμών των παραμέτρων όσο και οι τιμές που δίνονται ως παραδείγματα από την πλατφόρμα ανάλυσης χωρίς κάποιο αποτέλεσμα [**59**] .

Το πρόβλημα εντοπίζεται στα λάθη που ενδεχομένως υπάρχουν στην δομή του αρχείου. Η έλλειψη δεδομένων και η εμφάνιση των λαθών σε προγραμματιστικό κώδικα δυσκολεύει τον χρήστη να κατανοήσει που εμφανίζεται το πρόβλημα. Εκτός αυτού κατά το βήμα 1) ο χρήστης μπορεί να καταλάβει σε τι στάδιο βρίσκεται το ανέβασμα του αρχείου καθώς μια πράσινη ένδειξη εμφανίζει τα ποσοστά ολοκλήρωσης του ανεβάσματος. Στο βήμα 2) όπου αναφέρεται η λέξη progress δεν μπορεί να γίνει κατανοητό από τον χρήστη πόσος χρόνος ακόμα χρειάζεται για να ολοκληρωθεί η ανάλυση και η εμφάνιση των αποτελεσμάτων αυτού του βήματος [**59**] .

Συνεπώς, αυτές είναι κάποιες παρατηρήσεις που προκύπτουν από την ανάλυση στην πλατφόρμα Gene Pattern Notebook που δεν έγινε εφικτό να ολοκληρωθεί.

7.2.2 scGEAToolbox

Μια δεύτερη προσέγγιση δεδομένων τύπου από τεχνικές single-cell RNA-sequencing έγινε με την χρήση του εργαλείου scGEAToolBox. Πρόκειται για ένα σύνολο υπορουτινών σε περιβάλλον MATLAB. Η ανάλυση πραγματοποιήθηκε με την σύνδεση σε λογαριασμό στο MATLAB online [**58**] .

Αρχικά για την εγκατάσταση του, σε ένα αρχείο τύπου .m καταγράφεται ο παρακάτω κώδικας:

```
1   tic
2   fprintf('Installing scGEAToolbox...')
3   unzip('https://github.com/jamesjcai/scGEAToolbox/archive/main.zip');
4   addpath('./scGEAToolbox-main');
5   toc
6   if exist('scgeatool.m','file')
7       fprintf('scGEAToolbox installed!')
8   end
9   savepath;
10
```

“ ©[[**91**](Quick installation scGEAToolbox)] via Quick installation — scGEAToolbox documentation ”

Εικόνα 1: Κώδικας εγκατάστασης πακέτου scGEATool στο περιβάλλον του MATLAB

Με την εκτέλεση του παραπάνω κώδικα εγκαθίσταται ένα πακέτο αρχείων και προετοιμάζεται το περιβάλλον MATLAB για την ανάλυση των δεδομένων στην εφαρμογή scGEAToolbox. Μόλις ολοκληρωθεί η εγκατάσταση και εμφανιστεί στο Command Window το μήνυμα όπως φαίνεται στην γραμμή 7 (‘scGEAToolbox installed!’) πληκτρολογούμε την εντολή scgeatool όπως φαίνεται παρακάτω και αμέσως ένα παράθυρο αλληλεπίδρασης εμφανίζεται [**58**] .

```
Command Window
New to MATLAB? See resources for Getting Started.
>> scgeatool
>>
```

“ ©[[**91**](Getting Stared scGEAToolbox)] via Getting Started — scGEAToolbox documentation”

Εικόνα 2: Εντολή εκκίνησης πακέτου scGEATool στο περιβάλλον του MATLAB

Πλέον το περιβάλλον MATLAB είναι έτοιμο για ανάλυση.

Επειδή τα δεδομένα που χρησιμοποιούνται είναι προ-επεξεργασμένα, αυτό σημαίνει πως έχει συμβεί ο ποιοτικός έλεγχος, η κανονικοποίηση, το data correction και η μείωση της διάστασης των δεδομένων, ο χρήστης δεν χρειάζεται να εκτελέσει κάποιο από αυτά εκτός αν κρίνεται απαραίτητο. Σε περίπτωση που τα αρχεία χαρακτηριζόταν

ως raw data αυτό σημαίνει πως ο χρήστης θα έπρεπε να ακολουθήσει μια άλλη προσέγγιση πραγματοποιώντας τα παραπάνω βασικά για την ποιότητα των δεδομένων βήματα [**58**] .

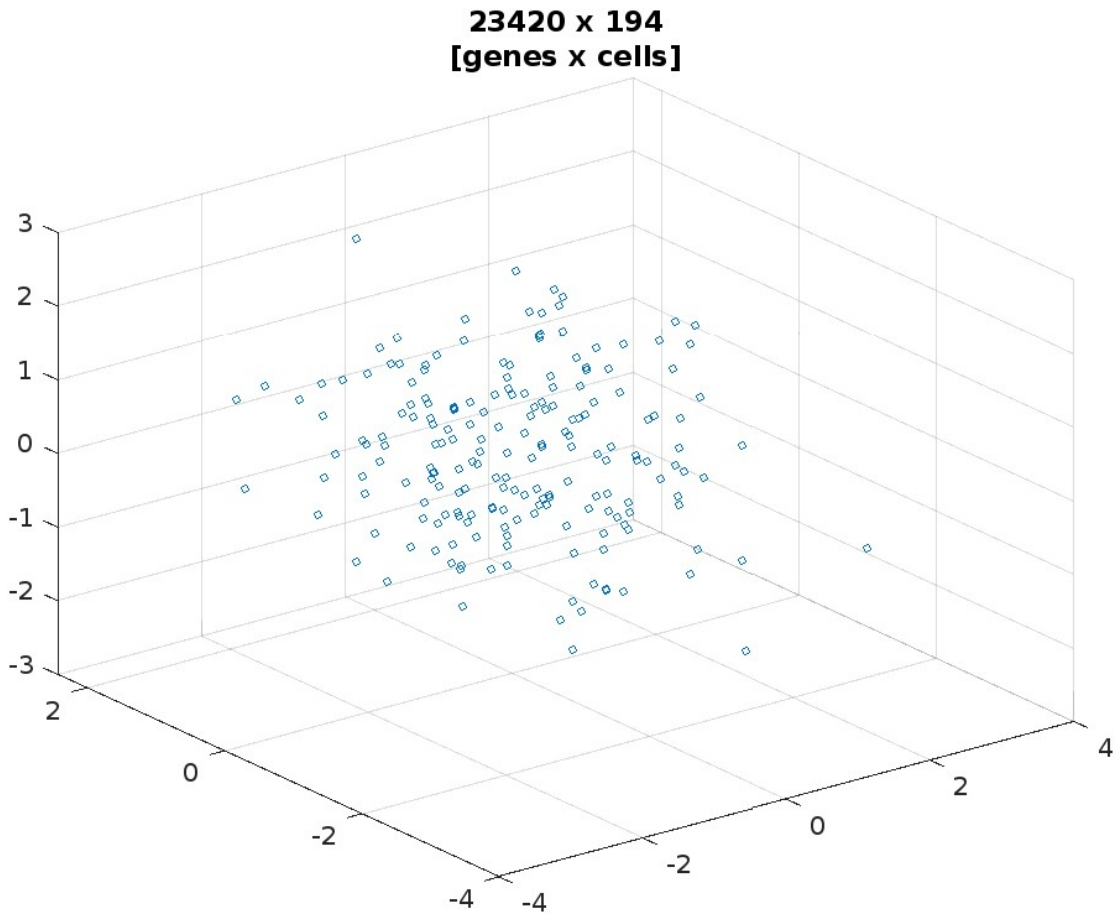
Μέσα στο φάκελο example_data ανέβασα το αρχείο που θέλω να πραγματοποιήσω την ανάλυση. Από το μενού επιλογών, επιλέγοντας το Home πάτησα την επιλογή Upload όπου επιλέγω τον φάκελο τοπικά στον υπολογιστή όπου βρίσκεται το αρχείο δεδομένων. Με αυτό τον τρόπο το MATLAB Drive διευκολύνει την εξοικονόμηση αποθηκευτικού χώρου καθώς ο χρήστης μπορεί να διαγράψει το αρχείο από τον υπολογιστή [**58**] .

Ο χρήστης πληκτρολογώντας ξανά την εντολή scgeatool, θα μπορέσει να επιλέξει μια από τις μορφές δεδομένων που δίνονται στο παράθυρο αλληλεπίδρασης για να ξεκινήσει η ανάλυση [**58**] .

GSE74923:

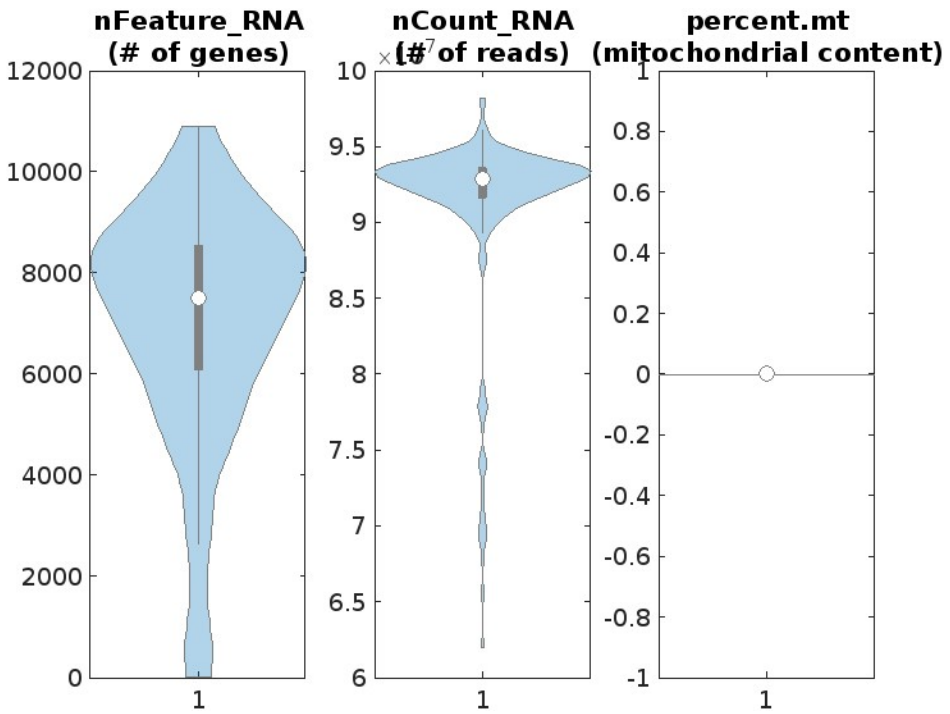
Το αρχείο που ανέβηκε, ήταν τύπου .csv και αφορούσε το πακέτο δεδομένων GSE74923.csv.

Το πρώτο γράφημα που εμφανίζεται, αφορά την απεικόνιση σε 3D των κυττάρων του εν λόγο πακέτου. Στο τίτλο εμφανίζεται το πλήθος των στοιχείων του πίνακα (23420 x 194, genes x cells). Αυτό σημαίνει πως ο πίνακας εισόδου περιλαμβάνει ως γραμμές το πλήθος κάθε γονιδίου που εμφανίζεται σε καθένα από τα κύτταρα [**54**] , [**58**] .



"©[(Results scGEAToolbox)] via scGEAToolbox"
Εικόνα 3: Διαστάσεις δεδομένων εισόδου του πακέτου "GSE74923"

Έπειτα, η επιλογή που γίνεται αφορά την διερεύνηση και εμφάνιση των αντίστοιχων plot που αφορούν τις λεπτομέρειες σχετικά τις παραμέτρους που εξετάζονται κατά τον ποιοτικό έλεγχο [**54**] , [**58**] .

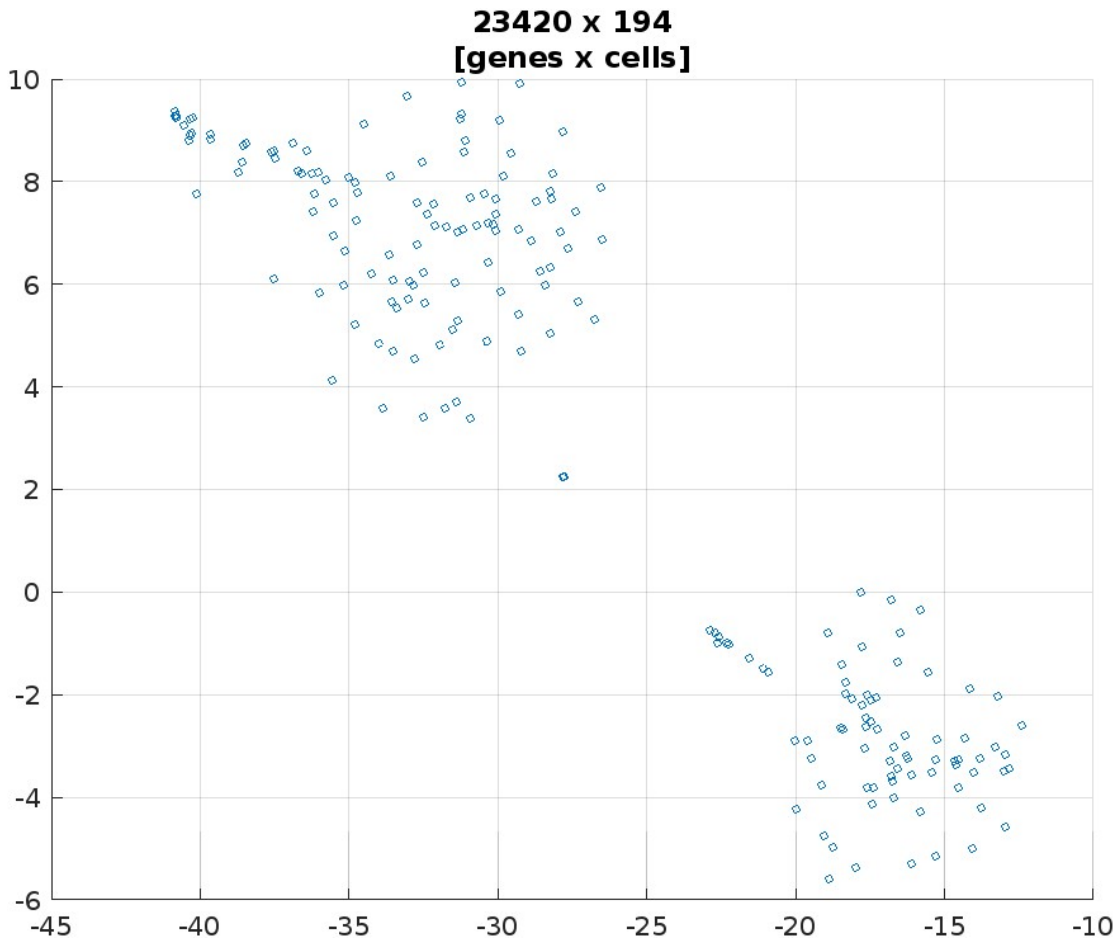


"©[(Results scGEAToolbox)] via scGEAToolbox "

Εικόνα 4: Παράμετροι ποιοτικού ελέγχου των δεδομένων εισόδου

Τα violin plots όπως χαρακτηρίζονται τα γραφήματα αυτής της μορφής, δίνουν πληροφορίες σχετικά με τον αριθμό των γονιδίων ανά κύτταρο (nFeature_RNA). Το πρώτο γράφημα λοιπόν μας ενημερώνει για το ότι το μεγαλύτερο μέρος του δείγματος αποτελείται από κύτταρα με τουλάχιστον λίγο κάτω από τα 5000 γονίδια και το πολύ κοντά στα 11000 γονίδια. Υπάρχει και ένα μικρό ποσοστό κυττάρων με λιγότερα από 4000 γονίδια και αυτό γίνεται κατανοητό από την πυκνότητα του γραφήματος. Το δεύτερο γράφημα αφορά τον αριθμό UMIs που εκφράζονται ανά κύτταρο(nCount_RNA). Από αυτό ο χρήστης κατανοεί πως τα περισσότερα κύτταρα εκφράζουν κατά μέσο όρο περίπου 9,3 UMIs. Τέλος, το τρίτο γράφημα που αφορά τα μιτοχονδριακά γονίδια (percent.mt) όπου παρατηρείται πως δεν υπάρχουν καθόλου μιτοχονδριακά γονίδια [**54**] , [**58**] .

Στη συνέχεια της ανάλυσης ο χρήστης με την εφαρμογή του αλγορίθμου t-SNE μπορεί να παρατηρήσει τα κύτταρα σε γράφημα 2 διαστάσεων 2D [**54**] , [**58**] .

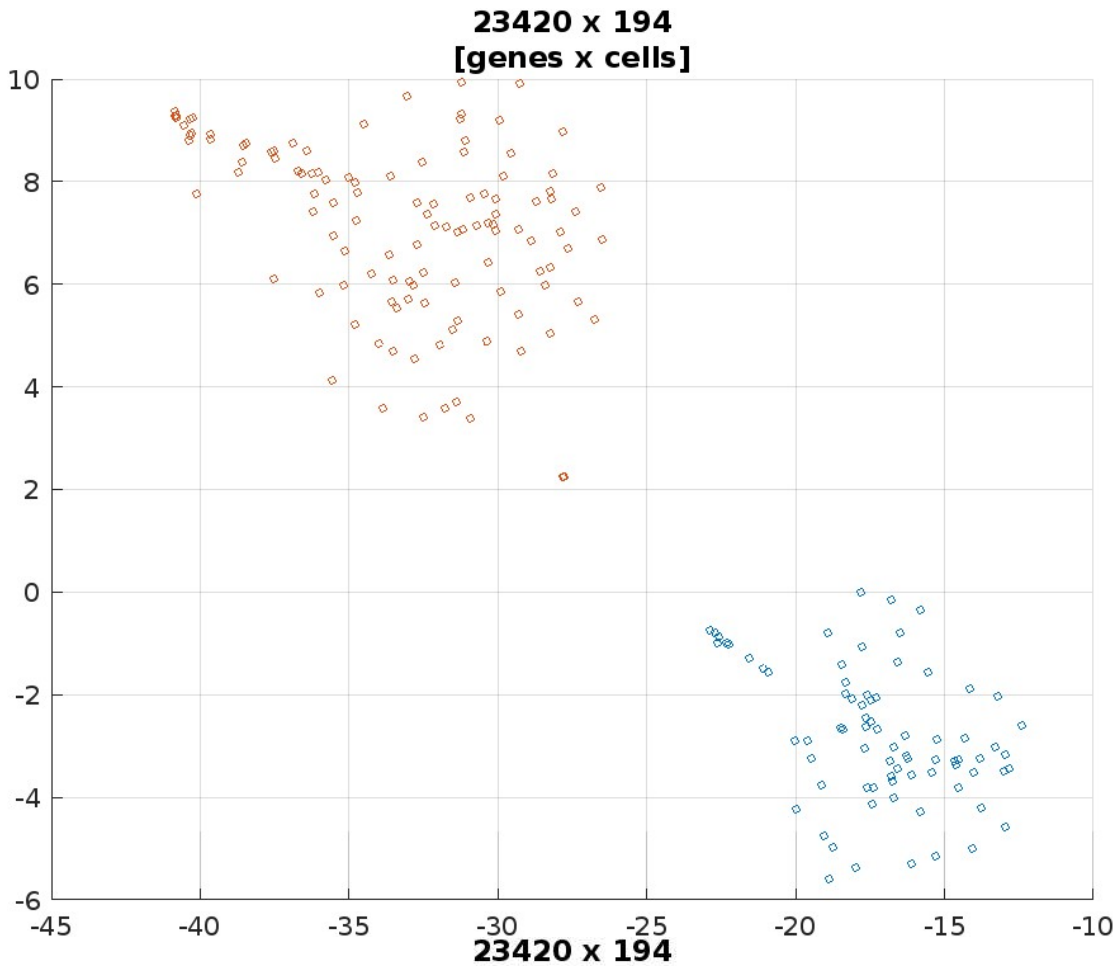


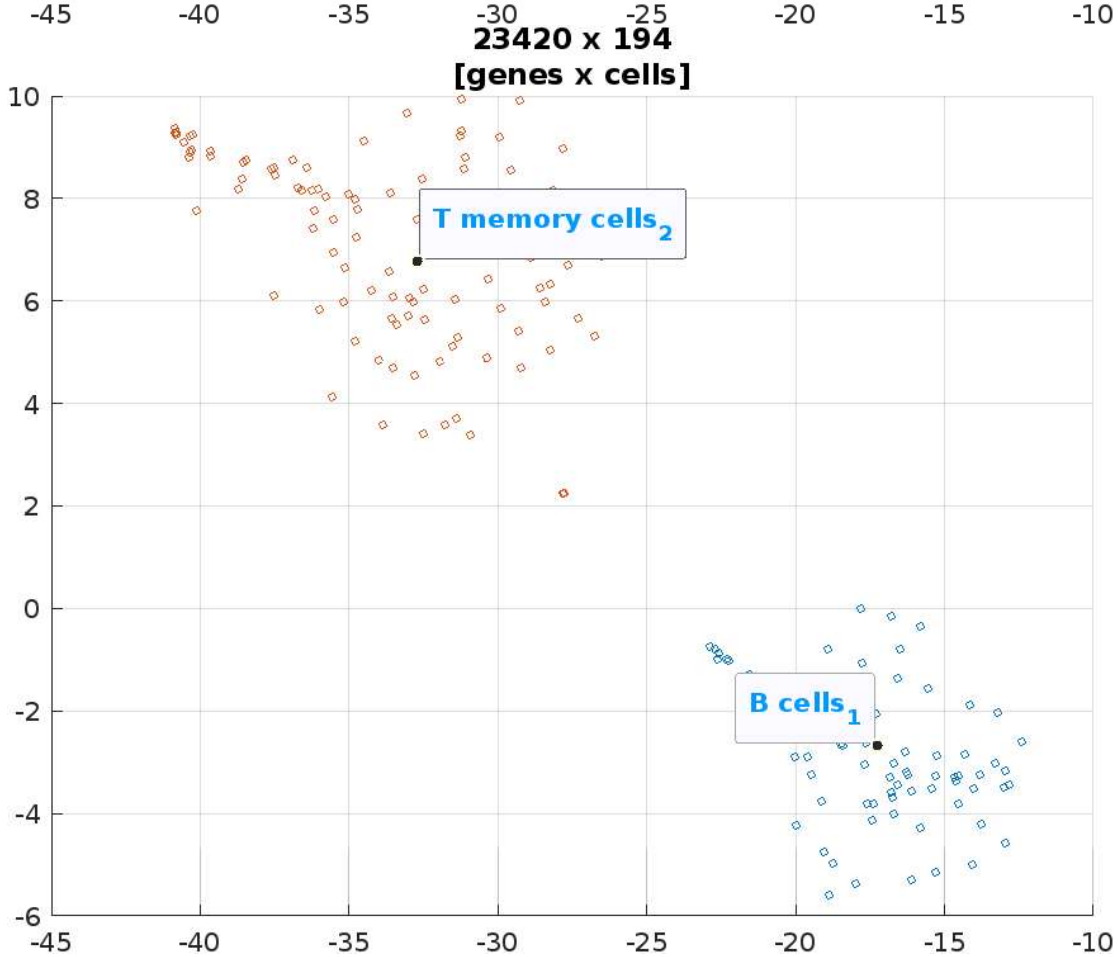
"©[(Results scGEAToolbox)] via scGEAToolbox "

Εικόνα 5: Εφαρμογή του αλγορίθμου t-SNE

Η μείωση της διάστασης διευκολύνει τον χρήστη να κατανοήσει πως τα δεδομένα μάλλον ταξινομούνται σε δυο συστάδες, από την μορφή που έχουν στον χώρο οπτικής απεικόνισης [**54**] , [**58**] .

Έτσι κατά την πραγματοποίηση της συσταδοποίησης, την επιλογή του πλήθους των συστάδων ως 2 και την ταυτοποίησης προκύπτουν τα ακόλουθα γραφήματα:


23420 x 194
[genes x cells]
10
8
6
4
2
0
-2
-4
-6
-45
-40
-35
-30
-25
-20
-15
-10


23420 x 194
[genes x cells]
T memory cells2
B cells1
10
8
6
4
2
0
-2
-4
-6
-45
-40
-35
-30
-25
-20
-15
-10

“©[(Results scGEAToolbox)] via scGEAToolbox”

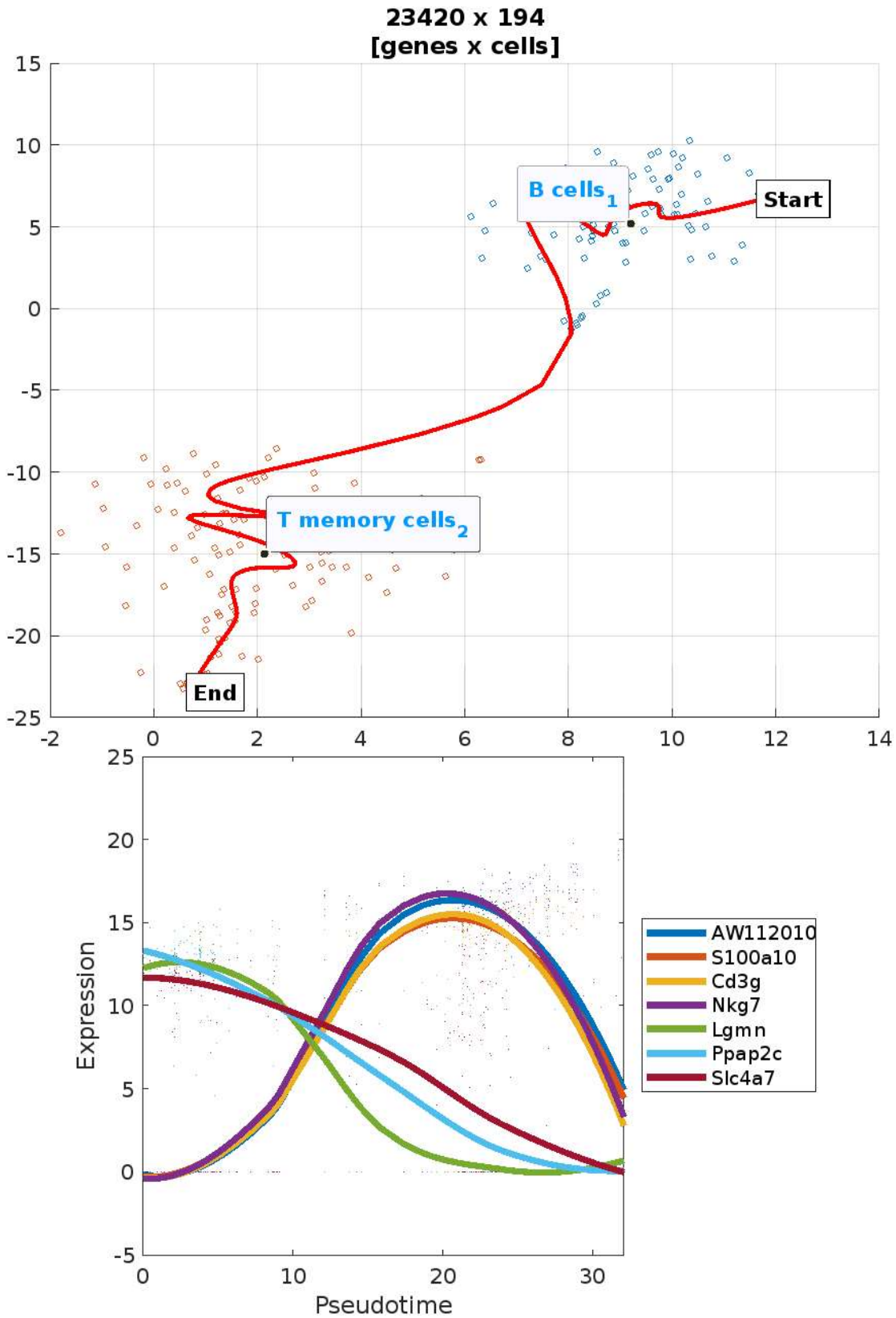
Εικόνα 6: Πρώτη εικόνα: Συσταδοποίηση σε 2 συστάδες, Δεύτερη εικόνα: Κυτταρική ταυτοποίηση των συστάδων

Από την ταυτοποίηση προέκυψε όπως φαίνεται ότι η ομάδα 1 αφορά Β-κύτταρα και η ομάδα 2 αφορά Τ-κύτταρα μνήμης [**54**] , [**58**] .

Τα marker genes είναι αυτά που βοηθούν τις υπολογιστικές πλατφόρμες να βρουν ποια κυτταρική ομάδα αφορά η εξέταση των δειγμάτων, καθώς καθεμία εκφράζει κάποιο συγκεκριμένο γονίδιο. Είναι σημαντικό να αναφερθεί πως το δείγμα είναι του οργανισμού Mus musculus και κατά την ταυτοποίηση ο χρήστης χρειάζεται να γνωρίζει αυτή την πληροφορία για τα δεδομένα που εξετάζει καθώς καλείται να επιλέξει ανάμεσα στις επιλογές Human, Mouse, Zebrafish [**54**] , [**58**] .

Σε επόμενο στάδιο διερευνάται η trajectory ανάλυση που αφορά την διερεύνηση των σταδίων ανάπτυξης και διαφοροποίησης. Σχετικά με αυτό υπάρχουν πολλές επιλογές όπως η ανάπτυξη του Wilcoxon rank sum test και του DeSeq2 που είναι οι πιο γρήγορες μέθοδοι και τα αποτελέσματα αυτών μπορούν είτε να αναπαρασταθούν σε γράφημα όπως θα παρουσιαστεί στην συνέχεια είτε ο χρήστης να επιλέξει τα αποτελέσματα να αποθηκευτούν σε ένα .txt, .xl αρχείο [**54**] , [**58**] .

Παρακάτω φαίνονται η αρχή και το τέλος και η πορεία ανάπτυξης των κυττάρων και τυχαία επιλέγεται να αναπαρασταθεί ένα πλήθος μόλις 7 γονιδίων και το τρόπος ανάπτυξης τους αποτυπωμένος σε ένα γράφημα με άξονα τον ψευδοχρόνο [**54**] , [**58**] .

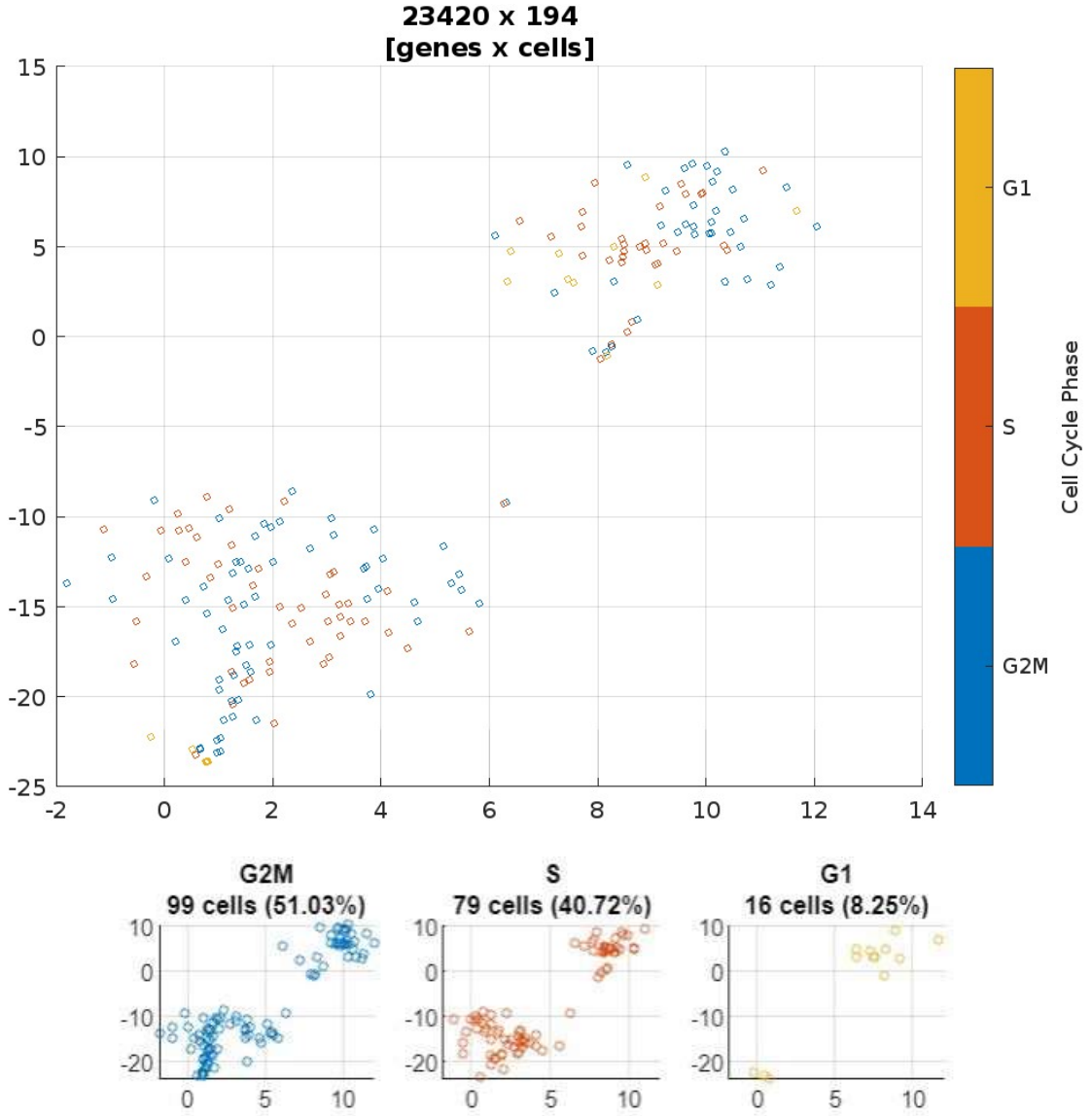


"©[(Results scGEAToolbox)] via scGEAToolbox"

Εικόνα 7: Πρώτη εικόνα: Trajectory analysis, Δεύτερη εικόνα: Έκφραση τυχαία επιλεγμένων γονιδίων σε άξονα ψευδοχρόνου

Το δεύτερο γράφημα αποτυπώνει τον τρόπο έκφρασης των γονιδίων, πότε η έκφραση τους αυξάνεται και πότε μειώνεται [**54**] , [**58**] .

Ακόμη, ο χρήστης έχει πολλές ακόμη επιλογές να επιλέξει για να διερευνήσει τα δεδομένα όπως:

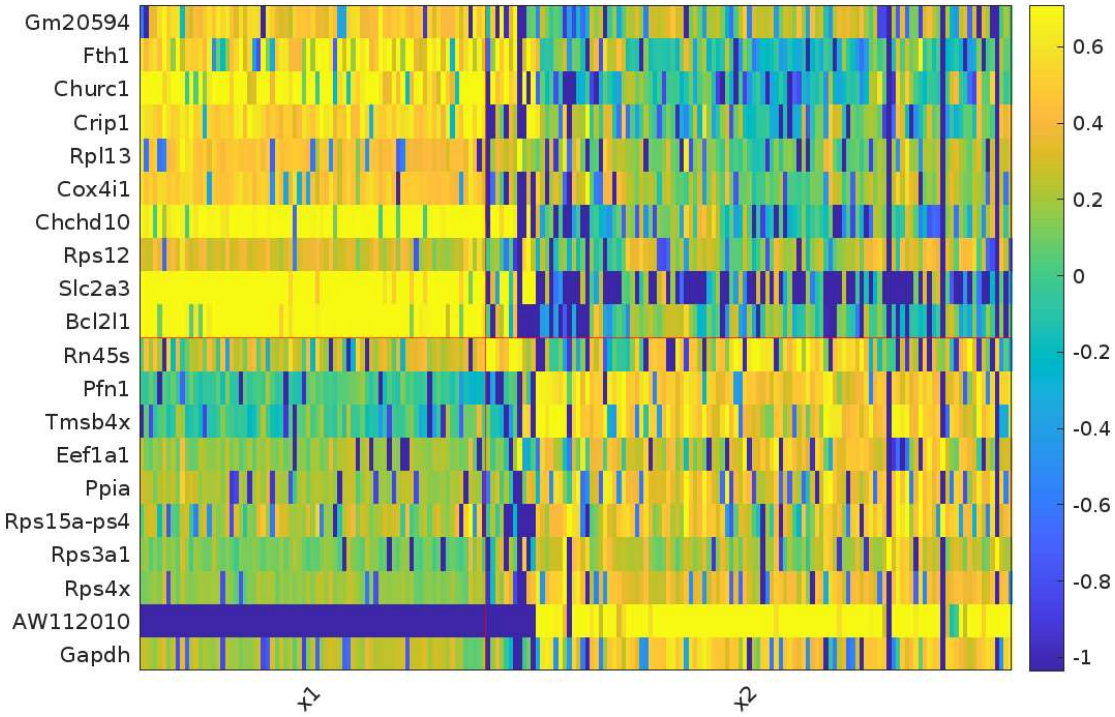


"©[(Results scGEAToolbox)] via scGEAToolbox"

Εικόνα 8: Πρώτη εικόνα: Κύτταρα και εύρεση του σταδίου τους κατά τον κυτταρικό κύκλο, Δεύτερη εικόνα: Πλήθος κυττάρων σε κάθε φάση του κυτταρικού κύκλου

Το παραπάνω γράφημα αποτυπώνει τις φάσεις ζωής των κυττάρων, καθώς επίσης μπορεί αναλυτικά να δει και το πόσα κύτταρα βρίσκονται σε κάθε φάση από τις G1, S, G2 [**54**] , [**58**] .

Επιπλέον ο χρήστης μπορεί να ανακαλύψει κι άλλες δυνατότητες όπως με την επιλογή Heatmap. Αυτό το είδος γραφημάτων είναι ο πιο εύκολος και παράλληλα ισχυρός στο να τον κατανοήσει κάποιος που τον μελετά [**54**] , [**58**] .



"©[(Results scGEAToolbox)] via scGEAToolbox"

Εικόνα 9: Heatmap, έκφραση τυχαίων γονιδίων στις 2 συστάδες

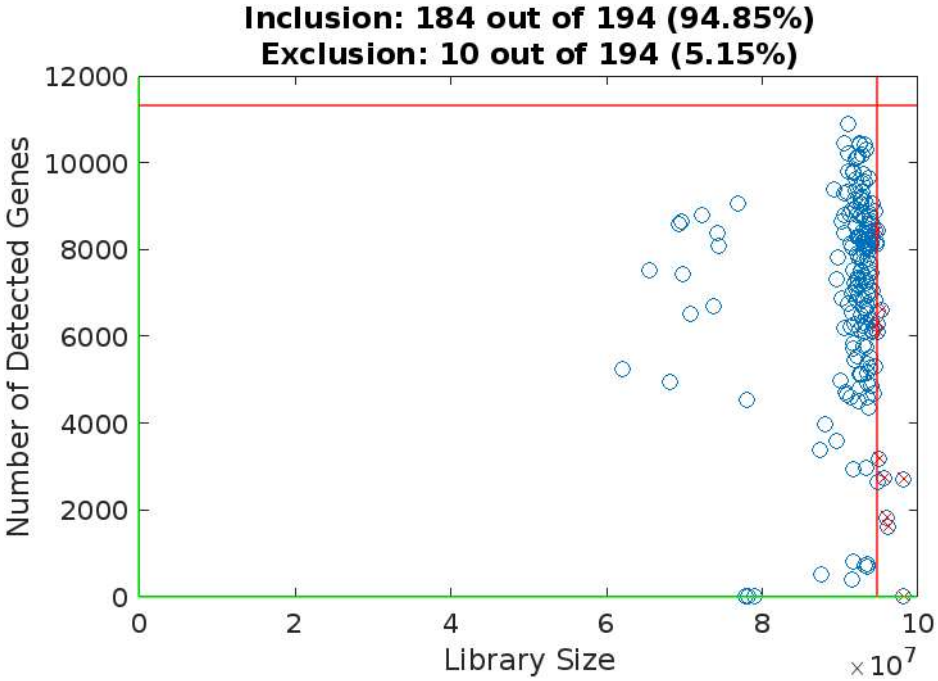
Από τον παραπάνω heatmap μαζί με το συνοδευτικό υπόμνημα ο χρήστης μπορεί να κατανοήσει σχετικά με την έκφραση των γονιδίων. Για παράδειγμα ενώ το γονίδιο με όνομα AW112010 εκφράζεται ελάχιστα στην πρώτη συστάδα, αυτή των Β-κυττάρων, στην δεύτερη συστάδα των Τ-κυττάρων μνήμης παρουσιάζει τόσο χαμηλή και μέτρια όσο και πολύ υψηλή έκφραση. Το υπόμνημα κατευθύνει τον χρήστη στο πως θα κατανοήσει την έκφραση των γονιδίων σύμφωνα με το χρώμα, με το μπλε να είναι πολύ χαμηλή και με το κίτρινο πολύ υψηλή η έκφραση [**54**] , [**58**] .

Σημαντικό στάδιο διερεύνησης είναι η εκτέλεση της Differential Expression analysis. Ανά 2 ομάδες όποιας επιλογής του χρήστη, ανάμεσα σε συστάδες ή κυτταρικούς τύπους ή μεταξύ φάσεων του κυτταρικού κύκλου μπορεί να συγκριθεί ο τρόπος που διαφοροποιείται η γονιδιακή έκφραση καθενός γονιδίου. Σε αυτή την περίπτωση ανάλυσης επιλέχθηκε να πραγματοποιηθεί DE ανάλυση με την μέθοδο DEseq2 για τις 2 συστάδες της μελέτης. Τα αποτελέσματα αποθηκεύτηκαν έπειτα από επιλογή σε ένα αρχείο .txt και ένα δείγμα 7 γονιδιίων θα παρουσιαστεί παρακάτω:

gene	p_val	avg_log2FC	abs_log2FC	pct_1	pct_2	p_val_adj
AW112010	0	-Inf	Inf	0	0.8803418803418	0
Cd3e	0	-Inf	Inf	0	0.871794871794872	0
Cd3g	0	-Inf	Inf	0	0.863247863247863	0
Ms4a6b	0	-Inf	Inf	0	0.863247863247863	0
Fyb	0	-Inf	Inf		0.84615384615384	0
B4galnt1	0	-Inf	Inf	0	0.846153846153846	0
Ms4a4b	0	-Inf	Inf	0	0.846153846153846	0

**τυχαίο δείγμα δεδομένων από το .txt αρχείο σχετικά με την DE analysis*

Μια άλλη δυνατότητα διερεύνησης προσφέρει το παρακάτω γράφημα:

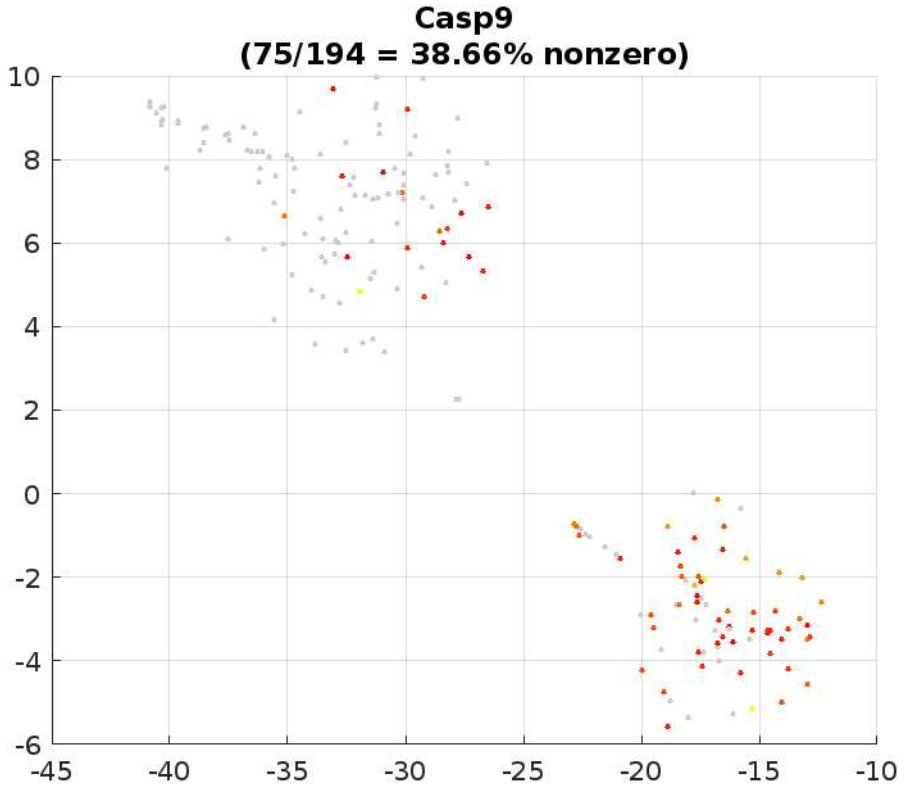


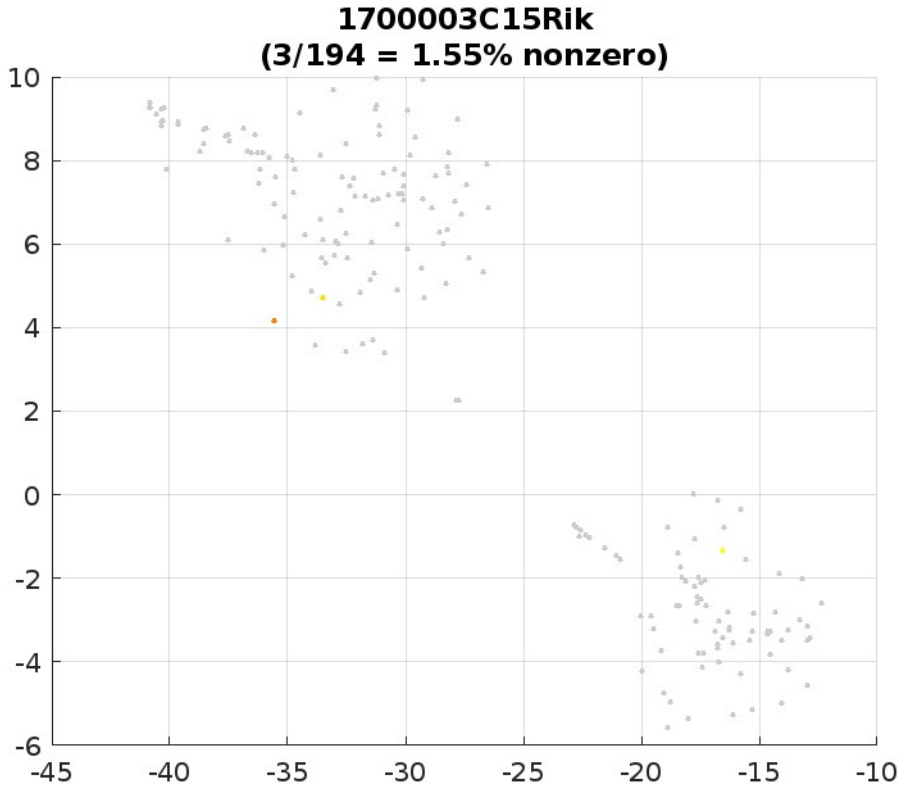
"©[(Results scGEAToolbox)] via scGEAToolbox"

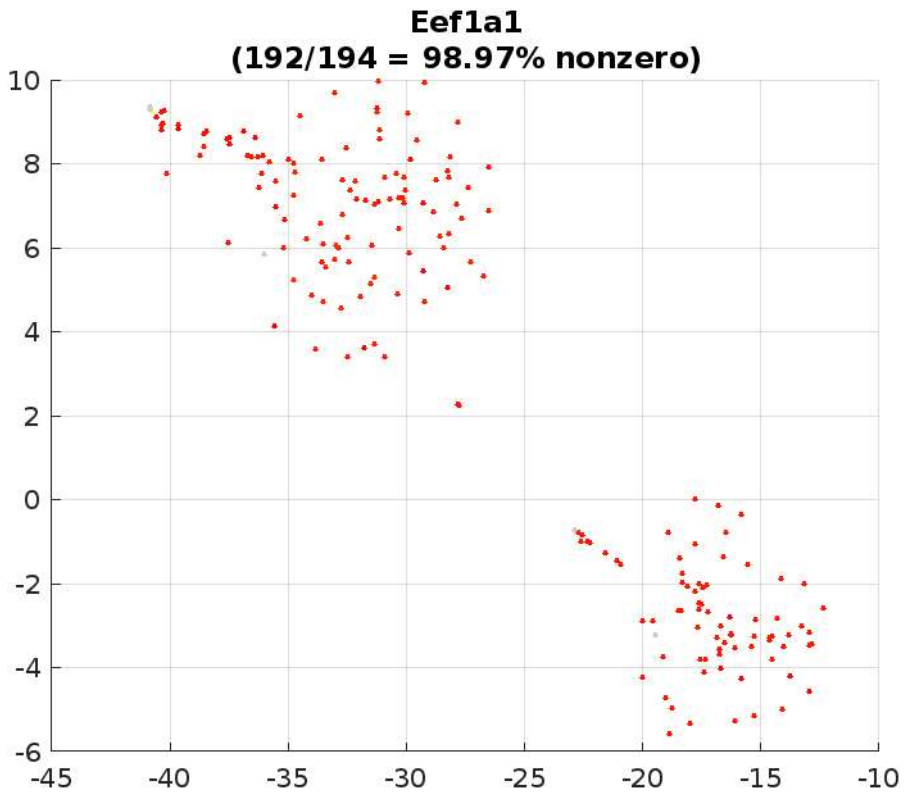
Εικόνα 10: Αριθμός γονιδίων που ανιχνεύτηκαν σχετικά με το μέγεθος βιβλιοθήκης του συγκεκριμένο πακέτου δεδομένων

Το γράφημα αυτό πληροφορεί για το πόσα κύτταρα περιλαμβάνουν τα γονίδια της cDNA βιβλιοθήκης της εν προκειμένω μελέτης. Μόλις 184/194 κύτταρα πληρούν τις προϋποθέσεις να συμπεριλαμβάνονται στην μελέτη. Τα υπόλοιπα 10 κύτταρα αποκλείονται καθώς εκφράζουν γονίδια που δεν χρειάζονται για την συγκεκριμένη μελέτη και αφαιρέθηκαν [**54**] , [**58**] .

Όσο αφορά τα γονίδια ο χρήστης έχει την δυνατότητα να επιλέξει να μάθει για την έκφραση των γονιδίων μεταξύ των τριών επιλογών: λίστα με αλφαβητική σειρά, μέση έκφραση και καθόλου έκφραση. Μερικά παραδείγματα βρέθηκαν τυχαία και αποτυπώνονται παρακάτω. Η επιλογή αφορά γονίδια με μέση έκφραση [**54**] , [**58**] .







"©[(Results scGEAToolbox)] via scGEAToolbox"

Εικόνα 11: Έκφραση τυχαία επιλεγμένων γονίδιων με κριτήριο την μέση τιμή έκφρασης τους

Επιπλέον πληροφορίες μπορεί να λάβει ο χρήστης για την έκφραση των γονιδίων. Διαλέγοντας την επιλογή Calculate Gene Expression Statistics ο χρήστης μπορεί να λάβει σε μορφή αρχείου .txt λεπτομέρειες για την μέση τιμή, το dropout κ.α. [**54**] , [**58**] .

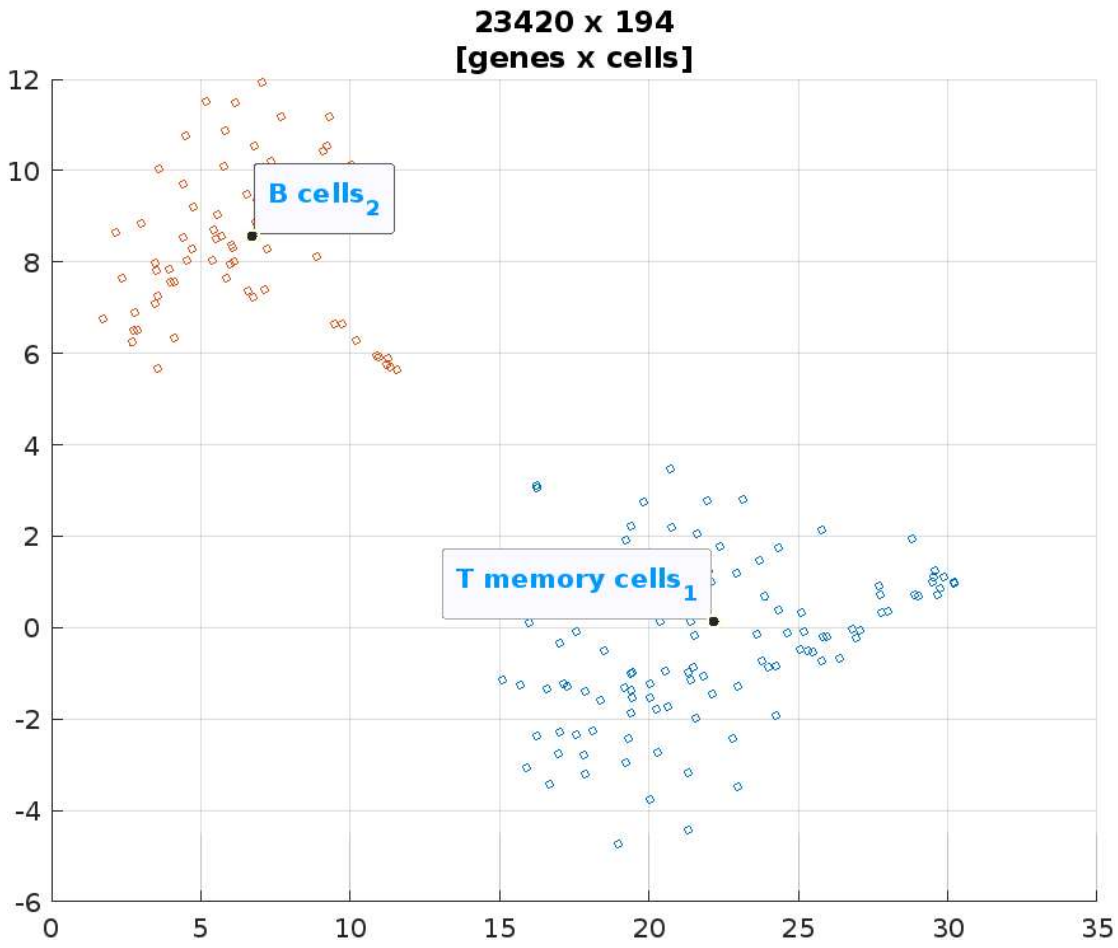
Η μορφή του αρχείου παρουσιάζεται μέσα από τα πρώτα 5 γονίδια:

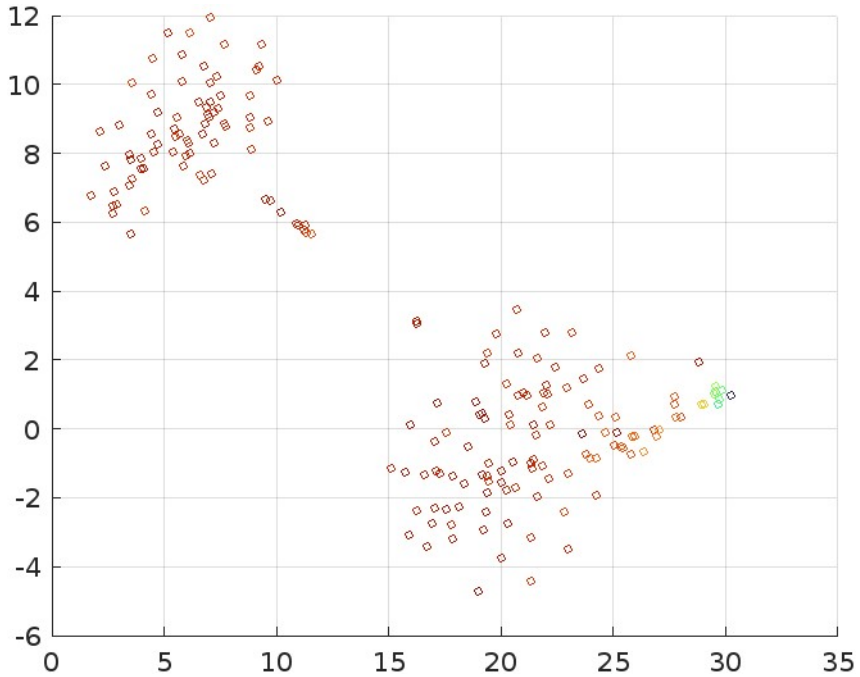
Gene	**Mean**	**CV**	**Dropout_rate**
Itm2a	**938.948453608247**	**3.89978002040782**	**0.783505154639175**
Sergef	**2368.61340206186**	**3.27046770651069**	**0.391752577319588**
Fam109a	**962.458762886598**	**3.97848327804267**	**0.664948453608248**
Dhx9	**2154.54639175258**	**0.887331566661594**	**0.11340206185567**
Ssu72	**15943.8556701031**	**0.846254324987634**	**0.0618556701030928**

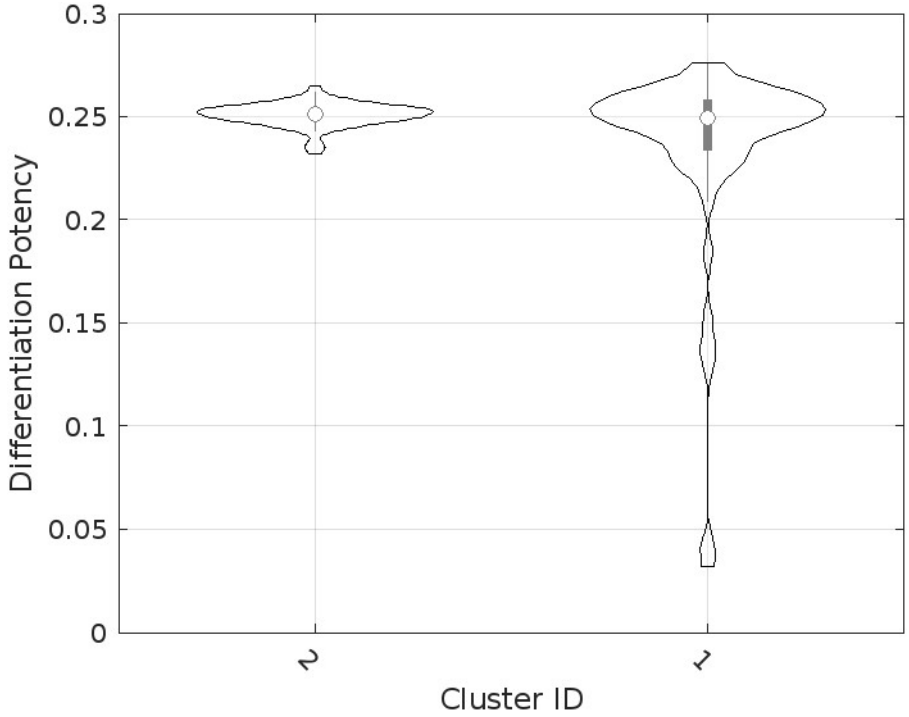
**τυχαίο δείγμα δεδομένων από το .txt αρχείο σχετικά με στατιστικά δεδομένα που αφορούν την έκφραση γονιδίων και τις παραμέτρους mean, cv, dropout rate*

Η τελευταία στήλη δείχνει πόσο συχνά ένα γονίδιο ενώ υπάρχει στο κύτταρο/α τελικά η έκφραση του δεν ανιχνεύεται [**54**] , [**58**] .

Ακόμη, το πόσο συχνά κάθε κυτταρικός τύπος εκφράζει τα γονιδιακά του προϊόντα (πρωτεΐνες) αλλάζει όχι μόνο από κύτταρο σε κύτταρο αλλά και από οργανισμό σε οργανισμό. Η πλατφόρμα δίνει την δυνατότητα να γνωρίζει ο χρήστης λεπτομέρειες για αυτό [**54**] , [**58**] .


23420 x 194
[genes x cells]
B cells2
T memory cells1
12
10
8
6
4
2
0
-2
-4
-6
0
5
10
15
20
25
30
35

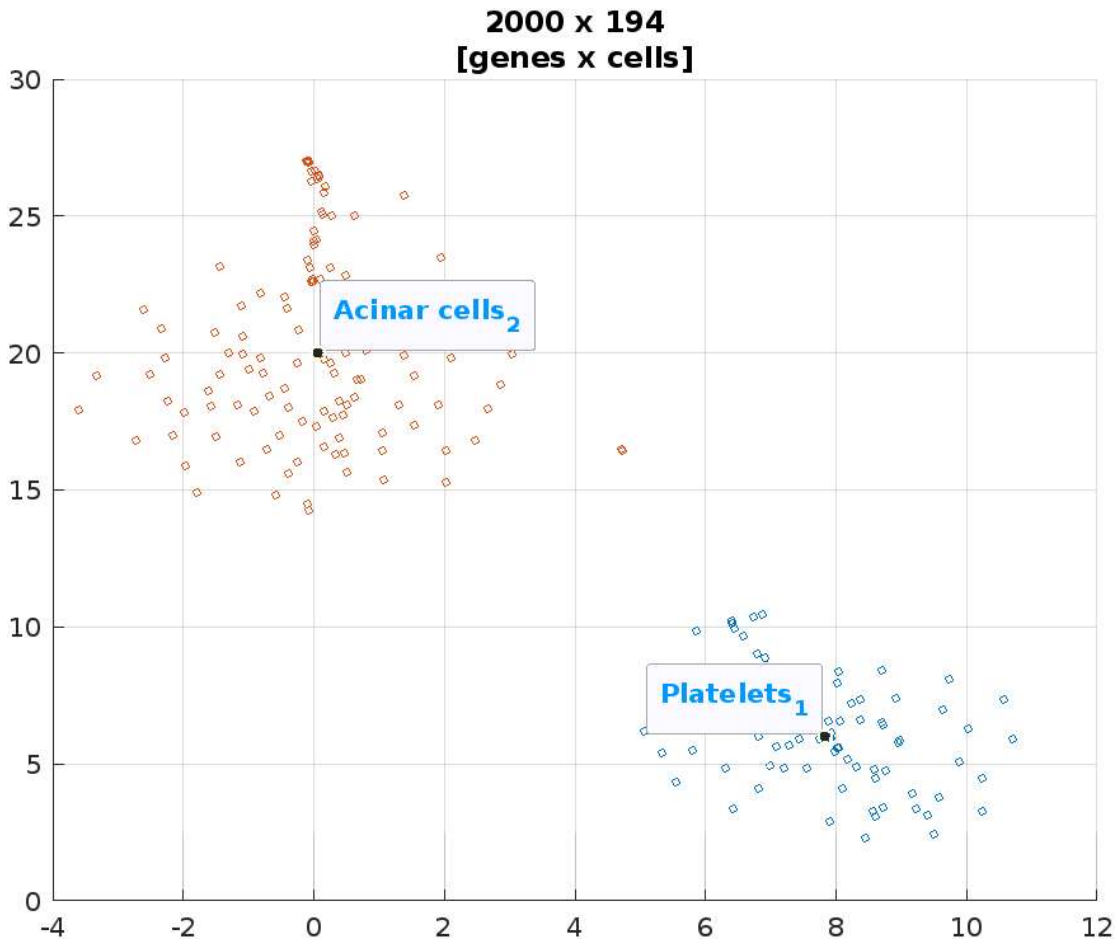

12
10
8
6
4
2
0
-2
-4
-6
0
5
10
15
20
25
30
35



“©[(Results scGEAToolbox)] via scGEAToolbox”

Εικόνα 12: Εμφάνιση της διαφοροποίησης της κυτταρικής έκφρασης μεταξύ διαφορετικών κυττάρων για την δεύτερη εικόνα και μεταξύ των συστάδων για την τρίτη εικόνα

Τέλος, ιδιαίτερα σημαντική πληροφορία αποτελούν τα HVGs τα οποία περιλαμβάνουν συνήθως 2000 γονίδια που εμφανίζουν πολλές διαφορές κατά τον τρόπο έκφρασής τους. Μια σχετική διερεύνηση στα παραπάνω δεδομένα έδωσε τα παρακάτω αποτελέσματα που έπειτα από κυτταρική ταυτοποίηση εμφάνισαν μια πληροφορία που χρήζει άλλης διερεύνησης [**54**] , [**58**] .



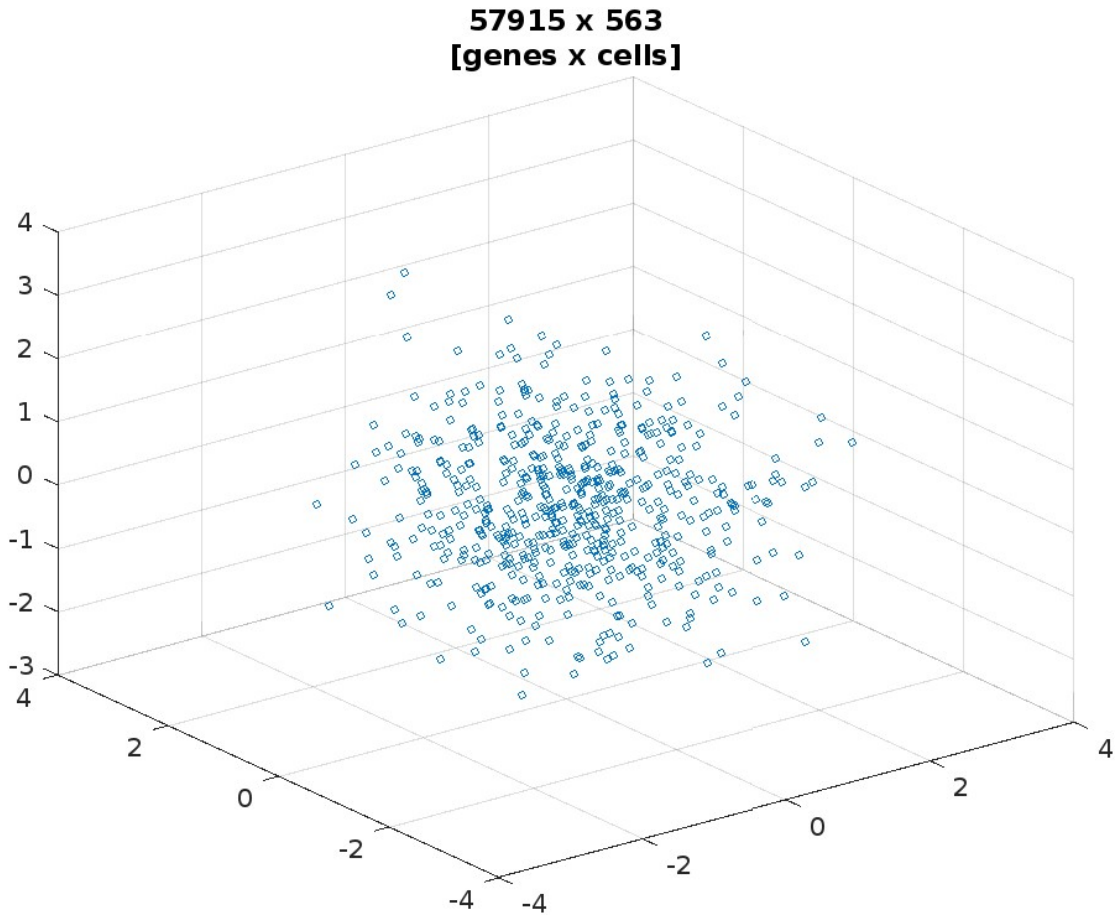
"©[(Results scGEAToolbox)] via scGEAToolbox"

Εικόνα 13: Συσταδοποίηση για τα 2000 HVGs και κυτταρική ταυτοποίηση

GSE75688:

Μια δεύτερη ανάλυση που πραγματοποιήθηκε αφορά το αρχείο GSE75688.csv [**55**] , [**58**] .

Μετά την εισαγωγή των δεδομένων λαμβάνουμε την παρακάτω εικόνα με τα δεδομένα:

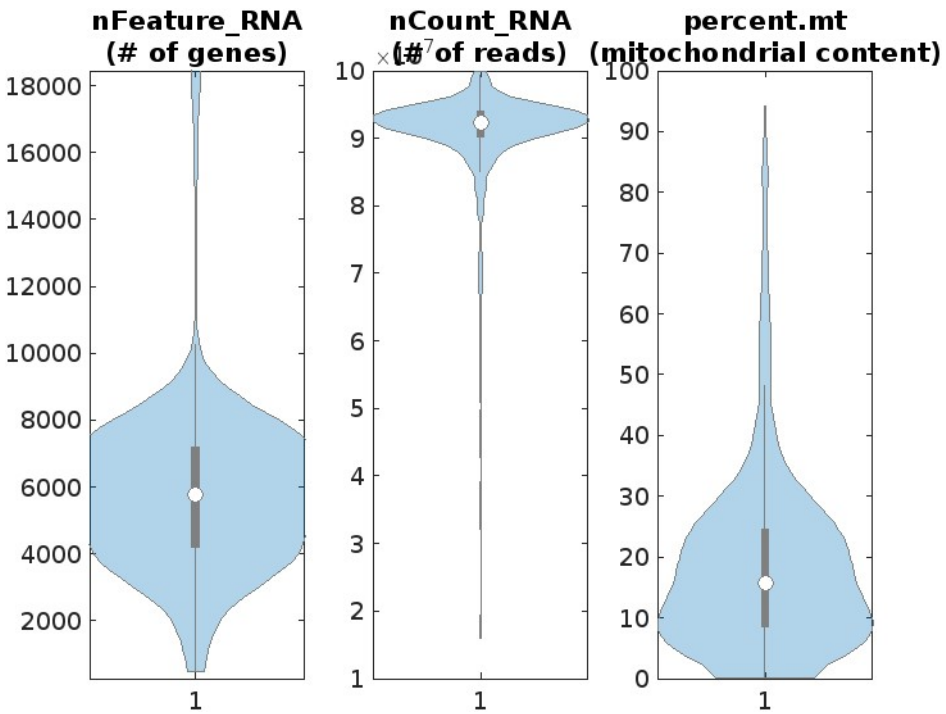


"©[(Results scGEAToolbox)] via scGEAToolbox"

Εικόνα 14: GSE75688 δεδομένα

Από την εικόνα πληροφορούμαστε πως το πακέτο δεδομένων GSE75688 αποτελείται από 57915 γονίδια και 563 κύτταρα [**55**] , [**58**] .

Στη συνέχεια και θέλοντας να ενημερωθούμε για τις 3 παραμέτρους του ποιοτικού ελέγχου όπως αναφέρονται στο σχετικό κεφάλαιο, λαμβάνουμε το εξής γράφημα που έχει την μορφή violin plot [**55**] , [**58**] .

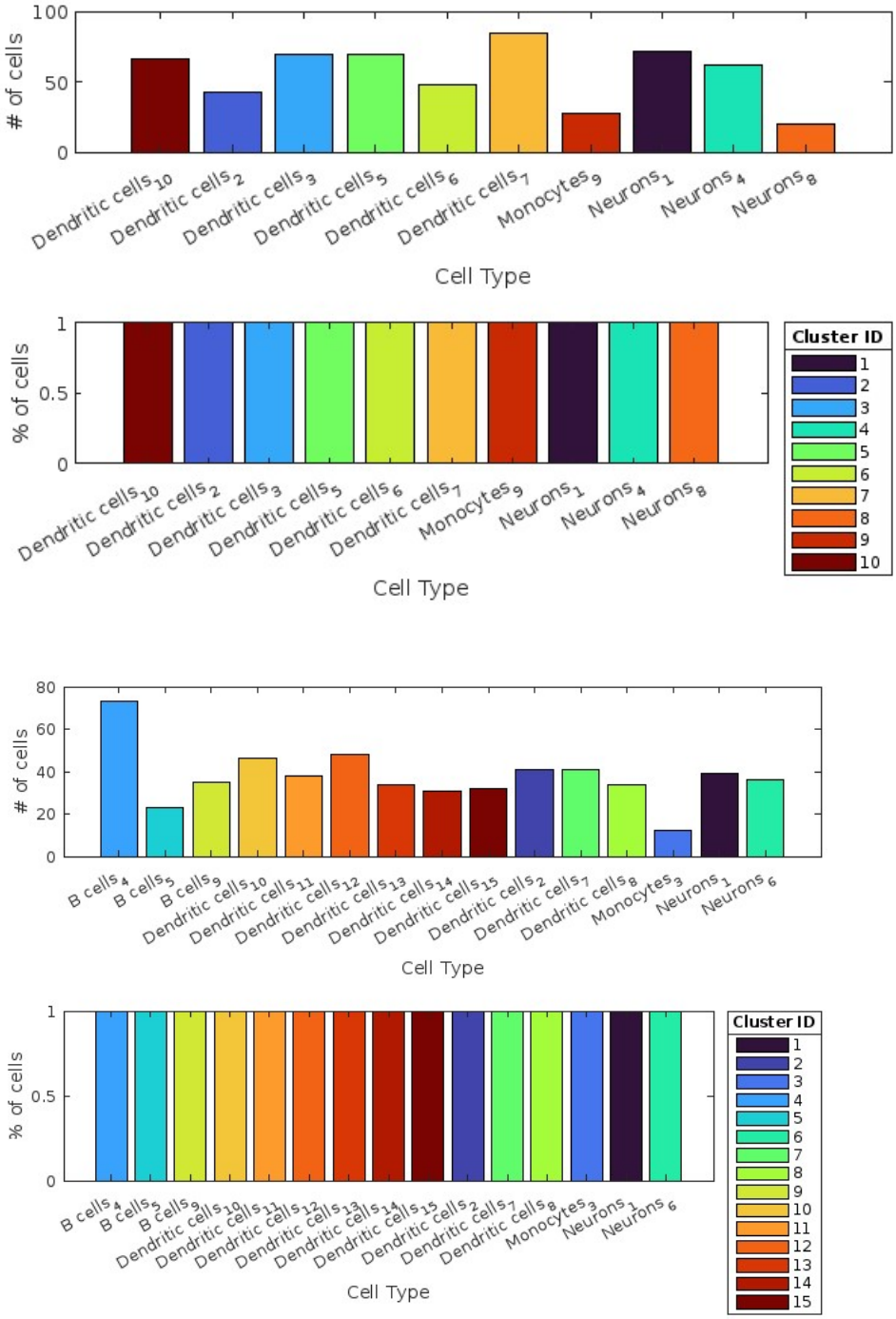


"©[(Results scGEAToolbox)] via scGEAToolbox"

Εικόνα 15: Τιμές παραμέτρων ποιοτικού ελέγχου

Επειδή τα δεδομένα αυτού του πακέτου είναι πολύ μεγάλα, ο αλγόριθμος t-SNE δεν ήταν δυνατόν να πραγματοποιηθεί. Για να συνεχίσουμε να λαμβάνουμε χρήσιμες πληροφορίες κατά την διερεύνηση του πακέτου, εκτελέστηκε συσταδοποίηση αρχικά για πλήθος 10 συστάδων και έπειτα για πλήθος 15 συστάδων [**55**] , [**58**] .

Τα υπομνήματα που λήφθηκαν παρουσιάζονται παρακάτω:

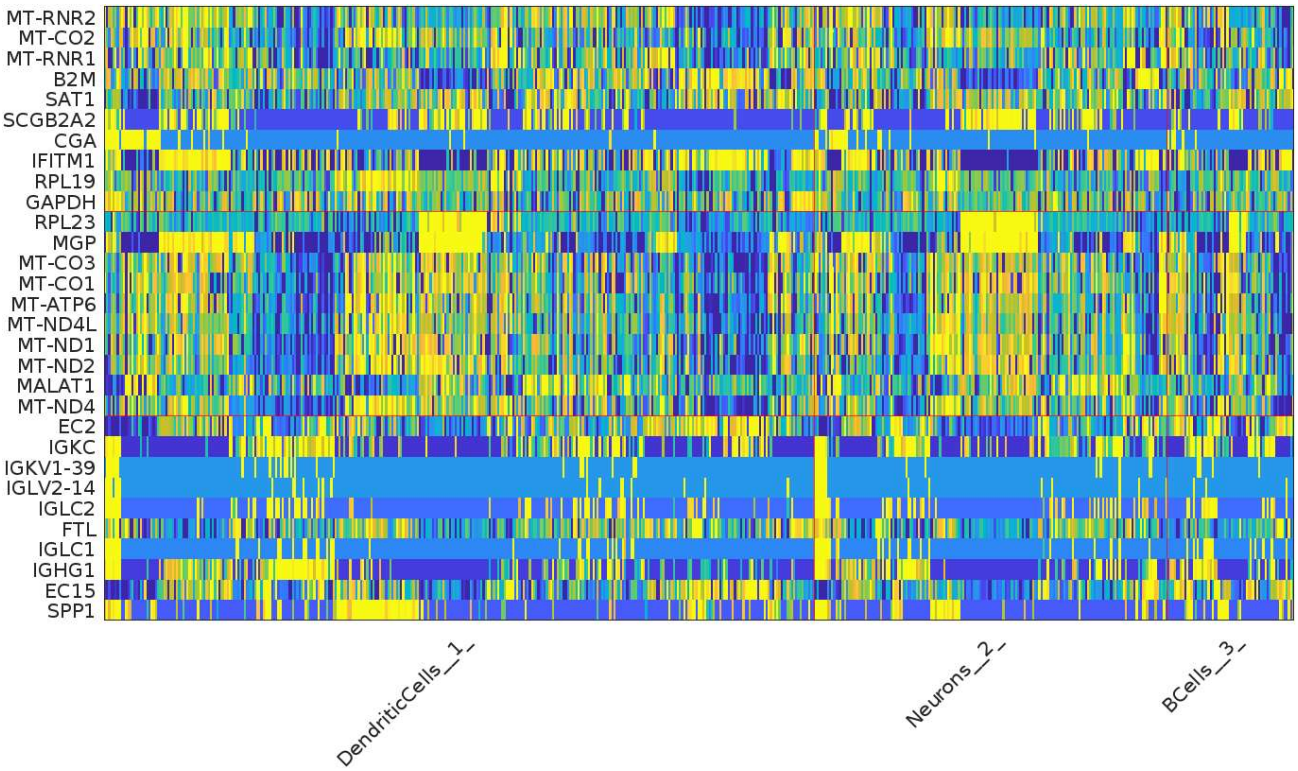


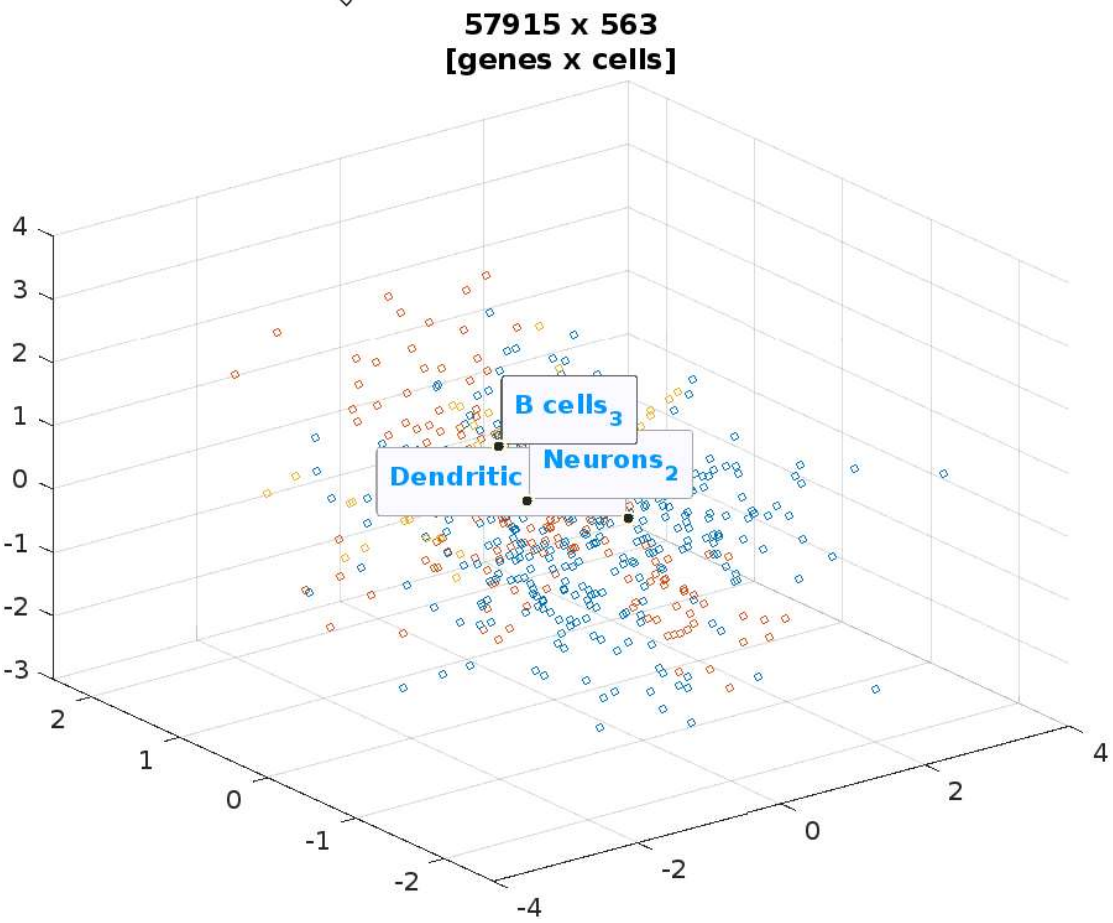
"©[(Results scGEAToolbox)] via scGEAToolbox"

Εικόνα 16: Πρώτη εικόνα: υπόμνημα για τις 10 συστάδες, Δεύτερη εικόνα: υπόμνημα για τις 15 συστάδες

Είναι φανερό πως εάν δεν είχε διερευνηθεί περαιτέρω το πλήθος των συστάδων, οι 3 ομάδες των Β-κυττάρων δεν θα είχαν ανακαλυφθεί [**55**] , [**58**] .

Στη συνέχεια, αναζητήθηκε η γονιδιακή έκφραση των marker genes μέσω του heatmap που φαίνεται παρακάτω:



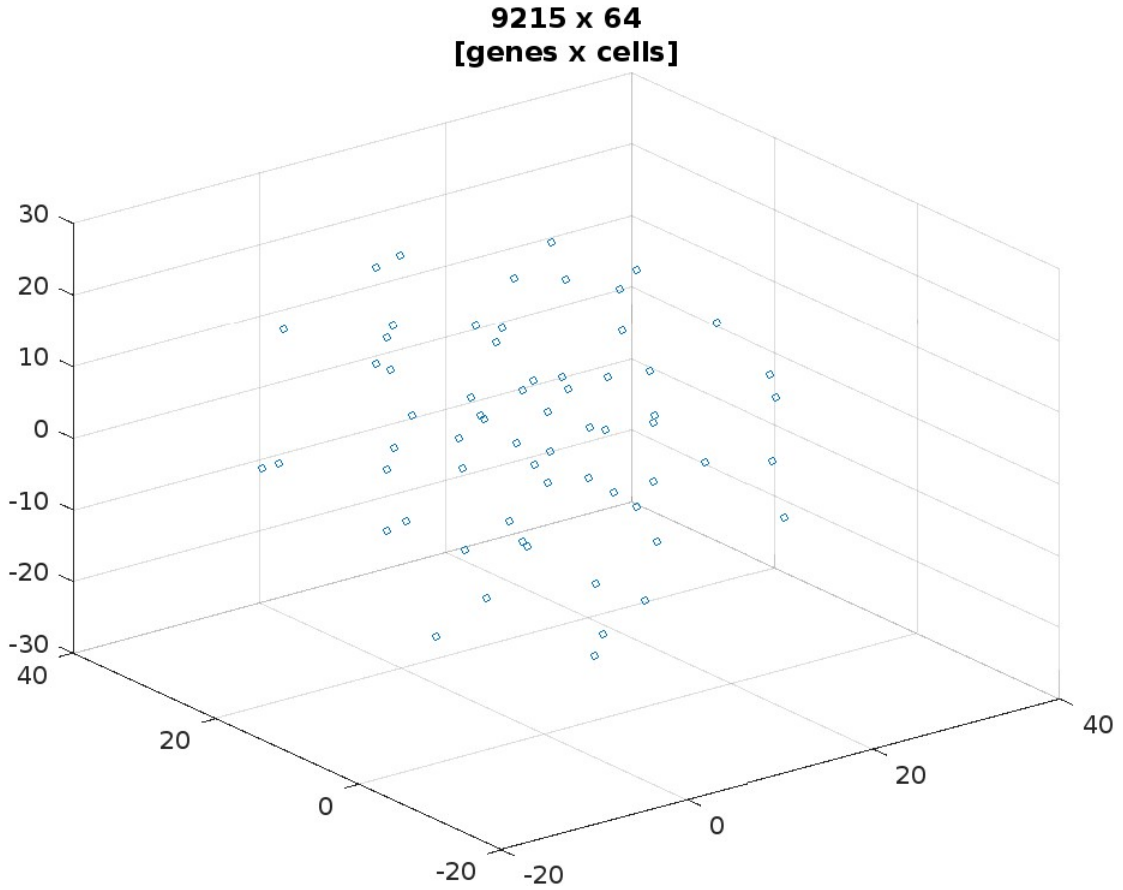


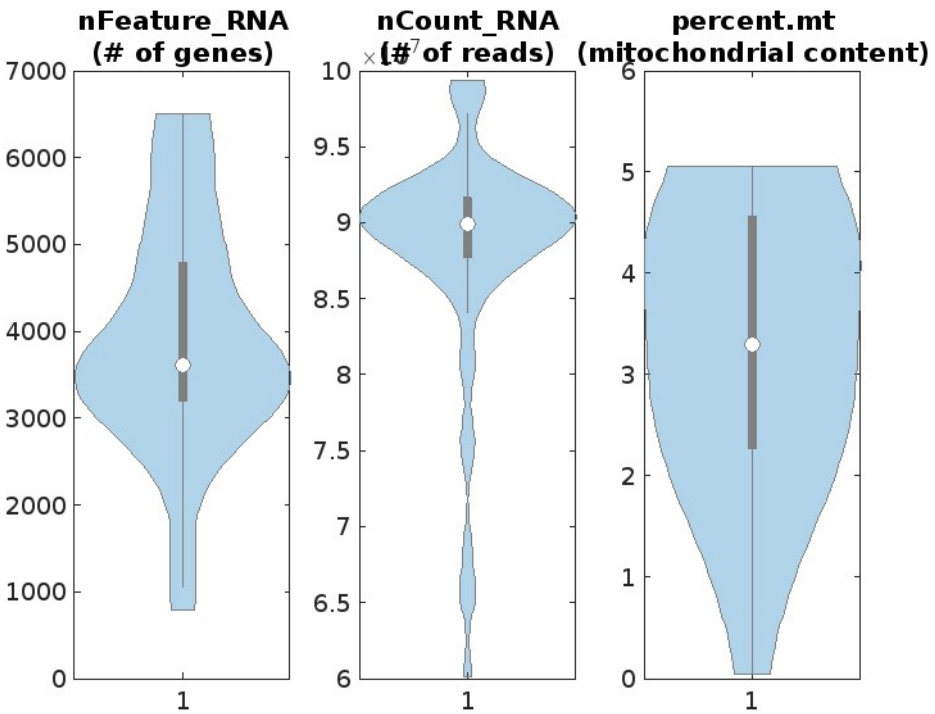
“©[(Results scGEAToolbox)] via scGEAToolbox”

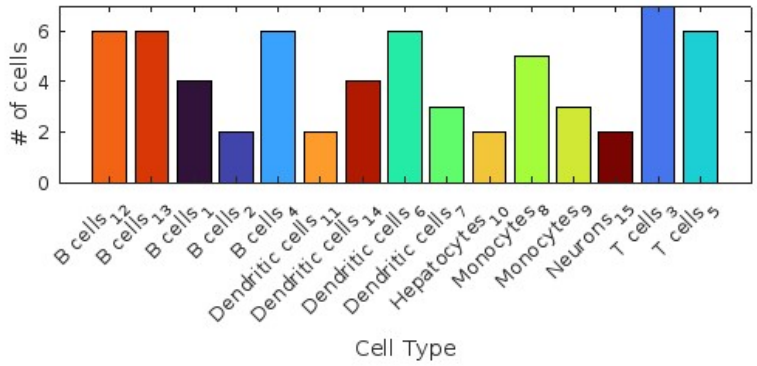
Εικόνα 17: Πρώτη εικόνα:Heatmap, Δεύτερη εικόνα: Merge 15 συστάδων

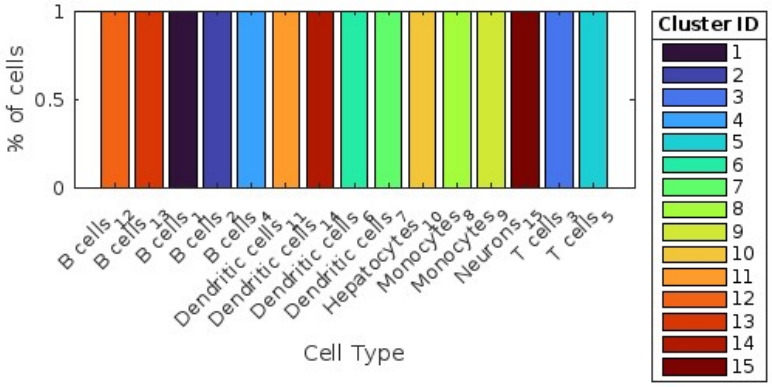
Αναλυτικότερα, επειδή η οπτικοποίηση τυχαίου πλήθους γονιδίων στον παραπάνω Heatmap δεδομένων των 15 συστάδων ήταν αδύνατη, τα μεγάλου όγκου δεδομένα ήταν δυσνόητα, η οπτικοποίηση έγινε με άλλη μέθοδο. Οι 15 συστάδες συμπιέστηκαν, δηλαδή κάθε συστάδα που εμφάνιζε χαρακτηριστικές διαφορές που την διαφοροποιούσαν από τον δίπλα κυτταρικό ιστό, όμως συνολικά οι συστάδες αυτές ανήκουν στον ίδιο κυτταρικό τύπο (υποσυστάδες) συμπιέστηκε στην πλησιέστερη. Αυτή η συμπίεση έδωσε ένα νέο πλήθος 3 συστάδων. Η δημιουργία Heatmap πλέον ήταν δυνατή και το αποτέλεσμα απεικονίζεται παραπάνω [**55**] , [**58**] .

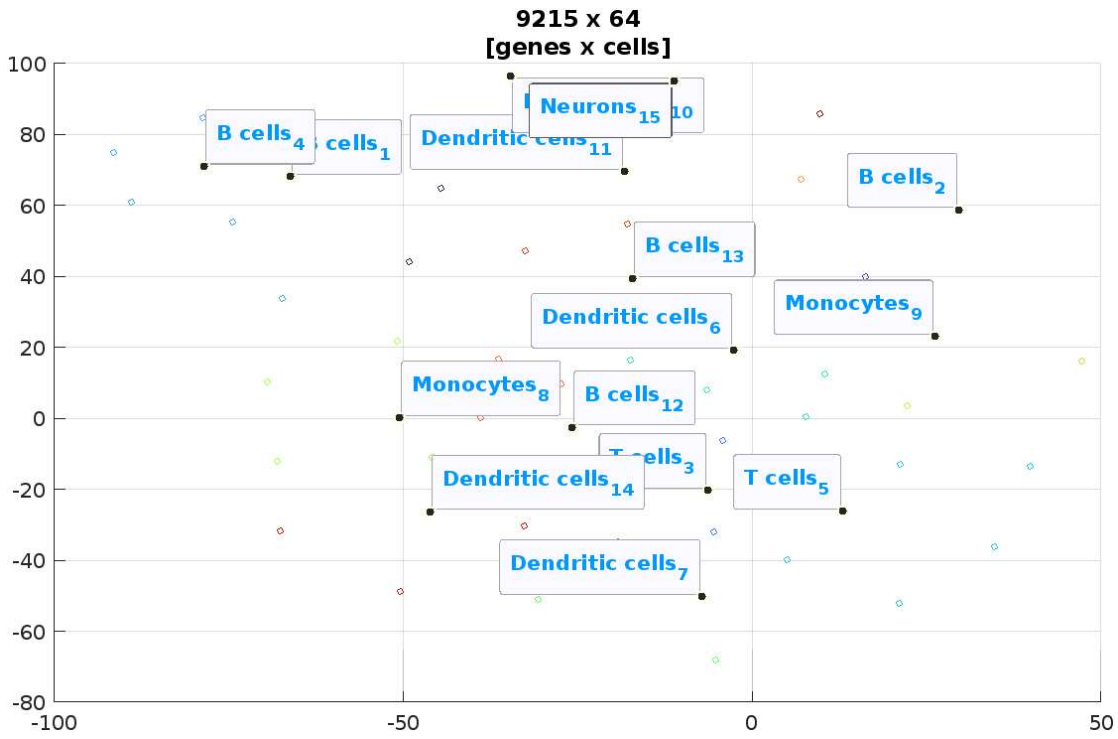
Επειδή το πρώτο γράφημα των παραμέτρων του ποιοτικού ελέγχου έδειξε πως υπάρχουν μιτοχονδριακά γονίδια, εφαρμόστηκε ένα νέο κατώφλι, της τάξεως του 0,05 και τα αποτελέσματα φαίνονται παρακάτω:




nFeature_RNA
(# of genes)
nCount_RNA
(# of reads)
percent.mt
(mitochondrial content)


# of cells
Cell Type


% of cells
Cell Type
Cluster ID

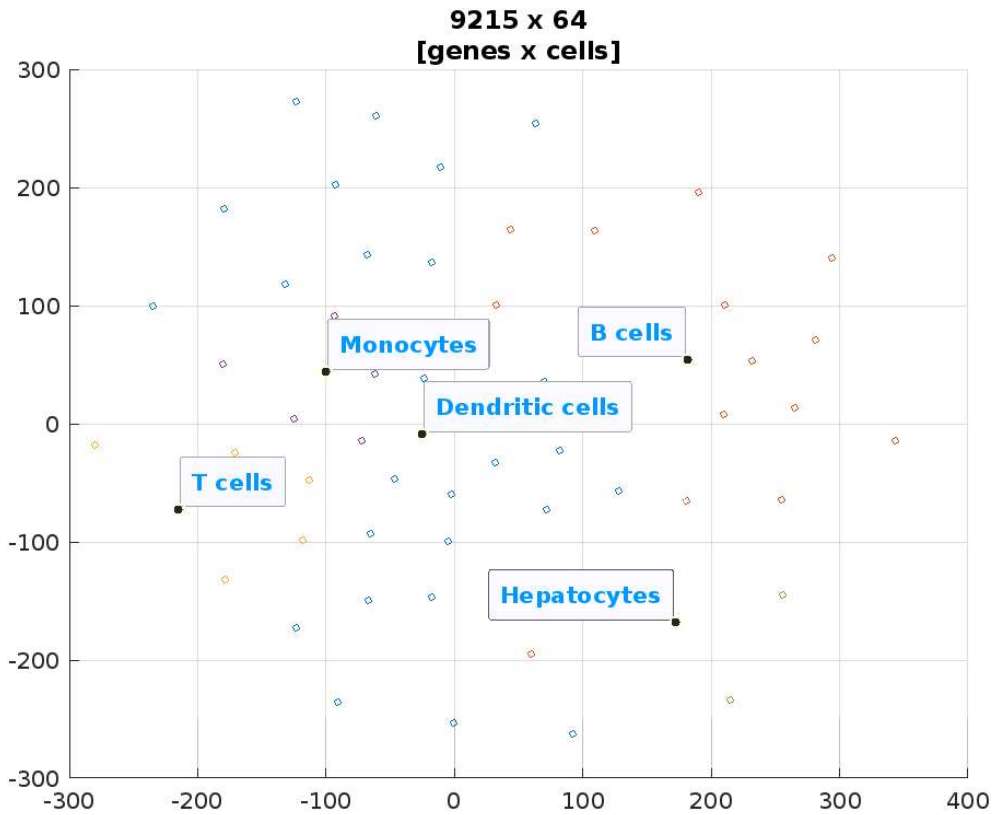


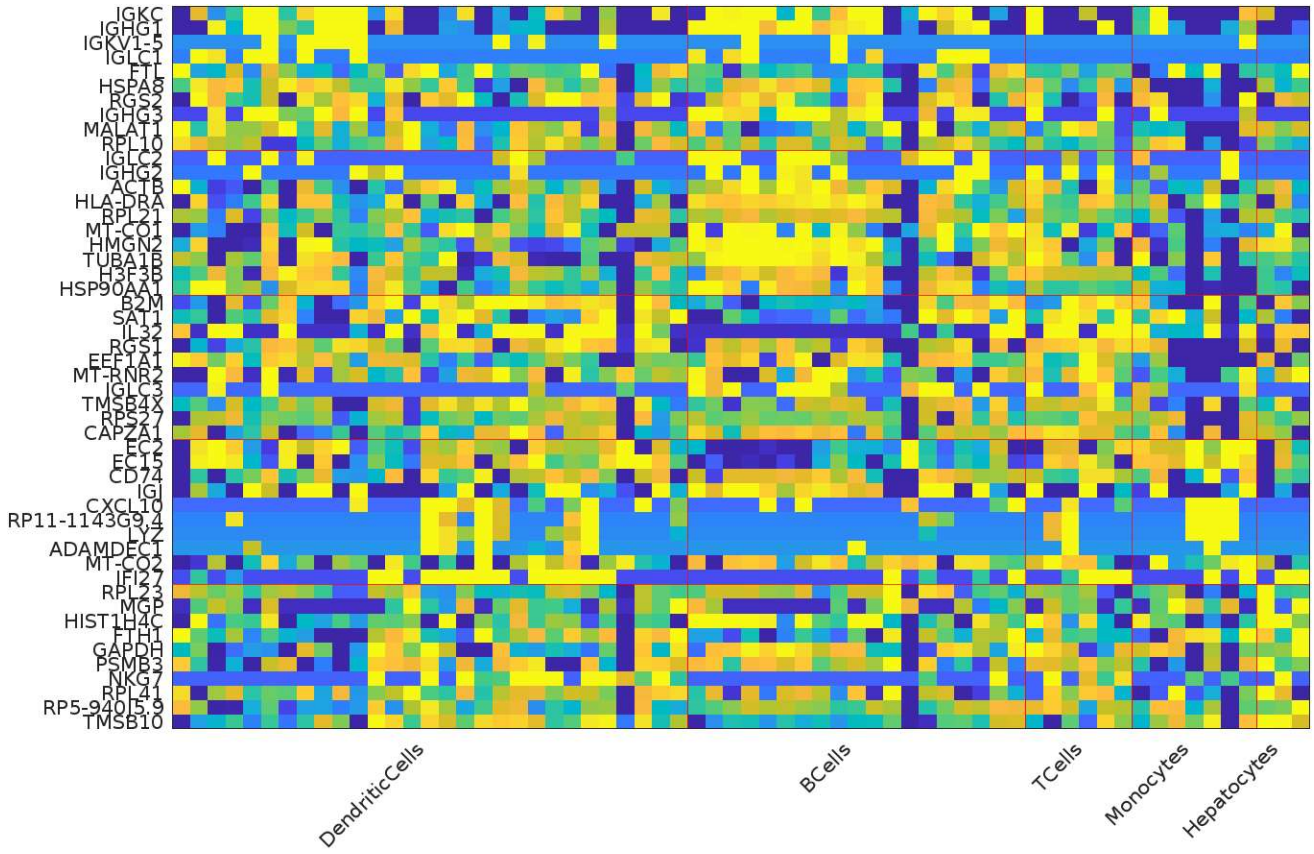
“©[(Results scGEAToolbox)] via scGEAToolbox”

Εικόνα 18: Πρώτη εικόνα: Γονίδια και κύτταρα που έχουν απομείνει μετά την εφαρμογή της τιμής 0,05 percent mito, Δεύτερη εικόνα: Τιμές παραμέτρων ποιοτικού ελέγχου, Τρίτη εικόνα: Υπόμνημα, Τέταρτη εικόνα: Συσταδοποίηση και ταυτοποίηση κυτταρικών τύπων

Τα κύτταρα μειώθηκαν κατά πολύ. Από περίπου 500 μειώθηκαν στα 60. Σε αυτή την περίπτωση λοιπόν ανακαλύφθηκε μια ακόμα κυτταρική ομάδα τα Hepatocytes, τα οποία όπως φαίνεται στον αριθμό είναι λίγα και άρα αξίζει να διερευνηθούν[**55**] , [**58**] .

Τέλος, θα αναζητηθεί ο Heatmap χάρτης της έκφρασης τυχαία επιλεγμένων γονιδίων. Όπως και παραπάνω, επειδή είναι αδύνατο κάτι τέτοιο να πραγματοποιηθεί για 10 συστάδες, αυτές και πάλι συμπιέστηκαν αυτό δίνοντας τον παρακάτω Heatmap:


9215 x 64
[genes x cells]
Monocytes
B cells
Dendritic cells
T cells
Hepatocytes
300
200
100
0
-100
-200
-300
-300
-200
-100
0
100
200
300
400


IGKC
IGHG1
IGKV1-5
IGLC1
FTL
HSPA8
RGS2
IGHG3
MALAT1
RPL10
IGLC2
IGHG2
ACTB
HLA-DRA
RPL21
MT-CO1
HMGN2
TUBA1B
H3F3B
HSP90AA1
B2M
SAT1
IL32
RGS1
EEF1A1
MT-RNR2
IGLC3
TMSB4X
RPS27
CAPZA1
EC2
EC15
CD74
IGJ
CXCL10
RP11-1143G9.4
LYZ
ADAMDEC1
MT-CO2
IFI27
RPL23
MGP
HIST1H4C
FTH1
GAPDH
PSMB3
NKG7
RPL41
RP5-940J5.9
TMSB10
DendriticCells
BCells
TCells
Monocytes
Hepatocytes

“©[(Results scGEAToolbox)] via scGEAToolbox”

Εικόνα 19: Πρώτη εικόνα: Συμπίεση των υποσυστάδων, Δεύτερη εικόνα: Τιμές έκφρασης γονιδίων με την μορφή Heatmap

Μάλιστα, κάποια γονίδια της ομάδας κυττάρων Hepatocytes αξίζει να διερευνηθούν καθώς εμφανίζουν υψηλή γονιδιακή έκφραση που συμβολίζεται με το κίτρινο χρώμα όπως έχει αναλυθεί στο προηγούμενο σετ δεδομένων [**55**] , [**58**] .

8 Συμπεράσματα

Συμπερασματικά λοιπόν η πλατφόρμα scGEATool είναι ένα ισχυρό εργαλείο διερευνητικής ανάλυσης. Εκτός του ότι μπορεί να τρέξει online και προσφέρει πάρα πολλές επιλογές αρχείων εισόδου μπορεί να απελευθερώσει και πολλούς υπολογιστικούς πόρους και αποθηκευτικό χώρο, μιας και διαθέτει διαθέσιμη μνήμη 20GB. Εκτελεί κάθε βήμα σε ένα διάστημα 3-20 δευτερόλεπτα. Ακόμη είναι μια πλατφόρμα παράλληλης ανάλυσης χωρίς να υποχρεώνει τον χρήστη να παραμένει στο ίδιο βήμα. Επιπλέον, η διερεύνηση μπορεί να αποκτήσει πολλές μορφές καθώς ο χρήστης έχει μια πληθώρα υπολογιστικών πακέτων στην ίδια εφαρμογή για να διερευνήσει από το ποσοσστό των μιτοχονδριακών γονιδίων μέχρι και την χαμηλή έκφραση κάποιου γονιδίου.

Από την άλλη, η πλατφόρμα δεν περιλαμβάνει κάποιο μέρος αποθήκευσης όλων των βημάτων με αποτέλεσμα ο χρήστης να μην μπορεί να μοιραστεί συνολικά την ροή της ανάλυσης απευθείας από την πλατφόρμα.

Ακόμη, είναι ιδιαίτερα σημαντικό να γίνεται κατανοητός και να κοινοποιείται ο σκοπός δημιουργίας του υπολογιστικού εργαλείου. Η διαδικτυακή κοινοποίηση των υπολογιστικών εργαλείων και των χαρακτηριστικών τους προσελκύει μια ομάδα χρηστών. Εάν τα πλαίσια δημιουργίας τους αφορούν κάποια εργασία και τίποτα άλλο αυτό θα πρέπει να γίνεται γνωστό. Ο χρήστης αναζητά προς χρήση εργαλεία που είναι διαθέσιμα για τον σκοπό αυτό. Αυτό σημαίνει πως θα ήταν καλό να υπάρχει η δυνατότητα να ανοίξει ένας δίαυλος επικοινωνίας ευνοϊκός και για τις 2 πλευρές. Ο χρήστης μέσω των ερωτήσεων αναζητά απαντήσεις για να μπορέσει να λύσει όλα τα προβλήματα κατά την διάρκεια ανάλυσης του. Ο δημιουργός/οι λαμβάνουν πληροφορίες σχετικά με το πως μπορούν να τροποποιήσουν και να βελτιώσουν την υπολογιστική εμπειρία των ερευνητών. Η έλλειψη αυτής της στοιχειώδους επικοινωνίας μεταφράζεται σε χαμένο χρόνο και άγχος από την πλευρά του ερευνητή που προσπαθεί να χρησιμοποιήσει ένα εργαλείο που εν τέλη φτιάχτηκε απλώς και μόνο για ακαδημαϊκούς σκοπούς.

Τέλος, ως μελλοντική διερεύνηση αποτελεί το ζήτημα των αρχείων δεδομένων. Δεδομένα διαφορετικών μορφών και αρχείων τύπων ζητούνται από κάθε πλατφόρμα. Εκτός αυτού, ακόμη και η ίδια μορφή αρχείων χρειάζεται τροποποίηση με αποτέλεσμα κάποια λάθη που προκύπτουν να αφορούν αυτό και μόνο, μια λανθασμένη μορφή των

δεδομένων εισόδου, που λόγω του τεράστιου όγκου δεν είναι εφικτή η εύρεση του σφάλματος. Συνεπώς η δημιουργία μιας πλατφόρμας/βάσης δεδομένων που θα περιλαμβάνει όλες τις μορφές των αρχείων θα διευκόλυνε την λήψη και διερεύνηση τους εύκολα και άμεσα χωρίς να προκύπτουν ζητήματα ήδη από τα αρχικά στάδια της διερευνητικής ανάλυσης.

Βιβλιογραφία

[1]. Λιάσσα, Μ. Δ. (2021). *Μέθοδοι κατηγοριοποίησης σε δεδομένα μεγάλου όγκου από τεχνικές single-cell RNA-sequencing* (Bachelor's thesis).

[2]. Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., & Waterston, R. H. (2017). DNA sequencing at 40: past, present and future. *Nature*, *550*(7676), 345-353.

[3]. Chaudhry, F., Isherwood, J., Bawa, T., Patel, D., Gurdziel, K., Lanfear, D. E., ... & Levy, P. D. (2019). Single-cell RNA sequencing of the cardiovascular system: new looks for old diseases. *Frontiers in Cardiovascular Medicine*, *6*, 173.

[4]. Stiftung für Innovative Medizin. (2016). *Single Cell RNA Sequencing - Finding a cure for DIPG* [video file]. Retrieved from: Single Cell RNA Sequencing - Finding a cure for DIPG - YouTube

[5]. Haque, A., Engel, J., Teichmann, S. A., & Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome medicine*, *9*(1), 1-12.

[6]. Zappia L, Phipson B, Oshlack A. (2018) "Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database", PLOS Computational Biology, DOI: 10.1371/journal.pcbi.1006245

[7]. Picelli, S. (2017). Single-cell RNA-sequencing: the future of genome biology is now. *RNA biology*, *14*(5), 637-650.

[8]. Method of the Year 2013. (2014) *Nat Methods* **11**, 1. https://doi.org/10.1038/nmeth.2801

[9]. Science magazine (2018) Retrieved from: Breakthrough of the Year 2018 (sciencemag.org)

[10]. Chen, Z., Wei, L., Duru, F., & Chen, L. (2020). Single-cell RNA sequencing: in-depth decoding of heart biology and cardiovascular diseases. *Current genomics*, *21*(8), 585-601.

[11]. Wang, M., Gu, M., Liu, L., Liu, Y., & Tian, L. (2021). Single-cell RNA sequencing (scRNA-seq) in cardiac tissue: applications and limitations. *Vascular Health and Risk Management*, *17*, 641-657.

[12]. Ma, S. X., & Lim, S. B. (2021). Single-Cell RNA Sequencing in Parkinson's Disease. *Biomedicines*, *9*(4), 368.

[13]. Jiang, J., Wang, C., Qi, R., Fu, H., & Ma, Q. (2020). scREAD: a single-cell RNA-seq database for Alzheimer's disease. *Iscience*, *23*(11), 101769.

[14]. IBM, IBM Cloud Learn Hub, Retrieved from: What is Exploratory Data Analysis? | IBM

[15]. Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular systems biology*, *15*(6), e8746.

[16]. Herzenberg, L. A., Sweet, R. G., & Herzenberg, L. A. (1976). Fluorescence-activated cell sorting. *Scientific American*, *234*(3), 108-118.

[17]. Harvard Medical School(2015), Drop-seq: Droplet barcoding of single cells [video] Retrieved from: (72) Drop-seq: Droplet barcoding of single cells - YouTube

[18]. Soares, M. B., Bonaldo, M. D. F., Jelene, P., Su, L., Lawton, L., & Efstratiadis, A. (1994). Construction and characterization of a normalized cDNA library. *Proceedings of the National Academy of Sciences*, *91*(20), 9228-9232.

[19]. Κωνσταντίνου, Ε. (2016). *Ανάπτυξη εναλλακτικής μεθοδολογίας κλωνοποίησης γονιδίων σε πλασμίδια* (Doctoral dissertation).

[20]. Yao, C., Bora, S. A., Chen, P., Goodridge, H. S., & Gharib, S. A. (2021). Sample processing and single cell RNA-sequencing of peripheral blood immune cells from COVID-19 patients. *STAR protocols*, *2*(2), 100582.

[21]. BMH learning(2022), Unique Molecular Identifiers| Unique Molecular Indices| Molecular Barcodes [video file] Retrieved from: (72) Unique Molecular Identifiers | Unique Molecular Indices | Molecular Barcodes | - YouTube

[22]. Chipster Tutorials (2020), Introduction to scRNA-seq data analysis [video file] Retrieved from: (72) Introduction to scRNA-seq data analysis - YouTube

[23]. Balzer, M. S., Ma, Z., Zhou, J., Abedini, A., & Susztak, K. (2021). How to Get Started with Single Cell RNA Sequencing Data Analysis. *Journal of the American Society of Nephrology*, *32*(6), 1279-1292.

[24]. Chipster Tutorials (2020), scRNA-seq: Quality control and filtering cells [video] Retrieved from: scRNA-seq: Quality control and filtering cells

[25]. Chipster Tutorials (2020), scRNA-seq: Identify highly variable genes [video] Retrieved from: scRNA-seq: Identify highly variable genes

[26]. Μολόχα, Ν. Μ. (2017). *Μέθοδοι για ανάλυση δεδομένων πειράματος μικροσυστοιχίων-κανονικοποίηση* (Master's thesis).

[27]. Data Science Cornwall(2020), Visualising High-Dimensional Data with t-SNE [video] Retrieved from: (72) Visualising High-Dimensional Data with t-SNE - YouTube

[28]. StatQuest with Josh Starmer(2017), StatQuest: t-SNE, Clearly Explained [video] Retrieved from: (72) StatQuest: t-SNE, Clearly Explained - YouTube

[29]. Abdullah Al Mamun(2019), t-SNE Clearly Explained [video] Retrieved from: (72) t-SNE: Clearly Explained - YouTube

[30]. Bajram, A. (2020). *Ανάλυση βιοπληροφορικών δεδομένων με τεχνικές μηχανικής μάθησης* (Master's thesis, Πανεπιστήμιο Πειραιώς).

[31]. Chomboon, K., Chujai, P., Teerarassamee, P., Kerdprasop, K., & Kerdprasop, N. (2015, March). An empirical study of distance metrics for k-nearest neighbor algorithm. In *Proceedings of the 3rd international conference on industrial application engineering* (pp. 280-285).

[32]. Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal*, *1*(6), 90-95.

[33]. Human Cell Atlas, Home, Mission Retrieved from: Human Cell Atlas – To create comprehensive reference maps of all human cells—the fundamental units of life—as a basis for both understanding human health and diagnosing, monitoring, and treating disease.

[34]. Chipster Tutorials(2017), Differential expression analysis [video] Retrieved from: (72) Differential expression analysis - YouTube

[35]. Gong, W., Kwak, I. Y., Pota, P., Koyano-Nakagawa, N., & Garry, D. J. (2018). DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC bioinformatics*, *19*(1), 1-10.

[36]. Qiu, P. (2020). Embracing the dropouts in single-cell RNA-seq analysis. *Nature communications*, *11*(1), 1-9.

[37]. Haghverdi, L., Lun, A. T., Morgan, M. D., & Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, *36*(5), 421-427.

[38]. Tran, H. T. N., Ang, K. S., Chevrier, M., Zhang, X., Lee, N. Y. S., Goh, M., & Chen, J. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome biology*, *21*(1), 1-32.

[39]. Li, X., Wang, K., Lyu, Y., Pan, H., Zhang, J., Stambolian, D., ... & Li, M. (2020). Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nature communications*, *11*(1), 1-14.

[40]. Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., ... & Schönhuth, A. (2020). Eleven grand challenges in single-cell data science. *Genome biology*, *21*(1), 1-35.

[41]. Zheng, K., Lin, L., Jiang, W., Chen, L., Zhang, X., Zhang, Q., ... & Hao, J. (2022). Single-cell RNA-seq reveals the transcriptional landscape in ischemic stroke. *Journal of Cerebral Blood Flow & Metabolism*, *42*(1), 56-73.

[42]. Wang, X., He, Y., Zhang, Q., Ren, X., & Zhang, Z. (2021). Direct comparative analyses of 10X genomics chromium and smart-seq2. *Genomics, proteomics & bioinformatics*, *19*(2), 253-266.

[43]. Zhang, Y., Kim, M. S., Reichenberger, E. R., Stear, B., & Taylor, D. M. (2020). Scedar: A scalable Python package for single-cell RNA-seq exploratory data analysis. *PLoS computational biology*, *16*(4), e1007794.

[44]. Papili Gao, N., Hartmann, T., Fang, T., & Gunawan, R. (2020). CALISTA: clustering and LINEAGE inference in single-cell transcriptional analysis. *Frontiers in bioengineering and biotechnology*, *8*, 18.

[45]. LiquidBrain Bioinformatics (2022), No Code Seurat Analysis | Azimuth First Look [video] Retrieved from: No Code Seurat Analysis | Azimuth First Look

[46]. Bioinformagician(2022), How to analyze single-cell RNA-Seq data in R | Detailed Seurat Workflow Tutorial [video] Retrieved from: How to analyze single-cell RNA-Seq data in R | Detailed Seurat Workflow Tutorial

[47]. Satijalab, Seurat, Getting Started with Seurat, Retrieved from: Getting Started with Seurat • Seurat (satijalab.org)

[48]. LiquidBrain Bioinformatics (2022), How to setup your computer for scanpy [video] Retrieved from: How to setup your computer for scanpy

[49]. Sanbomics(2022), Single cell analysis in python with Scanpy [video] Retrievd from: Single cell analysis in python with Scanpy

[50]. Mah, C. K., Wenzel, A. T., Juarez, E. F., Tabor, T., Reich, M. M., & Mesirov, J. P. (2018). An accessible, interactive GenePattern Notebook for analysis and exploration of single-cell transcriptomic data. *F1000Research*, *7*.

[51]. SCGEATOOL:: Single-cell Gene Expression Analysis Tool, Retrieved from: SCGEATOOL :: Single-cell Gene Expression Analysis Tool

[52]. Le, T., Phan, T., Pham, M., Tran, D., Lam, L., Nguyen, T., ... & Pham, S. (2020). BBrowser: Making single-cell data easily accessible. *BioRxiv*.

[53]. Yuan Cao, Junjie Zhu, Guangchun Han, Peilin Jia, Zhongming Zhao , scRNASeqDB: a database for gene expression profiling in human single cell by RNA-seq bioRxiv 104810; doi: https://doi.org/10.1101/104810 Website: scRNASeqDB (uth.edu)

[54]. Kimmerling RJ, Lee Szeto G, Li JW, Genshaft AS et al. A microfluidic platform enabling single-cell RNA-seq of multigenerational lineages. *Nat Commun* 2016 Jan 6;7:10220. PMID: 26732280

[55]. Chung W, Eum HH, Lee HO, Lee KM et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat Commun* 2017 May 5;8:15081. PMID: 28474673

[56]. GEO(Gene Expression Omnibus), Website: Home - GEO - NCBI (nih.gov)

[57]. Mercadante, A. A., Dimri, M., & Mohiuddin, S. S. (2019). Biochemistry, replication and transcription , Retrieved from: Biochemistry, Replication and Transcription - StatPearls - NCBI Bookshelf (nih.gov)

[58]. Cai, J. J. (2020). scGEAToolbox: a Matlab toolbox for single-cell RNA sequencing data analysis.

[59]. Reich, M., Tabor, T., Liefeld, T., Thorvaldsdóttir, H., Hill, B., Tamayo, P., & Mesirov, J. P. (2017). The GenePattern notebook environment. *Cell systems*, *5*(2), 149-151.