ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΔΙΙΔΡΥΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
"ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗΝ ΙΑΤΡΙΚΗ ΚΑΙ ΤΗ ΒΙΟΛΟΓΙΑ"

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Δημιουργία πακέτου R ανάλυσης γονιδιακής έκφρασης κυττάρων που αναγνωρίζει τις κυτταρικές καταστάσεις και ανακατασκευάζει ρυθμιστικά δίκτυα για τις πιθανές μεταβάσεις καταστάσεων με μη-εποπτευόμενη μηχανική μάθηση

Ευθυμία Χ. Μαλέσιου

Επιβλέπων: **Ηλίας Μανωλάκος,** Καθηγητής, Τμήμα Πληροφορικής και Τηλεπικοινωνιών, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

ΑΘΗΝΑ

ΔΕΚΕΜΒΡΙΟΣ 2019

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Δημιουργία πακέτου R ανάλυσης γονιδιακής έκφρασης κυττάρων που αναγνωρίζει τις κυτταρικές καταστάσεις και ανακατασκευάζει ρυθμιστικά δίκτυα για τις πιθανές μεταβάσεις καταστάσεων με μη-εποπτευόμενη μηχανική μάθηση

Ευθυμία Χ. Μαλέσιου

Α.Μ.: ΠΙΒ0171

ΕΠΙΒΛΕΠΩΝ: **Ηλίας Μανωλάκος,** Καθηγητής, Τμήμα Πληροφορικής και Τηλεπικοινωνιών, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:
Martin Reczko, Ειδικός Λειτουργικός Επιστήμονας Α', Ερευνητικού Κέντρου Βιοϊατρικών Επιστημών «Αλέξανδρος Φλέμινγκ»
Έμα Αναστασιάδου, Ερευνήτρια Δ', Ίδρυμα Ιατροβιολογικών Ερευνών, Ακαδημίας Αθηνών
Ηλίας Μανωλάκος, Καθηγητής, Τμήμα Πληροφορικής και Τηλεπικοινωνιών, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Δεκέμβριος 2019

ΠΕΡΙΛΗΨΗ

Η δυνατότητα ποσοτικοποίησης κι ανάλυσης των προφίλ γονιδιακής έκφρασης σε επίπεδο μονήρων κυττάρων (single-cells), έχει επιτρέψει τη μελέτη της ετερογένειας των κυτταρικών πληθυσμών στους ιστούς, την αναγνώριση σπάνιων καταστάσεων και τη διερεύνηση του ρόλου τους και των αποτελεσμάτων της αλληλεπίδρασής τους με το μικρο-περιβάλλον. Ιδιαίτερο ενδιαφέρον, παρουσιάζει η μελέτη των δυναμικών μεταβάσεων ή τροχιών που σχηματίζονται μεταξύ ζευγών κυτταρικών καταστάσεων. Πρόσφατα, αναπτύχθηκαν αρκετοί αλγόριθμοι για την ανακατασκευή τροχιών, οι κύριες διαφορές μεταξύ των οποίων, είναι η απαίτηση εκ των προτέρων πληροφορίας, ο τρόπος διαμόρφωσης της τοπολογίας, η διάταξη των κυττάρων και το μαθηματικό πλαίσιο στο οποίο βασίζονται.

Στη δημοσίευση των Τσακανίκα Π., Μανατάκη Δ. και Μανωλάκου Η.Σ., *«Machine learning methods to reverse engineer dynamic gene regulatory networks governing cell state transitions»*, bioRxiv, 2018 (DOI: http://dx.doi.org/10.1101/264671), περιγράφεται ένα πιθανοτικό πλαίσιο μη-εποπτευόμενης μηχανικής μάθησης για την ανακατασκευή δυναμικών γονιδιακών ρυθμιστικών δικτύων που καθοδηγούν τη μετάβαση μεταξύ κυτταρικών καταστάσεων, εισαγάγοντας, ταυτόχρονα, την έννοια των μικρο-καταστάσεων σε μία τροχιά. Για τη δημιουργία του προτύπου που περιγράφει το «επιγενετικό τοπίο», χρησιμοποιείται ένα μείγμα κανονικών κατανομών με τις εκ των υστέρων πιθανότητες που προκύπτουν να καθορίζουν τις κυτταρικές καταστάσεις και τις πιθανές μεταβάσεις μεταξύ τους. Περαιτέρω, σε κάθε τροχιά μετάβασης που σχηματίζεται (μετάβαση από την κατάσταση «έναρξης» προς την κατάσταση «προορισμού»), προσδιορίζονται διαδοχικές μικρο-καταστάσεις (φάσεις μετάβασης) κι αναγνωρίζονται τα κύρια γονίδια – ρυθμιστές, καταλήγοντας στη δημιουργία στοχευμένων αιτιατών γονιδιακών ρυθμιστικών δικτύων ανά μικρο-κατάσταση.

Η παρούσα διπλωματική εργασία, αφορά στη δημιουργία πακέτου R (MLscAN: Machine Learning single-cell ANalytics) που βασίζεται στη μεθοδολογία της παραπάνω δημοσίευσης (Tsakanikas P. et al., 2018), με δυνατότητα ευέλικτης εκτέλεσης όλων των βημάτων, με μόνη απαιτούμενη είσοδο, τα προ-επεξεργασμένα δεδομένα έκφρασης. Εκτός των προκαθορισμένων επιλογών, δίνεται η ευχέρεια στους χρήστες να ενσωματώσουν σε οποιοδήποτε βήμα της διαδικασίας, δικούς τους αλγορίθμους ή ήδη διαθέσιμα αποτελέσματα, αλλά, και να παρέμβουν μετά τη δημιουργία του προτύπου MLscAN, τροποποιώντας στοιχεία στοχευμένα. Το πακέτο, μπορεί να χρησιμοποιηθεί για την παραγωγή κι οπτικοποίηση των αποτελεσμάτων ανάλυσης σε διαφορετικά στάδια της ροής επεξεργασίας· από τη διερεύνηση του προ-επεξεργασμένου πίνακα δεδομένων έως τη μείωση της διαστατικότητας, τον προσδιορισμό των κυτταρικών καταστάσεων και των πιθανών μεταβάσεων, την εξαγωγή των τροχιών και των μικρο-καταστάσεων, την αναγνώριση των κύριων γονιδίων και την κατασκευή των αιτιατών γονιδιακών ρυθιστικών δικτύων στο επίπεδο της μικρο-κατάστασης, με χρήση μη-εποπτευόμενων μεθοδολογιών μηχανικής μάθησης.

Τέλος, το πακέτο R χρησιμοποιήθηκε στην εργασία, για την ανάλυση ενός δημοσιευμένου συνόλου δεδομένων που αφορά στην τροχιά από-διαφοροποίησης β-κυττάρων των νησιδίων του Langerhans ατόμων με σακχαρώδη διαβήτη τύπου 2, με στόχο τη διερεύνηση των αποτελεσμάτων που παραγάγονται σε σχέση με τις επιλεγμένες παραμέτρους.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: μηχανική μάθηση, ανάλυση δεδομένων μονήρων κυττάρων, βιοπληροφορική

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: μονήρη κύτταρα, μετάβαση μεταξύ καταστάσεων, επιγενετικό τοπίο, τροχιά, μικρο-κατάσταση, γονιδιακό ρυθμιστικό δίκτυο

ABSTRACT

Our ability to measure and analyze gene expression profiles at the single-cell level has enabled the study of the heterogeneity of cell populations in tissues, the identification of rare cell states, as well as their role and interaction with the micro-environment. Of special interest is the study of dynamic transitions or trajectories, formed between pairs of cell states. Recently, many trajectory inference algorithms have been proposed; their main differences lie in requiring or not prior information, the methodology applied to determine the topology, the ordering of the cells and the mathematical frameworks they are based upon.

In their recent paper, *"Machine learning methods to reverse engineer dynamic gene regulatory networks governing cell state transitions"*, bioRxiv, 2018 (DOI: http://dx.doi.org/10.1101/264671), Tsakanikas P., Manatakis D. and Manolakos E.S., have proposed a probabilistic machine learning framework for the reconstruction of dynamic gene regulatory networks (GRNs) governing cell state transitions, without supervision, while introducing the concept of a trajectory's micro-states. Furthermore, each transition's trajectory (from a "ground" cell-state to a "landing" cell-state), is partitioned into consecutive micro-states, and after the transition's key-genes are identified, a causal GRN can be inferred per micro-state.

The main objective of this thesis was the development of an R package (MLscAN: Machine Learning single-cell ANalytics) based on the methodology of the aforementioned article, to execute the full workflow, only requiring the pre-processed expression data as input. Besides the default settings, the users may incorporate, at each stage of the process, their own algorithms or previously generated results. Also, the users may focus on any object and specifically alter it. The package can be used to generate and visualize the results of the top-down analysis at different stages of the workflow, from the pre-processed data matrix exploration to dimensionality reduction, states and possible transitions identification, trajectories and micro-states extraction, key-genes identification and causal GRNs inference down to the micro-state level, based on unsupervised machine learning methods.

Finally, the developed R package was used to analyze a published dataset concerned with the dedifferentiation trajectory of β-cells of the islets of Langerhans of subjects with type 2 diabetes mellitus, aiming at exploring the results generated in conjunction with the selected parameters.

SUBJECT AREA: machine learning, single-cell data analysis, bioinformatics

KEYWORDS: single-cells, state transition, epigenetic landscape, trajectory, micro-state, gene regulatory networks

«…πᾶσά τε ἐπιστήμη χωριζομένη δικαιοσύνης καὶ τῆς ἄλλης ἀρετῆς πανουργία, οὐ σοφία φαίνεται.», Μενέξενος, Πλάτωνας

"All sentient beings should have at least one right – the right not to be treated as property", Gary L. Francione

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω, αρχικά, τον επιβλέποντα, κ. Ηλία Μανωλάκο, για την ευκαιρία που μου δόθηκε να ασχοληθώ με ένα σύγχρονο κι αρκετά ενδιαφέρον θέμα, για τον χρόνο που αφιέρωσε, την καθοδήγηση και τις ιδέες βελτίωσης.

Επίσης, θα ήθελα να ευχαριστήσω, τον κ. Πάνο Τσακανίκα και τον κ. Δημήτρη Μανατάκη· εκτός του ότι ήταν, μαζί με τον κ. Μανωλάκο, οι συγγραφείς του άρθρου στο οποίο βασίστηκε η διπλωματική εργασία, επίσης αφιέρωσαν χρόνο, πρόσφεραν χρήσιμες συμβουλές κι ιδέες που διαμόρφωσαν το τελικό αποτέλεσμα. Ακόμη, ήταν ιδιαίτερα βοηθητική κι η παροχή του κώδικα που χρησιμοποιήθηκε στο άρθρο, από τον κ. Τσακανίκα.

Τέλος, θα ήθελα να ευχαριστήσω, τον κ. Martin Reczko και την κ. Έμα Αναστασιάδου για τη συμμετοχή τους στην εξεταστική επιτροπή.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

1. ΕΙΣΑΓΩΓΗ

Στο κεφάλαιο αυτό, παρατίθενται, το πλαίσιο της έρευνας, η ισχύουσα κατάσταση σε ό,τι αφορά τις μεθόδους δημιουργίας τροχιών που σχηματίζονται μεταξύ των κυτταρικών καταστάσεων, η καινοτομίας της νέας μεθόδου στην οποία βασίζεται το πακέτο που αναπτύχθηκε, καθώς κι οι στόχοι της εργασίας ως προς το πακέτο R, την ανάλυση του δημοσιευμένου συνόλου δεδομένων, και τέλος, η οργάνωση των ενοτήτων.

1.1 Γενικά

1.1.1 Πλαίσιο της έρευνας

Το ποσό των πληροφοριών που μπορεί να «κρύβει» ένα κύτταρο είναι τεράστιο· το μεταγράφωμα (transcriptome), το πρωτεΐνωμα (proteome), το μεταβόλωμα (metabolome) κι η δομή του, είναι λίγες από τις διαστάσεις τους. Με δεδομένο ότι το σώμα του μέσου ανθρώπου, αποτελείται από περίπου $3*10^{13}$ κύτταρα [1], η πολυπλοκότητα αυξάνεται ακόμη περισσότερο, καθιστώντας εξαιρετικά δύσκολη τη σε βάθος και συνδυαστική μελέτη τους. Στο επίπεδο του μεταγραφώματος σε δεδομένη χρονική στιγμή (snapshot), με εκτίμηση της έκφρασης των γονιδίων από δεδομένα αλληλούχησης του RNA, η δυνατότητα διερεύνησης σε επίπεδο μονήρων κυττάρων (single-cells), που αναπτύσσεται ιδιαίτερα τα τελευταία χρόνια, δημιουργεί νέες προοπτικές, επιτρέποντας μεγαλύτερη ευκρίνεια.

Με την ομαδοποίηση (bulk) των κυττάρων, τα δεδομένα της έκφρασης, αποτελούσαν εκτίμηση μόνο της μέσης κατάστασης, που ίσως να μην αντιπροσωπευόταν από κανένα πραγματικό κύτταρο. Αυτή η ομαδοποίηση, αναγκαστικά, μείωνε και την ικανότητα να εντοπιστεί, στα κύτταρα ενός ιστολογικού τύπου, η παρουσία διακριτών υπο-πληθυσμών ή σπάνιων καταστάσεων, που ενδεχομένως να μην είναι γνωστά, αλλά ταυτόχρονα, να έχουν καίριο ρόλο σε φυσιολογικές ή παθολογικές διεργασίες. Επίσης, είναι σημαντική κι η δυνατότητα εστίασης στις αλληλεπιδράσεις μεταξύ τους, αλλά, και με το περιβάλλον τους, για την αποκάλυψη των μηχανισμών που ενέχονται στις διεργασίες αυτές.

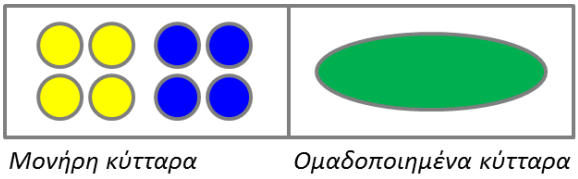
Ταυτόχρονα, ωστόσο, αναδύονται πρόσθετες δυσκολίες, όπως η υψηλή παρουσία θορύβου, ο τρόπος χειρισμού του εύρος του λόγου εντοπισμού της έκφρασης ενός γονιδίου στα διαθέσιμα κύτταρα (περίπου, μεταξύ 1% και 65% [2]), ο ποιοτικός έλεγχος, κατά τον οποίον είναι σύνηθες να απομακρύνεται σημαντικό μέρος των αρχικά επιλεγμένων κυττάρων, ο εντοπισμός ζευγών κυττάρων που θεωρούνται ένα (doublets) λόγω τεχνικού σφάλματος, η επιλογή των γονιδίων που θα χρησιμοποιηθούν κι ο μετασχηματισμός (transformation) του πίνακα έκφρασης.

Οι τροχιές (trajectories) ή οι μεταβάσεις (transitions) που σχηματίζονται μεταξύ των κυτταρικών καταστάσεων (cell-states) που εντοπίζονται, παρέχουν έναν άξονα διάταξης των κυττάρων, στο σύνολο ή απομονωμένα. Αποτελούν διατάξεις των κυττάρων με διαφορετικούς περιορισμούς και θα οριστούν με μεγαλύτερη ακρίβεια στο κεφάλαιο 2 (2.2.6, 2.2.7.1). Αυτές, θέτουν το πλαίσιο για να αναγνωριστούν οι σημαντικοί παράγοντες – γονίδια και να εκτιμηθούν, η βαρύτητά τους κι ο τύπος των μεταξύ τους αλληλεπιδράσεων – δηλαδή, να ανακατασκευαστεί το γονιδιακό ρυθμιστικό δίκτυο (GRN: Gene Regulatory Network).

Η συνολική θεώρηση των αποτελεσμάτων της παραπάνω διαδικασίας, οδηγεί στον σχηματισμό ενός «επιγενετικού τοπίου» (epigenetic landscape) [3]· μίας

αναπαράστασης των παραγόντων που αλληλεπιδρούν έχοντας ως αποτέλεσμα τροποποιήσεις στη γονιδιακή έκφραση που επηρεάζουν δυνητικά τις κυτταρικές λειτουργίες κι οδηγούν στη διαμόρφωση της «ποικιλομορφίας» των κυττάρων.

Για παράδειγμα, ένας όγκος μπορεί να αποτελείται από ένα σύνολο υπο-τύπων και κυττάρων με μεικτή έκφραση δεικτών (markers) [4]. Οι δυναμικές σχέσεις μεταξύ τους, η ιεραρχική τοποθέτηση των καταστάσεων, η εκτροπή ενός υπο-πληθυσμού με μετάβαση σε άλλη κατάσταση (π.χ. αποδιαφοροποίηση) και τα πρότυπα κλωνικής επέκτασης, είναι κάποιες από τις κύριες περιοχές ενδιαφέροντος για την πιθανή ανάκτηση περισσότερης και πιο εξειδικευμένης γνώσης. Εστιάζοντας στην απάντηση του ανοσοποιητικού, υπάρχουν πτυχές που παραμένουν άγνωστες· για παράδειγμα, δεν είναι γνωστό αν υπάρχει συγκεκριμένη στρατηγική αντιμετώπισης και πώς αυτή επηρεάζει την εξέλιξη [5] – αν κινητοποιούνται κύτταρα από την περιφέρεια, από εγγύς υγιείς ιστούς ή από την περιοχή του όγκου. Ακόμη, και χαρακτηριστικά όπως το μεταβολικό τους προφίλ, που σε πρώτο επίπεδο ενδεχομένως να μη θεωρούνται τόσο άμεσα σχετιζόμενα, μπορούν να επηρεάσουν την αποτελεσματικότητα της ανοσοθεραπείας [6]. Έτσι, η μελέτη των τροχιών μετάβασης προς κυτταρολυτικά Τ κύτταρα, ενδεχομένως να αποκαλύψει τα κύρια γονίδια που καθοδηγούν τη μετάβαση και το πώς αλληλεπιδρούν ώστε να ενισχυθεί αυτή η μετατροπή. Ακόμη, έχει δειχθεί ότι το δίκτυο που ανακτήθηκε από την αντιστρέψιμη μετάβαση μελανοκυττάρων σε κατάσταση που χαρακτηριζόταν από αντοχή σε φαρμακευτικό παράγοντα, έκανε δυνατή την αναστολή αυτής της μετάβασης [7]. Οπότε, δημιουργούνται νέες οδοί, για την κατανόηση, θεραπεία και πρώιμη διάγνωση ασθενειών.



Εικόνα 1.1: Αφαιρετική αναπαράσταση της εικόνας που προσφέρει η ανάλυση μονήρων κυττάρων σε σύγκριση με αυτήν που προκύπτει από την ομαδοποίησή τους.

1.1.2 Χρησιμότητα

1.1.2.1 Ισχύουσα κατάσταση

Ο αριθμός των μεθοδολογιών που αφορούν στη δημιουργία τροχιών από δεδομένα έκφρασης μονήρων κυττάρων, ξεπερνά τις 70 και για τις περισσότερες είναι δωρεάν διαθέσιμη υλοποίησή τους σε R ή Python. Σε πρόσφατο άρθρο [8] όπου συγκρίνονται 45 από αυτές, γίνεται εκτενής αξιολόγηση και σύγκρισή τους.

Βασικές διαφορές των μεθοδολογιών, είναι, η τοπολογία των τροχιών που χειρίζονται, το αν είναι προκαθορισμένη ή αποτέλεσμα συμπερασμού (inference) καθώς και το αν είναι εποπτευόμενες (supervised) ή μη. Αρκετές μέθοδοι, απαιτούν κάποιες πρόσθετες πληροφορίες, εκτός του πίνακα έκφρασης, και κυρίως, το κύτταρο έναρξης, την αρχική και τελική καταστάσεις, τον αριθμό των καταστάσεων και την κατεύθυνση της τροχιάς. Βέβαια, μπορεί να θεωρηθεί σε κάποιες περιπτώσεις βοηθητική η παροχή αυτών των πληροφοριών για την αύξηση της βιολογικής σημασίας των αποτελεσμάτων.

Ως προς την τοπολογία, το εύρος είναι μεγάλο, ξεκινώντας από απλή γραμμική τροχιά, στηριζόμενες συνήθως στην υπόθεση της ομαλής μετάβασης μεταξύ αρκετά

ομοιογενών κυττάρων, και καταλήγοντας σε μη-συνεκτικούς γράφους (disconnected graphs) [8]. Οι νεώτερες μέθοδοι, συνήθως δε λειτουργούν με προκαθορισμένη τοπολογία.

Επίσης, διαφέρουν στους ορισμούς και στις υποθέσεις στις οποίες στηρίζονται. Σε κάποιες επιτρέπεται μία κατάσταση ή ένα κύτταρο να ανήκουν σε πολλές τροχιές, με διαφορετικούς ψευδοχρόνους (pseudotimes), δηλαδή, προβολές στις τροχιές ή ακόμη δημιουργείται και καθολικός ψευδοχρόνος. Σε άλλες, δημιουργείται ένας σκελετός των τροχιών από οδηγά κύτταρα, μεταξύ των οποίων τοποθετούνται τα υπόλοιπα. Επιπλέον, διαφορές εντοπίζονται και στις προσεγγίσεις μείωσης της διαστατικότητας, υπολογισμού του αριθμού των καταστάσεων και του ψευδοχρόνου. Τέλος, σε ορισμένες, ενέχεται στοχαστικότητα και χρήση ευριστικών αλγορίθμων για τη δημιουργία των τροχιών.

Διαπιστώνεται [8] ότι δεν υπάρχει κάποια που να λειτουργεί το ίδιο καλά σε όλους τους συνδυασμούς αναμενόμενων τοπολογιών και διαθέσιμων δεδομένων, αν και περιορισμένος αριθμός είχε την τάση να ανταποκρίνεται καλά σε σημαντικό ποσοστό των συνόλων δεδομένων και τοπολογιών. Συμπεραίνεται ότι υπάρχει συμπληρωματικότητα μεταξύ τους κι είναι προτιμότερο να επιβεβαιώνονται τα αποτελέσματα μίας μεθόδου, χρησιμοποιώντας κι άλλες.

Όσον αφορά την υλοποίηση, σε πολλές περιπτώσεις, υπάρχει σημαντικό πρόβλημα στην κλιμάκωση, δηλαδή στην ικανότητα χειρισμού δεδομένων αυξημένης διαστατικότητας σε εύλογο χρονικό διάστημα και χωρίς υπερβολικές απαιτήσεις μνήμης [8]. Ακόμη, σημαντικές αδυναμίες, είναι η σχετική έλλειψη ευρωστίας, δηλαδή, ικανότητας παραγωγής αποτελεσμάτων με μικρές αποκλίσεις όταν η είσοδος έχει μικρές διαφορές, κι η έλλειψη φιλικότητας στον χρήστη.

Παρακάτω παρατίθεται η αριθμητική κατανομή των μεθόδων [8], σε σχέση με τον τύπο της τοπολογίας και τη διάθεση προηγούμενης γνώσης (εποπτευόμενες).

Αριθμός μεθόδων που μπορούν να χρησιμοποιηθούν όταν διατίθεται προηγούμενη γνώση / πληροφορία:

- συγκεντρωτικά, με οποιουδήποτε τύπου προηγούμενη γνώση: 47,

  κι αυτές, κατανέμονται περαιτέρω βάσει της αναμενόμενης τοπολογίας:

  - γραμμική: 41
  - κυκλική: 6
  - με διακλάδωση: 29
  - με πολλαπλές διακλαδώσεις: 27
  - δένδρο: 21
  - ακυκλικός γράφος: 3
  - συνδεδεμένος γράφος: 3
  - ασύνδετος γράφος (ελεύθερη τοπολογία): 2

- παρέχοντας μόνο τα αρχικά και τελικά κύτταρα: 38

- παρέχοντας μόνο για τους αριθμούς των αρχικών και τελικών καταστάσεων: 36
- παρέχοντας μόνο τη συσταδοποίηση: 32

Αριθμός μεθόδων που μπορούν να χρησιμοποιηθούν χωρίς τη διάθεση προηγούμενης γνώσης / πληροφορίας:

- 31,

  κι αυτές, κατανέμονται περαιτέρω βάσει της αναμενόμενης τοπολογίας:

  - γραμμική: 27
  - κυκλική: 4
  - με διακλάδωση: 15
  - με πολλαπλές διακλαδώσεις: 14
  - δένδρο: 14
  - ακυκλικός γράφος: 1
  - συνδεδεμένος γράφος: 1
  - ασύνδετος γράφος (ελεύθερη τοπολογία): 1

Από τις μεθόδους που εντάσσονται στις μη-εποπτευόμενες και με τη δυνατότητα η τοπολογία να είναι τουλάχιστον δενδροειδής (14), με την επιπλέον απαίτηση να εκτελείται η διαδικασία σχηματισμού των τροχιών, για 10.000 κύτταρα και 1.000 γονίδια, εντός μίας ώρας και με έως 10GB [8], απομένουν, οι ακόλουθες πέντε:

- cellTree VEM [9]
  - χρησιμοποιεί τη λανθάνουσα κατανομή Dirichlet [10]
  - δημιουργεί ιεραρχική δενδροειδή δομή
  - μπορεί να χρησιμοποιηθεί και με χωρικά δεδομένα
- Monocle DDRTree [11]
  - σχηματισμός των τροχιών μεταξύ των κεντροειδών (μετά από τη χρήση του αλγορίθμου των *k*-μέσων τιμών [12]) στον χώρο των αποτελεσμάτων μείωσης της διαστατικότητας, με χρήση ελάχιστου καλύπτοντος δένδρου (minimum spanning tree)
  - τα κύτταρα, μετακινούνται διαρκώς προς τις κοντινότερες κορυφές, του διαρκώς ανασχηματιζόμενου ελάχιστου καλύπτοντος δένδρου έως να επιτευχθεί σύγκλιση

- ο χρήστης, επιλέγει ένα κύτταρο που θεωρείται ρίζα του δένδρου για τον υπολογισμού του ψευδοχρόνου βάσει της απόστασης των κυττάρων από τη ρίζα

- RaceID / StemID [13]
  - δε δέχεται ούτε προαιρετικά τη χρήση οποιασδήποτε προηγούμενης γνώσης
  - εξειδικεύεται στην εντόπιση σπάνιων κυττάρων που άλλες μέθοδοι θεωρούν ακραία
  - η συσταδοποίηση γίνεται με τον αλγόριθμο των *k*-ενδιαμέσων τιμών [14]
  - τα κύτταρα τοποθετούνται στον ψευδοχρόνο των τροχιών, με προβολή του διανύσματος του κυττάρου με αρχή την ενδιάμεση τιμή της συστάδας που ανήκει, στο τμήμα που συνδέει την ενδιάμεση αυτή τιμή με την ενδιάμεση τιμή μίας άλλης συστάδας

- SLICE [15]
  - χρησιμοποιεί την κυτταρική εντροπία για τον σχηματισμό των τροχιών
  - επιλέγει αυτόματα τον αριθμό των καταστάσεων

- Slingshot [16]
  - βασίζεται στη δομή του ελάχιστου επικαλύπτοντος δένδρου για τον σχηματισμό της καθολικής δομής
  - προσαρμόζει κύριες καμπύλες (principal curves) [17] για τον σχηματισμό των ομαλών εξελίξεων στα επιμέρους τμήματα και την εξαγωγή του ψευδοχρόνου

1.1.2.2 Καινοτομία της νέας μεθόδου

Όταν απαιτείται μέγιστη ευελιξία, χωρίς την απαίτηση προηγούμενης γνώσης και χωρίς περιορισμούς στην τοπολογία, τότε, απομένει μόνο μία μέθοδος από τις πέντε που παρατέθηκαν στην προηγούμενη ενότητα· η RaceID / StemID [13]. Συγκριτικά, η νέα μέθοδος [18], έχει τα εξής πλεονεκτήματα:

- έχει ελάχιστες απαιτήσεις· συγκεκριμένα, τον προ-επεξεργασμένο πίνακα έκφρασης, και μπορεί να εκτελέσει όλα τα απαιτούμενα βήματα χωρίς καμία πρόσθετη παρέμβαση του χρήστη

- είναι μη-εποπτευόμενη, χωρίς να αποκλείει τη χρήση προηγούμενης γνώσης (π.χ. αριθμός καταστάσεων)
- η προκαθορισμένη ροή επεξεργασίας, δεν είναι στοχαστική, και κατά συνέπεια, αναμένεται να μη μεταβάλλονται τα αποτελέματα όταν η είσοδος είναι ίδια, με την εξαίρεση των περιπτώσεων επίλυσης ισοπαλιών, που ενδέχεται να προκύψουν, για παράδειγμα, κατά τη διαδικασία σχηματισμού των συστάδων
- δημιουργεί ένα πιθανοτικό (probabilistic) πρότυπο για τις μεταβάσεις μεταξύ των κυτταρικών καταστάσεων
- εξαγάγει τις τροχιές από τον χώρο των πιθανοτήτων
- ορίζει και χρησιμοποιεί τις μικρο-καταστάσεις σε μία τροχιά, δηλαδή, διαιρεί την τροχιά σε στάδια εξέλιξης
- επιτρέπει τη συμμετοχή μίας κατάστασης σε πολλές τροχιές, όμως, με διαφορετικό υπο-σύνολο κυττάρων σε καθεμία από αυτές, ώστε ένα κύτταρο είναι δυνατό να συμμετέχει το πολύ σε ένα ζεύγος αντίθετων τροχιών
- δε σταματά στη δημιουργία των τροχιών, αλλά, συνεχίζει έως και τη δημιουργία ενός GRN ανά μικρο-κατάσταση
- δεν αναμένει μόνο αρκετά ομοιογενή κύτταρα
- δεν αναμένει όλα τα κύτταρα να συμμετέχουν στην ίδια τροχιά
- δε βασίζεται σε κάποια προκαθορισμένη τοπολογία

1.2 Στόχοι της Εργασίας

1.2.1 Δημιουργία πακέτου R

Ο κύριος στόχος της διπλωματικής εργασίας, είναι, η δημιουργία ενός πακέτου R, που θα εκτελεί όλα τα βήματα της μεθοδολογίας που περιγράφονται στη δημοσίευση [18], έχοντας τις ελάχιστες δυνατές απαιτήσεις εισόδου. Ταυτόχρονα, θα πρέπει να παρέχονται εύκολοι τρόποι παρέμβασης κι εξατομίκευσης των παραμέτρων σε όλα τα στάδια, με επανακαθορισμό μόνο των μερών που επηρεάζονται, ώστε να μπορεί να προσαρμόζεται στις απαιτήσεις διαφορετικών συνόλων δεδομένων και να επιτρέπει τη σύγκριση των αποτελεσμάτων που παράγονται. Επιπλέον, θα πρέπει να εξαντλεί τα περιθώρια δημιουργίας αποτελεσμάτων, ανακάμπτοντας από καταστάσεις όπου η τυπική διαδικασία αδυνατεί να προχωρήσει. Πρόσθετα, θα πρέπει να ανταποκρίνεται σε δεδομένα υψηλής διασταστικότητας, τουλάχιστον της τάξης 10^5, να εκτελείται σε εύλογο χρονικό διάστημα, να γίνεται συστηματική πρόληψη λαθών με έλεγχο της εισόδου, να είναι εύκολη η χρήση του, το εγχειρίδιο τεκμηρίωσης μέσω παραδειγμάτων να συμβάλλει στην καλύτερη κατανόηση της ροής επεξεργασίας και των δυνατοτήτων, και να δημιουργούνται αυτόματα επαρκή και πληροφοριακά αρχεία εξόδου – διαγράμματα και αρχεία πληροφοριών – που θα συμβάλλουν στην ευκολότερη οπτική επισκόπηση και κατανόηση των αποτελεσμάτων.

1.2.2 Ανάλυση συνόλου δεδομένων με χρήση του πακέτου R

Δευτερεύων στόχος, είναι η χρησιμοποίηση του πακέτου R για την ανάλυση δημοσιευμένου συνόλου δεδομένων, για τη διερεύνηση και την επιβεβαίωση των αποτελεσμάτων που παράγονται και του πώς επηρεάζονται σε σχέση με τις επιλεγμένες παραμέτρους.

1.3 Οργάνωση της Εργασίας

Στο κεφάλαιο 2, αρχικά παρουσιάζεται συνοπτικά η μεθοδολογία, και στη συνέχεια παρατίθενται αναλυτικές πληροφορίες και το θεωρητικό υπόβαθρο κάθε βήματος (μείωση της διαστατικότητας (dimensionality reduction), εκτίμηση των εκ των υστέρων (posterior) πιθανοτήτων, διαμόρφωση των κυτταρικών καταστάσεων (states), ορισμός των τροχιών, ορισμός των μικρο-καταστάσεων (micro-states), αναγνώριση των κύριων γονιδίων (key-genes) της τροχιάς, δημιουργία των γονιδιακών ρυθιστικών δικτύων ανά μικρο-κατάσταση (GRN)). Στο κεφάλαιο 3, περιγράφονται, η δομή κι οι δυνατότητες του πακέτου R, μαζί με παραδείγματα χρήσης. Στο κεφάλαιο 4, παρουσιάζονται τα αποτελέσματα από τη χρήση του πακέτου R, για την ανάλυση δημοσιευμένου συνόλου δεδομένων, καταδεικνύοντας τη χρησιμότητα και τις προοπτικές της μελέτης των τροχιών σε δεδομένα έκφρασης μονήρων κυττάρων. Τέλος, στο κεφάλαιο 5, βρίσκονται τα γενικά συμπεράσματα.

2. ΜΕΘΟΔΟΛΟΓΙΑ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΓΙΑ ΤΗΝ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΜΟΝΗΡΩΝ ΚΥΤΤΑΡΩΝ

Σε αυτό το κεφάλαιο, αρχικά, συνοψίζεται η ροή επεξεργασίας των δεδομένων, που ακολουθείται στο πακέτο R, και στη συνέχεια αναλύεται καθένα από τα βήματα επεξεργασίας: μείωση της διαστατικότητας, εκτίμηση εκ των υστέρων πιθανοτήτων, διαμόρφωση των καταστάσεων, αναγνώριση των ακραίων καταστάσεων, δημιουργία των τροχιών, και τέλος, ορισμός των μικρο-καταστάσεων και των κύριων γονιδίων τους και κατασκευή των GRNs ανά μικρο-κατάσταση. Δίνεται εμφαση στην περιγραφή των μεθόδων που υλοποιήθηκαν, κι ιδίως όσων βελτιώνουν αυτές που παρουσιάζονται στο άρθρο [18]. Η περιγραφή του πακέτου και παραδείγματα χρήσης, παρουσιάζονται στα κεφάλαια 3 και 4.

2.1 Σύνοψη της μεθοδολογίας – ροή επεξεργασίας δεδομένων (workflow)

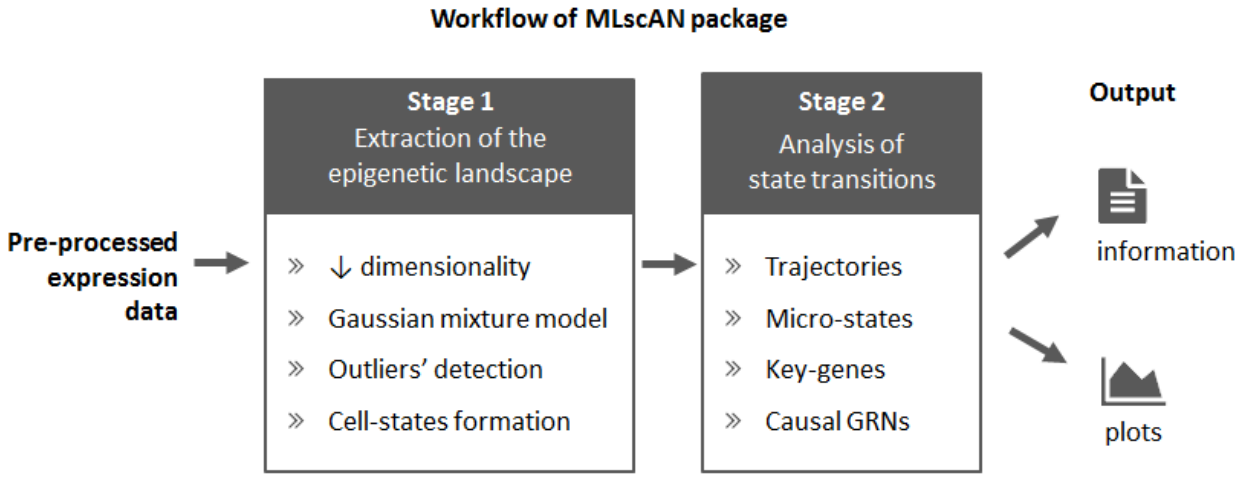
Ξεκινώντας από τα προ-επεξεργασμένα (pre-processed) δεδομένα έκφρασης και λαμβάνοντας υπόψη όποιες προαιρετικές παραμέτρους οριστούν, η ροή επεξεργασίας μπορεί να διακριθεί σε δύο στάδια· στο πρώτο, όπου πραγματοποιείται η εξαγωγή του επιγενετικού τοπίου (epigenetic landscape) και περιλαμβάνονται τα βήματα επεξεργασίας έως τη διαμόρφωση των κυτταρικών καταστάσεων (cell-states), και στο δεύτερο, όπου αναλύονται οι μεταβάσεις (transitions) μεταξύ των καταστάσεων, σχηματίζονται οι τροχιές (trajectories), οι μικρο-κατατάσεις (micro-states) των τροχιών, αναγνωρίζονται τα κύρια γονίδιά (key-genes) τους και δημιουργούνται τα GRNs ανά μικρο-κατάσταση. Στο τέλος, παραγάγονται αρχεία που συνοψίζουν και προβάλουν τα αποτελέσματα του προτύπου MLscAN (αρχεία πληροφοριών και διαγράμματα).

Στάδιο 1 – Εξαγωγή του επιγενετικού τοπίου:

- Μείωση της διαστατικότητας (dimensionality reduction)
- Δημιουργία του προτύπου μείξης κανονικών κατανομών (gaussian mixture model)
- Αναγνώριση κι αφαίρεση των ακραίων κυττάρων (outliers)
- Επαναδημιουργία του προτύπου μείξης κανονικών κατανομών, αν στο προηγούμενο βήμα αφαιρέθηκαν κύτταρα
- Σχηματισμός των κυτταρικών καταστάσεων

Στάδιο 2 – Ανάλυση των μεταβάσεων μεταξύ των καταστάσεων:

- Σχηματισμός των τροχιών
    - Προσδιορισμός των μικρο-καταστάσεων ανά τροχιά
    - Αναγνώριση των κύριων γονιδίων ανά τροχιά



Εικόνα 2.1: Σχηματική αναπαράσταση της ροής επεξεργασίας του πακέτου MLscAN

- Δημιουργία των γονιδιακών ρυθμιστικών δικτύων (GRNs) ανά μικρο-κατάσταση της τροχιάς

Έξοδος:

- **Δημιουργία αρχείων πληροφοριών, σχετικά με**
  - τα χαρακτηριστικά των κυττάρων
  - τα χαρακτηριστικά των γονιδίων
  - τα χαρακτηριστικά των τροχιών
  - τα χαρακτηριστικά των γονιδίων σε σχέση με τις τροχιές
  - το επιγενετικό τοπίο
  - τις συνοπτικές πληροφορίες του προτύπου MLscAN
- **Δημιουργία καταλόγων αρχείων διαγραμμάτων, σχετικά με**
  - τα δεδομένα έκφρασης
  - τα αποτελέσματα μείωσης της διαστατικότητας
  - τα γενικά στοιχεία του προτύπου MLscAN
  - τις κυτταρικές καταστάσεις
  - τους κυτταρικούς τύπους
  - τις τροχιές

2.2 Περιγραφή της ροής επεξεργασίας δεδομένων

2.2.1 Δεδομένα εισόδου (input)

Η μόνη απαιτούμενη είσοδος, είναι τα προ-επεξεργασμένα δεδομένα έκφρασης μονήρων κυττάρων (κύτταρα x γονίδια) από την αλληλούχηση (sequencing) του RNA. Συνεπώς, οποιοσδήποτε μετασχηματισμός ή η αφαίρεση γονιδίων και κυττάρων, θα πρέπει να έχει πραγματοποιηθεί πριν τη δημιουργία του προτύπου MLscAN.

Ταυτόχρονα, για καθένα από τα βήματα της ροής επεξεργασίας, μπορούν να οριστούν μία σειρά παραμέτρων ή να δοθούν ήδη διαθέσιμα αποτελέσματα (π.χ. μείωσης της διαστατικότητας), προκειμένου να προσαρμοστεί στις απαιτήσεις κάθε περίπτωσης ή να ελεγχθεί ο τρόπος που επηρεάζουν τα αποτελέσματα. Παρακάτω, θα αναφερθούν, ανά στάδιο, οι μέθοδοι κι οι προαιρετικές παράμετροι με τις προ-επιλεγμένες τιμές.

2.2.2 Μείωση της διαστατικότητας

Επειδή συνήθως ο πίνακας έκφρασης αποτελείται από χιλιάδες γονίδια και χιλιάδες κύτταρα, και μάλιστα με πολλά μηδενικά (αραιός πίνακας (sparse matrix)), είναι σημαντικό για τη μείωση της υπολογιστικής πολυπλοκότητας, να μειωθεί η διαστατικότητα, διατηρώντας, όμως, ταυτόχρονα το μεγαλύτερο δυνατό μέρος χρήσιμης πληροφορίας. Επιπλέον, αρκετές φορές είναι μικρός ο αριθμός των κυττάρων για το πλήθος των γονιδίων – συνθήκη που επηρεάζει τη δημιουργία των καταστάσεων, χρησιμοποιώντας για την ομαδοποίηση συγκριτικά λίγων παρατηρήσεων, πολλά χαρακτηριστικά. Τα αποτελέσματα στα επόμενα βήματα, είναι δυνατό να διαφέρουν σημαντικά ανάλογα με την επιλεγένη μέθοδο.

2.2.2.1 Μέθοδος & εναλλακτικές επιλογές

Η προ-επιλεγμένη μέθοδος μείωσης της διαστατικότητας, είναι η ΑΚΣ: Ανάλυση Κύριων Συνιστωσών (PCA: Principal Component Analysis), με εφαρμογή τυποποίησης στα δεδομένα έκφρασης. Στηρίζεται στην εύρεση των «πραγματικών» κι όχι των σχετικών αποστάσεων μεταξύ των δειγμάτων (για αυτό, είναι μειονέκτημα η παρουσία πολλών μεμονωμένων δειγμάτων – κυττάρων), αποκαλύπτει ακραίες ομάδες κυττάρων, ενώ αποτελεί το αρχικό βήμα άλλων μεθόδων. Ωστόσο, δεν είναι η καταλληλότερη επιλογή για πολύπλοκες δομές και με μη-γραμμικές σχέσεις.

Εναλλακτικά, μπορούν να δοθούν προϋπάρχοντα αποτελέσματα, από οποιαδήποτε άλλη μέθοδο.

Η PCA καθώς και κάποιες από τις κυριότερες επιλογές μείωσης της διαστατικότητας όταν χρησιμοποιούνται δεδομένα αλληλούχησης RNA μονήρων κυττάρων, αναφέρονται συνοπτικά στις επόμενες υπο-ενότητες.

2.2.2.1.1 Ανάλυση κύριων συνιστωσών (PCA: Principal Components Analysis) [19]

Είναι μία ευρέως χρησιμοποιούμενη γραμμική μέθοδος και σχετικά γρήγορη. Βασίζεται στην υπόθεση ότι η κατανομή των δεδομένων είναι κατά προσέγγιση κανονική. Έχοντας την ιδιότητα της διατήρησης τόσο των τοπικών όσο και των καθολικών αποστάσεων, επιτρέπει τη σύγκριση μεταξύ των συστάδων και διευκολύνει τον εντοπισμό των ακραίων.

Τα δεδομένα μετασχηματίζονται προβάλλοντάς τα σε νέο σύστημα γραμμικά ασυσχέτιστων κύριων συνιστωσών, με κριτήριο τη διακύμανση· η πρώτη κύρια συνιστώσα, αντιστοιχεί στη μέγιστη διακύμανση και κάθε επόμενη, επιλέγεται με το ίδιο κριτήριο δεδομένου ότι είναι κάθετη στις προηγούμενες.

Ωστόσο, η παρουσία μηδενικών είναι αυξημένη στα δεδομένα αλληλούχησης μονήρων κυττάρων. Έτσι, έχουν αναπτυχθεί τροποποιήσεις της PCA, όπως η ZIFA (Zero-Inflated Factor Analysis – ανάλυση παραγόντων παρουσία πολλών μηδενικών) [20] κι η ZINB-WaVE (Zero-Inflated Negative Binomial-based Wanted Variation Extraction – εξαγωγή επιθυμητής διακύμανσης βασισμένη σε αρνητική διωνυμική κατανομή παρουσία πολλών μηδενικών) [21], που ενσωματώνουν αυτήν την πληροφορία, αλλά, απαιτούν τη χρήση του πίνακα έκφρασης χωρίς καμία προ-επεξεργασία.

2.2.2.1.2 t-SNE (t-distributed Stochastic Neighbor Embedding) [22]

Πρόκειται για στοχαστική μέθοδο – κάτι που σημαίνει ότι όταν επαναλαμβάνεται η διαδικασία, τα αποτελέσματα θα είναι διαφορετικά, αν κι αναμένεται να μην είναι μεγάλες οι διαφορές. Έχει την ικανότητα να χειριστεί μη-γραμμικές σχέσεις ενώ διατηρεί τις τοπικές σχέσεις, οι οποίες θεωρούνται γραμμικές στην πολλαπλότητα (manifold). Έτσι, δεν επιτρέπει την εκτίμηση της απόστασης μεταξύ των συστάδων. Ακόμη, η συνάρτηση κόστους εμφανίζει αρκετά τοπικά ελάχιστα, αυξάνοντας τον κίνδυνο σύγκλισης σε κάποιο από αυτά.

Μία ευαίσθητη υπερπαράμετρος (hyperparameter) της μεθόδου, είναι ο βαθμός σύγχυσης τοπικών και καθολικών σχέσεων (perplexity), δηλαδή, η βαρύτητα που πρέπει να δοθεί στις τοπικές σχέσεις σε σύγκριση με τις καθολικές κι ορίζει κατά προσέγγιση τον θεωρούμενο αριθμό γειτόνων κάθε παρατήρησης (συνήθως χρησιμοποιούνται τιμές από το 5 έως το 50). Μάλιστα, με τον τρόπο που εκτιμώνται οι γειτονικές παρατηρήσεις, τα όρια γύρω από κάθε παρατήρηση δεν είναι σταθερά. Αρχικά, στον χώρο υψηλής διαστατικότητας, για κάθε παρατήρηση, υπολογίζεται η πιθανότητα θεώρησης οποιασδήποτε άλλης ως γειτονικής, με κατανομή ανάλογη της κανονικής και κέντρο την παρατήρηση αυτήν. Η τιμή της διασποράς σε αραιές περιοχές είναι αυξημένη ενώ σε πυκνές περιοχές μικρότερη, ενισχύοντας τις τοπικές σχέσεις. Έτσι, αποφεύγεται η συρροή παρατηρήσεων ενδιάμεσης απόστασης. Στη συνέχεια, γίνεται προσπάθεια να διατηρηθούν σχετικά σταθερές οι τοπικές αποστάσεις, χρησιμοποιώντας στον χώρο μειωμένης διαστατικότητας, την κατανομή t του Student [23].

Κατά βάση χρησιμοποιείται για την οπτικοποίηση σε 2 ή 3 διαστάσεις κι είναι σχετικά αργή.

2.2.2.1.3 Προσέγγιση και προβολή ομοιόμορφης πολλαπλότητας (UMAP: Uniform Manifold Approximation and Projection) [24]

Αυτή η μέθοδος είναι παρόμοια με την t-SNE, αλλά, έχει τα πλεονεκτήματα της ταχύτερης εκτέλεσης, της ενσωμάτωσης της καθολικής δομής σε υψηλότερο βαθμό, της καλύτερης διατήρησης της συνέχειας των συστάδων και της ικανότητας διάκρισης ακόμη και κυτταρικών πληθυσμών που διαφέρουν ελάχιστα [25].

Βασίζεται στη μάθηση της πολλαπλότητας. Αρχικά, δημιουργείται μία ασαφής τοπολογική αναπαράσταση, η οποία στη συνέχεια βελτιστοποιείται στον χώρο μειωμένης διαστατικότητας, προσπαθώντας να διατηρηθεί κατά το δυνατό, χρησιμοποιώντας ως μέτρο τη διεντροπία (cross-entropy).

2.2.2.1.4 Τοπικά-γραμμική εμβύθιση (LLE: Locally-Linear Embedding) [26]

Η τοπικά-γραμμική εμβύθιση (locally-linear embedding), αποτελεί επίσης μία μη-γραμμική μέθοδο. Ένα μειονέκτημά της είναι ότι δεν ανταποκρίνεται καλά όταν οι πυκνότητες των περιοχών δεν είναι ομοιόμορφες. Κατά μία έννοια, μοιάζει με την εκτέλεση μίας σειράς τοπικών PCA που συγκρίνονται καθολικά για την εύρεση της καλύτερης μη-γραμμικής εμβύθισης.

Αρχικά, για κάθε παρατήρηση, επιλέγεται το σύνολο των κοντινότερων γειτόνων της. Στη συνέχεια, υπολογίζεται ένα διάνυσμα βαρών για κάθε παρατήρηση, θεωρώντας τη γραμμικό συνδυασμό των γειτόνων της, με την επίλυση μίας γραμμικής εξίσωσης. Τελικά, υπολογίζονται οι συντεταγμένες της στον χώρο μειωμένης διαστατικότητας, χωρίς να προσαρμόζονται τα βάρη για τη μείωση της τιμής της συνάρτησης κόστους (χωρίς τον κίνδυνο τοπικών ελαχίστων), αλλά, οι συντεταγμένες. Το βήμα αυτό, πραγματοποιείται με μερική αποσύνθεση (partial decomposition) των ιδιοτιμών (eigenvalues) και διατήρηση ενός αριθμού ιδιοδιανυσμάτων (eigenvectors) που αντιστοιχούν στις μικρότερες μη-μηδενικές ιδιοτιμές.

2.2.2.1.5 Χάρτης διάχυσης (Diffusion map) [27]

Ο χάρτης διάχυσης, είναι μια μη-γραμμική μέθοδος, που βασίζεται στην υπόθεση ότι τα δεδομένα βρίσκονται σε πολλαπλότητα δεδομένου αριθμού διαστάσεων – μικρότερου του αρχικού, διατηρώντας τις τοπικές και καθολικές σχέσεις. Καλύτερα αποτελέσματα προκύπτουν όταν είναι διαθέσιμος μεγάλος αριθμός κυττάρων [28].

Αρχικά, μετασχηματίζεται ο πίνακας αποστάσεων μεταξύ των παρατηρήσεων, χρησιμοποιώντας έναν συμμετρικό πυρήνα (συχνά Gauss). Στη συνέχεια, ο πίνακας αυτός κανονικοποιείται για τον υπολογισμό του πίνακα τυχαίων περιπάτων (random walks) Markov, από τον οποίο προκύπτουν τα ιδιοδιανύσματα (κανονικοποιημένα με την αντίστοιχη ιδιοτιμή) που διατηρούνται για τον σχηματισμό του χώρου μειωμένης διαστατικότητας, αφαιρώντας αυτά με μικρές ιδιοτιμές.

Οι παρατηρήσεις, θεωρούνται κορυφές ενός συμμετρικού γράφου, με τα βάρη των ακμών να είναι αυτά του μετασχηματισμένου πίνακα αποστάσεων. Η απόσταση διάχυσης (diffusion distance), περιγράφει τη συνεκτικότητα (connectivity) του γράφου μεταξύ κάθε ζεύγους κορυφών, που χαρακτηρίζεται από την πιθανότητα μετάβασης μεταξύ τους.

Έτσι, είναι χρήσιμη όταν μελετώνται συνεχείς μεταβολές των κυττάρων [29].

2.2.3 Εκτίμηση των εκ των υστέρων πιθανοτήτων (a-posteriori probabilities)

2.2.3.1 Μέθοδος & εναλλακτικές επιλογές

Η προ-επιλεγμένη μέθοδος υπολογισμού των εκ των υστέρων πιθανοτήτων, είναι το πρότυπο μείξης κανονικών κατανομών [30].

Εκτός της προ-επιλεγμένης μεθόδου, μπορούν να δοθούν προϋπάρχοντα αποτελέσματα ή συνάρτηση που εκτελεί την επιθυμητή μέθοδο υπολογισμού των εκ των υστέρων πιθανοτήτων κι επιστρέφει τα αποτελέσματα.

2.2.3.1.1 Πρότυπο μείξης κανονικών κατανομών (GMM: Gaussian Mixture Model)

Για την εκτίμηση των εκ των υστέρων πιθανοτήτων, χρησιμοποιείται πρότυπο μείξης κανονικών κατανομών, υποθέτοντας πως οι συστάδες είναι σφαιρικές, ώστε να μην

υπάρχει επικάλυψη, κι ο όγκος τους άνισος, μιας και δεν υπάρχει λόγος να αναμένονται μόνο ίσου όγκου καταστάσεις, χρησιμοποιώντας συγκεκριμένο αριθμό διαστάσεων των αποτελεσμάτων μείωσης της διαστατικότητας των δεδομένων έκφρασης.

Για τον υπολογισμό των προτύπων, χρησιμοποιείται το πακέτο *mclust* (5.4.5) [31], που βασίζεται στον αλγόριθμο μεγιστοποίησης της προσδοκίας (Expectation-Maximization) [32] με αρχικοποίηση μέσω συσσωρευτικής ιεραρχικής (agglomerative hierarchical) συσταδοποίησης [33].

Τα αποτελέσματα είναι δυνατό να διαφέρουν σημαντικά ανάλογα με τον επιλεγμένο αριθμό διαστάσεων.

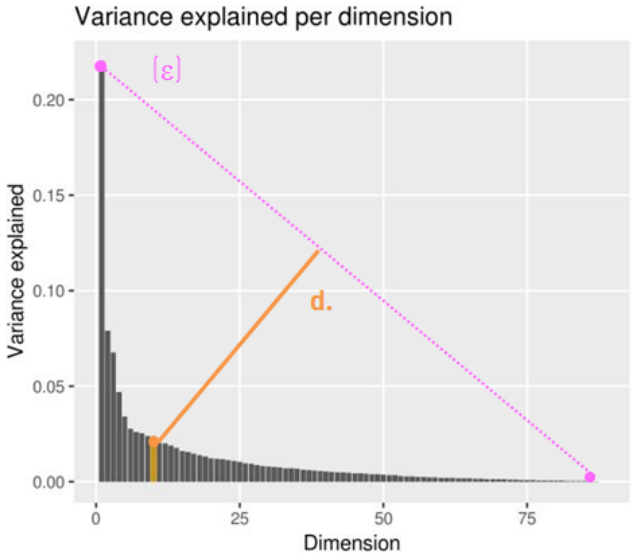
2.2.3.2 Επιλογή του αριθμού διαστάσεων των δεδομένων

Η προ-επιλεγμένη μέθοδος επιλογής του αριθμού των διαστάσεων, είναι, το σημείο γονάτου της διακύμανσης ανά διάσταση.

Εναλλακτικά, είναι δυνατό να προσδιοριστεί απευθείας ο αριθμός, να οριστεί το ελάχιστο ποσοστό της αθροιστικής διακύμανσης, να χρησιμοποιηθεί το κριτήριο Kaiser [34] ή να δοθεί συνάρτηση που εκτελεί την επιθυμητή μέθοδο κι επιστρέφει το αποτέλεσμα.

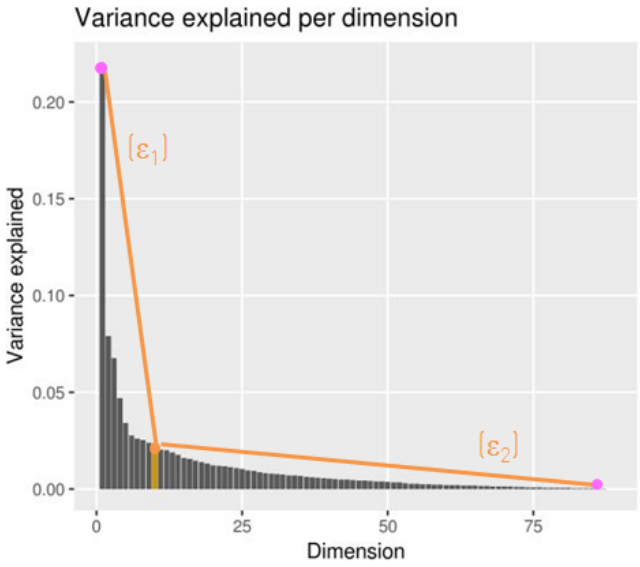
2.2.3.2.1 Σημείο γονάτου (knee-point) της διακύμανσης ανά διάσταση

Είναι διαθέσιμες δύο επιλογές για τον υπολογισμό του σημείου γονάτου· μία πιο αργή και μία πιο γρήγορη (προ-επιλογή).



Εικόνα 2.2: Απεικόνιση του τρόπου εύρεσης του σημείου γονάτου με την πιο γρήγορη μέθοδο (όπως περιγράφεται στην ενότητα 2.2.3.2.1)

Στην πιο γρήγορη μέθοδο, το σημείο γονάτου, είναι αυτό όπου η κάθετη απόσταση (*d.*) από την ευθεία (*ε*) που σχηματίζεται μεταξύ του πρώτου και του τελευταίου σημείων – διαστάσεων (εν σειρά), είναι μέγιστη.



Εικόνα 2.3: Απεικόνιση του τρόπου εύρεσης του σημείου γονάτου με την πιο αργή μέθοδο (όπως περιγράφεται στην ενότητα 2.2.3.2.1)

Στην πιο αργή μέθοδο [35], προσαρμόζονται δύο ευθείες (ε_1, ε_2), με αρχή κάθε σημείο – συνιστώσα (εκτός των δύο ακραίων), μία δεξία και μία αριστερά, κι υπολογίζεται το σφάλμα που προκύπτει από το άθροισμα των απόλυτων τιμών των διαφορών των πραγματικών τιμών από τις εκτιμώμενες – σημεία των ευθειών. Το σημείο γονάτου, είναι αυτό όπου το υπολογιζόμενο σφάλμα είναι ελάχιστο.

2.2.3.2.2 Αθροιστική διακύμανση

Συνήθως, δεν είναι δυνατό να προκύψει υψηλό ποσοστό αθροιστικής διακύμανσης (> 50%) με σχετικά μικρό αριθμό διαστάσεων, οπότε, θα πρέπει αυτή η παράμετρος να χρησιμοποιείται με προσοχή. Αν επιλεγεί πολύ μεγάλος αριθμός διαστάσεων, αναμένεται να υπάρξει σημαντική χρονική επιβάρυνση στη δημιουργία του προτύπου ή κι αδυναμία δημιουργίας του προτύπου (π.χ. επειδή προκύπτει μη-αντιστρέψιμος πίνακας).

2.2.3.2.3 Κριτήριο του Kaiser [34]

Σύμφωνα με το κριτήριο του Kaiser, διατηρούνται μόνο όσες συνιστώσες έχουν ιδιοτιμή > 1. Αυτό, είναι επίσης ένα κριτήριο που πρέπει να χρσιμοποιηθεί με προσοχή, καθώς έχει παρατηρηθεί η τάση να οδηγεί σε επιλογή σχετικά μεγάλου αριθμού διαστάσεων.

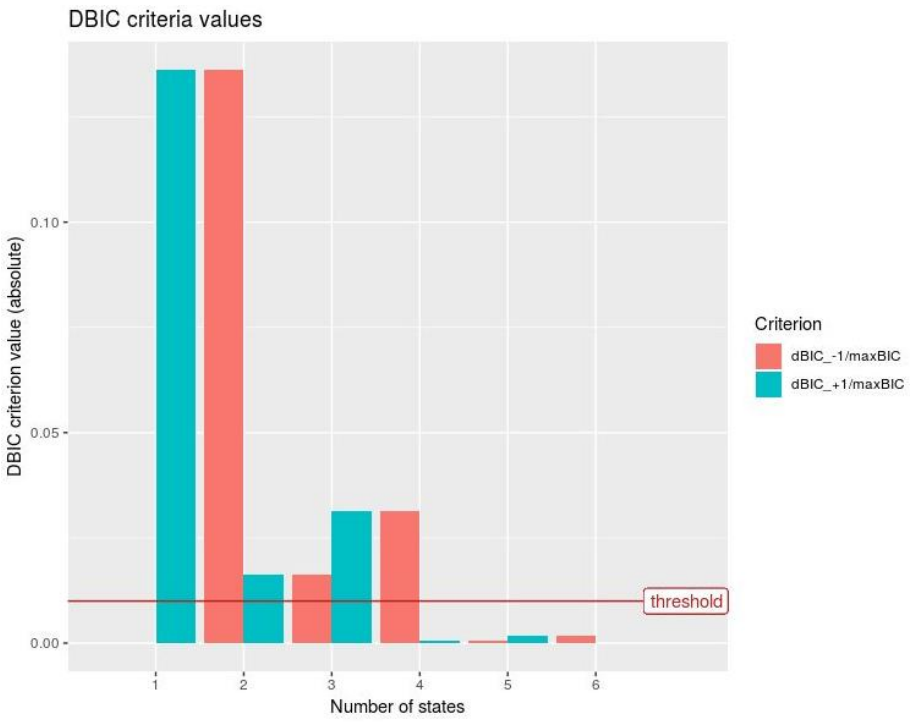
2.2.3.3 Επιλογή του αριθμού των καταστάσεων

Η προ-επιλεγμένη μέθοδος προσδιορισμού τού αριθμού των καταστάσεων, είναι, τα κριτήρια ΔBIC, που ορίζονται παρακάτω.

Εναλλακτικά, είναι δυνατό να προσδιοριστεί απευθείας ο αριθμός ή να δοθεί συνάρτηση που εκτελεί την επιθυμητή μέθοδο κι επιστρέφει το αποτέλεσμα.

2.2.3.3.1 Κριτήρια ΔBIC

Διαθέτοντας τις τιμές του BIC (Bayesian Information Criterion) [36] για ένα εύρος αριθμού κυτταρικών καταστάσεων (δηλαδή, ομάδων στο πρότυπο μείξης κανονικών



Εικόνα 2.4: Τιμές των κριτηρίων ΔBIC_1 (2.1) και ΔBIC_2 (2.2) για ένα πρότυπο μείξης κανονικών κατανομών με μία έως και 4 καταστάσεις. Η κόκκινη γραμμή, επισημαίνει την τελική τιμή του ορίου των κριτηρίων ΔBIC για την επιλογή αριθμού καταστάσεων.

κατανομών), επιλέγεται, αρχικά, ο μικρότερος αριθμός καταστάσεων, *x*, που ικανοποιεί και τα δύο κριτήρια, ΔBIC_1 και ΔBIC_2.

$$Κριτήριο\ \Delta BIC_1: \left|\frac{BIC_x - BIC_{x-1}}{BIC_{max.}}\right| < όριο \quad (2.1)$$

$$Κριτήριο\ \Delta BIC_2: \left|\frac{BIC_x - BIC_{x+1}}{BIC_{max.}}\right| < όριο \quad (2.2)$$

Τελικά, επιλέγονται, max{2, (x-1)} κυτταρικές καταστάσεις, προκειμένου να μειωθεί η πολυπλοκότητα. Το όριο που συμμετέχει στα κριτήρια, αρχικά έχει προ-επιλεγμένη τιμή, 0,01. Αν με αυτό το όριο δεν προκύπτει κάποια περίπτωση που να ικανοποιούνται και τα δύο κριτήρια, τότε, αυξάνεται σε κάθε επανάληψη, με την προσθήκη, 0,01. Εναλλακτικά, η τιμή του αρχικού ορίου, μπορεί να οριστεί απευθείας από τον χρήστη, καθώς και να επιλεγεί αν σε κάθε επανάληψη, μέχρι την επιλογή κάποιας περίπτωσης, θα προστίθεται κάποια τιμή στο τρέχον όριο και ποια θα είναι αυτή ή αν θα πολλαπλασιάζεται με κάποιον παράγοντα και ποια θα είναι η τιμή τού παράγοντα αυτού.

Στην εικόνα 2.4, φαίνεται ένα παράδειγμα των τιμών ΔBIC_1 και ΔBIC_2 που προέκυψαν για ένα εύρος αριθμού κυτταρικού καταστάσεων, από 1 έως και 6. Επειδή στα δύο άκρα, δεν μπορούν να υπολογιστούν και τα δύο κριτήρια, αυτές οι περιπτώσεις, δε συμμετέχουν στην επιλογή. Από τις υπόλοιπες, με την προ-επιλεγμένη μέθοδο, ο ελάχιστος (και μοναδικός) αριθμός κυτταρικών καταστάσεων που ικανοποιεί τα κριτήρια, είναι, 5. Συνεπώς, τελικά, θα προκύψει πρότυπο με: max{2, (5 - 1)} = 4 κυτταρικές καταστάσεις.

2.2.4 Καταστάσεις

2.2.4.1 Ορισμός

Μία κυτταρική κατάσταση ή απλώς κατάσταση, συνίσταται από το σύνολο των κυττάρων που η μέγιστη εκ των υστέρων πιθανότητά τους αντιστοιχεί στη δεδομένη συστάδα. Έτσι, κάθε κύτταρο, μπορεί να ανήκει σε μία μόνο κατάσταση.

Επιπλέον, μετά τη δημιουργία του προτύπου MLscAN, με τη σχετική μέθοδο ανάθεσης, ο χρήστης μπορεί να ονομάσει με τον επιθυμητό τρόπο τις καταστάσεις. Ο προκαθορισμένος τρόπος ονομασίας των καταστάσεων, είναι: «1», «2», κ.ο.κ.. Επιπλέον, είναι δυνατό να ονομαστούν αυτόματα με κάποιον από τους ακόλουθους τρόπους, χρησιμοποιώντας τους κυτταρικούς τύπους (χαρακτηριστικό, «cellType»), αν αυτή η πληροφορία είναι διαθέσιμη:

α) με την επιλογή, «mostFreqPerType», αν σε μία κατάσταση, καταλήγουν τα περισσότερα κύτταρα ενός τύπου σε σχέση με οποιαδήποτε άλλη κατάσταση, τότε, αυτή παίρνει το όνομα τύπου. Αν σε μία κατάσταση, καταλήγουν τα περισσότερα κύτταρα περισσότερων του ενός τύπων, τότε, αυτή θεωρείται μεικτή κι ονομάζεται «Mi», όπου i είναι ο αριθμός της κατάστασης στη διάταξη των μεικτών καταστάσεων με φθίνοντα τρόπο βάσει του αριθμού των τύπων ή απλώς, «M», αν είναι μόνο μία. Τέλος, οι υπόλοιπες καταστάσεις, ονομάζονται, «1», «2», κ.ο.κ..

β) με την επιλογή, «mostFreqPerState», βάσει της οποίας, χρησιμοποιείται το όνομα ενός τύπου για την ονομασία κάθε κατάστασης που τα κύτταρά του αποτελούν τουλάχιστον το 70% όλων των κυττάρων της. Εφόσον αυτό ισχύει για κάποιον τύπο, έστω A, για περισσότερες από μία καταστάσεις, προστίθεται το επίθημα της θέσης της κατάστασης στη διάταξή τους βάσει του ποσοστού των κυττάρων που είναι τύπου A, ξεκινώντας από αυτήν με το μεγαλύτερο ποσοστό. Για όσες καταστάσεις δεν είναι δυνατό να χρησιμοποιηθεί ένας τύπος, ακολουθείται η προκαθορισμένη ονομασία, «1», «2», κ.ο.κ.. Με αυτόν τον τρόπο, είναι πιο εύκολο να σχετιστούν οι καταστάσεις με τους τύπους για τον εντοπισμό πιθανών υπο-πληθυσμών και τη διάκριση των μεταβάσεων στις οποίες συμμετέχουν.

2.2.4.2 Ακραίοι υπο-πληθυσμοί καταστάσεων

Μετά τον σχηματισμό των καταστάσεων, κι εφόσον οι επιλεγμένες παράμετροι το επιτρέπουν, είναι δυνατό να αναγνωριστούν υπο-πληθυσμοί καταστάσεων που θεωρούνται ακραίοι και στη συνέχεια απομακρύνονται από τα επόμενα βήματα της ροής επεξεργασίας.

Εκτός της μεθόδου που αναφέρεται παρακάτω, μπορεί να δοθεί συνάρτηση που εκτελεί την επιθυμητή μέθοδο κι επιστρέφει το αποτέλεσμα.

Αν οι παράμετροι που αφορούν στον επιτρεπόμενο αριθμό ακραίων υπο-πληθυσμών και στο επιτρεπόμενο ποσοστό τους στο σύνολο των κυττάρων, επιτρέπουν την αφαίρεση ακραίων κυττάρων, τότε, επαναλαμβάνεται η δημιουργία του προτύπου μείξης κανονικών κατανομών, αφαιρώντας από τον προηγουμένως επιλεγμένο αριθμό καταστάσεων το πλήθος των ακραίων υπο-πληθυσμών που απομακρύνθηκαν.

Επιλέον, αν δίνονται απευθείας οι εκ των υστέρων πιθανότητες κι ο αριθμός των κυττάρων είναι μικρότερος αυτού τού πίνακα έκφρασης, τότε, θεωρείται πως τα κύτταρα που δεν περιλαμβάνονταν, είναι ακραία.

2.2.4.2.1 Μέθοδος

Αρχικά, σχηματίζονται οι υπο-πληθυσμοί κάθε κατάστασης. Τα κύτταρα που ανήκουν σε μία κατάσταση, ανήκουν και στον υπο-πληθυσμό της, εφόσον η τιμή της εκ των υστέρων πιθανότητας για αυτήν, είναι μεγαλύτερη του ορίου (προ-επιλογή: 0,5).

Στη συνέχεια, υπολογίζεται η τυπική απόκλιση ανά υπο-πληθυσμό και για το σύνολο των κυττάρων, ανά συνιστώσα των αποτελεσμάτων μειωμένης διαστατικότητας που χρησιμοποιήθηκε για να δημιουργηθεί το πρότυπο μείξης κανονικών κατανομών.

Αν για κάποιον υποπληθυσμό, η τυπική απόκλιση σε κάθε συνιστώσα είναι μεγαλύτερη της τυπικής απόκλισης του συνόλου των κυττάρων, τότε, θεωρείται πως αυτός ο υπο-πληθυσμός αποτελείται από ακραία κύτταρα.

Ανάλογα με τον επιτρεπόμενο αριθμό ακραίων υπο-πληθυσμών (προ-επιλογή: Inf) καθώς και τον επιτρεπόμενο λόγο ακραίων κυττάρων (προ-επιλογή: 0,1), αποφασίζεται αν τελικά θα αφαιρεθούν κάποιοι ακραίοι υπο-πληθυσμοί ή όχι. Αν δεν μπορούν να αφαιρεθούν όλοι, τότε, διατάσσονται με βάση το άθροισμα των τυπικών αποκλίσεων με φθίνοντα τρόπο, κι απομακρύνονται όσοι είναι δυνατό. Ωστόσο, αν ο λόγος των κυττάρων τους, ξεπερνά το όριο, αφαιρείται ο μέγιστος αριθμός ακραίων υπο-πληθυσμών που το άθροισμα των κυττάρων τους δεν ξεπερνά το όριο αυτό. Έτσι, είτε θα αφαιρεθούν όλα τα κύτταρα ενός ακραίου υπο-πληθυσμού είτε κανένα.

Παρόλαυτα, αποθηκεύονται στο σύνολό τους τα θεωρούμενα ακραία κύτταρα κι οι καταστάσεις με αυτούς τους υπο-πληθυσμούς, για διερεύνηση μετά τη δημιουργία του προτύπου.

2.2.5 Τροχιές

2.2.5.1 Ορισμός

Μία τροχιά, σχηματίζεται μεταξύ ενός ζεύγους καταστάσεων· της κατάστασης έναρξης (ground state) και της κατάστασης προορισμού (landing state), ονομαζόμενη: «groundState-to-landingState» [18].

2.2.5.2 Μέθοδος

Η τροχιά, αποτελείται από το σύνολο των κυττάρων που οι δύο μέγιστες εκ των υστέρων πιθανότητές τους, ανήκουν στις καταστάσεις έναρξης και προορισμού. Προκειμένου να θεωρηθεί ότι υπάρχει τροχιά (είναι έγκυρη), θα πρέπει ο αριθμός αυτών των κυττάρων να είναι τουλάχιστον ίσος με 6, προκειμένου να μην αποκλείει σε μία ακραία περίπτωση, με πολύ μικρό αριθμό κυττάρων, να σχηματιστούν 3 μικρο-καταστάσεις (αναφέρονται παρακάτω) και ταυτόχρονα, τουλάχιστον 3 κύτταρα να ανήκουν σε κάθε κατάσταση.

Τα κύτταρα της τροχιάς, διατάσσονται με φθίνοντα τρόπο, βάσει της εκ των υστέρων πιθανότητας για την κατάσταση έναρξης.

Μία παράμετρος που μπορεί να επηρεάσει τον αριθμό των κυττάρων της τροχιάς, μετά την προηγούμενη διαδικασία, είναι η απαιτούμενη μονοτονία των εκ των υστέρων πιθανοτήτων της κατάστασης προορισμού (προ-επιλογή: καμία απαίτηση). Δηλαδή, μπορεί να απαιτείται να αυξάνεται μονότονα ή αυστηρά μονότονα η εκ των υστέρων πιθανότητα για την κατάσταση προορισμού καθώς αυξάνεται η θέση διάταξης των κυττάρων στην τροχιά.

Έτσι, ένα κύτταρο, μπορεί να ανήκει το πολύ σε δύο τροχιές, χωρίς να ανήκει αναγκαστικά σε τουλάχιστον μία (αν έχει πιθανότητα, 1, για κάποια κατάσταση ή δεν μπορεί να σχηματιστεί τροχιά μεταξύ των καταστάσεων που αντιστοιχούν στις μέγιστες εκ των υστέρων πιθανότητές του ή αποκλείονται λόγω των απαιτήσεων στη μονοτονία των εκ των υστέρων πιθανοτήτων για την κατάσταση προορισμού).

2.2.6 Μεταβάσεις

Μία μετάβαση, σχηματίζεται μεταξύ ενός ζεύγους καταστάσεων· της κατάστασης έναρξης και της κατάστασης προορισμού, από το σύνολο των κυττάρων των οποίων οι δύο μέγιστες εκ των υστέρων πιθανότητες αντιστοιχούν στις καταστάσεις αυτές. Συνεπώς, ένα κύτταρο μπορεί να ανήκει το πολύ σε ένα ζεύγος μεταβάσεων.

Διαφέρουν από τις τροχιές, επειδή απουσιάζουν οι περιορισμοί που αναφέρονται στην ενότητα 2.2.5.1. Αρκεί, δηλαδή, η ύπαρξη ακόμη κι ενός κυττάρου στη μετάβαση.

Για κάθε κατάσταση, ορίζεται η τάση μετάβασης (transition propensity) των κυττάρων της σε μία άλλη κατάσταση – κατάσταση προορισμού, ως ο λόγος των κυττάρων της των οποίων η δεύτερη μεγαλύτερη εκ των υστέρων πιθανότητα αντιστοιχεί στη δεδομένη κατάσταση προορισμού. Συνεπώς, το άθροισμα αυτών των λόγων πτοσθέτοντας και τον λόγο των κυττάρων της κατάστασης που έχουν μηδενική εκ των υστέρων πιθανότητα για οποιαδήποτε άλλη κατάσταση, ισούται με 1.

2.2.7 Μικρο-καταστάσεις

2.2.7.1 Ορισμός

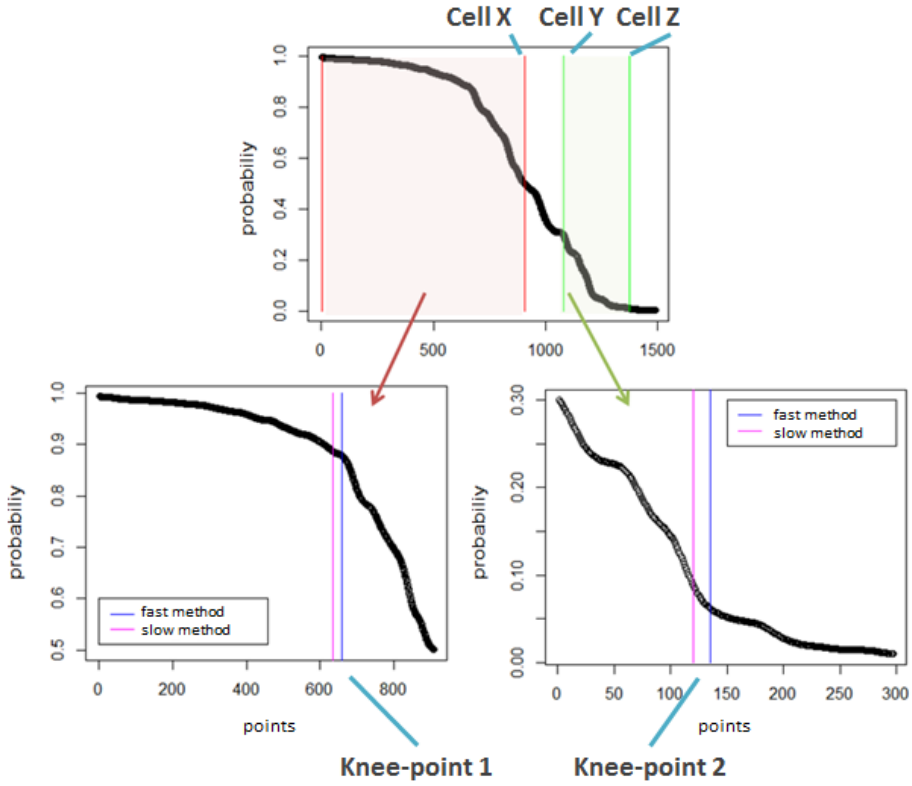
Οι μικρο-καταστάσεις, αποτελούν διαδοχικά υπο-σύνολα των διατεταγμένων κυττάρων της τροχιάς. Τυπικά, γίνεται προσπάθεια να οριστούν 3 μικρο-καταστάσεις, οι οποίες εν σειρά, είναι οι: μικρο-κατάσταση έναρξης (ground m-state), μικρο-κατάσταση μετάβασης (transitional m-state), μικρο-κατάσταση προορισμού (landing m-state) [18].

Εκτός της μεθόδου που αναφέρεται παρακάτω, είναι δυνατό να δοθεί συνάρτηση που εκτελεί την επιθυμητή μέθοδο κι επιστρέφει τα αποτελέσματα ή μετά τη δημιουργία του προτύπου MLscAN να οριστούν απευθείας τα δύο σημεία για τον διαχωρισμό των μικρο-καταστάσεων.

2.2.7.2 Μέθοδος

Αρχικά, από τα διατεταγμένα κύτταρα της τροχιάς, παραγάγονται δεκαπλάσια σημεία, χρησιμοποιώντας κυβική σφηνοειδή συνάρτηση (cubic spline function) [37] και το φίλτρο Hyman [38], για την εξασφάλιση μονοτονίας στα αποτελέσματα που προκύπτουν, μιας και τα κύτταρα διατάσσονται με τρόπο που εξασφαλίζει τη μονοτονία αυτή. Αυτό γίνεται επειδή προκύπτουν περιπτώσεις με λίγα κύτταρα ή / και μη-ομαλές μεταβολές στην τιμή των πιθανοτήτων.

Στη συνέχεια, προσδιορίζονται οι δύο περιοχές στις οποίες θα αναζητηθούν τα σημεία γονάτου (με τον τρόπο που περιγράφηκε στην ενότητα 2.2.3.2.1), τα οποία, θα προσαρμοστούν στην αρχική κλίμακα – κύτταρα της τροχιάς, για να οριστούν τελικά τα όρια κάθε μικρο-κατάστασης.



Εικόνα 2.5: Παράδειγμα επιλογής των περιοχών της καμπύλης της πιθανότητας της κατάστασης έναρξης των κυττάρων μίας τροχιάς (με κόκκινο χρώμα και με πράσινο χρώμα) για την αναζήτηση των σημείων γονάτων προκειμένου να οριστούν οι τρεις μικρο-καταστάσεις. Στη συνέχεια, τα σημεία των γονάτων, επιλέγονται με τη γρηγορότερη μέθοδο και την πιο αργή.

Οι περιοχές αναζήτησης των σημείων γονάτου, ορίζονται χρησιμοποιώντας τρεις τιμές εκ των υστέρων πιθανοτήτων της κατάστασης έναρξης (προ-επιλογή: 0,5, 0,3, 0,01). Η πρώτη περιοχή (με κόκκινο χρώμα στην εικόνα 2.5), εκτείνεται από το πρώτο σημείο στη διάταξη έως και το τελευταίο σημείο με πιθανότητα μεγαλύτερη ή ίση της πρώτης επιλεγμένης πιθανότητας. Η δεύτερη περιοχή (με πράσινο χρώμα στην εικόνα 2.5), εκτείνεται από το πρώτο σημείο με τιμή πιθανότητας μικρότερη ή ίση της δεύτερης επιλεγμένης πιθανότητας έως και το πρώτο σημείο με τιμή πιθανότητας μικρότερη ή ίση της τρίτης επιλεγμένης πιθανότητας.

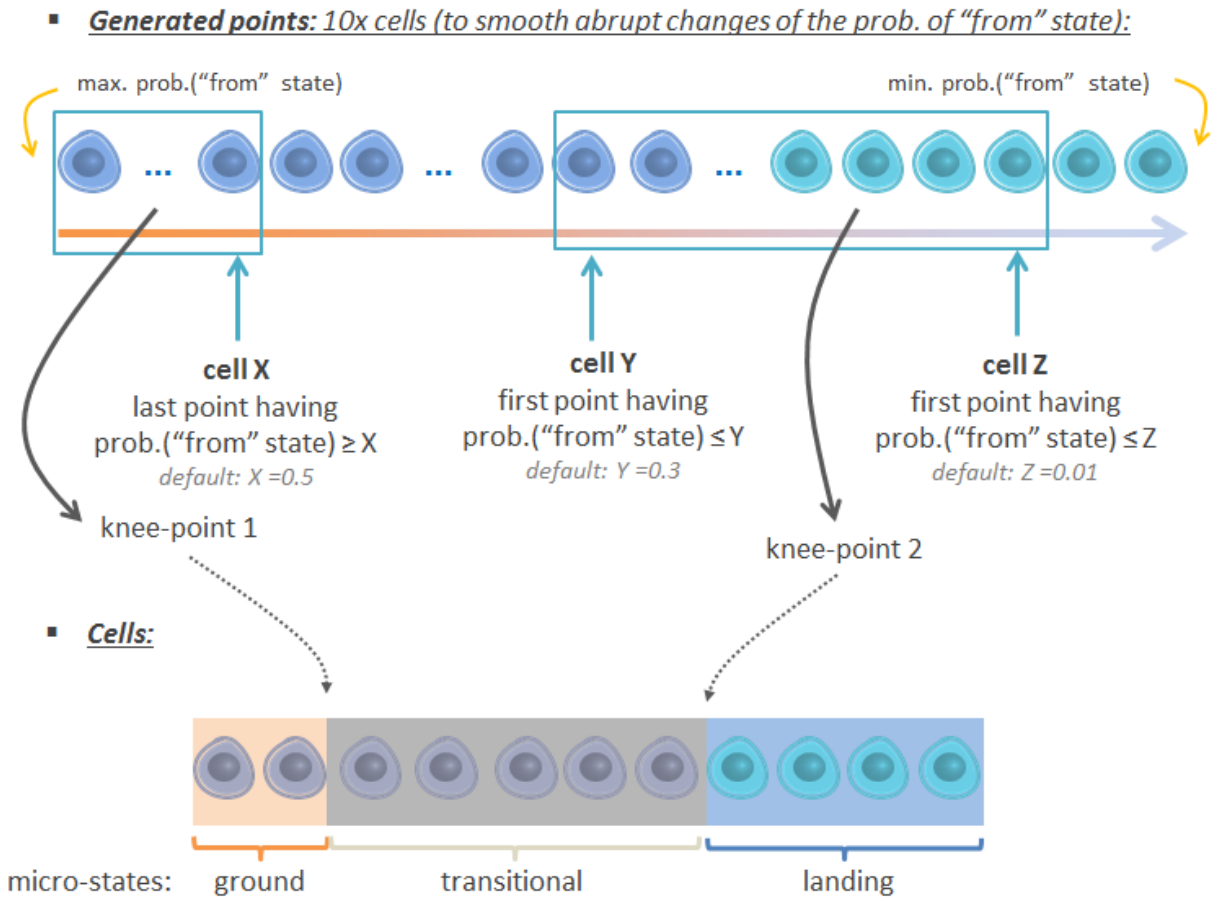
Σε αυτές τις περιοχές, εφόσον δεν υπάρχει επικάλυψη, προσδιορίζονται τα σημεία γονάτου και στη συνέχεια προσαρμόζοναι στην αρχική κλίμακα - αριθμό κυττάρων ώστε αντιστοιχίζονται σε πραγματικά κύτταρα, και διαμερίζουν τα κύτταρα της τροχιάς σε 3 διαδοχικές μικρο-καταστάσεις, ως εξής:

- **μικρο-κατάσταση έναρξης (ground m-state):**

  από το πρώτο κύτταρο έως και το πρώτο σημείο – κύτταρο γονάτου

- **μικρο-κατάσταση μετάβασης (transitional m-state):**

  από το (πρώτο σημείο – κύτταρο γονάτου + 1) έως και το δεύτερο σημείο – κύτταρο γονάτου



Εικόνα 2.6: Σχηματικά, η διαδικασία ορισμού των περιοχών αναζήτησης των σημείων γονάτων και του διαχωρισμού της τροχιάς σε τρεις μικρο-καταστάσεις, μετά την αύξηση των διαθέσιμων σημείων, χρησιμοποιώντας τις επιλεγμένες τιμές για την εκ των υστέρων πιθανότητα της κατάστασης έναρξης.

- **μικρο-κατάσταση προορισμού (landing m-state):**

  από το (δεύτερο σημείο – κύτταρο γονάτου + 1) έως και το τελευταίο κύτταρο

Επιπλέον, για να είναι έγκυρες οι μικρο-καταστάσεις, θα πρέπει καθεμία να έχει τουλάχιστον δύο κύτταρα, μιας και στη συνέχεια θα γίνει προσπάθεια δημιουργίας ενός GRN για αυτή.

Όταν δεν μπορούν να δημιουργηθούν 3 μικρο-καταστάσεις, τότε, ορίζονται 2· η μικρο-κατάσταση έναρξης κι η μικρο-κατάσταση προορισμού. Στη μικρο-κατάσταση έναρξης, περιλαμβάνονται όλα τα κύτταρα της τροχιάς που ανήκουν στην κατάσταση έναρξης και στη μικρο-κατάσταση προορισμού, περιλαμβάνονται όλα τα κύτταρα της τροχιάς που ανήκουν στην κατάσταση προορισμού.

2.2.8 Κύρια γονίδια

Τα κύρια γονίδια ή γονίδια-κλειδιά (key-genes), αποτελούν αυτά τα γονίδια που θεωρούνται πιο σημαντικά για τη δημιουργία των GRNs ανά μικρο-κατάσταση της τροχιάς. Δεν υπάρχει περιορισμός ως προς τον αριθμό τους, κι ανάλογα με την επιλεγμένη μέθοδο προσδιορισμού, μπορούν να θεωρηθούν όλα σημαντικά, κανένα ή οποιοσδήποτε ενδιάμεσος αριθμός.

Εκτός της μεθόδου που αναφέρεται παρακάτω, κι αναπτύχθηκε από τους συγγραφείς του άρθρου [18], είναι δυνατό να δοθεί συνάρτηση που εκτελεί την επιθυμητή μέθοδο κι επιστρέφει τα αποτελέσματα ή μετά τη δημιουργία του προτύπου MLscAN να οριστούν απευθείας. Η σειρά με την οποία εμφανίζονται / επιστρέφονται τα κύρια γονίδια, θεωρείται πως αντικατοπτρίζει το πόσο σημαντικά είναι.

2.2.8.1 Μέθοδος

Με τα κριτήρια της μεθόδου, που αναφέρονται παρακάτω, στόχος είναι να επιλέγονται γονίδια με διτροπική (bimodal) κατανομή έκφρασης και μάλιστα το επίπεδο έκφρασης να αλλάζει σημαντικά σε επαρκή λόγο των κυττάρων, συγκρίνοντας τα κύτταρα που ανήκουν στη μικρο-κατάσταση έναρξης και στη μικρο-κατάσταση προορισμού.

Για να θεωρηθεί ένα γονίδιο σημαντικό, θα πρέπει να ισχύουν όλες οι παρακάτω συνθήκες:

- **Συνθήκη 1:** η διακύμανση της έκφρασης δεν είναι 0

Μετά τη δημιουργία ενός προτύπου μείξης κανονικών κατανομών με τα κύτταρα της τροχιάς και με δύο μόνο ομάδες (G_0, G_1), που αντιστοιχούν στις ομάδες κυττάρων, με χαμηλή ή μη-υπάρχουσα έκφραση (μέση τιμή: μ_0, τυπική απόκλιση: σ_0, αναλογία: π_0) και με υψηλότερη έκφραση ή απλώς έκφραση (μέση τιμή: μ_1, τυπική απόκλιση: σ_1, αναλογία: π_1), και θα πρέπει να ισχύουν επιλέον:

- **Συνθήκη 2:**

  $BIC_{two\ groups} - BIC_{one\ group} < -kgMinBICDiff$ (2.3)

  (προ-επιλεγμένη τιμή *kgMinBICDiff* = 2)

- **Συνθήκη 3:**

  $|\pi_0 - \pi_1| < kgMaxPropDiff$ (2.4)

  (προ-επιλεγμένη τιμή *kgMaxPropDiff* = 0,4)

- **Συνθήκη 4:**

  $(\mu_1 - \mu_0) > kgSTDWeight * (\sigma_1 + \sigma_0)$

  (προ-επιλεγμένη τιμή *kgSTDWeight* = 1)

- **Συνθήκη 5:**

$$\frac{f_{ground\_A,1}}{f_{ground\_B,1}} < kgCriteriaThr \ \ και \ \ \frac{f_{ground\_A,0}}{f_{ground\_A,0}} > \frac{1}{kgCriteriaThr} \quad (2.5)$$

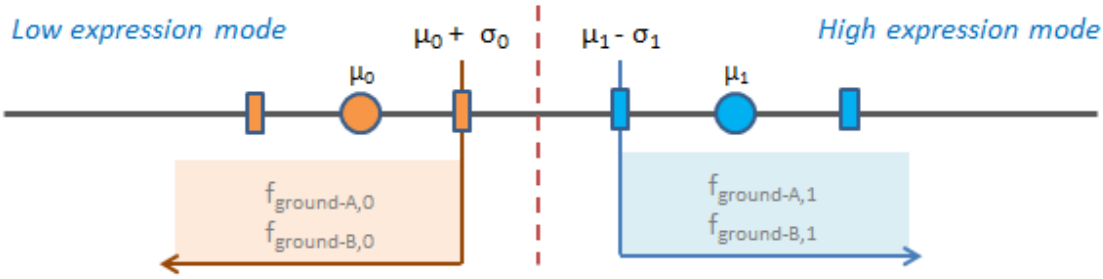
$$\frac{f_{ground\_A,1}}{f_{ground\_B,1}} > \frac{1}{kgCriteriaThr} \ \ και \ \ \frac{f_{ground\_A,0}}{f_{ground\_A,0}} < kgCriteriaThr \quad (2.6)$$

  με:

  - $f_{ground\_A,0}$ = λόγος κυττάρων της μικρο-κατάστασης έναρξης με έκφραση < $\mu_0 + kgSTDWeightCrit * \sigma_0$ (2.7)

- $f_{ground\_B,0}$ = λόγος κυττάρων της μικρο-κατάστασης προορισμού με έκφραση < μ_0 + kgSTDWeightCrit * σ_0 (2.8)
- $f_{ground\_A,1}$ = λόγος κυττάρων της μικρο-κατάστασης έναρξης με έκφραση > μ_1 - kgSTDWeightCrit * σ_1 (2.9)
- $f_{ground\_B,1}$ = λόγος κυττάρων της μικρο-κατάστασης προορισμού με έκφραση > μ_1 - kgSTDWeightCrit * σ_1 (2.10)

(προ-επιλεγμένη τιμή *kgSTDWeightCrit* = 1, προ-επιλεγμένη τιμή *kgCriteriaThr* = 0,5)



Εικόνα 2.7: Περιοχές στις οποίες εντοπίζονται τα κύτταρα μικρο-καταστάσεων έναρξης και προορισμού, με χαμηλή κι υψηλή έκφρασης κάποιου γονιδίου. Είναι οι ομάδες που χρησιμοποιούνται στον έλεγχο της συνθήκης 5 της προκαθορισμένης μεθόδου (2.2.8.1) αναγνώρισης των κύριων γονιδίων μίας τροχιάς.

Μετά την παραπάνω διαδικασία, αποθηκεύονται οι απόλυτες (n_{ij}) και σχετικές συχνότητες (f_{ij}) των κυττάρων των μικρο-καταστάσεων έναρξης και προορισμού (τιμή i = *G* ή *L*, αντίστοιχα), με υψηλή (τιμή j = 1) ή χαμηλή (τιμή j = 0) έκφραση:

n_{G0} = αριθμός κυττάρων της μικρο-κατάστασης έναρξης με έκφραση < μ_0 + kgSTDWeightCrit * σ_0 (2.11)

n_{G1} = αριθμός κυττάρων της μικρο-κατάστασης έναρξης με έκφραση > μ_1 - kgSTDWeightCrit * σ_1 (2.12)

n_{L0}= αριθμός κυττάρων της μικρο-κατάστασης προορισμού με έκφραση < μ_0 + kgSTDWeightCrit * σ_0 (2.13)

n_{L1} = αριθμός κυττάρων της μικρο-κατάστασης προορισμού με έκφραση > μ_1 - kgSTDWeightCrit * σ_1 (2.14)

Τα κύρια γονίδια, ταξινομούνται με φθίνοντα τρόπο βάσει της μέγιστης τιμής των λόγων των κριτηρίων της συνθήκης 5.

Όταν χρησιμοποιείται άλλη μέθοδος, θεωρείται πως η σειρά με την οποία επιστρέφονται, αντικατοπτρίζει τη σημαντικότητά τους.

2.2.8.2 Διαφορικά εκφρασμένα γονίδια

Επιπλέον, είναι διαθέσιμες πέντε επιπλέον μέθοδοι, για την αναγνώριση διαφορικά εκφρασμένων γονιδίων: *Seurat* > *bimod* (συνάρτηση, «FindMarkers», θέτοντας την παράμετρο, «test.use= "bimod"») [39, 40], *Seurat* > δοκιμασία *t* του Student (συνάρτηση, «FindMarkers», θέτοντας την παράμετρο, «test.use= "t"») [23] (συνάρτηση: keyGenesDEt), *Seurat* > *MAST* (συνάρτηση, «FindMarkers», θέτοντας την παράμετρο, «test.use= "MAST"») [41], *switchde* [42] και *edgeR* [43]. Στις περιπτώσεις που είναι διαθέσιμες αυτές οι πληροφορίες, απαιτείται να πληρούνται τα ακόλουθα κριτήρια: τιμή $p \leq 0{,}01$ και $\log_2$(fold-change) ≥ 2 (όλες εκτός της *switchde*) ή τιμή $q \leq 0{,}01$ (*switchde*). Όταν γίνεται σύγκριση δύο ομάδων κυττάρων (σε όλες τις μεθόδους εκτός της switchde), η σύγκριση γίνεται μεταξύ των κυττάρων που ανήκουν στη μικρο-κατάσταση έναρξης κι αυτών που ανήκουν στη μικρο-κατάσταση προορισμού.

Η συνάρτηση, keyGenesDEstr, χρησιμοποιεί το πακέτο R, *edgeR* [43], με γενικευμένα γραμμικά πρότυπα και δοκιμασία του λόγου της πιθανοφάνειας (likelihood ratio test), που είναι πιο κατάλληλες επιλογές για τα μονήρη κύτταρα.

Στη μέθοδο, *bimod* (συνάρτηση: keyGenesDEBimod), που είναι προσαρμοσμένη στα χαρακτηριστικά των μονήρων κυττάρων, ελέγχεται η αλλαγή τόσο της μέσης τιμής έκφρασης μεταξύ των συγκρινόμενων ομάδων κυττάρων (κύτταρα της μικρο-κατάσταση έναρξης και κύτταρα της μικρο-κατάστασης προορισμού) όσο κι η αναλογία των κυττάρων που εκφράζουν το γονίδιο, χρησιμοποιώντας τον λόγο της πιθανοφάνειας. Θεωρείται πως τη διτροπική (bimodal) έκφραση χαρακτηρίζουν, μία ομάδα που δεν εκφράζει το γονίδιο και μία με λογο-κανονική έκφραση.

Στη μέθοδο, *MAST* (συνάρτηση: keyGenesDEMAST), που έχει επίσης αναπτυχθεί για μονήρη κύτταρα, χρησιμοποιείται ένα γενικευμένο πρότυπο παλινδρόμησης δύο μερών (εμποδίου), προτυποποιώντας ξεχωριστά τη διακριτή τάξη έκφρασης (λογαριθμική παλινδρόμηση) και το συνεχές επίπεδο έκφρασης (κανονική κατανομή). Ο λόγος των γονιδίων που εκφράζονται ανά κύτταρο, θεωρείται πως αντανακλά τους τεχνικούς και βιολογικούς παράγοντες που επηρεάζουν την καθολική έκφραση κι είναι σημαντική πηγή της παρατηρούμενης μεταβλητότητας, αποτελώντας συνδιακυμαίνουσα μεταβλητή.

Στη μέθοδο, *switchde* (συνάρτηση: keyGenesDEswitchde), επιπλέον, χρησιμοποιείται ο ψευδοχρόνος των μονήρων κυττάρων στην τροχιά μετάβασης. Εδώ, ο ψευδοχρόνος κάθε τροχιάς, αντιστοιχεί στην εκ των υστέρων πιθανότητα να ανήκει το κύτταρο στην κατάσταση έναρξης. Άλλη μία διαφορά με τις άλλες μεθόδους, είναι πως χρησιμοποιούνται όλα τα κύτταρα της τροχιάς με τη μεταβολή στην έκφραση να είναι δυνατό να συμβεί σε οποιοδήποτε σημείο της. Υποθέτει ότι η έκφραση τού γονιδίου ακολουθεί σιγμοειδές πρότυπο κι εκ νέου χρησιμοποιείται δοκιμασία που στηρίζεται στον λόγο της πιθανοφάνειας.

2.2.8.2.1 Ταξινόμηση των κύριων γονιδίων

Τα κύρια γονίδια, ταξινομούνται με φθίνοντα τρόπο βάσει της προσαρμοσμένης τιμής p με τη μέθοδο BH (Benjamini and Hochberg) [44].

2.2.8.3 Χρήση πολλαπλών μεθόδων ταυτόχρονα

Χρησιμοποιώντας τη συνάρτηση, kg_voting, κύρια γονίδια, θεωρούνται αυτά που αναγνωρίζονται ως κύρια γονίδια από τουλάχιστον $\left\lceil \frac{\lceil \#\ επιλεγμένων\ μεθόδων}{2} \right\rceil$, από τις εφαρμοζόμενες μεθόδους· είναι διαθέσιμες έξι μέθοδοι (η προκαθορισμένη, MLscAN, κι

οι πέντε που αναφέρθηκαν στην ενότητα 2.2.8.2), αλλά, ο χρήστης μπορεί να ορίσει οποιοδήποτε σύνολο συναρτήσεων.

Σε αυτήν την περίπτωση, τα γονίδια διατάσσονται με φθίνοντα τρόπο, βάσει του αριθμού των μεθόδων που τα θεώρησαν κύρια γονίδια. Στην περίπτωση που υπάρχουν τουλάχιστον δύο γονίδια που θεωρήθηκαν κύρια γονίδια από τον ίδιο αριθμό μεθόδων, τότε, η διάταξη αυτού του υποσυνόλου, γίνεται χρησιμοποιώντας το άθροισμα της τιμής ανά μέθοδο, $\left[\frac{\#κύριων\ γονιδίων - θέση\ γονιδίου\ στη\ διάταξη + 1}{\#κύριων\ γονιδίων}\right]$, με φθίνοντα τρόπο.

2.2.9 Γονιδιακά ρυθμιστικά δίκτυα

Τα γονιδιακά ρυθμιστικά δίκτυα, δημιουργούνται ανά μικρο-κατάσταση της τροχιάς, χρησιμοποιώντας το σύνολο των κύριων γονιδίων της.

Στον πίνακα των βαρών που δημιουργείται κι αφορά στα γονίδια – στόχους (στήλες) σε σχέση με τα γονίδια – ρυθμιστές (γραμμές), θεωρείται πως η απόλυτη τιμή τους αντικατοπτρίζει τη σημασία / σπουδαιότητα του ρυθμιστή για τον στόχο. Με τα πρόσημα των βαρών, επισημαίνεται ο τύπος της αλληλεπίδρασης· το θετικό πρόσημο, υποδεικνύει ενισχυτική / ενεργοποιητική / διεγερτική δράση ενώ το αρνητικό πρόσημο, κατασταλτική / ανασταλτική.

Όταν δεν είναι δυνατό να δημιουργηθούν έγκυρα GRNs για κάθε σχηματισμένη μικρο-καταστάση της τροχιάς, γίνεται προσπάθεια δημιουργίας GRNs με διαφορετικά σύνολα κυττάρων, αλλά, με τα ίδια γονίδια – τα κύρια γονίδια της τροχιάς. Πιο συγκεκριμένα:

Αν η τροχιά έχει τρεις μικρο-καταστάσεις και δεν προκύπτουν τρία έγκυρα GRNs για αυτές, γίνεται προσπάθεια δημιουργίας GRNs, θεωρώντας πως υπάρχουν δύο νέες, παραγόμενες «μικρο-καταστάσεις», η groundG, που περιλαμβάνει τα κύτταρα της τροχιάς που ανήκουν στην κατάσταση έναρξης, κι η landG, που περιλαμβάνει τα κύτταρα της τροχιάς που ανήκουν στην κατάσταση προορισμού. Συνεπώς, αναμένεται να δημιουργηθούν δύο GRNs.

Αν η τροχιά έχει δύο μικρο-καταστάσεις και δεν προκύπτουν δύο έγκυρα GRNs για αυτές ή έχει αποτύχει η δημιουργία έγκυρων GRNs στις δύο προηγούμενες παραγόμενες «μικρο-καταστάσεις», γίνεται προσπάθεια δημιουργίας ενός GRN, θεωρώντας πως υπάρχει μία νέα παραγόμενη «μικρο-κατάσταση», η trajG, που περιλαμβάνει όλα τα κύτταρα της τροχιάς. Συνεπώς, αναμένεται να δημιουργηθεί ένα GRN.

Αν έχει αποτύχει η δημιουργία GRN στην παραγόμενη «μικρο-κατάσταση» trajG, γίνεται προσπάθεια δημιουργίας ενός GRN, θεωρώντας πως υπάρχει μία νέα παραγόμενη «μικρο-κατάσταση», η statesG, που περιλαμβάνει όλα τα κύτταρα των δύο καταστάσεων που σχηματίζουν την τροχιά. Συνεπώς, αναμένεται να δημιουργηθεί ένα GRN.

Εκτός της μεθόδου που αναφέρεται παρακάτω, μπορεί να δοθεί συνάρτηση που εκτελεί την επιθυμητή μέθοδο κι επιστρέφει τα αποτελέσματα (πίνακα βαρών με πρόσημα, για τις αλληλεπιδράσεις των γονιδίων) ενώ και μετά τη δημιουργία του προτύπου MLscAN είναι δυνατό να οριστούν απευθείας (μέσω πίνακα ή αρχείου).

2.2.9.1 Μέθοδος

2.2.9.1.1 Βάρη – ο αλγόριθμος GENIE3

Ο αλγόριθμος GENIE3 [45], έχει δειχθεί πως είναι αρκετά αποτελεσματικός [46, 47], και ταυτόχρονα, λειτουργεί χωρίς να στηρίζεται σε καμία υπόθεση για τον τύπο των αλληλεπιδράσεων, χειρίζεται μη-γραμμικές σχέσεις, δημιουργεί κατευθυνόμενες αλληλεπιδράσεις (όχι συμμετρικές) και δεν απαιτεί συγκεκριμένο τρόπο μετασχηματισμού των δεδομένων (αν κι αυτό, επηρεάζει τα αποτελέσματα).

Για τον υπολογισμό των βαρών, χρησιμοποιείται τυχαίο δάσος (προ-επιλογή) ή υπερ-τυχαιοποιημένα δέντρα, με δέντρα παλινδρόμησης (regression trees) δυαδικών κόμβων. Εκτός της μεθόδου, και το πλήθος των δέντρων (προ-επιλογή: 1.000), αποτελεί παράμετρο. Τα χαρακτηριστικά, δηλαδή τα γονίδια – ρυθμιστές σε κάθε δέντρο, είναι τυχαίο υπο-σύνολο της επιλεγμένης πληθικότητας (προ-επιλογή για το πλήθος των υποψήφιων ρυθμιστών: ρίζα του πλήθους των κύριων γονιδίων). Τα δεδομένα έκφρασης που λαμβάνονται υπόψη (κύτταρα x υποψήφια γονίδια – ρυθμιστές), τυποποιούνται, προκειμένου η διακύμανση να είναι μοναδιαία και να μπορούν να συγκριθούν τα βάρη μεταξύ τους, αφού σε αυτήν την περίπτωση δεν υπερισχύουν γονίδια με υψηλότερη μεταβλητότητα έκφρασης. Παρόμοια με τα γονίδια, σε κάθε δέντρο, επιλέγονται τυχαία, με επανατοποθέτηση, κύτταρα απο αυτά που παρέχονται.

Η κυριότερη διαφορά των δύο μεθόδων, εντοπίζεται στον τρόπο υπολογισμού της τιμής διαχωρισμού (split) σε κάθε κόμβο. Στο τυχαίο δάσος, επιλέγεται η περίπτωση του τοπικού ελαχίστου της συνάρτησης κόστους ενώ στα υπερ-τυχαιοποιημένα δέντρα, επιλέγεται τυχαία, μειώνοντας την υπολογιστική πολυπλοκότητα, αλλά, οδηγώντας σε λιγότερο καλά αποτελέσματα παρουσία θορύβου.

Η έξοδος κάθε δέντρου, είναι η εκτίμηση της συνεισφοράς των υποψήφιων γονιδίων – ρυθμιστών στην πρόβλεψη της έκφρασης του γονιδίου – στόχου. Η τελική τιμή του βάρους, είναι η μέση τιμή όλων των δέντρων, με μεγαλύτερες τιμές βαρών να αντιστοιχούν σε πιο πιθανές ρυθμιστικές σχέσεις

Στο τέλος, οι τιμές των βαρών ανά γονίδιο – στόχο, διαιρούνται με το άθροισμά τους, ώστε το άθροισμα των βαρών για αυτό να είναι ίσο με ένα. Συνεπώς, η ελάχιστη τιμή βάρους είναι 0 κι η μέγιστη 1.

2.2.9.1.2 Πρόσημα – τύποι αλληλεπίδρασης

Για την εξαγωγή του τύπου της αλληλεπίδρασης, χρησιμοποείται το πρόσημο του αποτελέσματος ανάλυσης της συσχέτισης μεταξύ κάθε ζεύγους γονιδίου – στόχου και γονιδίου – ρυθμιστή, με την επιλεγμένη μέθοδο συσχέτισης (προ-επιλογή: Spearman), με δειγματοληψία με αντικατάσταση. Ο συντελεστής συσχέτισης Spearman, έχει υψηλή τιμή, όταν συσχετίζονται μονότονα, χωρίς να απαιτείται η ύπαρξη γραμμικής σχέση. Αυτός ο τρόπος, βέβαια, αναμένεται να οδηγεί σε συμμετρία των τύπων αλληλεπιδράσεων, μεταξύ του γονιδίου 1 ως ρυθμιστή με το γονίδιο 2 ως στόχο και του γονιδίου 2 ως ρυθμιστή με το γονίδιο 1 ως στόχο.

Διαφορετικές προσεγγίσεις, δεν έχουν πολύ καλύτερη επίδοση από την τυχαία επιλογή και για τους δύο τύπους αλληλεπιδράσεων [49].

2.2.9.1.3 Πότε θεωρείται σημαντική η αλληλεπίδραση

Λόγω της δυσκολίας ορισμού μίας τιμής – ορίου προκειμένου να θεωρηθεί σημαντική η αλληλεπίδραση, καθώς και της σύγκρισης μεταξύ των βαρών διαφορετικών γονιδίων –

στόχων, κι ιδιαίτερα μεταξύ διαφορετικών μεθοδολογιών, έχει επιλεγεί να προβάλλονται στα διαγράμματα των GRNs που δημιουργούνται (θεωρώντας τες πιο σημαντικές) οι $\left[\sqrt{(\#key\_genes - 1)}\right]$ αλληλεπιδράσεις (με κορυφή το γονίδιο – στόχο) με τις μεγαλύτερες απόλυτες τιμές βαρών. Η παράμετρος αυτή μπορεί να μεταβληθεί, και τελικά να διατηρούνται $\min\{(\#\ key\_genes\ -\ 1),\ parameter\_value\}$ βάρη ανά γονίδιο – στόχο (παράδειγμα: εικόνα 3.54).

3. ΤΟ ΠΑΚΕΤΟ R MLscAN

Στο κεφάλαιο αυτό, παρουσιάζονται, η δομή του πακέτου, οι κλάσεις, οι μέθοδοι και συναρτήσεις που χρησιμοποιούνται, τα αρχεία εξόδου που δημιουργούνται, παραδείγματα χρήσης και τα αποτελέσματα του αναλυτή κατανομής / απόδοσης.

3.1 Δομή

Εκτός των απαιτούμενων στοιχείων: R (αρχεία κώδικα R), DESCRIPTION και NAMESPACE, στο πακέτο περιλαμβάνονται αρχεία τεκμηρίωσης (Rd) κάθε εξαγώμενης μεθόδου ή συνάρτησης, παραδείγματα χρήσης (vignette, demo) με δεδομένα που έχουν παραχθεί τυχαία κι αρχεία ελέγχου μονάδων (unit testing) χρησιμοποιώντας το πακέτο R, testthat [50].

Η δομή του πακέτου, χωρίς περιττές λεπτομέρειες, είναι η ακόλουθη:

- **data**

  Πίνακες των δεδομένων που χρησιμοποιούνται για τα παραδείγματα.

  - cellFeaturesRand.RData
  - exprRand.RData

- **demo**

  Αρχείο επίδειξης χρήσης.

  - MLscAN_demo.R

- **man**

  Αρχεία τεκμηρίωσης.

  - Rd files; one per exported method or function

- **R**

  - AllClasses.R
  - AllGenerics_Getters.R
  - AllGenerics_Setters.R
  - AllMethods_Constructors.R
  - AllMethods_Getters.R
  - AllMethods_Helpers.R
  - AllMethods_Plots.R
  - AllMethods_SaveInfo.R
  - AllMethods_Setters.R

  - cellFeaturesRand-data.R
  - exprRand-data.R
  - show.R

- **tests**

  Αρχεία ελέγχου μονάδων.

  - testthat [50]

- **vignettes**

  Αρχείο παραγωγής παραδείγματος χρήσης με παράθεση κώδικα και διαγραμμάτων.

  - MLscAN.Rmd

- **DESCRIPTION**
- **LICENSE**
- **NAMESPACE**
- **NEWS**

3.1.1 Κλάσεις

Στο πακέτο, έχουν δημιουργηθεί μόνο κλάσεις S4, που είναι ευρέως χρησιμοποιούμενες και διαθέτουν στοιχεία του κλασικού αντικειμενοστραφούς προγραμματισμού [48].

Συνολικά, δημιουργήθηκαν δέκα κλάσεις – οι εννέα σχετίζονται με τα βασικά βήματα της ροής επεξεργασίας (MLscANExpr: δεδομένα έκφρασης, MLscANDimRed: μείωση της διαστατικότητας, MLscANModel: πρότυπο καταστάσεων, MLscANSubpop: υπο-πληθυσμοί των καταστάσεων, MLscANOutliers: ακραίοι υπο-πληθυσμοί καταστάσεων, MLscANTraj: τροχιές, MLscANMicrost: μικρο-καταστάσεις, MLscANKeyGenes: κύρια γονίδια, MLscANGRN: GRN) κι η μία (MLscAN) περιλαμβάνει όλες τις υπόλοιπες καθώς και μεταβλητές γενικών επιλογών. Ακολουθεί η περιγραφή κάθε κλάσης και του τύπου των μελών της:

MLscAN

MLscANExpression="MLscANExpr"

MLscANDimRed="MLscANDimRed"

MLscANModel="MLscANModel"

MLscANGeneFeatures="matrix"

MLscANGeneFeaturesInInfo="character"

MLscANGeneFeaturesInFile="character"

MLscANCellFeatures="matrix"
MLscANCellFeaturesInFile="character"
MLscANCellFeaturesInInfo="character",
MLscANOutDir="character"
MLscANOutMode="character"
MLscANStopAt="character"
MLscANUseParallel="logical"

MLscANExpr

exprData="matrix"
exprInFile="character"
exprInInfo="character"

MLscANDimRed

dimRedData="matrix"
dimRedInFile="character"
dimRedInInfo="character"
dimRedDimVar="numeric"
dimRedMode="character"
dimRedCVar="numeric"
dimRedNumDim="numeric"
dimRedNumDimSelFun="function"
dimRedDimMode="character"
dimRedDimNames="character"

MLscANModel

modelMAPState2="matrix"
modelBICValues="numeric"
modelDBICValues="matrix"
modelDBICThr="numeric"
modelDBICFactor="numeric"
modelDBICStep="numeric"
modelDBICMode="character"
modelStateNameMode="character"
modelStatesSelFun="function"
modelNumStates="numeric"
modelProbsMode="character"

modelStatesMode="character"

modelInFile="character"

modelInInfo="character"

modelStatesNames="character"

modelOutl="MLscANOutliers"

modelSubpop="MLscANSubpop"

modelTrajectories="list"

modelEpigenetic="matrix"

MLscANOutliers

outlStates="logical"

outlPotentialStates="logical"

outlPotentialCells="character"

outlCells="character"

outlMaxStates="numeric"

outlMaxPercCells="numeric"

outlSelFun="function"

outlMode="character"

MLscANSubpop

subpop="list"

nonSubpop="list"

subpopThr="numeric"

MLscANTraj

trajCells="character"

trajStateFrom="character"

trajStateTo="character"

trajMStates="MLscANMicrost"

trajMonoMode="character"

trajKeyGenes="MLscANKeyGenes"

trajCandidate="logical"

trajValid="logical"

MLscANMicrost

msCells="list"

msStepProbs="numeric"

msKneePoints="numeric"

msKneePointsSelFun="function"

msKneePointsMode="character"

msGRNType="character"

msGRN="list"

msValid="logical"

msType="numeric"

MLscANKeyGenes

kgGenes="character"

kgCritValues="numeric"

kgNa0="numeric"

kgNa1="numeric"

kgNb0="numeric"

kgNb1="numeric"

kgFa0="numeric"

kgFa1="numeric"

kgFb0="numeric"

kgFb1="numeric"

kgSTDWeight="numeric"

kgSTDWeightCrit="numeric"

kgMaxPropDiff="numeric"

kgCriteriaThr="numeric"

kgGenesSelFun="function"

kgMode="character"

kgInFile="character"

kgInInfo="character"

kgValid="logical"

kgMinBICDiff="numeric"

MLscANGRN

grn="matrix"

grnInFile="character"

grnInInfo="character"

grnMode="character"

grnFun="function"

grnValid="logical"

grnK="character"

grnNTrees="numeric"

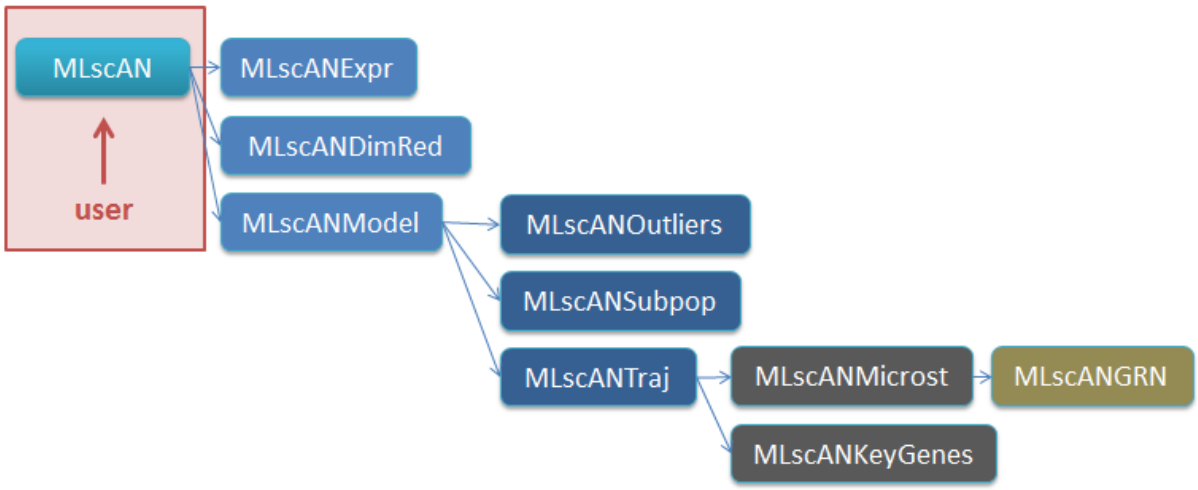
grnTreeMethod="character"

grnCorrMethod="character"

grnTargetNumRegulators="numeric"

grnType="character"

Σχηματικά η σύνδεση μεταξύ τους:



Εικόνα 3.1: Οι κλάσεις του πακέτου MLscAN κι η εμφώλευσή τους.

Παρότι δεν υπάρχει τρόπος να μην είναι προσβάσιμες κάποιες κλάσεις ή / και μέλη τους στον χρήστη, επιλέγεται να εξαχθεί μόνο η κλάση MlscAN· είναι η μόνη των οποίων τα αντικείμενα αναμένεται να χρησιμοποιήσει.

3.1.2 Μέθοδοι και συναρτήσεις

Για κάθε παράμετρο / μέλος των κλάσεων, υπάρχουν μέθοδοι / συναρτήσεις ανάκτησης και για επιλεγμένες, συναρτήσεις ανάθεσης. Επιπλέον, είναι διαθέσιμες βοηθητικές συναρτήσεις και συναρτήσεις για τη δημιουργία αρχείων (πληροφοριών ή διαγραμμάτων), μαζικά ή μεμονωμένα, δίνοντας μεγαλύτερο εύρος επιλογών στην οπτικοποίηση και στον τύπο των εξαγώμενων πληροφοριών.

3.1.2.1 Μέθοδοι Ανάκτησης (getters)

Οι μέθοδοι / συναρτήσεις ανάκτησης, έχουν όνομα, ίδιο με αυτό των μελών των κλάσεων.

Επιπλέον, για γρηγορότερη επιλογή σημαντικών χαρακτηριστικών, είναι διαθέσιμες οι συναρτήσεις:

- **geneNames:** επιστρέφει το σύνολο των ονομάτων των γονιδίων
- **cellNames:** επιστρέφει το σύνολο των ονομάτων των κυττάρων
- **cellTypes:** επιστρέφει το σύνολο των ονομάτων των κυτταρικών τύπων

- **cellFeaturesNames:** επιστρέφει τα ονόματα των χαρακτηριστικών των κυττάρων
- **geneFeaturesNames:** επιστρέφει τα ονόματα των χαρακτηριστικών των γονιδίων
- **stateCells:** επιστρέφει τα κύτταρα της επιλεγμένης κατάστασης
- **MAPPostProbs:** επιστρέφει την τιμή της μέγιστης εκ των υστέρων πιθανότητας ανά κύτταρο
- **MAP2PostProbs:** επιστρέφει την τιμή της δεύτερης μέγιστης εκ των υστέρων πιθανότητας ανά κύτταρο
- **trajNames:** επιστρέφει το σύνολο των τροχιών που σχηματίστηκαν
- **groundCells:** επιστρέφει τα κύτταρα της μικρο-κατάστασης έναρξης τής επιλεγμένης τροχιάς
- **transCells:** επιστρέφει τα κύτταρα της μικρο-κατάστασης μετάβασης της επιλεγμένης τροχιάς
- **landCells:** επιστρέφει τα κύτταρα της μικρο-κατάστασης προορισμού της επιλεγμένης τροχιάς
- **nonOutlCells:** επιστρέφει τα ονόματα των μη-ακραίων κυττάρων

3.1.2.2 Μέθοδοι Ανάθεσης (setters)

Είναι μέθοδοι / συναρτήσεις που επιτρέπουν την αλλαγή παραμέτρων εστιασμένα, όπως είναι για παράδειγμα, η αλλαγή της συνάρτησης επιλογής των κύριων γονιδίων για συγκεκριμένη τροχιά. Αυτές οι συναρτήσεις, είναι οι ακόλουθες:

- **MLscAN**
    - outDir (κατάλογος αποθήκευσης των δημιουργούμενων διαγραμμάτων κι αρχείων πληροφοριών)
- **MLscANDimRed**
    - dimRedDimNames (ονομασίες των διαστάσεων των αποτελεσμάτων μείωσης της διαστατικότητας)
- **MLscANModel**
    - statesNames (ονομασίες των καταστάσεων)
- **MLscANTraj**

- trajMonoMode (είδος μονοτονίας για τα την πιθανότητα της κατάστασης προορισμού)

- **MLscANMicrost**
  - msStepProbs (πιθανότητες για τον ορισμό των περιοχών εύρεσης των σημείων γονάτου)
  - msKneePointsSelFun (συνάρτηση εύρεσης των σημείων γονάτου)
  - msKneePointsMode (τρόπος εύρεσης των σημείων γονάτου)

- **MLscANKeyGenes**
  - keyGenes (κύρια γονίδια)
  - kgSTDWeight (τιμή της παραμέτρου, όπως περιγράφηκε στην ενότητα 2.2.8.1)
  - kgSTDWeightCrit (τιμή της παραμέτρου, όπως περιγράφηκε στην ενότητα 2.2.8.1)
  - kgMaxPropDiff (τιμή της παραμέτρου, όπως περιγράφηκε στην ενότητα 2.2.8.1)
  - kgMeansPropDiff (τιμή της παραμέτρου, όπως περιγράφηκε στην ενότητα 2.2.8.1)
  - kgCriteriaThr (τιμή της παραμέτρου, όπως περιγράφηκε στην ενότητα 2.2.8.1)
  - kgGenesSelFun (συνάρτηση επιλογής των κύριων γονιδίων)
  - kgMode (τρόπος επιλογής των κύριων γονιδίων)
  - kgMinBICDiff (τιμή της παραμέτρου, όπως περιγράφηκε στην ενότητα 2.2.8.1)

- **MLscANGRN**
  - grnWeights (πίνακας βαρών του GRN)
  - grnMode (τρόπος δημιουργίας του GRN)
  - grnFun (συνάρτηση δημιουργίας του GRN)
  - grnNTrees (αριθμός δένδρων για τη δημιουργία του GRN με τη μέθοδο GENIE3 [45])

- grnK (αριθμός υποψήφιων ρυθμιστών που για κάθε κόμβο των δένδρων που παραγάγονται για τη δημιουργία του GRN με τη μέθοδο GENIE3 [45])
- grnTreeMethod (τύπος δένδρων για τη δημιουργία του GRN με τη μέθοδο GENIE3 [45])
- grnCorrMethod (συντελεστής συσχέτισης για τον προσδιορισμό των τύπων των αλληλεπιδράσεων (πρόσημα) του GRN)
- grnTargetNumRegulators (αριθμός των γονιδίων – ρυθμιστών κάθε γονιδίου – στόχου που θεωρούνται σημαντικοί, και θα περιλαμβάνονται στα σχετικά διαγράμματα)

3.1.2.3 Βοηθητικές μέθοδοι

- **generateGRN:** δημιουργία GRN, χρησιμοποιώντας το σύνολο των κυττάρων μίας τροχιάς, το σύνολο των κυττάρων των δύο καταστάσεων που συμμετέχουν στη δημιουργία της τροχιάς ή κάποια από τις μικρο-καταστάσεις ή παραγόμενες «μικρο-καταστάσεις» (όπως περιγράφονται στην ενότητα 2.2.9) και τα κύρια γονίδια της επιλεγμένης τροχιάς
- **generateFiles:** δημιουργεί έναν κατάλογο με τα αρχεία εξόδου, βάσει των επιλεγμένων παραμέτρων, για ένα πρότυπο MLscAN.
- **getOverallKeyGenes:** επιστρέφει το σύνολο των γονιδίων που θεωρήθηκαν κύρια στο σύνολο των τροχιών
- **getTrajStates:** επιστρέφει τα ονόματα των καταστάσεων που σχηματίζουν μία τροχιά, δίνοντας το όνομα της τροχιάς και το διαχωριστικό των καταστάσεων, αν δε χρησιμοποιείται το προκαθορισμένο
- **createTrajNames:** επιστρέφει τα ονόματα όλων των πιθανών τροχιών (διατάξεις των καταστάσεων), δίνοντας τα ονόματα των καταστάσεων και το διαχωριστικό των καταστάσεων, αν δε χρησιμοποιείται το προκαθορισμένο
- **createTrajName:** επιστρέφει το όνομα μίας τροχιάς, δίνοντας τα ονόματα των καταστάσεων έναρξης και προορισμού και το διαχωριστικό των καταστάσεων, αν δε χρησιμοποιείται το προκαθορισμένο
- **getValidKeyGenesTrajs:** επιστρέφει το σύνολο των τροχιών στις οποίες αναγνωρίστηκε τουλάχιστον ένα κύριο γονίδιο
- **get2MStatesTrajs:** επιστρέφει το σύνολο των τροχιών με δύο μικρο-καστατάσεις

- **get3MStatesTrajs:** επιστρέφει το σύνολο των τροχιών με τρεις μικρο-καστατάσεις
- **confusionMatrixType:** πίνακας σύγχυσης, από τους κυτταρικούς τύπους (γραμμές) προς τις καταστάσεις (στήλες)
- **confusionMatrixState**: πίνακας σύγχυσης, από τις καταστάσεις (γραμμές) προς τους κυτταρικούς τύπους (στήλες)

3.1.2.4 Συναρτήσεις δημιουργίας αρχείων εξόδου

Μπορούν να διακριθούν στις ακόλουθες κατηγορίες:

- **Αρχείων πληροφοριών**
  - saveCellInfo
  - saveGeneInfo
  - saveInfoSummary
  - saveEpigeneticMatrix
  - saveTrajInfo
  - saveTrajGenesInfo
- **Διαγραμμάτων**
  - **Δεδομένων έκφρασης**
    - plotExprPCAVarInd
    - plotExprMeanSD
    - plotExprRatio0Genes
    - plotExprRatio0Cells
    - plotExprMinMeanMax
    - plotExprDeciles
    - plotExprHist
    - plotExprBoxplot
    - plotExprHeatmap
  - **Γενικά**

- plotEpigenetic
- plotTransitions
- plotTrajectories
- plotBIC
- plotCellFeature
- plotCellFeaturePie
- plotGeneFeature
- plotGeneFeaturePie
- plotMStates
- plotVarianceComb
- plotStatesComposition
- plotDecimalStates
- plotAlluvialState
- plotOverallKeyGenes
- plotNumPCStates
- plotSD

- **Αποτελεσμάτων μείωσης της διαστατικότητας**
  - plotDimRed
  - plotPCALoadings
  - plotPCALoadingsPairs
  - plotDimRedPairs
  - plotDimRed2Features
- **Καταστάσεων**
  - plotCircleState
  - plotDecimalState
  - plotDecimalStateAllStates
  - plotAlluvialState

- plotHeatmapState

- **Κυτταρικών τύπων**
  - plotCircleType
  - plotAlluvialState
  - plotHeatmapType
- **Τροχιών**
  - plotCircleTraj
  - plotProbTraj
  - plotHeatmapTraj
  - plotDotTraj
  - plotBarExprTraj
  - plotViolinTraj
  - plotViolinOverlayTraj
  - plotViolinTrajSmooth
  - plotViolinTrajMSSmooth
  - plotBoxplotTraj
  - plotKeyGenesCritValues
  - plotGRN
  - plotGRNBipartite
  - plotGRNHeatmap
  - plotGeneGRNHeatmap
  - plotGeneratedGRN

Περισσότερες πληροφορίες και παραδείγματα, δίνονται στην ενότητα 3.2.

Επιπλέον, είναι διαθέσιμη η συνάρτηση, generateFiles, για τη μαζική δημιουργία των επιλεγμένων τύπων διαγραμμάτων. Είναι ιδιαίτερα χρήσιμη όταν αλλαχθεί κάποια παράμετρος του προτύπου MLscAN μετά τη δημιουργία του.

3.2 Αρχεία εξόδου

3.2.1 Δεδομένα

3.2.1.1 Σύνοψη των πληροφοριών

Σε ένα απλό αρχείο κειμένου, αποθηκεύεται η σύνοψη των πληροφοριών – αποτελεσμάτων του αντικειμένου MLscAN, στις ακόλουθες ενότητες:

- **Γενικές πληροφορίες:**
    - οι διαστάσεις του πίνακα έκφρασης
    - ο αριθμός και το ποσοστό των ακραίων κυττάρων
    - ο αριθμός των διαστάσεων των αποτελεσμάτων μείωσης της διαστατικότητας που χρησιμοποιήθηκαν
    - η αθροιστική διακύμανση των διαστάσεων των αποτελεσμάτων μείωσης της διαστατικότητας που χρησιμοποιήθηκαν
    - η ονομασία των διαστάσεων των αποτελεσμάτων μείωσης της διαστατικότητας που χρησιμοποιήθηκαν
    - ο πίνακας σύγχυσης από τις καταστάσεις (γραμμές) προς τους κυτταρικούς τύπους (στήλες) κι αυτός από τους κυτταρικούς τύπους (γραμμές) προς τις καταστάσεις (στήλες)
- **Πληροφορίες ανά κυτταρικό τύπο:**
    - ο αριθμός των κυττάρων του τύπου και το ποσοστό τους στο σύνολο των κυττάρων
    - ο αριθμός των κυττάρων του τύπου, ανά κατάσταση και ποσοστό της κατάστασης που αποτελούν
- **Πληροφορίες ανά κυτταρική κατάταση:**
    - ο αριθμός των κυττάρων που ανήκουν στην κατάσταση και το ποσοστό τους στο σύνολο των κυττάρων
    - ο αριθμός των κυττάρων που ανήκουν στον υπο-πληθυσμό (όπως ορίστηκε στην ενότητα 2.2.4.2.1) της κατάστασης και το ποσοστό τους στο σύνολο των κυττάρων της κατάστασης

- ο αριθμός των κυττάρων της κατάστασης και το ποσοστό τους στο σύνολο των κυττάρων της κατάστασης, ανά κατάσταση που αντιστοιχεί η δεύτερη μεγαλύτερη εκ των υστέρων πιθανότητα
- ο αριθμός των κυττάρων και το ποσοστό τους (λαμβάνοντας υπόψη όλα τα κύτταρα) ανά διάστημα της εκ των υστέρων πιθανότητας για την κατάσταση

- **Πληροφορίες ανά τροχιά:**
  - ο αριθμός των κυττάρων της τροχιάς και το ποσοστό τους στο σύνολο των κυττάρων
  - ο αριθμός των μικρο-καταστάσεων
  - ο αριθμός των κυττάρων σε κάθε μικρο-κατάσταση, το ποσοστό τους στο σύνολο των κυττάρων της τροχιάς κι ένδειξη της παρουσίας έγκυρου GRN σε αυτήν
  - ένδειξη της ύπαρξης κύριων γονιδίων
  - ο αριθμός των κύριων γονιδίων και το ποσοστό τους στο σύνολο των γονιδίων

```
An S4 object of class `MLscAN`

##### GENERAL INFORMATION ####
- Initial expression data: 88 cells x 123 genes
- Expression data used: 88 cells x 123 genes
- Outlier states: 0, outlier cells: 0 (0.0%)
- No. dimensions of dimensionality reduction results used: 11
** Variance explained: 68.7%
** Dim. names: PC1, PC2, PC3, PC4, PC5, PC6, PC7, PC8, PC9, PC10, PC11

- Confusion matrix (%) - state:
                type `adult` type `T2D` type `child`
state `child`       5.00000    0.00000           95
state `T2D`         9.52381   90.47619            0
state `adult1`    100.00000    0.00000            0
state `adult2`     90.47619    9.52381            0

- Confusion matrix (%) - type:
             state `child` state `T2D` state `adult1` state `adult2`
type `adult`      2.083333    4.166667       54.16667       39.58333
type `T2D`        0.000000   90.476190        0.00000        9.52381
type `child`    100.000000    0.000000        0.00000        0.00000


##### CELL TYPES ####
- Type `adult`: 48 cells (54.5%)
** 1 cells in state `child` (5.0% of the state)
** 2 cells in state `T2D` (9.5% of the state)
```

Εικόνα 3.2: Μέρος των περιεχομένων του αρχείου σύνοψης των πληροφοριών που δημιουργείται.

3.2.1.2 Επιγενετικό τοπίο

Σε ένα απλό αρχείο κειμένου, αποθηκεύονται οι τιμές των λόγων των τάσεων μετάβασης μεταξύ των καταστάσεων (όπως ορίστηκαν στην ενότητα 2.2.6).

Στις γραμμές, βρίσκονται οι καταστάσεις της μέγιστης εκ των υστέρων πιθανότητας και στις στήλες, οι καταστάσεις της δεύτερης μέγιστης εκ των υστέρων πιθανότητας. Συνεπώς, η ελάχιστη τιμή είναι 0 κι η μέγιστη 1, και το άθροισμα των τιμών κάθε γραμμής είναι ίσο με 1.

	child	T2D	adult1	adult2
child	0.0000000	0.9000000	0.1000000	0.0000000
T2D	0.3809524	0.0000000	0.4761905	0.1428571
adult1	0.1538462	0.5000000	0.0000000	0.3461538
adult2	0.0000000	0.2380952	0.7619048	0.0000000

Εικόνα 3.3: Το περιεχόμενο του αρχείου του επιγενετικού τοπίου που δημιουργείται.

Παράδειγμα δημιουργίας:

```
saveEpigeneticMatrix(MLscAN_obj)
```

3.2.1.3 Χαρακτηριστικά των κυττάρων

Σε ένα αρχείο CSV ή TAB, αποθηκεύονται τα επιλεγμένα χαρακτηριστικά των επιλεγμένων κυττάρων (προ-επιλογή: όλα), από τα παρακάτω:

- **cellName:** το όνομα του κυττάρου
- **cellType:** η τιμή ανά διαθέσιμο χαρακτηριστικό των κυττάρων (από τον πίνακα που δίνεται ως παράμετρος)
- **expr:** η τιμή έκφρασης ανά γονίδιο
- **dimRed:** η τιμή ανά διάσταση των δεδομένων μειωμένης διαστατικότητας που χρησιμοποιήθηκε
- **state:** η κατάσταση στην οποία ανήκει
- **transitionState:** η κατάσταση της δεύτερης μέγιστης εκ των υστέρων πιθανότητας
- **inSubpop:** TRUE, αν ανήκει στον υπο-πληθυσμό της κατάστασης
- **isOutl:** TRUE, αν είναι ακραίο κύτταρο
- **isPotentialOutl:** TRUE, αν αναγνωρίστηκε ότι είναι ακραίο κύτταρο, αλλά, δεν αφαιρέθηκε από τα επόμενα στάδια της ροής επεξεργασίας επειδή αυτό δεν ήταν δυνατό

Τα κύτταρα καθώς και τα χαρακτηριστικά (μαζί με τα διαθέσιμα φίλτρα), αποτελούν παραμέτρους.

	Standard	Standard	Standard	Standard	S
1	cellName	cellType	donor	CLDN2	I
2	reads.16097	adult	adult_ABAF490	0.122868655561865	0
3	reads.26087	child	child_ICRH76	2.89726788490565	3
4	reads.26095	child	child_ICRH76	1.85772973874341	2
5	reads.26123	child	child_ICRH76	0.551206005341635	0
6	reads.26129	child	child_ICRH76	3.60537429335828	2
7	reads.29312	child	child_ICRH80	1.81173564126599	1
8	reads.29317	child	child_ICRH80	2.70748535875479	2

Εικόνα 3.4: Μέρος των περιεχομένων του αρχείου με τα χαρακτηριστικά των κυττάρων που δημιουργείται.

Παράδειγμα δημιουργίας:

```
saveCellInfo(MLscAN_obj)
```

3.2.1.4 Χαρακτηριστικά των γονιδίων

Σε ένα αρχείο CSV ή TAB, αποθηκεύονται τα επιλεγμένα χαρακτηριστικά των επιλεγμένων γονιδίων (προ-επιλογή: όλα), από τα παρακάτω:

- **cellName:** το όνομα του κυττάρου
- **cellType:** η τιμή ανά διαθέσιμο χαρακτηριστικό των γονιδίων (από τον πίνακα που δίνεται ως παράμετρος)

Τα γονίδια καθώς και τα χαρακτηριστικά (μαζί με τα διαθέσιμα φίλτρα), αποτελούν παραμέτρους.

	Standard
1	geneName
2	CLDN2
3	IL8
4	SPARC
5	PIGR
6	COL6A3
7	COL1A2
8	COL1A1

Εικόνα 3.5: Μέρος των περιεχομένων του αρχείου με τα χαρακτηριστικά των γονιδίων που δημιουργείται.

Παράδειγμα δημιουργίας:

```
saveGeneInfo(MLscAN_obj)
```

3.2.1.5 Πληροφορίες ανά τροχιά

Σε ένα αρχείο CSV ή TAB, αποθηκεύονται πληροφορίες για τις επιλεγμένες τροχιές (προ-επιλογή: όλες), από τα παρακάτω:

- **trajName:** το όνομα της τροχιάς
- **hasMStates:** ο αριθμός των μικρο-καταστάσεων της τροχιάς
- **haskeyGenes:** TRUE, αν έχει κύρια γονίδια
- **numKeyGenes:** ο αριθμός των κύριων γονιδίων της τροχιάς
- **numCellsTraj:** ο αριθμός των κυττάρων της τροχιάς
- **numCellsTrajGround:** ο αριθμός των κυττάρων της μικρο-κατάστασης έναρξης
- **numCellsTrajTrans:** ο αριθμός των κυττάρων της μικρο-κατάστασης μετάβασης
- **numCellsTrajLand:** ο αριθμός των κυττάρων της μικρο-κατάστασης προορισμού
- **validGRNGround:** TRUE, αν υπάρχει έγκυρο GRN στη μικρο-κατάσταση έναρξης
- **validGRNTrans:** TRUE, αν υπάρχει έγκυρο GRN στη μικρο-κατάσταση μετάβασης
- **validGRNLand:** TRUE, αν υπάρχει έγκυρο GRN στη μικρο-κατάσταση προορισμού
- **validGRNGroundG:** TRUE, αν υπάρχει έγκυρο GRN τύπου grounG (όπως περιγράφηκε στην ενότητα 2.2.9)
- **validGRNLandG:** TRUE, αν υπάρχει έγκυρο GRN τύπου landG (όπως περιγράφηκε στην ενότητα 2.2.9)
- **validGRNTrajG:** TRUE, αν υπάρχει έγκυρο GRN τύπου trajG (όπως περιγράφηκε στην ενότητα 2.2.9)
- **validGRNStatesG:** TRUE, αν υπάρχει έγκυρο GRN τύπου statesG (όπως περιγράφηκε στην ενότητα 2.2.9)
- **numCellsA:** ο αριθμός κυττάρων της κατάστασης A (έναρξης) που ανήκουν στην τροχιά A-to-B
- **numCellsStateA:** ο αριθμός των κυττάρων της κατάστασης A (έναρξης) συνολικά (τροχιά A-to-B)

- **numCellsB:** ο αριθμός των κυττάρων της κατάστασης B (προορισμού) που ανήκουν στην τροχιά A-to-B
- **numCellsStateB:** ο αριθμός των κυττάρων της κατάστασης B (προορισμού) συνολικά (τροχιά A-to-B)
- **probAtoB:** η πιθανότητα η 2η μέγιστη εκ των υστέρων πιθανότητα να αντιστοιχεί στην κατάσταση B (προορισμού), δεδομένου ότι η μέγιστη εκ των υστέρων πιθανότητα αντιστοιχεί στην κατάσταση A (έναρξης)
- **probBtoA:** η πιθανότητα η 2η μέγιστη εκ των υστέρων πιθανότητα να αντιστοιχεί στην κατάσταση A (έναρξης), δεδομένου ότι η μέγιστη εκ των υστέρων πιθανότητα αντιστοιχεί στην κατάσταση B (προορισμού)

Οι τροχιές, αποτελούν παράμετρο.

	Standard	Standard	Standard	Standard	Standard
1	trajName	hasMStates	hasKeyGenes	numCellsTraj	numCellsTrajG
2	child-to-T2D	2	TRUE	26	18
3	T2D-to-child	2	TRUE	26	8
4	T2D-to-adult1	3	TRUE	23	9
5	adult1-to-T2D	3	TRUE	23	12
6	T2D-to-adult2	2	TRUE	8	3
7	adult2-to-T2D	2	TRUE	8	5
8	adult1-to-adult2	3	TRUE	25	8

Εικόνα 3.6: Μέρος των περιεχομένων του αρχείου των πληροφοριών για τις τροχιές που δημιουργείται.

Παράδειγμα δημιουργίας:

```
saveTrajInfo(MLscAN_obj)
```

3.2.1.6 Πληροφορίες για τα γονίδια σε σχέση με τις τροχιές

Σε ένα αρχείο CSV ή TAB, αποθηκεύονται τα επιλεγμένα χαρακτηριστικά των γονιδίων (προ-επιλογή: όλα) σε σχέση με τις επιλεγμένες τροχιές (προ-επιλογή: όλες), από τα παρακάτω:

- **trajName:** το όνομα της τροχιάς
- **geneName:** το όνομα του γονιδίου
- **iskeyGene:** TRUE, αν είναι κύριο γονίδιο της τροχιάς
- **critValue:** η τιμή του κριτηρίου ταξινόμησης των γονιδίων (όπως περιγράφηκε στην ενότητα 2.2.8)

- **nA0:** αριθμός κυττάρων της μικρο-κατάστασης έναρξης με χαμηλό επίπεδο έκφρασης (όπως περιγράφηκε στην ενότητα 2.2.8.1)
- **nA1:** αριθμός κυττάρων της μικρο-κατάστασης έναρξης με υψηλό επίπεδο έκφρασης (όπως περιγράφηκε στην ενότητα 2.2.8.1)
- **nB0:** αριθμός κυττάρων της μικρο-κατάστασης προορισμού με χαμηλό επίπεδο έκφρασης (όπως περιγράφηκε στην ενότητα 2.2.8.1)
- **nB1:** αριθμός κυττάρων της μικρο-κατάστασης προορισμού με υψηλό επίπεδο έκφρασης (όπως περιγράφηκε στην ενότητα 2.2.8.1)
- **fA0:** λόγος των κυττάρων της μικρο-κατάστασης έναρξης με χαμηλό επίπεδο έκφρασης (όπως περιγράφηκε στην ενότητα 2.2.8.1)
- **fA1:** λόγος των κυττάρων της μικρο-κατάστασης έναρξης με υψηλό επίπεδο έκφρασης, στην τροχιά A-to-B (όπως περιγράφηκε στην ενότητα 2.2.8.1)
- **fB0:** λόγος των κυττάρων της μικρο-κατάστασης προορισμού με χαμηλό επίπεδο έκφρασης, στην τροχιά A-to-B (όπως περιγράφηκε στην ενότητα 2.2.8.1)
- **fB1:** λόγος των κυττάρων της μικρο-κατάστασης προορισμού με υψηλό επίπεδο έκφρασης, στην τροχιά A-to-B (όπως περιγράφηκε στην ενότητα 2.2.8.1)

Οι τροχιές και τα γονίδια, αποτελούν παραμέτρους.

	Standard	Standard	Standard	Standard	Standard	Standard	Star
1	trajName	geneName	isKeyGene	critValue	nS0	nS1	nE0
2	child-to-T2D	CLDN2	TRUE	NA	NA	NA	NA
3	T2D-to-child	IL8	TRUE	NA	NA	NA	NA
4	T2D-to-adult1	SPARC	TRUE	NA	NA	NA	NA
5	adult1-to-T2D	PIGR	FALSE	NA	NA	NA	NA
6	T2D-to-adult2	COL6A3	TRUE	NA	NA	NA	NA
7	adult2-to-T2D	COL1A2	TRUE	NA	NA	NA	NA
8	adult1-to-adult2	COL1A1	TRUE	NA	NA	NA	NA

Εικόνα 3.7: Μέρος των περιεχομένων του των πληροφοριών για τα γονίδια σε σχέση με τις τροχιές που δημιουργείται.

Παράδειγμα δημιουργίας:

```
saveTrajGenesInfo(MLscAN_obj)
```

3.2.2 Διαγράμματα

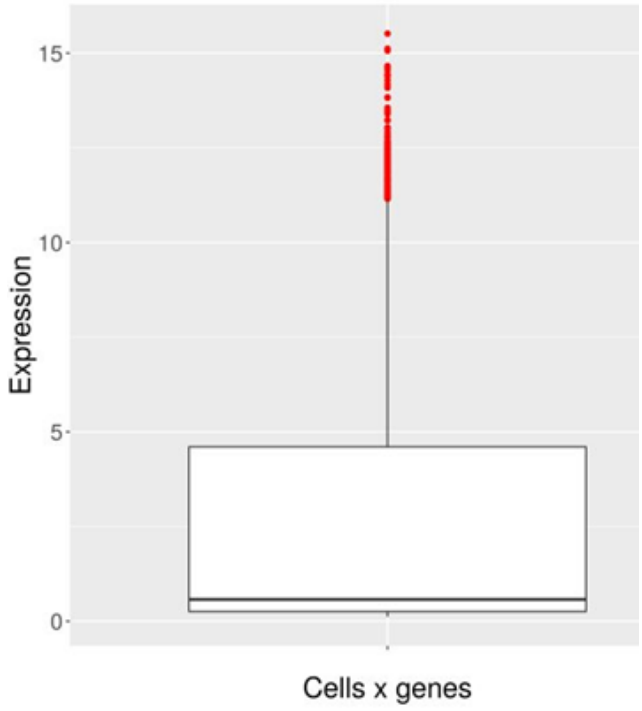
Τα διαγράμματα, ομαδοποιούνται κι αποθηκεύονται σε διαφορετικό κατάλογο. Διακρίνονται σε αυτά που αφορούν στον αρχικό πίνακα έκφρασης (Expr_plots), στα αποτελέσματα μείωσης της διαστατικότητας (DimRed_plots), στους κυτταρικούς τύπους

(Types_plots), στις καταστάσεις (States_plots), στις τροχιές (Trajs_plots), και σε γενικότερες πληροφορίες (General_plots).

Οι κυτταρικοί τύποι κι οι καταστάσεις, διατηρούν σε όλα τα διαγράμματα το ίδιο χρώμα.

3.2.2.1 Δεδομένων έκφρασης (Expr_plots)

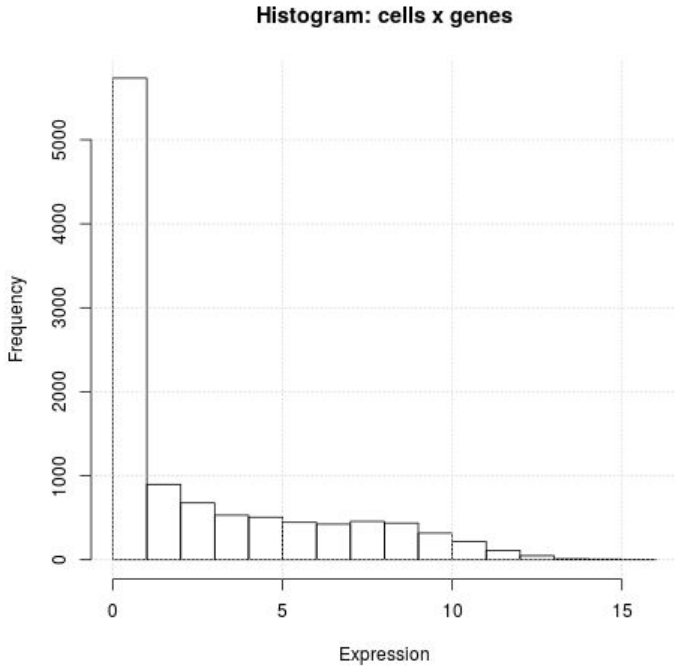
Με τα διαγράμματα αυτής της ενότητας, μπορούν να ελεχθούν βασικές πληροφορίες για τα δεδομένα του μετασχηματισμένου πίνακα δεδομένων, στο σύνολό τους (κύτταρα x γονίδια), ανά κύτταρο κι ανά γονίδιο.



Εικόνα 3.8: Θηκόγραμμα (boxplot) που προκύπτει από το σύνολο των τιμών του πίνακα έκφρασης (κύτταρα x γονίδια).

Παράδειγμα δημιουργίας:

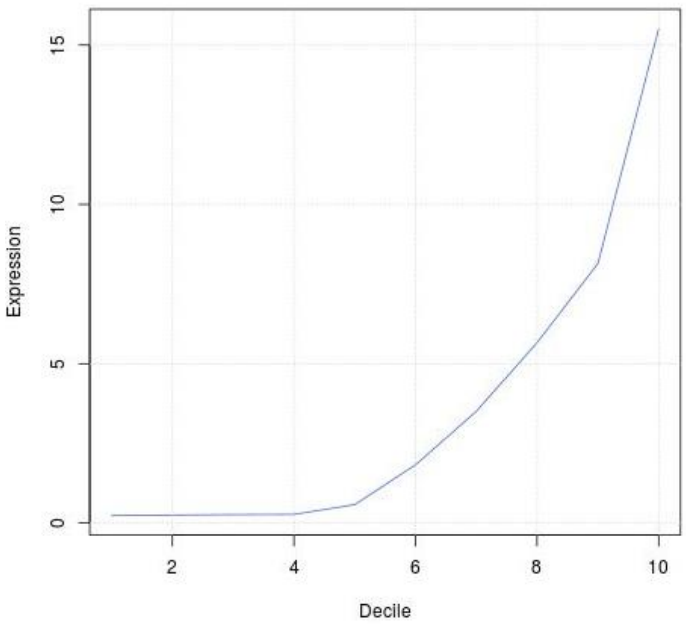
```
plotExprBoxplot(exprData(MLscAN_obj))
```



Εικόνα 3.9: Ιστόγραμμα (histogram) που προκύπτει από το σύνολο των τιμών του πίνακα έκφρασης (κύτταρα x γονίδια).

Παράδειγμα δημιουργίας:

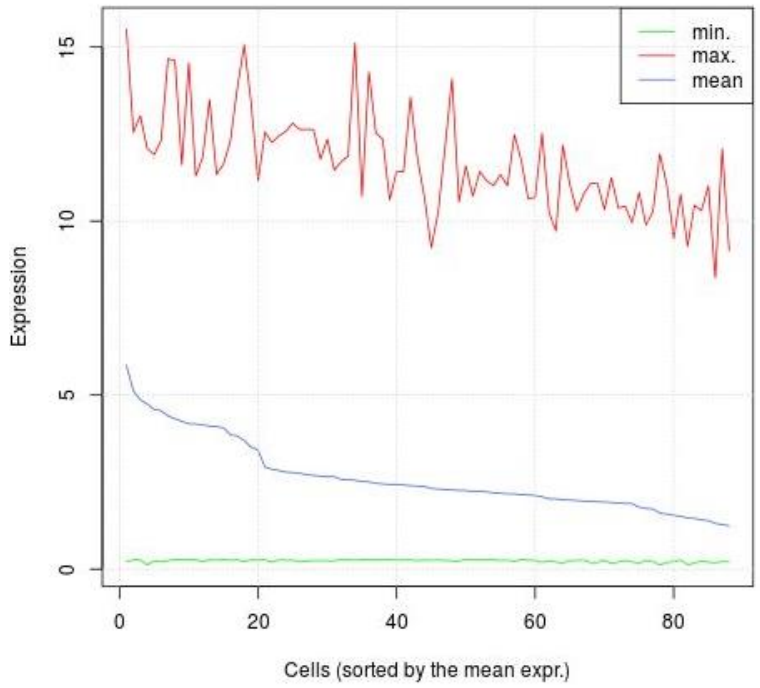
```
plotExprHist(exprData(MLscAN_obj))
```



Εικόνα 3.10: Διάγραμμα των τιμών έκφρασης ανά δεκατημόριο, οι οποίες προκύπτουν από το σύνολο των τιμών του πίνακα έκφρασης (κύτταρα x γονίδια).

Παράδειγμα δημιουργίας:

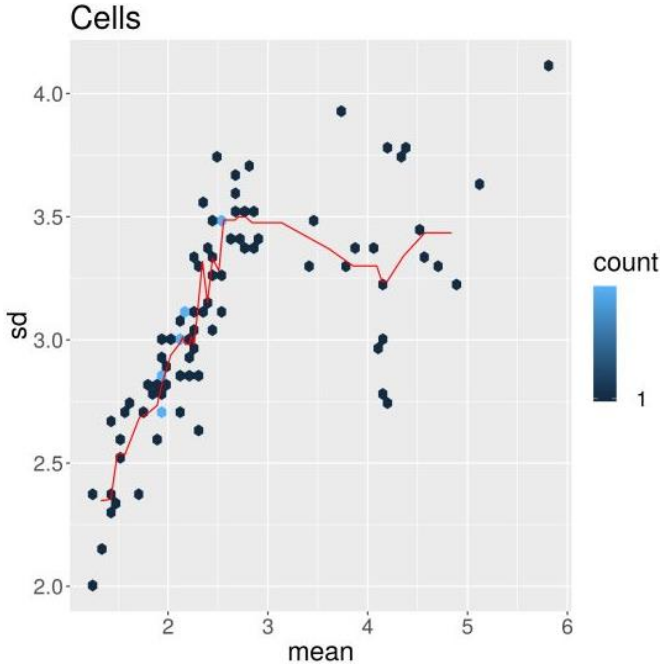
```
plotExprDeciles(as.vector(exprData(MLscAN_obj)))
```



Εικόνα 3.11: Διάγραμμα της ελάχιστης, της μέσης και της μέγιστης τιμών έκφρασης ανά κύτταρο. Τα κύτταρα διατάσσονται με φθίνοντα τρόπο βάσει της μέσης έκφρασης όλων των γονιδίων σε καθένα από αυτά.

Παράδειγμα δημιουργίας:

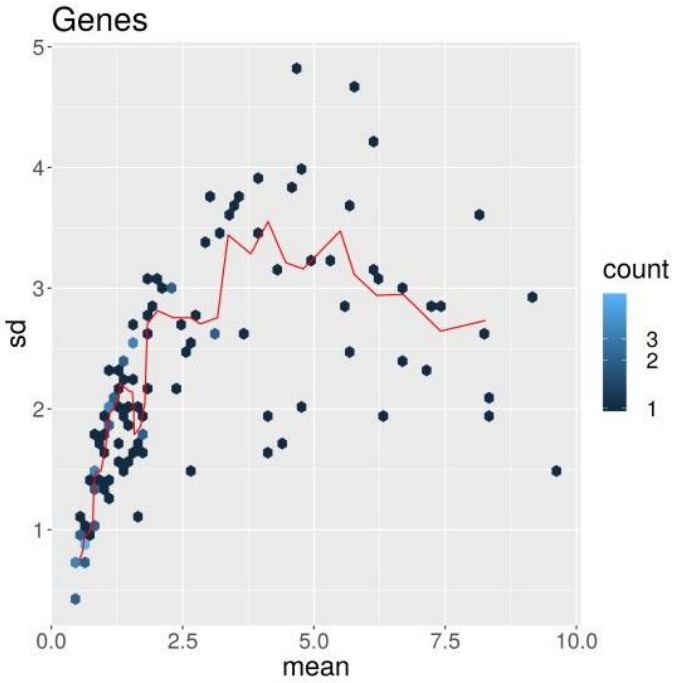
```
plotExprMinMeanMax(exprData(MLscAN_obj))
```



Εικόνα 3.12: Διάγραμμα της μέσης τιμής έκφρασης προς την τυπική απόκλιση, ανά κύτταρο. Οι κυψελιδικές περιοχές, χρωματίζονται ανάλογα με το πλήθος των σημείων που περιλαμβάνουν. Η κόκκινη γραμμή, αποτελεί την εκτίμηση της κινούμενης διάμεσης τιμής.

Παράδειγμα δημιουργίας:

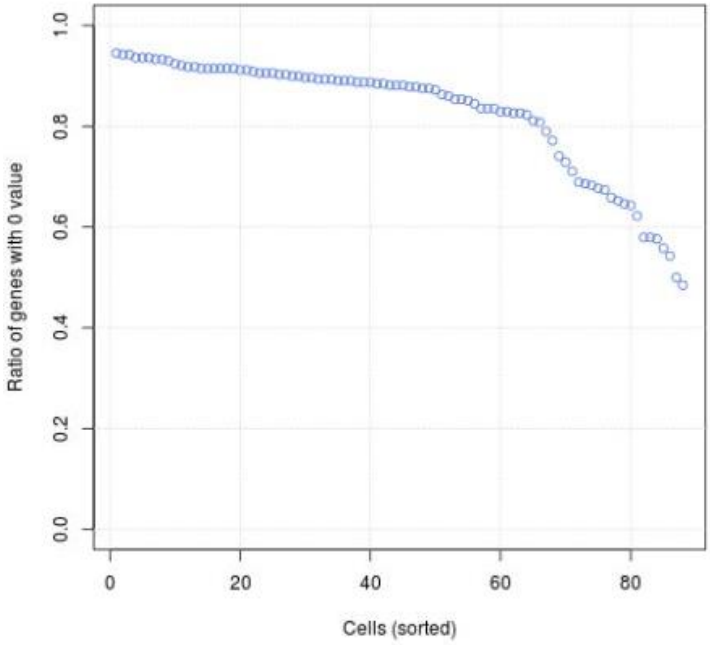
```
plotExprMeanSD(exprData(MLscAN_obj))
```



Εικόνα 3.13: Διάγραμμα της μέσης τιμής έκφρασης προς την τυπική απόκλιση, ανά γονίδιο (σε όλα τα κύτταρα). Οι κυψελιδικές περιοχές, χρωματίζονται ανάλογα με το πλήθος των σημείων που περιλαμβάνουν. Η κόκκινη γραμμή, αποτελεί την εκτίμηση της κινούμενης διάμεσης τιμής.

Παράδειγμα δημιουργίας:

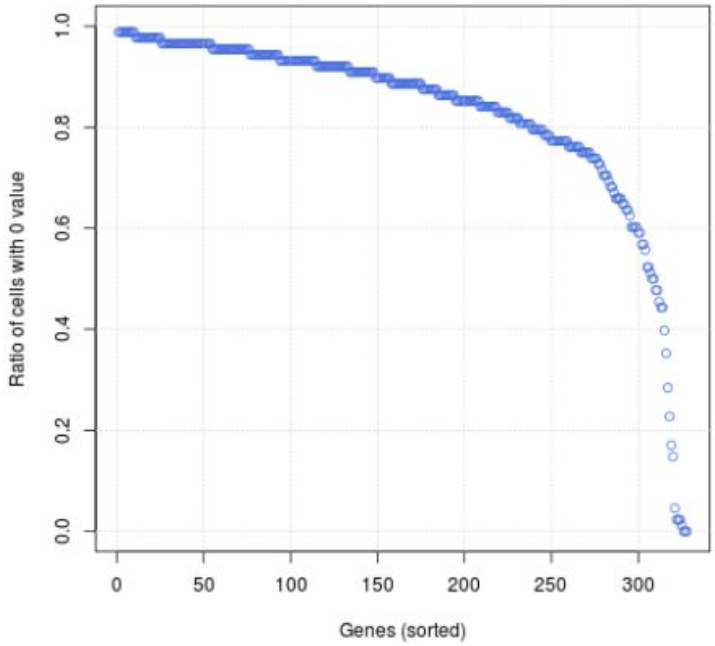
```
plotExprMeanSD(t(exprData(MLscAN_obj)))
```



Εικόνα 3.14: Διάγραμμα του λόγου των μηδενικών τιμών της έκφρασης ανά κύτταρο, διατάσσοντας τα κύτταρα με φθίνοντα τρόπο βάσει του λόγου των μηδενικών τιμών.

Παράδειγμα δημιουργίας:

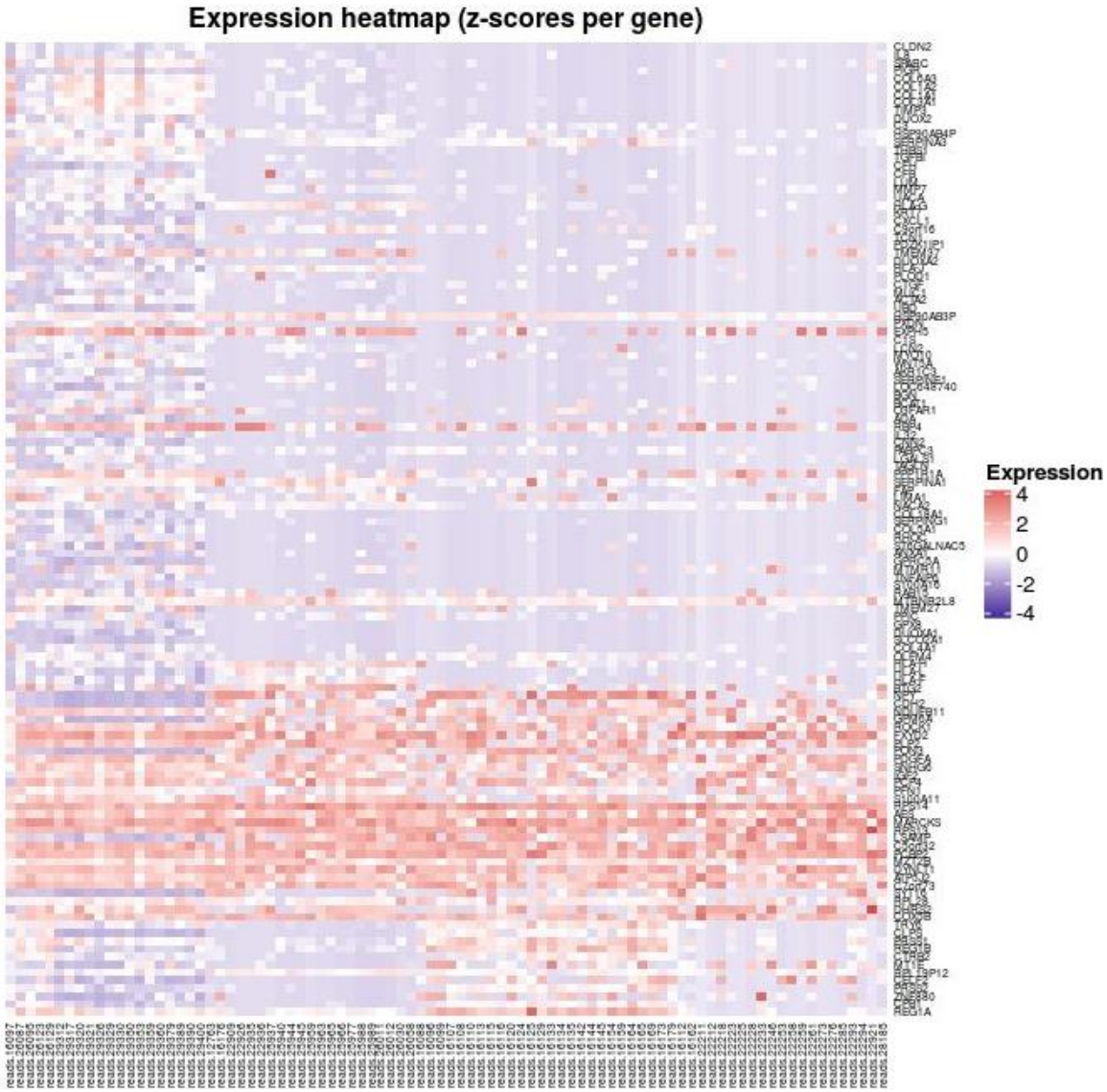
```
plotExprRatio0Cells(exprData(MLscAN_obj))
```

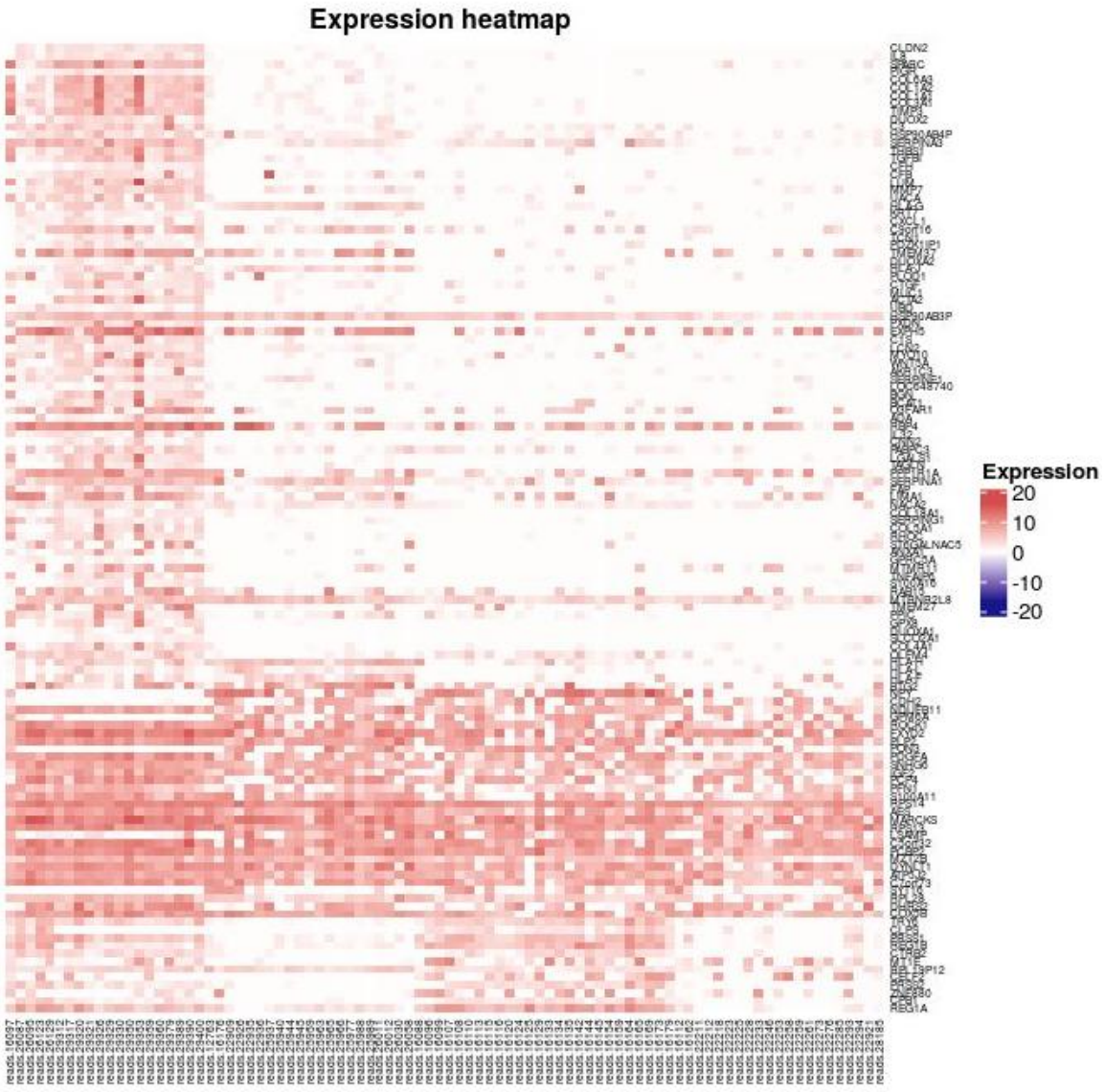


Εικόνα 3.15: Διάγραμμα του λόγου των μηδενικών τιμών ανά γονίδιο, διατάσσοντας τα γονίδια με φθίνοντα τρόπο βάσει του λόγου των μηδενικών τιμών.

Παράδειγμα δημιουργίας:

```
plotExprRatio0Genes(exprData(MLscAN_obj))
```


Expression heatmap (z-scores per gene)
Expression
4
2
0
-2
-4



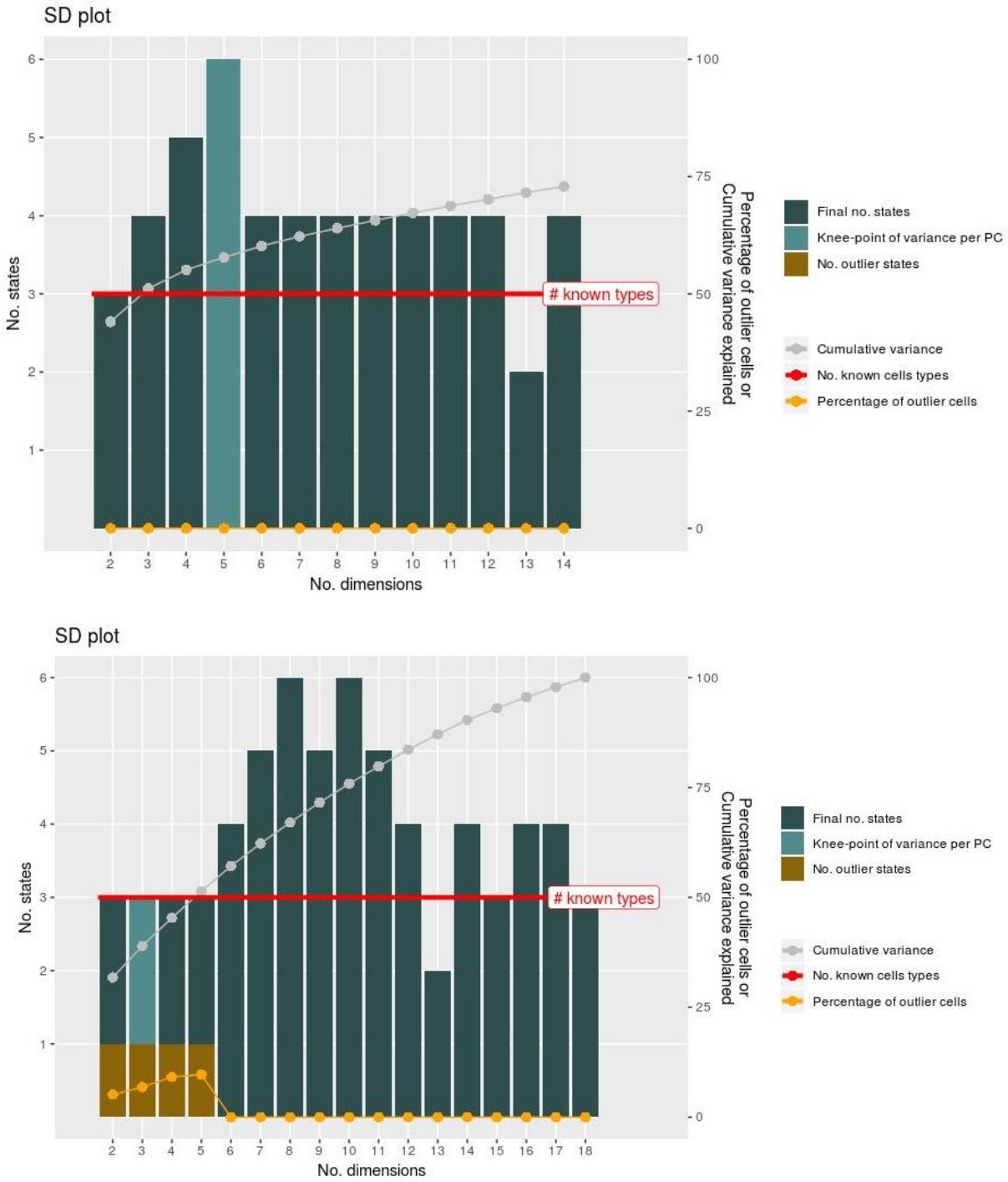
Εικόνα 3.16: Χάρτες θερμότητας (heatmap) της έκφρασης όλων των γονιδίων όλων των κυττάρων, με χρήση (πρώτο διάγραμμα) ή χωρίς χρήση (δεύτερο διάγραμμα) των τυπικών τιμών (z-scores) ανά γονίδιο.

Παραδείγματα δημιουργίας:

```
plotExprHeatmap(exprData(MLscAN_obj))
plotExprHeatmap(exprData(MLscAN_obj), z_scores=FALSE)
```

3.2.2.2 Γενικά διαγράμματα (Overview_plots)

Εδώ, περιλαμβάνονται διαγράμματα που συνδυάζουν χαρατηριστικά διαφορετικών βημάτων της ροής επεξεργασίας ή / και δεν εστιάζονται σε έναν κυτταρικό τύπο, μία κατάσταση ή μία τροχιά.



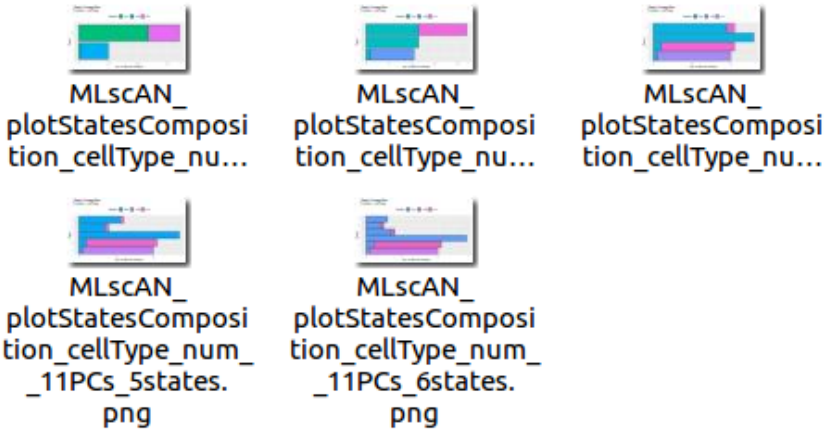
Εικόνα 3.17: Παραδείγματα διαγραμμάτων με συγκεντρωτικά στοιχεία για τα αποτελέσματα του προτύπου MLscAN, σε ένα εύρος χρησιμοποιούμενων συνιστωσών των αποτελεσμάτων μείωσης της διαστατικότητας, χρησιμοποιώντας σταθερές παραμέτρους. Επισημαίνονται, ανά περίπτωση, ο τελικός αριθμός των καταστάσεων, ο αριθμός των ακραίων υπο-πληθυσμών και το ποσοστό των κυττάρων τους στο σύνολο των κυττάρων, το σημείο γονάτου της διακύμανσης ανά

συνιστώσα κι η αθροιστική διακύμανση. Η κόκκινη γραμμή, αντιστοιχεί στον αριθμό των γνωστών τύπων κυττάρων (όταν είναι διαθέσιμη αυτή η πληροφορία), προκειμένου να συγκρίνονται άμεσα με αυτόν οι καταστάσεις που δημιουργούνται ανά περίπτωση.

Με τη χρήση αυτού του τύπου διαγράμματος και λαμβάνοντας υπόψη τον αριθμό των καταστάσεων που «κυριαρχεί» και τη σύσταση των καταστάσεων σε κάθε περίπτωση, μπορεί να επιλεγεί ο πλέον κατάλληλος αριθμός διαστάσεων, αυξάνοντας τη βιολογική σημασία των τροχιών που πιθανώς θα σχηματιστούν.

Παράδειγμα δημιουργίας:

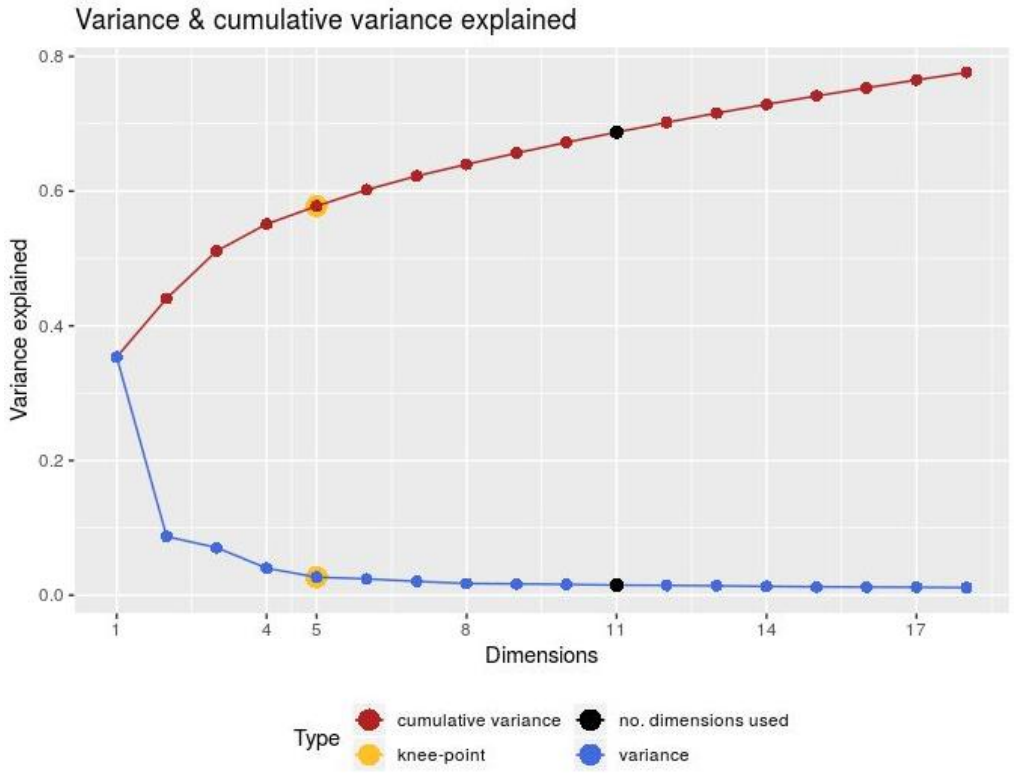
```
plotSD(expr, from=2, to=18,
       known_cellTypes=letters[seq(3)],
       plot_statesComp=TRUE)
```



Εικόνα 3.18: Διαγράμματα της σύστασης των καταστάσεων που σχηματίζονται, επιλέγοντας συγκεκριμένο αριθμό διαστάσεων των αποτελεσμάτων μείωσης της διαστατικότητας κι ένα εύρος αριθμού καταστάσεων.

Παράδειγμα δημιουργίας:

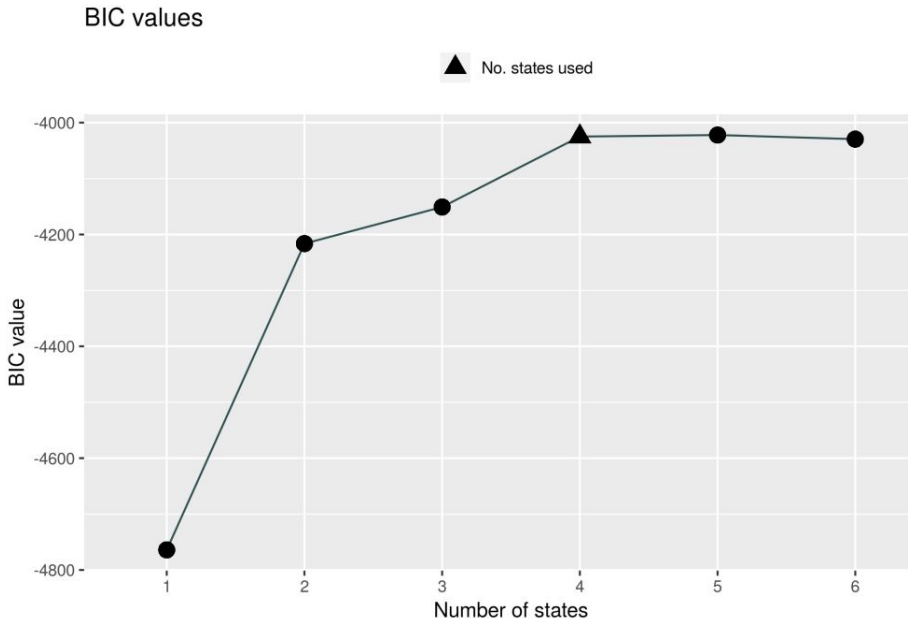
```
plotNumPCStates(expr, 11, from=2, to=6,
                cellType=data$cell_features[,"cellType"])
```



Εικόνα 3.19: Διάγραμμα της διακύμανσης κι αθροιστικής διακύμανσης ανά συνιστώσα των αποτελεσμάτων μείωσης της διαστατικότητας.

Παράδειγμα δημιουργίας:

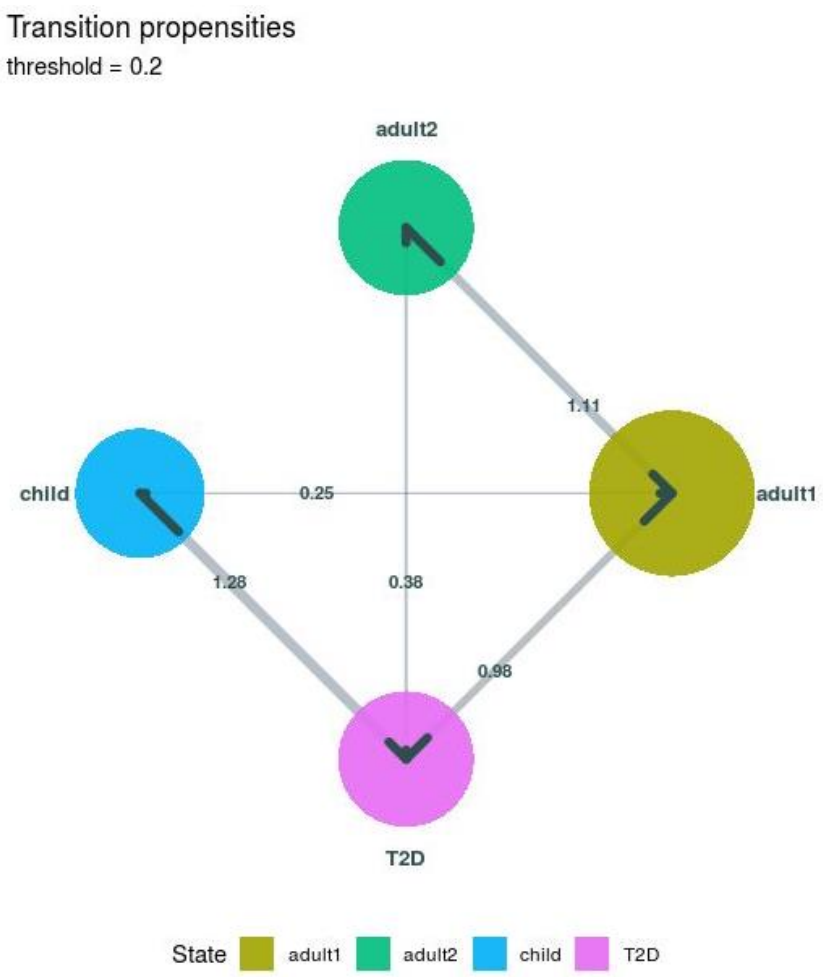
```
plotVarianceComb(MLscAN_obj, from=1, to=18)
```



Εικόνα 3.20: Διάγραμμα των τιμών του BIC σε σχέση με τον αριθμό των καταστάσεων, επισημαίνοντας τον αριθμό των καταστάσεων που έχει επιλεγεί.

Παράδειγμα δημιουργίας:

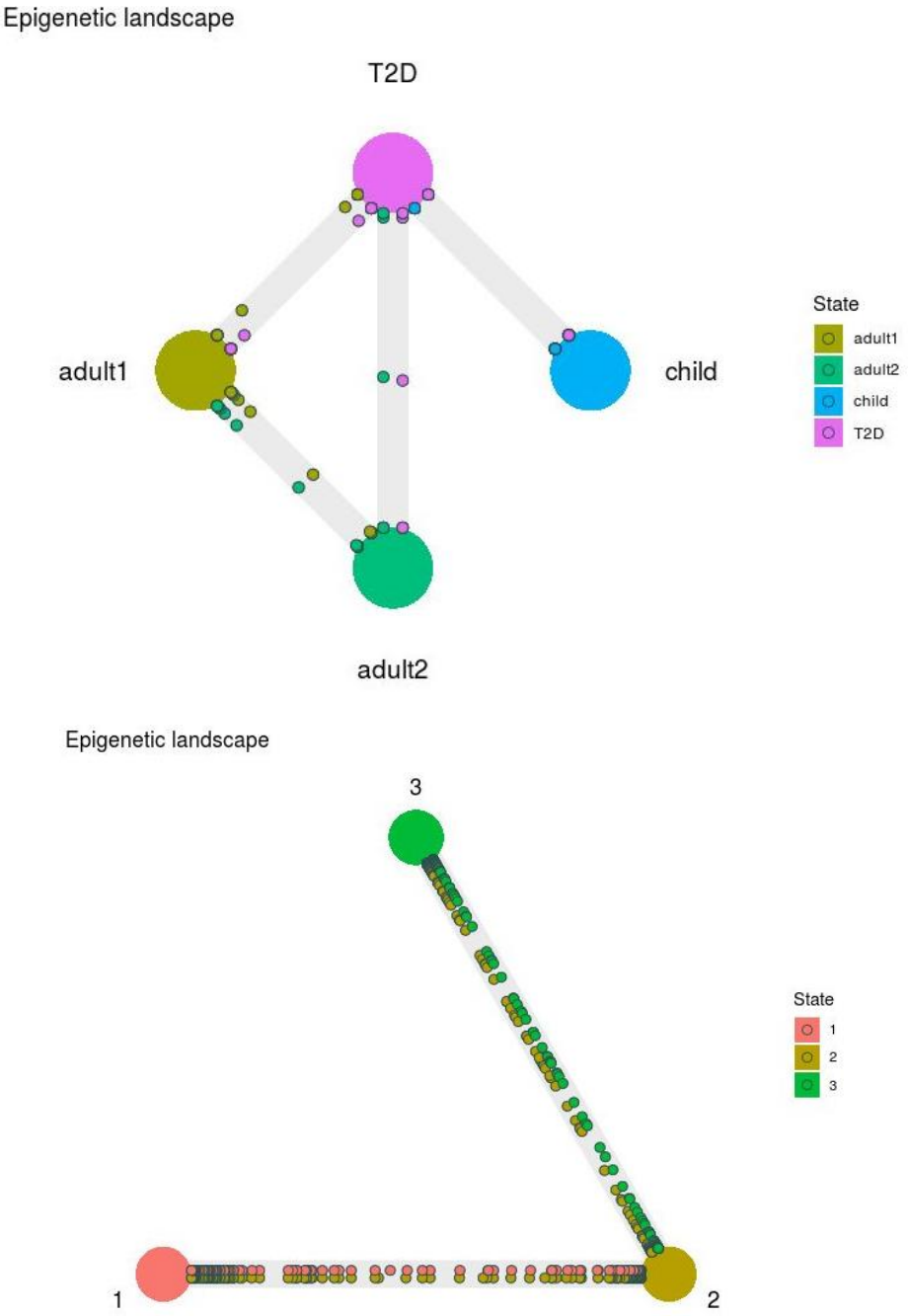
```
plotBIC(MLscAN_obj)
```



Εικόνα 3.21: Διάγραμμα των τάσεων μετάβασης (transition propensities) μεταξύ των καταστάσεων. Κάθε κατάσταση, επισημαίνεται με έναν δίσκο χαρακτηριστικού χρώματος και μεγέθους ανάλογου τού αριθμού των κυττάρων που ανήκουν σε αυτήν. Μεταξύ των καταστάσεων που υπάρχει έστω κι ένα κύτταρο με τις δύο μέγιστες εκ των υστέρων πιθανότητές του να αντιστοιχούν σε αυτές, προστίθεται μία ακμή, με μέγεθος ανάλογο του αθροίσματος των λόγων των κυττάρων κάθε κατάστασης που συμμετέχουν σε αυτήν τη μετάβαση. Η τιμή αυτή, συνεπώς, βρίσκεται στο διάστημα (0,2], κι αναγράφεται κατά μήκος τού τμήματος που συνδέει τις δύο καταστάσεις. Τα τμήματα των ακτίνων σε κάθε δίσκο, στην προέκταση του τμήματος που ενώνει τους δίσκους της μετάβασης, έχουν μήκος ανάλογο με τον λόγο των κυττάρων της κατάστασης που συμμετέχουν στη μετάβαση. Επίσης, μπορεί να τεθεί όριο (όπως εδώ το 0,2) για την ελάχιστη τιμή της τάσης μετάβασης προκειμένου να περιληφθεί η αντίστοιχη ακμή στο διάγραμμα.

Παράδειγμα δημιουργίας:

```
plotTransitions(MLscAN_obj)
```

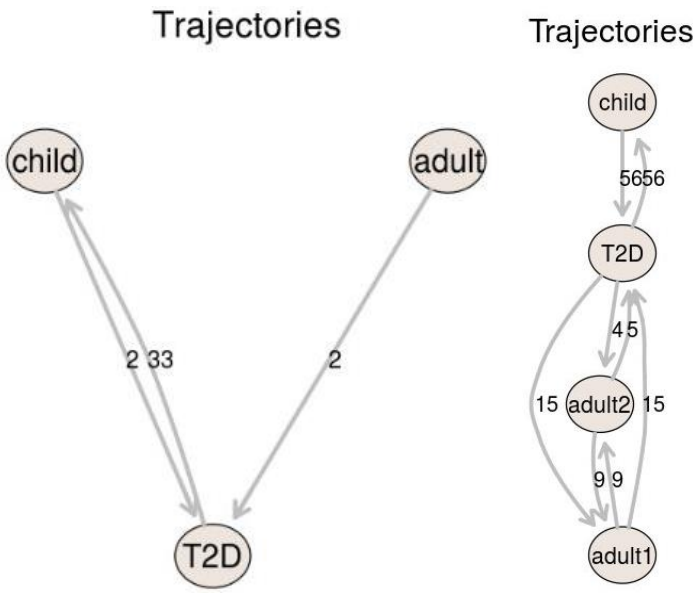


Εικόνα 3.22: Διαγράμματα των επιγενετικών τοπίων δύο προτύπων. Κάθε κατάσταση, επισημαίνεται με μία σφαίρα χαρακτηριστικού χρώματος και σταθερού μεγέθους. Μεταξύ των καταστάσεων που υπάρχει έστω κι ένα κύτταρο με τις δύο μέγιστες εκ των υστέρων πιθανότητές του να αντιστοιχούν σε αυτές, δημιουργείται ένα ζεύγος γραμμών (που αντιστοιχούν στο σχετικό ζεύγος μεταβάσεων) που συνδέει τις σχετικές σφαίρες. Κάθε γραμμή, σχετίζεται με μία κατάσταση του ζεύγους και τα κύτταρα που τοποθετούνται σε αυτήν, έχουν το χρώμα της κατάστασης αυτής. Όσο πιο κοντά βρίσκεται ένα κύτταρο της μετάβασης στη σφαίρα του ίδιου χρώματος, τόσο πιο μεγάλη είναι η εκ των υστέρων πιθανότητά του για την κατάσταση αυτήν.

Με αυτόν το διάγραμμα, γίνεται αντιληπτή η παρουσία των κυττάρων μίας τροχιάς σε κάθε στάδιο εξέλιξής της. Επειδή ενδέχεται περισσότερα του ενός κύτταρα να εντοπίζονται στην ίδια ή σε αρκετά κοντική θέση και να υπάρχη κάλυψη της παρουσίας τους, δεν μπορεί να χρησιμοποιηθεί για την εκτίμηση της αριθμητικής κατανομής τους.

Παράδειγμα δημιουργίας:

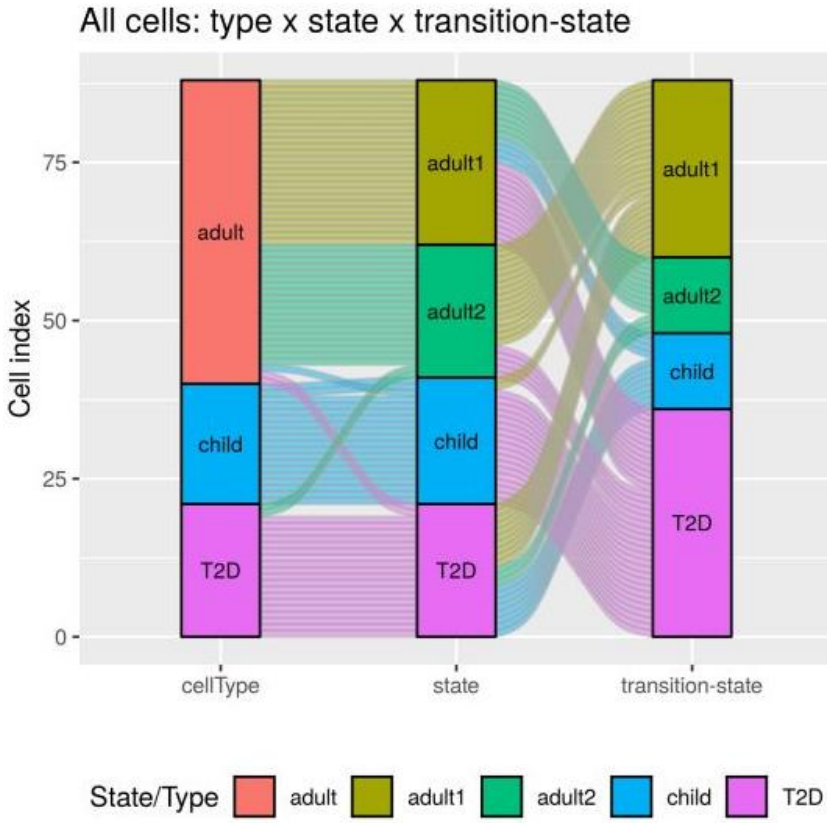
```
plotEpigenetic(MLscAN_obj)
```



Εικόνα 3.23: Διάγραμμα των τροχιών (κατευθυνόμενες ακμές) που αναγνωρίστηκαν ανάμεσα σε ζεύγη καταστάσεων (κόμβοι), με επισήμανση του αριθμού των κύριων γονιδίων τους.

Παράδειγμα δημιουργίας:

```
plotTrajectories(MLscAN_obj)
```

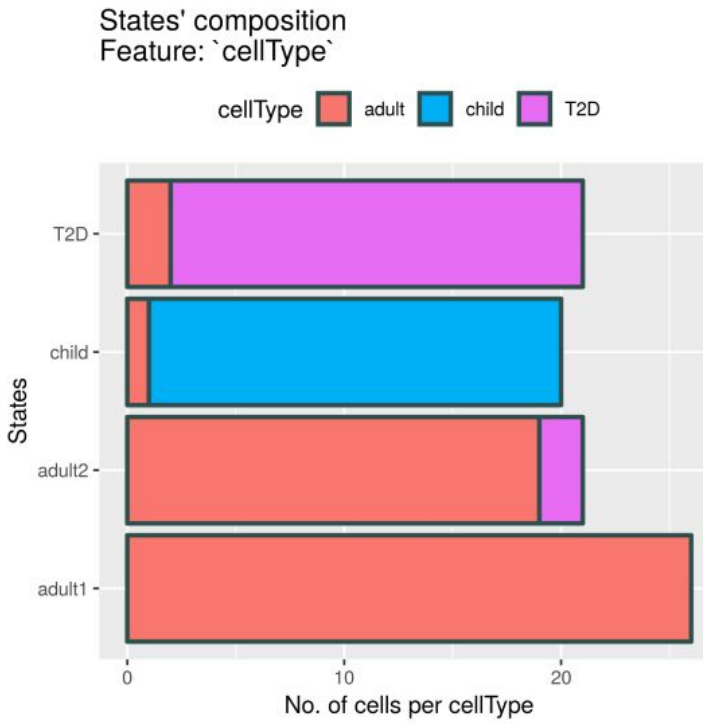


Εικόνα 3.24: Διάγραμμα που συνδέει κάθε κύτταρο με τα χαρακτηριστικά του στο πρότυπο MLscAN: τον κυτταρικό τύπο (αν είναι διαθέσιμη αυτή η πληροφορία), την κατάσταση που ανήκει και την κατάσταση μετάβασης (δηλ., την κατάσταση που αντιστοιχεί στη δεύτερη μεγαλύτερη εκ των υστέρων πιθανότητα).

Σε αυτό το διάγραμμα, είναι εύκολο να διακριθούν οι πληροφορίες που αφορούν στην πορεία των κυττάρων μίας κατάαστασης στις σχηματισμένες καταστάσεις, στη σύσταση των καταστάσεων από τα κύτταρα των διάφορων τύπων και στις μεταβάσεις που συμμετέχουν τα κύτταρα κάθε κατάστασης.

Παράδειγμα δημιουργίας:

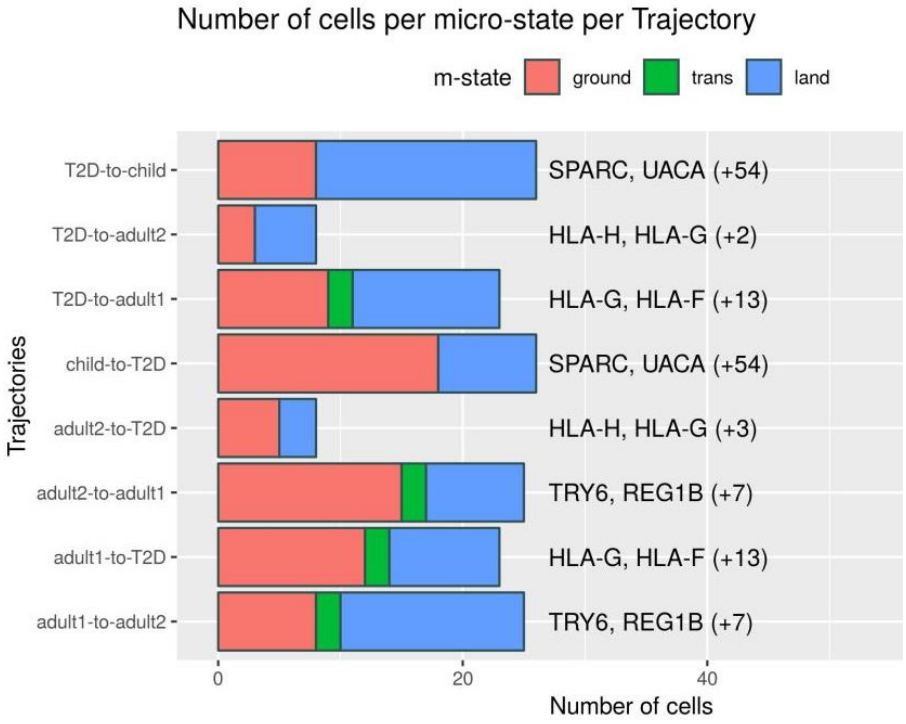
```
plotAlluvialState(MLscAN_obj)
```



Εικόνα 3.25: Ραβδόγραμμα της κυτταρικής σύστασης κάθε κατάστασης με βάση το επιλεγμένο χαρακτηριστικό των κυττάρων.

Παράδειγμα δημιουργίας:

```
plotStatesComposition(MLscAN_obj, feature="cellType")
```

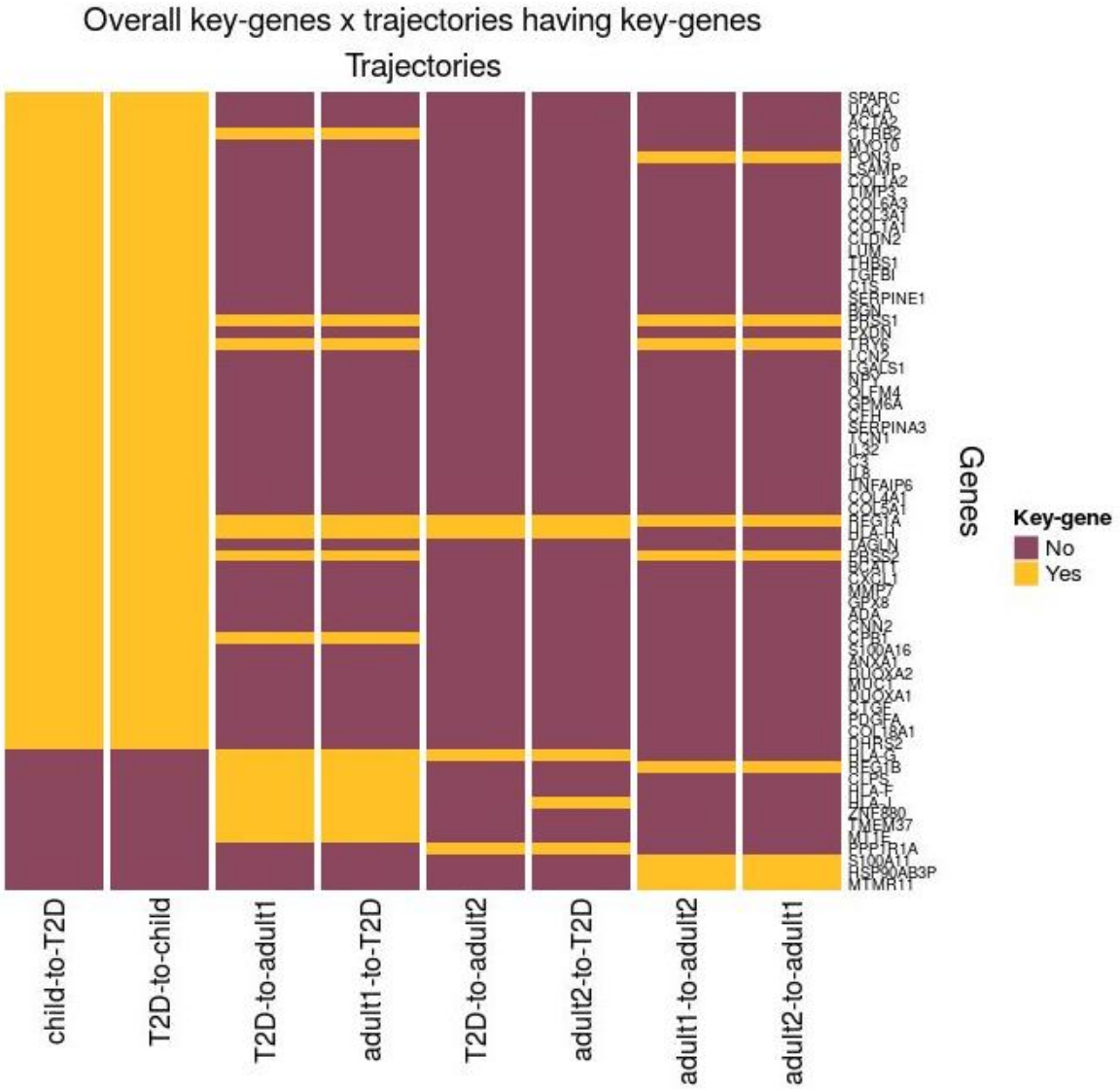


Εικόνα 3.26: Διάγραμμα των μικρο-καταστάσεων ανά τροχιά. Για κάθε τροχιά που αναγνωρίστηκε, απεικονίζεται ο αριθμός των κυττάρων ανά μικρο-κατάσταση, μαζί με τα κύρια

γονίδια (εφόσον υπάρχουν). Εμφανίζεται, επίσης, το όνομα των δύο πρώτων (βάσει της σημαντικότητάς τους για την τροχιά) το πολύ κύριων γονιδίων, και σε παρένθεση το πλήθος των υπόλοιπων κύριων γονιδίων της τροχιάς.

Παράδειγμα δημιουργίας:

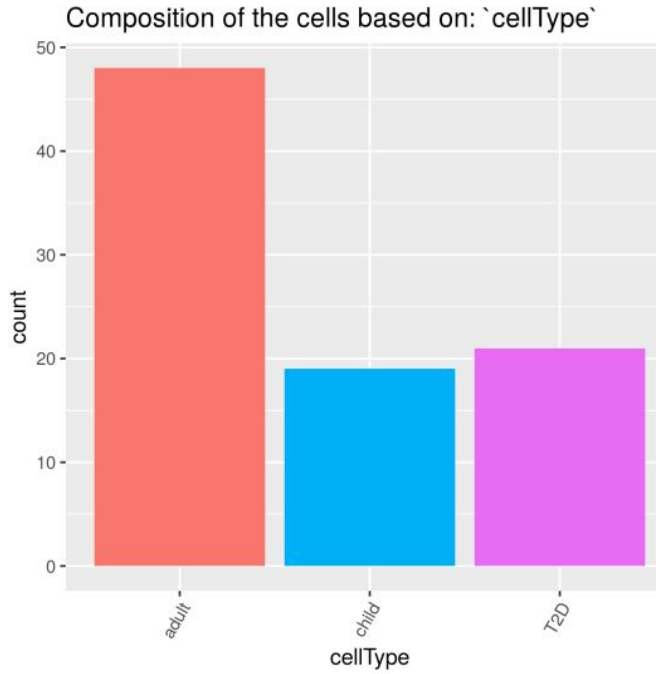
```
plotMStates(MLscAN_obj, mode="num")
```



Εικόνα 3.27: Χάρτης θερμότητας (heatmap) των γονιδίων που θεωρήθηκαν κύρια για τουλάχιστον μία τροχιά.

Παράδειγμα δημιουργίας:

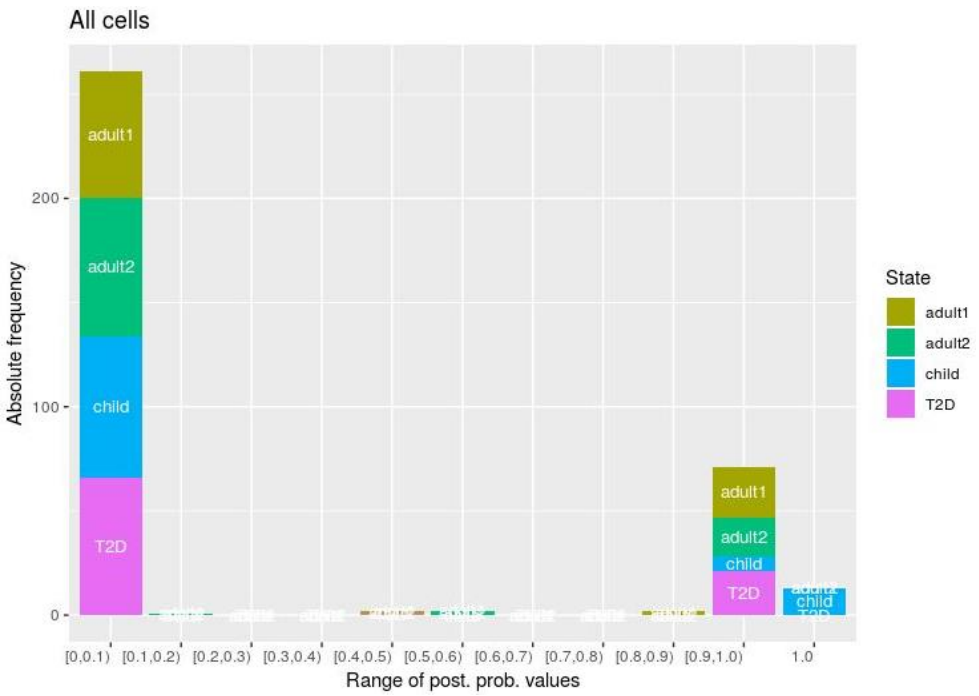
```
plotOverallKeyGenes(MLscAN_obj)
```

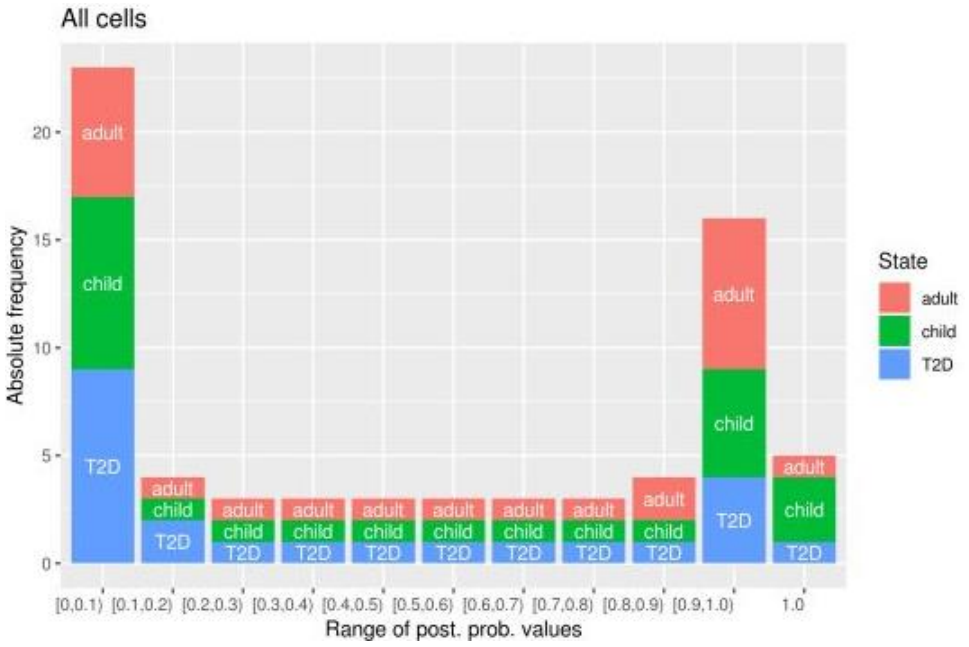


Εικόνα 3.28: Ιστόγραμμα της απόλυτης κάθε διακριτής τιμής του επιλεγμένου χαρακτηριστικού στο σύνολο των κυττάρων / γονιδίων.

Παραδείγματα δημιουργίας:

```
plotCellFeature(MLscAN_obj, feature="cellType")
plotCellFeaturePie(MLscAN_obj, feature="cellType")
plotGeneFeature(MLscAN_obj, feature="geneGroup")
plotGeneFeaturePie(MLscAN_obj, feature="geneGroup")
```





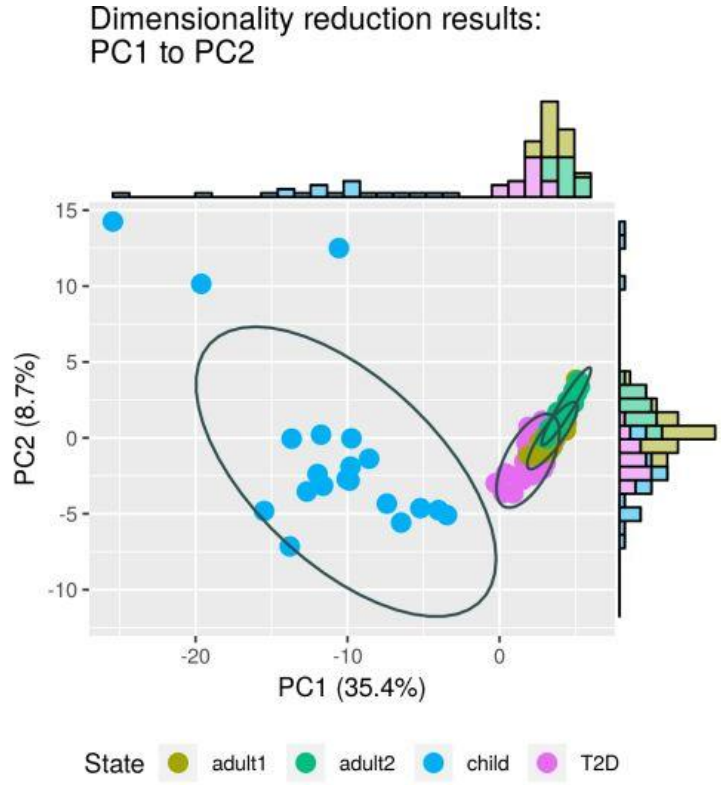
Εικόνα 3.29: Διάγραμμα της απόλυτης συχνότητας κάθε εύρους των εκ των υστέρων πιθανοτήτων ανά κατάσταση, για το σύνολο των κυττάρων.

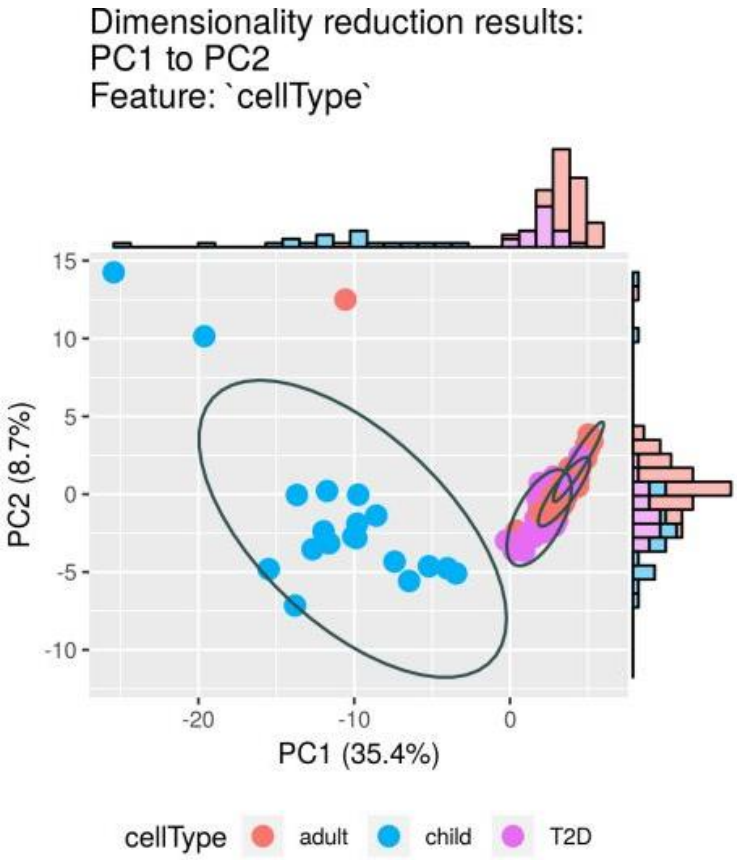
Παράδειγμα δημιουργίας:

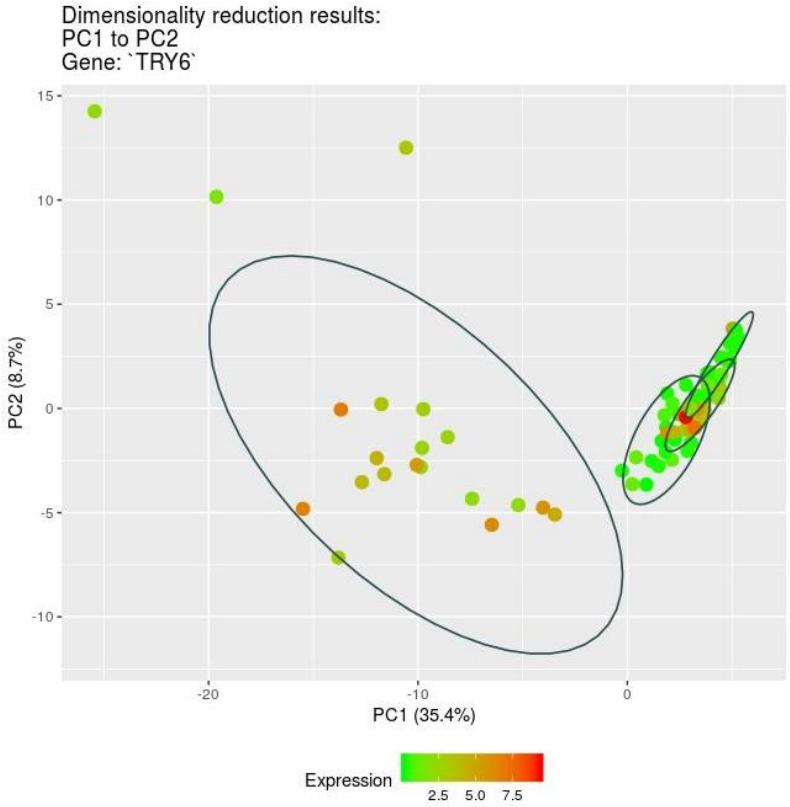
```
plotDecimalStates(MLscAN_obj)
```

3.2.2.3 Αποτελεσμάτων μείωσης της διαστατικότητας (DimRed_plots)

Στην ενότητα αυτή, περιλαμβάνονται διαγράμματα σχετικά με τα αποτελέσματα μείωσης της διαστατικότητας.


Dimensionality reduction results:
PC1 to PC2
15
10
5
0
-5
-10
PC2 (8.7%)
-20
-10
0
PC1 (35.4%)
State
adult1
adult2
child
T2D

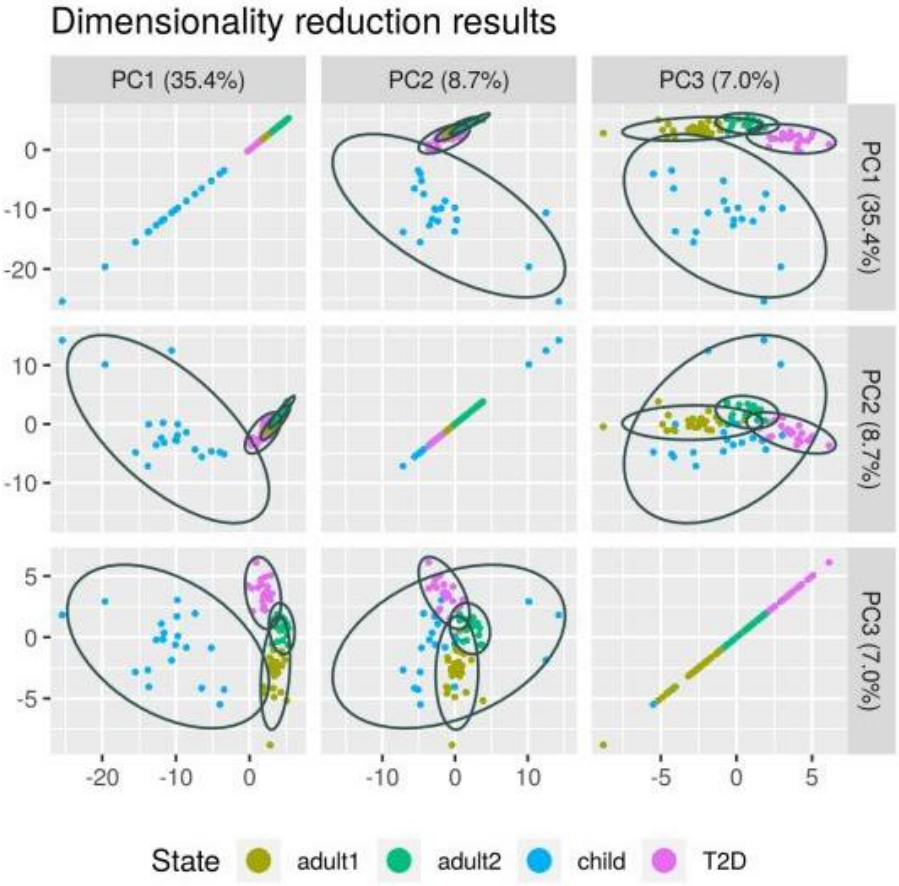

Dimensionality reduction results:
PC1 to PC2
Feature: `cellType`
15
10
5
0
-5
-10
PC2 (8.7%)
-20
-10
0
PC1 (35.4%)
cellType
adult
child
T2D



Εικόνα 3.30: Διαγράμματα προβολής των κυττάρων στο επιλεγμένο ζεύγος συνιστωσών των αποτελεσμάτων μείωσης της διαστατικότητας. Οι ελλείψεις (προ-επιλογή: κατανομή t, 95% επίπεδο εμπιστοσύνης), αντιστοιχούν στις κυτταρικές καταστάσεις σε όλες τις περιπτώσεις, αλλά, ο χρωματισμός των κυττάρων, εκτός από τις καταστάσεις, μπορεί να αφορά σε κάποιο από τα υπόλοιπα χαρακτηριστικά τους ή στην έκφραση επιλεγμένου γονιδίου. Με βάση το χαρακτηριστικό των κυττάρων, προκύπτουν και τα ιστογράμματα ανά διάσταση.

Παραδείγματα δημιουργίας:

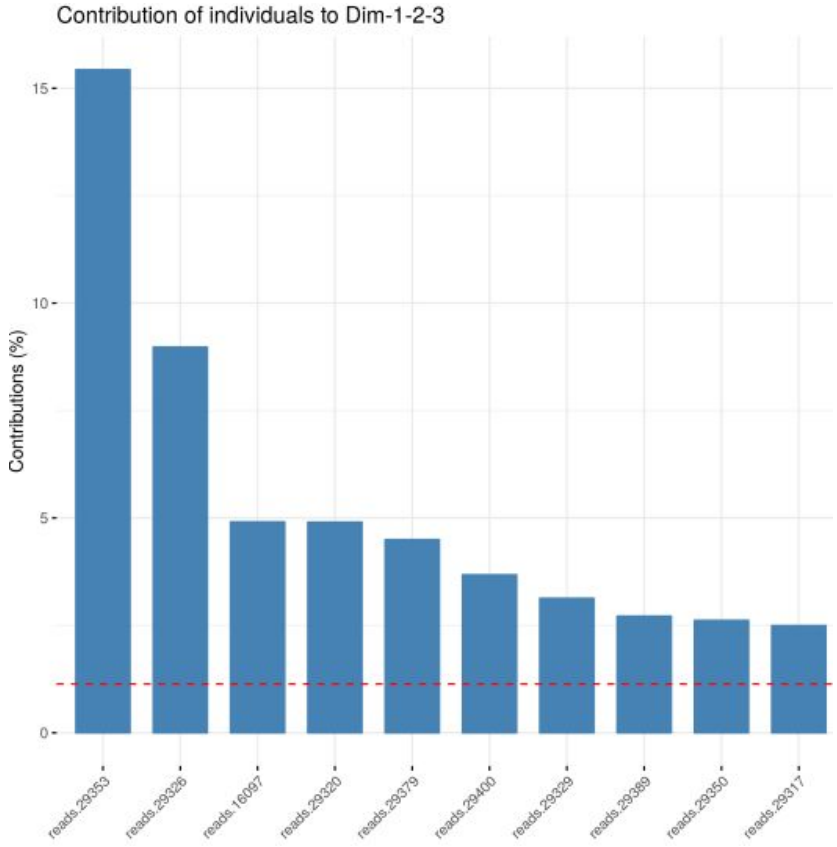
```
plotDimRed(MLscAN_obj, dim1="PC1", dim2="PC2")
plotDimRed(MLscAN_obj, dim1="PC1", dim2="PC2", feature="cellType")
plotDimRed(MLscAN_obj, dim1="PC1", dim2="PC2", gene="TRY6")
```



Εικόνα 3.31: Διάγραμμα προβολής των κυττάρων στις επιλεγμένες συνιστώσες των αποτελεσμάτων μείωσης της διαστατικότητας ανά δύο. Οι ελλείψεις (προ-επιλογή: κατανομή t, 95% επίπεδο εμπιστοσύνης), αντιστοιχούν στις κυτταρικές καταστάσεις σε όλες τις περιπτώσεις, αλλά, ο χρωματισμός των κυττάρων, εκτός από τις καταστάσεις, μπορεί να αφορά σε κάποιο από τα υπόλοιπα χαρακτηριστικά τους.

Παράδειγμα δημιουργίας:

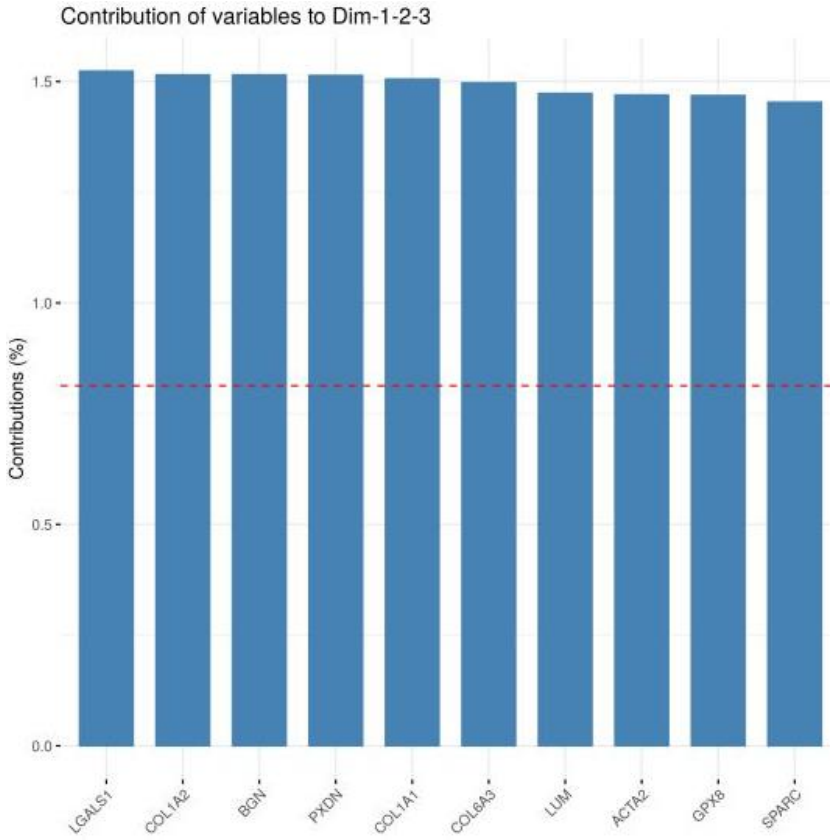
```
plotDimRedPairs(MLscAN_obj, dims=paste0("PC", seq(3)))
```



Εικόνα 3.32: Διάγραμμα των δέκα κυττάρων με τη μεγαλύτερη ποσοστιαία συνεισφορά (αθροιστικά) στη διακύμανση των τριών πρώτων κύριων συνιστωσών της PCA.

Παράδειγμα δημιουργίας:

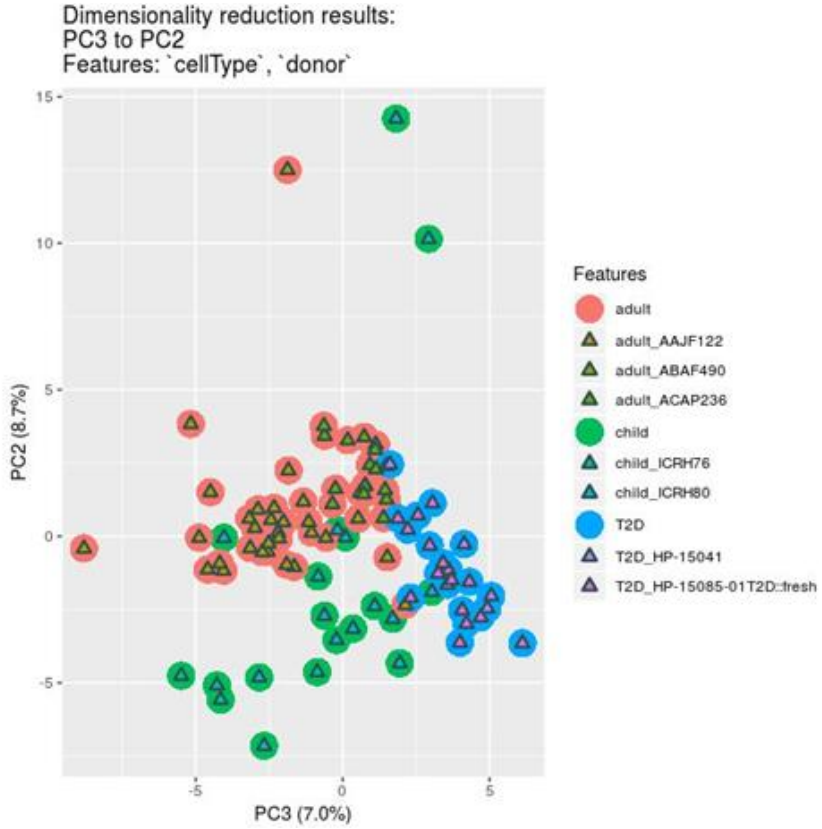
```
plotExprPCAVarInd(exprData(MLscAN_obj))
```



Εικόνα 3.33: Διάγραμμα των δέκα γονιδίων με τη μεγαλύτερη ποσοστιαία συνεισφορά (αθροιστικά) στη διακύμανση των τριών πρώτων κύριων συνιστωσών της PCA.

Παράδειγμα δημιουργίας:

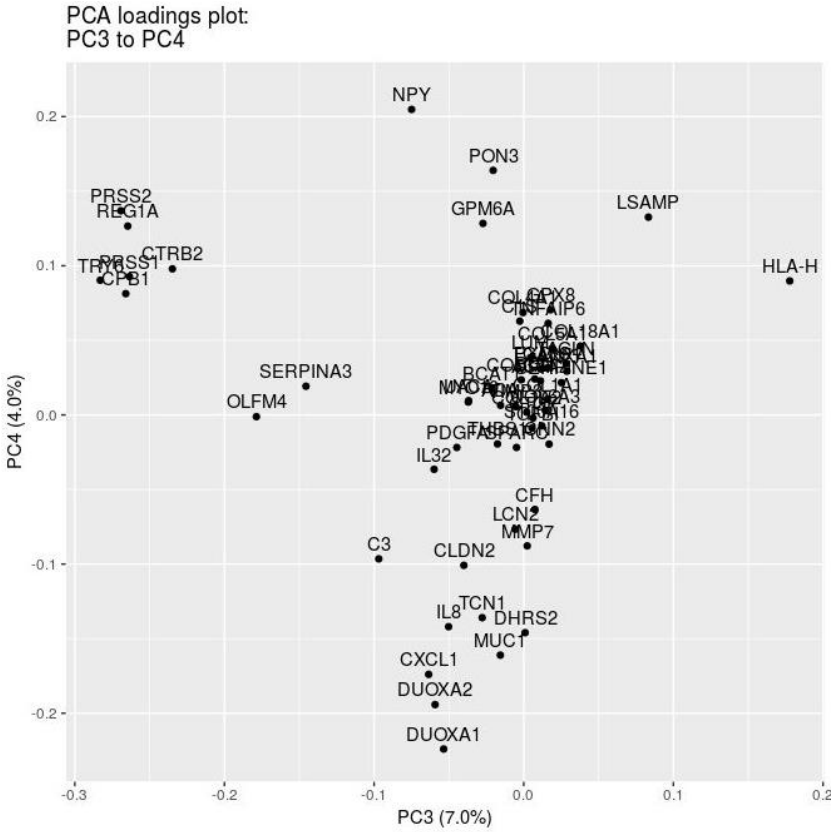
```
plotExprPCAVarInd(exprData(MLscAN_obj))
```



Εικόνα 3.34: Διάγραμμα προβολής των κυττάρων στο επιλεγμένο ζεύγος συνιστωσών των αποτελεσμάτων μείωσης της διαστατικότητας. Ο διττός χρωματισμός των σημείων – κυττάρων και το σχήμα τους, καθορίζονται από τις τιμές των δύο επιλεγμένων χαρακτηριστικών.

Παράδειγμα δημιουργίας:

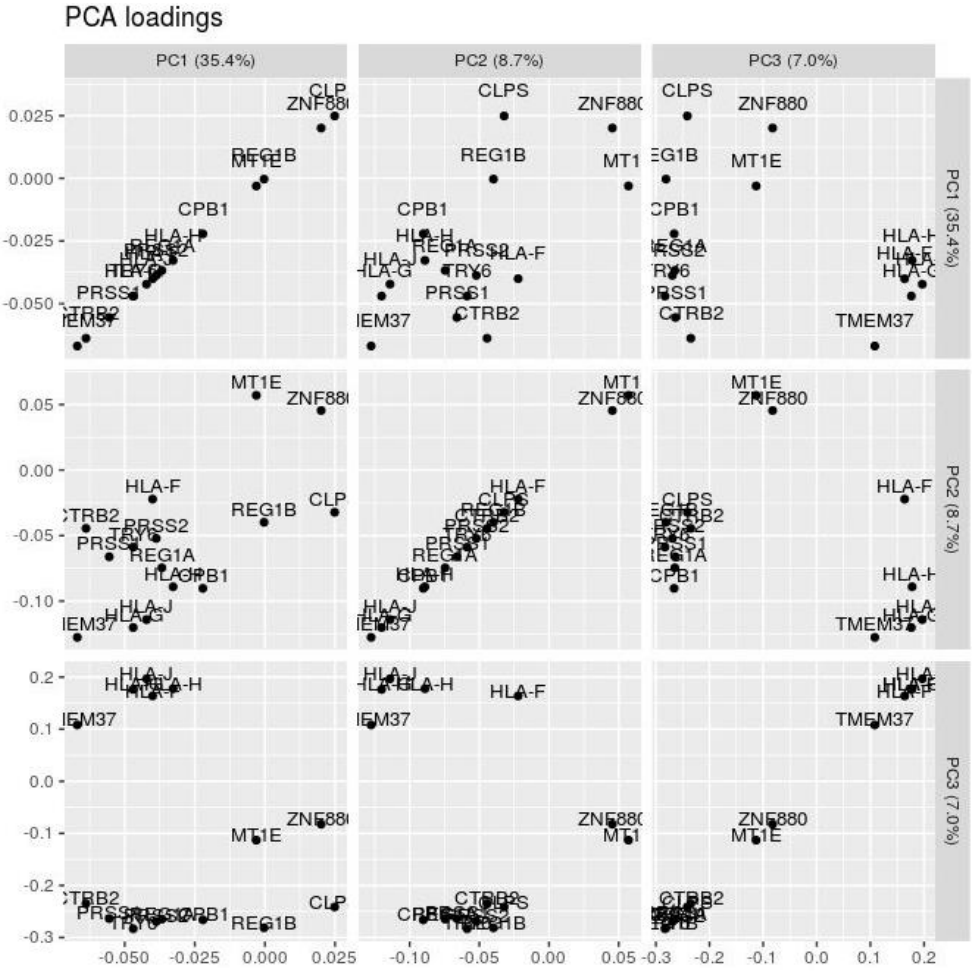
```
plotDimRed2Features(MLscAN_obj, dim1="PC2", dim2="PC3",
                    feature1="cellType", feature2="donor")
```



Εικόνα 3.35: Διάγραμμα προβολής των φορτώσεων (loadings) – των γονιδίων στο επιλεγμένο ζεύγος κύριων συνιστωσών της PCA.

Παράδειγμα δημιουργίας:

```
plotPCALoadings(MLscAN_obj, dim1="PC3", dim2="PC4",
                genes=keyGenes(MLscAN_obj, "adult1-to-adult2"))
```



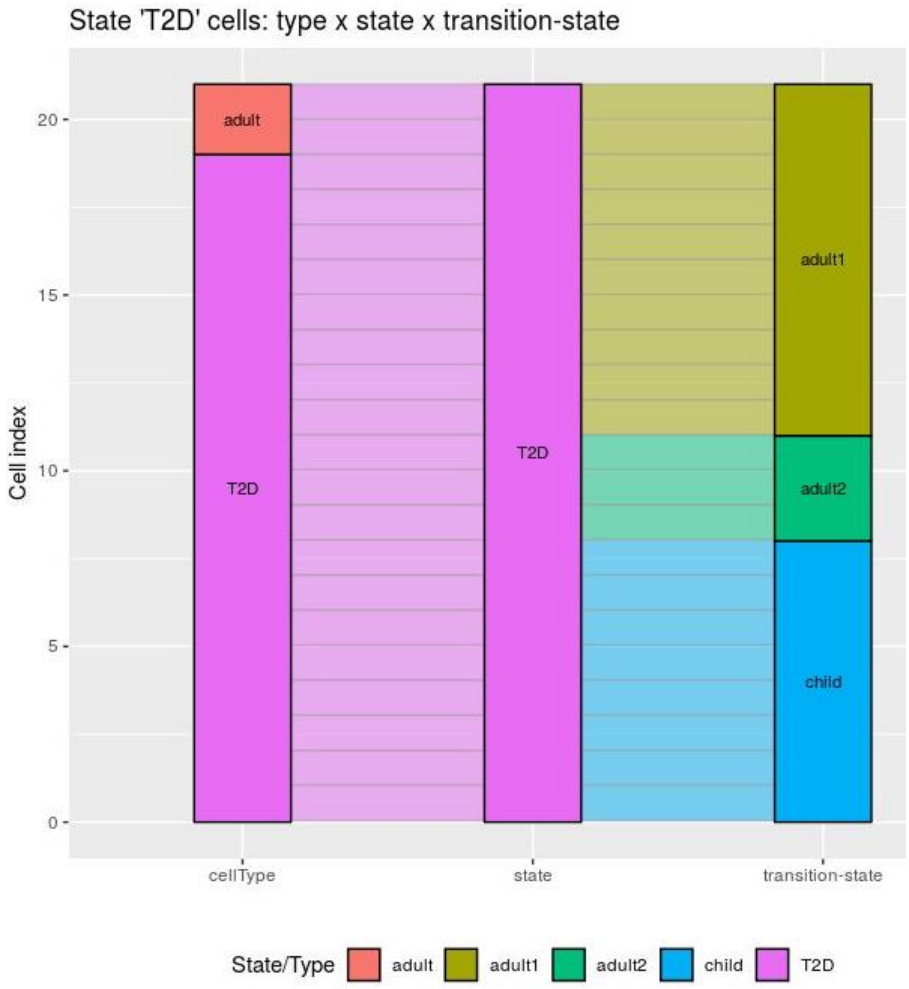
Εικόνα 3.36: Διάγραμμα προβολής των φορτώσεων (loadings) – των γονιδίων, των επιλεγμένων κύριων συνιστωσών της PCA ανά δύο.

Παράδειγμα δημιουργίας:

```
plotPCALoadingsPairs(MLscAN_obj, dims=paste0("PC", seq(3)),
                     genes=keyGenes(MLscAN_obj, "adult1-to-adult2"))
```

3.2.2.4 Διαγράμματα των καταστάσεων (States_plots)

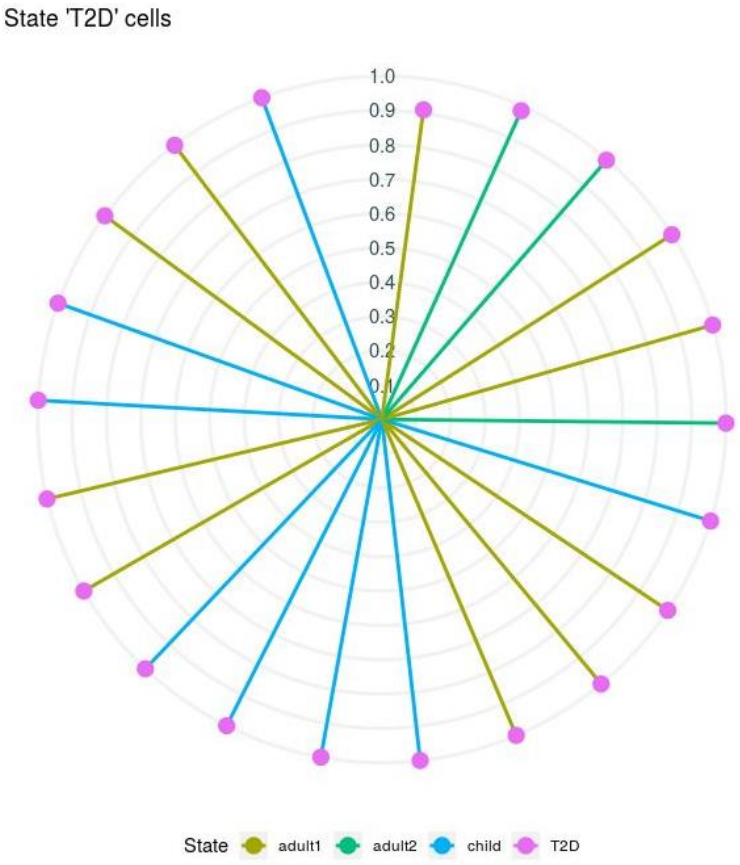
Εδώ, περιλαμβάνονται διαγράμματα που εστιάζονται σε μία κατάσταση.



Εικόνα 3.37: Διάγραμμα που συνδέει κάθε κύτταρο συγκεκριμένης κατάστασης με τα χαρακτηριστικά του στο πρότυπο MLscAN: τον κυτταρικό τύπο (αν είναι διαθέσιμη αυτή η πληροφορία), την κατάσταση που ανήκει και την κατάσταση μετάβασης (δηλ., την κατάσταση που αντιστοιχεί στη δεύτερη μεγαλύτερη εκ των υστέρων πιθανότητα).

Παράδειγμα δημιουργίας:

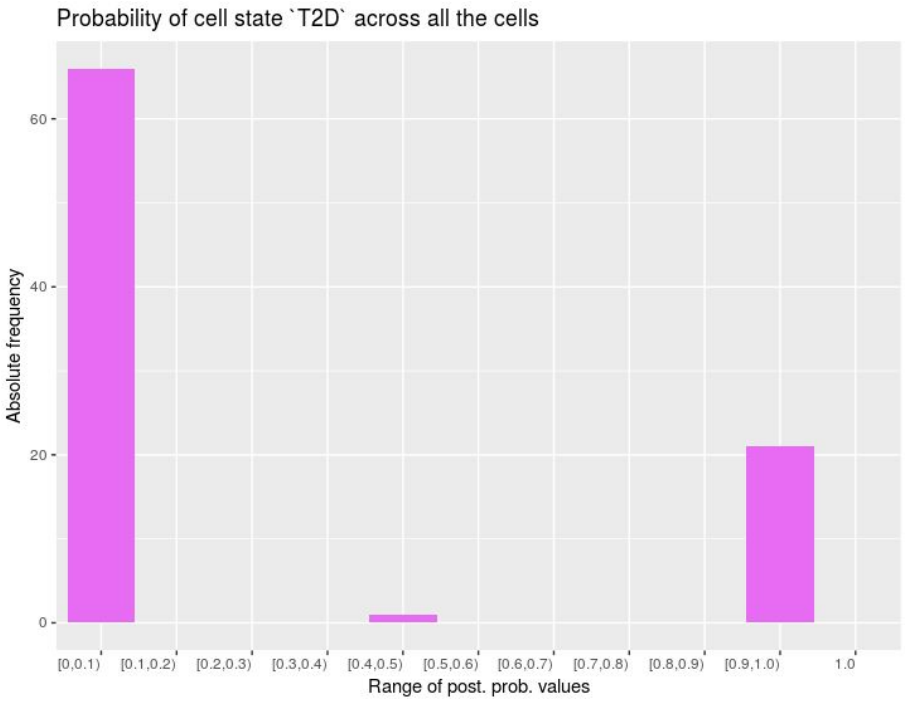
```
plotAlluvialState(MLscAN_obj, cellState="T2D")
```



Εικόνα 3.38: Στο διάγραμμα αυτό, οι σφαίρες, αντιστοιχούν στα κύτταρα που ανήκουν στην επιλεγμένη κατάσταση. Τοποθετούνται σε απόσταση από το κέντρο, ανάλογη της εκ των υστέρων πιθανότητας για την κατάσταση αυτήν. Οι γκρι ομόκεντροι κύκλοι βοηθούν στην αντίληψη της τιμής αυτής. Αντίστοιχα, τα κύτταρα διατάσσονται με φθίνοντα τρόπο βάσει της πιθανότητας αυτής, αντίστροφα από τη φορά των δεικτών του ρολογιού. Το χρώμα της ακτίνας, είναι αυτό της δεύτερης κατάστασης τής μετάβασης που συμμετέχουν.

Παράδειγμα δημιουργίας:

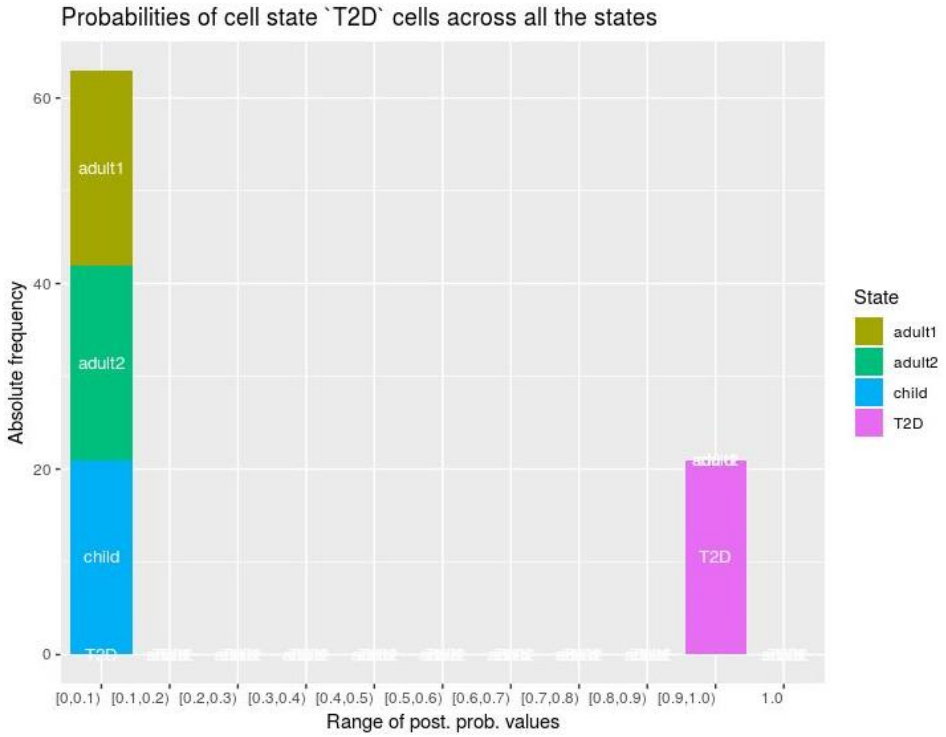
```
plotCircleState(MLscAN_obj, cellState="T2D")
```



Εικόνα 3.39: Διαγράμματα της απόλυτης συχνότητας των εκ των υστέρων πιθανοτήτων, σε κάθε εύρος, για την επιλεγμένη κατάσταση, λαβάνοντας υπόψη όλα τα κύτταρα (όχι μόνο όσα ανήκουν στην κατάσταση).

Παράδειγμα δημιουργίας:

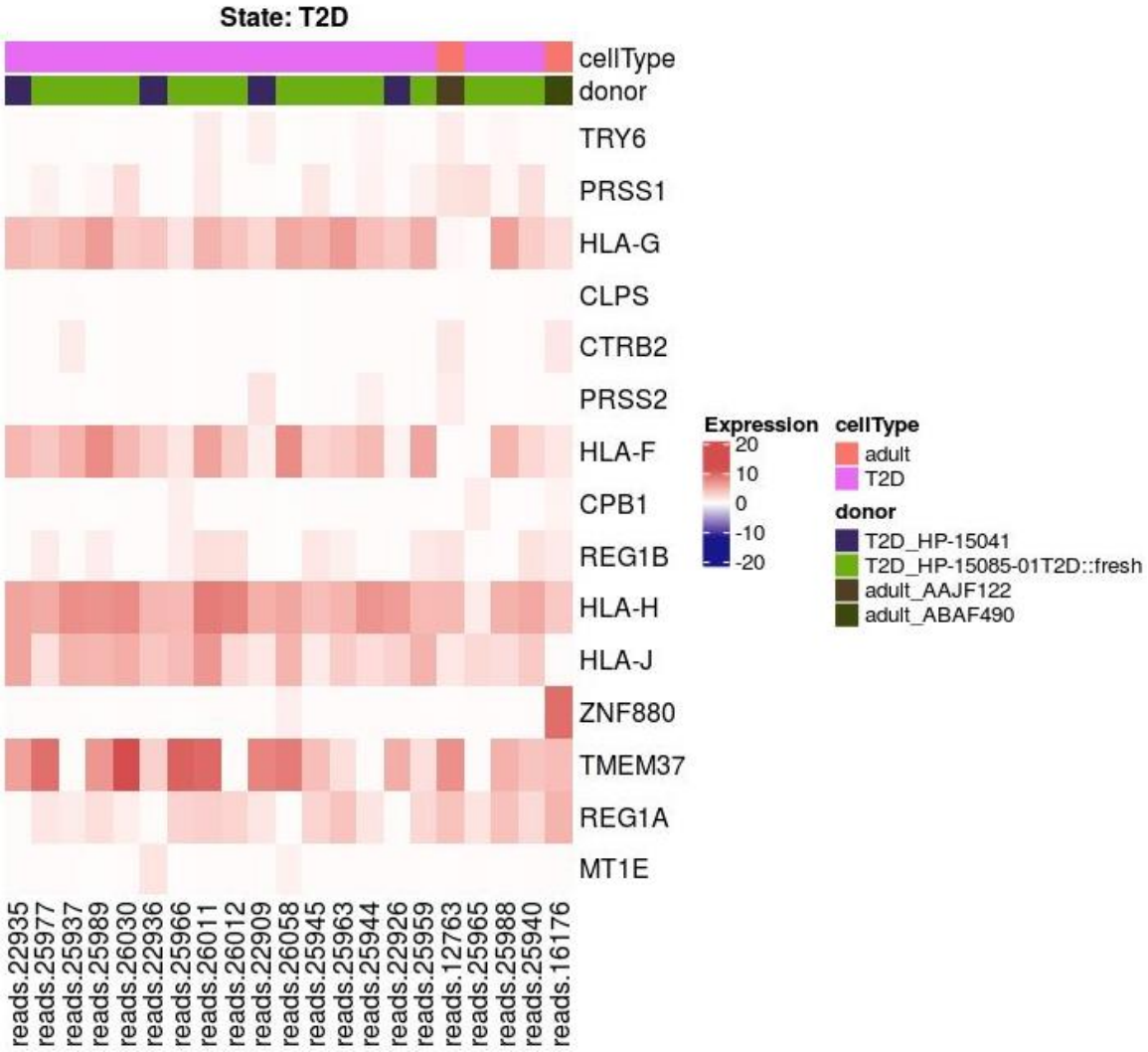
```
plotDecimalState(MLscAN_obj, cellState="T2D")
```

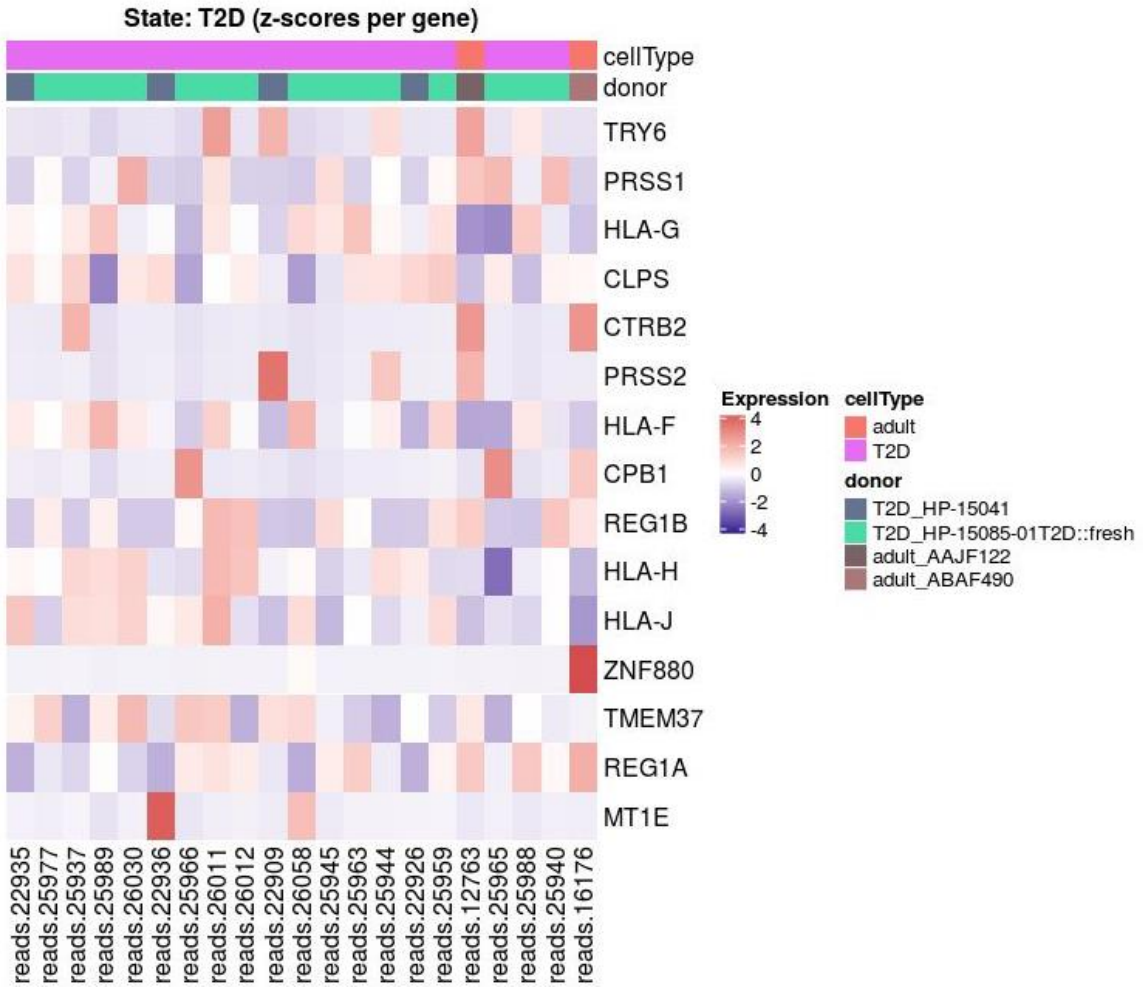


Εικόνα 3.40: Διαγράμματα της απόλυτης συχνότητας των εκ των υστέρων πιθανοτήτων, σε κάθε εύρος, για όλες τις καταστάσεις, λαβάνοντας υπόψη μόνο τα κύτταρα που ανήκουν στην επιλεγμένη κατάσταση.

Παράδειγμα δημιουργίας:

```
plotDecimalStateAllStates(MLscAN_obj, cellState="T2D")
```


State: T2D
cellType
donor
TRY6
PRSS1
HLA-G
CLPS
CTRB2
PRSS2
HLA-F
CPB1
REG1B
HLA-H
HLA-J
ZNF880
TMEM37
REG1A
MT1E
reads.22935
reads.25977
reads.25937
reads.25989
reads.26030
reads.22936
reads.25966
reads.26011
reads.26012
reads.22909
reads.26058
reads.25945
reads.25963
reads.25944
reads.22926
reads.25959
reads.12763
reads.25965
reads.25988
reads.25940
reads.16176
Expression
20
10
0
-10
-20
cellType
adult
T2D
donor
T2D_HP-15041
T2D_HP-15085-01T2D::fresh
adult_AAJF122
adult_ABAF490



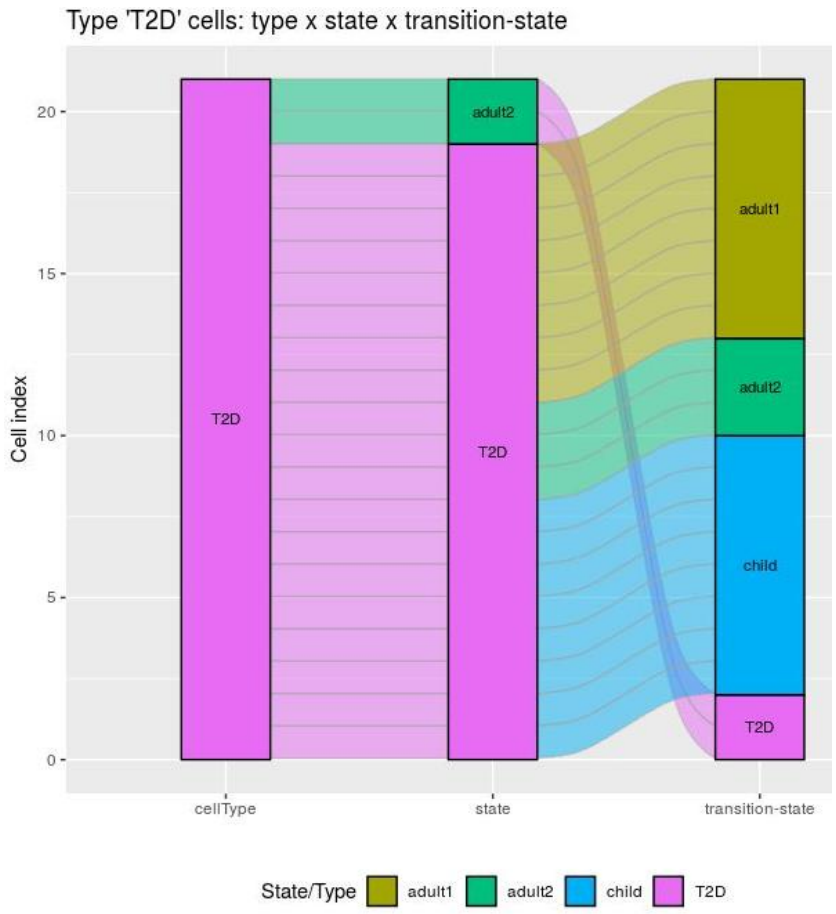
Εικόνα 3.41: Χάρτες θερμότητας (heatmap) των τιμών έκφρασης των γονιδίων ή των τυπικών τιμών (z-scores) τους ανά για όλα τα κύτταρα της επιλεγμένης κατάστασης και για τα επιλεγμένα γονίδια.

Παραδείγματα δημιουργίας:

```
plotHeatmapState(MLscAN_obj, state="T2D", z_scores=FALSE,
                 genes=keyGenes(MLscAN_obj, "adult1-to-T2D"),
                 features_cells=c("cellType", "donor"))

plotHeatmapState(MLscAN_obj, state="T2D",
                 genes=keyGenes(MLscAN_obj, "adult1-to-T2D"),
                 features_cells=c("cellType", "donor"))
```

3.2.2.5 Διαγράμματα των κυτταρικών τύπων (Types_plots)

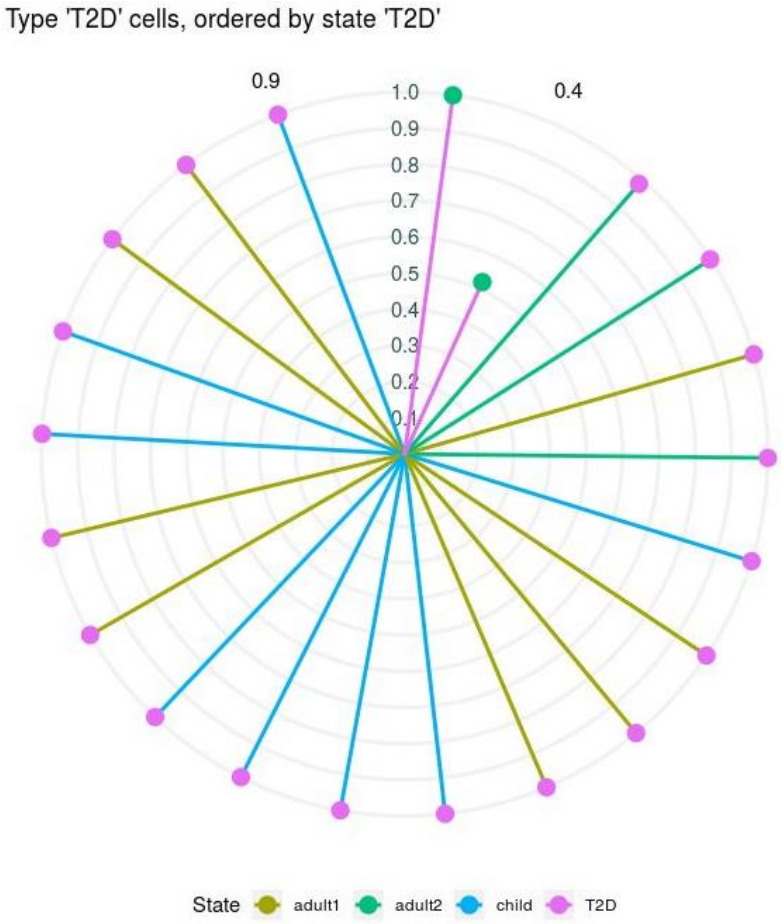
Εδώ, περιλαμβάνονται διαγράμματα που εστιάζονται σε έναν κυτταρικό τύπο.



Εικόνα 3.42: Διάγραμμα που συνδέει κάθε κύτταρο συγκεκριμένου τύπου (εκτός των ακραίων) με τα χαρακτηριστικά του στο πρότυπο MLscAN: τον κυτταρικό τύπο, την κατάσταση που ανήκει και την κατάσταση μετάβασης (δηλ., την κατάσταση που αντιστοιχεί στη δεύτερη μεγαλύτερη εκ των υστέρων πιθανότητα).

Παράδειγμα δημιουργίας:

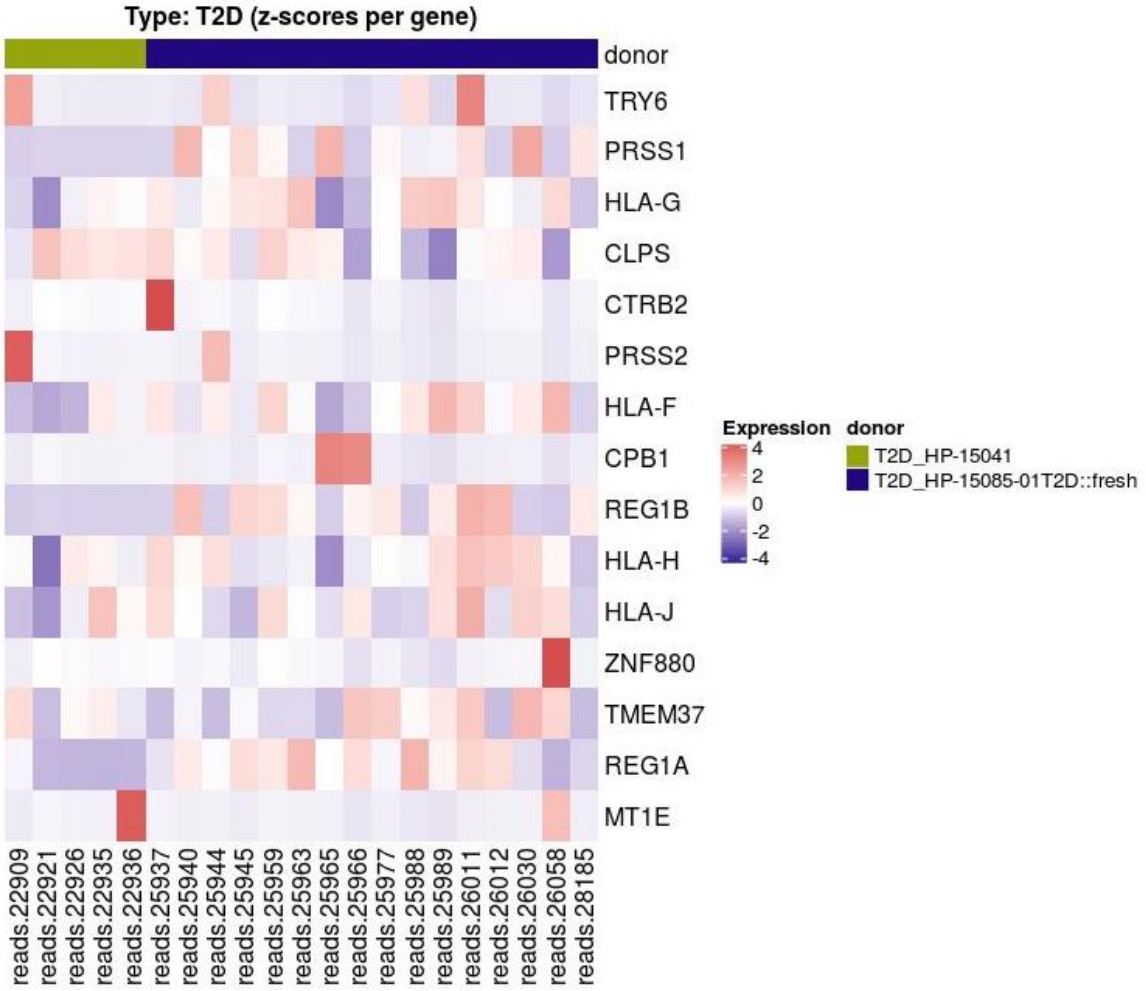
```
plotAlluvialState(MLscAN_obj, cellType="T2D")
```



Εικόνα 3.43: Στο διάγραμμα αυτό, οι σφαίρες, αντιστοιχούν στα κύτταρα του επιλεγμένου κυτταρικού τύπου και χρωματίζονται ανάλογα με την κατάσταση στην οποία ανήκουν. Τοποθετούνται σε απόσταση από το κέντρο, ανάλογη της εκ των υστέρων πιθανότητας για την κατάσταση στην οποία ανήκουν. Οι γκρι ομόκεντροι κύκλοι βοηθούν στην αντίληψη της τιμής αυτής. Τα κύτταρα, διατάσσονται με φθίνοντα τρόπο βάσει της πιθανότητας για την επιλεγμένη κατάσταση (αν δεν οριστεί, χρησιμοποιείται η κατάσταση που έχει το ίδιο όνομα με τον τύπο – εφόσον υπάρχει –), αντίστροφα από τη φορά των δεικτών του ρολογιού. Το χρώμα της ακτίνας, είναι αυτό της δεύτερης κατάστασης τής μετάβασης που συμμετέχουν. Εξωτερικά, στους κύκλους, επισημαίνονται οι θέσεις – κύτταρα όπου η τιμή της πιθανότητας για την επιλεγμένη κατάσταση αλλάζει ως προς το πρώτο δεκαδικό ψηφίο (δηλ., 0.9, 0.8, 0.7, κ.ο.κ.).

Παράδειγμα δημιουργίας:

```
plotCircleType(MLscAN_obj, cellType="T2D")
```



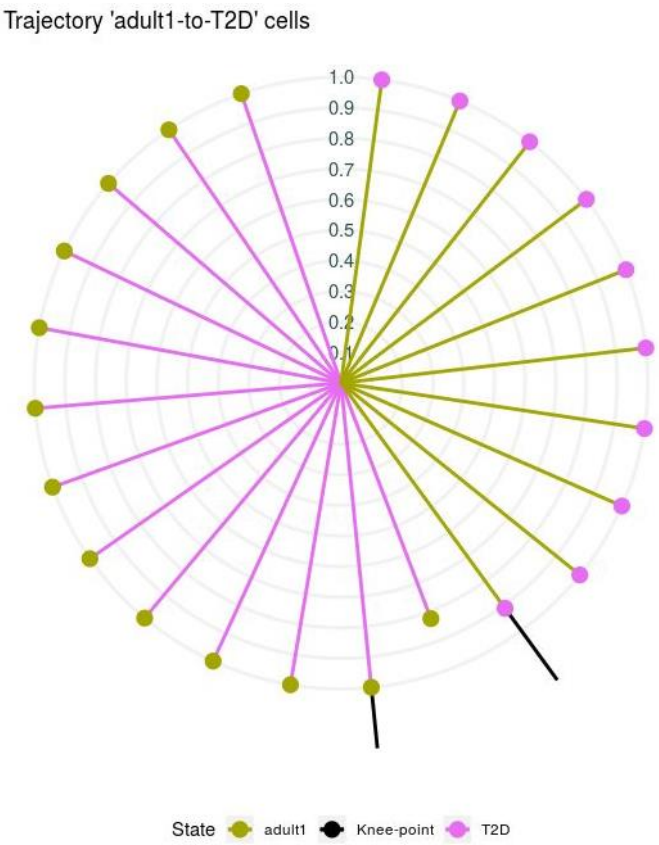
Εικόνα 3.44: Χάρτης θερμότητας (heatmap) της έκφρασης των κυττάρων του επιλεγμένου τύπου για τα επιλεγμένα γονίδια.

Παράδειγμα δημιουργίας:

```
plotHeatmapType(MLscAN_obj, type="T2D",
                keyGenes(MLscAN_obj, "adult1-to-T2D"),
                features_cells="donor")
```

3.2.2.6 Διαγράμματα των τροχιών (Traj_plots)

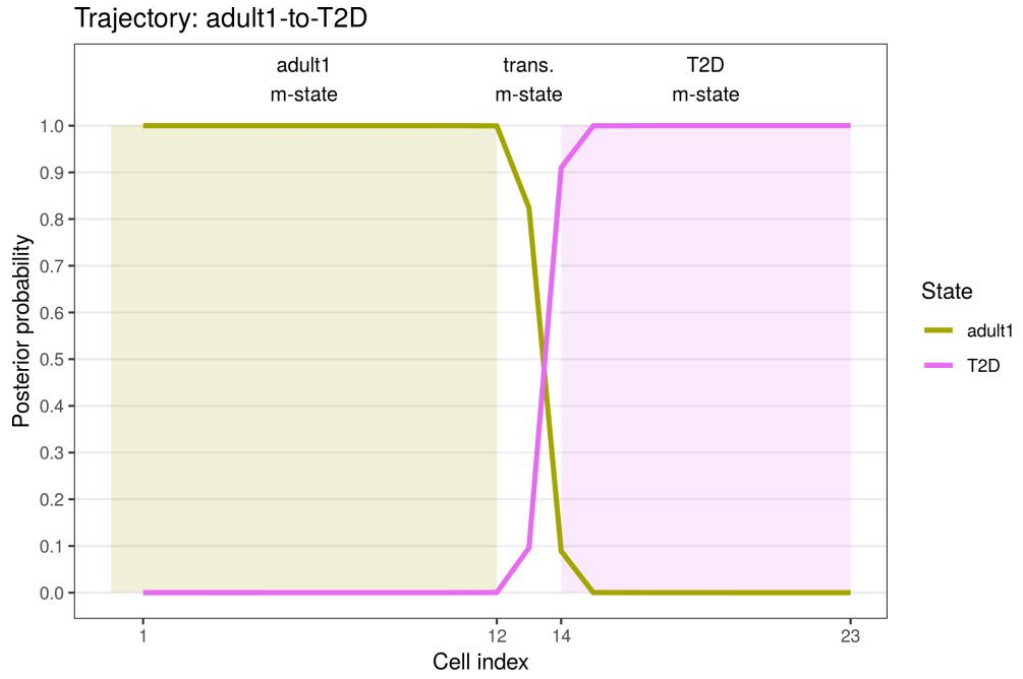
Εδώ, περιλαμβάνονται διαγράμματα που εστιάζονται σε μία τροχιά.



Εικόνα 3.45: Στο διάγραμμα αυτό, οι σφαίρες, αντιστοιχούν στα κύτταρα που συμμετέχουν στην επιλεγμένη τροχιά. Τοποθετούνται σε απόσταση από το κέντρο, ανάλογη της εκ των υστέρων πιθανότητας για την κατάσταση στην οποία ανήκουν. Οι γκρι ομόκεντροι κύκλοι βοηθούν στην αντίληψη της τιμής αυτής. Αντίστοιχα, τα κύτταρα διατάσσονται με φθίνοντα τρόπο βάσει της πιθανότητας για την κατάσταση έναρξης, αντίστροφα από τη φορά των δεικτών του ρολογιού (όπως, δηλαδή, διατάσσονται και στην τροχιά). Το χρώμα της ακτίνας, είναι αυτό της δεύτερης κατάστασης τής τροχιάς μετάβασης που συμμετέχουν.

Παράδειγμα δημιουργίας:

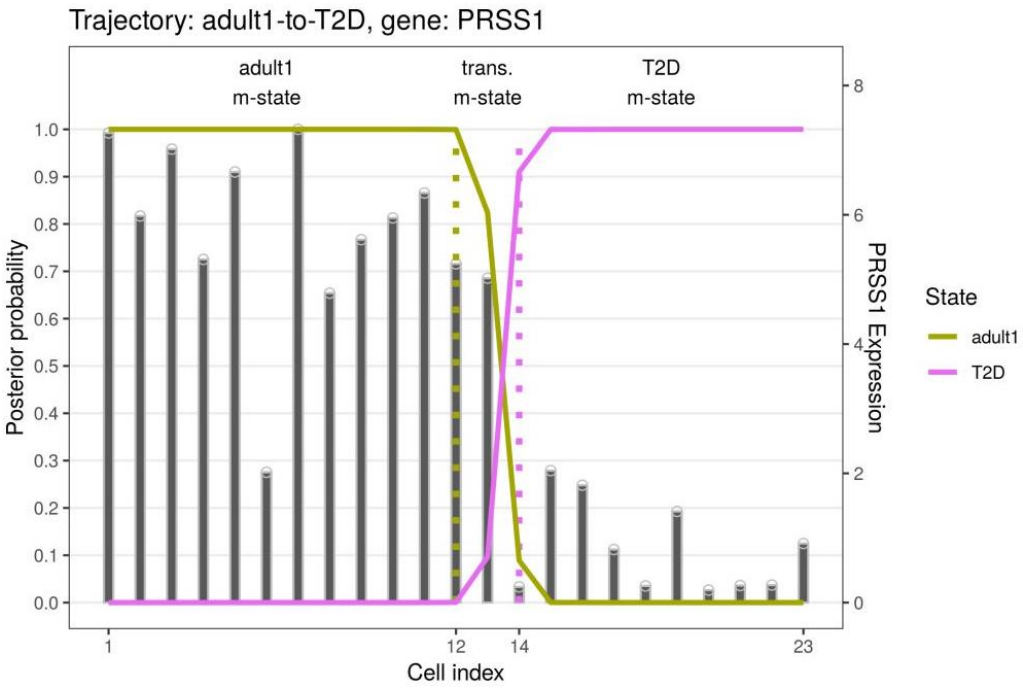
```
plotCircleTraj(MLscAN_obj, traj="adult1-to-T2D")
```



Εικόνα 3.46: Διάγραμμα των μεταβολών των εκ των υστέρων πιθανοτήτων για τις καταστάσεις της επιλεγμένης τροχιάς, διατηρώντας τη διάταξη των κυττάρων της τροχιάς κι επισημαίνοντας τις μικρο-καταστάσεις.

Παράδειγμα δημιουργίας:

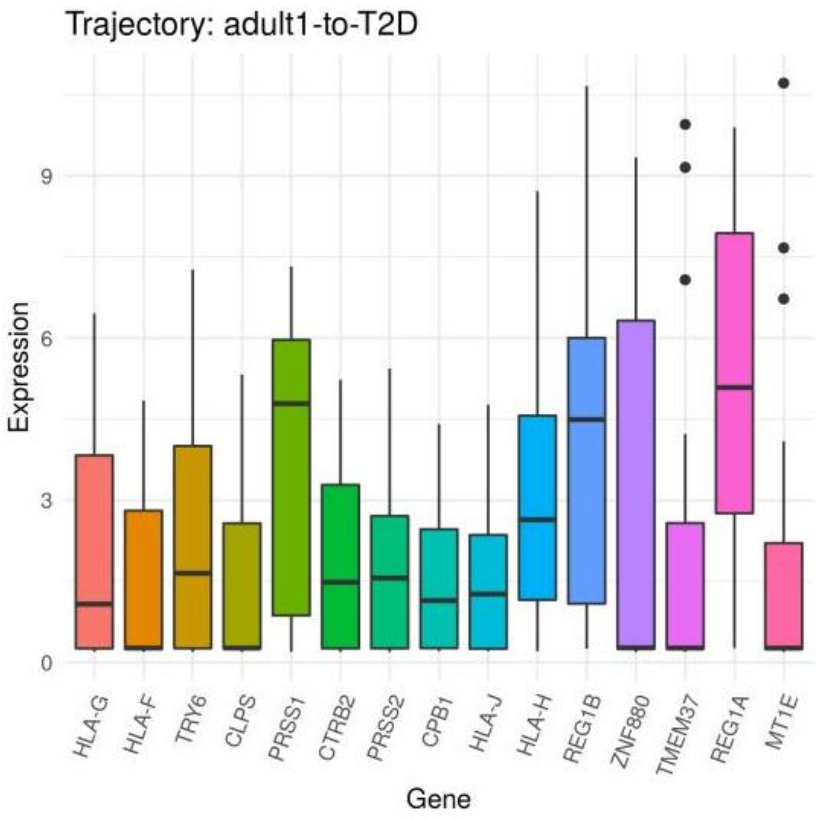
```
plotProbTraj(MLscAN_obj, traj="adult1-to-T2D")
```



Εικόνα 3.47: Διάγραμμα των μεταβολών των εκ των υστέρων πιθανοτήτων για τις καταστάσεις της επιλεγμένης τροχιάς και της έκφρασης ενός κύριου γονιδίου (ή άλλου επιλεγμένου γονιδίου), διατηρώντας τη διάταξη των κυττάρων της τροχιάς κι επισημαίνοντας τις μικρο-καταστάσεις.

Παράδειγμα δημιουργίας:

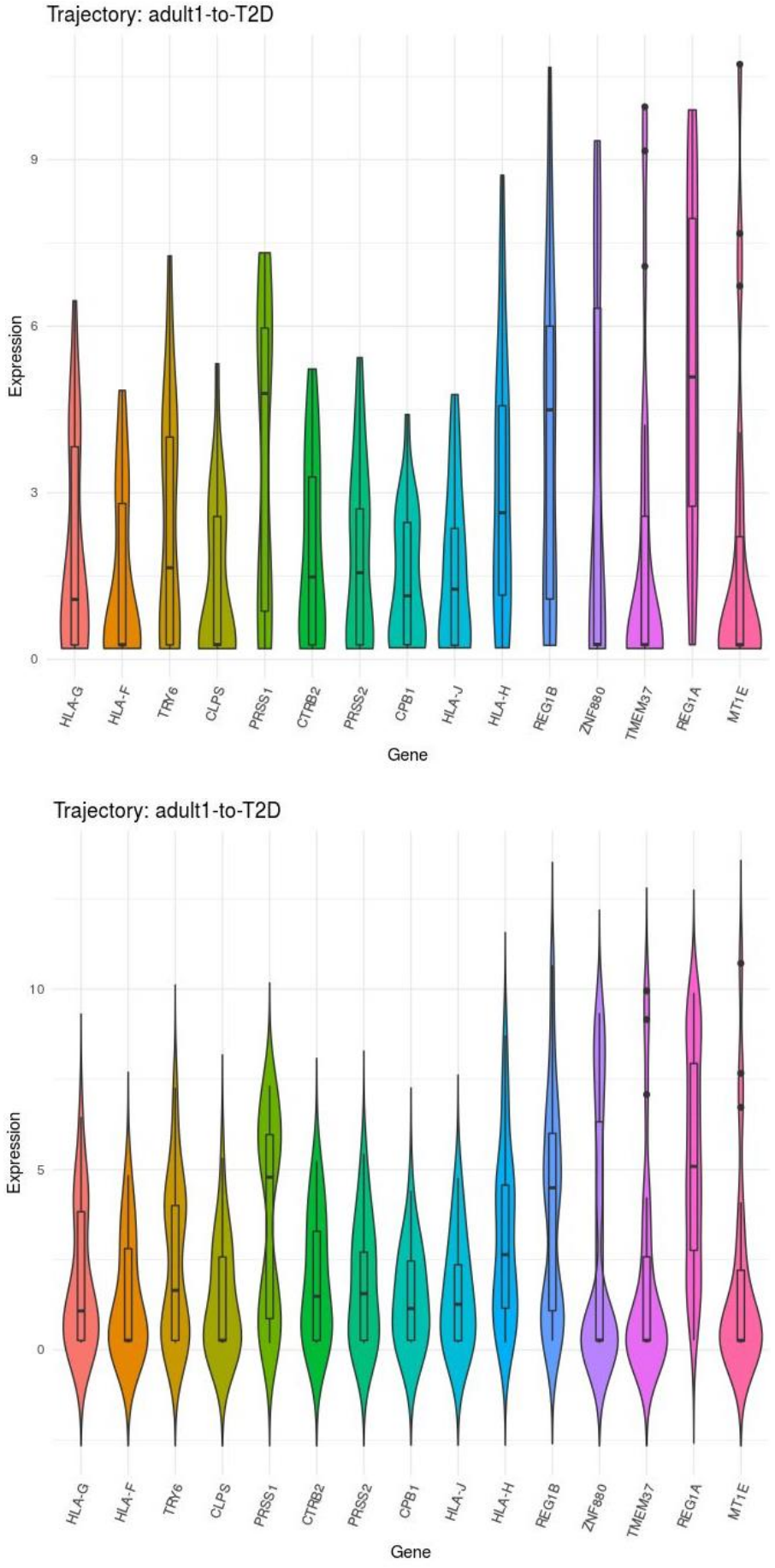
```
plotBarExprTraj(MLscAN_obj, traj="adult1-to-T2D", gene="PRSS1")
```



Εικόνα 3.48: Θηκογράμματα της έκφρασης των κύριων γονιδίων (ή άλλων επιλεγμένων γονιδίων) στα κύτταρα της επιλεγμένης τροχιάς.

Παράδειγμα δημιουργίας:

```
plotBoxplotTraj(MLscAN_obj, traj="adult1-to-T2D",
                genes=keyGenes(MLscAN_obj, traj="adult1-to-T2D"))
```



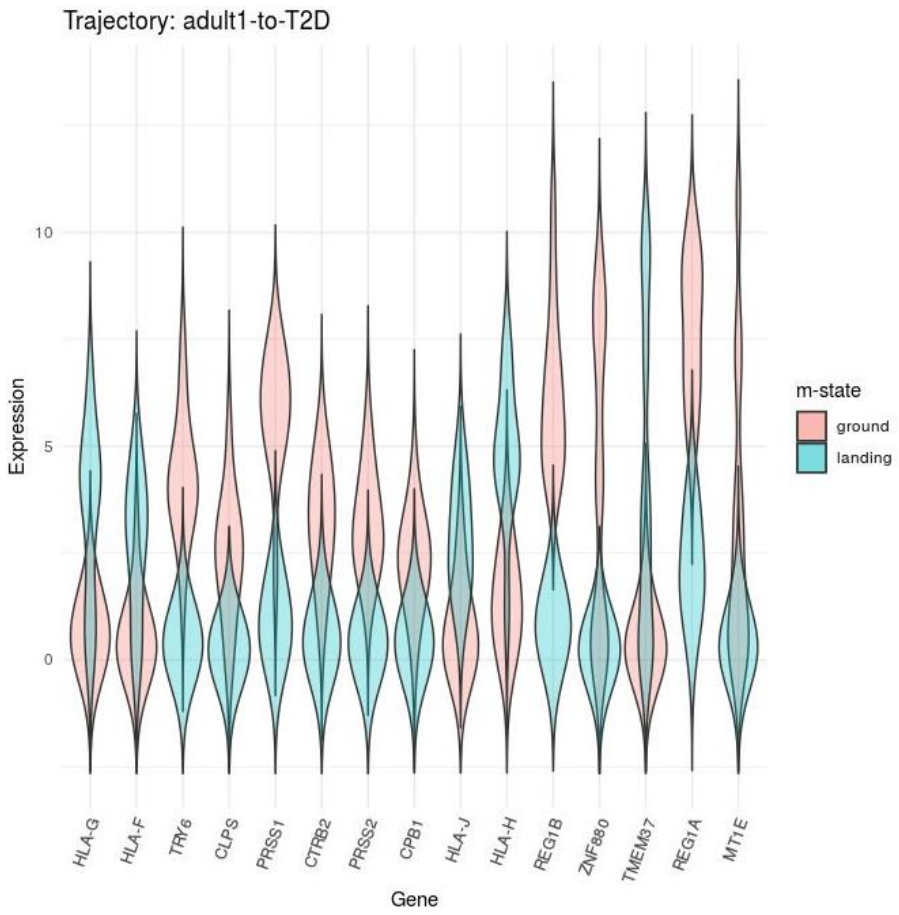
Εικόνα 3.49: Διαγράμματα βιολιού (violin plots) της έκφρασης των κύριων γονιδίων (ή άλλων επιλεγμένων γονιδίων) στα κύτταρα της επιλεγμένης τροχιάς. Εκτός από τη χρήση των

διαθέσιμων τιμών έκφρασης (πάνω διάγραμμα), μπορεί να επιλεγεί η εξομάλυνση τους (κάτω διάγραμμα), ειδικά όταν είναι μικρός ο αριθός των κυττάρων. Για την εξομάλυνση, χρησιμοποιείται ο πυρήνας Gauss με σταθερό εύρος ζώνης (0,95) για όλα τα γονίδια.

Η εξομάλυνση, είναι ιδιαίτερα βοηθητική όταν υπάρχουν λίγα κύτταρα, μιας και διαφορετικά, είναι διαθέσιμες λίγες, μη-συνεχείς παρατηρήσεις και δεν είναι εύκολα ορατές οι περιοχές συγκέντρωσης των τιμών έκφρασης.

Παραδείγματα δημιουργίας:

```
plotViolinTraj(MLscAN_obj, traj="adlt1-to-T2D",
               genes=keyGenes(MLscAN_obj, traj="adult1-to-T2D"))

plotViolinTrajSmooth(MLscAN_obj, traj="adlt1-to-T2D",
                     genes=keyGenes(MLscAN_obj, traj="adult1-to-T2D"))
```

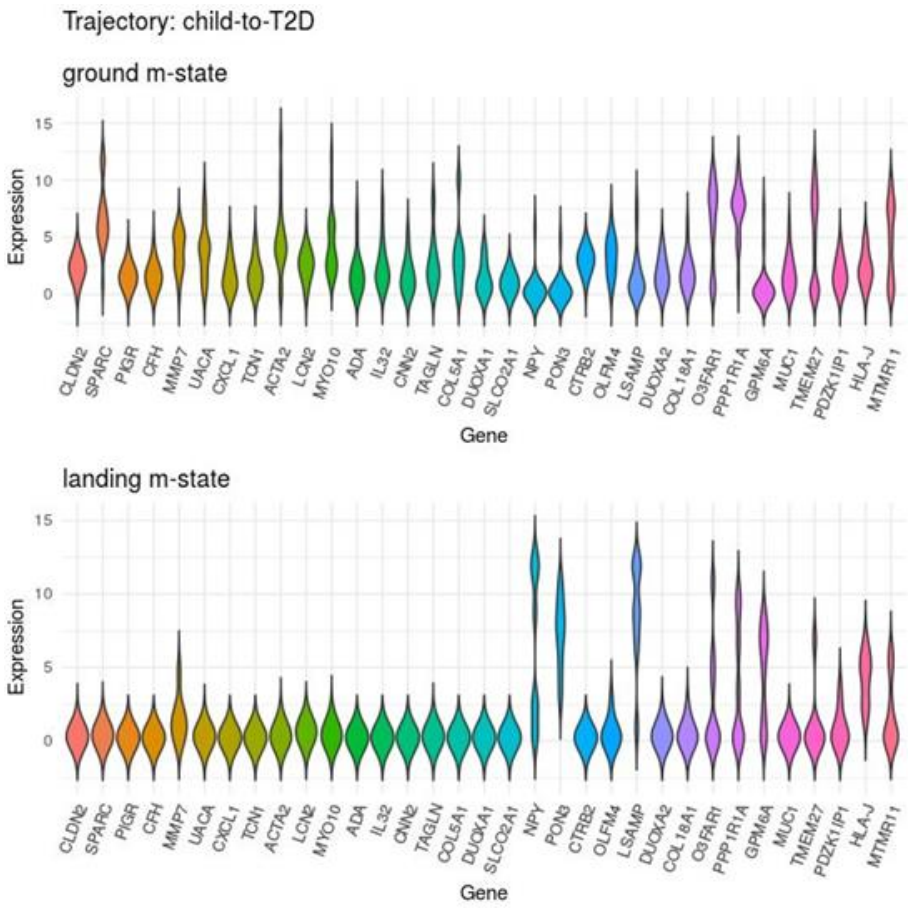


Εικόνα 3.50: Διάγραμμα βιολιού (violin plot) της έκφρασης των κύριων γονιδίων (ή άλλων επιλεγμένων γονιδίων) στα κύτταρα της μικρο-κατάστασης έναρξης και της μικρο-κατάστασης προορισμού της επιλεγμένης τροχιάς, με εξομάλυνση (πυρήνας Gauss, εύρος ζώνης: 0,95).

Σε αυτήν την περίπτωση, μπορεί πιο εύκολα να διακριθεί αν οι τιμές έκφρασης, των δύο ακραίων μικρο-καταστάσεων, εμφανίζουν κατά προσέγγιση διτροπική κατανομή.

Παραδείγματα δημιουργίας:

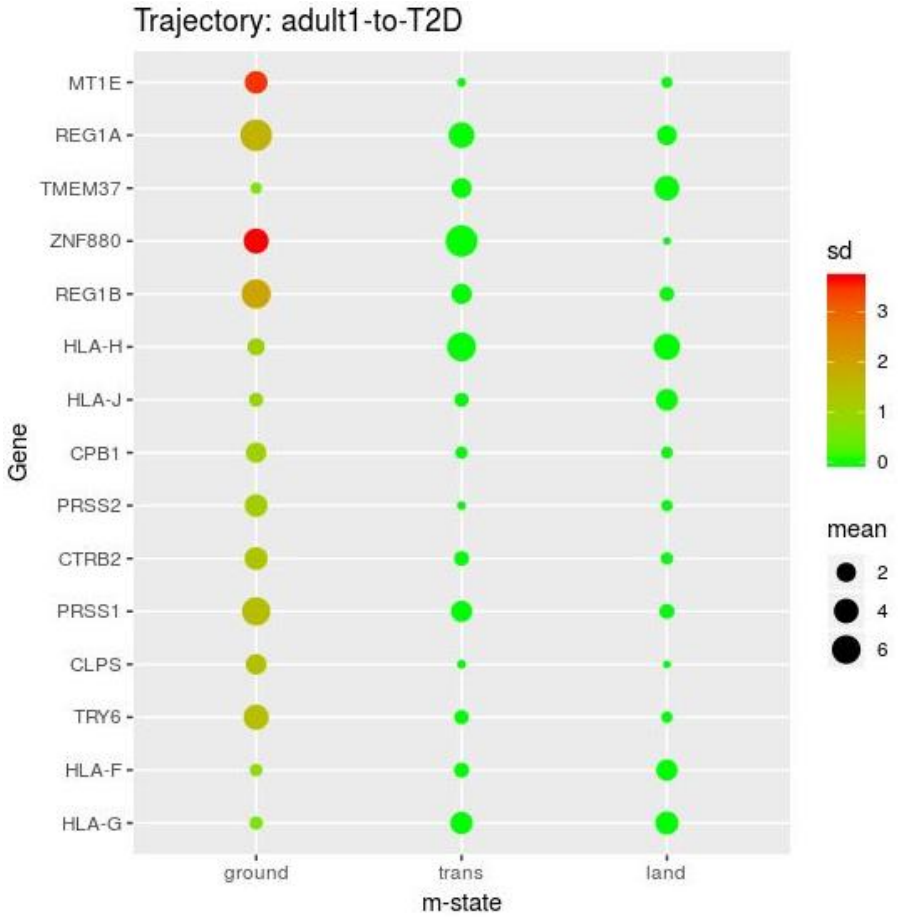
```
plotViolinOverlayTraj(MLscAN_obj, traj="adult1-to-T2D",
                      genes=keyGenes(MLscAN_obj, traj="adult1-to-T2D"))
```



Εικόνα 3.51: Διάγραμμα, αντίστοιχο με το κάτω διάγραμμα της εικόνα 3.49, με εξομάλυνση (πυρήνας Gauss, εύρος ζώνης: 0,95), όπου παρατίθενται τα αποτελέσματα ξεχωριστά για κάθε μικρο-κατάσταση της τροχιάς.

Παράδειγμα δημιουργίας:

```
plotViolinTrajMSSmooth(MLscAN_obj, traj="child-to-T2D",
                       genes=keyGenes(MLscAN_obj,
                                      traj="child-to-T2D")[seq(33)])
```

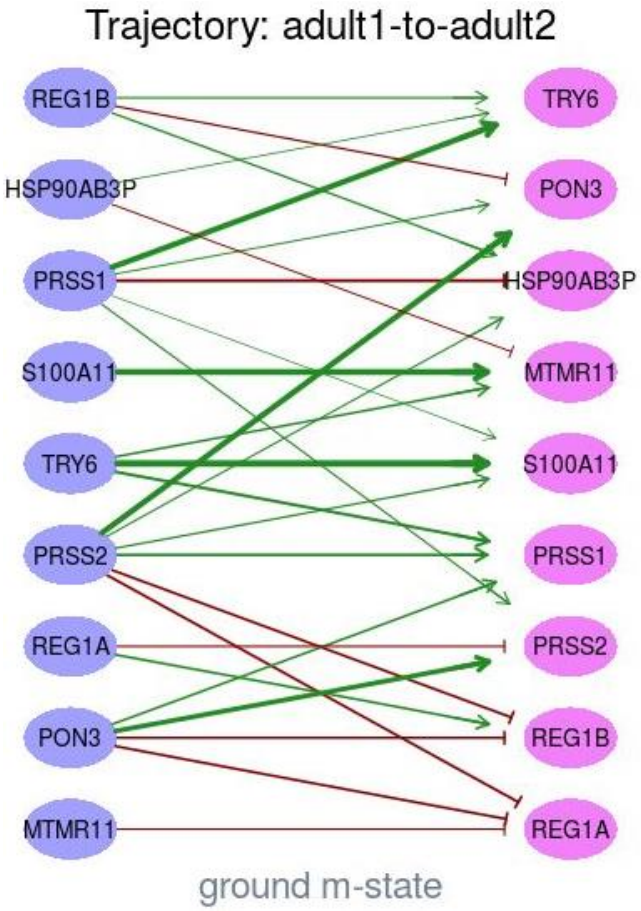


Εικόνα 3.52: Διάγραμμα της μέσης έκφρασης (μέγεθος των κύκλων) και της τυπικής απόκλισης (χρώμα των κύκλων) των κυττάρων, ανά κύριο γονίδιο (ή άλλο επιλεγμένο γονίδιο), για τα κύτταρα κάθε μικρο-κατάστασης της επιλεγμένης τροχιάς.

Στο διάγραμμα αυτό, επισημαίνονται οι αλλαγές στη μέση τιμή και το πόσο αυτή αντιπροσωπεύει την πλειονότητα των κυττάρων κάθε μικρο-κατάστασης.

Παράδειγμα δημιουργίας:

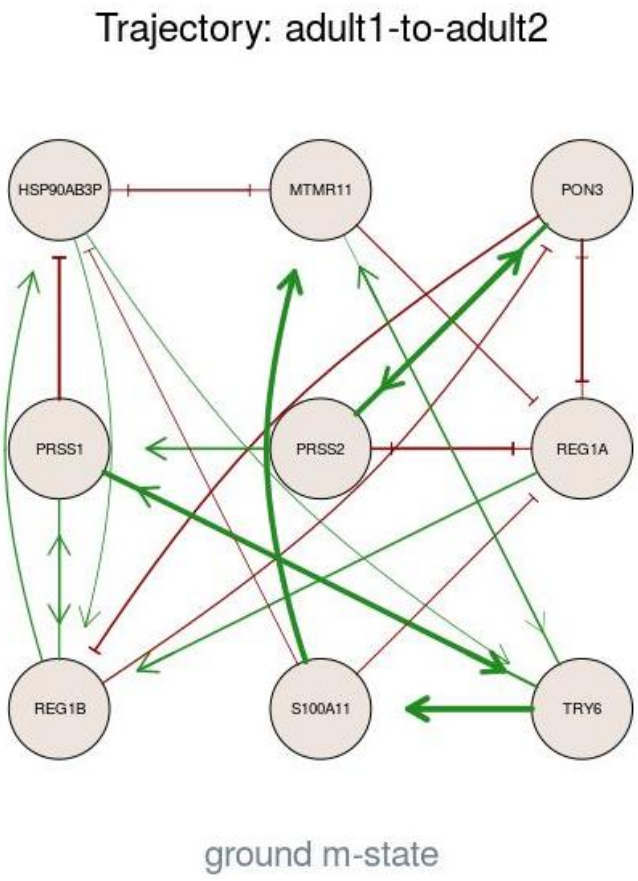
```
plotDotTraj(MLscAN_obj, traj="adult1-to-T2D",
            genes=keyGenes(MLscAN_obj, traj="adult1-to-T2D"))
```



Εικόνα 3.53: Διμερής γράφος του GRN της τροχιάς και της μικρο-κατάστασης που έχουν επιλεγεί. Οι κορυφές, αντιστοιχούν στα κύρια γονίδια. Το μέγεθος των κατευθυνόμενων ακμών, είναι ανάλογο του βάρους της αντίστοιχης αλληλεπίδρασης. Με κόκκινο χρώμα, επισημαίνεται κατασταλτική / ανασταλτική αλληλεπίδραση ενώ με πράσινο, ενισχυτική / ενεργοποιητική / διεγερτική.

Παράδειγμα δημιουργίας:

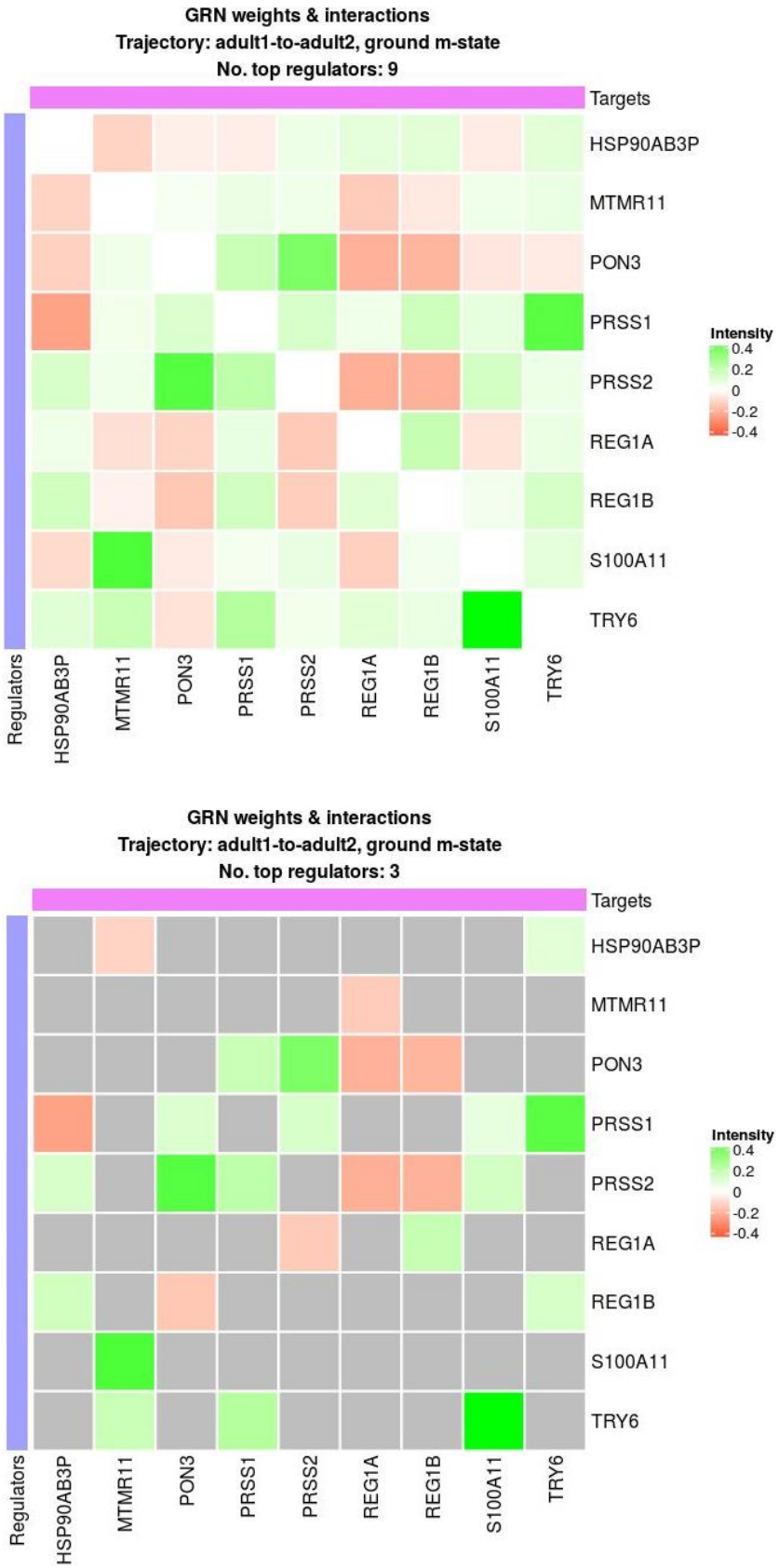
```
plotGRNBipartite(MLscAN_obj, traj="adult1-to-adult2", mstate="ground")
```



Εικόνα 3.54: Γράφος του GRN της τροχιάς και της μικρο-κατάτασης που έχουν επιλεγεί. Οι κορυφές, αντιστοιχούν στα κύρια γονίδια κι οι ακμές στις αλληλεπιδράσεις. Το μέγεθος των κατευθυνόμενων ακμών, είναι ανάλογο του βάρους της αντίστοιχης αλληλεπίδρασης. Με κόκκινο χρώμα, επισημαίνεται κατασταλτική / ανασταλτική αλληλεπίδραση ενώ με πράσινο, ενισχυτική / ενεργοποιητική / διεγερτική.

Παραδείγματα δημιουργίας:

```
plotGRN(MLscAN_obj, traj="adult1-to-adult2", mstate="ground")
plotGeneratedGRN(grn_weights_mat, numRegs=2)
```

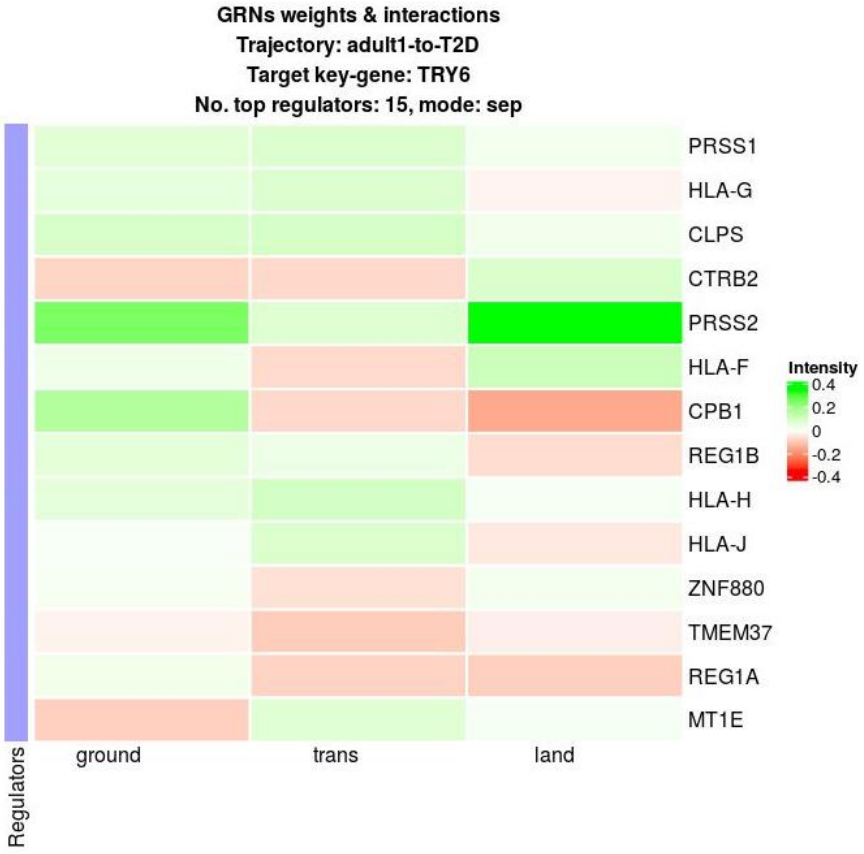


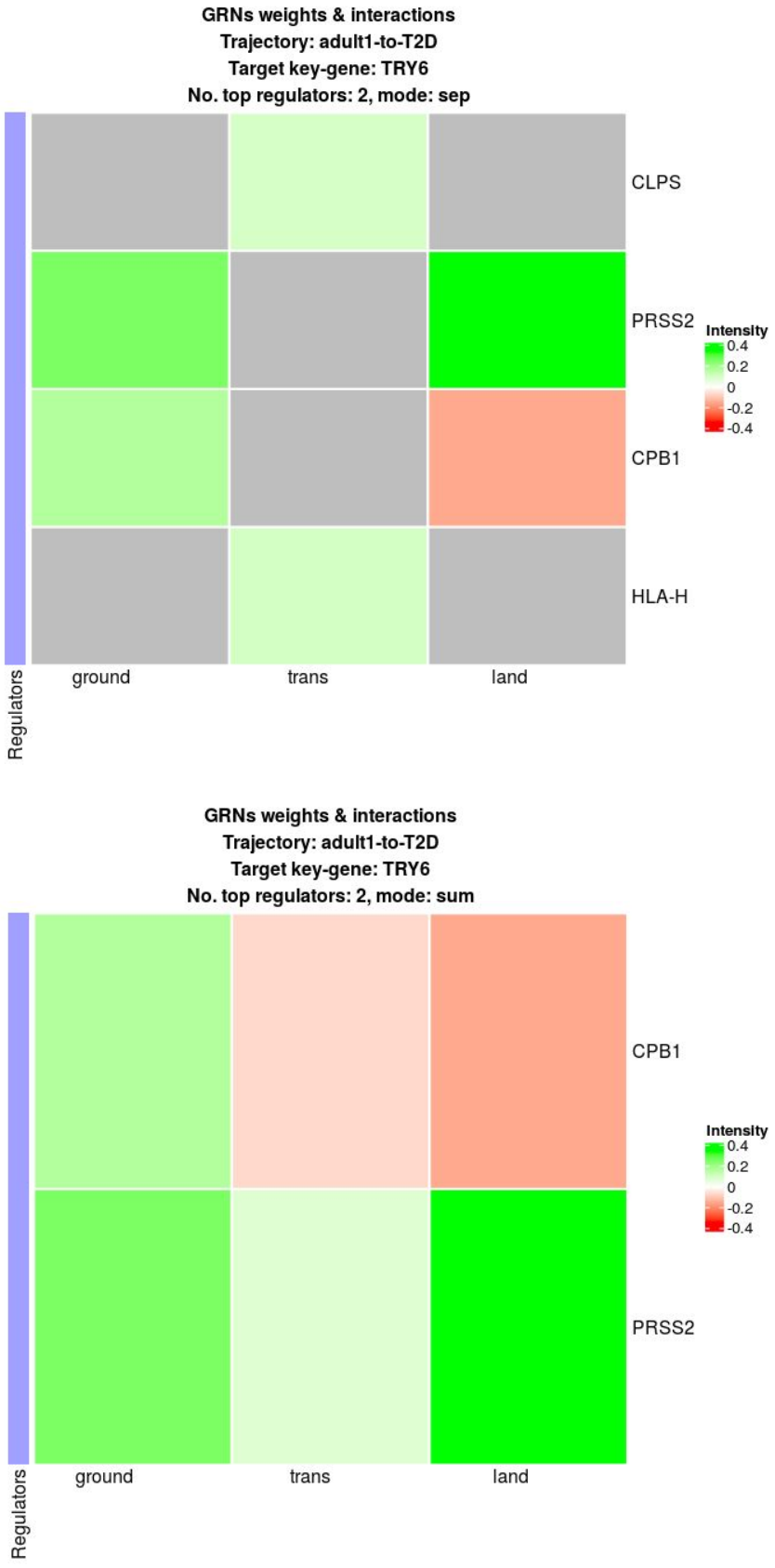
Εικόνα 3.55: Χάρτες θερμότητας (heatmap) των βαρών του GRN της τροχιάς και της μικρο-κατάτασης που έχουν επιλεγεί. Με κόκκινο χρώμα, επισημαίνεται κατασταλτική / ανασταλτική αλληλεπίδραση ενώ με πράσινο, ενισχυτική / ενεργοποιητική / διεγερτική. Μπορεί να οριστεί ο αριθμός των ρυθμιστών ανά γονίδιο-στόχο, για τους οποίους θα φαίνεται η τιμή του βάρους της αλληλεπίδρασης, με βάση την απόλυτη τιμή των βαρών, ξεκινώντας από αυτά με τη μεγαλύτερη

τιμή. Εφόσον ο επιλεγμένος αριθμός δεν επιτρέπει την προβολή όλων των βαρών, τα υπολοιπόμενα, επισημαίνονται με γκρι χρώμα. Αν κάποιο γονίδιο-ρυθμιστής, δεν ανήκει στο επιλεγμένο πλήθος των γονιδίων-στόχων, τότε, δεν προστίθεται στον χάρτη θερμότητας. Τα γονίδια, διατάσσονται με αλφαριθμητική σειρά, από αριστερά προς και δεξιά κι από πάνω προς τα κάτω.

Παραδείγματα δημιουργίας:

```
plotGRNHeatmap(MLscAN_obj, traj="adult1-to-adult2", mstate="ground")

plotGRNHeatmap(MLscAN_obj, traj="adult1-to-adult2", mstate="ground",
               numTopRegulators=3)
```





Εικόνα 3.56: Χάρτες θερμότητας (heatmap) των βαρών των GRNs της τροχιάς για κάθε μικρο-κατάσταση για ένα γονίδιο-στόχο που έχει επιλεγεί. Με κόκκινο χρώμα, επισημαίνεται κατασταλτική / ανασταλτική αλληλεπίδραση ενώ με πράσινο, ενισχυτική / ενεργοποιητική / διεγερτική. Μπορεί να οριστεί ο αριθμός των ρυθμιστών του γονιδίου-στόχου, για τους οποίους θα φαίνεται η τιμή του βάρους της αλληλεπίδρασης, με βάση την απόλυτη τιμή των βαρών ή το

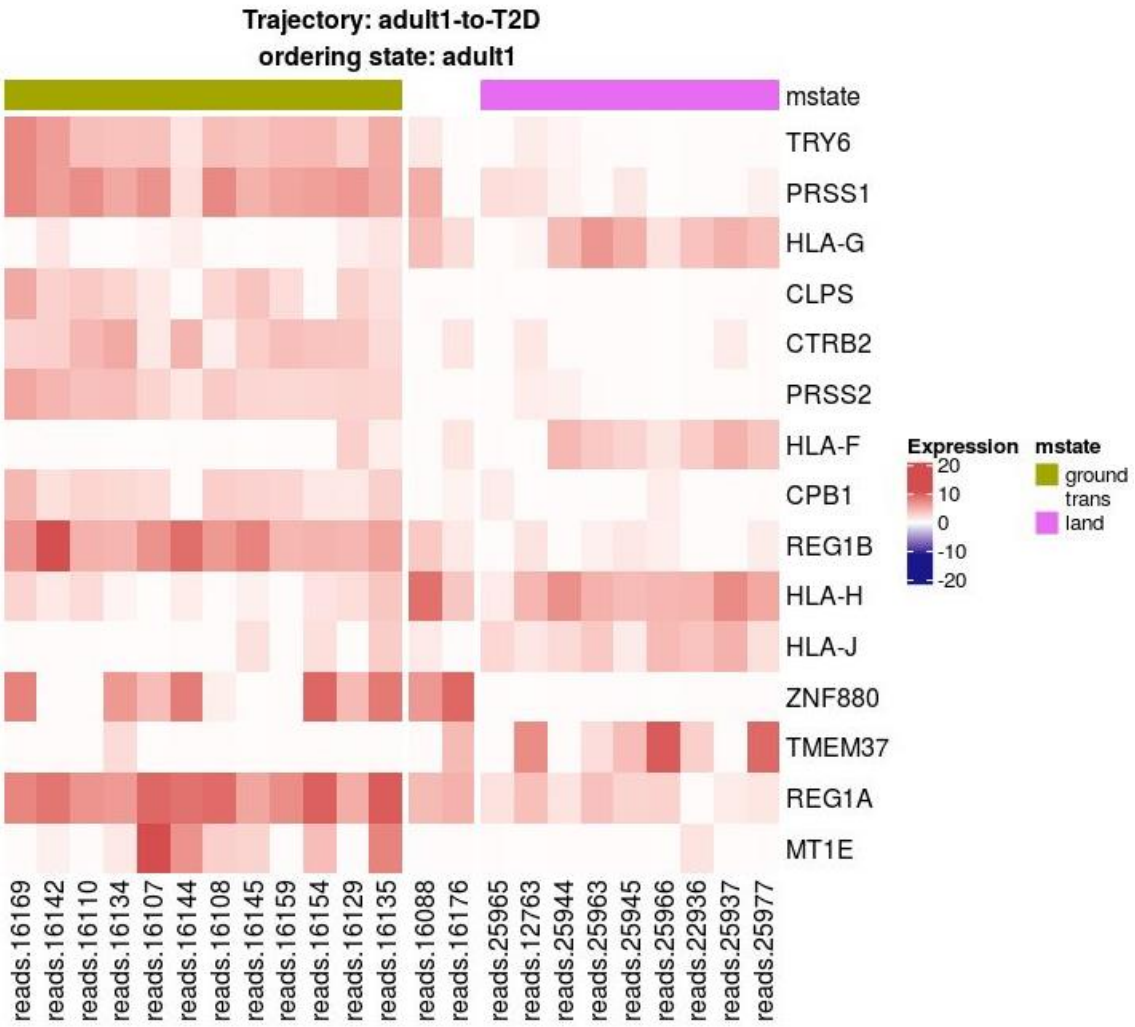
άθροισμα των απόλυτων τιμών των διαφορών των βαρών μεταξύ των διαδοχικών μικρο-καταστάσεων. Εφόσον ο επιλεγμένος αριθμός δεν επιτρέπει την προβολή όλων των βαρών, τα υπολοιπόμενα, επισημαίνονται με γκρι χρώμα. Αν κάποιο γονίδιο-ρυθμιστής, δεν ανήκει στο επιλεγμένο πλήθος του γονιδίου-στόχου, τότε, δεν περιλαμβάνεται στον χάρτη θερμότητας. Τα γονίδια, διατηρούν τη διάταξη των κύριων γονιδίων (βάσει της σημαντικότητας για την τροχιά), από πάνω προς τα κάτω.

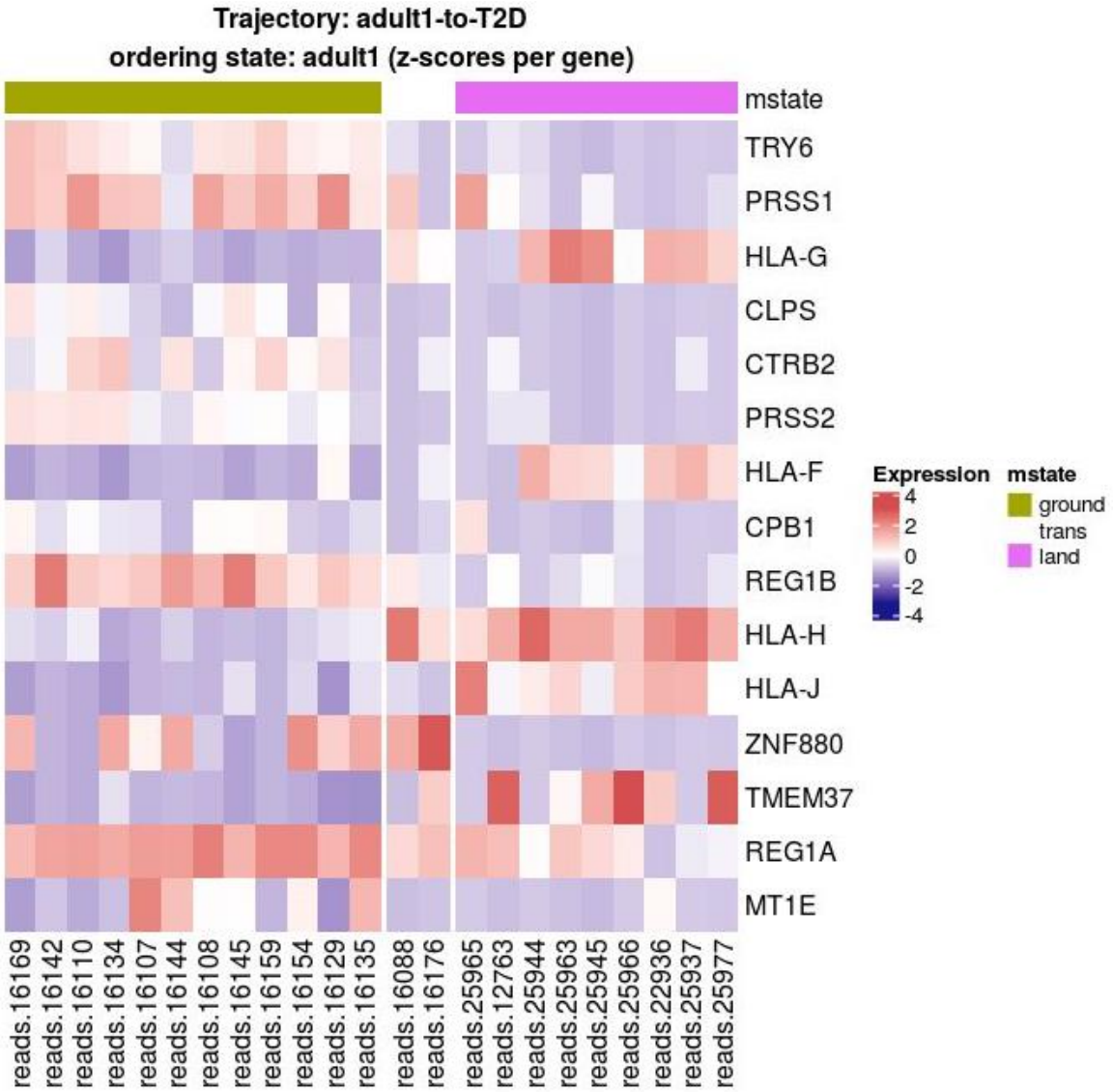
Παραδείγματα δημιουργίας:

```
plotGeneGRNHeatmap(MLscAN_obj, traj="adult1-to-T2D", gene="TRY6")

plotGeneGRNHeatmap(MLscAN_obj, traj="adult1-to-T2D", gene="TRY6",
                   numTopRegulators=3)

plotGeneGRNHeatmap(MLscAN_obj, traj="adult1-to-T2D", gene="TRY6",
                   numTopRegulators=3, topRegulatorsMode="sum")
```

Ως προς τις δύο προσεγγίσεις επιλογής των κύριων γονιδίων που θα θεωρηθούν σημαντικότεροι ρυθμιστές:

- Με την πρώτη, όπου επιλέγονται τα γονίδια – ρυθμιστές με τις μεγαλύτερες απόλυτες τιμές βαρών ανά μικρο-κατάσταση (topRegulatorsMode="sep"), η προσοχή εστιάζεται στους ρυθμιστές που συμπεραίνεται ότι έχουν σημαντικότερη επίδραση στα επίπεδα έκφρασης των γονιδίων – στόχων για τη δεδομένη φάση (μικρο-κατάσταση).
- Με τη δεύτερη, όπου επιλέγονται τα γονίδια – ρυθμιστές με τα μεγαλύτερα αθροίσματα απόλυτων διαφορών μεταξύ των διαδοχικών μικρο-καταστάσεων (topRegulatorsMode="sum"), η προσοχή εστιάζεται σε αυτά τα γονίδια για τα οποία συγκεντρωτικά άλλαξε περισσότερο η «επίδραση» στο γονίδιο – στόχο. Επίσης, παρατηρείται στο σύνολο των μικρο-καταστάσεων, η εξέλιξη της έντασης και του τύπου της αλληλεπίδρασης για δεδομένο ρυθμιστή. Αυτοί οι ρυθμιστές, ενδέχεται να θεωρηθούν ενδιαφέροντες για τη μελέτη των παραγόντων που μπορεί να έχουν σημαντική επίδραση στην καθοδήγηση της εξέλιξης της μετάβασης ή χαρακτηρίζουν κάποιο στάδιό της.

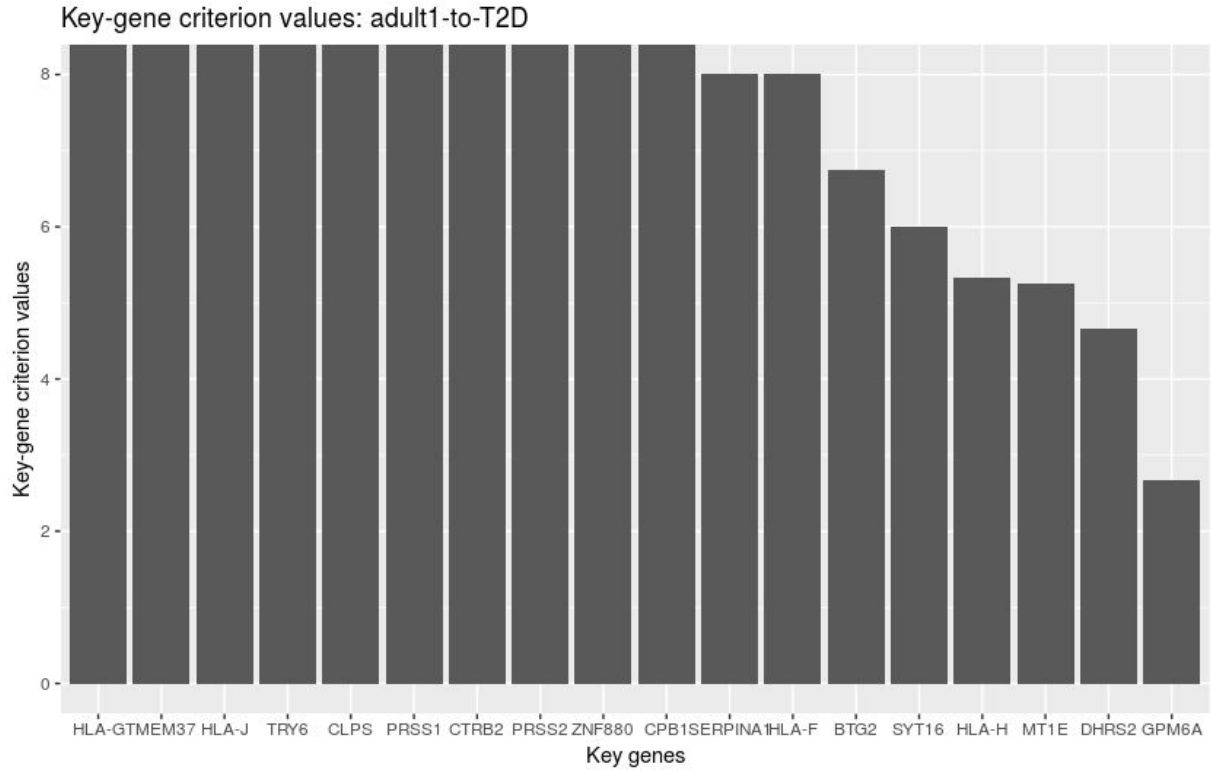

Trajectory: adult1-to-T2D
ordering state: adult1
mstate
TRY6
PRSS1
HLA-G
CLPS
CTRB2
PRSS2
HLA-F
CPB1
REG1B
HLA-H
HLA-J
ZNF880
TMEM37
REG1A
MT1E
Expression
20
10
0
-10
-20
mstate
ground
trans
land
reads.16169
reads.16142
reads.16110
reads.16134
reads.16107
reads.16144
reads.16108
reads.16145
reads.16159
reads.16154
reads.16129
reads.16135
reads.16088
reads.16176
reads.25965
reads.12763
reads.25944
reads.25963
reads.25945
reads.25966
reads.22936
reads.25937
reads.25977



Εικόνα 3.57: Χάρτες θερμότητας (heatmap) ως προς την έκφραση των κύριων γονιδίων (ή άλλων επιλεγμένων γονιδίων) της επιλεγμένης τροχιάς, χρησιμοποιώντας απευθείας τις τιμές αυτές ή τις τυπικές τους τιμές (υπολογισμός ανά γονίδιο για τα κύτταρα της τροχιάς). Τα γονίδια, διατηρούν τη διάταξη με την οποία παρέχονται, από πάνω προς τα κάτω. Στο επάνω μέρος, επισημαίνονται οι μικρο-καταστάσεις στις οποίες ανήκουν τα κύτταρα, που παραμένουν ταξινομημένα όπως και στην τροχιά (από αριστερά προς τα δεξιά).

Παραδείγματα δημιουργίας:

```
plotHeatmapTraj(MLscAN_obj, traj="adult1-to-T2D", z_scores=FALSE,
                genes=keyGenes(MLscAN_obj, traj="adult1-to-T2D"))

plotHeatmapTraj(MLscAN_obj, traj="adult1-to-T2D",
                genes=keyGenes(MLscAN_obj, traj="adult1-to-T2D"))
```



Εικόνα 3.58: Διάγραμμα της τιμής του κριτηρίου ταξινόμησης των κύριων γονιδίων (όπως αναφέρθηκε στην ενότητα 2.2.8.1.1) για την επιλμεγένη τροχιά.

Παράδειγμα δημιουργίας:

```
plotKeyGenesCritValues(MLscAN_obj, traj="adult1-to-T2D")
```

3.3 Τεκμηρίωση

Το εγχειρίδιο τεκμηρίωσης, δημιουργείται από τα αρχεία Rd του πακέτου MLscAN, για παράδειγμα, με την εντολή:

```
R CMD Rd2pdf MLscAN
```

3.4 Παραδείγματα χρήσης

Αρχικά, αν είναι επιθυμητό, μπορεί να χρησιμοποιηθεί το πακέτο, BiocParallel [51], για παραλληλοποίηση στο επίπεδο δημιουργίας των τροχιών (που περιλαμβάνει τη δημιουργία των μικρο-καταστάσεων, την επιλογή των κύριων γονιδίων και τη δημιουργία των GRNs) κι είναι συνήθως το πιο χρονοβόρο μέρος της ροής επεξεργασίας. Επισημαίνεται ότι αυτό δεν είναι δυνατό σε περιβάλλον Windows, γιατί στηρίζεται στη δημιουργία κλώνων των διεργασιών (forking). Οι όποιες επιλογές εκτέλεσης, πρέπει να οριστούν πριν αρχίσει η δημιουργία του προτύπου MLscAN, όπως για παράδειγμα συμβαίνει παρακάτω (καθολική επιλογή χρήσης 3 πυρήνων):

```
BiocParallel::register(BiocParalell::bpstart(BiocParalell::MulticoreParam(
                        workers=3)))
```

1. Ο απλούστερος τρόπος για τη δημιουργία του προτύπου MLscAN, απαιτεί μόνο τα δεδομένα έκφρασης:

```
MLscAN_obj <- MLscAN(exprData=data)
```

2. Αν χρησιμοποιηθεί και παραλληλοποίηση με το πακέτο BiocParalell:

```
MLscAN_obj <- MLscAN(exprData=data,
                     MLscANUseParallel=TRUE)
```

3. Ορίζοντας επιπλέον τον κατάλογο που θα αποθηκευτούν τα αρχεία εξόδου:

```
MLscAN_obj <- MLscAN(exprData=data,
                     MLscANOutDir="/home/user/Desktop")
```

4. Σταματώντας τη διαδικασία στη δημιουργία των καταστάσεων:

```
MLscAN_obj <- MLscAN(exprData=data,
                     MLscANStopAt="model")
```

Στη συνέχεια, είναι δυνατό να «συνεχιστεί» η δημιουργία του μοντέλου μέχρι τη δημιουργία των GRNs ανά μικρο-κατάσταση των τροχιών, εκμεταλλευόμενοι τα δεδομένα που έχουν ήδη παραχθεί, ιδιαίτερα εάν η διαδικασία είναι χρονοβόρα ή τα αποτελέσματα που προκύπτουν με χρήση του πακέτου mlust, δεν είναι ίδια μεταξύ επαναλαμβανόμενων κλήσεων (εξαιτίας της φύσης των δεδομένων έκφρασης σε σχέση με την εφαρμοζόμενη μέθοδο αρχικοποίησης, όταν προκύπτουν ισοπαλίες που επιλύονται τυχαία [31]):

```
MLscAN_obj <- MLscAN(exprData=data,
                     dimRedData=dimRedData(MLscAN_obj),
                     modelPostProbs=postProbs(MLscAN_obj))
```

5. Ορίζοντας την επιλογή διαστάσεων των αποτελεσμάτων μείωσης της διαστατικότητας που θα χρησιμοποιηθούν βάσει της αθροιστικής διακύμανσης ή παρέχοντας απευθείας τον επιθυμητό αριθμό:

```
MLscAN_obj <- MLscAN(exprData=data,
                     dimRedNumDim="0.5")

MLscAN_obj <- MLscAN(exprData=data,
                     dimRedNumDim="15")
```

6. Παρέχοντας και πίνακα χαρακτηριστικών των κυττάρων (στήλες: χαρακτηριστικά, γραμμές: κύτταρα):

```
MLscAN_obj <- MLscAN(exprData=data,
                     mscanCellFeatures=cell_features_mat)
```

7. Δίνοντας τα αποτελέσματα μείωσης της διαστατικότητας από ένα αρχείο τύπου CSV (με παρουσία ονομάτων των γραμμών και τοποθέτηση των κυττάρων στις γραμμές):

```
MLscAN_obj <- MLscAN(exprData=data,
                     dimRedInFile="plsr_results.csv",
                     dimRedInInfo="ftrc")
```

8. Επιλέγοντας τον αριθμό των καταστάσεων κι επιτρέποντας να ονομαστούν με βάση τους κυτταρικούς τύπους:

```
MLscAN_obj <- MLscAN(exprData=data,
                     MLscANCellFeatures=cell_features,
                     modelNumStates=3,
                     modelStateNameMode="mostFreqPerType")
```

9. Αλλάζοντας το όριο των κριτηρίων ΔBIC και του βήματος αύξησής του σε κάθε επανάληψη:

```
MLscAN_obj <- MLscAN(exprData=data,
                     modelDBICThr=0.02,
                     modelDBICStep=0.005)
```

10. Ορίζοντας τη συνάρτηση που θα χρησιμοποιηθεί για την αναγνώριση των πιθανών ακραίων υπο-πληθυσμών:

```
MLscAN_obj <- MLscAN(exprData=data,
                     outlSelFun=custom_outlFun)
```

11. Αλλάζοντας την τιμή του kgSTDWeight (περιγράφηκε στην ενότητα 2.2.8.1):

```
MLscAN_obj <- MLscAN(exprData=data,
                     kgSTDWeight=1.5)
```

12. Αλλάζοντας τον μέγιστο αριθμό των σημαντικότερων γονιδίων – ρυθμιστών ανά γονίδιο – στόχο,που θα χρησιμοποιηθεί για τη δημιουργία των διαγραμμάτων των GRNs:

```
MLscAN_obj <- MLscAN(exprData=data,
                     grnTargetNumregulators=3)
```

3.5 Αποτελέσματα αναλυτή κατανομής / απόδοσης (profiler)

Αρχικά, πρέπει να επισημανθεί ότι ορισμένα τμήματα της ροής επεξεργασίας, απαιτούν σημαντικά περισσότερο χρόνο σε σχέση με άλλα, όπως φαίνεται στην εικόνα 3.61, αλλά, τα χαρακτηριστικά των δεδομένων δρουν καίρια στη διαμόρφωση του συνολικού χρόνου (χρόνου χρήστη (user time)). Αν για παράδειγμα είναι μικρός ο αριθμός των γονιδίων ή δε σχηματιστούν τροχιές ή τα κύρια γονίδια κάθε τροχιάς είναι λίγα, ο χρόνος μειώνεται πολύ.

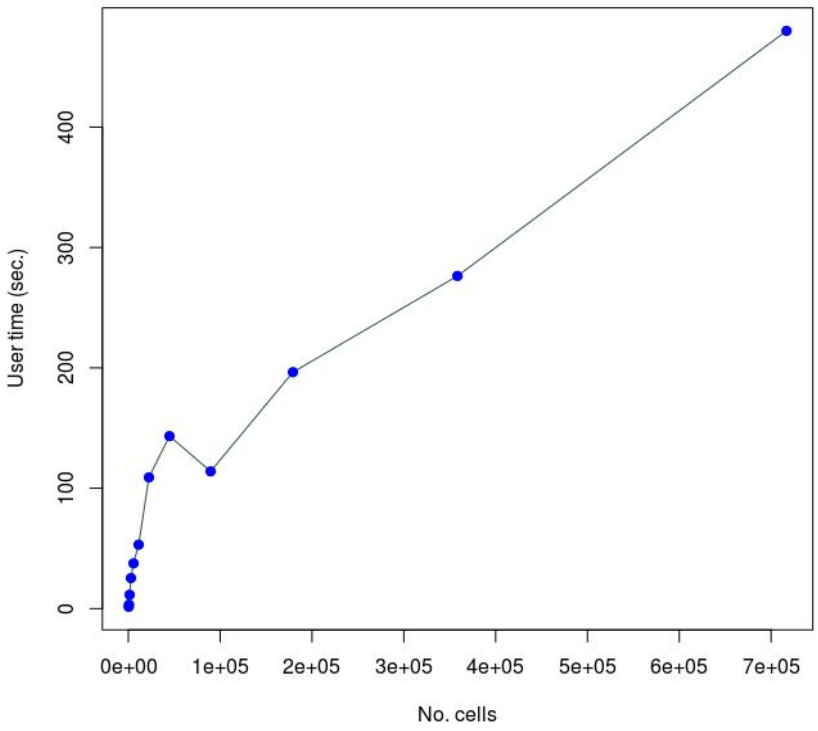
Ανά βήμα της προκαθορισμένης διαδικασίας, η πολυπλοκότητα χρόνου με βάση τους παράγοντες που καθορίζουν τον φόρτο στα απαιτητικά τμήματα, είναι:

- Μείωση της διαστατικότητας: $O(\#κυττάρων^2 * \#γονιδίων + \#γονιδίων^3)$
- Δημιουργία του προτύπου μείξης κανονικών κατανομών: $O(\#κυττάρων * \#καταστάσεων)$
- Σχηματισμός των μικρο-καταστάσεων των τροχιών: $O(\#τροχιών * \#κυττάρων\ τροχιάς)$
- Αναγνώριση των κύριων γονιδίων των τροχιών: $O(\#τροχιών * \#γονιδίων * \#κυττάρων\ τροχιάς)$
- Σχηματισμός των GRNs ανά μικρο-κατάσταση των τροχιών: $O(\#τροχιών * \#μικρο-καταστάσεων * \#κυττάρων\ τροχιάς * log(\#κυττάρων\ τροχιάς) * \#κύριων\ γονιδίων\ τροχιάς)$

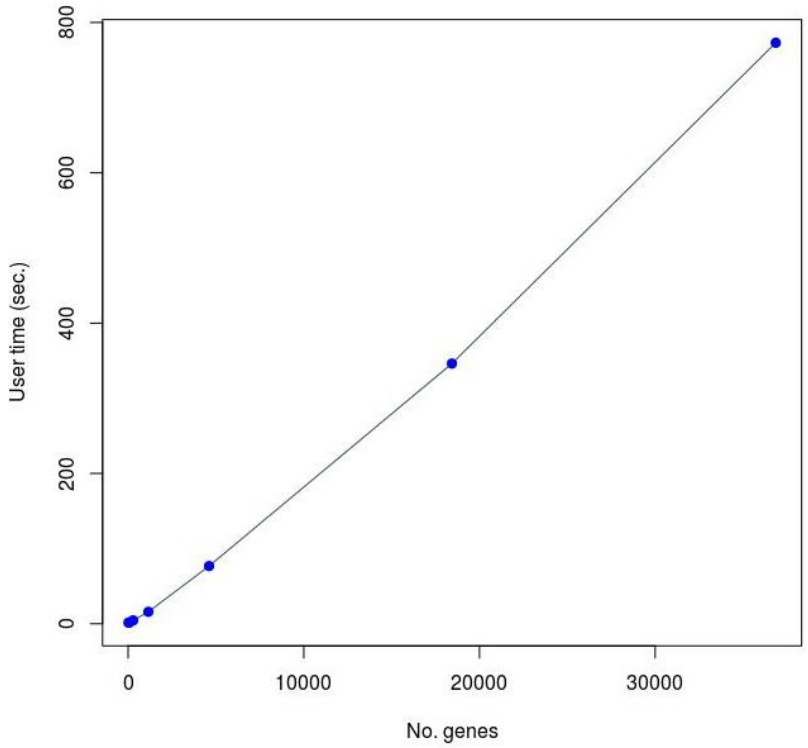
Όλα τα αποτελέσματα παρακάτω, δημιουργήθηκαν χωρίς παραλληλοποίηση, σε περιβάλλον Windows 7 / 2,70 GHz / 4,00 GB.

Για τους ελέγχους, χρησιμοποιήθηκαν 350 τυχαία κύτταρα και 20 γονίδια, επιλέγοντας να δημιουργούνται 3 καταστάσεις κι έχοντας ελέγξει προηγουμένως ότι δημιουργούνται δύο τροχιές με κύρια γονίδια. Για την αύξηση του αριθμού των κυττάρων ή των γονιδίων, δημιουργούνται αντίγραφα των αρχικών κυττάρων ή γονιδίων. Οι δοκιμές, πραγματοποιήθηκαν εν σειρά, χωρίς επαναλήψεις.

Όπως φαίνεται στις εικόνες 3.59 και 3.60, ο χρόνος χρήστη αυξάνεται σχεδόν γραμμικά με την αύξηση του αριθμού των κυττάρων ή των γονιδίων.



Εικόνα 3.59: Αποτελέσματα ελέγχου του χρόνου χρήστη σε σχέση με τον αριθμό των κυττάρων για τη δημιουργία του προτύπου MIscAN, διατηρώντας σταθερό τον αριθμό των γονδίων.

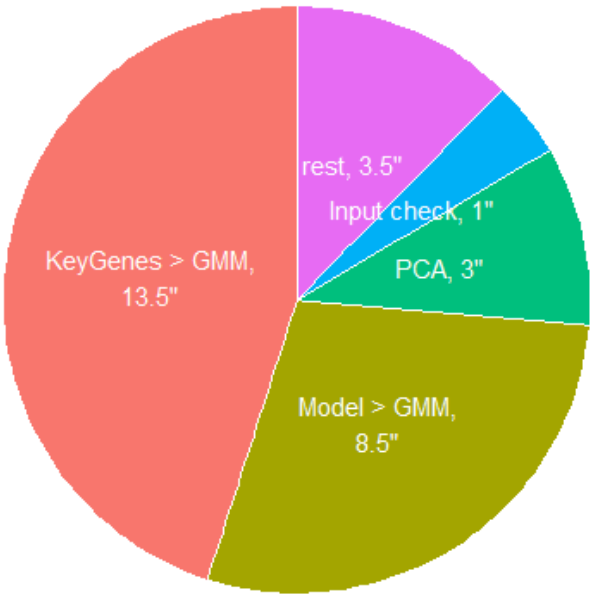


Εικόνα 3.60: Αποτελέσματα ελέγχου του χρόνου χρήστη σε σχέση με τον αριθμό των γονιδίων για τη δημιουργία του προτύπου MLSCAn, διατηρώντας σταθερό τον αριθμό των κυττάρων.

Στη συνέχεια, με τα ίδια αρχικά δεδομένα, και παραγάγοντας 5.600 κύτταρα και 320 γονίδια, αναλύθηκε η κατανομή του χρόνου στα διάφορα μέρη της ροής επεξεργασίας, χρησιμοποιώντας το πακέτο R, profvis [52].

Ο χρόνος χρήστη, ήταν 29,5 δευτερόλεπτα και στην εικόνα 3.61, παρουσιάζονται οι διεργασίες που διήρκησαν τουλάχιστον ένα δευτερόλεπτο (δηλ., 3% του συνολικού χρόνου).

Διαπιστώνεται ότι η δημιουργία των προτύπων μείξης κανονικών κατανομών, είναι η πλέον χρονοβόρα διαδικασία. Επαναλαμβανόμενη μία φορα για κάθε γονίδιο και για κάθε τροχιά, φτάνει να αποτελεί το 46% του συνολικού χρόνου κι ένα πρόσθετο 29% για τη δημιουργία των καταστάσεων. Η μείωση της διαστατικότητας με PCA, αποτελεί το 10% του συνολικού χρόνο, ο έλεγχος της εισόδου στις συναρτήσεις κατασκευής των κλάσεων το 3%, κι όλες οι υπόλοιπες διεργασίες, αποτελούν το 12% του συνολικού χρόνου.



Εικόνα 3.61: Η κατανομή του χρόνου στις διεργασίες της ροής επεξεργασίας για τη δημιουργία του προτύπου MLscAN, προβάλλοντας ξεχωριστά μόνο τα βήματα με διάρκεια τουλάχιστον ενός δευτερολέπτου.

3.6 Αποτελέσματα δοκιμών σε διαφορετικά λειτουργικά συστήματα

Οι δοκιμές, έγιναν σε περιβάλλον Windows 7 και Linux Ubuntu 14.01.

Ελέχοντας το πακέτο με:

- R CMD build
- R CMD CHECK
- BiocCheck

δεν προέκυψε κάποιο σφάλμα ή προειδοποίηση.

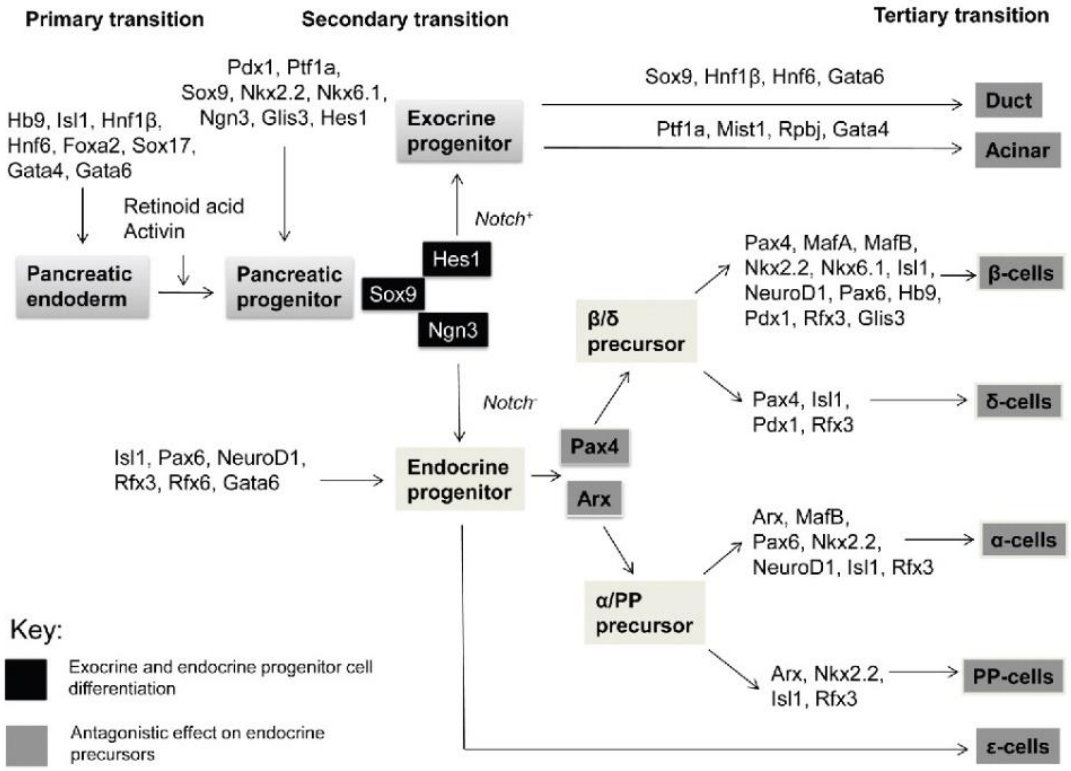
4. ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΧΡΗΣΙΜΟΠΟΙΩΝΤΑΣ ΤΟ ΠΑΚΕΤΟ MLscAN

Σε αυτό το κεφάλαιο, χρησιμοποιούνται δεδομένα από την ανάλυση της έκφρασης μονήρων κυττάρων από το πάγκρεας ανθρώπων, παιδιών κι ενηλίκων, με σακχαρώδη διαβήτη ή όχι, [53] για τη δημιουργία ενός προτύπου MLscAN. Πιο συγκεκριμένα, θα χρησιμοποιηθούν τα β-κύτταρα της ενδοκρινούς μοίρας του παγκρέατος παιδιών κι ενηλίκων χωρίς σακχαρώδη διαβήτη κι ενηλίκων με σακχαρώδη διαβήτη τύπου 2 (ΣΔΤ2).

4.1 Συνοπτικές πληροφορίες για την ενδοκρινή μοίρα του παγκρέατος και τον ΣΔΤ2

Τα κύτταρα της ενδοκρινούς μοίρας του παγκρέατος, αποτελούν μόνο το 2% της μάζας του οργάνου, αλλά, δέχονται περίπου το 15% της αιματικής ροής. Εντοπίζονται διάσπαρτα στην εξωκρινή μοίρα, κατά βάση οργανωμένα στα νησίδια του Langerhans, τα οποία έχουν πλούσια νεύρωση. Ένα νησίδιο Langerhans, αποτελεί μία εκκριτική μονάδα, με τέσσερις βασικούς τύπους κυττάρων, καθένας από τους οποίους παράγει τουλάχιστον μία χαρακτηριστική πρωτεΐνη / ορμόνη· τα α-κύτταρα, που παράγουν τη σωματοστατίνη κι αποτελούν περίπου το 20% των κυττάρων του νησιδίου, τα β-κύτταρα που παράγουν την ινσουλίνη κι αποτελούν περίπου το 70% των κυττάρων του νησιδίου, εντοπιζόμενα κυρίως στο κέντρο του, τα δ-κύτταρα που παράγουν τη σωματοστατίνη κι αποτελούν περίπου το 5% των κυττάρων του νησιδίου και τα κύτταρα PP που παράγουν το παγκρεατικό πολυπεπτίδιο κι αποτελούν περίπου το 1% των κυττάρων του νησιδίου. [54] Ο ρόλος τους, είναι κεντρικός για τη ρύθμιση των επιπέδων γλυκόζης.

Στην πορεία διαφοροποίησης αυτών των κυττάρων, από τα ολοδύναμα (pluripotent) έως τα τελικώς διαφοροποιημένα κύτταρα του παγκρέατος, που αντικατοπτρίζεται στην παρουσία ή απουσία δράσης συγκεκριμένων μεταγραφικών παραγόντων, οι κυτταρικοί τύποι μοιράζονται σε διάφορο βαθμό κοινούς «προγόνους», όπως παρουσιάζεται στην εικόνα 4.1 [55].



Εικόνα 4.1: Κύριοι μεταγραφικοί παράγοντες κατά τη διαφοροποίηση των κυττάρων του παγκρέατος [55].

Ο σακχαρώδης διαβήτης, αποτελεί ένα σύνολο μεταβολικών ασθενειών, με κοινό χαρακτηριστικό την παρουσία υπεργλυκαιμίας, που οφείλεται σε διαταραχές στην έκκριση ή στη δράση της ινσουλίνης. Περίπου 90% - 95% των ατόμων με σακχαρώδη διαβήτη, εντάσσονται στον τύπο 2. Σύμφωνα με την τρέχουσα κατηγοριοποίηση της Αμερικανικής Διαβητολογικής Εταιρείας, ο σακχαρώδης διαβήτης τύπου 2, αφορά σε προοδευτική απώλεια της έκκρισης ινσουλίνης από τα β-κύτταρα, συχνά στο έδαφος αντίστασης στην ινσουλίνη. Ο παθογενετικός μηχανισμός, ωστόσο, δεν είναι γνωστός. [56]

4.2 Πληροφορίες για το σύνολο δεδομένων και σημαντικά σημεία από το σχετικό άρθρο

Στο άρθρο από το οποίο προέρχονται τα δεδομένα που θα χρησιμοποιηθούν [53], τα σημαντικά σημεία σε ό,τι αφορά τα β-κύτταρα και τον ΣΔΤ2, είναι τα ακόλουθα:

- Υπάρχει ετερογένεια (δεν είναι όλα τα β-κύτταρα πανομοιότυπα) κι ιδιαίτερα σε συνθήκες μεταβολικού στρες.
- Παρατηρείται αρκετή μεταβλητότητα στην έκφραση ακόμη και των χαρατηριστικών ορμονών.
- Τα β-κύτταρα της ομάδας των παιδιών έχουν λιγότερο καλά ορισμένη γονιδιακή υπογραφή σε σχέση με την ομάδα των ενηλίκων. Μάλιστα, αρκετά γονίδια της

γονιδιακής υπογραφής των β-κυττάρων, δεν εκφράζονται στα β-κύτταρα της ομάδας των παιδιών.

- Αρκετά γονίδια της γονιδιακής υπογραφής των α-κυττάρων εκφράζονται στα β-κύτταρα της ομάδας των παιδιών.
- Τα β-κύτταρα της ομάδας των ατόμων με ΣΔΤ2 έχουν προφίλ έκφρασης με χαρακτηριστικά που παρατηρούνται σε αυτά της ομάδας των παιδιών, εμφανίζοντας ένα πιο «ανώριμο» προφίλ. Αυτό, υποδεικνύει την παρουσία μερικής αποδιαφοροποίησης, όπως υποδηλώνεται από την προς τα πάνω ρύθμιση γονιδίων που σχετίζονται με τον κυτταρικό κύκλο.
- Η βαθμολογία εμπλουτισμού του γονιδιακού συνόλου (gene-set enrichment score) των β-κυττάρων της ομάδας των ενηλίκων, είναι πολύ μικρότερη στην ομάδα των παιδιών σε σχέση με την ομάδα των ατόμων με ΣΔΤ2.
- Αν κι οι μηχανισμοί που ενέχονται στη διαδικασία αποδιαφοροποίησης δεν είναι γνωστοί, σίγουρα συμμετέχει ο μεταγραφικός παράγοντας FOXO1.
- Η διαδικασία ποιοτικού ελέγχου των κυττάρων (συνοπτική παρουσίαση παρακάτω), μπορεί να οδήγησε στην αφαίρεση κάποιων που βρίσκονταν στη διαδικασία διαδιαφοροποίησης ή σπάνιων προγονικών κυττάρων.

Προκειμένου να αναγνωριστεί ο τύπος κάθε κυττάρου και να απομακρυνθούν ζεύγη κυττάρων που εκλαμβάνονται ως ένα, εφαρμόστηκε μία διαδικασία πέντε βημάτων, ως εξής:

- Δημιουργία ενός προτύπου μείξης κανονικών κατανομών για κάθε γονίδιο, με 3 ομάδες χαρακτηριζόμενες από τα επίπεδα έκφρασης: χαμηλή έκφραση, ενδιάμεση έκφραση, υψηλή έκφραση. Σε αυτό το στάδιο, ανατίθεται αρχικά ένας τύπος (α, β, δ, ε, PP, κυψελιδικά, των πόρων, μεσεγχυματικά), στα κύτταρα που ταξινομούνται στην ομάδα υψηλής έκφρασης για τα γνωστά γονίδια – δείκτες, με την προϋπόθεση αυτό να συβαίνει μόνο για τους δείκτες ενός τύπου.
- Παραγωγή των γονιδιακών υπογραφών κάθε τύπου, συγκρίνοντας τα κύτταρα κάθε τύπου με το μεταγράφωμα του τύπου αυτού.
- Έλεγχος του βαθμού συσχέτισης κάθε κυττάρου με τη γονιδιακή υπογραφή κάθε τύπου. Αμιγή, θεωρούνται τα κύτταρα που έχουν υψηλή τιμή συσχέτισης με μία μόνο γονιδιακή υπογραφή.
- Με όσα κύτταρα δεν αφαιρέθηκαν στα προηγούμενα στάδια, επαναπροσδιορισμός των γονιδιακών υπογραφών των τύπων.

Αναγνωριστικό δότη	Ηλικία (έτη)	Φύλο	Εθνικότητα	ΔΜΣ	Καλλιέργεια (ημέρες)	Ομάδα	Αριθμός κυττάρων
AAJF122	52	Άνδρας	Ασιατική	29,1	6	Ενηλίκων	1
ABAF490	39	Γυναίκα	Λευκή	45,2	4	Ενηλίκων	30
ACAP236	21	Άνδρας	Λευκή	39	2	Ενηλίκων	17
HP-15041	57	Άνδρας	Αφροαμερικανική	23,98	4	ΣΔΤ2	5
HP-15085	37	Γυναίκα	Λευκή	39,3	4	ΣΔΤ2	16
ICRH76	2	Άνδρας	Λευκή	13,6	2	Παιδιών	4
ICRH80	1,6	Γυναίκα	Λευκή	18	3	Παιδιών	15

Εικόνα 4.2: Χαρακτηριστικά των κυττάρων των δοτών (GSE83139) [57].

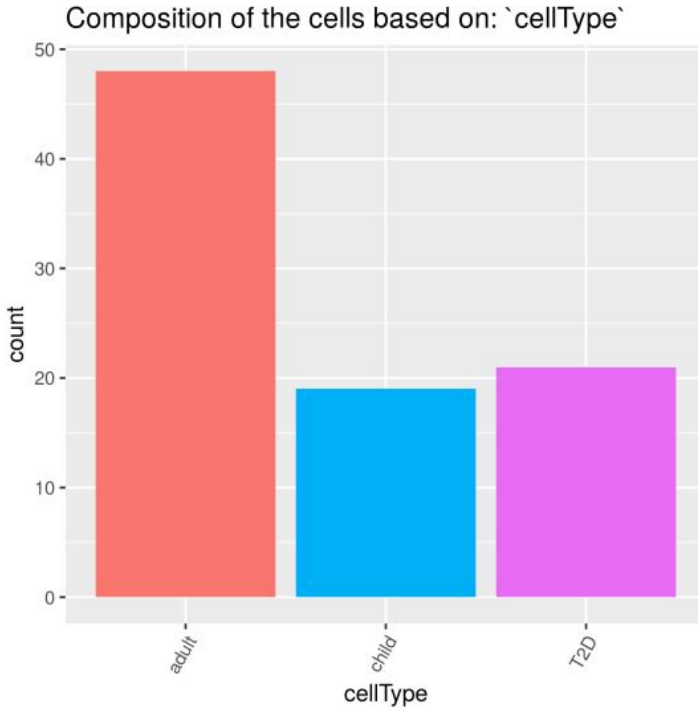
- Με την υπόθεση ότι τα ζεύγη κυττάρων που εκλαμβάνονται ως ένα, αποτελούν γραμμικό συνδυασμό δύο τύπων, δημιουργήθηκαν υπολογιστικά, μεικτά προφίλ έκφρασης για κάθε συνδυασμό τύπων. Έτσι, κύτταρα που το προφίλ έκφρασής τους μπορούσε να εξηγηθεί από ένα μεικτό προφίλ και δεν είχαν υψηλό βαθμό συσχέτισης με τη γονιδιακή υπογραφή ενός αμιγούς τύπου, απομακρύνθηκαν.

Με τον τρόπο αυτό, διατηρήθηκαν 430 από τα 635 κύτταρα όλων των τύπων.

Στο Gene Expression Omnibus [57], παρέχεται ο προ-επεξεργασμένος πίνακας έκφρασης, από τον οποίον διατηρήθηκαν μόνο τα β-κύτταρα, αφαιρώντας επιπλέον αυτά του ατόμου με σακχαρώδη διαβήτη τύπου 1 (αναγνωριστικό δότη: ACGI428) κι αυτά που καλλιεργήθηκαν περισσότερες ημέρες (αναγνωριστικό δότη: HP-15085:cultured). Έτσι, σε αυτό το στάδιο, ο πίνακας έκφρασης, συνίσταται από 88 κύτταρα και 19.949 γονίδια. Στην εικόνα 4.2, παρατίθενται τα διαθέσιμα χαρακτηριστικά των δοτών των κυττάρων αυτών.

Διακρίνουμε τους δότες σε τρεις ομάδες: ενήλικα άτομα χωρίς ΣΔΤ2 («ενήλικοι»), άτομα παιδικής ηλικίας χωρίς ΣΔ («παιδιά»), ενήλικα άτομα με ΣΔΤ2 («ΣΔΤ2»). Περισσότερα από τα μισά κύτταρα, προέρχονται από τους τρεις δότες της ομάδας των ενηλίκων (30 + 17 + 1 = 48 κύτταρα). Τα υπόλοιπα, προέρχονται σε σχεδόν ίση αναλογία από τους δύο δότες της ομάδας των παιδιών (15 + 4 = 19 κύτταρα) και τους δύο δότες αυτής των ατόμων με ΣΔΤ2 (16 + 5 = 21 κύτταρα). Συγκεντρωτικά, η σύσταση των κυττάρων βάσει των ομάδων αυτών, βρίσκεται στην εικόνα 4.3.

Αξίζει να σημειωθεί ότι κι οι τρεις ενήλικοι δότες της ομάδας ελέγχου έχουν δείκτη μάζας σώματος (ΔΜΣ) μεγαλύτερο του φυσιολογικού· με βάση την κατηγοριοποίηση του Παγκόσμιου Οργανισμού Υγείας [58], ένας εντάσσεται στην κατάσταση της προ-παχυσαρκίας, ένας στην τάξη ΙΙ της παχυσαρκίας κι ένας στην τάξη ΙΙΙ της παχυσαρκίας. Αυτό αυξάνει την πιθανότητα οι δότες αυτοί να εμφάνιζαν, μεταξύ άλλων, προ-διαβήτη [56].

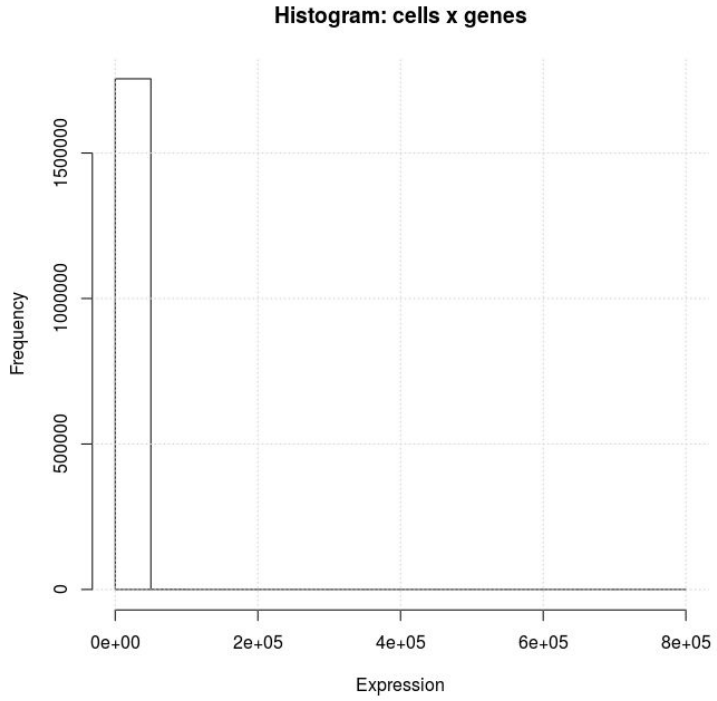


Εικόνα 4.3: Η κατανομή των β-κυττάρων που επιλέχθηκαν βάσει της ομάδας στην οποία ανήκουν.

Όπως αναφέρεται στο άρθρο, για την προ-επεξεργασία του πίνακα έκφρασης, ακολουθήθηκαν αδρά τα παρακάτω βήματα:

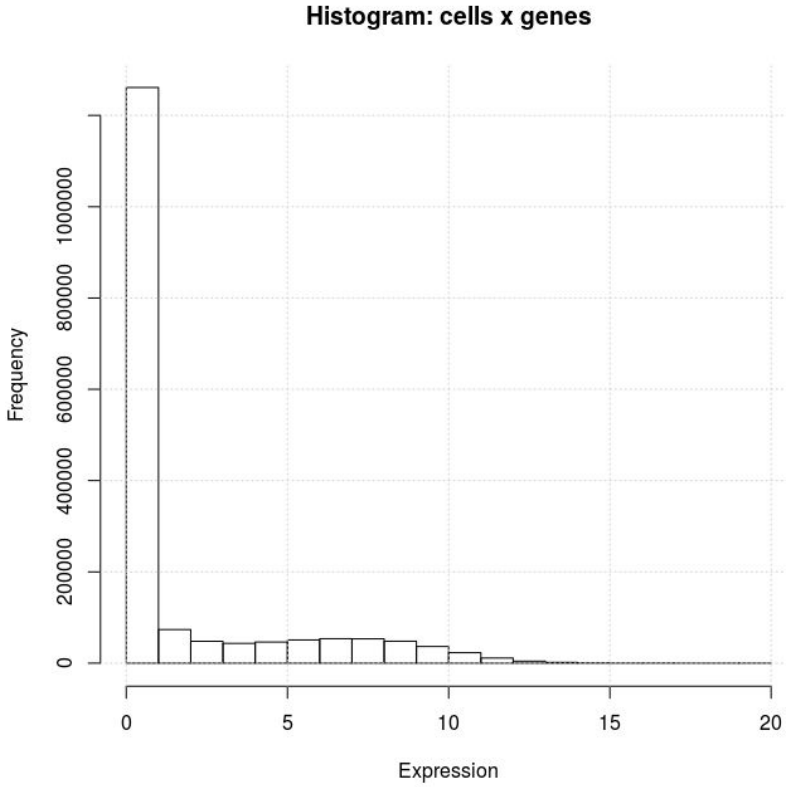
- μετατροπή των αναγνωσμάτων ανά γονίδιο (raw counts) σε αναγνώσματα ανά εκατομμύριο (Counts per million: CPM)
- προσθήκη ψευδο-αναγνωσμάτων (pseudo-counts)
- εφαρμογή παράγοντα κλιμάκωσης (scaling factor) χρησιμοποιώντας τη μέση τιμή έκφρασης των 10 κυττάρων με το μεγαλύτερο βάθος αλληλούχισης

Με αυτήν τη διαδικασία, όλες οι τιμές έκφρασης καταλήγουν να είναι μη-μηδενικές. Στην εικόνα 4.4, βρίσκεται το ιστόγραμμα των τιμών έκφρασης στα 88 β-κύτταρα που χρησιμοποίηθηκαν και σε όλα τα γονίδια (19.949). Είναι φανερό πως το εύρος έχει διαφορά επτά τάξεις μεγέθους, κι επίσης, μεγάλο ποσοστό αυτών έχει πολύ μικρή τιμή. Ενδεικτικά, το εύρος έκφρασης, είναι: 0,087 - 762.857, η μέση τιμή έκφρασης είναι: 117,7 και το 70% των τιμών έκφρασης είναι < 0,5. Αυτές οι πολύ χαμηλές τιμές, θεωρείται πως δεν αντιπροσωπεύουν εκφραζόμενο γονίδιο για τα αντίστοιχα κύτταρα.



Εικόνα 4.4: Ιστόγραμμα της έκφρασης όλων των γονιδίων κι όλων των κυττάρων, χωρίς περαιτέρω μετασχηματισμό.

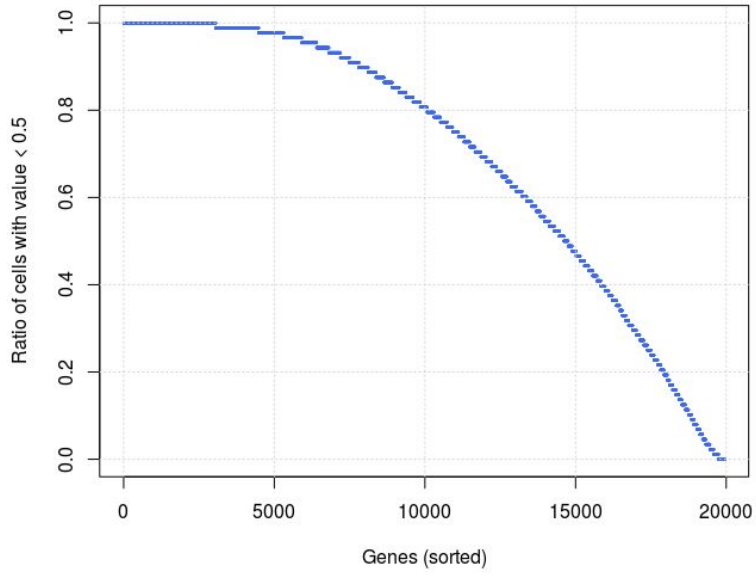
Προκειμένου να προσεγγίσει η κατανομή των εκφραζόμενων ανά κύτταρο γονιδίων την κανονική, εφαρμόζεται ο μετασχηματισμός $\log_2$(έκφραση + 1). Έτσι, εξακολουθεί να μην υπάρχει μηδενική τιμή στον πίνακα έκφρασης. Πλέον, το εύρος έκφρασης, είναι: 0,12 - 19,5, η μέση τιμή έκφρασης είναι: 1,8 και το 70% των τιμών έκφρασης είναι < 0,5. Στην εικόνα 4.5, βρίσκεται το νέο ιστόγραμμα.



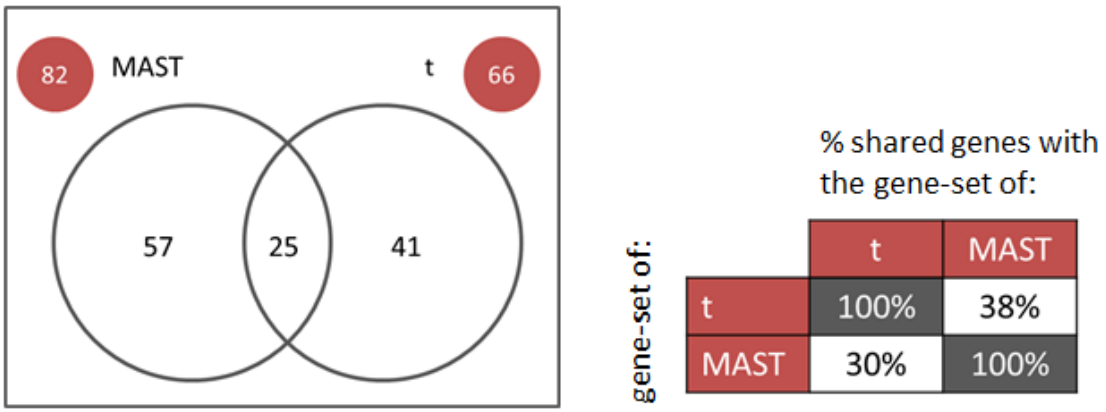
Εικόνα 4.5: Ιστόγραμμα της έκφρασης όλων των γονιδίων κι όλων των κυττάρων, μετά τον μετασχηματισμό.

4.3 Επιλογή του συνόλου γονιδίων και διερεύνηση των επιλεγμένων γονιδίων

Ακόμη και μετά τον μετασχηματισμό που έγινε, εξακολουθούν να υπάρχουν αρκετά γονίδια με πολύ χαμηλές τιμές έκφρασης, όπως φαίνεται στην εικόνα 4.6. Τα περισσότερα, αναμένεται να προσθέτουν θόρυβο, αλλά, κάποια ενδέχεται να έχουν βιολογικά χρήσιμη πληροφορία, δεδομένου και του σχετικά μικρού αριθμού διαθέσιμων κυττάρων.



Εικόνα 4.6: Λόγος των τιμών έκφρασης < 0,5 σε όλα τα κύτταρα, ανά γονίδιο.



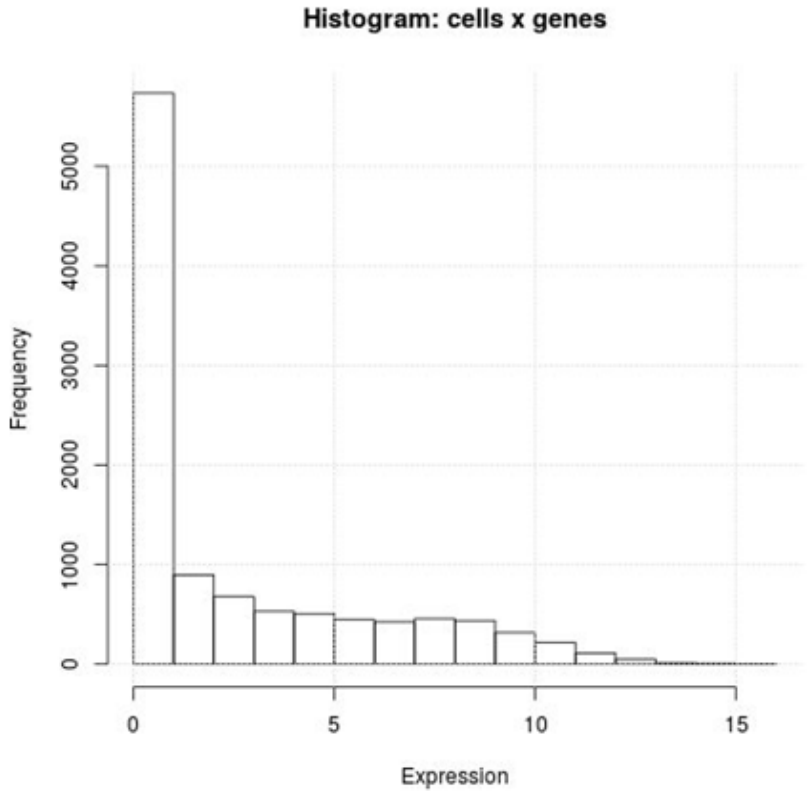
Εικόνα 4.7: Τα γονίδια που επιλέχθηκαν εφαρμόζοντας τις μεθόδους MAST [41] και t [23].

Προκειμένου να απομακρυνθεί μέρος του θορύβου, αλλά, κυρίως να είναι δυνατό να σχηματιστούν καταστάσεις που θα αποτελούνται σε μεγάλο ποσοστό από κύτταρα μίας ομάδας (ενηλίκων, ΣΔΤ2, παιδιών), ώστε να μελετηθεί η δημιουργία τροχιών με την υπόθεση της μερικής απο-διαφοροποίησης των β-κυττάρων στα άτομα με ΣΔΤ2, θα επιλεγεί ένα κατάλληλο υποσύνολο γονιδίων.

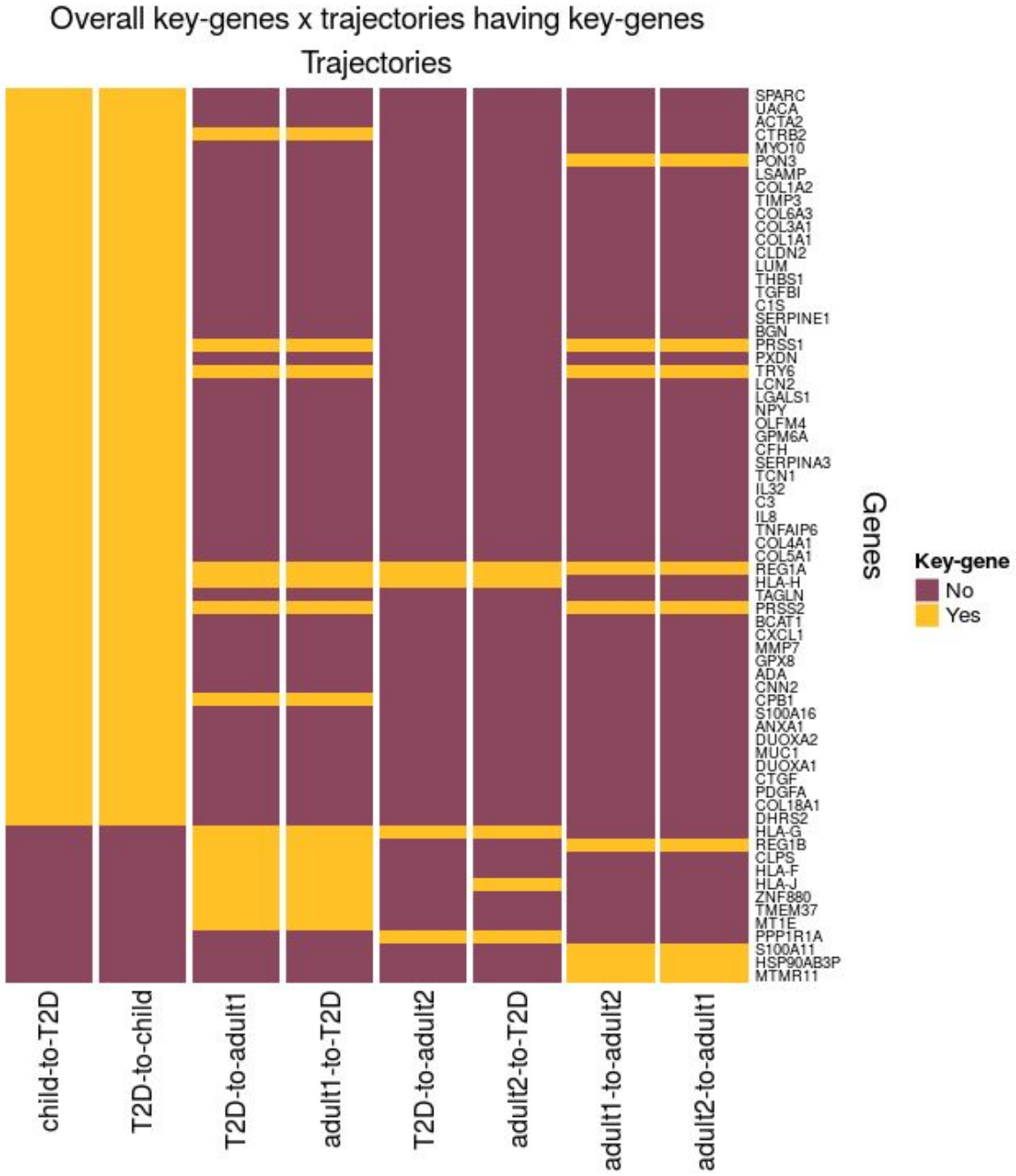
Για την επιλογή των γονιδίων, χρησιμοποιήθηκε η συνάρτηση, FindMarkers, του πακέτου R, Seurat (3.0.1) [39], εφαρμόζοντας τις μεθόδους, MAST και t (δοκιμασία t του Student), για τη σύγκριση των δύο ζευγών ομάδων κυττάρων: «παιδιά» και «ενήλικοι», «ΣΔΤ2» και «ενήλικοι». Οι συγκρίσεις, έγιναν με την ομάδα των «ενηλίκων» και καθεμία από τις άλλες δύο, επειδή από αυτήν προέρχονται περισσότερα από τα μισά κύτταρα (εικόνα 4.3). Από τα αποτελέσματα των δύο αυτών διαδικασιών και των δύο συγκρίσεων, διατηρήθηκαν, τελικά, τα γονίδια με προσαρμοσμένη τιμή p μικρότερη ή ίση του 0,1, σε οποιαδήποτε από τις δύο αυτές μεθόδους. Όπως φαίνεται στην εικόνα 4.7, με τη μέθοδο MAST, προέκυψαν 82 γονίδια και με τη μέθοδο t, 66, έχοντας 25 κοινά, και συνολικά 123 διακριτά γονίδια. Αυτό, μας παρέχει τη δυνατότητα να διαχωρίσουμε σε ικανοποιητικό βαθμό τις ομάδες των κυττάρων στις καταστάσεις, προσθέτοντας βιολογική σημασία στα αποτελέσματα. Αντίθετα, αυτό δε συνέβη όταν εφαρμόστηκαν κριτήρια επιλογής με βάση τη διακύμανση ανά κάδο ή τον λόγο παρουσίας μηδενικών τιμών με τον μετασχηματισμό $\log_2$(έκφραση).

Για τα 123 επιλεγμένα γονίδια, το ιστόγραμμα, βρίσκεται στην εικόνα 4.8, και διακρίνεται η αφαίρεση αρκετών εξαιρετικά χαμηλών τιμών.

Από τα 123 γονίδια που έχουν επιλεγεί, τα 68 θα αναγνωριστούν ως κύρια γονίδια για τουλάχιστον μία από τις τροχιές που θα σχηματιστούν (εικόνα 4.9), χρησιμοποιώντας όλες τις άμεσα διαθέσιμες μεθόδους μέσω της συνάρτησης *kg_voting* για την αναγνώρισή τους (όπως αναφέρθηκε στην ενότητα 2.2.8.3).



Εικόνα 4.8: Ιστόγραμμα της έκφρασης όλων των κυττάρων για τα 123 γονίδια που επιλέχθηκαν.



Εικόνα 4.9: Τα κύρια γονίδια για όλες τις τροχιές τού προτύπου MLscAN.

Διερευνώντας τα βασικά χαρακτηριστικά των γονιδίων που έχουν επιλεγεί, παρατηρείται ότι:

- Από το σύνολο των 1.105 γνωστών μεταγραφικών παραγόντων [59], περιλαμβάνονται:
    - στα 123 επιλεγμένα γονίδια: 1 (που δεν είναι το *FOXO1*)
    - στα 68 κύρια γονίδια: 0
- Από τα 3.381 γονίδια που σχετίστηκαν με τον ΣΔΤ2 μέσω μελέτης συσχέτισης σε επίπεδο γονιδιώματος (GWAS) [60], περιλαμβάνονται:

- στα 123 επιλεγμένα γονίδια: 19
- στα 68 κύρια γονίδια: 14

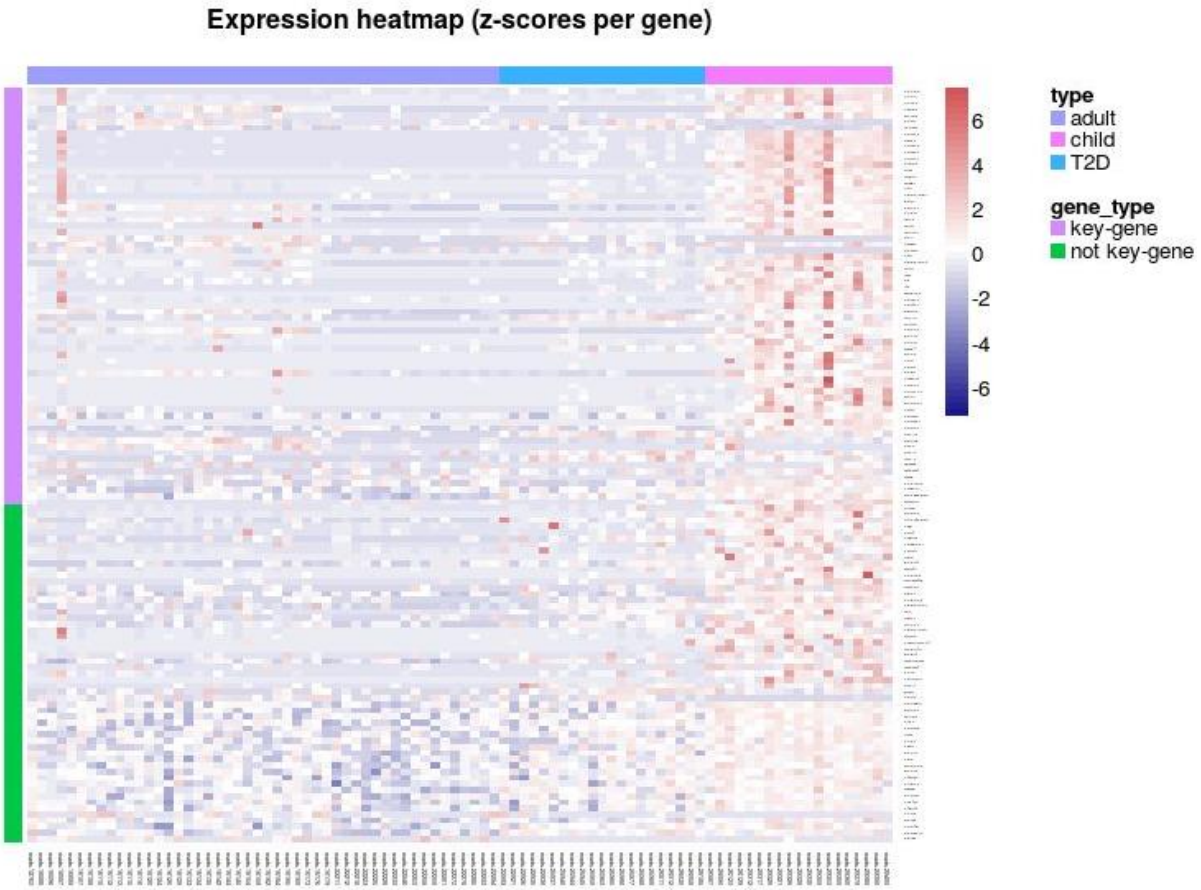
- Από τα 701 γονίδια που σχετίζονται με τον έλεγχο του κυτταρικού κύκλου [61], περιλαμβάνονται:
  - στα 123 επιλεγμένα γονίδια: 0
  - στα 68 κύρια γονίδια: 0
- Από τα 844 γονίδια που χαρακτηρίζονται στο άρθρο, άτυπα εκφρασμένα (misepressed, γονίδια που είτε εμφανίζουν υψηλή έκφραση στα κύτταρα των ομάδων των παιδιών και των ατόμων με ΣΔΤ2 κι όχι σε αυτήν των ενηλίκων είτε πρόκειται για γονίδια της υπογραφής των α- ή β-κυττάρων της ομάδας των ενηλίκων με αρκετά υψηλή έκφραση στα β- ή α-κύτταρα, αντίστοιχα, των κυττάρων της ομάδας των παιδιών ή των ατόμων με ΣΔΤ2), ή θεωρούνται γονίδια υπογραφής (γονίδια με χαρακτηριστικό πρότυπο έκφρασης σε ένα σύνολο κυττάρων), περιλαμβάνονται:
  - στα 123 επιλεγμένα γονίδια: 50
  - στα 68 κύρια γονίδια: 29
    - Γονίδια υπογραφής:
      - adult α (212 γονίδια): 5
      - adult β (376 γονίδια): 5
      - child & adult α (60 γονίδια): 2
      - child & adult β (242 γονίδια): 22
    - Άτυπα εκφρασμένα γονίδια:
      - Adult α, Child β (13 γονίδια): 1
      - Child α, T2D α (16 γονίδια): 2
      - Child β, T2D β (52 γονίδια): 2
- Γονίδια με αναφορά στην άμεση σχέση τους με τον ΣΔΤ2, στο GeneCards [62]:
  - στα 68 κύρια γονίδια: 5 (*CLPS*, *CTRB2*, *MT1E*, *NPY*, *PXDN*)
- Κύρια γονίδια που εμφανίζονται σε περισσότερες τροχιές από ένα ζεύγος αντίθετων τροχιών:

- σε 8 τροχιές: 1 (*REG1A*)
- σε 6 τροχιές: 4 (*HLA-H*, *PRSS1*, *PRSS2*, *TRY6*)
- σε 4 τροχιές: 5 (*CPB1*, *CTRB2*, *HLA-G*, *PON3*, *REG1B*)
- σε 3 τροχιές: 1 (*HLA-J*)

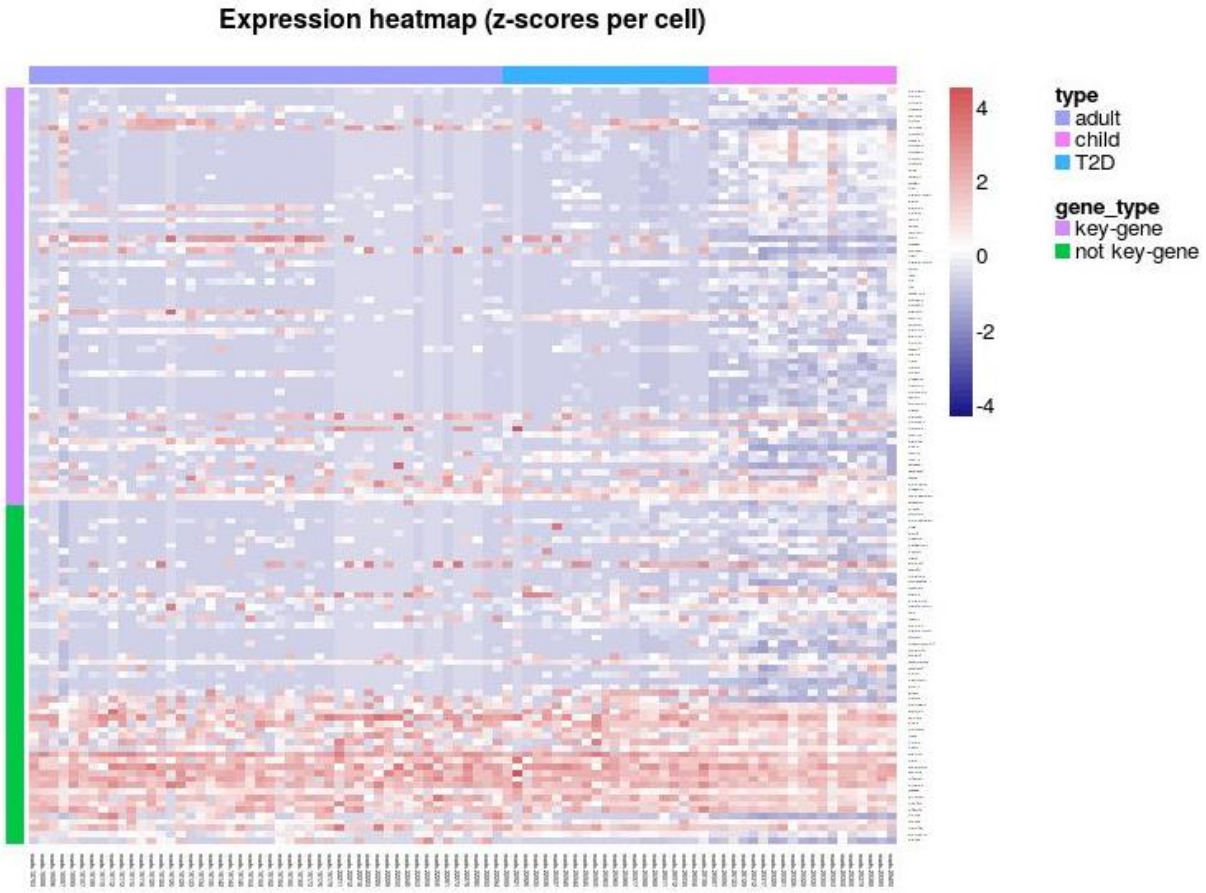
- Περίπου τα μισά κύρια γονίδια των τροχιών που δε συμμετέχει η κατάσταση των παιδιών, δεν είναι ταυτόχρονα και κύρια γονίδια των τροχιών που συμμετέχει η κατάσταση των παιδιών.

Στους χάρτες θερμότητας των εικόνων 4.10 και 4.11, περιλαμβάνονται τα 123 γονίδια που έχουν επιλεγεί, επισημαίνοντας αυτά που θεωρήθηκαν κύρια γονίδια για τουλάχιστον μία τροχιά και την ομάδα στην οποία ανήκουν τα κύτταρα.

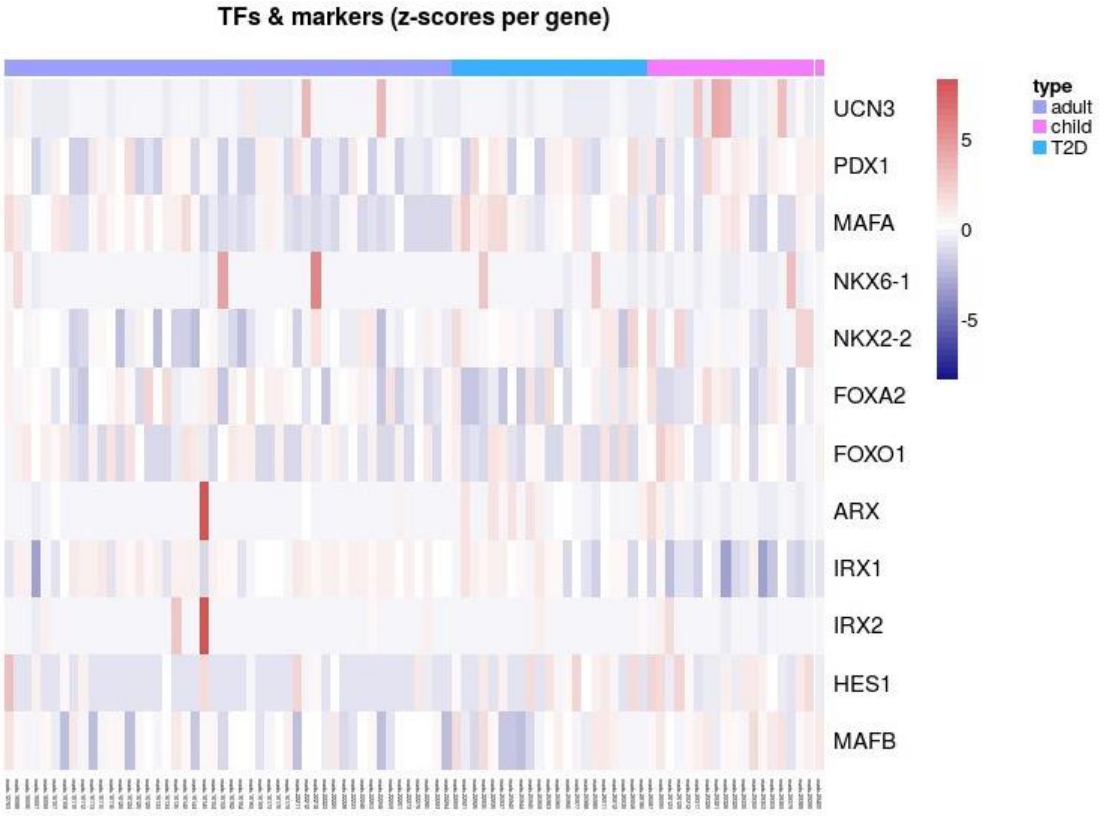
Σε ό,τι αφορά τους μεταγραφικούς παράγοντες, από τους χάρτες θερμότητας, με τους κυριότερους από αυτούς για τα α- και β-κύτταρα, δεν παρατηρείται κάποια συνεπής και σημαντική διαφορά μεταξύ των κυττάρων των τριών ομάδων (εικόνες 4.12 και 4.13). Σε αυτούς, περιλαμβάνεται κι ο FOXO1, που θεωρείται ότι συμμετέχει στη διαδικασία μερικής από-διαφοροποίησης των β-κυττάρων των ατόμων με ΣΔΤ2.



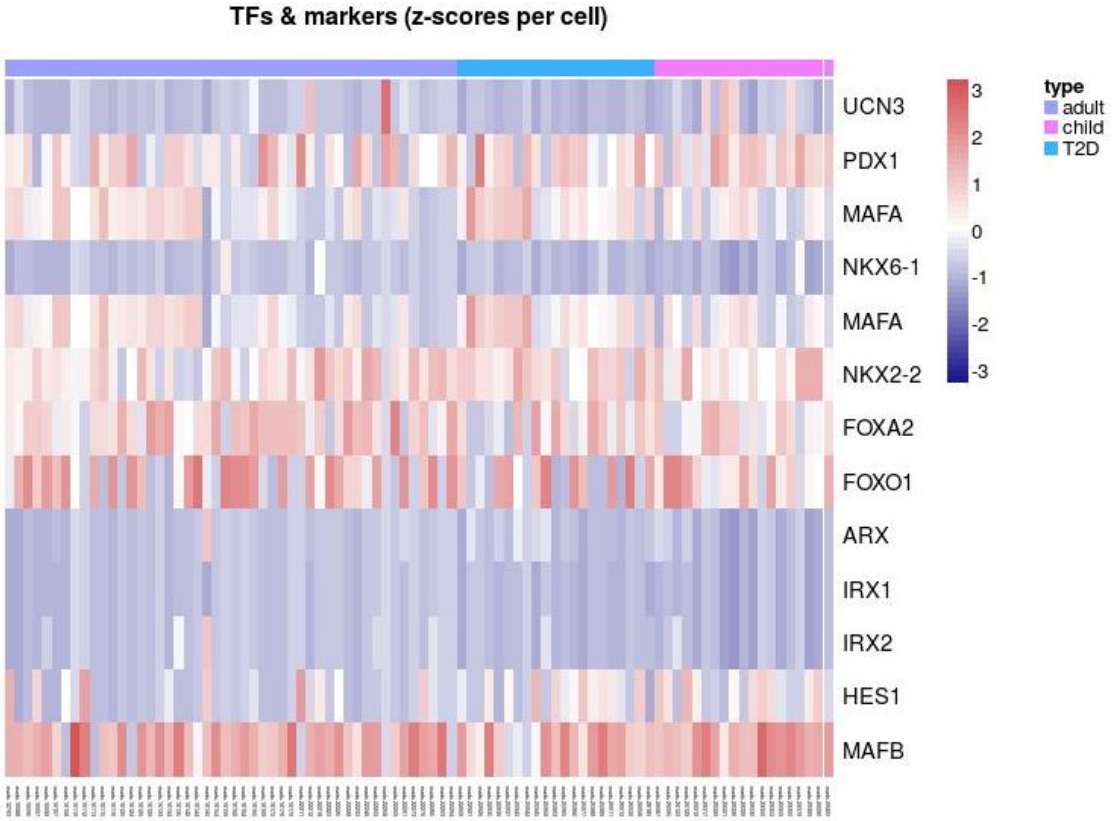
Εικόνα 4.10: Χάρτης θερμότητας της έκφρασης όλων των γονιδίων κι όλων των κυττάρων, χρησιμοποιώντας τις τιμές z ανά γονίδιο.



Εικόνα 4.11: Χάρτης θερμότητας της έκφρασης όλων των γονιδίων κι όλων των κυττάρων, χρησιμοποιώντας τις τιμές z ανά κύτταρο.



Εικόνα 4.12: Χάρτης θερμότητας της έκφρασης κύριων μεταγραφικών παραγόντων σε όλα τα κύτταρα, χρησιμοποιώντας τις τιμές z ανά γονίδιο.



Εικόνα 4.13: Χάρτης θερμότητας της έκφρασης κύριων μεταγραφικών παραγόντων σε όλα τα κύτταρα, χρησιμοποιώντας τις τιμές z ανά κύτταρο.

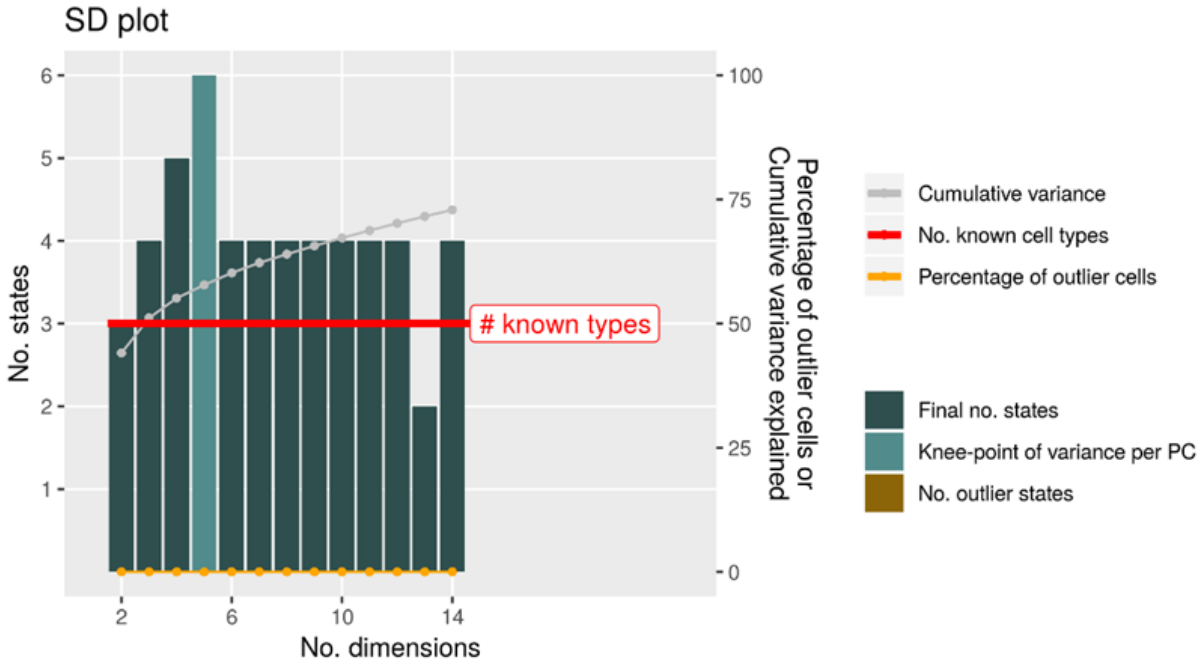
4.4 Επιλογή των υπόλοιπων παραμέτρων του προτύπου MLscAN

Προκειμένου να μελετηθει η δημιουργία τροχιών με την υπόθεση της μερικής απο-διαφοροποίησης των β-κυττάρων στα άτομα με ΣΔΤ2, είναι επιθυμητό να ξεκινά η δημιουργία των τροχιών έχοντας καταστάσεις που θα αποτελούνται σε μεγάλο ποσοστό από κύτταρα μίας ομάδας (ενηλίκων, ΣΔΤ2, παιδιών). Ο αριθμός των καταστάσεων, με την προαναφερθείσα συνθήκη, ενδέχεται να αποτελεί ένδειξη της παρουσίας υπο-πληθυσμών. Συνεπώς, δεν τίθεται κάποιος περιορισμός. Πρακτικά, ομάδες με αρκετά μικρό αριθμό κυττάρων, θα αποτελέσουν «θόρυβο» κι εξαρχής δυσχεραίνουν τη δημιουργία τροχιών, μιας κι απαιτείται ένας ελάχιστος αριθμός κυττάρων (τρία) από κάθε κατάσταση που συμμετέχει, και ταυτοχρόνως μπορεί να μειωθεί ο αριθμός των άλλων καταστάσεων που μπορούν να συμμετάσχουν σε άλλες τροχιές, αν κάποια «δεσμεύονται» σε μεταβάσεις με αυτές τις καταστάσεις.

Για να επιλεγεί ο καλύτερος αριθμός διαστάσεων, χρησιμοποιείται η συνάρτηση plotSD, από την οποία που δημιουργούνται τα διαγράμμα στις εικόνες 4.14 - 4.16:

```
plotSD(expr_data, from=2, to=14,
       know_cellTypes=unique(cell_features[,"cellType"]),
       plot_stateComp=TRUE,
       cellType=cell_features[,"cellType"])
```

Με αυτήν τη συνάρτηση, δημιουργείται το πρότυπο MLscAN, χρησιμοποιώντας ένα εύρος αριθμού κύριων συνιστωσών (PC) της PCA. Στην προκειμένη περίπτωση, από 2 έως και 14. Παρατηρείται ότι ακόμη και με 2 διαστάσεις, η αθροιστική διακύμανση είναι υψηλή (> 40%), οπότε, δε μας περιορίζει αυτό το χαρακτηριστικό ως προς την επιλογή αριθμού διαστάσεων. Επιπλέον, στις περισσότερες περιπτώσεις, προκύπτουν 4 καταστάσεις και κανένας υπο-πληθυσμός ακραίων κυττάρων.



Εικόνα 4.14: Διάγραμμα με συγκεντρωτικά στοιχεία για τα αποτελέσματα του προτύπου MLscAN, σε ένα εύρος χρησιμοποιούμενων συνιστωσών των αποτελεσμάτων μείωσης της διαστατικότητας (από 2 έως 14), χρησιμοποιώντας σταθερές παραμέτρους.

Εικόνα 4.15: Διαγράμματα της σύνθεσης των καταστάσεων των προτύπων MIscAN που προκύπτουν χρησιμοποιώντας ένα εύρος συνιστωσών των αποτελεσμάτων μείωσης της διαστατικότητας (2 έως 14), χρησιμοποιώντας σταθερές παραμέτρους.

Ως προς τη σύσταση των καταστάσεων, οι περιπτώσεις που έχουν σε μεγαλύτερο βαθμό τα επιθυμητά χαρακτηριστικά, είναι αυτές με χρήση, 9, 11, 12 και 14 κύριων συνιστωσών. Τελικά επιλέγεται να χρησιμοποιηθούν 11 κύριες συνιστώσες (εικόνα 4.16) – την περίπτωση στο ενδιαμεσο του εύρους, 9 - 14.

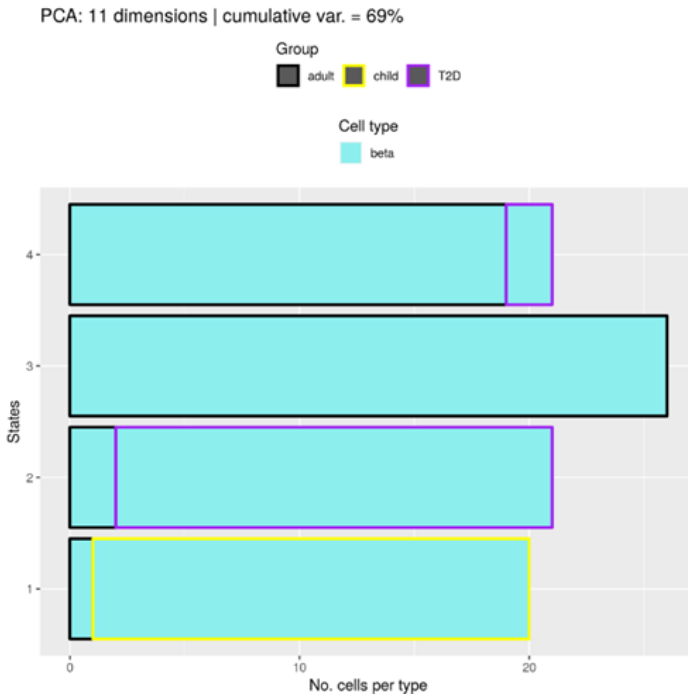
Έχοντας επιλέξει τις τιμές για τις απαραίτητες παραμέτρους, μπορεί να δημιουργηθεί το πρότυπο MLscAN έως τη δημιουργία των GRNs ανά μικρο-κατάσταση των τροχιών, με:

- χρήση 11 PC της PCA για την παραγωγή των εκ των υστέρων πιθανοτήτων
- επιλογή και διάταξη των κύριων γονιδίων κάθε τροχιάς χρησιμοποιώντας όλες τις άμεσα διαθέσιμες μεθόδους (όπως αναφέρθηκε στην ενότητα 2.2.8.3)
- αυτόματη ονομασία των καταστάσεων βάσει του τύπου των κυττάρων τους (επεξήγηση στην ενότητα 2.2.4.1)

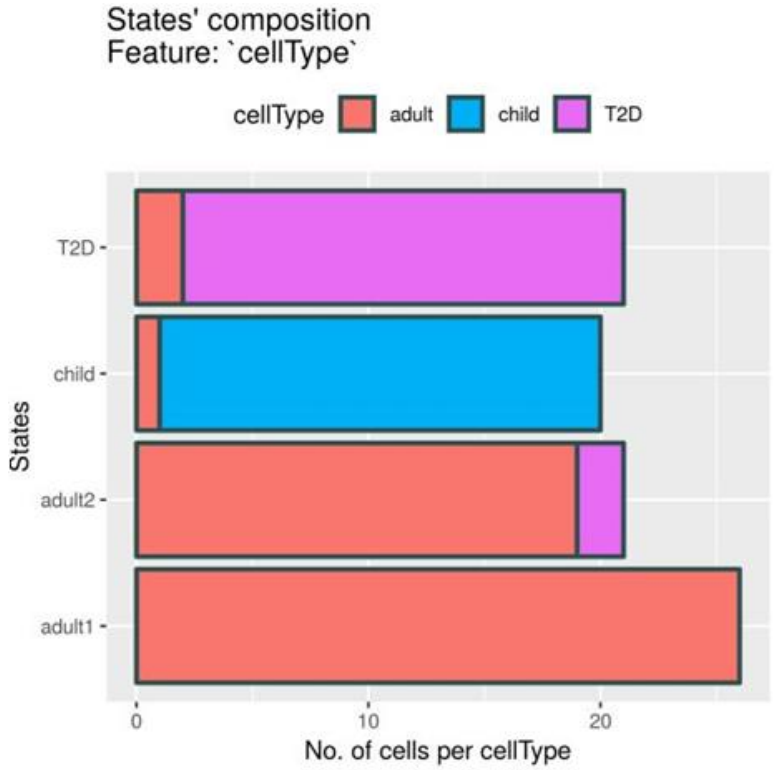
```
MLscAN_model <- MLscAN(exprData=expr_data,
                       MLscANCellFeatures=cell_features,
                       dimRedNumDim=11,
                       kgGenesSelFun=kg_voting(),
                       modelStateNameMode="mostFreqPerState")
```

4.5 Αποτελέσματα – Γενικά

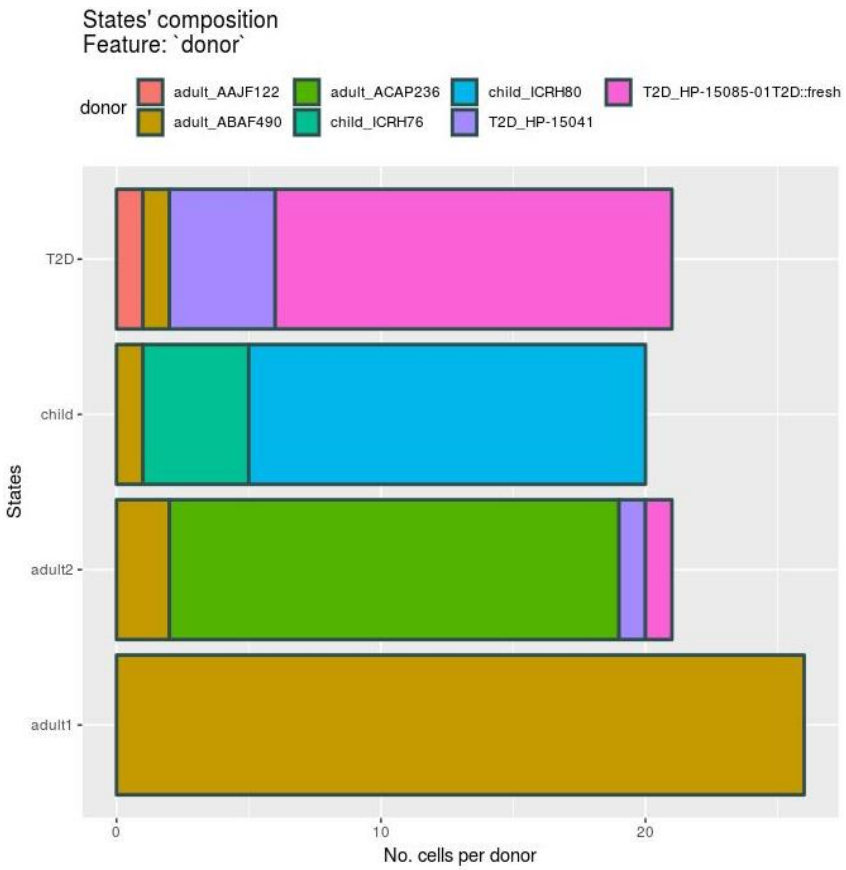
Είναι ήδη γνωστό ότι προκύπτουν 4 καταστάσεις (εικόνα 4.17) κι όπως παρατηρείται στην εικόνα 4.18, στις δύο καταστάσεις, «adult1» και «adult2», διαχωρίζονται τα κύτταρα των δύο δοτών της ομάδας των «ενηλίκων» με τα περισσότερα κύτταρα. Στην «adult1», βρίσκονται κυρίως τα κύτταρα της γυναίκας δότριας, με τον μεγαλύτερο ΔΜΣ.



Εικόνα 4.16: Η σύσταση των καταστάσεων χρησιμοποιώντας 11 κύριες συνιστώσες.



Εικόνα 4.17: Η σύσταση των καταστάσεων που προτύπου MLscAN βάσει του τύπου των κυττάρων.



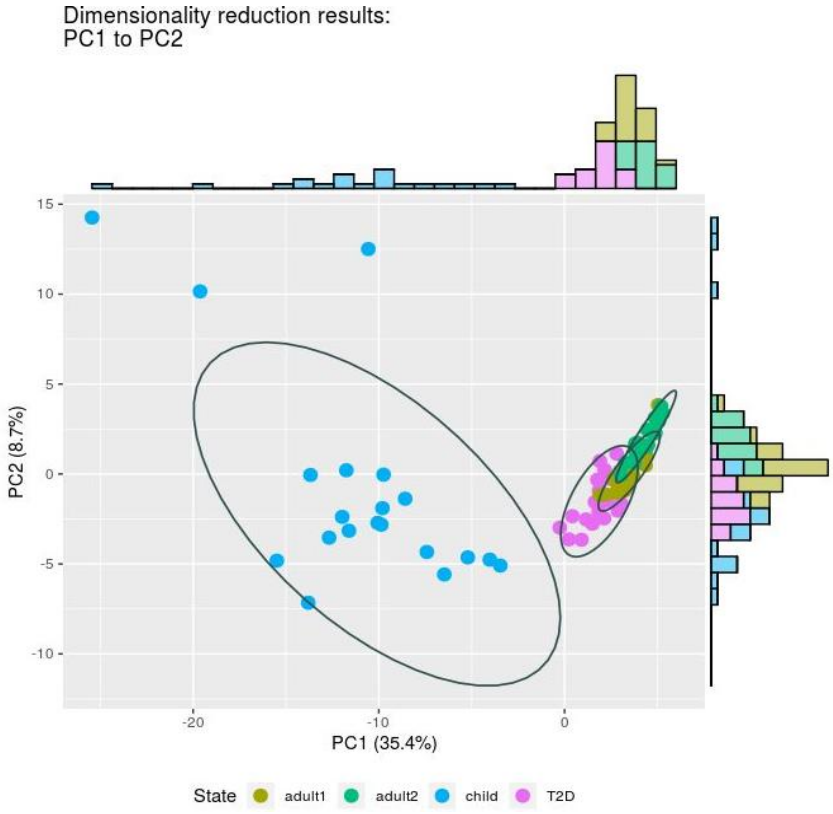
Εικόνα 4.18: Η σύσταση των καταστάσεων που προτύπου MLscAN βάσει των δοτών.

Στα διαγράμματα των αποτελεσμάτων μείωσης της διασταστικότητας (εικόνες 4.19 - 4.21), με τις τρεις πρώτες διαστάσεις ανά ζεύγος, φαίνεται ότι η κατάσταση «child» εμφανίζει μεγαλύτερη διασπορά και στις τρεις κύριες συνιστώσες. Με την PC1, μπορούν να διαχωριστούν όλες οι άλλες καταστάσεις από την «child», με τα κύτταρα της «T2D» να είναι εγγύτερά τους, ενώ με την PC2 ή την PC3, μπορεί να υπάρξει σχετικός διαχωρισμός των καταστάσεων, εκτός της «child», μεταξύ τους. Βέβαια, παρατηρείται αρκετή επικάλυψη και παρουσιάζουν, συγκριτικά, αρκετά μεγαλύτερη συνοχή.

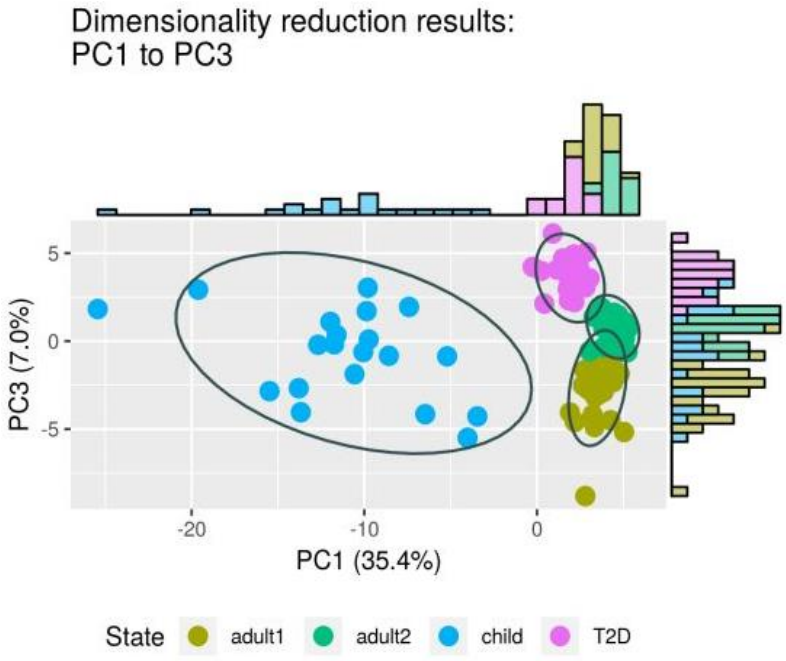
Τα δέκα πρώτα γονίδια, με τη μεγαλύτερη ποσοστιαία συμμετοχή στη διασπορά των PC1, PC2 και PC3, κατά σειρά, όπως διακρίνονται στα αποτελέσματα των διαγραμμάτων του τύπου της εικόνας 3.33 (επιλέγοντας κάθε φορά μία PC), είναι τα:

- PC1: *COL1A2*, *COL6A3*, *COL1A1*, *SPARC*, *COL3A1*, *TIMP3*, *THBS1*, *TGFBI*, *LUM*, *ACTA2*
- PC2: *GPX8*, *HSP90AB3P*, *COL5A1*, *LGALS1*, *TNFAIP6*, *RPS14*, *COL4A1*, *ANXA1*, *BGN*, *PXDN*
- PC3: *TRY6*, *REG1B*, *PRSS2*, *CPB1*, *REG1A*, *PRSS1*, *CLPS*, *CTRB2*, *HLA-J*, *OLFM4*

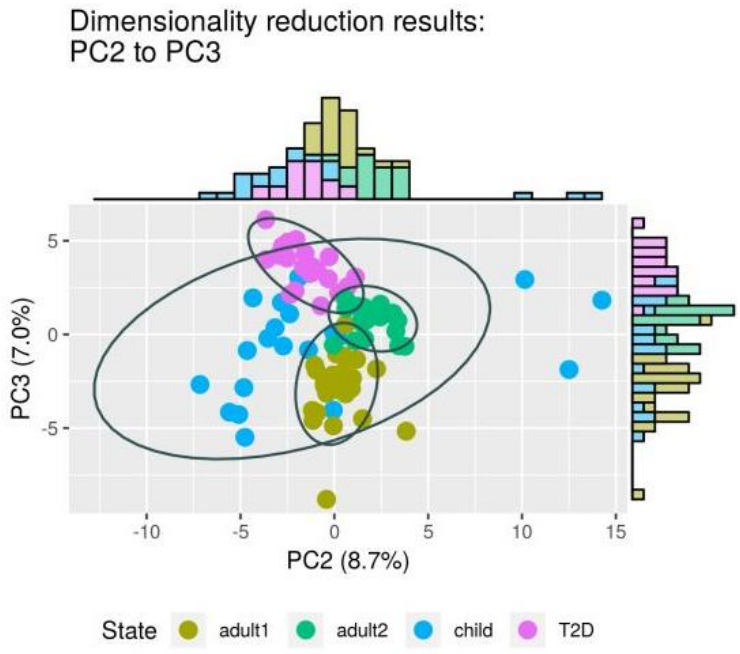
Σε αυτά της PC1 και δευτερευόντως της PC2, είναι έντονη η παρουσία γονιδίων που σχετίζονται με την εξωκυττάρια θεμέλια ουσία. Στης PC3, αντίθετα, κυριαρχούν γονίδια που έχουν σχέση με τον παγκρεατικό χυμό, που παράγεται από την εξωκρινή μοίρα, και τη φλεγμονώδη απάντηση.



Εικόνα 4.19: Προβολή των κυττάρων στις PC1 και PC2, χρωματισμένα βάσει της κατάστασης στην οποία ανήκουν.

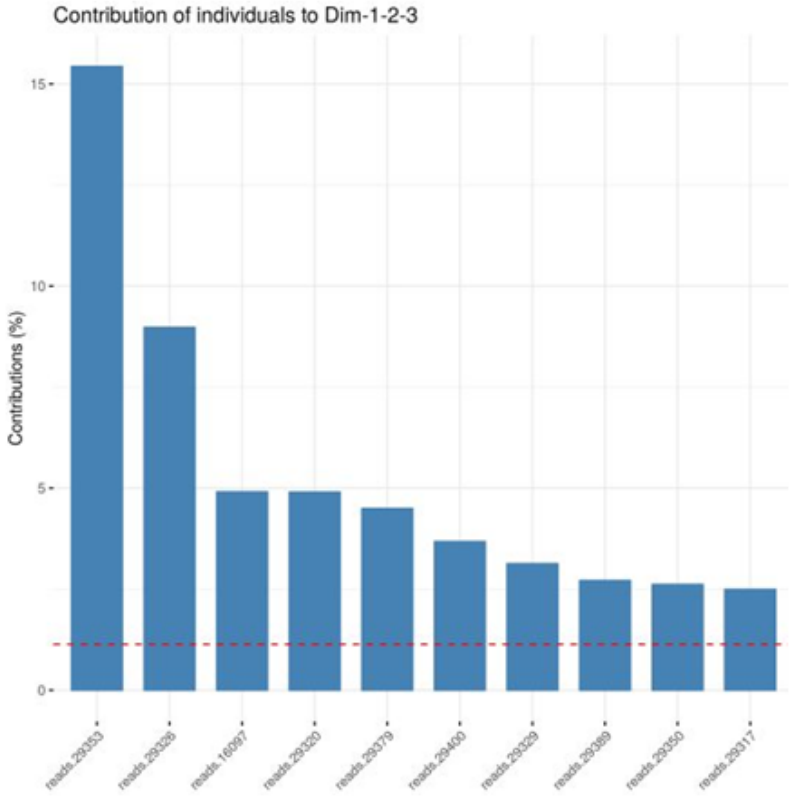


Εικόνα 4.20: Προβολή των κυττάρων στις PC1 και PC3, χρωματισμένα βάσει της κατάστασης στην οποία ανήκουν.

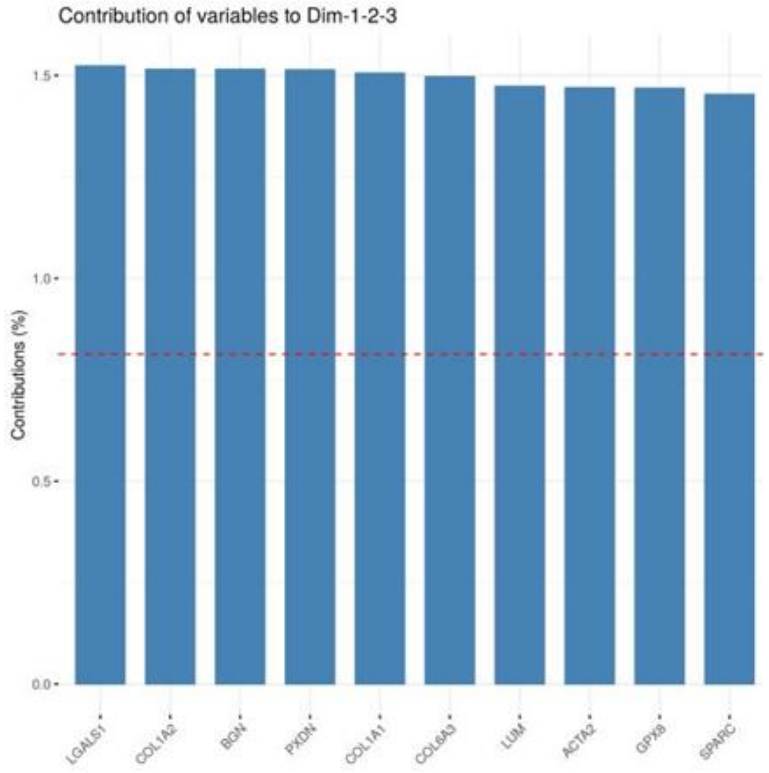


Εικόνα 4.21: Προβολή των κυττάρων στις PC2 και PC3, χρωματισμένα βάσει της κατάστασης στην οποία ανήκουν.

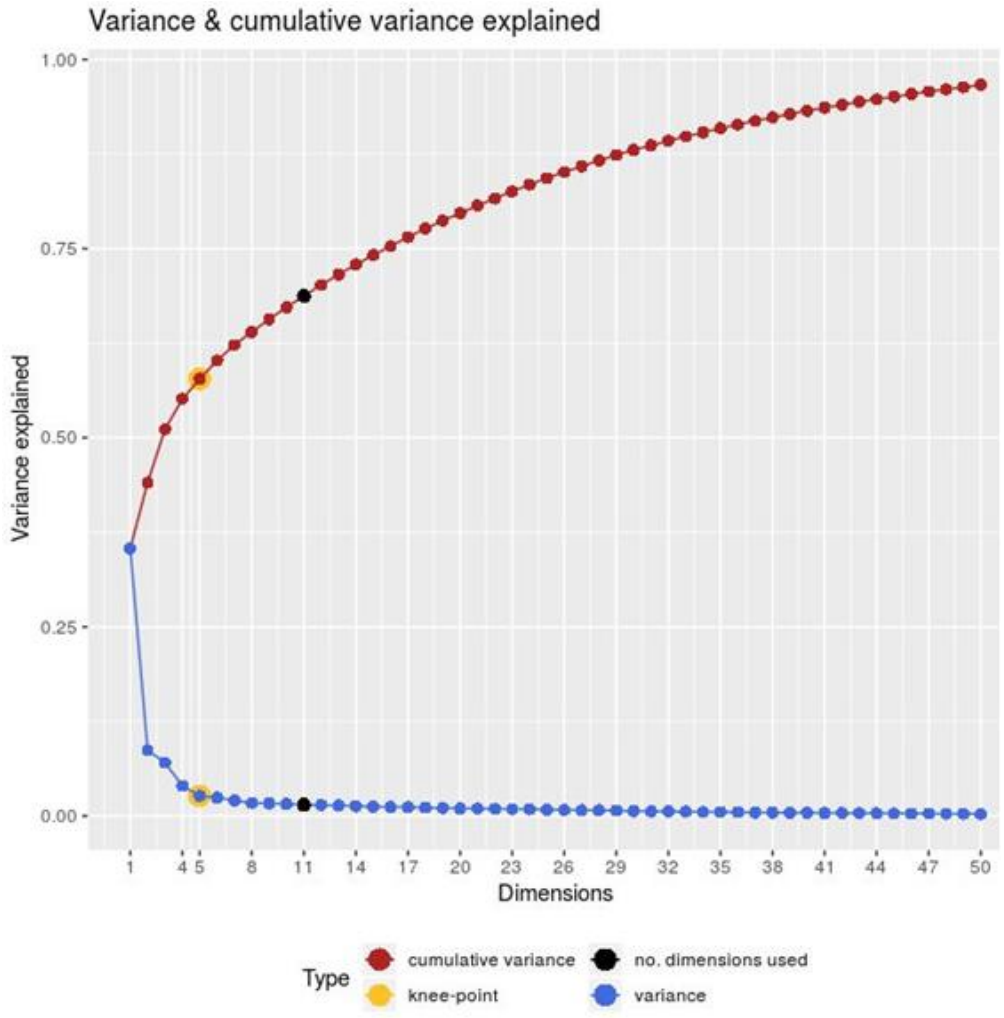
Αθροιστικά, στις PC1, PC2 και PC3, τα δέκα πρώτα κύτταρα και γονίδια βάσει τη συμμετοχής τους στη διασπορά, παρατίθενται στις εικόνες 4.22 και 4.23. Στην περίπτωση των κυττάρων, τα δύο πρώτα μόνο, αντιστοιχούν στο 25% περίπου. Πρόκειται για κύτταρα της κατάστασης «child», όπως ήταν αναμενόμενο, λόγω της μεγάλης διασποράς τους.



Εικόνα 4.22: Η ποσοστιαία συμβολή στη διακύμανση των τριών πρώτων PCs, των δέκα κυττάρων με τη μεγαλύτερη συνεισφορά.

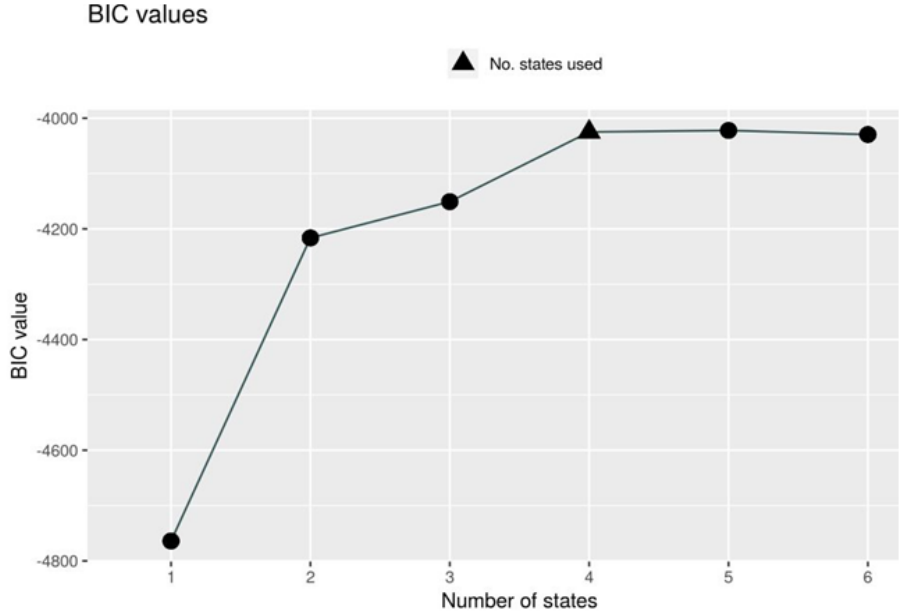


Εικόνα 4.23: Η ποσοστιαία συμβολή στη διακύμανση των τριών πρώτων PCs, των δέκα γονιδίων με τη μεγαλύτερη συνεισφορά.

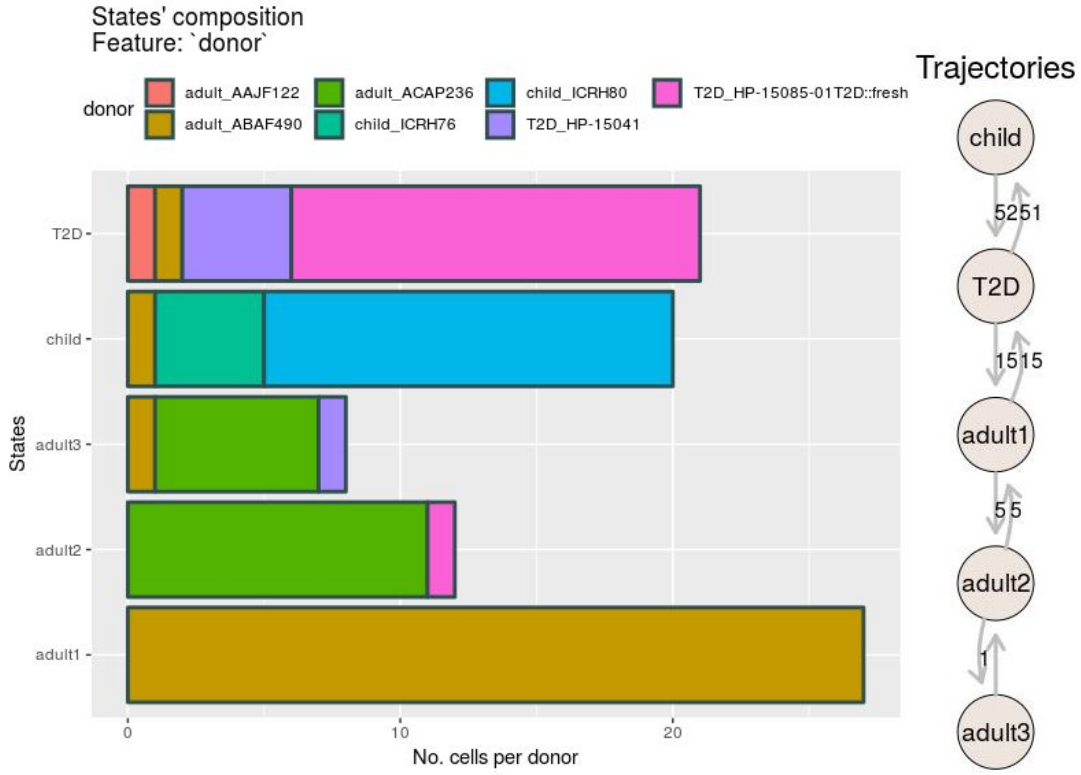


Εικόνα 4.24: Διάγραμμα της διακύμανσης κι αθροιστικής διακύμανσης ανά συνιστώσα των αποτελεσμάτων μείωσης της διαστατικότητας.

Η διακύμανση κι η αθροιστική διακύμανση ανά κύρια συνιστώσα, φαίνονται στην εικόνα 4.24. Το σημείο γονάτου, εντοπίζεται στις 5 PC, αλλά, στο πρότυπο, έχουν χρησιμοποιηθεί 11 PC, όπως αιτιολογήθηκε παραπάνω, με αθροιστική διακύμανση, 68,7%. Ακολούθως, χρησιμοποιώντας αυτές τις 11 PC και τον προκαθορισμένο τρόπο επιλογής αριθμού καταστάσεων, δημιουργούνται 4 καταστάσεις. Αν είχε χρησιμοποιηθεί μόνο η τιμή του BIC (εικόνα 4.25), θα προέκυπταν, 5 καταστάσεις, με τη σύσταση που φαίνεται στην εικόνα 4.26. Ουσιαστικά, αυξάνεται ο αριθμός των αριθμός των καταστάσεων της ομάδας των «ενηλίκων», με διάσπαση των κυττάρων ενός δότη, κάτι που δεν είναι επιθυμητό με την ύπαρξη λίγων κυττάρων.

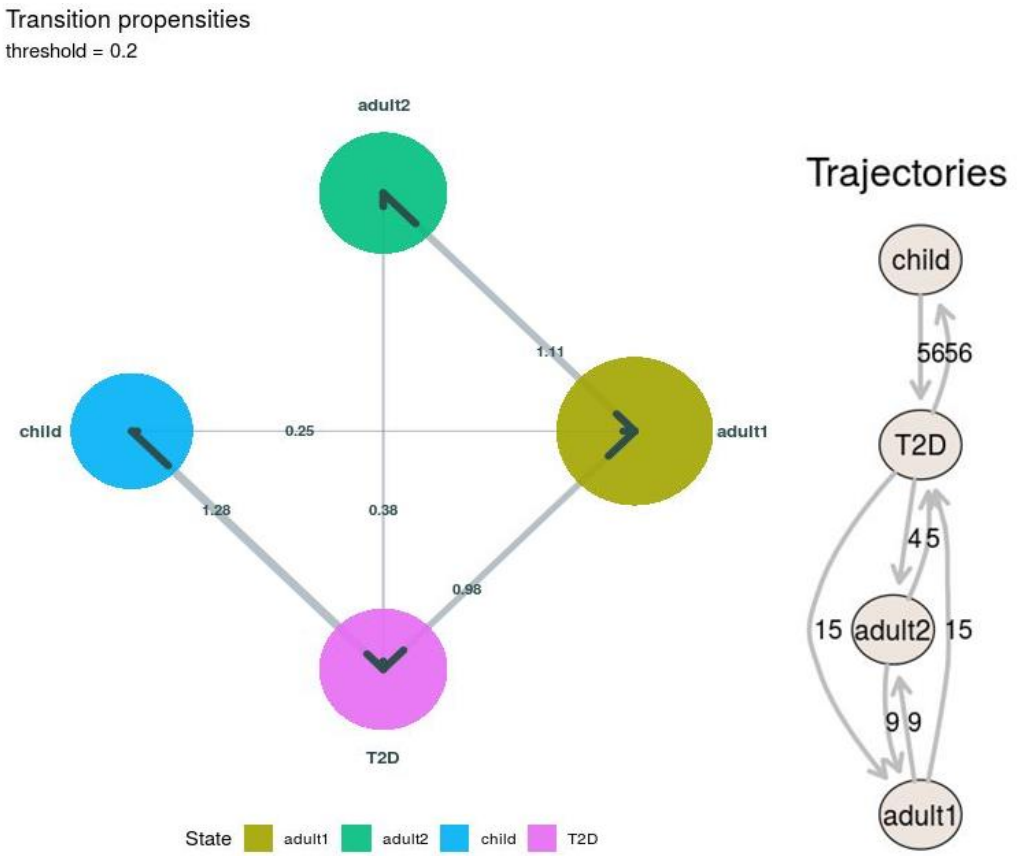


Εικόνα 4.25: Διάγραμμα των τιμών του BIC σε σχέση με τον αριθμό των καταστάσεων, επισημαίνοντας τον αριθμό των καταστάσεων που έχει επιλεγεί.



Εικόνα 4.26: Η σύσταση των καταστάσεων βάσει των δοτών κι οι τροχιές που σχηματίζονται με τον αντίστοιχο αριθμό κύριο γονιδίων.

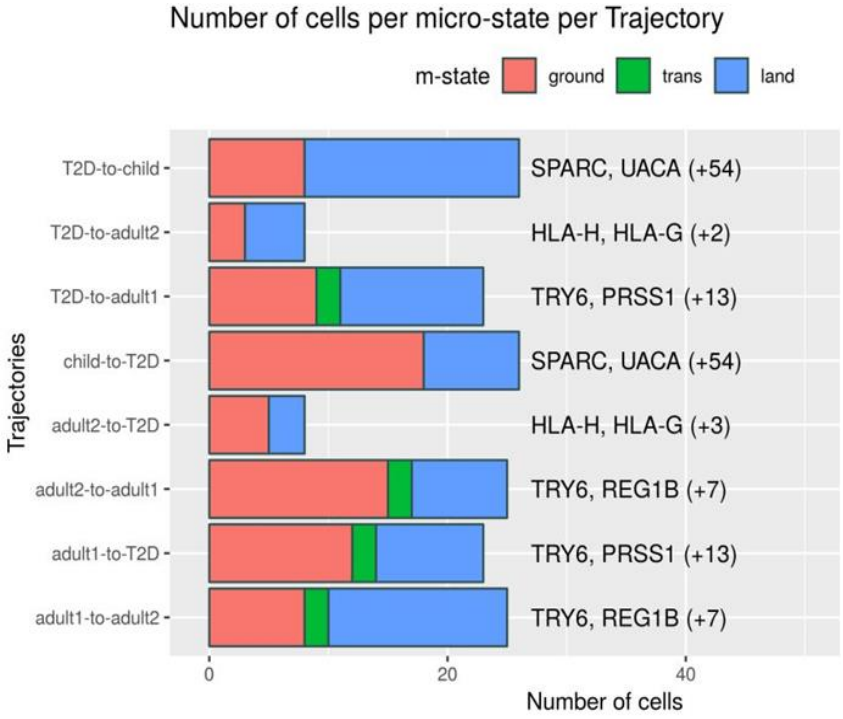
Οι μεταβάσεις (εικόνα 4.27), επιβεβαιώνουν την υπόθεση ότι τα κύτταρα των ατόμων με «ΣΔΤ2», βρίσκονται στο ενδιάμεσο της πορείας εξέλιξης από τα κύτταρα της ομάδας των «παιδιών» προς αυτά της ομάδας των «ενηλίκων». Πιθανώς κάποια κύτταρα των «παιδιών» είχαν λίγο πιο ώριμο προφίλ κι έτσι εντοπίζονται ορισμένα κύτταρα στη μετάβαση μεταξύ των καταστάσεων, «child» και «adult1», αν και δεν είναι αρκετά για να



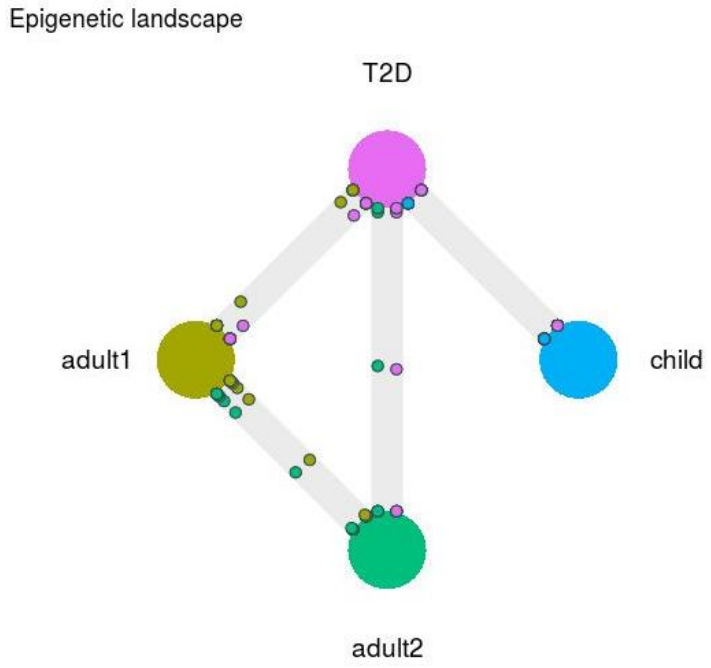
Εικόνα 4.27: Διάγραμμα των τάσεων μετάβασης μεταξύ των καταστάσεων κι οι τροχιές που σχηματίζονται με τον αντίστοιχο αριθμό κύριο γονιδίων.

σχηματιστούν τροχιές. Ο αριθμός των κύριων γονιδίων στις τροχιές που σχηματίζονται ανάμεσα στις καταστάσεις «child» και «T2D» (56), είναι πολύ μεγαλύτερος από τις υπόλοιπες (4, 5, 9, 15) – μία, ενδεχομένως, ένδειξη της παρουσίας περισσότερων ή / και μεγαλύτερων διαφορών στην έκφραση της κατάστασης «T2D» με την «child» σε σύγκριση με τις «adult1» και «adult2».

Παρατηρώντας τις μικρο-καταστάσεις των τροχιών (εικόνα 4.28), φαίνεται σχετική συμμετρία στην αναλογία τους μεταξύ των αντίθετων τροχιών. Μικρο-κατάσταση μετάβασης, δημιουργείται μόνο στις τροχιές που συμμετέχει η κατάσταση «adult1», χωρίς αυτό να σχετίζεται με τον αριθμό των κυττάρων. Στην εικόνα 4.29, φαίνεται πως σε όλες τις τροχιές, οι εκ των υστέρων πιθανότητες των κυττάρων στις τροχιές, έχουν κατά βάση πολύ μεγάλες τιμές για τις καταστάσεις στις οποίες ανήκουν και μόνο λίγα βρίσκονται στο ενδιάμεσο (π.χ. στην τροχιά «adult1-to-adult2»).

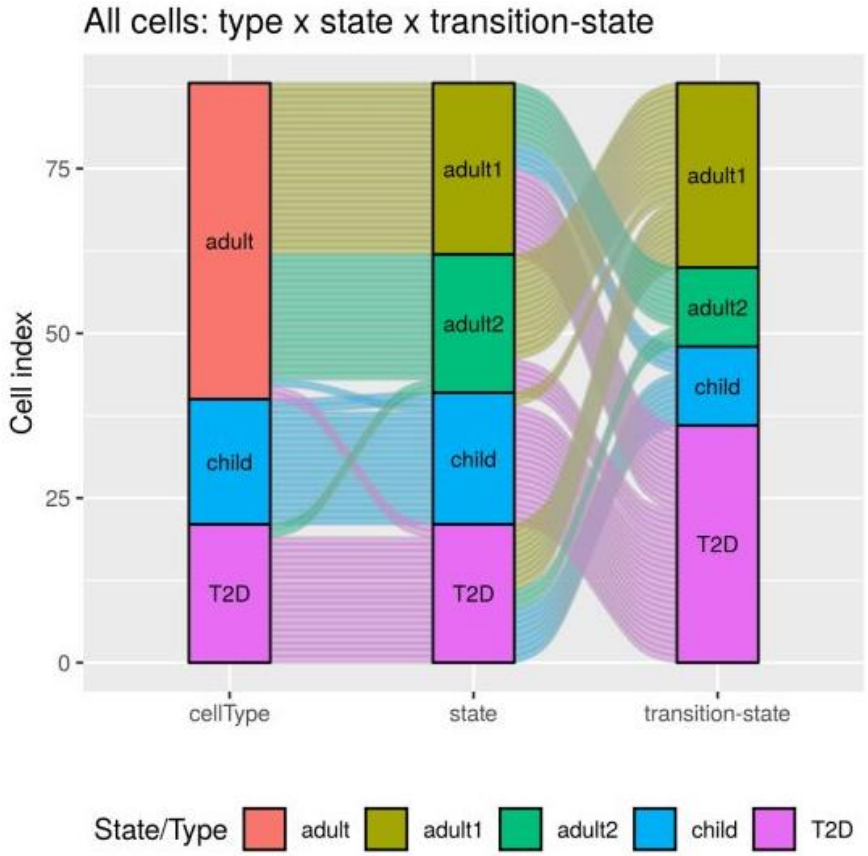


Εικόνα 4.28: Διάγραμμα των μικρο-καταστάσεων ανά τροχιά, με αναφορά των δύο σημαντικότερων κύριων γονιδίων ανά τροχιά κι επισήμανση του αριθμού των υπόλοιπων κύριων γονιδίων.



Εικόνα 4.29: Διάγραμμα του επιγενετικού τοπίου.

Συγκεντρωτικά, μπορούμε να δούμε την κατανομή των κυττάρων στις γνωστές ομάδες («cellType») και στις καταστάσεις που αντιστοιχούν στις δύο μεγαλύτερες εκ των υστέρων πιθανότητες (δηλ. την κατάσταση που ανήκουν και την κατάσταση με την οποία σχηματίζεται το ζεύγος μεταβάσεων ή τροχιών στο οποίο ανήκουν), στην εικόνα 4.30.

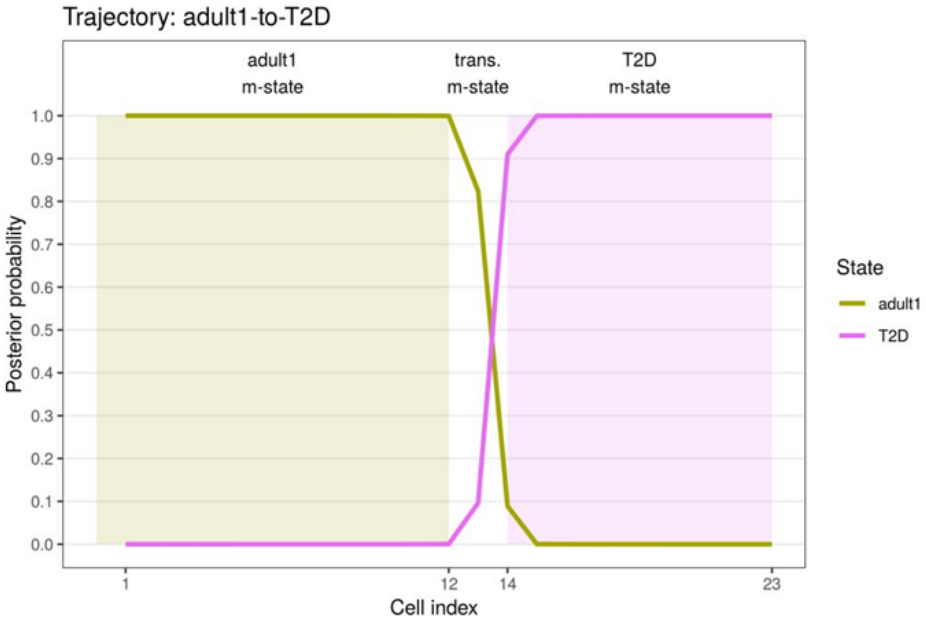


Εικόνα 4.30: Διάγραμμα που συνδέει κάθε κύτταρο με τα χαρακτηριστικά: τον κυτταρικό τύπο, την κατάσταση που ανήκει και την κατάσταση μετάβασης.

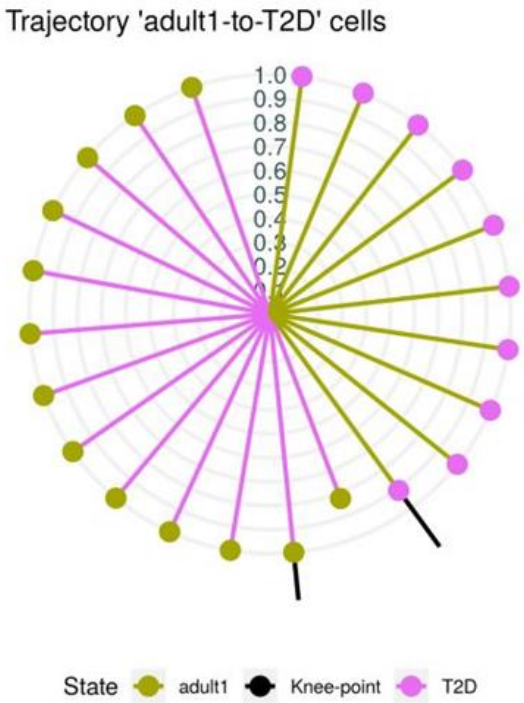
4.6 Αποτελέσματα – «Τροχιά adult1-to-T2D»

Στη συνέχεια, θα παρατεθούν αποτελέσματα για μία από τις τροχιές που αντιστοιχούν στη θεωρούμενη πορεία από-διαφοροποίησης· την «adult1-to-T2D».

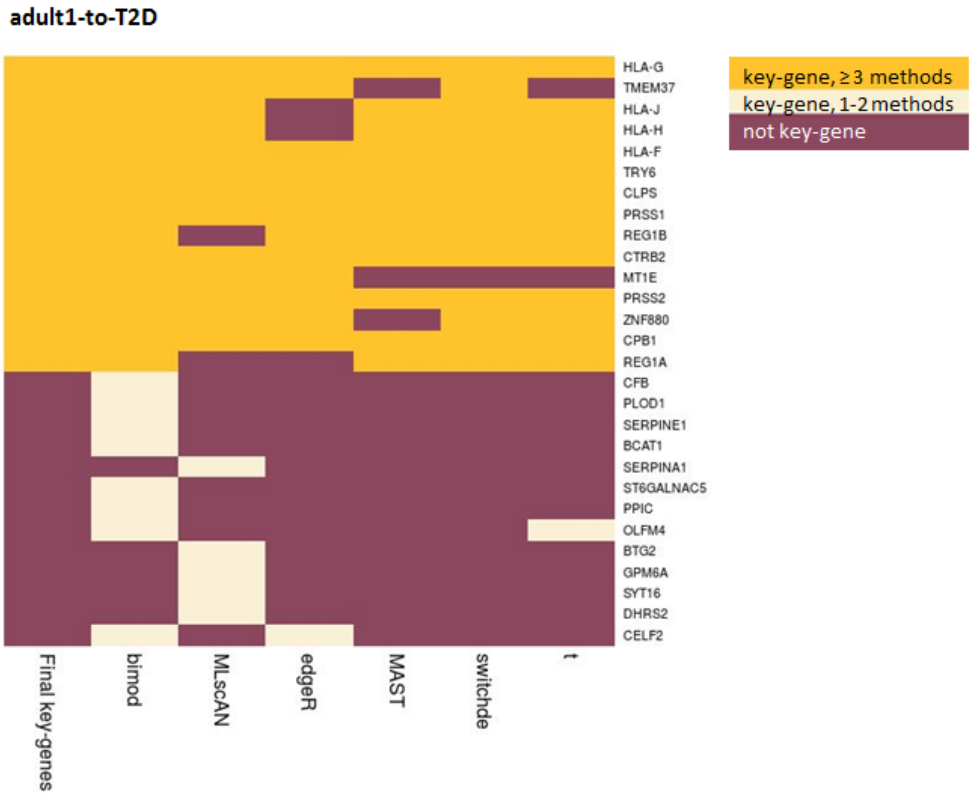
Στην εικόνα 4.31, φαίνεται η απότομη και «σύντομη» μετάβαση από την πολύ υψηλή πιθανότητα για την κατάσταση έναρξης και την πολύ χαμηλή πιθανότητα για την κατάσταση προορισμού, στη μικρο-κατάσταση έναρξης, στην αντίθετη κατάσταση, στη μικρο-κατάσταση προορισμού. Το ίδιο παρατηρείται και στην εικόνα 4.32, για κάθε κύτταρο ξεχωριστά.



Εικόνα 4.31: Διάγραμμα των μεταβολών των εκ των υστέρων πιθανοτήτων για τις καταστάσεις της επιλεγμένης τροχιάς, διατηρώντας τη διάταξη των κυττάρων της τροχιάς κι επισημαίνοντας την έκταση των μικρο-καταστάσεων.



Εικόνα 4.32: Στο διάγραμμα αυτό, οι σφαίρες, αντιστοιχούν στα κύτταρα που συμμετέχουν στην τροχιά «adult1-to-T2D». Τοποθετούνται σε απόσταση από το κέντρο, ανάλογη της εκ των υστέρων πιθανότητας για την κατάσταση στην οποία ανήκουν, έχοντας το χρώμα της κατάστασης στην οποία ανήκουν. Τα κύτταρα διατάσσονται με φθίνοντα τρόπο βάσει της πιθανότητας για την κατάσταση «adult1», αντίστροφα από τη φορά των δεικτών του ρολογιού (όπως, δηλαδή, διατάσσονται και στην τροχιά). Το χρώμα της ακτίνας, είναι αυτό της δεύτερης κατάστασης τής μετάβασης που συμμετέχουν.



Εικόνα 4.33: Τα κύρια γονίδια της τροχιάς «adult1-to-T2D» που αναγνωρίζονται από κάθε μέθοδο της συνάρτησης *kg_voting*.

Για να αναγνωριστούν τα κύρια γονίδια, χρησιμοποιήθηκε η συνάρτηση, *kg_voting* (ενότητα 2.2.8.3), με όλες τις διαθέσιμες μεθόδους. Στην εικόνα 4.33, διακρίνονται τα γονίδια που κάθε μέθοδος θεώρησε κύρια. Οι μέθοδοι *bimod* και *MLscAN*, έχουν αναγνωρίσει αρκετά περισσότερα κύρια γονίδια σε σχέση με τις υπόλοιπες ενώ η *bimod*, είναι η μόνη που περιλαβάνει στα αποτελέσματά της όλα τα γονίδια που τελικά θεωρούνται κύρια.

Τα 15 γονίδια που θεωρήθηκαν τελικά κύρια, είναι τα ακόλουθα, κατά σειρά σημαντικότητας, με πληροφορίες από το GeneCards [62]:

- *TRY6*
    - ψευδογονίδιο 2 της πρωτεάσης της σερίνης 3
    - φαίνεται πως το πρωτεϊνικό του προϊόν είναι παρόμοιο με το θρυψινογόνο (πρόδρομη μορφή παγκρεατικού ενζύμου του παγκρεατικού χυμού της εξωκρινούς μοίρας)
    - σχετίζεται με την αλκοολική παγκρεατίτιδα
- *PRSS1*
    - πρωτεάση σερίνης της οικογένειας της θρυψίνης

- μεταλλάξεις σε αυτό το γονίδιο σχετίζονται με την κληρονομική παγκρεατίτιδα

- *HLA-G*
  - μείζον σύμπλεγμα ιστοσυμβατότητας τάξης Ι
  - παράλογο των βαρέων αλυσίδων
- *CLPS*
  - Συμπαράγοντας της παγκρεατικής λιπάσης
  - η παραλλαγή Cys-109, σχετίζεται με αυξημένο κίνδυνο εμφάνισης σακχαρώδη διαβήτη τύπου 2
- *CTRB2*
  - χυμοθρυψινογόνο Β2
  - σχετίζεται με τον σακχαρώηδη διαβήτη τύπου 2
- *PRSS2*
  - πρωτεάση σερίνης της οικογένειας της θρυψίνης
  - εντοπίζεται στον παγκρεατικό χυμό σε υψηλά επίπεδα κι η προς τα πάνω ρύθμισή του είναι χαρακτηριστικό της παγκρεατίτιδας
- *HLA-F*
  - μείζον σύμπλεγμα ιστοσυμβατότητας τάξης Ι, αλυσίδα α F
  - σχετίζεται με τον σαχαρώδη διαβήτη τύπου 1
- *CPB1*
  - δείκτης οξείας παγκρεατίτιδας
- *REG1B*
  - Περιλαμβάνεται στον παγκρεατικό χυμό
  - το πρωτεϊνικό του προϊόν μοιάζει αρκετά με αυτό του *REG1A* (αναφέρεται παρακάτω)
- *HLA-H*
  - δυνητικά ανήκει στο μείζον σύμπλεγμα ιστοσυμβατότητας τάξης Ι, αλυσίδα α Η

- ψευδογονίδιο

- *HLA-J*
  - μείζον σύμπλεγμα ιστοσυμβατότητας τάξης Ι
  - ψευδογονίδιο, πιθανώς παραγόμενο από το *HLA-A*

- *ZNF880*
  - πρωτεΐνη 880 δακτυλίου ψευδαργύρου

- *TMEM37*
  - διαμεμβρανική πρωτεΐνη 37
  - συμμετέχει στο μονοπάτι της ενεργοποιούμενης από μιτογόνα πρωτεϊνικής κινάσης

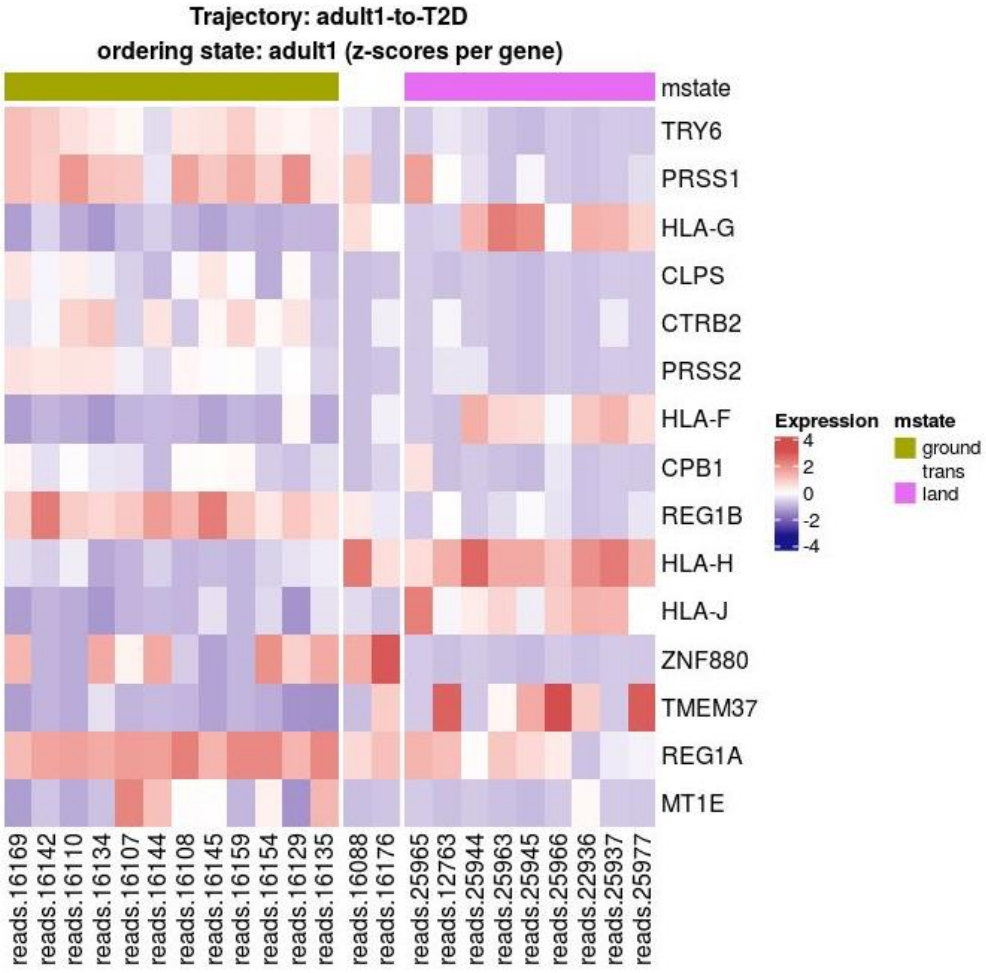
- *REG1A*
  - παραγώμενα από τα νησίδια πρωτεϊνη 1-άλφα
  - περιλαμβάνεται στον παγκρεατικό χυμό
  - σχετίζεται με την αναγέννηση των νησιδιακών κυττάρων κι ίσως έχει ρόλο στην παγκρεατική λιθογένεση
  - σχετίζεται με την παγκρεατίτιδα, την τροπική παγκρεατίτδα και τον όψιμης έναρξης σακχαρώδη διαβήτη των νέων

- *MT1E*
  - μεταλλοθειονίνη1Ε
  - πολυμορφισμοί του σχετίζονται με τον κίνδυνο εμφάνισης ΣΔΤ2 κι επιπλοκών του

Από αυτά, ήταν:

- Γονίδια υπογραφής:
  - adult α (212 γονίδια): 2 {*HLA-G*, *HLA-J*}
  - adult β (376 γονίδια): 1 {*TMEM37*}

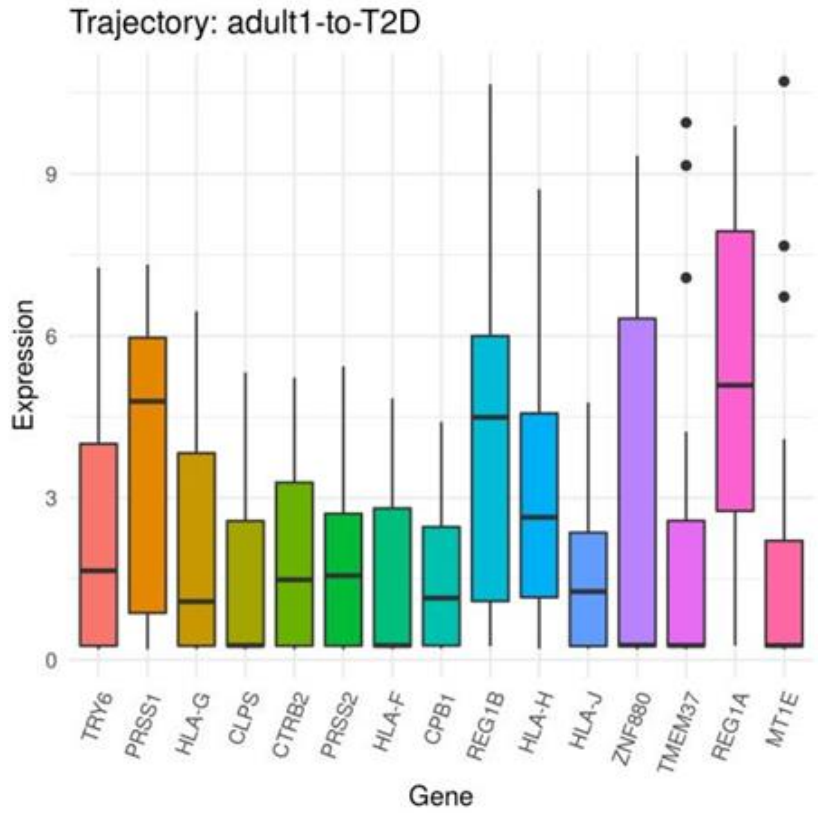
- Άτυπα εκφρασμένα γονίδια:
  - Adult α, T2D β (4 γονίδια): 1 {*HLA-G*}
  - Child β, T2D β (52 γονίδια): 2 {*HLA-G*, *HLA-J*}



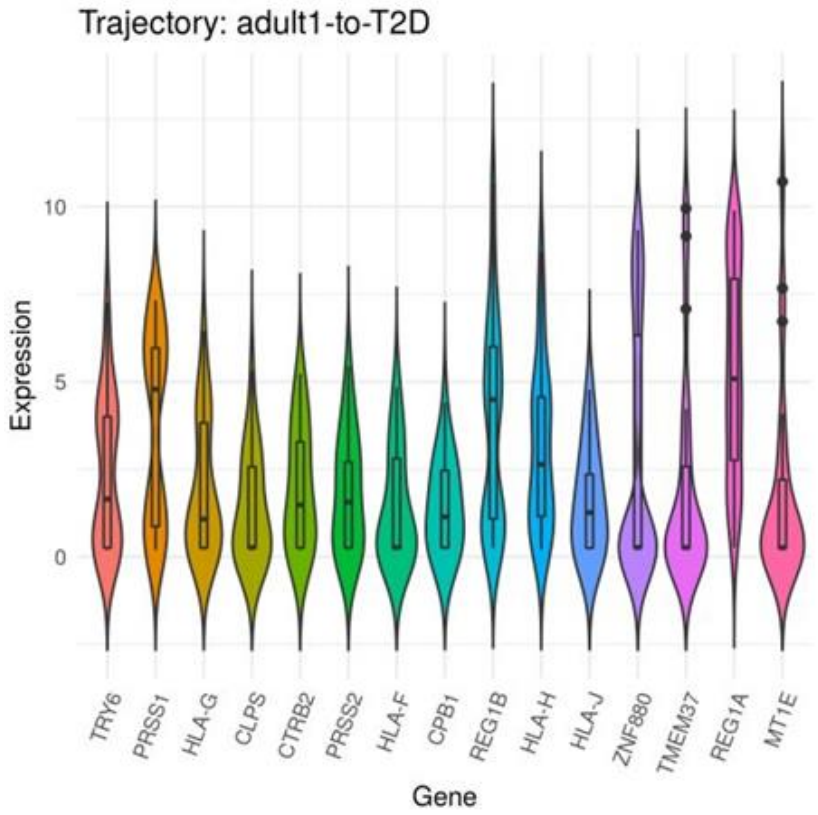
Εικόνα 4.34: Χάρτης θερμότητας των κύριων γονιδίων της τροχιάς «adult1-to-T2D» στα διαταγμένα κύτταρά της.

Στην εικόνα 4.34, βρίσκεται ο χάρτης θερμότητας για τα κύτταρα της τροχιάς ανά μικρο-κατάσταση. Στις περισσότερες περιπτώσεις (5 / 15), το γονίδιο εκφράζεται σε σημαντικά υψηλότερο βαθμό στα κύτταρα της μικρο-κατάστασης έναρξης.

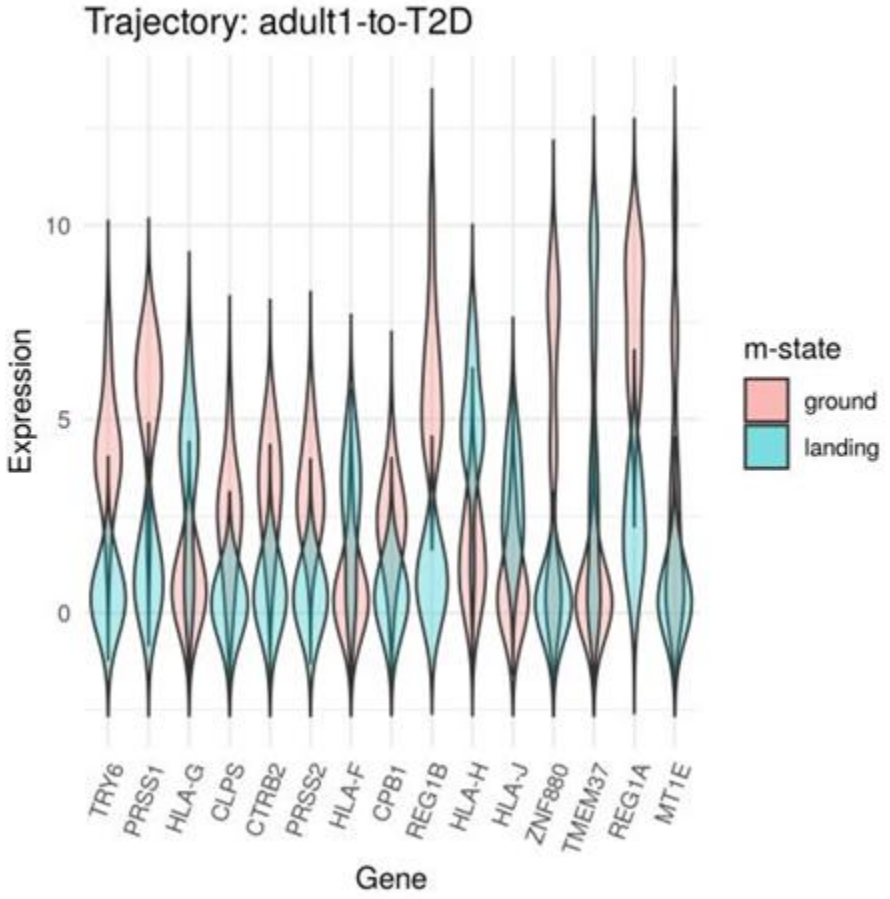
Στις εικόνες 4.35 - 4.37, βρίσκονται, τα θηκογράμματα, τα διαγράμματα βιολιού με ομαλοποίηση στο σύνολο των κυττάρων της τροχιάς και τα διαγράμματα βιολιού με ομαλοποίηση στις μικρο-καταστάσεις έναρξης και προορισμού, ανά κύριο γονίδιο. Υψηλότερα επίπεδα έκφρασης και σε περισσότερα κύτταρα, παρατηρείται στα γονίδια, *REG1A*, *REG1B* και *PRSS1*. Στα ίδια γονίδια, μαζί με το *TRY6*, εντοπίζεται περισσότερο το αναμενόμενο πρότυπο για διτροπική έκφραση – δηλαδή, συγκέντρωση της έκφρασης σε δύο απομακρυσμένες μεταξύ τους περιοχές, που αντανακλούν την έκφραση στις μικρο-καταστάσεις έναρξης και προορισμού.



Εικόνα 4.35: Θηκογράμματα των κύριων γονίδιων της τροχιάς «adult1-to-T2D» διαταγμένα με σειρά σημαντικότητας.



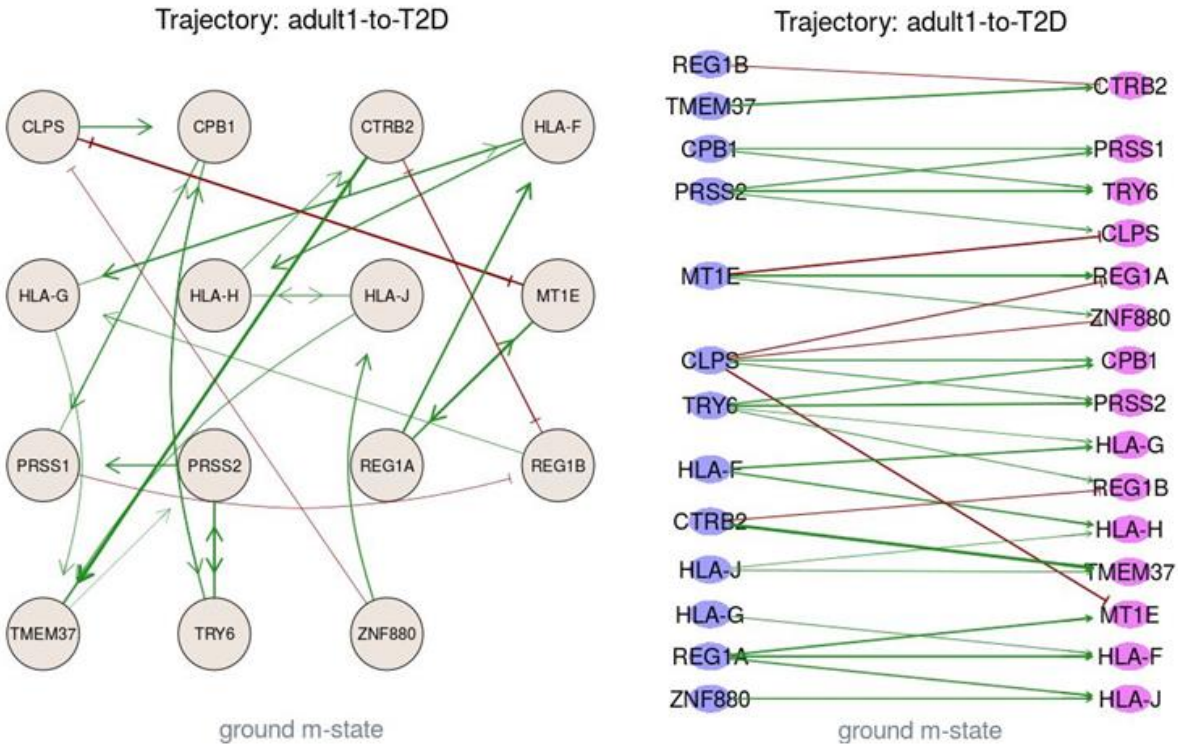
Εικόνα 4.36: Διαγράμματα βιολιού των κύριων γονίδιων της τροχιάς «adult1-to-T2D» διαταγμένα με σειρά σημαντικότητας, εφαρμόζοντας ομαλοποίηση (πυρήνας Gauss με σταθερό εύρος ζώνης=0,95).



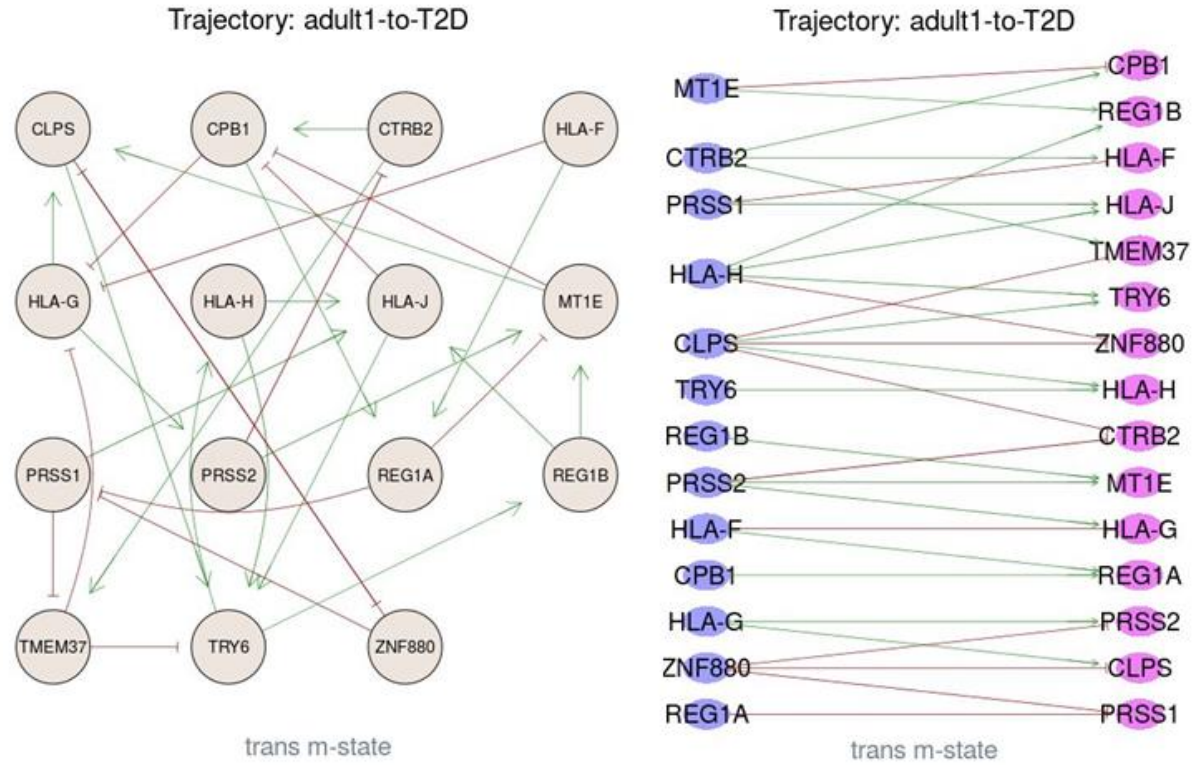
Εικόνα 4.37: Διαγράμματα βιολιού των κύριων γονίδιων της τροχιάς «adult1-to-T2D» διαταγμένα με σειρά σημαντικότητας, για τις μικρο-καταστάσεις έναρξης και προορισμού, εφαρμόζοντας ομαλοποίηση (πυρήνας Gauss με σταθερό εύρος ζώνης=0,95).

Στη συνέχεια, στα GRNs που δημιουργούνται ανά μικρο-κατάσταση (εικόνες 4.38 - 4.40), επιλέχθηκε να περιληφθούν οι ακμές μόνο για τους δύο σημαντικότερους ρυθμιστές ανά στόχο. Σε αυτά των μικρο-καταστάσεων έναρξης και μετάβασης, κυριαρχούν οι ενισχυτικές αλληλεπιδράσεις (24 / 30 και 18 / 30), ενώ στη μικρο-κατάσταση προορισμού, δεν υπερτερεί κάποια κατηγορία. Στη μικρο-κατάσταση μετάβασης, παρατηρούνται μικρότερες τιμές βαρών, ενώ στις άλλες δύο, υπάρχουν μεγαλύτερες διαφορές, κι ιδιαίτερα στη μικρο-κατάσταση προορισμού.

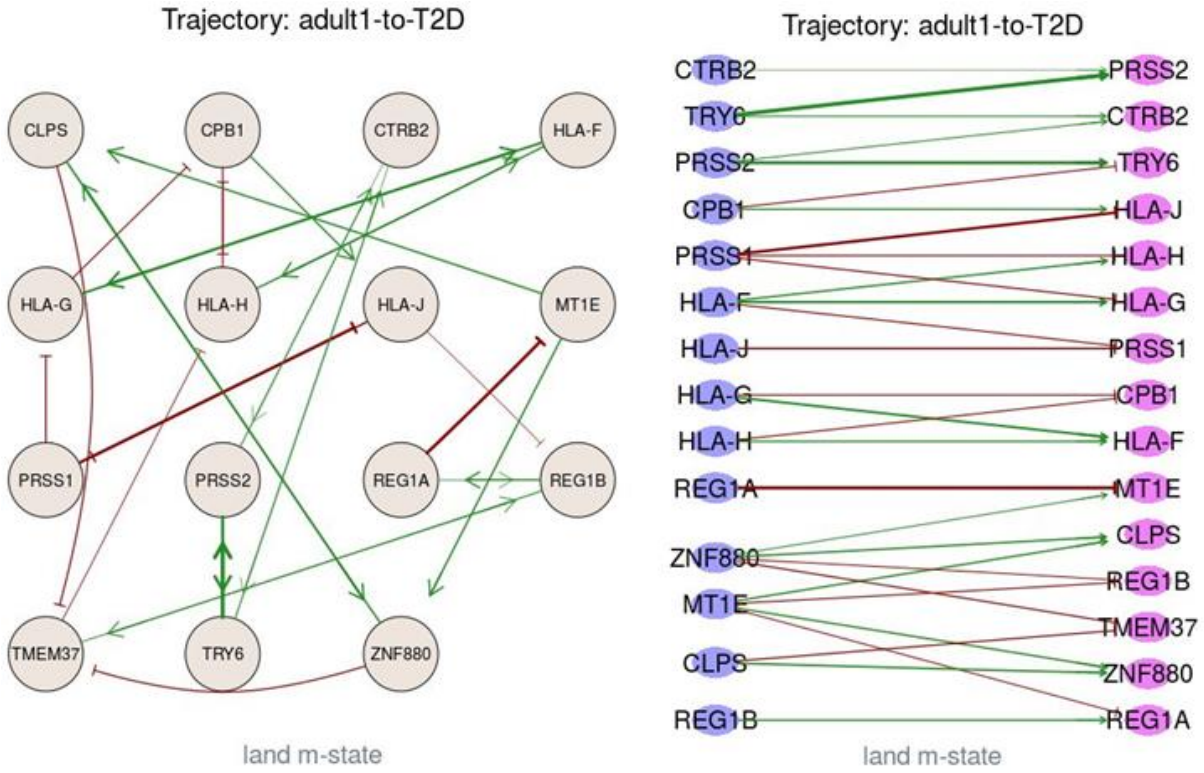
Εστιάζοντας, για παράδειγμα, στο γονίδιο *CTRB2* (εικόνα 4.41), αρχικά, παρατηρείται ότι στην τροχιά σταδιακά μειώνονται τα επίπεδα έκφρασης. Μάλιστα, το ίδιο πρότυπο έκφρασης, με αντίθετη πορεία, παρατηρείται στη θεωρούμενη συνέχεια της απο-διαφοροποίησης, στην τροχιά «T2D-to-child» (εικόνα 4.43). Οι πέντε ρυθμιστές με το μεγαλύτερο άθροισμα απόλυτων τιμών διαφορών στα βάρη μεταξύ των διαδοχικών μικρο-καταστάσεων, είναι τα γονίδια: *TMEM37*, *HLA-H*, *HLA-J*, *PRSS2* και *CLPS*. Σε όλα, διαπιστώνονται διαφοροποιήσεις στον τύπο της αλληλεπίδρασης στις διαφορετικές μικρο-καταστάσεις, με περισσότερες αλλαγές, στα γονίδια, *HLA-G* και *HLA-J*. Οι μεγαλύτερες απόλυτες τιμές των βαρών, εντοπίζονται στα γονίδια, *PRSS2* και *TMEM37* (εικόνα 4.42).



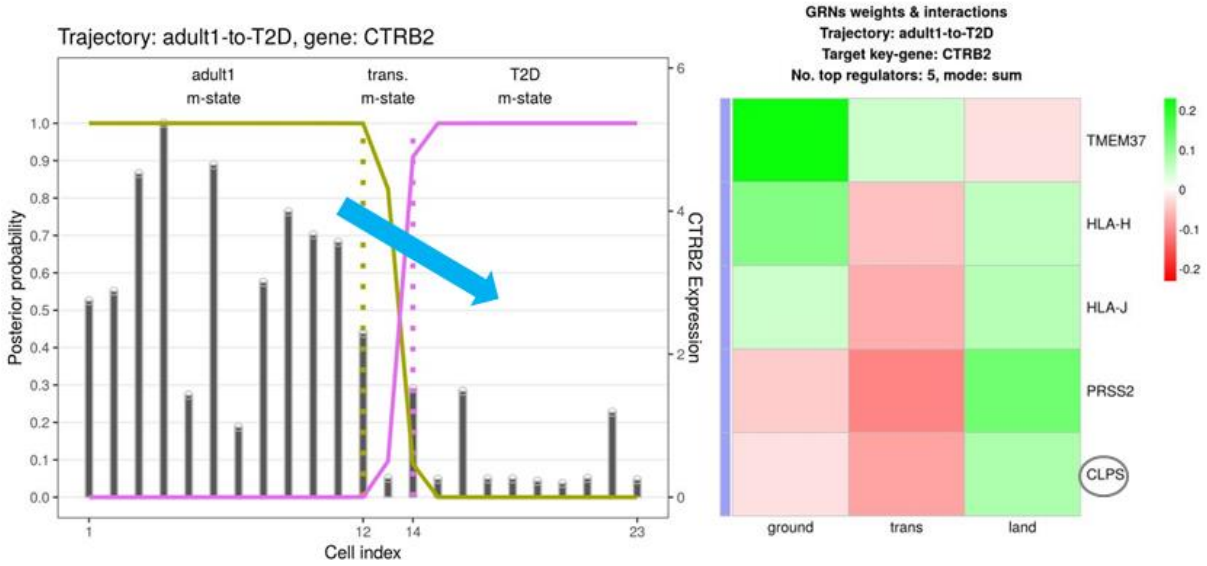
Εικόνα 4.38: GRN της μικρο-κατάστασης έναρξης της τροχιάς «adult1-to-T2D», εμφανίζοντας μόνο τις ακμές για τους δύο σημαντικότερους ρυθμιστές κάθε στόχου.



Εικόνα 4.39: GRN της μικρο-κατάστασης μετάβασης της τροχιάς «adult1-to-T2D», εμφανίζοντας μόνο τις ακμές για τους δύο σημαντικότερους ρυθμιστές κάθε στόχου.



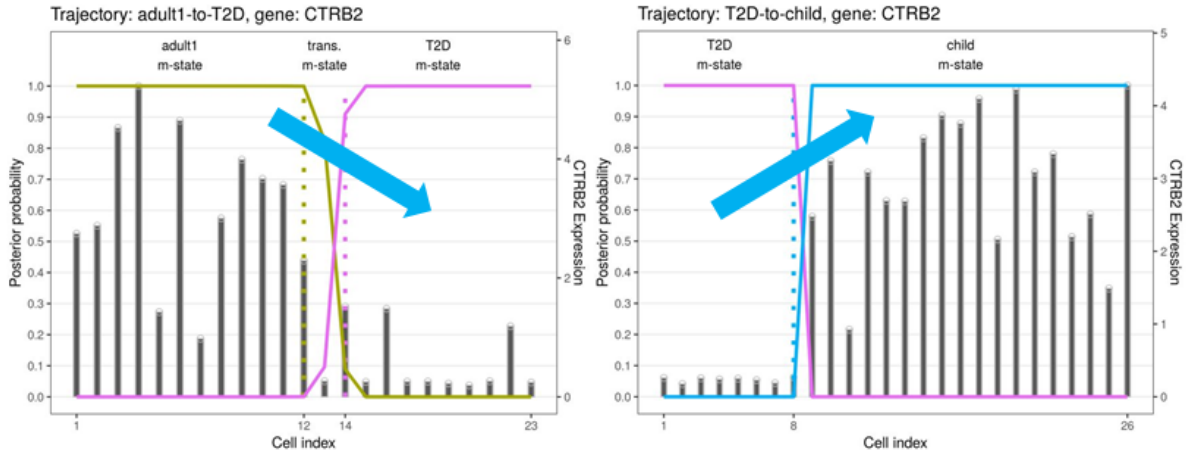
Εικόνα 4.40: GRN της μικρο-κατάστασης προορισμού της τροχιάς «adult1-to-T2D», εμφανίζοντας μόνο τις ακμές για τους δύο σημαντικότερους ρυθμιστές κάθε στόχου.



Εικόνα 4.41: Διάγραμμα της έκφρασης του γονιδίου *CTRB2* στα κύτταρα της τροχιάς «adult1-to-T2D» και χάρτης θερμότητας των βαρών, των πέντε ρυθμιστών με τις μεγαλύτερες μεταβολές, των GRN ανά μικρο-κατάσταση.

	ground	trans	land
TRY6	-0.04339637	-0.06250000	0.14338004
PRSS1	-0.05994597	-0.08266129	0.02003878
HLA-G	-0.02630877	-0.06653226	-0.02745599
CLPS	-0.02338363	-0.08266129	0.08059779
PRSS2	-0.04055036	-0.10483871	0.13641444
HLA-F	0.02726059	0.08064516	0.10431582
CPB1	-0.07769865	0.07258065	-0.09484815
REG1B	-0.11776774	-0.07862903	-0.05758311
HLA-H	0.11279857	-0.05443548	0.06186614
HLA-J	0.05266056	-0.06653226	0.06992575
ZNF880	0.02653063	0.07258065	0.08608788
TMEM37	0.23063139	0.05443548	-0.01962695
REG1A	-0.10347982	0.05645161	-0.05184824
MT1E	-0.05758695	-0.06451613	0.04601091

Εικόνα 4.42: Τα βάρη όλων των ρυθμιστών των GRN ανά μικρο-κατάσταση της τροχιάς «adult1-to-T2D» για το γονίδιο *CTRB2*.

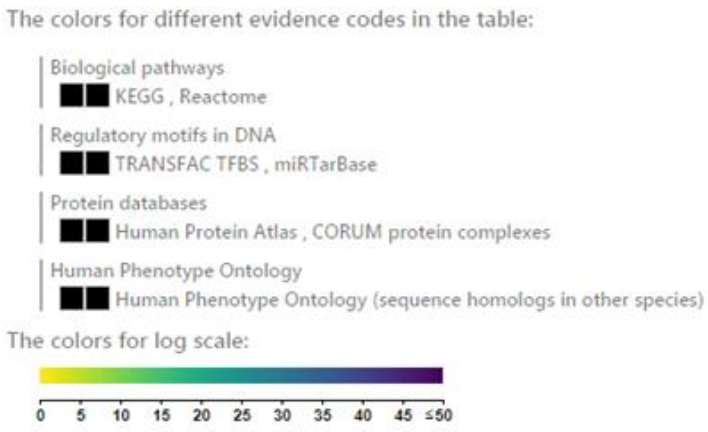


Εικόνα 4.43: Διαγράμματα της έκφρασης του γονιδίου *CTRB2*, στα κύτταρα των τροχιών «adult1-to-T2D» και «T2D-to-child».

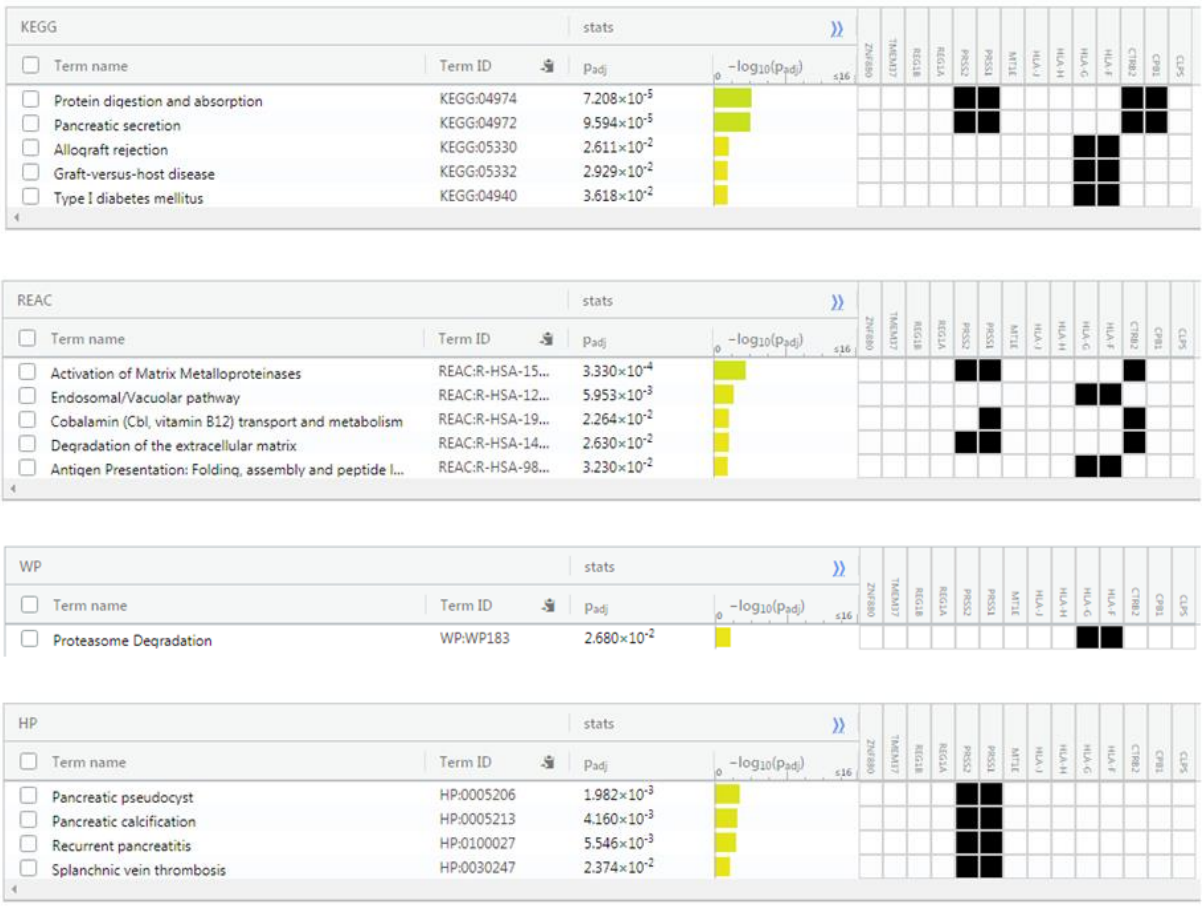
4.6.1 Περαιτέρω διερεύνηση των μονοπατιών και δικτύων που συμμετέχουν τα κύρια γονίδια

Στη συνέχεια, χρησιμοποιήθηκε ο κατάλογος των κύριων γονιδίων για τη διερεύνηση των μονοπατιών που συμμετέχουν, των σχετιζόμενων φαινοτύπων καθώς και των δικτύων αλληλεπίδρασης μεταξύ «απλών» πρωτεϊνών, μεταγραφικών παραγόντων και miRNA.

Διατηρώντας τις προκαθορισμένες ρυθμίσεις του g:Profiler [63], και λαμβάνοντας υπόψη το σχετικό υπόμνημα, της εικόνας 4.44, προκύπτουν τα αποτελέσματα της εικόνας 4.45, που αφορούν στα μονοπάτια και στους φαινοτύπους.



Εικόνα 4.44: Υπόμνημα για τα διαγράμματα των αποτελεσμάτων του g:Profiler [63].



Εικόνα 4.45: Τα αποτελέσματα του g:Profiler για τα μονοπάτια και τους φαινοτύπους [63].

Τα μονοπάτια, λαμβάνοντας υπόψη και τα γονίδια που συμμετέχουν, μπορούν να ενταχθούν σε τρεις κατηγορίες:

- Σχετιζόμενα με το εξωκρινές πάγκρεας
    - Πέψη κι απορρόφηση πρωτεϊνών
    - Παγκρεατική έκκριση
- Σχετιζόμενα με το ανοσοποιητικό σύστημα

  - Απόρριψη αλλομοσχεύματος
  - Νόσος μοσχεύματος έναντι του ξενιστή
  - Σακχαρώδης διαβήτης τύπου 1
  - Μονοπάτι ενδοσώματος / κενοτοπίου
  - Παρουσίαση αντιγόνου: δίπλωμα, συναρμολόγηση και φόρτωση πεπτιδίου του μείζονος συμπλέγματος ιστοσυμβατότητας τάξης Ι

- Σχετιζόμενα με την εξωκυττάρια θεμέλια ουσία
  - Ενεργοποίηση των μεταλλοπρωτεϊνασών της θεμέλιας ουσίας
  - Αποδόμηση της εξωκυττάριας θεμέλιας ουσίας

Επιπλέον, κι οι φαινότυποι, σχετίζονται κυρίως με την εξωκρινή μοίρα του παγκρέατος (παγκρεατική ψευδοκύστη, παγκρεατική ασβέστωση, επαναλαμβανόμενη παγκρεατίτιδα).

Διατηρώντας τις προκαθορισμένες ρυθμίσεις του NetworkAnalyst [64], προκύπτουν τα αποτελέσματα που συνοψίζονται στην εικόνα 4.46, κι αφορούν στα μονοπάτια και στις αλληλεπιδράσεις πρωτεϊνών, μεταγραφικών παραγόντων και miRNA. Στις εικόνες 4.47 - 4.52, βρίσκονται αντιπροσωπευτικά παραδείγματα – ένα από κάθε κατηγορία: μονοπάτια, αλληλεπιδράσεις μεταξύ πρωτεϊνών, αλληλεπιδράσεις γονιδίων – μεταγραφικών παραγόντων, αλληλεπιδράσεις γονιδίων – miRNA, κι αλληλεπιδράσεις γονιδίων – μεταγραφικών παραγόντων – miRNA.

Τα υπο-δίκτυα αλληλεπιδράσεων που περιλαμβάνονται, έχουν τουλάχιστον τρεις κόμβους κι είναι πρώτου βαθμού (αυτό σημαίνει ότι παραβλέπονται οι άμεσες συνδέσεις των κόμβων που αντιστοιχούν στα κύρια γονίδια). Το μέγεθος των κόμβων, αντιστοιχεί στον βαθμό του. Ο βαθμός (degree), είναι ο αριθμός των ακμών του κόμβου. Το χρώμα των κόμβων, αντιστοιχεί στην τιμή p (εικόνα 4.47) ή στην τιμή της κεντρικότητας ενδιαμεσότητας (εικόνες 4.49 – 4.52). Η κεντρικότητα ενδιαμεσότητας (betweeness centrality), είναι ο αριθμός των πιο σύντομων μονοπατιών που διέρχονται από τον κόμβο.

Σε ό,τι αφορά τα μονοπάτια (εικόνα 4.47), υπάρχει αρκετή επικάλυψη των αποτελεσμάτων, και μάλιστα αυτών με τις μικρότερες τιμές p – πιο σημαντικά, με αυτά του g:Profiler, όπως για παράδειγμα τα: πέψη κι απορρόφηση πρωτεϊνών, παγκρεατική έκκριση, απόρριψη αλλομοσχεύματος, ασθένεια μοσχεύματος έναντι του ξενιστή και σακχαρώδης διαβήτης τύπου 1. Στον γράφο, οι κόμβοι – μονοπάτια, συνδέονται μεταξύ τους μέσω των γονιδίων που μοιράζονται (το πολύ τρία), όπως φαίνεται στην εικόνα 4.48.

Protein-protein Interactions
IMEX Interactome

Networks	Nodes	Edges	Seeds
subnetwork1	30	31	7
subnetwork2	8	7	1
subnetwork3	3	2	1

GRN: Gene-miRNA Interactions
TarBase, miRTarBase

Networks	Nodes	Edges	Seeds
subnetwork1	23	22	1
subnetwork2	12	11	1
subnetwork3	4	3	1
subnetwork4	3	2	1

GRN: TF-gene Interactions
JASPAR

Networks	Nodes	Edges	Seeds
subnetwork1	57	95	15

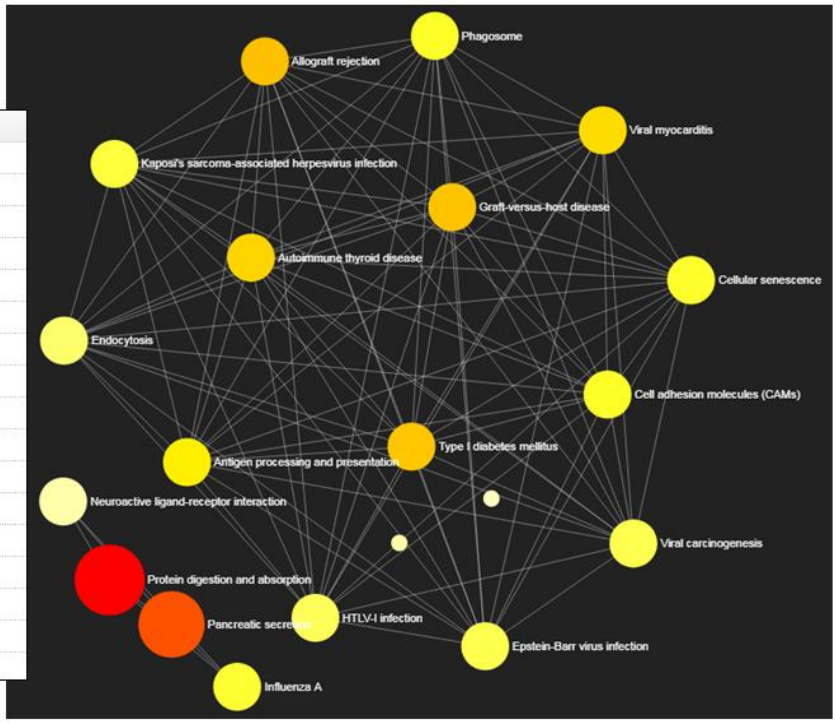
GRN: TF-miRNA Coregulatory Network
TarBase, miRTarBase

Networks	Nodes	Edges	Seeds
subnetwork1	67	66	6
subnetwork2	4	3	1
subnetwork3	3	2	1
subnetwork4	3	2	1

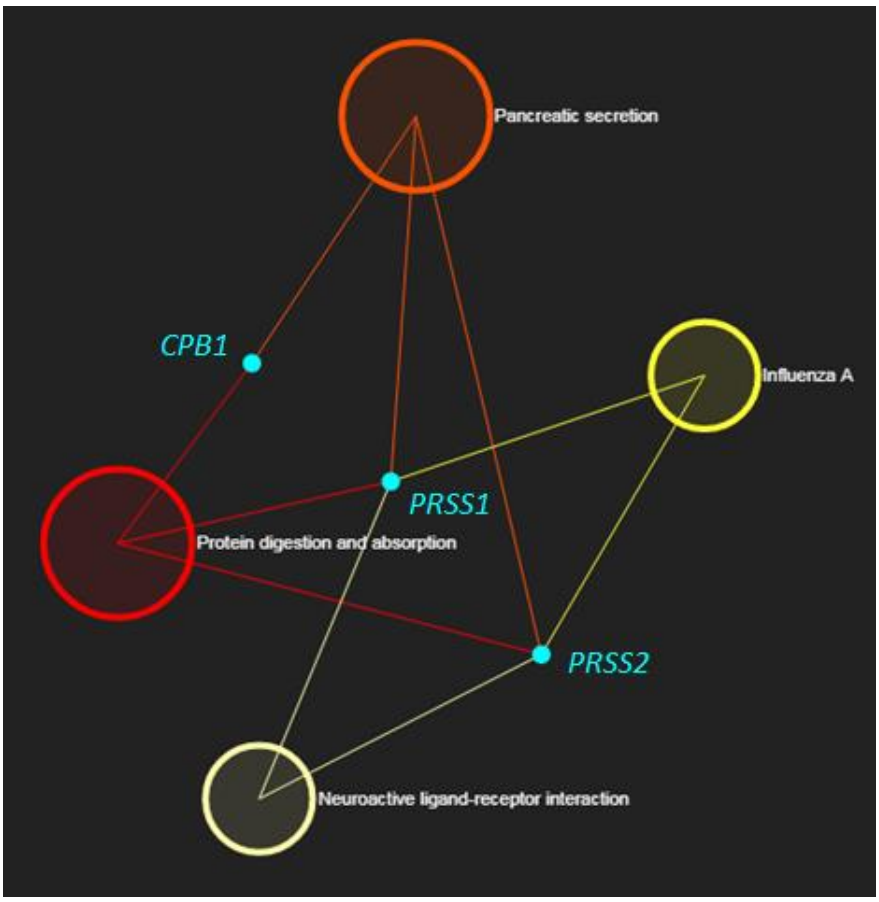
Εικόνα 4.46: Σύνοψη των αποτελεσμάτων των δικτύων αλληλεπιδράσεων του NetworkAnalyst [64]. Φύτρα (seeds), χαρακτηρίζονται τα γονίδια βάσει των οποίων γίνεται η αναζήτηση και στην προκειμένη περίπτωση είναι τα κύρια γονίδια.

Enrichment analysis
Database: KEGG
Operation: Union Submit

Name	Hits	Pval	AdjP
Protein digestion and abs	3/90	5.15e-5	0.0106
Pancreatic secretion	3/98	6.65e-5	0.0106
Allograft rejection	2/38	4.86e-4	0.0396
Graft-versus-host disease	2/41	5.66e-4	0.0396
Type I diabetes mellitus	2/43	6.23e-4	0.0396
Autoimmune thyroid dise	2/53	9.46e-4	0.0501
Viral myocarditis	2/59	0.00117	0.0532
Antigen processing and p	2/77	0.00199	0.079
Cell adhesion molecules (	2/146	0.00698	0.24
Phagosome	2/152	0.00755	0.24
Cellular senescence	2/160	0.00834	0.24
Influenza A	2/167	0.00906	0.24
Kaposi's sarcoma-associa	2/186	0.0111	0.273
Epstein-Barr virus infecti	2/201	0.0129	0.274
Viral carcinogenesis	2/201	0.0129	0.274
HTLV-I infection	2/219	0.0152	0.303
Endocytosis	2/244	0.0187	0.35



Εικόνα 4.47: Δίκτυο των μονοπατιών από τα αποτελέσματα του NetworkAnalyst [64].

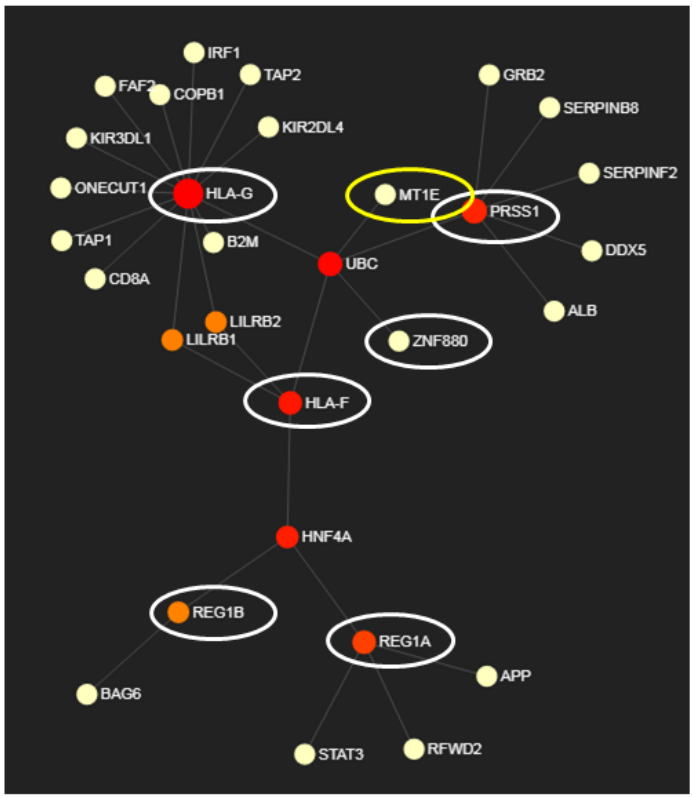


Εικόνα 4.48: Περιοχή του δικτύου των μονοπατιών από τα αποτελέσματα του NetworkAnalyst [64], όπου φαίνεται η σύνδεση των κόμβων–μονοπατιών με τα κύρια γονίδια της τροχιάς «adult1-to-T2D», τα οποία επισημαίνονται.

Protein-protein Interactions
IMEX Interactome

Name	Degree	Betweeness
HLA-G	13	244.5
PRSS1	6	130
UBC	5	210.3333
REG1A	4	81
HLA-F	4	163.5
HNF4A	3	146
REG1B	2	28
LILRB1	2	29.33333
LILRB2	2	29.33333
IRF1	1	0
STAT3	1	0
B2M	1	0
KIR3DL1	1	0
KIR2DL4	1	0
SERPINF	1	0

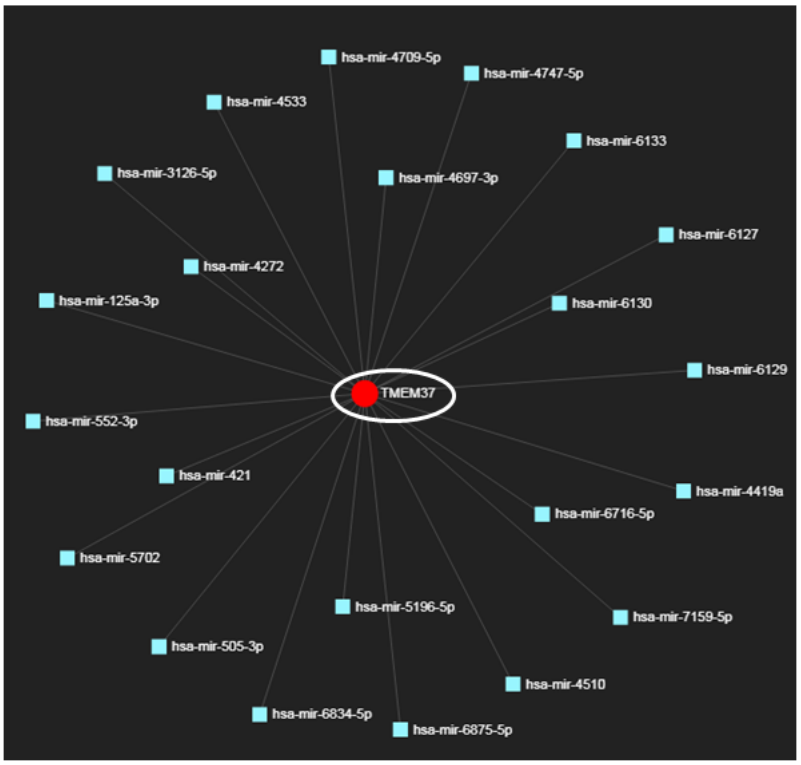
Name	Degree	Betweeness
HLA-G	13	244.5
UBC	5	210.3333
HLA-F	4	163.5
HNF4A	3	146
PRSS1	6	130
REG1A	4	81
LILRB1	2	29.33333
LILRB2	2	29.33333
REG1B	2	28
IRF1	1	0
STAT3	1	0
B2M	1	0
KIR3DL1	1	0
KIR2DL4	1	0
SERPINF	1	0



Εικόνα 4.49: Δίκτυο αλληλεπίδρασης μεταξύ πρωτεϊνών (βάση δεδομένων: IMEX Interactome) από τα αποτελέσματα του NetworkAnalyst [64] κι οι αντίστοιχοι πίνακες των τιμών του βαθμού και της ενδιαμεσότητας κεντρικότητας για τους κυριότερους – γονίδια.

GRN:
Gene-miRNA Interactions
TarBase, miRTarBase

Name	Degree	Betweeness
TMEM37	22	231
hsa-mir-505-3	1	0
hsa-mir-552-3	1	0
hsa-mir-421	1	0
hsa-mir-125a-	1	0
hsa-mir-3126-	1	0
hsa-mir-4272	1	0
hsa-mir-4419a	1	0
hsa-mir-4510	1	0
hsa-mir-4533	1	0
hsa-mir-4697-	1	0
hsa-mir-4709-	1	0
hsa-mir-4747-	1	0
hsa-mir-5196-	1	0
hsa-mir-5702	1	0

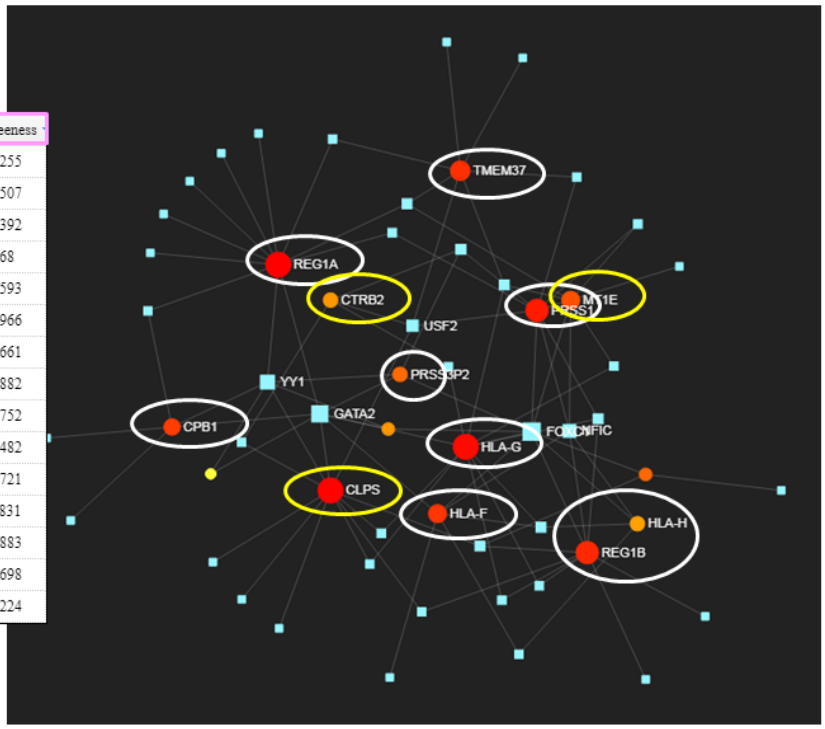


Εικόνα 4.50: Δίκτυο αλληλεπίδρασης γονιδίων και miRNA (βάσεις δεδομένων: TarBase, miRTarBase) από τα αποτελέσματα του NetworkAnalyst [64] κι οι αντίστοιχοι πίνακες των τιμών του βαθμού και της ενδιαμεσότητας κεντρικότητας για τους κυριότερους – γονίδια.

GRN:
TF-gene Interactions
JASPAR

Name	Degree	Betweeness
REG1A	11	378.1255
HLA-G	11	302.168
CLPS	11	351.5392
REG1B	9	179.5661
PRSS1	9	228.9966
FOXC1	9	266.8593
GATA2	8	369.0507
TMEM37	7	162.8882
MT1E	6	89.42883
HLA-F	6	137.6482
YY1	6	137.8752
CPB1	5	126.3721
NFIC	5	72.70698
HLA-H	4	22.12165
PRSS3P2	4	56.27057

Name	Degree	Betweeness
REG1A	11	378.1255
GATA2	8	369.0507
CLPS	11	351.5392
HLA-G	11	302.168
FOXC1	9	266.8593
PRSS1	9	228.9966
REG1B	9	179.5661
TMEM37	7	162.8882
YY1	6	137.8752
HLA-F	6	137.6482
CPB1	5	126.3721
USF2	4	118.1831
MT1E	6	89.42883
NFIC	5	72.70698
ELK1	3	63.67224

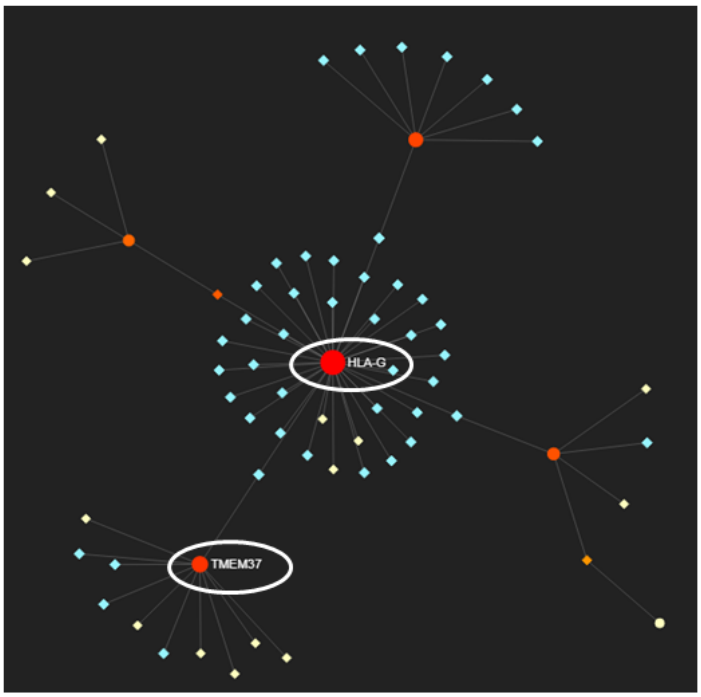


Εικόνα 4.51: Δίκτυο αλληλεπίδρασης γονιδίων και μεταγραφικών παραγόντων (βάση δεδομένων: JASPAR) από τα αποτελέσματα του NetworkAnalyst [64] κι οι αντίστοιχοι πίνακες των τιμών του βαθμού και της ενδιαμεσότητας κεντρικότητας για τους κυριότερους κόμβους – γονίδια.

GRN:
TF-miRNA Coregulatory Network
TarBase, miRTarBase

Name	Degree	Betweeness
HLA-G	37	2012
TMEM37	11	605
CLPS	8	434
CPB1	5	314
HLA-F	4	192
TBP	2	65
ATF2	2	248
hsa-miR-1	2	605
hsa-miR-5	2	464
hsa-miR-6	2	360
TFAP2C	1	0
TFAP2A	1	0
SPI1	1	0
SP1	1	0
RXRB	1	0

Name	Degree	Betweeness
HLA-G	37	2012
TMEM37	11	605
hsa-miR-1	2	605
hsa-miR-5	2	464
CLPS	8	434
hsa-miR-6	2	360
CPB1	5	314
ATF2	2	248
HLA-F	4	192
TBP	2	65
TFAP2C	1	0
TFAP2A	1	0
SPI1	1	0
SP1	1	0
RXRB	1	0



Εικόνα 4.52: Δίκτυο συρρύθμισης γονιδίων από μεταγραφικούς παράγοντες και miRNA (βάσεις δεδομένων: TarBase, miRTarBase) από τα αποτελέσματα του NetworkAnalyst [64] κι οι αντίστοιχοι πίνακες των τιμών του βαθμού και της ενδιαμεσότητας κεντρικότητας για τους κυριότερους κόμβους – γονίδια.

4.7 Συμπεράσματα

- Τροχιές:
    - Οι τροχιές φαίνεται πως επιβεβαιώνουν την τοποθέτηση των κυττάρων των ατόμων με «ΣΔΤ2», ανάμεσα στο προφίλ των κυττάρων των «ενηλίκων» κι αυτό των κυττάρων των «παιδιών».
    - Ελάχιστα κύτταρα των καταστάσεων «child» και «adult1», βρίσκονται στις μεταξύ τους σχηματιζόμενες μεταβάσεις. Σε αντίθεση, τα κύτταρα της «T2D», κατά 62% συμμετέχουν στις μεταβάσεις που σχηματίζουν με τις καταστάσεις «adult1» και «adult2», και κατά 38% σε αυτές με την «child».
    - Δε διακρίνονται υπο-πληθυσμοί στα κύτταρα του ίδιου δότη, αν και στην ομάδα «adult», κατανέμονται σε διαφορετικές καταστάσεις τα κύτταρα των δύο δοτών με συνεισφορά κυττάρων περισσότερων του ενός. Βέβαια, είναι και περιορισμένος ο συνολικός αριθμός κυττάρων (88).
- Κύρια γονίδια που σχετίζονται με την τροχιά «adult1-to-T2D»:
    - Αρκετά από αυτά τα γονίδια, σχετίζονται κυρίως με τη λειτουργία της εξωκρινούς μοίρας του παγκρέατος, του ανοσοποιητικού (τα αλλήλια δεν είναι γνωστά) και της εξωκυττάριας θεμέλιας ουσίας (αν και τα β-κύτταρα δεν έχουν τη δυνατότητα να παραγάγουν τα στοιχεία της βασικής μεμβράνης, αλλά, εκκρίνουν παράγοντες που προσελκύουν τα κύτταρα που τη συνθέτουν [65]). Πιθανώς η αρχική επιλογή του συνόλου των γονιδίων να λειτουργεί περιοριστικά.
    - Σε άλλο άρθρο [66], όπου χρησιμοποιούνται μονήρη κύτταρα ενηλίκων με ή χωρίς ΣΔΤ2, βρέθηκαν διαφορικά εκφρασμένα γονίδια που κατά 92% δεν είχαν προηγουμένως σχετιστεί με τη λειτουργία ή την ανάπτυξη των νησιδίων του Langerhans. Από αυτά, 4: *PPP1R1A* (στα β-κύτταρα, αυξημένη έκφραση στα άτομα με ΣΔΤ2), *HLA-H* (στα α-κύτταρα, μειωμένη έκφραση στα άτομα με ΣΔΤ2), *NPY* (στα α-κύτταρα, αυξημένη έκφραση στα άτομα με ΣΔΤ2) και *FXYD2* (στα β-κύτταρα, αυξημένη έκφραση στα άτομα με ΣΔΤ2), υπάρχουν στα 123 γονίδια που επιλέχθηκαν και 3 (*PPP1R1A*, *HLA-H*, *NPY*) στα κύρια γονίδια όλων των τροχιών.

- Μονοπάτια που σχετίζονται με την τροχιά «adult1-to-T2D»:
    - Όπως αναφέρθηκε και για τα κύρια γονίδια, τα μονοπάτια που θεωρούνται σημαντικά, σχετίζονται κυρίως με την εξωκρινή μοίρα του παγκρέατος, το ανοσοποιητικό και την εξωκυττάρια θεμέλια ουσία. Η έλλειψη περαιτέρω πληροφοριών (π.χ. οι γονότυποι) κι εξειδίκευσής τους στα β-κύτταρα, δεν επιτρέπουν να γίνουν δεκτά τα αποτελέσματα χωρίς σημαντικές επιφυλάξεις.
- Δίκτυα αλληλεπιδράσεων που σχετίζονται με την τροχιά «adult1-to-T2D»:
    - Αν και τα περισσότερα δίκτυα αλληλεπιδράσεων περιλαμβάνουν λίγα από τα κύρια γονίδια και πρόκειται για δίκτυα πρώτου βαθμού με δεδομένα τα γονίδια ενδιαφέροντος (κύρια γονίδια), σε τρία από αυτά, περιλαμβάνονται αρκετά κύρια γονίδια (6, 7 και 15 από τα 15), έχοντας ταυτόχρονα, μεγάλο βαθμό κι υψηλή ενδιαμεσότητα κεντρικότητας.

5. ΓΕΝΙΚΑ ΣΥΜΠΕΡΑΣΜΑΤΑ

Ως προς το πακέτο R, οι κύριοι στόχοι που τέθηκαν για την ανάπτυξή του, και περιγράφονται στην ενότητα 1.2.1, έχουν επιτευχθεί σε αρκετό βαθμό. Ωστόσο, υπάρχουν περιθώρια βελτίωσης, με κύρια τα ακόλουθα: α) ενσωμάτωση, στον κυρίως κώδικα, νεώτερων μεθοδολογιών, ιδίως για τη μείωση της διαστατικότητας και τον χειρισμό των πολλών μηδενικών τιμών που αναμένονται, που θα είναι άμεσα κι ευκολότερα διαθέσιμες στον χρήστη, β) ενσωμάτωση της χρήσης πακέτων που επιτρέπουν την ευχερέστερη διαχείριση αντικειμένων που απαιτούν πολλή μνήμη, όπως το *bigmemory* [67], γ) εκτέλεση εκτενέστερων ελέγχων της εισόδου με πιο πληροφοριακά μηνύματα σφαλμάτων, και δ) σύγκριση της αποτελεσματικότητας με άλλα πακέτα και πρότυπα διάφορων τοπολογιών.

Ως προς τη χρήση του πακέτου R για τη διερεύνηση της τροχιάς από-διαφοροποίησης των β-κυττάρων των νησιδίων του Langerhans, όπως αναφέρθηκε στο κεφάλαιο 4 κι ιδιαίτερα στην ενότητα 4.7, τουλάχιστον στο επίπεδο των τροχιών, φαίνεται να «επιβεβαιώνεται» η πορεία από-διαφοροποίησης, από τα κύτταρα των καταστάσεων των «ενηλίκων» προς την κατάσταση των ατόμων «ΣΔΤ2», χωρίς να υπάρχει καμία σύνδεση των «ενηλίκων» με την κατάσταση της ομάδας των «παιδιών». Με εξαίρεση ελάχιστα κύτταρα, το ίδιο ισχύει και για τις μεταβάσεις. Τα κύρια γονίδια, επηρεάζονται από το βήμα της αρχικής επιλογής ενός υπο-συνόλου τους, αλλά, έχει αναφερθεί η παρουσία «μη-αναμενόμενων» διαφορικά εκφρασμένων γονιδίων, σε αντίστοιχες συγκρίσεις κυττάρων [66], τα οποία μάλιστα αποτελούσαν και την πλειοψηφία. Η προ-επεξεργασία του πίνακα έκφρασης καθώς κι η επιλογή παραμέτρων, μπορούν να επηρεάσουν σημαντικά τα αποτελέσματα. Ενδεχομένως, η παρουσία περισσότερων κυττάρων ανά δότη κι από περισσότερους δότες, με ΔΜΣ στο φυσιολογικό εύρος για τα άτομα χωρίς ΣΔΤ2, να επέτρεπε να αποκαλυφθεί η παρουσία υπο-πληθυσμών που θα κατάφερναν να σχηματίσουν καταστάσεις και τροχιές μεταξύ τους, και με αποτελέσματα που θα θεωρούνταν λιγότερο επισφαλή.

ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	**Ελληνικός Όρος**
Agglomerative hierarchical	συσσωρευτική ιεραρχική
A-posteriori	εκ των υστέρων
A-priori	εκ των προτέρων
Bandwidth	εύρος ζώνης
Barplot	ραβδόγραμμα
Bayesian Information Criterion	κριτήριο πληροφορίας Bayes
Betweeness centrality	κεντρικότητα ενδιαμεσότητας
Bimodal	διτροπική
Boxplot	θηκόγραμμα
Bulk	ομαδοποιημένα
Cell state	κυτταρική κατάσταση
Components	συνιστώσες
Connectivity	συνεκτικότητα
Covariate	συνδιακυμαίνουσα
Cross entropy	διεντροπία
Differential expression	διαφορική έκφραση
Diffusion distance	απόσταση διάχυσης
Diffusion map	χάρτης διάχυσης
Digital object identifier	αναγνωριστικό ψηφιακού αντικειμένου
Dimensionality reduction	μείωση της διαστατικότητας
Doublet	ζεύγος
Eigenvalue	ιδιοτιμή
Eigenvector	ιδιοδιάνυσμα
Embedding	εμβύθιση
Epigenetic landsape	επιγενετικό τοπίο
Excitatory	διεγερτική
Expectation-Maximization algorithm	αλγόριθμος μεγιστοποίησης της προσδοκίας
Extra-randomized trees	υπερ-τυχαιοποιημένα δέντρα
Forking	δημιουργία κλώνου
Gaussian mixture model	πρότυπο μείξης κανονικών κατανομών
Gene regulatory network	γονιδιακό ρυθμιστικό δίκτυο
Gene-set enrichment score	βαθμολογία εμπλουτισμού του γονιδιακού συνόλου
Genome-wide association study	μελέτη συσχέτισης σε επίπεδο γονιδιώματος
Getters	μέθοδοι / συναρτήσεις ανάκτησης
Ground micro-state	μικρο-κατάσταση έναρξης
Heatmap	χάρτης θερμότητας
Heuristic	ευριστικός
Histogram	ιστόγραμμα
Homogeneous	ομοιογενής
Hurdle model	πρότυπο εμποδίου
Hyperparameter	υπερπαράμετρος
Inference	συμπερασμός
Infinite	άπειρος
Inhibitory	ανασταλτική
Input	είσοδος

Kernel	πυρήνας
Key-genes	κύρια γονίδια, γονίδια-κλειδιά, σημαντικά γονίδια
Knee-point	σημείο γονάτου, σημείο καμπής
Latent Dirichlet Allocation	λανθάνουσα κατανομή Dirichlet
Locally-Linear Embedding	τοπικά-γραμμική εμβύθιση
Landing micro-state	μικρο-κατάσταση προορισμού
Loading	φόρτωση
Log-normal	λογο-κανονική
Manifold	πολλαπλότητα
Marker	δείκτης
Maturity-onset diabetes of the youth	όψιμης έναρξης σακχαρώδης διαβήτης των νέων
Metabolome	μεταβόλωμα
Micro-state	μικρο-κατάσταση
Minimum spanning tree	ελάχιστο επικαλύπτον δένδρο
Misexpressed	άτυπα εκφρασμένα
Model	πρότυπο
Nesting	εμφώλευση
Normalization	κανονικοποίηση
Partial decomposition	μερική αποσύνθεση
Perplexity	βαθμός σύγχυσης
Pre-processing	προ-επεξεργασία
Principal Components Analysis	ανάλυση κύριων συνιστωσών
Principal curves	κύριες καμπύλες
Probabilistic	πιθανοτικό
Profiler	αναλυτής κατανομής / απόδοσης
Projection	προβολή
Proteome	πρωτεΐνωμα
Quantitative polymerase chain reaction	ποσοτική αλυσιδωτή αντίδραση πολυμεράσης
Pluripotent	ολοδύναμα
Progenitor	προγονικό
Propensity	τάση
Pseudo-count	ψευδοανάγνωσμα
Pseudotime	ψευδοχρόνος
Random forest	τυχαίο δάσος
Random walks	τυχαίοι περίπατοι
Regression tree	δέντρο παλινδρόμησης
Ribonucleic acid	ριβονουκλεϊκό οξύ
Running median	κινούμενη διάμεση τιμή
Scaling factor	παράγοντας κλιμάκωσης
Seeds	φύτρα
Sequencing	αλληλούχηση
Setters	μέθοδοι / συναρτήσεις ανάθεσης
Signature genes	γονιδιακή υπογραφή
Single-cells	μονήρη κύτταρα
Smoothing	εξομάλυνση
Snapshot	στιγμιότυπο
Sparse matrix	αραιός πίνακας
Spline function	σφηνοειδής συνάρτηση

Split	διαχωρισμός
Standardization	τυποποίηση
Supervised	εποπτευόμενη
t-distributed Stochastic Neighbor Embedding	στοχαστική εμβύθιση γειτόνων βάσει της κατανομής t
Trajectory	τροχιά
Transcriptome	μεταγράφωμα
Transformation	μετασχηματισμός
Transition	μετάβαση
Transitional micro-state	μικρο-κατάσταση μετάβασης
Two-part model	πρότυπο δύο μερών
Uniform Manifold Approximation and Projection	προσέγγιση και προβολή ομοιόμορφης πολλαπλότητας
Upregulation	προς τα πάνω ρύθμιση
User time	χρόνος χρήστη
Violin plot	διάγραμμα βιολιού
Workflow	ροή επεξεργασίας
World Health Organization	Παγκόσμιος Οργανισμός Υγείας
Zero-Inflated Factor Analysis	ανάλυση παραγόντων παρουσία πολλών μηδενικών
Zero-Inflated Negative Binomial-based Wanted Variation Extraction	εξαγωγή της επιθυμητής διακύμανσης βασισμένη σε αρνητική διωνυμική κατανομή παρουσία πολλών μηδενικών
Z-score	τυπική τιμή

ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

BIC	Bayesian Information Criterion
DOI	Digital Object Identifier
EM algorithm	Expectation-Maximization algorithm
GB	Giga bytes
GMM	Gaussian Mixture Model
GRN	Gene Regulatory Network
GWAS	Genome-wide association study
Inf	Infinite
LLE	Locally-Linear Embedding
m-state	micro-state
PC	Principal Component(s)
PCA	Principal Component Analysis
qPCR	quantitative polymerase chain reaction
RNA	ριβονουκλεϊκό οξύ
RNA-Seq	αλληλούχηση του ριβονουκλεϊκού οξέος
t-SNE	t-distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
WHO	World Health Organization
ZIFA	Zero-Inflated Factor Analysis
ZINB-WaVE	Zero-Inflated Negative Binomial-based Wanted Variation Extraction
ΑΚΣ	Ανάλυση Κύριων Συνιστωσών
Αλγόριθμος ΕΜ	αλγόριθμος μεγιστοποίησης της προσδοκίας
ΔΜΣ	Δείκτης μάζας σώματος
σδ	σύνολο δεδομένων
ΣΔΤ2	σακχαρώδης διαβήτης τύπου 2

ΑΝΑΦΟΡΕΣ

[1] R. Sender, S. Fuchs and R. Milo, "Revised Estimates for the Number of Human and Bacteria Cells in the Body", PLOS Biology, vol. 14, no. 8, p. e1002533, 2016. Available: 10.1371/journal.pbio.1002533 [Accessed 17 July 2019].

[2] S. Hicks, F. Townes, M. Teng and R. Irizarry, "Missing data and technical variability in single-cell RNA-sequencing experiments", Biostatistics, vol. 19, no. 4, pp. 562-578, 2017. Available: 10.1093/biostatistics/kxx053 [Accessed 17 July 2019].

[3] A. Goldberg, C. Allis and E. Bernstein, "Epigenetics: A Landscape Takes Shape", Cell, vol. 128, no. 4, pp. 635-638, 2007. Available: 10.1016/j.cell.2007.02.006 [Accessed 17 July 2019].

[4] Q. Nguyen et al., "Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity", Nature Communications, vol. 9, no. 1, 2018. Available: 10.1038/s41467-018-04334-1 [Accessed 17 July 2019].

[5] C. Zheng et al., "Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing", Cell, vol. 169, no. 7, pp. 1342-1356.e16, 2017. Available: 10.1016/j.cell.2017.05.035 [Accessed 17 July 2019].

[6] R. Kishton, M. Sukumar and N. Restifo, "Metabolic Regulation of T Cell Longevity and Function in Tumor Immunotherapy", Cell Metabolism, vol. 26, no. 1, pp. 94-109, 2017. Available: 10.1016/j.cmet.2017.06.016 [Accessed 17 July 2019].

[7] Y. Su et al., "Single-cell analysis resolves the cell state transition and signaling dynamics associated with melanoma drug-induced resistance", Proceedings of the National Academy of Sciences, vol. 114, no. 52, pp. 13679-13684, 2017. Available: 10.1073/pnas.1712064115 [Accessed 17 July 2019].

[8] W. Saelens, R. Cannoodt, H. Todorov and Y. Saeys, "A comparison of single-cell trajectory inference methods", Nature Biotechnology, vol. 37, no. 5, pp. 547-554, 2019. Available: 10.1038/s41587-019-0071-9 [Accessed 17 July 2019].

[9] D. duVerle, S. Yotsukura, S. Nomura, H. Aburatani and K. Tsuda, "CellTree: an R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data", BMC Bioinformatics, vol. 17, no. 1, 2016. Available: 10.1186/s12859-016-1175-6 [Accessed 13 December 2019].

[10] J.K. Pritchard, M. Stephens and P. Donnelly, "Inference of Population Structure Using Multilocus Genotype Data", Genetics, vol. 155, pp. 945–959, 2000.

[11] C. Trapnell et al., "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells", Nature Biotechnology, vol. 32, no. 4, pp. 381-386, 2014. Available: 10.1038/nbt.2859 [Accessed 14 December 2019].

[12] J.B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1. University of California Press, pp. 281–297, 1967.

[13] J. Herman, Sagar and D. Grün, "FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data", Nature Methods, vol. 15, no. 5, pp. 379-386, 2018. Available: 10.1038/nmeth.4662 [Accessed 13 December 2019].

[14] L. Kaufman and P.J. Rousseeuw, "Clustering by means of Medoids, in Statistical Data Analysis Based on the L1–Norm and Related Methods", apers prepared at the First International Conference on Statistical Data Analysis Based on the L1-norm and Related Methods, held in Neuchâtel, Switzerland, from August 31-September 4, 1987, pp. 405–416, 1987.

[15] M. Guo, E. Bao, M. Wagner, J. Whitsett and Y. Xu, "SLICE: determining cell differentiation and lineage based on single cell entropy", Nucleic Acids Research, p. gkw1278, 2016. Available: 10.1093/nar/gkw1278 [Accessed 13 December 2019].

[16] K. Street et al., "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics", BMC Genomics, vol. 19, no. 1, 2018. Available: 10.1186/s12864-018-4772-0 [Accessed 13 December 2019].

[17] T. Hastie and W. Stuetzle, "Principal curves", J Am Stat Assoc, vol. 84, no. 406, pp. 502-516, 1989.

[18] P. Tsakanikas, D. Manatakis and E. Manolakos, "Machine learning methods to reverse engineer dynamic gene regulatory networks governing cell state transitions", 2018. Available: 10.1101/264671 [Accessed 17 July 2019].

[19] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (PCA)", Computers & Geosciences, vol. 19, no. 3, pp. 303-342, 1993. Available: 10.1016/0098-3004(93)90090-r [Accessed 17 July 2019].

[20] E. Pierson and C. Yau, "ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis", Genome Biology, vol. 16, no. 1, 2015. Available: 10.1186/s13059-015-0805-z [Accessed 17 July 2019].

[21] D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit and J. Vert, "A general and flexible method for signal extraction from single-cell RNA-seq data", Nature Communications, vol. 9, no. 1, 2018. Available: 10.1038/s41467-017-02554-5.

[22] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE", Journal of Machine Learning Research, vol. 9, pp. 2579–2605, 2008.
[23] “Student” [William Sealy Gosset], "THE PROBABLE ERROR OF A MEAN", Biometrika, vol. 6, no. 1, pp. 1-25, 1908. Available: 10.1093/biomet/6.1.1 [Accessed 17 July 2019].
[24] L. McInnes, J. Healy, N. Saul and L. Großberger, "UMAP: Uniform Manifold Approximation and Projection", Journal of Open Source Software, vol. 3, no. 29, p. 861, 2018. Available: 10.21105/joss.00861.
[25] E. Becht et al., "Dimensionality reduction for visualizing single-cell data using UMAP", Nature Biotechnology, vol. 37, no. 1, pp. 38-44, 2018. Available: 10.1038/nbt.4314 [Accessed 4 December 2019].
[26] S. Roweis, "Nonlinear Dimensionality Reduction by Locally Linear Embedding", Science, vol. 290, no. 5500, pp. 2323-2326, 2000. Available: 10.1126/science.290.5500.2323 [Accessed 17 July 2019].
[27] R. Coifman and S. Lafon, "Diffusion maps", Applied and Computational Harmonic Analysis, vol. 21, no. 1, pp. 5-30, 2006. Available: 10.1016/j.acha.2006.04.006 [Accessed 4 December 2019].
[28] X. Qiu et al., "Reversed graph embedding resolves complex single-cell trajectories", Nature Methods, vol. 14, no. 10, pp. 979-982, 2017. Available: 10.1038/nmeth.4402 [Accessed 4 December 2019].
[29] P. Angerer, L. Haghverdi, M. Büttner, F. Theis, C. Marr and F. Buettner, "destiny: diffusion maps for large-scale single-cell data in R", Bioinformatics, vol. 32, no. 8, pp. 1241-1243, 2015. Available: 10.1093/bioinformatics/btv715 [Accessed 17 July 2019].
[30] R. Chellappa et al., "Gaussian Mixture Models", Encyclopedia of Biometrics, pp. 659-663, 2009. Available: 10.1007/978-0-387-73003-5_196 [Accessed 17 July 2019].
[31] L. Scrucca, M. Fop, T. Murphy and A. Raftery, "mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models", The R Journal, vol. 8, no. 1, p. 289, 2016. Available: 10.32614/rj-2016-021 [Accessed 17 July 2019].
[32] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", Journal of the Royal Statistical Society. Series B (Methodological), vol. 39, no. 1, pp. 1-38, 1977. Available: https://www.jstor.org/stable/2984875 [Accessed 17 July 2019].
[33] J. D. Banfield and A. E. Raftery, "Model-based Gaussian and non-Gaussian clustering", Biometrics, vol. 49, pp. 803-821, 1993.
[34] H. Kaiser, "The Application of Electronic Computers to Factor Analysis", Educational and Psychological Measurement, vol. 20, no. 1, pp. 141-151, 1960. Available: 10.1177/001316446002000116 [Accessed 17 July 2019].
[35] D. Kaplan, "Knee Point - File Exchange - MATLAB Central", Mathworks.com, 2019. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/35094-knee-point. [Accessed: 17- Jul- 2019].
[36] G. Schwarz, "Estimating the Dimension of a Model", The Annals of Statistics, vol. 6, no. 2, pp. 461-464, 1978. Available: 10.1214/aos/1176344136 [Accessed 17 July 2019].
[37] J. Ferguson, "Multivariable Curve Interpolation", Journal of the ACM, vol. 11, no. 2, pp. 221-228, 1964. Available: 10.1145/321217.321225 [Accessed 17 July 2019].
[38] R. Dougherty, A. Edelman and J. Hyman, "Nonnegativity-, monotonicity-, or convexity-preserving cubic and quintic Hermite interpolation", Mathematics of Computation, vol. 52, no. 186, pp. 471-471, 1989. Available: 10.1090/s0025-5718-1989-0962209-1 [Accessed 17 July 2019].
[39] A. McDavid et al., "Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments", Bioinformatics, vol. 29, no. 4, pp. 461-467, 2012. Available: 10.1093/bioinformatics/bts714.
[40] T. Stuart et al., "Comprehensive Integration of Single-Cell Data", Cell, vol. 177, no. 7, pp. 1888-1902.e21, 2019. Available: 10.1016/j.cell.2019.05.031 [Accessed 19 September 2019].
[41] G. Finak et al., "MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data", Genome Biology, vol. 16, no. 1, 2015. Available: 10.1186/s13059-015-0844-5 [Accessed 19 September 2019].
[42] K. Campbell, switchde: Switch-like differential expression across single-cell trajectories. 2019 [Accessed 19 September 2019].
[43] M. Robinson, D. McCarthy and G. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data", Bioinformatics, vol. 26, no. 1, pp. 139-140, 2009. Available: 10.1093/bioinformatics/btp616 [Accessed 17 July 2019].
[44] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing", Journal of the Royal Statistical Society: Series B (Methodological), vol. 57, no. 1, pp. 289-300, 1995. Available: 10.1111/j.2517-6161.1995.tb02031.x [Accessed 17 July 2019].
[45] V. Huynh-Thu, A. Irrthum, L. Wehenkel and P. Geurts, "Inferring Regulatory Networks from Expression Data Using Tree-Based Methods", PLoS ONE, vol. 5, no. 9, p. e12776, 2010. Available: 10.1371/journal.pone.0012776 [Accessed 17 July 2019].

[46] D. Marbach et al., "Wisdom of crowds for robust gene network inference", Nature Methods, vol. 9, no. 8, pp. 796-804, 2012. Available: 10.1038/nmeth.2016 [Accessed 17 July 2019].
[47] P. Bellot, C. Olsen, P. Salembier, A. Oliveras-Vergés and P. Meyer, "NetBenchmark: a bioconductor package for reproducible benchmarks of gene regulatory network inference", BMC Bioinformatics, vol. 16, no. 1, 2015. Available: 10.1186/s12859-015-0728-4 [Accessed 17 July 2019].
[48] M. Morgan and H. Pagès, "S4 classes and methods", Bioconductor.org, 2019. [Online]. Available: https://www.bioconductor.org/help/course-materials/2017/Zurich/S4-classes-and-methods.html#strengths-of-s4-compared-to-s3. [Accessed: 17- Jul- 2019].
[49] A. Pratapa, A. Jalihal, J. Law, A. Bharadwaj and T. Murali, "Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data", 2019. Available: 10.1101/642926 [Accessed 17 July 2019].
[50] H. Wickham, "testthat: Get Started with Testing", The R Journal, vol. 3, no. 1, p. 5, 2011. Available: 10.32614/rj-2011-002.
[51] M. Morgan, V. Obenchain, M. Lang and R. Thompson, "BiocParallel", Bioconductor, 2019. [Online]. Available: http://bioconductor.statistik.tu-dortmund.de/packages/3.5/bioc/html/BiocParallel.html. [Accessed: 17- Jul- 2019].
[52] W. Chang, J. Luraschi and T. Mastny, "Interactive Visualizations for Profiling R Code [R package profvis version 0.3.6]", Cran.r-project.org, 2019. [Online]. Available: https://cran.r-project.org/web/packages/profvis/index.html. [Accessed: 17- Jul- 2019].
[53] Y. Wang et al., "Single-Cell Transcriptomics of the Human Endocrine Pancreas", Diabetes, vol. 65, no. 10, pp. 3028-3038, 2016. Available: 10.2337/db16-0405 [Accessed 1 December 2019].
[54] A. Stevens and J. Lowe, Human Histology. Philadelphia, Pa: Elsevier Mosby, 2013.
[55] R. Dassaye, S. Naidoo and M. Cerf, "Transcription factor regulation of pancreatic organogenesis, differentiation and maturation", Islets, vol. 8, no. 1, pp. 13-34, 2015. Available: 10.1080/19382014.2015.1075687 [Accessed 1 December 2019].
[56] "2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2019", Diabetes Care, vol. 42, no. 1, pp. S13-S28, 2018. Available: 10.2337/dc19-s002 [Accessed 1 December 2019].
[57] "GEO Accession viewer", Ncbi.nlm.nih.gov, 2019. [Online]. Available: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE83139. [Accessed: 01- Dec- 2019].
[58] "Body mass index - BMI", Euro.who.int, 2019. [Online]. Available: http://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi. [Accessed: 01- Dec- 2019].
[59] K. Chawla, S. Tripathi, L. Thommesen, A. Lægreid and M. Kuiper, "TFcheckpoint: a curated compendium of specific DNA-binding RNA polymerase II transcription factors", Bioinformatics, vol. 29, no. 19, pp. 2519-2520, 2013. Available: 10.1093/bioinformatics/btt432 [Accessed 1 December 2019].
[60] A. Rouillard et al., "The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins", Database, vol. 2016, p. baw100, 2016. Available: 10.1093/database/baw100 [Accessed 1 December 2019].
[61] B. Giotti et al., "Assembly of a parts list of the human mitotic cell cycle machinery", Journal of Molecular Cell Biology, vol. 11, no. 8, pp. 703-718, 2018. Available: 10.1093/jmcb/mjy063.
[62] G. Stelzer et al., "The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses", Current Protocols in Bioinformatics, vol. 54, no. 1, 2016. Available: 10.1002/cpbi.5 [Accessed 1 December 2019].
[63] U. Raudvere et al., "g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)", Nucleic Acids Research, vol. 47, no. 1, pp. W191-W198, 2019. Available: 10.1093/nar/gkz369 [Accessed 1 December 2019].
[64] G. Zhou, O. Soufan, J. Ewald, R. Hancock, N. Basu and J. Xia, "NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis", Nucleic Acids Research, vol. 47, no. 1, pp. W234-W241, 2019. Available: 10.1093/nar/gkz240.
[65] K. Aamodt and A. Powers, "Signals in the pancreatic islet microenvironment influence β-cell proliferation", Diabetes, Obesity and Metabolism, vol. 19, pp. 124-136, 2017. Available: 10.1111/dom.13031 [Accessed 1 December 2019].
[66] Y. Xin et al., "RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes", Cell Metabolism, vol. 24, no. 4, pp. 608-615, 2016. Available: 10.1016/j.cmet.2016.08.018 [Accessed 1 December 2019]
[67] M. Kane, J. Emerson and S. Weston, "Scalable Strategies for Computing with Massive Data", Journal of Statistical Software, vol. 55, no. 14, 2013. Available: 10.18637/jss.v055.i14 [Accessed 5 December 2019].