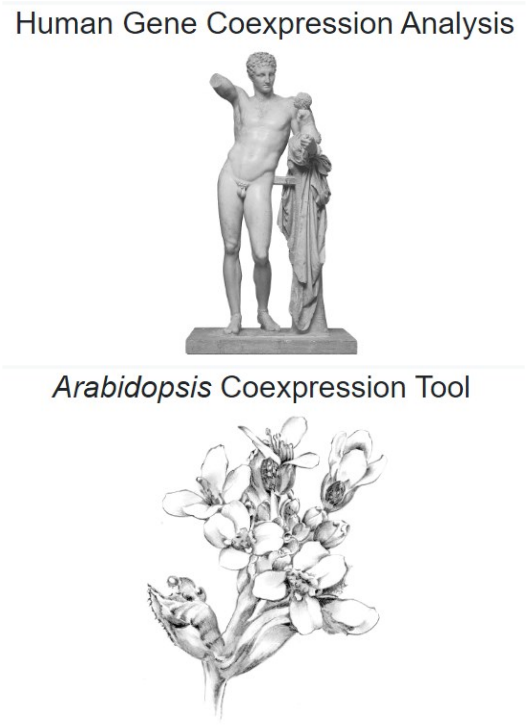
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικό και Καποδιστριακό
Πανεπιστήμιο Αθηνών

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
«ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ»

Δ Ι Π Λ Ω Μ Α Τ Ι Κ Η Ε Ρ Γ Α Σ Ι Α

HGCA και ACT: Εργαλεία για τη μελέτη της γονιδιακής συνέκφρασης στον άνθρωπο και το πρότυπο φυτό *Arabidopsis thaliana*



Βασίλειος Ζωγόπουλος

Πτυχιούχος Πληφοροφικής με Εφαρμογές στη Βιοϊατρική,
Πανεπιστήμιο Θεσσαλίας

ΑΘΗΝΑ 2020

Εξεταστική Επιτροπή:

1) Αναπληρώτρια Καθηγήτρια κα. Βασιλική Οικονομίδου

Τμήμα Βιολογίας, Πανεπιστήμιο Αθηνών

2) Καθηγητής κ. Παντελής Μπάγκος

Τμήμα Πληροφορικής με Εφαρμογές στην Βιοϊατρική, Πανεπιστήμιο Θεσσαλίας

3) Επίκουρος Καθηγητής κ. Βασίλειος Κουβέλης

Τμήμα Βιολογίας, Πανεπιστήμιο Αθηνών

Ευχαριστίες

Η παρούσα εργασία πραγματοποιήθηκε στο Ίδρυμα Ιατροβιολογικών Ερευνών της Ακαδημίας Αθηνών (IIBEAA) κάτω από την συνεχή επίβλεψη του Δρ. Ιωάννη Μιχαλόπουλου, χωρίς το όραμα και την προσπάθεια του οποίου η εργασία δεν θα μπορούσε ποτέ να ολοκληρωθεί. Ευχαριστώ το Elixir-GR (MIS: 5002780) για την προσφορά της μεταπτυχιακής υποτροφίας για τη δημιουργία των εργαλείων. Θα ήθελα να ευχαριστήσω θερμά την Αναπληρώτρια Καθηγήτρια κα Β. Οικονομίδου ως επιβλέπουσα καθηγήτρια της πτυχιακής εργασίας, η οποία έδωσε την άδεια για την συνεργασία με τον Δρ. Μιχαλόπουλο. Επιπλέον θα ήθελα να ευχαριστήσω τον Καθηγητή κ. Παντελή Μπάγκο και τον Επίκουρο Καθηγητή κ. Βασίλειο Κουβέλη που μου έκαναν την τιμή να είναι τα άλλα δύο μέλη της τριμελούς επιτροπής εξέτασης της πτυχιακής εργασίας. Για το εργαλείο ACT, ευχαριστώ ιδιαίτερα την Δρ. Γεωργία Σαξάμη για όλο το έργο της πάνω στο ACT, τον Δρ. Απόστολο Μαλατρά, για όλη του την συνεισφορά στην χρήση και ανάπτυξη αλγορίθμων που χρησιμοποιήθηκαν καθώς και τον Καθηγητή κ. Πολυδεύκη Χατζόπουλο και Δρ. Γεράσιμο Δάρρα για τις δοκιμές τους στο ACT και την συγγραφή ιστοριών επιτυχίας. Επίσης, θα ήθελα να ευχαριστήσω την υπόλοιπη συγγραφική ομάδα του ACT (Καθηγητής D.R. Westhead, Καθηγητής W.J. Duddy, Dr. J.H. Jen και Αντωνία Αγγελοπούλου) για τις συμβουλές τους στη βελτίωση του εργαλείου. Για το εργαλείο HGCA2, θα ήθελα να ευχαριστήσω τον Χρήστο Βασιλείου, για όλη του την έρευνα πάνω στο GTEx και τον Κωνσταντίνο Κυριακίδη για την επιλογή και εκτέλεση του αλγορίθμου επεξεργασίας των δεδομένων. Για το εργαλείο HGCA1.5 θα ήθελα να ευχαριστήσω πάλι τον Δρ. Μαλατρά για την λήψη και κανονικοποίηση των δειγμάτων.

Ευχαριστώ από καρδιάς όλους τους συνεργάτες του εργαστηρίου στο IIBEAA, καθώς και τον ερευνητή Δρ. Μάριο Αγγελόπουλο του IIBEAA για την στήριξη και τις επιστημονικές τους συμβουλές. Τέλος ευχαριστώ όλους τους φίλους και συμφοιτητές που με στήριξαν σε όλη την διάρκεια εκπόνησης της πτυχιακής και φυσικά τους γονείς μου οι οποίοι πίστεψαν σε εμένα και συνεχίζουν να με υποστηρίζουν σε κάθε βήμα.

Περίληψη

Γονίδια με παρόμοια πρότυπα έκφρασης τείνουν να συμμετέχουν σε σχετικές βιολογικές διεργασίες. Ο πιο αποτελεσματικός τρόπος για την μελέτη της γονιδιακής συνέκφρασης βασίζεται στην ανάλυση μεταγραφωμικών δεδομένων του συνόλου των πιο αντιπροσωπευτικών δειγμάτων από κάθε ιστό ή είδος κυττάρου. Η συνέκφραση γονιδίων που αποκαλύπτεται από μία ποικιλία πειραμάτων μεταγραφωμικής, που υπάρχουν διαθέσιμα σε δημόσια καταθετήρια, μπορεί να περιέχει πληροφορίες που υπερβαίνουν τον αρχικό σκοπό του κάθε πειράματος και αυτό μπορεί να αποτελέσει ένα πολύτιμο εργαλείο πρόβλεψης για τη λειτουργία γονιδίων και τη συμμετοχή τους σε βιολογικά μονοπάτια. Δημιουργήθηκαν 3 εργαλεία συνέκφρασης: το *Arabidopsis* Coexpression Tool (ACT) που μελετάει τη γονιδιακή συνέκφραση στο πρότυπο φυτό *Arabidopsis thaliana* και είναι βασισμένο σε 3500 δείγματα μικροσυστοιχιών από 3 διαφορετικές βάσεις δεδομένων, το Human Gene Coexpression Analysis (HGCA) 1.5 που μελετάει τη γονιδιακή συνέκφραση στον άνθρωπο και είναι βασισμένο σε 1959 δείγματα μικροσυστοιχιών από την GEO και το HGCA2, πάλι για τον άνθρωπο, που είναι βασισμένο σε 3500 δείγματα RNA-seq από την GTEx. Η επιλογή των δειγμάτων σε κάθε περίπτωση έγινε με λεπτομερή τρόπο, χρησιμοποιήθηκαν καινοτόμοι αλγόριθμοι επεξεργασίας των δεδομένων και η ομαδοποίηση των γονιδίων έγινε με ιεραρχική ομαδοποίηση. Η ανάπτυξη των ιστοτόπων έγινε χρησιμοποιώντας μοντέρνες τεχνολογίες. Εισάγοντας ένα γονίδιο-οδηγό, τα εργαλεία παρουσιάζουν ένα φυλογενετικό υποδέντρο του οποίου τα φύλλα αποτελούν γονίδια συνεκφρασμένα με το γονίδιο εισόδου. Ο χρήστης μπορεί επιπλέον να πραγματοποιήσει ποικίλες αναλύσεις εμπλουτισμού βιολογικών όρων πάνω στα γονίδια του υποδέντρου. Μελετώντας τη λίστα συνεκφρασμένων γονιδίων και τα αποτελέσματα υπερεκπροσώπησης ανακαλύπτονται λειτουργικοί συνεργάτες ή προσδίδονται ρόλοι σε μη χαρακτηρισμένα γονίδια, κοιτώντας τους γείτονές τους στο υποδέντρο συνέκφρασης.

Abstract

Genes with similar expression patterns tend to participate in related biological processes. The most efficient way to study gene coexpression is based on the transcriptomic data analysis of the subset of samples which contain the best representatives of each tissue or cell type. The coexpression of genes revealed from a variety of transcriptomic experiments stored in public repositories, may contain information far beyond the original scope of each constituent experiment, and can be a valuable predictive tool for gene function and pathway membership. Three tools were developed: *Arabidopsis* Coexpression Tool (ACT) studies gene coexpression in model plant *Arabidopsis thaliana*, which is based on 3500 microarray samples from three different public repositories, Human Gene Coexpression Analysis (HGCA) 1.5 studies gene coexpression in human, which is based on 1959 microarray samples from GEO and HGCA2, again for human, based on 3500 RNA-seq samples from GTEx. Meticulous sample selection was performed in each case, novel data processing algorithms were used and genes were grouped using hierarchical clustering. The websites were developed using modern libraries. By typing a driver-gene, the tools output a coexpression subtree whose leaves contain genes coexpressed with the input gene. The user can perform a variety of biological term enrichment analyses upon the list of gene of the coexpression subtree. By studying the list of coexpressed genes and the enrichment analysis results, gene functional partners to the gene of interest can be discovered or a function can be assigned to a gene of unknown role by examining its coexpressed genes in neighbouring leaves.

Table of Contents

Εισαγωγή 10
Μικροσυστοιχίες DNA 10
Περιγραφή πειραματικής διαδικασίας μικροσυστοιχιών DNA 11
Ολιγονουκλεοτιδικές μικροσυστοιχίες (Affymetrix GeneChip) 14
Αρχεία Affymetrix GeneChip 16
Αναλύσεις διαφορικής έκφρασης και γονιδιακής συνέκφρασης 17
Προεπεξεργασία μικροσυστοιχιών Affymetrix GeneChip 18
Διόρθωση υποβάθρου 18
Κανονικοποίηση 18
Διόρθωση PM 19
Σύνοψη ανιχνευτών 19
Μέθοδοι προεπεξεργασίας υλοποιημένες για μικροσυστοιχίες Affymetrix GeneChip 20
Microarray Suite 5.0 (MAS5.0) 20
Probe Logarithmic Intensity Error Estimation (PLIER) 20
Robust Multi-Array Analysis (RMA) 21
GeneChip RMA (GC-RMA) 21
Frozen Robust Multi-Array Analysis (fRMA) 21
Single Channel Array Normalisation (SCAN) 22
Minimum Information About a Microarray Experiment (MIAME) 22
Η βάση δεδομένων Gene Expression Omnibus (GEO) 23
Πλατφόρμες (Platforms) 24
Δείγματα (Samples) 24
Σειρές (Series) 24
Σύνολα Δεδομένων (Datasets) 25
Η βάση δεδομένων GTEx 26
Εισαγωγή ..... **Error! Bookmark not defined.**
Παραγωγή και διάθεση δεδομένων 27
Δεδομένα έκφρασης 27
Γονοτυπικά δεδομένα 27
Μέθοδοι ανάλυσης 28
Προεπεξεργασία 29
Στοίχιση RNA-seq 29
Γονοτυπική ανάλυση 29
Ποσοτικοποίηση Έκφρασης 30
Μεταγραφικό υπόδειγμα 30

Συμπτυγμένο γονιδιακό υπόδειγμα 30
Ποσοτικοποίηση 30
Φυλογενετικά δέντρα 32
Φυλογενετική ανάλυση 33
Μέθοδοι κατασκευής δέντρων 34
Μέθοδοι βασισμένες στην απόσταση 35
Μέθοδοι βασισμένες στους χαρακτήρες 37
Μορφοποίηση αρχείων δέντρων Newick 38
Μέθοδοι 40
Τεχνικές λεπτομέρειες υπολογιστή προγραμματιστικής ανάπτυξης 40
Apache Server 40
Σχεσιακό Μοντέλο Βάσεων Δεδομένων 40
MySQL Server 43
Δημιουργία των Βάσεων Δεδομένων 44
PHP 46
R & R Studio 46
Oracle Virtual Machine 47
Arabidopsis Coexpression Tool (ACT) 47
Συλλογή πρωταρχικών δεδομένων – Εισαγωγή 47
Συλλογή πρωταρχικών δεδομένων – Εκτέλεση 48
Ποιοτικός έλεγχος πρωτογενών δεδομένων και απόρριψη δεδομένων χαμηλής ποιότητας 50
Κανονικοποίηση των πρωτογενών δεδομένων με βέλτιστους αλγορίθμους και περιγραφές ανιχνευτών για την παραγωγή δευτερογενών δεδομένων 51
Αναζήτηση και ανάκτηση από διάφορες βάσεις δεδομένων όρων που περιγράφουν τα γονίδια του φυτού και επεξεργασία τους 52
Κατασκευή βάσης δεδομένων και αποθήκευση των δευτερογενών δεδομένων σε αυτή 52
Λεπτομερής ανάλυση τρόπου ανάκτησης δεδομένων από τις αντίστοιχες βάσεις 59
Δημιουργία του δέντρου συσχέτισης δειγμάτων και αυτόματη επιλογή των αντιπροσωπευτικών δειγμάτων 63
Δημιουργία του δέντρου γονιδιακής συσχέτισης γονιδίων 66
Συμφαινετική Συσχέτιση (Cophenetic Correlation) 67
Δημιουργία της διαδικτυακής διεπαφής χρήστη 68
Υλοποίηση ανάλυσης υπερεκπροσώπησης όρων 68
Human Gene Coexpression Analysis (HGCA) 69
Συλλογή πρωταρχικών δεδομένων 69
Επεξεργασία των δεδομένων 70

Αναζήτηση και ανάκτηση από διάφορες βάσεις δεδομένων όρων που περιγράφουν τα ανθρώπινα γονίδια και επεξεργασία τους 71
Κατασκευή βάσης δεδομένων και αποθήκευση των δευτερογενών δεδομένων σε αυτή 71
Λεπτομερής ανάλυση τρόπου ανάκτησης δεδομένων από τις αντίστοιχες βάσεις 76
Δημιουργία του δέντρου συσχέτισης δειγμάτων και αυτόματη επιλογή των αντιπροσωπευτικών δειγμάτων 80
Δημιουργία του δέντρου γονιδιακής συσχέτισης γονιδίων 82
Microarray Human Gene Coexpression Analysis (HGCA) 83
Συλλογή πρωταρχικών δεδομένων και κανονικοποίησή τους 83
Δημιουργία βάσης δεδομένων 84
Δημιουργία του δέντρου γονιδιακής συσχέτισης γονιδίων 84
Αποτελέσματα 86
Περιγραφή των ιστοτόπων 86
Αρχικές σελίδες 86
Αναζήτηση γονιδίου 88
Επιλογή γονιδίου και εκτέλεση ανάλυσης 90
Ανάλυση υπερεκπροσώπησης όρων 93
Παραδείγματα αποτελεσμάτων 95
ACT 95
Ριβοσωμικές Πρωτεΐνες 95
Πρωτεΐνες Θερμικού Σοκ (Heat Shock Proteins) 98
Αντίδραση στο κρύο (Response to cold) 99
Βιογένεση Κυτταρικού Τοιχώματος (Cell Wall Biogenesis) 100
Φωτοσύνθεση 106
Κιρκαδικός Ρυθμός 107
Χλωροπλαστικές και Μιτοχονδριακές Πρωτεΐνες 107
Ανθήρες και Γύρη 109
Ανάπτυξη Εμβρύου 112
HGCA 115
Πρωτεΐνες HLA 115
Μεταλλοθειονίνες 119
Γενική Ανοσολογική Απόκριση 120
Απόκριση σε Ιούς 122
Κυτοκίνες 124
Όσφρηση 126
Υποδοχείς COVID-19 127

HGCA1.5 130
Κυτοκίνες 130
Συζήτηση 132
ACT 132
Σύγκριση δυνατοτήτων ανταγωνιστικών εργαλείων με το ACT 133
Σύγκριση αποτελεσμάτων ανταγωνιστικών εργαλείων με το ACT 137
HGCA 143
Παρόμοια εργαλεία με το HGCA 146
Σύγκριση αποτελεσμάτων ανταγωνιστικών εργαλείων με το HGCA 149
Συμπέρασμα 154
Βιβλιογραφία 157

Εισαγωγή

Μικροσυστοιχίες DNA

Η τεχνολογία των μικροσυστοιχιών DNA επινοήθηκε στα μέσα της δεκαετίας του 1990 (Schena et al., 1995). Οι μικροσυστοιχίες DNA είναι μια διάταξη μεγάλου αριθμού μικροσκοπικών κηλίδων DNA πάνω σε μία στερεή επιφάνεια. Κάθε κηλίδα DNA ονομάζεται ανιχνευτής και μπορεί να αντιστοιχεί σε ένα τμήμα γονιδίου ή κάποιο άλλο στοιχείο DNA, το οποίο στη συνέχεια χρησιμοποιείται για να υβριδοποιηθεί με έναν στόχο (π.χ. cDNA, cRNA). Οι ανιχνευτές ακινητοποιούνται με ομοιοπολικούς δεσμούς στη στερεή επιφάνεια η οποία μπορεί να αποτελείται από πλαστικό, γυαλί ή σιλάνιο. Η δομή αυτή, διαστάσεων μικρότερων της ανθρώπινης παλάμης, που δημιουργείται με σύγχρονες τεχνικές νανοτεχνολογίας, αποτελεί τη μικροσυστοιχία. Οι μικροσυστοιχίες επιτρέπουν την ανάλυση της γονιδιακής έκφρασης, της ποικιλότητας, της αλληλουχίας του DNA με μαζική και παράλληλη επεξεργασία. Αυτή η μέθοδος υψηλής απόδοσης έρχεται να συμπληρώσει τις ήδη υπάρχουσες κοινές πειραματικές μεθόδους, καθώς μας δίνει τη δυνατότητα ανάλυσης ολόκληρου του μεταγραφώματος ενός ιστού σε ένα μόνο πείραμα. Μπορεί κανείς να εξάγει χρήσιμες πληροφορίες για τη βιολογική λειτουργία ενός οργανισμού, βρίσκοντας ποια γονίδια επάγονται ή καταστέλλονται σε κάποια φάση του κυτταρικού κύκλου, σε κάποια αναπτυξιακή στιγμή ή σε απόκριση σε ερεθίσματα του περιβάλλοντος, όπως π.χ. η απόκριση σε ορμόνες ή σε υψηλή θερμοκρασία. Ομάδες γονιδίων των οποίων η έκφραση αυξάνεται ή μειώνεται ταυτόχρονα, υπό τις ίδιες συνθήκες, είναι πιθανό να συμμετέχουν σε κοινές βιολογικές διεργασίες και σε κοινά μεταβολικά μονοπάτια ή να είναι στόχοι κοινών μεταγραφικών παραγόντων. Η γονιδιακή έκφραση είναι άμεσα συσχετιζόμενη με τις βιολογικές λειτουργίες και οι μικροσυστοιχίες προσφέρουν τεράστιο μέγεθος πληροφοριών πάνω σε ανθρώπινες ασθένειες, γήρανση, φαρμακευτική δράση, ορμονική δράση, διανοητικές ασθένειες, μεταβολισμό και σε αρκετά ακόμα κλινικά θέματα. Αν και η τεχνολογία των μικροσυστοιχιών προσφέρει μία πιο οικονομική και χρονικά βολική λύση στη μελέτη της γονιδιακής έκφρασης, πρέπει να λάβουμε υπ’ όψη ότι λόγω της μαζικής φύσης του πειράματος, τα αποτελέσματα που προκύπτουν από ένα πείραμα

μικροσυστοιχιών, θα πρέπει πάντα να επαληθεύονται με τις κλασικές δια χειρός πειραματικές μεθόδους, όπως την qPCR.

Η ανάπτυξη της νέας αυτής τεχνολογίας έδωσε νέες, ενδιαφέρουσες πληροφορίες και αύξησε εκθετικά τα διαθέσιμα δεδομένα για την κατανόηση των βιολογικών συστημάτων. Από την αρχική της εφαρμογή ως καινούργια τεχνική για μεγάλης κλίμακας χαρτογράφηση του DNA και την αρχική επιτυχία ως εργαλείο ανάλυσης μεταγράφων, η τεχνολογία των μικροσυστοιχιών εξαπλώθηκε σε πολλές περιοχές, προσαρμόζοντας τη βασική επινόηση και συνδυάζοντάς τη με άλλες τεχνικές.

Περιγραφή πειραματικής διαδικασίας μικροσυστοιχιών DNA

Ένα πείραμα με μικροσυστοιχίες είναι μια πολύπλοκη ακολουθία διεργασιών που πρέπει να περατωθούν με επιτυχία, για να εξασφαλιστεί η αποδεκτή ποιότητα των δεδομένων που θα προκύψουν και των συμπερασμάτων που θα εξαχθούν. Η πειραματική διαδικασία αφορά τα στάδια που πρέπει να ακολουθηθούν κατά τη διεξαγωγή ενός πειράματος μικροσυστοιχιών (Εικόνα 1). Τα πειράματα μικροσυστοιχιών για ανάλυση γονιδιακού προφίλ ελέγχουν ταυτόχρονα την έκφραση χιλιάδων γονιδίων για να μελετήσουν την επίδραση μίας θεραπείας, ασθένειας ή αναπτυξιακού σταδίου στην γονιδιακή έκφραση.

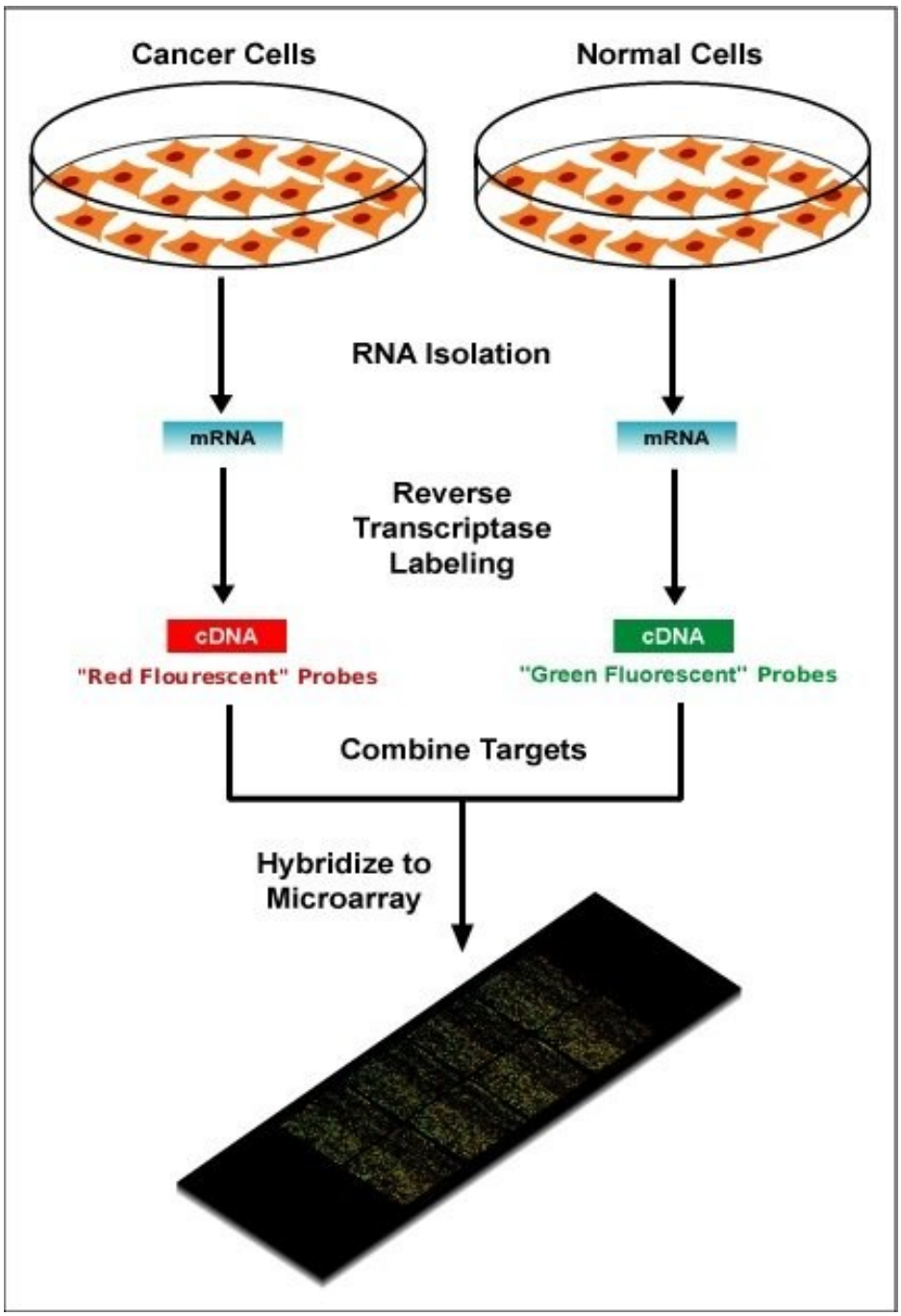
Υπάρχουν πολλά διαφορετικά είδη μικροσυστοιχιών DNA. Διαφέρουν ανάλογα με την κατασκευή των ανιχνευτών (Spotted or *In-Situ* synthesized arrays) αλλά και την ίδια τη δομή της μικροσυστοιχίας (One-channel or Two-channel arrays). Αρχικά, διατυπώνεται ένα βιολογικό ερώτημα το οποίο ελπίζουμε ότι θα απαντηθεί με το τέλος του πειράματος. Στη συνέχεια, ανάλογα με το ερώτημα γίνεται η κατάλληλη επιλογή της μικροσυστοιχίας. Ο τύπος της μικροσυστοιχίας επιλέγεται ως προς τον τρόπο προετοιμασίας του, αλλά και το ποια είναι η διάταξη και το είδος των ανιχνευτών που ακινητοποιούνται στην επιφάνεια. Παράλληλα, ετοιμάζεται το βιολογικό υλικό, δηλαδή απομονώνεται mRNA από δύο τύπους κυττάρων και σημαίνεται με διαφορετικές χρωστικές για κάθε δείγμα, χρησιμοποιώντας την αντίστροφη μεταγραφάση. Κατόπιν, ακολουθεί υβριδοποίηση των σημασμένων στόχων με τους ανιχνευτές της μικροσυστοιχίας. Το επόμενο βήμα είναι η σάρωση της επιφάνειας της μικροσυστοιχίας, απ' όπου προκύπτει μια ψηφιακή εικόνα που προέρχεται από τη διέγερση των μορίων σήμανσης που βρίσκονται στους στόχους και

φθορίζουν σε συγκεκριμένα μήκη κύματος. Οι διχρωματικές ή δικαναλικές μικροσυστοιχίες (Two-color microarrays or two-channel microarrays) συνήθως υβριδοποιούνται με cDNA από δύο δείγματα με σκοπό να συγκριθούν μεταξύ τους, π.χ. καρκινικός και υγιής ιστός, τα οποία σημαίνονται με δύο διαφορετικές χρωστικές ουσίες. Οι πιο συνηθισμένες για σήμανση cDNA περιλαμβάνουν την Cy3, που έχει ένα μήκος κύματος εκπομπής των 570 nm όταν διεγερθεί και αντιστοιχεί στο πράσινο χρώμα, και την Cy5 που έχει ένα μήκος κύματος εκπομπής των 670 nm όταν διεγερθεί και αντιστοιχεί στο κόκκινο χρώμα. Τα δύο δείγματα cDNA που έχουν σημανθεί με Cy, αναμιγνύονται και υβριδίζονται σε μία κοινή μικροσυστοιχία η οποία σαρώνεται σε ένα σαρωτή μικροσυστοιχιών, ώστε να απεικονιστεί η σήμανση των δύο ουσιών μετά την διέγερση από ακτίνα λέιζερ συγκεκριμένου μήκους κύματος. Οι συγκρινόμενες σχετικές εντάσεις κάθε κηλίδας χρησιμοποιούνται για να ταυτοποιηθούν υπερεκφρασμένα και υποεκφρασμένα γονίδια.

Λόγω του γεγονότος ότι η υβριδοποίηση γίνεται ταυτόχρονα και για τα δύο δείγματα, υπάρχει σαφής ανταγωνισμός μεταξύ των στόχων για τους ανιχνευτές. Οι στόχοι που βρίσκονται σε περίσσεια από το κάθε δείγμα για κάθε γονίδιο, θα υπερισχύσουν έναντι των λιγότερων άλλων και θα καταλάβουν περισσότερους ανιχνευτές. Στην παραχθείσα εικόνα μετά από σάρωση, Θα δούμε κόκκινο τον ανιχνευτή αν το συγκεκριμένο γονίδιο υπερεκφράζεται στα κύτταρα του πρώτου δείγματος, θα δούμε πράσινο τον ανιχνευτή αν ισχύει το αντίστοιχο για το δεύτερο δείγμα, θα δούμε κίτρινο τον ανιχνευτή αν η έκφραση είναι παρόμοια και, τέλος, θα δούμε μαύρο τον ανιχνευτή αν δεν υπάρχει καθόλου έκφραση. Όταν, τέλος, έχουμε την εικόνα, προχωράμε σε ανάλυσή της από την οποία προκύπτουν ποσοτικοποιημένα δεδομένα. Γίνεται επεξεργασία των δεδομένων αυτών στον ηλεκτρονικό υπολογιστή με πληθώρα αλγορίθμων, ώστε να απαλειφθούν τα σφάλματα και να δώσουν συμπεράσματα τα οποία ο άνθρωπος δε θα μπορούσε να εξάγει, λόγω του μεγάλου τους όγκου.

Στην περίπτωση των μονοχρωματικών ή μονοκαναλικών μικροσυστοιχιών (single-channel microarrays or one-color microarrays), οι μικροσυστοιχίες προσφέρουν δεδομένα απόλυτες τιμές έκφρασης για κάθε ανιχνευτή ή σύνολο ανιχνευτών αντίστοιχο με το επίπεδο υβριδοποίησης με το σημασμένο στόχο. Η σχετική αφθονία γίνεται εμφανής όταν συγκριθεί με άλλα δείγματα ή

καταστάσεις που έχουν υποστεί επεξεργασία στο ίδιο πείραμα. Μπορούν να συγκριθούν με μεγάλη ευκολία τα δείγματα από διαφορετικά πειράματα, με την προϋπόθεση ότι έχει γίνει η κατάλληλη μέριμνα για τα batch-effects (επίδραση του συνόλου παραγωγής). Τα batch-effects μπορούν να προκύψουν σε οποιοδήποτε στάδιο της πειραματικής διαδικασίας και αναφέρονται στις συστηματικές διαφορές που έχουν οι μετρήσεις από διαφορετικές ομάδες πειραμάτων.

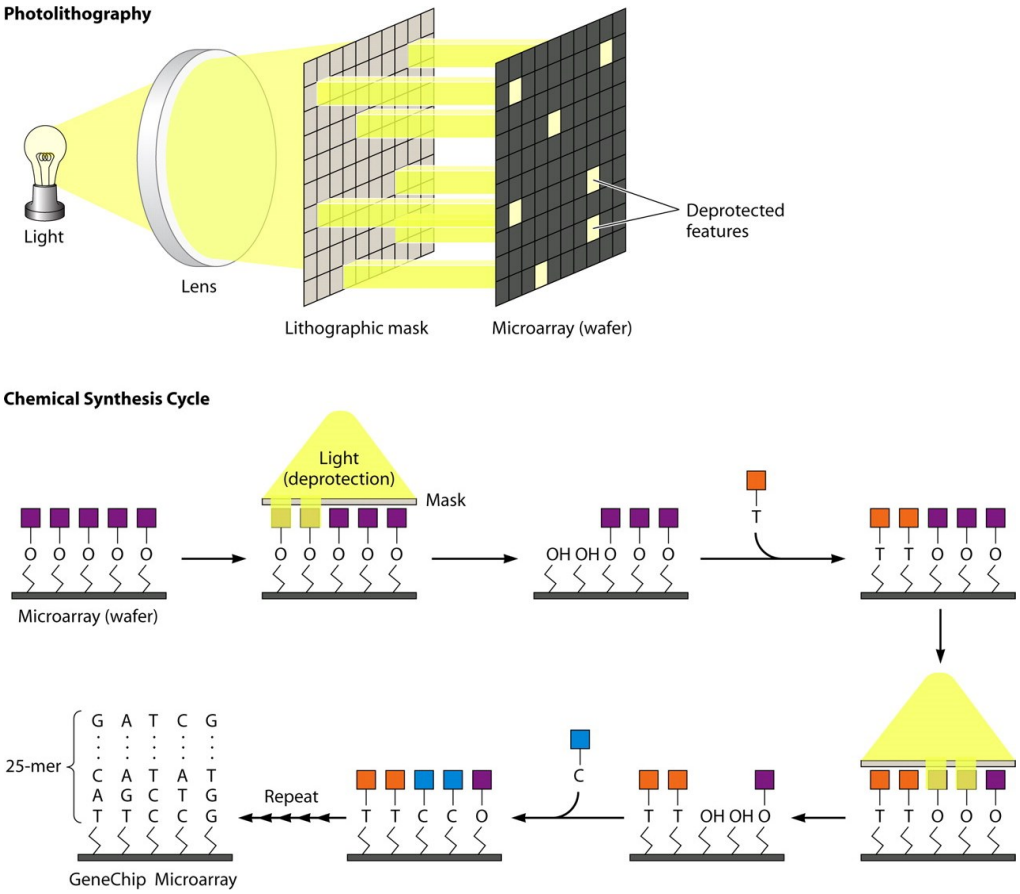


Εικόνα 1 – Τα βασικά βήματα σε ένα πείραμα δικαναλικών μικροσυστοιχιών cDNA

Ολιγονουκλεοτιδικές μικροσυστοιχίες (Affymetrix GeneChip)

Οι μικροσυστοιχίες GeneChip® της εταιρείας Affymetrix ανήκουν στις μονοκαναλικές μικροσυστοιχίες και αποτελούνται από μονόκλωνα 25μερή ολιγονουκλεοτίδια-ανιχνευτές, τα οποία συντίθενται στη στερεή επιφάνεια της μικροσυστοιχίας με τη μέθοδο της φωτολιθογραφίας (Εικόνα 2).

Η διαδικασία ξεκινάει με το γυάλινο πλακίδιο (wafer) όπου αποτελεί την στερεή επιφάνεια της μικροσυστοιχίας. Το πλακίδιο εμβαπτίζεται σε σιλάνιο (SiH_4) και τα μόρια του σιλανίου συνδυάζονται με το γυαλί. Ένα μόριο συνδέτης μαζί με ένα φωτοευαίσθητο μόριο προστίθενται σε κάθε μόριο σιλανίου και το μόριο συνδέτης είναι το σημείο έναρξης πρόσδεσης του πρώτου δεσοξυριβονουκλεοτιδίου. Κάθε νουκλεοτίδιο είναι φωτοχημικά τροποποιημένο αφού φέρει μια προστατευτική ομάδα η οποία απομακρύνεται μετά από επίδραση υπεριώδους ακτινοβολίας και αποτελεί το υπόστρωμα, όπου θα προσδεθεί το επόμενο νουκλεοτίδιο. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να σχηματιστούν συγκεκριμένα 25μερή ολιγονουκλεοτίδια σε κάθε ανιχνευτικό σημείο (Lipshutz et al., 1999).



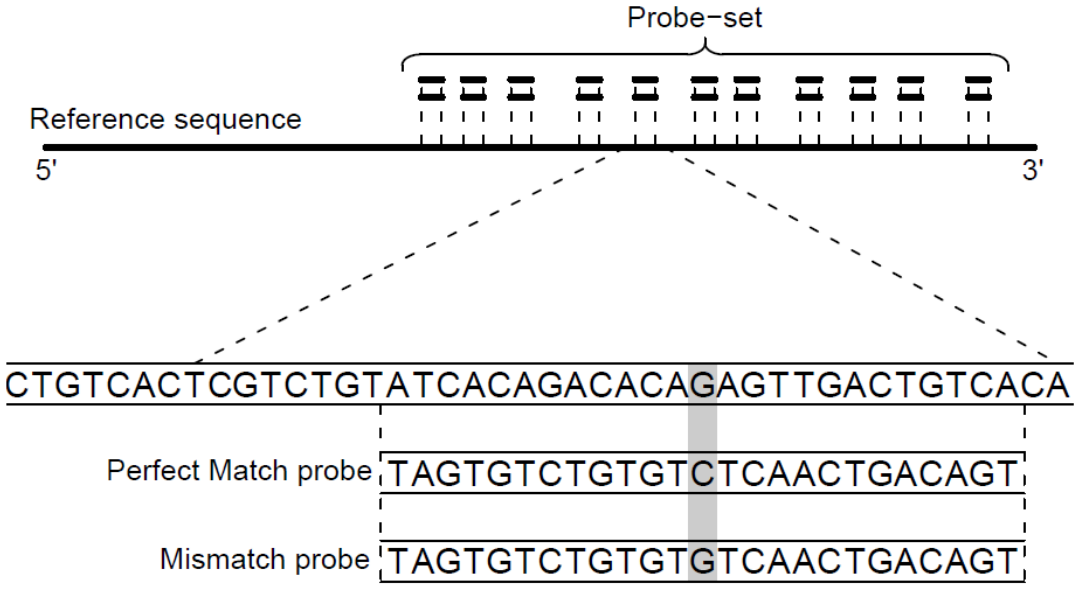
Εικόνα 2 - Μικροσυστοιχία ολιγονουκλεοτιδίων Affymetrix GeneChip®. Επάνω: Φωτολιθογραφία. Υπεριώδης ακτινοβολία διέρχεται μέσω της λιθογραφικής μάσκας που δρα ως φίλτρο, είτε για να

μεταδώσει, είτε να μπλοκάρει την ακτινοβολία από τη χημικά προστατευμένη επιφάνεια της μικροσυστοιχίας. Η διαδοχική εφαρμογή των ειδικών λιθογραφικών μασκών προσδιορίζει τη σειρά της σύνθεσης του ολιγονουκλεοτιδικού ανιχνευτή. Κάτω: Ο κύκλος της χημικής σύνθεσης. Η υπεριώδης ακτινοβολία αφαιρεί τις προστατευτικές ομάδες (τετράγωνα) από την επιφάνεια της μικροσυστοιχίας, επιτρέποντας την προσθήκη ενός μόνο φωτοχημικά προστατευμένου νουκλεοτιδίου. Διαδοχικοί κύκλοι αποπροστασίας με ακτινοβολία, αλλαγής στο μοτίβο φιλτραρίσματος των λιθογραφικών μασκών, και προσθήκη ενός είδους μονονουκλεοτιδίων σχηματίζουν μικροσυστοιχίες με συγκεκριμένα 25μερή ολιγονουκλεοτίδια-ανιχνευτές (Dalma-Weiszhausz et al., 2006).

Κάθε γονίδιο ή νουκλεοτιδική αλληλουχία προς εξέταση αντιπροσωπεύεται από 11 έως 20 μοναδικούς ανιχνευτές που δημιουργούνται μέσω υπολογιστή και είναι διάσπαρτοι στη μικροσυστοιχία, ώστε να αποφευχθεί η λανθασμένη εκτίμηση της ποσοτικοποίησης της έκφρασης λόγω της θέσης τους. Οι ανιχνευτές χρησιμεύουν ως ευαίσθητοι, μοναδικοί, ειδικών αλληλουχιών αισθητήρες. Συνήθως, οι ανιχνευτές υβριδοποιούνται σε ανεξάρτητες περιοχές της αλληλουχίας, όμως ορισμένες φορές μπορεί να έχουν μικρή αλληλοεπικάλυψη (εφόσον αυτό κρίνεται απαραίτητο). Η ομάδα των ανιχνευτών που αφορά συγκεκριμένο γονίδιο ή ομάδα παρόμοιων γονιδίων, είναι γνωστή ως σύνολο ανιχνευτών (probe set) το οποίο παρέχει, με υψηλή ακρίβεια, τον υπολογισμό της έκφρασης του γονιδίου-στόχου. Οι ολιγονουκλεοτιδικοί ανιχνευτές που αναγνωρίζουν τμήματα του 3΄ άκρου του γονιδίου, καλούνται ανιχνευτές τέλειου ταιριάσματος (Perfect Match ή PM). Ο μεγάλος αριθμός των ανιχνευτών για διαφορετικές περιοχές του ίδιου RNA, βελτιώνει σημαντικά το λόγο της έντασης σήματος ως προς το θόρυβο (λόγω του υπολογισμού της μέσης τιμής των εντάσεων των πολλαπλών ανιχνευτικών σημείων) και παρέχει ακρίβεια στην ποσοτικοποίηση του RNA, ενώ αποτρέπει φαινόμενα διασταυρούμενης υβριδοποίησης (crosshybridization effect) και μειώνει δραστικά τα ψευδώς θετικά σήματα (Lipshutz et al., 1999).

Επιπλέον έλεγχος γίνεται με την χρήση των ανιχνευτών ατελούς ταιριάσματος (Mismatch ή MM) (Εικόνα 3). Οι ανιχνευτές MM έχουν ακριβώς την ίδια νουκλεοτιδική ακολουθία με τα αντίστοιχα PM με τη διαφορά ότι στη 13η βάση υπάρχει η συμπληρωματική βάση. Οι ανιχνευτές MM λειτουργούν ως εξειδικευμένοι έλεγχοι που επιτρέπουν την αφαίρεση του υποβάθρου και της διασταυρούμενης υβριδοποίησης, και επιτρέπουν τη διάκριση μεταξύ πραγματικών σημάτων και εκείνων που οφείλονται σε μη-ειδικό υβριδισμό. Η υβριδοποίηση των μορίων στόχων RNA στα PM παράγει υψηλότερο σήμα από ότι στα MM, έχοντας ως αποτέλεσμα σταθερά πρότυπα τα οποία είναι απίθανο να έχουν προκύψει τυχαία. Ακόμη και σε χαμηλές συγκεντρώσεις RNA, η

υβριδοποίηση στα PM/MM αποδίδει αναγνωρίσιμα πρότυπα που μπορούν να ποσοτικοποιηθούν. Κάθε ανιχνευτής MM βρίσκεται στη διπλανή θέση από αυτή του αντίστοιχου PM για να αποκλεισθεί οποιαδήποτε επίδραση λόγω θέσης.



Εικόνα 3 –Σχεδιασμός ανιχνευτών μικροσυστοιχίας Affymetrix. Οι ολιγονουκλεοτιδικοί ανιχνευτές επιλέγονται βάσει κριτηρίων μοναδικότητας και κανόνων σχεδιασμού. Για τους ευκαρυωτικούς οργανισμούς, οι ανιχνευτές επιλέγονται τυπικά από το 3΄ άκρο του γονιδίου ή μεταγράφου (κοντά στην ουρά polyA), για να μειωθούν τα προβλήματα που μπορεί να προκύψουν από τη χρήση του μερικώς αποικοδομημένου mRNA. Η χρήση της διαφοράς των PM από τα MM μειώνει σημαντικά το θόρυβο υποβάθρου και της διασταυρούμενης υβριδοποίησης, και αυξάνει την ποσοτική ακρίβεια και επαναληψιμότητα των μετρήσεων.

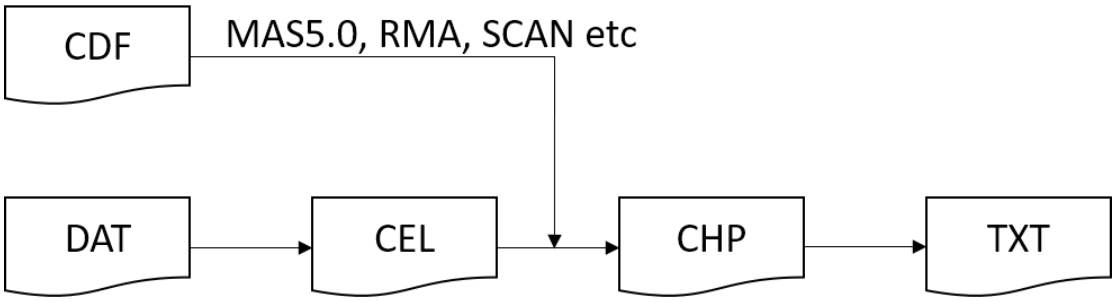
Αρχεία Affymetrix GeneChip

Το λογισμικό της Affymetrix για την επεξεργασία των δεδομένων από τις μικροσυστοιχίες GeneChip χρησιμοποιεί πολλά αρχεία δεδομένων με διάφορες μορφές (Εικόνα 4). Τα πιο κοινά είναι:

- Αρχείο DAT: Αποτελεί τη μη επεξεργασμένη οπτική εικόνα του υβριδοποιημένου πλακιδίου. Περιέχει τις τιμές έντασης των εικονοστοιχείων που συλλέγονται από ένα σαρωτή Affymetrix.
- Αρχείο CEL: Αποθηκεύει τα αποτελέσματα των υπολογισμών της έντασης κάθε ανιχνευτή από τις τιμές των εικονοστοιχείων του αρχείου DAT. Αυτά περιλαμβάνουν την τιμή της έντασης, την τυπική απόκλιση της έντασης και τον αριθμό των εικονοστοιχείων που χρησιμοποιούνται για τον υπολογισμό της έντασης κάθε ανιχνευτή. Το αρχείο αυτό αποτελεί τα πραγματικά πρωτογενή δεδομένα της τεχνολογίας αυτής.
- Αρχείο CDF (Chip Description File): Περιγράφει τη θέση των ανιχνευτών σε μια μικροσυστοιχία Affymetrix GeneChip. Πρακτικά, αποτελεί μία

χαρτογράφηση της συγκεκριμένης μικροσυστοιχίας. Αντιστοιχεί κάθε ανιχνευτή σε ένα σύνολο ανιχνευτών. Όλα τα ονόματα των συνόλων ανιχνευτών εντός μιας μικροσυστοιχίας είναι μοναδικά.

- Αρχείο CHP: Περιέχει τις τιμές έκφρασης ανά σύνολο ανιχνευτών που προκύπτουν από την ανάλυση του αρχείου CEL. Υπάρχουν διάφορες εκδόσεις του αρχείου CHP που δημιουργούνται από λογισμικά της Affymetrix χρησιμοποιώντας αλγορίθμους όπως MAS5.0, RMA, κλπ. Έχει δυαδική μορφή για χρήση σε διάφορα προγράμματα ανάλυσης των αποτελεσμάτων.
- Αρχείο TXT: Είναι το αρχείο CHP σε αναγνώσιμη μορφή (κείμενο).



Εικόνα 4 - Τα αρχεία της μικροσυστοιχίας Affymetrix Genechip και με ποια σειρά χρησιμοποιούνται.

Αναλύσεις διαφορικής έκφρασης και γονιδιακής συνέκφρασης

Χρησιμοποιώντας την τεχνολογία των μικροσυστοιχιών, υπάρχουν δύο βασικά είδη αναλύσεων που μπορούν να πραγματοποιηθούν:

- Ανάλυση διαφορικής έκφρασης

Στην ανάλυση διαφορικής έκφρασης οι εκφράσεις γονιδίων από δείγματα δύο διαφορετικών βιολογικών καταστάσεων συγκρίνονται, ώστε να ανακαλυφθούν γονίδια με στατιστικώς σημαντική διαφορά στην έκφρασή τους. Με βάση αυτή τη λίστα διαφορικώς εκφρασμένων γονιδίων, μπορούμε να εξάγουμε κάποιο συμπέρασμα για το ποια παίζουν βασικό ρόλο στη διαφοροποίηση μεταξύ των δύο συγκρινόμενων καταστάσεων. Σε κάθε πείραμα προτείνεται να έχει τουλάχιστον 3 τεχνικά αντίγραφα ώστε να μειωθούν οι πιθανότητες για τεχνικά σφάλματα. Επίσης, μπορούμε να έχουμε και βιολογικά αντίγραφα, που είναι η επανάληψη του πειράματος ίδιου σχεδιασμού με τις ίδιες συνθήκες αλλά σε διαφορετικό δείγμα.

- Ανάλυση Γονιδιακής Συνέκφρασης

Αντίθετα η ανάλυση γονιδιακής συνέκφρασης μελετάει το σύνολο της έκφρασης γονιδίων για έναν συγκεκριμένο οργανισμό από πολλά διαφορετικά δείγματα. Έτσι ανακαλύπτονται ομάδες γονιδίων με κοινά προφίλ έκφρασης, τα οποία μπορούμε να πούμε ότι συνδέονται λειτουργικά. Η ανάλυση γονιδιακής συνέκφρασης γίνεται χρησιμοποιώντας μονοκαναλικές μικροσυστοιχίες.

Προεπεξεργασία μικροσυστοιχιών Affymetrix GeneChip

Λόγω του σχεδιασμού των μικροσυστοιχιών της Affymetrix, τα βήματα που πρέπει να ληφθούν πριν από οποιαδήποτε ανάλυση έκφρασης είναι ελαφρώς πιο περίπλοκα από ότι για τις άλλες μικροσυστοιχίες cDNA, τα οποία θα περιγράψουμε παρακάτω.

Διόρθωση υποβάθρου

Γενικά το πρώτο βήμα είναι η διόρθωση υποβάθρου της έντασης του κάθε ανιχνευτή. Ο φθορισμός υποβάθρου μπορεί να προκύψει από πολλές πηγές, όπως η μη ειδική σύνδεση του σημασμένου δείγματος στην επιφάνεια της μικροσυστοιχίας, ή οι εναποθέσεις που παρέμειναν μετά το στάδιο της πλύσης ή από οπτικό θόρυβο από το σαρωτή. Κάποιο επίπεδο φθορισμού (θόρυβος υποβάθρου) θα εντοπιστεί από το σαρωτή, ακόμη και εάν μόνο αποστειρωμένο νερό είναι επισημασμένο και υβριδιστεί στη μικροσυστοιχία. Οι διάφοροι αλγόριθμοι χρησιμοποιούν διαφορετικές μεθόδους διόρθωσης υποβάθρου, για παράδειγμα, ο αλγόριθμος RMA χρησιμοποιεί τη συνέλιξη του σήματος και του θορύβου.

Κανονικοποίηση

Το επόμενο στάδιο είναι η κανονικοποίηση. Ο σκοπός αυτού του βήματος είναι να προσαρμόσουν τα δεδομένα στις τεχνικές διακυμάνσεις, σε αντίθεση με τις βιολογικές διαφορές μεταξύ των δειγμάτων. Υπάρχουν πάντα μικρές αποκλίσεις μεταξύ των διαδικασιών υβριδισμού για κάθε μικροσυστοιχία και αυτές οι αποκλίσεις τείνουν να οδηγούν σε μεγάλες αποκλίσεις μεταξύ των συνολικών εντάσεων για διαφορετικές μικροσυστοιχίες. Για παράδειγμα, η ποσότητα του RNA σε ένα δείγμα, ο χρόνος που ένα δείγμα υβριδίζεται ή ο

όγκος ενός δείγματος μπορεί να εισαγάγουν σημαντικές διακυμάνσεις. Ακόμη και ανεπαίσθητες υλικές διαφορές μεταξύ των μικροσυστοιχιών ή μεταξύ των σαρωτών που χρησιμοποιούνται για να σαρώσουν τις μικροσυστοιχίες, μπορεί να έχουν επίδραση στα αποτελέσματα. Με απλά λόγια, η κανονικοποίηση εξασφαλίζει ότι η σύγκριση επιπέδων έκφρασης διαφορετικών μικροσυστοιχιών, είναι όσο το δυνατόν, σύγκριση ομοειδών. Μελέτες έχουν δείξει ότι οι μέθοδοι κανονικοποίησης που χρησιμοποιούνται έχουν σημαντική επίδραση στα τελικά επίπεδα διαφορικής έκφρασης, γι' αυτό είναι ζωτικής σημασίας να επιλεγεί η κατάλληλη μέθοδος.

Διόρθωση PM

Όπως αναφέρθηκε προηγουμένως, ανιχνευτές PM της μικροσυστοιχίας GeneChip μετράνε τόσο τη σχετική αφθονία του αντίστοιχου γονιδίου όσο και την ποσότητα της μη-ειδικής δέσμευσης, η οποία προκύπτει όταν το mRNA δεσμεύεται σε έναν ανιχνευτή που δεν τον στοχεύει. Οι ανιχνευτές MM έχουν σχεδιαστεί για να μετράνε τη μη ειδική σύνδεση των αντίστοιχών τους ανιχνευτών PM. Οι τιμές MM θα πρέπει να αφαιρούνται από τις αντίστοιχες τους τιμές PM, ως πρώτο βήμα στη διαδικασία ανάλυσης. Στην πραγματικότητα, ωστόσο, αυτό δεν λειτουργεί, επειδή γενικά περίπου το 30% των τιμών MM είναι στην πράξη μεγαλύτερες από τις αντίστοιχες τιμές PM τους. Αυτό συμβαίνει επειδή, επιπρόσθετα της μέτρησης του σήματος υποβάθρου, μεγάλος όγκος του mRNA που αναγνωρίζεται από τους ανιχνευτές PM, τείνει να δεσμευτεί επίσης σε ανιχνευτές MM. Πολλές από τις πιο δημοφιλείς μεθόδους προεπεξεργασίας λύνουν αυτό το πρόβλημα απλά αγνοώντας τους ανιχνευτές MM εντελώς και οι τιμές PM διορθώνονται για τη μη ειδική σύνδεση χρησιμοποιώντας άλλες προσεγγίσεις.

Σύνοψη ανιχνευτών

Έχουμε ήδη δει πως οι μικροσυστοιχίες GeneChip λειτουργούν χρησιμοποιώντας 11 έως 20 διαφορετικούς ανιχνευτές PM που στοχεύουν σε 11 έως 20 χωριστά νουκλεοτιδικά τμήματα κάθε mRNA. Οι ανιχνευτές αυτοί αποτελούν ένα Σύνολο Ανιχνευτών (Probe set). Το τελικό βήμα στη προεπεξεργασία δεδομένων από μικροσυστοιχίες GeneChip, είναι να γίνει σύνοψη των δεδομένων από 11-20 χωριστούς ανιχνευτές σε μια τιμή έκφρασης για το εν λόγω mRNA. Για την επίτευξη αυτού υπάρχουν αρκετοί διαφορετικοί

τρόποι, αλλά το τελικό αποτέλεσμα είναι πάντα μια μοναδική τιμή έκφρασης για κάθε σύνολο ανιχνευτών σε κάθε μικροσυστοιχία.

Μέθοδοι προεπεξεργασίας υλοποιημένες για μικροσυστοιχίες Affymetrix GeneChip

Έχοντας εισάγει τη γενική μεθοδολογία που ακολουθείται για προεπεξεργασία δεδομένων μικροσυστοιχιών Affymetrix, θα περιγράψουμε μερικούς από τους πιο δημοφιλείς σύνθετους αλγορίθμους προεπεξεργασίας. Αυτοί οι αλγόριθμοι εφαρμόζουν τα τέσσερα βήματα προεπεξεργασίας που περιγράφονται ανωτέρω, δηλαδή κάνουν διόρθωση υποβάθρου και κανονικοποιούν τις τιμές έκφρασης για κάθε σύνολο ανιχνευτών σε κάθε μικροσυστοιχία.

Microarray Suite 5.0 (MAS5.0)

Ο αλγόριθμος MAS5.0 (Bolstad et al., 2005 ; Gentleman et al., 2004) αναπτύχθηκε από την Affymetrix. Αρχικά κάνει διόρθωση υποβάθρου τόσο στους ανιχνευτές PM, όσο και στους MM. Οι MM στη συνέχεια μετατρέπονται σε ιδανικά ατελή ταιριάσματα, όπου οι τιμές τους είναι πάντα μικρότερες από τις τιμές των αντίστοιχων τους PM. Θυμηθείτε ότι περίπου το 30% των MM τιμών είναι μεγαλύτερες από αυτές των αντίστοιχων PM τους. Αν MM < PM, τότε η MM τιμή παραμένει αμετάβλητη. Υπολογίζεται ένα σταθερός μέσος όρος από τις τροποποιημένες $\log_2$ διαφορές μεταξύ των PM και των ήδη υπολογισμένων ιδανικών ατελών ταιριασμάτων (IM). Οι τιμές έκφρασης κανονικοποιούνται ρυθμίζοντας τον μέσο όρο των αρχικών σημάτων της κάθε μικροσυστοιχίας με μία προδιαγεγραμμένη τιμή. Ως εκ τούτου, ο MAS5.0 κανονικοποιεί τα δεδομένα μετά από σύνοψη, και όχι πριν, όπως σε πολλούς άλλους αλγορίθμους.

Probe Logarithmic Intensity Error Estimation (PLIER)

Η Affymetrix υποστηρίζει ότι ο αλγόριθμος PLIER είναι βελτίωση του MAS5.0 εισάγοντας υψηλότερη επαναληψιμότητα σήματος (κατώτερος συντελεστής μεταβολής) χωρίς απώλεια ακρίβειας. Προσφέρει υψηλότερη ευαισθησία στις αλλαγές αφθονίας για στόχους κοντά στο υπόβαθρο και ισοσταθμίζει δυναμικά τους ανιχνευτές που περιέχουν περισσότερες

πληροφορίες από ένα σύνολο δεδομένων για τον προσδιορισμό του σήματος. Ο αλγόριθμος PLIER δεν κανονικοποιεί τα δεδομένα.

Robust Multi-Array Analysis (RMA)

Ο αλγόριθμος RMA είναι η ακαδημαϊκή εναλλακτική λύση για τους αλγορίθμους της Affymetrix για τη μετατροπή των δεδομένων από επίπεδο ανιχνευτών σε τιμές γονιδιακής έκφρασης. Αυτή η μέθοδος είναι διαφορετική από τις μεθόδους της Affymetrix, επειδή αγνοεί εντελώς τις τιμές των ανιχνευτών MM (Bolstad et al., 2005 ; Irizarry et al., 2003b) Οι εφευρέτες του αλγόριθμου υποστηρίζουν ότι οι ανιχνευτές MM εισάγουν περισσότερο θόρυβο. Παρότι αναγνωρίζουν ότι οι ανιχνευτές MM παρέχουν χρήσιμες πληροφορίες, μέχρι τη στιγμή της δημοσίευσης της μεθόδου, δεν βρήκαν κάποιο παραγωγικό τρόπο για να τους χρησιμοποιήσουν. Ο αλγόριθμος λειτουργεί προσαρμόζοντας τα σήματα για το θόρυβο υποβάθρου με μια πρωτογενή κλίμακα έντασης, η οποία δεν οδηγεί σε αρνητικές τιμές διόρθωσης υποβάθρου. Στη συνέχεια λαμβάνεται η $\log_2$ μετασχηματισμένη τιμή κάθε PM ανιχνευτή, διορθωμένη από το θόρυβο υποβάθρου, και αυτές οι τιμές κανονικοποιούνται χρησιμοποιώντας κανονικοποίηση ποσοστημορίων. Η ανάλυση των μικροσυστοιχιών προς εξέταση, διεξάγεται στη συνέχεια επί των ποσοστημορίων.

GeneChip RMA (GC-RMA)

Ο αλγόριθμος GCRMA βασίζεται σε μεγάλο βαθμό στον RMA και στην πραγματικότητα διαφέρει μόνο στο βήμα διόρθωσης υποβάθρου όπου χρησιμοποιεί τις ακολουθίες των ανιχνευτών για να εκτιμηθεί το υπόβαθρο καλύτερα. Επίσης, οι ανιχνευτές MM ρυθμίζονται ανάλογα με την συγγένεια των ανιχνευτών και μετά οι τιμές τους αφαιρούνται από τους ανιχνευτές PM. Αυτό οδηγεί σε βελτιωμένη στόχευση στην αναλογία μεγέθους, αλλά εις βάρος οριακά χαμηλότερης ακρίβειας.

Frozen Robust Multi-Array Analysis (fRMA)

Ο αλγόριθμος FRMA αποτελεί μία βελτίωση του αλγορίθμου RMA, από τους ίδιους δημιουργούς (McCall et al., 2010). Προσπαθεί να καλύψει τα κενά του RMA σε περιπτώσεις οπού στα δείγματα πρέπει να γίνει επεξεργασία ατομικά ή σε μικρές ομάδες ή όταν συλλογές δεδομένων που έχουν υποστεί

ξεχωριστή προεπεξεργασία δεν είναι συγκρίσιμες μεταξύ τους. Ο fRMA, επιτρέπει την ανάλυση δειγμάτων ατομικά ή σε μικρές ομάδες και στη συνέχεια συνδυάζει τα δεδομένα για την εκτέλεση της ανάλυσης. Βασίζεται πάνω σε ήδη υπολογισμένα δεδομένα μικροσυστοιχιών από δημόσιες βάσεις δεδομένων.

Single Channel Array Normalisation (SCAN)

Ο αλγόριθμος SCAN αποτελεί μία μέθοδο προεπεξεργασίας για μονοκαναλικές μικροσυστοιχίες (Piccolo et al., 2012) και το κύριο πλεονέκτημά του είναι ότι προσπαθεί να καλύψει για τις τεχνολογικές προκαταλήψεις καθώς και άλλους τυχαίους παράγοντες που μπορεί να προκύψουν κατά τη διάρκεια πειραμάτων μικροσυστοιχιών. Ο SCAN κανονικοποιεί κάθε μικροσυστοιχία ανεξάρτητα από τα υπόλοιπα δείγματα της σειράς για αυτό και προτιμάται όταν χρησιμοποιούμε δείγματα από πολλά διαφορετικά πειράματα. Περισσότερες λεπτομέρειες για τον SCAN αναφέρονται στο αντίστοιχο κομμάτι των Μεθόδων.

Minimum Information About a Microarray Experiment (MIAME)

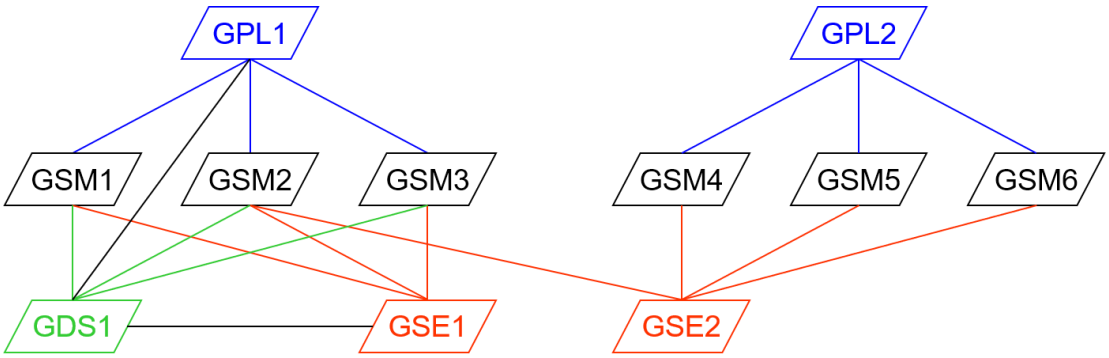
Οι ελάχιστες πληροφορίες για ένα πείραμα μικροσυστοιχιών (MIAME) είναι μια κοινή γλώσσα για την αναπαράσταση και την επικοινωνία των δεδομένων μικροσυστοιχιών (Brazma et al., 2001). Περιλαμβάνουν πληροφορίες σχετικά με το συνολικό πειραματικό σχεδιασμό, το σχεδιασμό της μικροσυστοιχίας (δηλαδή αναγνώριση του κάθε ανιχνευτή σε κάθε μικροσυστοιχία), την προέλευση του κάθε ανιχνευτή και τη μέθοδο σήμανσης, τις διαδικασίες και τις παραμέτρους υβριδισμού και τις διαδικασίες μέτρησης (συμπεριλαμβανομένων των μεθόδων κανονικοποίησης). Τα έξι πιο κρίσιμα στοιχεία του MIAME είναι τα παρακάτω:

- Πρωτογενή δεδομένα κάθε υβριδισμού (π.χ., αρχεία CEL ή GPR)
- Επεξεργασμένα (κανονικοποιημένα) δεδομένα για το σύνολο των υβριδισμών στην πειραματική μελέτη (π.χ., ο πίνακας δεδομένων γονιδιακής έκφρασης που χρησιμοποιείται για να εξαχθούν τα συμπεράσματα από τη μελέτη)
- Βασικοί σχολιασμοί των δειγμάτων, συμπεριλαμβανομένων των πειραματικών παραγόντων και των τιμών τους (π.χ., ουσία και δόση σε πείραμα απόκρισης σε δόση)

- Πειραματικός σχεδιασμός, συμπεριλαμβανομένων των σχέσεων δειγμάτων και δεδομένων (π.χ., ποιο αρχείο πρωτογενών δεδομένων σχετίζεται με ποιο δείγμα, ποιοι υβριδισμοί είναι τεχνικές και βιολογικές επαναλήψεις)
- Επαρκής σχολιασμός της μικροσυστοιχίας (π.χ. αναγνωριστικά γονιδίου, γονιδιωματικές συντεταγμένες, ακολουθίες ολιγονουκλεοτιδικών ανιχνευτών ή αριθμός καταλόγου εμπορικών μικροσυστοιχιών)
- Βασικά πρωτόκολλα εργαστηριακής επεξεργασίας και επεξεργασίας δεδομένων (π.χ., ποια μέθοδος κανονικοποίησης έχει χρησιμοποιηθεί για τη λήψη των τελικών επεξεργασμένων δεδομένων)

Η βάση δεδομένων Gene Expression Omnibus (GEO)

Η βάση δεδομένων Gene Expression Omnibus (GEO) (Barrett et al., 2013) του NCBI χρησιμεύει ως δημόσιο καταθετήριο για ένα ευρύ φάσμα πειραματικών δεδομένων υψηλής απόδοσης. Αυτά τα δεδομένα περιλαμβάνουν πειράματα μονοκάναλων (GeneChip της Affymetrix) και δικάναλων (cDNA) μικροσυστοιχιών mRNA, γονιδιωματικού DNA, πρωτεϊνών, και άλλων τεχνικών όπως η σειριακή ανάλυση έκφρασης γονιδίων (SAGE), η φασματομετρία μάζας πρωτεϊνικών δεδομένων και η μέθοδος εύρεσης αλληλουχιών επόμενης γενιάς. Επιπλέον σχεδόν πάντα μαζί με τα επεξεργασμένα δεδομένα υψηλής απόδοσης, κατατίθενται και τα μη επεξεργασμένα (πρωτογενή) αντίστοιχα δεδομένα τους. Στο βασικό επίπεδο οργάνωσης του GEO, υπάρχουν τέσσερις βασικοί τύποι οντότητας. Οι πρώτοι τρείς (Δείγμα, Πλατφόρμα, και Σειρά) συμπληρώνονται και παρέχονται στο GEO από τους χρήστες. Το προσωπικό του GEO καταρτίζει και επιμελείται τον τέταρτο τύπο, το Σύνολο Δεδομένων, χρησιμοποιώντας τα δεδομένα που έχουν υποβληθεί από τους χρήστες (Εικόνα 5).



Εικόνα 5 - Αναπαράσταση των εγγραφών της βάσης δεδομένων GEO

Πλατφόρμες (Platforms)

Μια εγγραφή Πλατφόρμας περιγράφει τα χαρακτηριστικά της μικροσυστοιχίας (π.χ. cDNA, probe sets ολιγονουκλεοτιδίων, ORFs, αντισώματα), τον κατάλογο των στοιχείων που μπορούν να ανιχνευθούν και ποσοτικοποιηθούν σε αυτό το πείραμα (π.χ., υπογραφές SAGE, πεπτίδια), κλπ. Κάθε εγγραφή Πλατφόρμας έχει έναν μοναδικό και σταθερό αριθμό καταχώρησης GEO που ξεκινάει πάντα με τα γράμματα “GPL” και ακολουθείται από αριθμούς (π.χ. Πλατφόρμα GPL96 περιγράφει την μικροσυστοιχία Affymetrix Human Genome U133A). Η Πλατφόρμα μπορεί να παραπέμπει σε πολλά δείγματα που έχουν υποβληθεί από διάφορους χρήστες.

Δείγματα (Samples)

Μια εγγραφή Δείγματος περιγράφει την προέλευση κάθε ξεχωριστού δείγματος, τα πειραματικά πρωτόκολλα συλλογής, εκχύλισης, σήμανσης, υβριδισμού και σάρωσης, την υπολογιστική επεξεργασία των ληφθέντων πρωτογενών δεδομένων καθώς και τη μέτρηση της αφθονίας κάθε στοιχείου που προκύπτει από το δείγμα. Κάθε εγγραφή Δείγματος έχει έναν μοναδικό και σταθερό αριθμό καταχώρησης GEO που ξεκινάει πάντα με τα γράμματα “GSM” και ακολουθείται από αριθμούς (π.χ. GSM845740). Κάθε Δείγμα πρέπει να παραπέμπει σε μια μόνο Πλατφόρμα και μπορεί να συμπεριληφθεί σε μία ή περισσότερες Σειρές.

Σειρές (Series)

Μια εγγραφή Σειράς καθορίζει μια συλλογή Δειγμάτων που ανήκουν σε μια ομάδα, εξηγεί πώς σχετίζονται τα δείγματα και πώς διευθετούνται. Η Σειρά είναι το κομβικό σημείο της περιγραφής του πειράματος συνολικά. Οι εγγραφές των Σειρών μπορεί επίσης να περιέχουν πίνακες που περιγράφουν εξαγόμενα

δεδομένα, συνοπτικά συμπεράσματα, ή αναλύσεις. Κάθε εγγραφή Σειράς έχει έναν μοναδικό και σταθερό αριθμό GEO που ξεκινάει πάντα με τα γράμματα “GSE” και ακολουθείται από αριθμούς (π.χ. GSE34248) και επίσης, μπορεί να περιλαμβάνει Δείγματα από διαφορετικές Πλατφόρμες.

Σύνολα Δεδομένων (Datasets)

Τα Σύνολα Δεδομένων GEO είναι επιμελημένες ομάδες Δειγμάτων GEO. Μια εγγραφή Συνόλου Δεδομένων είναι μια συλλογή από βιολογικά και στατιστικά συγκρίσιμα Δείγματα GEO και αποτελεί τη βάση της σουίτας διαδικτυακών εφαρμογών του GEO των εργαλείων ανάλυσης και προβολής δεδομένων. Τα Δείγματα κάθε Συνόλου Δεδομένων ανήκουν αποκλειστικά σε μια Πλατφόρμα. Οι τιμές των μετρήσεων για κάθε Δείγμα που ανήκει σε ένα Σύνολο Δεδομένων υπολογίζονται με ταυτόσημο τρόπο, δηλαδή, παράγοντες όπως η επεξεργασία υποβάθρου και η κανονικοποίηση είναι κοινές σε ολόκληρη την ομάδα δεδομένων. Επιπλέον πληροφορίες που αντανακλούν τον πειραματικό σχεδιασμό, παρέχονται μέσω υποσυνόλων Δειγμάτων. Η κάθε εγγραφή Συνόλου Δεδομένων έχει έναν μοναδικό και σταθερό αριθμό GEO που ξεκινάει πάντα με τα γράμματα “GDS” και ακολουθείται από αριθμούς.

Η βάση δεδομένων GTEx

Το GTEx (Genotype-Tissue Expression) είναι ένα ερευνητικό έργο που έχει δημιουργήσει μια τράπεζα ιστών και μια σχετιζόμενη βάση δεδομένων για την επιστημονική κοινότητα, για τη μελέτη της σχέσης μεταξύ της γενετικής διαφοροποίησης και της γονιδιακής έκφρασης στους μη νοσούντες ανθρώπινους ιστούς (GTEx Consortium, 2013). Ο στόχος του GTEx είναι να καθορίσει πώς η γενετική διαφοροποίηση επηρεάζει την φυσιολογική γονιδιακή έκφραση στους ανθρώπινους ιστούς και τελικά να εκτιμήσει πώς αυτή η σχέση συνδέεται με την ανάπτυξη κάποιας ασθένειας. Για την επίτευξη αυτού του στόχου, το πρόγραμμα συλλέγει πολλαπλούς διαφορετικούς ανθρώπινους ιστούς από καθέναν από τους εκατοντάδες δότες, απομονώνει νουκλεϊκά οξέα από τους ιστούς και διεξάγει γονοτυπική ανάλυση, ανάλυση προφίλ γονιδιακής έκφρασης, αλληλούχηση ολόκληρου του γονιδιώματος, αλληλούχηση RNA και αναλύει τα δεδομένα ώστε να προσδιοριστούν οι γενετικοί τόποι ποσοτικών χαρακτηριστικών που σχετίζονται με την έκφραση (eQTL). Οι επιστημονικοί στόχοι του προγράμματος απαιτούν οι δότες και τα βιοϋλικά τους να παρουσιάζονται χωρίς ένδειξη ασθένειας (Carithers et al., 2015).

Οι συσχετισμοί μεταξύ του γονότυπου και των επιπέδων έκφρασης των γονιδίων που είναι ειδικά για τον ιστό, θα βοηθήσουν στον εντοπισμό περιοχών του γονιδιώματος που επηρεάζουν το αν και πόσο εκφράζεται ένα γονίδιο. Το GTEx θα βοηθήσει τους ερευνητές να κατανοήσουν την κληρονομική ευαισθησία σε ασθένειες και είναι μια τράπεζα ιστών και μια βάση δεδομένων για πολλές μελέτες στο μέλλον.

Παρά την ταχεία πρόοδο που έχει επιτευχθεί χρησιμοποιώντας μελέτες γενετικής συσχέτισης σε γονιδιωματική κλίμακα (Genome Wide Assossiation Studies - GWAS) για τον εντοπισμό των γενετικών αλλαγών που συνδέονται με τις κοινές ασθένειες του ανθρώπου, όπως οι καρδιακές παθήσεις, ο καρκίνος, ο διαβήτης και το άσθμα, η μεγάλη πλειοψηφία αυτών των γενετικών αλλαγών βρίσκεται έξω από τις περιοχές κωδικοποίησης πρωτεϊνών των γονιδίων και συχνά ακόμη και εκτός των ίδιων των γονιδίων, καθιστώντας δύσκολη την ανίχνευση των γονιδίων που επηρεάζονται και με ποιο μηχανισμό. Η πλήρης αναγνώριση των ανθρωπίνων eQTL θα συμβάλλει σημαντικά στον προσδιορισμό των γονιδίων των οποίων η έκφραση επηρεάζεται από τη

γενετική ποικιλομορφία και θα αποτελέσει πολύτιμη βάση για να μελετηθεί ο μηχανισμός της γονιδιακής ρύθμισης.
Το πρόγραμμα θα περιλαμβάνει επίσης διαβούλευση και έρευνα για τα βιοηθικά ζητήματα που ανακύπτουν από την έρευνα, υποστήριξη για την ανάπτυξη στατιστικών μεθόδων και τη δημιουργία μιας βάσης δεδομένων για την κάλυψη των υφιστάμενων δεδομένων και των δεδομένων eQTL που παράγονται από το GTEx. Η βάση δεδομένων θα επιτρέπει στους χρήστες να βλέπουν και να λαμβάνουν τα υπολογισμένα αποτελέσματα eQTL και θα παρέχει ένα σύστημα ελεγχόμενης πρόσβασης για τον προσδιορισμένο γονότυπο, την έκφραση και τα κλινικά δεδομένα ατομικού επιπέδου. Το σχετικό αποθετήριο ιστών θα χρησιμεύει επίσης ως πόρος για πολλές πρόσθετες αναλύσεις.

Παραγωγή και διάθεση δεδομένων

Για λόγους προστασίας της ιδιωτικότητας, η πολιτική της NIH εμποδίζει τη διάθεση των πρωτογενών δεδομένων μέσω της πύλης του GTEx. Αυτά τα δεδομένα είναι διαθέσιμα μέσω του dbGaP (Tryka et al., 2014) και των SRA (Kodama et al., 2012) μετά από αίτηση. Σημειώνεται ότι αυτοί που μπορούν να κάνουν αίτηση για πρόσβαση στα πρωτογενή δεδομένα είναι οι Ανώτεροι ερευνητές (πρέπει να είναι μόνιμοι υπάλληλοι του ιδρύματός τους σε επίπεδο ισοδύναμο με καθηγητή προγραμματισμένης πανεπιστημιακής σταδιοδρομίας ή ανώτερο επιστήμονα με καθήκοντα που πιθανότατα περιλαμβάνουν τη διοίκηση και την επίβλεψη εργαστηρίου) και οι ερευνητές της NIH.
Μόνο αρχεία BAM είναι διαθέσιμα στο dbGap, αλλά είναι εύκολο να παραχθούν τα FASTQs από αυτά χρησιμοποιώντας τα εργαλεία Picard (Broad Institute, 2020).

Δεδομένα έκφρασης

Τα δεδομένα έκφρασης που παρέχονται είναι τα εξής:

- Illumina TruSeq RNA sequencing
- Affymetrix Human Gene 1.1 ST Expression Array (V3; 837 samples)

Γονοτυπικά δεδομένα

Τα γονοτυπικά δεδομένα που παρέχονται είναι τα εξής:

- Whole genome sequencing (HiSeq X; first batch on HiSeq 2000)

- Whole exome sequencing (Agilent or ICE target capture, HiSeq 2000)
- Illumina OMNI 5M Array or 2.5M SNP Array
- Illumina Human Exome SNP Array

Μέθοδοι ανάλυσης

Το RNA-seq εκτελέσθηκε με τη χρήση του Illumina TruSeq library construction protocol (non-stranded, polyA+ selection) (Illumina, 2014). Το ολικό RNA ποσοτικοποιήθηκε χρησιμοποιώντας το Quant-iTTM RiboGreen® RNA Assay Kit και κανονικοποιήθηκε σε 5 ng ανά μL. Ένα κλάσμα 200 ng για κάθε δείγμα μεταφέρθηκε σε παρασκεύασμα βιβλιοθήκης, το οποίο ήταν μια αυτοματοποιημένη παραλλαγή του Illumina TruSeqTM RNA sample preparation protocol (Illumina, 2010). Αυτή η μέθοδος χρησιμοποίησε σφαιρίδια oligo dT για να επιλέξει mRNA από το δείγμα ολικού RNA ακολουθούμενο από θραυσματοποίηση με θερμότητα και σύνθεση cDNA από το πρότυπο RNA. Το cDNA που προέκυψε, έπειτα πέρασε από προετοιμασία βιβλιοθήκης (τελική αποκατάσταση, προσθήκη βάσης A, πρόσδεση προσαρμογέα και εμπλουτισμός) με χρήση σχεδιασμένων από το Broad Institute, προσαρμογέων με ευρετηρίαση, υποκατεστημένων για πολυπλεξία. Μετά τον εμπλουτισμό, οι βιβλιοθήκες ποσοτικοποιήθηκαν με qPCR χρησιμοποιώντας το KAPA Library Quantification Kit for Illumina Sequencing Platforms και στη συνέχεια συγκεντρώθηκαν ισομοριακά. Η όλη διαδικασία διεξήχθη σε τρυβλία 96 φρεατίων και ολόκληρη η αναρρόφηση με μικροσυφώνια πραγματοποιήθηκε με χειριστές υγρών είτε Agilent Bravo είτε Hamilton Starlet με ηλεκτρονική παρακολούθηση καθ' όλη τη διάρκεια της διαδικασίας σε πραγματικό χρόνο, συμπεριλαμβάνοντας αριθμούς παρτίδων αντιδραστηρίων, ειδικές αυτοματοποιήσεις που χρησιμοποιήθηκαν, χρονικές σφραγίδες για κάθε στάδιο επεξεργασίας και της αυτόματης εγγραφής.

Οι συγκεντρωμένες βιβλιοθήκες κανονικοποιήθηκαν σε 2 nM και αποδιατάχθηκαν χρησιμοποιώντας 0.1 N NaOH πριν από την αλληλούχηση. Η ενίσχυση του συμπλέγματος κυττάρων ροής και η αλληλούχηση διεξήχθησαν σύμφωνα με τα πρωτόκολλα του κατασκευαστή χρησιμοποιώντας είτε το HiSeq 2000 είτε το HiSeq 2500. Η αλληλούχηση παρήγαγε αναγνώσεις 76bp με ζευγαρωμένα άκρα και ένα κώδικα ευρετηρίου 8 βάσεων και διεξήχθη με στόχο

κάλυψης των 50M αναγνώσεων (η διάμεσος που επιτεύχθηκε ήταν ~82M συνολικές αναγνώσεις).

Προεπεξεργασία

Στοίχιση RNA-seq

Η στοίχιση με το ανθρώπινο γονιδίωμα αναφοράς hg19/GRCh37 εκτελέσθηκε χρησιμοποιώντας το STAR v2.4.2a (Dobin et al., 2013), βασισμένο στο GENCODE v19 (Frankish et al., 2019). Οι μη στοιχισμένες αναγνώσεις διατηρήθηκαν στο τελικό αρχείο BAM. Ανάμεσα στις πολλαπλώς χαρτογραφημένες αναγνώσεις, μία ανάγνωση επισημαίνεται ως κύρια στοίχιση από το STAR. Η διαδικασία της στοίχισης είναι διαθέσιμη στη διεύθυνση:

https://github.com/broadinstitute/gtex-pipeline/tree/master/rnaseq

Γονοτυπική ανάλυση

Ο προσδιορισμός αλληλουχίας ολόκληρου του γονιδιώματος (WGS) πραγματοποιήθηκε από το Broad Institute’s Genomics Platform σε δείγματα DNA από 652 δότες GTEx με μέση κάλυψη 30X. Από αυτά, 68 δείγματα αλληλουχήθηκαν από το Illumina HiSeq 2000 χρησιμοποιώντας αναγνώσεις με ζευγαρωμένα άκρα 101 bp και 584 δείγματα από το Illumina HiSeq X χρησιμοποιώντας αναγνώσεις με ζευγαρωμένα άκρα 151 bp. Το DNA που απομονώθηκε από δείγματα αίματος που συλλέχθηκαν από κάθε δότη GTEx ήταν η πρωταρχική πηγή DNA που χρησιμοποιήθηκε για τον προσδιορισμό του γονότυπου (~100 ng DNA ανά δείγμα). Τα WGS BAM υποβλήθηκαν σε επεξεργασία μέσω μιας διαδικασίας με βάση το Picard χρησιμοποιώντας την επαναβαθμονόμηση της βαθμολογίας ποιότητας βάσης και την τοπική επαναστοίχιση σε γνωστά indels. Οι αναγνώσεις WGS στοιχήθηκαν με το ανθρώπινο γονιδίωμα αναφοράς hg19/GRCh37 που δημιουργήθηκε με το BWA-MEM (Li, 2013). Η κοινή κλήση μετάλλαξης εκτελέσθηκε σε όλα τα δείγματα WGS χρησιμοποιώντας το GATK’s HaplotypeCaller v3.4 (Poplin et al., 2017). Τα ποσοστά κλήσεων ανάλυσης γονότυπου ανά άτομο ξεπέρασαν το 98% για όλα τα δείγματα και όλα τα δείγματα πέρασαν ελέγχους γενετικών δακτυλικών αποτυπωμάτων και συμφωνίας φύλου. Για να αυξηθεί η ποιότητα μετάλλαξης κλήσεων, υπολογίσθηκαν οι πιθανότητες μεταγενέστερου γονότυπου για όλες τις κλήσεις που βασίστηκαν στη συχνότητα των

αλληλόμορφων στο 1000 Genomes Project Phase 3 version 1 και οι βαθμολογίες ποιότητας γονότυπου ενημερώθηκαν αναλόγως, χρησιμοποιώντας το GATK's CalculateGenotypePosteriors.

Ποσοτικοποίηση Έκφρασης

Μεταγραφικό υπόδειγμα

GENCODE19, με ονόματα χρωμοσωμάτων τροποποιημένα για να ταιριάζουν με τα ονόματα χρωμοσωμάτων του hg19.

Συμπτυγμένο γονιδιακό υπόδειγμα

Η ποσοτικοποίηση της έκφρασης στο γονιδιακό επίπεδο βασίστηκε στο GENCODE 19, συμπτυγμένο σε ένα μοντέλο απλού μεταγράφου για κάθε γονίδιο χρησιμοποιώντας μια διαδικασία σύμπτυξης ισομορφής, που περιλαμβάνει τα ακόλουθα βήματα:

- Εξαιρέθηκαν τα εξώνια που ανήκουν σε μετάγραφα που σχολιάσθηκαν ως "retained_intron" και "read_through".
- Τα διαστήματα εξωνίου που επικαλύπτονται μέσα σε ένα γονίδιο συγχωνεύθηκαν.
- Οι διασταυρώσεις των διαστημάτων εξωνίων μεταξύ των γονιδίων που επικαλύπτονται εξαιρέθηκαν.
- Τα υπολειπόμενα διαστήματα εξωνίων χαρτογραφήθηκαν στο αντίστοιχο αναγνωριστικό τους γονίδιο και αποθηκεύτηκαν σε μορφή GTF.

Ο κώδικας για την παραγωγή του συμπτυγμένου μοντέλου είναι διαθέσιμος στο:

https://github.com/broadinstitute/gtex-pipeline/tree/master/gene_model.

Ποσοτικοποίηση

Ποσοτικοποίηση γονιδιακού επιπέδου: οι μετρήσεις ανάγνωσης και οι τιμές TPM παρήχθησαν με το RNA-SeQC v1.1.9 (DeLuca et al., 2012), χρησιμοποιώντας τα παρακάτω φίλτρα σε επίπεδο ανάγνωσης:

- Οι αναγνώσεις χαρτογραφήθηκαν μονοσήμαντα (αντιστοίχιση σε ποιότητα χαρτογράφησης 255 για START BAMs).

- Οι αναγνώσεις στοιχήθηκαν σε κατάλληλα ζεύγη.
- Η απόσταση στοίχισης ανάγνωσης ήταν ≤ 6 (δηλ. οι στοιχίσεις δεν πρέπει να περιέχουν περισσότερες από έξι βάσεις μη-αναφοράς).
- Οι αναγνώσεις περιέχονται πλήρως εντός των ορίων του εξωνίου. Αναγνώσεις που επικαλύπτουν εσώνια δεν μετρήθηκαν.

Αυτά τα φίλτρα εφαρμόσθηκαν χρησιμοποιώντας την παράμετρο "-strictMode" στο RNA-SeQC. Οι τιμές TPM που μπορούν να μεταφορτωθούν δεν έχουν κανονικοποιηθεί ή διορθωθεί για όλες τις μεταβλητές.

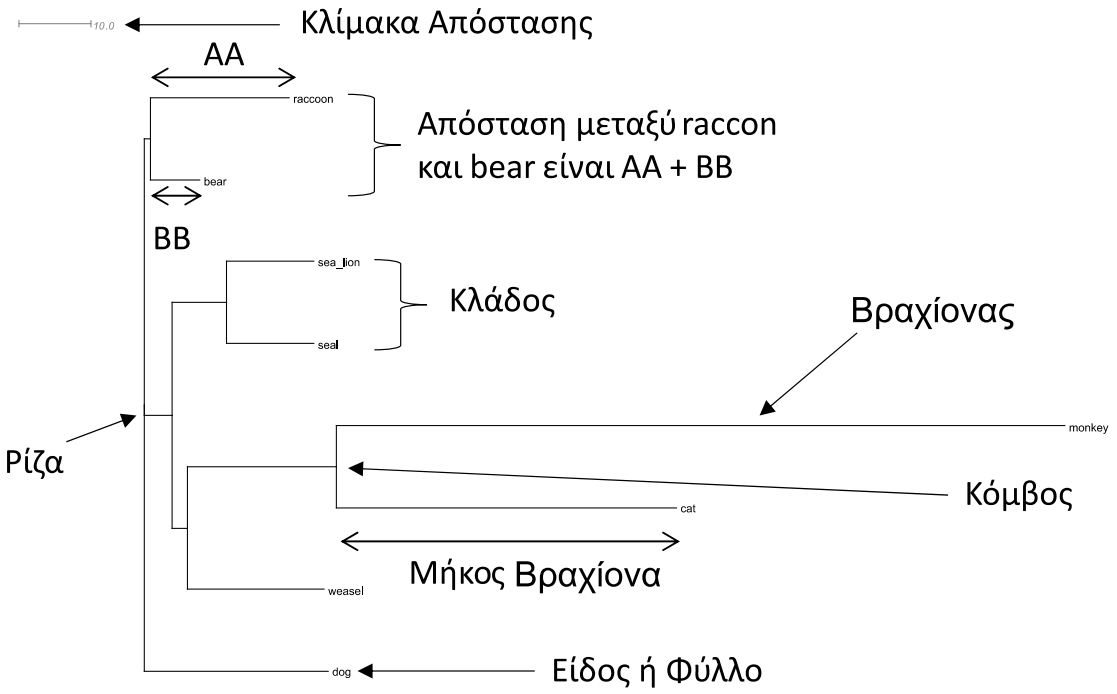
TPM: Η έκφραση γονιδίων και μεταγράφων στο GTEx portal παρουσιάζεται σε μετάγραφα ανά εκατομμύριο (Transcripts per Million), υπολογιζόμενα ως:

$$TPM = \frac{n_t}{\tilde{l}_t}\left(\sum_{k \in T}\frac{n_k}{\tilde{l}_k}\right)^{-1} \cdot 10^6$$

Όπου n_t είναι ο αριθμός των αναγνώσεων για το μετάγραφο/γονίδιο t, $\tilde{l}_t$ είναι το αποτελεσματικό μήκος μεταγράφου/γονιδίου (Li et al., 2010). Για γονίδια με μία ισομορφή, το $\tilde{l}_t$ είναι απλώς το μήκος αυτής της ισομορφής. Για εναλλακτικώς ματισμένα γονίδια, λαμβάνουμε το $\tilde{l}_t$ ως το μήκος της ένωσης όλων των γονιδιωματικών διαστημάτων που αντιστοιχούν σε εξώνια ισομορφών του γονιδίου t, όπως πραγματοποιείται στο πακέτο λογισμικού ERANGE (Mortazavi et al., 2008). Τ είναι το σύνολο των μεταγράφων/γονιδίων.

Φυλογενετικά δέντρα

Η φυλογενετική είναι κομμάτι της συστηματικής και προσπαθεί την περιγράψει την εξελικτική πορεία και τη σχέση μεταξύ οργανισμών ή ομάδων οργανισμών. Στην φυλογενετικές μελέτες, ο πιο βολικός τρόπος για να παρουσιαστούν οπτικά οι εξελικτικές σχέσεις ανάμεσα σε μία ομάδα οργανισμών, είναι μέσω απεικονίσεων που ονομάζονται φυλογενετικά δέντρα. Ένα φυλογενετικό δέντρο αποτελείται από κόμβους, που ο καθένας αντιπροσωπεύει μία ταξινομική μονάδα (είδη, πληθυσμοί, άτομα), και κλαδιά, τα οποία καθορίζουν τη σχέση μεταξύ των ταξινομικών μονάδων ως προς την καταγωγή και τους προγόνους τους (Εικόνα 6). Ένα κλαδί μπορεί να συνδέσει οποιουσδήποτε δύο γειτονικούς κόμβους. Το μοτίβο της διακλάδωσης του δέντρου ονομάζεται τοπολογία, και το μήκος των κλαδιών συνήθως αντιπροσωπεύει τον αριθμό των αλλαγών που έχουν επέλθει στον κλάδο. Αυτό ονομάζεται κλαδί υπό κλίμακα. Τα δέντρα υπό κλίμακα συχνά βαθμονομούνται ώστε να εκπροσωπούν το πέρασα του χρόνου. Η θεωρητική βάση τέτοιων δέντρων είναι ένα ή περισσότερα υπό ανάλυση γονίδια. Τα κλαδιά μπορούν επίσης να είναι χωρίς κλίμακα, που σημαίνει ότι το μήκος των κλαδιών δεν είναι ανάλογο με τον αριθμό των αλλαγών που έχουν συμβεί, αν και ο πραγματικός αριθμός μπορεί να αναφέρεται αριθμητικά κάπου στο κλαδί. Τα φυλογενετικά δέντρα μπορούν επίσης να είναι με ή χωρίς ρίζα. Στα δέντρα με ρίζα, υπάρχει ένας συγκεκριμένος κόμβος, που ονομάζεται ρίζα, που αντιπροσωπεύει έναν κοινό πρόγονο, από την οποία ένα μοναδικό μονοπάτι οδηγεί σε οποιονδήποτε άλλο κόμβο του δέντρου. Ένα δέντρο χωρίς ρίζα καθορίζει μόνο τη σχέση μεταξύ των ειδών, χωρίς να προσδιορίζει έναν κοινό πρόγονο, ή εξελικτική πορεία.



Εικόνα 6 – Δομή ενός απλού φυλογενετικού δέντρου

Φυλογενετική ανάλυση

Φυλογενετική ανάλυση ονομάζεται η διαδικασία εκτίμησης των εξελικτικών σχέσεων των οργανισμών, μέσα από τη μελέτη των αντίστοιχων βιολογικών αλληλουχιών τους (Bagos, 2015). Βέβαια, πέρα από την επιστήμη της φυλογενετικής των ειδών, είναι δυνατόν να δημιουργηθούν φυλογενετικά δέντρα και σε άλλες περιπτώσεις, όπως μεταξύ πρωτεϊνικών αλληλουχιών ή γονιδίων, με βάση την ομοιότητα στην έκφρασή τους. Υπάρχουν πολλοί αλγόριθμοι, προγράμματα και μέθοδοι για την εκτέλεση μίας φυλογενετικής ανάλυσης. Λαμβάνοντας υπ' όψη το είδος και τον όγκο των διαθέσιμων δεδομένων, π.χ. νουκλεοτιδικές ακολουθίες, πίνακες αποστάσεων κλπ., καθώς και το αντίστοιχο βιολογικό ερώτημα που θέλει ο κάθε ερευνητής να απαντήσει, υπάρχει μεγάλος όγκος μεθόδων προς χρήση Τα πλεονεκτήματα και τα μειονεκτήματα αυτών των διαφόρων μεθόδων είναι αντικείμενα πολλών επιστημονικών διαφωνιών, γιατί ο κίνδυνος της δημιουργίας λανθασμένων αποτελεσμάτων είναι μμεγαλύτερος στην υπολογιστική μοριακή φυλογενετική από ότι σε πολλούς άλλους τομείς της επιστήμης. Οι βασικές αρχές της φυλογενετικής ανάλυσης, μπορούμε να πούμε ότι στηρίζονται σε μερικές απλές παραδοχές (Brinkman and Leipe, 2001):

- Οποιαδήποτε ομάδα οργανισμών (ή αλληλουχιών) προέρχεται από κάποιον κοινό πρόγονο μέσω της εξέλιξης. Αν οι οργανισμοί (ή οι αλληλουχίες) είναι πολύ διαφορετικοί, ο κοινός πρόγονος υπάρχει αλλά βρίσκεται πολύ πίσω στον εξελικτικό χρόνο.
- Υπάρχει διχαλωτό πρότυπο στην εξέλιξη. Η διαδικασία της εξέλιξης οδηγεί πάντα σε διχοτόμηση ενός είδους ή μίας αλληλουχίας, έτσι ώστε να δημιουργούνται δύο βραχίονες κάτω από έναν κόμβο.
- Αλλαγή στα παρατηρήσιμα χαρακτηριστικά των οργανισμών εμφανίζεται μετά το πέρασμα πολλών γενιών.

Μία φυλογενετική ανάλυση και κατασκευή δέντρου αποτελείται από τέσσερα διακριτά σημεία:

- Μία πολλαπλή στοίχιση. Από αυτήν ξεκινάνε όλα, και όλα βασίζονται σε αυτή. Αν η αρχική στοίχιση είναι λάθος, όλες οι παρακάτω αναλύσεις θα είναι επισφαλείς. Γι' αυτό, πολλές φορές χρειάζεται εμπειρία και δια χειρός επεξεργασία.
- Καθορισμός του μοντέλου αντικατάστασης, δηλαδή του μαθηματικού μοντέλου της εξελικτικής αλλαγής. Αυτή είναι μια απαίτηση των περισσότερων μεθόδων, και χρειάζεται ιδιαίτερη προσοχή, καθώς ένα απλό μοντέλο μπορεί να κάνει εύκολους τους υπολογισμούς αλλά μπορεί να μην είναι ρεαλιστικό.
- Κατασκευή του δέντρου. Σε αυτό το σημείο, υπάρχουν οι βασικότερες διαφοροποιήσεις των αλγορίθμων. Κάποιες μέθοδοι είναι γρήγορες, άλλες πιο χρονοβόρες, άλλες κάνουν περισσότερες υποθέσεις κ.ο.κ.
- Αξιολόγηση του δέντρου. Αφού το δέντρο κατασκευαστεί, πρέπει να υπάρχει και ένας τρόπος να υπολογιστεί η αξιοπιστία του. Ανάλογα με τη μέθοδο κατασκευής, και με το χρησιμοποιούμενο λογισμικό, μπορεί να υπάρχουν και διαφορετικοί τρόποι ελέγχου της αξιοπιστίας του δέντρου.

Μέθοδοι κατασκευής δέντρων

Οι μέθοδοι κατασκευής φυλογενετικών δέντρων κατατάσσονται σε δύο κατηγορίες, τις μεθόδους βασισμένες στην απόσταση και στις μεθόδους βασισμένες στους χαρακτήρες. Στην παρούσα εργασία, μπορούν να

χρησιμοποιηθούν μόνο οι μέθοδοι της απόστασης, καθώς η συσχέτισης μεταξύ δειγμάτων ή μεταξύ γονιδίων γίνεται με τη χρήση του συντελεστή συσχέτισης Pearson, που εν τέλει παράγει έναν πίνακα αποστάσεων που θα χρησιμοποιηθεί από τις αντίστοιχες μεθόδους. Παρόλα αυτά αναφέρουμε και τις μεθόδους των χαρακτήρων.

Μέθοδοι βασισμένες στην απόσταση

Οι μέθοδοι αυτοί υπολογίζουν έναν πίνακα αποστάσεων για όλα τα ζευγάρια των αντικειμένων (είδη, αλληλουχίες, γονίδια, κλπ.) και στη συνέχεια κατασκευάζουν το δέντρο με βάση αυτόν. Έχουμε τις εξής μεθόδους

- UPGMA

Η UPGMA (Unweighted Pair Group Method with Arithmetic mean ή Μέθοδος Ομαδοποίησης Αστάθμητων Ζευγών με Αριθμητικούς Μέσους Όρους) ορίζει την απόσταση μεταξύ δυο ομάδων (Clusters) να είναι η μέση απόσταση μεταξύ ζευγών (Sokal and Michener, 1958). Ανήκει στην κατηγορία των αλγορίθμων ιεραρχικής ομαδοποίησης, μία μέθοδος στατιστικής ομαδοποίησης που προσπαθεί να δημιουργήσει μία ιεραρχία από ομάδες, με προσέγγιση από «κάτω προς τα πάνω», δηλαδή ξεκινάει δημιουργώντας ομάδες για κάθε παρατήρηση (ζεύγος αντικειμένων/γονίδια) οι οποίες ενώνονται όσο ανεβαίνουμε στην ιεραρχία, πάντα με βάση την απόσταση των αντικειμένων μεταξύ τους. Αρχικά υπολογίζεται η ομάδα με το ζεύγος αντικειμένων που έχουν την μικρότερη απόσταση και υπολογίζεται ένας καινούργιος πίνακας αποστάσεων με βάση την κοινή απόσταση της ομάδας από τα άλλα φύλλα. Αυτή η διαδικασία συνεχίζεται μέχρι το τέλος. Ο αλγόριθμος κατασκευάζει δέντρα με ρίζα σε χρόνο της τάξης του $O(n^2)$. Η συγκεκριμένη έκδοση της μεθόδου UPGMA ονομάζεται Average Linkage λόγω του γεγονότος ότι επιλέγεται η μέση απόσταση μεταξύ των ζευγών αντικειμένων. Αντίστοιχα έχουμε: την Simple Linkage όπου επιλέγεται η μικρότερη απόσταση εκ των δύο αντικειμένων του ζεύγους και την Complete Linkage όπου επιλέγεται η μεγαλύτερη. Επιπλέον, η UPGMA παράγει εξ ’ορισμού δέντρα στα οποία ισχύει η προσθετική ιδιότητα. Με αυτό, εννοούμε δέντρα στα οποία η απόσταση δύο οποιονδήποτε άκρων, είναι ίση με το άθροισμα των μηκών των ακμών που τα συνδέουν. Τέλος η απόσταση όλων

των φύλλων από τη ρίζα είναι η ίδια, πράγμα που οφείλεται στην υπόθεση του σταθερού ρυθμού εξελικτικών αλλαγών, η οποία είναι γνωστή ως το «μοριακό ρολόι».

- WPGMA

Η WPGMA (Weighted Pair Group Method with Arithmetic mean ή Μέθοδος Ομαδοποίησης Σταθμισμένων Ζευγών με Αριθμητικούς Μέσους Όρους) είναι ίδια με τη μέθοδο UPGMA, με τη μόνη διαφορά ότι παράγει δέντρα με βάρη.

- Neighbor-Join

Η μέθοδος ένωσης γειτόνων (Neighbor Join - NJ) (Saitou and Nei, 1987) κατατάσσει τα αντικείμενα σε ζευγάρια, χρησιμοποιώντας μία τροποποιημένη απόσταση, για να βρει τελικά τα πιο «γειτονικά». Στη συνέχεια, θα ενώσει τα δύο επόμενα κλαδιά με αντίστοιχο τρόπο. Η NJ παράγει φυλογενετικά δέντρα χωρίς ρίζα αλλά για την εύρεση αυτής μπορούμε να χρησιμοποιήσουμε σαν σημείο αναφοράς ένα «μακρινό» αντικείμενο (εξωομάδα - outgroup) το οποίο ξέρουμε ότι βρίσκεται πολύ μακριά σε απόσταση από όσα εξετάσαμε. Η μέθοδος είναι πολύ γρήγορη, και αυτό την κάνει ελκυστική, ειδικά για αναλύσεις μεγάλων συνόλων δεδομένων ή για εφαρμογή στατιστικών τεχνικών όπως το bootsrap. Σε σύγκριση με την UPGMA, είναι πιο αργή, αλλά αυτό αντισταθμίζεται από την χαλάρωση της απαίτησης του «μοριακού ρολογιού».

- Fitch-Margoliash

Η μέθοδος Fitch-Margoliash (Fitch and Margoliash, 1967) βασίζεται στη στατιστική τεχνική της γραμμικής παλινδρόμησης, δηλαδή, της ευθείας ελαχίστων τετραγώνων. Πρακτικά, αναζητούμε το δέντρο, τα μήκη των βραχιόνων του οποίου θα έχουν τη μικρότερη τετραγωνική απόκλιση των αποστάσεων, σε σχέση με όλα τα πιθανά μήκη μονοπατιών. Στις στενά συσχετιζόμενες ακολουθίες δίνεται μεγαλύτερο βάρος στην διαδικασία κατασκευής του δέντρου ώστε να καλυφθεί η αυξημένη αστοχία στον υπολογισμό αποστάσεων μεταξύ μακρινών ακολουθιών.

Μέθοδοι βασισμένες στους χαρακτήρες

Οι μέθοδοι που βασίζονται στους χαρακτήρες (character-based methods), σε αντίθεση με τις μεθόδους αποστάσεων, δεν μετασχηματίζουν τις αλληλουχίες, αλλά τις χρησιμοποιούν σε όλη τη διαδικασία της εκτίμησης, αντιμετωπίζοντας αυτές, όπως ακριβώς είναι: ακολουθίες διακριτών συμβόλων από ένα πεπερασμένο αλφάβητο, π.χ. στις νουκλεοτιδικές ακολουθίες DNA κάθε νουκλεοτίδιο είναι μία κατάσταση. Το φυλογενετικό δέντρο προκύπτει από την εξέταση των εξελικτικών σχέσεων των ακολουθιών του DNA σε κάθε νουκλεοτιδική θέση. Χωρίζονται σε δύο μεγάλες κατηγορίες:

- Μέγιστης φειδωλότητας (Maximum Parsimony)

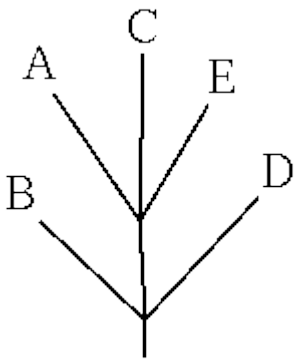
Οι μέθοδοι μέγιστης φειδωλότητας κάνουν διάκριση μεταξύ πληροφοριακών και μη-πληροφοριακών θέσεων στις αλληλουχίες, με τις πληροφοριακές θέσεις να είναι αυτές που παρουσιάζουν πολυμορφισμό (ύπαρξη πάνω από δυο ειδών νουκλεοτιδίων) τουλάχιστον δυο φορές. Εφαρμόζεται στην εξελικτική βιολογία προτού να εμφανιστεί η μοριακή φυλογένεια και έχει σκοπό να εξηγήσει τις εξελικτικές διάφορες με το μικρότερο αριθμό αλλαγών. Η μέθοδος γίνεται όλο και πιο ανακριβής όσο αυξάνεται ο αριθμός φύλλων ενός δέντρου. Σε περιπτώσεις με πάνω από 20 φύλλα πλέον αναφερόμαστε σε ευριστική αναζήτηση.

- Μέγιστης πιθανοφάνειας (Maximum Likelihood)

Οι μέθοδοι μέγιστης πιθανοφάνειας, χρησιμοποιούν ένα ξεκάθαρο μαθηματικό μοντέλο για την εξέλιξη. Ο υπολογισμός της πιθανοφάνειας γίνεται με κάποιον αλγόριθμο ο οποίος αθροίζει τις συνεισφορές για όλα τα πιθανά δέντρα, και ο πιο γνωστός αλγόριθμος που έχει προταθεί για αυτό το σκοπό, είναι αυτός του Felsenstein (Felsenstein, 1981). Αν και έχουν προταθεί πολλές παραλλαγές, σε γενικές γραμμές η διαδικασία που ακολουθείται για να υπολογιστεί η πιθανοφάνεια είναι η εξής: Ο αλγόριθμος εντοπίζει ένα πιθανό δέντρο και οι παράμετροι για αυτό το δέντρο αλλάζουν λίγο-λίγο με κάποια από τις επαναληπτικές διαδικασίες που αναφέραμε παραπάνω έως ότου βρεθεί το βέλτιστο δέντρο. Η διαδικασία αυτή επαναλαμβάνεται για όλα τα πιθανά δέντρα και αυτό που δίνει τη μέγιστη πιθανοφάνεια, από όλα τα δέντρα, επιλέγεται τελικά ως το δέντρο μέγιστης πιθανοφάνειας.

Μορφοποίηση αρχείων δέντρων Newick

Το πρότυπο Newick (Archie et al., 2008)για την αναπαράσταση δέντρων σε μορφή αναγνώσιμη από ηλεκτρονικούς υπολογιστές, κάνει χρήση της αντιστοιχίας ανάμεσα σε δέντρα και σε εμφωλευμένες παρενθέσεις. Υιοθετήθηκε στις 26 Ιουνίου 1986 σε μία άτυπη συνεδρίαση της επιτροπής κατά τη διάρκεια της συνεύρεσης της Κοινωνίας για τη Μελέτη της Εξέλιξης (Society for the Study of Evolution) στο Durham, New Hampshire που αποτελούταν από τους James Archie, William HE Day, Wayne Maddison, Christopher Meacham, James F. Rohlf, David Swofford και Joseph Felsenstein. Ο λόγος για το όνομα είναι ότι στη δεύτερη και τελική συνεδρίαση της επιτροπής, συναντήθηκαν στο εστιατόριο “Newick’s” στο Dover του New Hampshire. Η αναπαράσταση δέντρων ήταν μία γενίκευση που αναπτύχθηκε από τον Christopher Meacham το 1984 για τα προγράμματα σχεδίασης δέντρων που έγραψε για το πακέτο PHYLIP (Felsenstein, 2008) κατά την επίσκεψή του στο Seattle. Το πρότυπο Newick περιεγράφηκε το 1857 από τον διάσημο Άγγλο μαθηματικό Arthur Cayley. Αυτό το δέντρο με ρίζα:



αντιπροσωπεύεται από την ακόλουθη ακολουθία χαρακτήρων: (B, (A, C, E), D); Το δέντρο τελειώνει με ένα ελληνικό ερωτηματικό. Ο κατώτατος κόμβος σε αυτό το δέντρο είναι ένας εσωτερικός κόμβος και όχι μία ακμή (φύλλο). Οι εσωτερικοί κόμβοι αντιπροσωπεύονται από ένα ζεύγος ταιριασμένων παρενθέσεων. Μεταξύ των παρενθέσεων αναπαριστώνται οι κόμβοι που κατάγονται άμεσα από έναν κόμβο, χωρισμένοι με κόμματα. Στο παραπάνω δέντρο, οι άμεσοι απόγονοι είναι ο B, ένας ακόμη εσωτερικός κόμβος, και ο D. Ο άλλος εσωτερικός κόμβος αντιπροσωπεύεται από ένα ζεύγος παρενθέσεων, εσωκλείοντας τις αναπαραστάσεις των άμεσων απογόνων του, A, C, και E. Στο

παράδειγμά μας, αυτοί συμβαίνουν να είναι άκρες (φύλλα), αλλά σε γενικές γραμμές θα μπορούσαν επίσης να είναι εσωτερικοί κόμβοι και το αποτέλεσμα θα ήταν περαιτέρω εμφωλευμένες παρενθέσεις, σε οποιοδήποτε επίπεδο. Οι άκρες (φύλλα) αντιπροσωπεύονται από τα ονόματά τους. Ένα όνομα μπορεί να είναι οποιαδήποτε σειρά χαρακτήρων εκτός από κενά, άνω-κάτω τελείες, ερωτηματικά, παρενθέσεις και αγκύλες. Επειδή αν θέλουμε να συμπεριληφθεί ένα κενό σε ένα όνομα, θεωρείται ότι ένας χαρακτήρας υπογράμμισης ("_") σημαίνει ένα κενό: οποιοσδήποτε από τους χαρακτήρες υπογράμμισης σε ένα όνομα, θα μετατραπεί σε ένα κενό, όταν διαβάζεται. Κάθε όνομα μπορεί επίσης να είναι κενό: ένα δέντρο, όπως το παρακάτω είναι επιτρεπτό. (, (, ,) ,) ; Τα δέντρα μπορούν να διακλαδίζονται σε οποιοδήποτε επίπεδο. Τα μήκη των κλαδιών μπορούν να ενσωματωθούν σε ένα δέντρο βάζοντας μία άνω-κάτω τελεία και έναν πραγματικό αριθμό, με ή χωρίς υποδιαστολή, μετά από το κόμβο. Αυτό αντιπροσωπεύει το μήκος του κλάδου αμέσως κάτω από αυτό τον κόμβο. Έτσι, το παραπάνω δέντρο θα μπορούσε να έχει μήκη που παριστάνονται ως: (B:6.0,(A:5.0,C:3.0,E:4.0):5.0,D:11.0); Το δέντρο ξεκινά από την πρώτη γραμμή του αρχείου, και μπορεί να συνεχιστεί στις επόμενες γραμμές. Είναι καλύτερο να προχωρούμε σε μία νέα γραμμή αμέσως μετά από ένα κόμμα. Τα κενά μπορούν να εισαχθούν σε οποιοδήποτε σημείο εκτός από την μέση του ονόματος ενός είδους ή του μήκους ενός κλάδου. Η παραπάνω περιγραφή είναι στην πραγματικότητα ένα υποσύνολο του προτύπου Newick. Για παράδειγμα, οι εσωτερικό κόμβοι μπορούν να έχουν ονόματα στο εν λόγω πρότυπο. Τα ονόματα αυτά ακολουθούν την δεξιά παρένθεση για αυτόν τον εσωτερικό κόμβο, όπως σε αυτό το παράδειγμα: (B:6.0,(A:5.0,C:3.0,E:4.0)Ancestor1:5.0,D:11.0). Ως παράδειγμα, η Newick μορφοποίηση του δέντρου που απεικονίζεται στην Εικόνα 6 είναι:

```
((raccoon:19.19959,bear:6.80041):0.84600,((sea_lion:11.99
700,
seal:12.00300):7.52973,((monkey:100.85930,cat:47.14069):2
0.59201, weasel:18.87953):2.09460):3.87382,dog:25.46154);
```

Μέθοδοι

Τεχνικές λεπτομέρειες υπολογιστή προγραμματιστικής ανάπτυξης

Ως βάση της ανάπτυξης των διαδικτυακών εργαλείων χρησιμοποιήθηκε ένας σταθερός υπολογιστής με λειτουργικό σύστημα Linux Kubuntu 18.04. Ο υπολογιστής διαθέτει 64GB RAM και 12 πυρήνες στην κεντρική μονάδα επεξεργασίας. Το λειτουργικό σύστημα και οι βασικές εφαρμογές που αποτελούν μέρος του και περιγράφονται παρακάτω, είναι εγκατεστημένα σε σκληρό δίσκο τεχνολογίας SSD που επιτρέπει γρήγορες ταχύτητες στην προσπέλαση στα δεδομένα του δίσκου, άρα και γρηγορότερη εκτέλεση των προγραμμάτων. Όλα τα παραγόμενα δεδομένα μεγάλου όγκου αποθηκεύονται σε εσωτερικούς και εξωτερικούς δίσκους τεχνολογίας HDD μεγάλης χωρητικότητας.

Apache Server

Το Apache HTTP Server (https://httpd.apache.org/) είναι ένας διακομιστής HTTP ανοιχτού κώδικα για σύγχρονα λειτουργικά συστήματα, συμπεριλαμβανομένων των UNIX και των Windows. Είναι ένας ασφαλής, αποδοτικός και επεκτάσιμος διακομιστής που παρέχει υπηρεσίες HTTP με βάση τα τρέχοντα πρότυπα. Για την εγκατάστασή του αρκεί η εντολή root:

```
apt-get install apache2
```

Ο εξυπηρετητής ανοίγει με την εντολή root:

```
service apache2 start
```

Σχεσιακό Μοντέλο Βάσεων Δεδομένων

Υπάρχουν τέσσερα μοντέλα βάσεων δεδομένων: το ιεραρχικό, το δικτυακό, το σχεσιακό και το αντικειμενοστραφές. Σήμερα στις περισσότερες βάσεις κυριαρχεί το σχεσιακό μοντέλο. Η απλή και κατανοητή δομή του είναι δύο από τα αίτια της μεγάλης απήχησής του, τόσο σε κλασσικές, όσο και σε σύγχρονες εφαρμογές υψηλών απαιτήσεων. Οι βασικοί στόχοι του σχεσιακού μοντέλου είναι οι εξής:

- Η υποστήριξη της ανεξαρτησίας των δεδομένων, έτσι ώστε αλλαγές στη φυσική δομή και οργάνωση της βάσης δεδομένων να μην απαιτούν αλλαγές στα προγράμματα εφαρμογής.
- Η αποφυγή του πλεονασμού που εμφανίζεται όταν τα ίδια δεδομένα αποθηκεύονται πολλές φορές σε διαφορετικές περιοχές της βάσης δεδομένων.
- Η διατήρηση της ακεραιότητας και της συνέπειας των δεδομένων.
- Η υποστήριξη της ανάπτυξης γλωσσών χειρισμού δεδομένων, οι οποίες διευκολύνουν τη διατύπωση ερωτημάτων προς το Σύστημα Διαχείρισης Βάσεων Δεδομένων

Εδώ περιλαμβάνονται βάσεις δεδομένων με δυνατότητες ταυτόχρονου χειρισμού και σύνδεσης πολλών συλλογών από εγγραφές διαφορετικών τύπων, οργανωμένες σε πίνακες. Ο όρος «σχεσιακό» αναφέρεται σε μία συγκεκριμένη μέθοδο οργάνωσης των βάσεων δεδομένων, σύμφωνα με την οποία οι πίνακες της βάσης δεδομένων μπορούν να συσχετισθούν μεταξύ τους με αποτελέσματα οι πληροφορίες να κατανέμονται ομοιόμορφα σε όλη τη βάση. Για την αναπαράσταση δεδομένων, το σχεσιακό μοντέλο χρησιμοποιεί πίνακες. Ο κάθε πίνακας έχει ένα μοναδικό όνομα και προσδιορίζεται από ένα σύνολο γραμμών και ένα σύνολο στηλών. Οι στήλες του πίνακα ορίζουν τα χαρακτηριστικά της κάθε εγγραφής. Κάθε γραμμή του πίνακα αναπαριστά μία εγγραφή δεδομένων και ονομάζεται πλειάδα. Το πλήθος των χαρακτηριστικών της σχέσης καλείται βαθμός (degree), ενώ ο αριθμός των πλειάδων καλείται πληθικότητα. Για κάθε χαρακτηριστικό υπάρχει ένα σύνολο επιτρεπτών τιμών, το οποίο καλείται πεδίο ορισμού του χαρακτηριστικού. Οι τιμές που μπορεί να πάρει ένα χαρακτηριστικό προσδιορίζονται από το αντίστοιχο πεδίο ορισμού, ενώ επίσης είναι απαραίτητο να γνωρίζουμε τον τύπο δεδομένων (data type) και τη μορφοποίηση (format).

Οι βασικότερες ιδιότητες των πινάκων είναι οι εξής:

- Κάθε πίνακας της βάσης δεδομένων έχει ένα μοναδικό όνομα.
- Η τιμή κάθε χαρακτηριστικού σε κάθε πλειάδα είναι μία.
- Το κάθε χαρακτηριστικό έχει μοναδικό όνομα μέσα στον πίνακα.

- Δύο χαρακτηρίστηκα που ανήκουν σε διαφορετικούς πίνακες επιτρέπεται να έχουν ίδιο όνομα.
- Όλες οι τιμές ενός χαρακτηριστικού πρέπει να ανήκουν στο πεδίο ορισμού του χαρακτηριστικού.
- Η σειρά δήλωσης των χαρακτηριστικών ενός πίνακα δεν παίζει κανένα ρόλο.
- Δύο πλειάδες μίας σχέσης δεν επιτρέπεται να ταυτίζονται σε όλα τα χαρακτηριστικά.
- Στο σχεσιακό μοντέλο δεν μας ενδιαφέρει η σειρά των πλειάδων στον πίνακα. Ωστόσο, η σειρά αποθήκευσης των δεδομένων συνήθως επηρεάζει το χρόνο επεξεργασίας και επομένως λαμβάνεται υπόψη.

Ένα χαρακτηριστικό που διακρίνει τις διάφορες γραμμές του πίνακα (πλειάδες), ονομάζεται πρωτεύον κλειδί. Σε πολλές περιπτώσεις απαιτούνται περισσότερα από ένα χαρακτηριστικά για να συνθέσουν ένα κλειδί, οπότε το κλειδί καλείται σύνθετο. Τα υπόλοιπα χαρακτηριστικά του πίνακα λέγονται δευτερεύοντα κλειδιά. Η συνέπεια των δεδομένων μετά από εισαγωγές, διαγραφές και ενημερώσεις, διατηρείται με τη χρήση περιορισμών ακεραιότητας. Οι σημαντικότεροι είναι:

- Περιορισμός χρήσης κενών τιμών: Υπάρχουν περιπτώσεις όπου δεν γνωρίζουμε την τιμή ενός χαρακτηριστικού ή δεν μπορούμε να την προσδιορίσουμε. Σε αυτές τις περιπτώσεις αποδίδουμε στο χαρακτηριστικό την κενή τιμή (NULL).
- Περιορισμοί ακεραιότητας οντοτήτων: Είναι στενά συνδεδεμένη με την έννοια του κλειδιού μίας σχέσης. Κάθε γραμμή πρέπει να προσδιορίζεται μοναδικά τουλάχιστον από το πρωτεύον κλειδί του πίνακα.
- Περιορισμοί αναφορών: Αν το κλειδί k ενός πίνακα Α εξάγεται ως χαρακτηριστικό σε έναν άλλο πίνακα τότε λέγεται ότι το k αποτελεί ξένο κλειδί για τον δεύτερο πίνακα. Η ακεραιότητα αναφορών επιβάλλει η τιμή ενός ξένου κλειδιού να είναι η ίδια με αυτήν του αρχικού πίνακα.

Οι πίνακες συνδέονται μέσω των εξής κατευθυνόμενων συσχετίσεων: μία εγγραφή ενός πίνακα συσχετίζεται με μία και μόνο μία εγγραφή ενός άλλου πίνακα (1:1), μία εγγραφή ενός πίνακα συσχετίζεται με πολλές εγγραφές ενός

άλλου πίνακα αλλά και κάθε εγγραφή του δεύτερου πίνακα συσχετίζεται με μία και μόνο μία εγγραφή του πρώτου πίνακα (1:N) και σε μία εγγραφή ενός πίνακα αντιστοιχούν πολλές εγγραφές ενός άλλου πίνακα και σε κάθε εγγραφή του δεύτερου πίνακα αντιστοιχούν πολλές εγγραφές του πρώτου πίνακα (M:N). Υπάρχουν επίσης προαιρετικές προϋποθέσεις συμμετοχής των πινάκων (όπου ένας πίνακας δεν χρειάζεται να συσχετίζεται με κανένα άλλο πίνακα). Τα πρωτεύοντα κλειδιά του πίνακα από τον οποίο ξεκινάει η συσχέτιση εξάγονται ως ξένα κλειδιά στον πίνακα που καταλήγει η συσχέτιση. Εάν τα ξένα κλειδιά γίνονται πρωτεύοντα κλειδιά στον καταλήγοντα πίνακα, η συσχέτιση ονομάζεται ταυτοποιούσα, ενώ εάν γίνονται δευτερεύοντα, η συσχέτιση ονομάζεται μη ταυτοποιούσα.

MySQL Server

Η MySQL (https://www.mysql.com/) είναι ένα σύστημα διαχείρισης σχεσιακών βάσεων δεδομένων ανοιχτού κώδικα, την οποία εκμεταλλευτήκαμε για να αποθηκεύσουμε τα δεδομένα μας. Όπως και με τον Apache, χρειάστηκε η λήψη και η εγκατάσταση με την εντολή root:

```
apt-get install mysql-server mysql-client
```

Κατά την εγκατάσταση, μας ζητείται ένας κωδικός που θα επιτρέπει την είσοδο στον διαχειριστή (root) της βάσης. Μετά το τέλος της εγκατάστασης, η ενεργοποίηση της υπηρεσίας γίνεται με την εντολή root:

```
service mysql start
```

Για την είσοδο στο πρόγραμμα διαχείρισης βάσεων της MySQL χρειάζεται η εντολή:

```
mysql –u <user> <database> –p
```

όπου `user` είναι το όνομα λογαριασμού του χρήστη, όπως π.χ. `root`, και `database` το όνομα της βάσης δεδομένων που θέλουμε να συνδεθούμε. Στη συνέχεια, ζητείται η πληκτρολόγηση του μυστικού κωδικού που αντιστοιχεί στον εν λόγω χρήστη. Πλέον, έχει γίνει η σύνδεση στο τοπικό σύστημα της MySQL. Η έξοδος γίνεται πατώντας ^D. Η διαχείριση βάσεων δεδομένων MySQL καθίσταται δυνατή με την εκτέλεση εντολών στη Γλώσσα Δομημένων Ερωτημάτων (Structured Query Language - SQL). Επιτρέπεται η δημιουργία

μίας βάσης δεδομένων, η πρόσθεση, τροποποίηση και διαγραφή πινάκων και εγγραφών, καθώς και η αναζήτηση εγγραφών με συγκεκριμένα κριτήρια.

Δημιουργία των Βάσεων Δεδομένων

Για την σχεδίαση και απεικόνιση των βάσεων δεδομένων των προγραμμάτων ACT και HGCA χρησιμοποιήθηκε το πρόγραμμα MySQL Workbench (https://www.mysql.com/products/workbench/), ένα φιλικό προς το χρήστη εργαλείο σχεδιασμού και διαχείρισης σχεσιακών βάσεων δεδομένων, το οποίο διατίθεται δωρεάν από την ίδια τη MySQL. Με το MySQL Workbench, είναι δυνατή η σχεδίαση διαγραμμάτων Οντοτήτων-Συσχετίσεων (Entity Relation Diagram - ERD) που περιέχουν όλους τους πίνακες της βάσης δεδομένων και τις μεταξύ τους συσχετίσεις. Επιπλέον, είναι δυνατή η προσθαφαίρεση πινάκων και χαρακτηριστικών σε αυτούς. Σε κάθε χαρακτηριστικό ορίζεται το όνομα και ο τύπος του.

Το Διάγραμμα είναι ένα μοντέλο δεδομένων για την περιγραφή μιας βάσης δεδομένων με έναν περιληπτικό τρόπο. Το σύστημα χαρακτήρων και συμβόλων που χρησιμοποιείται κατά κόρον από τα εργαλεία δημιουργίας Διαγραμμάτων Οντοτήτων-Συσχετίσεων ονομάζεται Crow’s foot. Οι πίνακες αντιπροσωπεύονται ως κουτιά και οι συσχετίσεις με γραμμές μεταξύ των κουτιών. Τα διαφορετικά σχήματα στα άκρα αυτών των γραμμών αντιπροσωπεύουν τους τύπους των συσχετίσεων. Οι συνεχείς γραμμές αναπαριστούν τις ταυτοποιούσες συσχετίσεις ενώ, οι διακεκομμένες γραμμές τις μη ταυτοποιούσες.

Μία λειτουργία του MySQL Workbench είναι η δυνατότητα παραγωγής εντολών SQL δημιουργίας των πινάκων του διαγράμματος, που επικολλόνται στην γραμμή εντολών της MySQL.

Επιπλέον, είναι δυνατή και η αντίστροφη λειτουργία. Το MySQL Workbench μπορεί να δημιουργήσει το διάγραμμα Οντοτήτων-Συσχετίσεων μίας βάσης δεδομένων, συνδεόμενο στην εν λόγω βάση. Η μετατροπή δεν είναι πάντα τέλεια, οπότε όποια σφάλματα εμφανιστούν, θα πρέπει να διορθωθούν δια χειρός.

Δημιουργήθηκαν δύο βάσεις δεδομένων, μία για κάθε εργαλείο (με όνομα ACT και GTEx, αντίστοιχα). Η ενέργεια γίνεται με την εντολή mysql root:

```
CREATE DATABASE <databasename>;
```

όπου `databasename` είναι το όνομα της βάσης δεδομένων.

Στη συνέχεια πρέπει να δωθούν τα απαραίτητα δικαιώματα στον χρήστη με την εντολή mysql root:

```
GRANT ALL PRIVILEGES ON <databasename>.* TO
'<user>'@'localhost' IDENTIFIED BY '<password>'
```

όπου `databasename` είναι το όνομα της βάσης δεδομένων, `user` είναι το όνομα λογαριασμού του χρήστη και `password` είναι ο μυστικός κωδικός του χρήστη της συγκεκριμένης βάσης δεδομένων.

Επόμενο βήμα είναι η δημιουργία των πινάκων, όπως περιγράφονται στο διάγραμμα ERD. Για παράδειγμα, ο πίνακας Gene του HGCA δημιουργείται ως εξής:

```
CREATE TABLE IF NOT EXISTS `GTEx`.`Gene` (
  `ENSG` VARCHAR(40) NOT NULL,
  `HGNC` VARCHAR(45) NULL,
  `Description` VARCHAR(250) NULL,
  PRIMARY KEY (`ENSG`),
  CONSTRAINT `fk_Gene_Expression1`
    FOREIGN KEY (`ENSG`)
    REFERENCES `GTEx`.`Expression` (`ENSG`)
    ON DELETE NO ACTION
    ON UPDATE NO ACTION)
ENGINE = InnoDB;
```

Για την εισαγωγή δεδομένων από αρχείο μέσα στη βάση, πρέπει η MySQL να λειτουργεί με προσθέτοντας την επιλογή `--local-infile` στη τέλος της εντολής εισόδου στη MySQL:

```
mysql –u <user> –p <databasename> --local-infile
```

όπου `databasename` είναι το όνομα της βάσης δεδομένων και `user` είναι το όνομα λογαριασμού του χρήστη.

Η εισαγωγή των δεδομένων γίνεται πληκτρολογώντας στη γραμμή εντολών της MySQL την εντολή:

```
LOAD DATA LOCAL INFILE '<filename>' INTO TABLE
<table>;
```

όπου `filename` είναι το αρχείο προς εισαγωγή και `table` ο πίνακας στον οποίο θα εισαχθούν τα δεδομένα. Για να γίνει επιτυχώς η εισαγωγή, το αρχείο εισόδου θα πρέπει να είναι σε στηλοθετημένη με tab μορφή (flat files), όπου κάθε γραμμή αντιστοιχεί σε μία μοναδική εγγραφή του πίνακα.

PHP

Η γλώσσα προγραμματισμού που επιλέχθηκε για την ανάπτυξη της εφαρμογής είναι η ανοιχτού κώδικα γενικής χρήσης scripting γλώσσα PHP (http://php.net/). Η εγκατάστασή της στο περιβάλλον των Linux είναι εξίσου απλή, χρησιμοποιώντας την εντολή root:

```
apt-get install php-cli libapache2-mod-php php-mcrypt php-mysql
```

Μετά το πέρας της εγκατάστασης, η εκτέλεση αρχείων .php γίνεται μέσω της γραμμής εντολών, ως εξής:

```
php <file.php>
```

Η PHP μπορεί να αλληλοεπιδράσει με την MySQL, στέλνοντάς της ερωτήματα SQL και ανακτώντας τα δεδομένα και να επικοινωνήσει με τον εξυπηρετητή Apache, παράγοντας κώδικα HTML. Τα παραπάνω πακέτα που εγκαταστάθηκαν είναι απαραίτητα για την εκπλήρωση των αναφερθέντων λειτουργιών.

R & R Studio

Η γλώσσα προγραμματισμού R προσφέρει ένα περιβάλλον για ανάπτυξη εφαρμογών κυρίως για στατιστική και γραφικά. Για την αποτελεσματική εκτέλεση της γλώσσας R, εγκαθίσταται και το λογισμικό RStudio, ένα ολοκληρωμένο περιβάλλον ανάπτυξης για την R. Το RStudio διευκολύνει τη χρήση της γλώσσας, καθώς και την εγκατάσταση επιπλέον πακέτων και ενημέρωσης των ήδη υπαρχόντων. Η R στα Ubuntu Linux εγκαθίσταται χρησιμοποιώντας την εντολή root:

```
apt-get install r-base
```

Για την εγκατάσταση του RStudio επιλέγουμε την τελευταία έκδοση για το λογισμικό που έχουμε. Στην συγκεκριμένη περίπτωση επιλέξαμε την έκδοση για το Ubuntu 18.04:

https://download1.rstudio.org/desktop/bionic/amd64/rstudio-1.2.5033-amd64.deb

Η εγκατάσταση γίνεται με την εντολή root:

```
dpkg -i rstudio-1.2.5033-amd64.deb
```

Oracle Virtual Machine

Πολλές εφαρμογές δεν έχουν έκδοση διαθέσιμη για το λογισμικό των Linux. Για να μπορέσουμε να έχουμε διαθέσιμες αυτές τις εφαρμογές στον υπολογιστή μας, χρησιμοποιήσαμε το πρόγραμμα Oracle Virtual Box για να δημιουργήσουμε μία εικονική μηχανή με λειτουργικό Windows 10 64-bit. Επίσης, έγινε εγκατάσταση των Virtual Box Guest Additions, τα οποία προσφέρουν επιπλέον δυνατότητες στην εικονική μηχανή, όπως αλληλεπίδραση μεταξύ των δύο λειτουργικών και επικοινωνία των χώρων αποθήκευσης για εύκολη κοινή χρήση αρχείων.

Arabidopsis Coexpression Tool (ACT)

Συλλογή πρωταρχικών δεδομένων – Εισαγωγή

Για την συλλογή πρωτογενών δεδομένων μικροσυστοιχιών cDNA για το φυτικό πρότυπο *Arabidopsis thaliana* (αρχεία .CEL) έγινε αναζήτηση στα δημόσια καταθετήρια μεταγραφωμικών δεδομένων Gene Expression Omnibus (GEO) (Barrett et al., 2013) του NCBI, ArrayExpress (AE) (Kolesnikov et al., 2015) του EBI και NASCArrays (Craigon et al., 2004) του Nottingham Arabidopsis Stock Centre. Σε αυτή την αρχική συλλογή, παρατηρήθηκε ότι το κυρίαρχο chip μικροσυστοιχιών είναι το Arabidopsis ATH1 Genome Array (ATH1), της εταιρίας Affymetrix, με κωδικούς καταχώρησης πλατφόρμας GPL198 στην GEO και A-AFFY-2 στην AE. Οι αναλύσεις συσχέτισης απαιτούν ομοιογενή δεδομένα, για αυτό το σκοπό στην παρούσα μελέτη, θα χρησιμοποιηθούν δεδομένα μόνο από το συγκεκριμένο chip, καθώς αντιπροσωπεύει περίπου το 50% όλων των μεταγραφωμικών δεδομένων για το *Arabidopsis thaliana*. Παρότι η βάση δεδομένων NASCArrays δεν βρίσκεται σε λειτουργία, πραγματοποιήθηκε διεξοδική αναζήτηση στον κατάλογο των πρωτογενών δεδομένων τους. Επιπλέον, η αναζήτηση μικροσυστοιχιών στην βάση δεδομένων GEO ήταν ανεξάρτητη από την αναζήτηση στην βάση

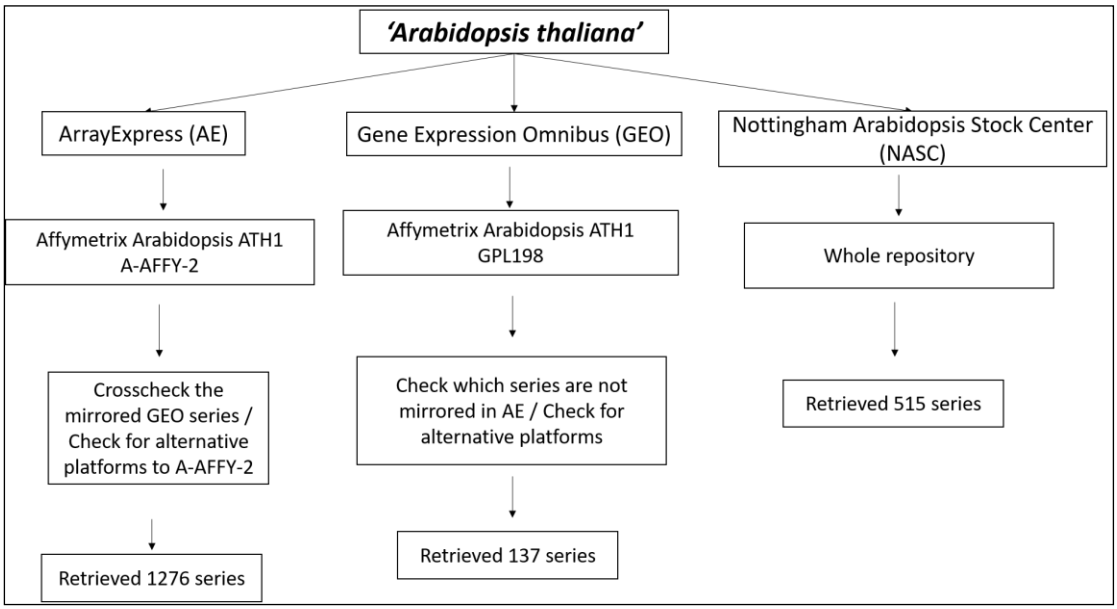
δεδομένων ΑΕ, παρότι η ΑΕ καταχωρεί κατοπτρικά αντίγραφα των δεδομένων της GEO. Ακόμα, έγινε ανάκτηση των ελάχιστων πληροφοριών περί πειραμάτων μικροσυστοιχιών (Minimum Information About a Microarray Experiment - MIAME) (Brazma et al., 2001), τα οποία περιλαμβάνουν πληροφορίες σχετικά με τον συνολικό πειραματικό σχεδιασμό, το σχεδιασμό της μικροσυστοιχίας, την προέλευση του κάθε ανιχνευτή, τη μέθοδο σήμανσης, τις διαδικασίες και τις παραμέτρους υβριδισμού και τις διαδικασίες μέτρησης.

Επίσης, έγινε αναζήτηση και ανάκτηση των πρωτογενών δεδομένων από τις εναλλακτικές πλατφόρμες των GPL198 / A-AFFY-2, οι οποίες χρησιμοποιούν το ίδιο chip μικροσυστοιχιών ATH1, αλλά τα δεδομένα τους έχουν κανονικοποιηθεί με διαφορετικό Αρχείο Περιγραφής Chip (Chip Description File - CDF) από το πρωτότυπο της Affymetrix, με το οποίο γίνεται η συσχέτιση συνόλων ανιχνευτών-γονιδίων. Η GEO και η ΑΕ καταχωρούν με διαφορετικό κωδικό κάθε εναλλακτική πλατφόρμα. Συχνά οι ερευνητές χρησιμοποιούν διαφορετικό αρχείο CDF από το πρωτότυπο για τη βελτίωση της συσχέτισης συνόλων ανιχνευτών και γονιδίων. Η GEO παρέχει έναν κατάλογο με εναλλακτικές πλατφόρμες προς την πλατφόρμα GPL198, αλλά συνήθως δεν περιλαμβάνει όλες τις εναλλακτικές πλατφόρμες. Ο καλύτερος τρόπος για την ταυτοποίηση αυτών, είναι μέσω αναζήτησης για εναλλακτικούς κωδικούς καταχώρησης στην ΑΕ που είναι δια χειρός επιμελημένη. Στη σελίδα περιήγησης της ArrayExpress αναζητήσαμε το chip ATH1 και ανακτήσαμε όλες τις εναλλακτικές πλατφόρμες.

Συλλογή πρωταρχικών δεδομένων – Εκτέλεση

Πραγματοποιήθηκε αναζήτηση στην ArrayExpress με ερώτημα που περιλάμβανε το "*Arabidopsis thaliana*", τον κωδικό καταχώρησης A-AFFY-2 και έγινε αναζήτηση για εναλλακτικές πλατφόρμες στο ATH1 από τη σελίδα περιήγησης. Στη συνέχεια, έγινε λήψη των πρωτογενών δεδομένων και μετα-δεδομένων από τον διακομιστή FTP της ΑΕ που περιέχει 1276 σειρές συμπεριλαμβανομένων των εναλλακτικών σειρών προς την ATH1. Με τη χρήση PHP scripts, έγινε έλεγχος εάν τα κατοπτρικά αντίγραφα από τη GEO είναι σωστά (ειδικά εάν έχουν μεταφερθεί σωστά τα πρωτογενή τους αρχεία). Ταυτοποιήθηκαν 14 κατοπτρικές σειρές GEO που περιλάμβαναν πρωτογενή

δεδομένα στη GEO τα οποία και ανακτήθηκαν. Στη συνέχεια, έγινε έλεγχος και λήψη όλων των δειγμάτων *Arabidopsis thaliana* από την GEO αναζητώντας για όλες τις σειρές της πλατφόρμας GPL198, περιλαμβάνοντας τις εναλλακτικές πλατφόρμες. Στη συνέχεια, εξετάστηκε ποιες από τις σειρές GEO περιέχουν κατοπτρικά αντίγραφα στην AE, οι οποίες και αφαιρέθηκαν. Βρέθηκαν 137 σειρές που ήταν μοναδικές μόνο για τη GEO (138 συμπεριλαμβανομένης μίας υπερσειράς η οποία υπήρχε στην AE οπότε και αγνοήθηκε). Τέλος, για τη βάση δεδομένων NASC ανακτήθηκαν όλες 515 σειρές από τον διακομιστή FTP τους, μαζί με τα μετα-δεδομένα τους. Όλη η διαδικασία έγινε προγραμματιστικά, χρησιμοποιώντας κώδικα γραμμένο στη γλώσσα PHP.



Εικόνα 7 – Διάγραμμα επιλογής μελετών από κάθε δημόσιο καταθετήριο

Σε δεύτερη φάση, έγινε έλεγχος για την ακεραιότητα των μελετών. Εν τέλει, οι 1183 από τις 1276 σειρές της AE, οι 135 από τις 137 της GEO και οι 396 από τις 515 της NASC περιείχαν πρωτογενή δεδομένα. Έγινε η μετατροπή των .CEL σε μορφοποίηση αρχείων κειμένου, με χρήση του προγράμματος apt-cel-convert.exe που προέρχεται από το πακέτο εντολών του Affymetrix Power Tools (APT) (Affymetrix, 2006), τα οποία στη συνέχεια ελέγχθηκαν για το αν όντως προέρχονται από το ATH1 chip. Αφαιρέθηκαν συνολικά 438 δείγματα που δεν ήταν ATH1. Σε επόμενο έλεγχο έγινε η ταυτοποίηση δειγμάτων των οποίων οι ανιχνευτές είχαν τιμές έντασης εκτός ορίων, οπότε αφαιρέθηκαν 54 επιπλέον δείγματα. Αυτά τα 54 δείγματα είχαν ληφθεί από την AE αλλά προέρχονταν αρχικά από την NASC. Έτσι έγινε επιπλέον έλεγχος στον παλιό

ιστότοπο της NASC, οπού και βρέθηκαν τα 50 από τα 54 δείγματα και μετονομάστηκαν αντίστοιχα, πραγματοποιώντας παράλληλα και επιπλέον επαλήθευση των δεδομένων. Σε τελευταίο στάδιο, έγινε έλεγχος για διπλότυπα αρχεία .CEL. Βρέθηκαν 5960 ίδια δείγματα, τα οποία αφαιρέθηκαν (AE: 1800, GEO: 507, NASC: 3653). Τελικά απέμειναν 19887 δείγματα από 1391 πειράματα (σειρές).

Ποιοτικός έλεγχος πρωτογενών δεδομένων και απόρριψη δεδομένων χαμηλής ποιότητας

Πραγματοποιήθηκε ποιοτικός έλεγχος (Quality Control - QC) στα πρωτογενή δεδομένα μικροσυστοιχιών CEL, χρησιμοποιώντας την σουίτα προγραμμάτων BioConductor (Gentleman et al., 2004; Huber et al., 2015) και συγκεκριμένα τα πακέτα: ‘simplyaffy’ (Miller CJ, 2018), ‘affyQCReport’ (Parman et al., 2018) και ‘affyPLM’ (Bolstad et al., 2005 ; Brettschneider et al., 2008), χρησιμοποιώντας τον αλγόριθμο κανονικοποίησης MAS 5.0 (Hubbell et al., 2002) και το αρχείο CDF της Affymetrix. Για κάθε δείγμα χρησιμοποιήθηκαν ποιοτικές μετρήσεις μοναδικών μικροσυστοιχιών Affymetrix, όπως το μέσο υπόβαθρο, ο συντελεστής κλίμακας, το ποσοστό των παρόντων γονιδίων (present) και ο λόγος υβριδισμού RNA 3’ προς 5’ των γονιδίων της β-ακτίνης και GADPH.

Χρησιμοποιήθηκαν δύο ποιοτικές μετρήσεις πολλαπλών μικροσυστοιχιών χρησιμοποιώντας τα θηκογράμματα RLE (Relative Log Expression) (Bolstad et al., 2005) και NUSE (Normalized Unscaled Standard Errors)(Bolstad et al., 2005). Οι τιμές RLE υπολογίζονται για κάθε σύνολο ανιχνευτών συγκρίνοντας την τιμή έκφρασης μιας μικροσυστοιχίας, σε σχέση με τη μέση τιμή έκφρασης για το συγκεκριμένο σύνολο ανιχνευτών μεταξύ των μικροσυστοιχιών κάθε μελέτης. Εφόσον η έκφραση των περισσοτέρων ανιχνευτών δεν μεταβάλλεται μεταξύ των μικροσυστοιχιών κάθε μελέτης, αναμένεται ότι η διάμεσος των λόγων έκφρασης των συνόλων ανιχνευτών σε όλα τα δείγματα μίας σειράς θα είναι περίπου 0 σε λογαριθμική κλίμακα. Τα θηκογράμματα RLE που παρουσιάζουν την κατανομή αυτών των λογαριθμημένων τιμών θα πρέπει να είναι κεντραρισμένα κοντά στο 0, ενώ τα δείγματα χαμηλής ποιότητας θα πρέπει να έχουν διασπορά μεγαλύτερη από 0.2. Οι μικροσυστοιχίες που είχαν

ακραίες τιμές ή ήταν υψηλότερες από τις καθορισμένες ουδούς δεν θα χρησιμοποιηθούν για περαιτέρω ανάλυση.

Επιπλέον, έγινε επιμελής αναζήτηση στα μετα-δεδομένα του κάθε πειράματος ώστε να ανακαλυφθούν όλα τα δείγματα που προέρχονται από διαφορετικό φυτικό είδος, από ολόκληρο το φυτό και από μεταλλαγμένα δείγματα. Οπότε έγινε αναζήτηση για λέξεις κλειδιά όπως `whole plant`, `whole organism`, `mutated` κτλπ και αφαιρέθηκαν όλα τα πειράματα σχετικά με αυτά. Τελικά καταλήξαμε σε 6933 υγιή δείγματα από διακριτά μέρη (ιστούς) του *Arabidopsis thaliana*.

Κανονικοποίηση των πρωτογενών δεδομένων με βέλτιστους αλγορίθμους και περιγραφές ανιχνευτών για την παραγωγή δευτερογενών δεδομένων

Στα πρωτογενή δεδομένα (CEL) των μικροσυστοιχιών που πέρασαν επιτυχώς τον ποιοτικό έλεγχο και τα άλλα κριτήρια επιλογής, πραγματοποιήθηκε επεξεργασία με τον αλγόριθμο SCAN (Piccolo et al., 2012). Ο αλγόριθμος SCAN κανονικοποιεί κάθε μικροσυστοιχία ανεξάρτητα από τα υπόλοιπα δείγματα της σειράς, πραγματοποιεί διόρθωση μεροληψίας του ποσοστού GC και μειώνει τις διακυμάνσεις των ανιχνευτών και των μικροσυστοιχιών από κάθε ξεχωριστό δείγμα, ενώ αυξάνει τον λόγο σήματος/θορύβου. Ο αλγόριθμος SCAN προτιμάται όταν συνδυάζονται δείγματα μικροσυστοιχιών από διαφορετικές σειρές ή εργαστήρια, καθώς άλλοι αλγόριθμοι επεξεργασίας, όπως ο RMA (Irizarry et al., 2003a) ή ο GC-RMA (Wu et al., 2004), αντλούν πληροφορίες από όλα τα δείγματα μαζί, για την κανονικοποίηση και με αυτό τον τρόπο δυνητικά εισάγουν εσφαλμένες συσχετίσεις. Επιπλέον χρησιμοποιήθηκε το αρχείο περιγραφής μικροσυστοιχιών της BrainArray (TAIRG version 22.0.0)(Dai et al., 2005) που υπερτερεί του default CDF της Affymetrix. Ο λόγος είναι ότι προ δεκαπενταετίας που σχεδιάστηκε το CDF της Affymetrix, οι γνώσεις για το γονιδίωμα και το μεταγράφωμα ήταν πρώιμες, οπότε αρκετά σύνολα ανιχνευτών είτε δεν αντιστοιχούν σε κανένα γονίδιο, είτε αντιστοιχούν σε περισσότερα από ένα γονίδια, ενώ και κάποια γονίδια αντιστοιχούν σε περισσότερα του ενός συνόλου ανιχνευτών. Η εκ νέου ανάλυση των δεδομένων χρησιμοποιώντας CDF, τα οποία ενημερώνονται συνεχώς, δίνει πιο αξιόπιστα δευτερογενή δεδομένα,

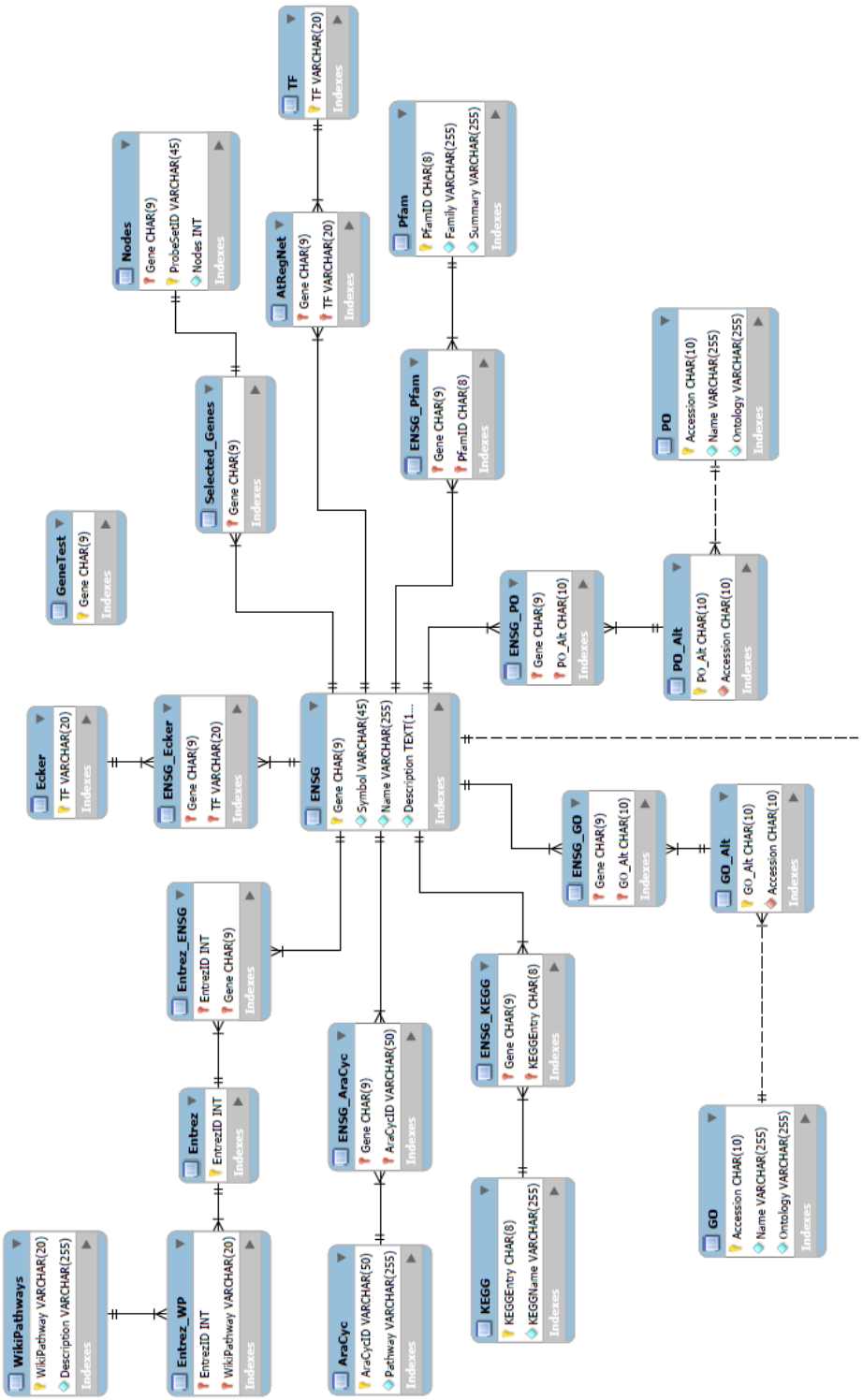
όπου κάθε σύνολο ανιχνευτών αντιστοιχεί σε ένα γονίδιο και κάθε γονίδιο σε ένα σύνολο ανιχνευτών.

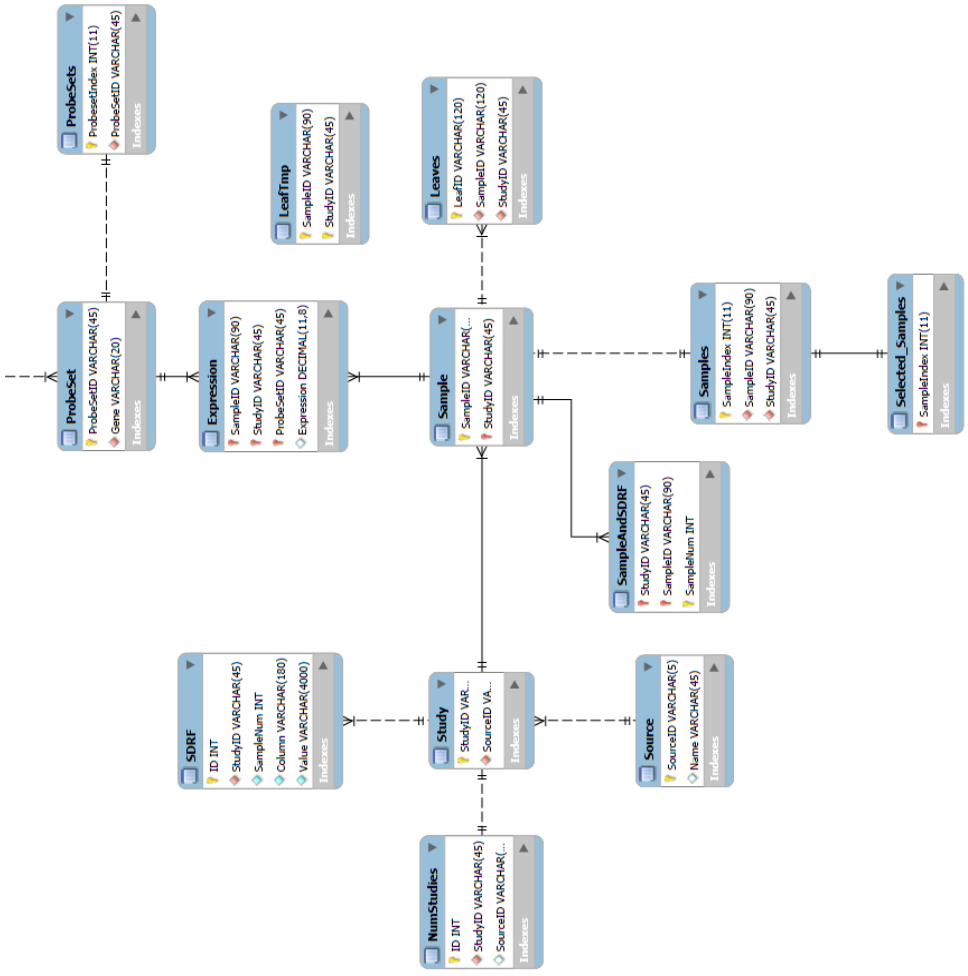
Αναζήτηση και ανάκτηση από διάφορες βάσεις δεδομένων όρων που περιγράφουν τα γονίδια του φυτού και επεξεργασία τους

Έγινε αναζήτηση βιολογικών όρων για Οντολογίες Γονιδίων (GeneOntology) (The Gene Ontology, 2019), Οντολογίες Φυτών (Plant Ontology) (Cooper et al., 2018), Βιολογικών Μονοπατιών (KEGG, AraCyc Pathways) (Kanehisa and Goto, 2000; Schlapfer et al., 2017), Μεταγραφικών Παραγόντων (Plant Cistrome Database, AtRegNet) (O'Malley et al., 2016; Yilmaz et al., 2011) και Οικογενειών Πρωτεϊνών (Pfam) (El-Gebali et al., 2019), οι οποίοι περιγράφουν τα γονίδια του *Arabidopsis thaliana*.

Κατασκευή βάσης δεδομένων και αποθήκευση των δευτερογενών δεδομένων σε αυτή

Η σχεσιακή βάση δεδομένων δημιουργήθηκε με σκοπό να αποθηκεύσει όλα τα απαραίτητα δεδομένα, τα οποία στη συνέχεια θα μπορούν να χρησιμοποιηθούν με διάφορους τρόπους, όπως για τη δημιουργία του πίνακα αποστάσεων μεταξύ γονιδίων ή για την παρουσίασή τους στον ιστότοπο. Η βάση σχεδιάστηκε με τη βοήθεια του MySQL Workbench, μέσω του οποίου δημιουργήθηκε και το διάγραμμα Οντοτήτων – Συσχετίσεων (Εικόνα 8).


TF
TF VARCHAR(20)
Indexes
Nodes
Gene CHAR(9)
ProbeSetID VARCHAR(45)
Nodes INT
Indexes
AtRegNet
Gene CHAR(9)
TF VARCHAR(20)
Indexes
Pfam
PfamID CHAR(8)
Family VARCHAR(255)
Summary VARCHAR(255)
Indexes
Selected_Genes
Gene CHAR(9)
Indexes
ENSG_Pfam
Gene CHAR(9)
PfamID CHAR(8)
Indexes
PO
Accession CHAR(10)
Name VARCHAR(255)
Ontology VARCHAR(255)
Indexes
GeneTest
Gene CHAR(9)
Indexes
ENSG_PO
Gene CHAR(9)
PO_Alt CHAR(10)
Indexes
PO_Alt
PO_Alt CHAR(10)
Accession CHAR(10)
Indexes
Ecker
TF VARCHAR(20)
Indexes
ENSG_Ecker
Gene CHAR(9)
TF VARCHAR(20)
Indexes
ENSG
Gene CHAR(9)
Symbol VARCHAR(45)
Name VARCHAR(255)
Description TEXT(1...
Indexes
ENSG_GO
Gene CHAR(9)
Go_Alt CHAR(10)
Indexes
GO_Alt
GO_Alt CHAR(10)
Accession CHAR(10)
Indexes
Entrez_ENSG
EntrezID INT
Gene CHAR(9)
Indexes
ENSG_KEGG
Gene CHAR(9)
KEGGEntry CHAR(8)
Indexes
ENSG_AraCyc
Gene CHAR(9)
AraCycID VARCHAR(50)
Indexes
Entrez
EntrezID INT
Indexes
GO
Accession CHAR(10)
Name VARCHAR(255)
Ontology VARCHAR(255)
Indexes
KEGG
KEGGEntry CHAR(8)
KEGGName VARCHAR(255)
Indexes
AraCyc
AraCycID VARCHAR(50)
Pathway VARCHAR(255)
Indexes
WikiPathways
WikiPathway VARCHAR(20)
Description VARCHAR(255)
Indexes
Entrez_WP
EntrezID INT
WikiPathway VARCHAR(20)
Indexes



Εικόνα 8 – Το ERD της βάσης δεδομένων του ACT

Η βάση δεδομένων περιέχει τους εξής κύριους πίνακες:

- ***Source***

  Ο πίνακας Source περιέχει τα ονόματα των δημόσιων αποθετηρίων από τα οποία έγινε λήψη πρωτογενών δεδομένων πειραμάτων. Περιλαμβάνει τα χαρακτηριστικά SourceID το οποίο αποτελεί το πρωτεύον κλειδί και περιέχει τον κωδικό της κάθε βάσης δεδομένων (GEO, NASC και AE) και Name που αναφέρει αναλυτικά το όνομα της κάθε βάσης δεδομένων (Gene Expression Omnibus, Nottingham Arabidopsis Stock Centre και ArrayExpress, αντίστοιχα).

- ***Study***

  Ο πίνακας Study περιέχει τα ονόματα των μελετών που συλλέχθηκαν. Περιλαμβάνει το StudyID που αναφέρει τον κωδικό της μελέτης και αποτελεί και το πρωτεύον κλειδί και το SourceID και περιέχει τον κωδικό της κάθε βάσης δεδομένων. Ο συγκεκριμένος πίνακας συσχετίζεται με μία προς πολλές (1:N) εγγραφές με τον πίνακα Sample.

- ***Sample***

  Ο πίνακας Sample περιέχει τα δείγματα των μελετών που συλλέχθηκαν. Περιλαμβάνει τα χαρακτηριστικά SampleID (πρωτεύον κλειδί) που περιλαμβάνει τον κωδικό του κάθε δείγματος και το StudyID που περιλαμβάνει τον κωδικό της κάθε μελέτης. Ο πίνακας Sample συσχετίζεται με μία προς πολλές (1:N) εγγραφές με τον πίνακα Expression.

- ***Samples***

  Ο πίνακας Samples περιλαμβάνει τα χαρακτηριστικά SampleID (πρωτεύον κλειδί) που περιλαμβάνει τον κωδικό του κάθε δείγματος και το StudyID που περιλαμβάνει τον κωδικό της κάθε μελέτης και το χαρακτηριστικό SampleIndex που περιλαμβάνει την αρίθμηση των δειγμάτων με αύξουσα σειρά. Χρησιμοποιείται ως ενδιάμεσος πίνακας μεταξύ Sample και Selected_Samples.

- ***Expression***

  Ο πίνακας Expression περιλαμβάνει τα χαρακτηριστικά SampleID (κωδικός του δείγματος), StudyID(κωδικός μελέτης), ProbeSetID που περιλαμβάνει τον κωδικό του κάθε συνόλου ανιχνευτών και Expression που περιέχει τις τιμές έκφρασης του ανιχνευτή για κάθε δείγμα.

- ***ProbeSet***

  Ο πίνακας ProbeSet περιέχει το σύνολο ανιχνευτών για κάθε γονίδιο για το chip μικροσυστοιχιών της Affymetrix για το *Arabidopsis thaliana*. Περιλαμβάνει τα χαρακτηριστικά ProbeSetID (κωδικός του κάθε συνόλου ανιχνευτών) και Gene που περιλαμβάνει τον κωδικό AGI του κάθε γονιδίου. Ο συγκεκριμένος πίνακας συσχετίζεται με μία προς πολλές (1:N) εγγραφές με τον πίνακα Expression.

- ***ProbeSets***

  Ο πίνακας ProbeSets περιλαμβάνει τα χαρακτηριστικά ProbeSetIndex (πρωτεύον κλειδί) που περιλαμβάνει την αρίθμηση των συνόλων ανιχνευτών με αύξουσα σειρά και ProbeSetID (κωδικός του κάθε συνόλου ανιχνευτών).

- ***Selected_Samples***

  Ο πίνακας Selected_Samples περιέχει όλα τα δείγματα που επιλέχθηκαν μετά τον έλεγχο ποιότητας. Περιλαμβάνει τα χαρακτηριστικά SampleIndex (πρωτεύον κλειδί) που περιλαμβάνει την αρίθμηση των δειγμάτων με αύξουσα σειρά και ProbeSetID (κωδικός του κάθε συνόλου ανιχνευτών).

- ***SDRF***

  Ο πίνακας SDRF περιέχει τα μετα-δεδομένα για το κάθε πείραμα. Περιλαμβάνει τα χαρακτηριστικά ID, που είναι ένας μοναδικός αύξων αριθμός για κάθε χαρακτηριστικό και πρωτεύον κλειδί, StudyID ως πρωτεύον και ξένο κλειδί και SampleNum, Column και Value, τα οποία χρησιμοποιούνται για να αποθηκεύσουν και να κατηγοριοποιήσουν τις διαφορετικές πληροφορίες του κάθε δείγματος σε κάθε μελέτη.

- ***ENSG***

  Ο πίνακας ENSG περιέχει τις πληροφορίες των γονιδίων. Περιλαμβάνει τα χαρακτηριστικά Gene που είναι ένα μοναδικό όνομα για κάθε γονίδιο και πρωτεύον κλειδί, Symbol που είναι η ονομασία-σύμβολο του γονιδίου κατά HGNC, Name το πλήρες όνομα του γονιδίου και Description που είναι η περιγραφή του.

- ***Selected_Genes***

  Ο πίνακας Selected_Genes περιέχει τα επιλεγμένα γονίδια του ACT. Περιλαμβάνει το χαρακτηριστικό Gene ως ξένο και πρωτεύον κλειδί.

- ***TF***

  Ο πίνακας TF περιέχει τα ονόματα όλων των διαθέσιμων μεταγραφικών παραγόντων για το *Arabidopsis thaliana*. Περιλαμβάνει το χαρακτηριστικό TF ως κύριο κλειδί που αποτελεί τα ονόματα των μεταγραφικών παραγόντων που έχουν δεδομένα με γονίδια στόχους. Τα δεδομένα προέρχονται από το AtRegNet.

- ***GO***

  Ο πίνακας GO περιέχει τα δεδομένα οντολογίας γονιδίων από το Gene Ontology. Περιλαμβάνει τα χαρακτηριστικά Accession ως πρωτεύον κλειδί, που είναι ο μοναδικός κωδικός για κάθε οντολογία, Name και Ontology, που αντίστοιχα είναι το όνομα και σε ποια κατηγορία οντολογίας γονιδίων από τις τρεις ανήκει (Biological Process, Cellular Component, Molecular Function)

- ***PO***

  Ο πίνακας PO περιέχει τα δεδομένα οντολογίας φυτών από τη βάση δεδομένων Planteome. Περιλαβάνει τα χαρακτηριστικά Accession ως πρωτεύον κλειδί, που είναι ο μοναδικός κωδικός για κάθε οντολογία, Name και Ontology, που είναι το όνομα και σε ποια κατηγορία οντολογίας φυτών από τις δύο ανήκει (Plant Anatomy, Plant Developmental Stage).

- ***Entrez***

  Ο πίνακας Entrez περιέχει του κωδικούς γονιδίων από τη βάση δεδομένων Entrez Genes. Περιλαμβάνει το χαρακτηριστικό EntrezID ως πρωτεύον κλειδί, που είναι οι μοναδικοί κωδικοί γονιδίων με βάση τα EntrezGenes. Χρησιμοποιείται για την περιγραφή των WikiPathways.

- ***WikiPathways***

  Ο πίνακας WikiPathways περιέχει δεδομένα για βιολογικά μονοπάτια γονιδίων. Περιλαμβάνει τα χαρακτηριστικά WikiPathway, που είναι ένας μοναδικός κωδικός για κάθε χαρακτηριστικό από τη βάση WikiPathways και πρωτεύον κλειδί και Description το οποίο είναι μία αναλυτική περιγραφή του χαρακτηριστικού.

- ***AraCyc***

  Ο πίνακας AraCyc περιέχει δεδομένα για μεταβολικά μονοπάτια γονιδίων φυτών. Περιλαμβάνει τα χαρακτηριστικά AraCyc, που είναι ένας μοναδικός κωδικός για κάθε χαρακτηριστικό και πρωτεύον κλειδί και Pathway το οποίο είναι μία αναλυτική περιγραφή του χαρακτηριστικού.

- ***KEGG***

  Ο πίνακας KEGG περιέχει δεδομένα για βιολογικά μονοπάτια γονιδίων από τη βάση KEGG Pathways. Περιλαμβάνει τα χαρακτηριστικά KEGGEntry που είναι ένας μοναδικός κωδικός για κάθε χαρακτηριστικό και πρωτεύον κλειδί και KEGGName το οποίο είναι το πλήρες όνομα του βιολογικού μονοπατιού.

- ***PFam***

  Ο πίνακας PFam περιέχει δεδομένα για τις οικογένειες πρωτεϊνών των γονιδίων. Περιλαμβάνει τα χαρακτηριστικά PfamID που είναι ένας μοναδικός κωδικός για κάθε εγγραφή και πρωτεύον κλειδί, Family το οποίο αναφέρεται στην ευρύτερη οικογένεια που ανήκει η εγγραφή και Summary το οποίο είναι μία αναλυτική περιγραφή της οικογένειας.

- ***AtRegNet***

  Ο πίνακας AtRegNet περιέχει δεδομένα μεταγραφικών παραγόντων και των γονιδίων στόχων τους από την ειδική βάση δεδομένων AtRegNet. Περιλαμβάνει τα χαρακτηριστικά Gene που είναι το γονίδιο και TF, ο μεταγραφικός παράγοντας που είναι στόχος του, και τα δύο ως πρωτεύοντα και ξένα κλειδιά.

- ***Ecker***

  Ο πίνακας Ecker περιέχει δεδομένα μεταγραφικών παραγόντων και των γονιδίων στόχων τους από την ειδική βάση δεδομένων Plant Cistrome Database. Περιλαμβάνει τα χαρακτηριστικά Gene που είναι το γονίδιο και TF, ο μεταγραφικός παράγοντας που είναι στόχος του, και τα δύο ως πρωτεύοντα και ξένα κλειδιά.

Οι υπόλοιποι πίνακες που εμφανίζονται στο διάγραμμα αποτελούν ενδιάμεσους πίνακες που συσχετίζουν τα δεδομένα των πινάκων που ενώνουν. Για παράδειγμα ο πίνακας ENSG_GO συσχετίζει όλα τα γονίδια με το κάθε όρο οντολογίας γονιδίων που περιγράφει το καθένα. Σε κάθε πίνακα έγινε εισαγωγή του αρχείου .txt με το αντίστοιχο όνομα. Η δημιουργία των flat files περιγράφεται παρακάτω.

Λεπτομερής ανάλυση τρόπου ανάκτησης δεδομένων από τις αντίστοιχες βάσεις

Thalemine Araport

Το Thalemine (Krishnakumar et al., 2015) αποτελεί την πλέον αναλυτικότερη και περιεκτική βάση δεδομένων σε ό,τι αφορά γονίδια για το *Arabidopsis thaliana*. Προσφέρει αναλυτικές πληροφορίες για κάθε γονίδιο, δίνει δυνατότητα λήψης πληροφοριών του κάθε γονιδίου με εξωτερικές βάσεις, όπως οντολογίες γονιδίων και φυτών και βιολογικών μονοπατιών. Ο ιστότοπος προσφέρει μία λειτουργία λήψης πληροφοριών, τόσο από την ίδια τη βάση δεδομένων Araport (Cheng et al., 2017), όσο και από εξωτερικές πηγές, παρόμοια με το Ensembl Biomart (Kinsella et al., 2011). Χάρις τη δυνατότητα σύνδεσης με αυτές τις εξωτερικές πηγές, είναι δυνατόν να παράγουμε λίστες που συνδέουν το ENSG ID του κάθε γονιδίου με το αντίστοιχο χαρακτηριστικό, π.χ. τις οντολογίες γονιδίων που τον περιγράφουν ή τα ονόματα HGNC του

κάθε γονιδίου. Επιπλέον, μπορούμε να παράγουμε αυτή τη λίστα με τις ίδιες ακριβώς παραμέτρους προγραμματιστικά, χωρίς να την κατεβάσουμε τοπικά, χρησιμοποιώντας τη λειτουργία XML που προσφέρεται από το Thalemine BioMart.

Gene Ontology

Οι οντολογίες γονιδίων που περιγράφουν τα γονίδια του *Arabidopsis thaliana* ελήφθησαν από τον ιστότοπο του geneontology.com. Έγινε χρήση ενός αρχείου μορφοποίησης .obo το οποίο περιέχει όλες τις οντολογίες που είναι διαθέσιμες. Κάθε εγγραφή έχει έναν μοναδικό κωδικό που λέγεται Accession και χαρακτηρίζεται από τα αρχικά GO και έναν επταψήφιο αριθμό ενωμένα με άνω-κάτω τελεία (π.χ. GO:0000005). Κάτω από τον μοναδικό κωδικό υπάρχουν επιπλέον πληροφορίες σχετικά με την οντολογία, όπως το πλήρες όνομά της, μία περιγραφή της καθώς και οι σχέσεις της με άλλες οντολογίες. Μία οντολογία μπορεί να αποτελεί ευρύτερο όρο από μία άλλη οπότε αποτελεί οντολογία-γονέα αυτής. Υπάρχουν 3 κατηγορίες οντολογιών, biological process, cellular component και molecular function, άρα όλες οι οντολογίες εμπίπτουν σε μία ή περισσότερες από τις 3 κατηγορίες. Τέλος, αναγράφεται αν μία οντολογία έχει καταργηθεί καθώς και η προτεινόμενη οντολογία που θα την αντικαταστήσει ή αν έχει εναλλακτικές οντολογίες.

Κατά την λήψη των δεδομένων δημιουργήθηκε το αρχείο GO.txt, το οποίο περιέχει όλες τις διαθέσιμες οντολογίες, ανάλογα με την κατηγορία που ανήκουν. Επιπλέον, δημιουργήθηκε και ένα αρχείο GO_Alt.txt το οποίο συνδέει το Accession κάθε οντολογίας με το εναλλακτικό ή το προτεινόμενο Accession που της αντιστοιχεί.

Το αρχείο που συνδέει το κάθε γονίδιο με κάποια οντολογία (ENSG_GO.txt) δημιουργήθηκε χρησιμοποιώντας την πλατφόρμα εξόρυξης δεδομένων που προσφέρεται από το Thalemine.

Για την κατασκευή του αρχείου που περιέχει τις οντολογίες και τις οντολογίες-γονείς της κάθε οντολογίας, θεωρήθηκε ότι γονείς είναι οι οντολογίες που περιγράφονται με τις εξής φράσεις:

```
is_a: GO:
intersection_of: part_of GO:
```

```
intersection_of: negatively_regulates GO:
intersection_of: occurs_in GO:
intersection_of: positively_regulates GO:
intersection_of: regulates GO:
relationship: part_of GO:
relationship: negatively_regulates GO:
relationship: occurs_in GO:
relationship: positively_regulates GO:
relationship: regulates GO:
```

Plant Ontology

Αντίστοιχα με τις οντολογίες γονιδίων, υπάρχουν και οι οντολογίες φυτών, που περιέχουν όρους και βιολογικές διεργασίες αποκλειστικές για φυτικούς οργανισμούς. Η κάθε οντολογία μπορεί να ανήκει σε μία από δύο κατηγορίες (Plant Anatomy και Plant Developmental Stage). Η διαφορά με τις οντολογίες γονιδίων είναι ότι τα αρχικά κάθε Accession είναι PO.

Η λήψη των δεδομένων έγινε με αντίστοιχο τρόπο με το Gene Ontology και δημιουργήθηκαν τα αρχεία, PO.txt, PO_Alt.txt και ENSG_PO.txt.

KEGG

Η βάση δεδομένων KEGG προσφέρει πληροφορίες για βιολογικά μονοπάτια στα οποία λαμβάνουν μέρος γονίδια. Τα μονοπάτια χωρίζονται ανά οργανισμό και επιπλέον κάθε μονοπάτι έχει έναν μοναδικό κωδικό στη βάση δεδομένων. Έγινε αναζήτηση και λήψη όλων των διαθέσιμων μονοπατιών για το *Arabidopsis thaliana*, που έχει τον μοναδικό κωδικό οργανισμού T00041 στη βάση δεδομένων, στο αρχείο KEGG.txt μέσω του συνδέσμου:

https://www.genome.jp/dbget-bin/get_linkdb?-t+pathway+gn:T00041

Τα δεδομένα συνδυασμού γονιδίων του φυτού και μονοπατιών της KEGG, λήφθηκαν από το Biomart του Thalemine, στο αρχείο ENSG_KEGG.txt.

WikiPathways

Η WikiPathways είναι μία βάση δεδομένων για βιολογικά μονοπάτια και συντηρείται και ανανεώνεται από την επιστημονική κοινότητα. Έγινε λήψη των δεδομένων βιολογικών μονοπατιών για το *Arabidopsis thaliana* από το σύνδεσμο:

http://data.wikipathways.org/current/gmt/Arabidopsis_thaliana

Επειδή τα δεδομένα από την WikiPathways χρησιμοποιούν τους κωδικούς Entrez για κάθε γονίδιο, έγινε πρώτα λήψη πληροφοριών για τα όλα γονίδια από τη βάση δεδομένων Entrez καθώς και η αντιστοίχισή τους με τους κωδικούς από την Ensembl. Παράχθηκαν τα αρχεία, Entrez.txt, Entrez_ENSG.txt, WikiPathways.txt και Entrez_WP.txt.

AraCyc

Στη βάση δεδομένων AraCyc περιέχονται δεδομένα για βιολογικά μονοπάτια συγκεκριμένα για το *Arabidopsis thaliana* καθώς αποτελεί παράρτημα της TAIR. Έγινε λήψη όλων των βιολογικών μονοπατιών από το σύνδεσμο:

https://pmn.plantcyc.org/groups/export?id=:ALL-PATHWAYS&tsv-type=FRAMES

και παράχθηκαν τα αρχεία AraCyc.txt και ENSG_AraCyc.txt το οποίο περιγράφει τη σύνδεση γονιδίων και μονοπατιών.

AtRegNet

Η βάση δεδομένων AtRegNet περιέχει πειραματικά επαληθευμένους στόχους μεταγραφικών παραγόντων πάνω σε γονίδια του *Arabidopsis thaliana*. Έγινε λήψη του αρχείου από το σύνδεσμο:

https://agris-knowledgebase.org/Downloads/AtRegNet.zip

και παράχθηκαν τα αρχεία AtRegNet.txt και TF.txt

Plant Cistrome Database

Η βάση δεδομένων Plant Cistrome Database περιέχει στόχους μεταγραφικών παραγόντων πάνω σε γονίδια του *Arabidopsis thaliana*, παραγόμενους μέσω πειραμάτων DAP-seq. Έγινε λήψη των δεδομένων ampDAP γονιδίων στόχων από το σύνδεσμο:

http://neomorph.salk.edu/dap_web/pages/dap_data_v4/fullset/dap_download_may2016_genes.zip

και παράχθηκαν τα αρχεία Ecker.txt και TF_Ecker.txt

Pfam

Η βάση δεδομένων Pfam είναι η κατεξοχήν επιλογή για δεδομένα οικογενειών πρωτεϊνών. Έγινε λήψη των δεδομένων από το σύνδεσμο:

ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/database_files/pfamA.txt.gz

και παράχθηκαν τα αρχεία Pfam.txt και ENSG_Pfam.txt

Δημιουργία του δέντρου συσχέτισης δειγμάτων και αυτόματη επιλογή των αντιπροσωπευτικών δειγμάτων

Στη διάθεσή μας μέσω του πίνακα Expression της βάσης δεδομένων, έχουμε την έκφραση του κάθε γονιδίου που περιλαμβάνονται στο chip της Affymetrix, σε κάθε ένα από τα δείγματα που έχουμε κατεβάσει από τα δημόσια καταθετήρια. Αυτό μας δίνει τη δυνατότητα να υπολογίσουμε τον συντελεστή συσχέτισης Pearson μεταξύ του κάθε ζεύγους από τα 6933 δείγματα συγκρίνοντας την έκφραση του κάθε δείγματος ανά γονίδιο.

Ο συντελεστής συσχέτισης Pearson (Pearson Correlation Coefficient, PCC) (Pearson, 1895) ή αλλιώς τιμή *r* (*r*-value), υπολογίζεται ως εξής:

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

Σε αυτήν την περίπτωση x_i είναι η τιμή έκφρασης ενός συγκεκριμένου γονιδίου *i* στο δείγμα *x* και y_i είναι η τιμή έκφρασης του ίδιου γονιδίου στο δείγμα *y*, με το *n* να είναι το σύνολο όλων των γονιδίων. Οι διαθέσιμες τιμές του *r*-value είναι από -1 μέχρι 1. Το 1 δείχνει ότι τα δύο στοιχεία, στην περίπτωσή μας τα δείγματα, που συγκρίνονται είναι απολύτως συσχετιζόμενα το ένα με το άλλο, ενώ η τιμή -1 δείχνει ότι τα δύο δείγματα είναι απολύτως αντίθετα συσχετιζόμενα. Οι ενδιάμεσες τιμές δείχνουν πόσο πολύ συσχετίζονται θετικά ή αρνητικά τα δύο στοιχεία, ενώ η τιμή 0 σημαίνει ότι δεν υπάρχει καμία απολύτως συσχέτιση. Εν τέλει, λαμβάνονται οι ανά ζεύγη συντελεστές συσχέτισης ανάμεσα στα δείγματα. Οι τιμές αυτές μπορούν να αναπαρασταθούν ως ένας πίνακας αποστάσεων μορφοποίησης Phylip (Felsenstein, 2008). Αρχικά, πρέπει να γίνει η μετατροπή των *r*-values σε τιμές αποστάσεων (distance values). Ο μαθηματικός τύπος που χρησιμοποιείται είναι:

$$d = 1 - r \text{ (Kassambara, 2017)}$$

Αφαιρώντας την κάθε *r*-value από το 1, έχουμε πλέον εύρος τιμών [0,2], με τη μικρότερη τιμή να δείχνει πλήρη συσχέτιση και τη μεγαλύτερη να δείχνει πλήρη αντι-συσχέτιση.

Τα αρχεία Phylip ακολουθούν τη μορφή τριγωνικού πίνακα. Στη βασική τους μορφή, η πρώτη γραμμή είναι ένας αριθμός με το σύνολο όλων των δειγμάτων. Οι υπόλοιπες γραμμές είναι χωρισμένες με tab, σε μορφή στηλών. Η πρώτη στήλη της κάθε γραμμής είναι και ένα διαφορετικό δείγμα και οι υπόλοιπες στήλες τις γραμμής είναι οι τιμές αποστάσεων εκείνου του δείγματος με τα υπόλοιπα συμπεριλαμβανομένου και του ιδίου, οπότε το σύνολο των στηλών είναι ίσο με τον αριθμό των δειγμάτων συν την πρώτη στήλη με το όνομα του διαφορετικού δείγματος κάθε φορά. Αντίστοιχα και το σύνολο των γραμμών είναι ίσο με τον αριθμό των δειγμάτων συν την πρώτη γραμμή που αναφέραμε. Επίσης, περιμένουμε η διαγώνιος του πίνακα να είναι μηδενική καθώς γίνεται σύγκριση κάθε δείγματος με τον εαυτό του, άρα και η τιμή απόστασης μεταξύ του ίδιου δείγματος θα είναι 0. Καθώς, λοιπόν, οι πίνακες μορφοποίησης Phylip έχουν τη μορφή που περιγράψαμε, εν τέλει μπορούμε να πούμε ότι είναι πρακτικά ένας άνω τριγωνικός πίνακας, που τα κατοπτρικά του στοιχεία χωρίζονται από τη μηδενική διαγώνιο. Τελικά, οι μοναδικές τιμές του πίνακα υπολογίζονται από τον τύπο:

$$\frac{N(N-1)}{2}$$

	6					
A	0	0.3	0.15	1.35	1.25	0.6
B	0.3	0	0.3	1.3	1.2	0.65
C	0.15	0.3	0	1.35	1.2	0.6
D	1.35	1.3	1.35	0	0.4	1.4
E	1.25	1.2	1.2	0.4	0	1.3
F	0.6	0.65	0.6	1.4	1.3	0

Πίνακας 1 – Παράδειγμα αρχείου Phylip

Η δημιουργία του αρχείου μορφοποίησης Phylip, πίνακα αποστάσεων, μεταξύ των 6933 δειγμάτων έγινε χρησιμοποιώντας το script *pcc.php*. Το αρχείο αυτό είχε 6934 γραμμές και 6934 στήλες και με συνολικό αριθμό μοναδικών τιμών PCC να είναι 24,029,778.

Χρησιμοποιώντας το αρχείο μορφοποίησης Phylip ως αρχείο εισόδου, οι αλγόριθμοι ιεραρχικής ομαδοποίησης μπορούν να δημιουργήσουν ένα φυλογενετικό δέντρο μορφοποίησης Newick. Το πακέτο phangorn (Schliep et al., 2017) της R προσφέρει πολλές δυνατότητες για φυλογενετικά δέντρα και δίκτυα, συμπεριλαμβανομένων και υλοποιήσεων των αλγορίθμων ιεραρχικής ομαδοποίησης Neighbor Joining (NJ) (Saitou and Nei, 1987) και Unweighted Pair Group Method with Arithmetic Mean (UPGMA) (Sokal and Michener, 1958). Η μέθοδος UPGMA χρησιμοποιήθηκε για τη δημιουργία του δέντρου συσχέτισης των 6933 δειγμάτων με τις εξής εντολές στην R:

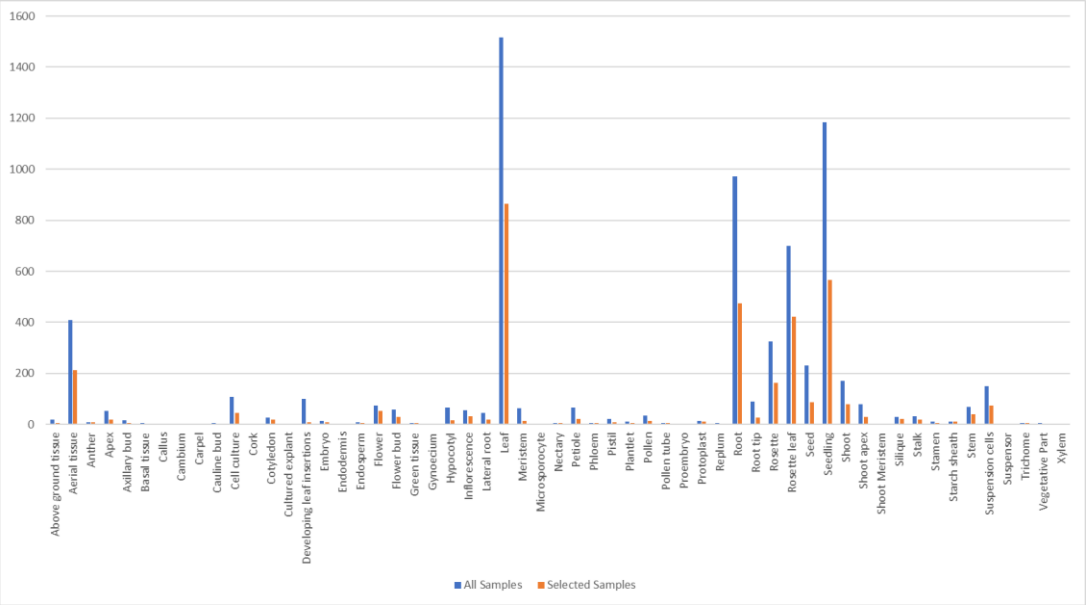
```
library("phangorn")
dm <- as.dist(read.table("samples_R_clean.phylip", sep =
"\t"))
upgma_tree <- upgma(dm)
write.tree(upgma_tree, "samples_R_clean_upgma.new")
```

με τη διαφορά στο αρχείο Phylip, ότι για να διαβαστεί από την R, θα πρέπει η πρώτη του γραμμή να μην είναι ο αριθμός των δειγμάτων αλλά όλα τα ονόματα των δειγμάτων χωρισμένα με tab. Η μετατροπή του έγινε χρησιμοποιώντας PHP.

Το αρχείο `samples_R_clean_upgma.new` είναι το δέντρο συσχέτισης των 6933 δειγμάτων σε μορφοποίηση Newick. Το κάθε φύλλο του δέντρο αντιστοιχεί σε ένα από τα 6933 δείγματα. Για την οπτικοποίηση του δέντρου χρησιμοποιήθηκε το πρόγραμμα προβολής δέντρων Dendroscope (Huson and Scornavacca, 2012). Επίσης με το ίδιο πρόγραμμα έγινε και η ταξινόμηση του δέντρου με αύξουσα (ascending) σειρά.

Για να μπορέσουμε να απομονώσουμε τα πιο αντιπροσωπευτικά δείγματα από τα 6933, χρησιμοποιήθηκε ένα αλγόριθμος αυτόματης κοπής φύλλων που δημιουργήθηκε από τον Δρ Μαλατρά και τροποποιήθηκε από εμάς, ο PhyloPrune. Σε ένα κλάδο με 2 φύλλα ο αλγόριθμος επιλέγει να κόψει το φύλλο με την μικρότερη απόσταση από τον κοινό κόμβο. Στην περίπτωση των δέντρων που έχουν δημιουργηθεί με UPGMA, τα φύλλα καταλήγουν στην ίδια τελικά απόσταση, οπότε σε αυτήν την περίπτωση ο PhyloPrune επιλέγει να κόψει το 1ο από τα 2 φύλλα με ίδια απόσταση στον ίδιο κλάδο. Αποφασίστηκε να μειωθεί ο αριθμός των δειγμάτων στο μισό, δηλαδή από τα 6933, να

παραμείνουν τα 3500 πιο αντιπροσωπευτικά από αυτά (Λίστα των δειγμάτων διαθέσιμη: https://www.michalopoulos.net/act/sample_table.php) (Εικόνα 9). Τα 3500 δείγματα αποτελούν, τελικά, τα μη-μεταλλαγμένα, ξεχωριστού ιστού selected_samples, βάσει των οποίων θα γίνει η ανάλυση γονιδιακής συνέκφρασης για το *Arabidopsis thaliana*.



Εικόνα 9 - Η κατανομή των ιστών πριν και μετά την αυτόματη περικοπή των δειγμάτων για το ACT

Δημιουργία του δέντρου γονιδιακής συσχέτισης γονιδίων

Καθώς μέσω του πίνακα Expression της βάσης δεδομένων έχουμε την έκφραση του κάθε γονιδίου σε κάθε ένα από τα δείγματα, με αντίστοιχο τρόπο με τον υπολογισμό των συσχετίσεων μεταξύ των δειγμάτων, μπορούμε να υπολογίσουμε τον συντελεστή συσχέτισης Pearson μεταξύ του καθενός από τα 21273 γονίδια, δηλαδή τα γονίδια που περιέχονται στο chip της Affymetrix, συγκρίνοντας την έκφραση του κάθε γονιδίου στα 3500 δείγματα που επιλέξαμε εκ των προτέρων. Ο υπολογισμός γίνεται με τον ίδιο τρόπο που περιεγράφηκε στην προηγούμενη ενότητα, μόνο που σε αυτήν την περίπτωση x_i είναι η τιμή έκφρασης ενός συγκεκριμένου γονιδίου σε ένα συγκεκριμένο δείγμα i και y_i είναι η τιμή έκφρασης ενός άλλου γονιδίου στο ίδιο δείγμα, με το n να είναι το σύνολο όλων των επιλεγμένων δειγμάτων. Έτσι παράγεται ένα αρχείο-πίνακας μορφοποίησης Phylip με τα γονίδια και τις τιμές συνέκφρασής τους. Το αρχείο Phylip δημιουργήθηκε χρησιμοποιώντας το pccg.php και είχε 21274 γραμμές

και 21274 στήλες και με συνολικό αριθμό μοναδικών τιμών PCC να είναι 226,259,628.

Για τη δημιουργία του δέντρου γονιδιακής συσχέτισης χρησιμοποιήθηκαν οι ίδιοι αλγόριθμοι και παράμετροι που περιεγράφηκαν στην προηγούμενη ενότητα. Το δέντρο που παράχθηκε (probeset_R_clean_1_minus_d_upgma_ladder.new) αποτελεί το τελικό προϊόν της ανάλυσης γονιδιακής συνέκφρασης για το *Arabidopsis thaliana*. Τα 21273 γονίδια αποτελούν τα φύλλα του δέντρου. Η απόσταση μεταξύ δύο γονιδίων δείχνει πόσο στενά συνδεδεμένη είναι η έκφραση του ενός γονιδίου με το άλλο. Γονίδια που συμμετέχουν σε παρόμοιες βιολογικές διαδικασίες ή ίδια μεταβολικά μονοπάτια βρίσκονται σε κοινούς κλάδους του δέντρου. Για καλύτερη απεικόνισή του, έγινε ταξινόμηση του δέντρου με φθίνουσα σειρά.

Συμφαινετική Συσχέτιση (Cophenetic Correlation)

Ο συμφαινετικός συντελεστής συσχέτισης μετρά πόσο πιστά ένα φυλογενετικό δέντρο διατηρεί τις ανά ζεύγη τιμές του αρχικού πίνακα αποστάσεων από τον οποίο προήλθε το δέντρο μέσω αλγορίθμων ομαδοποίησης (Sokal and Rohlf, 1962; Sokal and Michener, 1958). Χρησιμοποιώντας το ίδιο το δέντρο ως όρισμα, παράγεται ένας πίνακας τιμών που περιέχει τις ανά ζεύγη αποστάσεις, όπως αντιπροσωπεύονται από το δέντρο και ονομάζονται συμφαινετικές αποστάσεις ή συμφαινετικές τιμές. Ένας συμφαινετικός πίνακας αποστάσεων μπορεί να παραχθεί μέσω της συνάρτησης cophenetic του πακέτου stats του πυρήνα της R, από ένα δεδομένο φυλογενετικό δέντρο σε μορφοποίηση Newick, ως εξής:

```
upgma_tree <- read.tree("upgma_tree.new")
upgma_dm<-cophenetic(upgma_tree)
write.table(upgma_dm,file="cophenetic_upgma.dist",sep= "\t")
```

Στη συνέχεια, συγκρίνοντας τις ανά ζεύγη τιμές των ίδιων ζευγών από τον συμφαινετικό και τον αρχικό πίνακα αποστάσεων, υπολογίζεται ο συντελεστής συσχέτισης Pearson, όπου στην συγκεκριμένη περίπτωση ονομάζεται συμφαινετικός συντελεστής συσχέτισης (CPCC). Έτσι μπορούμε να δούμε

πόσο καλά αναπαριστώνται οι τιμές του αρχικού πίνακα αποστάσεων με την μέθοδο ομαδοποίησης που χρησιμοποιήθηκε -στην περίπτωσή μας την UPGMA. Εφαρμόζοντας όσα αναφέρθηκαν παραπάνω για το ACT, ο CPCC του δημιουργημένου με UPGMA δέντρου, υπολογίστηκε 0.592.

Δημιουργία της διαδικτυακής διεπαφής χρήστη

Για την δυνατότητα εύρεσης συνεκφραζόμενων γονιδίων αποφασίστηκε να δημιουργηθεί ένας ιστότοπος που παράλληλα θα προσφέρει και λειτουργία εμπλουτισμού όρων. Η ανάπτυξη του ιστοτόπου έγινε χρησιμοποιώντας την γλώσσα υπερκειμένου HTML5 σε συνδυασμό με την βιβλιοθήκη Bootstrap 4 καθώς και την γλώσσα στυλ CSS και τη γλώσσα JavaScript, πάνω σε έναν Apache εξυπηρετητή. Στον ιστότοπο εκτελούνται κομμάτια κώδικα PHP, η οποία επίσης επιτρέπει και την επικοινωνία με τη βάση δεδομένων χρησιμοποιώντας MySQL ερωτήματα.

Υλοποίηση ανάλυσης υπερεκπροσώπησης όρων

Η ανάλυση εμπλουτισμού (υπερεκπροσώπησης) βιολογικών όρων, μας επιτρέπει να βρούμε κάποια προεξάρχοντα χαρακτηριστικά που απαντώνται συχνότερα σε ομάδες λειτουργικά σχετιζόμενων γονιδίων σε σχέση με το σύνολο των γονιδίων που μελετάμε. Καθώς περιμένουμε ότι τα συνεκφραζόμενα γονίδια συμμετέχουν σε κοινές διεργασίες, τότε πιθανώς να υπάρχουν βιολογικοί όροι που περιγράφουν τις εν λόγω διεργασίες, οι οποίοι να υπερεκπροσωπούνται σε αυτό το υποσύνολο των συνεκφραζόμενων γονιδίων.

Η ανάλυση γίνεται με βάση την υπεργεωμετρική κατανομή. Η υπεργεωμετρική κατανομή είναι μια διακριτή συνάρτηση κατανομής τυχαίας μεταβλητής. Περιγράφει, σε ένα τυχαίο πείραμα με δυο πιθανά αποτελέσματα (επιτυχία - αποτυχία), την πιθανότητα $\Pr(X = k)$ να έχουμε k επιτυχίες σε n δοκιμές, χωρίς επανατοποθέτηση, από έναν πεπερασμένο πληθυσμό μεγέθους N που περιέχει ακριβώς K αντικείμενα με αυτό το συγκεκριμένο χαρακτηριστικό. Στην περίπτωσή μας, για έναν συγκεκριμένο βιολογικό όρο, επιτυχία θεωρείται να επιλέξουμε τυχαία ένα γονίδιο και αυτό το γονίδιο να περιγράφεται από τον εν λόγω όρο. Ο μαθηματικός τύπος είναι:

$$\Pr(X = k) = \frac{\binom{K}{k}\binom{N-k}{n-k}}{\binom{N}{n}}$$

Από τον τύπο προκύπτει μία P-value, η οποία μας δείχνει την πιθανότητα της μηδενικής υπόθεσης, η οποία στην περίπτωσή μας είναι ότι ο βιολογικός όρος που επιλέξαμε εμφανίζεται τυχαία ως υπερεκπροσωπημένος στα συνεκφραζόμενα γονίδια. Οπότε, αν για έναν όρο έχουμε P-value χαμηλότερο από μία ουδό, π.χ. 0.05, τότε απορρίπτεται η μηδενική υπόθεση, άρα ο συγκεκριμένος βιολογικός όρος είναι όντως υπερεκπροσωπημένος στην ομάδα των συνεκφραζόμενων γονιδίων. Επιπλέον, είναι απαραίτητη η διόρθωση των P-values, χρησιμοποιώντας την τροποποίηση False Discovery Rate (FDR) (Benjamini and Hochberg, 1995), όταν εκτελούνται πολλαπλές συγκρίσεις.

Human Gene Coexpression Analysis (HGCA2)

Συλλογή πρωταρχικών δεδομένων

Το εργαλείο HGCA2 πέρα από το γεγονός ότι μελετά την γονιδιακή συνέκφραση στον άνθρωπο, διαφέρει από το ACT στο γεγονός ότι χρησιμοποιούνται δεδομένα έκφρασης γονιδίων που παράχθηκαν από αλληλούχιση επόμενης γενιάς (Next Generation Sequencing - NGS) και πιο συγκεκριμένα από RNASeq. Τα δεδομένα προέρχονται από τη δημόσια βάση δεδομένων GTEx (Gene-Tissue Expression) (GTEx Consortium, 2013) και στην εργασία αυτή θα χρησιμοποιήσουμε τα δευτερογενή επεξεργασμένα δεδομένα που είναι διαθέσιμα στον ιστότοπό τους. Έγινε λήψη:

https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_reads.gct.gz

Το αρχείο RNASeQCv1.1.9_gene_reads.gct περιέχει τις γονιδιακές καταμετρήσεις, δηλαδή τον αριθμό των καταμετρημένων αναγνώσεων που αντιστοιχούν σε κάθε γονίδιο σε κάθε ένα από τα 17382 δείγματα.

https://storage.googleapis.com/gtex_analysis_v8/annotations/GTEx_Analysis_v8_Annotations_SubjectPhenotypesDS.txt

Το αρχείο SubjectPhenotypesDS.txt περιέχει πληροφορίες σχετικά με το φύλο, την ηλικία και τον τρόπο θανάτου των δοτών.

https://storage.googleapis.com/gtex_analysis_v8/annotations/GTEx_Analysis_v8_Annotations_SampleAttributesDS.txt

Το αρχείο SampleAttributesDS.txt περιέχει πληροφορίες σχετικά με τα δείγματα, τους ιστούς από τους οποίους αυτά λαμβάνονται για 714 διαφορετικούς ανθρώπινους οργανισμούς, καθώς και άλλα στοιχεία.

https://www.ebi.ac.uk/arrayexpress/files/E-MTAB-5214/E-MTAB-5214.sdrf.txt

Το αρχείο E-MTAB-5214.sdrf.txt περιέχει τα μεταδεδομένα για κάθε πείραμα που περιλαμβάνεται στη βάση δεδομένων GTEx.

Επίσης έγινε λήψη του αρχείου:

https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm.gct.gz

το οποίο περιέχει τις τιμές TPM των γονιδίων στα 17382 δείγματα. Ανακαλύφθηκαν ~280 γονίδια τα οποία είχαν μηδενική τιμή έκφρασης σε όλα τα δείγματα, οπότε και αποφασίστηκε να αγνοηθούν. Επίσης, από τα 17382 δείγματα, αφαιρέθηκαν αυτά που ήταν κυτταρικές σειρές. Εν τέλει, απέμειναν 16704 δείγματα.

Επεξεργασία των δεδομένων

Έγινε εισαγωγή του αρχείου των γονιδιακών καταμετρήσεων στην R, αφού πρώτα το αρχείο μετατράπηκε σε αντικείμενο του DeSeq2 (Love et al., 2014) και στη συνέχεια έγινε κανονικοποίηση των τιμών έκφρασης χρησιμοποιώντας τη μέθοδο κανονικοποίησης qsmooth του πακέτου YARN (Paulson et al., 2017) της R. Το qsmooth αποτελεί αλγόριθμο διακριτής ανά ιστό κανονικοποίησης ποσοστιμορίου, που έχει δοκιμαστεί και ελεγχθεί πάνω στα δεδομένα από τα δείγματα του GTEx. Τελικά, παράχθηκε το αρχείο qsmooth_expression.txt το οποίο περιέχει την κανονικοποιημένη τιμή έκφρασης κάθε γονιδίου σε κάθε ένα από τα δείγματα.

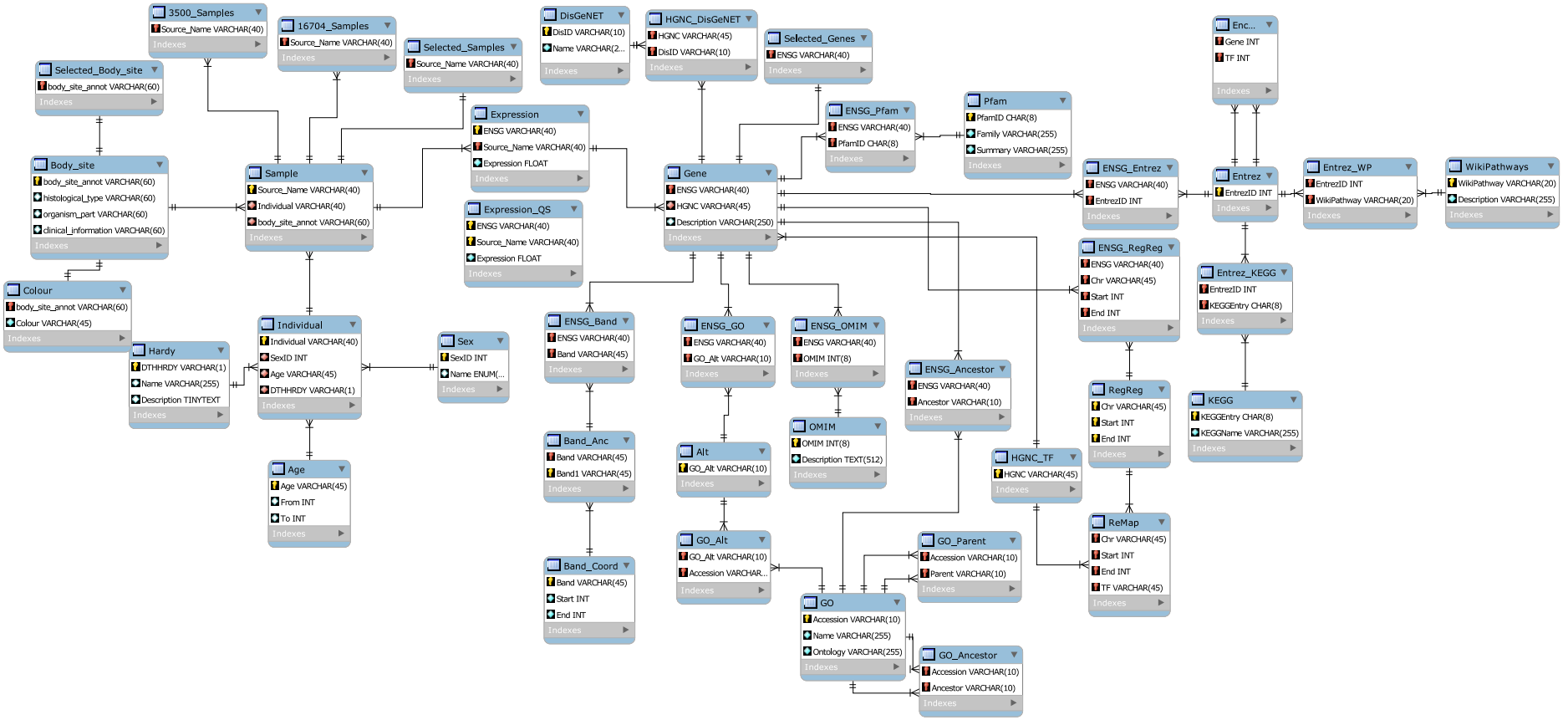
Από το αρχείο SubjectPhenotypesDS.txt δημιουργήθηκαν τα αρχεία Individual.txt, Sex.txt, Age.txt και Hardy.txt. Τα δεδομένα που περιέχονται στο Hardy.txt αναφέρονται στον τρόπο θανάτου του κάθε δότη και στηρίζονται στη χρήση δεδομένων ληφθέντων από την ιστοσελίδα του dbGaP, υπό phs000424.v7 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/variable.cgi?study_id=phs000424.v7.p2&phv=169092). Από το αρχείο SampleAttributesDS.txt δημιουργήθηκε το αρχείο Sample.txt. Τέλος, από το αρχείο E-MTAB-5214.sdrf.txt δημιουργήθηκε το αρχείο Body_site.txt.

Αναζήτηση και ανάκτηση από διάφορες βάσεις δεδομένων όρων που περιγράφουν τα ανθρώπινα γονίδια και επεξεργασία τους

Έγινε αναζήτηση βιολογικών όρων για Οντολογίες Γονιδίων (GeneOntology), Βιολογικών Μονοπατιών (KEGG, Wiki Pathways) Μεταγραφικών Παραγόντων (ENCODE, ReMap) (Cheneby et al., 2020; Consortium, 2011), συσχετίσεις ασθενειών και γονιδίων (OMIM, DisGeNET) (McKusick, 2007; Pinero et al., 2020), Οικογενειών Πρωτεϊνών (Pfam), καθώς και δεδομένα από ζώνες χρωμοσωμάτων και γονιδιακών συντεταγμένων μέσω του NCBI Genome Decoration Page. Ο σχολιασμός των γονιδίων του ανθρώπου έγινε χρησιμοποιώντας το Ensembl και το Genecards (Stelzer et al., 2016).

Κατασκευή βάσης δεδομένων και αποθήκευση των δευτερογενών δεδομένων σε αυτή

Όπως και στην διαδικασία δημιουργίας του ACT, δημιουργήθηκε μία σχεσιακή βάση δεδομένων για το εργαλείο HGCA2, της οποίας το διάγραμμα οντοτήτων-συσχετίσεων απεικονίζεται στην Εικόνα 10. Αρκετοί από τους πίνακες έχουν ακριβώς την ίδια σχεδίαση με τους αντίστοιχους πίνακες στη βάση δεδομένων του ACT, όπως π.χ. ο πίνακας Pfam, οπότε θα αναφερθούμε μόνο στους διαφορετικούς πίνακες της βάσης δεδομένων του HGCA2.



Εικόνα 10 – Το ERD για τη βάση δεδομένων του HGCA2

Η βάση δεδομένων περιέχει τους εξής κύριους πίνακες:

- ***Expression***

  Ο πίνακας Expression περιέχει τις κανονικοποιημένες με την μέθοδο qsmooth τιμές έκφρασης για κάθε γονίδιο σε κάθε δείγμα. Περιλαμβάνει τα χαρακτηριστικά ENSG, που είναι το κατά Ensembl κωδικός του κάθε γονιδίου και Source_Name που είναι ο κωδικός του κάθε δείγματος, τα οποία είναι πρωτεύοντα κλειδιά. Τέλος, έχει το χαρακτηριστικό Expression που είναι η αριθμητική τιμή έκφρασης γονιδίου.

- ***Sample***

  Ο πίνακας Study περιέχει πληροφορίες για τα δείγματα της GTEx. Περιλαμβάνει το Source_Name που είναι ο κωδικός του κάθε δείγματος και αποτελεί πρωτεύον κλειδί, Individual που είναι ο κωδικός του ατόμου από το οποίο προήλθε το δείγμα και Body_site_annotation που είναι ο ιστός στον οποίον ανήκει το δείγμα.

- ***Selected_Samples***

  Ο πίνακας Selected_Samples περιέχει τα δείγματα που τελικώς επιλέχθηκαν για την ανάλυση συνέκφρασης.

- ***Individual***

  Ο πίνακας Individual περιέχει πληροφορίες για τα άτομα-δότες των δειγμάτων του GTEx. Περιλαμβάνει τα χαρακτηριστικά Individual (πρωτεύον κλειδί) που είναι ένας μοναδικός κωδικός για κάθε άτομο, SexID, που είναι το φύλο (male/female), Age (ηλικία θανάτου σε ένα εύρος 10 ετών) και DTHHRDY που δηλώνει πόσο βίαιος ήταν ο θάνατος.

- ***Hardy***

  Ο πίνακας Hardy περιέχει πληροφορίες για την κλίμακα θανάτου Hardy. Περιλαμβάνει τα χαρακτηριστικά DTHHRDY (πρωτεύον κλειδί), Name και Description, που εξηγούν περαιτέρω τις κατηγορίες θανάτου.

- ***Age***

  Ο πίνακας Age περιέχει τις κατηγορίες της ηλικίας για τους δότες. Περιλαμβάνει το AgeID (πρωτεύον κλειδί) και τα χαρακτηριστικά From και To που είναι το κατώτερο και ανώτερο όριο ηλικίας σε κάθε κατηγορία.

- ***Sex***

  Ο πίνακας Sex περιέχει το φύλο για το δότη. Περιλαμβάνει το SexID (πρωτεύον κλειδί) και το Name. Στο SexID = 1 αντιστοιχεί το male και στο SexID = 2 το female.

- ***Body_site***

  Ο πίνακας Body_site περιέχει τους ιστούς από τους οποίους προέρχονται τα δείγματα του GTEx. Περιλαμβάνει τα χαρακτηριστικά body_site_annot (πρωτεύον κλειδί) που είναι το όνομα του ιστού, histological type που είναι η κατηγορία στην οποία ανήκει ο ιστός, organism_part που είναι το σημείο του σώματος που προήλθε και clinical information που είναι μία επιπλέον πληροφορία/επεξήγηση για την κατάσταση του ιστού.

- ***Selected_Body_site***

  Ο πίνακας Selected_Body_site περιέχει μόνο τους ιστούς που αποφασίσαμε να κρατήσουμε στη βάση δεδομένων. Επειδή υπήρχε διαφορά στην ονομασία των ιστών στις διαφορετικές εκδόσεις των αρχείων του GTEx, επιλέξαμε τους ιστούς που αντιστοιχούσαν στο σχολιασμό των δειγμάτων.

- ***Gene***

  Ο πίνακας Gene περιέχει πληροφορίες για τα ~55000 γονίδια που μελετώνται από το GTEx και για τα οποία έχουμε τιμές έκφρασης. Περιλαμβάνει τα χαρακτηριστικά ENSG (πρωτεύον κλειδί), HGNC που είναι το κατά HGNC σύμβολο του γονιδίου και Description που είναι αναλυτική περιγραφή του γονιδίου.

- ***OMIM***

  Ο πίνακας OMIM περιέχει τις ασθένειες από τη βάση δεδομένων OMIM. Περιλαμβάνει τα χαρακτηριστικά OMIM (πρωτεύον κλειδί) που είναι ο μοναδικός κωδικός της OMIM για κάθε ασθένεια και Description που είναι η περιγραφή της ασθένειας.

- ***DisGeNET***

  Ο πίνακας DisGeNET περιέχει περιέχει τις ασθένειες από τη βάση δεδομένων DisGeNET. Περιλαμβάνει τα χαρακτηριστικά DisID (πρωτεύον κλειδί) που είναι ο μοναδικός κωδικός της DisGeNET για κάθε ασθένεια και Name που είναι το πλήρες όνομα της ασθένειας.

- ***Encode***

  Ο πίνακας Encode περιέχει τους μεταγραφικούς παράγοντες και τα γονίδια-στόχους τους από την Encode. Περιλαμβάνει τα χαρακτηριστικά Gene και TF ως πρωτεύοντα κλειδιά που είναι το γονίδιο και ο μεταγραφικός παράγοντας που το στοχεύει, αντίστοιχα.

- ***RegReg***

  Ο πίνακας RegReg περιέχει τις μεταγραφικά ενεργές περιοχές του γονιδιώματος όπως προήλθαν από τη βάση δεδομένων ReMap. Περιλαμβάνει τα χαρακτηριστικά Chr, Start και End ως πρωτεύοντα κλειδιά τα οποία αντιπροσωπεύουν τις γονιδιακές συντεταγμένες (χρωμόσωμα, αρίθμηση αρχικού και τελικού νουκλεοτιδίου) της περιοχής.

- ***ReMap***

  Ο πίνακας ReMap περιέχει τους μεταγραφικούς παράγοντες και τις περιοχές που στοχεύουν, όπως προήλθαν από τη βάση δεδομένων ReMap. Περιλαμβάνει τα χαρακτηριστικά Chr, Start, End και TF ως πρωτεύοντα κλειδιά τα οποία αντιπροσωπεύουν τις γονιδιακές συντεταγμένες για κάθε μεταγραφικό παράγοντα TF.

- ***ENSG_RegReg***

  Ο πίνακας ENSG_RegReg περιέχει τα γονίδια και τις περιοχές τους πάνω στις οποίες προσδένονται μεταγραφικοί παράγοντες, όπως προέκυψαν από τα δεδομένα της βάσης ReMap. Περιλαμβάνει τα χαρακτηριστικά ENSG, Chr, Start και End ως πρωτεύοντα κλειδιά τα οποία αντιπροσωπεύουν τις γονιδιακές συντεταγμένες των περιοχών κάθε γονιδίου, όπου αποτελούν στόχους μεταγραφικών παραγόντων.

Οι υπόλοιποι πίνακες που εμφανίζονται στο διάγραμμα αποτελούν ενδιάμεσους πίνακες που συσχετίζουν τα δεδομένα των πινάκων που ενώνουν, είτε είναι ταυτόσημοι με πίνακες που είχαν αναφερθεί στην ενότητα του ACT. Σε κάθε πίνακα έγινε εισαγωγή του αρχείου .txt με το αντίστοιχο όνομα, όπου η δημιουργία των flat files περιγράφεται παρακάτω.

Λεπτομερής ανάλυση τρόπου ανάκτησης δεδομένων από τις αντίστοιχες βάσεις

Ensembl BioMart

Το Ensembl BioMart (Kinsella et al., 2011) αποτελεί ένα εργαλείο εξόρυξης δεδομένων από τη βάση δεδομένων γονιδίων Ensembl. Το εργαλείο αυτό μας επιτρέπει να κατεβάσουμε πληροφορίες σχετικά με μία λίστα γονιδίων που δίνεται από τον χρήστη. Προσφέρει μία μεγάλη ποικιλία πληροφοριών τόσο από την ίδια την Ensembl, όσο και από εξωτερικές πηγές. Χάρις τη δυνατότητα σύνδεσης με αυτές τις εξωτερικές πηγές, είναι δυνατόν να παράγουμε λίστες που συνδέουν το ENSG ID του κάθε γονιδίου με το αντίστοιχο χαρακτηριστικό, π.χ. τις οντολογίες γονιδίων που τον περιγράφουν ή τα ονόματα HGNC του κάθε γονιδίου. Επιπλέον, μπορούμε να παράγουμε αυτή τη λίστα με τις ίδιες ακριβώς παραμέτρους προγραμματιστικά, χωρίς να την κατεβάσουμε τοπικά, χρησιμοποιώντας τη λειτουργία XML που προσφέρεται από το BioMart.

Μέσω του BioMart, έγινε η λήψη όλων των μοναδικών κωδικών ENSG ID για τα γονίδια του ανθρώπου, το σύμβολό τους κατά HGNC, το όνομα του γονιδίου και η περιγραφή τους, δημιουργώντας το αρχείο ENSG.txt.

Gene Ontology

Οι οντολογίες γονιδίων που περιγράφουν τα γονίδια του ανθρώπου λήφθηκαν από τον ιστότοπο του geneontology.com, με αντίστοιχο τρόπο όπως περιεγράφηκε στην ενότητα λήψης δεδομένων για το ACT.

Δημιουργήθηκαν τα αρχεία GO.txt και GO_Alt.txt. Το αρχείο που συνδέει το κάθε γονίδιο με κάποια οντολογία (ENSG_GO.txt) δημιουργήθηκε χρησιμοποιώντας την λειτουργία xml από το Ensembl Biomart.

KEGG

Έγινε αναζήτηση και λήψη όλων των διαθέσιμων μονοπατιών για το *Homo sapiens*, που έχει τον μοναδικό κωδικό οργανισμού T01001 στη βάση δεδομένων, στο αρχείο KEGG.txt μέσω του συνδέσμου:

https://www.genome.jp/dbget-bin/get_linkdb?-t+pathway+gn:T01001

Τα δεδομένα συνδυασμού γονιδίων του ανθρώπου και μονοπατιών της KEGG, λήφθηκαν πάλι μέσα από τον ιστότοπο της KEGG, αλγοριθμικά χρησιμοποιώντας κώδικα PHP, στο αρχείο Entrez_KEGG.txt.

WikiPathways

Έγινε λήψη των δεδομένων βιολογικών μονοπατιών για τον άνθρωπο από το σύνδεσμο:

http://data.wikipathways.org/current/gmt/wikipathways-20200810-gmt-Homo_sapiens.gmt

Επειδή τα δεδομένα από την WikiPathways χρησιμοποιούν τους κωδικούς Entrez για κάθε γονίδιο, έγινε πρώτα λήψη πληροφοριών για τα όλα γονίδια από τη βάση δεδομένων Entrez καθώς και η αντιστοίχισή τους με τους κωδικούς από την Ensembl, μέσω του Biomart. Παράχθηκαν τα αρχεία, Entrez.txt, Entrez_ENSG.txt, WikiPathways.txt και Entrez_WP.txt.

ENCODE

Το επιστημονικό πρόγραμμα ENCODE αποτελεί συνεργασία από πολλές ερευνητικές ομάδες, που ασχολείται με την ανακάλυψη ρυθμιστικών στοιχείων στο ανθρώπινο γονιδίωμα, όπως μεταγραφικούς παράγοντες. Μέσω του ιστότοπου του Harmonizome (Rouillard et al., 2016), έγινε λήψη ήδη επεξεργασμένων δεδομένων γονιδίων και των αντίστοιχων μεταγραφικών

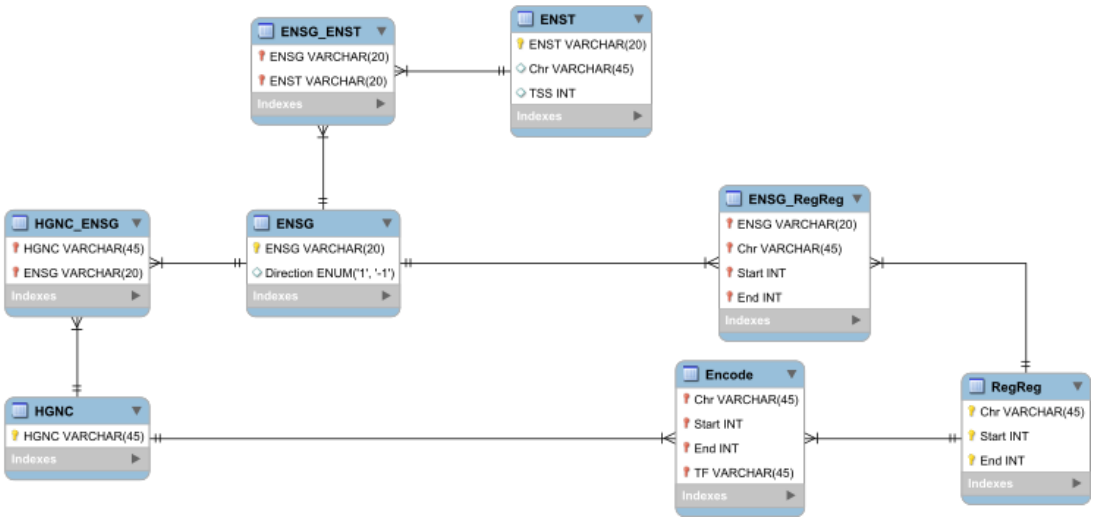
παραγόντων που τα στοχεύουν που προέρχονται από το Encode. Δημιουργήθηκε το αρχείο Encode.txt μετά την κατάλληλη επεξεργασία των δεδομένων.

ReMap

Η βάση δεδομένων ReMap αποτελεί ένα εκτενές καταθετήριο για ρυθμιστικές περιοχές στο ανθρώπινο γονιδίωμα. Για την τελευταία έκδοση του ReMap 2020, δεν υπάρχουν ήδη επεξεργασμένα δεδομένα που να περιέχουν μεταγραφικούς παράγοντες και τα γονίδια στόχους τους σε απλή μορφή. Αρχικά, έγινε η λήψη του:

http://remap.univ-amu.fr/storage/remap2020/hg38/MACS2/remap2020_crm_macs2_hg38_v1_0.bed.gz

το οποίο περιέχει όλες τις μεταγραφικώς ρυθμιστικές περιοχές των γονιδίων. Στη συνέχεια, έγινε η λήψη από το Ensembl Biomart όλων των περιοχών έναρξης και λήξης μεταγραφής των διαθέσιμων γονιδίων. Τέλος, επειδή τα ονόματα των μεταγραφικών παραγόντων είναι γραμμένα σε κανονικά ονόματα κατά HGNC στα αρχεία του ReMap, έγινε λήψη και αρχείου αντιστοίχισής τους με τους κωδικούς ENSG. Δημιουργήθηκε μία βάση δεδομένων (Εικόνα 11) ώστε να αποθηκευτούν όλα δεδομένα που αναφέραμε και να γίνεις δυνατός ο υπολογισμός των γονιδίων-στόχων ανά μεταγραφικό παράγοντα.



Εικόνα 11 – Το ERD για τη βάση δεδομένων του ReMap

Ο πίνακας RegReg περιέχει όλες τις περιοχές (συντεταγμένες του γονιδιώματος) που προσδένονται σε αυτές μεταγραφικοί παράγοντες. Ο πίνακας Encode προέρχεται από το αρχείο που κατεβάσαμε αφού πρώτα επεξεργαστήκαμε προγραμματιστικά τα δεδομένα ώστε να τα φέρουμε σε κατάλληλη μορφή αρχείου. Ο πίνακας ENSG_RegReg περιέχει τις μεταγραφικά ενεργές περιοχές του κάθε γονιδίου. Ο πίνακας ENST περιέχει τα διαφορετικά μετάγραφα του κάθε γονιδίου και τις περιοχές έναρξης μεταγραφής τους. Οι υπόλοιποι πίνακες περιέχουν πληροφορίες για τα διάφορα ονόματα των γονιδίων. Για τον υπολογισμό του αρχείου που εισέρχεται στον πίνακα ENSG_RegReg, χρησιμοποιήθηκε η εξής εντολή SQL:

```
select distinct ENSG_ENST.ENSG, RegReg.Chr, Start, End from ENST, ENSG_ENST, RegReg where ENST.ENST = ENSG_ENST.ENST and ENST.Chr=RegReg.Chr and TSS>=Start and TSS<=End
```

Πρακτικά ελέγχουμε αν οι περιοχές που αποτελούν στόχο μεταγραφικών παραγόντων περιέχουν την περιοχή έναρξης μεταγραφής του κάθε γονιδίου. Δημιουργήθηκαν τα αρχεία, RegReg.txt, ReMap.txt και ENSG_RegReg.txt.

Η βάση δεδομένων που δημιουργήσαμε είχε ως μοναδικό σκοπό την παραγωγή των παραπάνω αρχείων και δεν χρησιμοποιείται περαιτέρω.

OMIM

Η βάση δεδομένων OMIM (Online Mendelian Inheritance in Man) περιέχει πληροφορίες σύνδεσης γονιδίων με γενετικές ασθένειες. Έγινε λήψη όλων των εγγραφών των ασθενειών και όλες οι συσχετίσεις ασθενειών και γονιδίων. Παράχθηκαν τα αρχεία OMIM.txt και ENSG_OMIM.txt

DisGeNET

Η βάση δεδομένων DisGeNET περιέχει συσχετίσεις γονιδίων και πολυμορφισμών με ασθένειες. Έγινε λήψη των δεδομένων συσχετίσεων γονιδίων και ασθενειών από το σύνδεσμο:

https://www.disgenet.org/static/disgenet_ap1/files/downloads/all_gene_disease_associations.tsv.gz

και κρατώντας τα στοιχεία που χρειαζόμαστε από αυτό, δημιουργήθηκαν τα αρχεία DisGeNET.txt και HGNC_DisGeNET.txt.

Chromosome Bands

Οι γονιδιακές συντεταγμένες ενός γονιδίου μπορούν να περιγραφούν με τις ζώνες του χρωμοσώματος που ανήκουν. Έγινε λήψη 3 αρχείων
ideogram_9606_GCF_000001305.15_850_V1
ideogram_9606_GCF_000001305.15_550_V1
ideogram_9606_GCF_000001305.15_400_V1
από τον εξυπηρετητή FTP του NCBI Genome Decoration στον σύνδεσμο: ftp://ftp.ncbi.nlm.nih.gov/pub/gdp/
τα οποία χωρίζουν τα χρωμοσώματα σε 850, 550, και 400 ζώνες αντίστοιχα, ανάλογα με το πόσο λεπτομερείς θέλουμε να είμαστε. Οι επιπλέον ζώνες που περιλαμβάνονται στο αρχείο των 850, είναι υποσύνολα των ζωνών του αρχείου των 550 και το αρχείο των 400 έχεις τις πιο γενικές ζώνες. Έγινε λήψη των ζωνών που ανήκουν τα διαθέσιμα γονίδια από το Ensembl Biomart. Επίσης, κάθε ειδικότερη ζώνη αντιστοιχήθηκε με την αμέσως γενικότερή της. Ακόμη, κάθε ζώνη αντιστοιχήθηκε με τις αντίστοιχες γονιδιακές συντεταγμένες που αντιπροσωπεύει. Τελικά δημιουργήθηκαν τα αρχεία ENSG_Band.txt, Band_Coord.txt και Band_Anc.txt.

Pfam

Με ίδιο τρόπο με το ACT, έγινε λήψη των δεδομένων από το σύνδεσμο:

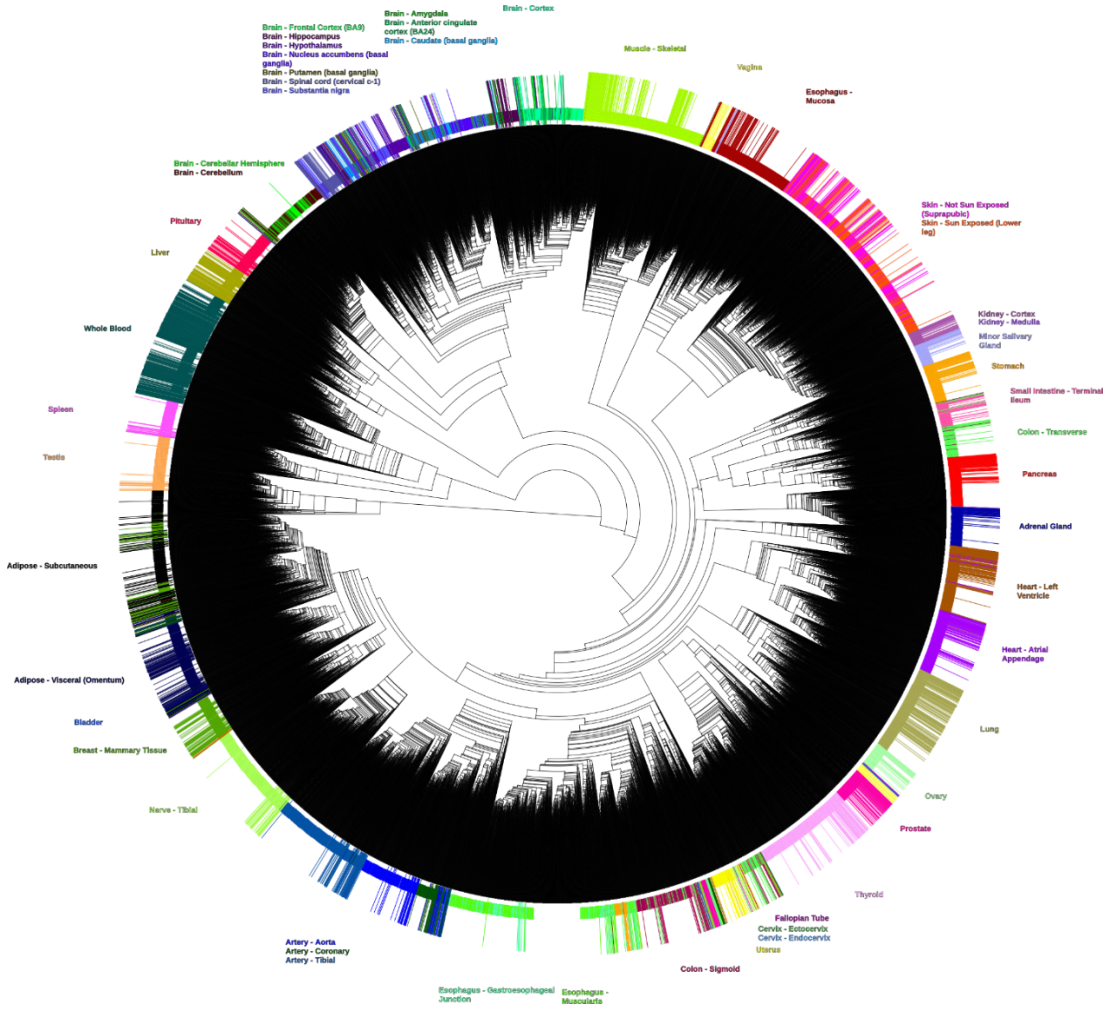
ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/database_files/pfamA.txt.gz

και παράχθηκαν τα αρχεία Pfam.txt και ENSG_Pfam.txt

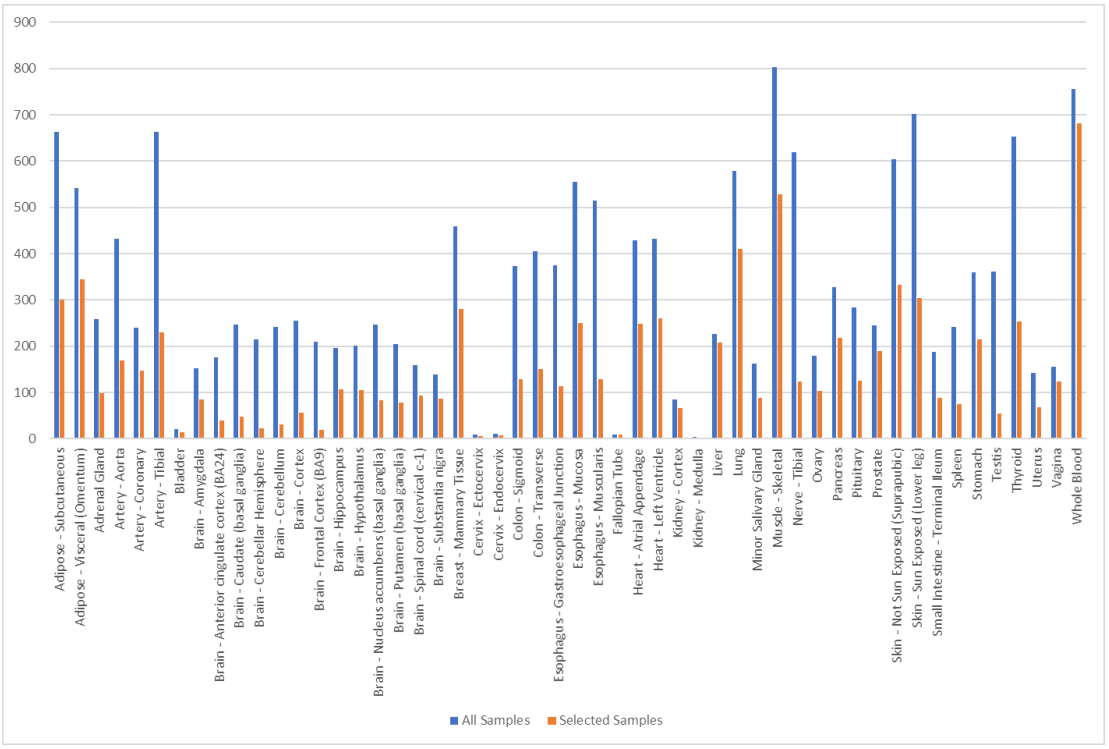
Δημιουργία του δέντρου συσχέτισης δειγμάτων και αυτόματη επιλογή των αντιπροσωπευτικών δειγμάτων

Με παρόμοιο τρόπο, όπως περιγράφηκε στην αντίστοιχη ενότητα για το εργαλείο ACT, έγινε η δημιουργία του πίνακα αποστάσεως σε μορφοποίηση Phylip των δειγμάτων του GTEx. Το αρχείο αυτό απότελείται από 16705 γραμμές (16704 δείγματα, μία γραμμή για το καθένα, και η πρώτη γραμμή με τον αριθμό των στοιχείων). Μετά την κατάλληλη μετατροπή του αρχείου, έγινε η δημιουργία του δέντρου συσχέτισης δειγμάτων, μέσω του πακέτου phangorn της R, χρησιμοποιώντας τον αλγόριθμο UPGMA. Στη συνέχεια, χρησιμοποιώντας τον αυτόματο αλγόριθμο αποκοπής φύλλων για UPGMA του PhyloPrune, έγινε η αυτόματη περικοπή του δέντρου δειγμάτων από 16704 στα

3500 πιο αντιπροσωπευτικά δείγματα (Εικόνα 12, Εικόνα 13). Στη συνέχεια, αυτά τα 3500 δείγματα εισήχθησαν στον πίνακα Selected_Samples της βάσης δεδομένων.



Εικόνα 12 – Το δέντρο συσχέτισης δειγμάτων. Τα φύλλα που συμβολίζονται με μακριές γραμμές είναι τα 3500 δείγματα που έμειναν μετά την αυτόματη περικοπή



Εικόνα 13 – Η κατανομή των ιστών πριν και μετά την αυτόματη περικοπή των δειγμάτων

Δημιουργία του δέντρου γονιδιακής συσχέτισης γονιδίων

Η διαδικασία έγινε παρομοίως με την αντίστοιχη του ACT. Μέσω του πίνακα Expression της βάσης δεδομένων έγινε η δημιουργία του πίνακα συνέκφρασης γονιδίων, συγκρίνοντας την έκφραση του κάθε ενός από τα 55431 γονίδια στα 3500 δείγματα που επιλέξαμε εκ των προτέρων. Το αρχείο Phylip που δημιουργήθηκε είχε 55432 γραμμές.

Για τη δημιουργία του δέντρου γονιδιακής συσχέτισης χρησιμοποιήθηκαν οι ίδιοι αλγόριθμοι και παράμετροι που περιεγράφηκαν στην προηγούμενη ενότητα. Το δέντρο που παράχθηκε (qsmooth_genes_one_minus_d_upgma_ladder.new) αποτελεί το τελικό προϊόν της ανάλυσης γονιδιακής συνέκφρασης για τον άνθρωπο. Τα 55431 γονίδια αποτελούν τα φύλλα του δέντρου. Για καλύτερη απεικόνισή του, έγινε ταξινόμηση του δέντρου με φθίνουσα σειρά. Επειδή το συγκεκριμένο δέντρο έχει πολύ μεγάλο μέγεθος για να διαβαστεί από το Dendroscope, η ταξινόμησή του έγινε χρησιμοποιώντας το πρόγραμμα Newick Utilities (Junier and Zdobnov, 2010).

Η διαδικτυακή επαφή για το HGCA2 και υλοποίηση της ανάλυσης υπερεκπροσώπησης όρων έγιναν με παρόμοιο τρόπο με του ACT.

Microarray Human Gene Coexpression Analysis (HGCA1.5)

Μετά τη δημιουργία του εργαλείου HGCA2 χρησιμοποιώντας τα δευτερογενή δεδομένα RNASeq από τη βάση δεδομένων GTEx, αποφασίστηκε να δημιουργηθεί και ένα τρίτο εργαλείο το οποίο θα έχει ακριβώς τα ίδια χαρακτηριστικά με το HGCA2, αλλά η ανάλυση συνέκφρασης θα βασίζεται σε δεδομένα μικροσυστοιχιών. Το παρόν εργαλείο αποτελεί πρακτικά μία αναβάθμιση στο ήδη υπάρχον εργαλείο Human Gene Coexpression Analysis που δημιουργήθηκε το 2012 (Michalopoulos et al., 2012) και είναι διαθέσιμο στον ιστότοπο:

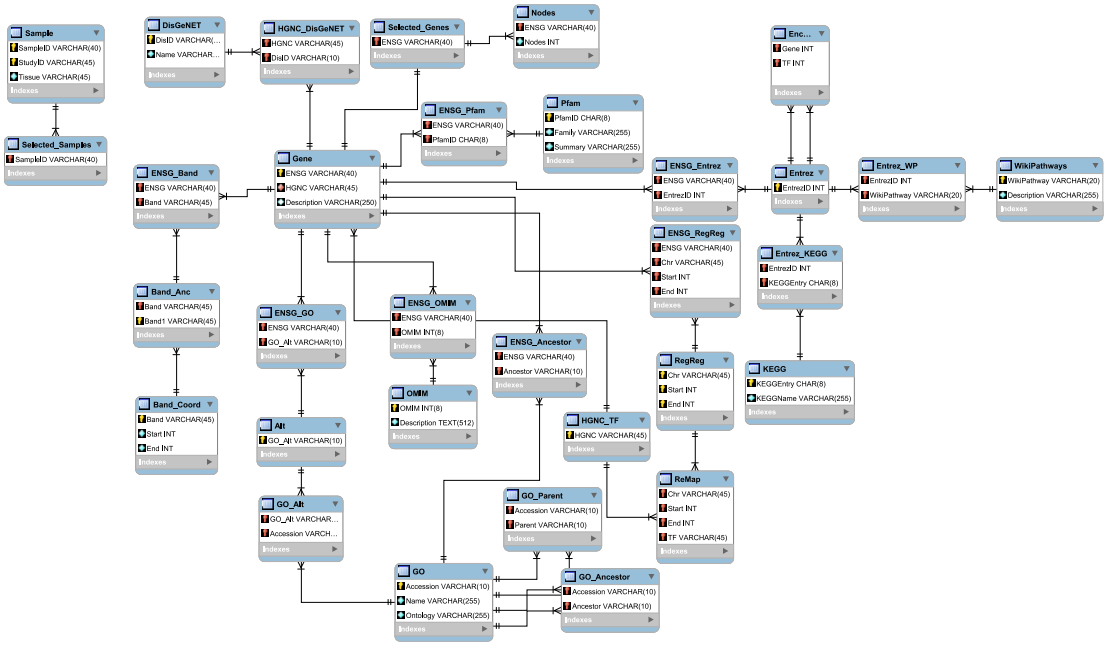
https://www.michalopoulos.net/coexpression/

Συλλογή πρωταρχικών δεδομένων και κανονικοποίησή τους

Ο κατάλογος των δειγμάτων αναφέρεται στο συμπληρωματικό υλικό της δημοσίευσης. Πρόκειται για 1959 δείγματα που προέρχονται από το chip Human Genome U133 Plus 2.0 της Affymetrix. Ενώ στην αρχική ανάλυση είχε χρησιμοποιηθεί ο αλγόριθμος κανονικοποίησης MAS5.0 σε συνδυασμό με το default CDF που δίνεται από την Affymetrix, σε αυτήν την ανάλυση, ακολουθείται η ίδια διαδικασία κανονικοποίησης με αυτή που αναφέρθηκε στην ενότητα για το ACT. Έγινε λήψη και χρήση της έκδοσης 24 για το CDF της Brainarray για το Human Genome U133 Plus 2.0 array chip και τα δεδομένα από τα 1959 δείγματα κανονικοποιήθηκαν χρησιμοποιώντας τον αλγόριθμο SCAN. Τελικά παράχθηκε το αρχείο SCAN_matrix_1959.txt που περιέχει τις τιμές έκφρασης κάθε ενός από τα 20001 γονίδια σε κάθε ένα από τα 1959 δείγματα. Στα 20001 γονίδια περιλαμβάνονται 62 γονίδια/ομάδες ανιχνευτών ελέγχου από την Affymetrix, που αναγνωρίζονται με το πρόθεμα AFFX. Επίσης 6 γονίδια αναφέρονται σε παλαιότερη έκδοση του γονιδιώματος και θεωρούνται παρωχημένα. Και στις 2 περιπτώσεις αγνοήσαμε τα εν λόγω γονίδια. Εν τέλει, τα επιλεγμένα γονίδια είναι 19933.

Δημιουργία βάσης δεδομένων

Η βάση δεδομένων για το εργαλείο HGCA1.5 (Εικόνα 14) με βάση τις μικροσυστοιχίες, αποτελεί πιστό αντίγραφο της βάσης δεδομένων για το HGCA2 με βάση τα RNASeq, με διαφορές στα δείγματα και τα γονίδια που περιγράφονται.



Εικόνα 14 – Το ERD για τη βάση δεδομένων του HGCA1.5

Δημιουργία του δέντρου γονιδιακής συσχέτισης γονιδίων

Η διαδικασία έγινε με παρόμοιο τρόπο με την αντίστοιχη του ACT και του HGCA2. Μέσω του πίνακα Expression της βάσης δεδομένων έγινε η δημιουργία του πίνακα συνέκφρασης γονιδίων, συγκρίνοντας την έκφραση του κάθε ενός από τα 19933 γονίδια στα 1959 δείγματα που είχαν επιλεγεί. Το αρχείο Phylip που δημιουργήθηκε είχε 19934 γραμμές.

Για τη δημιουργία του δέντρου γονιδιακής συσχέτισης χρησιμοποιήθηκαν οι ίδιοι αλγόριθμοι και παράμετροι που περιεγράφηκαν στην προηγούμενη ενότητα. Το δέντρο που παράχθηκε (upgma_19933_1959_one_minus_d_ladder.new) αποτελεί το τελικό προϊόν της ανάλυσης γονιδιακής συνέκφρασης για τον άνθρωπο με βάση πειράματα μικροσυστοιχιών σε αυτήν την περίπτωση. Τα 19933 γονίδια αποτελούν τα φύλλα του δέντρου. Για καλύτερη απεικόνισή του, έγινε ταξινόμηση του δέντρου με φθίνουσα σειρά.

Η διαδικτυακή διεπαφή για το HGCA1.5 και η υλοποίηση της ανάλυσης υπερεκπροσώπησης όρων έγιναν με παρόμοιο τρόπο με του ACT και του HGCA2.

Αποτελέσματα

Τα εργαλεία ACT, HGCA2 και HGCA1.5 είναι πλήρως λειτουργικά και διαθέσιμα στις παρακάτω διευθύνσεις:

ACT: https://www.michalopoulos.net/act/

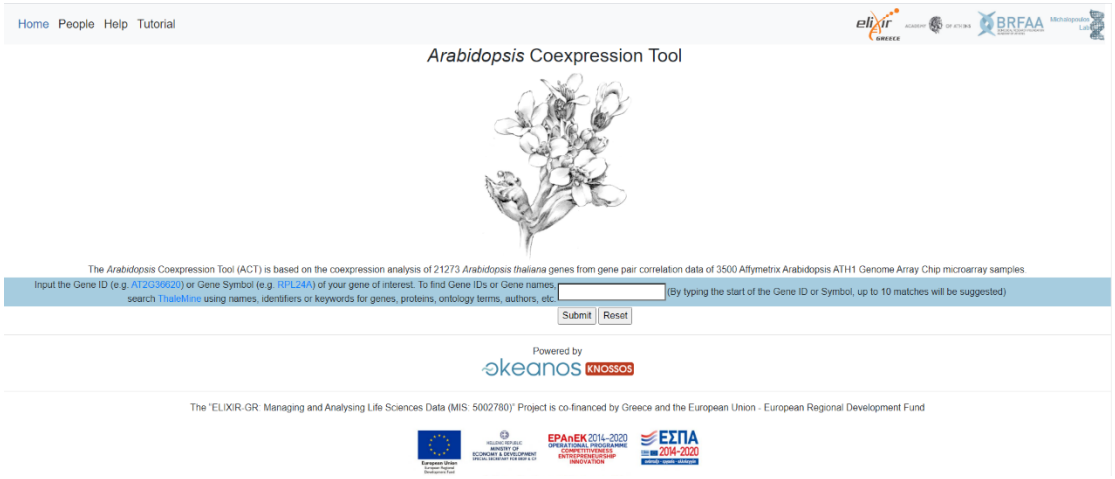
HGCA2: https://www.michalopoulos.net/hgca2.0/

HGCA1.5: https://www.michalopoulos.net/hgca1.5/

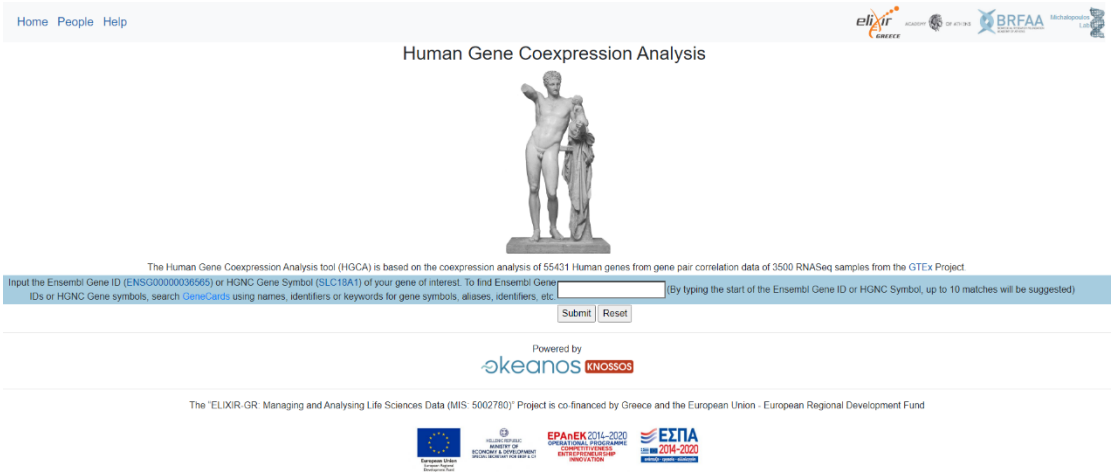
Περιγραφή των ιστοτόπων

Αρχικές σελίδες

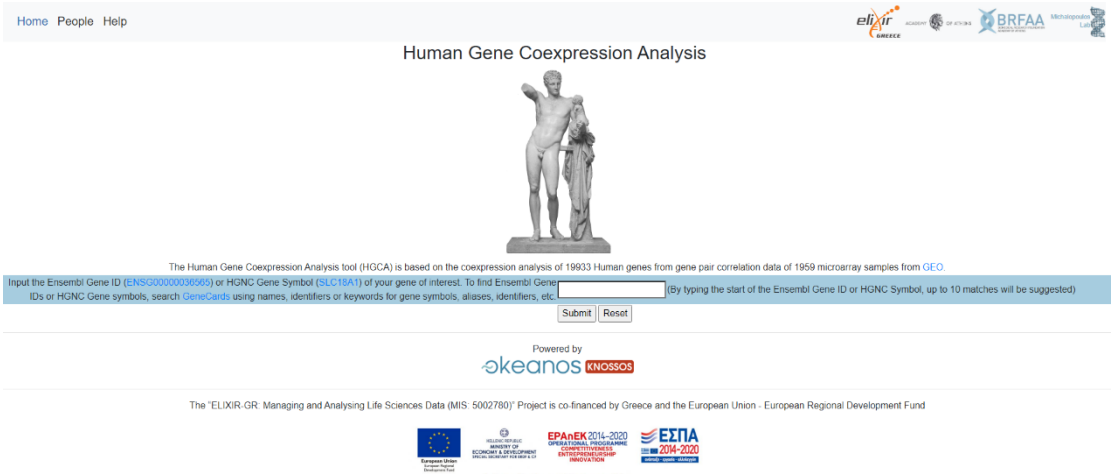
Η πρώτη σελίδα που συναντάει ο χρήστης μόλις επισκεφθεί την αντίστοιχη διεύθυνση, περιέχει τον τίτλο του αντίστοιχου εργαλείου καθώς και μία καλλιτεχνική εικόνα που είναι σχετική με τον οργανισμό που μελετάται, με το ACT (Εικόνα 15) να απεικονίζει ένα σκίτσο του *Arabidopsis thaliana* και τα HGCA να απεικονίζουν τον Ερμή του Πραξιτέλη (Εικόνα 16, Εικόνα 17). Αναφέρονται, ο συνολικός αριθμός των γονιδίων και δειγμάτων που χρησιμοποιήθηκαν στην ανάλυση, καθώς και το chip ή η βάση δεδομένων που προέρχονται τα δείγματα.



Εικόνα 15 – Η αρχική σελίδα του ACT

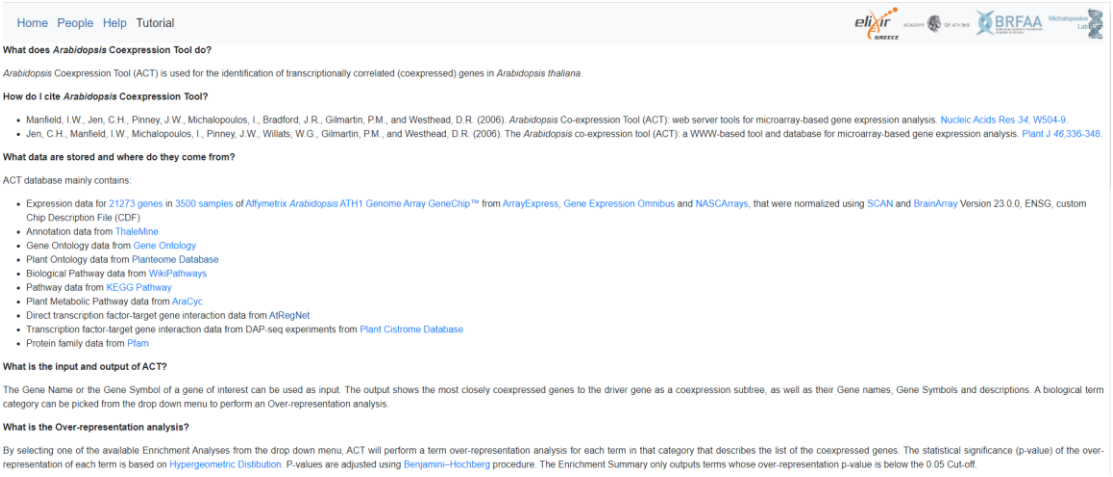


Εικόνα 16 – Η αρχική σελίδα του HGCA2



Εικόνα 17 – Η αρχική σελίδα του HGCA1.5

Στο πάνω μέρος της σελίδας υπάρχει μία μπάρα πλοήγησης με τους εξής συνδέσμους: Home, People, Help και στην περίπτωση του ACT, Tutorial. Κάνοντας κλικ πάνω στον σύνδεσμο Home, μεταφερόμαστε στην αρχική σελίδα του εργαλείου. Στη σελίδα People, αναφέρονται όλοι οι συντελεστές και συνεργάτες που δούλεψαν πάνω στο αντίστοιχο εργαλείο καθώς και στοιχεία επικοινωνίας για τον κύριο επιβλέποντα της εργασίας. Η σελίδα Help (Εικόνα 18) αναφέρει βασικές πληροφορίες και απαντάει σε απορίες σχετικά με την λειτουργία του ιστότοπου, καθώς και για την προέλευση των δεδομένων από τις διάφορες διαδικτυακές βάσεις.



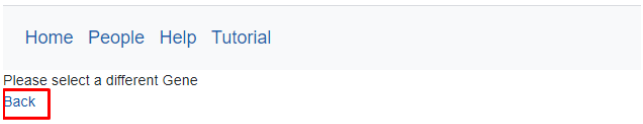
Εικόνα 18 – Απόσπασμα της σελίδας Help για το ACT

Στο Tutorial για το ACT (Εικόνα 19) περιέχεται ένα βίντεο στα αγγλικά, με οδηγίες βήμα βήμα για την εκτέλεση μίας ανάλυσης του ACT.

Εικόνα 19 – Το βίντεο οδηγιών για το ACT

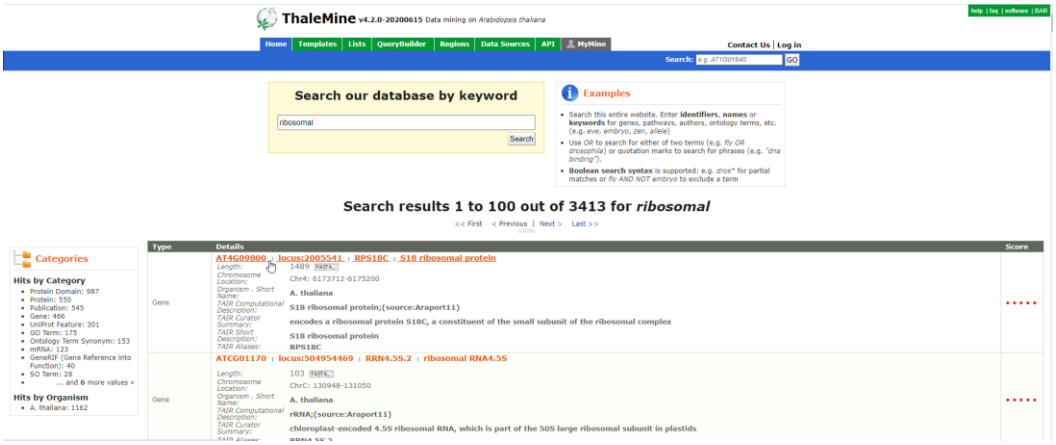
Αναζήτηση γονιδίου

Στην αρχική σελίδα υπάρχει ένα πεδίο αναζήτησης γονιδίων. Ο χρήστης μπορεί να πληκτρολογήσει τον κωδικό ENSG ή το κατά HGNC σύμβολο ενός γονιδίου για τον άνθρωπο, ή τον κωδικό AGI ή το σύμβολο ενός γονιδίου για το *Arabidopsis thaliana*. Αν το γονίδιο είναι διαθέσιμο, τότε η φόρμα θα προτείνει αυτόματη συμπλήρωση του πεδίου. Πατώντας το πλήκτρο Submit, επιτελείται η διαδικασία της ανάλυσης συνέκφρασης. Σε περίπτωση που το γονίδιο που συμπληρωθεί δεν υπάρχει ή έχουν πληκτρολογηθεί λάθος στοιχεία, τότε μετά το πάτημα του Submit, εμφανίζεται μήνυμα λάθους (Εικόνα 20) και ο χρήστης καλείται να επιστρέψει στην αρχική σελίδα.



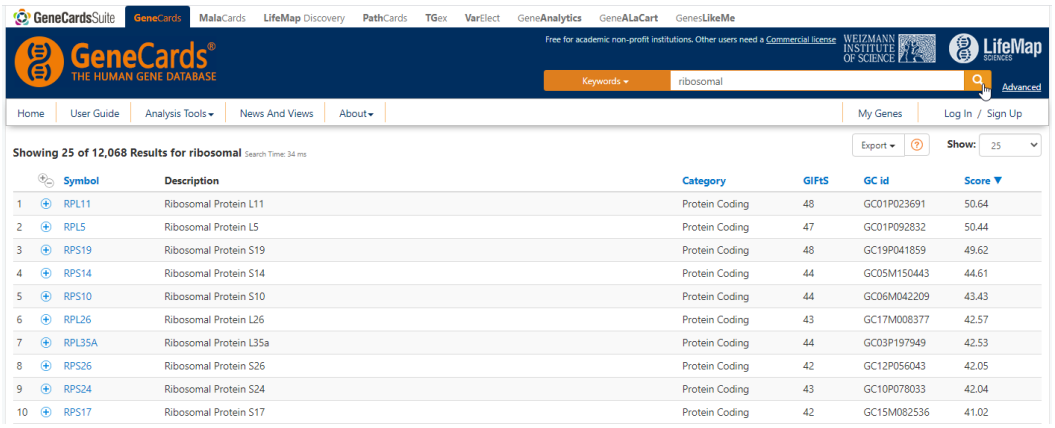
Εικόνα 20 – Μήνυμα λάθους και σύνδεσμος για επιστροφή στην Αρχική σελίδα

Επίσης προσφέρεται σύνδεσμος για είσοδο γονιδίου – παράδειγμα για την λειτουργία του κάθε εργαλείου. Σε περίπτωση που ο χρήστης δεν γνωρίζει το όνομα του γονιδίου που θα εισάγει για ανάλυση, προσφέρεται σύνδεσμος σε εξωτερική βάση δεδομένων, όπου μπορούν να αναζητηθούν γονίδια με βάση λέξεις κλειδιά που σχετίζονται με τη λειτουργία ή το βιολογικό τους ρόλο. Στην περίπτωση του ACT χρησιμοποιείται η βάση δεδομένων Thalemine (Εικόνα 21).



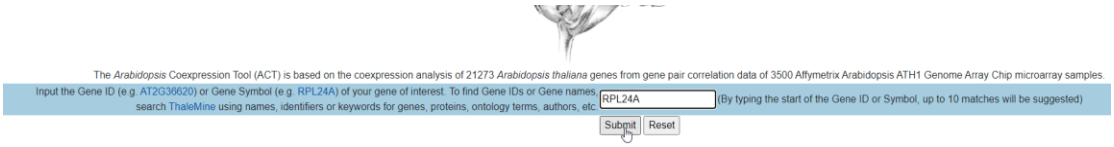
Εικόνα 21 – Αναζήτηση για τον όρο «ribosomal» στη σελίδα αναζήτησης γονιδίων του Thalemine. Φαίνεται ο κωδικός AGI και το κανονικό όνομα των γονιδίων που εμφανίζονται στα αποτελέσματα

Για τον άνθρωπο προτείνεται η βάση δεδομένων GeneCards, όπου επιτελείται παρόμοια αναζήτηση.

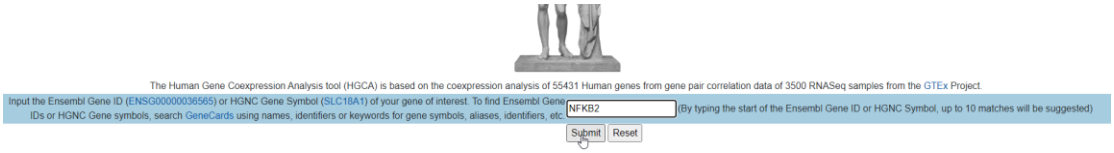


Εικόνα 22 - Αναζήτηση για τον όρο «ribosomal» στη σελίδα αναζήτησης γονιδίων του GeneCards

Ως παράδειγμα για αναζήτηση στο εργαλείο ACT θα χρησιμοποιήσουμε το ριβοσωμικό γονίδιο RPL24A (Εικόνα 23) του παραδείγματος, ενώ για τα εργαλεία HGCA θα χρησιμοποιήσουμε το γονίδιο NFKB2 (Εικόνα 24).



Εικόνα 23 – Επιλογή του γονιδίου RPL24A στο ACT



Εικόνα 24 – Επιλογή του γονιδίου NFKB2 στο HGCA2

Επιλογή γονιδίου και εκτέλεση ανάλυσης

Μετά το πάτημα του Submit εμφανίζεται το υποδέντρο γονιδιακής συνέκφρασης που περιέχει το γονίδιο που αναζητήσαμε (Εικόνα 25). Το δέντρο αυτό είναι ένα κομμάτι του αρχικού δέντρου συνέκφρασης που είχαμε υπολογίσει. Τα εργαλεία συνέκφρασης βρίσκουν τη θέση του γονιδίου αναζήτησης (που από εδώ και πέρα ονομάζουμε «γονίδιο-οδηγό») και εμφανίζουν στη οθόνη το κομμένο δέντρο που περιέχει 5 προγονικούς κόμβους από το φύλλο του γονιδίου-οδηγού. Ο χρήστης μπορεί να μειώσει ή να αυξήσει το μέγεθος του δέντρου, προσθέτοντας ή μειώνοντας του κόμβους με τους αντίστοιχους συνδέσμους. Επίσης είναι δυνατή η δια χειρός μεταβολή του αριθμού των κόμβων του δέντρου χρησιμοποιώντας το URL της ιστοσελίδας εκείνη τη στιγμή. Αλλάζοντας σε αυτή τη διεύθυνση:

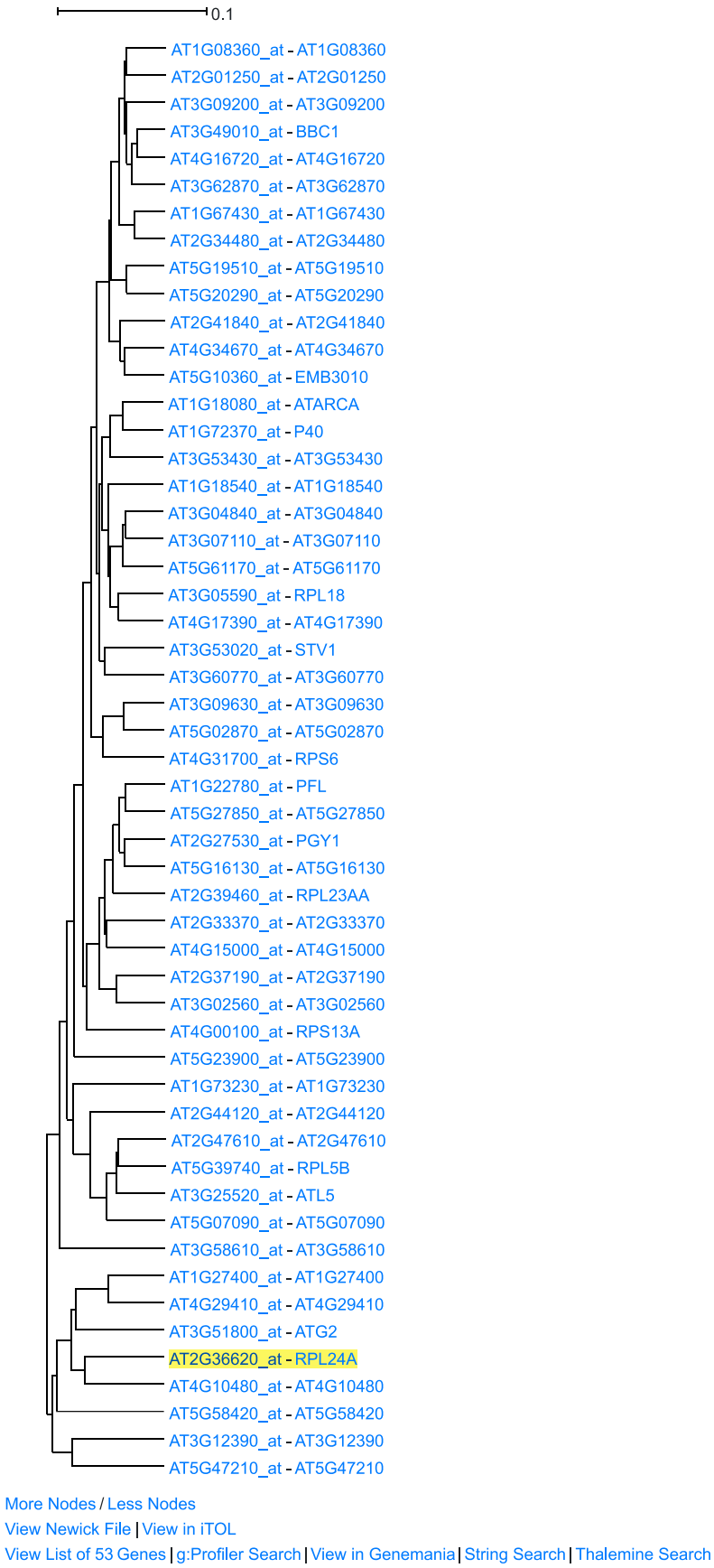
https://www.michalopoulos.net/act/trees/AraPort.php?probeset=AT2G36620_at&nodes=5

τον αριθμό στο σημείο nodes=5, για παράδειγμα, σε nodes=10, και πατώντας Enter, εμφανίζεται το αντίστοιχο δέντρο αλλά για 10 προγονικούς κόμβους.

Από το αρχικό ολοκληρωμένο δέντρο συνέκφρασης, αποκόπτεται με τη χρήση του προγράμματος PhyloPrune το σημείο που περιέχει το γονίδιο-οδηγό μέχρι και 5 κόμβους περαιτέρω, πράγμα που γίνεται αυτόματα με το πάτημα του

Submit. Αν με τους 5 κόμβους το υποδέντρο που εμφανιστεί ξεπερνάει έναν ορισμένο αριθμό φύλλων, τότε αυτόματα επιλέγεται ο αμέσως μικρότερος αριθμός κόμβων που να έχει ικανοποιητικό αριθμό φύλλων, κ.ο.κ. Στη συνέχεια, η σελίδα διαβάζει το αρχείο του υποδέντρου που έχει μορφοποίηση Newick (Archie et al., 2008) και το απεικονίζει μόνο με τη χρήση HTML. Τα φύλλα του δέντρου είναι τα γονίδια, που συμβολίζονται στο HGCA τόσο με τον μοναδικό κωδικό ENSG όσο και με το σύμβολό τους κατά HGNC, χωρισμένα με μία παύλα και στο ACT με τον κωδικό AGI και το σύμβολο του γονιδίου αν αυτό είναι διαθέσιμο. Το γονίδιο-οδηγός επισημαίνεται με κίτρινο φόντο. Κάνοντας κλικ στον κωδικό ENSG για το HGCA ή τον κωδικό AGI για το ACT, εκτελείται νέα ανάλυση χρησιμοποιώντας αυτό το γονίδιο ως οδηγό, χωρίς όμως να μεταβάλλεται το υποδέντρο συνέκφρασης που απεικονίζεται. Κάνοντας κλικ στο σύμβολο του γονιδίου, ο χρήστης μεταφέρεται στην εγγραφή του γονιδίου στο GeneCards ή Thalemine, ανάλογα αν βρίσκεται στα HGCA ή στο ACT.

Ακριβώς κάτω από υποδέντρο υπάρχουν οι επιλογές που προσφέρονται για περαιτέρω ανάλυση των γονιδίων του δέντρου. Αρχικά, ο χρήστης μπορεί να κατεβάσει το αρχείο μορφοποίησης Newick που περιέχει το εικονιζόμενο υποδέντρο, καθώς και να κατεβάσει τη λίστα των συμπεριλαμβανομένων γονιδίων. Επίσης, υπάρχουν επιλογές για αυτόματη εισαγωγή του δέντρου ή της λίστας γονιδίων σε εξωτερικά εργαλεία για δευτερεύουσα ανάλυση. Το δέντρο μπορεί να απεικονιστεί στο διαδικτυακό εργαλείο προβολής φυλογενετικών δέντρων Interactive Tree of Life (iToL) (Letunic and Bork, 2019). Η λίστα γονιδίων μπορεί να εξαχθεί στα εργαλεία: g:Profiler (Reimand et al., 2016) για ανάλυση εμπλουτισμού όρων, Genemania (Franz et al., 2018) και String (Szklarczyk et al., 2019) για δημιουργία δικτύων αλληλεπίδρασης μεταξύ γονιδίων και πρωτεϊνών και όσον αφορά το ACT, εξαγωγή στην ιστοσελίδα ανάλυσης λίστας γονιδίων του Thalemine (Εικόνα 26), όπου προσφέρει διάφορες αναλύσεις εμπλουτισμού όρων, όπως ανάλυση εμπλουτισμού βιβλιογραφίας ή πρωτεϊνικών δομικών-λειτουργικών περιοχών. Όσον αφορά τα HGCA (Εικόνα 27), προσφέρεται επιπλέον αναζήτηση στο Enrichr (Kuleshov et al., 2016) όπου περιλαμβάνει μεγάλη ποικιλία από όρους για ανάλυση εμπλουτισμού και τέλος το Pathway Commons (Rodchenkov et al., 2020) που αναζητεί βιολογικά και μεταβολικά μονοπάτια.



Εικόνα 25 – Το υποδέντρο συνέκφρασης για το γονίδιο RPL24A ως αποτέλεσμα στο εργαλείο ACT. Το δέντρο έχει τους default 5 κόμβους.

View Newick File | View in iTOL
View List of 54 Genes | g:Profiler Search | View in Genemania | String Search | Thalemine Search

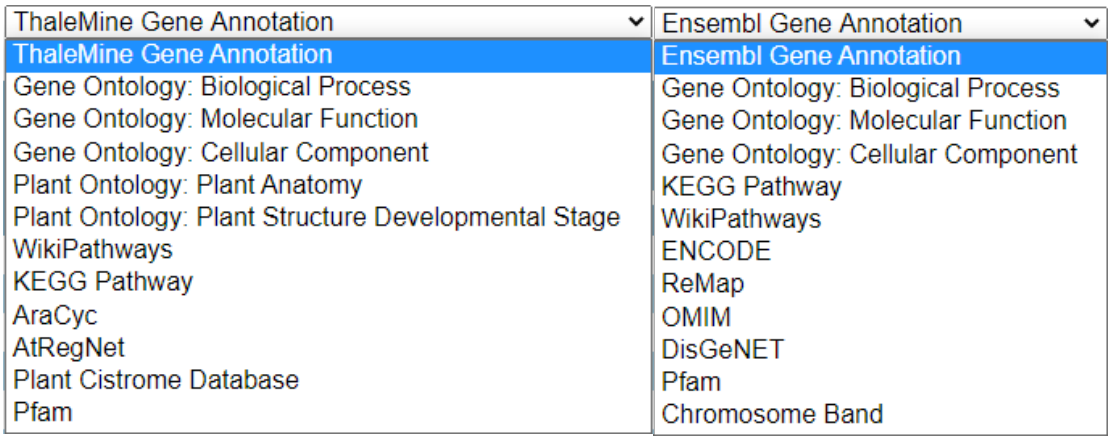
Εικόνα 26 - Εξωτερικοί σύνδεσμοι για επιπλέον αναλύσεις στο ACT

View Newick File | View in iTOL
View List of 23 Genes | g:Profiler Search | View in Genemania | Pathway Commons Search | String Search | Enrichr Search

Εικόνα 27 – Εξωτερικοί σύνδεσμοι για επιπλέον αναλύσεις στα HGCA

Ανάλυση υπερεκπροσώπησης όρων

Ο χρήστης μετά την εμφάνιση του υποδέντρου συνέκφρασης, μπορεί να επιλέξει να πραγματοποιήσει μία από τις αναλύσεις υπερεκπροσώπησης όρων των γονιδίων του υποδέντρου. Η προεπιλεγμένη επιλογή είναι ο σχολιασμός των γονιδίων. Παρουσιάζεται ένας πίνακας, όπου κάθε γραμμή είναι ένα από τα γονίδια του υποδέντρου με τη σειρά που εμφανίζονται σε αυτό, και στις άλλες στήλες είναι το σύμβολο του γονιδίου και η περιγραφή του. Σε αυτή την σελίδα δεν εκτελείται κάποια ανάλυση υπερεκπροσώπησης. Για να γίνει αυτό, πρέπει να επιλεχθεί μία κατηγορία εμπλουτισμού μέσω ενός αναπτυσσόμενου μενού (Εικόνα 28) δίπλα από την ένδειξη «Select an Enrichment Analysis:» που αναβοσβήνει.



Εικόνα 28 - Κατηγορίες ανάλυσης εμπλουτισμού για το ACT (αριστερά) και τα HGCA (δεξιά)

Επιλέγοντας μία κατηγορία, η σελίδα ανανεώνεται και εμφανίζονται σε έναν πίνακα «Enrichment Summary» τα αποτελέσματα της ανάλυσης υπερεκπροσώπησης. Ανάλογα με την κατηγορία, ο αριθμός των στηλών του πίνακα διαφέρει, όμως οι 4 πρώτες είναι πάντα οι ίδιες (Εικόνα 29). Η πρώτη στήλη, Rank, κατατάσσει τους όρους που εμφανίστηκαν στα αποτελέσματα της ανάλυσης εμπλουτισμού με βάση το μικρότερο P-value. Στην δεύτερη στήλη, FDR adj P-Value, αναγράφεται το τροποποιημένο P-value βάσει του False

Discovery Rate. Αναφέρεται ότι εμφανίζονται μόνο οι όροι με FDR adj P-value < 0.05 ($5.0 \cdot 10^{-2}$), καθώς αυτοί θεωρούνται στατιστικά σημαντικοί. Στην τρίτη στήλη, Hits, εμφανίζεται ο αριθμός των γονιδίων του υποδέντρου που περιγράφονται από αυτόν τον όρο προς τον αριθμό όλων των γονιδίων που έχουμε στη βάση δεδομένων μας που περιγράφονται από αυτόν το όρο, καθώς και το ποσοστό που αντιστοιχεί σε αυτό το πηλίκο, μέσα σε παρένθεση. Στην τέταρτη στήλη, Over-representation, αναγράφεται το πόσες φορές είναι υπερεκπροσωπημένος ο όρος στα γονίδια του υποδέντρου. Στις υπόλοιπες στήλες, αναφέρεται το όνομα και η περιγραφή του όρου, καθώς και ο μοναδικός κωδικός του στη βάση δεδομένων που προέρχεται. Στον κωδικό υπάρχει σύνδεσμος που μεταφέρει τον χρήστη στην εγγραφή του όρου στην αντίστοιχη βάση δεδομένων.

Gene Ontology: Biological Process - Enrichment Summary					
Rank	FDR adj p-value	Hits	Over-representation	Accession	Biological Process
1	$3.7 \cdot 10^{-62}$	46/425 (10.8%)	29.1	GO:0006414	translational elongation
2	$1.7 \cdot 10^{-56}$	47/632 (7.4%)	20.0	GO:0006412	translation
3	$1.7 \cdot 10^{-56}$	47/636 (7.4%)	19.8	GO:0043043	peptide biosynthetic process
4	$5.3 \cdot 10^{-56}$	47/655 (7.2%)	19.3	GO:0043604	amide biosynthetic process
5	$1.5 \cdot 10^{-54}$	47/705 (6.7%)	17.9	GO:0006518	peptide metabolic process
6	$1.8 \cdot 10^{-53}$	47/745 (6.3%)	16.9	GO:0043603	cellular amide metabolic process
7	$9.5 \cdot 10^{-47}$	48/1140 (4.2%)	11.3	GO:1901566	organonitrogen compound biosynthetic process
8	$1.4 \cdot 10^{-39}$	48/1608 (3.0%)	8.0	GO:1901564	organonitrogen compound metabolic process
9	$1.3 \cdot 10^{-27}$	49/3104 (1.6%)	4.2	GO:0034645	cellular macromolecule biosynthetic process
10	$3.1 \cdot 10^{-27}$	49/3169 (1.5%)	4.2	GO:0009059	macromolecule biosynthetic process
11	$3.5 \cdot 10^{-27}$	49/3183 (1.5%)	4.1	GO:0044271	cellular nitrogen compound biosynthetic process
12	$1.5 \cdot 10^{-26}$	49/3286 (1.5%)	4.0	GO:0010467	gene expression
13	$2.8 \cdot 10^{-25}$	47/3002 (1.6%)	4.2	GO:0044267	cellular protein metabolic process
14	$3.1 \cdot 10^{-24}$	48/3411 (1.4%)	3.8	GO:0019538	protein metabolic process
15	$7.3 \cdot 10^{-24}$	22/263 (8.4%)	22.5	GO:0042254	ribosome biogenesis
16	$3.0 \cdot 10^{-22}$	50/4375 (1.1%)	3.1	GO:0044249	cellular biosynthetic process
17	$3.1 \cdot 10^{-22}$	22/313 (7.0%)	18.9	GO:0022613	ribonucleoprotein complex biogenesis
18	$6.5 \cdot 10^{-22}$	50/4455 (1.1%)	3.0	GO:1901576	organic substance biosynthetic process
19	$3.7 \cdot 10^{-21}$	49/4290 (1.1%)	3.1	GO:0034641	cellular nitrogen compound metabolic process
20	$6.7 \cdot 10^{-21}$	50/4681 (1.1%)	2.9	GO:0006807	nitrogen compound metabolic process

Εικόνα 29 – Πίνακας αποτελεσμάτων εμπλουτισμού όρων Οντολογίας Γονιδίων για Βιολογικές Διεργασίες, για το γονίδιο RPL4A στο υποδέντρο των 5 κόμβων από το εργαλείο ACT.

Κάτω από τον πίνακα αποτελεσμάτων εμπλουτισμού όρων, υπάρχει και ένα δεύτερος πίνακας με τη λίστα των γονιδίων με τη σειρά που εμφανίζονται στο υποδέντρο, αντίστοιχος με τον πίνακα σχολιασμού γονιδίων που εμφανίζεται αρχικά μετά το πάτημα του Submit. Ανάλογα με την κατηγορία εμπλουτισμού που έχουμε επιλέξει, εμφανίζονται όλοι οι διαθέσιμοι όροι που περιγράφουν το κάθε γονίδιο σε αυτήν την κατηγορία (Εικόνα 30).

Probe Set	Accession	Gene Ontology: Biological Process Biological Process
AT1G08360_at	GO:0006412	translation
	GO:0006414	translational elongation
	GO:0006518	peptide metabolic process
	GO:0006807	nitrogen compound metabolic process
	GO:0008152	metabolic process
	GO:0009058	biosynthetic process
	GO:0009059	macromolecule biosynthetic process
	GO:0009987	cellular process
	GO:0010467	gene expression
	GO:0019538	protein metabolic process
	GO:0034641	cellular nitrogen compound metabolic process
	GO:0034645	cellular macromolecule biosynthetic process
	GO:0043043	peptide biosynthetic process
	GO:0043170	macromolecule metabolic process
	GO:0043603	cellular amide metabolic process
	GO:0043604	amide biosynthetic process
	GO:0044237	cellular metabolic process
	GO:0044238	primary metabolic process
	GO:0044249	cellular biosynthetic process
	GO:0044260	cellular macromolecule metabolic process
	GO:0044267	cellular protein metabolic process
	GO:0044271	cellular nitrogen compound biosynthetic process
	GO:0071704	organic substance metabolic process
	GO:1901564	organonitrogen compound metabolic process
	GO:1901566	organonitrogen compound biosynthetic process
	GO:1901576	organic substance biosynthetic process

Εικόνα 30 – Οι όροι βιολογικής διεργασίας οντολογίας γονιδίων που περιγράφουν το γονίδιο AT1G08360 από το εργαλείο ACT. Τα υπόλοιπα γονίδια του υποδέντρου παρουσιάζονται στον ίδιο πίνακα.

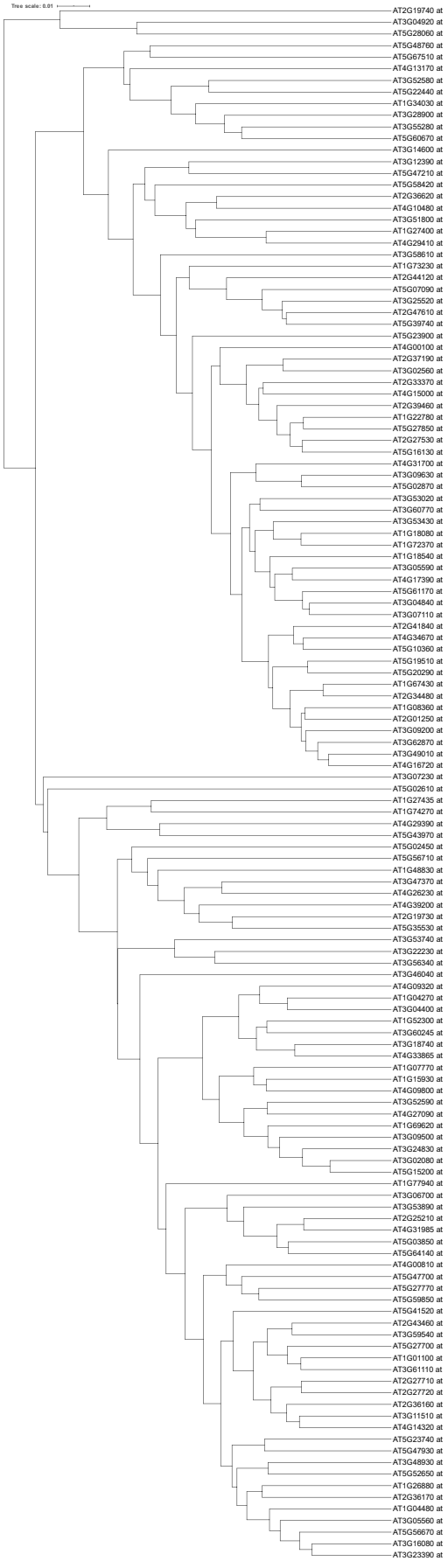
Παραδείγματα αποτελεσμάτων

ACT

Ριβοσωμικές Πρωτεΐνες

Η υπομονάδα του ριβοσώματος στο *Arabidopsis thaliana* αποτελείται από 80 ριβοσωμικές πρωτεΐνες (r-proteins), κάθε μία από 80 διαφορετικές οικογένειες που κωδικοποιούνται από 249 γονίδια. Κανένα από αυτά τα γονίδια δεν είναι μοναδικό αντίγραφο, γεγονός που σημαίνει ότι οι περισσότερες από τις ριβοσωμικές πρωτεΐνες κωδικοποιούνται από 3 ή 4 γονίδια. (Barakat et al., 2001). Το AT4G13170, γονίδιο που κωδικοποιεί μία ριβοσωμική πρωτεΐνη της οικογένειας L13, επιλέχθηκε ως γονίδιο-οδηγός σε μία ανάλυση του ACT. Το υποδέντρο συνέκφρασης που προέκυψε αποτελείται από 134 γονίδια-φύλλα (Εικόνα 31). Τα περισσότερα συνεκφραζόμενα γονίδια είναι δομικά συστατικά του ριβοσώματος. Για να επαληθευθεί η στατιστική σημαντικότητα του ευρήματος, έγινε ανάλυση εμπλουτισμού βιολογικών όρων μέσω του ACT (Πίνακας 2). Οι υπερεκπροσωπημένοι όροι από όλες τις πτυχές της Οντολογίας Γονιδίων σχετίζονται με ριβοσώματα και μετάφραση, παρουσιάζοντας πολύ μικρά τροποποιημένα P-values κατά FDR, με τιμές από ~10^{-176} μέχρι ~10^{-156}. Η ανάλυση βιολογικών μονοπατιών του KEGG, παρομοίως πρότεινε ρόλο συσχετιζόμενο με ριβοσώματα και η ανάλυση Pfam έδειξε εμπλουτισμό σε οικογένειες ριβοσωμικών πρωτεϊνών. Όροι συσχετιζόμενοι με «Κοτυληδόνα»

(cotyledon) και «Δομή Εμβρύου» (embryo structure) εμφανίστηκαν ως υπερεκπροσωπημένοι για την πτυχή της Ανατομίας Φυτών στις Οντολογίες Φυτών, ενώ ο όρος «Στάδιο Κοτυληδόνων Εμβρύου Φυτού» (plant embryo cotyledonary stage) εμφανίστηκε στην πτυχή Αναπτυξιακό Στάδιο Δομής Φυτού. Ανάλυση εμπλουτισμού Μεταγραφικών Παραγόντων, χρησιμοποιώντας τόσο τη βάση AtRegNet όσο και τη Plant Cistrome Database, αποκάλυψε δύο μεταγραφικούς παράγοντες, AT1G72740 (TRB5) and TRB2 που ανήκουν στην οικογένεια πρωτεϊνών homeodomain-like/winged-helix που προσδένονται στο DNA. Αυτό συμβαδίζει με την ανακάλυψη ότι η οικογένεια μεταγραφικών παραγόντων TRB ρυθμίζει την έκφραση των γονιδίων που σχετίζονται με την συγκρότηση του μηχανισμού μετάφρασης στα φυτά (Schrumpfova et al., 2016).



Εικόνα 31 – Το υποδέντρο συνέκφρασης για το γονίδιο AT4G13170, όπως απεικονίζεται από τον ιστότοπο iToL

Enrichment Summary for AT4G13170			
Category	p-value	Term ID	Description
GO: Biological Process	$1.2 \cdot 10^{-176}$	GO:0006414	translational elongation
	$7.0 \cdot 10^{-156}$	GO:0006412	translation
GO: Molecular Function	$5.9 \cdot 10^{-194}$	GO:0003735	structural constituent of ribosome
GO: Cellular Component	$2.8 \cdot 10^{-221}$	GO:0022626	cytosolic ribosome
PO: Plant Anatomy	$1.6 \cdot 10^{-25}$	PO:0020030	cotyledon
	$1.6 \cdot 10^{-25}$	PO:0025099	embryo plant structure
PO: Plant Structure Development Stage	$8.4 \cdot 10^{-17}$	PO:0001078	plant embryo cotyledonary stage
KEGG	$8.6 \cdot 10^{-155}$	ath03010	Ribosome - Arabidopsis thaliana (thale cress)
AtRegNet	$1.0 \cdot 10^{-22}$	AT1G72740	Homeodomain-like/winged-helix DNA-binding family protein
	$7.5 \cdot 10^{-13}$	TRB2	Homeodomain-like/winged-helix DNA-binding family protein
Pfam	$1.0 \cdot 10^{-6}$	PF01248	Ribosomal_L7Ae

Πίνακας 2 – Οι σημαντικότεροι εμπλουτισμοί όρων για το γονίδιο AT4G13170, όπως προκύπτουν από τον ιστότοπο του ACT. Οι περισσότεροι όροι συνδέονται με ριβοσώματα

Πρωτεΐνες Θερμικού Σοκ (Heat Shock Proteins)

Οι Πρωτεΐνες Θερμικού Σοκ (Heat shock proteins - HSP) είναι μία οικογένεια πρωτεϊνών που εκφράζονται ως αντίδραση σε στρεσογόνες καταστάσεις. Το γονίδιο Heat shock protein 101 (HSP101), που ανήκει στην οικογένεια HSP100 που είναι υπεύθυνη για την επιβίωση του *Arabidopsis thaliana* σε υψηλές θερμοκρασίες (Tonsor et al., 2008), χρησιμοποιήθηκε ως είσοδος στο ACT. Το υποδέντρο αναπτύχθηκε έως τους 11 προγονικούς κόμβους, αποτελούμενο από 44 γονίδια-φύλλα. Η ανάλυση Βιολογικής Διεργασίας Οντολογίας Γονιδίων έδειξε τον όρο «αντίσταση στη θερμότητα» (resistance to heat) ως κορυφαίο όρο και η ανάλυση Pfam κατέταξε πάνω από τα μισά γονίδια του υποδέντρου σε γονίδια που κωδικοποιούν πρωτεΐνες της οικογένειας HSP20. Επιπλέον, η ανάλυση του AtRegNet ανακάλυψε πέντε υπερεκπροσωπημένους μεταγραφικούς παράγοντες να στοχεύουν γονίδια του υποδέντρου. Οι μεταγραφικοί παράγοντες από τον δεύτερο μέχρι και τον πέμπτο είναι όλοι παράγοντες θερμικού σοκ (HSF3, HSFB2A, HSFC1 and AT-HSFB2B), ενώ για τον πρώτο μεταγραφικό παράγοντα, AT3G09735, δεν έχουμε αρκετά στοιχεία από την βιβλιογραφία για να τον χαρακτηρίσουμε. Στη συνέχεια, το γονίδιο Heat shock protein 90 (HSP90, AT5G56030) επιλέχθηκε ως γονίδιο-οδηγός. Αυξάνοντας το αρχικό υποδέντρο στα 66 γονίδια-φύλλα, η ανάλυση Βιολογικής Διεργασίας Οντολογίας Γονιδίων έδειξε τους όρους

«αντίσταση στη θερμότητα» (resistance to heat), «αντίσταση στο ερέθισμα θερμοκρασίας» (resistance to temperature stimulus), «αντίδραση σε υψηλή ένταση φωτός» (response to high light intensity), «αντίδραση σε αβιοτικό ερέθισμα» (response to abiotic stimulus) και «πρωτεϊνικό δίπλωμα» (protein folding) (Πίνακας 4), όροι που ταιριάζουν με τις γενικές πρωτεϊνικές λειτουργίες του HSP90 (Milioni and Hatzopoulos, 1997). Ο εμπλουτισμός βιολογικών μονοπατιών του KEGG, έδειξε τον όρο «Επεξεργασία πρωτεϊνών στο ενδοπλασματικό δίκτυο» (Protein processing in endoplasmic reticulum), πράγμα που επαληθεύεται από το ρόλο του HSP90 ως μοριακού συνοδού που βοηθάει στο δίπλωμα άλλων πρωτεϊνών και την σταθεροποίησή τους.

Enrichment Summary for HSP101			
Category	p-value	Term ID	Description
GO: Biological Process	$9.1 \cdot 10^{-45}$	GO:0009409	response to heat
Pfam	$1.5 \cdot 10^{-29}$	PF00011	Hsp20/alpha crystallin family
	$1.5 \cdot 10^{-5}$	PF00012	Hsp70 protein
AtRegNet	$2.3 \cdot 10^{-24}$	AT3G09735	S1FA-like DNA-binding protein
	$5.6 \cdot 10^{-24}$	HSF3	heat shock factor 3
	$2.3 \cdot 10^{-19}$	HSFB2A	heat shock transcription factor B2A
	$2.9 \cdot 10^{-17}$	HSFC1	heat shock transcription factor C1
	$4.9 \cdot 10^{-5}$	AT-HSFB2B	winged-helix DNA-binding transcription factor family protein

Πίνακας 3 - Αποτελέσματα της ανάλυσης υπερεκπροσώπησης για το HSP101.

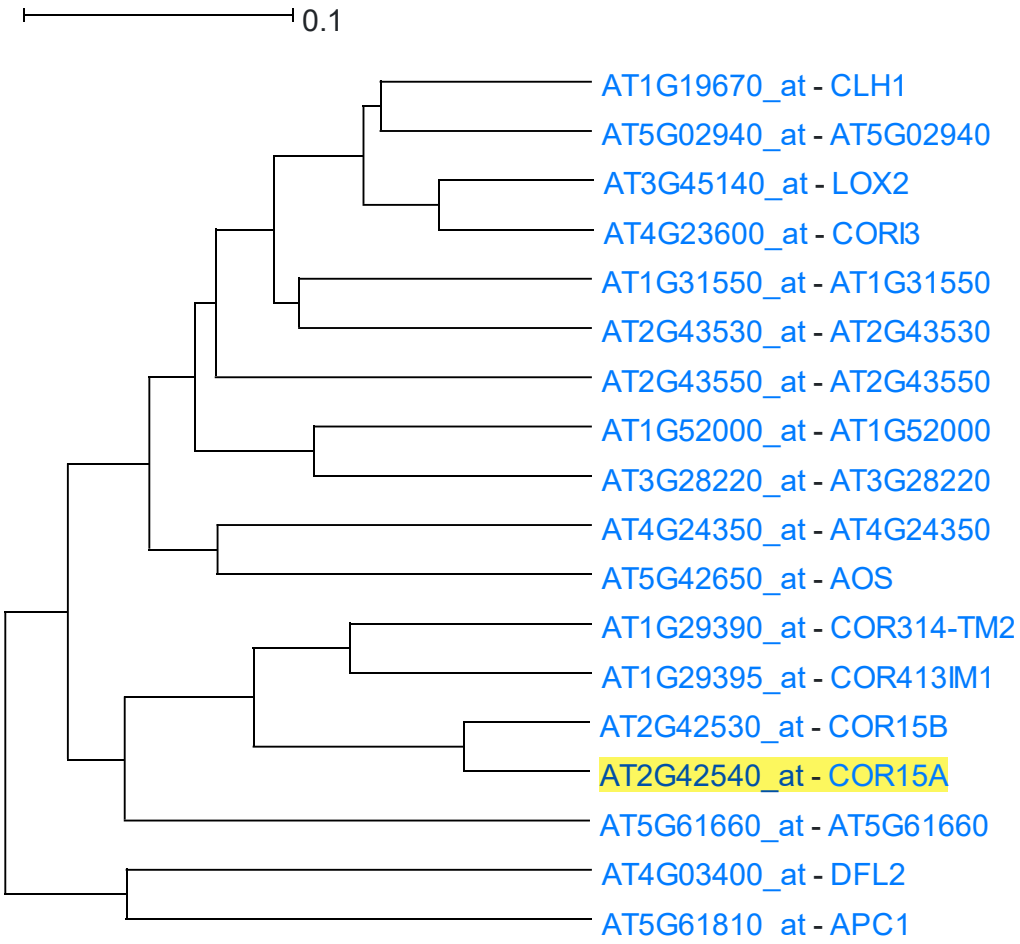
Enrichment Summary for HSP90			
Category	p-value	Term ID	Description
GO: Biological Process	$9.1 \cdot 10^{-45}$	GO:0009409	response to heat
	$7.3 \cdot 10^{-34}$	GO:0009266	response to temperature stimulus
	$6.1 \cdot 10^{-19}$	GO:0009644	response to high light intensity
	$1.7 \cdot 10^{-18}$	GO:0042542	response to hydrogen peroxide
	$3.7 \cdot 10^{-18}$	GO:0009628	response to abiotic stimulus
	$8.5 \cdot 10^{-17}$	GO:0006457	protein folding
KEGG	$9.6 \cdot 10^{-25}$	ath04141	Protein processing in endoplasmic reticulum

Πίνακας 4 – Αποτελέσματα της ανάλυσης υπερεκπροσώπησης για το HSP90.

Αντίδραση στο κρύο (Response to cold)

Το γονίδιο Cold-regulated 15a (*COR15A*) ενισχύει την αντίσταση στην ψύξη στο *Arabidopsis thaliana* (Artus et al., 1996; Wang and Hua, 2009). Το *COR15A* χρησιμοποιήθηκε ως γονίδιο-οδηγός στο ACT (Πίνακας 5). Το *COR15A* εμφανίστηκε σε διπλανό φύλλο από το ομόλογό του, *COR15B,* πάνω

στο υποδέντρο των 18 γονιδίων που δημιουργήθηκε. Τα γονίδια *COR314-TM2* και *COR413IM1*, τα οποία επίσης έχουν λειτουργία ρύθμισης για το κρύο, εμφανίστηκαν στον ίδιο κλάδο. Ανάλυση υπερεκπροσώπησης Οντολογίας Γονιδίων Βιολογικής Διεργασίας έδειξε τον όρο «εγκλιματισμός στο κρύο» (acclimation to cold) ως υπερεκπροσωπημένο.



Εικόνα 32 – Το υποδέντρο συνέκφρασης για το γονίδιο COR15A.

Enrichment Summary for COR15A			
Category	p-value	Term ID	Description
GO: Biological Process	$6.8 \cdot 10^{-6}$	GO:0009631	Cold acclimation

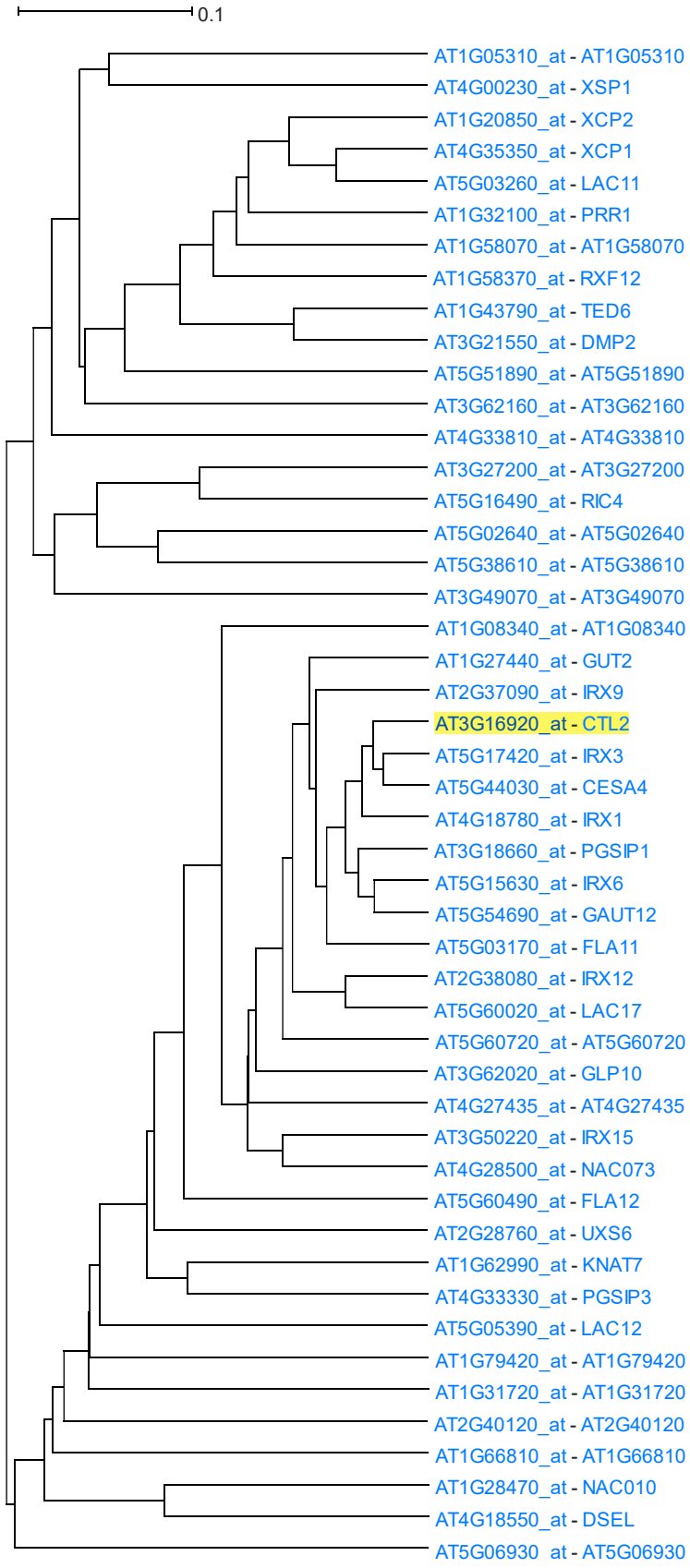
Πίνακας 5 - Αποτελέσματα της ανάλυσης υπερεκπροσώπησης για το COR15A.

Βιογένεση Κυτταρικού Τοιχώματος (Cell Wall Biogenesis)

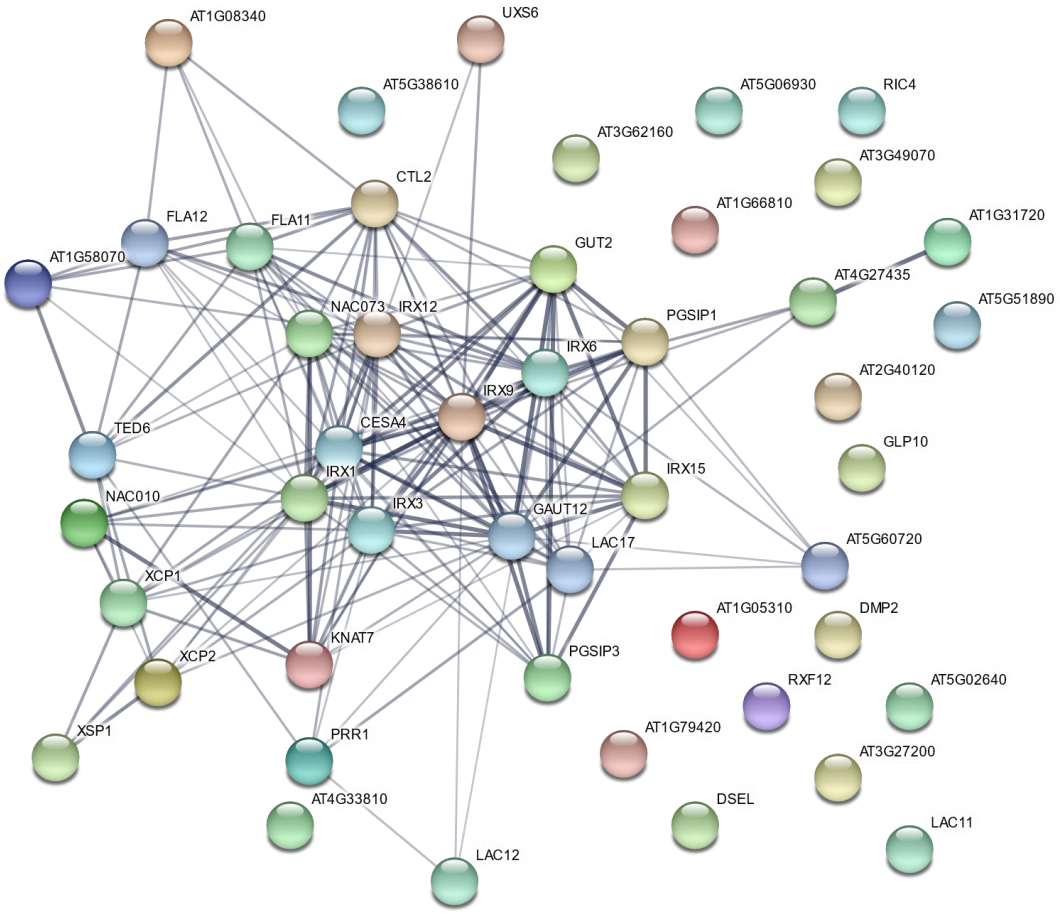
Το γονίδιο Chitinase-like protein 2 (*CTL2*), το οποίο έχει πιθανό ρόλο στην σύνθεση κυτταρικού τοιχώματος στο *Arabidopsis thaliana* (Hossain et al., 2010) χρησιμοποιήθηκε ως είσοδος στο ACT, και το παραγόμενο υποδέντρο αναπτύχθηκε έως τους 23 προγονικούς κόμβους (Εικόνα 33). Ανάλυση υπερεκπροσώπησης Βιολογικής Διεργασίας των 48 συνεκφρασμένων

γονιδίων κατέταξε τη «βιογένεση δευτεροταγούς τοιχώματος φυτικού τύπου» (plant-type secondary wall biogenesis) ως κορυφαίο όρο (p-value: $2.0 \cdot 10^{-29}$) με κοντινό δεύτερο (p-value: $2.0 \cdot 10^{-29}$) τον όρο «βιογένεση κυτταρικού τοιχώματος φυτικού τύπου» (plant-type cell wall biogenesis) και η ανάλυση AraCyc επίσης πρότεινε τη βιοσύνθεση κυτταρίνης ως υπερεκπροσωπημένο όρο (p-value: $6.2 \cdot 10^{-5}$). Ο όρος «καταβολική διεργασία λιγνίνης» (lignin catabolic process), ο οποίος επίσης εμφανίστηκε ως υπερεκπροσωπημένος, ταιριάζει με τον εύρημα ότι μία μετάλλαξη στο *CTL2* αυξάνει την συσσώρευση λιγνίνης σε φυτάρια του *Arabidopsis thaliana* που αναπτύχθηκαν στο σκοτάδι (Hossain et al., 2010). Επίσης, το ορθόλογο του *CTL2* στο βαμβάκι, εκφράζεται περισσότερο σε κύτταρα με δευτεροταγές κυτταρικό τοίχωμα (Zhang et al., 2004). Δίκτυο συσχέτισης πρωτεϊνών που δημιουργήθηκε μέσω του STRING (Εικόνα 34), χρησιμοποιώντας τα συνεκφραζόμενα γονίδια, έδειξε μεγάλη συσχέτιση μεταξύ του γονιδίου-οδηγού *CLT2* και γειτονικών γονιδίων-φύλλων σε αυτό, ιδιαίτερα εκείνων που κωδικοποιούν πρωτεΐνες της οικογένειας IRX. Τέλος, η ανάλυση AtRegNet, έδειξε τον *VND7* ως πρωταρχικό μεταγραφικό παράγοντα μεταξύ και άλλων υπερεκπροσωπημένων. Ο *VND7* ρυθμίζει τον τρόπο απόθεσης δευτεροταγούς κυτταρικού τοιχώματος στα αγγεία των φυτών (Yamaguchi et al., 2011) και επίσης είναι γνωστό ότι προσδένεται στους υποκινητές πολλών γονιδίων που σχετίζονται με τη βιοσύνθεση δευτεροταγούς κυτταρικού τοιχώματος (Taylor-Teeples et al., 2015). Επιπλέον, ένα μέλος της οικογένειας γονιδίων συνθάσης κυτταρίνης, το *CEV1* (*AT5G05170*) γνωστό και ως *CESA3*, χρησιμοποιήθηκε για μία άλλη ανάλυση στο ACT. Το *CEV1* είναι μία καταλυτική υπομονάδα συμπλόκων συνθάσης κυτταρίνης που εμπλέκεται στην δημιουργία του πρωτοταγούς κυτταρικού τοιχώματος (Burn et al., 2002; Daras et al., 2009). Το παραγόμενο υποδέντρο αναπτύχθηκε έως 7 προγονικούς κόμβους και έδειξε συνέκφραση με άλλα γονίδια συνθάσης κυτταρίνης και με πρωτεΐνες που εμπλέκονται στην κυτταρική επέκταση (cell expansion) όπως *COB, POM1* και *CSI1* το οποίο είναι πρωτεΐνη που αλληλεπιδρά με συνθάση κυτταρίνης (Εικόνα 35). Η ανάλυση Βιολογικής Διεργασίας των συνεκφραζόμενων γονιδίων ανέδειξε τους όρους «διεργασία βιοσύνθεσης πολυσακχαριτών» (polysaccharide biosynthetic process), «διεργασία βιοσύνθεσης κυτταρίνης» (cellulose biosynthetic process), «βιογένεση πρωτοταγούς κυτταρικού

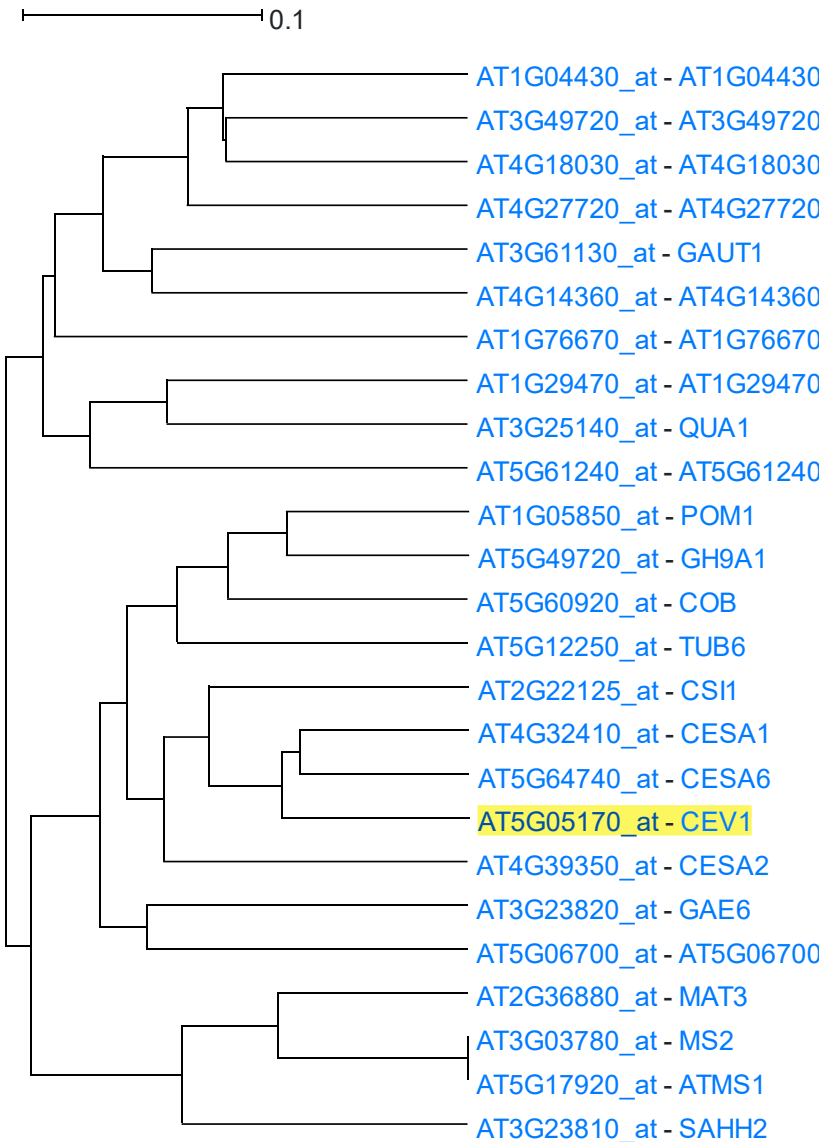
τοιχώματος φυτικού τύπου» (plant-type primary cell wall biogenesis) και «διεργασία βιοσύνθεσης β-γλυκανών» (beta-glucan biosynthetic process), ως κορυφαίους (Πίνακας 6). Επιπλέον, όσον αφορά τους κορυφαίους όρους στη Μοριακή Λειτουργία, έχουμε «ενεργότητα συνθάσης κυτταρίνης (δημιουργία UDP)» (cellulose synthase (UDP-forming) activity), «ενεργότητα συνθάσης κυτταρίνης» (cellulose synthase activity), «ενεργότητα S-μεθυλοτρανσφεράσης» (S-methyltransferase activity) και «ενεργότητα γλυκοτρασνφεράσης UDP» (UDP-glycosyltransferase activity), οι οποίοι επιβεβαιώνουν τους ρόλους των γονιδίων του υποδέντρου. Η ανάλυση Κυτταρικού Συστατικού ανέδειξε τους όρους «δίκτυο δια-Golgi» (trans-Golgi network), «υποδιαμέρισμα Golgi» (Golgi subcompartment) και «πλασματική μεμβράνη» (plasma membrane) ως υπερεκπροσωπημένους, γεγονός που υποστηρίζει τη λειτουργία αυτών των γονιδίων στις συγκεκριμένες υποκυτταρικές περιοχές (Wightman and Turner, 2010).



Εικόνα 33 - Το υποδέντρο συνέκφρασης για το γονίδιο CTL2.



Εικόνα 34 – Το δίκτυο αλληλεπίδρασης πρωτεϊνών που παράχθηκε από το STRING, με είσοδο τα συνεκφραζόμενα γονίδια με το CTL2.



Εικόνα 35 - Το υποδέντρο συνέκφρασης για το γονίδιο CEV1.

Enrichment Summary for CEV1			
Category	p-value	Term ID	Description
GO: Biological Process	$6.4 \cdot 10^{-14}$	GO:0000271	polysaccharide biosynthetic process
	$1.4 \cdot 10^{-13}$	GO:0030244	cellulose biosynthetic process
	$1.9 \cdot 10^{-13}$	GO:0009833	plant-type primary cell wall biogenesis
	$1.9 \cdot 10^{-13}$	GO:0033692	cellular polysaccharide biosynthetic process
	$4.4 \cdot 10^{-13}$	GO:0051274	beta-glucan biosynthetic process
GO: Molecular Function	$5.8 \cdot 10^{-7}$	GO:0016760	cellulose synthase (UDP-forming) activity
	$1.7 \cdot 10^{-6}$	GO:0016759	cellulose synthase activity
	$2.6 \cdot 10^{-6}$	GO:0008172	S-methyltransferase activity
	$3.0 \cdot 10^{-6}$	GO:0008194	UDP-glycosyltransferase activity
GO: Cellular Component	$2.1 \cdot 10^{-19}$	GO:0005802	trans-Golgi network
	$1.1 \cdot 10^{-18}$	GO:0098791	Golgi subcompartment

Πίνακας 6 - Αποτελέσματα της ανάλυσης υπερεκπροσώπησης για το CEV1.

Φωτοσύνθεση

Η πρωτεΐνη PSB28 είναι συστατικό του φωτοσυστήματος ΙΙ (photosystem II - PSII) το οποίο βοηθάει στην επιδιόρθωση και στην *de novo* σύνθεση των πρωτεϊνών του συμπλόκου PSII, ως αντίδραση σε επαγόμενο από ακραία υψηλό φως στρες (extreme high light induced-stress)(Parrine et al., 2018). Το παραγόμενο υποδέντρο πρότεινε την φωτοσύνθεση ως κορυφαία βιολογική διεργασία. Έγινε επέκταση του αρχικού υποδέντρου στο σημείο που εμφανίστηκαν οι ελάχιστες p-values. 41 προγονικοί κόμβοι παρήγαγαν υποδέντρο αποτελούμενο από 729 γονίδια, πάνω στο οποίο έγινε ανάλυση εμπλουτισμού όρων (Πίνακας 7). Οι κορυφαίοι όροι για Βιολογική Διεργασία, KEGG και AraCyc έδειξαν τον όρο «φωτοσύνθεση» (photosynthesis) ως υπερεκπροσωπημένο, ενώ η ανάλυση Κυτταρικού Συστατικού, πρότεινε το «πλαστίδιο» (plastid) ως το βασικό οργανίδιο, συμπίπτοντας με την «πρόσδεση σε χρωστική» (pigment binding) ως κορυφαίο όρο της ανάλυσης Μοριακής Λειτουργίας και τον όρο «αρχέγονη κοτυληδόνα» (cotyledon primordium) της Ανατομίας Φυτών. Ακόμα, η ανάλυση Pfam έδειξε τις πρωτεΐνες πρόσδεσης σε χλωροφύλλη ως κορυφαίο όρο και η ανάλυση AtRegNet ανακάλυψε τον *PIF4* (phytochrome interacting factor 4) ως κορυφαίο μεταγραφικό παράγοντα.

Enrichment Summary for PSB28			
Category	p-value	Term ID	Description
GO: Biological Process	$5.9 \cdot 10^{-123}$	GO:0015979	photosynthesis
	$2.2 \cdot 10^{-74}$	GO:0019684	photosynthesis, light reaction
GO: Molecular Function	$3.3 \cdot 10^{-12}$	GO:0031409	pigment binding
GO: Cellular Component	0	GO:0044434	chloroplast part
	0	GO:0044435	plastid part
PO: Plant Anatomy	$2.6 \cdot 10^{-125}$	PO:0000015	cotyledon primordium
	$2.6 \cdot 10^{-125}$	PO:0025432	cotyledon anlagen
KEGG	$1.4 \cdot 10^{-40}$	ath00195	Photosynthesis - Arabidopsis thaliana (thale cress)
AraCyc	$8.1 \cdot 10^{-24}$	PWY-101	photosynthesis light reactions
AtRegNet	$5.8 \cdot 10^{-5}$	PIF4	phytochrome interacting factor 4
Pfam	$1.2 \cdot 10^{-13}$	PF00504	Chlorophyll A-B binding protein

Πίνακας 7 - Αποτελέσματα της ανάλυσης υπερεκπροσώπησης για το PSB28

Κιρκαδικός Ρυθμός

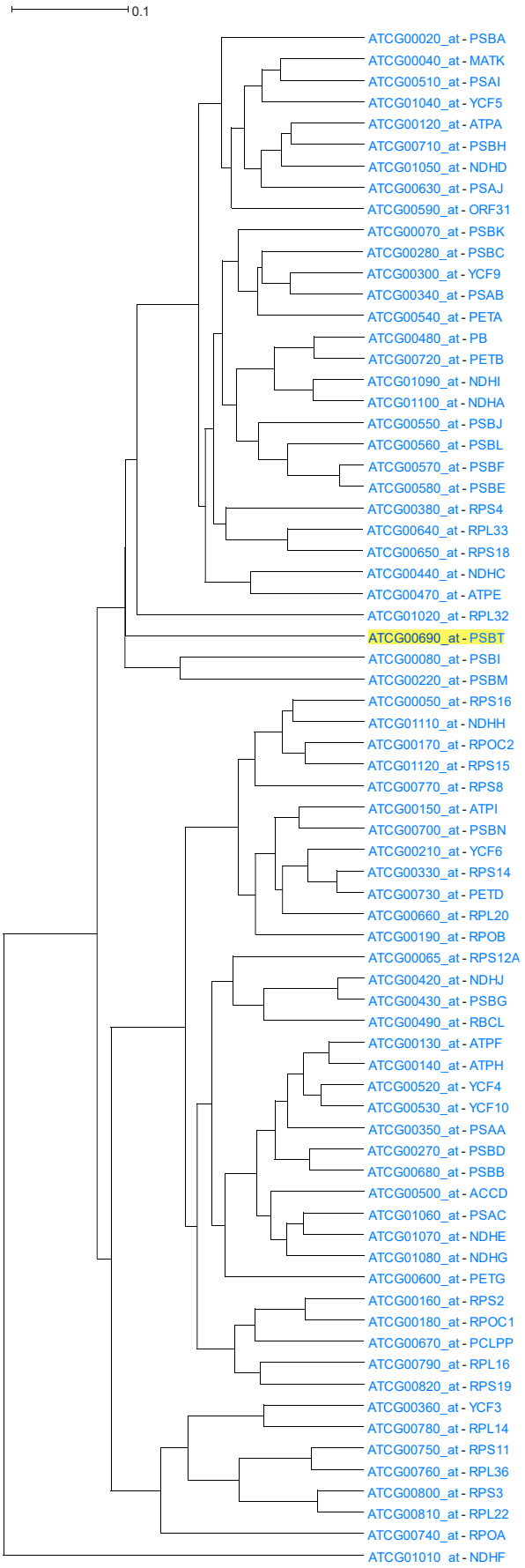
Το γονίδιο LATE ELONGATED HYPOCOTYL (LHY) παίζει ρόλο στο κιρκαδικό ρολόι του *Arabidopsis thaliana* (Lu et al., 2009). Χρησιμοποιώντας το σε ανάλυση του ACT και αφού έγινε ανάπτυξη του υποδέντρου έως 8 προγονικούς κόμβους (Πίνακας 8), ο κορυφαίος όρος Βιολογικής Διεργασίας σε αυτά τα 21 συνεκφραζόμενα γονίδια ήταν «ρυθμική διεργασία» (rhythmic process). Η ανάλυση μονοπατιών KEGG, επίσης πρότεινε τον «κιρκαδικό ρυθμό» (circadian rhythm) ως κορυφαίο όρο και η ανάλυση Pfam κατέταξε 8 από τα γονίδια ως μεταγραφικούς παράγοντες που ανήκουν στις οικογένειες πρωτεϊνών B-box zinc finger (zf-B_box) και Myb-like DNA-binding domain (Myb_DNA-binding). Η ανάλυση AtRegNet ανακάλυψε τον *TIMING OF CAB EXPRESSION 1* (*TOC1*), ο οποίος αποτελεί κύριο συστατικό στο κιρκαδικό ρολόι που ενσωματώνει πληροφορίες από το περιβάλλον για να οργανώσει κιρκαδικές αντιδράσεις (Perales and Mas, 2007), ως μεταγραφικό παράγοντα που στοχεύει τα συνεκφραζόμενα γονίδια. Αρκετά ενδιαφέρον είναι το γεγονός ότι δύο γονίδια μεταγραφικών παραγόντων, RVE8 και CCA1, που προσδένονται στον υποκινητή του TOC1 (Farinas and Mas, 2011), ήταν ανάμεσα στα συνεκφραζόμενα γονίδια.

Enrichment Summary for LHY			
Category	p-value	Term ID	Description
GO: Biological Process	$6.8 \cdot 10^{-12}$	GO:0048511	rhythmic process
KEGG	$5.6 \cdot 10^{-7}$	ath04712	Circadian rhythm - plant - Arabidopsis thaliana (thale cress)
Pfam	$1.8 \cdot 10^{-11}$	PF00643	B-box zinc finger
	$1.0 \cdot 10^{-4}$	PF00249	Myb-like DNA-binding domain

Πίνακας 8 - Αποτελέσματα της ανάλυσης υπερεκπροσώπησης για το LHY

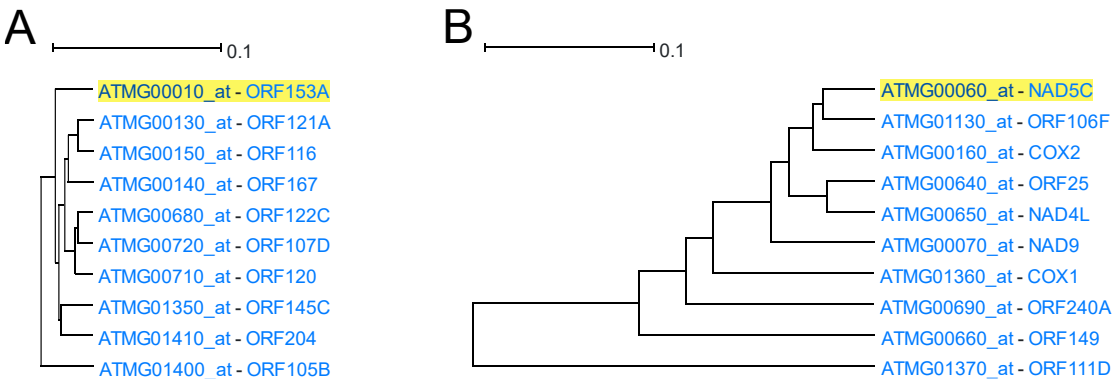
Χλωροπλαστικές και Μιτοχονδριακές Πρωτεΐνες

Το chip ATH1 περιέχει ανιχνευτές για 72 χλωροπλαστικά γονίδια. Χρησιμοποιώντας ένα από αυτά, το *photosystem II reaction centre protein T* (*PSBT*), ως γονίδιο-οδηγό στο ACT και μειώνοντας του προγονικούς κόμβους στους 4, παράγεται ένα υποδέντρο συνέκφρασης που περιέχει αποκλειστικά και μόνο τα 72 χλωροπλαστικά γονίδια (διακρίνονται από τα αρχικά ATCG στον κωδικό AGI του γονιδίου) (Εικόνα 36). Τα πρώτα 7 γονίδια του δέντρου σχετίζονται με μετάφραση ενώ τα υπόλοιπα κυρίως με φωτοσύνθεση.



Εικόνα 36 – Το υποδέντρο συνέκφρασης με τα 72 χλωροπλαστικά γονίδια

Όσον αφορά τα μιτοχόνδρια, το chip ATH1 περιέχει ανιχνευτές για 27 μιτοχονδριακά γονίδια (διακρίνονται με τα αρχικά ATMG στον κωδικό AGI του γονιδίου). Σε αντίθεση με τα χλωροπλαστικά γονίδια, τα μιτοχονδριακά δε βρίσκονται όλα συγκεντρωμένα σε ένα κοινό υποδέντρο, αλλά συναντώνται σε ομάδες. Υπάρχουν δύο υποδέντρα που περιέχουν ομάδες μιτοχονδριακών γονιδίων, ένα με 11 (Εικόνα 37 Α) και ένα με 10 φύλλα (Εικόνα 37 Β). Το πρώτο υποδέντρο περιέχει κυρίως υποθετικές πρωτεΐνες, ενώ το δεύτερο, πέρα από το γεγονός ότι έχει καλύτερα σχολιασμένα γονίδια, έδειξε στην ανάλυση Βιολογικής Διεργασίας, εμπλουτισμένους τους όρους «κυτταρική αναπνοή» (cellular respiration) (p-value: $1.3 \cdot 10^{-7}$) και «παραγωγή ενέργειας μέσω οξείδωσης οργανικών ενώσεων» (energy derivation by oxidation of organic compounds) (p-value: $1.5 \cdot 10^{-7}$). Τα υπόλοιπα μιτοχονδριακά γονίδια είναι συγκεντρωμένα σε ομάδες των δύο ή τριών.

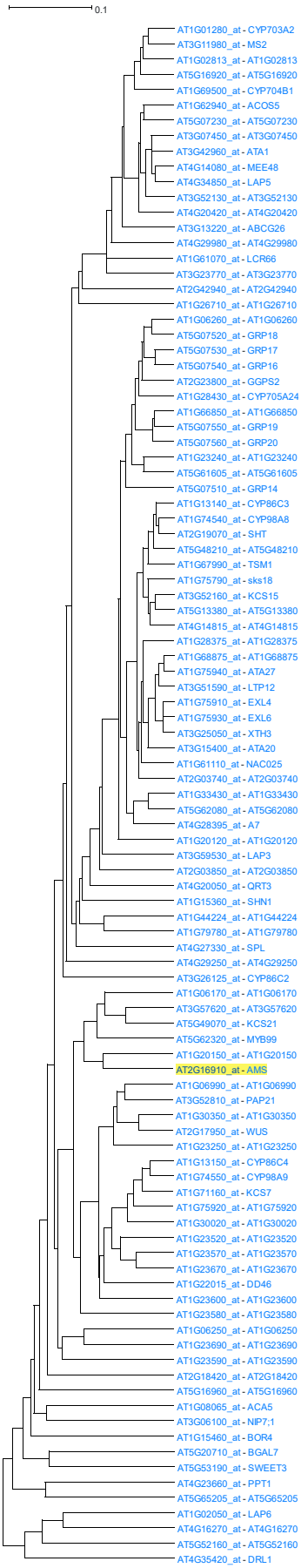


Εικόνα 37 – Τα υποδέντρα συνέκφρασης για τις μιτοχονδριακές πρωτεΐνες με τα 11 (Α) και 10 (Β) μιτοχονδριακά γονίδια-φύλλα

Ανθήρες και Γύρη

Το γονίδιο *ABORTED MICROSPORES (AMS)* παίζει ρόλο στην ανάπτυξη των κυττάρων τάπητα (έναν ειδικό τύπο θρεπτικών κυττάρων που βρίσκονται μέσα στον ανθήρα) (Xu et al., 2010). Χρησιμοποιήθηκε ως είσοδος σε μία ανάλυση του ACT και το παραγόμενο υποδέντρο αναπτύχθηκε έως 12 προγονικούς κόμβους, με 101 συνολικά γονίδια (Εικόνα 38). Ανάλυση υπερεκπροσώπησης Βιολογικής Διεργασίας, έδειξε τον όρο «συγκρότηση τοίχους γύρης» (pollen wall assembly) ως κορυφαίο, η ανάλυση Κυτταρικού Συστατικού έδειξε τον όρο «κάλυμμα γύρης» (pollen coat) και οι τρεις κορυφαίοι όροι για την ανάλυση Ανατομίας Φυτού ήταν «τοίχος σποραγγείων»

(sporangium wall), «τάπητας» (tapetum) και «ανθήρας» (anther). Όλοι οι αναφερόμενοι όροι σχετίζονται με ανθήρες.



Εικόνα 38 – Το υποδέντρο συνέκφρασης για το AMS

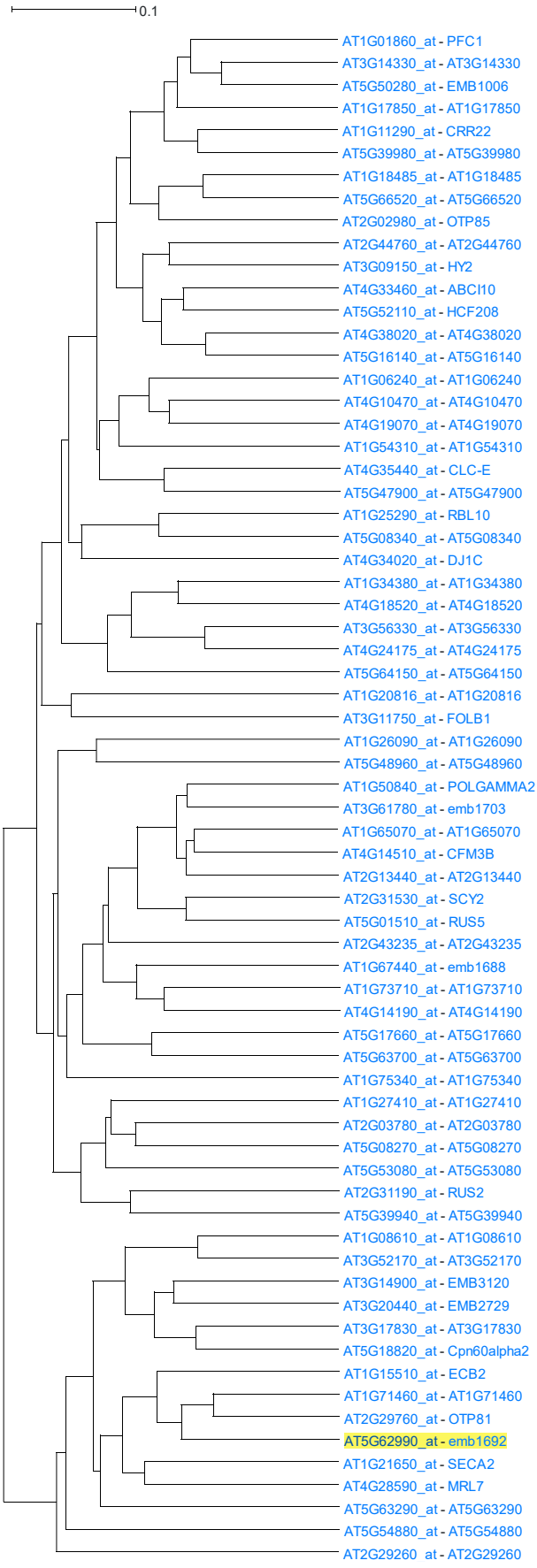
Enrichment Summary for AMS			
Category	p-value	Term ID	Description
GO: Biological Process	$7.9 \cdot 10^{-22}$	GO:0010208	pollen wall assembly
GO: Cellular Component	$1.5 \cdot 10^{-6}$	GO:0070505	pollen coat
PO: Plant Anatomy	$9.9 \cdot 10^{-17}$	PO:0025306	sporangium wall
	$1.2 \cdot 10^{-15}$	PO:0025313	tapetum
	$6.3 \cdot 10^{-15}$	PO:0009066	anther

Πίνακας 9 - Αποτελέσματα της ανάλυσης υπερεκπροσώπησης για το AMS

Ανάπτυξη Εμβρύου

Το γονίδιο *embryo defective 1692* (*emb1692*) παράγει μία πυρηνική πρωτεΐνη που προσδένεται σε RNA, η οποία συμμετέχει στο μάτισμα χλωροπλαστικών εσωνίων της ομάδας II και πυρηνικού πρώιμου mRNA. Η πρωτεΐνη ελέγχει την εμβρυϊκή και μετα-εμβρυϊκή ανάπτυξη και εκφράζεται τόσο στους πυρήνες όσο και στους χλωροπλάστες των μεριστωμάτων κυρίως (Daras et al., 2019). Η ανάλυση του ACT για το *emb1692*, παρήγαγε ένα υποδέντρο το οποίο αφού αναπτύχθηκε έως 8 προγονικούς κόμβους, περιείχε 68 γονίδια (Εικόνα 39). Η ανάλυση Pfam έδειξε τις οικογένειες PPR και DYW ως κορυφαίες (Πίνακας 10). Η οικογένεια DYW είναι μία υποομάδα της PPR και είναι απαραίτητη κυρίως για τροποποίηση του RNA σε οργανίδια (Okuda et al., 2009). Δηλαδή, αρκετά γονίδια ανήκουν στην οικογένεια πρωτεϊνών της επανάληψης τριακονταπενταπεπτιδίου (Pentatricopeptide repeat - PPR), τα οποία εμπλέκονται στον μεταβολισμό του RNA σε οργανίδια που έχουν βασικό ρόλο στη βιογένεσή τους και την ανάπτυξη του εμβρύου (Lurin et al., 2004). Επειδή το *emb1692* συμμετέχει στο μάτισμα χλωροπλαστικών εσωνίων της ομάδας II, η ύπαρξη ενός δικτύου με γονίδια PPR δικαιολογεί τον ρόλο των συνεκφρασμένων γονιδίων στον μεταβολισμό του RNA. Ανάλυση Οντολογίας Γονιδίων Βιολογικής Διεργασίας, παρουσίασε τους όρους «τροποποίηση RNA» (RNA modification), «ανάπτυξη εμβρύου» (embryo development), «ανάπτυξη καρπού» (seed development), «επεξεργασία RNA» (RNA processing) και «τροποποίηση χλωροπλαστικού RNA» (chloroplast RNA modification) που είναι στενά συνδεδεμένοι με τις λειτουργίες του γονιδίου. Επιπλέον, τα αποτελέσματα εμπλουτισμού για Αναπτυξιακό Στάδιο Δομής Φυτού, έδειξαν πολλούς διαφορετικούς όρους σχετιζόμενους με διάφορα στάδια εμβρυϊκής

ανάπτυξης, γεγονός που υποστηρίζει το ρόλο του *emb1692* και των συνεκφραζόμενων με αυτό γονιδίων στον έλεγχο της εμβρυϊκής ανάπτυξης.



Εικόνα 39 - Το υποδέντρο συνέκφρασης για το emb1692

Enrichment Summary for emb1692			
Category	p-value	Term ID	Description
GO: Biological Process	$1.5 \cdot 10^{-7}$	GO:0009451	RNA modification
	$1.1 \cdot 10^{-4}$	GO:0009790	embryo development
	$1.1 \cdot 10^{-4}$	GO:0048316	seed development
	$1.5 \cdot 10^{-4}$	GO:0006396	RNA processing
	$2.2 \cdot 10^{-4}$	GO:1900865	chloroplast RNA modification
PO: Plant Structure Developmental Stage	$5.9 \cdot 10^{-6}$	PO:0001078	plant embryo cotyledonary stage
	$6.6 \cdot 10^{-6}$	PO:0001081	mature plant embryo stage
Pfam	$1.5 \cdot 10^{-11}$	PPR_2	PPR repeat family
	$1.5 \cdot 10^{-11}$	PPR	PPR repeat
	$1.3 \cdot 10^{-8}$	DYW_deaminase	DYW family of nucleic acid deaminases

Πίνακας 10 - Αποτελέσματα της ανάλυσης υπερεκπροσώπησης για το emb1692

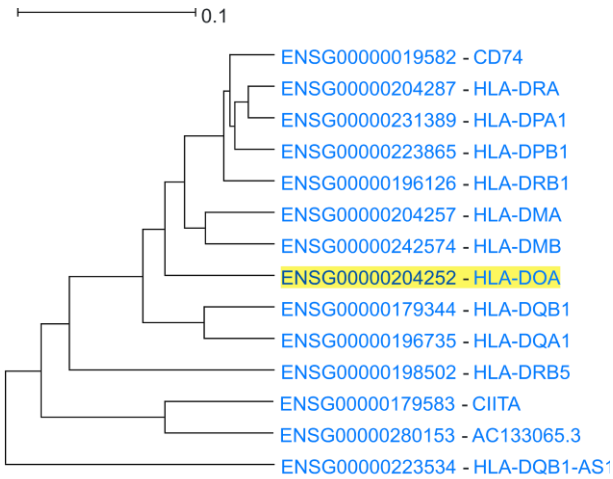
HGCA2

Πρωτεΐνες HLA

Τα γονίδια της οικογένειας HLA-DO και HLA-DM αποτελούν ένα σύνολο στενά συνεκφρασμένων γονιδίων που ο ρόλος τους είναι η φόρτωση πεπτιδίων στα τάξης ΙΙ μόρια μείζονος συμπλέγματος ιστοσυμβατότητας (Kropshofer et al., 1999). Το γονίδιο HLA-DOA (major histocompatibility complex, class II, DO alpha) χρησιμοποιήθηκε ως γονίδιο-οδηγός σε μία ανάλυση του HGCA2. Στο παραγόμενο υποδέντρο των 14 γονιδίων-φύλλων, τα 12 ήταν γονίδια HLA ή σχετιζόμενα με αυτά (Εικόνα 40). Η ανάλυση Βιολογικών Διεργασιών ανέδειξε όρους αντιγονοπαρουσίασης μέσω του MHC τάξης ΙΙ (antigen processing and presentation of exogenous peptide antigen via MHC class II, κλπ.), η Μοριακή Λειτουργία όρους πρόσδεσης στις πρωτεΐνες του MHC τάξης ΙΙ (MHC class II protein complex binding) και η ανάλυση Κυτταρικού Συστατικού επίσης έδειξε το σύμπλεγμα MHC τάξης ΙΙ ως κορυφαίο όρο. Η ανάλυση βιολογικών μονοπατιών KEGG, έδειξε την παρουσίαση και επεξεργασία αντιγόνων στον άνθρωπο ως κορυφαία λειτουργία και η ανάλυση Pfam παρουσίασε οικογένειες της α και β περιοχής του MHC τάξης ΙΙ και μία οικογένεια που αντιστοιχεί στη C1-set περιοχή της ανοσοσφαιρίνης ως υπερεκπροσωπημένες.

Αναπτύσσοντας το δέντρο έως τους 17 προγονικούς, προκύπτει ένα υποδέντρο συνέκφρασης με 97 γονίδια-φύλλα (Εικόνα 41). Μέσα σε αυτά περιέχονται γονίδια που σχετίζονται με ιντερλευκίνες (IL2RA, IL4I1, IL18BP), η ιντερφερόνη γ (IFNG), καθώς και γονίδια της οικογένειας TRGV (T-cell receptor

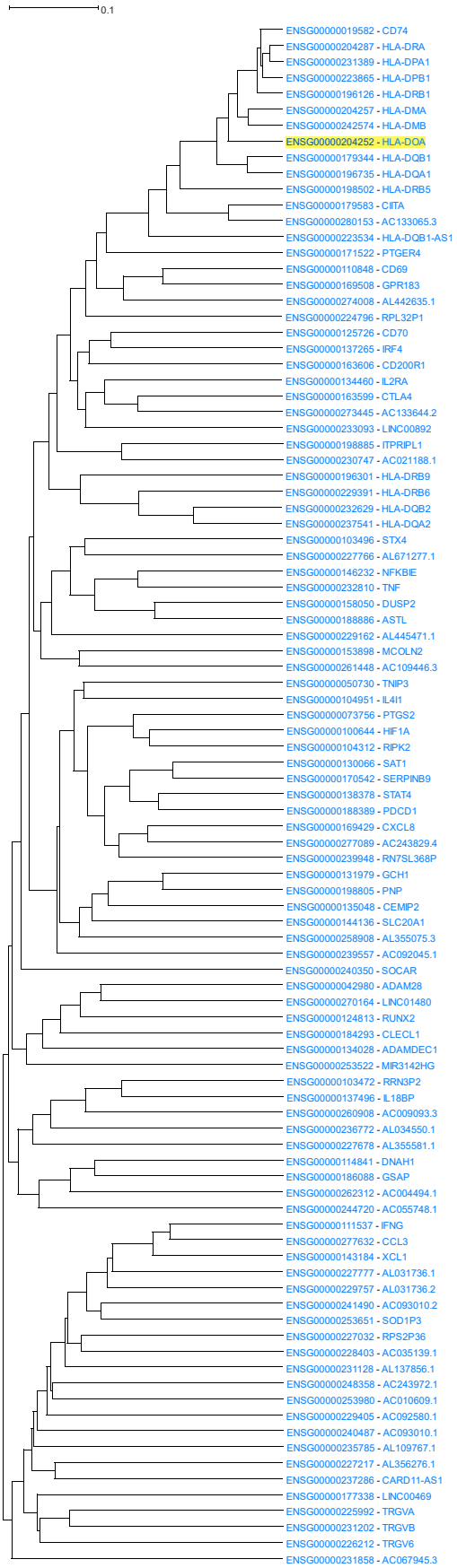
gamma). Οι αναλύσεις Βιολογικής Διεργασίας και Μοριακής λειτουργίας, εκτός από όρους σχετικούς με το MHCII, έδειξαν επιπλέον και όρους σχετικούς με την ανοσολογική απόκριση (immune response, immune receptor activity, κλπ.). Επιπλέον, τόσο οι αναλύσεις βιολογικών μονοπατιών, όσο και οι αναλύσεις ασθενειών, ανέδειξαν πολλές ασθένειες, ιδιαίτερα αυτοάνοσες, ως υπερεκπροσωπημένες, γεγονός που δείχνει το βασικό ρόλο των συνεκφραζόμενων γονιδίων στην ανοσολογική απόκριση (Πίνακας 12). Τέλος, δημιουργήθηκε και ένα δίκτυο αλληλεπίδρασης μεταξύ πρωτεϊνών μέσω του STRING για τα γονίδια του επεκταμένου υποδέντρου (Εικόνα 42).



Εικόνα 40 – Το υποδέντρο συνέκφρασης για το HLA-DOA

Enrichment Summary for HLA-DOA			
Category	p-value	Term ID	Description
GO: Biological Process	$1.9 \cdot 10^{-19}$	GO:0019886	antigen processing and presentation of exogenous peptide antigen via MHC class II
	$1.9 \cdot 10^{-19}$	GO:0002495	antigen processing and presentation of peptide antigen via MHC class II
	$1.9 \cdot 10^{-19}$	GO:0002504	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II
GO: Molecular Function	$1.8 \cdot 10^{-15}$	GO:0023026	MHC class II protein complex binding
	$1.2 \cdot 10^{-14}$	GO:0032395	MHC class II receptor activity
GO: Cellular Component	$1.9 \cdot 10^{-34}$	GO:0042613	MHC class II protein complex
KEGG	$2.1 \cdot 10^{-24}$	hsa04612	Antigen processing and presentation - Homo sapiens (human)
Pfam	$1.6 \cdot 10^{-25}$	C1-set	Immunoglobulin C1-set domain
	$4.4 \cdot 10^{-16}$	MHC_II_alpha	Class II histocompatibility antigen, alpha domain
	$1.0 \cdot 10^{-15}$	MHC_II_beta	Class II histocompatibility antigen, beta domain

Πίνακας 11 – Αποτελέσματα της ανάλυσης υπερεκπροσώπησης για το HLA-DOA



Εικόνα 41 – Το υποδέντρο συνέκρασης για το HLA-DOA με 17 προγονικούς κόμβους

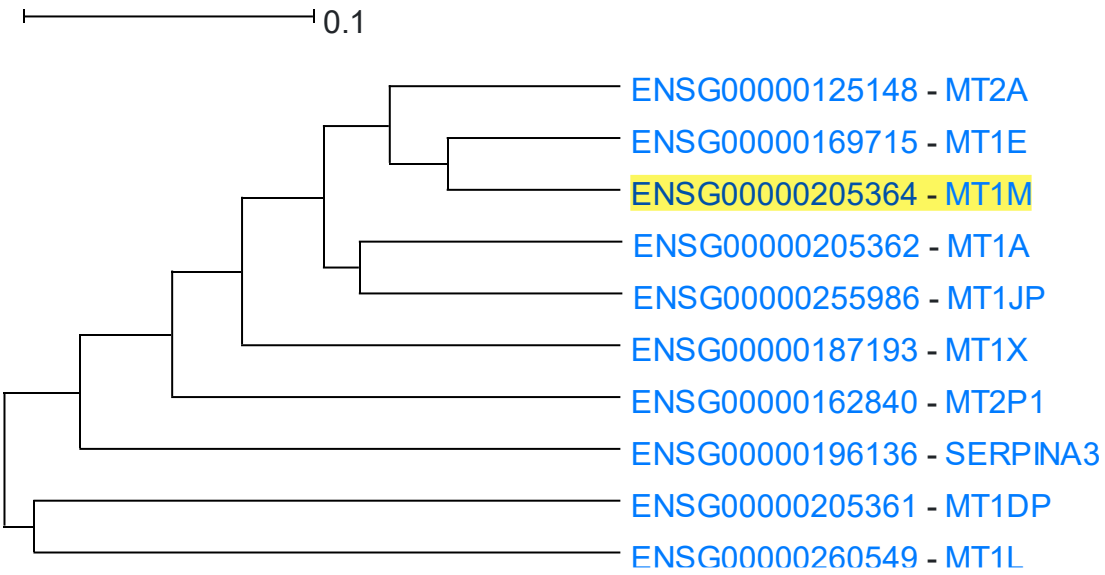
Enrichment Summary for HLA-DOA			
Category	p-value	Term ID	Description
GO: Biological Process	$1.5 \cdot 10^{-16}$	GO:0006955	immune response
	$2.6 \cdot 10^{-16}$	GO:0002250	adaptive immune response
	$3.1 \cdot 10^{-16}$	GO:0019221	cytokine-mediated signaling pathway
	$1.0 \cdot 10^{-15}$	GO:0002682	regulation of immune system process
GO: Molecular Function	$1.6 \cdot 10^{-10}$	GO:0140375	immune receptor activity
KEGG	$1.1 \cdot 10^{-20}$	hsa05330	Allograft rejection - Homo sapiens (human)
	$1.6 \cdot 10^{-10}$	hsa05332	Graft-versus-host disease - Homo sapiens (human)
	$1.6 \cdot 10^{-10}$	hsa05323	Rheumatoid arthritis - Homo sapiens (human)
WikiPathways	$1.5 \cdot 10^{-20}$	WP2328_r106557	Allograft Rejection
	$8.0 \cdot 10^{-12}$	WP4217_r101851	Ebola Virus Pathway on Host
DisGeNet	$2.2 \cdot 10^{-29}$	C0036202	Sarcoidosis
	$2.7 \cdot 10^{-17}$	C0004364	Autoimmune Diseases
	$1.5 \cdot 10^{-16}$		Rheumatoid Arthritis
	$1.5 \cdot 10^{-16}$		Ankylosing spondylitis

Πίνακας 12 – Επιπλέον όροι που εμφανίστηκαν με την επέκταση του δέντρου του HLA-DOA στους 17 προγονικούς κόμβους

Εικόνα 42 – Το String δίκτυο αλληλεπίδρασης για το HLA-DOA με 17 προγονικούς κόμβους. Αναγνωρίσθηκαν από το πρόγραμμα τα 53 από τα 97 γονίδια. Παρατηρούμε μία αναμενόμενη μεγάλη σύνδεση μεταξύ των HLA γονιδίων αλλά και μεγάλη συσχέτιση μεταξύ των υπολοίπων γονιδίων οφειλούμενη στην κοινή τους λειτουργία που είναι η ανοσολογική απόκριση.

Μεταλλοθειονίνες

Οι μεταλλοθειονίνες έχουν ένα μεγάλο ποσοστό καταλοίπων κυστεΐνης και προσδένονται σε διάφορα βαρέα μέταλλα. Ρυθμίζονται σε επίπεδο μεταγραφής τόσο από βαρέα μέταλλα όσο και από γλυκοκορτικοειδή (Murphy et al., 2008). Χρησιμοποιήθηκε το γονίδιο MT1M (Metallothionein 1M) ως γονίδιο-οδηγός. Το παραγόμενο υποδέντρο αναπτύχθηκε έως τους 7 προγονικούς κόμβους και περιείχε 10 γονίδια (Εικόνα 43). Τα 9 από αυτά ανήκουν στις μεταλλοθειονίνες, με 4 από αυτά να είναι ψευδογονίδια και ανεπαρκώς σχολιασμένα. Η ανάλυση Βιολογικής Διεργασίας κατέταξε όρους σχετικούς με αντίδραση και αποτοξίνωση ιόντων χαλκού και μετάλλου, καθώς και καδμίου και ψευδαργύρου πιο κάτω στη λίστα, ως κορυφαία υπερεκπροσωπημένους (Πίνακας 13). Η ανάλυση Μοριακής Λειτουργίας έδειξε επίσης όρους πρόσδεσης σε βαρέα μέταλλα. Η ανάλυση βιολογικών μονοπατιών KEGG πρότεινε τον όρο «απορρόφηση μετάλλων» (mineral absorption) στον άνθρωπο και η WikiPathways τους όρους «ομοιόσταση ψευδαργύρου» (Zinc homeostasis) και «ομοιόσταση χαλκού» (Copper homeostasis). Η ανάλυση Pfam ανέθεσε τις πρωτεΐνες των συνεκφρασμένων γονιδίων στην οικογένεια της Μεταλλοθειονίνης. Οι κορυφαίες ασθένειες που εμφανίστηκαν μέσω του DisGeNet ήταν η Μεταταρσαλγία και η έλλειψη μελατονίνης. Τέλος, η ανάλυση μεταγραφικών παραγόντων μέσω του ReMap αποκάλυψε δύο μεταγραφικούς παράγοντες της οικογένειας δακτύλων ψευδαργύρου (zinc finger protein 879 και 26) να στοχεύουν τα γονίδια του υποδέντρου.



Εικόνα 43 – Το υποδέντρο συνέκφρασης για το γονίδιο MT1M

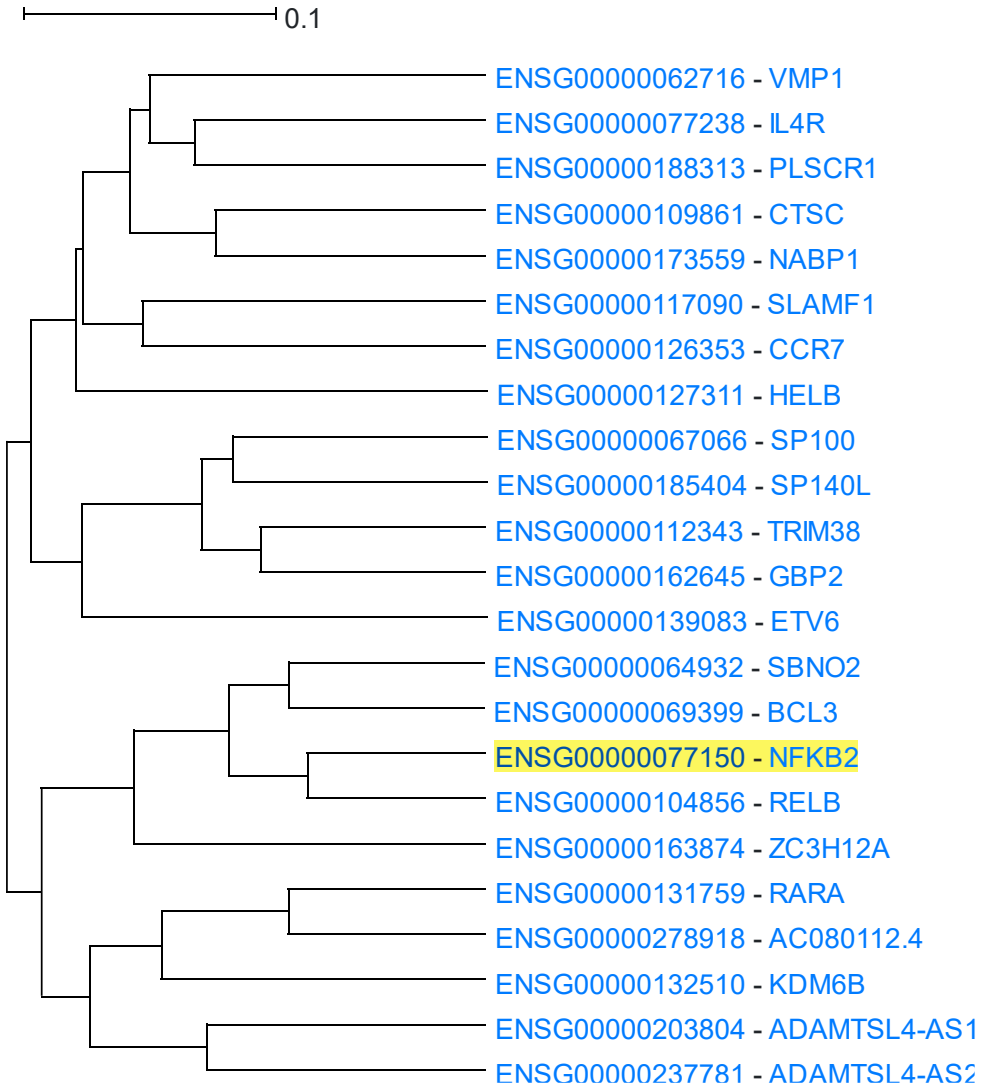
Enrichment Summary for HLA-DOA			
Category	p-value	Term ID	Description
GO: Biological Process	$2.6 \cdot 10^{-14}$	GO:1990169	stress response to copper ion
	$2.6 \cdot 10^{-14}$	GO:0010273	detoxification of copper ion
	$2.9 \cdot 10^{-14}$	GO:0097501	stress response to metal ion
	$2.9 \cdot 10^{-14}$	GO:0061687	detoxification of inorganic compound
GO: Molecular Function	$8.8 \cdot 10^{-6}$	GO:0008270	zinc ion binding
	$1.8 \cdot 10^{-5}$	GO:0046914	transition metal ion binding
KEGG	$1.8 \cdot 10^{-11}$	hsa04978	Mineral absorption - Homo sapiens (human)
WikiPathways	$7.9 \cdot 10^{-13}$	WP3529_r106738	Zinc homeostasis
	$3.9 \cdot 10^{-12}$	WP3286_r106367	Copper homeostasis
Pfam	$3.0 \cdot 10^{-16}$	Metallothio	Metallothionein
DisGeNet	$4.8 \cdot 10^{-16}$	C0025587	Metatarsalgia
	$6.9 \cdot 10^{-16}$	C4285716	Melatonin deficiency
ReMap	$7.6 \cdot 10^{-5}$	ZNF879	zinc finger protein 879
	$1.7 \cdot 10^{-2}$	ZNF26	zinc finger protein 26

Πίνακας 13 - Αποτελέσματα της ανάλυσης υπερεκπροσώπησης για το MT1M

Γενική Ανοσολογική Απόκριση

Το γονίδιο NFKB2 (Nuclear Factor Kappa B Subunit 2) κωδικοποιεί μία υπομονάδα του NF-*κ*B ο οποίος παίζει κεντρικό ρόλο ως ενεργοποιητής γονιδίων στην φλεγμονή και την ανοσοαπόκριση. Εισάγοντας τον NFKB2 στο HGCA2, παράγεται ένα υποδέντρο συνέκφρασης με 23 γονίδια-φύλλα. Παρατηρούμε ότι στο ακριβώς διπλανό φύλλο από το NFKB2, υπάρχει το ομόλογό του, RELB (Εικόνα 44). Στον ίδιο υποκλάδο βρίσκεται και το γονίδιο

BCL3, ένας μη-τυπικός αναστολέας του NF-κB (Oeckinghaus and Ghosh, 2009). Η ανάλυση Βιολογικής Διεργασίας, έδειξε ως κορυφαίο τον όρο «ανοσολογική απόκριση» (immune response), όπως και άλλους όρους σχετικούς με την άμυνα (Πίνακας 14). Η ανάλυση Κυτταρικού Συστατικού αναμενόμενα έδειξε δύο όρους σχετιζόμενους με συμπλέγματα της οικογένειας NF-κB ως κορυφαίους, καθώς και ανέδειξε τον πυρήνα ως την υπερεκπροσωπημένη περιοχή του κυττάρου που συναντώνται τα περισσότερα από τα συνεκφραζόμενα γονίδια. Οι δύο κορυφαίοι μεταγραφικοί παράγοντες που αναδείχθηκαν μέσω της ανάλυσης του Encode, ήταν ο WRNIP1 και ο RELA, όπου ο τελευταίος αποτελεί μέλος της οικογένειας του NF-κB.



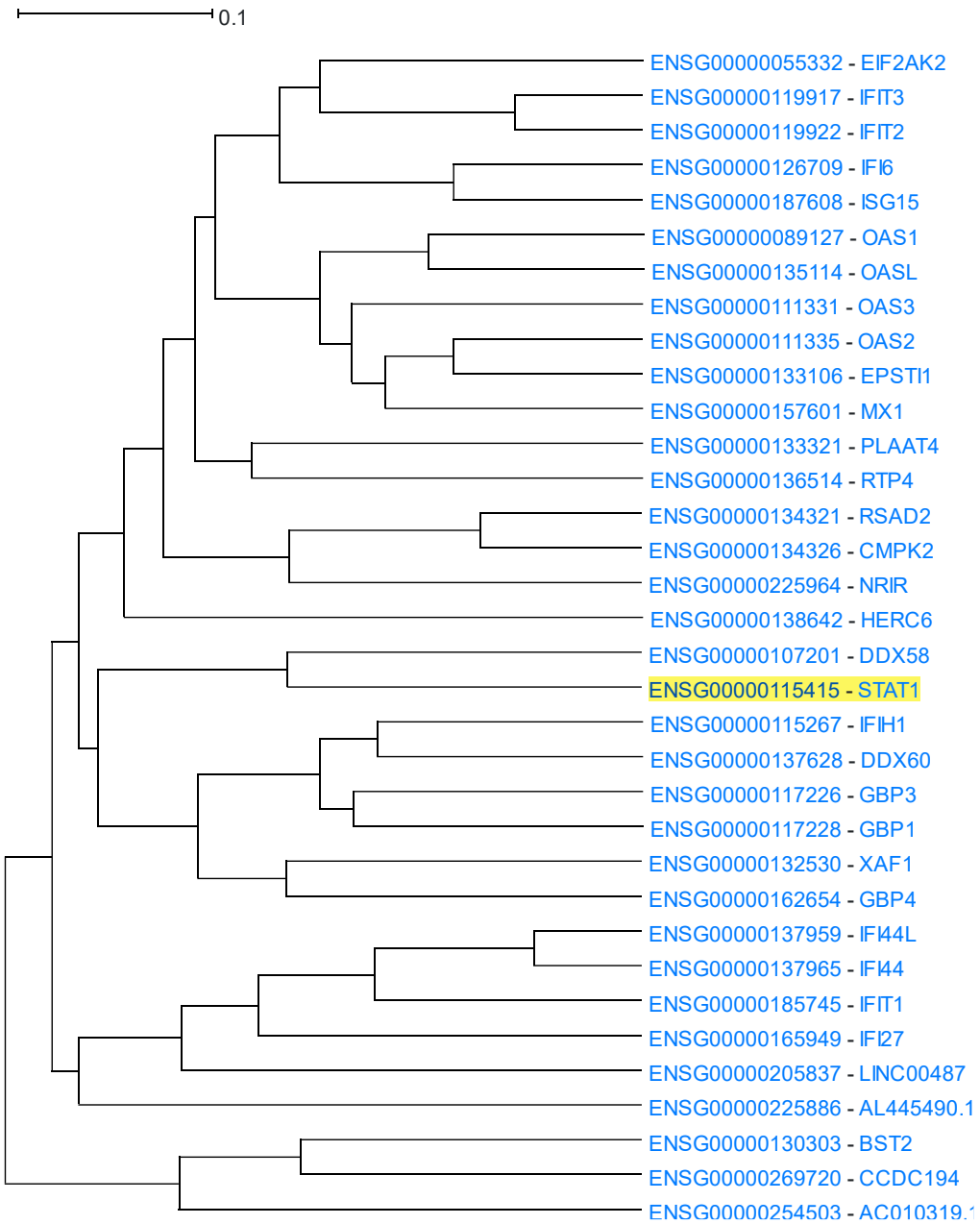
Εικόνα 44 - Το υποδέντρο συνέκφρασης για το γονίδιο NFKB2

Enrichment Summary for NFKB2			
Category	p-value	Term ID	Description
GO: Biological Process	$1.1 \cdot 10^{-7}$	GO:0006952	defense response
	$1.1 \cdot 10^{-7}$	GO:0006955	immune response
	$1.1 \cdot 10^{-7}$	GO:0051707	response to other organism
	$1.1 \cdot 10^{-7}$	GO:0043207	response to external biotic stimulus
GO: Cellular Component	$2.4 \cdot 10^{-6}$	GO:0033256	I-kappaB/NF-kappaB complex
	$3.5 \cdot 10^{-5}$	GO:0033257	Bcl3/NF-kappaB2 complex
	$1.2 \cdot 10^{-4}$	GO:0031981	nuclear lumen
	$6.0 \cdot 10^{-4}$	GO:0005634	nucleus
Encode	$2.3 \cdot 10^{-4}$	WRNIP1	WRN helicase interacting protein 1
	$2.3 \cdot 10^{-4}$	RELA	RELA proto-oncogene, NF-kB subunit

Πίνακας 14 - Αποτελέσματα της ανάλυσης υπερεκπροσώπησης για το NFKB2

Απόκριση σε ιούς

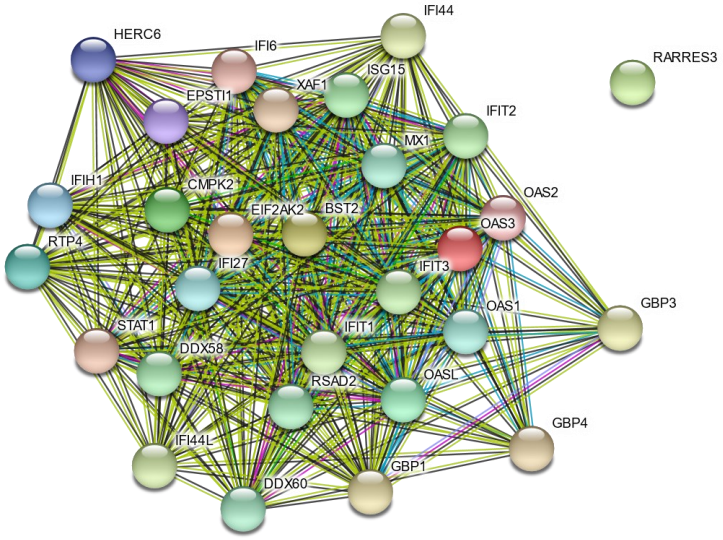
Το γονίδιο STAT1 (Signal Transducer And Activator Of Transcription 1) που κωδικοποιεί την αντίστοιχη πρωτεΐνη η οποία μεσολαβεί για την έκφραση διαφόρων γονιδίων που παίζουν ρόλο στην επιβίωση του κυττάρου ως απόκριση σε διάφορα ερεθίσματα και παθογόνα (Najjar and Fagard, 2010), χρησιμοποιήθηκε ως γονίδιο-οδηγός στο HGCA2. Παράχθηκε ένα υποδέντρο με 34 γονίδια-φύλλα, πολλά από τα οποία σχετίζονται με ιντερφερόνες (Εικόνα 45). Ανάλυση Βιολογικής Διεργασίας ανέδειξε ως κορυφαίους, όρους σχετικούς με την άμυνα ενάντια σε ιούς (defence response to virus). Οι αναλύσεις KEGG και DisGeNet έδειξαν συσχέτιση με διάφορες ιογενείς ασθένειες. Ενδιαφέρον έχουν τα αποτελέσματα της ανάλυσης WikiPathways, που ανακάλυψε τη συμμετοχή του STAT1, καθώς και άλλων γονιδίων του υποδέντρου, όπως τα γονίδια της οικογένειας OAS, στην απόκριση κατά των ανθρώπινων κοροναϊών (Πίνακας 15). Τέλος, η ανάλυση μεταγραφικών παραγόντων έδειξε τον παράγοντα STAT2 ως κορυφαίο μεταγραφικό παράγοντα που στοχεύει πάνω από τα 2/3 των γονιδίων του υποδέντρου. Ως επιπλέον ανάλυση, δημιουργήθηκε ένα δίκτυο αλληλεπίδρασης πρωτεϊνών μέσω του STRING, με είσοδο τα γονίδια του υποδέντρου (Εικόνα 46), το οποίο δείχνει μία στενή σχέση μεταξύ τους.



Εικόνα 45 - Το υποδέντρο συνέκφρασης για το γονίδιο STAT1

Enrichment Summary for STAT1			
Category	p-value	Term ID	Description
GO: Biological Process	$2.2 \cdot 10^{-33}$	GO:0051607	defense response to virus
	$4.1 \cdot 10^{-33}$	GO:0009615	response to virus
KEGG	$7.7 \cdot 10^{-11}$	hsa05160	Hepatitis C - Homo sapiens (human)
	$8.4 \cdot 10^{-11}$	hsa05164	Influenza A - Homo sapiens (human)
WikiPathways	$2.1 \cdot 10^{-11}$	WP4880_r109979	Host-pathogen interaction of human corona viruses - Interferon induction
	$3.4 \cdot 10^{-8}$	WP4868_r109974	Type I Interferon Induction and Signaling During SARS-CoV-2 Infection
Pfam	$6.0 \cdot 10^{-11}$	OAS1_C	2'-5'-oligoadenylate synthetase 1, domain 2, C-terminus
DisGeNet	$4.7 \cdot 10^{-28}$	C0021400	Influenza
	$6.3 \cdot 10^{-13}$	C0042769	Virus Diseases
	$1.8 \cdot 10^{-10}$	C0019196	Hepatitis C
Encode	$2.6 \cdot 10^{-40}$	STAT2	signal transducer and activator of transcription 2
ReMap	$5.9 \cdot 10^{-16}$	STAT2	signal transducer and activator of transcription 2

Πίνακας 15 - Αποτελέσματα της ανάλυσης υπερεκπροσώπησης για το STAT1

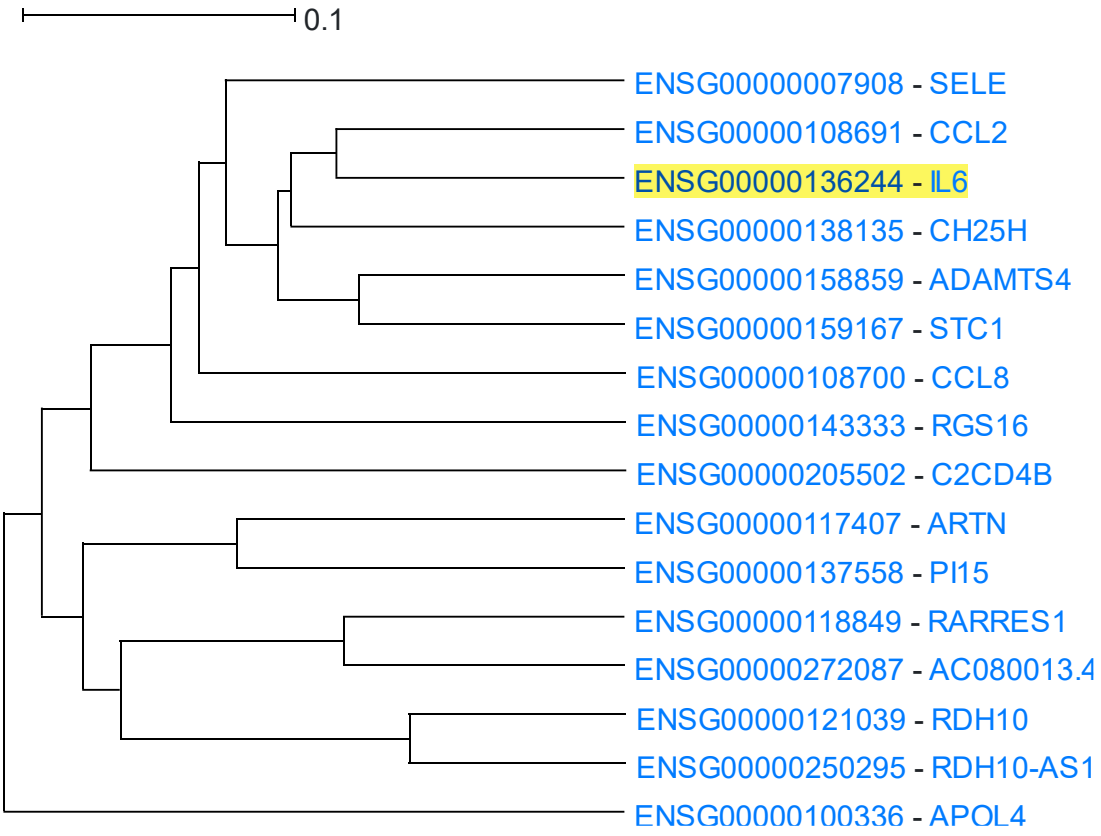


Εικόνα 46 – Το δίκτυο STRING για το γονίδιο STAT1

Κυτοκίνες

Το γονίδιο IL6 (Interleukin 6) που κωδικοποιεί μία κυτοκίνη η οποία έχει λειτουργικό ρόλο στην φλεγμονή και την ωρίμανση των Β λεμφοκυττάρων, χρησιμοποιήθηκε ως γονίδιο-οδηγός σε μία ανάλυση του HGCA2. Το παραγόμενο υποδέντρο αναπτύχθηκε έως τους 9 προγονικούς κόμβους, με 16 γονίδια-φύλλα (Εικόνα 47). Η ανάλυση Βιολογικής Διεργασίας έδειξε όρους σχετικούς με μετανάστευση και χημειοταξία λευκοκυττάρων και

λεμφοκυττάρων, η ανάλυση Μοριακής Λειτουργίας περιείχε τους όρους «ενεργότητα κυτοκίνης» (cytokine activity) και «ενεργότητα χημειοκίνης» (chemokine activity) και τα βιολογικά μονοπάτια KEGG έδειξαν επίσης όρο σχετιζόμενο με κυτοκίνη. Τα παραπάνω αποτελέσματα συμβαδίζουν με τις γενικές λειτουργίες του IL6, ιδιαίτερα όσον αφορά την χημειοταξία (Weissenbach et al., 2004). Επιπλέον, τόσο η KEGG όσο και η WikiPathways, έδειξαν ως υπερεκπροσωπημένο το μονοπάτι σηματοδότησης του TNFα. Τέλος, αρκετό ενδιαφέρον παρουσιάζει το γεγονός ότι στα αποτελέσματα, ως εξίσου εμπλουτισμένο εμφανίστηκε και το μονοπάτι μόλυνσης του COVID-19 (COVID-19 Adverse Outcome Pathway) (Πίνακας 16).



Εικόνα 47 - Το υποδέντρο συνέκφρασης για το γονίδιο IL6

Enrichment Summary for IL6			
Category	p-value	Term ID	Description
GO: Biological Process	$3.4 \cdot 10^{-4}$	GO:0050900	leukocyte migration
	$3.4 \cdot 10^{-4}$	GO:0072676	lymphocyte migration
	$1.3 \cdot 10^{-3}$	GO:0048247	lymphocyte chemotaxis
	$1.3 \cdot 10^{-3}$	GO:0030595	leukocyte chemotaxis
GO: Molecular Function	$3.2 \cdot 10^{-3}$	GO:0005125	cytokine activity
	$3.2 \cdot 10^{-3}$	GO:0008009	chemokine activity
KEGG	$2.2 \cdot 10^{-4}$	hsa04668	TNF signaling pathway - Homo sapiens (human)
	$2.3 \cdot 10^{-3}$	hsa04060	Cytokine-cytokine receptor interaction - Homo sapiens (human)
WikiPathways	$9.2 \cdot 10^{-4}$	WP231_r105836	TNF alpha Signaling Pathway
	$9.2 \cdot 10^{-4}$	WP4891_r109962	COVID-19 Adverse Outcome Pathway

Πίνακας 16 - Αποτελέσματα της ανάλυσης υπερεκπροσώπησης για το IL6

Όσφρηση

Οι οσφρητικοί υποδοχείς (olfactory receptors) αποτελούν οικογένεια γονιδίων που είναι υπαίτια για την αίσθηση της όσφρησης (Gaillard et al., 2004). Χρησιμοποιήθηκε το γονίδιο OR1D2 (Olfactory Receptor Family 1 Subfamily D Member 2) ως είσοδος στο HGCA2 και το παραγόμενο υποδέντρο αναπτύχθηκε έως τους 96 προγονικούς κόμβους και περιείχε 392 γονίδια. Η ανάλυση Βιολογικών Διεργασιών ανέδειξε όρους σχετικούς με ανίχνευση ερεθισμάτων ως υπερεκπροσωπημένους. Κορυφαίος όρος ήταν η «ανίχνευση χημικού ερεθίσματος που εμπλέκεται στην αισθητική αντίληψη της όσφρησης» (detection of chemical stimulus involved in sensory perception of smell), ο οποίος περιγράφει τα 180 από τα 392 γονίδια του υποδέντρου. Για τον συγκεκριμένο όρο, υπάρχουν συνολικά 387 γονίδια που περιγράφονται γενικώς από αυτόν, τα 180 από τα οποία (46.5%) βρίσκονται στο εν λόγω υποδέντρο. Αυτό έχεις ως αποτελέσματα να έχουμε ~35 φορές υπερεκπροσώπηση καθώς και ένα πάρα πολύ μικρό p-value ($7.2 \cdot 10^{-269}$). Αντίστοιχα, η Μοριακή Λειτουργία έδειξε τους όρους «ενεργότητα οσφρητικών υποδοχέων» (olfactory receptor activity) και «ενεργότητα υποδοχέων συζευγμένων με G-πρωτεΐνες» (G protein-coupled receptor activity) και η ανάλυση Κυτταρικού Συστατικού έδειξε τα γονίδια να αποτελούν μέρος μεμβράνης. Τα παραπάνω συμφωνούν με το γεγονός ότι οι οσφρητικοί υποδοχείς είναι μέλη της μεγάλης οικογένειας των υποδοχέων που συνδέονται

με G-πρωτεΐνες και συνεπώς είναι φυσικό να συσχετίζονται με την κυτταρική μεμβράνη. Η ανάλυση KEGG παρομοίως εμφάνισε τον όρο «Οσφρητική Μεταγωγή στον άνθρωπο» (Olfactory transduction - Homo sapiens (human)) και η ανάλυση Pfam κατέταξε τα ίδια 180 γονίδια στην οικογένεια των οσφρητικών υποδοχέων. Επιπλέον, η ανάλυση WikiPathways εμφάνισε αναμενόμενα ως κορυφαίους όρους σχετικούς με υποδοχείς που συνδέονται με G-πρωτεΐνες.

Enrichment Summary for OR1D2			
Category	p-value	Term ID	Description
GO: Biological Process	$7.2 \cdot 10^{-269}$	GO:0050911	detection of chemical stimulus involved in sensory perception of smell
	$9.4 \cdot 10^{-265}$	GO:0050907	detection of chemical stimulus involved in sensory perception
	$1.2 \cdot 10^{-261}$	GO:0007608	sensory perception of smell
GO: Molecular Function	$1.5 \cdot 10^{-285}$	GO:0004984	olfactory receptor activity
	$9.1 \cdot 10^{-236}$	GO:0004930	G protein-coupled receptor activity
GO: Cellular Component	$9.3 \cdot 10^{-65}$	GO:0016021	integral component of membrane
	$5.9 \cdot 10^{-63}$	GO:0031224	intrinsic component of membrane
	$9.2 \cdot 10^{-63}$	GO:0005886	plasma membrane
KEGG	$5.9 \cdot 10^{-225}$	hsa04740	Olfactory transduction - Homo sapiens (human)
WikiPathways	$1.2 \cdot 10^{-18}$	WP455_r106426	GPCRs, Class A Rhodopsin-like
	$3.3 \cdot 10^{-11}$	WP117_r107421	GPCRs, Other
Pfam	$1.2 \cdot 10^{-278}$	7tm_4	Olfactory receptor

Πίνακας 17 - Αποτελέσματα της ανάλυσης υπερεκπροσώπησης για το OR1D2

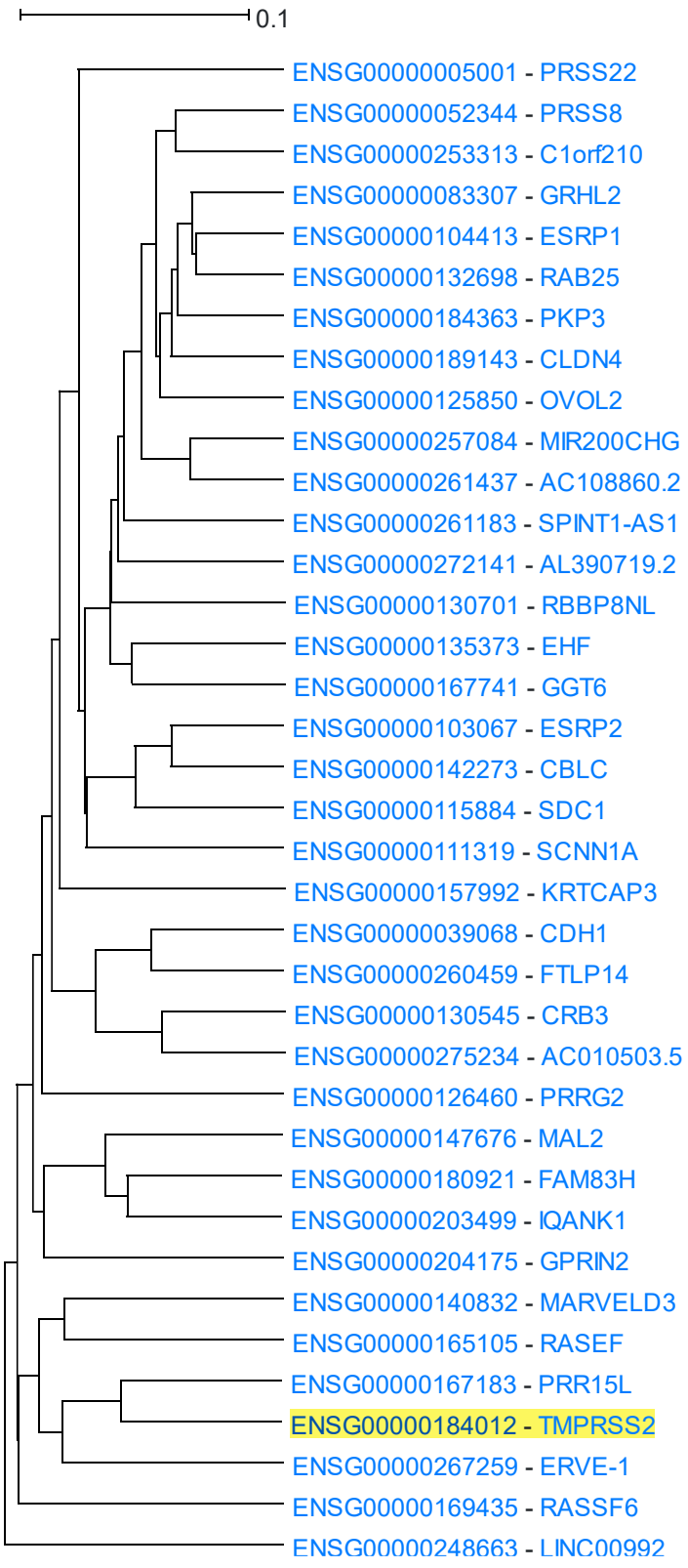
Υποδοχείς COVID-19

Έχει βρεθεί ότι ο ιός COVID-19 προσβάλει τον ανθρώπινο οργανισμό μέσω του υποδοχέα ACE2 σε συνδυασμό με τον TMPRSS2 (Mollica et al., 2020). Το γονίδιο TMPRSS2 (Transmembrane Serine Protease 2), χρησιμοποιήθηκε ως γονίδιο-οδηγός σε μία ανάλυση του HGCA2. Το παραγόμενο υποδέντρο αναπτύχθηκε έως τους 6 προγονικούς κόμβους και περιείχε 37 γονίδια-φύλλα (Εικόνα 48). Η ανάλυση Βιολογικής Διεργασίας ανέδειξε όρους σχετικά με επιθηλιακά κύτταρα και σύνδεση μεταξύ κυττάρων, πράγμα που συμβαδίζει με το γεγονός ότι ο COVID-19 προσδένεται σε επιθηλιακά κύτταρα. Οι αναλύσεις Κυτταρικού Συστατικού και βιολογικών μονοπατιών KEGG, έδειξαν επίσης όρους σχετικούς με σύνδεση μεταξύ

κυττάρων. Όσον αφορά τις αναλύσεις Μεταγραφικών Παραγόντων, η Encode έδειξε ανάμεσα στους τρεις κορυφαίους παράγοντες, δύο σχετικούς με την οικογένεια δακτύλων ψευδαργύρου αλλά και τον παράγοντα ESR1 (Estrogen Receptor 1). Η ανάλυση ReMap ανακάλυψε ότι ο ESR1 στοχεύει τα 36 από τα 37 γονίδια του υποδέντρου συνέκφρασης. Η ύπαρξη του ESR1 ως στοχεύον παράγοντας σε γονίδια συνεκφραζόμενα με το TMPRSS2 μπορεί να εξηγεί το γεγονός ότι ο COVID-19 είναι λιγότερο θανατηφόρος σε άτομα θηλυκού από ότι αρσενικού γένους (Jin et al., 2020).

Enrichment Summary for TMPRSS2			
Category	p-value	Term ID	Description
GO: Biological Process	$1.7 \cdot 10^{-4}$	GO:0007043	cell-cell junction assembly
	$3.2 \cdot 10^{-4}$	GO:0030855	epithelial cell differentiation
	$3.2 \cdot 10^{-4}$	GO:0045216	cell-cell junction organization
	$9.6 \cdot 10^{-4}$	GO:0060429	epithelium development
GO: Cellular Component	$1.4 \cdot 10^{-3}$	GO:0043296	apical junction complex
	$1.4 \cdot 10^{-3}$	GO:0005911	cell-cell junction
KEGG	$1.5 \cdot 10^{-4}$	hsa04514	Cell adhesion molecules (CAMs) - Homo sapiens (human)
	$1.5 \cdot 10^{-3}$	hsa04530	Tight junction - Homo sapiens (human)
Encode	$8.9 \cdot 10^{-5}$	ZNF217	zinc finger protein 217
	$4.2 \cdot 10^{-4}$	ESR1	estrogen receptor 1
	$4.3 \cdot 10^{-3}$	ZBTB7A	zinc finger and BTB domain containing 7A

Πίνακας 18 - Αποτελέσματα της ανάλυσης υπερεκπροσώπησης για το TMPRSS2

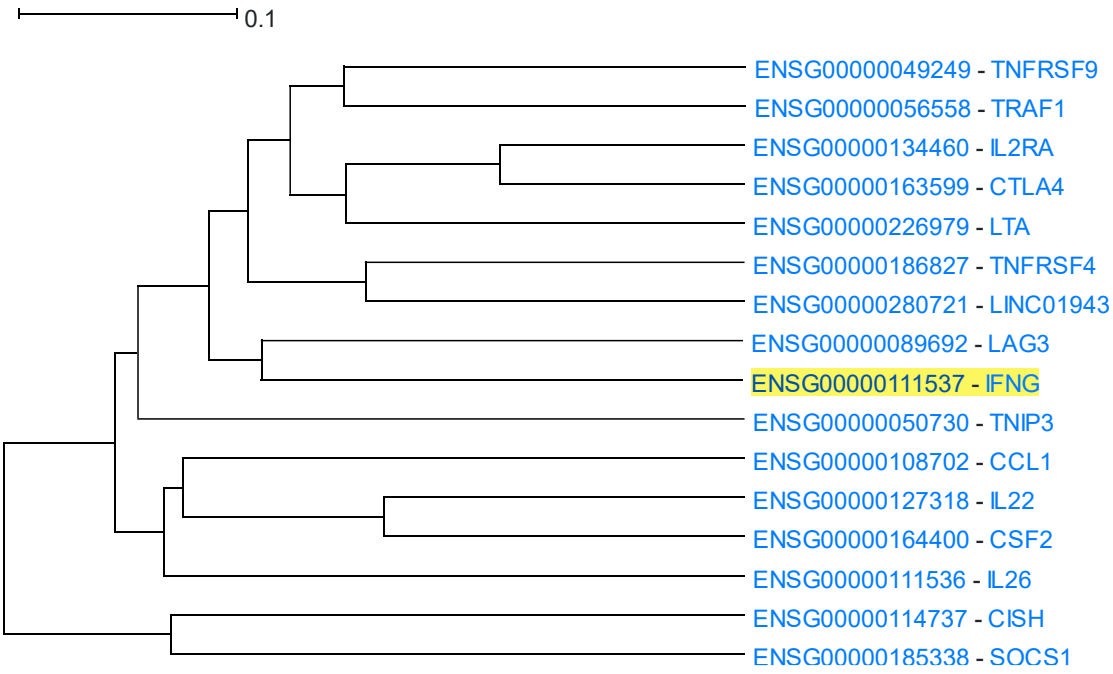


Εικόνα 48 - Το υποδέντρο συνέκφρασης για το γονίδιο TMPRSS2

HGCA1.5

Κυτοκίνες

Η IFNG (Interferon γ) είναι μία κυτοκίνη που έχει σημαντικό ρόλο στην φυσική αλλά και στην επίκτητη ανοσία, ενάντια σε διάφορες μολύνσεις (ιικές, βακτηριακές). Το IFNG χρησιμοποιήθηκε ως είσοδος στο HGCA1.5 και το παραγόμενο υποδέντρο περιείχε άλλα γονίδια κυτοκινών, καθώς και πολλές ιντερλευκίνες (Εικόνα 49). Πράγματι, οι αναλύσεις Βιολογικής Διεργασίας, Μοριακής Λειτουργίας και βιολογικών μονοπατιών KEGG έδειξαν όρους συσχετιζόμενους με κυτοκίνες ως υπερεκπροσωπημένους (Πίνακας 19). Παράλληλα, οι αναλύσεις ασθενειών τόσο από OMIM όσο και από DisGeNet, αποκάλυψαν σχέση των γονιδίων του υποδέντρου με την ασθένεια της φυματίωσης, κάτι αρκετά ενδιαφέρον καθώς έχει αποδειχθεί ότι ο IFNG βελτιώνει την αντίδραση των μακροφάγων ενάντια στην εν λόγω ασθένεια (Khan et al., 2016).



Εικόνα 49 - Το υποδέντρο συνέκφρασης για το γονίδιο IFNG από το εργαλείο HGCA1.5

Enrichment Summary for IFNG			
Category	p-value	Term ID	Description
GO: Biological Process	$2.8 \cdot 10^{-11}$	GO:0019221	cytokine-mediated signaling pathway
	$1.0 \cdot 10^{-9}$	GO:0071345	cellular response to cytokine stimulus
	$1.7 \cdot 10^{-9}$	GO:0034097	response to cytokine
GO: Molecular Function	$5.7 \cdot 10^{-7}$	GO:0005125	cytokine activity
	$7.3 \cdot 10^{-7}$	GO:0005126	cytokine receptor binding
KEGG	$2.9 \cdot 10^{-9}$	hsa04060	Cytokine-cytokine receptor interaction - Homo sapiens (human)
WikiPathways	$2.1 \cdot 10^{-11}$	WP4880_r109979	Host-pathogen interaction of human corona viruses - Interferon induction
	$3.4 \cdot 10^{-8}$	WP4868_r109974	Type I Interferon Induction and Signaling During SARS-CoV-2 Infection
OMIM	$9.2 \cdot 10^{-5}$	#607948	MYCOBACTERIUM TUBERCULOSIS, SUSCEPTIBILITY TOMYCOBACTERIUM TUBERCULOSIS, PROTECTION AGAINST, INCLUDED
DisGeNet	$2.4 \cdot 10^{-14}$	C1609538	Latent Tuberculosis

Πίνακας 19 - Αποτελέσματα της ανάλυσης υπερεκπροσώπησης για το IFNG από το εργαλείο HGCA1.5

Συζήτηση

ACT

Το ACT ταυτοποιεί συνεκφραζόμενα γονίδια με ένα γονίδιο επιλεγμένο από τον χρήστη. Η έξοδός του είναι ένα δέντρο του οποίου τα φύλλα αποτελούνται από το γονίδιο επιλογής και γονίδια με παρόμοιο προφίλ έκφρασης, γεγονός που υποθέτει συμμετοχή τους σε παρόμοιες βιολογικές διεργασίες και βιολογικά μονοπάτια. Με αυτόν τον τρόπο ταυτοποιούνται δυνητικοί λειτουργικοί συνεργάτες του γονιδίου ενδιαφέροντος. Επίσης, μπορούμε να συμπεράνουμε τα χαρακτηριστικά ενός γονιδίου με άγνωστη λειτουργία, μελετώντας τα υπόλοιπα γονίδια του υποδέντρου και τους στατιστικά σημαντικούς υπερεκπροσωπημένους όρους. Οι αναλύσεις του ACT επαληθεύονται από παραδείγματα τα οποία αναπαράγουν ήδη γνωστή βιολογία. Για παράδειγμα, όσον αφορά το σύμπλεγμα του ριβοσώματος, όλες οι πρωτεΐνες που απαρτίζουν ένα ριβοσωμικό σύμπλεγμα, αναμένουμε να εκφράζονται κατά την διάρκεια της βιογένεσης του ριβοσώματος. Έτσι, χρησιμοποιώντας μία ριβοσωμική πρωτεΐνη ως γονίδιο-οδηγό, ένα σύνολο από 134 γονίδια κωδικοποιούντα για δομικά συστατικά του ριβοσώματος ανακαλύφθηκαν ομαδοποιημένα σε ένα κοινό υποδέντρο. Επίσης, τα γονίδια στο χλωροπλαστικό DNA αναμένονται να είναι ομαδοποιημένα σε μία ανάλυση συνέκφρασης, καθώς υπάρχουν ιστοί που δεν περιέχουν χλωροπλάστες, όπως π.χ. ρίζες, και υπάρχουν οι πράσινοι ιστοί που περιέχουν, όπως π.χ. τα φύλλα. Η υπόθεση αυτή επαληθεύθηκε από το ACT, καθώς όλα τα χλωροπλαστικά γονίδια βρέθηκαν ομαδοποιημένα σε ένα κοινό κλάδο. Τέλος, μελετώντας ως γονίδιο-οδηγό το CTL2 του οποίου πιθανός ρόλος είναι η βιοσύνθεση του κυτταρικού τοιχώματος, βρέθηκε ομαδοποιημένο με πρωτεΐνες που εμπλέκονται στη σύνθεση κυτταρίνης και γενικά με γονίδια συσχετιζόμενα με τη βιογένεση του κυτταρικού τοιχώματος, και παράλληλα ανακαλύφθηκε ο VND7 ως βασικός μεταγραφικός παράγοντας που ρυθμίζει τα συνεκφραζόμενα γονίδια, ένα εύρημα το οποίο είχε ήδη επιβεβαιωθεί πειραματικά (Yamaguchi et al., 2011).

Σύγκριση δυνατοτήτων ανταγωνιστικών εργαλείων με το ACT

Υπάρχουν αρκετά εργαλεία γονιδιακής συνέκφρασης για το *Arabidopsis thaliana*, όπως: ATTED-II (Obayashi et al., 2018), EXPath (Chien et al., 2015), Genemania (Zuberi et al., 2013), Genevestigator (Hruz et al., 2008), SeedNet (Bassel et al., 2011), FlowerNet (Pearce et al., 2015) και GEM2Net (Zaag et al., 2015).

Χρησιμοποιήθηκαν νέες προσεγγίσεις που διαχωρίζουν το ACT από άλλα εργαλεία συνέκφρασης για το *Arabidopsis thaliana*. Η παλαιά έκδοση του ACT, ήταν βασισμένη αρχικά σε 322 δείγματα (Jen et al., 2006) τα οποία στην πορεία αυξήθηκαν σε περίπου 1400. Η έκδοση που περιγράφεται στην παρούσα εργασία, βασίζεται σε 3500 δείγματα, τα οποία επιλέχθηκαν από 19887 αρχικά δείγματα, μετά από εκτενή έλεγχο ποιότητας και διαλογή των πιο αντιπροσωπευτικών δειγμάτων για κάθε διαθέσιμο ιστό. Το ACT, βασίζεται σε δεδομένα από μια συγκεκριμένη πλατφόρμα μικροσυστοιχιών, ενώ άλλα εργαλεία όπως το ATTED-II και το EXPath, χρησιμοποιούν δεδομένα τόσο από μικροσυστοιχίες όσο και από RNA-Seq, που όμως παρουσιάζουν μεγάλες διαφορές στον υπολογισμό της συσχέτισης μεταξύ των δύο τύπων δεδομένων. Επιπλέον, το παλιό ACT χρησιμοποίησε τον MAS5.0 για επεξεργασία και κανονικοποιήση των δεδομένων. Όμως, λίγο καιρό μετά τη δημιουργία του παλιού ACT, η ίδια η Affymetrix ανακοίνωσε ότι ο MAS5.0 προτείνεται να χρησιμοποιείται κυρίως για να παραχθεί μία αρχική αναφορά σχετικά με την ποιότητα των μικροσυστοιχιών και για να ταυτοποιηθούν εμφανή προβλήματα, παρά ως κύρια μέθοδο κανονικοποίησης (Affymetrix, 2018; Dziuda, 2010). Ως εναλλακτικές μεθόδους, πρότειναν τους αλγορίθμους PLIER ή RMA. Πράγματι, τα περισσότερα εργαλεία συνέκφρασης χρησιμοποιούν τον αλγόριθμο RMA ο οποίος υποθέτει ότι η κατανομή των τιμών έντασης των ανιχνευτών είναι κοινή μεταξύ όλων των δειγμάτων (κανονικοποίηση ποσοστιμορίων). Αυτή η υπόθεση καθιστά τους αλγορίθμους κανονικοποίησης πολλαπλών μικροσυστοιχιών ακατάλληλους για μία ανάλυση συνέκφρασης, καθώς σε αυτήν την περίπτωση συγκεντρώνουμε δείγματα από διαφορετικούς ιστούς ή ερευνητικές ομάδες. Η παραπάνω παρατήρηση μπορεί να εξηγεί γιατί οι αλγόριθμοι κανονικοποίησης πολλαπλών μικροσυστοιχιών εισάγουν μεγάλο αριθμό ψευδών συσχετίσεων και γιατί δεδομένα κανονικοποιημένα από

αλγορίθμους κανονικοποίησης ξεχωριστών μικροσυστοιχιών, όπως ο MAS5.0, προσφέρουν καλύτερη βάση για δημιουργία δικτύων αλληλεπίδρασης μεταξύ πρωτεϊνών (Lim et al., 2007). Στην περίπτωσή μας, χρησιμοποιήθηκε ο καινοτόμος αλγόριθμος SCAN ο οποίος αναφέραμε ότι κανονικοποιεί κάθε δείγμα ξεχωριστά από τα υπόλοιπα.

Όσον αφορά τα αρχεία CDF, το παλιό ACT όπως και αρκετά σύγχρονα εργαλεία συνέκφρασης, χρησιμοποίησαν το προτεινόμενο από την Affymetrix CDF, ώστε να αντιστοιχίσουν τα σύνολα ανιχνευτών του ATH1 με τα γονίδια. Το default CDF δημιουργήθηκε το 2002 και περιέχει 22810 σύνολα ανιχνευτών, εκ των οποίων 64 είναι σύνολα ελέγχου. Από τα υπόλοιπα 22746 σύνολα, το 5.47% δεν αντιστοιχεί σε κάποιο γονίδιο και το 3.82% αντιστοιχεί σε περισσότερα τους ενός γονιδίου. Επιπλέον, ο συνολικός αριθμός γονιδίων που χαρτογραφούνται από το default CDF είναι 22168, τα 118 από τα οποία είναι παρωχημένα. Το κάθε εργαλείο χρησιμοποιεί έναν διαφορετικό τρόπο για να εξασφαλίσει την ένα-προς-ένα συσχέτιση μεταξύ γονιδίων και συνόλων ανιχνευτών. Το ATTED-II επέλεξε ένα μοναδικό σύνολο ανιχνευτών από το default CDF για κάθε γονίδιο και το EXPath δεν συμπεριέλαβε καμία διφορούμενη χαρτογράφηση. Αντιθέτως, το σύγχρονο ACT χρησιμοποιεί το επικαιροποιημένο CDF από την Brainarray, το οποίο σε αντίθεση με το default CDF της Affymetrix, εξασφαλίζει ότι κάθε σύνολο ανιχνευτών αντιστοιχεί σε ένα ακριβώς γονίδιο και το αντίθετο, φτάνοντας σε ένα τελικό σύνολο 21287 μη-παρωχημένων γονιδίων. Επιπλέον, το Brainarray CDF ενημερώνεται ετησίως, ορίζοντας τα σύνολα ανιχνευτών ανάλογα με τις τρέχουσες γονιδιακές και μεταγραφωμικές γνώσεις.

Για τον υπολογισμό των ανά ζεύγη αποστάσεων μεταξύ δειγμάτων ή γονιδίων, χρησιμοποιήθηκε ο τύπος μετατροπής των συντελεστών συσχέτισης Pearson σε απόσταση, $d = 1 - r$. Για την κατασκευή του δέντρου, χρησιμοποιήθηκε ο αλγόριθμος UPGMA. Η αναπαράσταση των συνεκφρασμένων δικτύων γονιδίων σε δέντρο, λαμβάνει υπ' όψη όλες τις βιολογικές πληροφορίες διαθέσιμες μέσω του πίνακα αποστάσεων και δεν βασίζεται σε ήδη αποφασισμένες αυθαίρετες ουδούς για τα r-value ή τα p-value, όπως γίνεται στην περίπτωση κατασκευής δικτύων γονιδίων με γράφους. Το ACT δίνει τη δυνατότητα στον χρήστη να βρει την βέλτιστη λίστα

συνεκφρασμένων γονιδίων, μέσω της αυξομείωσης του μεγέθους των φύλλων του δέντρου και στη συνέχεια βλέποντας και κρίνοντας από την τοπολογία των γονιδίων στο δέντρο και τις τιμές p-value των εμπλουτισμένων όρων.

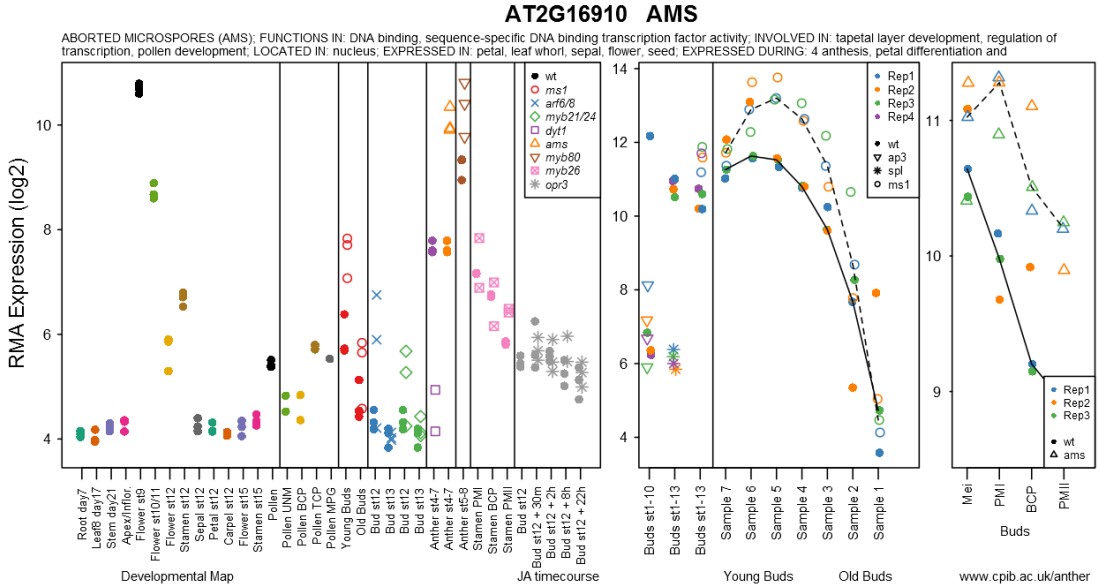
Το ACT διαθέτει μία απλή και συμπαγή σχεδίαση, λαμβάνοντας υπ' όψη τις ανάγκες των μοριακών βιολόγων που είναι και οι κύριοι χρήστες του εργαλείου. Οι έξοδοι που παράγονται είναι σε μία κατανοητή μορφή με βιολογικό ενδιαφέρον και αποφεύγεται ο καταιγισμός πληροφοριών που χαρακτηρίζει πολλά ανταγωνιστικά εργαλεία. Οι τιμές υπερεκπροσώπησης παρουσιάζονται με μία κοινώς κατανοητή μορφοποίηση και FDR adj p-values > 0.05 παραλείπονται ώστε να αποφευχθεί η παρουσίαση μη-στατιστικώς σημαντικών όρων. Οι πίνακες εμπλουτισμού όρων παράγονται με ευκολία, προσφέρουν μεγάλη ποικιλία όρων, καθώς και συνδέσμους στην αντίστοιχη εξωτερική βάση από την οποία προήλθαν.

Τα περισσότερα εργαλεία πραγματοποιούν ανάλυση εμπλουτισμού Οντολογίας Γονιδίων. Μερικά εργαλεία προσφέρουν επιπλέον επιλογές, όπως ανάλυση βιολογικών μονοπατιών (EXPath, ATTED-II) ή εύρεση μεταγραφικών στοιχείων (ATTED-II). Όμως, ενώ το ATTED-II ανακαλύπτει αυτά τα ρυθμιστικά στοιχεία χωρίς να βρίσκει τους μεταγραφικούς παράγοντες που προσδένονται πάνω σε αυτά τα μοτίβα, το ACT χρησιμοποιεί πειραματικά επαληθευμένα δεδομένα μεταγραφικών παραγόντων και των γονιδίων στόχων τους, μία κρίσιμη πληροφορία που αποκαλύπτει τους πιθανούς μεταγραφικούς παράγοντες που καθοδηγούν την συρρύθμιση γονιδίων.

Τα εργαλεία SeedNet και GEM2Net διαφέρουν στην προσέγγισή τους από το ACT, καθώς κατηγοριοποιούν τα γονίδια με βάση την έκφρασή τους, τόσο θετική όσο και αρνητική, κατά την βλάστηση του σπέρματος ή κάτω από βιοτικό/αβιοτικό στρες, αντίστοιχα. Ενώ το SeedNet προσφέρει μία εκτενή απεικόνιση ενός δικτύου συνέκφρασης, λαμβάνοντας υπ' όψη τόσο τα συνεκφρασμένα όσο και τα αρνητικώς συσχετισμένα γονίδια, δεν προσφέρει καμία επιλογή ανάλυσης εμπλουτισμού. Επιπλέον, υπάρχουν λάθη στην αντιστοίχιση Συμβόλου γονιδίου και κωδικού AGI. Το GEM2Net είναι συγκεκριμένο στην ανάλυσή του, προσφέροντας πολλές διαφορετικές υποκατηγορίες καταστάσεων βιοτικού και αβιοτικού στρες, όπου κάθε μία

βασίζεται σε περιορισμένο αριθμό δειγμάτων που τις περισσότερες φορές δεν ξεπερνάει τα 100, περιορίζοντας την συνέπεια των αναλύσεων. Τέλος, τα αποτελέσματα των αναλύσεων εμπλουτισμού παρουσιάζονται με μία περίπλοκη μορφοποίηση για τον χρήστη και με ειδικευμένους όρους που αναφέρονται σε διάφορες κατηγορίες στρες, αντί να υπάρχει ένας κοινός γενικός πίνακας αποτελεσμάτων.

Αντιθέτως, το ACT μελετάει το προφίλ της γονιδιακής συνέκφρασης στο σύνολο των ιστών. Παρόλα αυτά, μία ανάλυση ACT, χρησιμοποιώντας ένα ιστοειδικό γονίδιο ως οδηγό, το AMS που εκφράζεται στους ανθήρες, παρήγαγε συγκρίσιμα αποτελέσματα με μία αντίστοιχη ανάλυση από το εργαλείο FlowerNet (Εικόνα 50, Πίνακας 20), ένα εργαλείο συνέκφρασης βασισμένο σε δείγματα από άνθη του *Arabidopsis*, γεγονός που δείχνει ότι το ACT είναι εξίσου καλό στο να εντοπίζει όχι μόνο γονίδια με γενικευμένη έκφραση σε όλους τους ιστούς, αλλά και γονίδια που παρουσιάζουν ιστοειδική έκφραση.



Εικόνα 50 – Αποτελέσματα του FlowerNet για το AMS, το οποίο κατετάχθη στο cluster 2061.

FlowerNet GO Analysis			
GO:ID	Term	Annotated	p-value
GO:0048658	anther wall tapetum development	6	0.00063
GO:0019722	calcium-mediated signaling	53	0.00556
GO:0019932	second-messenger-mediated signaling	69	0.00724
GO:0048653	anther development	77	0.00808
GO:0007267	cell-cell signaling	89	0.00933

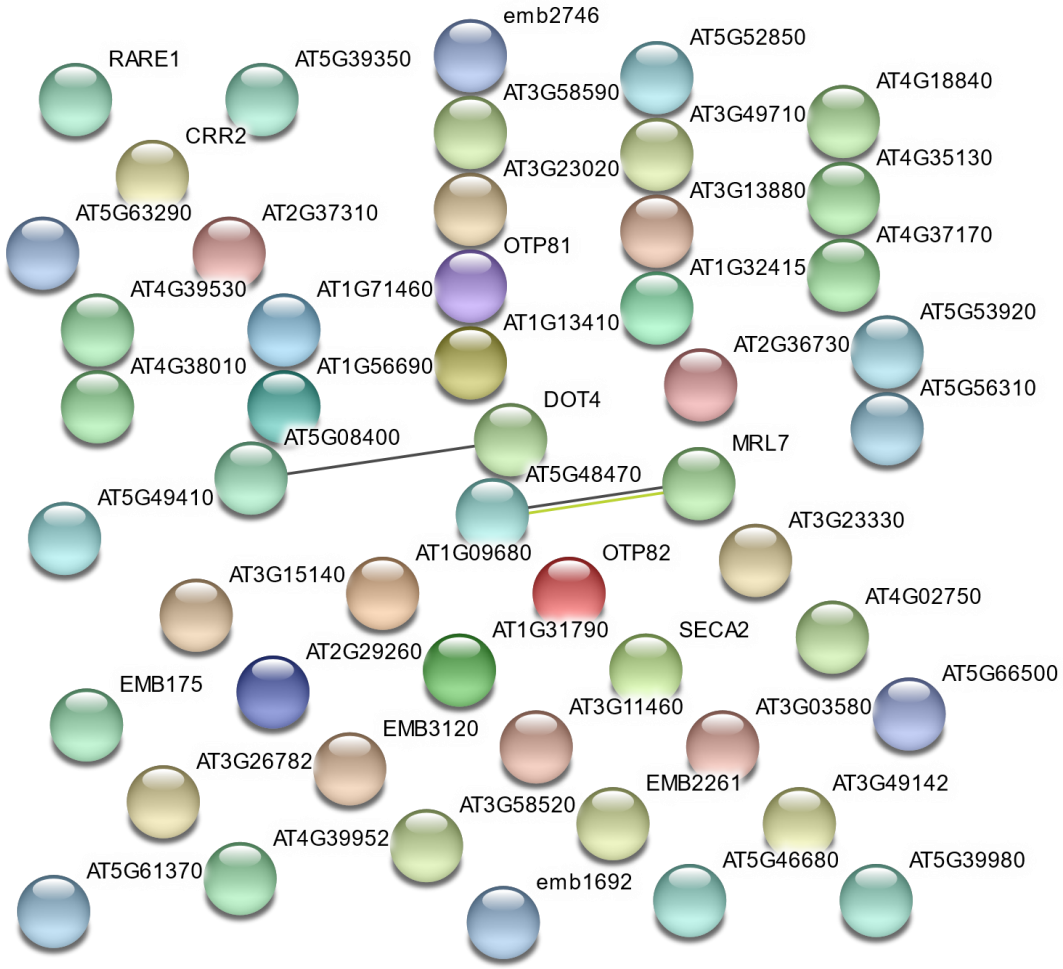
Πίνακας 20 - Ανάλυση εμπλουτισμού όρων Οντολογίας Γονιδίων του FlowerNet για το AMS (cluster 2061)

Σύγκριση αποτελεσμάτων ανταγωνιστικών εργαλείων με το ACT

Το γονίδιο emb1692 που έχει σχέση με την ανάπτυξη του εμβρύου κάτι που επιβεβαιώθηκε και από την ανάλυση του ACT, χρησιμοποιήθηκε ως είσοδος σε άλλα εργαλεία συνέκφρασης για το *Arabidopsis*. Στην αρχική λίστα των 50 συνεκφραζόμενων γονιδίων που παράγει το ATTED-II, περιέχονται και κάποια γονίδια που, από την περιγραφή τους, επίσης ανήκουν στην οικογένεια PPR, όπως είχε εντοπίσει και το ACT. Πέρα από το emb1692, στη λίστα του ATTED-II υπήρχαν 4 ακόμη embryo defective γονίδια σε μακρινή θέση από το γονίδιο-οδηγό. Το ACT περιείχε 5 embryo defective γονίδια και μερικά από αυτά βρίσκονται σε κοντινούς κλάδους με το emb1692. Η λίστα των 50 γονιδίων του ATTED-II και εκείνη του ACT, χρησιμοποιήθηκαν ως είσοδος σε αναλύσεις πρωτεϊνικών αλληλεπιδράσεων του STRING. Αν και στις δύο περιπτώσεις έχουμε αρκετά γονίδια, τα οποία δεν παρουσιάζουν κάποια σχέση, το δίκτυο του ACT (Εικόνα 51) έχει περισσότερες συσχετίσεις από εκείνο του ATTED-II (Εικόνα 52), γεγονός που σημαίνει ότι περισσότερα γονίδια στη λίστα του ACT συνδέονται λειτουργικά. Εδώ διαφαίνεται γιατί η χρήση του φυλογενετικού δέντρου που λαμβάνει υπ' όψη όλες τις συσχετίσεις των γονιδίων μεταξύ τους και όχι η χρήση της λίστας που λαμβάνει υπ' όψη τις συσχετίσεις ενός γονιδίου-οδηγού με τα υπόλοιπα, μπορεί τελικά να συνθέσει καλύτερα γονιδιακά δίκτυα. Επιπλέον, εκτελώντας περεταίρω ανάλυση εμπλουτισμού για τις δύο λίστες μέσω του g:Profiler, βρέθηκε και στις δύο κορυφαίος εμπλουτισμένος ο όρος «τροποποίηση RNA» (RNA modification), με τη λίστα του ATTED-II να έχει πολύ χαμηλό p-value (~10^{-40}) σε σχέση με το ACT (~10^{-10}), όσον αφορά την ανάλυση Βιολογικής Διεργασίας. Αυτή η παρατήρηση συμβαδίζει με την σύγκριση των δύο διαφορετικών λειτουργιών του αρχικού εργαλείου HGCA1.0, στο οποίο προέκυψε το συμπέρασμα ότι η λειτουργία κατασκευής δέντρου συνέκφρασης αντιπροσωπεύει καλύτερα γονίδια με μεγάλα r-values (>0.6), ενώ η λειτουργία παραγωγής λίστας συνεκφρασμένων γονιδίων προσφέρει καλύτερα απότελέσματα για γονίδια που έχουν χαμηλότερα r-values (https://www.michalopoulos.net/hgca1.0/FAQ.php). Η λίστα του ACT ανέδειξε στην κατηγορία Κυτταρικού Συστατικού τον χλωροπλάστη ως κορυφαίο οργανίδιο (p-value ~10^{-19}), κάτι που στο ATTED-II είχε p-value της τάξης του 10^{-3}, και στην Μοριακή Λειτουργία έδειξε τον όρο RNA methyltransferase

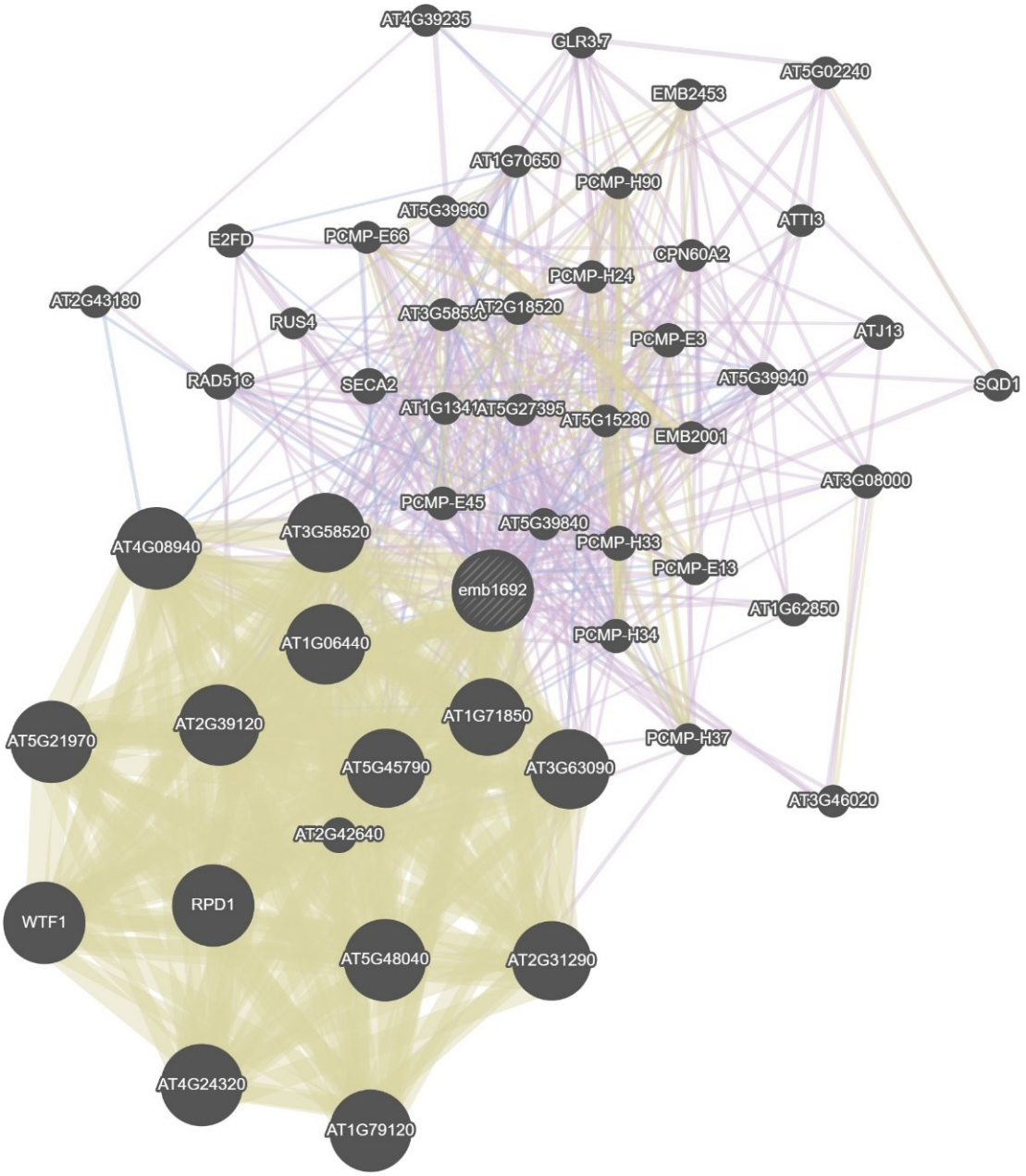
activity που ταιριάζει με τους γνωστούς ρόλους του γονιδίου, ενώ η λίστα του ATTED-II έδειξε κυρίως όρους σχετικούς με πρόσδεση σε ιόντα βαρέων μετάλλων.

Εικόνα 51 – Δίκτυο STRING με βάση τα συνεκφραζόμενα γονίδια με το emb1692 του ACT

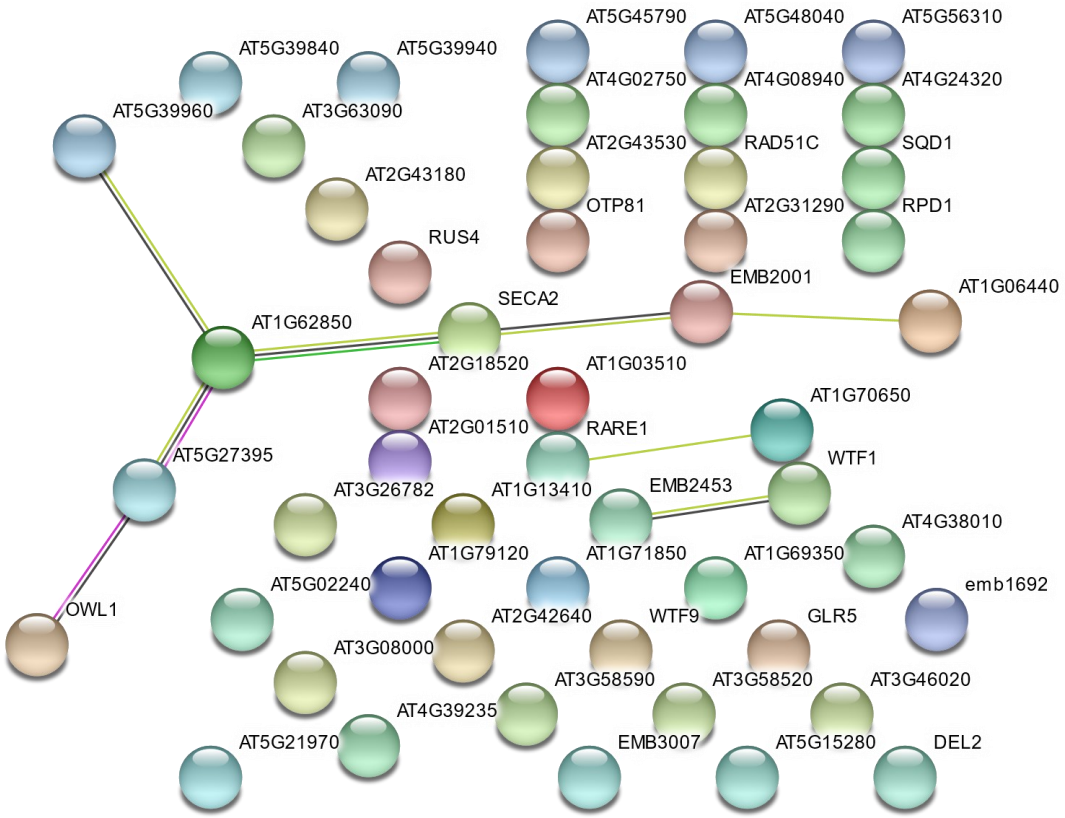


Εικόνα 52 – Δίκτυο STRING με βάση τα συνεκφραζόμενα γονίδια με το emb1692 του ATTED-II

Εισάγοντας το γονίδιο emb1692 στο Genemania και διαλέγοντας την επιλογή για δημιουργία δικτύου με 50 γονίδια, τα γονίδια που προκύπτουν διαχωρίζονται σε δύο υποδίκτυα με το emb1692 να αποτελεί τον κεντρικό κόμβο σύνδεσης μεταξύ τους (Εικόνα 53). Το ένα δίκτυο βασίζεται κυρίως στην συνέκφραση ενώ το άλλο το οποίο είναι περισσότερο πυκνό, βασίζεται στις κοινές δομικές και λειτουργικές περιοχές των πρωτεϊνών τους. Χρησιμοποιώντας αυτήν την λίστα ως είσοδο στο STRING, παράγεται ένα δίκτυο που παρουσιάζει μία σχέση μεταξύ επτά γονιδίων (Εικόνα 54), αλλά δεν είναι τόσο πυκνό όσο το αντίστοιχο του ACT (Εικόνα 51). Η ανάλυση μέσω του g:Profiler, δεν ανακάλυψε κάποιον στατιστικώς σημαντικό εμπλουτισμένο όρο στη λίστα του Genemania, σε αντίθεση με την ανάλυση στην λίστα παραγόμενη από το ACT, όπως είδαμε παραπάνω.



Εικόνα 53 – Το δίκτυο των 50 κόμβων του Genemania με είσοδο το emb1692



Εικόνα 54 – Το δίκτυο STRING με τη λίστα του Genemania για το emb1692

Το εργαλείο SeedNet δεν περιείχε το γονίδιο emb1692, αλλά χρησιμοποιώντας ένα γειτονικό γονίδιο στο υποδέντρο του ACT, το OTP81 (AT2G29760) το οποίο ανήκει στην οικογένεια PPR της οποίας τα μέλη εκτελούν τροποποιήσεις του RNA στους χλωροπλάστες (Hammani et al., 2009), εμφανίστηκαν 20 γειτονικά γονίδια στο δίκτυο του SeedNet, όπου κανένα δεν ανήκε στην οικογένεια PPR. Η ανάλυση της συγκεκριμένης λίστας, μέσω g:Profiler, εμφάνισε στατιστικώς σημαντικούς όρους μόνο στην κατηγορία Κυτταρικού Συστατικού, οι οποίοι ήταν σχετικοί με χλωροπλάστες και ο κορυφαίος είχε p-value ~10^{-6}.

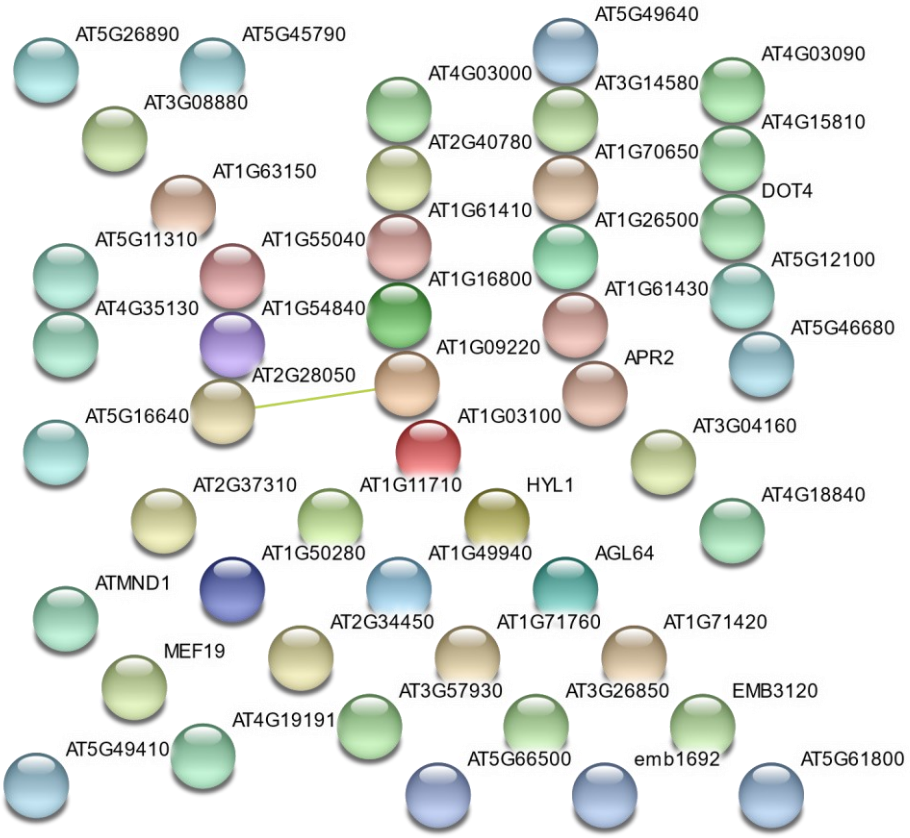
Το FlowerNet δεν κατέταξε το emb1692 σε κάποιο cluster, ενώ το GEM2Net κατέταξε το εν λόγω γονίδιο σε 4 διαφορετικές ομάδες, 3 αβιοτικές (Άζωτο, Αλάτι και Θερμοκρασία) και μία βιοτική (Βιοτροφικά Βακτήρια) (Πίνακας 21). Η ομάδα της θερμοκρασίας (cluster 32) αν και αποτελείται από ~700 γονίδια, παρουσίασε τις πιο χαμηλές τιμές p-value στην εσωτερική ανάλυση εμπλουτισμού του GEM2Net, με κορυφαίους εμπλουτισμένους όρους όπως «Αναπτυξιακή Διεργασία» και «μεταβολισμός RNA και DNA» οι οποίοι

αποτελούν πολύ πιο γενικευμένους όρους των εμπλουτισμένων όρων του ACT “embryo development” και “RNA modification”, αντίστοιχα.

Stress categories with MAP classification rule				
		Abiotic Stress		Biotic Stress
Gene Name	NITROGEN	SALT	TEMPERATURE	BIOTROPHIC BACTERIA
Total per Stress	1	1	1	1
AT5G62990	cluster_48	cluster_20	cluster_32	cluster_51

Πίνακας 21 – Οι κατηγορίες και οι ομάδες που κατέταξε το GEM2Net το γονίδιο emb1692

Τέλος το Genevestigator, σε μία ανάλυση θετικής συνέκφρασης για το emb1692, με την επιλογή παραγωγής λίστας με 50 συνεκφραζόμενα γονίδια, ενώ περιείχε γονίδια PPR, δεν ανακάλυψε κανένα embryo defective γονίδιο μέσα σε αυτή. Η ανάλυση STRING έδειξε ένα δίκτυο με ελάχιστη αλληλεπίδραση (Εικόνα 55). Η εσωτερική ανάλυση εμπλουτισμού όρων Οντολογίας Γονιδίων του Genevestigator, έδειξε το όρο «ριβόσωμα» ως κορυφαίο, χωρίς ιδιαίτερα χαμηλό p-value.



Εικόνα 55 - Το δίκτυο STRING με τη λίστα του Genevestigator για το emb1692

Η σύγκριση των διάφορων εργαλείων δείχνει ότι τα αποτελέσματα του ACT, τουλάχιστον όσον αφορά το γονίδιο emb1692 του οποίου μόλις πρόσφατα ο ρόλος άρχισε να διασαφηνίζεται (Daras et al., 2019), υπερτερούν σε σχέση με τα άλλα εργαλεία. Όσον αφορά το ATTED-II που αποτελεί βασικό ανταγωνιστή στον τομέα γονιδιακής συνέκφρασης στο *Arabidopsis thaliana*, τα αποτελέσματα συνέκφρασης είναι συγκρίσιμα με του ACT το οποίο παρουσιάζει μεγαλύτερη επιτυχία στην εύρεση λειτουργικών συνεργατών μέσω του παραγόμενου υποδέντρου συνέκφρασης, λόγω του ότι παράγει δέντρα αντί για λίστες.

HGCA2

Το εργαλείο HGCA2 διαθέτει όλα τα θετικά στοιχεία του σχεδιασμού και των δυνατοτήτων ανάλυσης του ACT, με τη μόνη διαφορά ότι η ανάλυση συνέκφρασης πραγματοποιήθηκε χρησιμοποιώντας δεδομένα RNA-Seq. Τα δείγματα προήλθαν από τη βάση δεδομένων GTEx, από πληθώρα διαφορετικών, υγειών ιστών του ανθρώπινου σώματος. Αυτό οδήγησε στην αποτελεσματική ομαδοποίηση των δειγμάτων και στην βέλτιστη επιλογή των 3500 πιο αντιπροσωπευτικών που χρησιμοποιήθηκαν στην ανάλυση, όπως φαίνεται στην Εικόνα 12. Επιπλέον, η χρήση δεδομένων από αλληλούχηση νέας γενιάς, προσφέρει μία επιστημονική εξέλιξη πάνω στην παλαιότερη τεχνολογία των μικροσυστοιχιών και αποτελεί μία γερή βάση για περισσότερες μελλοντικές βελτιστοποιήσεις, όσο θα βελτιώνονται οι υπάρχοντες μέθοδοι και θα δημιουργούνται καινούργιοι αλγόριθμοι επεξεργασίας και κανονικοποίησης των δεδομένων RNA-seq. Το γεγονός ότι υπάρχουν ~55000 γονίδια διαθέσιμα, επιτρέπει στο HGCA2 να αξιοποιήσει στο μέγιστο την ικανότητα για χαρακτηρισμό γονιδίων με άγνωστη λειτουργία, τα οποία ανήκουν σε κλάδους με ήδη χαρακτηρισμένα γονίδια. Τέλος, η ύπαρξη έκδοσης HGCA τόσο με δεδομένα RNA-Seq όσο και με δεδομένα μικροσυστοιχιών, επιτρέπει την ταυτόχρονη σύγκριση των αποτελεσμάτων κάθε έκδοσης του εργαλείου και πιθανώς αποτελεί ένα μέτρο αξιολόγησης των μεθόδων επεξεργασίας των RNA-seq έναντι των αντίστοιχων μεθόδων των μικροσυστοιχιών.

Είδαμε ότι χρησιμοποιώντας το γονίδιο NFKB2 ως είσοδο στο HGCA2, το υποδέντρο των συνεκφρασμένων γονιδίων χαρακτηρίστηκε με όρους

σχετικούς με ανοσολογική απόκριση, γεγονός που συμπίπτει με την λειτουργία του γονιδίου. Το ίδιο γονίδιο χρησιμοποιήθηκε ως είσοδος, τόσο στο HGCA1.5 που περιεγράφηκε στην παρούσα εργασία, όσο και με το παλαιό HGCA1 που δημιουργήθηκε το 2012 (Michalopoulos et al., 2012). Τα δύο εργαλεία είναι βασισμένα στα ίδια ακριβώς δείγματα μικροσυστοιχίων. Το HGCA1 προσφέρει δυνατότητα παραγωγής λίστας συνεκφρασμένων γονιδίων και δυνατότητα παραγωγής υποδέντρου συνέκφρασης. Το HGCA1.5 έχει κανονικοποιήσει τα δείγματα με αλγόριθμο SCAN, χρησιμοποίησε το CDF της Brainarray και ομαδοποίησε τα γονίδια με βάση τον UPGMA. Το αρχικό HGCA, κανονικοποίησε τα δείγματα με MAS5.0, χρησιμοποίησε το default CDF και έκανε ομαδοποίηση με Neighbour-Join. Οι διαφορές είναι εμφανείς, καθώς το HGCA1 ζητάει από τον χρήστη να επιλέξει το σύνολο ανιχνευτών του γονιδίου που θέλει να μελετήσει και στο παραγόμενο υποδέντρο ή λίστα συνέκφρασης, το ίδιο γονίδιο εμφανίζεται πολλές φορές, λόγω της πολλά προς πολλά σχέσης μεταξύ συνόλου ανιχνευτών και γονιδίων. Χρησιμοποιώντας τη λειτουργία δημιουργίας δέντρου, με οδηγό το σύνολο ανιχνευτών 209636_at, το HGCA1 ομαδοποίησε το NFKB2 στον ίδιο κλάδο με γονίδια της ίδιας οικογένειας (NKFB1, RELB, REL). Η ανάλυση υπερεκπροσώπησης Βιολογικής Διεργασίας έδειξε ως κορυφαίους, όρους σχετικούς με απόπτωση και κυτταρικό θάνατο, στην τάξη των 10^{-5}. Υπάρχουν πολλοί διαφορετικοί όροι υπερεκπροσωπημένοι πέρα από τους κορυφαίους. Όροι σχετικοί με ανοσολογική απόκριση υπάρχουν μέσα σε αυτούς με p-value από 10^{-4} – 10^{-2}. Η ανάλυση Μοριακής Λειτουργίας έδειξε όρους σχετικούς με ενεργότητα μεταγραφικών παραγόντων της τάξης του 10^{-3} και η ανάλυση Κυτταρικού Συστατικού έδειξε όρους των συμπλεγμάτων NF-*κ*B, με p-value ~10^{-4}. Αντίθετα, η λειτουργία δημιουργίας λίστας, παρήγαγε μία λίστα 50 ξεχωριστών γονιδίων (70 συνόλων ανιχνευτών) σχετικά χαμηλής συνέκφρασης σε σχέση με το γονίδιο-οδηγό (r-value < 0.46). Κορυφαίο γονίδιο στη λίστα ήταν το RELB, αλλά το REL που υπήρχε στην λειτουργία δέντρου, δεν εμφανίστηκε μέσα σε αυτή (αλλά υπάρχει στην επεκταμένη λίστα των 100 ξεχωριστών γονιδίων). Οι αναλύσεις υπερεκπροσώπησης παρουσίασαν γενικά χαμηλότερα p-values στην λίστα από ότι στο δέντρο, με την ανάλυση Βιολογικής Διεργασίας να αναδεικνύει κορυφαίους όρους με ανοσολογική απόκριση με p-values ~10^{-8}. Αξίζει να αναφερθεί ότι το γονίδιο BCL3 που αναφέραμε ότι είναι ένας μη-τυπικός

αναστολέας του NF-κB, εμφανίστηκε τόσο στη λίστα όσο και στο δέντρο, αλλά σε σχετικά μακρινές θέσεις. Όσον αφορά το HGCA1.5, το παραγόμενο υποδέντρο αναπτύχθηκε έως τους 6 κόμβους με 49 γονίδια. Σε αυτό το υποδέντρο, ο NFKB2 βρίσκεται στον ίδιο κλάδο με τους RELB και NFKB1, ενώ ο BCL3 δεν εμφανίζεται κάπου μέσα σε αυτό. Παράλληλα στο ίδιο υποδέντρο υπάρχει ένα κλάδος με 7 γονίδια HLA. Η ανάλυση εμπλουτισμού Βιολογικής Διεργασίας, ανέδειξε όρους σχετικούς με το MHC I (p-value ~10^{-14}), κάτι που προκύπτει από την παρουσία των γονιδίων HLA στο δέντρο, ενώ όροι σχετικοί με ανοσολογική απόκριση εμφανίζονται επίσης, με p-value ~10^{-8}, κάτι που συμπίπτει με τους εμπλουτισμούς της λίστας του HGCA1. Αντίστοιχα και οι περισσότεροι εμπλουτισμένοι όροι στις υπόλοιπες κατηγορίες, περιγράφουν περισσότερο τα γονίδια HLA.

Σε γενικές γραμμές το εργαλείο HGCA2, παρουσίασε υψηλότερα p-values στους εμπλουτισμούς σε σχέση με τα αποτελέσματα του HGCA1.5 και τη λειτουργία λίστας του HGCA1, όμως οι υπερεκπροσωπημένοι όροι ήταν πιο σχετικοί με τις λειτουργίες του NFKB2, σε σχέση με το HGCA1.5. Τα στενά συνδεδεμένα γονίδια HLA που εμφανίστηκαν στο HGCA1.5 μπορεί να προκαλούν τις χαμηλές τιμές στους εμπλουτισμένους όρους οι οποίοι όμως περιγράφουν την εν λόγω οικογένεια. Η λειτουργία δέντρου του HGCA1, αν και έβαλε τα γονίδια της οικογένειας NFKB όλα μαζί, δεν παρουσίασε ικανοποιητικά εμπλουτισμένους όρους. Η λειτουργία λίστας του HGCA1 παρουσιάζει τους πιο σφαιρικούς εμπλουτισμούς, πράγμα που οφείλεται στο γεγονός ότι τα συνεκφρασμένα γονίδια με το NFKB2 έχουν σχετικά χαμηλά r-values. Μπορούμε να πούμε ότι στην περίπτωση του NFKB2, και τα τρία εργαλεία παρουσίασαν συγκρίσιμα αποτελέσματα, με τη λειτουργία δέντρου του HGCA1 να έχει τις χειρότερες υπερεκπροσωπήσεις όρων. Το HGCA2 έκανε καλύτερη ομαδοποίηση των γονιδίων καθώς δεν τοποθέτησε τόσο κοντά τα HLA με το NFKB2, ενώ παράλληλα εμφάνισε τον BCL3 σε γειτονικό κλάδο, κάτι που δεν έγινε στο HGCA1.5. Τέλος ως κορυφαίο συνεκφραζόμενο γονίδιο στο NFKB2 όλα εμφάνισαν το RELB.

Παρόμοια εργαλεία με το HGCA2

Υπάρχουν αρκετά εργαλεία γονιδιακής συνέκφρασης για τον άνθρωπο βασισμένα τόσο σε δεδομένα μικροσυστοιχιών όσο και RNA-seq, όπως: COXPRESdb (Okamura et al., 2015), Genemania, ARCHS^4 (Lachmann et al., 2018), GeneFriends (van Dam et al., 2015), MEM (Adler et al., 2009) και SEEK (Zhu et al., 2015).

Η ροή ARCHS^4 επεξεργάζεται δεδομένα RNA-seq από τις βάσεις GEO και SRA, ώστε να υποστηρίξει δευτερεύουσες αναλύσεις. Ο ιστότοπος του ARCHS^4 προσφέρει πολλούς διαφορετικούς τρόπους για πρόσβαση των επεξεργασμένων δεδομένων γονιδιακής έκφρασης. Γίνεται απεικόνιση των δειγμάτων τα οποία υπέστησαν επεξεργασία καθώς και όλων των γονιδίων του ανθρώπου και του ποντικού με βάση την ομοιότητα στην συνέκφρασή τους (Globe Visualisation). Μπορεί να γίνει εύρεση στα εν λόγω δείγματα με βάση τα μετα-δεδομένα τους. Επίσης, ο χρήστης μπορεί να επιλέγει ένα σύνολο γονιδίων, ώστε να εξαχθεί στο EnrichR για ανάλυση εμπλουτισμού βιολογικών όρων. Βέβαια, το ίδιο το ARCHS^4 προσφέρει πίνακες με εμπλουτισμένους όρους, όπως Μεταγραφικούς Παράγοντες, Οντολογίες Γονιδίων, Μεταβολικά Μονοπάτια, κλπ. Εισάγοντας ένα γονίδιο πραγματοποιείται ανάλυση γονιδιακής συνέκφρασης και παρουσιάζονται τα 100 πιο κοντινά συνεκφρασμένα γονίδια, χρησιμοποιώντας τον συντελεστή συσχέτισης Pearson. Αν και το ARCHS^4 δεν χρησιμοποιεί δεδομένα από την GTEx στη βάση δεδομένων του, υποστηρίζει ότι οι πίνακες συνέκφρασης που προέκυψαν με τα διάφορα δεδομένα από GEO/SRA παράγουν καλύτερα αποτελέσματα από αντίστοιχους πίνακες συνέκφρασης του GTEx.

Το GeneFriends πραγματοποιεί ανάλυση συνέκφρασης με είσοδο ένα ή περισσότερα γονίδια. Βασίζεται σε έναν χάρτη γονιδιακής συνέκφρασης που περιγράφει ποια γονίδια τείνουν είτε να εκφράζονται είτε να υποεκφράζονται ταυτόχρονα σε ένα μεγάλο όγκο δειγμάτων RNA-seq. Επειδή υπάρχει μεγάλη ποικιλία στις καταστάσεις των δειγμάτων, αυτός ο χάρτης αντικατοπτρίζει τα γονίδια που υφίστανται κοινή μεταγραφική ρύθμιση. Αυτή η πληροφορία μπορεί να χρησιμοποιηθεί, αξιοποιώντας την προσέγγιση ότι λειτουργικά όμοια γονίδια συνεκφράζονται. Έτσι, μπορούν να βρεθούν καινοτόμα υποψήφια γονίδια που

εμπλέκονται σε βιολογικές διεργασίες ή συμπλέγματα ασθενειών, όπου μία λίστα γονιδίων που συσχετίζονται με αυτήν τη διεργασία ή την ασθένεια είναι διαθέσιμη. Επίσης, μπορεί να χρησιμοποιηθεί για να αποδοθεί ένα ρόλος σε γονίδια με άγνωστη λειτουργία, τα οποία συνεκφράζονται με άλλα γνωστά. Ο ανθρώπινος χάρτης συνέκφρασης κατασκευάστηκε από 46475 δείγματα και του ποντικού από 34322. Επίσης, κατασκευάστηκε ένας χάρτης συνέκφρασης από δείγματα του TCGA (Hutter and Zenklusen, 2018) και ένας από δείγματα του GTEx. Το GeneFriends παρουσιάζει τα συνεκφραζόμενα γονίδια, τους συνεκφραζόμενους μεταγραφικούς παράγοντες, προσφέρει ένα αρχείο δικτύου για χρήση με το Cytoscape, καθώς και εμπλουτισμό όρων μέσω του David. Η λίστα με τα συνεκφραζόμενα γονίδια και τους μεταγραφικούς παράγοντες κατατάσσεται με βάση το Mutual Rank, αλλά υπάρχει και ο συντελεστής συσχέτισης Pearson. Τα 50 κορυφαία γονίδια παρουσιάζονται, αλλά η πλήρης λίστα είναι διαθέσιμη.

Το MEM αποτελεί ένα διαδικτυακό εργαλείο που επιτρέπει την ανάλυση έκφρασης γονιδίων σε πολλαπλά πειράματα και την απεικόνισή τους. Τα δεδομένα προέρχονται από δημόσια πειράματα κατατεθειμένα στην ArrayExpress. Τα δείγματα είναι από ποικίλους ιστούς, καταστάσεις και ασθένειες και είναι ομαδοποιημένα με βάση την πλατφόρμα τους. Με είσοδο ένα μοναδικό γονίδιο, το MEM κατατάσσει άλλα γονίδια με βάση την ομοιότητα στην έκφρασή τους σε κάθε ξεχωριστό σύνολο δεδομένων. Ένα μοναδικό κοινό σκορ σημαντικότητας δημιουργείται λαμβάνοντας υπ’ όψη τα υπόλοιπα ατομικά σκορ. Τα δεδομένα μικροσυστοιχιών κανονικοποιήθηκαν με RMA. Η λίστα των γονιδίων είναι διαθέσιμη κάτω από την εικόνα του Heatmap και τα αποτελέσματα μπορούν να ληφθούν με μορφοποίηση NetCDF (Rew and Davis, 1990). Τέλος, υπάρχει διαθέσιμη ανάλυση εμπλουτισμού μέσω του g:Profiler.

Το SEEK είναι μία μηχανή αναζήτησης συνεκφρασμένων γονιδίων. Η είσοδος μπορεί να είναι είτε ένα γονίδιο είτε μία λίστα γονιδίων. Η έξοδος είναι μία λίστα με τα 100 κορυφαία συνεκφραζόμενα γονίδια και το σκορ συσχέτισής τους και η ομοιότητά τους ανάμεσα στα διαφορετικά σύνολα δεδομένων απεικονίζεται μέσω ενός Heatmap. Τα γονίδια έχουν καταταχθεί με βάση ένα z-score που προκύπτει από τον συντελεστή συσχέτισης του γονιδίου,

βεβαρημένο ανάλογα με το πόσο σχετικό είναι κάθε διαφορετικό σύνολο δεδομένων (τιμές: -3 έως +3). Υπάρχει διαθέσιμη ανάλυση εμπλουτισμού όρων, με πολλές επιλογές. Τα δεδομένα προέρχονται τόσο από μικροσυστοιχίες από διάφορες πλατφόρμες (4918 δείγματα) όσο και από RNA-seq (278 δείγματα). Σε κάθε περίπτωση χρησιμοποιήθηκε ο κατάλληλος αλγόριθμος κανονικοποίησης. Οι κανονικοποιημένες τιμές έκφρασης είναι επίσης διαθέσιμες μέσω του SEEK.

Η βάση COXPRESdb αποτελεί την αντίστοιχη έκδοση της ATTED-II, αλλά για ζωικούς οργανισμούς. Δεδομένα Illumina RNA-seq, ελήφθησαν από την DDBJ Sequence Read Archive. Η COXPRESdb παρέχει πολλές δυνατότητες. Γίνεται ανάλυση συνέκφρασης με είσοδο ένα γονίδιο και παραγωγή της λίστας των συνκεφρασμένων γονιδίων και πάνω σε αυτά μπορεί να πραγματοποιηθεί ανάλυση εμπλουτισμού όρων Οντολογίας Γονιδίων και Βιολογικών Μονοπατιών KEGG. Επίσης, προσφέρονται πολλές εκδόσεις δεδομένων από διάφορες πλατφόρμες, τόσο RNA-seq όσο και μικροσυστοιχιών και επιπλέον, για κάθε γονίδιο αναφέρεται το ομόλογό του σε κάποιον άλλο οργανισμό (γονίδιο του ανθρώπου με εκείνο του ποντικού, κλπ). Επιπλέον, προσφέρεται δημιουργία δικτύου συνέκφρασης, καθώς και σχεδιασμός φυλογενετικού δέντρου, το οποίο όμως παράγεται από μία δεδομένη από το χρήστη λίστα γονιδίων και όχι αυτόματα με είσοδο ένα γονίδιο.

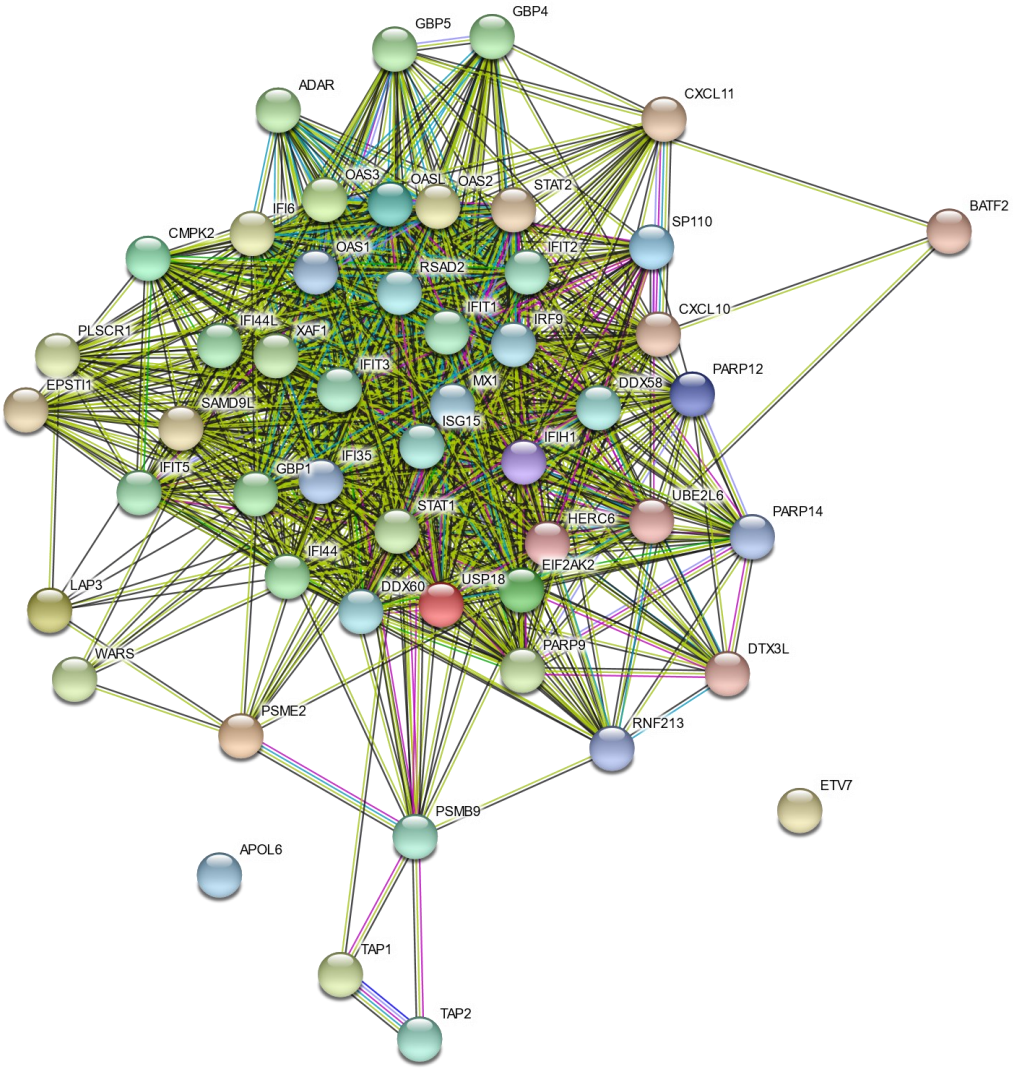
Το GeneMania μοιάζει περισσότερο με το STRING παρά με τα υπόλοιπα εργαλεία συνέκφρασης. Περιέχει πολλαπλά δεδομένα δικτύων από διάφορες πηγές. Εκατοντάδες σύνολα δεδομένων και δεδομένα αλληλεπιδράσεων συλλέχθηκαν από GEO, BioGRID, iRefIndex και I2D. Το εργαλεία δημιουργεί ένα δίκτυο που συνδέει παρόμοια γονίδια. Το δίκτυο μπορεί να κατασκευαστεί είτε από μία λίστα γονιδίων είτε εισάγοντας ένα γονίδιο και ανακαλύπτοντας τα κοινά με εκείνο γονίδια. Τα δεδομένα αλληλεπιδράσεων περιέχουν πρωτεϊνικές και γονιδιακές συσχετίσεις, βιολογικά μονοπάτια, συνέκφραση και πρωτεϊνικές δομικές και λειτουργικές περιοχές. Επίσης περιλαμβάνει πολλούς διαφορετικούς οργανισμούς, όπως άνθρωπο, ποντικό, δροσόφιλα, κλπ.

Σύγκριση αποτελεσμάτων ανταγωνιστικών εργαλείων με το HGCA2

Χρησιμοποιώντας το γονίδιο STAT1 (σύνολο ανιχνευτών 200887_s_at) ως είσοδο, έγινε σύγκριση των αποτελεσμάτων του HGCA1.5 και HGCA1 με το HGCA2. Το HGCA1.5 παρήγαγε ένα υποδέντρο συνέκφρασης των 18 γονιδίων, τα 16 από τα οποία ήταν στο υποδέντρο που παράχθηκε από το HGCA2. Παρομοίως, οι αναλύσεις εμπλουτισμού παρουσίασαν κοινούς όρους. Αντίστοιχα, το παραγόμενο δέντρο από το HGCA1, είχε κοινά γονίδια με τα δέντρα των άλλων δύο HGCA (οικογένεια OAS και γονίδια σχετιζόμενα με ιντερφερόνες). Οι αναλύσεις εμπλουτισμού του δέντρου του HGCA1, επίσης έδειξαν κορυφαίους όρους σχετικούς με ιούς και άμυνα εναντίων των ιών αν και με υψηλότερα p-values. Τέλος, η λειτουργία λίστας του HGCA1 έδειξε στενή σχέση με τα 10 πρώτα γονίδια (r-value > 0.6) και ανέδειξε κορυφαίους όρους σχετιζόμενους γενικά με την ανοσολογική απόκριση, ενώ οι όροι σχετικοί με άμυνα στους ιούς ήταν σε χαμηλότερη θέση.

Έγινε σύγκριση του HGCA2 με το εργαλείο COXPRESdb, χρησιμοποιώντας το γονίδιο STAT1 ως είσοδο. Το COXPRESdb εξάγει αρχικά μία λίστα με τα 50 πιο συνεκφρασμένα γονίδια με το STAT1, η οποία μπορεί να επεκταθεί μέχρι και τα 2000. Υπάρχουν αρκετές εμφανείς ομοιότητες με τα συνεκφραζόμενα γονίδια του HGCA2. Οι κύριοι όροι KEGG των γονιδίων στο COXPRESdb έδειξαν ιογενείς ασθένειες, κάποιες που εμφανίστηκαν και στα αποτελέσματα του HGCA2, όπως Υπατίτιδα C και Ιλαρά. Τέλος, έγινε σύγκριση ενός δικτύου STRING που κατασκευάστηκε με τα γονίδια του COXPRESdb (Εικόνα 56) και εκείνου με τα γονίδια του υποδέντρου συνέκφρασης του HGCA2 (Εικόνα 46). Τα αποτελέσματα, αν εξαιρέσουμε το διαφορετικό αριθμό των γονιδίων σε κάθε περίπτωση, έδειξαν μία ιδιαίτερα στενή σχέση και στις 2 περιπτώσεις, με το HGCA2 να έχει σχετικά περισσότερες αλληλεπιδράσεις μεταξύ των συνεκφρασμένων γονιδίων, καθώς στο δίκτυο του COXPRESSdb, πολλές περιφερειακές πρωτεΐνες διαχωρίζονται από την κλειστή κεντρική ομάδα: Η πυκνότητα γράφου (Coleman and Moré, 1983) υπολογίστηκε για το δίκτυο του HGCA2 ως 0.8546, ενώ στο COXPRESdb ως 0.5772. Πρέπει να αναφερθεί, ότι ενώ μόνο το COXPRESdb έδειξε το γονίδιο STAT2 να περιέχεται στη λίστα των συνεκφρασμένων γονιδίων, το HGCA2 ανακάλυψε τον ίδιο

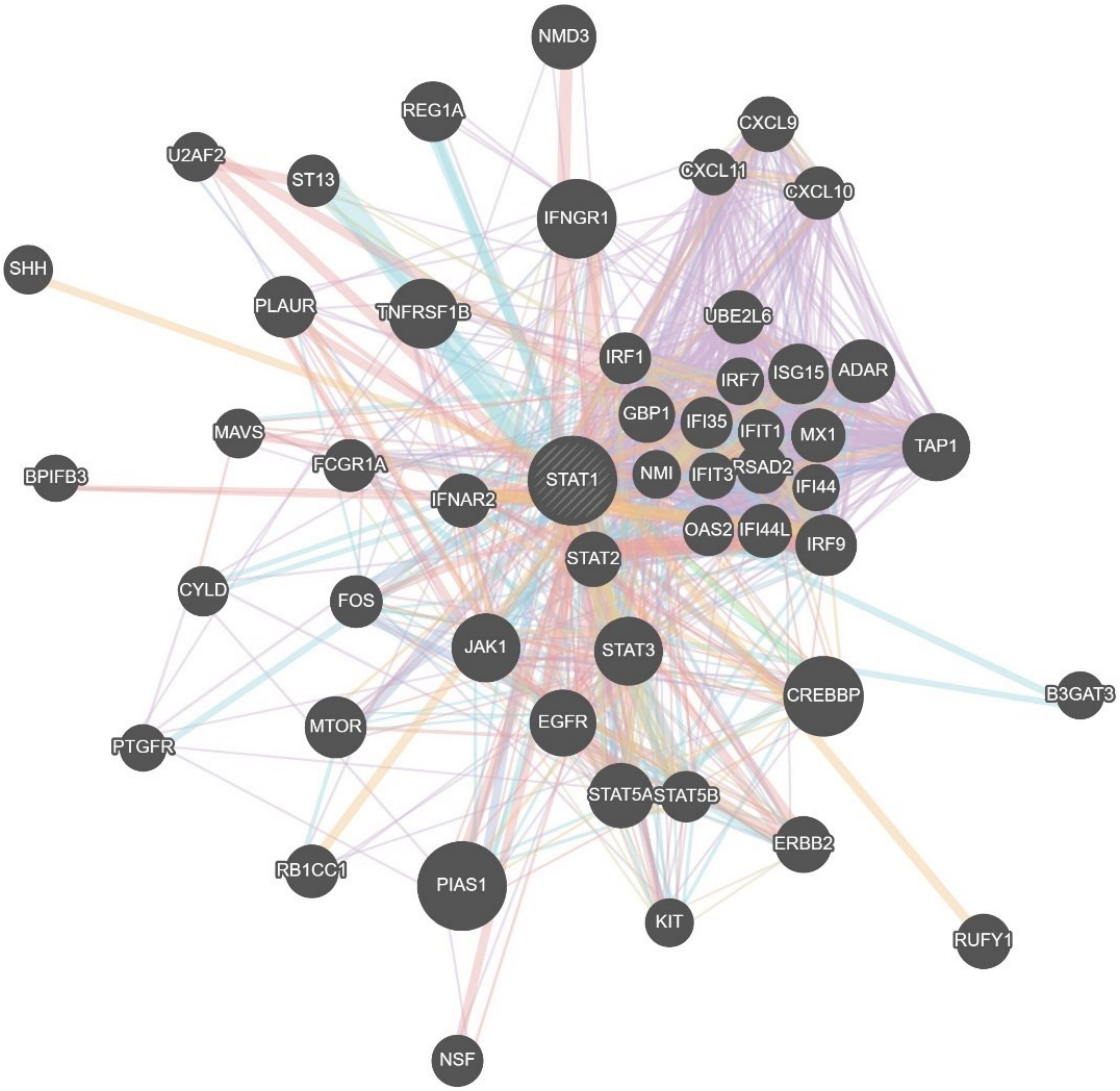
παράγοντα ως κορυφαίο υπερεκπροσωπημένο με πολύ μικρό p-value (~10^{-40}) στην αντίστοιχη ανάλυση εμπλουτισμού μεταγραφικών παραγόντων.



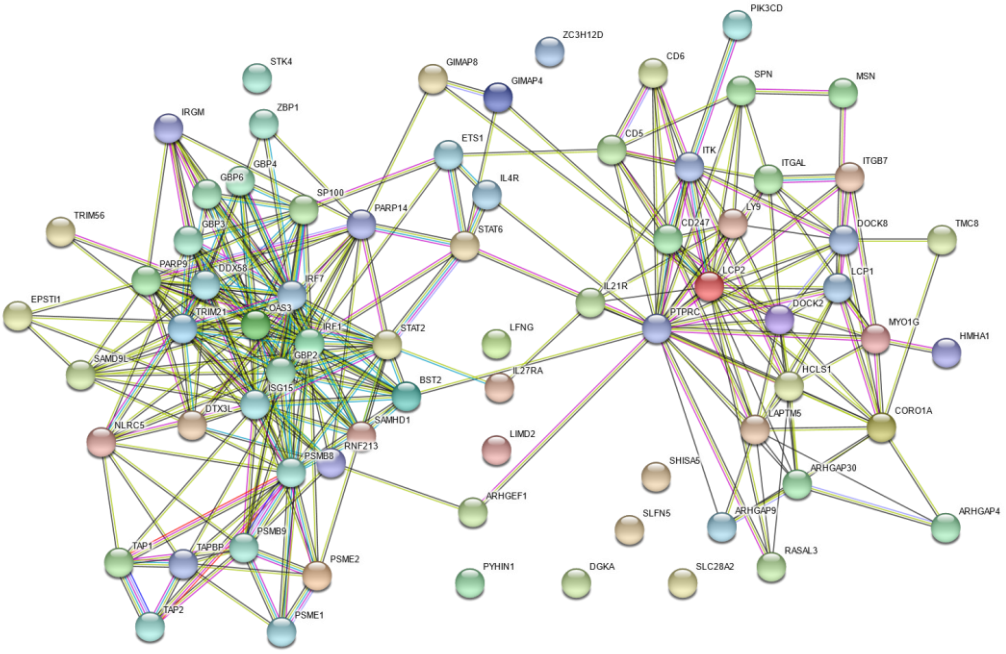
Εικόνα 56 – Το δίκτυο STRING με τα γονίδια συνέκφρασης του COXPRESdb, με είσοδο το STAT1

Όσον αφορά το GeneMania, έγινε χρήση της λειτουργίας δημιουργίας δικτύου με είσοδο ενός μοναδικού γονιδίου, στη συγκεκριμένη περίπτωση πάλι το STAT1. Το αποτέλεσμα, με επιλογή για δίκτυο αποτελούμενο από 50 γονίδια (μέγιστο μέγεθος τα 100), παρουσίασε περισσότερο συνεργάτες με φυσική αλληλεπίδραση και δευτερευόντως με βάση τη συνέκφραση (Εικόνα 57). Αυτό εξηγεί και τις διαφορές ανάμεσα στις λίστες των γονιδίων, όπως την ύπαρξη παραγόντων της ίδιας οικογένειας (STAT3, STAT5) στη λίστα του GeneMania, καθώς, ενώ ανήκουν στην ίδια οικογένεια γονιδίων, οι βιολογικές τους διεργασίες είναι διαφορετικές. Γενικώς, η χρήση του GeneMania είναι

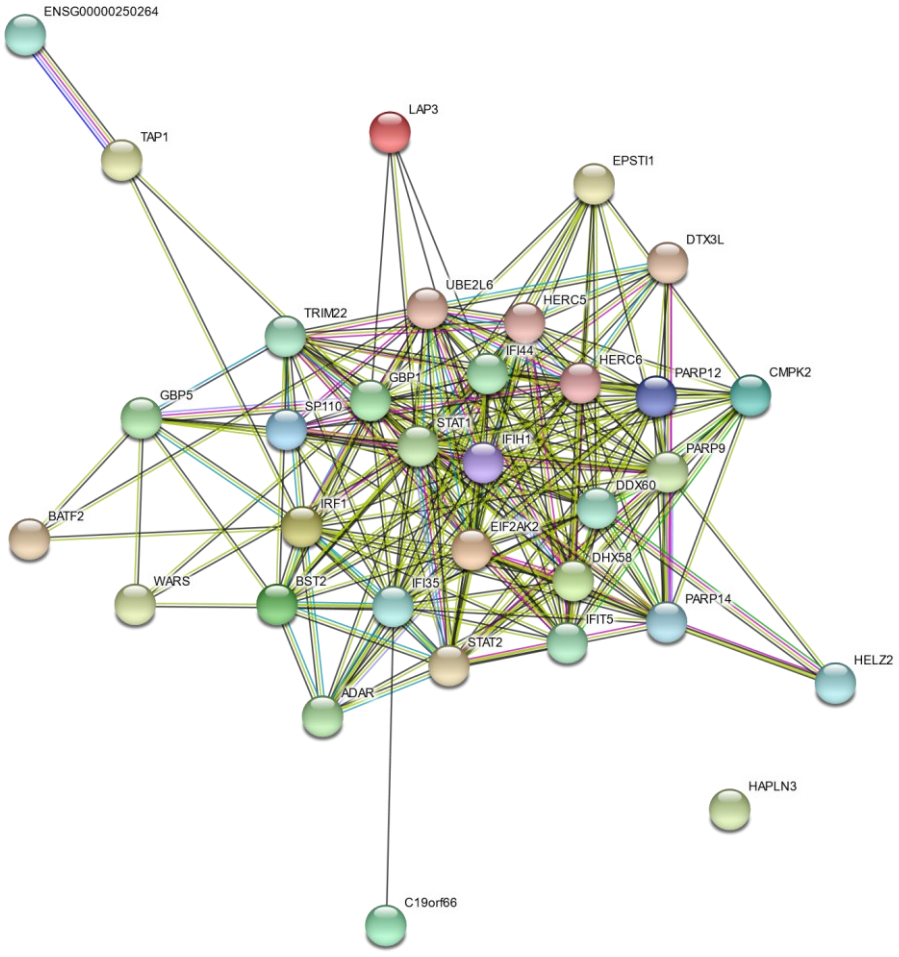
περισσότερο συγκρίσιμη με εκείνη του STRING, δηλαδή η δημιουργία δικτύων αλληλεπίδρασης με είσοδο μία λίστα γονιδίων.



Εικόνα 57 – Το δίκτυο του Genemania με είσοδο το STAT1



Εικόνα 58 – Το δίκτυο STRING από τα αποτελέσματα του ARCHS4 με είσοδο το STAT1

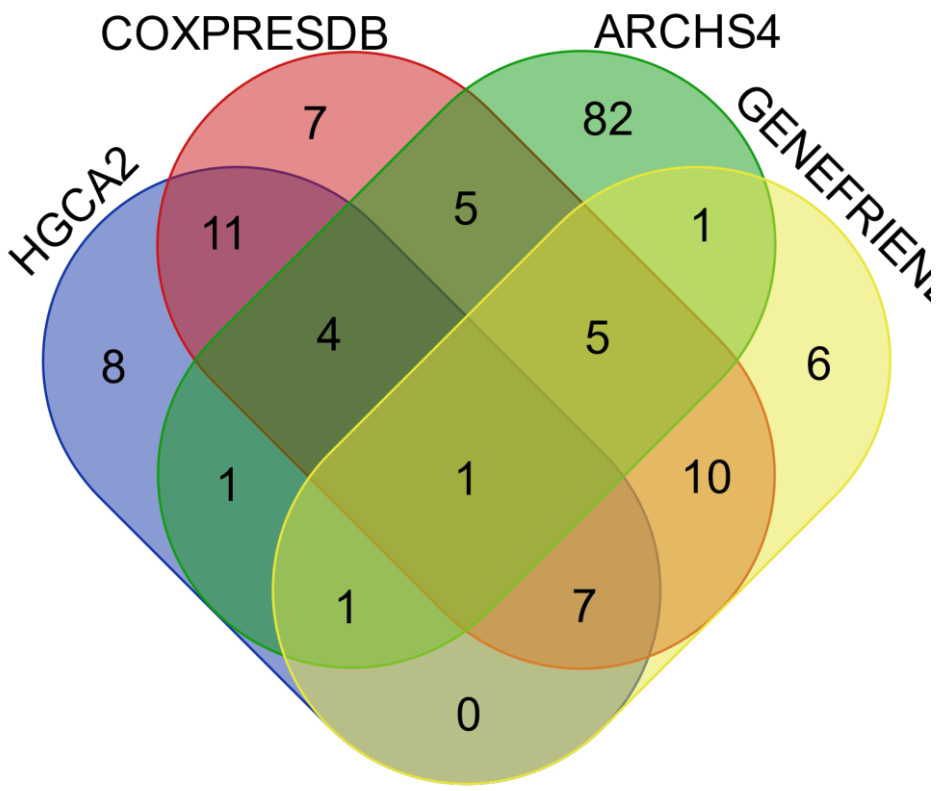


Εικόνα 59 – Το δίκτυο STRING με τα αποτελέσματα του GeneFriends με είσοδο το γονίδιο STAT1

Ανάλυση συνέκφρασης στο ARCHS[4] με είσοδο το STAT1, παρήγαγε αρκετά διαφορετική λίστα συνεκφρασμένων γονιδίων από αυτή του HGCA2. Η αυτόματη εξαγωγή στο EnrichR, έδειξε κυρίως εμπλουτισμό όρων που σχετίζονται με ιντερφερόνες τύπου Ι. Το ARCHS[4] παρουσίασε μία λίστα με τα 100 πιο συνεκφρασμένα γονίδια, η οποία χρησιμοποιήθηκε επίσης για ανάλυση στο STRING (Εικόνα 58). Σε αντίθεση με τα υπόλοιπα εργαλεία, το δίκτυο που προέκυψε δεν είναι ιδιαίτερα πυκνό. Στην συγκεκριμένη λίστα, φαίνεται ότι τα συνεκφραζόμενα γονίδια διαχωρίζονται σε δυο εντελώς ξεχωριστά λειτουργικά δίκτυα που συνδέονται μέσω ενδιάμεσων κόμβων.

Τέλος, το GeneFriends, παρουσιάζει ως έξοδο στην ιστοσελίδα τα 50 πιο συνεκφρασμένα γονίδια με το STAT1, αλλά είναι δυνατόν να γίνει λήψη της λίστας που συσχετίζει όλα τα διαθέσιμα γονίδια με βάση τη συνέκφρασή τους με το γονίδιο εισόδου. Χρησιμοποιώντας την λίστα των 50 γονιδίων ως είσοδο στο STRING , προκύπτει ένα δίκτυο το οποίο έχει αρκετά κοινά στοιχεία με εκείνα των HGCA2 και COXPRESdb, αλλά έχει ελάχιστα λιγότερο στενές συσχετίσεις μεταξύ των πρωτεϊνών, όπως φαίνεται από το παραγόμενο δίκτυο STRING (Εικόνα 59).

Συγκρίνοντας τις παραγόμενες λίστες συνεκφρασμένων γονιδίων από τα εργαλεία HGCA2, COXPRESdb, ARCHS[4] και GeneFriends (Εικόνα 60), παρατηρούμε ότι υπάρχουν αρκετές διακυμάνσεις μεταξύ των εργαλείων, με το ARCHS[4] να διαφέρει περισσότερο, όπως αναφέραμε, δεδομένου ότι είχε και τη λίστα με τα περισσότερα γονίδια (100). Μόλις ένα γονίδιο ήταν κοινό μεταξύ όλων των εργαλείων, το EPSTI1 του οποίου πιθανός ρόλος είναι η φωσφορυλίωση και ο πυρηνικός εντοπισμός των RELA και STAT1 κατά την ενεργοποίηση των μακροφάγων.



Εικόνα 60 - Διάγραμμα VENN με τις διαφορετικές λίστες συνεκφρασμένων γονιδίων από κάθε εργαλείο

Συγκρίνοντας τα αποτελέσματα, βλέπουμε ότι όπως και στην περίπτωση του ACT, το αδελφό εργαλείο του ATTED-II για τον άνθρωπο, το COXPRESdb, αποτελεί τον κύριο ανταγωνιστή του HGCA2, με το GeneFriends να παράγει αρκετά συγκρίσιμα αποτελέσματα και με τα άλλα δύο εργαλεία. Τα υπόλοιπα εργαλεία δεν παρήγαγαν καλύτερα αποτελέσματα από το HGCA2. Κυρίως μέσω της ανάλυσης του STRING, μπορούμε να πούμε ότι το HGCA2 μπορεί να παράγει καλύτερες λίστες με λειτουργικώς σχετιζόμενα συνεκφρασμένα γονίδια.

Συμπέρασμα

Το κύριο χαρακτηριστικό που ξεχωρίζει τα εργαλεία ACT και HGCA με τα άλλα ανταγωνιστικά είναι ότι δημιουργήθηκε εξαρχής έχοντας κατά νου τη χρήση από Μοριακούς Βιολόγους και όχι από Βιοπληροφορικούς, για αυτό και εξηγείται η λιτότητα του σχεδιασμού, η απλότητα λειτουργίας τους και κυρίως η παρουσίαση βιολογικά σχετικών αποτελεσμάτων. Τα ATTED-II και COEXPRESSdb χαρακτηρίζονται από πολύ περίπλοκη σχεδίαση: απαιτείται από τον χρήστη ιδιαίτερη παρατηρητικότητα για να βρει τον κατάλληλο

σύνδεσμο για να εμφανιστεί η λίστα συνεκφρασμένων γονιδίων και παράλληλα χρειάζονται πολλά κλικ μέχρι να εμφανιστεί το δίκτυο συνεκφρασμένων γονιδίων και οι όροι που το περιγράφουν. Στα εργαλεία μας, το ίδιο αποτέλεσμα γίνεται με μία μόνο κίνηση και όλα τα αποτελέσματα είναι στην ίδια σελίδα. Επιπλέον, τα περισσότερα εργαλεία που δημιουργούν λίστα συνεκφρασμένων γονιδίων, βασίζονται σε εξωτερικά εργαλεία εμπλουτισμού όρων, όπως g:Profiler και EnrichR, για πιο εξειδικευμένες κατηγορίες υπερεκπροσώπησης. Στην περίπτωσή μας, όλες οι κατηγορίες εμπλουτισμού που περιέχουν τα πιο καινούργια δεδομένα από τις καλύτερες βάσεις δεδομένων του κάθε είδους, είναι διαθέσιμες εσωτερικά στα εργαλεία. Το ACT περιέχει επιβεβαιωμένους πειραματικά στόχους μεταγραφικών παραγόντων από AtRegNet και Plant Cistrome Database και το HGCA περιέχει τους στόχους από τα Encode και ReMap2020, κάτι το οποίο τα ξεχωρίζει από τα ήδη υπάρχοντα εργαλεία συνέκφρασης, που είτε δεν έχουν καθόλου αυτήν την κατηγορία, είτε απλά ανακαλύπτουν κοινά ρυθμιστικά μοτίβα στα συνεκφραζόμενα γονίδια. Μόνο μέσω των εργαλείων μας μπορεί ο χρήστης να ανακαλύψει τους μεταγραφικούς παράγοντες υπεύθυνους για την συρρύθμιση γονιδίων.

Τα εργαλεία μας είναι τα μοναδικά που προσφέρουν την παραγωγή και παρουσίαση των συνεκφρασμένων γονιδίων με μορφή φυλογενετικού δέντρου. Τα περισσότερα εργαλεία προτιμούν την παραγωγή λίστας ή δικτύου συνέκφρασης. Το Genemania δημιουργεί δίκτυο γονιδίων, το οποίο όμως δεν βασίζεται αποκλειστικά στην συνέκφραση, αλλά κυρίως σε πειραματικά επαληθευμένες αλληλεπιδράσεις μεταξύ γονιδίων, κάτι που περιορίζει το στοιχείο της ανακάλυψης.

Η χρήση δειγμάτων από πολλά διαφορετικά εργαστήρια, κυρίως στην περίπτωση της τεχνολογίας RNA-seq που βρίσκεται ακόμα σε πρωταρχικό στάδιο, τείνει να εισάγει μεγάλο αριθμό από επιδράσεις συνόλων παραγωγής (batch-effects). Όπως είδαμε στην περίπτωση του ARCHS[4], που είχε λάβει δεδομένα τόσο από GEO όσο και από SRA, τα αποτελέσματά του ήταν κατώτερα από οποιοδήποτε παρόμοιο εργαλείο συγκρίθηκε. Το HGCA2 αντιμετώπισε αυτό το πρόβλημα μαζεύοντας δεδομένα GTEx που είναι από κοινή πηγή, κοινό εργαστήριο με συγκεκριμένες και βέλτιστες πειραματικές μεθόδους και έχουν συγκεκριμένα χαρακτηριστικά προέλευσης δειγμάτων. Αν

και στην περίπτωση των ACT και HGCA1.5 υπήρχε μεγάλη ποικιλία στην προέλευση δειγμάτων, οι αλγόριθμοι κανονικοποίησης μικροσυστοιχιών βρίσκονται σε αρκετά ώριμο στάδιο, όπως φαίνεται με τον SCAN που σε κάποιο βαθμό μεριμνά για την απαλοιφή των batch-effects, και η επιλογή δειγμάτων έγινε με πολύ λεπτομερή τρόπο, ώστε να αποφύγει όσο το δυνατόν περισσότερες προκαταλήψεις. Τέλος, τα HGCA1.5 και ACT, χρησιμοποιούν το συνεχώς ανανεωμένο CDF της Brainarray, την ύπαρξη του οποίου τα περισσότερα εργαλεία συνέκφρασης με βάση τις μικροσυστοιχίες ακόμα αγνοούν.

Τα εργαλεία ACT και HGCA προσφέρουν πολλές δυνατότητες, μελετώντας βασικούς οργανισμούς-μοντέλα, παρουσιάζοντας κατανοητά και επαληθεύσιμα αποτελέσματα και χρησιμοποιώντας σύγχρονα δεδομένα και τεχνολογίες. Πιστεύουμε ότι αποτελούν χρήσιμα εργαλεία στην κοινότητα της Μοριακής Βιολογίας και ότι μπορούν να εδραιωθούν περισσότερο με μελλοντικές τροποποιήσεις.

Βιβλιογραφία

Adler, P., Kolde, R., Kull, M., Tkachenko, A., Peterson, H., Reimand, J., and Vilo, J. (2009). Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. Genome Biol *10*, R139.

Affymetrix (2006). Affymetrix Power Tools.

Affymetrix (2018). Expression Console™ Software v1.4 USER GUIDE.

Archie, J., Day, H.E.W., Felsenstein, J., Maddison, W., Meacham, C., Rohlf, F.J., and Swofford, D. (2008). The Newick tree format.

Artus, N.N., Uemura, M., Steponkus, P.L., Gilmour, S.J., Lin, C., and Thomashow, M.F. (1996). Constitutive expression of the cold-regulated Arabidopsis thaliana COR15a gene affects both chloroplast and protoplast freezing tolerance. Proc Natl Acad Sci U S A *93*, 13404-13409.

Bagos, P. (2015). Βιοπληροφορική (ΣΥΝΔΕΣΜΟΣ ΕΛΛΗΝΙΚΩΝ ΑΚΑΔΗΜΑΪΚΩΝ ΒΙΒΛΙΟΘΗΚΩΝ).

Barakat, A., Szick-Miranda, K., Chang, I.F., Guyot, R., Blanc, G., Cooke, R., Delseny, M., and Bailey-Serres, J. (2001). The organization of cytoplasmic ribosomal protein genes in the Arabidopsis genome. Plant Physiol *127*, 398-415.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M.*, et al.* (2013). NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res *41*, D991-995.

Bassel, G.W., Lan, H., Glaab, E., Gibbs, D.J., Gerjets, T., Krasnogor, N., Bonner, A.J., Holdsworth, M.J., and Provart, N.J. (2011). Genome-wide network model capturing seed germination reveals coordinated regulation of plant cellular phase transitions. Proc Natl Acad Sci U S A *108*, 9709-9714.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Series B Stat Methodol *57*, 289-300.

Bolstad, B.M., Collin, F., Brettschneider, J., Simpson, K., Cope, L., Irizarry, R.A., and Speed, T.P. (2005). Quality Assessment of Affymetrix GeneChip Data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor, R. Gentleman, V.J. Carey, W. Huber, R.A. Irizarry, and S. Dudoit, eds. (Springer, New York, NY.).

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C.*, et al.* (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet *29*, 365-371.

Brettschneider, J., Collin, F., Bolstad, B.M., and Speed, T.P. (2008). Quality Assessment for Short Oligonucleotide Microarray Data. Technometrics *50*, 241-264.

Brinkman, F.S., and Leipe, D.D. (2001). Phylogenetic analysis. Methods Biochem Anal *43*, 323-358.

Broad Institute (2020). Picard Toolkit.

Burn, J.E., Hocart, C.H., Birch, R.J., Cork, A.C., and Williamson, R.E. (2002). Functional analysis of the cellulose synthase genes CesA1, CesA2, and CesA3 in Arabidopsis. Plant Physiol *129*, 797-807.

Carithers, L.J., Ardlie, K., Barcus, M., Branton, P.A., Britton, A., Buia, S.A., Compton, C.C., DeLuca, D.S., Peter-Demchok, J., Gelfand, E.T.*, et al.* (2015). A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. Biopreserv Biobank *13*, 311-319.

Cheneby, J., Menetrier, Z., Mestdagh, M., Rosnet, T., Douida, A., Rhalloussi, W., Bergon, A., Lopez, F., and Ballester, B. (2020). ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. Nucleic Acids Res *48*, D180-D188.

Cheng, C.Y., Krishnakumar, V., Chan, A.P., Thibaud-Nissen, F., Schobel, S., and Town, C.D. (2017). Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant J *89*, 789-804.

Chien, C.H., Chow, C.N., Wu, N.Y., Chiang-Hsieh, Y.F., Hou, P.F., and Chang, W.C. (2015). EXPath: a database of comparative expression analysis inferring metabolic pathways for plants. BMC Genomics *16 Suppl 2*, S6.

Coleman, T.F., and Moré, J.J. (1983). Estimation of Sparse Jacobian Matrices and Graph Coloring Problems. SIAM J Math Anal *20*, 187-209.

Consortium, E.P. (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol *9*, e1001046.

Cooper, L., Meier, A., Laporte, M.A., Elser, J.L., Mungall, C., Sinn, B.T., Cavaliere, D., Carbon, S., Dunn, N.A., Smith, B.*, et al.* (2018). The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. Nucleic Acids Res *46*, D1168-D1180.

Craigon, D.J., James, N., Okyere, J., Higgins, J., Jotham, J., and May, S. (2004). NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. Nucleic Acids Res *32*, D575-577.

Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., Jones, E.G., Bunney, W.E., Myers, R.M., Speed, T.P., Akil, H.*, et al.* (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res *33*, e175.

Dalma-Weiszhausz, D.D., Warrington, J., Tanimoto, E.Y., and Miyada, C.G. (2006). The affymetrix GeneChip platform: an overview. Methods Enzymol *410*, 3-28.

Daras, G., Rigas, S., Alatzas, A., Samiotaki, M., Chatzopoulos, D., Tsitsekian, D., Papadaki, V., Templalexis, D., Banilas, G., Athanasiadou, A.M.*, et al.* (2019). LEFKOTHEA Regulates Nuclear and Chloroplast mRNA Splicing in Plants. Dev Cell *50*, 767-779 e767.

Daras, G., Rigas, S., Penning, B., Milioni, D., McCann, M.C., Carpita, N.C., Fasseas, C., and Hatzopoulos, P. (2009). The thanatos mutation in Arabidopsis thaliana cellulose synthase 3 (AtCesA3) has a dominant-negative effect on cellulose synthesis and plant growth. New Phytol *184*, 114-126.

DeLuca, D.S., Levin, J.Z., Sivachenko, A., Fennell, T., Nazaire, M.D., Williams, C., Reich, M., Winckler, W., and Getz, G. (2012). RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics *28*, 1530-1532.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Dziuda, D.M. (2010). Basic Analysis of Gene Expression Microarray Data. In Data Mining for Genomics and Proteomics (Hoboken: John Wiley & Sons, Inc.), pp. 17-93.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A.*, et al.* (2019). The Pfam protein families database in 2019. Nucleic Acids Res *47*, D427-D432.

Farinas, B., and Mas, P. (2011). Histone acetylation and the circadian clock: a role for the MYB transcription factor RVE8/LCL5. Plant Signal Behav *6*, 541-543.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution *17*, 368-376.

Felsenstein, J. (2008). Distance matrix programs.

Fitch, W.M., and Margoliash, E. (1967). Construction of phylogenetic trees. Science *155*, 279-284.

Frankish, A., Diekhans, M., Ferreira, A.M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J.*, et al.* (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res *47*, D766-D773.

Franz, M., Rodriguez, H., Lopes, C., Zuberi, K., Montojo, J., Bader, G.D., and Morris, Q. (2018). GeneMANIA update 2018. Nucleic Acids Res *46*, W60-W64.

Gaillard, I., Rouquier, S., and Giorgi, D. (2004). Olfactory receptors. Cell Mol Life Sci *61*, 456-469.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J.*, et al.* (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol *5*, R80.

GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet *45*, 580-585.

Hammani, K., Okuda, K., Tanz, S.K., Chateigner-Boutin, A.L., Shikanai, T., and Small, I. (2009). A study of new Arabidopsis chloroplast RNA editing mutants reveals general features of editing factors and their target sites. Plant Cell *21*, 3686-3699.

Hossain, M.A., Noh, H.N., Kim, K.I., Koh, E.J., Wi, S.G., Bae, H.J., Lee, H., and Hong, S.W. (2010). Mutation of the chitinase-like protein-encoding AtCTL2 gene enhances lignin accumulation in dark-grown Arabidopsis seedlings. J Plant Physiol *167*, 650-658.

Hruz, T., Laule, O., Szabo, G., Wessendorp, F., Bleuler, S., Oertle, L., Widmayer, P., Gruissem, W., and Zimmermann, P. (2008). Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. Adv Bioinformatics *2008*, 420747.

Hubbell, E., Liu, W.M., and Mei, R. (2002). Robust estimators for expression analysis. Bioinformatics *18*, 1585-1592.

Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T.*, et al.* (2015). Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods *12*, 115-121.

Huson, D.H., and Scornavacca, C. (2012). Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol *61*, 1061-1067.

Hutter, C., and Zenklusen, J.C. (2018). The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. Cell *173*, 283-285.

Illumina (2010). TruSeq™ RNA Sample Preparation Guide.

Illumina (2014). TruSeq™ RNA and DNA Library Preparation Kits v2.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., and Speed, T.P. (2003a). Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res *31*, e15.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003b). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics *4*, 249-264.

Jen, C.H., Manfield, I.W., Michalopoulos, I., Pinney, J.W., Willats, W.G., Gilmartin, P.M., and Westhead, D.R. (2006). The Arabidopsis co-expression tool (ACT): a WWW-based tool and database for microarray-based gene expression analysis. Plant J *46*, 336-348.

Jin, J.M., Bai, P., He, W., Wu, F., Liu, X.F., Han, D.M., Liu, S., and Yang, J.K. (2020). Gender Differences in Patients With COVID-19: Focus on Severity and Mortality. Front Public Health *8*, 152.

Junier, T., and Zdobnov, E.M. (2010). The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics *26*, 1669-1670.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res *28*, 27-30.

Kassambara, A. (2017). Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning (STHDA).

Khan, T.A., Mazhar, H., Saleha, S., Tipu, H.N., Muhammad, N., and Abbas, M.N. (2016). Interferon-Gamma Improves Macrophages Function against M. tuberculosis in Multidrug-Resistant Tuberculosis Patients. Chemother Res Pract *2016*, 7295390.

Kinsella, R.J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A*., et al.* (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database (Oxford) *2011*, bar030.

Kodama, Y., Shumway, M., Leinonen, R., and International Nucleotide Sequence Database, C. (2012). The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res *40*, D54-56.

Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y.A., Williams, E., Dylag, M., Kurbatova, N., Brandizi, M., Burdett, T*., et al.* (2015). ArrayExpress update--simplifying data submissions. Nucleic Acids Res *43*, D1113-1116.

Krishnakumar, V., Hanlon, M.R., Contrino, S., Ferlanti, E.S., Karamycheva, S., Kim, M., Rosen, B.D., Cheng, C.Y., Moreira, W., Mock, S.A*., et al.* (2015). Araport: the Arabidopsis information portal. Nucleic Acids Res *43*, D1003-1009.

Kropshofer, H., Hammerling, G.J., and Vogt, A.B. (1999). The impact of the non-classical MHC proteins HLA-DM and HLA-DO on loading of MHC class II molecules. Immunol Rev *172*, 267-278.

Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A*., et al.* (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res *44*, W90-97.

Lachmann, A., Torre, D., Keenan, A.B., Jagodnik, K.M., Lee, H.J., Wang, L., Silverstein, M.C., and Ma'ayan, A. (2018). Massive mining of publicly available RNA-seq data from human and mouse. Nat Commun *9*, 1366.

Letunic, I., and Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res *47*, W256-W259.

Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A., and Dewey, C.N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics *26*, 493-500.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv *1303*.

Lim, W.K., Wang, K., Lefebvre, C., and Califano, A. (2007). Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. Bioinformatics *23*, i282-288.

Lipshutz, R.J., Fodor, S.P., Gingeras, T.R., and Lockhart, D.J. (1999). High density synthetic oligonucleotide arrays. Nat Genet *21*, 20-24.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550.

Lu, S.X., Knowles, S.M., Andronis, C., Ong, M.S., and Tobin, E.M. (2009). CIRCADIAN CLOCK ASSOCIATED1 and LATE ELONGATED HYPOCOTYL function synergistically in the circadian clock of Arabidopsis. Plant Physiol *150*, 834-843.

Lurin, C., Andres, C., Aubourg, S., Bellaoui, M., Bitton, F., Bruyere, C., Caboche, M., Debast, C., Gualberto, J., Hoffmann, B.*, et al.* (2004). Genome-wide analysis of Arabidopsis pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. Plant Cell *16*, 2089-2103.

McCall, M.N., Bolstad, B.M., and Irizarry, R.A. (2010). Frozen robust multiarray analysis (fRMA). Biostatistics *11*, 242-253.

McKusick, V.A. (2007). Mendelian Inheritance in Man and its online version, OMIM. Am J Hum Genet *80*, 588-604.

Michalopoulos, I., Pavlopoulos, G.A., Malatras, A., Karelas, A., Kostadima, M.A., Schneider, R., and Kossida, S. (2012). Human gene correlation analysis (HGCA): a tool for the identification of transcriptionally co-expressed genes. BMC Res Notes *5*, 265.

Milioni, D., and Hatzopoulos, P. (1997). Genomic organization of hsp90 gene family in Arabidopsis. Plant Mol Biol *35*, 955-961.

Miller CJ (2018). simpleaffy: Very simple high level analysis of Affymetrix data.

Mollica, V., Rizzo, A., and Massari, F. (2020). The pivotal role of TMPRSS2 in coronavirus disease 2019 and prostate cancer. Future Oncol.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods *5*, 621-628.

Murphy, B.J., Kimura, T., Sato, B.G., Shi, Y., and Andrews, G.K. (2008). Metallothionein induction by hypoxia involves cooperative interactions between metal-

responsive transcription factor-1 and hypoxia-inducible transcription factor-1alpha. Mol Cancer Res *6*, 483-490.

Najjar, I., and Fagard, R. (2010). STAT1 and pathogens, not a friendly relationship. Biochimie *92*, 425-444.

O'Malley, R.C., Huang, S.C., Song, L., Lewsey, M.G., Bartlett, A., Nery, J.R., Galli, M., Gallavotti, A., and Ecker, J.R. (2016). Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell *165*, 1280-1292.

Obayashi, T., Aoki, Y., Tadaka, S., Kagaya, Y., and Kinoshita, K. (2018). ATTED-II in 2018: A Plant Coexpression Database Based on Investigation of the Statistical Property of the Mutual Rank Index. Plant Cell Physiol *59*, e3.

Oeckinghaus, A., and Ghosh, S. (2009). The NF-kappaB family of transcription factors and its regulation. Cold Spring Harb Perspect Biol *1*, a000034.

Okamura, Y., Aoki, Y., Obayashi, T., Tadaka, S., Ito, S., Narise, T., and Kinoshita, K. (2015). COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems. Nucleic Acids Res *43*, D82-86.

Okuda, K., Chateigner-Boutin, A.L., Nakamura, T., Delannoy, E., Sugita, M., Myouga, F., Motohashi, R., Shinozaki, K., Small, I., and Shikanai, T. (2009). Pentatricopeptide repeat proteins with the DYW motif have distinct molecular functions in RNA editing and RNA cleavage in Arabidopsis chloroplasts. Plant Cell *21*, 146-156.

Parman, C., Halling, C., and Gentleman, R. (2018). affyQCReport: QC Report Generation for affyBatch objects. R package version 1580.

Parrine, D., Wu, B.S., Muhammad, B., Rivera, K., Pappin, D., Zhao, X., and Lefsrud, M. (2018). Proteome modifications on tomato under extreme high light induced-stress. Proteome Sci *16*, 20.

Paulson, J.N., Chen, C.Y., Lopes-Ramos, C.M., Kuijjer, M.L., Platig, J., Sonawane, A.R., Fagny, M., Glass, K., and Quackenbush, J. (2017). Tissue-aware RNA-Seq processing and normalization for heterogeneous and sparse data. BMC Bioinformatics *18*, 437.

Pearce, S., Ferguson, A., King, J., and Wilson, Z.A. (2015). FlowerNet: a gene expression correlation network for anther and pollen development. Plant Physiol *167*, 1717-1730.

Pearson, K. (1895). VII. Note on regression and inheritance in the case of two parents. Proc R Soc Lond *58*, 240-242.

Perales, M., and Mas, P. (2007). A functional link between rhythmic changes in chromatin structure and the Arabidopsis biological clock. Plant Cell *19*, 2111-2123.

Piccolo, S.R., Sun, Y., Campbell, J.D., Lenburg, M.E., Bild, A.H., and Johnson, W.E. (2012). A single-sample microarray normalization method to facilitate personalized-medicine workflows. Genomics *100*, 337-344.

Pinero, J., Ramirez-Anguita, J.M., Sauch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., and Furlong, L.I. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res *48*, D845-D855.

Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A., Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D.*, et al.* (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv, 201178.

Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., and Vilo, J. (2016). g:Profiler-a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res *44*, W83-89.

Rew, R., and Davis, G. (1990). NetCDF: an interface for scientific data access. IEEE Computer Graphics and Applications *10*, 76-82.

Rodchenkov, I., Babur, O., Luna, A., Aksoy, B.A., Wong, J.V., Fong, D., Franz, M., Siper, M.C., Cheung, M., Wrana, M.*, et al.* (2020). Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. Nucleic Acids Res *48*, D489-D497.

Rouillard, A.D., Gundersen, G.W., Fernandez, N.F., Wang, Z., Monteiro, C.D., McDermott, M.G., and Ma'ayan, A. (2016). The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database (Oxford) *2016*.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol *4*, 406-425.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*, 467-470.

Schlapfer, P., Zhang, P., Wang, C., Kim, T., Banf, M., Chae, L., Dreher, K., Chavali, A.K., Nilo-Poyanco, R., Bernard, T.*, et al.* (2017). Genome-Wide Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants. Plant Physiol *173*, 2041-2059.

Schliep, K., Potts, A.J., Morrison, D.A., and Grimm, G.W. (2017). Intertwining phylogenetic trees and networks. Methods in Ecology and Evolution *8*, 1212-1220.

Schrumpfova, P.P., Vychodilova, I., Hapala, J., Schorova, S., Dvoracek, V., and Fajkus, J. (2016). Telomere binding protein TRB1 is associated with promoters of translation machinery genes in vivo. Plant Mol Biol *90*, 189-206.

Sokal, R., and Rohlf, F. (1962). The comparison of dendrograms by objective methods. Taxon *11*, 33-40.

Sokal, R.R., and Michener, C.D. (1958). A Statistical Methods for Evaluating Systematic Relationships. Univ Kansas Sci Bull *38*, 1409-1438.

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T.I., Nudel, R., Lieder, I., Mazor, Y.*, et al.* (2016). The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Curr Protoc Bioinformatics *54*, 1 30 31-31 30 33.

Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P.*, et al.* (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res *47*, D607-D613.

Taylor-Teeples, M., Lin, L., de Lucas, M., Turco, G., Toal, T.W., Gaudinier, A., Young, N.F., Trabucco, G.M., Veling, M.T., Lamothe, R.*, et al.* (2015). An Arabidopsis gene regulatory network for secondary cell wall synthesis. Nature *517*, 571-575.

The Gene Ontology, C. (2019). The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res *47*, D330-D338.

Tonsor, S.J., Scott, C., Boumaza, I., Liss, T.R., Brodsky, J.L., and Vierling, E. (2008). Heat shock protein 101 effects in A. thaliana: genetic variation, fitness and pleiotropy in controlled temperature conditions. Mol Ecol *17*, 1614-1626.

Tryka, K.A., Hao, L., Sturcke, A., Jin, Y., Wang, Z.Y., Ziyabari, L., Lee, M., Popova, N., Sharopova, N., Kimura, M.*, et al.* (2014). NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic Acids Res *42*, D975-979.

van Dam, S., Craig, T., and de Magalhaes, J.P. (2015). GeneFriends: a human RNA-seq-based gene and transcript co-expression database. Nucleic Acids Res *43*, D1124-1132.

Wang, Y., and Hua, J. (2009). A moderate decrease in temperature induces COR15a expression through the CBF signaling cascade and enhances freezing tolerance. Plant J *60*, 340-349.

Weissenbach, M., Clahsen, T., Weber, C., Spitzer, D., Wirth, D., Vestweber, D., Heinrich, P.C., and Schaper, F. (2004). Interleukin-6 is a direct mediator of T cell migration. Eur J Immunol *34*, 2895-2906.

Wightman, R., and Turner, S. (2010). Trafficking of the plant cellulose synthase complex. Plant Physiol *153*, 427-432.

Wu, Z., Irizarry, R.A., Gentleman, R., Martinez-Murillo, F., and Spencer, F. (2004). A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. Journal of the American Statistical Association *99*, 909-917.

Xu, J., Yang, C., Yuan, Z., Zhang, D., Gondwe, M.Y., Ding, Z., Liang, W., Zhang, D., and Wilson, Z.A. (2010). The ABORTED MICROSPORES regulatory network is required for postmeiotic male reproductive development in Arabidopsis thaliana. Plant Cell *22*, 91-107.

Yamaguchi, M., Mitsuda, N., Ohtani, M., Ohme-Takagi, M., Kato, K., and Demura, T. (2011). VASCULAR-RELATED NAC-DOMAIN7 directly regulates the expression of a broad range of genes for xylem vessel formation. Plant J *66*, 579-590.

Yilmaz, A., Mejia-Guerra, M.K., Kurz, K., Liang, X., Welch, L., and Grotewold, E. (2011). AGRIS: the Arabidopsis Gene Regulatory Information Server, an update. Nucleic Acids Res *39*, D1118-1122.

Zaag, R., Tamby, J.P., Guichard, C., Tariq, Z., Rigaill, G., Delannoy, E., Renou, J.P., Balzergue, S., Mary-Huard, T., Aubourg, S.*, et al.* (2015). GEM2Net: from gene expression modeling to -omics networks, a new CATdb module to investigate Arabidopsis thaliana genes involved in stress response. Nucleic Acids Res *43*, D1010-1017.

Zhang, D., Hrmova, M., Wan, C.H., Wu, C., Balzen, J., Cai, W., Wang, J., Densmore, L.D., Fincher, G.B., Zhang, H.*, et al.* (2004). Members of a new group of chitinase-like genes are expressed preferentially in cotton cells with secondary walls. Plant Mol Biol *54*, 353-372.

Zhu, Q., Wong, A.K., Krishnan, A., Aure, M.R., Tadych, A., Zhang, R., Corney, D.C., Greene, C.S., Bongo, L.A., Kristensen, V.N.*, et al.* (2015). Targeted exploration and analysis of large cross-platform human transcriptomic compendia. Nat Methods *12*, 211-214, 213 p following 214.

Zuberi, K., Franz, M., Rodriguez, H., Montojo, J., Lopes, C.T., Bader, G.D., and Morris, Q. (2013). GeneMANIA prediction server 2013 update. Nucleic Acids Res *41*, W115-122.